# Search for new scalar resonances using proton-proton collisions recorded by the CMS experiment at the Large Hadron Collider

**Roguljić, Matej**

University of Zagreb

FACULTY OF SCIENCE

Matej Roguljić

# Search for new scalar resonances using proton-proton collisions recorded by the CMS experiment at the Large Hadron Collider

DOCTORAL DISSERTATION

Zagreb, 2022.

University of Zagreb

FACULTY OF SCIENCE

Matej Roguljić

# Search for new scalar resonances using proton-proton collisions recorded by the CMS experiment at the Large Hadron Collider

DOCTORAL DISSERTATION

Supervisor:
Devdatta Majumder, PhD

Zagreb, 2022.

Sveučilište u Zagrebu

PRIRODOSLOVNO MATEMATIČKI FAKULTET

Matej Roguljić

# Potraga za novim skalarnim rezonancijama u sudarima protona zabilježenima detektorom CMS na velikom hadronskom sudarivaču

DOKTORSKI RAD

Mentor:
dr. sc. Devdatta Majumder

Zagreb, 2022.

## Supervisor's biography

Devdatta Majumder received his doctorate degree in high energy physics from the Tata Institute of Fundamental Research in Mumbai, India, in 2011. After finishing his doctorate degree, he joined the National Taiwan University in Taipei, Taiwan, as a postdoctoral fellow. In 2014, he joined the University of Kansas in Lawrence, USA, as a research associate until 2019 when he accepted the staff scientist position at the assistant professor level at the Ruđer Bošković Institute in Zagreb, Croatia.

Devdatta's research focuses on the high energy physics measurements using the LHC data collected by the CMS experiment. He participated in the searches for vector-like quarks, the searches for physics beyond the Standard model using Higgs bosons and the development of the Lorentz-boosted Higgs boson identification algorithms. He was also involved in the SM study of W$\gamma$ events at the LHC and the measurement of its inclusive production cross section with first LHC data.

# Abstract

In this thesis, the search for the resonant production of a new massive scalar particle X decaying into a new light scalar particle Y and a standard model Higgs boson H through the process $X \to YH \to b\bar{b}b\bar{b}$ is presented. Data from CERN LHC proton-proton collisions at a centre-of-mass energy of 13 TeV are used, collected by the CMS detector in 2016–2018 and amounting to 138 $fb^{-1}$ of integrated luminosity. The search is performed in mass ranges of X (0.9–4 TeV) and Y (60–600 GeV) where both the Y and the H are highly Lorentz-boosted. In this kinematic regime, their b quark-antiquark daughter particles are collimated enough to allow the reconstruction of H and Y using single large-area jets each. The mass of one of the large-area jets is required to be compatible with that of the Higgs boson, which is 125 GeV. A scan is performed in a two dimensional plane spanned by the mass of the other jet, associated to Y, and the invariant mass of both large-area jets used to reconstruct X. The results are interpreted in the context of scalar resonances predicted in the next-to-minimal supersymmetric standard model and also in an extension of the standard model with two additional singlet scalar fields. Upper limits are placed on the production cross section of the process as a function of the masses of X and Y in the 0.1–150 fb range. This is the first search for this process using Lorentz-boosted event topologies and significantly extends the constraints on the studied model.

A novel calibration method of the ParticleNet algorithm designed to recognize the decays of a massive particle into a pair of b quarks is also presented. The calibration consists of selecting events with a large-area jet with high momentum and measuring the strength of the Z boson peak on a smooth hadronic background. Previously used method performed the calibration on a set of jets originating from gluon fragmentation into $b\bar{b}$, with special selection applied to make them more akin to the jets originating from the decays of a massive particle, such as Z or H. The presented measurement demonstrates the possibility of a direct calibration of the tagger and provides a validity test for previously used, indirect measurements.



Keywords: LHC, CMS, standard model, Higgs boson, NMSSM, TRSM, boosted objects, heavy-flavor tagging, ParticleNet

# Prošireni sažetak rada

## Uvod

Fizika elementarnih čestica bavi se najmanjim, nedjeljivim česticama. Ona teži identificiranju takvih čestica i proučavanju njihovih interakcija. Standardni model (SM) fizike elementarnih čestica je teorija koja opisuje sve poznate elementarne čestice i tri od četiri fundamentalne sile: elektromagnetsku, slabu i jaku. Unatoč mnogim pokušajima, gravitacijska sila još uvijek nije uspješno opisana matematičkim aparatom SM-a, kvantnom teorijom polja. Međutim, gravitacijska je sila zanemariva na skalama elementarnih čestica stoga SM može iznimno precizno opisati interakcije među česticama koristeći tri spomenute sile.

Unutar SM-a, nalazi se 12 čestica materije, zvanih fermioni, i 5 čestica medijatora sila, bozoni. Od pet bozona, njih se četiri povezuju s tri fundamentalne sile, dok je peti bozon, Higgsov bozon (H), ključan za uvođenje mehanizma pomoću kojeg čestice dobivaju masu. Skoro je pedeset godina prošlo od postuliranja Higgsovog bozona do njegovog otkrića koje predstavlja veliki trijumf SM-a. Također su i mnoga druga mjerenja u fizici elementarnih čestica potvrdila slaganje između predviđanja SM-a i eksperimenta. Ipak, postoji nekoliko eksperimentalnih rezultata koji ukazuju na nedostatke teorije. Stoga se predlažu teorije izvan standardnog modela (engl. Beyond the standard model, BSM) koje daju objašnjenje za neka od tih opažanja. Jedan od problema SM-a je već spomenuta nemogućnost opisivanja gravitacije postojećom teorijom. Također, astronomska mjerenja zaključuju da se tek 5% ukupne energije u svemiru može pripisati sadržaju SM-a. Oko 26% energije bi trebala činiti tamna materija koja djeluje jedino gravitacijskom i potencijalno slabom silom. Trenutna teorija ne predlaže česticu koja bi mogla biti kandidat za tamnu materiju. Preostalih 69% energije u svemiru pripisuje se pozadinskoj energiji vakuuma. Predviđanje energije vakuuma SM-a je 120 redova veličine veće od vrijednosti dobivene astronomskim opažanjima. Osim eksperimentalnih rezultata, neka svojstva teorije ukazuju na moguće probleme u trenutnom opisu fizike. Jedan od njih je izračun mase Higgsovog bozona. U tom se izračunu javljaju jake kvantne korekcije te se vrijednosti parametara teorije moraju precizno namjestiti da se računom dobije izmjerena masa. Drugi je primjer ujedinjenje sila. Vrijednosti parametara koje određuju jakosti sila ovise o energijskoj skali interakcije

te se međusobno približavaju na viskom energijama. To daje naznaku da bi se na nekoj (visokoj) energijskoj skali tri sile mogle ujediniti što se ne događa u SM-u.

Jedan od primjera BSM teorije je minimalni-supersimetrični standardni model (MSSM). U njemu se prirodno ponište velike kvantne korekcije na masu Higgsovog bozona i postiže se izjednačavanje jakosti sila na visokim energijama. Također, ta teorija predviđa i postojanje čestice koja bi mogla odgovarati tamnoj materiji. Navedena svojstva teorije su pružila snažnu motivaciju za brojnim potragama za česticama koje MSSM predviđa. Budući da model predviđa dodatne čestice u Higgs sektoru, neke od tih potraga bile su za rezonancijama spina 0. Nijedno mjerenje do sada nije pronašlo novu česticu te su neki dijelovi faznog prostora parametara MSSM-a isključeni. Sljedbenik MSSM-a je sljedeći-minimalni-supersimetrični standardni model (engl. next-to-minimal supersymmetric standard model, NMSSM) koji je predložen jer pruža objašnjenje za, naoko "neprirodnu", vrijednost jednog od parametara kojeg uvodi MSSM, tzv. $\mu$ problem. NMSSM sadrži 7 bozona u Higgsovom sektoru (jedan od njih odgovara Higgsovom bozonu unutar SM-a). U tom se modelu mogu dogoditi raspadi teške skalarne čestice, X, u drugu, lakšu skalarnu česticu Y i Higgsov bozon, X$\rightarrow$YH. Asimetrični raspad teške rezonancije X je zanimljiv jer NMSSM predviđa da vezanja čestice Y na čestice SM-a mogu biti potisnuta te bi X$\rightarrow$YH mogao biti vodeći način produkcije čestice Y. Također, takvi su procesi još uglavnom neistraženi. Najveći omjer grananja za H i za Y, dok je masa Y manja od dvostruke mase t kvarka, je u par b kvarkova stoga se u ovom radu predstavlja potraga za $X \rightarrow YH \rightarrow b\bar{b}b\bar{b}$ procesom. Potraga se fokusira na ultra-relativističku topologiju gdje H i Y poprimaju velike količine gibanja zbog čega su produkti njihovih raspada međusobno bliski. H i Y su tada svaki rekonstruirani kao jedan široki hadronski mlaz. Cilj ove potrage je pronaći signal takvog procesa ili postaviti gornje granice na udarni presjek na širokom rasponu masa X i Y, gdje su ultra-relativističke topologije moguće.

**Eksperimentalni postav**

Međudjelovanja elementarnih čestica promatraju se ubrzavanjem čestica i njihovim sudaranjem čime se one potiču na interakcije. Trenutno najveći ubrzivač čestica je Veliki Hadronski Sudarivač (engl. Large Hadron Collider, LHC). LHC je kružni ubrzivač koji dovodi dva snopa protona do energije od 6.5 TeV u suprotnim smjerovima, odnosno sudara protone na energiji od 13 TeV u sustavu centra mase. Sudari se događaju na četiri točke

koje odgovaraju pozicijama četiriju velikih detektora: ATLAS, CMS, ALICE i LHCb.

U ovom se radu koriste podaci prikupljeni CMS detektorom u vremenskom periodu 2016–2018 koji odgovaraju 138 $\text{fb}^{-1}$ integriranog luminoziteta. CMS detektor je jedan od dva detektora opće namjene (drugi je ATLAS) na LHC-u. CMS se sastoji od više pod-detektorskih sustava koji su posloženi u slojevima oko točke sudara protona. Prvi se sustav bavi detekcijom tragova nabijenih čestica. Nakon njega se nalaze redom elektromagnetski kalorimetar, u kojem se zaustavljaju elektroni i fotoni, i hadronski kalorimetar, u kojem se zaustavljaju hadroni. Navedeni se sustavi nalaze unutar magnetske zavojnice načinjene od supravodljivog materijala koja pruža homogeno magnetsko polje od 3.8T. Detektorski sustav najudaljeniji od točke sudara su mionske komore koje se nalaze izvan magneta stoga je povratno magnetsko polje nešto slabije nego u ostalim dijelovima detektora. Do njih dopiru samo mioni i neutrini. Budući da je učestalost međudjelovanja neutrina s detektorskim materijalima iznimno mala, u njima se detektiraju samo mioni.

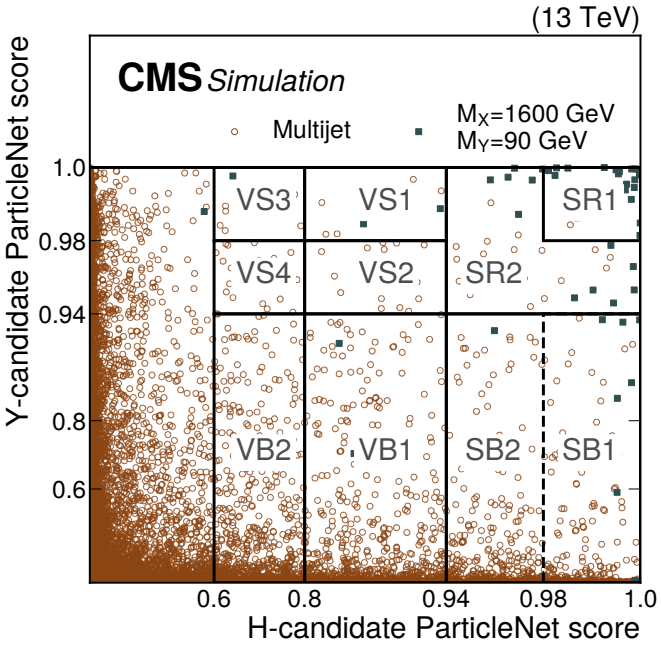
Važno oruđe u analizi podataka na CMS-u je i simulacija samoga detektora. Vrsta i energija čestica koje nastaju u sudarima mogu se predvidjeti koristeći SM ili jednu od BSM teorija. Propagacija tih, "izlaznih", čestica kroz detektor simulira se pomoću Geant4 alata čime se naposljetku dobije procjena očekivanih rezultata mjerenja. Izmjereni se rezultati zatim mogu usporediti s očekivanjima u SM ili BSM teorijama.

**Analiza podataka**

Selekcija događaja u analizi podataka temelji se na očekivanoj topologiji signalnog procesa. Potpis signalnog procesa u ovoj analizi je prisutnost dva široka hadronska mlaza visoke energije. Procesi SM-a koji mogu producirati događaje nalik signalnima zovemo pozadinom. Značajan doprinos ovoj analizi čine više-mlazni događaji nastali procesima kvantne kromodinamike (QCD događaji) i produkcija para t kvarkova ($\text{t}\bar{\text{t}}$ događaji).

Osim zahtjeva za dvama mlazovima visokih energija, u selekciji događaja također se koristi i činjenica da jedan od mlazova dolazi od raspada Higgsovog bozona čija je masa poznata. Stoga je postavljen i uvjet na postojanje hadronskog mlaza čija je masa u intervalu $[110, 140]$ GeV. Taj se mlaz naziva H kandidatom, a drugi mlaz Y kandidatom. Posljednji korak u selekciji događaja je primjena ParticleNet algoritma za prepoznavanje mlazova čija su svojstva konzistentna s hipotezom da potječu iz raspada teške čestice u

$b\bar{b}$ par. Njegova primjena značajno poboljšava omjer očekivanih doprinosa signala i pozadine. Primjer raščlanjivanja pozadine i signala uporabom ParticleNet algoritma prikazan je na Slici 1. Potraga za signalom se odvija u signalnim regijama označenim SR1 i SR2.




Slika 1: Distribucija vrijednosti ParticleNet algoritma H i Y kandidata za signalne ($M_{\mathrm{X}}$ =1600 i $M_{\mathrm{Y}}$ =125GeV) i QCD događaje.

Varijable koje se koriste za potragu su invarijantna masa dvaju mlazova, $M_{\mathrm{JJ}}$, i masa Y kandidata, $M_{\mathrm{J}}^{\mathrm{Y}}$. Za signalni događaj, one odgovaraju masama X i Y čestica pa će prisutnost signala u $M_{\mathrm{J}}^{\mathrm{Y}}$–$M_{\mathrm{JJ}}$ ravnini biti vidljiva kao nakupina (višak) događaja oko ($M_{\mathrm{Y}}$,$M_{\mathrm{X}}$) koordinate.

Za procjenu QCD pozadine koristi se metoda koja se zasniva na korištenju izmjerenih podataka u područjima u kojima dominiraju QCD događaji, na Slici 1 označeni SB1 i SB2. Budući da u tim područjima dominantan doprinos daju QCD događaji, procjena oblika QCD distribucija u tim područjima iz podataka je veoma pouzdana. Nadalje, definira se prijenosna funkcija koja odgovara omjeru QCD distribucija u SR1(2) i SB1(2) područjima čiji je oblik nepoznat te se isti određuje prilikom prilagodbe modela na podatke. Pretpostavka metode je da su oblici $M_{\mathrm{J}}^{\mathrm{Y}}$–$M_{\mathrm{JJ}}$ distribucija QCD događaja u SR1(2) i SB1(2) slični stoga se prijenosna funkcija može opisati jednostavnim umnošcima polinoma varijabli $M_{\mathrm{J}}^{\mathrm{Y}}$ i $M_{\mathrm{JJ}}$. Ova metoda, koja se zasniva na podacima, pouzdanije procjenjuje QCD distribucije jer trenutne simulacije QCD događaja koriste najniži red računa smetnje. Također, ParticleNet algoritam snažno filtrira mlazove nastale QCD procesima pa

je količina simuliranih QCD događaja u signalnim regijama nedovoljna za procjenu QCD distribucija.

Nasuprot tome, simulacije $t\bar{t}$ događaja koriste prvi viši red računa smetnje stoga se mogu koristiti za procjenu doprinosa pozadini. Međutim, usporedno uz prilagodbu modela na podatke u signalnim područjima, mjerimo i korekcije na simulaciju $t\bar{t}$ događaja koristeći polu-leptonski kanal $t\bar{t}$ raspada. Takvi se događaji izoliraju traženjem triju znakova leptonskog raspada t kvarka: prisutnost leptona (elektron ili mion) visoke energije, hadronski mlaz koji dolazi od raspada b kvarka i nedostajuća transverzalna energija koja se pripisuje neutrinu. Ako su te tri sastavnice pronađene, na suprotnoj se strani detektora traga za postojanjem hadronskog mlaza visoke energije za kojeg se tada sa visokom sigurnošću može reći da potječe od hadronskog raspada t kvarka. Takvi se mlazovi nazivaju "ispitni" mlazovi jer njima ispitujemo svojstva hadronskog raspada t kvarka. Prilagođavanjem spektra mase simuliranih ispitnih mlazova na podatke, dobivaju se korekcijski faktori na simulaciju koji se tada mogu koristiti za ispravljanje simulacije $t\bar{t}$ pozadine u signalnim područjima.

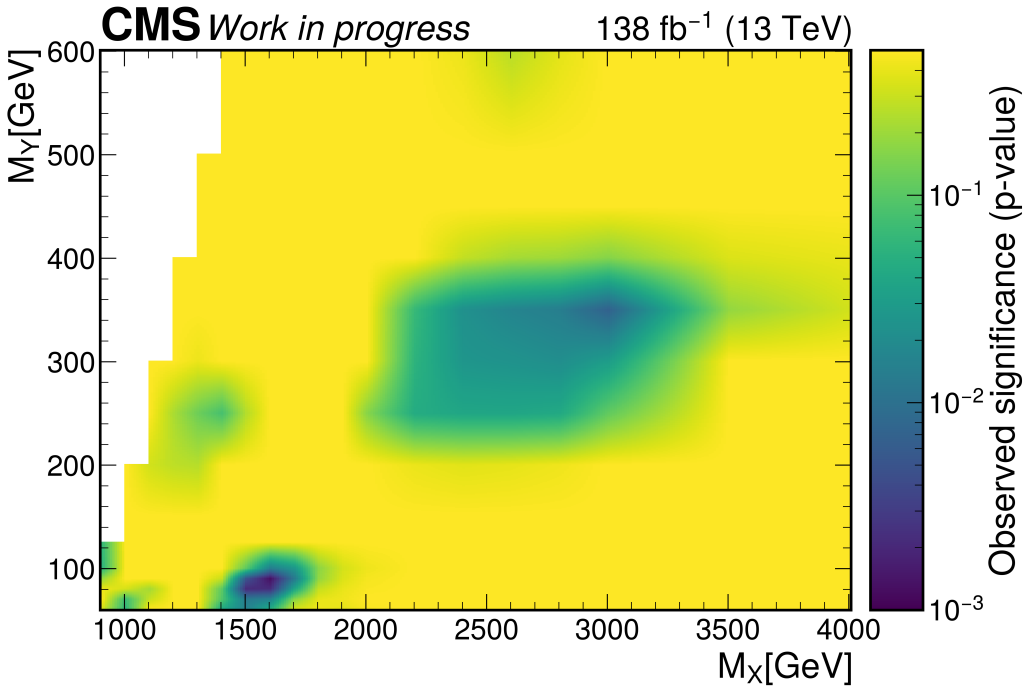
Očekivana distribucija signala dobivena je simulacijom. Generacija događaja napravljena je unutar NMSSM modela. Međutim, kinematički parametri modela su vrlo generični te se rezultati mogu interpretirati i u kontekstu drugih modela koji predviđaju promatrani kaskadni raspad.

Prilagodba modela na podatke istodobno se izvršava u signalnim i polu-leptonskim područjima analize koristeći funkciju maksimalne vjerodostojnosti. Sistematske nesigurnosti na signal i pozadinske procese su uključene u statistički model. Sistematska nesigurnost s najvećim utjecajem jest nesigurnost efikasnosti ParticleNet algoritma na signalne mlazove. To je vodeća nesigurnost jer je ukupna efikasnost selekcije za signalne događaje otprilike kvadratno proporcionalna s efikasnošću ParticleNet algoritma (tražimo da dva mlaza zadovoljavaju prag ParticleNet algoritma).

Prije prilagodbe na podatke u signalnim područjima, učinkovitost metode ispitana je i potvrđena na kontrolnim područjima (koja su označena kao VS1 i VS2 na Slici 1) u kojima su puno manji omjeri očekivanog signala i pozadine.

**Rezultati**

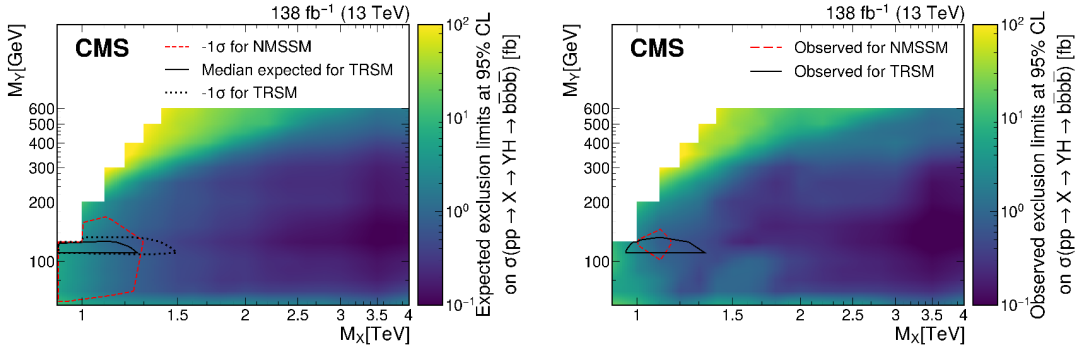
Nakon potvrđivanja učinkovitosti modela procjenjivanja pozadine, isti je "zamrznut" te je napravljena prilagodba modela na podatke u signalnim područjima. Područja u $M_{\mathrm{J}}^{\mathrm{Y}}$–$M_{\mathrm{JJ}}$ ravnini gdje se pojavljuje višak zabilježenih događaja u odnosu na procijenjenu pozadinu se ispituju kao naznaka signalnih događaja. Za kvantificiranje moguće prisutnosti signala koristi se test značajnosti koji uspoređuje vrijednost maksimalne vjerodostojnosti dviju hipoteza. Prva se vrijednost dobije prilagodbom u kojoj je jakost signala stavljena na nulu (hipoteza u kojoj signalni proces ne postoji), a druga se vrijednost dobijem stavljanjem jakosti signala na onu vrijednost koja maksimizira funkciju vjerodostojnosti. Slika 2 prikazuju dvo-dimenzionalni prikaz značajnosti signala, izražen u $p$-vrijednostima.




Slika 2: Značajnosti opaženih ekscesa u podacima, izražene kao $p$-vrijednost u $M_{\mathrm{X}}$–$M_{\mathrm{Y}}$ ravnini. Hipoteze signala za koje nije zabilježen višak događaja imaju $p$-vrijednost od 0.5. Najniža opažena $p$-vrijednost je $1.1 \times 10^{-3}$ za $M_{\mathrm{X}} = 1600\mathrm{GeV}$, $M_{\mathrm{Y}} = 90\mathrm{GeV}$.

Kombinacija masa koja dovodi do najveće vrijednosti značajnosti signala jest $M_{\mathrm{X}} = 1.6\mathrm{TeV}$ i $M_{\mathrm{Y}} = 90\mathrm{GeV}$. Ta hipoteza opažena je s $3.1\sigma$ značajnosti ($p$-vrijednost = 0.001). Međutim, moramo uzeti u obzir to da se u potrazi istovremeno traga za 260 različitih kombinacija masa signala. Očekivano je da će se opaziti povećana razina značajnosti signala za neke signalne točke zbog statističkih fluktuacija pozadine. Relevantna veličina za otkriće novog signala je "globalna" značajnost koja se definira kao vjerojatnost opažanja

$p$-vrijednosti jednake ili manje od najmanje opažene na bilo kojoj promatranoj kombinaciji masa, pod pretpostavkom da signal ne postoji. Globalna značajnost za opaženi rezultat ovoga mjerenja iznosi $0.7\sigma$ ($p$-vrijednost=0.25) što ukazuje na to da hipoteza koja uključuje samo pozadinu dobro opisuje podatke. Također se računaju i gornje granice na udarni presjek signala u ovisnosti o $M_{\mathrm{X}}$ i $M_{\mathrm{Y}}$, prikazane na Slici 3.




Slika 3: Očekivane (lijevo) i opažene (desno) gornje granice udarnog presjeka procesa $\sigma(\mathrm{pp} \to \mathrm{X} \to \mathrm{YH} \to \mathrm{b\bar{b}b\bar{b}})$ sa sigurnošću intervala pouzdanosti od 95% za različite vrijednost $M_{\mathrm{X}}$ i $M_{\mathrm{Y}}$. Područja unutar crvenih i crnih kontura predstavljaju područja gdje je opažena granica na udarni presjek niža od dopuštenih udarnih presjeka u dvije teorije.

Gornje se granice na udarni presjek signalnog procesa postavljaju korištenjem $CL_S$ metode sa sigurnošću intervala pouzdanosti od 95%. Dobivene gornje granice uspoređene su s najvećim dopuštenim udarnim presjecima u NMSSM i teoriji SM-a proširenog s dva skalarna singleta (engl. Two-real-scalar-singlet extension of the SM, TRSM). Područja gdje su opažene gornje granice udarnog presjeka niže od predviđenih u teoriji dovodi do isključenja dijela faznog prostora parametara teorije.

Za NMSSM, ne postoji raspon masa gdje je očekivano isključenje parametara. Međutim, opažene gornje granice dovode do isključenja u rasponu 1.00–1.15 TeV za $M_{\mathrm{X}}$ i 101–145 GeV za $M_{\mathrm{Y}}$. Za TRSM, očekivano područje isključenja je u granicama $0.90 < M_{\mathrm{X}} < 1.26$TeV i $100 < M_{\mathrm{Y}} < 126$GeV, dok je opaženo unutar $0.95 < M_{\mathrm{X}} < 1.33$TeV i $110 < M_{\mathrm{Y}} < 132$GeV.

### Kalibracija algoritma za označavanje

U ovom je radu predstavljena i nova metoda kalibracije ParticleNet algoritma koristeći raspade Z bozona na dva b kvarka. Za kalibraciju efikasnosti algoritma koji označava

mlazove koji potječu iz raspada teške čestice u dva b kvarka, potrebno je izolirati takve mlazove u podacima. To je problematično zbog visoke pozadine mlazova koji potječu iz QCD interakcija. Stoga su dosadašnje metode kalibracije koristile upravo mlazove iz QCD interakcija uz posebnu selekciju koja odabire mlazove čija su svojstva bliska svojstvima signalnih mlazova. Metoda predstavljena u ovom radu izvršava kalibraciju direktno na signalnim mlazovima, demonstrirajući mogućnost takvog mjerenja, te može poslužiti kao potvrda dosadašnjih mjerenja.

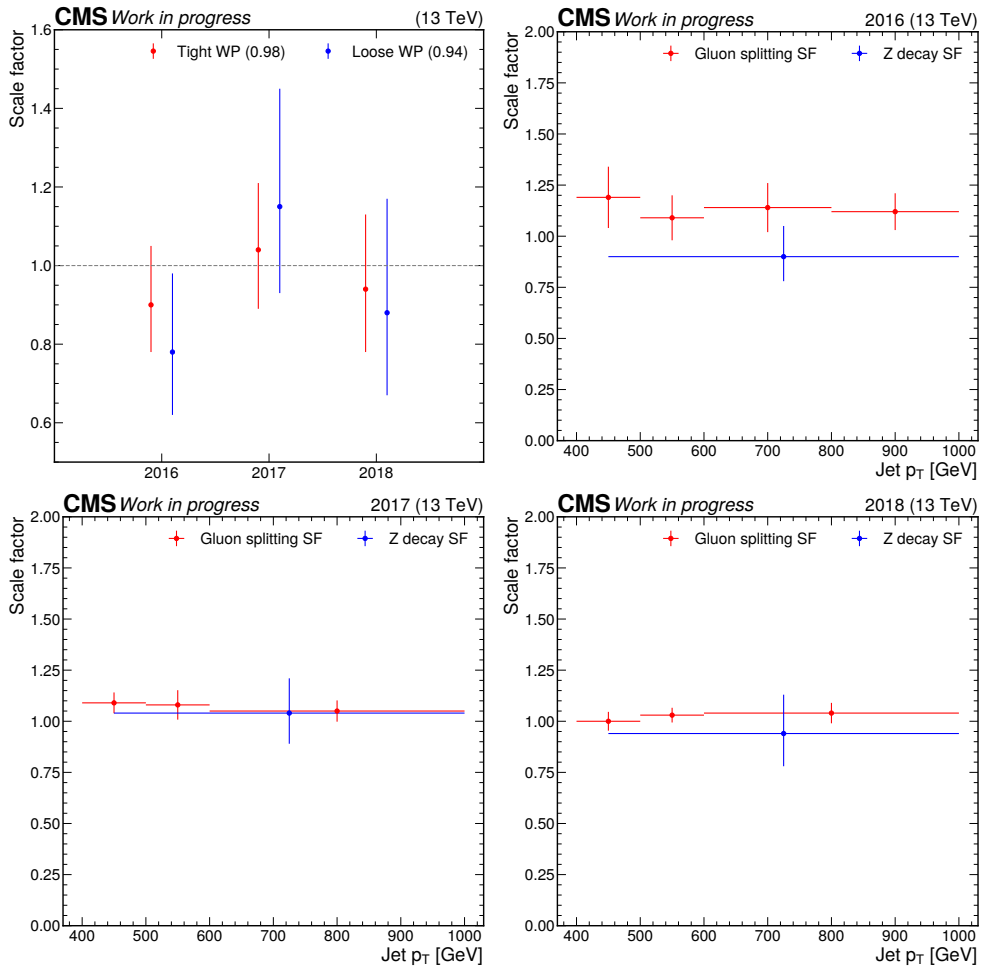
Za mjerenje su odabrani događaji gdje vodeći (po transverzalnom impulsu, $p_T$) široki mlaz ima visoki $p_T$, $>$ 450GeV, dok je zahtjev na $p_T$ drugog mlaza nešto manji $p_T$ >200GeV. Zahtjev na $p_T$ drugog mlaza stavljen je da se odaberu primarno dvo-mlazni događaji. Također se primjenjuju i dva veta. Prvi je veto na prisutnost elektrona ili miona u događaju, a drugi je veto na prisutnost hadronskog mlaza, nasuprot vodećem, koji potječe od hadronizacije b kvarka. Potonji veto je stavljen s ciljem smanjenja doprinosa $t\bar{t}$ produkcije pozadini. Naposljetku se definiraju dva područja, područje "prolaza" i "pada", gdje vodeći mlaz prolazi ili pada postavljenju radnu točku ParticleNet algoritma. Prolazno područje je obogaćeno signalom, dok je područje pada u potpunosti dominirano QCD događajima. Kalibracija je napravljena za dvije radne točke, zasebno: 0.94 i 0.98. Te su iste radne točke korištene u glavnoj analizi ovoga rada.

Kalibracija se sastoji od mjerenja Z rezonancije u ravnini koju čine masa i transverzalni impuls vodećih mlazova u odabranim događajima. Mlazovi nastali raspadom Z u dva b kvarka vidljivi su kao rezonancija u varijabli mase, oko 90GeV (masa Z bozona). Pozadinu većinom čine QCD mlazovi te u manjoj mjeri mlazovi nastali raspadom W bozona. QCD pozadina se, kao i u glavnoj analizi predstavljenoj u ovom radu, procjenjuje iz podataka koristeći prijenosnu funkciju koja procjenjuje doprinos QCD pozadine u prolaznom području iz područja pada. Doprinos pozadini koji potječe od W mlazova procijenjen je iz simulacije.

Simulacija W i Z događaja koristi najniži red računa smetnje jer jedino takva simulacija pruža dovoljno velik broj simuliranih događaja gdje bozoni imaju visoki $p_T$. Međutim, za ovu je kalibraciju važno imati što točnije teorijsko modeliranje signala jer se indirektno mjeri omjer efikasnosti ParticleNet algoritma u podacima i simulaciji. Naime, za direktno mjerenje efikasnosti bilo bi potrebno izmjeriti broj mlazova u području prolaza

i području pada, ali to nije moguće napraviti u području pada zbog prevelike pozadine QCD događaja. Stoga se efikasnost mjeri na način da precizno odredimo teorijsko predviđanje doprinosa Z događaja u području prolaza, a razliku između predviđanja i mjerenja pripišemo razlici u efikasnosti ParticleNet algoritma. Za dobiti što bolja predviđanja, na simulaciju koja koristi najniži red računa smetnje primjenjujemo korekcije izračunate pomoću sljedećeg višeg reda računa smetnje.

Prilagodba modela na podatke istovremeno se radi u području prolaza i pada, posebno za svaku godinu prikupljanja podataka i za svaku od dvije radne točke. Rezultati su prikazani na Slici 4, izraženi kao omjer efikasnosti u podacima i simulaciji (engl. data-to-simulation scale factor, SF). Rezultati na radnoj točki 0.98 uspoređeni su s rezultatima dobivenim prijašnjom metodom te je vidljivo dobro slaganje između dvije metode.




Slika 4: Izmjeren omjer efikasnosti ParticleNet algoritma između podataka i simulacije, koristeći $Z \to b\bar{b}$ mlazove, prikazan je na gornjem lijevom grafu. Ostala tri prikazuju usporedbu omjera efikasnosti dobivenih metodama koje koriste $Z \to b\bar{b}$ mlazove i QCD mlazove za tri godine prikupljanja podataka, zasebno.

**Zaključak**

U ovom je radu opisana potraga za kaskadnim raspadom teške skalarne čestice X u drugu skalarnu česticu Y i Higgsov bozon H gdje se obje dalje raspadaju u par b kvarkova. Raspadi čestice X u dvije skalarne čestice različitih masa predviđeni su u mnoštvu BSM teorija, ali su uglavnom neistraženi na LHC-u. Potraga je napravljena u rasponu masa 0.9–4 TeV za X i 60–600 GeV za Y, fokusirajući se na kinematički režim gdje Y i H imaju visoki $p_T$ te se produkti njihovih raspada mogu rekonstruirati kao jedinstveni široki mlazovi. Potraga koristi podatke proton-proton sudara na energiji od 13 TeV u sustavu centra mase, prikupljene CMS eksperimentom na LHC-u. Podaci su prikupljeni u 2016–2018 i ukupna količina podataka iznosi 138 $\text{fb}^{-1}$.

Od 260 različitih testiranih ($M_\text{X}$,$M_\text{Y}$) hipoteza, ona s najvećom značajnošću opažanja od $3.1\sigma$ je $M_\text{X} = 1.6\text{TeV}$ i $M_\text{Y} = 90\text{GeV}$. Globalna značajnost takvog opažanja je $0.7\sigma$. Također su postavljene i gornje granice na udarni presjek traženog procesa sa sigurnošću intervala pouzdanosti od 95% u rasponu 0.1–150 fb, ovisno o $M_\text{X}$ i $M_\text{Y}$. Postavljene granice su uspoređene s najvećim dopuštenim udarnim presjecima u NMSSM i TRSM teorijama što dovodi do isključenja dijela faznog prostora teorija za određene $M_\text{X}$ i $M_\text{Y}$. Za NMSSM, područje isključenja obuhvaća raspone $M_\text{X}$ 1000–1150 GeV i 101–145 GeV za $M_\text{Y}$. Za TRSM, područje isključenja je unutar $950 < M_\text{X} < 1330\text{GeV}$ i $110 < M_\text{Y} < 132\text{GeV}$. Gornje granice su također uspoređene s drugim relevantnim mjerenjima i pokazano je da predstavljeni rezultati značajno poboljšavaju prethodne rezultate.

Također je u radu predstavljena i kalibracija ParticleNet algoritma za dvije radne točke. Kalibracija se sastoji od mjerenja Z rezonancije na pozadini QCD događaja. Ključna komponenta mjerenja su korekcije idućeg reda računa smetnje primijenjene na simulaciju Z i W događaja koja koristi najniži red računa smetnje. Rezultati dobiveni ovom, novom, metodom uspoređeni su s postojećim rezultatima koji koriste same QCD događaje za kalibraciju. Rezultati potvrđuju slaganje između dviju metoda.

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction and motivation

The idea that all matter is composed of tiny, indivisible components is more than two thousand years old. In the 5th century B.C., Democritus postulated an atomic theory of the universe by naming these components atoms (from Greek *atomon* "uncuttable, indivisible"). Ruđer Bošković published his "Theory of natural philosophy" (*Theoria philosophiae naturalis redacta ad unicam legem virium in natura existentium"*) in 1758 encapsulating many fields of science, but the most important work is that contributing to the understanding of the structure of matter. He postulated that the matter is composed of points with no internal structure and the points interact with a force which changes behaviour from repulsive to attractive multiple times, starting from repulsive at the smallest distances and ending up attractive at large distances, the latter being congruent with the gravitational force. In a way, this is a prototype of the forces between atoms in a molecule or the nuclear force between the nucleons. In the modern science, it is established that the atoms are the basic particles that compose a chemical element. Interestingly enough, atom is a composite particle and therefore divisible, however, the name is kept out of convention. Truly indivisible particles are called elementary particles and are studied by the field of elementary particle physics.

The theory describing the elementary particles and the interactions between them is the standard model (SM). One of the ways of studying the interactions between the elementary particles is to force them to interact by accelerating them to high energies and colliding them. The world's largest and most powerful particle accelerator is the

Large Hadron Collider (LHC). The main motivation for constructing the LHC was the potential for the discovery of the Higgs boson, which was at the time the only unobserved particle postulated by the SM. The announcement of the discovery of the Higgs boson by the ATLAS and CMS experiments, nearly 50 years after the postulation of the Higgs boson, was a great triumph of the theory. However, this discovery did not end the physics program at the LHC. Even though the SM shows an agreement with a remarkable number of measurements, there are some strong indications that it is not a complete theory. For example, the astronomical observations conclude that only approximately 5% of the energy in the universe is attributable to the content of the SM. Therefore, theories beyond the standard model (BSM) are proposed in order to provide explanation for phenomena otherwise not described by the SM. The data from the collisions at the LHC are analyzed to look for the signs of BSM theories by searching for the presence of new particles or deviations of the parameters from their SM values.

In this thesis, a search for two new scalar particles, X and Y, is presented. The measurement is looking for the signs of a process $X \to YH \to b\bar{b}b\bar{b}$, where H is the Higgs boson of the SM. It is performed in the kinematic regime where the H and Y are imparted with large momenta such that their decay products, $b\bar{b}$ pairs, are collimated. Therefore, the H and Y are each reconstructed as a large-area hadronic jet. This search is motivated by the BSM theories such as the next-to-minimal supersymmetric standard model (NMSSM) which provide explanation for some of the shortcomings of the SM and, among others, postulate additional scalar particles.

The rest of this chapter aims to introduce the theory of elementary particle physics, focusing on its aspects relevant to the Higgs boson sector as it is directly related to the search for a signature of new physics presented in this thesis. A brief overview of the theory describing elementary particle physics are given in Sec. 1.1. The underlying mechanism of the theory, gauge invariance, is shown on a simple example of electromagnetic interaction in Sec. 1.2. The electroweak unification is described in Sec. 1.3, providing the motivation for electroweak symmetry breaking and the incorporation of the Higgs field in the theory, which is given in Sec. 1.4. Some theoretical and experimental shortcomings of the current theory are discussed in Sec. 1.5, inviting the searches for new particles predicted by models which attempt to resolve some of these shortcomings. Finally, several physics

models which increase the particle content of the Higgs sector are described in Sec. 1.6, providing the motivation for the search presented in this thesis.

## 1.1 The standard model of elementary particle physics

The SM of elementary particle physics is the theory describing all known elementary particles and three of the four known fundamental forces, the electromagnetic, weak and strong. It is based on the framework of quantum field theory (QFT) which treats particles as the excited states of their quantum fields. The interaction between particles is governed by the Lagrangian which contains terms describing the interactions between their quantum fields.

Particles of the SM are divided into bosons and fermions. Bosons are particles with integer spin, following Bose-Einstein statistics, and they act as force carriers, mediating interactions between particles. The bosons of the SM are the gluons, photon, W+, W-, Z and the Higgs boson, H. Fermions are particles of half-integer spin and follow Fermi-Dirac statistics. They are associated with "ordinary" matter. Each fermion has a corresponding antiparticle, which is a particle with the same properties, but with opposite physical charges.

The electromagnetic force is mediated by photons which are electrically neutral spin-1 bosons that interact with electrically charged particles. Quantum Electrodynamics (QED) is the theory describing these interactions. It is the first theory which fully reconciled quantum mechanics and special relativity. It is also famous for extremely accurate predictions, for example the prediction of the anomalous magnetic moment of the electron [1]. Photons are massless, allowing the range of the electromagnetic force to be infinite.

The weak force between particles carrying weak isospin charge is mediated by one neutral, Z, and two electrically charged, $W^+$, $W^-$, spin-1 bosons . The bosons of the weak force are massive ($\approx$90GeV) and short-lived with a lifetime of around $10^{-24}$s. This makes the range of the weak force short (R$\approx 10^{-18}$m). In the standard model, the weak and the electromagnetic force are described within a single framework describing "electroweak"

(EWK) interactions. It is said that the two forces are unified at high energies.

Gluons are electrically neutral spin-1 bosons that carry the color charge and mediate the strong force between color-charged particles. This interaction is described by the theory called Quantum Chromodynamics (QCD). The three color charges are labeled red, green and blue. Since they carry the charge of the interaction which they mediate, they can interact with one another. Gluons are also massless. However, due to the nature of QCD, the range of strong interaction is short (R$\approx 10^{-15}$m).

The final boson in the SM, the Higgs boson, is introduced to explain the generation of the mass of the weak-force bosons through the Higgs mechanism. Without the Higgs mechanism, these bosons would be massless in the SM, contrary to observations. The introduction of the Higgs boson in the SM can also explain the origin of fermion masses through the so-called Yukawa couplings. The mechanism was proposed in 1964 [2, 3, 4] and the discovery of a new particle consistent with the Higgs boson in 2012, almost 50 years later, was a great triumph of the SM. The Higgs mechanism and Yukawa coupling are briefly described in Sec. 1.4.

Elementary fermions are split into three generations, each comprising two leptons and two quarks. The division of fermions into quarks and leptons is based on whether they do (quarks) or do not (leptons) interact with the strong force. Particles in higher generations have greater masses, causing them to decay into lower-generation particles through weak interactions. Therefore, all stable matter is made of first-generation fermions. Each generation of leptons comes in pairs which are: electron ($e$) and electron neutrino ($\nu_e$), muon ($\mu$) and muon neutrino ($\nu_\mu$), tau ($\tau$) and tau neutrino ($\nu_\tau$). All leptons carry weak isospin charge and thus interact with the weak force. Electrons, muons and taus carry one unit of electric charge so they also interact with the electromagnetic force.

Pairs of quarks in a generation consist of a particle with +2/3 charge and -1/3 charge. The three generations of quark consist of: up and down, charm and strange, top and bottom quarks. All quarks carry color, weak isospin and electric charge. Therefore, they interact with all three forces in the SM. An interesting phenomenon of the strong force, the so-called confinement, is that quarks cannot be observed as free particles. This is due to the nature of the strong force, where the potential energy between quarks increases with distance. At a certain distance, it becomes energetically more favourable to create a

quark-antiquark pair from vacuum which will couple with the original quarks into colorless bound states. Due to this, quarks are always found in bound states. An exception to this is the top quark which decays before the process off hadronization.

The elementary particles of the SM are shown in Fig. 1.1

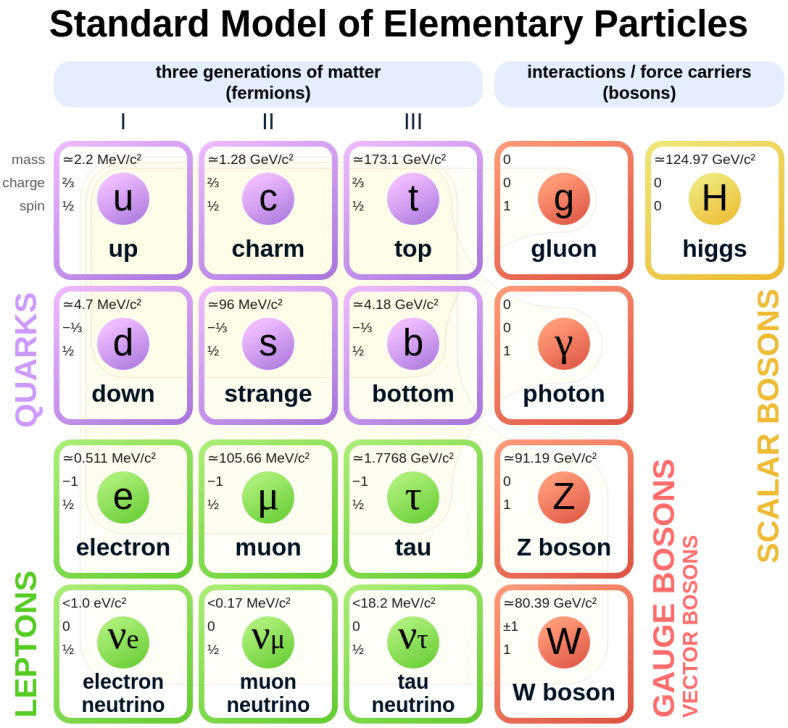



Figure 1.1: A diagram showing the 12 fundamental fermions and 5 fundamental bosons in the SM. Figure taken from Ref. [5]

## 1.2 Gauge invariance in the standard model

As stated in Sec. 1.1, the SM uses the formalism of QFT where particles are described in terms of fields and their excitations. The "physics" of the SM is encoded in a Lagrangian density, $\mathcal{L}_{SM}$, containing terms describing the fields and their interactions. Throughout the rest of the thesis, Lagrangian density is shortened to just Lagrangian.

A key aspect of the SM is that it is a gauge theory, meaning that the Lagrangian is invariant under local 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.

To show an example of the importance of local gauge invariance, we can have a look at the simplest case, the electromagnetic interaction which follows the $U(1)$ symmetry. We can start from the Dirac equation of motion. It describes a free field for particles of spin $1/2$ with mass $m$.

$$(i\gamma^{\mu}\partial_{\mu} - m)\psi(x) = 0, \tag{1.1}$$

where $\gamma^{\mu}$ are the Dirac matrices, $\partial_{\mu}$ is shorthand notation for $\frac{\partial}{\partial x^{\mu}}$ $\psi(x)$ and $\psi(x)$ is the four-component spinor. The Lagrangian giving rise to the Dirac equation is

$$\mathcal{L}_{free} = \overline{\psi}(i\gamma^{\mu}\partial_{\mu} - m)\psi(x) \tag{1.2}$$

The local $U(1)$ gauge invariance would require that the Lagrangian is invariant under $\psi(x) \rightarrow \psi'(x) = e^{iQ\theta(x)}\psi(x)$. However, this is not satisfied because an extra term will arise from:

$$\partial_{\mu}\psi(x) \rightarrow e^{iQ\theta}(\partial_{\mu} + iQ\partial_{\mu}\theta)\psi(x). \tag{1.3}$$

To satisfy the local gauge invariance requirement, an additional term must be added to the Lagrangian, cancelling the extra term from Eq. 1.3. We can introduce a new field, $A_{\mu}(x)$, which by definition transforms as:

$$A'_{\mu}(x) = A_{\mu}(x) + \frac{1}{e}\partial_{\mu}\theta, \tag{1.4}$$

and define a covariant derivative

$$D_{\mu}\psi(x) = [\partial_{\mu} - ieQA_{\mu}(x)]\psi(x). \tag{1.5}$$

The newly constructed Lagrangian can now be written as

$$\mathcal{L} = i\overline{\psi}(x)\gamma^{\mu}D_{\mu}\psi(x) - m\overline{\psi}(x)\psi(x) = \mathcal{L}_{free} + eQA_{\mu}(x)\overline{\psi}(x)\gamma^{\mu}\psi(x), \tag{1.6}$$

and is invariant under local $U(1)$ transformations. The interesting part is that a new (gauge) field needed to be added to the Lagrangian in order to satisfy the local gauge invariance requirement. Furthermore, the additional term 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. To complete the Lagrangian, a gauge-invariant kinematic term needs to be added. This is the Lagrangian giving rise to Maxwell equations:

$$\mathcal{L}_{kin} = -\frac{1}{4} F^{\mu\nu} F_{\mu\nu}, \tag{1.7}$$

where $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$ is the antisymmetric field tensor for $A_\mu(x)$. Another interesting aspect is that a mass term for the gauge field, which would be written as $L_{mass} = -\frac{1}{4} m^2 A^\mu A_\mu$, would violate the gauge invariance and is thus forbidden. This is in line with the observation that the photon is a massless particle.

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. This leads to the introduction of the fields which can be related to the gauge bosons described in Sec. 1.1.

## 1.3 Electroweak unification

It is experimentally observed that weak interactions do not conserve parity (P). The weak interaction therefore includes different couplings to the two parity states. The Dirac spinor is decomposed into two components using the chirality operators, $P_L$ and $P_R$, giving the left- and right-handed chirality components, $\psi_L$ and $\psi_R$. The weak interaction is generated by the $SU(2)_L$ group. The underscript $L$ signifies that left-handed fields transform as: $\psi_L \rightarrow e^{i\theta_j \tau_j} \psi_L$, where $\tau_j$ are the three Pauli matrices divided by 2 (the so-called generators of the $SU(2)$ group), while the right-handed fields remain invariant under $SU(2)_L$ transformations.

The left-handed states are grouped into lepton, $\binom{\nu_L}{\ell_L}$, and quark, $\binom{q_L}{q'_L}$, doublets. Each

group represents one generation of leptons or quarks. A new quantum number is introduced for this interaction, the weak isospin, $I$. Its projection represents the charge corresponding to the $SU(2)_L$ symmetry and is $I_3 = +1/2$, for $\nu_L$ and $q_L$, and $I_3 = -1/2$ for $\ell_L$ and $q'_L$. The corresponding charge for right-handed fermions is $I = 0$. Thus, the right-handed fermions form singlets.

