INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2022IPPAX137

Thèse de doctorat

Study of vector boson scattering in events with four leptons and two jets with the CMS detector at the LHC

Thèse de doctorat de l'Institut Polytechnique de Paris et de l'Université de Zagreb
préparée à l'Ecole Polytechnique

École doctorale de l'Institut Polytechnique de Paris (ED IP Paris) n°626
Spécialité de doctorat: Physique des particules

Thèse présentée et soutenue à Palaiseau, le 9 Decembre 2022, par

GILJANOVIĆ DUJE

Composition du Jury :

Isabelle Wingerter-Seez Directrice de recherche, CPPM	Rapporteur
Vuko Brigljević Professeur des universités, Rudjer Boskovic Institute	Rapporteur
Olivier Drapier Directeur de recherche, École polytechnique (LLR)	Président du Jury
Riccardo Bellan Professeur des universités, Università degli Studi di Torino	Examinateur
Krešimir Kumerički Professeur des universités, University of Zagreb, PMF	Examinateur
Mathieu Pellen Chargée de recherche, University Freiburg	Examinateur
Claude Charlot Directeur de recherche, École polytechnique (LLR)	Co-directeur de thèse
Damir Lelas Professeur assistant, University of Split, FESB	Co-directeur de thèse

University of Zagreb

GILJANOVIĆ DUJE

Study of vector boson scattering in events with four leptons and two jets with the CMS detector at the LHC

International dual doctorate

Examinator	Isabelle Wingerter-Seez
Examinator	Brigljević Vuko
President of the Jury	Olivier Drappier
Referee	Riccardo Bellan
Referee	Krešimir Kumericki
Referee	Mathieu Pellen
Thesis supervisor	Claude Charlot
Thesis co-supervisor	Damir Lelas

Palaiseau, France, 2022

University of Zagreb

GILJANOVIĆ DUJE

Recherche de la diffusion de boson de jauge dans les événements avec le détecteur CMS auprès du LHC

Cotutelle internationale

Examinateur	Isabelle Wingerter-Seez
Examinateur	Brigljević Vuko
Président du Jury	Olivier Drappier
Rapporteur	Riccardo Bellan
Rapporteur	Krešimir Kumericki
Rapporteur	Mathieu Pellen
Directeur de thèse	Claude Charlot
Co-directeur de thèse	Damir Lelas

Palaiseau, France, 2022

University of Zagreb

GILJANOVIĆ DUJE

Proučavanje raspršenja vektorskih bozona u događajima s četiri leptona i dva mlaza koristeći detektor CMS na LHC-u

Međunarodni dvojni doktorat

Ocjenjivač	Isabelle Wingerter-Seez
Ocjenjivač	Brigljević Vuko
Predsjednik Povjerenstva	Olivier Drappier
Povjerenstvo	Riccardo Bellan
Povjerenstvo	Krešimir Kumericki
Povjerenstvo	Mathieu Pellen
Mentor	Claude Charlot
Komentor	Damir Lelas

Palaiseau, Francuska, 2022

Acknowledgments

Only now when the time has come to list all of you who have helped me reach this monumental achievement do I see how truly lucky I was on this journey. Please forgive me if you don't find your name explicitly written. I thank you for being a part of this personal milestone!

First and foremost, I couldn't have done this without my mentors Claude and Damir. I thank you sincerely for your patience, guidance, trust and understanding you showed me. I thank you for helping me with your expertise and knowledge. I thank you for supporting me and guiding me until the very end. To list everything you did for me I would need to write another dissertation just for that purpose. Thank you Claude for taking care of everything during my LLR visits: from doing paperwork with me to bringing me bed sheets and pillows. Tank you Damir for being the first "line of defence" and talking with me during my lowest moments. Thank you both for many fruitful discussions that we had.

I thank my LLR colleagues for helping me fit in and making me feel welcome. I thank all of you. Special thanks, however, goes to my LLR officemates: Jonas, Matteo and Chiara. Without you taking care of my apartment keys, Jonas, I would have literally slept on the streets the first day I arrived. Without you, Matteo and Chiara, our PhD launches wouldn't have been nearly as fun. I thank you Jacques for your unwavering kindness.
I also thank my friends and colleagues at FESB for making this journey easier. Thank you Ivica for introducing me to the field of experimental particle physics, first through my Bachelor's thesis, and then through my Master's thesis. I thank Darko, Dunja, Ilja, Nikola, Suri, Bojan and Stipe for helping with my teaching obligations. Thank you Marko for helping me with trigger shifts and introducing me to the world of CMSSW. Thank you Toni (Šarić) for being a fantastic colleague. Thank you Marina for our many talks and for helping me fill so many documents and forms. Special thanks go to team Šćulac -Toni and Ana, with whom I shared my office, tasks, concerns and dilemmas.
An entire paragraph would be needed to even starting thanking you, teta Anita. You have been my friend, psychologist, psychiatrist, a go-to person for absolutely any problem that comes to mind. I don't see how I could have reached this point without your constant help with every-day tasks, documents, forms, apartments, finances and everything else. In the end, I really did need an entire paragraph just to start thanking you. It is still no enough, I know. It will never be.
I also have to thank Marko Hum for helping me with all formal steps. Thank you for your patience and promptness.

I will use this oportunity to thank you, Tamara and Roje. You have been by my side since the high school days. You know my every victory and defeat, all my problems, struggles, doubts and plans. You have been there for me whenever I needed you and you have shown an extraordinary amount of patience with me for more than a decade. Thank you for everything!
I was blessed to have many friends by my side during all these years. For brevity sake, I will just mention those of you who I have known the longest. Thank you Roga, Bela, Mršo, Ciki, Vrci, Marina and Kiki. No, I didn't forget about you. I just have to finally stop writing. I thank you as well!
I thank Tomislav and Zoran for letting me write a good portion of this manuscript on one of the most beautiful places

in Croatia - Mosor observatory. You have made me feel welcome and like at home. Thank you Zvjezdano selo Mosor! I also thank you, Fanica Barbaroša, for being the best high school physics teacher ever. Without you, I doubt I would ever have enough courage to pursuit a physics degree, let alone a PhD.

I cannot think of a person in my life that has been a source of happiness as bright as you Anja, and I am lucky to have found you in the best possible moment. Thank you for always being there for me, for never doubting me and for always encouraging me. This is not my success alone - it is yours as well!
Finally, the most important thank you goes to my family. You have been my support for an eternity, and you have shown nothing but patience and love for me. It was never hard for you to listen to my problems and help me to solve them. You have shared my happiness, sadness, struggles and all my achievements. Without you, there is no me, there is no physics, and there is no PhD. It all starts with you, and it is only fair that my acknowledgments end with you. Thank you for everything! I love you more than you know!

Abstract

Studying Vector Boson Scattering is crucial for understanding the electroweak symmetry breaking mechanism and it provides a complementary tool for measuring Higgs boson couplings to vector bosons. In addition, using the effective field theory (EFT) framework, one can probe the Beyond Standard Model physics through modifications of certain quartic gauge couplings. This thesis reports the first evidence, with the CMS detector, of electroweak (EW) production of leptonically decaying Z boson pair accompanied by two hadronic jets with a vector boson scattering topology. The study analyses $137 fb^{-1}$ of proton-proton collisions produced at CERN Large Hadron Collider (LHC) at 13 TeV centre-of-mass energy. Additionally, a prospective study is presented on the longitudinal scattering in the same channel at High-Luminosity (HL-LHC) and High-Energy LHC (HE-LHC) conditions, corresponding to 14 and 27 TeV centre-of-mass energy, respectively, with full event kinematics simulated.

Although this channel is characterised by a fully reconstructable final state, the small cross section of EW signal compared to the QCD-induced background makes it challenging to measure. Efficient identification of final state leptons is essential since efficiencies on their measurement enter the analysis with a power of four. Measurement of electron selection efficiencies and derivation of scale factors for 2016, 2017 and 2018 data-taking periods was performed. Electron identification is done at CMS using the multivariate approach with a multivariate classifier retrained, for all three periods, using the ExtremeGradient Boost software and with electron isolation included in the training. Uncertainties on both electron selection efficiencies and scale factors were reduced across the p_T spectrum with special care towards reducing the uncertainties at low p_T.

The EW signal was extracted at 13 TeV using the Matrix Element Likelihood Approach (MELA) and the performance was cross-checked with the boosted decision tree (BDT) classifier. The EW production of two jets in association with two Z bosons was measured with an observed (expected) significance of 4.0 (3.5) standard deviations. The cross sections for the EW production were measured in three fiducial volumes and is $0.33^{(+0.11)}_{(-0.10)}(stat)^{(+0.04)}_{(-0.03)}(syst)\ fb$ in the most inclusive volume, in agreement with the Standard Model (SM) prediction of $0.275 \pm 0.021\ fb$. Limits on the anomalous quartic gauge couplings were derived in terms of EFT operators T0, T1, T2, T8, and T9.

The extraction of the longitudinal component of the Z bosons at the HL- and HE-LHC was performed using two multivariate approaches. A combined-background BDT was trained to separate the $Z_L Z_L$ signal from the mixture of $Z_L Z_T$, $Z_T Z_T$ and QCD-induced backgrounds. In addition, a more complex approach, referred to as the 2D BDT, was designed to increase signal sensitivity. Two BDTs were trained simultaneously to separate the $Z_L Z_L$ signal from the QCD-induced backgrounds and the $Z_L Z_L$ signal from the mixture of $Z_L Z_T$ and $Z_T Z_T$ backgrounds. The effect on signal significance when increasing electron acceptance from $|\eta| = 3$ to $|\eta| = 4$ was studied as well. With an increased electron acceptance, the longitudinal component is expected to be measured with a significance of 1.4 standard deviations at 14 TeV and with an integrated luminosity of $3000 fb^{-1}$. A measurement of the longitudinal scattering in the ZZ channel is expected at 27 TeV, corresponding to an integrated luminosity of $15000 fb^{-1}$, with

a signal significance of 4.6 standard deviations. With the extended electron acceptance, the first observation is expected with a significance of 5.4 standard deviations. Hence, this study demonstrates a significant benefit of further energy increase at the LHC for understanding the EW sector of the SM.

Résumé

La diffusion de bosons de jauges (VBS) constitue un moyen fondamental pour étudier le mécanisme de brisure spontannée de la symmétrie électrofaible (EW), à travers laquelle les bosons de jauges faibles acquièrent une masse dans le modèle standard. Cette thèse présente la première indication pour la production EW d'une paire de bosons Z, se désintégrant ensjuite en qautre leptons (électrons ou muons), accompagné par deux jets hadroniques. L'analyse porte sur 137 fb^{-1} de collisions proton-proton collectées par le détecteur CMS au LHC à une énergie de 13 TeV dans le centre de masse. De plus, les premiéres études prospectives pour la diffusion longitudinale dans ce canal pour la phase de haute luminosité (HL-LHC) et de haute énergie (HE-LHC), correspondant à des énergies dans le centre de masse de 14 TeV et 27 TeV, respectivement, sont présentées avec une simulation complète de la cinématique des événements.

Introduction au modèle standard et à la diffusion des bosons de jauges

Le modèle standard est, à l'heure actuelle, la théorie la plus complète pour décrire les particules élémentaires et leurs intéractions. C'est une théorie des champs quantiques relativistes reposant sur le groupe de symmétrie $SU(3)_C \times SU(2)_L \times U(1)_Y$, où le premier terme décrit la symmétrie dees intéractions fortes (QCD), le second celle des intéractions faibles, et le dernier celle des intéractions électromagnétiques.
Une des classifications possibles des particules du modèle standard est basée sur la propriété quantique appelée *spin*. De cette façon on distingue les particules de spin demi-entier, les *fermions*, des particules de spin entier, les *bosons*. Toute la matière dans l'univers est constituée de fermions, alors que les bosons véhiculent les intéractions entre eux. Les fermions sont regroupés en trois classes de quarks et de leptons. La première classe de leptons inclue l'*electron* (e) et l'*electron neutrino* (ν_e), alors que la premiére classe de quarks est constituée du *up quark* (u) et du *down quark* (d). Les deux autres classes sont composées de copies plus massives: les leptons μ, ν_μ, τ, et ν_τ et les quarks c, s, t, and b. En plus des fermions, le modèle standard décrit les anti-fermions qui ont des nombres quantiques opposés. Contrairement aux leptons, les quarks ne sont pas détectables individuellement, mais toujours regroupés en paires quark-antiquark (nommées "*mésons*") ou en triplet de quarks (nommés "*baryons*"). Les baryons et les mésons sont collectivement dénommés *hadrons*.
Un atome est un état lié de nucléons et d'électrons, où la force électromagnétique, transportée par une particule de spin 1 appeléee photon, joue un rôle crucial. Les nucléons du noyau de l'atome sont liés par la force forte, transmise par huit gluons, qui ont aussi un spin 1. Finalement, la force faible, responsable de la désintégration β est médiée par trois bosons de jauges: $W^\pm$ et Z^0.

Avant 1964, il y avait un désaccord entre les prédictions du modéle standard, qui suggére que les bosons de jauges sont sans masse, et les expériences qui indiquaient le contraire. Une façon de résoudre ce problème a été proposée en 1964 et confirmée par la découverte de boson de Higgs au CERN en 2012. C'est le mécanisme de Brout-Englert-Higgs (BEH), qui est basé sur l'introduction d'un doublet de champs complexes scalaires ayant

une valeur moyenne non nulle dans le vide. Il s'ensuit une brisure spontannée de la symmétrie, qui conduisent ía présence de quatre bosons de Goldstone. De façon a préserver la symmétrie locale SU(2), les champs des bosons de Goldstone se combinent avec les champs des bosons $W^{\pm}$ et Z^0 sans masse, résultant en la génération de masses pour ces bosons.
Contrairement au bosons W et Z sans masse, qui ont seulement une composante de polarisation transverse, les bosons W et Z massifs acquièrent un degré de liberté de polarisation supplémentaire: une composante longitudinale de polarisation. La différence importante de comportement des bosons W et Z polarisé longitudinalement en comparaison des bosons polarisés transversalement se manifeste dans le comportement divergent de la section efficace de diffusion à haute énergie pour les bosons longitudinaux. Le boson de Higgs joue un rôle crucial ici, avec ses couplages aux bosons vecteurs qui permettent de restaurer l'unitarité. L'étude de la diffusion des bosons vecteurs longitudinaux constitue donc un outil supplémentaire pour étudier les propriétés du boson de Higgs et le mécanisme de brisure spontannée de la symmétrie EW. Elle permet également détudier la structure non-abélienne des intéractions EW à travers l'étude des couplages quartiques. Enfin, des indications de physique au-delà du modèle standard peuvent se manifester par des modifications de certains couplages quartiques.

Le grand collisioneur de hadrons et l'expérience CMS

Le grand collisioneur de hadrons (LHC) est le plus grand accélérateur du monde, avec une circonférence de 27 km, et est géré par l'organisation européenne pour la recherche en physique des particules (CERN). Situé à la frontière franco-suisse, il constitue létape finale de l'accélération des protons, permettant d'atteindre une énergie de 13 TeV dans le centre de masse des collisions proton-proton. Dans le tunnel du LHC, deux faisceaux de protons circulent et se croisent à des endroits bien déterminés. Pour augmenter le taux de collisions, les protons sont regroupés en paquets au nombre de 2000 environ dans chaque faisceau qui circule dans la machine en permanence. Chaque paquet de protons contient approximativement 10^{11} protons. En plus des deux détecteurs généralistes CMS et ATLAS, le LHC héberge également plusieurs autres détecteurs pour étudier la nouvelle physique, incluant LHCb, ALICE, TOTEM, MoEDAL, and LHCf.
Les données utilisées dans cette thèse ont été collectée par le détecteur CMS situé près du village de Cessy, à environ 100 m de profondeur. Le détecteur fait 21 m de long et 15 m de haut et de large, et est conçu autour d'un aimant solénoidal supraconducteur qui fournit un champ magnétique de 4 T. CMS consiste en plusieurs sous-détecteurs conçu pour des tâches spécifiques. Le sous-détecteur le plus proche du point où les faisceaux de protons entrent en collisions (i.e. le point d'intéraction) est le *détecteur de traces*, qui mesure l'impulsion des particules chargées par la courbure de leur trajectoire dans le champ magnétique. Ensuite vient le calorimètre électromagnétique (ECAL), qui mesure l'énergie des électrons et des photons. Le matériau actif du ECAL est constitué par des cristaux transparents de PbWO4. Lorsque les électrons ou les photons traversent les cristaux du ECAL, ils produisent des gerbes d'électrons et de photons de plus basse énergie et les électrons produits perdent leur énergie en produisant des petits flashs de lumiére appelée scintillation. ces flashs de lumière sont collectés par des photodétecteurs situés à l'arrière des cristaux, et convertis en un signal électrique, qui est transmis pour traitement. Après le ECAL se trouve le calorimètre hadronique (HCAL), qui mesure l'énergie des hadrons et fonctionne avec le détecteur de trace et le ECAL pour mesurer l'énergie des hadrons. Enfin le sous-détecteur le plus éloigné est constitué des chambres à muons, qui permettent d'identifier les muons et de mesurer leur impulsion.
Toutes les 25 ns, en moyenne, il y a une collision de proton au LHC qui génère environ 50 TB de données par seconde. Comme il n'est pas possible d'enregistrer un si gros volume de données, une grande fraction de ces événements ne sera pas enregistrée. Pour éviter d'éliminer des événements interessants, un système de déclenchement a été développé, dont le rôle est de préserver environ 1-2 kHz des 40 MHz initiaux. Les événements qui sont sauvegardés sont processés et les objets physiques sont reconstruits. Des algorithmes complexes ont été développés à cette

fin, et, en utilisant tous les sous-détecteurs, ils reconstruisent les électrons, muons, jets hadroniques, et l'énergie transverse manquante qui indique la présence de neutrinos. Les analyses qui les utilisent doivent utiliser des critères additionnels (appelés sélection) pour minimiser les événements de bruits de fonds.
En 2018, le CERN a achevé la phase dite de Run 2 des opérations au LHC, pendant laquelle 140 fb^{-1} de données ont été collectées. Les préparations pour le Run 3 ont commencées depuis 2019 avec l'objectif de collecter environ 300 fb^{-1} à l'énergie de 13,6 TeV. Pour continuer les recherches en physique des particules élémentaires, une phase de haute luminosité (HL-LHC) est prévue après le Run 3, avec l'objectif de collecter 3000 fb^{-1} à l'énergie de 14 TeV. Le projet suivant du CERN est le futur collisionneur circulaire (FCC), un anneau de 100 km de diamètre avec une énergie de collision de 100 TeV. Entre le HL-LHC et le FCC, une phase de haute énergie du LHC (HE-LHC) est envisagée avec pour but de collecter environ 15000 fb^{-1} à l'énergie de 27 TeV dans le centre de masse.

Reconstruction des électron et mesure des efficacités

Comparée à la reconstruction des muons, la reconstruction des électrons est significativement plus complexe en raison du rayonnement de bremsstrahlung lors de la traversée du détecteur de traces. La première conséquence de ce rayonnement est une augmentation du nombre de cristaux, suivant la direction ϕ, dans lesquels les électrons déposent leur énergie. La seconde conséquence est l'augmentation de la courbure de la trajectoire des électrons. Pour tenir compte de cela, des algorithmes complexes ont été implémentés pour la reconstruction des électrons.
La première étape de la reconstruction des électrons consiste à collecter tous les dépôts d'énergie créés par le passage des électrons dans le milieu actif, ainsi que tous les dépôts d'énergie des photons émis, dans des aggrégats. Pour augmenter l'efficacité de reconstruction, les aggrégats sont regroupés en super-agrrégats. En parallèle, la reconstruction de la trace est effectuée. L'étape suivante est l'estimation de la charge de l'électron en utilisant la courbure de la trace et d'autres méthodes complémentaires. Bien que l'idée essentielle de la reconstruction des aggrégats est de collecter tous les dépôts d'énergie, cette procédure n'est en général pas efficace à 100%. L'énergie reconstruite est donc corrigée à l'aide de simulations et de techniques d'intelligence artificielle. Toute l'information recueillie est utilisées dans l'algorithme de particle-flow, qui classifie les objets reconstruits comme électrons ou photons.
La partie essentielle de ce travail consiste dans la sélection des électrons, c.à.d la mesure de l'efficacité de sélection des électrons. Pour réduire le bruit de fond (par ex. les électrons faussement reconstruits à partir des jets hadroniques, la sélectionc des électrons s'effectue en plusieurs étapes:

1. la selection des électrons primaires (c.à.d. les électrons qui ne sont pas originaires de conversions de photons, prompt ou venant de désintégrations de pions neutres)
2. identification des électrons
3. isolation des électrons

Depuis 2017, l'algorithme d'identification a été mis à jour et les variables utilisées pour l'isolation des électrons sont maintenant incluses dans l'algorithme d'identification. L'identification des électrons est basé sur un algorithme multivarié.
De façon à mesurer l'efficacité de sélection des électrons, une méthode dite de "tag and probe" a été développée, qui utilise les événements $Z \to l^+l^-$ pour obtenir un échantillon pur de candidats électrons. tout d'abord, un électron de signal est sélectionné dans chaque événement avec un critère strict (électron dit "tag"). Ensuite, dans le même événement, un électron (dit "probe") est recherché en utilisant des critères moins stricts de sélection dont on veut mesurer l'efficacité, qui, avec le "tag", correspond à la masse du boson Z. L'efficacité de la sélection est alors définie comme la rappport du nombre d'électrons "probe" qui passent la sélection au nombre total d'électrons "probe", Dans

le cas où le bruit de fond est faible, l'efficacité de sélection peut être obtenue par un simple comptage. Dans le cas contraire, un ajustement est effectué dans chaque bin de moment transverse momentum (p_T) et de pseudorapidité (η) pour évaluer le nombre d'événements du signal dans les distributions où l'électron "probe" passe la sélection et dans celles où l'électron ne passe pas la sélection testée. Le nombre d'événements signal où l'électron "probe" passe la sélection et où l'électron ne passe pas la sélection sont ensuite utilisés pour évaluer l'efficacité de la sélection. L'efficacité de sélection est mesurée à la fois dans les données et dans la simulation de fac on à corriger toutes les différences entre les données et la simulation, séparément pour chaque bin en p_T et η, par des facteurs de correction obtenus par ces mesures.
L'efficacité de reconstruction des électrons et les facteurs de corrections ont été mesurés tout d'abord pour les données de 2018 en utilisant l'algorithme d'identification entrainé sur les données de 2017, où les variables d'isolation étaient utilisées dans l'identification. Par la suite, l'identification des électrons pour 2016 a été réévaluée pour les données 2016 en entrainant l'algorithme sur les données de 2016. La même chose a été faite pour la période 2018. L'efficacité de sélection et les facteurs de correction ont été réévalués pour les trois années de fac on a réduire l'incertitude sur la mesure des électrons en particulier à bas moment transverse ($< 10GeV$), et également pour mieux comprendre la structure en η de l'efficacité de sélection.
Pour mieux réduire l'incertitude sur l'efficacité de sélection des électrons à bas p_T, des condtions plus stricts ont été appliquées sur le tag dans les événements où le probe est à bas p_T. Cela a résulté en un pic mieux visible dans les distributions où l'électron probe ne passe pas la sélection, ce qui a amélioré la précision pour les trois périodes. La sélection plus stricte du tag a pour conséquence un changement de la forme de la distribution dans la région de basse masse de m_{ee} pour les électrons probe qui ne passent pas la sélection. En étudiant ce phénomène, il a été compris que cela est dû à une migration d'électrons de signal entre les électrons qui passent et ceux qui ne passent pas le critère de sélection. Pour effectuer un ajustement correct, il a été nécessaire de modifier la fonction ajustée. En conséquence les incertitudes sur les mesures d'efficacité ont été encore réduite. En plus des conditions plus strictes sur l'électron tag, le nombre de bins en η et p_T a été augmenté pour mieux étudier la structure en η de l'efficacité de sélection. Une structure en forme de parapluie a été observée en conséquence de la moins bonne modélisation par la simulation dansd les régions à grand η. Ces résultats ont été utilisés dans l'analyse $H \to ZZ \to 4l$ utilisant toutes les données du Run 2 et les améliorations apportées sont maintenant utilisées de façon standard pour la mesure des efficiacité de sélection des électrons dans CMS.

Recherche du processus VBS dans le canal ZZ $\to$ 4l2j

Dans la recherche du processus VBS dans le canal $ZZ \to 4l2j$, les données de CMS de 2016 à 2018 ont été utilisées. Comme les particules de l'état final, les électrons, muons et jets hadroniques peuvent être complètement reconstruits dans CMS, ce canal sera probablement l'un des plus importants dans l'étude de la diffusion longitudinale à l'avenir, permettant de mieux comprendre la théorie des intéractions EW dans le modèle standard. De plus ce canal est sensible a certains opérateurs dans le formalisme de théorie effective (T0, T1, T8 et T9). et permet l'étude de couplages quartiques anormaux et donc l'étude d'une physique au-delà du modèle standard.

De façon à mieux décrire les processus du signal et du bruit de fonds, une attention particulière a été portée aux simulations par Monte Carlo. La sélection des leptons a été spécialement conçue pour réduire le taux de bruit de fond. Une sélection peu stricte a été imposée aux candidats leptons pour réduire le bruit de fond tout en maintenant une très grande efficacité pour le signal, un nombre suffisant d'événements et une signification statistique optimale. Les leptons qui satisfont aux critères de sélections sont utilisés pour former des bosons Z, en combinant les leptons de même saveur et de charge opposée (e^+e^-, $\mu^+\mu^-$). Chaque candidat Z doit satisfaire un emsemble de critères spécifiques pour réduire la probabilité de bruit de fond. Parmi les deux bosons Z reconstruits dans l'événement, le Z_1

est le boson Z boson de plus grand p_T, alors que le Z_2 est l'autre boson Z. Dans les événements avec plus de deux bosons Z possibles, la paire dont la somme scalaire des p_T des deux leptons est la plus grande est choisie pour le Z_2. Pour l'analyse statistique, une sélection (nommée sélection ZZjj inclusive) est défine en demandant, en plus des citères ci-dessus, que les deux jets ont une masse invariante plus grande que 100 GeV. Cela défini la sélection de base pour le signal dans l'analyse statistique. Des critères alternatifs (dénommés sélection VBS relachée et sélection VBS stricte) sont définis pour sélectionner un espace de phase moins strict ou plus strict.

Une approche par des éléments de matrices (MELA) a été utilisée pour évaluer la section efficace pour la production EW et EW+QCD $ZZ \to 4l2j$. La même méthode a été utilsée pour évaluer la force du signal et la significance statistique. MELA est basé sur l'observation que la cinématique de la production $ZZ \to 4l2j$ peut être décrite par un nombre limité d'observables: la masse invariante 4-leptons, les masses invariantes des bosons Z_1 et Z_2, l'angle polaire et l'angle azimutal entre le faisceau incident et la direction du Z_1 dans le référentiel au repos des quatre leptons, l'angle azimutal entre les plans de désintégration des bosons Z_1 et Z_2 et l'angle azimutal entre le plan défini par le faisceau incident et la direction du Z_1 et le plan de d;'esintégration du Z_1, mesuré dans le référentiel des quatre leptons, l'angle entre la direction du Z_1 et la direction du lepton négatif issus du Z_1 dans le référentiel au repos du Z_1, et l'angle entre la direction du Z_2 et la direction du lepton négatif issus du Z_2 dans le référentiel au repos du Z_2. La production EW de deux jets associée à deux bosons Z est mesurée avec une significativité statistique observée (attendue) de 4.0 (3.5) écarts standards. La section efficace pour la production de ce processus est mesurée dans trois régions fiducielles et est $0.33^{+0.11}_{-0.10}(stat)^{+0.04}_{-0.03}(syst)\ fb$ dans la région fiducielle la plus inclusive, en accord avec la prédiction du modèle standard de $0.275 \pm 0.021\ fb$.
Une méthode alternative pour évaluer la significance basée sur une extraction du signal au moyen d'arbres de décisions boostés a été utilisée. Un arbre de décision est une technique d'apprentissage supervisée utilisée pour des problèmes de classification ou de régression. Dans cette étude, les arbres de décision ont été utilisés pour classifier les événements comme signal ou bruit de fond. Un arbre de décision consiste en des noeuds, des feuilles et des branches. Chaque noeud représente une règle qui est utilisée pour séparer les données. La séparation est répétée jusquà la dernière feuille qui permet de classer de façon non ambigue l'événement comme signal ou bruit de fond. Une fois que l'arbre de décision est entrainé, sa performance est testée sur au nouvel échantillon (déjà connu). Cette phase est appelée test. En plus de vérifier la performance, le test est également utilisée pour détecter des signes d'un ajustement surcontraint. La surcontrainte est un phénomène où le mdoèle d'apprentissage ajuste les données de façon trop proche et induit un manque de capacité de généralisation. L'utilisation d'un arbre de décision unique peut donner un résultat instable. Pour éviter cela différentes méthodes ont été développées. Une d'elles est l'algorithme de 'boosting', qui utilise un grand nombre d'arbres de décision, où l'erreur de la décision précédente est utilisée pour améliorer l'apprentissage de l'arbre suivant. Les arbres de décisions 'boostés' (BDT) sont utilisés dans cette étude.
L'efficacité de l'arbre de décision dépend des variables utilisées pour l'entrainement. Pour maximiser la performance de l'analyse, deux approches ont été suivies pour l'extraction du signal:

1. BDT7: Un arbre de décision boosté qui utilise 7 variables
2. BDT28: Un arbre de décision boosté qui utilise 28 variables

Alors que le BD7 est utilisé comme l'alternative de base pour l'extraction du signal, le BDT28 est utilisé pour évaluer de combien le pouvoir discriminant augmente avec l'introduction de nouvelles variables. La significance du signal EW a été mesurée avec une valeur observée (attendue) de 5.1 (3.8) déviations standards avec le BDT7 et de 4.7 (3.9) déviations standards avec le BDT28. L'étude montre que a significance pour le BDT7 est comparable à MELA et qu'il n'y a pas de gain significatif à utiliser 28 variables plutôt que 7.

Des limites sur les coefficients de Wilson associés aux opérateurs de dimension 8 dans le cadre de la thérie effective (EFT) sont établies en utilsant la distribution de m_{4l}. Pour cette mesure, un échantillon Monte Carlo spécial a été généré dans lequel les valeurs des couplages quartiques des bosons de jauges sont modifiés par rapport aux valeurs du modèle standard. La conséquence est une augmentation de la section efficace de production du processus $ZZ \to 4l2j$ aux grandes valeurs de m_{4l}. Si le même comportement était observé dans les données, ce serait une indication forte de nouvelle physique. Dans cette étude les limites les plus fortes (compétitives) à ce jour sur les coefficients pour les opérateurs T0, T1, T2 et T8 sont établies. **Études prospectives pour la diffusion longitudinale des bosons vecteurs au HL-LHC et au HE-LHC**

Cette thèse montre que CMS est entré dans l'ère de la mesure pour la diffusion des bosons de jauges dans le canal $ZZ \to 4l2j$. Cependant, la nouveauté après l'introduction du mécanisme de brisure spontanée de la symmétrie EW est l'apparition d'une masse pour les bosons vecteurs faibles et, par conséquent, l'apparition d'une composante de polarisation longitudinale. Avec le LHC il n'est pas possible d'étudier la diffusion des bosons de jauges longitudinaux dans le canal $ZZ \to 4l2j$ en raison de la section efficace extrêmement faible de ce processus. Néanmoins, la plus grande énergie et la plus grande luminosité intégrée prévues pour la phase de haute luminosité (HL-LHC) et de haute énergie (HE-LHC) rendent la question de la possibilité de cette mesure pertinente. Pour cette étude, une simulation complète de la cinématique de tous les états de polarisation et des bruits de fond dans le canal $ZZ \to 4l2j$ est réalisée pour la première fois.
Les processus de signal et de bruits de fond sont simulés avec le générateur *Madgraph_AMC@NLO*, et l'hadronisation est simulée avec *Pythia8*. Au lieu d'une simulation complète des effets du détecteur par *GEANT4*, une solution plus rapide utilisant *Delphes* est choisie.
Pour réduire le bruit de fond le plus possible, les leptons doivent être isolé, ce qui est réalisé dans Delphes en demandant qu'il n'y ait pas d'activité dans un cône de rayon *R* autour du lepton. Dans les conditions du HL- et du HE-LHC, le nombre de collisions proton-proton additionnelles dans chaque croisement de paquets jouera un rôle crucial. Ce phénomène est appelé "pileup" (PU). En moyenne, 200 événements de PU sont attendus à chaque croisement de faisceau dans les conditions du HL-et du HE-LHC. Différents algorithmes développés par la collaboration CMS ont été utilisés dans l'analyse pour réduire l'impact du PU. Ce sont les algorithmes dits de "charged hadron subtraction" (CHS) et de "pileup per particle identification" (PUPPI). De plus, chaque pair de leptons doit avoir une masse invariante comprise entre 60 GeV et 120 GeV pour être considérée comme un candidat Z. Le boson Z de plus grand moment transverse est désigné Z_1, alors que le boson Z de moment transverse inférieur est nommé Z_2. Ces critères sont désignés par la sélection ZZ. Pour l'analyse statistique, deux autres conditions sont introduites:

- La selection "baseline", en plus des critères de la sélection ZZ, demande que la masse invariante des deux jets soit au moins de 100 GeV.
- La selection "VBS", en plus des critères de la sélection ZZ, demande que la masse invariante des deux jets soit au moins de 400 GeV et une différence en pseudorapidité entre les deux jets de au moins 2.4 en valeur absolue.

Avant de procéder à l'extraction du signal, l'impact des gerbes partoniques et du PU sur la sélection des deux jets de VBS dans l'événement est étudié et confirme la robustesse de l'analyse. Pour maximiser la sensibilité de l'analyse au signal, deux méthodes pour extraire le signal ont été développées, toutes les deux basées sur un algorithme de BDT. La plus simple, dite "combined-background BDT", traite les événements où les deux bosons Z sont longitudinaux (événements LL) comme signal, et les événements avec un boson Z polarisé longitudinalement et un transversalement (événements LT), deux bosons Z polarisés transversalement (événements TT), et les événements où l'état final est obtenu par des processus impliquant des vertex QCD (avec deux quarks ou avec deux gluons dans l'état initial)

comme le bruit de fond. La méthode plus complexe, appelée "2D BDT", consiste en deux étapes. Dans la première étape, un BDT est utilisé pour séparer le signal LL du bruit de fond qui consiste en les processus impliquant des vertex QCD avec deux quarks dans l'état initial. L'objectif est de séparer la production EW $ZZjj$ production de la production QCD du même état final. Un second BDT est utilisé ensuite pour séparer le signal LL d'un mélange des bruits de fonds constitu¡e par les événements LT et TT.
La même procédure a été utilisée pour la phase HL-LHC correspondant à une énergie dans le centre de masse de 14 TeV et une luminosité intégrée de 3000 fb^{-1} et pour les conditions HE-LHC correspondant à une énergie dans le centre de masse de 27 TeV et une luminosité intégrée de 15000 fb^{-1} Pour le HL-LHC, le signal LL est mesuré avec une significance attendue de 0.98 déviations standards avec la méthode "combined-background BDT" et de 1.32 déviations standard avec la méthode "2D BDT". Pour le HE-LHC, le signal LL est mesuré avec une significance attendue de 3.55 déviations standards avec la méthode "combined-background BDT" et de 4.66 déviations standard avec la méthode "2D BDT".
Durant la mise à niveau du détecteur CMS en vue de la phase HL-LHC, il est envisagé d'étendre la couverture angulaire pour les électrons, actuellement limitée à $|\eta| < 3$, à une région plus grande définie par $|\eta| < 4$. En prenant en compte une telle extension de l'acceptance pour les électrons, une plus grande sensibilité à la composante longitudinale de la diffusion ZZ est attendue. Dans les conditions du HL-LHC avec une extension de l'acceptance pour les électrons, le signal LL est mesuré avec une significance attendue de 1.04 déviations standards avec la méthode "combined-background BDT" et de 1.39 déviations standard avec la méthode "2D BDT". Dans les conditions du HE-LHC, le signal LL est mesuré avec une significance attendue de 4.66 déviations standards avec la méthode "combined-background BDT" et de 5.35 déviations standard avec la méthode "2D BDT".
Ces études prospectives démontrent le potentiel important d'une augmentation de l'énergie des collisions et de la luminosité intégrée pour l'étude de la diffusion des bosons Z longitudinaux en comparaison avec les accélérateurs actuels.

Conclusion

L'étude de la diffusion des bosons de jauges fourni un moyen extraordinaire pour une meilleure compréhension du mécanisme de brisure de la symmétrie EW, qui donne une masse aux bosons $W^{\pm}$ et Z^0. De plus, les processus VBS offrent un moyen d'étude de la structure non-Abélienne des intéractions EW à travers l'étude de couplages quartiques anormaux, fournissant ainsi une méthode pour rechercher de la physique au=delà du modèle standard dans le cadre EFT. Dans cette thèse, l'analyse du canal $ZZ \to 4l2j$ est présentée, démontrant que CMS est entrée dans lère de la mesure pour ces processus. Néanmoins, l'étude de VBS avec des bosons Z polarisés longitudinalement est encore hors d'atteinte au LHC. Cette étude montre un potentiel important pour l'étude de la diffusion longitudinale auprès d'accélérateurs à plus grande énergie et avec des plus grandes luminsotiés intégrées comme prévus dans le futur.

Prošireni sažetak

Raspršenje vektorskih bozona (VBS) daje fundamentalan uvid u mehanizam spontanog narušenja elektroslabe simetrije kojim baždarni bozoni u teoriji Standardnog Modela dobivaju masu. Ovaj rad prikazuje prvi dokaz elektroslabe proizvodnje para Z bozona, koji se raspada u četiri leptona (elektrona i/ili miona), praćenog s dva hadronska mlaza s topologijom raspršenja vektorskih bozona. Studija analizira 137 fb^{-1} proton-proton sudara prikupljenih CMS detektorom na LHC-u pri energiji 13 TeV u centru mase. Dodatno, iznesene su prve prospektivne studije o mogućnosti mjerenja longitudinalnog raspršenju u istom kanalu raspada u uvjetima LHC-a visokog luminoziteta (HL-LHC) i visoke energije (HE-LHC), koje odgovaraju energiji centra mase od 14 odnosno 27 TeV, sa simuliranom potpunom kinematikom navedenih događaja.

Uvod u Standardni Model i raspršenje vektorskih bozona

Standardni Model je, u trenutku nastanka ovog rada, najpotpunija teorija koja opisuje elementarne čestice i njihove interakcije. To je relativistička kvantna teorija polja opisana grupom simetrija $SU(3)_C \times SU(2)_L \times U(1)_Y$ pri čemu prvi član opisuje simetriju kvantne kromodinamike (QCD), drugi član opisuje simetriju u teoriji slabih interakcija, dok je posljednji član grupa simetrija u teoriji elektromagnetskih interakcija.
Jedna od mogućih podjela čestica u Standardnom Modelu bazirana je na kvantnomehaničkoj observabli *spin*. Prema ovoj podjeli, u prirodi razlikujemo čestice polucjelobrojnog spina koje nazivamo *fermionima* i čestice cjelobrojnog spina koje nazivamo *bozonima*. Sva materija u svemiru građena je od fermiona, dok su bozoni odgovorni za interakcije među njima. Fermioni se dijele na kvarkove i leptone grupirane u 3 razreda. U prvom razredu leptona nalaze se *elektron* (e) i *elektronski neutrino* (ν_e), dok prvi razred kvarkova sačinjavaju *gore kvark* (u) i *dolje kvark* (d). Preostala dva razreda sačinjena su od njihovih masivnijih kopija: μ, ν_μ, τ i ν_τ leptona te c, s, t i b kvarkova. Uz navedene fermione, u Standardnom Modelu postoje i anti-fermioni koji se od njih razlikuju u suprotnim vrijednostima kvantnih brojeva. Za razliku od leptona, kvarkove u prirodi nikada ne nalazimo individualno, već uvijek grupirane u kvark-antikvark parove (takozvani "*mezoni*") ili u trojke kvarkova (takozvani "*barioni*"). Barioni i mezoni se jednom riječju nazivaju *hadroni*.
Atom je vezano stanje jezgre i elektrona, pri čemu ključnu ulogu ima *elektromagnetska sila* koju prenosi čestica spina 1 (takozvani baždarni bozon) koju nazivamo *foton*. Jezgru atoma na okupu drži *jaka sila* koju prenosi osam *gluona* koji također imaju spin 1. Konačno, slaba sila, odgovorna za beta raspade, prenosi se trima baždarnim bozonima: $W^\pm$ i Z^0.

Prije 1964. godine postojalo je neslaganje u predviđanju Standardnog Modela, kojim bi baždarni bozoni trebali biti bez mase, i eksperimenata koji su ukazivali suprotno. Jedan način razrješenja ovog problema ponuđen je 1964. godine i potvrđen otkrićem Higgsovog bozona na CERN-u 2012. godine. Radi se o Brout-Englert-Higgs (BEH) mehanizmu koji se temelji na uvođenju dubleta kompleksnog skalarnog polja s neiščezavajućom vrijednošću vakuuma. Rezultat je spontano narušenje simetrije elektroslabih interakcija koje rezultira pojavom četiri Goldstoneova

bozona. Kako bi se očuvala lokalna SU_2 simetrija, polja Goldstoneovih bozona se kombiniraju s poljima nemasivnih $W^{\pm}$ i Z^0 bozona što za posljedicu ima generiranje mase istih.
Za razliku od bezmasivnih W i Z bozona koji imaju samo transverzalnu komponentu polarizacije, masivni W i Z bozoni dobivaju dodatni stupanj slobode: longitudinalnu komponentu polarizacije. Značajna razlika u ponašanju longitudinalno polariziranih W i Z bozona u odnosu na transverzalno polarizirane očituje se u divergentnom ponašanju amplitude raspršenja longitudinalnih vektorskih bozona pri visokim energijama. Ovdje ključnu ulogu ima Higgsov bozon čije vezanje na vektorske bozone omogućava unitarizaciju. Proučavanje raspršenja longitudinalnih vektorskih bozona, dakle, predstavlja dodatni alat za proučavanje svojstava Higgsovog bozona i mehanizma spontanog narušenja elektroslabe simetrije. Također, ono omogućava proučavanje neabelove strukture elektroslabih interakcija kroz proučavanje kvadrilinearnih verteksa. Konačno, naznake fizike van Standardnog Modela (BSM) mogu se manifestirati kroz modifikacije određenih kvartičkih baždarnih vezanja.

Veliki hadronski sudarivač i CMS eksperiment

Veliki hadronski sudarivač (LHC) je najveći kompleks akceleratora na svijetu s opsegom od 27 km kojim upravlja Europsko vijeće za nuklearna istraživanja (CERN). Nalazi se na Francusko-Švicarskoj granici i posljednja je faza ubrzavanja protona pri čemu se postiže energija od približno 13 TeV u centru mase proton-proton sudara. U tunelima LHC-a protoni kruže u dva protonska snopa koji se sudaraju na unaprijed predviđenim mjestima. Kako bi se povećala vjerojatnost da se dva protona sudare, LHC-om u svakom trenutku kruži 2000 gustih paketića protona u svakom protonskom snopu. Svaki protonski paketić sadrži približno 10^{11} protona. Uz dva eksperimenta općenite namjene, *CMS* i *ATLAS*, na LHC-u novu fiziku traže *LHCb*, *ALICE*, *TOTEM*, *MoEDAL* i *LHCf* eksperimenti.
U ovom radu korišteni su podaci prikupljeni CMS detektorom koji se nalazi nedaleko od Francuskog sela Cessy, približno 100 metara ispod zemlje. Detektor je dug 21 metar te širok i visok 15 metara, a dizajniran je oko supravodljivog solenoidalnog magneta koji generira magnetsko polje do 4 T. CMS se sastoji od nekoliko pod-detektorskih sustava dizajniranih za obavljanje specifičnih zadataka. Najbliže točki u kojoj se sudaraju protoni (i.e. interakcijskoj točki) nalazi se *detektor tragova* čija je uloga mjerenje količine gibanja nabijenih čestica koristeći zakrivljenosti putanja čestica u magnetskom polju. Slijedi elektromagnetski kalorimetar (ECAL) koji se koristi za mjerenje energije elektrona i fotona. Aktivni medij ECAL-a su transparentni kristali olovnog volframata. Elektroni i fotoni prilikom prolaska kroz kristale ECAL-a stvaraju pljusak elektrona i fotona nižih energija. Elektroni gube svoju energiju u obliku kratkih bljeskova svjetla što nazivamo scintilacijom. Ovi bljeskovi svjetlosti prikupljaju se na fotodetektorima koji se nalaze na kraju svakog kristala i pretvaraju se i električni signal koji se šalje na obradu. Iza ECAL-a postavljen je hadronski kalorimetar (HCAL) čija je uloga mjerenje energije hadrona te u suradnji s ECAL-om i detektorom tragova omogućava mjerenje količine gibanja svih hadrona. Najudaljeniji pod-detektor CMS-a su mionske komore čija je uloga identifikacija i mjerenje količine gibanja miona.
Svakih 25 ns, u prosjeku, dogodi se sudar protona na LHC-u što generira približno 50 TB podataka u svakoj sekundi. Pošto ovoliku količinu podataka nije moguće pohraniti, veliki dio je odbačen. Kako se ne bi odbacili fizikalno zanimljivi događaji, razvijen je, takozvani, sustav za okidanje (eng. trigger), čija je uloga od početnih 40 MHz podataka sačuvati samo približno 1-2 kHz. Događaji koji su sačuvani idu na daljnju obradu prilikom koje se rekonstruiraju fizikalni objekti. Za ovo su razvijeni kompleksni algoritmi koji koristeći sve pod-detektorske sustave rekonstruiraju elektrone, mione, hadronske mlazove (eng. *jet*) i nedostajuću transverzalnu energiju (eng. missing E_T) koja ukazuje na prisustvo neutrina. Analize koje koriste ove objekte moraju implementirati dodatne zahtjeve (i.e. selekciju) kako bi maksimalno smanjili pozadinske događaje (eng. background) u svojim podacima.
Tijekom 2018. godine CERN je priveo kraju Run 2 fazu rada LHC-a tijekom koje je sakupljeno 140 fb^{-1} podataka. Pripreme za Run 3 fazu krenule su 2019. godine s ciljem prikupljanja približno 300 fb^{-1} pri energiji 13.6 TeV u centru mase proton-proton sudara. Kako bi se osigurao napredak u području fizike elementarnih čestica, nakon Run

3 faze predviđena je faza visokog luminoziteta (HL-LHC) s ciljem prikupljanja 3000 fb^{-1} podataka pri energiji 14 TeV u centru mase proton-proton sudara. Sljedeći veliki projekt CERN-a je Budući kružni sudarivač (FCC) opsega 100 km s planiranom energijom 100 TeV u centru mase proton-proton sudara. Između HL-LHC faze i FCC-a planirana je faza visoke energije LHC-a (HE-LHC) s ciljem prikupljanja 15000 fb^{-1} pri energiji 27 TeV u centru mase proton-proton sudara.

Rekonstrukcija elektrona i mjerenje efikasnosti

U odnosu na rekonstrukciju miona, rekonstrukcija elektrona je osjetno kompleksnija zbog zakočnog zračenja (bremsstrahlung) kojem elektroni podliježu prilikom kretanja kroz aktivni medij detektora. Prva posljedica zakočnog zračenja je povećanje broja kristala, duž ϕ smjera, u kojima elektroni ostavljaju svoju energiju. Druga posljedica zakočnog zračenja je povećanje zakrivljenosti putanje elektrona. Kako bi se ovomu doskočilo, implementirani su složeni algoritmi za rekonstrukciju elektrona.
Prvi korak u rekonstrukciji je prikupljanje svih energijskih depozicija nastalih uslijed prolaska elektrona kroz aktivni medij detektora, kao i energijske depozicije izračenih fotona, u takozvane klastere. Kako bi se povećala efikasnost rekonstrukcije, klasteri se grupiraju u takozvane superklastere. Paralelno s klasteriranjem radi se rekonstrukcija tragova u detektoru tragova. Slijedi procjena naboja objekta koristeći zakrivljenost putanje čestice u kombinaciji s drugim metodama. Iako je osnovna ideja grupiranja klastera u superklastere upravo prikupljanje svih energijskih depozicija, ovo najčešće nije 100 % efikasno. Iz tog razloga se rekonstruirana energija korigira koristeći simulacije i metode strojnog učenja. Sve dobivene informacije koriste se kao ulazni parametri u, takozvani, *particle-flow* algoritam koji klasificira objekt kao elektron ili foton.
Poseban naglasak u ovom radu stavljen je na selekciju elektrona, odnosno, mjerenje efikasnosti selekcije elektrona. Kako bi se smanjila pozadina (primjerice, elektroni iz hadronskih mlazova čestica ili signali iz detektora koji su pogrešno rekonstruirani kao elektroni), selekcija elektrona se odvija u nekoliko koraka:

1. selekcija primarnih elektrona (i.e. elektrona koji ne potiču iz, primjerice, konverzija fotona ili piona)
2. identifikacija elektrona
3. izolacija elektrona

Od 2017. godine algoritam je unaprijeđen tako da se varijable korištene za izolaciju elektrona iskoriste za identifikaciju elektrona. Sama identifikacija elektrona radi se uz pomoć strojnog učenja.
Kako bi se izmjerila efikasnost selekcije elektrona, razvijena je standardna metoda "označi i ispitaj" (eng. Tag and Probe) koja koristi događaje iz $Z \to l^+l^-$ kanala kako bi se dobio uzorak elektronskih kandidata. Prvo se iz dobivenih podataka u svakom događaju odabere signalni elektron koristeći vrlo stroge uvjete (takozvani "tag" elektron). U istom događaju se zatim traži, korištenjem blažih uvjeta čiju efikasnost ispitujemo, elektron (takozvani "probe" elektron) koji u paru s "tag" elektronom ima masu blisku masi Z bozona. Efikasnost selekcije je definirana kao omjer broja "probe" elektrona koji zadovoljavaju uvjete selekcije i ukupnog broja "probe" elektrona. U slučaju kada nema značajnog doprinosa pozadine, efikasnost selekcije se može dobiti jednostavnom metodom prebrojavanja (eng. cut and count). U suprotnom se koristi složenija metoda u kojoj se podaci grupiraju u razrede transverzalne količine gibanja (p_T) i pseudorapiditeta (η), a zatim se u svakom razredu matematičkom prilagodbom podacima (eng. fitting) pronađe krivulja za "probe" elektrone koji prolaze selekciju, kao i za one koji ne prolaze selekciju. Površina ispod krivulje daje broj "probe" elektrona koji prolaze, odnosno ne prolaze, selekciju što se zatim koristi za izračun efikasnosti selekcije. Efikasnost selekcije mjeri se kako na stvarnim podacima, tako i na simuliranim podacima. Bilo kakva razlika u efikasnosti u simulaciji i podacima se korigira, za svaki p_T i η razred odvojeno, primjenom faktora skaliranja (eng. scale factors).

Efikasnost selekcije elektrona i faktori skaliranja prvo su izračunati za 2018. godinu koristeći elektronsku identifikaciju dobivenu metodama strojnog učenja primijenjenim na podatke iz 2017. godine pri čemu su izolacijske varijable korištene u identifikaciji. Nakon toga, elektronska identifikacija je iznova napravljena za 2016. godinu primjenjujući metode strojnog učenja na podatke iz 2016. godine. Isto je napravljeno i za 2018. godinu. Efikasnost selekcije i faktori skaliranja nanovo su izračunati za sve tri godine kako bi se smanjila nesigurnost mjerenja za elektrone s malim vrijednostima transverzalne količine gibanja (< 10 GeV), ali i da bi se proučila η struktura u efikasnostima selekcije. Kako bi se dodatno smanjila nesigurnost mjerenja selekcije za elektrone s malim vrijednostima transverzalne količine gibanja, zadani su stroži uvjeti na selekciju "tag" elektrona u događajima s niskim vrijednostima transverzalne količine gibanja "probe" elektrona. Ovo je rezultiralo nešto jasnijim vrhom distribucije (eng. peak) u okolici mase Z bozona što je poboljšalo preciznost i smanjilo nesigurnost mjerenja selekcije za sva tri razdoblja. Kao posljedica strože selekcije "tag" elektrona pojavila se nakupina događaja (eng. bump) u području niskih vrijednosti m_{ee} distribucije za "probe" elektrone koji ne zadovoljavaju uvjete selekcije. Proučavanjem ove pojave zaključeno je kako se radi o migraciji signalnih elektrona u skupinu elektrona koji ne zadovoljavaju uvjete selekcije. Kako bi se nova distribucija uspješno fitala, bilo je potrebno modificirati funkciju korištenu u fitu. Posljedično, nesigurnosti mjerenja u ovom području su dodatno smanjene. Osim strožih uvjeta na selekciju "tag" elektrona, povećan je i broj η i p_T podjela kako bi se pobliže proučila η struktura u efikasnosti selekcije. Uočena je struktura kišobrana koja je rezultat zahtjevnije rekonstrukcije i identifikacije elektrona za velike vrijednosti η.
Ovi rezultati korišteni su u publikaciji $H \to ZZ \to 4l$ analize s Run 2 podacima i koriste se kao standardni recept za selekciju elektrona s transverzalnom količinom gibanja između 5 i 10 GeV. Također, ključni su za uspješnost VBS $ZZ \to 4l2j$ analize prezentirane u ovom radu.

Potraga za VBS-om u $ZZ \to 4l2j$ kanalu

Potraga za VBS-om u $ZZ \to 4l2j$ kanalu provedena je korištenjem podataka prikupljenim u periodu 2016. - 2018. godine koristeći CMS detektor. Pošto se finalne čestice u ovom kanalu, elektroni, mioni i hadronski mlazovi, mogu u potpunosti rekonstruirati na CMS-u, za očekivati je kako će baš ovaj kanal postati među najvažnijima za proučavanje raspršenja longitudinalnih vektorskih bozona u budućnosti što će omogućiti bolji uvid u teoriju elektroslabih interakcija u sklopu Standardnog Modela. Štoviše, ovaj kanal je osjetljiv na određene operatore u formalizmu efektivne teorije polja (T0, T1, T2, T8 i T9), pa omogućava i proučavanje u anomalnih kvartičkih baždarnih vezanja, a time i alat za proučavanje fizike izvan granica Standardnog Modela.

Kako bi se što bolje opisali signalni i pozadinski procesi, posebna pažnja pridana je Monte Carlo (MC) simulacijama. Uvjeti selekcije elektrona posebno su dizajnirani za ovu analizu kako bi se smanjio udio događaja pozadine. Na selekciju leptona su postavljeni blagi uvjeti kako bi se smanjila pozadina, a istovremeno zadržala maksimalna efikasnost signala pazeći pritom da se broj događaja previše ne reducira i time smanji statistička moć analize. Leptoni koji zadovolje uvjete selekcije kombiniraju se u Z bozone pri čemu se sparuju isključivo leptoni istog okusa i različitog naboja (i.e. e^+e^-, $\mu^+\mu^-$). Svaki Z kandidat mora zadovoljiti niz kriterija kako bi se smanjila vjerojatnost da Z bozone koji nastaju u pozadinskim procesima zamijenimo za signalne Z bozone. Od dva odabrana Z bozona u događaju, Z_1 je onaj Z bozon s većim iznosom transverzalne količine gibanja, dok je Z_2 preostali Z bozon. U događajima s više od dva Z bozona odabiremo onaj par čiji će skalarni zbroj transverzalne količine gibanja dvaju leptona koji čine Z_2 biti veći. Za potrebe statističke analize je definiran kriterij (eng. inclusive ZZjj selection) koji, uz navedeno, zahtijeva da dva hadronska mlaza imaju masu veću od 100 GeV. Ovim je definiran osnovni set zahtijeva na signalne događaje korišten u statističkoj analizi. Definirani su i dodatni kriteriji (eng. loose VBS selection i tight VBS selection) kako bi se odabrali fazni prostori s većim udjelom događaja koji dolaze od VBS procesa.

Matrix Element Likelihood Approach (MELA) diskriminanta je korištena kako bi se izračunao udarni presjek elektroslabe (EWK) i EWK+QCD $ZZ \to 4l2j$ produkcije. Ista metoda je korištena za izračun jakosti i značajnosti EWK signala. MELA se temelji na spoznaji da se kinematika $ZZ \to 4l2j$ kanala može opisati koristeći nekoliko observabli: invarijantnu masu četiriju leptona u finalnom stanju, invarijantnu masu Z_1 i Z_2 bozona, polarne i azimutalne kuteve između osi zrake protona i smjera kretanja Z_1 bozona u sustavu mirovanja četiriju leptona, azimutalni kut između vektora normale na ravnine produkata Z_1 i Z_2 bozona, azimutalni kut između ravnine definirane upadnim protonskim snopom i Z_1 bozonom te ravnine produkata raspada Z_1 bozona mjerenim u sustavu mirovanja četiriju leptona, kut između smjera putanje Z_1 bozona i vektora količine gibanja negativnog produkta raspada Z_1 bozona u sustavu mirovanja Z_1 bozona te kut između smjera putanje Z_2 bozona i vektora količine gibanja negativnog produkta raspada Z_2 bozona u sustavu mirovanja Z_2 bozona. Elektroslaba proizvodnja dvaju mlazova u kombinaciji s dva Z bozona izmjerena je s opaženom (očekivanom) značajnošću od 4.0 (3.5) standardne devijacije. Udarni presjeci za elektroslabu proizvodnju navedenog procesa izmjereni su u tri vrste selekcije s $0.33^{+0.11}_{-0.10}(stat)^{+0.04}_{-0.03}(syst)\ fb$, u najinkluzivnijem slučaju selekcije, a u skladu s predviđanjem Standardnog Modela od $0.275 \pm 0.021\ fb$.
Kao alternativna metoda izračuna signifikantnosti korištena su stabla odluke (eng. decision trees). Stablo odluke je nadzirana metoda strojnog učenja koja se koristi kako za probleme klasifikacije, tako i probleme regresije. U ovom radu stabla odluke su korištena kako bi se događaji klasificirali ili kao signal ili kao pozadina. Stablo odluke se sastoji od čvorova, listova i grana. Svaki čvor predstavlja pravilo koje se koristi za grananje podataka. Grananja se ponavljaju sve dok se ne dođe do čvora koji jednoznačno svrstava događaj kao signal ili kao pozadinu. Treniranje bilo koje metode strojnog učenja je proces u kojem se koriste poznati podaci kako bi se konstruirao set pravila koji najtočnije svrstava podatke u signal i pozadinu. Nakon što je stablo odluke istrenirano, potrebno je testirati njegov učinak na novom (ali poznatom) setu podataka. Ova faza se naziva testiranje stabla odluke. Osim provjere efikasnosti, testiranje se koristi i kako bi se otkrile naznake *"pretreniranja"*. Pretreniranje je pojava pri kojoj je odabranoj metodi strojnog učenja dozvoljeno da predetaljno prouči svojstva skupa podataka na kojima trenira zbog čega metoda gubi moć generalizacije. Korištenjem samo jednog stabla odluke dobivaju se nerobustni rezultati. Kako bi se ovo izbjeglo, razvijene su različite metode. Jedna od njih je "boost" algoritam koji koristi velik broj stabala odluke pri čemu se pogreške prethodnog stabla odluke koriste kako bi se poboljšalo učenje sljedećeg stabla. U ovom radu su korištena upravo boostana stabla odluke (eng. boosted decision trees, BDTs).
Učinkovitost stabla odluke ovisi o varijablama koje se koriste za treniranje. Kako bi statistička moć analize bila što veća, osmišljena su dva pristupa ekstrakcije signala koristeći stabla odluke:

1. BDT7: stablo odluke koje koristi 7 ulaznih varijabli za treniranje
2. BDT28: stablo odluke koje koristi 28 ulaznih varijabli za treniranje

Dok je BDT7 korišten kao primarna alternativna metoda ekstrakcije signala, BDT28 je korišten da se provjeri koliko se moć razlučivanja signala poveća uvođenjem dodatnih varijabli. Značajnost EWK signala izmjerena je s opaženom (očekivanom) značajnošću od 5.1 (3.8) standardne devijacije koristeći BDT7, odnosno 4.7 (3.9) standardne devijacije koristeći BDT28. Ovim je pokazano kako MELA dobro opisuje kinematiku $ZZ \to 4l2j$ događaja te da se ne dobiva značajno na moći razlučivanja signala korištenjem BDT28 u odnosu na BDT7.

Granične vrijednosti Wilsonovih koeficijenata pridruženih 8-dimenzionalnim operatorima anomalnih kvartičkih bždarnih vezanja izračunate su u formalizmu efektivne teorije polja (EFT) koristeći m_{4l} distribuciju. Za potrebe ovog izračuna generiran je poseban MC uzorak pri čemu su modificirane vrijednosti kvartičkih vezanja vektorskih bozona u odnosu na nominalne vrijednosti iz Standardnog Modela. Posljedica ovih modifikacija je povećanje udarnog presjeka $ZZ \to 4l2j$ procesa pri velikim vrijednostima m_{4l}. Ukoliko bi se isto ponašanje uočilo u stvarnim podacima, ovo bi bila snažna naznaka nove fizike. U ovom radu su postavljene najstrože granične vrijednosti, dostupne u vrijeme istraživanja, Wilsonovih koeficijenata za operatore T0, T1, T2, T8 i T9.

Prospektivne studije za raspršenje longitudinalnih vektorskih bozona pri HL-LHC i HE-LHC

Ovaj rad pokazuje kako je CMS ušao u fazu mjerenja raspršenja vektorskih bozona u $ZZ \to 4l2j$ kanalu raspada. Ipak, novitet nakon uvođenja mehanizma spontanog narušenja elektroslabe simetrije je pojava mase vektorskih bozona i, posljedično, pojava longitudinalne komponente polarizacije istih. Pri postojećim uvjetima koji vladaju na LHC-u nije moguće proučavati raspršenja longitudinalnih vektorskih bozona u $ZZ \to 4l2j$ kanalu raspada zbog iznimno malog udarnog presjeka ovih procesa. Međutim, puno veća energija i integrirani luminozitet planirani u uvjetima LHC-a visokog luminoziteta (HL-LHC) i visoke energije (HE-LHC) čine pitanje mogućnosti mjerenja ovih procesa relevantnim. Za potrebe ovog istraživanja prvi put je napravljena simulacija potpune kinematike svih stanja polarizacije u VBS $ZZ \to 4l2j$ kanalu.
Signalni i pozadinski procesi simulirani su korištenjem *Madgraph_AMC@NLO* alata, dok je hadronizacija fragmenata sudara simulirana *Pythia8* paketom. Umjesto potpune simulacije utjecaja detektora na propagaciju čestica koristeći *GEANT4*, odabrano je brže i jednostavnije rješenje u vidu *Delphes* paketa.
Kako bi se što više reducirala pozadina, postavljen je niz zahtjeva na fizikalne objekte i događaje. Tražilo se da fizikalni objekti budu izolirani što se u Delphes-u postiže zahtjevom da u konusu radijusa *R* oko smjera putanje leptona nema dodatne aktivnosti. U HL- i HE-LHC uvjetima veliku će ulogu imati broj dodatnih proton-proton sudara pri svakom sudaru paketića protona. Ova pojava naziva se "pileup" (PU). U prosjeku, očekuje se 200 PU događaja u svakom sudaru paketića protona pri HL- i HE-LHC uvjetima. U analizi su korišteni različiti algoritmi dizajnirani od strane CMS kolaboracije kako bi se utjecaj PU-a smanjio. Riječ je o, takozvanim, "*charged hadron subtraction*" (CHS) i *pileup per particle identification* (PUPPI) algoritmima. Uz navedeno, od svakog para leptona traženo je da ima invarijantnu masu između 60 i 120 GeV kako bi bio kandidat za sparivanje u Z bozon. Z bozon s najvećim iznosom transverzalne količine gibanja u događaju proglašen je Z_1, a preostali Z bozon Z_2. Ovaj skup zahtjeva naziva se *ZZ selekcija*. Za potrebe statističke analize definirana su i dva dodatna uvjeta:

- "baseline" selekcija koja, povrh zahtjeva ZZ selekcije, traži da dva hadronska mlaza u događaju imaju invarijantnu masu od, minimalno, 100 GeV
- "VBS" selekcija koja, povrh zahtjeva ZZ selekcije, traži da dva hadronska mlaza u događaju imaju invarijantnu masu od, minimalno, 400 GeV i da budu razmaknuti u apsolutnoj vrijednosti pseudorapiditeta za, minimalno, 2.4

Prije postupka ekstrakcije signala provjeren je utjecaj partonskih pljuskova i PU-a na odabir dva vodeća hadronska mlaza u događaju kako bi se potvrdilo da je analiza stabilna.

Kako bi se maksimizirala osjetljivost analize na signalne događaje, razvijena su dva postupka ekstrakcije signala, pri čemu se oba temelje na BDT algoritmu strojnog učenja. Jednostavnija od dvije metode, "combined-background BDT" metoda, tretira događaje s dva longitudinalna Z bozona (i.e. LL događaje) kao signal, a mješavinu događaja s jednim longitudinalnim i jednim transverzalnim Z bozonom (i.e. LT događaje), s oba transverzalna Z bozona (i.e. TT događaje) i događaje u kojima je identično finalno stanje dobiveno u procesima s QCD vertexima (bilo s dva kvarka ili dva gluona u početnom stanju) kao pozadinu. Kompleksnija metoda, takozvana "2D BDT", odvija se u dva paralelna koraka. U prvom koraku se koristi BDT kako bi se LL događaje (signal) odvojilo od pozadine u kojoj je identično finalno stanje nastalo u procesima s QCD vertexima s dva kvarka u početnom stanju (pozadina). Ovo za zadaću ima odvojiti EWK produkciju $ZZjj$ finalnog stanja od QCD produkcije istog finalnog stanja. Istovremeno, koristi se novi BDT kako bi se LL događaje (signal) odvojilo od mješavine LT i TT događaja (pozadina).
Iste su procedure korištene pri HL-LHC uvjetima koji odgovaraju energiji od 14 TeV u centru mase proton-proton sudara i integriranom luminozitetu od 3000 fb^{-1} i pri HE-LHC uvjetima koji odgovaraju energiji od 27 TeV u centru

mase proton-proton sudara i integriranom luminozitetu od 15000 fb^{-1}. U HL-LHC uvjetima LL signal je izmjeren s očekivanom značajnošću od 0.98 standardnih devijacija korištenjem "combined-background BDT" metode i 1.32 standardne devijacije korištenjem "2D BDT" metode. U HE-LHC uvjetima LL signal je izmjeren s očekivanom značajnošću od 3.55 standardnih devijacija korištenjem "combined-background BDT" metode i 4.66 standardnih devijacija korištenjem "2D BDT" metode.
Za vrijeme trajanja nadogradnje CMS detektora tijekom pripreme za HL-LHC fazu razmatra se mogućnost proširenja zone prihvaćanja elektrona, koja je trenutno ograničena na $|\eta| < 3$, na širu zonu definiranu s $|\eta| < 4$. Ako se uzme u obzir proširena zona prihvaćanja elektrona, tada se može očekivati i veća osjetljivost na longitudinalnu komponentu raspršenja Z bozona. U HL-LHC uvjetima s proširenom zonom prihvaćanja elektrona LL signal je izmjeren s očekivanom značajnošću od 1.04 standardne devijacije korištenjem "combined-background BDT" metode i 1.39 standardnih devijacija korištenjem "2D BDT" metode. U HE-LHC uvjetima s proširenom zonom prihvaćanja elektrona LL signal je izmjeren s očekivanom značajnošću od 4.66 standardnih devijacija korištenjem "combined-background BDT" metode i 5.35 standardnih devijacija korištenjem "2D BDT" metode.
Ovim prospektivnim studijama pokazan je veliki potencijal za proučavanje raspršenja longitudinalnih Z bozona na akceleratorima s većim energijama i integriranim luminozitetima u odnosu na postojeće akceleratore.

Zaključak

Proučavanje raspršenja vektorskih bozona predstavlja izvanredan alat za bolje razumijevanje mehanizma spontanog narušenja elektroslabe simetrije kojim baždarni bozoni Standardnog Modela dobivaju masu. Također, proučavanje VBS procesa daje uvid u neabelovu strukturu elektroslabih interakcija kroz proučavanje kvartičkih baždarnih vezanja, ali i predstavlja metodu kojom bi se mogle pronaći naznake fizike izvan granica Standardnog Modela u formalizmu efektivne teorije polja. U ovom radu predstavljena je CMS analiza u $ZZ \to 4l2j$ kanalu raspada kojom je pokazano da CMS ulazi u fazu uspješnog mjerenja ovih procesa. Unatoč tome, raspršenje longitudinalne komponente Z bozona s VBS topologijom i dalje je van dometa na LHC-u. Konačno, ova studija pokazuje i veliki potencijal za proučavanje raspršenja longitudinalnih Z bozona na akceleratorima s većim energijama i integriranim luminozitetima predviđenim u nadolazećim projektima pod okriljem CERN-a.

Contents

Chapter 1

The Standard Model and vector boson scattering

1.1 Preface to the chapter

This chapter discusses the theoretical foundations essential to follow the work presented in this thesis. The chapter starts with a short overview of the Standard Model. In sections 1.2.1 and 1.2.2 the Lagrangians of the quantum electrodynamics and quantum chromodynamics are derived from the local gauge invariance requirement. The next section discusses the unification of electromagnetic and weak interactions and derives the Lagrangian for the theory of electroweak interactions. In section 1.3 the origin of the weak vector boson masses is discussed through the mechanism of electroweak symmetry breaking and the Brout-Englert-Higgs mechanism. Section 1.4 introduces the theoretical concepts and phenomenology of the vector boson scattering. The emergence of the longitudinal polarization of vector bosons after the electroweak symmetry breaking is a greatly important concept and is discussed in detail. Section 1.4.2 introduces the effective field theory framework within which the beyond Standard Model physics is searched for via measurements of anomalous quartic gauge couplings. The chapter ends with a chronological overview of the most important results published thus far by the CMS and ATLAS collaborations on the topic of vector boson scattering in various channels and at different energies. This section is envisioned as a compact summary of available results and will help reader see the contribution of this thesis work as a part of the important ongoing endeavour toward better understanding fundamental physics at the smallest scale.

1.2 Introduction to the Standard Model

The most complete theory, to date, of elementary particles and interactions between them, is given by the Standard Model (SM) of particle physics. This is a relativistic quantum field theory with an underlying $SU(3)_C \times SU(2)_L \times U(1)_Y$ structure where the first term denotes a group symmetry of the quantum chromodynamics (QCD), the second term defines a group symmetry of the week sector of the theory while $U(1)$ is a group symmetry of the quantum electrodynamics (QED). The subscripts C, L, and Y refer to *color*, *left*, and *hypercharge*. The building blocks of the theory are quantum fields whose excitations are identified as elementary particles.

One possible classification of elementary particles in the SM is based on the quantum mechanics observable *spin*. Therefore, elementary particles are divided into the half-integer spin particles called *fermions*, and the integer spin particles called *bosons*. All matter in the universe consists of fermions, while bosons govern the interactions

between them. Fermions are divided into leptons and quarks which are further grouped into three flavour generations. The first lepton generation consists of an electron (e) and an electron neutrino (ν_e). The other two generations, namely, muon (μ) with corresponding muon neutrino (ν_μ) and tau (τ) with corresponding tau neutrino (ν_τ), are then more massive replicas of the former. While the first generation is stable, the other two are not and consequentially decay very quickly into their first-generation counterparts.
Similarly, quarks are grouped into three generations. The first generation is comprised of the up (u) and down (d) quarks. The other two generations comprise the charm (c) and strange (s) quarks and the top (t) and bottom (b) quarks. Similarly to leptons, only the first generation of quarks is stable.
In addition to each fermion mentioned above, the SM recognizes anti-fermions which differ from their matter counterparts by opposite quantum numbers (e.g. electric charge and lepton numbers). The anti-particle is usually denoted with a *bar* above the particle designation. For example, an antimatter pair for the u quark is the anti-u quark denoted simply as $\bar{u}$.
Unlike leptons, which can be observed in nature as excitations of underlying fields, this is not the case for quarks. A single quark has never been observed in nature. Instead, only combinations of a (quark, anti-quark) pair, called *mesons*, or quark triplets, called *baryons*, have been observed. An example of a baryon are the (u, u, d) triplet or the (u, d, d). The former is known as the proton and the latter as the neutron. Baryons and mesons are usually referred to as *hadrons*. Together with an electron, they make up the atom which is the fundamental building block of life.

The electron is held in a bound state with a nucleus via electromagnetic interaction mediated through gauge bosons of the electromagnetic interaction. These are called *photons* and are massless spin-1 bosons. On the other hand, the nucleus of the atom is held together by means of the strong force. The mediators of the strong force are massless, spin-1 gauge bosons called *gluons*. Heavy hadrons and leptons decay through the exchange of massive spin-1 gauge bosons of the weak force. These are $W^\pm$ and Z^0 bosons. Unlike the strong interaction, which affects only those particles which possess the colour charge, or electromagnetic interaction which only affects fermions with non-vanishing electric charge, a weak interaction affects all aforementioned particles. In the high energy limit, the electromagnetic force and the weak force are unified into a single force - the electroweak force.

The final member of the elementary particle zoo is the massive, spin-0 particle predicted in theory in 1964 and discovered at CERN in 2012. As we will see in the following sections, this particle was introduced in order to resolve an important disagreement between the theory and the measured reality. Namely, according to the "original" SM, the gauge bosons should be massless. Although this is true for gauge bosons of the electromagnetic and the strong interactions, it contradicts the mass measurements of the gauge bosons in the weak sector of the theory. The particle is known as the Higgs boson and its origin will be discussed in section 1.3.2.

The full list of elementary particles in the SM is summarized in Fig. 1.1.

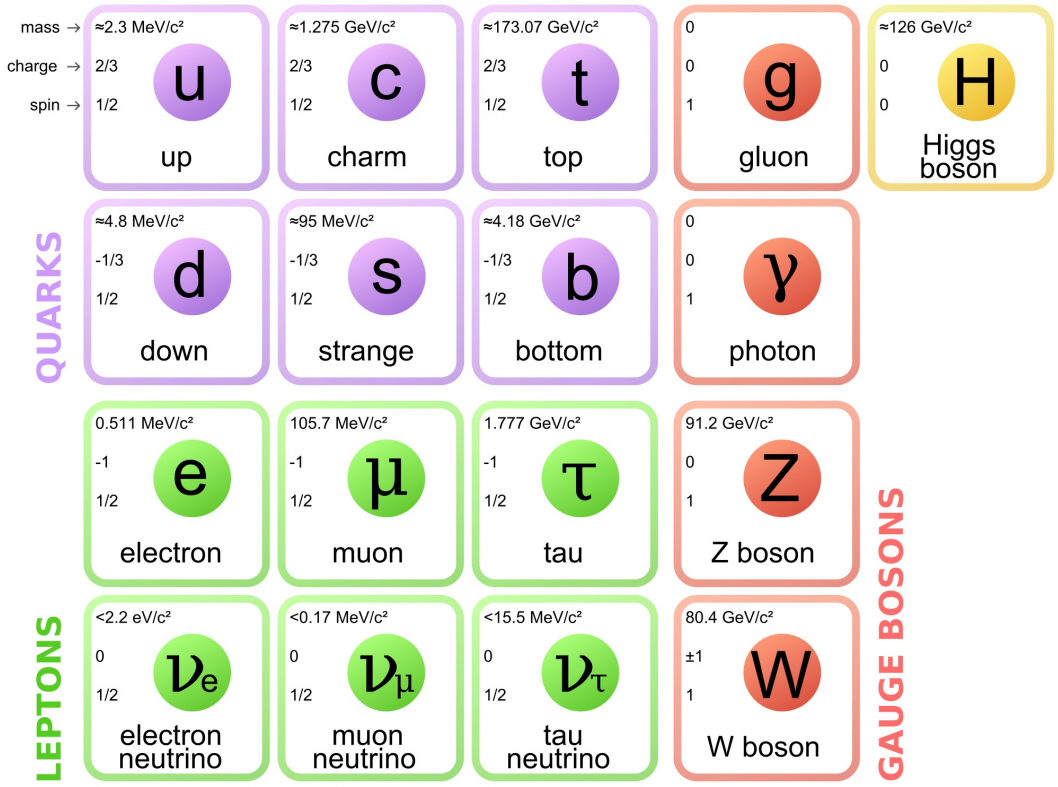


Figure 1.1: The list of known elementary particles in the Standard Model.

1.2.1 The Lagrangian of the quantum electrodynamics

The starting point for constructing the Lagrangian density (henceforth Lagrangian) of QED is the free Dirac fermion:

$$\mathcal{L}_{free} = i\bar{\psi}(x)\gamma^{\mu}\partial_{\mu}\psi(x) - m\bar{\psi}(x)\psi(x). \tag{1.2.1}$$

It can be easily checked that $\mathcal{L}_{free}$ is invariant under *global* U(1) transformation

$$\psi(x) \to \psi'(x) = e^{iQ\theta}\psi(x), \tag{1.2.2}$$

where $Q\theta$ is an arbitrary real constant, by plugging the transformation 1.2.2 into $\mathcal{L}_{free}$:

$$\begin{aligned}\mathcal{L}_{free} &= i\bar{\psi}'(x)\gamma^{\mu}\partial_{\mu}\psi'(x) - m\bar{\psi}'(x)\psi'(x) \\ &= ie^{-iQ\theta}\bar{\psi}(x)\gamma^{\mu}\partial_{\mu}e^{iQ\theta}\psi(x) - me^{-iQ\theta}\bar{\psi}(x)e^{iQ\theta}\psi(x) \\ &= ie^{-iQ\theta}\bar{\psi}(x)\gamma^{\mu}e^{iQ\theta}\partial\psi(x) - m\bar{\psi}(x)\psi(x) \\ &= i\bar{\psi}(x)\gamma^{\mu}\partial_{\mu}\psi(x) - m\bar{\psi}(x)\psi(x).\end{aligned} \tag{1.2.3}$$

One would also like to have a similar behaviour of the Lagrangian if the phase θ was an explicit function of the space-time coordinate $\theta = \theta(x)$. However, this is not the case because

$$\partial\psi_{\mu}(x) \to e^{iQ\theta}\ (\partial_{\mu} + iQ\partial_{\mu}\theta)\psi(x). \tag{1.2.4}$$

If one wants the $U(1)$ phase invariance to hold locally, a requirement known as the "gauge principle", one has to add another piece to the Lagrangian in a way that an additional term in 1.2.4 ($\partial_\mu\theta$) will cancel out. This can be achieved by introducing a new spin-1 field $A_\mu(x)$ which transforms as

$$A_\mu(x) \to A'_\mu(x) \equiv A_\mu(x) + \frac{1}{e}\partial_\mu\theta \tag{1.2.5}$$

and replacing the usual derivative (∂_μ) with the *covariant derivative*

$$D_\mu\psi(x) \equiv [\partial_\mu - ieQA_\mu(x)]\,\psi(x) \tag{1.2.6}$$

that transforms in the same way as the field itself:

$$D_\mu\psi(x) \to (D_\mu\psi)'(x) \equiv e^{iQ\theta}D_\mu\psi(x). \tag{1.2.7}$$

The new Lagrangian

$$\mathcal{L} \equiv i\bar{\psi}(x)\gamma^\mu D_\mu\psi(x) - m\bar{\psi}(x)\psi(x) = \mathcal{L}_{free} + eQA_\mu(x)\bar{\psi}(x)\gamma^\mu\psi(x) \tag{1.2.8}$$

is now invariant under local $U(1)$ transformations. The second term in Eq. 1.2.8 defines an interaction between the Dirac spinor $\psi(x)$ and the gauge field A_μ. Finally, in order for A_μ to be a true propagating field, one needs to add a gauge-invariant kinetic term in the Lagrangian:

$$\mathcal{L}_{kin} = -\frac{1}{4}F_{\mu\nu}(x)F^{\mu\nu}(x), \tag{1.2.9}$$

where $F_{\mu\nu}(x) \equiv \partial_\mu A_\nu - \partial_\nu A_\mu$ is the electromagnetic field strength tensor.
The complete Lagrangian describing an electron and a massless vector boson (photon) of spin 1 can be written as

$$\mathcal{L}_{QED} = \mathcal{L}_{free} + eQA_\mu(x)\bar{\psi}(x)\gamma^\mu - \frac{1}{4}F_{\mu\nu}(x)F^{\mu\nu}(x)\psi(x). \tag{1.2.10}$$

One could be tempted to introduce a mass term $\frac{1}{2}m^2A_\mu A_\nu$, but this is not possible since the gauge invariance of the Lagrangian would be violated. As a result, the photon field remains massless. The Lagrangian that was derived gives raise to a set of equations

$$\partial_\mu F^{\mu\nu} = J^\nu, \tag{1.2.11}$$

where $J^\nu = -eQ\bar{\psi}\gamma^\nu\psi$ is the fermion electromagnetic current. These are known as Maxwell's equations for electromagnetism. Therefore, by only using gauge symmetry requirements, one can deduce the right QED Lagrangian from which the Maxwell equations follow. This points to the possibility that the QCD Lagrangian could be derived in a similar manner.

1.2.2 The Lagrangian of quantum chromodynamics

Let q_f^α be a quark flavor field f and a color charge α. Using a vector notation in the colour space, $q_f^T \equiv (q_f^1, q_f^2, q_f^3)$, the free QCD Lagrangian reads

$$\mathcal{L}_{free} = \sum_f \bar{q_f}(i\gamma^\mu\partial_\mu - m_f)q_f. \tag{1.2.12}$$

The Lagrangian is invariant under global $SU(3)_C$ transformations in colour space

$$q_f^\alpha \to (q_f^\alpha)' = U_\beta^\alpha q_f^\beta, \tag{1.2.13}$$

where U are unitary $SU(3)$ matrices that can be written as

$$U = e^{i\frac{\lambda^a}{2}\theta_a}. \tag{1.2.14}$$

Here, θ_a are arbitrary parameters, while $\frac{1}{2}\lambda^a$ $(a = 1, 2, ..., 8)$ are traceless Gell-Mann matrices that represent eight generators of the $SU(3)_C$ group. The matrices λ^a satisfy the commutation relations

$$\left[\frac{\lambda^a}{2},\ \frac{\lambda^b}{2}\right] = if^{abc}\frac{\lambda_c}{2}, \tag{1.2.15}$$

f^{abc} being the $SU(3)$ structure constant.
As was done in the derivation of the QED Lagrangian, one would like the QCD Lagrangian to be invariant under *local* $SU(3)_C$ transformations. Requiring again $\theta = \theta(x)$, one is forced to replace ordinary quark derivatives with covariant derivatives. Since the $SU(3)_C$ group has eight generators of symmetry, eight different gauge bosons, $G_a^\mu(x)$, are needed. These are identified with eight gluons that mediate the strong interaction. Hence,

$$D^\mu q_f \equiv \left[\partial^\mu - ig_s\frac{\lambda^a}{2}G_a^\mu(x)\right] q_f \equiv \left[\partial^\mu - ig_s G^\mu(x)\right] q_f, \tag{1.2.16}$$

where $[G^\mu(x)_{\alpha\beta}] \equiv \left(\frac{\lambda^a}{2}\right)_{\alpha\beta} G_a^\mu(x)$ and g is a dimensionless coupling strength.
Similarly to what was done in the QED case, one requires that covariate derivatives of the colour vectors, $D^\mu q_f$, transform as vectors themselves which fixes the transformation properties of the gauge fields:

$$\begin{aligned} D^\mu &\to (D^\mu)' = UD^\mu U^\dagger \\ G^\mu &\to (G^\mu)' = UG^\mu U^\dagger - \frac{i}{g_s}(\partial^\mu U)U^\dagger. \end{aligned} \tag{1.2.17}$$

One can show that, for the infinitesimal $SU(3)_C$ transformation, the gauge fields transform as

$$G_a^\mu \to (G_a^\mu)' = G_a^\mu + \frac{1}{g_s}\partial^\mu(\delta\theta_a) - f^{abc}(\delta\theta_b)G_c^\mu. \tag{1.2.18}$$

Because the $SU(3)_C$ group is not commutative, the gauge transformation of the gluon fields is more complicated than that obtained in QED for the photon field. In addition, the non-commutativity of the $SU(3)_C$ gives rise to an additional term involving the gluon fields themselves. Finally, the coupling constant, g, which describes the strength of the interaction between the gluon fields and quarks, is constant.

In order to construct a kinetic term for the gluon fields, the corresponding field strengths are introduced:

$$G^{\mu\nu}(x) \equiv \frac{i}{g_s}[D^\mu, D^\nu] = \partial^\mu G^\nu - \partial^\nu G^\mu - ig_s[G^\mu, G^\nu] \equiv \frac{\lambda^a}{2}G_a^{\mu\nu}(x), \tag{1.2.19}$$

where $G_a^{\mu\nu}(x) = \partial^\mu G_a^\nu - \partial^\nu G_a^\mu + g_s f^{abc}G_b^\mu G_c^\nu$
Using a gauge transformation

$$G^{\mu\nu} \to (G^{\mu\nu})' = UG^{\mu\nu}U^\dagger \tag{1.2.20}$$

and taking a proper normalization for the gluon kinetic term, one can derive the $SU(3)_C$-invariant QCD Lagrangian:

$$\mathcal{L}_{QCD} = \frac{1}{4} G_a^{\mu\nu} G^a_{\mu\nu} + \sum_f \bar{q}_f (i\gamma^\mu D_\mu - m_f) q_f. \tag{1.2.21}$$

By decomposing the Lagrangian into separated components, one can get a better insight into the structure of QCD:

$$\begin{aligned} \mathcal{L}_{QCD} = &- \frac{1}{4}(\partial^\mu G_a^\nu - \partial^\nu G_a^\mu)(\partial_\mu G^a_\nu - \partial_\nu G^a_\mu) + \sum_f \bar{q}_f^\alpha (i\gamma^\mu \partial_\mu - m_f) q_f^{alpha} \\ &+ g_s G_a^\mu \sum_f \bar{q}_f^\alpha \gamma_\mu \left(\frac{\lambda^a}{2}\right)_{\alpha\beta} q_f^\beta \\ &- \frac{g_s}{2} f^{abc} (\partial^\mu G_a^\nu - \partial^\nu G_a^\mu) G^b_\mu G^c_\nu - \frac{g_s^2}{4} f^{abc} f_{ade} G_b^\mu G_c^\nu G_\mu^d G_\nu^e. \end{aligned} \tag{1.2.22}$$

The first line contains the correct kinetic terms for the different fields which give rise to the quark propagators. The second line describes the interaction between quarks and gluons. The last line is a consequence of the non-Abelian structure of the $SU(3)_C$ group and includes the cubic and the quartic gluon self-interaction terms. The gluon self-interaction is a unique feature of the QCD theory whereby no such terms exist in the QED. This is the source of the emergent phenomena in the theory of QCD interactions such as the asymptotic freedom and colour confinement. The former ensures that the strong interaction becomes weaker at small distances, while the latter ensures that the strong interaction becomes stronger as quarks are being separated which results in only colour-neutral states to be observed in nature.

1.2.3 Unification of the electromagnetic and the weak interaction

In the last two sections, the application of gauge invariance on the $SU(3)_C$ and $U(1)$ group led to the Lagrangian of the gauge fields of the QED and QCD sectors of the SM. It then makes sense to try the same approach to obtain the Lagrangian of the electroweak interaction. It is known that left- and right-handed fields exhibit different behaviour. In addition, left-handed fermions appear in doublets, while right-handed fermions appear as singlet states. In addition, the theory should produce three massive gauge bosons, corresponding to $W^\pm$ and Z^0 bosons, as well as the massless photon field. The simplest symmetry group that satisfies these conditions is

$$SU(2)_L \otimes U(1)_Y, \tag{1.2.23}$$

where L and Y stand for *left* and *hypercharge*, respectively.
For a single family of quarks, we would have the following representation

$$\psi_1(x) = \begin{pmatrix} u \\ d \end{pmatrix}_L \qquad \psi_2(x) = u_R \qquad \psi_3(x) = d_R. \tag{1.2.24}$$

The same can be applied to leptons

$$\psi_1(x) = \begin{pmatrix} \nu_e \\ e^- \end{pmatrix}_L \qquad \psi_2(x) = \nu_{eR} \qquad \psi_3(x) = e^-_R. \tag{1.2.25}$$

We start from the free Lagrangian

$$\mathcal{L}_{free} = i\bar{u}(x)\gamma^\mu\partial_\mu u(x) + i\bar{d}(x)\gamma^\mu\partial_\mu d(x) = \sum_{j=1}^{j=3} i\bar{\psi}_j(x)\gamma^\mu\partial_\mu\psi_j(x) \tag{1.2.26}$$

which is invariant under global transformations in the flavour space:

$$\begin{aligned}
\psi_1(x) &\to \psi_1'(x) \equiv e^{iy_1\beta}U_L\psi_1(x),\\
\psi_2(x) &\to \psi_2'(x) \equiv e^{iy_2\beta}\psi_2(x),\\
\psi_3(x) &\to \psi_3'(x) \equiv e^{iy_3\beta}\psi_3(x).
\end{aligned} \tag{1.2.27}$$

Here, the parameters y_i are hypercharges and $U_L \equiv e^{i\frac{\sigma_i}{2}\alpha^i}$ $(i = 1, 2, 3)$ is the non-Abelian matrix representing the $SU(2)_L$ transformation acting on the doublet field ψ_1.
Similarly to the QED case, we require the Lagrangian to be invariant under local gauge transformations by having $\alpha^i = \alpha^i(x)$ and $\beta^i = \beta^i(x)$. The first step is to introduce the covariant derivative. Since there are four generators of symmetry for $SU(2)_L \otimes U(1)_Y$, there will also be four gauge parameters:

$$\begin{aligned}
D_\mu\psi_1(x) &\equiv \left[\partial_\mu - ig\tilde{W}_\mu(x) - ig'y_1B_\mu(x)\right]\psi_1(x),\\
D_\mu\psi_2(x) &\equiv \left[\partial_\mu - ig'y_2B_\mu(x)\right]\psi_2(x),\\
D_\mu\psi_3(x) &\equiv \left[\partial_\mu - ig'y_3B_\mu(x)\right]\psi_3(x).
\end{aligned} \tag{1.2.28}$$

For easier readability, $SU(2)_L$ matrix field $\widetilde{W}_\mu(x) = \frac{\sigma_i}{2}W_\mu^i(x)$ was introduced. The three gauge fields W_μ and the additional field B_μ give exactly four gauge fields as needed. However, these are not, at this point, identically identified with the $W^\pm$ and Z^0 bosons.
Similarly to what was done in the derivation of the QCD Lagrangian, one wants the $D_\mu\psi_j(x)$ to transform in the same manner as the $\psi_j(x)$ fields which fixes the transformation properties of the gauge fields:

$$\begin{aligned}
\widetilde{W}_\mu \to \widetilde{W}_\mu' &\equiv U_L(x)\widetilde{W}_\mu U_L^\dagger(x) - \frac{i}{g}\partial_\mu U_L(x)U_L^\dagger(x),\\
B_\mu(x) \to B_\mu'(x) &\equiv B_\mu(x) + \frac{1}{g'}\partial_\mu\beta(x),
\end{aligned} \tag{1.2.29}$$

where $U_L \equiv e^{i\frac{\sigma_i}{2}\alpha^i(x)}$. One can see that the B_μ field transform exactly in the same way as the photon field of QED, while the W_μ^i fields transform in a similar way to the gluon field of the QCD.
Finally, the free Lagrangian

$$\mathcal{L}_{free} = \sum_{j=1}^{j=3} i\bar{\psi}_j(x)\gamma^\mu D_\mu\psi_j(x) \tag{1.2.30}$$

is now invariant under local $SU(2)_L \otimes U(1)_Y$ gauge transformations. If one wants to build a gauge-invariant kinetic

term, one can introduce corresponding field strengths:

$$\begin{aligned} B_{\mu\nu} &\equiv \partial_\mu B_\nu - \partial_\nu B_\mu \\ \widetilde{W}_{\mu\nu} &\equiv \frac{\sigma_i}{2} W^i_{\mu\nu}, \end{aligned} \tag{1.2.31}$$

where $W^i_{\mu\nu} = \partial_\mu W^i_\nu - \partial_\nu W^i_\mu + g\epsilon^{ijk} W^j_\mu W^k_\nu$. The field $B_{\mu\nu}$ will remain invariant under the local gauge transformation, while $\widetilde{W}_{\mu\nu}$ will transform covariantly:

$$B_{\mu\nu} \to B_{\mu\nu}, \qquad \widetilde{W}_{\mu\nu} \to U_L \widetilde{W}_{\mu\nu} U_L^\dagger. \tag{1.2.32}$$

The kinetic part of the Lagrangian is then

$$\mathcal{L}_{kin} = -\frac{1}{4} B_{\mu\nu} B^{\mu\nu} - \frac{1}{4} W^i_{\mu\nu} W_i^{\mu\nu}. \tag{1.2.33}$$

Finally, the full Lagrangian for the electroweak interaction is then

$$\mathcal{L} = \sum_{j=1}^{j=3} i\bar{\psi}_j(x)\gamma^\mu D_\mu \psi_j(x) - \frac{1}{4} B_{\mu\nu} B^{\mu\nu} - \frac{1}{4} W^i_{\mu\nu} W_i^{\mu\nu} \tag{1.2.34}$$

Because the Lagrangian contains quadratic terms in $W^i_{\mu\nu}$, the cubic (ZWW, γWW) and quartic ($ZZWW$, γZWW, $\gamma\gamma WW$ and $WWWW$) self-interaction among gauge fields arise directly. The strength of these interactions is determined by the $SU(2)_L$ coupling g. One can notice that there is always, at least, a pair of charged W bosons in the self-interaction terms since the non-Abelian structure of the $SU(2)_L$ doesn't generate neutral vertices containing only photons and Z bosons. The Lagrangian contains the interaction of the fermion fields with the gauge bosons and the $W^\pm$ and the Z^0 boson are obtained through linear combinations of the gauge bosons.

$$\begin{aligned} W^\pm_\mu &= \frac{1}{\sqrt{2}} \left(W^1_\mu \mp W^2_\mu\right) \\ Z^0_\mu &= cos\theta_w W^3_\mu - sin\theta_w B_\mu \end{aligned}, \tag{1.2.35}$$

and the photon field A_μ

$$A_\mu = sin\theta_w W^3_\mu - cos\theta_w B_\mu. \tag{1.2.36}$$

Additionally, one can see that the gauge invariance forbids the massive fermionic fields since this would give raise to a mixture of the left- and right-handed fields through term $m(\bar{\psi}_L\psi_R + \bar{\psi}_R\psi_L)$ which would explicitly break the gauge symmetry of the Lagrangian. In the same manner, one cannot introduce the mass term for the gauge fields without explicitly breaking the gauge symmetry. Thus, the $SU(2)_L \otimes U(1)_Y$ only contains massless gauge fields. Before 1964, the origin of the gauge boson masses in the SM framework was one of the most urgent issues to be resolved since the experiments estimated the mass of the W and the Z bosons to be around $80.30\ GeV$ and $91.19\ GeV$ respectively. This was resolved through the mechanism of spontaneous symmetry breaking discussed in the next section.

1.3 Electroweak symmetry breaking

1.3.1 Spontaneous symmetry breaking and the Goldstone theorem

We start by introducing a complex scalar field ϕ, with Lagrangian

$$\mathcal{L} = \partial_\mu \phi^\dagger \partial^\mu \phi - V(\phi), \tag{1.3.1}$$

where

$$V(\phi) = \mu^2 \phi^\dagger \phi + h \left(\phi^\dagger \phi\right)^2, \tag{1.3.2}$$

and the Lagrangian is invariant under global phase transformations of the scalar field

$$\phi(x) \to \phi'(x) \equiv e^{i\theta} \phi(x). \tag{1.3.3}$$

In order to find the ground state of the potential one can simply evaluate

$$\frac{\partial V(\psi)}{\partial \phi} \equiv \mu^2 \phi^\dagger + 2h\phi^\dagger \phi \phi^\dagger = 0 \Longrightarrow \phi^\dagger \left(\mu^2 + 2h\phi^\dagger \phi\right) = 0. \tag{1.3.4}$$

In order to have a ground state one must bound it from below and thus $h > 0$. The Eq. (1.3.4) can then only hold if

1. $\mu^2 > 0$: $|\phi_0| = 0$
2. $\mu^2 < 0$: $|\phi_0| = \sqrt{\frac{-\mu^2}{2h}} \equiv \frac{v}{\sqrt{2}} > 0$.

The first solution is a trivial one which describes a scalar particle with mass μ and coupling h. This solution corresponds to the potential shape shown on the left-hand side of Fig. 1.2.
The solution with $\mu < 0$ is more interesting and is shown on the right-hand side of the same figure. Due to the phase invariance of the Lagrangian, there is an infinite number of degenerate states of minimum energy, $\phi_0 = \frac{v}{\sqrt{2}} e^{i\theta}$, each corresponding to a single point in the "minima" valley depicted as the dashed circle in the figure. By choosing any specific solution, the symmetry of the ground state will be broken. This is referred to as *spontaneous symmetry breaking*. One can choose $\theta = 0$ and parametrize the excitations above the ground state as

$$\phi(x) = \frac{1}{\sqrt{2}} \left[v + \phi_1(x) + i\phi_2(x)\right], \tag{1.3.5}$$

where ϕ_1 and ϕ_2 are real fields. The field ϕ_1 corresponds to the oscillations in the radial direction around the specified minimum. On the other side, the field ϕ_2 corresponds to the oscillation in the angular direction around the specified minimum, i.e. along the dashed circle in the figure. This then results in the potential of the form

$$V(\phi) = V(\phi_0) - \mu^2 \phi_1^2 + hv\phi_1(\phi_1^2 + \phi_2^2) + \frac{h}{4}(\phi_1^2 + \phi_2^2)^2. \tag{1.3.6}$$

One can immediately notice that the mass term, $m_{\phi_1}^2 = -2\mu^2$, for the ϕ_1 field, describing the radial oscillations, pops out. At the same time, an additional massless field, ϕ_2, associated with the angular oscillations around the minimum, emerges as a consequence of the spontaneous symmetry breaking. In other terms, in addition to a massive particle, a massless particle emerged as a result of spontaneously broken symmetry.

This finding is generalized through *Nambu-Goldstone theorem*: if the Lagrangian is invariant under a continuous group with M generators, but the vacuum is invariant only under a subgroup with N generators $(M > N)$, then

there must appear $M - N$ massless, spin-0 particles. In other words, one massless, spin-0 particle must appear for each generator of the symmetry that was lost. This particle is known as the *Nambu-Goldstone boson*, or, simply, the *Goldstone boson*. The idea of the spontaneous symmetry breaking, and, consequently, the emergence of Goldstone bosons, is central to the generation of masses of the $W^{\pm}$ and Z^0 bosons. This is discussed in the next section.

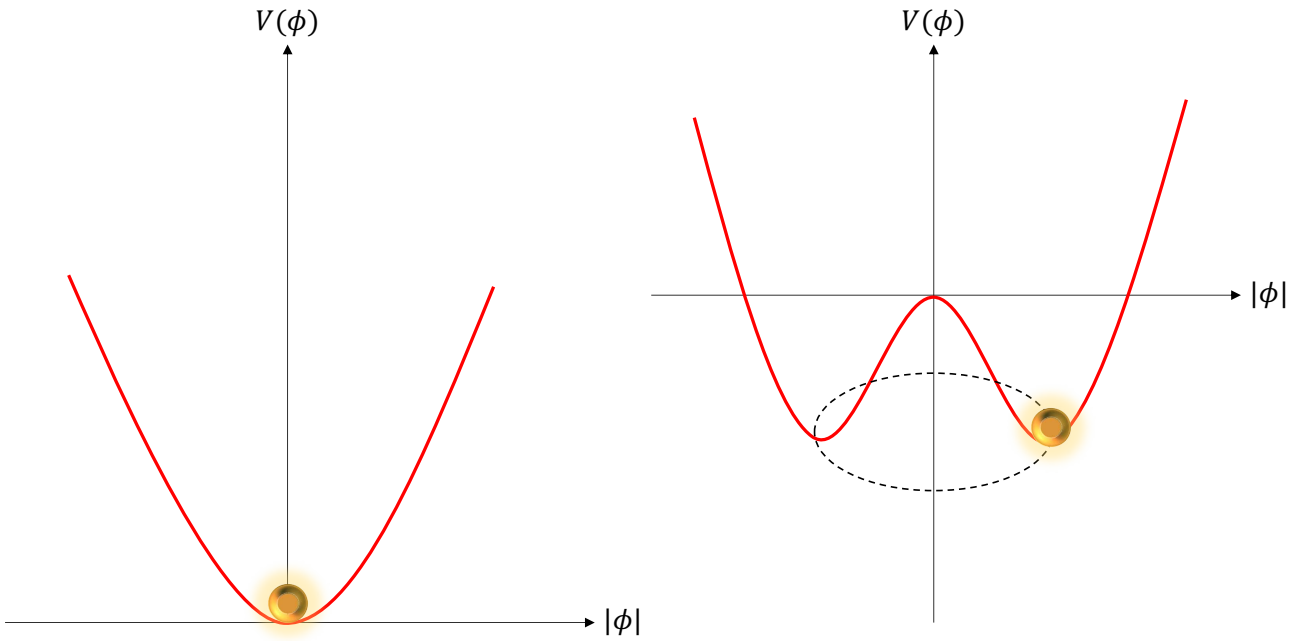


Figure 1.2: The shape of the scalar potential for $\mu^2 > 0$ (left) and $\mu^2 < 0$. The latter features an infinite set of degenerate vacua, corresponding to different phases θ, connected through a massless field excitation ϕ_2.

1.3.2 The Brout-Englert-Higgs mechanism

In order to explain the origin of masses of the weak gauge bosons, one must consider a doublet of complex scalar fields

$$\phi(x) = \begin{pmatrix} \phi^{(+)} = \phi_1 + i\phi_2 \\ \phi^{(0)} = \phi_3 + i\phi_4 \end{pmatrix}. \tag{1.3.7}$$

The two components of the charged field, ϕ_1 and ϕ_2 will give raise to two Goldstone bosons that will be incorporated into two massive W bosons, while the ϕ_4 component of the neutral field will give raise to a third Goldstone boson that will be incorporated into a massive Z boson.
The corresponding Lagrangian for the the doublet of complex scalar fields than reads

$$\mathcal{L} = (D_\mu \phi)^\dagger D^\mu \phi - \mu^2 \phi^\dagger \phi - h(\phi^\dagger \phi)^2, \tag{1.3.8}$$

where

$$D^\mu \phi = \left[\partial^\mu - ig\widetilde{W}^\mu - \frac{ig'}{2} B^\mu \right] \phi, \tag{1.3.9}$$

is invariant under local $SU(2)_L \otimes U(1)_Y$ symmetry. As before, one requires that the potential be bound from bellow so that $h > 0$ and by choosing $\mu < 0$ obtains the potential similar to the one considered before. This gives rise to an

infinite set of degenerate ground states defined by

$$|\phi_0^{(0)}| = \sqrt{\frac{-\mu^2}{2h}} \equiv \frac{v}{\sqrt{2}}, \tag{1.3.10}$$

where one must bear in mind that only a neutral field can acquire a vacuum expectation value due to electric charge conservation. By choosing any specific ground state one spontaneously breaks the $SU(2)_L \otimes U(1)_Y$ symmetry into electromagnetic subgroup $U(1)_Y$. Since the original symmetry with four generators has been broken into the symmetry with only one generator, the Goldstone theorem mandates that three Goldstone bosons must appear. One can now, similarly to before, parametrize the excitations above the ground state as

$$\phi(x) = e^{i\frac{\sigma_i}{2}\theta^i(x)} \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + H(x) \end{pmatrix}, \tag{1.3.11}$$

where σ^i are generators of the $SU(2)_L$ algebra, $\theta^i(x)$ are the three massless Goldstone bosons, and $H(x)$ is the *Higgs field* whose excitation corresponds to the *Higgs boson*.
If one chooses the unitary gauge $\theta^i(x) = 0$, the kinetic part of the Lagrangian in Eq. 1.3.8 then becomes

$$(D_\mu \phi)^\dagger) D^\mu \phi \to \frac{1}{2}\partial_\mu H \partial^\mu H + (v + H)^2 \left[\frac{g^2}{4} W_\mu^\dagger W^\mu + \frac{g^2}{8cos^2\theta_w} Z_\mu Z^\mu \right]. \tag{1.3.12}$$

As one can notice, the massless Goldstone boson fields have been incorporated into new, massive, W and Z boson fields as a consequence of the non-vanishing vacuum value of the neutral scalar field and after a choice of an appropriate gauge requirement. After the spontaneous breaking of the electroweak symmetry (EWSB), the mass of the electroweak bosons is

$$\begin{aligned} M_Z &= \frac{vg}{2cos\theta_w} \\ M_W &= \frac{1}{2} vg. \end{aligned} \tag{1.3.13}$$

The photon remained massless after the EWSB because the $U(1)_{QED}$ is an unbroken symmetry. Before the EWSB the Lagrangian contained massless $W^\pm$ and Z^0 bosons which gives $3 \times 2 = 6$ degrees of freedom since massless, spin-1 fields can only have two values of polarization, 1 and -1, corresponding to the two transverse polarizations. However, after the EWSB, three Goldstone bosons have been "eaten" by the weak bosons giving them mass and, consequently, an additional degree of freedom: longitudinal polarization.

1.4 Vector boson scattering

A new feature emerging from the EWSB mechanism is the longitudinal polarization of massive gauge bosons in the weak sector. By comparing the polarization vectors of the transversely polarized vector bosons

$$\begin{aligned} \epsilon_+^\mu &= \frac{1}{\sqrt{2}}(0\ 1\ i\ 0)^\mu \\ \epsilon_-^\mu &= \frac{1}{\sqrt{2}}(0\ 1\ -i\ 0)^\mu, \end{aligned} \tag{1.4.1}$$

where ϵ_+^μ and ϵ_-^μ correspond to the right and left helicity states of the transverse polarization, respectively,

to the polarization vector of the longitudinally polarized vector boson of mass m and momentum $k^\mu = \frac{1}{m}(k_z\ 0\ 0\ E)^\mu$

$$\epsilon_L^\mu = \frac{1}{m}(k_z\ 0\ 0\ E)^\mu, \tag{1.4.2}$$

one notices a striking difference between the two. While the transverse components remain constant as the scattering energy increases, the longitudinal component scales with scattering energy as E/m. The reason for the difference in the high-energy behaviour of the two polarizations stems from the different origins of the two. The transverse polarization exists in the theory of the weak sector prior to the EWSB and corresponds to the massless gauge bosons. On the other hand, the longitudinal component of the polarization is the consequence of incorporating the Goldstone bosons of the EWSB into the gauge boson fields as a result of local symmetry and unitarity gauge requirement.

The difference in the behaviour of the two polarizations in the high-energy limits implies that, at high energies, a longitudinal component can be disentangled from the transverse component. While the longitudinal states are equivalent to the Goldstone Bosons of the EWSB, the transverse states correspond to the original electroweak gauge bosons. Using the Goldstone boson equivalence theorem [1–3], one can say that the scattering of the longitudinal vector bosons at high energy is equivalent to the scattering of Goldstone bosons.
The importance of this statement becomes clear if one looks, for example, at the scattering amplitude of the $W_L^+W_L^- \to W_L^+W_L^-$ process [4]:

$$\mathcal{M}_{SM}(W_L^+W_L^- \to W_L^+W_L^-) \approx \mathcal{M}_{SM}(w^+w^- \to w^+w^-) \approx -i\frac{m_H^2}{v^2}\left[2 + \frac{m_H^2}{s - m_H^2} + \frac{m_H^2}{t - m_H^2}\right], \tag{1.4.3}$$

where $w^\pm$ are the Goldstone bosons, and s and t are the Mandelstam variables.
In the high-energy limit where $s,\ |t| >> m_H^2$, this expression becomes constant and the cross section falls proportionally with the scattering energy ($\sigma \sim \frac{1}{s}$).
On the other hand, without the Higgs boson in the SM ($m_H \to \infty$), the matrix element becomes

$$\mathcal{M}_{Higgsless}(W_L^+W_L^- \to W_L^+W_L^-) \approx i\frac{s+t}{v^2}. \tag{1.4.4}$$

This shows that without the Higgs boson, in the high-energy limit, the cross section will diverge and, therefore, the unitarity of the theory will be violated. This points to the significance of the cancellations between the contributions from pure gauge diagrams and Higgs interactions. Moreover, not only does the unitarity violation occur if there is no Higgs boson, but it occurs also around the energy of $\approx 1.2\ TeV$ if the couplings of the Higgs bosons to the vector bosons differ from the SM prediction [2,5]. This result was the primary argument for the yet-unobserved physics at the TeV scale. It showed that either the Higgs boson will have to be found at the LHC, or some other phenomena would have to appear at the TeV scale in order to preserve unitarity.

With the Higgs boson discovered, the problem of the unitarity violation was resolved given that the coupling of the Higgs boson to the vector bosons is as predicted by the SM. Thus, the scattering of longitudinal vector bosons proves to be an important tool for probing the scalar sector of the SM and studying the EWSB mechanism. In addition, VBS enables one to study the non-Abelian structure of the electroweak (EW) sector by probing the quartic vertices. Finally, a beyond SM (BSM) phenomena could manifest themselves, for example, in modifications to the quartic gauge couplings that increase the production cross section. This will be discussed, in more detail, in section 1.4.2.

1.4.1 Characteristics of VBS processes

A typical example of a VBS process are two gauge bosons mediated by the two separate quark lines which then interact. VBS is defined at a tree-level as $\mathcal{O}(\alpha^6)$ process when including the decay products of the heavy gauge

bosons. Depending on the decay products of the two vector bosons, VBS processes may be looked for through the *fully leptonic* decay channel with four leptons and two hadronic jets in the final state, the *semi-leptonic* channel with two leptons and four jets in the final state and the *fully hadronic* decay channel with six jets in the final state. Some example Feynman diagrams of the processes that lead to the $4l2j$ final state are shown in Figs. 1.3 - 1.6 where the vector bosons are denoted as V, fermions with f and quarks q.

Fig. 1.3 shows representative Feynman diagrams for the VBS production of the $4l2j$ final state. The top row shows the EW VBS production through the quartic (top-left) and two trilinear (top-right) coupling diagrams. The bottom diagram shows the production of the same final state through the t-channel exchange of the Higgs boson. The latter ensures unitarity through Higgs boson coupling to the vector boson fields.

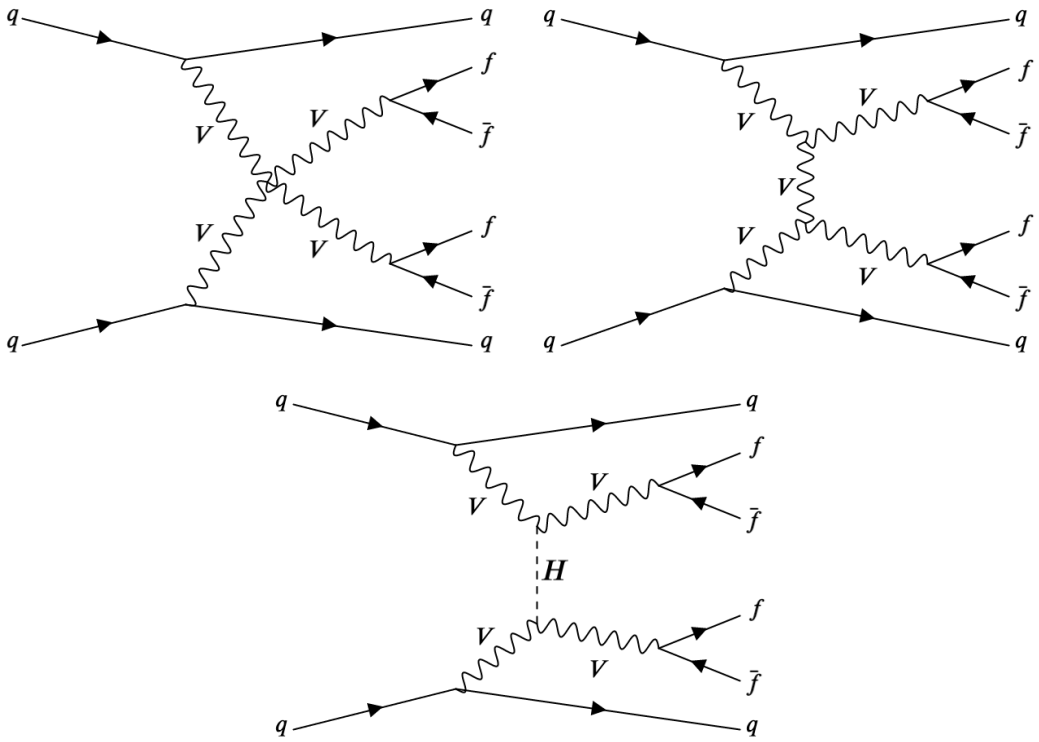


Figure 1.3: Representative Feynman diagrams for the VBS production of the *VVjj* final state with a scattering topology including the quartic (top-left) and two trilinear (top-right) vertices together with the t-channel exchange of the Higgs boson.

Fig. 1.4 represents the non-VBS production that can be suppressed by proper selection criteria. These criteria are defined in Chapter 4. The left-hand side diagram is an example of an off-shell boson splitting into two final state leptons, while on the right-hand side diagram one vector boson is radiated from the quark line. Both diagrams must be included in the simulations in order to ensure the gauge invariance.

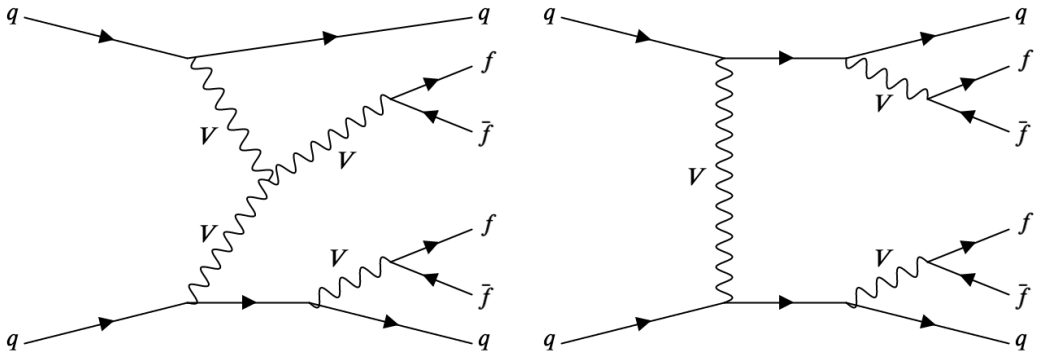


Figure 1.4: Representative Feynman diagrams for the non-VBS production of the VVjj final state.

Fig. 1.5 shows the pure EW diagrams that are not relevant for the study presented in this thesis and can be suppressed by appropriate phase space selection. The diagram shown on the left-hand side is an example of the non-resonant production of the 4l2j final state, while the right-hand side shows the triboson production where one of the gauge bosons decays hadronically. The hadronic jets originating from such a process will have the dijet mass around 80 or 91 GeV. For this reason, the $m_{jj} > 100\ GeV$ cut will be applied in the analysis.

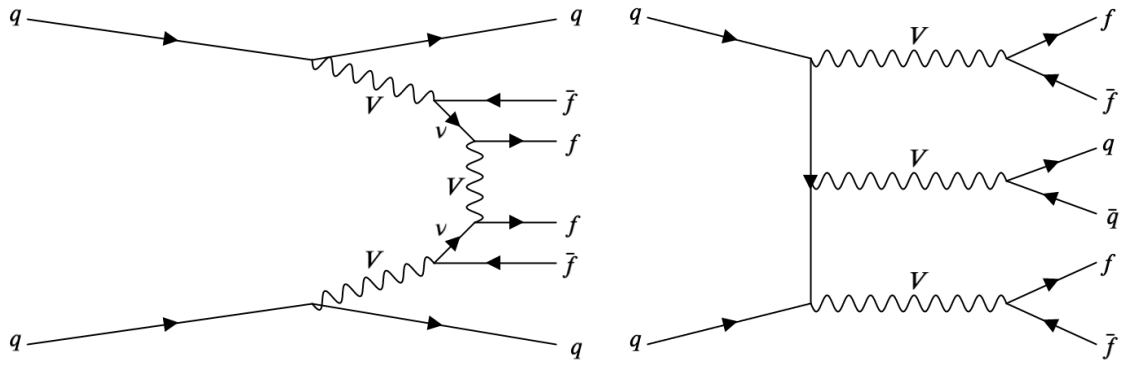


Figure 1.5: Feynman diagrams for the non-resonant (left) and triboson (right) production of the *4l2j* final state. These processes can be suppressed by appropriate phase space selection.

Finally, Fig. 1.6 shows two examples of the QCD-induced production of the $4l2j$ final state.

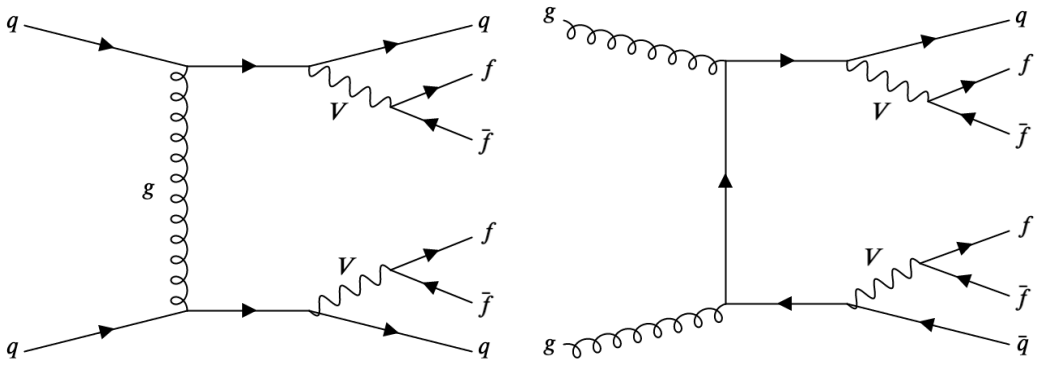


Figure 1.6: Representative Feynman diagrams for the QCD-induced production of the *VVjj* final state.

In addition to the purely EW contributions at order $\mathcal{O}(\alpha^6)$, VVjj final states also include reducible contributions of order $\mathcal{O}(\alpha^5\alpha_s)$ and $\mathcal{O}(\alpha^4\alpha_s^2)$. These are referred to as the *VBS signal*, *interference* and *QCD background* respectively. Because EW and QCD contributions behave differently, the maxima of dijet mass distributions will peak at different parts of the detector. This is shown in Fig. 1.7.

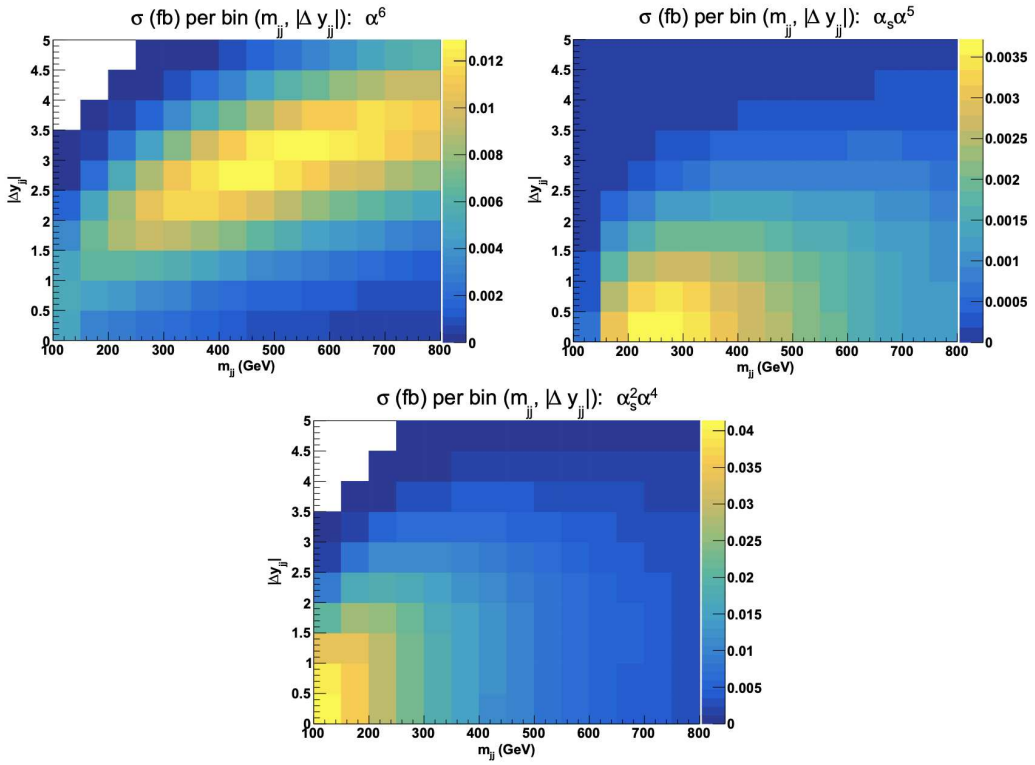


Figure 1.7: Double-differential distributions in the variables m_{jj} and $|\Delta y_{jj}|$ for the three LO contributions of orders $\mathcal{O}(\alpha^6)$ (top-left), $\mathcal{O}(\alpha^5\alpha_s)$ (top-right) and $\mathcal{O}(\alpha^4\alpha_s^2)$ (bottom) corresponding to the VBS signal, interference and QCD background contribution respectively. The figure is taken from [6] where one can also find the details on the selection criteria which was applied.

Since the EW contribution doesn't feature QCD exchanges between the two quark lines while the QCD component does, the differential cross section as a function of the dijet invariant mass or the rapidity difference between the two outgoing hadronic jets is different for the two components.
As one can see, the distinctive feature of the VBS processes are the two hadronic jets (referred to as the tagging jets) with large invariant masses and a large pseudorapidity gap between them. The latter can be also seen by looking at the expression for the square of the scattering amplitude:

$$|\mathcal{A}|^2 \sim \frac{p_1 \cdot p_2 \cdot p_3 \cdot p_4}{(q_1^2 - m_Z^2)^2 (q_2^2 - m_Z^2)^2}, \tag{1.4.5}$$

where p_1 and p_2 are the momenta of the incoming quarks, p_3 and p_4 are the momenta of the outgoing quarks, and $q_1 = p_1 - p_3$ and $q_2 = p_2 - p_3$ are the momenta of the intermediate gauge bosons.
The scattering amplitude is large if $m_{jj} \equiv p_3 \cdot p_4$ is large. One can show that

$$m_{jj} \approx 2 \cdot p_T(j_1) p_T(j_2) \left[cosh(\eta_{j_1} - \eta_{j_2}) - cos(\phi_{j_1} - \phi_{j_2}) \right]. \tag{1.4.6}$$

Given the constant momenta of the outgoing jets, this expression is largest when the pseudorapidity gap between the jets is large and when the jets are back-to-back ($\phi_{j_1} - \phi_{j_2} \to \pi$). This agrees with conclusions inferred from the distributions shown in Fig. 1.7. Fig. 1.8 shows the event display with a typical VBS topology in which a W and a Z boson are produced in association with two forward jets.

