University of Cyprus

DEPARTMENT OF PHYSICS

SEARCH FOR CHARGED HIGGS BOSONS WITH HADRONIC DECAYS OF TOP AND BOTTOM QUARKS AT THE CMS EXPERIMENT USING THE LHC RUN II pp COLLISION DATA

DOCTOR OF PHILOSOPHY DISSERTATION

SOTIROULLA KONSTANTINOU

2023

University of Cyprus

DEPARTMENT OF PHYSICS

SEARCH FOR CHARGED HIGGS BOSONS WITH HADRONIC DECAYS OF TOP AND BOTTOM QUARKS AT THE CMS EXPERIMENT USING THE LHC RUN II pp COLLISION DATA

SOTIROULLA KONSTANTINOU

A Dissertation Submitted to the University of Cyprus in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

May 2023

VALIDATION PAGE

Doctoral Candidate:
Sotiroulla Konstantinou

Doctoral Thesis Title:
Search for charged Higgs bosons with hadronic decays of top and bottom quarks at the CMS experiment using the LHC Run II pp collision data

The present Doctoral Dissertation was submitted in partial fulfillment of the requirements for the Degree of Doctor of Philosophy at the Department of Physics at the University of Cyprus. It was approved on May 17, 2023 by the members of the Examination Committee.

Examination Committee:

Ekaterini Tzamariudaki
Comittee member
Director of Research, National Centre for Scientific Research *Demokritos*

Konstantinos Kordas
Comittee member
Professor, Aristotle University of Thessaloniki

Nicolaos Toumbas
Committee head
Associate Professor, University of Cyprus

Halil Saka
Committee member
Lecturer, University of Cyprus

Fotios Ptochos
Research Advisor
Professor, University of Cyprus

DECLARATION OF DOCTORAL CANDIDATE

The present doctoral dissertation was submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy of the University of Cyprus. It is a product of original work of my own, unless otherwise mentioned through references, notes, or any other statements.

Sotiroulla Konstantinou

ABSTRACT

Particle physics aims to explain our current understanding of the subatomic nature of the universe. Our current theory, the Standard Model (SM) of particle physics, describes the nature of the elementary constituents of matter and their interaction through the strong, weak and electromagnetic forces. Furthermore, it predicts a mechanism that is responsible for the non-zero masses of particles and introduces an additional particle, the Higgs boson. The discovery of the Higgs boson by the ATLAS and CMS collaborations in 2012 was the last missing piece of the puzzle that confirmed the SM. However, there are many mysteries of the nature that the SM cannot explain, including the existence of dark energy, dark matter and the matter-antimatter asymmetry. Several theories seek to complete the SM. Some of them propose additional electrically charged and neutral Higgs bosons and the discovery of such particles would be a sign of new physics. After the discovery of the Higgs boson, the CMS and ATLAS experiments at the Large Hadron Collider (LHC) oriented the experimental studies to the observation of any possible extensions of the SM, which could potentially provide answers to a variety of currently unanswered questions in the field.

This work presents a search for charged Higgs bosons with a mass larger than that of the top quark, that decay to a top and bottom quark-antiquark pair, in the all-jet final state. The search uses proton-proton collision data recorded by the CMS detector during 2016-2018 at a center-of-mass energy of 13 TeV that corresponds to an integrated luminosity of 138 fb^{-1}. Simulated data are also used to describe the behavior of signal and background events. The events are reconstructed using dedicated algorithms that identify physics objects, by exploiting information from all the sub-components of the CMS detector.

Two analyses are performed, each targeting different regions of the signal parameter space. The resolved analysis is optimized for charged Higgs bosons with low masses that decay to top quarks with moderate transverse momentum (low Lorentz boost) identified as three separate small-radius jets. The semi-boosted analysis targets charged Higgs bosons with masses larger than 1 TeV, and utilizes collimated hadronically decaying Lorentz-boosted top quarks. The analyses use multivariate analysis techniques for particle identification and signal-to-background discrimination, and data-driven background estimation methods.

The analyses report upper exclusion limits on the product of the charged Higgs boson production cross-section and the branching ratio of its decay to a top and a bottom quark at a 95% confidence level. The resolved analysis sets limits of 8.375 to 0.107 pb for charged Higgs boson masses in the range of 200 to 1250 GeV. The semi-boosted analysis uses the data recorded in 2018, corresponding to 59.8 fb^{-1}. The preliminary results exhibit best sensitivity for large charged Higgs boson masses above 800 GeV. For both analyses, the upper limits show a significant improvement compared to the results published with the 2016 data.

ΠΕΡΙΛΗΨΗ

Η φυσική στοιχειωδών σωματιδίων στοχεύει στην κατανόηση της υποατομικής φύσης του σύμπαντος. Η επικρατούσα θεωρία, το Καθιερωμένο Πρότυπο (ΚΠ) της σωματιδιακής φυσικής, περιγράφει τη φύση των στοιχειωδών συστατικών της ύλης και την αλληλεπίδρασή τους μέσω της ισχυρής, ασθενούς και ηλεκτρομαγνητικής δύναμης. Επιπλέον, προβλέπει το μηχανισμό που είναι υπεύθυνος για τις μη μηδενικές μάζες των σωματιδίων και εισάγει ένα επιπλέον σωματίδιο, το μποζόνιο Higgs. Η ανακάλυψη του μποζονίου Higgs από τα πειράματα ATLAS και CMS το 2012 αποτελεί το τελευταίο κομμάτι του παζλ που επιβεβαιώνει το ΚΠ. Ωστόσο, υπάρχουν πολλά μυστήρια της φύσης που το ΚΠ αδυνατεί να εξηγήσει, όπως η ύπαρξη σκοτεινής ύλης και σκοτεινής ενέργειας, και η ασυμμετρια ύλης-αντιύλης. Αρκετές θεωρίες επιδιώκουν να συμπληρώσουν το ΚΠ. Μερικές από αυτές προτείνουν επιπρόσθετα ηλεκτρικά φορτισμένα και ουδέτερα μποζόνια Higgs, η ανακάλυψη των οποίων θα αποτελούσε ένδειξη νέας φυσικής. Μετά την ανακάλυψη του μποζονίου Higgs, τα πειράματα CMS και ATLAS στον Μεγάλο Αδρονικό Επιταχυντή (LHC) προσανατολίζουν τις πειραματικές τους μελέτες στην παρατήρηση πιθανών επεκτάσεων του ΚΠ, οι οποίες θα μπορούσαν να δώσουν απαντήσεις σε μια πληθώρα από αναπάντητα ερωτήματα στο πεδίο.

Η παρούσα διατριβή παρουσιάζει την αναζήτηση φορτισμένων μποζονίων Higgs με μάζα μεγαλύτερη από αυτή του top quark που διασπάται σε ένα ζεύγος top και bottom quark-antiquark, στην πλήρως αδρονική τελική κατάσταση. Η αναζήτηση χρησιμοποιεί δεδομένα σύγκρουσης πρωτονίου-πρωτονίου που καταγράφηκαν από τον ανιχνευτή CMS κατά την περίοδο 2016-2018 σε ενέργεια κέντρου μάζας 13 TeV η οποία αντιστοιχεί σε ολοκληρωμένη φωτεινότητα 138 fb^{-1}. Προσομοιωμένα δεδομένα χρησιμοποιούνται επίσης για να περιγράψουν τη συμπεριφορά των γεγονότων σήματος και υποβάθρου. Τα γεγονότα ανακατασκευάζονται χρησιμοποιώντας αλγόριθμους ταυτοποίησης αντικειμένων, αξιοποιώντας πληροφορίες από όλα τα υποσυστήματα του ανιχνευτή CMS.

Στην παρούσα εργασία εκτελούνται δύο αναλύσεις, καθεμία από τις οποίες στοχεύει σε διαφορετικές περιοχές με διαφορετική ευαισθησία στο μέγεθος της μάζας του αναζητούμενου μποζονίου. Η διακρίσιμη ή "resolved" ανάλυση έχει βελτιστοποιηθεί για την εύρεση φορτισμένων μποζονίων Higgs με μικρές μάζες που διασπώνται σε top quarks μέτριας εγκάρσιας ορμής (χαμηλής ώθησης) που ανακατασκευάζονται ως τρεις ξεχωριστοί πίδακες (jets) μικρής ακτίνας. Η δεύτερη ανάλυση στοχεύει στην εύρεση φορτισμένων μποζονίων Higgs με μάζες μεγαλύτερες από 1 TeV, και χρησιμοποιεί ανακατασκευασμένα top quarks χαμηλής ώθησης καθως επίσης συγχωνευμένους πίδακες μεγάλης ακτίνας που προέρχονται από αδρονικές διασπάσεις των top quarks με μεγάλη εγκάρσια ορμή (υψηλή ώθηση). Η ανάλυση αυτή ονομάζεται "semi-boosted". Οι αναλύσεις χρησιμοποιούν τεχνικές πολλαπλών μεταβλητών (multivariate analysis) για την ταυτοποίηση σωματιδίων και το διαχωρισμό σήματος-υποβάθρου, και μεθόδους εκτίμησης υποβάθρου που στηρίζονται σε

πραγματικά γεγονότα σύγκρουσης πρωτονίων.

Οι αναλύσεις θεσπίζουν ανώτερα όρια αποκλεισμού στο γινόμενο της ενεργού διατομής παραγωγής φορτισμένου μποζονίου Higgs και του ποσοστού διακλάδωσης της διάσπασής του σε top και bottom quarks σε επίπεδο εμπιστοσύνης 95%. Η resolved ανάλυση θέτει ανώτερα όρια από 8.375 έως 0.107 pb για μάζες φορτισμένων μποζονίων Higgs στην περιοχή από 200 έως 1250 GeV. Η semi-boosted ανάλυση χρησιμοποιεί τα δεδομένα που καταγράφηκαν το 2018, που αντιστοιχούν σε 59.8 fb^{-1}. Τα προκαταρκτικά αποτελέσματα παρουσιάζουν μεγαλύτερη ευαισθησία για μάζες φορτισμένων μποζονίων Higgs πάνω από 800 GeV. Και για τις δύο αναλύσεις τα ανώτερα όρια που επιτυγχάνονται παρουσιάζουν σημαντική βελτίωση σε σχέση με τα αποτελέσματα που δημοσιεύτηκαν με τα δεδομένα του 2016.

Στους γονείς μου, Κώστα και Παντελίτσα

ACKNOWLEDGEMENTS

First of all, I would like to express my gratitude to my supervisor, Fotis Ptochos, for his help and support throughout these years. His enthusiasm and dedication have always been a great source of inspiration and motivation for me. He believed in my abilities long before I did, and without his infinite patience, these pages would have never been written. Working with him has been an invaluable journey of continuous learning, and I hope that he also learnt something from me.

I want to thank my Ph.D. defense committee, Ekaterini Tzamariudaki, Konstantinos Kordas, Nicolaos Toumbas, Halil Saka and Fotios Ptochos for evaluating my work and providing me with insightful feedback.

I express my gratitude to the *Research & Innovation Foundation* for co-funding my research work and their financial support in enabling my visits to CERN.

Special thanks to all the members of my group at the University of Cyprus: Halil Saka for all the fruitful discussions during our weekly meetings. Mohsan Waseem Ather for his help in transferring me his knowledge on Top HLT validation and the moments filled with laughter at CERN. Gouranga Kole, the Monte Carlo expert for helping me during my first steps as a Monte Carlo contact and Anton Stepennov for all the interesting discussions during the lunch breaks. Alexandros Attikis for the countless discussions we had on physics (and music) in his office. His perfectionism has helped me to enhance my coding skills. I am truly grateful for his priceless support and his impact on my academic growth. Konstantinos Christoforou, Christos Leonidou and Leonidas Paizanos for their support and for always bringing joy at the office. Your presence has made the work environment truly enjoyable. Finally, my beloved friends, Marina Kolosova, Marina Toumazou and Mikela Haraki. Marina Kolosova, this journey wouldn't be the same without your endless support and encouragement. I greatly admire your dedication and I am forever grateful for always bringing me a smile. Marina Toumazou you are an endless source of laughter and joy. Thank you for always brightening up the atmosphere and for always being so generous. Your objective advice has made me stronger! Mikela, you have taught me the importance of standing firm in my beliefs. I admire your kindness and your enthusiasm for the little things in life.

I would also like to thank Sami Lehti for his insightful comments at the meetings and his help with the framework, and Tommaso Diotalevi for our incredible collaboration as Monte Carlo contacts. I thank Loukas Gouskos for his valuable input on machine learning and neural networks. I really appreciate his words during our corridor conversation at the University of Cyprus.

I truly thank my friends Despoina, Elena, Ioanna, Christina, Sozos and Andrea for your love and support throughout all these years. Your presence in my life has been a source of strength and happiness.

Τελος, θα ήθελα να ευχαριστήσω απο καρδιάς την οικογένεια μου. Τη μητέρα μου

Παντελίτσα και τον πατέρα μου Κώστα για τη στήριξη, την κατανόηση την διακριτικότητα και κυριως την αγάπη τους. Τα αδέρφια μου Χριστοθέα, Ραφαέλα και Γιάννη καθως επίσης τον Κωνσταντίνο και τον Άγγελο που κάνουν τη ζωή μου πιο χαρούμενη, και μια υπόσχεση να επανορθώσω για τους καφέδες που δεν ήπιαμε και τις ταινίες που δεν είδαμε (Γιάννη). Αφιερώνω αυτή την εργασία στην οικογένεια μου και στη γιαγιά μου Σωτήρα.

Author's contribution

Publications

- CMS Collaboration, "Search for charged Higgs bosons decaying into a top and a bottom quark in the all-jet final state of pp collisions at $\sqrt{s} = 13$ TeV", *JHEP*, vol. 07, p. 126, 2020. DOI: 10.1007/JHEP07(2020) 126. arXiv: 2001.07763 [hep-ex].
- CMS Collaboration "Search for a charged Higgs boson decaying into a heavy neutral Higgs boson and a W boson in proton-proton collisions at $\sqrt{s} = 13$ TeV", accepted by *JHEP*. arXiv: 2207.01046 [hep-ex].

Conferences

- Plenary talk. "Searches for heavy resonances decaying into Z, W and Higgs bosons at CMS" at "BOOST2022: 14th International Workshop on Boosted Object Phenomenology, Reconstruction, Measurements and Searches in HEP, 15-19 Aug 2022. Hamburg (Germany)"
- Parallel talk. "Searches for BSM Higgs bosons at CMS" at "TeVPA2021: TeV Particle Astrophysics, 24-29 Oct 2021, Chengdu (China)"
- Talk. "Supersymmetry and search for Charged Higgs boson" at "Third Pancyprian Conference of Physics, Mar 2019, Nicosia (Cyprus)"

CMS Internal Notes

- "Search for charged Higgs bosons with the $\mathrm{H}^{\pm} \to \mathrm{tb}$ decay in the fully hadronic final state at $\sqrt{s} = 13$ TeV with full Run II data" (CMS AN-2021/060)
- "Resolved top quark identification using deep neural network" (CMS AN-2021/019)
- "Search for charged Higgs bosons in the neutral Higgs boson and W boson decay channel at $\sqrt{s} = 13$ TeV with full Run II data" (CMS AN-2020/107)
- "Search for charged Higgs bosons with the $\mathrm{H}^{\pm} \to \mathrm{tb}$ decay in the fully hadronic final state with $\sqrt{s} = 13$ TeV data (CMS AN-2018/010)

CMS Internal Talks

- "$\mathrm{H}^{\pm} \to \mathrm{tb}$ (0ℓ, boosted + resolved) update" at "Beyond Two Generations (B2G) Meeting" (17 Feb 2023)
- "$\mathrm{H}^{\pm} \to \mathrm{tb}$ (0ℓ) status" at "Beyond Two Generations (B2G) Meeting" (25 Nov 2022)
- "Status of resonant MC samples production (Run 2)" at "Higgs Meeting - HH meeting" (07 Nov 2022)

- "Status of the $\mathrm{H}^{\pm} \to \mathrm{tb}$ fully hadronic analysis" at "Higgs Meeting - Extended Higgs Sectors" (13 Sep 2021)
- "Resolved top tagger for HIG-21-010" at "JetMET Algorithms and Reconstruction (JMAR) Meeting" (01 Jun 2021)
- "Status of HExtended MC requests" at "Higgs Meeting - Extended Higgs Sectors" (29 Mar 2021)
- "Status of the $\mathrm{H}^{\pm} \to \mathrm{tb}$ fully hadronic analysis" at "Higgs Meeting - Extended Higgs Sectors" (04 Jan 2021)
- "HLT emulation in Ultra Legacy MC: validation by TSG" at "PPD General Meeting" (04 Apr 2019)
- "Ultra Legacy MC campaign - Top HLT Validation" at "Trigger Studies Group meeting" (22 May 2019)

Activities within the CMS Collaboration

- Beyond Two Generations (B2G) Monte Carlo contact (Sep 2022 -)
- Extended Higgs sectors (HExtended) Monte Carlo contact (Dec 2020 - Sep 2022)
- Jet flavour tagging maintenance (Apr 2019 - Dec 2020)
- Top Trigger Validation and Data Quality Monitoring (DQM) (Jan 2018 - Sep 2019)

Contents

List of Tables

List of Figures

1 Theoretical background

1.1 Standard model

The Standard Model (SM) is a gauge theory based on the symmetry group $\mathrm{SU(3)_C} \otimes \mathrm{SU(2)_L} \otimes \mathrm{U(1)_Y}$ that summarizes our current understanding of the particle physics, the elementary constituents of the universe and the way they interact through the strong, weak and electromagnetic forces [1–5]. The elementary constituents of our nature are illustrated in Fig. 1.1.

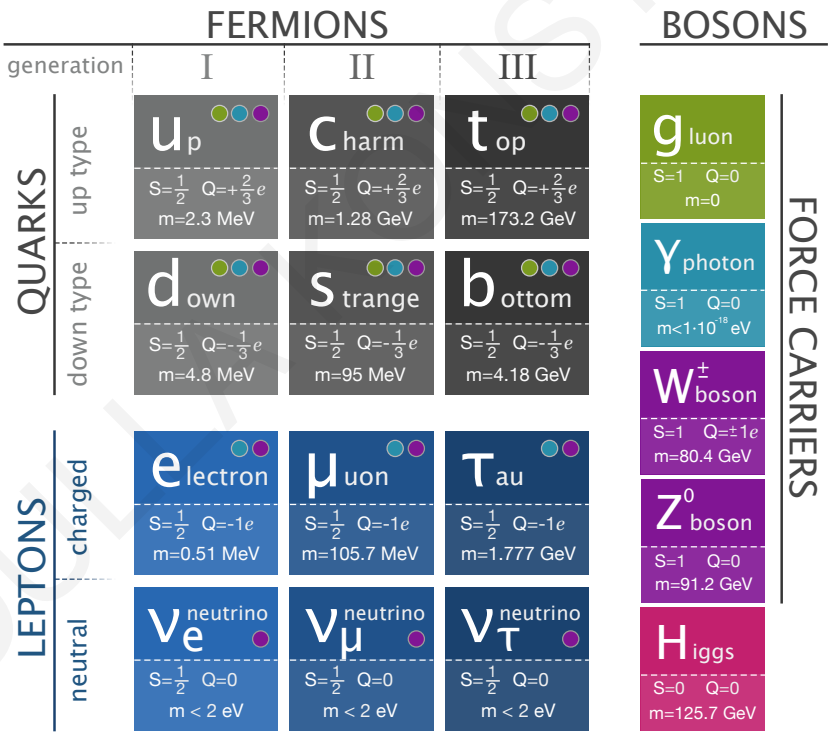


Figure 1.1: The Standard Model elementary particles.

In the SM, matter consists of elementary spin-1/2 particles, the fermions. Fermions are further categorized into leptons and quarks, which are organized in a three-generation structure or three flavors as shown in Fig. 1.1. The three generations share the same properties, excluding their flavor quantum number and their mass which increases from the first generation to the last. Electrically-charged leptons, electron (e), muon (μ) and tau (τ), carry an electric charge equal to the electron charge $|e|$. The electrically-neutral leptons are the neutrinos, which appear in pairs with the three electrically-charged lepton flavors: ν_e, ν_μ and

ν_τ. Pairs of quarks that belong to each generation have electric charge +2/3$|e|$ and -1/3$|e|$. The six quarks are the up-down (u, d), charm-strange (c, s) and top-bottom (t, b). Quarks carry also a color charge (three colors) and they can interact through the strong force. Unlike the six leptons, quarks are not observed as free particles in nature, they are confined by the strong force to form bound states called hadrons. The hadrons carry zero total color charge and a total electric charge that is integer-multiple of the $|e|$. The most common types of hadrons are the mesons that consist of a quark-antiquark pair of the same flavor, and the baryons that are bound states of three quarks. Protons and neutrons are baryons that consist of two up and one down quark (uud) and an up and two down quarks (udd), respectively. Other exotic states of hadrons observed the recent years in particle colliders are tetraquarks and pentaquarks.

For each of the twelve particles, there is a corresponding antiparticle with exactly the same properties but opposite electric charge. Ordinary matter consists of the first-generation fermions u, d, e and ν_e. Other particles can be produced from cosmic rays, nuclear reactions or particle collisions in accelerators.

In the SM, fermions obey the Fermi-Dirac statistics [6]. They are described as left-handed doublet fields or right-handed singlet fields with the form:

$$\begin{pmatrix} q_u \\ q_d \end{pmatrix}_L, \quad \begin{pmatrix} \nu_\ell \\ \ell \end{pmatrix}_L, \quad \ell_R, \quad q_{uR} \quad q_{dR} \tag{1.1}$$

where q_u and q_d are the up-type and down-type quarks, ν_ℓ and ℓ are the neutrino and lepton of a specific flavor and the indices L and R refer to the left and right chirality. The Lagrangian of a free Dirac-fermion is given by:

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

where ψ and $\bar{\psi}$ are the Dirac spinor field and its conjugate, m is the particle's mass, ∂_μ is the partial derivative in the four space-time coordinates and γ^μ are the Dirac matrices:

$$\gamma^0 = \begin{pmatrix} \mathrm{I} & 0 \\ 0 & -\mathrm{I} \end{pmatrix}, \qquad \gamma^i = \begin{pmatrix} 0 & \sigma^i \\ -\sigma^i & 0 \end{pmatrix}, \tag{1.3}$$

In quantum field theory [7], fermions interact with the exchange of spin-1 particles, the gauge bosons. The electromagnetic interactions between electrically charged particles are mediated by the massless photon γ and are described by quantum electrodynamics (QED) [8]. Weak charged- and neutral-current interactions are mediated with the exchange of a charged $W^\pm$ or a neutral Z^0 boson. All twelve fermions can experience weak interactions as they carry the charge of the weak force, the isospin. Strong interactions are described

by the quantum chromodynamic theory (QCD) [9, 10], with the exchange of a massless gluon g between particles with color charge. The fourth fundamental force is gravity, which is negligible in particle interactions and it is not yet described by any quantum theory. The relative strength of the interactions are characterized by the coupling constants. Table 1.1 shows the relative strength of the four forces with respect to the strong coupling, at subatomic level.

Table 1.1: The four fundamental forces and their relative strength at subatomic level.

Force	Gauge Boson	Coupling constant
Strong	gluon (g)	1
Electromagnetic	photon (γ)	$\frac{1}{137}$
Weak	$W^{\pm}$, Z^0 bosons	10^{-6}
Gravity	Graviton?	10^{-39}

In the 1060's S.L. Glashow [11], S. Weinberg [12] and A. Salam [13] developed a theory that showed that electromagnetic and weak forces can be unified as a single electroweak theory. In 1964 R. Brout, F. Englert [14] and P. Higgs [15] introduced a mechanism that generates the masses of the fermions and the gauge bosons of the weak interaction through the spontaneous electroweak symmetry breaking and gives rise to a scalar (spin-0) particle, the Higgs boson.

1.1.1 Quantum chromodynamics

The QCD is a non-abelian theory based on the $SU(3)_C$ symmetry and describes the strong interaction [16]. The quarks are represented as triplets in the color space ψ_f^{α} where α is the color number and f is the quark flavor. The Lagrangian is given by:

$$\mathcal{L}_{QCD} = \sum_f \bar{\psi_f}(i\gamma^{\mu}D_{\mu} - m_f)\psi_f - \frac{1}{4}G_i^{\mu\nu}G_{\mu\nu}^i \tag{1.4}$$

where the first term is the free Dirac-fermion Lagrangian and the second is the gluon kinetic term. The covariant derivative D_{μ} ensures invariance under local $SU(3)_C$ transformations:

$$D_{\mu}\psi_f(x) = \left[\partial_{\mu} + ig_s\frac{\lambda_i}{2}G_{\mu}^i(x)\right]\psi_f(x) \tag{1.5}$$

where $\frac{\lambda_i}{2}$ are the eight generators of the $SU(3)_C$ and G_{μ}^i the eight gluon fields. The strong coupling constant g_s shows that all the quark flavors interact with gluons with the same strength. In the kinetic term, the gluon-field tensor is:

$$G_{\mu\nu}^i(x) = \partial_{\mu}G_{\nu}^i - \partial_{\nu}G_{\mu}^i - g_s f_{abc}G_{\mu}^b G_{\nu}^c. \tag{1.6}$$

The last term is a result of the non-commutative symmetry and gives rise to gluon self-interaction terms which are responsible for many QCD properties such as the asymptotic freedom. Asymptotic freedom describes the behavior of the strong interaction, which increases at larger distances or lower energies and confines quarks into bound states of colorless hadrons.

The absence of quadratic gluon-field terms in the Lagrangian indicates massless gluons. Such terms are forbidden since they violate the $SU(3)_C$ symmetry. Expanding Eq. (1.4), quark-gluon interactions are described by:

$$\mathcal{L}_{int} = -g_s G^i_\mu \sum_f \bar{\psi}^\alpha_f \gamma^\mu \left(\frac{\lambda_i}{2}\right)_{\alpha\beta} \psi^\beta_f \tag{1.7}$$

1.1.2 Electroweak interactions

The electroweak theory is based on the symmetry $SU(2)_L \otimes U(1)_Y$, where L implies that the weak $SU(2)_L$ interaction acts on the left-handed doublet fields only and Y denotes the weak hypercharge. The free Dirac-fermion Lagrangian is:

$$\mathcal{L}_{free} = \sum_{j=1}^{3} i\bar{\psi}_j(x)\gamma^\mu D_\mu \psi_j(x) \tag{1.8}$$

where j refers to the left-handed and right-handed fields of a specific flavor:

$$\psi_1(x) = \begin{pmatrix} u \\ d \end{pmatrix}_L, \quad \psi_2(x) = u_R, \quad \psi_3(x) = d_R \tag{1.9}$$

The Lagrangian does not include mass terms as it would mix left- and right-handed fields. The covariant derivative D_μ in Eq. (1.8) includes the four different gauge bosons and it transforms as:

$$\begin{aligned} D_\mu\psi_1(x) &= \left[\partial_\mu + ig\widetilde{W_\mu}(x) + ig' y_1 B_\mu(x)\right]\psi_1(x) \\ D_\mu\psi_2(x) &= \left[\partial_\mu + ig' y_2 B_\mu(x)\right]\psi_2(x) \\ D_\mu\psi_3(x) &= \left[\partial_\mu + ig' y_3 B_\mu(x)\right]\psi_3(x) \end{aligned} \tag{1.10}$$

where y_i are the hypercharges and g, g' the couplings of $SU(2)_L$ and $U(1)_Y$. The corresponding fields are the B_μ and $\widetilde{W_\mu}$ were:

$$\widetilde{W_\mu} \equiv \frac{\sigma_i}{2} W^i_\mu = \frac{1}{2}\begin{pmatrix} W^3_\mu & W^1_\mu - iW^2_\mu \\ W^1_\mu + iW^2_\mu & -W^3_\mu \end{pmatrix} \tag{1.11}$$

Combining Eq. (1.8) and Eq. (1.10), the $\mathcal{L}_{free}$ becomes:

$$\mathcal{L}_{free} = \sum_{j=1}^{3} i\bar{\psi}_j \gamma^\mu \partial_\mu \psi_j + g\bar{\psi}_1 \gamma^\mu \widetilde{W}_\mu \psi_1 - g' B_\mu \sum_{j=1}^{3} y_j \bar{\psi}_j \gamma^\mu \psi_j \tag{1.12}$$

The second and third terms describe interactions of the fermions with the gauge bosons. Charged current interactions arise from the second term, where the left-handed fermions interact with the $W^\pm$ boson fields:

$$W_\mu^\pm = \frac{W_\mu^1 \mp i W_\mu^2}{\sqrt{2}} \tag{1.13}$$

Neutral-current interactions with the W_μ^3 and B_μ neutral fields arise from the second and third terms. Since the photon must interact with all fermion chiralities with the same strength $g' y_i$, it cannot be represented by the B_μ. The Z^0 and γ mediators are expressed as combinations of the two fields using a rotational transformation:

$$Z_\mu = W_\mu^3 \cos\theta_W - B_\mu \sin\theta_W \tag{1.14}$$

$$A_\mu = W_\mu^3 \sin\theta_W + B_\mu \cos\theta_W \tag{1.15}$$

Equation (1.12) leads to the QED Lagrangian with the following conditions for the electromagnetic coupling g_e and the weak hypercharge:

$$g \sin\theta_W = g' \cos\theta_W = g_e \tag{1.16}$$

$$Y = Q - T_3 \tag{1.17}$$

Equation (1.17), known as the Gell-Mann-Nishijima formula, relates the weak isospin Y, electric charge Q and $T_3 \equiv \sigma_3/2$. The physical Z_μ and A_μ bosons can be then expressed in terms of the coupling constants g, g':

$$Z_\mu = \frac{g W_\mu^3 - g' B_\mu}{\sqrt{g^2 + g'^2}} \tag{1.18}$$

$$A_\mu = \frac{g' W_\mu^3 + g B_\mu}{\sqrt{g^2 + g'^2}} \tag{1.19}$$

The gauge-invariant kinetic Lagrangian term is given by

$$\mathcal{L}_{kin} = -\frac{1}{4} B_{\mu\nu} B^{\mu\nu} - \frac{1}{4} W_{\mu\nu}^i W_i^{\mu\nu} \tag{1.20}$$

and the tensors $B_{\mu\nu}$ and $W^i_{\mu\nu}$ have the following form:

$$B_{\mu\nu} \equiv \partial_\mu B_\nu - \partial_\nu B_\mu \tag{1.21}$$

$$W^i_{\mu\nu} \equiv \partial_\mu W^i_\nu - \partial_\nu W^i_\mu - g\epsilon^{ijk} W^j_\mu W^k_\nu. \tag{1.22}$$

The Lagrangian in Eq. (1.20) does not include mass terms of the gauge bosons as they would violate the gauge symmetry. The presence of quadradic terms of $W^i_{\mu\nu}$ in Eq. (1.22) gives rise to self-interactions between the gauge fields.

1.1.3 The Brout-Englert-Higgs mechanism

The unification of the weak and electromagnetic interactions, proposed about a century after Maxwell's unified theory of the electric and magnetic forces, was a breakthrough in modern physics. It provided a complete picture of the interaction of fermions through the weak and electromagnetic force, as well as self-interactions of the gauge bosons. However, the theory predicts massless fermions and γ, $W^\pm$ and Z^0 gauge bosons.

According to the Brout-Englert-Higgs mechanism (or Higgs mechanism), particles acquire mass through their interaction with the Higgs field, which gives rise to a scalar particle, the Higgs boson, through a spontaneous symmetry breaking of the $SU(2)_L \otimes U(1)_Y$.

A spontaneous symmetry breaking occurs when the symmetry of a system is not in its ground state. The Lagrangian of such system must be invariant under a local gauge symmetry and have degenerate states with minimum potential energy. Thus, the symmetry breaks spontaneously when the system goes to one of the minima without the action of any external force. Such potential is described by:

$$V(\phi) = \mu^2\phi^2 + \lambda\phi^4 \tag{1.23}$$

with $\lambda > 0$ to ensure an absolute minimum. For $\mu^2 < 0$ the potential is shown in Fig. 1.2 and has infinite minima at

$$\phi_0 = \sqrt{-\frac{\mu^2}{\lambda}} \equiv u \tag{1.24}$$

The $SU(2)_L$ doublet of complex scalar fields ϕ is given by:

$$\phi(x) = \begin{pmatrix} \phi^+(x) \\ \phi^0(x) \end{pmatrix} = \begin{pmatrix} \phi_1 + i\phi_2 \\ \phi_3 + i\phi_4 \end{pmatrix}. \tag{1.25}$$

After the symmetry breaking only the neutral scalar can acquire a vacuum expectation value u. The selection of an arbitrary vacuum will break the $SU(2)_L \otimes U(1)_Y$ symmetry to the electromagnetic $U(1)_{QED}$ symmetry. Perturbations around the vacuum can be parameterized

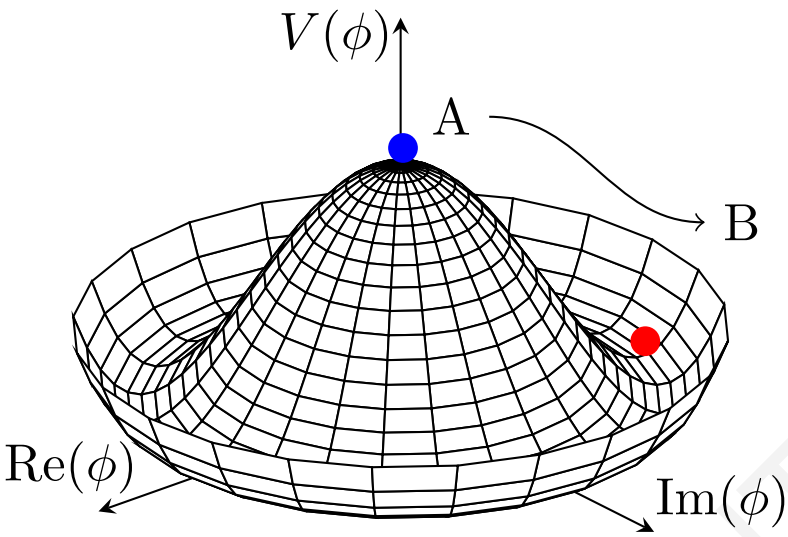


Figure 1.2: The Higgs potential.

based on four real fields, $\theta_1(x)$, $\theta_2(x)$, $\theta_3(x)$ and $H(x)$ as follows:

$$\phi(x) = \frac{1}{\sqrt{2}} e^{i\frac{\sigma_i}{2}\theta_i(x)} \begin{pmatrix} 0 \\ u + H(x) \end{pmatrix} \tag{1.26}$$

The Lagrangian is then given by:

$$\mathcal{L}_{scalar} = (D_\mu \phi)^\dagger (D^\mu \phi) - \mu^2 \phi^\dagger \phi - \lambda (\phi^\dagger \phi)^2 \tag{1.27}$$

and the covariant derivative $D_\mu \phi$ gives:

$$(D_\mu \phi) = \left[\partial_\mu + i(g\widetilde{W}_\mu + g' B_\mu)\right] \phi = \partial_\mu \phi + \frac{(u+H)}{2\sqrt{2}} \begin{pmatrix} g(W_\mu^1 - iW_\mu^2) \\ i(-gW_\mu^3 + g' B_\mu) \end{pmatrix} \tag{1.28}$$

$$(D_\mu \phi) = \partial_\mu \phi + \frac{(u+H)}{2\sqrt{2}} \begin{pmatrix} \sqrt{2} W_\mu^+ \\ -i\sqrt{g^2 + g'^2} Z_\mu \end{pmatrix}$$

From Eq. (1.27) and Eq. (1.28), the Lagrangian takes the form:

$$\mathcal{L}_{scalar} = \frac{1}{2} \partial_\mu H \partial^\mu H + \frac{(u+H)^2}{8} \left[g^2 W_\mu^\dagger W^\mu + (g^2 + g'^2) Z_\mu Z^\mu + 0 A_\mu A^\mu \right] - V(\phi) \tag{1.29}$$

Due to the local invariance under $SU(2)_L$ rotations, the Lagrangian is independent of the $\theta_i(x)$ fields. Based on the J. Goldstone's theorem [17], the spontaneous breaking of a continuous symmetry gives rise to a massless spin-0 particle for each broken generator, the so-called Goldstone bosons. The three Goldstone bosons $\theta_i(x)$ are absorbed by the $W^\pm$ and

Z^0 bosons which acquire an additional degree of freedom (d.o.f). The quadratic terms of the physical $W^{\pm}$ and Z^0 bosons, determine their masses:

$$M_{W^{\pm}} = \frac{ug}{2} \tag{1.30}$$

$$M_{Z^0} = \frac{u\sqrt{g^2 + g'^2}}{2} \tag{1.31}$$

The absence of a quadratic $A_\mu A^\mu$ term, states a massless photon. The Higgs field gives rise to a scalar particle H, the Higgs boson. The expansion of Eq. (1.29) includes interaction terms between the gauge bosons and the Higgs field. These interactions are proportional to the mass or the squared mass of the gauge bosons.

The non-zero mass of the fermions is also a consequence of the Higgs mechanism. The gauge-invariant Lagrangian of the interaction of leptons to the Higgs field has the form:

$$\mathcal{L}_{\ell\ mass} = -g_\ell[\bar{\psi}_L \phi \psi_R + \bar{\psi}_R \bar{\phi} \psi_L] \tag{1.32}$$

where ψ_L and ψ_R are the left-handed doublet and right-handed singlet respectively and g_ℓ is the coupling constant of a lepton to the Higgs field, the Yukawa coupling. After the spontaneous symmetry breaking Eq. (1.32) gives:

$$\mathcal{L}_{\ell\ mass} = -\frac{g_\ell}{\sqrt{2}} \left[\begin{pmatrix} \bar{\nu}_\ell & \bar{\ell} \end{pmatrix}_L \begin{pmatrix} 0 \\ u+H \end{pmatrix} \ell_R + \bar{\ell}_R \begin{pmatrix} 0 & u+H \end{pmatrix} \begin{pmatrix} \nu_\ell \\ \ell \end{pmatrix}_L \right] \tag{1.33}$$

$$\begin{aligned} &= -\frac{g_\ell}{\sqrt{2}} \left[\bar{\ell}_L (u+H) \ell_R + \bar{\ell}_R (u+H) \ell_L \right] \\ &= -\frac{g_\ell (u+H)}{\sqrt{2}} \left[\bar{\ell}_L \ell_R + \bar{\ell}_R \ell_L \right] = -\frac{g_\ell (u+H)}{\sqrt{2}} \bar{\ell}\ell \\ &= -\frac{g_\ell u}{\sqrt{2}} \bar{\ell}\ell + -\frac{g_\ell H}{\sqrt{2}} \bar{\ell}\ell \end{aligned} \tag{1.34}$$

where the first and second term show the mass and the lepton-Higgs boson interaction respectively. As in the case of the weak-gauge bosons, the coupling of the Higgs field with the fermions is proportional to their mass. The Lagrangian in Eq. (1.32) gives mass to the charged fermions but leaves the neutrinos massless. The same Lagrangian can be used to describe the interaction of the Higgs field with the down-type quarks. For up-type quarks, the conjugate doublet field ϕ_c is used which is defined as:

$$\phi_c = i\sigma_2 \phi^* = -\frac{1}{2} \begin{pmatrix} u+H \\ 0 \end{pmatrix} \tag{1.35}$$

and the corresponding Lagrangian term is:

$$\mathcal{L}_{up\ mass} = -g_{up}\bar{\psi}_L\phi_c\psi_R + h.c \tag{1.36}$$

where h.c stands for hermitian conjugate.

1.1.4 The discovery of the Higgs boson

The Higgs boson was discovered in July 2012 by the ATLAS [18] and CMS [19] collaborations and it is one of the most significant discoveries of the century as it completes the SM of particle physics. The two experiments combined results from different decay modes of the Higgs boson, with the most sensitive ones being the decay into two photons $H \to \gamma\gamma$ or to two Z bosons in the four-lepton final state $H \to ZZ \to 4\ell$. The decay into a pair of vector bosons with a total electric charge equal to zero indicates the production of a neutral particle. The $\gamma\gamma$ decay is a sign of a spin-0 boson, as dictated by the Landau-Yang theorem [20, 21] that states that spin-1 particles cannot decay into massless spin-1 bosons. This is because the two photons are produced in opposite directions resulting in a total orbital angular momentum equal to zero. Based on their helicities, photons can be either parallel or antiparallel and therefore, the symmetric state of the two photons can give a total angular momentum equal to zero or two. The spin-2 hypothesis has been excluded experimentally with a confidence level exceeding 99.9% [22].

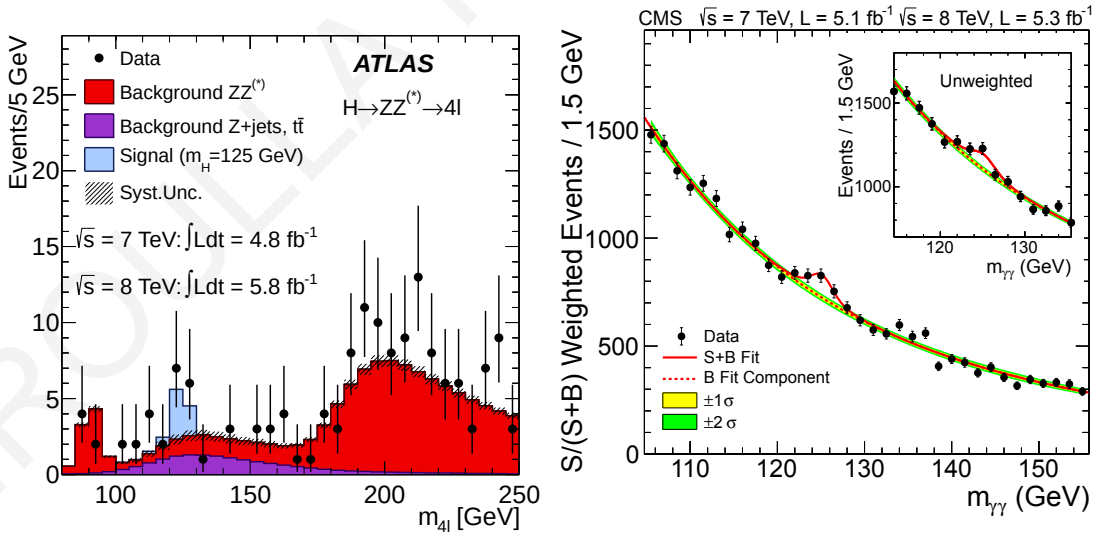


Figure 1.3: Observation of the Higgs boson. Left: Results from the ATLAS experiment in the decay four-lepton final state [18]. Right: Results from the CMS experiment in the diphoton final state [19].

To date, all the measurements of the newly discovered particle are compatible with the properties of the Higgs boson predicted by the SM. Figure 1.4 displays the most recent measurements of the couplings of the Higgs boson to the SM particles by the CMS experiment [23]. The ratio with respect to the expected values agrees with the SM predictions.

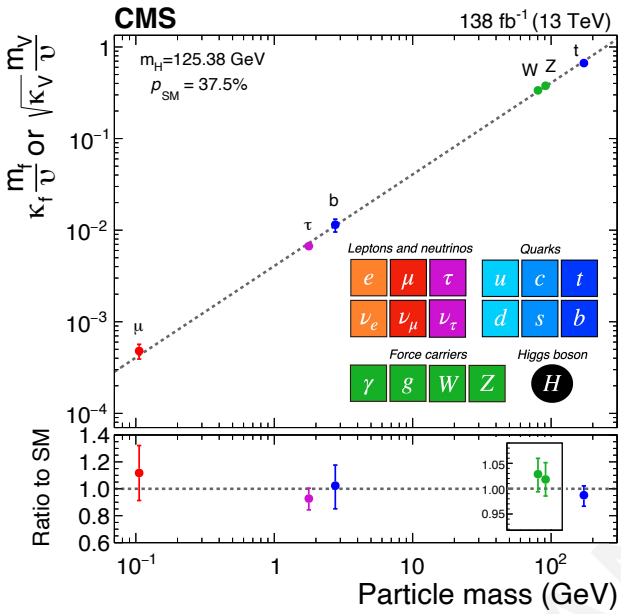


Figure 1.4: CMS measurements of the coupling constants between the Higgs boson and SM particles. The κ_f and κ_V are the coupling modifiers to fermions and heavy gauge bosons respectively which are equal to one in the SM [23].

In 2013 P. Higgs and F. Englert, were awarded the Nobel Prize in Physics for the theoretical prediction of the Higgs mechanism.

1.2 Physics beyond the SM

The SM is one of the greatest achievements of modern physics which provides precise predictions of the experimental observations. However, there are still many unanswered questions arising from the limitations of the model to explain various experimental measurements. Many theories beyond the Standard Model (BSM) have been developed, aiming to shed light on the mysteries of nature. Some of them, extend their Higgs sector with additional multiplet fields. The minimal possible extension is the introduction of a second Higgs-doublet and is discussed in Section 1.2.2.

1.2.1 Limitations of the SM

The lack of a quantum theory that describes the gravitational force at subatomic level is one of the major shortcomings of the SM. Additionally, according to the CPT symmetry model [24] first proposed by M. A. Markov [25], the Big Bang produced equal amounts of matter and antimatter. The dominance of matter over antimatter in the observable universe, also known as matter-antimatter or baryon asymmetry may be a result of the violation of charge conjugation and parity symmetries (CP violation). The CP violation needs to be much larger than what is observed so far, so the source of this extra CP violation is not known.

In the SM, quantum loop corrections to the Higgs boson mass from virtual particle-antiparticle loop processes are extremely large. Such process is shown in Fig. 1.5 for top-antitop quark pairs. The corrections are expressed in terms of the loop-momentum cut-off Λ:

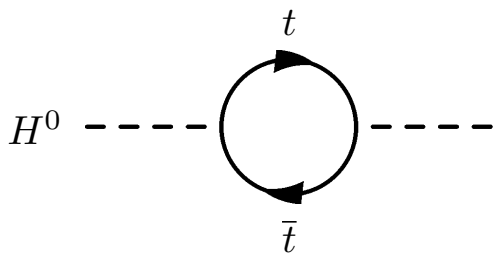


Figure 1.5: Loop corrections to the Higgs boson mass from top quarks.

$$(\Delta m_H^2)^{top} = -\frac{3}{8\pi^2}\lambda_t^2\Lambda^2 \tag{1.37}$$

where λ_t is the top quark-to-Higgs field Yukawa coupling. For Λ of the order of the Planck scale (10^{19}) where the SM is expected to be valid, the Higgs-boson mass diverges from the electroweak scale. This phenomenon is described as the hierarchy problem. BSM theories introduce new particles that cancel these unnaturally large corrections, with contributions of opposite signs.

Another big limitation of the SM is the assumption of massless neutrinos. This contradicts experimental observations of neutrino oscillations that imply non-zero neutrino masses. The first evidence of neutrino flavor conversion arose from the "solar neutrino problem", describing the anomalously low measured flux of ν_e arriving from the Sun and the "atmospheric neutrino anomaly" that refers to the deviation of the observed muon-to-electron neutrinos flux ratio which is expected to be $\sim$2. The discovery of neutrino oscillations with the detection of solar neutrinos was announced by the Sudbury Neutrino Observatory (SNO) Collaboration [26, 27] with simultaneous measurement of the relative rate of neutrino-deuteron reactions producing a pair of protons, mostly occurring with electron-neutrinos, and neutrino-deuteron producing a proton and a neutrino of all flavors. The results showed consistency with the predicted solar-neutrino flux and conversion of about two-thirds of the electron-neutrinos into muon- or tau-neutrinos.

Evidence of neutrino oscillations in the context of atmospheric neutrinos from cosmic rays was presented by the Super-Kamiokande Collaboration [28]. The experiment detected electron and muon neutrinos through Cherenkov radiation. A deficit in the number of up-going high energy muon-neutrinos was observed, indicating that up-going neutrinos were created at the opposite side of the Earth and converted into tau-neutrinos before reaching the detector. In 2015, A. B. McDonald (SNO) and T. Kajita (Super-Kamiokande) were awarded the Nobel Prize in Physics for the discovery of neutrino oscillations [29].

The existence of dark matter (DM) and dark energy (DE) is undoubted evidence of new

physics [30–32]. The existence of DM arose first from the observation that the rotational velocity of luminous objects is higher than the expected of an object experiencing only the gravitational force from other visible objects. Furthermore, supernova measurements showed that the universe's expansion is accelerating, implying the existence of a form of dark energy with negative pressure. According to the Lambda cold dark matter (ΛCDM) model, only 5% of the Universe forms ordinary baryonic matter, while 27% forms the DM and the remaining 68% is DE. Many theories introduce hypothetical candidates of the DM mediator, including primordial black holes, massless axions, sterile $SU(2)\otimes U(1)_Y$ singlet neutrinos and weakly interacting massive particles (WIMP) with a mass between 10 GeV up to a few TeV.

1.2.2 Two-Higgs-doublet-models

The two-Higgs-doublet-models (2HDM) is the simplest extension of the SM which introduces a second $SU(2)_L$ doublet of complex fields [33]. The motivation behind the 2HDM is that the addition of a scalar doublet can produce CP-violation in the Higgs sector which could be a solution to the problem of baryon asymmetry. Additionally, specific 2HDM predict massless axions, a potential candidate of dark matter.

The most general scalar potential for two $SU(2)_L$ doublet scalar fields with hypercharge +1 is given by:

$$\begin{aligned} V &= m_{11}^2\Phi_1^\dagger\Phi_1 + m_{22}^2\Phi_2^\dagger\Phi_2 - m_{12}^2(\Phi_1^\dagger\Phi_2 + \Phi_2^\dagger\Phi_1) \\ &+ \frac{\lambda_1}{2}(\Phi_1^\dagger\Phi_1)^2 + \frac{\lambda_2}{2}(\Phi_2^\dagger\Phi_2)^2 + \lambda_3(\Phi_1^\dagger\Phi_1)(\Phi_2^\dagger\Phi_2) + \lambda_4(\Phi_1^\dagger\Phi_2)(\Phi_2^\dagger\Phi_1) \\ &+ \left[\frac{\lambda_5}{2}(\Phi_1^\dagger\Phi_2)^2 + \lambda_6(\Phi_1^\dagger\Phi_1)(\Phi_1^\dagger\Phi_2) + \lambda_7(\Phi_2^\dagger\Phi_2)(\Phi_1^\dagger\Phi_2) + h.c\right] \end{aligned} \tag{1.38}$$

where the mass matrix elements m_{11}, m_{22} and Higgs-self couplings λ_1, λ_2, λ_3, λ_4 are real parameters and m_{12}, λ_5,λ_6,λ_7 are complex, resulting to 14 parameters. The vacuum expectation values of the two doublets Φ_1, Φ_2 are

$$\langle\Phi_1\rangle_0 = \begin{pmatrix} 0 \\ \frac{v_1}{\sqrt{2}} \end{pmatrix}, \quad \langle\Phi_2\rangle_0 = \begin{pmatrix} 0 \\ \frac{v_2}{\sqrt{2}} \end{pmatrix} \tag{1.39}$$

Perturbations around the minima introduce eight fields, $\phi_{12}^\pm$, ρ_{12} and η_{12}. The two complex SU(2) scalar fields can be written:

$$\Phi_\alpha = \begin{pmatrix} \phi_\alpha^\pm \\ \frac{(v_\alpha+\rho_\alpha+i\eta_\alpha)}{\sqrt{2}} \end{pmatrix}, \quad \alpha = 1, 2. \tag{1.40}$$

The quadratic terms in the 2HDM potential describe the masses of the eight fields.

From Eq. (1.39) the mass terms are

$$\mathcal{L}_{\phi^\pm\ mass} = [m_{12}^2 - (\lambda_4+\lambda_5)v_1v_2](\phi_1^-, \phi_2^-) \begin{pmatrix} \frac{v_2}{v_1} & -1 \\ -1 & \frac{v_1}{v_2} \end{pmatrix} \begin{pmatrix} \phi_1^- \\ \phi_2^+ \end{pmatrix} \tag{1.41}$$

$$\mathcal{L}_{\eta\ mass} = \frac{m_A^2}{v_1^2+v_2^2}(\eta_1, \eta_2) \begin{pmatrix} v_2^2 & -v_1v_2 \\ -v_1v_2 & v_1^2 \end{pmatrix} \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} \tag{1.42}$$

$$\mathcal{L}_{\rho\ mass} = -(\rho_1, \rho_2) \begin{pmatrix} m_{12}^2\frac{v_2}{v_1} + \lambda_1 v_1^2 & -m_{12}^2 + \lambda_{345}v_1v_2 \\ -m_{12}^2 + \lambda_{345}v_1v_2 & m_{12}^2\frac{v_1}{v_2} + \lambda_2 v_2^2 \end{pmatrix} \begin{pmatrix} \rho_1 \\ \rho_2 \end{pmatrix} \tag{1.43}$$

where $\lambda_{345} = \lambda_3 + \lambda_4 + \lambda_5$ and the three 2×2 mass-squared matrices defined as $\mathcal{M}_\phi^\pm$, $\mathcal{M}_\eta$ and $\mathcal{M}_\rho$. The physical particles are mass eigenstates and can be obtained with simultaneous diagonalization of the mass-squared matrices. The mass-square matrix of the scalars can be diagonalized with a rotation by an angle α while the mass-squared matrices of the charged scalar and the pseudoscalar fields can be simultaneously diagonalized with a rotation by an angle β. The parameters α, β play a special role in the context of electroweak symmetry breaking because they determine the interactions of the Higgs fields with the vector bosons and the fermions. The tangent of angle β is defined as:

$$\tan\beta \equiv \frac{v_2}{v_1} \tag{1.44}$$

The physical mass eigenstates are given by:

$$\begin{pmatrix} \cos\beta & \sin\beta \\ -\sin\beta & \cos\beta \end{pmatrix} \begin{pmatrix} \phi_1^\pm \\ \phi_2^\pm \end{pmatrix} = \begin{pmatrix} \phi_1^\pm\cos\beta + \phi_2^\pm\sin\beta \\ -\phi_1^\pm\sin\beta + \phi_2^\pm\cos\beta \end{pmatrix} = \begin{pmatrix} G^\pm \\ H^\pm \end{pmatrix} \tag{1.45}$$

$$\begin{pmatrix} \cos\beta & \sin\beta \\ -\sin\beta & \cos\beta \end{pmatrix} \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \begin{pmatrix} \eta_1\cos\beta + \eta_2\sin\beta \\ -\eta_1\sin\beta + \eta_2\cos\beta \end{pmatrix} = \begin{pmatrix} G^0 \\ A^0 \end{pmatrix} \tag{1.46}$$

$$\begin{pmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{pmatrix} \begin{pmatrix} \rho_1 \\ \rho_2 \end{pmatrix} = \begin{pmatrix} \rho_1\cos\alpha + \rho_2\sin\alpha \\ -\rho_1\sin\alpha + \rho_2\cos\alpha \end{pmatrix} = \begin{pmatrix} H^0 \\ h^0 \end{pmatrix} \tag{1.47}$$

The diagonalization of $\mathcal{M}_{\phi}^{\pm}$, $\mathcal{M}_{\rho}$ and $\mathcal{M}_{\eta}$ lead to the following masses:

$$m_{G^{\pm}}^2 = 0 \tag{1.48}$$

$$m_{H^{\pm}}^2 = \left[\frac{m_{12}^2}{v_1 v_2} - \frac{\lambda_4 + \lambda_5}{2}\right](v_1^2 + v_2^2) \tag{1.49}$$

$$m_{G^0}^2 = 0 \tag{1.50}$$

$$m_{A^0}^2 = \left[\frac{m_{12}^2}{v_1 v_2} - 2\lambda_5\right](v_1^2 + v_2^2) \tag{1.51}$$

$$m_{H^0}^2 = \frac{1}{2}\left[\mathcal{M}_{\rho 11}^2 + \mathcal{M}_{\rho 12}^2 + \sqrt{(\mathcal{M}_{\rho 11}^2 - \mathcal{M}_{\rho 22}^2)^2 + 4(\mathcal{M}_{\rho 12}^2)^2}\right] \tag{1.52}$$

$$m_{h^0}^2 = \frac{1}{2}\left[\mathcal{M}_{\rho 11}^2 + \mathcal{M}_{\rho 12}^2 - \sqrt{(\mathcal{M}_{\rho 11}^2 - \mathcal{M}_{\rho 22}^2)^2 + 4(\mathcal{M}_{\rho 12}^2)^2}\right] \tag{1.53}$$

Three out of the eight fields are the massless Goldstone bosons $G^{\pm}$ are G^0. The $G^{\pm}$ and G^0 are "eaten" to give mass to the $W^{\pm}$ and Z^0 bosons respectively, while the remaining five are the physical scalar Higgs bosons. The electroweak symmetry breaking predicts the formation of two charged Higgs bosons ($H^{\pm}$), a light (h^0) and a heavy (H^0) CP-even Higgs boson and one CP-odd (pseudoscalar) (A^0).

1.2.3 Yukawa couplings

One of the problems of the 2HDM is the possibility of the existence of tree-level flavor-changing neutral currents (FCNC) which are mediated by the neutral scalars. This can be explained as follows. The Yukawa couplings between down-type quarks and the Higgs fields are given by:

$$\mathcal{L}_Y = y_{ij}^1 \bar{\psi}_i \psi_j \Phi_1 + y_{ij}^2 \bar{\psi}_i \psi_j \Phi_2 \tag{1.54}$$

and the mass matrix is:

$$M_{ij} = y_{ij}^1 \frac{v_1}{\sqrt{2}} + y_{ij}^2 \frac{v_2}{\sqrt{2}} \tag{1.55}$$

where i, j define the three fermion generations. Unlike SM, in 2HDM the diagonalization of the mass matrix does not lead to flavor diagonal Yukawa couplings y^1, y^2. However, under specific assumptions, FCNC can be eliminated. According to the Paschos-Glashow-Weinberg theorem [34, 35], if fermions with the same quantum numbers couple to the same Higgs multiplets, the FCNC are negated. This means that all right-handed quarks must couple to a single Higgs multiplet. In the 2HDM this can be accomplished by the introduction of discrete or continuous symmetries.

Based on the couplings of the Higgs bosons to fermions, the 2HDM are divided into different categories. In type-I 2HDM all fermions couple to just one Higgs doublet, the Φ_2, by introducing the discrete symmetry $\Phi_1 \to -\Phi_1$. In type-II 2HDM, the introduction of

the discrete symmetry $\Phi_1 \to -\Phi_1$, $d_R^i \to -d_R^i$, $\ell_R^i \to -\ell_R^i$ imposes right-handed up-type fermions to couple to the Φ_2 and right-handed down-type fermions to Φ_1. The Minimal Supersymmetric Standard Model (MSSM) is a special case of a type-II model. In the lepton-specific or type-X 2HDM, the right-handed leptons couple to Φ_1 while right-handed quarks to the Φ_2 enforced by the $\Phi_1 \to -\Phi_1$, $\ell_R^i \to -\ell_R^i$ symmetry. Finally, as in the type-II model, in the flipped or type-Y 2HDM, right-handed up-type and down-type fermions couple to different Higgs doublets but right-handed leptons couple to Φ_2. A summary of the couplings and symmetries of all the different models is given in Table 1.2.

Table 1.2: 2HDM models with the specific couplings of the Higgs-doublets to the right-handed SM fermions and the discrete symmetries introduced to eliminate all FCNC.

	u_R^i	d_R^i	ℓ_R^i	Discrete symmetry
Type-I	Φ_2	Φ_2	Φ_2	$\Phi_1 \to -\Phi_1$
Type-II	Φ_2	Φ_1	Φ_1	$(\Phi_1, d_R^i, \ell_R^i) \to -(\Phi_1, d_R^i, \ell_R^i)$
Type-X	Φ_2	Φ_2	Φ_1	$(\Phi_1, \ell_R^i) \to -(\Phi_1, \ell_R^i)$
Type-Y	Φ_2	Φ_1	Φ_2	$(\Phi_1, d_R^i) \to -(\Phi_1, d_R^i)$

1.2.4 Supersymmetry

Supersymmetry (SUSY) [36, 37] is a theoretical BSM framework that relates bosons with fermions. Each SM fermion or boson corresponds to a SUSY boson or fermion, respectively, called a "superpartner". In an exact SUSY theory, the superpartners have exactly the same properties as the SM particles except for their spin which differs by a half-integer. However, since no fundamental particles with the same masses have been detected, SUSY cannot be an exact theory and is instead considered a broken symmetry. The breaking of the SUSY symmetry shifts the masses of the superpartners to higher values. The SUSY generators "$\mathcal{Q}$" transform the fermions into bosons and vice-versa as follows:

$$\mathcal{Q}|Fermion\rangle = |Boson\rangle, \qquad \mathcal{Q}|Boson\rangle = |Fermion\rangle \tag{1.56}$$

and satisfy an algebra of the following anticommutation and commutation relations:

$$\left\{Q_\alpha, Q^\dagger_{\dot\beta}\right\} = 2\sigma^\mu_{\alpha\dot\beta} P_\mu \tag{1.57}$$

$$\left\{Q_\alpha, Q_\beta\right\} = \left\{Q^\dagger_{\dot\alpha}, Q^\dagger_{\dot\beta}\right\} = 0 \tag{1.58}$$

$$\left[Q_\alpha, P_\mu\right] = \left[Q^\dagger_{\dot\alpha}, P_\mu\right] = 0 \tag{1.59}$$

where the $\alpha, \dot{\beta}$ are spinor indices that act on left-handed and right-handed Weyl spinors respectively. The P_μ is the four-momentum generator of the spacetime translation.

The particles are grouped into supermultiplets in which the number of fermionic d.o.f matches the bosonic d.o.f. The simplest case contains a Weyl spin-1/2 field ψ_i (2 d.o.f) and a SUSY complex scalar field S_i (2 d.o.f) and it is called the "chiral" or "matter" or "scalar" supermultiplet. Another simple case of a supermultiplet contains a spin-1 massless gauge field A_μ^α (2 d.o.f) and its superpartner Weyl spin-1/2 field λ_α (2 d.o.f), the "gaugino". The index α is the gauge index of the associated gauge symmetry group.

SUSY gives answers to many unresolved problems of the SM. It introduces stable and neutral weakly interacting particles which are good mediator candidates of the dark matter. Furthermore, it explains the baryon asymmetry problem and solves the hierarchy problem. In the presence of scalars S with mass m_S that couple to the Higgs boson, quantum loop corrections contribute to the Higgs boson mass corrections described by:

$$\Delta m_H^2 \approx \frac{\lambda_f^2 N_f}{4\pi^2}\left[(m_f^2 - m_S^2)\log\left(\frac{\Lambda}{m_S}\right) + 3m_f^2\log\left(\frac{m_S}{m_f}\right)\right] + \mathcal{O}(\frac{1}{\Lambda^2}) \tag{1.60}$$

where the subscripts f and S refer to the fermions and the SUSY scalars respectively. The N_f refers to the repetition number of the fermion f. As seen in Eq. (1.60) the quadratic terms of Λ that appear in Eq. (1.37) vanish while only a logarithmic dependency is present. Depending on the masses of the superpartners, Δm_H^2 can be small even for values of Λ in the Planck scale. However, for SUSY particles that are too heavy compared to the SM fermion masses, Δm_H^2 is still large and the hierarchy problem preserves. For corrections of the same order as the tree–level Higgs boson mass, the SUSY particles should not exceed the TeV scale.

Minimal Supersymmetric Standard Model

The simplest SUSY model introduces the smallest possible number of new particles and it is called the Minimal Supersymmetric Standard Model (MSSM). MSSM is based on the SM gauge symmetry $SU(3)_C \otimes SU(2)_L \otimes U(1)_Y$ and it is a 2HDM type-II model. For each spin-1 gauge field, it predicts a spin-1/2 gaugino, the bino $\widetilde{B}$ of the $U(1)$, three winos $\widetilde{W}_\alpha$ of the $SU(2)_L$ and eight gluinos $\widetilde{G}_\alpha$ of the $SU(3)_C$. Furthermore, for each SM fermion, there is a left- or right-handed scalar SUSY particle that belongs to the same supermultiplet. These particles are called squarks and sleptons. The two doublet scalar fields H^u, H^d give masses to the up and down fermions respectively through the spontaneous electroweak symmetry breaking, resulting in the five 2HDM Higgs bosons. The superpartners of the Higgs bosons are the spin-1/2 higgsinos $\widetilde{H}^u$, $\widetilde{H}^d$, which mix with the $\widetilde{W}_\alpha$ and $\widetilde{B}$ to generate masses to the physical mass eigenstates, the charginos $\chi^\pm_{1,2}$ and neutralinos $\chi^0_{1,2,3,4}$. The lightest SUSY particle χ^0_1 is a good candidate for dark matter as it is predicted to be stable, electrically

neutral and interact weakly with the SM particles. Table 1.3 summarizes the supermultiplets with the SM particles and the associated MSSM superpartners.

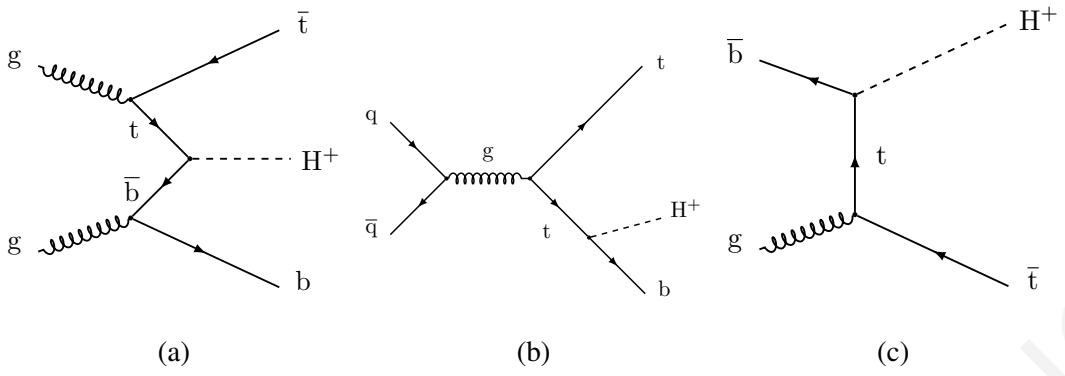
Table 1.3: Supermultiplets in the MSSM.

Supermultiplet	Superfield	Bosonic field	Fermionic field
gluon / gluino	$\hat{G}_a$	G^a_μ	$\widetilde{G}^a_\mu$
W boson / wino	$\hat{W}_a$	W^a_μ	$\widetilde{W}^a_\mu$
B boson / bino	$\hat{B}$	B_μ	$\widetilde{B}_\mu$
slepton / lepton	$\hat{L}$	$(\widetilde{\nu}_L, \widetilde{e}_L)$	$(\nu, e)_L$
	$\hat{E}^c$	$\widetilde{e}^*_R$	$e^\dagger_R$
squark / quark	$\hat{Q}$	$(\widetilde{u}_L, d_L)$	$(u, d)_L$
	$\hat{U}^c$	$\widetilde{u}^*_R$	u^*_L
	$\hat{D}^c$	$\widetilde{d}^*_R$	d^*_L
Higgs / higgsino	$\hat{H}_1$	(H^0_d, H^-_d)	$\widetilde{H}^0_d, \widetilde{H}^-_d$
	$\hat{H}_2$	(H^+_u, H^0_u)	$\widetilde{H}^+_u, \widetilde{H}^0_u$

1.3 Charged Higgs boson production in proton-proton colliders

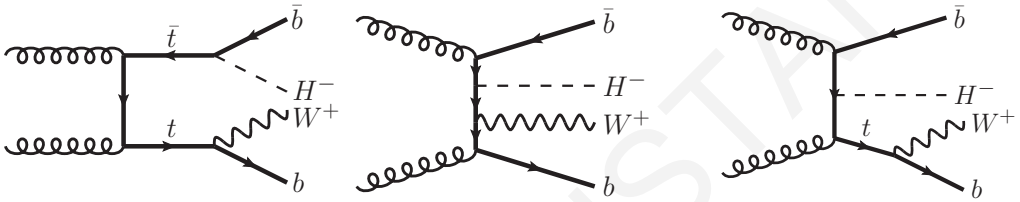
The experimental observation of a $H^\pm$ would be strong evidence of new physics as the SM does not include any electrically charged scalar particles. In proton-proton (pp) colliders, the production of the $H^\pm$ can take place in different modes, depending on the $H^\pm$ mass ($m_{H^\pm}$) [30, 38]. Searches on the $H^\pm$ distinguish a "light $H^\pm$" with a mass less than the difference of the top- and the bottom-quark mass ($m_{H^\pm} < m_t - m_b$), an "intermediate $H^\pm$" with a mass close to the mass of the top quark ($m_{H^\pm} \sim m_t$) and a "heavy $H^\pm$" with a mass greater than the sum of the top- and the bottom-quark mass ($m_{H^\pm} > m_t + m_b$).

In the light $H^\pm$ scenario, the $H^\pm$ arises mainly from the decays of a top quark in association with a bottom quark ($t \to bH^\pm$). In the heavy $H^\pm$ scenario, the dominant production mechanism is in association with a top quark and a down-type quark X, mainly a bottom quark.($pp \to tH^\pm + X$). Depending on whether bottom quarks are involved in the hard scattering, the production can take place through different modes. In the four-flavor scheme (4FS), the bottom quark is considered as heavy and it does not participate in the hard scattering. Thus the production occurs through gluon-gluon fusion ($gg \to tbH^\pm$) or quark-antiquark annihilation, the so-called Higgsstrahlung ($q\bar{q} \to tbH^\pm$). In the five-flavor scheme



(a) (b) (c)

Figure 1.6: Tree-level Feynman diagrams for the heavy charged Higgs boson production in pp collisions. The direct $\mathrm{gg} \to \mathrm{tbH}^{\pm}$ production process (a), the Higgs-strahlung $\mathrm{q\bar{q}} \to \mathrm{tbH}^{\pm}$ process in the 4FS (b) and the gluon splitting process $\mathrm{gb} \to \mathrm{tH}^{\pm}$ process in the 5FS (c).



(a) (b) (c)

Figure 1.7: Tree-level Feynman diagrams of the light $H^{\pm}$ production through the decay of a top quark (a), the intermediate $H^{\pm}$ production through the non-resonant top quark (b) and the heavy $H^{\pm}$ production in association with a top and a bottom quark (c).

(5FS) the bottom quark is treated as a massless parton and can be involved in the initial state. The production in the 5FS takes place also through a gluon and a bottom quark interaction ($gb \to tH^{\pm}$). The Feynman diagrams of the heavy $H^{\pm}$ are illustrated in Fig. 1.6.

In the case of an intermediate $m_{H^{\pm}}$, the production through the non-resonant top quark $pp \to H^{\pm}W^{\mp}bb$ contributes to the interference between the $t \to bH^{\pm}$ and $pp \to tH^{\pm} + X$ production modes. Figure 1.7 shows the Feynman diagrams of the production of a light $H^{\pm}$ from the top-quark decay, the intermediate $pp \to H^{\pm}W^{\mp}bb$ process and the production of the heavy $H^{\pm}$ in association with a top and a bottom quark.

1.3.1 Cross section

The cross section of the light $H^{\pm}$ production is estimated from the top-antitop quark ($\mathrm{t\bar{t}}$) pair production cross section times the branching ratio $t \to bH^{\pm}$. In pp colliders, the $\mathrm{t\bar{t}}$ production occurs through quark-antiquark annihilation or gluon-gluon fusion. For the calculation of the heavy $H^{\pm}$ cross section, the 4FS and 5FS are combined with the Santander matching formula [38–42] to provide a unique cross section in the asymptotic limits $m_{H^{\pm}}/m_b \to 1$ for the 4FS and $m_{H^{\pm}}/m_b \to \infty$ for the 5FS. The combined cross section is given by:

$$\sigma^{matched} = \frac{\sigma^{4FS} + w\sigma^{5FS}}{1+w} \tag{1.61}$$

where w is a weight which depends on the relative importance on the $m_{H^\pm}$:

$$w = \ln \frac{M_{H^\pm}}{m_b} - 2 \tag{1.62}$$

The matched cross section is shown in Fig. 1.8 for different values of $\tan\beta$ and $m_{H^\pm}$.

