University of Zagreb
Faculty of Science
Department of Physics

Saswat Mishra

Measurement of the Cross section for associated production of the Higgs boson and an electroweak boson in final states with two b quarks and two leptons in proton proton collisions at the Large Hadron Collider

DOCTORAL THESIS

Zagreb, 2023

University of Zagreb
Faculty of Science
Department of Physics

Saswat Mishra

Measurement of the Cross section for associated production of the Higgs boson and an electroweak boson in final states with two b quarks and two leptons in proton proton collisions at the Large Hadron Collider

DOCTORAL THESIS

Supervisor:
Dr. sc. Dinko Ferenček

Zagreb, 2023

Sveučilište u Zagrebu
Prirodoslovno-matematički fakultet
Fizički odsjek

Saswat Mishra

Mjerenje udarnog presjeka za zajedničku tvorbu Higgsovog bozona i elektroslabog bozona u konačnim stanjima s dva b kvarka i dva leptona u proton proton sudarima na Velikom hadronskom sudarivaču

DOKTORSKI RAD

Mentor:
Dr. sc. Dinko Ferenček

Zagreb, 2023.

Supervisor information

Dinko FERENČEK

Education

2006–2011	**Ph.D.**, **Physics**, University of Maryland, USA
2001–2006	**M.Sc.**, **Physics**, University of Zagreb, Croatia

Work Experience

2021–Present	**Senior Research Associate**, Ruđer Bošković Institute, Zagreb, Croatia
2015–2021	**Research Associate**, Ruđer Bošković Institute, Zagreb, Croatia
2011–2015	**Postdoctoral Research Associate**, Rutgers, The State University of New Jersey, USA
2007–2011	**Research Assistant**, University of Maryland, USA
2006–2007	**Teaching Assistant**, University of Maryland, USA

Selected Publications

1. Ferenček D., Roguljić M., Starodumov A., "Production, calibration, and performance of the layer 1 replacement modules for the CMS pixel detector", *Proceedings of The 29th International Workshop on Vertex Detectors (VERTEX2020)*, JPS Conf. Proc. **34** (2021) 010023

2. Roguljić M., Starodumov A., Karadzhinova-Ferrer A., Ferenček D., Ahmed A. A., Jara-Casas L. M., "Low dose rate ^{60}Co facility in Zagreb", *Proceedings of The 28th International Workshop on Vertex Detectors (VERTEX2019)*, *PoS* Vertex2019 (2020) 066

3. Majer M., Roguljić M., Knežević, Ž., Starodumov A., Ferenček D., Brigljević V., Mihaljević B., "Dose mapping of the panoramic ^{60}Co gamma irradiation facility at the Ruđer Bošković Institute – Geant4 simulation and measurements", *Appl. Radiat. Isot.* **154** (2019) 108824

4. CMS Collaboration, "Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV", *JINST* **13** (2018) P05011, arXiv:1712.07158

5. CMS Collaboration, "Search for dijet resonances in proton-proton collisions at $\sqrt{s} =$ 13 TeV and constraints on dark matter and other models", *Phys. Lett. B* **769** (2017) 520, arXiv:1611.03568

6. CMS Collaboration, "Search for narrow resonances in dijet final states at $\sqrt{s} = 8$ TeV with the novel CMS technique of data scouting", *Phys. Rev. Lett.* **117** (2016) 031802, arXiv:1604.08907

7. CMS Collaboration, "Search for heavy resonances decaying to two Higgs bosons in final states containing four b quarks", *Eur. Phys. J. C* **76** (2016) 371, arXiv:1602.08762

8. CMS Collaboration, "Search for narrow resonances decaying to dijets in proton-proton collisions at $\sqrt{s} = 13$ TeV", *Phys. Rev. Lett.* **116** (2016) 071801, arXiv:1512.01224

9. CMS Collaboration, "Search for pair-produced vector-like B quarks in pp collisions at $\sqrt{s} = 8$ TeV", *Phys. Rev. D* **93** (2016) 112009, arXiv:1507.07129

10. CMS Collaboration, "Search for vector-like T quarks decaying to top quarks and Higgs bosons in the all-hadronic channel using jet substructure", *JHEP* **06** (2015) 080, arXiv:1503.01952

11. CMS Collaboration, "Search for resonances and quantum black holes using dijet mass spectra in proton-proton collisions at $\sqrt{s} = 8$ TeV", *Phys. Rev. D* **91** (2015) 052009, arXiv:1501.04198

12. CMS Collaboration, "Search for pair-produced resonances decaying to jet pairs in proton-proton collisions at $\sqrt{s} = 8$ TeV", *Phys. Lett. B* **747** (2015) 98, arXiv:1412.7706

13. CMS Collaboration, "Search for narrow resonances and quantum black holes in inclusive and b-tagged dijet mass spectra from pp collisions at $\sqrt{s} = 7$ TeV", *JHEP* **01** (2013) 013, arXiv:1210.2387

14. CMS Collaboration, "Search for First Generation Scalar Leptoquarks in the $evjj$ Channel in pp Collisions at $\sqrt{s} = 7$ TeV", *Phys. Lett. B* **703** (2011) 246, arXiv:1105.5237

15. CMS Collaboration, "Missing transverse energy performance of the CMS detector", *JINST* **6** (2011) 09001, arXiv:1106.5048

16. CMS Collaboration, "Search for Pair Production of First-Generation Scalar Leptoquarks in pp Collisions at $\sqrt{s} = 7$ TeV", *Phys. Rev. Lett.* **106** (2011) 201802, arXiv:1012.4031

17. V. Brigljević et. al., "Study of di-boson production with the CMS detector at LHC", *J. Phys. G* **34** (2007) N269-N295

Acknowledgements

This work would not have been possible without the support and funding of the Croatian Science Foundation under the project IP-2016-06-3321 and DOK-2018-01-9182. I would like to express my heartfelt gratitude for the sincere cooperation and help from the members of the CMS group in Rudjer Boskovic Institute led by Dr. Vuko Brigljevic. I would personally like to acknowledge the efforts of my supervisor, Dr. Dinko Ferenček whose steady assistance helped me with the efforts in the analysis. My humble thanks to Dr. Vuko Brigljevic for being the nice teacher he is. I will always cherish those long evening physics discussions we used to have. Your insights to some of my confusions has helped me grow as a person and as a physicist.
I would like to thank the members of $VH, H \to b\bar{b}$ analysis working group in CMS and further all members of the CMS collaboration whose collective efforts towards the collaboration made this analysis a success.
I would like to extend my gratitude towards my parents and my family members for their constant belief and emotional support during hard times in my PhD. Thanks to Om Prakash and Anirudh for providing so many good memories worth cherishing during my time in Zagreb. Thanks to Naveen for being a nice roommate and thanks to Bhakti and Matej for being great colleagues by providing much needed breathers from time to time during work at IRB. Also thanks to Priyanka and Hrishikesh for being such great friends.
I would like to acknowledge my friends in India who are also pursuing PhD in physics. Chatting about hardships during PhD and sharing common issues always gave me sense of delight and helped me cope up with difficult times. Thank you Sameer, Ayesha and Sarthak. Also thanks to all my childhood friends and my three cousins (Vivek, Vibhanshu and Anurag) who, even though couldn't provide as much help, but helped me relax from time to time by making me forget my worries.
Last but not the least, thanks to my fiancee Sonali for her constant support, care and love while still being in India. I look forward to reciprocate in similar fashion !!

Abstract

This thesis summarizes the analysis in which we measure the cross-section for the production of the standard model Higss boson of 125 GeV in association with an electroweak boson (W or Z). The analysis is performed in the final state where the Higgs boson decays into a pair of b quarks and the electroweak boson decays lectonically resulting in three channels based on the number of charged leptons in the final state (0, 1 or 2 leptons). The analysis uses data recorded by the CMS experiment from proton-proton collisions at $\sqrt{s}$= 13 TeV in the LHC during the full Run 2 data taking period (2016-2018). The recorded data corresponds to an integrated luminosity of 138 fb^{-1}. The analysis searches for 2 b-jets produced from b-quarks originating from the Higgs boson along with lepton candidates decaying from the vector boson. The mass of the Higgs boson is reconstructed from the four momenta of the b-jets which are identified using b-tagging algorithms.

To account for events with a Lorentz boosted Higgs boson, a single large cone jet is reconstructed which consists of the two b-jets merged together due to the boost. Taking this into account, separate analysis called the boosted analysis is performed along with the nominal resolved analysis to improve sensitivity in high p_T phase space. Both resolved and boosted analysis are combined together in the final fit to enhance precision of the measurement.

To reduce the dependency on theoretical uncertainties this measurement is performed in the simplified template cross-section (STXS) scheme. This also allows for straightforward comparison of theoretical models using such measurements. Under this scheme, the cross section measurement is done in regions delineated by type of vector boson (W or Z), vector boson transverse momentum (p_T), and the presence of additional jets.

Keywords: LHC, CMS, standard model, Higgs boson, boosted objects, DeepCSV, DeepAK8, STXS

Prošireni sažetak

Uvod

Područje proučavanja fizike elementarnih čestica temeljni su gradivni elementi materije i njihove interakcije. Standardni model (SM) fizike čestica teorija je koja je konstruirana kako bi opisala ponašanje elementarnih čestica i utvrdila njihova svojstva razumijevanjem njihovih interakcija putem tri temeljne sile, jakom i slabom nuklearnom silom i elektromagnetskom silom. U ovom trenutku gravitacijska sila nije obuhvaćena SM-om i to je jedno od njegovih ograničenja. Međutim, budući da je gravitacijska sila iznimno slaba na skali elementarnih čestica, SM može opisati interakcije čestica pomoću preostale tri temeljne sile s velikom preciznošću.
SM kakav poznajemo grupira elemntarne čestice u dvije skupine, fermione i bozone. Među fermione spadaju sve čestice koje čine svu poznatu materiju u svemiru. Postoji dvanaest fermiona u trenutnom standardnom modelu koji se dalje dijele na kvarkove i leptone. S druge strane, među bozone spadaju četiri čestice koje predstavljaju tri temeljne sile opisane SM-om i Higgsov bozon. Svi otkriveni bozoni nositelji sile su čestice spina 1 i stoga se nazivaju vektorski bozoni. Higgsov bozon bio je najnoviji dodatak postojećem SM-u i bio je to prvi skalarni bozon (spina 0) koji je otkriven. Pretpostavljalo se da Higgsov bozon postoji gotovo 50 godina prije nego što su ga 2012. godine zajedno otkrili eksperimenti ATLAS i CMS.
10 godina nakon otkrića, fizičari čestica na Velikom hadronskom sudarivaču (LHC) dodatno su ispitali svojstva Higgsovog bozona i do sada je otkriven u pet konačnih stanja, ZZ(4l), WW, $\tau\tau$, $\gamma\gamma$ i $b\bar{b}$. Nedavno mjerenje također je pokazalo dokaze da se Higgsov bozon raspada u par miona, što je bilo moguće samo zbog izvanredne količine proton-proton sudara koji su se odvijali u LHC-u. Također je potvrđeno da je vezanje Higgsovog bozona na bilo koju fundamentalnu česticu izravno proporcionalno masi čestice.
Preciznije mjerenje mase Higgsovog bozona utvrdilo je vrijednost od 125,18 $\pm$ 0,16 pomoću eksperimenta CMS s razinom preciznosti od 0,12%. Koristeći izmjerenu masu m_H kao ulazni parametar u teorijski model, može se vidjeti da Higgsov bozon koji se raspada u par b-kvarkova

ima najveći udio grananja od 58% među svim kanalima raspada Higgsovog bozona. Proces $H \to b\bar{b}$ otkriven je u eksperimentu CMS s opaženom (očekivanom) signifikantnošću od 5,6σ (5,5σ) i u eksperimentu ATLAS s opaženom (očekivanom) signifikantnošću od 5,4 σ (5,5σ) 2018. godine.
U $H \to b\bar{b}$ procesu, 78% od svih Higgsovih bozona proizvedeno je putem mehanizma gluonske fuzije ($gg \to H$). Međutim, precizna mjerenja u ovom proizvodnom kanalu nisu bila moguća zbog ogromne količine pozadine s više hadronskih mlazova u konačnom stanju. Način proizvodnje ZH i WH, s druge strane, precizniji je pri mjerenju procesa $H \to b\bar{b}$. To je zbog mogućnosti iskorištavanja prisutnosti leptona u konačnom stanju koji su produkti raspada Z ili W bozona. Stoga su 2018. godine za otkriće raspada $H \to b\bar{b}$ ZH i WH kanali produkcije bili ključni kanali za postizanje preciznosti potrebne za otkriće.
U ovoj analizi mjerimo udarni presjek za zajedničku tvorbu Higgsovog bozona i masivnoga elektroslabog bozona (W ili Z) gdje se Higgsov bozon raspada u par b-kvarka dok se elektroslabi bozon raspada leptonski. Udarni presjek mjeri se u shemi simplificiranih predložaka udarnog presjeka (STXS) koja ima za cilj mjerenja fiducijalnog udarnog presjeka u potrazi za novom fizikom te se također smanjenjuje ovisnost o teorijskim nesigurnostima. Prema ovoj shemi, mjerenje udarnog presjeka provodi se u područjima određenim tipom vektorskog bozona (W ili Z), transverzalnim momentom vektorskog bozona (p_T) i prisutnošću dodatnih hadronskih mlazova.

Eksperimentalni postav

Analiza je provedena na podacima prikupljenim eksperimentom CMS u proton-proton sudara koji se odvijaju u LHC-u. LHC je najsnažniji sudarač trenutno u funkciji i dizajniran je za sudaranje protona i teških iona. Većina sudara u LHC-u događa se između protona, dok drugi teški ioni poput olova također doprinose malom udjelu ukupnih sudara koji se događaju u LHC-u. Sudari se odvijaju na četiri točke duž LHC tunela gdje su postavljena četiri velika detektora za bilježenje sudara. Eksperimenti ATLAS i CMS dva su detektora opće namjene koji su smješteni u dvije od ove četiri točke sudara. Osim toga, eksperiment ALICE nalazi se u jednoj od točaka sudara za proučavanje sudara teških iona, a na četvrtoj točki sudara smješten je eksperiment LHCb posebno dizajniran za proučavanje b-fizike. Trenutno LHC radi na rekordnoj energiji sudara $\sqrt{s}$ od 13,6 TeV za proton-proton sudare počevši od 2022. godine kada je započeo treći ciklus prikupljanja podataka (Run 3). Ova analiza provedena je na podacima prikupljenim eksperimentom CMS tijekom drugog ciklusa prikupljanja podataka (Run 2) od 2016. do 2018. godine na $\sqrt{s}$ od 13 TeV što odgovara 138 fb^{-1} integriranog luminoziteta.

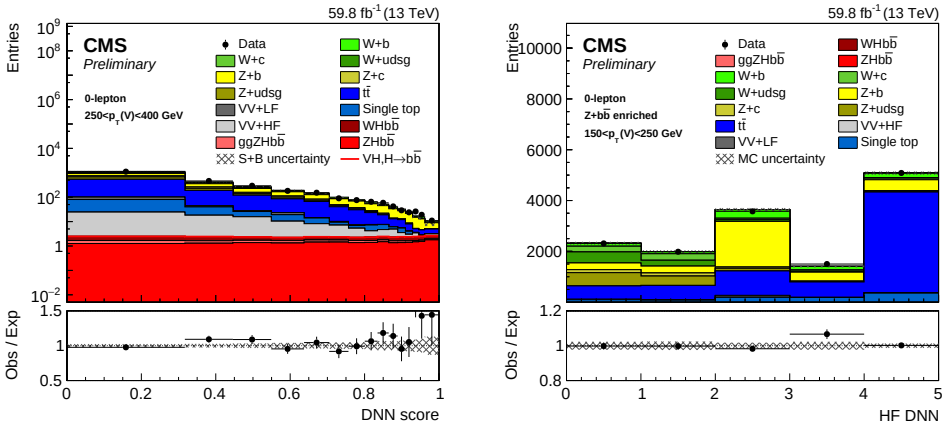
Detektor CMS, što je skraćenica od Compact Muon Solenoid, cilindrični je detektor s mnogo poddetektora raspoređenih u slojeve kako bi se detektirali različiti tipovi čestica u različitim dijelovima detektora. U neposrednoj blizini točke sudara nalazi se sustav za detekciju tragova koji je odgovoran za mjerenje tragova nabijenih čestica i koristan je za rekonstrukciju kratkoživućih čestica, npr. Higgsov bozon, iskorištavanjem podataka o putanjama produkata raspada. Sustav za detekciju tragova okružen je s dva kalorimetra, elektromagnetskim kalorimetrom (ECAL) i hadronskim kalorimetrom (HCAL), koji mjere ukupnu energiju upadnih čestica uz pomoć detekcije pljuska čestica koji upadna čestica kreira unutar ovih kalorimetara. Sustav za detekciju tragova i kalorimetri zatvoreni su unutar supravodljivog solenoida koji osigurava homogeno magnetsko polje od 3,8 T. Prisutnost jakog magnetskog polja omogućuje precizno određivanje impulsa nabijenih čestica i njihovog naboja na osnovu njihove putanje u magnetskom polju. Izvan solenoida postavljen je niz mionskih poddetektora za detekciju miona s čistim signalom jer samo mioni putuju velike udaljenosti bez deponiranja puno energije u bilo kojem drugom podsustavu u unutarnjem dijelu detektora.
Sve su čestice rekonstruirane na osnovu svojih karakterističnih potpisa u detektoru. Jedna iznimka su naravno neutrini koji ne ostavljaju nikakav trag u detektoru za kasniju rekonstrukciju. Stoga se energija neutrina u događaju rekonstruira izračunavanjem nedostajuće transverzalne energije (MET) događaja koristeći zakon očuvanja impulsa. Za generiranje simuliranih događaja koristi se alat GEANT4 (GEometry And Tracking) za simulaciju geometrije detektora i njegovog odgovora na prolaz različitih čestica. Procedura rekonstrukcije prikupljenih i simuliranih podataka identična je kako bi se moglo osigurati dobro slaganja između podataka i simulacije.

Postupak analize

U analizi se traže dva kandidata za hadronski mlaz iz b kvarka koji odgovaraju Higgsovom bozonu i 0, 1 ili 2 leptona koji odgovaraju produktima raspada Z ili W bozona. Ova je analiza stoga podijeljena u tri kanala na temelju broja nabijenih leptona među produktima raspada Z ili W bozona kako bi se uzelo u obzir različite količine pozadine u različitim kanalima. Razni pozadinski procesi koji imaju sličan potpis konačnog stanja kao VHbb proces također su uzeti u obzir i modelirani. Neki od glavnih pozadinskih procesa koji doprinose su događaji V+jets, $t\bar{t}$ i diboson.
B-mlazovi u ovoj analizi rekonstruirani su korištenjem algoritma za grupiranje čestica u mlazove. Ova analiza koristi Anti-kT algoritam s polumjerom stošca od 0,4 za grupiranje u mla-

zove. Kako bi se uzeo u obzir Higgsov bozon s visokim Lorentzovim boostom, provodi se zasebna analiza koja traži jedan široki mlaz većeg polumjera stošca od 0,8 koji se sastoji od dva b-mlaza umjesto dva odvojena b-mlaza radijusa stošca od 0,4. Rekonstruirani mlazovi u topologijama s razlučenim b-mlazovima i širokim mlazom nazivaju se AK4 odnosno AK8 mlazovi radi praktičnosti. Rekonstruirani mlazovi zatim prolaze kroz algoritam b-označavanja koji razlikuje b-mlazove od mlazova nastalih iz lakših kvarkova i gluona. B-mlazovi identificirani u topologiji s razlučenim b-mlazovima označavaju se pomoću algoritma DeepCSV, dok su spojeni b-mlazovi u topologiji s širokim mlazom označeni pomoću algoritma DeepAK8. I DeepCSV i DeepAK8 su algoritmi temeljeni na dubokim neuronskim mrežama (DNN) koji iskorištavaju razlike u kinematičkim informacijama b-mlazova i mlazova nastalih iz lakših kvarkova i gluona. DeepCSV oslanja se na informacije o sekundarnom verteksu i udarnim parametrima tragova, dok je DeepAK8 dizajniran za identifikaciju širokih mlazova koji potječu od dva b-hadrona koristeći informacije o podstrukturi mlaza.
Događaji su odabrani uzimajući u obzir potpis za VHbb proces u tri leptonska kanala. U odabiru događaja s VHbb topologijom koriste se kinematička svojstva b-mlazova i izoliranih leptona zajedno s MET-om. U odabiru se također primjenjuje ocjena algoritma za b-označavanja na kandidatima za b-mlazove kako bi se odabrali kandidati za Higgsov bozon veće čistoće. Kriteriji odabira su isti u topologijama s razlučenim b-mlazovima i širokim mlazom, osim kriterija za odabir kandidata za b-mlazove koji se razlikuju između dviju topologija. Na kraju, obje se topologije statistički kombiniraju u konačnoj prilagodbi kako bi se povećala osjetljivost analize. Kako bi se ograničile glavne pozadine, kontrolna područja su dizajnirana primjenom sličnih kriterija odabira kao područje signala, ali ortogonalno u faznom prostoru u jednoj ili više varijabli. $t\bar{t}$ kontrolno područje izgrađeno je traženjem dodatnih mlazova u događaju, kontrolno područje za proces V+laki kvarkovi izgrađeno je iz regije neprolazne ocjene algoritma za b-označavanje b-mlazova dok je kontrolno područje za process V+teški kvarkovi izgrađena uzimajući u obzir bočni pojas oko prozora mase Higgsovog bozona.

Kako bi se dobilo precizno mjerenje, dobro odvajanje signala od pozadine osigurano je upotrebom različitih multi-varijantnih tehnika u ovoj analizi. Za odabir signala koristi se duboka neuronska mreža (DNN) za klasifikaciju signala i pozadine. DNN koji se koristi u ovoj analizi treniran je na simuliranim uzorcima i na temelju ocjene svaki događaj dobiva klasifikaciju koja označava je li DNN sličan signalu (DNN ocjena bliže 1) ili pozadini (DNN ocjena bliže 0) (Slika 1 lijevo). Za odabir signala u topologiji sa širokim mlazom, pojačana stabla odlučivanja (BDT) koriste se za klasifikaciju signala i pozadine. U kontrolnom području V+teški kvarkovi za 0 i 1-leptonske kanale, klasifikator s više klasa osposobljen je za odvajanje različitih komponenti pozadine: V+jets (zajednička proizvodnja vektorskog bozona i lakih, c ili b kvarkova),



Slika 1: Predlošci duboke neuralne mreže u području signala (lijevo) i području V+teški kvarkovi (desno) za kanal s 0 leptona pokazuju izvrsno slaganje podataka sa simulacijom.

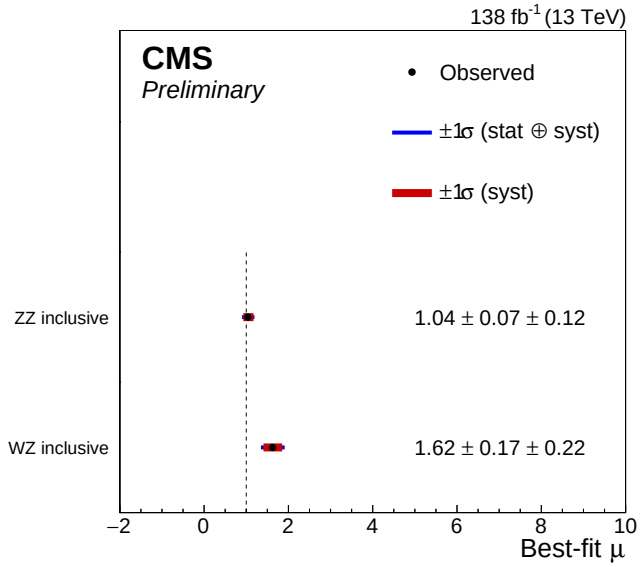
pojedinačni top kvark i $t\bar{t}$ (Slika 1 desno). Ulazne značajke koje se koriste u treniranju DNN-a obuhvaćaju kinematička svojstva konačnog stanja: mlazova, kandidata za vektorske bozone, leptona, mase, momenti i kutevi sustava dva mlaza. Također se koristi multiplicitet rekonstruiranih mlazova. Ove varijable odabrane su korištenjem iterativnog postupka optimizacije, počevši od velikog broja potencijalno diskriminirajućih varijabli. Također se provjerava modeliranje ovih varijabli u podacima, prije nego što se izvrši prilagodba podacima.

Rezultati

Budući da se udarni presjeci mjere u shemi STXS, konačni skup događaja u kanalu i u svakoj regiji (signal ili kontrola) dalje se dijele prema shemi predloška STXS. Prilagodba maksimalne vjerodostojnosti izvodi se istovremeno u signalnim i kontrolnim područjima za svako STXS područje kako bi se dobio modifikator jačine signala (μ) koji označava omjer promatranog broja VHbb događaja u odnosu na onaj koji se očekuje u SM-u. μ=1 predstavlja udarni presjek u skladu s SM-om.

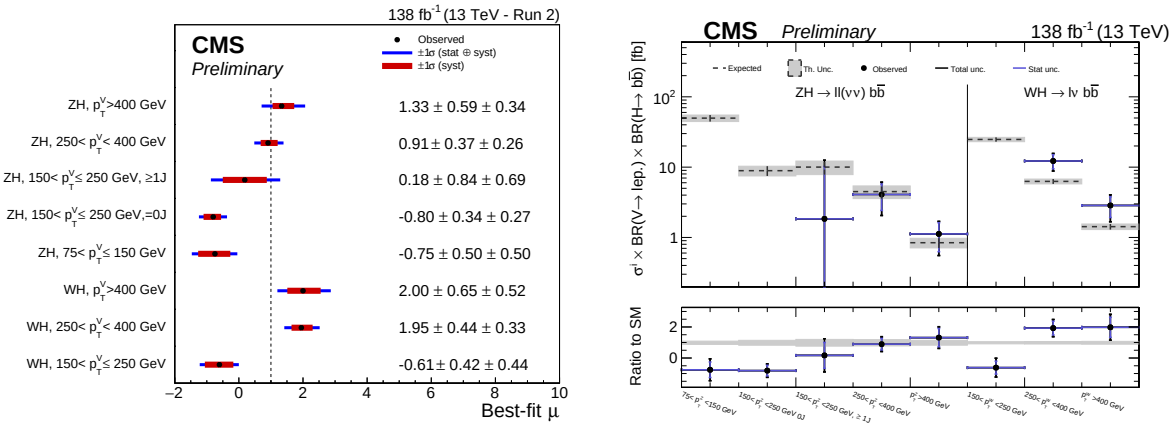
Analiza udarnog presjeka za VZbb

Za provjeru valjanosti analize, proces VH, $H \to b\bar{b}$ zamijenjen je s VZ, $Z \to b\bar{b}$ mijenjanjem prozora mase kako bi se uključila masa Z bozona. Posebni MVA-ovi se treniraju za VZ, $Z \to b\bar{b}$ održavajući strategiju prilagodbe istom kao i u glavnoj analizi. Dobivene snage signala za ZZ i WZ procese prikazane su na Slici 5.1 za sve kanale kada se koristi skup podataka od 2016. do



Slika 2: Rezultat za VZ, $Z \to b\bar{b}$ kanal korištenjem punog uzorka podataka iz Run 2 i za WZ i ZZ način produkcije.

2018. godine. Inkluzivna opažena VZ, $Z \to b\bar{b}$ snaga signala je $\mu = 1,16 \pm 0,13$ što odgovara opaženoj i očekivanoj signifikantnosti znatno iznad 5 standardnih devijacija.



Slika 3: Izmjerene snage STXS signala iz prilagodbe (lijevo). Izmjerene vrijednosti $\sigma \times \mathcal{B}$ u istim STXS područjima kao i za jačine signala, kombinirajući sve godine (desno).

STXS mjerenje za $VH, H \to b\bar{b}$

Signal VH, $H \to b\bar{b}$ dobiven je za svako STXS područje iz prilagodbe koja kombinira skupove podataka od 2016. do 2018. godine. Inkluzivna snaga signala u odnosu na standardni model ($\mu = 1$) izmjerena je na $\mu = 0,58^{+0,19}_{-0,18}$ što odgovara opaženoj (očekivanoj) signifikantnosti od 3,3 (5,2) standardne devijacije. Slika 5.6 (lijevo) prikazuje izmjerene jačine signala u svakom STXS području. Ovi se rezultati dalje tumače kao udarni presjeci za VH produkciju pomnoženi omjerom grananja ($\sigma \times \mathcal{B}$) za V $\to$ leptoni i H $\to$ bb na Slici 5.6 (desno). Kako bi se rezultati predstavili kao udarni presjeci za produkciju, teorijske nesigurnosti koje mijenjaju ukupni udarni presjek pojedinačnih STXS područja ili inkluzivni udarni presjek uklonjene su iz prilagodbe.

Ključne riječi: LHC, CMS, standardni model, Higgsov bozon, ultrarelativistički objekti, DeepCSV, DeepAK8, STXS

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The idea that all matter in the universe is composed of tiny indivisible particles is a primitive one and dates back to as early as 6^{th} century BC. Some of the early foundations were laid by Jains in India between 9^{th} and 5^{th} century BC. According to some of the founders of Jainism religion, the ajiva (non living part of universe) consists of matter or pudgala, of definite or indefinite shape which is made up of tiny uncountable and invisible particles called permanu. Permanu occupies space-point and each permanu has definite colour, smell, taste and texture. Infinite varieties of permanu unite and form pudgala. Some of the philosophical theories on atom and its nature were also studied by the Greek philosophers such as Leucippus, Democritus, and Epicurus. In the 5th century BC, Democritus postulated an atomic theory of the universe by naming these components atoms (from Greek atomon "uncuttable, indivisible"). Sometime around 2^{nd} to 4^{th} century BC ancient Indian philosopher Kanada founded the Vaisheshika school of Indian philosophy which was centered around studying "naturalism" or atomism in natural philosophy.

Although profound theories existed before the era of modern physics, the fundamentals of such theories were abstract and based on philosophical reasoning rather than experimental observations. In the early 1800's John Dalton postulated the fundamentals of atomic physics using principles of stoichiometry. His theories on atom were based on laws of conservation of mass and constant composition, i.e., a pure compound will always have the same proportion of the same elements throughout. By the end of 19^{th} century, J.J Thompson discovered the electron as a negatively charged particle situated inside the atom. He made the discovery by passing high voltage in a cathode ray tube that resulted in emission of electrons or "cathode rays" from cathode to anode. On introducing a magnetic field, the electron ray got deflected, proving that the discovered particle is a charged particle. In later years, the nucleus and its constituents, the proton and the neutron were also discovered. This was the beginning of atomic physics era.

During the same period, radioactivity was also discovered by Henri Becquerel and Marie Curie, while working with phosphorescent materials which led to the foundation of principles of particle interactions. Beginning of the 20^{th} century also marked early developments in the field of quantum mechanics and the concept of particle-wave duality. Today's theoretical formulations of particle physics are based on quantum field theory which results from the unification of quantum mechanics and the theory of relativity.

The Standard Model of particle physics, as we know it, is the theory of fundamental particles and interactions between them. In particle physics, such a fundamental theory was required to systematize all the new fundamental particles which were discovered in the last century. In the early 1900's the particle nature of light was proposed by Albert Einstein when he tried to explain the photoelectric effect. This theory was further solidified by A. H. Compton in 1923 with the discovery of Compton scattering. This led to a more definite approach to particle interactions via electromagnetic force. Electron neutrinos were also proposed shortly after, to resolve the apparent violation of conservation of energy in beta-decay. The theory of strong nuclear force was proposed by Yukawa in 1934. He assumed that the protons in a nucleus must be attracted to one another by exchanging a "Yukawa particle". However, no such particle was experimentally observed till then to support his claim. This puzzle was resolved in 1946 when pions (π) were discovered from cosmic ray experiments and for a long time pions were "considered to be Yukawa particles that keep the nucleus bound. Muons (μ) were also discovered in one cosmic ray experiment and behave in every way like a heavier version of the electron and properly belongs in the lepton family. Similar to the electron neutrino, the muon neutrino was also proposed by applying momentum conservation principles on muons decaying into electrons. It was observed that electrons on average carry only a third of the muon momentum leading to believe there must be two missing particles, one electron neutrino and one muon neutrino.

Cosmic rays proved to be a reliable source for discovering new particles and during the 1940's many new hadrons like K's, Λ's, Σ's, and Θ's were discovered. In 1952, the first modern particle accelerator began operating and for the first time it was possible to create particles in a laboratory rather than relying on cosmic ray experiments. With so many new particles being discovered in quick succession, physicists had a hard time accommodating all the hadrons in a single pattern, a periodic table of particles. Gell-Mann and Zweig, in 1964, made a proposal that all hadrons are composed of even more elementary particles. This led to the birth of the quark model and until 1974 the known quarks were up, down and strange quark. In 1975, with the discovery of the J/ψ meson, the charm quark was discovered. The Tau lepton (τ) along with

its associated neutrino were also discovered in the same year. The Standard Model was nearly complete by that time but now there was an anomaly. There were 3 generations of leptons but only 2 generations of quarks. The Bottom quark was shortly discovered in 1977 which led particle physicists to believe that there should exist one last quark which is yet to be discovered. Meanwhile, the gluon, the mediator of the strong nuclear force (not π's as proposed by Yukawa) was also discovered during this era in 1979 while the $W^{\pm}$ and Z bosons, the last of gauge bosons mediating weak nuclear interactions, were discovered in 1983. Finally in 1995, the case of the last quark to be discovered was settled with the discovery of the top quark, making the three generations of quarks complete. By the beginning of the 21^{st} century, the Standard Model of particle physics was considered complete (for the time being), with all the known discovered leptons, quarks and gauge bosons until the Higgs boson was discovered in 2012 independently by the CMS and the ATLAS experiment using data collected from proton-proton collisions at the Large Hadron Collider (LHC). Details of the Standard Model of particle physics will be discussed in the next chapter.

1.1 Standard model

The Standard Model (SM) of particle physics was constructed to explain fundamental interactions of particles involving electromagnetic, strong nuclear and weak nuclear forces. Its foundation is laid upon the principles of quantum field theory (QFT) which considers particles to be excitations of the respective quantum field, e.g. the electron is an excitation in the electron field. Various interactions in QFT are represented by which are Lagrangians that describe interactions between the respective quantum fields. The current established SM does not account for the gravitational force. Particles that constitute the SM can be broadly divided into two groups, namely fermions and bosons.

The Higgs mechanism was introduced to explain how the mass of the weak-force bosons are generated. Contrary to observations, these bosons would not have mass in the SM without the Higgs mechanism. Through the so-called Yukawa couplings, the Higgs boson's introduction into the Standard Model can also explain how fermion masses came to be. A major victory for the SM came with the finding of a new particle, that is consistent with the Higgs boson, about 50 years after the mechanism was first postulated [22, 23, 24]. In 2012, it was discovered jointly by the CMS [15] and ATLAS [25] experiments. In section 1.2.2, a brief explanation of the Higgs mechanism and Yukawa coupling is described. Fig.1.1 shows an overview of the standard model.

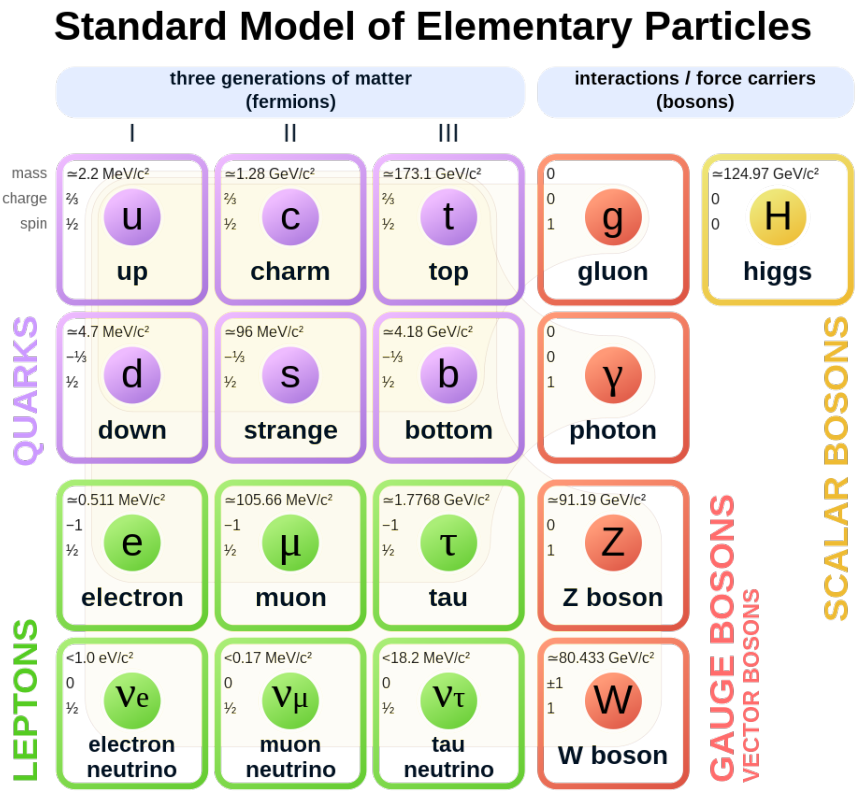


Figure 1.1: The Standard Model of particle physics, with the Higgs boson as the latest addition. Figure taken from [1].

1.1.1 Feynman diagrams

Feynman diagrams were introduced in 1948 by Richard Feynman to visually represent otherwise complex particle interactions. It is extensively used as a tool in theoretical particle physics to construct new particle interaction models and compute complex properties of interactions. The rules of the diagrams are quite simple and easy to visualize. Fig. 1.2 shows the simplest feynman diagram with an electron-positron (straight line with arrows) annihilating to release a photon (wiggly line). One can observe that particles are shown by arrows pointing in the direction of time while anti-particles are denoted to be backward in time. The progression in time is horizontally from left to right. All electromagnetic and weak nuclear interactions are denoted by wiggly lines (Fig. 1.3) and mediated by W and Z bosons while QCD interactions are denoted by curly lines (Fig. 1.4) and these represent propagators in QFT formalism. In my thesis, I will be using Feynman diagrams to explain various processes.