The local gauge invariance under the $SU(2)_L$ transformations requires the addition of three gauge boson fields, $W^{\mu,i}$. Two of them can be associated to the two charged bosons mediating the weak force with the following transformation:

$$W^{\pm,\mu} = \frac{1}{\sqrt{2}}(W^{\mu,1} \mp W^{\mu,2}). \tag{1.8}$$

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 EW theory can be constructed by requiring gauge invariance under $SU(2)_L$ and the $U(1)_Y$ transformations, leading to gauge fields $W^{\mu,3}$ and $B^\mu$. The related covariant derivative which ensures the invariance of the Lagrangian takes the form of:

$$D_\mu = \partial_\mu + ig\tau^i W^i_\mu + ig'\frac{Y}{2}B_\mu, \tag{1.9}$$

where g and g' are the coupling strengths of the electroweak interaction to the $W^i$ and $B$ fields, respectively. We have also introduced a new quantum number, the weak hypercharge $Y$. It is connected to the electrical charge, $Q$, and the weak isospin through the relation $Q = I_3 + Y/2$. A linear transformation between the $W^{\mu,3}$ and $B^\mu$ can be made:

$$\begin{pmatrix} W^{3,\mu} \\ B^\mu \end{pmatrix} = \begin{pmatrix} \cos\theta_W & \sin\theta_W \\ -\sin\theta_W & \cos\theta_W \end{pmatrix} \begin{pmatrix} Z^\mu \\ A^\mu \end{pmatrix}. \tag{1.10}$$

The mixing angle, $\theta_W$, also called the Weinberg angle, is selected so that

$$\theta_W = \tan^{-1}(g'/g), \tag{1.11}$$

where $g$ and $g'$ are the coupling strengths of $SU(2)_L$ and $U(1)_Y$ in the EW Lagrangian, respectively. The motivation for such a choice is that it causes the $Z^\mu$ to only couple to isospin and $A^\mu$ only to the electrical charge. The two fields then correspond to the $Z$ boson and the photon, respectively, and can be expressed as:

$$\begin{aligned} Z^\mu &= \frac{-g'B_\mu + gW_\mu^3}{\sqrt{g^2 + g'^2}} \\ A^\mu &= \frac{gB_\mu + g'W_\mu^3}{\sqrt{g^2 + g'^2}} \end{aligned} \tag{1.12}$$

## 1.4 Electroweak symmetry breaking

The EW Lagrangian constructed with the procedure described in Sec.1.3 does not contain mass terms for the W and Z bosons, which is not surprising since we stated in Sec. 1.2 that such terms would violate the gauge symmetry. Moreover, unlike the EM Lagrangian of Sec. 1.2, the EW Lagrangian also does not (yet) have mass term for fermions, which would be:

$$\sim m\overline{\psi}\psi = m(\overline{\psi}^R\psi^L + \overline{\psi}^L\psi^R). \tag{1.13}$$

This term cannot be gauge invariant under $SU(2)_L$ since only the left-handed fields would get transformed, while the right-handed fields would stay unchanged. However, the observed masses of W, Z bosons and fermions need to be included in the Lagrangian to have a theory which agrees with the experiment.

The generation of the mass terms can be achieved with the so-called spontaneous symmetry breaking (SSB), which we can introduce with a simple example. Let us consider a Lagrangian for a complex scalar field, $\phi(x)$, with a potential of the form:

$$\mathcal{L} = \partial_\mu\phi^\dagger\partial^\mu\phi - V(\phi) = \partial_\mu\phi^\dagger\partial^\mu\phi - \mu^2\phi^\dagger\phi - \lambda(\phi^\dagger\phi)^2 \tag{1.14}$$

The parameter $\lambda$ is assumed positive because otherwise the potential would have no state

of minimum energy. We can distinguish two possible shapes of the potential depending on the values of the parameter $\mu$.

If $\mu^2 > 0$, the $\phi = 0$ state will be the minimum of the potential and the Lagrangian describes a scalar particle with mass $\mu$, and quartic coupling, $\lambda$. However, this case does not lead to the phenomenon we would like to see.

In the other case, $\mu^2 < 0$, the minimum is no longer at zero and is degenerate, owing to the Lagrangian invariance under global $\phi(x) \to e^{i\theta}\phi(x)$ transformations. The states satisfying the following equation define the local minimum:

$$\phi^\dagger\phi = \frac{-\mu^2}{2\lambda} \equiv \frac{v^2}{2}, \tag{1.15}$$

where we have defined $v^2 = -\mu^2/\lambda$. This is a common parameter in many SSB models and is called the vacuum expectation value, or vev. The minimum state can therefore be written as:

$$\phi_{min}(x) = \frac{v}{\sqrt{2}}e^{i\theta} \tag{1.16}$$

Once a particular minimum state is selected, for example $\theta = 0$, the field $\phi$ no longer exhibits symmetry under the U(1) rotation, and the symmetry is said to be spontaneously broken.

To understand the particle content of the Lagrangian, the excitations around the ground state can be parametrized as two fields:

$$\phi(x) = \frac{1}{\sqrt{2}}(v + \eta(x) + i\xi(x)). \tag{1.17}$$

The potential now takes the form of:

$$V(\phi) = V(\phi_0) - \mu^2\eta^2 + \lambda v\eta(\eta^2 + \xi^2) + \frac{\lambda^2}{4}(\eta^2 + \xi^2)^2. \tag{1.18}$$

The interesting part is that a mass term arises for the $\eta$ field, giving it mass $m_\eta^2 = -2\mu^2$.

The mass term for $\xi$ is not present, indicating that it is the field of a massless particle. The fields can be interpreted as excitations around the vev in the directions where the potential rises ($\eta$) and along the minimum, where the potential is constant ($\xi$). Since the latter excitations do not cost any energy, they correspond to a massless state.

### 1.4.1 Higgs mechanism in the Electroweak theory

In the previous section, we have shown how breaking of a symmetry can lead to the generation of additional fields which may have mass. This is known as the Higgs mechanism. This can now be applied to the electroweak Lagrangian, which exhibits the $SU(2)_L \otimes U(1)_Y$ invariance, following a similar procedure with some adaptations to account for the more complex symmetry group.

Since we are working with an $SU(2)_L$ theory, the minimum Higgs mechanism consists of a doublet of complex scalar fields with four degrees of freedom:

$$\phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} = \begin{pmatrix} \phi_1 + i\phi_2 \\ \phi_3 + i\phi_4 \end{pmatrix}. \tag{1.19}$$

The Lagrangian for the doublet then takes the form

$$\begin{aligned} \mathcal{L} &= (D_\mu \phi)^\dagger (D^\mu \phi) - V(\phi), \\ V(\phi) &= \mu^2 \phi^\dagger \phi + \lambda (\phi^\dagger \phi)^2. \end{aligned} \tag{1.20}$$

where $D_\mu$ is the covariant derivative for the electroweak theory, given in Eq. 1.9. Similarly to the case with which we introduced the SSB, when $\mu^2 < 0$, the local minimum is degenerate for $\phi^\dagger \phi = -\mu^2/\lambda$.

Through gauge rotations, we set the expected values of field $\phi_{1,2,4}$ to zero and rewrite the doublet as:

$$\phi = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + H \end{pmatrix}, \tag{1.21}$$

where we expanded $\phi_3$ around its vev, $\phi_3 = v + H$. Inserting this field into the kinematic term of the Lagrangian in Eq. 1.20 gives rise to the mass terms for the gauge bosons in

the form of:

$$\frac{v^2}{8}\left(g^2(W_\mu^1 W^{1,\mu} + W_\mu^2 W^{2,\mu}) + (g' B_\mu - g W_\mu^3)^2\right) \tag{1.22}$$

By inserting the transformations between fields of the Eq. 1.22 and the $W^{\pm}$, Z and the photon, given in Eq. 1.8, we can recover the mass terms from the Lagrangian:

$$\begin{aligned} M_W^2 =& \frac{1}{4} g^2 v^2 \\ M_Z^2 =& \frac{1}{4}(g^2 + g'^2) v^2 \\ M_A^2 =& 0 \end{aligned} \tag{1.23}$$

In summary, the inclusion of a complex scalar doublet field $\phi$ with its potential having a minimum energy state differing from $\phi = 0$ enabled us to write a Lagrangian whose kinematic term contains the mass terms for the W and Z gauge bosons. It needs to be noted that the mass terms for the W and Z bosons originated from the application of covariant derivative onto the new field, which differs from the previous case. In the previous case, the generated mass term for the field describing excitations around the ground state, $\eta$, arose from the potential. This would correspond to the mass term generated for the $H$ field in the example of SM EWK symmetry breaking, i.e., the mass of the Higgs boson.

The mass terms for fermions can be obtained in a similar fashion with the Yukawa coupling. The following terms are invariant to the $SU(2)_L$ transformation and can be added (together with their hermitian conjugates):

$$\mathcal{L} = c_1(\bar{u}, \bar{d})_L \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} d_R + c_2(\bar{u}, \bar{d})_L \begin{pmatrix} \phi^{0*} \\ -\phi^- \end{pmatrix} u_R + c_3(\bar{\nu}_e, \bar{e})_L \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} e_R + h.c. \tag{1.24}$$

Replacing the doublet term with the form in 1.21, the Lagrangian takes a simpler form:

$$\mathcal{L} = \frac{1}{\sqrt{2}}(v + H)\left[c_1 \bar{d} d + c_2 \bar{u} u + c_3 \bar{e} e\right] \tag{1.25}$$

Generated masses of the fermions can now be read out as:

$$\begin{aligned} m_d &= -c_1 \frac{v}{\sqrt{2}}, \\ m_u &= -c_2 \frac{v}{\sqrt{2}}, \\ m_e &= -c_3 \frac{v}{\sqrt{2}} \end{aligned} \tag{1.26}$$

Coefficients $c_i$ are parameters of the theory so the fermion masses are arbitrary. More details of the Higgs mechanisms and Electroweak symmetry breaking can be found in Refs. [3, 2, 4, 6, 7, 8, 9]

## 1.5 Physics beyond the standard model

The standard model is an extremely successful theory. Numerous precision measurements have been carried out which show excellent agreement between the SM prediction and the measurement. However, there are some observations which cannot be explained with the SM. Various physical extensions of the SM have been proposed which may give explanation to some of them, while still being consistent with the existing measurements. Such models are categorized as BSM models. Some of the experimental issues which BSM models try to resolve are:

- **Gravity** - The SM does not include gravity as one of the fundamental forces. Furthermore, the theory describing gravity, general relativity, is considered incompatible with the QFT, the framework of the SM. Some BSM theories propose the existence of an additional boson, graviton, as a carrier of the gravitational force, but it is yet undiscovered.
- **Dark matter** - Cosmological observations conclude that only 5% of the energy in the universe can be attributed to SM. About 26% of the energy should come from dark matter [10] which only interacts with gravitational force and possibly with another force that needs to be weak in strength. The SM does not provide a particle which may be a candidate for dark matter.

- **Dark energy** - The remaining 69% of the energy in the universe (from the previous point) should come from the energy density of the vacuum [10]. The calculation of the vacuum energy density using the SM leads to a mismatch of 120 orders of magnitude [11].

- **Matter-antimatter asymmetry**: The standard model predicts that matter and antimatter should have been created in almost equal amounts during the Big Bang, while we observe that the universe is almost exclusively made out of matter. Therefore, it is likely that there was a physical law acting differently on matter and antimatter. Although there are mechanisms in the SM breaking this symmetry, they cannot sufficiently explain the apparent abundance of matter over antimatter.

There are other, more technical, problems with the standard model not coming from a direct observation.

- **Fine-tuning problem** - The calculation of the mass of the Higgs boson involves large quantum corrections due to the presence of virtual particles. The corrections are much larger than the mass of the Higgs boson and have to cancel out almost perfectly. Some BSM models introduce a more "natural" way for the corrections to lead to the observed mass of the Higgs boson.

- **Force unification** - Since the electroweak unification, it is widely believed that the three fundamental forces described by the SM are just different manifestations of a single underlying force which would eventually unify to a common value at some higher energy scale. This is indicated by the strength of the strong force weakening with increasing energy scales. However, in the SM, such a unification does not occur.

- **Free parameters of the SM** - The SM is determined by 19 numerical parameters, the values of which are determined in the experiments. There are no underlying principles giving explanations for the values of these parameters.

## 1.6 Higgs sector beyond the standard model

In spite of its enormous success in predicting a wide range of phenomena and observables in the last several decades, the SM is considered to have several theoretical as well as experimental shortcomings, some of which are listed in the previous section, sparking the searches for potential new physics. The discovery of the Higgs boson at the CERN LHC [12, 13] is considered to be the capstone of the SM. Correspondingly, searches for manifestations of physics beyond the SM involving this newly discovered particle have intensified since its discovery. Among the plethora of theoretical constructs to address many of the SM issues, many predict the existence of new particles in the Higgs sector. Some of them are introduced in this section.

In Sec. 1.4, a doublet scalar field was introduced to include the Higgs mechanism in the SM. However, this is just the minimal case for the inclusion of the Higgs mechanism and more complex realizations of the Higgs mechanism are allowed. A particular class of common extensions of the SM is the Two Higgs Doublet Model (2HDM) which, as suggested by the name, contains two Higgs doublets instead of just one [14]. The introduction of the second doublet leads to a richer phenomenology. The Higgs sector contains five scalars, instead of just one as is in the SM. These are: the two charge-parity (CP) even neutral Higgs bosons, which differ in mass and with one of them corresponding to the SM Higgs boson, one CP odd pseudoscalar (A) and two charged Higgs bosons $H^{\pm}$. Correspondingly, the number of parameters describing the Higgs sector is increased with respect to the SM which uses only two parameters, the mass of the Higgs boson and the vev. The parameters in the 2HDM are the four Higgs masses, the ratio of the vevs of the two doublet fields ($\tan\beta$) and the mixing angle ($\alpha$) that diagonalizes the mass matrix of the neutral CP even Higgses.

Besides the Higgs sector, the Yukawa couplings to fermions are also modified with the introduction of the second doublet. In the 2HDM, fermions have a possibility to couple to either of, or both, doublets. However, if a 2HDM model with no flavour changing neutral currents (FCNC) is to be constructed, all fermions with the same quantum numbers need to couple to the same Higgs doublet [15]. Based on the choice of coupling, four different type of 2HDM models exist [15]. In the type–I models, the charged leptons and quarks

only couple to one of the two Higgs doublets while in the type–II models the charged leptons and down-type quarks couple to one Higgs doublet and the up-type quarks to the other. The flipped model is similar to type-II, the charged leptons and up-type quarks couple to the same doublet and the down-type quarks to the other. Finally, in the so-called lepton-specific model, quarks of both types couple to one, and leptons to the other doublet. This is summarized in Table 1.1

Table 1.1: Four types of 2HDM models without tree–level FCNC and the couplings of quarks and charged leptons to the Higgs doublets. By convention, the doublet to which the up-type quarks couple is labeled as $\phi_1$.

| Model type | Up quarks | Down quarks | Ch. leptons |
|---|---|---|---|
| Type-I | $\phi_1$ | $\phi_1$ | $\phi_1$ |
| Type-II | $\phi_1$ | $\phi_2$ | $\phi_2$ |
| Flipped model | $\phi_1$ | $\phi_2$ | $\phi_1$ |
| Lepton-specific model | $\phi_1$ | $\phi_1$ | $\phi_2$ |

The Type-II 2HDM Higgs sector arises in the minimal supersymmetric extension of the standard model (MSSM) [16, 17]. The proposal of the MSSM is motivated by some of the shortcomings of the SM which it solves. In the supersymmetric models, the problem of the fine-tuning of the Higgs mass, described in Sec. 1.5, is resolved. It also unifies the three interactions into one at high energy scales. Finally, one of the additional (super)particles proposed by supersymmetric models is a candidate for dark matter. Such strong motivation lead to many searches for particles proposed by the MSSM. Among them, searches for a heavy neutral or a charged Higgs boson in different final states were performed [18, 19, 20, 21]. Depending on the $\tan\beta$ parameter of the MSSM, searches for a heavier scalar decaying to $\tau$ leptons or top quark pairs have set a limit on its mass as high as $M_{\mathrm{X}} > 2.0\mathrm{TeV}$ [19, 18, 22].

The NMSSM [23, 24] has been proposed to solve the "$\mu$ problem" of the MSSM [25] where the Higgs mass parameter is many orders of magnitude smaller than the Planck scale. The NMSSM contains an extra complex scalar field, compared to the MSSM, and therefore a total of seven Higgs bosons: three neutral scalars, two neutral pseudoscalars, and two charged scalars. One of the neutral scalars is associated with the observed 125GeV Higgs boson H. The mixing of the doublet and singlet scalar fields may lead to very small couplings of the neutral scalars to the SM fermions, suppressing their production

cross section [23]. Certain scenarios lead to enhanced "Higgs-to-Higgs" decays, where the heaviest scalar X decays to a H and the lighter neutral scalar Y [26], with the decay $X \rightarrow YH$ being the most dominant process [27]. Within the NMSSM, the maximum branching fractions of both H and Y are to a pair of b quark-antiquarks giving the final state $X \rightarrow YH \rightarrow b\bar{b}b\bar{b}$, for Y with a mass less than twice that of the top quark [27]. The second dominant process is $X \rightarrow YH \rightarrow b\bar{b}\tau\tau$, which has been searched for by the CMS Collaboration [28], excluding $0.4 < M_X < 0.6$TeV and $50 < M_Y < 200$GeV, based on the model parameters.

Another interesting new physics model is the two-real-scalar-singlet extension of the SM (TRSM) [29], which introduces two additional scalar fields to the SM. This simplified model, onto which more complicated theories can be mapped, has nine degrees of freedom: three masses and vacuum expectation values (vevs) of the scalar fields, and three mixing angles. In the scenario where all three vevs are non-zero, the three fields give rise to three massive scalars, one of which can be associated with the discovered Higgs boson. This provides many interesting possible decay signatures: single production of a scalar boson decaying to SM particles, production of a heavy scalar boson with subsequent decay into two identical or different scalars and even final states with three or four scalars.

Excepting the $X \rightarrow YH \rightarrow b\bar{b}\tau\tau$ search, the cascade decays of a scalar resonance into two scalars of unequal masses are largely unexplored at the LHC. However, we have seen that theoretical motivation for such searches are provided by the NMSSM and TRSM models. This thesis aims to search for the $X \rightarrow YH \rightarrow b\bar{b}b\bar{b}$ process, shown in Fig. 1.2 using proton-proton (pp) collision data at the LHC collected by the CMS experiment between 2016 and 2018 and corresponding to an integrated luminosity of $138\text{fb}^{-1}$. The final state will consist of four b quarks in two pairs from the decays of the H and Y, collimated into a single jet each because of the large Lorentz-boost imparted to the H and Y bosons by the decay of the massive parent particle X.

The motivation for the search can be further bolstered with the searches for a massive resonance decaying to HH in the four b quark final states conducted by the ATLAS [30] and CMS [31] experiments. These $H_{125}H_{125}$ searches primarily searched for signatures of extradimensional models where the massive resonances would be either a spin-0 radion [32, 33, 34] or a spin-2 bulk graviton [35, 36]. The masses explored were in the range of 750GeV

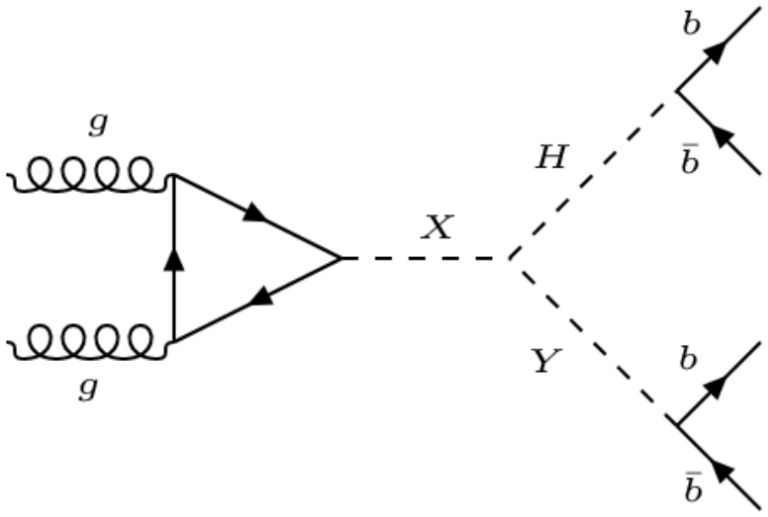



Figure 1.2: A Feynmann diagram showing the gluon-gluon production mode of a heavy scalar X followed by its decay process X → YH → $b\bar{b}b\bar{b}$.

to 3TeV, where the boosted Higgs boson reconstruction technique is the most sensitive. The usage of these techniques in fact leads to these searches yielding the most stringent limits on the extradimensional models [37] [38].

The reasons for mentioning these searches can be found once we recognize that they are a special case of the X → YH → $b\bar{b}b\bar{b}$ search with the $M_Y = 125$GeV condition. First, they can be used as a performance benchmark for the search. Second, the X → YH → $b\bar{b}b\bar{b}$ search results can be reinterpreted in the context of models considered by the resonant di-Higgs searches, increasing the impact of the search.

# Chapter 2

# Experimental Setup

This chapter contains a section giving an overview of the Large Hadron Collider (LHC), followed by a section on the Compact Muon Solenoid Detector (CMS)

## 2.1 The Large Hadron Collider

Particle interactions at high energies are studied by accelerating particles and colliding them. The collision can happen with a fixed target or with another accelerated beam, usually accelerated to the same energy. The benefit of the latter is that the center of mass energy ($\sqrt{s}$) is much higher in that case than with the fixed target accelerator. The Large Hadron Collider is such an accelerator, designed to accelerate two proton beams up to 7 TeV energy each, i.e. colliding them at $\sqrt{s} = 14$ TeV, making it the most powerful particle collider ever built. The construction of the LHC was approved in 1995, while the first proton tests started in September, 2008. During the first data-taking run (Run I), the LHC collided beams with $\sqrt{s} = 7$ TeV and $\sqrt{s} = 8$ TeV. The analysis of data from this run resulted in the observation of the Higgs boson [13, 12, 39], already achieving the main goal of the LHC. However, LHC was built to also look for signs of new physics proposed by numerous BSM models and it still continues to provide a wealth of physics results. During the second data-taking run (Run II), finished in 2018, the LHC collided beams at $\sqrt{s} = 13$ TeV. Subsequent runs are planned with the nominal or close to nominal center of mass energy, $\sqrt{s} = 14$ TeV, and increased instantaneous beam luminosities, next of

which (Run III) is started in July of 2022.

Other design choices, besides going for two beams instead of a fixed target accelerator, were necessary in order to achieve such high energies. One of them is that the LHC is a circular accelerator which makes it possible to incrementally accelerate particles in each lap they make around the accelerator. The alternative choice, a linear accelerator, would mean that the acceleration needs to happen in one pass in the space between the source and the collision point which would necessitate an unfeasibly long accelerator with the current technologies. Conveniently, in the vicinity of Geneva, there is a 27 km long underground circular tunnel previously used by the Large Electron Positron Collider (LEP) in which the LHC is now located. The second design choice is to accelerate and collide hadrons (protons), unlike its predecessor LEP which was colliding electrons and positrons. The switch from electrons to protons significantly increased the achievable beam energies due to the synchrotron radiation effect where a charged particle irradiates energy when in circular motion. The power of the energy loss is proportional to $(\frac{E}{m})^4$, where $E$ and $m$ are the energy and mass of the particle. Since the proton is roughly 2000 times heavier than the electron, the synchrotron radiation effect is weaker by 13 orders of magnitude. This is reflected in the large difference in the design beam energy of the LHC and LEP, whose energy per-beam topped at 104.5 GeV. Another motivation for accelerating protons, besides the increased beam energy, is that they are composite particles, containing quarks and gluons. In the head-on proton collisions, the collision happens between the proton constituents, each carrying a fraction of the proton energy. Therefore, protons (or composite particles in general) produce collisions in a much wider energy spectrum than the collisions of elementary particles like electron-positron collisions. This is suitable for the searches for new particles with a-priori unknown masses. Besides protons, LHC can also collide heavy ions (lead nuclei) to explore phenomena like the quark-gluon plasma, but we will be focusing on protons as these are collisions of interest for this thesis.

To achieve the required energy, hadrons are accelerated in several stages before the injection into the LHC ring. The full energy ramping chain consists of other, smaller accelerators which increase the beam energy in steps before injecting it into the LHC ring. The acceleration chain can be seen in Fig.2.1 and is described below.

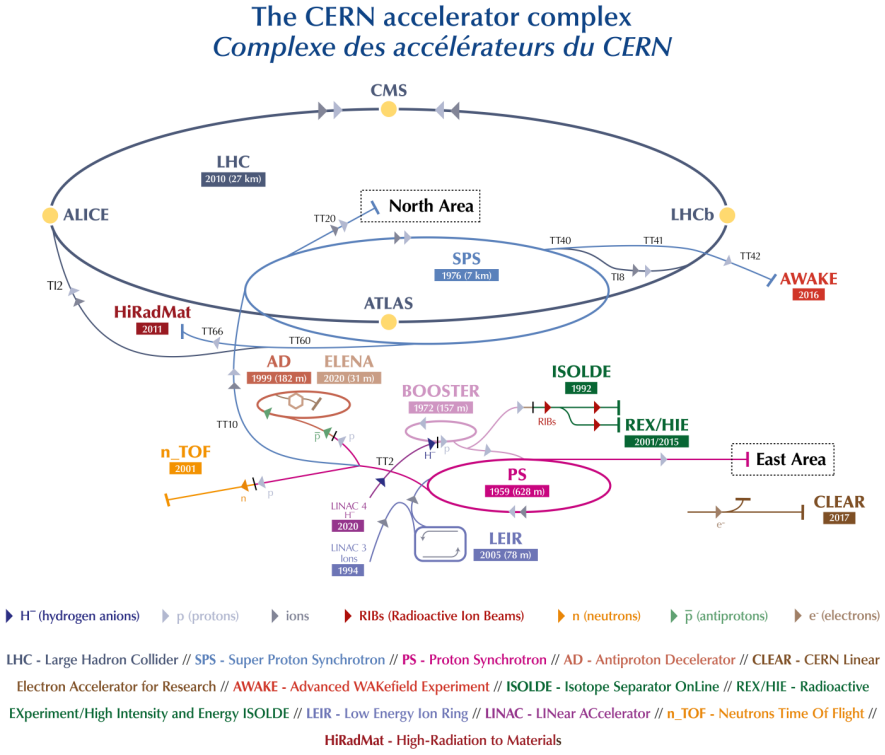



Figure 2.1: The CERN accelerator complex. The proton injection chain consists of Linac 4 (replacing Linac 2), PSB (denoted on the Figure as "Booster"), PS and SPS. The ion injection chain starts with Linac 3, LEIR and then continues with PS and SPS like the proton chain. Figure taken from [40]

The injection chain for the protons consists of:

- Linac 2 - Linear accelerator 2 is the starting point for the protons used in experiments at CERN. The hydrogen is passed through an electric field to strip off its electrons, leaving only protons to enter the accelerator. By the time they reach the other end, the protons reach the energy of 50 MeV. In 2020, the Linac 2 was replaced by a newer accelerator, Linac 4, which reaches the energy of 160 MeV. The rest of the chain is described as it was with the use of Linac 2.

- PSB - The Proton Synchrotron Booster is made up of four superimposed synchrotron rings that receive beams of protons from Linac 2 at 50 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. For a brief period of time, in the 1960s, the PS was the world's highest energy particle accelerator. In 1970s its principal role became to supply particles to the new machines. The accelerator boosts protons up to 26 GeV.

- SPS - The Super Proton Synchrotron is a nearly 7 km long circular accelerator and the second-largest machine at CERN. It provides proton or ion beams to the LHC by taking particles from the PS and accelerating them up to 450 GeV. It also provides beam 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 particles while running as a proton-antiproton collider.

The injection chain for the ions consists of:

- Linac 3 - Linear accelerator 3 is the starting point for the ions used in experiments at CERN. It provides lead ions for the LHC and for fixed-target experiments. At the particles' origin and during acceleration through Linac 3, electrons are stripped away. Eventually, all of the electrons are removed and the lead is transformed into bare nuclei.

- LEIR - The Low Energy Ion Ring (LEIR) receives lead ions from Linear accelerator 3 (Linac 3) and like PSB in the proton injection chain, it transforms the ion beams into bunches suitable for injection to the next ramping step. After LEIR, the ions are fed to the same accelerators as protons, namely the PS and SPS.

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 shape the proton bunches. Protons which are ahead of the rest in its 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.

To keep the protons on a circular path, LHC deploys superconducting dipole magnets which are required to have magnetic fields as strong as 8.3 T. To reach such strong magnetic fields, the superconducting magnets achieve currents of up to 11080 Amperes. In order to keep the material superconductive, the magnets are cooled by liquid helium. Because the LHC accelerates equal charge beams and the beams travel in opposite directions in separate beam pipes, the magnets need to have opposite field directions for each beam. The cross section of the LHC beam pipe inside the dipole magnet is shown in Fig. 2.2. Each dipole is 15 metres long, with the mass of 35 tonnes. Besides the 1232 main dipoles, LHC uses quadrupole magnets to keep the beam narrow and injector magnet systems called "inner triplets", which consists of three successive quadrupoles, to further bunch the protons together just before the interaction point. Inner triplets tighten the beam from 0.2 millimetres down to 16 micrometres across.

The collision happen at the four locations, corresponding to the four large detectors: ATLAS, CMS, Alice and LHCb at the rate of 40 MHz.

Besides the center of mass energy, another important property of any collider is the instantaneous luminosity, $\mathcal{L}$, which measures the number of particle collisions per unit

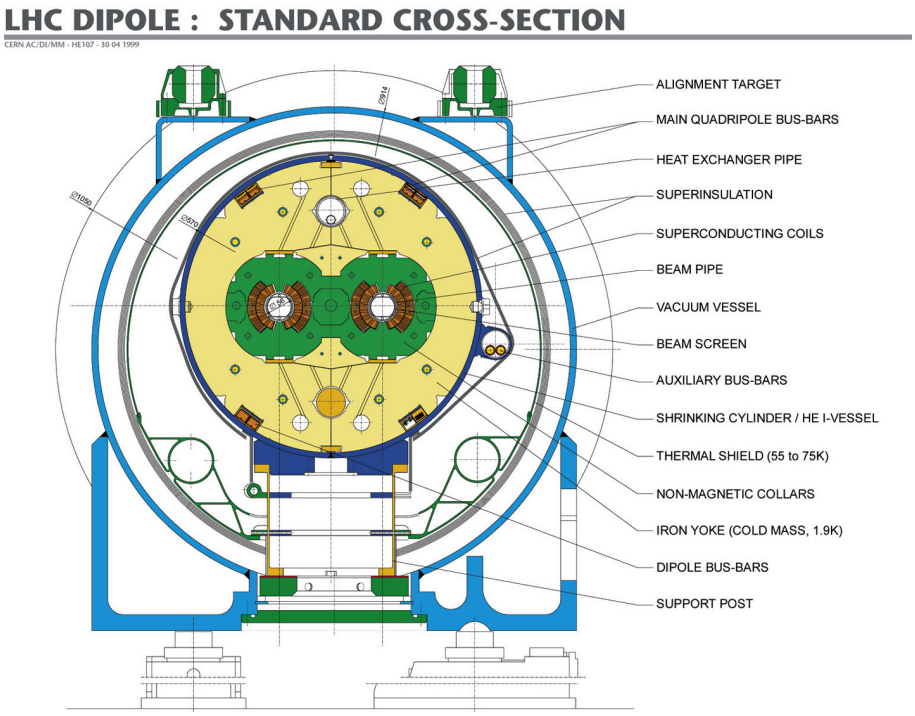



Figure 2.2: Schematic cross section of the LHC dipole. Figure taken from [41]

of time. The designed instantaneous luminosity at the LHC, with $\sqrt{s} = 14$ TeV, is $10^{34}$ cm$^2$s$^{-1}$. Important related quantity is the integrated luminosity which is just the integral of the luminosity over time,

$$\mathcal{L}_{int} = \int \mathcal{L} dt \tag{2.1}$$

The expected number of times a process with the production cross section, $\sigma$, occurred is given by the relation

$$N = \sigma \mathcal{L}_{int} \tag{2.2}$$

The instantaneous luminosity of a particle collider can be expressed as:

$$\mathcal{L} = \frac{N_b^2 n_b f \gamma_r}{4\pi \epsilon_n \beta^*} F \tag{2.3}$$

In the numerator are the quantities concerned with the rate at which protons enter the interaction region. $N_b$ is the number of particles per bunch, $n_b$ is the number of colliding bunches per beam, $f$ is the frequency of revolution and $\gamma_r$ is the relativistic gamma factor

of the beam. The denominator represents the geometrical cross section of the interaction region. $\epsilon_n$ is the normalized emmitance of the beam measuring the spread of the beam in position and momentum space, $\beta^*$ is the beta function at the collision point, related to the transverse size of the particle beam. The relativistic correction factor, $F$, determines the reduction in the luminosity in the case that the bunches collide at some angle. It is given by the expression

$$F = \frac{1}{\sqrt{1 + \left(\frac{\alpha\sigma_z}{2\sigma_t}^2\right)}},$$

where $\alpha$ is the beam crossing angle, $\sigma_z$ the bunch length and $\sigma_t$ is the transverse width of the bunch [42].

The determination of the beam parameters needed to calculate the luminosity are done with Van Der Meer scans [43] which involve 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. The integrated luminosity delivered by LHC to the CMS detector during Run II is shown in Fig.2.3.

## 2.2 The Compact Muon Solenoid

The Compact Muon Solenoid [45] is one of the four large particle detectors at the LHC. The other three are listed below:

- **A Toroidal LHC Apparatus (ATLAS)** is a general-purpose detector designed to observe a wide range of possible new physics at the LHC.
- **A Large Ion Collider Experiment (ALICE)** is a detector dedicated to heavy-ion physics. It studies the physics of strongly interacting matter at extreme energy densities, where a phase of matter called quark-gluon plasma forms, by collecting data from lead nuclei collisions.

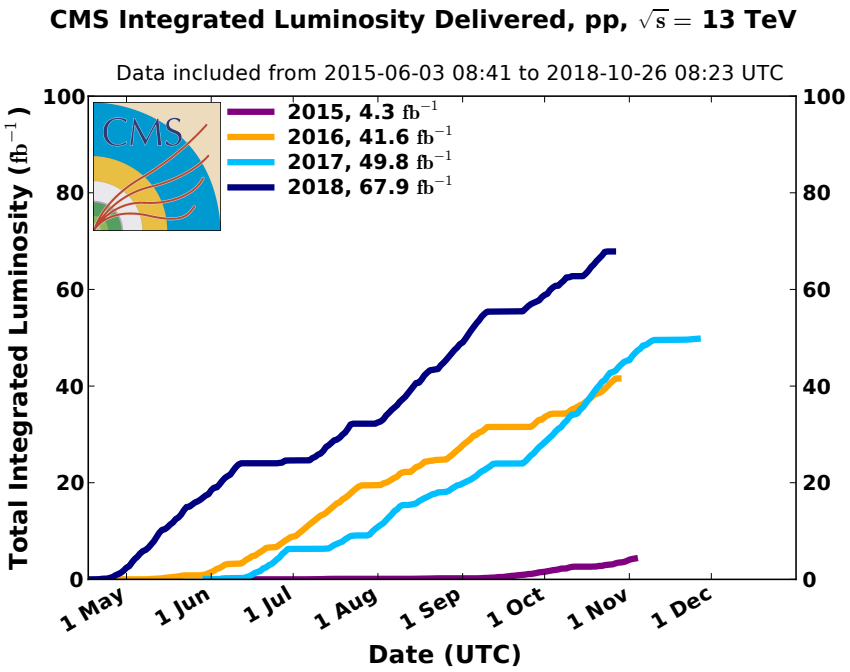



Figure 2.3: Integrated luminosity delivered by LHC to CMS versus time for 2015-2018 (proton data only). Figure taken from [44]

- **LHC beauty (LHCb)** detector specialized in investigating the matter and anti-matter asymmetry by studying the rare decays of hadrons containing b quarks.

Like ATLAS, CMS is a general-purpose particle detector. It has a broad physics program ranging from the discovery of the Higgs boson to exploring new physics and performing precise SM measurements. The detector is cylindrically shaped, centered around the interaction point. It has a length of 21 metres and height and width of 15 metres [46]. The CMS experiment uses a right-handed coordinate system with the origin at the nominal collision point. The x-axis points to the centre of the LHC ring, the y-axis upwards, perpendicular to the LHC plane, and the z axis points along the anticlockwise beam direction. Radial distance, $r$, is defined as the distance from the z-axis. The two common angular coordinates are the azimuthal angle ($\phi$), measured from the positive x-axis in the x-y plane, and the pseudorapidity ($\eta$). Pseudorapidity is the approximation of the physical quantity called rapidity ($y$) whose differences, $\Delta y$ are invariant under Lorentz boosts along the $z$ axis. Rapidity of a particle is defined as

$$y = \frac{1}{2} \ln \left( \frac{E + p_z}{E - p_z} \right), \tag{2.4}$$

where $E$ is the energy of the particle and $p_z$ is the $z$ component of its momentum. Pseudorapidity is defined as

$$\eta = -\ln\left[\tan\left(\frac{\theta}{2}\right)\right], \tag{2.5}$$

where $\theta$ is the polar angle, measured from the positive z-axis. The conversion between $\eta$ and $\theta$ is shown in Fig. 2.4. In the limit when the particle momentum is much greater than the mass of the particle, the pseudorapidity converges to rapidity and is more commonly used since it only relies on the polar angle, $\theta$.

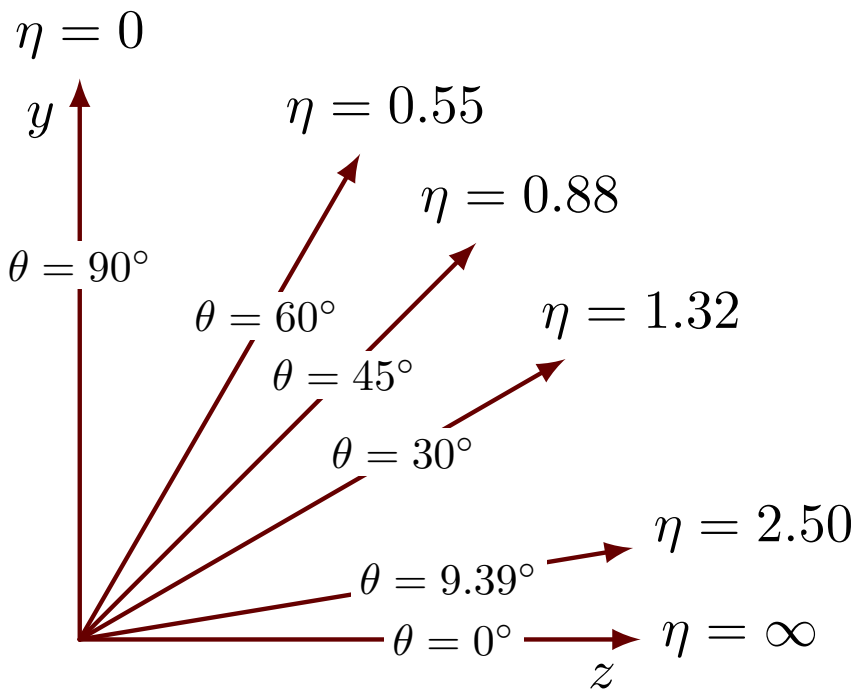



Figure 2.4: Sketch showing the relationship between pseudorapidity $\eta$ and the polar angle $\theta$.

The CMS detector is made of subdetector systems, each having a specialized role. A cross section of the detector is shown in Fig. 2.5 in which the main subdetector systems can be seen. The systems are listed below.

- **Inner Tracker** is the subdetector system closest to the interaction point and is often just called "Tracker". It records the passage of charged particles which enables the reconstruction of their tracks. The tracker consists of the silicon pixel detector

and the silicon strip detector which are located closer and further from the beam, respectively. The coverage of the tracker extends up to $|\eta| = 2.5$.

- **Electromagnetic Calorimeter (ECAL)** measures the energy of photons and electrons which are fully stopped in this detector subsystem. Scintillating lead tungsten crystal are used to measure their energies.

- **Hadron Calorimeter (HCAL)** is the second calorimeter subsystem which stops the hadrons and records their released energy. It consists of alternating layers of brass and the plastic scintillators.

- **Solenoidal Magnet** is the largest superconducting magnet ever build. It has a diameter of 6 metres and a length of 12.5 metres, weighing 12000 tonnes. It is built around the previously listed systems and provides internal uniform magnetic field of 3.8 T along the beam direction. It is embedded in an iron return yoke and provides a residual magnetic field of $\approx 1.6$ T outside the magnet, in the opposite direction than in the interior.

- **Muon Tracker** is placed outside the superconducting magnet because, unlike the other particles recorded by CMS, muons do not stop in the calorimeters (nor in the magnet). Muon tracks are reconstructed by looking at the hits among the four muons stations and also using the information from the silicon tracker system.

A more detailed description of the subdetector systems is given in the following sections.

### 2.2.1 Tracker

The Tracker is the subdetector system closest to the collision point. It is used to reconstruct the trajectories of charged particles originating from beam collisions. The precise reconstruction of trajectories is necessary to determine the position of the interaction point from which the particles originate, the vertex. In the collision events, we are only interested in the particles coming from the head-on pp collisions ("hard-scatter"). The vertex with the largest value of summed squared transverse momenta ($p_T^2$) of the related tracks is taken to be the primary pp interaction vertex. Additional pp interactions within the same or nearby bunch crossings (pileup) contaminate the event so the reconstruction

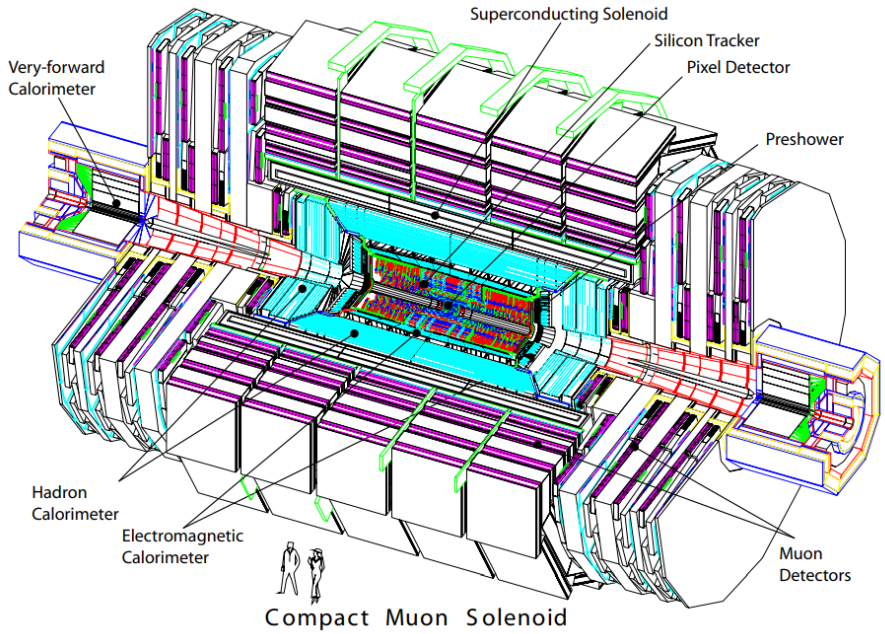



Figure 2.5: The cross section view of the CMS detector showing its main subdetector systems. Figure taken from [45]

of pileup vertices is important in order to subtract their contribution from the primary event. The expected average number of pileup interactions per bunch crossing is in the range of 23–32 at the LHC beam conditions between 2016–2018. It is also important to reconstruct secondary vertices which may, for example, arise from the decay of b–hadrons or photon conversion to electrons. Their reconstruction helps determine the precise evolution of the event in the detector.

The tracker is made of a small silicon pixel detector and a larger, surrounding silicon strip tracker. They are contained within a cylinder of 5.8 m in length with a 2.5 m diameter.

The configuration of the pixel detector in 2016, called "Phase-0" pixel detector, included three barrel layers and two forward disks, located at the ends of the cylinder, extending the tracker acceptance to $|\eta| < 2.4$. The pixel detector was replaced in 2017 with the "Phase-1" configuration, adding a layer in the barrel region and a disk in the forward regions, further extending the tracker acceptance to $|\eta| < 2.5$. The comparison of Phase0 and Phase1 layouts of the detector is shown in Fig. 2.6.

The Phase-1 pixel detector consists of four concentric barrel layers at radii of 29, 68, 109, and 160 mm, and three disks on each end at distances of 291, 396, and 516 mm from the

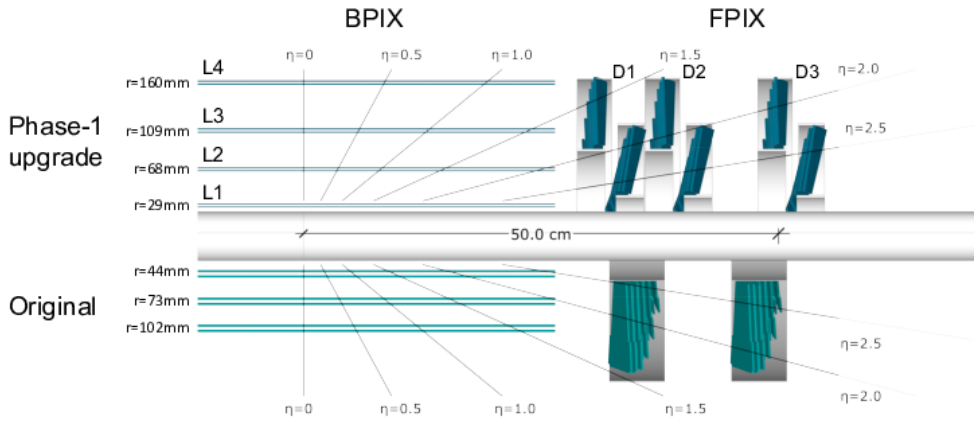



Figure 2.6: Layout of the Phase1 pixel detector (upper) compared to the Phase0 layout (lower) in longitudinal view. Figure taken from [47]

center of the detector [47].

The building block of the detector is a silicon sensor module. In the barrel part of the pixel detector (BPIX) 1184 modules are used and 672 modules in the forward disks (FPIX). A module is built from a planar, highly segmented, silicon sensor with a size of $18.6 \times 66.6\,\mathrm{mm}^2$ and a thickness of 285 $\mu$m, bump-bonded to an array of $2 \times 8$ read out chips (ROCs). The bump-bond is a small solder bump, providing the electrical connection between the sensor and the ROC. Each ROC is segmented into $52 \times 80$ pixel cells with dimensions of $100 \times 150\ \mu\mathrm{m}^2$. A high-density interconnect (HDI) printed circuit is glued to the other side of the sensor and wire-bonded to the ROCs. A token bit manager chip (TBM), located on the HDI, controls the readout of a group of ROCs. The described module components are shown in Fig. 2.7.

To detect the passage of a charged particle, a reverse polarization voltage is applied to the sensor. The electron-hole pairs created in the silicon by the charged particle will drift to the edges of the sensor and will be collected by the ROC. Since the modules are located in a magnetic field, a Lorentz force will be exerted on the charges, resulting in their drift towards neighboring pixels. This increases the charge sharing between the neighbouring pixels. The shape of the collected charge is exploited to enhance the resolution significantly below the pixel size, the measured resolution is approximately 10 $\mu$m in the $z$ direction and 20$\mu$ in the $r - \phi$ direction [47].

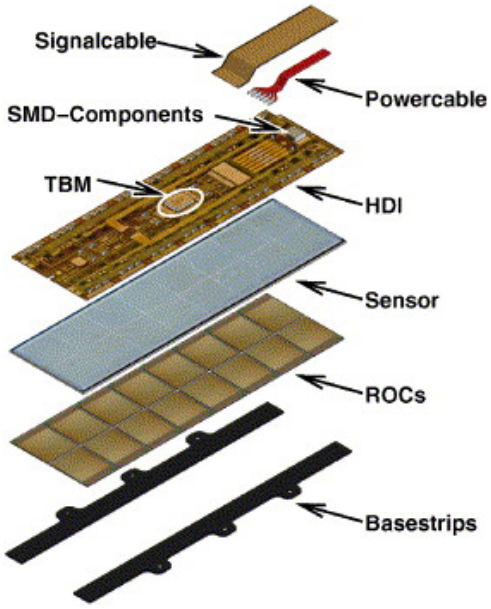



Figure 2.7: Exploded view of a barrel pixel detector module. Figure taken from Ref. [48]

Around the pixel detector is the silicon strip tracker. The strip detector consists of ten layers in the barrel region and twelve in the forward region. It is made of 15148 silicon modules, with an active area of about 200 $m^2$ [49]. It consists of four subsystems. The Tracker Inner Barrel (TIB) and Disks (TID) cover $r < 55\,\text{cm}$ and $|z| < 118\,\text{cm}$. The former is composed of four barrel layers and the latter of three disks at each end. The TIB and TID provide measurements in the $r - \phi$ direction with a resolution of 13–38 $\mu$m. The Tracker Outer Barrel (TOB) covers $r > 55$ cm and $|z| < 118$ cm. It consist of six barrel layers and provides measurements in the $r - \phi$ direction with a resolution of 18–47 $\mu$m. Finally, the Tracker Endcaps (TEC) cover the $124 < |z| < 282$ cm region. Each TEC is composed of nine disks and yields a range of resolutions similar to that of the TOB.

The silicon strips in the barrel are oriented parallel to the beam axis and the distance between the strips (pitch) varies from 80 $\mu$m in the inner TIB layers to 183 $\mu$m in the outer TOB layers. The endcap disks use wedge-shaped sensor with radial strips. Their pitch varies from 81 $\mu$m at small radii to 205 $\mu$m at large radii.

To provide the measurement of the other coordinate, a second strip detector is mounted back-to-back to some modules. Such modules are called stereo modules as they are rotated on a stereo angle of 100 mrad with respect to the regular modules. The hits from these module pairs can be combined to provide a measurement of the $z$ coordinate in barrel

and $r$ on the disks. The achieved single-point resolution of this measurement is an order of magnitude worse than in $r-\phi$ [49]. Stereo modules are mounted on the innermost two layers of both the TIB and the TOB and on the first two rings of the TID and rings 1, 2 and 5 of the TEC. The layout of different strip detector subsystems is shown in Fig. 2.8

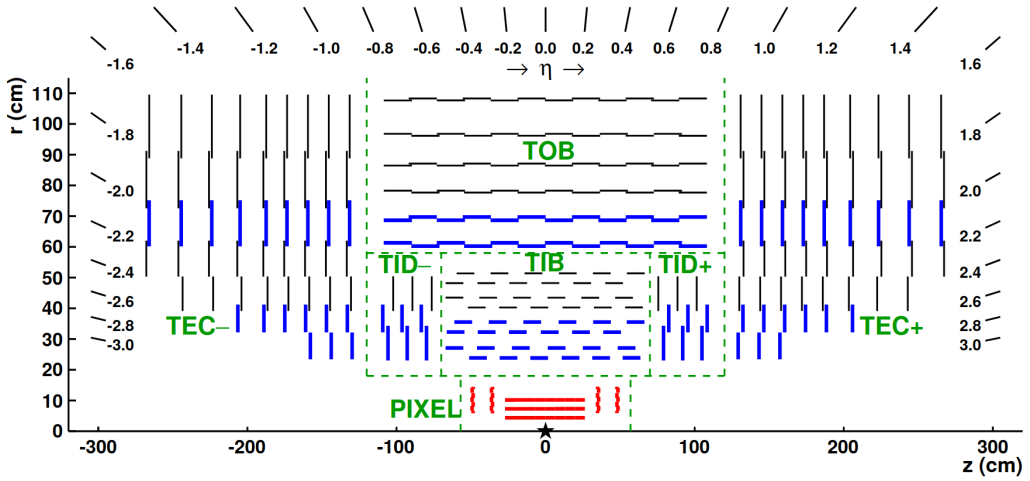



Figure 2.8: Schematic cross section through the CMS tracker in the r-z plane, showing the top half of the tracker. Strip tracker modules that provide 2-D hits are shown by thin, black lines, while those permitting the reconstruction of hit positions in 3-D (stereo modules) are shown by thick, blue lines. The pixel detector in the Phase-0 configuration is shown in red. Figure taken from Ref. [49]

The design of the tracker reflects the requirements needed to achieve sufficient track resolution which are the low detector occupancy and large redundancy of the measured points per track. The high granularity of the tracker detector provides the low occupancy of the detector. The granularity of the pixel detector is finer than the granularity of the strip detector since it is closer to the interaction point and therefore the flux of the particles is greater in the pixel detector. The redundancy in the number of measured particle positions (hits) is achieved by building multiple detection layers. However, it is desirable to reduce the material thickness of the tracker to minimum as the interaction of the particles with the material may change their trajectories, deteriorating the quality of track reconstruction. This limits the number of layers (material) of the tracking system. The material budget is expressed in units of radiation length ($X_0$) or the hadronic interaction length ($\lambda_I$). Radiation length is defined as the mean distance over which a high-energy electron losses all but $1/e$ of its energy by bremsstrahlung. It is also the 7/9 of the mean free path for the pair production of electrons by a high-energy photon. This makes it an

appropriate length scale for describing electromagnetic cascades. The hadronic interaction length is the expected distance that a hadron travels before a nuclear interaction occurs which makes it an appropriate scale describing the evolution of hadronic particle showers. The material budget of the tracker in the units of $X_0$ and $\lambda_I$ is shown in Fig. 2.9.