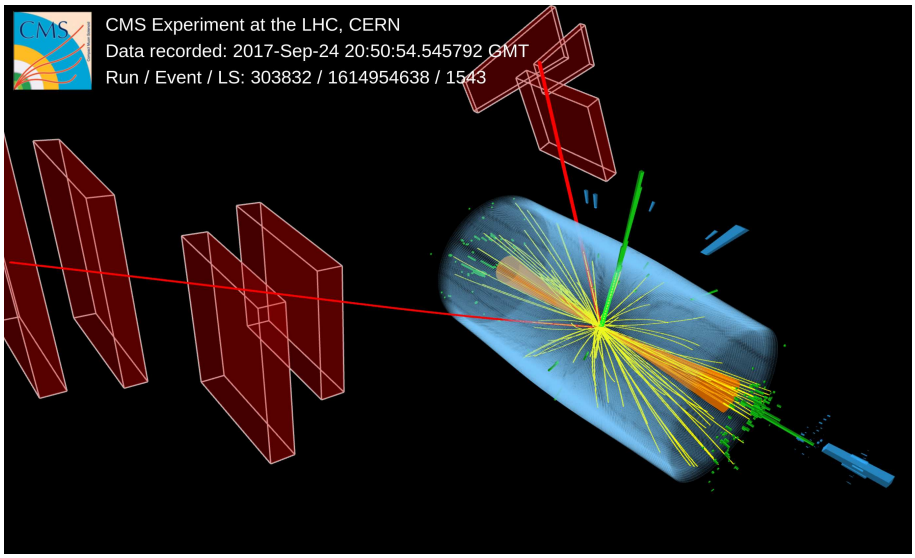


Figure 1.8: Event display with a typical VBS topology in which a W and a Z boson are produced in association with two forward jets. Trajectory and energy depositions of an electron from a W boson decay are shown in green, while the trajectory and energy depositions of two muons from a Z boson decay are shown in red. The two hadronic jets are represented by the orange cones. The illustration is taken from Ref. [7].

Additionally, the expression for the square of the scattering amplitude can be large if the denominator is small which occurs for small values of q_i. It can be shown that the square of q_1 can be written in terms of the scattering angle, θ_1, between $\vec{p}_1$ and $\vec{p}_3$, the energy of the incoming (E_1) and outgoing (E_3) quarks, and the transverse momentum of the outgoing quark ($p_{T,3}$):

$$q_1^2 = -\frac{2}{1 + cos\theta_1} \frac{E_1}{E_3} p_{T,3}^2. \tag{1.4.7}$$

This expression is smallest when the scattering angle is small ($\theta_1 \to 0$) or when the transverse momentum of the outgoing quark is small. Since the quarks will recoil against the vector bosons upon radiation, and since enough energy is needed to create the on-shell Z boson of the final state, the p_T of the outgoing jets will be of the order of the Z boson mass $p_T(j) \approx m_Z$.

An additional feature of the VBS processes is the kinematics of the vector bosons with respect to the tagging jets. While the jets are found, preferably, at low scattering angles, the gauge bosons tend to be found in the pseudorapidity gap between them.
Finally, due to the absence of colour flow between the interacting partons, hadron activity in the central region of the detector is suppressed [8].

1.4.2 Effective field theory

As was discussed in section 1.2.3, the $SU(2)_L \otimes U(1)_Y$ gauge symmetry gives rise to the trilinear and quartic gauge couplings of the vector bosons. Therefore, studying these interactions can further confirm the theoretical predictions, or point to some deviations from SM predictions that would give a hint to possible new physics at a higher scale. In particular, modifications of the vector boson couplings, either amongst themselves, or to the Higgs boson, could result in imperfect cancellation between the Feynman amplitudes including quartic gauge boson interactions, trilinear gauge boson couplings, and Higgs exchange. This would result in the increase of the cross section with energy that could be observed as an excess of events compared to the SM prediction.

One way to search for beyond SM (BSM) physics exploits the effective field theory (EFT), which comes in two flavours. In the model-dependent *top-down* approach one starts with an ultraviolet-complete (UV-complete) theory, finds its low-energy behaviour, and, finally, tries to match it to the SM. On the other hand, in the *bottom-up* approach one starts with the SM and builds towards the UV regime. Although this approach does not provide concrete predictions for the BSM scenarios, it gives us tools to search for new physics through modifications of certain gauge couplings. [9–12].

When building the bottom-up approach one starts from the SM Lagrangian with underlying $SU(3)_C \otimes SU(2)_L \otimes U(1)_Y$ symmetry with all operators with dimension of up to four. Next, one seeks to extend the theory by adding operators of higher dimensions with coefficients of inverse power of mass, therefore, lifting the restriction of renormalizability. The EFT Lagrangian can be written as

$$\mathcal{L}_{EFT} = \mathcal{L}_{SM} + \sum_{d>4} \sum_{i} \frac{f_i^{(d)}}{\Lambda^{d-4}} \mathcal{O}_i^{(d)}, \tag{1.4.8}$$

where $\mathcal{O}_i^{(d)}$ are the d-dimensional BSM operators invariant under the symmetries of the SM, f_i are the corresponding Wilson coefficients or *coupling strengths* and Λ is the typical scale of new physics. Any evidence for a non-zero Wilson coefficient would represent a clear sign of new physics.
The dominant operator in the expansion will be the one with dimension five and is responsible for generating Majorana masses for neutrinos [13]. However, all odd-power operators lead to lepton or baryon number violation and are not considered here [14]. We are, thus, left with dimension-6 operators followed by dimension-8 operators.
The former are responsible for the emergence of anomalous triple gauge couplings (aTGCs), while the latter give rise to anomalous quartic gauge couplings (aQGCs). The study presented in this thesis will focus on the operators that modify the quartic gauge couplings while simultaneously leaving trilinear gauge couplings unchanged. The reason is that the VBS channel explored in this thesis is most sensitive to probing aQGCs. The list of all aQGC operators in the

linear Higgs-doublet representation [15] can be found in Table 1.1. The modified field strength tensors are given by

$$\begin{aligned} \hat{W}^{\mu\nu} &= ig_w \frac{\sigma^j}{2} W^{j,\,\mu\nu} \\ \hat{B}^{\mu\nu} &= \frac{ig}{2} B^{\mu\nu}. \end{aligned} \tag{1.4.9}$$

Table 1.2 shows which vertices are modified by individual operators.
The aQGCs that only involve the EW fields are given by the tensor operators $\mathcal{O}_T$, whereby the $ZZjj$ channel explored in this thesis work is most sensitive to the charged-current operators $\mathcal{O}_{T,0,1,2}$ as well as the neutral-current operators $\mathcal{O}_{T,8,9}$.

Class	Definition
Scalar involve only the scalar field	$\mathcal{O}_{S,0} = \left[(D_\mu\phi)^\dagger D_\nu\phi\right] \times \left[(D^\mu\phi)^\dagger D^\nu\phi\right]$
	$\mathcal{O}_{S,1} = \left[(D_\mu\phi)^\dagger D_\mu\phi\right] \times \left[(D_\nu\phi)^\dagger D^\nu\phi\right]$
	$\mathcal{O}_{S,2} = \left[(D_\mu\phi)^\dagger D_\nu\phi\right] \times \left[(D^\nu\phi)^\dagger D^\mu\phi\right]$
Tensor involve only the field strength tensor	$\mathcal{O}_{T,0} = Tr\left[\hat{W}_{\mu\nu},\, \hat{W}^{\mu\nu}\right] \times Tr\left[\hat{W}_{\alpha\beta},\, \hat{W}^{\alpha\beta}\right]$
	$\mathcal{O}_{T,1} = Tr\left[\hat{W}_{\alpha\nu},\, \hat{W}^{\mu\beta}\right] \times Tr\left[\hat{W}_{\mu\beta},\, \hat{W}^{\alpha\nu}\right]$
	$\mathcal{O}_{T,2} = Tr\left[\hat{W}_{\alpha\mu},\, \hat{W}^{\mu\beta}\right] \times Tr\left[\hat{W}_{\beta\nu},\, \hat{W}^{\nu\alpha}\right]$
	$\mathcal{O}_{T,5} = Tr\left[\hat{W}_{\mu\nu},\, \hat{W}^{\mu\nu}\right] \times \hat{B}_{\alpha\beta}\hat{B}^{\alpha\beta}$
	$\mathcal{O}_{T,6} = Tr\left[\hat{W}_{\alpha\nu},\, \hat{W}^{\mu\beta}\right] \times \hat{B}_{\mu\beta}\hat{B}^{\alpha\nu}$
	$\mathcal{O}_{T,7} = Tr\left[\hat{W}_{\alpha\mu},\, \hat{W}^{\mu\beta}\right] \times \hat{B}_{\beta\nu}\hat{B}^{\nu\alpha}$
	$\mathcal{O}_{T,8} = \hat{B}_{\mu\nu}\hat{B}^{\mu\nu} \times \hat{B}_{\alpha\beta}\hat{B}^{\alpha\beta}$
	$\mathcal{O}_{T,9} = \hat{B}_{\alpha\mu}\hat{B}^{\mu\beta} \times \hat{B}_{\beta\nu}\hat{B}^{\nu\alpha}$
Mixed involve the field strength tensor and the scalar field	$\mathcal{O}_{M,0} = Tr\left[\hat{W}_{\mu\nu},\, \hat{W}^{\mu\nu}\right] \times \left[(D_\beta\phi)^\dagger D^\beta\phi\right]$
	$\mathcal{O}_{M,1} = Tr\left[\hat{W}_{\mu\nu},\, \hat{W}^{\nu\beta}\right] \times \left[(D_\beta\phi)^\dagger D^\mu\phi\right]$
	$\mathcal{O}_{M,2} = \hat{B}_{\mu\nu}\hat{B}^{\mu\nu} \times \left[(D_\beta\phi)^\dagger D^\beta\phi\right]$
	$\mathcal{O}_{M,3} = \hat{B}_{\mu\nu}\hat{B}^{\nu\beta} \times \left[(D_\beta\phi)^\dagger D^\mu\phi\right]$
	$\mathcal{O}_{M,4} = (D_\mu\phi)^\dagger \hat{W}_{\beta\nu} D^\mu\phi \times \hat{B}^{\beta\nu}$
	$\mathcal{O}_{M,5} = (D_\mu\phi)^\dagger \hat{W}_{\beta\nu} D^\nu\phi \times \hat{B}^{\beta\mu}$
	$\mathcal{O}_{M,7} = (D_\mu\phi)^\dagger \hat{W}_{\beta\nu}\hat{W}^{\beta\mu} D^\nu\phi$

Table 1.1: Scalar, tensor and mixed dimension-eight operators in the EFT approach. Limits on the Wilson coefficients for the $\mathcal{O}_{T,0,1,2}$ as well as the $\mathcal{O}_{T,8,9}$ operators are derived in chapter 4. The table is taken from [16].

	$\mathcal{O}_{S,0}$ $\mathcal{O}_{S,1}$ $\mathcal{O}_{S,22}$	$\mathcal{O}_{M,0}$ $\mathcal{O}_{M,1}$ $\mathcal{O}_{M,7}$	$\mathcal{O}_{M,2}$ $\mathcal{O}_{M,3}$ $\mathcal{O}_{M,4}$ $\mathcal{O}_{M,5}$	$\mathcal{O}_{T,0}$ $\mathcal{O}_{T,1}$ $\mathcal{O}_{T,2}$	$\mathcal{O}_{T,5}$ $\mathcal{O}_{T,6}$ $\mathcal{O}_{T,7}$	$\mathcal{O}_{T,8}$ $\mathcal{O}_{T,9}$
$WWWW$	✓	✓		✓		
$WWZZ$	✓	✓	✓	✓	✓	
$ZZZZ$	✓	✓	✓	✓	✓	✓
$WWZ\gamma$		✓	✓	✓	✓	
$WW\gamma\gamma$		✓	✓	✓	✓	
$ZZZ\gamma$		✓	✓	✓	✓	✓
$ZZ\gamma\gamma$		✓	✓	✓	✓	✓
$Z\gamma\gamma\gamma$				✓	✓	✓
$\gamma\gamma\gamma\gamma$				✓	✓	✓

Table 1.2: List of vertices modified by a given aQGC operator

1.5 Overview of the experimental searches for vector boson scattering

The following section presents a chronological overview of the most important results, obtained by the CMS and ATLAS collaborations, on the vector boson scattering in different channels and centre-of-mass energies. This section will help the reader understand the progress in the field and will put in perspective the work presented in this thesis. For brevity's sake, many details are omitted. An interested reader can find an in-depth discussion on the selection criteria, fiducial region definitions, signal extraction methods and other details in the corresponding papers. A summary of results discussed in this section is given in Table 1.3.

The first results on the scattering of two vector bosons in the VBS topology channels were reported by the CMS and ATLAS collaborations in 2014 at $8\ TeV$ centre-of-mass energy.
CMS reported an observed (expected) significance of 2.0 (3.1) standard deviations for the **same-sign W boson** production accompanied by two hadronic jets with an integrated luminosity of $19.4\ fb^{-1}$. In addition, the cross section in fiducial region for $W^{\pm}W^{\pm}$ and WZ processes was also measured giving $\sigma_{fid}(W^{\pm}W^{\pm}jj) = 4.0^{+2.4}_{-2.0}(stat)^{+1.1}_{-1.0}(syst)\ fb$ with an expectation of $5.8 \pm 1.2\ fb$ for the former, and $\sigma_{fid}(WZjj) = 10.8 \pm 4.0(stat) \pm 1.3(syst)\ fb$ with an expectation of $14.4 \pm 4.0\ fb$ for the latter. Limits on aQGC operators S_0, S_1, M_0, M_1, M_6, M_7, T_0, T_1 and T_2 were reported as well [17].
The ATLAS collaboration reported the first evidence for the $\boldsymbol{W^{\pm}W^{\pm}jj}$ production and **electroweak-only** $\boldsymbol{W^{\pm}W^{\pm}jj}$ production with observed significance of 4.5 and 3.6 standard deviations respectively at an integrated luminosity of $20.3\ fb^{-1}$. In addition, the cross section measurements in the two fiducial regions were reported as well: inclusive and VBS. The former is defined by requiring $p_T^l > 25\ GeV$, $|\eta| < 2.5$ and $\Delta R_{ll} > 0.3$. In addition, two jets with $p_T > 30\ GeV$ and $||\eta < 4.5$, separated from leptons by $\Delta R_{jl} > 0.3$ are required. Jets are also required to have an invariant mass above $500\ GeV$. The VBS region is defined by requiring the two jets with the largest p_T to be separated in rapidity by $|\Delta y_{jj} > 2.4|$. Cross sections of $\sigma_{fid} = 2.1 \pm 0.5(stat) \pm 0.3(syst)\ fb$ in the inclusive and $\sigma_{fid} = 1.3 \pm 0.4(stat) \pm 0.2(syst)\ fb$ in the VBS region are reported. The measured cross section in the VBS region was used to set limits on aQGCs affecting vertices with four interacting W bosons [18].

Year	Channel	Collaboration	$\sqrt{s}\,[TeV]$	$\mathcal{L}\,[fb^{-1}]$	Comment
2014	SS $W^{\pm}W^{\pm}jj$	CMS	8	19.4	first VBS results in CMS
2014	SS $W^{\pm}W^{\pm}jj$	ATLAS	8	20.3	first VBS results in ATLAS
2016	$W^{\pm}Z$	ATLAS	8	20.3	
2016	SS $W^{\pm}W^{\pm}jj$	ATLAS	8	20.3	35 % improvement in sensitivity to gague coupling parameters α_4 and α_5 w.r.t. the previous ATLAS result
2016	$W\gamma jj$	CMS	8	19.7	the first limits on the gauge coupling parameters $f_{M,4}$ and $f_{T,5-7}$
2017	$Z\gamma jj$	CMS	8	20	first evidence
2017	$Z\gamma jj$	ATLAS	8	20.2	
2017	SS $W^{\pm}W^{\pm}jj$	CMS	13	35.9	first observation
2017	$ZZjj$	CMS	13	35.9	first measurement
2018	$W^{\pm}Z$	ATLAS	13	36.1	first observation
2019	$W^{\pm}Z$	CMS	13	35.9	
2019	SS $W^{\pm}W^{\pm}jj$	ATLAS	13	36.1	
2019	$WW/WZ/ZZ$	CMS	13	35.9	search for anomalous EW production
2020	$ZZjj$	ATLAS	13	137	preliminary results
2020	$W_L^{\pm}W_L^{\pm}jj$	CMS	13	137	first measurement of polarized scattering
2020	$W\gamma jj$	CMS	13	138	improved sensitivity to the interference between the SM and the O_{3W} contribution
2021	$Z\gamma jj$	CMS	13	137	most stringent (or competitive) constraints on the EFT dimension-8 operators

Table 1.3: Summary of the experimental searches for VBS in CMS and ATLAS collaborations.

In 2016 the ATLAS collaboration published their measurement of $\boldsymbol{W^{\pm}Z}$ production cross sections in pp collisions at $\sqrt{s} = 8\ TeV$ corresponding to the integrated luminosity of $20.3\ fb^{-1}$. In the VBS phase-space, the cross section was reported to be $\sigma_{VBS}(W^{\pm}Zjj) = 0.29^{+14}_{-12}(stat)^{+0.09}_{-0.1}(syst)\ fb$, where the SM prediction gives $0.13 \pm 0.01\ fb$. In addition, limits on anomalous gauge boson self-couplings were reported as well [19].
Another study was done by ATLAS in the same year aiming for the measurement of the $\boldsymbol{W^{\pm}W^{\pm}}$ production in events with two leptons (e or μ) with the same electric charge and at least two jets using the pp collision data at $\sqrt{s} = 8\ TeV$ and an integrated luminosity of $20.3\ fb^{-1}$. In addition, the goal was to put more stringent limits on the aQGCs. The measured fiducial cross-section of $\sigma_{fid}^{incl.}(W^{\pm}W^{\pm}jj) = 2.3 \pm 0.6(stat) \pm 0.3(syst)\ fb^{-1}$ in the inclusive region was reported. The same was also measured in the VBS region giving $\sigma_{fid}^{VBS}(W^{\pm}W^{\pm}jj) = 1.5 \pm 0.5(stat) \pm 0.2(syst)\ fb^{-1}$. The expected sensitivity to α_4 and α_5 was improved significantly, compared to the previous ATLAS result, by selecting a phase-space region that is more sensitive to anomalous contributions to the $WWWW$ vertex. The paper reports the following expected (observed) limits: $-0.06 < \alpha_4 < 0.07$ ($-0.14 < \alpha_4 < 0.15$) and $-0.10 < \alpha_5 < 0.11$

($-0.22 < \alpha_5 < 0.22$). The result constitutes a 35 % improvement in the expected aQGC sensitivity with respect to the previous results [20].
In the same year, the CMS collaboration reported a study on the electroweak-induced production of $\boldsymbol{W\gamma}$ with two jets in pp collisions at $\sqrt{s} = 8\ TeV$ and an integrated luminosity of $19.7\ fb^{-1}$. Limits on the anomalous quartic gauge couplings were set as well. For the EW signal, the observed (expected) significance was found to be 2.7 (1.5) standard deviations, while for the EW+QCD signal significance of 7.7 (7.5) standard deviations was observed. The measured cross section in the fiducial region was found to be $10.8 \pm 4.1(stat) \pm 3.4(syst) \pm 0.3(lumi)\ fb$ for the EW-induced $W\gamma$ + 2jets production and $23.2 \pm 4.3(stat) \pm 1.7(syst) \pm 0.6(lumi)\ fb$ for the total $W\gamma$ + 2jets production. Exclusion limits for aQGC parameters $f_{M,0-7}/\Lambda^4$, $f_{T,0-2}/\Lambda^4$ and $f_{T,5-7}/\Lambda^4$ were set at 95 % confidence level. This study reported the first limits on the $f_{M,4}/\Lambda^4$ and $f_{T,5-7}/\Lambda^4$ parameters [21].

In 2017 both the CMS and ATLAS collaborations reported the first measurements on the VBS in the $\boldsymbol{Z\gamma}$ channel at $\sqrt{s} = 8\ TeV$ with the data corresponding to roughly $20\ fb^{-1}$.
The CMS reported an evidence for **EW $\boldsymbol{Z\gamma jj}$** production with an observed (expected) significance of 3.0 (2.1) standard deviations. The fiducial cross section for EW $Z\gamma jj$ production was measured to be $\sigma_{fid}(Z\gamma) = 1.86^{+0.90}_{-0.75}(stat)^{+0.34}_{-0.26}(syst) \pm 0.05(lumi)\ fb^{-1}$. The fiducial cross section for combined EW and QCD $Z\gamma jj$ production of $\sigma_{fid}(Z\gamma) = 5.94^{+1.53}_{-1.35}(stat)^{+0.43}_{-0.37}(syst) \pm 0.13(lumi)\ fb^{-1}$ was reported as well. Both measurements are consistent with the theoretical predictions. In addition to previously imposed limits on the $f_{M0,1,2,3}$ and $f_{T0,1,2,9}$ parameters, the first observed (expected) limits on the neutral aQGC parameter f_{T8} were reported: $-1.8 < f_{T8}/\Lambda^4 < 1.8$ ($-2.7 < f_{T8}/\Lambda^4 < 2.7$). The limits on aQGC parameters are expressed in TeV^{-4} [22].
The ATLAS collaboration reported 2.0σ (1.8σ) observed (expected) significance for the production of the **EW $\boldsymbol{Z\gamma jj}$** with the fiducial cross section of $\sigma_{fid}(Z\gamma) = 1.1 \pm 0.5(stat) \pm 0.4(syst)\ fb^{-1}$. The EWK+QCD cross section was also reported and quoted to be $\sigma_{fid}(Z\gamma) = 3.4 \pm 0.3(stat) \pm 0.4(syst)\ fb^{-1}$. Limits on the aQGCs are also discussed in the paper [23].
The first measurement of the **same-sign $\boldsymbol{WW}$** production at $\sqrt{s} = 13\ TeV$ was made by the CMS collaboration in 2017 using the data that corresponds to the integrated luminosity of $35.9\ fb^{-1}$. The observed significance of 5.5 standard deviations was reported, whereas a significance of 5.7 standard deviations was expected based on the standard model predictions. The ratio of the measured event yields to that expected from the SM at the leading order was measured to be 0.90 ± 0.22. In addition, bounds were given on the structure of quartic vector boson interactions in the framework of dimension-eight effective field theory operators and on the production of doubly charged Higgs bosons [24].
In the same year, the CMS collaboration did, for the first time, a search for the VBS in the **fully leptonic ZZjj** channel at $\sqrt{13}\ TeV$. The process is measured with an observed (expected) significance of 2.7 (1.6) standard deviations. A fiducial cross section for the EW production is measured to be $\sigma_{EW}(ZZjj) = 0.40^{+0.21}_{-0.16}(stat)^{+0.13}_{-0.09}(syst)\ fb$ which is consistent with the SM prediction. Limits on the anomalous quartic gauge couplings were determined in terms of the EFT operators $f_{T0,1,2,8,9}$. These are shown in Table 1.4. More details on this study can be found in [25]

In 2018 the ATLAS collaboration reported their efforts in measuring the **EW $\boldsymbol{WZ}$** boson pair production in association with two jets in pp collisions at $\sqrt{s} = 13\ TeV$, corresponding to an integrated luminosity of $36.1 fb^{-1}$. The observed (expected) significance of 5.3 (3.2) standard deviations was reported. The measured fiducial cross section for EW production, including interference effects, was measured to be $\sigma_{fid}(W^{\pm}Z) = 0.57^{+0.14}_{-0.13}(stat)^{+0.05}_{-0.04}(syst)^{+0.01}_{-0.01}(lumi)\ fb$ [26].

Coupling	Exp. lower	Exp. upper	Obs. lower	Obs. upper	Unitarity bound
f_{T_0}/Λ^4	-0.53	0.51	-0.46	0.44	2.9
f_{T_1}/Λ^4	-0.72	0.71	-0.61	0.61	2.7
f_{T_2}/Λ^4	-1.4	1.4	-1.2	1.2	2.8
f_{T_8}/Λ^4	-0.99	0.99	-0.84	0.84	1.8
f_{T_9}/Λ^4	-2.1	2.1	-1.8	1.8	1.8

Table 1.4: Expected and observed lower and upper 95% CL limits on the couplings of the quartic operators T0, T1, and T2, as well as the neutral current operators T8 and T9. The unitarity bounds are also listed. All coupling parameter limits are in TeV^{-4}, while the unitarity bounds are in TeV. The table is taken from [25]

In the following year the CMS collaboration published their results on the measurement of the **EW WZ** boson production and search for new physics in $WZ+\,2jet$ events in pp collisions at $\sqrt{s} = 13\,TeV$. The measured (expected) significance of 2.2 (2.5) standard deviations was reported. The best-fit value for the $WZjj$ signal strength was used to obtain a cross section in the tight fiducial region and was measured to be $\sigma_{fid}^{tight}(WZjj) = 3.18^{+0.57}_{-0.52}(stat)^{+0.43}_{-0.36}(syst)\ fb$. This is compatible with the SM prediction of $\sigma_{pred} = 3.27^{+0.39}_{-0.32}(scale) \pm 0.15(PDF)\ fb$. In addition, results were also obtained in a looser fiducial region to simplify comparisons with theoretical calculations. The resulting WZjj loose fiducial cross section was measured to be $\sigma_{fid}^{loose}(WZjj) = 4.39^{+0.78}_{-0.72}(stat)^{+0.60}_{-0.50}(syst)\ fb$. This can be compared to the predicted value of $\sigma_{pred} = 4.51^{+0.78}_{-0.72}(scale) \pm 0.18(PDF)\ fb$. Finally, constraints on charged Higgs boson production and on aQGCs in terms of dimension-eight EFT operators were presented as well [27].
Measurement of the **same-sign W** boson pair production at $\sqrt{s} = 13\,TeV$ by ATLAS collaboration was published in the same year. The background-only hypothesis was rejected with the significance of 6.5σ. The measurement of the fiducial cross section was reported as well giving $\sigma_{fid}(W^{\pm}W^{\pm}) = 2.89^{+0.51}_{-0.48}(stat)^{+29}_{-28}(syst)\ fb$ [28].
Another important paper in 2019 was published by the CMS collaboration where a report on a search for anomalous EW production of **WW**, **WZ**, **and ZZ** boson pairs in association with two jets in proton-proton collisions at $\sqrt{s} = 13\,TeV$ was presented. No excess of events, with respect to the SM background predictions, was observed. The events in the signal region were used to constrain aQGCs in the EFT framework. The study reported new constraints on operators $T_{S0,1}$, $T_{M0,1,6,7}$ and $T_{T0,1,2}$. This study was the first one to search for anomalous EW production of WW, WZ, and ZZ boson pairs in WV and ZV semi-leptonic channels at $13\,TeV$ and it improved the sensitivity of the previous CMS results at 13 TeV in fully leptonic decay channel by factors of up to seven, depending on the operator [29].

While the analysis presented in chapter 4 was being prepared for publishing, the ATLAS collaboration presented their preliminary results on VBS in the channel with **two leptonically decaying Z bosons** accompanied by the two hadronic jets using the full Run 2 data at $\sqrt{s} = 13\,TeV$ corresponding to the integrated luminosity of $137\ fb^{-1}$. The background-only hypothesis was rejected with an observed (expected) significance of 5.5σ (4.3σ). The fiducial cross section was reported to be $\sigma_{fid}(ZZjj) = 1.27 \pm 0.14\ fb$ where $1.14 \pm 0.04(stat) \pm 0.20(syst)$ was expected [30].
Another important study in 2020 was the first measurements of the **polarized same-sign W** boson pairs in association with two jets in pp collisions at $\sqrt{s} = 13\,TeV$ presented by the CMS collaboration. The signal was measured with an observed (expected) significance of 2.3 (3.1) standard deviations. An observed 95% confidence level upper limit on the production cross section for longitudinally polarized same-sign W boson production was reported to be $0.32^{+0.42}_{-40}\ fb$ in the $W^{\pm}W^{\pm}$ centre-of-mass frame and $0.24^{+0.40}_{-0.37}\ fb$ in the parton-parton centre-of-mass frame. Both measurements agree with theoretical predictions [31].
In the same year, the CMS collaboration reported measurement of $W\gamma$ differential cross sections at $\sqrt{s} = 13\,TeV$

c.o.m. energy and integrated luminosity of $138\ fb^{-1}$. Constraints on the C_{3W} coefficient, affecting the $WW\gamma$ vertex, in the EFT framework via a parametrization of the fiducial cross section in photon transverse momentum (p_T^{γ}) and azimuthal angle of the charged lepton ($|\phi_f|$) were also set. This 2D approach resulted in a tenfold improvement in sensitivity to the interference between the SM and the O_{3W} contribution compared to using the transverse momentum alone [32].

Finally, in 2021 the CMS collaboration measured the EW production of $Z\gamma$ associated with two jets at $\sqrt{s} = 13\ TeV$ with both expected and observed signal significance greater than five standard deviations. The fiducial cross section was reported to be $\sigma_{fid}(EWZ\gamma) = 5.21 \pm 0.52(stat) \pm 0.56(syst)\ fb$. Exclusion limits on the dimension-eight operators M_{0-7} and $T_{0-2,5-9}$ in the EFT framework at 95% confidence level was reported as well [33].

The work presented in this thesis work shows the first measurement, in the CMS collaboration, of the EW production of two leptonically decaying Z bosons accompanied by two jets. In addition, it shows the first prospective studies for the scattering of the longitudinal Z bosons at 14 as well as $27\ TeV$ conditions expected in future LHC runs.

Chapter 2

The Large Hadron Collider and the CMS experiment

2.1 Preface to the chapter

In the first section of this chapter, I will make a historical overview of events that led to the construction of the World's largest particle detector in history; the Large Hadron Collider (LHC) at CERN. Next, I will briefly introduce the LHC machine and the largest experiments designed to collect and analyse data produced in the proton-proton collisions at the LHC.
The analysis presented in this thesis uses data collected by the Compact Muon Solenoid (CMS) detector described in section 2.3. I will first describe the coordinate system used throughout this document. Next, I will briefly describe each of the sub-detectors. I will finish the section with a discussion of the trigger system used at CMS. In section 2.4 I will describe how muons and jets are reconstructed at CMS followed by a description of the particle-flow algorithm. Electron reconstruction is described, in more detail, in the next chapter. Finally, in section 2.5, I will introduce a reader with future plans for the LHC and the CMS detector. This section is a basis for the discussion presented in Chapter 5.

2.2 The Large Hadron Collider (LHC)

2.2.1 The LHC machine and physics experiments

Only a brief overview of the LHC operation procedure is discussed in this section. A much more detailed description can be found elsewhere [34]. A full LHC accelerator complex is illustrated on Fig. 2.1.
The process of particle acceleration starts by stripping electrons from the compressed hydrogen. This leaves only protons which are accelerated in the electric field. A DC voltage cannot be used as particles would be accelerated through the gap, but decelerated elsewhere. Thus, oscillating voltage is needed so that particles see accelerating voltage across the gap and, at the same time, the voltage has to cancel out as the particle goes around the rest of the accelerator. For this, radio frequency (RF) systems are used [36]. Initially, the protons are accelerated in the linear accelerator. In the following phase, protons enter the Proton Synchrotron Booster where they are accelerated to the speed of 91.6 % of the speed of light. The next phase of acceleration is taking place in the Proton Synchrotron (PS) where protons gain a speed of 99.9 % of the speed of light. The final phase of acceleration before the LHC is Super Proton Synchrotron (SPS) which increases the energy of protons to 450 GeV.
Finally, protons are injected into the LHC which is placed roughly 170 m below the surface and has a circumference

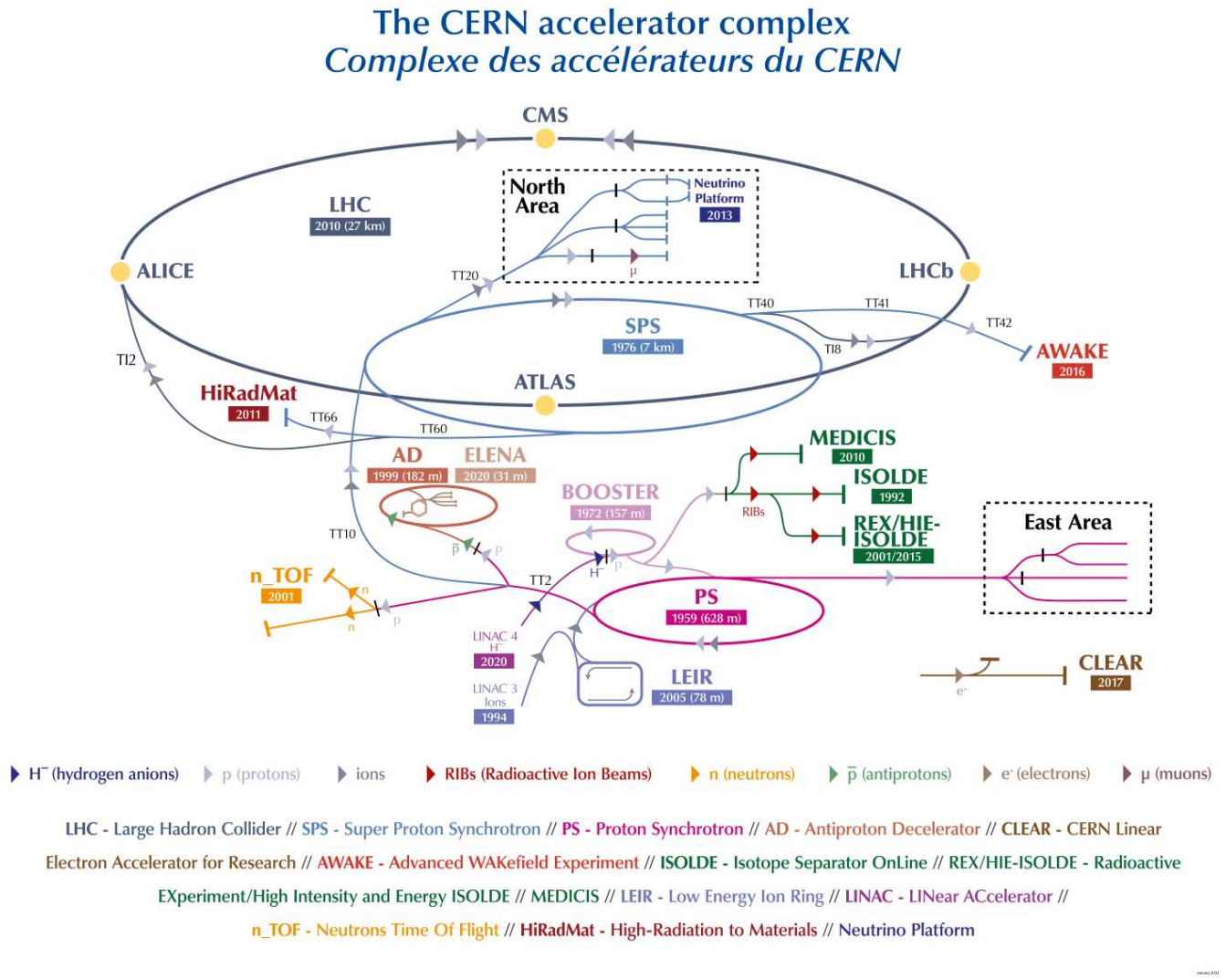


Figure 2.1: The LHC accelerator complex. Protons are first accelerated through the linear accelerator LINAC2. Before entering the largest ring, LHC, protons go through Proton Synchrotron Booster (PSB), the Proton Synchrotron (PS), and the Super Proton Synchrotron (SPS). The illustration is taken from [35]

of 27 km. The LHC machine consists of two tubes in which protons circulate in opposite directions. In four locations the tubes cross and protons are collided.
To fill the LHC with protons, 12 cycles of SPS are needed. Each cycle of SPS required 3 to 4 cycles of PS. Since SPS and PS cycle times are 21.6 and 3.6 seconds respectively, LHC filling time is then around 4 minutes per beam. Since LHC requires additional 4 SPS cycles for the injection setup, and LHC operators need at least 2 minutes to adjust the machine settings, the injection time per beam for LHC then becomes approximately 16 minutes [37].

Protons are not spread uniformly along the beam, but are, instead, grouped together in, so-called, *bunches*. Each bunch contains around $1.15 \cdot 10^{11}$ protons and is roughly 7.5 cm long and focused using quadrupole magnets into the area of $16 \times 16\ \mu m^2$. At any given time, there are approximately 2000 proton bunches in a single beam.
Because of the small cross section of processes studied at the LHC, one would like to maximize the rate of events which depends on the cross section and the instantaneous luminosity, $\mathcal{L}$:

$$\mathcal{L} = \gamma \frac{f n_b N^2}{4\pi \epsilon_n \beta^*} R,$$

where $\gamma = \frac{E}{m}$ is the relativistic factor for protons, f is the revolution frequency, n_b is the number of bunches, N is the number of protons per bunch, ϵ_n is the normalized transverse beam emittance [38], β^* is the beam beta function at the collision point and R is a reduction factor due to the beam crossing angle at the interaction point [39]. Assuming nominal beam parameters, this yields instantaneous luminosity of order $10^{34}\ cm^2 s^{-1}$, two orders of magnitude larger than that of the Tevatron collider. The spacing between the two bunch crossings at the LHC is around 25 ns, which corresponds to the bunch crossing rate of 40 MHz.

Physics experiments at the LHC

One of the first meetings dedicated to physics experiments at the LHC was held in Barcelona in 1989 where the first predecessor of the Experiment for Accurate Gamma, Lepton and Energy measurements (EAGLE) experiment started forming. The next important workshop, Towards the LHC Experimental Programme, was held in Evian in 1992 where proto-collaborations described respective detector plans. In total, 12 proposals were made.
Four of those were made for general-purpose experiments: EAGLE, Apparatus with Superconducting Toroids (ASCOT), L3+1 and Compact Muon Solenoid (CMS). Next, three b-physics experiments were competing for approval: a Collider Beauty Experiment (COBEX), the LHB collaboration envisaged as a fixed-target experiment dedicated to the study of beauty hadrons and a CP-violation gas jet experiment (GAJET).
Three experiments were proposed for heavy-ion experiments: the one that will later be known as A Large Ion Collider (ALICE), the one that wanted to use the DELPHI detector from LEP, and the one that suggested a heavy-ion program for the CMS detector.
Amongst four multi-purpose detectors, only two would be accepted at the LHC. One of them was CMS. The other was formed by merging ASCOT and EAGLE into A Toroidal LHC Apparatus (ATLAS) experiment. In January 1996, CMS and ATLAS were approved and the approval for construction was given on January 31^{st} 1997. The last large pieces of CMS and ATLAS were lowered into the experimental caverns on July 23^{rd} and February 29^{th} 2008, respectively.
In January 1994, the CERN LHC Experiments Committee (LHCC) recommended that COBEX, GAJET and LHB form a single collaboration. In September 1998, a technical proposal for the newly formed collaboration, LHCb, was accepted. Finally, ALICE was approved in February 1997.
After the four big experiments, **CMS**, **ATLAS**, **LHCb** and **ALICE**, were approved, three smaller experiments submitted a Letter of Intent: the Total Cross Section, Elastic Scattering and Diffraction Dissociation Measurement at the LHC (**TOTEM**) experiment in 1997, Monopole and Exotics Detector at the LHC (**MoEDAL**) in 1998 and **LHCf** in 2003.
In the following sections, the design of the CMS detector is discussed.

2.3 The CMS experiment

The CMS detector is one of the two largest detectors at the CERN LHC. It is a general-purpose detector located roughly 100 meters below the surface near the French village of Cessy, between Lake Geneva and the Jura mountains. Many goals of the LHC include understanding the mechanism of EWSB and the Higgs mechanism and the search for new physics that could manifest itself in terms of extra dimensions, forces, and symmetries. These, and many other phenomena, present strong arguments to investigate a TeV energy scale at the LHC.
Apart from the high energy conditions, a very high luminosity is expected at the LHC as well, with estimated 10^9 proton-proton collisions every second. This will result in around 1000 charged particles emerging from the interaction point every 25 ns. Because of this, very high levels of radiation are expected requiring radiation-hard detectors and

front-end electronics. Finally, the greatest challenge for the LHC now and in the future is the *pileup*, i.e. the average number of events per bunch crossing. The challenging environment expected at the LHC requires careful detector design.
In order to meet the goals of the LHC physics programme, the CMS detector has to satisfy several requirements which can be summarized as the following [40]:

- good muon identification and momentum resolution over a wide range of momenta in region $|\eta| < 2.5$, dimuon mass resolution of order 1 % at $100\ GeV/c^2$ and reliable charge measurement of muons with $p < 1\ TeV$
- good charged particle momentum resolution and reconstruction efficiency in the inner tracker
- good electromagnetic energy resolution, diphoton and dielectron mass resolution of order 1 % at $100\ GeV/c^2$, geometric coverage up to $|\eta| = 2.5$, correct localization of the primary interaction vertex, π^0 rejection and efficient photon and lepton isolation at high luminosities
- good resolution for the dijet mass and missing transverse energy measurements

An illustration of the detector is given in Fig. 2.2. The 21 meters long, 15 meters wide and 15 meters high detector is designed around the superconducting magnet responsible for bending trajectories of charged particles within the detector. The closest subdetector system to the interaction point is the silicon tracker used to measure the momentum of particles via trajectory curvature in the magnetic field followed by the Electromagnetic Calorimeter (ECAL) used to collect energy depositions of electrons and photons, and to capture one part of energy depositions of hadrons. Another layer enclosed by the solenoid magnet is the Hadron Calorimeter (HCAL) with the main task of completing the energy collection of hadrons. Ideally, all particles (except for neutrinos and muons) will leave their energy in the CMS calorimeters. In order to obtain a clean muon signal, the muon detector is, therefore, placed furthest from the interaction point. A more detailed description of each subdetector system is given in the following sections.

CMS coordinate system

In order to follow the discussions presented in this thesis, it is essential to define the coordinate system. The coordinate system, illustrated in Fig. 2.3, has the origin at the interaction point, the z- axis along the proton beam direction, the y-axis pointing vertically upward and the x-axis pointing radially inward towards the centre of the LHC. Many observables are defined in the cylindrical coordinate system where the transverse plane is given by the $x - y$ plane. By this convention, the *transverse momentum* of a particle, p_T, is defined as the projection of the particle momentum onto the transverse plane and its magnitude is defined as

$$p_T = \sqrt{p_x^2 + p_y^2}$$

Another property of a particle is its *rapidity*, y, defined as

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

When the mass of the particle is much smaller than its energy, which will always be the case for electrons and muons in this analysis, it is useful to define its *pseudorapidity* as

$$\eta = -ln \left(tg \frac{\theta}{2} \right)$$

For particles with coordinates $(\theta_1,\ \phi_1)$ and $(\theta_2,\ \phi_2)$ we define *angular distance*, ΔR, as

$$\Delta R = \sqrt{(\theta_1 - \theta_2)^2 + (\phi_1 - \phi_2)^2}$$

Finally, transverse missing energy is defined as the negative transverse momentum sum of all reconstructed particles projected onto the transverse plane.

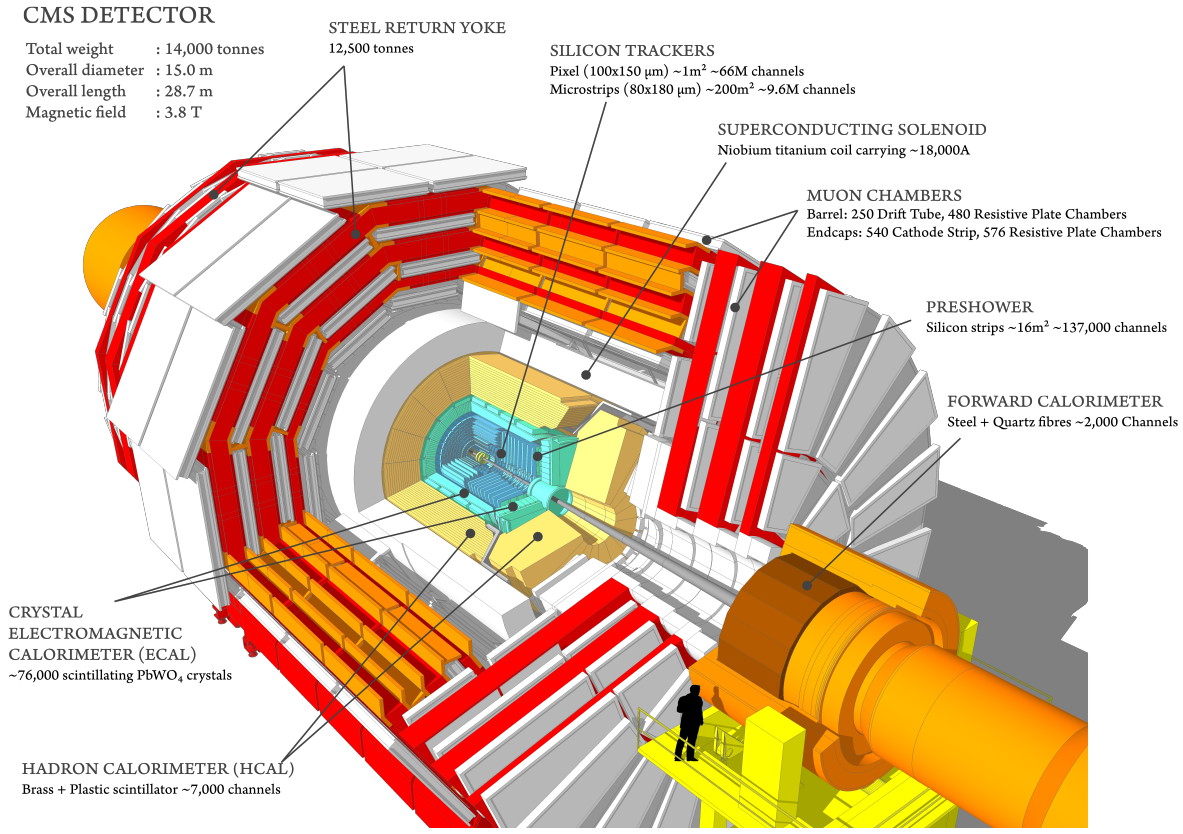


Figure 2.2: A sectional view of the CMS detector. The LHC beams travel in opposite directions along the central axis of the CMS cylinder colliding in the middle of the CMS detector.

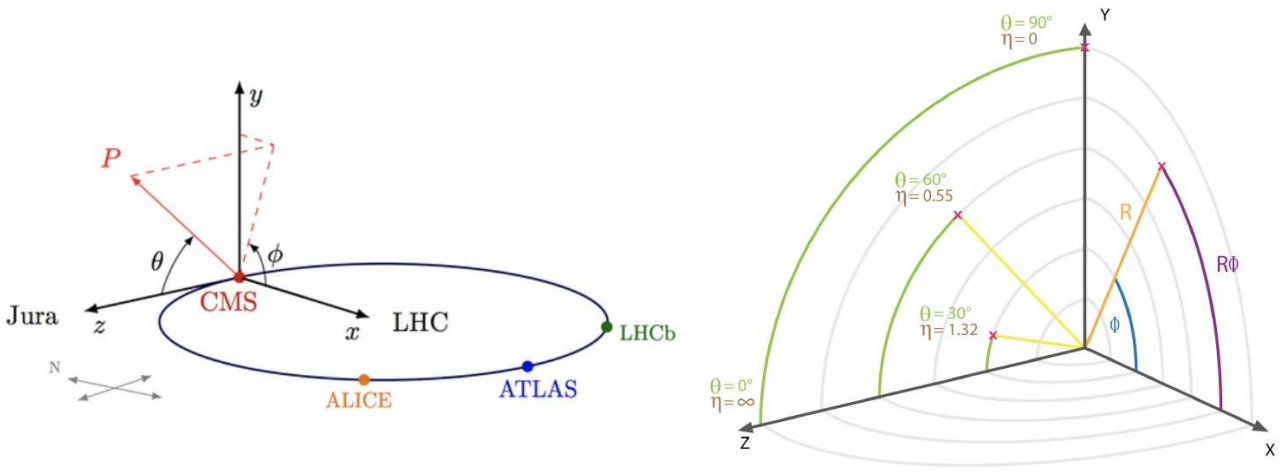


Figure 2.3: Illustration of the coordinate system used throughout the thesis. The figure is taken from [41]

2.3.1 The Silicon Tracker

The role of the *tracker*, is to measure the trajectory of charged particles and to reconstruct the primary vertices and secondary decay vertices. Because of the magnetic field, the trajectory of charged particles in the tracker is curved which is used to measure their momentum. Because of the large number of proton-proton interactions per bunch crossing, the tracker is required to resolve a large number of pileup interactions from the hard processes. Failing to do so would result in quality degradation of the physics measurements.
Fig. 2.4 shows the illustration of the cross section of the CMS tracker. It consists of the *barrel* (up to $|\eta| < 0.9$) and two *endcaps*, together covering the region up to $|\eta| < 2.5$. The tracking is achieved by successive layers of active material. When a charged particle traverses the active material of the tracker, it causes ionization in the material that is registered by the readout electronics. In total, the sensitive area of the tracker is over 200 m^2 and it allows for fast and precise measurements with temporal and spatial resolutions that fulfil the challenges posed by the high luminosity LHC collisions, which occur at a frequency of 40 MHz.

The tracker consists of two sub-detectors: the *pixel detector* and the *strip detector*. The pixel detector has a surface area of 1.1 m^2 and is built from 66 million pixels of size $100\ \mu m \times 150\ \mu m$ and $285\ \mu m$ thick. The main reason for the small pixel size is the need to separate the hits from particles that are nearby and to resolve the z coordinate of particles coming from different vertices to suppress pileup. The pixel detector has three layers in the barrel region at distances of 4.3 cm, 7.2 cm and 11 cm from the interaction point, and two disks (endcap regions) on each side of the barrel at 34.5 cm and 46.5 cm from the interaction point. The pixel detector contains 15840 read-out chips (ROCs) arranged into modules which transmit data via 1312 read-out links.
The pixel detector is surrounded by the strip detector segmented into 9.6 million strips with a thickness of $320\ \mu m$ $(500\ \mu m)$ in the inner (outer) layers. The strip detector has 10 layers in the barrel region at the distance from 25 cm up to 110 cm from the interaction point and up to 120 cm along the z-axis. Four layers are located in the inner barrel (TIB) and six layers are in the outer barrel (TOB). In addition, there are 12 disks in the endcap region at a distance of up to 110 cm from the interaction point and up to 280 cm along the z-axis: three inner discs (TID) inside, and nine endcap discs (TEC) outside the TOB. [42].

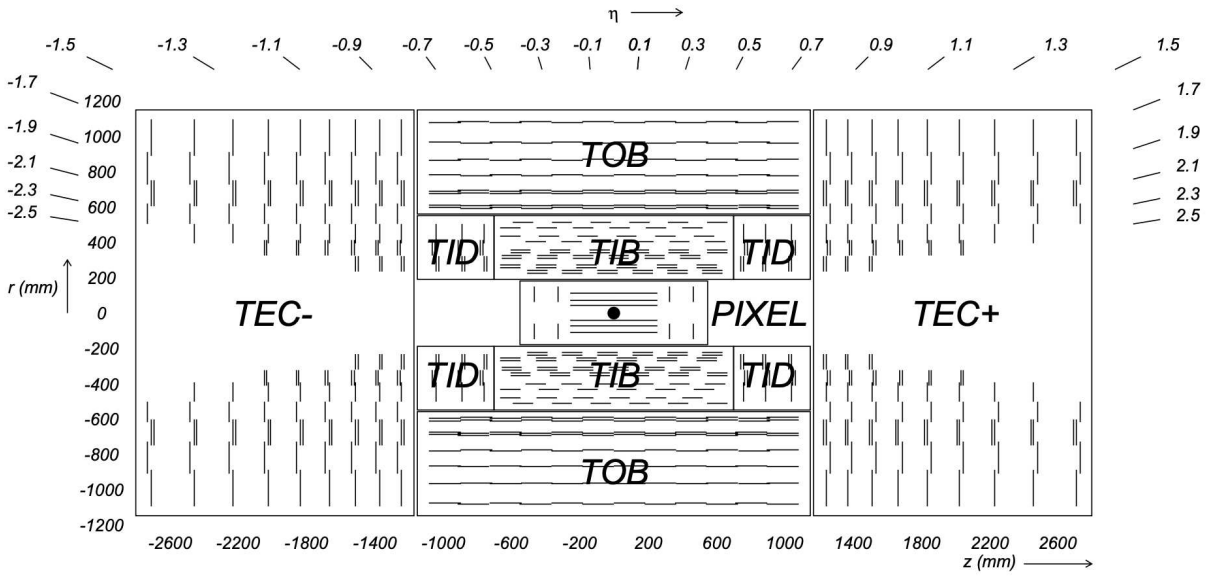


Figure 2.4: Cross-section view through the CMS tracker. Each line represents a detector module. Double lines indicate back-to-back modules which deliver stereo hits. The figure is taken from [43]

2.3.2 The Electromagnetic Calorimeter

The CMS ECAL is a hermetic, homogeneous, fine grained calorimeter made of 61200 lead tungstate ($PbWO_4$) crystals arranged in the central *barrel* region (EB) and surrounded by 7324 crystals in each of the two *endcaps* (EE). A cross section of the ECAL is shown in Fig. 2.5.

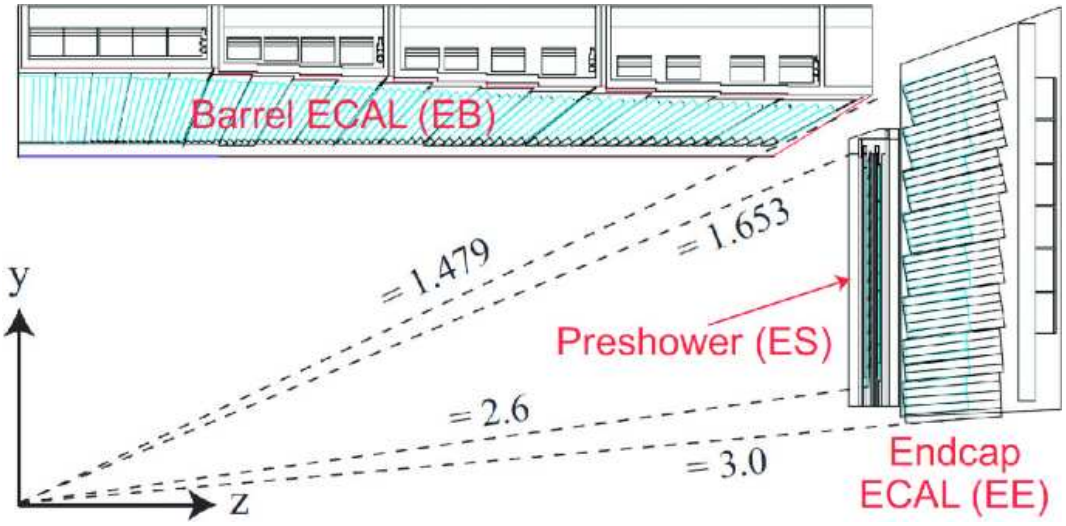


Figure 2.5: Illustration of the CMS ECAL. The figure is taken from [44]

ECAL is one of the most important components of the CMS detector optimised to detect the decay of Higgs boson into photon pairs. For this, it had to be designed for efficient photon and electron identification as well as high energy and position resolution. The design requirements are [45]:

- excellent energy and position/angle resolution up to $|\eta| < 2.5$ in order to match the tracker coverage
- hermeticity, compactness and high granularity
- fast response ($\sim 25\ ns$) and precise particle identification and isolation at the trigger level
- capability of measuring electron and photon energies in range from 5 GeV to 5 TeV
- high radiation tolerance

ECAL design

With high density ($2.82\ g/cm^3$), short radiation length ($0.89\ cm$) and small Molière radius ($2.2\ cm$), $PBWO_4$ crystals are the perfect candidates for designing a fine granularity and a compact calorimeter. The crystals emit blue-green scintillation light when electrons and photons pass through them. The scintillation decay time is close to the LHC bunch crossing time with roughly 80 % of the light emitted in 25 ns. A picture of the crystal is shown in Fig. 2.6. The picture on the right-hand side of the figure shows the photodetector attached. Due to high radiation levels at the LHC, ECAL crystals suffer from wavelength-dependent loss of light transmission which is tracked and corrected by monitoring the optical transparency with injected laser light.

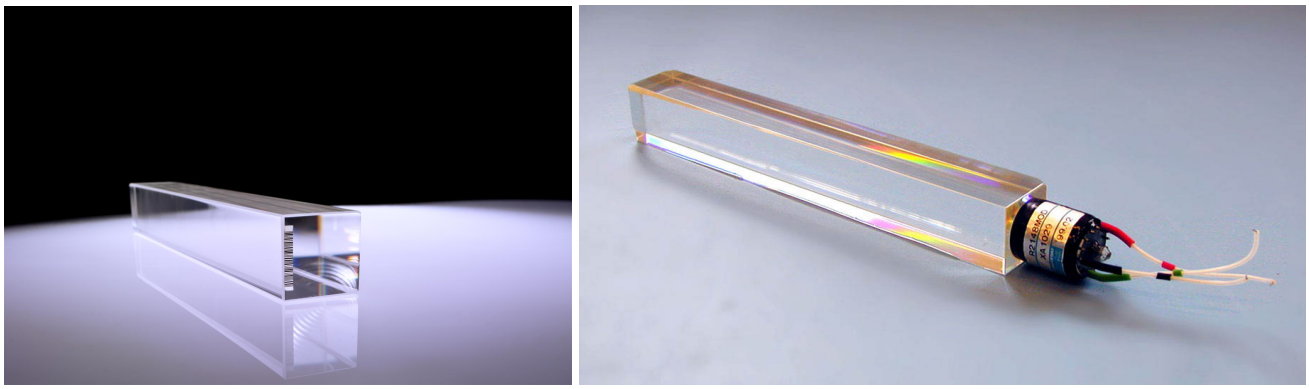


Figure 2.6: Picture of the ECAL $PBWO_4$ crystal. The right-hand side shows the photodetector on one of the crystals used in the ECAL endcap.

In the ECAL barrel, covering the pseudorapidity range $|\eta| < 1.479$, crystals with a tapered shape, varying slightly in η, are mounted in a quasi-projective geometry to avoid cracks aligned with particle trajectories. Thus, their axis is slanted by $3°$ with respect to the vector from the nominal interaction vertex. At the front face, the surface of the crystal is $22 \times 22\ mm$, while at the back face it is $26 \times 26\ mm$, corresponding to an area of approximately 0.0174×0.0174 in η. The crystal length is $23\ cm$ with front faces at the distance of $1.29\ m$ from the beam. The total crystal volume in the EB is $8.14\ m^3$ weighing around 67.4 tonnes. Crystals are grouped in, so-called, submodules. To reduce the number of different crystals, each submodule contains only a pair of shapes, one being a mirror image of the other. The submodules are further assembled into modules of different types, according to the position in η, with each module containing 400 or 500 crystals. Finally, four modules are grouped into a supermodule which contains 1700 crystals. Left-hand isde in Fig. 2.7 shows one crystal module.

In the ECAL endcaps, covering the pseudorapidity range $1.479 < |\eta| < 3.0$, identically shaped crystals are grouped into 5×5 arrays called *supercrystals* (SCs). Each endcap is divided into 2 halves referred to as the *Dee*, with each Dee holding 3662 crystals grouped into 138 standard SCs and 18 partial SCs on the inner and outer circumference. The SCs are arranged into a rectangular grid as shown on the right in Fig. 2.7. At the front face, the surface of the crystal is $28.62 \times 28.62\ mm$, while at the back it is $30 \times 30\ mm$. The length of crystals is $22\ cm$. The total crystal volume in each EE is $2.9\ m^2$ weighing around 24 tonnes.
Because the number of scintillation photons emitted by the crystals and the amplification of the photodetectors depends on temperature, ECAL temperature has to be maintained constant. This requires a cooling system capable of extracting the heat dissipated by the read-out electronics and of keeping the temperature of crystals and photodetectors stable within $\pm 0.05\ °C$. For optimal performance, the nominal operating temperature of the CMS ECAL is $18\ °C$.

The Preshower detector

In order to identify neutral pions in the endcaps, help identify electrons against minimum ionizing particles, and improve the position determination of electrons and photons, the Preshower detector is placed in front of EEs within $1.653 < |\eta| < 2.6$. It is a sampling calorimeter made of two planes of lead each followed by a plane of strip silicon sensors. Photons and electrons passing through lead create electromagnetic showers that are measured by the silicon strip sensors. The total thickness of the detector is $20\ cm$. The material thickness of the Preshower traversed at $\eta = 1.653$, before reaching the first sensor plane is two radiation lengths, X_0, followed by an additional 1

Figure 2.7: Left: Front view of an ECAL module equipped with crystals. Right: An endcap Dee of the CMS ECAL, fully equipped with supercrystals. The figures are taken from [43].

X_0 before reaching the second plane. Therefore, roughly 95 % of incident photons will produce electromagnetic showers before reaching the second plane. The lead planes are arranged into two Dees, one on each side of the beam pipe, with the same orientation as the crystal Dees. The surface area of each silicon sensor is $63 \times 63\ mm^2$ with an active area divided into 32 strips. The micromodules are placed on baseplates in groups of 7, 8 or 10 and, together with the electronics system motherboard placed above the micromodules, form a ladder. In total, there are 500 ladders corresponding to a total of around 4300 micromodules and 137000 individual read-out channels.

ECAL energy resolution

For energies below 500 GeV, the energy resolution can be parametrized as

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

where S is the stochastic term, N is the noise term, and C is the constant term. The stochastic term comes from event-to-event fluctuations in the lateral shower containment, a photostatistics contribution of $2.1\ \%$ and fluctuations in the energy deposited in the preshower lead absorber with respect to what is measured in the preshower silicon detector. The photostatistics contribution is given by $a_{pe} = \sqrt{F/N_{pe}}$, where N_{pe} is the number of primary photoelectrons released in the photodetector per GeV, and F is the excess noise factor which parametrizes fluctuations in the photodetector gain. The most important contributions to the constant term are the non-uniformity of the longitudinal light collection, inter-calibration errors, and leakage of energy from the back of the crystal. The noise term comes from the electronics noise, digitization noise, and pileup noise [43]. It was shown (see Ref. [46]), using the test beam results on two 3x3 crystal arrays, that, on average, $S = 2.8\ \%$, $N = 12\ \%$ and $C = 0.3\ \%$.

2.3.3 The Hadron Calorimeter

The main task of the CMS Hadron Calorimeter (HCAL) is to measure long-lived hadrons that form hadronic jets, and neutrinos or exotic particles resulting in missing transverse energy. It is responsible for collecting energy depositions of hadrons that traversed ECAL and it complements the momentum measurement of the tracker for charged hadrons. Fig. 2.8 shows the longitudinal view of the HCAL geometry.

The hadron barrel calorimeter (HB) is a sampling calorimeter covering the pseudorapidity range $|\eta| < 1.3$. It consists of 36 identical azimuthal wedges which form the two half-barrels (HB+ and HB-) that use the flat brass plates to absorb energy from particle showers. Plastic scintillator tiles are attached to a $0.94 - mm$-diameter green double-cladded wavelength-shifting fibres which transmit the signal to multi-channel hybrid photodiodes for readout.

The hadron calorimeter endcaps (HE) provides additional coverage up to $|\eta| < 3$, a region containing about 34 % of the final-state particles. The choice of the absorber material in the HE is dictated by a strong magnetic field of the solenoid. In addition, it is required to have a maximum number of interaction lengths to contain hadronic showers, good mechanical properties and reasonable cost. It was found that brass satisfies all the requirements. The scintillation light is collected by wavelength shifting (WLS) fibres that are spliced, at one end, to clear optic fibres which transfer signals to the photodetectors.

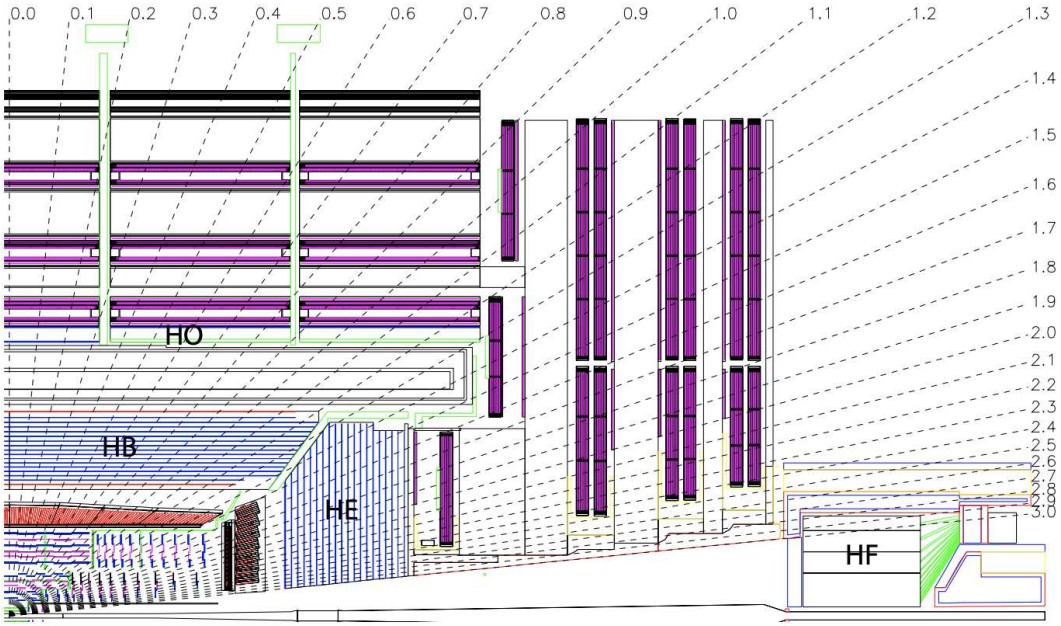


Figure 2.8: Longitudinal view of the CMS detector showing the locations of the hadron barrel (HB), endcap (HE), outer (HO) and forward (HF) calorimeters. The figure is taken from [43].

The hadron outer (HO) calorimeter is located outside the solenoid and it provides coverage up to $|\eta| < 1.26$. Its physical impact was studied (see Ref. [47]) using a simulation of the CMS detector. Single pions of fixed energies were shot at specific η values and the resulting energy deposits in the ECAL and HCAL were combined to measure particle energy. It was found that, without the HO, a certain amount of energy managed to leak outside ECAL and HCAL. However, with HO included, this energy was successfully recovered.

The design of the forward calorimeter (HF) was guided by the necessity to survive very-high radiation levels anticipated at high-η regions. Because of this, quartz fibres were chosen as the active medium which collects Cherenkov radiation generated by charged shower particles passing through the iron absorber. The HF is a cylindrical steel structure with an outer radius of $130\ cm$ with a front face located $11.2\ m$ from the interaction point. It is subdivided into 20° modular wedges. Thirty-six such wedges make up the HF calorimeters offering coverage up to $|\eta| < 5$.

2.3.4 The solenoid magnet

The solenoid magnet is the central feature of the CMS detector. Its task is to bend the trajectories of charged particles emerging from the interaction point. Depending on the charge of the particle, its trajectory will bend in a different way. Thus, the magnet enables the particle charge measurement. In addition, particles with higher momentum will bend less in the magnetic field than those with lesser momentum. Therefore, the magnetic field, together with information from other subdetector systems, enables a very accurate measurement of particle momentum.
The 12.5 tonnes magnet generates a magnetic field of up to $4\ T$, 100000 times stronger than the magnetic field of Earth. This is achieved by running $18500\ A$ of current through the superconductive coil cooled down to $4.6\ K$. The steel return yoke, composed of five barrel wheels and six disks grouped in two endcaps, increases the field homogenity in the Tracker and reduces the stray field by returning the magnetic flux of the solenoid.

2.3.5 The muon chambers

The muon system was designed to perform muon identification, momentum measurement, and triggering as well as to have the capability of reconstructing the momentum of muons over the entire kinematic range of the LHC. Since its instrumented area is roughly $25000\ m^2$, the gaseous detectors were found to be the best option because of their cost-efficiency, robustness and reliability. Three types of gaseous detectors are used: the *Drift Tube* (DT) system, the *Chatode Strip Chambers* (CSC) and the *Resistive Plate Chamber* (RPC) system. Following the design of the other subdetector systems, muon chambers are split into a barrel and two endcap regions. The longitudinal view of the muon system is shown in Fig. 2.9.

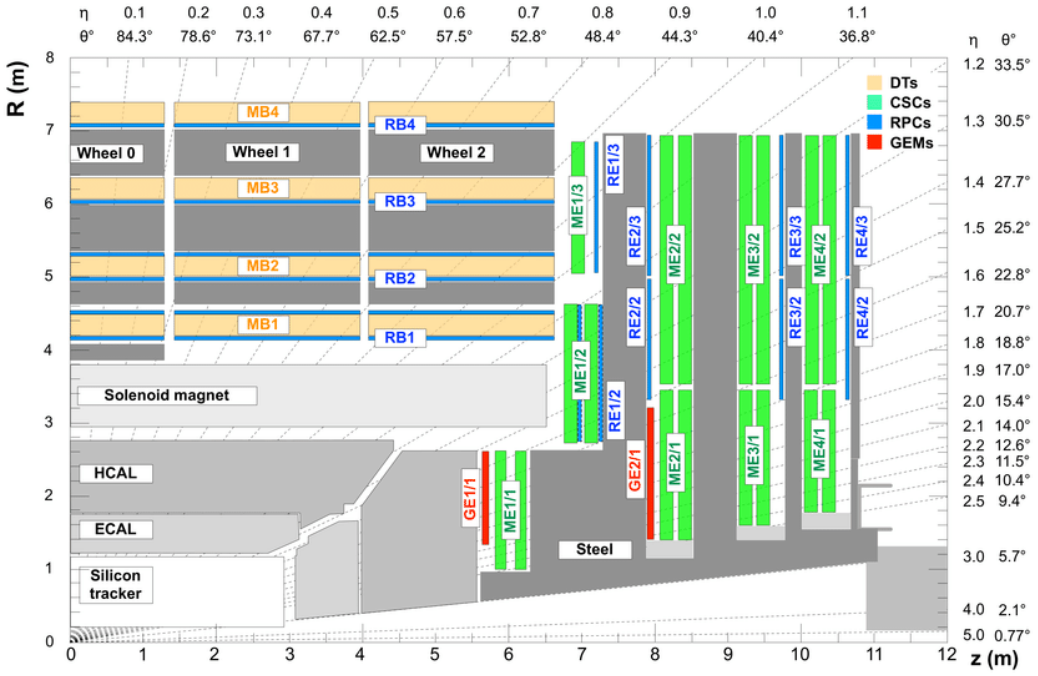


Figure 2.9: Longitudinal view of the muon system showing the position of the Drift Tube (DT) detectors in the barrel, the Cathode Strip Chambers (CSC) detectors in the endcap and the Resistive Plate Chambers (RPC) used for triggering. The figure is taken from [48].

In the muon barrel region, DT system is used with coverage up to $|\eta| < 1.2$. It consists of 4 chambers (MB1, MB2, MB3 and MB4) forming concentric cylinders around the beam line. Three inner cylinders incorporate 60 drift tubes

each, while the outer cylinder has 70. Each chamber, except MB4, is made of 3 independent units, a so-called superlayer (SL), and a thick honeycomb plate glued together. Each SL is composed of 4 layers of drift tubes, with all wires parallel. Each drift tube has a surface area of $43 \times 13\ mm^2$ and consists of a stretched wire immersed in the 85 - 15 % mixture of argon and carbon dioxide. When muon enters a drift tube, it ionizes the gas inside the chamber which causes electrons to drift towards the anode due to the applied electric field. By registering the time needed for electrons to reach the anode, and knowing where along the wire the hit is registered, it is possible to pinpoint where the muon passed [49].

The endcap region of the muon system is equipped with the CSCs that provide coverage between 0.8 and 2.4 in $|\eta|$ and enable precise time and position measurements. Each chamber is trapezoidal in shape and consists of six gas gaps. Each gap is equipped with a plane of radial cathode strips and a plane of anode wires. The passage of muon through the CSC causes the gas ionization and subsequent electron avalanche that produces a charge on the anode wire and an image charge on a group of cathode strips. By interpolating among the strips a very fine spatial resolution of $50\ \mu m$ is obtained.

The RPCs are located in both the barrel and in the endcaps and they offer very fine temporal resolution ($\sim 1\ ns$). Because of this, they guarantee a precise bunch crossing assignment of the muon tracks and are used for prompt trigger decisions. The barrel region is equipped with two RPC layers for the innermost stations (RB1 and RB2) and one layer for the outer stations (RB3 and RB4). The endcap region is equipped with one RPC layer per station for a total of 4 layers. The RPCs are made of two parallel bakelite plates placed at a distance of $2\ mm$ and filled with a $96 - 3.5 - 0.5\ \%$ mixture of tetraflouretan, isobutane and sulfur hexafluoride gas. The high voltage is applied to the outer graphite-coated surface of the bakelite plates which provides the electric field inside the gas mixture. When a passing muon ionizes the gas, the resulting electron avalanche induces a signal on the aluminium strips placed outside the gap. The signal is quickly picked up by external metallic strips which enable quick measurement of the muon momentum used by the trigger system to decide whether to keep the particular or not.

2.3.6 The trigger system

CMS is confronted with $40\ MHz$ of data which could contain signs of interesting, and possibly new, physics. To put it into perspective, this would be roughly 50 terabytes of data each second. At the same time, it can record only ~ 1000 events every second. Therefore, a lot of data has to be rejected. In order to prevent losing interesting collision data, a potent system had to be devised. This is the task of CMS *trigger*. The trigger is a two-level system. The *Level-1 Trigger*, or, simply, *L1 trigger*, has to reduce data from $40\ MHz$ to roughly $100\ KHz$ within $3.8\ \mu s$ from the pp collision. The High Level Trigger, often referred to as the *HLT*, further reduces data to roughly $1\ kHz$ in the next $300\ ms$.

The L1 trigger is implemented through very fast, specially designed, electronics known as Field-programmable gate arrays (FPGAs) where possible, but application-specific integrated circuits (ASICs) and programmable memory lookup tables (LUT) are also widely used. The schematic of the L1 trigger is shown in Fig. 2.10.

The trigger primitives, based on energy deposits in calorimeter trigger towers and track segments or hit patterns in muon chambers, are processed in several steps.
The L1 calorimeter trigger comprises of a *regional calorimeter trigger* (RCT) and the *global calorimeter trigger* (GCT). The RCT collects the information from ECAL and HCAL to determine ranked and sorted trigger objects such as electron or muon candidates. The rank reflects the level of confidence attributed to the L1 parameter measurement. The output ($e\gamma$ candidates and regional E_T sums based on 4×4 towers) is sent to the GCT which

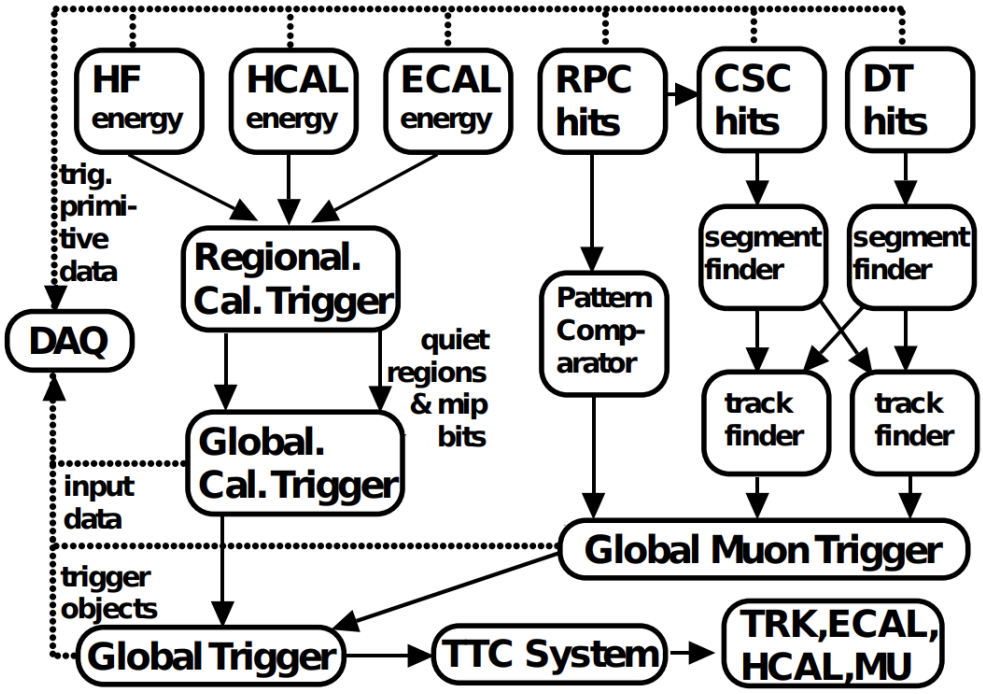


Figure 2.10: Overview of the CMS L1 trigger system. The figure is taken from [50].

sorts $e\gamma$ candidates, finds jets using E_T sums and calculates quantities such as E_T^{miss}.
Similarly, each muon detector system participates in the L1 muon trigger. For the DT and CSC systems, track segments are identified from the hit information and are transmitted via optical fibres to regional track finders to identify muon candidates and measure their momentum. The hits from the RPCs are directly sent to pattern comparator trigger (PACT) logic boards that identify muon candidates. Three muon regional track finder sorts the identified muon candidates and transmits them to the *global muon trigger* (GMT). Each muon candidate is then assigned a p_T and quality code as well as an $(\eta,\ \phi)$ position in the muon system. The GMT then merges overlapping muon candidates passing multiple-muon triggers.
The final step of the L1 trigger is the *global trigger* (GT) which receives outputs of the GCT and GMT such as $e\gamma$ objects, muons, central and forward hadronic jets, τ jets and global quantities [43]. Objects representing particles and jets are ranked and sorted followed by the decision of whether to reject an event or to accept it for further evaluation by the HLT.

The HLT is a software system implemented on a farm of about one thousand commercial processors. It is designed as a menu made of over 600 different paths targeting a broad range of physics signatures and purposes. Each HLT path is a sequence of increasingly complex algorithms similar to those used in offline analysis. In order to achieve the best performance, the faster algorithms are run first and the time-consuming algorithms, such as the Particle Flow algorithm (for more details on Particle Flow algorithm see section 2.4.5), are run at the end of the path. Unlike offline analysis where information from all subdetector systems is used for object identification and reconstruction, the HLT applies the algorithms only to a region of interest in the sub-detectors and reconstructs the physics object only for what is needed to select the event. If a filter fails at any step, the remaining part of the path is skipped in order to minimize CPU usage. If an event passes one of the HLT paths, it will be permanently stored and transferred to the CERN Tier-0 (T0) in one or more data streams. Data streams gather similar trigger paths that are commonly used by offline analysis. [51–54].

2.4 Physics objects reconstruction

Physics analyses don't use raw data stored by the HLT but, instead, high-level objects, such as electrons, muons and jets, amongst others. Thus, an intermediate step is needed before physics analysis can be performed. Object reconstruction comes into play here. In essence, this is a set of algorithms applied offline on raw data in order to fully exploit all CMS subdetectors to build *physic objects* from energy deposits as efficiently as possible and with the highest purity possible.
Object reconstruction based on the single subdetector is efficient when dealing, for example, with prompt electrons. Here, ECAL could be successfully used. However, inefficiencies may arise when dealing with objects such as hadronic jets for which the energy measured had traditionally been based on calorimeter clusters. This is due to the vastly different measurement efficiency for individual objects that constitute jets. The energy in a jet is split between charged hadrons, photons and long-lived neutral hadrons. Even though the energy resolution for photons is of the order of a few per cent, this is not the case for the charged hadrons for which the energy resolution can be of the order of several tens of per cent. Thus, the uncertainty in the measurement of hadron energy would dominate the measurement of the jet energy completely. In order to optimize object reconstruction, the *paricle-flow* (*PF*) algorithm was designed. This algorithm uses all available information from underlying subdetectors to optimize object reconstruction. This section describes CMS tracking and clustering algorithms followed by the discussion on the muon and jet reconstruction. The procedure for electron reconstruction is more complex and is described, in full detail, in the next chapter. Finally, the PF algorithm is described in section 2.4.5.

2.4.1 Tracking

Since the basic elements of the *PF* algorithm are charged particle tracks obtained from the silicon tracker and energy clusters from the ECAL and HCAL, it is instrumental that trajectories of the charged particles be measured as precisely as possible. The tracking software in CMS is commonly referred to as the Combinatorial Track Finder (CTF) which is an adaptation of the Kalman Filter (KF) [55–58]. The procedure is done iteratively in order to maximize the tracking efficiency, while, at the same time, reducing the rate of fake tracks as much as possible. Ten iteration steps are made, each more complex than the previous. Every iteration is done in four steps:

- Seed generation which provides the initial track candidates.
- Track finding which extrapolates the seed trajectories along the expected flight path of a charged particle, searching for additional hits that can be assigned to the track candidate.
- The track-fitting module used to provide the best possible estimate of the parameters of each trajectory.
- Track selection used to minimize the number of fake tracks.

Seed generation is done using only 2-3 hits and provides the initial trajectory parameters and associated uncertainties of potential tracks. Charged particles follow a helical path in the magnetic field that can be described with five parameters. These are obtained either from three 3-D hits, or two 3-D hits and a constraint on the origin of the trajectory assuming that the particle originated near the beam spot. Here, the 3-D hit is any hit that provides a 3-D position measurement. Some weak restrictions (e.g. minimum p_T and their consistency with originating from the pp interaction region) are imposed on seeds in order to reduce the number of hit combinations.
Although tracker seeds could be built starting from either the outer or inner layers of the tracker, the latter is used. The reason for this is the higher granularity of the crystals in the inner pixel tracker which results in higher efficiency of generated tracks.

The track-finding algorithm is based on the KF and begins with a rough estimate of the track parameters provided by the trajectory seed. Then it builds track candidates by adding hits from successive detector layers, updating the parameters at each layer. The track-finding is implemented in four steps.
The first step uses the parameters of the track candidate to determine which layers of the detectors will be intersected by extrapolation of the trajectory.
The second step searches for the compatible silicon modules in the layers returned by the previous step. Because of the design of the silicon tracker, sensors often slightly overlap with their neighbours. This means that it can happen that a particle crosses two sensors in the same layer. To solve this problem, modules in each layer are divided into groups in such a way that a particle passing through one group in a layer can't physically pass through another group in the same layer. This is used in the next step.
In the third step, the groups of hits are formed, each of which is defined by the collection of all the hits from one of the module groups. The hit positions and uncertainties are refined using the trajectory direction on the sensor surface, to calculate more accurately the Lorentz drift of the ionization-charge carriers inside the silicon bulk. In order to check which of the hits is compatible with the extrapolated trajectory, a χ^2 test is applied.
In the last step, the trajectories are updated. From each of the original track candidates, new track candidates are formed by adding exactly one of the compatible hits from each module grouping. Finally, trajectory parameters are then updated at the location of the module surface, by combining the information from the added hits with the extrapolated trajectory of the original track candidate.
For each trajectory, the track-finding algorithm yields a collection of hits as well as an estimate of the track parameters. Still, the full information about the trajectory is only available at the final hit of the trajectory when all hits are known. In addition, constraints such as a beam spot constraint applied to the trajectory during the seeding stage can bias the estimate. Thus, the trajectory is refitted using a Kalman filter and smoothened. This procedure is referred to as track-fitting.

A track-fitting procedure usually results in a large number of fake tracks in case of many hits. The fake track is defined as a reconstructed track that is not associated with a charged particle. To mitigate this problem, tracks are selected on the basis of the number of layers that have hits, the quality of the fit and how likely they originate from a primary interaction vertex. A more in-depth discussion on CMS tracking can be found at [59].

2.4.2 Clustering

The effect of the magnetic field on charged particles is used in the tracker to obtain information about their momentum. However, the trajectory of neutral particles, such as photons and neutral hadrons, will not bend while passing through the tracker and thus tracker alone is not enough for the successful reconstruction of physics objects. For this reason, another important piece of the puzzle in the *PF* algorithm are the energy depositions in CMS calorimeters.
While passing through crystals in ECAL and HCAL, particles (except neutrinos and muons) lose most of their energy and leave it within the calorimeter crystals. Individual crystals in which energy was deposited are grouped in, so-called, *clusters* for further processing. This is a task of *clustering* algorithm in the CMS which

- detects and measures the energy and direction of stable neutral particles, e.g. photons and neutral hadrons
- separates energy depositions originating from charged and neutral particles
- reconstructs and identifies electrons together with all accompanying bremsstrahlung photons
- helps with the energy measurement of charged hadrons for which the track parameters were not determined accurately

The clustering is performed separately for each of the sub-detectors. For this reason, the parameters of clustering, summarized in Table 2.1, are different for each subdetector. A threshold of 150 MeV is specially applied on E_T in the ECAL endcap in order to cope with high noise in that region. It should be noted that no clustering is performed in the HF so that each cell results in one cluster. The procedure can be divided into three main steps:

1. Clustering procedure starts with identifying cluster seeds. A cluster seed is defined as the cell with the highest energy deposition in its closest vicinity and is required to have an energy above the specified threshold.

2. Topological clusters are grown from the seeds by aggregating cells with at least a corner in common with a cell already in the cluster. Each of these cells is also required to have an energy above the threshold set to twice the noise level.

3. An *expectation-maximization* algorithm based on a *Gaussian-mixture model* (GMM) is then used to reconstruct the clusters within a topological cluster. In essence, the GMM is an extension of the K-means clustering algorithm so that the clusters are modelled with Gaussian distributions. GMM postulates that the energy deposits in the M individual cells of the topological cluster arise from N Gaussian energy deposits. Here, N is the number of cluster seeds. The model parameters needed to be fit are the amplitude and the mean of each Gaussian in the (ϕ, η) plane. The with of the Gaussian, σ, is fixed to a different value depending on the considered calorimeter. The fitting is done using an expectation-maximization algorithm that performs the maximum likelihood fit on the parameters. This is an iterative process that is repeated until convergence is achieved.
   Obtained parameters of the fit are used to define *PF clusters*. Finally, these are calibrated to compensate for the bias in the energy arising from the finite cell thresholds during topological clustering as well as the energy loss in the dead material between ECAL and HCAL.

	ECAL		**HCAL**		**preshower**
Cell E threshold $[MeV]$	80	300	800	800	0.06
Number of closest cells to seed	8	8	4	4	8
Seed E threshold $[MeV]$	230	600	800	1100	0.12
Seed E_T **threshold** $[MeV]$	0	150	0	0	0
Gaussian width σ $[cm]$	1.5	1.5	10.0	10.0	0.2

Table 2.1: Parameters used in the clustering algorithm for the ECAL, HCAL and preshower.

2.4.3 Muons

Because muons are much less likely to undergo bremsstrahlung than electrons ($\sim 10^{10}$ smaller probability), their reconstruction is much less challenging. Additionally, other than neutrinos that don't leave their trace in the tracker, muons are the only particles to traverse all subdetector systems. Therefore, combining the information from the tracker and muon chambers provides an exquisite measurement quality of muons.

Muon reconstruction

In the standard CMS procedure, tracks are first reconstructed independently in the inner tracker (tracker track) and in the muon system (standalone-muon track). Three muon classes are reconstructed with the PF algorithm:

- *standalone muon tracks* are built by clustering hits inside DTs and CSCs. The final standalone muon track is obtained by adding the information from the RPC.
- *tracker muon tracks* are built "inside-out" by propagating tracker tracks to the muon system. If at least one DT or CSC segment is geometrically matched to the inner track with $p_T > 0.5\ GeV$ and a total momentum $p > 2\ GeV$, the inner track is considered a tracker muon track.
- *global muon tracks* are built "outside-in" by matching standalone-muon tracks with tracker tracks.

Due to the very high efficiency of the tracker track and muon segment reconstruction, $\sim 99\ \%$ of muons produced within the geometrical acceptance of the muon system are reconstructed as either a global muon track or as a tracker muon track, and very often as both. If the global muons and tracker muons share the same tracker track, they are merged into a single candidate. While global muons improve the momentum resolution with respect to the tracker-only fit, tracker muons recover efficiency for very low-p_T muons that do not always fully traverse the CMS detector [60].

Muon identification

In order to fit the needs of different analyses, several muon identification (ID) criteria have been defined [60]. The algorithm uses variables such as the track fit quality (χ^2), the number of hits per track, or the degree of matching between tracker tracks and standalone-muon tracks (for global muons), and returns a value in a range between 0 and 1. Three main identification classes used in CMS are

- *Loose muon ID* aims to identify prompt muons originating at the primary vertex and muons from light and heavy flavour decays, as well as to maintain a low rate of the misidentification of charged hadrons as muons. A loose muon is a PF candidate that is either a tracker or a global muon.
- *Medium muon ID* is optimized for prompt muons and for muons from the heavy flavour decays. A medium muon is a loose muon with a tracker track that uses hits from more than 80 % of the inner tracker layers it traverses. If the muon is only reconstructed as a tracker muon, the muon segment compatibility must be greater than 0.451. If it is reconstructed as both a tracker muon and a global muon, then
  - its segment compatibility must be greater than 0.303
  - the global fit is required to have goodness-of-fit per degree of freedom (χ^2/dof) less than 3
  - the position match between the tracker muon and standalone-muon must have $\chi^2 < 12$

  - the maximum χ^2 computed by the *kink-finding* algorithm must be less than 20. The kink-finding algorithm splits the tracker track into two separate tracks at several places along the trajectory and, for each split, evaluates whether the two are incompatible with being a single track.

- *Tigh muon ID* aims to suppress muon from decays in flight and misidentifying hadron shower remnants reaching the innermost muon station (punch-through). A tight muon is a loose muon with a tracker track that uses hits from at least six layers of the inner tracker including at least one pixel hit. In addition, it must be reconstructed as both a tracker muon and a global muon, where the tracker muon must have segment matching in at least two of the muon stations and a global muon fit must have $\chi^2/dof < 10$ and include at least one hit from the muon system. Finally, a tight muon must come from the primary vertex.

2.4.4 Jets

Because of the colour confinement, individual quarks and gluons cannot be observed. Instead, they hadronize into more complex objects called hadronic jets which can be reconstructed. Quality of jet reconstruction is crucial, especially in analyses that study processes with VBS topology, such as the one presented in this thesis. Since jet reconstruction is a very involved procedure, only the key insights are discussed here. More can be find elsewhere [61–63].

Jet reconstruction and identification

Jets are reconstructed from PF candidates using the anti-k_T algorithm [64] and the Cambridge-Aachen algorithm [65, 66], as implemented in FastJet version 3.0.1 [67]. In this thesis, the anti-k_T algorithm was used with parameter distance parameter ΔR set to 0.4. In simple terms, the anti-kt algorithm can be understood as the following. One can assume that within an event there is a number of well-separated hard particles with transverse momenta $kt1$, $kt2$, $kt3$... and many soft particles. If there are no other hard particles closer than $2R$ around a given hard particle, then all soft particles in a circle of radius R will be clustered together with the hard particle resulting in a perfect conical jet. If there are two hard particles with distance $R < d_{ij} < 2R$ between them, then there will be two hard jets. If $k_{t1} \gg k_{t2}$ then only jet 1 will be conical since the second jet will miss a part that overlaps with the first jet. If $k_{t1} = k_{t2}$ then neither of the jets will be conical with the overlapping part being equally divided between the two.

In order to obtain a clean jet collection, jets coming from the detector noise (referred to as the noise jets), and fake jets arising from spurious energy depositions in a single sub-detector or instances in which particles from different interactions deposit their energy in the proximity of each other in the detector are cleaned. The fractions of the jet energy carried by certain types of PF candidates clustered into a jet (PF jet energy fractions), together with the number of PF candidates clustered into a jet are used in order to discriminate between noise/fake jets and physical jets. The PF jet ID criteria is summarized in Table 2.2 where two working points are defined: *tight* and *tight lepton veto*. The tight working point is chosen to remove jets originating from calorimetric noise, while the tight lepton veto working point additionally rejects the potential background from miss-reconstructed electron and muon candidates which helps resolving also the ambiguity between isolated lepton candidates and jets reconstructed from single lepton candidates. Prior to 2017, an additional working point, referred to as the *loose* working point, was used. However, since the efficiency of the tight working point is above 99 %, the loose working point is not recommended anymore. In the table, reconstructed electrons are referred to as "charged EM" while photons are referred to as "neutral EM".
In order to successfully recover jet energy, additional calibrations are needed. Since different corrections applied on jets are usually analysis-dependent, this is discussed, in more detail, in section 4.3.

Jet Variables	$\|\eta\| \leq 2.6$	$2.6 < \|\eta\| \leq 2.7$	$2.7 < \|\eta\| \leq 3.0$	$3.0 < \|\eta\| \leq 5.0$
Neutral Hadron Fraction	< 0.90		-	> 0.2
Charged Hadron Fraction	> 0	-		
Neutral EM Fraction	< 0.90	< 0.99	> 0.02 and < 0.99	< 0.9
Charged EM Fraction $\star$	< 0.80		-	
Muon Fraction $\star$	< 0.80		-	
Number of Constituents	> 1	-		
Number of Neutral Particles	-		> 2	> 10
Charged Multiplicity	> 0		-	

Table 2.2: Definition of the PF jet ID criteria. Charged EM fraction refers to the fraction of the jet energy carried by electrons, while neutral EM fraction refers to the fraction of energy carried by photons. Variables with $\star$ are only applied for the tight lepton veto working point.

2.4.5 Particle-flow (PF) link algorithm

The inputs of the PF algorithm, often referred to as the *elements*, are the tracks and clusters discussed in the previous two sections. Any given particle in the detector is expected to produce several PF elements in various subdetectors which is why the reconstruction proceeds with *link algorithm* that connects the PF elements from different subdetectors. The idea behind the link algorithm is to identify elements that are likely to originate from the same particle and should thus be grouped together. In order to prevent the computing time of the link algorithm from growing quadratically with the number of particles, the pairs of elements considered by the link procedure are restricted to the nearest neighbours in the $(\phi,\ \eta)$ plane.

If the two elements are found to be linked, the link algorithm then produces *PF blocks* of elements. This is depicted in Figure 2.11 which shows an example of an event display in which a jet is produced. The track T_1 is linked to ECAL cluster E_1 and HCAL clusters H_1 and H_2. At the same time, the track T_2 is only linked to HCAL clusters H_1 and H_2. These two tracks form the first PF block composed of five elements: T_1, E_1, H_1, T_2 and H_2. The first three correspond to the generated π^- and the last two to the π^+ meson.

Finally, the PF reconstruction proceeds to reconstruct the PF candidates by applying a sequence of lengthy algorithms on the PF blocks. A detailed description of this procedure is not needed in order to follow the analysis presented in this thesis and is thus omitted. Further details on the PF link algorithm can be found in [68]. The output of the PF algorithm is the list of PF candidates that are then used for further processing, e.g., jet reconstruction, sophisticated particle-flow isolation, or the calculation of event-level quantities like missing transverse energy.

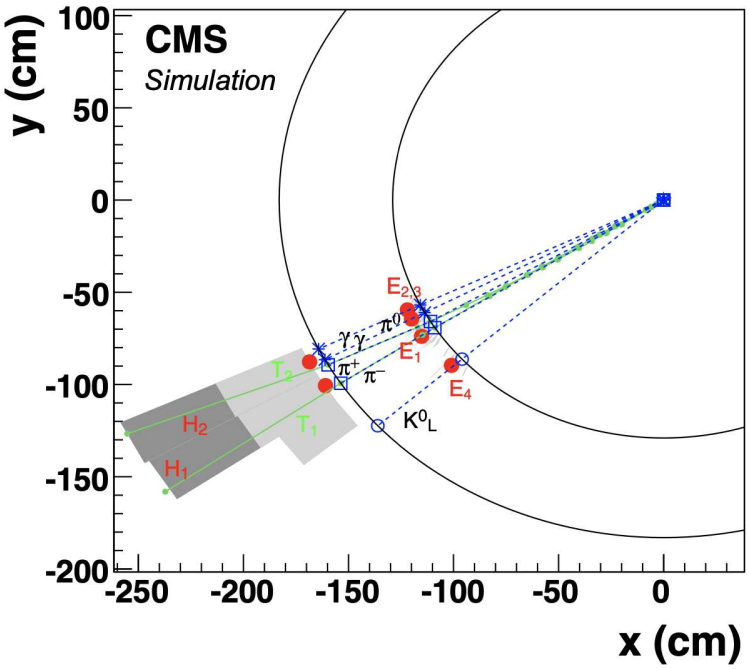


Figure 2.11: An illustration of a jet made of five particles shown in the $(x,\ y)$ plane. ECAL and HCAL surfaces are represented as circles centred around the interaction point. The K_L^0, π^- and two photons from the π^0 decay are detected as four well separated ECAL clusters $E_1,\ E_2,\ E_3$ and E_4. The π^+ does not create a cluster in the ECAL. The two charged pions are reconstructed as charged-particle tracks T_1 and T_2 appearing as green circular arcs. These tracks point towards two HCAL clusters H_1 and H_2. The cluster positions are represented by dots, the simulated particles by dashed lines, and the positions of their impacts on the calorimeter surface by various open markers.

2.5 The future of LHC and CMS

2.5.1 High-luminosity LHC

In 2018, LHC finished Run 2 with an integrated luminosity of roughly $140\ fb^{-1}$ which provided valuable pp collision data used in many analyses to drive our knowledge of fundamental building blocks of nature. In 2019, preparations for Run 3 began, by the end of which an integrated luminosity is foreseen to be doubled in comparison to Run 2. However, further statistical gain without significant luminosity increase will become marginal with the running time needed to half the statistical error of the order of a decade. In order to maintain scientific progress, the LHC will need to be upgraded with a large boost in integrated luminosity. This is the main goal of the *High-Luminosity LHC* (HL-LHC) programme foreseen after Run 3. The LHC timetable is shown in Fig. 2.12. The end of 2024 marks the start of Phase 2 of the LHC operation and the preparation for the start of the HL-LHC in 2031. The projected LHC performance through 2038 is shown in Fig. 2.13 with $3000\ fb^{-1}$ integrated luminosity at $14\ TeV$ c.o.m. energy expected at the end of the HL-LHC phase. Sensitivity increase for the $Z_L Z_L jj$ scattering with VBS topology at the end of the HL-LHC phase is explored in Chapter 5.

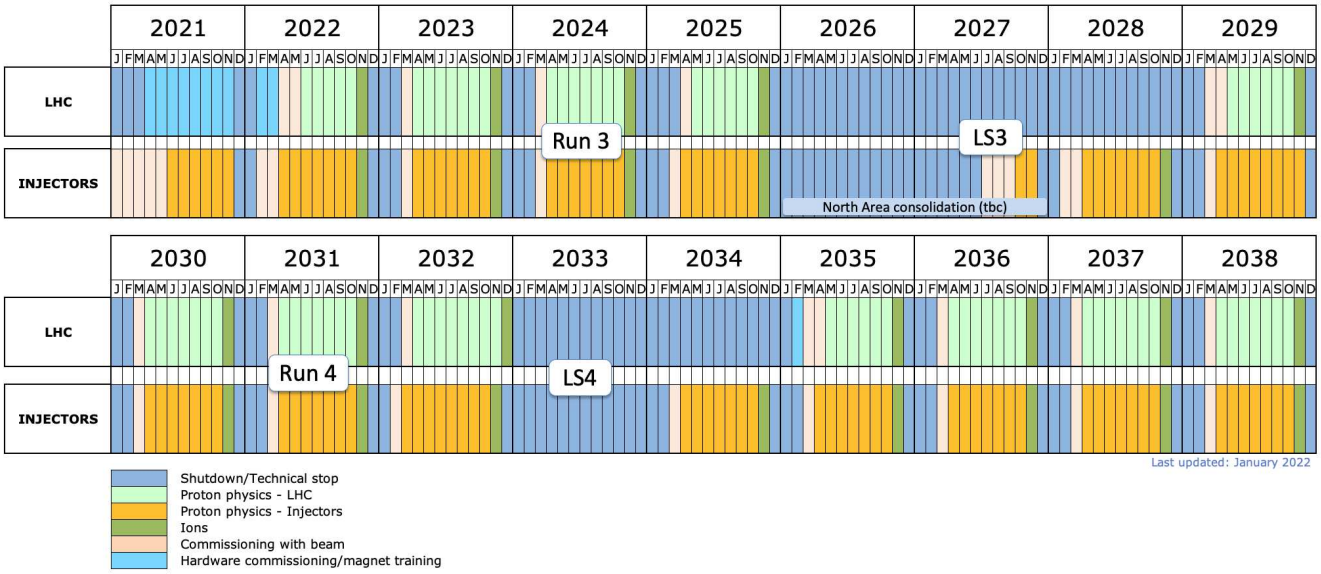


Figure 2.12: Planned schedule for the future operations of the LHC.

In order to cope with a new and harsher environment, the LHC will need to be upgraded. At about $300\ fb^{-1}$ some components of the triplet quadrupoles and their corrector magnets will have received a radiation dose of $30\ MGy$. Although quadrupoles are designed for luminosities of $400 - 700\ fb^{-1}$, some corrector magnets are expected to wear out already above $300\ fb^{-1}$. Since system failure will manifest through a sudden electric breakdown, requiring serious and long repairs, replacement of the inner triplet magnets must be foreseen before the damage occurs. In addition, LHC main dipole will be replaced with a dipole of equal bending strength ($121\ T \cdot m$) obtained by a higher field ($11\ T$) and shorter length ($11\ m$) compared to current LHC dipoles ($8.3\ T$ and $14.2\ m$). Naturally, the full list of LHC upgrades foreseen for the HL-LHC is substantial and is out of the scope of this thesis. An exhaustive description of all accelerator upgrades can be found in other sources [69].

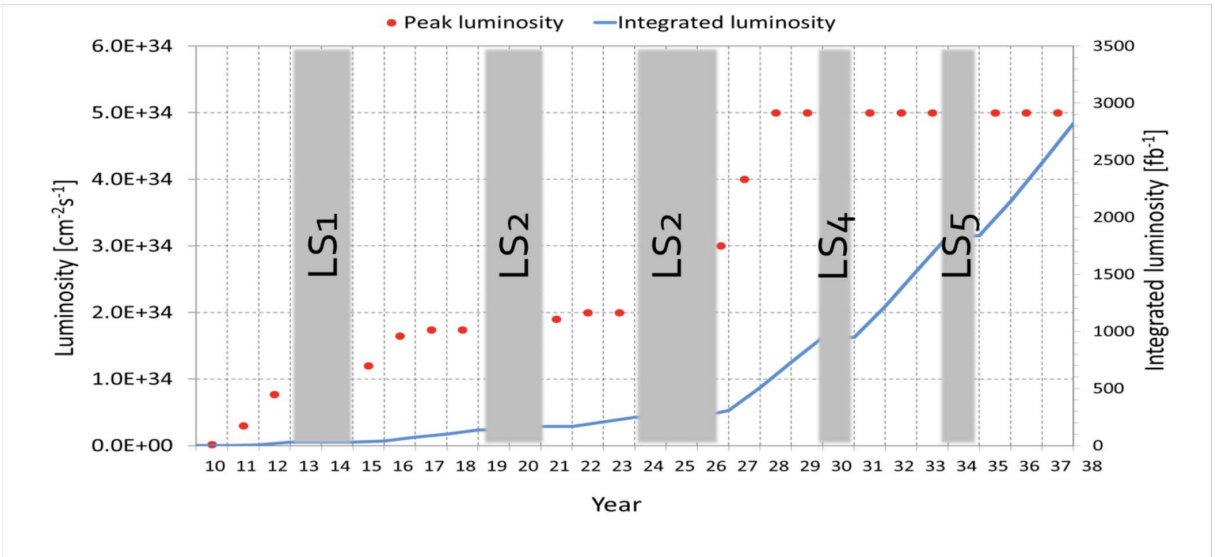


Figure 2.13: Projected LHC performance through 2038, showing preliminary dates for long shutdowns (LS) of the LHC and projected luminosities. Figure is taken from [70].

Since the LHC will produce collisions at a rate of about $5 \times 10^9/s$, the levels of radiation in the material of the detectors and the on-board electronics will cause significant damage and could result in a progressive degradation of the detector performance. Although the CMS detector was designed to withstand severe radiation levels in the LHC environment, radiation levels in a single year at HI-LHC will be similar to the total dose of all operations from the beginning of the LHC program to the start of LS3. Predicted levels of radiation at HL-LHC in the CMS detector are given in Fig. 2.14 which shows the distribution of absorbed dose over the CMS detector for an integrated luminosity of $3000\ fb^{-1}$. The radiation dose will be highest in the most forward regions of the detector, with radiation varying from subdetector to subdetector. For silicon detectors, radiation produces defects in the silicon lattice that change the bulk electrical properties of the silicon resulting in lower signals. The main effect on calorimeters, mainly made of scintillating $PbWO_4$ crystals or plastic scintillating tiles with wavelength-shifting fibres embedded in them, is the loss of transparency. This results in signal losses with up to $90\ \%$ in some cases, and, consequentially, a reduction in the resolution.
Under the foreseen luminosities at the HL-LHC, the PU will become a major challenge for the experiments. At the nominal luminosity of the HL-LHC, the average number of interactions in a single crossing will be approximately 140 and is expected to go up to 200. Most of these interactions are "soft" and do not contribute to the search for new physics at the 0.1-few TeV scale. Only a relatively small fraction of all collisions are "hard" collisions that contain high transverse momentum particles. PU produces many more hits in the tracking detectors, leading to mismeasured or misidentified tracks. In addition, it adds extra energy depositions in calorimeters. Since many analyses, including the one presented in this thesis, require isolated leptons with very little activity around them, energy depositions from PU can cause isolated leptons to appear non-isolated. PU makes triggering and offline reconstruction more challenging. Finally, it increases the amount of data that has to be read out in each bunch crossing so much that, at the HL-LHC, most of the data read out will be associated with the PU rather than the hard scattering collisions. In order to prepare for the intense HL-LHC environment, the CMS detector will be upgraded during LS3. Some important upgrades are described below, while many others are discussed, in detail, elsewhere [71–74].

High Granularity Calorimeter

One of the most emphasized goals of the HL-LHC upgrade programme is the replacement of the existing endcap calorimeters with a *High Granularity Calorimeter* (HGCAL) since the endcap calorimeters are among the subdetectors that will be exposed to the highest radiation levels. Under HL-LHC conditions, current endcaps would degrade very quickly in performance.
HGCAL is a sampling calorimeter consisting of the electromagnetic part (EE) and two hadronic parts (FH and BH) providing a coverage of $1.5 < |\eta| < 3$. It is sometimes also referred to as *CE*, and, therefore, the electromagnetic part as CE-E, and hadronic parts as CE-H. An illustration of the HGCAL is shown in Fig. 2.15.

Silicon is the main active material and is used throughout EE and in the innermost sections of the FH. The EE consists of 28 layers of tungsten and copper plates interleaved with silicon sensors for a total of 26 radiation lengths (X_0) and 1.5 interaction lengths (λ). The FH consists of 12 layers of brass and copper plates interleaved with silicon sensors for a total of roughly $3.5\ \lambda$. Additional 12 layers of the brass-plastic scintillator ($5\ \lambda$) will be added in the BH to ensure the full containment of showers.
Active elements in HGCAL are hexagonal silicon wafers of different sizes and thicknesses, depending on the pseudorapidity. For $1.48 < |\eta| < 2.15$, $1\ cm^2$ wafers with $300\ \mu m$ active material thickness in $|\eta| < 1.75$ region and $200\ \mu m$ active material thickness otherwise. For $2.15 < |\eta| < 3$, $0.5\ cm^2$ wafers with $120\ \mu m$ active material thickness will be used. Sensors are mounted on printed circuit boards (PCB), with a front-end chip bonded to it, and glued on the other face to a copper-tungsten baseplate to form a module. Modules will be mounted on a $6mm$-thick copper

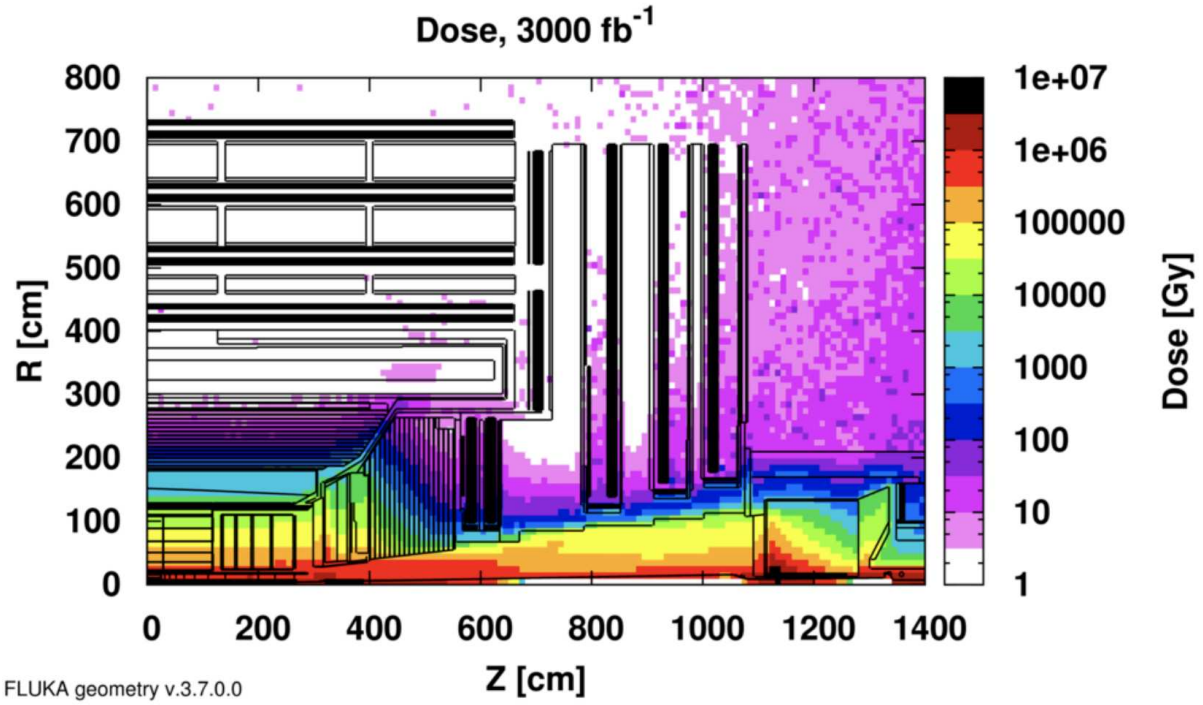


Figure 2.14: Absorbed dose in the CMS cavern after an integrated luminosity of $3000\ fb^{-1}$. R is the transverse distance from the beamline and Z is the distance along the beamline from the Interaction Point at $Z = 0$. The figure is taken from [70]

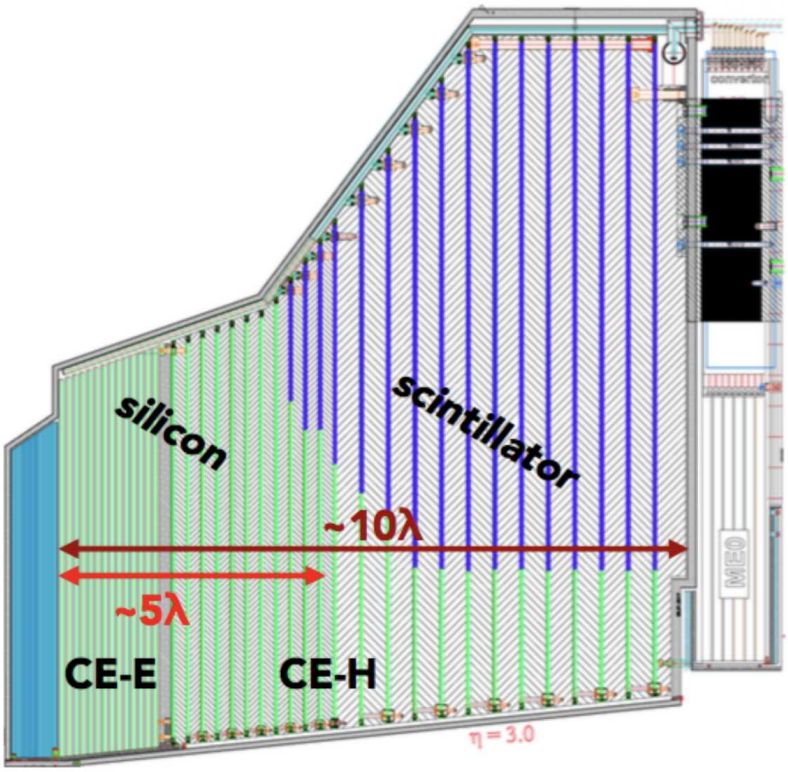


Figure 2.15: Schematic view of the High Granularity Calorimeter design. The figure is taken from [75]

plate with embedded stainless steel pipes for cooling to make a *cassette*. This is illustrated in Fig. 2.16. HGCAL will be cooled down to $-30\ ^{\circ}C$ via evaporating CO_2 system to mitigate leakage current in silicon sensors due to radiation damage. Finally, cassettes are mounted into 12 $30\ ^{\circ}$-sectors.
In the outermost regions of FH and in the BH, radiation levels are lower and plastic scintillating tiles with SiPM readout will be used [76, 77].

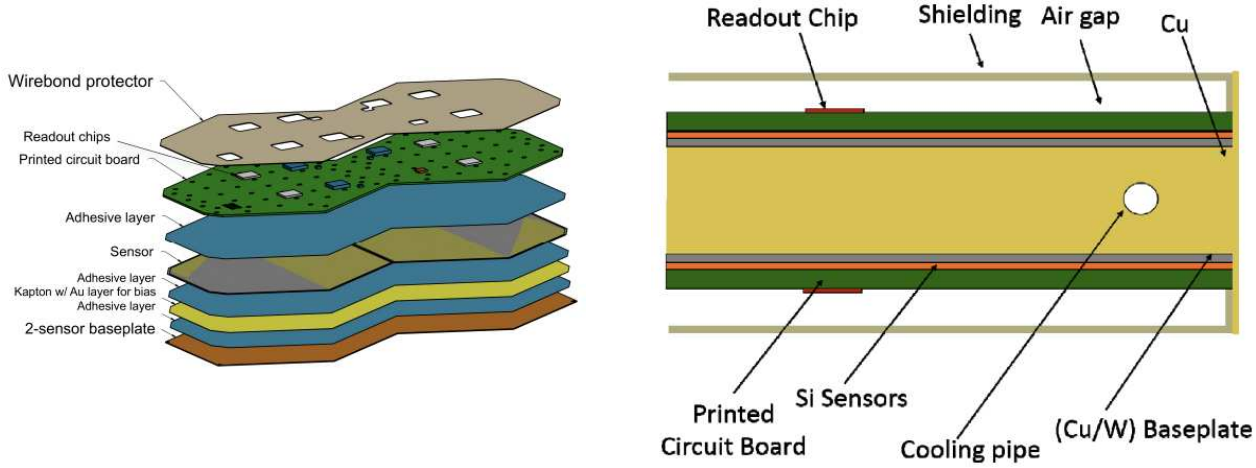


Figure 2.16: Left: Module consisting of the printed circuit board, silicon sensors and baseplate. Right: Sketch of a cassette with modules mounted on either side of the copper cooling plate. The figure is taken from [76]

HF nose

In order to better exploit the coverage up to $|\eta| \approx 4$ of the tracker, an additional detector upgrade feature, referred to as the "*HF nose*", is being considered for the HL-LHC upgrade. This high granularity extension of the HF detector would enable the measurement of electrons up to $|\eta| = 4$. Design-wise, it is based on the HGCAL with 8-inch, $0.5\ cm^2$ hexagonal sensors mounted on the multilayer module. The cassette design is also taken from the HGCAL, together with a design for the cooling system, on-detector electronics and backend electronics. For the EE part, stainless steel (SS) clad Pb absorber is proposed with 6 sampling layers and a total thickness of $0.95\ \lambda$. The hadronic part would consist of the SS plates and 2 sampling layers resulting in a total thickness of $0.9\ \lambda$.
Together with the tracker extending up to $|\eta| = 4$, HF nose will help identify vertices from hard scattering processes and, therefore, suppress PU contribution. In addition, it will help with PF jet reconstruction. Finally, it is expected to improve on electron energy resolution by around 50 % [78]. The HF nose option is considered in Chapter 5 to access the gain in the measurement sensitivity of the scattering of the longitudinal polarization of vector bosons.

2.5.2 High-energy LHC

The European Strategy for Particle Physics (ESPP) update 2013 stated that *"To stay at the forefront of particle physics, Europe needs to be in a position to propose an ambitious post-LHC accelerator project at CERN by the time of the next Strategy"*. Soon, the *Future Circular Collider* (FCC) study [79] was launched as a world-wide international collaboration hosted by CERN. The study covers the option for the FCC hadron collider (FCC-hh) and the lepton collider (FCC-ee), both exploiting 100 km tunnel infrastructure to reach energies of up to 100 TeV. As an intermediate step towards the 100 TeV circular accelerator, the *High-Energy LHC* (HE-LHC) projects had been considered. This section will cover the main ideas behind the HL-LHC project needed to follow the analysis presented in Chapter 5.
At the heart of the HE-LHC project is a pp collider, designed to operate at 27 TeV c.o.m. energy and to deliver $15000\ fb^{-1}$ of collision data during 20 years of operation. The HE-LHC is the next step after the HL-LHC and will use the same tunnel as the HL-LHC. Although its performance is well below the 100 TeV target of the FCC-hh, it is nevertheless a massive upgrade with respect to the HL-LHC bringing a significant increase in both energy and integrated luminosity. The beam parameters for the HE-LHC will be essentially the same as those used at HL-LHC with $2.2 \cdot 10^{11}$ protons within a single bunch and $25\ ns$ between the two bunch crossings. The number of bunches circulating the HE-LHC at any given time will be 2808 - the same as for the LHC and HL-LHC. Compared to the HL-LHC which will use $8.33\ T$ dipole magnets, the HE-LHC will be able to exploit the advancements in the dipole technology and will use $16\ T$ dipoles designed for the FCC-hh.
For the designed bunch spacing of $25ns$, the peak PU in the HE-LHC is expected to be around 460. If needed, this could be reduced to around 200 by levelling, as is planned for the HL-LHC, or by reducing the bunch spacing. By halving the bunch spacing from $25\ ns$ to $12.5\ ns$ the PU would be also halved. In study presented in Chapter 5, 200 PU scenario is assumed.

The main goals of the HE-LHC [79] include

1. Extending the HL-LHC reach in direct searches for new particles by approximately doubling the reach in mass.
2. Providing deeper insights into the nature of the EWSB mechanism and the EW sector of the SM.
3. Improving the precision of the HL-LHC measurements in the EW and flavour sectors
4. Providing higher sensitivity to elusive final states such as the one presented in this thesis.
5. Exploring, in greater detail, the properties of possible future LHC discoveries, confirming preliminary signs of discovery from the LHC, or identifying the underlying origin of new phenomena revealed indirectly (e.g. the flavour anomalies) or in experiments other than those of the LHC (e.g. dark matter or neutrino experiments)

The HE-LHC will bring a lot to the analyses that are, currently, statistically limited. This can be seen from Fig. 2.17 showing the expected increase in statistics at HE-LHC compared to the HL-LHC, for final states of a given mass M produced through various partonic initial states. One can see that the higher energy of HE-LHC is particularly beneficial in the case of the heaviest objects. On the other hand, for the study of low-mass systems, luminosity is the key factor.

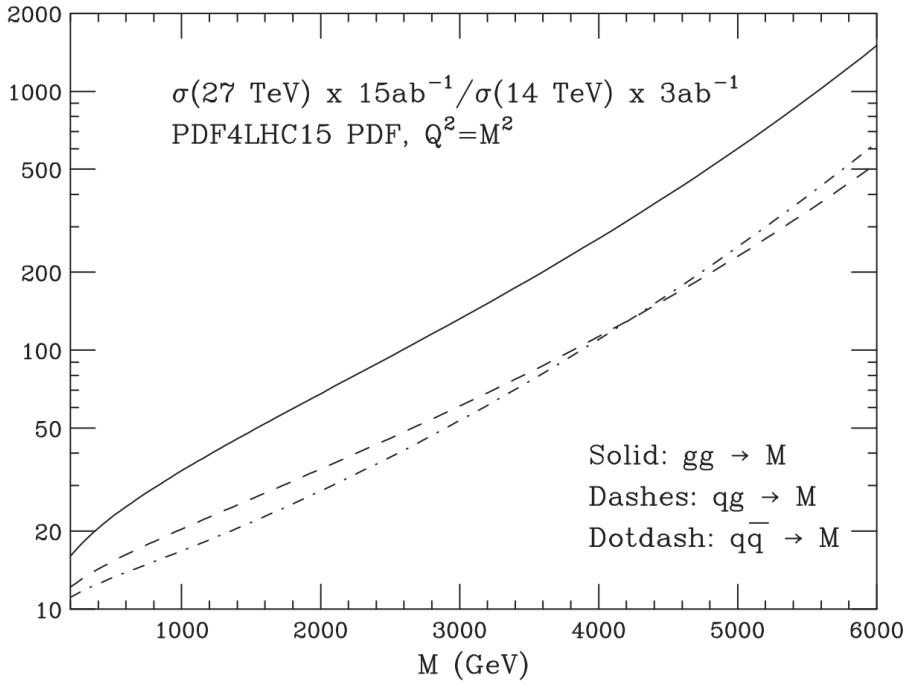


Figure 2.17: Statistics increase at the HE-LHC, relative to the HL-LHC, for the production of a system of mass M, in the three production channels. The figure is taken from [79].

Chapter 3

Electron reconstruction and identification

3.1 Preface to the chapter

After discussing muons and jets at the end of the previous chapter, in this chapter, I will discuss electron reconstruction and selection. The focal point of this chapter is my work on electron efficiency measurements and the derivation of electron scale factors for the full Run 2 period. These results, presented in section 3.4, are an important contribution to the HZZ working group and were used in the $H \to ZZ \to 4l$ analysis. The results presented here are also used in the VBS $ZZ \to 4l2j$ analysis presented in chapter 4 since the electron selection in the two analyses is identical.
In section 3.2.1 I will describe the formation of ECAL clusters and the importance of the superclustering algorithm. In sections 3.2.2 through 3.2.7 I present an overview of the algorithms used to reconstruct electron trajectories, measure electron charge, momentum and energy. Finally, I describe energy corrections, combining momentum and energy measurements as well as the incorporation of discussed algorithms into the particle flow framework.
In section 3.3 I will describe vertex and impact parameter requirements on electrons as well as the identification and isolation algorithms. These are all used to define the electron selection criteria. Finally, in section 3.5 I will summarize the results of the electron efficiency measurements and scale factors.