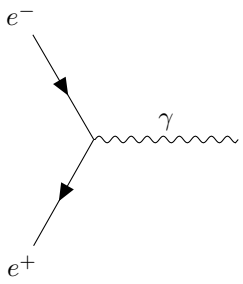


Figure 1.2: Feynman diagram of electron-positron annihilation to release a photon.

1.1.2 Fermions

Fermions are spin half particles which make up all visible matter in the universe. They obey Pauli's exclusion principle and follow Fermi-Dirac statistics and hence the name fermions. Fermions can be further classified into leptons and quarks, and they come in 3 generations based on hierarchy of masses, from the lightest to the heaviest.

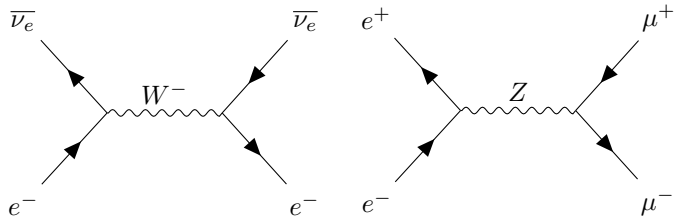


Figure 1.3: Feynman diagrams of weak nuclear interactions representing the two kinds of electroweak bosons, the W and the Z boson. A W^- boson is produced between an electron and electron neutrino (left) and a Z boson produced from electron-positron annihilation which further decays into a pair of muons (right).

Quarks

Quarks are the class of fermions which follow the QFT framework designed for strong nuclear force called Quantum Chromodynamics (QCD). Besides that they also interact via electroweak mechanisms. Quarks can be categorized into 2 types based on electric charge and further into 3 generations based on mass, making it a total of 6 quarks in the standard model. The distinct feature of each of these quarks is that they possess fractional charge of $+\frac{2}{3}$ for up-type and $-\frac{1}{3}$ for down-type quarks. QCD allows for quarks to possess a kind of charge similar to electric charge and this is termed as color charge. Each quark comes in 3 different color charges, namely red, blue and green. In QCD, color charge should always be conserved, therefore quarks with color

don't exist as free particles and are always confined in bound states with other quarks or anti-quarks forming hadrons which is "white" in color. This phenomena is called color confinement. Proton and neutron are well known examples of hadrons which are a bound state of 3 quarks, 2 up and 1 one down for the proton and 1 up and 2 down quarks for the neutron. Such hadrons are called baryons. There are also hadrons which are made up of 2 quarks. They are called mesons and π meson is one good example of such a bound state.

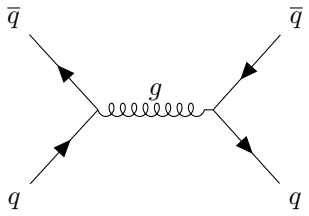


Figure 1.4: Feynman Diagrams of QCD interactions showing gluon (curly line) exchanged being two pairs of quarks.

Leptons

Leptons, unlike quarks, don't carry a color charge and purely interact via electoweak interactions and therefore, can be observed as free particles. Similarly to quarks, they are also classified into 2 types based on electric charge, and in 3 mass generations. Leptons possess integer values of electric charge. Electron, muon and tau are the 3 generation of charged leptons that differ from one another in terms of mass, i.e. electron being the lightest and tau being the heaviest. Neutral leptons are called neutrinos and they are named after their respective charged lepton in that generation, e.g. electron neutrino (denoted by ν_e). They have feeble electroweak interactions unlike charged leptons. They are also extremely light particles and for a long time they were considered massless. It is only in the last decade that upper limits on the mass of neutrinos are measured.

1.1.3 Bosons

Bosons are integer spin particles which do not obey Pauli's exclusion principle and follow Bose-Einstein statistics. These particles emerge as mediating particles for various interactions. Electromagnetic interactions are mediated by photons and QCD interactions are mediated by gluons. Weak nuclear interactions are mediated by $W^{\pm}$ and Z bosons. All the Lagrangians which are represented by the force carrier bosons are invariant under gauge symmetry. There-

fore, these bosons are also called gauge bosons. They also have the characteristic feature of having integer spin values. Higgs boson is the only observed scalar (spin 0) boson.

1.2 The electroweak theory

The SM Lagrangian is a gauge theory meaning it is invariant under local gauge transformations which form certain Lie groups. The Lie groups which give rise to the interactions described by the SM are the SU(2)$\otimes$ U(1) and SU(3), corresponding to electroweak and strong interactions respectively. More generally speaking, any process which can be described by the SM preserve charge, parity and time symmetry with few exceptions. If we take the simplest example of electromagnetic interactions which follows U(1) symmetry and apply it on Dirac equation, it can be seen that it is invariant under gauge transformation. The Lagrangian from which the Dirac equation is derived is

$$\mathcal{L}_{free} = \overline{\Psi}(i\gamma^{\mu}\partial_{\mu} - m)\Psi(x) \tag{1.1}$$

Applying local U(1) gauge invariance requires the Lagrangian to be invariant under $\psi'(x) \rightarrow e^{iQ\theta}\psi(x)$ transformation. To achieve this symmetry, a new gauge field $A_{\mu}(x)$ is added to the Dirac Lagrangian which transforms as:

$$A'_{\mu}(x) = A_{\mu}(x) + \frac{1}{e}\partial_{\mu}(\theta) \tag{1.2}$$

We also define a covariant derivative

$$D_{\mu}\psi(x) = [\partial_{\mu} - ieQA_{\mu}(x)]\psi(x) \tag{1.3}$$

Applying U(1) transformation accounting for the $A_{\mu}(x)$ term, the newly constructed Lagrangian is

$$\mathcal{L} = i\overline{\psi}(x)\gamma^{\mu}D_{\mu}\psi(x) - m\overline{\psi}(x)\psi(x) = \mathcal{L}_{free} + eQA_{\mu}(x)\overline{\psi}(x)\gamma^{\mu}\psi(x) \tag{1.4}$$

and is invariant under local U(1) transformations. This additional field term ($A_{\mu}(x)$) describes the interaction between this field and the fermions. We can recognize that this field can be attributed to the gauge boson of the electromagnetic interaction, the photon.

The Lagrangians describing the strong and weak interactions are constructed using similar methods, requiring local gauge invariance under SU(3) and SU(2) $\otimes$ U(1) symmetry groups,

respectively.

1.2.1 Weak interaction and electroweak unification

The motivation for weak interaction was driven by nuclear fission and fusion reactions. The electroweak theory describes the electromagnetic force and the weak force in a unified theory based on a local $SU(2)_L \otimes U(1)_\gamma$ gauge symmetry. $U(1)_\gamma$ group represents the electromagnetic interactions mediated by photon, while weak interaction is generated by $SU(2)_L$ group. The subscript L in $SU(2)_L$ group signifies the transformation in left-handed fields while right-handed fields remain invariant under $SU(2)_L$ transformation. This implies that weak interactions do not conserve parity. A new quantum number is introduced for weak interactions, the weak isospin, I. Its projection represents the charge corresponding to the $SU(2)_L$ symmetry while another quantum number by the name of hypercharge is associated to $U(1)_\gamma$.

The chiral operators $(1-\gamma_5)/2$ and $(1+\gamma_5)/2$ project fermions into their respective left/right-handed components. Left-handed fermions transform as weak isospin-doublets under SU(2).

$$\chi_L = \quad l_L = \begin{pmatrix} \nu_e \\ e_L \end{pmatrix} \; or \; Q_L = \begin{pmatrix} u_L \\ d_L \end{pmatrix} \tag{1.5}$$

and the right-handed fermions

$$\Psi_R = e_R \; or \; u_R \; or \; d_R, \tag{1.6}$$

as singlets (e and ν_e for leptons and u, d for quarks). The corresponding local transformations are

$$\chi_L \rightarrow \chi'_L = e^{-ig\vec{\alpha}\vec{T} - ig'\beta\frac{\gamma}{2}}\chi_L, \tag{1.7}$$

$$\psi_R \rightarrow \psi'_R = e^{-ig'\beta\frac{\Upsilon}{2}}\psi_R \tag{1.8}$$

where $\vec{T} = \frac{1}{2}\vec{\sigma}$ are the generators of the $SU(2)_L$ group with $\vec{\sigma}$ being the Pauli matrices and Υ is the hypercharge operator. The couplings g' and g are gauge couplings of $U(1)_\Upsilon$ and $SU(2)_L$ respectively.

The field strength tensors are defined as,

$$B^{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu \tag{1.9}$$

$$\mathcal{W}^a_{\mu\nu} = \partial_\mu \mathcal{W}^a_\nu - \partial_\nu \mathcal{W}^a_\mu - g\varepsilon^{abc}\mathcal{W}^b_\mu \mathcal{W}^a_\nu \tag{1.10}$$

where ε^{abc} is the Levi-Civita [26] tensor and $\overrightarrow{\mathcal{W}_\mu} = \left(W^1_\mu, W^2_\mu, W^3_\mu\right)$ and B_μ represent the gauge fields. Two of them can be associated to the two charged bosons mediating the weak force with the following transformation:

$$W^{\pm,\mu} = i\frac{1}{\sqrt{2}}(W^{\mu,1} \mp W^{\mu,2}). \tag{1.11}$$

The third boson is a neutral gauge boson which is reminiscent of the neutral gauge boson in the electromagnetic Lagrangian. This is a hint that the electromagnetic and weak forces may be unified into a single force, the electroweak (EW) force.

The associated covariant derivative for electroweak theory that ensures the invariance of the Lagrangian can be defined as,

$$D_\mu = \partial_\mu + ig\overrightarrow{T}.\overrightarrow{\mathcal{W}}_\mu + ig'\frac{\Upsilon}{2}B_\mu \tag{1.12}$$

A linear transformation between the $W^{\mu,3}$ and B^μ can be expressed as:

$$\begin{pmatrix} W^{\mu,3} \\ B^\mu \end{pmatrix} = \begin{pmatrix} \cos\theta_W & \sin\theta_W \\ -\sin\theta_W & \cos\theta_W \end{pmatrix} \begin{pmatrix} Z^\mu \\ A^\mu \end{pmatrix} \tag{1.13}$$

The mixing angle, θ_W, also called the Weinberg angle, is selected so that

$$\theta_W = tan^{-1}(\frac{g'}{g}). \tag{1.14}$$

The motivation for such a choice is that it causes the Z^μ to only couple to isospin and A^μ only to the electrical charge. The two fields then correspond to the Z boson and the photon, respectively, and can be expressed as:

$$\begin{aligned} Z^\mu &= \frac{-g'B_\mu + gW^3_\mu}{\sqrt{g^2+g'^2}} \\ A^\mu &= \frac{gB_\mu + g'W^3_\mu}{\sqrt{g^2+g'^2}} \end{aligned} \tag{1.15}$$

Combining electroweak Lagrangian with the QCD Lagrangian, we finally obtain the Lagrangian density invariant under SU(3)$_c$ $\times SU(2)_L \times U(1)_\Upsilon$ and is expressed as follows,

$$\begin{aligned} \mathcal{L} = i\overline{L}_{iL}\not{D}L_{iL} + i\overline{Q}_{iL}\not{D}Q_{iL} + i\overline{e}_{iR}\not{D}e_{iR} + i\overline{\mu}_{iR}\not{D}u_{iR} + i\overline{d}_{iR}\not{D}d_{iR} \\ -\frac{1}{4}G^a_{\mu\nu}G^{\mu\nu}_a - \frac{1}{4}\overrightarrow{W}^{\mu\nu}.\overrightarrow{W}_{\mu\nu} - \frac{1}{4}\overrightarrow{B}^{\mu\nu}.\overrightarrow{B}_{\mu\nu}, \\ D_\mu = \partial_\mu + ig_s\frac{\lambda_a}{2}\mathcal{A}^a_\mu + ig\overrightarrow{T}.\overrightarrow{W}_\mu + ig'\frac{\Upsilon}{2}B_\mu. \end{aligned} \tag{1.16}$$

which describes a self-consistent massless theory of strong and electroweak interactions.

1.2.2 Electroweak symmetry breaking and Higgs mechanism

In the previous section, we constructed a Lagrangian which is massless but invariant under SU(3)$_c$ $\times SU(2)_L \times U(1)_\Upsilon$ transformations. However, this is not true since all the known quarks and fermions had masses. Including the W and Z bosons which were predicted to be massive particles and were also observed experimentally to be massive when they were discovered. An extra term is added to the electroweak Lagrangian to account for mass of the fermions which would be,

$$\tilde{m}\overline{\psi}\psi = m(\overline{\psi}^R\psi^L + \overline{\psi}^L\psi^R) \tag{1.17}$$

However, as we saw in the previous section, this term cannot be gauge invariant under $SU(2)_L$ since only left-handed fields would be affected by the transformation. The electroweak symmetry breaking (EWSB) an idea which is deeply rooted in condensed matter physics on global symmetries, introduces mass terms for local symmetries.
A Lagrangian density is added:

$$\mathcal{L}_{Higgs} = (D^\mu\phi)^\dagger(D_\mu\phi) - V(\phi), \tag{1.18}$$

where D_μ,

$$D_\mu = \partial_\mu + ig\overrightarrow{T}.\overrightarrow{W}_\mu + ig'\frac{\Upsilon}{2}B_\mu, \tag{1.19}$$

allows for the gauge invariance of the $SU(2)_L \times U(1)_\Upsilon$ Lagrangian to produce the physical symmetries and masses for W and Z bosons while leaving the photon massless. The covariant derivative of a Higgs scalar complex field ϕ, which is a $SU(2)_L$ doublet of the following form,

$$\phi = \frac{1}{\sqrt{2}} \begin{pmatrix} \phi_1 & +i\phi_2 \\ \phi_3 & +i\phi_4 \end{pmatrix} = \begin{pmatrix} \phi^\dagger \\ \phi^0 \end{pmatrix}, \tag{1.20}$$

describes the couplings to gauge fields. The procedure of EWSB, requires the potential $V(\phi)$ to have an infinite number of equivalent minima. Here the following form is considered,

$$V(\phi) = \mu^2 \phi^\dagger \phi + \lambda(\phi^\dagger \phi)^2, \text{ with } \mu^2 < 0. \tag{1.21}$$

The shape of this potential is depicted in Fig. 1.5. Picking one minimum, i.e. $\phi^0_{vac} = 0$ and $\phi^\dagger_{vac} = v$ the symmetry of the vacuum is simultaneously broken. Here v is the vacuum expectation value of the Higgs field [27] and expressed as follows,

$$v = \sqrt{\frac{-\mu^2}{2\lambda}} \approx 246.222 \text{ GeV} \tag{1.22}$$

Applying the minima conditions transforms the fields as follows,

$$W^{\pm\mu} = \frac{1}{\sqrt{2}}(W^{1\mu} \mp iW^{2\mu}) \quad \rightarrow \quad W^\pm \quad bosons \tag{1.23}$$

$$Z^\mu = -B^\mu \sin\theta_w + W^{3\mu} \cos\theta_w \quad \rightarrow \quad Z \quad boson \tag{1.24}$$

$$A^\mu = -B^\mu \cos\theta_w - W^{3\mu} \sin\theta_w \quad \rightarrow \quad \gamma \quad photon \tag{1.25}$$

where θ_w is the weak mixing angle as shown in Eq. 1.14. $\sin^2\theta_w$ is measured experimentally to be 0.23121 ± 0.00004 and the boson masses are related by, $m_W = \frac{1}{2}vg$, $m_Z = \frac{1}{2}v\sqrt{g^2 + g'^2}$. An excitation in the Higgs field is defined as,

$$\phi = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + H \end{pmatrix}; \tag{1.26}$$

replacing it in Eq. 1.18 gives rise to the Higgs boson mass term with $m_H = \sqrt{2\lambda v}$ and the trilinear and quartic self-coupling of the Higgs boson.
To add the fermionic masses, the following Lagrangian density terms are added,

$$\mathcal{L}_{HF} = -y^u_{ij}\overline{Q}_{iL}\widetilde{\phi}u_{jR} - y^d_{ij}\overline{Q}_{iL}\phi d_{jR} - y^e_{ij}\overline{L}_{LiL}\phi e_{jR} + h.c., i, j, = 1, ..., 3; \tag{1.27}$$

where y_{ij} are the Yukawa-couplings and $\widetilde{\phi} = i\sigma_2\phi^*$. The resulting fermionic masses, in the fermion mass eigenstate basis, have the following form,

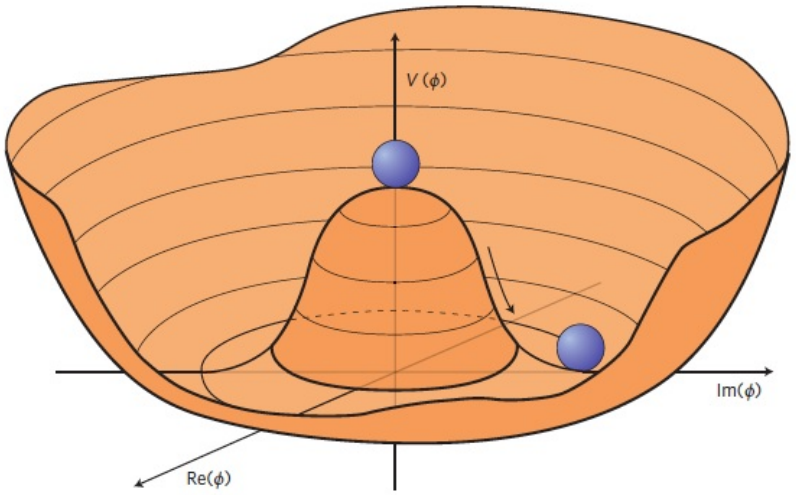


Figure 1.5: Shape of the Higgs potential. Picture taken from [2].

$$m_i^u = \frac{y_i^u v}{\sqrt{2}}, m_i^d = \frac{y_i^d v}{\sqrt{2}}, m_i^e = \frac{y_i^e v}{\sqrt{2}}. \tag{1.28}$$

This demonstrates that for various fermions, the ratio of interaction strengths (to Higgs) equals the ratio of their masses. More information of Higgs mechanisms and Electroweak symmetry breaking can be found in [22, 23, 24, 28].

1.3 Higgs boson phenomenology at the LHC

In this section, we discuss the production and decay modes of Higgs bosons produced at LHC given by theoretical predictions.

1.3.1 Higgs boson production modes

In the proton-proton collisions taking place at the center of CMS detector, the Higgs boson has 4 major production modes, namely gluon-gluon fusion (*ggF*), vector-boson fusion (*VBF*), associated production with a vector boson (*VH*), and associated production with a pair of top quarks (*ttH*). The corresponding Feynman diagrams to all these production modes are shown in Fig. 1.6. Fig. 1.7 shows the calculated value of different Higgs production modes at the LHC.

- In the ggF production mode, gluons don't couple directly to the Higgs boson, but through a loop where virtual quarks are exchanged. Since the coupling of Higgs boson to fermions is directly proportional to particle mass, the quark loop is usually a loop mediated by the

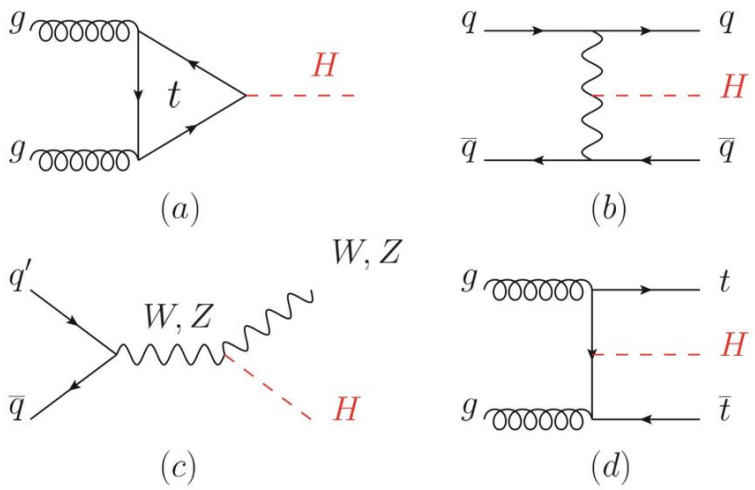


Figure 1.6: Feynman diagrams of the Higgs production at LHC: (a) gluon-gluon fusion, (b) vector-boson fusion, (c) associated production with a vector boson, and (d) associated production with top quarks.

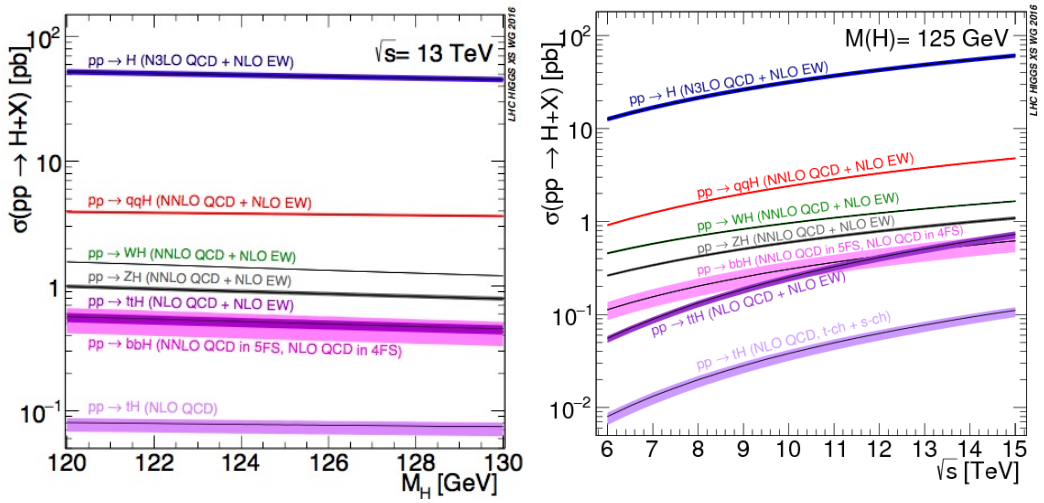


Figure 1.7: Calculated production cross-section of the Higgs boson via various production modes as a function of m_H (right) and as a function of $\sqrt{s}$ (left). Figure taken from [2].

heaviest quarks (top and bottom quarks). This is the dominant production mode of Higgs boson at the LHC. However it suffers from an overwhelming amount of background in form of QCD jets from the hard scattering of proton collisions.

- In the VBF mode, a pair of quarks radiate a vector boson (W or Z boson) while getting slightly scattered in forward direction. The radiated vector bosons "fuse" together to produce a massive particle such as a Higgs boson. The scattered quarks result into two hard jets in the forward regions of the detector giving a major signature of this mode.

VBF is the second largest cross-section process in the Higgs production at LHC.

- The VH production mode starts off with a quark anti-quark pair interacting weakly to form a virtual vector boson. If the radiated vector boson has high enough energy it radiates a Higgs boson. The kinematic information of the decaying leptons from the vector boson help in suppressing large QCD backgrounds obtaining a higher signal purity even though having the third largest cross-section among all Higgs production modes.
- In the ttH process, two gluons collide, with each decaying to a top-antitop quark pair. A top quark and an antitop quark from each pair form a Higgs boson together. The ttH process has the fourth largest cross-section.

1.3.2 Higgs boson decay modes

The latest measurement of mass of Higgs boson resultedin a value of to be 125.18 ± 0.16 . With the Higgs mass m_H treated as an input to the theoretical model, the prediction on the branching ratio of the Higgs decay channels can be made, as shown in Fig.1.8. Different decay products

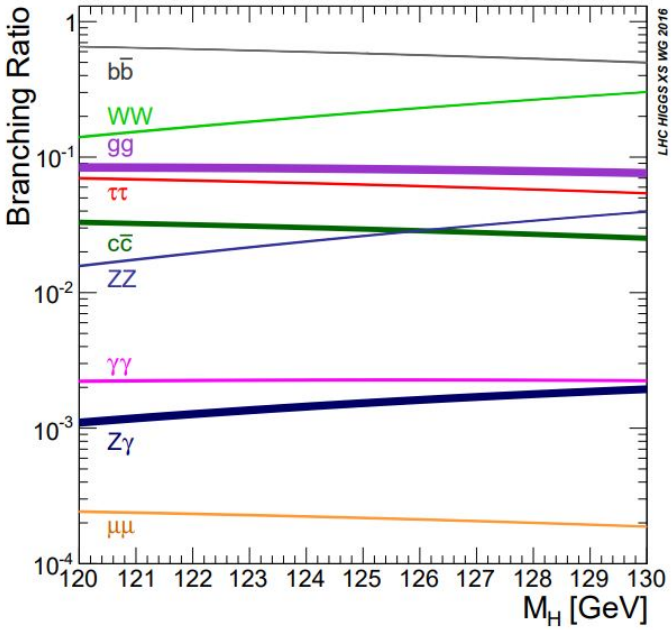


Figure 1.8: The branching ratios of the Higgs boson decays near m_H = 125 GeV . The theoretical uncertainties are represented as bands. Figure taken from [2].

of the Higgs boson result in different features in the decay channel. $H \to b\bar{b}$ has the highest branching fraction among all Higgs boson decays. However, the measurement of $H \to b\bar{b}$

process highly depends on the resolution of b jets and efficiency of tagging b-jets. $H \to \gamma\gamma$ and $H \to ZZ \to 4l$ are known to be golden channels and the first Higgs boson discovery were made possible through these channels even though having a lower branching ratio than other Higgs decays. This is because all particles in their final state can be well reconstructed making m_H to be measured with excellent resolution. Contrary to e.g. $H \to W^+W^- \to l^+\nu_l l^-\overline{\nu}_l$ and $H \to \tau^+\tau^-$ decay modes that suffer from the energy loss due to the presence of neutrinos and large backgrounds, thus, their m_H resolutions are relatively poor.

1.3.3 VHbb production at the LHC

Higgs boson decaying into a pair of b quarks ($H \to b\bar{b}$) has a branching ratio of 58% among all Higgs decays. As already stated both gluon-fusion and vector boson fusion modes have a higher production cross-section than VH mode but suffer from large mulitjet background. The VH production mode decaying into b quarks ($VH \to b\bar{b}$) relies on triggers based on leptons coming from decays of the W or the Z boson making VH production cleaner in terms of background rejection. The W boson can be reconstructed from its leptonic decay $W \to l\nu$ (l = e or μ), while the Z boson can be reconstructed from the decay of $Z \to e^+e^-$, $\mu^+\mu^-$ or $\nu\overline{\nu}$ where the presence of a neutrino is inferred from the missing transverse momentum observed in the detector. The Higgs boson can be reconstructed from a pair of b-jets which are identified using b-tagging algorithms. The main challenges in measuring $VH \to b\bar{b}$ come from background modeling, efficiency in tagging b-jets and measuring its momentum and energy resolution.

1.4 Limitations of the SM

So far SM has been consistent with what has been experimentally observed and in the last decades it has been proved to be a great asset to particle physicists however, there are still many open ended questions in field of particle physics which are not answered by the current SM. Listed below are few of many phenomena which are not well explained by the current SM.

- **Gravity in SM**

  It is understood for a long time that our nature consists of four fundamental forces. While the electromagnetic, weak and strong nuclear forces are well explained by the SM, gravity is not accounted for. It also fails to explain why gravitational force is so weak compared to the other 3 forces. Graviton is the hypothetical particle that mediates gravitational interactions, however it has not yet been experimentally observed.

- **Dark matter and dark energy**

  Dark matter is another elusive topic which is not explained by the SM. Through cosmological experimental observations it has been established that ordinary matter constitutes 5% while dark matter constitutes 26 % of the known universe [29]. The rest of the 69% is occupied by dark energy [29]. However they at best feebly interact with SM particles and therefore have not yet been experimentally observed. There are theories and models explaining coupling of Higgs boson to dark matter particles but they are beyond SM formalism.

- **Matter antimatter asymmetry**

  Since the Big Bang and the early creation of the universe, it is predicted that matter and antimatter should have been created in equal proportions throughout the universe. However, it is observed that the amount of baryons far exceeds the amount of anti-baryons leading to baryon asymmetry in the universe. The current SM fails to explain this asymmetry. It is also unable to predict any underlying theories which could be leading to this phenomena. So far, charge parity symmetry violation (CP violation) in baryons is one requiremnt which could cause this asymmetry and it was experimentally observed in 1964 with neutral kaons leading to Nobel prize in physics in the year 1980. Although CP violation is currently allowed in SM, it is alone insufficient to justify the baryon asymmetry. Another necessary condition for the asymmetry involves baryon number violation mechanism which is yet to be observed experimentally.

Chapter 2

Detector and Experiment

2.1 Large hadron collider

Large Hadron Collider (LHC) is world's largest and most energetic particle collider. It is a circular collider with a total circumference of 27 km situated 175 m underground on the national border of Switzerland and France. Fig. 2.1 shows an aerial view of the Swiss-French border with a schematic of the LHC ring situated underground. As described in previous chapter, hadrons are bound state of quarks bound together via strong interactions by exchanging gluons. Most of the hadron collisions that happens at LHC are proton-proton collisions. Besides that, fraction of collisions also happen in combination with heavier ions, e.g. lead. Proton-lead and lead-lead collisions happen for around one month in a year of total LHC collisions. LHC started circulating proton beams for the first time on September 10th, 2008. By end of 2009, it had already crossed Tevatron's record energy of 0.98 TeV making it the most energetic particle accelerator in the history. By March of 2010, the beams had ramped up to an energy of 3.5 TeV and first proton-proton collisions happened at LHC at a record collision centre of mass energy ($\sqrt{s}$) of 7 TeV. In 2012, the beam energy was ramped up to 4 TeV and collisions happened at $\sqrt{s}$ = 8 TeV. This is known as the Run-1 era of the LHC. After that a period of 1^{st} long shutdown from 2012 to 2015 was introduced for various R&D activities in order to achieve higher collision energies. LHC restarted collisions in 2015 and continued operating until 2018 at a record $\sqrt{s}$ of 13 TeV. This marked the Run-2 era of LHC. Recently LHC restarted collisions with Run-3 in July, 2022 with an even higher collision energy of $\sqrt{s}$ = 13.6 TeV after the period of 2^{nd} long shutdown. Current Run-3 is supposed to happen till 2025. Besides $\sqrt{s}$ of colliding protons, another important quantity in any collider experiment is instantaneous luminosity denoted by $\mathcal{L}$. This quantity measures the number of particle collisions per unit of time. Therefore, while increasing center of mass energy, LHC also has to ensure steady increase

Figure 2.1: Traversed path of LHC ring as observed from an aerial view of Swiss-French border. Figure taken from [3].

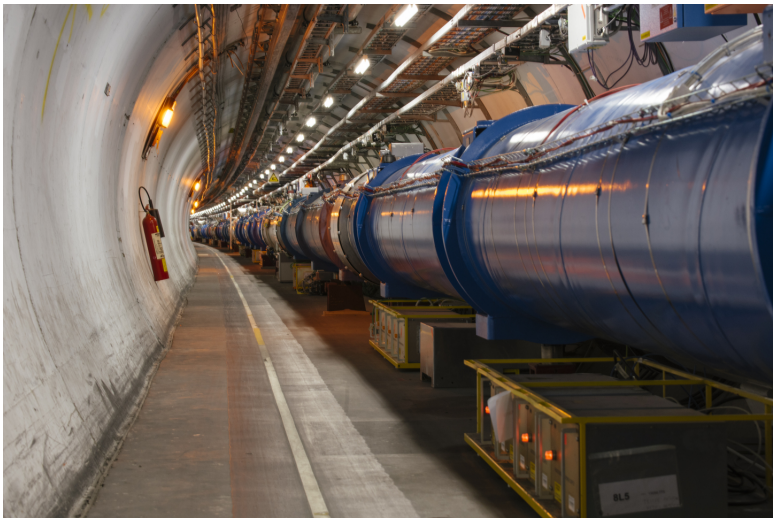


Figure 2.2: LHC tunnel situated underground which contains the LHC beam pipes

in luminosity to achieve higher statistical precision in collision data. The general expression for $\mathcal{L}$ is the following:

$$\mathcal{L} = \frac{n_b f_{rev} N_1 N_2}{4\pi\sigma_x\sigma_y}\mathcal{F} \tag{2.1}$$

where f_{rev} is revolution frequency, n_b is number of colliding bunch pairs, $N_{1,2}$ are the beam intensities while $\sigma_{x,y}$ are transverse sizes of the beam at the collision point assuming beam

direction is along z-axis. $\mathcal{F}$ is the geometric factor. In LHC, N_1 and N_2 are the same, so they can be replaced be a common intensity term N_b^2. σ_x and σ_y term can be expressed as,

$$\sigma_x = \sqrt{\beta^* \varepsilon_x \gamma^{-1}} \quad \sigma_y = \sqrt{\beta^* \varepsilon_y \gamma^{-1}} \tag{2.2}$$

where, ε is normalized transverse emittance along x and y axis, β^* is the beta function at collision point while γ is the relativistic factor. In LHC, $\varepsilon_x = \varepsilon_y$, so they can be replaced by a common emittance term, ε_n. Putting values of σ_x and σ_y in equation 2.1 we obtain,

$$\mathcal{L} = \frac{n_b f_{rev} N_b^2 \gamma}{4\pi \varepsilon_n \beta^*} \mathcal{F} \tag{2.3}$$

The geometric factor $\mathcal{F}$ is a relativistic correction term, that determines the change in luminosity in case the beams collide at an angle. It is expressed as,

$$\mathcal{F} = \frac{1}{\sqrt{1 + \frac{\alpha \sigma_z^2}{2\sigma_t}}} \tag{2.4}$$

where α is the beam crossing angle, σ_z is the bunch length and σ_t is the transverse width of the bunch. Determination of beam parameters is a necessary step towards luminosity measurements at LHC and this is performed with Van der Meer scans. It involves scanning the LHC beams through one another to determine the size of the beams at their point of collision. These measurements, when combined with information on the number of circulating protons, allow the determination of an absolute luminosity scale. Integrated luminosity is another related quantity which measures the number of collisions over a period of time. Mathematically, it is the time integration of instantaneous luminosity as shown in Eq. 2.5. Integrated luminosity is typically expressed in unit of inverse barns. This is a unit of area widely used by particle and nuclear physicists to represent cross-sectional area of particle interactions. 1 barn converts to 10^{-28} m^2 of area and 1 fb^{-1} (inverse femtobarns) of integrated luminosity corresponds to 1 collision per fb of area which is equal to 10^{-43} m^2 of area in SI units. In simple terms, 1 fb^{-1} of collision data is equivalent of approximately 10^{12} p-p collisions.

$$\mathcal{L}_{int} = \int \mathcal{L} dt \tag{2.5}$$

Total integrated luminosity delivered by LHC in the CMS detector from 2015-2018 is shown is Fig. 2.3.

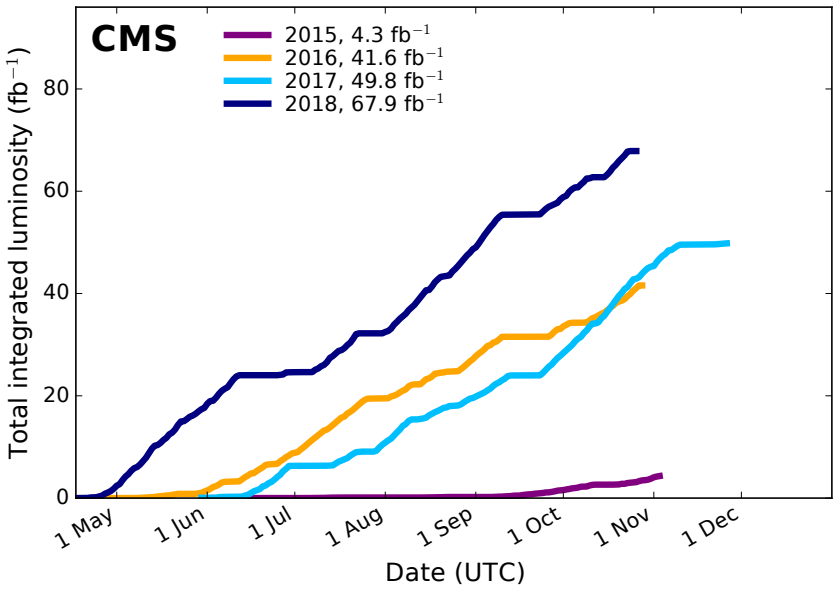


Figure 2.3: Luminosity delivered by LHC year by year in the CMS detector in Run 2 era. Figure taken from [4].

2.1.1 Accelerator Design

As mentioned in the introduction, LHC is built underground at a depth of 175 m. Below this depth exists the LHC tunnel which hosts the LHC beam pipes as shown in Fig. 2.2. Two beam pipes circulate inside the LHC tunnel carrying proton beams in opposite direction to one another. This tunnel was previously home to the Large Electron-Positron (LEP) collider where instead of hadrons electron-positron collisions were happening. The maximum collision energy achieved at LEP was 209 GeV. This is a much lower collision energy than today's LHC collisions owing to energy loss through synchrotron radiation which is much more pronounced in case of electrons. Synchrotron radiation is a form of electro-magnetic radiation which is emitted when a relativistic charged particle is moving perpendicular to a magnetic field. Since the radiation energy is inversely proportional to the fourth power of mass of the charge particle ($\propto \frac{1}{m^4}$), electrons radiate approximately 10^{13} times more energy compared to protons. This is one of the primary reasons for using protons as the colliding particles at the LHC collisions.

To achieve the required collision energy at LHC, protons are accelerated in several stages before getting injected into the LHC ring. For this reason, previous particle accelerators with lower energies are utilized. All of these accelerators are located in the vicinity of LHC and are in-

terconnected to ramp up the acceleration of the colliding proton beams. The course of protons during ramp up involves passing through the following stages:

- **Linac 4 -** Linear accelerator 4 or Linac stage 4 is the first stage where hydrogen gas is ionized in presence of electric field to generate free protons for acceleration. Protons then enter the accelerator and by end of the Linac tunnel protons are accelerated up to 160 MeV. Linac 4 started its operation in 2020, following its predecessor Linac 2 which had an acceleration energy of 50 MeV.

- **PSB -** The Proton Synchrotron Booster is made up of four superimposed synchrotron rings that receive beams of protons from Linac 4 at 160 MeV and accelerate them to 1.4 GeV for injection into the Proton Synchrotron (PS).