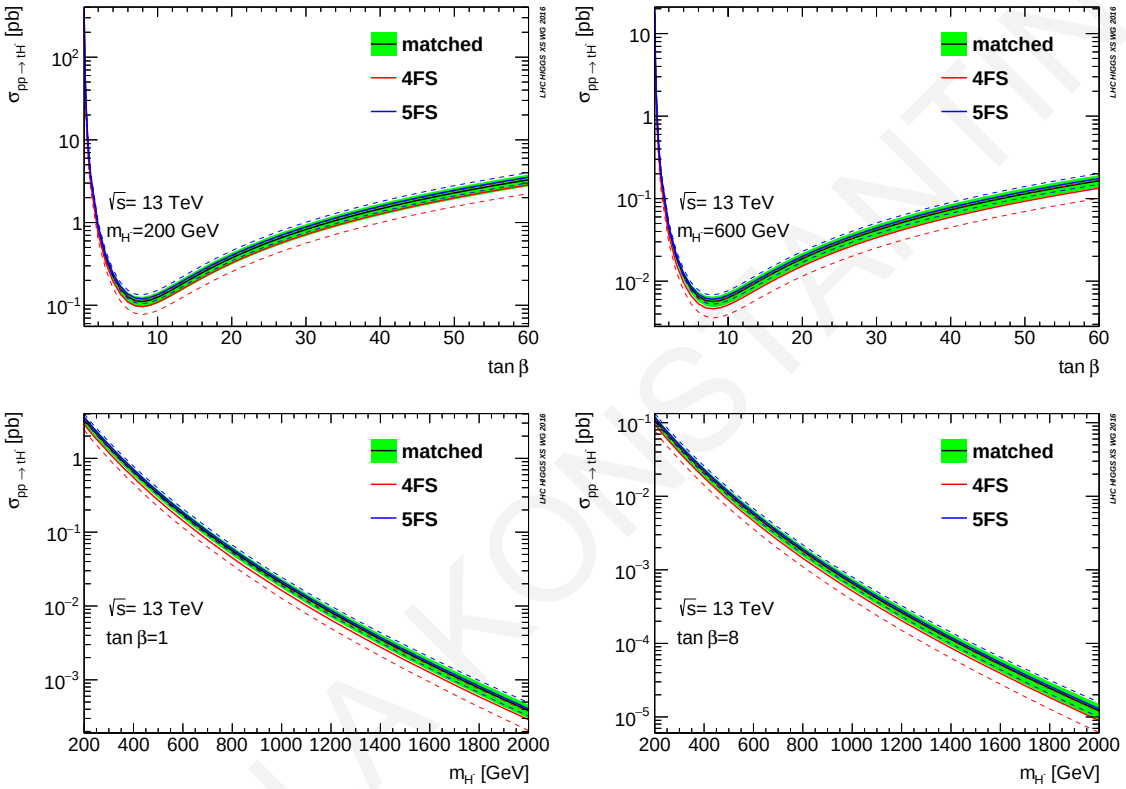


Figure 1.8: Cross section for $tH^\pm + X$ production, after matching the 4FS and 5FS results. The upper row shows the cross section as a function of the $\tan\beta$ for different $m_{H^\pm}$ values. The lower row shows the cross section as a function of the $m_{H^\pm}$ for different values of $\tan\beta$ [38].

In the intermediate $m_{H^\pm}$ region, finite top-quark width effects play an essential role in the estimation of the cross section. In [43] a first-time precise prediction is established.

1.3.2 Charged Higgs boson decay

The $H^\pm$ can decay to fermion-antifermion pairs of the same generation or pairs of bosons including the charged $W^\pm$ and a Z^0, γ or a lighter neutral Higgs boson. Decays to non-SM particles are also possible. The probability of each decay mode is defined as the branching ratio or branching fraction ($\mathcal{B}$) and it depends on the $m_{H^\pm}$ since in specific mass ranges, some decays are not kinematically allowed. Furthermore, as discussed in Section 1.2.3, Higgs boson to fermion couplings are determined by different models. This means that the branching ratio depends also on the theoretical model and the corresponding model parameters.

The decay modes of the $H^{\pm}$ are displayed in Figs. 1.9 and 1.10 as a function of the mass spectrum for the *hMSSM* and M_h^{125} theoretical benchmarks [44–48]. The *hMSSM* scenario, assumes that the lightest CP-even Higgs boson is the SM Higgs boson with mass 125 GeV and the SUSY particles appear with very high masses of the order of the TeV scale. As seen in Fig. 1.9 (left) in the low mass region the main decay channel is $H^{\pm} \to \tau^{\pm}\nu_{\tau}$ while for heavy $H^{\pm}$, the dominant channel is the decay into a top and a bottom quark ($H^{\pm} \to tb$). The branching ratio of $H^{\pm} \to tb$ in the high mass region which is the topic of this work remains large for a wide range of values of the parameter $\tan\beta$ values as seen in Fig. 1.9 (right). In the M_h^{125}, all the superparticles are heavy and the coupling with the light Higgs boson is

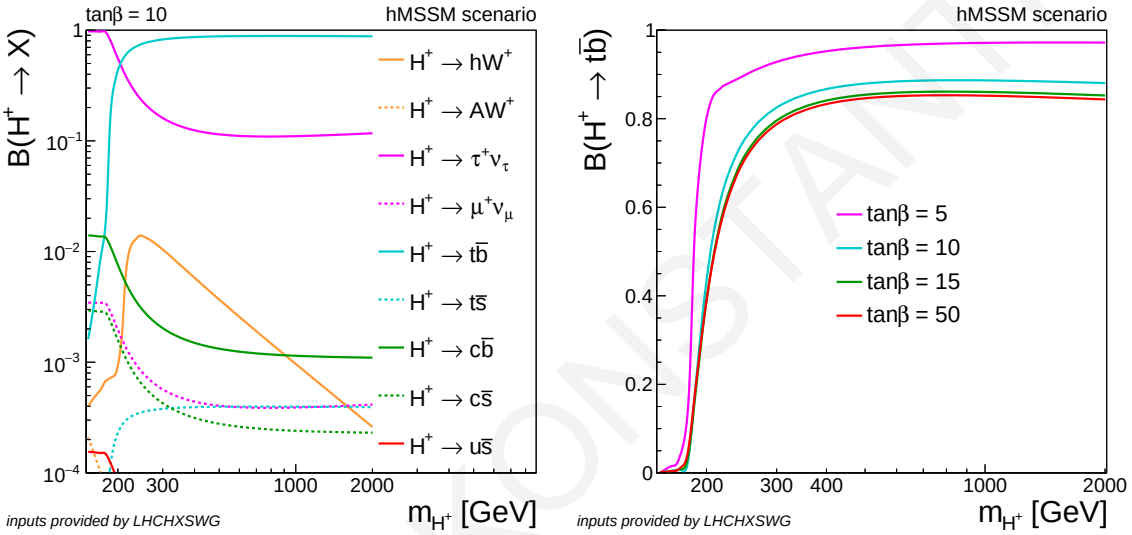


Figure 1.9: Branching ratios of the $H^{\pm}$ decay modes for $\tan\beta = 10$ (left) and branching ratio of $H^{\pm} \to tb$ for different values of $\tan\beta$ (right) in the *hMSSM* benchmark scenario.

negligible. Heavy Higgs bosons with masses up to 2 TeV decay only to SM particles. Above this value, the contribution from SUSY particles becomes significant. The branching ratios in the context of the M_h^{125} scenario are shown in Fig. 1.10 (left). As in the *hMSSM*, the decay of the $H^{\pm}$ to the third-generation fermions dominates for a wide range of $\tan\beta$ values. This is shown in Fig. 1.10 (right) for the $H^{\pm} \to tb$. The largest branching ratio in the light $H^{\pm}$ mass region corresponds to $H^{\pm} \to \tau^{\pm}\nu_{\tau}$ while for higher masses the main decay mode is $H^{\pm} \to tb$.

At the alignment limit $\sin(\beta - \alpha) \to 1$ where decays of the $H^{\pm}$ to a $W^{\pm}$ and neutral scalars h^0, H^0 or A is suppressed, the branching fractions for all the 2HDM types are shown in Fig. 1.11, for $m_{H^{\pm}} = 250$ GeV (heavy $H^{\pm}$ region) and degenerate $m_{H^{\pm}}$, m_{H^0} and m_A [49]. For type-X, the branching fraction to $\tau^{\pm}\nu_{\tau}$ starts to dominate over the tb channel for large values of $\tan\beta$. For the rest types of 2HDM, the tb decay dominates for the entire $\tan\beta$ range.

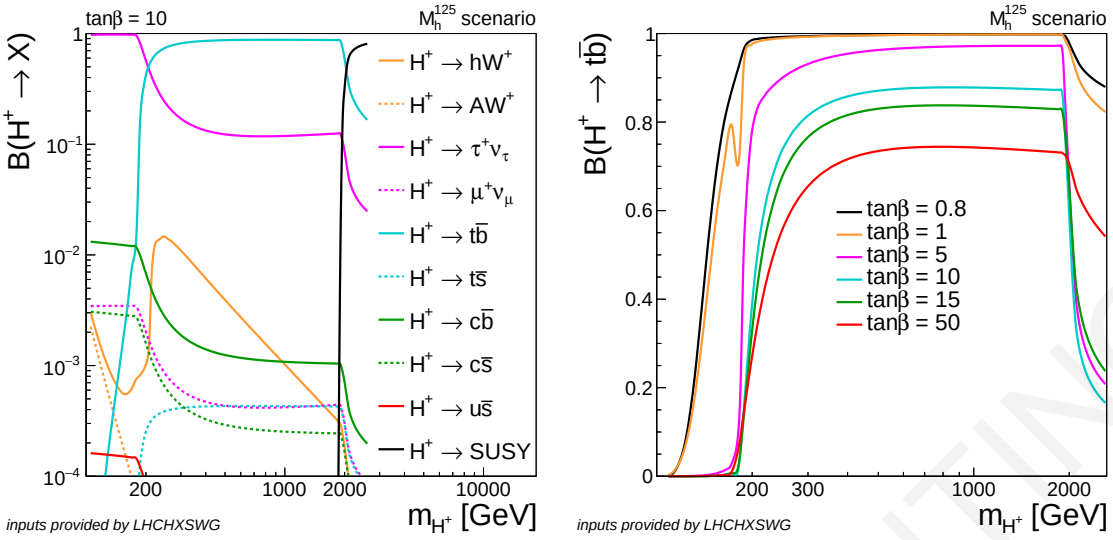


Figure 1.10: Branching ratios of the $H^{\pm}$ decay modes for $\tan\beta$ = 10 (left) and branching ratio of $H^{\pm} \to tb$ for different values of $\tan\beta$ (right) in the M_h^{125} benchmark scenario.

1.3.3 Experimental results

The ATLAS and CMS experiments set upper limits on the production cross section of the $H^{\pm}$ in different decay modes and final states.

Searches for a $H^{\pm}$ with the ATLAS [50] and CMS [51] experiments set upper limits on the $H^{\pm} \to \tau^{\pm}\nu_{\tau}$ in the light and heavy $H^{\pm}$ regimes at a center-of-mass energy of 13 TeV that correspond to an integrated luminosity of $\sim$ 36 fb^{-1}. In the low mass region, ATLAS extracts upper limits on the $\mathcal{B}(t \to bH^{\pm}) \times \mathcal{B}(H^{\pm} \to \tau\nu_{\tau})$ while CMS sets upper limits on the $\sigma_{H^{\pm}} \times \mathcal{B}(H^{\pm} \to \tau\nu_{\tau})$ assuming $\sigma_{H^{\pm}} = 2\sigma_{t\bar{t}}\mathcal{B}(t \to bH^{\pm})(1 - \mathcal{B}(t \to bH^{\pm}))$. The results obtained from the CMS experiment correspond to $\mathcal{B}(t \to bH^{\pm}) \times \mathcal{B}(H^{\pm} \to \tau\nu_{\tau})$ values between 0.36% and 0.08% in the mass range of 80-160 GeV. The results provided by ATLAS range from 0.25% to 0.031% for masses 90-160 GeV. In the heavy $H^{\pm}$ region, ATLAS assumes $\sigma(pp \to tbH^{\pm})$ while in CMS the production mode is not specified. The results of the ATLAS and CMS experiments are shown in Figs. 1.12 and 1.13 respectively.

Searches for a heavy $H^{\pm}$ that decays to a top and a bottom quark have been performed with the CMS experiment in the dilepton, single lepton and hadronic final states [52, 53] at $\sqrt{s}$ = 13 TeV and with an integrated luminosity of $\sim$ 36 fb^{-1}. The combined upper limits on the $\sigma(H^{\pm}tb)\mathcal{B}(H^{\pm} \to tb)$ exclude values above 9.25 pb and 0.005 pb in the range of 200-3000 GeV. Figure 1.14 (left) shows the combined limits along with the median expected limits of each final state. The single-lepton final state shows the best sensitivity in the whole $m_{H^{\pm}}$ range. A recent paper from ATLAS collaboration [54] set upper limits in the single lepton final state at $\sqrt{s}$ = 13 TeV, using the data collected during the 2016-2018 period that correspond to 139 fb^{-1}. Exclusion limits are set in the mass range 200-2000 GeV that correspond to 3.6 pb and 0.035 pb respectively. The results are shown in Fig. 1.14 (right) and are compared to the limits extracted with 36 fb^{-1} (red lines). Theoretical predictions are

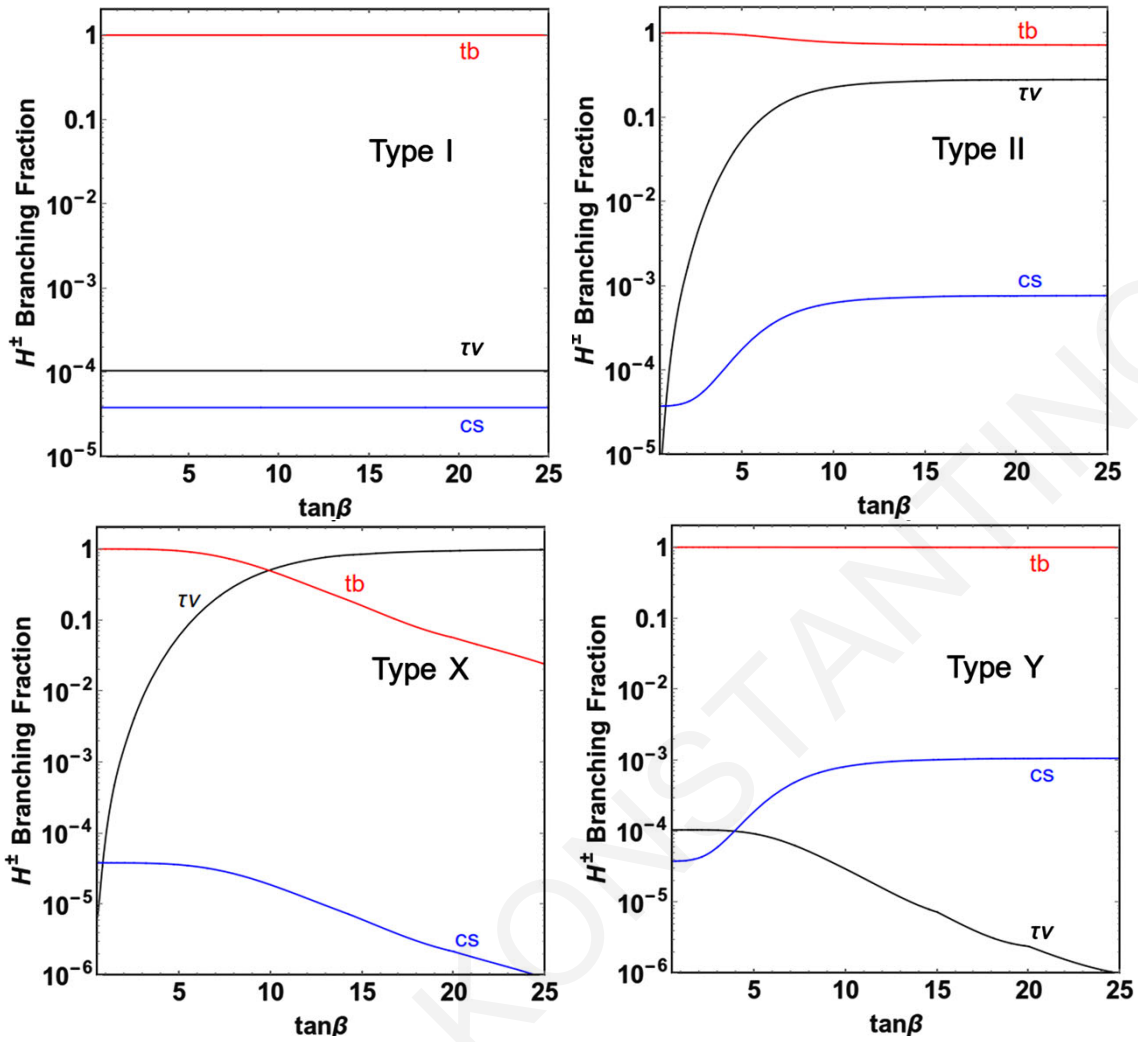


Figure 1.11: Branching fractions of the $H^{\pm}$ into the dominant third generation fermions as a function of $\tan\beta$ for $m_{H^{\pm}}$ = 250 GeV. The alignment limit $\sin(\beta-\alpha) \to 1$ and degenerate $m_{H^{\pm}}$, m_{H^0} and m_A are considered to suppress the $H^{\pm} \to W^{\pm}h^0$, $H^{\pm} \to W^{\pm}H^0$ and $H^{\pm} \to W^{\pm}A^0$ contributions [49].

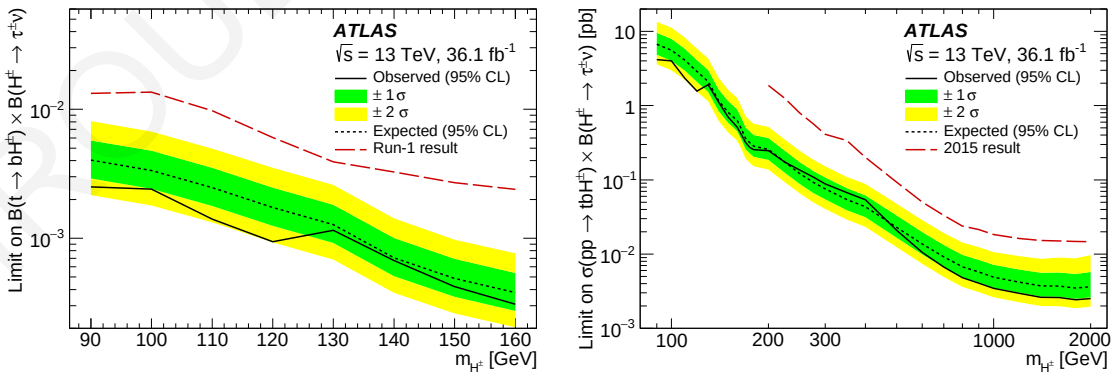


Figure 1.12: Exclusion limits of the ATLAS experiment on the $\mathcal{B}(t \to bH^{\pm}) \times \mathcal{B}(H^{\pm} \to \tau\nu_{\tau})$ (left) and $\sigma(pp \to tbH^{\pm}) \times \mathcal{B}(H^{\pm} \to \tau\nu_{\tau})$ (right) as a function of the $m_{H^{\pm}}$ obtained with $\sim$ 36 fb^{-1} at $\sqrt{s}$ = 13 TeV [50].

also illustrated in the *hMSSM* benchmark scenario and for $\tan\beta$ values of 0.5 and 1.

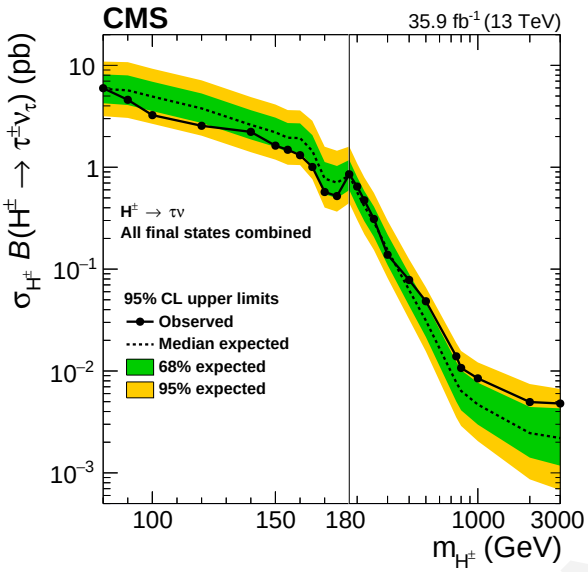


Figure 1.13: Exclusion limits of the CMS experiment on the $\sigma_{H^\pm} \times \mathcal{B}(H^\pm \rightarrow \tau\nu_\tau)$ as a function of the $m_{H^\pm}$ obtained with ~ 36 fb^{-1} at $\sqrt{s} = 13$ TeV [51].

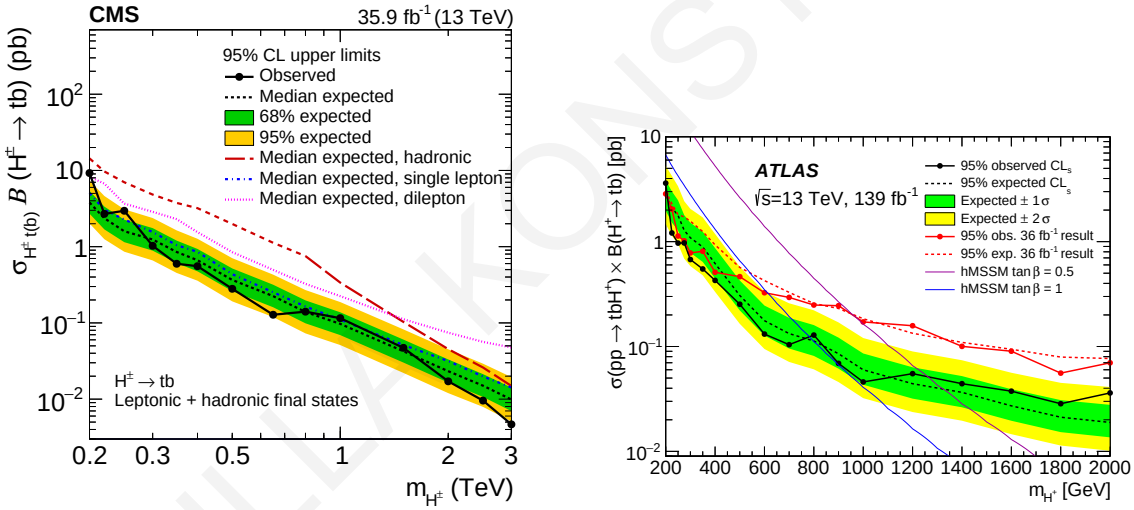


Figure 1.14: Exclusion limits on the $\sigma(H^\pm tb)\mathcal{B}(H^\pm \rightarrow tb)$ at $\sqrt{s} = 13$ TeV. Left: CMS limits obtained with ~ 36 fb^{-1} [53]. Right: ATLAS limits obtained with ~ 139 fb^{-1} [54].

1.3.4 Experimental constrains

Experimental results of $H^\pm$ studies put constraints on the $(m_{H^\pm}, \tan\beta)$ plane [49]. Searches on the $e^+e^- \rightarrow Z/\gamma \rightarrow H^+H^-$ performed in the LEP experiment [55] exclude $m_{H^\pm} \lesssim 80$ GeV in type-II and $m_{H^\pm} \lesssim 72.5$ GeV in type-I 2HDM. Strong indirect constrains are set from B meson decays $B \rightarrow X_S\gamma$ [56] that exclude $m_{H^\pm} \lesssim 580$ GeV. For type-II and type-Y the exclusion region is independent of $\tan\beta$ while in type-I and type-X, the exclusion limits affect the low $\tan\beta$ region. These results are combined with the latest CMS results of $H^\pm \rightarrow \tau\nu_\tau$ and $H^\pm \rightarrow tb$ to calculate exclusion regions in the light $m_{H^\pm}$ and heavy $m_{H^\pm}$ scenarios respectively. The exclusion regions are shown in Fig. 1.15 in the alignment region and for different types of 2HDM. The red regions represent the CMS

upper limits at $\sqrt{s}$ = 13 TeV and the green regions are excluded with the CMS upper limits at $\sqrt{s}$ = 8 TeV. The regions below the dashed black line are excluded by $\mathcal{B}(B \rightarrow X_S\gamma)$ and the solid black line defines the exclusion region of $\mathcal{B}(B_S \rightarrow \mu^+\mu^-)$.

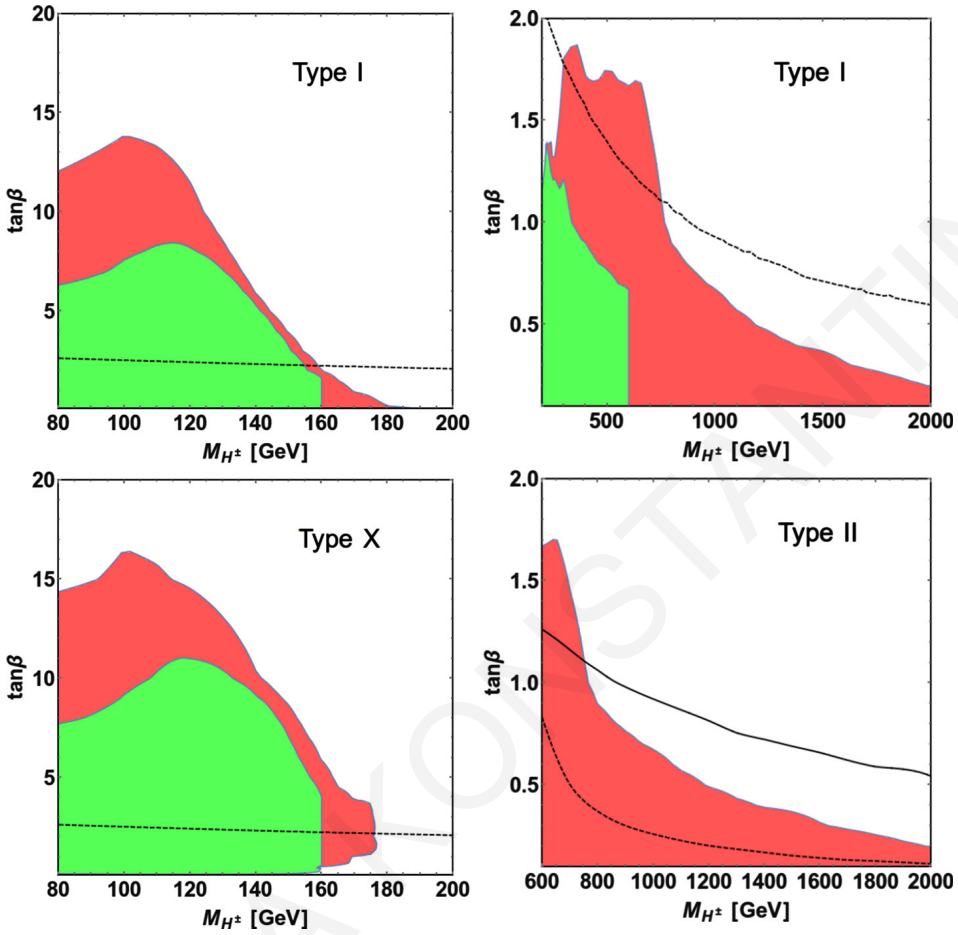


Figure 1.15: Branching ratios of the $H^\pm$ decay modes for $\tan\beta$ = 10 (left) and branching ratio of $H^\pm \rightarrow tb$ for different values of $\tan\beta$ (right) in the *hMSSM* benchmark scenario [49].

2 The experimental setup: the LHC accelerator and the CMS detector

2.1 Accelerators

The first accelerators were used in the early 1930s and were intended to study the structure of the atom's nucleus. To date, experiments in accelerators have studied various aspects of particle physics and led to big discoveries. The operation of the accelerators is based on enhancing the kinetic energy of a beam of charged particles, leptons or hadrons, with the use of electromagnetic field. The particle beam acquires a speed comparable to the speed of light and collides either with a fixed target or with another particle beam within a detector.

The most important components of a particle accelerator are the maximum energy and luminosity that can be reached. High energies are required to produce heavy particles, at a minimum center-of-mass energy equal to the rest energy of the particle. The center-of-mass energy is given by $\sqrt{s}$ where s is a Mandelstam variable calculated from the four-momenta of the colliding particles p_1, p_2:

$$s = (p_1 + p_2)^2. \tag{2.1}$$

Luminosity describes the rate of events produced (dN) per area and is expressed by:

$$\mathcal{L} = \frac{1}{\sigma}\frac{dN}{dt} \tag{2.2}$$

where σ is the transition cross section from the accelerating-particles scattering to a specific process. High luminosity allows the detection of rare processes, and the collection of a large amount of data can enhance the precision of their properties' measurements.

Based on their geometry, accelerators are categorized into linear or circular ones. In circular accelerators, a magnetic field B is applied to direct the beam in a circular orbit. Unlike linear accelerators, particle beams in circular accelerators can rotate multiple times until they reach sufficient energy. This means that the maximum energy that can be achieved is not limited solely by the physical size of the ring. However, accelerating charged particles in circular motion emit energy in the form of synchrotron radiation. The energy loss per revolution is given by:

$$\Delta E = \frac{4\pi}{3}\frac{q^2}{R}\left(\frac{E}{mc^2}\right)^4 \tag{2.3}$$

where R is the radius of the circular path, E, m, q are the particle energy, mass and charge respectively and c is the speed of light. The dependence of the energy loss on m^{-4} implies that the achieved energy is limited in the case of light particles, such as electrons.

The utilization of leptons provides a well-defined collision energy and a cleaner environment in terms of activity inside the detector, owing to the fact that leptons are elementary particles. These characteristics render the lepton colliders ideal for precision measurements. On the other hand, circular hadron colliders are used for discoveries at high energies because they can reach high center-of-mass energies. Due to the composite structure of hadrons, collisions occur between the hadron constituents, the quarks and gluons. The collision energy depends on the fraction of the total hadron energy carried by each particle constituent and the interaction of the particles with the detector results in a high-level collision activity.

2.2 Large Hadron Collider

The Large Hadron Collider (LHC) [57–59] is located at the European Organization for Nuclear Research (CERN) on the Franco-Swiss border and it is the largest and most powerful circular accelerator in the world. It was first commissioned in September 2008 and its main goal was the discovery of the Higgs boson and the discovery of new physics. The LHC is a 27 km ring, located 50-175m underground. The maximum operating center-of-mass energy is $\sqrt{s}$=14 TeV and can achieve a maximum luminosity of 10^{34} cm$^{-2}s^{-1}$. It is built with superconducting dipole NbTi magnets, which generate an 8.3 T magnetic field and direct two proton beams in a circular motion with opposite directions. To maintain the magnets, superfluid Helium is used as a coolant at a temperature of 1.9K. The magnet system is shown in Fig. 2.1. It consists of two pipes with separate vacuum chambers and dipole magnets for the two proton beams, while a shared cryostat and powering infrastructure are used. Additional quadrupole magnets are implemented to align and focus the beams. Around the interaction points strong quadrupole magnets are installed that squeeze the beams in high densities to increase the probability of the collision. Sextupole, octupole, and decapole magnets are used to correct small imperfections in the magnetic field.

The acceleration of the proton beams is done using radiofrequency (RF) cavities. For each beam there are eight cell cavities, providing a maximum accelerating voltage of 2 MV. The voltage generates an oscillating electric field of 400MHz. The protons passing through each cavity feel the electric field and get a push along the direction of their motions that accelerates or decelerates them so that they are synchronized. The synchronization of the protons in the bunches is very important as it ensures that the traveling protons will not collide with each other inside the accelerator. The protons are then grouped in packets called “bunches”. Each proton beam contains up to 2808 bunches of 1.15×10^{11} protons. Consecutive bunches have a distance of 7.5 m which corresponds to 25 ns spacing.

The LHC ring consists of eight arcs approximately 528 m long each which are connected

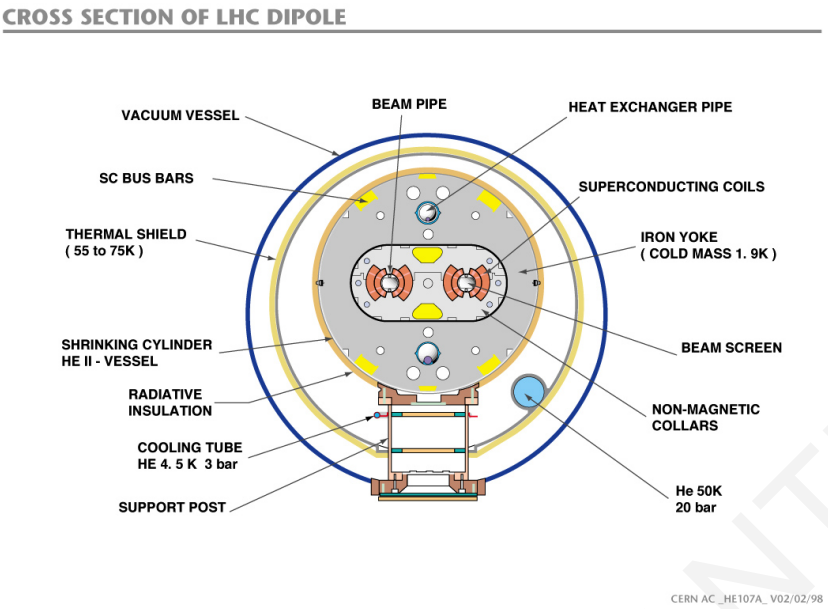


Figure 2.1: Cross section of the LHC dipole magnet [60].

by straight sections as illustrated in Fig. 2.2. The connection points are labeled as points 1-8 or interaction regions (IR). Four of them host the detectors of four of the experiments running at the LHC : A Toroidal LHC ApparatuS (ATLAS) [61] at point 1, Compact Muon Solenoid (CMS) [62] at point 5, A Large Ion Collider Experiment (ALICE) [63] at point 2, and Large Hadron Collider beauty (LHCb) [64] at point 8. These are the four points

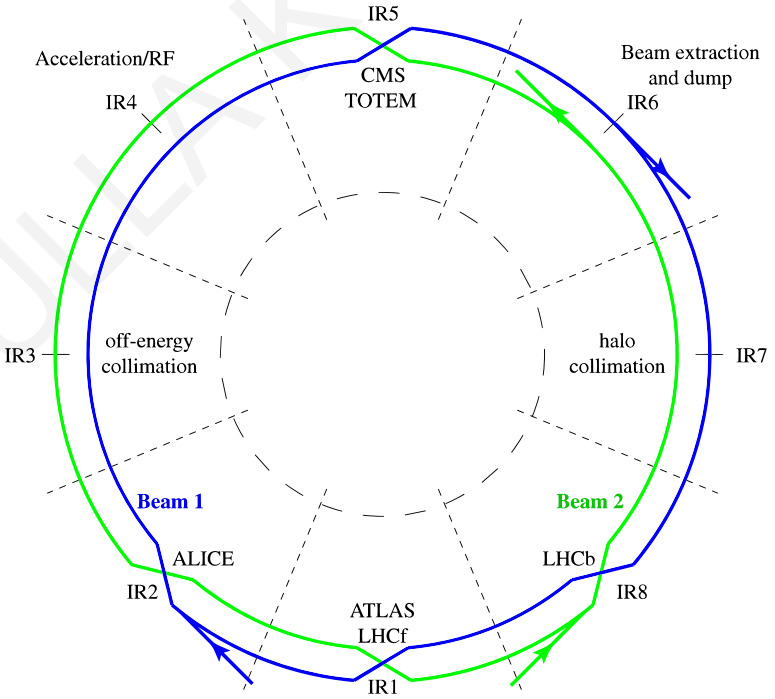


Figure 2.2: The LHC ring with the eight arc sections [59].

where the beams cross. The remaining four points are used as machine utilities. At point 3, off-momentum particles are caught by collimators and absorbers. At point 7, a collimation system removes halo particles to prevent damage to the LHC machines and minimize the

impact of radioactive activity. The RF cavity system is placed at point 4. Finally, at point 6, a beam dump system is installed to protect from beam loss. The response of the system is extremely fast and removes the beam within a single orbit.

The proton beam is produced by ionizing hydrogen gas. Before entering the LHC, the beams pass through a series of pre-accelerators. Leaving the hadron chamber, beams are injected in the LINAC 2 linear accelerator where they acquire energy of 50 MeV. The beams are then sent to the Proton Synchrotron Booster (PSB) to be accelerated to 1.4 GeV energy and then enter the Proton Synchrotron (PS). In the PS, the energy of the protons reaches 25 GeV. The last stop is the Super Proton Synchrotron (SPS) where the beams are cleaned from "halo" particles that are away from the beam to prevent damage to the LHC, beam losses and background radiation. Finally, the beams are transferred to the two pipes of the LHC with an initial energy of 450 GeV. The two beams are circulated in opposite directions inside the two pipes until they reach the desired energy of a maximum 7 TeV and collide near the four interaction points. The instantaneous luminosity for head-on collisions can be expressed as:

$$\mathcal{L} = \frac{n_b N_p^2 f_{rev}}{4\pi\sigma_x\sigma_y} \tag{2.4}$$

where n_b is the number of bunches containing N_p protons each, f_{rev} is the revolution frequency and $\sigma_x,\ \sigma_y$ are the transverse beam sizes in the horizontal and vertical directions respectively which express the overlap cross section at the interaction point. For a Gaussian beam distribution, the beam size is determined by the standard deviation (RMS) and is equal to:

$$\sigma = \sqrt{\frac{\epsilon_n\beta}{\gamma}}. \tag{2.5}$$

In the above equation, ϵ_n is the normalized emittance assuming a proton velocity $u \approx c$ and represents the beam size in the transverse phase space, γ is the relativistic Lorentz factor and β is the amplitude function [65]. At the interaction point the beam is squeezed to small β values, denoted as β^*. The crossing angle between the two beams at the interaction point causes a luminosity reduction which is given by the geometric luminosity reduction factor:

$$F = 1/\sqrt{1 + (\frac{\theta_c\sigma_z}{2\sigma^*})} \tag{2.6}$$

with θ_c being the crossing angle, σ_z the RMS bunch size in the longitudinal direction and σ^* the transverse RMS bunch size at the interaction point.

The instantaneous luminosity can be written in terms of the LHC parameters:

$$\mathcal{L} = \gamma\frac{n_b N_p^2 f_{rev}}{4\pi\beta^*\epsilon_n}F \tag{2.7}$$

The number of produced events of a given process can be estimated from the product of the

process cross section and the time integral of the instantaneous luminosity:

$$N_{exp} = \sigma \cdot \int \mathcal{L}(t)dt \tag{2.8}$$

The nominal beam and machine parameters as well as the parameters of the second operating period of the LHC (Run II) that took place from 2016 to 2018 are summarized in Table 2.1.

Table 2.1: Designed and operating parameters of the LHC for 2016-2018 [66].

Parameter	Designed	2016	2017	2018
Energy, $\sqrt{s}$ (TeV)	14	13	13	13
Number of bunches, n_b	2808	2220	2556	2556
Amplitude function, β^* (cm)	55	40	40→30	30→25
Normalized emittance, ϵ_n (μm)	3.75	1.8-2	1.8-2.2	1.8-2.2
Peak luminosity, $\mathcal{L}$ (10^{34}cm^{-2}s^{-1})	1.0	1.5	2.0	2.1
Half crossing angle, $\theta_c/2$ (μrad)	142.5	185→140	150→120	160→130

Additional inelastic pp interactions within the same or nearby bunch crossings may occur, overlapping the main proton collision. These interactions are called pileup and they are characterized by low momentum. In the presence of large pileup events, the environment inside the detector becomes harsh making the analysis challenging. The pileup interactions increase as the instantaneous luminosity increases, following a Poisson distribution. The distributions of the integrated luminosity recorded by the CMS detector and the average number of interactions per beam crossing are shown in Fig. 2.3 for the data-taking period of 2015 to 2018 as well as the 2022 period, the beginning of the Run III era with an increased center-of-mass energy of $\sqrt{s}$ = 13.6 TeV.

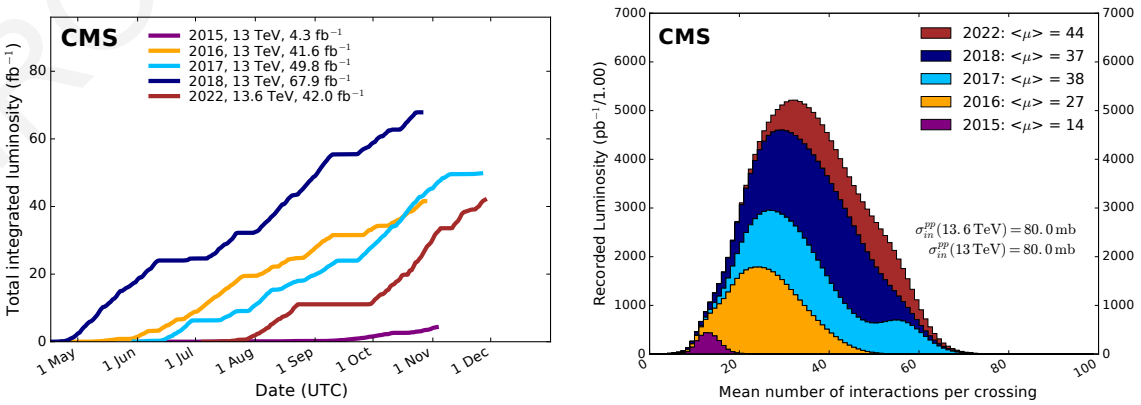


Figure 2.3: Right: Total integrated luminosity as a function of the day delivered to the CMS detector. Left: Recorded luminosity as a function of the mean number of interactions per bunch crossing.

2.3 The CMS detector

The CMS detector [62] is a cylindrical apparatus 22 m in length and 15 m in diameter. As shown in Fig. 2.4 it is composed of a series of concentric layers of subdetectors that are designed to exploit various properties of particles, such as their charge, position, energy, and momentum, as well as their interactions with matter, and prove information about the nature of the particles. In the central region, the subdetectors are arranged in a barrel geometry, while in the longitudinal direction, they are organized into disks or endcaps.

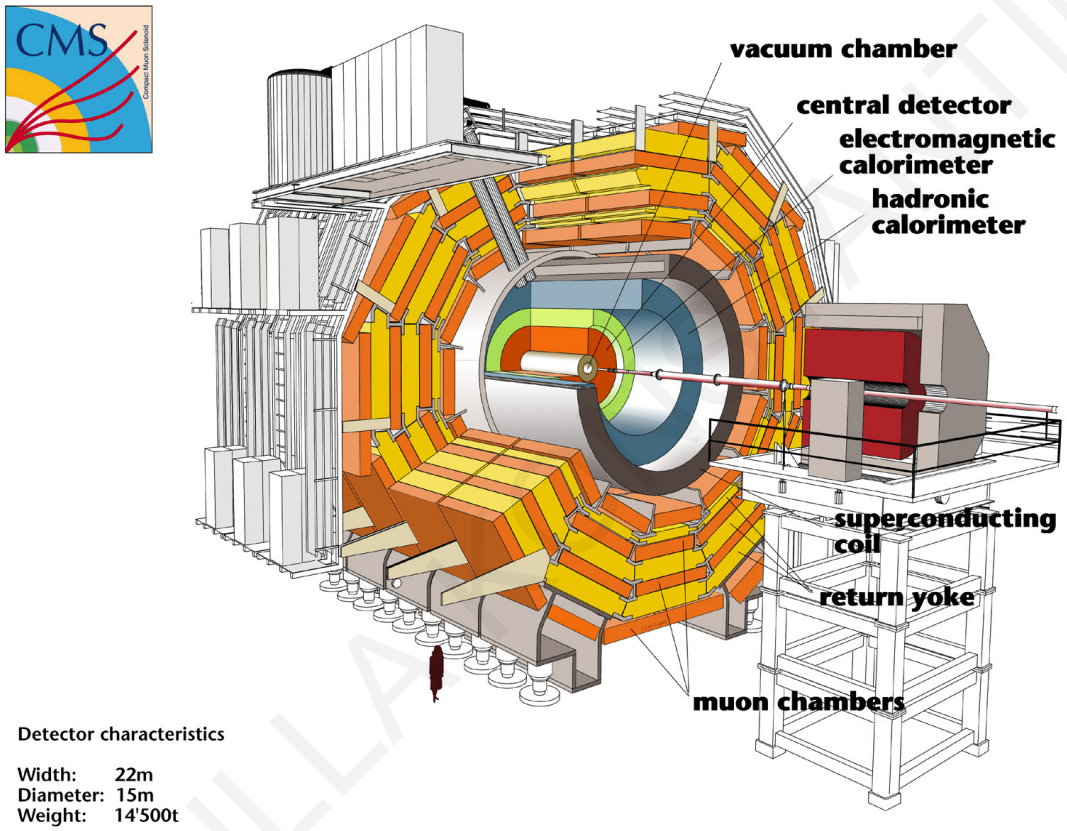


Figure 2.4: The CMS detector.

The CMS experiment uses a special coordinate system for the definition of spatial and angular variables, and is illustrated in Fig. 2.5. Considering the collision point as the origin, the positive x-direction is pointing towards the center of the LHC ring, the z-axis is parallel to the direction of the counter-clockwise proton beam and the y-axis is normal to the other two axes, pointing towards the surface. The z-axis determines the longitudinal direction while the $x-y$ plane defines the transverse plane. The azimuthal angle in the $x-y$ plane and the polar angle measured from the positive z-direction are defined as ϕ and θ angles respectively. Pseudorapidity η is a Lorentz invariant variable that is usually used instead of the polar angle θ and it is defined as:

$$\eta = -\ln\left[\tan\left(\frac{\theta}{2}\right)\right] \tag{2.9}$$

The relation between the θ angle and η is illustrated in Fig. 2.6. Pseudorapidity, approximates the rapidity y of a relativistic particle when its mass becomes negligible:

$$y = \frac{1}{2} \ln \left(\frac{E + p_Z}{E - p_Z} \right) \tag{2.10}$$

where E, p_Z are the energy and the longitudinal momentum of the particle. The momentum of a particle in the transverse plane is defined by the transverse momentum p_{T} as:

$$p_T = \sqrt{p_x^2 + p_y^2} \tag{2.11}$$

In conclusion, the coordinates of an object in the detector can be completely determined by the η and ϕ variables and the distance between two objects in the $\eta - \phi$ space is given by:

$$\Delta R \equiv \sqrt{\Delta\phi^2 + \Delta\eta^2} \tag{2.12}$$

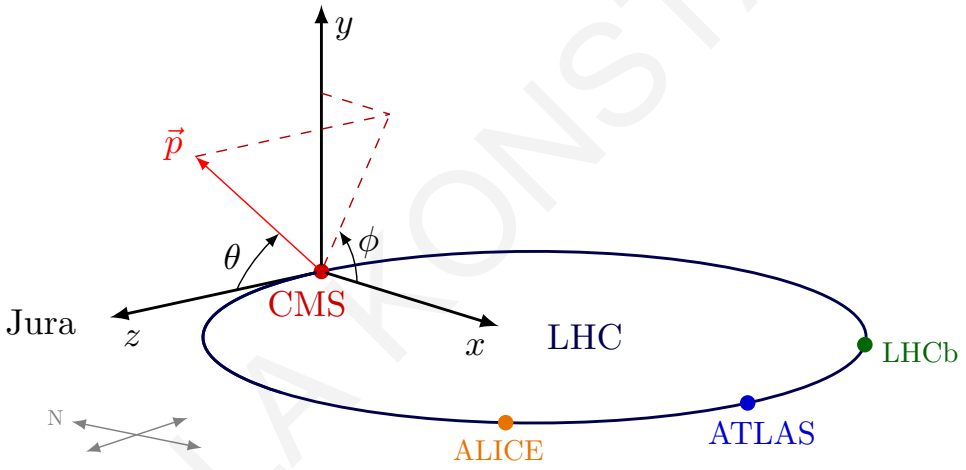


Figure 2.5: CMS coordination system.

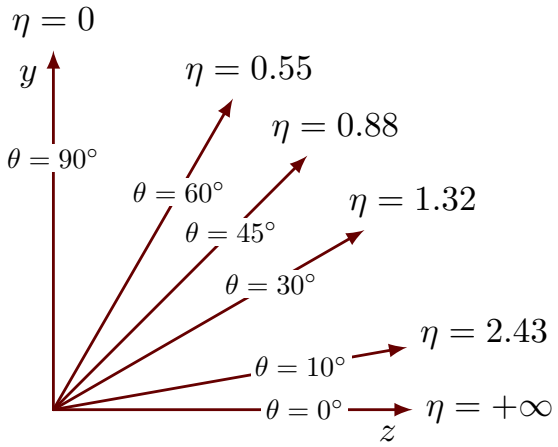


Figure 2.6: Pseudorapidity η for different values of the polar angle θ.

Starting from the collision point and moving towards the outer part of the detector, parti-

cles pass through subdetectors made of relatively light material where highly interacting particles are absorbed. Less interacting particles end up in high-density material layers where they deposit their energy and stop. Entering the detector, particles first encounter the tracking system, followed by the electromagnetic and the hadronic calorimeter. All these subdetectors are surrounded by a superconducting solenoid magnet [67] which generates a uniform magnetic field of 3.8 T. The magnetic field is enclosed by a steel magnetic yoke, which consists of five layers in the barrel and three layers in the endcaps regions. The strong magnetic field bends the trajectory of highly-energetic charged particles and allows the accurate measurement of their momentum. Muons interact very weakly with matter and manage to escape from all the previous layers and end up in the outer part of the detector, the muon system. Located outside of the solenoid magnet, the muon system is exposed to a non-uniform magnetic field. Figure 2.7 shows a transverse slice of the CMS detector and the trajectory of charged and neutral particles passing through the detector layers.

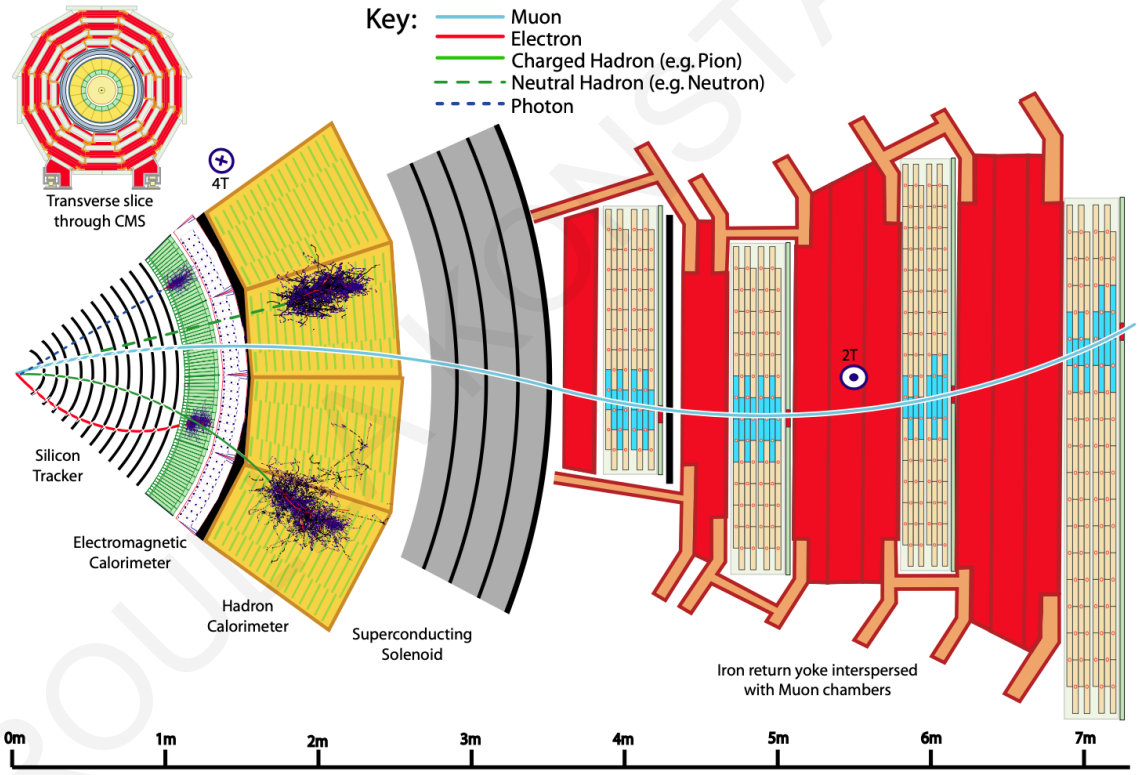


Figure 2.7: Cross section of the CMS detector.

2.3.1 Tracking system

The tracking system [68] is the subdetector located closest to the interaction point. Its purpose is the reconstruction of the trajectories of the electrically charged particles which allows the precise measurement of their momentum and charge, as well as the reconstruction of the position of the interaction point and impact parameters. Traces of the trajectories are left on the material of the detector as hits which reveal the particle position. These hits are then fitted

to reconstruct the tracks of the charged particles. The tracking system consists of the silicon pixel and the silicon strip tracker detectors, arranged in concentric cylindrical geometries. The selection of the material arises from the need to minimize the radiation damage caused by the enormous particle density that enters the detector and the bremsstrahlung radiation produced by the incoming electrons. Furthermore, it provides a fast response on the scale of 25ns allowing the control of contributions of additional inelastic interactions per bunch crossing. A schematic view of the tracker system is shown in Fig. 2.8.

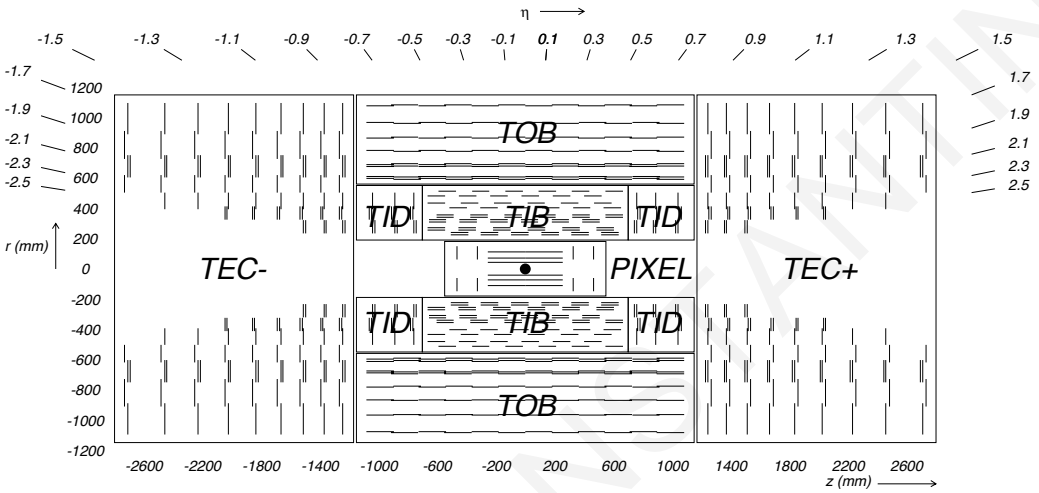


Figure 2.8: The CMS tracking system.

Pixel tracker detector

The Pixel tracker detector is the innermost part of the tracker that provides a three-dimensional track reconstruction with a pseudorapidity coverage of $|\eta| < 2.5$. It consists of about 66 million silicon sensor modules of size 100μm $\times$ 150μm, resulting in spatial resolution of the same magnitude. The original detector consisted of three barrel layers (BPIX) at a radial region between 4.4 cm to 10.2 cm and two endcap disks (FPIX) on each side. With the upgrade of the CMS during the first technical stop at the end of 2016, the pixel detector comprises four barrel layers and three endcap disks [69]. The installation of the new beam pipe with a smaller radius allowed the placement of the innermost barrel layer closer to the interaction point and at a radius of 2.9 cm improving the pattern recognition at smaller radii. The layers are arranged such that the Lorentz angle includes charge sharing between neighboring pixels.

Strip tracker detector

The silicon strip tracker is located at radii 20 cm to 116 cm away from the interaction point [70]. It is composed of the tracker inner barrel and disks (TIB, TID) which occupy the innermost part of the strip tracker with a maximum radius 55 cm, and it is surrounded by the outer barrel and endcaps (TOB, TEC). TIB consists of four barrel layers of silicon

micro-strip sensors parallel to the beam axis and have a thickness of 320μm. The strip pitch in the first two layers is 80μm and increases to 120μm in the third and fourth layers. TID consists of three disks at each end of the barrel. The silicon strip sensors are radially oriented and the pitch varies from 100μm to 141μm. In the outer strip tracker, TOB consists of six barrel layers of micro-strip sensors, with 500μm thickness and a strip pitch between 183μm to 122μm. The TEC includes nine disks on each end of the TOB, with up to seven rings of radial silicon strips. The strips on the rings have a thickness between 320μm to 500μm and an average pitch of 97μm to 184μm. The acceptance of the strip tracker is $|\eta| < 2.4$.

2.3.2 Calorimeters

Calorimeters are used for the measurement of the energy and position of the incoming particles. They are made of high-density material to stop and absorb most of the particles arising from the collisions. Upon the interaction of the particles with the detector material, a shower of particles is initiated, and eventually, particles deposit their energy in the calorimeters. The composition and spatial extension of the shower depend on the type and energy of the interacting particles as well as the detector's material, allowing the identification of the incident particles.

Electrons traveling in a medium, radiate bremsstrahlung photons which then split into electron-positron pairs e^-e^+. The consecutive bremsstrahlung radiation and e^-e^+ production results in a cascade of photons, electrons and positrons referred to as an electromagnetic shower. The dimensions of the shower are characterized by the radiation length X_0 i.e. the average distance in which an electron loses $1/e \approx 63.2\%$ of its total energy due to radiation, and the Molière radius (R_M) that is the transverse dimension of the shower at 90% of the energy deposition. The electromagnetic shower stops when the average particle energy is less than the critical energy E_c. The critical energy is defined as the energy in which the bremsstrahlung and ionization energy loss are equal.

Hadron showers can be produced in the detector calorimeters from charged and neutral hadrons that interact with the nucleus of the medium via strong interaction, giving rise to a cascade of particles. The depth of hadronic showers is parameterized by the nuclear interaction length λ_I defined as the mean distance traveled by a hadron before interacting. The nuclear interaction length is significantly larger than the radiation length. A diagram of an electromagnetic and hadronic shower is shown in Fig. 2.9.

Based on their construction, calorimeters can be either "homogeneous" or "sampling". Sampling calorimeters are made by alternating layers of a passive absorber and active material. The absorber is a dense material where the particle shower is produced and stopped (absorbed) in a limited space. The energy signal is then measured by the active detector. In homogeneous calorimeters, the absorption and detection of the signal are performed by the same material.

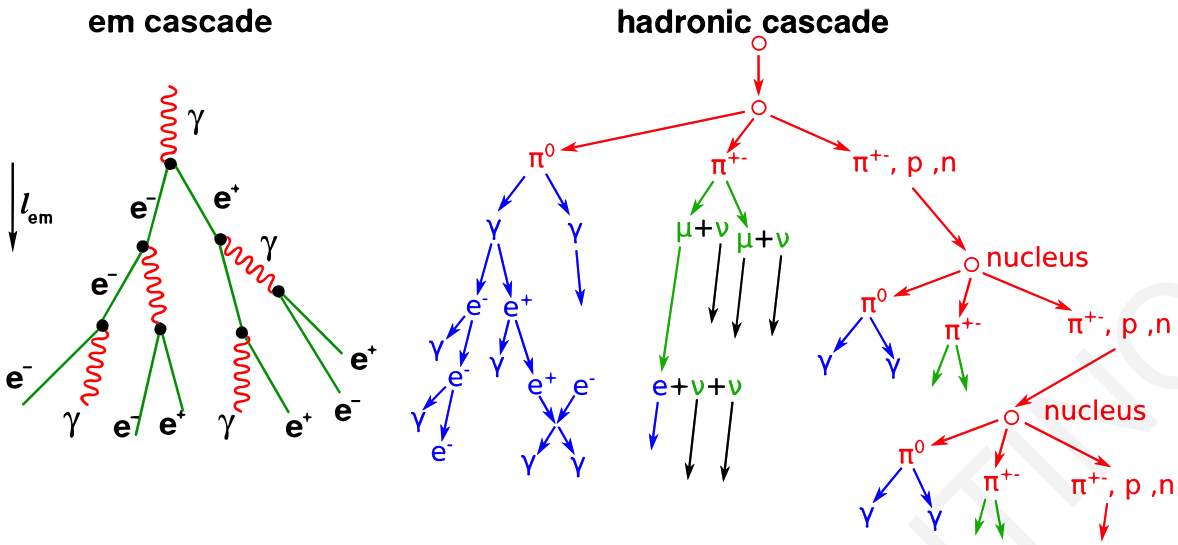


Figure 2.9: Electromagnetic (left) and hadronic (right) showers.

Electromagnetic Calorimeter

The Electromagnetic Calorimeter (ECAL) [71] is a homogeneous calorimeter used for the measurement of the energy and position of electrons and photons. It is made of high-density lead tungstate ($PbWO_4$) crystals with scintillator properties. The crystal is a transparent material with a short radiation length X_0 = 0.89 cm and a small Molière radius R_M = 22 mm, allowing the limitation of the electromagnetic shower to a compact area. As seen in Fig. 2.10 ECAL is composed of the barrel region (EB) and two endcaps (EE) providing a pseudorapidity acceptance of $|\eta| < 1.479$ and $|\eta| <= 3.0$ respectively. The $PbWO_4$ crystals have a very fast time response. The spatial granularity is $\Delta\eta \times \Delta\phi = 0.0175 \times 0.0175$ in the EB and increases up to $\Delta\eta \times \Delta\phi = 0.05 \times 0.05$ in the EE. For greater granularity, the ECAL features the preshower detectors (ES) positioned in front of the two endcaps at a pseudorapidity region of $1.653 < |\eta| < 2.61$. It is a sampling calorimeter that consists of two detector layers, a lead radiator followed by a silicon strip disk. The strip detectors provide a granularity of 1.9 mm in the direction perpendicular to the strips and 6 mm in the direction parallel to the strips. The primary purpose of the ES was to distinguish high-energy from low-energy photons.

The EB crystals have front face dimensions of 22×22 mm^2 and back face dimensions of 26×26mm^2 which matches the R_M size. The thickness of the crystals is 23 cm corresponding to $25.8X_0$. In the EE, the surface of the crystals is 28.6×28.6 mm^2 and 30×30 mm^2 in the front and back face dimensions respectively. The presence of the ES allows the use of thinner crystals of 22 cm which is translated to $24.7X_0$.

The light produced by the scintillators is proportional to the total energy of the original electron or photon. The scintillator light is converted to an electrical signal and amplified by avalanche photodiodes in the barrel region and vacuum phototriodes in the endcaps.

The energy resolution of the ECAL in the barrel region [72] is expressed as a function of the particles' energy and it was measured:

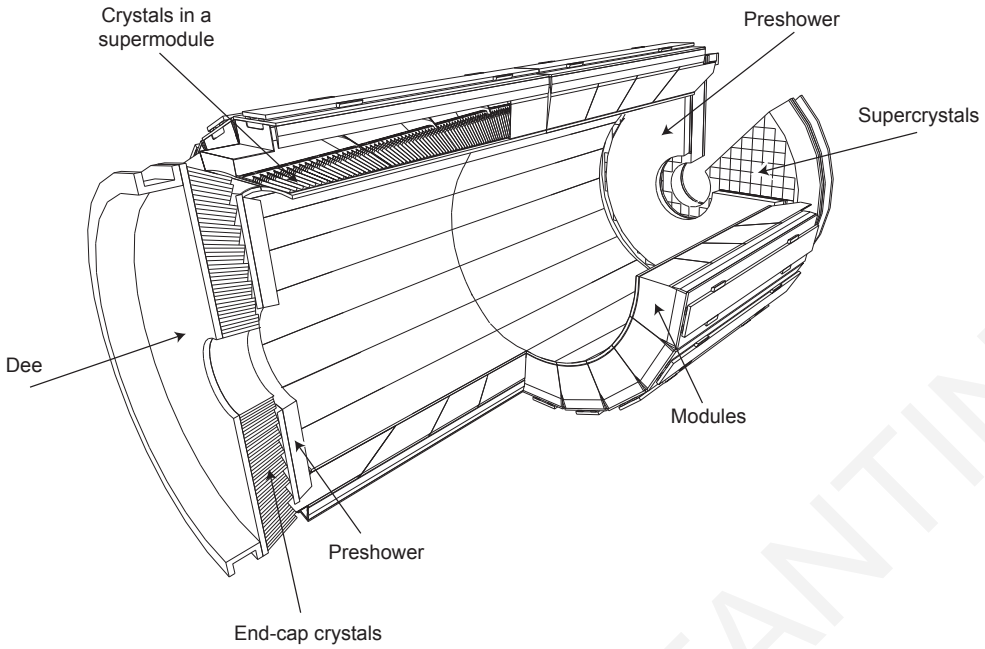


Figure 2.10: The CMS ECAL detector.

$$\left(\frac{\sigma}{E}\right)^2 = \left(\frac{2.8\%}{\sqrt{E[\mathrm{GeV}]}}\right)^2 + \left(\frac{12\%}{E[\mathrm{GeV}]}\right)^2 + (0.30\%)^2, \tag{2.13}$$

The three terms correspond to the stochastic, noise, and constant terms. The stochastic term describes statistical fluctuations in the number of photons. The noise term accounts for electronic noise and pileup energy. The constant term contains different systematic effects like calibration errors and damage to the detector.

Hadron Calorimeter

Strongly interacting particles deposit most of their energy in the Hadron Calorimeter (HCAL) [73]. HCAL is used for the energy measurement of the hadrons and the missing transverse energy. It is a sampling calorimeter that consists of the HCAL barrel (HB), the HCAL endcap (HE), the HCAL outer (HO) and the HCAL forward (HF). The HCAL detector is shown in Fig. 2.11.

HB and HE are made of active plastic scintillator layers inserted between brass absorber layers. The HB covers a pseudorapidity acceptance of $|\eta| < 1.4$ and provides a granularity of $\Delta\eta \times \Delta\phi$ = 0.087×0.087 while HE covers region of $1.3 < |\eta| < 3$ and the granularity is $\Delta\eta \times \Delta\phi$ = 0.17×0.17. In the HB, the absorber material consists of a front stainless steel layer of 40 mm thickness, eight 50.5 mm-thick brass plates, six 56.5-mm-thick brass plates and a back 75 mm-thick stainless steel layer. At the transverse plane, the thickness corresponds to $5.82\lambda_I$ and increases with the polar angle as $1/\sin\theta$, up to $10.6\lambda_I$. In the HE the absorber layers consist of nine brass plates of 79 mm-thickness. The wavelength of the light emitted by the plastic scintillator is shifted towards the red (Wavelength-Shifted WLS)

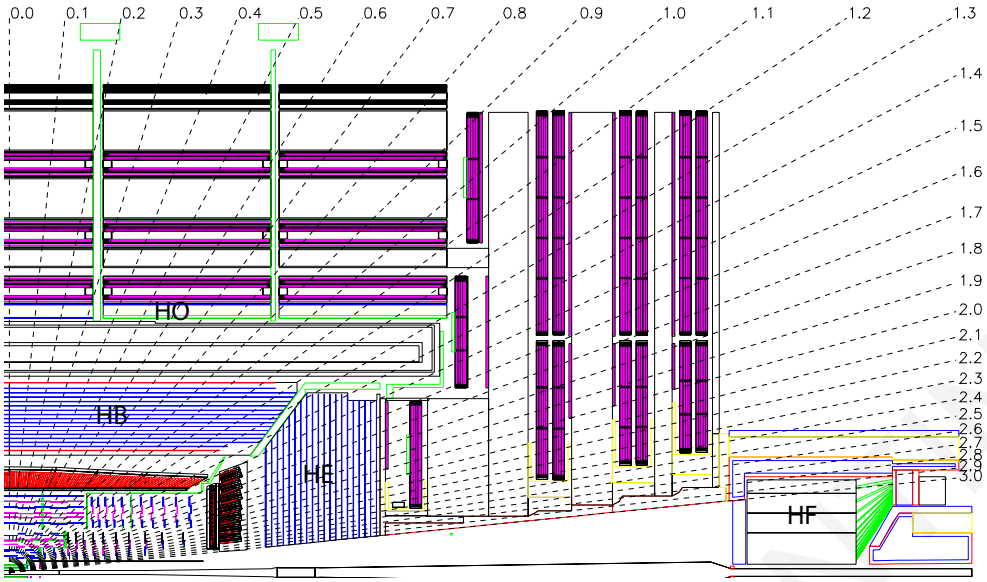


Figure 2.11: The CMS HCAL detector.

and collected by optical fibers. The signal is then transported to hybrid photodiodes (HPD) to be converted into an electrical signal.

The HO is located outside of the solenoid covering a pseudorapidity region of $|\eta| < 3$. It acts as a tail-catcher for hadronic showers providing additional nuclear interaction lengths of $1.4/\sin\theta\lambda_I$. It uses plastic scintillators of $\Delta\eta \times \Delta\phi = 0.087 \times 0.087$ as active material and the return yoke plates as an absorber. The scintillation light is collected with WLS fibers and read with HPD.

The HF is installed 6m away from the HE and extends the pseudorapidity coverage of the HCAL up to five. It is located in a high-rate environment of $3.0 < |\eta| < 5.0$. The HF is made of active medium of quartz fibers and steel absorber. Inside the active medium the incoming particles release energy through Cherenkov radiation which is collected by photomultiplier tubes (PMT).

The ECAL and HCAL calorimeters provide a maximum nuclear interaction length of approximately $12\lambda_I$ and $10\lambda_I$ in the barrel and endcap regions respectively. The combined energy resolution of the two calorimeters was measured in test beams [74] and is given by:

$$\left(\frac{\sigma_E}{E}\right)^2 = \left(\frac{84.7\%}{\sqrt{E[\mathrm{GeV}]}}\right)^2 + (7.4\%)^2 \tag{2.14}$$

where the first and second terms represent the stochastic fluctuations and the noise of the calorimeters respectively.

In the first operating period of the LHC, the PMT photomultipliers of the HF showed some signal of anomalously large energy deposits. As the luminosity increased, the identification and rejection of this artificially large signal became more challenging. To address this issue, during the technical stop at the end of 2016, the PMTs were replaced by higher-granularity multianode tubes [75]. The new photomultipliers can identify single-particle

anomalous signals, allowing the recovery of the response instead of rejecting the energy deposit entirely. Furthermore, the HB and HE HPD were replaced with silicon photomultipliers (SiPM) which offer significantly improved energy resolution. The previous HPDs were prone to electrical discharges when high voltage was applied.

2.3.3 Muon system

Muons interact very weakly with matter and can travel several meters through dense material before they stop. They are the only charged particles that escape the previous subdetectors and reach the outermost part of the detector, the muon system [76]. The muon system is used for muon identification and triggering. As seen in Fig. 2.12, it is divided into three types of detectors: the barrel drift tubes (DT), endcap cathode strip chambers (CSC) and the resistive plate chambers (RPC).

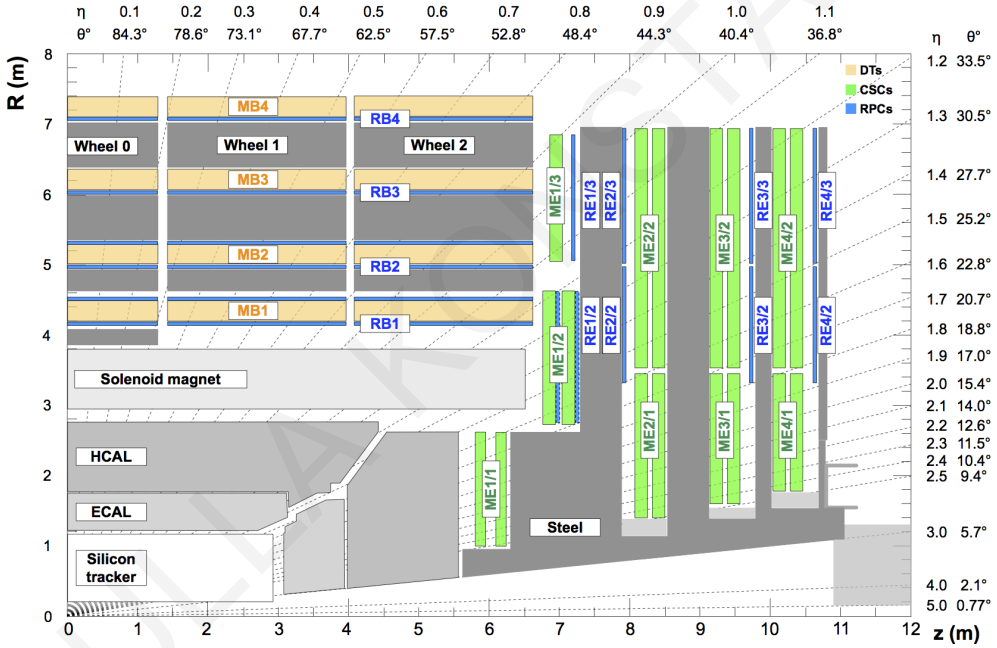


Figure 2.12: The CMS Muon system.