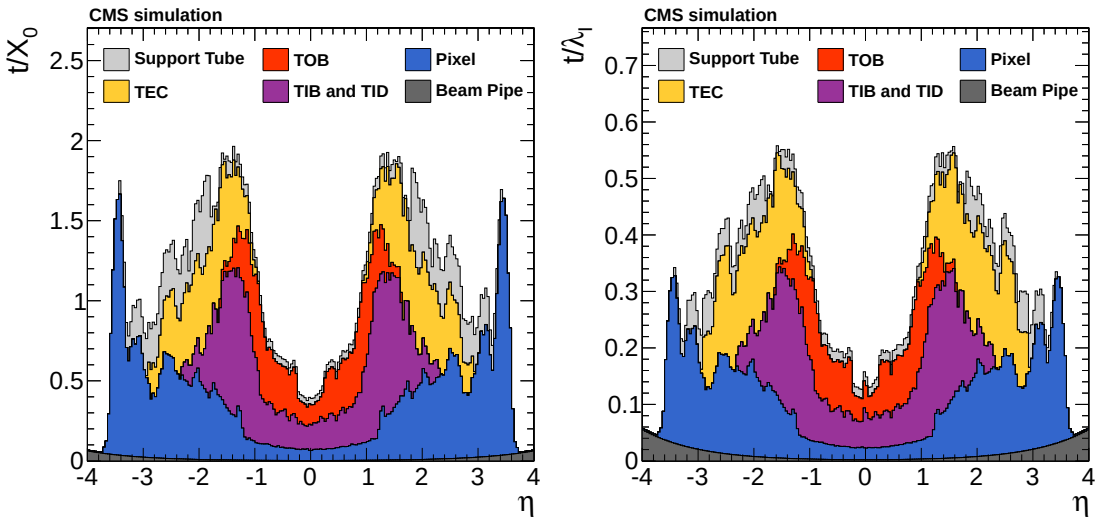



Figure 2.9: The tracker material budget expressed in units of radiation length (left) and hadronic interaction length (right). Figure taken from [49].

### 2.2.2 The Electromagnetic Calorimeter

The electromagnetic calorimeter (ECAL) is the second subdetector system, following the Tracker. Its primary goal is to measure the energies of electrons and photons. It consists of approximately 75000 lead tungstate ($PbWO_4$) scintillating crystals which produce light with the passage of ionizing particles. The main properties of the scintillating crystals are high density (8.28 g/cm$^3$), short radiation length ($X_0 = 0.85$ cm) and small Moliere radius, measuring the radius of a cylinder containing on average 90% of the electromagnetic shower's energy deposition, $R_M = 2.19$ cm. These properties allow the realization of a compact calorimeter with high granularity. Another important property is fast signal response, where 80% of the scintillation light is emitted in 25 ns which is the time separation between collisions. [50]

The two drawbacks of lead tungstate crystals are the reduced light yield, producing only $\sim$ 100 photons per MeV in a 23 cm long crystal, and a strong light yield dependence on

temperature ($-2\%/^{\circ}C$) at $\sim 18^{\circ}C$. The former necessitates the use of a photodetector readout with internal gain. Avalanche photodiodes are used in the barrel and vacuum photo-triodes in the endcap region. The light yield dependence on temperature is resolved by keeping the crystals at a stabilized temperature of 18°C.

The energy resolution of the crystals has been studied using beam tests [50]. The resolution for a central impact of electrons on a $3 \times 3$ crystal array was measured to be

$$\frac{\sigma_E}{E} = \frac{2.8\%\sqrt{\text{GeV}}}{\sqrt{E}} \oplus \frac{12.8\%\text{GeV}}{E} \oplus 0.3\%,$$

where $E$ is the measured energy in GeV. The first term is the statistical error coming from the stochastic nature of electromagnetic shower evolution and it dominates at low energies. The source of the second term is the electronic noise. The third, constant, term comes from detector non-uniformities. It dominates at high energies. The response of the crystals also changes with irradiation which leads to reduced crystal transparency (leading to reduced light collection). This damage can be partially recovered by keeping the crystals at room temperature for a few hours. The effect is known as thermal annealing. The irradiation damage is monitored by laser calibration.

The crystals are arranged in a central barrel section (EB), covering pseudorapidities up to $|\eta| = 1.48$, closed by the two endcaps (EE) which extend the coverage up to $|\eta| = 3.0$. The crystals are of trapezoidal shape, with a length of 23 cm and frontal and rear bases of $22 \times 22\,\text{mm}^2$ and $26 \times 26\,\text{mm}^2$ in the EB. These dimensions correspond to 25.8 $X_0$ in longitudinal length and 1 Moliere radius in the transverse plane. The crystals in the EE are 22 cm long with a frontal base of $28.6 \times 28.6\,\text{mm}^2$ and rear base of $30 \times 30\,\text{mm}^2$ [46].

A preshower detector (PS) is installed on the inner side of the endcaps. The goal of the PS is to enable the ECAL to separate showers originating from the decay of a neutral pion into two photons, $\pi^0 \to \gamma\gamma$, which would otherwise be absorbed in the same crystal and reconstructed as a single photon. The PS is based on a two-layer lead absorber and silicon strip sensor detector. It is able to identify a photon-initiated shower in the lead and resolve the shower in the silicon strips which have a resolution of 1–10 mm. The layout of the ECAL and its major components is shown in Fig. 2.10

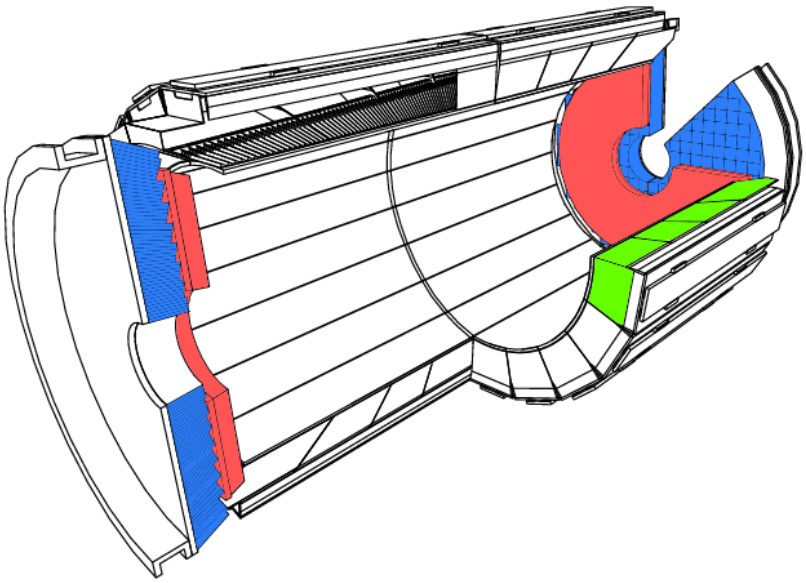


Figure 2.10: Layout of the CMS ECAL, showing the barrel modules (green), the two endcaps (blue), and the pre-shower detectors (red). Figure taken from Ref. [51]

### 2.2.3 The Hadronic Calorimeter

The hadronic calorimeter (HCAL) measures the energy of hadrons which penetrate the ECAL while depositing only a small fraction of their energies. Unlike ECAL, in which the entire volume of the crystals is sensitive and contributes to a signal, HCAL is built of alternating layers of a dense material, absorber, and a sensitive detector, scintillator. Calorimeters in which the material that produces the particle shower is distinct from the material that measures the deposited energy are called "sampling" calorimeters. The idea is to force the hadrons to interact with the nuclei of the absorber, creating multiple lower energy hadrons. The energy of the products of these interactions is recorded with the scintillators. An advantage of this is that a very dense material can be used to produce a shower that evolves quickly in a limited space, even if the material is unsuitable for measuring the energy deposited by the shower. HCAL uses brass as the absorber material, it has a radiation length of $X_0 = 1.49\,\text{cm}$ and interaction length $\lambda_I = 16.42\,\text{cm}$.

A disadvantage is that some of the energy is deposited in the wrong material and is not measured; thus the total shower energy must be estimated.

The HCAL and ECAL constitute a complete calorimetric system, with HCAL extending the acceptance to $|\eta| < 5$. Because of this almost hermetic coverage, the calorimetric sys-

tem can also provide indirect measurement of the non-interacting neutrinos by measuring the missing momentum in the transverse plane. This is usually referred to as the missing transverse energy ($E_T^{miss}$).

The HCAL is divided into four subdetectors: Barrel (HB), Endcap (HE), Outer (HO) and Forward (HF). The layout of one quarter of the HCAL can be seen in Fig. 2.11

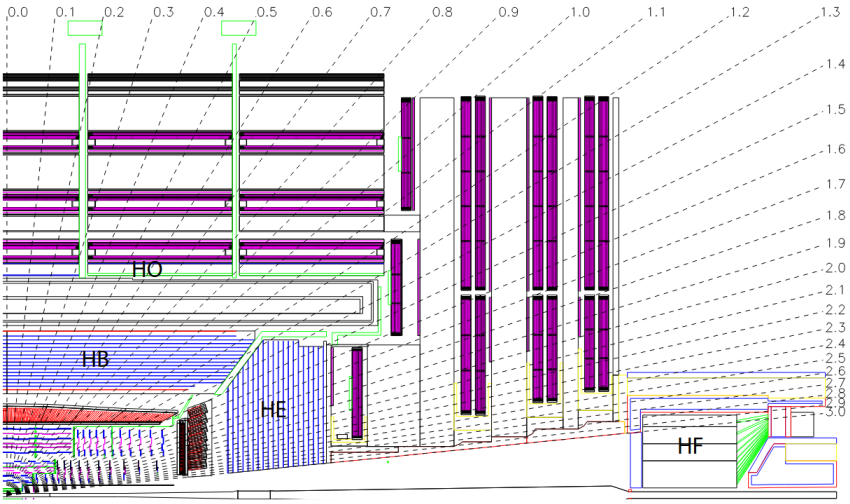



Figure 2.11: Layout of the CMS HCAL showing the four components: HCAL Barrel (HB), HCAL Outer (HO), HCAL Endcap (HE) and HCAL Forward (HF). Figure taken from Ref. [52]

The Barrel Hadronic Calorimeter covers the pseudorapidity range $|\eta| < 1.3$. It consists of 36 identical azimuthal wedges which form the two halves of the barrel (HB+ and HB-). Each wedge is divided into four sectors along the azimuthal angle, $\phi$. Furthermore, the plastic scintillators in each half-barrel are divided along $\eta$ into 16 sectors. This results in a segmentation of 2304 calorimeter towers ($36 \times 4 \times 16$) with the segmentation of $(\Delta\eta, \Delta\phi) = (0.087, 0.087)$ [52]. The absorber of each calorimeter tower consists of a front steel plate, followed by eight brass plates of 50.5 mm thickness, six brass plates of 56.5 mm thickness and the steel back plate. Each tower also contains 17 layers of plastic scintillators with a 3.7 mm thickness. The exception is the inner and outermost layers which are 9 mm thick to capture hadronic showers developing in the inert material between EB and HB and late developing particle showers leaking out of the HB, respectively. The total absorber thickness of the HB ranges from $5.82\lambda_I$ at $|\eta| = 0$ to $10.6\lambda_I$ at $|\eta| = 1.3$. The ECAL in front of the HB adds about 1.1 $\lambda_I$ of material.

The Endcap Hadronic Calorimeter (HE) covers the $1.3 < |\eta| < 3$ range. It shares the same

working principle and has a similar design as the HB. The absorber plates in the HE are 79 mm thick with 9 mm gaps which accommodate the scintillators. The total ECAL and HCAL endcap material amounts to about 10 $\lambda_I$. The empty region between the HB and HE is used to pass readout cables.

The combined stopping power of EB and HB in the central pseudorapidity region does not provide sufficient containment for hadron showers. For this reason, the hadron calorimeter is extended outside the solenoid with the HCAL Outer calorimeter (HO) [52]. The solenoidal magnet serves as an additional absorber equalling $1.4/sin\theta$ $\lambda_I$. Since the minimum of the absorber depth is at $\eta = 0$, the HO consists of two layers of plastic scintillators centered at $z = 0$ positions, while the other four HO sections, centered at $z = \pm 2.686, \pm 5.342$ m, have a single layer. The total depth of the calorimeter system (ECAL + HCAL) is therefore extended to a minimum of 11.8 $\lambda_I$, excepting the barrel-endcap boundary region.

Finally, the HCAL forward calorimeter (HF) [52] extends the calorimeter coverage to $|\eta| < 5$. It uses different materials than the rest of the HCAL due to much higher particle fluxes in the forward region. The projected radiation damage does not permit the use of plastic scintillators used in the rest of the HCAL. The detector uses quartz fibers as the active material, embedded in a 1.65 m long steel absorber (approximately $10\lambda_I$). The signal in the fibers is generated by the passage of a charged particle above the Cherenkov threshold ($E > 190$ keV for electrons) and is manifested as the emission of Cherenkov light. Hence, the detector is mostly sensitive to the electromagnetic component of the particle showers.

The hadronic energy resolution of the barrel HCAL, when combined with ECAL, is

$$\frac{\sigma}{E} = \frac{0.85\sqrt{GeV}}{\sqrt{E}} \oplus 7.4\%$$

The energy resolution in the endcaps is similar to that in the barrel. The resolution in HF is found to be

$$\frac{\sigma}{E} = \frac{200\sqrt{GeV}}{\sqrt{E}} \oplus 9\%$$

Since the forward jets typically have very high energies, the energy resolution of the HF is still satisfactory even with the large stochastic term [53].

### 2.2.4 Muon chambers

The system of muon chambers is the outermost CMS detector system. The material in between the collision point and the muon chambers filters electrons, photons and hadrons. 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. [54]

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. As shown in Fig. 2.12, the system consists of three subsystems: drift tubes (DT), resistive plate chambers (RPC) and cathode strip chambers (CSC). The operating principle is similar in all three subsystems. A muon ionizes the gas in the detector and the charge is collected by wires or strips, signalling the presence of a transversing charged particle. The muon system is divided into stations which are assemblies of muon chambers around a fixed value of $r$ in barrel and $z$ in the endcap. There are 4 stations in the barrel (labelled MB1-4) and in each endcap (labelled ME1-4). The barrel is also divided along $z$ into 5 wheels. Wheel 0 is centered at $z = 0$, wheels W+1, W+2 are located in the $+z$ direction and W-1, W-2 in the $-z$ direction. Similarly, the endcap is divided into three rings along the $r$ direction. The division into wheels and rings is also shown in Fig. 2.12.

The DT chambers cover the pseudorapidity region $|\eta| < 1.2$. They are organized into 12 $\phi$ segments per wheel, forming 4 stations at different radii as shown in Fig. 2.12. The DT systems consist of rectangular drift cells with a transverse size of $13 \times 42\,\mathrm{mm}^2$ and 2 to 4 m length, shown in Fig. 2.13. Four layers of parallel cells form a superlayer (SL). Each DT chamber consists of 2 SLs that measure the $r - \phi$ coordinates (wires parallel to the beam line) and an orthogonal SL that measures the $r - z$ coordinate. [54]. The spatial

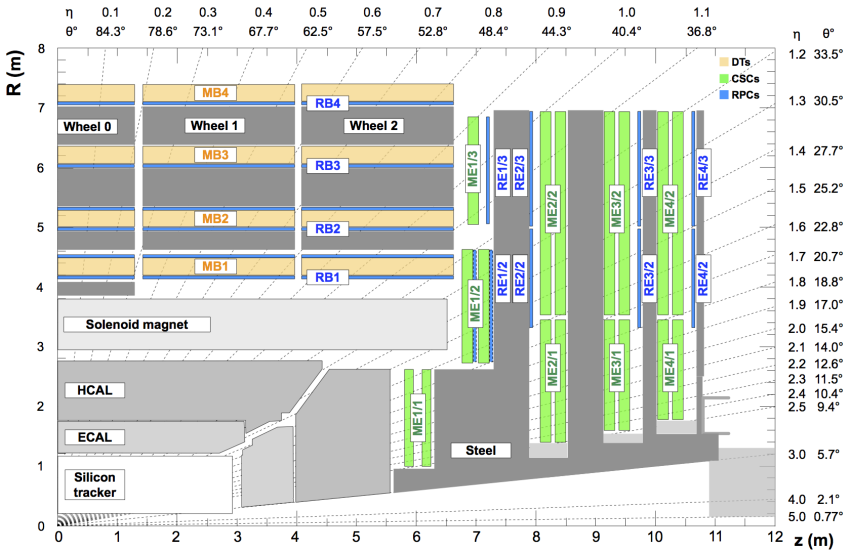



Figure 2.12: $r - z$ quadrant of the CMS detector highlighting the three CMS muon subdetectors: DT in yellow, CSC in green and RPC in blue. Barrel Wheel 0 and two Wheels at positive $z$ axes are shown as well as the separation into rings in the endcap. Figure taken from Ref. [55]

resolution per chamber is 80–120 $\mu$m. [54].

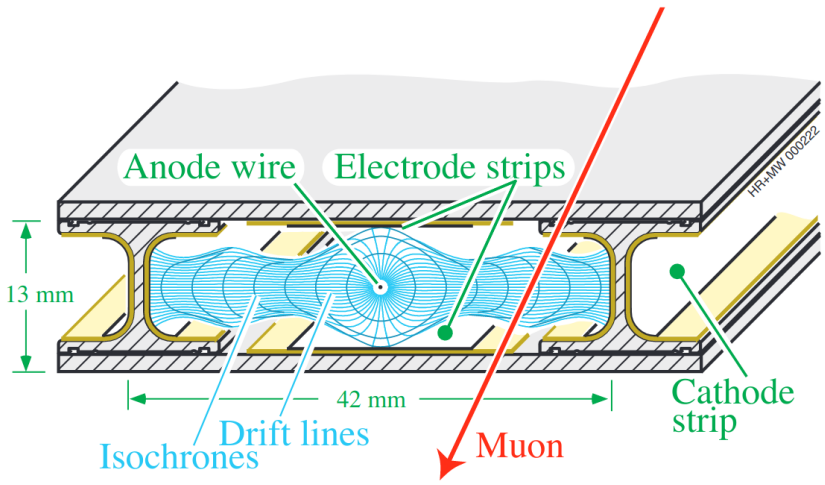



Figure 2.13: Section of a drift tube cell showing the geometry of the cell, drift lines and isochrones. Figure taken from Ref. [54]

The large particle flux at high $\eta$ and a relatively high magnetic field in the endcap region do not allow the usage of DT. Therefore, the usage of CSCs was accepted for the endcap region [54]. CSCs are multiwire proportional chambers filled with a gas mixture of Ar (40%), $CO_2$ (50%) and $CF_4$ (10%). 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 $\mu$m range [54].

A crucial property of the DT and CSC systems is that their timing resolution is well below the bunch crossing spacing of 25 ns. This enables them to identify the collision bunch crossing that generated the muon and trigger on the presence of muons with good efficiency [54].

The third muon subsystem, the RPCs, provide a complementary muon triggering system which reinforces the measurement of the correct bunch crossing time with their excellent time resolution of just one nanosecond. RPCs are located in both the barrel and endcap, covering the pseudorapidity range $|\eta| < 1.6$ as shown in Fig. 2.12. 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 [54].

### 2.2.5 The Trigger

The LHC provides a collision rate of 40 MHz, however, not all of the collisions are recorded. On the contrary, most of them are rejected for two reasons. First, the 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. The second reason is that most of the events are not really interesting for the physics goals set by CMS as they originate from well-known processes. As shown in Fig. 2.14, 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.

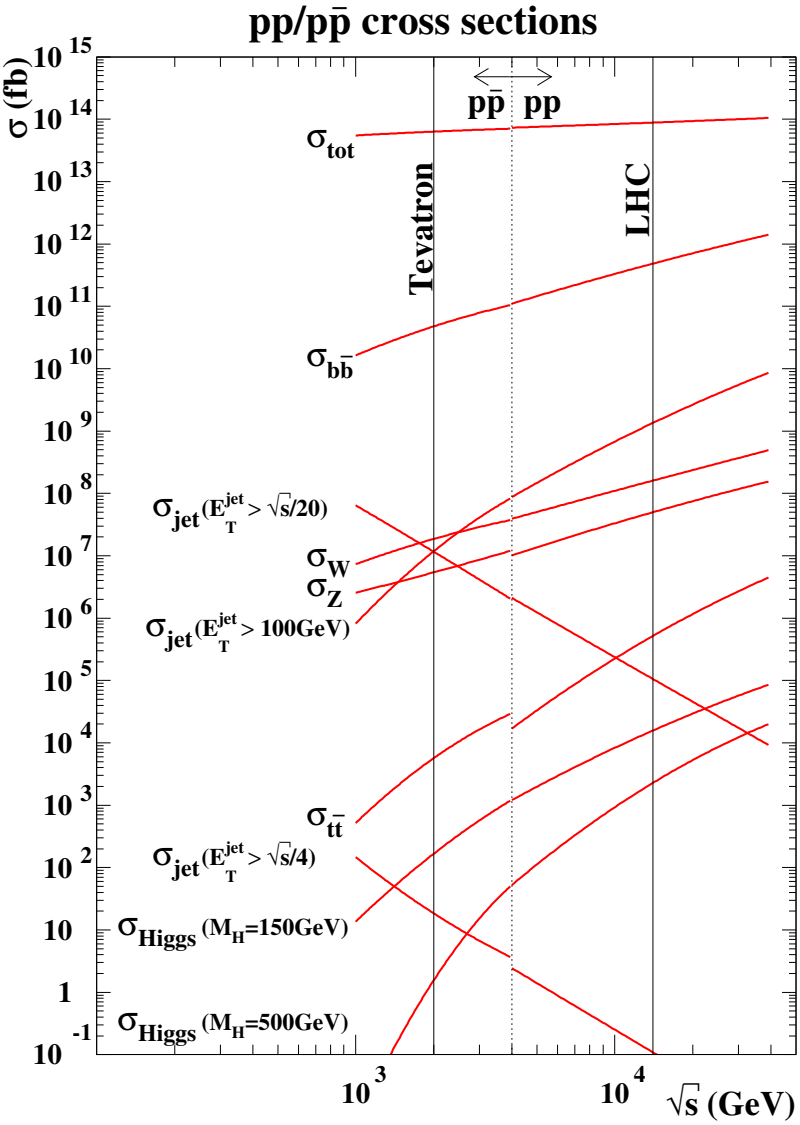



Figure 2.14: SM cross sections at hadron colliders as a function of the center of mass energy, $\sqrt{s}$, for several processes. Figure taken from Ref. [56]

Therefore, even if it was feasible to store all collision events, vast majority of them would not be of interest. Events of interest are selected using a two-tiered trigger system. 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 100kHz within a fixed latency of about 4 $\mu$s [57]. 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 1kHz before data storage [58].

**Level-1 trigger**

The L1 trigger uses trigger primitives (TP) from ECAL and HCAL and from the muon detectors. The information from the tracker is not used since it cannot be read out and processed fast enough. The TPs are processed in several steps before the combined event information is used to make a decision whether to pass the event to HLT or not.

The L1 calorimeter trigger contains two stages, a regional and a global calorimeter trigger (RCT and GCT). The RCT looks for cluster of signals collected both by ECAL and HCAL. It sends four isolated and four nonisolated $e/\gamma$ candidates and regional sums of transverse energy to the GCT. The GCT further sorts the $e/\gamma$ candidates and classifies the regional $E_T$ sums into central, forward and tau jets [58]. It sends as output to the L1 Global trigger four $e/\gamma$ candidates each of two types, isolated and nonisolated, four each of central, tau, and forward jets, and several global quantities such as the missing transverse energy.

The L1 muon trigger system uses all three types of muon chambers. DTs and CSCs trigger electronics separately build muon segments which are further processed by track finding algorithms to create muon candidates. Theses candidates are assigned $p_T$ and a quality flag and send to the Global Muon Trigger. Unlike the DTs and CSCs, the RPC trigger electronics directly uses the detector hits for muon trigger candidate recognition. The pattern comparator trigger logic compares signals from all RPC chamber layers to predefined hit patterns in order to find muon candidates. The RPCs assign the muon $p_T$, charge, $\eta$ and $\phi$ to the matched pattern.

The output of the muon detectors is combined in the Global Muon Trigger to create a final set of muons. The global muon trigger merges or cancels duplicate candidates and sorts the candidate set according to their quality and $p_T$. Four best candidates are send to the L1 Global Trigger.

The L1 Global trigger decides whether to readout the rest of the event and pass it to the HLT or reject it. Certain conditions are applied to the various objects passed to the Global trigger (isolated and nonisolated $e/\gamma$ object, muons and central, forward or tau jets) such as $p_T$ being above a certain threshold, $\eta$ and/or $\phi$ being within a selected window. More complex conditions can also be calculated, for example the $\Delta\eta$ between two objects, $E_T^{miss}$ or the $p_T$ sum of jets ($H_T$). Several conditions are combined by simple combinatorial logic to form up to 128 algorithms [58].

**High-level trigger**

The event selection at the HLT uses objects which are similar to those used in the offline processing, meaning more complex and accurate reconstruction than at L1. For example, HLT uses information from the tracker which is not available to L1. Object reconstruction for offline processing is described in Chapter 3. Some difference between the offline processing and HLT is introduced in order to comply with the time requirement to process an event within about 300 ms.

Similar to L1, HLT uses a set of different selection criteria (HLT paths), based on the properties of reconstructed objects (electrons, photons, muons and jets), to select potentially interesting events. The trigger menu in Run II contained more than 400 different HLT paths. Each path is a sequence of reconstruction and filtering modules. The modules within a path are arranged in blocks of increasing complexity, so that faster algorithms are run first [59].

# Chapter 3

# Event simulation and reconstruction

Besides studying SM processes at high energies, in HEP experiments we also compare the observation to the SM predictions in order to search for discrepancies which may point to new physics. It is difficult to directly relate theory predictions to measured quantities due to many detector effects which are hard to calculate. We therefore rely upon simulating the effect of the detector on physical observables and compare the simulated distributions with the data. The simulation of collision events predicted by SM or BSM theories is described in steps in Sec. 3.1. The simulation includes not only the generation of events from pp collisions, but also the propagation and interaction of the produced particles with the detector. Another example of the simulation of propagation and interaction of particles with material, for the purpose of determining dose rates in a radiation facility, is given in Appendix A.

The goal of event reconstruction is to identify and measure particles present in the event using the signals left in the detector. In CMS, a dedicated algorithm, called Particle-flow [60] (PF), is used for the event reconstruction. An overview of the PF algorithm is presented in Sec 3.2. A more detailed description can be found in Ref. [60]. This analysis is mainly concerned with the reconstruction of b quarks since the final state of interest contains four of them. However, quarks and gluons, collectively referred to as partons, cannot be directly observed in the detector. Instead, the successive splittings associated with parton shower and hadronization give rise to a shower of collimated particles which we recombine into an object called hadronic jet (or just jet). Since jets are of utmost

importance for this analysis, the description of the PF algorithm is followed by two sections focusing on the hadronic jets at CMS. A description of the jet reconstruction algorithms and jet calibration methods is given in Sec. 3.3. The usage of flavor tagging to select jets which contain one or two b quarks is given in Sec. 3.4.

## 3.1 Event simulation

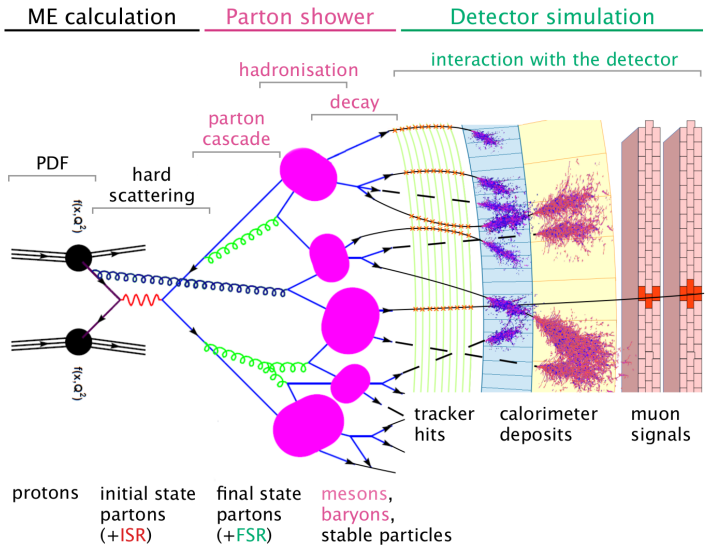



Figure 3.1: Illustration of the key steps of the event simulation procedure. Figure taken from [61]

The key steps in the event simulation are shown in Fig. 3.1. The process of event simulation starts with the incoming protons. The parton distribution functions, described in Sec. 3.1.1, 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. 3.1.2. It determines the type and kinematic properties of the final state particles. The parton-showering step, described in Sec. 3.1.3 describes the interaction of the remaining constituents of the incoming protons, not participating in the hard interaction. It also deals with initial- and final-state radiation as well as the shower evolution and hadronization of strongly interacting particles. Finally, the interaction of stable particles with the detector is simulated to obtain the detector response, briefly described in Sec. 3.1.4.

### 3.1.1 Parton Distribution Functions

Protons are composite particles made of three valence quarks (two u and one d quark), bound by gluons which can produce short lived quark-antiquark pairs of all flavors, the so-called "sea" quarks. A consequence of this is that in pp collisions, the hard interaction actually occurs between single constituents of each proton. Any of the constituents (partons) may be involved in the hard scattering of a given pp collision, be it valence quarks, sea quarks or gluons. The interacting parton only carries a fraction of the proton's momentum, $p = x \cdot P$. 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 one the energy scale $Q^2$ the proton is probed at. Therefore, PDFs are denoted as $f_i(x, Q^2)$, where $i$ represents the parton flavor (gluon or one of the quarks).

The usage of PDFs is justified by the QCD factorization theorem [62]. The theorem separates the cross section calculations in QCD into a contribution related to the hard process, which can be calculated using perturbative QCD, and the contribution from the internal structure of the proton, expressed through the PDFs. The factorization depends on an arbitrary energy scale which separates the two regimes, the factorization scale $\mu_F^2$. The theorem enables us to write the cross section for a pp $\to X$ collision process:

$$\sigma(\text{pp} \to X) = \int_0^1 dx_i dx_j \sum_{i,j} f_i(x_i, \mu_F^2) f_j(x_j, \mu_F^2) \sigma(ij \to X)(Q^2, \mu_F^2), \tag{3.1}$$

where $x_{i,j}$ are the the fractions of the proton momentum carried by partons $i, j$, $f_{i,j}$ are the proton PDFs for partons $i, j$ and $\sigma(ij \to X)(Q^2, \mu_F^2)$ is the cross section of the hard-scatter process of two partons obtained with perturbative calculation.

Since PDFs cannot be calculated with perturbative QCD, they are parametrized from experimental data of many collider experiments. The dependence of the PDFs on the scale can, however, be described with the Dokshitser-Gribov-Lipatov-Altarelli-Parisi (DGLAP) evolution equations [63, 64, 65, 66]. DGLAP evolution makes it possible to use the measured PDFs at one scale to calculate the PDFs at different scales. An example of a

difference between PDFs at two different $\mu_F$ can be seen in Fig. 3.2.

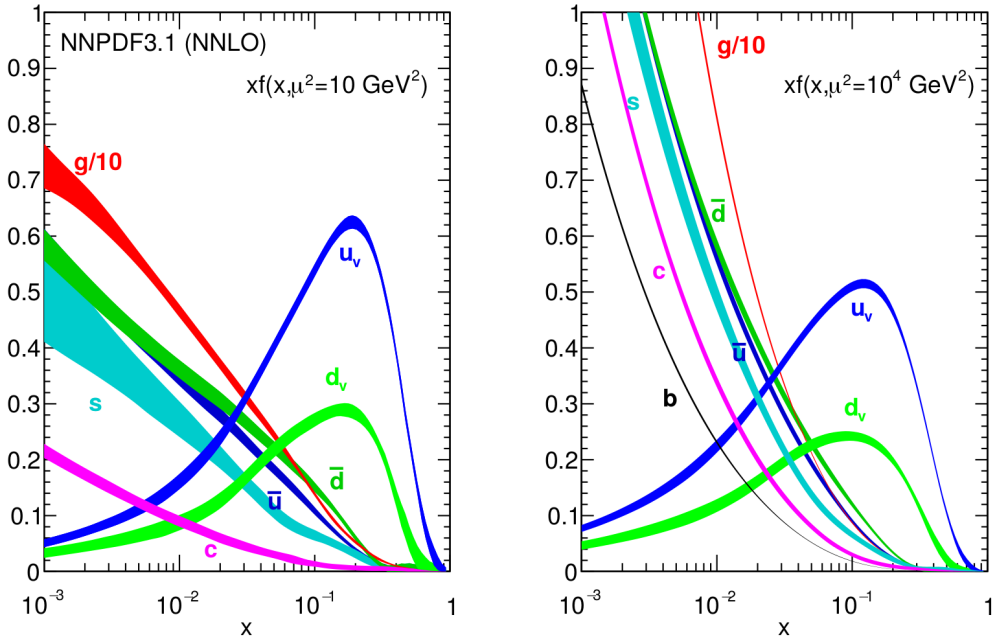



Figure 3.2: The PDFs evaluated at $\mu_F^2 = 10\,\mathrm{GeV}^2$(left) and $\mu_F^2 = 10^4\,\mathrm{GeV}^2$(right). The PDFs are part of the NNPDF3.1 set[67]. Figure taken from [67]

### 3.1.2 Matrix Element

The main characteristics of an event are given by the process with the largest momentum transfer, the hard process. In event simulation, the hard process is represented by the calculation of the matrix element. It is related to the probability of the transformation from the initial to the final state of a particular process. The initial state is given by two incoming, interacting partons and the final state by a varying number of particles.

Computation of the matrix element involves the calculation of Feynman diagrams up to the desired accuracy in perturbation theory. The first contribution in the expansion, the leading order (LO), is relatively simple to calculate and provides a good, albeit coarse, estimate of the matrix element. Subsequent orders, next-to-leading order (NLO), next-to-next-to-leading order (NNLO) and so on, improve the accuracy of the matrix element calculation but the Feynman diagrams become more complicated and the number of diagrams drastically increases. For these reasons, NLO and NNLO calculations are available only for a small number of processes and LO matrix elements are still widely used. The calculations at finite orders come with some restrictions. Divergences occur when pro-

cesses containing low-momenta partons or with collinearly emitted partons with respect to the radiating partons. Theses divergences are known as infrared and collinear divergences. The phase-space covered by the matrix element calculation is chosen to avoid the divergences by limiting the collinearity of partons and setting a lower threshold on their momenta. The excluded phase space needs to be recovered in the later steps of event simulation.

This analysis uses MadGraph [68] and POWHEG [69, 70, 71] event generators to calculate the matrix elements for event generation.

### 3.1.3 Parton showering and hadronization

Parton showering (PS) algorithms cover the initial (ISR) and final state radiation (FSR) of partons which can occur in a form of: $q \rightarrow qg$, $g \rightarrow qq$, $g \rightarrow gg$. The partons created by ISR and FSR can undergo further splitting in a cascade process, leading to the creation of a large number of partons. This process is known as parton showering. The parton showering procedure starts from the hard process final state, obtained by the matrix element, and evolves the event with successive random splittings. The splitting evolution is modelled with the Sudakov form factors [72] which, for each parton, give the probabilities for no emissions taking place between two momentum scales. Each subsequent splitting lowers the energy scale of the partons until the energies of all partons are individually less than a cutoff scale, $\Lambda_{QCD}$. At this scale the effects of color confinement start to dominate, causing partons to form color-neutral hadrons in a processs called hadronization. Before moving on to hadronization, the interfacing of the matrix element and parton showering algorithms needs to be mentioned. Since both the matrix element and parton showering simulate the emission of additional partons, matching algorithms need to be applied to prevent double counting. This can be done by introducing cutoffs which can ensure that emissions above particular momentum or a certain angle are described by matrix element and the rest of the phase space by the parton shower. Two examples of matching algorithms are MLM [73] and CKKW [74].

The hadronization process takes place at low energy scales, in a regime where perturbative calculations are no longer valid. Therefore, it is described by phenomenological models,

like the Lund string model [75]. The model introduces QCD field lines between quarks, which can be interpreted as strings storing potential energy. With the increasing distance between the quarks, the string is stretched until the potential energy becomes large enough to create a quark-antiquark pair from the vacuum. The two new quarks form the strings with the original quarks so that the total potential energy of the two new strings and the energy needed to create the two quarks is lower than the energy of the string before breaking. The Lund string model is used by the PYTHIA [76] software which is used in this analysis for parton showering and hadronization.

The PS algorithms also take into account objects produced in association with final states from hard process, called the underlying event (UE). The UE originates from the interactions of the remaining constituents of collided hadrons, not involved in the hard interaction. The UE description is complicated because of the combination of perturbative and non-perturbative processes. The models describing UE are configured by measuring variables such as particle multiplicities and matching the simulation prediction to observation [77, 78]. A set of configuration parameters used to describe the UE is referred to as the tune.

### 3.1.4 Detector simulation

The final step in event generation is the simulation of the detector's response to the transversing particles. A simulation of the entire detector is based on a toolkit for the simulation of the passage of particles through matter, GEANT4 [79]. 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.

The GEANT4 simulation software is widely used in other HEP experiments and even in other areas such as radiation medicine. One example is detailed in Appendix A which

describes the GEANT4 simulation setup developed to estimate dose rates within a $^{60}$Co facility at the Ruđer Bošković Institute (RBI).

## 3.2 Particle-flow algorithm

The Particle flow algorithm (PF) reconstructs and identifies individual particles by combining information for the various CMS subdetectors. There are five classes of PF particle candidates: electrons, photons, charged and neutral hadrons, and muons.

The passage of each of these particles through the CMS detector results in different detection signatures, as shown in Fig. 3.3. The differences in the signatures can then be exploited for the classification of the particles. An electron leaves hits in the inner tracking system and deposits its energy in the ECAL. A photon also deposits its energy in the ECAL, but it does not produce hits in the tracker since it is electrically neutral. Similar relation exists between charged and neutral hadrons. Both of them deposit their energy mostly in the HCAL, while only charged hadrons produce hits in the tracker. Finally, a muon produces hits in both the tracker and the muon detectors while depositing only small amounts of energy in the calorimeters.

The PF algorithm starts by building the basic elements which will be used to reconstruct particle candidates. These are the reconstructed charged-particle (inner) tracks, ECAL and HCAL calorimeter clusters and muon tracks. A given particle may (and usually does) give rise to multiple PF elements in various CMS subdetectors. For example, a charged hadron is expected to leave a charged-particle track and a cluster in the HCAL.

The next step in the PF algorithm is the formation of links between PF elements from different subdetectors. The criteria for linking two elements are described in the following paragraphs. If a link is established between two elements, a distance between them is calculated that measures the quality of the link. The link algorithm produces PF blocks of elements which can be associated either by a direct link or an indirect link through common linked elements.

To establish a link between an inner tracker track and a calorimeter cluster, the track is extrapolated from its last measured tracker hit to the corresponding calorimeter. The

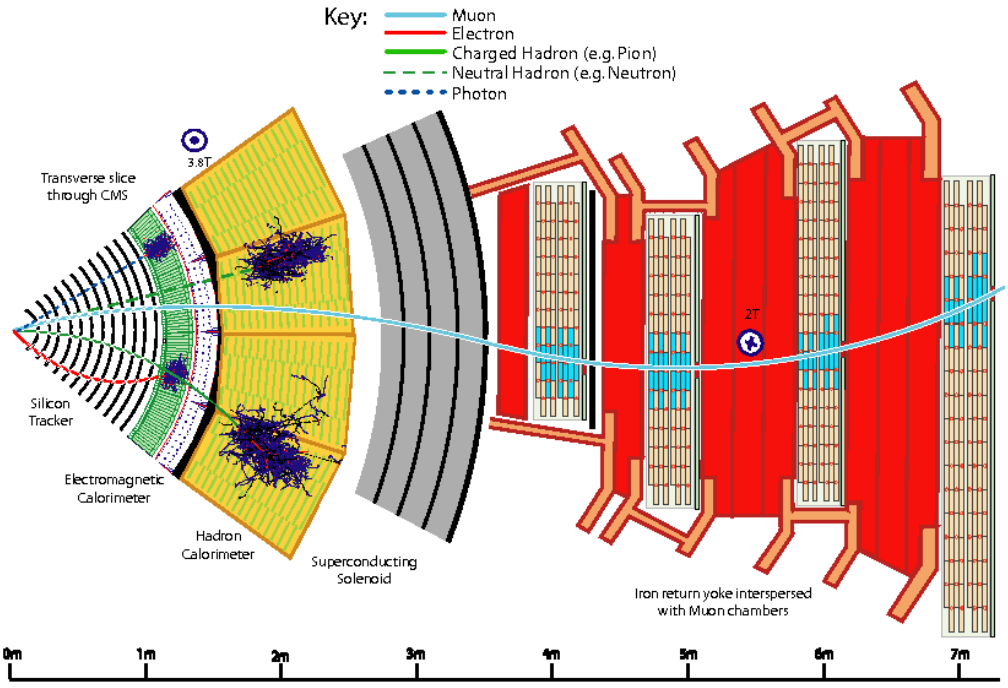



Figure 3.3: A sketch of the specific particle interactions in a transverse slice of the CMS detector, from the beam interaction region to the muon detector. The muon and the charged pion are positively charged, and the electron is negatively charged. Figure taken from [60]

track is linked to a cluster if its extrapolated position is within the cluster area. The link distance an between inner tracker track and a calorimeter cluster is defined as the distance in the $\eta - \phi$ plane between the track extrapolated to the calorimeter and the calorimeter cluster. If several HCAL clusters are linked to the same track, only the link with the smallest distance is kept. Similarly, if several tracks are linked to the same ECAL cluster, only the link with the smallest distance is kept.

Tracks may also be linked together through a common secondary vertex. Displaced vertices are required to have at least three tracks, where no more than one may be an incoming track. An incoming track is the one which is reconstructed with hits between the primary vertex and the displaced vertex. The primary vertex is taken to be the vertex corresponding to the hardest scattering in the event, evaluated using tracking information alone, as described in Section 9.4.1 of Ref. [80]. All the tracks sharing a displaced vertex are linked together.

A link can be formed between ECAL and HCAL clusters if the cluster position in the ECAL is within the cluster envelope of the HCAL. In that case, the distance is defined

as their distance in $\eta - \phi$ plane. If multiple HCAL clusters are linked to the same ECAL cluster, only the link with the smallest distance is kept.

For links between an inner tracker track and a muon track, the distance is the $\chi^2$ of a global-muon track fit, described later.

PF blocks are the input for the particle reconstruction and identification part of the PF algorithm in the following order. First, muon candidates are identified and reconstructed. The corresponding PF elements are removed from the PF block. The reconstruction and identification of electron candidates is done next. An attempt is made at identifying bremsstrahlung photons and adding them to the electrons. Any remaining isolated photons are also identified. The corresponding PF elements, tracks (electrons only) and ECAL clusters, are removed from the PF block. The remaining elements in the block are used to identify hadrons and nonisolated photons. Once all blocks have been processed, the stored collection of PF candidates represent a global description of the event. This collection is used in the construction of higher-level objects used in physics analyses.

### 3.2.1 Muons

The muon subdetectors allow muons to be identified with high efficiency. Their position also grants a high purity since the upstream calorimeters absorb other particles. Muon tracks are also reconstructed with the help of the inner tracker, providing more precise measurements of the muon momenta. There are three categories of muon tracks reconstructed by the PF algorithm.

- Standalone muons: Muon tracks which are formed using only the hits from the muon subdetectors.
- Global muons: If there is a compatible track in the inner tracker, a standalone muon track is matched to it. The hits in both the inner tracker and the standalone-muon track are combined and fit to form a global muon.
- Tracker muons: Reconstructed by extrapolating the inner track to the muon system. If at least one muon segment matches the extrapolated track, the inner track is qualified as tracker muon track.

Global and tracker muon track reconstructions are complementary. Tracker muon reconstruction is more efficient at low momenta, $p < 5$ GeV, because it only requires a single muon segment in the muon system. On the other hand, global muon reconstruction is designed to have high efficiency for higher momenta muons which leave hits in more than one muon station.

The identification of muons is done by applying a set of selections on global and tracker muon properties. Isolated global muons are selected by applying a requirement on an isolation variable which is computed by adding contributions of additional inner tracks and calorimeter depositions with $\Delta R < 0.3$ to the muon direction in the $\eta - \phi$ plane. The $p_T$ sum of the tracks and of the deposits in the calorimeters is required to be less than 10% of the muon $p_T$. This criterion is sufficient to adequately reject hadrons that may be misidentified as muons. This may happen if a muon is created a within hadron shower or if some of the hadron shower reaches the muon system, known as "punch-through".

For nonisolated global muons, the tight-muon selection is applied. Tight muons require a global muon track fit which includes at least one muon-chamber hit and that the fit returns $\chi^2/N_{dof} < 10$. The candidate should also be a tracker muon with at least two matched muon segments in different muon stations. Its inner track needs to be reconstructed from at least five inner-tracking layers, out of which at least one needs to be a pixel detector layer. Finally, conditions are placed on the track's transverse impact parameter, $d_{xy} < 2$ mm, and longitudinal impact parameter, $d_z < 5$ mm, with respect to the primary vertex.

Muons failing the tight-muon selection are salvaged if the standalone muon track fit is of high quality and associated with a large number of hits in the muon detectors. However, muons reconstructed only as standalone-muon tracks have worse momentum resolution and higher admixture of cosmic-ray muons than the Global and Tracker Muons and are usually not used in physics analyses [81].

The PF elements of identified muons are removed from the corresponding PF blocks and are not further processed as building elements for other particles.

In the presented analysis we use the above-described tight-muon selection to identify muons from leptonic decay of the t quark (Sec. 4.3.3). Loose muon identification is used to veto the presence of muons in the hadronic category of analysis where no leptons

are expected(Sec. 4.3.2). The loose muon identification only requires the muon to be reconstructed as Tracker or Global Muon.

Additional isolation criteria are imposed to reject hadrons that would be misidentified as muons. The muon isolation is defined as the sum of $p_T$ of the particles inside a cone around the muon with the radius $\Delta R = 0.4$ divided by the muon's $p_T$. The two isolation working points used in the analysis are the very loose and tight criteria which require the isolation to be $< 0.4$ and $< 0.15$, respectively.

### 3.2.2 Electrons and isolated photons

Electron reconstruction is based on the combined information from the inner tracker and the energy deposits in the ECAL. It is done simultaneously with the photon reconstruction.

Due to significant tracker material thickness, electrons will often emit bremsstrahlung photons before reaching the ECAL. To capture those emissions, ECAL clusters (EC) within a small window in $\eta$ and extended window in $\phi$ around the electron direction are grouped into "superclusters" (SC). Furthermore, these emissions make it more difficult to reconstruct electron tracks. For this reason, electron tracks are fitted using a "Gaussian sum filter" (GSF) [82] algorithm, instead of a "Kalman Filter" (KF) used for other tracks. This method allows for fitting a particle's trajectory with sudden, substantial energy losses. The GSF track fitting algorithm is too CPU intensive to be run over all tracker hits. The electron track reconstruction begins with the identification of a hit pattern ("seed") that may be part of an electron trajectory. The electron track seeds can be either "ECAL-driven" or "tracker-driven". The "ECAL-driven" approach uses energetic EC to infer the position of expected electron hits in the innermost tracker layer and uses them as seeds. The "tracker-driven" approach uses KF tracks which are compatible with an EC to obtain electron seeds. The ECAL-driven approach performs better for isolated electron with high $p_T$, while the tracker-driven approach recovers the tracking efficiency of low $p_T$ or non-isolated electrons. The GSF tracking algorithm is run on all ECAL- and tracker-driven seeds to reconstruct electron tracks.

In a given PF block, an electron candidate is seeded from a GSF track if the corresponding

ECAL cluster is linked to no more than two additional tracks. Electron candidates must satisfy additional identification criteria based on seven identification variables, defined in Ref. [83]. This analysis uses the "veto" and "tight" cut-based identification working points for electrons with approximately 95% and 70% efficiency [83], respectively.

A photon candidate is seeded from an ECAL supercluster with $E_T$ larger than 10 GeV, with no link to a GSF track. It is also required to be isolated from other calorimeter clusters in the event and that the ratio of HCAL and ECAL energy clusters are compatible with a photon shower. Photons are not used in the presented analysis.

All tracks and clusters in the PF block used to reconstruct electrons and photons are removed from further processing.

### 3.2.3 Hadrons and nonisolated photons

The last particles to be identified in the processing of the PF blocks are the charged hadrons ($K^{\pm}$,$\pi^{\pm}$ and protons), neutral hadrons (e.g. $K^0_L$ or neutrons) and nonisolated photons (e.g. coming from $\pi^0$ decays)

Non-isolated photons and neutral hadrons arise from ECAL and HCAL clusters not linked to any track. Within the tracker acceptance ($|\eta| < 2.5$), all such ECAL clusters are considered photons and HCAL clusters neutral hadrons. This is due to the observation that in hadronic jets, 25% of the energy is carried by the photons, while neutral hadrons deposit only 3% of the energy in the ECAL. Outside of the tracker acceptance, charged and neutral hadrons cannot be distinguished and together they leave 25% of the jet energy in the ECAL. For this reason, remaining ECAL clusters outside the tracker acceptance are assigned to photons only if there is no link between it and an HCAL cluster. If there is such a link, it is assumed that both energy clusters arise from the same (charged or neutral) hadron shower.

Charged hadrons are formed from remaining HCAL cluster that can be linked to tracks which may in turn be linked to the remaining ECAL clusters.

The final state of this analysis contains four b quarks. Through the process of hadronization, their signature in the event will be a localized production of hadrons and other

particles which we call a jet. The following section describes the methods for clustering and calibration of jets.

## 3.3 Jet clustering and calibration

Jets are collimated sprays of energetic hadrons and other particles originating from the fragmentation and hadronization of quarks or gluons. The definition of jets depends on the algorithm used to cluster particles into jets. An important desired property of the jet clustering algorithms is the infrared and collinear (IRC) safety. This means that the set of jets obtained with the use of a particular algorithm should be insensitive to low energy particles and small-angle splittings. A clustering algorithm is infrared unsafe if the set of jets changes with the addition of a low energy parton into the system. Similarly, if the splitting of a higher energy particle into two particles separated by a small-angle yields a different set of jets, the algorithm is not collinearly safe.

Cone-type algorithms, which sum the momenta of all particles within a cone of fixed size in the $\eta - \phi$ space, are all infrared unsafe, except SISCone [84]. CMS therefore uses a family of sequential recombination algorithms. This family of algorithms assumes that the particles inside jets have small differences in transverse momenta so the grouping of particles is based, in addition to their position, also on momentum space. A consequence of this may be that these jets can have varying sizes, as will be seen later. The workflow of these algorithms starts by defining an inter-particle distance, $d_{ij}$, for each particle pair in the given set and also by defining a distance $d_{iB}$ between each particle and the beam. These are defined as:

$$d_{ij} = \min(p_{T,i}^{2k}, p_{T,j}^{2k})\frac{\Delta R^2}{R},\ d_{iB} = p_{T,i}^{2k}, \tag{3.2}$$

where $p_{T,i}$ is the transverse momentum of the i-th particle and $\Delta R$ is the distance in the rapidity-azimuthal angle space, $\Delta R^2 = (\Delta\phi)^2 + (\Delta y)^2$. There are also two tunable, dimensionless parameters, $k$ and $R$, which will be explained later. The algorithm finds the smallest distance among all $d_{ij}$ and $d_{iB}$. If the smallest distance is $d_{iB}$, particle $i$ is removed from the set of particles and added to the set of jets. Otherwise, if $d_{ij}$ is the

smallest distance, particles $i$ and $j$ are then merged into a new particle and removed from the list.