- **PS -** The Proton Synchrotron accelerates either protons delivered by PSB or heavy ions delivered by the Low Energy Ion Ring (LEIR). It consists of 277 magnets located in a ring of 628 meters and is CERN's first synchrotron. In 1960s, the PS was the world's highest energy particle accelerator. The accelerator boosts protons up to 26 GeV.

- **SPS -** The Super Proton Synchrotron is a nearly 7 km long circular accelerator and is the second-largest accelerator at CERN. It provides proton or ion beams to the LHC by taking particles from the PS accelerating them up to 450 GeV. Besides, acting as a mediator collider between PS and LHC, it also serves as an independent collider providing beam collisions to other experiments at CERN: NA61, NA62 and COMPASS. The SPS was switched on in 1976 and played a crucial role in 1983 in the discovery of W and Z boson while running as a proton-antiproton collider.

Once the protons are inserted into the LHC ring, they need to be further accelerated to the designed beam energies. Sixteen Superconducting Radio Frequency (RF) cavities (8 per beam) are used for acceleration, applying a 400 MHz oscillating electrical field parallel to the beam line. After the beam reaches its nominal energy, the RF cavities provide the beam with the energy lost due to synchrotron radiation. The oscillating electrical field of the cavities also shapes the proton bunches. Protons which are ahead of the rest in the bunch will be decelerated, while the protons at the back of the bunch will get accelerated, centering the proton bunch. Each bunch contains about 110 billion protons and is approximately 7.5 cm long. The time separation between the bunches is 25 ns, corresponding to 7.5 m distance at the speed of light.

LHC Dipole

Around 1200 magnetic dipoles are placed in beam pipes to keep the hadrons in circular path. Additionally 4$\tilde{0}$0 magnetic quadrupoles are placed to keep the beams focused while stronger quadrupoles are placed at collision points to maximize collisions at crossing. Since the magnets are superconducting, the temperature required to maintain superconductivity is achieved using approximately 96 tonnes of liquid Helium. Overall temperature around the superconducting magnet coil is 1.9 K (Fig. 2.4). An alloy of Niobium and Titanium metals is used for making the superconducting magnets used in LHC dipoles. Around 10,000 superconducting magnets are installed in the LHC beam pipe combining all the magnets from dipoles, quadrupoles and some higher multipole magnets used in the LHC beam pipe.

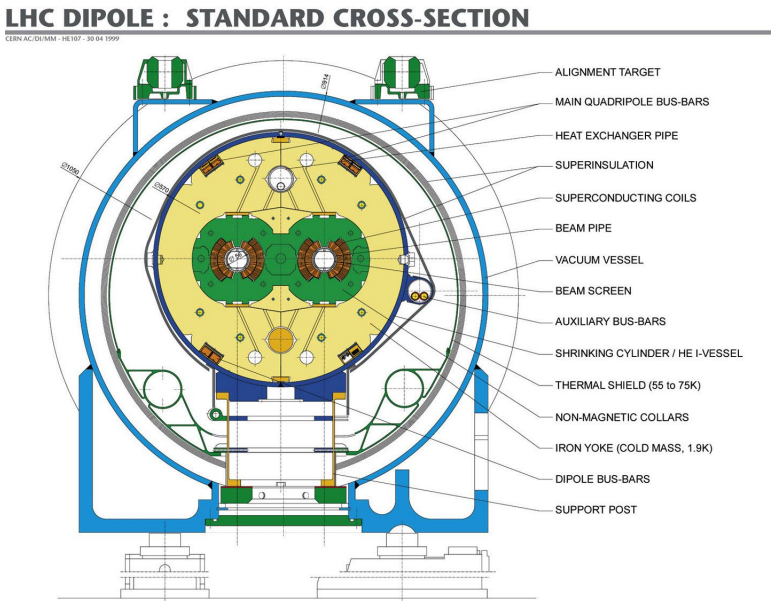


Figure 2.4: Transverse cross-section view of the LHC beam pipe consisting of two beam carrying tubes surrounded by superconducting dipoles. Figure taken from [5]

Collisions at LHC happen at four points along the LHC ring where four major experiments of LHC are situated. These experiments are situated to detect energetic particles originating from p-p collisions at these points in LHC. Our analysis uses data collected by the Compact Muon Solenoid (CMS) experiment which is located at the collision point 5 of the LHC. ATLAS, ALICE and LHCb are the other LHC experiments which are built at different locations on the LHC ring to record collision events similar to CMS experiment. Higgs boson was observed using collision data collected from the period of 2010-2012 (Run 1). This was the first major success of the LHC and in years to come, both CMS and ATLAS promises wealth of data to be collected which would lead to new discoveries of particles associated to new physics theories.

2.2 Compact Muon Solenoid experiment

The Compact Muon Solenoid (CMS) detector is a cylindrical detector built around one of the 4 proton-proton collision points at the LHC. The purpose of such an arrangement is to record energetic particles from collision in all 3 dimensions. It is one of the general purpose detectors designed to observe possible new physics phenomena that LHC might produce. Fig. 2.5 show a schematic diagram of the CMS detector.

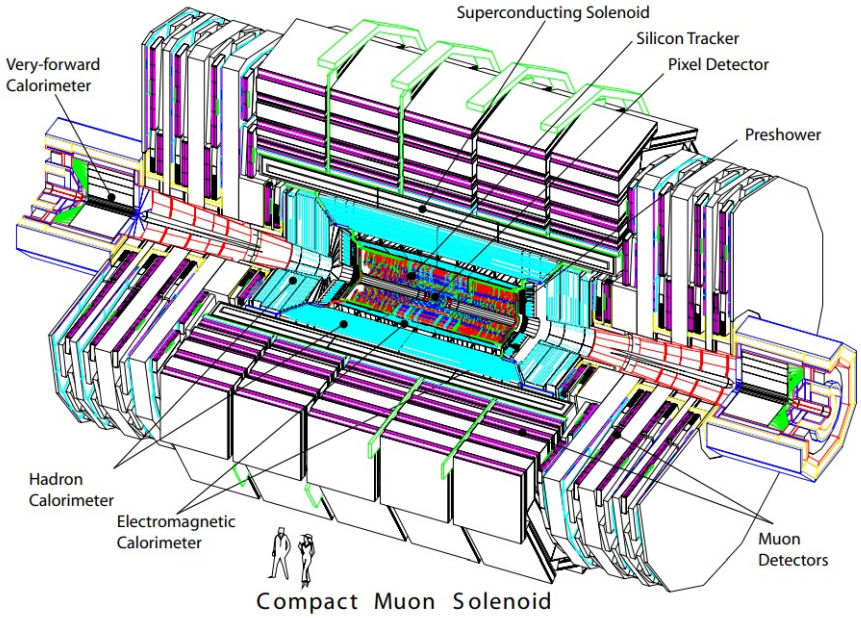


Figure 2.5: A schematic diagram of the CMS detector. Figure taken from [6].

2.2.1 Detector Geometry

CMS detector is a cylindrical detector centered around the beam pipe of the LHC. It uses right-handed coordinate system with origin centered around the nominal collision point. The z-axis is considered along the beam pipe while the transverse cross-section of the detector is on the x-y plane. Since protons are composite particles, it is a priori impossible to determine the momentum of the colliding partons which lead to a certain process. Also since the collisions happen at relativistic speeds, nominal measurements of particle properties would be altered due to presence of Lorentz boost. Therefore physical quantities which are invariant under Lorentz transformation are used. Cylindrical coordinates are used for measuring position of objects in the detector. While ϕ is the azimuthal angle that determines the angle in positive x-axis along the x-y plane, the polar angle (θ) is the angle between the positive z-axis and the direction of

the particle momentum. In practice, θ is not a Lorentz invariant physical quantity, so instead of θ we use a quantity called pseudorapidity (η) which is defined as follows

$$\eta = -ln\left[tan\left(\frac{\theta}{2}\right)\right] \tag{2.6}$$

where θ is the polar angle mentioned above. The conversion between η and θ is shown in Fig. 2.6. For massless particles and in the limit of momentum much greater than the particle mass, η is equivalent to the physical quantity called rapidity (y) whose differences, Δy are invariant under Lorentz boosts along the z axis. Rapidity of a particle is defined as

$$y = \frac{1}{2}ln\left(\frac{E+p_z}{E-p_z}\right) \tag{2.7}$$

where E is the energy of the particle and p_z is the z component of its momentum. Nevertheless, η is more commonly used since it only depends on the polar angle, θ.

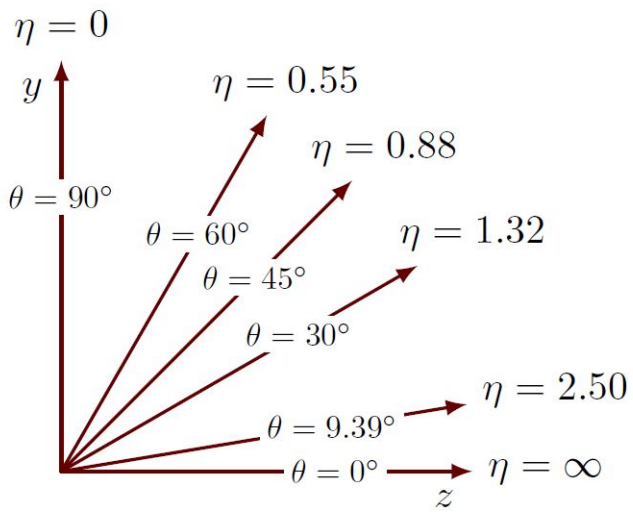


Figure 2.6: Sketch showing the relationship between pseudorapidity η and the polar angle θ.

With a height of 15 m and a length of 21 m is rather compact compared to the ATLAS detector and it was designed to detect muons very accurately and features a 3.8 T solenoid magnet. There are many parts of the detector which are put in place at different locations starting from the point of collision to measure different kinds of properties of particles coming from the hadron collisions in LHC.

2.2.2 Tracker

The tracker of the CMS detector is a cylindrical, full silicon based system with an outer radius of 1.20 m and a length of 5.6 m. It is designed to be extremely granular in order to offers separation of closely-spaced particle trajectories in jets. The barrel (endcaps) comprises of four (three) layers of pixel detectors and surrounded by ten (twelve) layers of micro-strip detectors. Around 16,500 silicon sensor modules are finely segmented into 124 million pixels [30] of dimensions 150 $\times$ 100 μm and 9.6 million strips of pitch of 80 $\times$ 180 μm. The pixel modules have a slight overlap to their adjacent modules in the overall arrangement in order to ensure a circular cross-section of the whole tracker system. Fig. 2.7 shows the arrangement of the tracker system in longitudinal cross-section of the CMS detector.

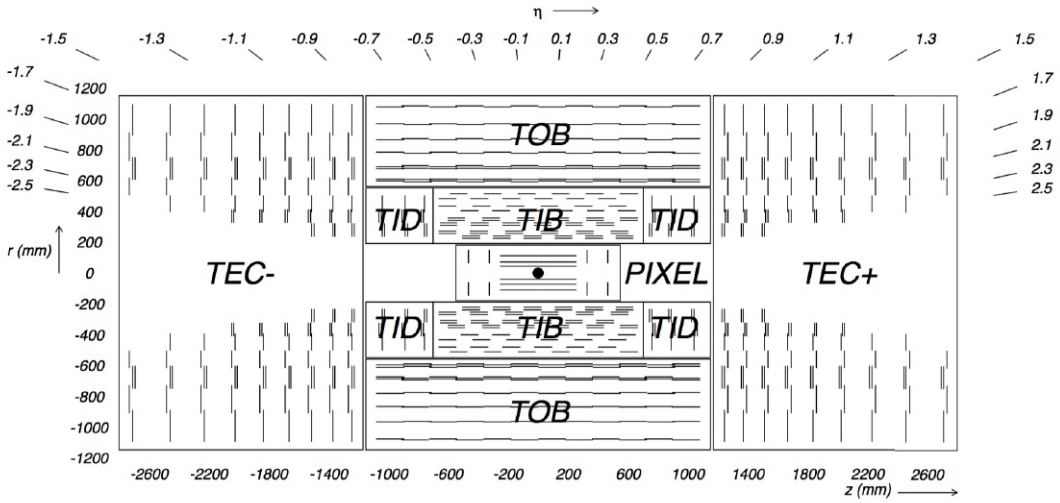


Figure 2.7: A longitudinal section view of the CMS tracker showing the position of the modules and the components. Tracker inner barrel (TIB), tracker outer barrel (TOB), tracker inner discs (TID), and tracker endcaps (TEC) are marked in the relevant position in the figure. Figure taken from [6].

Pixel Detector

The inner tracker is made of thin silicon pixel detectors, placed in 4 cylindrical layers and segmented in both z and ϕ directions to allow for good spatial resolution. The main purpose of the pixel layers is to measure the position of the interaction vertices and to create seeds for tracking. The interaction vertex can be a primary vertex (PV) from a pp collision or a secondary vertex (SV) from the decay of an unstable particle, for example a B hadron. Measuring these secondary vertices is important for the identification of b and c quarks produced in p-p collisions.

Strip Detector

The silicon strip tracker consists of 10 layers of silicon sensors with a total area of 200 m^2. The thickness of the sensors varies from 320 to 500 μm depending on the module position. The strip detectors are placed after the pixel detectors and provide coarser seeds for particle tracks while reducing the number of front-end electronic channels to be processed which is the case with the inner layers of pixel detectors.

2.2.3 Calorimeters

CMS hosts 2 sets of calorimeters whose primary goal is to do energy measurements by producing showers from incoming leptons, photons and hadrons after they have passed the tracker system. Each of these calorimeters work by the principle of scintillation and for targeting different particles different kinds of scintillators are put in place. The electron calorimeter or ECAL, measures energy from electrons and photons while the hadron calorimeter or HCAL does the same for hadrons.

Electromagnetic Calorimeter

Electromagnetic Calorimeter (ECAL) uses lead tungstate ($PbWO_4$) crystals, a material with high density that produces scintillation light in fast, small, well-defined electron and photon showers. Since, the yield of scintillation is low and in order to measure the energy, the scintillation light is are captured by photodetectors, converted to an electrical signal and then amplified. The ECAL barrel covers a pseudo-rapidity range of $|\eta| < 1.5$ and the two endcap disks cover a range of $1.5 < |\eta| < 3.0$. The barrel (endcap) crystal length of 23 (22) cm corresponds to 25.8 (24.7) radiation lengths, which is sufficient to contain more than 98% of the energy of electrons and photons up to 1 TeV. It also amounts to about one interaction length for hadrons, causing about two thirds of the hadrons to start showering in the ECAL before entering the HCAL.
Salient features of the ECAL are:

- Crystals measure $2.2\times2.2\times23$ cm^3 in the barrel and $3\times3\times22$ cm^3 in the endcaps
- There are around 76,000 crystals in the ECAL
- The density of lead tungstate is 8.3 g/cm^3

Hadron Calorimeter

The Hadronic Calorimeter (HCAL), which measures energy by stopping the particles through interactions, is a calorimeter like the ECAL. It is a sampling calorimeter because it is constructed with layers of absorber (brass) and scintillating (plastic tiles) material spaced apart from one another [31].

The HCAL features extensions that allow it to absorb energy leakage from the outer layer, also known as the HCAL barrel and annotated with HO in Fig. 2.8. The HCAL endcap (annotated with HE in Fig. 2.8), similar to the HB, uses brass as absorber and plastic tiles as scintillators; these pieces provide a $|\eta|$ = 5.2 coverage. The HO uses magnet and iron yoke as absorbers and plastic tiles as scintillators. The forward sections, annotated with HF in Fig. 2.8, use iron absorbers and quartz fibers parallel to the beam as scintillators.

The energy resolution of HCAL combined with ECAL for hadrons, from [32], is

$$\frac{\sigma_E}{E} = \frac{100\%}{\sqrt{E(GeV)}} \oplus 5\% \tag{2.8}$$

where the $\oplus$ symbol means that the uncertainties to be added in quadrature.

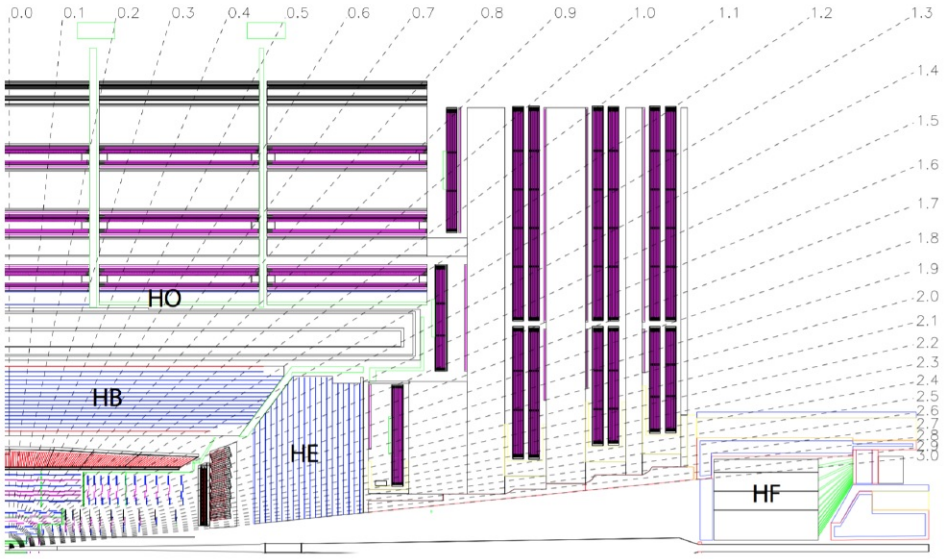


Figure 2.8: Longitudinal cross-section of the CMS HCAL showing the four components: HCAL Barrel (HB), HCAL Outer (HO), HCAL Endcap (HE) and HCAL Forward (HF). Figure taken from [6].

2.2.4 The CMS Magnet

The central feature of the CMS experiment design is a large superconducting solenoid magnet. It delivers an axial and uniform magnetic field of 3.8 T over a length of 12.5 m and a free-bore radius of 3.15 m. This radius is large enough to accommodate the tracker and the calorimeters, thereby minimizing the amount of material in front of the calorimeters. This feature is an advantage for reconstruction of particles, as it eliminates the energy losses before the calorimeters caused by particles showering in the coil material and facilitates the link between tracks and calorimeter clusters. At normal incidence, the bending power of 3.8 T magnet to the inner surface of the calorimeter system provides good separation between charged and neutral particle energy deposits.

2.2.5 Muon Detector

Outside the solenoid coil, the magnetic flux is returned through a yoke consisting of three layers of steel interleaved with four muon detector planes. The amount of absorbing material before the first muon station reduces the contribution of punch-through particles to about 5% of all muons reaching the first station and to about 0.2% of all muons reaching further muon stations.

The goal of the muon chambers is to record the passage of muons through the detector. The recorded hits are combined with the information from the Tracker and used to precisely reconstruct muon tracks. Even though it is located outside the solenoid, the strong return magnetic field in the iron yokes curves the muons and helps the determination of their momenta. Fig. 2.9 shows the overall placement of all the sub-detectors in the muon system. The whole muon subsystem can be divided broadly into 4 sub-detectors:

- **Drift Tubes (DT) -** Drift Tubes are placed in the central part of the CMS detector primarily to capture muon position in the central part. DT cover a pseudo-rapidity region of $|\eta| < 1.2$. The whole DT system consists of rectangular drift cells of transverse size of $13 \times 42\ m^2$ while being 2 to 4 m long. Each drift cell consist of a gas chamber and a centrally passing anode wire which is charged with high voltage. When a muon passes through the gas, it ionizes the gas releasing electrons. These electrons then drift towards the determines the location of the passing muon in the DT. Four layers of DT are arranged in orthogonal orientation to each other in order to measure muon position across orthogonal planes. Two layers measure the $r-\phi$ coordinates while 2 layers measure the $r-z$ coordinates. Each such arrangement consisting of four layers of DT forms a superlayer.

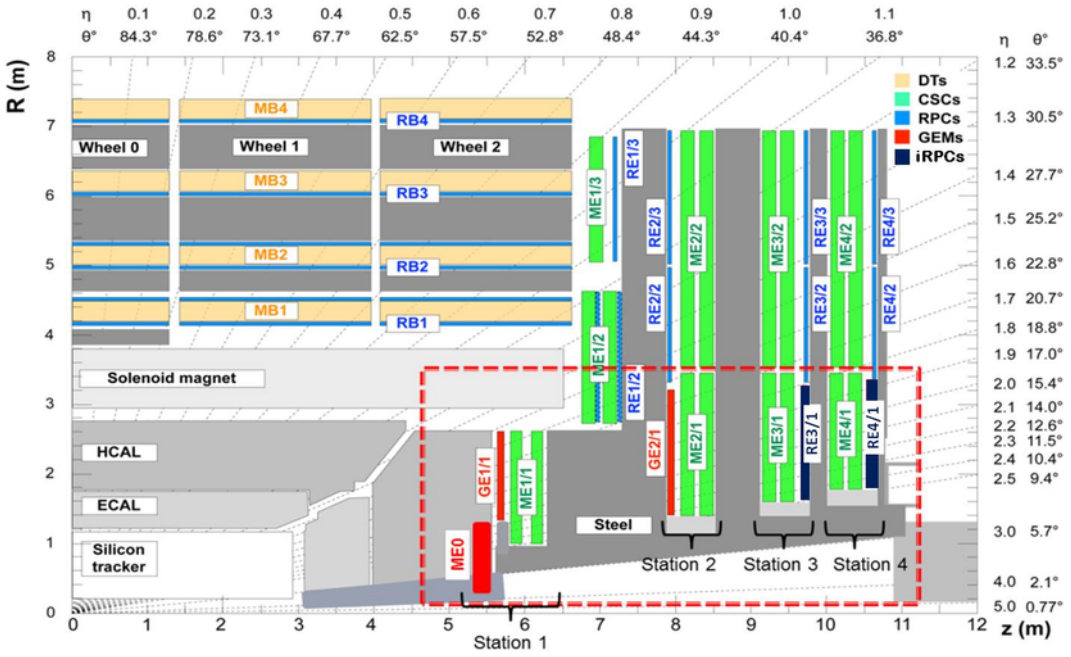


Figure 2.9: Placement of all the muon detectors in the $r-z$ quadrant of the CMS detector highlighting the four CMS muon subdetectors: DT in yellow, CSC in green, RPC in blue and the two newly placed GEM chambers in red. Barrel Wheel 0 and two Wheels at positive z axes are shown as well as the separation into rings in the endcap. Figure taken from [7].

Twelve superlayers cover the whole ϕ region in each wheel of the muon system. Fig. 2.10 shows a schematic diagram of one of the DT chambers in CMS experiment.

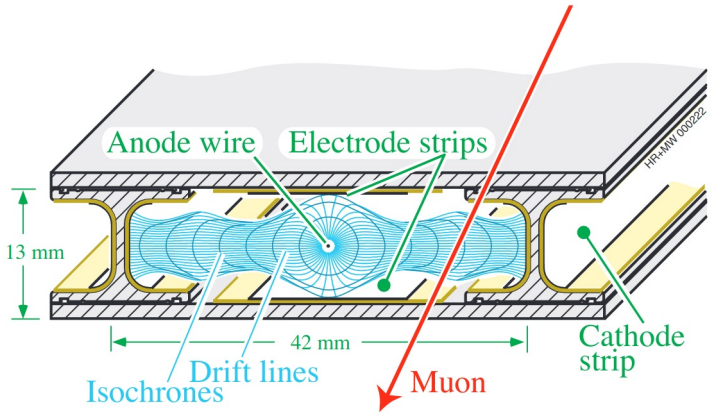


Figure 2.10: Mechanism inside the Drift Tubes for measuring muon position. Figure taken from [8].

- **Cathode Strip Chambers (CSC)** Due to presence of higher and inconsistent electric field on the end caps of the CMS experiment DTs cannot be used there. Therefore an

alternate detector in form of Cathode Strip Chambers (CSC) were used to cover the CMS end-caps. This is because they do not operate in very high magnetic field like the DTs. Each CSC chamber consists of a mixture of Ar (40%), CO2 (50%) and CF4 (10%) gases. They consist of arrays of positively-charged anode wires crossed with negatively-charged copper cathode strips within a gas volume. The directions of the wires and the strips are orthogonal to each other, allowing the measurement of two coordinates. The shorter drift paths of the charge carriers, when compared to DTs, makes them suitable for regions with higher flow of charge particles and strong non-homogeneous magnetic fields. Each endcap has 4 stations containing CSC chambers. A CSC chamber includes 6 CSC layers and the CSC chambers cover the $0.9 < |\eta| < 2.4$ region. The spatial resolution of the CSCs is in the 40 - 150 μm range.

- **Resistive Plate Chamber (RPC) -** Resistive Plate Chambers or RPCs are additional set of muon detectors put in place to improve timing resolution of the muon subsystem. Compared to DTs and CSCs, RPCs have excellent time resolution of 1 ns which makes them ideal for muon matching with corresponding bunch crossings. They are situated both in the barrel and the endcaps along with DTs and CSCs covering a pseudorapidity range $|\eta| < 1.6$ as shown in Fig. 2.9. RPCs consist of two parallel plates, an anode and a cathode, both made of a very high resistivity material and separated by a gas volume. Electrons, created by the ionization of the gas by the passage of a muon, get accelerated and in turn further ionize the gas, causing an avalanche of electrons. The large amount of generated charge induces an image charge on the external metallic readout strips which is readout as the electrical signal. The spatial resolution of the RPCs is in the 0.8–1.2 cm range.

- **Gas Electron Multipliers (GEM) -** GEMs are the newest addition to the CMS Muon system and currently they occupy two disks on the endcap. There are plans to add two more GEM disks during 2024-26 period. GEMs are gaseous detectors like all other Muon detectors and are filled with a mixture of Ar/CO_2. CO_2 in the mixture is responsible of getting ionized when a charge particle passes through it. Inside the gas chambers exists the GEM foil, which consists of a 50 micrometer-thick insulating polymer (polyimide) surrounded on the top and bottom with copper conductors. Throughout the foil, microscopic holes are etched in a regular hexagonal pattern. A potential difference applied across the foils generates sharp electric fields in the holes. The electrons created during the ionisation process drift towards the foils and are multiplied in the holes. The resulting electron avalanche induces a readout signal on the finely spaced strips.

Chapter 3

Event Simulation and Reconstruction

3.1 Trigger in CMS experiment

The LHC delivers a collision rate of 40 MHz to the CMS experiment. However, it is not feasible to record all the collisions given their high rate, and overwhelming volume of generated data. To provide a perspective, an average size of an event is roughly 1 MB so if all events were stored, that would amount to 40 TB of data per second which is impossible to store with the currently available technology. Secondly, most of the events are not really interesting for the physics goals set by the CMS as they originate from well-known processes. As shown in Fig. 3.1, interesting SM processes such as the pair production of top quarks, or the production of the Higgs boson have cross sections several orders of magnitude below the inclusive pp cross section. Therefore CMS has a 2-tiered trigger system in place that provides efficient and reliable way to select relevant data and discard the rest. The first level (L1), composed of custom hardware processors, uses information from the calorimeters and muon detectors to select events at a rate of around 100 kHz within a fixed latency of about 4 μs [33]. The second level, known as the high-level trigger (HLT), consists of a farm of processors running a version of the full event reconstruction software optimized for fast processing, and reduces the event rate to around 1 kHz before data storage [34]. The event reconstruction takes roughly 40 seconds in CMS where the L1 step takes roughly 4 ms and HLT around 300 ms. These time scales rely on incorporation of Field Programmable Gate Arrays (FPGA), Application Specific Integrated Circuits (ASIC) and high performance computing solutions chosen by the CMS collaboration. The collected raw data is further processed to reconstruct the physics objects from detector information. the next sections discuss the physics objects.

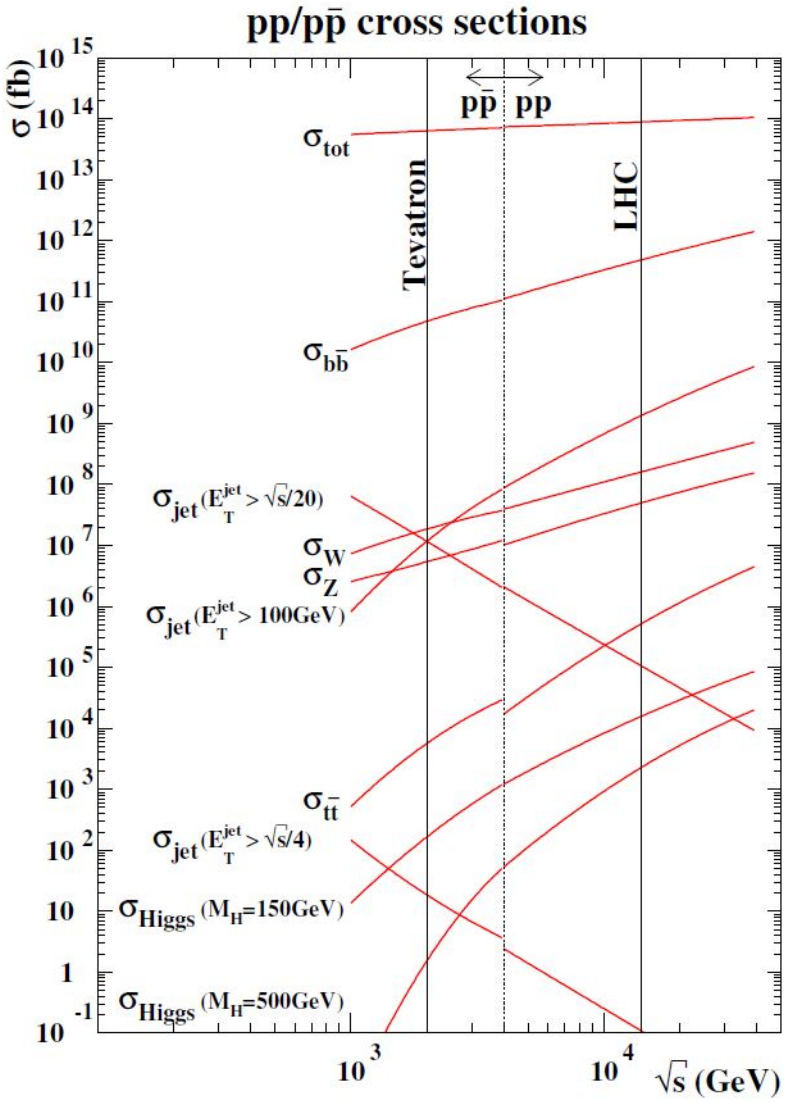


Figure 3.1: SM cross sections at hadron colliders as a function of the center of mass energy, $\sqrt{s}$, for several processes.

3.2 Event simulation

Computer-generated simulations of the necessary physics processes (generator level) and the detector behavior in reaction to the particles formed by these processes (reconstruction-level) are needed to examine the data produced by collisions in the CMS detector. A common approach entails comparing the simulated data with collision data in order to quantify the characteristics of recognized SM processes or to search for novel physics by looking for differences between the two sets of data. Multiple levels of complexity must be taken into account in the

proton-proton collision simulation.

Fig 3.2 demonstrates the approach used by general purpose Monte Carlo generators used in particle physics. The red circle in the center of the graph denotes the hard process, in which calculations are performed in perturbative regime, it is the result of the collision of the hardest momentum partons (constituents of hadrons, here the incident proton) that carry a fraction of the momentum of the proton. The behavior of these fractions are determined experimentally, and they are formulated as Parton Distribution Functions (PDF). More information about PDFs is described in section 3.2.1. The red branched out structures, including the red circle, are radiations and splittings from the hard process. The blue lines represent the Initial and Final State radiations (ISR and FSR), all of which are simulated with parton showers. The final state hadrons are indicated with green circles. The process of hadronization is modeled with phenomenological and effective models as in the confinement regime perturbation methods cannot be applied. Additional Multi-Parton Interactions (MPI) often have small momentum transfers and are simulated with similar models, these are shown as purple circle and lines. The general steps involved in event simulation are shown in Fig. 3.3. It starts with incoming protons for collision. The PDFs provide the probabilities for finding a parton in the proton at specific momenta. The hard interaction between the incoming partons is done by the matrix element calculation, described in Sec. 3.2.2. It determines the type and kinematic properties of the final state particles. The parton-showering step, described in Sec. 3.2.3 deals with the shower evolution of strongly interacting particles which undergo the process of hadronization. Finally, the interaction of stable particles with the detector is simulated to obtain the detector response, briefly described in Sec. 3.2.4.

3.2.1 Parton distribution function

Protons are hadrons which are made up of 3 valence quarks (2 up quark and one down quark) bound by gluons. Therefore collision between 2 protons actually occurs between single constituents of each proton. Any of the constituents (partons) may be involved in the hard scattering of a given p-p collision, be it valence quarks, gluons or short lived quark-antiquark pairs of all flavors called the "sea" quarks. Hence, the colliding partons carry a fraction of proton's total momentum $p = x.P$. Therefore, the energy at the center of mass of the two protons, $\sqrt{s}$, is not the center of mass energy of the interacting partons. The latter varies from collision to collision. The probability of finding a certain parton with the momentum fraction x is given by the parton distribution functions (PDFs). The PDFs not only depend on the type of the parton and the momentum fraction, but also on the energy scale Q^2 the proton is probed at. Therefore, PDFs

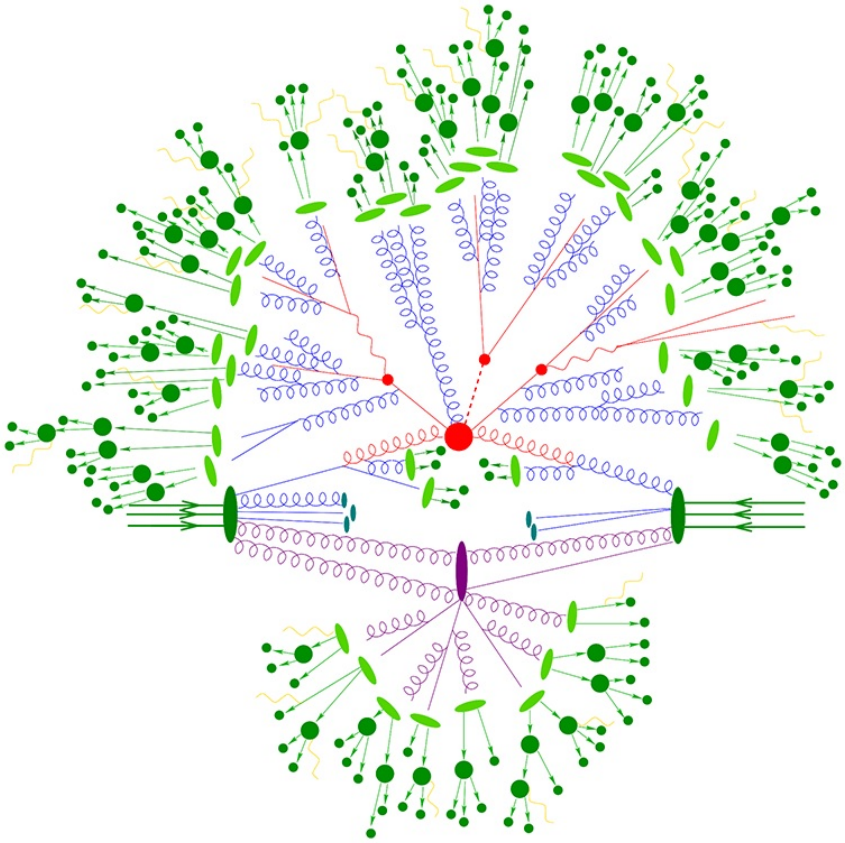


Figure 3.2: Schematic of a proton-proton collision Monte Carlo simulation. Figure taken from [9].

are denoted as $f_i(x, Q^2)$, where i represents the parton flavor (gluon or one of the quarks).

3.2.2 Matrix element

The main characteristics of an event is determined by the process with the greatest momentum transfer, known as the hard process. The calculation of the matrix element in event simulation serves as a representation of the hard process. It has to do with the probability of a process changing from its initial state to its final state. Two arriving, interacting partons determine the initial state, while a variable number of particles determine the final state.
Leading order (LO), the first contribution to the expansion, is comparatively easy to compute and offers a decent approximation of the matrix element, albeit a rough one. The precision of the calculation of the matrix elements is improved by subsequent orders, such as the next-to-leading order (NLO), next-to-next-to-leading order (NNLO), and so forth, but the complexity and quantity of Feynman diagrams significantly increase. Due to these factors, LO matrix ele-

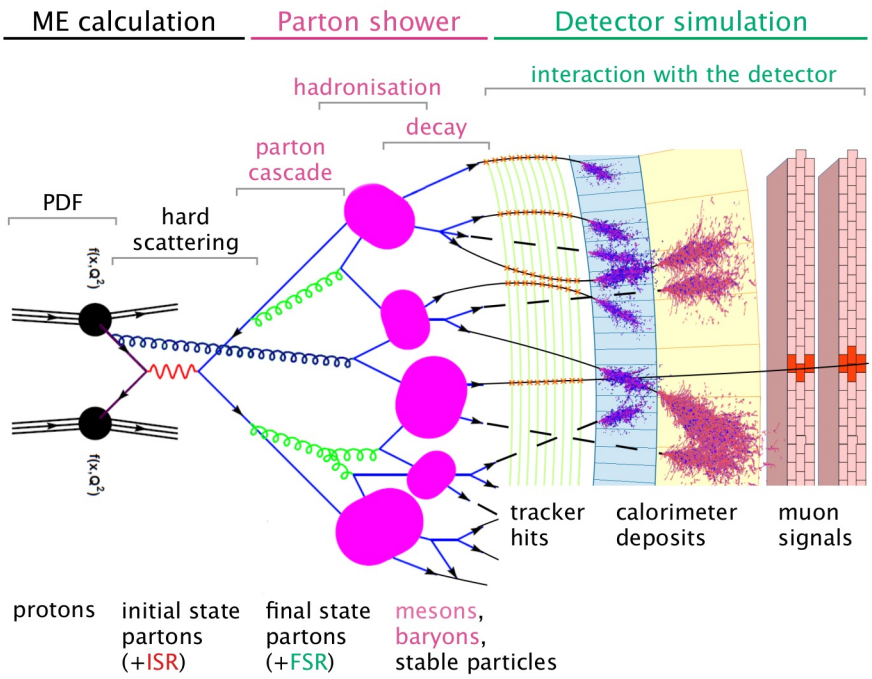


Figure 3.3: Illustration of the key steps of the event simulation procedure. Figure taken from [10].

ments are still commonly employed while NLO and NNLO calculations are only available for a select few processes.