The DT is mostly contained within the return yoke and it is placed at $|\eta| < 1.2$. It is comprised of four cylindrical stations placed among the layers of the return yoke. The stations are constructed with drift chambers which include 60 chambers in the three inner stations and 70 chambers in the outer station. The chambers contain a gas mixture of 85% Ar and 15% CO_2 which becomes ionized when a muon passes through the detector. The resulting signals allow the measurement of the muon's coordinates based on the bending in the $r - \phi$ plane, as well as the position in the z-direction. Hits from the different stations are then combined to reconstruct the muon tracks. To eliminate inefficiencies or dead spots and allow for precise time measurement, the drift cells of each chamber are positioned with a half-cell width offset from their neighbor cells. The DT provides a spatial resolution of approximately 250 μm in ϕ direction and 500 μm in θ direction.

The CSC consists of four endcaps with four stations of cathode strip chambers. They are multiwire proportional chambers of six layers with radially positioned cathodes and transverse strips with respect to the anodes' wires. The radial direction of the cathodes allows a precise measurement of the azimuthal direction. It is placed outside the solenoid and exposed to a large magnetic field and high muon and background rate and covers a pseudorapidity region of $0.9 < \eta < 2.4$. These chambers use a gas mixture of 40% Ar, 50% CO_2, and 10% CF_4. The CSC detector is characterized by fine granularity, fast time response and radiation resistance. The spatial resolution in each station depends on whether a hit is near the edge of a strip and it varies between 45 μm and 139 μm.

RPC is mounted in four barrel stations and three endcap stations which cover a region of $|\eta| < 1.9$. The plane chambers are filled with a gas mixture of 96.2% $C_2H_2F_4$, 3.5% C_4H_{10}, and 0.3% SF_6. It is characterized by its excellent time resolution of the order of a few nanoseconds and modest spatial resolution. In combination with the DT and CSC, it provides very good muon triggering.

2.3.4 Trigger

At the LHC interaction points, proton beams collide every 25ns with approximately 30-40 pileup interactions taking place at each bunch crossing. This results in millions of data per second, and the storage of such a massive volume of events is impossible. The physics related to the majority of those events has been already studied and therefore they do not need to be recorded. In CMS, a real-time triggering system is used to reduce the number of stored events and to select the ones that are interesting for physics analysis. The selection is done in two steps, by applying the level-1 (L1) [77] and the high level trigger (HLT) [78].

The L1 trigger is a hardware-based algorithm. It uses information from the energy deposits in the calorimeters as well as track segments and hit patterns from the muon system, called trigger primitive generators (TPG), to perform a fast but rough reconstruction of various physics-object candidates, such as electrons, photons, jets and muons as well as global event variables. The trigger objects are ranked based on energy and quality. The highest-ranked objects are determined by the Global Calorimeter Trigger and Global Muon Trigger, and are transferred to the Global Trigger which takes the final decision to accept or reject the event. The selected events are kept in pipelined memories in the front-end electronics. The L1 trigger reduces the flow of information from 40 MHz at 25 ns bunch crossing to 100 kHz, which corresponds to 1/400 events per second.

The HLT is a software-based trigger that processes all the events stored by the L1 trigger to perform complex calculations using information from all the subdetectors of the CMS. The HLT uses a simplified version of the particle flow algorithm (discussed in Chapter 4) to reconstruct more accurately the physics objects and properties using trigger paths. The trigger paths determine the criteria that an event must fulfill to be accepted. These criteria are related to the presence of a physics object with requirements on their kinematic properties

or requirements on global event variables. The HLT reduces the number of events by a factor of 100 and eventually stores about 1 kHz events for offline analysis.

3 Data and simulated pp collisions

3.1 CMS integrated luminosity

This work analyzes the data collected with the CMS detector at $\sqrt{s}$ = 13 TeV during Run II. The luminosity delivered by LHC during stable proton collisions is shown in Fig. 3.1 with azure color, as a function of the time delivered and corresponds to 158.7 fb^{-1}. The total luminosity recorded by the CMS detector corresponds to 146.5 fb^{-1} and is shown with orange color in Fig. 3.1.

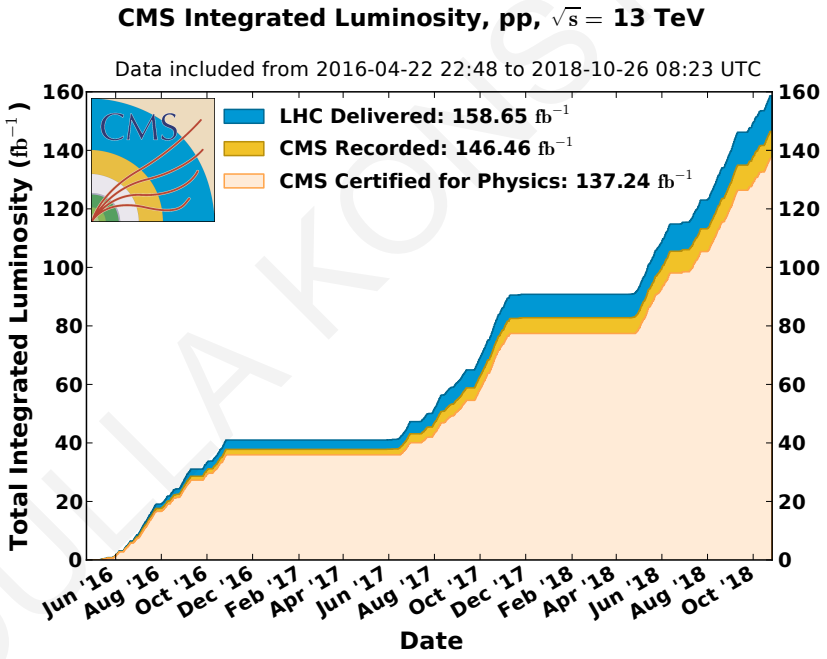


Figure 3.1: The cumulative curves for the luminosity delivered by LHC (azure), recorded by CMS (orange) and certified as good for physics analysis during stable proton beams (light orange).

The quality of the data collected by the CMS detector is monitored online, at the time of the collisions. This is a crucial task since problems in the detector can be spotted and resolved immediately, maintaining high efficiency of good quality data taking. The events that pass the L1 trigger are then certified offline in detail by experts that check the quality histograms related to each subdetector system. Events are certified as good if all the subdetectors, triggers and physics objects show the expected performance and are used for physics analyses. The luminosity of the certified data during the Run II data-taking period is 137.6 fb^{-1} and is shown in Fig. 3.1 with light orange color. This corresponds to a luminosity

of 36.3 fb^{-1}, 41.5 fb^{-1} and 59.8 fb^{-1} for the data-taking periods of 2016, 2017 and 2018.

At the beginning of the 2016 data-taking period, a decrease in the signal-to-noise ratio was observed in the silicon strip tracker. This was a result of a dynamic hit efficiency loss on the tracks, due to effects caused by highly ionizing particles (HIP). The problem was related to saturation effects in the pre-amplifier of the analogue pipeline voltage mode (APV25) readout chips of the tracker [79]. The problem was recovered in August 2016 and the affected period is referred to as "APV" or "preVFP" and previously known as "HIP mitigation (HIPM)". This period corresponds to an integrated luminosity of 19.5 fb^{-1}. To better describe the effect, data from the "APV" and "non-APV" 2016 period are analyzed separately [80].

The data samples used in this work are listed in Tables 3.1 to 3.4.

Table 3.1: Collision data from 2016 APV era.

Datasets	**Luminosity** (pb^{-1})
JetHT_Run2016B_HIPM_UL2016_MiniAODv2_v2_273150_275376	5828.41
JetHT_Run2016C_HIPM_UL2016_MiniAODv2_v2_275420_276283	2601.67
JetHT_Run2016D_HIPM_UL2016_MiniAODv2_v2_276315_276811	4285.85
JetHT_Run2016E_HIPM_UL2016_MiniAODv2_v2_276824_277420	4064.64
JetHT_Run2016F_HIPM_UL2016_MiniAODv2_v2_277932_278800	2717.34
Total:	19497.91

Table 3.2: Collision data from 2016 non-APV era.

Datasets	**Luminosity** (pb^{-1})
JetHT_Run2016F_UL2016_MiniAODv2_v2_278801_278808	410.83
JetHT_Run2016G_UL2016_MiniAODv2_v2_278816_280385	7652.81
JetHT_Run2016H_UL2016_MiniAODv2_v2_281613_284044	8739.88
Total:	16803.52

3.2 Event simulation

In this analysis, simulated data are used to study the kinematic properties of signal and background events, develop algorithms and evaluate their performance, as well as compare with experimental data to extract signal in the presence of the expected backgrounds.

Table 3.3: Collision data from 2017 era.

Datasets	**Luminosity** ($\mathbf{pb^{-1}}$)
JetHT_Run2017B_UL2017_MiniAODv2_v1_297050_299329	4803.14
JetHT_Run2017C_UL2017_MiniAODv2_v1_299368_302029	9574.02
JetHT_Run2017D_UL2017_MiniAODv2_v1_302031_302663	4247.68
JetHT_Run2017E_UL2017_MiniAODv2_v1_303825_304797	9313.75
JetHT_Run2017F_UL2017_MiniAODv2_v1_305044_306460	13534.53
Total:	41473.12

Table 3.4: Collision data from 2018 era.

Datasets	**Luminosity** ($\mathbf{pb^{-1}}$)
JetHT_Run2018A_UL2018_MiniAODv2_v1_315257_316995	14026.95
JetHT_Run2018B_UL2018_MiniAODv2_v1_317080_319310	7060.79
JetHT_Run2018C_UL2018_MiniAODv2_v1_319337_320065	6894.78
JetHT_Run2018D_UL2018_MiniAODv2_v2_320413_325172	31834.89
Total:	59817.41

The simulation of pp collision events is a multistage process. The first step describes the hard scattering at the parton level [30, 81, 82]. This is followed by the parton shower evolution and hadronization as well as the decay of unstable particles into final-state particles. The next step is the description of pileup events and low-momentum secondary interactions. Furthermore, the detector simulation describes the interaction between particles with the detector, the trigger, and digitization of the signals, taking into account the detector's acceptance and resolution effects.

3.2.1 Parton distribution functions

In high energies, protons consist of the three "valence" quarks (uud) and a sea of quarks and gluons arising from QCD interactions. The structure of a hadron cannot be explained by the QCD perturbation theory. Instead, the quark and gluon constituents are described with parton distribution functions (PDFs). The PDFs describe the probability of finding a parton of type α (quark or gluon) carrying a momentum fraction x with respect to the mother hadron h. The momentum scale Q is usually selected to be in the scale of the hadron, typically equal to the hadron's mass: $Q^2 \equiv \mu^2$. The PDF is denoted as $f_\alpha^h(x, \mu_F)$, where μ_F is the factorization scale parameter that is used to separate the hard process, which occurs at high momentum

transfer, from the soft process with low momentum transfer.

PDFs can be extracted experimentally with deep inelastic scattering experiments in short distances or phenomenologically with lattice QCD calculations. Figure 3.2 shows the PDFs for μ^2 = 10 GeV^2 (left) and μ^2 = 10^4 GeV^2 (right) provided by the NNPDF collaboration using collision data from multiple experiments [83]. In both cases, gluons and (anti)quarks from gluon splitting $g \to q\bar{q}$ contribute to low-x values while the PDF of the valence quarks increases in large values of x. Based on the measured PDFs at a given energy scale, a prediction of PDFs to different scales can be described by the Dokshitzer-Gribov-Lipatov-Altarelli-Parisi (DGLAP) evolution equation [84–86].

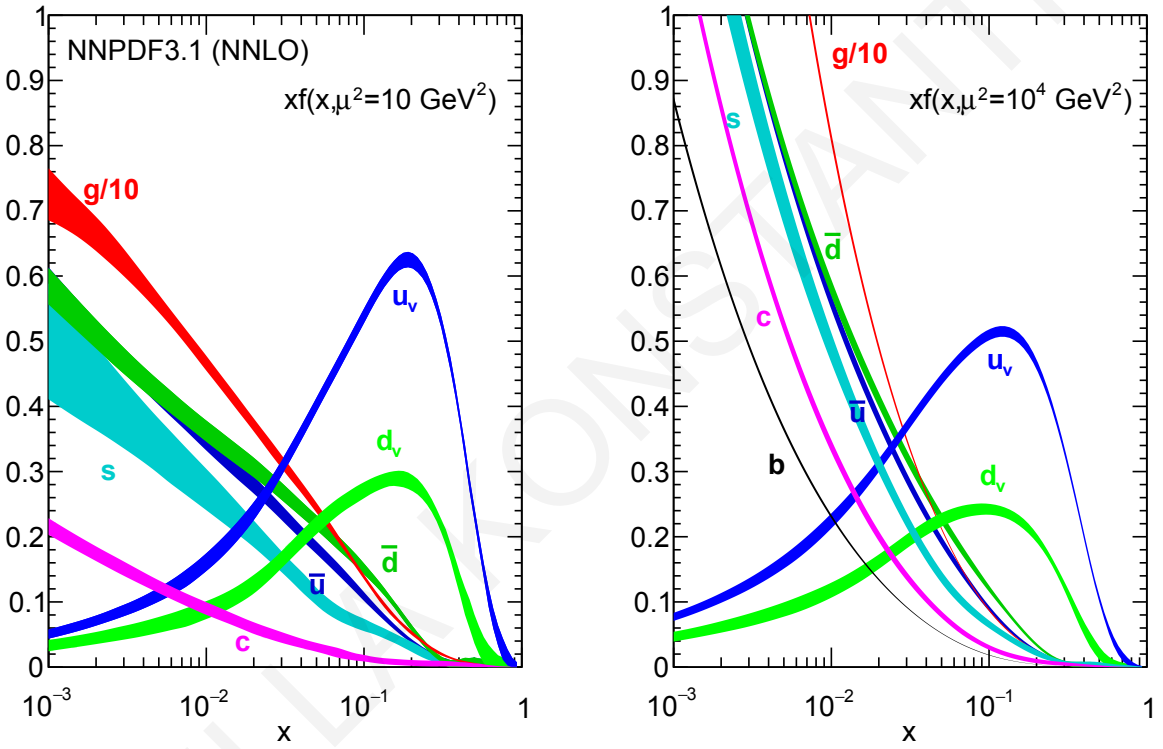


Figure 3.2: PDFs measured by the NNPDF collaboration at μ^2 = 10 GeV^2 (left) and μ^2 = 10^4 GeV^2 (right). Figures from [83].

3.2.2 Hard scattering

The hard scattering in hadron collisions describes deep inelastic scatterings with large momentum transfers. It is a short-distance process of asymptotically free particles that can be described by perturbation theory.

In collisions at a squared center-of-mass energy $s \equiv (P_1 + P_2)^2$, the partonic squared center-of-mass energy $\hat{s}$ is:

$$\hat{s} = (x_1 P_1 + x_2 P_2)^2 \approx 2 x_1 x_2 P_1 P_2 \approx x_1 x_2 s \tag{3.1}$$

where x_1 and x_2 are the fraction momentum carried by the two partons inside the hadrons with total momentum P_1 and P_2 respectively.

At the leading-order (LO) perturbation theory, the cross section of a scattering process of two hadrons $h_1 h_2 \to X$ can be computed by the factorization theorem:

$$\sigma_{h_1 h_2 \to X} = \sum_{i,j=q,\bar{q},g} \int dx_1 dx_2 f_i^{h_1}(x_1, \mu_F) f_j^{h_2}(x_2, \mu_F) \hat{\sigma}_{ij\to X}(x_1 x_2 s, \mu_R, \mu_F) \tag{3.2}$$

In the above equation, the sum runs over all the possible kinds of partons inside the two hadrons. The last term $\hat{\sigma}_{ij\to X}(x_1 x_2 s, \mu_R, \mu_F)$ represents the parton-level cross section for the production X through the initial partons i and j on the factorization and normalization scale μ_F and μ_R. The renormalization scale μ_R is a parameter used to absorb the divergences in higher-order perturbative calculations. The partonic cross section can be expressed as an integral in the differential final-state phase space of n particles Φ_n as:

$$\hat{\sigma}_{ij\to X} = \frac{1}{2\hat{s}} \cdot \int d\Phi_n \sum_{h,c} |\mathcal{M}_{ij\to X}|^2 \tag{3.3}$$

where $\mathcal{M}_{ij\to X}$ refers to the matrix elements and $\frac{1}{2\hat{s}}$ is the hadron flux. The matrix elements are calculated by the sum of all the Feynman diagrams and the parton helicities and colors. The differential phase space element over the n final-state particles is given by:

$$d\Phi_n = \prod_{k=1}^{n} \left[\frac{d^3 p_k}{(2\pi)^3 2E_k} \right] \cdot (2\pi)^4 \delta^4 (x_1 P_1 + x_2 P_2 - \sum_{k=1}^{n} p_k) \tag{3.4}$$

Several Monte Carlo (MC) event generators are available to calculate the hard process. For instance, MADGRAPH [87] calculates the matrix elements at LO precision, while MADGRAPH5_aMC@NLO [88] provides matrix element calculation at next-to-leading order (NLO) QCD corrections. POWHEG [89–94] incorporates both LO and NLO perturbative QCD corrections. All these generators provide no parton shower or hadronization modeling and therefore they need to be combined with a parton shower generator like PYTHIA [95]. PYTHIA is a multi-purpose generator that simulates all aspects of event generation and provides matrix element calculations at LO precision.

3.2.3 Parton shower

The parton shower can be described as a generic initial hard scattering process followed by an arbitrary sequence of splitting probabilities of $q \to qg$, $g \to q\bar{q}$, $g \to gg$. In an almost-collinear radiation splitting of the form $i \to jk$, the differential cross section of the k^{th} splitting is given by:

$$d\sigma_{k+1} \approx d\sigma_k \frac{\alpha_S(q^2)}{2\pi} \frac{d\theta^2}{\theta^2} dz d\phi P_{i,jk}(z) \tag{3.5}$$

where θ, ϕ are the opening and azimuthal angles of the splitting, $P_{i,jk}$ the flavor-dependent DGLAP splitting functions and z the energy fraction carried by the parton j.

The energy scale or the "hardness" of each splitting vertex can be defined as the virtuality of the splitting parton q^2:

$$q^2 = z(1-z)\theta^2 E^2 \tag{3.6}$$

which fulfills the condition:

$$\frac{dq^2}{q^2} = \frac{d\theta^2}{\theta^2} \tag{3.7}$$

The parton shower is terminated when the virtuality falls to the hadronization scale Q_0^2 which is of the order of 1 GeV^2.

To describe the evolution of a parton shower the differential cross section is multiplied by the probability of not splitting during evolution from the n^{th} to the m^{th} splitting vertex, which is given by the Sudakov form factor:

$$\Delta_i(q_n^2, q_m^2) = \exp\left[-\int_{q_m^2}^{q_n^2} \frac{\mathrm{d}q^2}{q^2} \frac{\alpha_S(q^2)}{2\pi} \int_{Q_0^2/q^2}^{1-Q_0^2/q^2} dz P_{i,jk}(z)\right] \tag{3.8}$$

Parton shower MC generators start with each parton in the primary process with a given initial scale Q^2 and solve the equation $r = \Delta_i(Q^2, q_n^2)$, where r is a uniform random number between 0 and 1. The solution is a value of the hardness of the next branch q_m. The energy fraction z, as well as the parton flavors j, k are generated with a probability proportional to $P_{i,jk}$. The procedure continues recursively by solving the equations $r = \Delta_i(q_n^2, q_m^2)$ and the evolution of each branch stops when the hardness reaches the hadronization scale Q_0^2.

Initial-state radiation

Prior to the hard scattering, initial partons emit radiation. This radiation creates a shower of partons and ultimately one of them participates in the hard scattering. Unlike final-state radiation, the incoming partons begin with low virtuality which gradually increases through splitting. The shower ends with the hard scattering, where virtuality reaches Q^2. In the simulation, initial-state radiations are evolved backward, starting from the hard process and moving opposite in time. The splitting is now described as $j \to ik$ and after each emission the partons "gain" energy. At the hard scattering, the energy fraction x is determined by PDFs. The shower is evolved with the modified Sudakov form factor:

$$\Delta_i(q_n^2, q_m^2) = \exp\left[-\int_{q_m^2}^{q_n^2} \frac{\mathrm{d}q^2}{q^2} \frac{\alpha_S(q^2)}{2\pi} \int_{Q_0^2/q^2}^{1-Q_0^2/q^2} dz P_{j,ik}(z) \frac{x/z f_j(x/z, q^2)}{x f_i(x, q^2)}\right] \tag{3.9}$$

where x/z describes the momentum fraction of "j" before splitting to i, k. The PDF of the parton "j" is calculated with the initial energy scale x/z while the PDF of the parton "i" is calculated with an energy x that the parton started with.

3.2.4 Hadronization

The parton shower stops at $\sim$1 GeV energy scales, where colored partons are confined into colorless hadrons. These hadrons can in turn decay into other stable hadrons. Hadronization is a non-perturbative process that can be described by various phenomenological models such as the string model and the cluster model.

The string model, also called the Lund model is based on the fact that at large distances the potential energy of color-singlet states, such as quark-antiquark pairs, increases linearly as they move apart. This is described as a quark and an antiquark placed at the edges of a gluonic string. As the quarks move apart the string stretches and the potential energy increases. When the potential energy reaches a critical value of the order of a typical hadron mass the string breaks and two new strings are formed, each attached to one of the quarks, forming new quark-antiquark pairs. The string simulation model is used by PYTHIA generator.

The cluster model is based on the "preconfinement" property of the perturbative QCD. According to this property, partons in a shower are clustered in colorless groups that are treated as color-singlet objects. The invariant mass distribution of the clusters depends only on the energy scale q and the fundamental QCD scale Λ. These groups form unstable hadrons that can decay into final-state hadrons. The cluster model is used by the HERWIG [96] generator.

3.2.5 Underline and pileup events

Underline events denote any additional activity beyond the hard scattering which is mainly associated with diffractive scattering or multiple parton interactions that do not participate in the hard scattering process. The cross section of the 2→2 parton-parton interactions is given by:

$$\sigma_{2\to 2} = \langle n \rangle \sigma_{pp} \tag{3.10}$$

where σ_{pp} is the pp cross section and $\langle n \rangle$ the average number of parton-parton interactions with $p_T > p_{\perp min} \approx \Lambda_{QCD}$. The probability for n scatterings is determined using a Poisson distribution. The PYTHIA modeling orders the additional interactions in p_{T} and requires the sum of all the energy fractions x, including the hard scattering to be less than or equal to unity. More details about the modeling of secondary parton interactions can be found in [81].

Pileup events are simulated by superimposing additional minimum bias events from inclusive soft QCD interactions to the hard scattering event. The number of pileup events and their energy density are controlled by the pileup distribution derived from data.

3.2.6 Detector simulation

Once an event is generated, a simulation of the detector is required to replicate a realistic simulation of the data that can be compared to the actual data collected by the CMS detector. To achieve this, the GEANT4 (GEometry ANd Tracking) software toolkit [97] is used. The GEANT4 generator provides a full simulation of the CMS detector geometry and materials, signal readout electronics, magnetic field, noise, detector response and other relevant factors. All these factors are combined with the physics of particle interactions with the detector. Hits on the detector are then digitized to reflect the output of the actual detector. The software emulates the response of all subdetectors' readout electronics. The simulation events are passed through a L1 Trigger Emulator (L1TEMU) which imitates the acceptance of the L1 trigger. The output of the L1TEMU is then transferred to the HLT. After digitization, the simulated data are provided in the same format as the data allowing the full event reconstruction and direct comparison between the two. Additional generator-level information is included in the simulation, such as the type of the particles, their origin and kinematic properties as predicted by the MC generator.

3.2.7 Simulated datasets

Signal and background contributions are simulated by various MC event generators described above. The matrix elements are calculated with the MADGRAPH, MADGRAPH5_aMC@NLO, POWHEG and PYTHIA generators using the NNPDF3.1 PDF sets. Parton showering and hadronization are generated with PYTHIA8 and underline events are tuned with the CP5 [98] parameters. The MADSPIN package [99] is used to model spin-correlation effects. The number of simulated events is normalized to the highest-order available cross section, while for signal, a normalization of $\sigma(\mathrm{pp} \to \mathrm{tbH}^{\pm}) \times \mathcal{B}(\mathrm{H}^{\pm} \to \mathrm{tb}) = 1\,\mathrm{pb}$ is used. Signal samples are generated with 17 different values of $m_{\mathrm{H}^{\pm}}$ ranging between 200 GeV to 3000 GeV. The production campaign of all the samples is the latest recommended ultra-legacy `RunIISummer20ULMiniAODv2`. The list of signal and background MC-generated samples used in this work are shown in Table 3.5.

Table 3.5: Simulated samples used to model signal and background processes.

Process	Generator	$\sigma \times \mathcal{B}$(pb)	Dataset (RunIISummer20ULMiniAODv2)
$tbH^{\pm}$, $H^{\pm} \rightarrow tb$	MG5_aMC@NLO + PYTHIA8	1.0	ChargedHiggs_HplusTB_HplusToTB_M-*_TuneCP5_13TeV_amcatnlo_pythia8 $m_{H^{\pm}}$ = [200, 220, 250, 300, 350, 400, 500, 600, 700, 800, 1000, 1250, 1500, 1750, 2000, 2500, 3000] GeV
$t\bar{t}$, hadronic	POWHEG + PYTHIA8	377.96 (NNLO [100–112])	TTToHadronic_TuneCP5_13TeV-powheg-pythia8
$t\bar{t}$, semileptonic	POWHEG + PYTHIA8	365.46 (NNLO [100–112])	TTToSemiLeptonic_TuneCP5_13TeV-powheg-pythia8
$t\bar{t}$, leptonic	POWHEG + PYTHIA8	88.34 (NNLO [100–112])	TTTo2L2Nu_TuneCP5_13TeV-powheg-pythia8
t, t channel, 4F	POWHEG + MADSPIN + PYTHIA8	136.02 (NNLO [108–116])	ST_t-channel_top_4f_InclusiveDecays_TuneCP5_13TeV-powheg-madspin-pythia8
$\bar{t}$, t channel, 4F	POWHEG + MADSPIN + PYTHIA8	80.95 (NNLO [108–116])	ST_t-channel_antitop_4f_InclusiveDecays_TuneCP5_13TeV-powheg-madspin-pythia8
t, tW channel, 5F	POWHEG + PYTHIA8	35.85 (NNLO [108–116])	ST_tW_top_5f_inclusiveDecays_TuneCP5_13TeV-powheg-pythia8
$\bar{t}$, tW channel, 5F	POWHEG + PYTHIA8	35.85 (NNLO [108–116])	ST_tW_antitop_5f_inclusiveDecays_TuneCP5_13TeV-powheg-pythia8
t, t channel, 5F	POWHEG + PYTHIA8	136.02 (NNLO [108–116])	ST_t-channel_top_5f_InclusiveDecays_TuneCP5_13TeV-powheg-pythia8
$\bar{t}$, t channel, 5F	POWHEG + PYTHIA8	80.95 (NNLO [108–116])	ST_t-channel_antitop_5f_InclusiveDecays_TuneCP5_13TeV-powheg-pythia8
$t\bar{t}t\bar{t}$	MG5_aMC@NLO + PYTHIA8	9.103×10^{-3} (NLO [117])	TTTT_TuneCP5_13TeV-amcatnlo-pythia8
$t\bar{t}W(qq)$ + jets	MG5_aMC@NLO + MADSPIN + PYTHIA8	0.371 (NLO [118])	TTWJetsToQQ_TuneCP5_13TeV-amcatnloFXFX-madspin-pythia8
$t\bar{t}Z(qq)$ + jets	MG5_aMC@NLO + PYTHIAeight 8	0.601 (NLO [118])	TTZToQQ_TuneCP5_13TeV-amcatnlo-pythia8
$t\bar{t}\gamma$ + jets	MG5_aMC@NLO + MADSPIN + PYTHIA8	3.697 (NLO [117])	TTGJets_TuneCP5_13TeV-amcatnloFXFX-madspin-pythia8
$t\bar{t}H, H(125) \rightarrow b\bar{b}$	POWHEG + PYTHIA8	0.295 (NLO [38])	ttHTobb_M125_TuneCP5_13TeV-powheg-pythia8
$t\bar{t}H, H(125) \nrightarrow b\bar{b}$	POWHEG + PYTHIA8	0.212 (NLO [38])	ttHToNonbb_M125_TuneCP5_13TeV-powheg-pythia8
$t\bar{t}$ WH(125)	MADGRAPH + PYTHIA8	1.349×10^{-3} (LO [117])	TTWH_TuneCP5_13TeV-madgraph-pythia8
$t\bar{t}$ ZH(125)	MADGRAPH + PYTHIA8	1.243×10^{-3} (LO [117])	TTZH_TuneCP5_13TeV-madgraph-pythia8
$Z/\gamma^{*}(qq)$ + jets	MADGRAPH + PYTHIA8	1006.0 (LO [117])	ZJetsToQQ_HT-200to400_TuneCP5_13TeV-madgraphMLM-pythia8
		114.2 (LO [117])	ZJetsToQQ_HT-400to600_TuneCP5_13TeV-madgraphMLM-pythia8
		25.53 (LO [117])	ZJetsToQQ_HT-600to800_TuneCP5_13TeV-madgraphMLM-pythia8
		12.87 (LO [117])	ZJetsToQQ_HT-800toInf_TuneCP5_13TeV-madgraphMLM-pythia8
W(qq) + jets	MADGRAPH + PYTHIA8	2542.0 (LO [117])	WJetsToQQ_HT-200to400_TuneCP5_13TeV-madgraphMLM-pythia8
		278.4 (LO [117])	WJetsToQQ_HT-400to600_TuneCP5_13TeV-madgraphMLM-pythia8
		58.71 (LO [117])	WJetsToQQ_HT-600to800_TuneCP5_13TeV-madgraphMLM-pythia8
		28.72 (LO [117])	WJetsToQQ_HT-800toInf_TuneCP5_13TeV-madgraphMLM-pythia8
WW	PYTHIA8	118.7 (NLO [119])	WW_TuneCP5_13TeV-pythia8
WZ	PYTHIA8	47.13 (NLO [119])	WZ_TuneCP5_13TeV-pythia8
ZZ	PYTHIA8	16.523 (NLO [119])	ZZ_TuneCP5_13TeV-pythia8
WWW, 4F	MG5_aMC@NLO + PYTHIA8	0.2158 (NLO [117])	WWW_4F_TuneCP5_13TeV-amcatnlo-pythia8
WWZ, 4F	MG5_aMC@NLO + PYTHIA8	0.1707 (NLO [117])	WWZ_4F_TuneCP5_13TeV-amcatnlo-pythia8
WZZ	MG5_aMC@NLO + PYTHIA8	0.05709 (NLO [117])	WZZ_TuneCP5_13TeV-amcatnlo-pythia8
ZZZ	MG5_aMC@NLO + PYTHIA8	0.01476 (NLO [117])	ZZZ_TuneCP5_13TeV-amcatnlo-pythia8
VH, $H(125) \nrightarrow b\bar{b}$	MG5_aMC@NLO +MADSPIN + PYTHIA8	0.9425 (NLO [38])	VHToNonbb_M125_TuneCP5_13TeV-amcatnloFXFX_madspin_pythia8
Z(qq)H, $H(125) \rightarrow b\bar{b}$	POWHEG + PYTHIA8	0.3599 (NLO [38])	ZH_HToBB_ZToQQ_M-125_TuneCP5_13TeV-powheg-pythia8
Z($\nu\nu$)H, $H(125) \rightarrow b\bar{b}$	POWHEG + PYTHIA8	0.1030 (NLO [38])	ZH_HToBB_ZToNuNu_M-125_TuneCP5_13TeV-powheg-pythia8

4 Object reconstruction and identification

The physics objects are reconstructed and identified using the particle flow (PF) [120] algorithm that exploits information from the signal hits from all the parts of the detector. The use of the PF algorithm in a hadronic collider detector is accomplished by the fine spatial granularity of the detector layers and the strong magnetic field maintained by the superconducting solenoid that allows the separation between energy deposits of charged and neutral particles.

4.1 Particle flow elements

The basic elements used as seeds to describe pp collision events in the CMS detector are the tracks and their origins (vertices) as well as the energy clusters recorded in neighboring cells of the calorimeters. The description of these PF seeds is discussed in the following sub-sections.

4.1.1 Tracks and vertices

Charged particle tracks are reconstructed using a track finder algorithm based on Kalman filtering [121]. The procedure starts with the generation of initial seeds of at least two hits from subsequent tracker layers, compatible with trajectories of charged particles. The trajectories are extrapolated to all the tracker layers to build tracks. The tracks are then fitted to determine the vertex, transverse momentum and direction of the track. Quality criteria are set on the fit χ^2, number of hits, p_{T} and the radial distance from the beam axis, to select the reconstructed tracks. The algorithm is performed multiple times. After each iteration, the hits associated to the selected tracks are removed and the procedure is repeated using relaxed selection criteria such that the track finding efficiency increases while maintaining a low mis-reconstruction rate.

The points of pp collisions inside the detector that are compatible with the center of the beam pipe are called the primary vertices (PV). The PVs are reconstructed using a deterministic annealing algorithm [122] to distinguish between hard scattering and pileup vertices. Displaced-track with respect to the PVs may arise from the decay of long-lived particles or interactions of secondary particles with the detector material. Vertices associated to the displaced tracks are called secondary vertices (SV).

4.1.2 Electron tracks

Tracks from electrons are reconstructed using two different approaches, the ECAL-based and the tracker-based seeding methods. ECAL-based tracks use energetic clusters in the ECAL as seeds. The position of the cluster is used to extrapolate the position of the associated hits in the innermost tracker layers. To take into account bremsstrahlung photons emitted due to the interaction of electrons with the tracker material, ECAL clusters are reconstructed as superclusters in small windows in the η direction but extended windows in the ϕ direction, to account for the bending of the electrons due to the magnetic field. ECAL-based electrons show good performance for highly-energetic electrons and electrons isolated from additional energy deposits in the ECAL. In the case of non-isolated electrons, the cluster-to-track association is affected by overlapping energy deposits, while in the presence of less-energetic electrons, the creation of superclusters cannot be accomplished as the bending of the electrons in the azimuthal direction increases. Non-reconstructed ECAL-based electrons are identified using the tracker-based approach, based on the iterative track finder algorithm. Radiating electrons are recovered during the track finder iterations with relaxed requirements on the number of hits and the p_{T} of the track. The probability of the electrons radiating inside the tracker material is also exploited to distinguish between electrons and charged hadrons. Electrons that radiate low energy in the tracker are reconstructed when the χ^2 of the track has good quality and can be associated with the closest cluster in an ECAL layer while the ratio of the cluster energy and the track momentum must be close to unity. Tracks with high energy loss due to radiation can often contain a small number of hits or have a large χ^2 value and thus, a set of supplemental quality selections is applied. The selected tracks are fitted with a Gaussian-sum filter (GSF) [123] to account for sudden and substantial energy losses along the trajectory. As a final step, a multivariate analysis (MVA) based on a boosted decision tree (BDT) is developed to select good-quality electron tracks. The BDT utilizes information from the GSF and KF fit parameters, the number of hits and the distance between the extrapolation of the track to the ECAL inner surface and the closest ECAL cluster. The final electron seeds are a combination of tracker-based and ECAL-based seeds. The introduction of the tracker-based seeds increases the electron efficiency by a factor of ~ 2 and allows the reconstruction of low-energy electrons with p_{T} down to 2 GeV.

4.1.3 Muon tracks

Muon tracks are reconstructed with the use of the tracker, the muon detector, or both detector systems and they are split into three types; standalone muons, global muons and tracker muons. Standalone muons are seeded with hits within the DT and CSC detectors that form track segments in the muon detectors. These seeds are then combined with all the DT, CSC and RPC hits, and are fitted to reconstruct the standalone-muon tracks. Standalone-muon tracks that are matched to a track in the inner tracker are combined and fit with the inner

track to form global-muon tracks. Global muons target high-p_{T} muons that pass through multiple layers of the muon system, requiring at least two segments on the muon detector planes. The tracker muon tracks are created by extrapolating each track in the inner tracker with $p_{\mathrm{T}} > 0.5$ GeV and momentum p > 2.5 GeV to the muon system. The extrapolated track must match at least one muon segment to be considered as a tracker muon. Tracker muons are efficient for tracks with p_{T} up to 10 GeV because of the requirement of only one segment in the muon detector.

Global muons and tracker muons provide an accurate reconstruction of about 99% of the muons produced within the geometrical acceptance of the muon system. Standalone-muons show worse momentum resolution and higher entanglement with muons from cosmic radiation.

4.1.4 Calorimeter clusters

A calorimeter clustering algorithm is developed to measure the energy and direction of the neutral hadrons and photons, distinguish neutral from charged hadrons, reconstruct and identify bremsstrahlung electrons and correct the energy measurement of highly energetic charged hadrons with low-quality tracks. The algorithm is performed separately in the two calorimeters, in the barrel and endcap regions, and the two preshower layers. The algorithm is seeded by cells with energy deposits larger than a given threshold, and larger than the energy of the neighboring cells, called cluster seeds. Topological clusters are then grown by adding to the cluster seeds neighboring cells with an energy excess of at least twice the noise level.

To reconstruct the clusters within the topological cluster, an expectation-maximization algorithm based on a Gaussian-mixture model is used which assumes Gaussian energy deposits in each cluster seed. For N cluster seeds, the expected energy fraction f_{ji} arising from the i^{th} Gaussian energy deposit in each individual cell at position $\vec{c}_j$ and energy E_j is given by:

$$f_{ji} = \frac{A_i e^{-(\vec{c}_j - \vec{\mu}_i)^2/(2\sigma^2)}}{\sum_{k=1}^{N} A_k e^{-(\vec{c}_j - \vec{\mu}_k)^2/(2\sigma^2)}}, \tag{4.1}$$

where the Gaussian amplitude A_i and the coordinates of the mean $\vec{\mu_i}$ are the model parameters that need to be adjusted. The f_{ji} is calculated using fixed values of A_i, $\vec{\mu_i}$. The model parameters are then determined using an analytical maximum-likelihood fit:

$$A_i = \sum_{j=1}^{M} f_{ji} E_j, \qquad \vec{\mu_i} = \sum_{j=1}^{M} f_{ji} E_j \vec{c}_j. \tag{4.2}$$

The calculation is repeated until convergence.

The reconstruction of neutral particles is completely determined by the calorimeter clusters. In the presence of overlapping charged hadrons, the energy deposit of the neutral particles becomes extremely challenging. To correct the energy scale of neutral particles and maximize the reconstruction probability while keeping the mis-reconstruction rate small, ECAL and HCAL calorimeter clusters are calibrated. The calibration is first done with test-beam data and subsequently with large simulated samples processed through a GEANT4 simulation [97] of the CMS detector. For the ECAL, a single photon simulated sample is used, in which photons that convert before entering the ECAL are excluded, to deal with the calibration of single clusters. HCAL clusters are calibrated with simulated single neutral hadrons, taking into account the calibrated ECAL clusters.

4.2 Particle identification and reconstruction

The reconstruction of the particles is obtained using the link algorithm that connects the PF elements from different subdetectors. Possible links are created using the nearest elements in the (η, ϕ) plane with the use of k-dimensional trees [124]. The distance between two linked elements is measured to quantify the quality of the link.

The event reconstruction starts with the identification of the muons and the corresponding PF elements are removed from the list of elements to reconstruct the remaining objects. Electrons and isolated photons are then reconstructed at the same step. The remaining PF elements are used to reconstruct the charged and neutral hadrons and photons from parton fragmentation, hadronization, and decays in jets. When all particles have been identified, the full reconstruction is revisited by a post-processing step to minimize the probability of misidentified and mis-reconstructed particles that lead to artificially large energy imbalance in the transverse plane.

4.2.1 The link algorithm

Links between tracks from the central tracker and a calorimeter cluster are established by extrapolating the track from the last hit to two layers of the preshower, ECAL or HCAL. The extrapolated track must be within a cluster area, at a depth that corresponds to the expected maximum of a typical electron shower in the ECAL or one interaction length in the HCAL. The distance between the extrapolated track and the cluster position determines the link distance. If more than one track is linked to the same ECAL cluster or a track is linked to several HCAL clusters, only the one with the smallest link distance is considered. To reconstruct photons emitted by electron bremsstrahlung, tangents to the electron tracks on each tracker layer are extrapolated to the ECAL. Links of potential bremsstrahlung photons are created if the extrapolated tangent position is within a cluster area. Photon candidates that convert to a e^+e^- pair inside the tracker are described using a conversion finder [125] that

creates track-to-track pairs. The sum of the momenta of the two tracks indicates the direction of the photon. A link of a potential converted photon is created between the original track and a track pair if the direction of the photon is compatible with one of the track tangents.

Links of clusters between HCAL and ECAL, as well as ECAL and preshower are created when the position of the cluster of the more granular calorimeter is within the boundaries of the second cluster. Links between ECAL clusters and superclusters are also established when at least one shared cell is found. When more than one HCAL or ECAL cluster is linked to a single ECAL or preshower cluster respectively, only the one with the smallest distance is kept.

Charged-particle tracks may also be linked through a common SV. At least three tracks must be associated with the SV, of which at least one is an incoming track that includes hits between the PVs and the SVs.

Global and tracker muons are links between tracks from the tracker detector and the muon system.

4.2.2 Muons

Isolated muons are reconstructed and identified from global or track muons as described in Section 4.1.3, with additional quality and isolation criteria applied to reject misidentified muons.

Non-isolated muons such as muons from semileptonic heavy-flavor decays are reconstructed by global muons associated with calorimeter clusters. To reject misidentified non-isolated muons mainly from charged hadron where residuals from the hadron shower pass through the muon detector layers (punch-through), a set of tight quality criteria is applied, including the requirement of at least three matching track segments in the muon detectors or that the calorimeter cluster associated with the track are compatible with originating from a muon.

For muons with $p_{\mathrm{T}} < 200$ GeV, the momentum resolution is dominated by the inner track which determines the reconstructed muon p_{T}. For higher values of p_{T}, the momentum is determined by the fit with the minimum χ^2 among the following tracks: tracker only, tracker and first muon detector plane, global, and global without the muon detector planes with high occupancy.

4.2.3 Electrons and isolated photons

Electron candidates are reconstructed from GSF tracks if the corresponding ECAL cluster is not linked to three or more additional tracks. Photon candidates are seeded from an ECAL supercluster that is not linked to a GSF track and has a transverse energy greater than 10 GeV.

Bremsstrahlung radiation and photon conversions are seeded by ECAL-based candidates for which the energy sum in the HCAL cells with a distance to the supercluster position

smaller than 0.15 in the (η, ϕ) is less than 10% of the supercluster energy. Additionally, all the ECAL clusters linked to either the supercluster or to one of the track tangents must be associated with the candidate. Tracks and HCAL clusters linked to these ECAL clusters must also be compatible with the electron hypothesis.

The total energy of the collected ECAL clusters is corrected for the energy missed in the association process, with analytical functions. The energy of the electrons is obtained from a combination of the corrected ECAL energy and the momentum of the GSF track. The electron direction is taken from the GSF track. The corrected ECAL energy is also assigned to the photons and the direction is determined by the position of the supercluster.

Additional identification criteria are applied to the electrons, based on an MVA that combines information from the electron track, the associated ECAL and HCAL clusters, the number of hits and the KF and GSF fit parameters.

Quality criteria are applied also to the photons which are required to be isolated from other tracks and calorimeter clusters, while the ECAL cell energy distributions and the HCAL-to-ECAL energy ratio must be compatible with those expected from a photon shower.

4.2.4 Hadrons and non-isolated photons

The PF elements used in the reconstruction of the muons, electrons, and isolated photons are masked and the remaining elements are used to identify the hadrons from jet fragmentation and hadronization and non-isolated photons. Within the tracker acceptance ($|\eta| < 2.5$), ECAL and HCAL clusters not linked to any track give rise to photons and neutral hadrons respectively. This is because neutral hadrons leave only 3% of the total energy inside the ECAL. Outside the tracker acceptance, charged and neutral hadrons cannot be distinguished and the reconstruction is established based on the calorimeter clusters. ECAL clusters linked to an HCAL cluster are assumed to arise from the same hadron shower, while ECAL clusters not linked to an HCAL cluster are identified as photons. The energy of the reconstructed hadrons and photons is calibrated as described in Section 4.1.4.

The remaining HCAL clusters linked to at least one track, which in turn are linked to some of the remaining ECAL clusters may give rise to additional hadrons, photons or muons from charged-hadron decays. These HCAL and ECAL clusters are calibrated from the energy of the specific HCAL cluster and the total energy of the ECAL clusters.

If the calibrated calorimetric energy is higher than the sum of the remaining track momenta, the excess is considered as coming from photons and neutral hadrons. If the calibrated calorimetric energy and the sum of the track momenta are compatible, a charged hadron is identified, with its momentum being measured from both the tracker and the calorimeter. Finally, in the case of calibrated calorimetric energy smaller than the sum of the track momenta by at least three standard deviations, a muon search is performed with relaxed requirements. After the identification of additional muons, if the track momentum sum is still in excess, it is likely to arise from residual mis-reconstructed tracks. These tracks are sorted in decreasing

uncertainty in p_{T} and are sequentially masked either until no such tracks remain or until the momentum excess disappears, whichever comes first.

4.3 Jets

Jets are collimated sprays of particles produced from the fragmentation and the hadronization of quarks and gluons. In CMS searches, they are reconstructed by clustering all PF particles using the anti-k_{T} [126, 127] algorithm, implemented by the FastJet package [128]. The anti-k_{T} algorithm is a generalization of the k_{T} [129] and Cambridge/Aachen [130] algorithms which introduce the distances d_{ij} between two objects, particles or pseudo-jets, with indices i, j and d_{iB} between the particle i and the beam position. The distances are defined as follows:

$$d_{ij} = \min(k_{Ti}^{2p}, k_{Tj}^{2p}) \frac{\Delta^2 ij}{R^2}, \tag{4.3}$$

$$d_{iB} = k_{Ti}^{2p}. \tag{4.4}$$

In the above equations $\Delta^2 ij \equiv (y_i - y_j)^2 + (\phi_i - \phi_j)^2$ and k_{Ti}, y_i, ϕ_i are the transverse momentum, rapidity and azimuth of the object i. The parameter R is the radius of the reconstructed jet, while p is an additional parameter that tunes the relative power between the energy and the geometrical scales. The reconstruction starts by quantifying all the distances d_{ij} and d_{iB} and identifying the smallest among all combinations of objects. If d_{ij} gets the minimum value, objects i and j are recombined into a single pseudo-jet. If d_{iB} is the smallest distance then object i is considered a single jet and it is removed from the list of objects. The procedure is repeated with the updated list until no objects are left.

For p = 1 and p = 0 in Eqs. (4.3) and (4.4), the algorithm corresponds to the inclusive k_{T} and Cambridge/Aachen algorithms respectively. For positive values of p, the treatment of soft radiation is similar to the one of the k_{T} algorithm.

The anti-k_{T} algorithm uses p = -1. For negative values of p the shapes of the jets are determined by the hard particles exclusively. From Eq. (4.3) and for $p < 0$, the distance d_{ij} between a hard and a soft particle depends on the transverse momentum of the hard particle and their separation Δ. In the case of two soft particles, d_{ij} gets large values and this leads to the clustering of the soft particles with the hard ones before they cluster among each other.

The jet reconstruction is compared for the k_{T}, Cambridge/Aachen, SISCone [131] and anti-k_{T} clustering algorithms, for a parton-level event of $\sim 10^4$ random soft particles and is illustrated in Fig. 4.1. For the k_{T} and Cambridge/Aachen algorithms, the jet borders are highly affected by the soft particles. The SISCone algorithm performs iterative jet searches of stable cones with a split-merge step to disentangle overlapping cones. The algorithm provides regular jet shapes for single-particle jets, but more complex shapes for composite jets. Finally, in the anti-k_{T} only soft-jets' shapes are influenced by soft radiation, while hard

particle jets have a conical shape of radius R, and the hard particles are close to the jet axis.

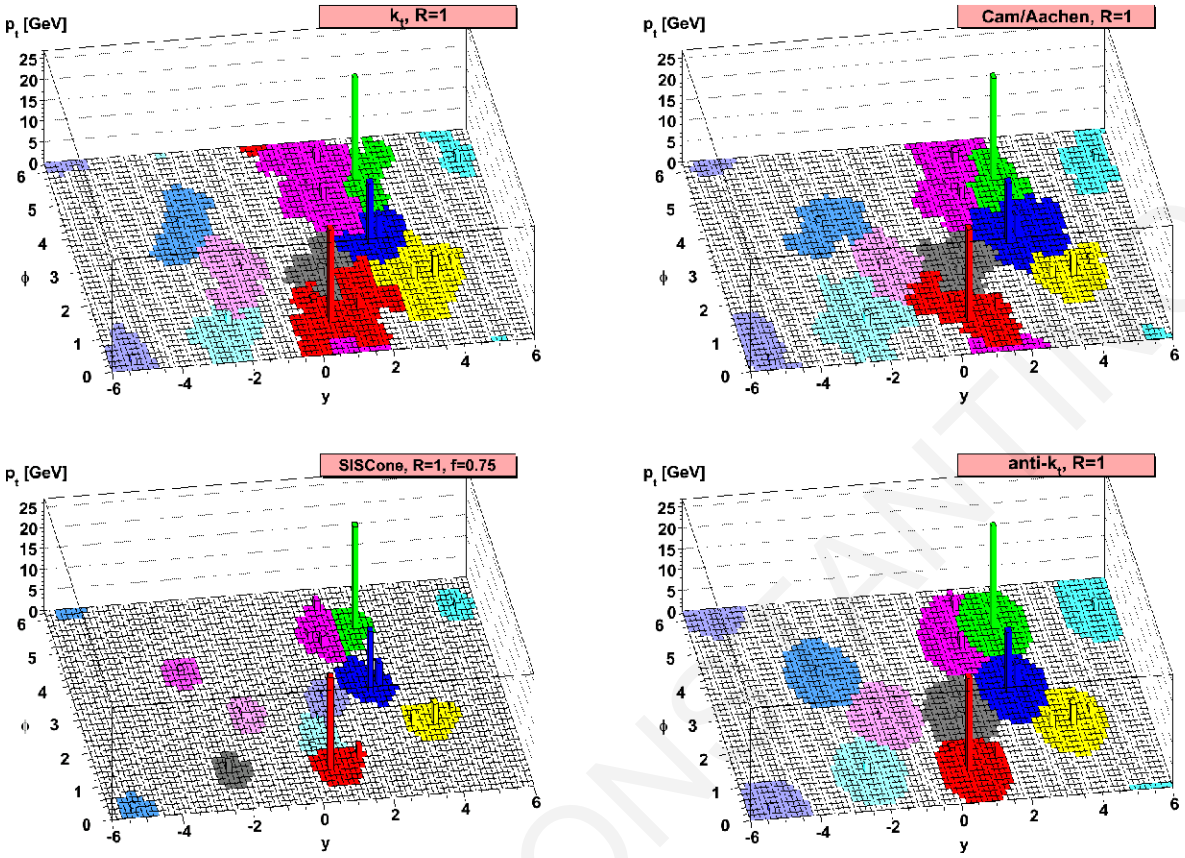


Figure 4.1: A sample parton-level event, together with random soft jets, clustered with k_{T}, Cambridge/Aachen, SISCone and anti-k_{T} jets algorithms [126].

In CMS analyses, jets from the hadronization of a single quark or gluon are reconstructed using the anti-k_{T} algorithm with a cone of radius R = 0.4 (AK4). Lorentz-boosted heavy particles with transverse momentum greater than their rest mass, pass their momentum to their decay products resulting in collimated particles. Because of their small distance, jets from hadronic decays of Lorentz-boosted particles are contained into a single large-radius jet, and cannot be reconstructed with the standard algorithm. A dedicated collection of jets is produced for these jets, using a radius of 0.8 (AK8) or 1.5 (AK15). Such jets can represent hadronic decays of $W^{\pm}$, Z^{0}, top quarks, Higgs bosons, etc.

Additional inelastic pp interactions produce additional tracks and energy clusters, which affect the measurement of the jet energy and momentum. For AK4 jets, the impact of the pileup offset is mitigated with the charged hadron subtraction (CHS) method [132] that removes charged PF particles identified as not originating from PVs before the jet clustering. For large-radius jets, the contribution from pileup is mitigated using the pileup per particle identification (PUPPI) method. For each particle, a local, shape-variable α is defined which exploits information from the p_{T} and the distance from all the particles in the vicinity as well as tracking information when available. The value of α is indicative of the particle originating from the hard scattering or pileup interactions. Based on the value of α a weight is assigned to each particle that scales the particle's four-momentum to correct for pileup at

particle-level.

Additional tight identification (TightID) criteria related to the jet substructure are applied to increase the efficiency of good-quality jets. The identification criteria used in Run II for jets with $|\eta| < 2.4$ are summarized in Table 4.1.

Table 4.1: Tight jet identification criteria for jets with $|\eta| < 2.4$, for the Run II data.

PF Jet TightID	
Neutral Hadron Fraction	< 0.90
Neutral EM Fraction	< 0.90
Number of Constituents	> 1
Muon Fraction	-
Charged Hadron Fraction	> 0
Charged Multiplicity	> 0
Charged EM Fraction	-
Number of Neutral Particles	-

The jet energy scale (JES) and jet energy resolution (JER) are calibrated to account for a series of effects. The pileup offset is calculated in simulated QCD-dijet events processed with and without pileup and the average difference in p_{T} is subtracted from both, data and simulation. Residual offset corrections between data and detector simulation are derived from minimum-bias data and simulated events using clustering particles in randomly placed cones. The aforementioned offset corrections are applied to the CHS jets only.

Effects due to the simulated detector response are corrected using QCD-dijet simulated events with the pileup-offset corrections applied. The jet response is defined as the ratio of the average p_{T} between the reconstructed jet and the matched particle-level jet:

$$R_{ptcl} = \frac{<p_T>}{<p_{T,ptcl}>} \tag{4.5}$$

1) Residual corrections that account for differences between data and detector simulation are used to correct jets in data. The response is quantified using two different methods. The missing-E_T projection fraction method (MPF) considers the response of the total hadronic activity to normalize the jet. The response is given by:

$$R_{jet,MPF} = 1 + \frac{\vec{p}_T^{\,miss} \cdot \vec{p}_{T,ref}}{p_{T,ref}^2} \tag{4.6}$$

where $\vec{p}_T^{\,miss}$ is the recoiling missing transverse momentum and $\vec{p}_{T,ref}$ is the p_{T} of reference objects. In dijet and multijet events, the reference object is a well measured jet in the

barrel region with $|\eta| < 1.3$. In the direct-balance (DB) method, the response is the jet-over-reference p_{T} ratio:

$$R_{DB} = \frac{p_{T,jet}}{p_{T,ref}} \tag{4.7}$$

2) Relative η-dependent corrections are estimated in dijet events with the MPF method, where the reference object is a jet of similar p_{T} within the barrel region. Additional data-to-simulation corrections are measured to correct the absolute jet p_{T} scale. Both MPF and DB methods are used in Z +jets events, where Z decays leptonically, γ+jets, multijet and hadronic $\mathrm{t\bar{t}}$ events. In Z/γ+jets, the Z or γ is used as reference, while in multijet and $\mathrm{t\bar{t}}$ events, the reference object is a jet with $|\eta| < 1.3$. The jet energy corrections (JEC) measurement with Run II CMS data is presented in [133].

4.4 B-tagged jets

Jets originating from the hadronization of heavy-flavor quarks such as bottom or charm are characterized by the presence of a hadron that carries a large fraction of the quark momentum. Heavy-flavor hadrons have a large lifetime and can travel a distance of a few millimeters before they decay. This results in displaced tracks originating from displaced vertices (SV). Several algorithms that take advantage of the heavy-flavor hadron properties to discriminate against light-flavor jets from light quarks or gluons have been developed in CMS. In this analysis, jets from the hadronization of bottom quarks (b jets) are identified with the DeepJet algorithm [134]. DeepJet is a deep neural network (DNN)-based algorithm seeded with $\sim$650 features from four collections: global variables, charged PF candidates, neutral PF candidates and SVs. The global variables include kinematic properties of the jet, the number of SVs and tracks inside the jet and the number of PVs in the event to reject pileup interactions. Charged PF-candidate features contain kinematic properties of the tracks, the track fit quality and displacement from the PV. Information from neutral PF candidates is exploited from the fraction of the jet momentum carried, the fraction of the energy deposited in the HCAL and the distance to the jet axis. Finally, the flight distance, the number of associated tracks, the track fit quality criteria and other kinematic properties are associated with the SVs. The full list of input features is provided in [134]. The algorithm is a multiclassifier that classifies the jets into six categories, three of them compatible with a jet originating from b hadrons while the remaining three, tag the jets as originating from charm, light quarks, or gluons. The three b-flavor categories refer to jets containing two b hadrons (bb), jets containing b hadrons decaying leptonically (b_{lep}), and jets with b hadrons decaying hadronically (b). The final b tagger is a combination of the three b-tag probabilities. The performance of the DeepJet was measured using 41.9 fb^{-1} and compared to the previous b tagging algorithm, the DeepCSV [135]. In Fig. 4.2, DeepJet shows significant improvement in the probabil-

ity of misidentifying light or charm jets as b jets with respect to the efficiency of correctly identifying b-tagged jets, using jets with $p_{\mathrm{T}} > 30\,\mathrm{GeV}$ and $|\eta| < 2.5$ [136]. Three selection working points (WPs) are defined that correspond to 10%, 1% and 0.1% probability of a light jet being misidentified as a b jet.

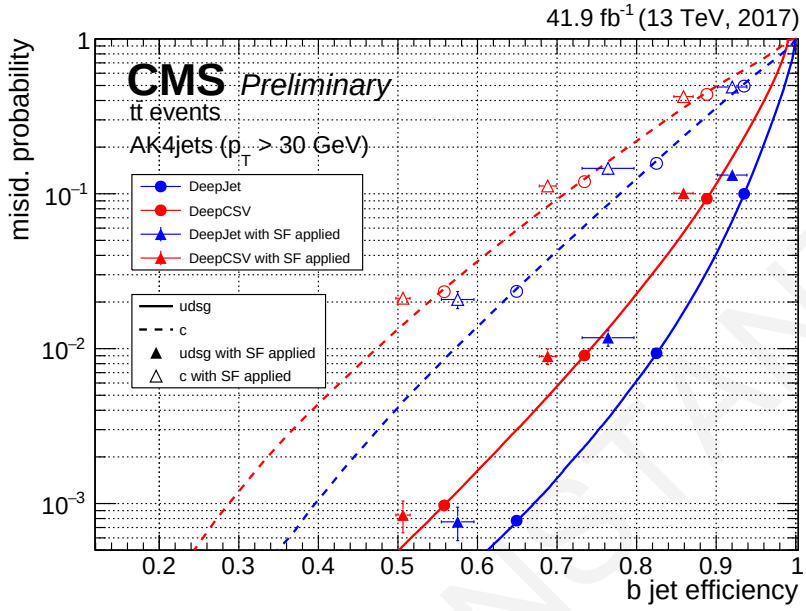


Figure 4.2: Performance of the DeepJet and DeepCSV algorithms using simulated jets with $p_{\mathrm{T}} > 30\,\mathrm{GeV}$ and $|\eta| < 2.5$ from $\mathrm{t\bar{t}}$ simulation [136].

4.5 Hadronic τ jets

Decays of τ-leptons can be either leptonic (τ_ℓ), producing a muon or an electron and two neutrinos, or hadronic (τ_h) with the presence of a ν_τ. Dominant hadronic decay modes include one charged hadron and up to two neutral pions (one-prong) or three charged hadrons (three-prong). The identification of τ_h-jets in CMS is established with the hadron-plus-strip (HPS) algorithm [137, 138]. The algorithm is seeded by AK4 jets which are tested for compatibility with isolated jets of low particle multiplicity with energy deposits in both ECAL and HCAL, indicating the presence of charged hadrons, and strips of ECAL energy deposits compatible with the decay of neutral pions to a pair of photons ($\pi^0 \to \gamma\gamma$). These strips are narrow in η and wide in ϕ to allow for the broadening of ECAL energy deposits from photon conversion within the tracker. Reconstructed τ_h-jet candidates are required to have a charge equal to ± 1 and a mass consistent with the corresponding decay mode.

Hadronic decays of τ leptons are discriminated against quarks or gluons jets, electrons or muons using a multiclass DNN-based algorithm, the DeepTau [139]. The algorithm exploits information from particle-level features from all the parts of the detector, and high-level information of the reconstructed τ-candidate to estimate the probability of the candidate belonging to each particle type.

4.6 Missing transverse momentum

Weakly interacting particles, such as neutrinos or BSM particles, escape the material layers of the detector without being detected. These particles can only be detected by the energy imbalance in the transverse direction, $\vec{p}_T^{\,\text{miss}}$. As a consequence of the energy and momentum conservation laws, $\vec{p}_T^{\,\text{miss}}$ is given by the negative vectorial sum of all the reconstructed particle's transverse momentum:

$$\vec{p}_T^{\,\text{miss}} = -\sum_{i}^{reco} \vec{p}_{T,i} \tag{4.8}$$

The JEC are propagated to the measurement of the corrected $\vec{p}_T^{\,\text{miss}}$:

$$\vec{p}_T^{\,\text{miss}} = \vec{p}_T^{\,\text{raw}} + \sum_{i}^{\text{jet}} \vec{p}_{T,i}^{\,\text{raw}} - \sum_{i}^{\text{jet}} \vec{p}_{T,i}^{\,\text{JEC}} \tag{4.9}$$

In the above equation, the superscript "raw" indicates that the JEC is not applied, while "JEC" denotes the corrected jets. $\vec{p}_T^{\,\text{raw}}$ refers to the $\vec{p}_T^{\,\text{miss}}$ in Eq. (4.8). The total $\vec{p}_T^{\,\text{miss}}$ combines the transverse momentum of the invisible particles, as well as detector inefficiencies and mismeasurement of the reconstructed particles.

4.7 Scalar transverse momentum

The scalar transverse momentum is defined as the scalar sum of the transverse momenta of all the reconstructed jets:

$$H_T = \sum_{i}^{jets} |\vec{p}_{T,i}| \tag{4.10}$$

The H_{T} is measured after applying all the identification criteria and the JES and JER corrections.

5 Machine learning

Machine learning (ML) algorithms [140] in high energy physics (HEP) are increasingly used in many aspects in recent years. The expanding amount of collision data delivered by the CMS detector and the complexity of the event topologies make the analyses more challenging and time-consuming. From event triggering and object identification to detector simulation and offline data analysis, the application of ML is introduced to enhance HEP searches [141–143].

Traditional data-analysis searches in HEP use cut-based approaches in which a sequence of selection requirements is applied to discriminating variables to distinguish signal from background events. Such approaches overlook correlations between variables and interesting events may be rejected if just one variable fails the selection criteria. Unlike cut-based methods, ML techniques combine information from multiple variables into one final discriminator. The main ML methods used in HEP are the BDT and DNN. In this analysis, DNNs have been developed for particle-flavor identification and signal extraction. A brief description of ML and DNN is given in the following sections.

5.1 Machine Learning Techniques

ML is a subset of artificial intelligence (AI) that enables computer systems to solve problems based on acquired knowledge. This is achieved by extracting patterns from raw data and making decisions that seem subjective. An ML algorithm is said to learn when its performance at a given problem or task improves with experience. The input of an algorithm is a collection of samples, called dataset. For each sample, a set of features is given, each of them providing useful information to the learning algorithm. A sample that consists of n features can be represented as a vector $\mathbf{x} \in \mathbb{R}^n$. The dataset used for the learning process is called training dataset.

Based on the kind of experience that ML algorithms obtain, the learning can be characterized as supervised, unsupervised or reinforcement. In supervised learning, each sample is associated with a label or a target. The output of the algorithm is a prediction of the value of the target. In unsupervised algorithms, no label is related to the input features and the algorithm is supposed to discover useful properties and patterns from the features of the input dataset. Finally, reinforcement algorithms learn via trial and error to interact with an environment by taking a series of actions that maximize the notion of reward. Reinforcement

learning is applied in technologies of self-driving cars, AI gaming and automated robots.

The most common types of supervised tasks are classification and regression. In classification tasks, the algorithm aims to classify the inputs in k categories. Each category is labeled by a numerical value y and the output value of the classifier $\hat{y} = f(\mathbf{x})$ specifies in which of the k categories a sample $\mathbf{x}$ belongs. In classification tasks of two categories, y usually takes the values 0 and 1 and $\mathbf{x}$ belongs to the category with $y = 0$ if the output value $\hat{y} = f(\mathbf{x})$ is close to 0 and vice versa. The output of the classifier can be either a distinct value or a continuous value that can be considered as a probability distribution over the classes. In HEP searches, classification is commonly used to discriminate signal ($y = 1$) from background ($y = 0$) samples for object or event identification. In regression tasks, the algorithm aims to predict several target features, given a set of input features $\mathbf{x}$. In collider physics, it is used to predict and hence correct the measurements of a particle's properties such as the energy, momentum, position or mass.

The performance of an ML algorithm is tested on a dataset that is unknown to the algorithm. This dataset is orthogonal to the one used for training and it is called the test dataset. A reliable algorithm is considered when training and test datasets give similar performance. If a significant performance degradation appears during the testing, the algorithm is said to be overtrained.

5.2 Deep Learning

Deep learning is an ML method that represents complex concepts as a nested sequence of simpler concepts [144, 145]. The models are composed of multiple processing layers to learn representations of data with multiple-level features. Deep learning models are also known as artificial neural networks (ANN), deep neural networks (DNN) or neural networks (NN) because they are inspired by the information processing of the biological brain. Just like the biological brain, deep learning models consist of neurons that perform parallel computations very fast and they can learn different tasks through experience.

Mathematically, a DNN can be expressed as $y = f(\mathbf{x}; \boldsymbol{\theta})$, where f is the learner (e.g. classifier or regressor), $\mathbf{x}$ is the vector of the input features, and $\boldsymbol{\theta}$ is the set of the parameters that need to be adjusted such that the deviation between the $f(\mathbf{x}; \boldsymbol{\theta})$ and the real value of the target y is minimized. The function f is determined by the architecture of the DNN. Fully-connected DNNs consist of the input layer which takes as input the $\mathbf{x}$, the hidden layers and the output layer. A fully-connected DNN can be visualized in Fig. 5.1. The input is a dataset with four features, the two hidden layers consist of five and six neurons each and the output layer returns the response of the DNN.

The input of each hidden layer is a transformation of the output of the previous layer and is given by:

$$\mathbf{z} = g(\mathbf{W}^{\mathbf{T}}\mathbf{x} + \mathbf{b}) \tag{5.1}$$

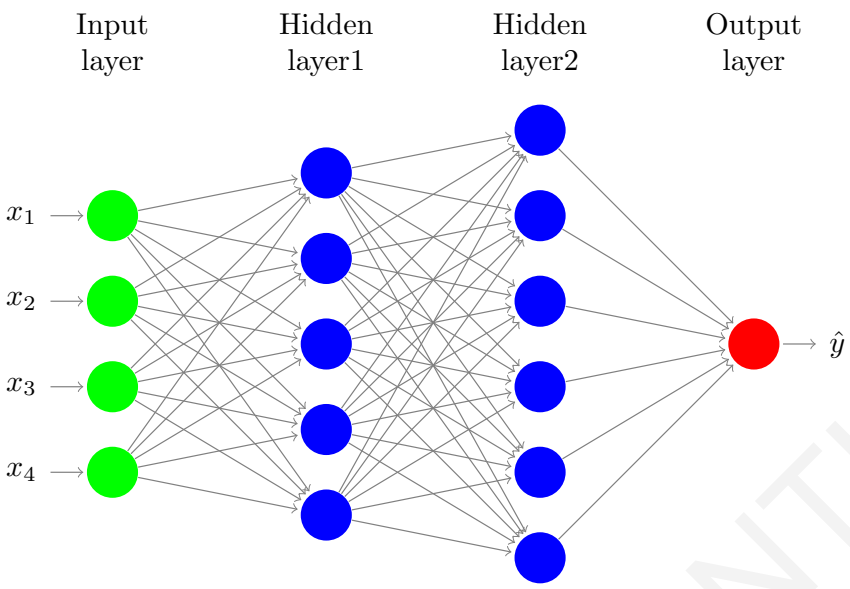


Figure 5.1: Schematic view of a fully connected neural network

where **W** is an array with the trainable parameters or weights and **b** is a bias term that shifts the value of the function so that it does not need to pass through the origin. The transformation g is called activation function. The size of the output vector **z** of each layer is given by the number of neurons or nodes of the layer. The output layer returns the predicted value $\hat{y}$ of the DNN.

$$\hat{y} = g(\sum_i W_i \mathrm{x}_i + b) \tag{5.2}$$

The training of a DNN occurs in multiple iterations through the entire dataset, called epochs. The dataset is split into subsets or batches that are used to update the trainable parameters within an epoch. The deviation between the response of the learner and the truth value y can be estimated by a loss function $\mathcal{L}(\mathbf{x}, \boldsymbol{\theta})$ and the minimization of the deviation can be found by solving the gradient of the $\mathcal{L}$:

$$\nabla_{\boldsymbol{\theta}} \sum_i^m \mathcal{L}(\mathbf{x}^{(i)}; \boldsymbol{\theta}) = 0 \tag{5.3}$$

with m being the batch size. The minimization of the deviation between y and $\hat{y}$ is estimated using sophisticated algorithms that calculate the extreme values of a function, called optimizers.

For classification tasks, the most common loss function is the cross-entropy, defined as:

$$CE = -\sum_i^n y_i \log(\hat{y}_i(\mathbf{x}; \boldsymbol{\theta})) \tag{5.4}$$

where i runs over the n categories of the classifier. For binary classification (e.g. signal or background), the cross-entropy becomes:

$$BCE = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y}) \tag{5.5}$$

where the label y is equal to 1 for signal and 0 for background. For regression tasks the most common loss functions are the mean squared error (MSE) and the mean squared logarithmic error (MSLE):

$$MSE = \frac{1}{m} \sum_{i}^{m} (\hat{y}^{(i)} - y^{(i)})^2 \tag{5.6}$$

$$MSLE = \frac{1}{m} \sum_{i}^{m} (\log(\hat{y}^{(i)} + 1) - \log(y^{(i)} + 1))^2 \tag{5.7}$$

The dataset seen during the training is split into training and validation datasets. The training dataset is used to fit the model and learn the parameters, while the validation dataset is used as a guide to update specific hyperparameters accordingly. At the end of each epoch, the model iterates over the validation dataset to evaluate the validation loss and other metrics, and the performance is compared to the one of the training datasets. An application of the validation process is the use of early stopping, where the training stops when the model performance converges or worsens in terms of specific metrics (loss function, area under curve, etc.).

6 Statistical methods

In this section, the statistical methods used to extract possible signal in the presence of background is discussed. In the absence of an observed signal, this is done by setting upper exclusion limits on the signal strength using a modified frequentist method [146, 147]. The level of incompatibility of data with the simultaneous presence of signal and background is expressed as a confidence level. In searches for new resonances exclusion upper limits are often expressed on the product of the production cross section and the branching fraction. In this case, the signal strength μ is:

$$\mu = \sigma_{\mathrm{pp}\to\mathrm{t(b)H}^{\pm}} \times \mathcal{B}(\mathrm{H}^{\pm} \to \mathrm{tb}). \tag{6.1}$$

The expected signal and background yields are subject to multiple systematic uncertainties which are described as a set of nuisance parameters θ and can be expressed as functions of θ: $s(\theta)$ and $b(\theta)$. For searches that bin the event yields in the discriminant variable, each bin can be treated as statistically independent and the number of signal and background events that fall in bin i are denoted as $s_i(\theta)$ and $b_i(\theta)$. At the LHC ATLAS and CMS experiments, the upper limits are determined by a maximum likelihood fit of a discriminant variable [148], where the likelihood function $\mathcal{L}$ is defined as follows:

$$\mathcal{L}\left(\text{data}|\mu,\theta\right) = \text{Poisson}\left(\text{data}|\mu\cdot s(\theta)+b(\theta)\right)\cdot\rho\left(\tilde{\theta}|\theta\right) \tag{6.2}$$

where "data" refers to either observed or simulated events. In the case of a binned likelihood, the Poisson distribution is expressed as a product of probabilities to observe n_i data in the i^{th} bin:

$$\text{Poisson}\left(n_i|\mu\cdot s_i(\theta)+b_i(\theta)\right) = \prod_i \frac{\left(\mu\cdot s_i(\theta)+b_i(\theta)\right)^{n_i}}{n_i!} e^{-(\mu\cdot s_i(\theta)+b_i(\theta))}. \tag{6.3}$$

In the above equations, $\mu\cdot s_i(\theta)+b_i(\theta)$ is the total event yields in bin i, in the signal+background hypothesis, where the signal events are scaled by the signal strength μ. The last term in Eq. (6.2) represents the systematic uncertainty probability density function (PDF) that determines the probability for the true parameter values θ to coincide with the best-guess values $\tilde{\theta}$. Following the Bayesian theorem, $\rho(\tilde{\theta}|\theta)$ can be reinterpreted as posteriors arising from some real or imaginary measurements of $\tilde{\theta}$ as:

$$\rho(\tilde{\theta}|\theta) \sim p(\tilde{\theta}|\theta) \cdot \pi_\theta(\theta) \tag{6.4}$$

where $\pi_\theta(\theta)$ functions are hyper-priors for the "measurements" and p are the auxiliary measurement PDFs. The choice of the p function depends on the type of systematic uncertainties. In this analysis, for rate systematic uncertainties, the log-normal PDF is used. In the case of shape systematic uncertainties, the final fit discriminant with the systematic uncertainty variations applied is provided as a template.