The above procedure is repeated until there are no more particles left. We can immediately understand why this family of algorithms is IRC safe. In the case of collinear splitting ($\Delta R \to 0$) will give $d_{ij} \to 0$ so the two particles will be merged. Infrared emissions will also have no impact on the properties of the jet since their momenta is by definition small.

The size of the jets is determined by parameter $R$. When the distance between particles becomes such that $\Delta R_{ij} > R$, the beam distance becomes smaller than the distance between particles so particles are no longer recombined. For that reason, $R$ is called the jet radius. If the jets are constructed with a small radius, the jet size may be too small to capture all hadronized particles. On the other hand, larger jets are more prone to capturing particles not originating from the target parton.

The parameter $k$ controls the order in which the particles are merged into jets. The originally proposed sequential recombination algorithm, the "$k_T$" algorithm used $k = 1$. This choice merges particles with lower transverse momenta first. Another notable choice is $k = 0$, called the Cambridge-Aachen (CA) [85] algorithm, which removes the ordering dependence on the momenta and clusters particles solely by their angular distance. Finally, the anti$-k_T$ algorithm (AK) [86] uses $k = -1$ and combines the particles with highest $p_T$ first. The comparison in jet shapes between different algorithms is shown in Fig. 3.4.

This analysis uses jets reconstructed with the anti$-k_T$ algorithm from the PF candidates. The two typical choices for the radius are $R = 0.4$ (AK4 jets) and $R = 0.8$ (AK8 jets), giving slimmer and wider jets, respectively. We are mainly interested in AK8 jets due to the event topologies of the process under consideration. The hadronic two-body decay of a heavy resonance at rest produces two back-to-back jets which can be well reconstructed using AK4 jets. However, if the heavy resonance has significant transverse momentum in the laboratory frame, the separation $\Delta R$ between the daughters is decreased. The angular separation can be roughly estimated to be $\Delta R \approx \frac{2M}{p_T}$. When the momentum of the resonance is sufficiently high, we enter the so-called boosted regime where the jets originating from the two products of the decay start overlapping. The efficiency of reconstructing decay products for such jets may be reduced for AK4 jets, while wider jets will be able to contain the constituents of both jets. Applying this to the process of

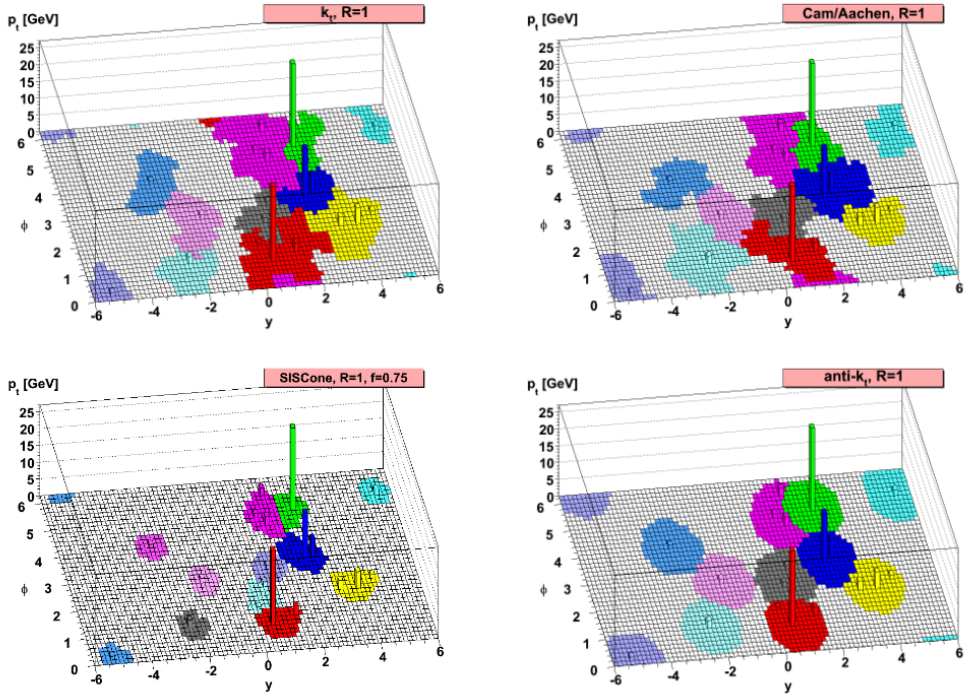



Figure 3.4: An example of jet clustering with four different algorithms applied to the same event. Shown are the $k_T$ (upper left), CA (upper right), SISCone (lower left) and anti$-k_T$ (lower right) algorithms. The colored regions correspond to different reconstructed jets. Figure taken from [86]

interest for this search, the X $\rightarrow$ YH $\rightarrow$ b$\overline{\text{b}}$b$\overline{\text{b}}$ decay will yield boosted events if the mass difference between X and the two lighter scalars is sufficiently high such that X imparts large enough momentum to the two scalars. This analysis is indeed primarily focused on such mass hypotheses ($M_{\text{X}}$, $M_{\text{Y}}$). Therefore, AK8 jets are used in the search for the signal. However, AK4 jets (in addition to AK8 jets) are used to trigger on data events. AK4 jets are also used in a semileptonic control region, described in Sec. 4.3.3, helping to reconstruct leptonic decays of the t quark.

### 3.3.1 Jet pileup removal

Particles originating from the additional proton-proton interactions can end up in the vicinity of particles from the hard-scattering which may contaminate the reconstructed jets. These unwanted contributions can be removed by actively working to identify and remove particles contaminating each jet. To mitigate this effect, the "Charged Hadron Subtraction" (CHS) technique is used. Charged particles identified to be originating from

pileup vertices are discarded and an offset correction is applied to correct for remaining contributions from neutral particles [87].

Another pileup removal algorithm is the Pileup Per Particle Identification (PUPPI) algorithm [88]. Similar to CHS, it discards charged particles identified to originate from pileup vertices. However, the PUPPI algorithm was developed to subtract pileup contribution of neutral particles from the jets on a per-particle basis. For each particle (both charged and neutral), PUPPI calculates a quantity

$$\alpha_i = \log \sum_j \xi_{ij} \times \Theta(R_{min} < \Delta R_{ij} < R_0), \tag{3.3}$$

where $\xi_{ij} = \frac{p_{Tj}}{\Delta R_{ij}}$ is the ratio of the transverse momentum of $j-$th particle in GeV and the angular distance between particles $i$ and $j$. The Heaviside step function ($\Theta$) ensures that only pairs of particles for which the separation is between $R_{min}$ and $R_0$ are taken into account. The two parameters are usually set to $R_{min} = 0.02$ and $R_{max} = 0.3$. It is clear that, generally, particles with other particles in their vicinity will have a higher $\alpha$. Since hadronization creates multiple particles within a small angular area, it is expected that the particles of a hadronic jet will have other particles around them. Furthermore, particles coming from the primary vertex are more energetic, compared to particles from other interactions (pileup), and will therefore have a higher transverse momentum. Therefore, pileup particles are expected to have lower values of $\alpha$ than particles associated with the primary vertex. The summation over $j$ can be made for two sets of particles: a set of charged particles from the primary vertex and a set of all particles in the event. The former is available only within the tracker acceptance, which is the central part of the detector, and is denoted as $\alpha^C$. In the forward part of the detector there is no information about charged particle tracking so the summation is done over the set of all particles and denoted as $\alpha^F$. Using only charged particles from the central part of the detector, the following distributions are made: $\alpha^C_{LV}, \alpha^C_{PU}, \alpha^F_{LV}$ and $\alpha^F_{PU}$, where $LV$ denotes $\alpha$ distributions for particles originating from the primary (leading) vertex, and $PU$ distributions for particles coming from the pileup. This distinction can be made since we plotted the $\alpha$ distributions for particles which left hits in the tracker. The distributions for neutral particles are assumed to be the same as for charged particles. The justification

for such an assumption is based on simulation and can be seen in Fig. 3.5. The median ($\bar{\alpha}$) and RMS ($\sigma$) of $\alpha_{PU}^{C}$ and $\alpha_{PU}^{F}$ are used to assign weights to the particles. In order to define weights, the following quantity is first introduced:

$$\chi_i^2 = \Theta(\alpha_i - \bar{\alpha}) \times \frac{(\alpha_i - \bar{\alpha}_{PU})^2}{\sigma_{PU}^2}. \tag{3.4}$$

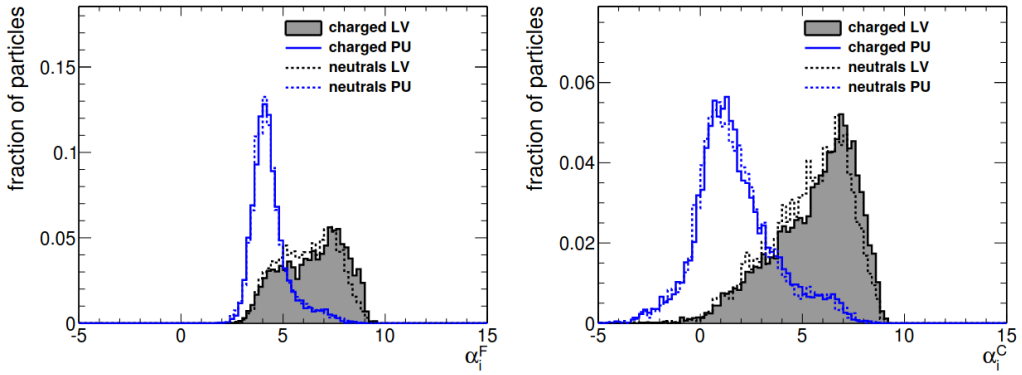



Figure 3.5: The distribution of $\alpha_i$ for particles from the leading vertex (gray filled) and particles from pileup (blue) in a dijet sample. The left figure shows $\alpha_i^F$ where the sum is done over all particles, and the right shows $\alpha_i^C$ where only the charged particles are summed over. Dotted and solid lines show distributions for neutral and charged particles, respectively. Figure taken from [88]

Equation 3.4 uses the $\alpha^C$ variant in the tracker acceptance area, while the $\alpha^F$ variant is used outside of this area. The defined quantity measures how far $\alpha_i$ fluctuates from the pileup median. Anything below the median is considered pileup and is assigned zero weight. Large fluctuations above median will receive a weight close to one, while intermediate fluctuations above the median get assigned fractional weights between zero and one. The distribution of $\alpha$ for pileup is roughly Gaussian-like, which is why the $\chi^2$ distribution for PU particles resembles a $\chi^2_{NDF=1}$ distribution (suggested by the notation). Finally, particles are assigned a weight:

$$w_i = F_{\chi^2_{NDF=1}}(\chi_i^2), \tag{3.5}$$

where $F_{\chi^2}$ is the cumulative distribution function of the $\chi^2$ distribution. This weight is used to rescale the particles' four-momenta prior to the application of jet clustering algorithms.

### 3.3.2 Jet energy corrections

The measured energy of the jets needs to be calibrated in order to match the energies of their parent particles. Jet energy corrections (JEC) [87] account for several effects and are applied in several stages.

The first layer of jet energy corrections accounts for pileup by subtracting the average transverse momentum contribution of the pileup interactions to the jet's cone area. The correction is derived from simulation by looking at events with and without pileup overlay. Corrections for residual differences between data and detector simulation are determined using events collected with a random (zero-bias) trigger. These events are not triggered by any hard-interaction so their main sources of energy deposits are pileup and detector noise.

Pileup corrections are followed by the "truth-matching" corrections, obtained using simulation. The reconstructed PF jets are compared with the jets assembled using the collection of generator-level particles after the parton shower simulation step, i.e. before detector simulation, as explained in Sec. 3.1. The latter jets are called "truth" jets. The correction on the energy scale is expressed as a function of $\eta$ and $p_T$ and brings the energy of the PF jets closer to the energy of the truth jets [87].

The final set of corrections, the residual corrections, are meant to account for remaining small differences in jet response between data and simulation, and are applied to data only. The corrections are determined using $X$ + jet events, where X is either a $Z$ boson (reconstructed from two leptons), a photon, or another jet. They are derived using momentum balancing between a jet and and a well-measured reference object X [87].

### 3.3.3 Jet grooming

Besides pileup contributions, hadronic jets may be contaminated by other contributions from the underlying event. Large area jets (AK8) are more susceptible to this contamination, which complicates the jet finding procedure and worsens the resolution in jet observables such as the jet mass. Jet grooming techniques are used to remove other unwanted jet contributions and are used in combination with the pileup removal. Grooming

is a "post-processing" treatment of large radius jets where unwanted particles are removed from the jet, based on the assumption that there is usually only one hard scattering per event. In other words, the sources of contamination are usually much softer. The two main grooming algorithms are jet trimming and soft drop.

For a given jet, the jet trimming algorithm reclusters the constituents into subjets using the anti $-k_T$ algorithm with a different radius ($R_{sub} < R$). The contribution of a given subjet $i$ is removed if $p_{T,i} < f_{cut} \cdot \Lambda_{hard}$, where $f_{cut}$ is a fixed dimensionless parameter and $\Lambda_{hard}$ is a scale chosen depending upon the kinematics of the event. The remaining subjets are finally assembled into the trimmed jet. At CMS, the trimming algorithm is used with $R_{sub} = 0.01$, $f_{cut} = 0.03$ and $\Lambda_{hard}$ being the $p_T$ of the non-groomed jet.

For the soft drop grooming, the CA algorithm is used to re-cluster the constituents of the jet on which we want to perform grooming. However, the last step of the clustering sequence, in which the two last subjets are merged, is not performed. The transverse momenta of the two subjets are used to evaluate the grooming condition. The softer (lower $p_T$) subjet is removed unless the following condition is satisfied:

$$\text{Soft drop condition} : \frac{min(p_{T1}, p_{T2})}{p_{T1} + p_{T2}} > z_{cut} \left( \frac{\Delta R_{12}}{R_0} \right)^{\beta}, \tag{3.6}$$

where $p_{T,i}$ is the transverse momentum of the $i$-th subjet, $\Delta R_{1,2}$ is the angular distance between the two subjets and $R$ is the size of the original jet. There are also two tunable parameters, $z_{cut}$ and $\beta$. The former determines the minimum $p_T$ fraction that the jet constituents need to have in order not to be removed, while the latter scales the $p_T$ fraction threshold as a function of the distance between the jet constituents. At CMS, the algorithm is used with radiation fraction parameter $z_{cut}$=0.1 and $\beta$=0. The process of removing subjets in softdrop grooming is repeated until the soft drop condition is satisfied. In this analysis, the mass of the AK8 jets is calculated using the soft drop algorithm, in conjunction with the PUPPI pileup removal technique.

## 3.4 Jet flavor tagging

As previously stated, this analysis is interested in final states with four b quarks. Moreover, since we are working in the boosted regime, the $b\bar{b}$ pairs are highly collimated and are to be reconstructed as two AK8 jets. Therefore, the events of interest are those with two AK8 jets, each coming from the hadronization of a $b\bar{b}$ pair. Furthermore, the analysis uses a semileptonic control region to constrain the $t\bar{t}$ background. For the purpose of tagging leptonically decaying t quarks, we will be looking for an AK4 jet originating from the hadronization of a b quark.

Generally, it is not possible to unambiguously determine the flavor of parton(s) which lead to the hadronic jet. However, there are some properties of b-jets which can be used to separate them from other jets. Perhaps the most important one is that the hadrons originating from b quark hadronization (b-hadrons) have a relatively long lifetime, $\tau \approx 1.5 \cdot 10^{-12}$ s. This makes it possible for b-hadrons to travel several millimeters in the detector before decaying, which is seen as a presence of a secondary vertex and tracks displaced from the primary vertex, as illustrated in Fig. 3.6.

Besides the long lifetime, b-hadrons have other distinguishing features like the relatively high mass ($\approx$ 5 GeV), high track multiplicity, and large semileptonic decay branching ratios. Using these features, b-tagging algorithms were developed with the goal of separating b-jets from other jets. There are several b-tagging algorithms used within the CMS. One of them is the "DeepJet" tagger [90]. The strategy of DeepJet consists in using low-level features from as many jet constituents as possible as opposed to selecting a few that are well identified, which was the approach of earlier taggers. There are approximately 650 input variables used. They can be divided into four categories: global jet variables, charged PF candidates features, neutral PF candidate features, and SV features associated with the jet. The full list of variables can be found in Ref. [90].

To process the large number of features, a custom neural network architecture was developed. This resulted in a significant gain in performance when compared to previous taggers. Fig. 3.7 shows the comparison in performance with the "DeepCSV" tagger [91]. In this analysis, we consider an AK4 jet b-tagged if it passes the medium working point, at which the mistag rate for light-flavored jets, originating from u, d or s quarks, is 1%

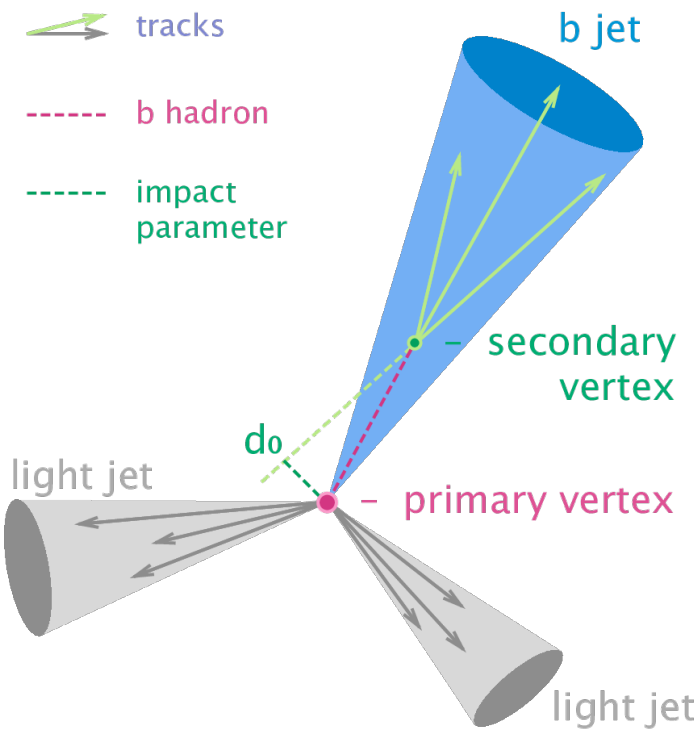



Figure 3.6: An illustration of a b jet (blue). The figure depicts the creation of a secondary vertex originating from the decay of a b-hadron whose displacement from the primary vertex is large enough to be measured. This also leads to tracks being displaced from the primary vertex, shown as the increased value of the impact parameter, $d_0$, for one of the particles. Figure taken from [89]

and the efficiency for b jets is 68% [91].

A different algorithm, called ParticleNet, is used in order to tag AK8 jets coming from the decay of a massive particle into two b quarks. The algorithm is further expanded upon in Ch. 4, Sec. 4.3.

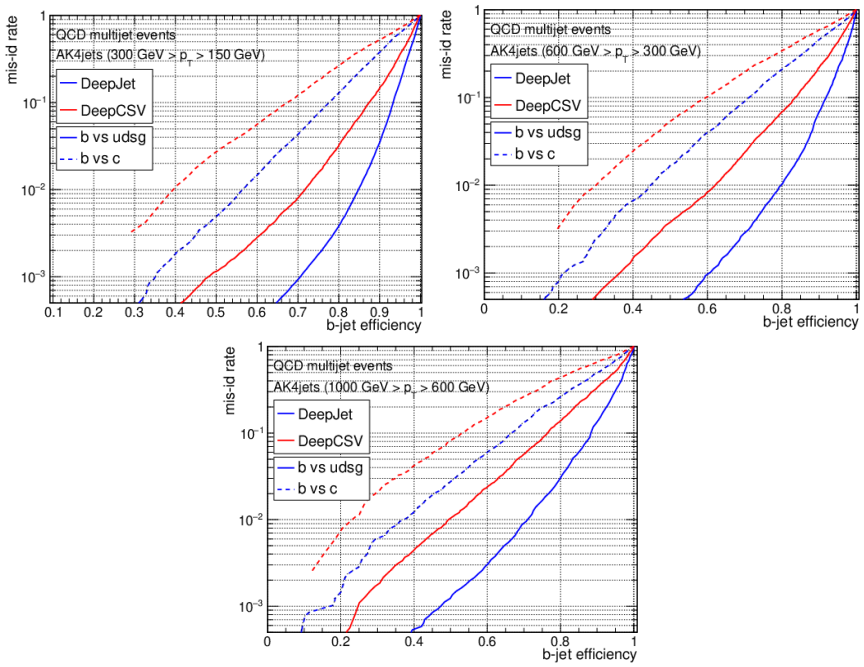



Figure 3.7: The performance of DeepJet and DeepCSV in three different jet $p_T$ ranges in QCD multijet events. The performance is shown for both b vs. c (dashed lines), and b vs. light (solid lines). Figures taken from [90]

# Chapter 4

# Data analysis

In this chapter, the search for the $X \to YH \to b\bar{b}b\bar{b}$ decay in the Lorentz-boosted topology is presented. The general search strategy is outlined in Sec. 4.1. The event simulation and event selections are described in Sec. 4.2 and Sec. 4.3, respectively. The methods used to estimate the background and the systematic uncertainties on the estimated background and signal are given in Sec. 4.4 and Sec. 4.5. Finally, the tests used to validate the background estimation model, before unblinding the data in the search (signal) regions, are shown in Sec. 4.6. The search results in the signal regions, after unblinding the data, are presented in Ch. 5.

## 4.1 Search strategy

As stated in Ch. 1, this analysis aims to explore the $X \to YH \to b\bar{b}b\bar{b}$ process using pp collision data at the LHC collected by the CMS experiment. This search focuses on the kinematic regions where $M_X$ is sufficiently larger than $M_Y$ and $M_H$ such that both Y and H are imparted with considerable momenta and therefore their decay products, i.e. the $b\bar{b}$ pairs, are highly collimated. The mass ranges $0.9 < M_X < 4$ TeV and $60 < M_Y < 600$ GeV are explored in this analysis, complementing the resolved $X \to YH \to b\bar{b}\tau\tau$ search [28].

The final state consists of four b quarks grouped into two pairs coming from the decays of the H and Y bosons. Each pair is reconstructed as a single large-area jet (AK8 jet). The signature of a signal event is the presence of two AK8 jets with high $p_T$ in the central

part of the detector. However, events with this topology can also arise from various SM processes with much higher cross sections. These processes are collectively referred to as background. The main background processes for this analysis are QCD multijet events and the pair production of top quarks, $\mathrm{t\bar{t}}$. The known mass of the H is exploited to reduce the amount of background by requiring that one of the two jets is consistent with $M_H = 125$ GeV. This jet is called the H-candidate, while the other AK8 jet is then called the Y-candidate. Finally, jet tagging techniques are employed to identify jets consistent with the decay of a massive resonance into a pair of b quarks, further improving the signal-to-background ratio and increasing the sensitivity of the search.

The search for the presence of signal is performed in a 2-dimensional plane, defined by the mass of the Y-candidate jet, $M_{\mathrm{J}}^{\mathrm{Y}}$, and the invariant mass of the two candidate jets, $M_{\mathrm{JJ}}$. For signal events, the two search variables correspond to the masses of Y and X, respectively. Therefore, the presence of a signal in the search plane is manifested as an excess of data, localized around $M_{\mathrm{Y}}$ and $M_{\mathrm{X}}$.

The multijet background in the search region is estimated using data-driven techniques, which provide better modelling than the simulation. The estimate is based on defining a non-search region which is enriched in multijet events. The shape and yield of the multijet background in this region can be measured in data. Transfer functions are used to estimate the multijet shape and yield in the search regions from the data in the non-search regions.

The $\mathrm{t\bar{t}}$ background is estimated using simulation. Corrections on the simulation yields are measured using a dedicated control region. This control region is designed to select semileptonic $\mathrm{t\bar{t}}$ decays and is referred to as the semileptonic control region. The data-to-simulation corrections are based on measuring the mass distribution of the hadronically decaying t quark, reconstructed as an AK8 jet, and comparing it with simulation.

Other sources of backgrounds like the single top quark production, Higgs boson in association with a top quark pair or a W or Z boson were found to have negligible contributions and are not further considered.

The analysis is performed using proton-proton collision data collected with the CMS detector at $\sqrt{s} = 13$ TeV during the RunII data-taking period, from 2016 to 2018. The integrated luminosities per year is given in Table 4.1. The data is grouped into pri-

mary datasets based on the triggers fired by the events. The analysis uses three primary datasets. The JetHT primary dataset is based on high jet activity in the event and is used to select signal-like events. The SingleMuon and SingleElectron datasets are based on the presence of a muon or an electron in the event and are used in the semileptonic control region selection.

Table 4.1: Integrated luminosities per data-taking year.

| Year | $\mathcal{L}(\text{fb}^{-1})$ |
|---|---|
| 2016 | 36.3 |
| 2017 | 41.5 |
| 2018 | 59.8 |
| Total RunII | 137.6 |

## 4.2 Simulated datasets

To estimate the behavior of the signal and background processes and develop the analysis tools, Monte Carlo simulations are used. The generation of simulated events includes the calculation of the matrix-element, the parton showering and hadronization, and the propagation of particles through the detector, as described in Ch. 3.

The signal process, $X \to YH \to b\bar{b}b\bar{b}$, is simulated with a mass width of 1 MeV for all the three scalars at the leading order (LO) precision using the MadGraph [68] event generator, version 2.6.5. The NMSSM model [92, 93] is used to produce the simulated samples. The kinematic parameters of the model are generic enough to allow the interpretation of the results under other BSM scenarios. The signal datasets are generated for 260 different ($M_X$,$M_Y$) hypotheses, as shown in Fig. 4.1.

The $t\bar{t}$+jets events with hadronic top quark decays are modelled using the POWHEG 2.0 [69, 70, 71] generator, at next-to-leading order (NLO). A sample of semileptonic $t\bar{t}$ decays, with one of the top quarks decaying via $t \to Wb \to \ell\nu b$, $\ell$ being a lepton (electron or muon), is simulated using the same generator. The latter sample is used in the semileptonic control region to derive data-to-simulation correction factors. The simulated $t\bar{t}$+jets event yields are scaled using a cross section of $832^{+46}_{-52}$ pb, calculated at next-to-next-to-leading order in QCD [94].

Multijet samples, containing two to four jets, are simulated at LO using the MadGraph event generator. However, these samples are not used in estimating the multijet background for which data-driven methods are used. They are instead used to develop the tools for the analysis.

The showering and hadronization of partons are simulated with pythia8.2 [76]. The matrix element-parton matching for the signal and the multijets background uses the MLM [73] scheme. The parameters for the underlying event description in pythia are given by the CP5 tune [78] for all samples, except for the $t\bar{t}$ and multijet samples in 2016 which use the CUETP8M2T4 [95] and CUETP8M1 [77] tunes, respectively.

The PDFs from the NNPDF3.0 [96] NLO and NNPDF3.1 [67] NNLO sets are used. They are included in the lhapdf6 PDF library [97].

All generated events are processed through a Geant4 [79] simulation of the CMS detector. The effects of pileup are modelled assuming a total inelastic pp cross section of 69.2 mb [98]. Simulated event samples are weighted to match the distribution of the expected pileup profile of data.

## 4.3 Event selection

The analysis uses two main categories of events:

1. Hadronic category: events with only jets. This event category is subdivided into various regions based on further event selection.

2. Semileptonic category: events with one isolated lepton (an electron or a muon). These events are used to measure the data-to-simulation corrections for the $t\bar{t}$ background. This category is also subdivided into various regions based on jet tagging requirements.

These two categories are mutually exclusive based on whether there is an isolated lepton present in the event or not.

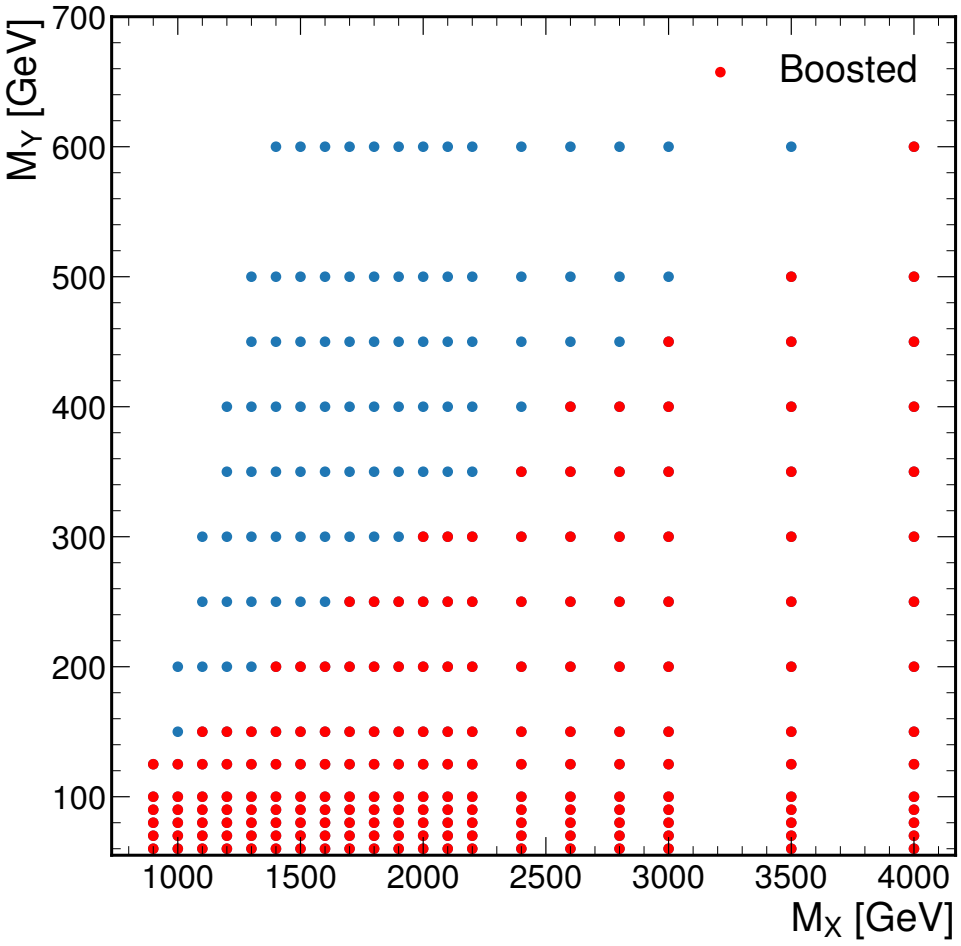



Figure 4.1: Grid of considered signal hypotheses. Sets of $M_{\mathrm{X}}$, $M_{\mathrm{Y}}$ which satisfy the boosted condition for both the scalars, $\frac{M_{\mathrm{X}}-M_{\mathrm{Y}}-M_{\mathrm{H}}}{2} > \frac{2\,M_{heavier}}{0.8}$, where $M_{heavier}$ is the mass of the heavier scalar between H and Y, are marked in red.

### 4.3.1 Online selection

The CMS triggering system, responsible for selecting data to be stored, is described in Chapter 2. A set of triggers based on requirements on jet or lepton $p_T$ are used for online event selection. The trigger paths, at the HLT level, used to select data for this analysis are listed in Table 4.2. The logical OR of HLT decisions for each channel of the two categories in a year are used to select collision data events. The description of the triggers listed in Table 4.2, with their names in parentheses, is given below, starting with the triggers used in the hadronic events category:

One trigger criterion required a single AK8 jet with a $p_T$ >450 and >500 GeV in 2016

and in 2017–2018, respectively (AK8PFJet). A second trigger required that the scalar sum $p_T$ of all AK4 jets with $p_T$ >30 GeV and $|\eta|$ <2.5 ($H_T$) be >800 or 900 GeV in 2016, depending on the LHC beam luminosity. In 2017 and 2018, $H_T$ >1050 GeV was required (PFHT).

The third trigger algorithm required an AK8 jet with a trimmed mass >30 GeV and with $p_T$ >360 GeV in 2016 (AK8Jet_TrimMass), while in 2017–2018 the AK8 jet $p_T$ threshold was raised to 400 and 420 GeV, depending on the LHC beam instantaneous luminosity, keeping the trimmed mass criterion the same. The fourth trigger required $H_T > 650$ or 700 GeV (in 2016) and $> 800$ GeV (in 2017–2018), together with an AK8 jet having trimmed mass $> 50$ GeV (AK8PFHT_Trim).

The following three algorithms were only used in 2016: (1) two AK8 jets with $p_T$ >280 and >200 GeV and with one of them having a trimmed mass >30 GeV (AK8DiPFJet_TrimMass); (2) had the same requirements as (1) with one of the AK8 jets passing a loose b-tagging criterion using the "combined secondary vertex" algorithm [91] with efficiency $\epsilon = 81\%$ (AK8DiPFJet_TrimMass_BTagCSV); (3) $H_T$ ≥650 GeV with a pair of AK4 jets having an invariant mass above 900 GeV and a pseudorapidity separation $|\Delta\eta| < 1.5$ (PFHT_WideJetMJJDEtaJJ).

The triggers in the semileptonic category required events to have either an isolated muon of $p_T > 24$ or 27 GeV (IsoMu or IsoTkMu); or an isolated electron having $p_T > 27$, 32, or 35 GeV (Ele_WPTight_Gsf); or a photon with $p_T > 175$ or 200 GeV (Photon). The photon trigger path is added due to the reduced online electron identification efficiency which may result in electrons being identified as photons at the trigger level. The thresholds change between data-taking years.

The HLT requirements are not applied to simulated events, which are instead reweighted by the measured trigger efficiency in the data.

The hadronic category trigger efficiency is measured in the data using the baseline trigger HLT_PFJet260. This trigger is prescaled over much of the run period, meaning that only every n-th event satisfying the condition raises the trigger flag, where n is the prescaling factor. The low threshold of this trigger necessitates the prescaling, as otherwise, the trigger rate would be too high for recording. Even though it is prescaled, this trigger path

Table 4.2: The HLT paths used for each year for the hadronic and the semileptonic event categories. The triggers are used as a logical OR of all the rows for each year in the hadronic selection. In the semileptonic selection, IsoMu (and IsoTkMu in 2016) are used in the muon channel, while the Ele_WPTight and Photon triggers are used in the electron channel.

| Hadronic | $\mathcal{L}_{eff}(\text{fb}^{-1})$ | Semileptonic | $\mathcal{L}_{eff}(\text{fb}^{-1})$ |
|---|---|---|---|
| 2016 | | | |
| PFHT800 | 27.71 | IsoMu24 | 36.47 |
| PFHT900 | 36.47 | IsoTkMu24 | 36.47 |
| PFHT650_WideJetMJJ900DEtaJJ1p5 | 36.47 | Ele27_WPTight_Gsf | 36.47 |
| AK8PFHT650_TrimR0p1PT0p03Mass50 | 20.20 | Photon175 | 36.47 |
| AK8PFHT700_TrimR0p1PT0p03Mass50 | 36.47 | | |
| AK8PFJet450 | 33.64 | | |
| AK8PFJet360_TrimMass30 | 36.47 | | |
| AK8DiPFJet280_200_TrimMass30 | 36.47 | | |
| AK8DiPFJet280_200_TrimMass30_BTagCSV_p20 | 36.47 | | |
| 2017 | | | |
| PFHT1050 | 41.54 | IsoMu27 | 41.54 |
| AK8PFHT800_TrimMass50 | 36.75 | Ele35_WPTight_Gsf | 41.54 |
| PFJet500 | 41.54 | Photon200 | 41.54 |
| AK8PFJet500 | 41.54 | | |
| AK8PFJet400_TrimMass30 | 36.75 | | |
| AK8PFJet420_TrimMass30 | 36.75 | | |
| 2018 | | | |
| PFHT1050 | 59.96 | IsoMu24 | 59.96 |
| AK8PFHT800_TrimMass50 | 59.96 | Ele32_WPTight_Gsf | 59.96 |
| PFJet500 | 59.96 | Photon200 | 59.96 |
| AK8PFJet500 | 59.96 | | |
| AK8PFJet400_TrimMass30 | 59.96 | | |
| AK8PFJet420_TrimMass30 | 59.96 | | |

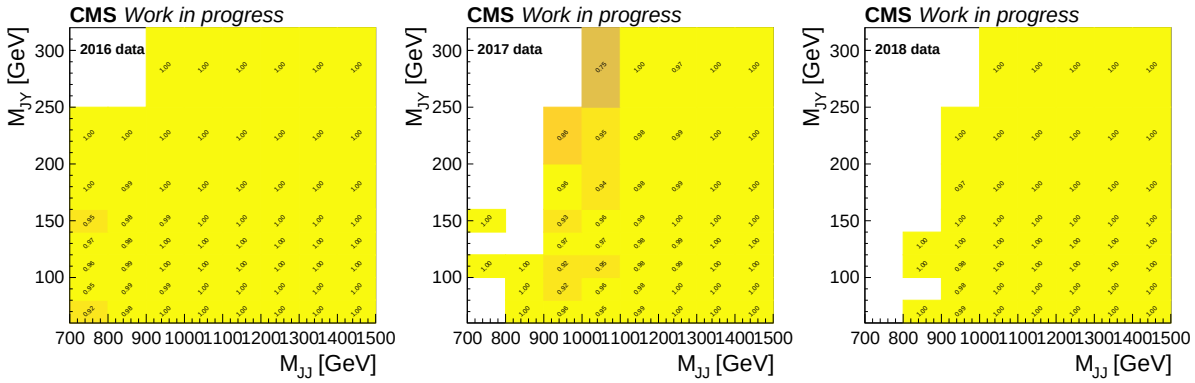



Figure 4.2: The trigger efficiency in the hadronic category as a 2D function of dijet invariant mass $M_{\text{JJ}}$ and Y-candidate soft-drop mass $M_{\text{J}}^{\text{Y}}$ in the 2016 (left), 2017 (middle) and 2018 (right) data.

provides enough events for the measurement of efficiencies. Events passing the baseline trigger are further required to pass selection criteria close to the signal selection in the actual analysis. The criteria are based on the properties of the two leading AK8 jets:

- Leading two (in $p_T$) AK8 jets in the event with $p_T > 350(450)$GeV and $|\eta| < 2.4(2.5)$ for 2016 (2017, 2018)

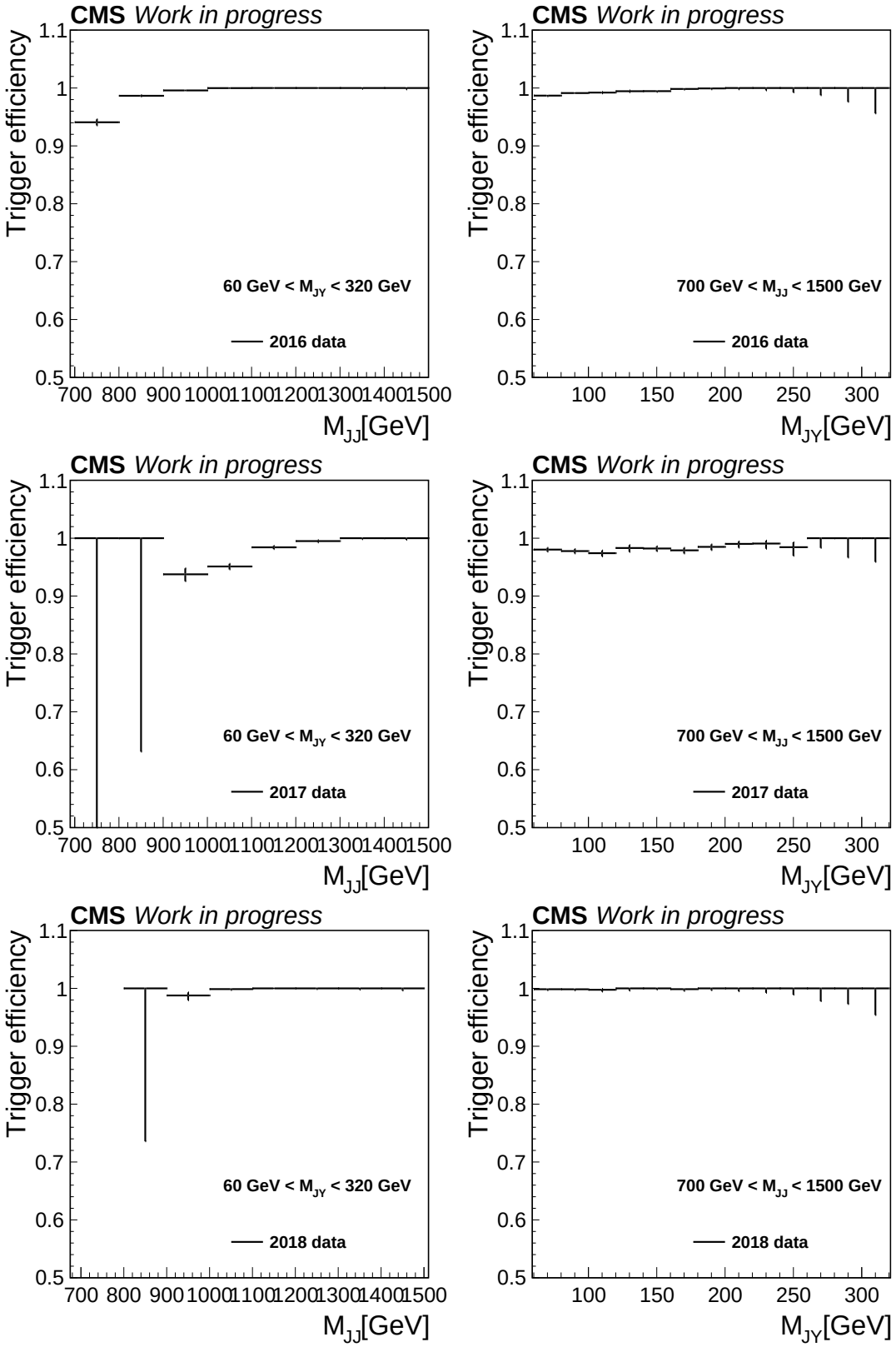



Figure 4.3: The trigger efficiency in the hadronic category as a function of dijet invariant mass $M_{\mathrm{JJ}}$ and Y-candidate soft-drop mass $M_{\mathrm{J}}^{\mathrm{Y}}$ in the 2016 (upper), 2017 (middle) and 2018 (lower) data.

- The softdrop mass ($M_{SD}$) of at least one of the two jets is $110 < M_{SD} < 140$ GeV and the mass of the other $M_{SD} > 60$ GeV, with all necessary jet mass corrections applied

- Difference in $\eta$, $|\Delta\eta_{\mathrm{JJ}}| < 1.3$, for the two leading AK8 jets

- Dijet invariant mass of the two leading AK8 jets $M_{\mathrm{JJ}} > 700$ GeV.

The 2-dimensional measured trigger efficiency as a function of $M_{\mathrm{JJ}}$ and the softdrop mass of the Y-candidate $M_{\mathrm{J}}^{\mathrm{Y}}$ is shown in Fig. 4.2. It was found that the efficiencies are mostly flat in $M_{\mathrm{J}}^{\mathrm{Y}}$ and close to 100%. Hence simulated events were reweighted using trigger efficiencies as a function of $M_{\mathrm{JJ}}$ only. These efficiencies are shown in Fig. 4.3. The combined set of triggers reaches full efficiency at $M_{\mathrm{JJ}} > 1000$ GeV in 2016, $M_{\mathrm{JJ}} > 1300$ GeV in 2017 and $M_{\mathrm{JJ}} > 1100$ GeV in 2018. The signal search is performed for $M_{\mathrm{JJ}} > 900$ GeV. Since the trigger efficiency is measured in data that mostly consist of the two considered background processes, an additional check needs to be made to confirm whether the measured efficiencies can be applied to signal events. A trigger efficiency comparison between signal and background simulation was made for that reason. The comparison is shown in Fig. 4.4. Efficiencies of signal and background in simulation agree and are 100%. Trigger efficiencies do not change with signal mass point choice as can be seen in Fig. 4.5.

The semileptonic event trigger efficiencies were measured in a sample of $\mathrm{Z}\to\ell\ell$ events by the CMS Muon and EGamma physics object groups. The efficiencies are shown in Fig. 4.6.

### 4.3.2 Offline selection: Hadronic category

The offline selection is based on the expected topology of the signal events, i.e. two AK8 jet with high $p_T$. For the 2016 (2017, 2018) analysis, the two $p_T$ leading AK8 jets need to have $p_T > 350$ (450) GeV and $|\eta| < 2.4$ (2.5). The $p_T$ threshold is increased in 2017 and 2018 in order for the trigger selection to be (nearly) fully efficient for events passing the offline selection.

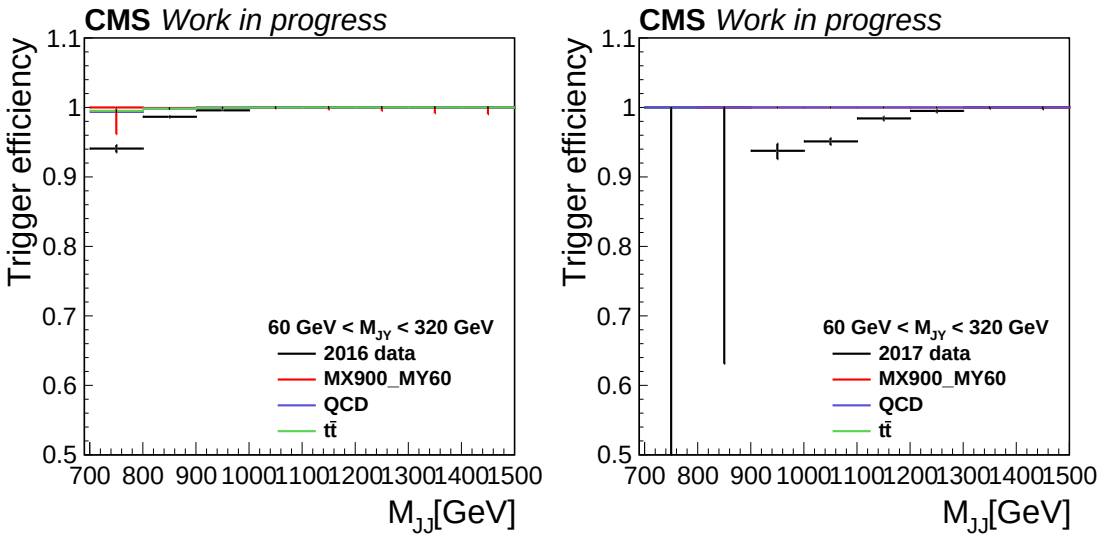



Figure 4.4: Comparison of the trigger efficiency in simulation for background and signal and trigger efficiency measured in data, as a function of dijet invariant mass $M_{\mathrm{JJ}}$, for 2016 (left) and 2017 (right).

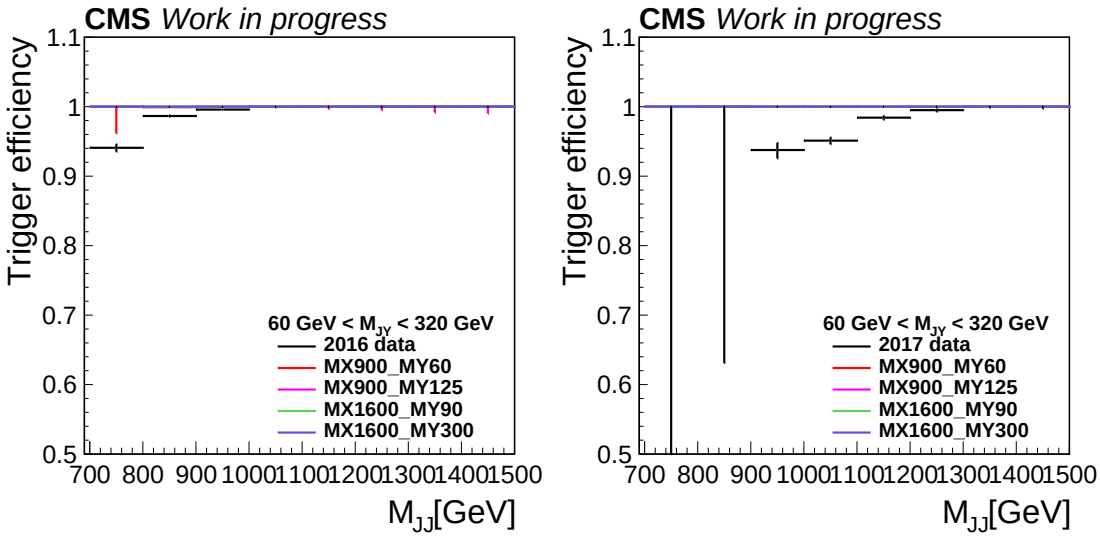



Figure 4.5: Comparison of the trigger efficiency in simulation for different signal samples and trigger efficiency measured in data, as a function of dijet invariant mass $M_{\mathrm{JJ}}$, for 2016 (left) and 2017 (right).

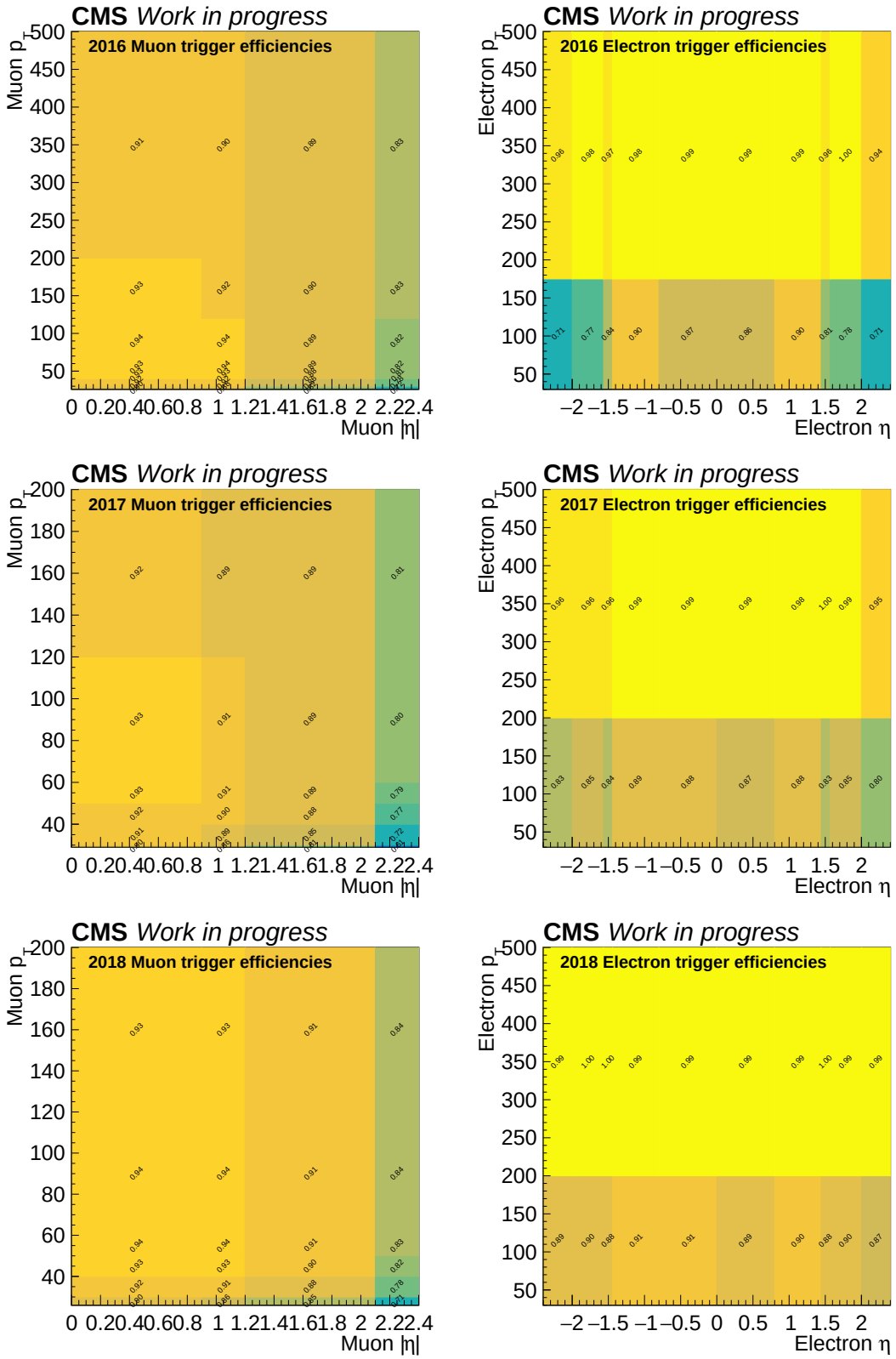



Figure 4.6: The trigger efficiency in the semileptonic channel as a function of lepton $p_T$ and $\eta$. Efficiencies are shown for muon (left) and electron (right) channels for 2016 (upper), 2017 (middle), 2018 (lower).