This analysis uses MadGraph [35] and POWHEG [36] event generators to calculate the matrix elements for event generation.

3.2.3 Parton showering and hadronization

Additional partons from p-p collisions are created from ISR and FSRs and can occur in the form of: $q \to qg$, $g \to qq$ and $g \to gg$. A huge number of partons can be produced by further splitting the generated partons in a cascade, as indicated in the Fig. 3.3. Parton showering is the term for this activity. The parton showering approach evolves the event with successive random splittings from the hard process final state, which was obtained by the matrix element. Each succeeding splitting reduces the partons' energy scale until each parton's energy is below a cutoff scale called Λ_{QCD}. At this scale, the effects of color confinement begin to dominate, forcing partons to undergo a process known as hadronization in which they produce color-neutral hadrons. Matching methods must be used to prevent double counting because both the matrix element and parton showering imitate the emission of extra partons. To do this, cutoffs that ensure emissions over a specific momentum or angle are represented by a matrix element

and the remaining portion of the phase space by a parton shower can be included. Hadronization occurs at low energy scales, outside the realm of applicability of perturbative calculations. One example of a Monte Carlo generator based on parton showering is PYTHIA [37]. It is used to simulate many of the Monte Carlo samples for this analysis. The hadronization process in the PYTHIA event generator is based on the Lund string model, which introduces the concept of QCD field lines between quarks, which can be seen as strings holding energy. The string is stretched as the gap between the quarks widens until the potential energy is high enough to produce a quark-antiquark pair out of vacuum.

3.2.4 Detector response simulation

After particles are generated following parton showering and hadronization, their response is simulated in the CMS detector to compare real data recorded from p-p collisions. This is done using GEANT4 (GEometry ANd Tracking) [38] toolkit which is used for simulation of particle response in matter. The simulation includes the geometry of the detector presented in Ch. 2 with the appropriate material for its individual components. The toolkit simulates the interactions with the material based on the cross sections of electromagnetic and hadronic processes. It takes the magnetic field into account when simulating the trajectories of the particles and it also allows the creation of new particles from the interaction with the detector which are also further propagated. The electronic responses of the various detector modules are determined and calibrated to match the observed data.

3.3 Event reconstruction

CMS uses a set of advanced algorithms called particle flow (PF) [39] to reconstruct particles whose properties are measured by different parts of the detector. Fig. 3.4 shows signature of different particles in different parts of the detector. Its purpose is to distinguish and reconstruct particles of different types as physics objects. Each physics object has a specific algorithm designed to reconstruct it as efficiently as possible. It reconstructs the tracks, extrapolates them to the energy deposits in the calorimeters, and links this data along with the identification of muons and electrons. It also reconstructs photons, charged and neutral hadrons and energy of neutrinos in terms of missing transverse energy (MET). The reconstructed particles from PF are further used for reconstruction of jets and other composite objects.

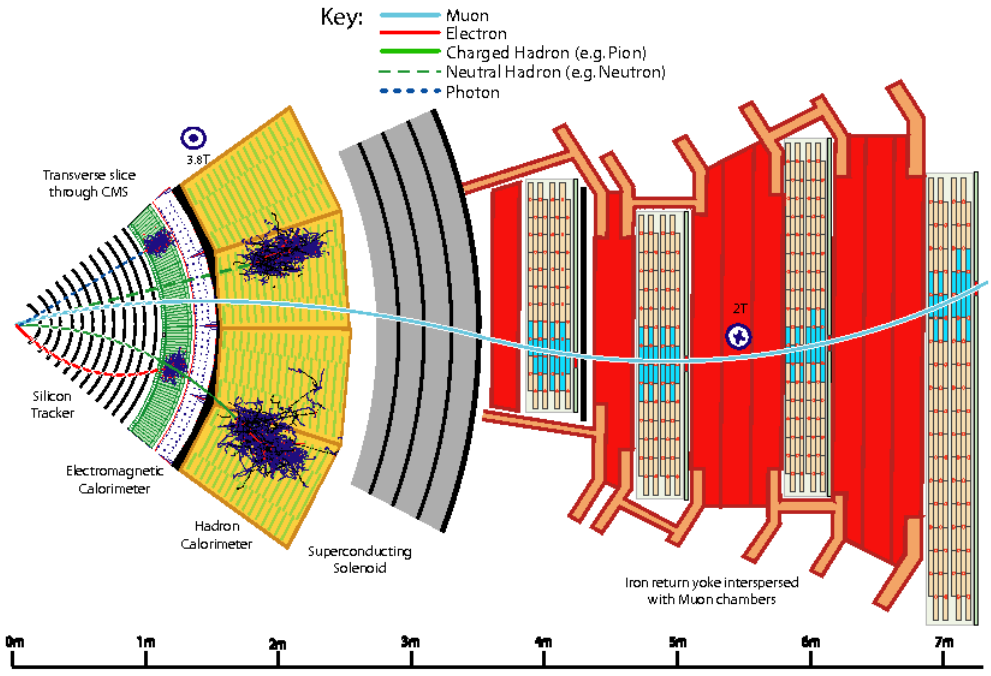


Figure 3.4: Transverse cross-section view of the CMS detector with signatures of different particles in different parts of the detector. Figure taken from [11].

3.3.1 Charged-particle tracks and vertices

Initially, charged-particle track reconstruction was used to tag b quark jets, measure the momentum of energetic and isolated muons, and detect energetic and isolated hadronic τ decays. Thus, tracking was restricted to well-measured tracks and focused exclusively on energetic particles. Since the magnetic field inside the tracker is very homogeneous, a track will travel in a helix-shaped path. The parameters used to describe a track are:

- d_0, z_0: transverse and longitudinal impact parameters
- ϕ: azimuthal angle
- θ: polar angle
- p_T: transverse momentum of the track based on curvature of track in magnetic field

where d_0 is given by $d_0 = -y_0 cos\phi + x_0 sin\phi$. Here x_0, y_0 and z_0 are 3D coordinates of the point of closest approach (PCA) to the center of the detector. An iterative procedure involving 4 steps is carried out while reconstructing tracks and are listed below:

- **track seeding**: find triplets/doublets of 3D hits in the pixel detector and add constraint for beam-spot position. This allows the estimation the 5 parameters needed to describe the helix-shaped track.
- **track finding**: extrapolation of seeds in the outward (and then later inwards) direction using a Kalman filter and association of new hits along the track
- **track fitting**: re-fitting of all found hits with Kalman filter and smoother applied including more sophisticated method to take inhomogeneous magnetic field into account and to remove a possible bias from the beam-spot constraint
- **track selection**: rejection of fake tracks with various quality requirements, such as number of layers with hits or χ^2/ndf of the fit

Each iteration follows removal of hits associated to already reconstructed tracks. This is done to reduce complexity in reconstructing difficult (e.g. displaced) tracks in subsequent iterations.

Primary vertices

Primary vertex (PV) refers to the position of hard collision along the beam axis. This is the point where partons from the collision hadronize and form jets. Therefore, PV is also associated to jets in a collision event.
It is reconstructed in a 3 step procedure:

- **selection** of tracks matching the beam spot
- **clustering** of similar tracks in z using simulated annealing algorithm
- **fitting** the position of vertex from tracks clustered in the previous step

The leading PV with the highest sum of squares of transverse momenta is selected as the PV of the jet. Subleading PVs with lower momenta are classified as pileup vertices if they are compatible with the luminous region and have at least 4 tracks associated.

Secondary vertices

The inclusive vertex finder (IVF) algorithm is used for reconstructing secondary vertices (SV). Based on their separation in three dimensions, it clusters tracks around seeds with high impact parameter significance. Then, an outlier-resistant fit of a common vertex of all the tracks in a cluster yields the SV position. Tracks are then re-associated to either the primary or the

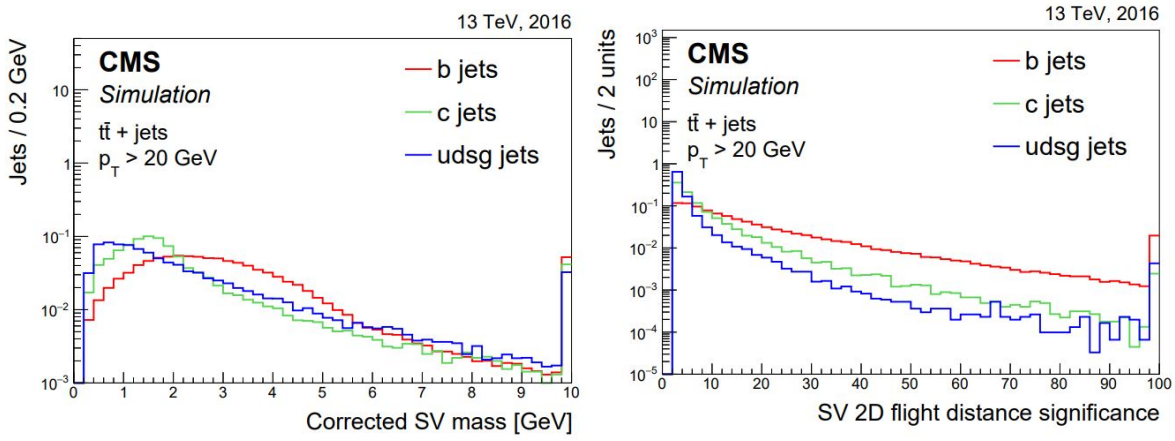


Figure 3.5: Secondary vertex properties (corrected SV mass, flight distance significance) to discriminate between b and light flavor jets. Figure taken from [12].

secondary vertex based on their compatibility and the SV position is fitted again using only the remaining tracks if there are at least two tracks remaining. SVs aren't employed directly in this study, but in the event that they can be rebuilt, they offer significant jet flavor differentiation and are thus used in b-tagging discriminants like DeepCSV. As example two of the SV properties are shown in figure 3.5 for different jet flavors.

3.3.2 Jets

Jets are a spray of stable particles in a conical shape formed from quarks and energetic gluons after they undergo hadronization and fragmentation. Reconstruction of jets from detected particles helps in understanding the properties and nature of its original parton.

Jet clustering

Many different algorithm exists for jet reconstruction, and each algorithm prioritizes some properties of jet over others, especially when multiple jets are close together or partially overlapping. One of the popular and widely implemented type of algorithm is sequential recombination algorithm, which takes list of particles associated to the jet and defines a measure of distance between each of them with reference to the beam axis:

$$d_{ij} = min(k_{ti}^{2p}, k_{tj}^{2p}) \frac{\Delta_{ij}^2}{R^2} \tag{3.1}$$

$$d_{iB} = k_{ti}^{2p} \tag{3.2}$$

where,

$$\Delta_{ij}^2 = (y_i - y_j)^2 + (\phi_i - \phi_j)^2 \tag{3.3}$$

with transverse momenta k_{ti}, rapidity y_i and azimuth angles ϕ_i. Nearby particles are iteratively clustered together until the distance exceeds a threshold. Each iteration involves the following steps,

- Finding the smallest between d_{ij} and d_{iB}
- If d_{ij}, recombine i and j into a pseudoparticle (by adding their 4-momenta)
- If d_{iB}, call i a jet and remove from list of (pseudo)particles
- Repeat step 1 until until no pseudoparticles left

Depending on the value of p from Eq. 3.1 and 3.2 different jet clustering algorithms are chosen which are listed as follow,

- p = 1 $\rightarrow$ k_T algorithm
- p = 0 $\rightarrow$ Cambridge-Aachen algorithm
- p = -1 $\rightarrow$ Anti-k_T algorithm

Jets reconstructed from various algorithms are shown in 3.6.

Anti-k_T algorithm

This analysis uses the anti-k_T algorithm for clustering jets. The anti-k_T algorithm (AK) is obtained by setting p = -1 in the above equation. This algorithm preserves shape of the jets and keeps particles for reconstruction within a desired jet cone radius. This reduces the smearing of the jet momenta from hard scattering when pile up events are added. In this analysis, we select jet with a cone radius of 0.4 for 2 separate b-jet candidates (AK4 b-jets).

Jet flavor tagging

Identifying the flavor of a jet's originating parton is crucial for the VHbb analysis since it allows for the removal of a significant portion of background data by requiring the presence of jets that originate from b-quarks induced hadrons. Each jet is given a score using a b-tagging algorithm that indicates how likely it is that the jet originated from a b-quark rather than a light quark. Modern taggers are implemented as multi-classifiers that may differentiate between light, c and b-quarks, as well as occasionally gluons or different decay channels of b-hadrons. The

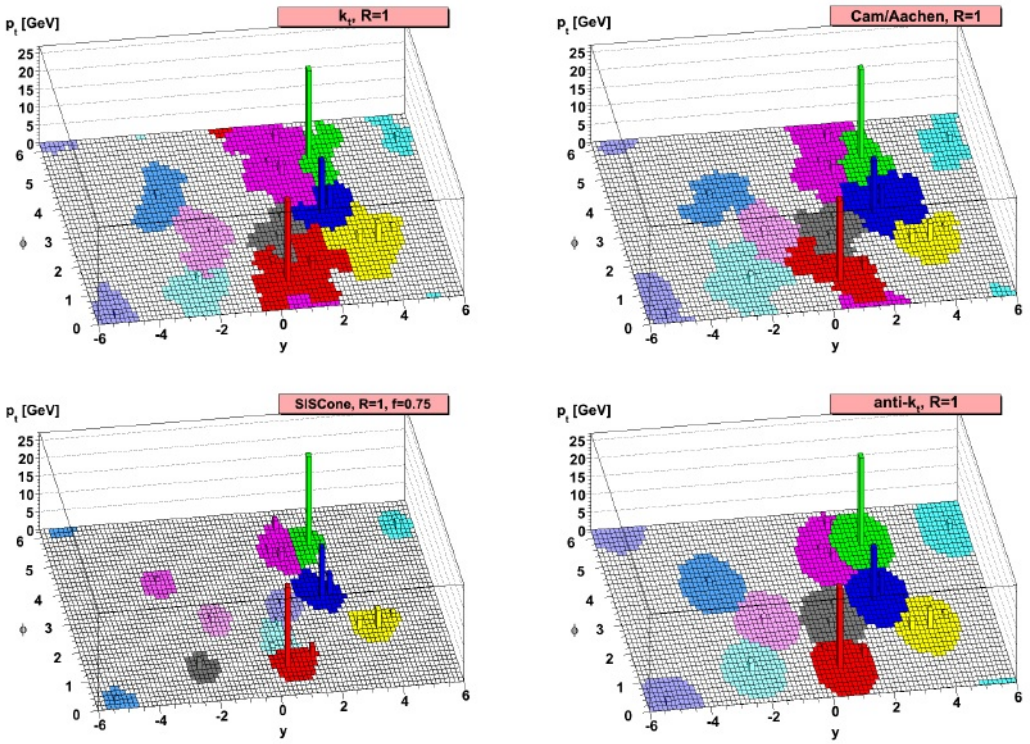


Figure 3.6: Jet reconstructed using various jet clustering algorithms with jet cone radius, R = 1. This analysis uses jets clustered using anti-k_T algorithm.

DeepCSV technique [40], which is based on SV and track data, is used in this research and uses a deep neural network (DNN). It performs a multi-class classification and outputs the scores (probabilities) of 5 different classes:

- b: one b hadron
- bb: at least two b hadrons
- c: one c hadron, no b hadrons
- cc: at least two c hadrons, no b hadron
- light: no c and no b hadrons

A b-tagging score between 0 and 1 is created by combining the probabilities for the b and bb class. Working points are cut positions with the given light and c flavoured jet contamination on this score (mis-tagged jets). Working points that are tight, medium, and loose have mis-tag rates of 0, 1, and 10%, respectively.Figure 3.7 shows a comparison in performance of DeepCSV algorithm with its predecessor b-tagging algorithms for the Phase 1 of the detector.

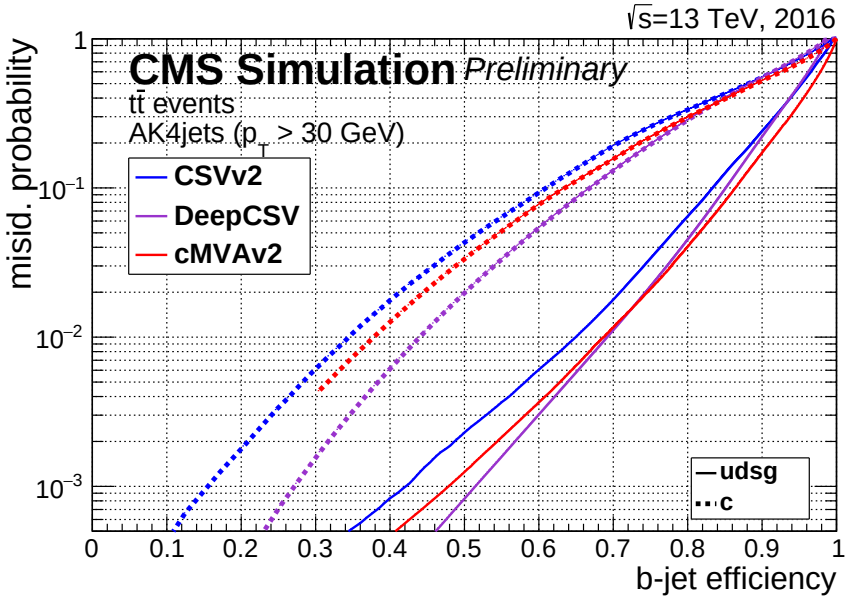


Figure 3.7: Performance of the DeepCSV b-tagging algorithm compared to its predecessor b-tagging algorithms. The curves are obtained on simulated $t\bar{t}$ events using jets within tracker acceptance with $p_T > 30$ GeV, b jets from gluon splitting to a pair of b quarks are considered as b jets. Figure taken from [13].

For Lorentz boosted part of the analysis DeepAK8 algorithm [14] is used for characterizing fat b-jets from other jets originating from a charm or lighter quarks/gluons. It is also a multiclass DNN classifier based algorithm trained on AK8 jets to distinguish Lorentz boosted resonances. It divides AK8 jets into W/Z/H/T and other categories, with subclasses for the various decay channels (e.g. H to bb). The DNN architecture uses up to 100 particles per jet, each of which has 42 attributes totaling essential kinematic properties including p_T, charge, energy deposits, and angular observables. Additionally, up to 7 SVs worth of secondary vertex features are utilized. The network design consists of a fully connected layer after 14 (10) 1D CNN layers for the particle (SV) features as shown in Fig. 3.8. The jet mass is a very effective discriminating variable between the various categories. Even if the jet mass is not explicitly stated, the network may nonetheless calculate it from the low level input properties, shaping the jet mass distribution for the various output categories. A mass decorrelated variant is trained using an adversarial method to reduce this sculpting as shown in Fig. 3.9. This analysis uses the Hbb output node of the mass-decorrelated DeepAK8 tagger (bb-score). In contrast to the DeepCSV tagger, no calibration of the simulation to match data for the full range from 0 to 1 is available, the tagger shape is reduced to 3 bins: 0-0.8, 0.8-0.97, 0.97-1. Scale factors for the signal, computed from gluon splitting to bb, with a transfer function applied to match them to H$b\bar{b}$ are available for

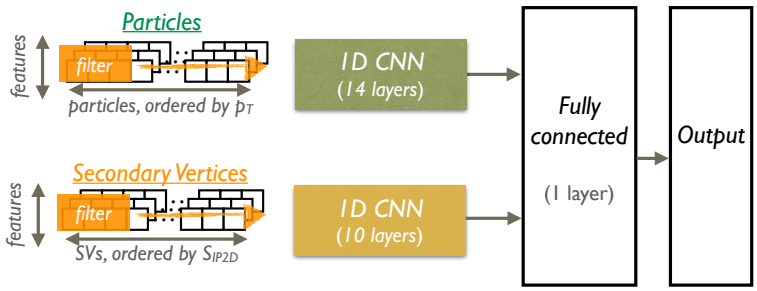


Figure 3.8: The network architecture of the DeepAK8 algorithm. Figure taken from [14].

the 0.8-0.97 and 0.97-1 bins. Scale factors for the backgrounds are measured in-situ and imple-

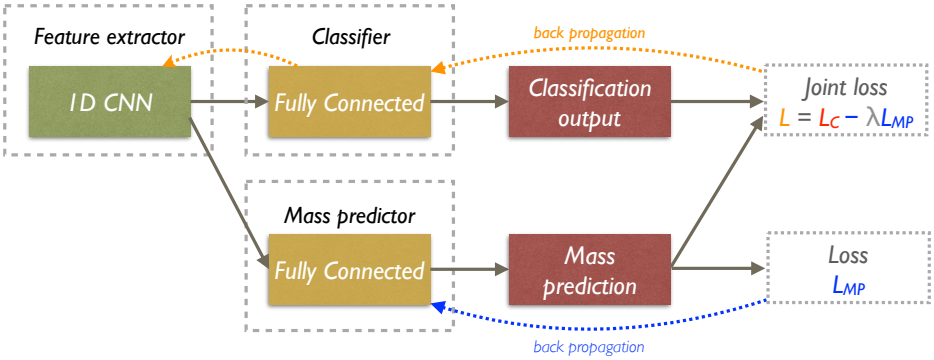


Figure 3.9: The network architecture of mass decorelated DeepAK8 algorithm. Figure taken from [14].

mented as free floating normalization parameters applied on top of the process normalization parameters that are global in the final fit. The four parameters for V+light jets, V+c jets, V+b jets and TT are constrained by the difference in the data to MC ratio in the boosted control regions.

3.3.3 Isolated leptons

The relative isolation to select clean and prompt leptons is defined as:

$$I_{PF,rel.} \equiv \frac{1}{p_T^l}\left(\sum p_T^{charged} + max\left[0, \sum p_T^{neutral} + \sum p_T^{\gamma} - p_T^{PU}(l)\right]\right) \tag{3.4}$$

Here p_T terms corresponds to p_T of various PF objects. p_T^l corresponds to p_T of leptons, $p_T^{charged}$ and $p_T^{neutral}$ refers to p_T of charged and neutral hadrons respectively. p_T^{γ} refers to p_T of photons and $p_T^{PU}(l)$ corresponds to p_T of hadrons originating from pileup.

Isolated electrons

Electrons compatible with the PV are reconstructed using the Gaussian Sum Filter (GSF electrons) algorithm (d_{xy} 0.05 cm, d_z 0.2 cm). To lessen fakes, an MVA discriminator is applied to electrons. For this MVA, two working points — loose (90% efficiency, WP90), and tight (80% efficiency, WP80) — are defined. Both loose and tight WP are needed for the selection of the two electrons in 2-leptons channel, while a tight WP is used for the MVA to choose electrons in the 1-lepton channel. For the 2 electrons channel, a relative isolation of less than 0.15 is required, and for the 1 electron channel, less than 0.06. On MC simulated samples, electron efficiency corrections scale factors are applied based on the corresponding working point, isolation, and trigger.

Isolated muons

Muons are reconstructed using the hit information from the tracker system and muon chambers (global muons). Compatibility with the primary vertex is required by cuts $d_{xy} < 0.5$ cm and $d_z < 1$ cm. A cut-based identification is applied, which cuts on χ^2 of global tracks, track impact parameter and the number of muon-chamber, tracker and pixel hits. The relative isolation is required below 0.25 for the 2-muon channel and below 0.06 for the 1-muon channel. Efficiency correction scale factors derived for the respective working point and trigger are applied to the MC simulation.

3.3.4 Missing transverse energy

Missing transverse energy (MET) refers to the amount of energy, which has not been detected in the detector but is expected from energy and momentum conservation. This energy accounts for the energy of the neutrinos in an event, since there is no direct way to measure properties of neutrinos in our detector. MET is particularly relevant in case of 0-lepton channel where the Z boson decays into pair of neutrinos. Apart from accounting energy of neutrino, MET can also be produced in following scenarios:

- limited detector acceptance or efficiency
- detector malfunction or mis-reconstruction
- cosmic rays or beam halo particles

Since the longitudinal momentum of the two colliding partons is not known, the missing energy is only reconstructed in the transverse plane, where the total momentum is expected to be zero

for perfect reconstruction of all particles. MET is then calculated as follows:

$$MET = -\left| \sum_{particles} \overrightarrow{p}_T \right| \tag{3.5}$$

This analysis uses two approaches to reconstruct MET:

- (raw) PF MET : full particle-flow reconstruction, so

$$PF\,MET = -\left| \sum_{PF\,candidates} \overrightarrow{p}_T \right| \tag{3.6}$$

- Type 1 PF MET : like raw PF MET but with jet energy corrections applied given by:

$$Type\,1\,PF\,MET = -\left(\sum_{jets} \overrightarrow{p}_T^{\,corr} + \sum_{e,\mu} \overrightarrow{p}_T + \sum_{unclusterd\,PF\,candidates} \overrightarrow{p}_T \right) \tag{3.7}$$

Chapter 4

Analysis Strategy

4.1 Motivation

In chapter 1, we discussed various decays modes of the Higgs boson and advantages of VH, $H \to b\bar{b}$ production mode of Higgs boson in $H \to b\bar{b}$ decay mode. The evidence of $H \to b\bar{b}$ process was first seen after analyzing data recorded by CMS experiment in 2016 at $\sqrt{s}$= 13 TeV and data collected during Run 1 (2009-2012) at $\sqrt{s}$ of 7 & 8 TeV with an observed (expected) significance of 3.8σ (3.8σ). The first evidence of VH, $H \to b\bar{b}$ process was found in 2018 after analyzing LHC data recorded in 2017 at $\sqrt{s}$ of 13 TeV with an observed (expected) significance of 3.3σ (3.1σ). Combined with all other previous (2016 and Run 1) VH, $H \to b\bar{b}$ searches, the observed (expected) significance was 4.8σ (4.9σ). A combination of all Higgs boson production channels along with VH channel led to discovery of $H \to b\bar{b}$ process at CMS experiment with an observed (expected) significance of 5.6σ (5.5σ) in 2018.

This analysis is performed on the full Run 2 data recorded by the CMS experiment from the period of 2016 to 2018. The recorded data was obtained from p-p collisions at $\sqrt{s}$ of 13 TeV. A total integrated luminosity of 138 fb^{-1} worth of collisions was recorded by the CMS experiment during this period. Due to an overwhelming amount of data collected in the last years by the LHC, the analysis aims at measuring VH, $H \to b\bar{b}$ cross-sections in STXS framework [21] which will be discussed in upcoming sections.

4.2 General strategy

The general analysis strategy is to determine a signal strength modifier (μ) from observed data by doing simultaneous fit in all signal and background regions. Each region is constructed by

applying a set selections on an observable and depending on whether an event passes the set of selections, it is assigned to either the signal region or one of the background control regions. The selections are applied both on data and on Monte Carlo (MC) simulated processes. An observable with appropriate binning is chosen in each region and after filling the events, the agreement between data and total Monte Carlo yield (signal + all background processes stacked) is observed bin by bin. This is called as template. The signal region templates are designed to have a higher purity in signal process while various background templates are modeled to obtain purity in a particular background process. Selections for signal and all the background regions are applied to ensure enrichment of that process in the corresponding region and each region is orthogonal in phase space to one another. Systematic uncertainties called nuisance parameters are accounted in the fit to correct either shape of the MC or overall yield to match data. Additional scale factors for most dominant backgrounds are introduced in the fit that are free-floating (unconstrained) in the fit and are used for the normalization of the respective backgrounds. This means the normalization of those backgrounds is measured from data, which reduces the dependency on the simulation.

The value of signal strength modifier (μ) obtained after doing the fit (post-fit) is the most-important result of this analysis and it denotes the ratio of observed (data) and standard model predicted (MC simulation) cross-sections of the signal. In this analysis, we fit exclusive μ's in different P_T region of the Z or $W^{\pm}$ boson produced in association with the Higgs boson.

To obtain a low uncertainty on the post-fit μ, a good separation of the signal and background in the signal region is needed. This is achieved by using the output of a multi-variate discriminator based on neural networks, as observable for the signal templates.

4.2.1 Treatment of Lorentz boosted phase space

When a Higgs boson is highly Lorentz boosted (with high p_T) the angle of separation between the two b-jet candidates becomes small and a sufficiently boosted Higgs boson can decay into a pair of b-jets which are highly collimated leading into a single jet with large cone radius. The angular separation between two b jets is inversely proportional to p_T of the Higgs boson and is given by equation 4.1.

$$\Delta R_{b,\bar{b}} \approx \frac{2}{\gamma} = \frac{2m_H}{p_T} \tag{4.1}$$

Following this equation, it can be seen that $\Delta R_{b,\bar{b}}$ is ≈ 1 at p_T= 250 GeV and further reduces to 0.8 and lower with Higgs boson p_T > 310 GeV. Since nominally jets are clustered with a

cone radius of 0.4, obtaining $\Delta R_{b,\bar{b}} < 0.8$ makes it difficult to reconstruct individual jets with a great resolution. Therefore for the highest p_T(V) category (> 250 GeV) a dedicated boosted jet reconstruction with a larger radius of 0.8 is performed, which can be used to reconstruct the Higgs boson out of a single so called AK8 fat jet instead of two resolved AK4 jets. A dedicated analysis is conducted for this phase space and is referred to as boosted analysis in subsequent sections.

4.3 Signal and background processes

4.3.1 Signal

Vector boson produced along with the Higgs boson, can be a W boson or a Z boson. Quark induced VH process can be both ZH and WH while gluon induced VH process can only be the ZH process. One thing to note is gluon induced Z boson can only occur via quark loops where heavy quarks (e.g. top quark) have higher cross-section in the loop. Following the different decay modes of the vector boson, the analysis is divided into three channels,

- 0-lepton (Znn): Z boson decays to two neutrinos (ν)
- 1-lepton (Wln): W boson decays to lepton and neutrino of same flavor
- 2-lepton (Zll): Z boson decays to two leptons of opposite charge and same flavor

The 1 and 2-lepton channels are further divided into electron and muon channels. Tau leptons are not reconstructed in the analysis because it introduces additional complexity in the final decays. Fig 4.1 shows all the possible leading order Feynman diagrams for VHbb process.

4.3.2 Background

Final states of VHb$\bar{b}$ process involves 2 b-jets and 2 leptons (excluding taus). Based on these final states, following processes have similar final state signature and are accounted to be major backgrounds in this analysis:

- V+jets
- Top quarks ($t\bar{t}$ and single top)
- Pair of vector bosons
- QCD multi-jet events

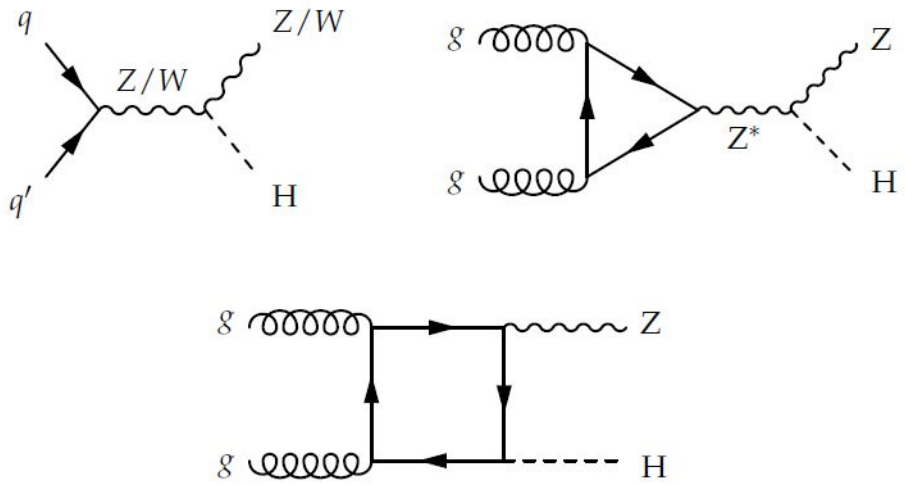


Figure 4.1: The leading order Feynman diagrams corresponding to the VH$b\bar{b}$ signal process. The gluon induced production mode contributes to the zero and two lepton channels (top right and bottom diagrams).

V+jets

Two quarks from proton collisions can produce a vector boson along with gluon spatially opposite to one another. The vector boson decays into a pair of leptons while the gluon decays into a pair of quark-antiquark jets. Relevant Feynman diagram is shown in Fig 4.2. Such process can lead to an exact replication of our signal signature if the pair of quarks from the irradiated gluon are a pair of b-quarks. Hence this process serves as the most important background in this analysis. Kinematically, the jets and the vector boson produced in such a process have a lower p_T and the di-jet invariant mass (m_{jj}) of the 2 jets from from the gluon falls outside the mass window of the signal which is intended to select only Higgs boson candidates. Another way to differentiate V+jet background from signal is to use the score from b-tagging algorithm to identify b-jets from jets originating from lighter quarks. Despite all kinds of selection implemented to obtain signal with higher purity, V+jets or primarily $V + b\bar{b}$ is still an irreducible background in the signal phase phase and is predicted properly via simulation. In the 2-lepton channel, this is the dominating background.

Based on the type of quark jets emerging from the energetic gluon, we categorize V+jets into 3 categories namely, V+b jets (gluon decays in b-quarks), V+c jets (gluon decays into c quarks) and V+light jets (gluon decay into lighter quarks or more gluons) based on counting B- and D-hadrons above 25 GeV within detector acceptance ($|\eta| < 2.6$). If multiple hadrons of different flavor are present, the flavor is defined by the heaviest quark. Full list is mentioned in table 4.1.

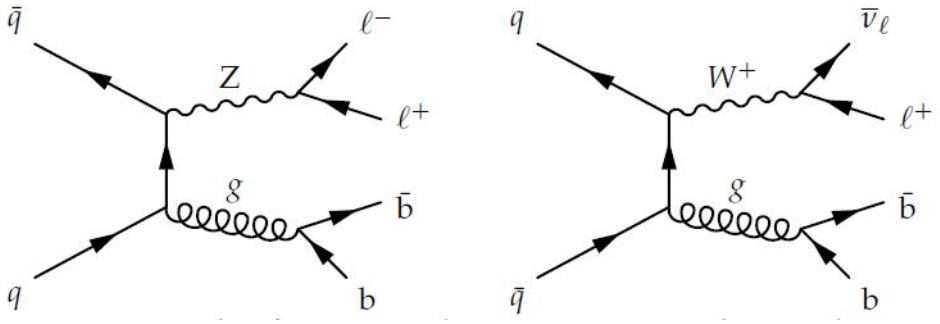


Figure 4.2: An example of Feynman diagrams corresponding to the Z + jets(left) and W+ jets (right) background processes.

Jet flavor	D-hadrons	B-hadrons
V+light	=0	=0
V+c	>0	=0
V+b		>0

Table 4.1: V+jet flavor definitions where B- and D-hadrons must be in detector acceptance, which is defined as: $p_T > 25$ GeV and $|\eta| < 2.6$.

Top quarks

Top quarks contribute in several ways to the background and primarily enriches 0 and 1-lepton channels. They are produced both as a pair of top quarks ($t\bar{t}$) and single top quarks. Single top quarks are produced via electroweak interaction in three channels namely, s-channel, t-channel and associated production with a W quark (tW channel) and the same is accounted in simulation. Fig. 4.4 show the relevant Feynman diagrams of the production channels of single top and fig. 4.3 shows the Feynman diagrams for all possible $t\bar{t}$ productions. In $t\bar{t}$ process, one of the W bosons can decay leptonically and produce a lepton signature similar to WH signal. The b quarks in this case have then an energy of around 65 GeV in the restframe of the t quarks and if the decay products of the second W are outside of the acceptance, the final state looks very similar to WH. If both W bosons decay leptonically, it looks like a ZH signal where the Z decays into two leptons, but with a flat distribution of the invariant mass. For both W bosons decaying hadronically, $t\bar{t}$ events can look like a signal in the 0-lepton channel, but with a higher number of additional jets. A veto on number of additional jets, checking the invariant mass of the 2 leptons which is not compatible to the Z mass and based on the angle between the vector boson and the di-jet system (which is less back-to-back than for signal) and by reconstruction

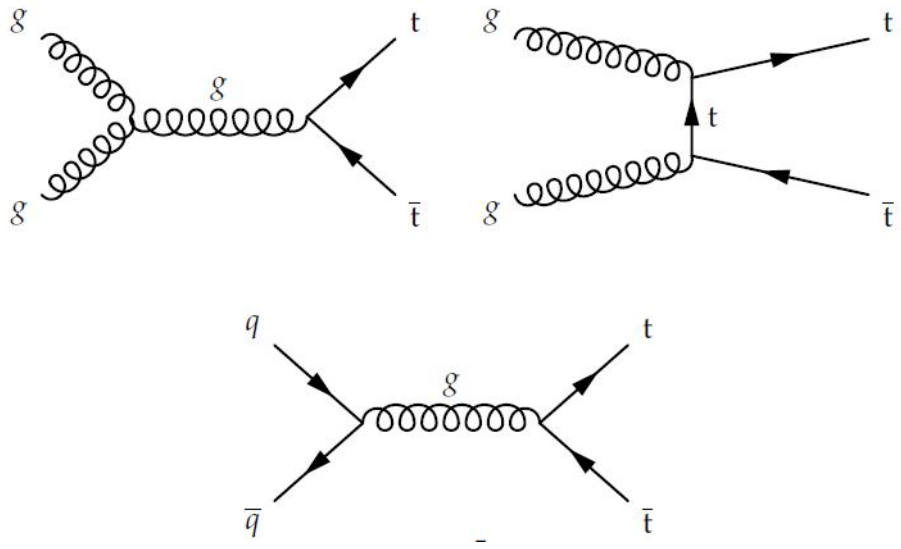


Figure 4.3: Leading order diagrams for $t\bar{t}$ production, the top quark decaying to a W boson and a b-quark, with the W decaying to a lepton and neutrino creates signatures imitating the signal process.

of the top mass in the 1-lepton channel can distinguish $t\bar{t}$ from signal. Single top background

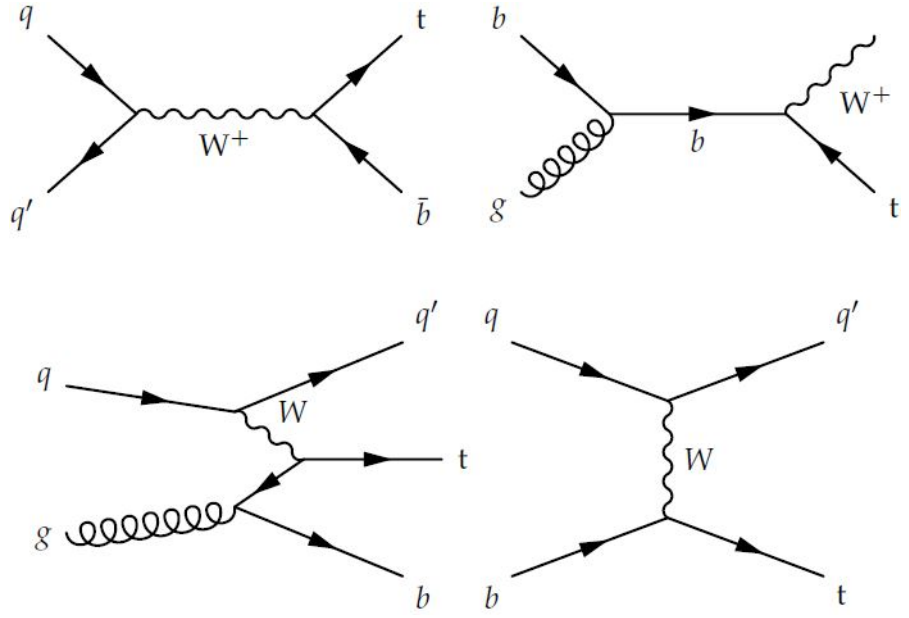


Figure 4.4: Production modes of the single top represented by Feynman diagrams.

is a suppressed background compared to top quark pair process. This is primarily because of coupling of the electroweak interaction compared to the strong coupling in $t\bar{t}$ events produced by QCD. However they are kinematicaly enhanced and are more difficult to distinguish from signal compared to $t\bar{t}$ events as the latter consists of more additional jets in its final state. This

results in a sizable contribution of 10-20% to the total top-induced background.