6.1 Observed limits

To discriminate signal-like from background-like events, a test statistic $\tilde{q}_\mu$ is designed as the likelihood ratio that follows a modified frequentistic statistical approach [149]:

$$\tilde{q}_\mu = -2 \ln \frac{\mathcal{L}\left(\text{data}|\mu, \hat{\theta}_\mu\right)}{\mathcal{L}\left(\text{data}|\hat{\mu}, \hat{\theta}\right)} \tag{6.5}$$

with $\hat{\mu}, \hat{\theta}$ being the global maximum of the likelihood and $\hat{\theta}_\mu$ is the conditional maximum likelihood estimator of θ for a given μ. The allowed values of μ are constrained in the interval $0 < \hat{\mu} < \mu$. The upper bound ($\hat{\mu} < \mu$) ensures a physical one-sided limit on the upper values of μ, allowing upward fluctuations without considering them as evidence against the signal hypothesis, while the lower bound ($\hat{\mu} > 0$) protects against negative signal rates. Maximizing the above equation, one can evaluate the observed value for a given μ under test, $\tilde{q}_\mu^{\text{obs}}$, to obtain the nuisance parameter values that describe best the data, in the signal+background and background-only (μ=0) hypotheses, denoted as $\hat{\theta}_\mu^{obs}$ and $\hat{\theta}_0^{obs}$ respectively. The probabilities for the $\tilde{q}_\mu^{\text{obs}}$ to be as or less compatible with the signal+background or the background-only hypothesis are defined by:

$$\begin{aligned} \text{CL}_{\text{s}+b} &= P\left(\tilde{q}_\mu \geq \tilde{q}_\mu^{\text{obs}}|\text{signal+background}\right) = \int_{\tilde{q}_\mu^{\text{obs}}}^{\infty} f\left(\tilde{q}_\mu|\mu, \hat{\theta}_\mu^{\text{obs}}\right) \text{d}\tilde{q}_\mu \\ \text{CL}_{\text{b}} &= P\left(\tilde{q}_\mu \geq \tilde{q}_\mu^{\text{obs}}|\text{background only}\right) = \int_{\tilde{q}_0^{\text{obs}}}^{\infty} f\left(\tilde{q}_\mu|0, \hat{\theta}_0^{\text{obs}}\right) \text{d}\tilde{q}_\mu. \end{aligned} \tag{6.6}$$

PDFs f are distributions generated with toy MC "pseudo-data". They are obtained for a given μ in the signal+background and μ = 0 for the background-only hypothesis, while keeping the $\hat{\theta}_\mu^{\text{obs}}$ or $\hat{\theta}_0^{\text{obs}}$ fixed to the maximum likelihood values [150]. An example of the f PDFs is shown in Fig. 6.1 for μ = 1 and μ = 0.

Finally, the ratio of CL_{s+b} and CL_b is used to set upper limits on μ,

$$\text{CL}_{\text{s}}(\mu) = \frac{\text{CL}_{\text{s}+b}(\mu)}{\text{CL}_{\text{b}}(\mu)} \leq \alpha. \tag{6.7}$$

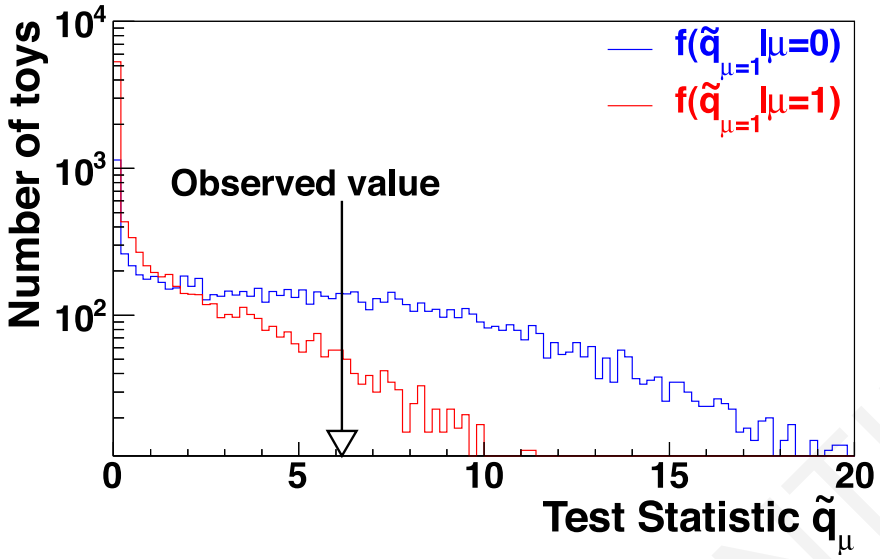


Figure 6.1: Test statistic distributions of pseudo-data generated for the signal+background and the background-only hypotheses. The figure is taken from [148].

The statement $\mathrm{CL_s}(\mu) < \alpha$ states exclusion of signal with strength μ with 1-α confidence level. This means that in a repetition of the experiment, a fraction of at most α experiments will falsely exclude the signal. With the choice of α = 0.05, 95% CL upper limits on μ are set.

6.2 Expected limits

The search for a new particle at the LHC begins with a data-blinded approach to minimize biases in the search strategy. After conducting a thorough cross-check of the methods employed, the last step is the unblinding of the data. At the blinded stage, expected limits are estimated in the background-only hypothesis using simulated background pseudo-data. The CL_s and $\mu^{95\%CL}$ are calculated following the methodology described to estimate the observed limits. The procedure is repeated to create a distribution of the $\mu^{95\%CL}$ results, which is then used to build a cumulative probability distribution of $\mu^{95\%CL}$. The value of $\mu^{95\%CL}$ that corresponds to 50% cumulative probability is the median expected value of $\mu^{95\%CL}$. The $\pm\,1\sigma$ and $\pm\,2\sigma$ bands of $\mu^{95\%CL}$ are also derived and correspond to 68% and 95% respectively.

6.3 Significance of an excess of events

The observation of an excess of events needs further investigation to claim a new particle discovery. The scenario of a fluctuation in the background is quantified using a p-value that describes the probability for the background to give an excess of events as large or larger than the observation. A test statistic is defined in the background-only hypothesis,

$$q_0 = -2 \ln \frac{\mathcal{L}\left(\text{data}|0, \hat{\theta}_0\right)}{\mathcal{L}\left(\text{data}|\hat{\mu}, \hat{\theta}\right)} \tag{6.8}$$

with the constraint $\hat{\mu} \geq 0$ and "data" being a toy MC "pseudo-data". The p-value that corresponds to q_0^{obs} is given by

$$p_0 = P\left(q_0 \geq q_0^{\text{obs}}\right) = \int_{q_0^{\text{obs}}}^{\infty} f\left(q_0|0, \hat{\theta}_0^{\text{obs}}\right) \mathrm{d}q_\mu. \tag{6.9}$$

The significance Z of an excess describes the number of standard deviations. It is obtained by converting the above p-value, with the convention of a one-sided normal distribution:

$$p = \int_Z^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = \frac{1}{2} P_{\chi_1^2}(Z^2) \tag{6.10}$$

where $P_{\chi_1^2}$ is a survival function of χ^2 for one degree of freedom. In the asymptotic limit the significance approximates:

$$Z \approx \sqrt{q_0^{\text{obs}}}. \tag{6.11}$$

A significance of 3σ (Z = 3) indicates significant evidence of possible signal and corresponds to a p-value of $\sim$0.00135. A discovery can be stated in the presence of an excess of 5σ with a probability of background fluctuation of 2.8×10^{-7}.

7 Resolved top quark identification

7.1 Introduction

A resolved top quark (t^{res}) algorithm that reconstructs and identifies hadronic decays of top quarks with moderate p_T is presented. Top quarks decay into a W and a b quark with a probability ~ 1, and in the hadronic final state, W decays into a quark-antiquark pair. The t^{res} tagger is a classifier based on fully connected DNNs [144, 145] that targets top quarks whose decay products are resolved as three separate anti-k_T [127, 128] jets with a distance parameter of 0.4. This MVA classifier utilizes high-level information from each of the three seed jets and the jet systems, such as invariant masses, angular separations, jet flavor, and jet shape variables. To prevent mass sculpting effects, the classifier's output is decorrelated from the top quark mass using the sample-reweighting technique. The misidentification rate and tagging efficiency of the algorithm are estimated to extract scale factors that correct the simulation. Systematic uncertainties that affect the performance of the algorithm have been measured.

7.2 Event Selection

Hadronically decaying resolved top quarks are studied in $t\bar{t}$ events decaying either fully hadronically or semileptonically where one of the top quarks decays to a muon and the other hadronically. Fully hadronic $t\bar{t}$ decays are selected using a combination of the HLT that collect (a) multijet events with the requirement of one or two b-tagged jets (b) events with at least one AK8 jet and (c) high-H_T events. To ensure a fully hadronic final state, events with isolated electrons or muons with $p_T > 10$ GeV and $|\eta| < 2.4$ and hadronically decaying τ leptons with $p_T > 20$ GeV and $|\eta| < 2.3$ are excluded. Events with semileptonic $t\bar{t}$ decays are triggered with the single muon `HLT_IsoMu24` for the 2016 and 2018 data-taking periods and `HLT_IsoMu27` for the data collected in the 2017 LHC run. As for the case of fully hadronic decays, events with isolated electrons and hadronically decaying τ leptons are vetoed using the same criteria. The presence of exactly one isolated muon is required, with $p_T > 29$ GeV and $|\eta| < 2.4$. In both final states, events with 5 or 6 AK4 jets identified as tight with $p_T > 40$ GeV, $|\eta| < 2.4$ are selected. At least one of the selected jets must be b-tagged, satisfying the medium WP of the DeepJet [151] b tagging algorithm. An additional requirement of $H_T > 500$ GeV is applied for the sample with fully hadronic

decays, which corresponds to the plateau of the trigger efficiency.

7.3 Algorithm

7.3.1 Training objects

A DNN is trained on $t\bar{t}$ simulated events to discriminate between trijet combinations originating from the decay of top quarks (signal) and other combinatorial trijet systems (background). The signal t^{res} candidates are matched to the generated top quark decay products, while t^{res} candidates with at least one non-matched jet are considered as background. A t^{res} candidate is considered as matched if each jet is matched to one of the quark originating from the top quark at generator level. The criteria that a jet (j) must fulfill to be regarded as matched to a parton (q) are based their proximity in the (η, ϕ) space, requiring $\Delta R < 0.3$ and their difference in p_{T}. To quantify the p_{T} requirement, a fit on the $\frac{\Delta p_T(j-q)}{p_T(q)}$ distribution is performed including jet-to-quark pairs that pass the $\Delta R(j,q) < 0.3$ requirement, in bins of the quark p_{T}. The analytic function used in the fit is a Gaussian+Crystal ball. The final selection corresponds to 2 standard deviations (2σ) for $p_T(q) > 40$ GeV and 3σ for $p_T(q) < 40$ GeV on the fitted distribution. The matching criteria are summarized in Eq. 7.1:

$$\Delta R(q,j) < 0.3 \text{ and } \frac{\Delta \mathrm{p_T(q,j)}}{\mathrm{p_{T,q}}} < \begin{cases} 3\sigma \approx 0.50, & p_{T,q} < 40 \text{ GeV} \\ 2\sigma \approx 0.33, & p_{T,q} > 40 \text{ GeV} \end{cases} \tag{7.1}$$

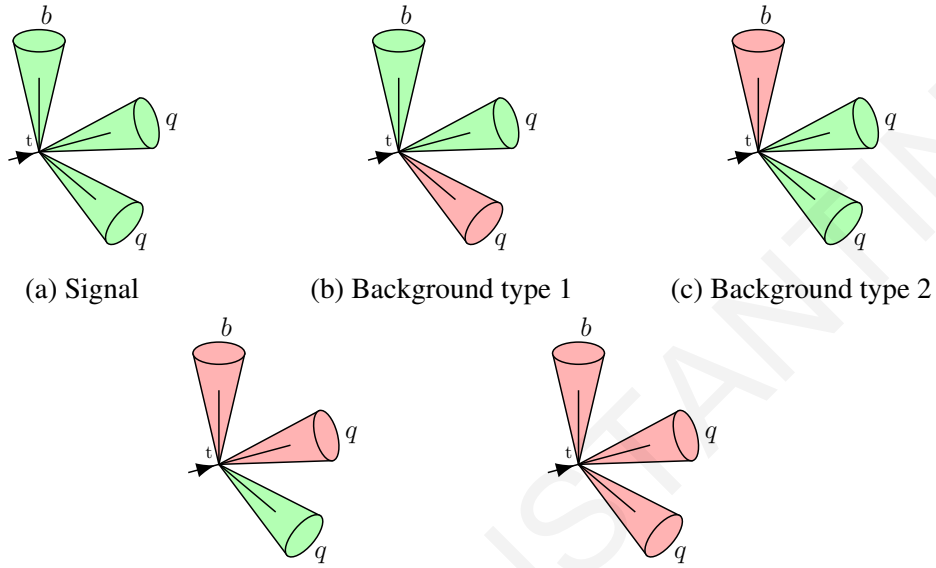
In background t^{res} candidates, each jet in the trijet system is characterized as a jet from the W boson decay or the b-quark jet based on the number of jets fulfilling the matching criteria or the flavor of the matched quark. The b quark and the quarks from the W boson are allowed to match any of the selected jets, regardless of being b-tagged or not. The assignment of the jets is done as follows:

- Background type 1: b-quark and one quark from the W-boson decay are matched:
  In the case where two jets are matched to the b-quark and a quark from the W boson decay, the remaining jet is considered as the second jet from the W-boson decay.

- Background type 2: W-boson products are matched:
  If exactly two jets are matched to the W-boson decay products then the remaining non-matched jet is considered as the jet from the b-quark (b jet)

- Background type 3: Exactly one quark is matched:
  If exactly one jet is matched to a quark from the W-boson decay, the b jet is selected to be the one with the highest b tagging discriminant value, between the remaining jets. If exactly one jet is matched to the b quark, the remaining jets belong to the W-boson decay.

- Background type 4: None of the top-quark decay products is matched:
  The jet with the highest b tagging discriminant value is considered as the b jet and the remaining jets belong to the W-boson decay if none of the jets are matched to the top-quark decay products.



(a) Signal (b) Background type 1 (c) Background type 2

(d) Background type 3 (e) Background type 4

Figure 7.1: Signal (a) and the four categories of background (b-e) t^{res} candidates considered. Truth-matched jets are shown with green color while non-matched jets are shown with red color.

The p_T distribution of the signal and background t^{res} candidates is shown in Fig. 7.2. The t^{res} algorithm targets t^{res} with $p_T \in$ (200-500) GeV. The low-p_T region is dominated by combinatorial t^{res} candidates while for high-p_T values above 500 GeV the decay products of the top quark start to coalesce in a single jet and cannot be reconstructed as three separated jets.

7.3.2 Input features

The input of the DNN is a set of 33 discriminating variables of the seed jets and the jet systems. The majority of the selected variables are uncorrelated to each other so that each of them gives independent information to the algorithm. Furthermore, all the variables are uncorrelated with the top mass to minimize possible correlations between the t^{res}-candidate mass ($m_{t^{res}}$) and the associated classifier output. The input variables are listed in Table 7.1 and the corresponding distributions are shown in Fig. 7.3 and Fig. 7.4.

The individual-jet variables include masses and kinematic properties, flavor discriminators and jet-shape properties [152]. The jet flavor discriminators are the DeepJet b tagging discriminator and the charm discriminators **CvsL** and **CvsB** that discriminate charm from light and b jets respectively. In both charm discriminators, charm-like jets tend to have an

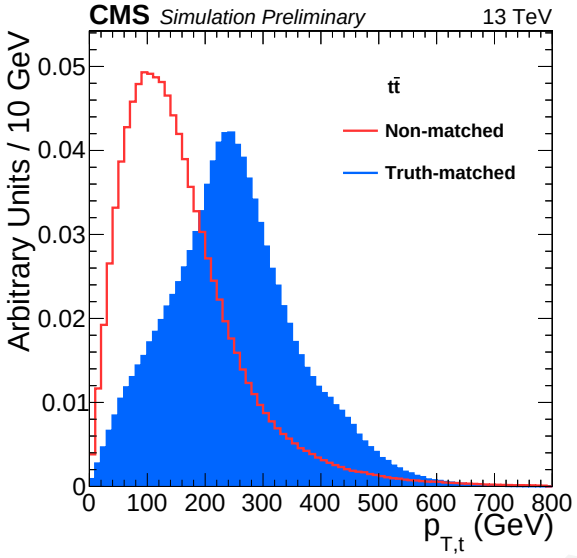


Figure 7.2: Distribution of the t^{res} p_T for signal and background candidates.

output score closer to 1. Jets originating from the W boson are expected to get higher CvsL values since the probability of a W boson decaying into cs quarks is $\sim$30%.

Three jet-shape variables are introduced to discriminate gluon from quark jets [152]. Approximating the shape of a jet as an ellipse on the $\eta-\phi$ plane, the **axis2** variable describes the minor axis of the jet. Jets originating from quarks are expected to be more narrow than gluon jets. The number of constituents composing a jet is described by the variable called **mult**. Quark jets tend to consist of a smaller number of sub-jets compared to jets from gluons splitting and therefore the quark jets exhibit lower constituent multiplicity. Taking into account that quark jets contain a small number of sub-jets, it is more likely that the constituents carry a large fraction of the jet energy. Thus, the fragmentation function $\boldsymbol{p_T D}$ defined as:

$$p_T D = \frac{\sqrt{\sum_i p_{T,i}^2}}{\sum_i p_{T,i}} \tag{7.2}$$

has a higher value for quark jets compared to gluon jets.

Other jet variables are the jet mass, the p_T of the jet in the center of mass of the t^{res} candidate (top CM) and the ratio of the absolute p_T difference between the jets and the t^{res} candidate divided by the scalar sum of their p_T.

Variables related to the jet systems describe the properties of the reconstructed resolved top quark (trijet) and W boson (dijet). The **mass** of the trijet and the dijet systems are calculated as the sum of the four-vectors of the jets composing the t^{res} and the W boson respectively. For all the mass-related variables, a narrow distribution with a peak at the mass value of the reconstructed object is expected when all the sub-jets are decay products of the same particle.

Table 7.1: List of input variables. The b-index refers to b-tagged jet properties, and $\mathrm{j}_1^{\mathrm{W}}(\mathrm{j}_2^{\mathrm{W}})$ indices refer to the leading (subleading) in p_T jet from the W boson decay.

Jet variables ($\mathrm{j}_1^{\mathrm{W}}$, $\mathrm{j}_2^{\mathrm{W}}$: W boson subjets, b: bjet)	
mass	($\mathrm{j}_1^{\mathrm{W}}$, $\mathrm{j}_2^{\mathrm{W}}$, b)
p_{T} in top CM	($\mathrm{j}_1^{\mathrm{W}}$, $\mathrm{j}_2^{\mathrm{W}}$, b)
$\frac{\lvert \mathrm{p_{T,jet}} - \mathrm{p_{T,top}} \rvert}{\mathrm{p_{T,jet}} + \mathrm{p_{T,top}}}$	($\mathrm{j}_1^{\mathrm{W}}$, $\mathrm{j}_2^{\mathrm{W}}$, b)
b tagging discriminator	(b)
charm to light (CvsL) discriminator	($\mathrm{j}_1^{\mathrm{W}}$, $\mathrm{j}_2^{\mathrm{W}}$, b)
charm to bottom (CvsB) discriminator	($\mathrm{j}_1^{\mathrm{W}}$, $\mathrm{j}_2^{\mathrm{W}}$, b)
minor jet axis (axis2)	($\mathrm{j}_1^{\mathrm{W}}$, $\mathrm{j}_2^{\mathrm{W}}$, b)
fragmentation function ($p_{\mathrm{T}}D$)	($\mathrm{j}_1^{\mathrm{W}}$, $\mathrm{j}_2^{\mathrm{W}}$)
constituent multiplicity	($\mathrm{j}_1^{\mathrm{W}}$, $\mathrm{j}_2^{\mathrm{W}}$)
Jet system variables	
top mass	
W mass	
$\mathrm{p_T}\Delta\mathrm{R_{top}}$ ($p_{\mathrm{T,top}}\Delta\mathrm{R(W,b)}$)	
$\mathrm{p_T}\Delta\mathrm{R_W}$ ($\mathrm{p_{T,W}}\Delta\mathrm{R}(\mathrm{j}_1^{\mathrm{W}}, \mathrm{j}_2^{\mathrm{W}})$)	
$\frac{\mathrm{p_{T,W}}}{\mathrm{p_{T,j_1^W}} + \mathrm{p_{T,j_2^W}}}$	
$\cos\omega(\mathrm{j_i},\mathrm{j_k}) = \frac{\tilde{\mathrm{p}}_{\mathrm{j_i}} \cdot \tilde{\mathrm{p}}_{\mathrm{j_k}}}{\lvert \mathrm{p_{j_i}} \rvert \lvert \mathrm{p_{j_k}} \rvert}$ in top CM	($\mathrm{j_i}$,$\mathrm{j_k}$ = $\mathrm{j}_1^{\mathrm{W}}$, $\mathrm{j}_2^{\mathrm{W}}$, b)
SoftDrop n2	
$\Delta\eta$(W,b) = $\lvert \eta_{\mathrm{W}} - \eta_{\mathrm{b}} \rvert$	

The variable $\mathrm{p_T}\Delta\mathrm{R}$ between two jets j_1, j_2 is defined as the product of the p_T of the j_1+j_2 system and the $\Delta\mathrm{R}$ distance between the two jets. The variable differentiates jets originating from gluon splitting, soft radiation and combinatorial jet combinations. For jets produced from top quarks or W bosons, a fixed value is expected that is approximately twice the mass of the top or W boson respectively.

The ratio of the vectorial (four-vector) over the scalar sum of the p_{T} of the jets composing the W-boson candidate is given by the variable $\frac{\mathbf{p_{T,W}}}{\mathbf{p_{T,j_1^W}} + \mathbf{p_{T,j_2^W}}}$. The higher the fraction is the smaller the distance between the two jets.

The three-dimensional angle between two jets j_i, j_k in the top CM can be expressed as the cosine of the angle between the momentum vectors of the two jets, $\mathbf{cos\omega}$. By definition, cosω is given by the inner product of the momentum vectors divided by their magnitude. The cosω variable is used for all the jet combinations in the trijet system.

The **SoftDrop n2** variable is inspired by the *softdrop* de-clustering algorithm [153]. It is used to distinguish final-state jets from wide-angle soft jets and soft collinear jets. The general condition is defined as:

$$\textit{softdrop} = \frac{\min\left(p_{\mathrm{T},1}, p_{\mathrm{T},2}\right)}{p_{\mathrm{T},1} + p_{\mathrm{T},2}} > z_{cut} \cdot \left(\frac{\Delta R_{12}}{R}\right)^{\beta} \tag{7.3}$$

where $p_{T,1}$ and $p_{T,2}$ are the transverse momenta of the two test jets, ΔR_{12} is the distance between the jets in the η-ϕ plane and z_{cut}, β are free parameters. The z_{cut} parameter determines the strength of the fractional p_{T} selection and β can be set to values greater than zero to soften the selection for collinear radiation. For $\beta > 0$, the algorithm rejects wide-angle radiation, while for $\beta < 0$, collinear jets fail the condition. The SoftDrop n2 variable that corresponds to $\beta = 2$ is used as an input variable for the t^{res} tagger.

The distance between the W-boson candidate and the b jet in the η direction is also used as an input to the t^{res} tagging algorithm.

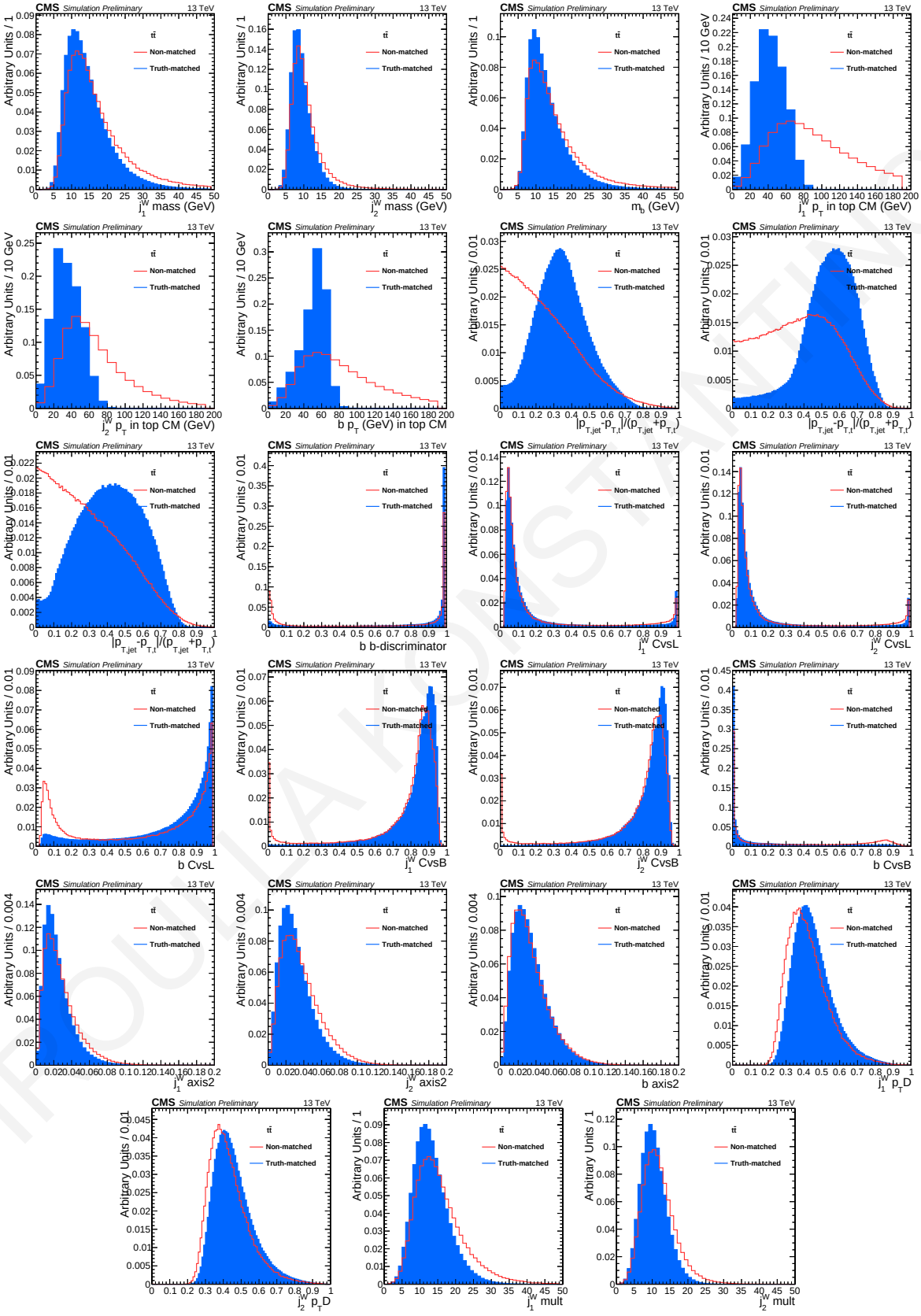


Figure 7.3: Distributions of the individual-jet variables used as input for the training of the t^{res} tagger.

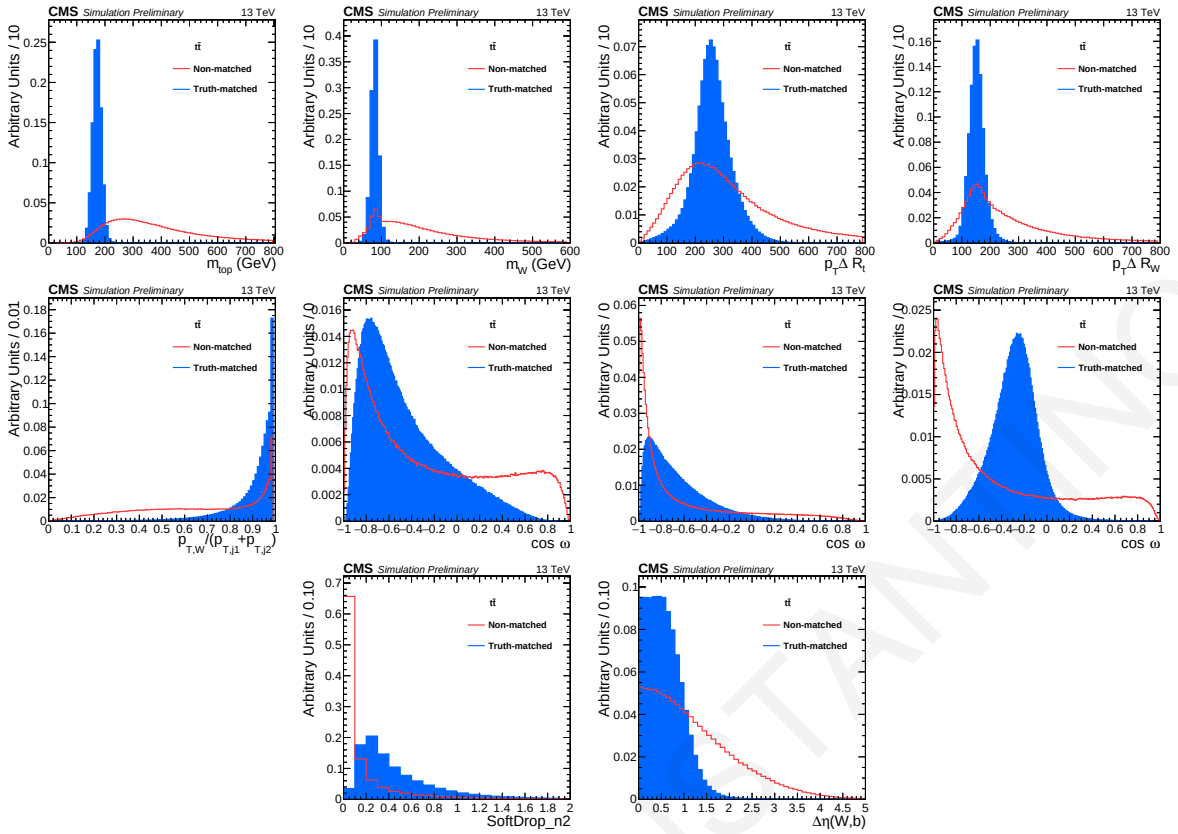


Figure 7.4: Distributions of the variables of the trijet and dijet systems used as input for the training of the t^{res} tagger.

7.3.3 Architecture

The design of a t^{res} tagger based on DNN architecture is implemented using the KERAS [154] and TENSORFLOW [155] software packages. A fully connected DNN is built and consists of three hidden layers with 32 neurons each. All hidden layers use the Rectified Linear Unit (ReLU = $\max(x, 0)$) activation function while the output layer follows the sigmoid activation function defined as:

$$\text{sigmoid} = \frac{1}{(1 + e^{-x})} \tag{7.4}$$

The Adam [156] optimizer is used with a learning rate of 0.001 to minimize the binary cross-entropy loss function $\mathcal{L}$ defined in Eq. (5.5).

A dataset of two million t^{res} candidates is used which consists of equal signal and background samples. For training and validation, 40% of the total samples are used and the remaining 20% of the samples are used for testing.

7.3.4 Data preprocessing

To improve the stability and performance of the learning algorithm the input features are transformed such that they are distributed in similar ranges and not influenced by outliers.

This is often necessary in the presence of non-normally distributed data when variables have long tails or the ranges of the input features differ over orders of magnitude. The t^{res} tagger uses the robust scaler preprocessing method via the SCIKIT-LEARN tool [157] which removes the median of each variable and scales the data by the interquartile range.

7.3.5 Mass decorrelation

One of the most usual problems encountered during the development of a particle tagger is that its output can be strongly correlated to the most discriminating input features, mainly the particle mass. This results in a modified background mass distribution that imitates the signal distribution and it is called the *mass sculpting effect*. Many methods have been developed to decorrelate the output of the tagger from the assumed particle's mass distribution. In this analysis, we use the sample-reweighting technique. This technique removes the mass information from all the input features by reweighting the combinatorial background data sets so that the $m_{t^{res}}$ distribution matches that of the signal. Each background sample in the training dataset is assigned a different weight that determines its contribution to the calculation of the loss function. The weights are expressed in bins of the $m_{t^{res}}$ and are defined by the ratio of the $m_{t^{res}}$ distributions in signal over the background distribution. The weights as a function of the $m_{t^{res}}$ shown in Fig. 7.5 demonstrate that background samples with a mass close to the top-quark mass get large weights. The weighted loss function is defined as:

$$\mathcal{L} = \frac{\sum_{i=1}^{n} w_i \mathcal{L}_i}{\sum_{i=1}^{n} w_i} \tag{7.5}$$

where n is the size of the sub-dataset (batch) that is used to calculate the loss and w_i, $\mathcal{L}_i$ are the per-sample weights and loss respectively.

7.4 Performance

The performance of the t^{res} tagger is expressed as a receiver operating characteristic (ROC) curve that shows the background mistagging rate against the signal efficiency. The ROC curve is measured using $t\bar{t}$ simulated events. The signal efficiency and background mistagging rate are defined as the number of tagged events over the total number of events in the signal and background datasets respectively. Figure 7.6 (left) illustrates the ROC curves of the t^{res} tagger. Three WPs are established corresponding to 10%, 5% and 1% background misidentification probability and are marked in Fig. 7.6 (left) with a triangle, a cross and a star respectively.

The performance has been also studied in terms of the dependence on the $m_{t^{res}}$. The mass distribution of the t^{res} candidates that pass the three WPs and the inclusive distribution

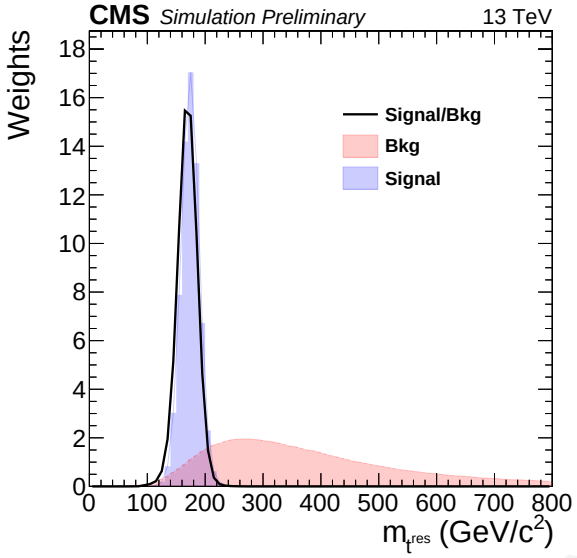


Figure 7.5: Weights assigned to each background sample in the training dataset to decorrelate the DNN output from the $m_{\mathrm{t}^{\mathrm{res}}}$. The weights are expressed in bins of the $m_{\mathrm{t}^{\mathrm{res}}}$.

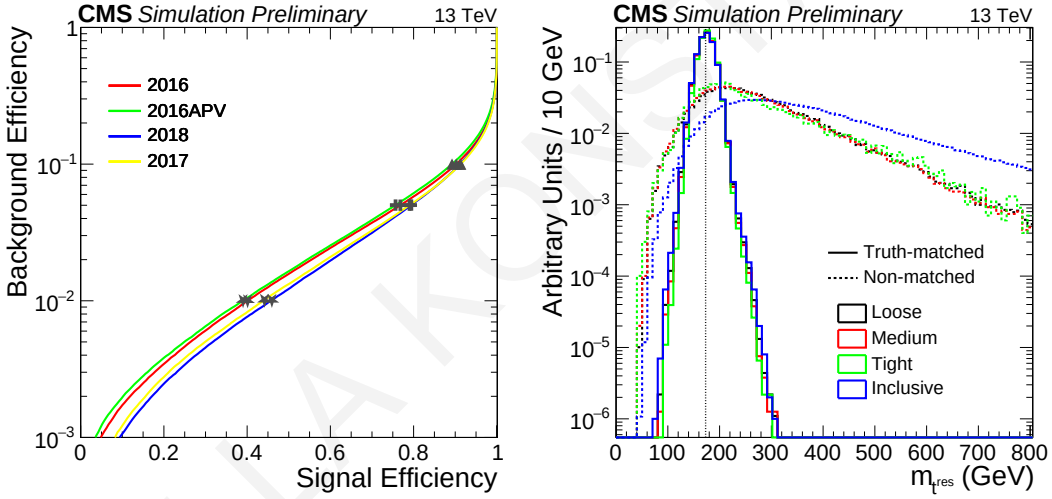


Figure 7.6: Left: ROC curve describes the performance of the $\mathrm{t}^{\mathrm{res}}$ tagger. Right: The mass distribution of the truth-matched and non-matched $\mathrm{t}^{\mathrm{res}}$ candidates that pass the loose, medium and tight WP of the $\mathrm{t}^{\mathrm{res}}$ tagger. The inclusive $m_{\mathrm{t}^{\mathrm{res}}}$ distribution is also shown.

are displayed in Fig. 7.6 (right). The results demonstrate the mass-decorrelated features of the tagger, which exhibits a smooth distribution for combinatorial $\mathrm{t}^{\mathrm{res}}$ and a peak-shaped distribution for real $\mathrm{t}^{\mathrm{res}}$.

7.5 Corrections to simulation

The misidentification rate and tagging efficiency of the $\mathrm{t}^{\mathrm{res}}$ tagger have been estimated and compared in data and simulation to extract data-to-simulation corrections, using a sample of lepton+jets, dominated by semileptonic $\mathrm{t\bar{t}}$ events. The events are collected with the `HLT_Mu50` trigger and they should contain exactly one reconstructed muon identified as tight, with $p_{\mathrm{T}} > 50$ GeV and $|\eta| < 2.4$. The events are also required to contain at least four

AK4 jets with $p_{\mathrm{T}} > 40$ GeV and $|\eta| < 2.4$, out of which at least one to be b tagged using the medium WP of the DeepJet b tagging algorithm. The b jet which is closest to the muon is considered as the b jet from the leptonic top-quark decay. A requirement of a ΔR distance less than 1.5 between the muon and the b jet from the leptonic top-quark is also applied. The trijet system with a mass closest to the top-quark mass is selected as the hadronic $\mathrm{t}^{\mathrm{res}}$ candidate. The distance between the muon and the hadronic $\mathrm{t}^{\mathrm{res}}$ candidate is $\Delta R > 2.0$. Based on the $E_{\mathrm{T}}^{\mathrm{miss}}$ and the muon mini-isolation, two orthogonal regions are defined. The low $E_{\mathrm{T}}^{\mathrm{miss}}$ (< 50 GeV) and high isolation (> 0.1) region is the background region (BR) and is expected to describe QCD multijet events and other non-top SM processes. The second region is characterized by high $E_{\mathrm{T}}^{\mathrm{miss}}$ (> 50 GeV) and tight mini-isolation (< 0.1) cuts. It is called signal region (SR) and is used to estimate the tagging and mistagging efficiencies.

To have simulated samples that describe well the data, QCD multijet and $\mathrm{t\bar{t}}$ samples are normalized in the BR and SR respectively in an iterative procedure. Starting from the BR, a QCD multijet normalization factor $\mathrm{f}^{\mathrm{QCD}}$ is calculated from the ratio of the number of QCD multijet events in data over simulation. The number of QCD multijet events in data is defined as:

$$\mathrm{N}_{\mathrm{BR}}^{\mathrm{QCD}} = \mathrm{N}_{\mathrm{BR}}^{\mathrm{data}} - \mathrm{N}_{\mathrm{BR}}^{\mathrm{t,t\bar{t}+X,EWK}} - \mathrm{f}^{\mathrm{t\bar{t}}} \times \mathrm{N}_{\mathrm{BR}}^{\mathrm{t\bar{t}}}. \tag{7.6}$$

The term $\mathrm{N}_{\mathrm{BR}}^{\mathrm{data}}$ indicates the number of data events in the BR. The term $\mathrm{N}_{\mathrm{BR}}^{\mathrm{t,t\bar{t}+X,EWK}}$ corresponds to simulated events from single top quark, $\mathrm{t\bar{t}}$ + X with X = (W, Z, γ, H and $\mathrm{t\bar{t}}$), and electroweak (EWK) contributions from V+jets, with V = (W or Z), diboson and triboson processes. The number of simulated $\mathrm{t\bar{t}}$ events in the BR $\mathrm{N}_{\mathrm{BR}}^{\mathrm{t\bar{t}}}$ is multiplied by the $\mathrm{t\bar{t}}$ normalization factor $\mathrm{f}^{\mathrm{t\bar{t}}}$ which takes an initial value equal to 1. Similarly, the $\mathrm{f}^{\mathrm{t\bar{t}}}$ is estimated in the SR taking into account the normalized QCD multijet contribution. The number of $\mathrm{t\bar{t}}$ events in data is defined as:

$$\mathrm{N}_{\mathrm{SR}}^{\mathrm{t\bar{t}}} = \mathrm{N}_{\mathrm{SR}}^{\mathrm{data}} - \mathrm{N}_{\mathrm{SR}}^{\mathrm{t,t\bar{t}+X,EWK}} - \mathrm{f}^{\mathrm{QCD}} \times \mathrm{N}_{\mathrm{SR}}^{\mathrm{QCD}} \tag{7.7}$$

This procedure is repeated using the updated values of $\mathrm{f}^{\mathrm{QCD}}$ and $\mathrm{f}^{\mathrm{t\bar{t}}}$ in the calculation of the $\mathrm{N}_{\mathrm{SR}}^{\mathrm{t\bar{t}}}$ and $\mathrm{N}_{\mathrm{BR}}^{\mathrm{QCD}}$ respectively. The iteration stops when $\mathrm{f}^{\mathrm{QCD}}$ and $\mathrm{f}^{\mathrm{t\bar{t}}}$ converge with a precision of $\mathcal{O}(10^{-5})$. For the calculation of the tagging and mistagging efficiencies, single t and $\mathrm{t\bar{t}}$ +X events are also corrected with the $\mathrm{f}^{\mathrm{t\bar{t}}}$ normalization factor.

The $\mathrm{t}^{\mathrm{res}}$ candidate's invariant mass for the inclusive sample in the SR before and after applying the medium WP of the $\mathrm{t}^{\mathrm{res}}$ algorithm is shown in Fig. 7.7 for the 2018 data. In both cases, in the vicinity of the top-quark mass, the selected events originate from truth-matched $\mathrm{t}^{\mathrm{res}}$. Beyond that region, contributions from combinatorial $\mathrm{t}^{\mathrm{res}}$ dominate. The transverse momentum of the $\mathrm{t}^{\mathrm{res}}$ candidates before and after applying the medium WP of the $\mathrm{t}^{\mathrm{res}}$ algorithm is shown in Fig. 7.8.

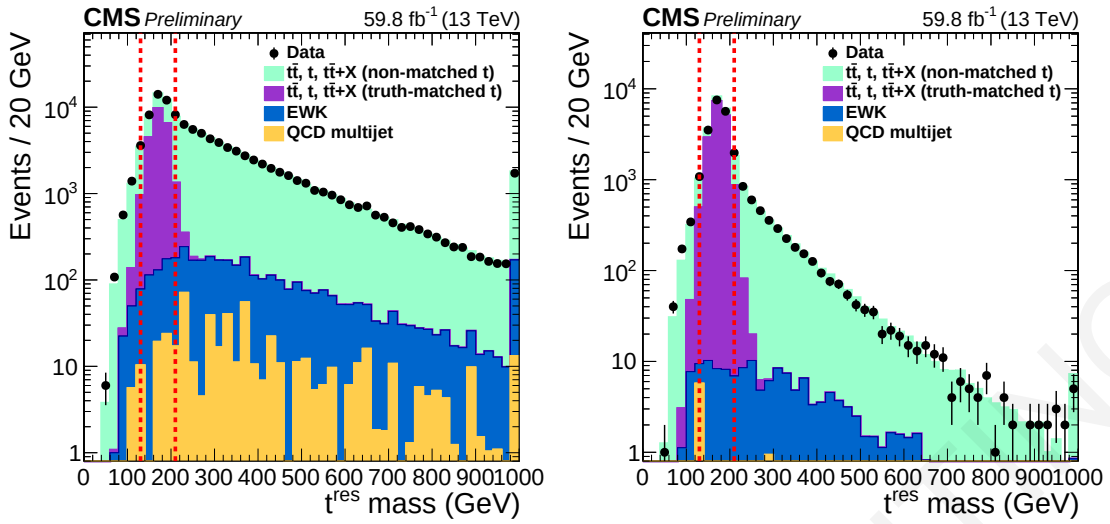


Figure 7.7: Invariant mass of the t^{res} candidates in the SR before (left) and after (right) the requirement of the medium WP of the t^{res} tagger, using 2018 data.

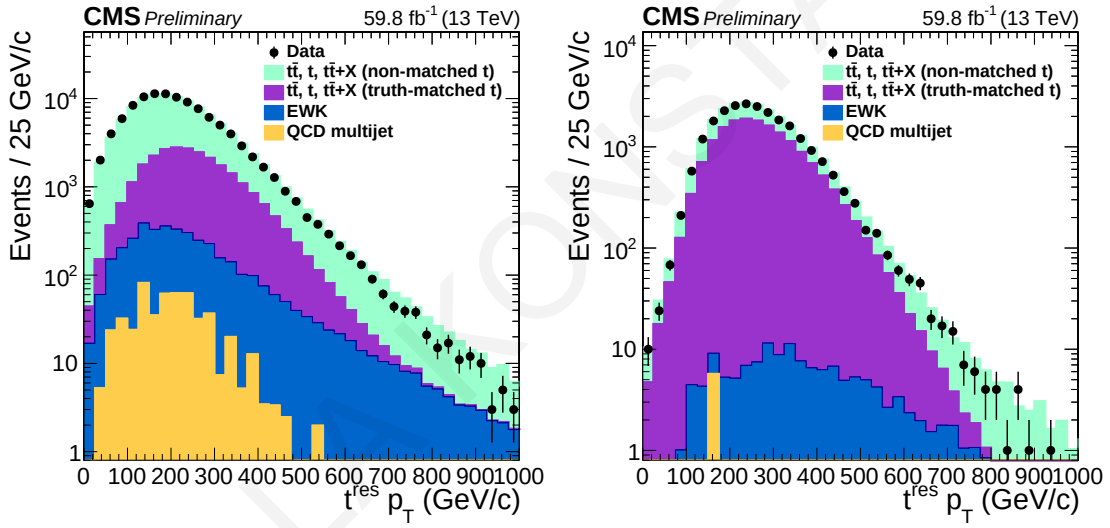


Figure 7.8: Transverse momentum of the t^{res} candidates in the SR before (left) and after (right) the requirement of the medium WP of the t^{res} tagger, using 2018 data.

7.5.1 Misidentification Rate

The misidentification rate is measured in events in the SR with the requirement the t^{res} candidate to have a mass that falls outside a mass window of 130-210 GeV, which corresponds to a 3σ window on the mass of the truth-matched t^{res} candidates as illustrated in Fig. 7.7 by the two vertical lines. This anti-selection ensures that the selected t^{res} candidate originates from random jet combinatorics which constitute the main source of mistagged t^{res}. The misidentification rate is given by the number of events where the t^{res} candidate passes a given WP of the t^{res} tagger over the total number of events and is expressed in bins of the t^{res} candidate's p_T. The misidentification rate is shown in Figs. 7.9 to 7.12 for the three WP and all data-taking eras. The scale factors (SFs) are derived from the data-over-simulation ratio for each WP and data-taking era.

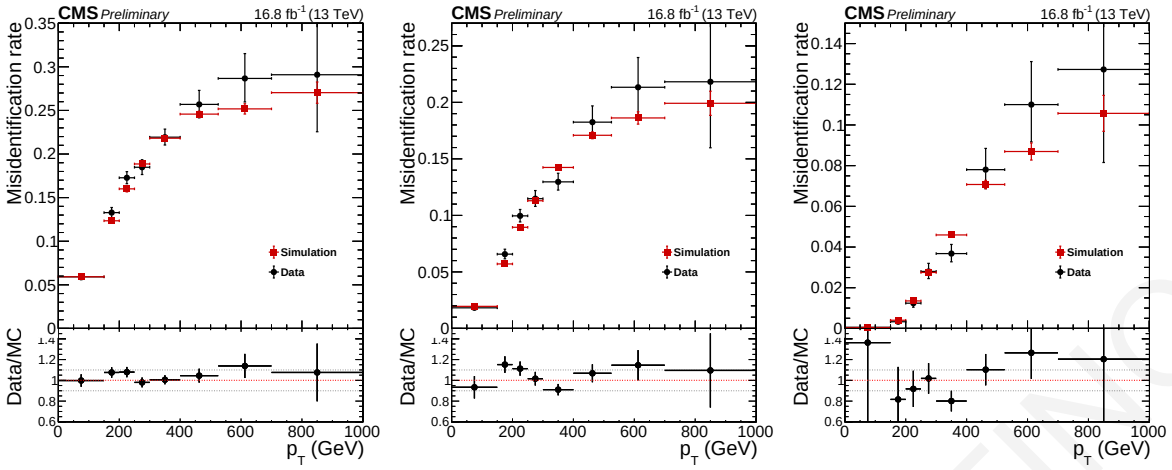


Figure 7.9: Misidentification rate in data and MC simulation, for the loose (left), medium (middle) and tight (right) WP for 2016 non-APV data.

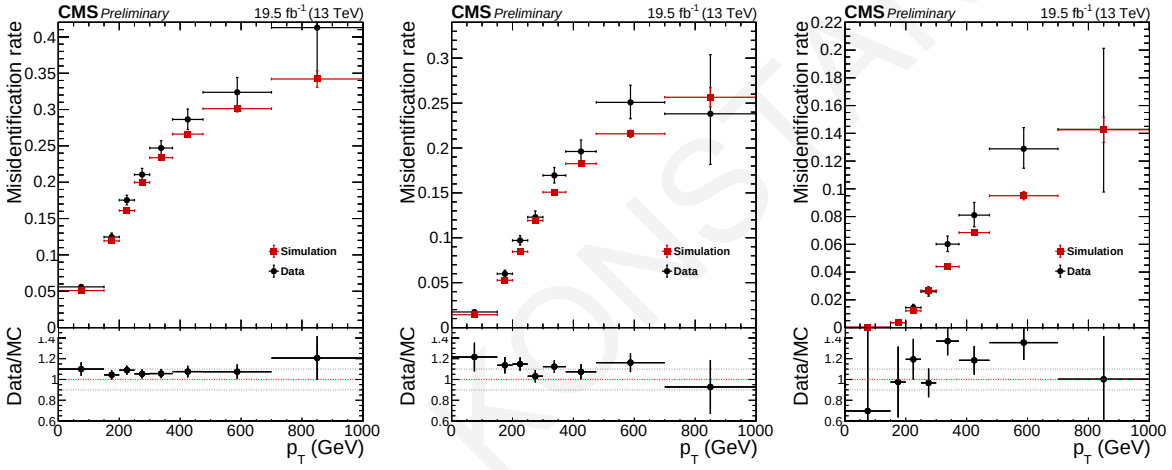


Figure 7.10: Misidentification rate in data and MC simulation, for the loose (left), medium (middle) and tight (right) WP for 2016 APV data.

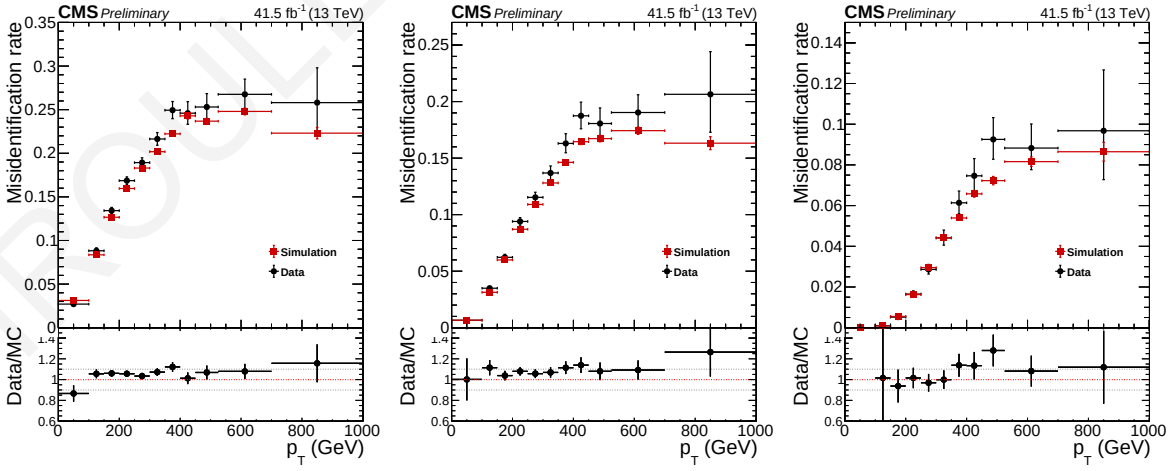


Figure 7.11: Misidentification rate in data and MC simulation, for the loose (left), medium (middle) and tight (right) WP for 2017 data.

7.5.2 Top-quark tagging efficiency

The tagging efficiency is measured in events from the SR requiring the t^{res} candidate to have a mass within 130-210 GeV. To minimize contributions of misidentified t^{res} candidates due to

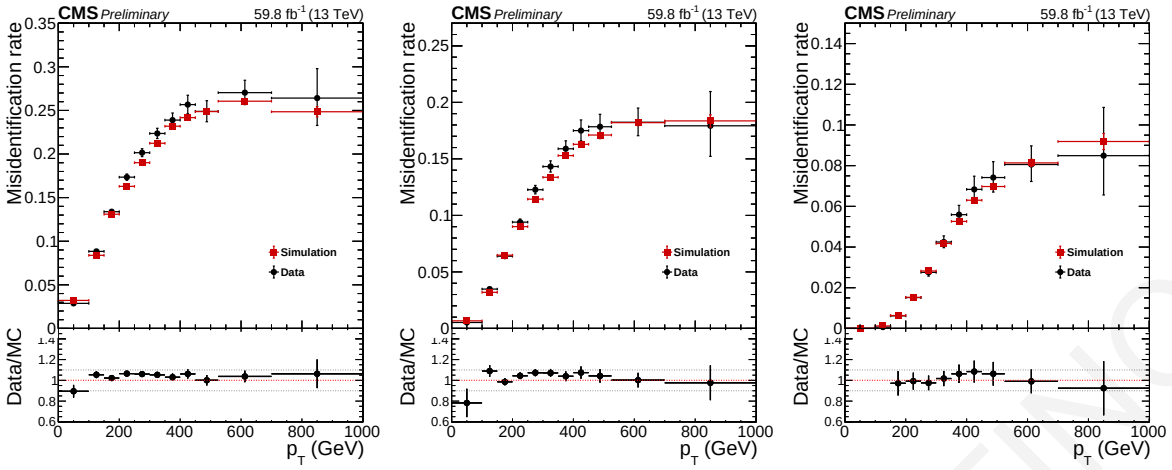


Figure 7.12: Misidentification rate in data and MC simulation, for the loose (left), medium (middle) and tight (right) WP for 2018 data.

combinatorial background, events with more than one t^{res} candidate inside the mass window are rejected. The t^{res}-tagging efficiency is estimated by subtracting from data, contributions from non-top-quark processes and combinatorial background from top-quark processes with misidentified t^{res} candidates. A t^{res} candidate is considered as misidentified if it fails the matching criteria described in Section 7.3.1. For this study the ΔR parton-jet matching is loosen to $\Delta R < 0.5$ for the entire p_{T} range. The mistagging rate SFs are also taken into account to correct the subtracted events.

The t^{res}-tagging efficiency as a function of the t^{res} candidate's p_{T} is shown in Figs. 7.13 to 7.16 for the three WP and all data-taking eras. Overall good agreement between data and simulation is observed while the differences observed in bins with a limited number of events are within the statistical uncertainties. The data-to-simulation t^{res} tagging efficiency SFs are derived for each WP and data-taking period.

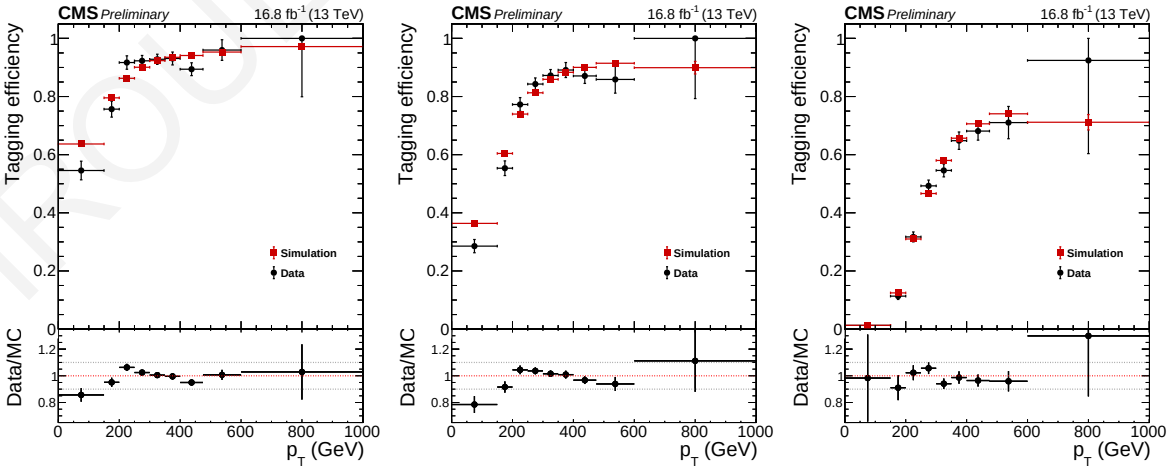


Figure 7.13: Tagging efficiency in data and MC, for the loose (left), medium (middle) and tight (right) WP for 2016 non-APV data.

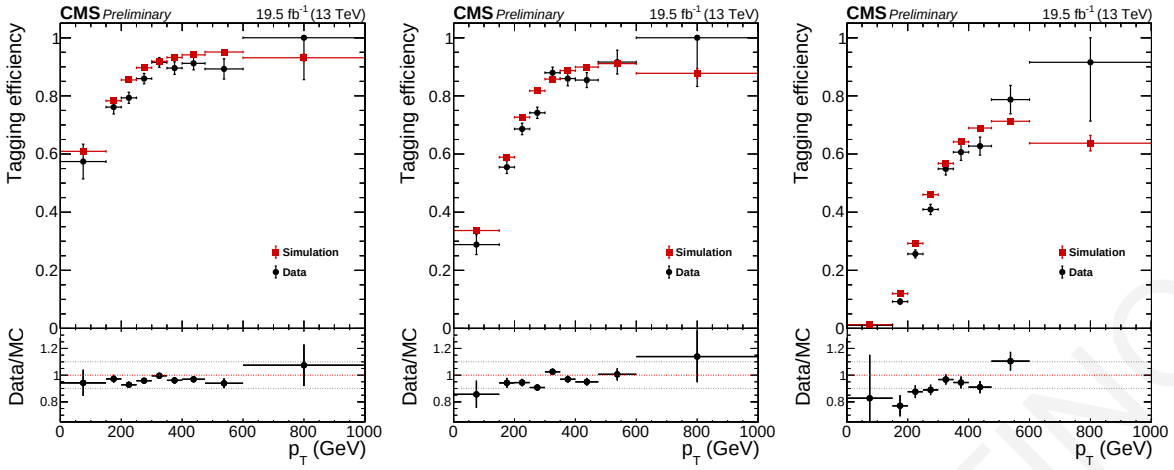


Figure 7.14: Tagging efficiency in data and MC, for the loose (left), medium (middle) and tight (right) WP for 2016 APV data.

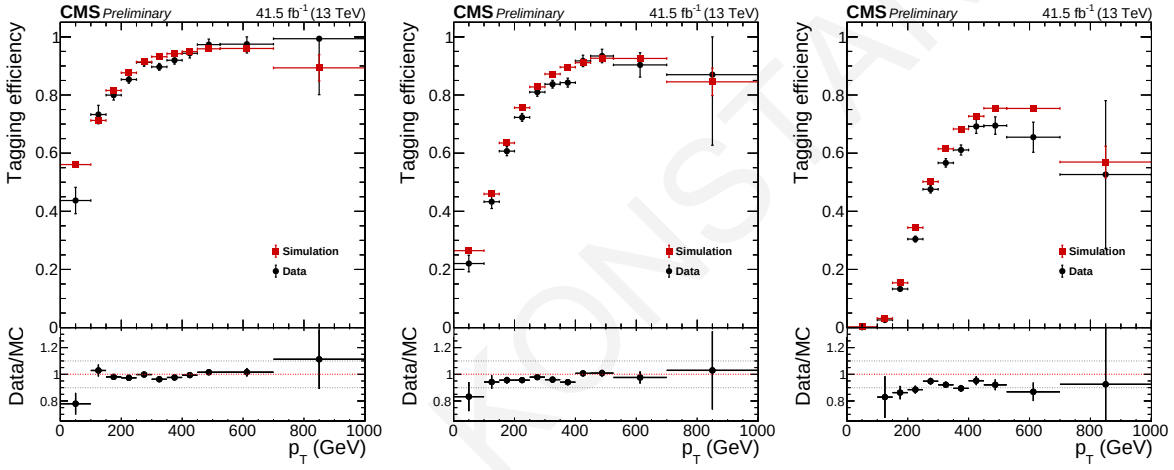


Figure 7.15: Tagging efficiency in data and MC, for the loose (left), medium (middle) and tight (right) WP for 2017 data.

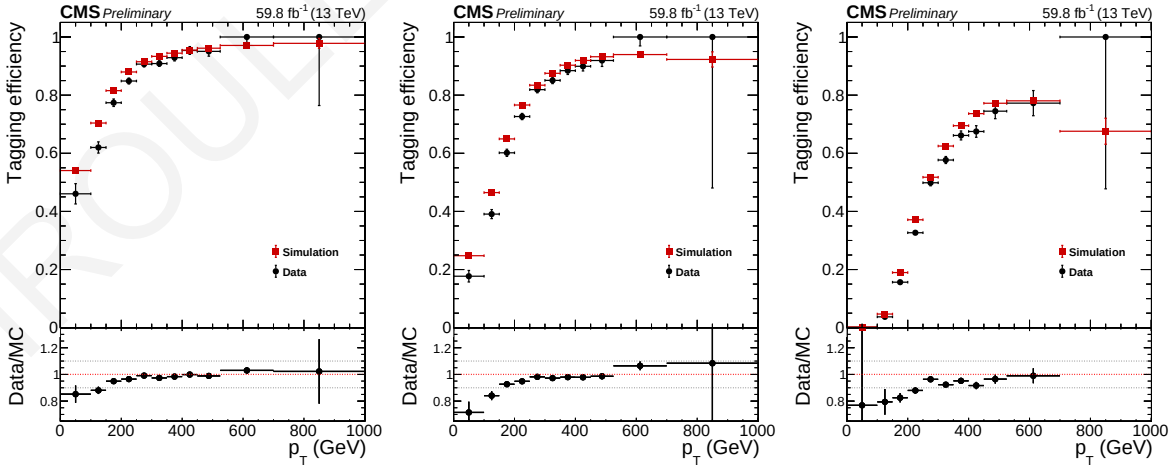


Figure 7.16: Tagging efficiency in data and MC, for the loose (left), medium (middle) and tight (right) WP for 2018 data.

7.6 Systematic uncertainties

Various sources of systematic uncertainties that affect the t^{res} tagging efficiency and mistagging rate have been studied.

The mistagging rate SF are affected by the JES [158] and JER [159]. To quantify the effect, variations of $\pm\ 1\sigma$ on JES and JER are applied. The results are presented in Fig. 7.17 for the medium WP and all data-taking eras. Differences in the mistagging rate SFs are less than 10%. The systematic uncertainties considered are calculated independently. The final up and down $\sigma^{\pm}_{SF\ mistag}$ variations are calculated separately, resulting in asymmetric uncertainties and their magnitude is the sum of all variations from each direction in quadrature:

$$\sigma^{+}_{SF\ mistag} = \sqrt{\sum_{j(\epsilon_j>\epsilon)}^{all\ effects} \left(1 - \frac{\epsilon_j}{\epsilon}\right)^2} \tag{7.8}$$

$$\sigma^{-}_{SF\ mistag} = \sqrt{\sum_{j(\epsilon_j<\epsilon)}^{all\ effects} \left(1 - \frac{\epsilon_j}{\epsilon}\right)^2} \tag{7.9}$$

with ϵ_j being the mistagging rate with the effect j and ϵ the nominal efficiency. The mistagging rate SFs are illustrated in Fig. 7.18 for the medium WP with the statistical uncertainties shown with the error bars and systematic uncertainties $\sigma^{\pm}_{SF\ mistag}$ shown with the gray bands.

The t^{res} tagging efficiency can be also affected by the JES and JER. The systematic uncertainties are calculated by applying up and down variations and also propagating their impact on the misidentification rate SFs shown in Fig. 7.17. The t^{res} tagging SF systematic uncertainties from JES and JER variations are shown in Fig. 7.19. Their impact on the tagging SFs does not exceed 10%, excluding the bins with large statistical uncertainties.

The dependence of the t^{res} tagging SFs from the combinatorial background that is subtracted from data to measure the t^{res} tagging efficiency is also studied by altering the definition of the misidentified t^{res} candidates. The t^{res} tagging efficiency has been measured using the events from top-quark processes where the t^{res} candidate falls inside the mass window without asking to pass the matching criteria. Therefore, the subtracted background contains non-top processes but not events with top-quarks with non-matched t^{res} candidates. The impact of the combinatorial background subtraction to the t^{res} tagging SFs is negligible as displayed in Fig. 7.20.

Finally, the systematic uncertainties caused by the top-quark modeling have been calculated, using $t\bar{t}$ simulation samples with different configurations to describe different effects that the top-quark modeling is sensitive to. The systematic effects studied are the following:

- the first shower emission, modeled by the hdamp parameter,
- the scale radiation by changing the early resonance decays (ERD) parameter,
- the color reconnection strength (CR1, CR2),
- the top-quark mass and
- the underlying events by changing the TuneCP5 parameters.

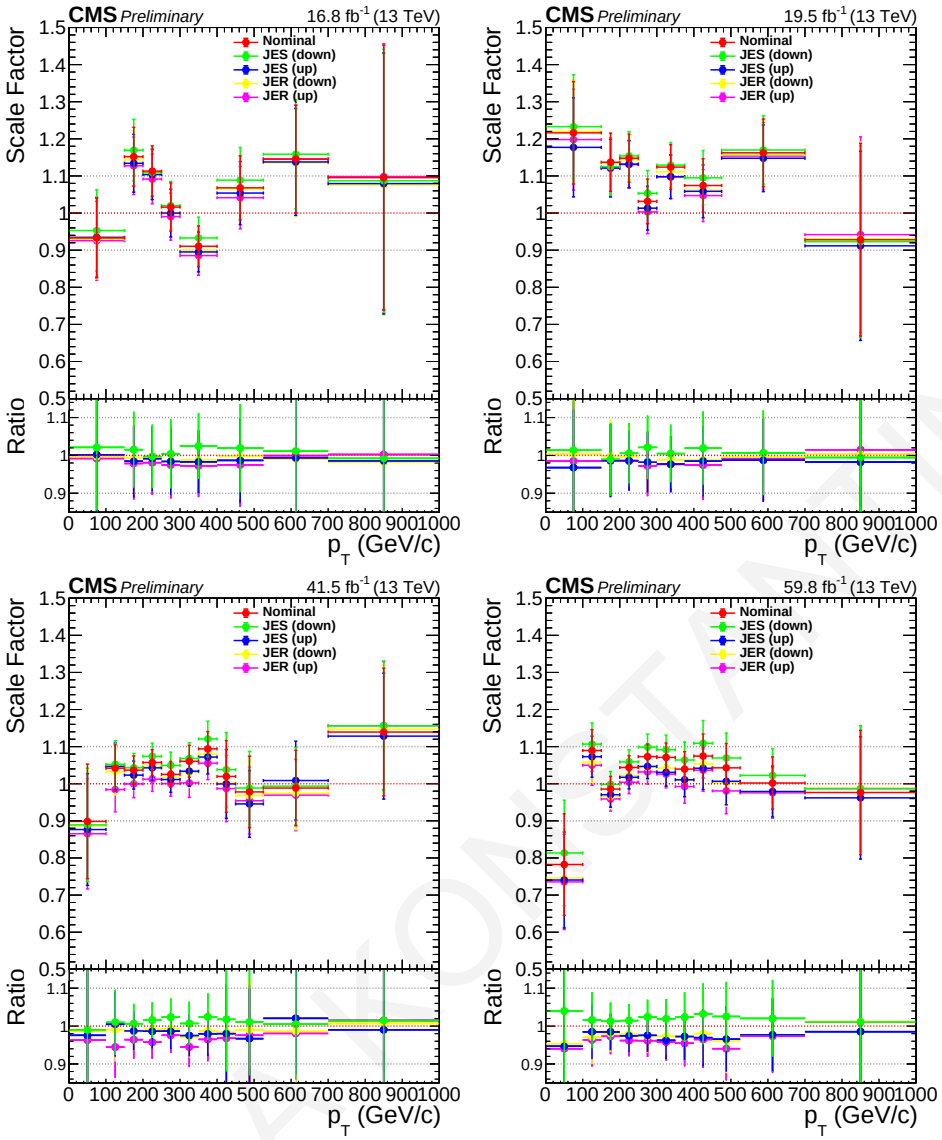


Figure 7.17: Mistagging rate SF systematic uncertainties from JES and JER, for the medium WP and all data-taking eras.

The impact of each of the aforementioned effects is shown in Fig. 7.21 for the medium WP.

The systematic uncertainties of the t^{res} tagging SFs originating from the JES and JER variations, the combinatorial background subtraction and the top-quark modeling are accounted for independently. The total systematic uncertainty is quantified by Eqs. (7.8) and (7.9) and displayed in Fig. 7.22 for the medium WP.