3.2 Electron reconstruction

Data obtained using a 120 GeV electron test beam showed that an electron impinging directly on the centre of the ECAL crystal will leave 97% of its energy in a 5x5 crystal array centred around the hit crystal [80]. However, due to a large material budget in front of the ECAL, a single electron will often produce a shower of particles through bremsstrahlung and photon conversions before reaching it. The energy loss due to bremsstrahlung is directly dependent on the thickness of the material the electron traverses. An electron will lose, on average, 33% of its energy before reaching ECAL if it propagates through the region with the least material budget that corresponds to $\eta \approx 0$. On the other hand, this goes up to 86%, on average, for electrons traversing through the region with the highest material budget, around $|\eta| \approx 1.4$.
The first effect of bremsstrahlung is the spread of electron energy depositions in ECAL along the ϕ direction due to the magnetic field produced by the solenoid. In order to cope with this, several algorithms were studied in CMS. Additionally, radiation results in a sizable change of curvature of the electron trajectory along the preshower and tracker detectors. All this makes an energy measurement associated with the original electron a challenging task.

3.2.1 Clustering

In order to measure the energy of the primary electron, it is imperative to collect the energy of all particles from the shower produced by its interaction with the detector material. Due to the solenoidal magnetic field, the energy reaching ECAL will be spread along the ϕ direction. The spread in η will usually be negligible except for very low-p_T electrons ($p_T < 3\ GeV$). Two algorithms have been developed to recover the energy spread in ϕ: the "hybrid" algorithm for the ECAL barrel and the "multi-5x5" algorithm for the ECAL endcaps.
The hybrid algorithm exploits the geometry of the ECAL barrel and the shape of the shower to collect the energy deposits in a small window in η and an extended window in ϕ [81]. The algorithm starts from the most energetic crystal in a region that has a transverse energy deposit larger than a predefined threshold ($E_T^{seed} > E_{T,\ min}^{seed}$). The crystal is referred to as the *seed crystal*. From here, 5x1 crystal "dominos" are added around the seed crystal in $\phi > 0$ and $\phi < 0$ direction as long as the transverse energy contained in the domino is larger than another threshold ($E_T^{5x1\ domino} > E_{T,\ min}^{5x1\ domino}$). Contiguous dominos around the seed crystal that contain energy greater than a threshold $E_{min}^{domino-array}$ are grouped within, so-called, *clusters*.
The multi-5x5 algorithm starts by finding crystal seeds defined as the ones having the highest energy amongst the four direct neighbours. Around each seed, starting with the one containing the highest energy, the energy is collected in clusters of 5x5 crystals. Since crystals in different clusters can overlap, a Gaussian shower profile is used to determine the fraction of the energy deposit to be assigned to each of the clusters [82].

In order to collect all the energy contained in the shower, corresponding to the energy of the original electron and spread in the ϕ direction, one final step must be done. In this step, clusters spread in ϕ are joined into *superclusters* (SCs). Two algorithms are combined in CMS for this task.
The first of the two, the "moustache" algorithm, especially useful for measuring very low energy deposits, relies solely on the information from the ECAL and the preshower detector. The algorithm starts by identifying the *seed cluster* around which other clusters are added if they fall in a certain $\Delta\eta$ - $\Delta\phi$ window. Because of the solenoid magnet, the $\Delta\eta$ - $\Delta\phi$ region has a slight bend since the energy spread is more pronounced along ϕ than η, hence the algorithm name. The region defined by the moustache SC is optimized to contain 98% of the shower energy in several bins of cluster seed energy and position along the detector [83].
The second superclustering algorithm, the "refined" algorithm, exploits the tracker information to extrapolate the trajectories of bremsstrahlung photons and the tracks of converted electron pairs in order to decide whether a given cluster should belong to the SC. Although it uses a moustache algorithm as a starting point, it is capable of increasing or decreasing the number of clusters in the SC. The refined algorithm ultimately determines all ECAL-based quantities of electron and photon objects. An illustration of the superclustering algorithm is shown in the top row of Fig. 3.1. The bottom row shows the reconstructed to generated energy ratio with and without the superclustering algorithm in the barrel and endcap regions.

3.2.2 Track reconstruction

When the energy loss due to bremsstrahlung radiation is significant, a classic KF approach will not be able to follow the changes in the curvature of the track and, thus, the tracks reconstruction efficiency will suffer. In order to better cope with non-Gaussian bremsstrahlung radiation losses, a dedicated algorithm, based on the *Gaussian Sum Filtering* (GSF), has been developed [85]. In essence, unlike KF which uses a single Gaussian to model the radiation loss, the GSF approach relies on mixing multiple Gaussians to approximate the energy loss distribution. In essence, the electron trajectory is reconstructed by collecting the hits that belong to a track and fitting the track parameters using the GSF algorithm. In the end, the backward fit is applied in order to optimize the trajectory parameters.

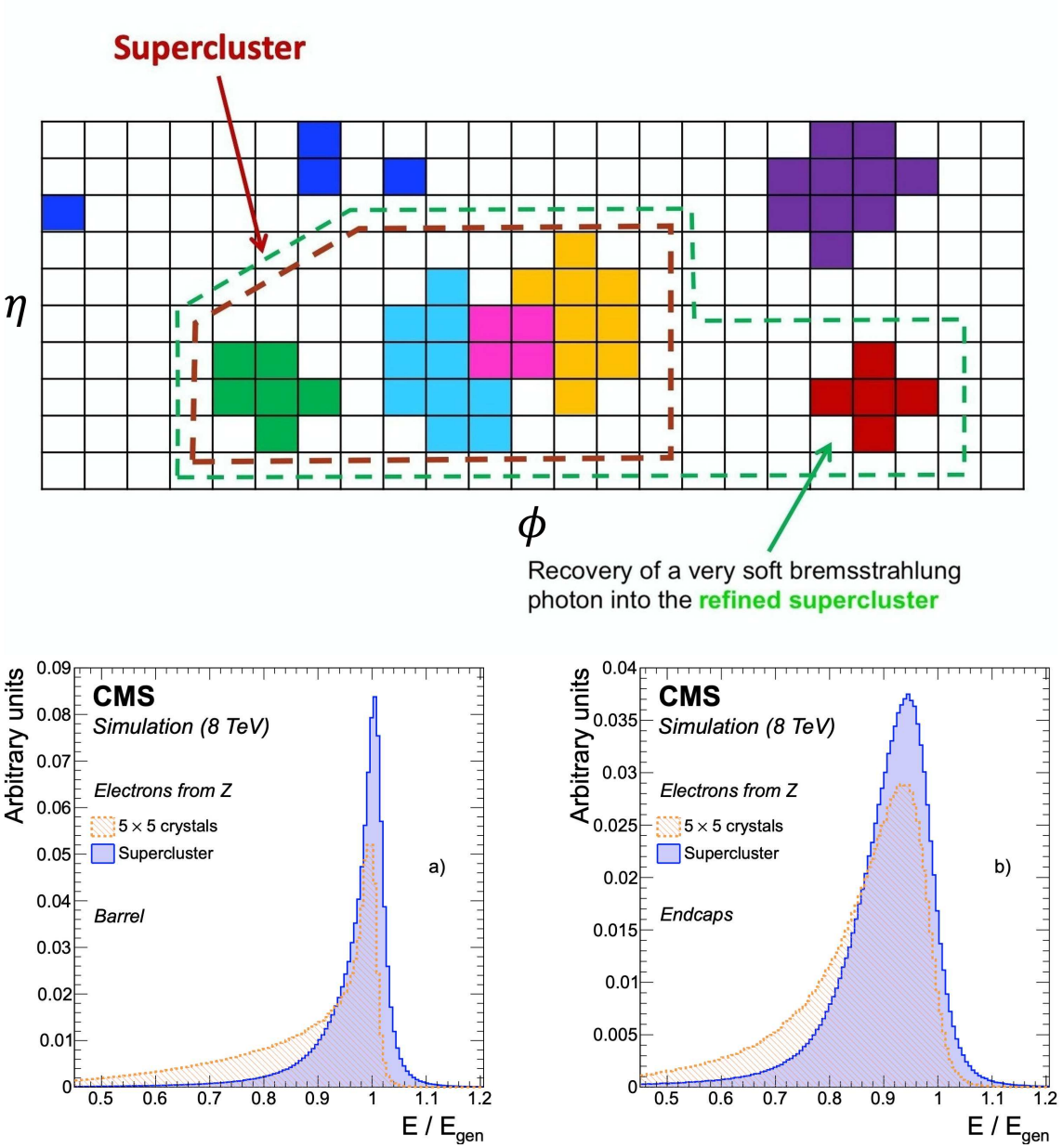


Figure 3.1: The top row shows an illustration of the superclustering algorithm. The bottom shows the comparison of the distributions of the ratio of reconstructed and generated energy for simulated electrons from the Z boson decays in the barrel (left), and the endcaps (right), for energies reconstructed using superclustering (solid histogram) and a matrix of 5x5 crystals (dashed histogram). No energy correction is applied to any of the distributions. The bottom plot is taken from Ref. [84]

Seeding

Due to the more complex and, thus, CPU-intensive nature of the GSF algorithm, the track parameter estimation cannot be performed on all tracks reconstructed in the tracker. The first step in the track reconstruction is finding two or three hits in the tracker from which the track can be initiated. This is referred to as the *track seeding* and is of high importance since it can affect the reconstruction efficiency. The trajectory seeding can be either "ECAL-driven" or "tracker-driven".

The ECAL-driven approach first selects mustache SCs with transverse energy $E_{SC,\,T} > 4\ GeV$ and with $H/E_{SC} < 0.15$ where the E_{SC} is the SC energy and H is the sum of the HCAL tower energies within a cone of $\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2} = 0.15$ centered at the SC position. Hits in the pixel layers are predicted using the energy-weighted position of SCs, assuming the helical trajectory of electrons in the magnetic field (and therefore no radiation losses) [84]. Here, both positive and negative charge hypothesis is tested. The first hit is searched for starting from the innermost pixel layer outward until it is found. When two hits of a tracker seed are matched within a certain $\Delta z \times \Delta\phi$ ($\Delta r \times \Delta\phi$) window for the barrel pixel detectors (forward pixel disks and endcap tracker) to the SC-predicted trajectory, they are selected for seeding a GSF track. The $\Delta z \times \Delta\phi$ ($\Delta r \times \Delta\phi$) windows are defined to take into account the fact that the trajectories of electrons deviate from perfect helices due to radiation losses.

The tracker-driven trajectory seeding starts by going through all generic tracks (not limited to electrons) with $p_T > 2\ GeV$ obtained using the KF approach. A multivariate algorithm is then used to check whether any of these tracks are compatible with either SC position. If so, their seeds are used to initiate a GSF track.

The ECAL-driven approach is more suited for the high-p_T isolated electrons while the tracker-driven approach is designed to recover efficiency for low-p_T or nonisolated electrons. In the end, the two approaches are combined to give an overall $> 95\%$ seeding efficiency for simulated electrons originating from the Z boson decay. The performance of the seeding algorithms is checked with the data showing a good agreement [84].

Trajectory building

The collection of trajectory seeds obtained by combining the ECAL-driven and tracker-driven approach is used to initiate the reconstruction of electron tracks. Starting from each track seed, compatible hits in the next layers are searched for using the KF algorithm to iteratively build the electron trajectory, with the electron energy loss modelled using a Bethe-Heitler distribution [86]. This is done until the last tracker layer unless no hit is found in the consecutive layers. A minimum of five hits is required to create a track. For each layer, the compatibility between the predicted and measured hit is calculated using the χ^2 test. No cut on the χ^2 is imposed for electrons. Instead, many trajectories are grown in parallel and only the two best candidates, with the smallest values of χ^2, are kept in the end. It can happen that a tracker hit is assigned to multiple electron trajectories. In this case, the trajectory with fewer hits is dropped. Alternatively, if the number of hits is the same, the track with higher χ^2 is dropped.

Track parameter estimation

When all the hits are collected, the GSF fit is performed to estimate the track parameters. For each hit, the GSF algorithm uses the parameters of all Gaussians that enter the mixture to model the energy loss in that layer. One possible approach for the electron momentum estimate is to take the weighted mean of all the components. An alternative is to take only the most probable value (i.e. the mode) of the probability density function. The "weighted mean" approach provides the best sensitivity to the momentum change along the track due to radiation emission, while the "mode" approach is better suited for obtaining an estimation, least affected by bremsstrahlung emission, of the most probable track parameters [87]. The two approaches are compared in Figure 3.2 using the p_T/p_T^{gen} ratio for simulated electrons from the Z boson decay [82]. As can be seen from the figure, the peak of the GSF mean distribution is slightly biased towards the higher values of the p_T/p_T^{gen} spectrum. This shows that the bulk of the non-radiating electrons will have a wrongly assigned value of the transverse momentum in this approach. On the other hand, the GSF mode approach gives a better resolution around the peak. In addition, even though the p_T/p_T^{gen} distribution shows a pronounced tail towards the lower values of the spectrum, which is expected since

photon emission results in a more curved track than predicted from the most probable value, it is peaking exactly at unity meaning that, for electrons that don't radiate a lot, it assigns the correct value of the transverse momentum. For these reasons, the mode approach is used to characterize all the parameters of electron tracks.

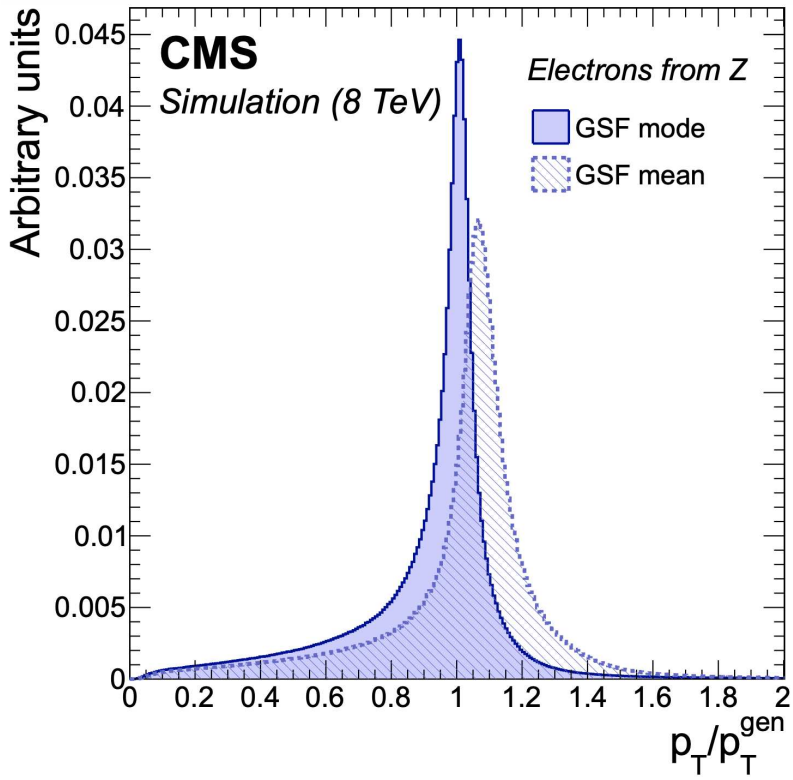


Figure 3.2: Distribution of the ratio of reconstructed over generated electron p_T in simulated $Z \to e^+e^-$ events reconstructed through the most probable value of the GSF track components (solid histogram) and its weighted mean (dashed histogram). The figure is taken from Ref. [84].

Since the described trajectory-building approach enables the collection of hits up to the outermost layers of the tracker, it is possible to extract track parameters close to the surface of the ECAL. This is used to assess the fraction of the energy lost due to bremsstrahlung radiation using the momentum at the innermost layer position (p_{in}) and the momentum at the outermost layer position (p_{out}). This variable, defined as $f_{brem} = 1 - \frac{p_{out}}{p_{in}}$, is used to define electron classes (see section 3.2.4) and, also, in the MVA-based electron identification (see section 3.3.2). Finally, it is used to assess whether the material budget is simulated properly as a function of η (since it measures the amount of bremsstrahlung).

3.2.3 Charge estimation

The electron charge measurement can become more complex in the case of early bremsstrahlung followed by photon conversion. The resulting electromagnetic showers can lead to very complex hit patterns, and the contributions from conversion legs can be wrongly included in the fitting of the electron track. Thus, three methods are combined in CMS to minimize the probability of mismeasuring electron charge:

1. sign of the GSF track curvature
2. curvature of the associated KF track matched to a GSF track when at least one hit is shared in the innermost region

3. sign of the difference in ϕ between the vector joining the beam spot to the SC position and the vector joining the beam spot and the first hit of the electron GSF track

The electron charge is the majority vote of the three charge measurements. The misidentification probability is predicted by the simulation to be 1.5% for reconstructed electrons from Z boson decays and is an improvement by a factor of two with respect to GSF track curvature measurement only. In addition, misidentification probability at $|\eta| > 2$ are predicted to be below 7%. Even higher purity can be achieved, at the price of a p_T- and η-dependent efficiency loss, by requiring all three charge measurements to agree. In that case, misidentification probabilities of less than 0.2% in the central part of the barrel, less than 0.5% in the outer part of the barrel, and less than 1% in the endcaps are achieved. This comes at the price of $\approx 7\%$ efficiency loss for electrons coming from the Z boson decays. All predictions discussed above closely match the observations in data [84].

3.2.4 Classification

The previously defined variable, f_{brem}, together with a bremsstrahlung fraction in the ECAL defined as $f_{brem}^{ECAL} = 1 - \frac{E_{ele}^{PF}}{E_{SC}^{PF}}$ are used to define five classes of electrons. Here, E_{ele}^{PF} and E_{SC}^{PF} are the electron-cluster energy and SC energy measured with the PF algorithm respectively.

1. The "golden" electrons are those with little bremsstrahlung and thus will provide the most accurate estimation of momentum. They are defined by a SC built from a single ECAL cluster and $f_{brem} < 0.5$.

2. "Big-brem" electrons have a large amount of bremsstrahlung radiated in a single step, either very early or very late along the electron trajectory. They are defined by a SC built from a single ECAL cluster and $f_{brem} > 0.5$.

3. "Showering" electrons have a large amount of bremsstrahlung radiated all along their trajectory. They are defined by a SC built from several ECAL clusters.

4. "Crack" electrons are defined by a SC seed crystal adjacent to an η boundary between the modules of the ECAL barrel, between the ECAL barrel and endcaps, or at the high $|\eta|$ edge of the endcaps.

5. "Bad track" electrons are defined by a significantly larger calorimetric bremsstrahlung fraction compared to the track bremsstrahlung fraction ($f_{brem}^{ECAL} - f_{brem} > 0.15$). These are electrons with a poorly fitted track in the innermost part of the trajectory.

3.2.5 Energy corrections

The idea behind clustering energy deposits in SCs is to reduce energy losses due to bremsstrahlung and photon conversions and thus improve upon the energy estimation of the primary electron. However, several effects can impact the estimation of SC energy. These are the energy leakage in ϕ or η outside the SC, the energy leakage into the gaps between the crystals, modules, and supermodules, as well as the transition region between the barrel and the endcaps, the energy leakage into the HCAL, the energy loss due to interactions in the material before the ECAL and the additional energy coming from pileup interactions. All these effects result in systematic variations of the energy measured in the ECAL and degrade the electron energy measurement. In order to improve the resolution, different multivariate techniques have been developed in CMS. The regression technique uses simulated events only, while the energy scale and resolution corrections are based on the comparison between data and simulation. Since the details of this procedure are not essential for understanding the work presented in this thesis, only the key elements are discussed here. An interested reader can find more details in Ref. [82].

Energy corrections with multivariate regressions

The multivariate regression for the SC energy correction defines as a target the ratio between the true energy of an electron and its reconstructed energy. Therefore, the regression prediction is used as the correction factor applied to the measured energy to obtain the best estimate of the true energy. The regression is implemented via a gradient-boosted decision tree (BDTG) (for details on BDTG see section 4.6.2) with a double-sided Crystal ball (DSCB) function [88] used in the regression algorithm. Through the training phase, the regression algorithm performs an estimate of the parameters of the DSCB probability density as a function of the input vector of the object and event characteristics. The electron energy correction is obtained by applying the regression algorithm in three steps. A first regression gives the correction of the SC energy, a second regression gives an estimate of the SC energy resolution and the last regression yields the final energy value correcting the combined energy estimate from the SC and the electron track information.

Energy scale and smearing corrections

Even after introducing energy corrections with the multivariate approach discussed above, small differences remain between the data and the simulation an example being a resolution which is better in the simulation than in the data. Hence, additional smearing has to be applied to the electron energy in simulations so that the peak position of the Z boson mass in the simulation matches that in the data. The electron energy scale is corrected by varying the scale in the data to match that observed in simulated events. The magnitude of the final correction is below 1.5% with uncertainty as small as 0.1% for the barrel and 0.3% for the endcap.
These corrections are obtained using the "fit method" and the "smearing method", both developed in Run 1. In the former, an analytic fit is performed to the invariant mass distribution of the Z boson by convoluting the Breit-Wigner (BW) and the one-sided Crystal ball (OSCB) function. The latter utilizes the simulated Z boson invariant mass distribution as a PDF in a maximum likelihood fit to the data. The difference in width between the data and simulation is described by an energy smearing function applied to the simulation.

The final electron energy resolution, after all corrections are applied, ranges from 2 - 5% depending on the electron η and the amount of energy lost due to the bremsstrahlung. The performance of energy corrections in data is shown in Figure 3.3 with the $Z \to ee$ mass distribution before and after corrections. The result is a peak in data that is better matched to the one in the simulation. The improvement is more pronounced in the endcap region. Additionally, one can see in the same figure an improvement in the energy resolution after applying energy corrections.

3.2.6 Combining energy and momentum measurements

The electron momentum estimate can be improved by combining the corrected energy measurements with the track momentum measurement. At low electron energies ($\lesssim 15\ GeV$), and for electrons near gaps in detectors, the track momentum is, in general, more precisely measured than the ECAL SC energy. The two approaches are combined using a regression technique that defines a weight w that multiplies the track momentum in a linear combination with the estimated SC energy as $p = wp + E_{SC} \cdot (1 - w)$. The variables used to train the regression BDT are the corrected ECAL energy, the track momentum estimate, the uncertainties of the two, the ratio of the corrected ECAL energy over the track momentum as obtained from the track fit, the uncertainty in this ratio, and the electron category, based on the amount of bremsstrahlung [89].

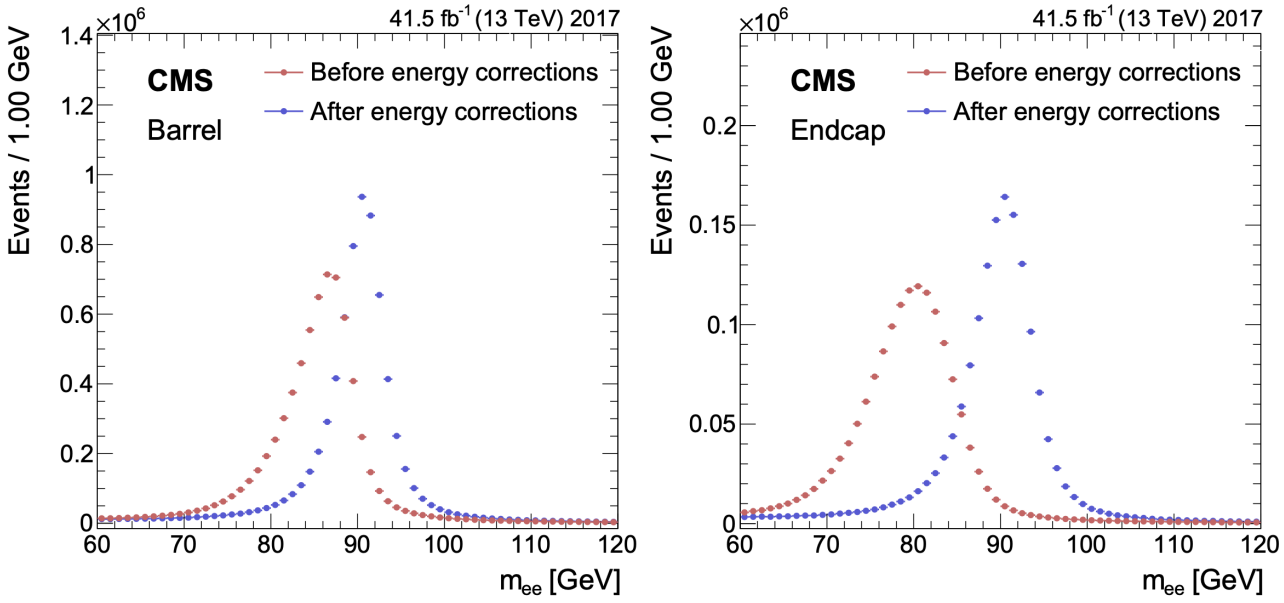


Figure 3.3: Dielectron invariant mass distribution in data before and after energy corrections (regression and scale corrections) for barrel (left) and endcap (right) regions for $Z \to ee$ events. The figure is taken from Ref. [82]

After combining the two estimates, the bias in the electron momentum is reduced in all regions and in all electron classes. An exception are the showering electrons in the endcaps, where the bias becomes slightly worse. The effective resolution, defined as the smallest interval around the peak position containing $\approx 68\%$ of the distribution, in the combined electron momentum can be seen in Figure 3.4 as a function of its p_T compared to the effective resolution of the corrected SC energy for golden electrons in the barrel and for showering electrons in the endcaps. The improvement is around 25% for electrons with $p_T \approx 15\ GeV$ in the barrel. For the golden electrons with $p_T < 10\ GeV$, this can reach 50%. More details on this topic can be found in Ref. [84].

3.2.7 Integration with particle-flow framework

Contrary to the Run 1, where different reconstruction algorithms were used for electrons, electron reconstruction in CMS is now fully integrated into the PF framework. ECAL clusters, SCs, GSF tracks and generic tracks associated with electrons, as well as the conversion tracks and associated clusters, are all imported into the PF algorithm that links the elements together into blocks of particles. These blocks are resolved into electron and photon objects, starting from either a GSF track or a SC, respectively. No difference between electrons and photons exists at this stage. Electron and photon objects are built from the refined SCs based on loose selection criteria (for clarification on selection criteria see section 3.3). All objects that pass the selection criteria, and have an associated GSF track, are labelled as electrons. Objects that pass the selection criteria but don't have a GSF track associated with them are identified as photons. This collection is referred to as the e/γ collection.
To separate electrons and photons from hadrons in the PF framework, a tighter selection is applied to decide if they are accepted as an electron or an isolated photon. If the object passes both the electron and the photon selection criteria, its object type is determined by whether it has a GSF track with a hit in the first layer of the pixel detector. If it fails the electron and photon selection criteria, its ECAL clusters and generic tracks are considered to form neutral hadrons, charged hadrons or nonisolated photons in the PF framework.

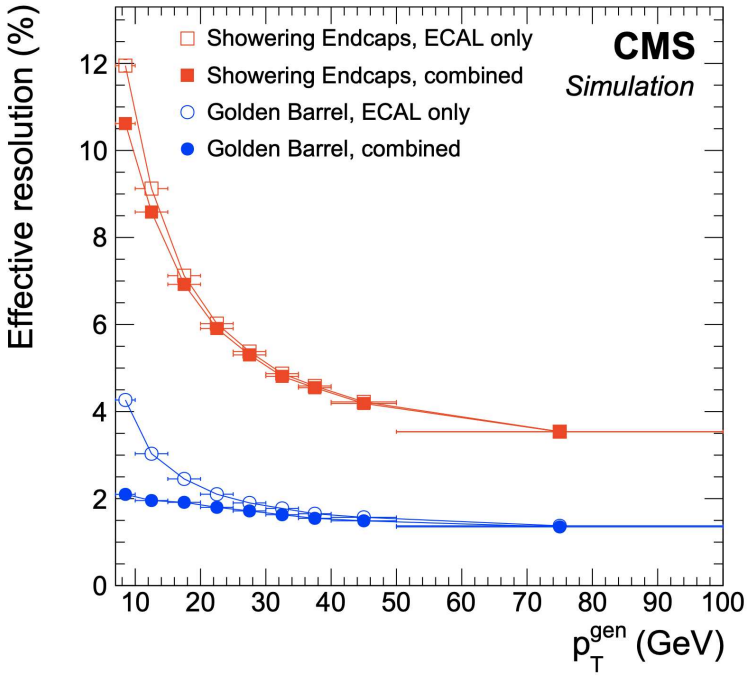


Figure 3.4: The effective resolution, as a function of the generated electron p_T, in electron momentum after combining the corrected SC energy and momentum estimates (solid symbols) compared to that of the corrected SC energy (open symbols). Golden electrons in the barrel (circles) and showering electrons in the endcaps (squares) are shown as examples. Electrons are generated with uniform distributions in η and ϕ and the resolution is shown after applying the spreading corrections. The figure is taken from Ref. [84]

3.3 Electron selection

The main goal of electron selection is to reduce the rate of fake electrons coming from various sources and thus contaminating the analysis. The selection criteria described in this section are used for the $H \to ZZ^* \to 4l$ analysis, where the lepton efficiency enters the selection with the power of four. Full details on selection criteria can be found in [90, 91]. Only the main points needed to understand the electron efficiency measurements discussed in section 3.4 are outlined here. In general, electron selection can be split into three blocks: kinematic and impact parameter selection, electron identification and electron isolation.

3.3.1 Kinematic and impact parameter selection

Because of the tracker acceptance, only electrons with $|\eta| < 2.5$ are considered in the analysis. Additionally, in order to mitigate the effect of the background, especially in the very low p_T region, as well as to account for the difficulties in reconstructing tracks and measuring momentum in this region, only electrons with $p_T > 7\ GeV$ are kept.

Loose vertex requirements defined as

$$|d_{xy}| < 0.5\ cm$$

$$|d_z| < 1\ cm$$

where $|d_{xy}|$ refers to the absolute value of the impact parameter, with respect to the primary vertex, in the transverse plane, and $|d_z|$ is the absolute value of the impact parameter along the z axis are imposed on electron candidates.

A further selection on the impact parameter is introduced in order to reduce the background that doesn't originate from the primary vertex but, rather, from bremsstrahlung photons, photon conversions and heavy flavor decays. In general, tracks of these secondary electron candidates (background in this analysis) will not point to the primary vertex and this can be used to separate them from primary electrons. The *impact parameter*, IP_{3D}, is defined as the algebraic distance, in the 3-dimensional space, between an electron candidate and the primary vertex. However, instead of the impact parameter, the significance of the impact parameter, SIP_{3D}, is used by dividing the impact parameter by its uncertainty. The selection then requires

$$|SIP_{3D}| = \frac{|IP_{3D}|}{\sigma_{IP_{3D}}} < 4$$

3.3.2 Identification

By imposing the selection on the significance of the impact parameter, backgrounds originating from secondary vertices are suppressed. However, hadronic jets (and remaining photon conversions) can mimic genuine electron energy depositions in the calorimeter. In order to distinguish signal electrons from the backgrounds such as reconstructed tracks from $\pi^{\pm}$ in vicinity of an electromagnetic cluster from $\pi^0 \to \gamma\gamma$, a complex *electron identification* algorithm was designed. In the CMS, two approaches are used for the electron identification: the *cut-based* approach and the *MVA-based* approach.

Cut-based electron identification

In the cut-based approach, one applies cuts on a set of tracker and ECAL-related variables. Four working points, corresponding to different signal efficiencies, are used in CMS. The "veto" working point corresponds to an average signal efficiency of about 95%. The "loose" working point corresponds to a signal efficiency of around 90% and is used in analyses with low backgrounds to electrons. The "medium" working point corresponds to an average signal efficiency of around 80%. Finally, the "tight" working point corresponds to roughly 70% signal efficiency and is used in analyses where large background contamination is expected.

MVA-based electron identification

Since the $H \to ZZ^* \to 4l$ channel requires a high signal efficiency, a loose ID, capable of reducing fake electrons, in particular in the low-p_T region, was developed. It uses a set of variables, summarized in Table 3.1, to produce a single MVA classifier using boosted decision tree (BDT) techniques. Three main categories of variables enter the training of the BDT:

- observables based on the shape of the ECAL clusters, an example being the width of the cluster, specifically in the η direction
- observables based on the tracking information such as f_{brem} describing the energy lost through bremsstrahlung
- observables that describe the quality of the matching between the supercluster and the track, an example being the ratio of the supercluster energy over the track momentum

The output of the BDT training is the score for each electron candidate, which is peaking close to unity for signal electrons and to zero for background electrons.

	Observable	Definition
Cluster shape	$\sigma_{i\eta i\eta}$	Energy-weighted standard deviation along η within a 5×5 block of crystals centered on the highest energy crystal of the seed cluster
	$\sigma_{i\phi i\phi}$	Similar to $\sigma_{i\eta i\eta}$ but in the ϕ direction
	η width	SC width along η
	ϕ width	SC width along ϕ
	$1 - E_{5\times1}/E_{5\times5}$	$E_{5\times5}$ is the energy computed in the 5×5 block of crystals centered on the highest energy crystal of the seed cluster, and $E_{5\times1}$ is the energy computed in the η strip of crystals containing it
	R_9	Energy sum in the 3×3 block of crystals centred on the highest energy crystal, divided by the SC energy
	H/E	Energy collected by the HCAL towers within a cone of $\Delta R = 0.15$ centred on the SC position, divided by the SC energy
	E_{PS}/E_{raw}	Energy fraction deposited in the preshower detector divided by the raw SC energy
Tracking	$f_{brem} = 1 - p_{out}/p_{in}$	Fractional momentum loss as measured by the GSF fit. The momenta p_{in} and p_{out} are the innermost and outermost estimates respectively.
	N_{KF}	Number of hits of the KF track (when reconstructed)
	N_{GSF}	Number of hits of the GSF track
	χ^2_{KF}	χ^2 of the KF track (when reconstructed)
	χ^2_{GSF}	χ^2 of the GSF track
	$N_{miss.\ hits}$	Number of expected but missing inner hits in the first tracker layers
	$P_{conv.}$	Fit probability for a conversion vertex associated with the electron track
Track-cluster matching	E_{SC}/p_{in}	Ratio of the supercluster energy to the track momentum at the innermost track position
	E_{ele}/p_{out}	Ratio of the energy of the cluster closest to the electron track and the track momentum at the outermost track position
	$\frac{1}{E_{SC}} - \frac{1}{p}$	Deviation of the SC energy from the electron momentum obtained by combining ECAL and tracker information
	$\Delta\eta_{in} = \lvert\eta_{SC} - \eta_{in}\rvert$	Distance between the energy-weighted center of the SC and the expected shower position as extrapolated from the GSF trajectory state at the vertex
	$\Delta\phi_{in} = \lvert\phi_{SC} - \phi_{in}\rvert$	Same as $\Delta\eta_{in}$, but in the ϕ direction
	$\Delta\eta_{seed} = \lvert\eta_{seed} - \eta_{out}\rvert$	Distance between the η of the seed cluster and the expected shower position as extrapolated from the GSF trajectory state of the outermost hit

Table 3.1: List of input variables, divided into three categories, that enter the BDT training for the MVA-based electron identification used in the $H \to ZZ \to 4l$ analysis.

3.3.3 Isolation

Fake electrons from hadronic jets can be mitigated by means of *isolation*. Prompt electrons are characterized by the absence of activity around them. The isolation can be defined using the PF candidates reconstructed with a momentum direction within a predefined isolation cone.
The isolation variables are obtained by summing the transverse momenta of charged hadrons, neutral hadrons and photons within an isolation cone defined by $\Delta R = 0.3$ and subtracting the contribution of the pileup. The combined per-electron isolation is constructed by combining different isolation-related observables:

$$I = \sum_{\substack{charged \\ hadrons}} p_T + max\left[0, \sum_{\substack{neutral \\ hadrons}} p_T + \sum_{photons} p_T - p_T^{PU}\right]$$

where $p_T^{PU} = \rho \times A_{eff}$ is the pileup correction for electrons calculated following the *FASTJET* technique [92–94].

The problem with using the isolation variable as defined above comes from the consideration of fake electrons in the background. For example, the p_T of the fake lepton inside a jet increases with the energy of the jet. If the energy of the jet is small, the activity surrounding the fake electron will be small and cutting simply on the p_T could lead to the fake electron being wrongly classified as an isolated electron. Therefore, the thresholds applied to the isolation quantities should depend on the particle energy. For this reason, the *relative isolation*, $I_{rel} = I/p_T^e$, is used. As discussed in section 3.4.2, electron isolation is included in the training of the MVA-based ID.

3.4 Electron efficiency measurements

In the previous section, electron selection requirements were defined. Depending on the analysis, one may need different selection criteria, which lead to different electron efficiency. Therefore, it is crucial to quantify the efficiency of the chosen selection criteria since these effects have to be included in the analysis. The same has to be done for the reconstruction procedure discussed in the first part of the chapter. One approach can be to estimate efficiencies using the simulations. However, because the detector effects aren't described perfectly by the simulation, this can lead to undesired bias in the estimation of the reconstruction or selection efficiency. In order to circumvent this issue, efficiencies are extracted directly from the data using the *Tag and Probe* (TnP) approach. For the electron efficiency measurements, the $Z \to ee$ channel is used to estimate the electron selection efficiencies.
In addition, the agreement between efficiencies in the data and simulation varies between the different regions of the detector and for different values of the electron p_T. This results in some disagreement, in most variables used in the analysis, between the simulation and the data. The differences in efficiency between the data and simulation are measured in various η and p_T bins using the TnP approach and *scale factors* are obtained by dividing the efficiency in the data by that in the simulation. These are applied to the simulation in order to correct for the efficiency difference.

3.4.1 Tag and Probe method

In order to measure the efficiency of the desired selection, one needs a pure sample of electrons. This can be achieved by using the decay products of a familiar resonance such as the Z boson which ensures a high purity. The Tag and Probe (TnP) approach is used in this analysis to measure electron selection efficiency.

The TnP method starts with selecting a set of Z bosons that decay into pairs of oppositely charged electrons. These pairs of electrons are required to have a mass within a window of $60\ GeV < m_{ee} < 120\ GeV$ which

ensures that genuine $Z \to ee$ decays are selected. However, some background events, coming mainly from the W+jets or QCD multijet processes, may pass this requirement as well. In order to make sure that the efficiency is measured for signal electrons, one electron, referred to as the *tag*, is required to pass a very tight selection. The corresponding opposite sign electron, referred to as the *probe*, is used to probe the efficiency of the selection under consideration. The efficiency of the selection criteria is defined as the number of probes that pass the selection with respect to the total number of probes:

$$\epsilon_{sel.} = \frac{N_P}{N_P + N_F}$$

where N_P is the number of the passing probes and N_F is the number of the failing probes. The probes are first split into several p_T and η bins defined in a way that ensures enough statistics inside every bin. The efficiency is then calculated for each bin separately.

One way to implement the efficiency measurement is to use the cut-and-count approach in which one simply counts the number of probes passing the selection and the number of probes that fail the selection. The efficiency is then easily calculated from the expression above. This can be a good approach when one is certain that there is no background contamination. Since this is the case in the simulation, the cut-and-count approach is used as the nominal method to estimate the efficiency in the simulation.
However, this is, in general, not the case in the data since very loose requirements are imposed on the probe. Therefore, another technique is used as the nominal signal efficiency measurement approach in the data. In this approach, both passing and failing probes are fitted, for each bin separately, using either an analytical function or the template extracted from the simulation. The nominal signal model is based on the Drell-Yan simulation used to obtain the template which is convoluted with a Gaussian distribution to account for the differences in resolution between the simulation and the data.

If no kinematic restrictions would be imposed on the tag and probe pairs, the dilepton mass distribution away from the resonance would be described nicely by a falling exponential function. However, cuts imposed on kinematic variables distort the invariant mass, m_{ee}, distribution in every bin in a way that is accounted for by using an error function. Thus, the background is described by a falling exponential function multiplied with an error function:

$$f(m_e e) = \mathrm{erf}(a - m_{ee}) \cdot e^{-d \cdot (m_{ee} - c)}$$

where a and c (b and d) are expressed in units of GeV (GeV^{-1}) and are free parameters in the fit.

The uncertainty on each efficiency measurement is obtained from the quadratic sum of the statistical uncertainty obtained from the fit and systematic uncertainty. The leading source of systematic uncertainty is the modelling of the signal and background contributions. The uncertainty in the signal model is obtained by replacing the template fit with a Breit-Wigner function convoluted with a one-sided Crystal ball (OSCB) function, while the uncertainty in the background model is obtained by using a falling exponential function instead of the product of a falling exponential function and an error function. For some low-p_T bins, a Chebyshev polynomial multiplied by a Gaussian (nominal signal), a Gaussian convoluted by a CB function (alternative signal), or a Gaussian multiplied by an exponential function (alternative background) was used in order to obtain a better fit.

The number of passing and failing probes in each bin is defined by the area between the signal and background functions. Examples of the nominal signal fits in the data are shown on the top of Fig. 3.5 for the (passing probe, failing probe) distributions for two different $(p_T,\ \eta)$ bins. The alternative signal fits in the simulation are shown

at the bottom of the figure for the (passing probe, failing probe) distributions in the same $(p_T,\ \eta)$ bins. The fitted signal contributions are shown in red, while the fitted background contributions are shown in blue.

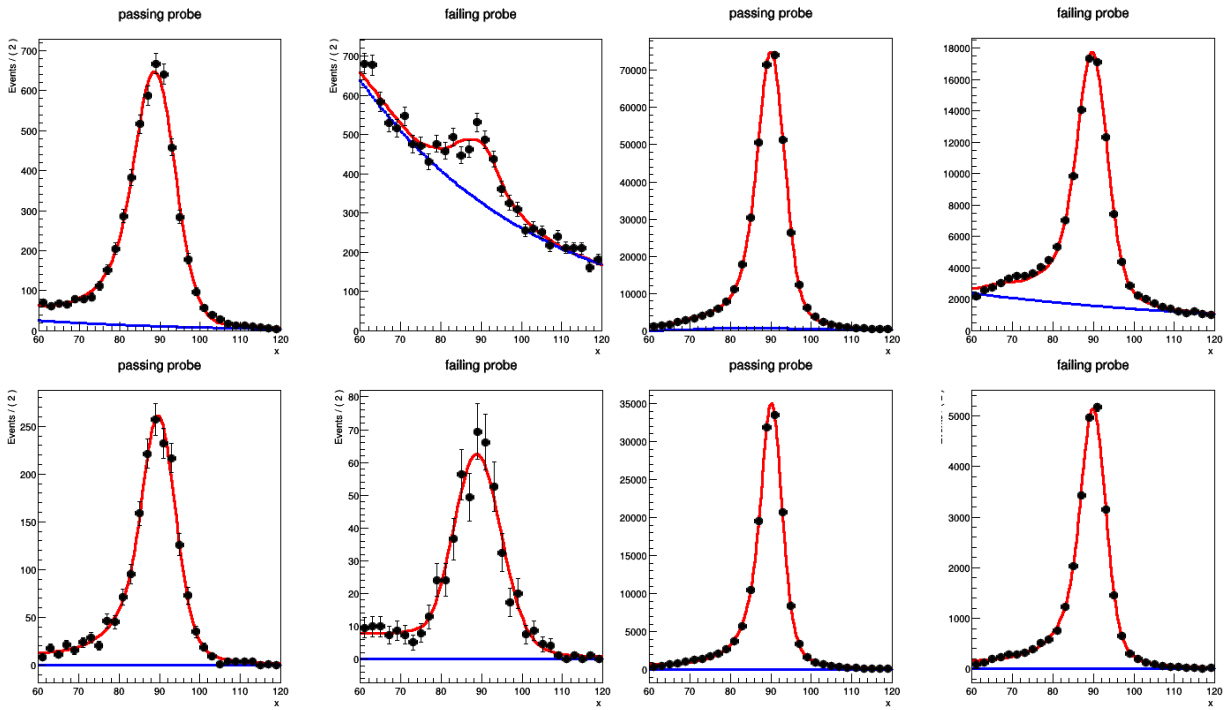


Figure 3.5: Example of the nominal signal fits in the data are shown on the top of the figure for the (passing probe, failing probe) distributions for two different $(p_T,\ \eta)$ bins. The alternative signal fits in the simulation are shown at the bottom of the figure for the (passing probe, failing probe) distributions in the same $(p_T,\ \eta)$ bins. The left-hand side plots show the (passing probe, failing probe) distributions in the $(2.00 < |\eta| < 2.5,\ 7\ GeV < p_T < 11\ GeV)$ bin and the right-hand side plots show the same in the $(2.00 < |\eta| < 2.5,\ 20\ GeV < p_T < 35\ GeV)$ bin. The fitted signal contributions are shown in red, while the fitted background contributions are shown in blue.

The efficiency measurements in each bin for the data and the simulation are used to derive *scale factors* (SFs) which are defined as the per-bin ratio of the efficiency under study obtained in the data divided by the efficiency in the simulation:

$$SF(p_T,\ \eta) = \frac{\epsilon_{data}(p_T,\ \eta)}{\epsilon_{MC}(p_T,\ \eta)}$$

These are used to scale the simulations to account for the different efficiency between the data and the simulation and therefore mitigate any discrepancies between the two left from the imperfect modelling.

Finally, the overall electron efficiency can be expressed as the product of the trigger efficiency, reconstruction efficiency and selection efficiency. The discussion about the trigger is rather involved and is not needed to follow the study presented here. An interested reader can find the details on the trigger performance in [95, 96]. The reconstruction efficiency is also measured using the TnP technique where the tag is an electron coming from the decay of the Z boson, and the second leg of the TnP are the SCs used to measure the efficiency (probes). One then counts the number of SCs that are promoted to electron (passing probe) with respect to the total number of probes. The largest source of uncertainty in the reconstruction efficiency measurements comes from the association of the SCs to the track. Since every analysis in CMS uses reconstruction efficiencies, these are produced centrally by the CMS collaboration and provided to all analyses containing electrons in the final state.

3.4.2 Electron selection efficiency in 2016, 2017 and 2018

The selection efficiency was derived for each data-taking period separately. The working point (WP) for the electron ID was optimized for the 2016 data-taking period in a way that corresponds to around 98% signal efficiency. The WPs for the 2017 and 2018 IDs were adjusted to reproduce the same signal efficiency. For all three data-taking periods, the electron ID included the isolation variables in the training of the multivariate classifier.

A first contribution to the electron ID was the measurement of the electron selection efficiency for the 2018 data-taking period using the recently improved MVA-based electron ID. Prior to this, the MVA training for the electron ID was based on the Toolkit for MultiVariate Analysis (TMVA) tool [97] and did not include isolation variables.
The retrained ID was obtained using the eXtreme Gradient Boosting (XGBoost) package [98] with the isolation variables included in the training. While the performance of the retrained ID was already demonstrated for the 2017 data-taking period [99], this was not yet done for the 2016 and 2018 periods. The efficiency of the retrained ID for the 2018 period was prepared and presented, for the first time, for the 2019 Moriond conference. An improvement for the 2017 data-taking period was presented at the conference as well. The efficiencies of the retrained electron IDs for the 2017 and 2018 periods are first discussed in this section.

Table 3.2 shows the list of data and MC samples used for both the 2017 and 2018 periods. The nominal MC efficiencies for both periods are evaluated from the leading order (LO) MadGraph [100] Drell-Yan sample, corresponding to a generic $q\bar{q} \to Z/\gamma^* \to e^+e^-$ production, while the next-to-leading order (NLO) MadGraph_AMCatNLO sample is used to assess the systematic uncertainty related to the generator being used.
For both the 2017 and 2018 periods the same requirements on the tag are imposed:

- trigger matched to HLT_Ele32_WPTight_Gsf_L1DoubleEG_v*
- $p_T^{tag} > 30\ GeV$, $|\eta_{SC}^{tag}| < 2.17$ and $q^{tag} \cdot q^{probe} < 0$

The first bullet ensures the geometrical matching of the tag to the leg of a single electron HLT object, ensuring that probes do not have any trigger selection cuts. Otherwise, the measurement of ID efficiency would be biased. The second bullet defines the p_T and η cut on the tag and requires an opposite-sign electron pair. Since the single electron trigger is restricted to $|\eta_{SC}| < 2.17$ because of the high background rates in the forward region of the detector, the same cut is imposed on the tag selection.

2017	
data	
/SingleElectron/Run2017B-17Nov2017-v1/MINIAOD	
/SingleElectron/Run2017C-17Nov2017-v1/MINIAOD	
/SingleElectron/Run2017D-17Nov2017-v1/MINIAOD	
/SingleElectron/Run2017E-17Nov2017-v1/MINIAOD	
/SingleElectron/Run2017F-17Nov2017-v1/MINIAOD	
MC	
sample	**usage**
/DYJetsToLL_M-50_TuneCP5_13TeV-madgraphMLM-pythia8/RunIIFall17MiniAOD-RECOSIMstep_94X_mc2017_realistic_v10-v1/MINIAODSIM	nominal
/DYJetsToLL_M-50_TuneCP5_13TeV-madgraphMLM-pythia8/RunIIFall17MiniAOD-RECOSIMstep_94X_mc2017_realistic_v10_ext1-v1/MINIAODSIM)	nominal
/DYJetsToLL_M-50_TuneCP5_13TeV-amcatnloFXFX-pythia8/RunIIFall17-MiniAODv2-PU2017_12Apr2018_94X_mc2017_realistic_v14-v1/MINIAODSIM	systematics
2018	
data	
/EGamma/Run2018A-17Sep2018-v2/MINIAOD	
/EGamma/Run2018B-17Sep2018-v2/MINIAOD	
/EGamma/Run2018C-17Sep2018-v2/MINIAOD	
/EGamma/Run2018D-17Sep2018-v2/MINIAOD	
MC	
sample	**usage**
/DYJetsToLL_M-50_TuneCP5_13TeV-madgraphMLM-pythia8/RunII-Autumn18MiniAOD-102X_upgrade2018_realistic_v15-v1/MINIAODSIM	nominal
DYJetsToLL_M-50_TuneCP5_13TeV-amcatnloFXFX-pythia8/RunII-Spring18MiniAOD-100X_upgrade2018_realistic_v10-v1/MINIAODSIM	systematics

Table 3.2: Data and MC samples used for the measurement of the electron selection efficiency for the 2017 and 2018 data-taking periods.

For the low p_T bins of the probe ($< 20\ GeV$), additional requirements were imposed in order to reject electrons coming from the W boson decays:

$$\text{Tighter tag ID } (MVA_{tag} > 0.92)$$

$$m_T = \sqrt{2 \cdot E_T^{miss} \cdot p_T^{tag} \cdot \Big[1 - cos(\phi_{E_T^{miss}} - \phi_{tag})\Big]} < 45\ GeV$$

For both periods, the selection under study is defined by the $H \to ZZ \to 4l$ MVA-based ID (mvaEleID-Fall17-iso-V2-wpHZZ) and the requirements on the vertex parameters and SIP as defined in section 3.3.1. Since electrons that end up in the region between the barrel and the endcap (henceforth referred to as the *gap electrons*) are expected to

be reconstructed with lower efficiency, they are treated separately in the efficiency measurements. Therefore, the selection efficiency and SFs are first derived for the non-gap electrons followed by the same analysis for the gap electrons only. The same selection on the tag and probe pairs is imposed in both cases.

Figure 3.6 shows the measured selection efficiencies (top pad in the figure) and SFs (bottom pad in the figure) in the different p_T bins for the two periods. The binning in η was chosen to be the same as the one used in the 2017 results already approved by CMS prior to this analysis. Gap electrons are excluded.

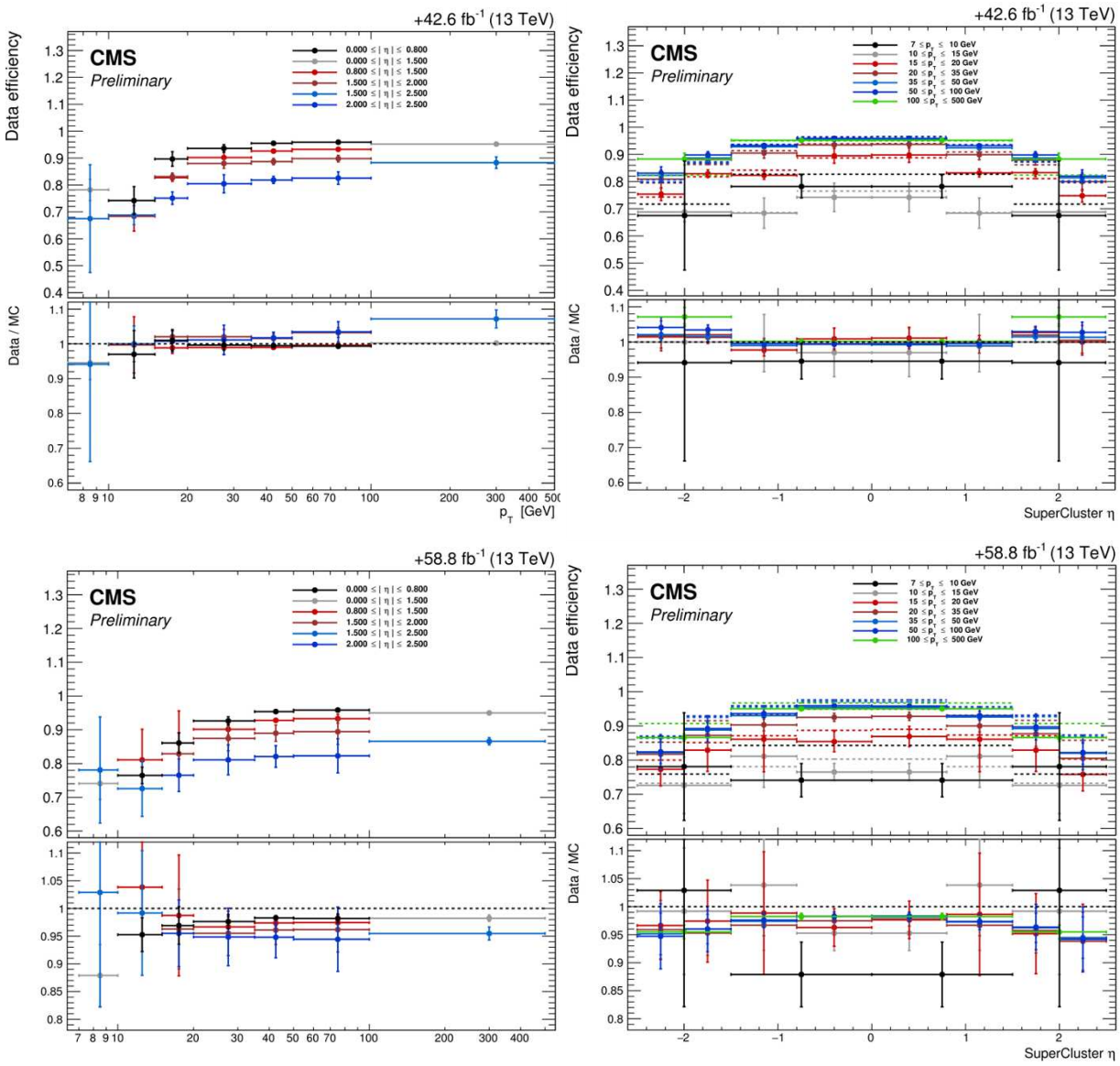


Figure 3.6: Electron selection efficiencies (top pad in the figures) and SFs (bottom pad in the figures) for the 2017 (top row) and 2018 (bottom row) data-taking periods. The left-hand side plots show the results for different p_T bins, while the right-hand side plots show the same for different η bins.

Due to lower statistics in the very low-p_T bin ($< 10\ GeV$) and high-p_T bin ($> 100\ GeV$), the efficiencies and SFs were calculated only for the combined barrel (light grey histogram) and the endcap (light blue histogram) region. The middle-p_T range is split into several η bins in order to gain insight into the possible η-dependent structure of SFs.

One feature that can be seen on the efficiency plots versus the electron p_T is the increase of the efficiency in the low-p_T region until the plateau is reached. This is the consequence of the bremsstrahlung which causes the loss of efficiency at low values of p_T.
An additional feature, especially pronounced in the 2018 period, is a consistent offset of SFs from unity over the entire p_T range. This was studied and traced back to the $|SIP| < 4$ cut. If the SIP cut requirement is removed from the selection, keeping other things unchanged, this feature disappears. This can be seen on Fig. 3.7 where the SFs are now consistent with unity. This behaviour was afterwards cured by the Ultra legacy (UL) reprocessing of the data and the MC samples.

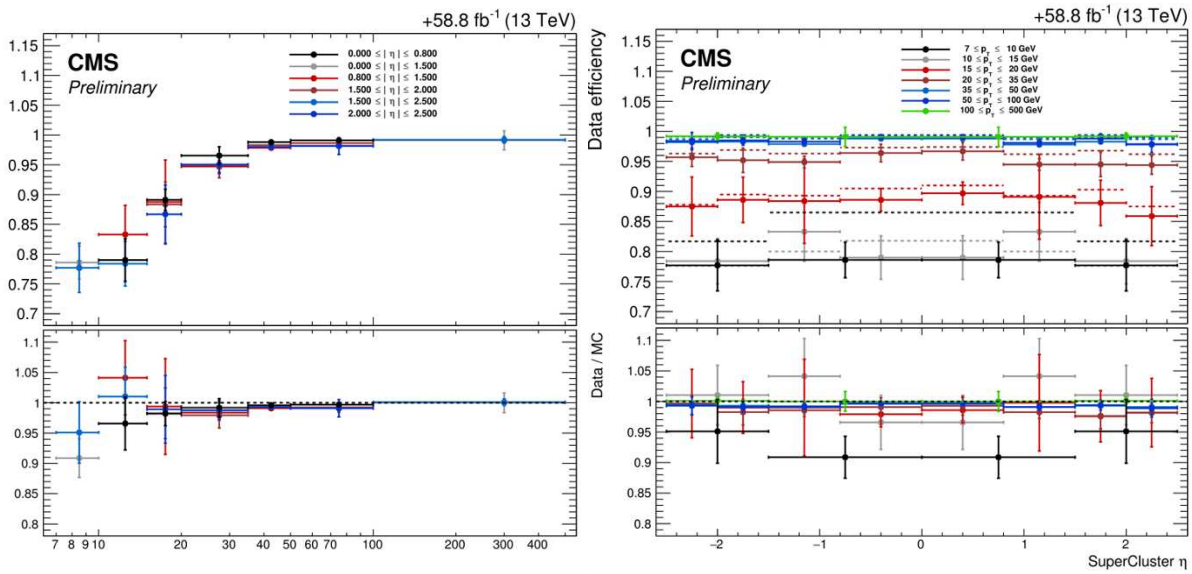


Figure 3.7: Electron selection efficiencies (top pad in the figures) and SFs (bottom pad in the figures) for the 2018 data-taking periods. The left-hand side plot shows the results for different p_T bins, while the right-hand side plot shows the same for different η bins. The only change with respect to the bottom row plots in Fig. 3.6 is the removal of the $|SIP| < 4$ cut.

Comparing the uncertainties obtained for the 2017 and 2018 periods, one can see that these are larger for the latter. This can be more easily seen on Fig. 3.8 showing the SFs and corresponding uncertainties in all p_T and η bins.

Fig. 3.9 and Fig. 3.10 show the selection efficiency, scale factors and corresponding uncertainty for the gap electrons in the 2018 data-taking period. The same plots for the 2017 period were obtained in CMS prior to this analysis and are thus omitted. Only three p_T bins were used in order to keep sufficient statistics in each bin. In addition, on the right-hand side plot, the bins are split in $|\eta|$, rather than in η.

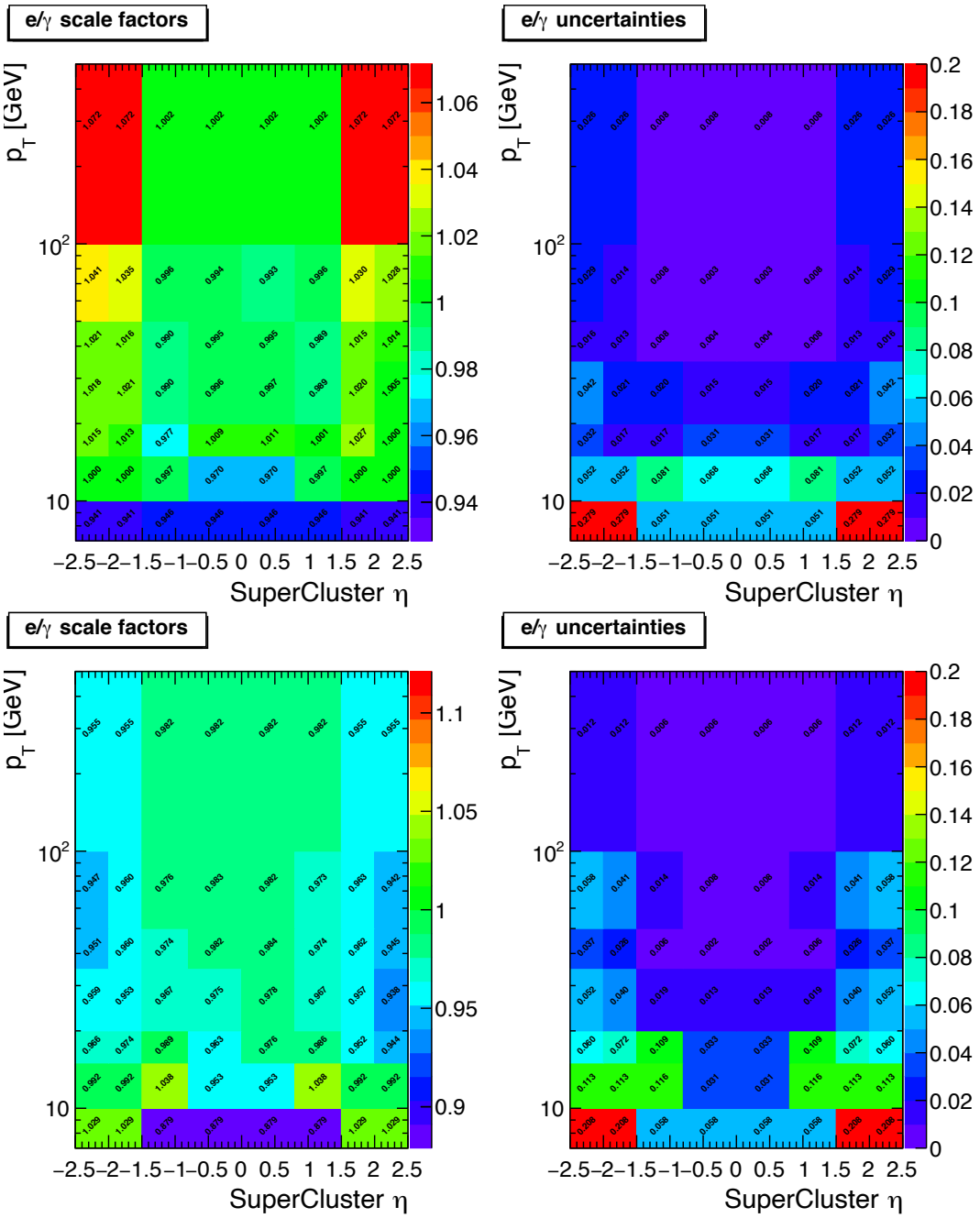


Figure 3.8: Electron SFs (left row) and corresponding overall uncertainty (right row) for all p_T and η bins shown in Fig. 3.6. Results for the 2017 (2018) period are shown in the top (bottom) row.

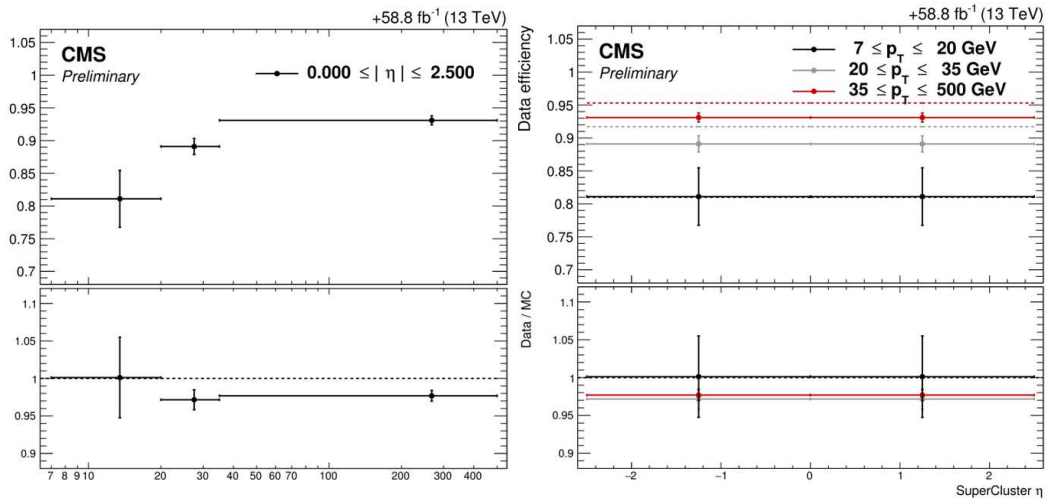


Figure 3.9: Electron selection efficiencies (top pad in the figures) and SFs (bottom pad in the figures) for the 2018 data-taking periods. The left-hand side plot shows the results for different p_T bins, while the right-hand side plot shows the same for different η bins. Only gap electrons are considered.

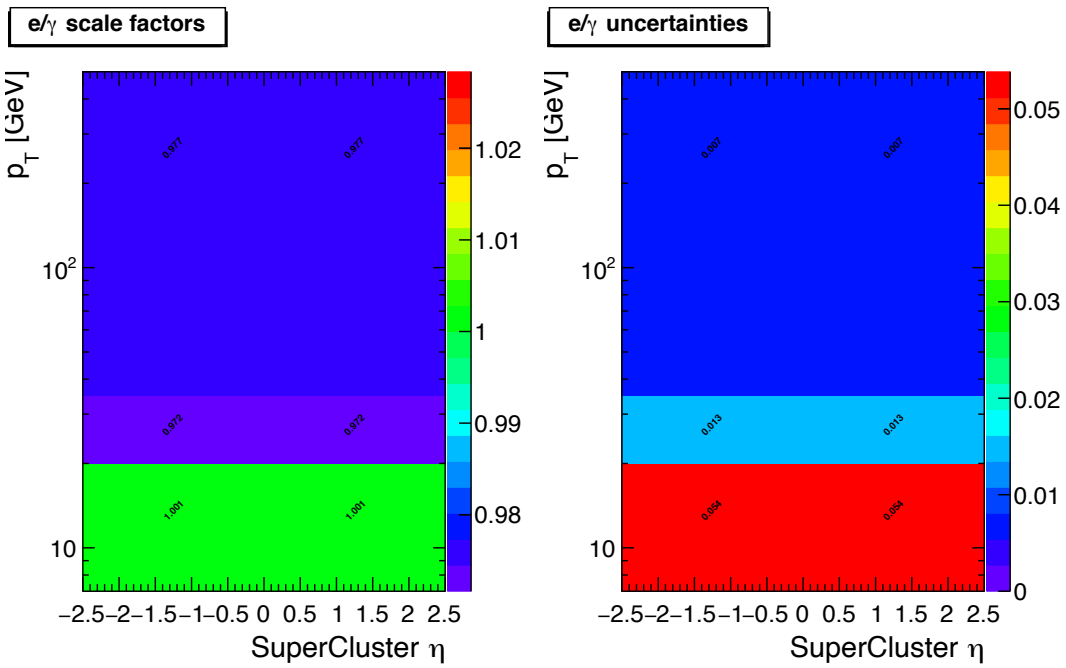


Figure 3.10: Electron SFs (left row) and corresponding overall uncertainty (right row) for all p_T and η bins shown in Fig. 3.9. Results for the 2018 period gap electrons are shown.

In order to prepare for the $H \to ZZ \to 4l$ Run 2 legacy paper [101], it was decided to retrain the electron ID for the 2016 period. This meant replacing the older ID which didn't incorporate an isolation variable in the training and which was trained using the TMVA package with the new ID (mvaEleID-Summer16-ID-ISO-HZZ) that included isolation in the training and was trained using the XGBoost package.
In the analysis discussed thus far in this section, the retrained ID used for the 2017 data-taking period was also used for the 2018 period. A dedicated ID retrained for the 2018 period was not essential at the time because it was shown that the performance of the 2017 training on the 2018 data was satisfactory. However, in the meantime, a dedicated ID was retrained also for the 2018 period by the CMS collaboration for consistency's sake. Since the training of the IDs is not a direct contribution to this thesis work, the WPs and corresponding signal and background efficiencies for all three Run 2 periods are merely summarized in Table 3.3.

The electron efficiency measurements and the SFs discussed in the following part of the section were rederived using the retrained electron IDs for 2016, 2017 and 2018 periods. The goal of this analysis was to further reduce the uncertainty in the low-p_T region and study the η-dependent structure of SFs. The former was especially needed since the leading source of uncertainty in the $H \to ZZ \to 4l$ analysis is the uncertainty on electron efficiency measurements that mostly originates from the measurement uncertainty of the low-p_T electrons that are present in the analysis due to the off-shell Z boson.
Data and simulations used in the analysis are listed in Table 3.4 for all three periods. For the 2016 period, the nominal MC efficiencies are evaluated from the leading order (LO) MadGraph Drell-Yan sample, while the next-to-leading order (NLO) MadGraph_AMCatNLO sample is used to access the systematic uncertainty. The only change in the 2018 period, with respect to the previously discussed analysis, is the use of the *POWHEG* [102–104] sample for accessing the systematic uncertainties instead of the (NLO) MadGraph_AMCatNLO sample. The reason for this change is the higher statistics in the POWHEG sample. As before, efficiency measurements for the non-gap electrons are shown first, followed by the measurements for the gap electrons.

2016 (mvaEleID-Summer16-ID-ISO-HZZ)			
$\|\eta\| < 0.8$			
	WP	ϵ_{sig} [%]	ϵ_{bkg} [%]
$5\ GeV < p_T < 10\ GeV$	1.8409	81.64	3.93
$p_T > 10\ GeV$	0.3902	97.44	2.17
$0.8 < \|\eta\| < 1.479$			
$5\ GeV < p_T < 10\ GeV$	1.7830	80.31	3.63
$p_T > 10\ GeV$	0.3484	96.68	2.75
$\|\eta\| > 1.479$			
$5\ GeV < p_T < 10\ GeV$	1.7559	74.37	3.06
$p_T > 10\ GeV$	-0.6518	96.62	7.66
2017 (mvaEleID-Fall17-iso-V2-wpHZZ)			
$\|\eta\| < 0.8$			
	WP	ϵ_{sig} [%]	ϵ_{bkg} [%]
$5\ GeV < p_T < 10\ GeV$	1.4499	81.64	5.66
$p_T > 10\ GeV$	0.0081	97.44	3.26
$0.8 < \|\eta\| < 1.479$			
$5\ GeV < p_T < 10\ GeV$	1.4856	80.31	4.74
$p_T > 10\ GeV$	-0.0374	96.68	4.05
$\|\eta\| > 1.479$			
$5\ GeV < p_T < 10\ GeV$	1.6901	74.37	3.59
$p_T > 10\ GeV$	-0.7497	96.62	8.10
2018 ((mvaElectronID_Autumn18_ID_ISO)			
$\|\eta\| < 0.8$			
	WP	ϵ_{sig} [%]	ϵ_{bkg} [%]
$5\ GeV < p_T < 10\ GeV$	0.8962	81.64	5.66
$p_T > 10\ GeV$	0.0279	97.45	3.28
$0.8 < \|\eta\| < 1.479$			
$5\ GeV < p_T < 10\ GeV$	0.9070	80.31	4.69
$p_T > 10\ GeV$	-0.0024	96.68	4.12
$\|\eta\| > 1.479$			
$5\ GeV < p_T < 10\ GeV$	0.9396	74.37	3.26
$p_T > 10\ GeV$	-0.5983	96.62	8.06

Table 3.3: Working points together with corresponding signal and background efficiencies for the BDT training of the electron ID for the three data-taking periods.

2016	
data	
/SingleElectron/Run2016B-17Jul2018_ver2-v1/MINIAOD	
/SingleElectron/Run2016C-17Jul2018-v1/MINIAOD	
/SingleElectron/Run2016D-17Jul2018-v1/MINIAOD	
/SingleElectron/Run2016E-17Jul2018-v1/MINIAOD	
/SingleElectron/Run2016F-17Jul2018-v1/MINIAOD	
/SingleElectron/Run2016G-17Jul2018-v1/MINIAOD	
/SingleElectron/Run2016H-17Jul2018-v1/MINIAOD	
MC	
sample	**usage**
/DYJetsToLL_M-50_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/RunII-Summer16-MiniAODv2-PUMoriond17_80X_mcRun2_asymptotic_2016_TrancheIV_v6_ext1-v2/MINIAODSIM	nominal sample
/DYJetsToLL_M-50_TuneCUETP8M1_13TeV-amcatnloFXFX-pythia8/RunIISummer16-MiniAODv2-PUMoriond17_80X_mcRun2_asymptotic_2016_TrancheIV_v6_ext2-v1/MINIAODSIM	systematics
2017	
data	
same as in Table 3.2	
MC	
same as in Table 3.2	
2018	
data	
same as in Table 3.2	
MC	
sample	**usage**
same as in Table 3.2	nominal sample
/DYToEE_M-50_NNPDF31_TuneCP5_13TeV-powheg-pythia8/RunIIAutumn18MiniAOD-102X_upgrade2018_realistic_v15-v1/MINIAODSIM	systematics

Table 3.4: Data and MC samples used for the measurement of the electron selection efficiency and SFs for the $H \to ZZ \to 4l$ Run 2 legacy paper.

The same selection on the tag is applied to three periods and is, for the most part, the same as defined before. In order to try to reduce the uncertainties, the p_T requirement on the tag was increased to 50 GeV for the lower p_T bins of the probe ($< 20\ GeV$). In addition, the requirement that all three charge measurements, defined in section 3.2.3, agree was required for the same bins. Finally, the coarser binning of the m_{ee} distribution, using 30 bins instead of 60, was used in order to further stabilize the fits.
The new requirements on the tag resulted in a slightly more clear peak around the nominal Z boson mass which

resulted in better precision and lower uncertainty for these bins. This can be seen in the right column in Fig. 3.11 which shows the nominal signal fit in data for one low-p_T bin ($11\ GeV < p_T < 15\ GeV$ and $0 < \eta < 0.5$). It was checked that no bias is introduced in the efficiency measurement by doing so.

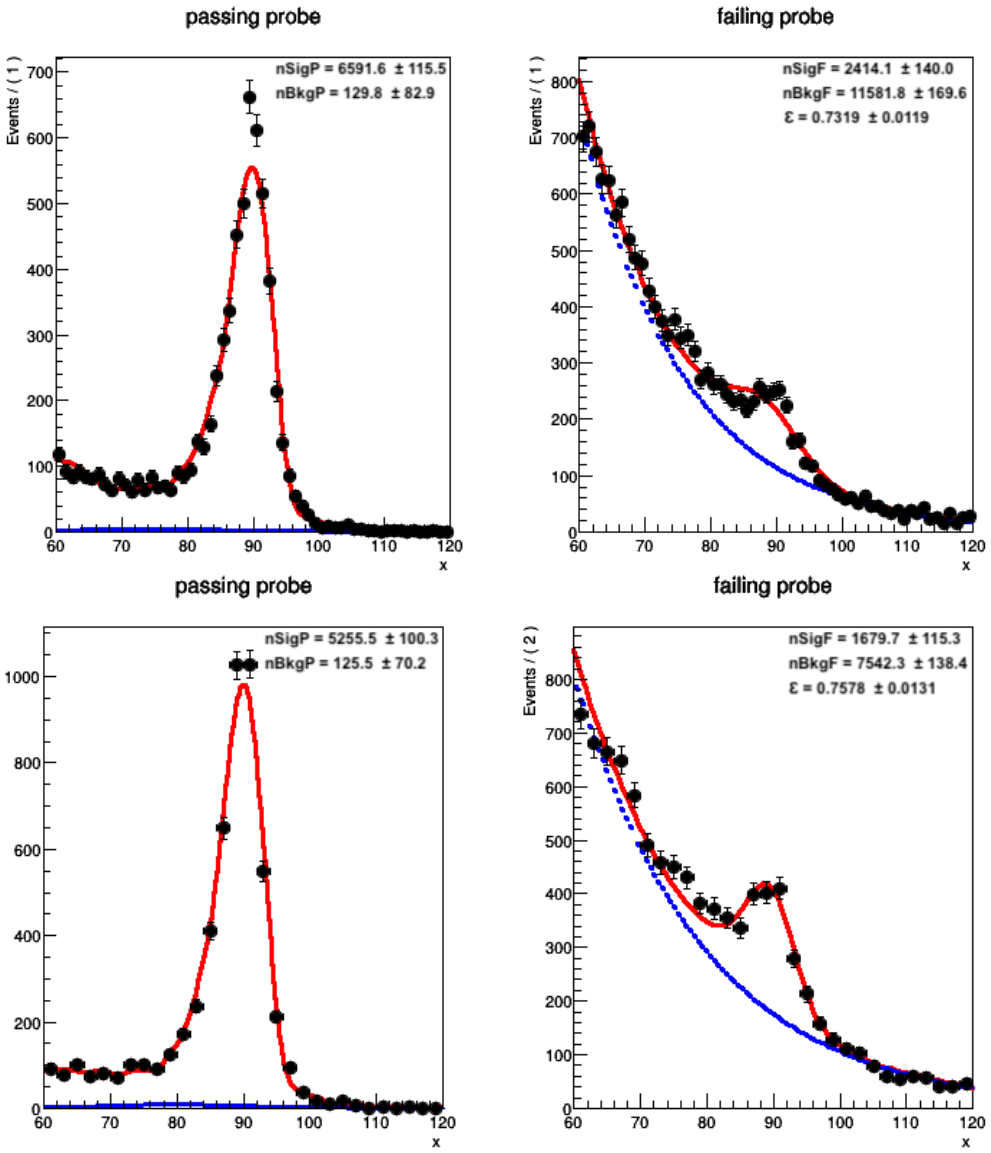


Figure 3.11: The m_{ee} distribution for one low-p_T bin ($11\ GeV < p_T < 15\ GeV$ and $0 < \eta < 0.5$) before (top row) and after (middle row) tightening the tag selection for the low-p_T bins of probe. The nominal fit in the data is shown in both figures.

Another consequence of the tighter tag selection was the appearance of the excess of events (the "bump") in the low mass tail of the m_{ee} distribution of the failing probes for $15 < p_T < 20$ GeV bins. This bump comes from the signal electrons that migrated from the passing probe group before tightening the cut to the failing probe group after tightening the cut. In order to successfully fit the bump, the function for the signal model in the failing probes had to be modified. It was found that a good fit for the signal can be achieved with a help of additional Gaussian. To achieve the convergence of the fit for the background, a default model was modified by introducing a Chebyshev polynomial. This is shown in Fig. 3.12 for the nominal fit in data for one bin ($15\ GeV < p_T < 20\ GeV$ and $1 < \eta < -0.5$). The left-hand side plot shows the bad fit in the failing probes before the modification of the fitting function, while the right-hand side

plot shows the improved fit.

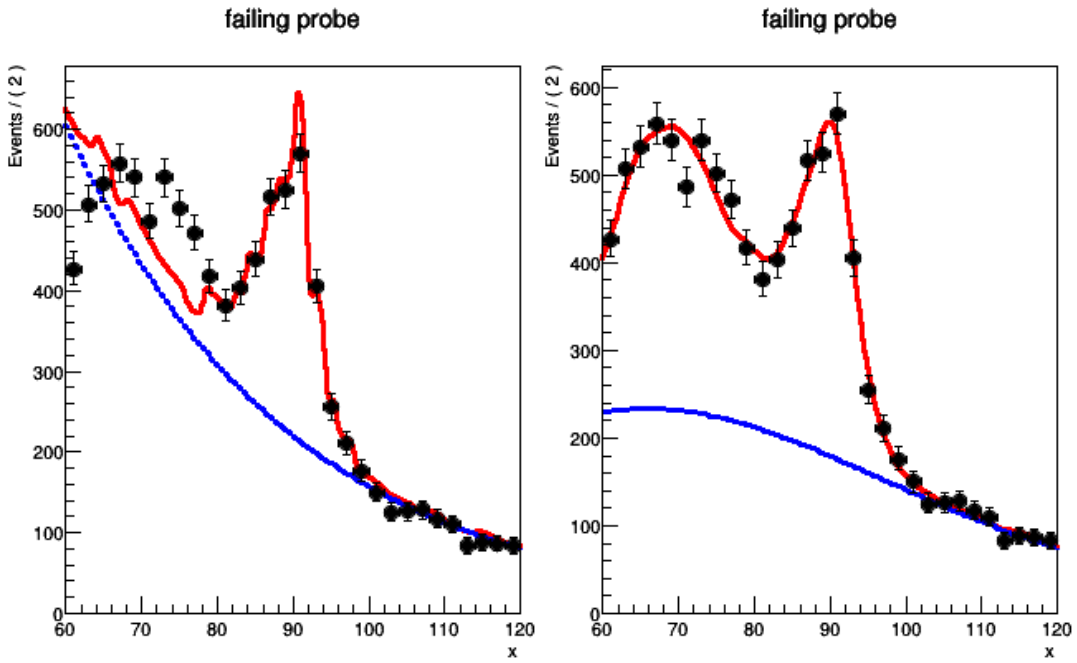


Figure 3.12: The m_{ee} distribution for one bin ($15\ GeV < p_T < 20\ GeV$ and $1 < \eta < -0.5$) after tightening the tag selection for the low-p_T bins of probe. The bad fit for the failing probes (left) was resolved by adding an additional Gaussian in the signal model and introducing a Chebyshev polynomial in the background model (right).

Fig. 3.13 shows that this treatment reduced uncertainties in the selection efficiency measurement, especially for the low-p_T and high-η region. Here, the gap electrons are not included.

In addition to the tighter requirements on the tag, one can see that the binning has been changed in order to try improving on the η dependency of the SFs. This feature is visible on the bottom-right plot in Fig. 3.13.
While studying different binning scenarios for the 2017 data-taking period, it was found that better results can be achieved by using a finer η binning. This is shown in the top row in Fig. 3.14 where a more pronounced η structure in SFs is observed. The "umbrella" shape in efficiency (top pad on the figure) is the result of inefficiencies in electron reconstruction and identification in the more forward regions of the detector. The top row shows the results for the 2017 period, while the bottom row shows the results for the 2016 period. Fig. 3.15 shows the SFs and the corresponding uncertainty for the three data-taking periods.

Finally, Figs. 3.16 and 3.17 show the efficiency, SFs and the overall uncertainty for the gap electrons for the 2016, 2017 and 2018 data-taking period.

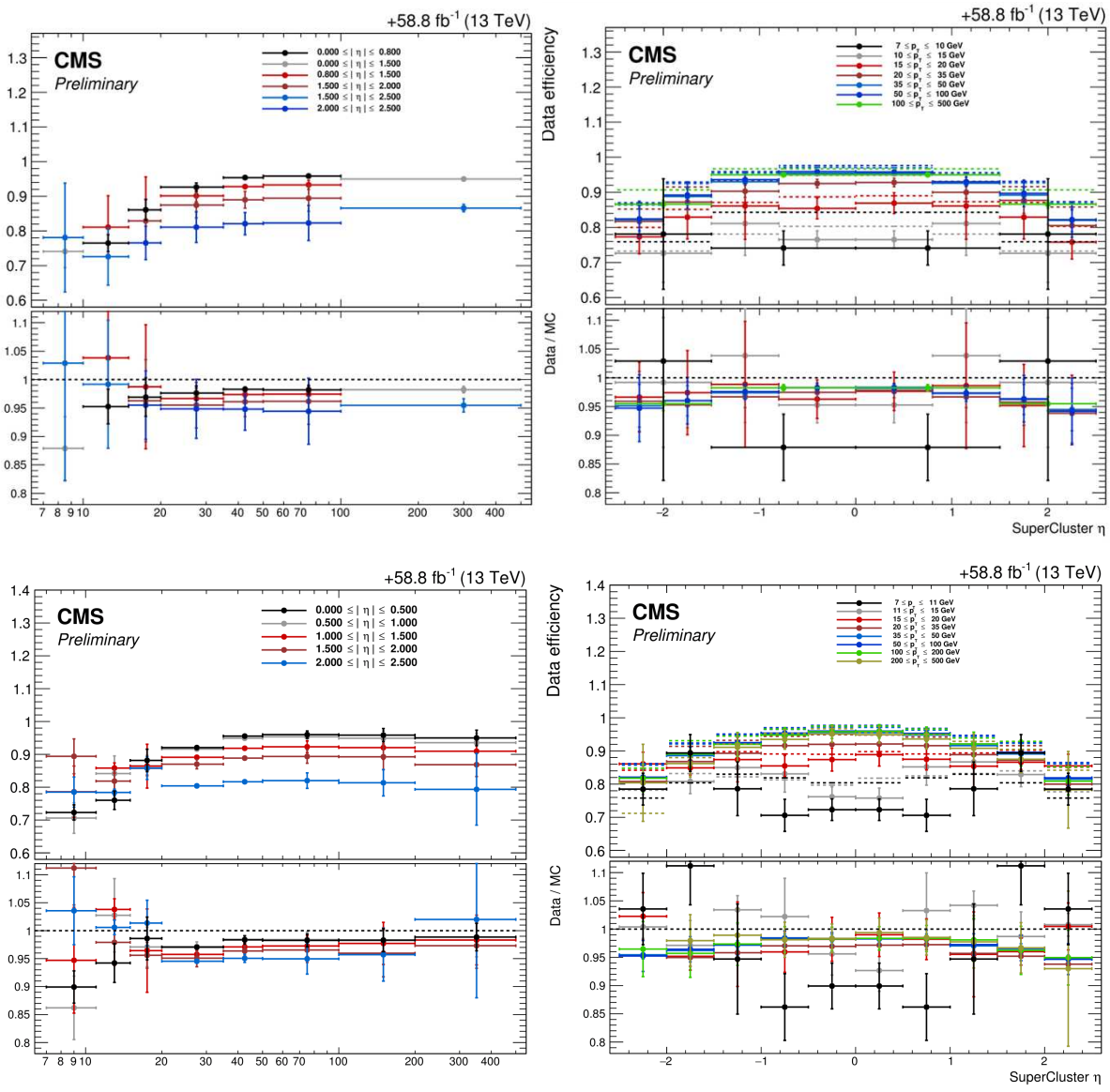


Figure 3.13: Electron selection efficiencies (top pad in the figures) and SFs (bottom pad in the figures) for the 2018 period with the original tag selection (top row) and the same period with tighter tag selection introduced for the low-p_T bins of the probe (bottom row). The left-hand side plots show the results for different p_T bins, while the right-hand side plots show the same for different η bins.

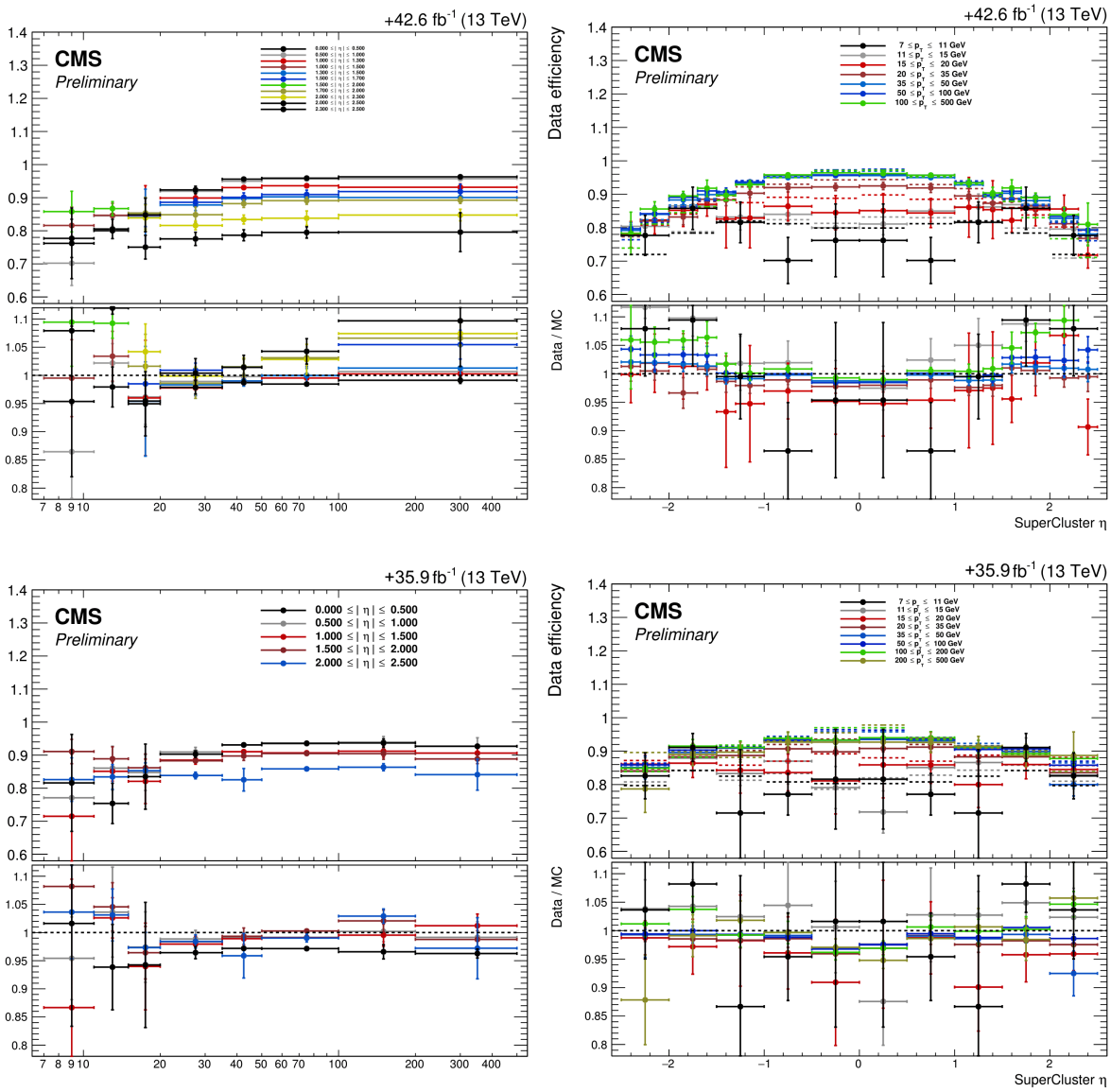


Figure 3.14: Electron selection efficiencies (top pad in the figures) and SFs (bottom pad in the figures) for the 2017 (top row) and 2016 (bottom row) periods using the retrained electron ID and the tighter tag selection for the low p_T bins of the probe. The left-hand side plots show the results for different p_T bins, while the right-hand side plots show the same for different η bins.

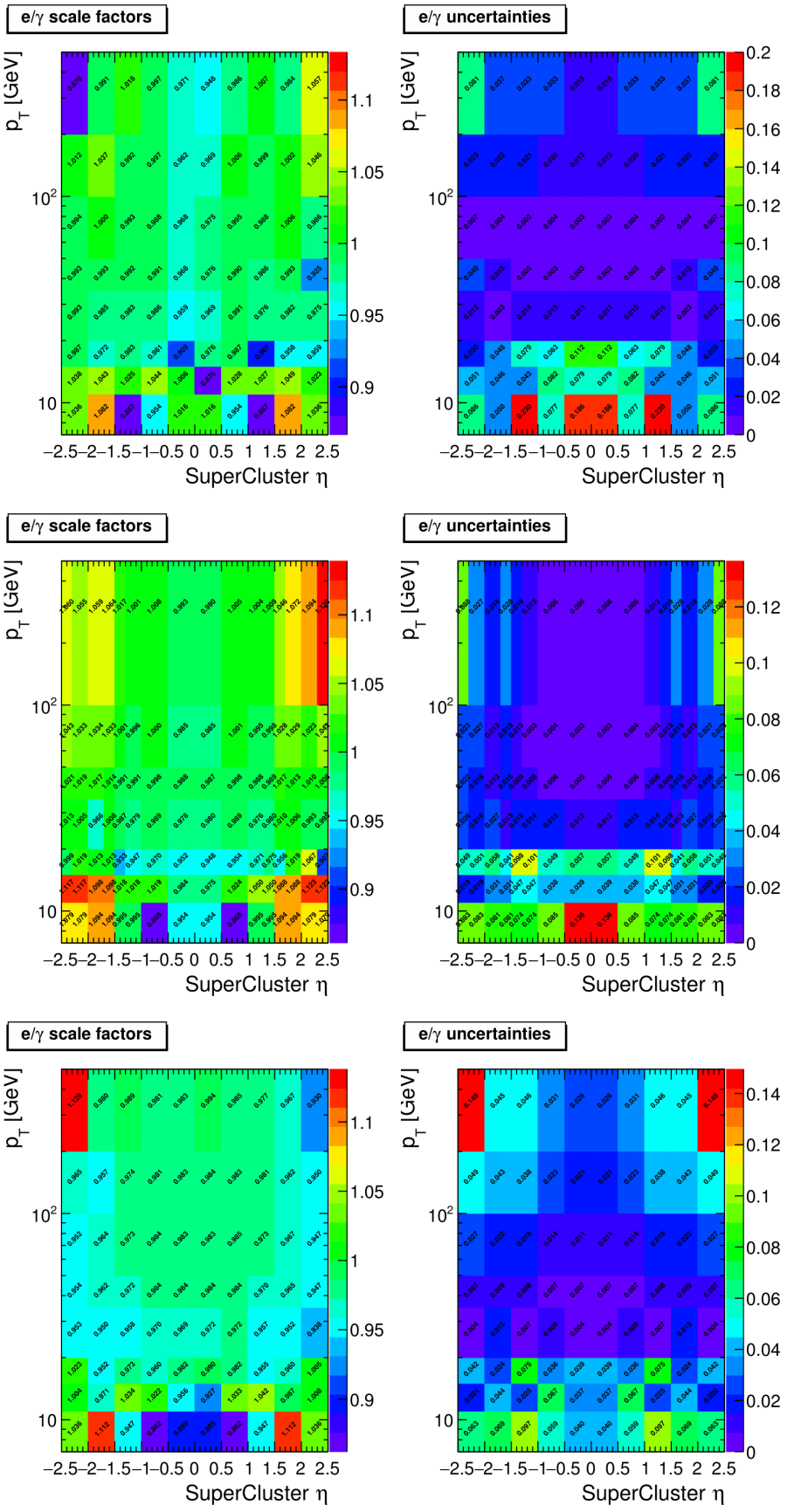


Figure 3.15: Electron SFs (left column) and corresponding overall uncertainties (right column) for all p_T and η bins shown in the bottom row in Fig. 3.13 and in Fig. 3.14. Results for 2016, 2017 and 2018 periods are shown in the top, middle and bottom rows respectively.

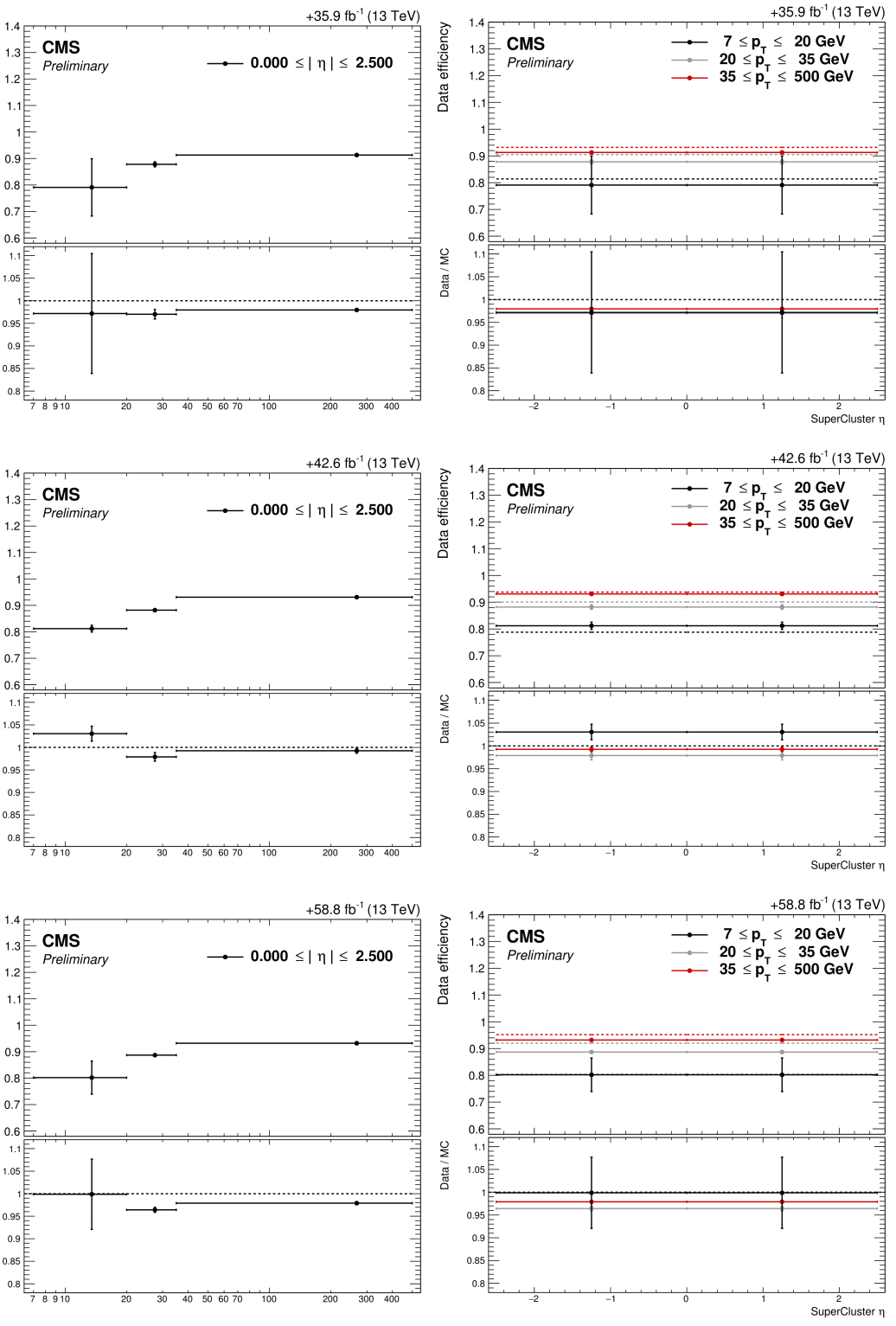


Figure 3.16: Electron selection efficiencies (top pad in the figures) and SFs (bottom pad in the figures) for the 2016 (top row), 2017 (middle row) and 2018 (bottom row) periods using the retrained electron IDs and the tighter tag selection for the low p_T bins of the probe.. The left-hand side plots show the results for different p_T bins, while the right-hand side plots show the same for different η bins. Results for gap electrons are shown.

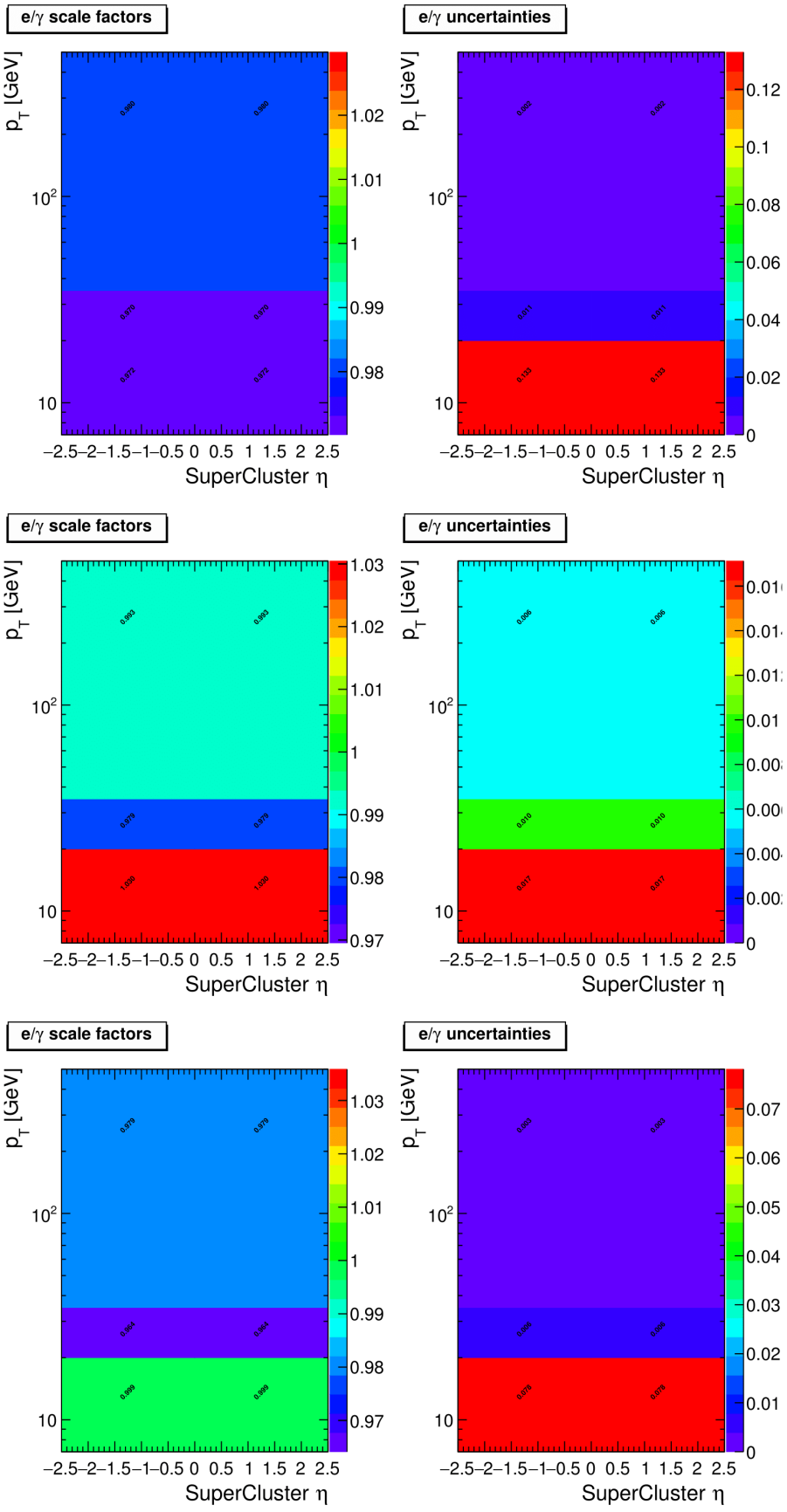


Figure 3.17: Electron SFs (left column) and corresponding overall uncertainties (right column) for all p_T and η bins shown in Fig. 3.16 for the 2016 (top row), 2017 (middle row) and 2018 (bottom row) period. Results for gap electrons are shown.

3.5 Summary

An overview of electron reconstruction and identification (ID) in CMS was discussed followed by the measurements of the electron selection efficiency for the 2016, 2017 and 2018 data-taking periods using the Tag and Probe (TnP) method.
The efficiency measurements and the scale factors (SFs) were first derived for the 2018 period using the electron ID trained on the 2017 data. The training of the ID was done centrally in CMS, the novelty being the incorporation of the isolation variables in the training of the multivariate classifier. The new ID training was improved by switching to the XGBoost package instead of the TMVA.
In order to prepare for the $H \rightarrow ZZ \rightarrow 4l$ Run 2 legacy paper, it was decided to retrain the electron ID for the 2016 data-taking period using the XGBoost package and include the isolation variables in the training. The same was done for the 2018 period which improved the performance of the ID. The retraining of the IDs for both data-taking periods was done centrally in CMS. New electron efficiency measurements, together with the SFs, were rederived for all three periods with the goal of reducing the uncertainties in the low-p_T region and studying the η structure in SFs.

In order to reduce the uncertainties in the low-p_T region, a tighter selection on the tag was applied for the low-p_T bins of the probe. The requirement that all three charge measurements must agree was also added. The new requirements on the tag gave rise to a slightly more clear peak around the nominal Z boson mass. This resulted in better precision and lower uncertainty for these bins and was first shown for the 2018 period. The same conclusion was found to be true for the 2016 and 2017 periods as well. An additional consequence of the tighter tag selection was the appearance of the "bump" in the low mass tail of the m_{ee} distribution of the failing probes, in particular for the $15 < p_T < 20\ GeV$ bins. It was found that the bump was populated by signal electrons that migrated to the failing probe group after tightening the tag selection. In order to fit the bump in the m_{ee} distribution for the failing probes, the fitting function for the signal and background contributions had to be modified. A better fit further reduced the uncertainty in these bins.
In addition to tightening the tag selection, the binning for the 2018 period was changed in order to try to improve on the η dependency of the SFs. This was more studied for the 2017 period where it was shown that a further improvement can be achieved by choosing even finer η binning. The expected "umbrella" shape in the efficiencies due to a more challenging reconstruction and identification of electrons in the forward regions of the detector was observed with better accuracy.

The results presented in this chapter were used in the publication of the $H \rightarrow ZZ \rightarrow 4l$ analysis with Run 2 data and are used as a default recipe for the selection of electrons with p_T bellow 10 GeV. In addition, these are an integral part of the VBS $ZZ \rightarrow 4l2j$ analysis discussed in the next chapter.

Chapter 4

Search for the VBS in the 4l final state using Run 2 data

4.1 Preface to the chapter

This chapter covers published results on the search for the VBS in the $ZZ \to 4l2j$ channel using full Run 2 data and is a continuation of a previous study within the CMS diboson group in the same channel that used 2016 data to extract the EWK signal [105].
The biggest challenge of the analysis is a small signal cross section, being one of the smallest ever measured at the LHC with only $0.3\ fb$. Another feature of this channel is a large contribution from the QCD-induced production of two jets and two Z bosons.
However, unlike final states containing W bosons, this channel is characterized by a fully reconstructable final state. Because of this, it is expected to be amongst the most important channels to separate the longitudinal polarization of the Z boson in the future. In addition, it is the most sensitive channel for studying certain anomalous quartic gauge couplings (aQGCs), specifically f_{T8} and f_{T_9}. Lastly, it had not yet been observed in CMS.

My main contribution to the published paper, apart from SFs discussed in the previous chapter, is the development of the BDT classifier used as an alternative signal extraction method. This is described in section 4.6. In addition, I derived the limits on the anomalous quartic gauge couplings in the EFT approach. The procedure for deriving aQGCs is presented in section 4.7.

I begin the chapter by describing the data sets and Monte Carlo simulations used in the analysis. The following section defines the event selection. In section 4.4 I define the variables used for the signal extraction with the BDT and to check the agreement between the data and the simulation.
In the published paper the MELA discriminant was used as the main tool for signal extraction and is discussed in section 4.5.1. In the same section, I describe how the VBS significance and the cross section in VBS and VBS+QCD fiducial regions were calculated. Section 4.8 will discuss the systematic uncertainties used in both the MELA and the BDT signal extraction approaches and also in the derivation of the limits on the aQGCs.
In section 4.9 I present results on the VBS signal significance using both the MELA and the BDT approaches and compare the two. The results on the aQGCs are reported here as well. The key points of the chapter are summarized in section 4.10.

4.2 Monte Carlo simulations and data sets

4.2.1 Monte Carlo samples

Several Monte Carlo (MC) samples have been produced and are used in this analysis to optimize the event selection, evaluate signal efficiency and acceptance, optimize the search strategy for the VBS as well as for a search for anomalous quartic gauge couplings (aQGCs).

Signal

In this analysis, the signal is defined as the purely electroweak (EWK) production of the two jets and the two leptonically decaying Z bosons. It was simulated at leading order (LO) using the *MadGraph5_aMCatNLO* (henceforth *MG5*) tool [106] by requiring explicitly the number of QCD vertices to be zero:

$$generate\ pp > zzjj\ QCD = 0,\ z > l + l-$$

Z bosons are only allowed to decay into electrons and muons. This is performed using the *MadSpin* tool in order to preserve the spin correlations between the leptons. The resulting sample includes contributions from the SM Higgs boson produced in vector boson fusion (VBF) as well as from the interference with non-Higgs diagrams and diagrams featuring triboson production with one hadronically decaying W boson. The latter is suppressed by requiring the dijet invariant mass, m_{jj} to be greater than 100 GeV.
An additional sample was produced using the *Phantom* tool [107] which includes off-shell Z boson decays and was used to cross-check the sample produced with *MG5*.

Irreducible backgrounds

The dominant, irreducible background in the analysis is the QCD-induced $pp \rightarrow ZZ$ production (henceforth qqZZ). This process was simulated at next-to-leading order (NLO) with up to two jets using MG5 and merged with parton showers using the FxFx scheme.

$$generate\ pp > l + l - l + l -\ [QCD]\ @0$$

$$add\ process\ pp > l + l - l + l - j\ [QCD]\ @1$$

The idea behind the FxFx jet merging scheme is to remove the overlap between jets produced at matrix elements (ME) and those produced by parton showers (PS) and thus removing the double counting of jets [108, 109]. This is the nominal sample for the qqZZ background in this analysis.
In order to study the interference between the signal and the irreducible background, an additional sample was generated using MG5 [110]. It was shown, by comparing the event yields and distribution shapes between the signal sample and the interference sample, that the yield ratio is between 1% and 6%. This was taken into account in the analysis via proper scaling.

Additional background to the signal is the gluon loop-induced ZZ production process (henceforth ggZZ). Although this process is suppressed by two additional strong couplings, it nevertheless contributes to $ZZ + 2j$ production at around 10% level.
A dedicated sample was studied and produced with *MG5* [111] specially for this analysis [112]. The process is

simulated at LO with up to 2 jets modelled from matrix-element and matched to PS using the MLM matching scheme [113] for the first time:

$$generate\ gg > zz\ [noborn = QCD]$$

$$add\ process\ pp > zzj\ [noborn = QCD]$$

$$add\ process\ pp > zzjj\ [noborn = QCD]$$

The requirement in the square brackets instructs MG5 to only consider loop diagrams. One can see that, for the 0j sample, a gg initial state was used, while qq initial state was used for the $1j$ and $2j$ samples. For the gg case, this is equivalent since there are no extra loop-induced diagrams from qq or gg. However, it is important to use pp initial state for $1j$ and $2j$ samples in order to include the Initial State Radiation (ISR) processes. In this case, a quark will first transform to a gluon through the ISR, after which it will be involved in the hard process. This results in significantly more diagrams from which only genuine loop-induced diagrams should be kept. This is achieved using a "diagram filter" designed especially for this purpose [114]. The filter requires that the loop does not contain any gluon line, such that the vertex- and box-correction diagrams are discarded. Additionally, the loop must attach to, at least, one Z boson, W boson, or photon to avoid diagrams concerning the gluon self-energy correction through quark lines and diagrams mediated by the Higgs boson. After the filter is applied, only genuine loop-induced diagrams survive. It must be emphasized that $1j$ and $2j$ diagrams are not simply a $0j$ diagram with some ISR decoration. These processes include some new diagrams with different structures that can't evolve from the $0j$ sample. Some examples are shown in Figure 4.1 where jets are emitted directly from the loop.

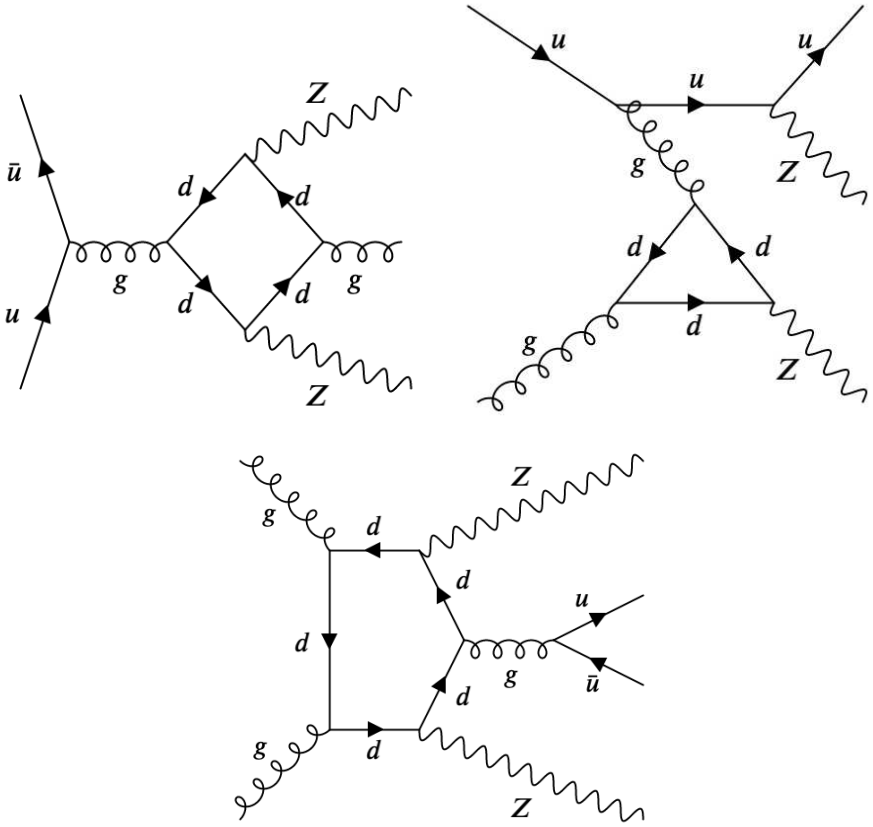


Figure 4.1: Example diagrams of loop induced ggZZ 1/2-jet process which can't evolve from the 0j sample

Although time-consuming, this simulation is expected to describe this background much better than the $0j$ sample where two jets are modelled from the PSs. However, since *MadSpin* generator can't decay particles generated in the loop-induced processes from the ME calculation, the decaying of the Z bosons is implemented in *Pythia8* such that spin correlations between the outgoing leptons are not included. The MLM matching scheme was applied to avoid double counting when merging jets modelled with ME and those modelled by parton shower.

The dijet phase-space produced from the loop-induced process is expected to be more accurately modelled with this sample compared to an alternative approach using the MCFM generator [115]. The difference between the new *MG5* ggZZ production and the *MCFM* production is especially visible in the p_T spectrum of the two leading jets. This is shown in Figure 4.2 for different jet multiplicities. The difference is most notable in the softer p_T spectrum for the 0,1,2 jet merged sample (purple) produced with *MG5*. This leads to a lower efficiency after applying the inclusive ZZjj selection (for details on event selection criteria see section 4.3). An additional effect is the higher mass of the ZZ pair.

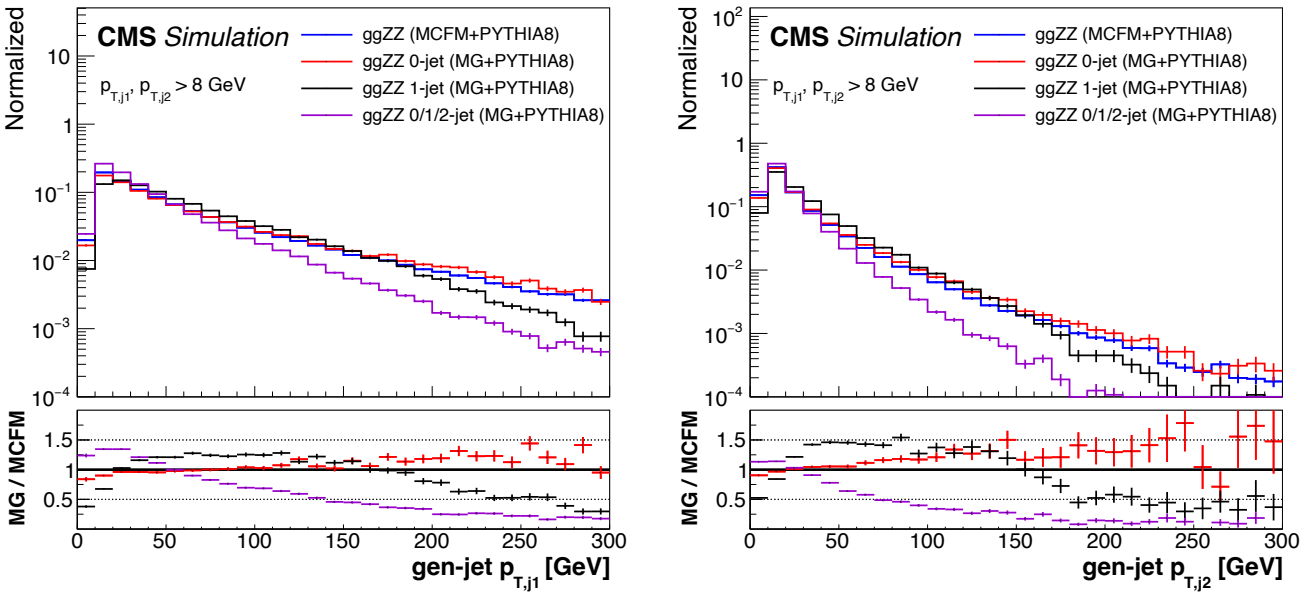


Figure 4.2: The p_T spectrum of the two leading jets in the QCD loop-induced samples generated with *MCFM* and the new, state-of-the-art samples with up to two jets merged with the MLM matching scheme and generated for the first time for this analysis using the *MG5*. The figure is taken from Ref. [112]

In addition to the state-of-the-art gg sample, an additional sample was generated to validate the former. The simulation was done at LO with 1 jet using MG5 and the following syntax:

$$generate\ gg > zzj\ [noborn = QCD],\ z > l + l$$

Pythia8 was again used to perform the decay of the Z bosons and thus the correlation between the spin of the decay leptons is ignored. For this reason another sample was produced at LO using *MCFM* 7.0 [115].

The *Pythia8* package was used for parton showering and hadronization for all MC samples, with parameters set by the CUETP8M1 tune [116] for the 2016 and the CP5 tune [117] for the 2017 and 2018 data-taking periods.
A NNPDF 3.0 parton distribution function (PDF) was used for all 2016 samples and NNPDF 3.1 for all 2017 and 2018 samples [118]. MC samples are reweighted with the true number of interactions in each event to match the level of PU observed in the data.

All simulated backgrounds are summarized in Table 4.1. The dijet mass for WZZ and ZZZ at the generator level is required to be smaller than 100 GeV in order to avoid double counting with the signal sample.

Process	Generator	Cross section [fb]	Remarks
signal samples for 2016, (2017 and 2018)			
$ZZ \to 4l\ +\ 2\ jets$	MadGraph (LO)	0.441 (427)	$m_{jj} > 100\ GeV$
$ZZ \to 4\mu\ +\ 2\ jets$	Phantom (LO)	0.418	used to cross-check MadGraph sample
$ZZ \to 4e\ +\ 2\ jets$	Phantom (LO)	0.418	used to cross-check MadGraph sample
$ZZ \to 2e2\mu\ +\ 2\ jets$	Phantom (LO)	0.836	used to cross-check MadGraph sample
irreducible background samples for 2016, 2017 and 2018			
$ZZ \to 4l\ +\ 0,1\ jets$	MadGraph (NLO)	1218	
$gg \to ZZ \to 4l\ +$ $0,1,2\ jets$	MadGraph (LO)	5.84	cross section computed at $\mu = m_{ZZ}/2$
$gg \to ZZ \to 4l\ +$ $1\ jet$	MadGraph (LO)	4.45	used to cross-check nominal sample
$gg \to ZZ \to 4\mu$	MCFM (LO)	1.59	used to cross-check MG5 samples
$gg \to ZZ \to 4e$	MCFM (LO)	1.59	used to cross-check MG5 samples
$gg \to ZZ \to 2e2\mu$	MCFM (LO)	3.19	used to cross-check MG5 samples
minor background samples for 2016, 2017 and 2018			
$t\bar{t}Z \to 4l2\nu$	MadGraph	253	
$WWZ\ +\ jets$	MadGraph (NLO)	165.1	
$WZZ\ +\ jets$	MadGraph (NLO)	55.7	inclusive decays, $m_{jj} < 100\ GeV$
$ZZZ\ +\ jets$	MadGraph (NLO)	14.0	inclusive decays, $m_{jj} < 100\ GeV$

Table 4.1: List of signal and background samples used in the analysis for the 2016, 2017 and 2018 data-taking periods.

Although the qqZZ background is simulated at NLO, the cross section has been computed at NNLO [119]. Thus, NNLO/NLO k-factors for the qqZZ process are applied to the MG5 sample as a function of $m(ZZ)$.
In addition, NLO EWK corrections dependent on the initial-state quark flavour and kinematics are applied to the qqZZ background in the region $m(ZZ) > 2m(Z)$ [120].
For the ggZZ background, the NLO/LO (NNLO/NLO) k-factor of 1.53 (1.64) extracted from [121, 122] was applied.

Reducible background

The reducible background for the $ZZ \rightarrow 4l$ analysis, henceforth Z+X, comes from processes which contain one or more non-prompt leptons in the four-lepton final state. The main source of such leptons are non-isolated electrons and muons coming from the decays of the heavy-flavour mesons, mis-reconstructed jets usually coming from the light-flavour quarks, and photon conversions. Any such occurrence will be referred to as the "fake lepton".

The contribution from the Z+X background is minor ($\approx 1\%$) and is estimated by measuring the ratios of fake electrons and fake muons which also pass the final selection criteria over those which do pass the loose selection criteria. The selection criteria are discussed in section 4.3. These ratios, referred to as the *fake rates*, are used to extract the expected background yields in the signal region.
A detailed description of the procedure is not needed to follow the analysis presented in this chapter and is left out. However, an interested reader can find a detailed discussion on the measurement of fake rates elsewhere [90, 112].

4.2.2 Data samples

This analysis uses the data collected in 2016, 2017 and 2018 data-taking periods corresponding to an integrated luminosity of $137\ fb^{-1}$. Only the data that passed the quality certification by all detector subsystems, stored in so-called golden JSON files, are used in the analysis. These are processed and stored in file formats that are easier to use in the analyses. One such format, known as the MINIAOD [123], was used here.

The analysis relies on five different primary data sets (PDs): DoubleEG, DoubleMu, MuonEG, SingleElectron, and SingleMuon. Each of these PDs combines a certain collection of HLT paths with exact requirements dependent on the data-taking period. Two primary data sets, DoubleEG and SingleElectron, were merged in 2018 into EGamma PD. Run periods used, together with reconstruction versions, are listed in Table 4.2.
The HLT paths used in the three data-taking periods are shown in Tables 4.3 - 4.5.