Dibosons

Diboson or VZ,Z→ $b\bar{b}$ (VZbb) is a very similar process to the VH$b\bar{b}$ signal except it is Z boson which produces pair of b quarks instead of the Higgs boson as is the case in the signal process. Both ZZ$b\bar{b}$ and WZ$b\bar{b}$ contribute to the respective channels based on the lepton final states from the Z or W as is the case in ZHbb or WHbb signal process. We can eliminate this background by looking at the di-jet invariant mass of the b-jets. In case of diboson the invariant mass will peak at Z boson peak which is ≈ 90 GeV. Therefore, setting a suitable mass window for selecting VHbb events reduces contamination from VZbb process in signal region. The leading order production modes are shown in Fig. 4.5. Since VZbb bears much resemblance with VHbb

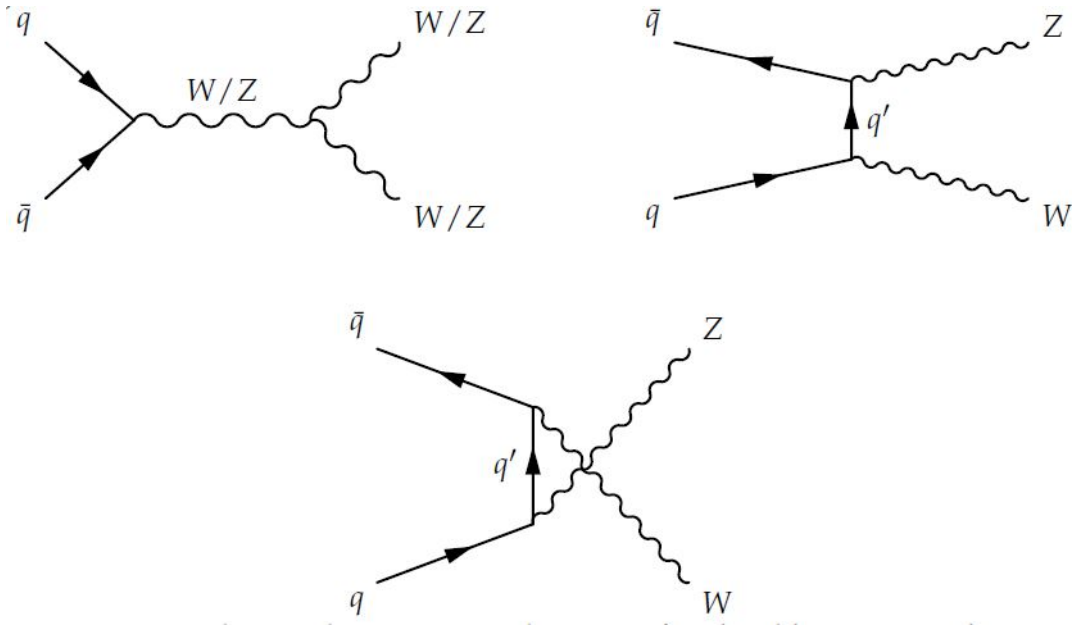


Figure 4.5: Leading order Feynman diagrams for the diboson contributions, s-channel at the top left, t-channel at top right and u-channel at the bottom.

process in terms of kinematics, a cross-check analysis is performed keeping the same analysis strategy and treating VZbb process as signal by shifting the m_{jj} window to target Z boson mass peak instead of Higgs boson. Since VZbb has a higher production cross-section than VHbb process, such a cross check analysis helps with validating the analysis strategy before fitting the VHbb signal strength.

QCD Multi-jet

Strong interactions (QCD) at high energies can form pair of b-jets and has a very high production cross-section. In situations where jet energy is mis-reconstructed, e.g. due to detector imperfections, can create missing energy which is representation of neutrinos. Such a signature, which 2 b-jets and a high missing tranverse energy (MET) can end up in 0-lepton channel. There is difficulty in simulating this background mainly because of very high cross-section. Also our analysis selection is robust enough to reject most QCD events even in 0-lepton channel.Therefore, we do not account for or simulate this process in our background modeling.

4.4 Observed and simulated data

4.4.1 Data trigger selection

The CMS trigger system was introduced in Section 3.1. Multiple HLT triggers are used to select data which corresponds to the signature of the signal processes in their respective channels.
For the zero-lepton channel, the same thresholds for the MET and the MHT ($\not{H}_T$) were applied during the reconstruction at the HLT level to trigger the data acquisition. These thresholds are 110 GeV in 2016 and 120 GeV in the 2017 and 2018 data-taking periods.
For one-lepton channel, a muon p_T threshold of 24 GeV for 2016 and 2018 and a threshold of 27 GeV for the 2017 data-taking period has been applied. Similarly for electrons a threshold of 27 GeV for 2016 and of 32 GeV for 2017 and 2018 on p_T was used.
For the two-lepton channel, the muons have p_T thresholds of 17 and of 8 GeV and the electrons have 23 and 12 GeV thresholds due to the needed coincidence of two leptons needed.
The lepton triggers also require isolation of the leptons and in case of the double muon trigger a minimum invariant mass of the muons. Table 4.2 shows the list of triggers used for selecting events in each channel for the analysis.

4.4.2 Simulated data

Event simulation and simulation of detector response for various physics objects was discussed in section 3.2.4. In this section, the simulated Monte Carlo datasets (referred to as samples) used for signal and background processes with their cross-sections are listed. Refer Table 4.3 for signal processes and Table 4.4 for background process. To reduce the statistical uncertainties on the samples, higher number of simulated events are produced than predicted by the process's production cross-section and expected to appear in the collected data. To retrieve the correct

channel	dataset	L1 seeds (OR)	HLT paths (OR)
Z(νν)H	MET	L1_ETM110 L1_ETMHF120 L1_ETMHF110_HTT60er	HLT_PFMET120_PFMHT120_IDTight HLT_PFMET120_PFMHT120_IDTight_PFHT60
W($\mu\nu$)H	SingleMuon	L1_SingleMu22	HLT_IsoMu27
W(eν)	SingleElectron	L1_SingleEG38 L1_SingleIsoEG30 L1_SingleIsoEG28er2p1 L1_DoubleEG 25 12	HLT_Ele32_WPTight_Gsf_L1DoubleEG
Z($\mu\mu$)	DoubleMuon	L1_DoubleMu 12 5	HLT_Mu17_TrkIsoVVL_Mu8_TrkIsoVVL_DZ_Mass3p8 HLT_Mu17_TrkIsoVVL_Mu8_TrkIsoVVL_DZ_Mass8∗
Z(ee)H	DoubleEG	L1_SingleEG30 L1_SingleIsoEG22er L1_SingleIsoEG24 L1_DoubleEG 15 10	HLT_Ele23_Ele12_CaloIdL_TrackIdL_IsoVL

Table 4.2: Triggers and datasets used for the 2017 data VHbb analysis. ∗ used as replacement trigger for periods where the main trigger was not available.

Sample	σ (pb)	k-factor	Event Generator
$pp \to ggZH; H \to b\bar{b}, Z \to l^+l^-, M_H$ = 125GeV	0.01437	1.0	POWHEG v2
$pp \to ggZH; H \to b\bar{b}, Z \to \nu\bar{\nu}, M_H$ = 125GeV	0.01437	1.0	POWHEG v2
$pp \to ZH; H \to b\bar{b}, Z \to l^+l^-, M_H$ = 125GeV	0.04718	1.0	POWHEG v2 + MiNLO
$pp \to ZH; H \to b\bar{b}, Z \to \nu\bar{\nu}, M_H$ = 125GeV	0.09322	1.0	POWHEG v2 + MiNLO
$pp \to W^-H; H \to b\bar{b}, M_H$ = 125GeV	0.10899	1.0	POWHEG v2 + MiNLO
$pp \to W^+H; H \to b\bar{b}, M_H$ = 125GeV	0.17202	1.0	POWHEG v2 + MiNLO

Table 4.3: Summary of Monte Carlo datasets for signal processes (All hadronized by PYTHIA8), where k-factors are multiplicative factors calculated to correct the leading order (LO) cross-sections to next to leading order (NLO).

number of events, a weight is assigned to each event which is calculated as follows:

$$\mathcal{W}_{MCevent} = \sigma \times \mathcal{L} \times \frac{\mathcal{W}_{generator}}{\sum \mathcal{W}_{generator}} \tag{4.2}$$

where σ is the production cross-section of a process and $\mathcal{L}$ is the integrated luminosity. $\mathcal{W}_{generator}$ is assigned by the Monte Carlo generator for each generated event and are not always constant, in some Monte Carlo generators these weights include negative values e.g. events generated with next-to-leading order (NLO) accuracy.

The signal samples for the quark induced production of ZH and WH are generated by the POWHEG v2 [36, 41, 42] event generator extended with the MiNLO procedure [43, 44] at NLO accuracy. The gluon induced signal samples on the other hand have leading order accuracy and are produced with POWHEG v2 [36, 41, 42], see Table 4.3.

Diboson processes namely ZZ, WZ and WW were produced using MADGRAPH5_aMC@NLO v2.3.3 at NLO generator using the FxFx merging scheme [45]. This generator was also used to generate QCD multijet and the V+jets processes at LO accuracy with the MLM matching scheme [46]. POWHEG v2 is used to generate the $t\bar{t}$ and the single top sample in the t-channel

Sample	σ (pb)	k-factor	Event Generator
$Z^0/\gamma \to l^+l^- + B-Jets, 100 < P_T(Z) < 200GeV$	3.206	1.23	MADGRAPH5_aMC@NLO
$Z^0/\gamma \to l^+l^- + B-Jets, P_T(Z) > 200GeV$	0.3304	1.23	MADGRAPH5_aMC@NLO
$Z^0/\gamma \to l^+l^- + Jets, 100 < P_T(Z) < 200GeV$	2.662	1.23	MADGRAPH5_aMC@NLO
$Z^0/\gamma \to l^+l^- + Jets, P_T(Z) > 200GeV$	0.3949	1.23	MADGRAPH5_aMC@NLO
$Z^0/\gamma \to l^+l^- + Jets, 100 < H_T < 200GeV$	160.8	1.23	MADGRAPH5_aMC@NLO
$Z^0/\gamma \to l^+l^- + Jets, 200 < H_T < 400GeV$	48.63	1.23	MADGRAPH5_aMC@NLO
$Z^0/\gamma \to l^+l^- + Jets, 400 < H_T < 600GeV$	6.982	1.23	MADGRAPH5_aMC@NLO
$Z^0/\gamma \to l^+l^- + Jets, 600 < H_T < 800GeV$	1.756	1.23	MADGRAPH5_aMC@NLO
$Z^0/\gamma \to l^+l^- + Jets, 800 < H_T < 1200GeV$	0.8094	1.23	MADGRAPH5_aMC@NLO
$Z^0/\gamma \to l^+l^- + Jets, 1200 < H_T < 2500GeV$	0.1931	1.23	MADGRAPH5_aMC@NLO
$Z^0/\gamma \to l^+l^- + Jets, H_T > 2500GeV$	0.003513	1.23	MADGRAPH5_aMC@NLO
$Z^0/\gamma \to l^+l^- + Jets, M_{Z^0/\gamma^*} > 50GeV$	5343.0	1.23	MADGRAPH5_aMC@NLO
Multijet - QCD, 200 < H_T < 300GeV	1547000.0	1.0	MADGRAPH5_aMC@NLO
Multijet - QCD, 300 < H_T < 500GeV	322600.0	1.0	MADGRAPH5_aMC@NLO
Multijet - QCD, 500 < H_T < 700GeV	29980.0	1.0	MADGRAPH5_aMC@NLO
Multijet - QCD, 700 < H_T < 1000GeV	6334.0	1.0	MADGRAPH5_aMC@NLO
Multijet - QCD, 1000 < H_T < 1500GeV	1088.0	1.0	MADGRAPH5_aMC@NLO
Multijet - QCD, 1500 < H_T < 2000GeV	99.11	1.0	MADGRAPH5_aMC@NLO
Multijet - QCD, H_T > 2000GeV	20.23	1.0	MADGRAPH5_aMC@NLO
Single Top production (s-channel)	3.74	1.0	POWHEG
Single anti-Top production (t-channel)	80.95	1.0	POWHEG
Single Top production (t-channel)	136.02	1.0	POWHEG
Single anti-Top production (tW-channel inclusive)	35.85	1.0	POWHEG
Single anti-Top production (tW-channel leptonic)	19.56	1.0	POWHEG
Single Top production (tW-channel inclusive)	35.85	1.0	POWHEG
Single Top production (tW-channel leptonic)	19.56	1.0	POWHEG
Hadronic $t\bar{t}$	377.96	1.0	POWHEG
$t\bar{t} \to l\nu$	88.29	1.0	POWHEG
Semi-leptonic $t\bar{t}$	365.34	1.0	POWHEG
$W \to l\nu + B-Jets, 100 < P_T(W) < 200GeV$	5.527	1.21	MADGRAPH5_aMC@NLO
$W \to l\nu + B-Jets, P_T(W) > 200GeV$	0.7996	1.21	MADGRAPH5_aMC@NLO
$W \to l\nu + Jets, 70 < H_T < 100GeV$	1353.0	1.21	MADGRAPH5_aMC@NLO
$W \to l\nu + Jets, 100 < H_T < 200GeV$	1392.0	1.21	MADGRAPH5_aMC@NLO
$W \to l\nu + Jets, 200 < H_T < 400GeV$	410.3	1.21	MADGRAPH5_aMC@NLO
$W \to l\nu + Jets, 400 < H_T < 600GeV$	57.85	1.21	MADGRAPH5_aMC@NLO
$W \to l\nu + Jets, 600 < H_T < 800GeV$	12.95	1.21	MADGRAPH5_aMC@NLO
$W \to l\nu + Jets, 800 < H_T < 1200GeV$	5.45	1.21	MADGRAPH5_aMC@NLO
$W \to l\nu + Jets, 1200 < H_T < 2500GeV$	1.084	1.21	MADGRAPH5_aMC@NLO
$W \to l\nu + Jets, H_T > 2500GeV$	0.008067	1.21	MADGRAPH5_aMC@NLO
$W \to l\nu + Jets, 100 < P_T(W) < 200GeV$	20.49	1.21	MADGRAPH5_aMC@NLO
$W \to l\nu + Jets, P_T(W) > 200GeV$	2.935	1.21	MADGRAPH5_aMC@NLO
$Z \to \nu\bar{\nu} + B-Jets, 100 < P_T(Z) < 200GeV$	6.195	1.23	MADGRAPH5_aMC@NLO
$Z \to \nu\bar{\nu} + B-Jets, P_T(Z) > 200GeV$	0.6293	1.23	MADGRAPH5_aMC@NLO
$Z \to \nu\bar{\nu} + Jets, 100 < P_T(Z) < 200GeV$	1.679	1.23	MADGRAPH5_aMC@NLO
$Z \to \nu\bar{\nu} + Jets, P_T(Z) > 200GeV$	0.2468	1.23	MADGRAPH5_aMC@NLO
$Z \to \nu\bar{\nu} + Jets, 100 < H_T < 200GeV$	303.4	1.23	MADGRAPH5_aMC@NLO
$Z \to \nu\bar{\nu} + Jets, 200 < H_T < 400GeV$	91.71	1.23	MADGRAPH5_aMC@NLO
$Z \to \nu\bar{\nu} + Jets, 400 < H_T < 600GeV$	13.1	1.23	MADGRAPH5_aMC@NLO
$Z \to \nu\bar{\nu} + Jets, 600 < H_T < 800GeV$	3.248	1.23	MADGRAPH5_aMC@NLO
$Z \to \nu\bar{\nu} + Jets, 800 < H_T < 1200GeV$	1.496	1.23	MADGRAPH5_aMC@NLO
$Z \to \nu\bar{\nu} + Jets, 1200 < H_T < 2500GeV$	0.3425	1.23	MADGRAPH5_aMC@NLO
$Z \to \nu\bar{\nu} + Jets, H_T > 2500GeV$	0.005263	1.23	MADGRAPH5_aMC@NLO
WW	117.6	1.0	MADGRAPH5_aMC@NLO
WZ	48.1	1.0	MADGRAPH5_aMC@NLO
ZZ	17.2	1.0	MADGRAPH5_aMC@NLO

Table 4.4: Summary of Monte Carlo Samples for background processes (All hadronized by PYTHIA8), where k-factors are calculated multiplicative factors to correct the leading order (LO) cross-sections to next to leading order (NLO).

while POWHEG v1 is used to generate the single top quark samples in the tW and s-channel.

4.4.3 V+jets MC datasets

The V+jets MC datasets that are used are created in various transverse hadronic momenta H_T bins as well as in vector boson p_T bins for the b-enrichment configurations. Two b-enrichment configurations are used to create the V+jets samples; one increases the number of b quarks in the samples by generating only the matrix-element b-quark decays, and the other produces this enrichment at the level of parton showers by taking into account the showers that contribute to the b-quark final states. The b-enriched MC datasets are used to boost statistics in regions with heavy flavor. Similar generator level phase-spaces are reweighted to match the anticipated SM cross-section in order to prevent double counting in the regions. Fig. 4.6 illustrates this in practice by color coding each V+jets MC dataset. The HT-binned MC datasets are generated in 8 bins (shown in Fig. 4.6 legend starting with "HT"), and the b-enriched MC datasets are generated in 2 p_T(V) bins for each of the two configurations (shown in Fig. 4.6 legend starting with "BGen" referring to the b-enrichment using matrix-element level b-enrichment and "BJet" referring to the parton-shower level b-enrichment).

NLO V+jets MC datasets

Since the V+jets background process makes significant non-reducible contributions to the signal regions, correct modeling of this process is crucial. NLO order MC datasets from Table 4.5 were used to improve the modeling.

For the 2016 data taking period, the leading order MC datasets are reweighted to the NLO accuracy in bins of p_T(V) and number of b-hadrons (nB) in order to enhance the data and MC agreement. The corrections are obtained by fitting polynomial functions to the ratio of $\Delta\eta_{bb}$ (of the two leading b-jets) for the $Z \to \nu\overline{\nu}$ + jets samples and linear functions to the ratio of (NLO/LO) of the p_T(V) for all other V+jet samples. This decision was made in order to use the greater statistics of the leading order MC datasets in comparison to the NLO MC datasets' accessible MC statistics for the 2016 data taking period.

For the data collection periods in 2017 and 2018, the NLO samples were used exclusively. The region with $\Delta R(jj) < 1$ (of the two leading selected jets in p_T) is one of the poorly modeled regions in the NLO samples. For this region, a reweighting is calculated per lepton channel and p_T(V) bin in the light flavor control region and extrapolated to the signal and heavy fla-

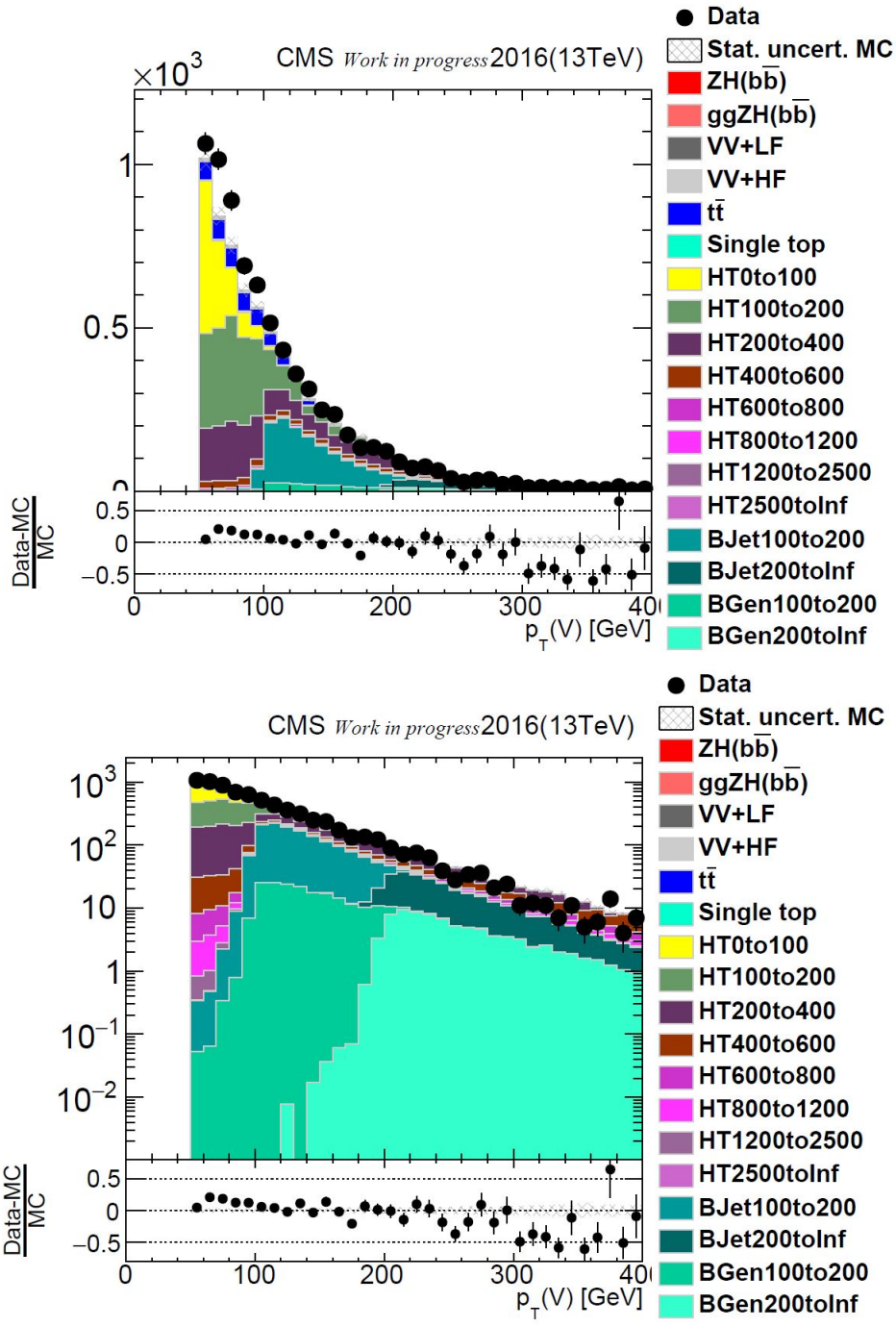


Figure 4.6: The pT(V) in the 2-lepton heavy flavour control region linear (top) and logarithmic (below) histograms for the 2016 data taking period, with V+jets samples shown with separate colors to demonstrate the stitching between different samples.

vor control regions. A two-dimensional reweighting using the DeepCSV scores for the leading and sub-leading jets as the dimensions resolves the other region affected by modeling concerns,

Sample	σ (pb)	k-factor	Event Generator
$Z^0/\gamma \to l^+l^- + 1-Jet, 50 < PT(Z) < 150$ GeV	316.6	1.0	MADGRAPH5_aMC@NLO
$Z^0/\gamma \to l^+l^- + 1-Jet, 150 < PT(Z) < 250$ GeV	9.543	1.0	MADGRAPH5_aMC@NLO
$Z^0/\gamma \to l^+l^- + 1-Jet, 250 < PT(Z) < 400$ GeV	1.098	1.0	MADGRAPH5_aMC@NLO
$Z^0/\gamma \to l^+l^- + 1-Jet, PT(Z) > 400$ GeV	0.1193	1.0	MADGRAPH5_aMC@NLO
$Z^0/\gamma \to l^+l^- + 2-Jets, 50 < PT(Z) < 150$ GeV	169.6	1.0	MADGRAPH5_aMC@NLO
$Z^0/\gamma \to l^+l^- + 2-Jets, 150 < PT(Z) < 250$ GeV	15.65	1.0	MADGRAPH5_aMC@NLO
$Z^0/\gamma \to l^+l^- + 2-Jets, 250 < PT(Z) < 400$ GeV	2.737	1.0	MADGRAPH5_aMC@NLO
$Z^0/\gamma \to l^+l^- + 2-Jets, PT(Z) > 400$ GeV	0.4477	1.0	MADGRAPH5_aMC@NLO
$Z^0/\gamma \to l^+l^- + 0-Jets, inclusive$	5333.0	1.0	MADGRAPH5_aMC@NLO
$Z^0/\gamma \to l^+l^- + 1-Jets, inclusive$	965.0	1.0	MADGRAPH5_aMC@NLO
$Z^0/\gamma \to l^+l^- + 2-Jets, inclusive$	362.0	1.0	MADGRAPH5_aMC@NLO
$Z^0/\gamma \to l^+l^- + Jets, 50 < PT(Z) < 100$ GeV	409.8	1.0	MADGRAPH5_aMC@NLO
$Z^0/\gamma \to l^+l^- + Jets, 100 < PT(Z) < 250$ GeV	97.26	1.0	MADGRAPH5_aMC@NLO
$Z^0/\gamma \to l^+l^- + Jets, 250 < PT(Z) < 400$ GeV	3.764	1.0	MADGRAPH5_aMC@NLO
$Z^0/\gamma \to l^+l^- + Jets, 400 < PT(Z) < 650$ GeV	0.5152	1.0	MADGRAPH5_aMC@NLO
$Z^0/\gamma \to l^+l^- + Jets, PT(Z) > 650$ GeV	0.0483	1.0	MADGRAPH5_aMC@NLO
$W \to l\nu + 0-Jets, inclusive$	54500.0	1.0	MADGRAPH5_aMC@NLO
$W \to l\nu + 1-Jets, inclusive$	8750.0	1.0	MADGRAPH5_aMC@NLO
$W \to l\nu + 2-Jets, inclusive$	3010.0	1.0	MADGRAPH5_aMC@NLO
$W \to l\nu + Jets, 50 < PT(W) < 100$ GeV	3570.0	1.0	MADGRAPH5_aMC@NLO
$W \to l\nu + Jets, 100 < PT(W) < 250$ GeV	770.8	1.0	MADGRAPH5_aMC@NLO
$W \to l\nu + Jets, 250 < PT(W) < 400$ GeV	28.06	1.0	MADGRAPH5_aMC@NLO
$W \to l\nu + Jets, 400 < PT(W) < 650$ GeV	3.591	1.0	MADGRAPH5_aMC@NLO
$W \to l\nu + Jets, PT(W) > 650$ GeV	0.5495	1.0	MADGRAPH5_aMC@NLO
$Z \to \nu\bar{\nu} + 1-Jet, 50 < PT(Z) < 150$ GeV	596.3	1.0	MADGRAPH5_aMC@NLO
$Z \to \nu\bar{\nu} + 1-Jet, 150 < PT(Z) < 250$ GeV	17.98	1.0	MADGRAPH5_aMC@NLO
$Z \to \nu\bar{\nu} + 1-Jet, 250 < PT(Z) < 400$ GeV	2.045	1.0	MADGRAPH5_aMC@NLO
$Z \to \nu\bar{\nu} + 1-Jet, PT(Z) > 400$ GeV	0.2243	1.0	MADGRAPH5_aMC@NLO
$Z \to \nu\bar{\nu} + 2-Jets, 50 < PT(Z) < 150$ GeV	325.7	1.0	MADGRAPH5_aMC@NLO
$Z \to \nu\bar{\nu} + 2-Jets, 150 < PT(Z) < 250$ GeV	29.76	1.0	MADGRAPH5_aMC@NLO
$Z \to \nu\bar{\nu} + 2-Jets, 250 < PT(Z) < 400$ GeV	5.166	1.0	MADGRAPH5_aMC@NLO
$Z \to \nu\bar{\nu} + 2-Jets, PT(Z) > 400$ GeV	0.8457	1.0	MADGRAPH5_aMC@NLO

Table 4.5: Summary of the NLO V+jets Monte Carlo Samples.

which is the region where the DeepCSV score is lower than the loose working point value when $\Delta R(jj) < 1$. Due to the requirement of b-tagging, other regions are unaffected by this correction, which is derived in the light flavor control region's $\Delta R(jj) > 1$ and extrapolated to the $\Delta R(jj) < 1$ light flavor control region. Each time a correction is made, dedicated systematic uncertainties are added, but the effect on the signal strength is seen to be minimal.

4.5 Statistical procedure

The typical strategy for employing the frequentist statistical test to look for novel phenomena begins with formulating the null hypothesis, H_0, which represents only known processes of the SM (called backgrounds in the context of particle physics). This hypothesis will be evaluated against the alternative hypothesis H_1 (containing both backgrounds and the new phenomenon we call signal in this context). The p-value is typically used to quantify how well the observed data fit the provided hypothesis H. It can be converted using the one-sided Gaussian (denoted

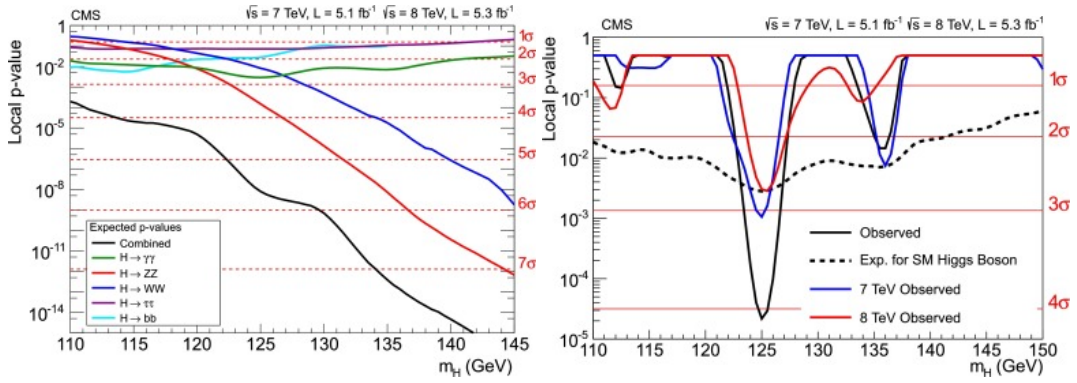


Figure 4.7: Some local p-values and the corresponding significance for the Higgs boson discovery in 2012, expected on the left and observed on the right. Figure taken from [15].

N) tail convention to the number of standard deviations (significance Z). (Eq 4.3).

$$p-value = \int_{Z}^{\inf} N(x;0,1)dx. \tag{4.3}$$

Figure 4.7 demonstrates some local p-values and the corresponding significance for the Higgs boson discovery [15].

4.5.1 Likelihood function and ratio

In this thesis, a Likelihood function is used for fitting observed data with SM expected data in form of MC samples. The likelihood, given a set of measurements n_j with s_j signal, b_j backgrounds and Gaussian distributed nuisances ($\vec{\theta}$) is given as follows:

$$L(\mu,\vec{\theta}) = \prod_j \frac{(\mu s_j + b_j)^{n_j}}{n_j!} e^{-(\mu s_j + b_j)} \prod_k e^{-\frac{1}{2}\theta_k^2}. \tag{4.4}$$

For testing an alternative hypothesis, we use likelihood ratios as test statistic. According to Neyman- Pearson lemma, for a hypothesis test of two simple hypotheses H_0: $(\mu,\vec{\theta}) = (\mu_0,\vec{\theta}_0)$ (null hypothesis) and H_1: $(\mu,\vec{\theta}) = (\mu_1,\vec{\theta}_1)$ (alternative hypothesis) and $\vec{\theta}$ being the nuisance parameters, the ratios of likelihood is defined as,

$$\lambda(x) = \frac{L(\mu_0,\vec{\theta}_0|x)}{L(\mu_1,\vec{\theta}_1|x)}. \tag{4.5}$$

4.5.2 Profile likelihood method

The method used to compute limits and significances at the LHC experiments is based on the profiled likelihood ratio as test statistic. The parameter of interest (POI) in the fit is the signal strength, defined as,

$$\mu = \frac{\sigma}{\sigma_{SM}}. \tag{4.6}$$

Setting $\mu = 0$ corresponds to the background-only model whereas $\mu = 1$ is the SM expectation. The list of parameters $\vec{\alpha}$ is conventionally split into the POI(s) (here only μ) and nuisance parameters $\vec{\theta}$.

$$\vec{\alpha} = (\mu, \vec{\theta}). \tag{4.7}$$

The profile likelihood ratio can then be constructed by

$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\vec{\theta}}})}{L(\hat{\mu}, \hat{\vec{\theta}})}. \tag{4.8}$$

where $\hat{\mu}$ and $\hat{\vec{\theta}}$ are the Maximum Likelihood Estimations (MLEs) for the POI and the nuisances and $\hat{\hat{\vec{\theta}}}$ is the MLE for the nuisances keeping μ fixed.
For discoveries the p-value for this test statistic for the background-only hypothesis is computed. Since the number of signal events has to be positive, the test statistic for discovery can be written as

$$\tilde{q_0} = \begin{cases} -2ln\lambda(\mu) & \text{for } \hat{\mu} > 0 \\ 0 & \text{for } \hat{\mu} \leq 0 \end{cases} \tag{4.9}$$

and its distribution $f(\tilde{q_0}|0, \hat{\hat{\vec{\theta}}})$ can be computed with pseudo-experiments using Monte Carlo techniques (toys). The p-value of the background-only hypothesis for the observed value of the test statistic on data $\tilde{q_{0,obs}}$ is then given by,

$$p_0 = \int_{\tilde{q_{0,obs}}}^{\inf} f(\tilde{q_0}|0, \hat{\hat{\vec{\theta}}}(\mu = 0, obs)) d\tilde{q_0}. \tag{4.10}$$

This can be converted to the quantile of a unit gaussian:

$$Z = \Phi^{-1}(1 - p_0). \tag{4.11}$$

where Φ is the cumulative distribution of a unit gaussian. This conversion is shown for a few example values in table 4.6.

p-value	Z [σ]	
1.59×10^{-1}	1.00	
2.28×10^{-2}	2.00	
1.35×10^{-3}	3.00	evidence
3.17×10^{-5}	4.00	
2.87×10^{-7}	5.00	discovery

Table 4.6: Conversion of p-values to quantiles for some specific points and commonly used High Energy Physics (HEP) definitions for evidence and discovery.

4.5.3 Approximations for discovery significance and the Asimov dataset

For the special case of the test statistic for a discovery, equation 4.9 reduces to

$$q_0 = \begin{cases} \hat{\mu}^2/\sigma^2 & \text{for } \hat{\mu} > 0 \\ 0 & \text{for } \hat{\mu} \leq 0. \end{cases} \tag{4.12}$$

It follows a Gaussian distribution

$$f(q_0|0) = \frac{1}{2}\delta(q_0) + \frac{1}{2}\frac{1}{\sqrt{2\pi}}e^{-q_0/2}. \tag{4.13}$$

and the significance can easily be calculated from the p-value to

$$Z_0 = \sqrt{q_0} \tag{4.14}$$

In the important special case of a counting experiment with known background (from simulation) the likelihood is given by

$$L(\mu) = \frac{(\mu s + b)^n}{n!}e^{-(\mu s+b)} \tag{4.15}$$

and by plugging above Poisson distribution into the expression for the likelihood ratio and using the Gaussian approximation, the significance can be computed to

$$Z_0 = \sqrt{q_0} = \begin{cases} \sqrt{2(n\ln\frac{n}{b} + b - n)} & \text{for } n \geq b \\ 0 & \text{for } n < b. \end{cases} \tag{4.16}$$

Using the so-called "Asimov dataset" for $\mu = 1$, in which every observation is set to the expected nominal value (so in this simple case $n = s + b$) the median of the significance to reject the

background hypothesis given an observation at the exact nominal value can be written as

$$median[Z_0|1] = \sqrt{q_{0,A}} = \sqrt{2((s+b)ln(1+\frac{s}{b})-s)} \tag{4.17}$$

which when limited to $s \ll b$ simplifies to $s/\sqrt{b}$.

4.6 Analysis selection

To increase signal purity and reject multi-jet background, an initial layer of preselections are applied throughout the analysis. The same selections are used for both real data and Monte Carlo simulated samples. This decreases the volume of data to process for subsequent steps. The preselection requires at least two central jets over 20 GeV, at least one isolated lepton above 20 GeV, or no such lepton, along with missing transverse energy above 150 GeV in the event. Following this, several selections are made depending on the kinematic characteristics of the source objects (leptons, jets), the lepton channel (0, 1 or 2 lepton), and the derived reconstructed objects (V and H candidates), as indicated in the following section. This is mostly a cut on the transverse momenta and acceptance in η, but also on a dedicated MVA to reduce contribution from pileup jets.