7.7 Summary

A DNN-based resolved top-quark tagger is developed to distinguish between trijet systems originating from hadronic top-quark decays of moderate p_{T} from combinatorial trijet systems. To eliminate the mass sculpting effect, the DNN algorithm has been decorrelated from the top-quark mass. The tagger exhibits good discrimination power against misidentified t^{res} candidates. The tagging efficiency and misidentification rate have been evaluated and

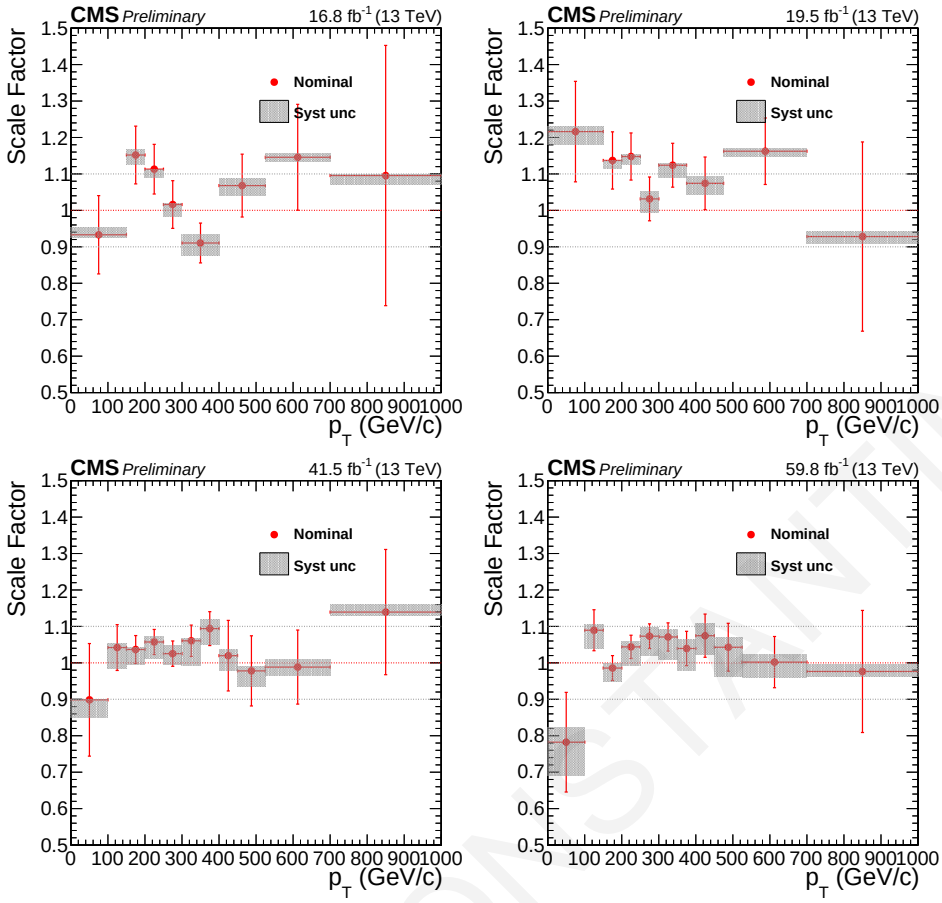


Figure 7.18: Total mistagging rate SF systematic uncertainties for the medium WP and all data-taking eras.

compared in data and simulation to extract SFs in bins of the t^{res} candidate's p_T. Good agreement is observed between data and simulation and any observed differences are within the statistical uncertainties.

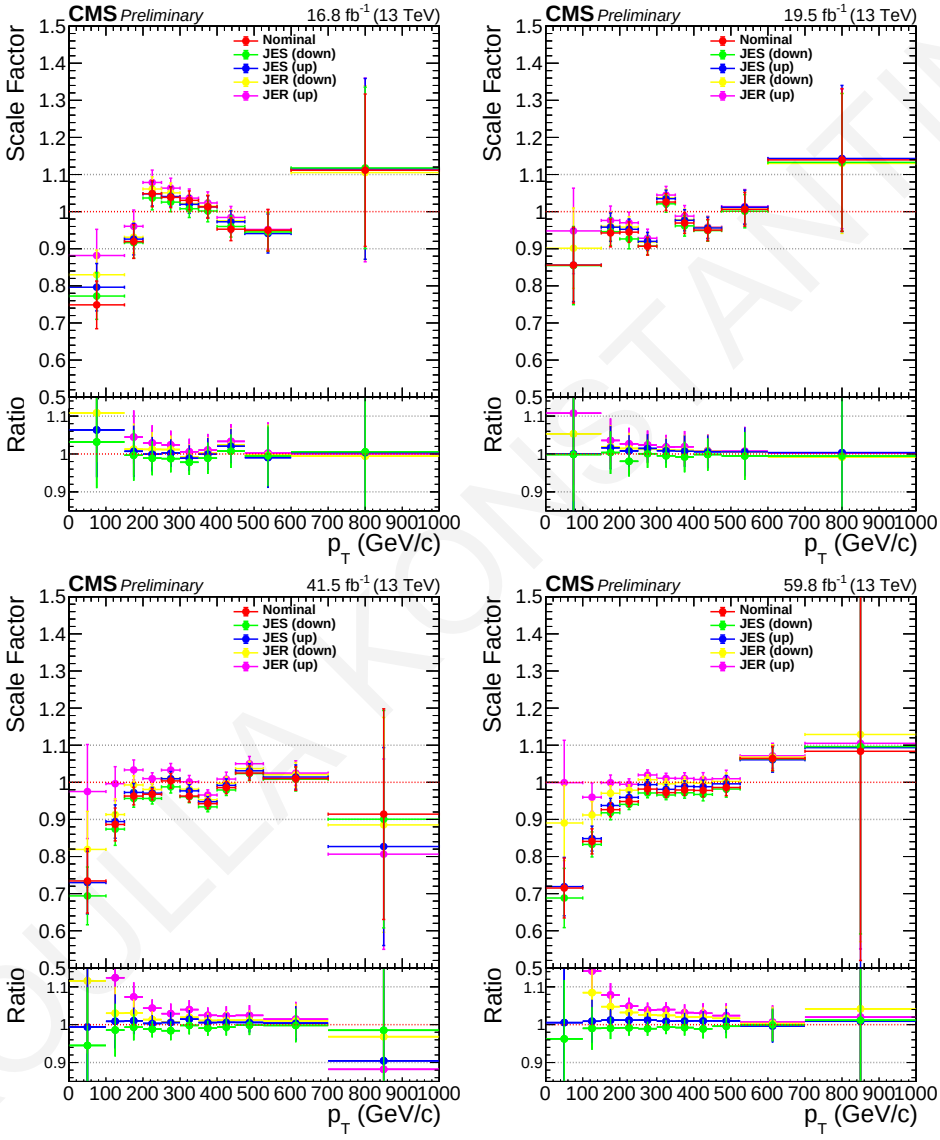


Figure 7.19: Top-quark tagging SF systematic uncertainties from JES and JER, for the medium WP and all data-taking eras.

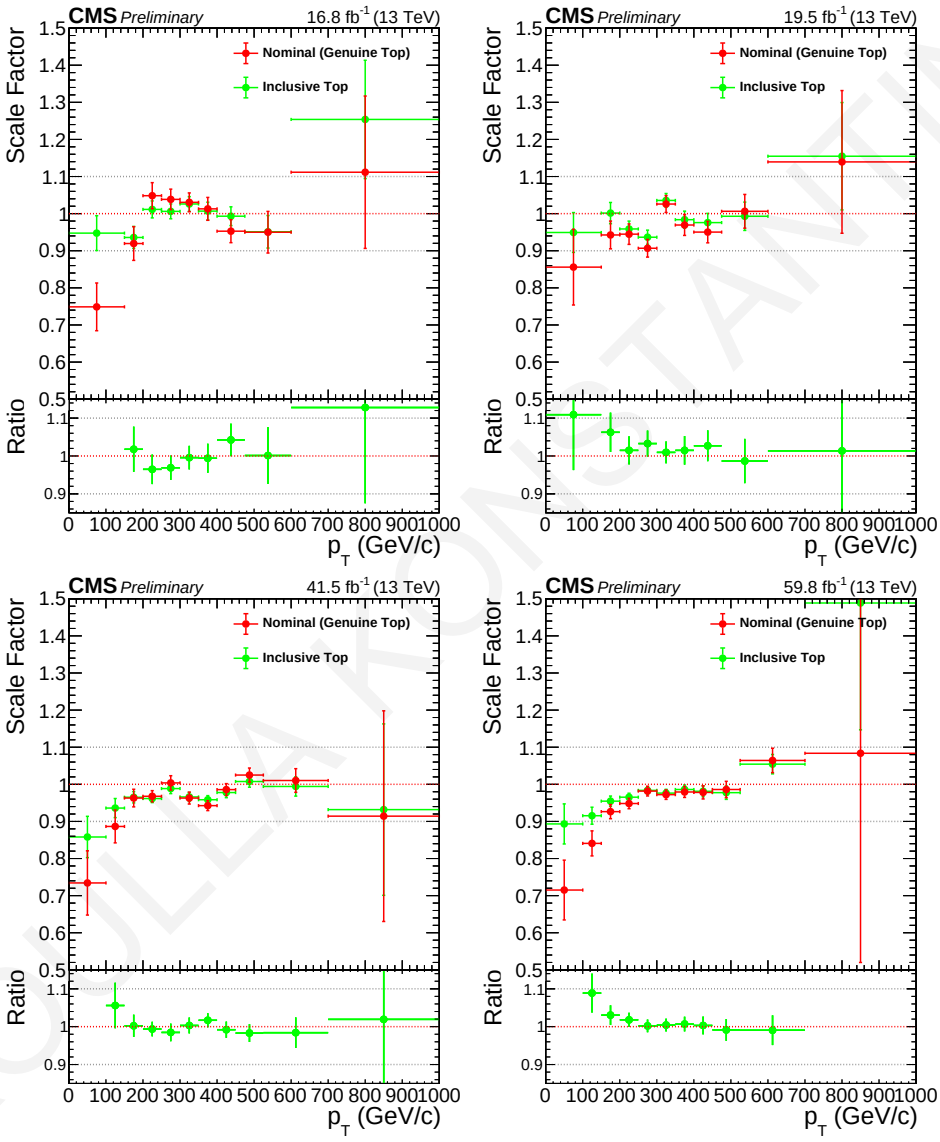


Figure 7.20: Top-quark tagging SF systematic uncertainties from the combinatorial background subtraction, for the medium WP and all data-taking eras.

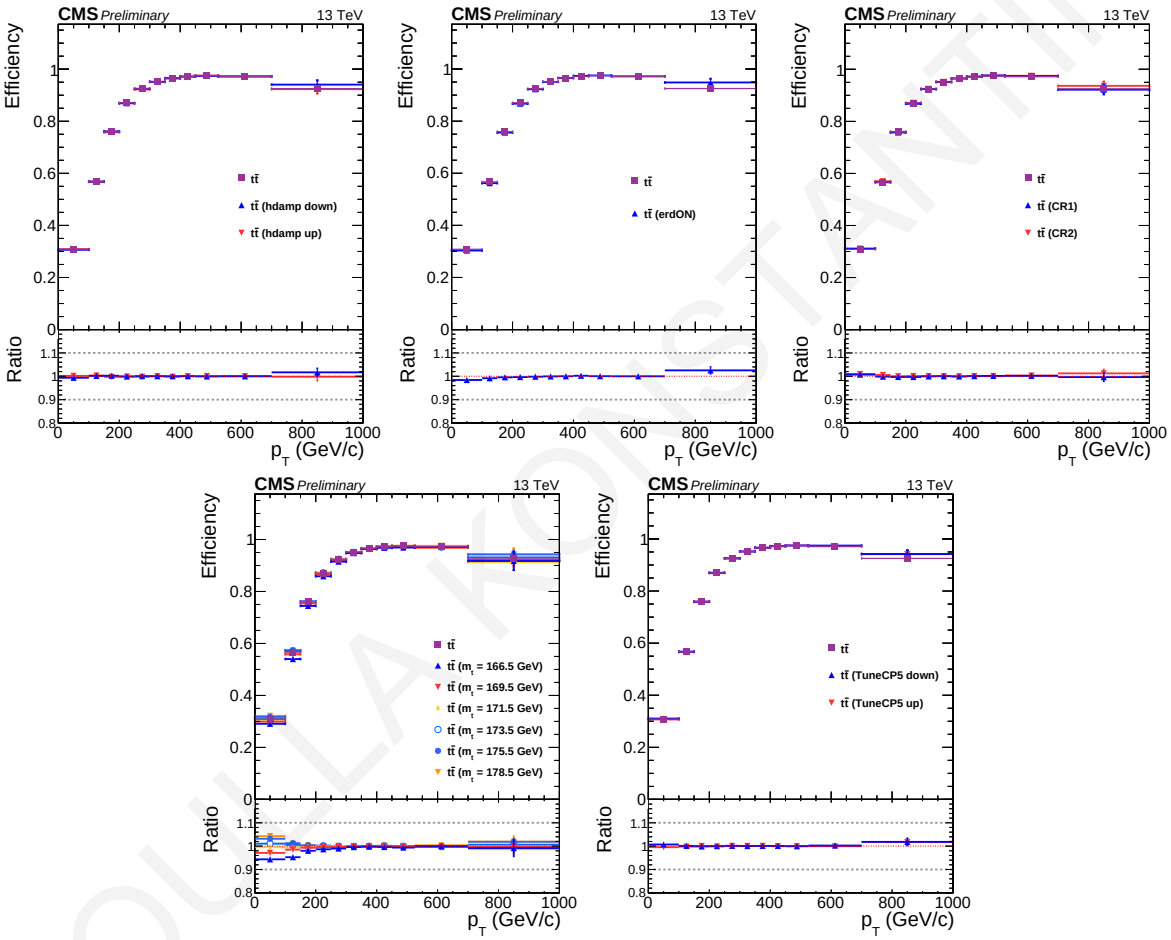


Figure 7.21: Tagging efficiency as determined in the nominal $t\bar{t}$ MC sample and samples with different configurations for the medium WP for 2017 data. The ratio plots express the systematic uncertainties of the top-quark tagging SFs.

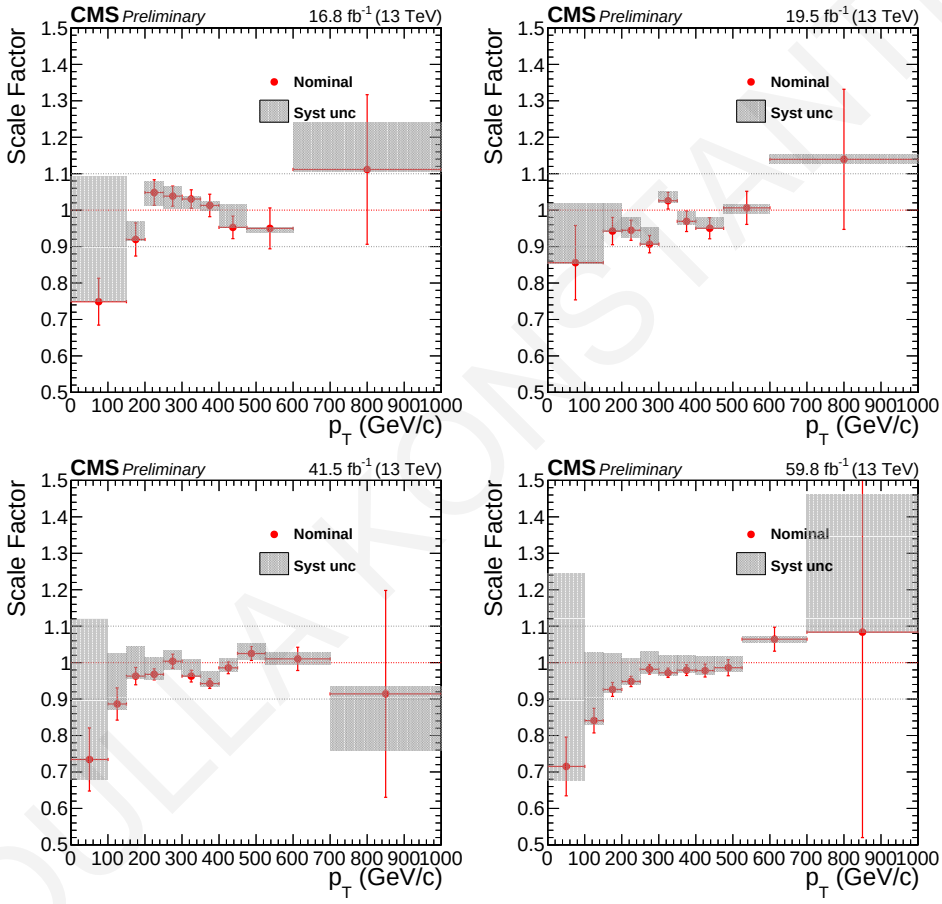


Figure 7.22: Total top-quark tagging SF systematic uncertainties for the medium WP and all data-taking eras.

8 Search for $H^{\pm} \rightarrow tb$ in fully hadronic final state

A search for a heavy charged Higgs boson produced in association with a top and a bottom quark, decaying to top and bottom quark ($H^{\pm} \rightarrow tb$) is presented, using pp collision data recorded by the CMS detector at $\sqrt{s}$ = 13 TeV during the Run II data-taking period with a total luminosity of 138 fb^{-1}. The final state signature of the $H^{\pm} \rightarrow tb$ production channel is determined by the decay of the W bosons since the top quark decays with nearly 100% probability to a W boson and a bottom quark. A fully hadronic final state has $\mathcal{B} \sim 45\%$ and in the associated production mechanism contains eight quarks, of which four are b quarks. Large H_{T} and low $\vec{p}_T^{\,\mathrm{miss}}$ due to the absence of neutrinos are some of the characteristics of the final state. The tree-level Feynman diagram of the associated production of $H^{\pm} \rightarrow tb$ in the fully hadronic final state is shown in Fig. 8.1.

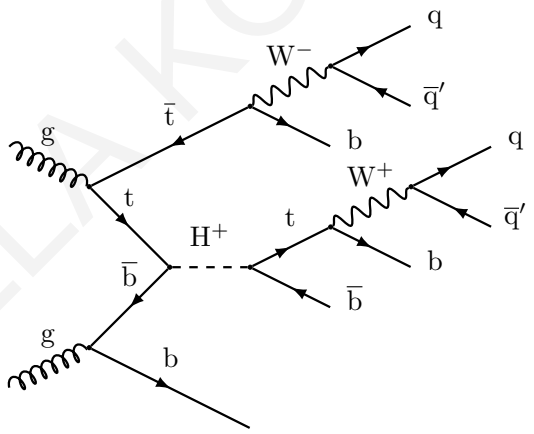


Figure 8.1: Direct production of $H^{\pm}$ in association with a top and a bottom quark in the fully hadronic final state.

Despite the large branching fraction of the hadronic decay, the presence of large multijet background produced via ordinary QCD multiparton processes (QCD multijet) and $t\bar{t}$ background make the search highly challenging. Other minor backgrounds include the production of a single top quark, $t\bar{t}$ + X with X = (W, Z, γ, H and $t\bar{t}$), and electroweak (EWK) contributions from V+jets, with V = (W or Z), diboson and triboson processes. Figure 8.2 shows the Feynman diagrams of the dominant QCD multijet and $t\bar{t}$ background processes in the all-jet final state.

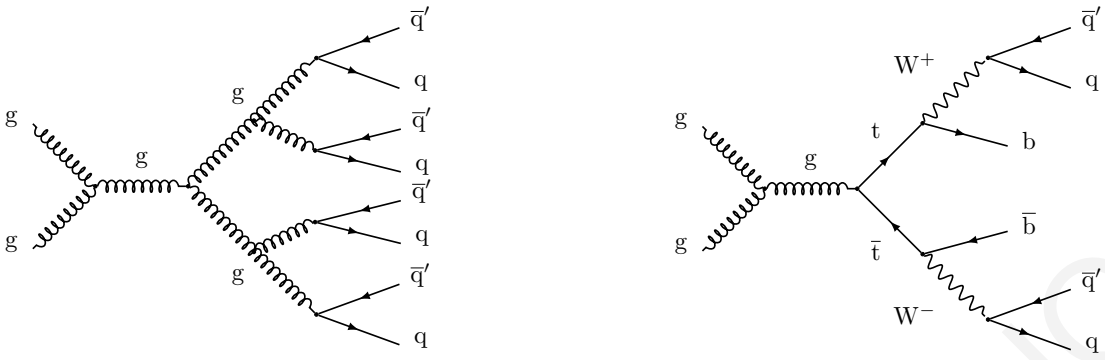
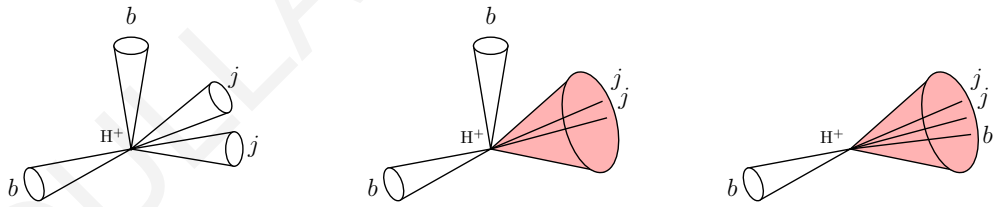


Figure 8.2: Feynman diagrams of QCD multijet (left) and $t\bar{t}$ pair (right) in the fully hadronic final state.

8.1 Resolved and Boosted topology

The topological and kinematic characteristics of an event are strongly dependent on the mass of the charged Higgs boson. When the $m_{\mathrm{H}^{\pm}}$ is large, the decay products of the $\mathrm{H}^{\pm}$ acquire large transverse momentum and become collimated starting to coalesce in large-radius boosted jets through hadronic decays. In such events, jets from the quarks of a W boson or the top-quark decay may merge to a signle W jet or a top jet respectively. The diagrams in Fig. 8.3 show the decay of the $\mathrm{H}^{\pm}$ into resolved and boosted jets. This analysis focuses on the resolved regime, where all jets are resolved into separate jets, and the semi-boosted regime, which features one resolved and one merged top quark.



(a) resolved top quark (b) boosted W boson (c) boosted top quark

Figure 8.3: Resolved and boosted topologies.

9 Search for charged Higgs bosons in the resolved regime

This section presents a search for a $\mathrm{H}^{\pm} \rightarrow \mathrm{tb}$ in the all-jet final state. The trijet systems originating from the top-quark decays are reconstructed with the resolved top-quark tagger described in Chapter 7. The dominant QCD multijet background is estimated with a data-driven method and is discussed in Section 9.4. To extract signal in the presence of SM background an event-based DNN is developed which is parameterized with respect to the $\mathrm{H}^{\pm}$ mass hypothesis. Model-independent upper limits at 95% CL are set on the product of the production cross section and the branching ratio into a top and a bottom quark-antiquark pair $\sigma(\mathrm{pp} \rightarrow \mathrm{tbH}^{\pm}) \times \mathcal{B}(\mathrm{H}^{\pm} \rightarrow \mathrm{tb})$. The results are summarized in Section 10.6.

9.1 Event selection

A set of selection criteria is applied to data and simulation to select signal-like events and reject as many background events as possible. These criteria are described in Sections 9.1.1 to 9.1.8 and summarized in Section 9.1.9.

9.1.1 Signal trigger

Events used in the analysis are required to pass specific HLT triggers. Fully hadronic multijet triggers were designed to select events with large jet multiplicity with a threshold on the H_{T} and at least one or two b-tagged jets reconstructed at the HLT with the CSV [160] or DeepCSV [135] algorithms. Additional triggers are used to select events with the presence of AK8 jets with requirements on the p_{T} and the jet mass, and high-H_{T} events. The full list of triggers is briefly described in Table 9.1. Studies on the efficiency of the signal triggers are presented in Section 9.2.

9.1.2 Primary vertex

The selected PVs are reconstructed as described in Section 4.1.1 and must fulfill a set of selection criteria. The number of d.o.f indicating the number of associated tracks must be $\mathrm{N}_{\mathrm{dof}} \geq 4$. The position of the PV along the beam axis must not exceed 24 cm and the radial position in the transverse plane is less than 2 cm. In events with multiple reconstructed

Table 9.1: Signal HLT triggers and their requirements.

2016	
HLT_PFHT400_SixJet30_DoubleBTagCSV_p056	$H_T > 400$ GeV, ≥ 6 AK4 jets with $p_T > 30$ GeV, 2 b tagged jets (CSV)
HLT_PFHT450_SixJet40_BTagCSV_p056	$H_T > 450$ GeV, ≥ 6 AK4 jets with $p_T > 40$ GeV, 1 b tagged jets (CSV)
HLT_PFHT900	$H_T > 900$ GeV
HLT_AK8PFJet450	≥ 1 AK8 jet with $p_T > 450$ GeV
HLT_AK8PFJet360_TrimMass30	≥ 1 AK8 jet with $p_T > 360$ GeV, M > 30 GeV
2017	
HLT_PFHT380_SixPFJet32_DoublePFBTagCSV_2p2	$H_T > 380$ GeV, ≥ 6 AK4 jets with $p_T > 32$ GeV, 2 b tagged jets (CSV)
HLT_PFHT380_SixPFJet32_DoublePFBTagDeepCSV_2p2	$H_T > 380$ GeV, ≥ 6 AK4 jets with $p_T > 32$ GeV, 2 b tagged jets (DeepCSV)
HLT_PFHT430_SixPFJet40_PFBTagCSV_1p5	$H_T > 430$ GeV, ≥ 6 AK4 jets with $p_T > 40$ GeV, 1 b tagged jets (CSV)
HLT_PFHT1050	$H_T > 1050$ GeV
HLT_AK8PFJet500	≥ 1 AK8 jet with $p_T > 500$ GeV
HLT_AK8PFJet400_TrimMass30	≥ 1 AK8 jet with $p_T > 400$ GeV, M > 30 GeV
HLT_AK8PFHT800_TrimMass50	$H_T > 800$ GeV, M > 50 GeV
2018	
HLT_PFHT380_SixPFJet32_DoublePFBTagCSV_2p2	$H_T > 380$ GeV, ≥ 6 AK4 jets with $p_T > 32$ GeV, 2 b tagged jets (CSV)
HLT_PFHT380_SixPFJet32_DoublePFBTagDeepCSV_2p2	$H_T > 380$ GeV, ≥ 6 AK4 jets with $p_T > 32$ GeV, 2 b tagged jets (DeepCSV)
HLT_PFHT400_SixPFJet32_DoublePFBTagDeepCSV_2p94	$H_T > 400$ GeV, ≥ 6 AK4 jets with $p_T > 32$ GeV, 2 b tagged jets (DeepCSV)
HLT_PFHT430_SixPFJet40_PFBTagCSV_1p5	$H_T > 430$ GeV, ≥ 6 AK4 jets with $p_T > 40$ GeV, 1 b tagged jets (CSV)
HLT_PFHT430_SixPFJet40_PFBTagDeepCSV_1p5	$H_T > 430$ GeV, ≥ 6 AK4 jets with $p_T > 40$ GeV, 1 b tagged jets (DeepCSV)
HLT_PFHT450_SixPFJet36_PFBTagDeepCSV_1p59	$H_T > 450$ GeV, ≥ 6 AK4 jets with $p_T > 36$ GeV, 1 b tagged jets (DeepCSV)
HLT_PFHT1050	$H_T > 1050$ GeV
HLT_AK8PFJet500	≥ 1 AK8 jet with $p_T > 500$ GeV
HLT_AK8PFJet400_TrimMass30	≥ 1 AK8 jet with $p_T > 400$ GeV, M > 30 GeV
HLT_AK8PFHT800_TrimMass50	$H_T > 800$ GeV, M > 50 GeV

PVs, the one with the largest value of the scalar quadratic p_T sum of the associated tracks is considered the PV of the hard scattering.

9.1.3 E_T^{miss} filters

A set of clean-up E_T^{miss} filters is applied to data and simulated events to reject events with artificially high-p_T^{miss} due to reconstruction failures or malfunctioning detectors. The list of filters used in this analysis is shown in Table 9.2. Details about each filter can be found in [161].

9.1.4 Electrons, muons, and hadronic τ veto

Events with isolated electrons, muons, and hadronically decaying τ's are excluded to ensure a fully hadronic final state selection.

The reconstruction and identification of isolated muons is described in Section 4.2.2. Global muons are required to pass the loose criteria for the identification of prompt muons and muons from heavy and light quark decays, as described in [162]. The mini-isolation algorithm is used to quantify the amount of measured energy in a cone relative to the transverse momentum of the lepton. The relative isolation is given by:

Table 9.2: Summary of the $E_{\mathrm{T}}^{\mathrm{miss}}$ filters applied to data and simulation.

good primary vertex filter
beam halo filter
HBHE noise filter
HBHE iso noise filter
ECAL dead cell trigger primitive fiter
Bad PF Muon Filter
Bad PF Muon Dz Filter
Bad EE supercluster filter
ECAL bad calibration filter (2017 and 2018)

$$\mathrm{Iso} = \sum p_{\mathrm{T}}^{\text{ch from PV}} + \max(0, \sum \mathrm{E}_{\mathrm{T}}^{\text{nh}} + \sum \mathrm{E}_{\mathrm{T}}^{\gamma} - \frac{1}{2} \sum \mathrm{p}_{\mathrm{T}}^{\mathrm{PU}}) / \mathrm{p}_{\mathrm{T}}(\ell) \tag{9.1}$$

where $p_{\mathrm{T}}^{\text{ch from PV}}$ is the p_{T} of charged hadrons from the PV, $\mathrm{E}_{\mathrm{T}}^{\text{nh}}$ and $\mathrm{E}_{\mathrm{T}}^{\gamma}$ is the transverse energy of neutral hadrons and photons around the muon and $p_{\mathrm{T}}^{\mathrm{PU}}$ is the p_{T} from pileup. The term $p_{\mathrm{T}}(\ell)$ in the denominator refers to the p_{T} of the lepton. The mini isolation uses a p_{T}-dependent size of the cone radius so that in boosted topologies, the smallest possible radius is selected to reduce overlap with jets. The radius is defined as:

$$R = \max(0.05, \min(0.2, 10/p_{\mathrm{T}}(\ell))) \tag{9.2}$$

and results in the following radius values:

$$R = \begin{cases} 0.2, & p_{\mathrm{T}}(\ell) \leq 50 \text{ GeV} \\ \frac{10 \text{ GeV}}{p_T(\ell)}, & 50 \text{ GeV} < p_{\mathrm{T}}(\ell) \leq 200 \text{ GeV} \\ 0.05 & p_{\mathrm{T}}(\ell) \geq 200 \text{ GeV} \end{cases} \tag{9.3}$$

Muons identified as loose with relative mini isolation less than 0.4, $p_{\mathrm{T}} > 10$ GeV and $|\eta| < 2.4$ are excluded from further analysis.

Electrons reconstructed as discussed in Section 4.2.3 are required to pass a cut-based identification algorithm provided by the E/γ physics object group with a working point that corresponds to 95% signal efficiency. Similarly to the muon veto, electrons with relative mini isolation less than 0.4, $p_{\mathrm{T}} > 10$ GeV and $|\eta| < 2.4$ are rejected.

Events with τ_h reconstructed following the method in Section 4.5 are vetoed if the selected τ_h jets have $p_{\mathrm{T}} > 20$ GeV and $|\eta| < 2.1$. The selected τ_h jets must also fulfill the iden-

tification criteria of the DeepTau algorithm against the hypotheses of misidentified muons, electrons or jets from quarks or gluons.

9.1.5 Hadronic jet selection

The hadronic jet reconstruction is described in Section 4.3. The selected events are required to have at least seven jets with $|\eta| < 2.4$ and pass the tight identification criteria listed in Table 4.1. The jets are sorted in descending p_{T} order. The sixth and seventh jets are required to have a p_{T} greater than 40 GeV and 30 GeV respectively. The scalar sum of the p_{T} of all the selected jets H_{T} must be above 500 GeV.

9.1.6 B-tagged jets

Jets originating from the hadronization of bottom quarks (b jets) are identified with the DeepJet b-tagging algorithm as discussed in Section 4.4. At least three out of the selected jets with $p_{\mathrm{T}} > 40$ GeV are required to be b-tagged, using the WP that corresponds to 1% misidentification rate.

9.1.7 Resolved top quarks

The $\mathrm{t}^{\mathrm{res}}$ tagger described in Chapter 7 is used to identify top quarks whose decay products are three separate small cone jets. All selected jets are used to create trijet combinations and only the ones that have a mass between 130 to 210 GeV and at least one b-tagged jet are considered as inputs to the DNN classifier to evaluate the output score. The $\mathrm{t}^{\mathrm{res}}$ candidates are sorted in descending DNN score and are cross cleaned so that each of them does not have shared jets with any $\mathrm{t}^{\mathrm{res}}$ candidate of higher DNN score. Events are then categorized based on the number of medium-tagged $\mathrm{t}^{\mathrm{res}}$ candidates. Two categories are considered in this analysis, one containing exactly two medium $\mathrm{t}^{\mathrm{res}}$ candidates ($2\mathrm{M}_{\mathrm{t}^{\mathrm{res}}}$) and one with exactly one medium and at least one loose-not-medium $\mathrm{t}^{\mathrm{res}}$ candidates ($1\mathrm{M}1\mathrm{L}_{\mathrm{t}^{\mathrm{res}}}$).

9.1.8 Charged Higgs boson reconstruction

The charged Higgs boson invariant mass (m_{tb}) is reconstructed by forming a four-jet system with the leading in p_{T} $\mathrm{t}^{\mathrm{res}}$ candidate and the leading in p_{T} b jet that is not used in the reconstruction of the two $\mathrm{t}^{\mathrm{res}}$ candidates. In the $1\mathrm{M}1\mathrm{L}_{\mathrm{t}^{\mathrm{res}}}$ category, the $\mathrm{t}^{\mathrm{res}}$ candidate assigned as the top quark coming from the charged Higgs boson decay is required to be medium-tagged, otherwise, the event is rejected. The subleading in p_{T} $\mathrm{t}^{\mathrm{res}}$ candidate is considered to be the associated top quark.

9.1.9 Summary

The event selection criteria are summarized in Table 9.3.

Table 9.3: Summary of the criteria used to select the signal analysis sample in the resolved analysis.

1	Signal trigger	AK4 multijet, AK8 jet, Inclusive H_{T}
2	MET Filters	Reject artificially high $\vec{p}_T^{\,\mathrm{miss}}$
3	PV	$\mathrm{N_{dof}} \geq 4$, |z| $<$ 24 cm, $\mathrm{r_{xy}} <$ 2 cm
4	= 0 electrons	$p_{\mathrm{T}} >$ 10 GeV, $\|\eta\| <$ 2.4, miniIso $<$ 0.4, cut-based ID (95% efficiency)
5	= 0 muons	$p_{\mathrm{T}} >$ 10 GeV, $\|\eta\| <$ 2.4, miniIso $<$ 0.4, loose ID
6	= 0 τ_{h}	$p_{\mathrm{T}} >$ 20 GeV, $\|\eta\| <$ 2.1, DeepTau (very loose $e \to \tau_{\mathrm{h}}$, medium $\mu \to \tau_{\mathrm{h}}$, loose $j \to \tau_{\mathrm{h}}$)
7	$\geq$ 7 jets	$\mathrm{p_T^{6th}} >$ 40 GeV, $\mathrm{p_T^{7th}} >$ 30 GeV, $\|\eta\| <$ 2.4, TightID, $H_{\mathrm{T}} >$ 500 GeV
8	$\geq$ 3 b jets	$p_{\mathrm{T}} >$ 40 GeV, DeepJet (1% misID rate)
9	= 2 $\mathrm{t^{res}}$	130 $< m_{\mathrm{t^{res}}} <$ 210 GeV, with $\geq$ 1 b-tagged subjet == 2 medium tagged $\mathrm{t^{res}}$ ($\mathrm{2M_{t^{res}}}$) **OR** == 1 medium tagged $\mathrm{t^{res}}$ and $\geq$ 1 loose-not-medium tagged $\mathrm{t^{res}}$ ($\mathrm{1M1L_{t^{res}}}$)
10	m_{tb}	Leading in p_{T} medium tagged $\mathrm{t^{res}}$ + leading in p_{T} free b jet

9.2 Trigger Performance

The efficiency of the HLT trigger paths used in this analysis is studied as a function of different jet-related variables that are part of the trigger requirement. The trigger efficiency is defined as the fraction of events that pass the reference trigger and the offline selection and in addition satisfy the OR of the signal triggers listed in Table 9.1.

$$\varepsilon_{HLT} = \frac{\text{Offline Selection \&\& Reference Trigger \&\& Signal Trigger}}{\text{Offline Selection \&\& Reference Trigger}} \tag{9.4}$$

The efficiency of the trigger is measured in both data and simulated MC events using a μ+jets sample collected with an orthogonal trigger and with selection criteria similar to the signal event selection discussed in Section 9.1. The events are required to pass the orthogonal single muon trigger `HLT_IsoMu24` for the 2016 and `HLT_IsoMu27` for the 2017 and 2018 data-taking eras. This muon trigger is used as reference trigger to determine the efficiency of the desired trigger. The offline selections require the presence of an isolated muon, identified as loose and with $p_{\mathrm{T}} > 26$ GeV for the 2016 or $p_{\mathrm{T}} > 29$ GeV for the 2017 and 2018 eras, and $|\eta| \leq 2.4$. As per the signal-event selections, events with electrons and τ_h are excluded. At least seven jets identified as tight and within $|\eta| \leq 2.4$ are required, with the six leading in p_{T} jets having $p_{\mathrm{T}} > 40$ GeV and the remaining jets having $p_{\mathrm{T}} > 30$ GeV. The events are categorized based on the number of b-tagged jets. The two categories contain exactly two or at least three b jets with $p_{\mathrm{T}} > 40$ GeV and satisfy the medium WP of the DeepJet algorithm which corresponds to 1% mistagging rate.

The efficiency is studied as a function of the H_{T} and the p_{T} of the sixth jet, and is shown

in Fig. 9.1 and Fig. 9.2 respectively, for all data-taking eras. In the two 2016 eras, negligible discrepancies are observed.

The differences observed in the low p_{T} and H_{T} regions in the 2017 data are caused by three factors. Firstly, during a part of the `Run2017C` era the single b tagging trigger was accidentally prescaled, affecting a luminosity of 6.2 fb^{-1}. Secondly, the high L1 $\mathrm{H_T}$ thresholds at the beginning of the data-taking period affected $\sim$6.2 fb^{-1} of the integrated luminosity. Thirdly, at the end of the `Run2017C` era, a copy of the single b tagging trigger was introduced with the online DeepCSV b tagging algorithm (HLT_PFHT380_SixPFJet32_DoublePFBTagDeepCSV_2p2). This period corresponds to a luminosity of $\sim$27.1 fb^{-1}

The differences observed in the low p_{T} and H_{T} regions in the 2018 data can be attributed to the use of the 2017 multijet trigger paths at the beginning of the data-taking period. These triggers were eventually replaced by the corresponding 2018 HLT trigger paths during the `Run2018B` era. However, simulated samples contain the HLT menu used at the end of the run and therefore the 2017 HLT triggers are not included.

To address the trigger efficiency discrepancies observed in the 2017 and 2018 data, SFs are derived to correct simulation, as will be discussed in Section 9.3.2.

9.3 Corrections to simulation

Differences between data and simulated events are taken into account, by correcting the simulation as described in the following subsections.

9.3.1 Pileup reweighting

The number of pp interactions changes as a function of the instantaneous luminosity during data taking. In simulation, the pileup is simulated as a distribution with fixed average number of interactions. To correct for this effect, the pileup distribution is reweighted according to the real data distribution. In data, the mean number of interactions per bunch crossing is estimated from the product of the measured luminosity in each bunch crossing and the average total pp inelastic cross section of 69.2 mb.

9.3.2 Trigger scale factors

As described in Section 9.2 differences between data and simulation seen in the trigger performance studies are taken into account by deriving SFs to correct simulation. The SFs are taken from the data-to-simulation efficiency ratio and are expressed as a function of the $\mathrm{H_T}$ of the event and the p_{T} of the sixth p_{T}-ordered jet.

To address the high L1 H_{T} thresholds at the beginning of the 2017 data-taking period, the prescale of the single b tagging trigger and the introduction of the HLT trigger with the

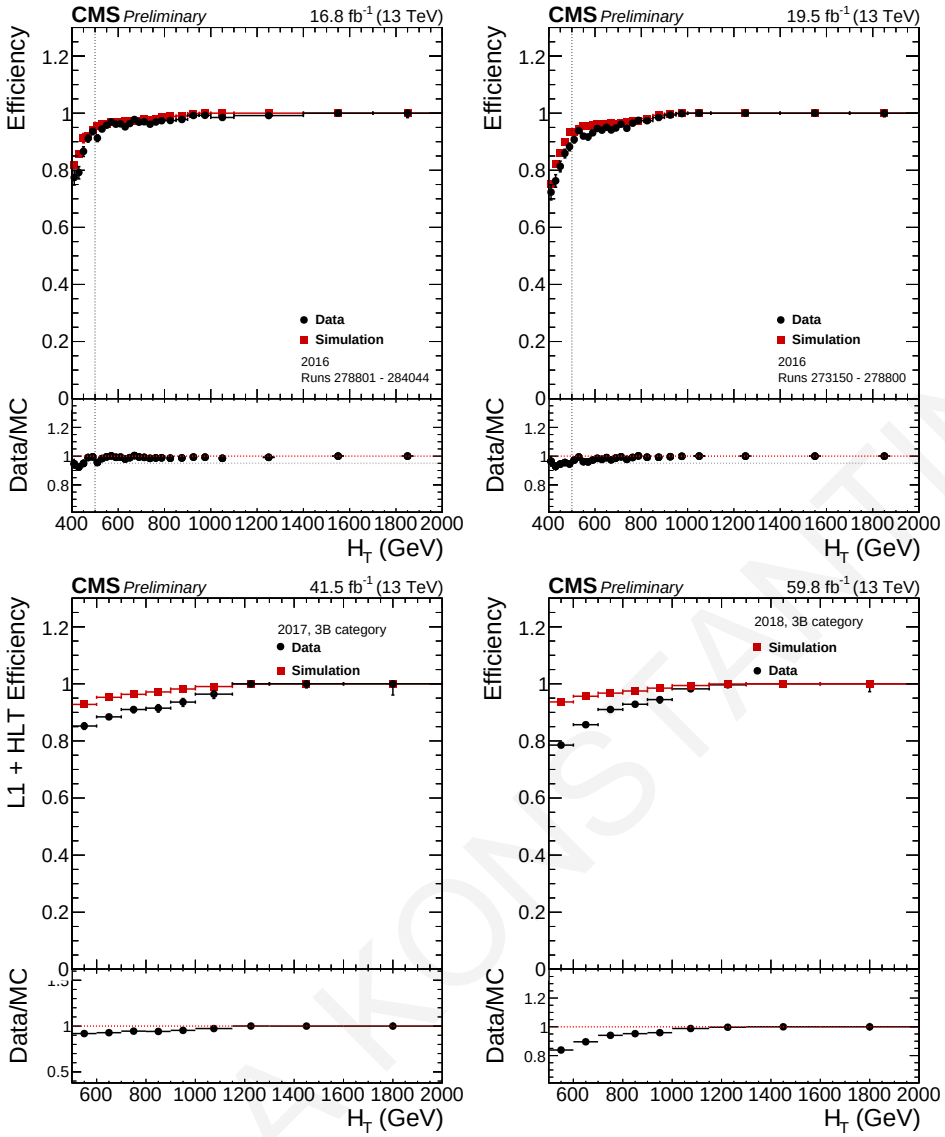


Figure 9.1: Trigger efficiency of the OR of the triggers used in the analysis as a function of the H_{T} of the event for all the data taking eras.

online DeepCSV b tagging algorithm, four scale factors are calculated that correspond to different run periods. The final SF is calculated as the sum of the individual SFs, weighed by the corresponding integrated luminosity. For the 2018 data, a single SF is derived for the entire period.

For the 2016 data-taking period, data and simulation agree and the changes do not exceed 5%. For this reason, no trigger SF was adopted and a flat 5% uncertainty is applied.

9.3.3 b-tagging scale factors

The b tagging efficiency measured in data differs from the efficiency determined in simulated events. These differences are taken into account with SFs which reweigh the simulated events. Per-jet b tagging efficiency SFs are provided by the b tagging physics object group for b and light flavor jets. The per-jet SFs depend on the p_{T}, η and flavor ($\mathcal{F}$) of each jet.

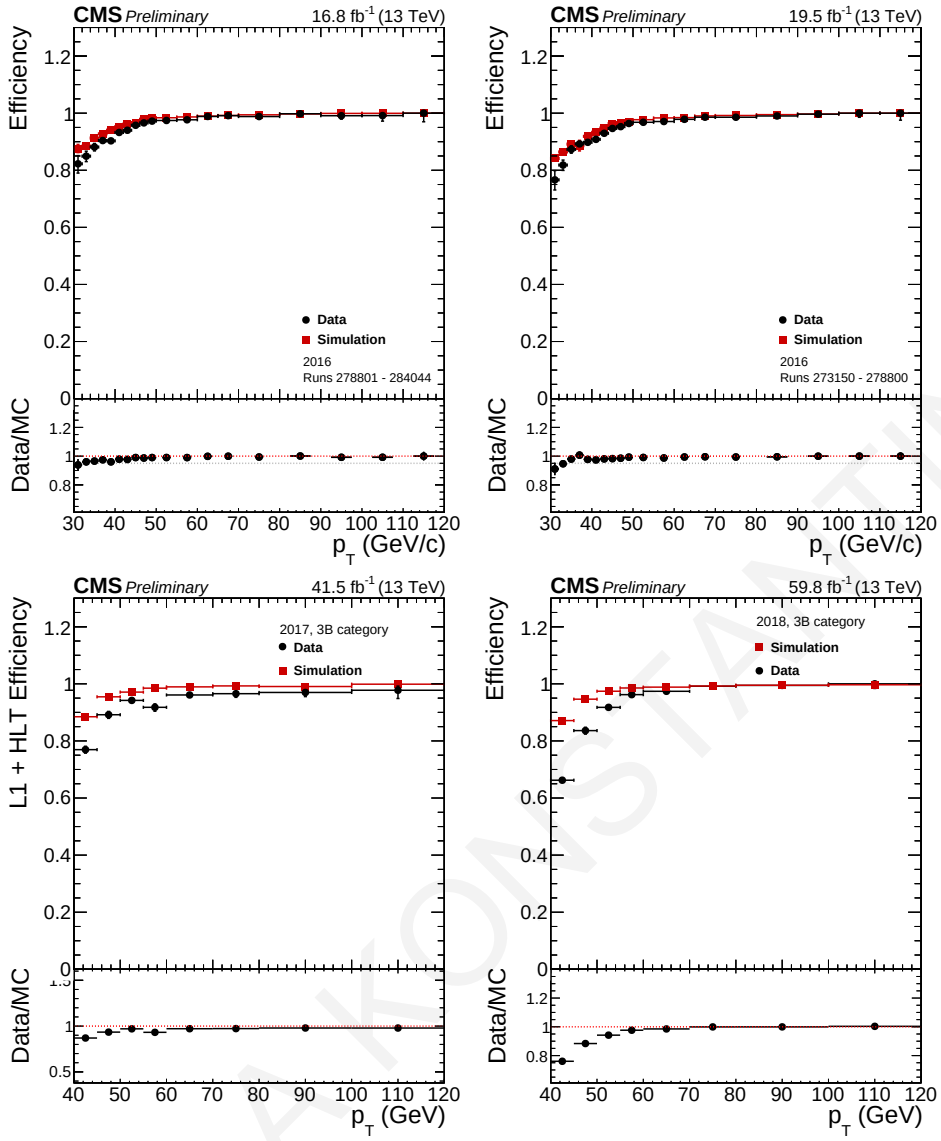


Figure 9.2: Trigger efficiency of the OR of the triggers used in the analysis as a function of the p_{T} of the sixth in p_{T} jet for all the data taking eras.

The per-jet tagging (f^{tag}) and mistagging (f^{mistag}) SFs are difined as:

$$f_{\mathrm{tag}}\left(p_{\mathrm{T}}\right) = \frac{\varepsilon^{\mathrm{tag}}_{\mathrm{Data}}\left(p_{\mathrm{T}}\right)}{\varepsilon^{\mathrm{tag}}_{\mathrm{MC}}\left(p_{\mathrm{T}}\right)} \tag{9.5}$$

$$f_{\mathrm{mistag}}\left(p_{\mathrm{T}}\right) = \frac{\varepsilon^{\mathrm{mistag}}_{\mathrm{Data}}\left(p_{\mathrm{T}}\right)}{\varepsilon^{\mathrm{mistag}}_{\mathrm{MC}}\left(p_{\mathrm{T}}\right)} \tag{9.6}$$

where $\varepsilon^{\mathrm{tag}}$ and $\varepsilon^{\mathrm{mistag}}$ are the per-jet b tagging and mistagging efficiencies. The probability for the event to pass the b tagging selection is:

$$\begin{aligned} \mathrm{P} = & \prod_{i=1}^{N_{\text{b-flavor jets tagged}}} \varepsilon_i^{\text{tag}} \prod_{j=1}^{N_{\text{b-flavor jets not tagged}}} (1-\varepsilon_j^{\text{tag}}) \\ & \times \prod_{k=1}^{N_{\text{light flavor jets tagged}}} \varepsilon_k^{\text{mistag}} \prod_{l=1}^{N_{\text{light flavor jets not tagged}}} (1-\varepsilon_l^{\text{mistag}}). \end{aligned} \tag{9.7}$$

The per-event SFs that reweigh the simulated events can then be calculated as:

$$\mathrm{SF} = \frac{P\,(\text{data})}{P\,(\text{MC})}. \tag{9.8}$$

With Eq. (9.5) and Eq. (9.7), the per-event SF of Eq. (9.8) can be written in terms of the per-jet SFs and b (mis)tagging efficiencies of jets:

$$\begin{aligned} \mathrm{SF} = & \prod_{i}^{N_{\text{bc tagged}}} f_i^{\text{tag}}(p_\mathrm{T},\eta,\mathcal{F}) \cdot \prod_{j}^{N_{\text{bc not tagged}}} \left(\frac{1-f_j^{\text{tag}}(p_\mathrm{T},\eta,\mathcal{F})\varepsilon_j^{\text{bc}}(p_\mathrm{T})}{1-\varepsilon_j^{\text{bc}}(p_\mathrm{T})} \right) \\ & \times \prod_{k}^{N_{\text{udsg tagged}}} f_k^{\text{mistag}}(p_\mathrm{T},\eta,\mathcal{F}) \cdot \prod_{l}^{N_{\text{udsg not tagged}}} \left(\frac{1-f_l^{\text{mistag}}(p_\mathrm{T},\eta,\mathcal{F})\varepsilon_l^{\text{udsg}}(p_\mathrm{T})}{1-\varepsilon_l^{\text{udsg}}(p_\mathrm{T})} \right). \end{aligned} \tag{9.9}$$

In the above equation, the b tagging and mistagging efficiencies refer to the simulated efficiencies and are measured in a sample orthogonal to the one used in the measurement of the per-jet SF. Specifically, simulated QCD multijet, $\mathrm{t\bar{t}}$, W+jets events with at least 6 jets and 2 b jets are used to measure the b tagging and mistagging efficiencies.

9.3.4 $\mathrm{t}^{\mathrm{res}}$ scale factors

Differences in the performance of the $\mathrm{t}^{\mathrm{res}}$-tagger between data and simulation are corrected by reweighting the simulated events with a per-event SF. This SF is calculated by applying the SF of each $\mathrm{t}^{\mathrm{res}}$ candidate in the event. Four categories characterize the $\mathrm{t}^{\mathrm{res}}$ candidates based on their tagging and matching status:

- tagged and truth-matched (t|gen-t)
- tagged and unmatched (t|!gen-t)
- not tagged and truth-matched (!t|gen-t)
- not tagged and unmatched (!t|!gen-t).

The simulated tagging efficiency and misidentification rates are denoted as $\varepsilon_i\,(\mathrm{t} \mid \text{gen-t})$ and $\varepsilon_i\,(\mathrm{t} \mid \text{!gen-t})$ respectively and depend on the p_T of the $\mathrm{t}^{\mathrm{res}}$ candidate with index i. The

probability of a simulated event passing the t^{res}-tagging selections is the product of the individual efficiencies of all the t^{res} candidates falling into the four categories:

$$P\,(\text{MC}) = \prod_{i=1}^{\text{t|gen-t}} \varepsilon_i\,(\text{t} \mid \text{gen-t}) \times \prod_{j=1}^{\text{t|!gen-t}} \varepsilon_j\,(\text{t} \mid \text{!gen-t}) \times \prod_{k=1}^{\text{!t|gen-t}} \left[1 - \varepsilon_k\,(\text{t} \mid \text{gen-t})\right] \times \prod_{l=1}^{\text{!t|!gen-t}} \left[1 - \varepsilon_l\,(\text{t} \mid \text{!gen-t})\right] \tag{9.10}$$

The probability of an event in data is expressed in terms of the simulated efficiencies and per-t^{res}-candidate SFs as:

$$P\,(\text{data}) = \prod_{i=1}^{\text{t|gen-t}} \text{SF}_i\,(\text{t} \mid \text{gen-t})\,\varepsilon_i\,(\text{t} \mid \text{gen-t}) \times \prod_{j=1}^{\text{t|!gen-t}} \text{SF}_j\,(\text{t} \mid \text{!gen-t})\,\varepsilon_j\,(\text{t} \mid \text{!gen-t}) \times \prod_{k=1}^{\text{!t|gen-t}} \left[1 - \text{SF}_k\,(\text{t} \mid \text{gen-t})\,\varepsilon_k\,(\text{t} \mid \text{gen-t})\right] \times \prod_{l=1}^{\text{!t|!gen-t}} \left[1 - \text{SF}_l\,(\text{t} \mid \text{!gen-t})\,\varepsilon_l\,(\text{t} \mid \text{!gen-t})\right] \tag{9.11}$$

Finally, the event weight to correct the simulation is given as an expression of products of the SF and MC tagging efficiency and misidentification rate:

$$w = \frac{P\,(\text{data})}{P\,(\text{MC})} = \prod_{i=1}^{\text{t|gen-t}} \text{SF}_i\,(\text{t} \mid \text{gen-t}) \times \prod_{j=1}^{\text{t|!gen-t}} \text{SF}_j\,(\text{t} \mid \text{!gen-t}) \times \prod_{k=1}^{\text{!t|gen-t}} \frac{\left[1 - \text{SF}_k\,(\text{t} \mid \text{gen-t})\,\varepsilon_k\,(\text{t} \mid \text{gen-t})\right]}{\left[1 - \varepsilon_k\,(\text{t} \mid \text{gen-t})\right]} \times \prod_{l=1}^{\text{!t|!gen-t}} \frac{\left[1 - \text{SF}_l\,(\text{t} \mid \text{!gen-t})\,\varepsilon_l\,(\text{t} \mid \text{!gen-t})\right]}{\left[1 - \varepsilon_l\,(\text{t} \mid \text{!gen-t})\right]}. \tag{9.12}$$

9.4 Background estimation

In the resolved regime, the dominant QCD multijet background is measured with a data-driven method while all the remaining SM processes are modeled using simulation. The data-driven method splits the phase space into four orthogonal regions by inverting two uncorrelated event selection criteria, defining 3 control regions (CRs) enriched in QCD multijet events. The layout of the four regions comprising the QCD multijet measurement method is shown in Fig. 9.3 (right). The CRs are determined by inverting at least one of the conditions of the MVA score of the t^{res} candidate from the $H^{\pm}$ decay and the $m_{t^{res}}$ of the associated t^{res} candidate. CRs with a tagged t^{res} candidate are labeled with "t" while CRs containing

non-tagged t^{res} candidate are labeled with "!t". Similarly, CRs where the associated t^{res} candidate is on or off the mass window, are labeled as "on-m" and "off-m" respectively. Events that pass all the selection criteria belong to the SR. In the inverted $m_{t^{res}}$ region labeled as $\mathrm{CR(off-m, t)}$, the mass of the associated t^{res} candidate falls within one of the two mass sidebands (SBs), defined by $m_{t^{res}} \in [50, 130]$ GeV and $m_{t^{res}} \in [210, 260]$ GeV as displayed in Fig. 9.3 (left). This CR is used to obtain the shapes of the observables, as the QCD multijet events possess similar kinematic properties to the events in the SR. To predict the number of QCD multijet events in the SR, two additional regions $\mathrm{CR(on-m, !t)}$ and $\mathrm{CR(off-m, !t)}$ are used to calculate the transfer factors (TF) that normalize the $\mathrm{CR(off-m, t)}$ to the SR. The TF are defined by the $\mathrm{CR(on-m, !t)}$-to-$\mathrm{CR(off-m, !t)}$ ratio. In both "!t" CRs, the t^{res} candidate from the $H^{\pm}$ decay fails the medium WP of the t^{res} tagger. To eliminate any kinematic differences between "on-m" and "off-m" t^{res} candidates, the method is performed in bins of the t^{res} candidate's p_T as follows:

- $\mathrm{1M1L_{t^{res}}}$: $p_T < 175$ GeV, $p_T > 175$ GeV
- $\mathrm{2M_{t^{res}}}$: $p_T < 100$ GeV, $p_T \in [100, 300]$ GeV, $p_T > 300$ GeV

The expected yields of QCD multijet events in the SR are estimated by:

$$\mathrm{N_{QCD}^{SR}} = \sum_{i}^{p_{T,i}} \mathrm{N_{QCD,i}^{CR(off-m,t)}} \cdot \left(\frac{\mathrm{N_{QCD,i}^{CR(on-m,!t)}}}{\mathrm{N_{QCD,i}^{CR(off-m,!t)}}} \right) \tag{9.13}$$

where index i runs over all the p_T bins of the associated t^{res} candidate and $\mathrm{N_{QCD}}$ is the number of QCD multijet events in each region which is estimated after subtracting from data, the contributions from the simulated SM background:

$$\mathrm{N_{QCD}} = \mathrm{N_{Data}} - \mathrm{N_{t\bar{t}}} - \mathrm{N_{t,t\bar{t}+X,EWK}} \tag{9.14}$$

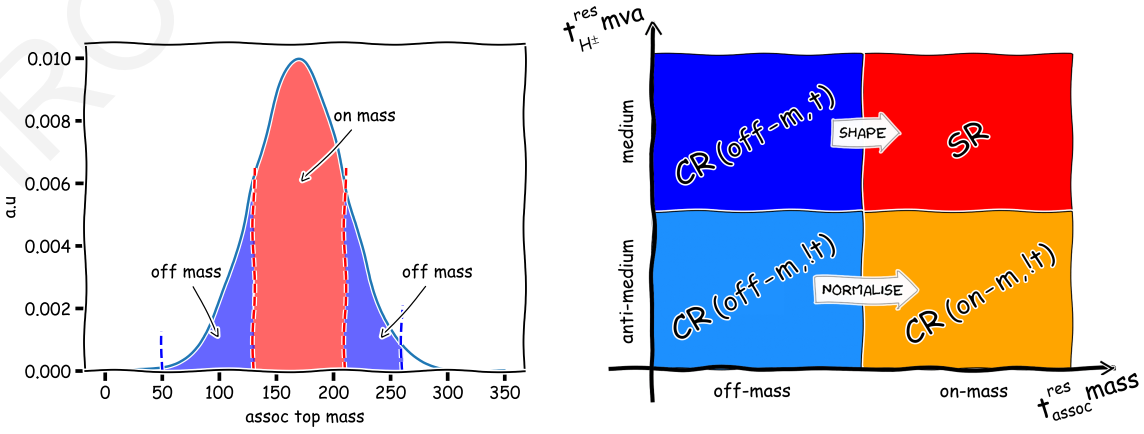


Figure 9.3: Left: Diagram showing the associated t^{res} candidate's mass ranges that define the SR (red) and the SBs (blue). Right: Schematic diagram of the SR and the four CRs used in the QCD multijet background estimation method.

To verify the validity of the background estimation method, the assumption of orthogonality of the two variables that define the CRs is studied. Figure 9.4 illustrates the MVA score shape of the t^{res} candidate from the $H^{\pm}$ decay in the SR, low sideband SB^{-}, upper sideband SB^{+} and the inclusive SB region, in QCD multijet events measured in data and $t\bar{t}$ simulated events. The distributions are created with the 2018 data and are normalized to unity. In QCD multijet events, the two variables show no correlation. In $t\bar{t}$, a small correlation between the SR and the SB^{-}, SB^{+} is observed in the SB-to-SR ratio, which is compensated when combining the two SBs.

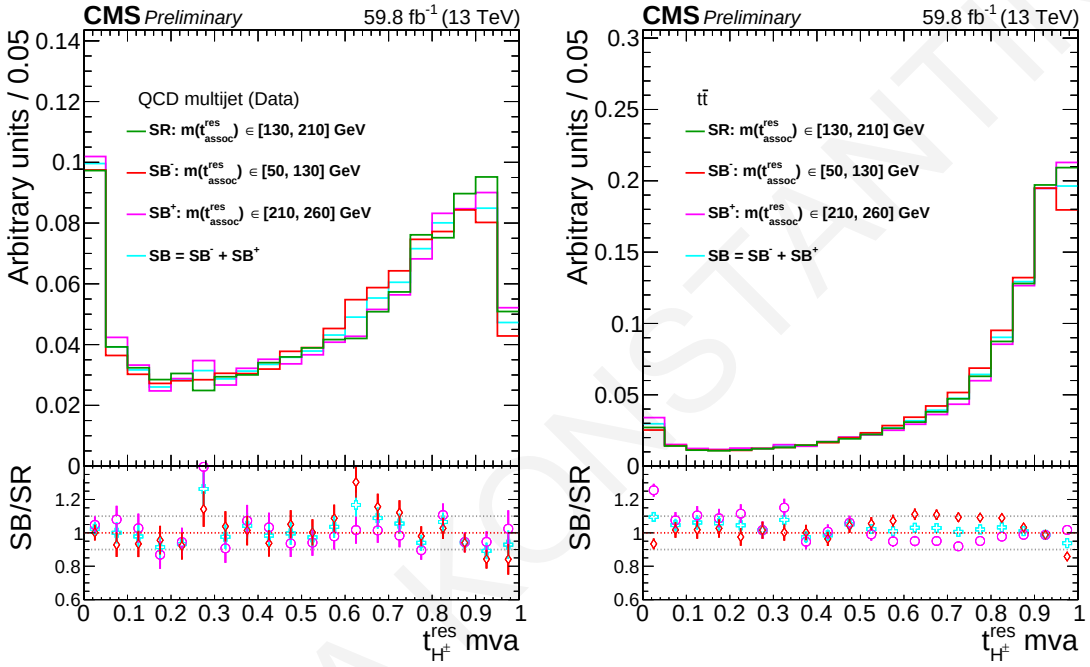


Figure 9.4: The MVA score of the t^{res} candidate from the $H^{\pm}$ decay in the SR and SB, normalized to unity for QCD multijet events measured in data (left) and simulated $t\bar{t}$ events (right).

9.4.1 Closure in validation region

The method used for the measurement of the QCD multijet background is validated in two validation regions (VR) which are orthogonal to the SR. The VRs fulfill the selection criteria described in Section 9.1 except for the b tagging selection that requires exactly two b-tagged jets. The two b-tagged jets are forced to be inside the two t^{res} candidates. Since no extra b jets are left for the reconstruction of the $H^{\pm}$ candidate, the leading free jet is used to suppress further any signal contamination.

The first VR is enriched in QCD multijet events and is labeled as VR_{QCD}. In this region the mass of the t^{res} candidate from the $H^{\pm}$ decay is outside a tighter mass window of $155 < m_{t^{res}} < 195\,\text{GeV}$. To enhance the purity in QCD multijet events, the minimum ΔR distance between a b-b pair ($\Delta R_{\text{min}}(\text{bb})$), including the jet used for the reconstruction of the $H^{\pm}$ candidate, is less than 1.0. This requirement keeps collinear $b\bar{b}$ pairs from gluon

splitting.

The second VR is dominated by $t\bar{t}$ events and is labeled as $\mathrm{VR}_{t\bar{t}}$. In the $\mathrm{VR}_{t\bar{t}}$, the mass of the t^{res} candidate from the $H^{\pm}$ decay is $155 < m_{t^{res}} < 195$ GeV and the minimum $\Delta R_{\mathrm{min}}(bb)$ is greater than 1.2.

The background estimation method is applied in $\mathrm{VR}_{\mathrm{QCD}}$ and $\mathrm{VR}_{t\bar{t}}$ to measure the QCD multijet background. The performance of the method is validated by comparing the distributions of the observables in data and background. Figures 9.5 and 9.6 show the reconstructed invariant mass of the $H^{\pm}$ candidate (m_{tb}) for the 2018 observed data, estimated background and expected signal of $m_{H^{\pm}} = 800$ GeV in $\mathrm{VR}_{\mathrm{QCD}}$ and $\mathrm{VR}_{t\bar{t}}$ respectively. In $\mathrm{VR}_{t\bar{t}}$, the contribution from $t\bar{t}$ is $\sim 65\%$ in the $1\mathrm{M1L}_{t^{res}}$ and $\sim 85\%$ in the $2\mathrm{M}_{t^{res}}$ category. In $\mathrm{VR}_{\mathrm{QCD}}$, QCD multijet events comprise $\sim 85\%$ and $\sim 65\%$ of the events in the $1\mathrm{M1L}_{t^{res}}$ and $2\mathrm{M}_{t^{res}}$ categories respectively. The presence of signal is negligible in all VRs and categories. Data and estimated background show good agreement and the result verifies the validity of the method.

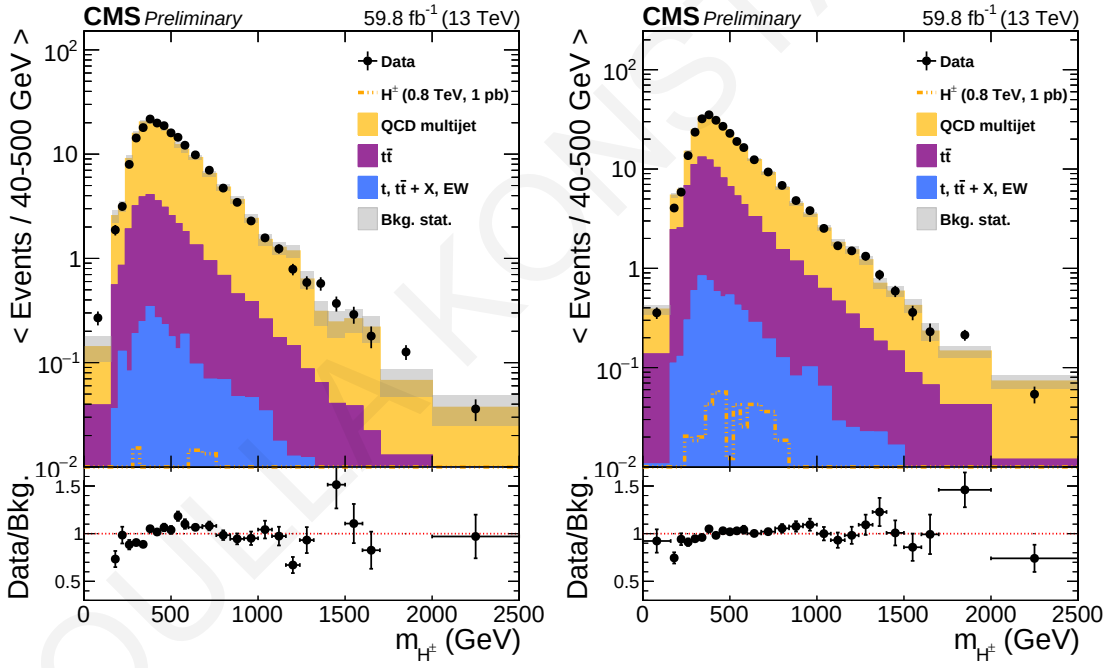


Figure 9.5: The m_{tb} distribution for the observed 2018 data, estimated background and expected signal of $m_{H^{\pm}} = 800$ GeV in $\mathrm{VR}_{\mathrm{QCD}}$ region, for the $1\mathrm{M1L}_{t^{res}}$ (left) and $2\mathrm{M}_{t^{res}}$ (right) category.

9.5 Signal extraction

A DNN-based discriminant has been developed to enhance the separation between signal of $H^{\pm}$ from background events. This event-based DNN is parameterized in respect of the $m_{H^{\pm}}$ to account for dependencies of the kinematic features from $m_{H^{\pm}}$. The idea of a parameterized DNN [163] is to generalize its performance in a wide range of the physics parameters' values by extending its input to include also the set of desired parameters. In simple DNNs, the

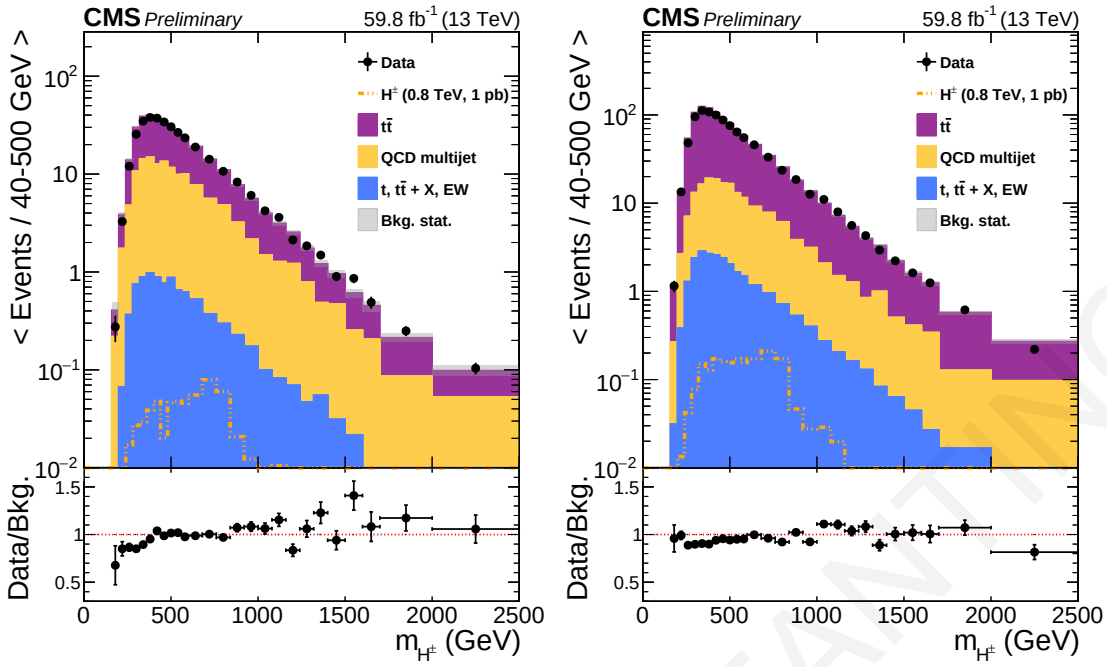


Figure 9.6: The m_{tb} distribution for the observed 2018 data, estimated background and expected signal of $m_{\mathrm{H}^{\pm}} = 800\,\mathrm{GeV}$ in $\mathrm{VR}_{\mathrm{t\bar{t}}}$ region, for the $1\mathrm{M}1\mathrm{L}_{\mathrm{t}^{\mathrm{res}}}$ (left) and $2\mathrm{M}_{\mathrm{t}^{\mathrm{res}}}$ (right) category.

input set of features $\bar{x}$ is described by specific values of a parameter set $\bar{\theta}$ and the output of the DNN is a function of $\bar{x}$, $f(\bar{x})$. Trivial solutions to expand the discrimination of a DNN in different choices of parameters are the development of a set of parameter-specific DNNs or the use of a mixture of samples from different parameters as input to a single DNN. These approaches would result in reduced performance, especially away from the parameter values that were used during the training and lead to discontinuities in the performance between the individual networks. On the contrary, parameterized DNNs receive as input the set of features and the distinct parameter values and associate them, providing a conditional output in terms of both $\bar{x}$ and $\bar{\theta}$, $f(\bar{x}, \bar{\theta})$. The construction of a parameterized DNN is visualized in Fig. 9.7 where on the left diagram, individual networks are trained with different parameter values $\theta = \theta_{\alpha}, \theta_{\beta}$, while in the right diagram, the parameterized DNN receives all the possible values of θ. The decision of the algorithm differs between parameter hypotheses. In the case of a mass-parameterized DNN, a specific event might be signal-like in a single $m_{\mathrm{H}^{\pm}}$ region but background-like in different choices of $m_{\mathrm{H}^{\pm}}$.

Parameterized networks have interpolation power between values of parameters that are not exposed to during the training. Moreover, they exploit correlations from the full parameter-dependent dataset resulting in improved performance compared to simple DNNs which remains optimal under systematic variations in the total parameter space.