To avoid duplicate events from different primary data sets, events are taken:

- from DoubleEG
  - if events pass the diEle trigger (HLT EleXX EleYY CaloIdXX TrackIdXX IsoXX(DZ))
  - or if events pass the triEle trigger (HLT EleXX EleYY EleZZ CaloIdXX TrackIdXX)
- from DoubleMuon
  - if events pass the diMuon trigger (HLT MuXX TrkIsoVVL MuYY TrkIsoVVL)
  - or if events pass the triMuon trigger (HLT TripleMu XX YY ZZ)
  - and if events fail the diEle and triEle triggers
- from MuEG
  - if events pass the MuEle trigger (HLT MuXX TrkIsoXX EleYY CaloIdYY TrackIdYY IsoYY)
  - or if events pass MuDiEle trigger (HLT MuXX DiEleYY CaloIdYY TrackIdYY)
  - or if events pass DiMuEle trigger (HLT DiMuXX EleYY CaloIdYY TrackIdYY)
  - and if events fail diEle, triEle, diMuon and triMuon triggers

- from SingleElectron
  - if events pass the singleElectron trigger (HLT EleXX etaXX WPLoose/Tight(Gsf))
  - and if events fail all triggers above
- from SingleMuon
  - if events pass the singleMuon trigger (HLT IsoMuXX OR HLT IsoTkMuXX)
  - and if events fail all triggers above

where XX, YY and ZZ are year-dependent thresholds.

Primary data set	**Run and reconstruction version**
DoubleMuon DoubleEG MuonEG SingleMuon SingleElectron	Run2016B-17Jul2018-v1
	Run2016C-17Jul2018-v1
	Run2016D-17Jul2018-v1
	Run2016E-17Jul2018-v1
	Run2016F-17Jul2018-v1
	Run2016G-17Jul2018-v1
	Run2016H-17Jul2018-v1
DoubleMuon	Run2017B-31Mar2018-v1
DoubleEG	Run2017C-31Mar2018-v1
MuonEG	Run2017D-31Mar2018-v1
SingleMuon	Run2017E-31Mar2018-v1
SingleElectron	Run2017F-31Mar2018-v1
DoubleMuon	Run2018A-17Sep2018-v1
MuonEG	Run2018B-17Sep2018-v1
SingleMuon	Run2018C-17Sep2018-v1
EGamma	Run2018D-PromptReco-v2

Table 4.2: The list of data samples used in the analysis. All runs for each of the data streams are used, for a total of 76 primary data sets in the MINIAOD format.

HLT path	prescale	primary data set
HLT_Ele17_Ele12_CaloIdL_TrackIdL_IsoVL_DZ	1	DoubleEG
HLT_Ele23_Ele12_CaloIdL_TrackIdL_IsoVL_DZ	1	DoubleEG
HLT_DoubleEle33_CaloIdL_GsfTrkIdVL	1	DoubleEG
HLT_Ele16_Ele12_Ele8_CaloIdL_TrackIdL	1	DoubleEG
HLT_Mu17_TrkIsoVVL_Mu8_TrkIsoVVL	1	DoubleMuon
HLT_Mu17_TrkIsoVVL_TkMu8_TrkIsoVVL	1	DoubleMuon
HLT_TripleMu_12_10_5	1	DoubleMuon
HLT_Mu8_TrkIsoVVL_Ele17_CaloIdL_TrackIdL_IsoVL	1	MuonEG
HLT_Mu8_TrkIsoVVL_Ele23_CaloIdL_TrackIdL_IsoVL	1	MuonEG
HLT_Mu17_TrkIsoVVL_Ele12_CaloIdL_TrackIdL_IsoVL	1	MuonEG
HLT_Mu23_TrkIsoVVL_Ele12_CaloIdL_TrackIdL_IsoVL	1	MuonEG
HLT_Mu23_TrkIsoVVL_Ele8_CaloIdL_TrackIdL_IsoVL	1	MuonEG
HLT_Mu8_DiEle12_CaloIdL_TrackIdL	1	MuonEG
HLT_DiMu9_Ele9_CaloIdL_TrackIdL	1	MuonEG
HLT_Ele25_eta2p1_WPTight	1	SingleElectron
HLT_Ele27_WPTight	1	SingleElectron
HLT_Ele27_eta2p1_WPLoose_Gsf	1	SingleElectron
HLT_IsoMu20 OR HLT_IsoTkMu20	1	SingleMuon
HLT_IsoMu22 OR HLT_IsoTkMu22	1	SingleMuon

Table 4.3: HLT paths for the 2016 data-taking period

HLT path	prescale	primary data set
HLT_Ele23_Ele12_CaloIdL_TrackIdL_IsoVL_*	1	DoubleEG
HLT_DoubleEle33_CaloIdL_GsfTrkIdVL	1	DoubleEG
HLT_Ele16_Ele12_Ele8_CaloIdL_TrackIdL	1	DoubleEG
HLT_Mu17_TrkIsoVVL_Mu8_TrkIsoVVL_DZ_Mass3p8	1	DoubleMuon
HLT_Mu17_TrkIsoVVL_Mu8_TrkIsoVVL_DZ_Mass8	1	DoubleMuon
HLT_TripleMu_12_10_5	1	DoubleMuon
HLT_TripleMu_10_5_5_D2	1	DoubleMuon
HLT_Mu23_TrkIsoVVL_Ele12_CaloIdL_TrackIdL_IsoVL	1	MuonEG
HLT_Mu8_TrkIsoVVL_Ele23_CaloIdL_TrackIdL_IsoVL_DZ	1	MuonEG
HLT_Mu12_TrkIsoVVL_Ele23_CaloIdL_TrackIdL_IsoVL_DZ	1	MuonEG
HLT_Mu23_TrkIsoVVL_Ele12_CaloIdL_TrackIdL_IsoVL_DZ	1	MuonEG
HLT_DiMu9_Ele9_CaloIdL_TrackIdL_DZ	1	MuonEG
HLT_Mu8_DiEle12_CaloIdL_TrackIdL	1	MuonEG
HLT_Mu8_DiEle12_CaloIdL_TrackIdL_DZ	1	MuonEG
HLT_Ele35_WPTight_Gsf_v*	1	SingleElectron
HLT_Ele38_WPTight_Gsf_v*	1	SingleElectron
HLT_Ele40_WPTight_Gsf_v*	1	SingleElectron
HLT_IsoMu27	1	SingleMuon

Table 4.4: HLT paths for the 2017 data-taking period

HLT path	prescale	primary data set
HLT_Ele23_Ele12_CaloIdL_TrackIdL_IsoVL_v*	1	EGamma
HLT_DoubleEle25_CaloIdL_MW_v*	1	EGamma
HLT_Mu17_TrkIsoVVL_Mu8_TrkIsoVVL_DZ_Mass3p8_v*	1	DoubleMuon
HLT_Mu23_TrkIsoVVL_Ele12_CaloIdL_TrackIdL_IsoVL_v*	1	MuonEG
HLT_Mu8_TrkIsoVVL_Ele23_CaloIdL_TrackIdL_IsoVL_DZ_v*	1	MuonEG
HLT_Mu12_TrkIsoVVL_Ele23_CaloIdL_TrackIdL_IsoVL_DZ_v*	1	MuonEG
HLT_DiMu9_Ele9_CaloIdL_TrackIdL_DZ_v*	1	MuonEG
HLT_Ele32_WPTight_Gsf_v*	1	EGamma
HLT_IsoMu24_v*	1	SingleMuon

Table 4.5: HLT paths for the 2018 data-taking period

4.3 Event selection

The final state in this analysis consists of at least two Z bosons decaying into pairs of oppositely charged leptons accompanied by two hadronic jets. The hallmark sign of the signal events are the two hadronic jets with a large pseudorapidity gap between them. In order to maximize the measurement sensitivity, a set of selection criteria was used.
The object reconstruction is based on the PF algorithm which uses information from all CMS subdetectors to identify individual particles within an event. These, so-called, PF candidates are then classified as either electrons, muons, photons, neutral hadrons or charged hadrons. Higher-level objects such as jets and isolated leptons are created from PF candidates [124, 125].

Electrons

Reconstructed electrons with $p_T > 7\ GeV$ and $|\eta| < 2.5$ that also satisfy a loose primary vertex constraint defined by $|d_{xy}| < 0.5\ cm$ and $|d_z| < 1\ cm$, so called *loose electron*s, are considered for the analysis. Requirements on SIP parameter, presented in section 3.3.1, were imposed as well. In addition, electrons coming from the decaying Z bosons are required to pass the MVA ID discussed in section 3.3.3. To account for the detector effects on electron momentum and energy, corrections were applied to MC simulations using the information from the data. $Z \to ee$ sample was used to match the reconstructed dielectron mass spectrum in data to the one in the simulation. This was discussed in section 3.2.5. Discrepancies between the data and MC samples are corrected as presented in section 3.4.1.
Those electrons that pass all presented requirements, so-called *tight electrons*, are considered candidates from which a Z bosons can be built.

Muons

Loose muons are defined with $p_T > 5\ GeV$, $|\eta| < 2.4$, $|d_{xy}| < 0.5\ cm$ and $|d_z| < 1\ cm$. The same requirements on the SIP parameter, as for electrons, are required.
Unlike electrons, muon identification and isolation are done separately. Loose muons with $p_T < 200\ GeV$ are considered identified muons if they also pass the PF muon ID, while loose muons with $p_T > 200\ GeV$ are considered identified muons if they pass the PF ID or the Tracker High-p_T ID [112].
Muons are required to be isolated and this is done using PF-based isolation. Muon isolation is defined by the parameter R_{iso} which measures activity in the cone of radius ΔR around the lepton and is defined as

$$R_{iso} = \left[\sum_{\substack{charged \\ hardons}} p_T + max\ (0, \sum_{\substack{neutral \\ hadrons}} E_T + \sum_{photons} E_T - \Delta\beta) \right] \Big/ p_T^l$$

where the sum runs over the charged and neutral hadrons and photons in the cone of radius ΔR around the lepton. The $\Delta\beta$ correction defined as $\Delta\beta = \frac{1}{2} \sum_{PU}^{chargedhad.} p_T$ gives an estimate of the energy deposit of neutral particles from the PU vertices and is used to remove the PU contribution for muons. The parameter ΔR is set to 0.3, and the isolation requirement is satisfied if $R_{iso} < 0.35$. The muon momentum scale is measured in data by fitting a CB function to the di-muon mass spectrum around the Z boson peak in the $Z \to \mu\mu$ control region. Like for electrons, the discrepancy between the data and MC is cured by applying SFs obtained using the TnP method.
Those muons that pass all presented requirements, so-called *tight muons*, are considered candidates from which a Z bosons can be built

FSR recovery

The Final State Radiation (FSR) recovery algorithm, used to recover jets coming from FSR, was simplified since the Run 1, without degrading the performance. Since the effect of FSR on this analysis is small, the details of the algorithm itself are omitted. An interested reader can find the full description elsewhere [90].

Jets

Jets are reconstructed from PF candidates using the $anti-k_t$ algorithm with a distance parameter $R = 0.4$ [126], after rejecting charged hadrons that are associated with a PU primary vertex. In order to be included in the analysis, all jets must have a corrected p_T larger than 30 GeV and should be within $|\eta| < 4.7$
In order to achieve a good reconstruction efficiency and to mitigate background and PU effects, tight ID criteria was applied on jets. In order to mitigate the PU contamination, a multivariate variable, the pileup jet ID (PUJetID), based on the compatibility of the associated tracks with the primary vertex and the topology of the jet shape, was applied. Additionally, jets are cleaned from any tight lepton and FSR photons by requiring $\Delta R(j, l/\gamma) > 0.4$.

Since the detector response to particles is not linear, the energy of the reconstructed jets does not correspond to the true particle-level energy. For this reason, the reconstructed jet energy is corrected to take into account effects such as interactions with matter, PU, and detector response and response. These corrections are derived from simulations and are crosschecked by studying energy balance in dijet, multijet, γ + jet and leptonic Z/γ + jet events [21, 127].

Unpredicted issues occurred during the three data-taking periods, which impact the quality of the reconstructed jets. In order to remedy the situation, additional requirements were imposed on jets. In 2018 it was noticed that a significant fraction of ECAL trigger primitives (TPs) in the forward region were wrongly associated with the previous bunch crossing. This was due to the degraded transparency of the ECAL crystals in the forward regions which resulted in the distorted shape of the electrical signal. Consequently, signals from the N^{th} bunch crossing were, in some cases, wrongly associated to the $N-1^{th}$ bunch crossing. If the early fired L1 object has E_T above the threshold, a previous event will be sent to the HLT instead of the current event that will be rejected. This feature in the 2016 and 2017 data-taking periods is called L1 prefiring and was mitigated by calculating the probability that the event didn't prefire and then applying this as a weight to the simulations. This was corrected in 2018 by a recalibration of the ECAL [96].

An increase in the ECAL noise in the 2017 data-taking period caused the appearance of peaks (henceforth "horns") in the jet η distributions around $2.5 < |\eta_{jet}| < 3$. The effect of these horns on the analysis was tested by removing soft jets with $p_T < 50\ GeV$ in $2.65 < |\eta| < 3.139$ region. No significant impact was observed.

ZZ selection

The four lepton candidates are built from the tight leptons discussed earlier. An additional lepton cleaning is performed by removing electrons that are within $\Delta R < 0.05$ of the selected muon. This removes fake electrons that arise from a muon track being wrongly matched to the electromagnetic cluster coming from an FSR emission along with the muon.

A *Z candidate* is defined as the pair of same-flavor, and opposite charge leptons (e^+e^- or $\mu^+\mu^-$) with dilepton invariant mass within $60\ GeV < m_{ll} < 120\ GeV$. The Z boson mass includes FSR photons if identified.

A *ZZ candidate* is defined as a pair of non-overlapping Z candidates and satisfies the following requirements:

1. **Ghost removal**: $\Delta R(\eta, \phi) > 0.02$ between each of the four leptons.
2. **lepton p_T**: two out of the four selected leptons should satisfy $p_T(l_1) > 20\ GeV$ and $p_T(l_2) > 10\ GeV$.
3. **Z mass**: the mass of both Z_1 and Z_2 must be larger than $60\ GeV$ in order to comply with MC samples that do not describe the off-shell ZZ^* distributions. Here, the Z boson with mass closest to the nominal Z mass is denoted Z_1.
4. **four-lepton invariant mass**: $m_{4l} > 180\ GeV$ in order to comply with MC samples that do not describe the off-shell ZZ^* distributions.
5. **QCD suppression**: regardless of flavor, all four opposite-sign pairs that can be built from the four leptons must satisfy $m_{ll} > 4\ GeV$. Selected FSR photons are not used in the calculation because a dilepton coming from QCD processes (e.g. J/Ψ) may have photons in vicinity (e.g. from π^0).
6. **"smart cut"**: defining Z_a and Z_b as the mass-sorted alternative pairing Z candidates (Z_a is the one with mass closest to the nominal Z mass), that satisfy $NOT(|m_{Za} - m_Z| < |m_{Z1} - m_Z|\ AND\ m_{Zb} < 12GeV)$. Here, the FSR photons are not included in calculation of the m_Z. This cut removes $4e$ and 4μ candidates where the alternative pairing looks like an on-shell Z boson accompanied by a low-mass lepton pair.

Only events containing at least one selected ZZ candidate are kept. If more ZZ candidates pass the selection requirements, the pair with the largest scalar p_T sum of the leptons constituting the Z_2 candidate is selected. This is because the false ZZ candidates are likely to be built from fake leptons which are more prominent at low p_T.

Inclusive and VBS selections

In order to select a VBS signal enriched phase space, an additional set of requirements is imposed. At least two jets with $|\eta| < 4.7$ and $p_T > 30\ GeV$ are required in an event. In case more than two jets are present in an event, the two with the highest p_T, referred to as the *tagging jets*, are taken. The tagging jets are required to have an invariant mass above 100 GeV in order to suppress hadronic W decays.
This set of requirements, on top of the ZZ selection, is referred to as the *inclusive selection* and was used to measure the signal significance, the total fiducial cross sections and to set limits on the aQGCs.

In addition, two more selections were defined to perform the measurement of the VBS and VBS+QCD cross sections. A *loose VBS selection* requires, on top of the ZZ selection, $m_{jj} > 400\ GeV$ and $|\Delta\eta| > 2.4$. A *tight*

VBS selection requires $m_{jj} > 1\ TeV$ and $|\Delta\eta| > 2.4$ on top of the ZZ selection.

A control region used to check the agreement between the data and MC, is defined by requiring events to pass the ZZjj inclusive selection, but to fail at least one condition of the loose VBS selection.
All selection criteria are summarized in Table 4.6.

lepton candidates	$p_T^e > 7GeV \qquad p_T^\mu > 5GeV$ $\|\eta\|^e < 2.5 \qquad \|\eta\|^\mu < 2.4$ $\|d_{xy}\| < 0.5\ cm$ $\|d_z\| < 1\ cm$ $\|SIP_{3D}\| < 4$ ID passed iso. in $ID \qquad R_{iso}^\mu < 0.35$
jet candidates	$p_T > 30\ GeV \qquad \|\eta\| < 4.7$ $\Delta R(j, l/\gamma) > 0.4$ ID passed L1 prefiring correction
Z candidate	tight lepton pair (e^+e^- or $\mu^+\mu^-$) $60\ GeV < m_{ll} < 120\ GeV$
ZZ selection	require pair of non-overlapping Z bosons $\Delta R(\eta, \phi) > 0.02$ between each of the four leptons $p_T(l_1) > 20\ GeV \qquad p_T(l_2) > 10\ GeV$ $m_{Z1} > 60\ GeV \qquad m_{Z2} > 60\ GeV$ $m_{4l} > 180\ GeV$ QCD suppression cut "smart" cut
inclusive ZZjj selection	ZZ selection + $m_{jj} > 100GeV$
loose VBS selection	ZZ selection + $m_{jj} > 400GeV + \|\Delta\eta_{jj}\| > 2.4$
tight VBS selection	ZZ selection + $m_{jj} > 1\ TeV + \|\Delta\eta_{jj}\| > 2.4$
control region	ZZ selection + ($m_{jj} < 400GeV$ or $\|\Delta\eta_{jj}\| < 2.4$)

Table 4.6: Summary of the analysis selection criteria.

4.4 VBS observables

The smoking gun sign of VBS are the two hadronic jets separated by a large pseudorapidity gap. Therefore, the most important kinematic variables describing a VBS process are the dijet invariant mass, m_{jj} and the difference in pseudorapidity between the two tagging jets, $\Delta\eta_{jj}$.
Variables $\eta^*(Z_1)$ and $\eta^*(Z_2)$, so-called Zeppenfeld variables defined in Table 4.7, were first introduced as a means of isolating events with no gluon emissions between the tagging jets in the vector boson fusion (VBF) processes [128]. Therefore, they measure activity between the two tagging jets.
Other variables used to isolate VBS are the ratio between the p_T of the tagging jet system and the scalar p_T sum of the tagging jets ($R^{jet}_{p_T}$) and the event balance ($R^{hard}_{p_T}$) defined as the transverse component of the vector sum of the Z bosons and leading jets momenta normalized to the scalar p_T sum of the same objects. The $qgtagger(j_i)$ are probabilities that jets are originating from quarks rather than gluons.
A full list of variables considered for signal extraction is shown in Table 4.7.

variable	definition
m_jj	invariant mass of the two leading jets
$\Delta\eta_{jj}$	pseudorapidity separation of the two leading jets
m_{4l}	invariant mass of the ZZ pair
$\eta^*(Z_1)$	direction of the Z_1 relative to the leading jets: $\eta^*(Z_1) = \eta(Z_1) - \frac{\eta(j_1)+\eta(j_2)}{2}$
$\eta^*(Z_2)$	direction of the Z_2 relative to the leading jets: $\eta^*(Z_2) = \eta(Z_2) - \frac{\eta(j_1)+\eta(j_2)}{2}$
$R^{hard}_{p_T}$	transverse component of the vector sum of the two leading jets and four leptons normalized to the scalar p_T sum of the same objects $R^{hard}_{p_T} = \frac{(\sum_{i=4l,\ 2j} \vec{V_i})_{transverse}}{\sum_{4l,\ 2j} p_T(i)}$
$R^{jet}_{p_T}$	transverse component of the vector sum of the two leading jets normalized to the scalar p_T sum of the same objects $R^{jet}_{p_T} = \frac{(\sum_{i=2j} \vec{V_i})_{transverse}}{\sum_{2j} p_T(i)}$
$p_T(j_1)$	transverse momentum of the leading jet
$p_T(j_2)$	transverse momentum of the second-leading jet
$y(j_1)$	rapidity of the leading jet: $y(j_1) = \frac{1}{2} ln \left[\frac{E(j_1)+p_L(j_1)}{E(j_1)-p_L(j_1)}\right]$
$y(j_2)$	rapidity of the second-leading jet: $y(j_2) = \frac{1}{2} ln \left[\frac{E(j_2)+p_L(j_2)}{E(j_2)-p_L(j_2)}\right]$
$\eta(j_1)$	pseudorapidity of the leading jet
$\eta(j_2)$	pseudorapidity of the second-leading jet
$\lvert\eta_{min}(j)\rvert$	smallest absolute value of the jet pseudorapidity
$\lvert\eta_{max}(j)\rvert$	largest absolute value of the jet pseudorapidity

$\sum \eta(j)$	sum of the pseudorapidity of selected jets
$\sum \lvert\eta(j)\rvert$	sum of the absolute value of the pseudorapidity of selected jets
$m_{jj}/\Delta\eta(jj)$	quotient of the invariant mass and the pseudorapidity gap of the two leading jet
$qgtagger(j_1)$	probability that the leading jet is coming from a quark rather than a gluon
$qgtagger(j_2)$	probability that the second-leading jet is coming from a quark rather than a gluon
$p_T(l_3)$	transverse momentum of the third-leading lepton
$\lvert\eta_{min}(lep)\rvert$	smallest absolute value of the lepton pseudorapidity
$\lvert\eta_{max}(lep)\rvert$	largest absolute value of the lepton pseudorapidity
$p_T(Z_1)$	transverse momntum of the Z_1
$p_T(Z_2)$	transverse momntum of the Z_2
$y(Z_1)$	rapidity of the Z_1: $y(Z_1) = \frac{1}{2} ln \left[\frac{E(Z_1)+p_L(Z_1)}{E(Z_1)-p_L(Z_1)}\right]$
$y(Z_2)$	rapidity of the Z_2: $y(Z_2) = \frac{1}{2} ln \left[\frac{E(Z_2)+p_L(Z_2)}{E(Z_2)-p_L(Z_2)}\right]$
$\Delta\phi(Z_1, Z_2)$	angular separation between the two Z bosons

Table 4.7: Set of 28 variables used to check the agreement between the data and MC.

Distributions of variables m_{jj} and $|\Delta\eta_{jj}|$, used to define the control region, are shown in Fig. 4.3 for all three data-taking periods and demonstrate a good agreement between the data and the simulation.
A good agreement between the data and the simulation is also observed in the signal region for a full set of variables used to extract the signal. This can be seen in Fig. 4.4 for the 2018 data-taking period with the baseline selection applied. All distributions for the 2016 and 2017 data-taking periods can be found in Appendix A.

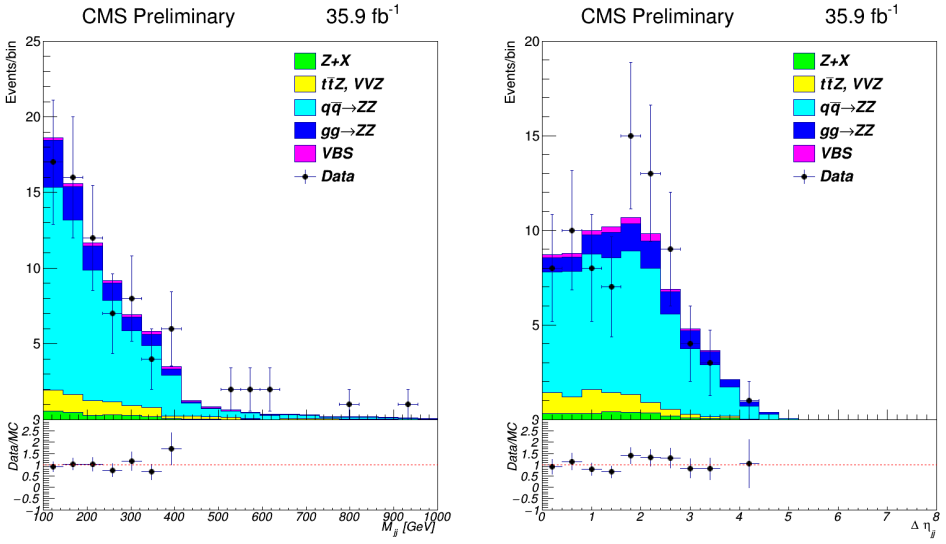


Figure 4.3

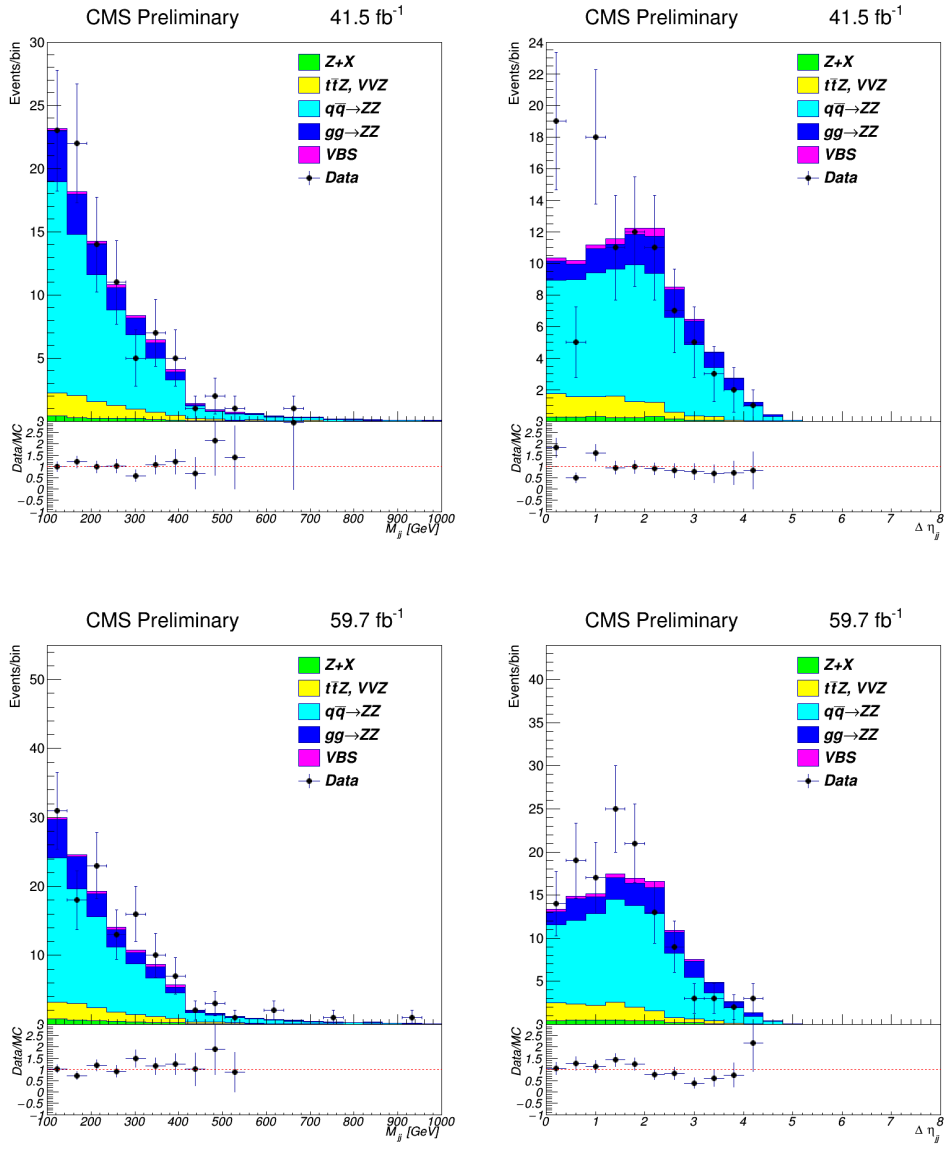


Figure 4.3: A comparison of data to the background and signal estimations in 2016 (top row), 2017 (middle row) and 2018 (bottom row) samples in the control region.

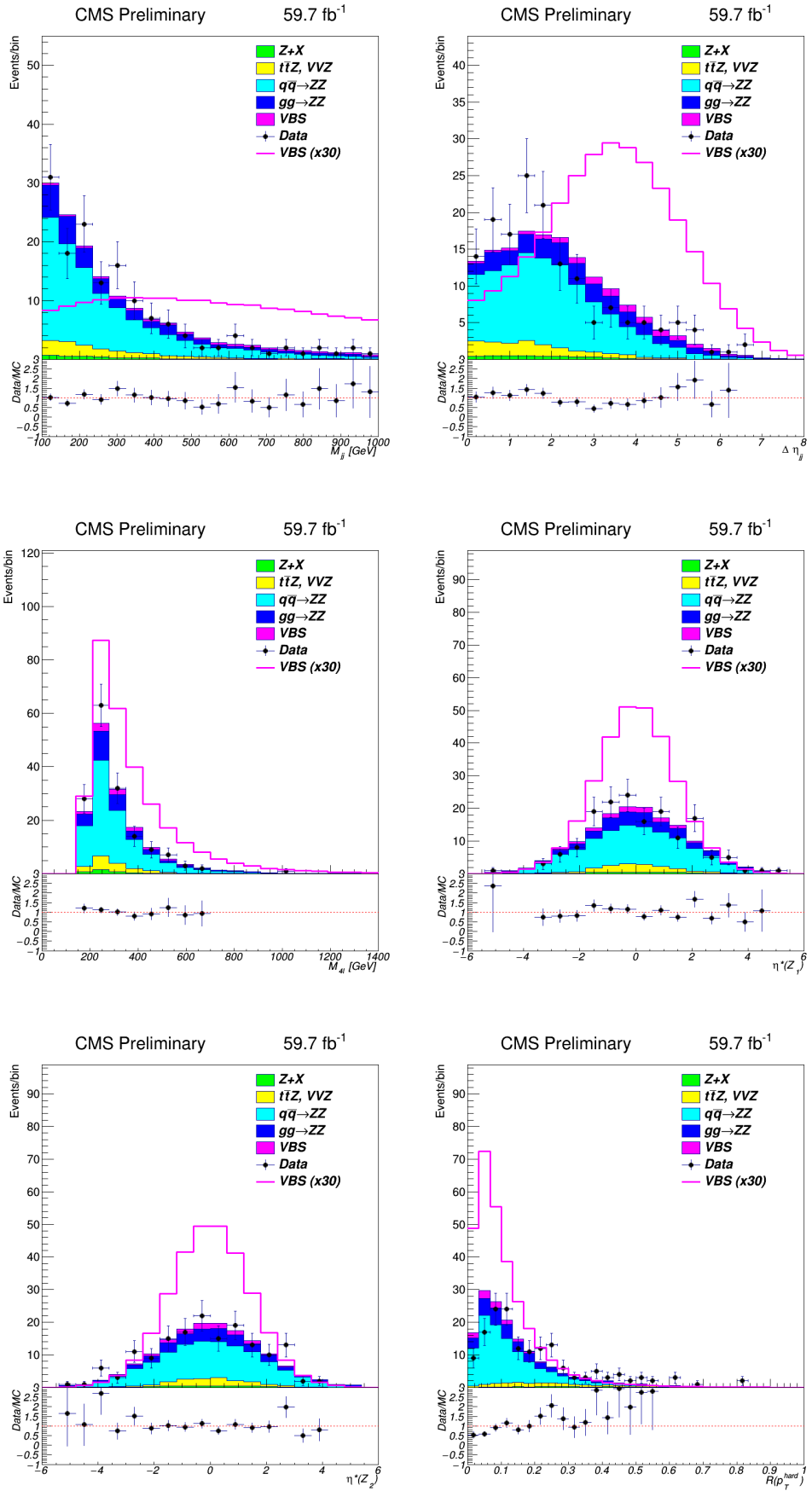

CMS Preliminary 59.7 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
M_jj [GeV]
Δ η_jj
M_4l [GeV]
η*(Z_1)
η*(Z_2)
R(p_T^hard)

Figure 4.4

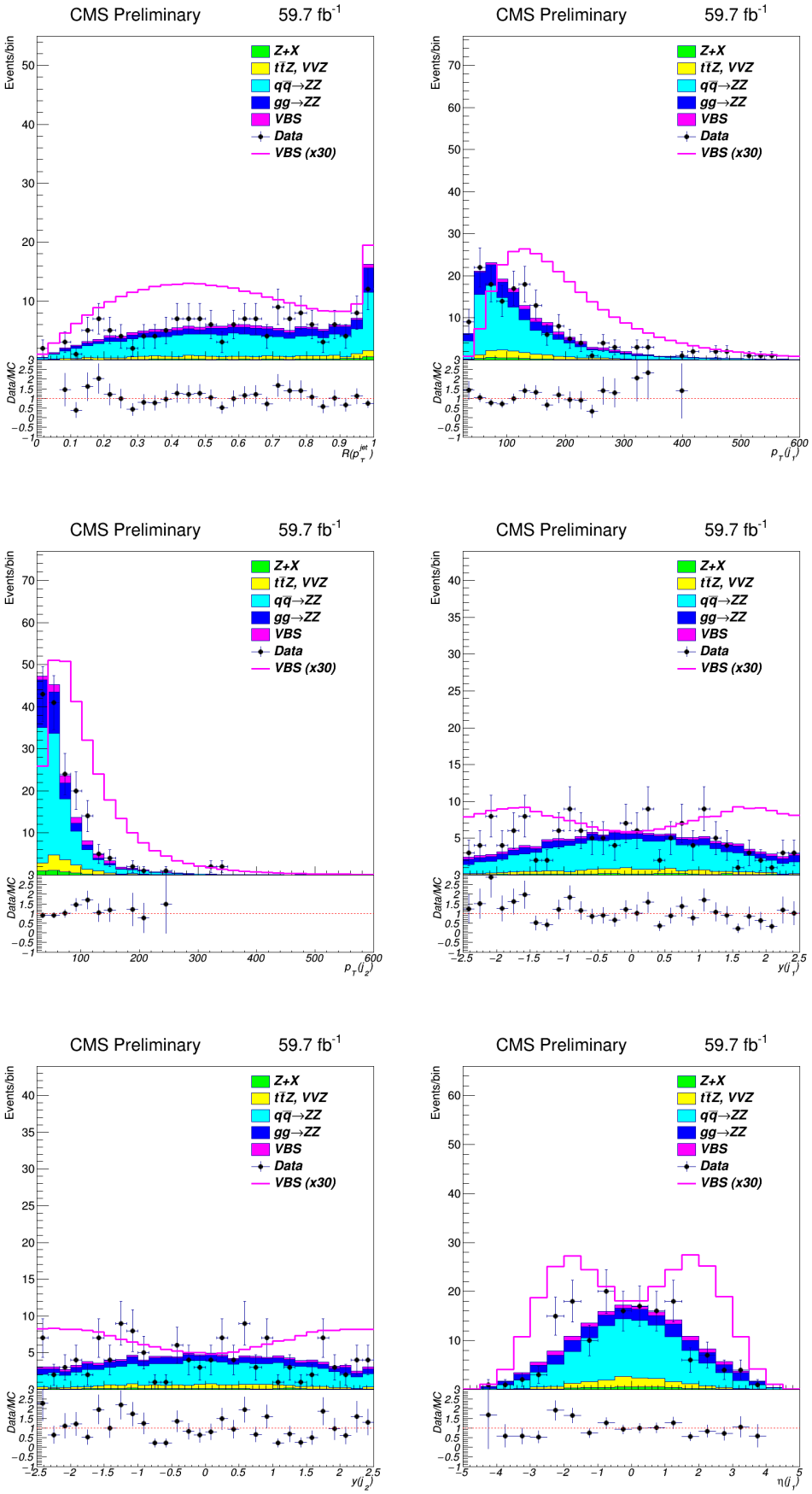

CMS Preliminary
59.7 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
R(p_T^jet)
p_T(j_1)
p_T(j_2)
y(j_1)
y(j_2)
η(j_1)

Figure 4.4

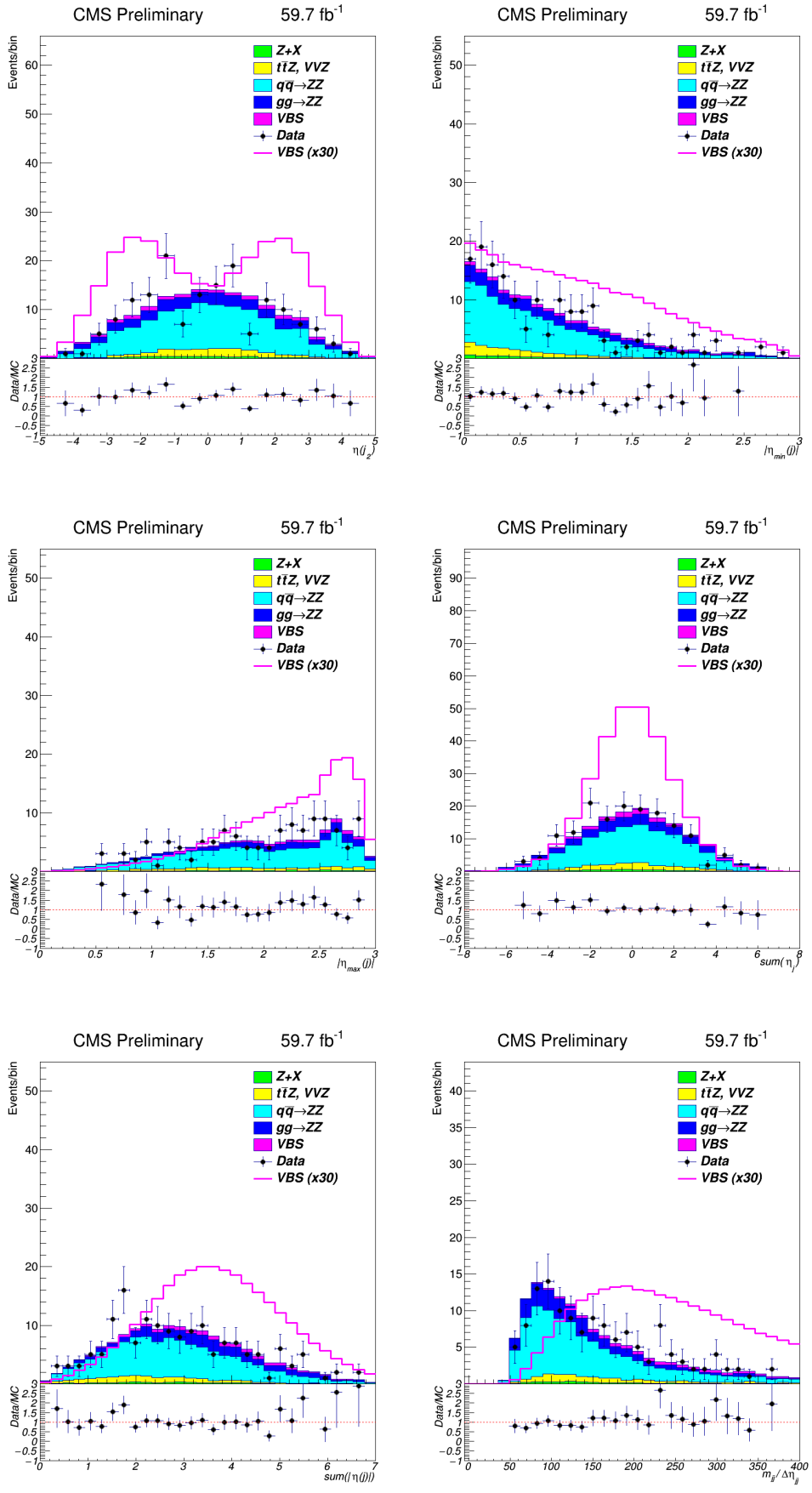

CMS Preliminary 59.7 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
η(j_z)
|η_min(j)|
|η_max(j)|
sum(η_j)
sum(|η(j)|)
m_jj / Δη_jj

Figure 4.4

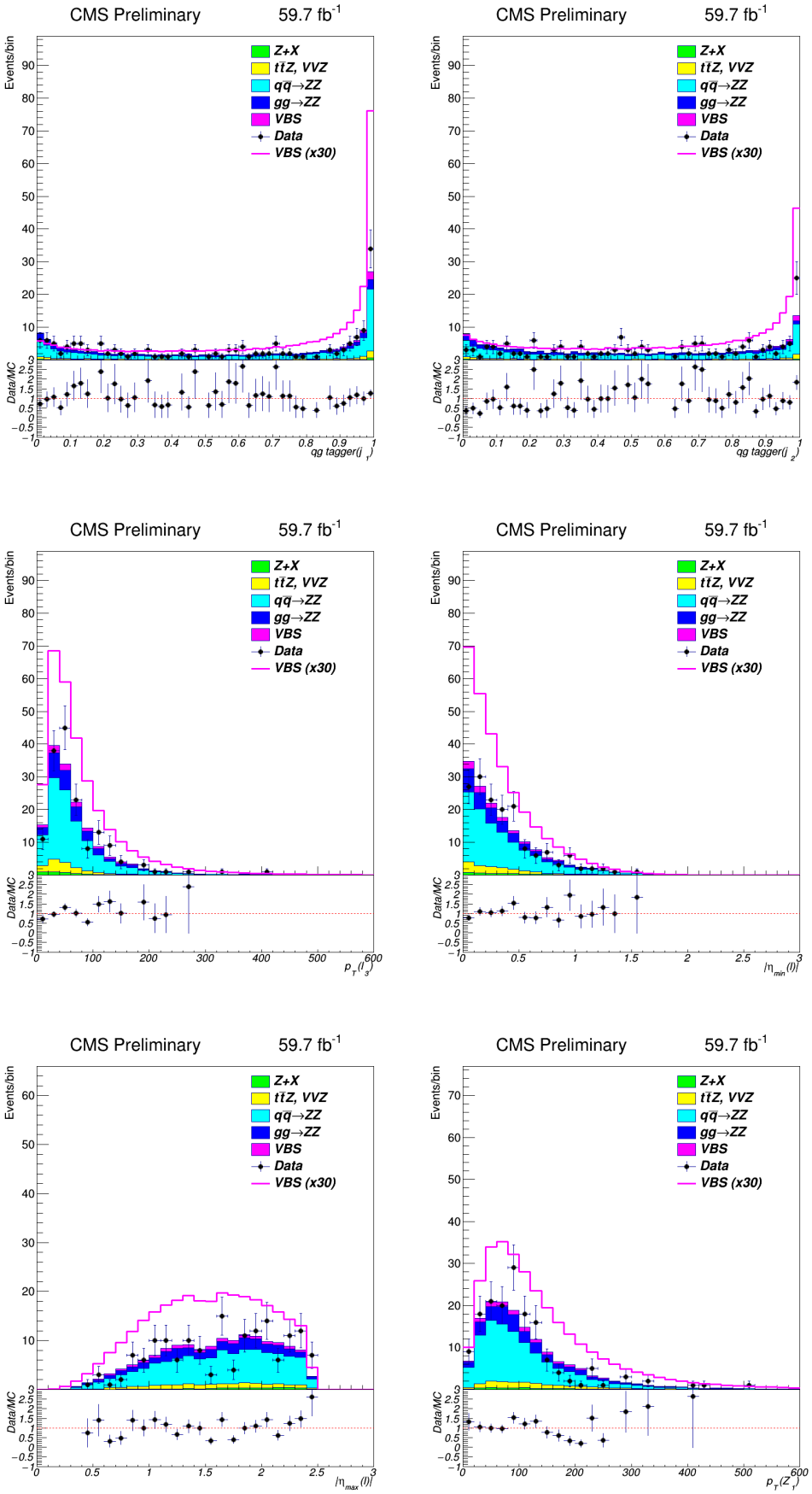

CMS Preliminary
59.7 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
qg tagger(j1)
qg tagger(j2)
pT(l3)
|η_min(l)|
|η_max(l)|
pT(Z1)

Figure 4.4

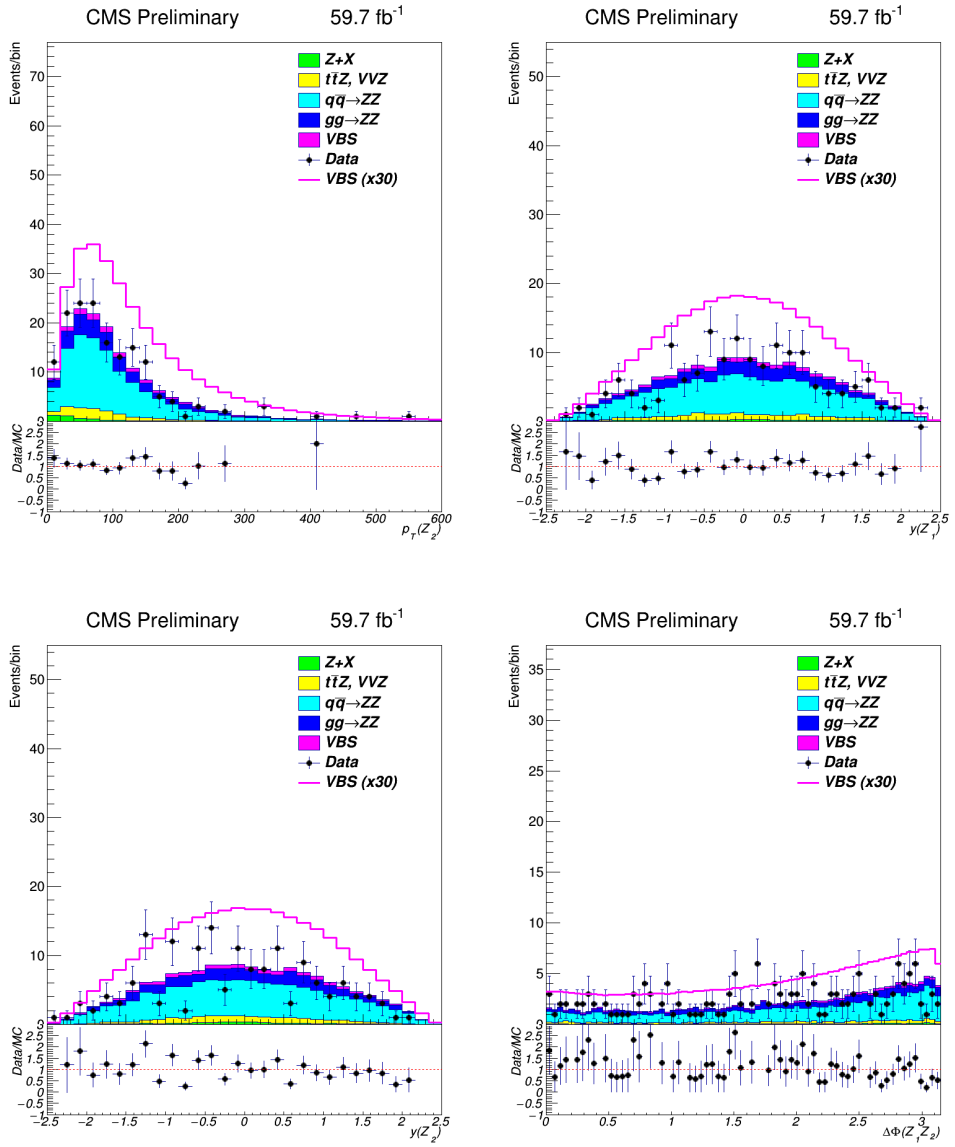


Figure 4.4: Comparison of data to the background and signal estimations in 2018 samples used in the analysis. All 28 variables from Table 4.7 are shown.

4.5 Signal extraction and the cross section measurement using the MELA discriminant

4.5.1 The MELA discriminant

In the published paper [110] on the search for VBS in the $4l$ final state using the Run 2 data, the signal extraction approach was based on a kinematic discriminant (MELA) that uses *MCFM* matrix elements for the EWK signal and the main $qqZZ$ background to describe process probabilities. At the hearth of MELA lies the fact that the kinematics of the VBS $4l$ final state coming from the decay of vector bosons can be fully described by the set of variables summarized in Table 4.8 and illustrated in Fig. 4.5 [129–131].

variable	description
m_{4l}	invariant mass of the 4 final-state leptons
m_{Z_1}	invariant mass of the Z_1
m_{Z_2}	invariant mass of the Z_2
θ^*	angle between the Z_1 boson and the z axis
Φ	angle between the normal vectors of the decay planes of Z_1 and Z_2
Φ_1	angle between the beam axis and the plane of the Z_1 decay products in the $4l$ rest frame
θ_1	angle between Z_1 direction and momenta of the decay lepton in Z_1 rest frame
θ_2	angle between Z_2 direction and momenta of the decay lepton in Z_2 rest frame

Table 4.8: The set of eight variables needed to fully characterize the $4l$ final state originating from the decay of Z bosons. All eight variables are used to construct the kinetic discriminant K_D.

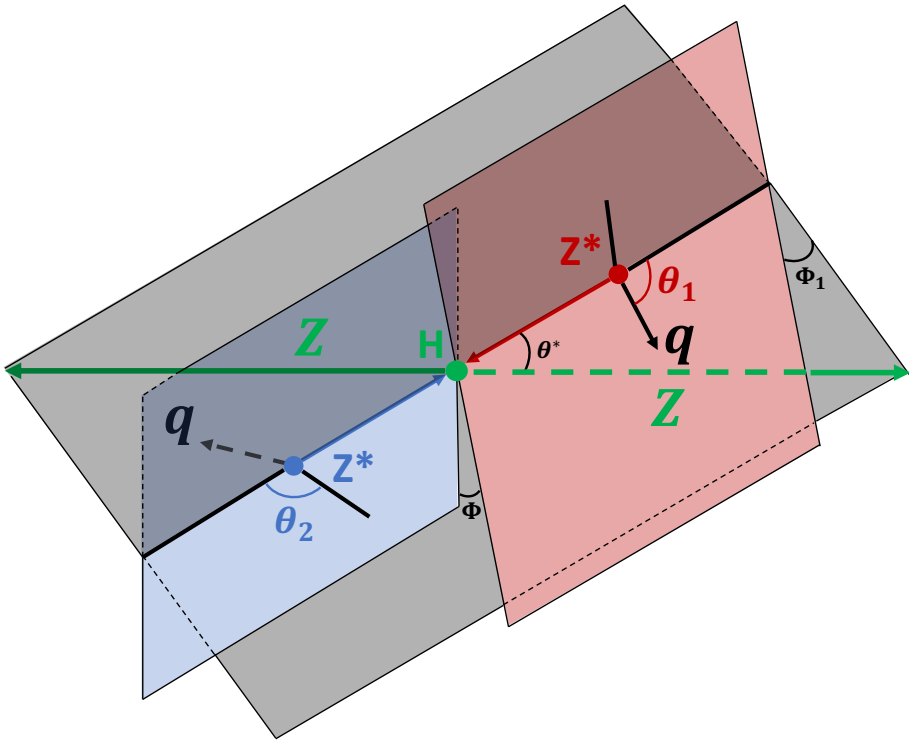


Figure 4.5: Aangles defined in Table 4.8 used, together with the three invariant masses, to build the kinematic discriminant K_D. Illustration shows a Higgs boson decay into two Z bosons that further decay into quarks. This can be modified to show a VBS process with outgoing quarks and two Z bosons by treating quarks in bold as the incoming particles.

From the set of mentioned variables, a kinematic discriminant, K_D is constructed:

$$K_D = \left[1 + c(m_{4l}) \cdot \frac{P_{QCD-JJ}(\vec{\Omega}^{4l+JJ}|m_{4l})}{P_{VBS+VVV}(\vec{\Omega}^{4l+JJ}|m_{4l})}\right]$$

In the expression above, $P_{VBS+VVV}$ represents the probability, obtained from the *MCFM* matrix elements, of an event coming from EWK processes. Similarly, P_{QCD-JJ} is the probability, obtained in the same way, that event originated from the QCD-induced production of the $4l2j$ final state. $\vec{\Omega}$ represents the set of invariant mass and angle variables from Table 4.8. Finally, $c(m_{4l})$ is an m_{4l}-dependent constant that is used to bound the distribution in the range [0, 1].

Figure 4.6 shows a good agreement, in the control region, of the K_D distribution between the data and the simulation for all three data-taking periods.

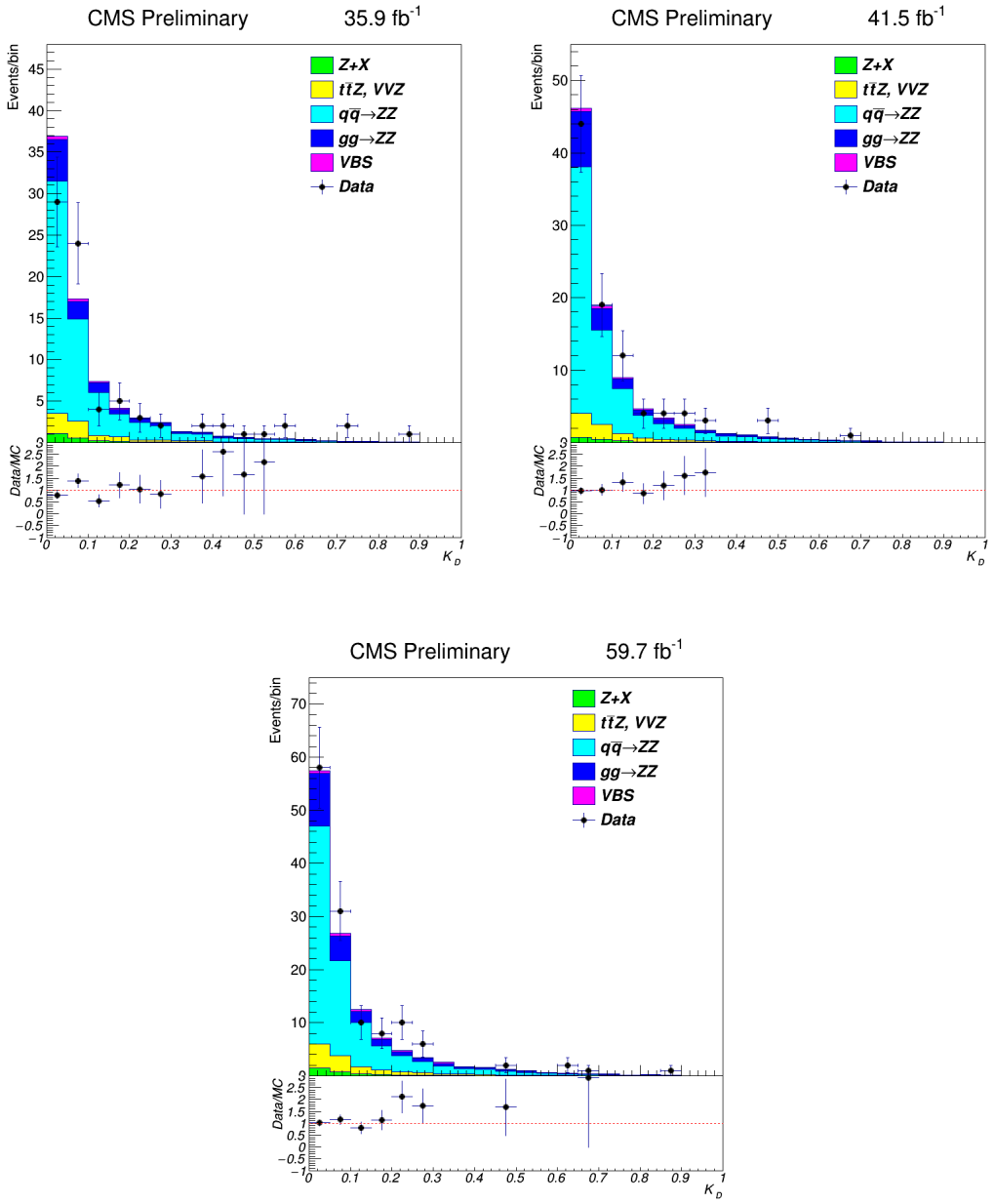


Figure 4.6: Comparison of data to the background and signal estimations, in the control region, for the kinematic discriminant, K_D, in all three data-taking periods. A good agreement between the data and the simulation is observed.

Figure 4.7 shows the performance of the variable in discriminating EWK signal from the backgrounds. The EWK signal is visible in the region with large values of K_D. The baseline selection was applied. The figure also shows a good agreement between the data and the simulation in the VBS-enriched region.

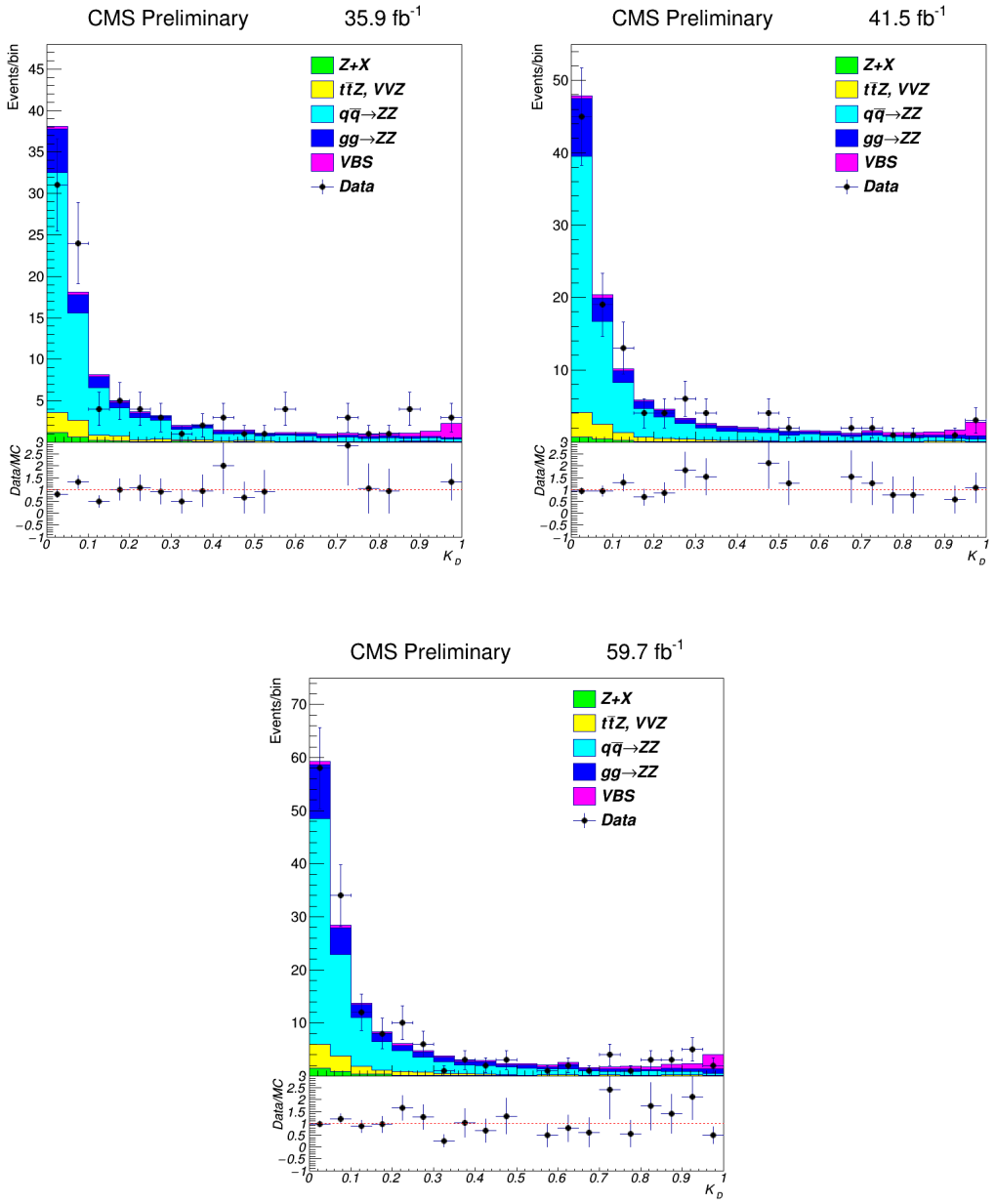


Figure 4.7: Comparison of data to the background and signal estimations, in the signal region, for the kinematic discriminant, K_D, in all three data-taking periods. Plots show the performance of the variable in discriminating between the signal and background distributions. The EWK signal is visible in the region with large values of K_D.

4.5.2 Significance and cross section measurement

The expected and observed significance for the three data-taking periods, as well as the combined significance, was calculated using the "*combine*" tool. This tool was designed to provide the user with a command-line interface to fit a signal and background models to the data.
The previously defined matrix element discriminant, K_D, was used to produce a histogram to model each contribution of interest. These histograms, together with event yields of every process, were fed into the combine tool in the form of configuration files called datacards. All histograms were used as a template to perform a maximum likelihood fit to the observed data. This procedure was done for each year separately. The expected distributions for the signal and the irreducible backgrounds were obtained from the MC simulation. The reducible background was estimated from data.
In each of the three datacards, the systematic uncertainties from all the sources were specified and treated as nuisance parameters in the fit. The sources of the systematic uncertainties are discussed in section 4.8. In order to constrain the QCD-induced production from the background-dominated region of the K_D distribution, the shape and normalization of each contribution were allowed to vary up and down in the fit. The signal significance for the integrated luminosity $\mathcal{L} = 137.1\ fb^{-1}$ was obtained by combining all three periods. This is done simply in *combine* by merging the individual datacards and performing a new fit.

The EWK and EWK+QCD cross sections were estimated in the fiducial regions defined in Table 4.9. These were defined very closely to the selection criteria at the reco level. The same fit used to obtain the signal significance was also used to calculate the signal strength, μ, defined as the ratio of the measured cross section to the SM expectation $\mu = \frac{\sigma}{\sigma_{SM}}$.
Since the K_D spectrum was optimized to separate the EWK signal from the backgrounds, the cross section for the EWK component was obtained by exploiting the shape of the MELA discriminant. The procedure here was identical to the one used to obtain the EWK signal significance. On the other hand, fits that only use event counts in the three fiducial regions were used to obtain the EWK+QCD cross section. This is possible because the EWK+QCD determination is, mostly, background-free.

The next section will describe an alternative signal extraction approach using boosted decision trees. The results for both approaches are discussed and compared in section 4.9.

Particle type	Selection
ZZjj inclusive	
Leptons	$p_T(l_1) > 20\ GeV$ $p_T(l_2) > 10\ GeV$ $p_T(l) > 5\ GeV$ $\|\eta(l)\| < 2.5$
Z and ZZ	$60 < m_{ll} < 120\ GeV$ $m_{4l} > 180\ GeV$
Jets	at least 2 $p_T(j) > 30\ GeV$ $\|\eta(j)\| < 4.7$ $m_{jj} > 100\ GeV$ $\Delta R(j,l) > 0.4$ for each $j,\ l$
VBS-enriched (loose)	
Leptons	same as ZZjj inclusive
Jets	ZZjj inclusive + $\|\Delta\eta_{jj}\| > 2.4$ $m_{jj} > 400\ GeV$
VBS-enriched (tight)	
Leptons	same as ZZjj inclusive
Jets	all above + $m_{jj} > 1\ TeV$

Table 4.9: Particle-level selections used to define the fiducial regions for EWK and EWK+QCD cross sections

4.6 Signal extraction using Boosted Decision Trees

4.6.1 A Tool for MultiVariate Analysis (TMVA)

The signal extraction discussed in the following sections is based on a multivariate approach. For this, the *Toolkit for MultiVariate Analysis* (*TMVA*) was used. The *TMVA* project started in 2005 with the goal of building a consistent, feature-rich framework for multivariate analysis (MVA). It provides a ROOT-integrated [132] environment for processing, parallel evaluation and application of classification and regression techniques. All MVA techniques implemented in the tool are based on supervised learning:

- Fisher
- Linear description (LD)
- Functional description analysis (FDA)

- Projective likelihood
- Cuts
- Probability density estimator—range search (PDE-RS)
- Probability density estimator—foam (PDE-foam)
- Neuronal network (MLP)
- Boosted decision trees (BDT)
- Support vector machine (SVM)
- Rule ensembles (RuleFit)

with each of the techniques implemented in C++/ROOT [133].

Apart from the techniques listed above, *TMVA* offers auxiliary tools such as parameter fitting and various data set transformations. It also provides training, testing and performance evaluation algorithms. Finally, it has implemented a Graphical User Interface (GUI) which enables users to easily obtain desired output plots [97].
The *TMVA* logic for both a classification and a regression problem is as follows

1. The data is fed to *TMVA* via ROOT TTrees or from an ASCII file.
2. The user defines variables from the input file that will be used during the training and test phase.
3. If needed, selection cuts and event weights are defined. At this stage, *TMVA* gives the user a convenient way to select desired preprocessing technique (normalisation, decorrelation, principal components analysis or gaussianisation).
4. The user chooses to do either classification or regression.
5. The desired MVA technique is selected.
6. The hyperparameters are defined for the selected MVA technique.
7. The training is initiated on one part of the available data sample followed by the implementation of the training to the unknown set (i.e. the test set)
8. *TMVA* evaluates the chosen MVA method(s) and produces the result in various formats: Receiver Operating Characteristic (ROC) curve, the curve of signal efficiencies and corresponding background rejection rates for each point on the ROC curve, signal significance, signal purity and a classifier distribution for the signal and background. Each value on the classifier distribution (henceforth the *cut value*) can be used to obtain a pair of (signal efficiency, background rejection) values (henceforth the *working point*).
9. *TMVA* stores the training result in the form of weights available in the "weight" file
10. Saved weights are used for the application of the training on individual signal and background samples

In this analysis, the *Boosted Decision Trees* (*BDT*) classifier was used to extract EWK signal from the backgrounds.

4.6.2 Introduction to Boosted Decision Trees

A decision tree is a supervised machine learning method that can be used in either classification or regression problems. In this thesis, we are interested in labelling each event as either signal or background. Thus, decision trees are used here to solve a classification problem.
A decision tree is a data structure that consists of the root node, decision nodes, leaf nodes and branches. By definition, the root node is simultaneously a decision node. Every decision tree is built starting from the root node. From here, the data are split, using some conditions, into different sub-trees. The process is completed when every branch has only leaf nodes. One simple decision tree is shown in Figure 4.8. Some "buzzwords" used in decision trees theory are summarized in Table 4.10.
In order to understand how decision trees work, a simple example is prepared. The input data described by only two features, i.e. input variables, is shown in Figure 4.9. Every red ball represents a signal and every blue ball represents a background. A green ball represents new data that will be classified once the decision tree is trained. It is not used in the forthcoming calculations.
The first step in the training of the decision tree is to load all training data in the tree. From here the root node is created. The idea behind every node is to consider all available features and select the one that does the best job in splitting the data into signal and background groups. For example, the data in Figure 4.9 can be split in two ways:

1. $x \leq -1$ or $x > -1$
2. $x \leq -4$ or $x > -4$

How does the tree decide which of the two lines it should use to split the data? This is done using the *Attribute Selection Measure* (*ASM*). The two examples of such tools are

- Gini index
- Information gain (IG)

Both give similar results, so only IG will be described here.
The IG is based on the minimization of the entropy obtained after the split. The entropy calculation is based on the information theory:

$$S = -\sum p_i \cdot log(p_i)$$

where p_i represents the probability of finding any class within a subgroup. The base of the logarithm can be arbitrarily chosen and is set to 2. In the beginning, there are 10 red balls out of 20 balls in total (the green ball is not included in the training) which gives the probability of selecting a red ball 50 %. The same argument holds for the blue balls. Thus, the entropy has the maximal value in the beginning

$$S = -\left[0.5 \cdot log(0.5) + 0.5 \cdot log(0.5)\right] = 1$$

The entropy of the region defined by $x \leq -1$ is then

$$S = -\left[\frac{4}{9} \cdot log\left(\frac{4}{9}\right) + \frac{5}{9} \cdot log\left(\frac{5}{9}\right)\right] = 0.99$$

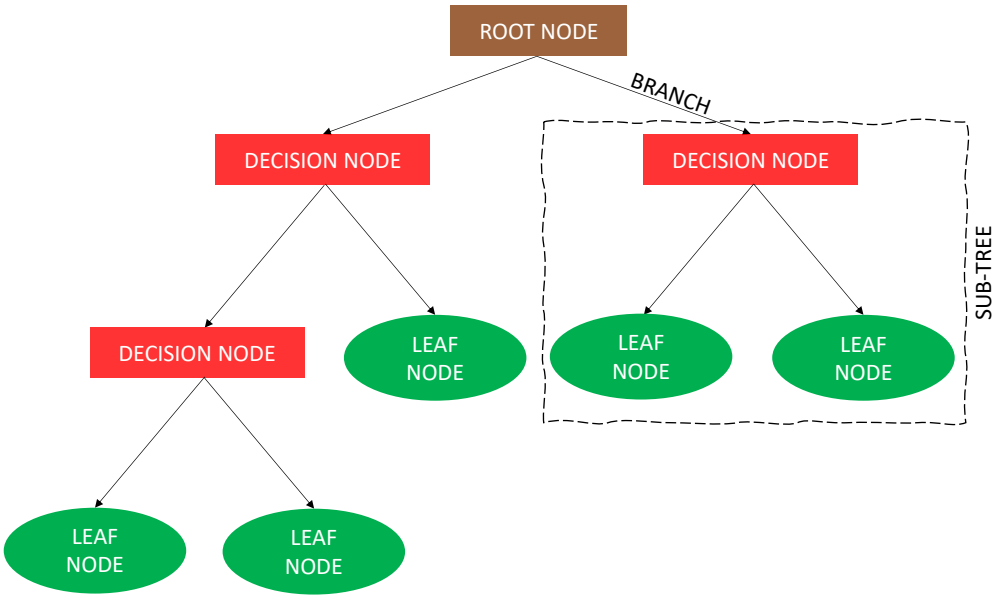


Figure 4.8: Illustration of the simple decision tree with a basic structure.

structure	definition
feature	input variable used in the training of the multivariate classifier
decision node	algorithm that splits the data depending on the value of the feature(s). Each decision node splits the structure and, therefore, creates an additional sub-tree
root node	the first decision node from which the decision tree is built
leaf node	ending node of a branch where all the data is classified as either signal or background. The branching of the tree ends at the leaf node.
branch	decision rule which creates a sub-tree.

Table 4.10: Definition of the basic decision tree structures

For the region defined by $x > -1$ this would be

$$S = -\left[\frac{6}{11} \cdot log\left(\frac{6}{11}\right) + \frac{5}{11} \cdot log\left(\frac{5}{11}\right)\right] = 0.99$$

Now one should calculate the IG from this split:

$$G = S_{parent} - \sum w_i \cdot S_{child}$$

where the factor w_i is the weight defined as the total number of balls in the region of interest divided by the total number of balls. Thus,

$$G_{-1} = 1 - \left[\frac{9}{20} \cdot 0.99 + \frac{11}{20} \cdot 0.99\right] = 0.01$$

For the region defined by $x \leq -4$ we have

$$S = -\left[\frac{1}{5} \cdot log\left(\frac{1}{5}\right) + \frac{4}{5} \cdot log\left(\frac{4}{5}\right)\right] = 0.72$$

and for the region defined by $x > -4$ we have

$$S = -\left[\frac{9}{15} \cdot log\left(\frac{9}{15}\right) + \frac{6}{15} \cdot log\left(\frac{6}{15}\right)\right] = 0.97$$

Thus, the IG for this split is

$$G_{-4} = 1 - \left[\frac{5}{20} \cdot 0.72 + \frac{15}{20} \cdot 0.97\right] = 0.09$$

From this it can be seen that splitting the data with line $X = -4$ results in a larger gain. For this reason the root node will split the data based on the condition $x \leq -4$ or $x > -4$. This is not to say that this is the best splitting option in this example. It was used merely for demonstration purposes.

Although simple, this example describes exactly how the decision tree splits the data using the available features. At every node, the tree will find the best feature using the *ASM* to split the data until nothing is left but leaves.
Before the tree performance is tested on new data, a method of simplifying the tree by means of deleting unnecessary nodes, called pruning, is applied.
Finally, we want to test the performance of our tree by introducing new data (the green ball) and calculating how efficient the tree is in classifying it. The features of the ball will traverse through the entire tree, starting at the root node, until the leaf is reached. When this is done, our green ball will be classified as either a signal or a background.

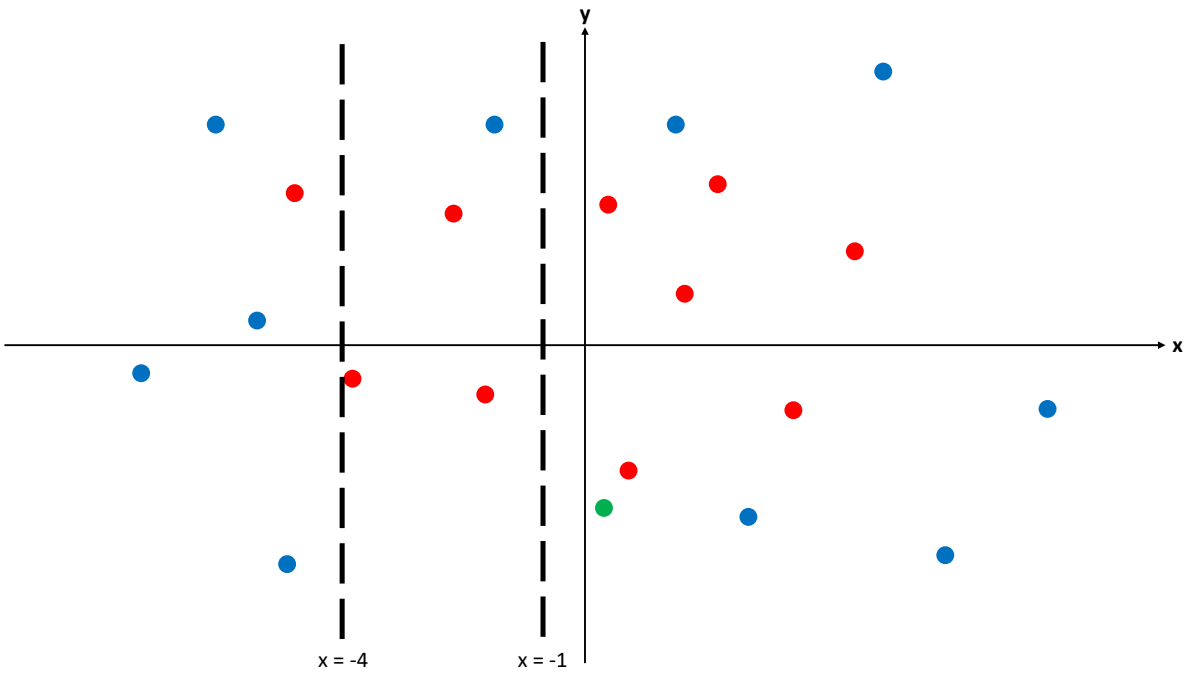


Figure 4.9: Illustration of a decision tree with a simple structure.

Boosted decision trees

The problem with using just a single decision tree, like in the example above, for doing a classification, or regression for that matter, is that a single decision tree has a tendency to overfit the data. This means that a decision tree is focusing on the noise in the data instead of the general behaviour. This results in poor performance in presence of new data.
This problem is solved by means of boosting. Boosting is a method that relies on using many weak learners, instead of just one strong learner, to perform the task at hand. A weak learner is just a simple decision tree with a small number of leaf nodes. Boosting must not be confused with another method, called bagging, also used to combine several decision trees into one strong learner. The difference between bagging and boosting lies in the way information from many decision trees are combined into the final decision. In algorithms based on bagging, such as the random forest algorithm, each tree is independent of the previous tree and the final decision is made by aggregating the predictions from all the trees.
The idea behind boosting lies in using the mistake of a previous tree to improve the prediction upon building another tree. Several boosting algorithms are available today, the most famous being AdaBost, Gradient boost and XGboost. *Gradient Boosted Decision Tree*, henceforth referred to as the *BDTG*, starts the classification training from some starting prediction. The residuals array, or simply errors array, is built next by calculating the prediction error with respect to each entry in the training set. This is calculated using some loss function. For a classification problem, this can be achieved using the logarithmic loss function, amongst others. These residuals go into the training of the second decision tree. In this way, the prediction error of the previous decision tree is passed on to the next decision tree. This process is continued either until the maximum number of decision trees is reached, or there are no improvements from adding additional trees. At each step, the previous tree is modified by the new one. How large modifications are is defined by the parameter called the learning rate. If the learning rate is too small the *BDTG* algorithm will need a lot of time to converge. On the other hand, a too large learning rate may result in jumping around the minimum of the loss function and never reaching it. The small learning rate can be compensated by increasing the maximum number of trees to be built. However, one must be careful because increasing the number of available trees also increases the probability of overtraining.

4.6.3 Algorithm setup for the signal extraction

The EWK signal extraction discussed in this chapter was based on two approaches:

1. a BDT classifier with gradient boosting (*BDTG*) that uses the first seven variables from the Table 4.7. The performance of these variables in separating the VBS contribution was presented in the previous study in this channel using 2016 data [105]. This approach is referred to as the *BDT7*
2. a BDTG classifier that uses all variables from Table 4.7. This is referred to as the *BDT28* and it was used to check the signal significance gain when using additional variables.

Regardless of the approach used, the setup for the classifier training was the same.
The first step was to prepare the data to be inserted into the *TMVA* tool for the training of the classifier. For this, the baseline selection was applied and the data was stored in root files. Together with the data passing the baseline selection, weights were stored as well for each event. These incorporate L1 prefiring probability as well as the MC and PU weights, trigger efficiency, luminosity, cross section, scale factors and K-factors for the qqZZ and ggZZ backgrounds. All weights were applied independently of the year with an exception of luminosity and L1 prefiring weights.

BDT classifier training

The EWK signal was trained only against the qqZZ background. Since the kinematics of the ggZZ events is rather similar to that of the qqZZ events, the gain of using it in the training would not be significant. Other backgrounds used in the analysis are minor and were not used in the training either. While using available background samples in the training would increase separation slightly, at the same time, it would reduce the robustness of the model. The result of the training, however, was applied to all samples.
The available signal and background data were equally split in the training set and the test set used to check the performance of the classifier on new data. The training and test samples were weighted, as previously discussed, in order to account for the difference in the distribution shapes between the different contributions. The hyperparameters used in the training are summarized in Table 4.11. It was checked that the training is stable under changes in hyperparameters.
After the training is completed, *TMVA* stores the result in "weight" files that are used to apply the training on the EWK signal and qqZZ, ggZZ, $t\bar{t}Z + VVZ$ and Z+X backgrounds. For each contribution, every event, correctly weighted, is passed through the BDT and its BDT score is evaluated. This is then used to produce a stacked BDT response histogram.
As a final step, the expected and the observed signal significance is calculated. This is done using the "*combine*" tool which performs a maximum likelihood fit to the data. The expected significance is calculated by assuming an Asimov data set [134] on top of the prediction. Distribution shapes and event yields for each contribution, together with systematic uncertainties, are provided in a file called the *datacard*. The systematic uncertainties are discussed in section 4.8. The "combine" tool also provides the user with an option to exclude systematic uncertainties when performing the fit.

parameter	value	parameter meaning
NTrees	1000	number of trees
MinNodeSize	2.5	minimum percentage of training events required in a leaf node
Shrinkage	0.1	learning rate
nCuts	20	number of grid points used in finding optimal cut in node splitting
maxDepth	2	maximum allowed depth of the decision tree

Table 4.11: Hyperparameters used in the training of the *BDT7* and *BDT28* classifiers.

4.6.4 Signal extraction using the BDT7

Distributions of input variables in the training of the BDT7 classifier are shown for the 2018 data-taking period in Figure 4.10. The same distributions for the 2016 and 2017 data-taking periods can be found in the Appendix A.

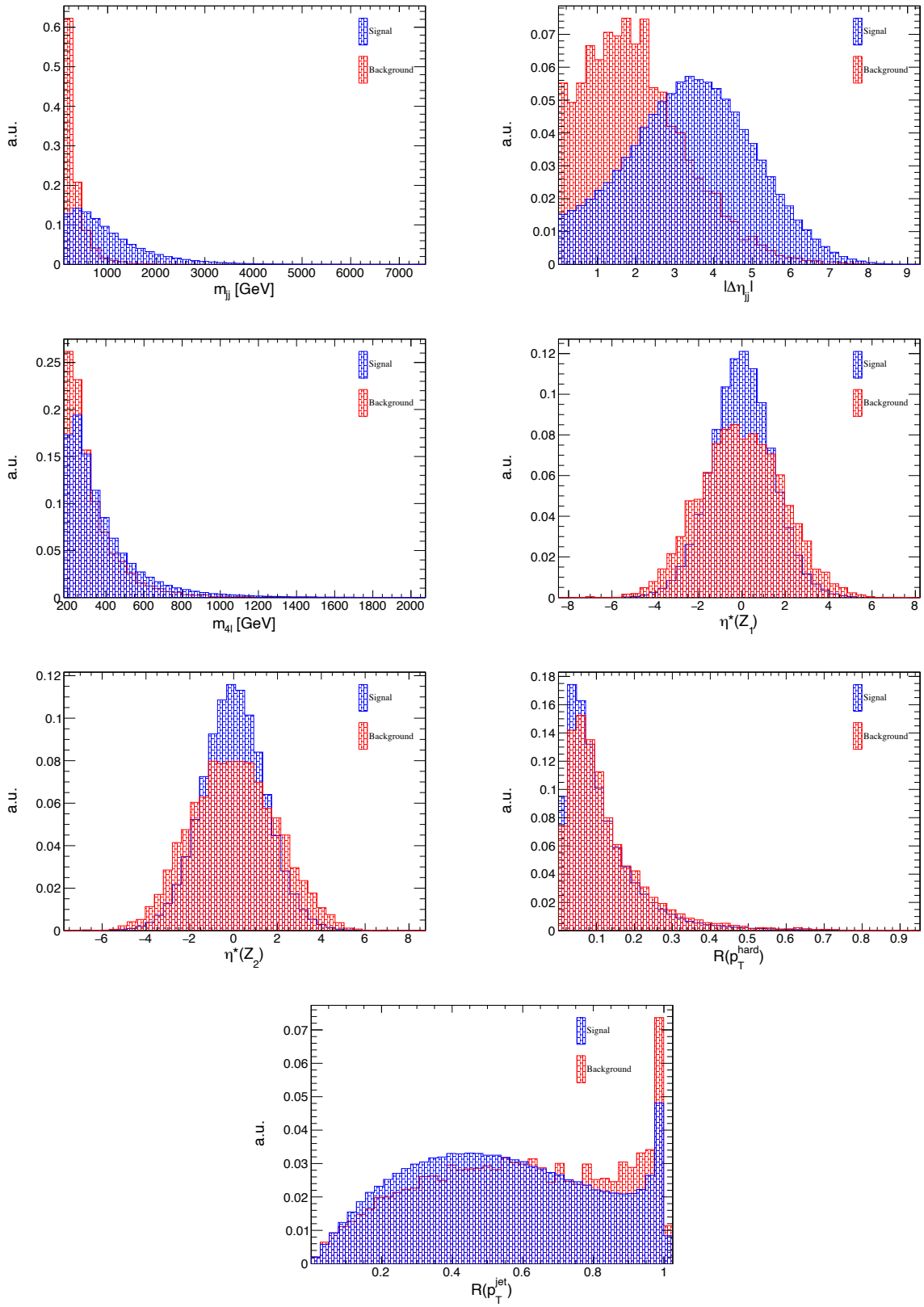


Figure 4.10: BDT input distributions for the EWK signal (in blue) and the qqZZ background (in red) to the *BDT7* training for the 2018 period.

The *BDT7* output distributions for the training and test samples, together with the overtraining check, for all three periods, are shown in Figure 4.11. Finally, Figure 4.12 shows the BDT response distribution for all three data-taking periods, and for the three periods combined, where each contribution is stacked on the previous ones. The agreement between data and the MC prediction is within the uncertainties, which are large in the right tail of the distribution due to limited statistics in that region.

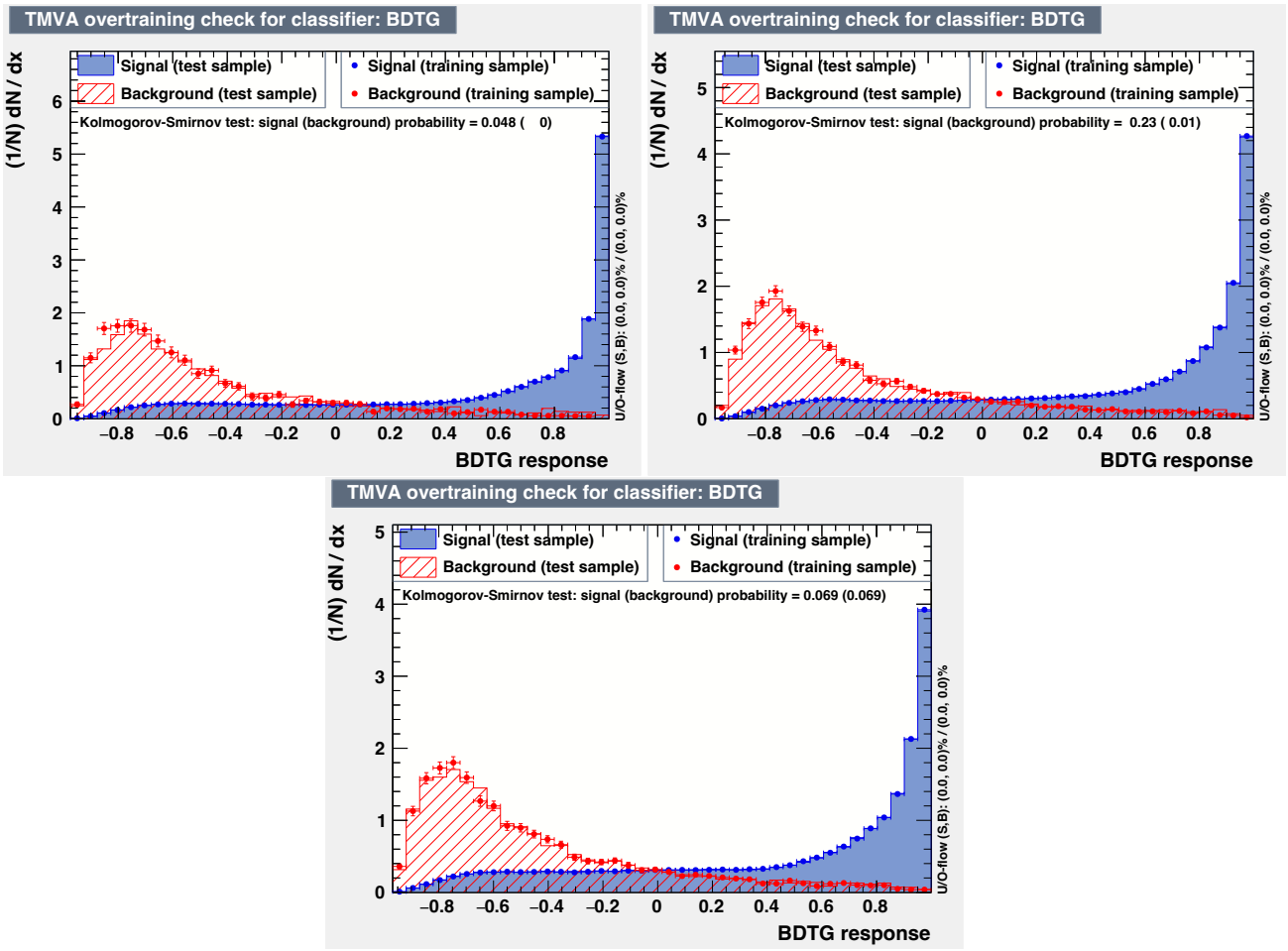


Figure 4.11: The BDT7 output distribution, together with the overtraining check for the 2016 (top-left), 2017 (top-right) and 2018 (bottom) data-taking periods.

4.6.5 Signal extraction using the BDT28

Input variables used in the training of the *BDT28* classifier are shown in Figure 4.13 for the 2018 data-taking period. The same distributions for the 2016 and 2017 data-taking periods can be found in the Appendix A.
Looking at the new variables introduced to the BDT28 training one can notice a great separation power of variables $y(j_1)$, $y(j_2)$, $\eta(j_1)$ and $\eta(j_2)$. However, these are correlated with the $\Delta\eta_{jj}$ variable already present in the BDT7. The same is true for the other jet variables as well. Variables $p_T(Z_1)$ and $p_T(Z_2)$ also show good separation power, but are correlated to m_{4l} and $R_{p_T}^{hard}$. For these reasons, one would not expect to gain a lot in terms of the BDT28 performance with respect to the BDT7.

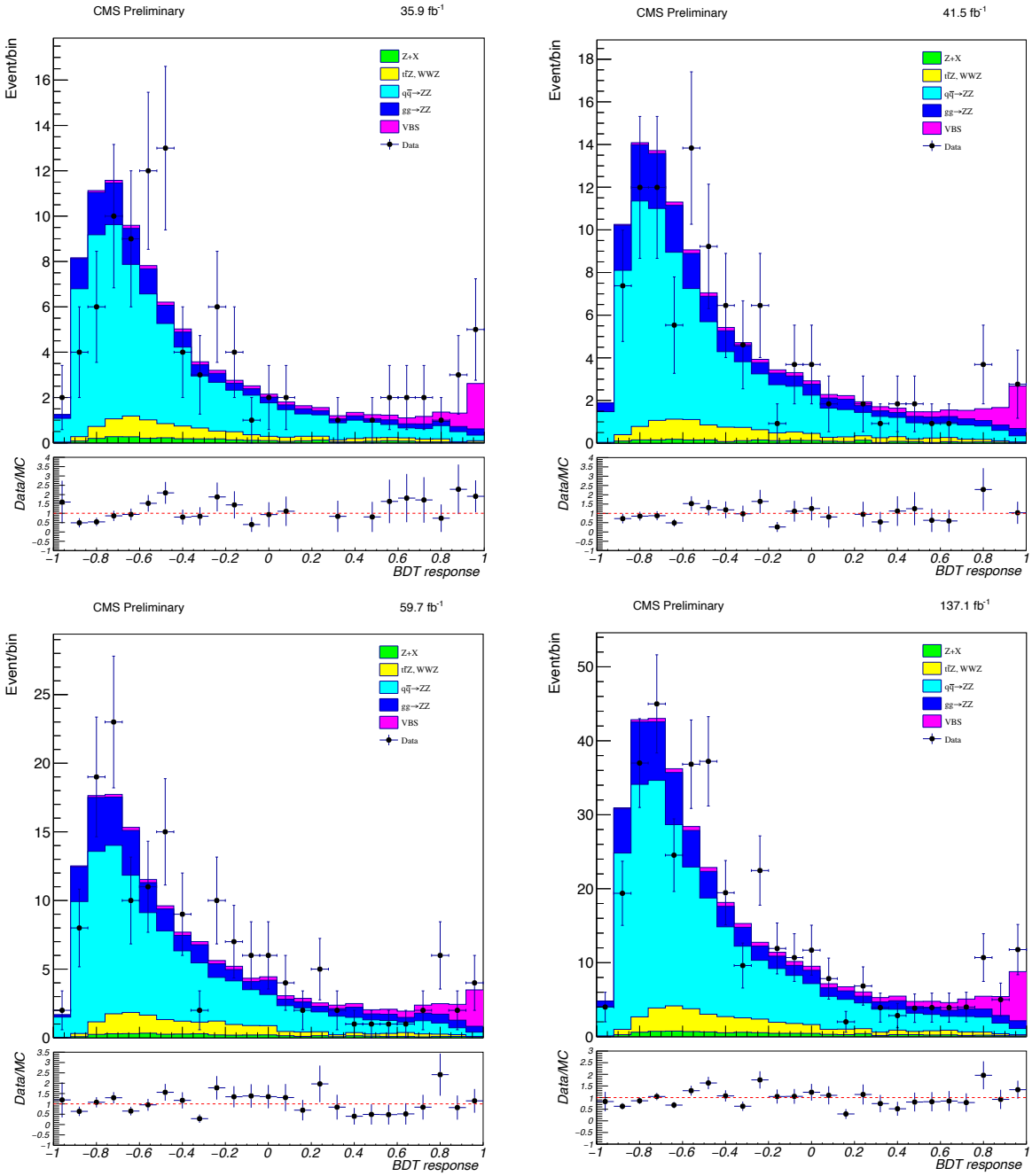


Figure 4.12: BDT output distribution for each contribution after the *BDT7* training for the 2016 (top-left), 2017 (top-right) and 2018 (bottom-left) period together with the period-combined distribution (bottom-right). Each contribution is stacked on top of the previous one starting with the Z+X sample. Bottom: comparison between data and MC expectation.

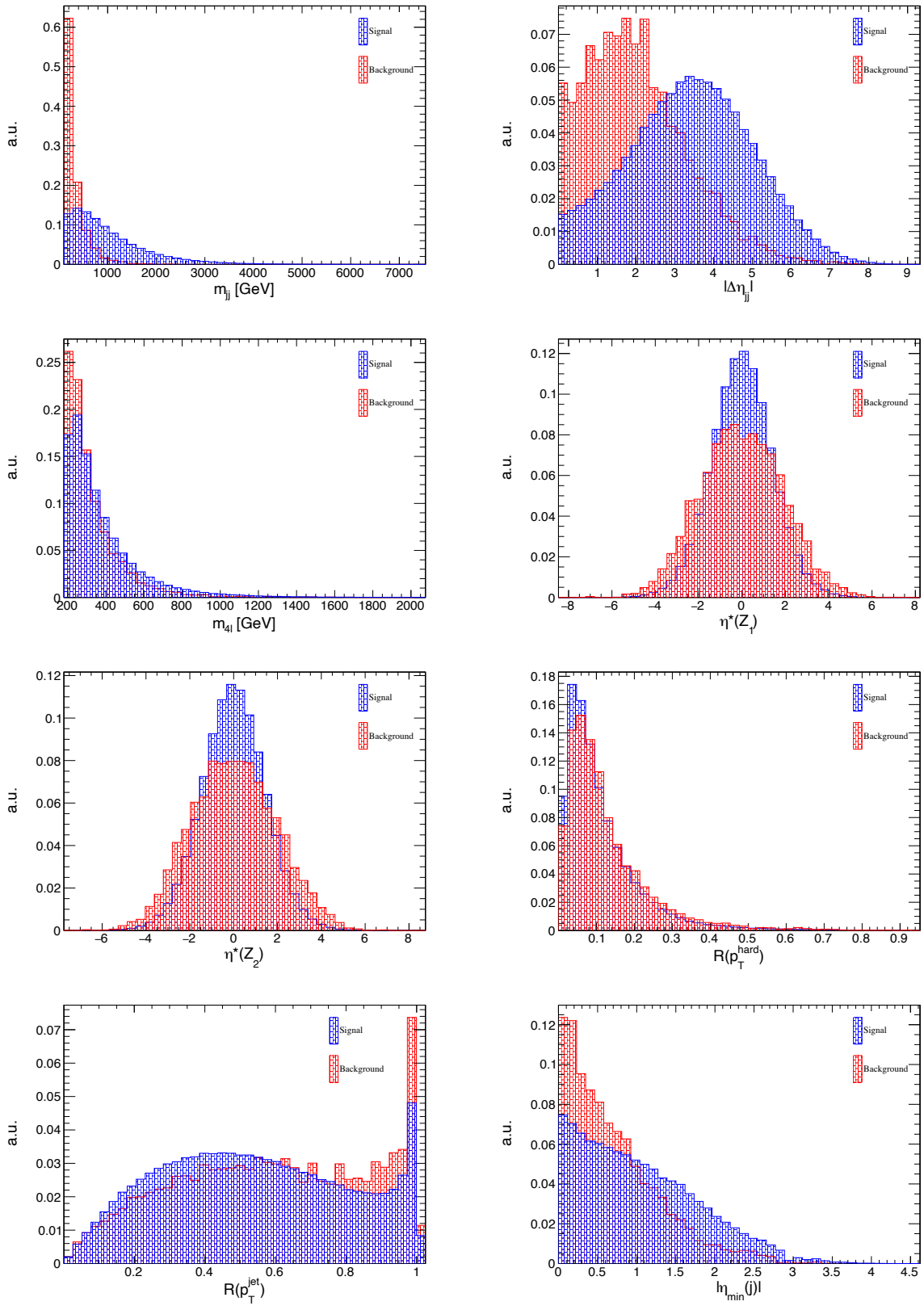

Signal
Background
a.u.
m_{jj} [GeV]
$|\Delta\eta_{jj}|$
m_{4l} [GeV]
$\eta^*(Z_1)$
$\eta^*(Z_2)$
$R(p_T^{hard})$
$R(p_T^{jet})$
$|\eta_{min}(j)|$

Figure 4.13

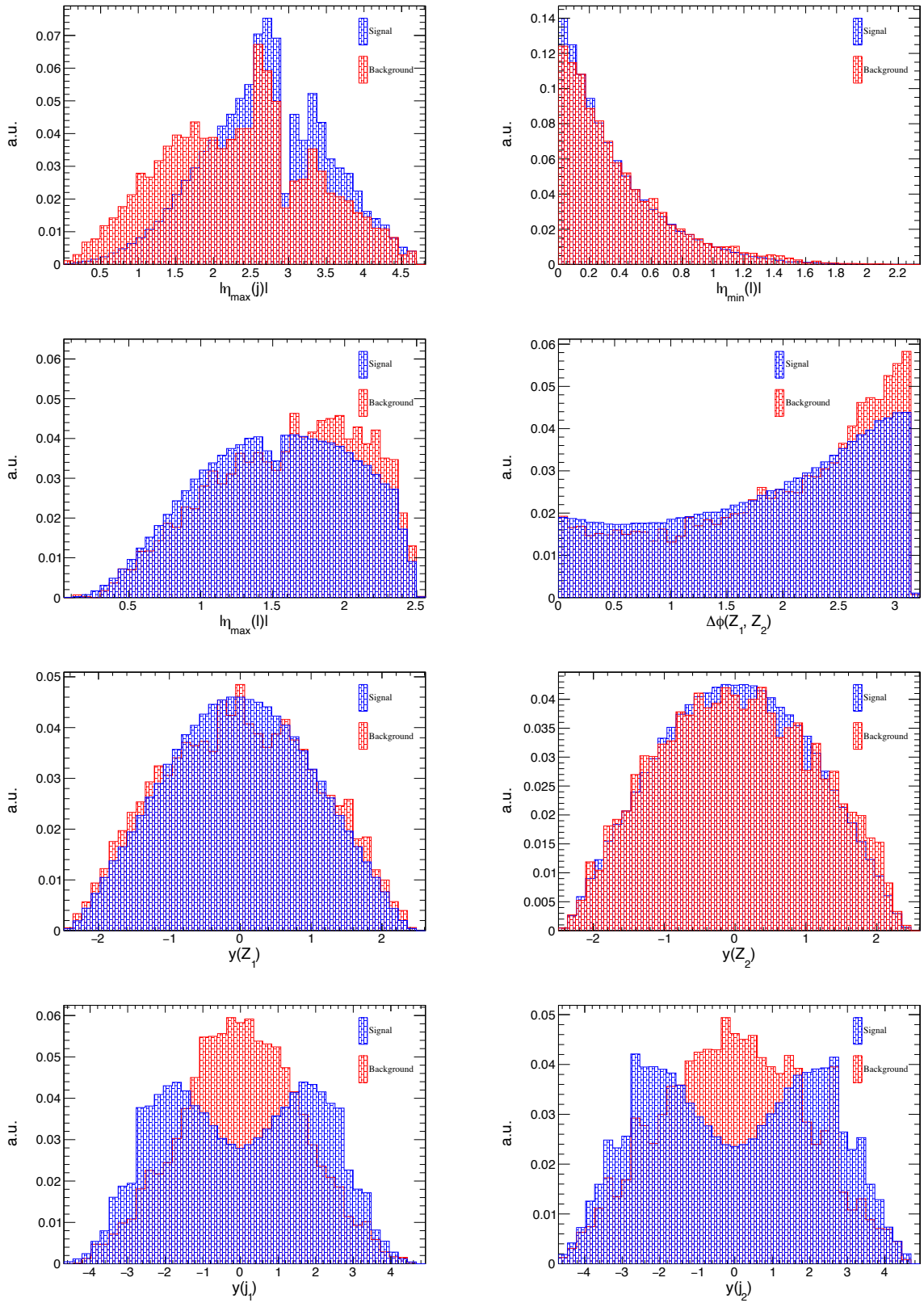

Signal
Background
a.u.
$|\eta_{max}(j)|$
$|\eta_{min}(l)|$
$|\eta_{max}(l)|$
$\Delta\phi(Z_1, Z_2)$
$y(Z_1)$
$y(Z_2)$
$y(j_1)$
$y(j_2)$

Figure 4.13

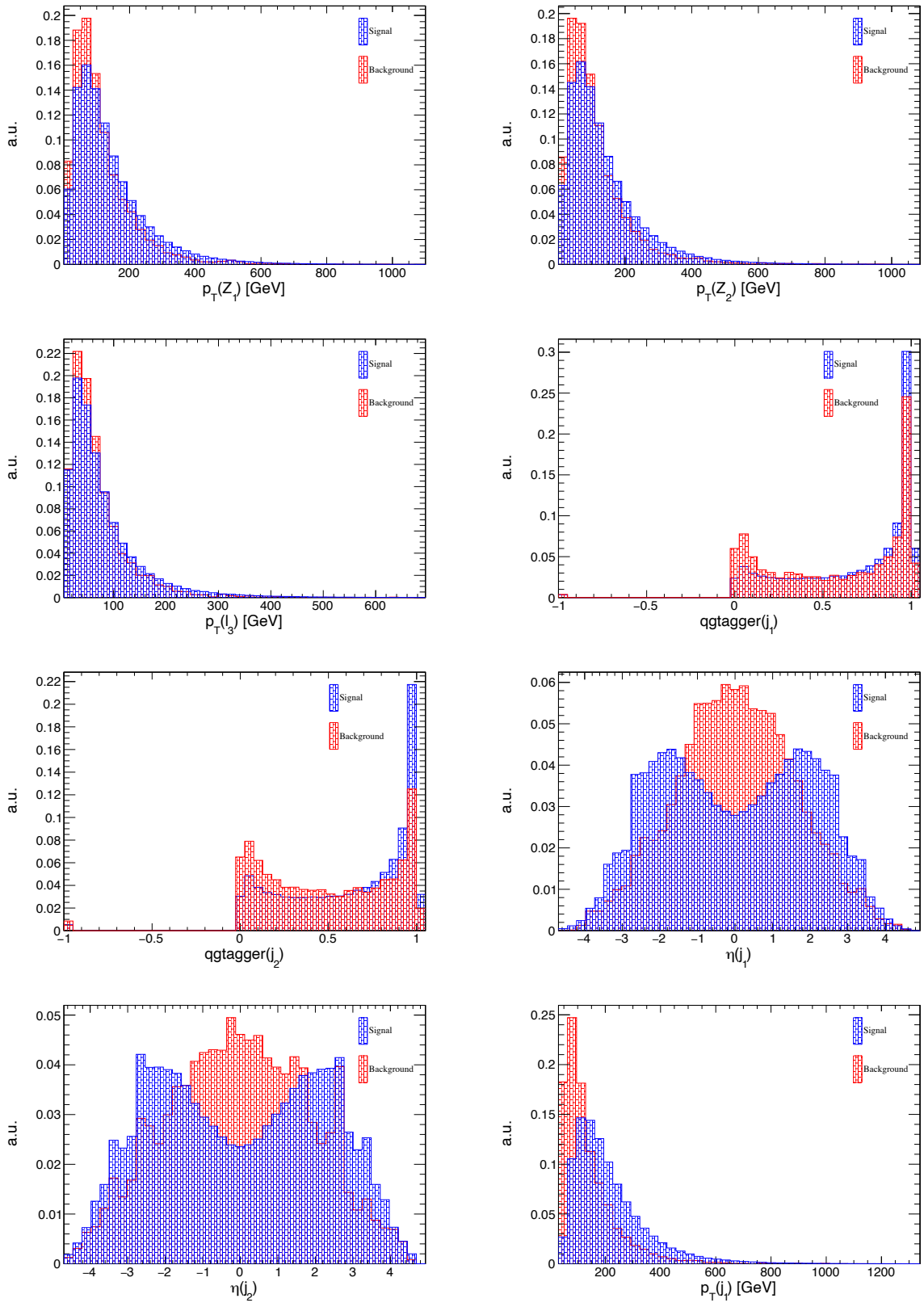

a.u.
Signal
Background
$p_T(Z_1)$ [GeV]
$p_T(Z_2)$ [GeV]
$p_T(l_3)$ [GeV]
qgtagger(j_1)
qgtagger(j_2)
$\eta(j_1)$
$\eta(j_2)$
$p_T(j_1)$ [GeV]

Figure 4.13

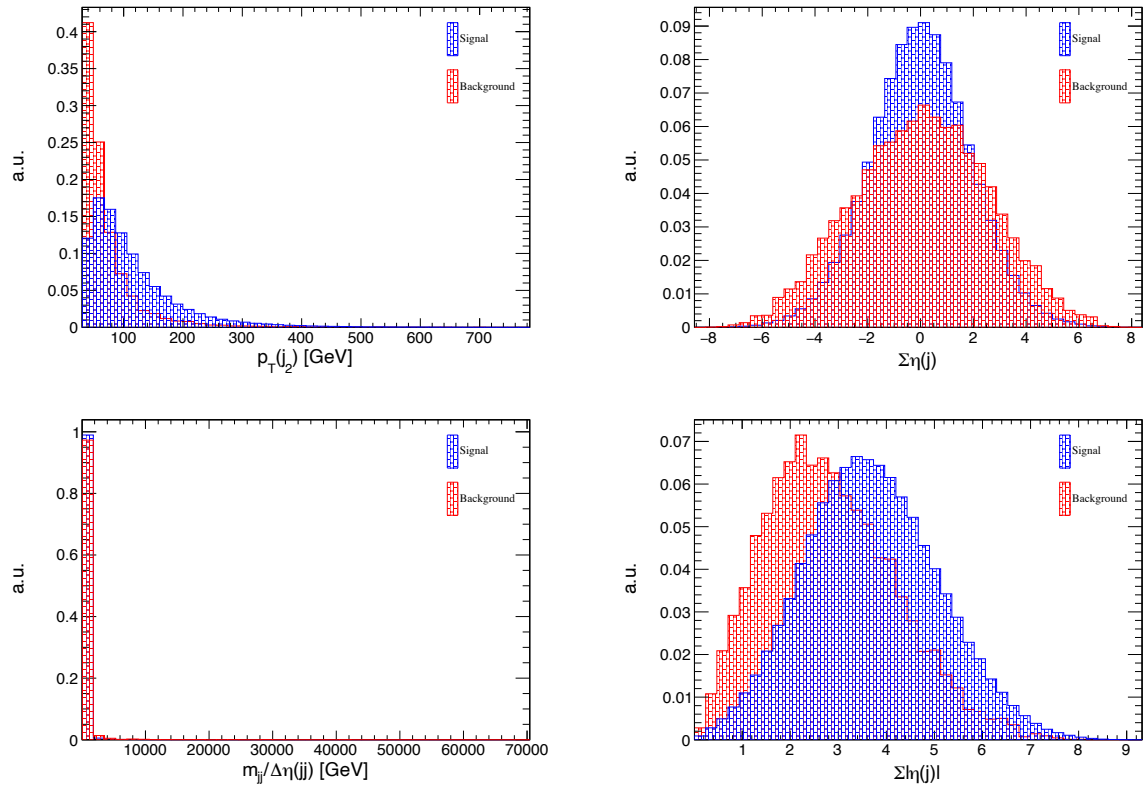


Figure 4.13: BDT input distributions for the EWK signal (in blue) and the qqZZ background (in red) to the *BDT28* training for the 2018 period.

The *BDT28* output distributions for the training and test samples, together with the overtraining check, for all three periods are shown in Figure 4.14. Finally, Figure 4.15 shows the BDT response distribution for all three data-taking periods, and for the three periods combined, where each contribution is stacked on the previous ones. The agreement between data and the MC prediction is within the uncertainties.

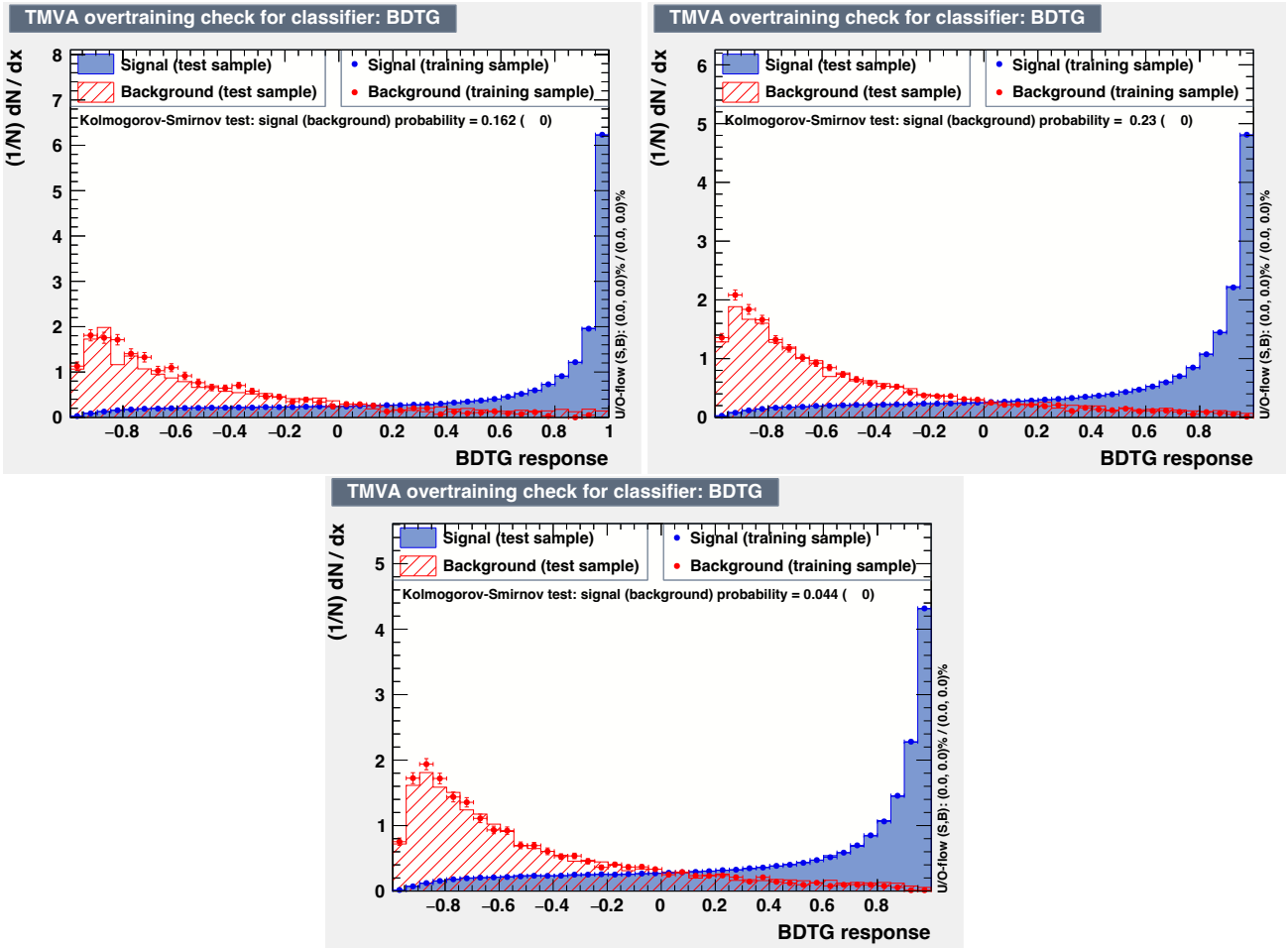


Figure 4.14: The BDT28 output distribution, together with the overtraining check for the 2016 (top-left), 2017 (top-right) and 2018 (bottom) data-taking periods.

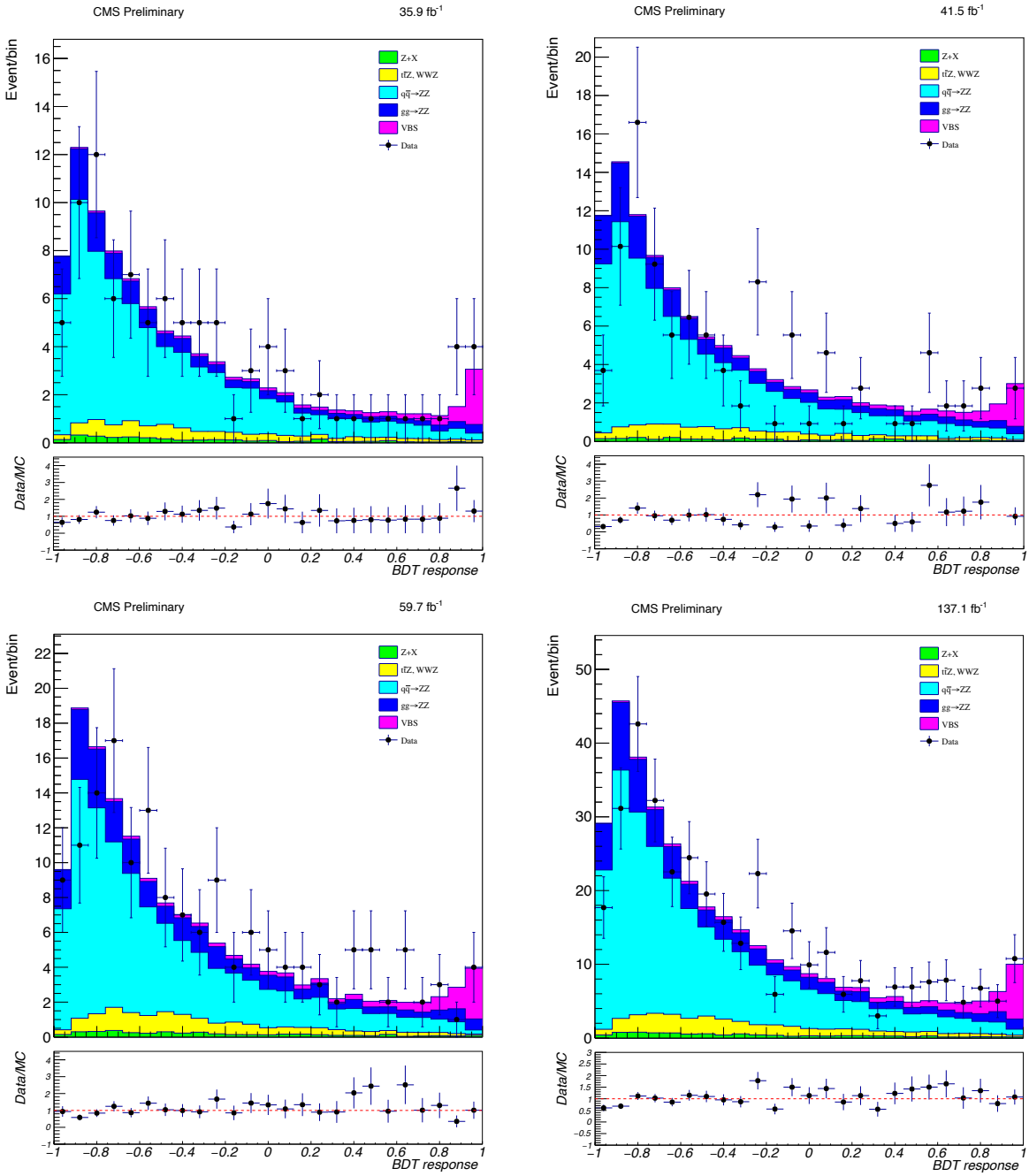


Figure 4.15: BDT output distribution for each contribution after the *BDT28* training for the 2016 (top-left), 2017 (top-right) and 2018 (bottom-left) period together with the period-combined distribution (bottom-right). Each contribution is stacked on top of the previous one starting with the Z+X sample. Bottom: comparison between data and MC expectation.

4.7 Setting limits on anomalous quartic gauge couplings

Limits on anomalous quartic gauge couplings (aQGCs) are derived in the effective field theory framework where the 8-dimensional operators originate from the covariant derivatives of the Higgs doublet and the charged and neutral field strength tensors. The latter generates eight independent operators which correspond to the couplings of the transverse degrees of freedom, T_i, of the gauge fields [110].
The ZZjj channel exploited in this analysis is particularly sensitive to the neutral-current operators T_8 and T_9 as well as the charged-current operators T_0, T_1 and T_2 [15] which enhance the production cross section at large values of m_{ZZ}.
Limits on the aQGC parameters f_{T_i}, corresponding to the Wilson coefficients of the operators, are derived based on the m_{4l} distribution following the previous analysis of the anomalous couplings in this channel [105]. The reason for choosing the m_{4l} distribution lies in the fact that the m_{4l} is Lorentz invariant and thus less sensitive to the higher-order corrections. This is crucial since the effect is dominant in the far tail of the distribution.

A dedicated MG sample was produced for the aQGC analysis:

$$generate\ p\ p > z\ z\ j\ j\ QED = 4\ QCD = 0\ NP = 1$$

The reweighting functionality of the *MG5* was used to obtain the expected distributions for different values of the couplings without needing to produce additional samples. The method uses event weights, w_{new}, to reweigh the nominal event sample to the alternative hypotheses of the coupling strength:

$$w_{new} = w_{old} \cdot \frac{|\mathcal{M}_{new}|^2}{|\mathcal{M}_{old}|^2}$$

where $\mathcal{M}_{new}$ and $\mathcal{M}_{old}$ are matrix elements with the modified coupling strength and the nominal matrix element respectively. The ratio of the aQGC to SM yields was calculated for several discrete coupling values and then fit with a quadratic function. The result is a semi-analytic description of the expected m_{ZZ} distribution for every bin as a function of the aQGC couplings. This is shown for the operator $T8$ in Figure 4.16 for the last four bins of the m_{4l} distribution. The overflow is included in the last bin. It can be seen that the effect on yields is rising towards the tail of the distribution. The same plots, corresponding to the last bin of the m_{4l} distribution for the $T0$, $T1$, $T2$ and $T9$ operators, are shown in Figure 4.17.

Figure 4.18 shows the expected m_{4l} distribution for the SM, with nuisance parameters set to their fitted values, and the expected distribution for one aQGC scenario, as well as the observed distribution. The fit was performed in the same way as for the EWK signal significance calculation, i.e., using the "*combine*" tool. The test statistic is the log-likelihood ratio with all systematic uncertainties profiled as nuisance parameters [135].
The 95% confidence level (CL) intervals were determined using Wilk's theorem assuming that the likelihood approaches the χ^2-distribution with one degree of freedom.
The expected limits were obtained using the pre-fit yields for the background and the EWK signal. The observed limits for the combined data set, setting the other coupling to zero, were obtained using the post-fit yields for the background and the signal expectations.
Finally, the unitarity limits were calculated using both the *VBFNLO* package [136] and a theoretical approach as suggested recently [137].

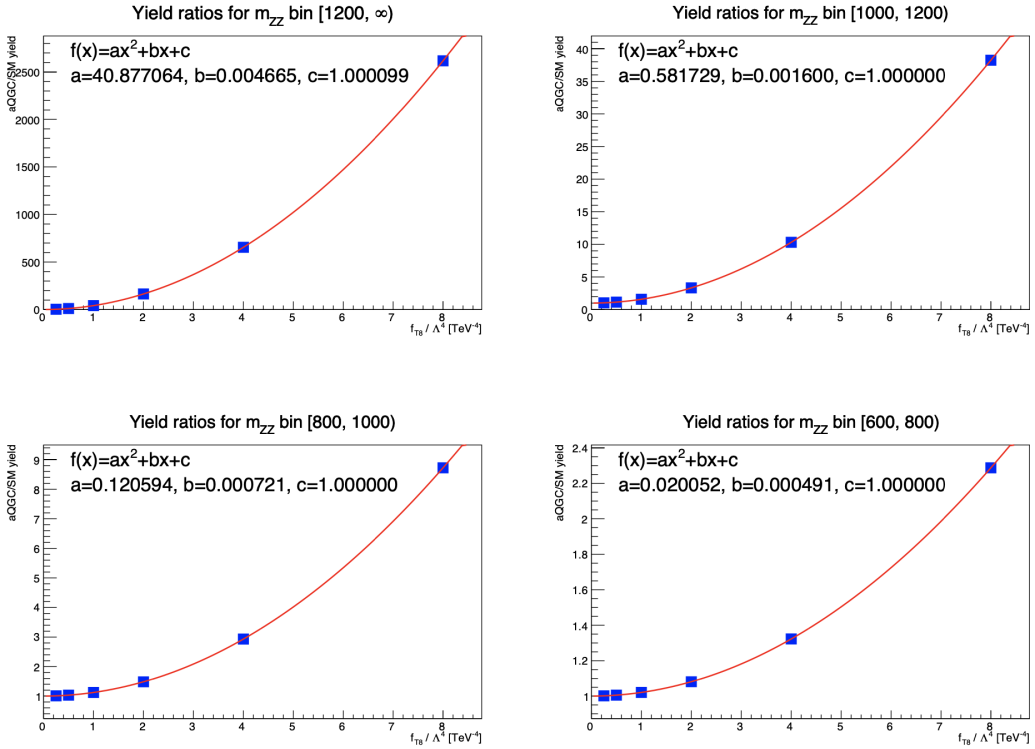


Figure 4.16: Yield ratios for a few values of the operator couplings, f_{T8}/Λ^4, obtained from the reweighing and the fitted quadratic interpolation for the most relevant mass bins used in the statistical analysis.

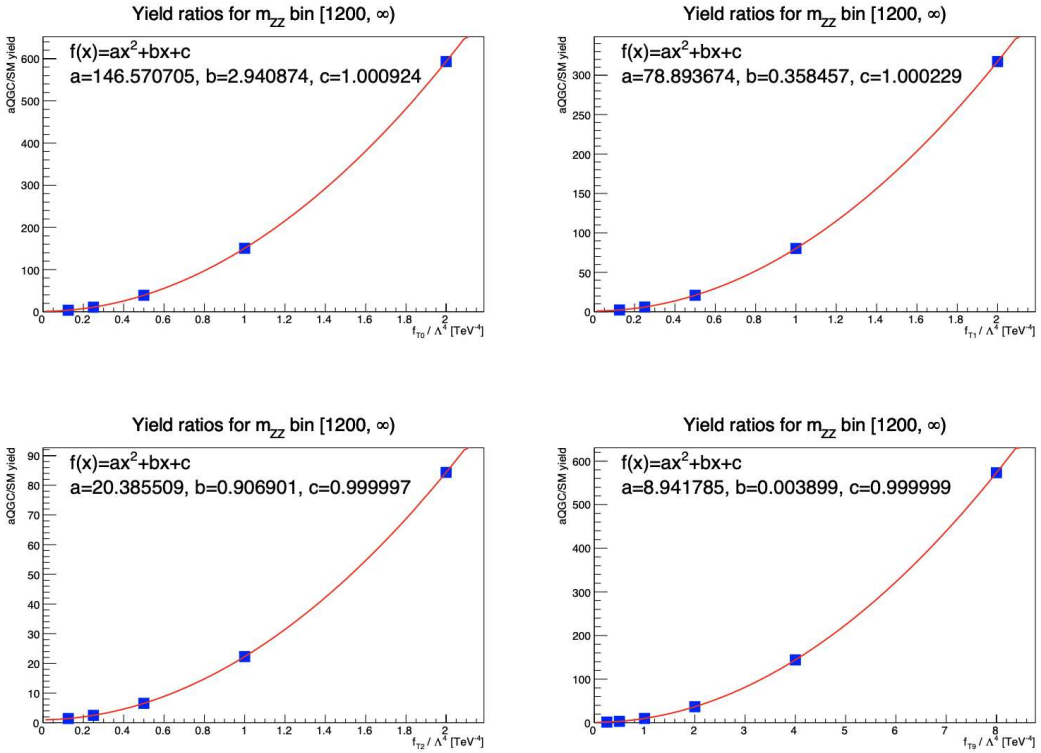


Figure 4.17: Yield ratios for a few values of the operator couplings obtained from the reweighing and the fitted quadratic interpolation for each of the mass bins used in the statistical analysis. The last bin of the m_{4l} distribution is shown for the f_{T0}/Λ^4 (top left), f_{T1}/Λ^4 (top right), f_{T2}/Λ^4 (bottom left) and f_{T9}/Λ^4 (bottom right) operators.