To reduce the contribution from multijet events, the two leading jets are required to be relatively close in $\eta$: $|\Delta\eta_{\mathrm{JJ}}| < 1.3$ [99, 100]. The AK8 jet pairs in multijet backgrounds tend to have a larger separation in pseudorapidity than the signal, for a given $M_{\mathrm{JJ}}$ range. Therefore the selection $|\Delta\eta_{\mathrm{JJ}}| < 1.3$ increases the ratio of signal to background.

Furthermore, since no leptons are present in the signal process final state, a veto on the presence of isolated leptons is imposed. The conditions for an isolated electron are: $p_T > 20$ GeV, $|\eta| < 2.4$ and the cut-based identification at the veto working point, as listed in Ch. 3, Sec. 3.2.2. A muon needs to satisfy $p_T > 20$ GeV, $|\eta| < 2.4$, loose PF identification and PF isolation at the "very loose" working point as detailed in Ch. 3, Sec. 3.2.1.

One of the two leading jets should be consistent with the mass of the Higgs boson, $110 < M_{SD} < 140$ GeV (H candidate), while the other is only required to have $M_{SD} > 60$ GeV (Y candidate), which is the lowest $M_{\mathrm{Y}}$ hypothesis considered. If both leading jets satisfy the Higgs jet mass window requirement, a Higgs jet candidate is chosen at random. Events without an H or a Y jet are rejected. The two variables used to localize the signal are the mass of the Y jet, $M_{\mathrm{J}}^{\mathrm{Y}}$, and the invariant mass of the H and Y jets, $M_{\mathrm{JJ}}$ with approximately 15% and 9% resolution, respectively.

The ParticleNet algorithm [101], based on a graph convolutional neural network and using the properties of the jet PF constituents as features, is used to identify the boosted $\mathrm{H} \to \mathrm{b\bar{b}}$ or $\mathrm{Y} \to \mathrm{b\bar{b}}$ decays against a background of other jets. The multiclassifier ParticleNet algorithm outputs several variables, each in the range 0–1, which can be interpreted as the probability of a jet to arise from a certain decay, such as from the decay of a massive resonance ($P(\mathrm{R}\to\mathrm{b\bar{b}})$) or the decay of a light-flavoured quark or gluon ($P(QCD)$). In this thesis, the discriminant $P(\mathrm{R}\to\mathrm{b\bar{b}})/(P(\mathrm{R}\to\mathrm{b\bar{b}}) + P(QCD))$ is used.

The ParticleNet algorithm is trained [102] on AK8 jets using simulated Lorentz-boosted spin-0 particles, with a wide range of masses, decaying to a pair of b quarks as the signal. QCD multijet samples are used for the background. The wide signal mass range in the training sample ensures that the background rejection rate is decorrelated from the mass of the jet [102]. This allows the definition of background enriched regions, by requiring the jets to have low ParticleNet scores, which have the same mass spectra as the background in the signal regions. This property is used in the estimation of the multijet background,

as will be described in Sec. 4.4.

The ParticleNet tagger was selected based on its performance compared to other taggers. For example, its background rejection is a factor of two larger than that of the previously used [31] mass-decorrelated DeepAK8 (DeepAK8-MD) algorithm [103], in the $p_T$ 500–1000 GeV range [102]. For this analysis, which selects two AK8 jets per event, this translates to approximately a factor of four improvement in background rejection for the same signal efficiency.

The ParticleNet scores used for selecting the $H \to b\bar{b}$ and the $Y \to b\bar{b}$ candidates ("signal jets") are required to be >0.98 (tight requirement) or >0.94 (loose requirement). Depending on the jet $p_T$, the former has an efficiency of 62–72% and a misidentification rate of 0.45%, while the latter has an efficiency of 80–85% and a misidentification rate of 1%, per jet.

Due to potential discrepancies between simulation and data, the efficiency of the ParticleNet classifier in data may differ from simulation. It therefore needs to be measured in data. The calibration was done using a sample of jets originating from gluon fragmentation to $b\bar{b}$ [104]. Since $g \to b\bar{b}$ jets differ from the jets originating from the decay of a massive particle, a special selection is applied to select "signal-like" (proxy) jets. A boosted decision tree classifier (BDT) is trained to separate jets in QCD multijet events with a clean composition of quarks from those with large contamination of extra gluons. This is schematically shown in Fig. 4.7

The value of the applied BDT cut is chosen so that the ParticleNet score distribution of the proxy jets resembles that of Y/H jets. A systematic uncertainty is assigned to account for the uncertainty in the choice of the BDT cut value.

The data-to-simulation corrections depend on the data-taking year and the jet $p_T$. They are listed, with their uncertainties, in Table 4.3. A different method of calibrating the ParticleNet efficiency, based on measuring $Z \to b\bar{b}$ decays, is presented in Ch. 6.

The ParticleNet scores of the H and Y jets are used to classify events into either signal, sideband or validation categories. A layout of the different regions is shown in Fig. 4.8, with their descriptions given in Table 4.4.

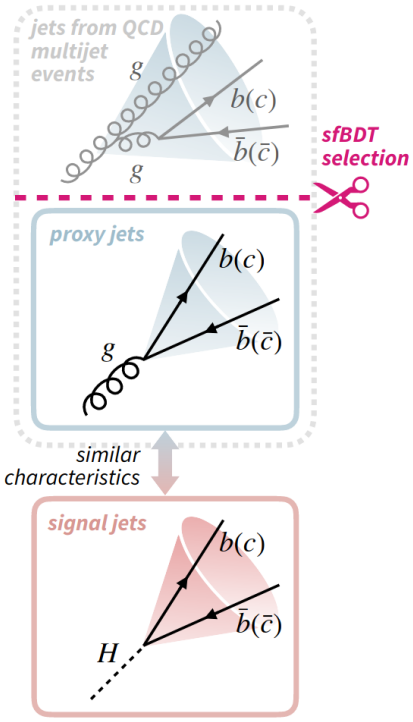



Figure 4.7: Illustration of the jet topologies selected by the BDT for the purposes of ParticleNet efficiency calibration in data. The BDT is trained to select those jets from QCD multijet events which have similar characteristics as the signal jets. Figure taken from Ref. [104]

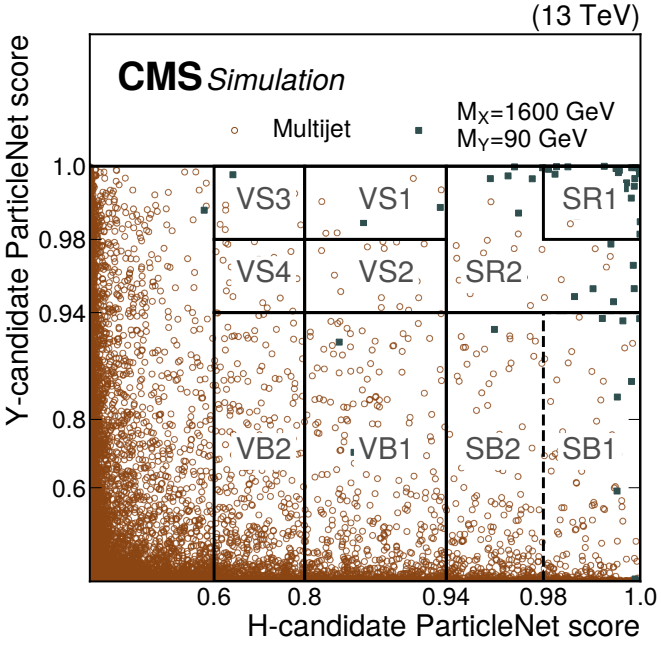



Figure 4.8: The distribution of the ParticleNet tagger scores for the Higgs and the Y jet candidates for the signal ($M_X$ =1600 and $M_Y$ =125 GeV), and QCD multijets.

Table 4.3: ParticleNet data-to-simulation efficiency corrections.

| 2016 | | | | | |
|---|---|---|---|---|---|
| WP | 300–400 GeV | 400–500 GeV | 500–600 GeV | 600-800 GeV | >800 GeV |
| 0.98 | $1.18 \pm 0.20$ | $1.34 \pm 0.37$ | $1.07 \pm 0.17$ | $1.24 \pm 0.28$ | $1.12 \pm 0.37$ |
| 0.94 | $1.15 \pm 0.21$ | $1.13 \pm 0.26$ | $1.12 \pm 0.21$ | $1.11 \pm 0.26$ | $1.18 \pm 0.44$ |

| 2017 | | | | |
|---|---|---|---|---|
| WP | 300–400 GeV | 400–500 GeV | 500–600 GeV | >600 GeV |
| 0.98 | $1.20 \pm 0.06$ | $1.15 \pm 0.09$ | $1.22 \pm 0.10$ | $1.31 \pm 0.21$ |
| 0.94 | $0.90 \pm 0.10$ | $0.85 \pm 0.07$ | $0.86 \pm 0.09$ | $0.89 \pm 0.11$ |

| 2018 | | | | |
|---|---|---|---|---|
| WP | 300–400 GeV | 400–500 GeV | 500–600 GeV | >600 GeV |
| 0.98 | $1.19 \pm 0.10$ | $1.14 \pm 0.08$ | $1.21 \pm 0.15$ | $1.36 \pm 0.24$ |
| 0.94 | $0.87 \pm 0.05$ | $0.89 \pm 0.06$ | $0.89 \pm 0.07$ | $0.95 \pm 0.14$ |

Two signal regions are defined using the loose and the tight ParticleNet scores (Fig. 4.8). The "signal region 1" (SR1) and the "signal region 2" (SR2) are statistically exclusive. The SR1 has a higher signal-to-background ratio and is thus more sensitive to the presence of signal. The purpose of SR2 is to improve the sensitivity for signal mass points with low background by increasing the total signal efficiency.

Two "sideband regions", corresponding to the two signal regions, are defined and labelled as "Sideband 1" (SB1) and "Sideband 2" (SB2) in Fig. 4.8. They are defined to help estimate the multijet background in the signal regions. The SB1 is a subset of the SB2 region in order to provide better sideband region characteristics for estimating the multijets background in SR2.

Furthermore, six signal-free "validation regions" are used to validate the background estimation method in data. They are grouped into two sets of three regions: labelled VS1, VS2, VB1, and VS3, VS4, VB2 in Fig. 4.8. All these regions are enriched in QCD multijet events with much smaller signal-to-background ratios than the signal regions.

Table 4.4: Definition of the signal, sideband, and validation regions used for background estimation. The regions are defined in terms of the ParticleNet discriminators of the H and Y candidate jets, as shown in Fig. 4.8.

| Region name and label in Fig. 4.8 | ParticleNet discriminator | |
|---|---|---|
| | H Jet | Y jet |
| Signal region 1 (SR1) | >0.98 | >0.98 |
| Signal region 2 (SR2) (excludes SR1) | >0.94 | >0.94 |
| Sideband 1 (SB1) | >0.98 | <0.94 |
| Sideband 2 (SB2) | >0.94 | <0.94 |
| Validation signal-like 1 (VS1) | | >0.98 |
| Validation signal-like 2 (VS2) | 0.8–0.94 | 0.94–0.98 |
| Validation sideband 1 (VB1) | | <0.94 |
| Validation signal-like 3 (VS3) | | >0.98 |
| Validation signal-like 4 (VS4) | 0.6–0.8 | 0.94–0.98 |
| Validation sideband 2 (VB2) | | <0.94 |

The signal selection efficiencies range from 1.7% to 12.6% in SR1 and 1.3% to 5.6% in SR2, depending on the signal mass hypothesis. Based on simulation, the background composition is about an equal proportion of $t\bar{t}$+jets and QCD multijets in signal regions and corresponding validation regions, VS1–VS4. The multijet events account for approximately 90% of events in the sideband and validation sideband regions.

### 4.3.3 Offline selection: Semileptonic category

A dedicated control region is defined to use semileptonically decaying $t\bar{t}$ events to help normalize the event yields of the second major background in the signal regions, which originate from the hadronically decaying $t\bar{t}$. The control region is defined by the following requirements:

- One lepton with (an electron or a muon) $p_T > 40\,\text{GeV}$ and $|\eta| < 2.4$;
- AK4 jet exists satisfying: $p_T > 50\,\text{GeV}$, $\Delta R\,(\texttt{lepton,jet}) < 1.5$ and required to be b-tagged;

- MET $> 60$ GeV;
- $H_{\mathrm{T}} > 500$ GeV;
- AK8 jet exists satisfying: $p_T > 350\,(450)$ GeV, $|\eta| < 2.4\,(2.5)$ in 2016(2017,2018) and $\Delta R\,(\texttt{lepton,AK8 jet}) > 2.0$.

where $H_{\mathrm{T}}$ is the scalar sum of $p_T$ of all AK4 jets with $p_T > 30$ GeV and $|\eta| < 2.4\,(2.5)$. The DeepJet algorithm is used to select AK4 jets which originate from a b quark [90], which we call a "b-tagged jet". The loose DeepJet working point, with a mistag rate of 10% and approximately 90% efficiency for b jets, is used.

The requirements for finding a lepton, a b-tagged jet close to the lepton and the $E_T^{miss}$ requirement constitute the three components of the leptonic decay of the t quark. On the other side of the event, a signature of a hadronic decay of the t quark is looked for in the form of an AK8 jet with the same requirements as the AK8 jets in the hadronic selection. Such a jet is called the "probe" jet as we are probing the properties of the hadronic decay of the t quark using these jets.

## 4.4 Background estimate

The analysis searches for a narrow signal in the 2-dimensional plane spanned by $M_{\mathrm{JJ}}$ and $M_{\mathrm{J}}^{\mathrm{Y}}$, in the signal regions 1 and 2. The distributions of the multijet events are estimated with data-driven techniques, with a pass-to-fail ratio method using the sideband regions of Fig. 4.8. The multijet background is estimated for the 3 years combined. This choice is made in order to increase the event statistics in the various signal and sideband regions and thus improve the fit quality.

The $\mathrm{t\bar{t}}$ background is taken from simulations, for each year separately, with the event yields normalized using the data in the semileptonic event category.

The multijets and the $\mathrm{t\bar{t}}$ background estimation methods are explained below. The total background estimation method is validated using data in the validation regions and also using generated (toy) data in the signal regions. The full background estimation also takes into account all systematic uncertainties, which are described in Sec. 4.5.

### 4.4.1 Multijet Background Estimate

We estimate the multijet background in the signal regions using a data-driven method which exploits the sideband regions SB1 and SB2 as defined in Section 4.3.2. First, a general example is described below, then the details of the implementation for this analysis are given.

The method is based on the ratio of multijet event distributions in the $M_{\mathrm{JJ}}$–$M_{\mathrm{J}}^{\mathrm{Y}}$ plane between ParticleNet tagger-passing (signal) and tagger-failing (sideband) regions, $R_{P/F}(M_{\mathrm{JJ}}, M_{\mathrm{J}}^{\mathrm{Y}})$. The goal of the method is to measure the ratio of these distributions in data, $R_{P/F}^{data}(M_{\mathrm{JJ}}, M_{\mathrm{J}}^{\mathrm{Y}})$. A good starting point is to estimate this ratio from MC simulation, $R_{P/F}^{MC}$. In that case, we only need to model a correction function that accounts for the differences between the ratios in simulation and data, as expressed in Eq. 4.1. This is the preferred method as the ratio of pass-to-fail ratios in data and simulation is expected to be easier to model than that the $R_{P/F}^{data}$.

$$R_{P/F}^{data} = R_{P/F}^{MC}\, R_{ratio} \Rightarrow R_{ratio} = \frac{R_{P/F}^{data}}{R_{P/F}^{MC}} \tag{4.1}$$

Due to the combinatorial nature of the multijet processes, both the data and MC pass-to-fail $(M_{\mathrm{JJ}}, M_{\mathrm{J}}^{\mathrm{Y}})$ ratios in Eq.4.1 are smooth. Their ratio, $R_{ratio}(M_{\mathrm{JJ}}, M_{\mathrm{J}}^{\mathrm{Y}})$ will therefore be smooth as well. This ratio-of-ratios can be used to correct for differences between the pass-to-fail ratio in simulation and data, while it should have a less complex shape to parametrize than the $R_{P/F}^{data}(M_{\mathrm{JJ}}, M_{\mathrm{J}}^{\mathrm{Y}})$. The number of multijet events in the i-th bin of the passing category is then given by:

$$n_{Pass,QCD}(i) = n_{Fail,QCD}(i)\, R_{P/F}^{MC}(M_{\mathrm{JJ}}, M_{\mathrm{J}}^{\mathrm{Y}})\, R_{ratio}(M_{\mathrm{JJ}}, M_{\mathrm{J}}^{\mathrm{Y}}), \tag{4.2}$$

where $n_{Fail,QCD}(i)$ is the number of failing multijet events in the i-th bin. It is estimated by subtracting the simulated contribution of other backgrounds from the data yield in that particular bin. The failing regions are constructed so that they are dominated by multijet events making this a good estimate for $n_{Fail,QCD}$. The $R_{ratio}$ is modelled as a product of polynomials in $M_{\mathrm{J}}^{\mathrm{Y}}$ and $M_{\mathrm{JJ}}$. The parameters of the polynomials are determined during the simultaneous fitting procedure, described in Sec. 4.4.3, of the background to the

observed data in both the passing and the failing region.

The implementation of this method for the presented analysis slightly differs from the above description. Measuring the $R_{P/F}^{MC}$ from simulation is not straightforward due to the strong background rejection power of the ParticleNet tagger, which limits the statistics of the multijet background in the ParticleNet-passing regions. This results in large statistical fluctuations of $R_{P/F}^{MC}$. This is further enhanced by the fact that we are working in two dimensions. One way of dealing with this is to employ a smoothing method on the simulated $(M_{\mathrm{JJ}}, M_{\mathrm{J}}^{\mathrm{Y}})$ distributions in the signal regions such as the kernel density estimate, used in the resonant HH $\rightarrow$ 4b analysis [31]. However, the requirement on the H-candidate to fall into the Higgs mass window introduces a dip in the $M_{\mathrm{J}}^{\mathrm{Y}}$ distribution at 100–140 GeV. This feature prevents us from using smoothing methods as they work best with smoothly falling backgrounds. The problem of low statistics in simulation might be remedied by estimating the $R_{P/F}^{MC}$ as a function of $M_{\mathrm{J}}^{\mathrm{Y}}$ only. Even then, the simulation yields in the SR1 and SR2 regions proved to be too low to get a reliable estimate.

A modified version of this method was developed for this analysis. A data-driven estimate of the pass-to-fail ratio, which we call the "initial pass-to-fail ratio" $R_{P/F}^{init}(M_{\mathrm{JJ}}, M_{\mathrm{J}}^{\mathrm{Y}})$, is obtained in the validation regions and applied to the measurement regions. This is done by subtracting the simulation prediction of other backgrounds ($\mathrm{t\bar{t}}$) from the data in regions VS1, VS2 and VB1 and taking the ratios, VS1-to-VB1 and VS2-to-VB1. The two ratios are an initial estimate of the transfer functions used to estimate the multijet background in SR1 and SR2 from the yields in SB1 and SB2 sidebands, respectively. Again, due to limited statistics, the estimate is done only in one dimension $R_{P/F}^{init}(M_{\mathrm{J}}^{\mathrm{Y}})$. In that case, the role of $R_{ratio}(M_{\mathrm{JJ}}, M_{\mathrm{J}}^{\mathrm{Y}})$ is, in addition to correcting for the differences in the $R_{P/F}$ between the two sets of regions, to account for the $M_{\mathrm{JJ}}$ dependence of the pass-to-fail ratio. The $R_{ratio}$ is modelled as a product of two polynomials in $M_{\mathrm{J}}^{\mathrm{Y}}$ and $M_{\mathrm{JJ}}$. The order of the polynomials is determined using the Fisher's F-test [105].

An F-test is any statistical test in which the test statistic is distributed according to the F-distribution under the null hypothesis. The F-distribution with parameters ($d_1$,$d_2$) arises as the ratio of two chi-squared variates: $X = \frac{\chi_1/d_1}{\chi_2/d_2}$, where $\chi_1$ and $\chi_2$ are independent and follow chi-squared distribution with $d_1$ and $d_2$ degrees of freedom. In the F-test we perform, two fit models are compared where the difference between them is in the order of

the polynomials describing the $R_{ratio}$, thus differing in the number of degrees of freedom. The base model (model 1), with $p_1$ parameters, is a subset of the alternative model with $p_2$ degrees of freedom ($p_1 < p_2$). We construct the F-statistic as:

$$F = \frac{-2\log(\lambda_1/\lambda_2)/(p_2 - p_1)}{-2\log\lambda_2/(n_{bins} - p_2)}, \tag{4.3}$$

where $n_{bins}$ is the number of bins in our maximum likelihood fit, described later in Sec. 4.4.3, and $\lambda_i$ is the likelihood value of model $i$. Under the null hypothesis that the alternative model does not provide a significantly better fit than the base model, both the logarithm of the likelihood ratio in the numerator and the logarithm of the likelihood in the denominator follow the chi-square distribution. The expression in the numerator does so according to Wilks' theorem [106]. The expression in the denominator asymptotically follows the chi-square distribution and is a test statistic often used for goodness of fit tests [107]. Therefore, under the null hypothesis, $F$ will follow the F-distribution with $(p_2 - p_1, n_{bins} - p_2)$ parameters. If the observed $F$ test score is in the tail of the F-distribution ($p$-value$<$0.05), we reject the null hypothesis and take the alternative model as the new base model.

The starting base model is a product of first-order polynomials in both variables. It is compared to the two alternative models where one of the two polynomials is promoted to a higher order. In case one of them is shown to fit the data better, it would be taken as the new base polynomial and compared with two other higher-order polynomials until the F-test shows no improvement in the fit quality with the introduction of higher-order polynomials. All the F-tests performed for this analysis showed that the product of first-order polynomials describes the data sufficiently well and it was used to model $R_{Ratio}$.

### 4.4.2 $t\bar{t}$ background estimate using semileptonic control region

The selection of the semileptonic control region, described in Section 4.3.3, results in a very pure ($> 90\%$) sample of hadronically decaying t quark "probe" jets. This category is used to adjust the normalization of the $t\bar{t}$ background in the hadronic category by fitting the simulated $t\bar{t}$ templates in the semileptonic category to the data. The observable in the

control region is the mass of the probe jet. The fitting templates, obtained in simulation, are split into three categories based on the parton content of the probe jet. This allows the measurement of separate correction factors as it is expected that the data-to-simulation corrections may differ between jet categories.

The three categories are: fully-merged top, semi-merged top, and others. Fully-merged jets are required to enclose all partons of the hadronic top decay: the b quark from the top decay and both quarks from the hadronic decay of the W boson (bqq). The semi-merged jets contain the b quark from the top and one of the quarks from the W decay (bq). The category of "other jets" contains jets that are reconstructed from the decay products of the W and jets matching neither of these categories. This categorization of the $t\bar{t}$ events is also applied in the hadronic regions, based on the content of the Y-candidate jet.

The electron and muon channels, defined by the flavor of the lepton produced in the decay chain of the t quark, are added together and the template fit is performed in two regions: tight and loose, based on whether the probe jet passes the tight or loose (but not tight) working point of the ParticleNet tagger. In each of the two regions, we fit two normalization parameters for the fully-merged and semi-merged top jets. The parameters are also separated by year, giving six parameters in total in each region. The same normalization parameters directly change the scale of the fully-merged and semi-merged $t\bar{t}$ categories in the hadronic regions (based on the content of the Y-candidate jet). The parameters are used to implicitly model data-to-simulation corrections. The two major possible sources of the corrections are the $t\bar{t}$ cross section uncertainty and the ParticleNet data-to-simulation efficiency correction for the top quark jets.

### 4.4.3 Maximum likelihood fit

The likelihood function (often simply called likelihood) is the joint probability of the observed data as a function of the parameters of the chosen statistical model. If the function $f(x|\theta)$ describes the probability density for an observable $x$, given parameters $\theta$, the likelihood of observing a series of $\boldsymbol{x} = (x_1, x_2, ..., x_n)$ would be $\mathcal{L}(\boldsymbol{x}|\theta) = \prod_{i=1}^{i=n} f(x_i|\theta)$.

The maximum likelihood estimate of the parameters $r$ and $\theta$ are the values for which $\mathcal{L}(\boldsymbol{x}|r, \theta)$ is at its global maximum. Here the parameter $r$ is the parameter of interest, in

our case that is the signal strength. It is separated out from other parameters, $\theta$, which are called nuisance parameters. A binned maximum likelihood fit to the observed data is jointly performed in the hadronic and semileptonic regions.

In the hadronic regions, the fit is performed in the ($M_{\mathrm{JJ}}$,$M_{\mathrm{J}}^{\mathrm{Y}}$) plane to the combined data of the three data-taking years. The prediction consists of the sum of signal, $\mathrm{t\bar{t}}$ and multijet distributions. A single multijet distribution is used for all the years, while the signal and $\mathrm{t\bar{t}}$ are split by data-taking years. Furthermore, the $\mathrm{t\bar{t}}$ contribution is split into categories based on the parton content as described in Section 4.4.2.

In the semileptonic regions, the mass distribution of the probe jets is used to fit the data, separated into six regions. The regions are defined by the data-taking year and also by the ParticleNet score of the probe jet.

Thus, the fit is performed simultaneously in four hadronic regions (two signal, SR1 and SR2, and two sideband, SB1 and SB2) and six semileptonic regions. The constructed likelihood can be written as

$$\begin{aligned}\mathcal{L}(\mathrm{data}|r,\theta) = &\prod_{i,j} \mathrm{Poisson}(N^{\mathrm{data}}_{SB,i,j}|N^{\mathrm{Multijet}}_{SB,i,j} + N^{\mathrm{t\bar{t}}}_{SB,i,j}) \\ &\times \prod_{i,j} \mathrm{Poisson}(N^{\mathrm{data}}_{SR,i,j}|N^{\mathrm{Multijet}}_{SR,i,j} + N^{\mathrm{t\bar{t}}}_{SR,i,j} + rN^{\mathrm{signal}}_{SR,i,j}) \\ &\times \prod_{k} \mathrm{Poisson}(N^{\mathrm{data}}_{CR,k}|N^{\mathrm{bqq}}_{CR,k} + N^{\mathrm{bq}}_{CR,k} + N^{\mathrm{other}}_{CR,k}) \times \mathrm{Nuisances}(\theta|\theta_0,\sigma_\theta),\end{aligned} \tag{4.4}$$

with the following definitions:

- $\mathrm{Poisson}(y|x)$ is the Poisson probability of observing $y$ events, while expecting $x$. That is, $\mathrm{Poisson}(y|x) = x^y e^{-x}/y!$,
- $N^{data}_{SB(SR),i,j}$ is the observed number of events in the $i^{th}$ $M_{\mathrm{J}}^{\mathrm{Y}}$ and $j^{th}$ $M_{\mathrm{JJ}}$ bin of the sideband (signal) hadronic region,
- $N^{process}_{SB(SR),i,j}$ is the expected number of events for each process in the $i^{th}$ $M_{\mathrm{J}}^{\mathrm{Y}}$ and $j^{th}$ $M_{\mathrm{JJ}}$ bin of the sideband (signal) hadronic region,

- $N_{CR,k}^{data}$ is the observed number of probe jets in the $k^{th}$ mass bin of a semileptonic region,

- $N_{CR,k}^{category}$ is the expected number of probe jets of each jet category in the $k^{th}$ mass bin of a semileptonic region,

- Nuisances$(\theta|\theta_0, \sigma_\theta)$ represent the penalty term coming from the constraints imposed on the nuisance parameters.

The nuisance parameters model the effects of the systematic uncertainties which are described in the following section.

## 4.5 Systematic uncertainties

Several sources of systematic uncertainty are included in the fit model, affecting the ($M_{\mathrm{JJ}}$,$M_{\mathrm{J}}^{\mathrm{Y}}$) shapes and yields of the signals and backgrounds. An example of the impact of the systematic uncertainties is given for the signal with $M_{\mathrm{X}} = 1.6$ TeV and $M_{\mathrm{Y}} = 150$ GeV:

The main uncertainties on the signal are:

- **ParticleNet efficiency scale factor**: the uncertainty in the correction applied to the simulation to match the efficiency of the ParticleNet discriminator in data. The uncertainty is 7–37%, depending on the AK8 jet $p_T$, and leads to a 15% uncertainty on the signal strength.

- **Jet mass scale**: the uncertainty is modelled as a 5% shift in the AK8 jet soft-drop mass. It is decorrelated between the bqq jets, the bq jets, and the signal jets. Besides changing the shape of distributions, it also affects the yields because of the Higgs mass condition applied to one of the jets. It impacts the signal by 13%. The JMS uncertainty on the $\mathrm{t\bar{t}}$+jets background is reduced by including the semileptonic control region.

- **Jet energy scale and resolution**: the uncertainties of the measured jet energy scale and resolution are applied to both AK4 and AK8 jets. They are fully correlated between the two sets of jets. The signal is impacted by 5%.

- **Jet mass resolution**: the simulated AK8 jet masses are smeared to match their distributions in the data. The level of resolution smearing is based on studies using Lorentz-boosted W $\rightarrow$ qq$'$ (W boson jets). The nominal simulated jet mass resolution is used to model the downward uncertainty while a 20% resolution smearing [108] is taken as upward variation to the AK8 jet mass resolution. It results in 4% impact on the signal strength.

The following uncertainties affect only the backgrounds:

- $\mathrm{t\bar{t}}$ **normalization**: the uncertainties in the parameters used to describe the data-based correction for the $\mathrm{t\bar{t}}$+jets background range from 6% to 16%.

- **Top quark** $p_T$ **modelling**: an uncertainty is associated with the top quark $p_T$ modelling in Monte Carlo simulations of the $\mathrm{t\bar{t}}$ process [109], resulting in a 2% uncertainty in the $\mathrm{t\bar{t}}$+jets background.

- **Pass-to-fail ratio uncertainty**: the main source of uncertainty for the multijet background comes from the uncertainty in the $R^{init}_{P/F}(M_{\mathrm{J}}^{\mathrm{Y}})$ fit and is proportional to the statistical uncertainty because of the sample size in the VS1, VS2 and VB1 regions. Its impact of 7–11% on the multijet background is evaluated by calculating the yield change when applying the $R^{init}_{P/F}(M_{\mathrm{J}}^{\mathrm{Y}})$ to the failing regions with uncertainties.

Other considered sources of systematic uncertainties with minor impact are:

- **Trigger efficiency**: an uncertainty equal to half of the difference between unity and the measured trigger efficiency is assigned. This is larger than the statistical uncertainty of the measurement in order to cover the jet energy scale uncertainties in the trigger selections. The maximum value of this uncertainty is 3%.

- **Trigger timing correction**: During the 2016 and 2017 data taking, a gradual shift in the timing of the inputs of the ECAL hardware level trigger caused a specific

trigger inefficiency. Based on the simulation of its effect on the signal and $t\bar{t}$ yield, a 2% normalization uncertainty is applied to these processes in 2016 and 2017.

- **Integrated luminosity**: the uncertainty on the total RunII integrated luminosity [110, 111] is 1.6%.

- **Pileup**: the assumed value of the pp total inelastic cross section (69.2 mb), that is used in the simulation of pileup events, is varied upwards and downwards by its uncertainty of 4.6% [112].

- **PDF and scale uncertainties**: the combined impact of the PDF and the QCD factorization and renormalization scale uncertainties on the signal acceptance and selection is estimated to be 1.0%. The former is derived using the PDF4LHC procedure [113], and the latter by changing the renormalization ($\mu_R$) and factorization ($\mu_F$) scales in simulation by factors of 0.5 and 2.

- **Sample size of sideband regions**: the statistical uncertainties associated with the number of events in the sideband regions SB1 and SB2 impact the estimated multijets background in the SR1 and SR2 signal regions. These uncertainties are small compared with the uncertainty in the $R^{init}_{P/F}(M_{\mathrm{J}}^{\mathrm{Y}})$ and are included in the Likelihood using the Barlow–Beeston Lite prescription [114, 115].

- **Lepton ID and isolation**: the lepton identification and isolation efficiency data-to-simulation correction factors have uncertainties that affect the event yields by 1–2% in the semileptonic selection.

- b**-tagging scale factor uncertainty**: a 2% uncertainty on the correction applied in the simulation to match the shape of the DeepJet discriminator in data is applied and affects the $t\bar{t}$ sample in the semileptonic region.

All uncertainties are decorrelated between years, except for the PDF and scale uncertainties, pileup, luminosity and top quark $p_T$ modelling uncertainties.

## 4.6 Validation of the background estimation method

In order to avoid any bias, the data in the signal regions are blinded until the background estimation methods are settled upon. Therefore, before unblinding the data in the signal regions, a validation fit is performed to check for the robustness of the fit method and its ability to achieve satisfying closure to the data. The fit to data in the validation regions is described in Sec. 4.6.1. An additional check, described in Sec. 4.6.2, is made by generating toy data in the signal regions and performing the fit to this toy data. The main purpose of this test is to inject varying amount of signal in the toy data in order to check if the ML fit will find the proper value for the signal strength parameter. In both fits, the goodness-of-fit test was performed and the test passed the p-value threshold of 0.05, confirming good agreement between the model and the data.

### 4.6.1 Validation using hadronic validation regions

A background estimation fit is performed in validation regions, VS1 and VS2, to check if the method can well describe the data. These regions are proxies, with a lower signal-to-background ratio, for the SR1 and SR2 signal regions. The corresponding sideband region, for both VS1 and VS2, is the VB1 region.

The 1-dimensional, initial pass-to-fail ratios are derived from the VS3 and VS4 regions. Two ratios are derived:

(1) $R_{P/F}^{VS3}$, which is the ratio of the $M_{\mathrm{J}}^{\mathrm{Y}}$ distributions in the VS3 to the VB2 regions.

(2) $R_{P/F}^{VS4}$, which is the ratio of the $M_{\mathrm{J}}^{\mathrm{Y}}$ distributions in the VS4 to the VB2 regions.

The $R_{P/F}^{VS4}$ is used as the nominal initial pass-to-fail ratio for predicting the multijet background in the VS2 region. Likewise, the $R_{P/F}^{VS3}$ serves as the initial input of the pass-to-fail ratio for the multijet background estimate in the VS1 region. They are modelled as second-order polynomials in $M_{\mathrm{J}}^{\mathrm{Y}}$. Figure 4.9 shows the $R_{P/F}^{VS4}$ and $R_{P/F}^{VS3}$ $M_{\mathrm{J}}^{\mathrm{Y}}$ distributions and the fitted functions.

Both $R_{Ratios}$ are products of first order polynomials in $M_{\mathrm{J}}^{\mathrm{Y}}$ and $M_{\mathrm{JJ}}$. The order of the polynomials is determined with the F-test procedure explained in Sec. 4.4. Their pa-

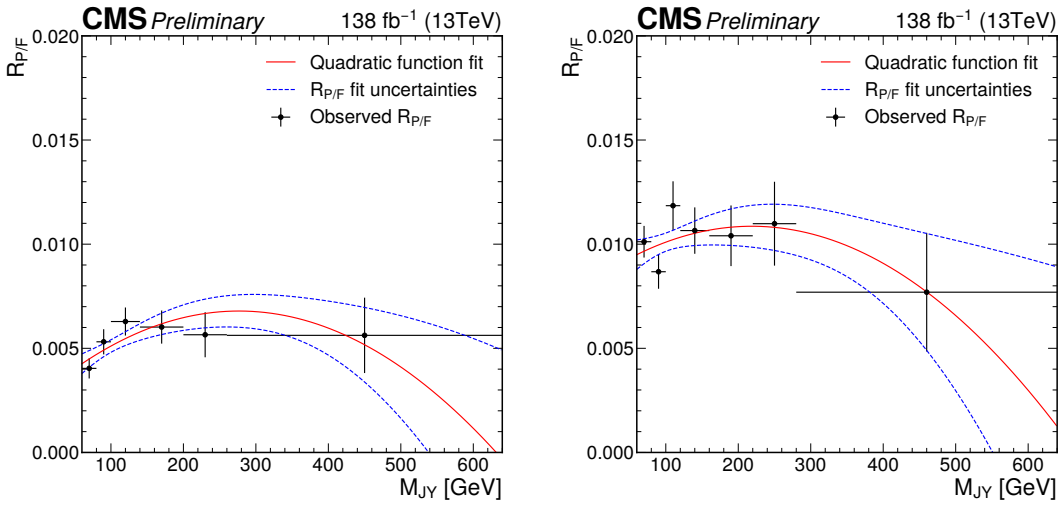



Figure 4.9: Data driven $R^{init}_{P/F}$ fits for VS3 (left) and VS4 (right) regions for RunII data. Simulation prediction of the $t\bar{t}$ in all three years is subtracted from the combined observed data and a second order polynomial is fitted to the result.

rameters are determined in the combined maximum-likelihood fit, using the full RunII data in hadronic and semileptonic regions, as described in Sec. 4.4.3. The systematic uncertainties, listed in Sec. 4.5, are also included in the fit.

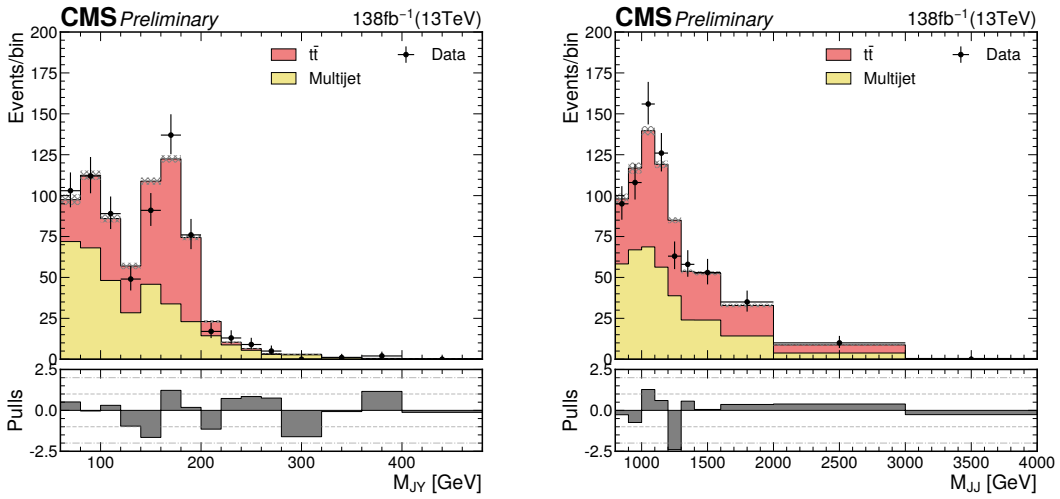



Figure 4.10: Projections on $M_J^Y$ and $M_{JJ}$ in VS1 region after a joint hadronic and semileptonic fit using full RunII data.

The projections of the $M_{JJ}$–$M_J^Y$ postfit distributions for the VS1 and VS2 regions are shown in Fig. 4.10 and Fig. 4.11. The $R_{Ratio}$ and the final $R_{P/F}$, which is the product of the initial estimate and the $R_{Ratio}$ are given in Fig. 4.12

The postfit distributions of the semileptonic $t\bar{t}$ regions for the 3 separate years and the

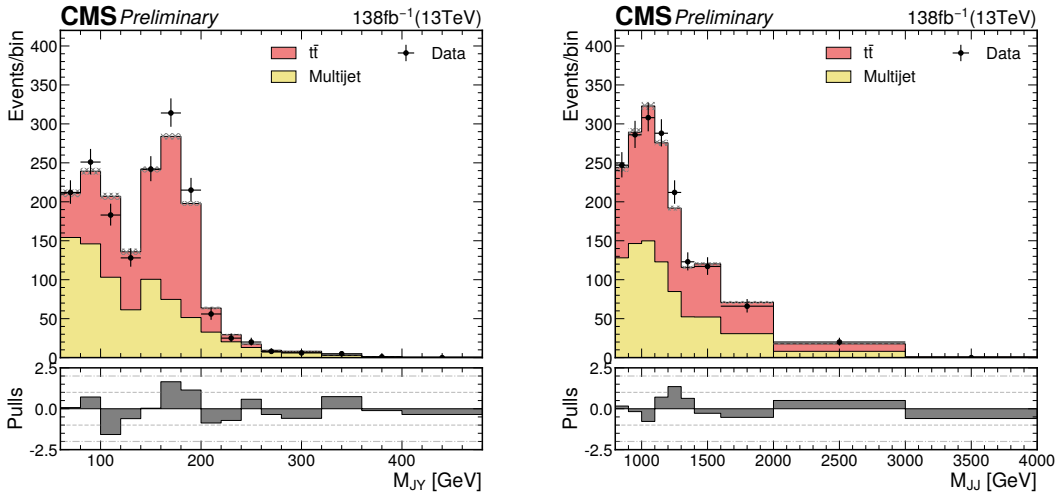



Figure 4.11: Projections on $M_{\mathrm{J}}^{\mathrm{Y}}$ and $M_{\mathrm{JJ}}$ in VS2 region after a joint hadronic and semileptonic fit using full RunII data.

Table 4.5: Postfit values of the rate parameters affecting the fully- and semi-merged $\mathrm{t\bar{t}}$ yield in VS1 (T) and VS2 (L) hadronic (semileptonic) regions.

| Normalization scale | 2016 | 2017 | 2018 |
|---|---|---|---|
| Semi-merged loose | $1.06 \pm 0.07$ | $1.32 \pm 0.15$ | $1.13 \pm 0.11$ |
| Semi-merged tight | $1.17 \pm 0.10$ | $1.21 \pm 0.17$ | $1.32 \pm 0.15$ |
| Fully-merged loose | $0.82 \pm 0.06$ | $0.86 \pm 0.07$ | $0.93 \pm 0.06$ |
| Fully-merged tight | $0.78 \pm 0.07$ | $0.89 \pm 0.08$ | $0.88 \pm 0.07$ |

2 regions are shown in Fig. 4.13. The measured data-to-simulation correction factors for the $\mathrm{t\bar{t}}$ background are reported in Table 4.5. The level of agreement is tested with a goodness-of-fit test using the logarithm of the likelihood, $\log\lambda$, mentioned in Sec. 4.4. To estimate the distribution of the test statistic, 500 toy datasets are used. The distribution of the test statistic is compared with the observed value. The test passes the critical p-value of 0.05, leading to the conclusion that the method achieves satisfactory closure in data.

### 4.6.2 Validation using toy data in signal regions

An additional validation is performed to test the method performance in the signal region. Since the masses of X and Y are unknown, the entire $M_{\mathrm{J}}^{\mathrm{Y}}$–$M_{\mathrm{JJ}}$ planes of the SR1 and SR2 regions are considered as signal regions and cannot be used before unblinding. The background estimate in signal regions (before unblinding) therefore needs to come from

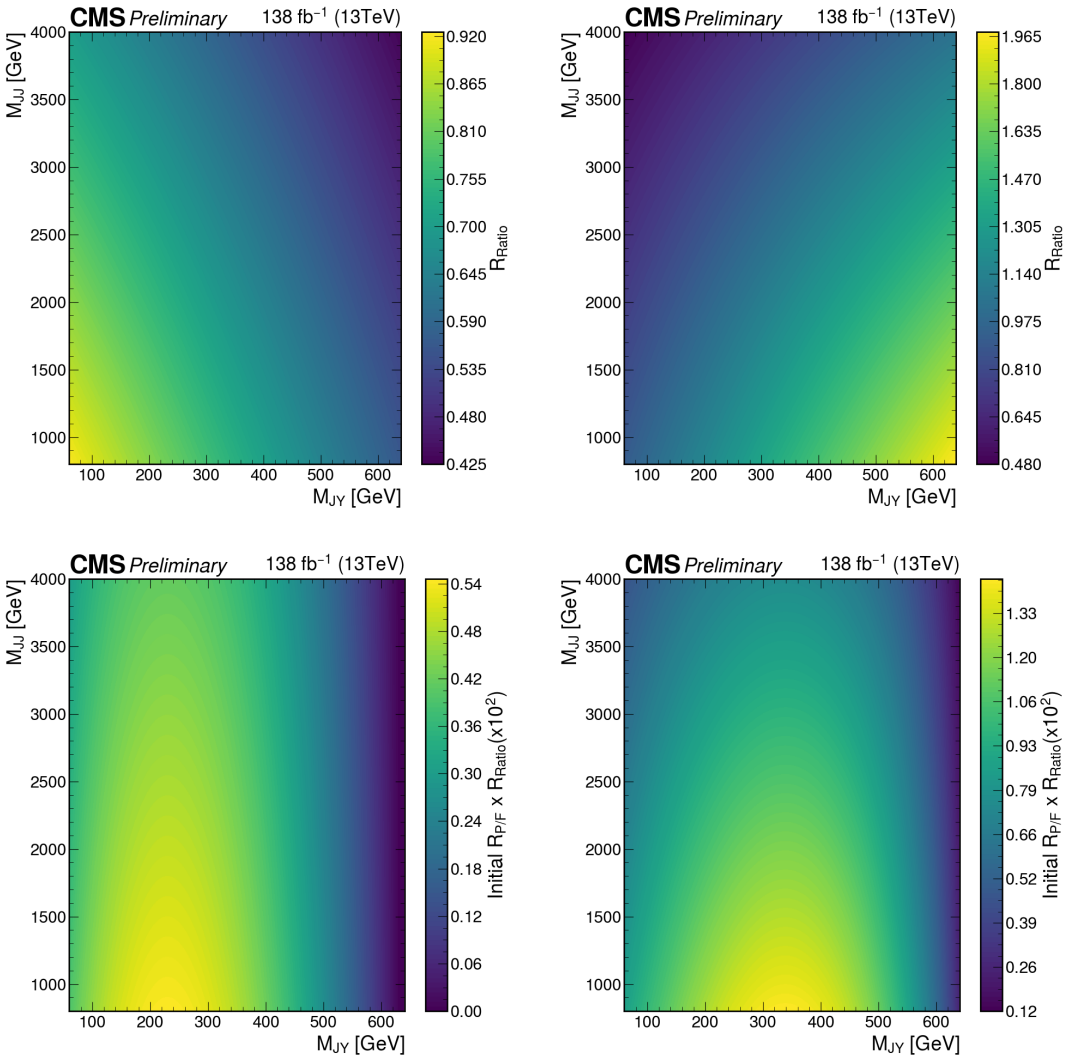



Figure 4.12: $R_{Ratio}$ (upper) and the final $R_{P/F}$ (lower) for the VS1 (left) and VS2 (right) regions after a joint hadronic and semileptonic fit using full RunII data.

either simulation or with the help of other regions. As stated before, due to the ParticleNet tagger's strong rejection of multijet events, the amount of generated statistics is not sufficient for a reliable estimate. For this reason, a toy multijet background dataset is generated in the signal region as follows. We take the fitted pass-to-fail ratio in the validation region, $R_{P/F}^{VS1,2}(M_{\mathrm{JJ}}, M_{\mathrm{J}}^{\mathrm{Y}})$, shown on the lower plots in Fig. 4.12. These pass-to-fail ratios are applied to the SB1 and SB2 regions in data to estimate the multijet background in the SR1 and SR2 regions.

The $\mathrm{t\bar{t}}$ sample contains sufficient statistics to estimate the $\mathrm{t\bar{t}}$ contribution in the signal regions from simulation. On top of the simulation, the corrections on the normalization of fully-merged and semi-merged $\mathrm{t\bar{t}}$ jets, obtained in the validation fit described in the

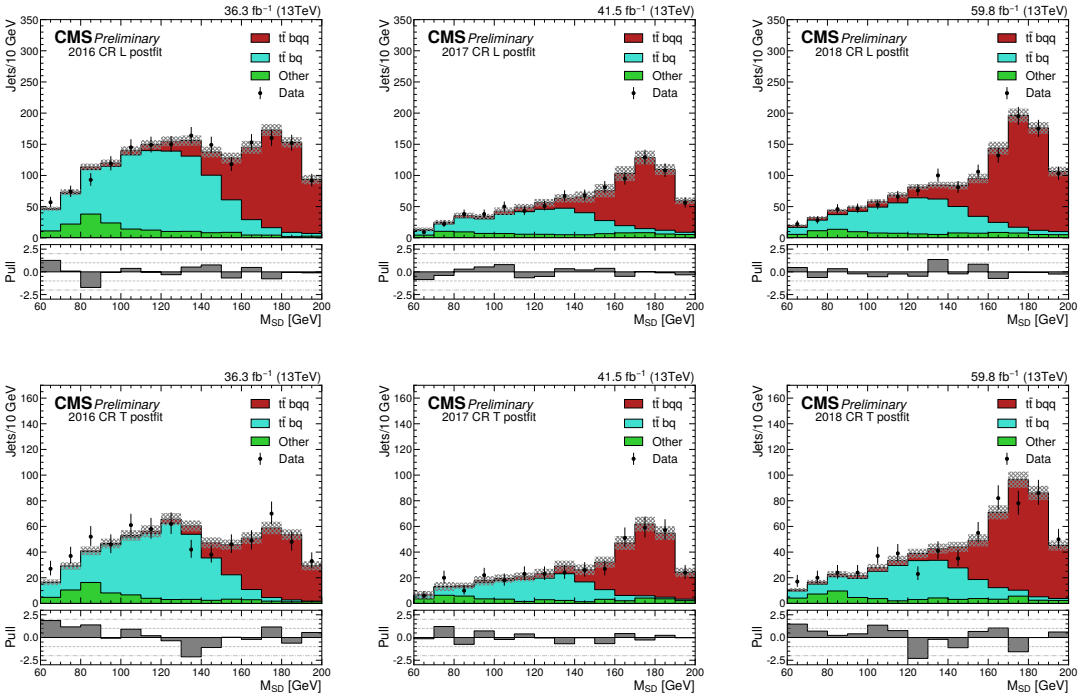



Figure 4.13: Semileptonic control region postfit plots in the loose and tight regions for 2016, 2017 and 2018 data.

previous section and shown in Table 4.5, are applied.