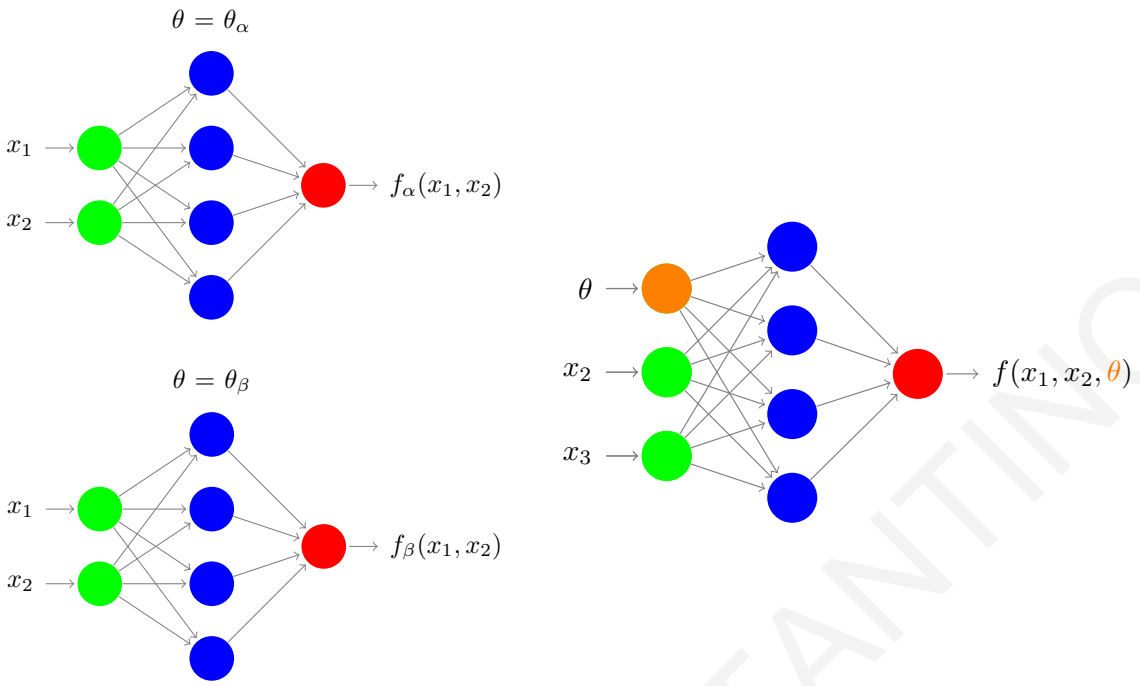


Figure 9.7: Individual networks trained with a single value of θ parameter are shown on the left diagram. The right diagram presents the parameterized DNN diagram, trained with the input features and the possible values of θ.

9.5.1 Training dataset

The parameterized DNN algorithm is a classifier that distinguishes $H^{\pm}$ signal events of different mass hypotheses from $t\bar{t}$ and combinatorial background. The input dataset is a combination of events from both $1M1L_{t^{res}}$ and $2M_{t^{res}}$ categories. The signal dataset is selected from events passing all the SR selections, using 6 different simulated samples of $m_{H^{\pm}}$ = [220, 350, 600, 1000, 1500, 2500] GeV. This selection is optimized so that it covers as much $m_{H^{\pm}}$ spectrum as possible. In the background dataset, both $t\bar{t}$ and combinatorial events are measured using $t\bar{t}$-simulated events. The $t\bar{t}$ events pass all the selection requirements of the SR. Combinatorial QCD multijet-like background contains events from the $\mathrm{CR}(\mathrm{off-m,t})$ in which none of the t^{res} candidates is truth-matched to a top-quark at generator level. The proportion of $t\bar{t}$ and QCD multijet-like events is set to match the background composition of the analysis in each of the two event categories. During the training, all the events are weighted using the corrections described in Section 9.1. Combinatorial events are also corrected using the p_{T} - binned TF discussed in Section 9.1, to take into account kinematic differences between $\mathrm{CR}(\mathrm{off-m,t})$ and SR. The total dataset consists of equal signal and background samples and the signal dataset is equally split into the 6 $m_{H^{\pm}}$ values. For orthogonality purposes, the DNN used to evaluate the data of each data-taking era is trained with the data of different eras. To check the consistency of the output between different eras, overtraining tests are performed between training and testing samples.

9.5.2 Input features

The input of the event-based tagger is a set of 12 discriminating features that include kinematic properties of the reconstructed jet systems, angular separations and event shape variables. The input is expanded to include also the true $m_{\mathrm{H}^\pm}$ parameter which is assigned to each event. For the background where $m_{\mathrm{H}^\pm}$ value is not physically meaningful, a random value is assigned to each event according to the values used for the signal, following a uniform distribution. The list of the 13 inputs is presented in Table 9.4.

Table 9.4: Summary of the input features and parameter used for the development of the parameterized event discriminator.

1	m_{tb}
2	$\Delta\theta(\mathrm{t}_{\mathrm{H}^\pm}, \mathrm{b}_{\mathrm{H}^\pm})$ in $\mathrm{H}^\pm$ CM
3	$\mathrm{p}_{\mathrm{T}}^{\mathrm{Asym}}(\mathrm{H}^\pm, \mathrm{b}_{\mathrm{H}^\pm})$
4	$\mathrm{H}_{\mathrm{T,3b}} = p_{\mathrm{T}}\,(\mathrm{b}_{\mathrm{t}^{\mathrm{res}}_{assoc}}) + p_{\mathrm{T}}\,(\mathrm{b}_{\mathrm{t}^{\mathrm{res}}_{\mathrm{H}^\pm}}) + p_{\mathrm{T}}\,(\mathrm{b}_{\mathrm{H}^\pm})$
5	$\mathrm{p_T}(\mathrm{bb}_{\Delta\mathrm{Rmin}})$
6	$m(\mathrm{bb}_{max\ p_{\mathrm{T}}})$
7	$\mathrm{p}_{\mathrm{T,b}(\mathrm{H}^\pm)}/\mathrm{H}_{\mathrm{T,3b}}$
8	$\mathrm{y23} = \mathrm{p}^2_{\mathrm{T,j3}}/(\mathrm{p}_{\mathrm{T,j1}} + \mathrm{p}_{\mathrm{T,j2}})^2$
9	N medium $\mathrm{t}^{\mathrm{res}}$
10	Circularity
11	Sphericity
12	Aplanarity
13	True mass parameter $m_{\mathrm{H}^\pm}$

Variables related to the $\mathrm{H}^\pm$ and the objects used for the reconstruction of the $\mathrm{H}^\pm$ candidate, the leading in p_{T} $\mathrm{t}^{\mathrm{res}}$ candidate ($\mathrm{t}_{\mathrm{H}^\pm}$) and the b jet with the highest p_{T} ($\mathrm{b}_{\mathrm{H}^\pm}$), include:

- the $\mathrm{H}^\pm$ invariant mass, $\boldsymbol{m_{\mathrm{tb}}}$
- the polar-angle separation in the $\mathrm{H}^\pm$ center-of-mass, $\boldsymbol{\Delta\theta(\mathrm{t}_{\mathrm{H}^\pm}, \mathrm{b}_{\mathrm{H}^\pm})}$.
- the absolute p_{T} difference between the $\mathrm{H}^\pm$ and the b jet from the $\mathrm{H}^\pm$ over their p_{T} sum defined as $\boldsymbol{\mathrm{p}_{\mathrm{T}}^{\mathrm{Asym}}(\mathrm{H}^\pm, \mathrm{b}_{\mathrm{H}^\pm})}$.

Four variables are associated with the three b jets from the $\mathrm{t}^{\mathrm{res}}$ candidates and $\mathrm{H}^\pm$ decay:

- the scalar sum of their p_{T}, $\boldsymbol{\mathrm{H}_{\mathrm{T,3b}}}$
- the p_{T} of the bb pair with the minimum ΔR distance, $\boldsymbol{\mathrm{p_T}(\mathrm{bb}_{\Delta\mathrm{Rmin}})}$

- the invariant mass of the bb pair with the maximum p_{T}, $\boldsymbol{m(\mathbf{bb}_{max\ p_{\mathrm{T}}})}$
- the fraction of the $\mathrm{H}_{\mathrm{T,3b}}$ carried by the b jet from the $\mathrm{H}^{\pm}$ decay, $\mathbf{p_{T,b(H^{\pm})}/H_{T,3b}}$.

Two additional variables used as input of the event-based tagger are:

- the third highest p_{T} jet resolution is determined by the variable, **y23**
- the number of medium-tagged $\mathrm{t}^{\mathrm{res}}$ candidates, used to increase the discrimination power in the two categories $1\mathrm{M}1\mathrm{L}_{\mathrm{t}^{\mathrm{res}}}$ and $2\mathrm{M}_{\mathrm{t}^{\mathrm{res}}}$.

Finally, three event-shape variables are defined by the eigenvalues of the normalized sphericity tensor:

$$M_{\alpha\beta} = \frac{\sum_{i}^{\text{jets}} p_{i,\alpha} p_{i,\beta}}{\sum_{i}^{\text{jets}} |\vec{p}_i|^2} \tag{9.15}$$

where index i runs over all the jets in the event, α, β run through the x,y,z components. The eigenvalues fulfill the condition $\lambda_1 \geq \lambda_2 \geq \lambda_3$ with $\sum_{i}^{3} \lambda_i = 1$. Based on the eigenvalues, sphericity and aplanarity are defined as follows.

$$S = \frac{3}{2}(\lambda_2 + \lambda_3) \qquad 0.0 \leq S \leq 1.0 \quad \text{Sphericity} \tag{9.16}$$

$$A = \frac{3}{2}(\lambda_3) \qquad 0.0 \leq A \leq 0.5 \quad \text{Aplanarity} \tag{9.17}$$

The sphericity of an event is a measurement of the summed squared transverse momentum of all jets $p_{\perp}^2$ with respect to the event axis and characterizes an event as isotropic if $S \approx 1$. Aplanarity is the transverse momentum component out of the event plane and characterizes an event as planar when $A \approx 0$ or isotropic if $A \approx \frac{1}{2}$. Following the definition of the sphericity tensor, we define the two-dimensional tensor in the transverse plane (α, β = x,y). The circularity is defined as:

$$\mathcal{C} = 2 \times \min \frac{\lambda_1, \lambda_2}{\lambda_1 + \lambda_2} \quad \text{Circularity} \tag{9.18}$$

and is independent from boosts along the z direction. Additionally, the normalization by the sum of the particle momenta makes it independent from energy correction effects. Circular events have large circularity values and as the value decreases the event becomes linear.

The distributions of the 12 variables are shown in Fig. 9.8 and the correlation matrix for signal mass hypotheses of 350 GeV and 1000 GeV is presented in Fig. 9.9.

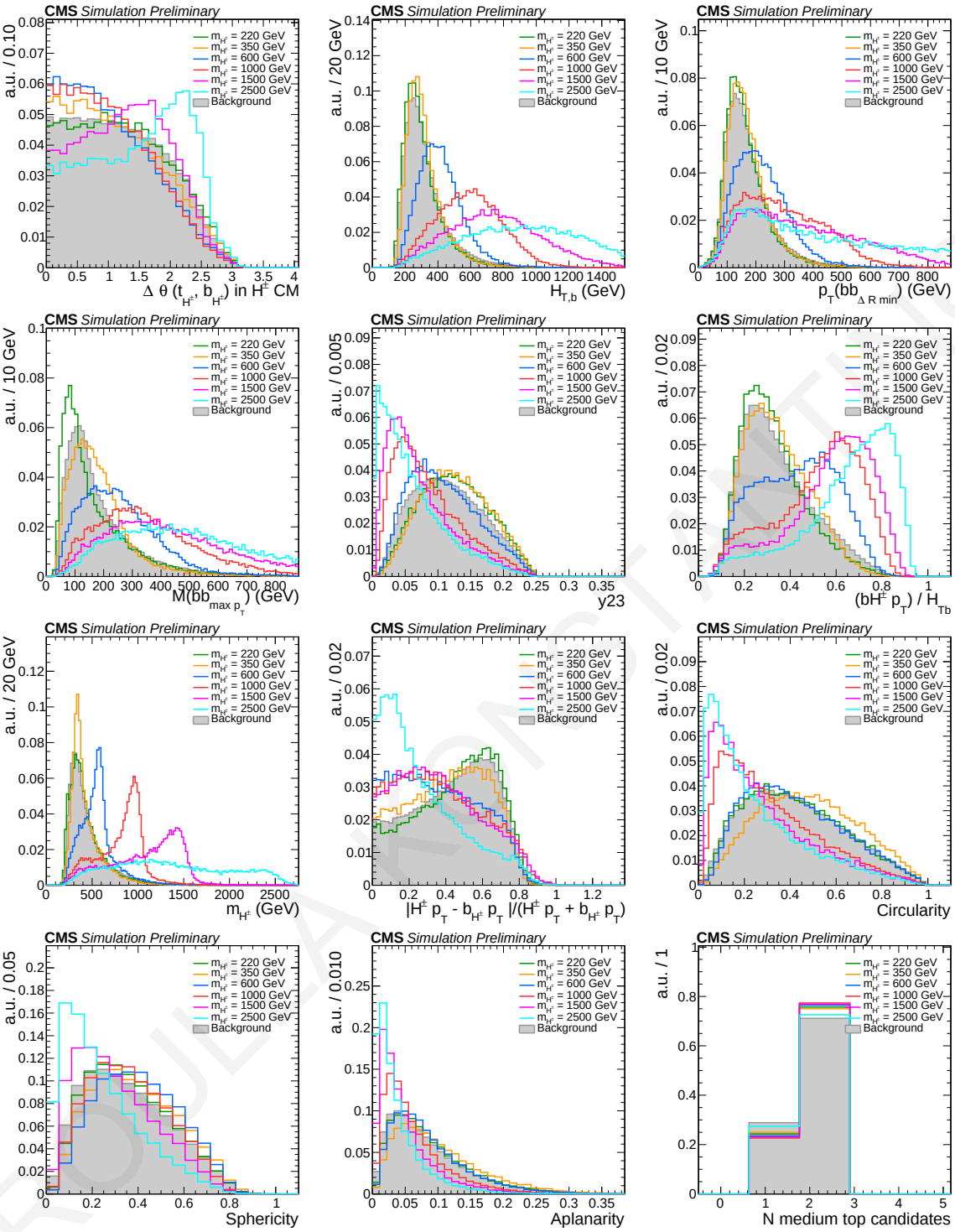


Figure 9.8: Distributions of the input variable used for the training of the parameterized DNN.

9.5.3 Architecture and data processing

The DNN method is provided by the KERAS [154] and TENSORFLOW [155] software packages. The parameterized DNN is a fully connected neural network that consists of 3 hidden dense layers with 32 neurons each. A batch normalization layer is introduced between all layers to accelerate the training procedure and improve the performance of the algorithm. The batch normalization layer applies a transformation on the input batch that keeps its

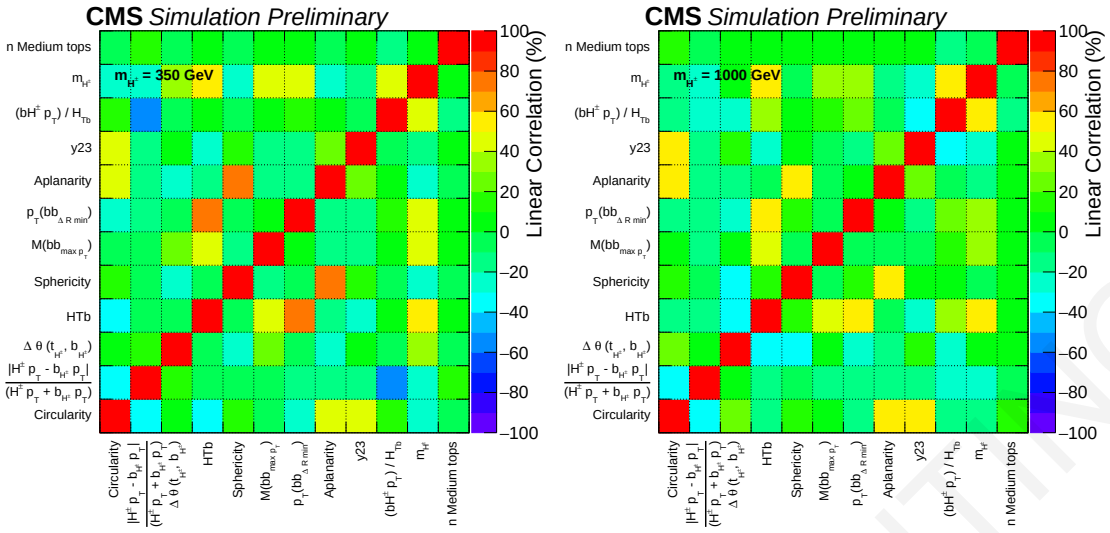


Figure 9.9: The correlation matrix shows the pairwise correlation of all the variables, for signal mass of 350 GeV (left) and 1000 GeV (right).

mean close to zero and the standard deviation close to one. Mathematically, the transformed output of the batch normalization layer is expressed by:

$$\text{output} = \frac{\text{batch} - \text{mean}(\text{batch})}{\text{var}(\text{batch}) + \varepsilon} \cdot \gamma + \beta \tag{9.19}$$

where ε is a small float that prevents division by zero and β, γ are an offset and a scaling trainable parameters respectively. The dense hidden layers and the output layer use the ReLU and sigmoid activation functions respectively. The Adam [156] optimizer is used to minimize the binary cross-entropy loss function with a learning rate of 0.001. The network is trained for 1000 epochs or until no improvement is observed in the area under the ROC curve with batches of 32768 events.

9.5.4 Performance

The performance of the algorithm is expressed in terms of the ROC curve which is produced for all the available signal mass points. For each mass point, the parameterized DNN is evaluated at the true mass of the generated signal. A set of individual DNNs of fixed $m_{\mathrm{H}^\pm}$ value has been trained using the same architecture to compare its performance with that of the parameterized network. The ROC curve in Fig. 9.10 presents the signal-to-background efficiency for the parameterized DNN with a solid line and the individual networks with a dashed line. The results demonstrate that the performance of the parametrized DNN for various $m_{\mathrm{H}^\pm}$ masses even the ones not used in its training, is compatible to the performance of the DNNs trained with fixed $m_{\mathrm{H}^\pm}$ values. This verifies the interpolation power of the parameterized DNN.

The DNN score is evaluated with the training and test datasets as seen in Fig. 9.11 for all training mass hypotheses. Similar performance is observed between the two datasets,

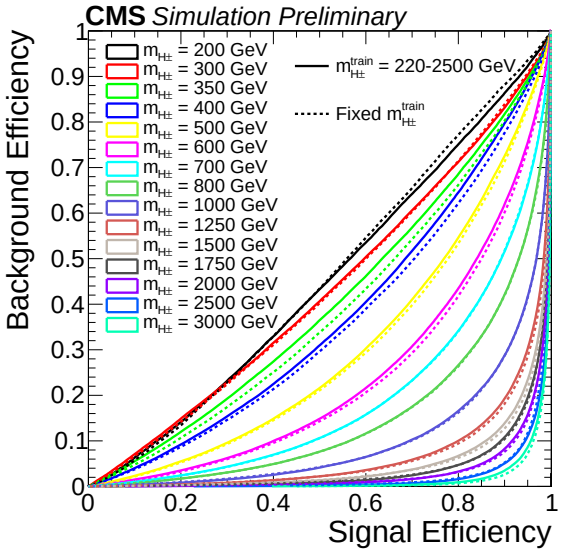


Figure 9.10: The ROC curves of the parameterized DNN, evaluated at the true mass of the generated signal. The performace is compared to the one obtained by individual DNNs trained with a fixed $m_{\mathrm{H}^\pm}$.

indicating that the algorithm is not overtrained. The distributions corresponding to higher $m_{\mathrm{H}^\pm}$ hypotheses display greater separation between the signal and background, while small discrimination is observed for the very low masses.

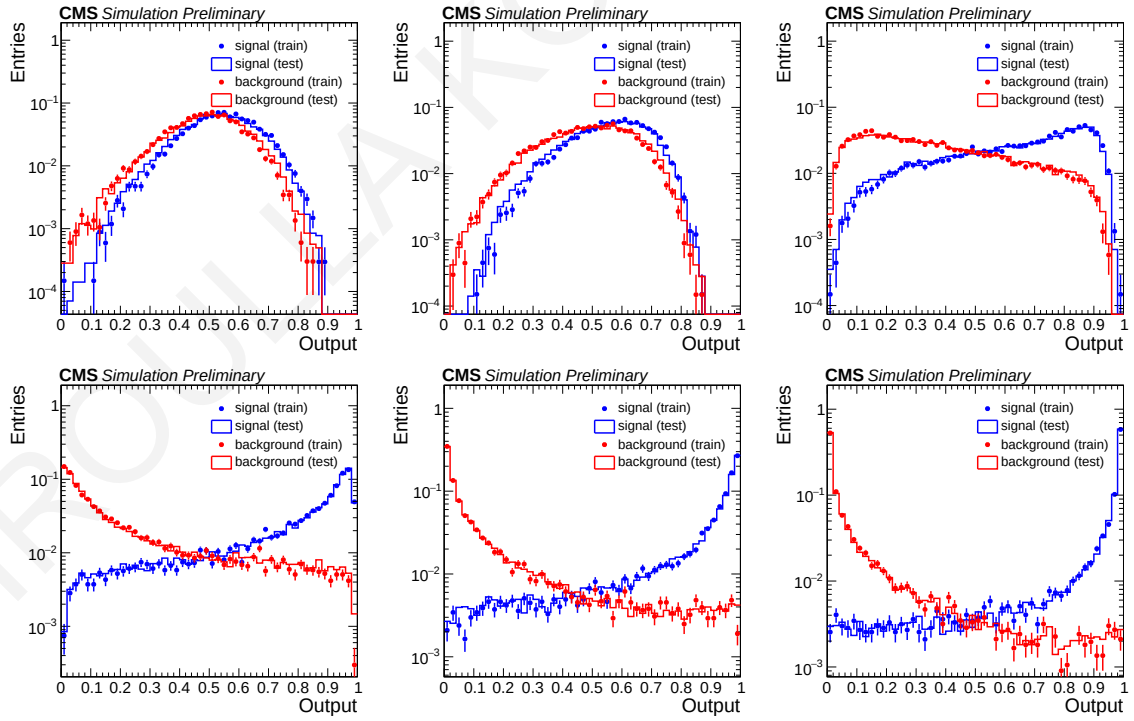


Figure 9.11: Overtraining test of the parameterized DNN, for all $m_{\mathrm{H}^\pm}$ used during the training. Top row: Overtraining test for mass hypotheses of 220 GeV (left), 250 GeV (middle) and 600 GeV (right). Bottom row: Overtraining test for mass hypotheses of 1000 GeV (left), 1500 GeV (middle) and 2500 GeV (right).

9.6 Systematic uncertainties

Systematic uncertainties are introduced as nuisance parameters in the final maximum likelihood fit to data. This section discusses the sources of systematic uncertainties which may affect the rate of the final yields, the shape of the fit discriminant, or both the rate and shape, and their contribution to the total uncertainty.

9.6.1 Luminosity uncertainties

The integrated luminosity for each data-taking era was measured as described in [164] with total systematic uncertainties of 1.2%, 2.3% and 2.5% for the 2016, 2017 and 2018 data sets respectively [164–166]. The uncertainties are treated as partially correlated among the eras.

9.6.2 Trigger efficiency uncertainties

Trigger efficiency uncertainties are determined by the statistical uncertainties of the measured trigger efficiency SFs discussed in Section 9.2 and Section 9.3.2. The SF up and down variations are propagated through the analysis, affecting both the shape and the rate of the final fit discriminant. For the 2016 data where no trigger SF is applied, a flat systematic uncertainty of 5% is adopted, affecting only the rate of the expected yields.

9.6.3 Pileup reweighting uncertainties

The uncertainties due to the pileup modeling are estimated by shifting the total inelastic pp interactions cross section by $\pm 5\%$ [167]. Pileup reweighting uncertainties affect both the shape and rate of the final fit distribution.

9.6.4 Lepton veto uncertainties

Uncertainties due to vetoing electrons, muons and hadronically decaying τ-leptons are calculated as:

$$\theta_\ell = \left(\frac{N_\ell^{\text{fail}}}{N_\ell^{\text{pass}}} \right) \times \Delta_{\ell\,\text{ID}} \tag{9.20}$$

where N_ℓ^{fail} and N_ℓ^{pass} correspond to the number of events that fail or pass the ℓ (e, μ, τ_h) veto. The misidentification rates $\Delta_{e\text{-ID}}$, $\Delta_{\mu\text{-ID}}$, and $\Delta_{\tau_\text{h}\text{-ID}}$ are equal to 2%, 1%, and 3%, for electrons, muons and τ_h's respectively.

9.6.5 JES uncertainties

Uncertainties related to the measurement of the JES are determined by varying the energy of the jets by $\pm$ 1 standard deviation around the nominal correction. The variation affects also

the measured $E_{\mathrm{T}}^{\mathrm{miss}}$. The events are re-analyzed and observed variations affect both the rate and shape of the final fit discriminant.

9.6.6 JER uncertainties

To account for the difference in JER between simulation and data, the jets in the simulation are smeared so that the p_{T} resolution describes the data. This is done by increasing and decreasing the energy of the jets in simulation around the nominal value. The effect is propagated through the analysis and results in both shape and rate variation.

9.6.7 B tagging and mistagging uncertainties

Uncertainties due to b tagging and mistagging efficiencies are determined by variations on the per-jet SFs described in Section 9.3.3. These variations are quantified by applying error propagation on each of the four terms in Eq. (9.9). The per-event SF is calculated based on:

$$
\begin{aligned}
SF^{\pm} = & \prod_{i}^{N_{\text{bc tagged}}} \left(f_i^{\text{tag}} \pm \Delta^{\pm} f_i^{\text{tag}}\right) \\
& \times \prod_{j}^{N_{\text{bc not tagged}}} \left[\left(\frac{1 - f_j^{\text{tag}}\varepsilon_j^{\text{bc}}}{1 - \varepsilon_j^{\text{bc}}}\right) \pm \sqrt{\left(\frac{-\varepsilon_j^{\text{bc}}}{1 - \varepsilon_j^{\text{bc}}}\Delta^{\pm} f_j^{\text{tag}}\right)^2 + \left(\frac{1 - f_j^{\text{tag}}}{\left(1 - \varepsilon_j^{\text{bc}}\right)^2}\Delta^{\pm}\varepsilon_j^{\text{bc}}\right)^2}\right] \\
& \times \prod_{k}^{N_{\text{udsg tagged}}} \left(f_k^{\text{mistag}} \pm \Delta^{\pm} f_k^{\text{mistag}}\right) \\
& \times \prod_{l}^{N_{\text{udsg not tagged}}} \left[\left(\frac{1 - f_l^{\text{mistag}}\varepsilon_l^{\text{udsg}}}{1 - \varepsilon_l^{\text{udsg}}}\right) \pm \sqrt{\left(\frac{-\varepsilon_l^{\text{udsg}}}{1 - \varepsilon_l^{\text{udsg}}}\Delta^{\pm} f_l^{\text{mistag}}\right)^2 + \left(\frac{1 - f_l^{\text{mistag}}}{\left(1 - \varepsilon_l^{\text{udsg}}\right)^2}\Delta^{\pm}\varepsilon_l^{\text{udsg}}\right)^2}\right]
\end{aligned}
\tag{9.21}
$$

where per-jet SFs f^{tag} and f^{mistag} as well as per-jet SFs and efficiencies are uncorrelated. In the final fit, b tagging and mistagging uncertainties are treated as two separate sources of shape and rate uncertainties.

9.6.8 Resolved top tagging and mistagging uncertainties

Systematic uncertainties related to the resolved top tagging and mistagging efficiencies depend on the per-t$^{\text{res}}$ SF uncertainties described in Section 7.6. The uncertainty on the per-event SF is calculated by applying error propagation on each term in Eq. (9.12). The tagging and mistagging uncertainties are treated as two separate sources of systematic uncertainties and they affect both the shape and rate of the final discriminant.

9.6.9 QCD multijet measurement uncertainties

Various sources of systematic uncertainties that affect the QCD multijet background measurement have been studied. Differences in the predicted background are split into uncertainties in the transfer factors (TF) and uncertainties that arise by varying the definition of the control regions (CRs), the purity of the QCD multijet background in the CRs and the correction of the transverse momentum of the associated t^{res} candidate. The impact on the analysis of each source of uncertainty, is estimated by the differences in the yields and shape of the fit discriminant.

Systematic uncertainties on the TF are calculated with error propagation on the $N_{QCD,i}^{CR(on-m,!t)}/N_{QCD,i}^{CR(off-m,!t)}$ ratio. The errors are determined by the statistical uncertainties in each region. Variations in the TF result in both rate and shape differences.

Uncertainties arising due to the definition of the CRs are estimated by varying the selection requirements of the associated t^{res} candidate's mass and the t^{res} tagging score of the t^{res} candidate from the $H^{\pm}$ side in the CRs. In the nominal sideband (SB) regions from either side of the $m_{t^{res}}$ region, the two mass SBs are defined by requiring that $m_{t^{res}} \in [50, 130] \cap [210, 260]$ GeV. Variations in the t^{res} candidate's mass are measured by narrowing the SBs, either by excluding the edges or using a "buffer region" between the selections of the SR and the SB. The alternative SBs are determined by $m_{t^{res}} \in [100, 130] \cap [210, 240]$ GeV and $m_{t^{res}} \in [50, 110] \cap [230, 260]$ GeV. Variations from the selection of the anti-tagged t^{res} candidates are calculated by shrinking the "!t" CRs. This is done by selecting events where the t^{res} candidate from the $H^{\pm}$ decay fails the loose, very loose (vloose) or very very loose (vvloose) WP of the t^{res} tagger, with the last two WPs being selected arbitrarily for the purpose of these studies. The dependence of the purity of the QCD multijet background in the CRs is quantified by altering the rate of the subtracted background from simulated events by $\pm$ 10%. Finally, the impact of the associated t^{res} candidate's p_T corrections is quantified by performing the data-driven method without binning the estimated background in p_T regions.

For each of the four aforementioned sources, the score of the parameterized DNN is predicted for 17 $m_{H^{\pm}}$ hypotheses that correspond to the signal mass points used in this analysis. The variation in the shape of the fit discriminant due to the sources of systematic uncertainties are illustrated in Fig. 9.12 ($1M1L_{t^{res}}$) and Fig. 9.13 ($2M_{t^{res}}$) which show the predicted QCD multijet background for a mass hypothesis of $m_{H^{\pm}}$ = 800 GeV using the 2018 data.

The nominal-over-varied ratio of the parameterized DNN score is fitted to a zero-order polynomial for all mass hypotheses and sources. The flat shape of the ratios implies rate uncertainties for the four sources of systematic variations. The impact of each source is taken as a symmetric variation equal to the maximum variation with respect to the nominal selections. The total systematic uncertainty affects the rate of the fit discriminant and includes all the sources which are summed in quadrature.

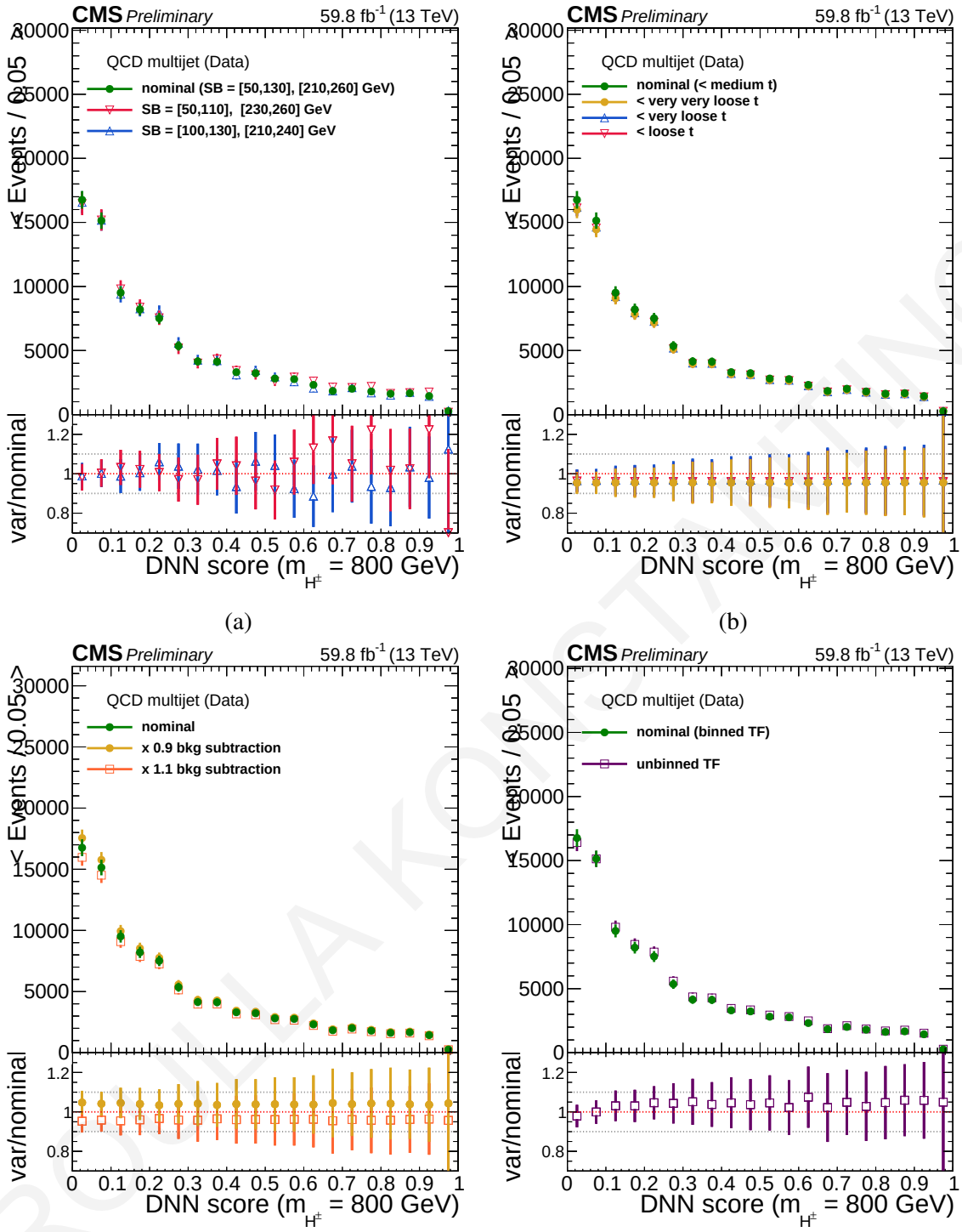


Figure 9.12: Sources of systematic uncertainties of the estimated QCD multijet background for the $1\mathrm{M1L}_{\mathrm{t}^{\mathrm{res}}}$ category. The effect of (a) the selection of the associated $\mathrm{t}^{\mathrm{res}}$ candidate's mass, (b) the $\mathrm{t}^{\mathrm{res}}$ tagging score of the $\mathrm{t}^{\mathrm{res}}$ candidate from the $\mathrm{H}^{\pm}$ decay, (c) the purity of the QCD multijet background in the CRs and (d) the correction of the transverse momentum of the associated $\mathrm{t}^{\mathrm{res}}$ candidate. The predicted background is expressed as a function of the parameterized DNN score for $m_{\mathrm{H}^{\pm}} = 800$ GeV, using the 2018 data.

9.6.10 Theoretical uncertainties

Theoretical uncertainties on the production cross section of various processes are treated as rate uncertainties. They originate from factorization and renormalization scale uncertainties

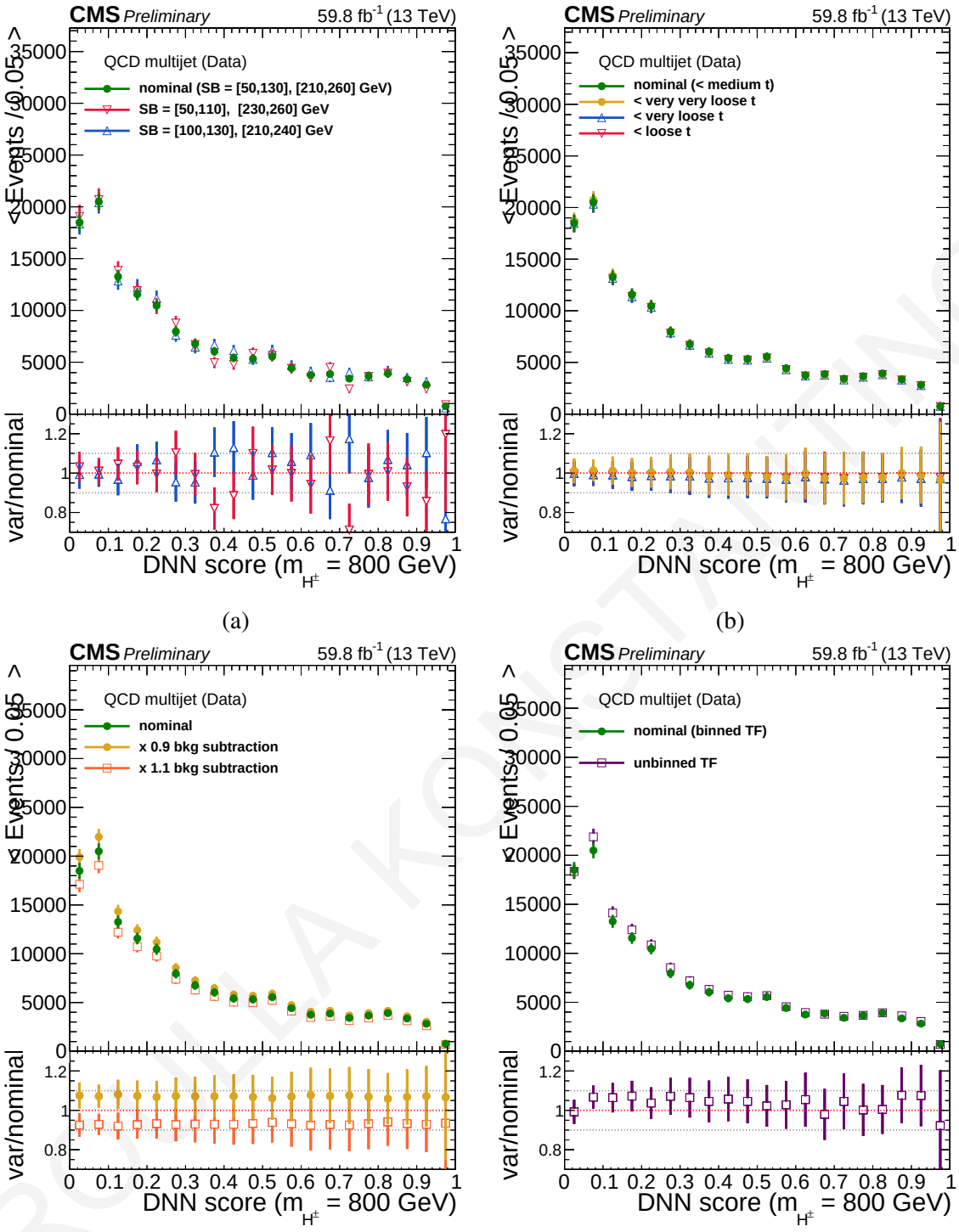


Figure 9.13: Sources of systematic uncertainties of the estimated QCD multijet background for the $2\mathrm{M}_{\mathrm{t}^{\mathrm{res}}}$ category. The effect of (a) the selection of the associated $\mathrm{t}^{\mathrm{res}}$ candidate's mass, (b) the $\mathrm{t}^{\mathrm{res}}$ tagging score of the $\mathrm{t}^{\mathrm{res}}$ candidate from the $\mathrm{H}^{\pm}$ decay, (c) the purity of the QCD multijet background in the CRs and (d) the correction of the transverse momentum of the associated $\mathrm{t}^{\mathrm{res}}$ candidate. The predicted background is expressed as a function of the parameterized DNN score for $m_{\mathrm{H}^{\pm}}$ = 800 GeV, using the 2018 data.

due to missing higher-order QCD corrections, uncertainties in the parton distribution functions (PDF) and the strong coupling constant α_S. Single t and $\mathrm{t\bar{t}}$ production cross sections

are also affected by the top-quark mass. Systematic effects on the acceptance of the events due to the renormalization and factorization scales are evaluated by varying the scales independently by a factor of two, excluding extreme cases in which they differ by more than a factor of two from each other. The scale uncertainty is defined by taking the envelope of the maximum variation with respect to the nominal fit distribution. Similarly, the uncertainty on the acceptance due to the choice of PDF sets is taken into account. Cross section uncertainties are treated as correlated between different data-taking eras.

A summary of the systematic uncertainties are shown in Table 9.5 for the $1\mathrm{M}1\mathrm{L}_{\mathrm{t}^{\mathrm{res}}}$ category and Table 9.6 for the $2\mathrm{M}_{\mathrm{t}^{\mathrm{res}}}$ category and for $m_{\mathrm{H}^{\pm}}$ = 800 GeV, using the 2018 data. The uncertainties that affect the shapes of signal and background are indicated with (S).

Table 9.5: Summary table of the systematic uncertainties for the $1\mathrm{M1L}_{\mathrm{t}^{\mathrm{res}}}$ category and $m_{\mathrm{H}^{\pm}}$ = 800 GeV, using the 2018 data.

Uncertainty Source ($1\mathrm{M1L}_{\mathrm{t}^{\mathrm{res}}}$)	$m_{\mathrm{H}^{\pm}}$ = 800 GeV	QCD multijet	$\mathrm{t\bar{t}}$	Single t, $\mathrm{t\bar{t}}$ +X, EWK
luminosity (2018)	2.5	–	2.5	2.5
luminosity correlated (2018)	2.0	–	2.0	2.0
muon veto eff.	< 0.1	–	< 0.1	< 0.1
electron veto eff.	< 0.1	–	< 0.1	< 0.1
tau veto eff.	0.2	–	0.2	< 0.1
QCD background	–	5.4	–	–
pileup reweighting (S)	0.3	0.2	0.4	0.3
jet energy scale (S)	7.5	4.9	12	8.9
jet energy resolution (S)	4.9	0.4	3.9	1.9
trigger efficiency (S)	1.3	0.8	2.2	2.1
b tagging (S)	6.5	1.2	2.9	3.7
b mistagging (S)	4.2	3.6	9.5	8.3
t tagging (S)	9.8	< 0.1	9.3	5.4
t mistagging (S)	1.7	< 0.1	1.9	2.6
QCD background TF0 (S)	–	2.0	–	–
QCD background TF1 (S)	–	1.2	–	–
QCD background TF2	–	–	–	–
QCD scale ($\mathrm{t\bar{t}}$)	–	–	$^{+2.4}_{-3.5}$	–
PDF ($\mathrm{t\bar{t}}$)	–	–	4.2	–
top-quark mass ($\mathrm{t\bar{t}}$)	–	–	–	–
QCD scale (single t)	–	–	–	2.5
PDF (single t)	–	–	–	4.7
top-quark mass (single t)	–	–	–	2.2
QCD scale (EW)	–	–	–	3.2
PDF (EW)	–	–	–	4.4
Theoretical α_S ($\mathrm{t\,\bar{t}\,H}$)	–	–	–	1.0
QCD scale acceptance ($\mathrm{H}^{\pm}$)	$^{+6.7}_{-4.4}$	–	–	–
PDF acceptance ($\mathrm{H}^{\pm}$)	0.1	–	–	–
QCD scale acceptance ($\mathrm{t\bar{t}}$)	–	–	3.2	3.2
PDF acceptance ($\mathrm{t\bar{t}}$)	–	–	0.1	0.1

Table 9.6: Summary table of the systematic uncertainties for the $2M_{t^{res}}$ category and $m_{H^\pm}$ = 800 GeV, using the 2018 data.

Uncertainty Source ($2M_{t^{res}}$)	$m_{H^\pm}$ = 800 GeV	QCD multijet	$t\bar{t}$	Single t, $t\bar{t}$ +X, EWK
luminosity (2018)	2.5	–	2.5	2.5
luminosity correlated (2018)	2.0	–	2.0	2.0
muon veto eff.	< 0.1	–	< 0.1	< 0.1
electron veto eff.	< 0.1	–	< 0.1	< 0.1
tau veto eff.	0.2	–	0.2	< 0.1
QCD background	–	9.6	–	–
pileup reweighting (S)	< 0.1	0.3	0.4	0.3
jet energy scale (S)	6.8	7.4	11	11
jet energy resolution (S)	13	0.2	7.5	5.7
trigger efficiency (S)	1.1	1.0	1.6	1.6
b tagging (S)	6.4	1.9	2.9	4.0
b mistagging (S)	3.9	5.4	9.4	7.6
t tagging (S)	10	< 0.1	11	6.8
t mistagging (S)	3.7	< 0.1	4.0	5.2
QCD background TF0 (S)	–	0.6	–	–
QCD background TF1 (S)	–	1.7	–	–
QCD background TF2 (S)	–	0.6	–	–
QCD scale ($t\bar{t}$)	–	–	$^{+2.4}_{-3.5}$	–
PDF ($t\bar{t}$)	–	–	4.2	–
top-quark mass ($t\bar{t}$)	–	–	–	–
QCD scale (single t)	–	–	–	2.5
PDF (single t)	–	–	–	4.7
top-quark mass (single t)	–	–	–	2.2
QCD scale (EW)	–	–	–	3.2
PDF (EW)	–	–	–	4.4
Theoretical α_S ($t\,\bar{t}\,H$)	–	–	–	1.0
QCD scale acceptance ($H^\pm$)	$^{+4.9}_{-3.5}$	–	–	–
PDF acceptance ($H^\pm$)	0.1	–	–	–
QCD scale acceptance ($t\bar{t}$)	–	–	2.1	2.1
PDF acceptance ($t\bar{t}$)	–	–	0.1	0.1

9.7 Results

Figures 9.14 to 9.17 and Figs. 9.18 to 9.21 display the pre-fit distributions of the final fit discriminant, the score of the parameterized DNN, for the $2M_t^{res}$ and $1M1L_t^{res}$ regions respectively and for nine mass hypotheses. The expected signal overlaid corresponds to the true mass hypothesis used to evaluate the discriminant. The observed data is blinded in the high-DNN score region where the signal significance increases. In the low-DNN score data and simulation are in agreement within statistical and systematic uncertainties.

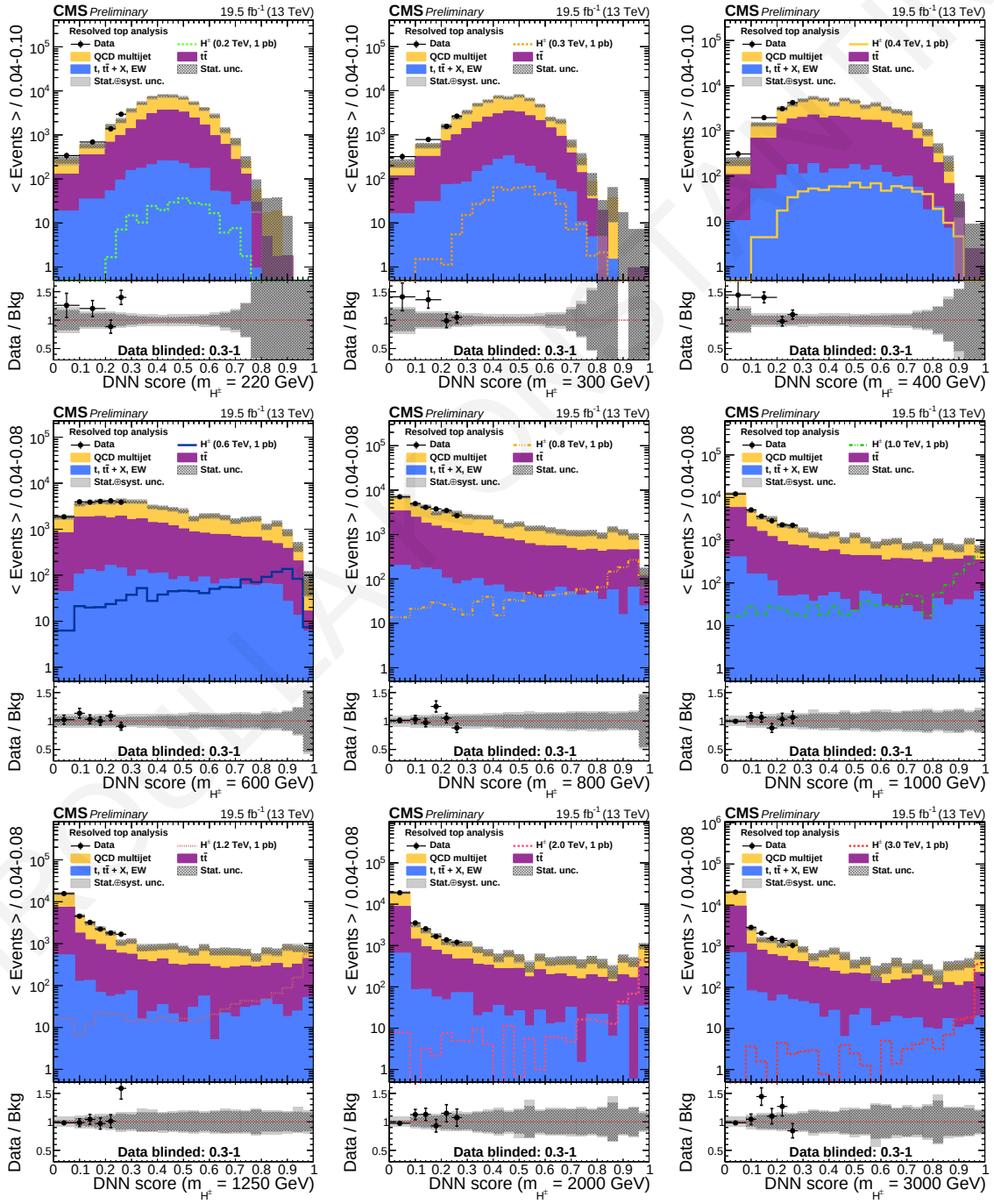


Figure 9.14: The output score of the parameterized DNN in the $1M1L_t^{res}$ category with the 2016APV data.

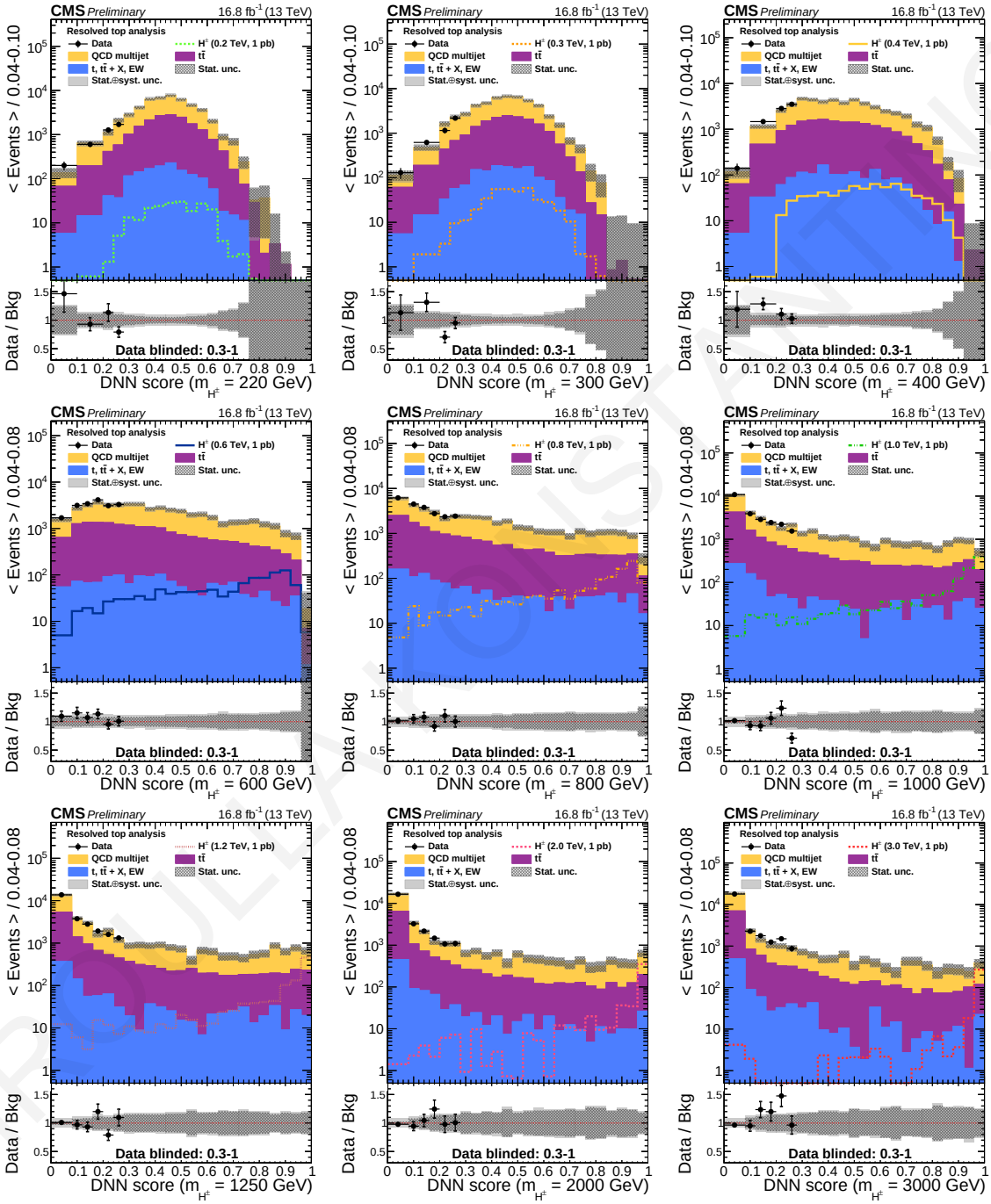


Figure 9.15: The output score of the parameterized DNN in the $1\mathrm{M}1\mathrm{L}_{\mathrm{t}}^{\mathrm{res}}$ category with the 2016 non-APV data.

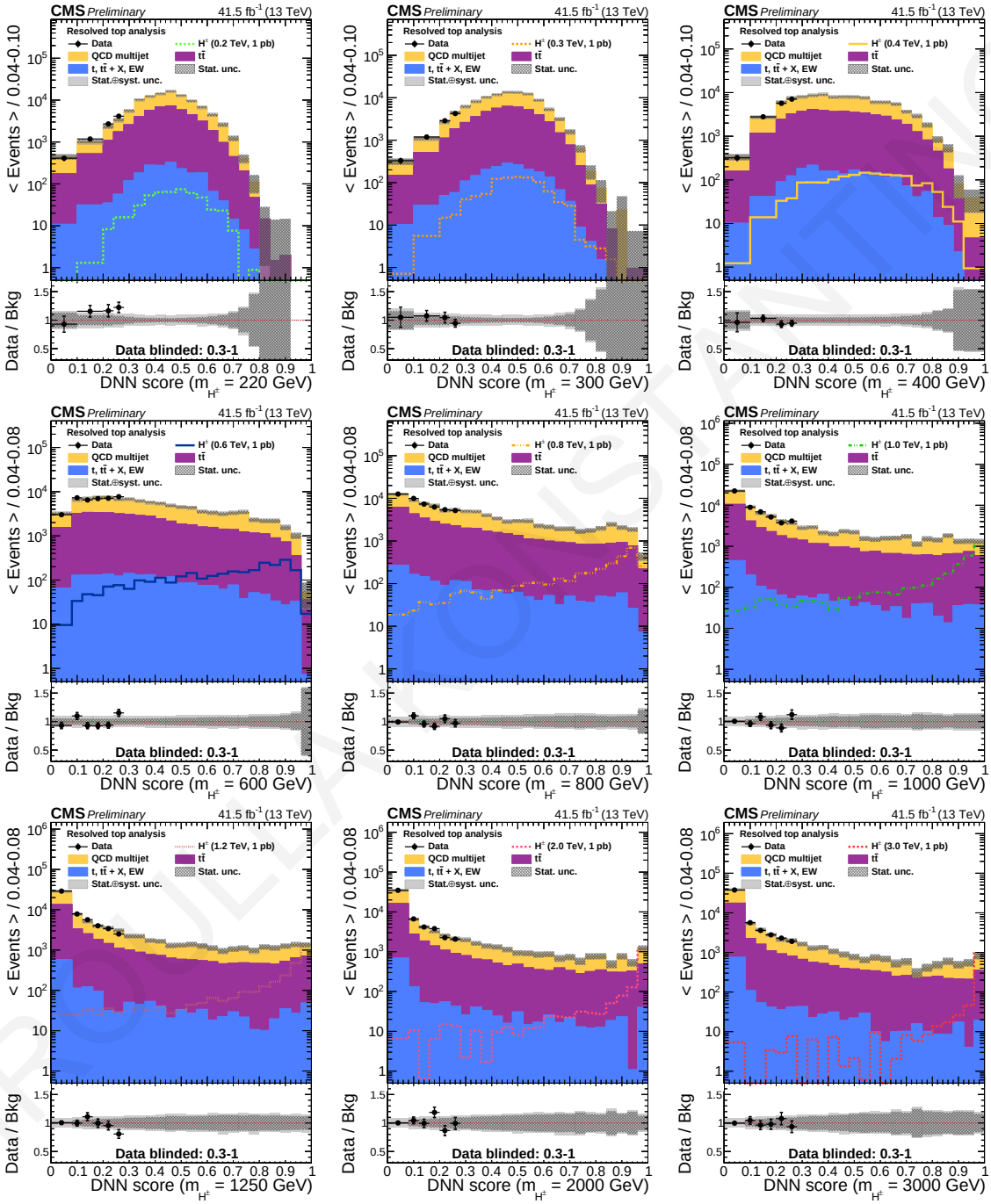


Figure 9.16: The output score of the parameterized DNN in the $1M1L_t^{res}$ category with the 2017 data.

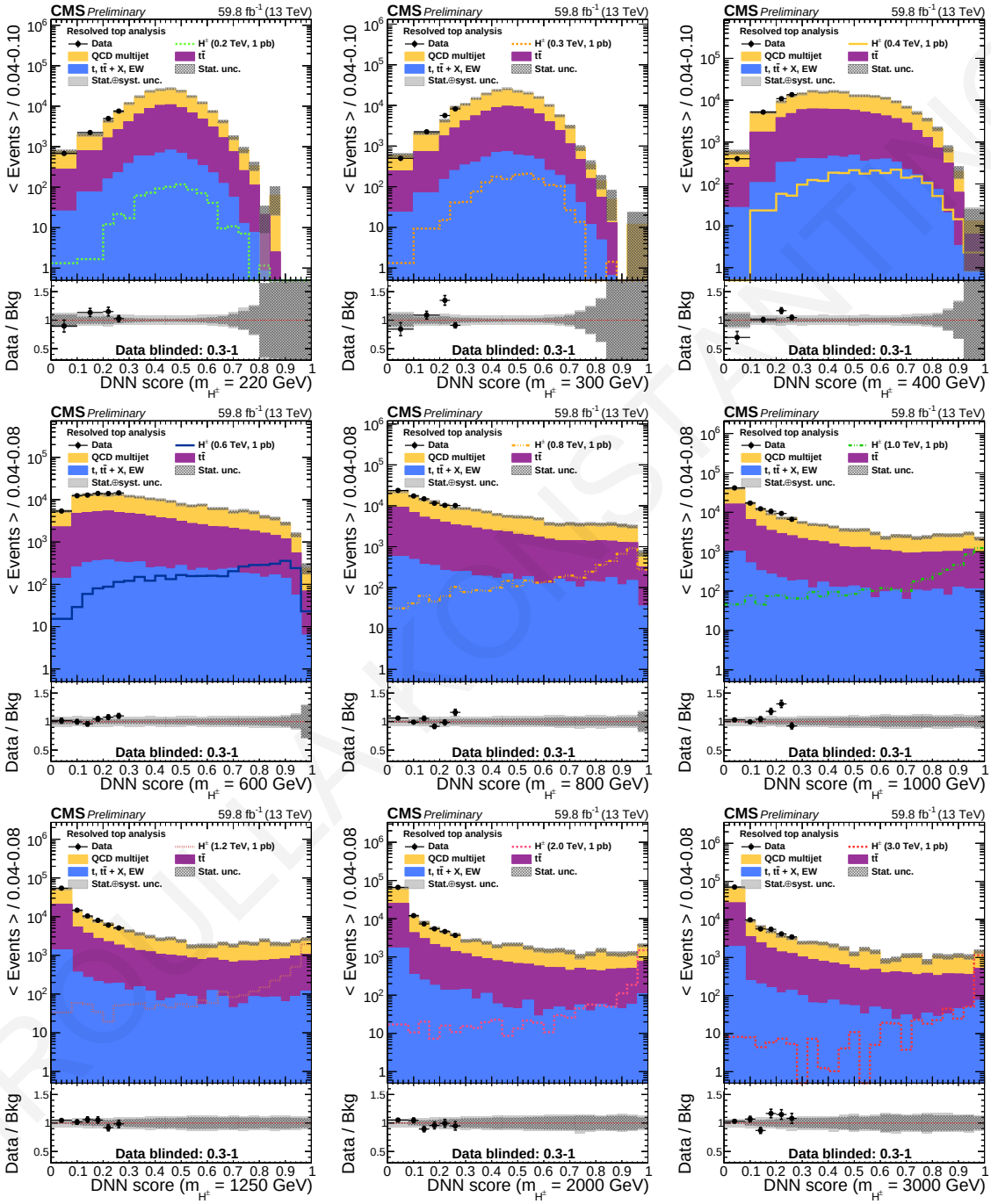


Figure 9.17: The output score of the parameterized DNN in the $1M1L_t^{res}$ category with the 2018 data.

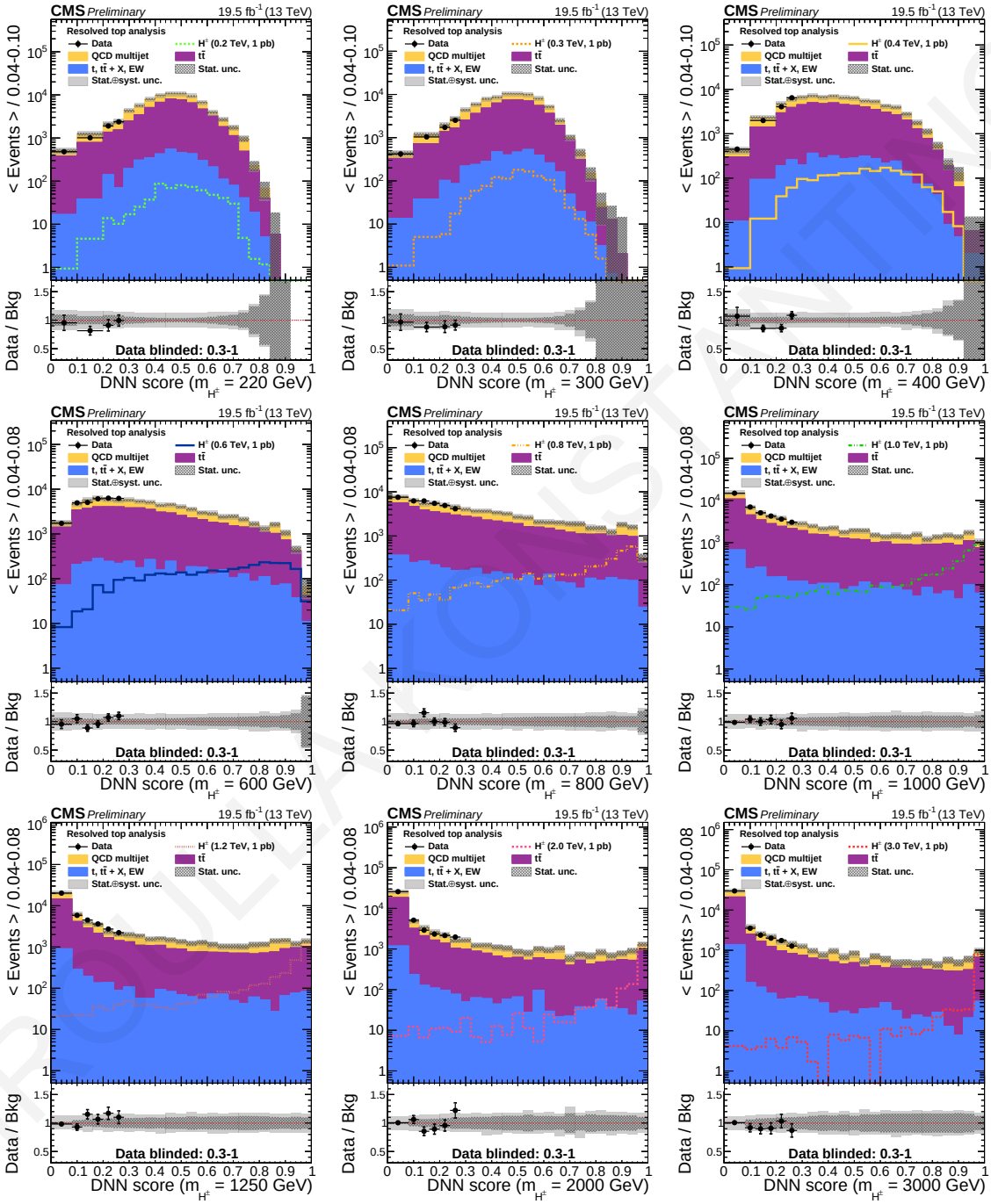


Figure 9.18: The output score of the parameterized DNN in the $2M_t^{res}$ category with the 2016APV data.

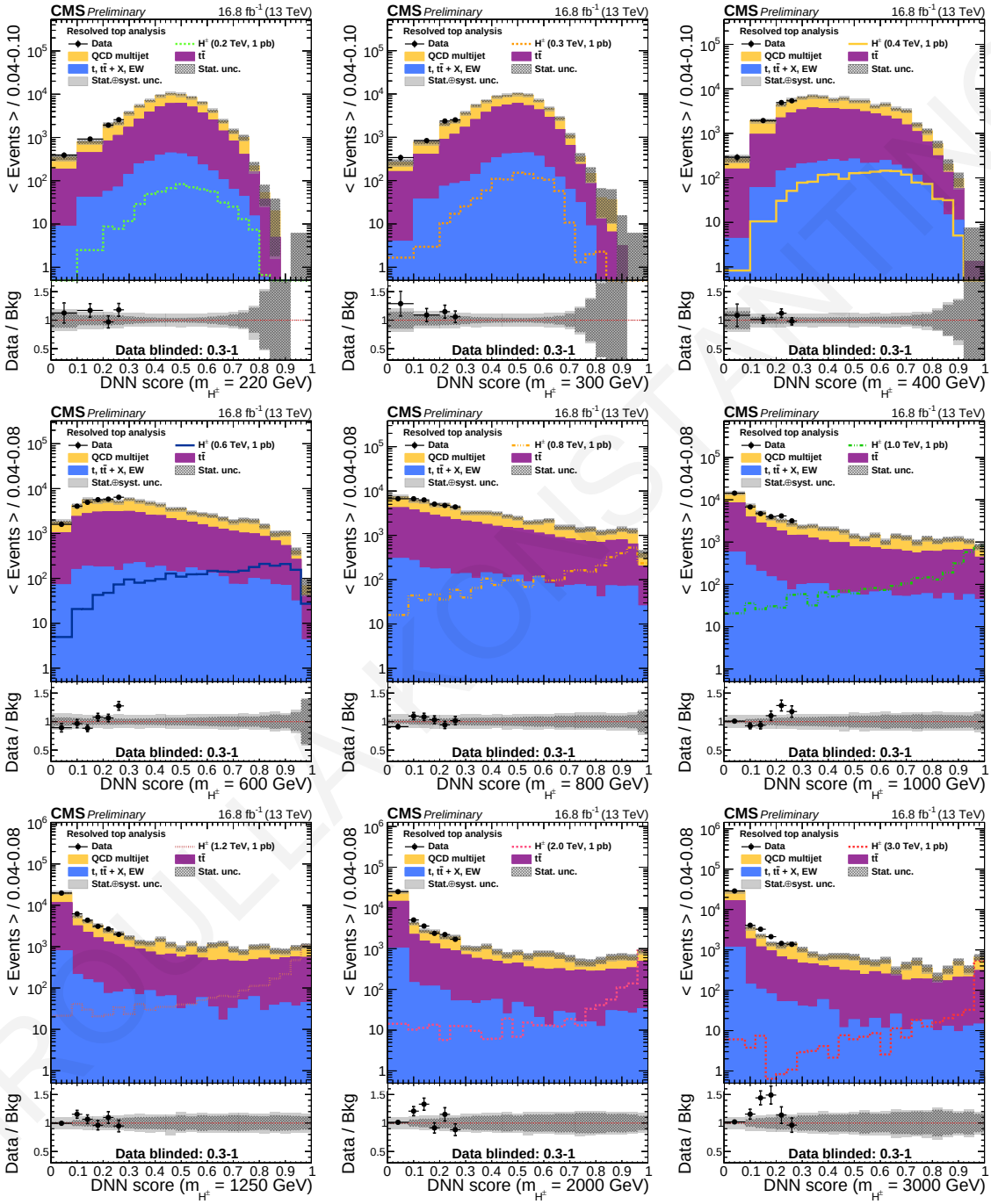


Figure 9.19: The output score of the parameterized DNN in the $2M_t^{res}$ category with the 2016 non-APV data.

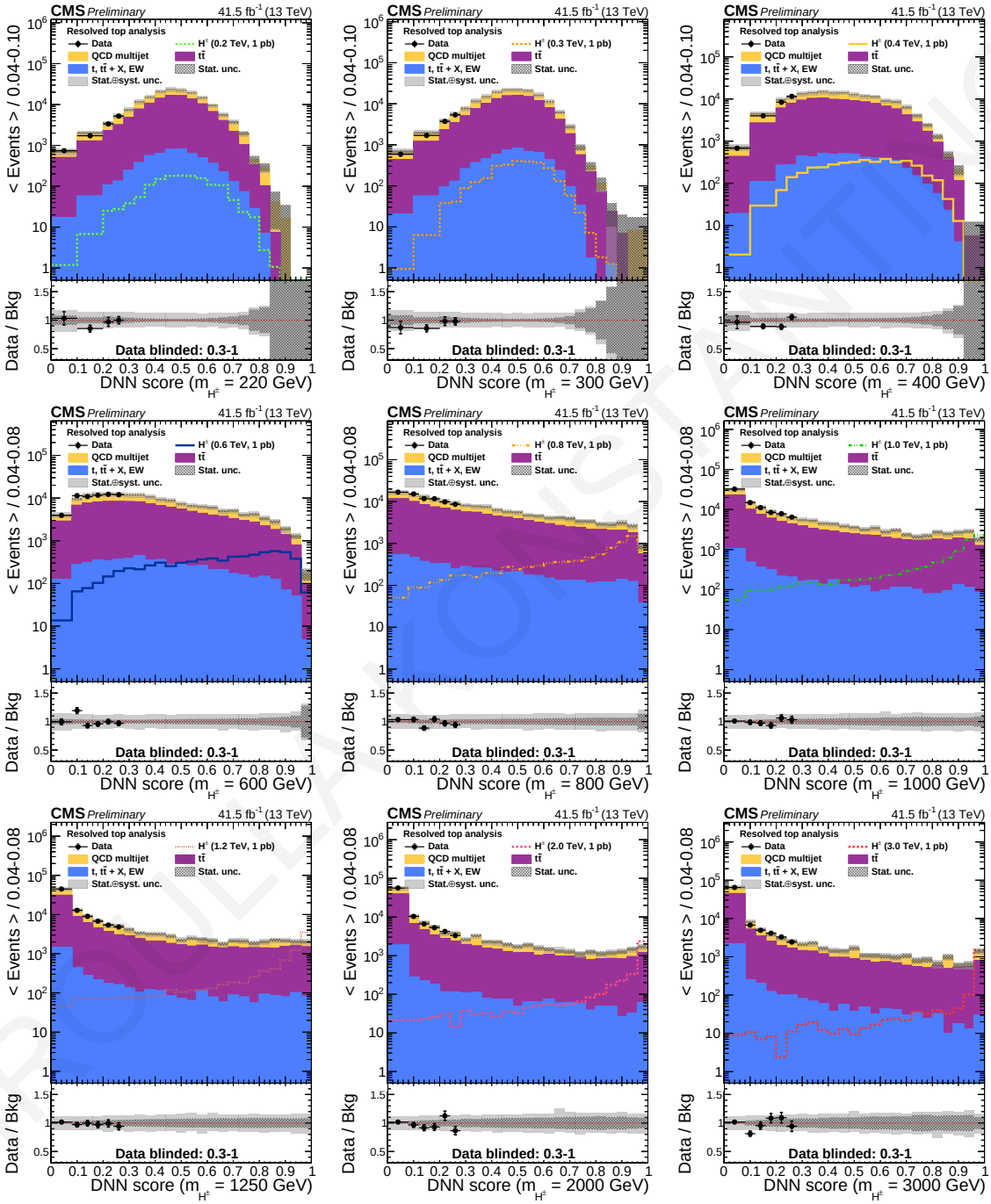


Figure 9.20: The output score of the parameterized DNN in the $2M_t^{res}$ category with the 2017 data.