4.6.1 Channel based preselection

After the initial preselection is applied on the total events, a selection based on lepton channels are applied to distribute events between the 3 different lepton final states. This is a loose preselection and it is not very specific to the VH signal process but rather to make sure all the needed objects like jets and leptons are present. For the jets there is a minimum threshold of the transverse momentum applied for the leading (j_1) and subleading (j_2) b-tagged jets. This ensures well reconstructed pair of jets and removes some of the pileup jets. All the channel preselections are shown in table 4.7.

Anti QCD selection

In 0-lepton channel, there are no isolated leptons which can be tagged to identify events of interest from multi-jet QCD events. Therefore an additional layer of selection specifically to reject QCD events is applied in 0-lepton channel selection which are defined as follows:

- $p_T > 30$ GeV

- tight jet ID (jet selection with high purity)
- pile-up rejection for jets
- $\Delta\phi$ (jets, MET) > 0.5

Anti QCD cut mentioned in the next sections refers to these set of cuts.

Variable	0-lepton	1-lepton	2-leptons
$p_{T,V}$	-	> 150 GeV	> 75 GeV
MET	> 170 GeV	-	-
min(MET, $\not{H}_T$)	> 100 GeV	-	-
$p_T\, j_{max}$	> 60 GeV	> 25 GeV	> 20 GeV
$p_T\, j_{min}$	> 35 GeV	> 25 GeV	> 20 GeV
$p_{T,\mu}$	-	> 25 GeV	> 20 GeV
$p_{T,e}$	-	> 30 GeV	> 20 GeV
Isolation$_{rel,\mu}$	-	< 0.06	< 0.25
Isolation$_{rel,e}$	-	< 0.06	< 0.15
$\mid \eta_\mu \mid$	-	< 2.4	< 2.4
$\mid \eta_e \mid$	-	< 2.5	< 2.5

Table 4.7: Preselection for all the 3 channels.

Following initial preselection at the level of the lepton channels, events are filtered to pass through either the resolved analysis selection or the boosted analysis selection. Here j_{max} and j_{min} are defined as follows:

$$p_T\, j_{max} = max(p_T\, j_1, p_T\, j_2) \tag{4.18}$$

$$p_T\, j_{min} = min(p_T\, j_1, p_T\, j_2) \tag{4.19}$$

Each of the two analyses has a set of signal regions appropriate for extracting VH, $H \to b\bar{b}$ signal from the MVA-based template, as well as equivalent control regions to confine the main backgrounds.

4.6.2 Simplified Template Cross-section (STXS) scheme

The simplified template cross sections (STXS) scheme [21], which has been established by theoretical physicists at LHC is designed to lessen the effect of theory dependence in the measurements and to make it simple to compare theoretical models with the observations, provide a consistent schema for these measurements. By carefully analyzing the Higgs boson production characteristics, this method assesses whether there are any departures from the Standard

Model's predictions that might point to the existence of novel physics. Measurements of the $H \to b\bar{b}$ decay channel are anticipated to dominate the sensitivity to the VH production STXS bins. The ATLAS experiment has made a STXS measurement in this channel using data gathered between 2015 and 2018 [47]. It expands on the method previously used to measure a single inclusive signal strength by measuring several signal strengths (or cross-sections if one wants to reduce theoretical uncertainty) in bins defined by generator level information of the signal. These bins are inclusive in the decays of the Higgs boson and are defined separately for each of its production channels. Fig 4.8 defines the bins for the VH production mode. With the current amount of data this analysis is not sensitive to all of the STXS bins defined in Fig. 4.8, so the following changes are made:

- qqZH and ggZH contributions are merged
- for WH, 0-jet and $\geq$ 1-jet contributions are merged
- 0-75 GeV bins are not used, 75-150 GeV only for ZH

Therefore, the final bins considered are:

- WH process:
  - $150 < p_T(\mathrm{V}) < 250$ GeV (one-lepton channel)
  - $250 < p_T(\mathrm{V}) < 400$ GeV (one-lepton channel, both resolved and boosted topologies contribute)
  - $p_T(\mathrm{V}) \geq 400$ GeV (one-lepton channel, boosted topology contribution is dominant)
- Quark induced and gluon induced processes (qqZH and ggZH):
  - $75 < p_T(\mathrm{V}) < 150$ GeV (two-lepton channel)
  - $150 < p_T(\mathrm{V}) < 250$ GeV with zero additional jets (zero/two-lepton channel)
  - $150 < p_T(\mathrm{V}) < 250$ GeV with one or more additional jets (zero/two lepton channel)
  - $250 < p_T(\mathrm{V}) < 400$ GeV (zero/two-lepton channel, both resolved and boosted topologies contribute)
  - $p_T(\mathrm{V}) \geq 400$ GeV (zero/two-lepton channel, boosted topology contribution is dominant)

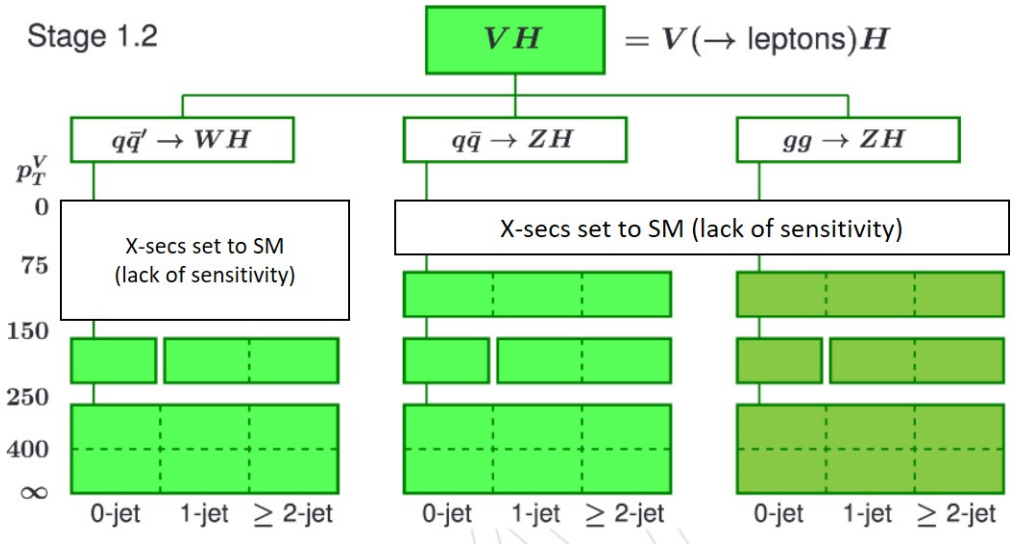


Figure 4.8: Stage 1.2 STXS scheme for VH production. Figure taken from [16]

4.6.3 Resolved analysis selection

A signal region selection is applied to segregate signal from background while maintaining a higher signal purity at the same time. Two important event variables for separating signal from background are the reconstructed di-jet invariant mass (m_{jj}) and the b-tag discriminator scores for the jets in the event. An MVA-classifier (DNN) used in the signal region further subdivides it into multiple bins of varying S/B ratio. Therefore for the signal region itself rather loose cuts are used and the selection of higher purity signal phase-space is left to the multivariate classifier. In the 0- and 1-lepton channel, there is also a cut on the number of additional jets being less than two applied to reduce the contribution from $t\bar{t}$ events.

Selections for zero lepton channel

When Z boson decays into a pair of neutrinos, the event is categorized under zero lepton channel. Therefore the signature in this channel consists of two b-jets coming from the Higgs boson and large amount of missing transverse energy (MET) originating from undetected neutrinos from the Z boson decay. For the signal region, the selection requires the events to not have any additional high-p_T prompt leptons. To reject the QCD multi-jet background, the Anti-QCD selection described in Section 4.6.1, is applied.

Variable	0-lepton	1-lepton	2-leptons
b-tag max	> medium	> medium	> medium
b-tag min	> loose	> loose	> loose
m_{jj}	$\in$ [90, 150] GeV	$\in$ [90, 150] GeV	$\in$ [90, 150] GeV
$p_{T,jj}$	> 120 GeV	> 100 GeV	-
$p_{T,V}$	> 170 GeV	> 150 GeV	>75 GeV
m_{ll}	-	-	$\in$ [75, 105] GeV
$n_{add.lep}$	= 0	= 0	-
$n_{add.jet}$	-	$\leq$1	-
$\Delta\phi(V,H)$	> 2.0	> 2.5	-
$\Delta\phi(MET,TkMET)$	< 0.5	-	-
$\Delta\phi(MET,lep)$	-	< 2.0	-
Anti-QCD	True	-	-
min(MET,$\not{H}_T$)	> 100 GeV	-	-

Table 4.8: Signal-region selection cuts for the resolved topology.

Selections for one lepton channel

The recoil of the W boson, which decays to a lepton and neutrino against the pair of b-quarks originating from the Higgs boson, is a distinctive aspect of the one-lepton channel signal signature. For the signal region, only one more jet and no additional leptons in the event is required.

Selections for two lepton channel

In the two-lepton channel the Z bosons decay to two electrons or muons. Kinematically, Z boson decay system is produced back to back against a pair of b-quark jets associated with the Higgs candidate. Owing to Z boson's decay products, MET contribution is minimal in this channel due to absence of neutrinos in final state. Due to excellent resolution of leptons in the CMS detector, a kinematic fit is performed, utilizing the kinematic information from the leptons and the Z boson to reconstruct the MET back to b-jets and improve the resolution of the Higgs mass (further details in Section 4.6.7). For the signal region, the mass of the dilepton system must be similar to that of a Z boson.

Table 4.8 shows all the cuts applied for event selection in signal region across three lepton channels. Tables 4.9 to 4.11 show the event selection cuts for all three control regions which are orthogonal to respective signal regions in each lepton channel.

4.6.4 Boosted Analysis Selection

As discussed previously in section 4.2.1, taking boosted topology into account improves sensitivity in probing $VH, H \to b\bar{b}$ process and is most relevant for high p_T bins of the STXS scheme

Variable	0-lepton	1-lepton	2-leptons
b-tag max	> medium	> tight	> tight
b-tag min	> loose	-	> loose
m_{jj}	$\in$ [50, 500] GeV	$\in$ [50, 250] GeV	> 50 GeV
$p_{T,jj}$	> 120 GeV	> 100 GeV	-
m_{ll}	-	-	$\notin$ [0, 10], $\notin$ [74, 120] GeV
$n_{add.jet}$	≥ 2	$\geq$2	-
$\Delta\phi(V,H)$	> 2.0	-	-
min $\Delta\phi(MET, jet)$	< 1.57	-	-
Anti-QCD	True	-	-

Table 4.9: $t\bar{t}$ control region selection cuts for the resolved topology.

Variable	0-lepton	1-lepton	2-leptons
b-tag max	> medium	< medium & > loose	< loose
b-tag min	> loose	> loose	< loose
m_{jj}	$\in$ [50, 500] GeV	$\in$ [50, 250] GeV	$\in$ [90, 150] GeV
$p_{T,jj}$	> 120 GeV	> 100 GeV	-
m_{ll}	-	-	$\in$ [75, 105] GeV
$n_{add.jet}$	<2	-	-
$\Delta\phi(V,H)$	> 2.0	-	> 2.5
$\Delta\phi(MET, TkMET)$	< 0.5	-	-
Anti-QCD	True	-	-

Table 4.10: V+LF control region selection cuts for the resolved topology.

Variable	0-lepton	1-lepton	2-leptons
b-tag max	> medium	> medium	> medium
b-tag min	> loose	> loose	> loose
m_{jj}	$\notin$ [90, 150] GeV	$\notin$ [90, 150] GeV	$\notin$ [90, 150] GeV
$p_{T,jj}$	> 120 GeV	> 100 GeV	-
m_{ll}	-	-	$\in$ [85, 97] GeV
$n_{add.jet}$	=0	< 2	-
$\Delta\phi(V,H)$	> 2.0	-	> 2.5
$\Delta\phi(MET, TkMET)$	< 0.5	-	-
Anti-QCD	True	-	-

Table 4.11: V+HF control region selection cuts for the resolved topology.

while reducing multi-jet background events.

Composition of AK8 jets in boosted topology

A comprehensive study was carried out to understand the composition of AK8 jets by exploiting the generator level information in Monte Carlo simulated processes and looking at the ΔR

between the reconstructed fat jets and the generator level jets. For ZH signal in 0 and 2-leptons channels, the AK8 fat jet was predominantly composed of jets from 2 b-hadrons coming from the V+jets background. Therefore, the study was more focused towards understanding the fat jet composition in background processes in 1-lepton channel.
In signal process of 1-lepton channel i.e. $WH, H \rightarrow b\bar{b}$ process, it was observed that the fat jet candidate was almost always (> 95 %) composed of two jets originating from the b-hadrons. This was studied by ΔR matching between the reconstructed fat jet and the generator level leading and subleading b-jets.
For $t\bar{t}$ process in 1-lepton channel, it was observed that the fat jet was composed of a single b-jet in a combination with a light flavored jet. The probability to find two b-jets inside a fat jet is low (< 5%). The composition of fat jet was almost similar for W+$b\bar{b}$ process, except the probability to find two generator level b-jets inside the fat jet was slightly higher ($\approx$ 15%) than that for $t\bar{t}$ process. Around 40% of the time a single b-jet constructed a single fat jet. For V + light flavored process, the fat jet was almost always constructed out of multiple light flavored jets with a minor probability of being constructed of a single b-jet in combination with light flavored jets.

Selection definition for Boosted analysis

For constructing the signal and the control regions in boosted topology, the initial preselection is applied to the boosted analysis along with some additional requirements. These requirements are soft-drop mass [48] (m_{SD}) > 50 GeV, p_T(V) > 250 GeV, and H_{pt} > 250 GeV. An additional selection is applied on 0-lepton channel requiring $p_T(MET)$ > 250 GeV. Similar to resolved topology, events in the signal region are characterized by applying selection on the DeepAK8 double b-tagger discriminant score and setting mass window on the M_{sd} of the fat jet candidate to target the mass of the Higgs boson (Table 4.12). Control regions are constructed orthogonal to the signal region similar to resolved analysis (Tables 4.13 - 4.15).

Variable	0-lepton	1-lepton	2-leptons
DeepAK8 score	> 0.8	> 0.8	> 0.8
m_{jj}	$\in$ [90, 150] GeV	$\in$ [90, 150] GeV	$\in$[90,150] GeV
$n_{add.lep}$	= 0	= 0	-
$n_{add.jet}$	= 0	= 0	-
V_{mass}	-	-	$\in$ [75, 105] GeV
Anti-QCD	True	-	-

Table 4.12: Signal region selection cuts for the boosted topology.

Variable	0-lepton	1-lepton	2-leptons
DeepAK8 score	> 0.8	> 0.8	> 0.8
m_{jj}	$\notin$ [90, 150] GeV	$\notin$ [90, 150] GeV	$\notin$[90,150] GeV
$n_{add.lep}$	= 0	= 0	-
$n_{add.jet}$	= 0	= 0	-
V_{mass}	-	-	$\in$ [75, 105] GeV
Anti-QCD	True	-	-

Table 4.13: V+HF control region selection cuts for the boosted topology.

Variable	0-lepton	1-lepton	2-leptons
DeepAK8 score	< 0.8	< 0.8	< 0.8
m_{jj}	> 50 GeV	> 50 GeV	> 50 GeV
$n_{add.lep}$	= 0	= 0	-
$n_{add.jet}$	= 0	= 0	-
V_{mass}	-	-	$\in$ [75, 105] GeV
Anti-QCD	True	-	-

Table 4.14: V+LF control region selection cuts for the boosted topology.

Variable	0-lepton	1-lepton	2-leptons
DeepAK8 score	> 0.8	> 0.8	> 0.8
m_{jj}	> 50 GeV	> 50 GeV	> 50 GeV
$n_{add.lep}$	>0	>0	-
$n_{add.jet}$	>1	>1	-
V_{mass}	-	-	$\notin$ [75, 105] GeV
Anti-QCD	-	-	-

Table 4.15: $t\bar{t}$ control region selection cuts for the boosted topology.

The DeepAK8 algorithm is calibrated by comparing data efficiency to Monte Carlo simulations in phase spaces enriched in boosted b-jets from gluon splitting events ($g \to b\bar{b}$). Because light, c, and b boosted jets are present in top-quark decays in the V+LF, V+HF, and $t\bar{t}$ control regions, and there are no dedicated studies on the efficiencies of the DeepAK8 algorithm in these regions, free floating rate-parameters are assigned to these regions to account for the algorithm's efficiency in these regions. In the context of this analysis, these rate parameters are referred to as "in-situ" scale factors, and they are constrained in the simultaneous fit. Figure 4.10 shows the inclusive m_{jj} distribution for signal regions across all the lepton channels. The signal region in 2-lepton channel has least amount of statistics compared to the other two channels but is considered the cleanest channel owing to a single source of background enrichment in the name of Z+$b\bar{b}$ process. Figure 4.11 shows the m_{jj} distribution of all the control regions across all the lepton channels. It can be observed that enrichment of different background processes

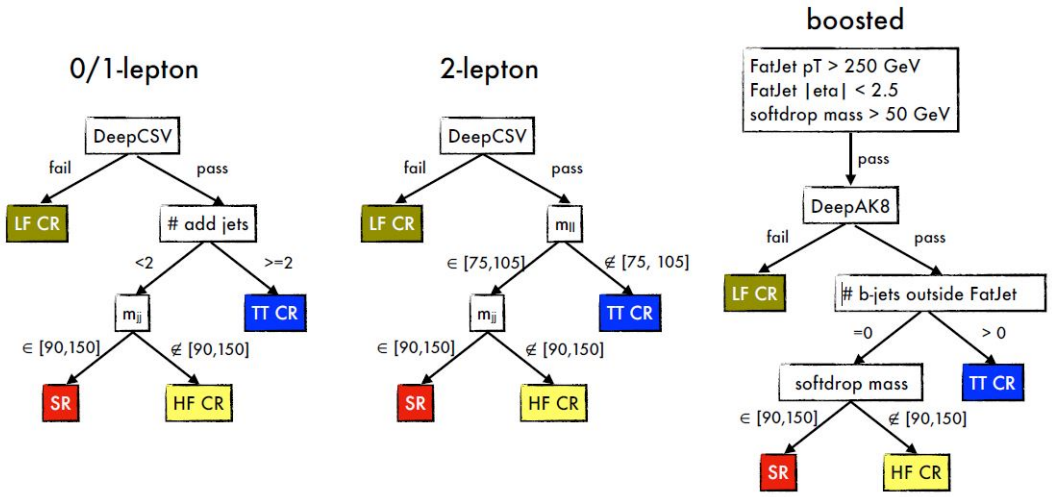


Figure 4.9: Flowchart based description of the selections used for differentiating the signal and all the control regions for the resolved and the boosted analysis. Figure taken from [17].

vary drastically across different lepton channels and each background process is enriched in their respective control regions. Overall the data/MC has a considerable agreement in all the plots.

4.6.5 Overlap between resolved and boosted topology

Since some events can be reconstructed in both the resolved and boosted analyses (called overlap events), they must be placed in either the resolved or boosted regions to avoid double counting. Four basic strategies for dealing with the overlap events depicted in figure 4.12 were examined, and the estimated uncertainty on the signal strength (μ) in the STXS bins was used as a criterion to select the best one.

Overall the second scheme from right marked with a blue box, was the final choice for the treatment of the overlap events as it provides the lowest uncertainty on μ in the high p_T(V) STXS bin among all the schemes as shown in table 4.16. In this scheme overlap events are assigned to the resolved categories, unless the event would move from the boosted signal region to a resolved control region, then it is assigned to the boosted signal region. Putting all overlap events in the resolved categories is slightly ($\approx$5%) worse than the selected scheme, but $\approx$20% better than putting all overlap events in the boosted categories. Overlap events contribute a large fraction to all boosted events in the p_T(V) range the analysis is sensitive in, which can be seen in figure 4.13.

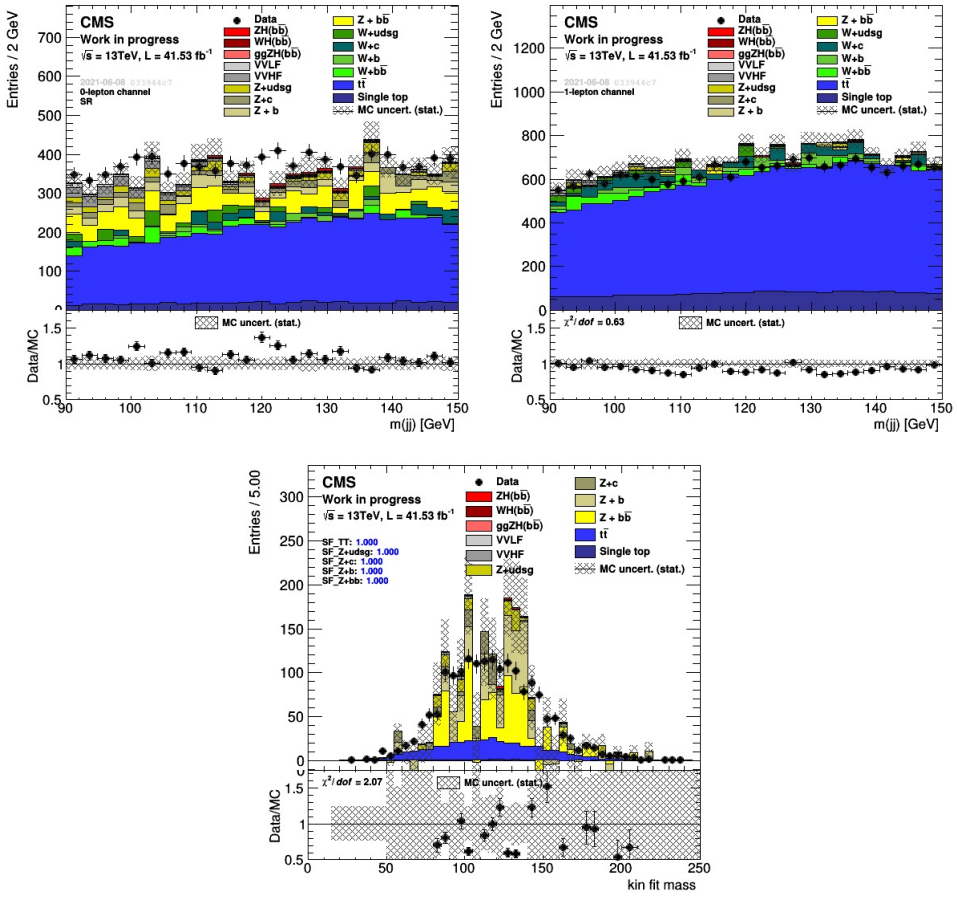


Figure 4.10: Di-jet invariant mass spectrum for data and Monte Carlo simulation for the three lepton channels inclusively in resolved topology. 0-lepton channel (top left) and 1-lepton channel (top right) have a higher background enrichment compared to 2-lepton signal region (bottom). Simulation of relevant background processes are color coded differently and marked accordingly.

	scheme 1	scheme2	scheme 3	scheme 4
ZH $p_T(V) > 250$	0.59	0.45	0.57	0.47
WH $p_T(V) > 250$	0.69	0.47	0.67	0.49

Table 4.16: Statistical uncertainties on μ in the high STXS bins. The order of the schemes is the same as in figure 4.12. Uncertainties on the STXS bins not mentioned in this table are not affected.

4.6.6 Top quark reconstruction

In the events with one lepton and MET in the final states, it is helpful to reconstruct the top quark mass by assuming that the lepton and MET stem from the W boson decay and combine their four-momenta with that of the b-jet spatially close to this system to reconstruct the four-momentum of the top quark. Then the top quark mass estimate is used as an input to the

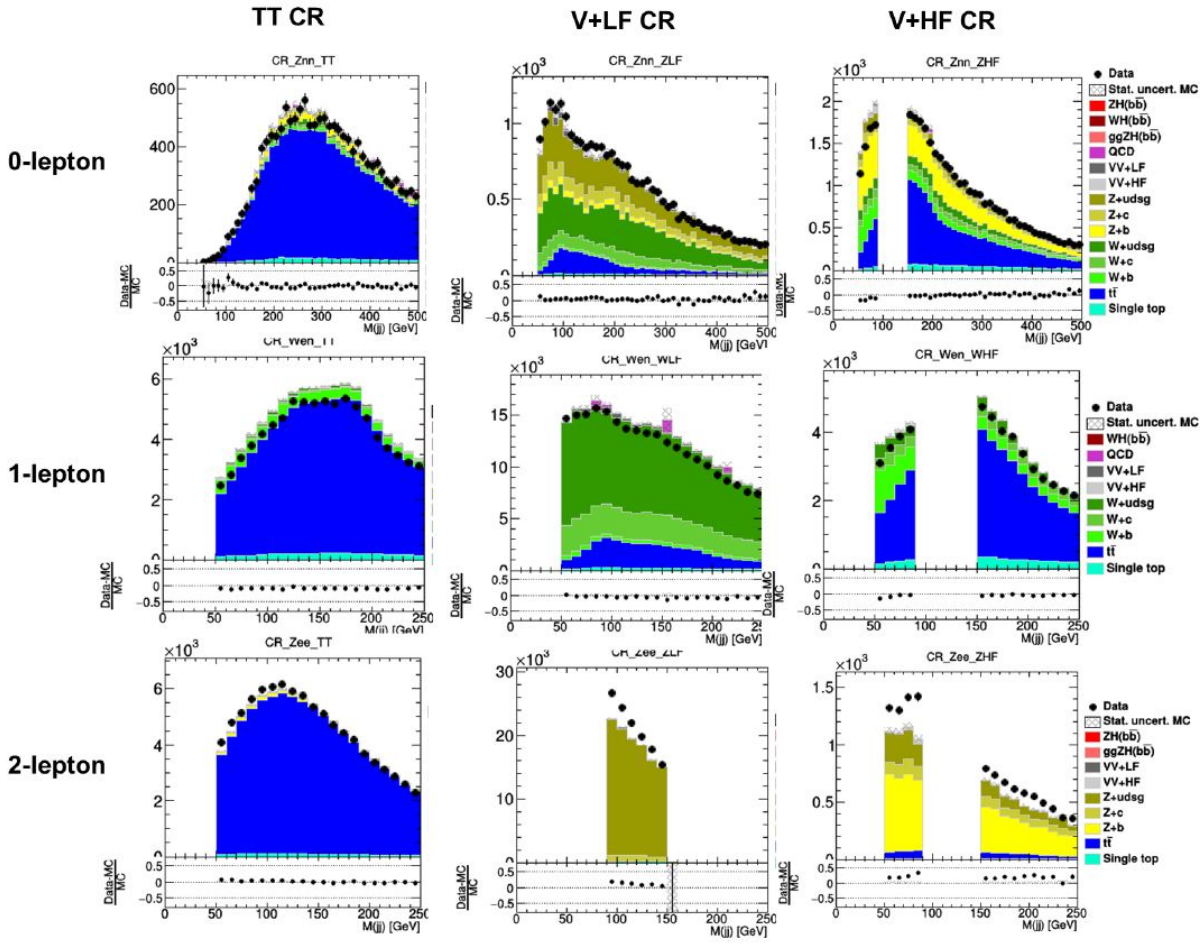


Figure 4.11: Di-jet invariant mass spectrum for data and Monte Carlo simulation for the three control regions (columns) across three lepton channels (rows) in the resolved analysis. Simulation of relevant background processes are color coded differently and marked on the right hand side of the figure.

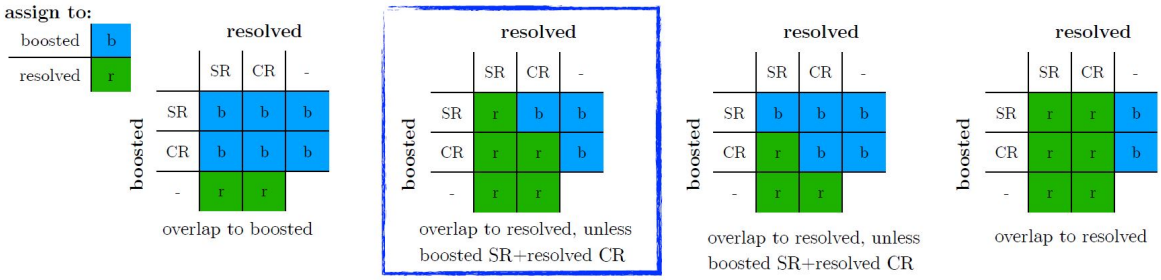


Figure 4.12: The four different overlap treatment schemes between the resolved and boosted analysis which have been studied. The scheme marked with a blue box was finally selected. Figure taken from [17].

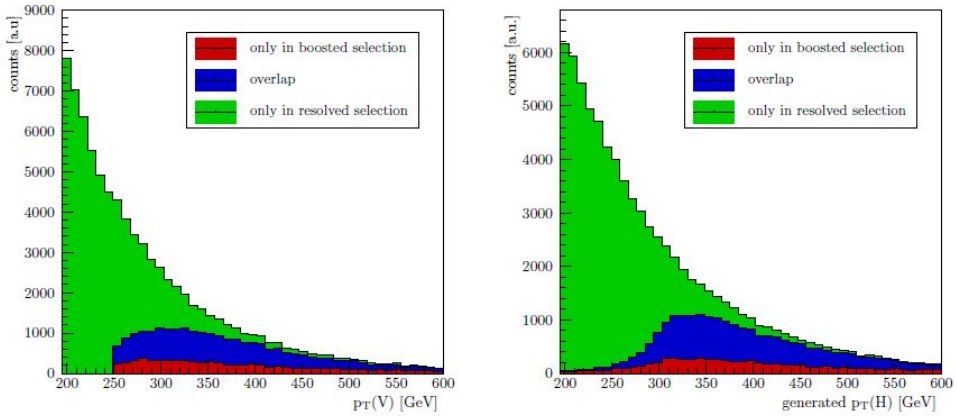


Figure 4.13: Purely resolved, overlap and purely boosted events vs. reconstructed p_T(V) (left) and generated p_T(H) (right). Picture taken from [17].

multivariate method described in Section 4.7. An example of the reconstructed top quark mass distribution is shown in Fig. 4.14.

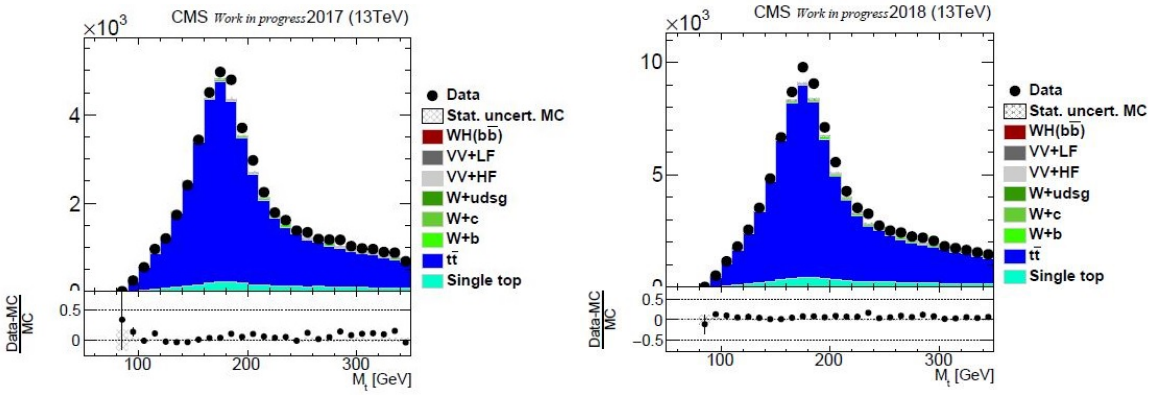


Figure 4.14: The distribution of the top quark mass in the $t\bar{t}$ enriched region, for the single electron channel with 2017 data (left) and the single muon channel with 2018 data (right).

4.6.7 Higgs boson reconstruction

The Higgs boson candidate is reconstructed from the four-vectors of the two highest pT b-jets (the b-jets selection in the signal regions is described in Section 4.6.3) that pass all the selections. In order to improve the accuracy of this reconstruction, jets with $p_T > 30$ GeV that are within a ΔR cone of less than 0.8 around any of the two selected b-jets are attributed to FSR jets. The FSR jets four-momenta are added to the four-momenta of the selected b-jets. Furthermore, the b-jet regression and finally a kinematic fit for the two-lepton channel only,

which corrects the MET and jet pT using the information from the Z boson, complete the Higgs candidate reconstruction. In Fig. 4.15 the distributions are fitted with the double shouldered crystal ball function [49] to find the width and mean.

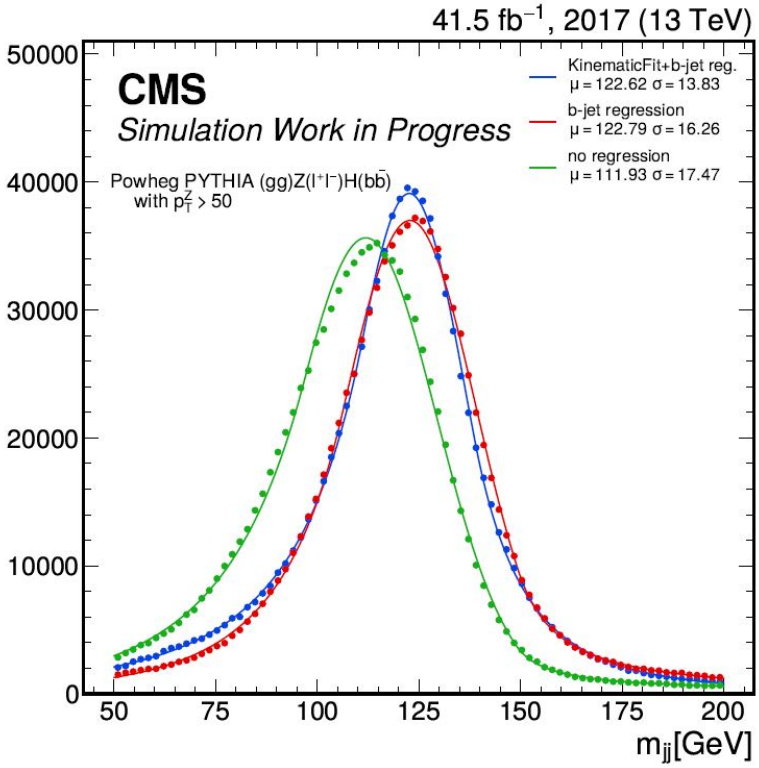


Figure 4.15: Comparison of Higgs candidate dijet system mass before and after applying corrections. Comparing the mean values of the fit, the b-jet regression pushes the Higgs candidate mass closer to the expected mass for Higgs. The kinematic fit significantly improves the resolution as one can see from the σ values of the fitted distributions.

4.7 Multi-variate methods

In order to obtain a good separation between signal and background in the signal templates, and in order to constrain certain backgrounds, multi-variate methods are used in this analysis. Deep Neural Networks (DNN) and Boosted Decision Trees (BDTs) are used for classification in signal region templates while multi-classifier DNN is used in V + heavy flavor templates in 0 and 1 lepton channels. For V+ heavy flavor region in 2 leptons channel, binned working points of b-tagging score of leading and sub leading b-jet is combined and used.