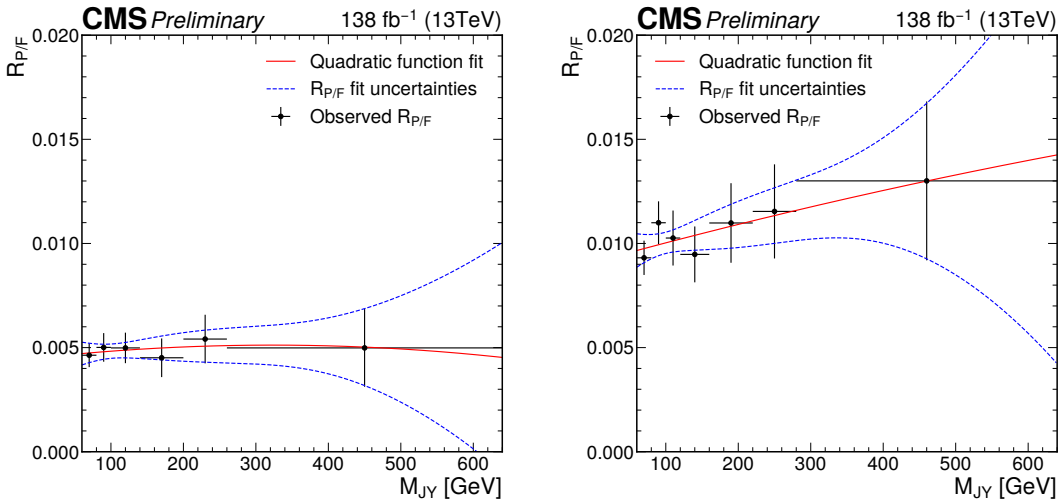



Figure 4.14: Data driven $R^{init}_{P/F}$ fits for VS1 (left) and VS2 (right) regions for RunII data. Simulation prediction of the $t\bar{t}$ in all three years is subtracted from the combined observed data and a second order polynomial is fitted to the result.

The toy data is generated using the sum of the estimated multijet and $t\bar{t}$ backgrounds. When applying the background estimation to the generated toy data, the $R_{P/F}$ measurements from the validation regions VS1 and VS2, shown in Fig. 4.14, are used as $R^{init}_{P/F}$.

The same ratios are also provided as the initial $R_{P/F}$ in the later unblinded fit to the data in the signal regions.

The primary goal of the background estimate using the toy data is to run the signal injection tests. In the signal injection test, the toys are generated using the estimated background in conjunction with signal shapes, obtained from simulation, at specified strength. Maximum likelihood fits are performed to these toys and the distribution of the obtained signal strengths is extracted. If there is no bias introduced by the method, the distribution of $(r - r_{\text{injection}})/\sigma_r$ should be a Gaussian with mean zero and unit width. A set of tests for a fixed mass hypothesis of $M_{\text{X}} = 1600$ GeV and $M_{\text{Y}} = 150$ GeV is shown in Fig. 4.15. No bias is observed, except in the case when no signal is injected. This is understood to be coming from the low levels of background in the signal regions. A repeated test, using background inflated by a factor of 100, returns no bias, confirming this to be the case.

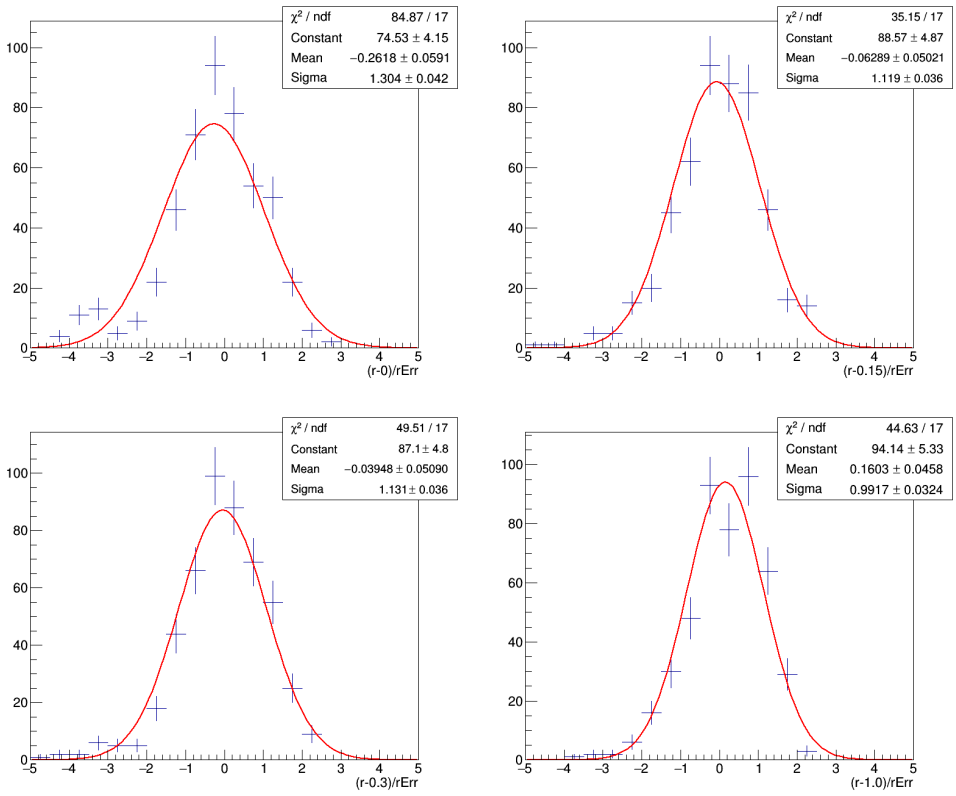



Figure 4.15: Signal injection tests for r = 0, 0.15, 0.3 and 1.0 with 500 toys, using $M_{\text{X}}$ = 1600 GeV and $M_{\text{Y}}$ = 150 GeV signal mass point. r=1.0 corresponds to signal cross section of 1 fb (0.3 fb is the expected exclusion limit). The bias in the case of no signal injected is due to low background levels in signal region. Repeated test using the inflated background shows no bias.

# Chapter 5

# Results

Having done the validation tests of Ch. 4, the data is unblinded and the maximum likelihood fit to data is performed in the signal regions. The observed data and predicted background, from the background-only fit, in the two signal regions are shown in Figs. 5.1 and 5.2. The background prediction is in good agreement with the data which is also confirmed by the goodness-of-fit test. The search for excesses in data, that is, the search for the presence of signal is described in Sec. 5.1. The procedure of setting the upper limits on the signal process cross section outlined in Sec. 5.2. Finally, the results of this search are compared with related analyses in Sec. 5.3.

## 5.1 Signal search and significances

The search for the presence of signal is done by comparing the signal plus background hypothesis ($H_1$) with the background-only hypothesis ($H_0$) for 260 different mass points shown in Fig. 4.1. The significance of the observed result for a given mass point is calculated using a ratio of profile likelihoods of the two hypotheses, $q$,

$$q = 2\ln[\mathcal{L}_{s+b}/\mathcal{L}_b] = 2\ln[\mathcal{L}(\text{data}|r=\hat{r},\hat{\theta}_r)/\mathcal{L}(\text{data}|r=0,\hat{\theta}_0)], \tag{5.1}$$

where $\hat{r}$ is the value of the signal strength which maximizes the likelihood under $H_1$, while $\hat{\theta}_0$ and $\hat{\theta}_r$ nuisance parameters maximizing the likelihood under $H_0$ and $H_1$, respectively.

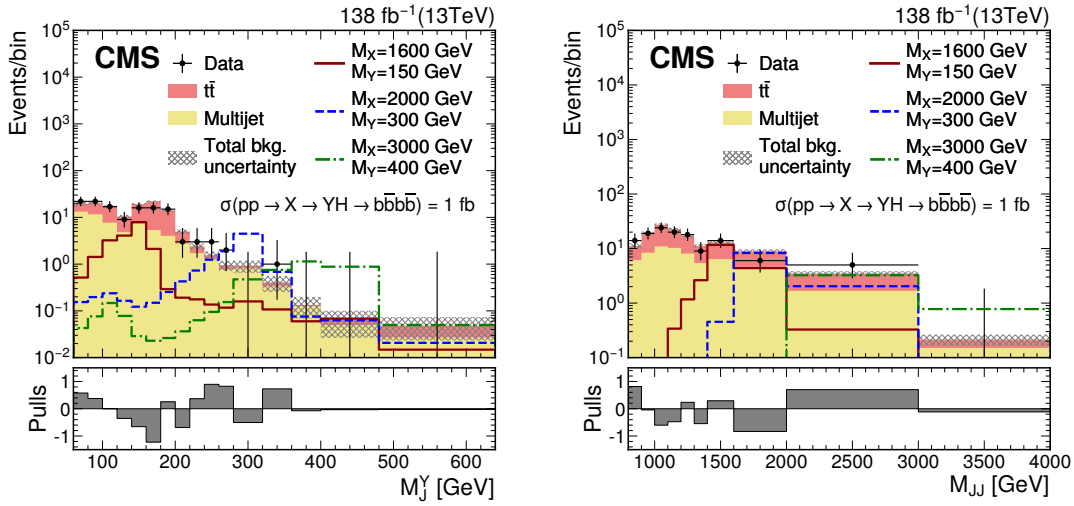



Figure 5.1: The $M_{\mathrm{J}}^{\mathrm{Y}}$ (left) and $M_{\mathrm{JJ}}$ (right) distributions for the number of observed events (black markers) compared with the estimated backgrounds (filled histograms) in the signal region 1. The distributions expected from the signal under three $M_{\mathrm{X}}$ and $M_{\mathrm{Y}}$ hypotheses and assuming a cross section of 1 fb are also shown. The lower panels show the "Pulls" defined as (observed events−expected events)$/\sqrt{\sigma_{obs}^2 - \sigma_{exp}^2}$, where $\sigma_{obs}$ and $\sigma_{exp}$ are the statistical and total uncertainties in the observation and the background estimation, respectively.

Wilk's theorem states that, under the $H_0$ hypothesis, $q$ follows the chi-square distribution with one degree of freedom. This allows us to calculate the significance of the observed result. The 2D map of observed local significance is shown in Fig. 5.3.

The signal hypothesis $M_{\mathrm{X}} = 1.6$ TeV and $M_{\mathrm{Y}} = 90$ GeV gives the highest observed local significance of $3.1\sigma$ ($p$-value$\,=0.001$). However, the global significance, defined as the probability of observing $p \leq p_{min}^{obs}$ under the no-signal hypothesis on at least one of all considered (260) mass scenarios, is much greater than $p_{min}^{obs}$. The fact that many mass hypotheses are considered needs to be taken into accounts since it is expected that there will be significant deviations for a certain number of them. This is known as the look-elsewhere effect (LEE). This effect can be estimated by throwing toy datasets and counting the frequency of having $p < p_{min}^{obs}$. This may require a large number of toy datasets so the estimation of the LEE is done using the method described in Ref. [116].

The method requires the definition of "excursion sets". This is a set of signal mass points for which the test-statistic is greater than some level, $q > u$. On these sets, for reasons which will get described below, the Euler characteristic, $\phi$, is calculated. For 2D shapes, this can be regarded as the number of disconnected components minus the number of

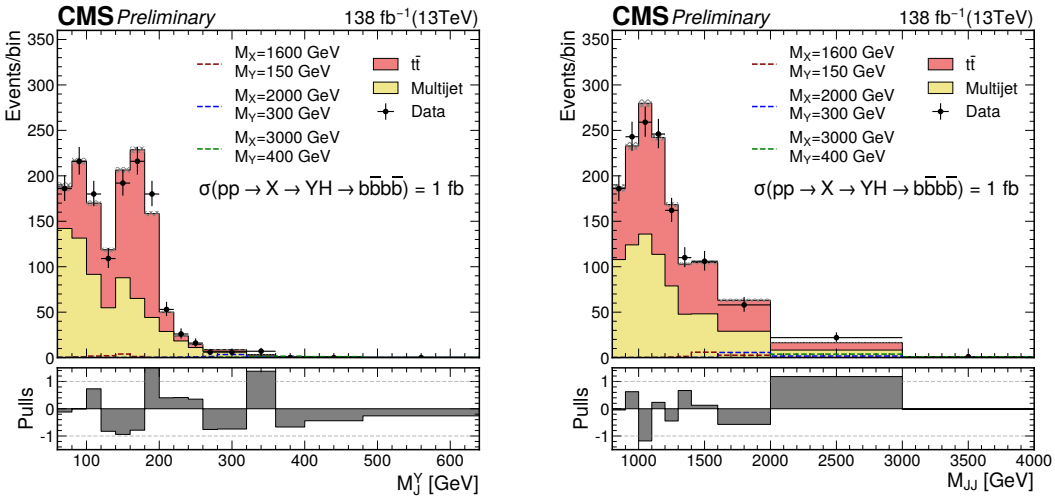



Figure 5.2: The $M_{\mathrm{J}}^{\mathrm{Y}}$ (left) and $M_{\mathrm{JJ}}$ (right) distributions for the number of observed events (black markers) compared with the estimated backgrounds (filled histograms) in the signal region 2. The distributions expected from the signal under three $M_{\mathrm{X}}$ and $M_{\mathrm{Y}}$ hypotheses and assuming a cross section of 1 fb are also shown. The lower panels show the "Pulls" defined as (observed events−expected events)/$\sqrt{\sigma_{obs}^2 - \sigma_{exp}^2}$, where $\sigma_{obs}$ and $\sigma_{exp}$ are the statistical and total uncertainties in the observation and the background estimation, respectively.

"holes", as shown in Fig. 5.4. As the level of the excursion set, $u$, is increased, the number of mass points satisfying this condition becomes smaller so the sets usually contain only a few disconnected regions. For even higher levels, the sets are mostly empty ($\phi = 0$) and rarely contain only one region ($\phi = 1$). Its expectation value therefore converges asymptotically to the probability of observing the test-statistic above the set threshold. In other words, $\mathbf{E}\left[\phi(u_{max}^{obs})\right]$ gives us the global significance corresponding to the largest observed local significance. The crux of the method is therefore the estimation of the expected value of $\phi(u)$ with which one can obtain the global significance for low $p-$value without throwing a large number of toy datasets.

The estimation of the $\mathbf{E}\left[\phi(u_{max}^{obs})\right]$ can be done using a known expression for a $\chi^2$ random field with one degree of freedom and two search dimensions [116]. The expected value is given by relation

$$\mathbf{E}\left[\phi(u)\right] = \mathbf{P}(\chi^2 > u) + e^{-u/2}(N_1 + \sqrt{u}N_2), \tag{5.2}$$

where $N_1$ and $N_2$ are unknown coefficients. We measure the $\phi(u)$ for two different levels,

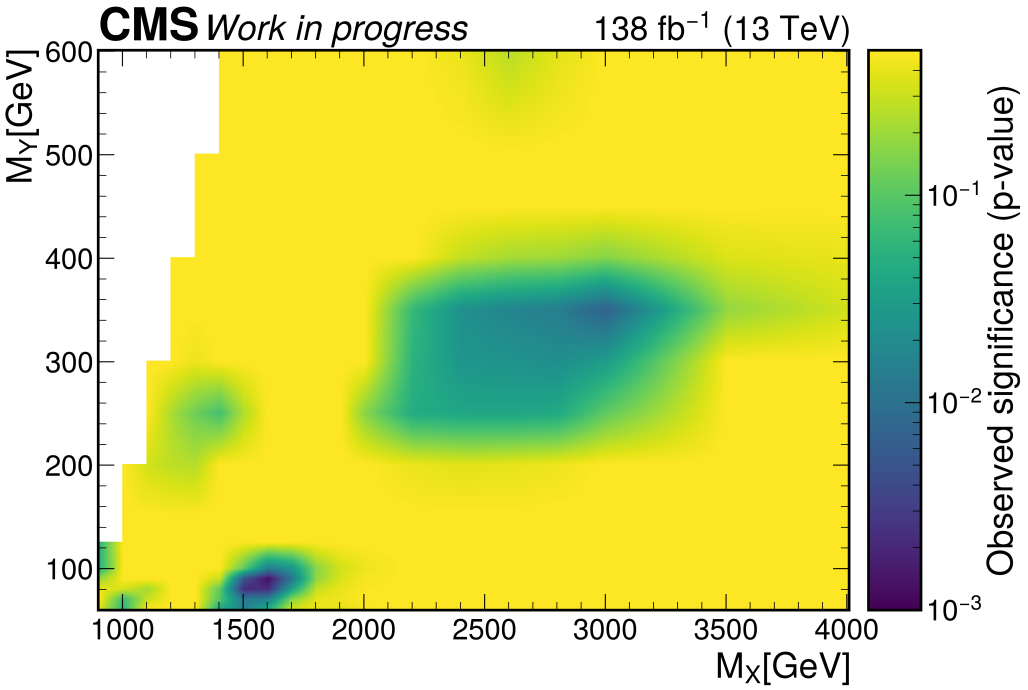



Figure 5.3: The significance of the observed data, expressed as the p-value. The mass points for which no excess is observed have their p-values set to 0.5. The lowest p-value observed is $1.1 \times 10^{-3}$ for $M_X = 1600$ GeV, $M_Y = 90$ GeV

$u_1 = 1, u_2 = 4$, on a set of N=100 toys and calculate their averages to estimate $\mathbf{E}\left[\phi(u_{1,2})\right]$. This returns two sets of equations which can be solved to extract $N_1$ and $N_2$. The chosen levels are selected to be smaller than the largest observed local significance, but large enough so that the regions in the excursion sets are disjointed. An example of excursion sets for the two chosen levels on one toy dataset are shown in Fig. 5.5.

The average values and the standard deviation divided by $\sqrt{N}$ of the Euler characteristics are $\phi(u = 1) = 6.56 \pm 0.21$ and $\phi(u = 4) = 2.51 \pm 0.22$. The corresponding parameters are $N_1 = 2.73$, $N_2 = 7.82$. This finally gives

$$\mathbf{E}(u_{max}^{obs} = 9.34) = 0.25$$

Therefore, the global $p-$value is estimated to be $0.68\,\sigma$ ($p-$value $= 0.25$). The value is large enough to be able to look at the generated toys to cross-check the result. If the probability of observing $p \leq p_{min}^{obs}$ on at least one of all the considered mass scenarios is 0.25, we would expect to see around 25 toy datasets (out of 100) satisfying this condition. The number of toys passing the condition is 31 which is in good agreement with the

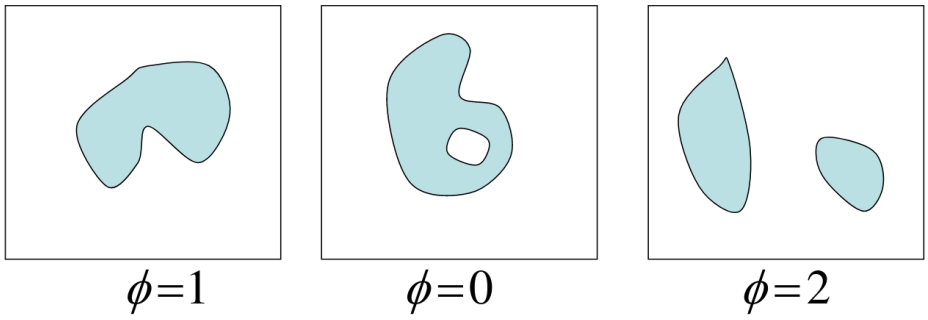



Figure 5.4: Illustration of the Euler characteristic of some 2-dimensional bodies. Taken from Ref. [116]

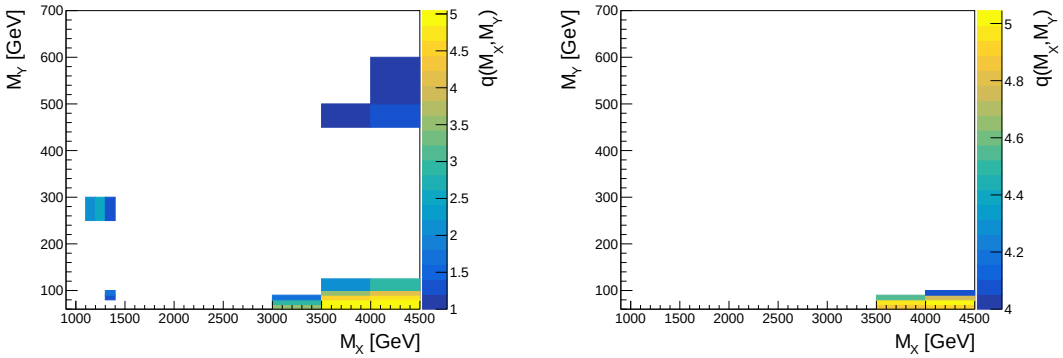



Figure 5.5: Excursion sets for one toy dataset and excursion levels $u_1 = 1$ (left) and $u_2 = 4$ (right). The corresponding Euler characteristics are $\phi(u = 1) = 4$ and $\phi(u = 4) = 1$.

prediction.

## 5.2 Exclusion limits

The estimated background is in good agreement with the observed data. Upper limits for the production cross section of pp $\rightarrow$ HY $\rightarrow$ $\mathrm{b\bar{b}b\bar{b}}$ are calculated for various hypothesized values of $M_\mathrm{X}$ and $M_\mathrm{Y}$. The upper limits are computed with a modified frequentist approach, using the $\mathrm{CL_S}$ criterion [117, 118] with the profile likelihood ratio $q$, introduced with the significance test, used as the test statistic with the asymptotic approximation [119] at 95% confidence level. The $\mathrm{CL_S}$ is defined as

$$\mathrm{CL_S}(r) = \frac{\mathrm{CL_{s+b}}(r)}{\mathrm{CL_b}},$$

where $CL_{s+b}$ and $CL_b$ are the probabilities to observe the value of the test statistic, $q$, as large or larger than the observed $q_{obs}$:

$$CL_{s+b}(r) = P_r(q > q_{obs}|r) = \int_{q_{obs}}^{+\infty} dq\, f(q|r),$$

$$CL_b(r = 0) = P_0(q > q_{obs}|r = 0) = \int_{q_{obs}}^{+\infty} dq\, f(q|r = 0),$$

where $f(q|r)$ and $f(q|r = 0)$ are the probability density functions for the test statistic, $q$, under signal plus background and background only hypotheses, respectively.

One approach in signal exclusion at 95% confidence level would be, for example, to exclude signal hypothesis if $CL_{s+b}$ is less than 0.05. In other words, the exclusion is stated when the signal-plus-background hypothesis is not compatible with the observation. However, $CL_S$ is preferably used over $CL_{s+b}$ because it prevents the exclusion in the regions where the analysis is not sensitive to the presence of signal. For example, let us assume that in some experiment 100 background and 8 signal events are expected for some nominal signal strength, while the observation is 75. Using $CL_{s+b}$ criterion would lead to exclusion of the signal hypothesis for large values of signal strength, however, it is clear that there is a large downward fluctuation in the background. The large fluctuation in the background also makes the background-only hypothesis less compatible with the observation which is reflected in the $CL_b$. The low value of $CL_b$ in denominator of $CL_S$ will increase its value and protect against the exclusion of signal hypothesis originating from the downward background fluctuation. More details on the $CL_S$ method can be found in Refs. [117],[118].

The expected and observed limits ranging from 0.1 fb to 150 fb as a function of $M_X$ and $M_Y$ are shown in Fig. 5.6. Comparing the cross section limits for the $pp \to X \to YH \to b\bar{b}b\bar{b}$ process with the maximally allowed predictions of NMSSM and TRSM leads to the exclusion of mass ranges for these models.

In the NMSSM, no mass range is excluded by the median expected limits. However, the observed limits exclude an area $1.00 < M_X < 1.15$ TeV and $101 < M_Y < 145$ GeV. For TRSM, an expected exclusion area with the bounds $0.90 < M_X < 1.26$ TeV and $100 < M_Y < 126$ GeV is found while the observed exclusion range spans $0.95 < M_X < 1.33$ TeV and $110 < M_Y < 132$ GeV.

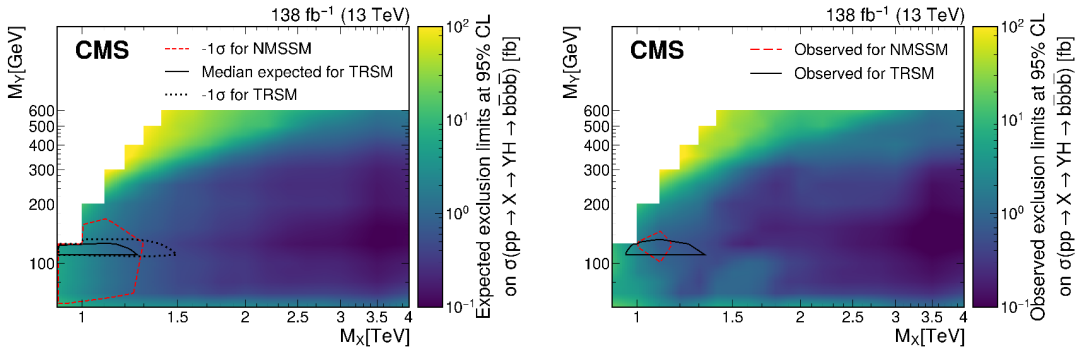



Figure 5.6: The 95% confidence level expected (left) and observed (right) upper limits on $\sigma(\mathrm{pp} \to \mathrm{X} \to \mathrm{YH} \to \mathrm{b\bar{b}b\bar{b}})$ for different values of $M_\mathrm{X}$ and $M_\mathrm{Y}$. The areas within the red and black contours represent the regions where the cross sections predicted by NMSSM and TRSM, respectively, are larger than the experimental limits. The areas within the dashed and dotted contours on the left show the excluded masses at $-1$ standard deviation from the expected limits.

## 5.3 Comparison with related results

The results of search for the $\mathrm{X} \to \mathrm{YH} \to \tau\tau\mathrm{b\bar{b}}$ decay by the CMS experiment [28] are shown in Fig. 5.7. The excluded mass range is for $M_\mathrm{Y} < 200$ GeV and $400 < M_\mathrm{X} < 600$ GeV, which is outside the mass range probed by the presented search. Therefore, the results presented in the previous Section complement the ones shown in Fig. 5.7. Additionally, the limits set in the boosted regime by the 4b search are generally lower than the ones set by the $\tau\tau\mathrm{b\bar{b}}$ final state search. This is partly helped by the higher branching ratio of Higgs decaying to $\mathrm{b\bar{b}}$, leading to the improvement in the exclusion limits set on $\sigma(\mathrm{X} \to \mathrm{YH})$ by as much as 2 orders of magnitude.

For the purpose of comparing the analysis results with the $\mathrm{pp} \to \mathrm{HH} \to \mathrm{b\bar{b}b\bar{b}}$ searches, the observed limits shown in Fig. 5.6 are displayed with a $M_\mathrm{Y} = 125$ GeV condition in Fig. 5.8.

The observed limits are in the 10-0.1 fb range for $M_\mathrm{X}$ in the 0.9–4.0 TeV range. The results of the boosted resonant di-Higgs, in the four b quark final state, search by CMS [31] collaboration is shown in Fig. 5.9. The observed limits are in the 10–0.3 fb range for $M_\mathrm{X}$ within 1.0–3.0 TeV. The current analysis improves the limits on $\mathrm{pp} \to \mathrm{HH} \to \mathrm{b\bar{b}b\bar{b}}$ approximately by a factor of two compared to Ref.[31], owing to the improved background rejection of the ParticleNet algorithm compared to the older DeepAK8-MD algorithm.

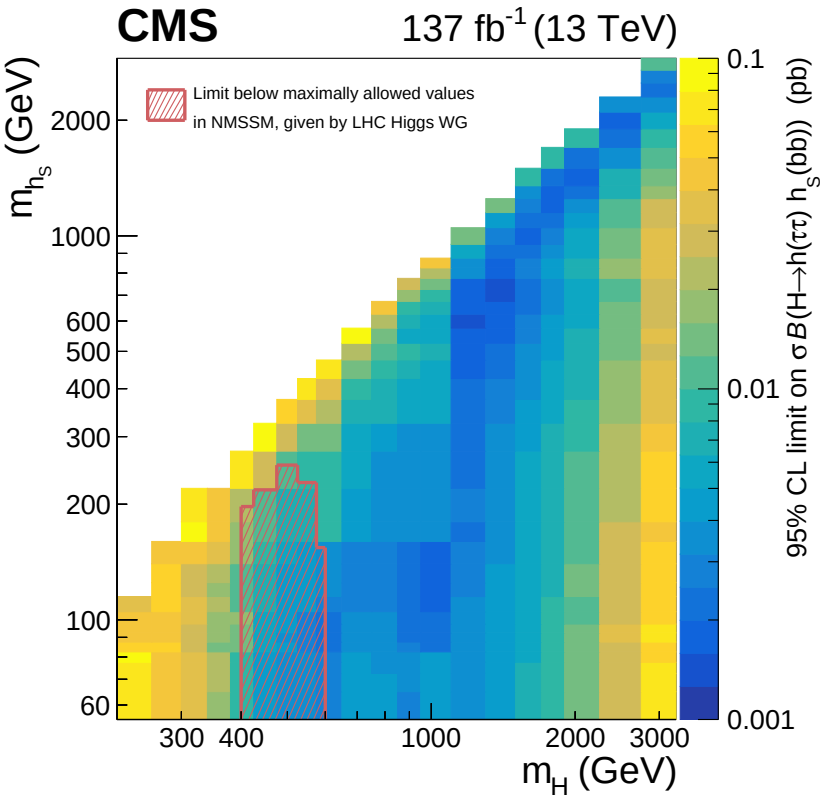



Figure 5.7: The 95% confidence level observed upper limits on $\sigma(\mathrm{X} \to \mathrm{YH} \to \tau\tau\mathrm{b\bar{b}})$ for different values of $M_\mathrm{X}$ and $M_\mathrm{Y}$. The region where the observed limits fall below the maximally allowed cross section in the NMSSM are shown in the red hatched area. Figure taken from Ref. [28].

These results can, therefore, be interpreted in the context of resonant di-Higgs search and improve upon the theoretical constraints relevant for the di-Higgs production.

The results of the CMS [18] and ATLAS [19] searches for a heavier scalar X decaying to a pair of $\tau$ leptons are shown in Fig. 5.10. The limits are expressed as a function of $M_\mathrm{X}$ and the tan$\beta$ parameter of the MSSM theory as the production cross section depends on it. The results from ATLAS give stronger constraints than the results from CMS due to higher processed integrated luminosity. ATLAS [19] set a limit on the resonance mass at $M_\mathrm{X} < 1.5(1)$ TeV for tan$\beta = 21(8)$. The NMSSM favours low tan$\beta$ (typically tan$\beta < 6$) where the current $M_\mathrm{X}$ bounds are the weakest. Therefore, the exclusion of $M_\mathrm{X} \approx 1.1$TeV in the context of NMSSM, set by this analysis, can be reinterpreted to further increase the mass exclusion in MSSM for low tan$\beta$.

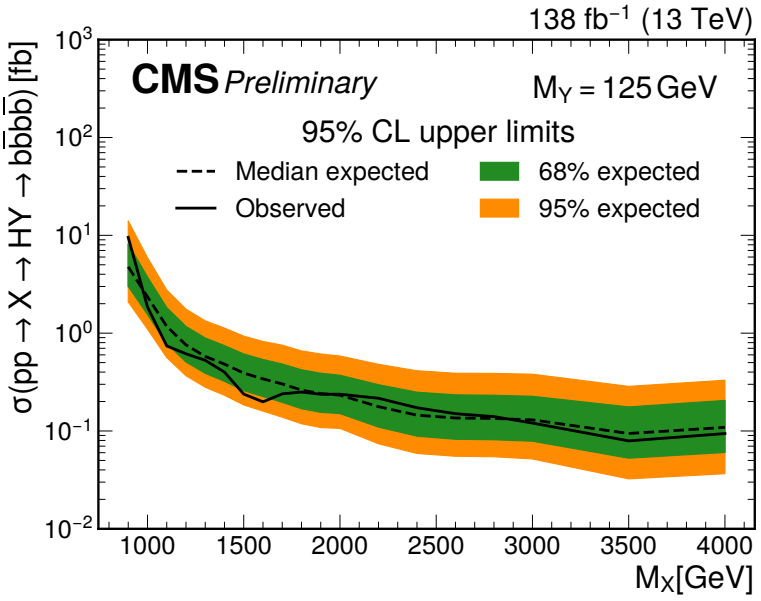



Figure 5.8: The 95% confidence level expected and observed upper limits on $\sigma(\mathrm{pp} \to \mathrm{X} \to \mathrm{YH} \to \mathrm{b\bar{b}b\bar{b}})$ for different values of $M_\mathrm{X}$ and $M_\mathrm{Y} = 125\,\mathrm{GeV}$.

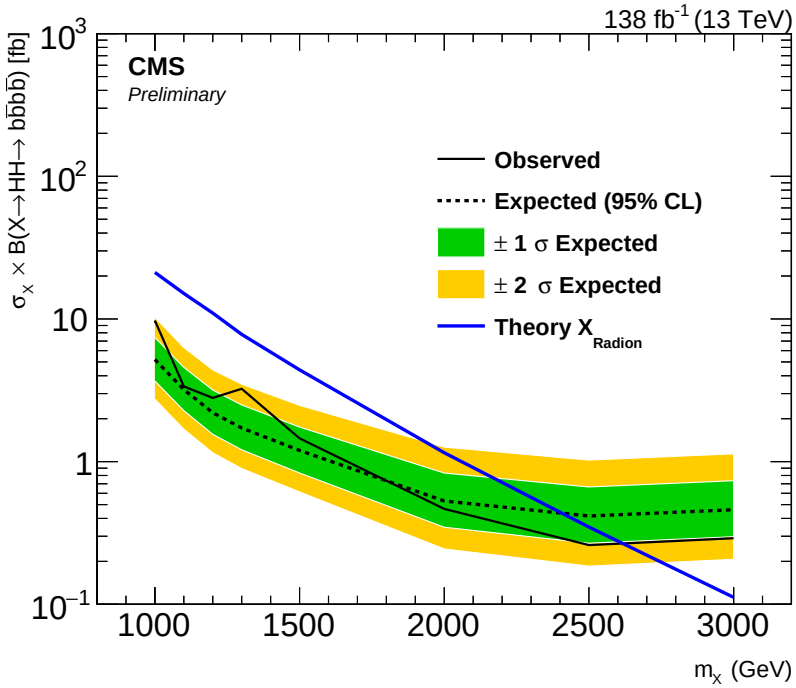



Figure 5.9: Observed and expected limits at 95% CL on the process of spin-0 radion decaying into $\mathrm{H_{125}H_{125}}$ in the final state with four b quarks as measured by the CMS experiment [31]. The limits are given as a function of the resonance mass. Figure taken from Ref. [31].

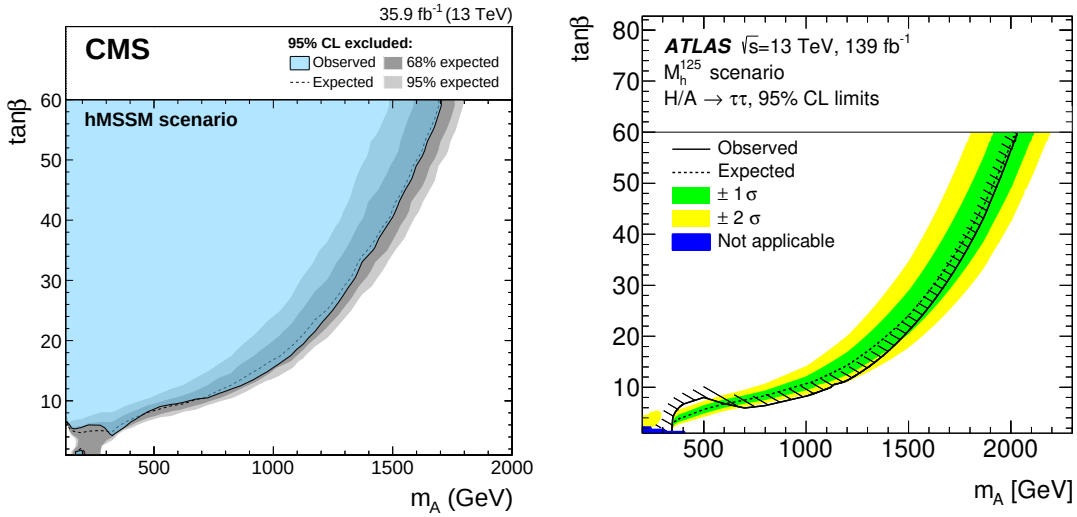



Figure 5.10: Expected and observed 95% CL exclusion contours for a pseudoscalar resonance decaying into two $\tau$ leptons as measured by the CMS (left) and ATLAS (right) experiments. The limits are given as a function of the resonance mass and $\tan\beta$, a parameter of the MSSM theory. Figure taken from Refs. [18] and [19].

# Chapter 6

# Calibration of the ParticleNet efficiency in data using $Z \to b\bar{b}$ decays

Tagging algorithms such as ParticleNet are trained on simulated events. Due to potential differences in simulation and data, their performance (efficiency) needs to be verified in data. Measuring the efficiency for a tagger targeting jets originating from a decay of H into $b\bar{b}$ would require isolating a sample of jets originating from H decays in the data. The low cross section of H production leads to unfavourable signal-to-background ratios making this task very difficult. Calibration measurements in cases like these therefore use jets that have similar characteristics as the jets which the algorithm is trying to select, called "proxy-jets". One example of such a method is the efficiency measurement of the ParticleNet tagger performed on jets originating from a fragmentation of a gluon into a $b\bar{b}$ pair, described in Ref. [104]. It relies on applying a BDT selection on the jets, which makes the proxy jet distribution of the ParticleNet scores similar to the distribution of the target jets, i.e. those originating from a massive particle decaying into a $b\bar{b}$ pair. The uncertainty in the BDT cut that defines the proxy jets is included as a systematic uncertainty in the measurement and is supposed to cover any efficiency differences between the proxy and target jets.

An alternative method, presented in this chapter, uses Lorentz-boosted $Z \to b\bar{b}$ decays reconstructed as single AK8 jets to perform the efficiency measurement. Since Z is a massive particle, the jets directly correspond to signal jets on which the mass decorrelated

ParticleNet tagger was trained. Therefore, no special selection needs to be applied to select proxy jets as is done in the method with jets originating from gluon splitting. However, a measurement based on jets from Z decays comes with less statistics, compared to gluon splitting jets, and is affected by a sizeable QCD multijet background. The method therefore reqiores to identify the Z peak on the non-resonant hadronic background. The event selection employed for this method is described in Sec. 6.1, followed by the derivation of the NLO corrections applied to simulated V + jets samples in Sec. 6.2. The method for estimating the background and extracting the data-to-MC tagging efficiency scale factor (SF) is given in Sec. 6.3. The results are shown in Sec. 6.4.

## 6.1 Event selection

The online event selection uses trigger algorithms that place requirements on the jet transverse momentum $p_T$, the jet groomed mass or b-tagging, or the scalar sum of jet transverse energies $H_{\mathrm{T}}$. The HLT algorithms used in the hadronic category of the $X \to YH \to b\bar{b}b\bar{b}$ analysis, given in Sec. 2.2.5, are also used for the $Z \to b\bar{b}$ SF measurement. The HLT requirement is not applied to simulated events which are instead reweighted by the measured trigger efficiency in the data.

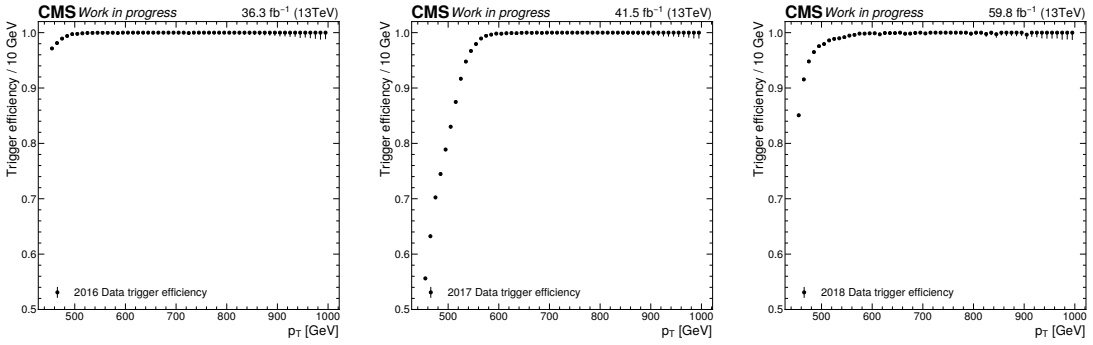



Figure 6.1: The trigger efficiency, as a function of the $p_T$ of the leading jet in the 2016 (left), 2017 (middle) and 2018 (right) data.

The hadronic category trigger efficiency is measured in data using the baseline trigger HLT_PFJet260. Events passing the baseline trigger are further required to pass offline selection criteria. The trigger efficiencies as a function of the leading AK8 jet's $p_T$ are shown for the three years in Fig. 6.1.

The offline selection criteria are:

- Leading AK8 jet $p_T > 450\,\text{GeV}$
- Subleading AK8 jet $p_T > 200\,\text{GeV}$
- $N_e = 0$ and $N_\mu = 0$
- No b-tagged AK4 jets satisfying: $p_T > 30\,\text{GeV}$, $\Delta R\,(\texttt{AK4,leading AK8jet}) > 0.8$

The criteria for selecting electrons and muons for the lepton veto, as well as the criterion for b-tagged AK4 jets are identical to those in the $X \to YH \to b\bar{b}b\bar{b}$ analysis outlined in Sec. 4.3.2 and Sec. 4.3.3, respectively. The latter veto was introduced to suppress the $t\bar{t}$ events.

The mass decorrelated ParticleNet score of the leading jet is used to separate events into passing and failing regions, depending on whether the score is greater or lower than the considered working point, respectively. Two working points, 0.94 (loose) and 0.98 (tight), are used.

The comparison between data and simulation of the events passing selection is shown in the passing and failing regions for the three years in Figs. 6.2 and 6.3.

## 6.2 NLO corrections on the V + jets samples

Since this method measures the Z peak on a non-resonant hadron-rich background, it is necessary to obtain the best theoretical prediction of the Z peak shape and size as it directly influences the results. The production of W and Z with up to 4 additional jets is simulated at LO in different $H_\text{T}$ bins, providing high statistics at large $p_T$. To get a more accurate theory prediction, correction factors are applied to the V + jets samples to match the generator-level $p_T$ distributions with those predicted by NLO in QCD. Further corrections are applied to incorporate the additional changes to cross section due to NLO electroweak effects (EWK).

W and Z production with up to two additional jets, simulated at NLO in QCD using the MadGraph5_aMC@NLO [68] event generator, is used to derive the NLO QCD

correction factors. The ratio of the NLO and LO generator level $p_T$ distributions of the W and Z is fitted to a third order polynomial in $\ln(p_T)$. The $p_T$ distributions at LO, before and after applying the correction, and the distributions at NLO in QCD are shown in Fig. 6.4

The NLO EWK effects are not accounted for in any existing simulated CMS samples. These corrections are therefore taken directly from Ref. [120]. The EWK corrections are also given as a function of generator level $p_T$ of the vector boson. The uncertainties on both the QCD and EWK NLO corrections are also taken from Ref. [120]. The individual QCD and EWK NLO corrections, together with the related uncertainties are shown in Fig. 6.5. The considered uncertainties are labelled in Ref. [120] as: `d1K_NLO, d2K_NLO, d1kappa_EW, d2kappa_EW, d3kappa_EW`, where the last two are uncorrelated across W and Z processes. The dominating uncertainty, as can be seen in Fig. 6.5 is the `d1K_NLO`, which accounts for the uncertainty associated with the renormalization and factorization scales ($\mu_R$, $\mu_F$). The uncertainties are discussed in detail in Ref. [120].

By applying the described NLO corrections to LO samples, we benefit from the necessary high statistics of the LO samples, while the cross section and one of the most important dependencies ($p_T$) are matched to higher order precision.

## 6.3 Fitting model

The fitting of the SM processes to the data is performed in passing and failing regions simultaneously, in a 2D plane defined by the $M_{SD}$-$p_T$ of the leading jet. The $M_{SD}$ axis is split into 5 GeV wide bins in a 50–150 GeV range. The $p_T$ axis is split into three $p_T$ bins determined by [450,500,600,2000] GeV edges, yielding roughly similar yields in each $p_T$ bin.

From the comparison of different process yields in simulation, which can be seen in Figs. 6.2 and 6.3, it is evident that the only background processes with significant contribution in the passing region are the QCD multijet production and the production of W in conjunction with jets. The other considered backgrounds (single top and tt̄ production) are excluded from the background modelling.

The V + jets processes are split into several categories based on the generator-level particle content of the leading AK8 jet. The only considerable component of the W + jets background is the W → cs and is taken from simulation with the NLO corrections applied.

The QCD multijet background is modelled with a data-driven method in order to obtain more accurate prediction. The method is based on the ratio of multijet event distributions in the $M_{SD}$–$p_T$ plane between ParticleNet passing and failing regions,
$R_{P/F}(M_{SD}, p_T)$.

$$n_{Pass,QCD}(i) = n_{Fail,QCD}(i)\, R_{P/F}(M_{SD}, p_T) \tag{6.1}$$

The failing region is overwhelmingly dominated by QCD events, so it can be directly estimated from data by subtracting simulated predictions from other processes. The $R_{P/F}$ is modelled as a two-dimensional polynomial. The parameters of the polynomial are determined during the fit. The $R_{P/F}$ for the tight working point is modelled as a polynomial of order (n,m) defined as:

$$R^T_{P/F} = k_0 \left(1 + \sum_{i=1}^{m} k_i M^i_{SD}\right) \left(1 + \sum_{j=1}^{n} k_j p^j_T\right), \tag{6.2}$$

where $k_i$ are the $n + m + 1$ parameters of the polynomial. The order of the polynomial which describes the data sufficiently well is determined with a Fisher's F-test. Order ($n = 1$,$m = 2$) is selected in all three years.

The simple $R_{P/F}$ form used for the tight working point does not model the QCD background sufficiently well for the loose working point. It is therefore modelled slightly differently. A polynomial of order $o$ is defined as:

$$R^L_{P/F} = \sum_{m,n=0}^{m+n \leq o \,\wedge\, n<3} k_{m,n} M^m_{SD} p^n_T, \tag{6.3}$$

where $k_{m,n}$ are the parameters of the polynomial. The exponent in the $p_T$ is kept below three since there are only three $p_T$ bins in which the fit is performed. The F-test selected polynomials of order $o = 2$, 3 and 5 in 2016, 2017 and 2018, respectively.

The Z + jets process is determined from simulation with NLO corrections applied. The two significant Z + jets categories in the passing region are the $b\bar{b}$ and $c\bar{c}$, with the former comprising $\sim 85\%$ of the Z + jets yield. The two categories are merged together in the fit because they peak at the same position and their shapes cannot be distinguished. Other categories are not considered.

The parameters of interest during the fit are the two SFs for the Z + jets and W + jets processes, corresponding to the data-to-MC correction on the efficiency of the jets coming from the $Z \to b\bar{b}$ decay and the data-to-MC mistag rate correction for the jets originating from W decay. The SFs allow the migration of V + jets events between the passing and failing regions. The two SFs are modelled without $p_T$ dependence. The contribution of the W + jets process is low in the passing region which makes the sensitivity for its SF low. It is therefore not reported in the results.

The following sources of systematic uncertainties are also included in the fit.

- **NLO correction uncertainties**: the uncertainties on the NLO corrections applied to the V + jets are taken from Ref. [120]. They account for the renormalization and factorization scale variations and shape uncertainties of the NLO QCD corrections. For the NLO EWK corrections, uncertainties account for higher-order Sudakov logarithms, hard NNLO emission effects and the limitations of Sudakov approximation. Details are given in Sec. 4 of Ref. [120]. The combined effect of the uncertainties results in approximately 10% yield uncertainty.

- **Jet mass scale**: the mass scale shift of 2% is applied as the uncertainty on the jet mass separately in each of the three $p_T$ bins.

- **Jet mass resolution**: the nominal resolution from simulation is taken as the downward uncertainty while a 20% resolution smear is applied as the upward uncertainty. The jet mass resolution uncertainty is separately applied to each of the three $p_T$ bins.

- **Jet energy scale and resolution**: the jet energy scale and resolution corrections are applied to match the simulation to the data [87]. The recommended uncertainties are used to make shape based uncertainties.

- **Luminosity**: an uncertainty of 1.2%, 2.3% and 2.5% is applied in 2016, 2017 and 2018 respectively. [121]

- **Trigger timing correction**: a 2% normalization uncertainty is applied to the V + jets simulation to account for the gradual shift in the timing of the inputs of the ECAL hardware level trigger during the 2016 and 2017 data taking.

- **Trigger efficiency uncertainty**: measured statistical uncertainty, increased by 1%, is applied. This is larger than just the statistical uncertainty of the measurement in order to cover the jet energy scale uncertainties in the trigger selections.

- **PDF**: the PDF uncertainties are implemented using the PDF4LHC [113] procedure with the NNPDF3.1 PDF sets.

- **Pileup**: a pileup correction is applied to simulated events based on the true number of interactions in the event. A shape uncertainty is generated by varying the inelastic pp cross section at 13 TeV (69.2 mb) by 4.6% [112].

A maximum likelihood fit is performed to extract the two SFs. The results are shown in Sec. 6.4

## 6.4 Postfit results

The postfit $M_{SD}$ distributions for each year and the two working points are shown in Figs. 6.6–6.9. The postfit $R_{P/F}$ values for each year and the two working points are shown in Fig. 6.10. Table 6.1 summarises the scale factors derived using the $Z \to b\bar{b}$ strategy, for the data-taking years 2016, 2017, and 2018 separately, for two ParticleNet bb-tagging working points. The results are also summarized in Fig. 6.12 and the results for the tight working point are compared to the ones obtained using the $g \to b\bar{b}$ proxy jet method in Fig. 6.12. The results of the two methods are in good agreement and this measurement could be considered a validation of the $g \to b\bar{b}$ proxy jet method. The comparison is not made for the loose working point since the $g \to b\bar{b}$ uses exclusively loose tagged jets (tight jets excluded), while the loose category in the $Z \to b\bar{b}$ method by

definition also includes jets tagged as tight in order to help with the signal purity in the loose measurement.