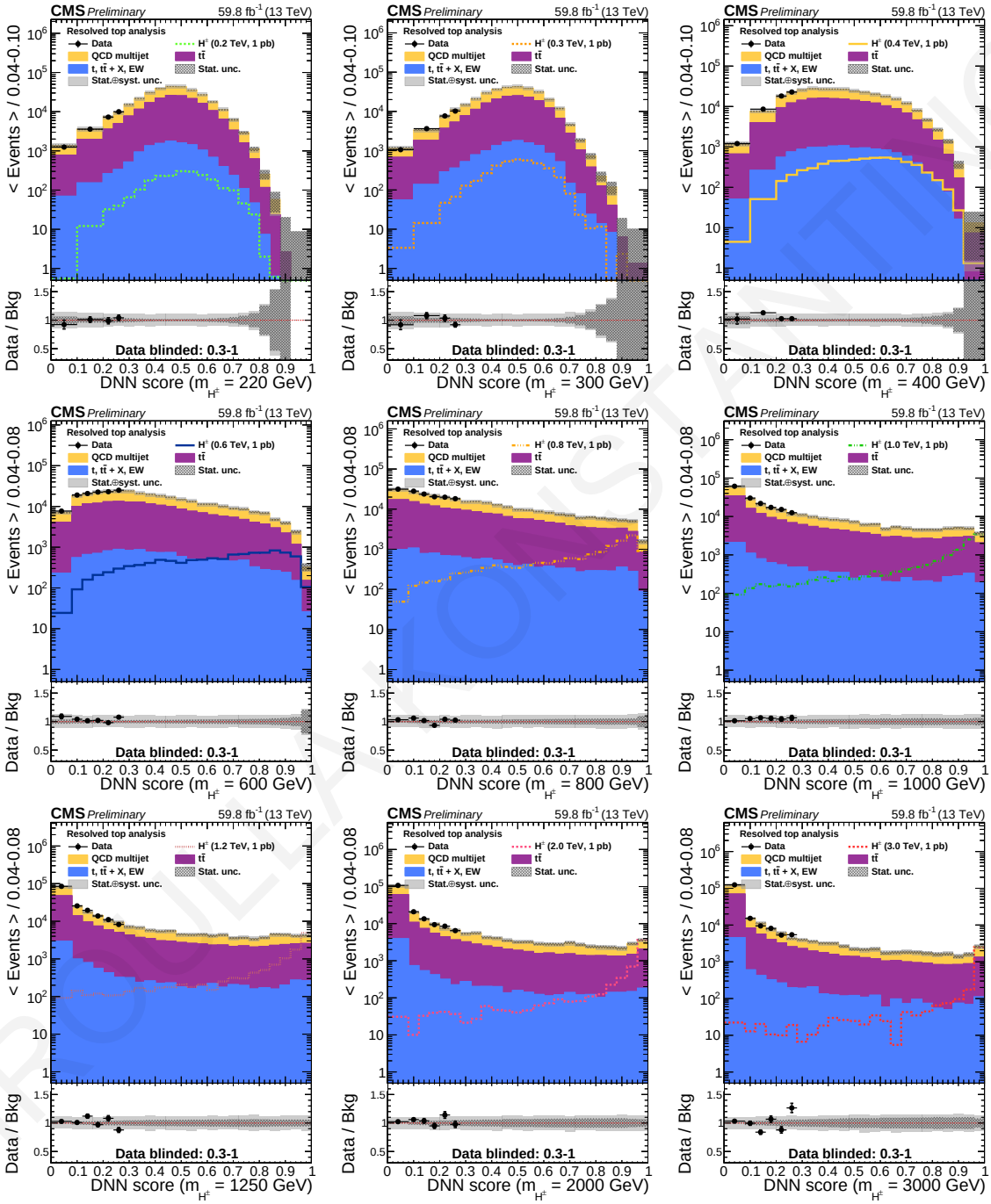


Figure 9.21: The output score of the parameterized DNN in the $2M_t^{res}$ category with the 2018 data.

9.7.1 Expected limits

Expected upper limits are set on the $\sigma(pp \to tbH^{\pm}) \times \mathcal{B}(H^{\pm} \to tb)$ at 95% CL for each category and data-taking period. The combined median expected limit and the limits from each contributing category and era are illustrated in Fig. 9.22. The $2M_{t^{res}}$ shows best sensitivity in the entire $m_{H^{\pm}}$ region. Figure 9.23 shows the combined median expected upper limits (dashed line), and the 68% (inner green band) and 95% (outer yellow band) confidence interval for the entire Run II and both categories. For comparison, the limits extracted from each data-taking period which combine the $1M1L_{t^{res}}$ and $2M_{t^{res}}$ categories are shown with solid lines. The analysis exhibits good sensitivity for $m_{H^{\pm}}$ values up to 1250 GeV, while beyond that value, the $H^{\pm}$ decay products become increasingly boosted and the limit comes to a plateau. The numerical values of the combined expected exclusion limits are listed in Table 9.7.

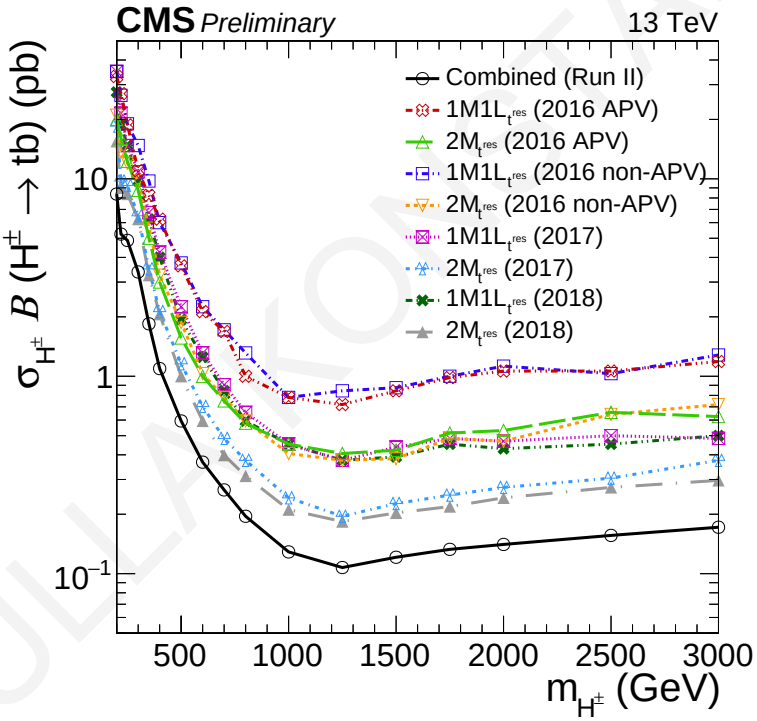


Figure 9.22: Median expected limit on the $\sigma(pp \to tbH^{\pm}) \times \mathcal{B}(H^{\pm} \to tb)$ for each data-taking period and category, and their combination for the entire Run II data taking period.

A comparison between the expected limit derived from the 2016 data and the expected limit of the resolved analysis published by the CMS collaboration (HIG-18-015) using the same data [53] is shown in Fig. 9.24. The results demonstrate significant improvement, which can be attributed to various factors. One of the primary sources of improvement is the use of the DeepJet b tagging algorithm, which exhibits superior performance compared to the CSVv2 algorithm used in the previous analysis. Furthermore, the BDT-based t^{res} tagging algorithm has been replaced by the DNN-based mass decorrelated t^{res} tagger which possesses better discrimination power when selecting t^{res} candidates with a mass close to the top-quark mass. The decorrelation of the t^{res} tagger's score from the candidate's mass

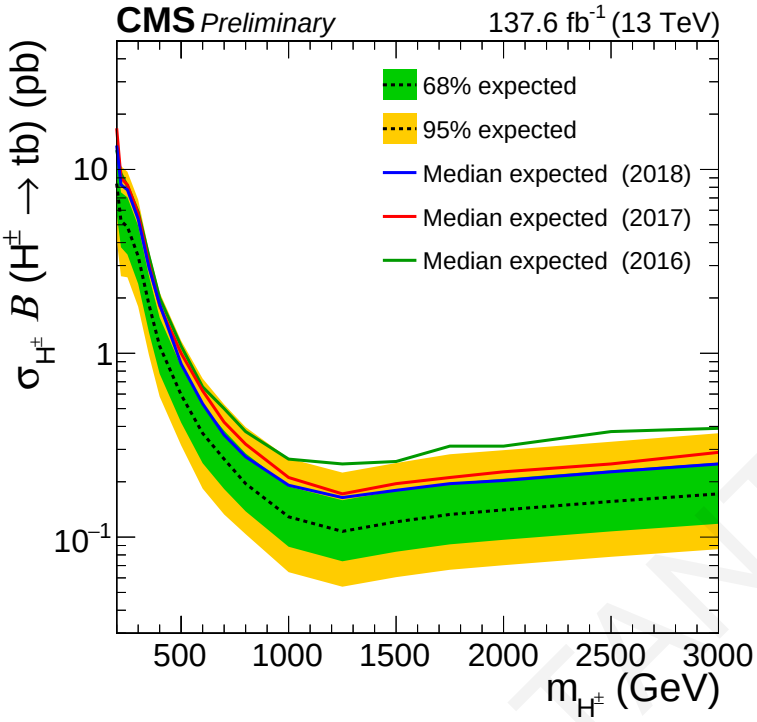


Figure 9.23: Expected combined upper limits on the $\sigma(\mathrm{pp} \to \mathrm{tbH}^{\pm}) \times \mathcal{B}(\mathrm{H}^{\pm} \to \mathrm{tb})$. The dotted black line shows the median expected limit of the entire Run II data and both categories. The bands correspond to 68% (green) and 95% (yellow) confidence intervals. The solid lines show the median expected limits obtained with the 2016 (green), 2017 (red) and 2018 (blue) data for both categories.

allows for the utilization of mass sidebands, which are used to estimate the QCD multijet background through a data-driven method. This new approach provides an improved agreement between the data and background compared to the previous method, which estimates the background events with misidentified b-tagged jets. Additionally, the signal extraction is performed using a mass-parameterized multivariate discriminant that exploits information from 13 event features, including the reconstructed mass of the $\mathrm{H}^{\pm}$ used as a fit discriminant in the published analysis. Finally, the sensitivity is enhanced with the addition of a second category, the $1\mathrm{M}1\mathrm{L}_{\mathrm{t}^{\mathrm{res}}}$.

Table 9.7: Upper limit at 95% CL on $\sigma(\mathrm{pp} \to \mathrm{tbH}^{\pm}) \times \mathcal{B}(\mathrm{H}^{\pm} \to \mathrm{tb})$ for the entire Run II data.

$m_{\mathrm{H}^{\pm}}$	Expected limit					Observed
(GeV)	-2σ	-1σ	median	$+1\sigma$	$+2\sigma$	limit
200	4.18750	5.75781	8.37500	12.38097	17.08382	Blinded
220	2.62500	3.77344	5.25000	7.51009	10.41629	Blinded
250	2.58984	3.44678	4.87500	7.05138	9.72327	Blinded
300	1.79297	2.38623	3.37500	4.93554	6.85005	Blinded
350	0.97949	1.30359	1.84375	2.63747	3.61171	Blinded
400	0.58105	0.77332	1.09375	1.56460	2.14254	Blinded
500	0.31543	0.41980	0.59375	0.84935	1.17056	Blinded
600	0.18359	0.25244	0.36719	0.52526	0.72852	Blinded
700	0.13281	0.18262	0.26562	0.37997	0.52701	Blinded
800	0.10376	0.13809	0.19531	0.28562	0.39641	Blinded
1000	0.06445	0.08862	0.12891	0.18851	0.26799	Blinded
1250	0.05371	0.07385	0.10742	0.15880	0.22437	Blinded
1500	0.06055	0.08325	0.12109	0.17902	0.25293	Blinded
1750	0.06641	0.09131	0.13281	0.19634	0.28227	Blinded
2000	0.07031	0.09668	0.14062	0.20789	0.29716	Blinded
2500	0.07812	0.10742	0.15625	0.23099	0.33017	Blinded
3000	0.08594	0.11816	0.17188	0.25409	0.36739	Blinded

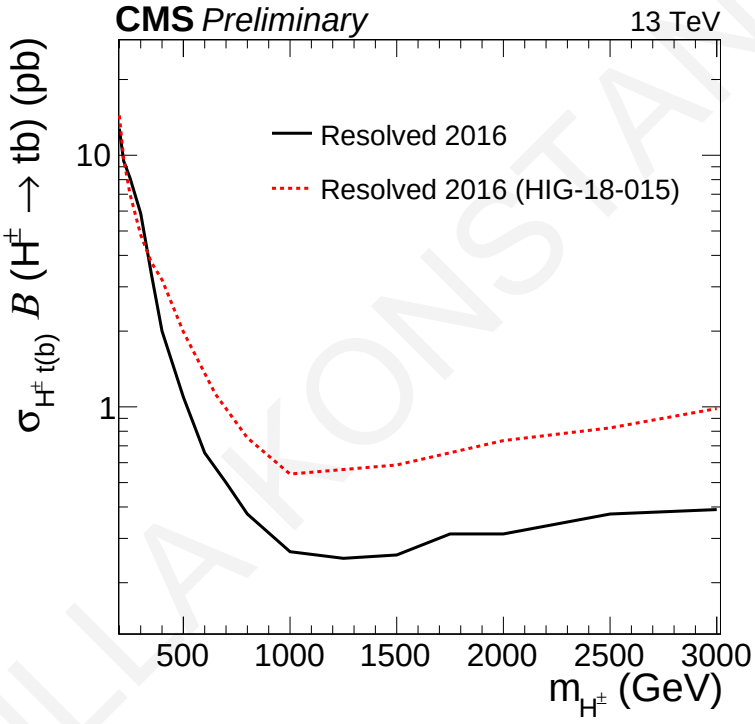


Figure 9.24: Comparison of the median expected limit on the $\sigma(\mathrm{pp} \to \mathrm{tbH}^{\pm}) \times \mathcal{B}(\mathrm{H}^{\pm} \to \mathrm{tb})$ using the 2016 data and the CMS published median expected limit [53] of the resolved analysis.

9.7.2 Impacts of the nuisance parameters

The impacts and pulls of the most significant nuisance parameters on the $\sigma(pp \to tbH^{\pm}) \times \mathcal{B}(H^{\pm} \to tb)$ for $m_{H^{\pm}} = 800\,\text{GeV}$ are shown in Figs. 9.25 and 9.26 for the $1M1L_{t^{res}}$ and $2M_{t^{res}}$ respectively, using the 2018 data. The impact of the nuisance parameters are defined by the shift $\Delta\hat{r}$ on the parameter of interest $\hat{r}$ with respect to the best fit due to $\pm\,1\,\sigma$ shifts in the value of a given nuisance parameter θ. The pulls refer to the effect that the nuisance parameters can have on the parameter $\hat{r}$ and if this is biased towards one direction. The nuisance parameters show no significant bias from the nominal value and their impact does not overconstrain the $\hat{r}$.

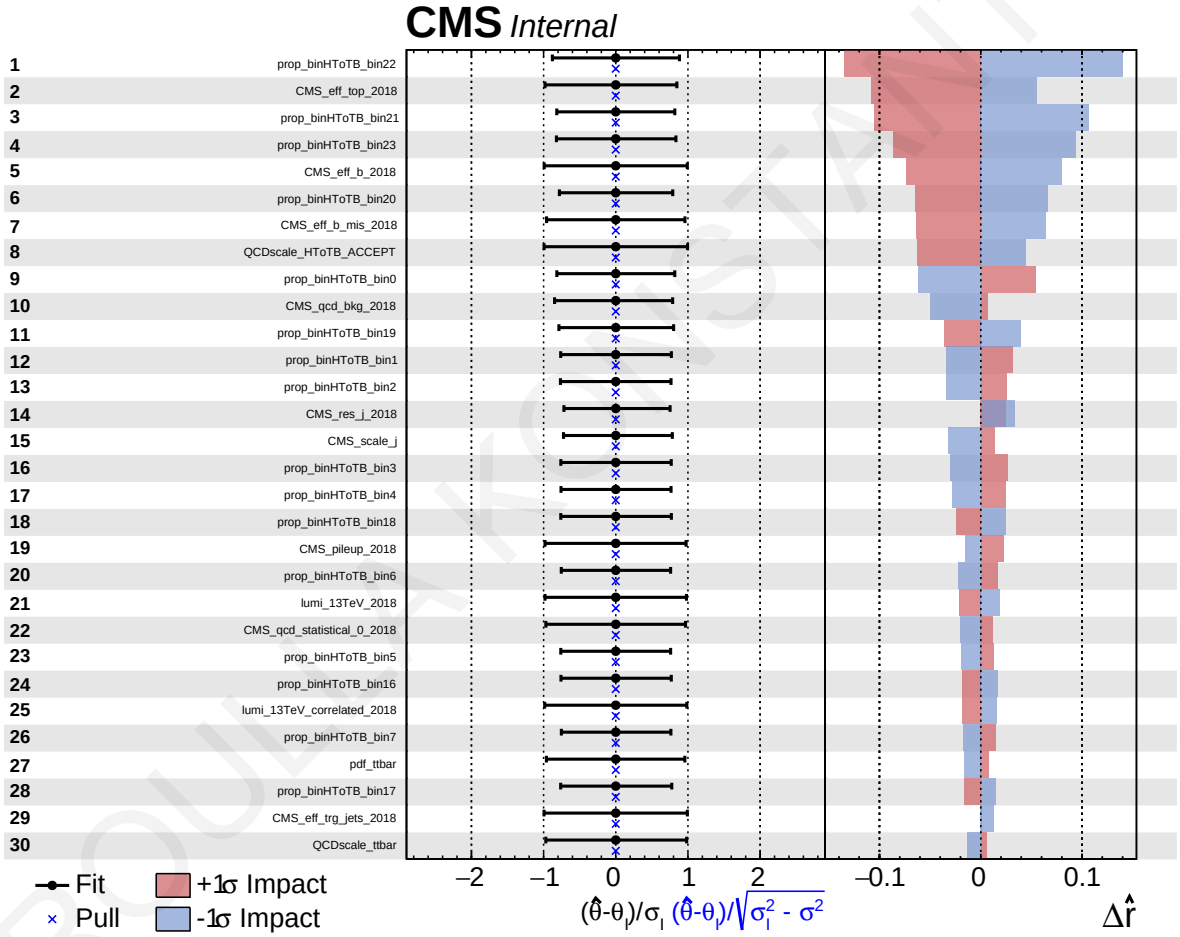


Figure 9.25: Pulls (left column) and impacts (right column) for the $1M1L_{t^{res}}$ category, using 2018 data and for the $m_{H^{\pm}} = 800\,\text{GeV}$ mass hypothesis.

9.7.3 Goodness of fit

A goodness-of-fit test is performed using the saturated method and a sample of 5000 toy events. The results are shown in Figs. 9.27 and 9.28 for the $1M1L_{t^{res}}$ and $2M_{t^{res}}$ categories respectively, and for different values of $m_{H^{\pm}}$, using the 2018 data. The values obtained from data are compatible with the expectation from the toy events. The worst tests give p-values

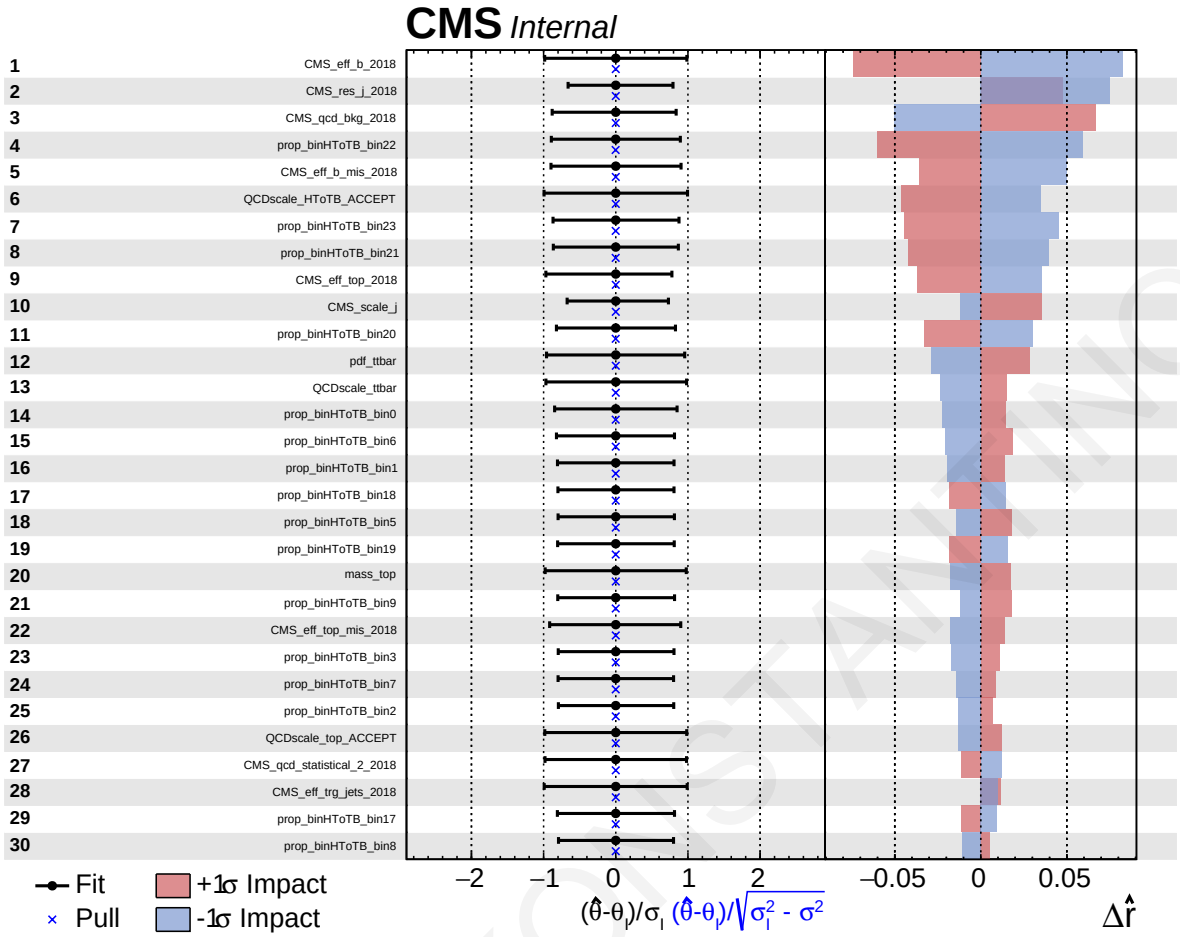


Figure 9.26: Pulls (left column) and impacts (right column) for the $2\mathrm{M}_{\mathrm{t}}^{\mathrm{res}}$ category, using 2018 data and for the $m_{\mathrm{H}^{\pm}}$ = 800 GeV mass hypothesis.

equal to 0.01 for $m_{\mathrm{H}^{\pm}}$ = 400 GeV in the $1\mathrm{M}1\mathrm{L}_{\mathrm{t}}^{\mathrm{res}}$ category and 0.33 for $m_{\mathrm{H}^{\pm}}$ = 500 GeV in the $2\mathrm{M}_{\mathrm{t}}^{\mathrm{res}}$ category.

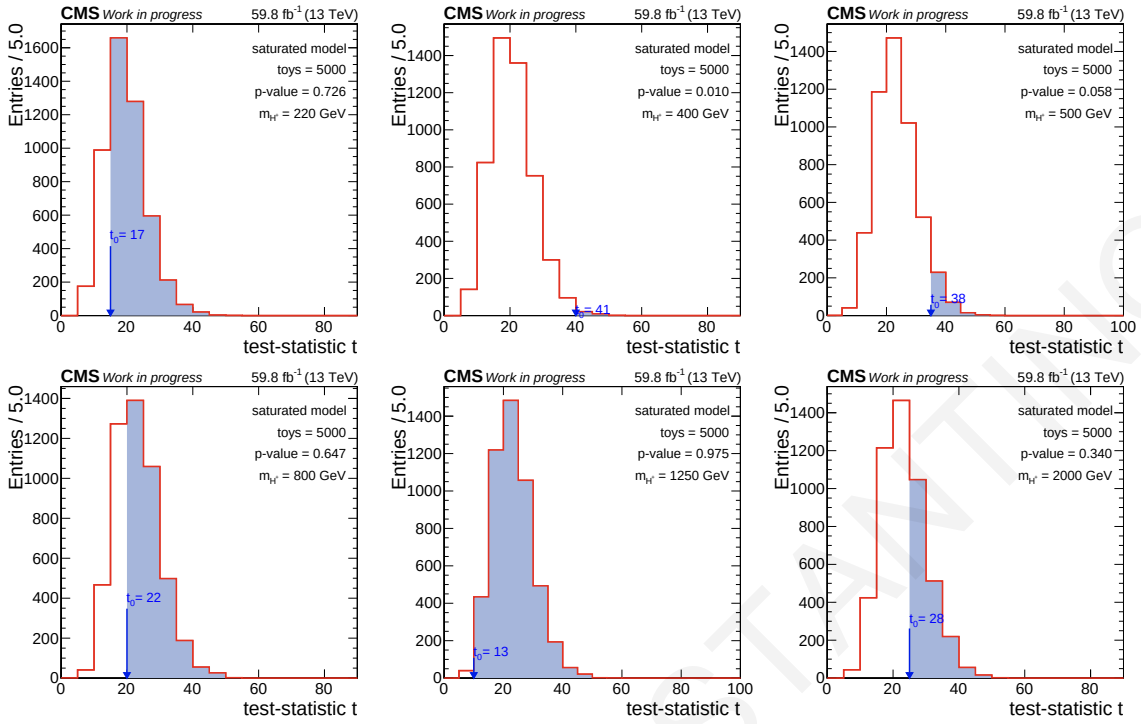


Figure 9.27: Goodness-of-fit test for different values of $m_{H^\pm}$ in the $1M1L_{t^{res}}$ category, using the 2018 data.

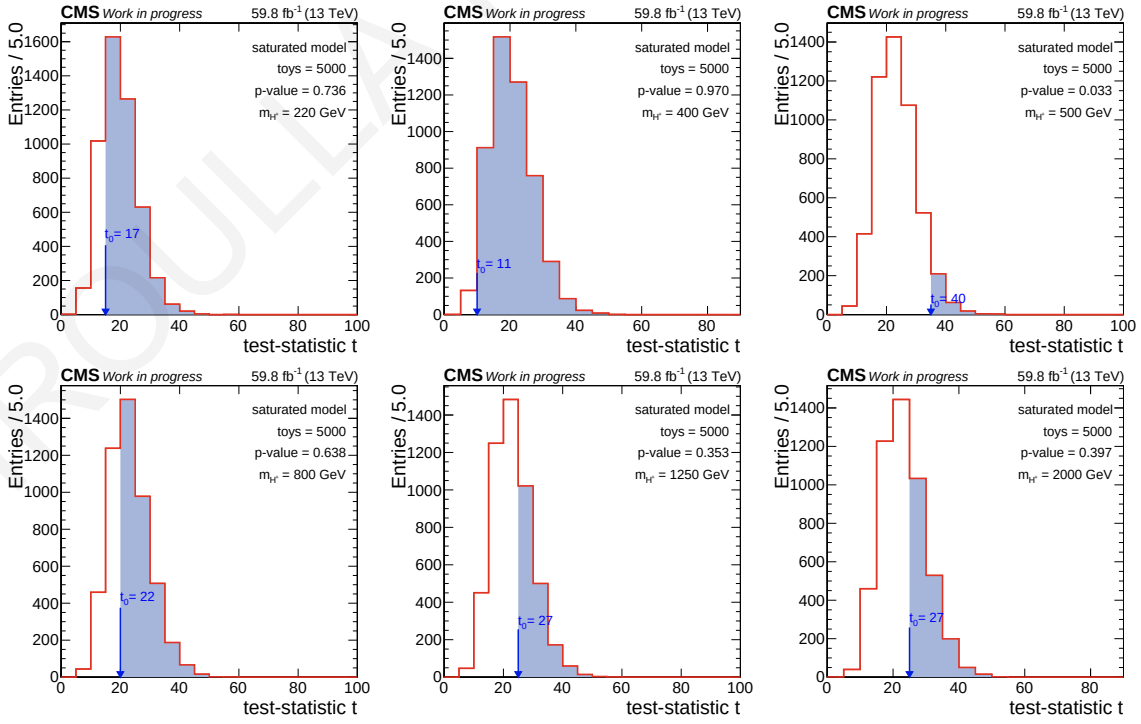


Figure 9.28: Goodness-of-fit test for different values of $m_{H^\pm}$ in the $2M_{t^{res}}$ category, using the 2018 data.

10 Search for charged Higgs bosons in the semi-boosted regime

A preliminary search for a $H^{\pm} \to tb$ in the all-jet final state, in the semi-boosted topology is presented in this section. The signal events contain exactly one Lorentz-boosted large-radius jet identified as a top-quark with the ParticleNet tagger [168] which is described in Section 10.1. The analysis searches for $H^{\pm}$ in the two-dimensional space of the boosted top-jet mass and the reconstructed $H^{\pm}$ invariant mass ($m_{t^{bst}}$, m_{tb}). The analysis developped is applied to the 2018 data and sets preliminary model-independent upper limits on the $\sigma(pp \to tbH^{\pm}) \times \mathcal{B}(H^{\pm} \to tb)$.

10.1 Boosted jet flavour identification

The flavor of Lorentz-boosted heavy particles is identified using MVA-based techniques. During Run II, the CMS Collaboration has done enormous progress in jet-flavor tagging. The first taggers utilized high-level features of the boosted jets using fully-connected DNNs. Later, more complex DNN architectures such as convolutional neural networks (CNN) were used to process low-level information of the jet substructure, including PF particles and secondary vertices.

The DeepAK8 multiclass tagger [169] combined convolutional and fully-connected layers. In the convolutional layers, a filter or kernel of given dimensions passes through the input sequences of the PF particles or SVs that are sorted in p_{T} or in the two-dimensional impact parameter significance respectively. The kernel convolves over a single feature x_i to exploit local information of the neighboring inputs. The convolved features have the form:

$$z_m = \sum_j k_{i,j} \cdot x_{i,m+j-1} \tag{10.1}$$

where j runs over the size of the kernel, i denotes the feature and m is the index of the convolved feature while passing through the entire input sequences. Transformations that combine all the features of the nearby particles sum over the particles i.

The use of CNN in jet-flavor identification techniques showed significant improvement in performance compared to traditional MVA methods. The most recent jet-flavor tagger, the ParticleNet, uses a state-of-the-art architecture of graph neural networks and exploits low-level features of the jet constituents.

10.1.1 The ParticleNet tagger

ParticleNet is a multiclass tagger that identifies hadronically-decaying boosted particles such as Higgs, top, W, Z, or QCD jet, and classifies different decay modes as displayed in Fig. 10.1. The algorithm represents the jets as "particle clouds" which are unordered permutation-invariant sets of particles, continuously distributed in space. The particle cloud data consist of PF particles and SVs and is processed using a dynamic graph convolutional neural network (DGCNN) [170]. The output is a probability score for each jet that indicates its likelihood of belonging to each category.

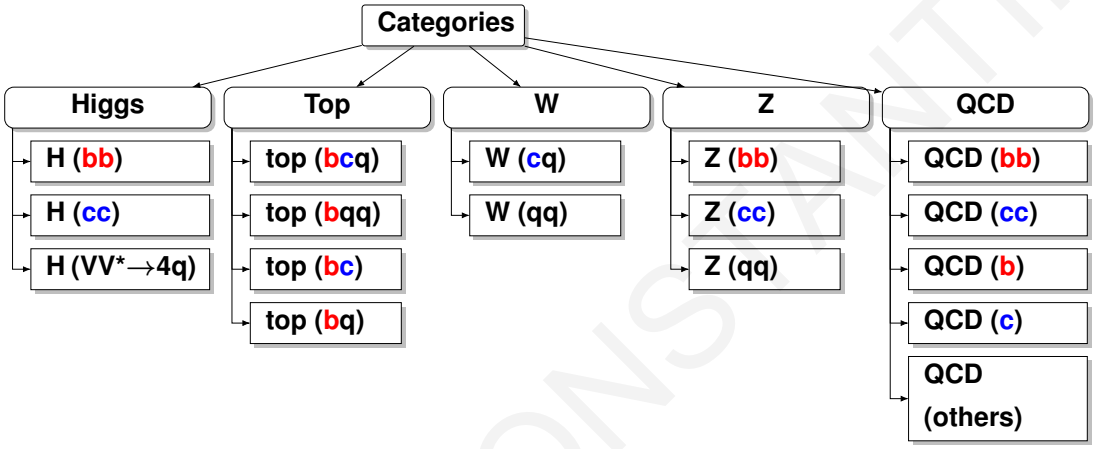


Figure 10.1: Summary of the DeepAK8 and ParticleNet output classes.

Edge convolution

The ParticleNet algorithm uses convolution-like operations for particle cloud data. This is achieved with the edge convolution (EdgeConv) operator [170] which represents the particle clouds as graphs. Each particle of the graph corresponds to a vertex and connections between the particles and their k-nearest neighbors are the edges of the graph.

In an N-dimensional space, the feature vector of each particle i is denoted by $\boldsymbol{x_i} \in \mathbb{R}^N$ and the indices of its k-nearest neighbors are given by $\{i_1, i_2, ...i_k\}$. For each $\boldsymbol{x_i}$ point, the EdgeConv operation gives:

$$\boldsymbol{x_i} = \square_{j=1}^{k} \mathbf{h_\Theta}(\boldsymbol{x_i}, \boldsymbol{x_{i_j}}) \tag{10.2}$$

where $\square$ is a channel-wise symmetric aggregation operation (e.g. $\sum$, $\max$, mean) while $\mathbf{h_\Theta}$ is the edge function parameterized by a set of learnable parameters $\boldsymbol{\Theta}$ and determines the edge features $e_{ij} = \mathbf{h_\Theta}(\boldsymbol{x_i}, \boldsymbol{x_{i_j}})$. The operation of the EdgeConv can be visualized in Fig. 10.2.

The selection of $\square$ and $\mathbf{h_\Theta}$ determines the properties of the EdgeConv. ParticleNet uses the mean as an aggregation operation and a specialized form of the edge function:

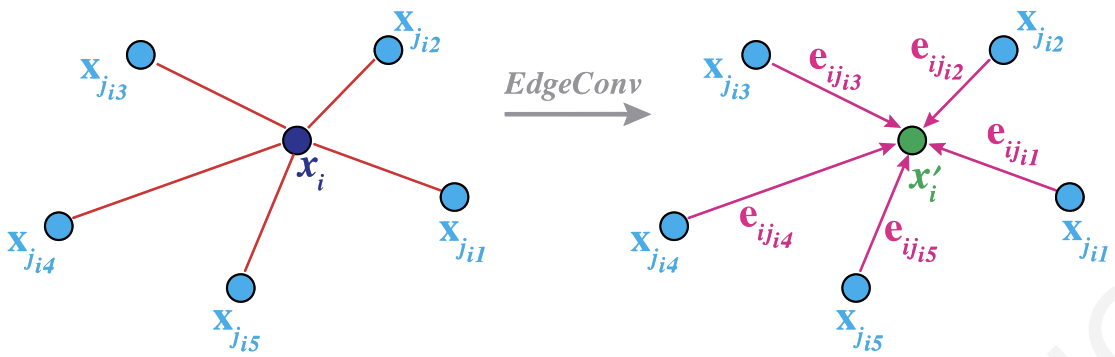


Figure 10.2: The EdgeConv operation calculates the edges that associate each vertex with its k-nearest neighbors.

$$\square = \frac{1}{k}\sum \tag{10.3}$$

$$\mathbf{h}_{\mathbf{\Theta}}(\boldsymbol{x}_i, \boldsymbol{x}_{i_j}) = \bar{\mathbf{h}}_{\mathbf{\Theta}}(\boldsymbol{x}_i, \boldsymbol{x}_i - \boldsymbol{x}_{i_j}) \tag{10.4}$$

The choice of $\bar{\mathbf{h}}_{\mathbf{\Theta}}$ combines global information captured by $\boldsymbol{x}_i$ and local neighboring information from $\boldsymbol{x}_i - \boldsymbol{x}_{i_j}$. The $\bar{\mathbf{h}}_{\mathbf{\Theta}}$ is implemented as a multilayer perceptron (MLP) with shared parameters among the edges.

Architecture

The architecture of the ParticleNet tagger consists of three EdgeConv blocks, followed by a global average pooling operation that aggregates the features extracted by the convolutional filters, a fully connected hidden layer and the output layer. Each EdgeConv block is implemented as a three-layer MLP: one linear transformation layer, followed by a batch normalization layer and an activation layer. The first EdgeConv block uses the spacial (η,ϕ) coordinates to find the distances between the vertices. The subsequent blocks use as input the transformed features of the previous layer which serve as "coordinates" to recompute the new k-nearest neighbors in the feature space. This dynamic update of the graphs is a powerful characteristic of the algorithm that improves its performance and prevents non-local diffusion of information through the particle-cloud data.

10.1.2 Top Tagging

Hadronically decaying Lorentz-boosted top quarks (t^{bst}) are identified by the probability of a jet being classified as a top quark against the probability to be initiated from QCD. This is defined as:

$$\text{TvsQCD} = \frac{\text{prob}(\text{t} \to \text{bcq}) + \text{prob}(\text{t} \to \text{bqq})}{\text{prob}(\text{t} \to \text{bcq}) + \text{prob}(\text{t} \to \text{bqq}) + \text{prob}(\text{QCD})} \tag{10.5}$$

where $\mathrm{prob}(\mathrm{t} \to \mathrm{bcq})$ and $\mathrm{prob}(\mathrm{t} \to \mathrm{bqq})$ are the raw probabilities evaluated by the output of the ParticleNet multiclassifier and $\mathrm{prob}(\mathrm{QCD})$ is the sum of all the QCD decay mode probabilities (bb, cc, b, c, others). Figure 10.3 shows the ROC curve of the TvsQCD discriminant for the ParticleNet and DeepAK8 tagger [171]. The AK8 jets have SoftDrop mass (m_{SD}) between 105 and 210 GeV while generated top quarks in signal and light quarks or gluons in background have $500 < p_{\mathrm{T}} < 1000$ GeV and $|\eta| < 2.4$. The performance of the ParticleNet algorithm shows sizeable improvement compared to the previous DeepAK8 tagger. Three standard WPs are used, namely loose, medium and tight, corresponding to 5.0%, 1.0% and 0.5% efficiency of misidentifying a QCD jet as a $\mathrm{t}^{\mathrm{bst}}$.

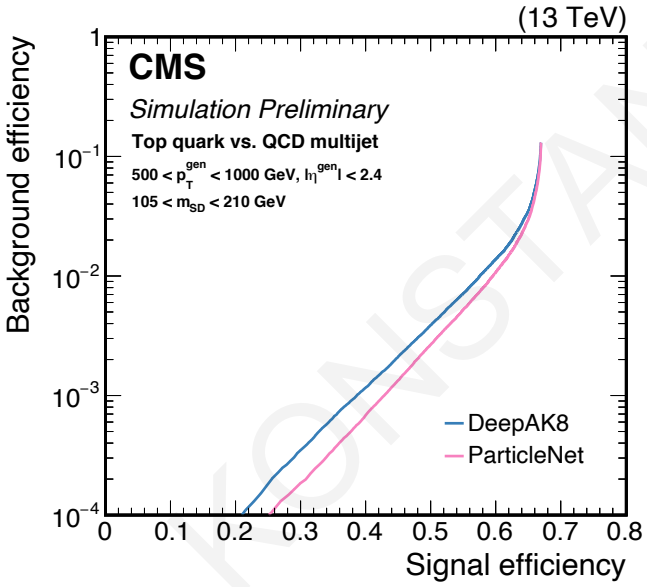


Figure 10.3: Performance of the ParticleNet and DeepAK8 TvsQCD tagger [171].

10.1.3 Mass regression

The jet mass is one of the most powerful variables that discriminate signal from background jets. However, the measurement of the reconstructed mass can be influenced by soft radiation and pileup jets. Several techniques have been developed to mitigate the effect, including the SoftDrop algorithm that subtracts wide-angle and soft collinear constituents. Nevertheless, this subtraction causes inefficiencies that affect the resolution of the jet mass.

A DGCNN has been developed to reconstruct the mass of hadronically decaying boosted jets. The algorithm is a regressor that uses similar network architecture and input features as the ParticleNet tagger to give a prediction of the jet mass. The training sample consists of QCD and Higgs boson events with equal amounts of $\mathrm{H} \to \mathrm{bb}/\mathrm{cc}/\mathrm{qq}$ (q=u,d,s). To prevent mass sculpting effects, the Higgs boson events are generated with a flat mass spectrum between 15 GeV and 250 GeV [172].

The jet mass resolution is defined as the ratio of the regressed mass over the target mass. For Higgs boson events, the target mass is defined as the mass of the generated Higgs boson

while for QCD events is the SoftDrop mass of the associated truth particle-level jet. The resolution is shown in Fig. 10.4 (left) for jets with $p_{\mathrm{T}} > 400\,\mathrm{GeV}$ and $100 < M_{target} < 150\,\mathrm{GeV}$ in the H$\rightarrow$ bb sample and exhibits significant improvement compared to the m_{SD} resolution. The effective resolution defined as half of the minimum interval containing 68% of the area under the mass resolution is illustrated in Fig. 10.4 (right) as a function of the M_{target}. The mass regression shown with solid line provides improved resolution with respect to the SoftDrop mass shown with dashed lines, in all the decay modes and for the entire mass range.

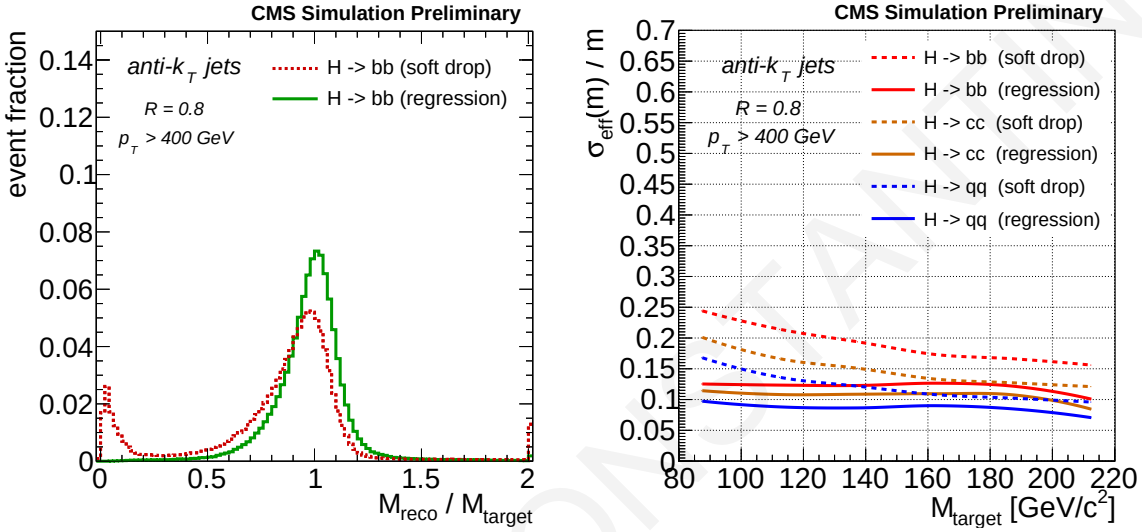


Figure 10.4: Left: Jet mass resolution for H$\rightarrow$ bb jets with $p_{\mathrm{T}} > 400\,\mathrm{GeV}$ and $100 < M_{target} < 150\,\mathrm{GeV}$. Right: Effective resolution as a function of the M_{target}. Figures from [172].

10.2 Designed decorrelated tagger

A designed decorrelated method is applied to decorrelate the output of the ParticleNet TvsQCD tagger from the mass of the jet [173]. The idea of this method is to parameterize the tagger's WP in bins of the jet p_{T} and the ρ variable defined as:

$$\rho = 2\ln(m_{reg}/p_{\mathrm{T}}) \tag{10.6}$$

where m_{reg} is the jet mass regression.

The designed decorrelated tagger (DDT) is a transformation of the tagger's raw score X for a specific misidentification rate $Y\%$. In each (ρ, p_{T}) bin, the value of X that corresponds to $Y\%$ misidentification rate, X(Y%), defines a per-bin WP. For example, Fig. 10.5 shows the ParticleNet TvsQCD score evaluated in QCD multijet events for (ρ, p_{T}) = (-4.5, 588 GeV) (left) and (ρ, p_{T}) = (-2.7, 412 GeV) (right). The X(0.5%) is shown with a vertical dashed line and corresponds to a raw score of ~0.17 for (-4.5, 588 GeV) and ~0.78 for (-2.7, 412 GeV). The X(Y%) scores are evaluated for each (ρ, p_{T}) bin to create a transformation map as shown in Fig. 10.6 for 5.0% misidentification rate.

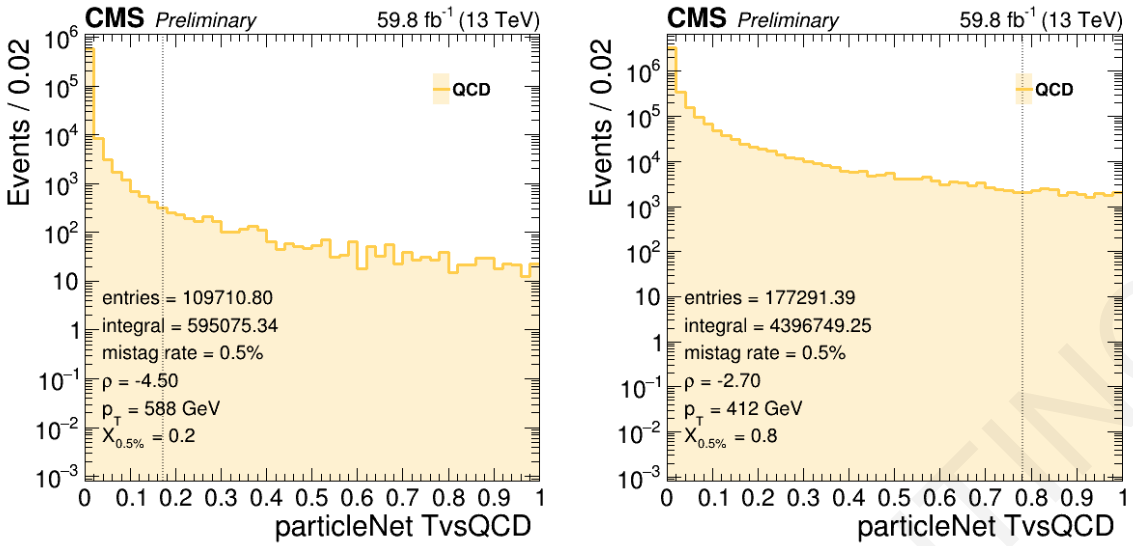


Figure 10.5: ParticleNet TvsQCD score distribution in two (ρ, p_{T}) bins.

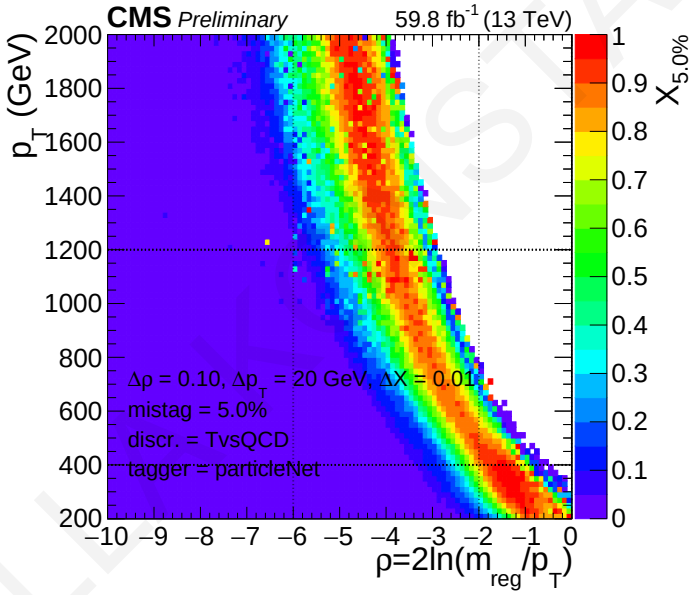


Figure 10.6: DDT transformation map of the ParticleNet TvsQCD tagger, for 5% misidentification rate.

The transformed DDT score $\mathrm{X}^{\mathrm{Y}\%}_{\mathrm{DDT}}$ is then defined as:

$$\mathrm{X}^{\mathrm{Y}\%}_{\mathrm{DDT}} = \mathrm{X} - \mathrm{X}(\mathrm{Y}\%) \tag{10.7}$$

A selection of $\mathrm{X(DDT)} > 0$ ensures constant Y% of simulated QCD multijet events, irrespective of the jet's ρ and p_{T}.

Figure 10.7 shows the regression mass of the AK8 jets with $p_{\mathrm{T}} > 400\,\mathrm{GeV}$ and $|\eta| < 2.4$ that pass the three WPs of the nominal ParticleNet TvsQCD, and the $\mathrm{X}^{5\%}_{\mathrm{DDT}}$. The left and right plots show the distributions for $\mathrm{t\bar{t}}$ and QCD multijet MC simulated samples, respectively. For comparison, the inclusive distribution of $\mathrm{t}^{\mathrm{bst}}$ candidates is also shown, where no $\mathrm{t}^{\mathrm{bst}}$ tagging requirement is applied. In the $\mathrm{t\bar{t}}$ simulation, the inclusive distribution exhibits

two peaks, one around the mass of the W boson and one around the top-quark mass. However, when applying the raw scores of the TvsQCD tagger, the W boson peak disappears, indicating sculpting around the top-quark mass. By using the $X^{5\%}_{DDT}$, the W boson and top-quark mass peaks are well separated. In the QCD multijet m_{reg} distribution, the WPs of the nominal TvsQCD tagger give peaked distributions that imitate the top-quark mass shape. On the contrary, using the $X^{5\%}_{DDT}$ score results in a distribution which is similar to the inclusive distribution. Figure 10.8 compares the regression mass of the t^{bst} candidates that pass or fail the $X^{5\%}_{DDT}$ to the inclusive distribution in QCD multijet events. The three distributions share the same shape, indicating that the DDT approach eliminates mass sculpting effects.

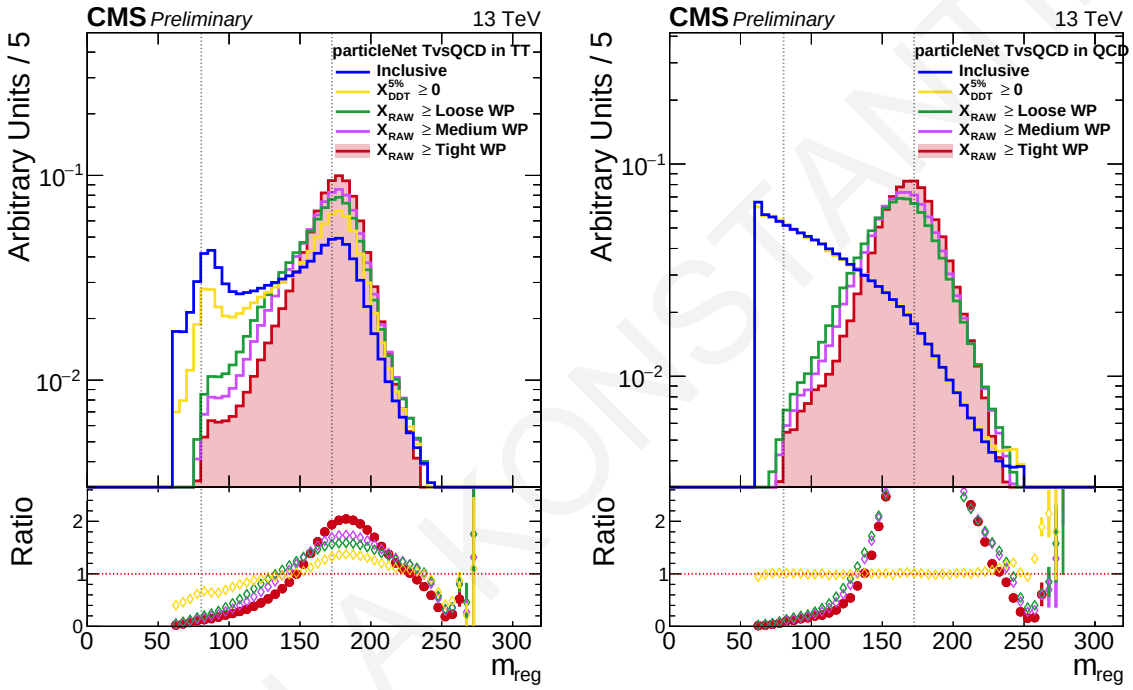


Figure 10.7: Regression mass of the t^{bst} candidates that pass the three WPs of the ParticleNet TvsQCD, the $X^{5\%}_{DDT}$ and the inclusive distribution without any top tagging requirements, in the $t\bar{t}$ (left) and QCD multijet (right) MC samples.

10.3 Event selection

This section presents the event selections applied to data and simulation to select $H^{\pm}$ signal in the semi-boosted topology. The semi-boosted and resolved analyses share the same HLT triggers and the requirements on the reconstructed PVs, E_T^{miss} filters and lepton veto discussed in Sections 9.1.1 to 9.1.4. The t^{bst} candidates are reconstructed as TightID AK8 jets with $p_T > 400$ GeV, $|\eta| < 2.4$ and $m_{reg} > 50$ GeV. The jet flavor is identified with the ParticleNet DDT TvsQCD. In this analysis, the selected events are required to contain exactly one t^{bst} candidate that fulfills the $X^{5\%}_{DDT}$. At least four TightID AK4 jets with $p_T > 40$ GeV, $|\eta| < 2.4$ and $H_T > 500$ GeV are selected in the signal events, of which at least two are b-tagged, using the DeepJet WP that corresponds to 1% misidentification rate. The selected

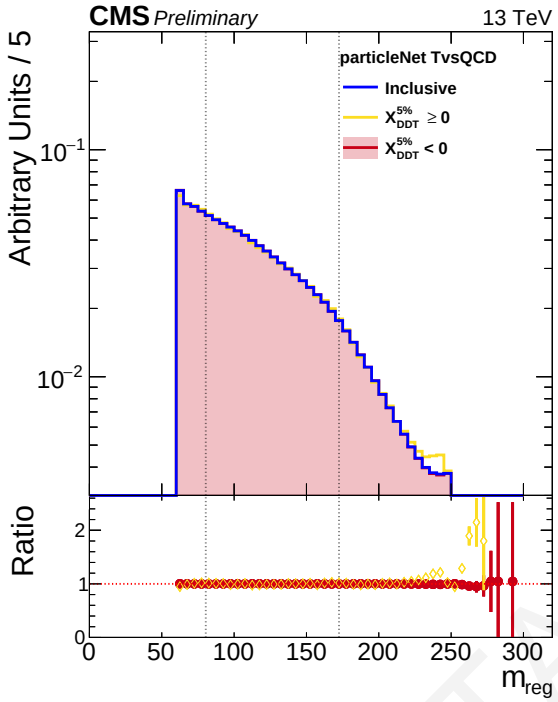


Figure 10.8: Regression mass of the t^{bst} candidate that pass or fail the $X_{DDT}^{5\%}$, and the inclusive distribution without any top tagging requirements, in the QCD multijet MC sample.

events must contain less than two loose-tagged t^{res} candidates with a mass between 50 GeV and 260 GeV to ensure orthogonality with the resolved analysis. Three kinematic requirements are applied to enhance the signal significance. In the case of $m_{H^\pm} \gtrsim 1$ TeV, the $H^\pm$ has a small transverse momentum and its decay products are back to back. Thus, a $\Delta R > 1.2$ is required between the t^{bst} candidate and the leading in p_T b-tagged jet (b^{ldg}). Furthermore, to reduce QCD multijet contributions with b $\bar{b}$ pairs produced via gluon splitting, the distance between the two most-distant b jets must be greater than 1.0 while the mass of the heaviest b b system is required to be at least 200 GeV. The complete set of event selection criteria is summarized in Table 10.1.

This analysis targets $H^\pm$ with large mass ($m_{H^\pm} \sim 1$ TeV) whose decay products are highly collimated and energetic. Therefore, the invariant mass of the $H^\pm$ is reconstructed by the t^{bst} candidate and the leading in p_T b jet (m_{tb}).

10.4 Background estimation

This section describes the measurement of the dominant QCD multijet and $t\bar{t}$ background in the semi-boosted regime. In both cases, the shape and yields of the MC simulation events are corrected using the two-dimensional distribution ($m_{t^{bst}}$, m_{tb}). Other subdominant SM processes are estimated from simulation.

While all backgrounds are non-resonant in m_{tb}, backgrounds containing top quarks, W and Z bosons can have a resonant $m_{t^{bst}}$. Therefore, the $t\bar{t}$ background can be decomposed into three categories based on the jet-to-parton matching: the merged-t jet, merged-W jet,

Table 10.1: Summary of the event selection steps taken to select the signal analysis sample in the semi-boosted analysis.

1	Signal trigger	AK4 multijet, AK8 jet, Inclusive H_{T}
2	MET Filters	Reject artificially high $\vec{p}_T^{\,\mathrm{miss}}$
3	PV	$N_{dof} \geq 4$, |z| < 24 cm, $\mathrm{r_{xy}} < 2$ cm
4	= 0 electrons	$p_{\mathrm{T}} > 10$ GeV, $\|\eta\| < 2.4$, miniIso < 0.4, cut-based ID (95% efficiency)
5	= 0 muons	$p_{\mathrm{T}} > 10$ GeV, $\|\eta\| < 2.4$, miniIso < 0.4, loose ID
6	= 0 τ_{h}	$p_{\mathrm{T}} > 20$ GeV, $\|\eta\| < 2.1$, DeepTau (very loose $e \to \tau_{\mathrm{h}}$, medium $\mu \to \tau_{\mathrm{h}}$, loose $j \to \tau_{\mathrm{h}}$)
7	= 1 $\mathrm{t}^{\mathrm{bst}}$	AK8, $p_{\mathrm{T}} > 400$ GeV, $\|\eta\| < 2.4$, TightID, ParticleNet DDT TvsQCD (5% misID rate)
8	≥ 4 jets	AK4, $p_{\mathrm{T}} > 40$ GeV, $\|\eta\| < 2.4$, TightID, $H_{\mathrm{T}} > 500$ GeV
9	≥ 2 b jets	AK4 jet, DeepJet (1% misID rate)
10	< 2 $\mathrm{t}^{\mathrm{res}}$	$50 < m_{\mathrm{t^{reg}}} < 260$ GeV, with ≥ 1 b-tagged subjet
11	$\Delta R(t^{bst}, b^{ldg}) > 1.2$	
12	$\max(\Delta R_{bb}) > 1.0$	
13	$\max(m_{bb}) > 200$ GeV	

or non-merged. A reconstructed $\mathrm{t}^{\mathrm{bst}}$ belongs to the merged-t category if it matches a generated top quark ($\mathrm{t}^{\mathrm{gen}}$) with $\Delta R(\mathrm{t}^{\mathrm{bst}}, \mathrm{t}^{\mathrm{gen}}) < 0.8$ and all three quarks from the $\mathrm{t}^{\mathrm{gen}}$ are within $\Delta R = 0.8$ of the axis of the $\mathrm{t}^{\mathrm{bst}}$ candidate. Non-merged-t events with a $\mathrm{t}^{\mathrm{bst}}$ candidate matched to a generated W boson are considered merged-W events. As in the merged-t category, the matching criteria are $\Delta R(\mathrm{t}^{\mathrm{bst}}, \mathrm{W}^{\mathrm{gen}}) < 0.8$, with the two quarks from the $\mathrm{W}^{\mathrm{gen}}$ decay enclosed inside the $\mathrm{t}^{\mathrm{bst}}$ candidate. The remaining events that do not belong to the merged-t and merged-W categories are categorized as non-merged events. The m_{reg} distribution is illustrated in Fig. 10.9 in MC simulated samples. The three $\mathrm{t\bar{t}}$ categories shown with purple colors exhibit mass peaks corresponding to the W boson and top-quark resonances.

To determine the contribution of QCD multijet, merged-t, merged-W and non-merged $\mathrm{t\bar{t}}$ in the SR, MC simulation events are used, and the normalization of each category is left floating in the final ($m_{\mathrm{t^{bst}}}$, m_{tb}) fit. The two-dimensional ($m_{\mathrm{t^{bst}}}$, m_{tb}) template is shown in Fig. 10.10 for $\mathrm{t\bar{t}}$ events. The $m_{\mathrm{t^{bst}}}$ is divided into slices. In each $m_{\mathrm{t^{bst}}}$ slice the QCD multijet and the three $\mathrm{t\bar{t}}$ categories are fitted simultaneously to predict the shape of the m_{tb}.

10.5 Systematic uncertainties

A preliminary set of rate systematic uncertainties is introduced as nuisance parameters in the final maximum likelihood fit to data. Uncertainties related to the integrated luminosity, pileup reweighting, lepton veto, as well as theoretical uncertainties are described in Section 9.6. Additional flat systematic uncertainties arise from the trigger efficiency SF, $\mathrm{t}^{\mathrm{bst}}$ (mis)tagging SF, JES and JER.

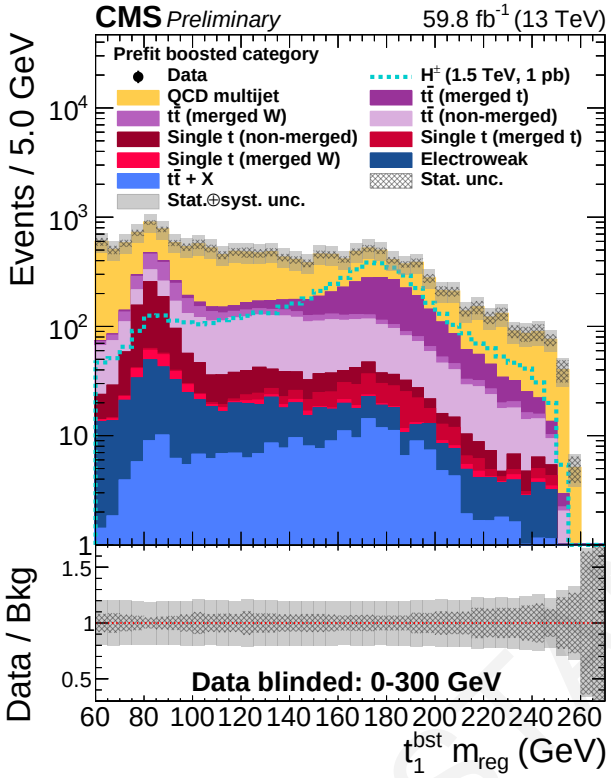


Figure 10.9: Prefit distribution of the regression mass of the t^{bst} candidate in the SR. The signal of $m_{H^{\pm}}$ = 1.5 TeV is also shown.

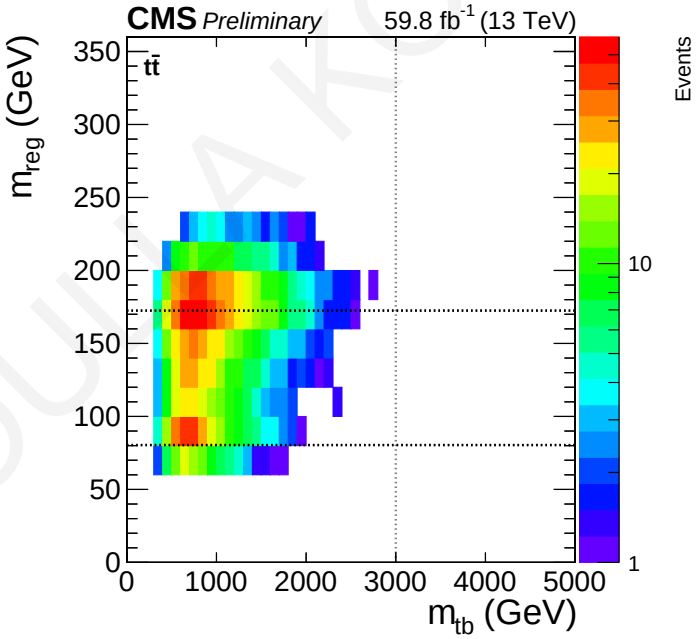
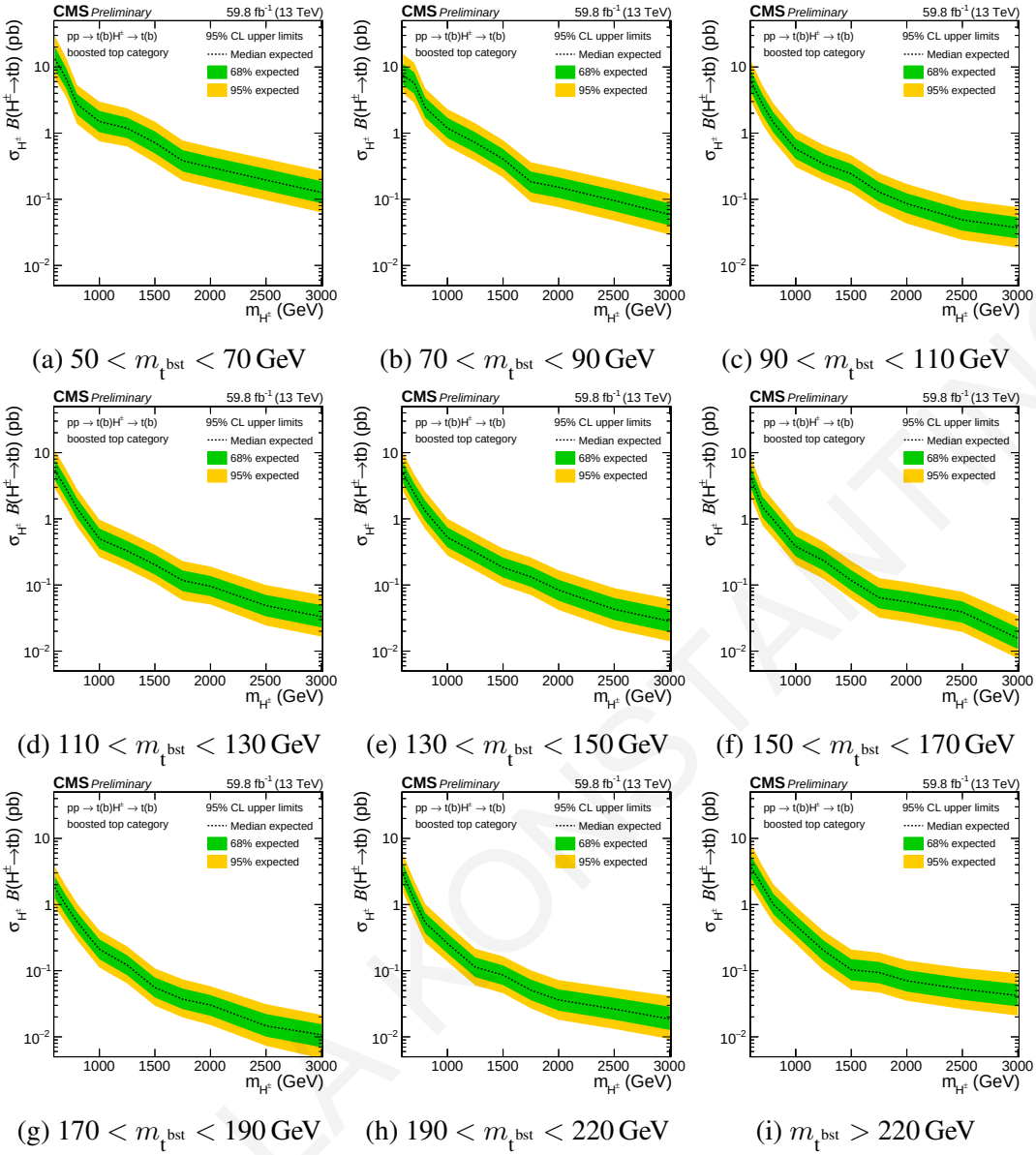


Figure 10.10: Two-dimensional template of the mass of the t^{bst} and the reconstructed invariant mass of the $H^{\pm}$.

10.6 Results

To extract signal in the presence of SM background, a simultaneous fit for signal and background is performed in the ($m_{t^{bst}}$, m_{tb}) plane. The m_{tb} distribution is projected in nine slices



(a) $50 < m_{t^{bst}} < 70$ GeV (b) $70 < m_{t^{bst}} < 90$ GeV (c) $90 < m_{t^{bst}} < 110$ GeV

(d) $110 < m_{t^{bst}} < 130$ GeV (e) $130 < m_{t^{bst}} < 150$ GeV (f) $150 < m_{t^{bst}} < 170$ GeV

(g) $170 < m_{t^{bst}} < 190$ GeV (h) $190 < m_{t^{bst}} < 220$ GeV (i) $m_{t^{bst}} > 220$ GeV

Figure 10.11: Upper lumits of the nine semi-boosted categories that correspond to different slices in $m_{t^{bst}}$.

of $m_{t^{bst}}$ that define the nine SRs of the semi-boosted analysis. These categories correspond to $m_{t^{bst}}$ bins with the following lower edges: $m_{t^{bst}}$ = [50, 70, 90, 110, 130, 150, 170, 190, 220] GeV. The m_{tb} is fitted in the nine categories to extract expected upper limits. The expected upper limits in each category are displayed in Fig. 10.11. The regions of $m_{t^{bst}}$ closer to the true top-quark mass exhibit the highest sensitivity. Finally, the results are combined to calculate the 95% CL upper limits on the $\sigma(\mathrm{pp} \to \mathrm{tbH}^{\pm}) \times \mathcal{B}(\mathrm{H}^{\pm} \to \mathrm{tb})$.

Figure 10.13 compares the limits of the resolved analysis obtained with the 2018 data with the limits of the semi-boosted analysis. The semi-boosted analysis exhibits the best sensitivity for $m_{\mathrm{H}^{\pm}}$ values larger than 800 GeV while the resolved analysis limits are the most stringent at lower masses. A comparison between the expected limit of the semi-boosted analysis and the expected limit of the CMS published results in the boosted regime is also performed. As shown in Fig. 10.14 the new results exhibit improved sensitivity compared to

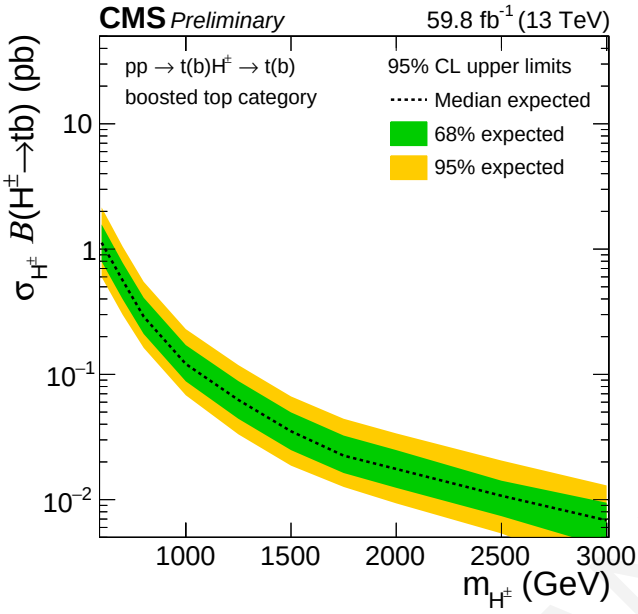


Figure 10.12: Combined upper lumits of the $\sigma(\mathrm{pp} \to \mathrm{tbH}^{\pm}) \times \mathcal{B}(\mathrm{H}^{\pm} \to \mathrm{tb})$ in the semi-boosted analysis.

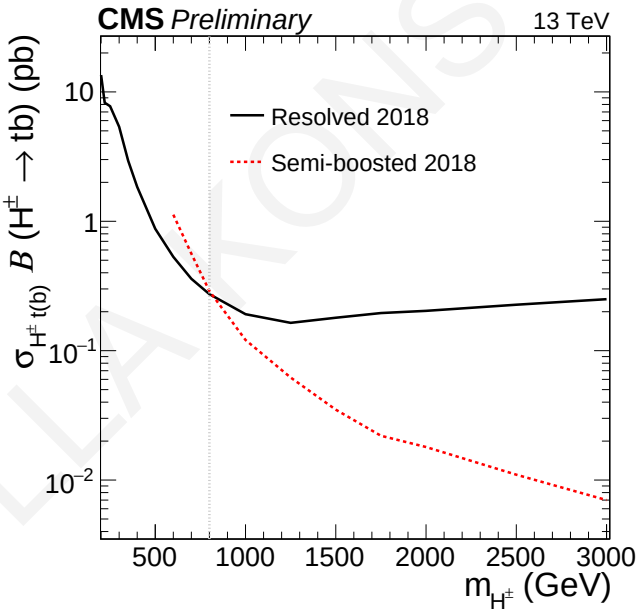


Figure 10.13: Comparison of the median expected limit on the $\sigma(\mathrm{pp} \to \mathrm{tbH}^{\pm}) \times \mathcal{B}(\mathrm{H}^{\pm} \to \mathrm{tb})$ of the resolved and semi-boosted analysis using the 2018 data.

the published results, taking into account the difference in the luminosity between the 2016 and 2018 data. This enhancement is primarily attributed to the utilization of an advanced boosted top tagging algorithm which replaces traditional cut-based methods used for the jet-flavor identification. Additionally, the semi-boosted analysis presented in this work follows a notably simplified approach in contrast to the published analysis, which uses 96 categories.