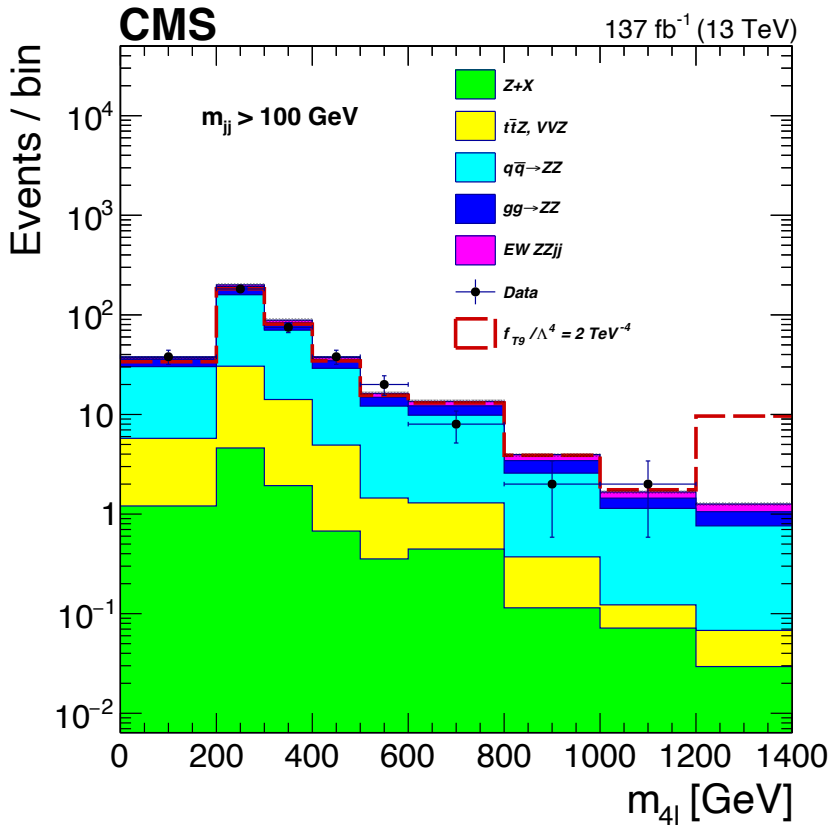


Figure 4.18: Postfit distributions of the four-lepton invariant mass for events satisfying the ZZjj inclusive selection. Points represent the data, filled histograms the fitted signal and background contributions, and the grey band the uncertainties derived from the fit covariance matrix. As an example, the expected distribution for $f_{T9}/\Lambda^4 = 2\ TeV^{-4}$ is also shown.

4.8 Systematic uncertainties

Regardless of the signal extraction method used, the MELA discriminant or the BDTs discussed in section 4.6, the same set of systematic uncertainties is applied. QCD scale and PDF uncertainties are originating from the incomplete theoretical description of the underlying physics. The rest described below come from an imperfect description of the detector effects or simulation.

When calculating the cross section of the desired process, one can find that, sometimes, it can't be explicitly done due to the divergences that appear. The nature of such infinities can be twofold:

1. ultraviolet (UV) divergences which arise due to the large momentum transfers in a loop of Feynman diagrams representing the process amplitude
2. infrared (IR) divergences that can arise either because a massless particle radiates another massless particle or because a virtual or real particle reaches zero momentum

In order to solve the UV divergences, the renormalization scale, μ_R, is introduced. Consequently, the running coupling constant, α_s, becomes a function of parameter μ_R.
If the IR divergences appear because of a massless particle being radiated by another massless particle, they can

be cured by introducing a factorization scale, μ_F. Consequently, the parton distribution functions (PDFs) and the fragmentation functions, which defines the evolution of the collision fragments, become a function of μ_F [138].
QCD scale uncertainties are estimated using the common procedure of varying the normalization and factorization scales up and down by a factor of two (excluding the extreme cases) with respect to the nominal value. Unlike for the EWK signal where the uncertainty is shape-dependent, a constant uncertainty, between 9% and 14%, is used for qqZZ and ggZZ backgrounds.

Uncertainties related to the choice of the PDFs and the strong coupling constant α_s are evaluated from the variations of the respective eigenvalues set [118]. Although different PDFs were used for different data-taking periods, the associated uncertainties are very similar. A constant uncertainty, between 3.3% and 6.6%, was used for different samples [110, 112].

The uncertainty in the LHC integrated luminosity is taken from [139] and is 2.3-2.5%. Since the correlated component amongst years is small, and because the overall effect of systematic uncertainties in the measurements is also small, the uncertainty in the luminosity between the years is assumed to be uncorrelated.

The uncertainty in the data-driven reducible background estimate is dominated by the statistical uncertainties because of the limited number of events in the control regions and ranges from 33% to 45% depending on the final state.

Processes estimated from the simulation are limited by the statistics of the MC sample. This is taken as a source of the shape-dependent, year-uncorrelated systematic uncertainty. For the cut-and-count analyses (i.e. calculation of the EWK and EWK+QCD cross sections, and a derivation of the limits on the aQGCs) integrated uncertainties of the MC sample were used, while for the template analysis (i.e. signal extraction using MELA) the $autoMCstats$ feature of the "combine" tool was used to obtain the shape-dependent uncertainty profile. For the calculation of limits on aQGCs, the uncertainties were enlarged because the sensitivity comes from the high-m_{ZZ} bins only.

Uncertainties coming from the trigger and lepton reconstruction and selection range from 2.5% to 9% depending on the final state and those coming from the PU reweighting range between 0.2% and 2.7% depending on the sample and year [140].

The jet energy scale (JES) uncertainty ranges from 4.9 - 11.4% for the QCD qqZZ background and 0.7% - 1.2% for the EWK signal. The jet energy resolution (JER) ranges from 2.2% - 6.3% and 0.2% - 0.4% for QCD qqZZ background and EWK signal respectively [21].

L1 prefiring weight variations range from 0.6% to 3.0% depending on the sample.

Systematic uncertainties are summarized in Table 4.12.

Systematic source	qqZZ	ggZZ	VBS	Z+X	Shape	Years correlated
QCD scales [%]	10 - 12	9 - 14	6	-	+	+
PDF + α_s [%]	3.2	5	6.6	-		+
Lepton trigger, reco, sel. [%]	2.5 - 9	2.5 - 9	2.5 - 9	-		+
L1 prefiring [%]	0.6 - 1.0	0.6	1.8 - 3.0	-	+	
Luminosity [%]	2.3 - 2.5	2.3 - 2.5	2.3 - 2.5	-		
JES [%]	4.9 - 11.4	3.6 - 10.2	0.7 - 1.2	-	+	
JER [%]	2.2 - 6.3	1.0 - 2.2	0.2 - 0.4	-		
MC samples [%]	2.5-4.2 (11-28)	3.2 (17-22)	$\ll 1$	-	+	
Pileup [%]	0.2 - 2.6	0.4 - 2.7	0.3 - 1.7	-		
Reducible background [%]	-	-	-	33 - 45		

Table 4.12: Systematic uncertainties on the signal and background yields. Minor backgrounds, for which the systematics is dominated by the MC sample size (19% - 24%), are not shown. The numbers in parentheses refer to the uncertainties used in the derivation of limits on aQGCs.

4.9 Results

Table 4.13 shows the expected and observed event yields for the ZZjj inclusive selection as well as the two VBS-enriched regions. A good agreement between the predicted and measured event yields is reported for all data-taking periods.
Measured cross sections and the corresponding SM predictions in the three fiducial regions obtained using the MELA discriminant for both EWK and EWK+QCD are summarized in Table 4.14. The same table shows the measured and expected EWK signal strength. The total uncertainty is quoted for all the measurements with statistical only separated in parentheses. SM predictions were extracted from the generated events in the MC samples used in the analysis including the K-factors where applicable. For the EWK ZZjj inclusive region, in addition to the higher-order calculations at NLO in QCD [141, 142] and theoretical predictions at LO in QCD, NLO EWK corrections [143] were included. Uncertainties in all SM predictions come from variations of the factorization and renormalization scales. $PDF + \alpha_s$ variation uncertainties are summed in quadrature, except from the prediction from [143].

Year	EWK signal	Z+X	$q\bar{q} \to ZZjj$	$gg \to ZZjj$	$t\bar{t}Z + VVZ$	Tot. predict.	Data
ZZjj inclusive							
2016 ($35.9\ fb^{-1}$)	6.3 ± 0.07	2.8 ± 1.1	65.6 ± 9.5	13.5 ± 2.0	8.4 ± 2.2	96 ± 13	95
2017 ($41.5\ fb^{-1}$)	7.4 ± 0.8	2.4 ± 0.9	77.7 ± 11.2	20.3 ± 3.0	9.6 ± 2.5	117 ± 15	111
2018 ($59.7\ fb^{-1}$)	10.4 ± 1.1	4.1 ± 1.6	98.1 ± 14.2	29.1 ± 4.3	14.2 ± 3.8	156 ± 20	159
All ($137.1\ fb^{-1}$)	24.1 ± 2.5	9.4 ± 3.6	241.5 ± 34.9	62.9 ± 9.3	32.2 ± 8.5	370 ± 48	365
VBS signal-enriched (loose)							
2016 ($35.9\ fb^{-1}$)	4.2 ± 0.4	0.4 ± 0.2	9.7 ± 1.4	3.2 ± 0.5	1.1 ± 0.3	18.7 ± 2.3	21
2017 ($41.5\ fb^{-1}$)	4.9 ± 0.5	0.5 ± 0.2	13.5 ± 1.9	5.5 ± 0.8	1.2 ± 0.3	25.5 ± 3.1	17
2018 ($59.7\ fb^{-1}$)	6.9 ± 0.7	0.8 ± 0.3	14.9 ± 2.2	8.3 ± 1.2	1.7 ± 0.5	32.6 ± 3.9	30
All ($137.1\ fb^{-1}$)	16.0 ± 1.7	1.6 ± 0.6	38.1 ± 5.5	17.0 ± 2.5	4.1 ± 1.1	76.8 ± 9.3	68
VBS signal-enriched (tight)							
2016 ($35.9\ fb^{-1}$)	2.4 ± 0.3	0.10 ± 0.04	1.3 ± 0.2	0.7 ± 0.1	0.24 ± 0.06	4.8 ± 0.5	4
2017 ($41.5\ fb^{-1}$)	2.7 ± 0.3	0.05 ± 0.02	1.9 ± 0.3	1.2 ± 0.2	0.14 ± 0.04	6.0 ± 0.7	3
2018 ($59.7\ fb^{-1}$)	3.9 ± 0.4	0.17 ± 0.06	2.0 ± 0.3	1.5 ± 0.2	0.30 ± 0.08	7.8 ± 0.9	10
All ($137.1\ fb^{-1}$)	9.0 ± 1.0	0.32 ± 0.12	5.3 ± 0.8	3.3 ± 0.5	0.68 ± 0.18	18.6 ± 2.1	17

Table 4.13: Predicted signal and background yields with total uncertainties, and the observed number of events for the ZZjj inclusive selection as well as the VBS loose and tight signal-enriched selections. Integrated luminosities per data set are reported in parentheses.

	SM σ $[fb]$	**Measured σ $[fb]$**	μ_{exp}	μ_{obs}
ZZjj inclusive				
EWK	LO: $0.275 \pm 0.021_{th.}$ NLO QCD: $0.278 \pm 0.017_{th.}$ NLO EWK: $0.242^{+0.015_{th.}}_{-0.013_{th.}}$	$0.33^{+0.11\ (+0.04)}_{-0.10\ (-0.03)}$	$1.00^{+0.43\ (+0.39)}_{-0.36\ (-0.34)}$	$1.21^{+0.47}_{-0.40}$
EWK+QCD	$5.35 \pm 0.51_{th.}$	$5.29^{+0.31\ (+0.46)}_{-0.30\ (-0.46)}$	$1.00^{+0.13\ (+0.06)}_{-0.12\ (-0.06)}$	$0.99^{+0.13}_{-0.12}$
VBS signal-enriched (loose)				
EWK	LO: $0.186 \pm 0.015_{th.}$ NLO QCD: $0.197 \pm 0.013_{th.}$	$0.200^{+0.078\ (+0.023)}_{-0.067\ (-0.013)}$	$1.00^{+0.45\ (+0.40)}_{-0.38\ (-0.35)}$	$1.08^{+0.47}_{-0.38}$
EWK+QCD	$1.21 \pm 0.09_{th.}$	$1.00^{+0.12\ (+0.06)}_{-0.11\ (-0.05)}$	$1.00^{+0.16\ (+0.13)}_{-0.15\ (-0.12)}$	$0.83^{+0.15}_{-0.13}$
VBS signal-enriched (tight)				
EWK	LO: $0.104 \pm 0.008_{th.}$ NLO QCD: $0.108 \pm 0.007_{th.}$	$0.09^{+0.04\ (+0.02)}_{-0.03\ (-0.02)}$	$1.00^{+0.52\ (+0.50)}_{-0.44\ (-0.41)}$	$0.87^{+0.48}_{-0.39}$
EWK+QCD	$0.221 \pm 0.014_{th.}$	$0.20^{+0.05\ (+0.02)}_{-0.04\ (-0.02)}$	$1.00^{+0.42\ (+0.40)}_{-0.34\ (-0.32)}$	$0.92^{+0.39}_{-0.32}$

Table 4.14: SM cross sections in the three fiducial regions together with the fitted value of the signal strength. The total uncertainty is quoted for all measurements with the statistical only contribution in parentheses. The theory uncertainty for the expected SM cross section is also quoted. For the EWK ZZjj inclusive region, NLO EWK corrections [143] are quoted in addition to the higher-order calculations at NLO in QCD [141, 142] and theoretical predictions at LO in QCD.

The significance of the EWK signal using the MELA classifier was obtained by calculating the probability of the background-only hypothesis (*p*-value) as the tail integral of the test statistic evaluated at $\mu_{EWK} = 0$ under the asymptotic approximation [144]. The background-only hypothesis was excluded with $4.0\ \sigma$ ($3.5\ \sigma$ expected).

The expected significance using the two BDTs was calculated for the separated data-taking periods as well as for the combined period. The results are summarized in Table 4.15. An expected significance of $3.9\ \sigma$ (stat. only) and $3.8\ \sigma$ (stat. + sys.) is reported for the combined period using the *BDT7*. The value of $3.8\ \sigma$ obtained using the *BDT7* classifier is comparable to the MELA result. Similar performance of the *BDT7* and MELA is also confirmed by comparing the ROC curves in Figure 4.19.

In order to assess the potential gain of using 28 variables in the BDT training, the EWK signal significance was calculated for the *BDT28* as well. For the *BDT28*, an expected significance of $4.0\ \sigma$ (stat. only) and $3.9\ \sigma$ (stat. + sys.) is reported for the combined period. A small increase in sensitivity is obtained at the expense of a loss of model robustness. This shows that the *BDT7* is capable of capturing and exploiting the kinematical difference between signal and background without the need for additional variables.
The observed signal significance for the three periods, as well as for the combined period, for the *BDT7* and *BDT28* is also reported in Table 4.15. An upward fluctuation in the data can be seen from the bottom right plots in Figure 4.12 and Figure 4.15. This is reflected in the increase of observed signal significance for the combined period in both *BDT7* and *BDT28*.

A possible gain in sensitivity was also looked for by using the training events that passed the VBS loose selection instead of the ZZjj baseline. This was done for the 2016 data-taking period with the *BDT7* training and a negligible increase (< 0.4%) in the signal sensitivity was observed while, at the same time, losing some signal events.

Year	Exp. significance [σ]	Obs. significance [σ]
BDT7		
2016 ($35.9\ fb^{-1}$)	2.07 (2.12)	4.08 (4.05)
2017 ($41.5\ fb^{-1}$)	2.8 (2.14)	1.79 (1.69)
2017 ($59.7\ fb^{-1}$)	2.44 (2.53)	2.90 (3.12)
All ($137.1\ fb^{-1}$)	3.77 (3.93)	5.09 (5.19)
BDT28		
2016 ($35.9\ fb^{-1}$)	2.13 (2.18)	3.24 (3.29)
2017 ($41.5\ fb^{-1}$)	2.14 (2.20)	2.02 (1.91)
2017 ($59.7\ fb^{-1}$)	2.46 (2.55)	2.85 (3.08)
All ($137.1\ fb^{-1}$)	3.85 (4.01)	4.69 (4.81)

Table 4.15: Expected and observed EWK signal significance for the three data-taking periods as well as the combined period. Both results for BDT7 and BDT28 are reported. The results with only statistical uncertainties are shown in parentheses.

The expected and observed lower and upper 95% CL limits on the couplings of the charged-current operators T_0, T_1 and T_2 as well as of the neutral-current operators T_8 and T_9 are shown in Table 4.16. Results with only statistical uncertainties included are shown in parentheses. The unitarity limits obtained using both the *VBFNLO* package and the approach suggested in the recent publication [137] are also shown. These were the most stringent limits, at the time, on the neutral-current operators T8 and T9. A recent study by CMS collaboration in the $Z\gamma$ channel provided slight improvements on the measurement of the operator T_9 [145].

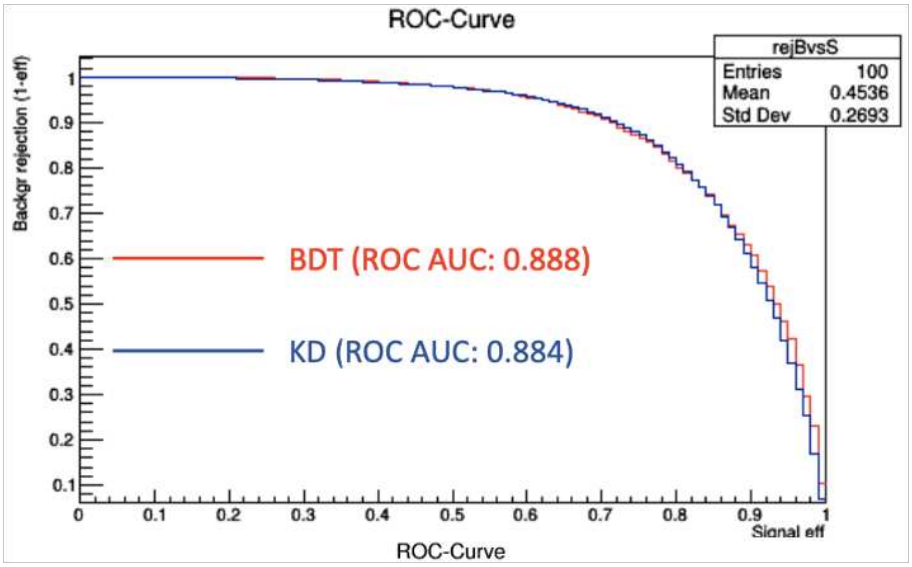


Figure 4.19: Performance of the *BDT7* compared to the MELA using the ROC curve and area under curve (AUC).

Coupling	Exp. lower	Exp. upper	Obs. lower	Obs. upper	Unit. limit (VBFNLO)	Unit. limit (Eboli)
f_{T_0}/Λ^4	-0.37	0.35	-0.24	0.22	2.9	2.4
f_{T_1}/Λ^4	-0.49	0.49	-0.31	0.31	2.7	2.6
f_{T_2}/Λ^4	-0.98	0.95	-0.63	0.59	2.8	2.5
f_{T_8}/Λ^4	-0.68	0.68	-0.43	0.43	1.8	1.8
f_{T_9}/Λ^4	-1.46	1.46	-0.92	0.92	1.8	1.8

Table 4.16: Observed and expected lower and upper 95 % CL limits on the coupling of the quartic tensor operators T_0, T_1 and T_2 as well as the neutral-current operators T_8 and T_9. The unitarity limits are also reported. All couplings are expressed in TeV^{-4} while the unitarity limits are expressed in TeV. Results are obtained using the postfit distributions.

4.10 Summary

A search for VBS in the $qq \to ZZ \to 4ljj$ channel using CMS data from the full Run 2 was presented in this chapter. Because of the fully reconstructable final state, this channel is expected to be amongst the most sensitive for the extraction of the longitudinal component of the ZZ scattering, thus providing a better insight into the scalar sector of the SM in the future. Since the channel is sensitive to the neutral-current operators, it enables to probe into the anomalous quartic gauge coupling phenomena and provides a tool for the exploration of physics the beyond the SM. In order to provide the best possible description of signal and background processes, special care was given to the MC simulations. This is especially true for the QCD loop-induced background which was simulated using *MG5* with up to two hadronic jets modelled at the matrix element and matched to the parton shower using the MLM matching scheme for the first time.

The Matrix Element Likelihood Approach (MELA) discriminant was used to measure the electroweak (EWK) and EWK+QCD cross sections in three fiducial regions defined to be as close as possible to the reco-level selection. The measurements were done using the MELA distribution as a base for a maximum likelihood fit to the observed data and a cut-and-count approach for the EWK and EWK+QCD cross section measurements, respectively. The EWK signal strength measurement in the three regions was reported as well. The background-only hypothesis was rejected with significance of 4.0 σ (3.5 σ expected).

A Boosted Decision Tree (BDT) classifier was used as an alternative signal extraction method in order to gauge possible gain in the sensitivity, with respect to MELA. The nominal BDT classifier used seven input variables to extract the EWK signal from the main QCD-induced background and is referred to as the *BDT7*. An additional BDT was built using the a of 28 variables, referred to as the *BDT28*, in order to assess possible gain when using a larger set of variables.
The shape of the BDT classifier was used as the template for the maximum likelihood fit. The background-only hypothesis was rejected using the *BDT7* with expected significance of $3.77\ \sigma$ for the combined data-taking period. This can be compared to the significance of $3.5\ \sigma$ obtained using MELA. It shows that MELA is able to capture the full kinematics of the event and a small gain in the significance (< 6%) is obtained which was not deemed enough to change the methodology. Observed EWK signal significance of 5.1 σ using the *BDT7* is reported.
The observed (expected) significance using BDT28 was found to be 4.69 σ (3.85 σ). This shows only a marginal gain in sensitivity ($\approx 2\%$) was achieved compared to BDT7.

The expected and observed lower and upper 95% CL limits on the anomalous quartic gauge couplings for the charged-current operators T_0, T_1 and T_2 as well as the neutral-current operators T_8 and T_9 are also reported. The limits obtained for the neutral-current operators T_8 and T_9 and discussed in this chapter were the tightest bounds available for these couplings at the time of the publication.

Chapter 5

Prospective studies for the High-Lumi and High-Energy LHC

5.1 Preface to the chapter

The previous chapter showed that the Run 2 data had opened the door for the measurement of the VBS processes with two Z bosons accompanied by the two jets coming from EWK vertices. However, the measurement of the individual vector boson polarizations remains out of reach because of the low cross section of these processes. At the same time, it is the longitudinal polarization of vector bosons that is directly connected to the EWSB and the Higgs mechanism. In 2018, a study was done, using the 13 TeV LHC data [146], to project the measurement sensitivity of the longitudinal polarization of the Z bosons for the HL- and HE-LHC conditions. This was done by simply scaling the measured yields with luminosity and cross section expected at future LHC conditions. This provided a motivation to simulate detailed kinematics at 14 and 27 TeV and do a more in-depth analysis. This analysis is presented here.

In the first sections, the reader will familiarize themselves with the MC simulations of the signal and background processes that were prepared for this analysis. Event selection is defined in section 5.3. I studied the effect of extended acceptance for electrons that is considered for the detector upgrade at the HL-LHC phase that would extend the HGCAL with silicon layers in front of HF (referred to as the HF nose). Since no additional treatment for the HF nose option was needed in the analysis, it is only referred to in the last section when the results are discussed. Section 5.4.1 will describe the lepton-jet cleaning algorithm that I designed in order to remove lepton duplicates from the jet collection. The origin and the effect of these on the analysis are also discussed. I studied the effect of parton showers and PU on the selection of the leading and the subleading jets. This is presented in sections 5.4.2 and 5.4.3. In order to maximize the signal sensitivity measurement, I designed two signal extraction algorithms: the combined-background BDT and the 2D BDT. This will be covered in section 5.6.1.
Next, the kinematics for the signal and the background processes and the application of the signal extraction techniques on 14 and 27 TeV samples will be shown. The results are presented in section 5.7 followed by a summary of the key points discussed in the chapter.

5.2 Simulations of the signal and backgrounds

The first step in the analysis is the simulation of the hard processes of interest. This was done using *MG5* package for all EWK processes and irreducible QCD $pp \to ZZ$ background. The gluon loop-induced QCD background was simulated using *MCFM*.
The next step in the simulation is the parton showering and hadronization of the outgoing particles and the simulation of the detector effects. The former is done using the PYTHIA8 framework and the latter using the DELPHES tool.
PYTHIA8 is a standalone tool used to generate events in high-energy collisions. However, in this analysis, it has been used in conjunction with MG5 through the usage of Les Houches Event (LHEF) files [147]. This is a standard file format used in high-energy physics to store process and event information obtained from event generators. The matrix element calculation is done by MG5, and the output is stored in the standard LHEF format. This is then used by PYTHIA to simulate the parton showering and hadronization [148].
Beforementioned MG5 and PYTHIA8 tools deal with event production based on purely theoretical considerations. However, to do a proper analysis, one cannot dismiss the importance of the interaction between matter and radiation with the detector. Whenever such analysis requires a high level of accuracy, these interactions are simulated using the GEANT4 package. It is important to note that, although this tool provides the most sophisticated simulation of the detector effects, it is also very complex and time-consuming. For this analysis, such a level of precision is not required. Thus, detector effects were simulated using the DELPHES tool which was designed by the LHC collaborations to be two to three orders of magnitude faster than GEANT4. This is done by propagating particles emerging from hard processes to the calorimeters in the uniform magnetic field parallel to the beam direction. The energies and momenta of long-lived particles are smeared to match the detector response. To take into account the CMS measurement efficiencies in different η regions, an efficiency parametrization from the full detector simulation is used. All these effects are stored in configuration files that must be forwarded to Delphes at runtime. Standard configuration files used by the CMS collaboration were used for generating all processes. CMS_PhaseII_0PU_v02.tcl was used for 0 PU, while CMS_PhaseII_200PU_v03.tcl was used for 200 PU samples.

In Delphes, it is assumed that electrons and photons leave all their energy in the electromagnetic calorimeter (ECAL) and forward calorimeters (FCAL). At the same time, neutral and charged hadrons leave all their energy in the hadron calorimeter (HCAL) and FCAL. Finally, the sharing of particle energy between two or more neighbouring cells, in case the particle hits a cell near its edge, is not implemented. Electrons and muons are identified in Delphes with no fake rates. For both, the efficiency is exactly zero outside the tracker acceptance. Both final electrons and final muons are obtained by smearing their 4-momentum.
In the analysis, the final states are dominated by jets. As such, it is important to identify them correctly. It is possible in Delphes to produce jets by starting from different collections. These can be generated jets, calorimeter jets or particle-flow jets. Generated jets are obtained by clustering generator-level (henceforth gen level) particles after parton shower and hadronization. Calorimeter jets are reconstructed by using calorimeter towers which are overlayed collections of cells from ECAL and HCAL. Energy-flow jets are obtained by combining the information from particle-flow tracks and particle-flow towers. Particle-flow tracks are reconstructed tracks from the ECAL and HCAL originating from the charged hadrons.
At last, there are six different jet clustering algorithms in Delphes that can be used to reconstruct jets: CDF jet clusters, CDF MidPoint, Seedless Infrared Safe Cone, Longitudinally invariant kt jet, Cambridge/Aachen jet and Anti-kt jet algorithm. In this analysis, the Anti-kt jet algorithm was used [149, 150].

5.2.1 Simulations of the EWK signal

In this analysis, the signal is the purely electroweak production of the two longitudinally polarized, leptonically decaying Z bosons accompanied by two hadronic jets originating from electroweak vertices. In the rest of the text, this process will be referred to as simply LL. It was simulated at LO by explicitly requiring that the number of QCD vertices be zero:

$$generate\ pp > z\{0\}z\{0\}jj\ QCD = 0,\ z > l + l-$$

Samples for both HL-LHC and HE-LHC configurations were simulated by requiring 7 TeV and 13.5 TeV beam energy respectively.
An important parameter to be set is the parton distribution function (PDF). A PDF is defined as the probability of finding a parton within a proton with a given fraction of the total proton energy. In the MC simulation of the signal, the cteq6l1 PDF set was used [151].
In addition, 10 GeV and 3 GeV cuts were imposed on the p_T of the jets and leptons, respectively. Cuts on pseudorapidity for both jets and leptons have been left open to enable a study of the effect of the future hadronic nose (henceforth HF nose) upgrade on the measurement sensitivity. Finally, a cut of 100 GeV is imposed on the di-jet system mass to suppress the tri-boson contribution. Samples with and without parton showering were simulated at both 14 TeV and 27 TeV to check the effect of parton showering on the tagging jets. In addition, zero PU samples were produced to check the effect of PU. In the end, 200 PU samples with parton showering included were used to obtain the signal significance.

5.2.2 Simulations of the EWK backgrounds

The EWK background in this analysis is the purely EWK production of the two leptonically decaying Z bosons accompanied by two hadronic jets originating from electroweak vertices where at least one Z boson has transverse polarization. These processes will be referred to as LT and TT in the following chapters.

$$generate\ pp > z\{0\}z\{T\}jj\ QCD = 0,\ z > l + l- \qquad (LT\ and\ TL\ polarisation)$$

and

$$generate\ pp > z\{0\}z\{T\}jj\ QCD = 0,\ z > l + l- \qquad (TT\ polarisation)$$

Generator level cuts are identical to those used in the signal simulations. The cross sections of EWK samples used in the analysis are given in Table 5.1

	EWK LL 14 TeV	**EWK LT 14 TeV**	**EWK TT 14 TeV**	**EWK LL 27 TeV**	**EWK LT 27 TeV**	**EWK TT 27 TeV**
$\boldsymbol{\sigma[fb]}$	0.033	0.189	0.317	0.115	0.669	1.142

Table 5.1: Cross sections, at the generator level, for all EWK processes at the HL-LHC and HE-LHC.

From the table, one can see that the LL contribution is only $\approx 6\%$ of the total at 14 TeV which is also the case at 27 TeV. The cross section of each contribution rises by a factor ≈ 3.5 when going from 14 TeV to 27 TeV.
Examples of Feynman diagrams showing the EWK production of two Z bosons and 2 hadronic jets can be seen in Fig. 5.1. Figure also shows an interference diagram with the Higgs boson which ensures the unitarization of the theory.

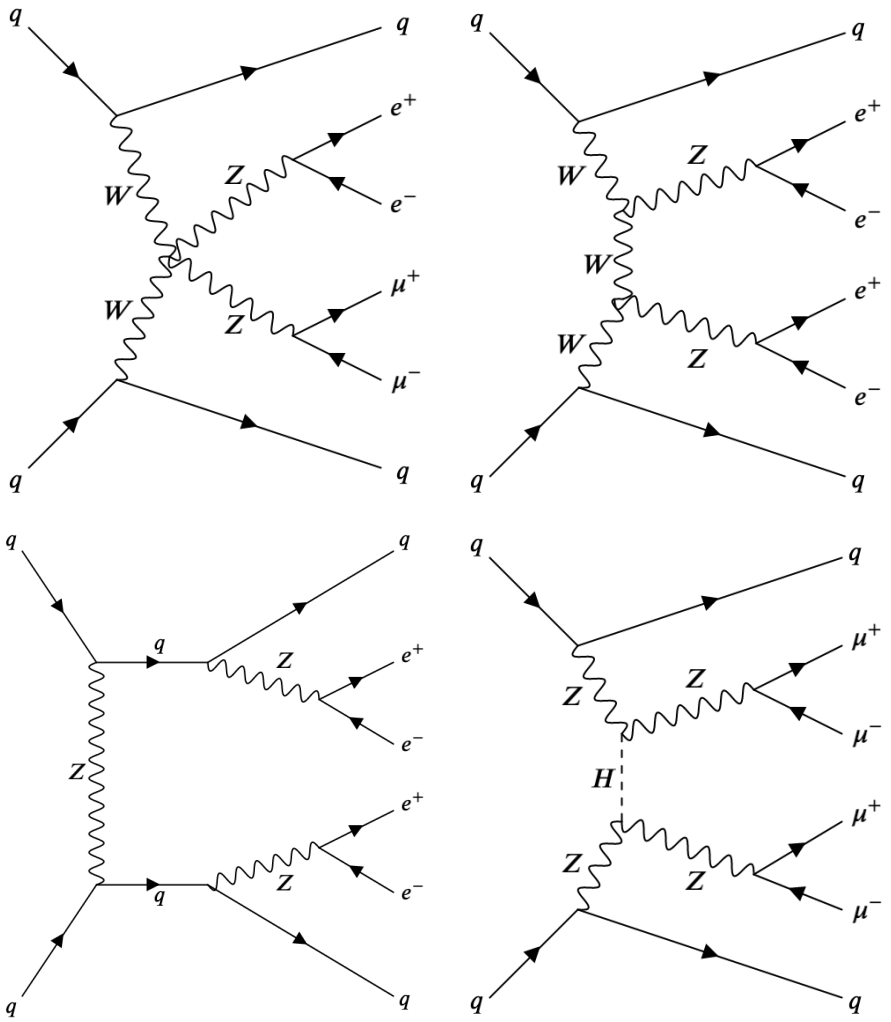


Figure 5.1: Example diagrams for the EWK production of two jets and two Z bosons decaying leptonically. The interference of the bottom-right diagram featuring the Higgs boson exchange with the processes depicted in the top row ensures the unitarization of the theory.

5.2.3 Simulations of the QCD backgrounds

The dominant background for this analysis is QCD-induced $pp \to ZZ$ process with up to 2 extra parton emissions coming from the QCD vertices. This is an irreducible background since the final state is identical to that coming from the signal process. Henceforth, the main QCD background will be referred to as the *qq* background.
As for the EWK processes, the qq background was simulated at HL- and HE-LHC conditions with zero PU as well as 200 PU and with the parton showering included as well as without it. In addition, 1,2-jet samples were simulated at LO and a 1-jet sample was simulated at the NLO. All the samples were simulated with MG5 with the following syntax:

$$generate\ pp > zzj\ QCD = 1\ QED = 2,\ z > l + l - \qquad (1j@LO)$$

$$generate\ pp > zzjj\ QCD = 2\ QED = 2,\ z > l + l - \qquad (2j@LO)$$

$$generate\ pp > zz > l + l - l + l - j\ [QCD] \qquad (1j@NLO)$$

The same set of generator-level cuts was used for the LO samples. The jet p_T was set to 10 GeV, while jet η was

set to 5. For the leptons, the p_T cut was set to 3 GeV, while η was left open. The di-jet mass of the 2j@LO sample was set to 100 GeV. For the 1j@NLO sample, the jet p_T cut was set to 15 GeV. Other cuts were later defined at the reconstruction level.
The LO samples were used to assess the effect of PU on the leading and the subleading jets in the case of a single jet and two jets produced at ME. The 1j@NLO sample with a parton shower included is the nominal sample and was used in the analysis. To simulate the 1j@NLO sample without the parton shower, one must specify this in the main Delphes configuration file:

$$PartonLevel : ISR = off$$

$$PartonLevel : FSR = off$$

As for the EWK samples, the cteq6l1 PDF was used for the LO samples. For the NLO sample, NN23NLO PDF was used.

Finally, there is also a gluon loop-induced ZZ production simulated at LO and hence denoted gg background. Although it contributes only at around 10% level with respect to the main background, it is nevertheless included to obtain better projections of the signal sensitivity. It is simulated at both HL- and HE-LHC conditions using the MCFM package. To simulate the desired process in MCFM, one must define a process number in the configuration card. This was set to 132 which corresponds to a LO production of the $gg \to ZZ$ processes with 4 leptons in the final state. In order to faster simulate a gg contribution, only $2e2\mu$ final state was included, therefore omitting the $4e$ and 4μ final states. Effectively, only half of the phase space is simulated this way which is reflected in the event counts. To counter this, expected event counts are doubled before performing multivariate analysis. For this process, the NN2.3NL PDF set was used. In the previous chapter, the state-of-the-art gg simulation was used. This was not done here since that level of precision was not needed.
Examples of the QCD background diagrams are shown in Fig. 5.2. The cross sections from the QCD samples used in the analysis are given in Table 5.2. At 14 (27) TeV, the cross section of the qq background is ≈ 14 (≈ 11) times larger than the gg cross section. The increase in cross section when moving to 27 TeV is more pronounced in the gg background (≈ 3 times) than in the gg background (2.3 times).

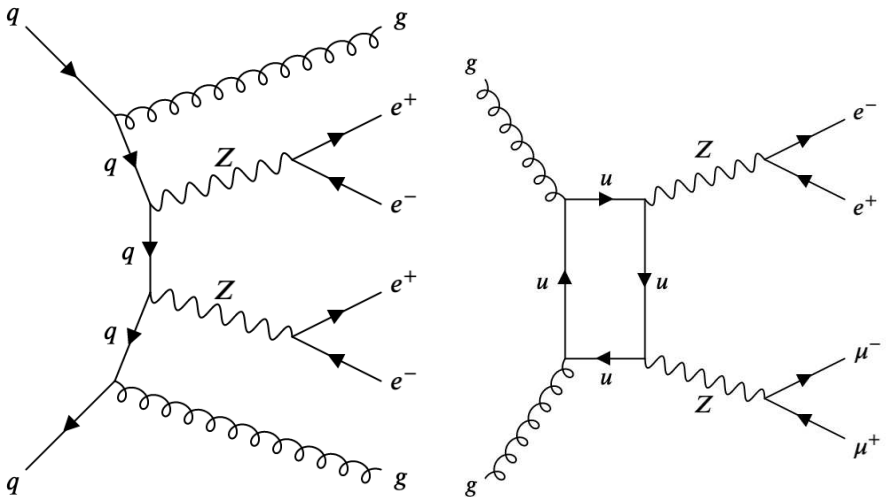


Figure 5.2: Example diagrams of the QCD-induced production of two Z bosons in the fully leptonic decay channel. The left figure shows the irreducible background with 2 jets in the final state. The right figure shows loop-induced production of the gg background.

	QCD qq 14 TeV	**QCD gg 14 TeV**	**QCD qq 27 TeV**	**QCD gg 27 TeV**
$\sigma[fb]$	49.6	3.57	116	10.9

Table 5.2: Cross sections, at generator level, for the QCD qq 1j@NLO and QCD gg@LO samples at HL-LHC and HE-LHC energies.

5.3 Event selection

The $ZZ \to 4l2j$ channel is a good candidate for the study of the EWSB mechanism due to the clean final state that can be fully reconstructed. For this reason, it is also a great fit for studying the scattering of longitudinal vector bosons. However, a small cross section of the LL processes, compared to the background, makes it challenging to measure. In order to suppress the background as effectively as possible, a set of efficient selection criteria has to be put in place. This was done in several steps:

1. In the analysis we require isolated objects. In the Delphes framework an object, such as an electron or a muon, is said to be isolated if the activity in a cone of radius R around the lepton direction is small enough. This is precisely defined with the variable I as

$$I = \frac{\sum_{i \neq P}^{\Delta R(i)<R,\ p_T(i)>p_T^{min}} p_T(i)}{p_T(P)}$$

   where the nominator sums over the p_T of all particles that are in the cone around the particle of interest, P, and the denominator is the p_T of the particle P. Values of $I > I_{min}$ indicate that the particle is isolated. The parameters $R,\ p_T^{min}$ and I_{min} are set to 0.5, 0.1 GeV and 0.1 respectively.
   For isolated electrons (muons), the p_T is required to be above 7 (5) GeV, while the $|\eta|$ is required to be less than 3 (2.8). Additionally, the extended acceptance for electrons was considered for which the $|\eta|$ acceptance for electrons was increased to 4.

2. With both HL- and HE-LHC conditions, PU is expected to affect the analyses and the CMS collaboration has been working on a number of algorithms to mitigate its effects. One such technique, the charged-hadron subtraction (CHS), was designed to remove charged particles coming from pileup vertices from the reconstructed objects and was used to treat leptons in this analysis. This method is not efficient enough when the PU contributions come from neutral hadrons. For this reason, a new PU mitigation approach, the pileup per particle identification (PUPPI), was devised. This technique was built on top of the CHS algorithm, and it estimates the probability that the neutral particle comes from the PU. It then scales the energy of such particles based on the calculated probability [152]. The PUPPI algorithm was used to reduce the effect of PU on jets.
   In this analysis, cuts were set to 25 GeV for the jet p_T and 4.7 for the jet $|\eta|$

3. Final state leptons are coming from the decay of Z bosons. Each Z boson candidate is reconstructed from a pair of oppositely charged electrons or muons with a dilepton mass in the window $60\ GeV < m_{ll} < 120\ GeV$. Each event is required to have a pair of non-overlapping Z bosons where the leading Z boson is chosen as the one with the highest p_T.

This set of requirements is referred to as the *ZZ selection*. For the analysis, two regions of interest are defined by additional selections:

- a baseline selection is built on top of the ZZ selection by requiring $m_{jj} > 100GeV$.
- a VBS selection is defined by requiring $m_{jj} > 400GeV$ and $|\Delta\eta_{jj}| > 2.4$

A summary of the selection criteria used in the analysis is shown in Table 5.3

lepton candidates	$p_T^e > 7GeV$ $\quad p_T^\mu > 5GeV$ $\|\eta\|^e < 3\ (4)$ $\quad \|\eta\|^\mu < 2.8$
jet candidates	at least two jets in the event with $p_T > 25\ GeV$ $\quad \|\eta\| < 4.7$
ZZ selection	lepton pair (e^+e^- or $\mu^+\mu^-$) with $60\ GeV < m_{ll} < 120\ GeV$ pair of non-overlapping Z bosons Z_1 defined as the one with the highest p_T Z_2 defined as the one with the next-to-highest p_T
baseline selection	ZZ selection + $m_{jj} > 100GeV$
VBS selection	ZZ selection + $m_{jj} > 400GeV + \|\Delta\eta_{jj}\| > 2.4$

Table 5.3: Summary of the selection criteria used in the analysis. The number in parentheses for the electron p_T is referring to the extended HGCAL option.

The efficiencies, defined for each contribution as the number of events passing the selection over the number of generated events, after the ZZ selection, baseline selection and VBS selection are reported in Table 5.4.

	ZZ selection		**Baseline selection**		**VBS selection**	
	14 TeV	27 TeV	14 TeV	27 TeV	14 TeV	27 TeV
Z_LZ_L efficiency [%]	51.5	44.2	44.3	38.4	30.3	27.5
Z_LZ_T efficiency [%]	53.9	47.2	47.8	42.5	31.1	29.8
Z_TZ_T efficiency [%]	59.0	52.6	52.6	47.8	32.7	32.3
qq efficiency [%]	44.9	36.6	9.80	11.1	1.40	1.90
gg efficiency [%]	42.1	40.9	13.7	16.7	3.50	4.80

Table 5.4: Signal and background efficiencies for the ZZ selection, baseline selection and VBS selection.

Different effect of the ZZ selection on the EWK and QCD contributions is mainly due to differences in the jet p_T spectrum, while the main drivers of differences in the remaining two regions are the m_{jj} and $\Delta\eta_{jj}$ spectra of the two leading jets for the EWK and QCD processes. This is discussed in more detail in section 5.5 where distributions for the different polarizations, as well as for the qq and gg backgrounds, are shown.

5.4 Cleaning of lepton-jets and effect of parton showering and pileup on the leading and subleading jets

5.4.1 Lepton-jet cleaning

Signal events in this analysis are characterized by two hadronic jets with a high pseudorapidity gap between them. It is thus imperative to build an analysis that will be as effective as possible in identifying such jets.
However, it was found that Delphes populates the jet collection with objects previously reconstructed as leptons and stored in the lepton collections. If untreated, this will lead to double counting of objects in an event and wrong interpretation of analysis results. Leptons that are found in the jet collection will be referred to as the *lepton-jets*.

When comparing the η spectrum of leptons and two leading jets, one would expect to see clearly distinctive distributions. However, the top two rows in Fig. 5.3 show that the η spectrum of leptons and jets is similar. In addition, one would expect a large pseudorapidity gap between the tagging jets in the signal sample which is not the case as can be seen on the bottom plot. This points to the lepton contamination of the jet collection used in the analysis. To check this, events were examined before the baseline selection. Example of one such event is given in Table 5.5. Along with four leptons, this event has five jets. However, it can be seen that e_1 and j_1 are the same objects stored once in the electron collection and once in the jet collection. The same is true for e_2 and j_3 and μ_1 and j_2. By looking at the LHEF file, it can be confirmed that these are the same final-state particles. The small discrepancy in the kinematics seen in Table 5.5 comes from the smearing of energy and momentum in Delphes. The detector response is different for leptons and jets and thus the applied smearing is also different.

One major issue with lepton-jets is that the analysis is sensitive to kinematic distributions and this information is used to extract the signal. In addition, event selection requires at least two jets in the event with a p_T of at least 25 GeV and a dijet mass of at least 100 GeV. When lepton-jets are removed from the event discussed in Table 5.5, the event does not pass the selection. Therefore, without removing lepton-jets, the final event counts will be wrong. Finally, if the lepton is wrongly identified as a jet, a softer jet candidate coming from PU will have a lower probability of becoming a leading or subleading jet and thus the effect of PU will be underestimated.

The lepton-jet cleaning algorithm is as follows:

1. Loop over each object in the jet collection.
2. For each jet, loop over every object in the electron collection. Calculate the distance, ΔR_{jl}, between the lepton and the jet.
3. Remove the jet closest to electron if $\Delta R_{jl} < 0.1$
4. Apply steps 1-3 also for objects in the muon collection.

The performance of the lepton-jet cleaning algorithm was thoroughly checked by going through dozens of events one-by-one and checking whether the algorithm removed fake jets while leaving others untouched. Next, all important kinematic variables were plotted to make sure that lepton-jets were removed. The same set of kinematic variables shown in 5.3 is shown in Fig. 5.4 after applying lepton-jet cleaning algorithm. The distributions show the expected difference between the lepton and the jet kinematics, as well as the expected pseudorapidity separation between the tagging jets.

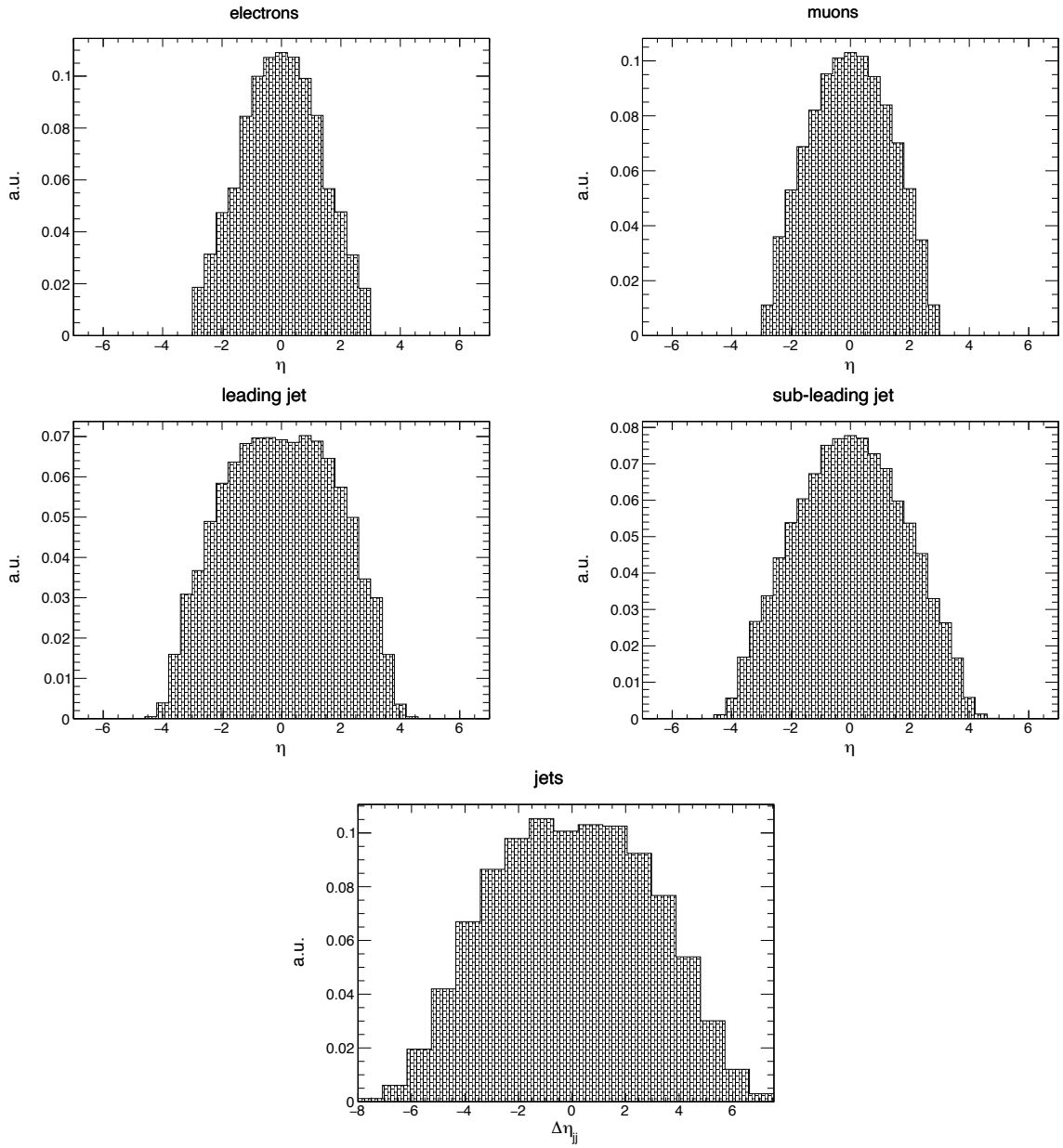


Figure 5.3: Top row: pseudorapidity spectrum of leptons. Middle row: pseudorapidity spectrum of the two leading jets. Bottom: the pseudorapidity difference between the two leading jets. Distributions for the signal sample, obtained before implementing the lepton-jet cleaning algorithm, are shown. Samples are simulated at 14 TeV c.o.m. energy and the baseline selection was applied.

	e_1	e_2	μ_1	μ_2	j_1	j_2	j_3	j_4	j_5
p_T **[GeV]**	121.7	20.6	81.7	14.9	125	85	20.6	18.6	16.8
η	-0.43	-0.29	-0.80	1.29	-0.42	-0.80	-0.29	2.52	-1.64
ϕ	2.99	0.63	-0.31	-0.91	2.98	-0.32	0.63	-1.39	1.37

Table 5.5: An example of values of the two jet kinematic variables in a single event before the baseline selection.

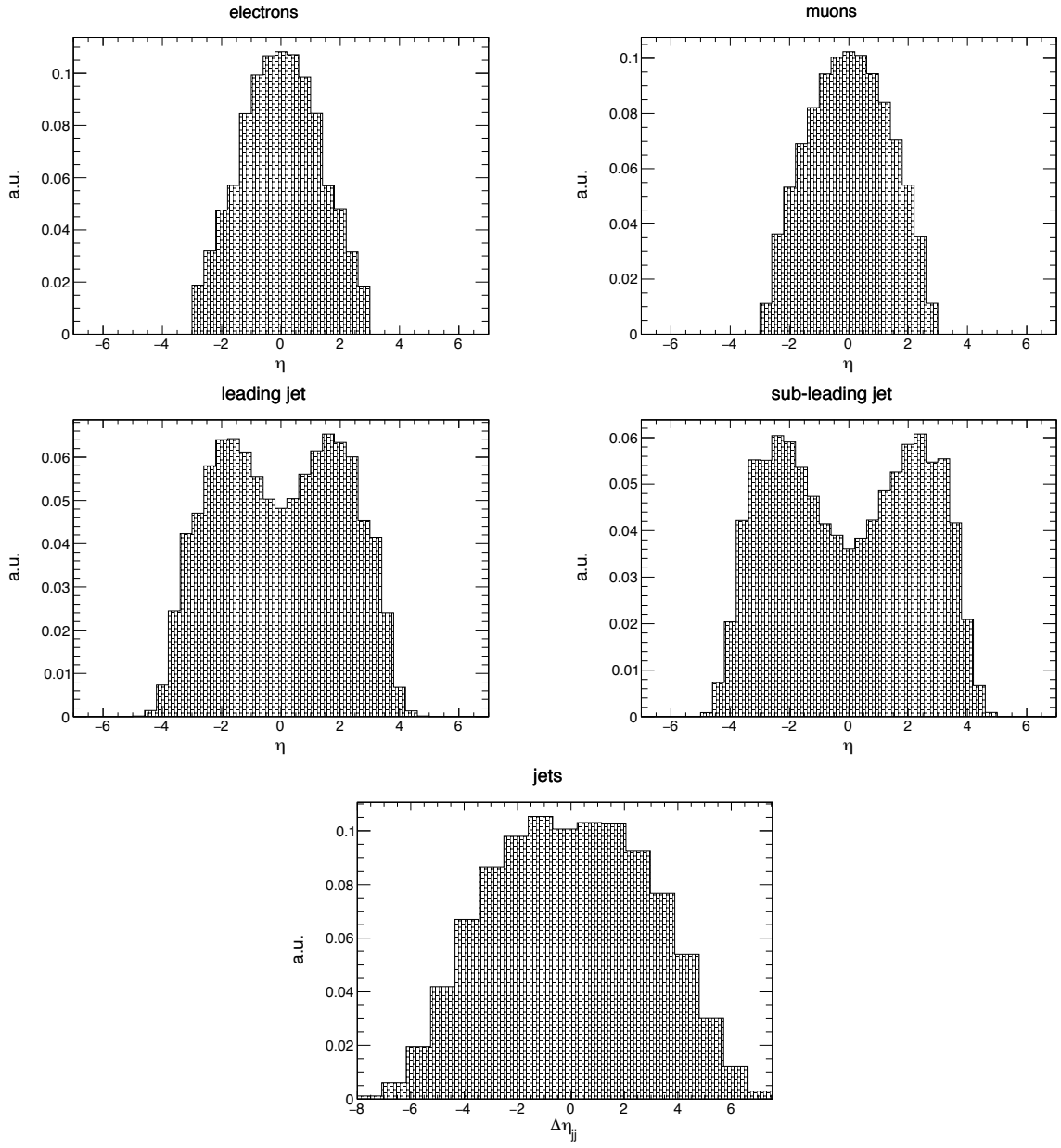


Figure 5.4: Top row: pseudorapidity spectrum of leptons. Middle row: pseudorapidity spectrum of the two leading jets. Bottom: the pseudorapidity difference between the two leading jets. Distributions for the signal sample, obtained after implementing the lepton-jet cleaning algorithm, are shown. Samples are simulated at 14 TeV c.o.m. energy with the baseline selection applied.

5.4.2 Effect of parton showering on the leading and subleading jets

The parton showering is introduced via initial state radiation (ISR) and final state radiation (FSR) simulations which are modelled with differential equations that give the probability of emitting radiation as the parton shower evolves with time. For the FSR, this is done by replacing a mother particle with two daughter particles at each branching. Contrary to the FSR where the parton shower evolves forwards in physical time, the ISR is simulated by starting from hard scattering partons and successively reconstructing prior branchings in the rising sequence of parton energies. In other words, the ISR evolution is modelled backwards in physical time [148, 153].
This section shows the effects of parton showering on the choice of the leading and the subleading jets. For this, the zero-PU (henceforth PU0) samples were used to prevent mixing the effects of PU and parton showers. The effect is shown for the VBS signal and the main QCD background. The study was done in several steps:

1. Run Delphes twice to obtain
    - sample with parton showering switched off (henceforth no-showering sample)
    - sample with parton showering switched on (henceforth showering sample)
2. record events that pass the baseline selection in the showering sample.
3. record events from the non-showering sample that were also recorded in step 2. This ensures that the same events are being compared.
4. Compare the two leading jets from step 2 to the two leading jets from step 3 and check
    - how often only the leading jet was changed by the parton showering
    - how often only the subleading jet was changed by the parton showering
    - how often either of the two jets was changed by the parton showering
    - how often both the leading and the subleading jets were changed by the parton showering
    - if the leading and the subleading jets simply swapped places or a new jet was introduced

The primary effect of parton showering is the increase of jet multiplicity within the event. This is shown in Fig. 5.5 which compares the number of jets within the same events for non-showering and showering samples of the VBS signal and the main QCD background.

In addition, parton showering can change the leading (subleading) jet. The leading (subleading) jet is said to be changed by parton showering if its distance, ΔR, to the leading (subleading) jet after parton showering is greater than 0.5. This is illustrated in Fig. 5.6. In the first case, both the leading jet (blue marble) and the subleading jet (red marble) remained the same after parton showering. In the second case, the parton showering caused the leading jet to be replaced by a new jet (green marble). In the third case, the parton showering caused the subleading to jet be replaced by a new jet (green marble). The fourth case depicts two possible scenarios in which both leading jets are changed by parton showering. In the first scenario, the two jets simply swapped places. However, in the second scenario, both leading jets have been replaced by new jets.

This effect is summarized for the VBS signal and the main QCD background in Table 5.6. In 1.5 % (0.8 %) events parton showering caused the leading jet in the signal (main background) sample to be replaced by a new jet coming from parton showers. This happened to the subleading jet in 16 % (29 %) of events. On the other hand, in 23 % (21 %) of events parton showering changed both leading jets! However, in 69 % (56 %) of those events, the two jets

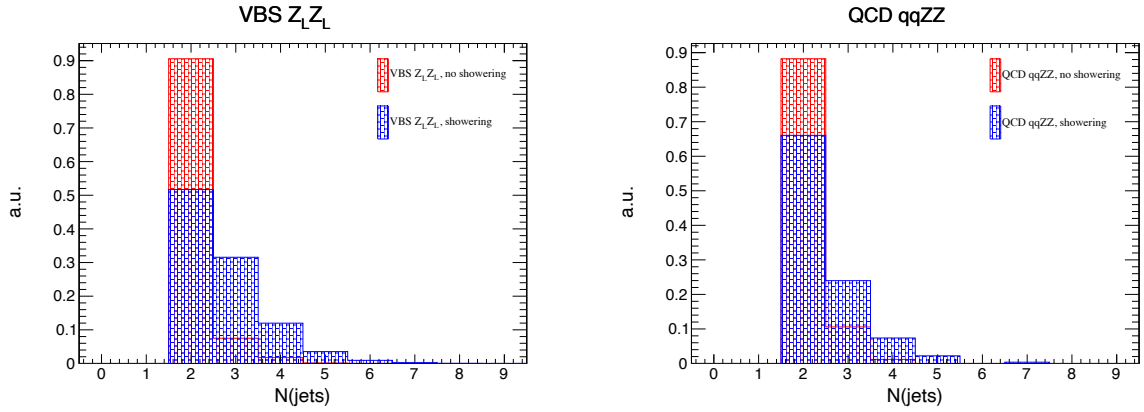


Figure 5.5: The effect of parton showering on the number of jets within the same event for the LL signal (left) and the qq background (right).

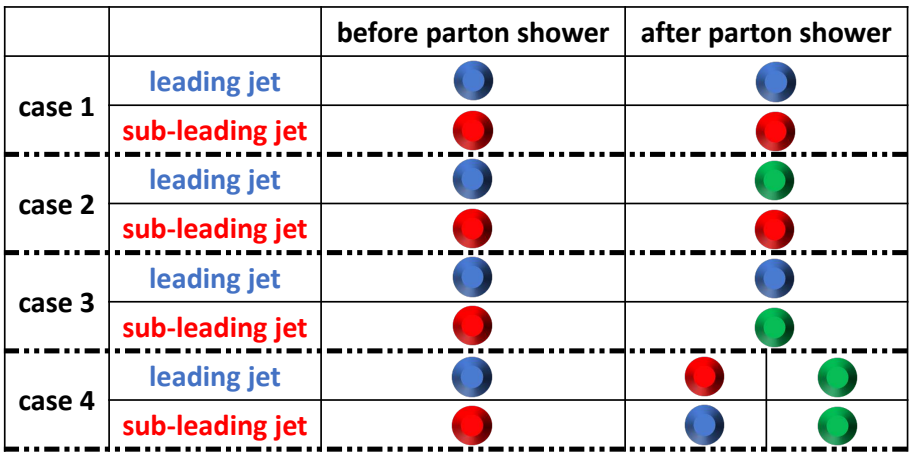


Figure 5.6: An illustration of the effect of parton showering on the two leading jets in an event. Parton showering can either simply swap the two leading jets, or it can introduce a new jet (green marble).

were simply swapped. Either of the two jets was changed in 40 % (51 %) of events. The results indicate that parton showering significantly affects the selection of the leading and the subleading jets.

VBS signal		**QCD qq**	
jets changed [%]	jet replaced	jets changed [%]	jet replaced
1.5	only first	0.8	only first
16	only second	29	only second
23	both	21	both
40	any	51	any

Table 5.6: The left-hand side of the table shows how often jets coming from parton showers interchange or replace tagging jets. The right-hand side of the table shows the same for the leading jets of the main QCD background.

Comparing the cross section weighted event counts for VBS processes after the baseline selection for the non-showering and showering samples in Table 5.7 one can see that, for the VBS signal, the difference is bellow 10

%. The effect of parton showers on the VBS background is similar. To further show that parton showering is under control, a set of lepton and jet plots for VBS signal is shown in Fig. 5.7.

	number of cross section weighted events after the baseline selection	
	non-showering samples	**showering sample**
LL	2.56	2.79
LT	6.90	7.38
TT	13.8	14.6

Table 5.7: The number of cross section weighted events for the VBS processes after the baseline selection at 14 TeV. Both non-showering and showering samples were produced from the same gen level output so that the effect of parton showering can be isolated and quantified.

5.4.3 Effect of pileup on the leading and subleading jets

In 2018, LHC has reported a mean PU of 32 at 13 TeV c.o.m. energy. This number is expected to be around 200 for the HL-LHC at 14 TeV. PU makes physical analyses more difficult by adding a large background noise and, therefore, must be treated carefully. The study of PU effects was also done in several steps:

1. Run Delphes twice to obtain
    - sample without pileup (henceforth PU0 sample)
    - sample with 200 pileup (henceforth PU200 sample)
2. record events that pass the baseline selection in the PU200 sample.
3. record events from the PU0 sample that were also recorded in step 2. This ensures that the same events are being compared.
4. Compare the two leading jets from step 2 to the two leading jets from step 3 and check
    - how often only the leading jet was changed by PU
    - how often only the subleading jet was changed by PU
    - how often either of the two jets were changed by PU
    - how often both the leading and the subleading jets were changed by PU
    - if the leading and the subleading jets simply swapped places, or a new jet was introduced

As for the previous study, the leading (subleading) jet is said to be changed by PU if its distance, ΔR, to the leading (subleading) jet after PU is greater than 0.5.
Table 5.8 summarizes the effect of PU on the leading and the subleading jets for the VBS signal and the qq background. In 0.3 % (1.1 %) of events, PU caused the leading jet in the signal (main background) sample to be replaced by a new jet coming from PU. This happened to the subleading jet in 9 % (15 %) of events. In 11 % (12 %) of events, PU changed both leading jets! However, in 81 % (67 %) of those events, the two jets were simply swapped. Either jet was changed in 20 % (29 %) of events. This result is especially significant for the VBS signal where the tagging jets are replaced by the PU jets in around 10 % of events. Although this effect is not extreme, it is sizeable.

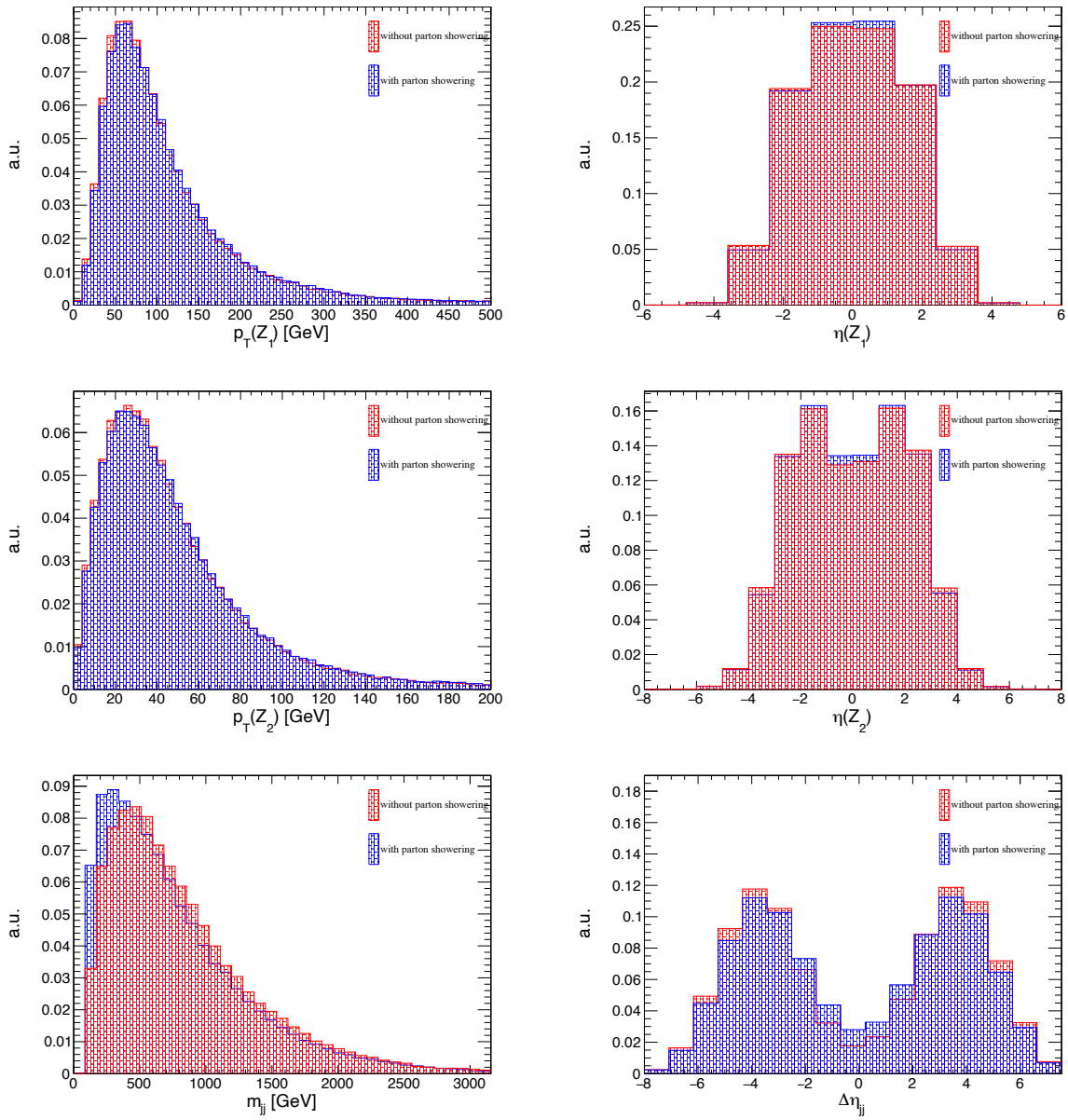


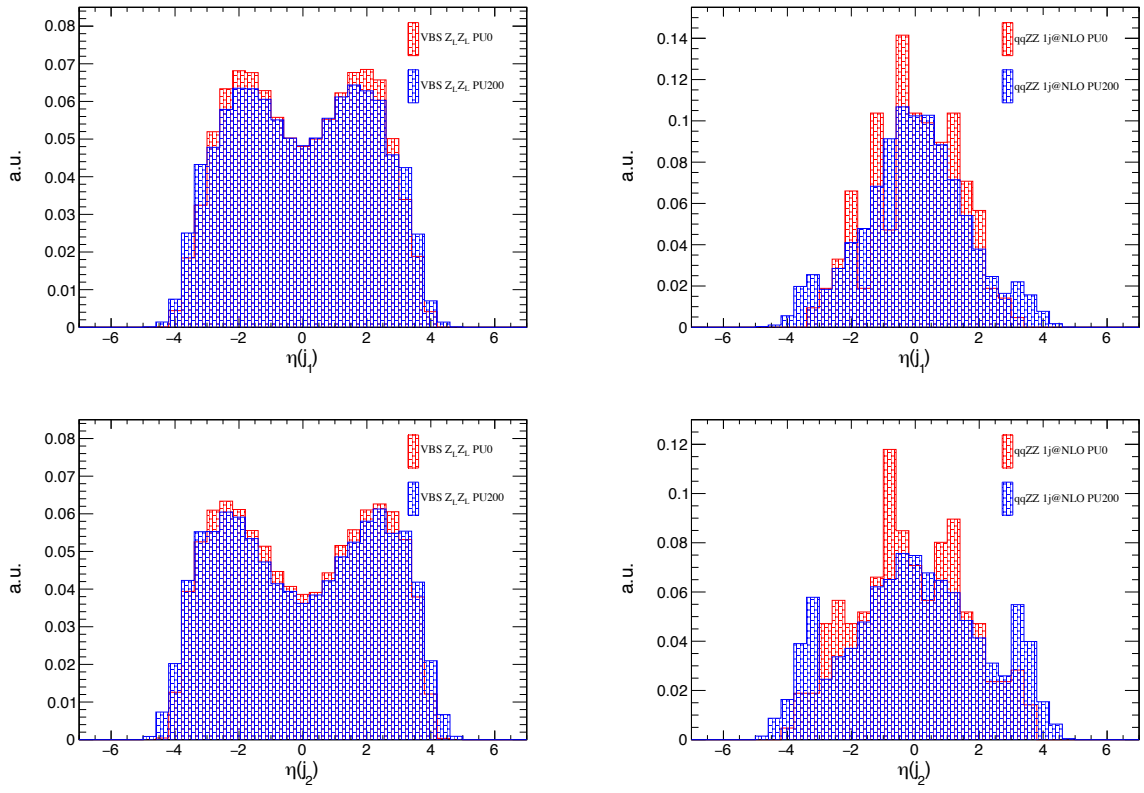
Figure 5.7: Lepton and jet kinematic distributions for the non-showering and showering samples after the baseline selection at 14 TeV. Distributions for the VBS signal are shown.

Another interesting PU feature can be seen by looking at the jet η distribution for the LL signal and the qq background shown in Fig. 5.8. The left-hand side plots show the pseudorapidity and the pseudorapidity separation between the two tagging jets in the signal sample, while the right-hand side plots show the same distributions for the two leading jets in the background sample. The effect of PU on the shape of jet distributions is especially pronounced in the qq sample.

VBS signal		**QCD qq**	
jets changed [%]	which jet replaced	jets changed [%]	which jet replaced
0.3	only first	1.1	only first
9	only second	15	only second
11	both	12	both
20	any	29	any

Table 5.8: The left-hand side of the table shows how often jets coming from PU interchange or replace tagging jets. The right-hand side of the table shows the same for the leading jets of the main QCD background. Both samples are simulated with parton showers included.

The distributions of both the leading and the subleading jets show two horns in the $3 < |\eta| < 4$ region. The low statistics of the 1j@NLO PU0 sample makes this harder to see. For this reason, the right-hand side distributions are also shown in Fig. 5.9 using the 1j@LO high-statistics sample. This feature is more pronounced in the η distribution of the subleading jets of the background samples compared to the signal sample because the subleading jet in the background sample is generally softer than the second tagging jet of the signal sample and is more affected by PU. One can recall from the previous chapter that horns were observed, in both data and the simulation, in the 2017 data-taking period and it was traced back to the noisy crystals. Here, the PU represents the noise in the analysis that causes horns to appear. Importantly, it was shown that these horns have a small impact on the analysis.



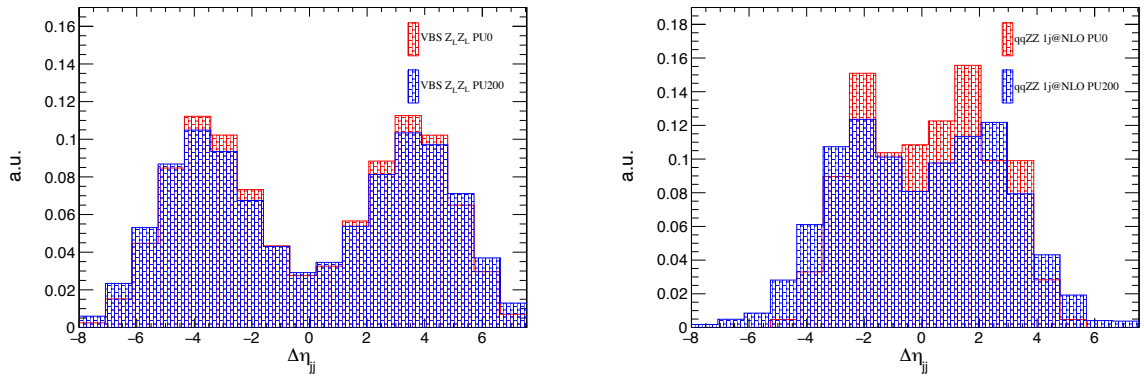


Figure 5.8: The left-hand side plots show the effect of pileup on the pseudorapidity for the two tagging jets in the LL signal samples as well as the pseudorapidity gap between them. The right-hand side shows the same distributions for the QCD qq1j@NLO background. All samples were produced at 14 TeV with parton showers included in the simulations.

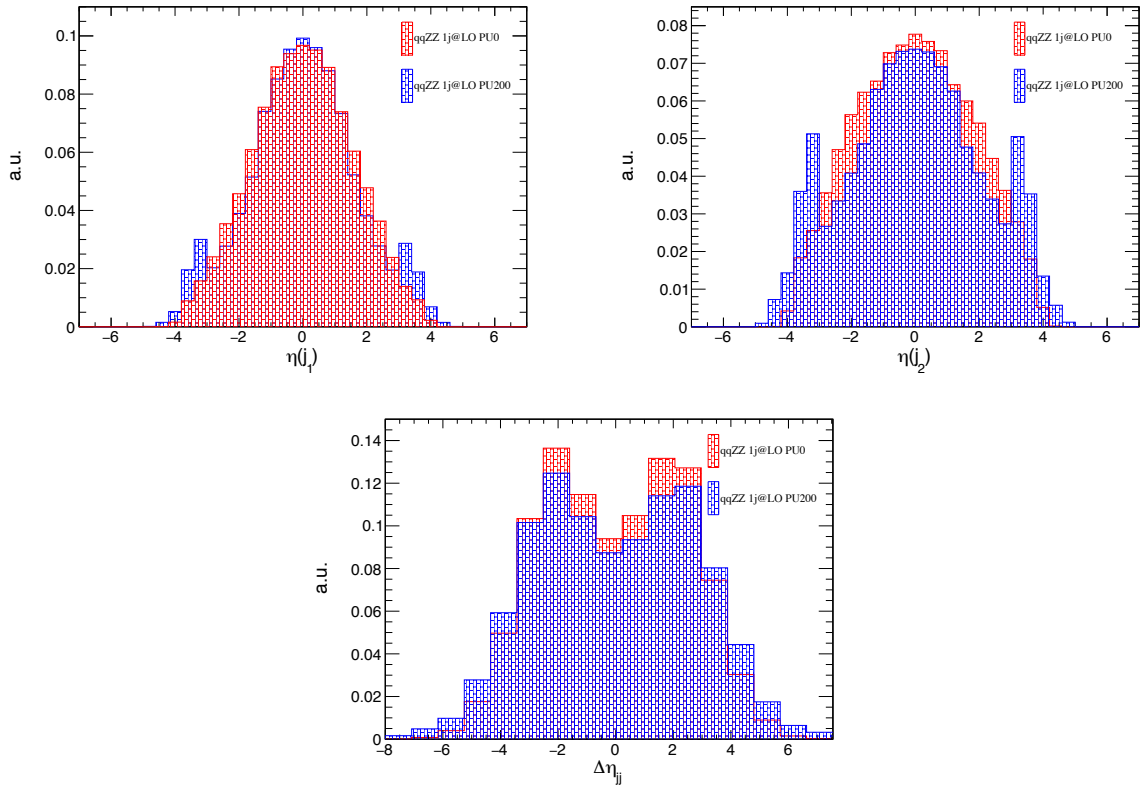
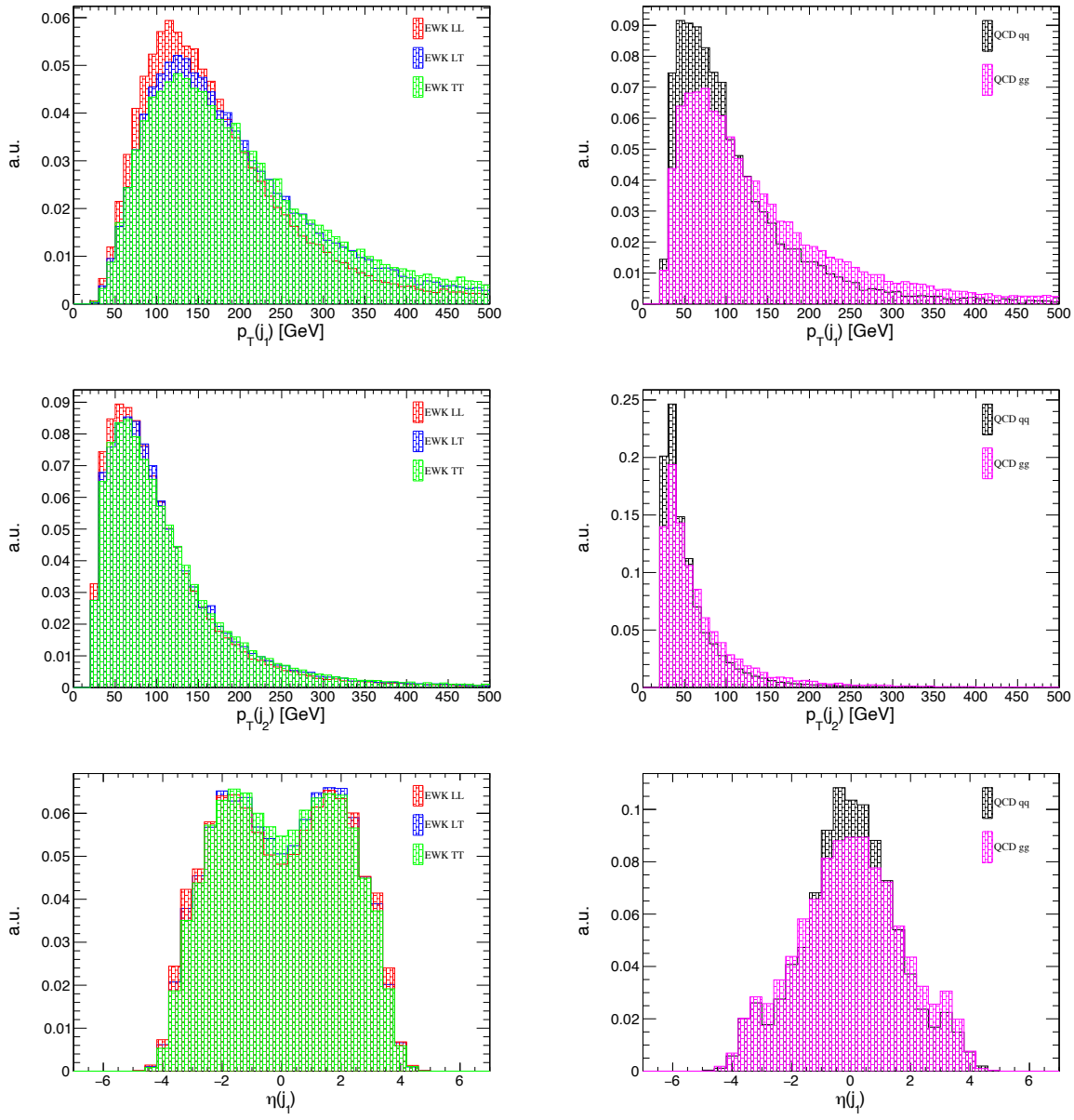


Figure 5.9: The effect of PU on the pseudorapidity for the two tagging jets in the QCD qq1j@LO samples as well as the pseudorapidity gap between them. Distributions are shown as a supplement for the right-hand plots in Fig. 5.8 since the LO samples were produced with higher statistics. All samples are produced at 14 TeV c.o.m. energy with parton showers included in the simulations.

5.5 Kinematics at 14 and 27 TeV

It was shown in section 5.3 (see Table 5.4) that the QCD background is more affected by the ZZ selection than the VBS contributions. This is mainly due to the requirements on the jets which have a harder p_T spectrum for the latter. This is shown on Fig. 5.10. The same figure shows the η distribution of the two leading jets for all contributions. It can be seen that the tagging jets in the LL signal have a somewhat softer p_T spectrum and are more forward than the LT and TT backgrounds. The two leading jets in the VBS samples have a harder m_{jj} spectrum compared to the leading jets in the qq and gg samples. This is reflected in the large drop in efficiency for the QCD samples after the baseline selection ($m_{jj} > 100\ GeV$). This is shown in the top row of Fig. 5.11. Finally, the VBS selection exploits the fact that the leading jets in the VBS samples have larger pseudorapidity separation compared to the leading jets of the qq and gg samples. This is shown in the bottom row of Fig.5.11.



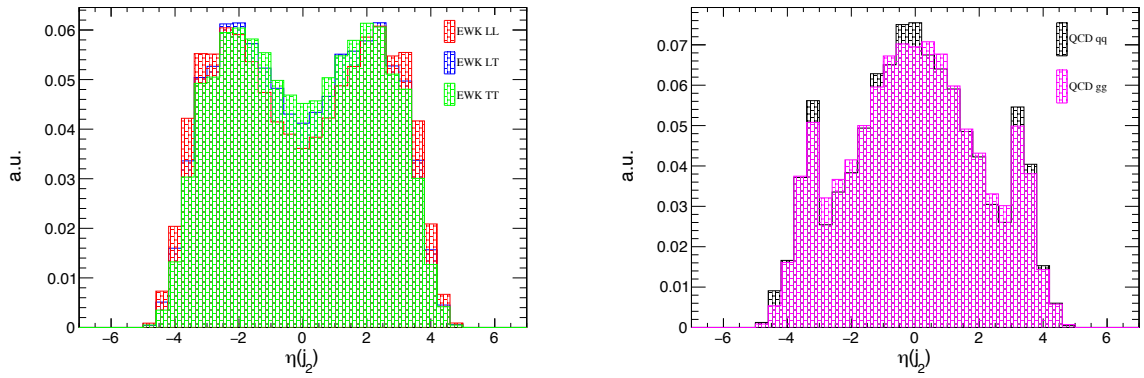


Figure 5.10: Transverse momentum and pseudorapidity of the leading jets for the VBS (left) and QCD (right) processes at 14 TeV. The baseline selection was applied.

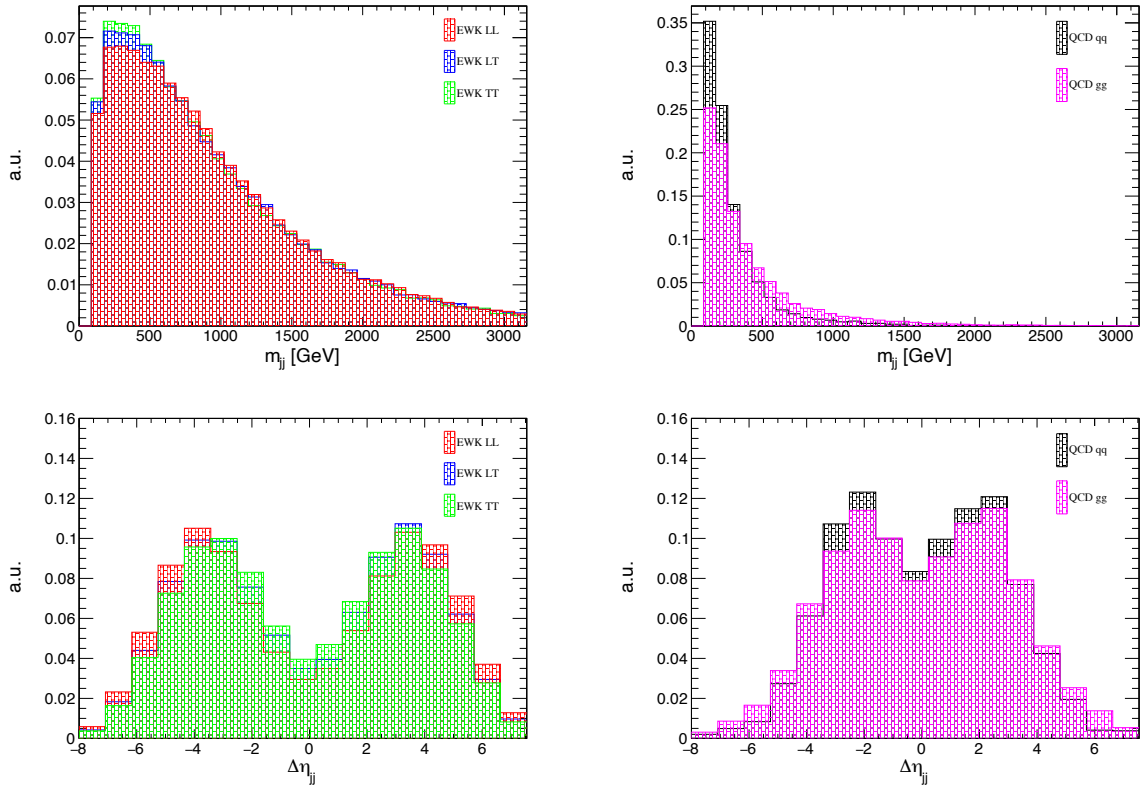


Figure 5.11: Dijet mass and pseudorapidity separation for the VBS (left) and QCD (right) processes at 14 TeV. The baseline selection was applied.

The same set of distributions is shown in Figs. 5.12 and 5.13 for 27 TeV. A bigger loss in the efficiency for the VBS contributions at 27 TeV comes mostly from the jet kinematics that shows a harder m_{jj} spectrum with more forward jets at 27 TeV compared to 14 TeV. All distributions at 14 and 27 TeV, along with 14 and 27 TeV distributions overlaid for easier comparison, are shown in Appendix B.

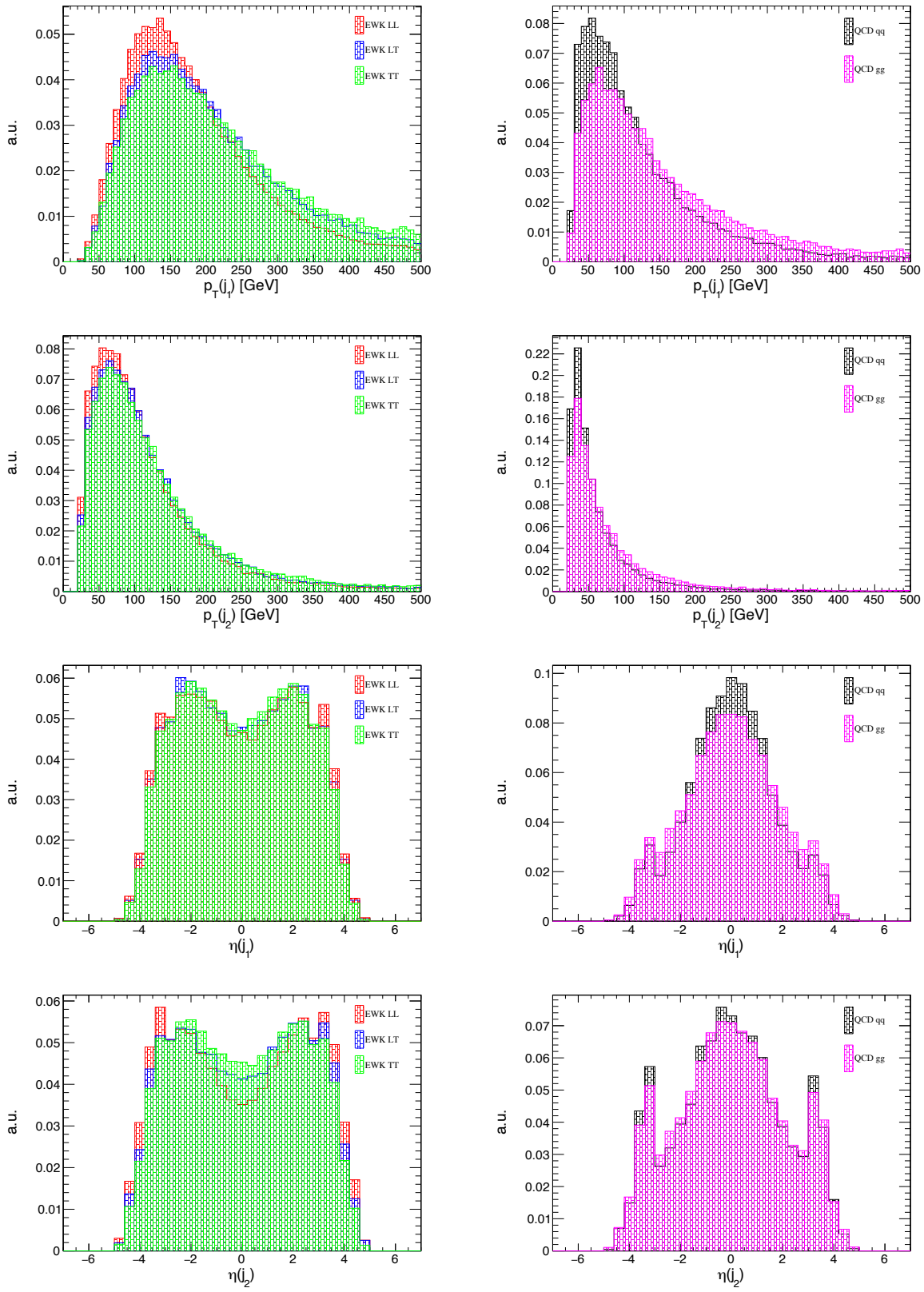


Figure 5.12: Transverse momentum and pseudorapidity of the leading jets for the VBS (left) and QCD (right) processes at 27 TeV. The baseline selection was applied.

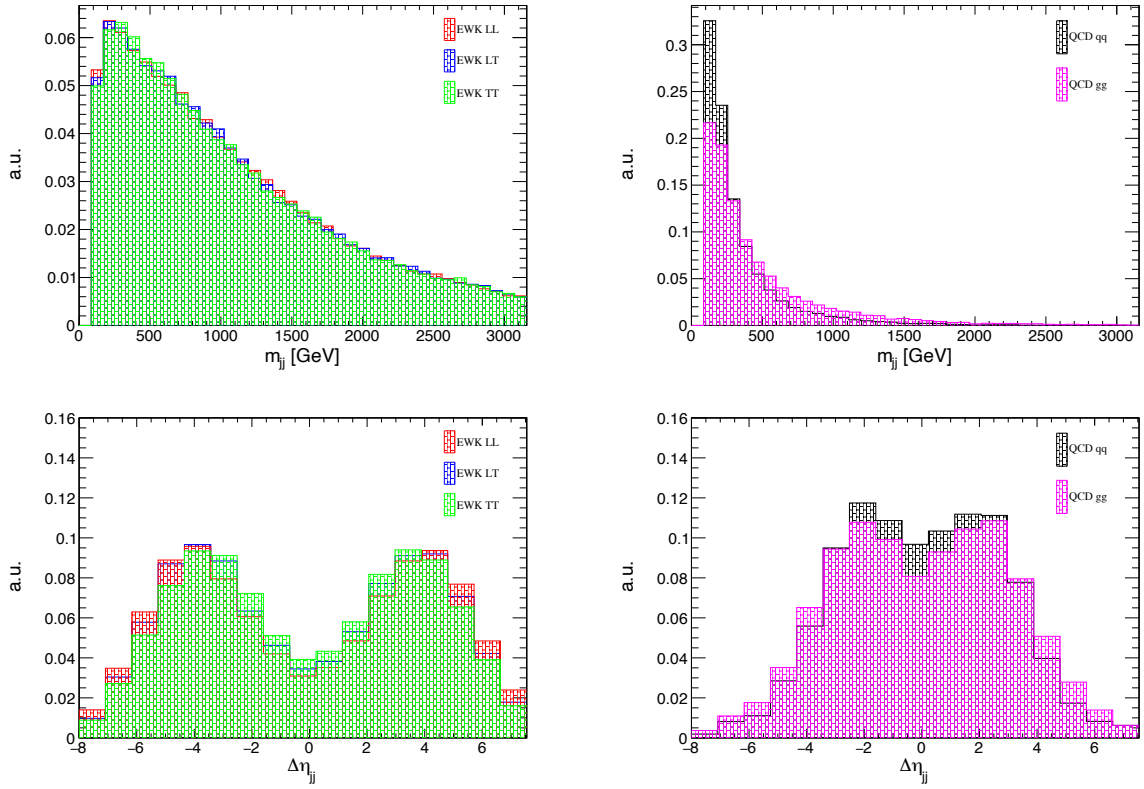


Figure 5.13: Dijet mass and pseudorapidity separation for the VBS (left) and QCD (right) processes at 27 TeV. The baseline selection was applied.

The emphasis of this chapter is on the extraction of the longitudinal polarization from the LT and TT polarizations and from the QCD backgrounds. The set of variables used to extract the LL signal is summarized in Table 5.9.
The first seven variables were shown in the previous chapter to separate well the VBS contribution from the QCD.

Along with p_T and η of the two Z bosons, variables $\theta^*(Z_i)$, defined as the angle between the momentum direction of the negatively charged lepton in the Z_i rest frame and the momentum direction of the Z_i in the laboratory frame, were found to separate well the LL signal from the LT and TT backgrounds. Angles $\theta^*(Z_i)$ are shown in Fig. 5.14.
Fig. 5.15 shows the distribution of the last six variables from Table 5.9 used to separate LL signal from the LT and TT backgrounds. It can be shown [154] that, when calculating decay rates, the matrix elements for transverse and longitudinal polarizations of vector bosons are

$$|M_-|^2 \approx (1 + cos\theta^*)^2 \qquad |M_+|^2 \approx (1 - cos\theta^*)^2 \qquad |M_L| \approx sin^2\theta^*$$

where $|M_-|$ and $|M_+|$ correspond to the left and right helicity states of the transverse polarization, respectively. From here, one would expect a very different angular distribution for transverse and longitudinal polarizations. This is shown in the bottom row of Fig. 5.15.
As a consequence, the p_T spectrum of the longitudinally polarized Z bosons is softer than the p_T spectrum of the transversely polarized Z bosons and the longitudinal component is produced at larger η values.
The same set of plots for 27 TeV is shown in Fig. 5.16

variable	definition
m_{jj}	invariant mass of the two leading jets
$\Delta\eta_{jj}$	pseudorapidity separation between the two leading jets
m_{4l}	invariant mass of the ZZ pair
$\eta^*(Z_1)$	η direction of the Z_1 relative to the leading jets: $\eta^*(Z_1) = \eta(Z_1) - \frac{\eta(j_1)+\eta(j_2)}{2}$
$\eta^*(Z_2)$	η direction of the Z_2 relative to the leading jets: $\eta^*(Z_2) = \eta(Z_2) - \frac{\eta(j_1)+\eta(j_2)}{2}$
$R_{p_T}^{hard}$	module of the transverse component of the vector sum of the two leading jets and four leptons in the event normalized to the scalar p_T sum of the same objects $R_{p_T}^{hard} = \frac{\lvert(\sum_{i=4l,\ 2j} \vec{V_i})_{transverse}\rvert}{\sum_{4l,\ 2j} p_T(i)}$
$R_{p_T}^{jet}$	module of the transverse component of the vector sum of the two leading jets and four leptons in the event normalized to the scalar p_T sum of the same objects $R_{p_T}^{jet} = \frac{\lvert(\sum_{i=2j} \vec{V_i})_{transverse}\rvert}{\sum_{2j} p_T(i)}$
$p_T(Z_1)$	transverse momentum of the Z_1
$\eta(Z_1)$	pseudorapidity of the Z_1
$p_T(Z_2)$	transverse momentum of the Z_2
$\eta(Z_2)$	pseudorapidity of the Z_2
$cos\theta^*(Z_1)$	angle between the momentum direction of the negatively charged lepton in the Z_1 rest frame and the momentum direction of the Z_1 in the laboratory frame
$cos\theta^*(Z_2)$	angle between the the momentum direction of the negatively charged lepton in the Z_2 rest frame and the momentum direction of the Z_2 in the laboratory frame

Table 5.9: Set of 13 variables used to separate the LL signal from the LT and TT polarizations and from the QCD backgrounds.

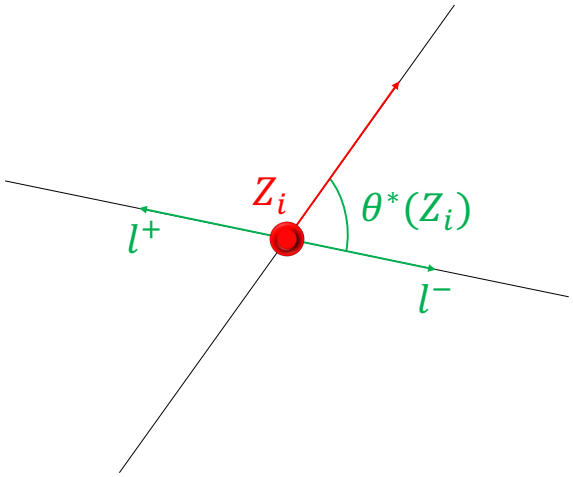


Figure 5.14: An illustration of angles $\theta^*(Z_i)$ defined in Table 5.9. The red arrow represents the momentum of the Z boson in the laboratory frame, while the green arrows represent the lepton momentum in the Z boson rest frame.

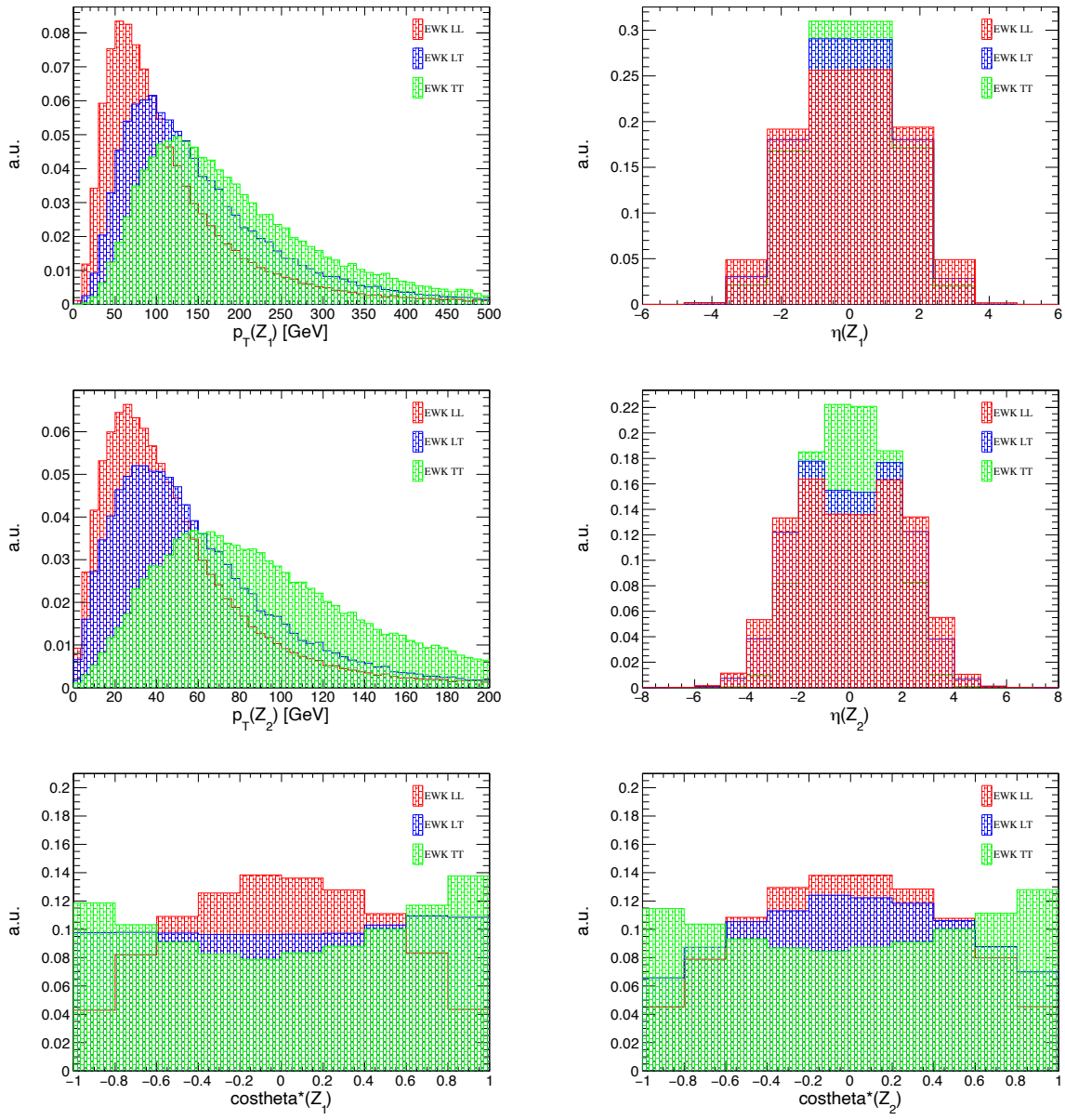


Figure 5.15: Distributions of the six variables from Table 5.9 used to extract the LL signal from LT and TT backgrounds. Plots for 14 TeV are shown, and the baseline selection was applied.

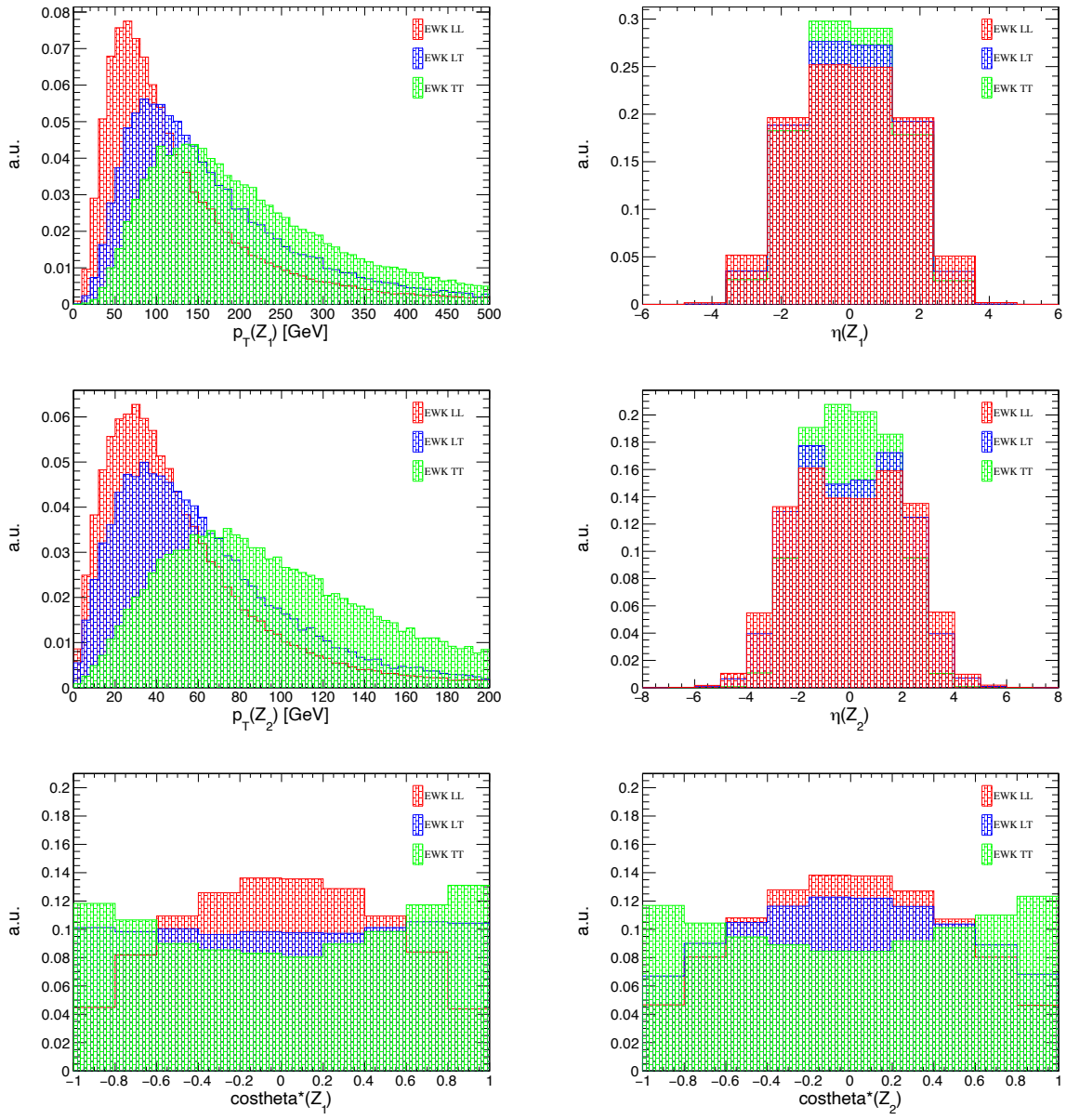


Figure 5.16: Distributions of the six variables from Table 5.9 used to extract the LL signal from LT and TT backgrounds. Plots for 27 TeV are shown, and the baseline selection was applied.

5.6 Signal extraction using a BDT and signal significance measurements

5.6.1 The combined-background BDT and the 2D BDT methods for signal extraction

Because of the low cross section of the LL signal, it is important to devise a signal extraction method that will keep as many signal events as possible while maximally reducing the background. Unfortunately, none of the signal distributions alone is discriminating enough to accomplish such a task. For this reason, a more complex method must be used. Two such methods were studied in order to obtain the maximum signal sensitivity:

1. Combined-background BDT
    - Train the BDT classifier on the events that pass the baseline selection to discriminate the LL signal from the mixture of all backgrounds
2. 2D BDT
    - QCD BDT: Train the BDT classifier on the events that pass the baseline selection to discriminate the LL signal from the qq background.
    - VBS BDT: Train the BDT classifier on the events that pass the baseline selection to discriminate the LL signal from the mixture of LT and TT backgrounds.

Regardless of the method used, the same set of 13 variables shown in Table 5.9 was used to train the BDT. For both methods, the gradient boosting was used with the hyperparameters listed in Table 5.10.

parameter	value	parameter meaning
NTrees	1000	number of trees in the forest
MinNodeSize	2.5	minimum percentage of training events required in a leaf node
Shrinkage	0.1	learning rate
nCuts	20	number of grid points used in finding optimal cut in node splitting
maxDepth	2	maximum allowed depth of the decision tree

Table 5.10: Hyperparameters used in the BDT training for the combined-background BDT and the 2D BDT at both 14 TeV and 27 TeV.

The combined-background BDT

In the combined-background BDT method, the first step, after selecting the appropriate set of variables and hyperparameters, is to properly weight each contribution. The weight for each contribution was calculated by dividing its cross section by the cross section of the qq background.
Although different, the kinematics of the loop-induced background is close to that of the main QCD background. In addition, the gg simulation used in this analysis is not state-of-the-art and using it in the BDT training at this stage would not gain much. For these reasons, the gg kinematics was not used in any training in this analysis. However, the result of the BDT training is always applied to the gg sample which is included in the calculation of the signal significance. Thus, in the combined-background BDT approach, the properly weighted LL signal is trained against the weighted mixture of the VBS backgrounds and the main QCD background.

An example of the BDT output distribution for the combined-background BDT is shown in Fig. 5.17. There is no sign of overtraining. To find the WP that maximizes the significance, $S/\sqrt{B}$, the cut values on the BDT output distribution that give the signal efficiencies in the range [10 %, 70 %] are calculated. Then, for the given signal efficiency, the efficiency of each background is calculated followed by the calculation of the signal significance.

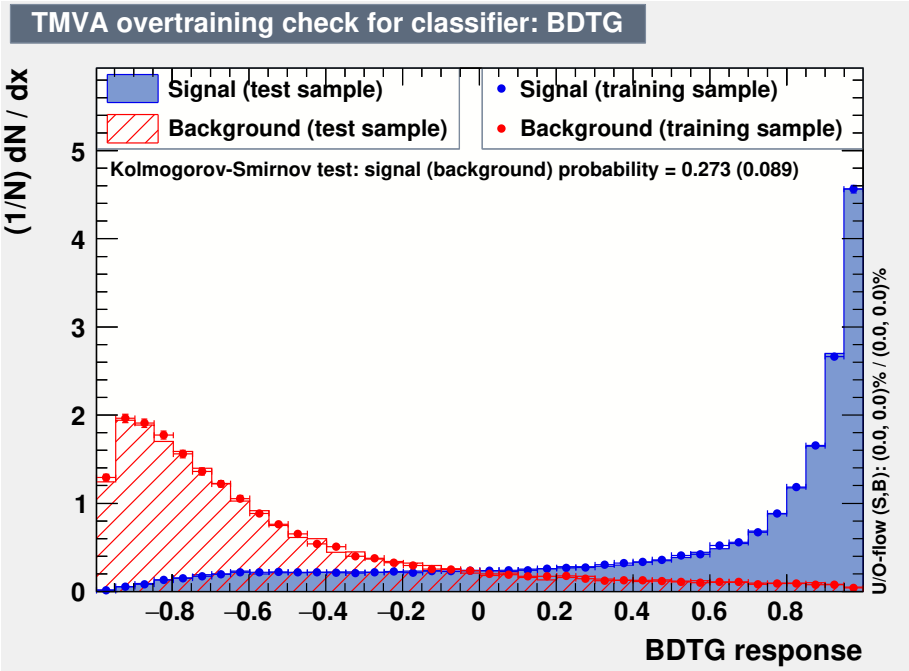


Figure 5.17: An example of the BDT output distribution of the training and test samples for the combined-background BDT approach.

The 2D BDT

The procedure of LL signal extraction with the 2D BDT method is illustrated in Fig. 5.18. Both the QCD BDT and the VBS BDT are trained in parallel on the same set of events that passed the baseline selection (referred to as the "original samples" in the rest of this section). The same set of weights, as used in the combined-background BDT, was also used here. For demonstration purposes, an example of the BDT output distributions obtained after training the QCD BDT and the VBS BDT is shown in Fig. 5.19.

The cut value on the QCD BDT output distribution is chosen so that 10% signal efficiency is obtained. The training of the QCD BDT is now applied to the original samples from which one calculates the efficiency of each background contribution. The efficiencies of the signal and each of the backgrounds after the QCD BDT are used later. At the same time, the BDT score is calculated for each event in the original samples. If the BDT score is greater than the cut value, an event is stored in a new sample (red discs in the illustration). In parallel, the cut values on the VBS BDT output distribution are chosen so that signal efficiencies in the range [10 %, 70 %] are obtained. For the sake of clarity, let's consider only a single cut value. This cut value is used to calculate the signal and background efficiencies, after the VBS BDT, in the new samples (red discs in the illustration). To obtain the expected yields after the QCD BDT and the VBS BDT training, one applies both QCD BDT and VBS BDT efficiencies to the expected

yields after the baseline selection. Now one can calculate the signal significance as $\frac{S}{\sqrt{B}}$.

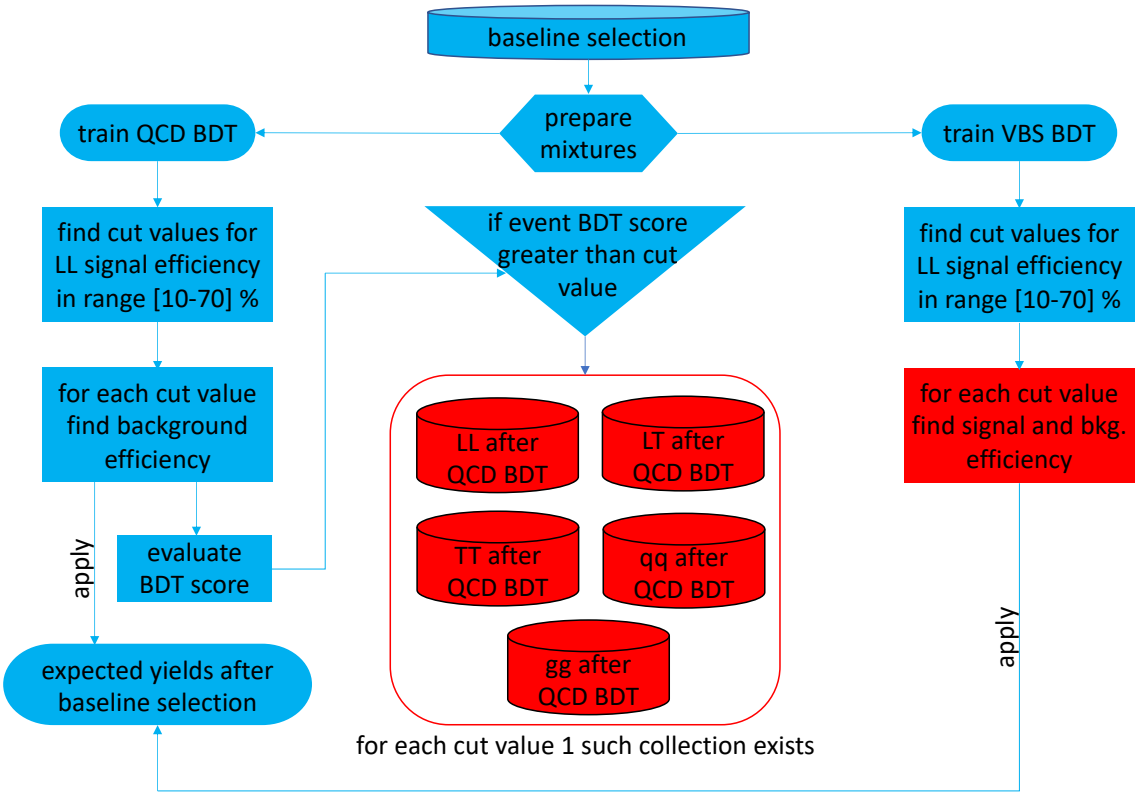


Figure 5.18: Illustration of the 2D BDT signal extraction approach. The blue colour marks data sets obtained after the baseline selection, as well as any operation applied to them. The red colour marks new data sets obtained after applying the QCD BDT training, as well as any operation applied to them.

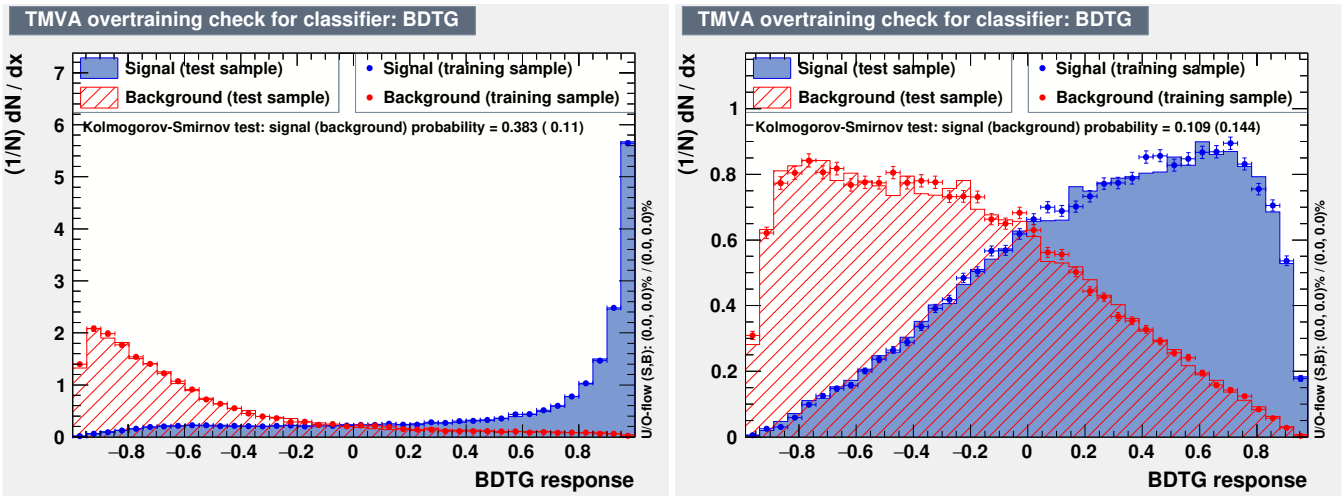


Figure 5.19: An example of the BDT output distributions of the training and test samples for the 2D BDT approach. The left-hand side plot top row shows the result of the QCD BDT training. The right-hand side plot shows the result of the VBS BDT training.

This procedure is repeated for other cut values in the VBS BDT output distribution. Finally, one will now choose several other cut values on the QCD BDT output distribution and repeat everything. This results in an array of VBS BDT signal and background efficiencies for each cut value in the QCD BDT output distribution, hence the name 2D BDT.

Because of the large cross section of the QCD background, the WP for the QCD BDT must be chosen such that it heavily suppresses the QCD contribution. The VBS QCD will further reduce the QCD contribution but at the expense of also reducing the LL signal.

5.6.2 Signal extraction and significance measurements at 14 TeV

Table 5.11 shows the number of generated events for all VBS and QCD processes at 14 TeV and for $3000\ fb^{-1}$ included in the analysis. For all contributions, the unweighted number of generated events is reported. In addition, for the VBS processes and the qq background, the number of events, weighted by the process cross section, is quoted as well. The events in the gg production are not weighted by the cross section but by unity. Very few events, less than 0.3 %, have a weight larger than 1. This is due to the setup of the computation grids where a balance between the precision and the time consumption was required. Such events have been rejected in the analysis and thus the unweighted number of events is slightly larger than the weighted number of events.

The expected number of events for the VBS and QCD qq contributions at luminosity L, expressed in inverse femtobarns, was calculated using the formula below:

$$N_{expected}^{L[fb^{-1}]} = \frac{N_{weighted}^{selection}}{N_{unweighted}^{generated}} \cdot 1000 \cdot L$$

The expected number of events for the QCD gg contribution at the luminosity L, expressed in inverse femtobarns, was calculated using the formula below:

$$N_{expected}^{L[fb^{-1}]} = 2 \cdot \sigma_{gen} \cdot \frac{N_{weighted}^{selection}}{N_{weighted}^{generated}} \cdot L\ ,$$

where the σ_{gen} is the cross section of the generated sample expressed in femtobarns. The factor 2 accounts for the fact that only $2e2\mu$ final state was simulated for the gg contribution, thus neglecting the $4e$ and 4μ final states and therefore also neglecting half of the available phase space.

Combined-background BDT

The distributions for the LL signal and the combined background, normalized to unit area, are shown in Fig. 5.20. The BDT output distributions for the training and test samples for the combined-background BDT are shown in Fig. 5.21. The BDT output distribution shows no signs of overtraining.

The top part of Table 5.12 shows the expected yields for the LL signal and all backgrounds after the baseline selection at 14 TeV for $3000\ fb^{-1}$. The bottom part shows the cut value chosen from the BDT output distribution together with the corresponding signal efficiency. For each signal efficiency, the efficiency of all contributions is reported. Table 5.13 shows expected yields corresponding to the efficiencies shown in Table 5.12 together with the signal significance for each WP.

	VBS LL	VBS LT	VBS TT	QCD qq	QCD gg
unweighted events	227731	94345	100000	502500	109731
weighted events	7.4	17.7	31.6	24942	109729
ZZ selection	3.8	9.5	18.7	11199	46189
baseline selection	3.3	8.5	16.6	2440	15080
VBS selection	2.3	5.5	10.3	353	3798
	expected yields at HL-LHC				
baseline selection	43	269	499	14569	2941
VBS selection	30	175	310	2106	741

Table 5.11: Top: unweighted and weighted number of generated events. For VBS and QCD qq processes events are weighted by the process cross section. Middle: weighted number of events after the selection. Bottom: expected number of events at 14 TeV and for $3000\ fb^{-1}$.