Table 6.1: Summary of the bb-tagging SFs derived using the $Z \to b\bar{b}$ method for the three data-taking years and two ParticleNet bb-tagging categories ($> 0.94$, $> 0.98$).

| WP | 2016 | 2017 | 2018 |
|---|---|---|---|
| 0.94 | $0.78^{+0.20}_{-0.16}$ | $1.15^{+0.30}_{-0.22}$ | $0.88^{+0.29}_{-0.21}$ |
| 0.98 | $0.90^{+0.15}_{-0.12}$ | $1.04^{+0.17}_{-0.15}$ | $0.94^{+0.19}_{-0.16}$ |

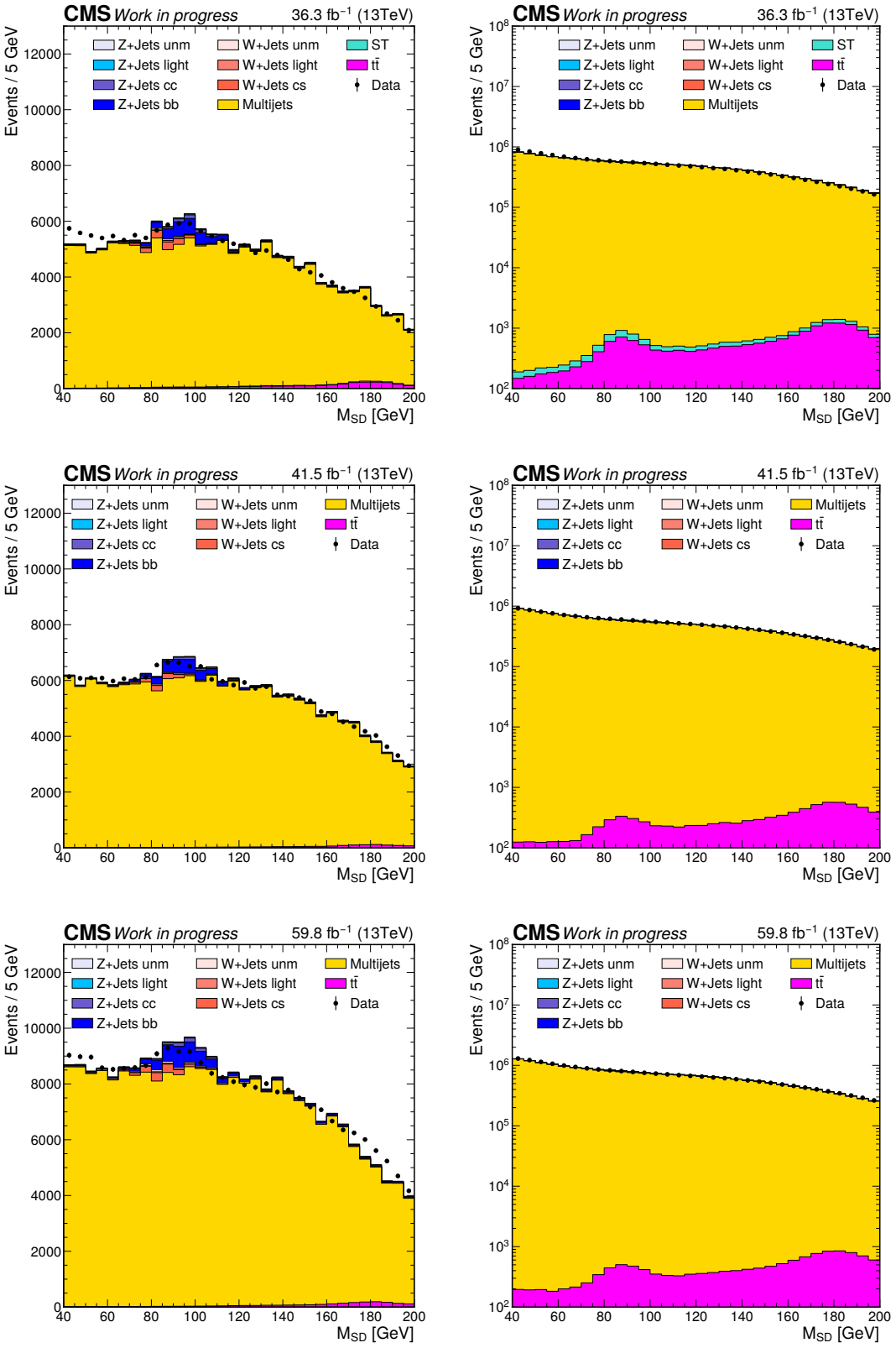



Figure 6.2: Comparison of the softdrop mass distribution of the leading jet for events passing selection in the pass (left) and fail (right) category for the loose working point in simulation and data. Multijet yields are scaled to match the overall simulation yield with the data. Distributions are shown for 2016 (upper), 2017 (middle) and 2018 (lower).

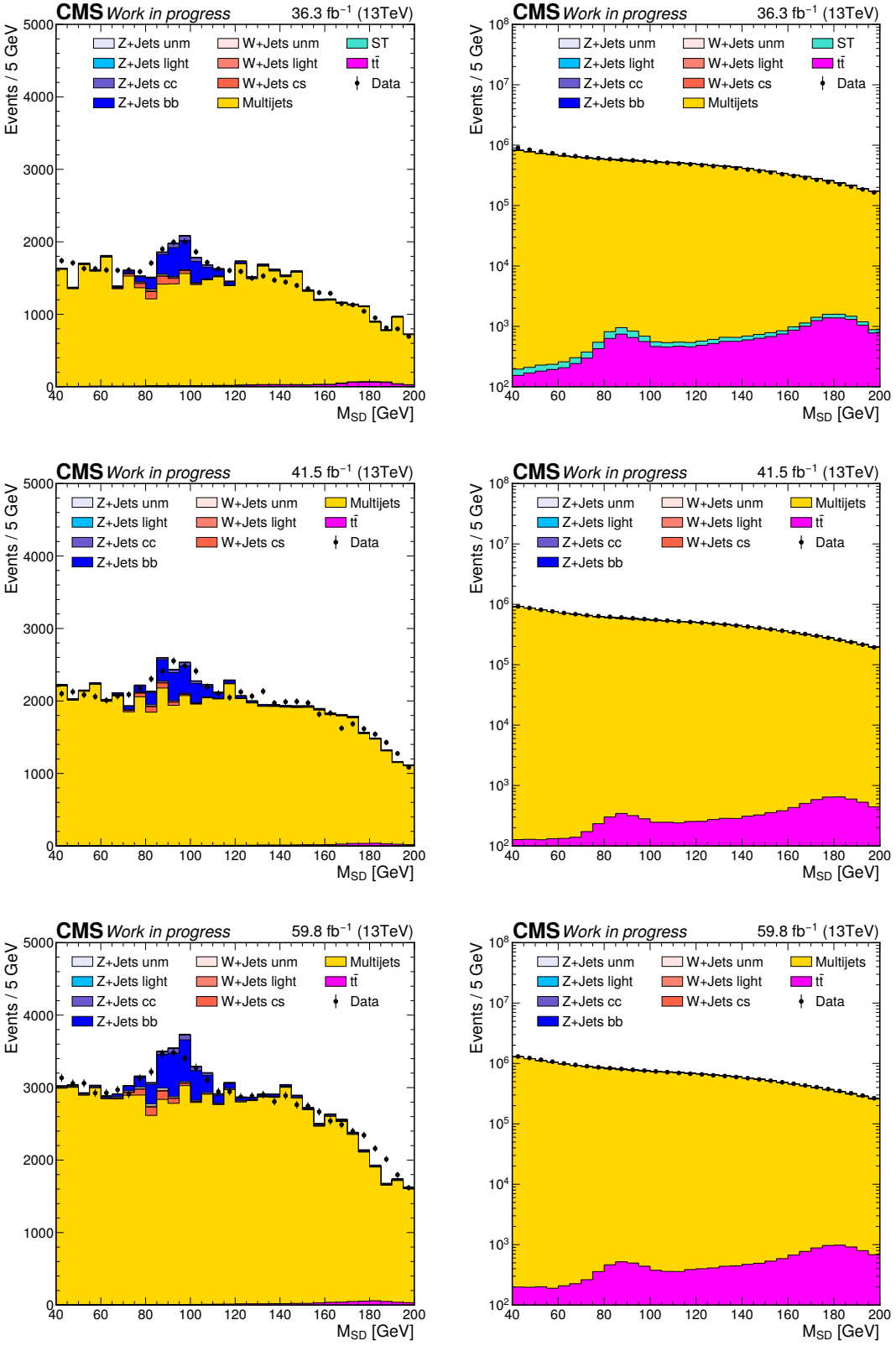



Figure 6.3: Comparison of the softdrop mass distribution of the leading jet for events passing selection in the pass (left) and fail (right) category for the tight working point in simulation and data. Multijet yields are scaled to match the overall simulation yield with the data. Distributions are shown for 2016 (upper), 2017 (middle) and 2018 (lower).

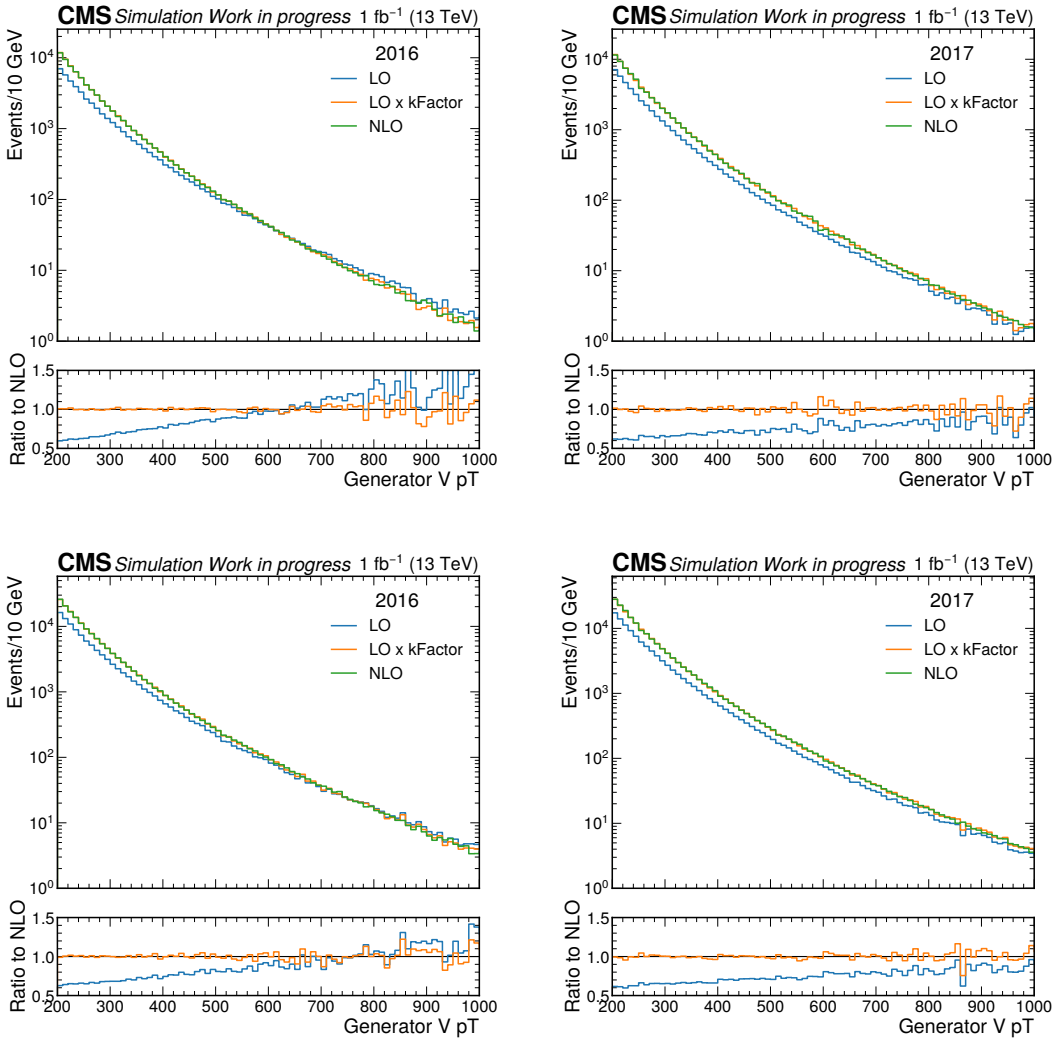



Figure 6.4: Generator level distribution of the Z (upper) and W (lower) at LO, with and without the derived correction applied, and NLO in QCD. The corrections are different for 2016 (left) and 2017/2018 (right) due to differences in the PDFs and PYTHIA tunes.

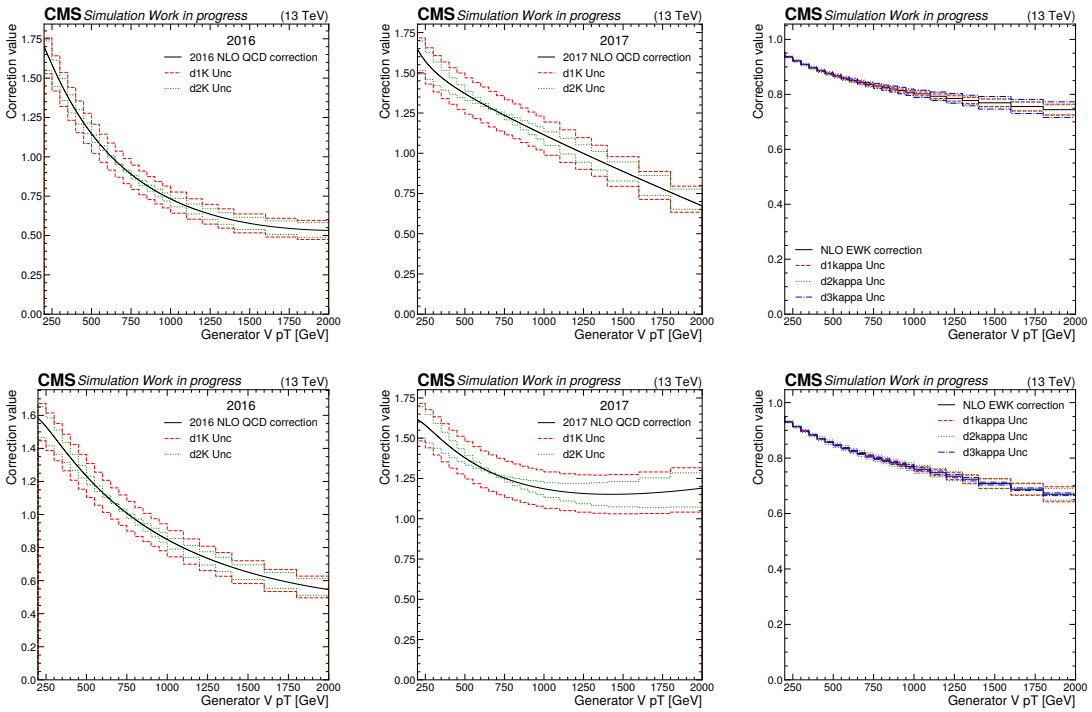



Figure 6.5: QCD (left, middle) and EWK (right) NLO corrections with corresponding uncertainties for Z (upper) and W (lower) jets.

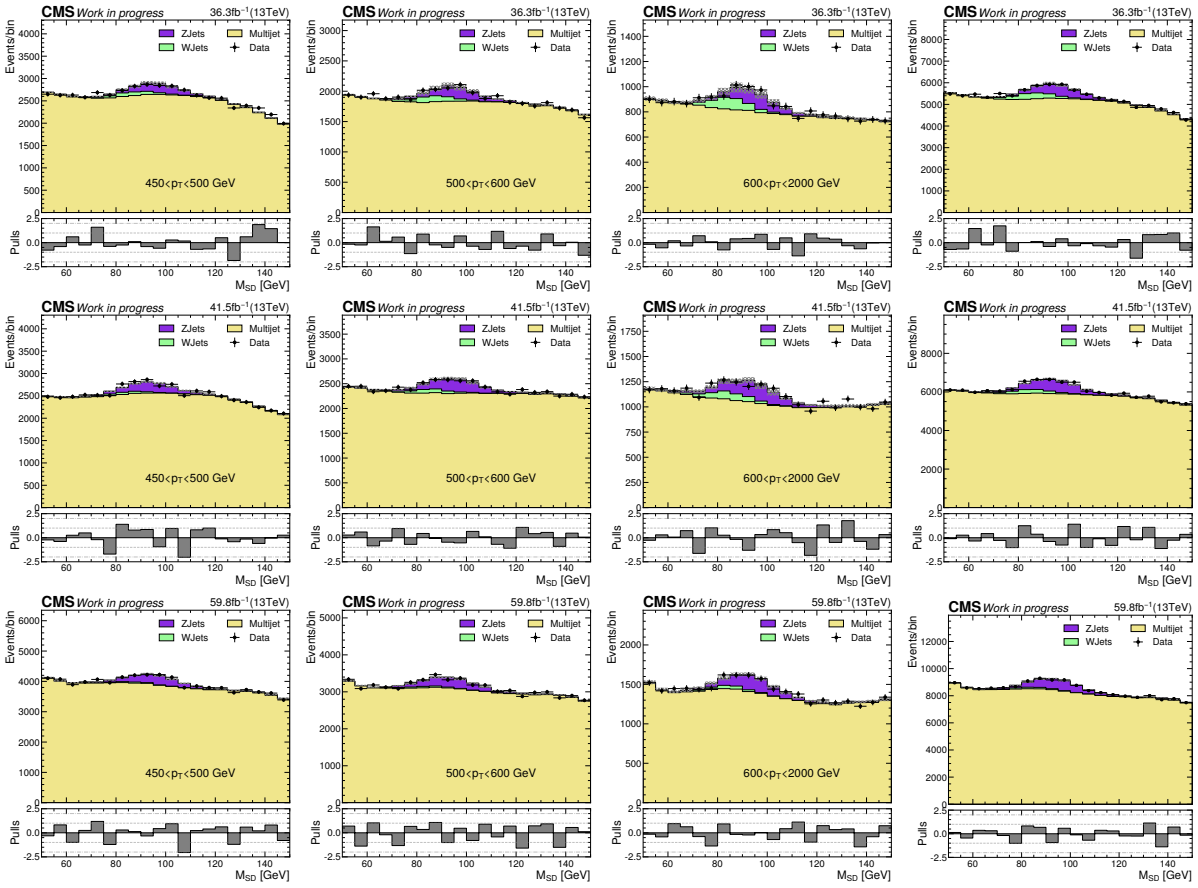



Figure 6.6: Postfit plots for the "pass" region with the 2016 (upper), 2017 (middle) and 2018 (lower) data, in the loose bb-tagging category. From left to right are the leading jet's $M_{SD}$ distributions in $p_T$ categories $[450, 500)$, $[500, 600)$, $[600, 2000)$, $[450, 2000)$ GeV.

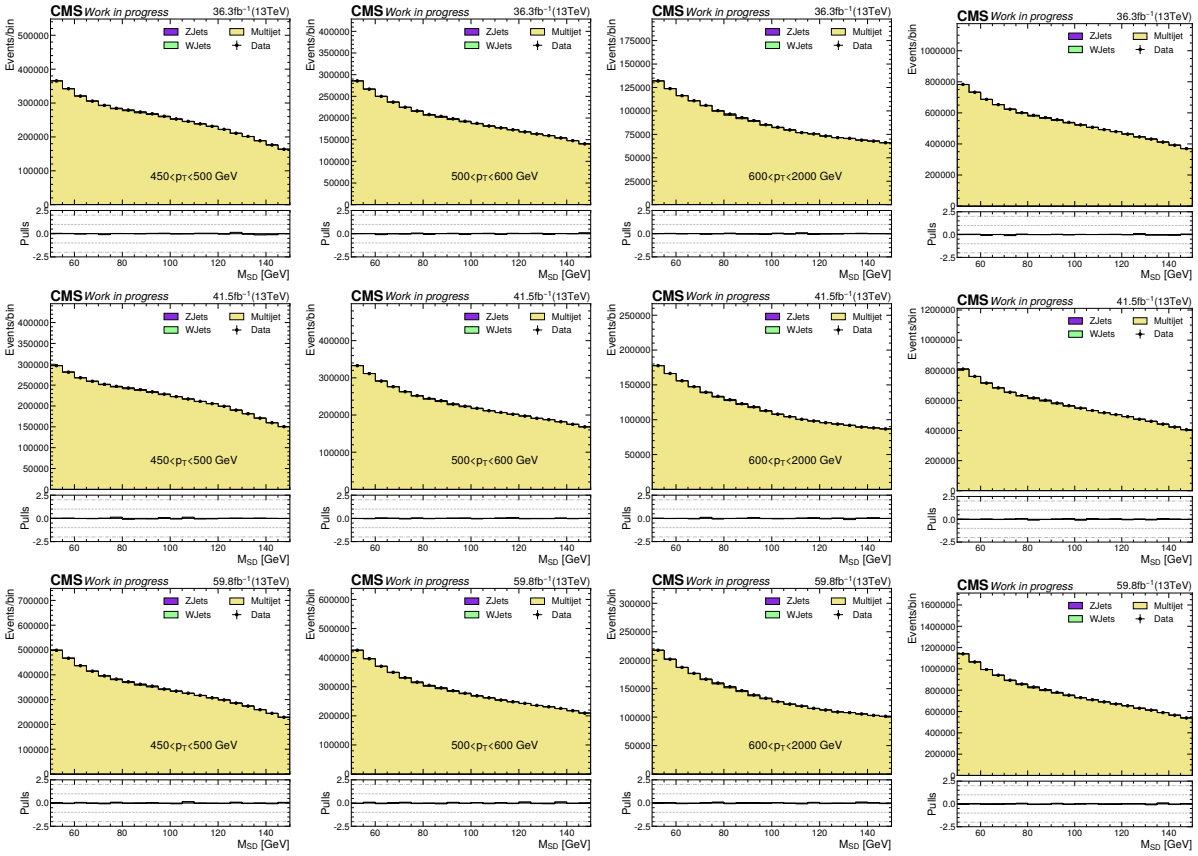



Figure 6.7: Postfit plots for the "fail" region with the 2016 (upper), 2017 (middle) and 2018 (lower) data, in the loose bb-tagging category. From left to right are the leading jet's $M_{SD}$ distributions in $p_T$ categories $[450, 500)$, $[500, 600)$, $[600, 2000)$, $[450, 2000)$ GeV.

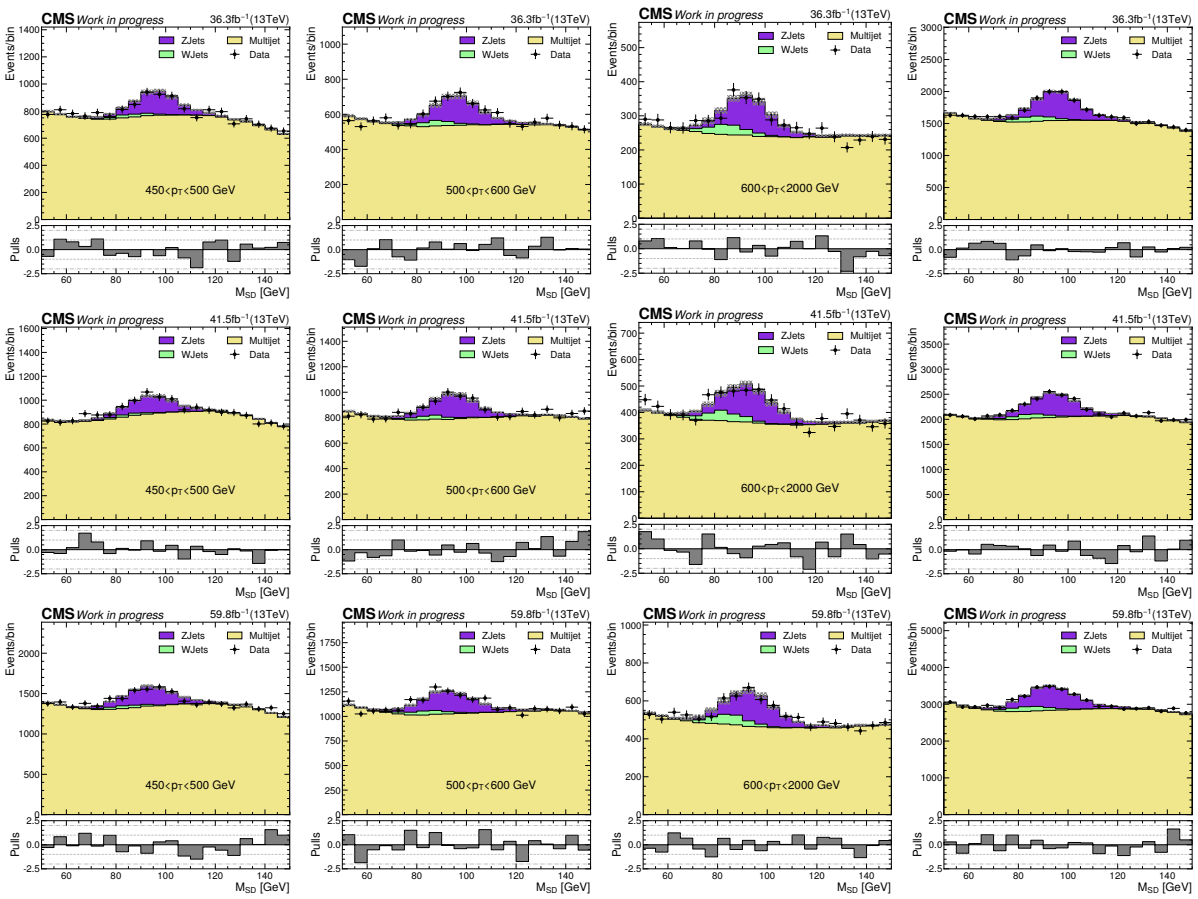



Figure 6.8: Postfit plots for the "pass" region with the 2016 (upper), 2017 (middle) and 2018 (lower) data, in the tight bb-tagging category. From left to right are the leading jet's $M_{SD}$ distributions in $p_T$ categories $[450, 500)$, $[500, 600)$, $[600, 2000)$, $[450, 2000)$ GeV.

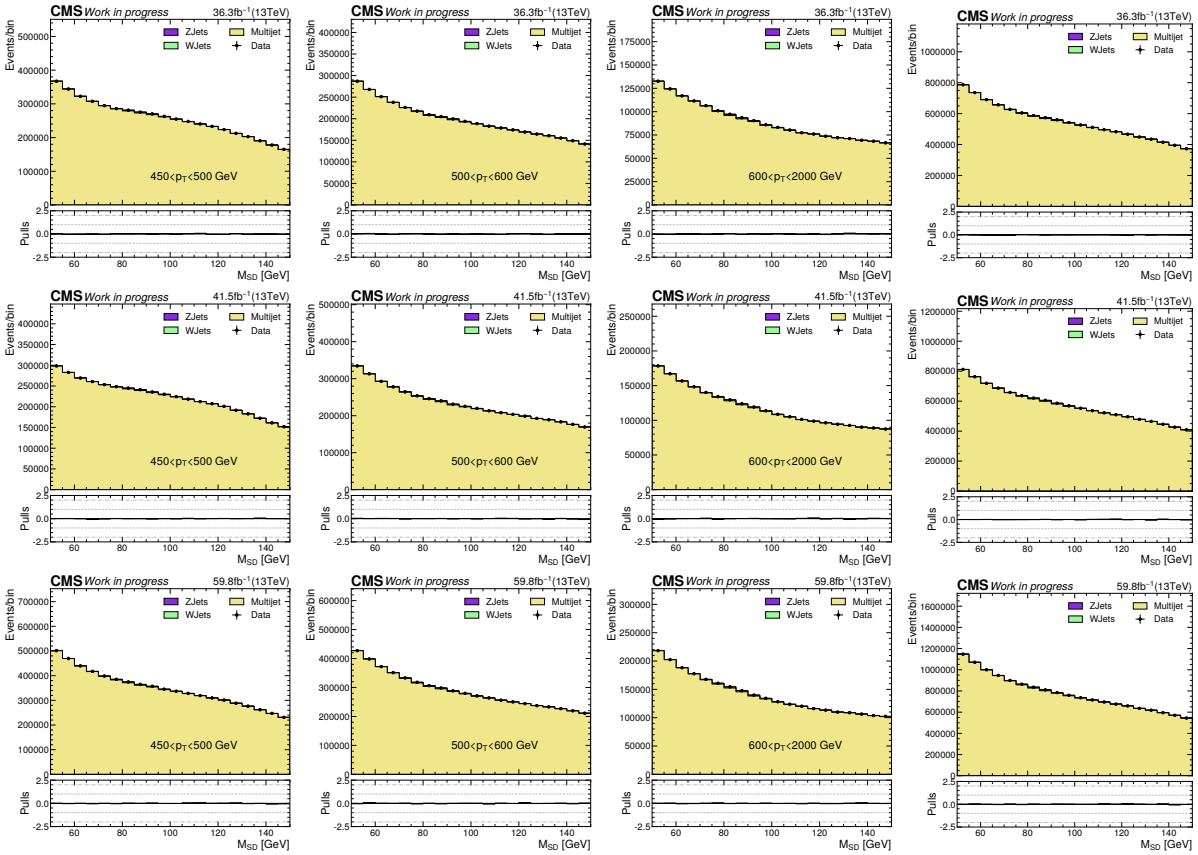



Figure 6.9: Postfit plots for the "fail" region with the 2016 (upper), 2017 (middle) and 2018 (lower) data, in the tight bb-tagging category. From left to right are the leading jet's $M_{SD}$ distributions in $p_T$ categories $[450, 500)$, $[500, 600)$, $[600, 2000)$, $[450, 2000)$ GeV.

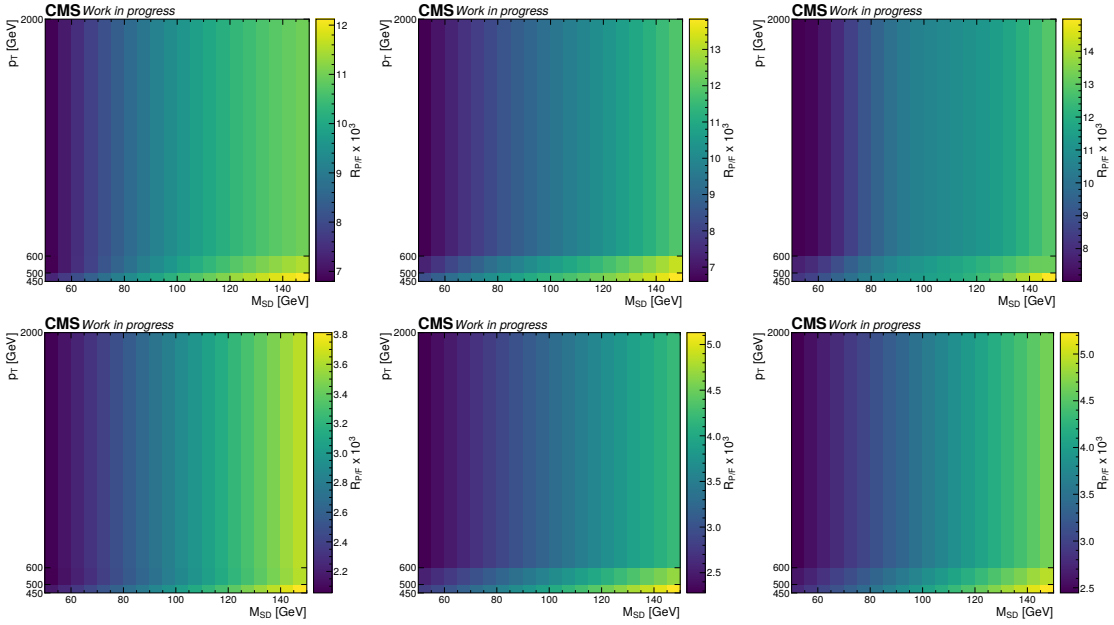



Figure 6.10: Postfit $R_{P/F}$ for the loose (upper) and tight (lower) categories for the 2016 (left), 2017 (middle) and 2018 (right) data.

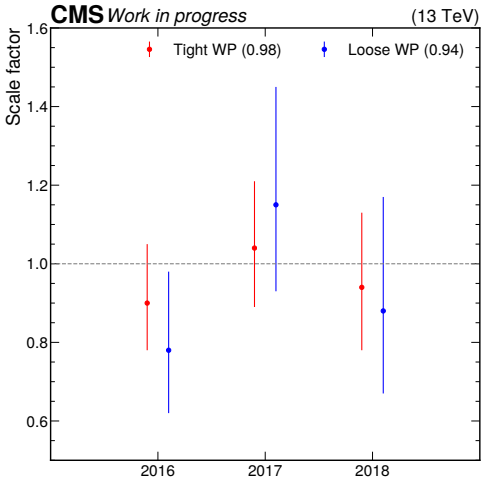



Figure 6.11: Summary of the bb-tagging SFs derived using the $Z \to b\bar{b}$ method for the three data-taking years and two ParticleNet bb-tagging working points loose ($> 0.94$) and tight($> 0.98$).

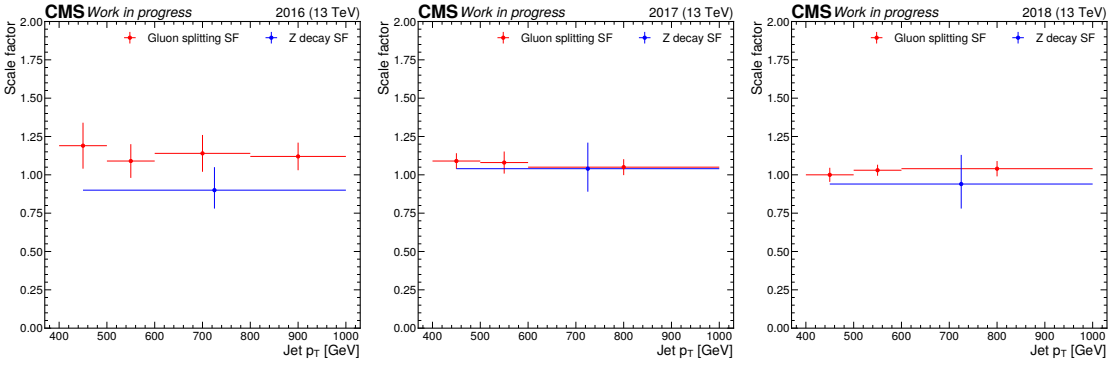



Figure 6.12: Comparison of the bb-tagging SFs derived using the $Z \to b\bar{b}$ and $g \to b\bar{b}$ methods for the three data-taking years in the tight ($> 0.98$) ParticleNet bb-tagging category.

# Chapter 7

# Summary

The search for the cascade decay of a heavy scalar X into another scalar Y and a Higgs boson H, both decaying into a pair of b quarks is presented. The decays of X into two scalars of uneven masses are motivated by several BSM theories, but are still largely unexplored at the LHC. The search is performed in mass ranges of 0.9–4 TeV for X and 60–600 GeV for Y, where both Y and H are Lorentz-boosted and reconstructed as a single large-area jet each. Measurements are performed using data from proton-proton collisions at a centre-of-mass energy of 13 TeV collected by the CMS experiment at the LHC. The data used was collected in 2016–2018 and corresponds to an integrated luminosity of 138 $\text{fb}^{-1}$.

The analysis uses two mutually exclusive categories of events, hadronic and semileptonic. The event selection for the hadronic category is designed to select signal-like events: two back-to-back large area jets in the central part of the detector with large $p_T$. The key variables for the search are the mass of a Y-candidate jet, $M_{\text{J}}^{\text{Y}}$, and the invariant mass of the two jets, $M_{\text{JJ}}$, corresponding to $M_{\text{Y}}$ and $M_{\text{X}}$, respectively. A graph convolutional neural network based algorithm, ParticleNet, is used to identify the boosted $\text{H} \to \text{b}\bar{\text{b}}$ or $\text{Y} \to \text{b}\bar{\text{b}}$ decays against a background of other jets. The two main backgrounds for the signal in the hadronic category are the QCD multijet production and the $\text{t}\bar{\text{t}}$ production. The multijet background is estimated using data-driven techniques based on pass-to-fail ratios. The novelty of this search is the application of the pass-to-fail technique in a 2D plane ($M_{\text{J}}^{\text{Y}}$–$M_{\text{JJ}}$) with unknown signal masses in either of the variables. Dedicated tests

were developed and performed to ensure the background model stability before unblinding the data in the signal regions. The other background, the $t\bar{t}$ production, is modelled using simulation with corrections simultaneously measured in data using a control region enriched in semileptonic $t\bar{t}$ decays.

The event selection of the semileptonic category is based on the semileptonic decay of $t\bar{t}$. A set of requirements is imposed to select leptonic decay of the top in order to measure the properties of the jets originating from the hadronically decaying t quarks on the opposite side of the detector. The mass distributions of these jets are used to obtain data-to-simulation corrections factors on the $t\bar{t}$ production.

The signal is modelled using simulation with the NMSSM model. However, the kinematic parameters are generic enough to interpret the results under other BSM scenarios. The signal search is performed for 260 different ($M_X$, $M_Y$) hypotheses. The highest observed local significance of 3.1 $\sigma$ is for the signal with $M_X = 1.6$ TeV and $M_Y = 90$ GeV, which becomes 0.7 $\sigma$ after accounting for the look-elsewhere effect. The upper limits at 95% CL on the signal cross section are computed. The observed limits range from 0.1 fb to 120 fb as a function of $M_X$ and $M_Y$. The cross section limits are compared with the maximally allowed cross sections in the NMSSM and TRSM models. Areas where the maximally allowed cross sections are above the observed limits are shown and lead to the exclusion of parameter space in the two models. For NMSSM, an exclusion area with maximum $M_X$ range of 1000–1150 GeV and $M_Y$ range of 101–145 GeV is observed. For TRSM, the observed exclusion range spans $950 < M_X < 1330$ GeV and $110 < M_Y < 132$ GeV. The cross section limits are also compared with the results of related searches and it is shown that the obtained results improve the current exclusion limits.

A calibration measurement of the ParticleNet tagger for two working points, loose and tight, in the three data-taking years is also presented. The method relies on the measurement of the Z peak on a non-resonant multijet background. The event selection targets boosted $Z \to b\bar{b}$ decays where the products are contained within a single AK8 jet. The dominant background is the QCD multijet production which is estimated using a data-driven method based on a pass-to-fail ratio. The second relevant background comes from the W + jets production, which is modelled from simulation. Both V + jets processes are simulated at LO with NLO corrections in both QCD and EWK applied to them as a

function of the generator-level $p_T$ of the W or the Z bosons. The measurement of the tagging efficiency for the tight working point is compared to the results obtained with another method, based on selecting $g \to b\bar{b}$ with "signal-like" characteristics. The two measurements are in agreement, validating the $g \to b\bar{b}$ method and also demonstrating that it is possible to perform the calibration measurement using Z decays.

# Bibliography


[1] B. Odom et al. "New Measurement of the Electron Magnetic Moment Using a One-Electron Quantum Cyclotron". In: *Phys. Rev. Lett.* 97 (3 2006), p. 030801. DOI: 10.1103/PhysRevLett.97.030801. URL: https://link.aps.org/doi/10.1103/PhysRevLett.97.030801 (cit. on p. 3).

[2] Peter W. Higgs. "Broken symmetries, massless particles and gauge fields". In: *Phys. Lett.* 12 (1964), p. 132. DOI: 10.1016/0031-9163(64)91136-9 (cit. on pp. 4, 13).

[3] F. Englert and R. Brout. "Broken Symmetry and the Mass of Gauge Vector Mesons". In: *Phys. Rev. Lett.* 13 (1964). Ed. by J. C. Taylor, p. 321. DOI: 10.1103/PhysRevLett.13.321 (cit. on pp. 4, 13).

[4] Peter W. Higgs. "Broken Symmetries and the Masses of Gauge Bosons". In: *Phys. Rev. Lett.* 13 (1964). Ed. by J. C. Taylor, p. 508. DOI: 10.1103/PhysRevLett.13.508 (cit. on pp. 4, 13).

[5] Wikimedia commons. *Standard Model of Elementary Particles*. https://commons.wikimedia.org/wiki/File:Standard_Model_of_Elementary_Particles.svg (cit. on p. 5).

[6] G. S. Guralnik, C. R. Hagen, and T. W. B. Kibble. "Global Conservation Laws and Massless Particles". In: *Phys. Rev. Lett.* 13 (1964). Ed. by J. C. Taylor, p. 585. DOI: 10.1103/PhysRevLett.13.585 (cit. on p. 13).

[7] Peter W. Higgs. "Spontaneous Symmetry Breakdown without Massless Bosons". In: *Phys. Rev.* 145 (1966), p. 1156. DOI: 10.1103/PhysRev.145.1156 (cit. on p. 13).

[8] T. W. B. Kibble. "Symmetry breaking in Non-Abelian gauge theories". In: *Phys. Rev.* 155 (1967). Ed. by J. C. Taylor, p. 1554. DOI: 10.1103/PhysRev.155.1554 (cit. on p. 13).

[9] S. Dawson. *Introduction to Electroweak Symmetry Breaking*. 1999. arXiv: hep-ph/9901280 [hep-ph] (cit. on p. 13).

[10] "Planck 2018 results. Overview and the cosmological legacy of Planck". In: *Astronomy & Astrophysics* 641 (2020), A1. DOI: 10.1051/0004-6361/201833880. URL: https://doi.org/10.1051%2F0004-6361%2F201833880 (cit. on pp. 13, 14).

[11] Steven Weinberg. "The cosmological constant problem". In: *Rev. Mod. Phys.* 61 (1 1989), pp. 1–23. DOI: 10.1103/RevModPhys.61.1. URL: https://link.aps.org/doi/10.1103/RevModPhys.61.1 (cit. on p. 14).

[12] CMS collaboration. "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC". In: *Phys. Let. B* 716.1 (2012), pp. 30–61. DOI: 10.1016/j.physletb.2012.08.021. URL: https://doi.org/10.1016/j.physletb.2012.08.021 (cit. on pp. 15, 19).

[13] ATLAS Collaboration. "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC". In: *Phys. Let. B* 716.1 (2012), pp. 1–29. DOI: 10.1016/j.physletb.2012.08.020. URL: https://doi.org/10.1016/j.physletb.2012.08.020 (cit. on pp. 15, 19).

[14] John F. Gunion and Howard E. Haber. "CP-conserving two-Higgs-doublet model: The approach to the decoupling limit". In: *Physical Review D* 67.7 (2003). ISSN: 1089-4918. DOI: 10.1103/physrevd.67.075019. URL: http://dx.doi.org/10.1103/PhysRevD.67.075019 (cit. on p. 15).

[15] A. Arhrib et al. "Two-Higgs-doublet type-II and -III models and $t \to ch$ at the LHC". In: *The European Physical Journal C* 76.6 (2016). DOI: 10.1140/epjc/s10052-016-4167-9. URL: https://doi.org/10.1140/epjc/s10052-016-4167-9 (cit. on p. 15).

[16] Howard E. Haber and Laurel Stephenson Haskins. "Supersymmetric Theory and Models". In: *Anticipating the Next Discoveries in Particle Physics* (2018). DOI: 10.1142/9789813233348_0006. URL: http://dx.doi.org/10.1142/9789813233348_0006 (cit. on p. 16).

[17] Hitoshi Murayama. "Supersymmetry Phenomenology". 2000. arXiv: hep-ph/0002232 [hep-ph] (cit. on p. 16).

[18] CMS Collaboration. "Search for additional neutral MSSM Higgs bosons in the $\tau\tau$ final state in proton-proton collisions at $\sqrt{s} = 13$ TeV". In: *JHEP* 09 (2018), p. 007. DOI: 10.1007/JHEP09(2018)007. arXiv: 1803.06553 [hep-ex] (cit. on pp. 16, 104, 106).

[19] ATLAS Collaboration. "Search for heavy Higgs bosons decaying into two tau leptons with the ATLAS detector using *pp* collisions at $\sqrt{s} = 13$ TeV". In: *Phys. Rev. Lett.* 125 (2020), p. 051801. DOI: 10.1103/PhysRevLett.125.051801. arXiv: 2002.12223 [hep-ex] (cit. on pp. 16, 104, 106).

[20] CMS Collaboration. "Search for charged Higgs bosons in the $H^{\pm} \to \tau^{\pm}\nu_{\tau}$ decay channel in proton-proton collisions at $\sqrt{s} = 13$ TeV". In: *JHEP* 07 (2019), p. 142. DOI: 10.1007/JHEP07(2019)142. arXiv: 1903.04560 [hep-ex] (cit. on p. 16).

[21] ATLAS Collaboration. "Search for charged Higgs bosons decaying via $H^{\pm} \to \tau^{\pm}\nu_{\tau}$ in the $\tau$+jets and $\tau$+lepton final states with 36 fb$^{-1}$ of *pp* collision data recorded at $\sqrt{s} = 13$ TeV with the ATLAS experiment". In: *JHEP* 09 (2018), p. 139. DOI: 10.1007/JHEP09(2018)139. arXiv: 1807.07915 [hep-ex] (cit. on p. 16).

[22] CMS Collaboration. "Search for heavy Higgs bosons decaying to a top quark pair in proton-proton collisions at $\sqrt{s} = 13$ TeV". In: *JHEP* 2004 (2019), p. 171. DOI: 10.1007/JHEP04(2020)171. arXiv: 1908.01115 (cit. on p. 16).

[23] Ulrich Ellwanger, Cyril Hugonie, and Ana M. Teixeira. "The Next-to-Minimal Supersymmetric Standard Model". In: *Phys. Rept.* 496 (2010), p. 1. DOI: 10.1016/j.physrep.2010.07.001. arXiv: 0910.1785 [hep-ph] (cit. on pp. 16, 17).

[24] M. Maniatis. "The Next-to-Minimal Supersymmetric extension of the Standard Model reviewed". In: *International Journal of Modern Physics A* 25 (2010), pp. 3505–3602. ISSN: 1793-656X. DOI: 10.1142/s0217751x10049827. URL: http://dx.doi.org/10.1142/S0217751X10049827 (cit. on p. 16).

[25] Jihn E. Kim and Hans Peter Nilles. "The $\mu$-problem and the Strong CP Problem". In: *Phys. Lett. B* 138 (1984), p. 150. DOI: 10.1016/0370-2693(84)91890-2 (cit. on p. 16).

[26] A. Djouadi et al. "Benchmark scenarios for the NMSSM". In: *JHEP* 07 (2008), p. 002. DOI: 10.1088/1126-6708/2008/07/002. arXiv: 0801.4321 [hep-ph] (cit. on p. 17).

[27] Ulrich Ellwanger and Matias Rodriguez-Vazquez. "Simultaneous search for extra light and heavy Higgs bosons via cascade decays". In: *JHEP* 11 (2017), p. 008. DOI: 10.1007/JHEP11(2017)008. arXiv: 1707.08522 [hep-ph] (cit. on p. 17).

[28] CMS Collaboration. "Search for a heavy Higgs boson decaying into two lighter Higgs bosons in the $\tau\tau$bb final state at 13 TeV". In: *Journal of High Energy Physics* 2021 (2021). DOI: 10.1007/jhep11(2021)057. URL: https://doi.org/10.1007/jhep11(2021)057 (cit. on pp. 17, 66, 103, 104).

[29] Tania Robens, Tim Stefaniak, and Jonas Wittbrodt. "Two-real-scalar-singlet extension of the SM: LHC phenomenology and benchmark scenarios". In: *Eur. Phys. J. C* 80 (2020), p. 151. DOI: 10.1140/epjc/s10052-020-7655-x. arXiv: 1908.08554 [hep-ph] (cit. on p. 17).

[30] ATLAS Collaboration. *Search for resonant pair production of Higgs bosons in the $b\bar{b}b\bar{b}$ final state using pp collisions at $\sqrt{s}$ = 13 TeV with the ATLAS detector*. Tech. rep. Geneva, 2021. URL: http://cds.cern.ch/record/2777861 (cit. on p. 17).

[31] CMS Collaboration. *Search for resonant Higgs boson pair production in four b quark final state using large-area jets in proton-proton collisions at $\sqrt{s} = 13$ TeV*. Tech. rep. Geneva: CERN, 2021. URL: http://cds.cern.ch/record/2777083 (cit. on pp. 17, 78, 84, 103, 105).

[32] O. DeWolfe et al. "Modeling the fifth dimension with scalars and gravity". In: *Phys. Rev. D* 62 (4 2000), p. 046008. DOI: 10.1103/PhysRevD.62.046008. URL: https://link.aps.org/doi/10.1103/PhysRevD.62.046008 (cit. on p. 17).

[33] Walter D. Goldberger and Mark B. Wise. "Modulus Stabilization with Bulk Fields". In: *Phys. Rev. Lett.* 83 (24 1999), pp. 4922–4925. DOI: 10.1103/PhysRevLett.83.4922. URL: https://link.aps.org/doi/10.1103/PhysRevLett.83.4922 (cit. on p. 17).

[34] Csaba Csáki et al. "Cosmology of brane models with radion stabilization". In: *Phys. Rev. D* 62 (4 2000), p. 045015. DOI: 10.1103/PhysRevD.62.045015. URL: https://link.aps.org/doi/10.1103/PhysRevD.62.045015 (cit. on p. 17).

[35] H. Davoudiasl, J. L. Hewett, and T. G. Rizzo. "Phenomenology of the Randall-Sundrum Gauge Hierarchy Model". In: *Phys. Rev. Lett.* 84 (10 2000), pp. 2080–2083. DOI: 10.1103/PhysRevLett.84.2080. URL: https://link.aps.org/doi/10.1103/PhysRevLett.84.2080 (cit. on p. 17).

[36] Kaustubh Agashe et al. "Warped gravitons at the CERN LHC and beyond". In: *Phys. Rev. D* 76 (3 2007), p. 036006. DOI: 10.1103/PhysRevD.76.036006. URL: https://link.aps.org/doi/10.1103/PhysRevD.76.036006 (cit. on p. 17).

[37] ATLAS Collaboration. "Combination of searches for Higgs boson pairs in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector". In: *Physics Letters B* 800 (2020), p. 135103. ISSN: 0370-2693. DOI: https://doi.org/10.1016/j.physletb.2019.135103. URL: https://www.sciencedirect.com/science/article/pii/S0370269319308251 (cit. on p. 18).

[38] CMS Collaboration. "Combination of Searches for Higgs Boson Pair Production in Proton-Proton Collisions at $\sqrt{s} = 13$ TeV". In: *Phys. Rev. Lett.* 122 (12 2019), p. 121803. DOI: 10.1103/PhysRevLett.122.121803. URL: https://link.aps.org/doi/10.1103/PhysRevLett.122.121803 (cit. on p. 18).

[39] CMS collaboration. "Observation of a new boson with mass near 125 GeV in pp collisions at $\sqrt{s}$=7 and 8 TeV". In: *Phys. Let. B* 2013.6 (2013). DOI: 10.1007/jhep06(2013)081. URL: https://doi.org/10.1007/jhep06(2013)081 (cit. on p. 19).

[40] Lyndon Evans and Philip Bryant. "LHC Machine". In: *Journal of Instrumentation* 3.08 (2008), S08001–S08001. DOI: 10.1088/1748-0221/3/08/s08001. URL: https://doi.org/10.1088/1748-0221/3/08/s08001 (cit. on p. 21).

[41] CERN. *Diagram of an LHC dipole magnet.* https://cds.cern.ch/record/40524. 1999 (cit. on p. 24).

[42] Michael Hostettler et al. "Impact of the Crossing Angle on Luminosity Asymmetries at the LHC in 2016 Proton Physics Operation". In: *Proceedings of the 8th Int. Particle Accelerator Conf.* (2017). DOI: 10.18429/JACoW-IPAC2017-TUPVA005. URL: https://cds.cern.ch/record/2289120 (cit. on p. 25).

[43] Vladislav Balagura. "Van der Meer scan luminosity measurement and beam–beam correction". In: *The European Physical Journal C* 81.1 (2021). DOI: 10.1140/epjc/s10052-021-08837-y. URL: https://doi.org/10.1140/epjc/s10052-021-08837-y (cit. on p. 25).

[44] CMS collaboration. *Public Results of CMS Luminosity Information*. 2022. URL: https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults (cit. on p. 26).

[45] CMS Collaboration. "The CMS experiment at the CERN LHC". In: *Journal of Instrumentation* 3.08 (2008), S08004–S08004. DOI: 10.1088/1748-0221/3/08/s08004. URL: https://doi.org/10.1088/1748-0221/3/08/s08004 (cit. on pp. 25, 29).

[46] CMS Collaboration. "The CMS experiment at the CERN LHC". In: *Journal of Instrumentation* 3.08 (2008), S08004–S08004. DOI: 10.1088/1748-0221/3/08/s08004. URL: https://doi.org/10.1088/1748-0221/3/08/s08004 (cit. on pp. 26, 34).

[47] W. Adam et al. "The CMS Phase-1 pixel detector upgrade". In: *Journal of Instrumentation* 16.02 (2021), P02027–P02027. DOI: 10.1088/1748-0221/16/02/p02027. URL: https://doi.org/10.1088/1748-0221/16/02/p02027 (cit. on p. 30).

[48] S. Konig et al. "Assembly of the CMSpixel barrel modules". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 565.1 (2006). Proceedings of the International Workshop on Semiconductor Pixel Detectors for Particles and Imaging, pp. 62–66. ISSN: 0168-9002. DOI: https://doi.org/10.1016/j.nima.2006.04.081. URL: https://www.sciencedirect.com/science/article/pii/S0168900206007182 (cit. on p. 31).