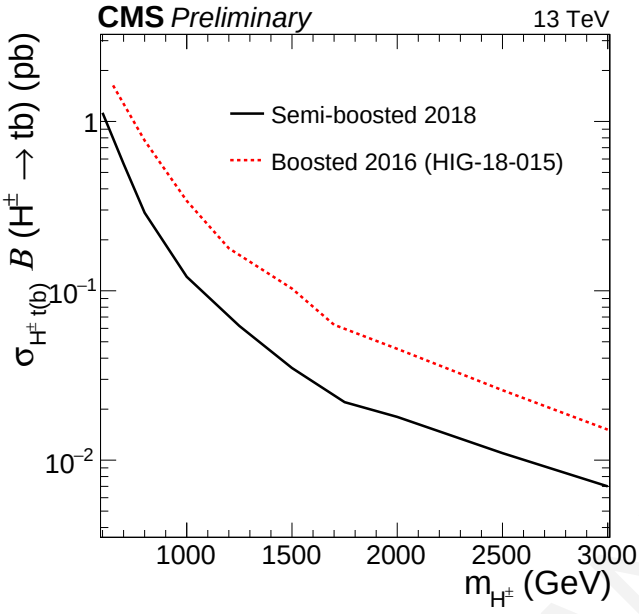


Figure 10.14: Comparison of the median expected limit on the $\sigma(\mathrm{pp} \to \mathrm{tbH}^{\pm}) \times \mathcal{B}(\mathrm{H}^{\pm} \to \mathrm{tb})$ of the semi-boosted analysis and the CMS published median expected limit [53] in the boosted regime.

10.7 Optimization

Several optimization studies have been performed to improve the sensitivity of the semi-boosted analysis. First, the event selections require tighter requirement on the DDT TvsQCD tagger with 1% misidentification rate ($\mathrm{X}^{1\%}_{\mathrm{DDT}}$) and a $m_{\mathrm{t}^{\mathrm{bst}}}$ between 60 to 240 GeV. Secondly, the binning of the $m_{\mathrm{t}^{\mathrm{bst}}}$ has been shrinked in the regions of large statistics, resulting into ten categories with lower edges: $m_{\mathrm{t}^{\mathrm{bst}}}$ = [60, 75, 90, 105, 120, 135, 150, 165, 180, 195] GeV. Finally, to improve the measurement of the QCD multijet background, a data-driven method was studied based on the use of a QCD-multijet enriched CR. This CR is defined by inverting the $\mathrm{X}^{1\%}_{\mathrm{DDT}}$ requirement ($\mathrm{X}^{1\%}_{\mathrm{DDT}} < 0$). Figure 10.15 shows the m_{reg} in the CR.

The shapes of the observables in QCD multijet events are anticipated to be similar in the SR and CR. This is due to the tagger's score being decorrelated from the jet's mass as shown in Fig. 10.8. To verify this assumption the shapes of the ($m_{\mathrm{t}^{\mathrm{bst}}}$, m_{tb}) in QCD multijet MC simulation from the SR are compared to the data in the CR. The two dimensional distributions are shown in Fig. 10.16. The SR-over-CR ratio illustrated in Fig. 10.17 shows no correlation between the shapes in the two regions. The comparison is also performed in bins of the $m_{\mathrm{t}^{\mathrm{bst}}}$, as a function of the m_{tb} normalized in unity, as demonstrated in Fig. 10.18. No discernible bias is observed and any differences are within statistical uncertainties. Moreover, to ensure that the data and QCD multijet events have similar shapes, a comparison between the two is performed. The shapes show good agreement as shown in Fig. 10.19.

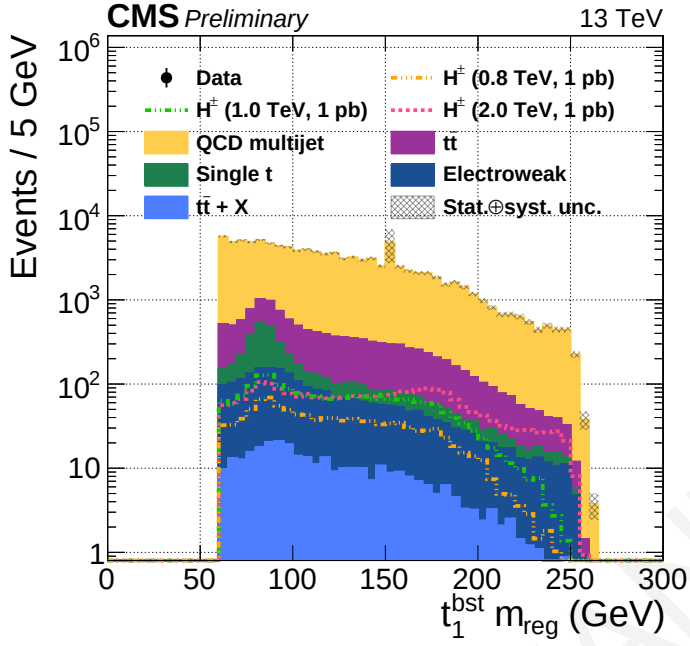


Figure 10.15: Regression mass of the t$^{\text{bst}}$ candidate in the QCD-multijet enriched CR.

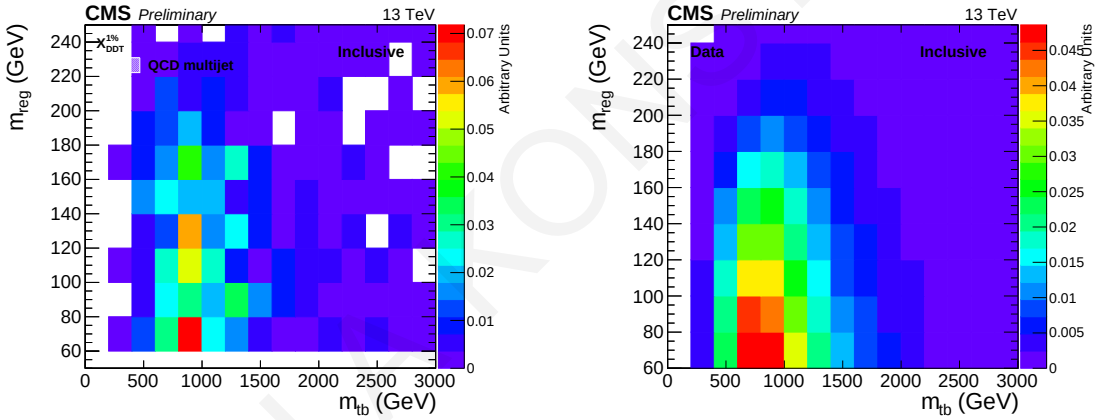


Figure 10.16: Distribution of the ($m_{\text{t}^{\text{bst}}}$, m_{tb}) in QCD multijet MC simulation from the SR (left) and data from the CR (right).

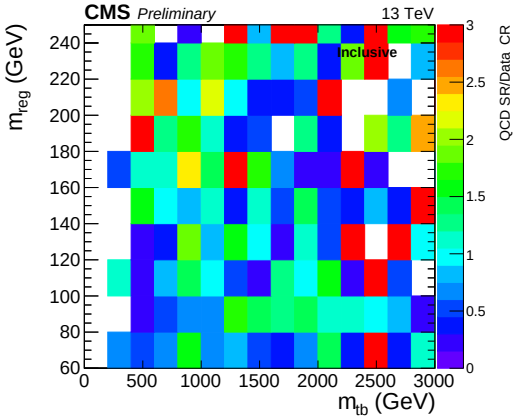


Figure 10.17: Ratio of the number of QCD multjet events in SR and the data in CR, in bins of the ($m_{\text{t}^{\text{bst}}}$, m_{tb}).

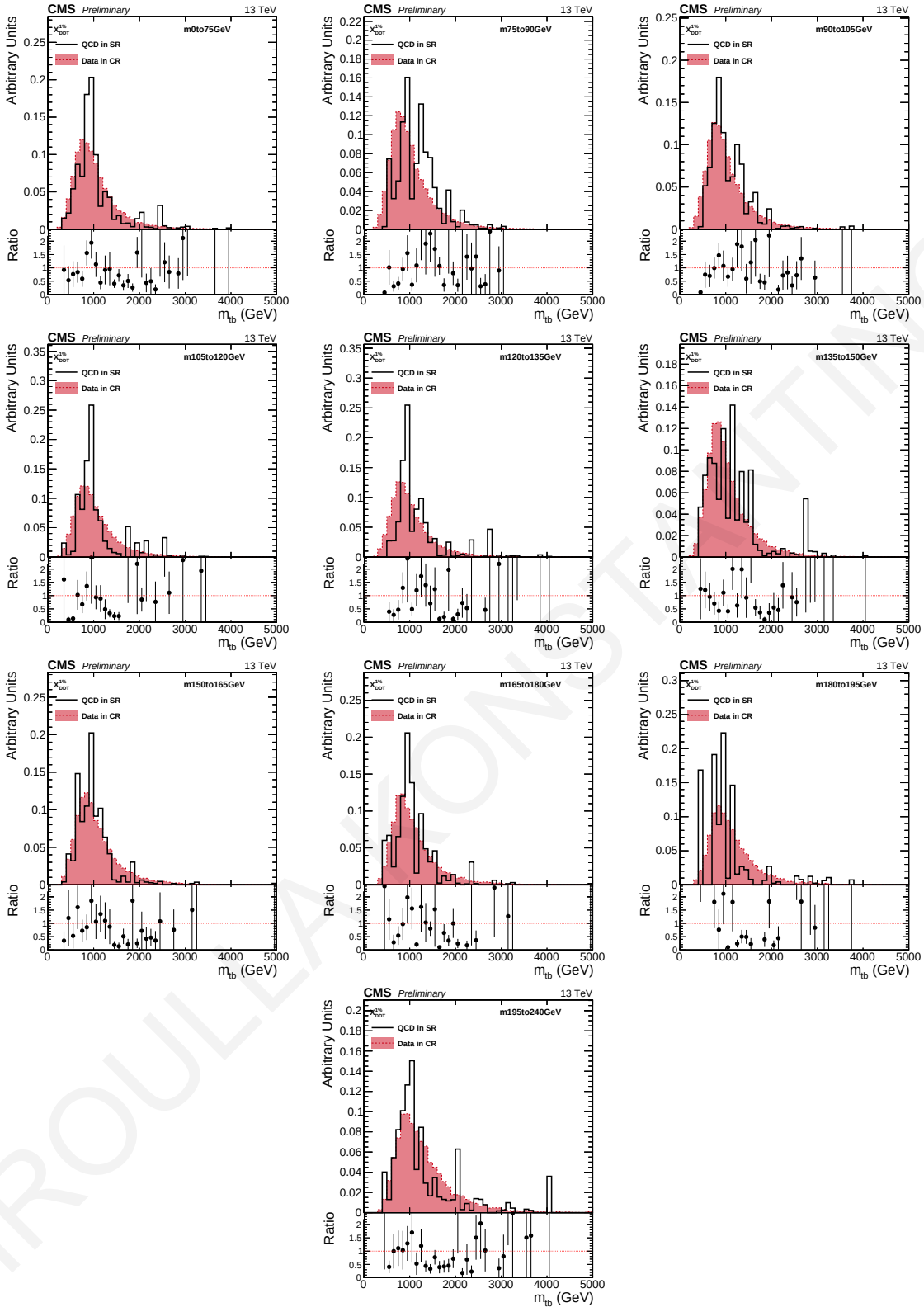


Figure 10.18: Distribution of the m_{tb} in bins of the $m_{\mathrm{t}^{\mathrm{bst}}}$ in QCD multjet events from the SR and data from the CR normalized to unity. The plots demonstrate the similarity of the m_{tb} distributions between the SR (QCD multijet) and CR (data).

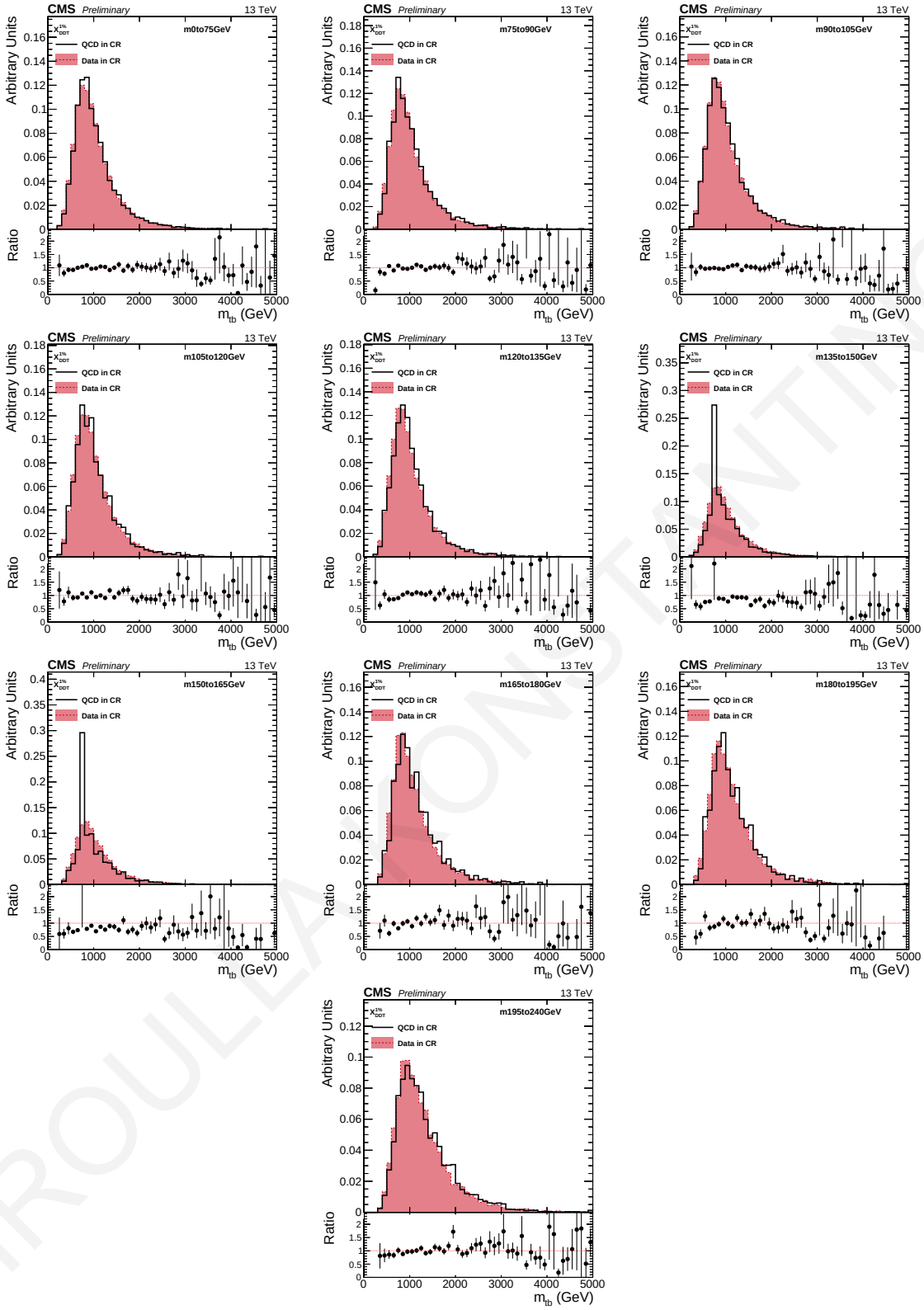


Figure 10.19: Distribution of the m_{tb} in bins of the $m_{\mathrm{t}^{\mathrm{bst}}}$ in QCD multjet events, and data from the CR normalized to unity. The plots demonstrate the similarity of the m_{tb} distributions between the QCD multjet and data.

The QCD multijet background in the SR is estimated from data in the CR, normalized to the expected number of QCD multijet events in the SR. The normalization factor is determined by taking the ratio of the number of simulated QCD multijet events in the SR to the number of QCD multijet events in the CR.

The preliminary expected limits are shown in Fig. 10.20 for the ten categories and the combined limit using the 2018 data. These first results indicate noteworthy improvent compared to the expected limits illustrated in Fig. 10.12.

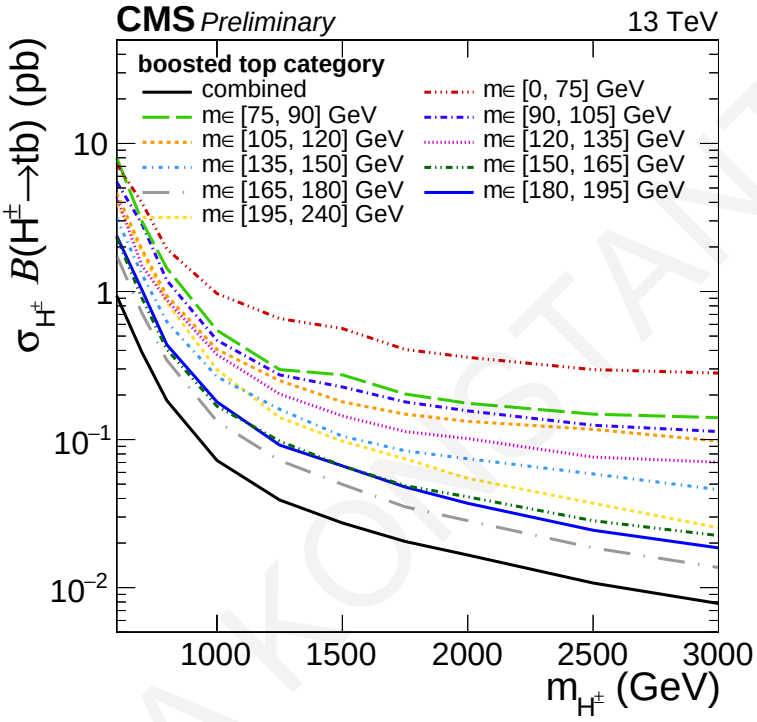


Figure 10.20: Limits using the optimized selections and categories, and the data-driven method to estimate the QCD multijet background.

11 Summary

A search for charged Higgs bosons ($H^{\pm}$) which decay to a top and a bottom quark in the all-jet final state has been presented. The proton-proton collision data used in this analysis were recorded by the CMS detector during the second operating period of the LHC (Run II) at a center-of-mass energy of 13 TeV which corresponds to an integrated luminosity of 138 fb^{-1}.

The search is performed in two distinct topologies. The resolved topology is characterized by $H^{\pm}$ of low masses whose decay products are moderately boosted. The hadronically decaying top quarks are identified using a multivariate top-quark tagger that distinguishes between trijet combinations originating from the decay of top quarks, and combinatorial trijets. Based on the identification criteria of the top-quark candidates, the analysis exploits two signal categories. The first category includes two top-quark candidates, tagged with a misidentification rate of 5%. In the second category, the tagging criteria of one top-quark candidate are loosened and correspond to a misidentification rate of 10%. The dominant QCD multijet background is estimated in a data-driven method while all the remaining Standard Model (SM) background processes with simulations of the corresponding processes. A mass-parameterized deep neural network (DNN) has been developed to extract the signal in the presence of the SM background.

The search is extended in the semi-boosted topology that is dedicated to $H^{\pm}$ with masses mainly above 1 TeV. In the boosted regime, the top quarks from the $H^{\pm}$ decay can be collimated and reconstructed as large-radius jets. The flavor of the Lorentz-boosted jets is identified using the ParticleNet top-quark tagger. The invariant mass of the $H^{\pm}$ is reconstructed from the Lorentz-boosted top quark candidate and the b jet with the largest transverse momentum. The signal is searched in the two-dimensional space of the mass of the Lorentz-boosted top quark candidate and the invariant mass of the $H^{\pm}$.

The analyses set upper limits of the product of the $H^{\pm}$ production cross section and its branching ratio into a top-bottom quark-antiquark pair. The resolved analysis sets limits of 8.375 to 0.107 pb for $H^{\pm}$ masses between 200 to 1250 GeV, using the entire Run II data. The limits obtained with the 2016 data that correspond to an integrated luminosity of 35.9 fb^{-1} show improved sensitivity compared to the CMS published results for the same channel in the resolved topology. The main sources of improvement are the use of the upgraded b and top-quark tagging algorithms, the data categorization, the improved data-driven background estimation method and the use of the high-performance mass parameterized DNN to enhance

the separation between signal and background.

The semi-boosted analysis is a new unexplored topology that exhibits promising preliminary results using the data recorded in 2018 that correspond to 59.8 fb^{-1}. Expected upper limits are extracted which show signal sensitivity for large $H^{\pm}$ masses above 800 GeV, where the resolved analysis sensitivity comes to a plateau. Better sensitivity is observed compared to the CMS published results in the boosted topology after scaling to the 2016 luminosity, due to the superior boosted top tagging algorithm and more all data-driven method. Furthermore, the proposed analysis is a lot simpler than the published one that is based on 96 categories and the current one does not use the two boosted categories yet.

The completion of the semi-boosted analysis and the inclusion of new topologies categories with boosted W boson candidates and more than one boosted top-quark candidates are expected to enhance further the sensitivity of the fully hadronic channel in a wide range of $H^{\pm}$ masses. Following all the comprehensive cross-checks of the employed methods, the data will be unblinded to investigate any potential excess beyond the expected SM background.

Bibliography

[1] M. Thomson, *Modern particle physics*. Cambridge University Press, 2013 (cit. on p. 1).

[2] D. Griffiths, *Introduction to elementary particles-david griffiths*, 2008 (cit. on p. 1).

[3] D. H. Perkins and D. H. Perkins, *Introduction to high energy physics*. CAMBRIDGE university press, 2000 (cit. on p. 1).

[4] R. Mann, *An introduction to particle physics and the standard model*. Taylor & Francis, 2010 (cit. on p. 1).

[5] A. Pich, "The standard model of electroweak interactions," 2012. DOI: 10.48550/ARXIV.1201.0537. [Online]. Available: https://arxiv.org/abs/1201.0537 (cit. on p. 1).

[6] W. Pauli, "The connection between spin and statistics," *Phys. Rev.*, vol. 58, pp. 716–722, 8 Oct. 1940. DOI: 10.1103/PhysRev.58.716. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRev.58.716 (cit. on p. 2).

[7] M. E. Peskin and D. V. Schroeder, *An introduction to quantum field theory (boulder, co*, 1995 (cit. on p. 2).

[8] P. A. M. Dirac and N. H. D. Bohr, "The quantum theory of the emission and absorption of radiation," *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, vol. 114, no. 767, pp. 243–265, 1927. DOI: 10.1098/rspa.1927.0039. eprint: https://royalsocietypublishing.org/doi/pdf/10.1098/rspa.1927.0039. [Online]. Available: https://royalsocietypublishing.org/doi/abs/10.1098/rspa.1927.0039 (cit. on p. 2).

[9] D. J. Gross and F. Wilczek, "Ultraviolet behavior of non-abelian gauge theories," *Phys. Rev. Lett.*, vol. 30, pp. 1343–1346, 26 Jun. 1973. DOI: 10.1103/PhysRevLett.30.1343. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.30.1343 (cit. on p. 3).

[10] H. D. Politzer, "Reliable perturbative results for strong interactions?" *Phys. Rev. Lett.*, vol. 30, pp. 1346–1349, 26 Jun. 1973. DOI: 10.1103/PhysRevLett.30.1346. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.30.1346 (cit. on p. 3).

[11] S. L. Glashow, "Partial-symmetries of weak interactions," *Nuclear Physics*, vol. 22, no. 4, pp. 579–588, 1961, ISSN: 0029-5582. DOI: https://doi.org/10.1016/0029-5582(61)90469-2. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0029558261904692 (cit. on p. 3).

[12] S. Weinberg, "A model of leptons," *Phys. Rev. Lett.*, vol. 19, pp. 1264–1266, 21 Nov. 1967. DOI: 10.1103/PhysRevLett.19.1264. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.19.1264 (cit. on p. 3).

[13] A. Salam, "Elementary particle theory," in *Elementary Particle Theory*, N. Svartholm, Ed., Stockholm: Almquist and Wiksells, 1969, pp. 367–377 (cit. on p. 3).

[14] F. Englert and R. Brout, "Broken symmetry and the mass of gauge vector mesons," *Phys. Rev. Lett.*, vol. 13, pp. 321–323, 9 Aug. 1964. DOI: 10.1103/PhysRevLett.13.321. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.13.321 (cit. on p. 3).

[15] P. W. Higgs, "Broken symmetries and the masses of gauge bosons," *Phys. Rev. Lett.*, vol. 13, pp. 508–509, 16 Oct. 1964. DOI: 10.1103/PhysRevLett.13.508. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.13.508 (cit. on p. 3).

[16] A. Pich, *Aspects of quantum chromodynamics*, 2000. DOI: 10.48550/ARXIV.HEP-PH/0001118. [Online]. Available: https://arxiv.org/abs/hep-ph/0001118 (cit. on p. 3).

[17] J. Goldstone, "Field theories with superconductor solutions," *Il Nuovo Cimento*, vol. 19, pp. 154–164, 1 1961. DOI: 10.1007/BF02812722 (cit. on p. 7).

[18] ATLAS Collaboration, "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC," *Phys. Lett. B*, vol. 716, pp. 1–29, 2012. DOI: 10.1016/j.physletb.2012.08.020. arXiv: 1207.7214 [hep-ex] (cit. on p. 9).

[19] CMS Collaboration, "Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC," *Phys. Lett. B*, vol. 716, pp. 30–61, 2012. DOI: 10.1016/j.physletb.2012.08.021. arXiv: 1207.7235 [hep-ex] (cit. on p. 9).

[20] L. D. Landau, "The moment of a 2-photon system," *Dokl. Akad. Nawk. USSR*, vol. 60, p. 207, 1948 (cit. on p. 9).

[21] C. N. Yang, "Selection rules for the dematerialization of a particle into two photons," *Phys. Rev.*, vol. 77, pp. 242–245, 2 Jan. 1950. DOI: 10.1103/PhysRev.77.242. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRev.77.242 (cit. on p. 9).

[22] ATLAS Collaboration, "Evidence for the spin-0 nature of the higgs boson using atlas data," *Physics Letters B*, vol. 726, no. 1, pp. 120–144, 2013, ISSN: 0370-2693. DOI: https://doi.org/10.1016/j.physletb.2013.08.026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0370269313006527 (cit. on p. 9).

[23] CMS Collaboration, "A portrait of the Higgs boson by the CMS experiment ten years after the discovery," *Nature*, vol. 607, no. 7917, pp. 60–68, 2022. DOI: 10.1038/s41586-022-04892-x. arXiv: 2207.00043 [hep-ex] (cit. on pp. 9, 10).

[24] A. D. Sakharov, "Violation of $\mathcal{CP}$ Invariance, C asymmetry, and baryon asymmetry of the universe," *Pisma Zh. Eksp. Teor. Fiz.*, vol. 5, pp. 32–35, 1967. DOI: 10.1070/PU1991v034n05ABEH002497 (cit. on p. 10).

[25] A. Sakharov, "Violation of cpin variance, casymmetry, and baryon asymmetry of the universe," *Usp. Fiz. Nauk*, vol. 161, pp. 61–64, 1991 (cit. on p. 10).

[26] Q. R. Ahmad *et al.*, "Measurement of the rate of $\nu_e + d \rightarrow p + p + e^-$ interactions produced by ^8B solar neutrinos at the sudbury neutrino observatory," *Phys. Rev. Lett.*, vol. 87, p. 071 301, 7 Jul. 2001. DOI: 10.1103/PhysRevLett.87.071301. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.87.071301 (cit. on p. 11).

[27] Q. R. Ahmad *et al.*, "Direct evidence for neutrino flavor transformation from neutral-current interactions in the sudbury neutrino observatory," *Phys. Rev. Lett.*, vol. 89, p. 011 301, 1 Jun. 2002. DOI: 10.1103/PhysRevLett.89.011301. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.89.011301 (cit. on p. 11).

[28] Y. Fukuda *et al.*, "Evidence for oscillation of atmospheric neutrinos," *Phys. Rev. Lett.*, vol. 81, pp. 1562–1567, 8 Aug. 1998. DOI: 10.1103/PhysRevLett.81.1562. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.81.1562 (cit. on p. 11).

[29] T. N. Prize, *The nobel prize in physics 2015*, 2015. [Online]. Available: https://www.nobelprize.org/uploads/2017/09/advanced-physicsprize2015.pdf (cit. on p. 11).

[30] R. L. Workman *et al.*, "Review of Particle Physics," *PTEP*, vol. 2022, p. 083C01, 2022. DOI: 10.1093/ptep/ptac097 (cit. on pp. 12, 17, 43).

[31] *Particle Dark Matter: Observations, Models and Searches*. Cambridge University Press, 2010. DOI: 10.1017/CBO9780511770739 (cit. on p. 12).

[32] G. Bertone, D. Hooper, and J. Silk, "Particle dark matter: Evidence, candidates and constraints," *Physics Reports*, vol. 405, no. 5-6, pp. 279–390, Jan. 2005. DOI: 10.1016/j.physrep.2004.08.031. [Online]. Available: https://doi.org/10.1016%5C%2Fj.physrep.2004.08.031 (cit. on p. 12).

[33] G. Branco, P. Ferreira, L. Lavoura, M. Rebelo, M. Sher, and J. P. Silva, "Theory and phenomenology of two-higgs-doublet models," *Physics Reports*, vol. 516, no. 1-2, pp. 1–102, Jul. 2012. DOI: 10.1016/j.physrep.2012.02.002. [Online]. Available: https://doi.org/10.1016%5C%2Fj.physrep.2012.02.002 (cit. on p. 12).

[34] S. L. Glashow and S. Weinberg, "Natural Conservation Laws for Neutral Currents," *Phys. Rev.*, vol. D15, p. 1958, 1977. DOI: 10.1103/PhysRevD.15.1958 (cit. on p. 14).

[35] E. A. Paschos, "Diagonal Neutral Currents," *Phys. Rev. D*, vol. 15, p. 1966, 1977. DOI: 10.1103/PhysRevD.15.1966 (cit. on p. 14).

[36] S. P. MARTIN, "A SUPERSYMMETRY PRIMER," in *Perspectives on Supersymmetry*, WORLD SCIENTIFIC, Jul. 1998, pp. 1–98. DOI: 10.1142/9789812839657_0001. [Online]. Available: https://doi.org/10.1142%5C%2F9789812839657_0001 (cit. on p. 15).

[37] A. DJOUADI, "The anatomy of electroweak symmetry breaking tome II: The higgs bosons in the minimal supersymmetric model," *Physics Reports*, vol. 459, no. 1-6, pp. 1–241, Apr. 2008. DOI: 10.1016/j.physrep.2007.10.005. [Online]. Available: https://doi.org/10.1016%5C%2Fj.physrep.2007.10.005 (cit. on p. 15).

[38] CERN, *Cern yellow reports: Monographs, vol 2 (2017): Handbook of lhc higgs cross sections: 4. deciphering the nature of the higgs sector*, en, 2017. DOI: 10.23731/CYRM-2017-002. [Online]. Available: https://e-publishing.cern.ch/index.php/CYRM/issue/view/32 (cit. on pp. 17–19, 49).

[39] C. Degrande, M. Ubiali, M. Wiesemann, and M. Zaro, *Heavy charged higgs boson production at the lhc*, 2015. DOI: 10.48550/ARXIV.1507.02549. [Online]. Available: https://arxiv.org/abs/1507.02549 (cit. on p. 18).

[40] M. Flechl, R. Klees, M. Kramer, M. Spira, and M. Ubiali, "Improved cross-section predictions for heavy charged higgs boson production at the LHC," *Physical Review D*, vol. 91, no. 7, Apr. 2015. DOI: 10.1103/physrevd.91.075015. [Online]. Available: https://doi.org/10.1103%5C%2Fphysrevd.91.075015 (cit. on p. 18).

[41] S. Dittmaier, M. Kramer, M. Spira, and M. Walser, "Charged-higgs-boson production at the LHC: Next-to-leading-order supersymmetric QCD corrections," *Physical Review D*, vol. 83, no. 5, Mar. 2011. DOI: 10.1103/physrevd.83.055005. [Online]. Available: https://doi.org/10.1103%5C%2Fphysrevd.83.055005 (cit. on p. 18).

[42] E. L. Berger, T. Han, J. Jiang, and T. Plehn, "Associated production of a top quark and a charged higgs boson," *Phys. Rev. D*, vol. 71, p. 115 012, 11 Jun. 2005. DOI: 10.1103/PhysRevD.71.115012. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevD.71.115012 (cit. on p. 18).

[43] C. Degrande, R. Frederix, V. Hirschi, M. Ubiali, M. Wiesemann, and M. Zaro, "Accurate predictions for charged higgs production: Closing the $m_{H^\pm} \sim m_t$ window," *Physics Letters B*, vol. 772, pp. 87–92, Sep. 2017. DOI: 10.1016/j.physletb.2017.06.037. [Online]. Available: https://doi.org/10.1016%5C%2Fj.physletb.2017.06.037 (cit. on p. 19).

[44] A. Djouadi and J. Quevillon, "The MSSM higgs sector at a high MSUSY: Reopening the low tan β regime and heavy higgs searches," *Journal of High Energy Physics*, vol. 2013, no. 10, Oct. 2013. DOI: 10.1007/jhep10(2013)028. [Online]. Available: https://doi.org/10.1007%5C%2Fjhep10%5C%282013%5C%29028 (cit. on p. 20).

[45] L. Maiani, A. Polosa, and V. Riquer, "Bounds to the higgs sector masses in minimal supersymmetry from LHC data," *Physics Letters B*, vol. 724, no. 4-5, pp. 274–277, Jul. 2013. DOI: 10.1016/j.physletb.2013.06.026. [Online]. Available: https://doi.org/10.1016%5C%2Fj.physletb.2013.06.026 (cit. on p. 20).

[46] A. Djouadi, L. Maiani, G. Moreau, A. Polosa, J. Quevillon, and V. Riquer, "The post-higgs MSSM scenario: Habemus MSSM?" *The European Physical Journal C*, vol. 73, no. 12, Nov. 2013. DOI: 10.1140/epjc/s10052-013-2650-0. [Online]. Available: https://doi.org/10.1140%5C%2Fepjc%5C%2Fs10052-013-2650-0 (cit. on p. 20).

[47] A. Djouadi, L. Maiani, A. Polosa, J. Quevillon, and V. Riquer, *Fully covering the mssm higgs sector at the lhc*, 2015. DOI: 10.48550/ARXIV.1502.05653. [Online]. Available: https://arxiv.org/abs/1502.05653 (cit. on p. 20).

[48] E. Bagnaschi *et al.*, "MSSM higgs boson searches at the LHC: Benchmark scenarios for run 2 and beyond," *The European Physical Journal C*, vol. 79, no. 7, Jul. 2019. DOI: 10.1140/epjc/s10052-019-7114-8. [Online]. Available: https://doi.org/10.1140%5C%2Fepjc%5C%2Fs10052-019-7114-8 (cit. on p. 20).

[49] P. Sanyal, "Limits on the charged higgs parameters in the two higgs doublet model using CMS $\sqrt{s}$ = 13 tev results," *The European Physical Journal C*, vol. 79, no. 11, Nov. 2019. DOI: 10.1140/epjc/s10052-019-7431-y. [Online]. Available: https://doi.org/10.1140%5C%2Fepjc%5C%2Fs10052-019-7431-y (cit. on pp. 20, 22–24).

[50] M. Aaboud *et al.*, "Search for charged Higgs bosons decaying via $H^{\pm} \to \tau^{\pm}\nu_{\tau}$ in the τ+jets and τ+lepton final states with 36 fb^{-1} of pp collision data recorded at $\sqrt{s} = 13$ TeV with the ATLAS experiment," *JHEP*, vol. 09, p. 139, 2018. DOI: 10.1007/JHEP09(2018)139. arXiv: 1807.07915 [hep-ex] (cit. on pp. 21, 22).

[51] 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," *JHEP*, vol. 07, p. 142, 2019. DOI: 10.1007/JHEP07(2019)142. arXiv: 1903.04560 [hep-ex] (cit. on pp. 21, 23).

[52] CMS Collaboration, "Search for a charged higgs boson decaying into top and bottom quarks in events with electrons or muons in proton-proton collisions at $\sqrt{s}$ = 13 tev," *Journal of High Energy Physics*, vol. 2020, no. 1, Jan. 2020. DOI: 10.1007/jhep01(2020)096. [Online]. Available: https://doi.org/10.1007%5C%2Fjhep01%5C%282020%5C%29096 (cit. on p. 21).

[53] CMS Collaboration, "Search for charged Higgs bosons decaying into a top and a bottom quark in the all-jet final state of pp collisions at $\sqrt{s}$ = 13 TeV," *JHEP*, vol. 07, p. 126, 2020. DOI: 10.1007/JHEP07(2020)126. arXiv: 2001.07763 [hep-ex] (cit. on pp. 21, 23, 130, 133, 149).

[54] ATLAS Collaboration, "Search for charged Higgs bosons decaying into a top quark and a bottom quark at $\sqrt{s}$ = 13 TeV with the ATLAS detector," *JHEP*, vol. 06, p. 145, 2021. DOI: 10.1007/JHEP06(2021)145. arXiv: 2102.10076 [hep-ex] (cit. on pp. 21, 23).

[55] Aleph, Delphi, L3, O. Collaborations, and t. L. w. g. f. H. b. searches the, *Search for charged higgs bosons: Combined results using lep data*, 2013. DOI: 10.48550/ARXIV.1301.6065. [Online]. Available: https://arxiv.org/abs/1301.6065 (cit. on p. 23).

[56] Heavy Flavor Averaging Group and A. et al., *Averages of b-hadron, c-hadron, and τ-lepton properties as of summer 2014*, 2014. DOI: 10.48550/ARXIV.1412.7515. [Online]. Available: https://arxiv.org/abs/1412.7515 (cit. on p. 23).

[57] L. Evans and P. Bryant, "LHC Machine," *Journal of Instrumentation*, vol. 3, no. 08, S08001, Aug. 2008. DOI: 10.1088/1748-0221/3/08/S08001. [Online]. Available: https://dx.doi.org/10.1088/1748-0221/3/08/S08001 (cit. on p. 26).

[58] L. Evans, "The large hadron collider," *New Journal of Physics*, vol. 9, no. 9, p. 335, Sep. 2007. DOI: 10.1088/1367-2630/9/9/335. [Online]. Available: https://dx.doi.org/10.1088/1367-2630/9/9/335 (cit. on p. 26).

[59] O. Bruning, H. Burkhardt, and S. Myers, "The Large Hadron Collider," *Progress in Particle and Nuclear Physics*, vol. 67, no. 3, pp. 705–734, 2012, ISSN: 0146-6410. DOI: https://doi.org/10.1016/j.ppnp.2012.03.001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0146641012000695 (cit. on pp. 26, 27).

[60] J.-L. Caron, "Cross section of LHC dipole. Dipole LHC: coupe transversale.," AC Collection. Legacy of AC. Pictures from 1992 to 2002., 1998. [Online]. Available: https://cds.cern.ch/record/841539 (cit. on p. 27).

[61] ATLAS Collaboration, "The ATLAS Experiment at the CERN Large Hadron Collider," *Journal of Instrumentation*, vol. 3, no. 08, S08003, Aug. 2008. DOI: 10.1088/1748-0221/3/08/S08003. [Online]. Available: https://dx.doi.org/10.1088/1748-0221/3/08/S08003 (cit. on p. 27).

[62] CMS Collaboration, "The CMS experiment at the CERN LHC," *Journal of Instrumentation*, vol. 3, no. 08, S08004, Aug. 2008. DOI: 10.1088/1748-0221/3/08/S08004. [Online]. Available: https://dx.doi.org/10.1088/1748-0221/3/08/S08004 (cit. on pp. 27, 30).

[63] ALICE Collaboration, "The ALICE experiment at the CERN LHC," *Journal of Instrumentation*, vol. 3, no. 08, S08002, Aug. 2008. DOI: 10.1088/1748-0221/3/08/S08002. [Online]. Available: https://dx.doi.org/10.1088/1748-0221/3/08/S08002 (cit. on p. 27).

[64] LHCb Collaboration, "The LHCb Detector at the LHC," *Journal of Instrumentation*, vol. 3, no. 08, S08005, Aug. 2008. DOI: 10.1088/1748-0221/3/08/S08005. [Online]. Available: https://dx.doi.org/10.1088/1748-0221/3/08/S08005 (cit. on p. 27).

[65] R. L. Workman *et al.*, "Review of Particle Physics: [Accelerator Physics and Technology: Colliders](https://pdg.lbl.gov/2022/reviews/rpp2022-rev-accel-phys-colliders.pdf)," *PTEP*, vol. 2022, p. 083C01, 2022. DOI: 10.1093/ptep/ptac097 (cit. on p. 28).

[66] B. Salvachua, "Overview of Proton-Proton Physics during Run 2," pp. 7–14, 2019. [Online]. Available: https://cds.cern.ch/record/2750272 (cit. on p. 29).

[67] CMS Collaboration, *The CMS magnet project: Technical Design Report* (Technical design report. CMS). Geneva: CERN, 1997. DOI: 10.17181/CERN.6ZU0.V4T9. [Online]. Available: https://cds.cern.ch/record/331056 (cit. on p. 32).

[68] CMS Collaboration, *The CMS tracker system project: Technical Design Report* (Technical design report. CMS). Geneva: CERN, 1997. [Online]. Available: https://cds.cern.ch/record/368412 (cit. on p. 32).

[69] CMS Collaboration, "The CMS Phase-1 Pixel Detector Upgrade," CERN, Geneva, Tech. Rep., Mar. 2020. [Online]. Available: https://cds.cern.ch/record/2745805 (cit. on p. 33).

[70] P. Azzurri, "The cms silicon strip tracker," *Journal of Physics: Conference Series*, vol. 41, no. 1, p. 127, May 2006. DOI: 10.1088/1742-6596/41/1/011. [Online]. Available: https://dx.doi.org/10.1088/1742-6596/41/1/011 (cit. on p. 33).

[71] CMS Collaboration, "CMS: The electromagnetic calorimeter. Technical design report," Dec. 1997 (cit. on p. 35).

[72] CMS Collaboration, "Energy resolution of the barrel of the CMS electromagnetic calorimeter," *JINST*, vol. 2, P04004, 2007. DOI: 10.1088/1748-0221/2/04/P04004 (cit. on p. 35).

[73] CMS Collaboration, "The CMS hadron calorimeter project: Technical Design Report," 1997. [Online]. Available: https://cds.cern.ch/record/357153 (cit. on p. 36).

[74] CMS Collaboration, "The CMS barrel calorimeter response to particle beams from 2-GeV/c to 350-GeV/c," *Eur. Phys. J. C*, vol. 60, pp. 359–373, 2009, [Erratum: Eur.Phys.J.C 61, 353–356 (2009)]. DOI: 10.1140/epjc/s10052-009-0959-5 (cit. on p. 37).

[75] CMS Collaboration, "CMS Technical Design Report for the Phase 1 Upgrade of the Hadron Calorimeter," J. Mans *et al.*, Eds., Sep. 2012. DOI: 10.2172/1151651 (cit. on p. 37).

[76] CMS Collaboration, *The CMS muon project: Technical Design Report* (Technical design report. CMS). Geneva: CERN, 1997. [Online]. Available: https://cds.cern.ch/record/343814 (cit. on p. 38).

[77] CMS Collaboration, *CMS TriDAS project: Technical Design Report, Volume 1: The Trigger Systems* (Technical design report. CMS). [Online]. Available: https://cds.cern.ch/record/706847 (cit. on p. 39).

[78] S. Cittolin, A. Racz, and P. Sphicas, *CMS The TriDAS Project: Technical Design Report, Volume 2: Data Acquisition and High-Level Trigger. CMS trigger and data-acquisition project* (Technical design report. CMS). Geneva: CERN, 2002. [Online]. Available: https://cds.cern.ch/record/578006 (cit. on p. 39).

[79] M. French *et al.*, "Design and results from the apv25, a deep sub-micron cmos front-end chip for the cms tracker," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 466, no. 2, pp. 359–365, 2001, 4th Int. Symp. on Development and Application of Semiconductor Tracking Detectors, ISSN: 0168-9002. DOI: https://doi.org/10.1016/S0168-9002(01)00589-7. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168900201005897 (cit. on p. 42).

[80] "Simulation of the Silicon Strip Tracker pre-amplifier in early 2016 data," 2020. [Online]. Available: https://cds.cern.ch/record/2740688 (cit. on p. 42).

[81] A. Buckley *et al.*, "General-purpose event generators for LHC physics," *Physics Reports*, vol. 504, no. 5, pp. 145–233, Jul. 2011. DOI: 10.1016/j.physrep.2011.03.005. [Online]. Available: https://doi.org/10.1016%5C%2Fj.physrep.2011.03.005 (cit. on pp. 43, 47).

[82] R. K. Ellis, W. J. Stirling, and B. R. Webber, *QCD and collider physics* (Cambridge monographs on particle physics, nuclear physics, and cosmology). Cambridge: Cambridge University Press, 2003, Photography by S. Vascotto. DOI: 10.1017/CBO9780511628788. [Online]. Available: https://cds.cern.ch/record/318585 (cit. on p. 43).

[83] R. D. Ball *et al.*, "Parton distributions from high-precision collider data," *The European Physical Journal C*, vol. 77, no. 10, Oct. 2017. DOI: 10.1140/epjc/s10052-017-5199-5. [Online]. Available: https://doi.org/10.1140%5C%2Fepjc%5C%2Fs10052-017-5199-5 (cit. on p. 44).

[84] G. Altarelli and G. Parisi, "Asymptotic freedom in parton language," *Nuclear Physics B*, vol. 126, no. 2, pp. 298–318, 1977, ISSN: 0550-3213. DOI: https://doi.org/10.1016/0550-3213(77)90384-4. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0550321377903844 (cit. on p. 44).

[85] Y. L. Dokshitzer, "Calculation of the Structure Functions for Deep Inelastic Scattering and e+ e- Annihilation by Perturbation Theory in Quantum Chromodynamics.," *Sov. Phys. JETP*, vol. 46, pp. 641–653, 1977 (cit. on p. 44).

[86] V. Gribov and L. Lipatov, "Deep inelastic electron scattering in perturbation theory," *Physics Letters B*, vol. 37, no. 1, pp. 78–80, 1971, ISSN: 0370-2693. DOI: https://doi.org/10.1016/0370-2693(71)90576-4. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0370269371905764 (cit. on p. 44).

[87] J. Alwall, M. Herquet, F. Maltoni, O. Mattelaer, and T. Stelzer, "MadGraph 5: Going beyond," *Journal of High Energy Physics*, vol. 2011, no. 6, Jun. 2011. DOI: 10.1007/jhep06(2011)128. [Online]. Available: https://doi.org/10.1007%5C%2Fjhep06%5C%282011%5C%29128 (cit. on p. 45).

[88] J. Alwall *et al.*, "The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations," *Journal of High Energy Physics*, vol. 2014, no. 7, Jul. 2014. DOI: 10.1007/jhep07(2014)079. [Online]. Available: https://doi.org/10.1007%5C%2Fjhep07%5C%282014%5C%29079 (cit. on p. 45).

[89] P. Nason, "A new method for combining NLO QCD with shower Monte Carlo algorithms," *JHEP*, vol. 11, p. 040, 2004. DOI: 10.1088/1126-6708/2004/11/040. arXiv: hep-ph/0409146 (cit. on p. 45).

[90] S. Frixione, P. Nason, and C. Oleari, "Matching NLO QCD computations with parton shower simulations: The POWHEG method," *JHEP*, vol. 11, p. 070, 2007. DOI: 10.1088/1126-6708/2007/11/070. arXiv: 0709.2092 [hep-ph] (cit. on p. 45).

[91] S. Alioli, P. Nason, C. Oleari, and E. Re, "A general framework for implementing NLO calculations in shower monte carlo programs: The POWHEG BOX," *JHEP*, vol. 06, p. 043, 2010. DOI: 10.1007/JHEP06(2010)043. arXiv: 1002.2581 [hep-ph] (cit. on p. 45).

[92] S. Alioli, K. Hamilton, P. Nason, C. Oleari, and E. Re, "Jet pair production in POWHEG," *JHEP*, vol. 04, p. 081, 2011. DOI: 10.1007/JHEP04(2011)081. arXiv: 1012.3380 [hep-ph] (cit. on p. 45).

[93] S. Alioli, P. Nason, C. Oleari, and E. Re, "NLO Higgs boson production via gluon fusion matched with shower in POWHEG," *JHEP*, vol. 04, p. 002, 2009. DOI: 10.1088/1126-6708/2009/04/002. arXiv: 0812.0578 [hep-ph] (cit. on p. 45).

[94] E. Bagnaschi, G. Degrassi, P. Slavich, and A. Vicini, "Higgs production via gluon fusion in the POWHEG approach in the SM and in the MSSM," *JHEP*, vol. 02, p. 088, 2012. DOI: 10.1007/JHEP02(2012)088. arXiv: 1111.2854 [hep-ph] (cit. on p. 45).

[95] T. Sjöstrand *et al.*, "An introduction to PYTHIA 8.2," *Comput. Phys. Commun.*, vol. 191, p. 159, 2015. DOI: 10.1016/j.cpc.2015.01.024. arXiv: 1410.3012 [hep-ph] (cit. on p. 45).

[96] M. Bahr *et al.*, "Herwig++ physics and manual," *The European Physical Journal C*, vol. 58, no. 4, pp. 639–707, Nov. 2008. DOI: 10.1140/epjc/s10052-008-0798-9. [Online]. Available: https://doi.org/10.1140%5C%2Fepjc%5C%2Fs10052-008-0798-9 (cit. on p. 47).

[97] S. Agostinelli et al., "Geant4-a simulation toolkit," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 506, no. 3, pp. 250–303, 2003, ISSN: 0168-9002. DOI: https://doi.org/10.1016/S0168-9002(03)01368-8. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168900203013688 (cit. on pp. 48, 53).

[98] CMS Collaboration, "Extraction and validation of a new set of CMS PYTHIA8 tunes from underlying-event measurements," *Eur. Phys. J. C*, vol. 80, no. 1, p. 4, 2020. DOI: 10.1140/epjc/s10052-019-7499-4. arXiv: 1903.12179 [hep-ex] (cit. on p. 48).

[99] P. Artoisenet, R. Frederix, O. Mattelaer, and R. Rietkerk, "Automatic spin-entangled decays of heavy resonances in Monte Carlo simulations," *JHEP*, vol. 03, p. 015, 2013. DOI: 10.1007/JHEP03(2013)015. arXiv: 1212.3460 [hep-ph] (cit. on p. 48).

[100] J. Gao *et al.*, "CT10 next-to-next-to-leading order global analysis of QCD," *Physical Review D*, vol. 89, no. 3, Feb. 2014. DOI: 10.1103/physrevd.89.033009. [Online]. Available: https://doi.org/10.1103%5C%2Fphysrevd.89.033009 (cit. on p. 49).

[101] M. Beneke, P. Falgari, S. Klein, and C. Schwinn, "Hadronic top-quark pair production with NNLL threshold resummation," *Nuclear Physics B*, vol. 855, no. 3, pp. 695–741, Feb. 2012. DOI: 10.1016/j.nuclphysb.2011.10.021. [Online]. Available: https://doi.org/10.1016%5C%2Fj.nuclphysb.2011.10.021 (cit. on p. 49).

[102] M. Cacciari, M. Czakon, M. Mangano, A. Mitov, and P. Nason, "Top-pair production at hadron colliders with next-to-next-to-leading logarithmic soft-gluon resummation," *Physics Letters B*, vol. 710, no. 4-5, pp. 612–622, Apr. 2012. DOI: 10.

1016/j.physletb.2012.03.013. [Online]. Available: https://doi.org/10.1016%5C%2Fj.physletb.2012.03.013 (cit. on p. 49).

[103] P. Barnreuther, M. Czakon, and A. Mitov, "Percent-level-precision physics at the tevatron: Next-to-next-to-leading order QCD corrections to $q\bar{q} \to t\bar{t} + X$," *Physical Review Letters*, vol. 109, no. 13, Sep. 2012. DOI: 10.1103/physrevlett.109.132001. [Online]. Available: https://doi.org/10.1103%5C%2Fphysrevlett.109.132001 (cit. on p. 49).

[104] M. Czakon and A. Mitov, "NNLO corrections to top-pair production at hadron colliders: The all-fermionic scattering channels," *Journal of High Energy Physics*, vol. 2012, no. 12, Dec. 2012. DOI: 10.1007/jhep12(2012)054. [Online]. Available: https://doi.org/10.1007%5C%2Fjhep12%5C%282012%5C%29054 (cit. on p. 49).

[105] M. Czakon and A. Mitov, "NNLO corrections to top pair production at hadron colliders: The quark-gluon reaction," *Journal of High Energy Physics*, vol. 2013, no. 1, Jan. 2013. DOI: 10.1007/jhep01(2013)080. [Online]. Available: https://doi.org/10.1007%5C%2Fjhep01%5C%282013%5C%29080 (cit. on p. 49).

[106] M. Czakon, P. Fiedler, and A. Mitov, "Total top-quark pair-production cross section at hadron colliders through $\mathcal{O}(\alpha_s^4)$," *Physical Review Letters*, vol. 110, no. 25, Jun. 2013. DOI: 10.1103/physrevlett.110.252004. [Online]. Available: https://doi.org/10.1103%5C%2Fphysrevlett.110.252004 (cit. on p. 49).

[107] M. Czakon and A. Mitov, "Top++: A Program for the Calculation of the Top-Pair Cross-Section at Hadron Colliders," *Comput. Phys. Commun.*, vol. 185, p. 2930, 2014. DOI: 10.1016/j.cpc.2014.06.021. arXiv: 1112.5675 [hep-ph] (cit. on p. 49).

[108] M. Botje *et al.*, *The PDF4LHC Working Group Interim Recommendations*, 2011. arXiv: 1101.0538 [hep-ph] (cit. on p. 49).

[109] A. D. Martin, W. J. Stirling, R. S. Thorne, and G. Watt, "Parton distributions for the LHC," *The European Physical Journal C*, vol. 63, no. 2, pp. 189–285, Jul. 2009. DOI: 10.1140/epjc/s10052-009-1072-5. [Online]. Available: https://doi.org/10.1140%5C%2Fepjc%5C%2Fs10052-009-1072-5 (cit. on p. 49).

[110] A. D. Martin, W. J. Stirling, R. S. Thorne, and G. Watt, "Uncertainties on α_S in global PDF analyses and implications for predicted hadronic cross sections," *The European Physical Journal C*, vol. 64, no. 4, pp. 653–680, Oct. 2009. DOI: 10.1140/epjc/s10052-009-1164-2. [Online]. Available: https://doi.org/10.1140%5C%2Fepjc%5C%2Fs10052-009-1164-2 (cit. on p. 49).

[111] R. D. Ball *et al.*, "Parton distributions with LHC data," *Nuclear Physics B*, vol. 867, no. 2, pp. 244–289, Feb. 2013. DOI: 10.1016/j.nuclphysb.2012.10.003. [Online]. Available: https://doi.org/10.1016%5C%2Fj.nuclphysb.2012.10.003 (cit. on p. 49).

[112] H.-L. Lai *et al.*, "New parton distributions for collider physics," *Physical Review D*, vol. 82, no. 7, Oct. 2010. DOI: 10.1103/physrevd.82.074024. [Online]. Available: https://doi.org/10.1103%5C%2Fphysrevd.82.074024 (cit. on p. 49).

[113] H.-L. Lai *et al.*, "New parton distributions for collider physics," *Physical Review D*, vol. 82, no. 7, Oct. 2010. DOI: 10.1103/physrevd.82.074024. [Online]. Available: https://doi.org/10.1103%5C%2Fphysrevd.82.074024 (cit. on p. 49).

[114] M. Aliev, H. Lacker, U. Langenfeld, S. Moch, P. Uwer, and M. Wiedermann, "HATHOR – HAdronic top and heavy quarks crOss section calculatoR," *Computer Physics Communications*, vol. 182, no. 4, pp. 1034–1046, Apr. 2011. DOI: 10.1016/j.cpc.2010.12.040. [Online]. Available: https://doi.org/10.1016%5C%2Fj.cpc.2010.12.040 (cit. on p. 49).

[115] P. Kant *et al.*, "HatHor for single top-quark production: Updated predictions and uncertainty estimates for single top-quark production in hadronic collisions," *Computer Physics Communications*, vol. 191, pp. 74–89, Jun. 2015. DOI: 10.1016/j.cpc.2015.02.001. [Online]. Available: https://doi.org/10.1016%5C%2Fj.cpc.2015.02.001 (cit. on p. 49).

[116] N. Kidonakis, *Top quark production*, 2013. arXiv: 1311.0283 [hep-ph] (cit. on p. 49).

[117] CMS Twiki, *How to compute cross sections with the genxsecanalyzer*, Available at https://twiki.cern.ch/twiki/bin/view/CMS/HowToGenXSecAnalyzer#Running_the_GenXSecAnalyzer_on_a. Revision 17. (cit. on p. 49).

[118] A. Kulesza, L. Motyka, D. Schwartlander, T. Stebel, and V. Theeuwes, "Associated production of a top quark pair with a heavy electroweak gauge boson at NLO+NNLL accuracy," *The European Physical Journal C*, vol. 79, no. 3, Mar. 2019. DOI: 10.1140/epjc/s10052-019-6746-z. [Online]. Available: https://doi.org/10.1140%5C%2Fepjc%5C%2Fs10052-019-6746-z (cit. on p. 49).

[119] J. M. Campbell, R. K. Ellis, and C. Williams, "Vector boson pair production at the LHC," *Journal of High Energy Physics*, vol. 2011, no. 7, Jul. 2011. DOI: 10.1007/jhep07(2011)018. [Online]. Available: https://doi.org/10.1007%5C%2Fjhep07%5C%282011%5C%29018 (cit. on p. 49).

[120] A. M. Sirunyan *et al.*, "Particle-flow reconstruction and global event description with the CMS detector," *JINST*, vol. 12, P10003, 2017. DOI: 10.1088/1748-0221/12/10/P10003. arXiv: 1706.04965 [physics.ins-det] (cit. on p. 50).

[121] W. Adam, B. Mangano, T. Speer, and T. Todorov, "Track Reconstruction in the CMS tracker," CERN, Geneva, Tech. Rep., 2006. [Online]. Available: https://cds.cern.ch/record/934067 (cit. on p. 50).

[122] K. Rose, "Deterministic annealing for clustering, compression, classification, regression, and related optimization problems," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2210–2239, 1998. DOI: 10.1109/5.726788 (cit. on p. 50).

[123] W. Adam, R. Fruhwirth, A. Strandlie, and T. Todorov, "Reconstruction of Electrons with the Gaussian-Sum Filter in the CMS traker at LHC," CERN, Geneva, Tech. Rep. N9, 2003, Talk from the 2003 Computing in High Energy and Nuclear Physics (CHEP03), La Jolla, Ca, USA, March 2003, LaTeX, 14 eps figures. PSN TULT009. DOI: 10.1088/0954-3899/31/9/N01. [Online]. Available: https://cds.cern.ch/record/687891 (cit. on p. 51).

[124] J. L. Bentley, "Multidimensional binary search trees used for associative searching," *Communications of the ACM*, vol. 18, no. 9, pp. 509–517, 1975 (cit. on p. 53).

[125] "Performance of photon reconstruction and identification with the cms detector in proton-proton collisions at $\sqrt{s}$=8 tev," *Journal of Instrumentation*, vol. 10, no. 08, P08010, Aug. 2015. DOI: 10.1088/1748-0221/10/08/P08010. [Online]. Available: https://dx.doi.org/10.1088/1748-0221/10/08/P08010 (cit. on p. 53).

[126] M. Cacciari, G. P. Salam, and G. Soyez, "The anti-kt jet clustering algorithm," *Journal of High Energy Physics*, vol. 2008, no. 04, p. 063, 2008 (cit. on pp. 56, 57).

[127] M. Cacciari and G. P. Salam, "Dispelling the N^3 myth for the k_t jet-finder," *Phys. Lett.*, vol. B641, pp. 57–61, 2006. DOI: 10.1016/j.physletb.2006.08.037. arXiv: hep-ph/0512210 [hep-ph] (cit. on pp. 56, 70).

[128] M. Cacciari, G. P. Salam, and G. Soyez, "FastJet User Manual," *Eur. Phys. J.*, vol. C72, p. 1896, 2012. DOI: 10.1140/epjc/s10052-012-1896-2. arXiv: 1111.6097 [hep-ph] (cit. on pp. 56, 70).

[129] S. D. Ellis and D. E. Soper, "Successive combination jet algorithm for hadron collisions," *Physical Review D*, vol. 48, no. 7, pp. 3160–3166, Oct. 1993. DOI: 10.1103/physrevd.48.3160. [Online]. Available: https://doi.org/10.1103%5C%2Fphysrevd.48.3160 (cit. on p. 56).

[130] Y. Dokshitzer, G. Leder, S. Moretti, and B. Webber, "Better jet clustering algorithms," *Journal of High Energy Physics*, vol. 1997, no. 08, pp. 001–001, Aug. 1997. DOI: 10.1088/1126-6708/1997/08/001. [Online]. Available: https://doi.org/10.1088%5C%2F1126-6708%5C%2F1997%5C%2F08%5C%2F001 (cit. on p. 56).

[131] G. P. Salam and G. Soyez, "A practical seedless infrared-safe cone jet algorithm," *Journal of High Energy Physics*, vol. 2007, no. 05, pp. 086–086, May 2007. DOI: 10.1088/1126-6708/2007/05/086. [Online]. Available: https://doi.org/10.1088%5C%2F1126-6708%5C%2F2007%5C%2F05%5C%2F086 (cit. on p. 56).

[132] V. Khachatryan, A. Sirunyan, A. Tumasyan, W. Adam, and E. A. et. at, "Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV," *Journal of Instrumentation*, vol. 12, no. 02, P02014–P02014, Feb. 2017. DOI: 10.1088/1748-0221/12/02/p02014. [Online]. Available: https://doi.org/10.1088%5C%2F1748-0221%5C%2F12%5C%2F02%5C%2Fp02014 (cit. on p. 57).

[133] "Jet energy scale and resolution measurement with Run 2 Legacy Data Collected by CMS at 13 TeV, 2021," 2021. [Online]. Available: https://cds.cern.ch/record/2792322 (cit. on p. 59).

[134] E. Bols, J. Kieseler, M. Verzetti, M. Stoye, and A. Stakia, "Jet flavour classification using DeepJet," *JINST*, vol. 15, P12012, 2020. DOI: 10.1088/1748-0221/15/12/P12012. arXiv: 2008.10519 [hep-ex] (cit. on p. 59).

[135] D. Guest, J. Collado, P. Baldi, S.-C. Hsu, G. Urban, and D. Whiteson, "Jet Flavor Classification in High-Energy Physics with Deep Neural Networks," *Phys. Rev. D*, vol. 94, no. 11, p. 112 002, 2016. DOI: 10.1103/PhysRevD.94.112002. arXiv: 1607.08633 [hep-ex] (cit. on pp. 59, 94).

[136] CMS Collaboration, "Performance of the DeepJet b tagging algorithm using 41.9 fb^{-1} of data from proton-proton collisions at 13 TeV with Phase 1 CMS detector," Nov. 2018. [Online]. Available: https://cds.cern.ch/record/2646773 (cit. on p. 60).

[137] V. Khachatryan *et al.*, "Reconstruction and identification of τ lepton decays to hadrons and ν_τ at CMS," *JINST*, vol. 11, P01019, 2016. DOI: 10.1088/1748-0221/11/01/P01019. arXiv: 1510.07488 [physics.ins-det] (cit. on p. 60).

[138] A. M. Sirunyan *et al.*, "Performance of reconstruction and identification of τ leptons decaying to hadrons and ν_τ in pp collisions at $\sqrt{s} = 13$ TeV," *JINST*, vol. 13, P10005, 2018. DOI: 10.1088/1748-0221/13/10/P10005. arXiv: 1809.02816 [hep-ex] (cit. on p. 60).

[139] A. M. Sirunyan *et al.*, "Identification of hadronic tau lepton decays using a deep neural network," Submitted to *JINST*, 2022. arXiv: 2201.08458 [hep-ex] (cit. on p. 60).

[140] T. M. Mitchell *et al.*, *Machine learning*. McGraw-hill New York, 2007, vol. 1 (cit. on p. 62).

[141] G. Karagiorgi, G. Kasieczka, S. Kravitz, B. Nachman, and D. Shih, *Machine learning in the search for new fundamental physics*, 2021. DOI: 10.48550/ARXIV.2112.03769. [Online]. Available: https://arxiv.org/abs/2112.03769 (cit. on p. 62).

[142] K. Albertsson *et al.*, *Machine learning in high energy physics community white paper*, 2018. DOI: 10.48550/ARXIV.1807.02876. [Online]. Available: https://arxiv.org/abs/1807.02876 (cit. on p. 62).

[143] D. Bourilkov, "Machine and deep learning applications in particle physics," *International Journal of Modern Physics A*, vol. 34, no. 35, p. 1 930 019, Dec. 2019. DOI: 10.1142/s0217751x19300199. [Online]. Available: https://doi.org/10.1142%5C%2Fs0217751x19300199 (cit. on p. 62).

[144] Y. B. Ian Goodfellow and A. Courville, "Deep learning," Book in preparation for MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org (cit. on pp. 63, 70).

[145] S. Haykin, *Neural networks: a comprehensive foundation*. Prentice Hall PTR, 1994 (cit. on pp. 63, 70).

[146] A. L. Read, "Presentation of search results: The CL_s technique," *J. Phys. G*, vol. 28, p. 2693, 2002. DOI: 10.1088/0954-3899/28/10/313 (cit. on p. 66).

[147] T. Junk, "Confidence level computation for combining searches with small statistics," *Nucl. Instrum. Meth. A*, vol. 434, p. 435, 1999. DOI: 10.1016/S0168-9002(99)00498-2. arXiv: hep-ex/9902006 [hep-ex] (cit. on p. 66).

[148] "Procedure for the LHC Higgs boson search combination in Summer 2011," CERN, Geneva, Tech. Rep., 2011. [Online]. Available: https://cds.cern.ch/record/1379837 (cit. on pp. 66, 68).

[149] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, "Asymptotic formulae for likelihood-based tests of new physics," *Eur. Phys. J. C*, vol. 71, p. 1554, 2011. DOI: 10.1140/epjc/s10052-011-1554-0. arXiv: 1007.1727 [physics.data-an] (cit. on p. 67).

[150] K. Cranmer, "Statistical challenges for searches for new physics at the lhc," in *Statistical problems in particle physics, astrophysics and cosmology*, World Scientific, 2006, pp. 112–123 (cit. on p. 67).

[151] E. Bols, J. Kieseler, M. Verzetti, M. Stoye, and A. Stakia, "Jet flavour classification using DeepJet," *JINST*, vol. 8, P12012–P12012, Dec. 2020. DOI: 10.1088/1748-0221/15/12/p12012. arXiv: 2008.10519. [Online]. Available: https://doi.org/10.1088/1748-0221/15/12/p12012 (cit. on p. 70).

[152] "Performance of quark/gluon discrimination in 8 TeV pp data," CERN, Geneva, Tech. Rep., 2013. [Online]. Available: https://cds.cern.ch/record/1599732 (cit. on pp. 72, 73).

[153] A. J. Larkoski, S. Marzani, G. Soyez, and J. Thaler, "Soft drop," *Journal of High Energy Physics*, vol. 2014, no. 5, p. 146, May 2014, ISSN: 1029-8479. DOI: 10.1007/JHEP05(2014)146. [Online]. Available: https://doi.org/10.1007/JHEP05(2014)146 (cit. on p. 75).

[154] F. Chollet *et al.*, *Keras*, https://keras.io, 2015 (cit. on pp. 77, 111).

[155] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous systems," Software available from tensorflow.org, 2015. arXiv: 1603.04467. [Online]. Available: https://www.tensorflow.org/ (cit. on pp. 77, 111).

[156] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, 2014. arXiv: 1412.6980 [cs.LG] (cit. on pp. 77, 112).

[157] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011 (cit. on p. 78).

[158] CMS Twiki, *Jet energy corrections: Official software tools for applying jec corrections and uncertainties*, Available at https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookJetEnergyCorrections. (cit. on p. 85).

[159] CMS Twiki, *Jer scaling factors and uncertainty for 13 tev (2015 and 2016)*, Available at https://twiki.cern.ch/twiki/bin/viewauth/CMS/JetResolution#JER_Scaling_factors_and_Uncertai. (cit. on p. 85).

[160] CMS Collaboration, "Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV," *JINST*, vol. 13, no. 05, P05011, 2018. DOI: 10.1088/1748-0221/13/05/P05011. arXiv: 1712.07158 [physics.ins-det] (cit. on p. 94).

[161] A. M. Sirunyan *et al.*, "Performance of missing transverse momentum reconstruction in proton-proton collisions at $\sqrt{s} = 13$ TeV using the CMS detector," *JINST*, vol. 14, P07004, 2019. DOI: 10.1088/1748-0221/14/07/P07004. arXiv: 1903.06078 [hep-ex] (cit. on p. 95).

[162] "Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at $\sqrt{s}$=13 tev," *Journal of Instrumentation*, vol. 13, no. 06, P06015–P06015, Jun. 2018. DOI: 10.1088/1748-0221/13/06/p06015. [Online]. Available: https://doi.org/10.1088%5C%2F1748-0221%5C%2F13%5C%2F06%5C%2Fp06015 (cit. on p. 95).

[163] P. Baldi, K. Cranmer, T. Faucett, P. Sadowski, and D. Whiteson, "Parameterized machine learning for high-energy physics," *arXiv:1601.07913*, 2016 (cit. on p. 106).

[164] A. M. Sirunyan *et al.*, "Precision luminosity measurement in proton-proton collisions at $\sqrt{s} = 13$ TeV in 2015 and 2016 at CMS," *Eur. Phys. J. C*, vol. 81, p. 800, 2021. DOI: 10.1140/epjc/s10052-021-09538-2. arXiv: 2104.01927 [hep-ex] (cit. on p. 114).

[165] 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. [Online]. Available: https://cds.cern.ch/record/2621960/ (cit. on p. 114).

[166] 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. [Online]. Available: https://cds.cern.ch/record/2676164/ (cit. on p. 114).

[167] CMS Twiki, *Estimating systematic errors due to pileup modeling*, Available at https://twiki.cern.ch/twiki/bin/view/CMS/PileupSystematicErrors. Revision 11. (cit. on p. 114).

[168] H. Qu and L. Gouskos, "Jet tagging via particle clouds," *Phys. Rev. D*, vol. 101, p. 056 019, 5 Mar. 2020. DOI: 10.1103/PhysRevD.101.056019. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevD.101.056019 (cit. on p. 137).

[169] C. Collaboration, "Identification of heavy, energetic, hadronically decaying particles using machine-learning techniques," *Journal of Instrumentation*, vol. 15, no. 06, P06005, Jun. 2020. DOI: 10.1088/1748-0221/15/06/P06005. [Online]. Available: https://dx.doi.org/10.1088/1748-0221/15/06/P06005 (cit. on p. 137).

[170] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, *Dynamic graph cnn for learning on point clouds*, 2019. arXiv: 1801.07829 [cs.CV] (cit. on p. 138).

[171] CMS Collaboration, "Identification of highly Lorentz-boosted heavy particles using graph neural networks and new mass decorrelation techniques," 2020. [Online]. Available: https://cds.cern.ch/record/2707946 (cit. on p. 140).

[172] CMS Collaboration, "Mass regression of highly-boosted jets using graph neural networks," 2021. [Online]. Available: https://cds.cern.ch/record/2777006 (cit. on pp. 140, 141).

[173] J. Dolen, P. Harris, S. Marzani, S. Rappoccio, and N. Tran, "Thinking outside the ROCs: Designing decorrelated taggers (DDT) for jet substructure," *Journal of High Energy Physics*, vol. 2016, no. 5, May 2016. DOI: 10.1007/jhep05(2016)156. [Online]. Available: https://doi.org/10.1007%5C%2Fjhep05%5C%282016%5C%29156 (cit. on p. 141).