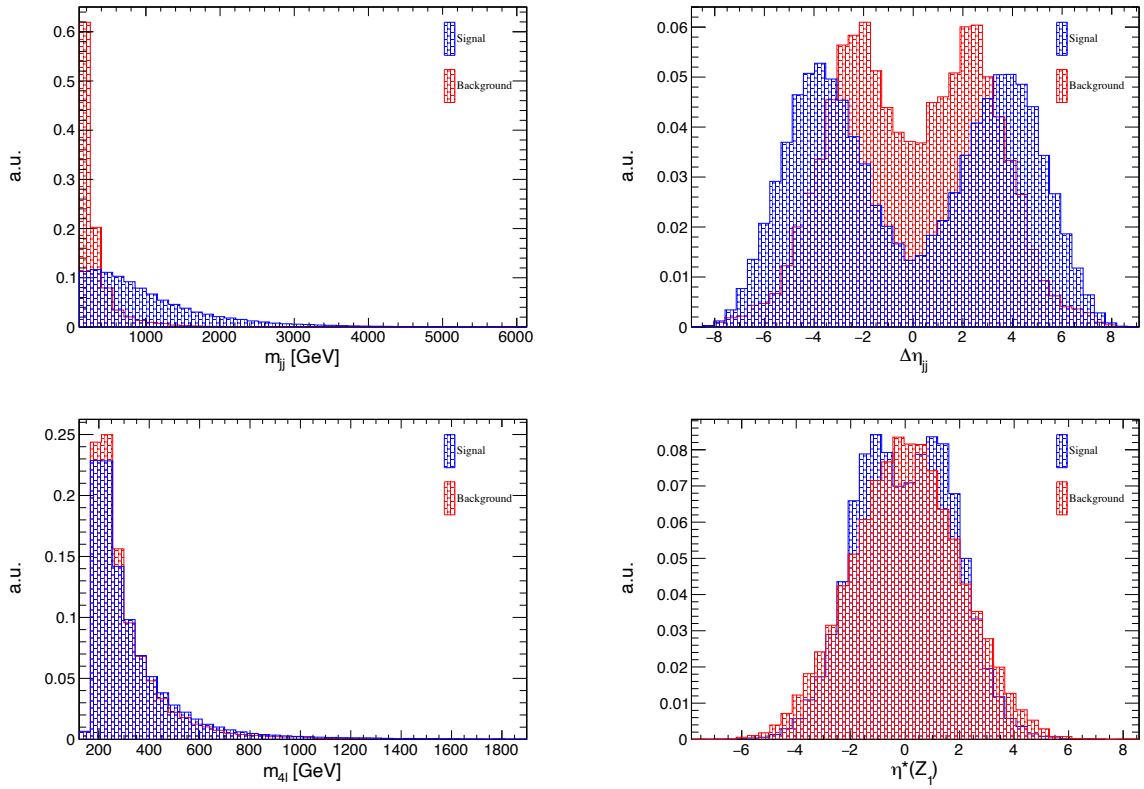


Figure 5.20

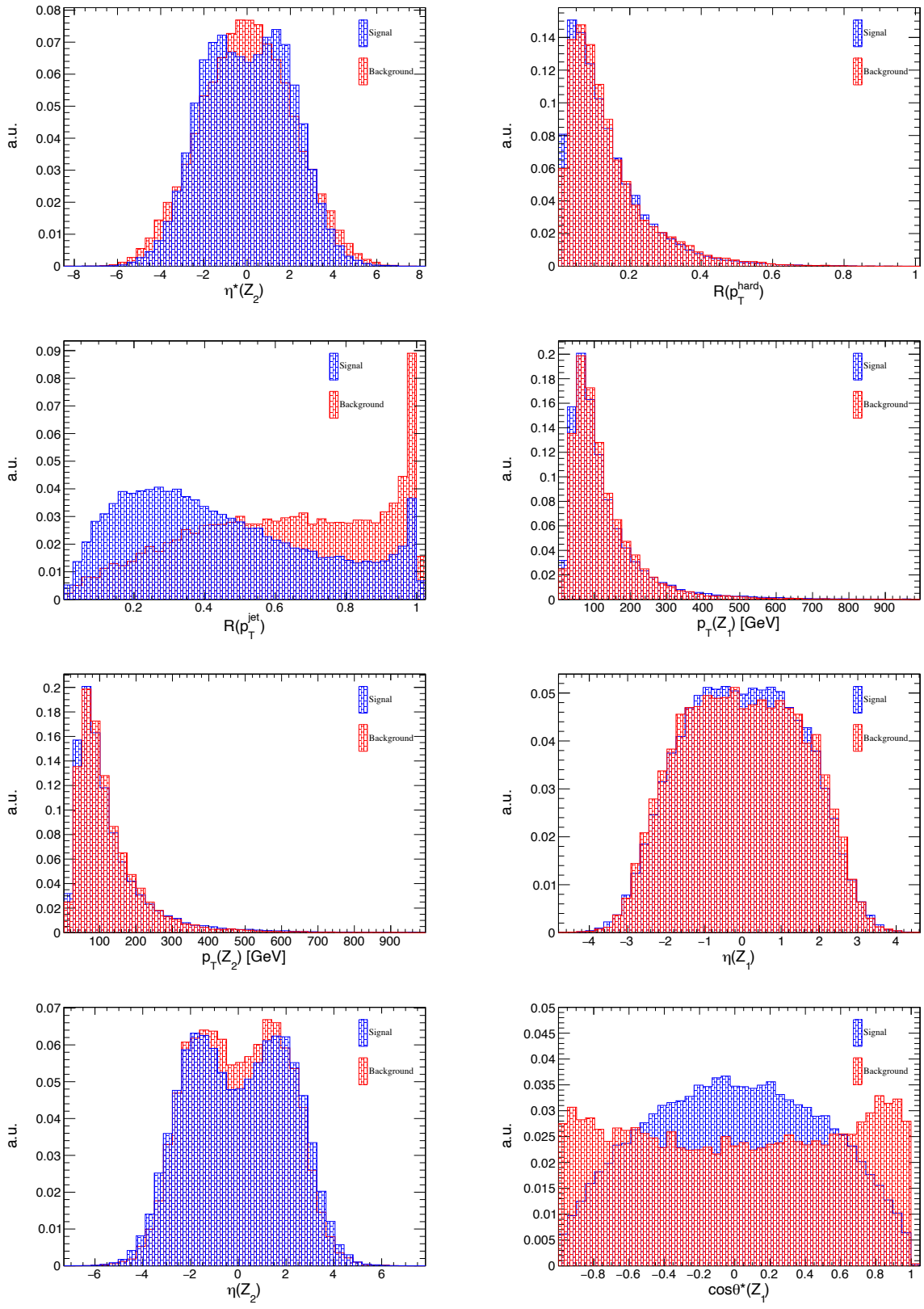

a.u.
Signal
Background
η*(Z2)
R(pTjet)
R(pThard)
pT(Z2) [GeV]
pT(Z1) [GeV]
η(Z2)
η(Z1)
cosθ*(Z1)

Figure 5.20

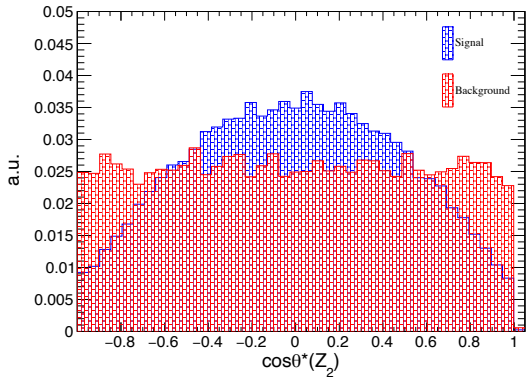


Figure 5.20: Input variables for the combined-background BDT training at 14 TeV. The *LL* signal is shown in blue and the mixture of backgrounds is in red.

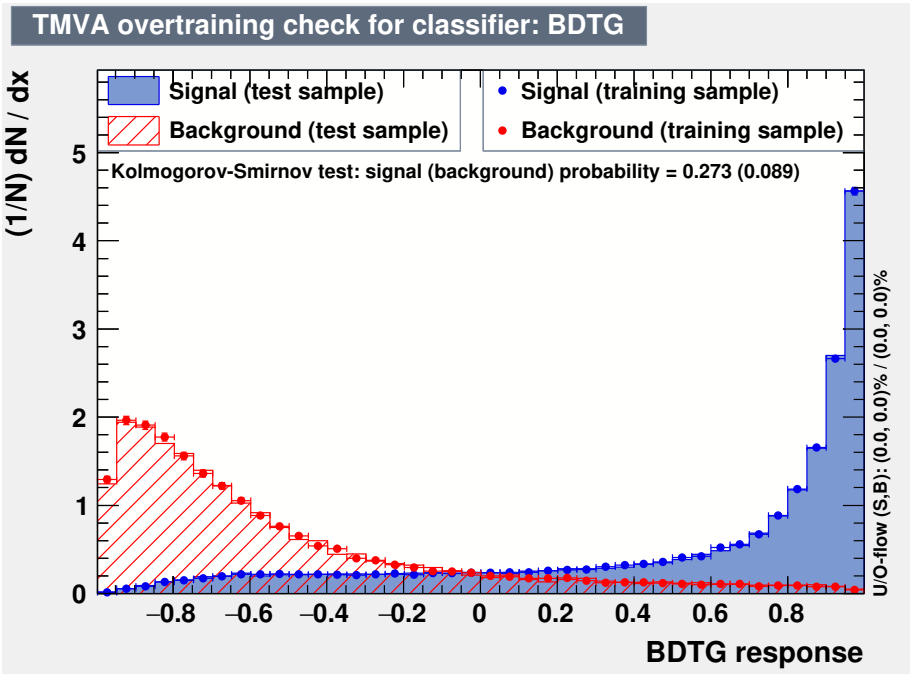


Figure 5.21: The BDT output distributions of the LL signal (in blue) and the mixture of backgrounds (in red) for the combined-background BDT training at 14 TeV.

	LL	**LT**	**TT**	**qq**	**gg**
expected yields after the baseline selection	43	269	499	14569	2941
signal efficiency [%]	**LL [%]**	**LT [%]**	**TT [%]**	**qq [%]**	**gg [%]**
45 (0.845)	45.0	33.5	22.7	0.70	3.30
40 (0.879)	40.0	28.5	18.3	0.50	2.30
35 (0.905)	35.0	23.9	14.5	0.30	1.70
30 (0.925)	30.0	19.4	11.1	0.20	1.20
20 (0.957)	20.0	11.3	5.60	0.06	0.50
15 (0. 969)	15.0	7.70	3.40	0.03	0.20

Table 5.12: Top: expected yields for the LL signal and all backgrounds after the baseline selection. Bottom: signal efficiencies and corresponding efficiencies for all contributions. Cut values corresponding to the signal efficiencies are shown in parentheses. Results are shown for the combined-background BDT training at 14 TeV and for $3000\ fb^{-1}$.

	Number of events					
signal efficiency [%]	**LL**	**LT**	**TT**	**qq**	**gg**	$S/\sqrt{B}$
45 (0.845)	19.4	90.0	113	108	97.1	0.96
40 (0.879)	17.2	76.6	91.5	81.2	69.0	0.96
35 (0.905)	15.0	64.3	72.6	48.2	49.5	0.98
30 (0.925)	12.9	52.1	55.6	29.4	35.5	0.98
20 (0.957)	8.60	30.4	28.0	9.20	13.8	0.95
15 (0.969)	6.40	20.7	16.7	5.00	6.00	0.93

Table 5.13: Expected yields for all contributions corresponding to efficiencies reported in Table 5.12. Cut values corresponding to the signal efficiencies are shown in parentheses. Results correspond to the combined-background BDT training at 14 TeV and for $3000\ fb^{-1}$.

2D BDT

Figs. 5.22 and 5.23 show the input variables for the QCD BDT and the VBS BDT. The QCD BDT output distribution for the training and test samples is shown in the top row of Fig. 5.24. The bottom row shows the same distributions for the VBS BDT training. No overtraining is observed for either case.

Table 5.14 shows the efficiencies of all contributions after the VBS BDT training for the fixed QCD BDT signal efficiency of 20%. Table 5.15 shows the expected yields corresponding to the efficiencies quoted in Table 5.14 together with the signal significance for each WP. Scanning of the 2D BDT significance space was performed to find the optimal working points for both QCD BDT and VBS BDT training. This is shown in Fig. 5.25.

A detailed discussion on the performance of the 2D BDT compared to the combined-background BDT at 14 TeV and for $3000\ fb^{-1}$ is presented in section 5.7.

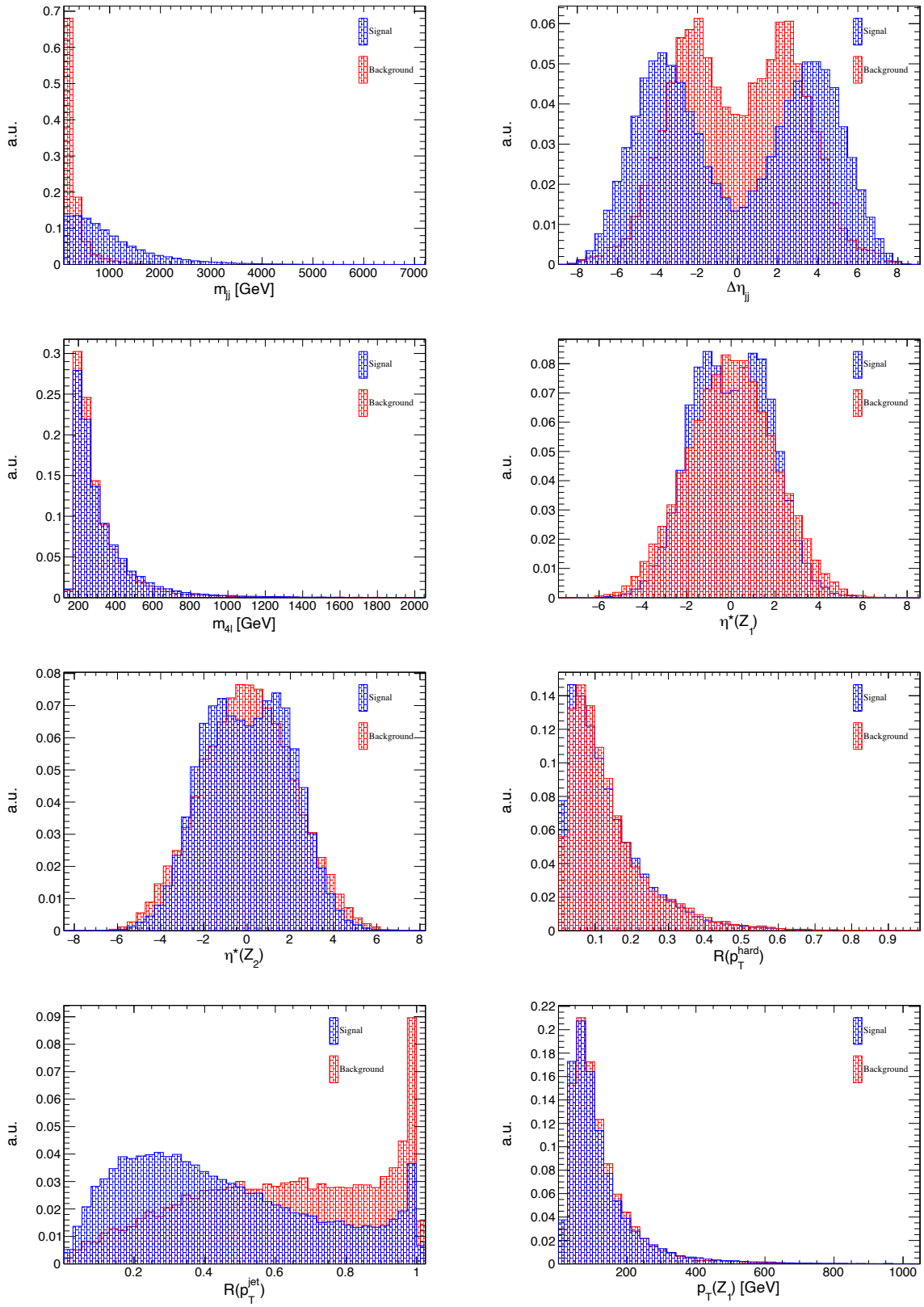

a.u.
Signal
Background
m_{jj} [GeV]
$\Delta\eta_{jj}$
m_{4l} [GeV]
$\eta^*(Z_1)$
$\eta^*(Z_2)$
$R(p_T^{hard})$
$R(p_T^{jet})$
$p_T(Z_1)$ [GeV]

Figure 5.22

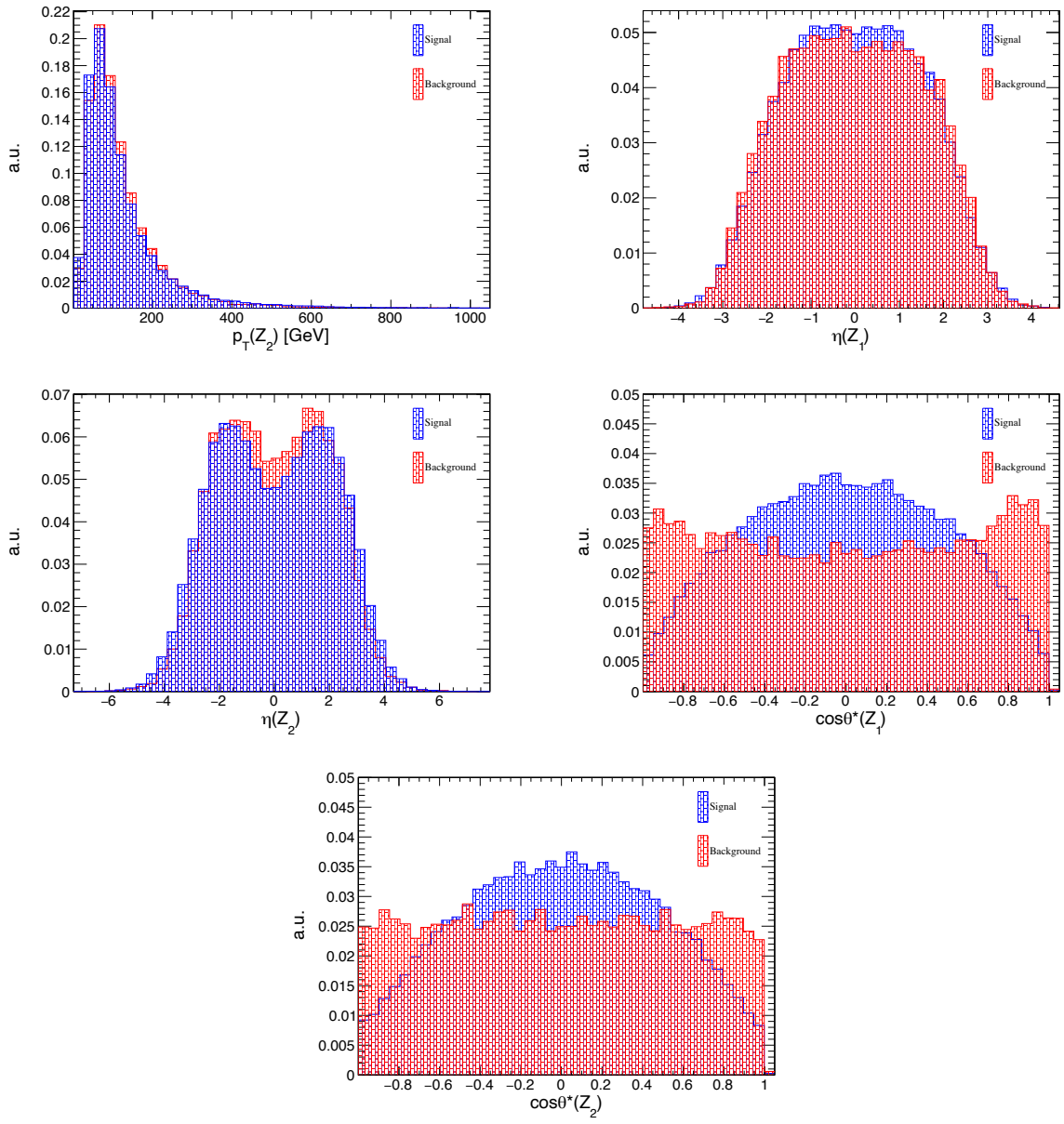


Figure 5.22: Input variables for the QCD BDT training at 14 TeV. The *LL* signal is shown in blue and the qq background is shown in red.

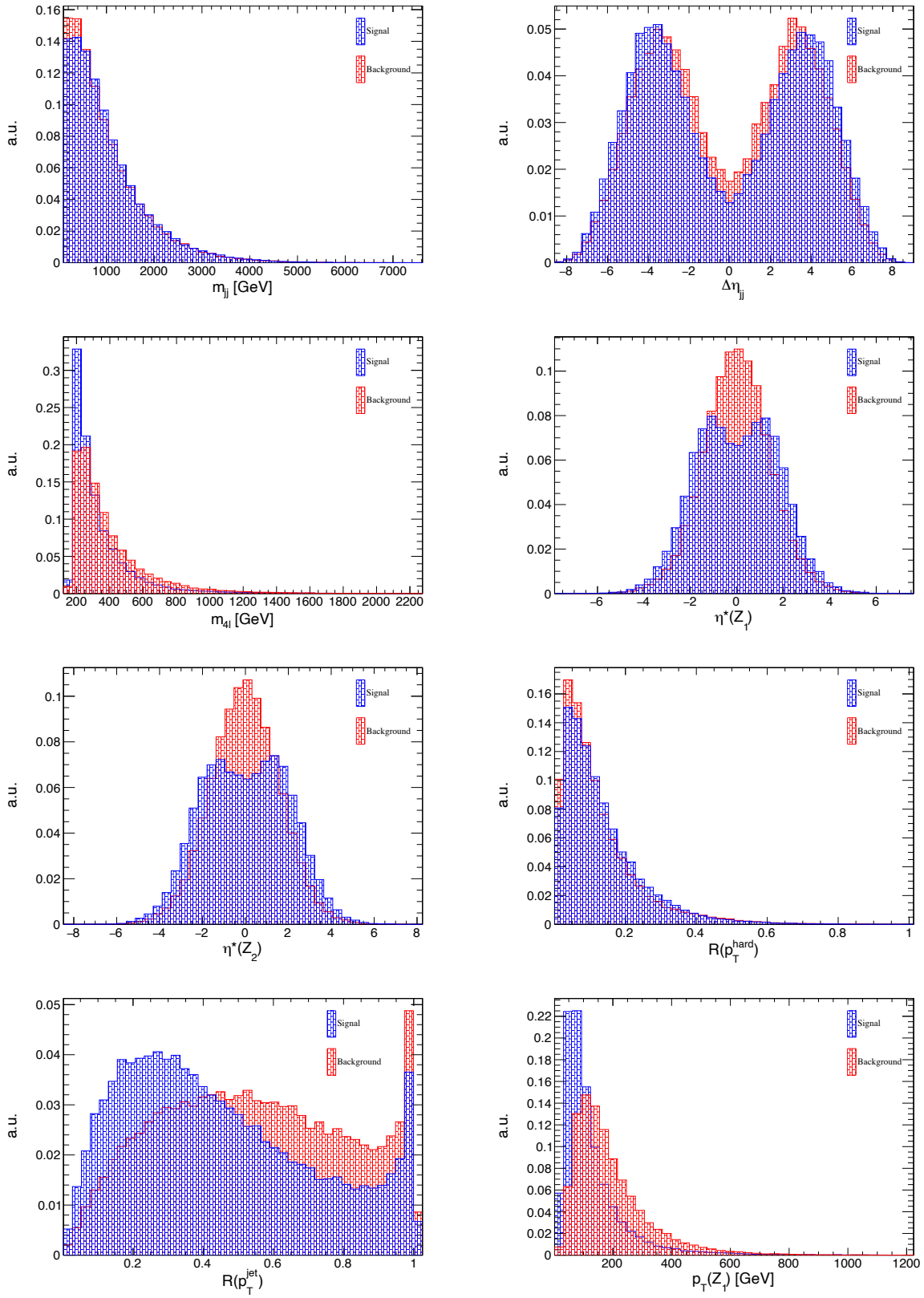

a.u.
Signal
Background
m_{jj} [GeV]
$\Delta\eta_{jj}$
m_{4l} [GeV]
$\eta^*(Z_1)$
$\eta^*(Z_2)$
$R(p_T^{hard})$
$R(p_T^{jet})$
$p_T(Z_1)$ [GeV]

Figure 5.23

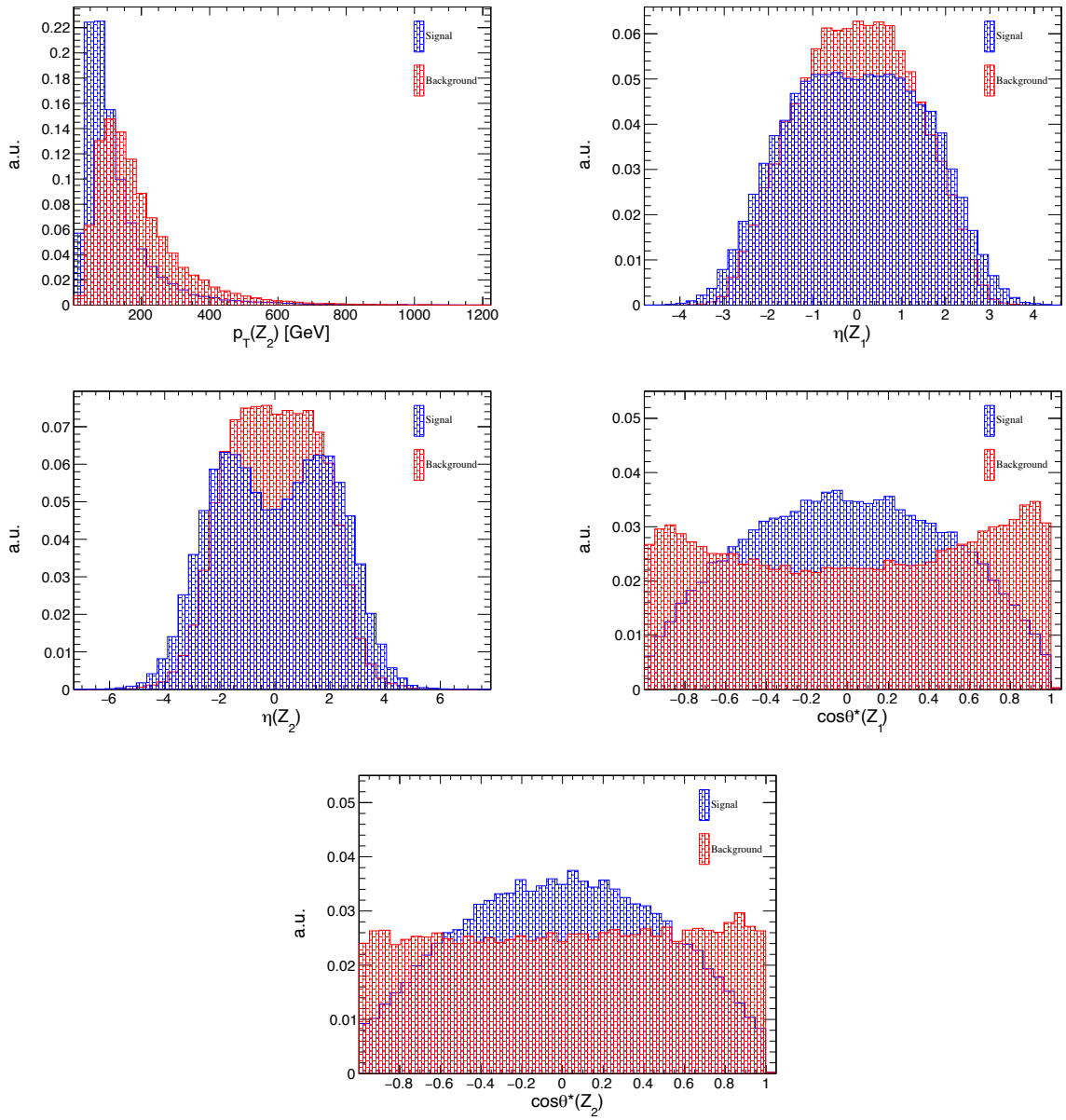


Figure 5.23: Input variables for the VBS BDT training at 14 TeV. The *LL* signal is shown in blue and the mixture of the LT and TT backgrounds is shown in red.

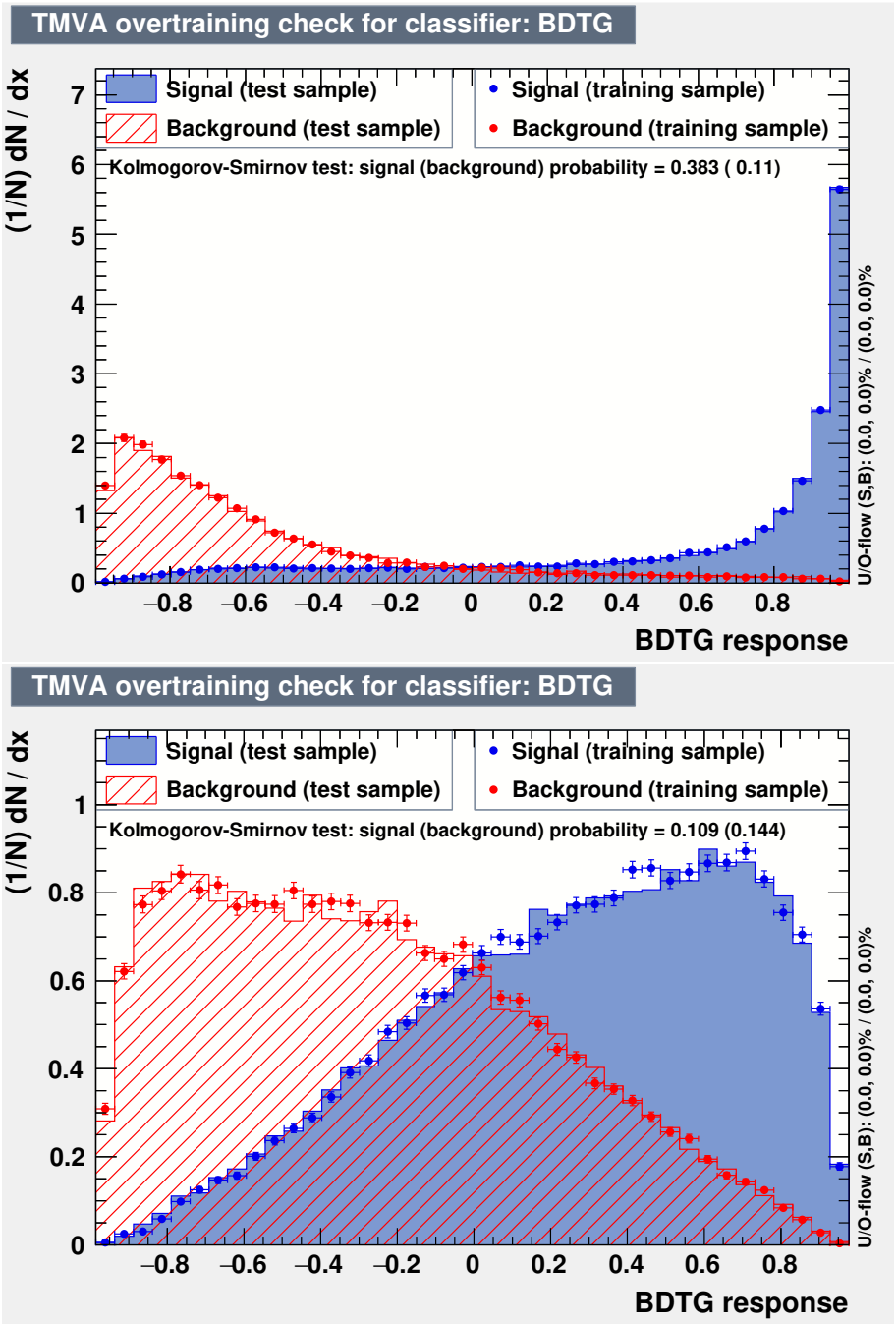


Figure 5.24: Top: the QCD BDT output distribution of the LL signal (in blue) and the qq background (in red) for the 2D BDT training at 14 TeV. Bottom: the VBS BDT output distribution of the LL signal (in blue) and the mixture of the LT and TT backgrounds (in red).

	LL [%]	**LT [%]**	**TT [%]**	**qq [%]**	**gg [%]**
efficiencies after the QCD BDT ($\epsilon_{sig} = 25\ \%$)	25.0	17.2	11.1	0.109	1.05
VBS BDT signal efficiencies [%]	**efficiencies after VBS BDT for QCD BDT** $\epsilon_{signal} = 25\ \%$				
	LL [%]	**LT [%]**	**TT [%]**	**qq [%]**	**gg [%]**
50 % (0.314)	70.9	36.8	14.2	35.1	24.7
45 % (0.377)	66.2	32.1	11.5	27.0	20.3
40 % (0.438)	61.3	27.7	8.81	21.6	13.9
35 % (0.498)	55.9	23.5	6.72	16.2	12.0
30 % (0.558)	50.3	19.4	4.90	13.5	10.8
25 % (0.615)	44.1	15.3	3.81	13.5	6.96

Table 5.14: Top: efficiencies of the LL signal and all backgrounds after the QCD BDT. The QCD BDT signal efficiency is fixed at 25 %. Bottom: signal efficiencies and corresponding background efficiencies after the VBS BDT for the 25 % QCD BDT signal efficiency. Several signal efficiencies, corresponding to the working points in the bottom-left plot in Fig. 5.24, were scanned to find the maximum signal significance. Cut values corresponding to the signal efficiencies are shown in parentheses. Results are shown for the 2D BDT training at 14 TeV and for $3000\ fb^{-1}$.

	LL	**LT**	**TT**	**qq**	**gg**	$S/\sqrt{B}$
expected yields after the QCD BDT ($\epsilon_{sig} = 25\ \%$)	10.8	46.4	55.3	15.9	30.8	0.88
VBS BDT signal efficiencies [%]	**expected yields after 2D BDT for QCD BDT** $\epsilon_{signal} = 25\ \%$					
	LL	**LT**	**TT**	**qq**	**gg**	$S/\sqrt{B}$
50 % (0.314)	7.60	17.1	7.90	5.60	7.60	1.24
45 % (0.377)	7.10	14.9	6.30	4.30	6.20	1.26
40 % (0.438)	6.60	12.8	4.90	3.40	4.30	1.31
35 % (0.498)	6.00	10.9	3.70	2.60	3.70	1.32
30 % (0.558)	5.40	9.00	2.70	2.10	3.30	1.30
25 % (0.615)	4.70	7.10	2.10	2.10	2.10	1.29

Table 5.15: Expected yields for all contributions corresponding to efficiencies reported in Table 5.14. Cut values corresponding to the signal efficiencies are shown in parentheses. Results are shown for the 2D BDT training at 14 TeV and for $3000\ fb^{-1}$.

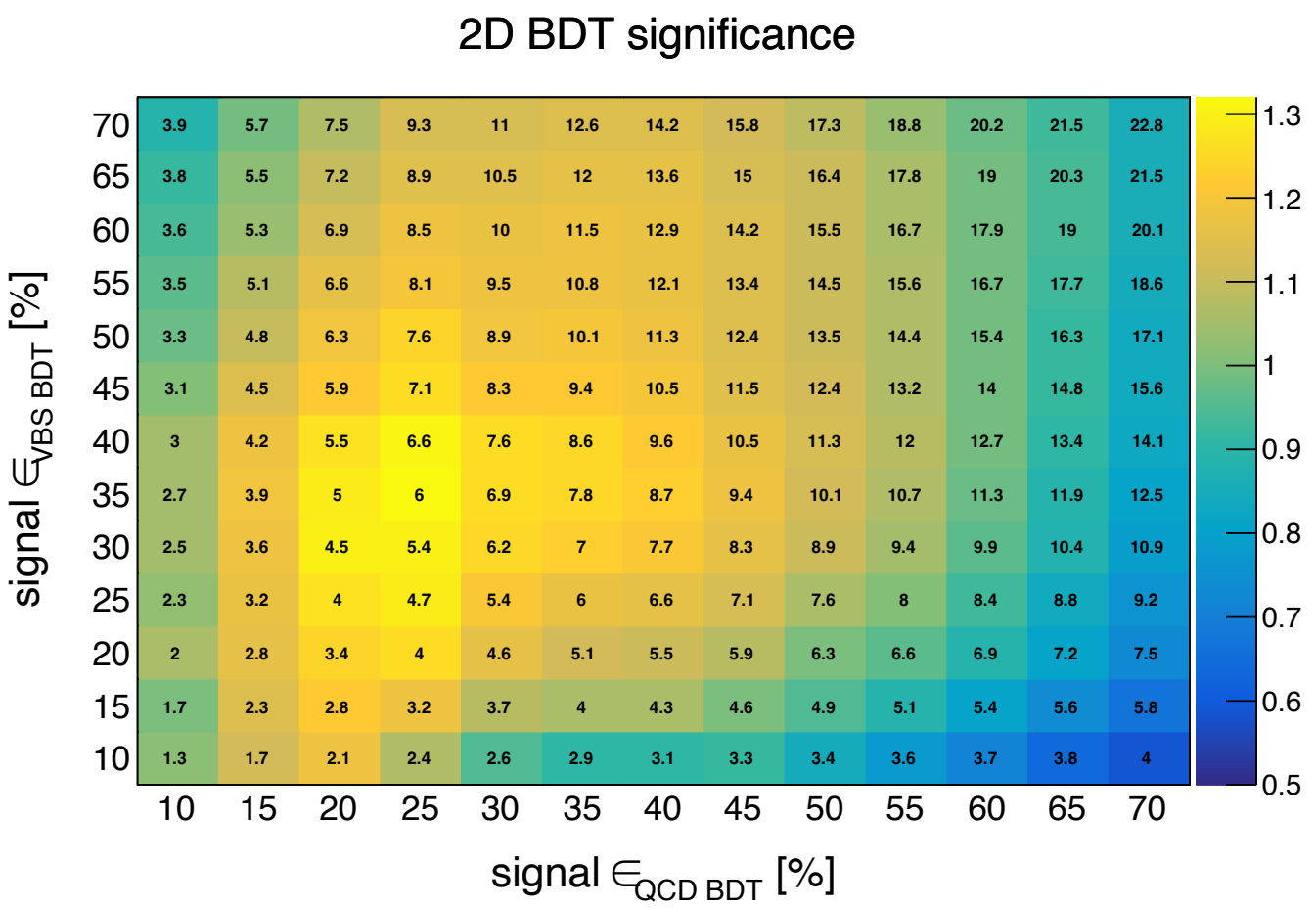


Figure 5.25: The 2D BDT significance plane used to scan for the optimal WP of the QCD BDT and the VBS BDT. Results at 14 TeV and for $3000\ fb^{-1}$ are shown. Brighter colours reflect the higher signal significance, while the numbers inside each bin show the expected LL yields.

5.6.3 Signal extraction and significance measurements at 27 TeV

Table 5.16 shows the number of generated events for all VBS and QCD processes at 27 TeV and for $15000\ fb^{-1}$ included in the analysis. The number of weighted events after each selection, together with the expected yields are shown.

Combined-background BDT

Distributions for the LL signal and the combined background, normalized to unit area, are shown in Fig. 5.26. The BDT output distributions for the training and test samples for the combined-background BDT are shown in Fig. 5.27. The BDT output distribution shows no signs of overtraining.

The top part of Table 5.17 shows the expected yields for the LL signal and all backgrounds after the baseline selection at 27 TeV for $15000\ fb^{-1}$. The bottom part shows the cut value chosen from the BDT output distribution together with the corresponding signal efficiency. For each signal efficiency, the efficiency of all contributions is reported. Table 5.18 shows expected yields corresponding to the efficiencies shown in Table 5.17 together with the signal significance for each WP.

	VBS LL	VBS LT	VBS TT	QCD qq	QCD gg
unweighted events	218909	80539	100000	592560	109632
weighted events	25.3	53.6	114.0	66747	109632
ZZ selection	11.2	25.3	60.0	24440	44876
baseline selection	9.70	22.8	54.5	7386	18264
VBS selection	6.94	16.0	36.8	1235	5211
	expected yields at HE-LHC				
baseline selection	664	4241	8178	187731	54503
VBS selection	576	2974	5521	31472	15551

Table 5.16: Top: unweighted and weighted number of generated events. For the VBS and QCD qq processes, events are weighted by the process cross section. Middle: weighted number of events after the selection. Bottom: expected number of events at 27 TeV and for $15000\ fb^{-1}$.

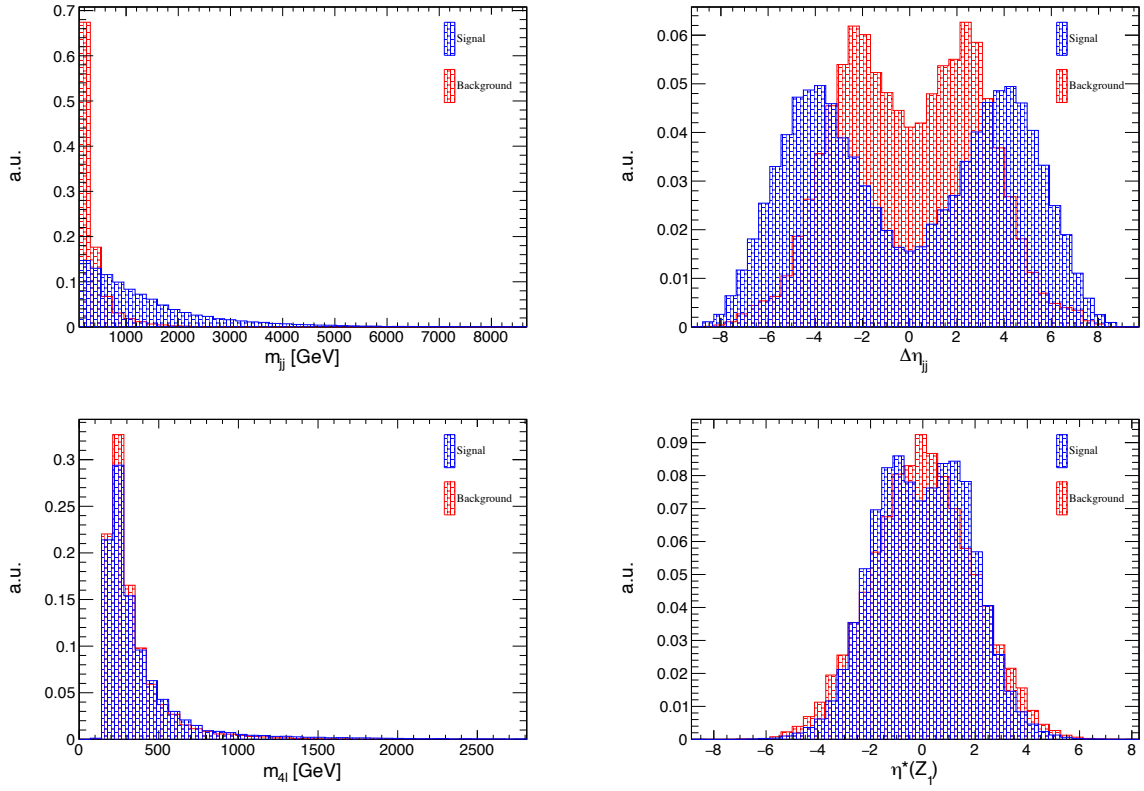


Figure 5.26

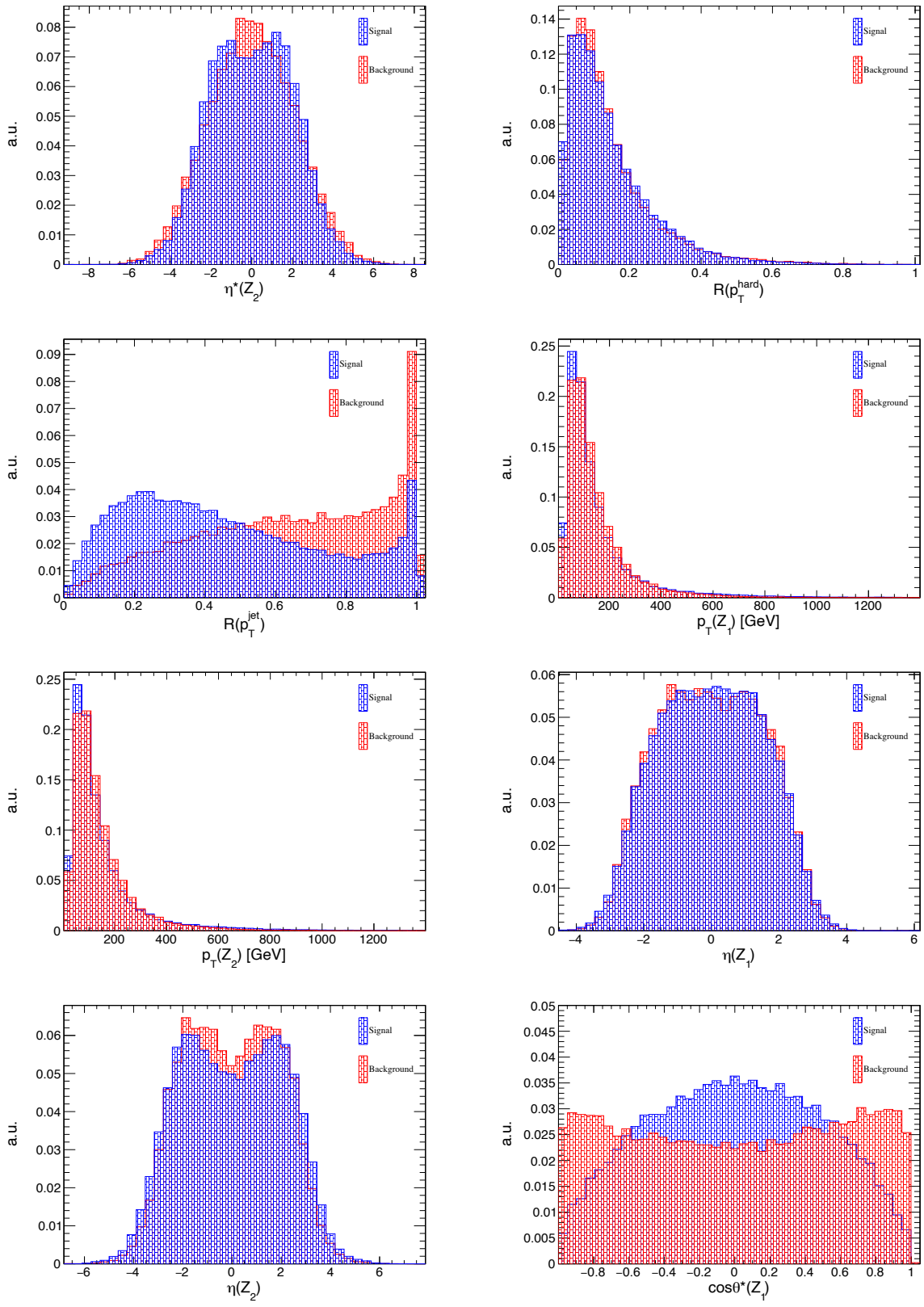

a.u.
Signal
Background
η*(Z2)
R(pTjet)
R(pThard)
pT(Z2) [GeV]
pT(Z1) [GeV]
η(Z2)
η(Z1)
cosθ*(Z1)

Figure 5.26

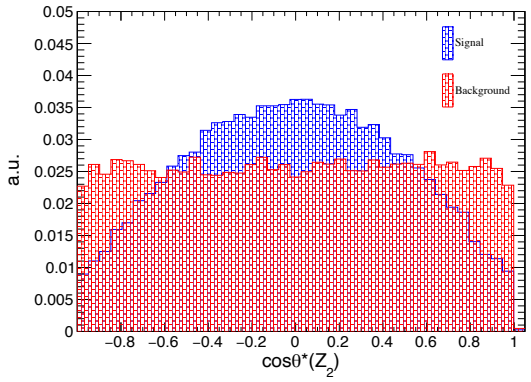


Figure 5.26: Input variables for the combined-background BDT training at 27 TeV. The *LL* signal is shown in blue and the mixture of backgrounds is in red.

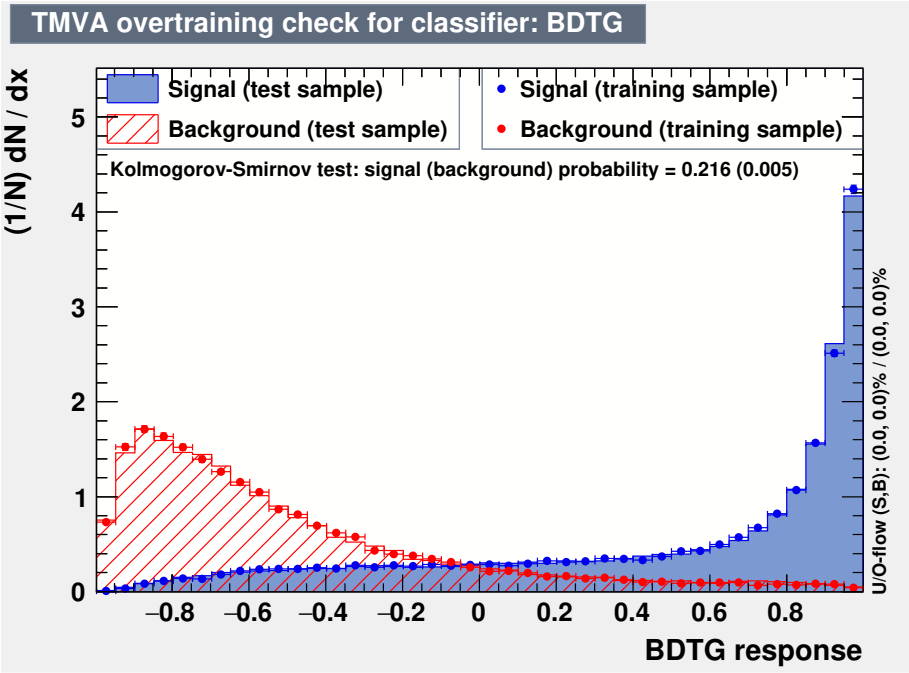


Figure 5.27: The BDT output distributions of the LL signal (in blue) and the mixture of backgrounds (in red) for the combined-background BDT training at 27 TeV.

	LL	**LT**	**TT**	**qq**	**gg**
expected yields after the baseline selection	664	4241	8178	187731	54503
signal efficiency [%]	**LL [%]**	**LT [%]**	**TT [%]**	**qq [%]**	**gg [%]**
45 (0.819)	45.0	36.6	27.6	0.86	3.43
40 (0.861)	40.0	31.4	22.7	0.59	2.43
35 (0.893)	35.0	26.3	18.4	0.38	1.76
30 (0.917)	30.0	21.5	14.4	0.23	1.19
20 (0.952)	20.0	13.1	7.67	0.09	0.52
15 (0.965)	15.0	9.32	5.03	0.04	0.27

Table 5.17: The: expected yields for the LL signal and all backgrounds after the baseline selection. Bottom: signal efficiencies and corresponding efficiencies for all contributions. Cut values corresponding to the signal efficiencies are shown in parentheses. Results are shown for the combined-background BDT training at 27 TeV and for $15000\ fb^{-1}$.

	Number of events					
signal efficiency [%]	**LL**	**LT**	**TT**	**qq**	**gg**	$S/\sqrt{B}$
45 (0.819)	299	1553	2256	1615	1868	3.50
40 (0.861)	266	1332	1857	1115	1322	3.54
35 (0.893)	232	1116	1505	705	961	3.55
30 (0.917)	199	914	1174	440	648	3.53
20 (0.952)	133	557	627	172	284	3.28
15 (0.965)	99.6	395	411	70.9	146	3.11

Table 5.18: Expected yields for all contributions corresponding to efficiencies reported in Table 5.17. Cut values corresponding to the signal efficiencies are shown in parentheses. Results are shown for the combined-background BDT training at 27 TeV and for $15000\ fb^{-1}$.

2D BDT

Figs. 5.28 and 5.29 show the input variables for the QCD BDT and the VBS BDT. The QCD BDT output distribution for the training and test samples is shown in the top row of Fig. 5.30. The bottom row shows the same distributions for the VBS BDT training. No overtraining is observed for either case.

Table 5.19 shows the efficiencies of all contributions after the VBS BDT training for the fixed QCD BDT signal efficiency of 40%. Table 5.20 shows the expected yields corresponding to the efficiencies quoted in Table 5.19 together with the signal significance for each WP. Once more, scanning of the 2D BDT significance space was performed to find the optimal working points for both QCD BDT and VBS BDT training. This is shown in Fig. 5.31.

A detailed discussion on the performance of the 2D BDT compared to the combined-background BDT at 27 TeV and for $15000\ fb^{-1}$ is presented in the next section.

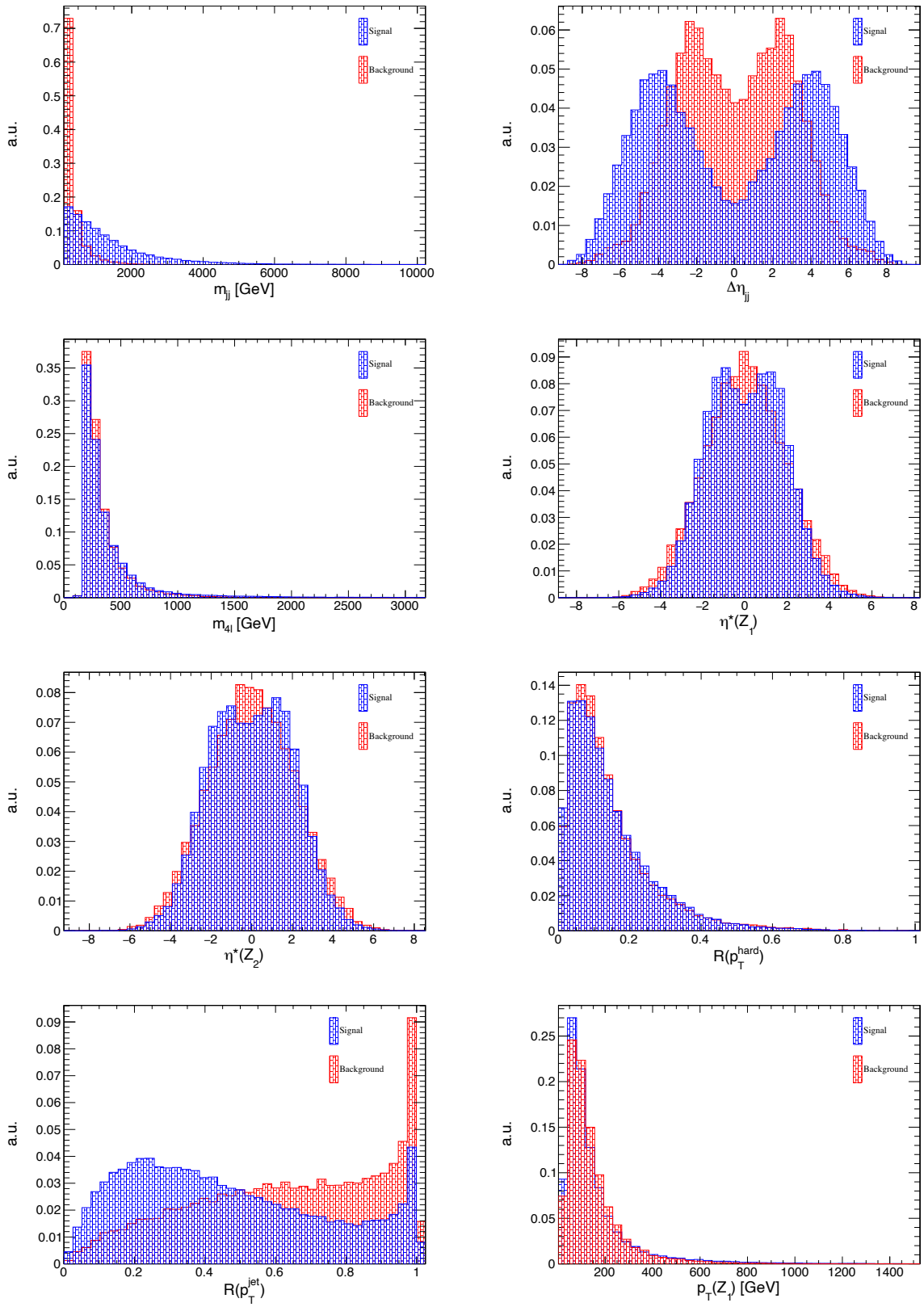

a.u.
Signal
Background
m_{jj} [GeV]
$\Delta\eta_{jj}$
m_{4l} [GeV]
$\eta^*(Z_1)$
$\eta^*(Z_2)$
$R(p_T^{hard})$
$R(p_T^{jet})$
$p_T(Z_1)$ [GeV]

Figure 5.28

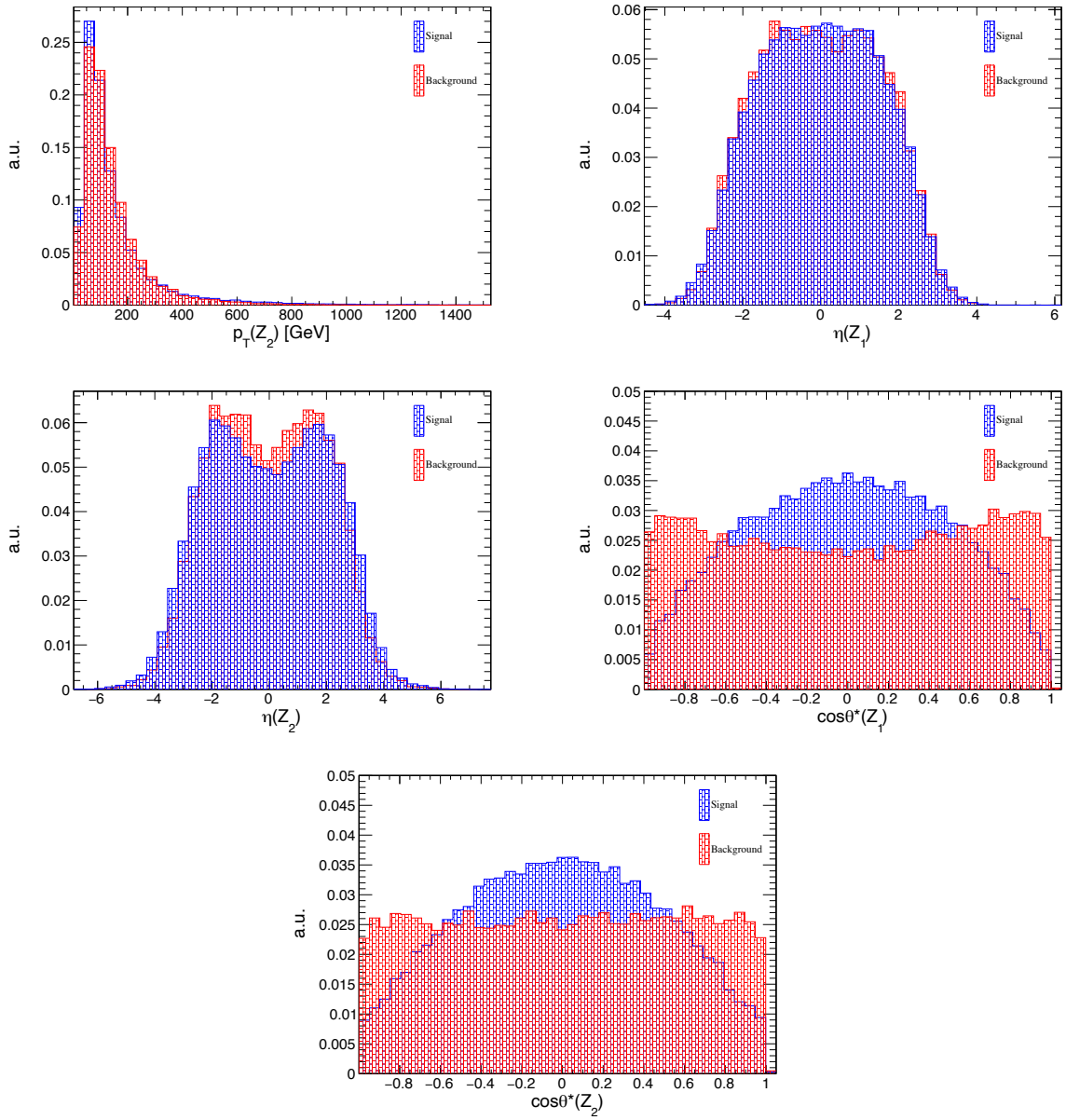


Figure 5.28: Input variables for the QCD BDT training at 27 TeV. The *LL* signal is shown in blue and the qq background is shown in red.

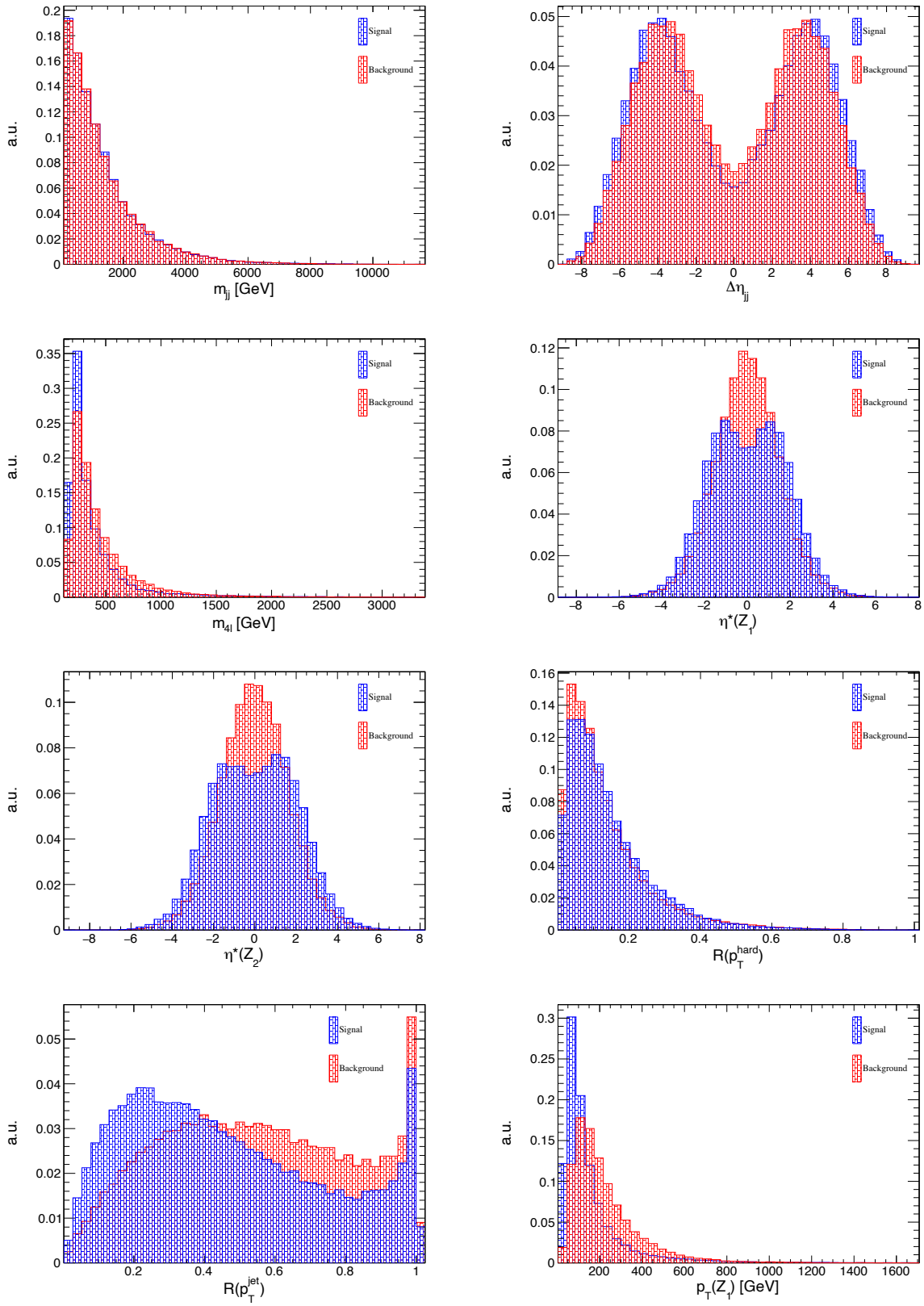

a.u.
Signal
Background
m_{jj} [GeV]
$\Delta\eta_{jj}$
m_{4l} [GeV]
$\eta^*(Z_1)$
$\eta^*(Z_2)$
$R(p_T^{hard})$
$R(p_T^{jet})$
$p_T(Z_1)$ [GeV]

Figure 5.29

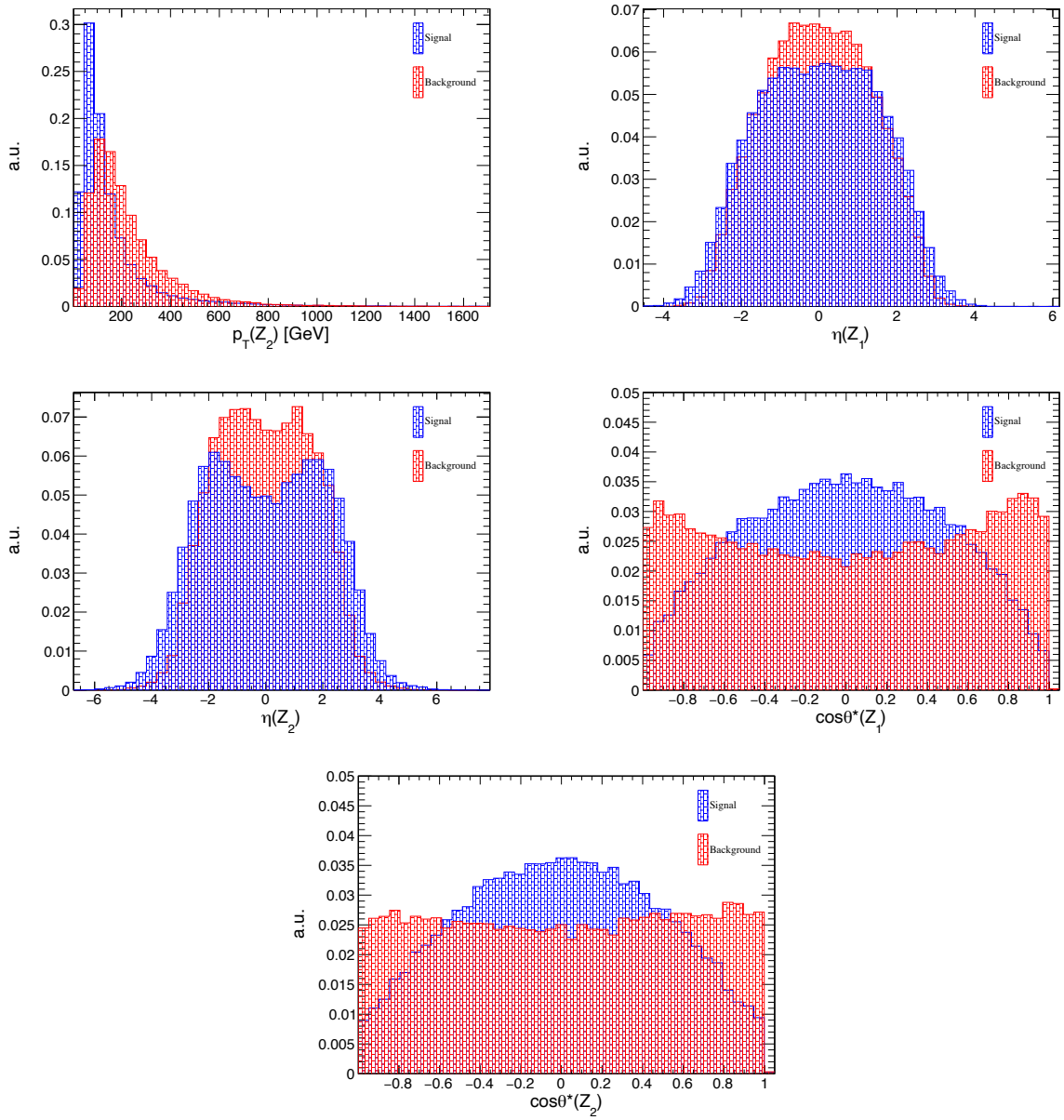


Figure 5.29: Input variables for the VBS BDT training at 27 TeV. The *LL* signal is shown in blue and the mixture of the LT and TT backgrounds is shown in red.

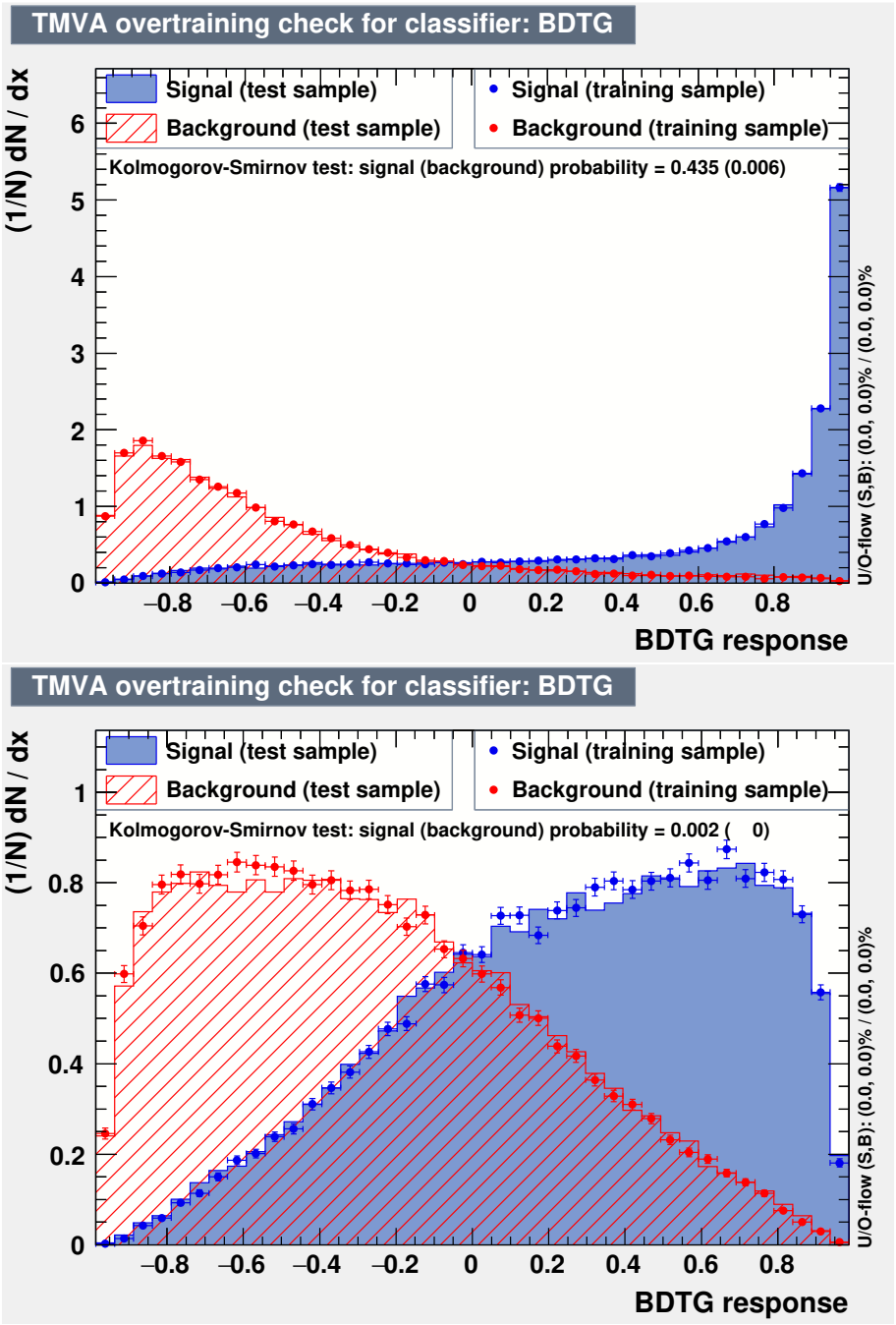


Figure 5.30: Top: the QCD BDT output distribution of the LL signal (in blue) and the qq background (in red) for the 2D BDT training at 27 TeV. Bottom: the VBS BDT output distribution of the LL signal (in blue) and the mixture of the LT and TT backgrounds (in red).

	LL [%]	**LT [%]**	**TT [%]**	**qq [%]**	**gg [%]**
efficiencies after the QCD BDT ($\epsilon_{sig} = 35\ \%$)	35.0	28.4	21.6	0.36	2.07
VBS BDT signal efficiencies [%]	**efficiencies after VBS BDT for QCD BDT** $\epsilon_{signal} = 35\ \%$				
	LL [%]	**LT [%]**	**TT [%]**	**qq [%]**	**gg [%]**
50 % (0.309)	64.3	31.3	11.2	41.0	19.0
45 % (0.374)	59.4	26.6	8.74	35.5	14.3
40 % (0.438)	54.3	22.0	6.96	32.2	11.4
35 % (0.500)	48.8	17.8	5.20	28.4	8.20
30 % (0.562)	43.0	14.2	3.86	26.2	5.56
25 % (0.622)	36.7	10.9	2.72	21.3	4.76

Table 5.19: Top: efficiencies of the LL signal and all backgrounds after the QCD BDT. The QCD BDT signal efficiency is fixed at 35 %. Bottom: signal efficiencies and corresponding background efficiencies after the VBS BDT for the 35 % QCD BDT signal efficiency. Several signal efficiencies, corresponding to the working points in the bottom-left plot in Fig. 5.30, were scanned to find the maximum signal significance. Cut values corresponding to the signal efficiencies are shown in parentheses. Results are obtained at 27 TeV and for $15000\ fb^{-1}$.

	LL	**LT**	**TT**	**qq**	**gg**	$S/\sqrt{B}$
expected yields after the QCD BDT ($\epsilon_{sig} = 35\ \%$)	232	1203	1764	682	1128	3.36
VBS BDT signal efficiencies [%]	**expected yields after 2D BDT for QCD BDT** $\epsilon_{signal} = 35\ \%$					
	LL	**LT**	**TT**	**qq**	**gg**	$S/\sqrt{B}$
50 % (0.309)	149.4	376.1	197.7	279.7	214.9	4.57
45 % (0.374)	138.0	320.2	154.1	242.4	161.1	4.66
40 % (0.438)	126.1	264.7	122.8	220.0	128.3	4.65
35 % (0.500)	113.4	214.3	91.70	193.9	92.50	4.66
30 % (0.562)	99.90	171.4	68.10	179.0	62.70	4.55
25 % (0.622)	85.30	131.3	47.90	145.4	53.70	4.38

Table 5.20: Expected yields for all contributions corresponding to efficiencies quoted in Table 5.19. Cut values corresponding to the signal efficiencies are shown in parentheses. Results are shown for the 2D BDT training at 27 TeV and for $15000\ fb^{-1}$.

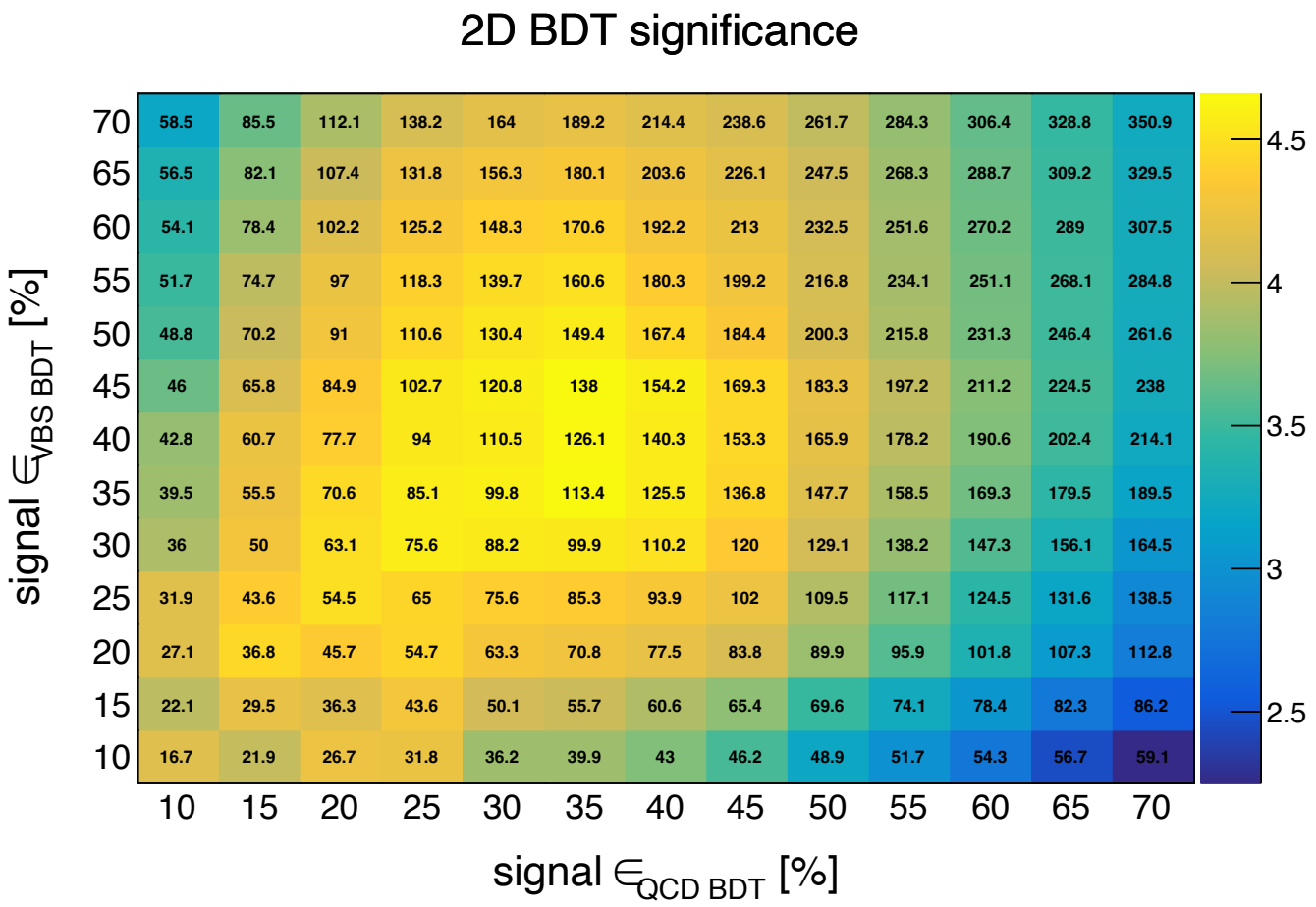


Figure 5.31: The 2D BDT significance plane used to scan for the optimal WP of the QCD BDT and the VBS BDT. Results at 27 TeV and for $15000\ fb^{-1}$ are shown. Brighter colours reflect the higher signal significance, while the numbers inside each bin show the expected LL yields.

5.7 Results

The significance of the LL signal at 14 and 27 TeV obtained using the combined-background BDT and the 2D BDT method is shown in Table 5.21. Events after the baseline selection were used as the foundation for the multivariate analysis. Signal significance with the inclusion of the HF nose option and the gain when moving from the combined-background BDT to the 2D BDT approach is also reported.

Using the combined-background BDT at 14 TeV, the confidence on the LL signal measurement is expected to reach the 1σ level. This can be improved using the 2D BDT method with the significance of the LL signal reaching 1.3σ. Extending the η acceptance for electrons up to 4 is expected to increase the significance to 1.4σ with the 2D BDT approach. The gain in LL sensitivity when moving from the combined-background BDT to the 2D BDT at HL-LHC is expected to be around 34 %.

At 27 TeV, the LL signal is expected to be measured with a significance of 4.7σ using the 2D BDT method. This is an improvement of roughly 31 % with respect to the simpler combined-background BDT. Most of the gain at 27 TeV, with respect to 14 TeV, comes from an increased luminosity which enables harder suppression of the QCD background. The importance of the HF nose option is especially noticeable at HE-LHC where the 5.4σ significance on the VBS LL measurement is expected if the 2D BDT approach is employed. This is due to the more forward kinematics at 27 TeV and $15000\ fb^{-1}$, with respect to 14 TeV and $3000\ fb^{-1}$.

5.8 Summary

Signal and background processes were simulated using MG5, MCFM, Delphes and Pythia8 tools at 14 and 27 TeV c.o.m. energies. Special care was given to jets since they dominate the final state and define the signal. The effect of parton showers on the leading jets was studied to make sure they are not affecting the identification of the tagging jets and thus making analysis unstable. With an increased luminosity, at HL- and HE-LHC conditions, the importance of pileup will increase as well. Thus, the effect of pile-up on the leading jets was studied as well. Both parton showers and pile-up were found to affect the leading jets at 10 % level.
Since no single kinematic variable is discriminating enough to separate individual polarizations, a multivariate approach was devised. Two such approaches were tested on both HL- and HE-LHC samples. The simpler of the two, the combined-background BDT, trained the LL signal against the proper mixture of VBS and QCD backgrounds to find the WP that maximizes signal sensitivity. The second approach exploits the difference in the kinematics of VBS and QCD processes to simultaneously train the BDT to separate the LL signal from the qq background and the LL signal from the mixture of the LT and TT backgrounds. This approach is referred to as the 2D BDT and is superior between the two. When compared to the simple cut-and-count approach (see expected yields after VBS selection quoted in Tables 5.11 and 5.16), as much as 160 % (120 %) can be gained in terms of signal sensitivity by exploiting the 2D BDT at 14 TeV and $3000\ fb^{-1}$ (27 TeV and $15000\ fb^{-1}$). Without the HF nose option, the LL signal is expected to be measured at 1.3σ (4.7σ) confidence level at 14 (27) TeV. Extending the lepton acceptance from $\eta = 3$ to $\eta = 4$, which corresponds to the HF nose upgrade, will increase the signal significance to 1.4σ (5.4σ) at 14 (27) TeV. These prospective studies show the great potential of the HE-LHC in observing the longitudinal component of the Z boson in the VBS $ZZ \to 4l2j$ channel.

	event counts at HL-LHC without the HF nose upgrade						
	LL	**LT**	**TT**	**qq**	**gg**	$S/\sqrt{B}$	**gain**
Combined-background BDT with $\epsilon_{signa} = 35\ \%$	15.1	64.3	72.6	48.2	49.5	0.98	34.7 %
2D BDT with $\epsilon_{signal}^{QCD\ BDT} = 25\ \%$ & $\epsilon_{signal}^{VBS\ BDT} = 35\ \%$	6.00	10.9	3.70	2.60	3.70	1.32	
	event counts at HL-LHC with the HF nose upgrade						
	LL	**LT**	**TT**	**qq**	**gg**	$S/\sqrt{B}$	**gain**
Combined-background BDT with $\epsilon_{signa} = 35\ \%$	16.1	68.2	77.1	47.9	48.6	1.04	33.7 %
2D BDT with $\epsilon_{signal}^{QCD\ BDT} = 20\ \%$ & $\epsilon_{signal}^{VBS\ BDT} = 35\ \%$	4.90	7.50	2.30	0.90	1.80	1.39	
	event counts at HE-LHC without the HF nose upgrade						
	LL	**LT**	**TT**	**qq**	**gg**	$S/\sqrt{B}$	**gain**
Combined-background BDT with $\epsilon_{signa} = 35\ \%$	232.4	1116	1501	704.8	960.9	3.55	31.3 %
2D BDT with $\epsilon_{signal}^{QCD\ BDT} = 35\ \%$ & $\epsilon_{signal}^{VBS\ BDT} = 45\ \%$	138.0	320.2	154.1	242.4	161.1	4.66	
	event counts at HE-LHC with the HF nose upgrade						
	LL	**LT**	**TT**	**qq**	**gg**	$S/\sqrt{B}$	**gain**
Combined-background BDT with $\epsilon_{signa} = 40\ \%$	293.2	1414	1888	1140.5	1310	3.87	38.2 %
2D BDT with $\epsilon_{signal}^{QCD\ BDT} = 30\ \%$ & $\epsilon_{signal}^{VBS\ BDT} = 40\ \%$	123.2	224.2	83.60	158.4	63.60	5.35	

Table 5.21: Event counts and corresponding signal significances for the combined-background BDT and the 2D BDT training at 14 and 27 TeV with an integrated luminosity of $3000\ fb^{-1}$ and $15000\ fb^{-1}$, respectively. Presented working points for both BDT training approaches give the most sensitive LL measurement. The gain using the 2D BDT compared to the simple combined-background BDT is also reported. The table shows both the results with and without the HF nose option.

Chapter 6

Conclusion and future prospects

Studying vector boson scattering (VBS) is a key element for understanding the mechanism of electroweak (EWK) symmetry breaking responsible for the generation of gauge boson masses through the Brout-Englert-Higgs mechanism. In addition, VBS is used as a tool to investigate the non-Abelian structure of the EWK sector of the Standard Model (SM) through a study of quartic gauge couplings. Finally, possible indications of physics beyond the SM can be found by probing anomalous quartic gauge couplings (aQGC) in the effective field theory (EFT) framework. This thesis reports the study of VBS in the fully leptonic $ZZ \to 4l2j$ channel.

Measurement of electron selection efficiency and derivation of scale factors for the full Run 2 period was done using the Tag and Probe method. Special care was given to reducing uncertainties in the low-p_T region and studying η structure in scale factors. These results were used in the VBS $ZZ \to 4l2j$ analysis presented in this thesis as well as in the publication of the $H \to ZZ \to 4l$ analysis with full Run 2 data. In addition, they are currently used as the standard for the selection of low-p_T electrons.
Full Run 2 data, corresponding to $137\ fb^{-1}$ at $13\ TeV$ centre of mass energy, collected with CMS detector was used to search for VBS in the $ZZ \to 4l2j$ channel. A kinematic discriminant was used as the main signal extraction tool and its performance was checked against the boosted decision tree (BDT) classifier. The EWK cross section was measured in three fiducial regions with $\sigma_{EW} = 0.33^{+0.11}_{-0.10}(stat)^{+0.04}_{-0.03}(sys)$ in the most inclusive volume. The total cross section and the signal strength were calculated in the three fiducial regions as well.
This analysis reported, for the first time with the CMS detector, the evidence for VBS in this channel with the background-only hypothesis rejected with an observed significance of $4.0\ \sigma$. Limits on dimension-8 tensor operators in the EFT framework were reported as well and are either World-best results thus far or competitive.
The final goal of all VBS analyses, and one of the most important tasks in EWK physics at the LHC, is the extraction of individual polarization components. While CMS is slowly entering the measurement era for the longitudinal scattering in the $W^{\pm}W^{\pm}jj$ channel with Run 2 data, this is not yet the case for the $ZZjj$ channel. At the same time, due to the fully reconstructable final state, this is expected to become the golden channel to separate the longitudinal component of polarization in the future. In order to assess the sensitivity to longitudinal scattering at higher energies and integrated luminosities, prospective studies for the High-Luminosity (HL) and High-Energy (HE) LHC were done. This corresponds to a 14 TeV centre of mass energy with $3000\ fb^{-1}$ of integrated luminosity for the former and 27 TeV centre of mass energy with $15000\ fb^{-1}$ of integrated luminosity for the latter. The longitudinal component was extracted using two BDT algorithms. The first algorithm was trained to separate the Z_LZ_L from the Z_LZ_T and Z_TZ_T components as well as the other backgrounds coming from the QCD-induced production of the same final state. The other, more sophisticated algorithm, was trained to simultaneously separate the Z_LZ_L component from the QCD-induced $pp \to ZZ$ processes with up to two extra parton emissions and the Z_LZ_L component from the mixture

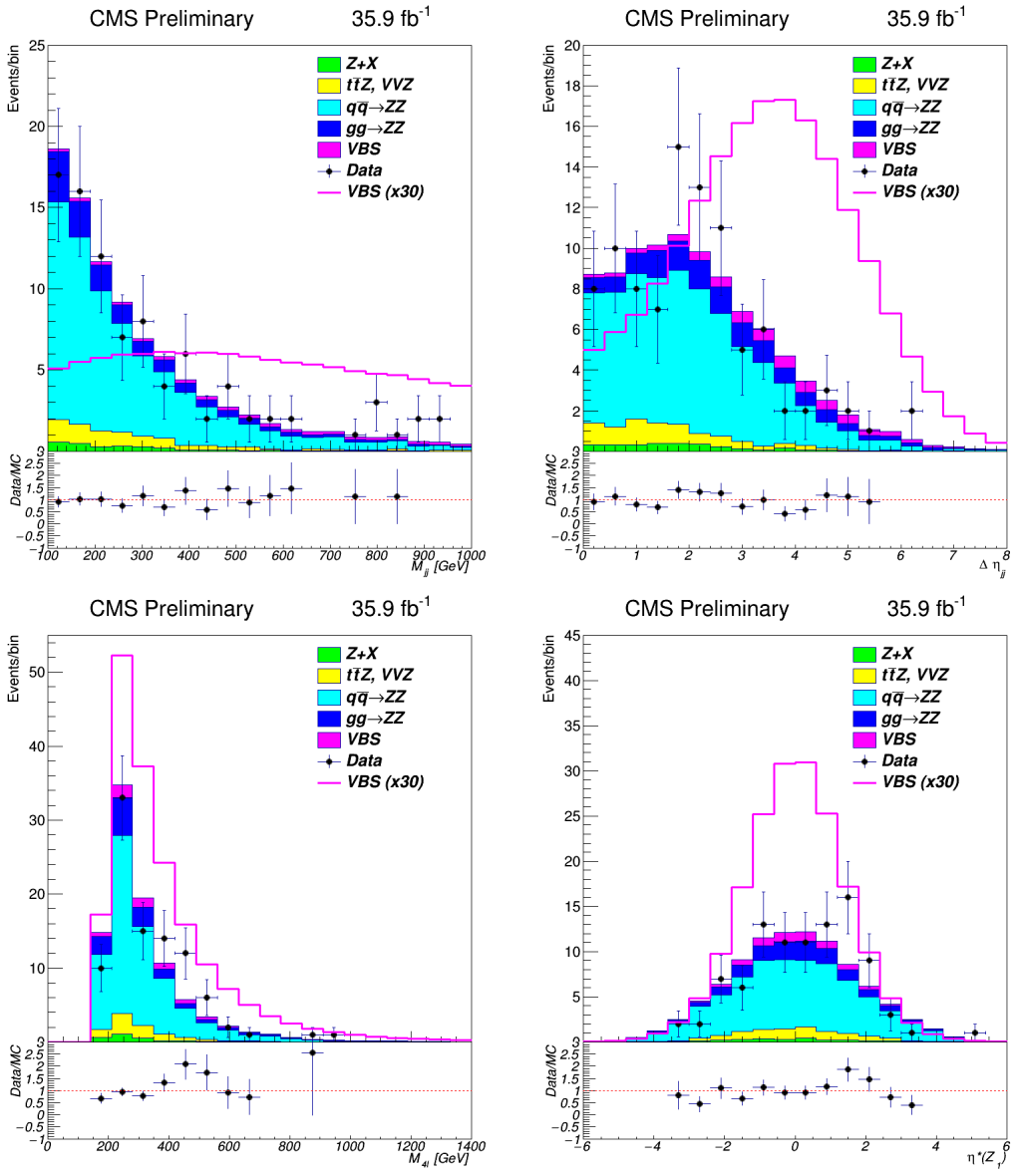
of $Z_L Z_T$ and $Z_T Z_T$ backgrounds. This algorithm is referred to as the 2D BDT and it was shown to outperform the simpler algorithm by up to almost $40\ \%$. At the same time, it outperforms the cut-and-count approach by almost $200\ \%$. Even with an outstanding performance of the 2D BDT algorithm, longitudinal scattering will remain out of reach at HL-LHC. However, it is expected to be finally measured at HE-LHC with significance passing the 5σ threshold.

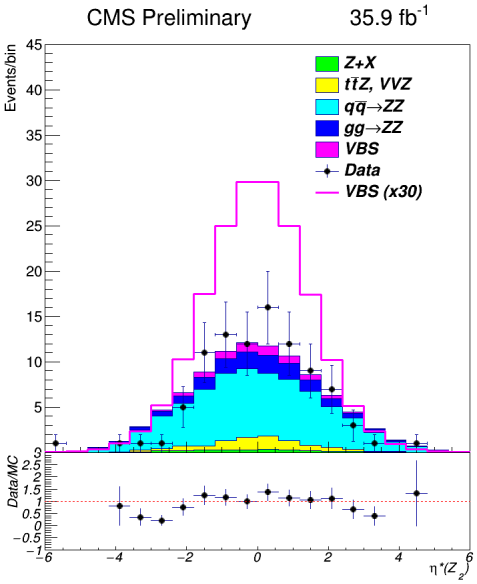
The prospective analysis presented in this thesis can be improved. Firstly, the fast simulation of detector effects used for these studies could be replaced by the full simulation with *GEANT4*. In addition, the loop-induced ZZ production was simulated at the leading order using *MCFM*. It was shown that a better description of the dijet phase space can be achieved with more involved simulation using *MadGraph5_aMC@NLO*. One of the biggest challenges at higher energies and luminosities will be the pileup (PU) which was found to affect leading jet selection at $10\ \%$ level. Further improvements in PU treatment could result in some gain in signal sensitivity. In addition, much progress has been made in the field of parton shower development in recent years which could improve signal sensitivity even more. Finally, deep neural networks (DNNs) have been shown to perform marvellously in a plethora of tasks. It would, therefore, seem most lucrative to try to replace current signal extraction tools with DNNs. However, even though some efforts in this direction have been made, the performance of DNNs in VBS analyses is still comparable to more standard signal extraction tools.

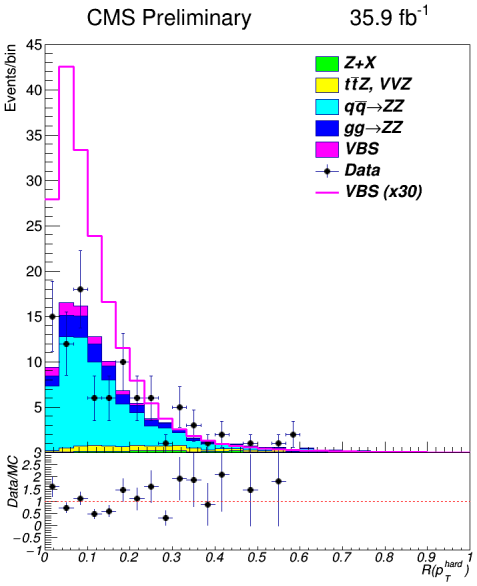
In the end, even though Run 2 was a huge success with significant new VBS results published by the CMS Collaboration, the long journey is still ahead of us. With Run 3 on its way, there will be roughly twice more collision data to analyse. This will enable us to do precision measurements of VBS in some channels (e.g. same sign $WWjj$, opposite sign $WWjj$, $Z\gamma$). On the other hand, we expect the first observation in the fully leptonic $ZZjj$ channel and, perhaps, the first observation of the longitudinal scattering in the $WWjj$ channel. In addition, a study of (non-VBS) diboson production at Run 3 will help us reduce systematic uncertainties, improve the modelling of certain processes and further develop machine learning techniques used for signal extraction. This will prepare us for analyses of more elusive VBS channels.
Following the end of Run 3, accelerator and detector update is planned to prepare for the installation of the HL-LHC which should begin with operation at the end of this decade. As was shown in this analysis, even $3000\ fb^{-1}$ of collision data expected at HL-LHC will not be sufficient to study longitudinal scattering in the $ZZjj$ channel. To fully unravel the mysteries of EWK physics, even higher energies and integrated luminosities, as expected at HE-LHC, will be needed. Furthermore, in 2020, *The European Strategy for Particle Physics* presented the long-term strategic plans for future particle colliders with the e^+e^- Higgs factory as one of the top priorities. Four proposals for the next-generation machines were made: two circular colliders (the *CEPC* and the *FCC*) and two linear colliders (the *ILC* and the *CLIC*). However, only the linear colliders can be extended into energy regions needed to study VBS. VBS analyses will benefit greatly from e^+e^- colliders because of the cleaner environment for the vector boson production as well as the higher coverage essential for measuring very-forward jets. With all this in mind, it is tempting to say that the study of the VBS processes has only just begun.

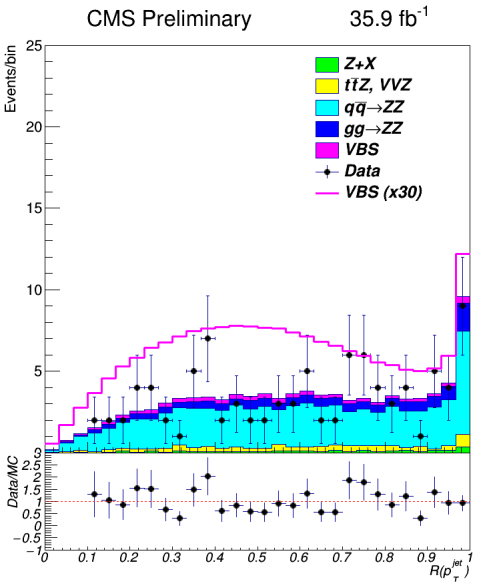
Appendix A: Supporting plots for the analysis presented in chapter 4

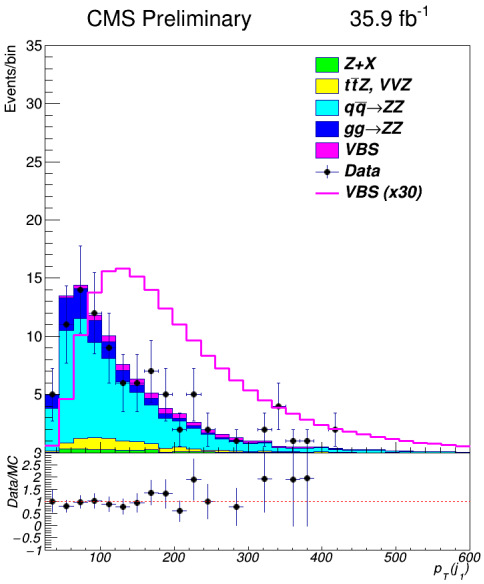
Figs. A.1 and A.2 show the distributions, defined in Table 4.7, for the 2016 and 2017 data-taking periods used to extract the VBS signal.

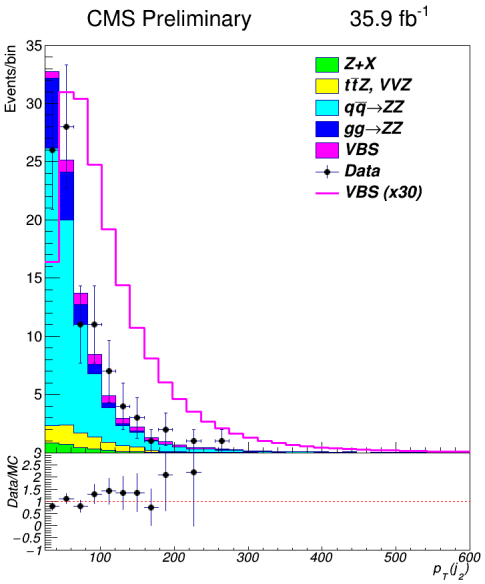


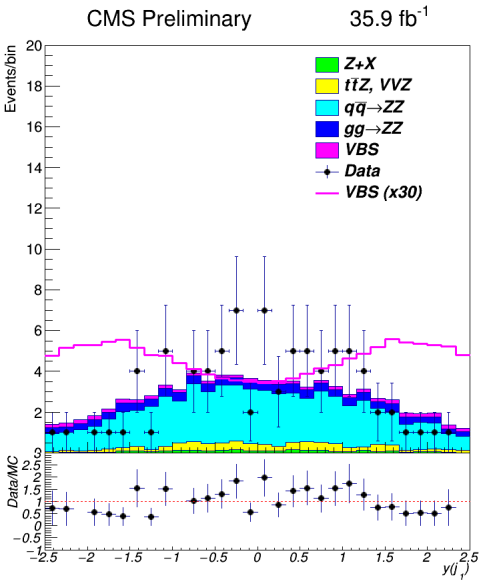

CMS Preliminary
35.9 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
η*(Z1)

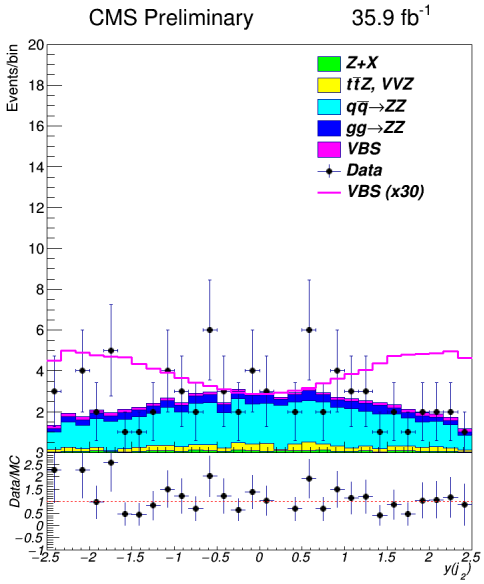

CMS Preliminary
35.9 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
R(pT hard)

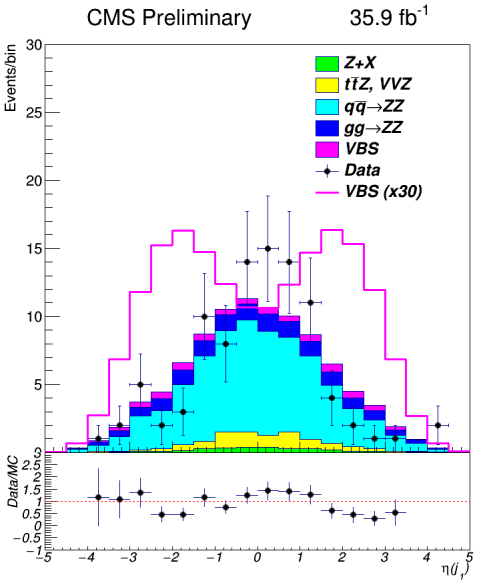

CMS Preliminary
35.9 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
R(pT jet)

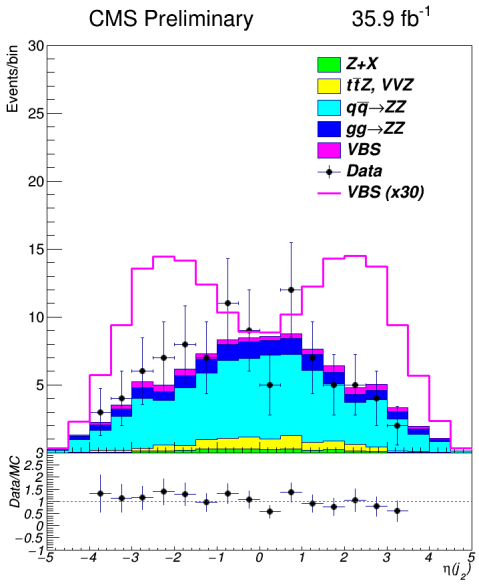

CMS Preliminary
35.9 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
pT(j1)

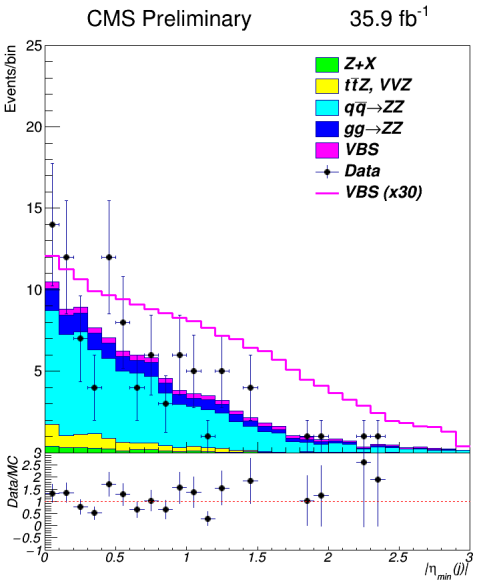

CMS Preliminary
35.9 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
pT(j2)

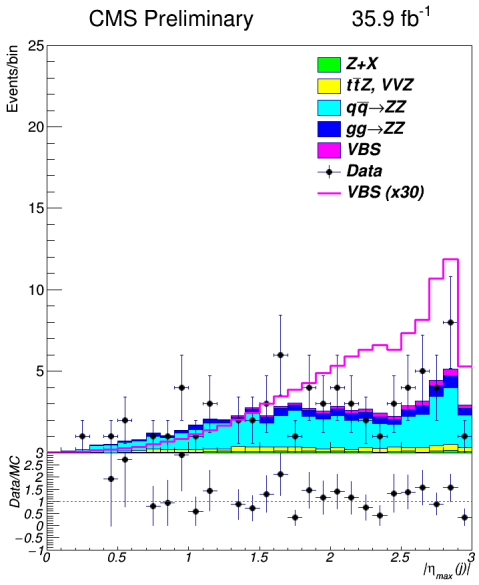

CMS Preliminary
35.9 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
y(j1)

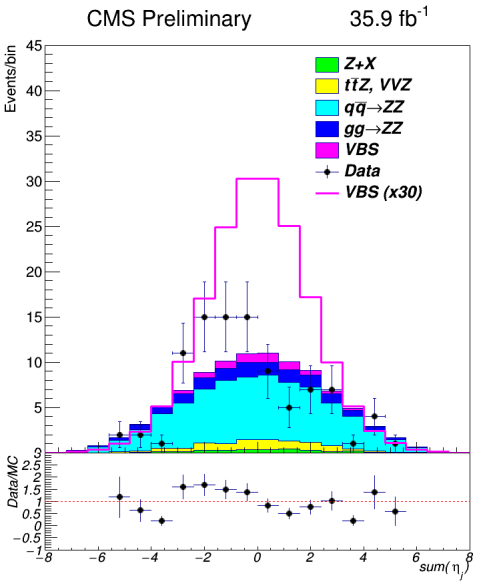

CMS Preliminary
35.9 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
y(j2)

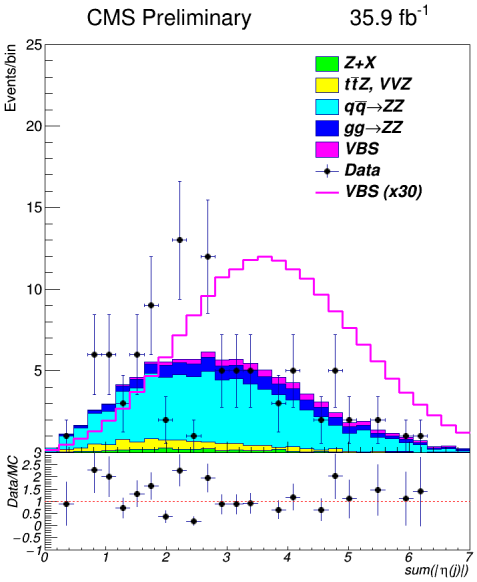

CMS Preliminary
35.9 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
η(j1)

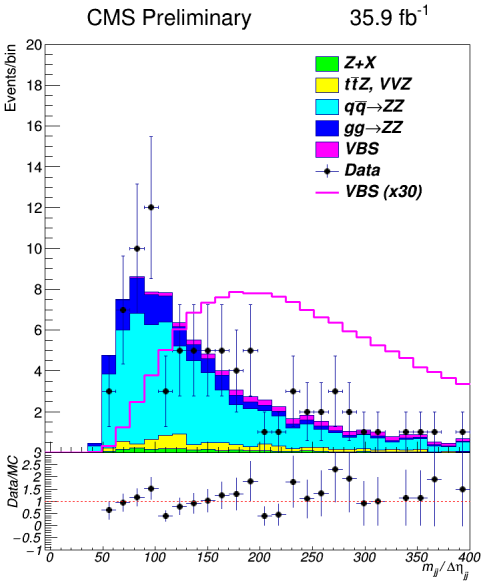

CMS Preliminary
35.9 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
η(j2)

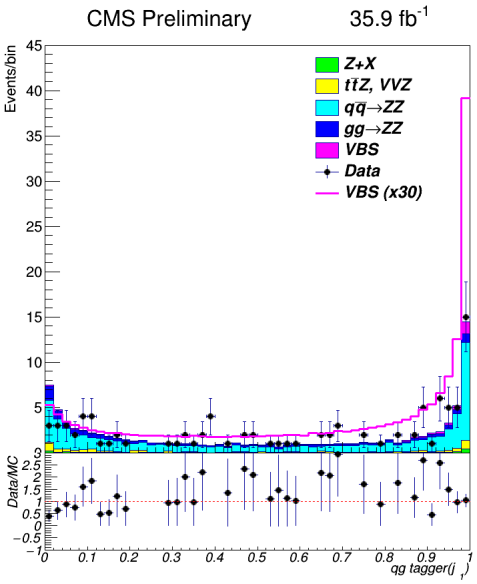

CMS Preliminary
35.9 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
|ηmin(j)|

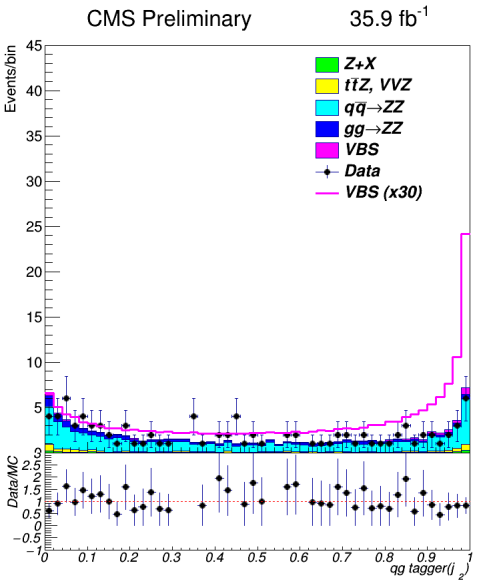

CMS Preliminary
35.9 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
|ηmax(j)|

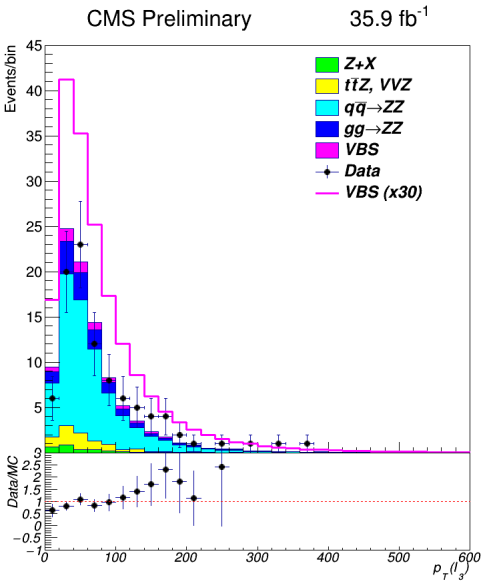

CMS Preliminary
35.9 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
sum(ηj)

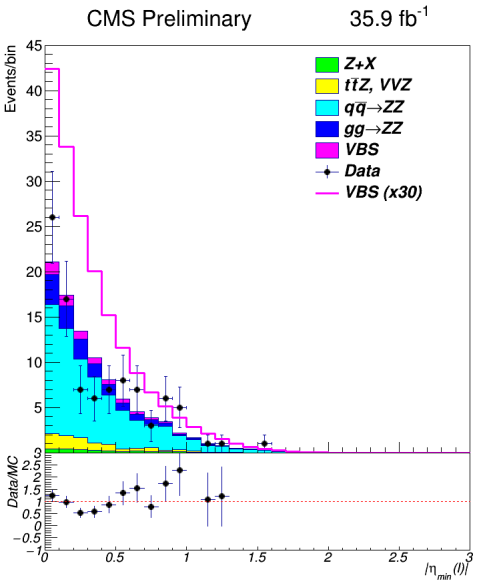

CMS Preliminary
35.9 fb-1
Events/bin
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Data/MC
sum(|η(j)|)


CMS Preliminary
35.9 fb-1
Events/bin
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Data/MC
mjj / Δηjj


CMS Preliminary
35.9 fb-1
Events/bin
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Data/MC
qg tagger(j1)


CMS Preliminary
35.9 fb-1
Events/bin
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Data/MC
qg tagger(j2)


CMS Preliminary
35.9 fb-1
Events/bin
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Data/MC
pT(l3)


CMS Preliminary
35.9 fb-1
Events/bin
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Data/MC
|ηmin(l)|

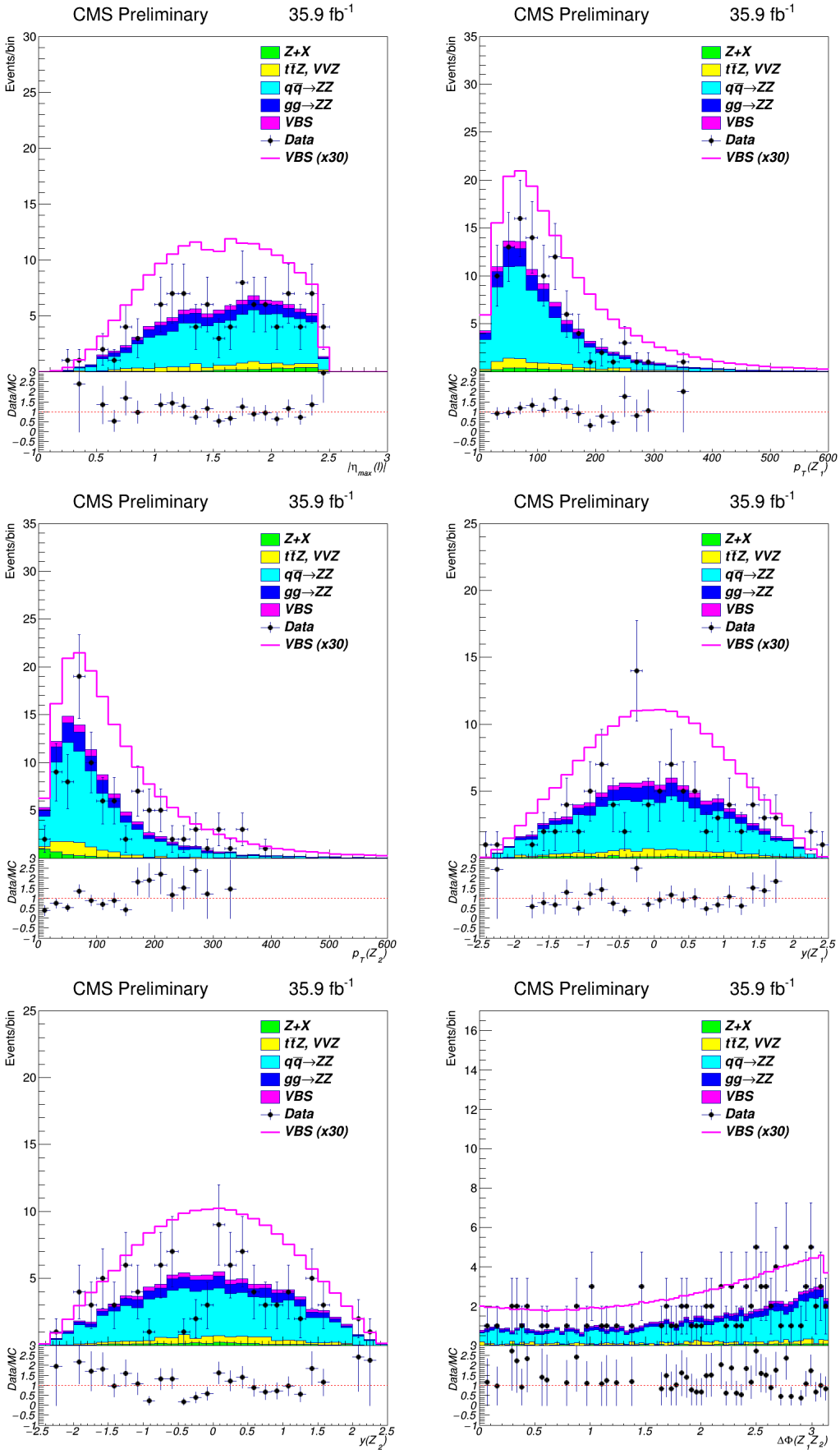
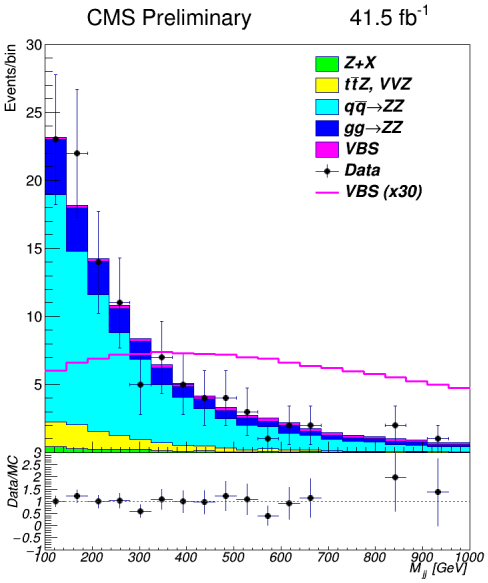
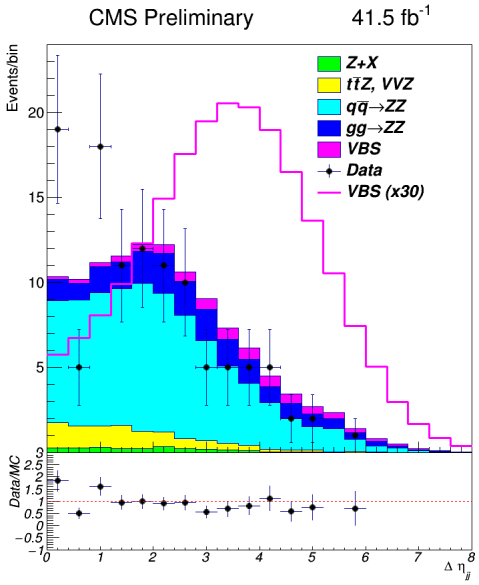
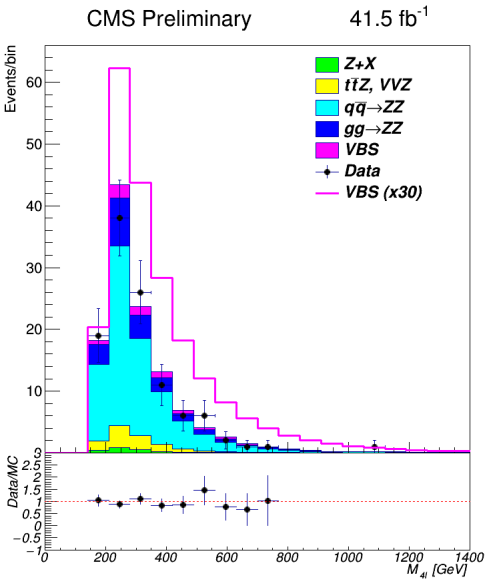
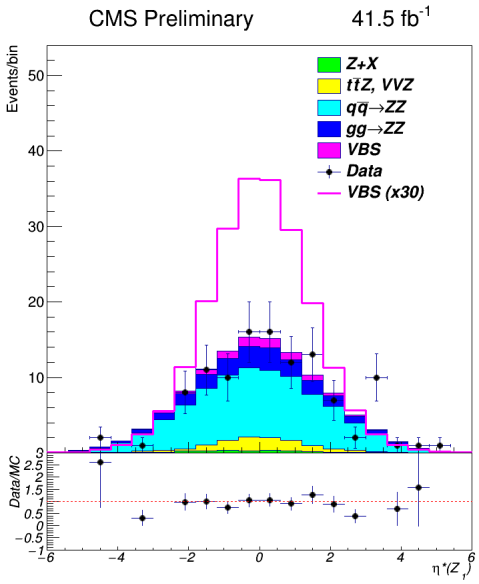


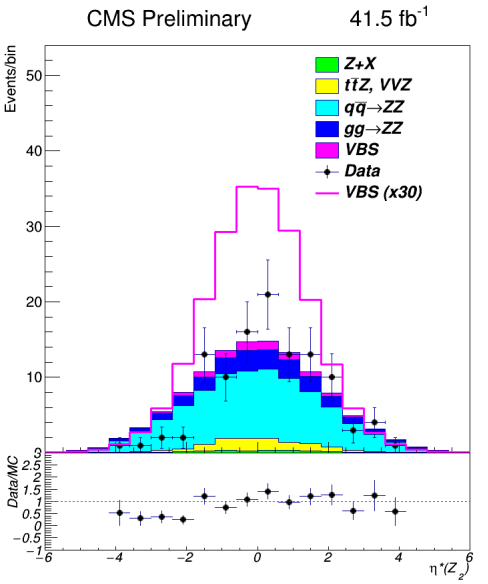
Figure A.1: Comparison of data to background and signal estimations in 2016 samples used in the analysis.