4.7.1 Deep neural networks (DNN)

Neural networks are set of algorithms which are designed to perform complex classifications or evaluate complex mathematical models. They are also used to perform regression corrections to simulated Monte Carlo to match with data. Such network consists of hidden layers which are intertwined to each other through connections in layers called nodes. Since neural network used in this analysis involves complex layers with numerous nodes for better classification, they are called Deep neural networks (DNN). Generally, a DNN is trained on training dataset and then evaluated on test datasets. In this analysis, DNNs are trained on a subset of Monte Carlo samples to classify signal or background and the choice of input features affect the performance of the trained DNN. The trained DNN is then evaluated on observed data and also on the testing dataset, which is the other subset of the Monte Carlo samples. The input variables used for training the DNN mainly includes kinematic features of the $VH, H \to b\bar{b}$ process such as di-jet invariant mass, separation between the reconstructed vector boson and Higgs boson etc. All the input variables are summarized in Table 4.17 for each channel. For training and evaluation of the DNNs, the TensorFlow [50] framework, provided as open-source software library by Google was used. The architecture of the DNNs consists of five hidden layers with sizes of 512, 256, 128, 64, 64 and 64 nodes. The DNN is used both for the multi-class and the binary signal background classification. A non-linear activation function applied to the output of each node enables the network to learn non-linear features. Leaky ReLU activation [51] is implemented on first 6 layers of the DNN which is defined as follows,

$$g(x) = \begin{cases} x & \text{if } x > 0 \\ 0.2x & \text{otherwise} \end{cases} \tag{4.20}$$

shows good results despite being very simple and fast to compute. The last layer is applied with softmax activation function which is defined by,

$$\sigma(\vec{r}_i) = \frac{e^{r_i}}{\sum_{j=1}^{m} e^{r_j}}. \tag{4.21}$$

to make the unnormalized output nodes interpretable as a probability. It can be shown that the probabilities for each class are then given by the softmax of the output nodes:

$$\vec{p} = \sigma(\vec{r}) \tag{4.22}$$

The structure of the DNN is schematically shown in Fig. 4.16. In case of binary classifica-

Variable	Description	0-lepton	1-lepton	2-lepton
M_{jj}	Dijet invariant mass	✓	✓	✓
p_{T}(jj)	Dijet transverse momentum	✓	✓	✓
p_{T}(MET)	Missing transverse momentum	✓	✓	✓
$\mathrm{M_T(V)}$	Transverse mass of the vector boson		✓	
p_T(V)	Transverse momentum of the vector boson		✓	✓
p_{T}(jj)/ p_T(V)	Ratio of transverse momenta of the vector boson and Higgs boson		✓	✓
$\Delta\phi$(V,H)	Azimuthal angle between the vector boson and the di-jet directions	✓	✓	✓
$\mathrm{btag_{max}}$	b tagging score of leading jet	✓	✓	✓
$\mathrm{btag_{min}}$	b tagging score of subleading jet	✓	✓	✓
$\Delta\eta$(jj)	Pseudorapidity difference between leading and sub-leading jet	✓	✓	✓
$\Delta\phi$(jj)	Azimuthal angle between leading and sub-leading jet	✓	✓	
$p_T^{\mathrm{max}}(\mathrm{j_1,j_2})$	Maximum transverse momentum of jet between leading and subleading jet	✓	✓	
SA5	Number of soft-track jets with momentum greater than 5 GeV	✓	✓	✓
$\mathrm{N_{aj}}$	Number of additional jets	✓	✓	
$\mathrm{btag_{max}}$ (add)	Maximum btagging discriminant score among additional jets	✓		
p_T^{max}(add)	Maximum transverse momentum among additional jets	✓		
$\Delta\phi(\mathrm{jet}, p_{\mathrm{T}}(MET))$	Azimuthal angle between additional jet and p_{T}(MET)	✓		
$\Delta\phi(\mathrm{lep}, p_{\mathrm{T}}(MET))$	Azimuthal angle between lepton and p_{T}(MET)		✓	
M_{t}	Reconstructed top quark mass		✓	
$p_{\mathrm{T}}(\mathrm{j_1})$	Transverse momentum of leading jet			✓
$p_{\mathrm{T}}(\mathrm{j_2})$	Transverse momentum of sub-leading jet			✓
M(V)	Reconstructed vector boson mass			✓
ΔR(V,H)	Angular separation between the vector boson and Higgs boson			✓
ΔR(V,H) (kin)	Angular separation between the vector boson (reconstructed after kinematic fit) and Higgs boson			✓
$\sigma(\mathrm{M(jj)})$	Resolution of dijet invariant mass			✓
N_{rec}	Number of recoil jets			✓

Table 4.17: Input variables used for the DNN training in the resolved SR of the 0-, 1- and 2-lepton channels. Reconstructed jets are classified as leading and subleading based on their b-tag score.

tion DNN which is used in signal region, all the signal process described in Section 4.3.1 are considered as one class (signal class) and background processes described in Section 4.3.2 are cumulatively classified as background class. The distribution of the signal-background classifier DNN is shown in Fig. 4.17. As mentioned above, V+HF region for 0 and 1 lepton channel uses a trained multi-classifier DNN as template. It adds additional control over the single-top and $t\bar{t}$ backgrounds while creating templates to improve modelling of the b and c-quark contri-

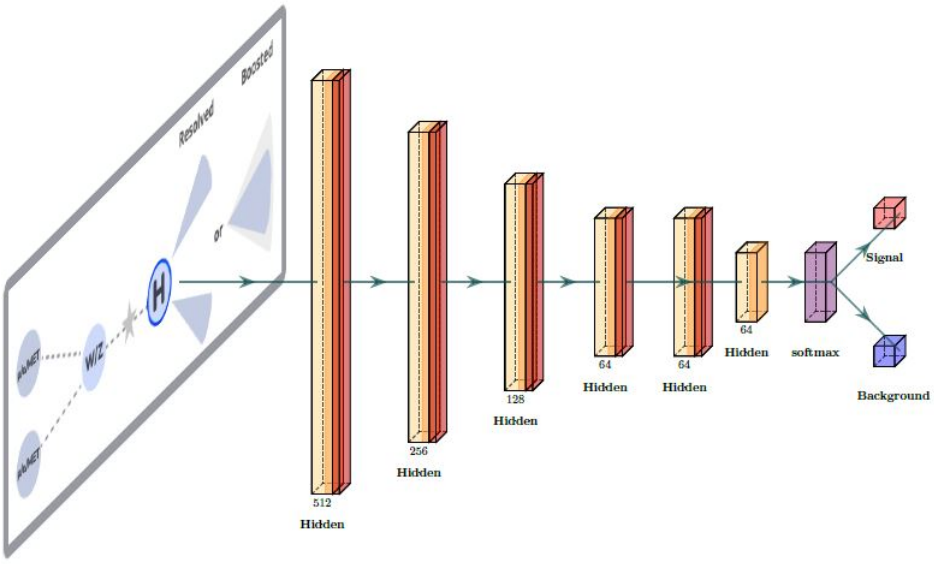


Figure 4.16: The architecture of the DNN, after each hidden layer a Leaky ReLU activation and on the last layer a softmax activation is used. Picture taken from [18].

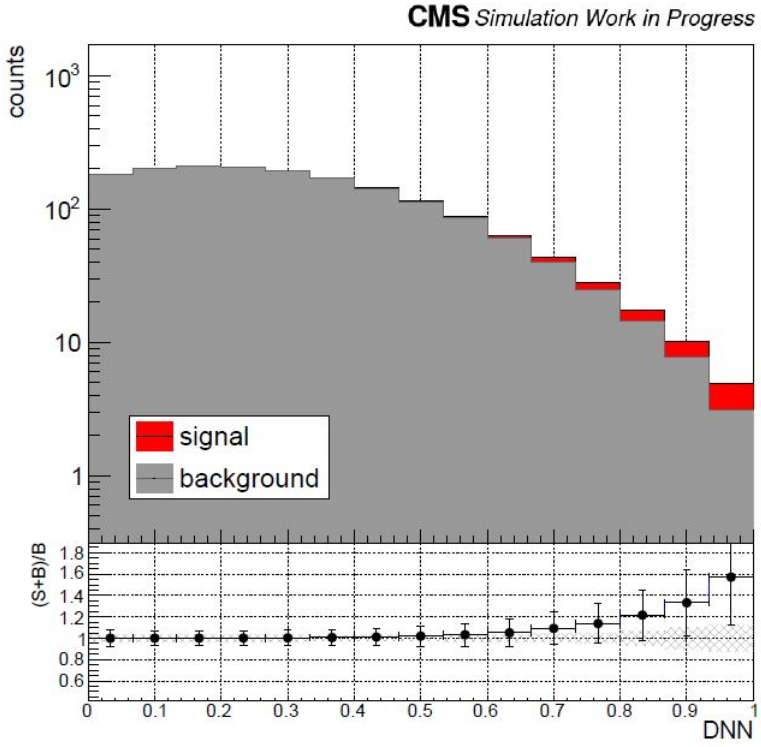


Figure 4.17: DNN output for the signal-background classification in 2-lepton high pT(V) region.

butions. Five classes of the classifier are defined and summarized in Table 4.18. Classes from 0-2 represent V+light flavored, V+charm flavored and V+b flavored jets. Classes 3 and 4 are

assigned to the single-top and the $t\bar{t}$ which correspond to the same named background processes.

For V+HF region in 2-lepton channel, the DeepCSV b-tagging working points are used to define a template to help the fit in constraining b and c-quark contributions. The labeling is described

0	V+udsg (VL)
1	V+c (VC)
2	V+b (VB)
3	single-top (ST)
4	$t\bar{t}$ (TT)

Table 4.18: Classes used for the 0/1-lepton V+HF multi-background classifier.

in Table 4.19 where "T" refers to the tight working point selection, "M" to the medium working point selection and "L" to the loose working point selection.

value	DeepCSV max	DeepCSV min
0	< T	< M
1	< T	> M
2	> T	< M
3	> T	> M, < T
4	> T	> T

Table 4.19: Class labelling used for template fit in 2-lepton HF control region.

4.7.2 Boosted decision tree

Decision tree is a powerful machine learning algorithm which is used widely for classification and regression purposes. They are advantageous compared to DNN in terms of resource and time required for training and evaluation. It consists of flowchart-like tree structures where each internal node represents a test on an input feature and each branch represent an outcome of the test. In boosted decision tree (BDT), each tree is dependent on outcome of prior trees. Therefore, BDTs tend to be more accurate than a conventional decision tree. For the boosted signal regions in all channels, BDTs are used for classification between signal and background. The classification is binary similar to DNN classification in resolved analysis. The input variables used in the training of the BDT are properties of the AK8 fat jet. To maintain a smooth transition between the resolved and boosted analysis selection and account for the overlap events some resolved kinematic features are used in the training. The full list of input variables used for training is summarized in Table 4.20. The distributions of all the BDT outputs for all channels across all years are shown in Fig. 4.18.

Variable	Description	Resolved	Boosted
M_{jj}	Dijet invariant mass	✓	
p_{T}(jj)	Dijet transverse momentum	✓	
p_{T}(MET)	Missing transverse momentum	✓	✓
$\Delta\phi(\mathrm{V},\mathrm{H})$	Azimuthal angle between the vector boson and the di-jet directions	✓	
$p_T^{\mathrm{max}}(\mathrm{j}_1,\mathrm{j}_2)$	Maximum transverse momentum of jet between leading and subleading jet	✓	
$\Delta\phi(\mathrm{jet}, p_{\mathrm{T}}(MET))$	Azimuthal angle between additional jet and p_{T}(MET)	✓	
N_{aj}	Number of additional jets	✓	
$\Delta\eta$(jj)	Pseudorapidity difference between leading and sub-leading jet	✓	
$\Delta\phi$(jj)	Azimuthal angle between leading and sub-leading jet	✓	
$p_{\mathrm{T}}(\mathrm{j}_1)$	Transverse momentum of leading jet	✓	
$p_{\mathrm{T}}(\mathrm{j}_2)$	Transverse momentum of sub-leading jet	✓	
p_{T}(Fj)	Transverse momentum of the leading AK8 jet		✓
M_{Fj}	Mass of the leading AK8 jet		✓
η_{Fj}	pseudo-rapidity of the leading AK8 jet		✓
$\Delta\phi$(V,Fj)	Azimuthal angle between vector boson and leading AK8 jet		✓
$\mathrm{btag}_{\mathrm{AK8}}$	binned DeepAK8 double b-tagged score of the leading AK8 jet		✓

Table 4.20: List of input variables used in training the BDT for boosted topology.

4.8 Systematic uncertainties

In this analysis, there are many sources of systematic uncertainties which are considered. These uncertainties play an important role in the fit affecting the Monte Carlo process template it is associated to. Therefore, correct assessment of the systematic uncertainty template is highly essential in the overall fit. Several systematic uncertainties have effects on the normalization of signal or background processes, many others on the shape of the key observables. All the systematic uncertainties accounted for in this analysis are listed and explained below.

4.8.1 Uncertainties affecting normalization only

Luminosity

Instantaneous luminosity is measured in the CMS detector in the following sub-detector parts,

- Pixel Luminosity Telescope (PLT) detector: silicon pixel based detector arranged in 16 telescope configurations with 3 layers each, placed outside of the pixel detector endcaps. It counts tracks with coincidences of hits in all three layers.

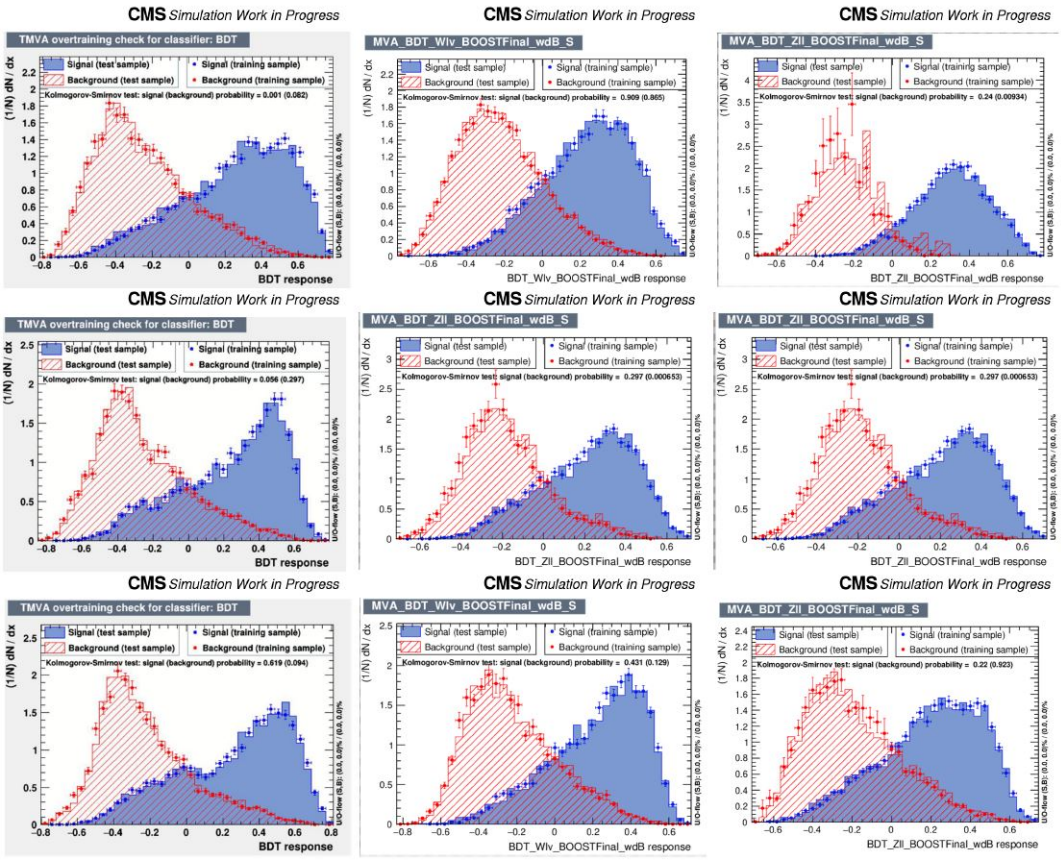


Figure 4.18: Overtraining tests for the boosted topology BDT for signal-background classification in all channels across three years. The columns represent BDTs trained in different lepton channels while the rows represents different years of data taking from 2016 to 2018 arranged top to bottom respectively.

- BCM1F detector: a fast beam condition monitor with a time resolution of 6.5 ns.
- HF calorimeters: total sum of transverse energy used
- DT: rate of tracks in barrel muon chambers
- pixel detector: counting of number of clusters

and calibrated by Van der Meer scans as mentioned in section 2.1. The uncertainty in the integrated luminosity measurement is 2.5% in 2016 and 2018, and 2.3% in 2017 [52, 53, 54]. These uncertainties are partially correlated between the different data-taking periods.

Signal theory uncertainties

For measuring μ, theoretical systematic uncertainties for signal are taken into account. These uncertainties are associated to the ratio of the measured over the cross-section computed from

theory. The sources of theory uncertainties for the signal originate from scale uncertainties, PDF uncertainties, cross section and branching ratio computation and electroweak corrections. Scale uncertainties are accounted for by varying the renormalization and factorization scales by factors of 0.5 and 2.0 for the down and up variations, respectively. Following are the full list of signal theory uncertainties, that are accounted in the analysis:

pdf qq ZH	1.6%
pdf qq WH	1.9%
BR($H \to b\bar{b}$)	0.5%
cross section qqZH	0.5%
cross section ggZH	22.0%
cross section WH	0.6%
electroweak correction	2.0%

Background theory uncertainties

For the three most important background processes namely $t\bar{t}$, Z+jets and W+jets normalization is measured in-situ in the fit by assigning respective process scale factors. These scale factors are derived from their respective control regions. Scale factors for V+jets are split further in jet flavor as V+b, V+c and V+light flavor to accurately estimate the respective V+jets components. For diboson and ST on the other hand, theory predictions are used and a normalization uncertainty is added for their cross-sections:

VV cross-section	15%
ST cross-section	15%

In addition, for all of the backgrounds, QCD renormalization and factorization scale uncertainties are taken into account by varying the scales by factors of 0.5 and 2.0 for the down and up variations, respectively. PDF uncertainties are implemented as normalization uncertainties for which the size has been determined from a set of PDF variations:

TT	0.5%
V+udsg	5.0%
V+c	5.0%
V+b	3.0%
V+b$\bar{b}$	2.0%
ZZ(LF)	3.0%
WZ(LF)	2.0%
VZ(HF)	2.0%

Lepton efficiencies

These set of uncertainties arise from measurement of leptons in the detector simulation. Uncertainties in the electron and muon ID, isolation, and trigger efficiencies amount to 2%. The uncertainties are determined by changing the parameters used to perform the efficiency measurement and define the range of the lepton scale factors. After that the effects of the variations are estimated in the analysis selection regions.

MET trigger efficiencies

Uncertainties in the MET trigger efficiency measurement amount to 1%, the number is estimated similar to the one for leptons.

4.8.2 Uncertainties affecting normalization and shape

Jet Energy Scale (JES)

The jet energy scale uncertainties are categorized as seen in Table 4.21. All the uncertainties are associated to measurement of the jet energy from various sub-detector parts. Some uncertainties also arise from various Monte Carlo generator differences. The uncertainties are given as function of the jet transverse momentum and pseudorapidity (η) and their variation in these variables is shown in figure 4.19. In general the relative uncertainty decreases with high pT and in the central region in η where tracking is available.

Uncertainty	Description
Absolute	Uncertainties on the absolute scale
Fragmentation	Difference of Fragmentation and UE between Pythia/Herwig
SinglePionECAL	Single-pion response in ECAL, 3%
SinglePionHCAL	Single-pion response in HCAL, 3%
Flavor	Variation of possible color mixtures
RelativeJER	Jet p_T resolution
RelativeBal	MPF (Missing Transverse Energy Projection Fraction) vs. p_T-balance
RelativeSample	Difference among dijet, Z+jets, g+jets
RelativeFSR	ISR+FSR correction
RelativeStat	Statistical uncertainty
PileUpDataMC	Data vs. MC simulation offset
PileUpP_t	Jet p_T-dependent offset

Table 4.21: The jet energy scale uncertainties groups used, each of which can be subdivided into several uncertainties, which covers distinct methodologies, samples, or detector locations, this grouping is based on [20].

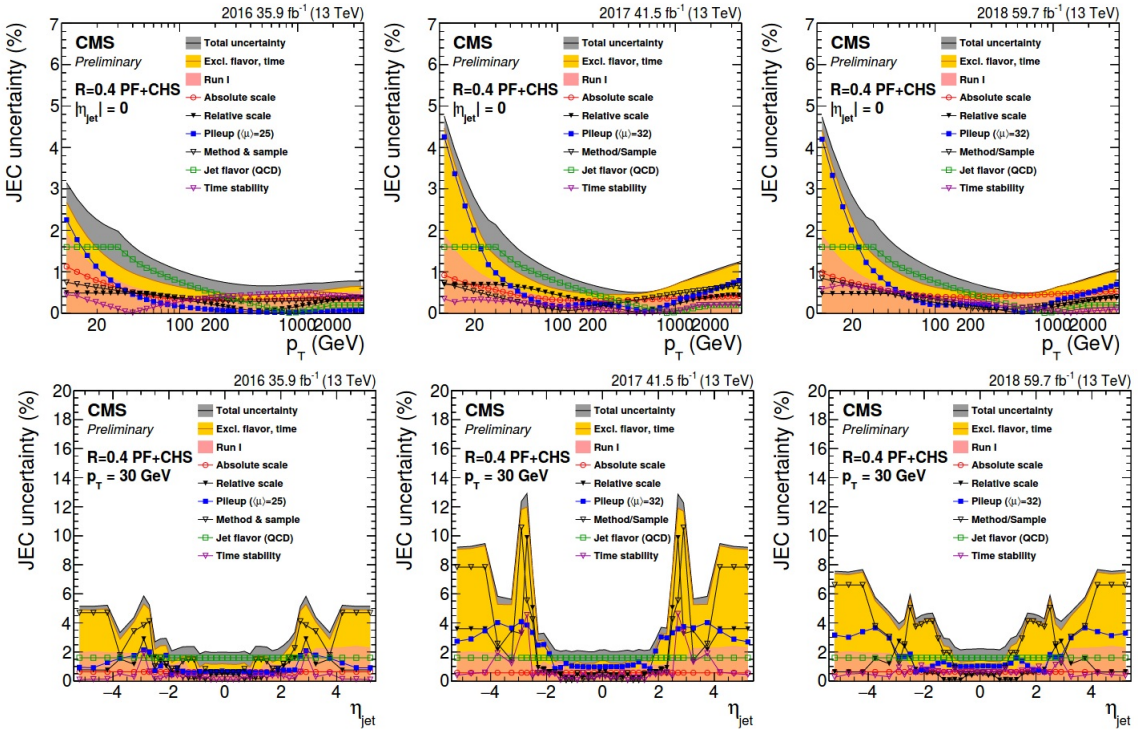


Figure 4.19: JES uncertainty sources and total uncertainty (quadratic sum of individual uncertainties) as a function of p_T^{Jet} (top) and η_{Jet} (bottom) for all three years. Figure taken from [19].

Jet Energy Resolution (JER)

Resolution of jet energy is different for recorded data and for simulated events. Therefore, resolution of simulated jets from Monte Carlo samples are smeared to match to data. The smearing is applied to the MC sample jets as follows,

$$p_{T,smeared} = s_{JER} \times (p_{T,gen} + (p_{T,reco} - p_{T,gen})(1 + s_{diff})). \tag{4.23}$$

where the $p_{T,gen}$ are the jet p_T without the simulated detector effects and the $p_{T,reco}$ after considering the detector effects. The s_{JER} and s_{diff} are listed in Table 4.22.

Year	Scaling (s_{JER})	Resolution Difference (s_{diff})
2016	0.998 ± 0.019	0.017 ± 0.060
2017	1.020 ± 0.023	0.088 ± 0.071
2018	0.985 ± 0.019	0.080 ± 0.073

Table 4.22: The smearing corrections for each data taking year as a percent of the jet's p_T.

b-tagging uncertainties

Uncertainties in the DeepCSV b-tagger's data/MC calibration are represented as shape changes for nine uncorrelated sources, which can be categorized into three groups:

- Jet energy scale uncertainties
- Flavor contamination, light+charm in heavy-flavor control region and b+charm in light-flavor
- Statistical uncertainties corresponding to each flavor of the tagged jets

Each of the calibrations and their associated uncertainties are derived in bins of jet p_T and η. For boosted topology, the DeepAK8 double b-tagger is used. However, since calibrations for DeepAK8 is not available, the efficiency and variations are accounted for by using "in-situ" scale factors for the background. These scale factors are unconstrained in the final fit. Shape uncertainties with variations up to 10% are used for signal processes in bins of the double b-tagger score and p_T^{jet} scores (from 0.8-0.97 and from 0.97-1.0).

Uncertainties due to limited Monte Carlo statistics

To account for systematic uncertainties originating from limited statistics of Monte Carlo simulated samples, Barlow-Beeston method [55] is used. Instead of using a separate nuisance parameter for each process in a bin, as an approximation a single nuisance parameter per bin can be used instead due to the independence of the processes. The contribution to the Negative Log Likelihood (NLL) reads:

$$-lnl_i = -nln\sum_j \beta_{ij}\mu_{ij} + \sum_j \beta_{ij}\mu_{ij} + \sum_j \frac{(\beta_{ij}-1)^2}{2\sigma^2_{\beta_{ij}}} \tag{4.24}$$

The resulting set of equations obtained by minimizing equation 4.24 can be solved numerically [56].

Chapter 5

Result and Summary

In this chapter, the results of the analysis described in this thesis are presented.

5.1 Signal strengths modifiers of the VH($b\bar{b}$) process

To extract the result, a combined simultaneous signal and background likelihood fit in control and signal regions is performed as described in Section 4.2. The templates used in each region of the fit are summarized in Table 5.1. The DNN, as stated in Section 4.7.1, is utilized in the

	SR	$t\bar{t}$ CR	V+LF CR	V+HF CR
zero-lepton, resolved	DNN	p_T(V)	p_T(V)	HFDNN
one-lepton, boosted	BDT	Double b-tagger	Double b-tagger	Double b-tagger
one-lepton, resolved	DNN	p_T(V)	p_T(V)	HFDNN
one-lepton, boosted	BDT	Double b-tagger	Double b-tagger	Double b-tagger
two-lepton, resolved	DNN	p_T(V)	p_T(V)	DeepCSV scores
two-lepton, boosted	BDT	Double b-tagger	Double b-tagger	Double b-tagger

Table 5.1: The variables for the distributions used in the fit for each signal and control region. The DNN and BDT distributions are used in the signal regions. The p_T(V) is used in the resolved control region for V+LF. The b-tagging discriminant distribution is used in the V+LF and V+HF boosted control regions as well as the V+HF two-lepton resolved control region, while the HFDNN is used for the remaining resolved topologies.

resolved signal regions, and the BDT distributions (Section 4.7.2) are used in the boosted signal regions. The p_T(V) variable is used for V+LF and $t\bar{t}$ resolved control region. The DeepAK8 discriminant is employed in the V+LF and V+HF boosted control regions, and the DeepCSV b-tagging discriminant distribution (binned in the working point cuts provided in Table 5.1) is used in the V+HF two-lepton resolved control region while in zero and one-lepton the trained V+HF

multi-classifier DNN is used. The final set of events in the channel and in each region (signal or control) are further divided according to the STXS template scheme since cross-sections are measured in the STXS scheme. Each STXS bin's signal and control areas are simultaneously fit using a binned maximum likelihood fit to provide a signal strength modifier (μ), which indicates the proportion of observed VHbb events to those predicted by the SM. The expected SM cross-section is represented by μ=1.
Additionally, the overall normalization and shapes for each template in the fit are modified within the scope of statistical and systematic uncertainties stated in Section 4.8. The free parameters associated with the normalization of significant background processes, notably the $t\bar{t}$, $V+udsg$, $V+c$, and $V+b$ processes in the fit (also called scale factors), are constrained by the control region fits, then extrapolated to the signal regions. These scale factors are inclusive in p_T(V). For the boosted topologies in the high p_T(V) regions, dedicated in-situ scale factors described in Section 4.6.4 are used.

5.1.1 VZ(Z$\rightarrow b\bar{b}$) and dijet mass cross-check analyses

VZ(Z$\rightarrow b\bar{b}$) analysis is used as a cross-check analysis to validate the overall VHbb analysis strategy. The VZ process, where the Z boson decays into a pair of b-quarks, has an identical final state as the VH process with $H \rightarrow b\bar{b}$. The simulated diboson sample is used as the training signal for the DNN and BDT discriminants in the resolved and boosted SRs. The background processes are all taken into consideration. The sole difference between the VZ(Z$\rightarrow b\bar{b}$) analysis and the VH analysis is the necessity that M_{jj} lies in the range of 60-120 GeV to define the SR for all channels. The extracted signal strengths for the ZZ and WZ processes are reported in Fig. 5.1 for all channels after analyzing the 2016–2018 data set. The observed and expected significance are both substantially over 5 standard deviations based on the inclusive observed VZ(Z$\rightarrow b\bar{b}$) $\mu = 1.16 \pm 0.13$. Measurement of exclusive μ dedicated to both the vector bosons yielded $\mu = 1.04 \pm 0.14$ for ZZ$\rightarrow b\bar{b}$ and $\mu = 1.62 \pm 0.28$ for WZ$\rightarrow b\bar{b}$ process.

Dijet mass analysis is another cross-check analysis which involves fitting the Dijet invariant mass spectrum of the two b-jet candidates. Figure 5.2 displays the combined dijet invariant mass distribution for all channels for the VH(H$\rightarrow b\bar{b}$) and VZ(Z$\rightarrow b\bar{b}$) processes, both with and without subtracting the background processes. Dedicated DNNs in the resolved SRs were used to obtain this distribution. For the purpose of not biasing the background shape, these DNNs do not use the dijet mass as an input feature. To emphasize the signal contribution to the distribution, all events are weighted according to S/(S+B). Following the fit to data, which includes applying the process scale factors, the amount of signal and background events in each bin of the DNN distribution's output is denoted by the letters S and B respectively.

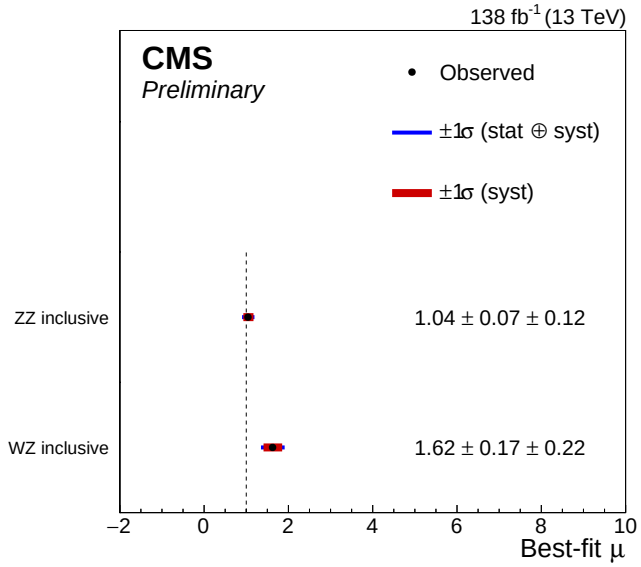


Figure 5.1: Result of the VZ, $Z \to b\bar{b}$ channel analysis using the full Run 2 dataset for both the WZ and ZZ production modes

The signal excess at m_H=125 GeV is consistent with the data. Since the DNN discriminant used for separating the VH signal from the total backgrounds does not take into consideration the dijet mass, which is a very potent signal-to-background separator, the sensitivity of this analysis is lower than the main STXS measurement. The fitted signal strength is obtained to be $\mu = 0.34 \pm 0.34$.

5.1.2 $VH, H \to b\bar{b}$ STXS Measurement

In the upper plot of Fig. 5.3, the signal regions for the reconstructed-level STXS categories for 2017 are represented as fractions; the patterns for the 2016 and 2018 are anticipated to be similar. This plot shows the migrations between the generator-level STXS categories and contamination from other signal processes. The STXS categorization is congruent with the reconstructed categories measured in this analysis, as shown in Fig. 5.3. The 2-lepton channel has a higher signal purity than the 0-lepton and 1-lepton channels. The correlation matrix of the signal strengths divided per STXS bin is displayed in Figure 5.3 (bottom) for the analysis of all data taking years combined. As anticipated, the correlation between the signal intensity for the medium p_T(V) STXS bins with 0 and at least 1 jet is the highest (-21%). Combining all three data-taking years, the simultaneous maximum-likelihood fit of the signal and control regions yielded the inclusive signal strength of $\mu = 0.57^{+0.19}_{-0.18}$ and the inclusive likelihood scan is shown in Figure 5.4. The signal strengths for each analysis channel are shown in Figure 5.5,

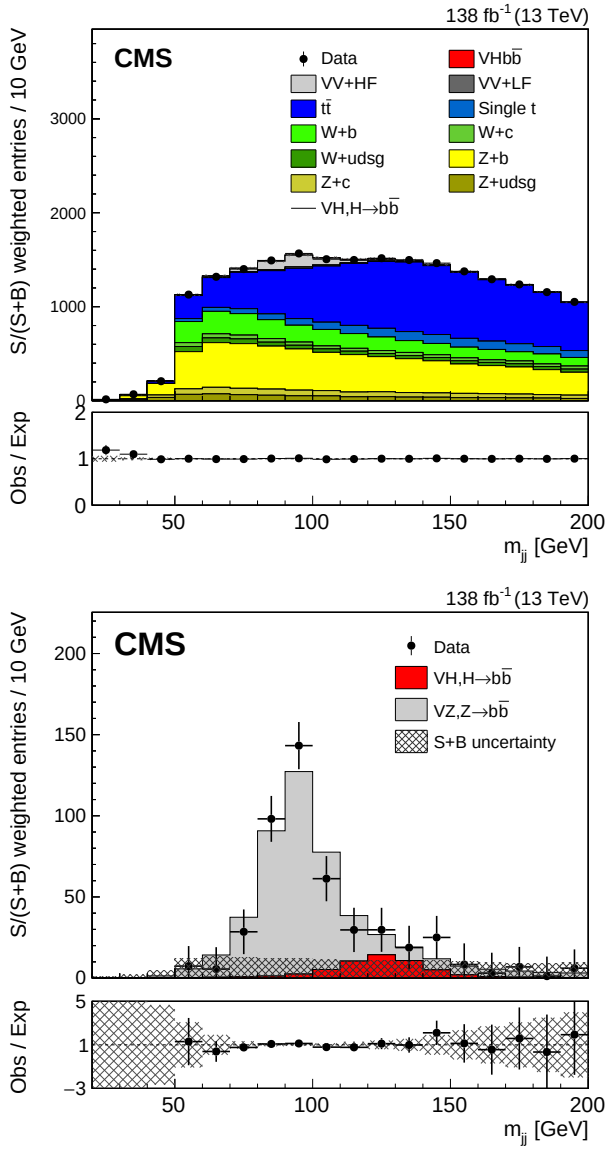


Figure 5.2: Dijet invariant mass distributions, combining all channels and data-taking periods, with events weighted according to S/(S+B). The distributions are evaluated after the fit to data and as a result, the fitted signal strength is utilized to scale the signal component. To display the invariant mass peaks of the VZ(Z→ $b\bar{b}$) and VH(H→ $b\bar{b}$) resonances, all background processes other than the VH and VZ contributions are also exhibited (top) or subtracted (bottom).

along with the signal strengths broken down by production mode (ZH or WH). The total significance of the three leptons channels' individual departures from the SM expectation (μ=1) is 2.9 standard deviations. The compatibility p-value between the Z(ll)H and Z($\nu\nu$)H decay modes against the inclusive ZH decay is 20%. Figure 5.6 (top) shows the measured signal strength in the different STXS bins, fitting all data-taking years (2016–2018). The branching fraction (sB)

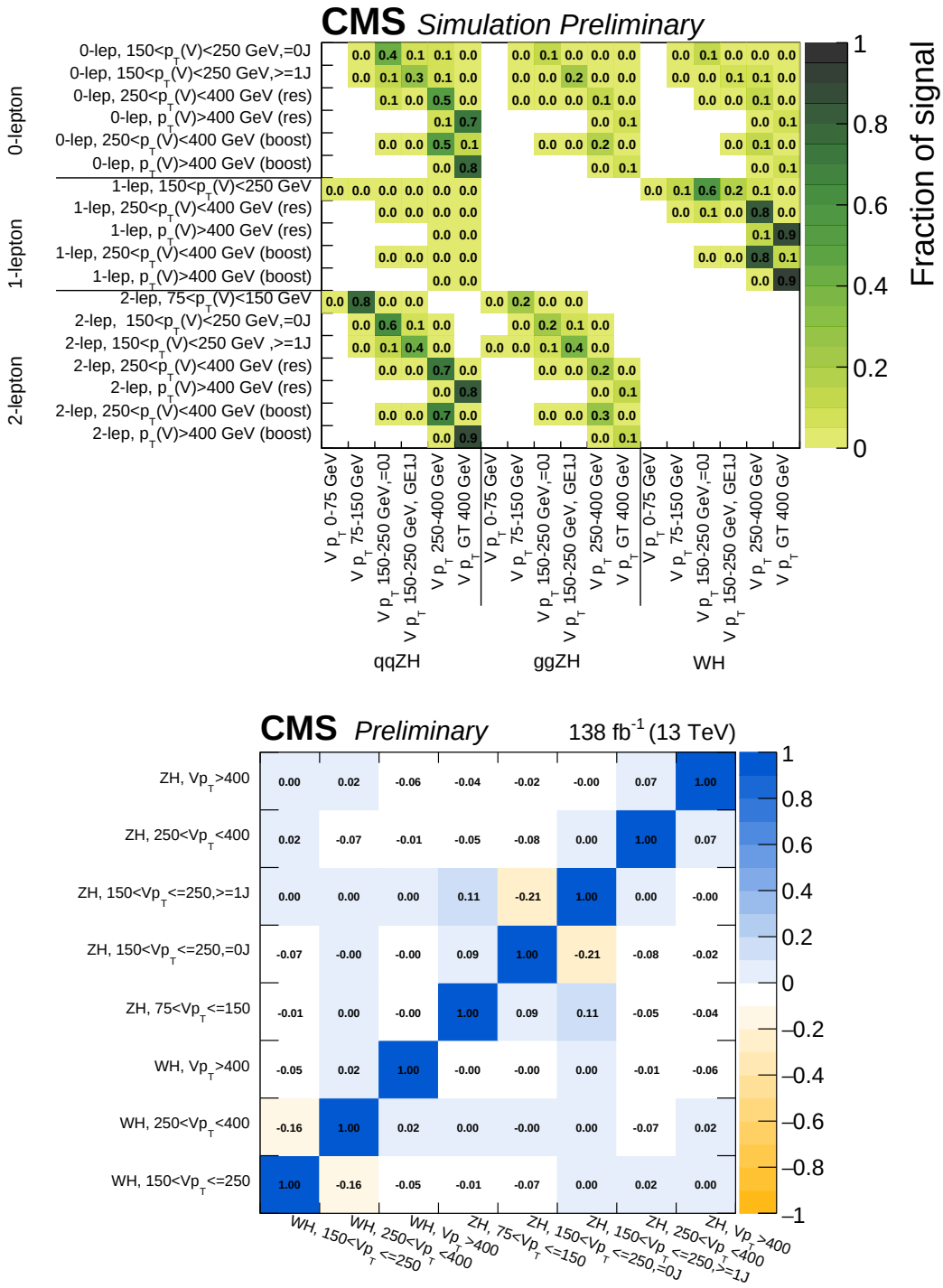


Figure 5.3: Contributions of the different STXS signal bins as a fraction of the total signal yield in each SR (upper). Correlation matrix of the parameters of interest in the STXS fit (lower).

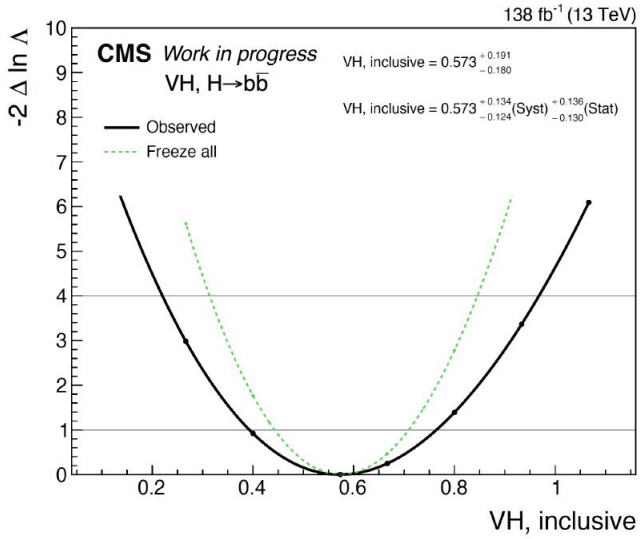


Figure 5.4: Measured inclusive signal strength from the fit.