[49] CMS Collaboration. "Description and performance of track and primary-vertex reconstruction with the CMS tracker". In: *Journal of Instrumentation* 9.10 (2014), P10009–P10009. DOI: 10.1088/1748-0221/9/10/p10009. URL: https://doi.org/10.1088/1748-0221/9/10/p10009 (cit. on pp. 31–33).

[50] Cristina Biino. "The CMS Electromagnetic Calorimeter: overview, lessons learned during Run 1 and future projections". In: *Journal of Physics: Conference Series* 587 (2015), p. 012001. DOI: 10.1088/1742-6596/587/1/012001. URL: https://doi.org/10.1088/1742-6596/587/1/012001 (cit. on pp. 33, 34).

[51] Badder Marzocchi. "Simulation of the CMS electromagnetic calorimeter response at the energy and intensity frontier". In: *Journal of Physics: Conference Series* 1162 (Jan. 2019), p. 012007. DOI: 10.1088/1742-6596/1162/1/012007 (cit. on p. 35).

[52] CMS Collaboration. "The CMS experiment at the CERN LHC". In: *JINST* 3 (2008), S08004. DOI: 10.1088/1748-0221/3/08/S08004 (cit. on pp. 36, 37).

[53] CMS Collaboration. "Performance of the CMS hadron calorimeter with cosmic ray muons and LHC beam data". In: *Journal of Instrumentation* 5.03 (2010), T03012–T03012. ISSN: 1748-0221. DOI: 10.1088/1748-0221/5/03/t03012. URL: http://dx.doi.org/10.1088/1748-0221/5/03/T03012 (cit. on p. 38).

[54] CMS Collaboration. "The performance of the CMS muon detector in proton-proton collisions at $\sqrt{s}$= 7 TeV at the LHC". In: *Journal of Instrumentation* 8.11 (2013), P11002–P11002. ISSN: 1748-0221. DOI: 10.1088/1748-0221/8/11/p11002. URL: http://dx.doi.org/10.1088/1748-0221/8/11/P11002 (cit. on pp. 38–40).

[55] CMS Collaboration. "Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at $\sqrt{s}$=13 TeV". In: *Journal of Instrumentation* 13.06 (2018), P06015–P06015. ISSN: 1748-0221. DOI: 10.1088/1748-0221/13/06/p06015. URL: http://dx.doi.org/10.1088/1748-0221/13/06/P06015 (cit. on p. 39).

[56] G. Weiglein et al. "Physics interplay of the LHC and the ILC". In: *Physics Reports* 426.2-6 (2006), pp. 47–358. ISSN: 0370-1573. DOI: 10.1016/j.physrep.2005.12.003. URL: http://dx.doi.org/10.1016/j.physrep.2005.12.003 (cit. on p. 41).

[57] Albert M Sirunyan et al. "Performance of the CMS Level-1 trigger in proton-proton collisions at $\sqrt{s}$ = 13 TeV". In: *JINST* 15 (2020), P10017. DOI: 10.1088/1748-0221/15/10/P10017. arXiv: 2006.10165 [hep-ex] (cit. on p. 42).

[58] CMS Collaboration. "The CMS trigger system". In: *Journal of Instrumentation* 12.01 (2017), P01020–P01020. ISSN: 1748-0221. DOI: 10.1088/1748-0221/12/01/p01020. URL: http://dx.doi.org/10.1088/1748-0221/12/01/P01020 (cit. on pp. 42, 43).

[59] Mia Tosi. *The CMS trigger in Run 2*. Tech. rep. Geneva: CERN, 2017. DOI: 10.22323/1.314.0523. URL: http://cds.cern.ch/record/2290106 (cit. on p. 43).

[60] CMS Collaboration. "Particle-flow reconstruction and global event description with the CMS detector". In: *JINST* 12.10 (2017), P10003–P10003. DOI: 10.1088/1748-0221/12/10/p10003. URL: https://doi.org/10.1088/1748-0221/12/10/p10003 (cit. on pp. 44, 51).

[61] Shawn Williamson. *Search for Higgs-Boson Production in Association with a Top-Quark Pair in the Boosted Regime with the CMS Experiment*. presented 11 Nov 2016. 2016. DOI: 10.5445/IR/1000068213. URL: http://cds.cern.ch/record/2300276 (cit. on p. 45).

[62] R.Keith Ellis et al. "Factorization and the parton model in QCD". In: *Physics Letters B* 78.2-3 (1978), pp. 281–284. DOI: 10.1016/0370-2693(78)90023-0. URL: https://doi.org/10.1016/0370-2693(78)90023-0 (cit. on p. 46).

[63] G. Altarelli and G. Parisi. "Asymptotic freedom in parton language". In: *Nuclear Physics B* 126.2 (1977), pp. 298–318. DOI: 10.1016/0550-3213(77)90384-4. URL: https://doi.org/10.1016/0550-3213(77)90384-4 (cit. on p. 46).

[64] Yuri L. Dokshitzer. "Calculation of the structure functions for deep inelastic scattering and $e^+e^-$ annihilation by perturbation theory in quantum chromodynamics." In: *Sov. Phys. JETP* 46 (1977). [Zh. Eksp. Teor. Fiz.73,1216(1977)], pp. 641–653 (cit. on p. 46).

[65] V.N. Gribov and L.N. Lipatov. "Deep inelastic electron scattering in perturbation theory". In: *Physics Letters B* 37.1 (1971), pp. 78–80. DOI: 10.1016/0370-2693(71)90576-4. URL: https://doi.org/10.1016/0370-2693(71)90576-4 (cit. on p. 46).

[66] L. N. Lipatov. "The parton model and perturbation theory". In: *Sov. J. Nucl. Phys.* 20 (1975). [Yad. Fiz.20,181(1974)], pp. 94–102 (cit. on p. 46).

[67] Richard D. Ball et al. "Parton distributions from high-precision collider data". In: *Eur. Phys. J. C* 77 (2017), p. 663. DOI: 10.1140/epjc/s10052-017-5199-5. arXiv: 1706.00428 [hep-ph] (cit. on pp. 47, 69).

[68] J. Alwall et al. "The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations". In: *JHEP* 07 (2014), p. 079. DOI: 10.1007/JHEP07(2014)079. arXiv: 1405.0301 [hep-ph] (cit. on pp. 48, 68, 109).

[69] Stefano Frixione, Giovanni Ridolfi, and Paolo Nason. "A positive-weight next-to-leading-order Monte Carlo for heavy flavour hadroproduction". In: *JHEP* 09 (2007), p. 126. DOI: 10.1088/1126-6708/2007/09/126. arXiv: 0707.3088 [hep-ph] (cit. on pp. 48, 68).

[70] Stefano Frixione, Paolo Nason, and Carlo Oleari. "Matching NLO QCD computations with Parton Shower simulations: the POWHEG method". In: *JHEP* 11 (2007), p. 070. DOI: 10.1088/1126-6708/2007/11/070. arXiv: 0709.2092 [hep-ph] (cit. on pp. 48, 68).

[71] Simone Alioli et al. "A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX". In: *JHEP* 06 (2010), p. 043. DOI: 10.1007/JHEP06(2010)043. arXiv: 1002.2581 [hep-ph] (cit. on pp. 48, 68).

[72] John C. Collins. "Sudakov Form Factors". In: *Adv.Ser.Direct.High Energy Phys* 5 (2003), pp. 573–614. DOI: 10.48550/ARXIV.HEP-PH/0312336. URL: https://arxiv.org/abs/hep-ph/0312336 (cit. on p. 48).

[73] J. Alwall et al. "Comparative study of various algorithms for the merging of parton showers and matrix elements in hadronic collisions". In: *The European Physical Journal C* 53 (2007), p. 473. DOI: 10.1140/epjc/s10052-007-0490-5. URL: https://doi.org/10.1140/epjc/s10052-007-0490-5 (cit. on pp. 48, 69).

[74] Stefano Catani et al. "QCD Matrix Elements + Parton Showers". In: *Journal of High Energy Physics* 2001.11 (2001), pp. 063–063. DOI: 10.1088/1126-6708/2001/11/063. URL: https://doi.org/10.1088/1126-6708/2001/11/063 (cit. on p. 48).

[75] B. Andersson et al. "Parton fragmentation and string dynamics". In: *Physics Reports* 97.2 (1983), pp. 31–145. ISSN: 0370-1573. DOI: https://doi.org/10.1016/0370-1573(83)90080-7. URL: https://www.sciencedirect.com/science/article/pii/0370157383900807 (cit. on p. 49).

[76] Torbjörn Sjöstrand et al. "An Introduction to PYTHIA 8.2". In: *Comput. Phys. Commun.* 191 (2015), p. 159. DOI: 10.1016/j.cpc.2015.01.024. arXiv: 1410.3012 [hep-ph] (cit. on pp. 49, 69).

[77] CMS Collaboration". "Event generator tunes obtained from underlying event and multiparton scattering measurements". In: *Eur. Phys. J. C* 76 (2016), p. 155. DOI: 10.1140/epjc/s10052-016-3988-x. arXiv: 1512.00815 [hep-ex] (cit. on pp. 49, 69).

[78] CMS Collaboration. "Extraction and validation of a new set of CMSPYTHIA8 tunes from underlying-event measurements". In: *Eur. Phys. J. C* 80 (2020), p. 4. DOI: 10.1140/epjc/s10052-019-7499-4. arXiv: 1903.12179 [hep-ex] (cit. on pp. 49, 69).

[79] John Allison et al. "Geant4 developments and applications". In: *IEEE Trans. Nucl. Sci.* 53 (2006), p. 270. DOI: 10.1109/TNS.2006.869826 (cit. on pp. 49, 69, 150).

[80] CMS Collaboration. *Technical proposal for the Phase-II upgrade of the Compact Muon Solenoid*. CMS Technical Proposal CERN-LHCC-2015-010, CMS-TDR-15-02. 2015. URL: http://cds.cern.ch/record/2020886 (cit. on p. 51).

[81] CMS Collaboration. "Performance of CMS muon reconstruction in pp collision events at $\sqrt{s} = 7$TeV". Submitted to *J. Inst.* 2012. arXiv: 1206.4071 [physics.ins-det] (cit. on p. 53).

[82] W Adam et al. "Reconstruction of electrons with the Gaussian-sum filter in the CMS tracker at the LHC". In: *Journal of Physics G: Nuclear and Particle Physics* 31.9 (2005), N9–N20. DOI: 10.1088/0954-3899/31/9/n01 (cit. on p. 54).

[83] CMS collaboration. "Electron and photon reconstruction and identification with the CMS experiment at the CERN LHC". In: *Journal of Instrumentation* 16.05 (2021), P05014. DOI: 10.1088/1748-0221/16/05/p05014. eprint: 2012.06888. URL: https://doi.org/10.1088%2F1748-0221%2F16%2F05%2Fp05014 (cit. on p. 55).

[84] Gavin P Salam and Gregory Soyez. "A practical seedless infrared-safe cone jet algorithm". In: *Journal of High Energy Physics* 2007.05 (2007). DOI: 10.1088/1126-6708/2007/05/086 (cit. on p. 56).

[85] CMS Collaboration. *A Cambridge-Aachen (C-A) based Jet Algorithm for boosted top-jet tagging*. Tech. rep. Geneva: CERN, 2009. URL: https://cds.cern.ch/record/1194489 (cit. on p. 57).

[86] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. "The anti-$k_t$ jet clustering algorithm". In: *JHEP* 04 (2008), p. 063. DOI: 10.1088/1126-6708/2008/04/063. arXiv: 0802.1189 [hep-ph] (cit. on pp. 57, 58).

[87] CMS Collaboration. "Jet energy scale and resolution in the CMSexperiment in pp collisions at 8 TeV". In: *JINST* 12.02 (2017), P02014–P02014. DOI: 10.1088/1748-0221/12/02/p02014. URL: https://doi.org/10.1088/1748-0221/12/02/p02014 (cit. on pp. 59, 61, 112).

[88] Daniele Bertolini et al. "Pileup per particle identification". In: *JHEP* 10 (2014), p. 059. DOI: 10.1007/JHEP10(2014)059. arXiv: 1407.6013 [hep-ph] (cit. on pp. 59, 60).

[89] Wikimedia commons. *b-jet Figure*. https://commons.wikimedia.org/w/index.php?curid=49738737 (cit. on p. 64).

[90] Emil Bols et al. "Jet Flavour Classification Using DeepJet". In: *JINST* 15 (2020), P12012. DOI: 10.1088/1748-0221/15/12/P12012. arXiv: 2008.10519 [hep-ex] (cit. on pp. 63, 65, 82).

[91] CMS collaboration. "Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV". In: *JINST* 13 (2018), P05011. DOI: 10.1088/1748-0221/13/05/P05011. eprint: 1712.07158 (cit. on pp. 63, 64, 71).

[92] David Curtin et al. "Exotic decays of the 125 GeV Higgs boson". In: *Physical Review D* 90 (2014). ISSN: 1550-2368. DOI: 10.1103/physrevd.90.075004. URL: http://dx.doi.org/10.1103/PhysRevD.90.075004 (cit. on p. 68).

[93] Adam Alloul et al. "FeynRules 2.0 - A complete toolbox for tree-level phenomenology". In: *Comput. Phys. Commun.* 185 (2014), p. 2250. DOI: 10.1016/j.cpc.2014.04.012. arXiv: 1310.1921 [hep-ph] (cit. on p. 68).

[94] Michal Czakon and Alexander Mitov. "Top++: A Program for the Calculation of the Top-Pair Cross-Section at Hadron Colliders". In: *Comput. Phys. Commun.* 185 (2014), p. 2930. DOI: 10.1016/j.cpc.2014.06.021. arXiv: 1112.5675 [hep-ph] (cit. on p. 68).

[95] CMS Collaboration". *Investigations of the impact of the parton shower tuning in Pythia 8 in the modelling of* $\mathrm{t\bar{t}}$ *at* $\sqrt{s} = 8$ *and 13 TeV*. Tech. rep. CMS-PAS-TOP-16-021. 2016. URL: https://cds.cern.ch/record/2235192 (cit. on p. 69).

[96] Richard D. Ball et al. "Parton distributions for the LHC Run II". In: *JHEP* 04 (2015), p. 040. DOI: 10.1007/JHEP04(2015)040. arXiv: 1410.8849 (cit. on p. 69).

[97] Andy Buckley et al. "LHAPDF6: parton density access in the LHC precision era". In: *Eur. Phys. J. C* 75 (2015), p. 132. DOI: 10.1140/epjc/s10052-015-3318-8. arXiv: 1412.7420 [hep-ph] (cit. on p. 69).

[98] CMS Collaboration. "Measurement of the inelastic proton-proton cross section at $\sqrt{s} = 13$TeV". In: *JHEP* 07 (2018), p. 161. DOI: 10.1007/JHEP07(2018)161. arXiv: 1802.02613 [hep-ex] (cit. on p. 69).

[99] 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". In: *JHEP* 01 (2013), p. 013. DOI: 10.1007/JHEP01(2013)013. arXiv: 1210.2387 [hep-ex] (cit. on p. 77).

[100] CMS Collaboration". "Search for massive resonances in dijet systems containing jets tagged as W or Z boson decays in pp collisions at $\sqrt{s} = 8$ TeV". In: *JHEP* 08 (2014), p. 173. DOI: 10.1007/JHEP08(2014)173. arXiv: 1405.1994 [hep-ex] (cit. on p. 77).

[101] Huilin Qu and Loukas Gouskos. "ParticleNet: Jet Tagging via Particle Clouds". In: *Phys. Rev. D* 101.5 (2020), p. 056019. DOI: 10.1103/PhysRevD.101.056019. arXiv: 1902.08570 [hep-ph] (cit. on p. 77).

[102] CMS Collaboration. *Identification of highly Lorentz-boosted heavy particles using graph neural networks and new mass decorrelation techniques*. CMS Physics Analysis Summary CMS-DP-2020-002. CERN, 2020. URL: https://cds.cern.ch/record/2707946 (cit. on pp. 77, 78).

[103] CMS Collaboration. "Identification of heavy, energetic, hadronically decaying particles using machine-learning techniques". In: *JINST* 15 (2020), P06005. DOI: 10.1088/1748-0221/15/06/P06005. arXiv: 2004.08262 [hep-ex] (cit. on p. 78).

[104] CMS collaboration. "Calibration of the mass-decorrelated ParticleNet tagger for boosted $b\bar{b}$ and $c\bar{c}$ jets using LHC Run 2 data". In: (2022). URL: https://cds.cern.ch/record/2805611 (cit. on pp. 78, 79, 107).

[105] Richard G Lomax and Debbie L Hahs-Vaughn. *Statistical concepts: a second course*. Taylor and Francis, 2012. URL: https://cds.cern.ch/record/1487958 (cit. on p. 84).

[106] Glen Cowan. *Statistical Data Analysis*. en. Oxford science publications. Oxford, England: Clarendon Press, 1998 (cit. on p. 85).

[107] Steve Baker and Robert D. Cousins. "Clarification of the use of CHI-square and likelihood functions in fits to histograms". In: *Nuclear Instruments and Methods in Physics Research* 221.2 (1984), pp. 437–442. DOI: 10.1016/0167-5087(84)90016-4. URL: https://doi.org/10.1016/0167-5087(84)90016-4 (cit. on p. 85).

[108] CMS Collaboration. "Pileup mitigation at CMSin 13 TeV data". In: *JINST* 15 (2020), P09018. DOI: 10.1088/1748-0221/15/09/p09018. arXiv: 2003.00503 [hep-ex] (cit. on p. 89).

[109] CMS Collaboration. "Measurement of the top quark mass with lepton+jets final states using pp collisions at $\sqrt{s} = 13$TeV". In: *Eur. Phys. J. C* 78 (2018), p. 891. DOI: 10.1140/epjc/s10052-018-6332-9. arXiv: 1805.01428 [hep-ex] (cit. on p. 89).

[110] CMS Collaboration. *CMS luminosity measurement for the 2017 data-taking period at $\sqrt{s} = 13$ TeV*. CMS Physics Analysis Summary CMS-PAS-LUM-17-004. 2018. URL: https://cds.cern.ch/record/2621960/ (cit. on p. 90).

[111] CMS collaboration. *CMS luminosity measurement for the 2018 data-taking period at $\sqrt{s} = 13$ TeV*. CMS Physics Analysis Summary CMS-PAS-LUM-18-002. 2019. URL: https://cds.cern.ch/record/2676164/ (cit. on p. 90).

[112] CMS Collaboration. "Measurement of the inelastic proton-proton cross section at $\sqrt{s} = 13$TeV". In: *JHEP* 07 (2018), p. 161. DOI: 10.1007/JHEP07(2018)161. arXiv: 1802.02613 [hep-ex] (cit. on pp. 90, 113).

[113] Jon Butterworth et al. "PDF4LHC recommendations for LHC Run II". In: *J. Phys. G* 43 (2016), p. 023001. DOI: 10.1088/0954-3899/43/2/023001. arXiv: 1510.03865 [hep-ph] (cit. on pp. 90, 113).

[114] Roger Barlow and Christine Beeston. "Fitting using finite Monte Carlo samples". In: *Comput. Phys. Commun.* 77 (1993), p. 219. ISSN: 0010-4655. DOI: 10.1016/0010-4655(93)90005-W (cit. on p. 90).

[115] J. S. Conway. "Incorporating Nuisance Parameters in Likelihoods for Multisource Spectra". In: *Proceedings, PHYSTAT 2011 Workshop on Statistical Issues Related to Discovery Claims in Search Experiments and Unfolding, CERN, Geneva, Switzerland 17-20 January 2011*. 2011, p. 115. DOI: 10.5170/CERN-2011-006.115. arXiv: 1103.0354 [physics.data-an] (cit. on p. 90).

[116] Ofer Vitells and Eilam Gross. "Estimating the significance of a signal in a multi-dimensional search". In: *Astropart. Phys.* 35 (2011), p. 230. DOI: 10.1016/j.astropartphys.2011.08.005. arXiv: 1105.4355 [astro-ph.IM] (cit. on pp. 98, 99, 101).

[117] A. L. Read. "Presentation of search results: the $CL_s$ technique". In: *J. Phys. G* 28 (2002), p. 2693. DOI: 10.1088/0954-3899/28/10/313 (cit. on pp. 101, 102).

[118] T. Junk. "Confidence level computation for combining searches with small statistics". In: *Nucl. Instrum. Meth. A* 434 (1999), p. 435. DOI: 10.1016/S0168-9002(99)00498-2. arXiv: hep-ex/9902006 [hep-ex] (cit. on pp. 101, 102).

[119] Glen Cowan et al. "Asymptotic formulae for likelihood-based tests of new physics". In: *The European Physical Journal C* 71.2 (2011). ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-011-1554-0. URL: http://dx.doi.org/10.1140/epjc/s10052-011-1554-0 (cit. on p. 101).

[120] J. M. Lindert et al. "Precise predictions for $V$+jets dark matter backgrounds". In: *The European Physical Journal C* 77.12 (2017). DOI: 10.1140/epjc/s10052-017-5389-1. URL: https://doi.org/10.1140%2Fepjc%2Fs10052-017-5389-1 (cit. on pp. 110, 112).

[121] *Precision luminosity measurement in proton-proton collisions at $\sqrt{s} = 13$ TeV in 2015 and 2016 at CMS*. Tech. rep. Geneva: CERN, 2021. arXiv: 2104.01927. URL: https://cds.cern.ch/record/2759951 (cit. on p. 113).

[122] Marija Majer et al. “Dose mapping of the panoramic 60Co gamma irradiation facility at the Ruđer Bošković Institute – Geant4 simulation and measurements”. In: *Applied Radiation and Isotopes* 154 (2019), p. 108824. ISSN: 0969-8043. DOI: https://doi.org/10.1016/j.apradiso.2019.108824. URL: https://www.sciencedirect.com/science/article/pii/S0969804319301071 (cit. on p. 143).

[123] Matej Roguljić et al. “Low dose rate $^{60}$Co facility in Zagreb”. In: *PoS* Vertex2019 (2020), p. 066. DOI: 10.22323/1.373.0066 (cit. on pp. 147, 148, 150–154).

# Appendix A

# Simulation of an irradiation facility

Dose mapping of an irradiation facility is an important task which ensures the application of precise radiation doses to target samples. It can be done experimentally by measuring the dose rates at various positions, or with the use of simulation. The latter is a very useful tool for the dose estimation in complex samples for which it may be difficult to ensure charged particle equilibrium (defined in the following sections) or to place a dosimeter. Furthermore, it enables detailed dose mapping for large samples, where the radiation field is not uniform across the sample. In this appendix, a brief summary of photon radiation interaction with matter is given in Sec. A.1, followed by Sec. A.2 in which the concepts related to the radiation dose are defined. The dose mapping results of the $^{60}$Co irradiation facility and their comparison to simulation, presented in Ref. [122], are summarized in Sec. A.3.

## A.1 Interaction of gamma radiation with matter

Cobalt-60 is a radioactive isotope of cobalt, produced in nuclear reactors, with a half-life of 5.2 years. It undergoes beta minus decay where one of its neutrons is converted into a proton, with the emission of an electron and its anti-neutrino

$$^{60}_{27}\mathrm{Co} \rightarrow ^{60}_{28}\mathrm{Ni}* + \mathrm{e}^{-} + \bar{\nu}_e \tag{A.1}$$

Nickel-60 is a stable nucleus, however, it is created in an excited state. Its de-excitation leads to the emission of photons. The decay scheme is shown in Fig. A.1. In most of the cases, two photons of 1.17 MeV and 1.33 MeV are emitted and one per-mill of decays result in only one photon of 1.33 MeV, but with the emission of a more energetic electron.

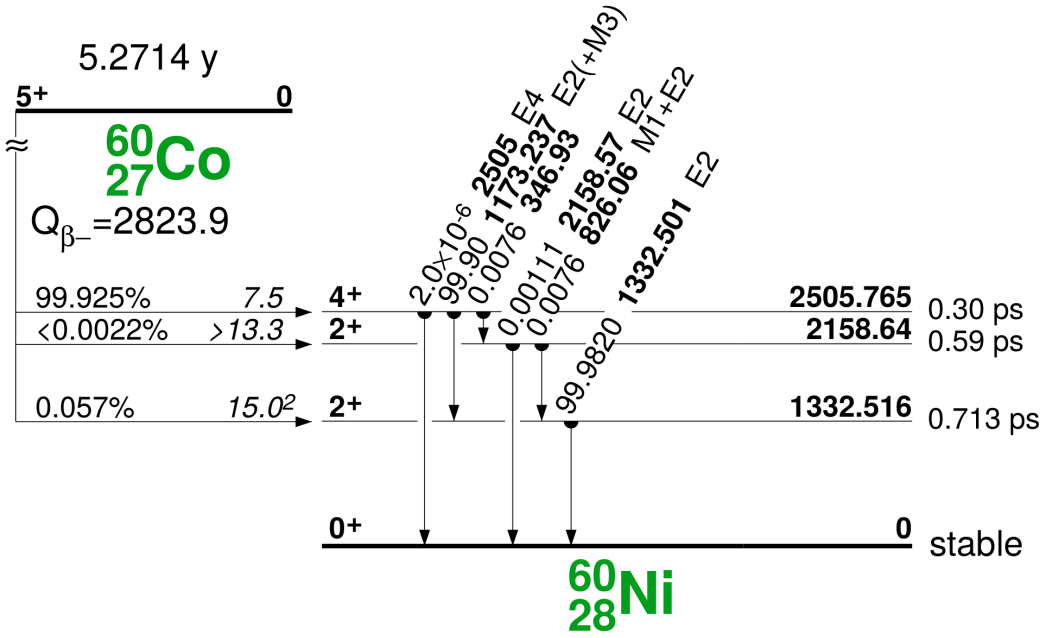



Figure A.1: Decay scheme of $^{60}$Co.

In radiation and nuclear physics, photons with energies above 100 keV are also called gamma rays or gamma radiation. Gamma radiation is much more penetrating, compared to alpha (Helium-4 nuclei) and beta (electrons or positrons) radiation. There are three most important processes through which ionizing photon radiation, i.e. radiation which is energetic enough to remove electrons from atoms and molecules, interacts with matter.

- The photoelectric effect involves the absorption of a photon by an electron which is consequently ejected from the atom. This is the dominant energy transfer mechanism for photons with $E < 100\,\text{keV}$.
- Compton scattering is a process where the incident photon ejects an electron from the atom (as in the photoelectric effect), but accompanied with the emission of a lower energy photon. The photon is emitted in a different direction from the incident photon which is why the effect is called scattering. Compton scattering is the dominant interaction mechanism in the $0.1 < E < 5\,\text{MeV}$ range. Therefore, this is the most relevant mechanism when discussing radiation emitted by $^{60}$Co decays.

- Electron-positron pair production is the mechanism which is possible when gamma energies exceed the mass of two electrons (1.02 MeV). In the interaction with the nucleus, the photon is converted into an electron-positron pair. Any excess energy appears as kinetic energy of the pair or as recoil of the nucleus. The effect becomes dominant around gamma energies of 5 MeV.

High energy gammas can also directly interact with nuclei, causing ejection of some nucleons (photodisintegration) or even the splitting of a nucleus (photofission). However, these processes are not of interest to us as the energies of considered gammas are not high enough.

In all the interaction processes, the incident gamma is removed from the beam. For this reason, the intensity of a gamma beam, defined as the power passing through unit area, traversing a material along z-direction, which is the beam direction, follows exponential decrease:

$$I(x, y, z) = I_0 f(x, y) e^{-n\sigma z}, \tag{A.2}$$

where $I_0$ is the starting intensity, $f(x, y)$ is the intensity profile, $n$ is the number of atoms per $\mathrm{cm}^3$ and $\sigma$ is the interaction cross section. The product of cross section and atom density, called the attenuation coefficient, is more often found in literature, $\mu = n\sigma$. It has a dependence on the material and the energy of the gammas.

## A.2 Total ionizing dose

The absorbed dose or total ionizing dose (TID) is a quantity that measures the energy deposited in matter by ionizing radiation per unit mass. Its SI unit is Gray (1 Gy=J/kg), but other units are also used, for example "rad", $1\,\mathrm{rad} = 0.01\,\mathrm{Gy}$. We can try calculating the dose along a thin rod, stretched along $z$, made of a material with mass density $\rho$. A small rod cross section, $A$, in the $x - y$ plane ensures that the intensity only depends on $z$. The energy lost by the beam in a small section, $\Delta z$, of this rod, with mass, $m = \rho \Delta z A$

is:

$$\begin{aligned} E &= \left[I(z) - I(z+\Delta z)\right] tA \\ &= I_0 e^{-\mu z}\left[1 - e^{-\mu \Delta z}\right] tA \\ &= I_0 e^{-\mu z} \mu \Delta z t A \end{aligned} \tag{A.3}$$

where $t$ is the time duration of irradiation. If all the energy lost by the beam is absorbed by the material at the point of interaction, the absorbed dose can then be expressed as:

$$D = I_0 \frac{\mu}{\rho} e^{-\mu z} t \tag{A.4}$$

The quantity $\mu/\rho$ is called the mass attenuation coefficient and it characterizes how effective a material is for the purpose of shielding, per unit of mass. From Eq. A.4, the absorbed dose can be calculated along each point $z$ of the thin rod. Furthermore, the beam intensity dependence can be easily reinstated by inserting the following substitution, $I_0 = I_0(x, y)$. However, Eq. A.4 assumes that all of the energy lost by the gamma beam gets absorbed in the same position where the interaction occurs, which is not exactly correct due to three reasons. First, photons which scatter from the material without any loss of energy are considered as energy lost by the beam. However, that energy is not transferred to the material. This is taken care of by replacing the mass attenuation coefficient with mass energy transfer coefficient, $\mu_{tr}$, before the exponential in the Eq. A.4. Eq. A.4 then describes the amount of kinetic energy released in the material per unit mass (Kerma). Second, not all kinetic energy released in the material is absorbed by the material (collision kerma), some of it may be irradiated (radiative kerma). Therefore, if we are interested in the absorbed dose, the mass energy absorption coefficient $\mu_{en}$ should be used instead. The third reason is that an electron ejected by the incoming gamma travels in the material and loses energy through the ionization of other atoms in the material. Therefore, the collision kerma generated at point $z$ will be spread out along the electron path.

Fig. A.2 shows an extremely simplified sketch of ejected electrons depositing the imparted energy over their path. Ejected electrons in any of the sections spread their energy across the next three sections in the direction of the beam ($+z$). Therefore, even though the loss

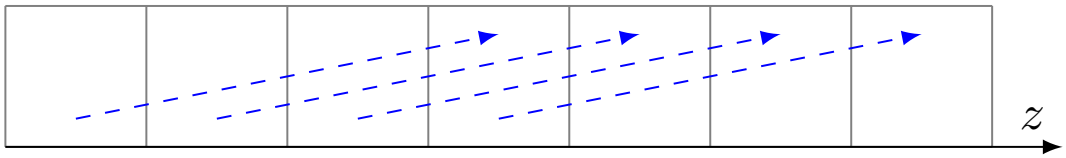



Figure A.2: Schema demonstrating gamma beam interaction point and the spread of energy through the ejected electrons, shown as blue arrows.

of gamma beam energy is uniform, the absorbed dose is not as there are electrons "leaking out" from one layer to the next few layers. The first few layers will then experience lower absorbed doses than what is given by Eq. A.4. The leakage of electrons in the bulk of the material are compensated by the incoming electrons from previous layers in which case the equivalence of absorbed dose and collision kerma holds. This balance is called the charged-particle equilibrium (CPE) and it occurs when the number of charged particles leaving a volume is equal to the number entering, for each energy and type of particle. When CPE exists in an irradiated medium, the absorbed dose in the volume is equal to the collision kerma.

Relationship between collision kerma and absorbed dose for a beam of photons is shown in Fig. A.3.

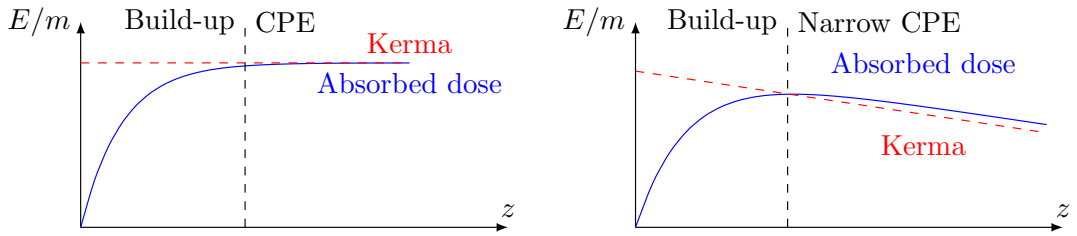



Figure A.3: Relationship between absorbed dose and collision kerma without (left) and with (right) beam attenuation considered.

If the beam attenuation is not taken into account, the CPE is established at some material depth, after the build-up region in which the absorbed dose is gradually rising and is lower than collision kerma. The build-up region, as can be deduced from Fig. A.2, depends on the mean range of the ejected electrons, which is closely related to the energy of the gammas in the beam, as shown in Fig. A.4. For $^{60}$Co gamma radiation, typical buildup length is around 1.5 mm [123].

The situation changes slightly when beam attenuation is taken into account. The CPE is only achieved in a small area where the effect of dose build-up is balanced with the

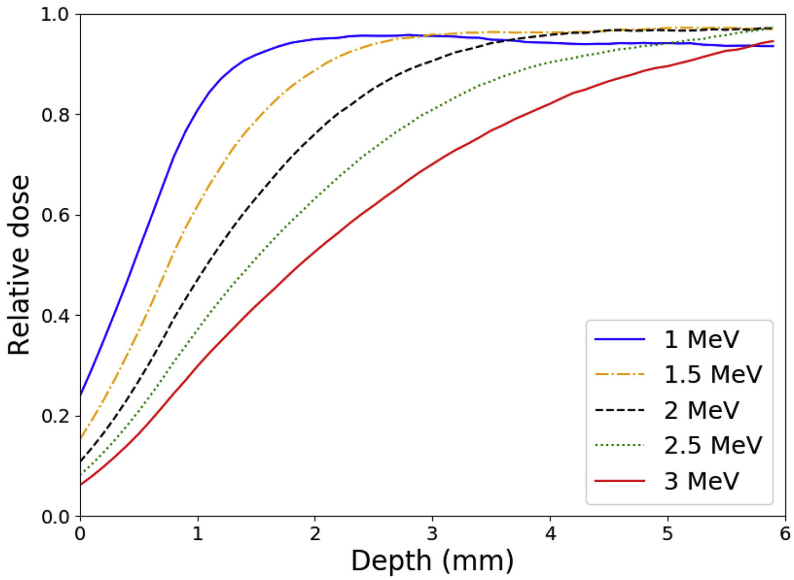



Figure A.4: Normalized dose profile of a gamma-irradiated block of silicon (20 cm × 20 cm, thickness 6 mm) for different photon energies simulated using Geant4. Taken from Ref. [123]

effect of attenuation, at the position of the dashed line in Fig. A.3. As a consequence of attenuation, after that point, collision kerma is lower than the absorbed dose since a given layer now receives more electrons from previous layers than the ejected electrons are leaking onto downstream layers. However, the mean range of the electrons released by $^{60}$Co gamma radiation is generally much lower than the attenuation length. We can take the build-up region length to be a measure of the mean range of electrons and compare it to the attenuation length. In water, for example, the two lengths are of the order of 1 mm and 10 cm for $^{60}$Co gamma irradiation. Therefore, the effect of attenuation can be neglected when considering CPE after the buildup region.

In conclusion, the absorbed dose in the first 1–2 mm of the material will be lower than the in the rest of the material where it is considered constant until the attenuation effects start becoming important. For thin samples, it is crucial to ensure CPE or to take into account the effects of buildup when calculating dose rates. The former can be done by placing additional material between the beam and the sample which will provide electrons ensuring CPE. Simulation of radiation can be used to measure the effects of buildup in irradiation cases in which ensuring CPE is difficult.

## A.3 Dose mapping of the Cobalt RBI irradiation facility

The $^{60}$Co gamma irradiation facility at the Ruđer Bošković Institute (RBI) is used both for commerical and scientific purposes. An example of the former is the irradiation of medical equipment to achieve sterilization. An example of the latter is the irradiation of sensors planned to be used in HEP experiments, allowing the assesment of their radaition hardness.

The facility consists of the control room and the irradiation chamber, both located in an underground vault. When there is no irradiation ongoing, the sources are stored in a lead container which is placed at the bottom of a 3.5 m deep well dug in the floor of the irradiation chamber. The container is shielded with lead bricks and gravel, ensuring safe levels of leakage radiation in the irradiation chamber when the sources are lowered.

During irradiation, sources are stored in an aluminum cylindrical rack, mounted on the floor of the irradiation chamber. The layout of the irradiation chamber and source rack with 24 guide tubes is shown in Fig. A.5 and Fig. A.6. The storage container and the rack are connected with 24 guide tubes made of stainless steel, inside which source assemblies can move between safe and working position.

The irradiation source itself is accordingly formed by 24 source assemblies arranged in a cylindrical shape as shown in Fig. A.6. Each source assembly consists of a stainless steel cylindrical holder containing four $^{60}$Co pencils. The source assemblies are hung on stainless steel cables, and the upper ends of the cables are hooked to a mechanism allowing for movement of the source assemblies between safe and working position. The dose absorbed during such movement is called transit dose and needs to be taken into account when irradiating to low doses. A facility timer is connected with the hoisting mechanism and gives measurements of the time duration during which the sources are in the working position. In the working position, the centre of each source assembly is elevated by 72 cm with respect to the chamber floor.

Mapping of the radiation field is performed to obtain the field dependence on the radial distance from the central vertical axis of the source rack ($r$), angular coordinate around

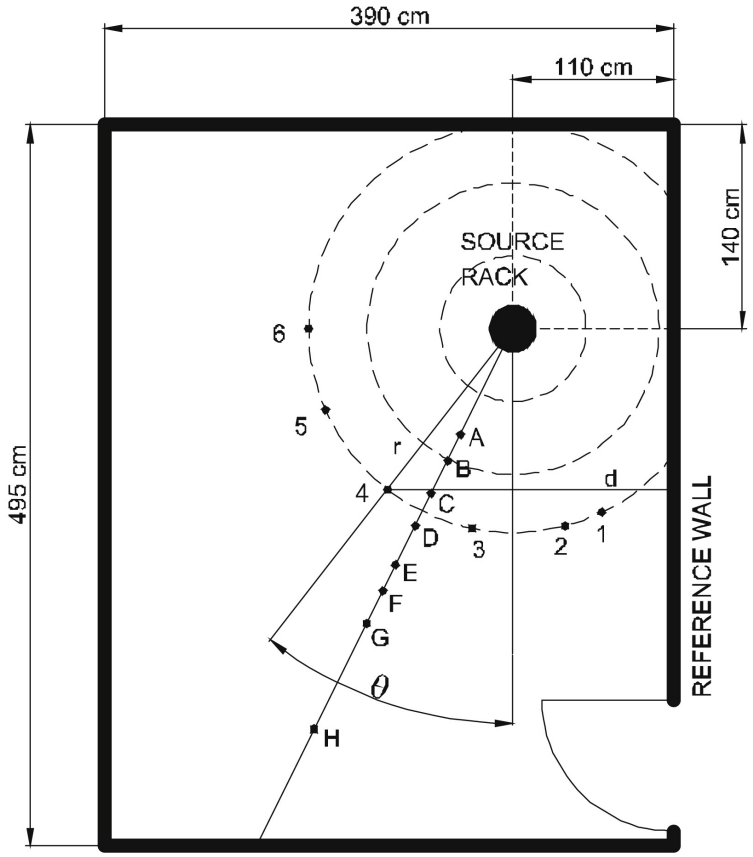



Figure A.5: Layout of the irradiation chamber. Circles around the source rack are drawn at distances of 50, 100 and 140 cm from the centre of the rack. Various positions for which the dose measurement and simulation were compared are denoted by 1–6 and A–H. Taken from Ref. [123].

the vertical axis ($\theta$) and height ($h$). Dose measurements are performed with a PTW Farmer ionisation chamber type 300313 which has a sensitive volume of 0.6 $cm^3$. The results are expressed as the "absorbed dose in water", i.e., the dose which would be absorbed by water in CPE at that location. The irradiation chamber is emptied during the dose mapping so that the only scattering is coming from the walls and the floor of the chamber.

A Monte Carlo simulation of the chamber is built using the GEANT4 [79] simulation toolkit. The geometry includes the source rack, consisting of 24 $^{60}$Co pencils, each surrounded by a cylindrical steel guide, its casing, and the covering lid. The aluminum casing consists of an inner and outer cylinder and a covering lid the top of the rack as can be seen in Fig. A.6. The walls of the room are also included in the geometry to account for the effect of backscattering.

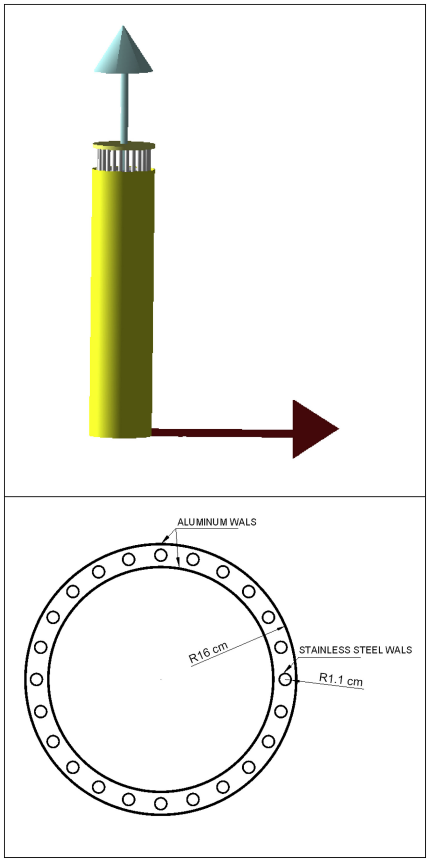



Figure A.6: Scheme of the source rack with the 24 guide tubes used for the Geant4 simulation (upper). Layout of the aluminum rack with 24 stainless steel guide tubes (lower). Source assemblies (24 cylindrical holders filled with $^{60}$Co pencils) are placed in guide tubes. Taken from Ref. [123].

In each simulated event, a photon originating in one of the $^{60}$Co pencils is emitted in a randomly selected direction following an isotropic angular distribution. Half of the photons are generated with an energy of 1.17 MeV and half with an energy of 1.33 MeV, corresponding to the energies of the photons emitted in $^{60}$Co decays. Photons are generated in any of the 24 pencils, uniformly across their volumes. A water sphere with a radius of 2 cm plays the role of a "dosimeter" in the simulation and the absorbed dose in the dosimeter sphere is computed. The statistical uncertainties on the total absorbed energy, corresponding to confidence level of 95%, are generally within 3% and never exceed 5%. The absorbed dose is defined as the total deposited energy in the sphere divided by its mass. It is converted to the dose rate through the relation:

$$\frac{dD}{dt} = 2\frac{AD}{N_{sim}}, \tag{A.5}$$

where $A$ is the summed activity of all source pencils defined as the number of decays per unit of time, $D$ is the absorbed dose of the water sphere in the simulation and $N_{sim}$ is the total number of generated photons. The additional factor of 2 accounts for the fact that two photons are emitted during each $^{60}$Co decay.

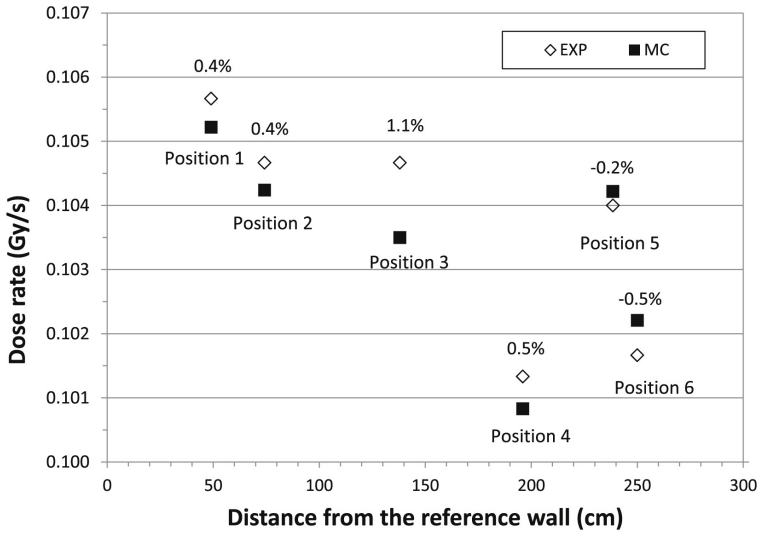



Figure A.7: Measured (EXP) and simulated (MC) angular dose rate dependence, measured on points 1–6 as shown in Fig. A.5. The relative difference between EXP and MC results are given for each position. Taken from Ref. [123].

Angular dependence of the radiation field is investigated by making the dose measurements at six positions on a circle around the source with $r = 140$ cm and $h = 72$ cm, labelled 1–6 in Fig. A.5. The dependence is shown in Fig. A.7. Although all measured points are at the same radial distance from the source and at the same height, their distances to the chamber walls are not the same. Therefore, the contributions from the scattered photons change. To confirm that this contribution is not negligible, a simulation without the chamber walls included in geometry is performed, which results in $\sim 6\%$ reduced dose rate.

The dependence of the radiation field on the radial distance is shown in Fig. A.8. The measurements are taken for $\theta = 19^\circ$ and $h = 72$ cm. The positions are labelled A–H in Fig. A.5. The dependence of the dose rate is fit by an inverse power law $\frac{dD}{dt} \propto r^{-k}$. As expected, the parameter $k$ is found to be close to 2, both for the measured and simulated values.

Finally, the dose rate is measured for different heights from the floor. Three sets of

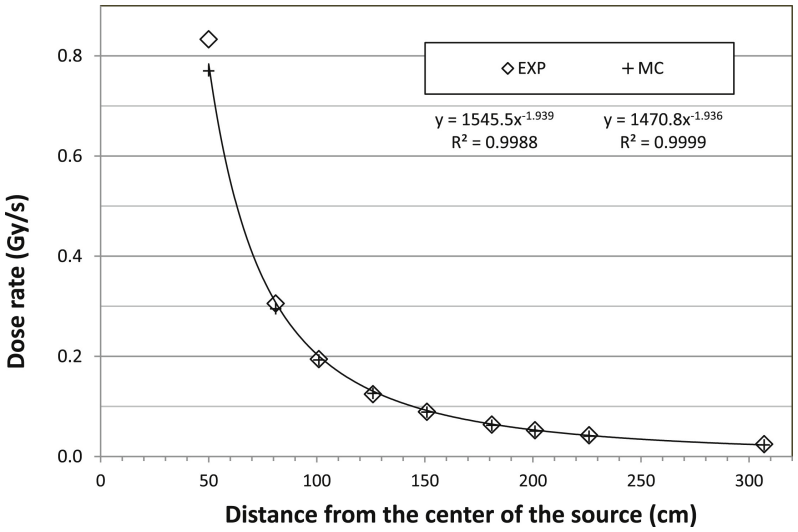



Figure A.8: Measured (EXP) and simulated (MC) dose rates as a function of the radial distance from the centre of the source. Taken from Ref. [123].

measurements are performed, at different radial distances from the centre of the source, $r = 50$, 100 and 140 cm. The results are shown in Fig. A.9. As expected, the dose rate variation with height is smaller at larger radii. Relative differences between measurement and simulation are up to 8%.

Overall, good agreement is observed between measured and simulated dose rate values with the exception of a few positions closest to the source. This may be remedied by the adjustment of geometry in simulation. For example, if the sources are shorter in simulation than they really are, it may lead to increased dose rates for positions around $h = 72$ cm and this effect would be more pronounced for small $r$, as is seen in Fig. A.9. Additionally, the gradient of the dose rate is increased closer to the source so a small difference in the positioning of the dosimeter will have a larger impact on the measured value. In summary, these results represent a solid foundation for the future simulation of the dose distribution in more complex materials and inside samples for which the simulation is crucial.

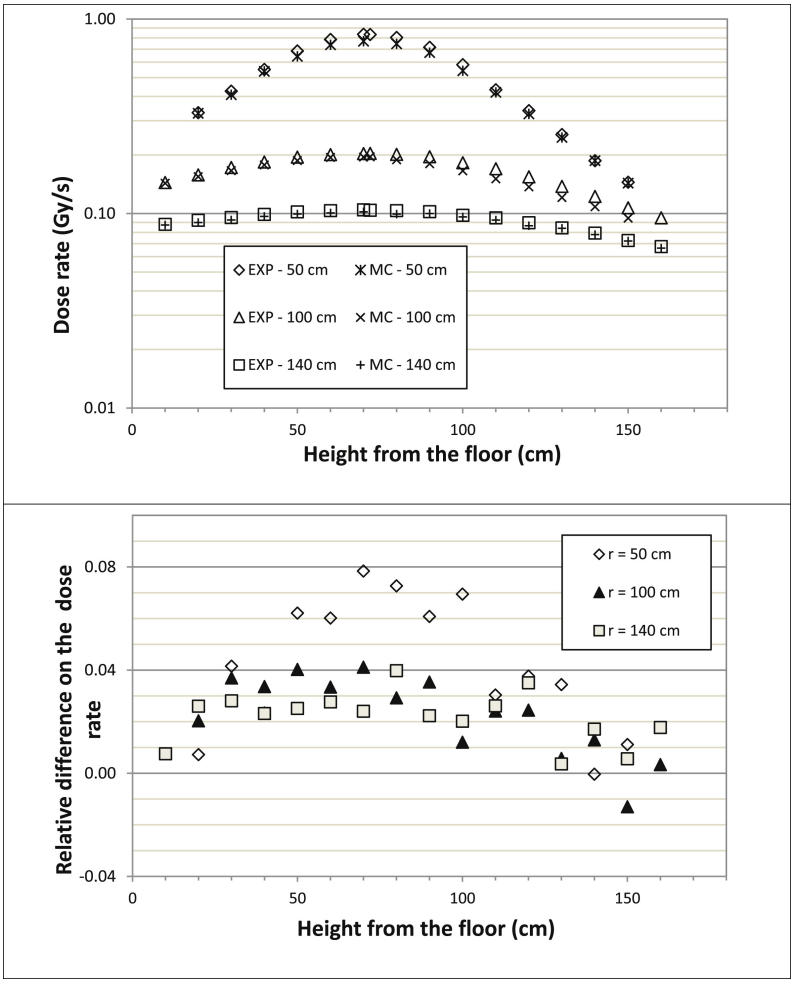



Figure A.9: Measured (EXP) and simulated (MC) dose rates (upper) and relative difference on dose rates (lower) as a function of the height from the floor for three sets of data: r = 50, 100 and 140 cm. Relative difference is calculated as (EXP-MC)/EXP. Taken from Ref. [123].

# Curriculum vitae

Matej Roguljić studied physics at the Faculty of Science at the University of Zagreb and attained his master's degree in 2017. After graduation he joined the high energy physics group at the Ruđer Bošković Institute, which is part of the CMS collaboration, and enrolled in the PhD course in physics at the University of Zagreb. During his PhD studies he worked on evaluating the radiation hardness and quality control of silicon detectors, rebuilding the first layer of the CMS pixel detector, a search for new scalar particles using data collected by CMS and the development of a method for calibrating an algorithm used to tag heavy-flavor hadronic jets.

## Publications

1. **"Production, Calibration, and Performance of the Layer 1 Replacement Modules for the CMS Pixel Detector"** D.Ferencek *et al.* [CMS Tracker Group],JPS Conf. Proc. **34** (2021), 010023

2. **"Low dose rate $^{60}$Co facility in Zagreb"** M.Roguljic *et al.*,PoS **Vertex2019** (2020), 066

3. **"Dose mapping of the panoramic 60Co gamma irradiation facility at the Ruđer Bošković Institute – Geant4 simulation and measurements"** M.Majer *et al.*,Appl. Radiat. Isot., **154**, (2019) 108824