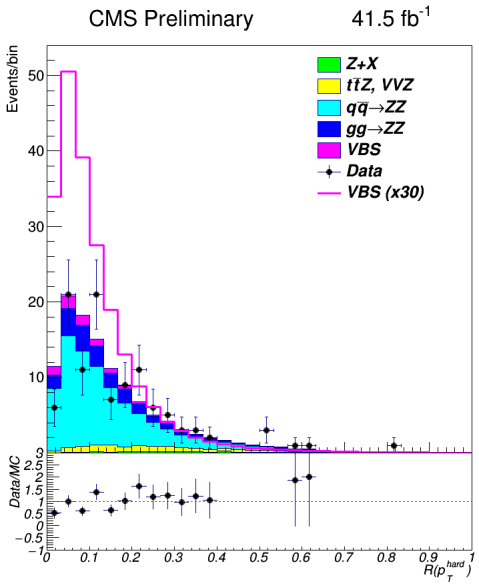

CMS Preliminary
41.5 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
Mjj [GeV]

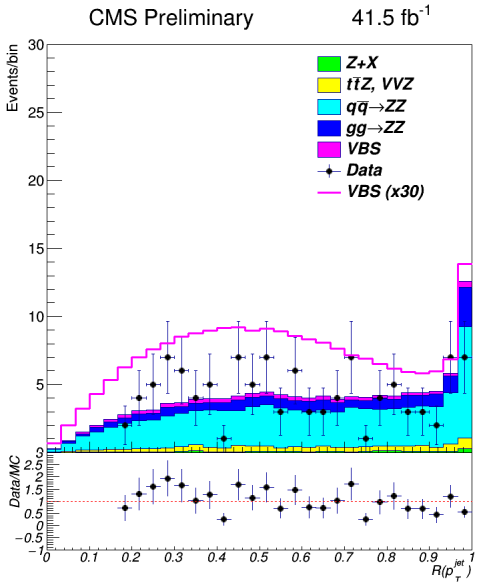

CMS Preliminary
41.5 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
Δ ηjj

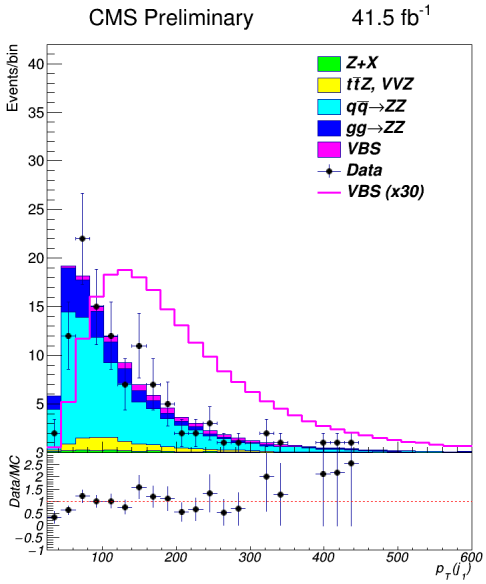

CMS Preliminary
41.5 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
M4l [GeV]

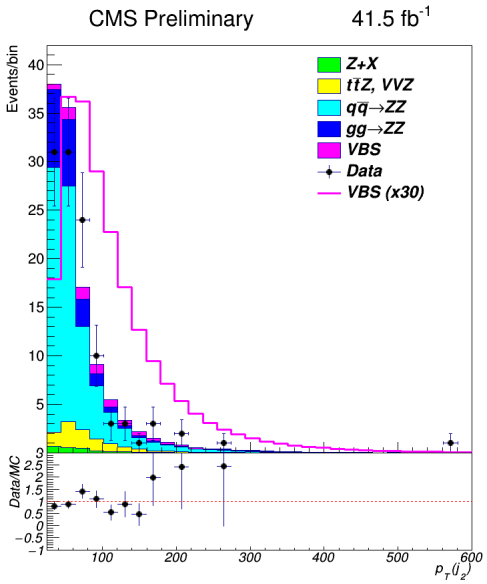

CMS Preliminary
41.5 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
η*(Z1)

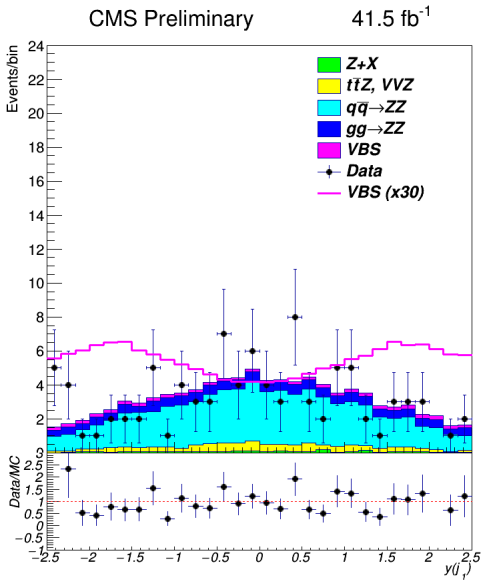

CMS Preliminary
41.5 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
η*(Z2)

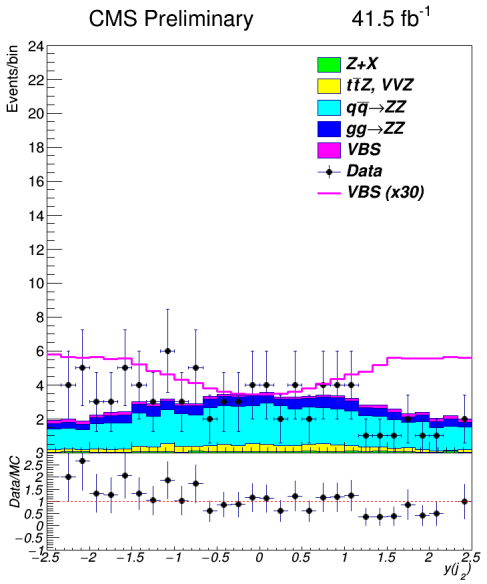

CMS Preliminary
41.5 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
R(pTHard)

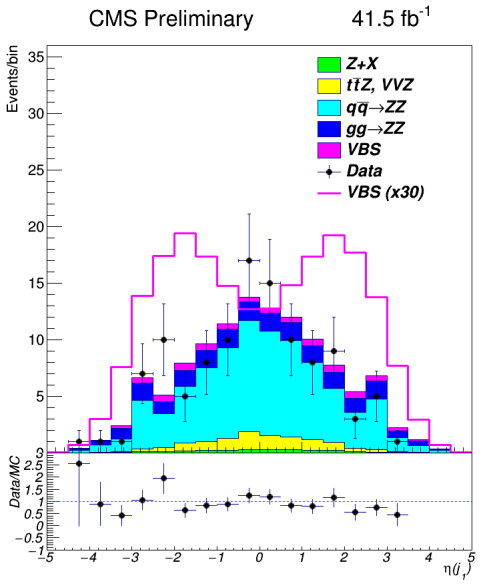

CMS Preliminary
41.5 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
R(p_T^jet)

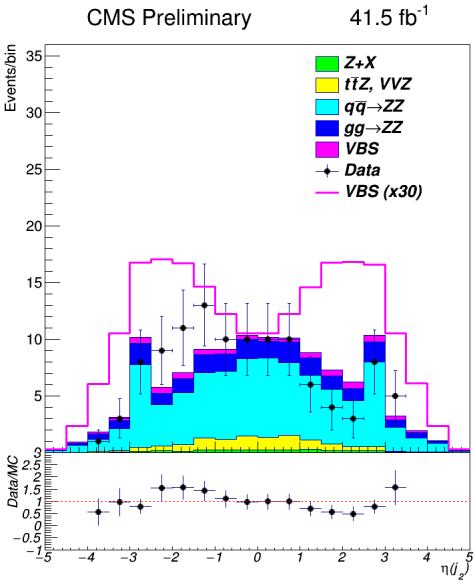

CMS Preliminary
41.5 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
p_T(j_1)

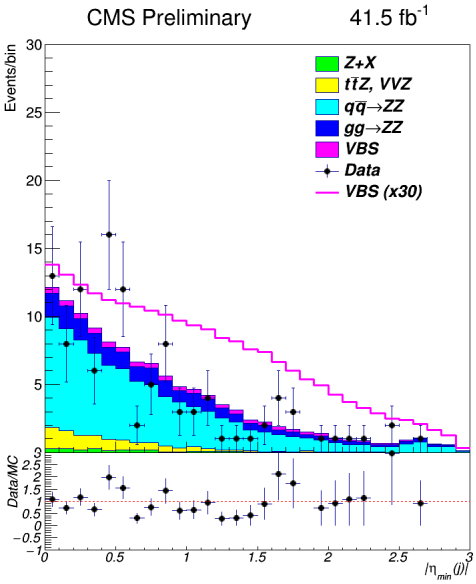

CMS Preliminary
41.5 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
p_T(j_2)

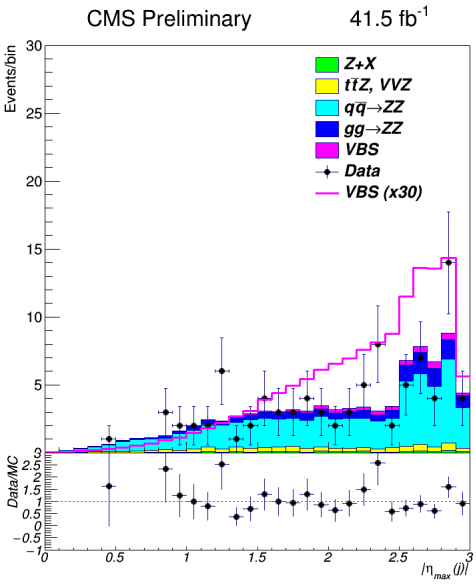

CMS Preliminary
41.5 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
y(j_1)

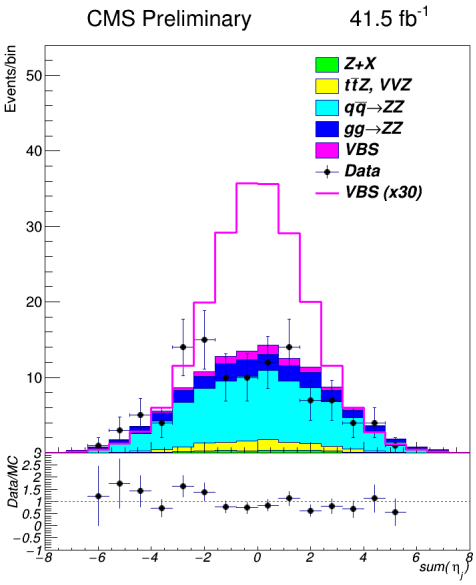

CMS Preliminary
41.5 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
y(j_2)

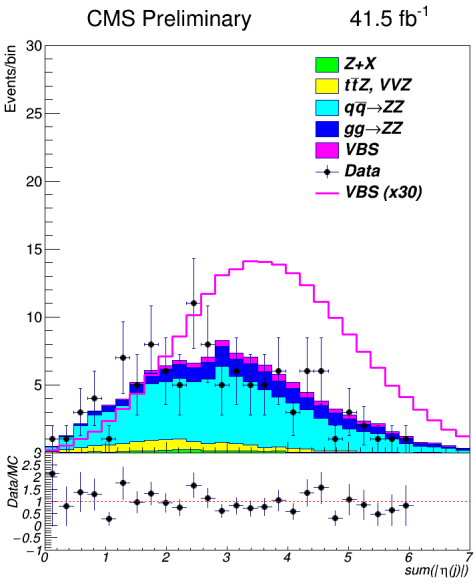

CMS Preliminary
41.5 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
η(j_1)

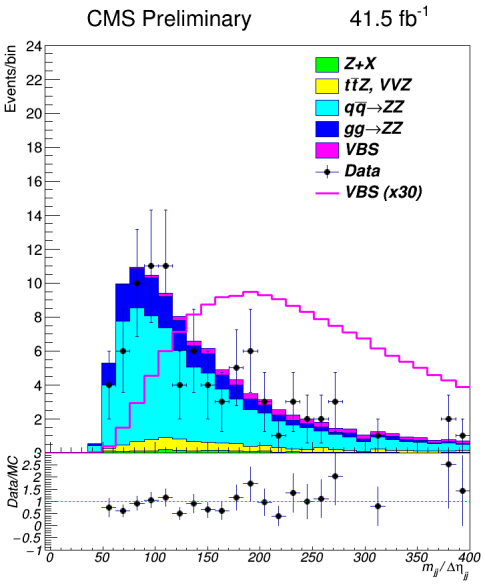

CMS Preliminary
41.5 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
η(j_z)

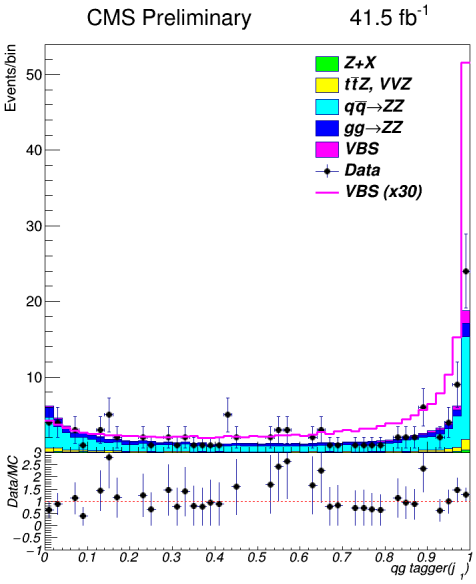

CMS Preliminary
41.5 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
|η_min(j)|

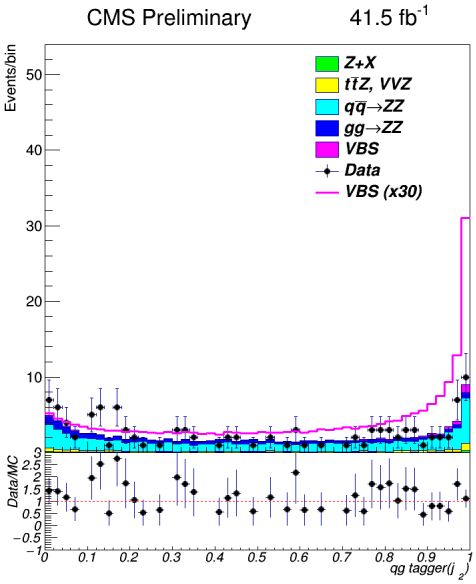

CMS Preliminary
41.5 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
|η_max(j)|

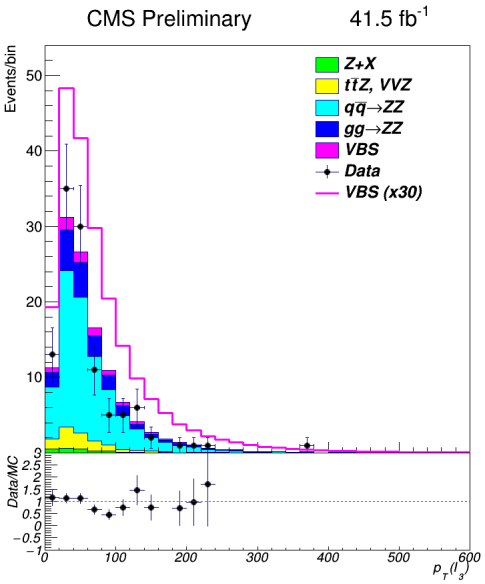

CMS Preliminary
41.5 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
sum(η_j)

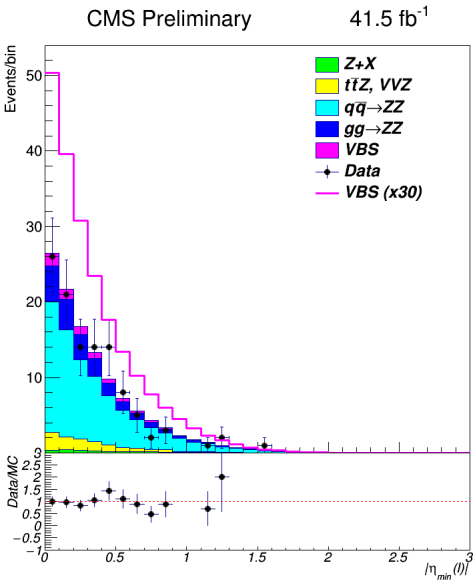

CMS Preliminary
41.5 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
sum(|η(j)|)

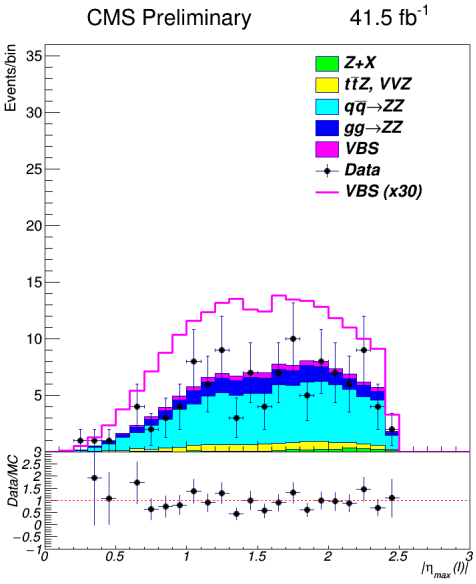

CMS Preliminary
41.5 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
m_jj/Δη_jj

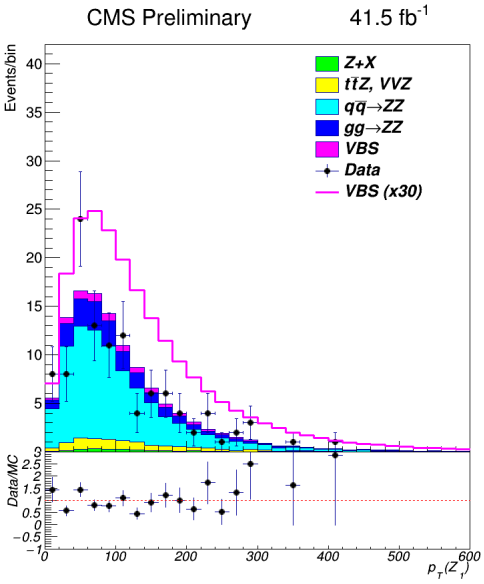

CMS Preliminary
41.5 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
qg tagger(j 1)


CMS Preliminary
41.5 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
qg tagger(j 2)


CMS Preliminary
41.5 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
pT(l3)


CMS Preliminary
41.5 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
|ηmin(l)|


CMS Preliminary
41.5 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
|ηmax(l)|


CMS Preliminary
41.5 fb-1
Z+X
tt̄Z, VVZ
qq̄→ZZ
gg→ZZ
VBS
Data
VBS (x30)
Events/bin
Data/MC
pT(Z1)

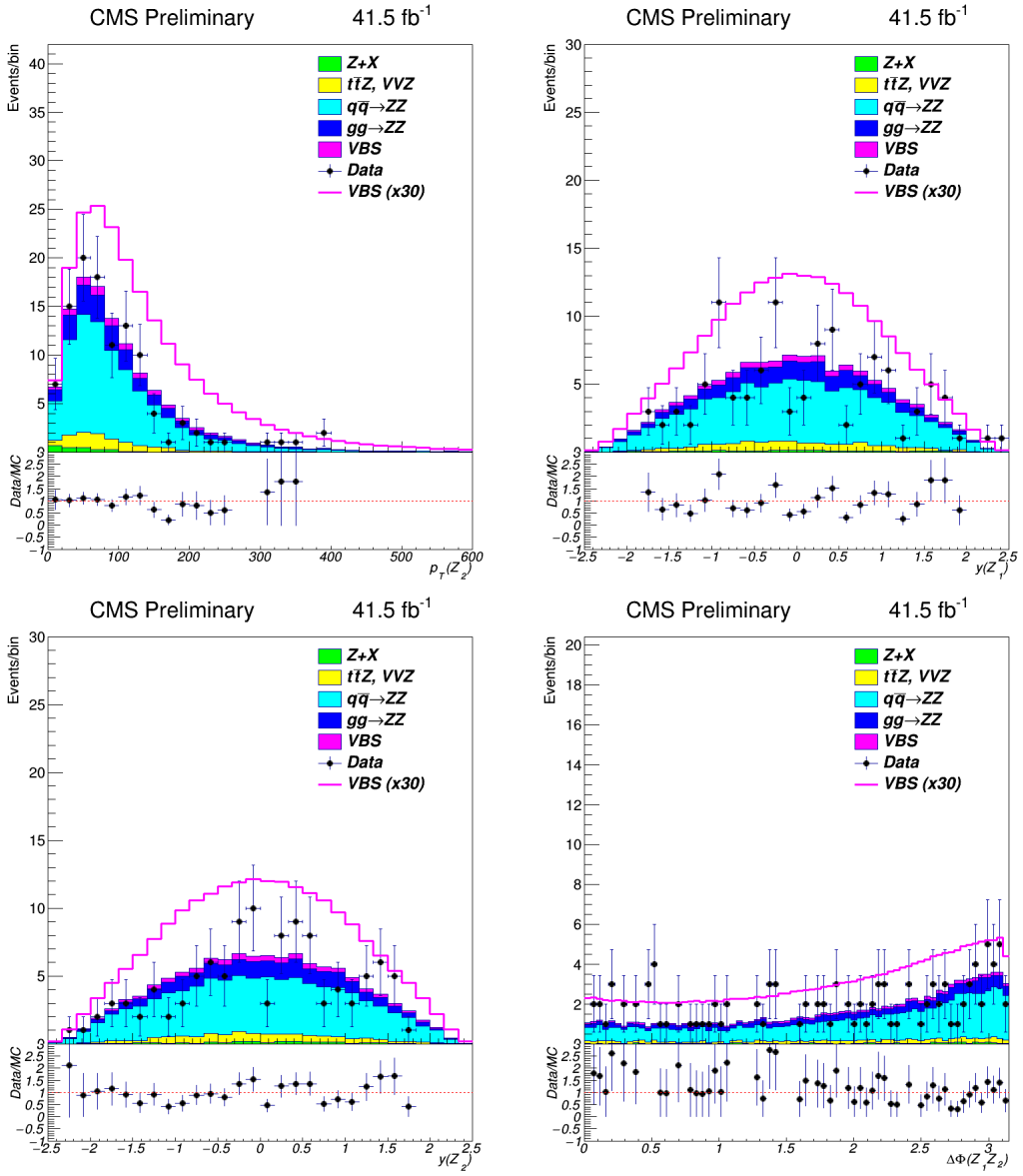


Figure A.2: Comparison of data to background and signal estimations in 2017 samples used in the analysis.

Figs. A.3 and A.4 show input distributions used for the *BDT7* training, for the 2016 and 2017 data-taking periods, respectively. EWK signal is shown in blue and the qqZZ background in red. Distributions used for the BDT28 training are shown in Figs. A.5 and A.6 for the 2016 and 2017 data-taking periods, respectively.

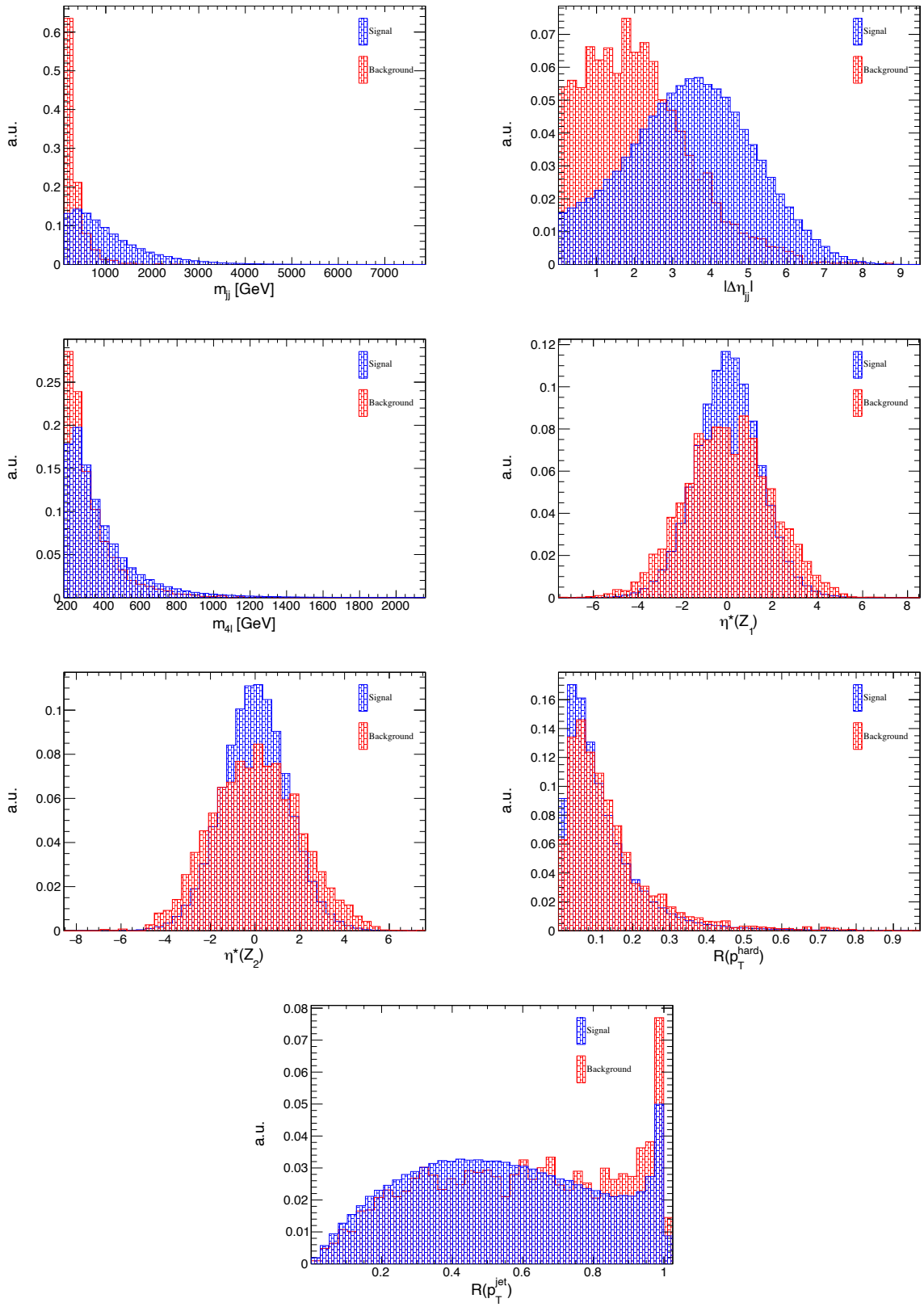


Figure A.3: BDT input distributions for the EWK signal (in blue) and the qqZZ background (in red) to the *BDT7* training for the 2016 period.

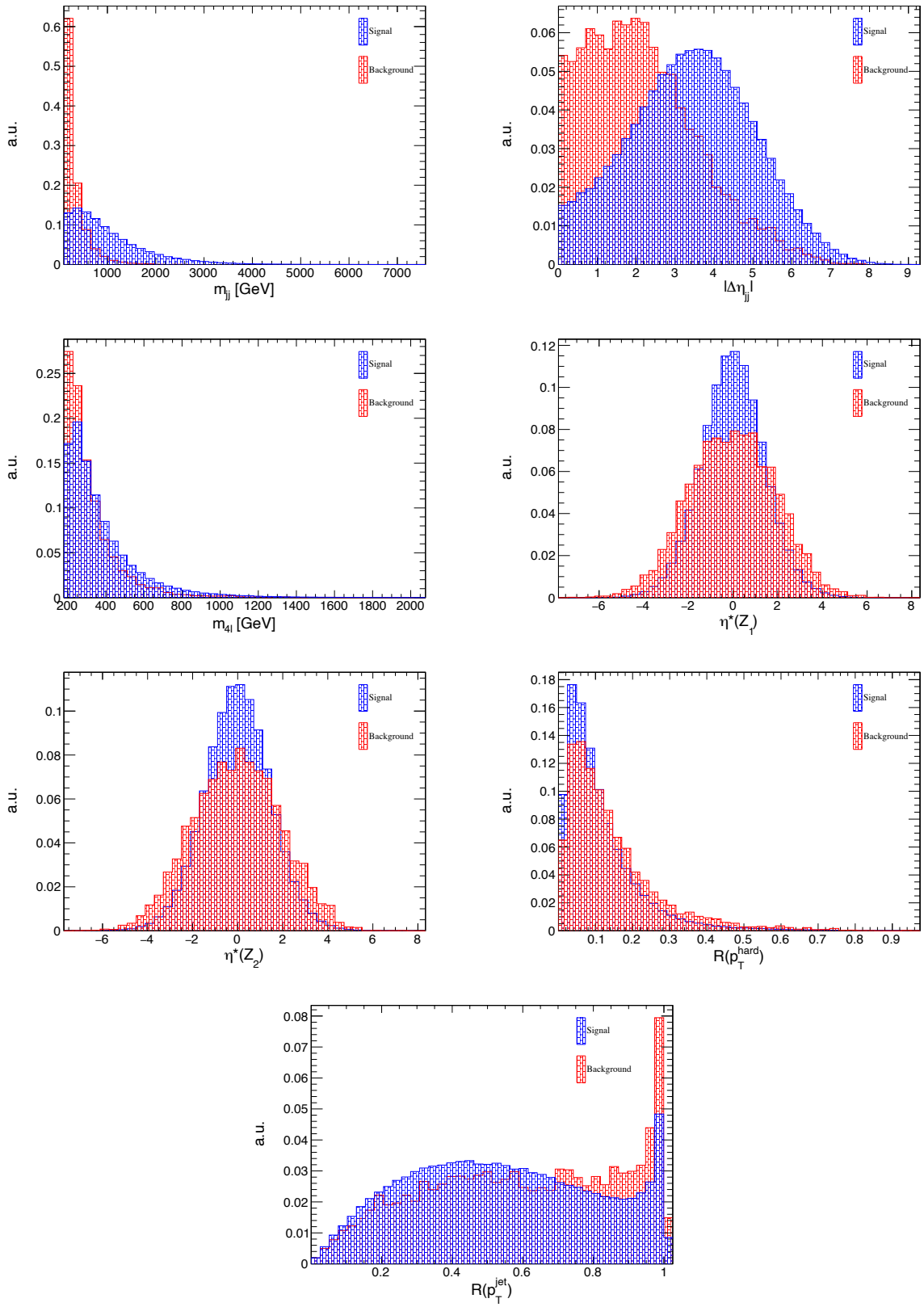


Figure A.4: BDT input distributions for the EWK signal (in blue) and the qqZZ background (in red) to the *BDT7* training for the 2017 period.

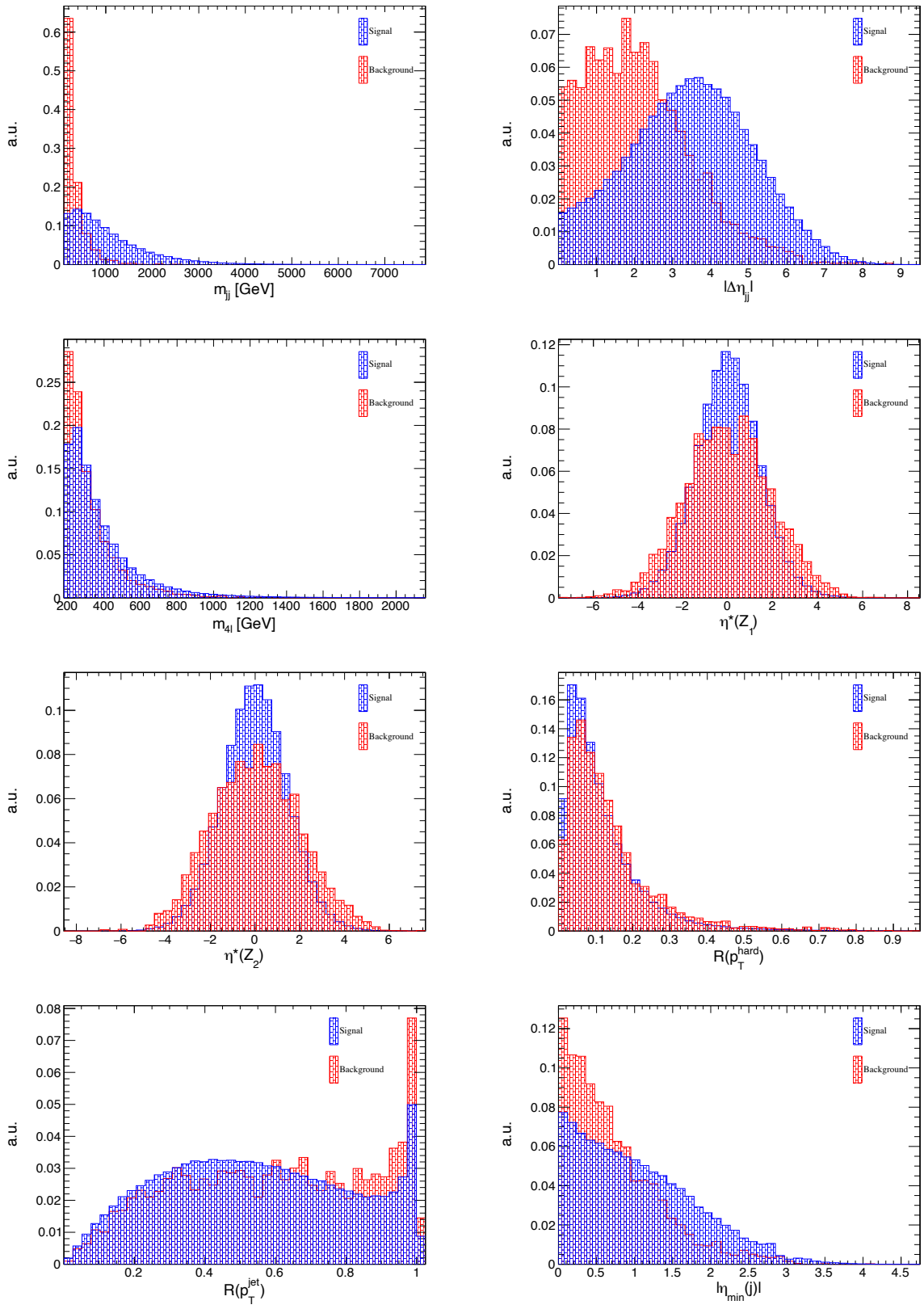

Signal
Background
a.u.
m_{jj} [GeV]
$|\Delta\eta_{jj}|$
m_{4l} [GeV]
$\eta^*(Z_1)$
$\eta^*(Z_2)$
$R(p_T^{hard})$
$R(p_T^{jet})$
$|\eta_{min}(j)|$

Figure A.5

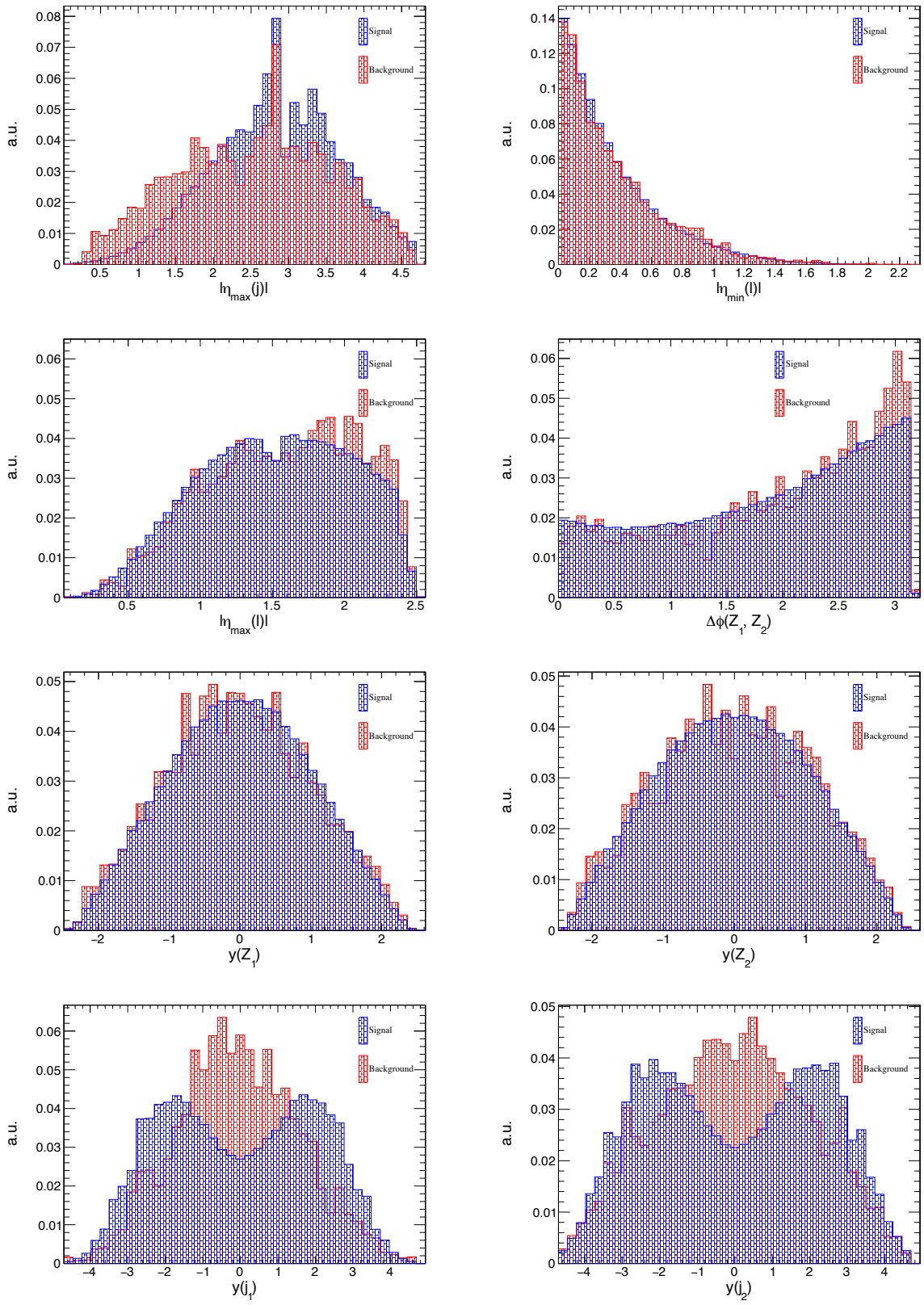

a.u.
Signal
Background
$|\eta_{max}(j)|$
$|\eta_{min}(l)|$
$|\eta_{max}(l)|$
$\Delta\phi(Z_1, Z_2)$
$y(Z_1)$
$y(Z_2)$
$y(j_1)$
$y(j_2)$

Figure A.5

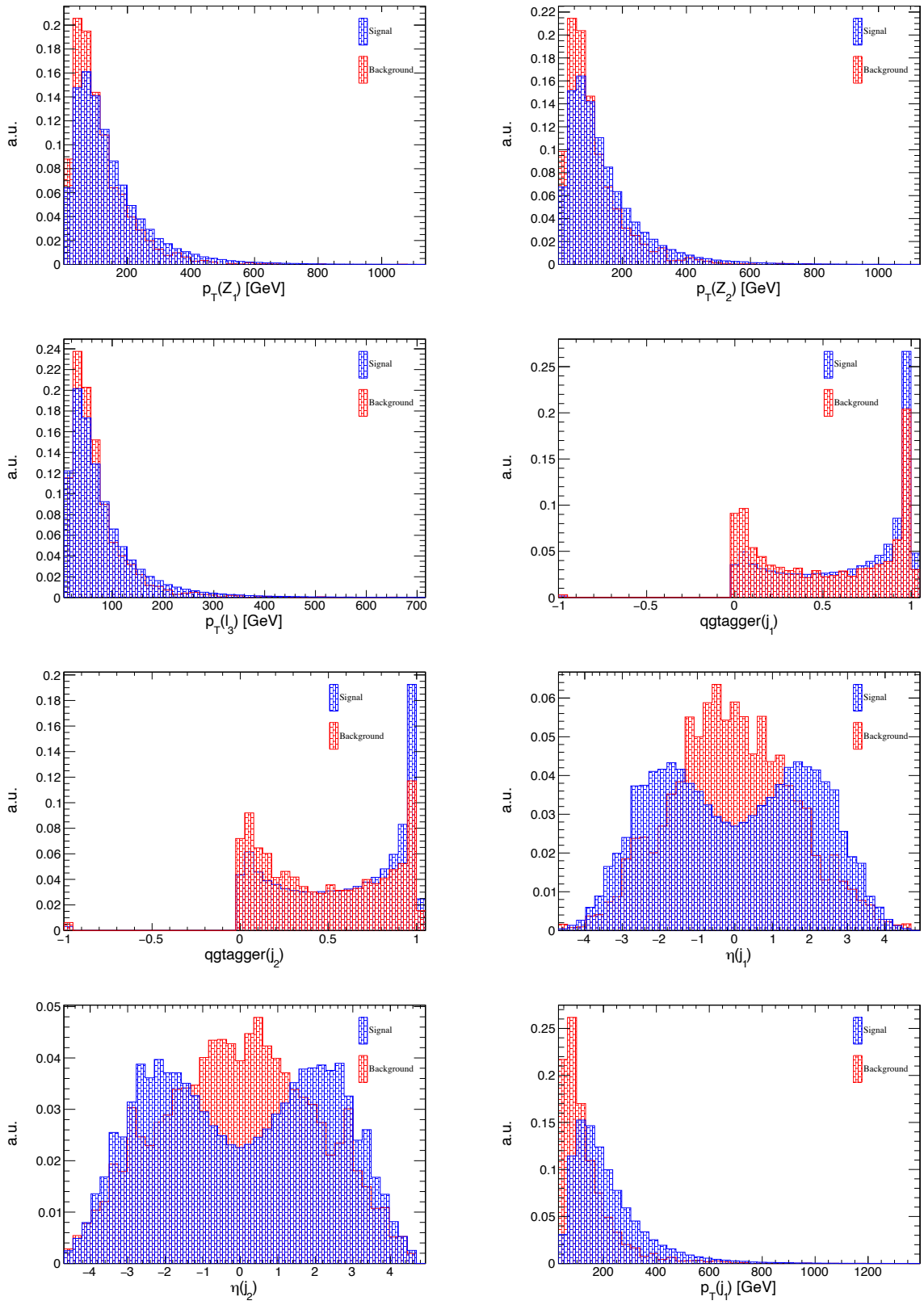

Signal
Background
a.u.
$p_T(Z_1)$ [GeV]
$p_T(Z_2)$ [GeV]
$p_T(l_3)$ [GeV]
qgtagger(j_1)
qgtagger(j_2)
$\eta(j_1)$
$\eta(j_2)$
$p_T(j_1)$ [GeV]

Figure A.5

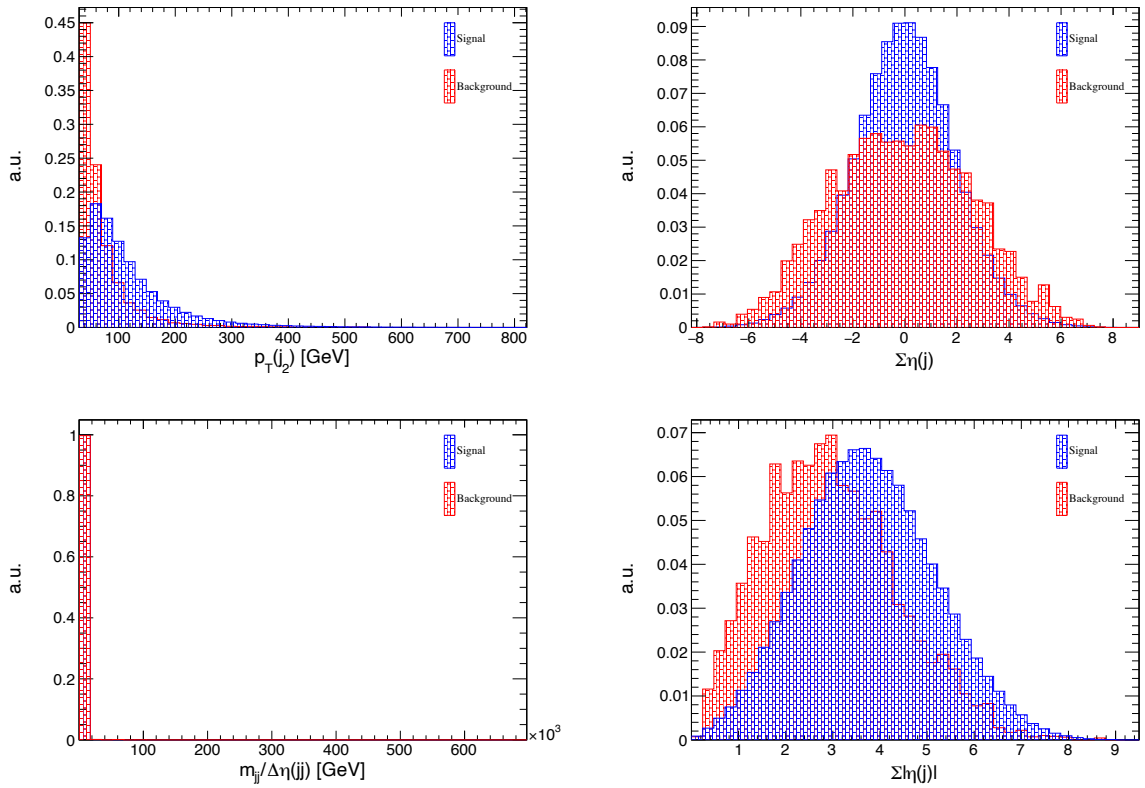


Figure A.5: BDT input distributions for the EWK signal (in blue) and the qqZZ background (in red) to the *BDT28* training for the 2016 period.

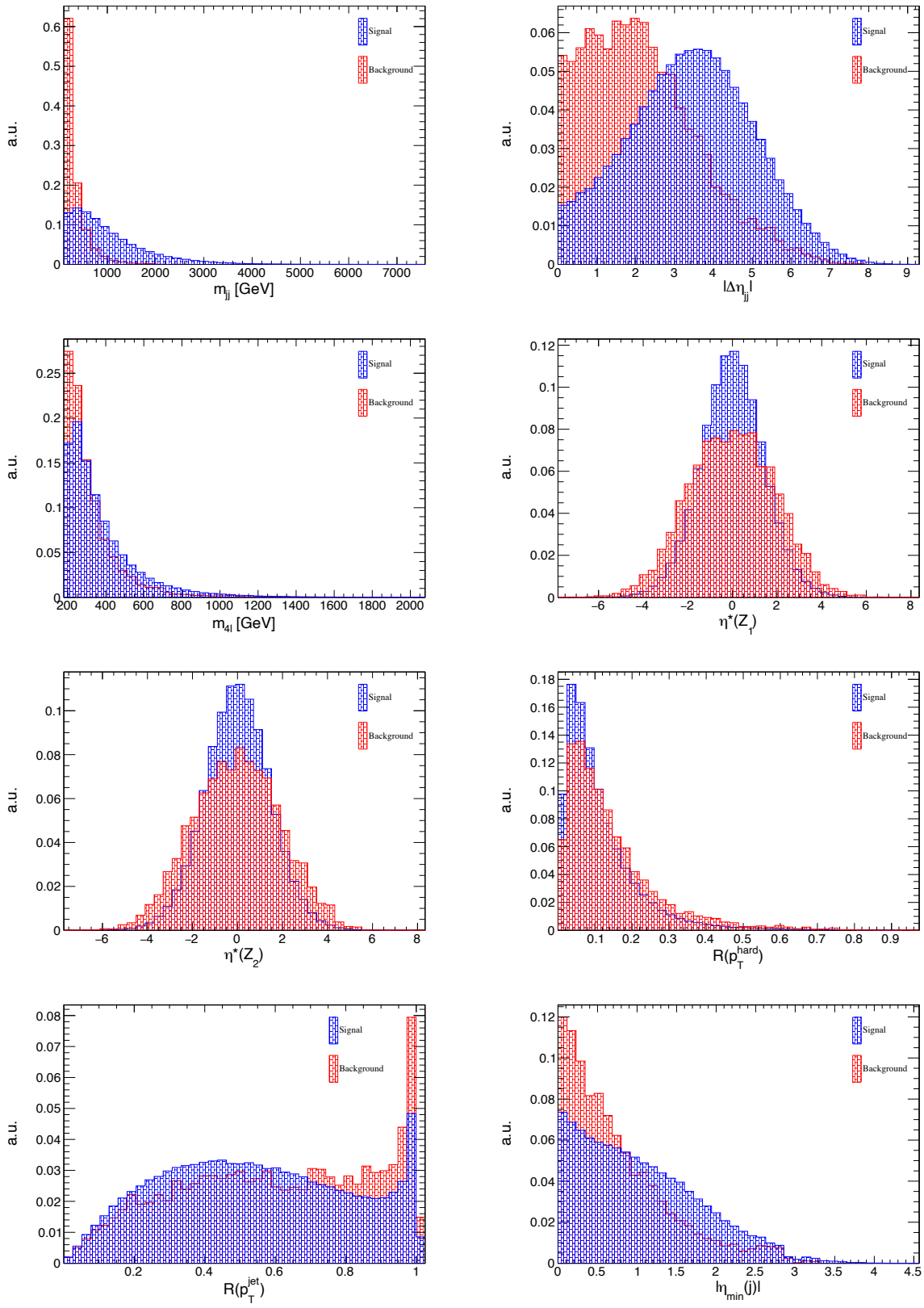

a.u.
Signal
Background
m_{jj} [GeV]
$|\Delta\eta_{jj}|$
m_{4l} [GeV]
$\eta^*(Z_1)$
$\eta^*(Z_2)$
$R(p_T^{hard})$
$R(p_T^{jet})$
$|\eta_{min}(j)|$

Figure A.6

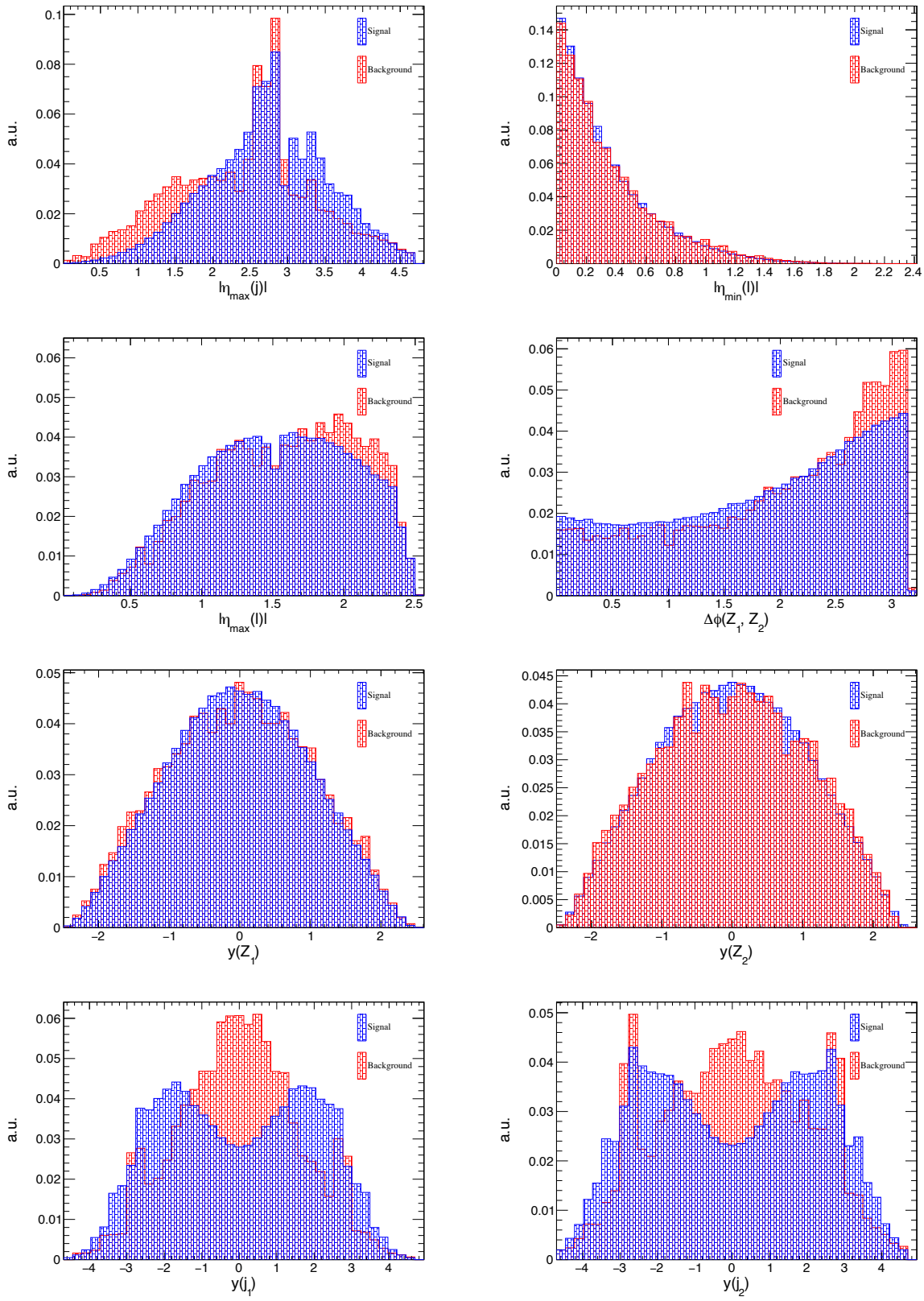

a.u.
Signal
Background
$|\eta_{max}(j)|$
$|\eta_{min}(l)|$
$|\eta_{max}(l)|$
$\Delta\phi(Z_1, Z_2)$
$y(Z_1)$
$y(Z_2)$
$y(j_1)$
$y(j_2)$

Figure A.6

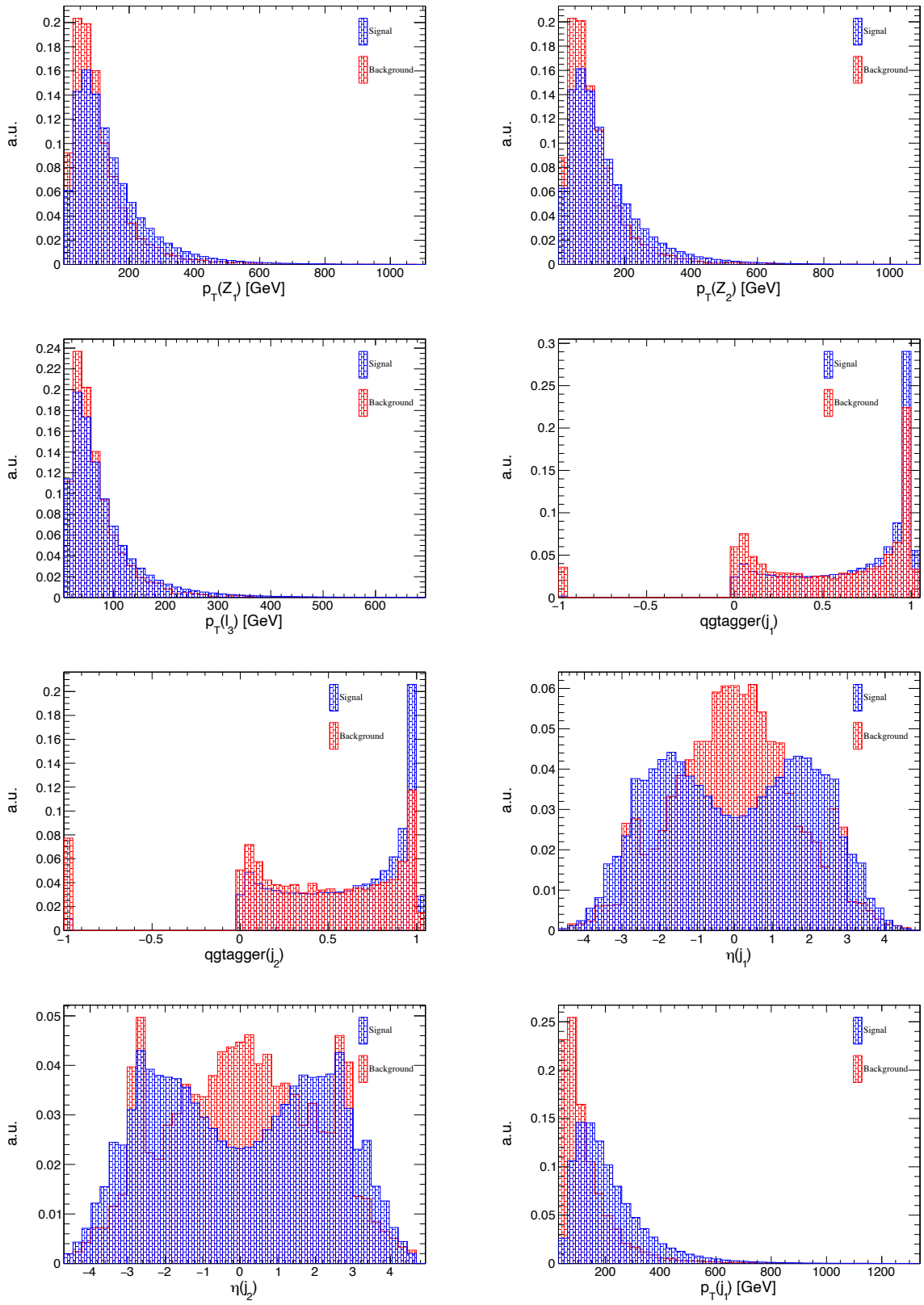

a.u.
Signal
Background
$p_T(Z_1)$ [GeV]
$p_T(Z_2)$ [GeV]
$p_T(l_3)$ [GeV]
qgtagger(j_1)
qgtagger(j_2)
$\eta(j_1)$
$\eta(j_2)$
$p_T(j_1)$ [GeV]

Figure A.6

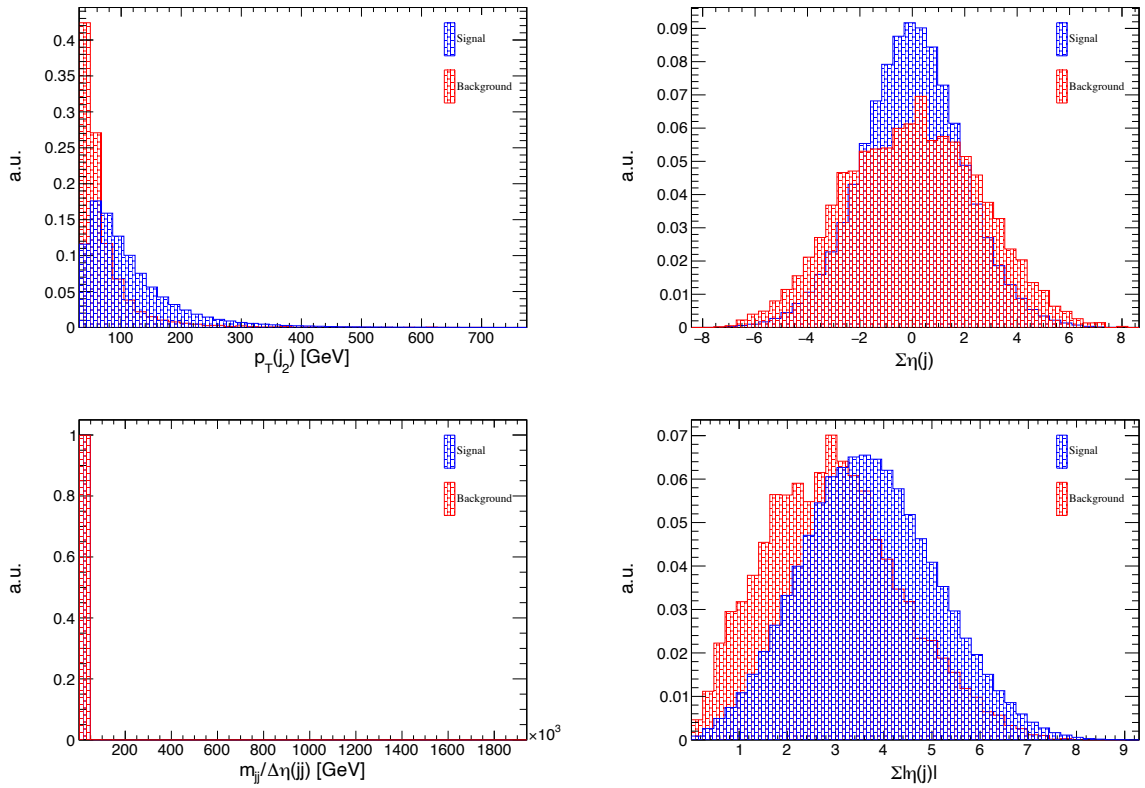
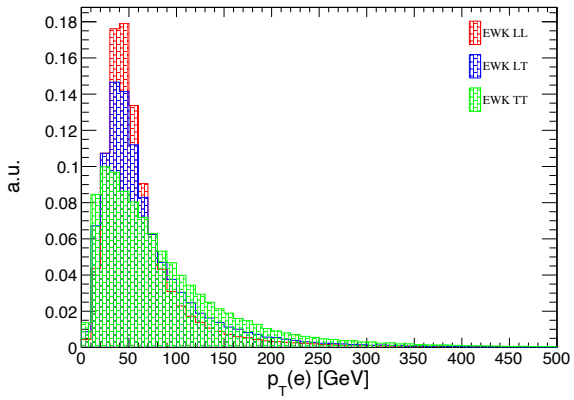
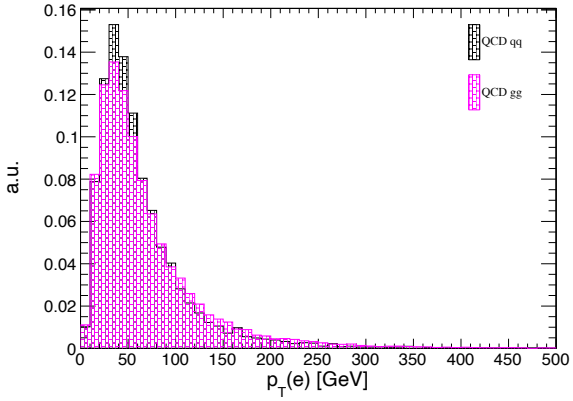


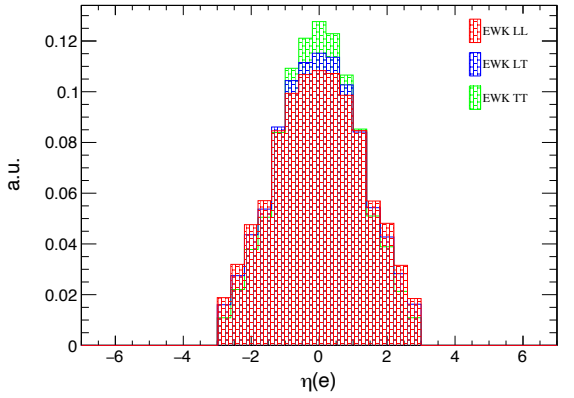
Figure A.6: BDT input distributions for the EWK signal (in blue) and the qqZZ background (in red) to the *BDT28* training for the 2017 period.

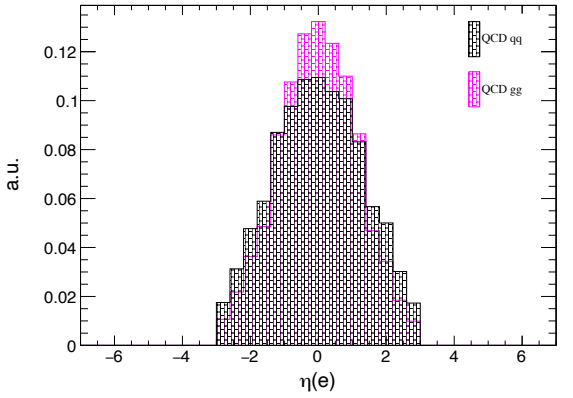
Appendix B: Supporting plots for the analysis presented in chapter 5

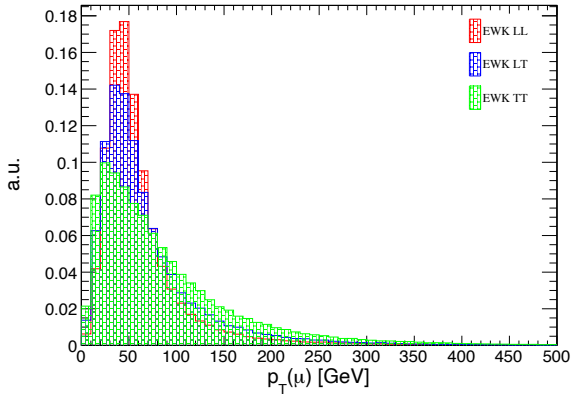
Figs. B.1 and B.2 show distributions at 14 and 27 TeV of all variables used in the analysis presented in chapter 5. Figs. B.3-B.7 compare the kinematics at 14 and 27 TeV for the LL, LT, TT, qq and gg contributions, separately.

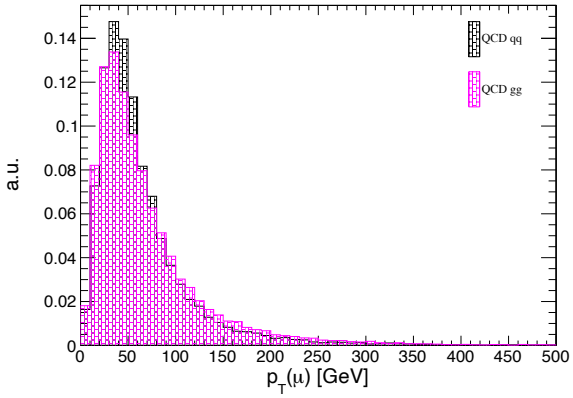


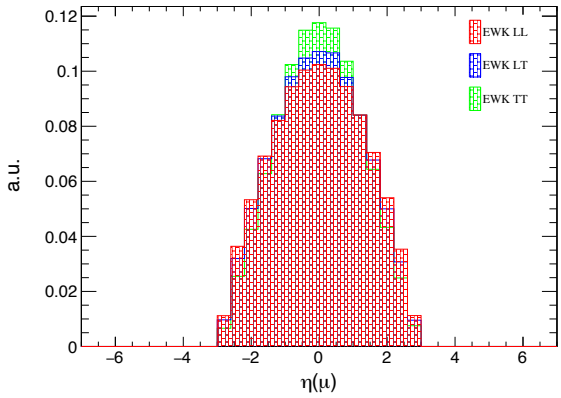


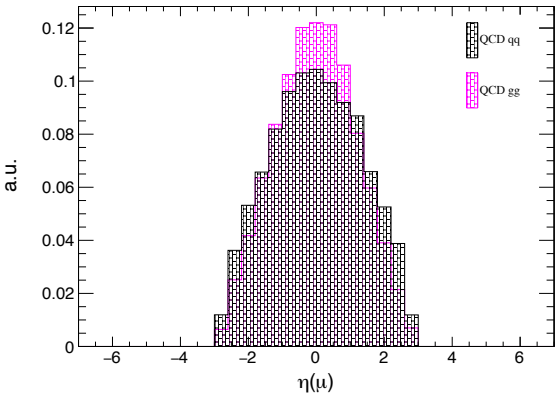


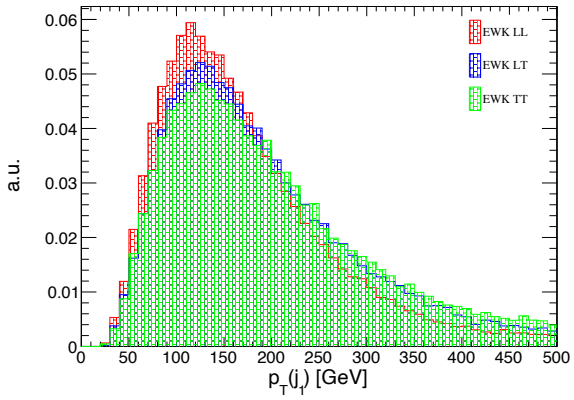


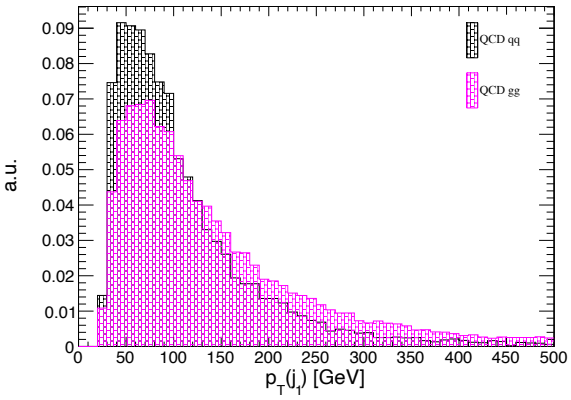


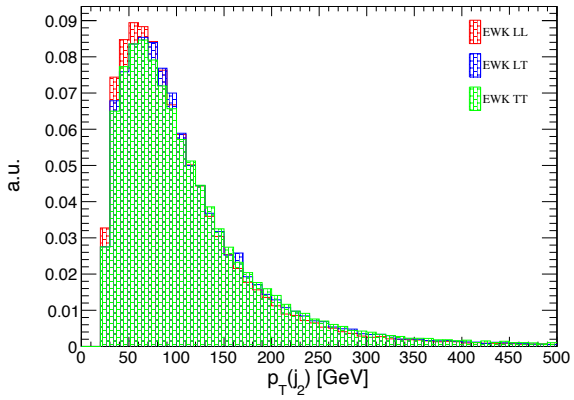


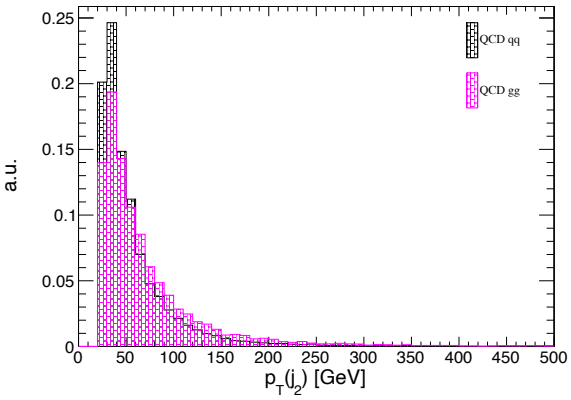

EWK LL
EWK LT
EWK TT
a.u.
η(μ)

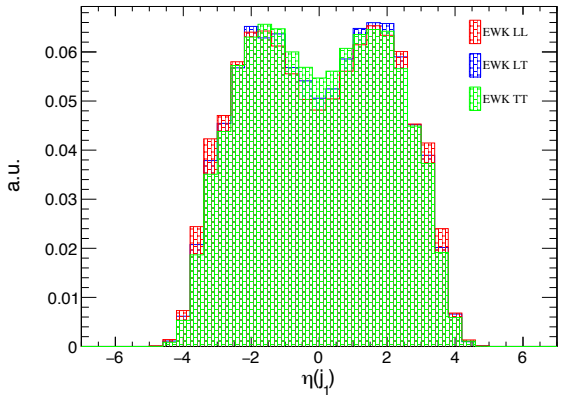

QCD qq
QCD gg
a.u.
η(μ)

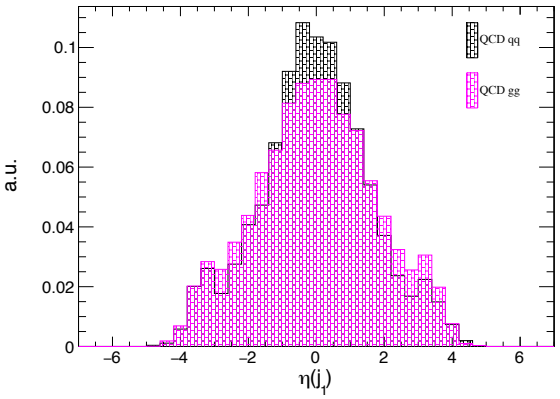

EWK LL
EWK LT
EWK TT
a.u.
p_T(j_1) [GeV]

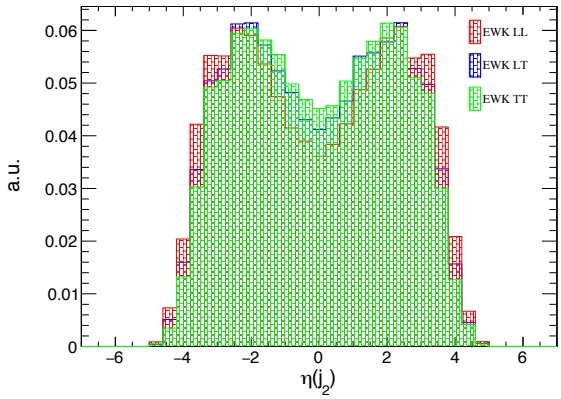

QCD qq
QCD gg
a.u.
p_T(j_1) [GeV]

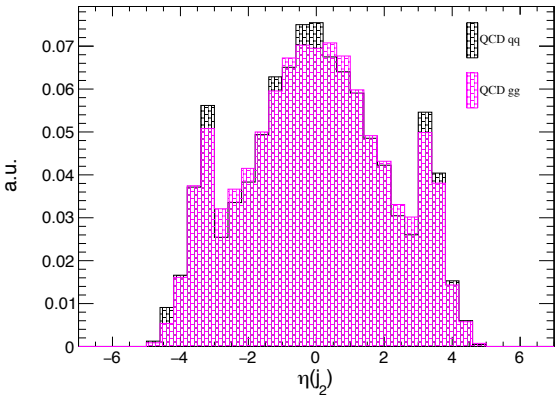

EWK LL
EWK LT
EWK TT
a.u.
p_T(j_2) [GeV]

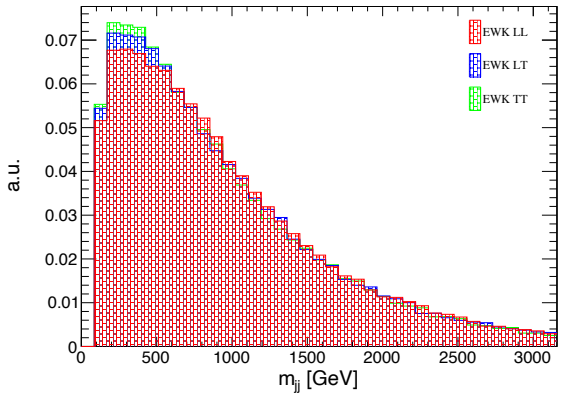

QCD qq
QCD gg
a.u.
p_T(j_2) [GeV]

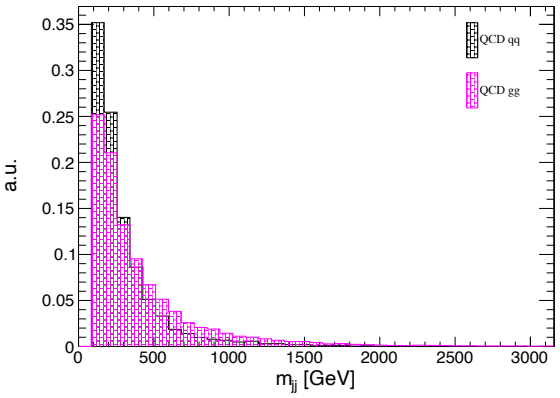

EWK LL
EWK LT
EWK TT
a.u.
η(j_1)

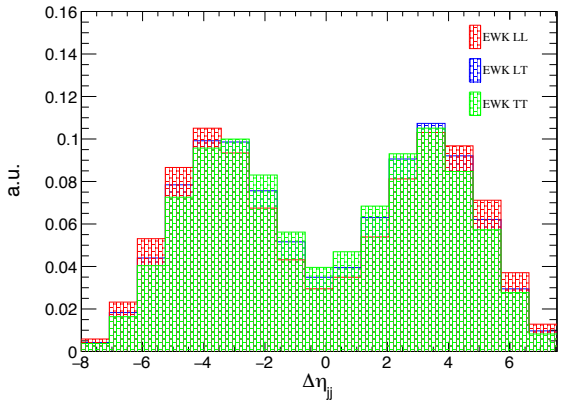

QCD qq
QCD gg
a.u.
η(j_1)

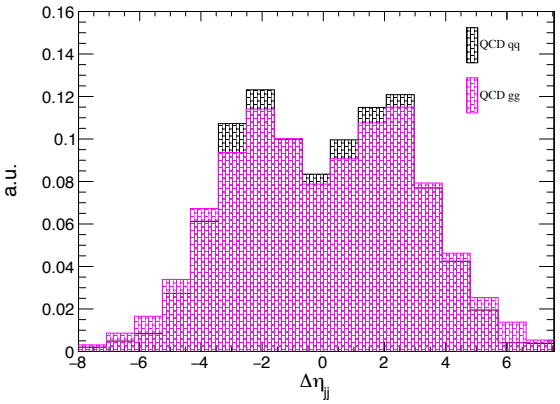

EWK LL
EWK LT
EWK TT
a.u.
η(j2)

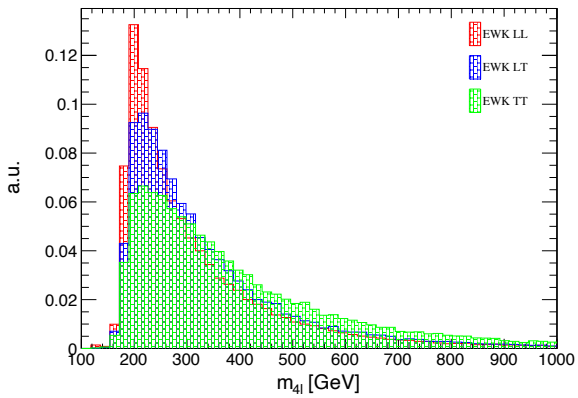

QCD qq
QCD gg
a.u.
η(j2)

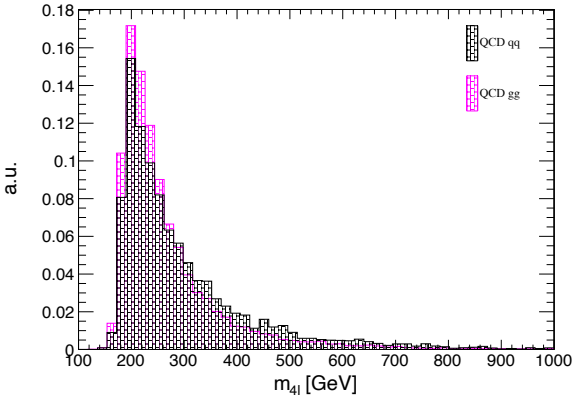

EWK LL
EWK LT
EWK TT
a.u.
mjj [GeV]

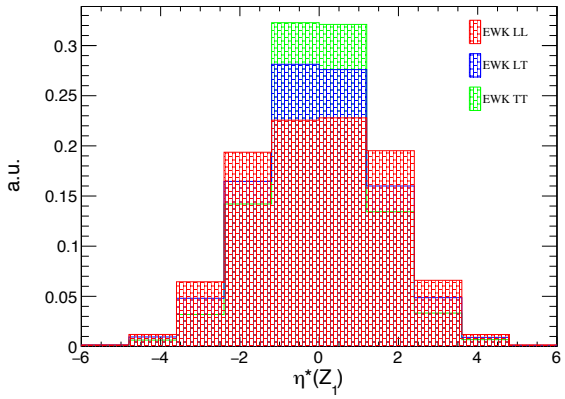

QCD qq
QCD gg
a.u.
mjj [GeV]

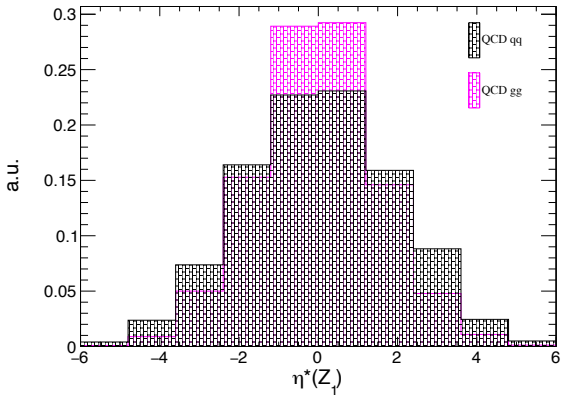

EWK LL
EWK LT
EWK TT
a.u.
Δηjj

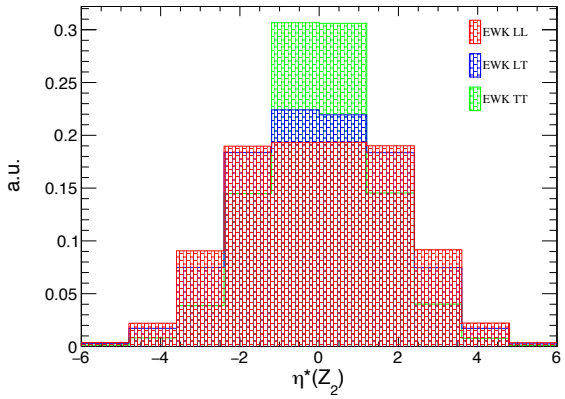

QCD qq
QCD gg
a.u.
Δηjj

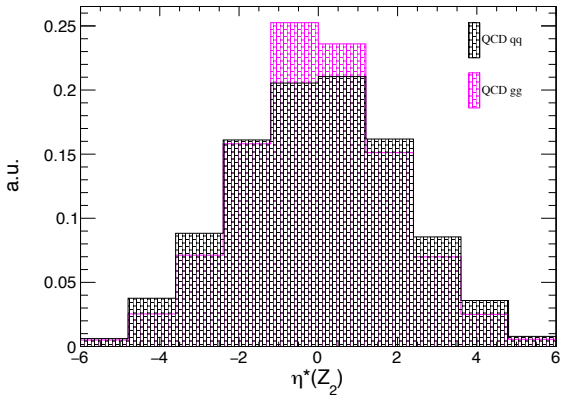

EWK LL
EWK LT
EWK TT
a.u.
m4l [GeV]

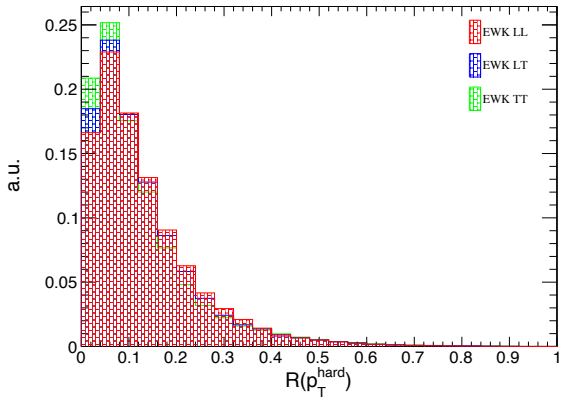

QCD qq
QCD gg
a.u.
m4l [GeV]

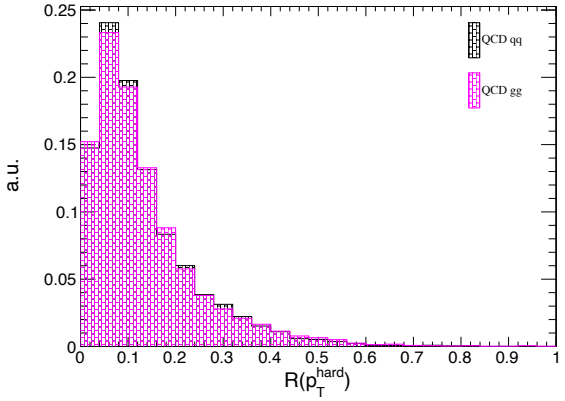

EWK LL
EWK LT
EWK TT
a.u.
η*(Z1)

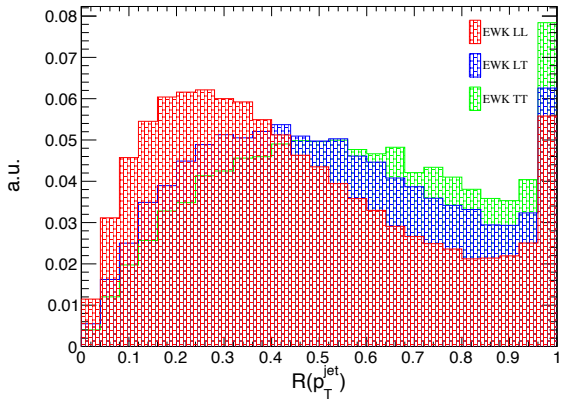

QCD qq
QCD gg
a.u.
η*(Z1)

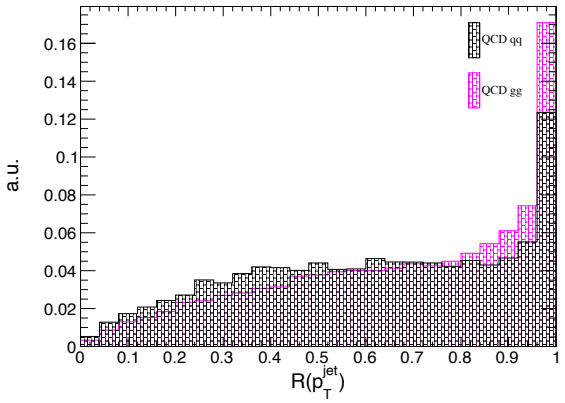

EWK LL
EWK LT
EWK TT
a.u.
η*(Z2)

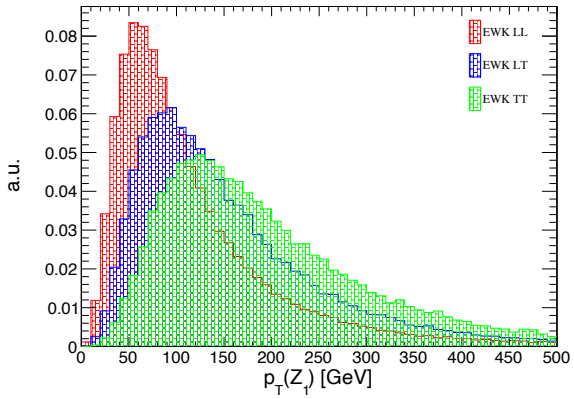

QCD qq
QCD gg
a.u.
η*(Z2)

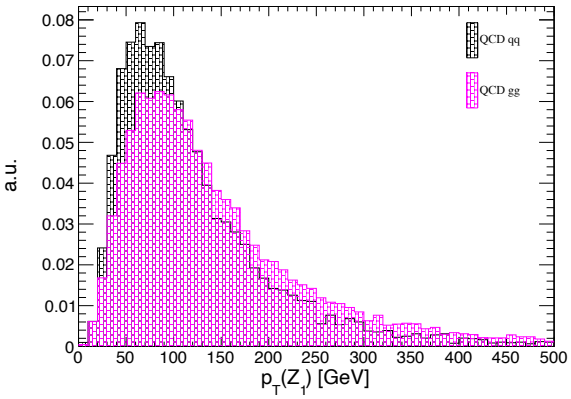

EWK LL
EWK LT
EWK TT
a.u.
R(pThard)

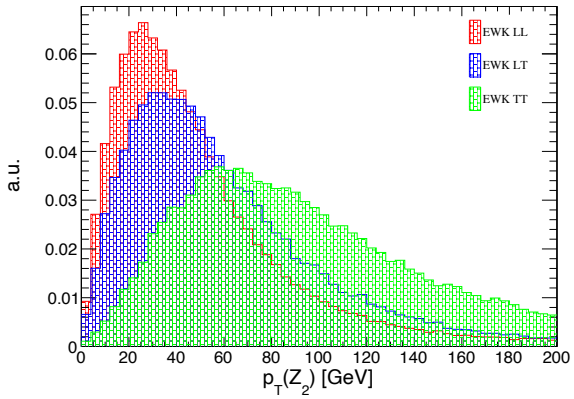

QCD qq
QCD gg
a.u.
R(pThard)

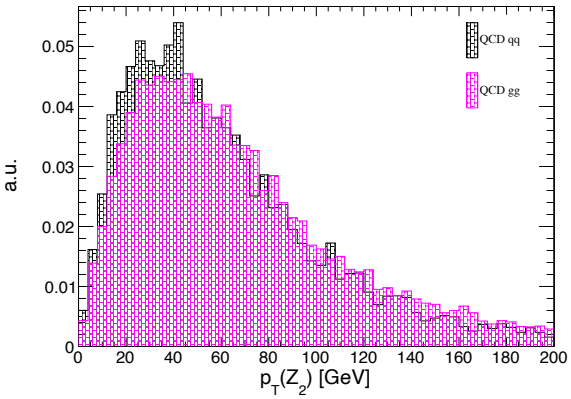

EWK LL
EWK LT
EWK TT
a.u.
R(pTjet)

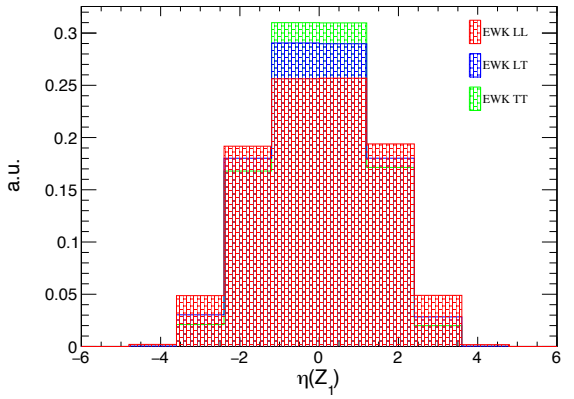

QCD qq
QCD gg
a.u.
R(pTjet)

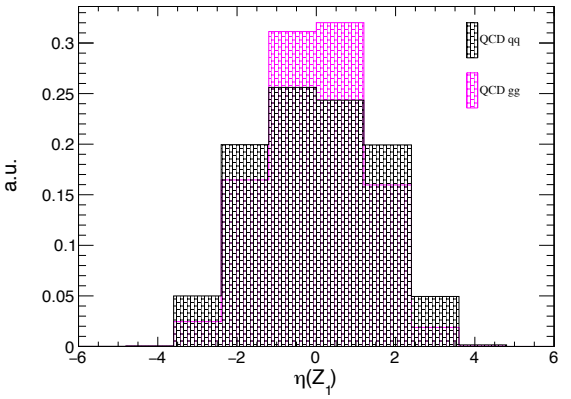

EWK LL
EWK LT
EWK TT
a.u.
$p_T(Z_1)$ [GeV]

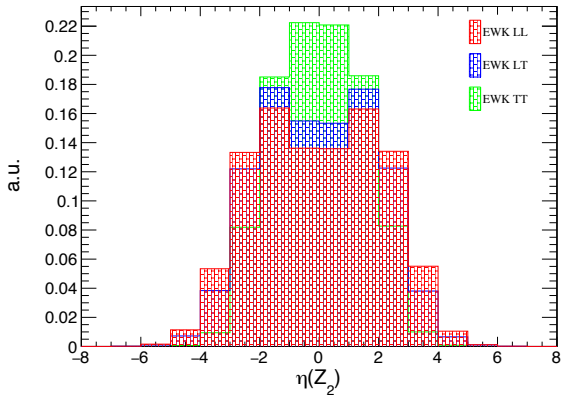

QCD qq
QCD gg
a.u.
$p_T(Z_1)$ [GeV]

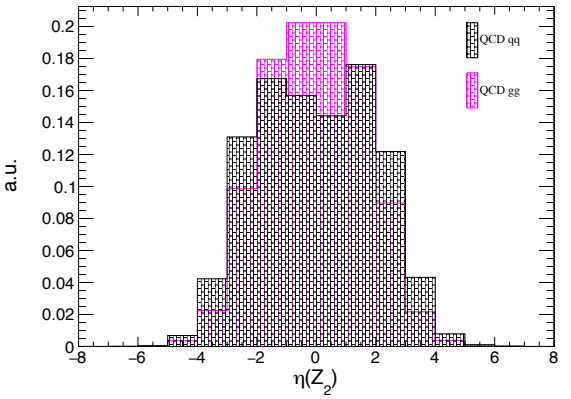

EWK LL
EWK LT
EWK TT
a.u.
$p_T(Z_2)$ [GeV]


QCD qq
QCD gg
a.u.
$p_T(Z_2)$ [GeV]


EWK LL
EWK LT
EWK TT
a.u.
$\eta(Z_1)$


QCD qq
QCD gg
a.u.
$\eta(Z_1)$


EWK LL
EWK LT
EWK TT
a.u.
$\eta(Z_2)$


QCD qq
QCD gg
a.u.
$\eta(Z_2)$

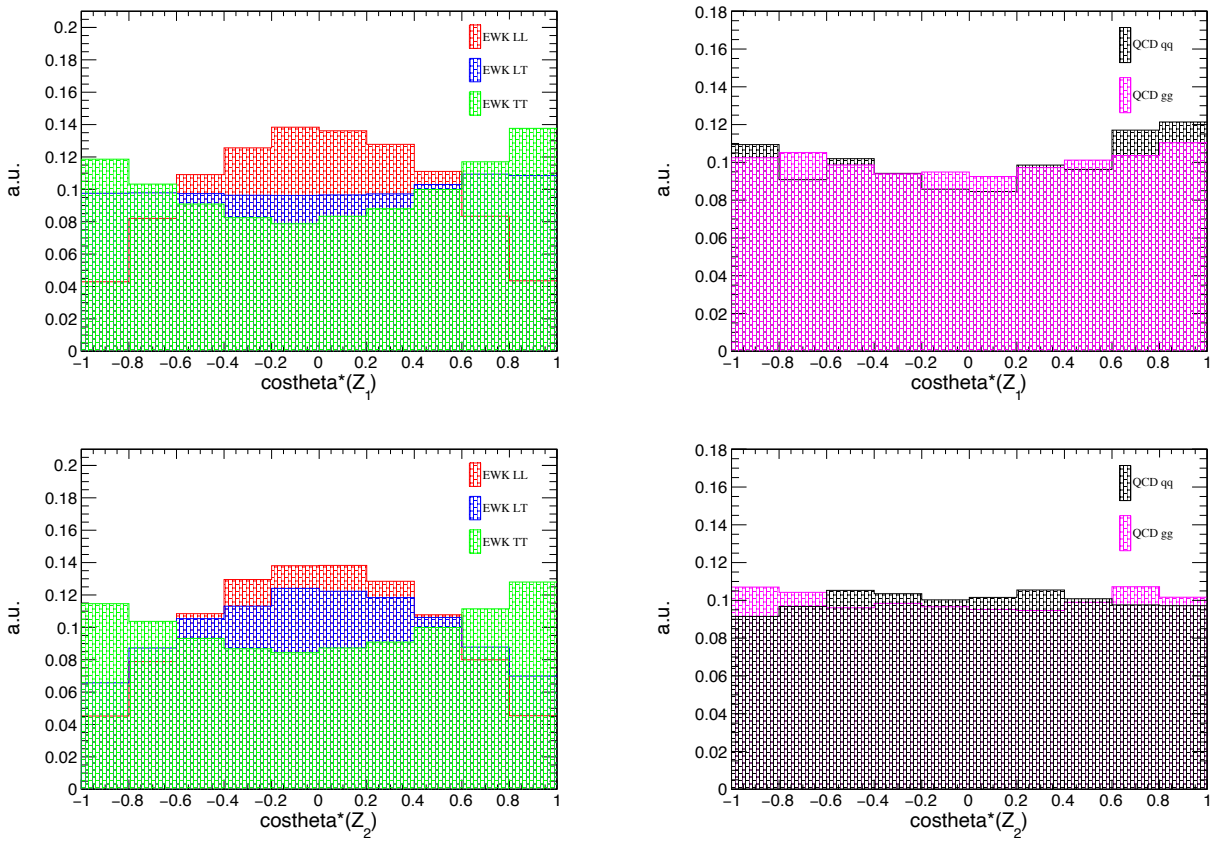
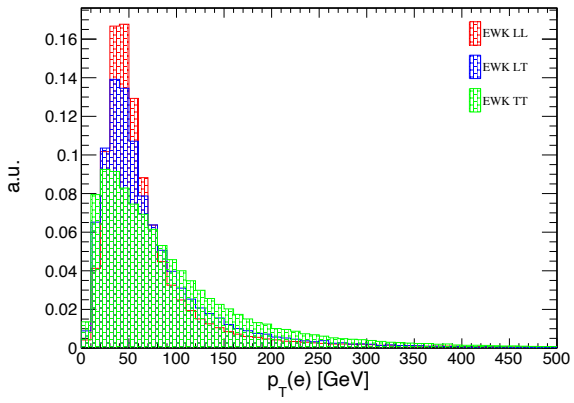
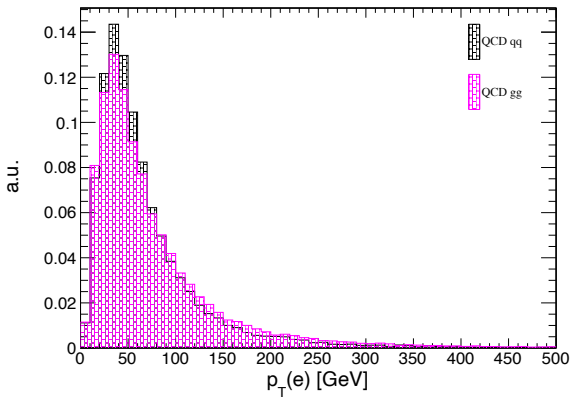
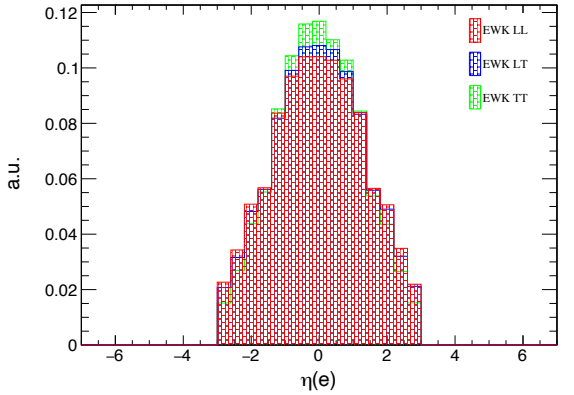
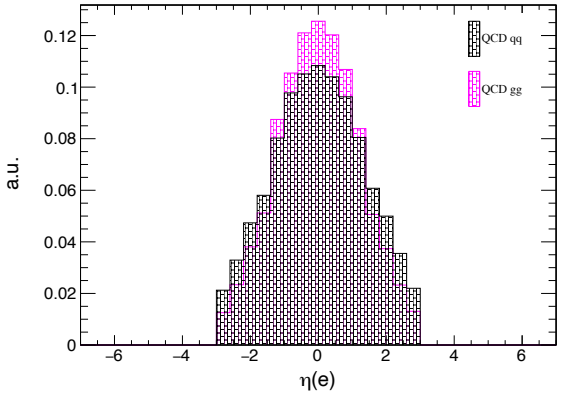


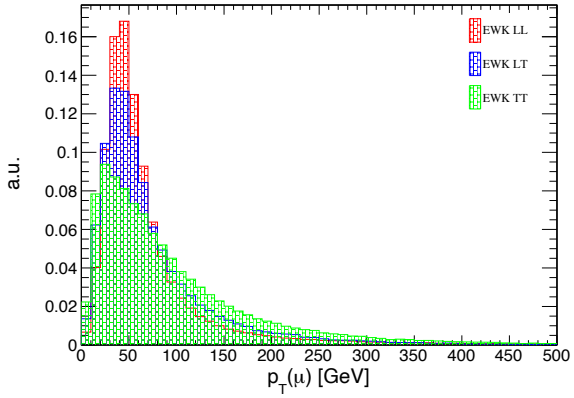
Figure B.1: Kinematics of VBS (left) and QCD (right) processes at 14 TeV after the baseline selection.

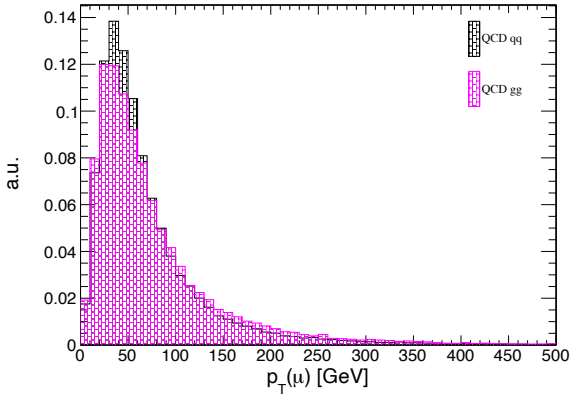

EWK LL
EWK LT
EWK TT
a.u.
p_T(e) [GeV]

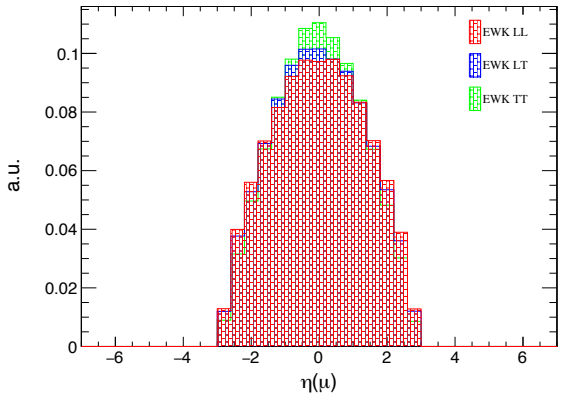

QCD qq
QCD gg
a.u.
p_T(e) [GeV]

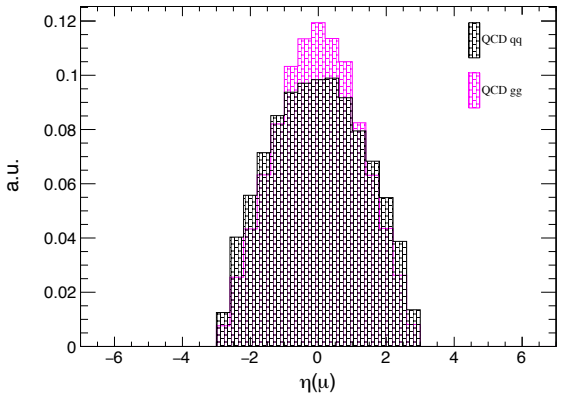

EWK LL
EWK LT
EWK TT
a.u.
η(e)

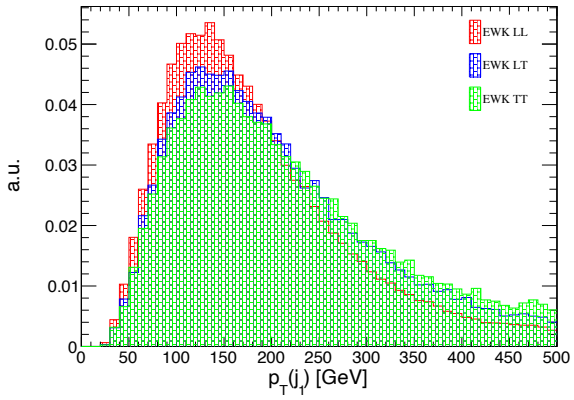

QCD qq
QCD gg
a.u.
η(e)

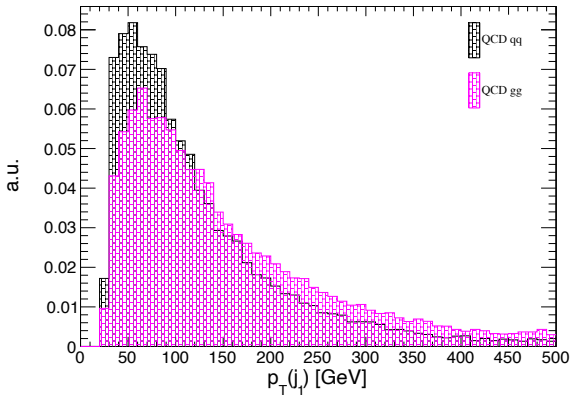

EWK LL
EWK LT
EWK TT
a.u.
p_T(μ) [GeV]

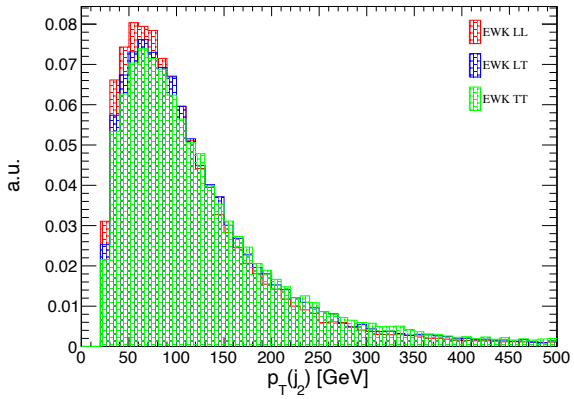

QCD qq
QCD gg
a.u.
p_T(μ) [GeV]

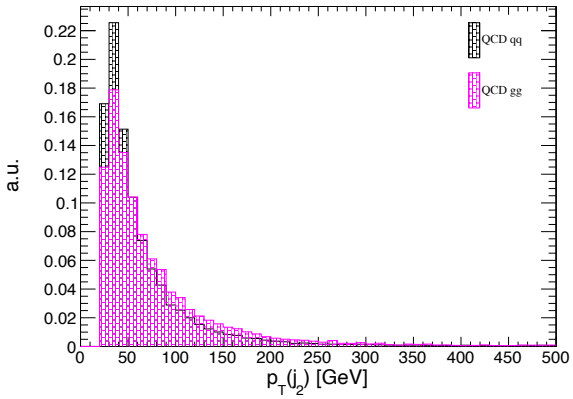

EWK LL
EWK LT
EWK TT
a.u.
η(μ)

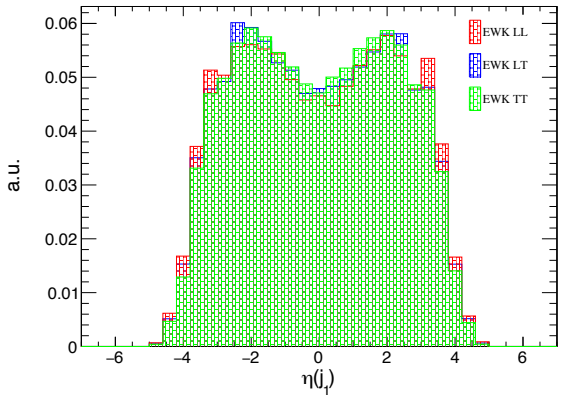

QCD qq
QCD gg
a.u.
η(μ)

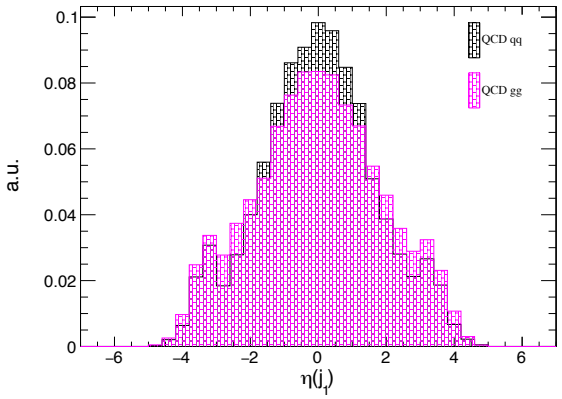

EWK LL
EWK LT
EWK TT
a.u.
p_T(j_1) [GeV]

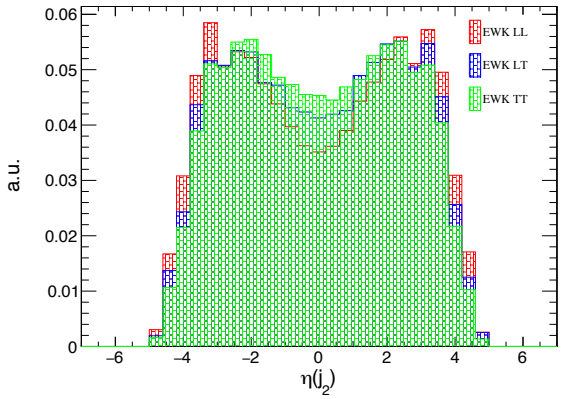

QCD qq
QCD gg
a.u.
p_T(j_1) [GeV]

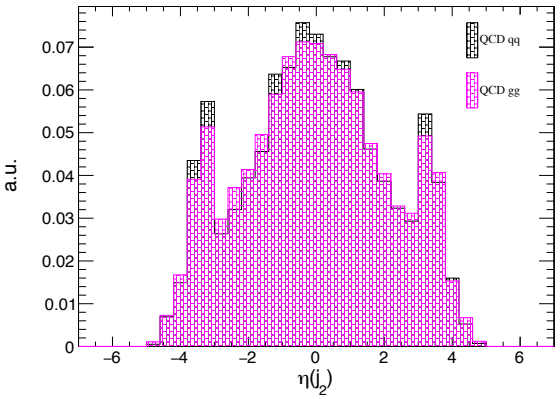

EWK LL
EWK LT
EWK TT
a.u.
p_T(j_2) [GeV]

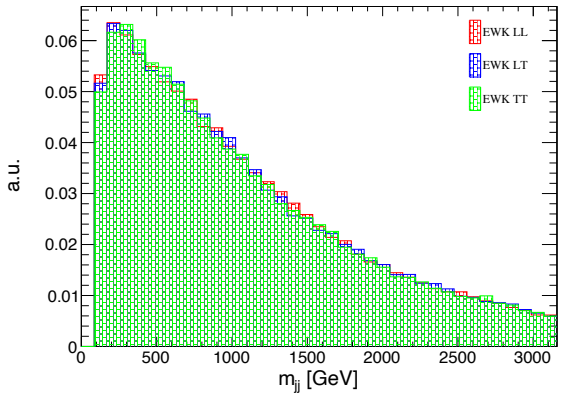

QCD qq
QCD gg
a.u.
p_T(j_2) [GeV]

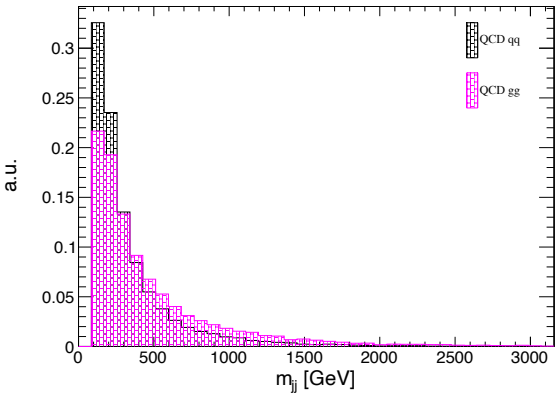

EWK LL
EWK LT
EWK TT
a.u.
η(j_1)

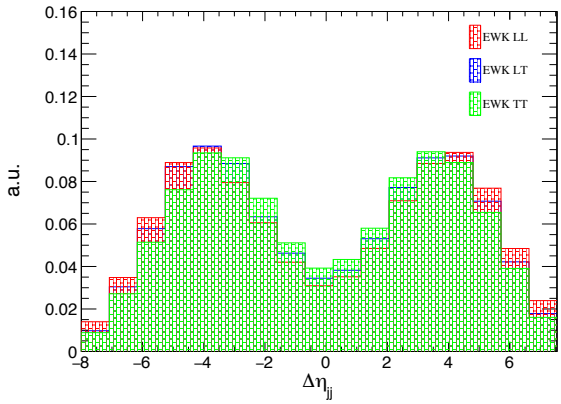

QCD qq
QCD gg
a.u.
η(j_1)

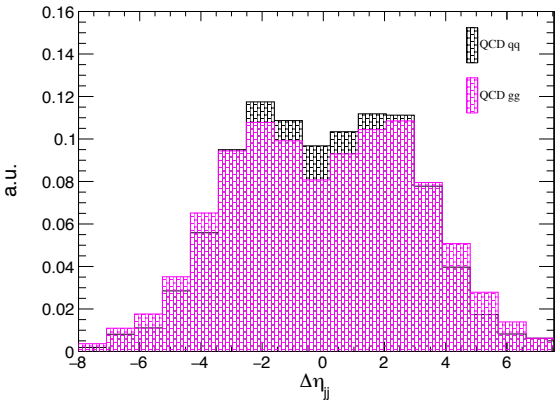

EWK LL
EWK LT
EWK TT
a.u.
η(j2)

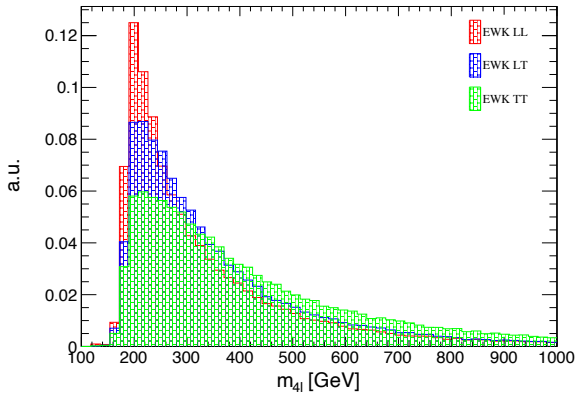

QCD qq
QCD gg
a.u.
η(j2)

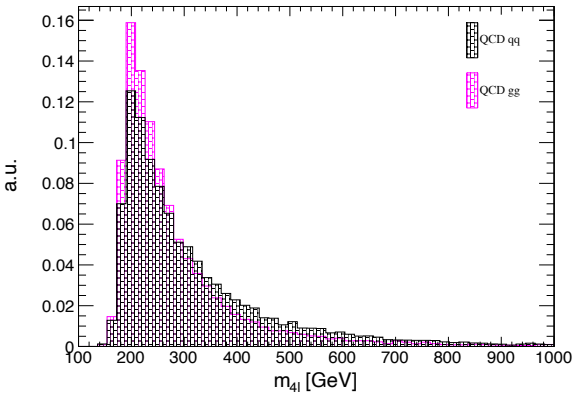

EWK LL
EWK LT
EWK TT
a.u.
mjj [GeV]

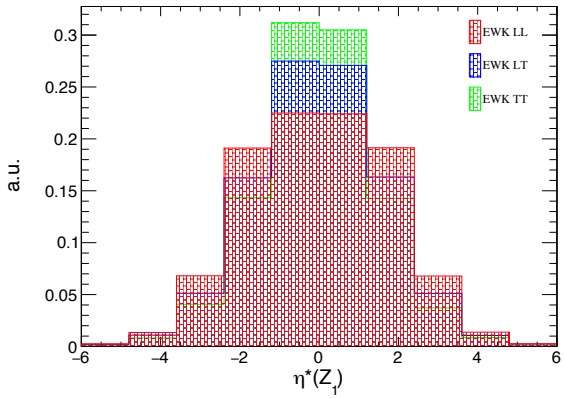

QCD qq
QCD gg
a.u.
mjj [GeV]

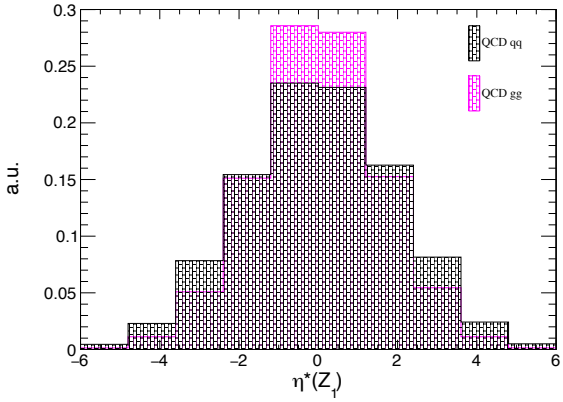

EWK LL
EWK LT
EWK TT
a.u.
Δηjj

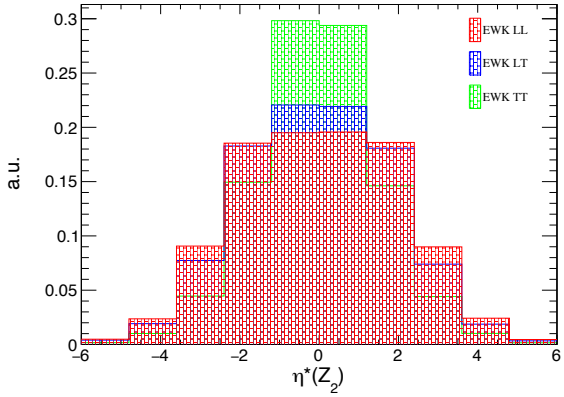

QCD qq
QCD gg
a.u.
Δηjj

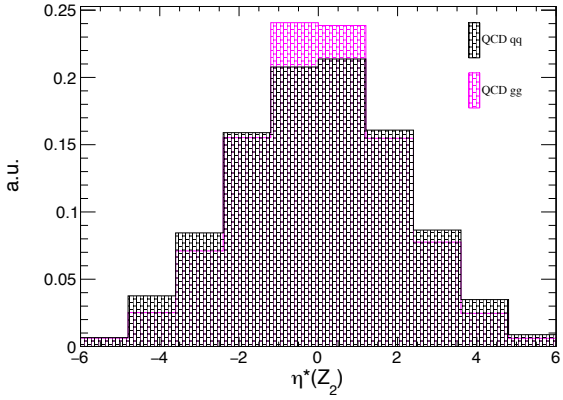

EWK LL
EWK LT
EWK TT
a.u.
m4l [GeV]

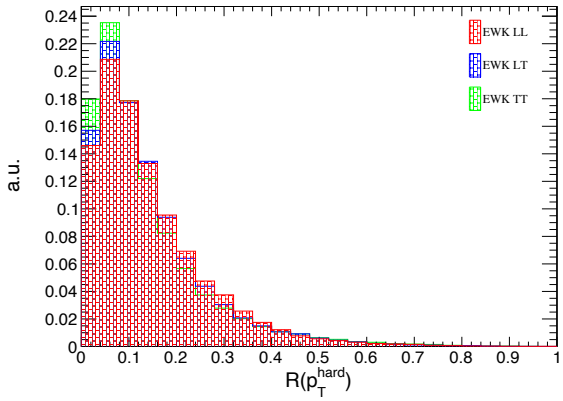

QCD qq
QCD gg
a.u.
m4l [GeV]

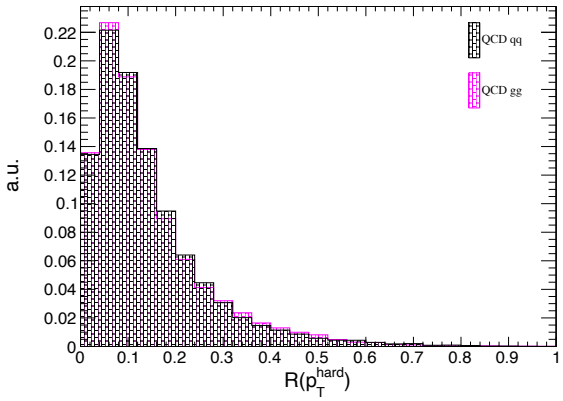

EWK LL
EWK LT
EWK TT
a.u.
η*(Z1)

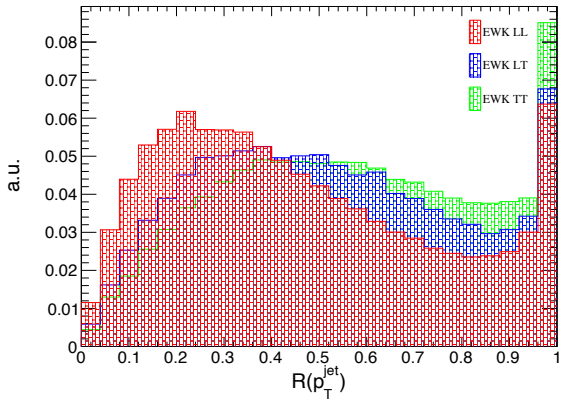

QCD qq
QCD gg
a.u.
η*(Z1)

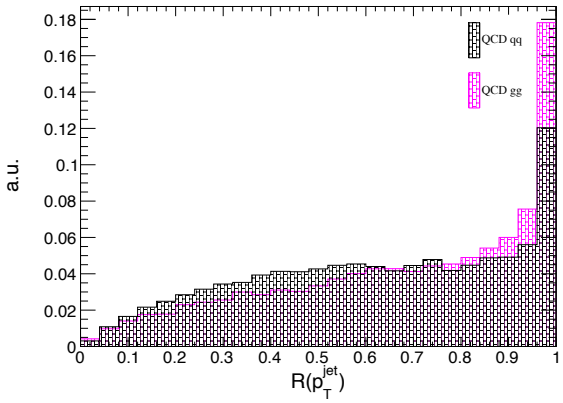

EWK LL
EWK LT
EWK TT
a.u.
η*(Z2)

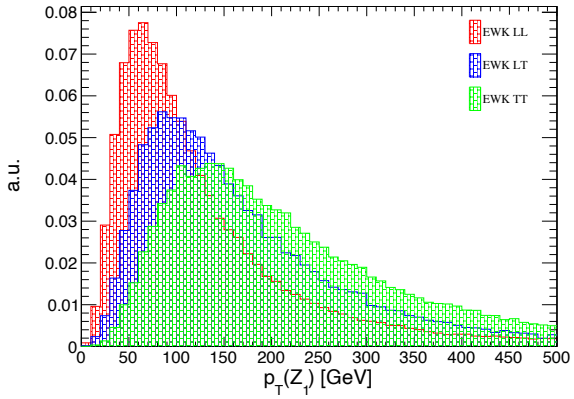

QCD qq
QCD gg
a.u.
η*(Z2)

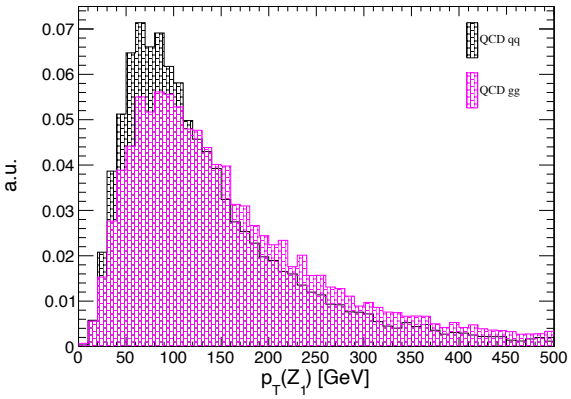

EWK LL
EWK LT
EWK TT
a.u.
R(pThard)

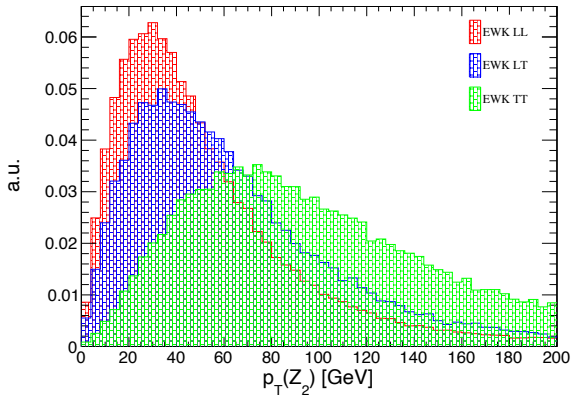

QCD qq
QCD gg
a.u.
R(pThard)

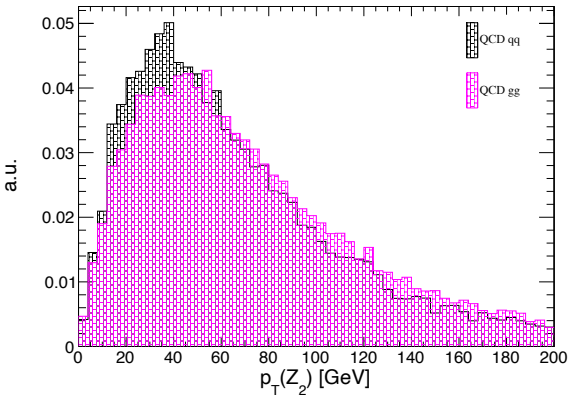

EWK LL
EWK LT
EWK TT
a.u.
R(pTjet)

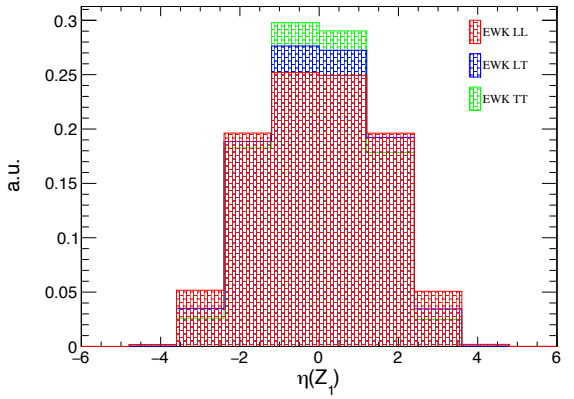

QCD qq
QCD gg
a.u.
R(pTjet)

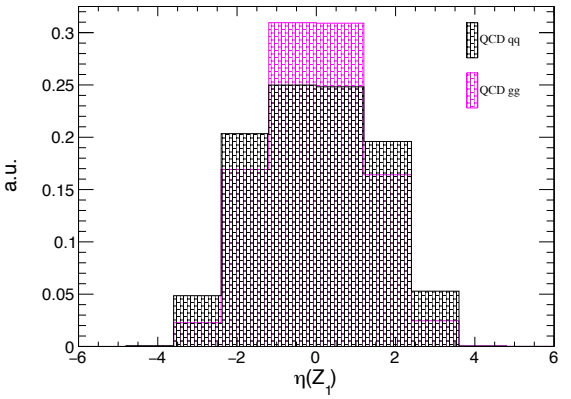

EWK LL
EWK LT
EWK TT
a.u.
$p_T(Z_1)$ [GeV]

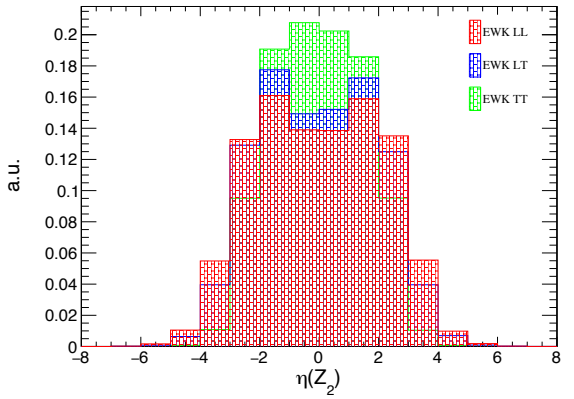

QCD qq
QCD gg
a.u.
$p_T(Z_1)$ [GeV]

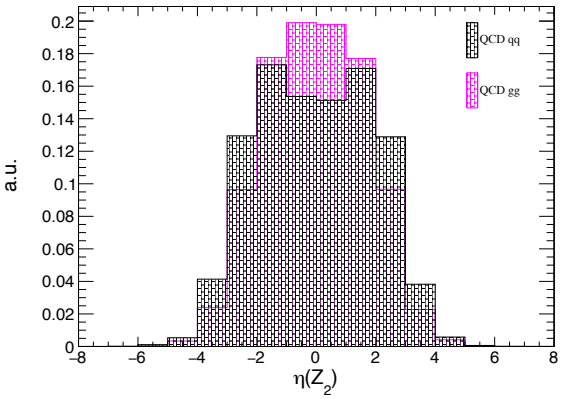

EWK LL
EWK LT
EWK TT
a.u.
$p_T(Z_2)$ [GeV]


QCD qq
QCD gg
a.u.
$p_T(Z_2)$ [GeV]


EWK LL
EWK LT
EWK TT
a.u.
$\eta(Z_1)$


QCD qq
QCD gg
a.u.
$\eta(Z_1)$


EWK LL
EWK LT
EWK TT
a.u.
$\eta(Z_2)$


QCD qq
QCD gg
a.u.
$\eta(Z_2)$

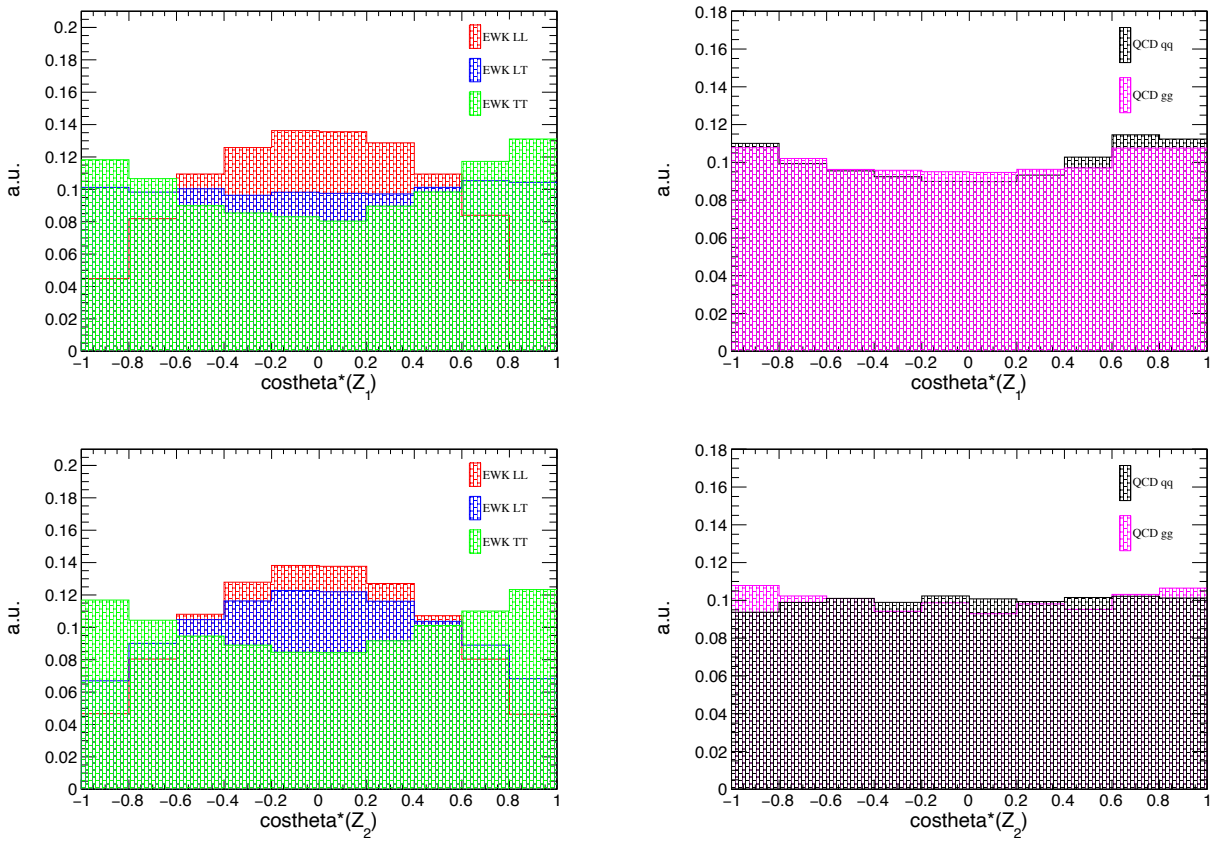
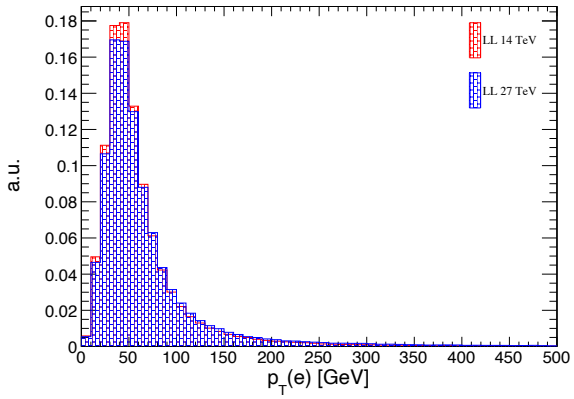
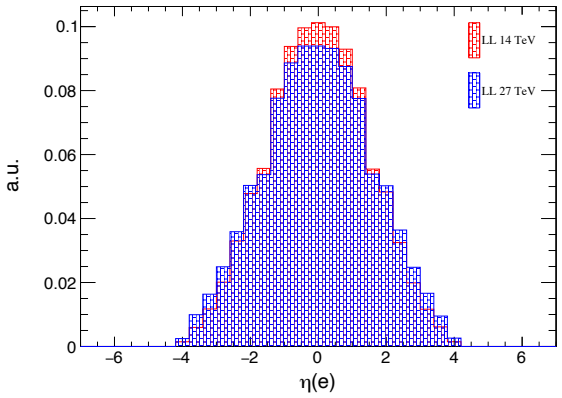
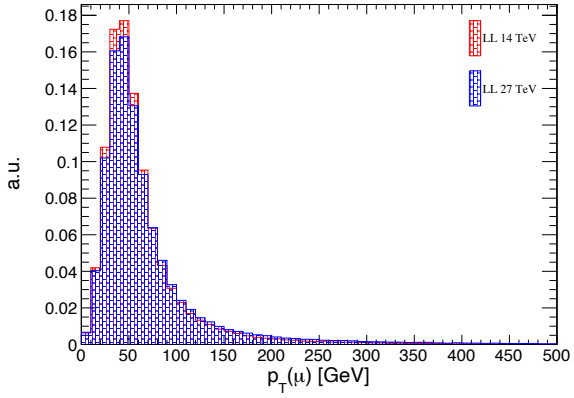
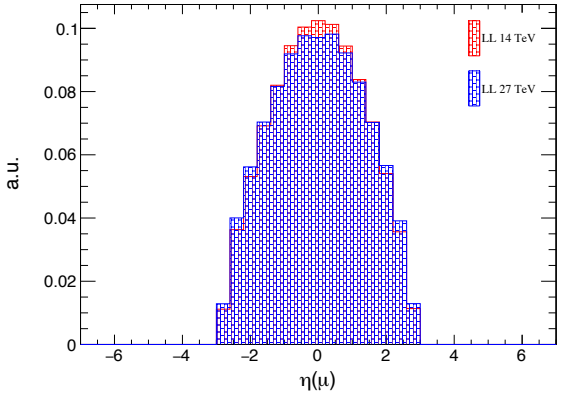


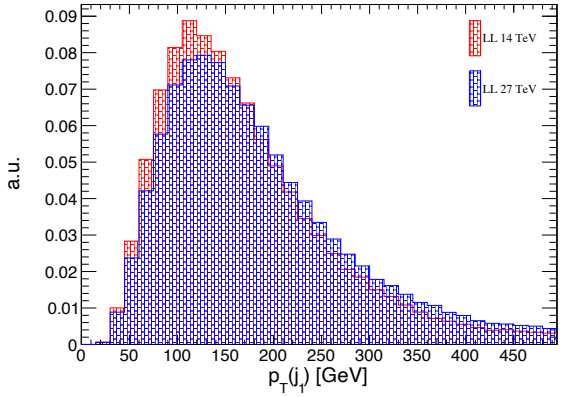
Figure B.2: Kinematics of VBS (left) and QCD (right) processes at 27 TeV after the baseline selection.