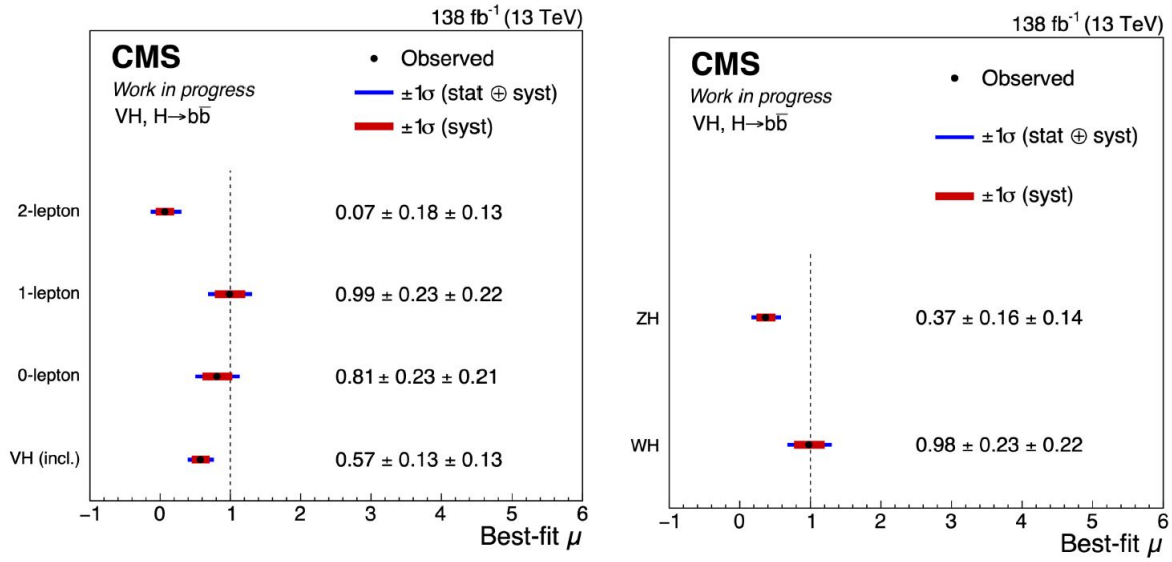


Figure 5.5: Signal strengths measured across each lepton channel (left) and split across ZH and WH channels (right).

of the V→ leptons and $H \to b\bar{b}$ in Figure 5.6 is used to interpret these results as VH production cross sections (bottom). The fit is adjusted to eliminate theoretical uncertainties that alter the inclusive cross section or the overall cross section of the individual STXS bins, which is how the results are represented as production cross sections. Additionally provided in Table 5.2 are these measured cross sections and the SM estimates. The contribution of the various sources of

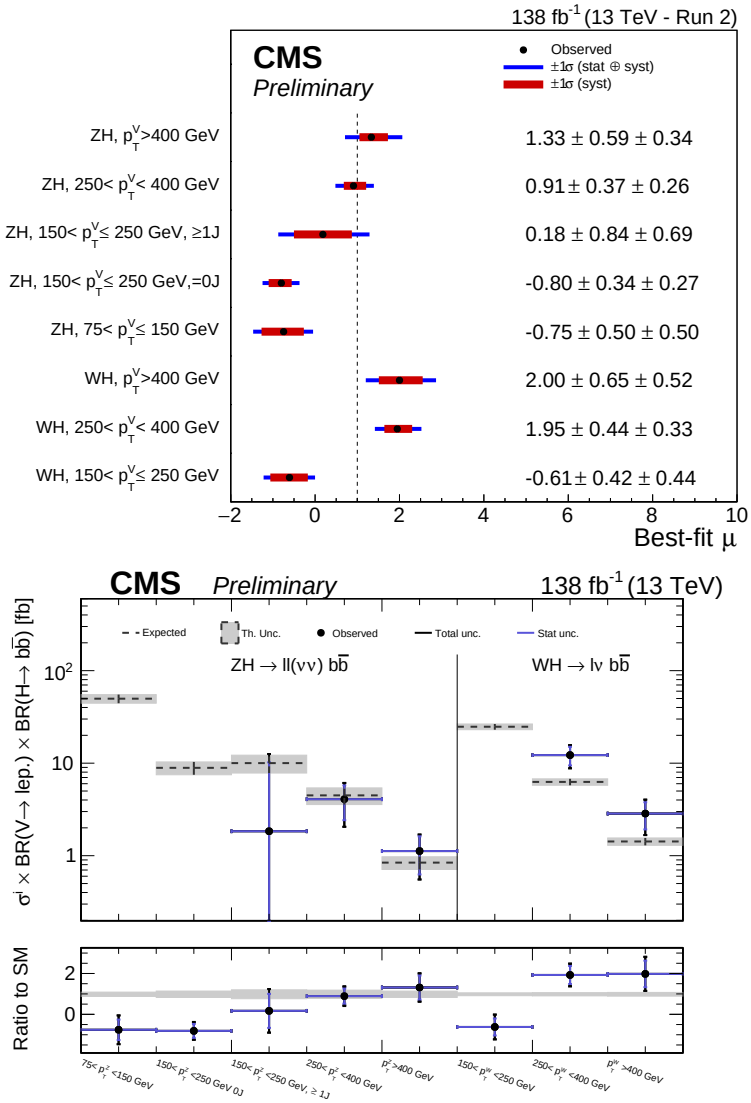


Figure 5.6: Measured STXS signal strengths from the fit (top). Measured values of $\sigma \times \mathcal{B}$ in the same STXS bins as for the signal strengths, combining all years (bottom).

systematic error to the uncertainty in the measured inclusive signal strength is shown in Table 5.3 as absolute uncertainties. The difference in quadrature between the total uncertainty in the signal strength and the uncertainty in the signal strength with the nuisance parameters of the corresponding group fixed to their best fit values is what is referred to as this contribution for a given group of uncertainties. The total systematic uncertainty is defined as the difference in quadrature between the total uncertainty in the signal strength and the total statistical uncertainty, whereas the total statistical uncertainty is defined as the uncertainty in the signal strength when all the constrained nuisance parameters are fixed to their best fit values. Following are the

STXS bin	Expected σB [fb]	Observed σB [fb]	σ/σ^{SM}
ZH $75 < p_T(Z) < 150$ GeV	50.0 ± 5.3	< 0	-0.7 ± 0.7
ZH $150 < p_T(Z) < 250$ GeV 0 jets	9.0 ± 1.4	< 0	-0.8 ± 0.4
ZH $150 < p_T(Z) < 250$ GeV ≥ 1 jets	10.1 ± 2.2	1.4 ± 10.9	0.2 ± 1.1
ZH $250 < p_T(Z) < 400$ GeV	4.5 ± 0.9	4.1 ± 2.1	0.9 ± 0.5
ZH $p_T(Z) > 400$ GeV	0.9 ± 0.1	1.2 ± 0.6	1.3 ± 0.7
WH $150 < p_T(W) < 250$ GeV	24.9 ± 1.8	< 0	-0.6 ± 0.6
WH $250 < p_T(W) < 400$ GeV	6.3 ± 0.5	12.4 ± 3.5	1.9 ± 0.5
WH $p_T(W) > 400$ GeV	1.4 ± 0.1	2.9 ± 1.2	2.0 ± 0.8

Table 5.2: The cross section values in the STXS binning for the VH process scheme multiplied by the branching fraction of $V \to$ leptons and $H \to b\bar{b}$. The SM predictions for each bin are calculated using the inclusive values reported in Ref. [21].

Source	$\Delta\mu^{\pm}$
Background (theory)	$+0.067\ -0.064$
Signal (theory)	$+0.082\ -0.060$
MC sample size	$+0.092\ -0.093$
Simulation modeling	$+0.070\ -0.066$
b-tagging	$+0.059\ -0.041$
Jet energy resolution	$+0.045\ -0.057$
Int. luminosity	$+0.041\ -0.034$
Jet energy scale	$+0.029\ -0.036$
Lepton ident.	$+0.016\ -0.002$
Trigger (MET)	$+0.001\ -0.001$

Table 5.3: Impacts of different nuisance parameter groups on the inclusive analysis signal strength.

major source of systematic uncertainties in the analysis,

- theoretical uncertainties in the signal and background components;
- limited size of simulated samples;
- simulation modeling, including uncertainty sources associated with the modeling of the V+jets background components. This modeling includes Δη(bb)-based LO-to-NLO reweighting uncertainties in the 2016 analysis, and specific ΔR(jj) corrections in NLO 2017/2018 analyses. Additionally, the p_T(V) migration uncertainties are considered in this category;
- experimental uncertainties (b-tagging, integrated luminosity, JES and JER, lepton identification, and trigger). The JES and JER components include the dedicated uncertainty on mass scale and smearing that is applied for jets subject to the b-jet energy regression.

5.1.3 Jackknife re-sampling with previous measurement

Owing to an observed inclusive μ which is incompatible ($> 2\sigma$) with the SM predictions, a Jackknife re-sampling study was performed to assess the compatibility between the current analysis and the previous analysis (observation of the H$\rightarrow b\bar{b}$ process [57]) both performed on data collected in 2017 by the CMS detector.

For the sake of simplicity, I will introduce few jargon exclusive to the CMS collaboration for naming each analysis framework and each dataset generation. The rest of the section will follow the same convention. The current analysis strategy is tagged as HIG-20-001 analysis while the previous analysis was tagged as HIG-18-016 analysis. The generation of 2017 datasets including MC samples used in HIG-18-016 is called V5 while datasets used in HIG-20-001 is termed as V11.

Jackknife re-sampling is a non-parametric cross-validation technique widely used for estimating uncertainty on a parameter by removing partitions from total event dataset. This study was performed to detect potential bias in analysis which could further explain an incompatible μ. Combining 2017 datasets from both the analyses (V5 + V11), the total collection of data is divided into g equal-sized orthogonal partitions.

For each iteration i corresponding to g Jackknife partitions, data events from the g$^{\text{th}}$ partition are excluded from the total data. After that, both the analyses are performed on that data leading up to the nominal maximum likelihood fit under both analyses strategies to obtain an inclusive μ_i from each analyses' fit. Consequently $\Delta\mu_i$ is calculated from the difference in μ from both the analyses. The variance on $\Delta\mu$ is then calculated from the variance of $\Delta\mu_i$ which is given as follows,

$$var(\Delta\mu) = \frac{g-1}{g}\sum_i(\Delta\mu_i - \overline{\Delta\mu_i})^2 = \frac{(g-1)^2}{g}var(\Delta\mu_i) \tag{5.1}$$

Finally disagreement on the two analyses is quantified by obtaining the σ on $\Delta\mu$ which is given as follows,

$$\sigma_{\Delta\mu} = \frac{\overline{\Delta\mu}}{\sqrt{var(\Delta\mu)}} = \frac{\overline{\Delta\mu}}{\frac{(g-1)}{\sqrt{g}} \times std.dev(\Delta\mu)} \tag{5.2}$$

We also obtain the correlation coefficient (ρ) on the two μ measurements for each iteration. The disagreements in the analyses in terms of $\sigma_{\Delta\mu}$ and correlation coefficient ρ are the two parameters of interests in this study.

The initial Jackknife test between HIG-18-016 and HIG-20-001 showed a correlation $\rho =$ 0.45 and calculated σ = 3.38 using equation 5.2. The results of the test are shown in figure 5.7. An intermediate analysis is carried out on V11 dataset, where features of the HIG-18-016

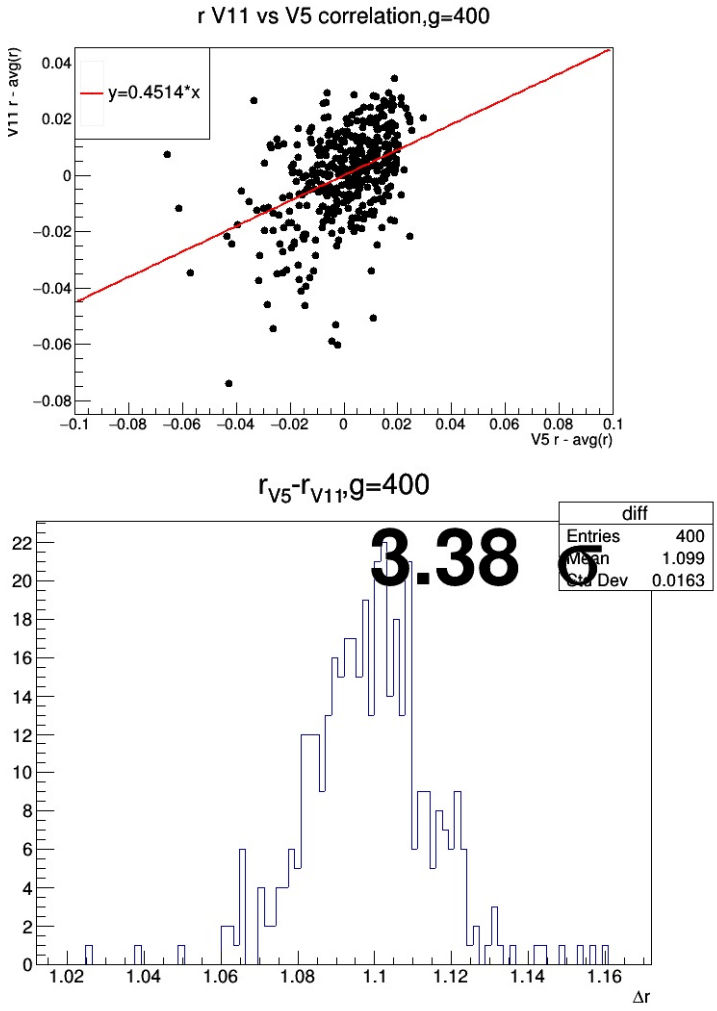


Figure 5.7: Jackknife study between HIG-20-001 and HIG-18-016 analysis.

analysis are kept while switching to some features of the HIG-20-001 analysis cumulatively in each Jackknife test. In each step, one feature of the HIG-18-016 analysis is swapped with the corresponding HIG-20-001 feature. The correlation between the intermediate analysis and the nominal HIG-18-016 analysis is obtained for each step to quantify which feature in HIG-20-001 analysis could be a potential source of mismodeling or bias in current analysis.

Owing to the difference in selections between HIG-18-016 and HIG-20-001 analysis, the created partitions may not equally impact the overall statistics for both the analyses. Figure 5.8 shows Venn diagrams of data events for both the analyses along with an intermediate analysis which is performed on V11 datasets but using HIG-18-016 analysis strategy (selection, MVAs etc.) for some of the signal regions. As seen in the figure 5.8, both the analyses have overlap events while also possessing events unique to their respective analyses. Table 5.4 shows total

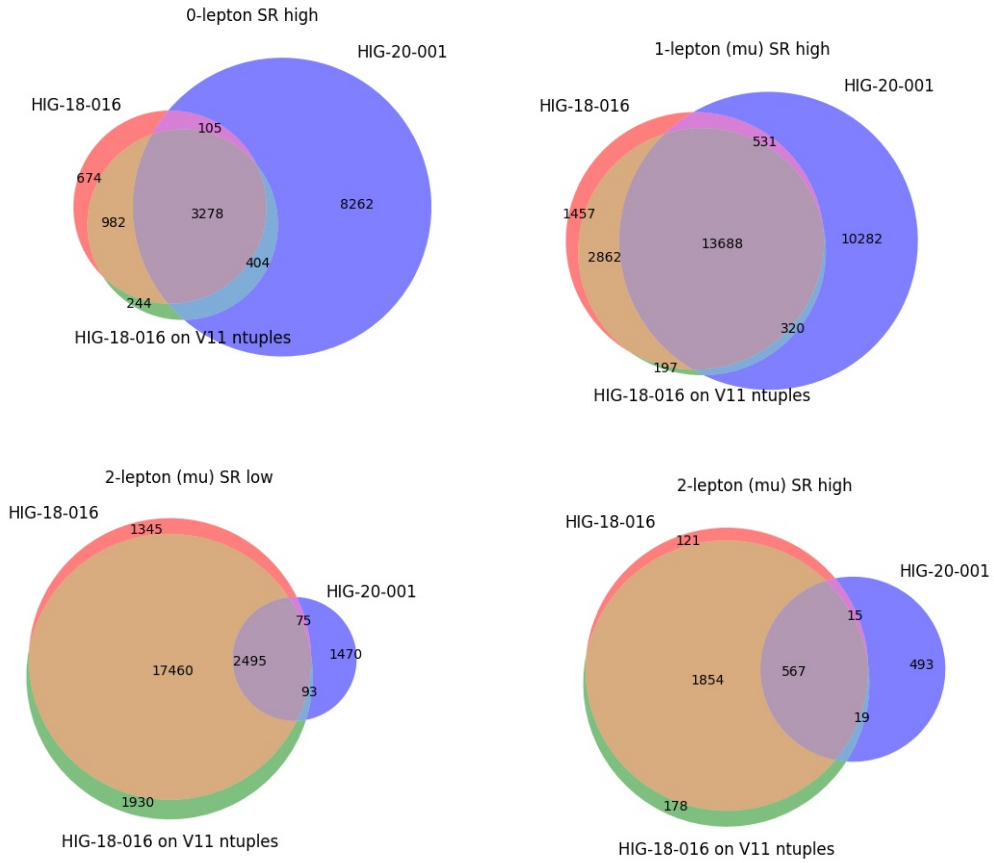


Figure 5.8: Venn diagrams of datasets used in previous analysis (HIG-18-016) and current analysis (HIG-20-001) and an intermediate analysis where HIG-18-016 style analysis (selection, MVAs etc.) is performed on 2017 dataset used in current analysis (V11) for signal regions across various channels. The numbers indicate the total data events in respective subsets.

data events in each analysis and events in the overlap of both the analysis for all the signal regions.

The intermediate analysis is constructed on V11 dataset but on HIG-18-016 analysis framework while changing few features to the HIG-20-001 analysis equivalent feature. Following HIG-20-001 features were added in each Jackknife iteration w.r.t. HIG-18-016 analysis:

- Different MC samples (already included in "intermediate" analysis).
- Division of regions by $p_T(V)$ and number of jets to correspond with STXS bins.
- Per-process scale factors and $p_T(V)$ category migrations instead of scale factors per channel and $p_T(V)$ category.

SR Event Comparison

	HIG-18-016 yield	HIG-20-001 yield	Yield of Overlap
All Regions	1230748	1063339	652031
SR Znn	4939	12049	3383
SR Wen	11268	16364	7955
SR Wmn	18538	24821	14219
SR Zee low	13325	4218	2372
SR Zee high	1826	1020	560
SR Zmm low	21375	4133	2570
SR Zmm high	2557	1094	582

Table 5.4: A comparison of total yield and overlap data events in signal regions between HIG-18-016 and HIG-20-001. HIG-20-001 signal regions are merged in $p_T(V)$ and number of jets to be consistent with HIG-18-016 binning.

- Signal region DNN binning with equal signal per bin instead of HIG-18-016 binning scheme.
- DNNs with HIG-20-001 training selection, architecture, and input variables.
- Observables for V+LF control region being $p_T(V)$ instead of b-tag score of subleading b-jet candidate, and 2 lepton V+HF control region using both jets' b-tag working point instead of b-tag score of subleading b-jet candidate.
- Updated V+Jets MC treatment. V+Jets reweighting to NLO using $p_T(V)$ in each channel, separate for HT and b-enriched samples, instead of $\Delta\eta(b\bar{b})$ NLO reweighting derived on inclusive samples. Additionally the process definitions of splitting V+Jets into different components was changed from counting generator jets containing b-hadrons in acceptance to counting b-hadrons within acceptance.
- Splitting of V+LF process to separate V+c and V+udsg processes, by assigning separate process scale factors and decorrelating nuisance parameters associated to V+LF process.
- Decorrelate process scale factors by lepton flavor
- Resolved category selection
- Addition of boosted analysis

Changes were added cumulatively to the intermediate analysis and the correlation of the inclusive signal strength with respect to the HIG-18-016 analysis was estimated with a blinded

Changes to intermediate analysis	Correlation w/HIG-18-016
HIG-18-016	1.0
HIG-20-001 MC Samples	0.9
p_T(V) and nJet Reco STXS bins	0.8
Process SFs & p_T(V) category migration unc.	0.85
HIG-20-001 SR binning scheme	0.75
HIG-20-001 Control region observables	0.80
Process SFs split by lepton flavor	0.80
HIG-20-001 V+Jets NLO reweighting	0.78
Split V+c/V+light processes	0.79
HIG-20-001 DNNs	0.54
Addition of Boosted Analysis	0.54
HIG-20-001	0.45

Table 5.5: Summary of correlations w.r.t. HIG-18-016 from Jackknife measurements.

Jackknife measurement. The full table of all Jackknife iterations involving changes to the intermediate analysis is shown in Tab. 5.5. From this study, the HIG-20-001 DNNs showed the most deviation in terms of correlation coefficient. This resulted in a cross-check study involving the DNNs of the analysis which is beyond the scope of this thesis.

5.2 Conclusions

The measurement of the cross-section of the VH(H$\rightarrow b\bar{b}$) process using the entire Run 2 dataset, which corresponds to an integrated luminosity of 138 fb^{-1}, is described in this thesis. The categories with the boosted Higgs decay topology and those where the Higgs boson is reconstructed from two resolved jets are both included in the VHbb analysis reported in this thesis. Measurements of STXS and inclusive signal strength are presented. The theoretical uncertainties from both the signal and background predictions, as well as the background MC statistical uncertainties, dominate the systematic uncertainties in the inclusive VHbb measurement.

This analysis can serve as a stepping stone for an Effective Field Theory (EFT) interpretation [67] analysis and also allows for an analysis to measure differential cross section for the VH(H$\rightarrow b\bar{b}$) process.

Some improvements could be done to this analysis in order to have a higher signal purity. One of the method could be to perform a DNN multi-classifier analysis to constrain major backgrounds instead of the selection based control regions defined and used in this analysis. Another method could be to use multi-variate methods to remove $t\bar{t}$ and single top backgrounds from the signal regions, specially relevant for 1-lepton channel.

This analysis is currently in Collaboration-wide review (CWR) in the CMS collaboration and there are efforts ongoing in the VHbb working group to understand a slightly incompatible observed VHbb μ from that of SM expected.

References

[1] "Standard Model of Elementary Particles". https://commons.wikimedia.org/wiki/File:Standard_Model_of_Elementary_Particles.svg.

[2] J. Ellis, "Higgs physics", 2013. doi:10.48550/ARXIV.1312.5672, https://arxiv.org/abs/1312.5672.

[3] R. E. Allen, "The Higgs Bridge", *Phys. Scripta* **89** (2013) 018001, doi:10.1088/0031-8949/89/01/018001, arXiv:1311.2647.

[4] CMS Collaboration, "Public results of cms luminosity information",.

[5] A. Team, "Diagram of an LHC dipole magnet. Schéma d'un aimant dipôle du LHC", 1999.

[6] CMS Collaboration, "The CMS Experiment at the CERN LHC", *JINST* **3** (2008) S08004, doi:10.1088/1748-0221/3/08/S08004.

[7] D. Abbaneo et al., "Quality control and beam test of GEM detectors for future upgrades of the CMS muon high rate region at the LHC", *JINST* **10** (2015), no. 03, C03039, doi:10.1088/1748-0221/10/03/C03039.

[8] CMS Collaboration, "The Performance of the CMS Muon Detector in Proton-Proton Collisions at $\sqrt{s} = 7$ TeV at the LHC", *JINST* **8** (2013) P11002, doi:10.1088/1748-0221/8/11/P11002, arXiv:1306.6905.

[9] T. Gleisberg et al., "Event generation with SHERPA 1.1", *Journal of High Energy Physics* **2009** (feb, 2009) 007–007, doi:10.1088/1126-6708/2009/02/007.

[10] S. Williamson, "Search for Higgs-Boson Production in Association with a Top-Quark Pair in the Boosted Regime with the CMS Experiment", 2016. presented 11 Nov 2016. doi:10.5445/IR/1000068213, http://cds.cern.ch/record/2300276.

[11] CMS Collaboration, "CMS slice raw illustrator files",.

[12] CMS Collaboration, "Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV", *JINST* **13** (2018), no. 05, P05011, doi:10.1088/1748-0221/13/05/P05011, arXiv:1712.07158.

[13] CMS Collaboration, "Heavy flavor identification at CMS with deep neural networks",.

[14] CMS Collaboration, "Machine learning-based identification of highly Lorentz-boosted hadronically decaying particles at the CMS experiment", technical report, CERN, Geneva, 2019.

[15] CMS Collaboration, "Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC", *Phys. Lett. B* **716** (2012) 30–61, doi:10.1016/j.physletb.2012.08.021, arXiv:1207.7235.

[16] F. Montiet al., "Modelling of the single-Higgs simplified template cross-sections (STXS 1.2) for the determination of the Higgs boson trilinear self-coupling", technical report, CERN, Geneva, 2022.

[17] P. Berger, "Measurement of the standard model Higgs Boson decay to b-quarks in association with a vector boson decaying to leptons, and module qualification for the CMS Phase-1 barrel pixel detector". Doctoral thesis, ETH Zurich, Zurich, 2021. doi:10.3929/ethz-b-000491182.

[18] H. Kaveh, "Simplified template cross-section measurement for Higgs boson decay to b-quarks in association with a vector boson with the full Run 2 CMS dataset". Dissertation, Universität Hamburg, Hamburg, 2022. Dissertation, Universität Hamburg, 2022. doi:10.3204/PUBDB-2022-02973.

[19] CMS Collaboration, "Jet energy scale and resolution performance with 13 TeV data collected by CMS in 2016-2018",.

[20] CMS Collaboration, "Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV", *JINST* **12** (2017), no. 02, P02014, doi:10.1088/1748-0221/12/02/P02014, arXiv:1607.03663.

[21] CERN, "Cern yellow reports: Monographs, vol 2 (2017): Handbook of lhc higgs cross sections: 4. deciphering the nature of the higgs sector", 2017. doi:10.23731/CYRM-2017-002, https://e-publishing.cern.ch/index.php/CYRM/issue/view/32.

[22] P. Higgs, "Broken symmetries, massless particles and gauge fields", *Physics Letters* **12** (1964), no. 2, 132–133, doi:https://doi.org/10.1016/0031-9163(64)91136-9.

[23] F. Englert and R. Brout, "Broken symmetry and the mass of gauge vector mesons", *Phys. Rev. Lett.* **13** (Aug, 1964) 321–323, doi:10.1103/PhysRevLett.13.321.

[24] P. W. Higgs, "Broken symmetries and the masses of gauge bosons", *Phys. Rev. Lett.* **13** (Oct, 1964) 508–509, doi:10.1103/PhysRevLett.13.508.

[25] ATLAS Collaboration, "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC", *Phys. Lett. B* **716** (2012) 1–29, doi:10.1016/j.physletb.2012.08.020, arXiv:1207.7214.

[26] J. R. Goodstein, "Ricci and levi-civita's tensor analysis paper: Lie groups: History, frontiers, and applications, volume 2. translated and edited by robert hermann. brookline, massachusetts (math. sci. press). 1975. 261 p. paper", *Historia Mathematica* **4** (1977), no. 2, 228, doi:10.1016/0315-0860(77)90126-4.

[27] Particle Data Group Collaboration, "Review of Particle Physics", *PTEP* **2020** (2020), no. 8, 083C01, doi:10.1093/ptep/ptaa104.

[28] G. S. Guralnik, C. R. Hagen, and T. W. B. Kibble, "Global conservation laws and massless particles", *Phys. Rev. Lett.* **13** (Nov, 1964) 585–587, doi:10.1103/PhysRevLett.13.585.

[29] Planck Collaboration et al., "Planck 2018 results - i. overview and the cosmological legacy of planck", *A&A* **641** (2020) A1, doi:10.1051/0004-6361/201833880.

[30] W. Adam et al., "The cms phase-1 pixel detector upgrade", *Journal of Instrumentation* **16** (feb, 2021) P02027, doi:10.1088/1748-0221/16/02/P02027.

[31] CMS Collaboration, "The CMS hadron calorimeter project: Technical Design Report", technical report, Geneva, 1997.

[32] CMS Collaboration, "Missing transverse energy performance of the CMS detector", *JINST* **6** (2011) P09001, doi:10.1088/1748-0221/6/09/P09001, arXiv:1106.5048.

[33] CMS Collaboration, "Performance of the CMS Level-1 trigger in proton-proton collisions at $\sqrt{s} = 13$ TeV", *JINST* **15** (2020), no. 10, P10017, doi:10.1088/1748-0221/15/10/P10017, arXiv:2006.10165.

[34] CMS Collaboration, "The CMS trigger system", *JINST* **12** (2017), no. 01, P01020, doi:10.1088/1748-0221/12/01/P01020, arXiv:1609.02366.

[35] J. Alwall et al., "The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations", *Journal of High Energy Physics* **2014** (jul, 2014) doi:10.1007/jhep07(2014)079.

[36] S. Frixione, P. Nason, and C. Oleari, "Matching NLO QCD computations with Parton Shower simulations: the POWHEG method", *JHEP* **11** (2007) 070, doi:10.1088/1126-6708/2007/11/070, arXiv:0709.2092.

[37] T. Sjöstrand et al., "An introduction to PYTHIA 8.2", *Comput. Phys. Commun.* **191** (2015) 159–177, doi:10.1016/j.cpc.2015.01.024, arXiv:1410.3012.

[38] GEANT4 Collaboration, "GEANT4–a simulation toolkit", *Nucl. Instrum. Meth. A* **506** (2003) 250–303, doi:10.1016/S0168-9002(03)01368-8.

[39] F. Beaudette, "The cms particle flow algorithm", doi:10.48550/ARXIV.1401.8155.

[40] D. Guest et al., "Jet flavor classification in high-energy physics with deep neural networks", *Physical Review D* **94** (dec, 2016) doi:10.1103/physrevd.94.112002.

[41] 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* **2010** (jun, 2010) doi:10.1007/jhep06(2010)043.

[42] P. Nason, "A new method for combining NLO QCD with shower monte carlo algorithms", *Journal of High Energy Physics* **2004** (nov, 2004) 040–040, doi:10.1088/1126-6708/2004/11/040.

[43] G. Luisoni, P. Nason, C. Oleari, and F. Tramontano, "HW ±/HZ 0 and 1 jet at NLO with the POWHEG BOX interfaced to GoSam and their merging within MiNLO", *Journal of High Energy Physics* **2013** (oct, 2013) doi:10.1007/jhep10(2013)083.

[44] K. Hamilton, P. Nason, and G. Zanderighi, "MINLO: multi-scale improved NLO", *Journal of High Energy Physics* **2012** (oct, 2012) doi:10.1007/jhep10(2012)155.

[45] R. Frederix and S. Frixione, "Merging meets matching in MC@NLO", *Journal of High Energy Physics* **2012** (dec, 2012) doi:10.1007/jhep12(2012)061.

[46] J. Alwall et al., "Comparative study of various algorithms for the merging of parton showers and matrix elements in hadronic collisions", *The European Physical Journal C* **53** (dec, 2007) 473–500, doi:10.1140/epjc/s10052-007-0490-5.

[47] ATLAS Collaboration, "Measurements of WH and ZH production in the $H \to b\bar{b}$ decay channel in pp collisions at 13 TeV with the ATLAS detector", *Eur. Phys. J. C* **81** (2021), no. 2, 178, doi:10.1140/epjc/s10052-020-08677-2, arXiv:2007.02873.

[48] A. J. Larkoski, S. Marzani, G. Soyez, and J. Thaler, "Soft Drop", *JHEP* **05** (2014) 146, doi:10.1007/JHEP05(2014)146, arXiv:1402.2657.

[49] S. Das, "A simple alternative to the crystal ball function", 2016. doi:10.48550/ARXIV.1603.08591, https://arxiv.org/abs/1603.08591.

[50] M. Abadi et al., "Tensorflow: Large-scale machine learning on heterogeneous distributed systems", 2016.

[51] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network", 2015.

[52] CMS Collaboration, "CMS Luminosity Measurements for the 2016 Data Taking Period", technical report, CERN, Geneva, 2017.

[53] CMS Collaboration, "CMS luminosity measurement for the 2017 data-taking period at $\sqrt{s} = 13$ TeV", technical report, CERN, Geneva, 2018.

[54] CMS Collaboration, "CMS luminosity measurement for the 2018 data-taking period at $\sqrt{s} = 13$ TeV", technical report, CERN, Geneva, 2019.

[55] R. J. Barlow and C. Beeston, "Fitting using finite Monte Carlo samples", *Comput. Phys. Commun.* **77** (1993) 219–228, doi:10.1016/0010-4655(93)90005-W.

[56] J. S. Conway, "Incorporating Nuisance Parameters in Likelihoods for Multisource Spectra", in *PHYSTAT 2011*, pp. 115–120. 2011. arXiv:1103.0354. doi:10.5170/CERN-2011-006.115.

[57] CMS Collaboration, "Observation of Higgs boson decay to bottom quarks", *Phys. Rev. Lett.* **121** (2018), no. 12, 121801, doi:10.1103/PhysRevLett.121.121801, arXiv:1808.08242.

Curriculum vitae

Saswat MISHRA

PERSONAL DATA

PLACE AND DATE OF BIRTH:	Puri	19 April 1994
ADDRESS:	Lokonath Lodge, Grand Road, Puri, Odisha- 752001
PHONE:	+91 7978049379, +385 976347842
EMAIL:	saswat.mishra@cern.ch

EDUCATION

June 2019 onwards	PhD in Particle Physics, department of physics, **Faculty of Science (PMF), University of Zagreb** Research Assistant, Laboratory for Particle Physics, **Rudjer Boskovic Institute, Zagreb, Croatia**
AUG 2017-18	Junior Research Fellow, GRAPES-3 experiment, **Tata Institute of Fundamental Research**, Mumbai
JULY 2015-17	Master of Science in PHYSICS, **Central University of Karnataka**, Gulburga
JULY 2012-15	Undergraduate Degree in PHYSICS (HONS), B.J.B Autonomous College, **Utkal University**, Bhubaneswar

APRIL 2010-12	CBSE 12^{th} at **ODM Public School**, Bhubaneswar
APRIL 2010	ICSE 10^{th} at **Blessed Sacrament High School**, Puri

WORK EXPERIENCE

June 2019 onwards (Phd work at Rudjer Boskovic Institute)

- Simplified template cross-section (STXS) measurement of higgs boson decaying into a pair of b quarks in association with a vector boson which decays leptonically in p-p collisions at $\sqrt{s}$ of 13 TeV, using full Run 2 data from the CMS experiment recorded at LHC (VHbb analysis using full Run 2 dataset)
- Central DQM shifter at the CMS experiment
- Central technical shifter at the CMS experiment
- Activities related to CMS pixel offline reconstruction

Aug 2017 - Sept 2018 (Junior Research Fellow at TIFR)

- GEANT4 simulation of the muon telescope based at GRAPES-3, which consisted of proportional counters arranged in 4 layers and responsible for studying angular orientation of the cosmic rays with the helps of secondary cosmic muons.
- GEANT simulation of GRAPES-3 based plastic scintillator detectors. A database consisting of detector response in terms of energy deposition in the scintiallators was created for different particles in the secondary cosmic rays of different momentum and oriented towards the detector at different zenith angles.
- Contribution towards development of air shower reconstruction algorithm for the GRAPES-3 experiment

Jan 2017- Apr 2017 M.Sc Thesis

The goal of my master thesis was to understand the detector response of proportional counter used in GRAPES-3 experiment. For that, detector geometry of proportional counter was simulated using GEANT4 and response of the detector in terms of energy deposition was studied

with the help of cosmic muons generated using CORSIKA air shower generator. The studies were presented at Central University of Karnataka, as an academic requirement for the M.Sc degree

POSTER & ORAL PRESENTATIONS

OCT 2022 Oral presentation on "**VH(bb) Simplified template cross section measurements with the CMS experiment**" at LHC Days in Split, 2022

JULY 2022 Oral presentation on "**Cross-section measurement of the VH $\rightarrow b\bar{b}$ process at the CMS experiment**" at XX FRASCATI SUMMER SCHOOL "BRUNO TOUSCHEK" in Nuclear, Subnuclear and Astroparticle Physics

MAY 2018 Poster presentation on "**GRAPES-3 shower reconstruction**" at DHEP annual meet 2018, Tata Institute of Fundamental Research

MAY 2018 Poster presentation on "**GEANT4 simulation of GRAPES-3 scintillator detector**" at DHEP annual meet 2018, Tata Institute of Fundamental Research

DEC 2016 Poster Presentation on "**Performance Metrics of a GPU Based Track Fitting Code for the INO Prototype Stack**" at XXII DAE-BRNS High Energy Physics Symposium -2016

NOV 2016 Poster Presentation on "**A Simulation Study of Beam Parameters in Proton Radiography Using Energy Measurements**" at AMPICON-2016

PROJECTS AND WORKSHOPS

FEB 2022	International Meeting on EFFECTIVE PATHWAYS TO NEW PHYSICS (IMEPNP), Institute of Physics, Bhubaneswar, Odisha
JUNE 2021	Thematic CERN School of Computing, Online
DEC 2019	Combine workshop and tutorial, CERN
SEPT 2019	CMS Physics Object School (CMSPOS), RWTH, Aachen, Germany
DEC 2016	Winter school in astroparticle physics (WAPP), Cosmic Ray Laboratory, TIFR, Ooty, Tamil Nadu
AUG-DEC 2015	P.G Diploma Course in Remote sensing and Digital Image Processing under **IIRS**, Dehradun
FEB 2015	Project on "Temperature Dependence on Resistance of a hot filament"as per disciplinary course under B.Sc Course

COMPUTER SKILLS

Software Skills:	ROOT, GEANT4, Higgs Combine, Nvidia Cuda, MS Office
Programming Skills:	C, C++ and python

LANGUAGES

ODIA:	Mothertongue
ENGLISH:	Fluent
HINDI:	Fluent

INTERESTS

Higgs physics, BSM physics, effective field theory, detector instrumentation and calibration

REFEREES

Dr. Vuko Briglevic
Senior Scientist,
Head of Laboratory for Particle Physics,
Laboratory for Particle Physics,
Rudjer Boskovic Institute,
email:- Vuko.Brigljevic@cern.ch

Dr. Dinko Ferencek
Senior Research Associate,
Laboratory for Particle Physics,
Rudjer Boskovic Institute,
email:- Dinko.Ferencek@cern.ch

Dr. Pravata K. Mohanty
Reader,
Tata Institute of Fundamental Research,
Mumbai
email:- pravata2006@gmail.com

Dr. Deepak Samuel
Associate Professor,
Central University of Karnataka,
Gulbarga
email:- deepaksamuel@cuk.ac.in