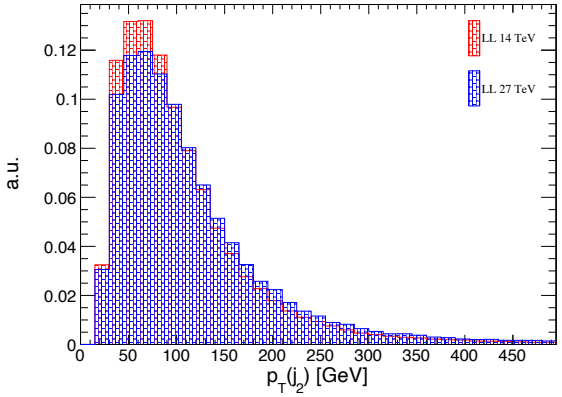

a.u.
LL 14 TeV
LL 27 TeV
$p_T(e)$ [GeV]

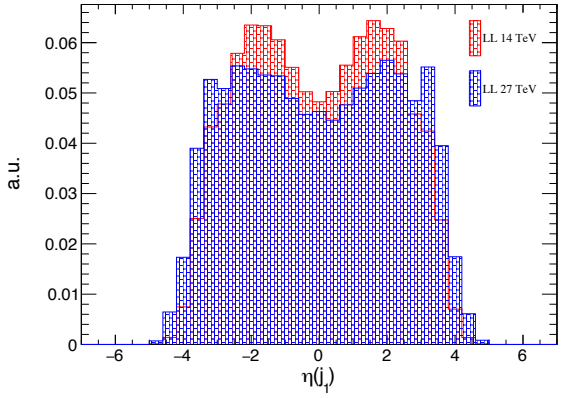

a.u.
LL 14 TeV
LL 27 TeV
η(e)

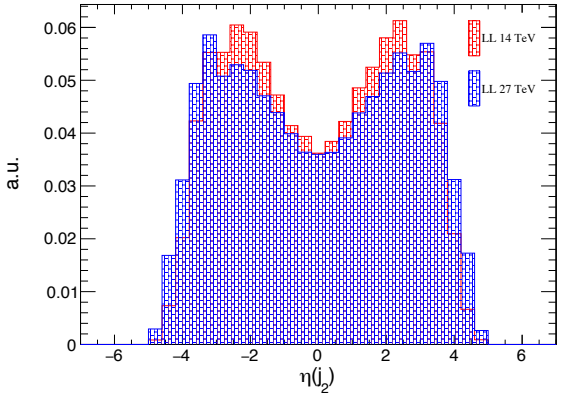

a.u.
LL 14 TeV
LL 27 TeV
$p_T(\mu)$ [GeV]

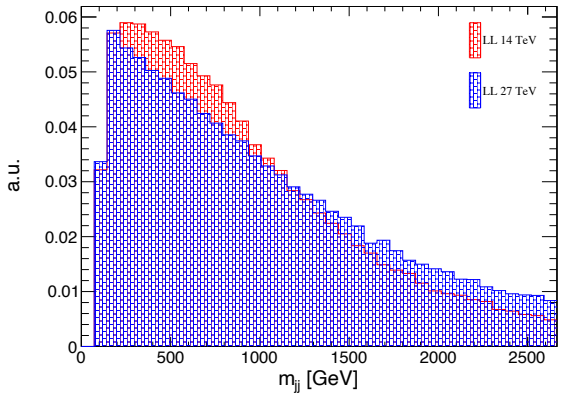

a.u.
LL 14 TeV
LL 27 TeV
η(μ)

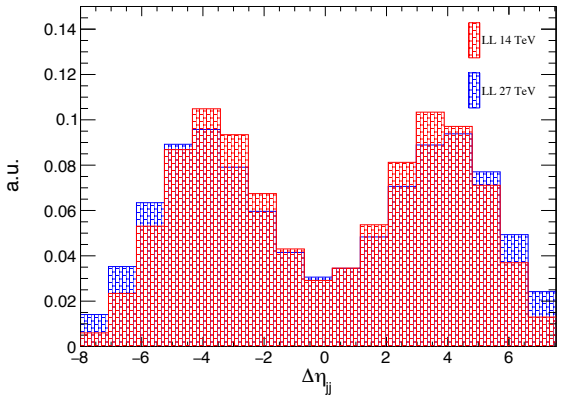

a.u.
LL 14 TeV
LL 27 TeV
$p_T(j_1)$ [GeV]

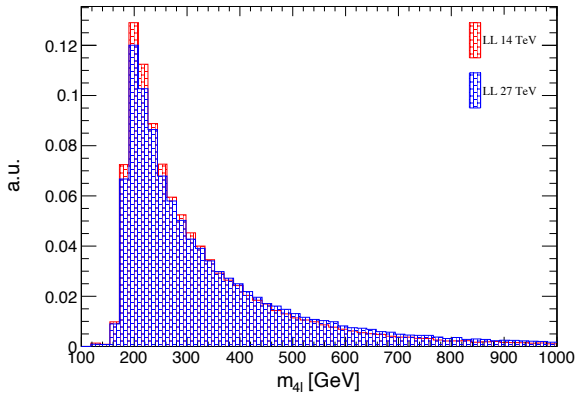

a.u.
LL 14 TeV
LL 27 TeV
$p_T(j_2)$ [GeV]

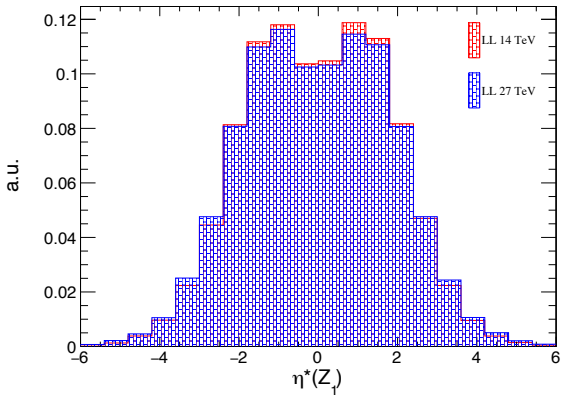

a.u.
LL 14 TeV
LL 27 TeV
$\eta(j_1)$

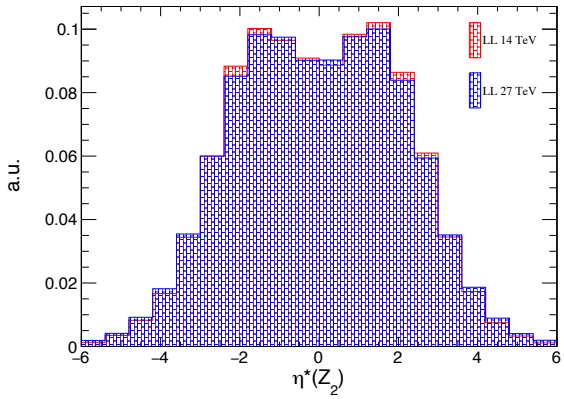

a.u.
LL 14 TeV
LL 27 TeV
$\eta(j_2)$

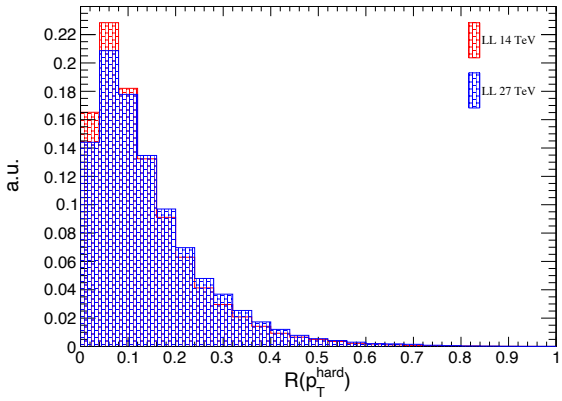

a.u.
LL 14 TeV
LL 27 TeV
m_{jj} [GeV]

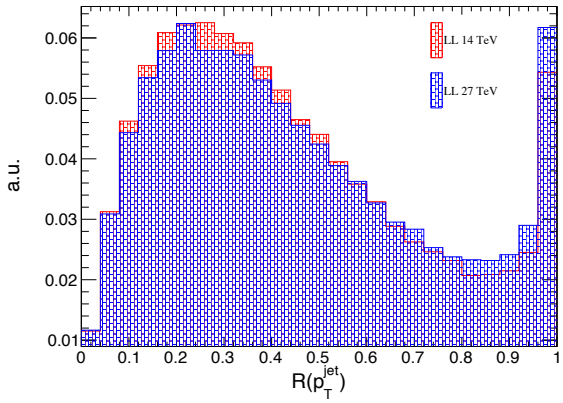

a.u.
LL 14 TeV
LL 27 TeV
$\Delta\eta_{jj}$

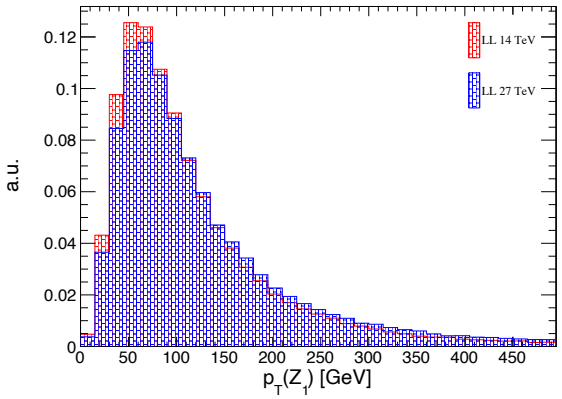

a.u.
LL 14 TeV
LL 27 TeV
m_{4l} [GeV]


a.u.
LL 14 TeV
LL 27 TeV
$\eta^*(Z_1)$


a.u.
LL 14 TeV
LL 27 TeV
$\eta^*(Z_2)$


a.u.
LL 14 TeV
LL 27 TeV
$R(p_T^{hard})$


a.u.
LL 14 TeV
LL 27 TeV
$R(p_T^{jet})$


a.u.
LL 14 TeV
LL 27 TeV
$p_T(Z_1)$ [GeV]

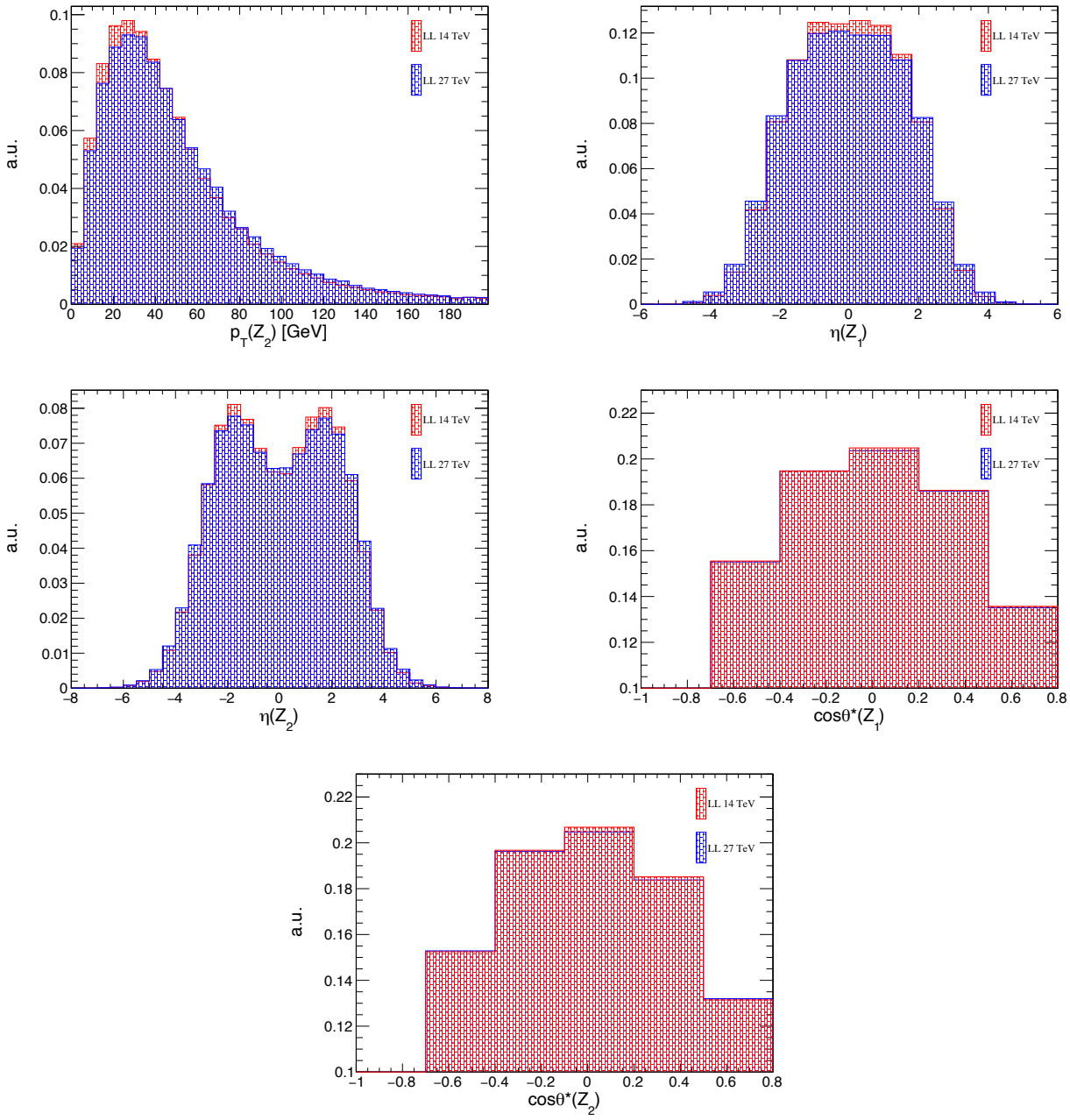
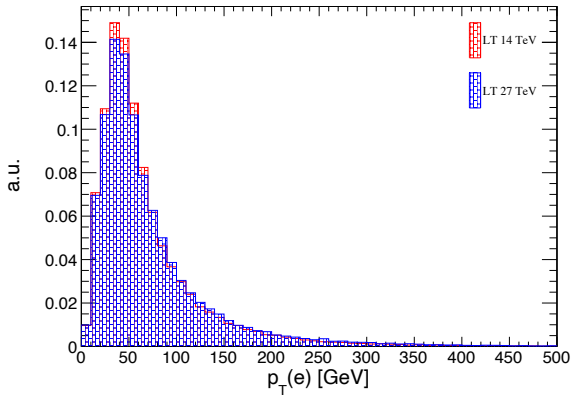
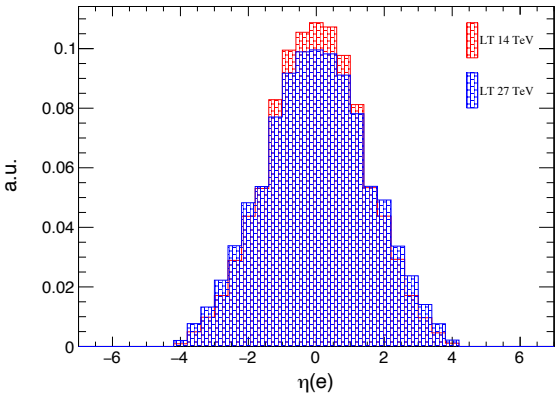
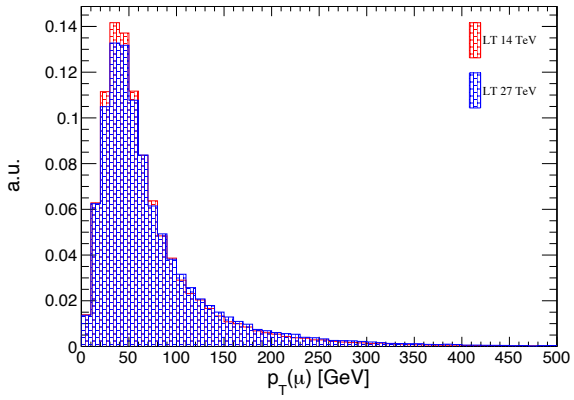
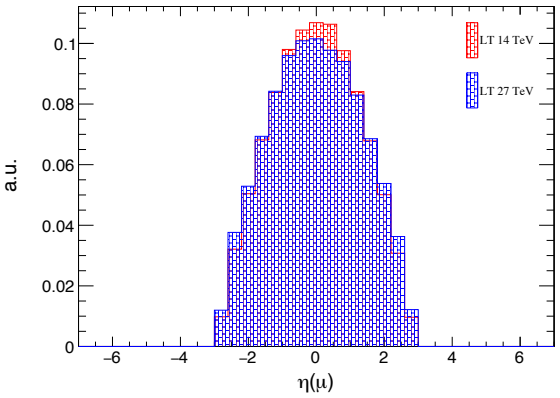


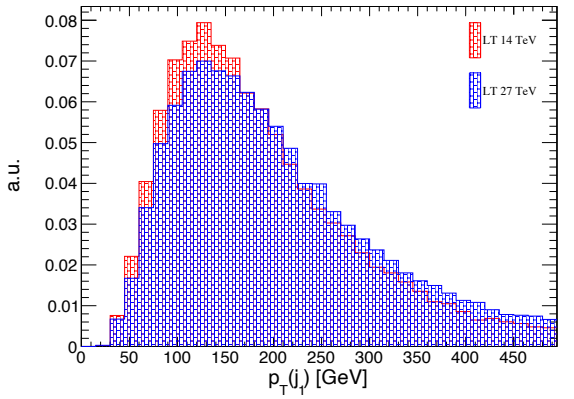
Figure B.3: Kinematics of the LL process at 14 and 27 TeV after the baseline selection.

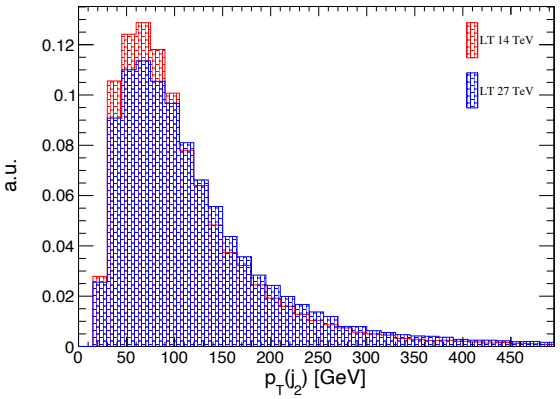

a.u.
LT 14 TeV
LT 27 TeV
p_T(e) [GeV]

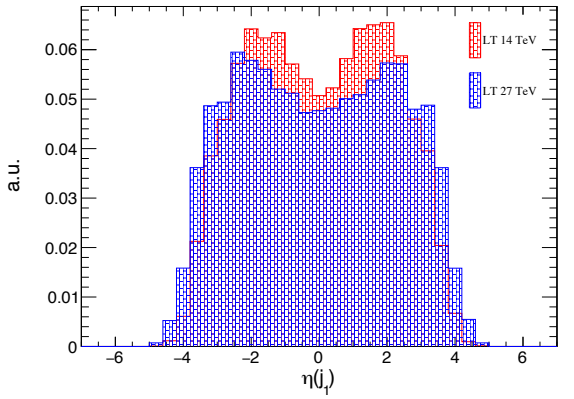

a.u.
LT 14 TeV
LT 27 TeV
η(e)

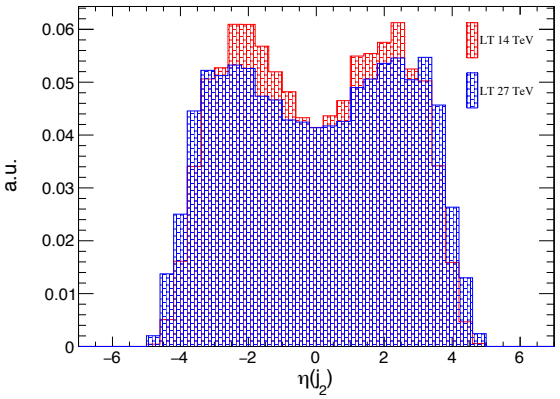

a.u.
LT 14 TeV
LT 27 TeV
p_T(μ) [GeV]

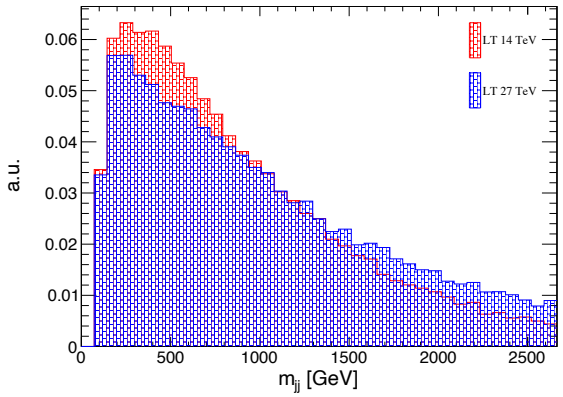

a.u.
LT 14 TeV
LT 27 TeV
η(μ)

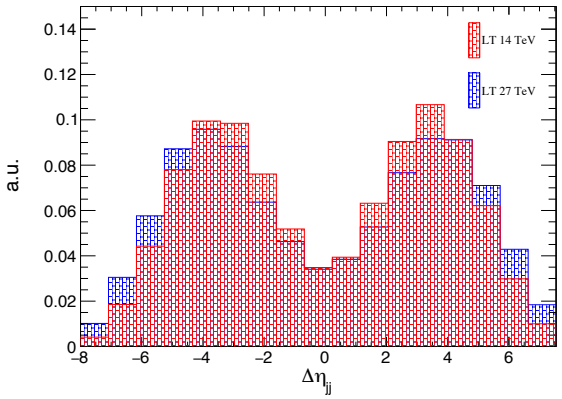

a.u.
LT 14 TeV
LT 27 TeV
$p_T(j_1)$ [GeV]

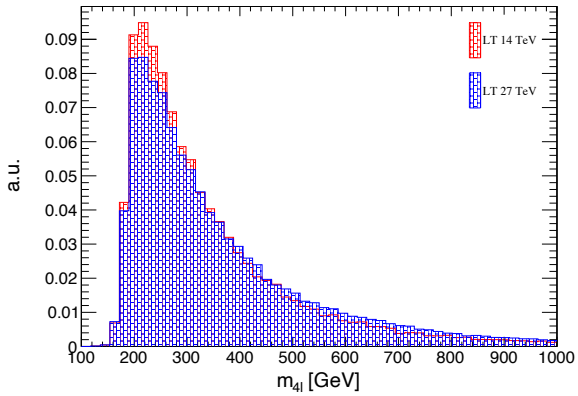

a.u.
LT 14 TeV
LT 27 TeV
$p_T(j_2)$ [GeV]

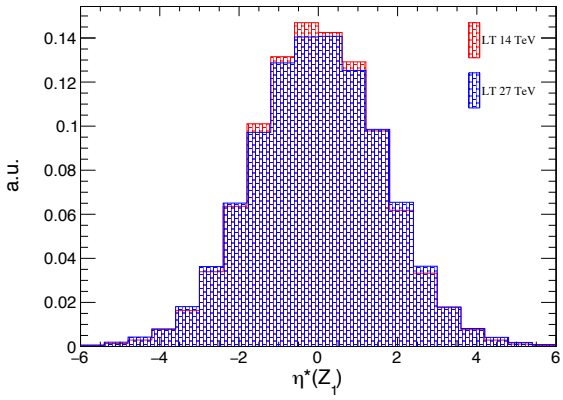

a.u.
LT 14 TeV
LT 27 TeV
η(j_1)

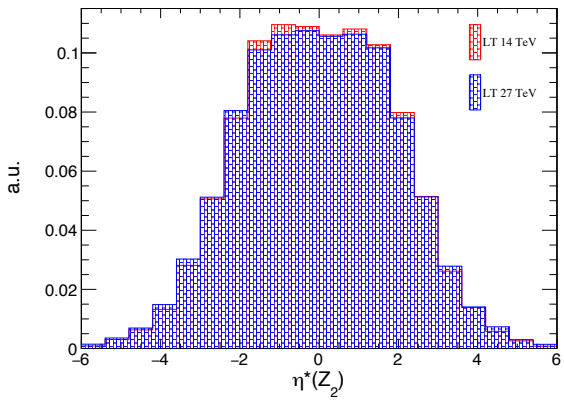

a.u.
LT 14 TeV
LT 27 TeV
η(j_2)

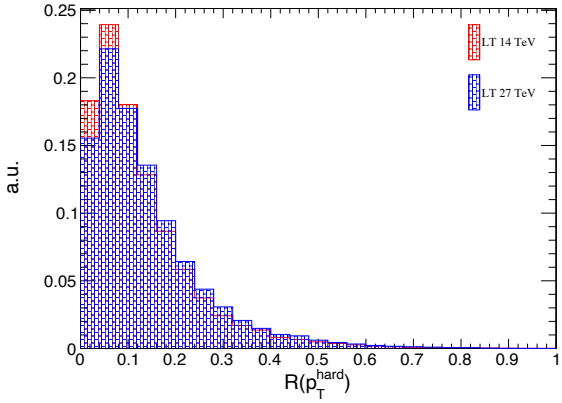

LT 14 TeV
LT 27 TeV
a.u.
m_{jj} [GeV]

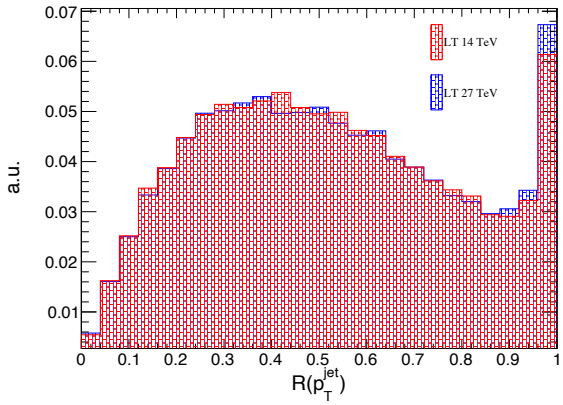

LT 14 TeV
LT 27 TeV
a.u.
$\Delta\eta_{jj}$

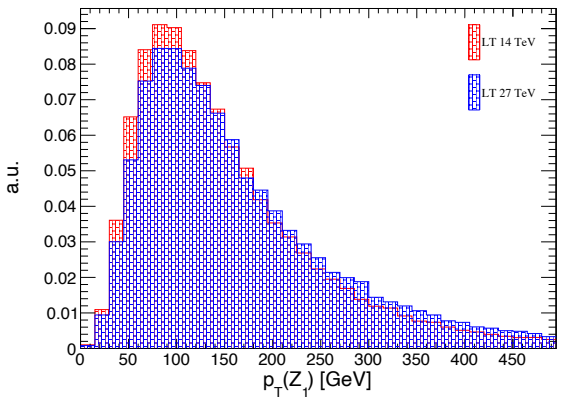

LT 14 TeV
LT 27 TeV
a.u.
m_{4l} [GeV]


LT 14 TeV
LT 27 TeV
a.u.
$\eta^*(Z_1)$


LT 14 TeV
LT 27 TeV
a.u.
$\eta^*(Z_2)$


LT 14 TeV
LT 27 TeV
a.u.
$R(p_T^{hard})$


LT 14 TeV
LT 27 TeV
a.u.
$R(p_T^{jet})$


LT 14 TeV
LT 27 TeV
a.u.
$p_T(Z_1)$ [GeV]

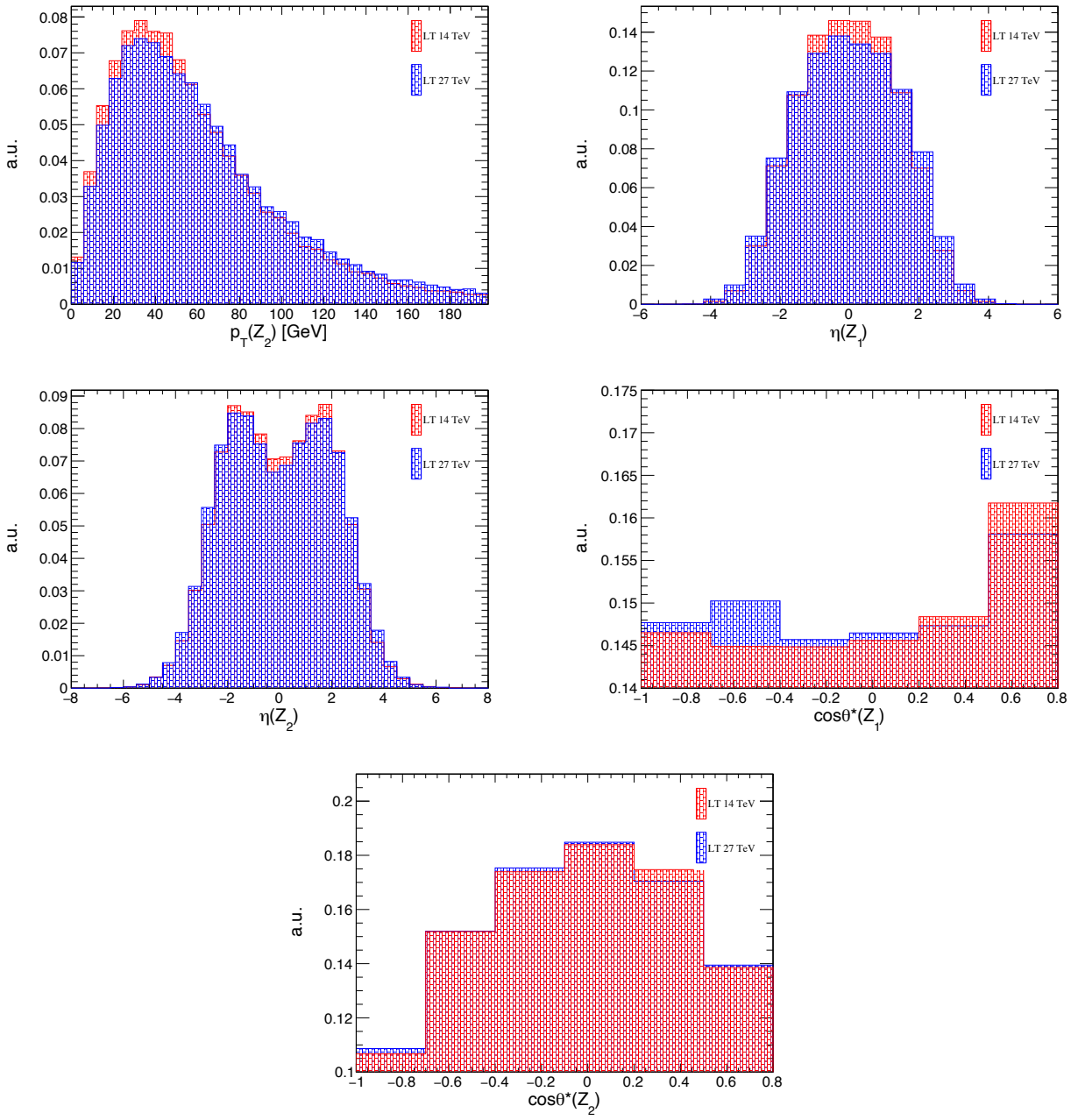
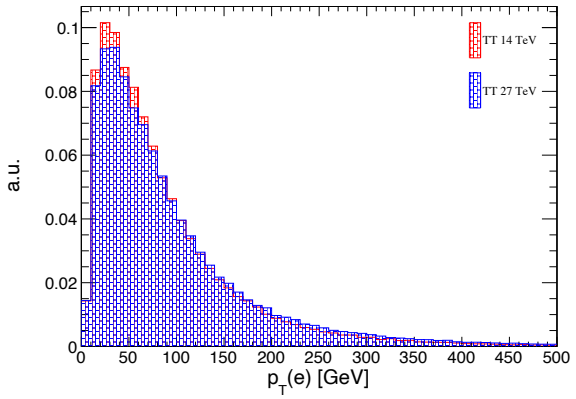
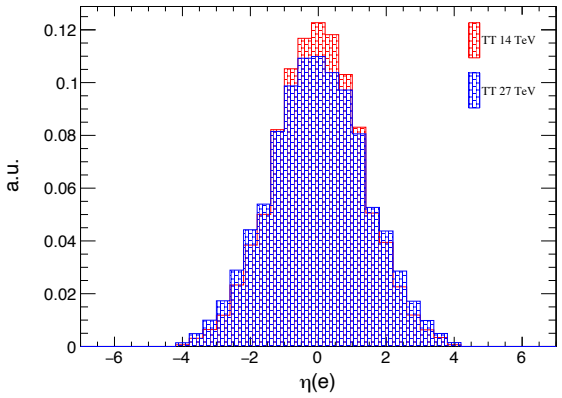
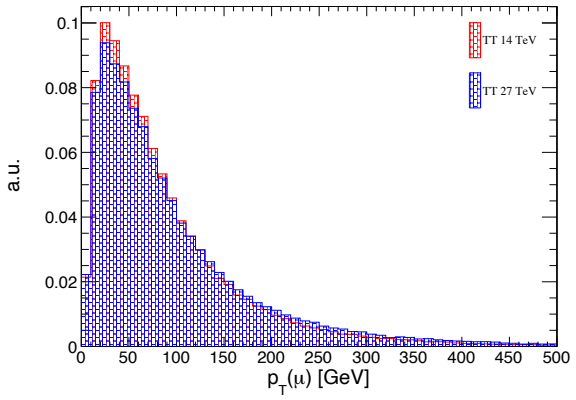
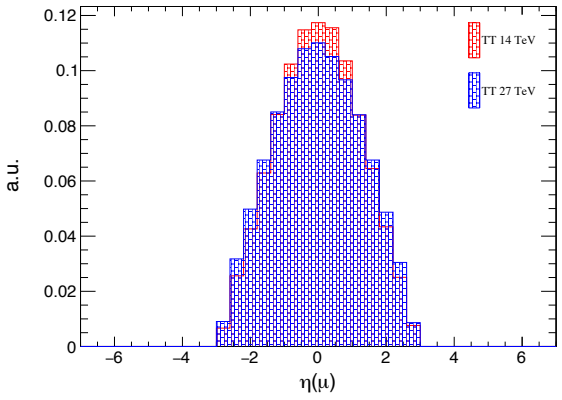


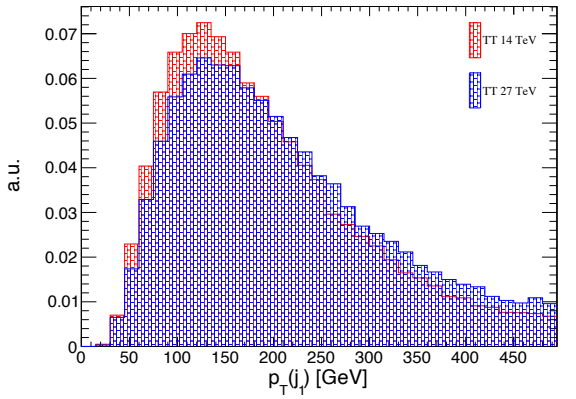
Figure B.4: Kinematics of the the LT process at 14 and 27 TeV after the baseline selection.

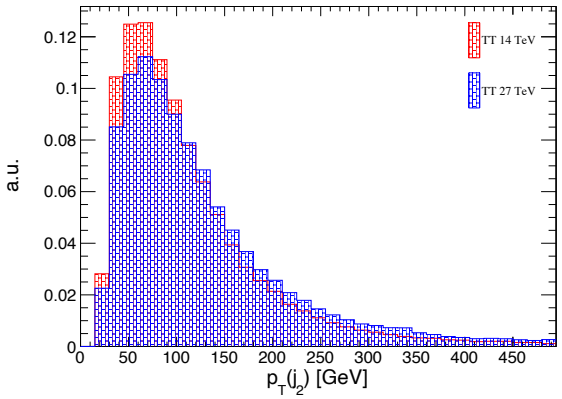

TT 14 TeV
TT 27 TeV
a.u.
$p_T(e)$ [GeV]

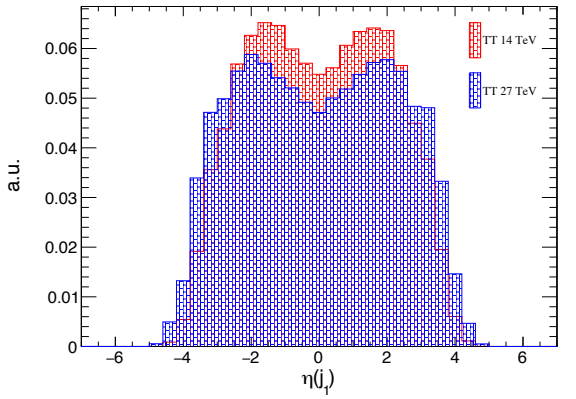

TT 14 TeV
TT 27 TeV
a.u.
η(e)

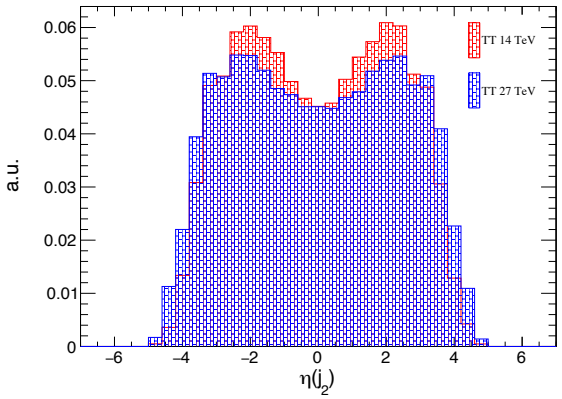

TT 14 TeV
TT 27 TeV
a.u.
$p_T(\mu)$ [GeV]

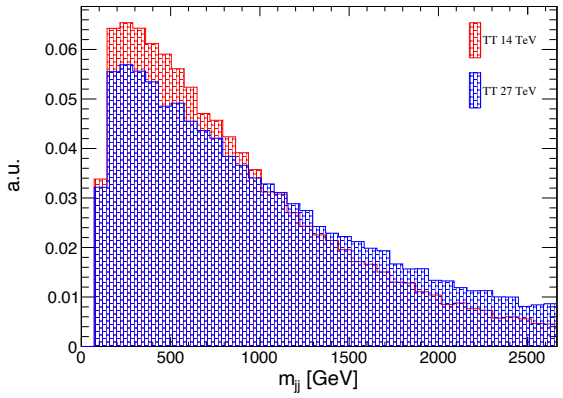

TT 14 TeV
TT 27 TeV
a.u.
η(μ)

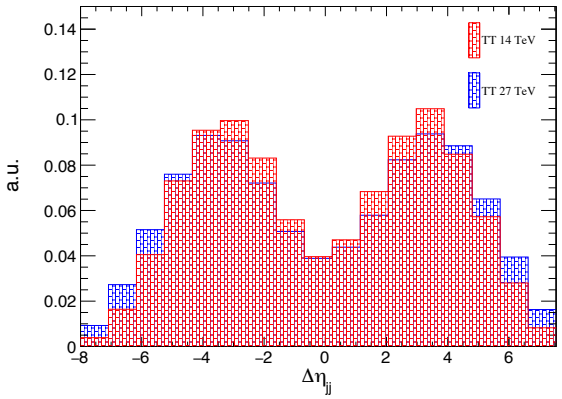

TT 14 TeV
TT 27 TeV
a.u.
$p_T(j_1)$ [GeV]

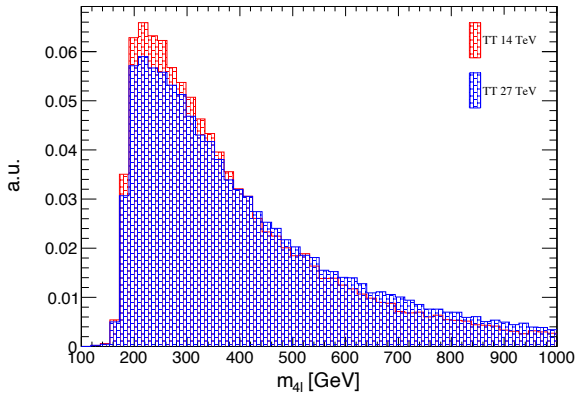

TT 14 TeV
TT 27 TeV
a.u.
$p_T(j_2)$ [GeV]

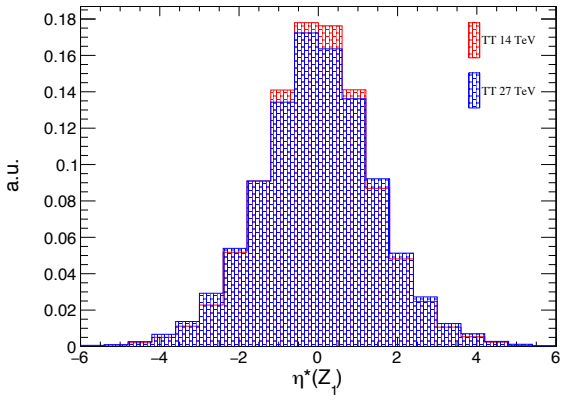

TT 14 TeV
TT 27 TeV
a.u.
$\eta(j_1)$

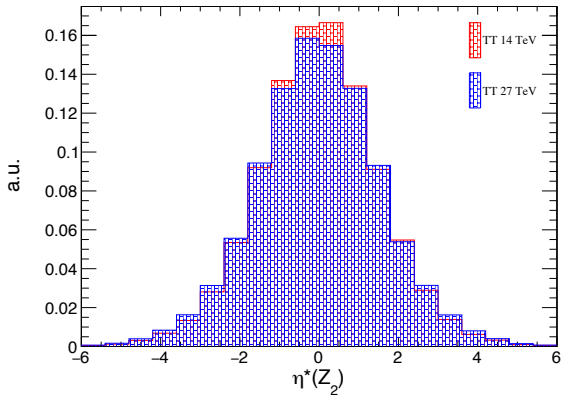

TT 14 TeV
TT 27 TeV
a.u.
$\eta(j_2)$

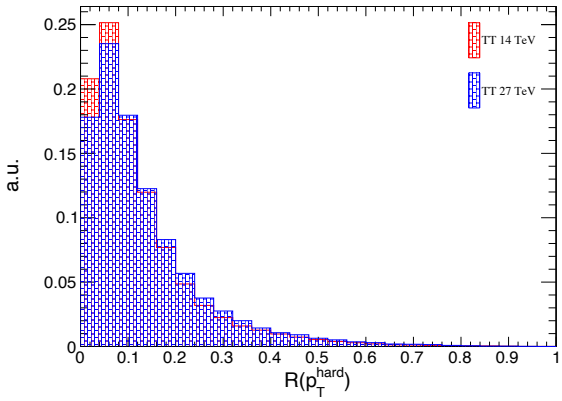

TT 14 TeV
TT 27 TeV
a.u.
m_{jj} [GeV]

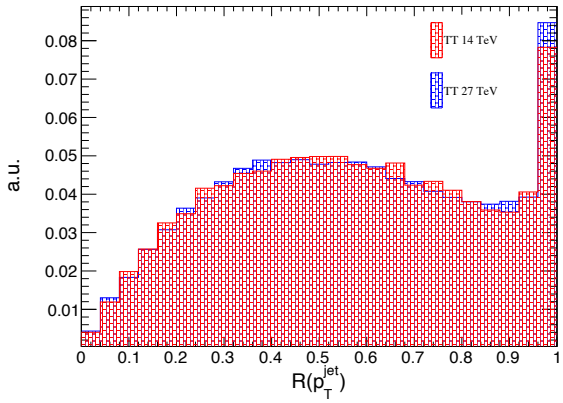

TT 14 TeV
TT 27 TeV
a.u.
$\Delta\eta_{jj}$

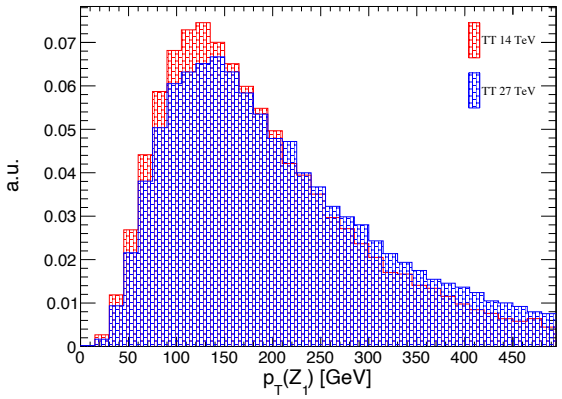

TT 14 TeV
TT 27 TeV
a.u.
m_{4l} [GeV]


TT 14 TeV
TT 27 TeV
a.u.
$\eta^*(Z_1)$


TT 14 TeV
TT 27 TeV
a.u.
$\eta^*(Z_2)$


TT 14 TeV
TT 27 TeV
a.u.
$R(p_T^{hard})$


TT 14 TeV
TT 27 TeV
a.u.
$R(p_T^{jet})$


TT 14 TeV
TT 27 TeV
a.u.
$p_T(Z_1)$ [GeV]

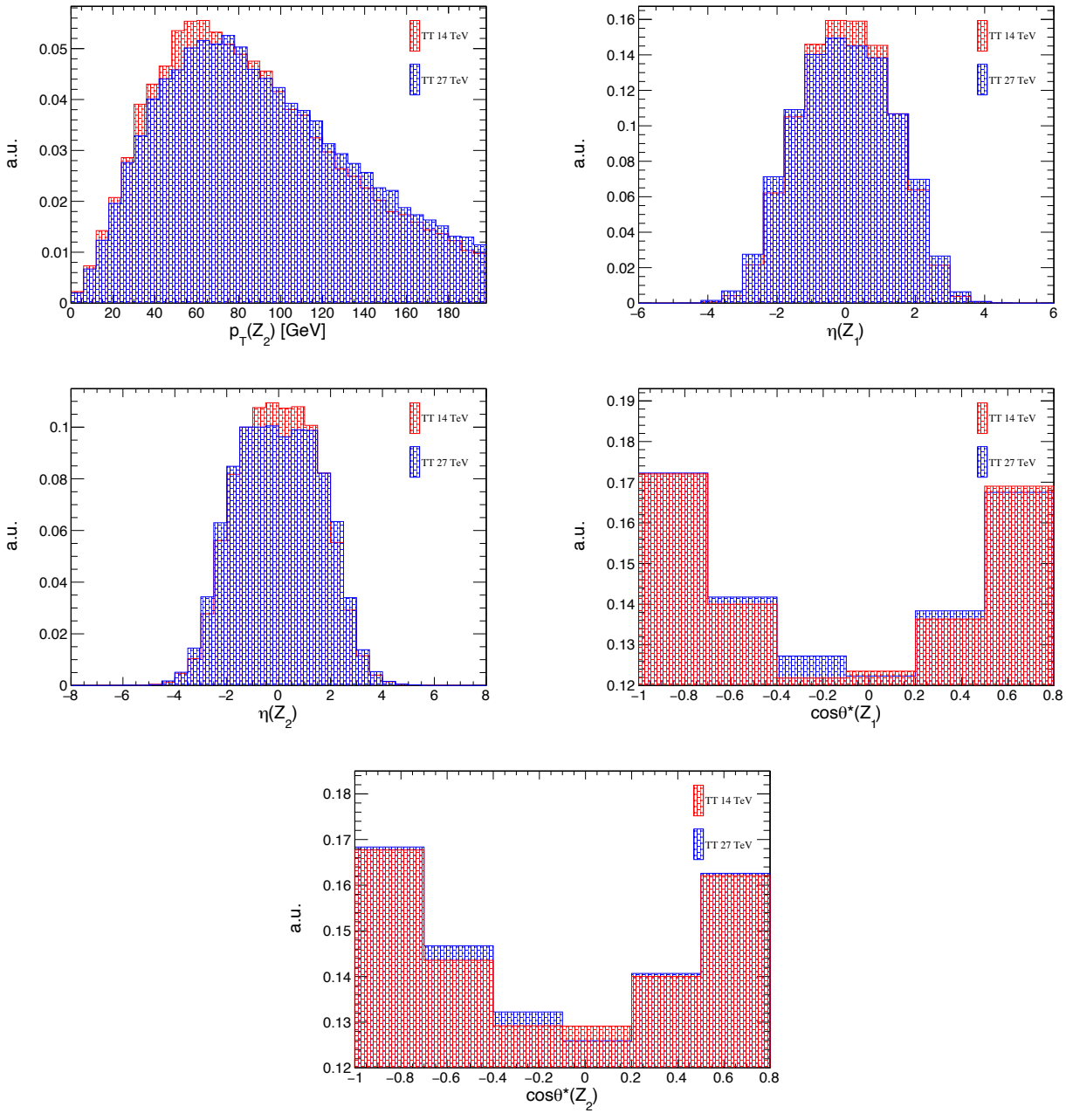
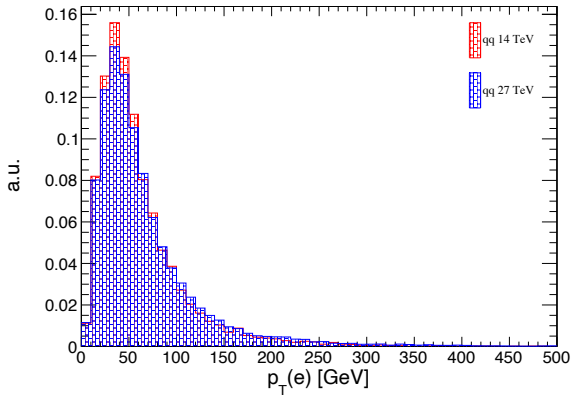
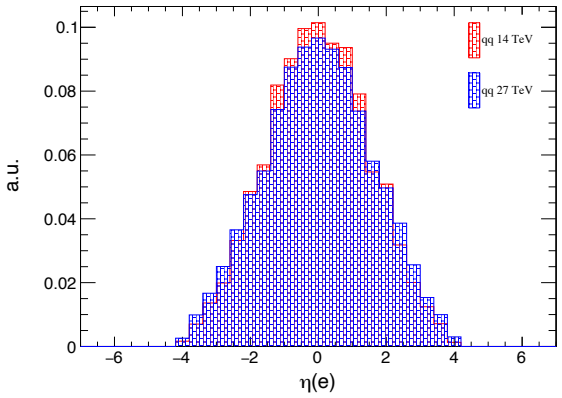
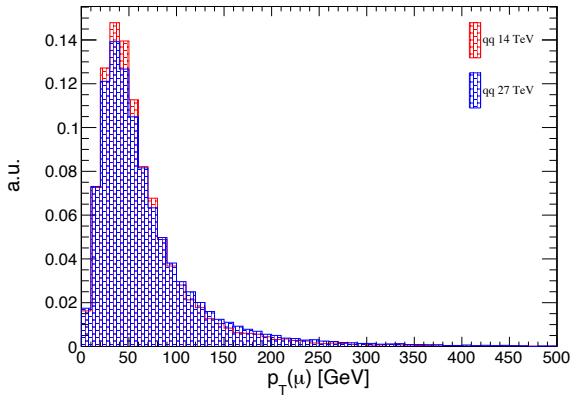
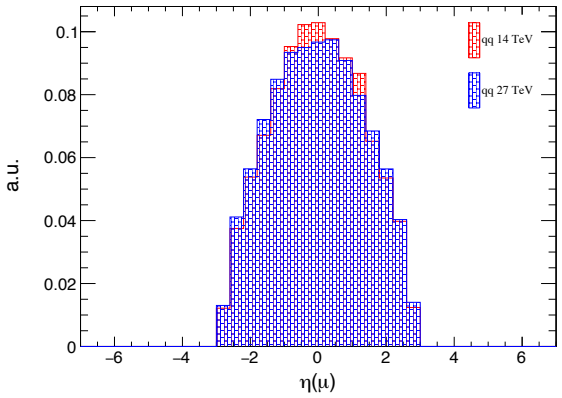


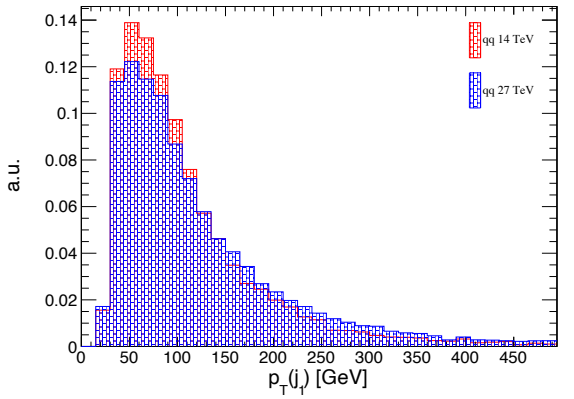
Figure B.5: Kinematics of the TT process at 14 and 27 TeV after the baseline selection.

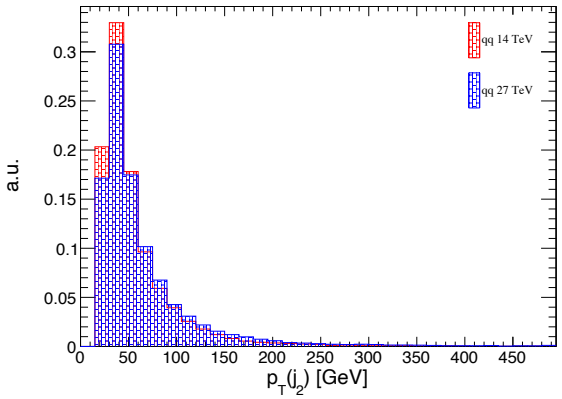

qq 14 TeV
qq 27 TeV
a.u.
p_T(e) [GeV]

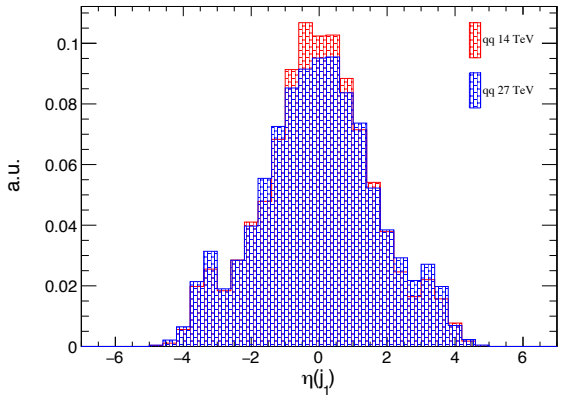

qq 14 TeV
qq 27 TeV
a.u.
η(e)

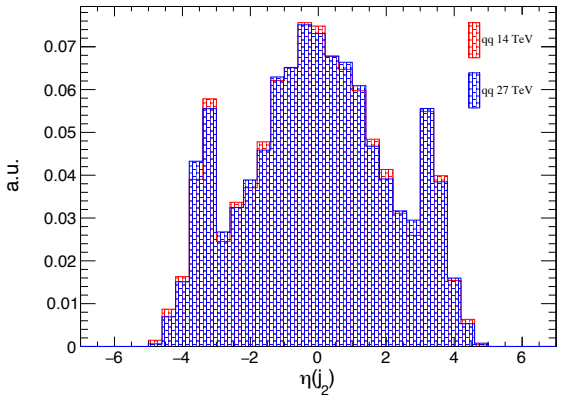

qq 14 TeV
qq 27 TeV
a.u.
p_T(μ) [GeV]

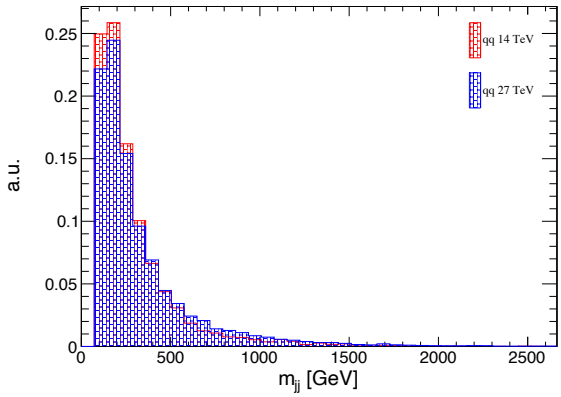

qq 14 TeV
qq 27 TeV
a.u.
η(μ)

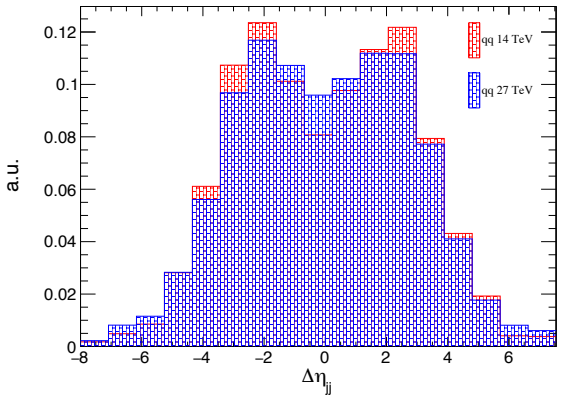

qq 14 TeV
qq 27 TeV
a.u.
$p_T(j_1)$ [GeV]

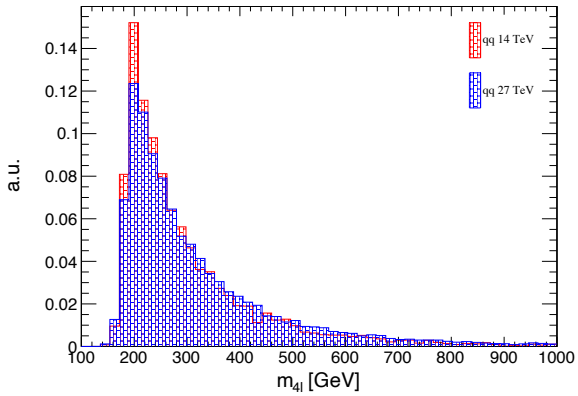

qq 14 TeV
qq 27 TeV
a.u.
$p_T(j_2)$ [GeV]

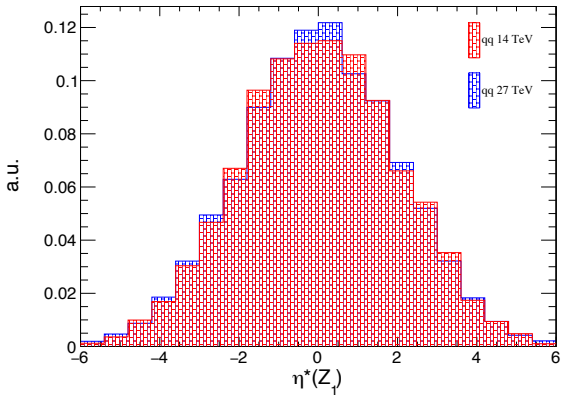

qq 14 TeV
qq 27 TeV
a.u.
η(j_1)

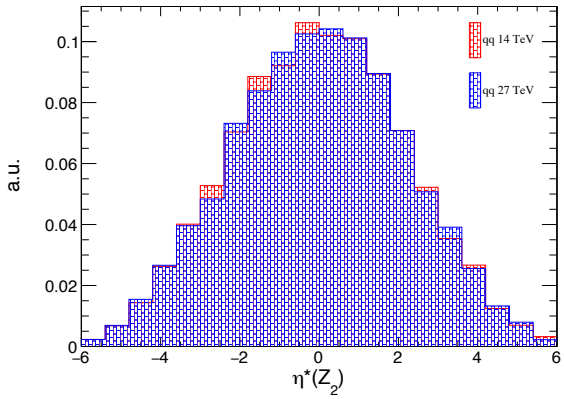

qq 14 TeV
qq 27 TeV
a.u.
η(j_2)

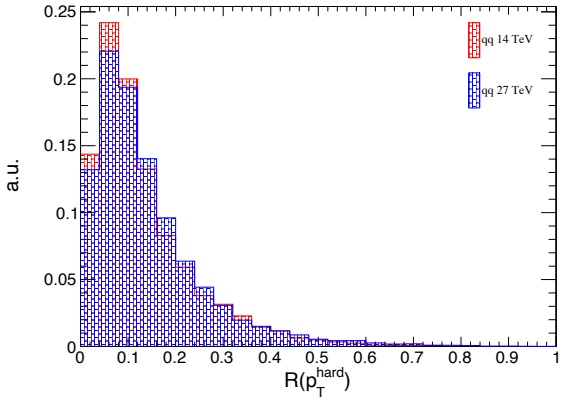

qq 14 TeV
qq 27 TeV
a.u.
m_{jj} [GeV]

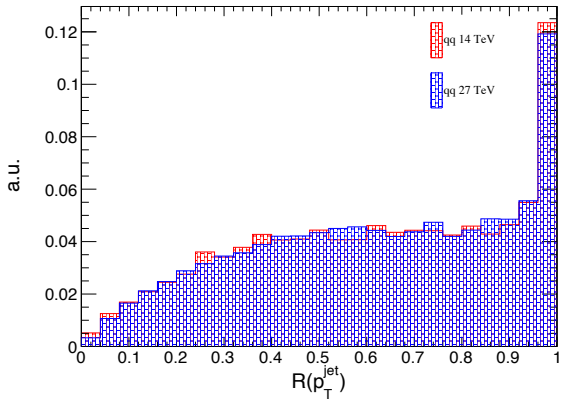

qq 14 TeV
qq 27 TeV
a.u.
$\Delta\eta_{jj}$

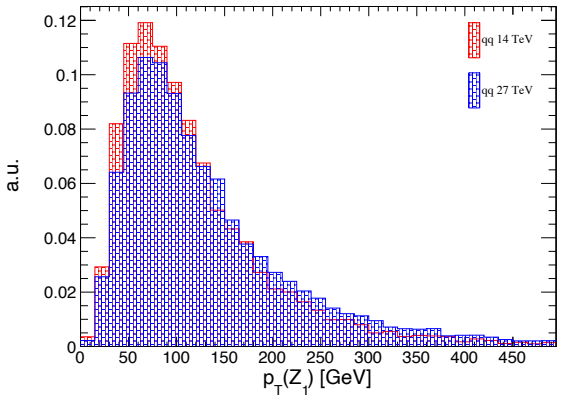

qq 14 TeV
qq 27 TeV
a.u.
m_{4l} [GeV]


qq 14 TeV
qq 27 TeV
a.u.
$\eta^*(Z_1)$


qq 14 TeV
qq 27 TeV
a.u.
$\eta^*(Z_2)$


qq 14 TeV
qq 27 TeV
a.u.
$R(p_T^{hard})$


qq 14 TeV
qq 27 TeV
a.u.
$R(p_T^{jet})$


qq 14 TeV
qq 27 TeV
a.u.
$p_T(Z_1)$ [GeV]

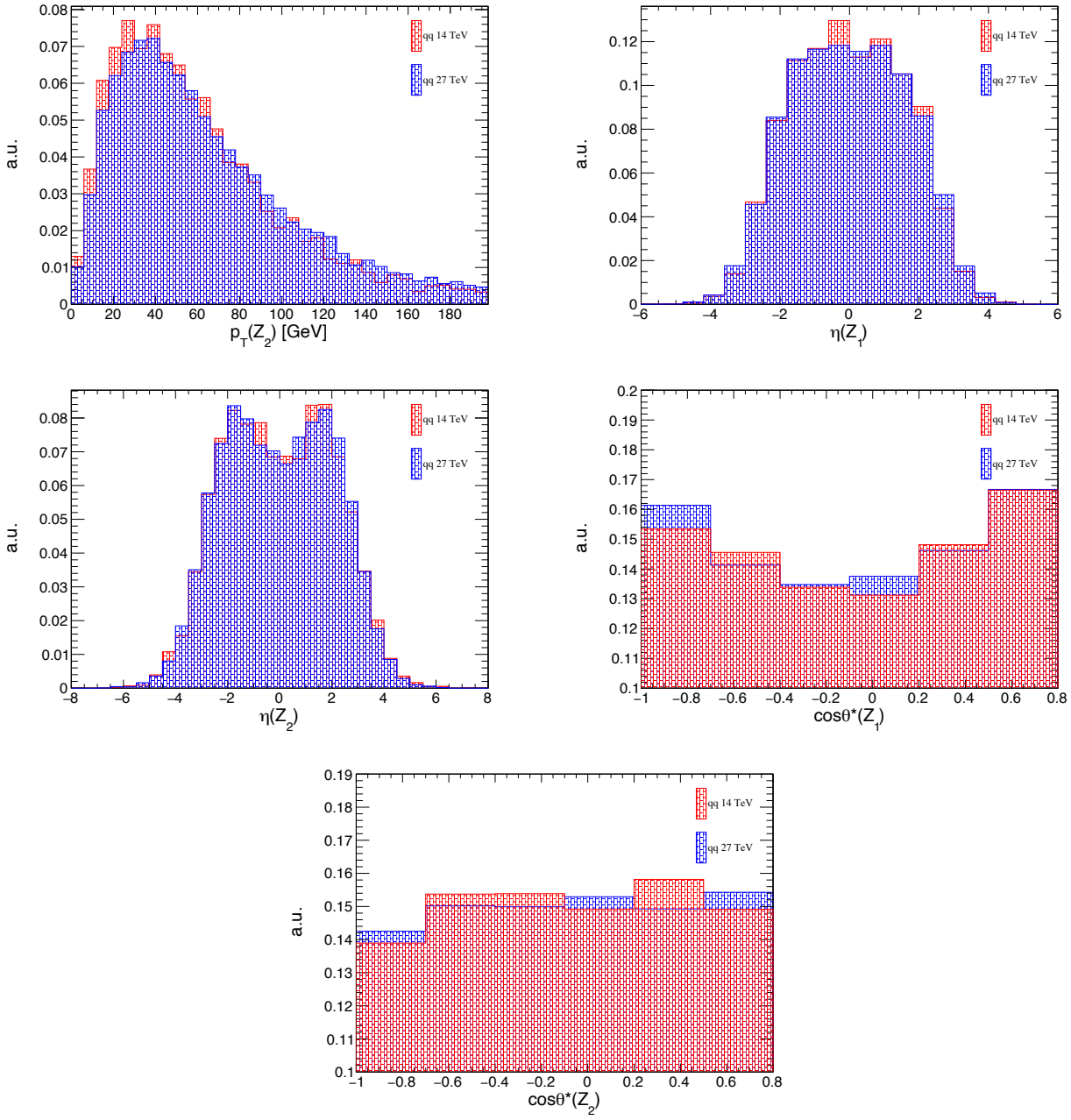
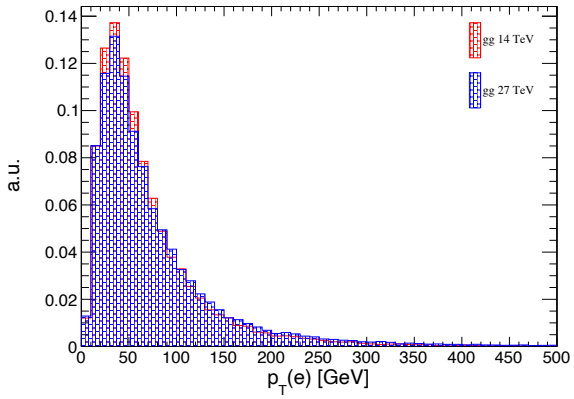
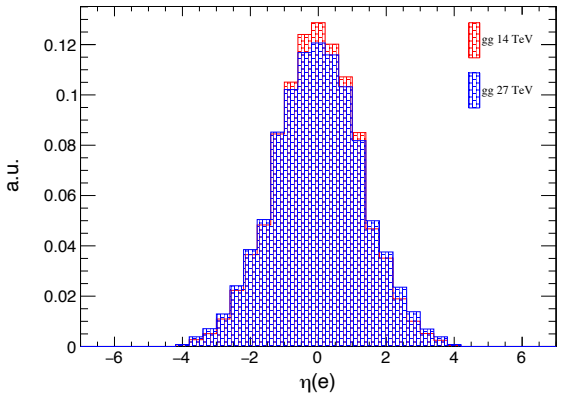
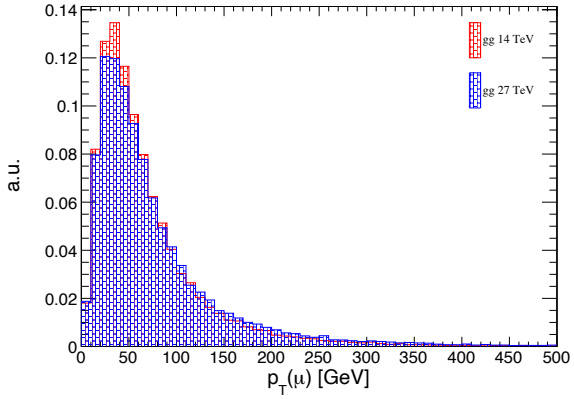
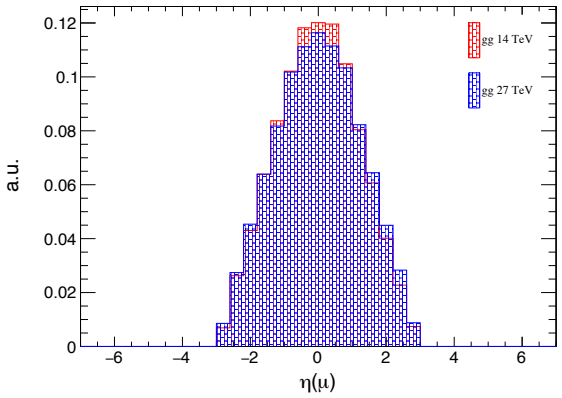


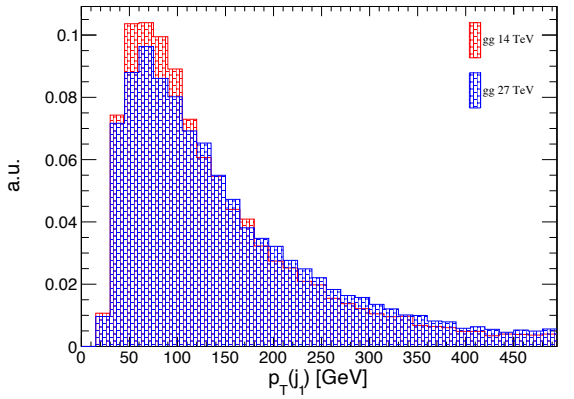
Figure B.6: Kinematics of the qq process at 14 and 27 TeV after the baseline selection.

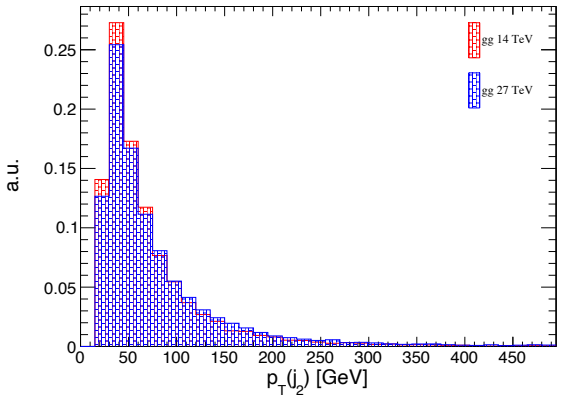

gg 14 TeV
gg 27 TeV
a.u.
$p_T(e)$ [GeV]

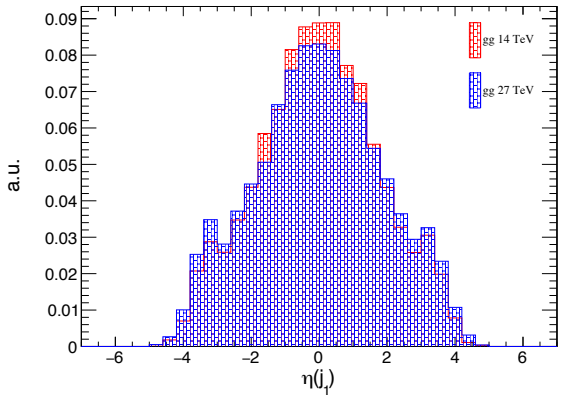

gg 14 TeV
gg 27 TeV
a.u.
η(e)

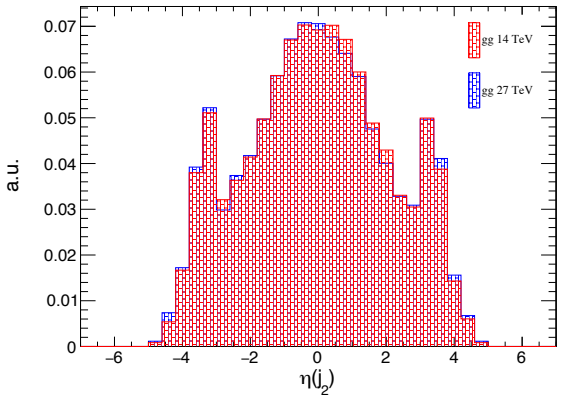

gg 14 TeV
gg 27 TeV
a.u.
$p_T(\mu)$ [GeV]

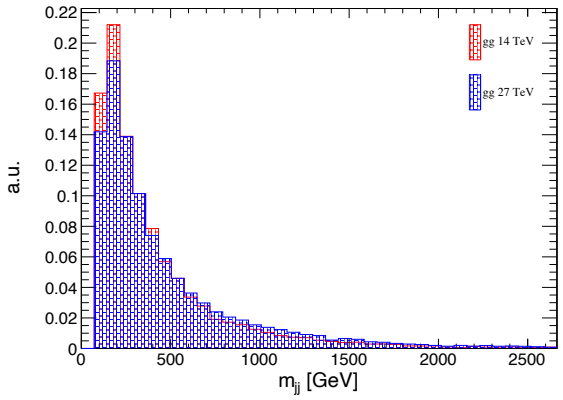

gg 14 TeV
gg 27 TeV
a.u.
η(μ)

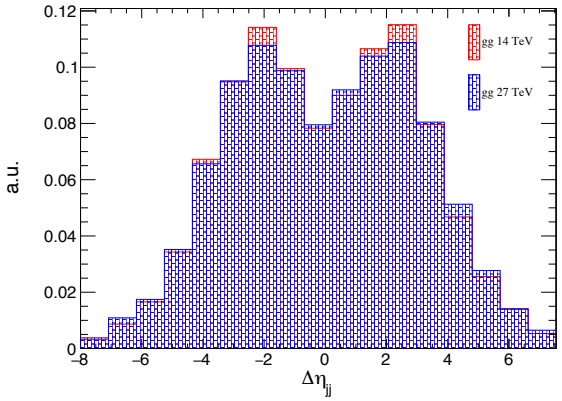

gg 14 TeV
gg 27 TeV
a.u.
$p_T(j_1)$ [GeV]

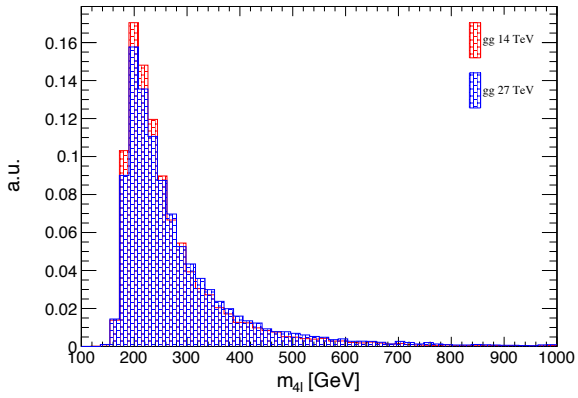

gg 14 TeV
gg 27 TeV
a.u.
$p_T(j_2)$ [GeV]

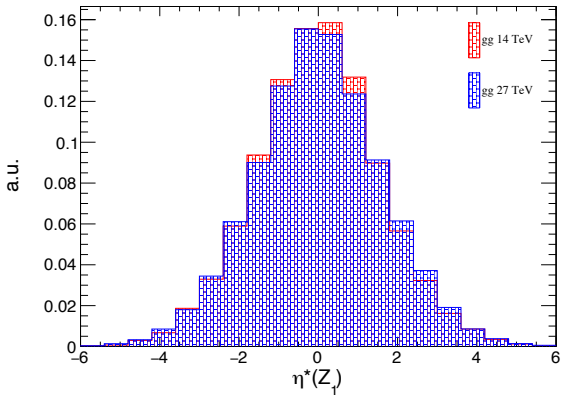

gg 14 TeV
gg 27 TeV
a.u.
η(j1)

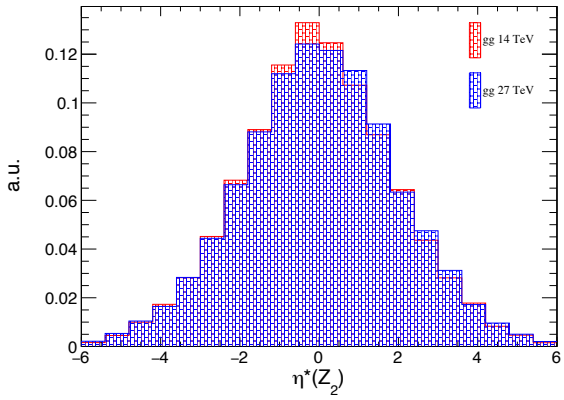

gg 14 TeV
gg 27 TeV
a.u.
η(j2)

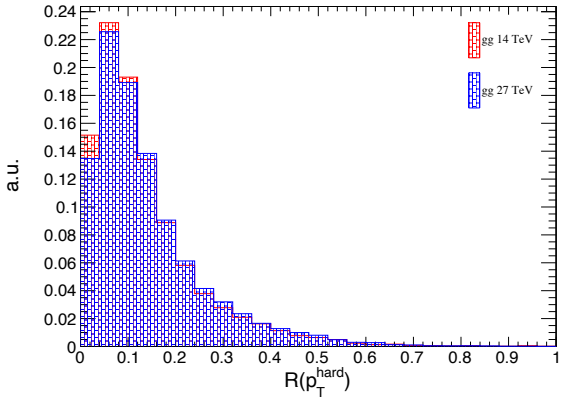

a.u.
gg 14 TeV
gg 27 TeV
m_{jj} [GeV]

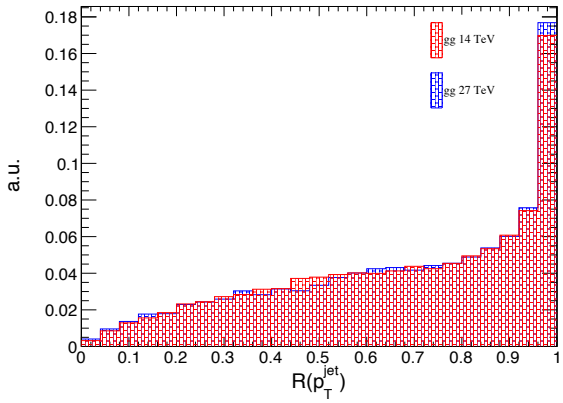

a.u.
gg 14 TeV
gg 27 TeV
$\Delta\eta_{jj}$

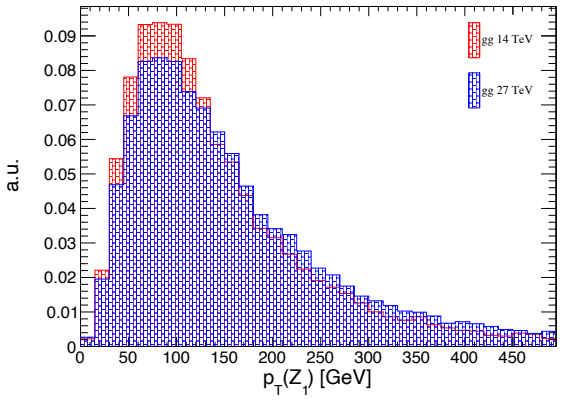

a.u.
gg 14 TeV
gg 27 TeV
m_{4l} [GeV]


a.u.
gg 14 TeV
gg 27 TeV
$\eta^*(Z_1)$


a.u.
gg 14 TeV
gg 27 TeV
$\eta^*(Z_2)$


a.u.
gg 14 TeV
gg 27 TeV
$R(p_T^{hard})$


a.u.
gg 14 TeV
gg 27 TeV
$R(p_T^{jet})$


a.u.
gg 14 TeV
gg 27 TeV
$p_T(Z_1)$ [GeV]

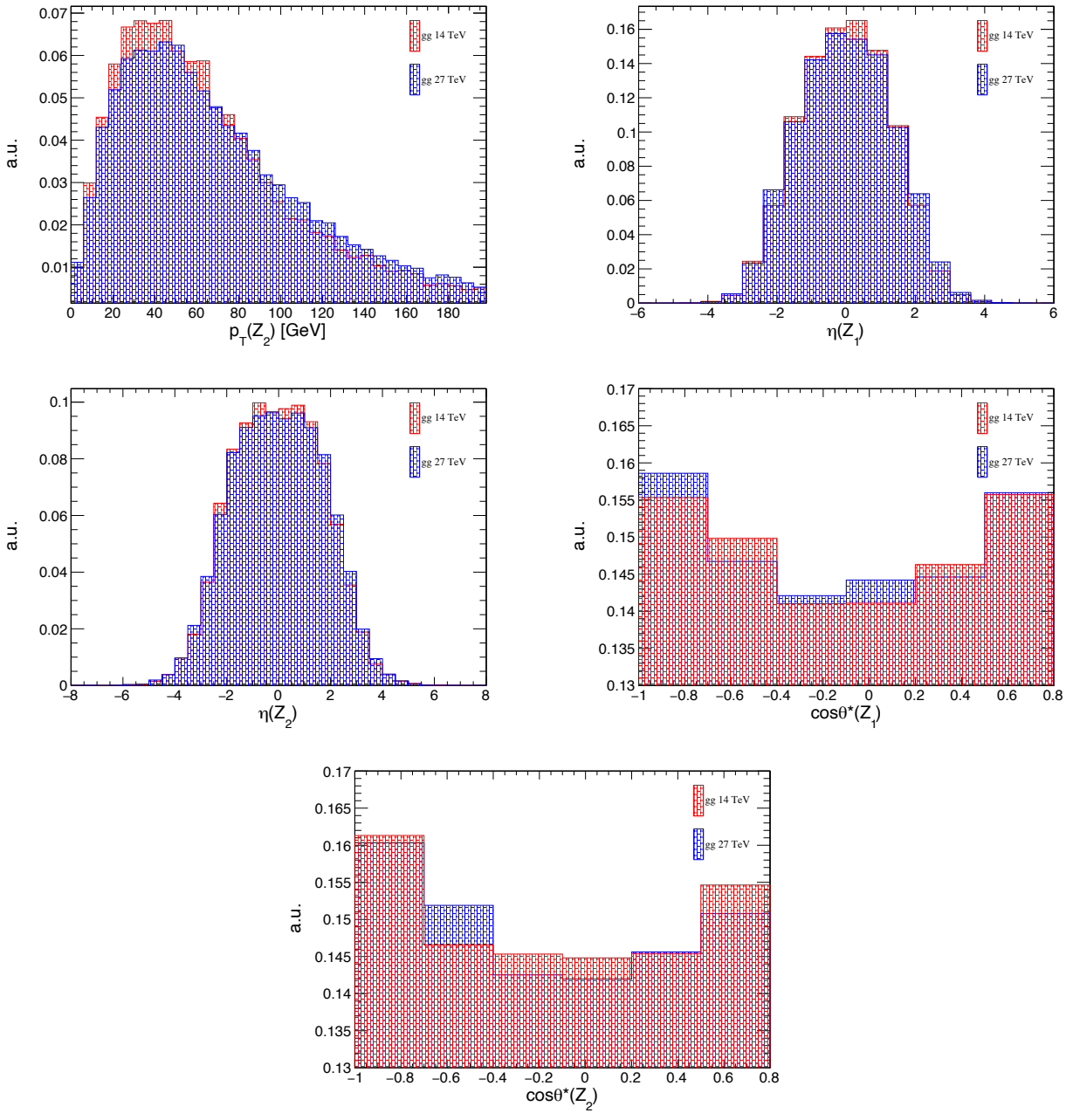


Figure B.7: Kinematics of the gg process at 14 and 27 TeV after the baseline selection.

Bibliography

[1] J. M. Cornwall, D. N. Levin, and G. Tiktopoulos, "Derivation of gauge invariance from high-energy unitarity bounds on the S matrix," *Phys. Rev. D*, vol. 10, pp. 1145–1167, Aug 1974.

[2] B. W. Lee, C. Quigg, and H. B. Thacker, "Weak interactions at very high energies: The role of the Higgs-boson mass," *Phys. Rev. D*, vol. 16, pp. 1519–1531, Sep 1977.

[3] M. S. Chanowitz and M. K. Gaillard, "The tev physics of strongly interacting W's and Z's," *Nuclear Physics B*, vol. 261, pp. 379–431, 1985.

[4] J. Brehmer, "Polarised WW Scattering at the LHC," Master's thesis, U. Heidelberg, ITP, 2014.

[5] B. W. Lee, C. Quigg, and H. B. Thacker, "Strength of Weak Interactions at Very High Energies and the Higgs Boson Mass," *Phys. Rev. Lett.*, vol. 38, pp. 883–885, Apr 1977.

[6] A. Ballestrero, B. Biedermann, *et al.*, "Precise predictions for same-sign W-boson scattering at the LHC," *The European Physical Journal C*, vol. 78, aug 2018.

[7] C. Collaboration and T. Mc Cauley, "Displays of WW or WZ production in association with two jets in the CMS detector." CMS Collection., 2020.

[8] D. Rainwater, R. Szalapski, and D. Zeppenfeld, "Probing color-singlet exchange in Z + 2-jet events at the LHC," vol. 54, pp. 6680–6689, dec 1996.

[9] R. Gomez-Ambrosio, "Study of VBF/VBS in the LHC at 13 TeV, the EFT Approach," 2016.

[10] C. Degrande, N. Greiner, W. Kilian, O. Mattelaer, H. Mebane, T. Stelzer, S. Willenbrock, and C. Zhang, "Effective field theory: A modern approach to anomalous couplings," *Annals of Physics*, vol. 335, pp. 21–32, aug 2013.

[11] G. Perez, M. Sekulla, and D. Zeppenfeld, "Anomalous quartic gauge couplings and unitarization for the vector boson scattering process $pp \to w^+w^+jjx \to l^+\nu_l l^+\nu_l jjx$," *The European Physical Journal C*, vol. 78, sep 2018.

[12] A. Dedes, P. Kozów, and M. Szleper, "Standard model EFT effects in vector-boson scattering at the LHC," *Phys. Rev. D*, vol. 104, p. 013003, Jul 2021.

[13] S. Weinberg, "Baryon- and lepton-nonconserving processes," *Phys. Rev. Lett.*, vol. 43, pp. 1566–1570, Nov 1979.

[14] M. Rauch, "Vector-Boson Fusion and Vector-Boson Scattering," 2016.

[15] O. J. P. É boli, M. C. Gonzalez-Garcia, and J. K. Mizukoshi, "$p\,p \to j\,j\,e+/-\,\mu+/-\,\nu\,\nu$ and $j\,j\,e+/-\,\mu+/-\,\nu\,\nu$ at $\mathcal{O}(\alpha_{em}^6)$ and $\mathcal{O}(\alpha_{em}^4\ \alpha_s^2)$ for the study of the Quartic Electroweak Gauge Boson Vertex at LHC," *Physical Review D*, vol. 74, oct 2006.

[16] M. Rauch, "Vector-Boson Fusion and Vector-Boson Scattering," 2016.

[17] The CMS Collaboration, "Study of Vector Boson Scattering and Search for New Physics in Events with Two Same-Sign Leptons and Two Jets," *Physical Review Letters*, vol. 114, feb 2015.

[18] The ATLAS Collaboration, "Evidence for Electroweak Production of $W^{\pm}W^{pm}jj$ in pp Collisions at $\sqrt{s} = 8\ TeV$ with the ATLAS Detector," *Physical Review Letters*, vol. 113, oct 2014.

[19] The ATLAS Collaboration, "Measurement of $W^{\pm}Z$ production cross sections in pp collisions at $\sqrt{s} = 8\ TeV$ with the ATLAS detector and limits on anomalous gauge boson self-couplings," *Physical Review D*, vol. 93, may 2016.

[20] The ATLAS Collaboration, "Measurement of $W^{\pm}W^{\pm}$ vector-boson scattering and limits on anomalous quartic gauge couplings with the ATLAS detector," *Physical Review D*, vol. 96, jul 2017.

[21] The CMS Colaboration, "Measurement of electroweak-induced production of $W\gamma$ with two jets in pp collisions at $\sqrt{s} = 8\ teV$ and constraints on anomalous quartic gauge couplings," *Journal of High Energy Physics*, vol. 2017, jun 2017.

[22] The CMS Collaboration, "Measurement of the cross section for electroweak production of Z gamma in association with two jets and constraints on anomalous quartic gauge couplings in proton-proton collisions at $\sqrt{s} = 8\ TeV$," *Physics Letters B*, vol. 770, pp. 380–402, jul 2017.

[23] The ATLAS Collaboration, "Studies of $Z\gamma$ production in association with a high-mass dijet system in pp collisions at $\sqrt{s} = 8\ TeV$ with the ATLAS detector," *Journal of High Energy Physics*, vol. 2017, jul 2017.

[24] The CMS Collaboration, "Observation of electroweak production of same-sign W boson pairs in the two jet and two same-sign lepton final state in proton-proton collisions at $\sqrt{s} = 13\ TeV$," *Physical Review Letters*, vol. 120, feb 2018.

[25] The CMS Collaboration, "Measurement of vector boson scattering and constraints on anomalous quartic couplings from events with four leptons and two jets in proton-proton collisions at $\sqrt{s} = 13\ TeV$," *Physics Letters B*, vol. 774, pp. 682–705, nov 2017.

[26] The ATLAS Collaboration, "Observation of electroweak $W^{\pm}Z$ boson pair production in association with two jets in pp collisions at $\sqrt{s} = 13\ TeV$ with the ATLAS detector," *Physics Letters B*, vol. 793, pp. 469–492, jun 2019.

[27] The CMS Collaboration, "Measurement of electroweak WZ boson production and search for new physics in $WZ+$ two jets events in pp collisions at $\sqrt{s} = 13\ TeV$," *Physics Letters B*, vol. 795, pp. 281–307, aug 2019.

[28] The ATLAS Collaboration, "Observation Observation of electroweak production of a same-sign W boson pair in association with two jets in pp collisions at $\sqrt{s} = 13\ TeV$ with the ATLAS detector," *Physical Review Letters*, vol. 123, oct 2019.

[29] The CMS Collaboration, "Search for anomalous electroweak production of vector boson pairs in association with two jets in proton-proton collisions at 13 TeV," *Physics Letters B*, vol. 798, p. 134985, nov 2019.

[30] "Observation of electroweak production of two jets in association with a Z-boson pair in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector," tech. rep., CERN, Geneva, 2019. All figures including auxiliary figures are available at https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/CONFNOTES/ATLAS-CONF-2019-033.

[31] CMS Collaboration, "Measurements of production cross sections of polarized same-sign W boson pairs in association with two jets in proton-proton collisions at $\sqrt{13}\ TeV$," 2020.

[32] The CMS Collaboration, "Measurement of $W^{\pm}\gamma$ differential cross sections in proton-proton collisions at $\sqrt{s} = 13\ TeV$ and effective field theory constraints," *Phys. Rev. D*, vol. 105, p. 052003, Mar 2022.

[33] The CMS Collaboration", "Measurement of the electroweak production of $Z\gamma$ and two jets in proton-proton collisions at $\sqrt{s} = 13$ TeV and constraints on dimension 8 operators," tech. rep., CERN, Geneva, 2021.

[34] "LHC Design Report. 3. The LHC injector chain," 12 2004.

[35] E. Mobs, "The CERN accelerator complex. Complexe des accélérateurs du CERN," 2016. General Photo.

[36] P. Baudrenghien, L. Arnaudon, T. Bohl, O. Brunner, A. Butterworth, P. Maesen, J. E. Muller, G. Ravida, E. Shaposhnikova, and H. Timko, "Status and commissioning plans for LHC Run 2. The RF system.," in *5th Evian workshop on LHC beam operation*, (Geneva), pp. 99–104, CERN, 2014.

[37] L. R. Evans and P. Bryant, "LHC Machine," *JINST*, vol. 3, p. S08001. 164 p, 2008. This report is an abridged version of the LHC Design Report (CERN-2004-003).

[38] L. Evans and P. Bryant, "LHC machine," *Journal of Instrumentation*, vol. 3, pp. S08001–S08001, aug 2008.

[39] O. Aberle, I. Béjar Alonso, *et al.*, *High-Luminosity Large Hadron Collider (HL-LHC): Technical design report*. CERN Yellow Reports: Monographs, Geneva: CERN, 2020.

[40] The CMS Collaboration, *CMS Physics: Technical Design Report Volume 1: Detector Performance and Software*. Technical design report. CMS, Geneva: CERN, 2006. There is an error on cover due to a technical problem for some items.

[41] "Signal/background discrimination for the VBF Higgs four lepton decay channel with the CMS experiment using Machine Learning classification techniques." `https://confluence.infn.it/pages/viewpage.action?pageId=53906361`. [Accessed 14-Oct-2022].

[42] V. Veszpremi, "Operation and performance of the CMS tracker," *Journal of Instrumentation*, vol. 9, pp. C03005–C03005, mar 2014.

[43] The CMS Collaboration, "The CMS experiment at the CERN LHC," *Journal of Instrumentation*, vol. 3, pp. S08004–S08004, aug 2008.

[44] C. Martin Perez, *The CMS Experiment at the LHC*, pp. 41–84. Cham: Springer International Publishing, 2022.

[45] C. Biino, "The CMS Electromagnetic Calorimeter: overview, lessons learned during Run 1 and future projections," *Journal of Physics: Conference Series*, vol. 587, p. 012001, feb 2015.

[46] "Energy calibration and resolution of the CMS electromagnetic calorimeter in pp collisions at $\sqrt{s} = 7\ TeV$," *Journal of Instrumentation*, vol. 8, pp. P09009–P09009, sep 2013.

[47] S. Banerjee, "Performance of Hadron Calorimeter with and without HO," tech. rep., CERN, Geneva, 1999.

[48] S. Mukhopadhyay, "Studies on the upgrade of the muon system in the forward region of the CMS experiment at LHC with GEMs," *Journal of Instrumentation*, vol. 9, 01 2014.

[49] P. Paolucci, "The CMS Muon system," tech. rep., CERN, Geneva, 2005.

[50] The CMS Collaboration, "The CMS trigger system," *Journal of Instrumentation*, vol. 12, pp. P01020–P01020, jan 2017.

[51] P. Govoni, "The CMS High-Level Trigger," in *IFAE 2006* (G. Montagna, O. Nicrosini, and V. Vercesi, eds.), (Milano), pp. 361–364, Springer Milan, 2007.

[52] V. Gori, "The CMS high level trigger," *International Journal of Modern Physics: Conference Series*, vol. 31, p. 1460297, jan 2014.

[53] D. Trocino, "The CMS High Level Trigger," *Journal of Physics: Conference Series*, vol. 513, p. 012036, jun 2014.

[54] H. Sert, "CMS Run 2 High Level Trigger Performance," *PoS*, vol. EPS-HEP2019, p. 165, 2020.

[55] P. Billoir, "Progressive track recognition with a Kalman-like fitting procedure," *Computer Physics Communications*, vol. 57, no. 1, pp. 390–394, 1989.

[56] P. Billoir and S. Qian, "Simultaneous pattern recognition and track fitting by the Kalman filtering method," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 294, no. 1, pp. 219–228, 1990.

[57] R. Mankel, "A concurrent track evolution algorithm for pattern recognition in the HERA-B main tracking system," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 395, no. 2, pp. 169–184, 1997.

[58] R. Frühwirth, "Application of Kalman filtering to track and vertex fitting," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 262, no. 2, pp. 444–450, 1987.

[59] T. C. Collaboration, "Description and performance of track and primary-vertex reconstruction with the CMS tracker," *Journal of Instrumentation*, vol. 9, pp. P10009–P10009, oct 2014.

[60] The CMS Collaboration, "Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at $\sqrt{s} = 13$ TeV," *Journal of Instrumentation*, vol. 13, pp. P06015–P06015, jun 2018.

[61] "Determination of the Jet Energy Resolutions and Jet Reconstruction Efficiency at CMS," tech. rep., CERN, Geneva, 2009.

[62] R. Atkin, "Review of jet reconstruction algorithms," *Journal of Physics: Conference Series*, vol. 645, p. 012008, oct 2015.

[63] "Jet algorithms performance in 13 TeV data," tech. rep., CERN, Geneva, 2017.

[64] M. Cacciari, G. P. Salam, and G. Soyez, "The anti-k_T jet clustering algorithm," *Journal of High Energy Physics*, vol. 2008, pp. 063–063, apr 2008.

[65] Y. Dokshitzer, G. Leder, S. Moretti, and B. Webber, "Better jet clustering algorithms," *Journal of High Energy Physics*, vol. 1997, pp. 001–001, aug 1997.

[66] M. Wobisch and T. Wengler, "Hadronization Corrections to Jet Cross Sections in Deep-Inelastic Scattering," 1999.

[67] M. Cacciari, G. P. Salam, and G. Soyez, "FastJet user manual," *The European Physical Journal C*, vol. 72, mar 2012.

[68] The CMS Collaboration, "Particle-flow reconstruction and global event description with the CMS detector," *Journal of Instrumentation*, vol. 12, pp. P10003–P10003, oct 2017.

[69] G. Apollinari, I. Béjar Alonso, *et al.*, "High-Luminosity Large Hadron Collider (HL-LHC). Technical Design Report V. 0.1,"

[70] B. Schmidt, "The High-Luminosity upgrade of the LHC: Physics and Technology Challenges for the Accelerator and the Experiments," *Journal of Physics: Conference Series*, vol. 706, p. 022002, 04 2016.

[71] A. L. Rosa, "The Upgrade of the CMS Tracker at HL-LHC," Journal of the Physical Society of Japan, jun 2021.

[72] A. Martelli, "The CMS HGCAL detector for HL-LHC upgrade," in *5th Large Hadron Collider Physics Conference*, 8 2017.

[73] C. Battilana, "Upgrades of the CMS muon detectors: from Run 3 towards HL-LHC. Upgrades of the CMS muon detectors: from Run-3 towards HL-LHC," tech. rep., CERN, Geneva, 2020.

[74] T. R. F. P. Tomei, "The CMS Trigger Upgrade for the HL-LHC," *EPJ Web of Conferences*, vol. 245, p. 01031, 2020.

[75] A. Martelli, "The CMS HGCAL detector for HL-LHC upgrade," 2017.

[76] C. Ochando, "HGCAL: A High-Granularity Calorimeter for the endcaps of CMS at HL-LHC," *J. Phys.: Conf. Ser.*, vol. 928, p. 012025. 4 p, 2017.

[77] A. Martelli, "The CMS HGCAL detector for HL-LHC upgrade," in *5th Large Hadron Collider Physics Conference*, 8 2017.

[78] "A High Granularity Nose for HF in HL-LHC?." `https://indico.cern.ch/event/735810/contributions/3045066/attachments/1673664/2686042/HG_HF_Nose_Jun18.pdf`. Accessed: 2022-10-22.

[79] F. Zimmermann, M. Benedikt, *et al.*, "HE-LHC: The High-Energy Large Hadron Collider: Future Circular Collider Conceptual Design Report Volume 4. Future Circular Collider," tech. rep., CERN, Geneva, 2019.

[80] Q. Ingram, "Energy resolution of the barrel of the CMS electromagnetic calorimeter," *Journal of Instrumentation*, vol. 2, pp. P04004–P04004, apr 2007.

[81] The CMS Collaboration, "CMS physics: Technical design report volume 1: Detector performance," *CMS Technical Design Report CERN-LHCC-2006-001*, 2006.

[82] The CMS Collaboration, "Electron and photon reconstruction and identification with the CMS experiment at the CERN LHC," *Journal of Instrumentation*, vol. 16, p. P05014, may 2021.

[83] D. Valsecchi, "Deep learning techniques for energy clustering in the CMS ECAL," 2022.

[84] "Performance of electron reconstruction and selection with the CMS detector in proton-proton collisions at $\sqrt{s} = 8$ TeV," vol. 10, pp. P06005–P06005, jun 2015.

[85] W. Adam, R. Frühwirth, A. Strandlie, and T. Todorov, "Reconstruction of electrons with the gaussian-sum filter in the CMS tracker at the LHC," *Journal of Physics G: Nuclear and Particle Physics*, vol. 31, pp. N9–N20, jul 2005.

[86] H. A. Bethe and L. C. Maximon, "Theory of Bremsstrahlung and Pair Production. I. Differential Cross Section," *Phys. Rev.*, vol. 93, pp. 768–784, Feb 1954.

[87] S. Baffioni, C. Charlot, *et al.*, "Electron reconstruction in CMS," *CMS-NOTE-2006-040*, Feb 2006.

[88] M. Oreglia, "A Study of the Reactions $\psi' \to \gamma\gamma\psi$," *SLAC Report SLAC-R-236*, 1980.

[89] The CMS Collaboration, "Measurement of the properties of a Higgs boson in the four-lepton final state," *Physical Review D*, vol. 89, may 2014.

[90] The CMS Collaboration, "Measurements of properties of the Higgs boson decaying into the four-lepton final state in pp collisions at $\sqrt{s} = 13$ TeV," *Journal of High Energy Physics*, vol. 2017, nov 2017.

[91] "Measurements of properties of the Higgs boson in the four-lepton final state in proton-proton collisions at $\sqrt{s} = 13$ TeV," tech. rep., CERN, Geneva, 2019.

[92] M. Cacciari and G. P. Salam, "Pileup subtraction using jet areas," *Physics Letters B*, vol. 659, pp. 119–126, jan 2008.

[93] M. Cacciari, G. P. Salam, and G. Soyez, "The catchment area of jets," *Journal of High Energy Physics*, vol. 2008, pp. 005–005, apr 2008.

[94] M. Cacciari, G. P. Salam, and G. Soyez, "FastJet user manual," *The European Physical Journal C*, vol. 72, mar 2012.

[95] Donato, Silvio, "CMS trigger performance," *EPJ Web Conf.*, vol. 182, p. 02037, 2018.

[96] The CMS Collaboration, "Performance of the CMS level-1 trigger in proton-proton collisions at $\sqrt{s} = 13$ TeV," *Journal of Instrumentation*, vol. 15, pp. P10017–P10017, oct 2020.

[97] A. Hocker *et al.*, "TMVA - Toolkit for Multivariate Data Analysis," 3 2007.

[98] "XGBoost Documentation." `https://xgboost.readthedocs.io/en/latest/index.html`.

[99] "Electron and Photon performance in CMS with the full 2017 data sample and additional 2016 highlights for the CALOR 2018 Conference," May 2018.

[100] J. Alwall, R. Frederix, *et al.*, "The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations," *Journal of High Energy Physics*, vol. 2014, jul 2014.

[101] The CMS Collaboration, "Measurements of production cross sections of the Higgs boson in the four-lepton final state in proton-proton collisions at $\sqrt{s} = 13$ TeV," *Eur. Phys. J. C 81 (2021) 488*, 2019.

[102] P. Nason, "A New Method for Combining NLO QCD with Shower Monte Carlo Algorithms," *Journal of High Energy Physics*, vol. 2004, pp. 040–040, nov 2004.

[103] S. Frixione, P. Nason, and C. Oleari, "Matching NLO QCD computations with parton shower simulations: the POWHEG method," *Journal of High Energy Physics*, vol. 2007, pp. 070–070, nov 2007.

[104] S. Alioli, P. Nason, C. Oleari, and E. Re, "A general framework for implementing NLO calculations in shower monte carlo programs: the POWHEG BOX," *Journal of High Energy Physics*, vol. 2010, jun 2010.

[105] P. Pigard, *Study of the EWK double Z production in the four leptons final state withthe CMS experiment at the LHC*. Theses, Université Paris-Saclay, July 2017.

[106] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H.-S. Shao, T. Stelzer, P. Torrielli, and M. Zaro, "The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations," *Journal of High Energy Physics*, vol. 2014, jul 2014.

[107] A. Ballestrero, A. Belhouari, G. Bevilacqua, V. Kashkan, and E. Maina, "PHANTOM: A Monte Carlo event generator for six parton final states at high energy colliders," *Comput. Phys. Commun.*, vol. 180, pp. 401–417, 2009.

[108] R. Frederix and I. Tsinikos, "On improving NLO merging for $t\bar{t}w$ production," *Journal of High Energy Physics*, vol. 2021, nov 2021.

[109] R. Frederix and S. Frixione, "Merging meets matching in MC@NLO," *Journal of High Energy Physics*, vol. 2012, dec 2012.

[110] The CMS Collaboration, "Evidence for electroweak production of four charged leptons and two jets in proton-proton collisions at s=13tev," *Physics Letters B*, vol. 812, p. 135992, 2021.

[111] J. Alwall, R. Frederix, *et al.*, "The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations," vol. 2014, no. 7, p. 79.

[112] R. Covarelli, Y. An, R. Bellan, M. Bonanomi, C. Charlot, E. Fontanesi, D. Giljanovic, H. He, D. Lelas, C. Li, Q. Li, G. Ortona, A. Savin, and T. Sculac, "Search for vector-boson scattering in the 4ljj final state with full Run2 data," *CMS AN 2019/172*.

[113] J. Alwall, S. Höche, *et al.*, "Comparative study of various algorithms for the merging of parton showers and matrix elements in hadronic collisions," *The European Physical Journal C*, vol. 53, pp. 473–500, dec 2007.

[114] V. Hirschi and P. Pigard, "Discussions on gluon-loop induced zz+2jets." `https://answers.launchpad.net/mg5amcnlo/+question/402723`, Oct. 2016.

[115] J. M. Campbell and R. Ellis, "MCFM for the tevatron and the LHC," *Nuclear Physics B - Proceedings Supplements*, vol. 205-206, pp. 10–15, aug 2010.

[116] The CMS Collaboration, "Event generator tunes obtained from underlying event and multiparton scattering measurements," *The European Physical Journal C*, vol. 76, mar 2016.

[117] The CMS collaboration, "Extraction and validation of a new set of CMS pythia8 tunes from underlying-event measurements," *The European Physical Journal C*, vol. 80, jan 2020.

[118] R. D. Ball, V. Bertone, *et al.*, "Parton distributions for the LHC run II," *Journal of High Energy Physics*, vol. 2015, apr 2015.

[119] M. Grazzini, S. Kallweit, and D. Rathlev, "ZZ production at the LHC: Fiducial cross sections and distributions in NNLO QCD," *Physics Letters B*, vol. 750, pp. 407–410, nov 2015.

[120] S. Gieseke, T. Kasprzik, and J. H. Kühn, "Vector-boson pair production and electroweak corrections in herwig++," 2014.

[121] F. Caola, K. Melnikov, *et al.*, "QCD corrections to zz production in gluon fusion at the LHC," *Physical Review D*, vol. 92, nov 2015.

[122] F. Caola, M. Dowling, *et al.*, "QCD corrections to vector boson pair production in gluon fusion including interference effects with off-shell higgs at the LHC," *Journal of High Energy Physics*, vol. 2016, jul 2016.

[123] G. Petrucciani, A. Rizzi, and C. Vuosalo, "Mini-AOD: A new analysis data format for CMS," *Journal of Physics: Conference Series*, vol. 664, p. 072052, dec 2015.

[124] The CMS Collaboration, A. M. Sirunyan, and A. Tumasyan, "Measurements of properties of the Higgs boson decaying into the four-lepton final state in pp collisions at $\sqrt{s}$ =13 TeV," vol. 2017, no. 11, p. 47.

[125] A. Sirunyan, A. Tumasyan, *et al.*, "Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at $\sqrt{s}$ =13 TeV," vol. 13, no. 06, pp. P06015–P06015.

[126] M. Cacciari, G. P. Salam, and G. Soyez, "FastJet user manual," vol. 72, no. 3, p. 1896.

[127] The CMS Collaboration, "Identification of b-quark jets with the CMS experiment," *Journal of Instrumentation*, vol. 8, pp. P04013–P04013, apr 2013.

[128] D. Rainwater, R. Szalapski, and D. Zeppenfeld, "Probing color-singlet exchange in $Z+2$-jet events at the CERN LHC," vol. 54, no. 11, pp. 6680–6689.

[129] Y. Gao, A. V. Gritsan, Z. Guo, K. Melnikov, M. Schulze, and N. V. Tran, "Spin determination of single-produced resonances at hadron colliders," *Physical Review D*, vol. 81, apr 2010.

[130] S. Bolognesi, Y. Gao, A. V. Gritsan, K. Melnikov, M. Schulze, N. V. Tran, and A. Whitbeck, "Spin and parity of a single-produced resonance at the LHC," *Physical Review D*, vol. 86, nov 2012.

[131] I. Anderson, S. Bolognesi, F. Caola, Y. Gao, A. V. Gritsan, C. B. Martin, K. Melnikov, M. Schulze, N. V. Tran, A. Whitbeck, and Y. Zhou, "Constraining anomalous HVV interactions at proton and lepton colliders," *Physical Review D*, vol. 89, feb 2014.

[132] R. Brun and F. Rademakers, "ROOT: An object oriented data analysis framework," *Nucl. Instrum. Meth. A*, vol. 389, pp. 81–86, 1997.

[133] P. Speckmayer, A. Höcker, J. Stelzer, and H. Voss, "The toolkit for multivariate data analysis, TMVA 4," *Journal of Physics: Conference Series*, vol. 219, p. 032057, apr 2010.

[134] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, "Asymptotic formulae for likelihood-based tests of new physics," *The European Physical Journal C*, vol. 71, feb 2011.

[135] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, "Asymptotic formulae for likelihood-based tests of new physics," *The European Physical Journal C*, vol. 71, feb 2011.

[136] K. Arnold, M. Bähr, *et al.*, "Vbfnlo: A parton level monte carlo for processes with electroweak bosons," *Computer Physics Communications*, vol. 180, pp. 1661–1670, sep 2009.

[137] E. da Silva Almeida, O. J. P. Eboli, *et al.*, "Unitarity constraints on anomalous quartic couplings," *Physical Review D*, vol. 101, jun 2020.

[138] T. Plehn, "Lhc phenomenology for physics hunters," 2008.

[139] The CMS Collaboration, "CMS luminosity measurement for the 2016/2017/2018 data-taking period at $\sqrt{s} = 13$ TeV,"

[140] C. Ochando, T. Sculac, M. Xiao, *et al.*, "Measurements of properties of the higgs boson in the four-lepton final state at $\sqrt{s} = 13$ TeV with full Run II data," *CMS AN 2019/139*.

[141] B. Jäger, A. Karlberg, and G. Zanderighi, "Electroweak ZZjj production in the standard model and beyond in the POWHEG-BOX v2," *Journal of High Energy Physics*, vol. 2014, mar 2014.

[142] S. Alioli, P. Nason, C. Oleari, and E. Re, "A general framework for implementing NLO calculations in shower monte carlo programs: the POWHEG BOX," *Journal of High Energy Physics*, vol. 2010, jun 2010.

[143] A. Denner, R. Franken, M. Pellen, and T. Schmidt, "NLO QCD and EW corrections to vector-boson scattering into ZZ at the LHC," *Journal of High Energy Physics*, vol. 2020, nov 2020.

[144] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, "Asymptotic formulae for likelihood-based tests of new physics," *The European Physical Journal C*, vol. 71, feb 2011.

[145] The CMS Collaboration, "Measurement of the electroweak production of $z\gamma$ and two jets in proton-proton collisions at $\sqrt{s} = 13$ TeV and constraints on anomalous quartic gauge couplings," *Physical Review D*, vol. 104, oct 2021.

[146] C. Charlot, D. Lelas, D. Giljanovic, and A. Savin, "Vector boson scattering prospective studies for the High-Luminosity LHC upgrade in the ZZ fully leptonic decay channel," *CMS AN 2018/072*.

[147] J. Alwall, A. Ballestrero, *et al.*, "A standard format for Les Houches Event Files.".

[148] T. Sjöstrand, S. Ask, *et al.*, "An Introduction to PYTHIA 8.2," vol. 191, pp. 159–177.

[149] S. Ovyn, X. Rouby, and V. Lemaitre, "Delphes, a Framework for Fast Simulation of a Generic Collider Experiment.".

[150] J. de Faverau, C. Delaere, *et al.*, "DELPHES 3, A Modular Framework for Fast Simulation of a Generic Collider Experiment," vol. 02, p. 057.

[151] J. Pumplin, D. R. Stump, *et al.*, "New Generation of Parton Distributions with Uncertainties from Global QCD Analysis," vol. 2002, no. 07, pp. 012–012.

[152] J. Mousa, A. Tumasyan, *et al.*, "Pileup mitigation at CMS in 13 TeV data," vol. 15, pp. P09018–P09018.

[153] T. Sjostrand, "A Model for Initial State Parton Showers," vol. 157, pp. 321–325.

[154] S. De, V. Rentala, and W. Shepherd, "Measuring the polarization of boosted, hadronic W bosons with jet substructure observables," 2020.

ECOLE
DOCTORALE

Titre: Recherche de la diffusion de boson de jauge dans les événements avec le détecteur CMS auprès du LHC

Mots clés: CMS, LHC, VBS, ZZ

Résumé: L'étude de la diffusion des bosons de jauges est un test crucial de la brisure de la symmétrie électrofaible et est un moyen complémentaire pour la mesure des couplages du boson de Higgs aux bosons vecteurs. Dans cette étude j'analyse $137 fb^{-1}$ de collisions proton-proton produites au grand collisionneur de protons (LHC) du CERN à une énergie de 13 TeV dans le centre de masse. Je présente également une étude prospective de la diffusion longitudinale dans le même canal pour la phase de haute luminosité (HL-LHC) et une phase de haute énergie (HE-LHC), correspondant respectivement à des énergies de 14 TeV et 27 TeV dans le centre de masse, avec une simulation complète de la cinématique des événements. Finalement, je présente mon travail sur l'amélioration de la mesure de l'efficacité de sélection des électrons pour les périodes de prise de données de 2016, 2017 et 2018.
La production électrofaible (EW) de 2 jets en association avec 2 bosons Z est mesurée avec une significance observée (attendue) de 4.0 (3.5) déviations standards. Les sections efficaces pour la production EW sont mesurées dans 3 régions fiducielles, et est $0.33^{(+0.11)}_{(-0.10)}(stat)^{(+0.04)}_{(-0.03)}(syst)\ fb$ dans la région la plus inclusive, en accord avec la prédiction du modèle standard (SM) de $0.275 \pm 0.021\ fb$. Des limites sur l'existence de couplages quartiques anormaux sont établies en terme des opérateurs de la théorie effective T0, T1, T2 , T8 et T9.
Une mesure de la diffusion longitudinale dans la canal ZZ est attendue à 27 TeV pour une luminosité intégrée de $15000 fb^{-1}$, avec une significance de 4.6 déviations standards. En étendant l'acceptance des électrons de $|\eta| = 3$ à $|\eta| = 4$, une première observation est attendue avec une significance de 5.4 déviations standards. Cette étude montre le bénéfice important d'une augmentation de l'énergie du LHC pour la compréhension du secteur EW du SM.

Title: Study of vector boson scattering in events with four leptons and two jets with the CMS detector at the LHC

Keywords: CMS, LHC, VBS, ZZ

Abstract: Studying Vector Boson Scattering is crucial for understanding the electroweak (EW) symmetry breaking mechanism and it provides a complementary tool for measuring Higgs boson couplings to vector bosons. In this study I analyse $137 fb^{-1}$ of proton-proton collisions produced at CERN Large Hadron Collider (LHC) at 13 TeV centre-of-mass energy. Additionally, I presented a prospective study on the longitudinal scattering in the same channel at High-Luminosity and High-Energy LHC conditions, corresponding to 14 and 27 TeV centre-of-mass energy, respectively, with full event kinematics simulated. Finally, I present my work on improving the measurement of electron selection efficiency for the 2016, 2017 and 2018 data-taking periods.
The EW production of two jets in association with two Z bosons was measured with an observed (expected) significance of 4.0 (3.5) standard deviations. The cross sections for the EW production were measured in three fiducial volumes and is $0.33^{(+0.11)}_{(-0.10)}(stat)^{(+0.04)}_{(-0.03)}(syst)\ fb$ in the most inclusive volume, in agreement with the Standard Model (SM) prediction of $0.275 \pm 0.021\ fb$. Limits on the anomalous quartic gauge couplings were derived in terms of EFT operators T0, T1, T2, T8, and T9.
A measurement of the longitudinal scattering in the ZZ channel is expected at 27 TeV, corresponding to an integrated luminosity of $15000 fb^{-1}$, with a signal significance of 4.6 standard deviations. By extending the electron acceptance from $|\eta| = 3$ to $|\eta| = 4$, the first observation is expected with a significance of 5.4 standard deviations. Hence, this study demonstrates a significant benefit of further energy increase at the LHC for understanding the EW sector of the SM.

Institut Polytechnique de Paris
91120 Palaiseau, France