Inclusive Search for a Lorentz-Boosted Higgs Boson Decaying into Bottom Quarks

A dissertation

submitted to the department of Physics

of Brown University

in partial fulfillment of the requirements

for the degree of

doctor of philosophy

Martin Kwok

October,2020

CERN-THESIS-2020-214
30/04/2020



© Copyright by Martin Kwok 2020

All Rights Reserved

This dissertation by Ka Hei Martin Kwok is accepted in its present form by the Department of Physics as satisfying the dissertation requirement for the degree of Doctor of Philosophy.

Date	Greg Landsberg Principal Adviser

Recommended to the Graduate Council

Date	David Cutts, Reader
Date	Jiji Fan, Reader

Approved by the Graduate Council

Date	Andrew G. Campbell, Dean of the Graduate School

Curriculum Vitae

Professional Experience

2019 - Present	LHC Physics Center Graduate Scholar, Fermilab
2016 - Present	Active member on CMS HCAL operation team, on-call DAQ expert
	CMS HCAL online software developer
2015 Summer	CMS HCAL Prompt Feed Back Group(PFG) member
2014-2015	Teaching Assistant, Brown University
2012-2014	Teaching Assistant, Chinese University of Hong Kong
2011	CERN Summer Student 2011

Education

2014 - Present	Ph. D. Physics, Brown University, Providence, R.I.
	Expected graduation: Summer, 2020
2012 - 2014	M. Phil. Physics, Chinese University of Hong Kong (CUHK)
	Thesis topic: "Sensitivity study of the Higgs decay channel $WH \to lv\bar{b}\bar{b}$ using the
	ATLAS detector in LHC Run II and Study of surface cosmic rays"
	Advisor: Prof. M. C. Chu
2009-2012	B.S. Physics, Chinese University of Hong Kong (CUHK)

Publication

2019 Apr	CMS Collaboration. Inclusive search for a highly boosted Higgs boson decaying
	to a bottom quark-antiquark pair at $\sqrt{s}$=13 TeV with 137 fb^{-1}, cds.cern.ch/record/2714916
2018 May	CMS Collaboration. Search for black holes and sphalerons in high-multiplicity
	final states in proton-proton collisions at $\sqrt{s}$ = 13 TeV, arXiv:1805.06013
2017 May	CMS Collaboration. Search for black holes and other new phenomena in high-multiplicity
	final states in proton-proton collisions at $\sqrt{s}$ = 13 TeV, Phys. Lett. B 774 (2017) 279

Conferences and Courses

2019 Dec	LPC Physics Forum, Fermilab
	Title: Search for new physics with boosted Higgs boson
2019 Aug	QCD at LHC, Buffalo, NY
	Title: VH(bb) and search for boosted ggH→bb in CMS
2016-2018	CMS Exotica Physics Analysis Group Workshop
2012	1st Asia-Europe-Pacific School of High Energy Physics in Fukuoka

Awards

2015	Award of Excellence as a Graduate Teaching Assistant, Brown U.
2013	Best Teaching Assistant Award, CUHK
2011	Faculty of Science Dean's List, CUHK

Preface and Acknowledgments

I embarked on this journey to pursue a Ph.D. in particle physics 6 years ago, with a humble heart, to learn what Mother Nature has in store for us. Although this journey has been at times lonely or even frustrating, ultimately it has been a gratifying experience, especially when it has brought me close to some of the most amazing people in my life, whom I must thank for helping me along the way.

First and foremost, I would like to thank my advisor, Greg Landsberg, who always shares his wisdom and vision with me when the path ahead is unclear. Thank you for being a role model for me to learn from and for all the opportunities you have introduced me to. I would like to thank the readers of my thesis, David Cutts and Jiji Fan, for providing valuable feedback.

I would also like to thank the graduate school classmates who provided great company during my relative short time at Brown, especially, John Leung, Ken Ma, Matthew Huang, Mary Hadley, Igor Prlina and Brian Kilpatrick (and his family).

It was an absolute privilege to have spent 3 years at CERN working with the CMS HCAL operation team. Thank you Josh Kunkle, Andrea Benaglia, John Hakala, Andrew Buccilli, Caleb Smith and Andrew Evans for staying up many times late

in the night with me, ensuring HCAL is up and running. I would like to thank Paolo Rumerio and Jay Dittmann for always appreciating the opertion team's effort including my own. And I have to especially thank my mentors and friends, Seth Cooper and Edward Liard, who taught me much more beyond software and HCAL knowledge. Finally, my life at CERN would not be complete without all the inspiring discussions with David Yu, Jared Sturdy, Brian Dorney, Rachel Bartek, Adam Davis and all other members of the YALT group. To Lucian Lo, Yi Ting Duh and Chia Yu Hsieh, thank you for always being my friends when I needed you.

To my colleague and friends at Fermilab, especially Cristina Mantilla, the DAZSLE group has been a very fun and vibrant group to work with. I have to thank Javier Duarte and Nick Smith for being patient teachers of all the technical work involved in the boosted Higgs analysis. I am fortunate to have met Nhan Tran and Jim Hirschauer, who has served as my mentor and supervisor at Fermilab.

To my special friends in Chicago, Clare Huang and Andy Li, thank you for being the best possible friends and the emotional support during the last mile of my Ph.D. journey. Thank you Alice Wong, without whom I would not have been able to finish this thesis during quarantine. I will be forever grateful for the joy and inspiration you have brought into my life.

Last but not least, I cannot express enough gratitude to my parents and family, who provided endless support for me during all these years despite being far away.

Contents

Preface and Acknowledgments **vi**

1 Introduction **1**

2 The Standard Model **3**

2.1 Particle content 3

2.1.1 Bosons 5

2.1.2 Fermions 7

2.2 Local gauge invariance 9

2.2.1 The Quantum Electrodynamics 9

2.2.2 Yang-Mills Theory 11

2.3 Quantum Chromodynamics 13

2.4 Electroweak theory 15

2.4.1 Higgs Mechanism 18

2.5 Higgs boson at the LHC 23

2.5.1 Production and decay 23

2.5.2 Fermionic couplings 24

2.5.3 Boosted Higgs boson 28

3 The Compact Muon Solenoid Experiment 32

3.1 The Large Hadron Collider 33

3.2 The CMS detector 38

3.2.1 Coordinate System 41

3.2.2 The Pixel Detector 41

3.2.3 Silicon Strip Tracker 42

3.2.4 Electromagnetic Calorimeter (ECAL) 44

3.2.5 Hadronic Calorimeter (HCAL) 46

3.2.6 Muon System 54

3.2.7 Trigger System 58

3.3 Physics Objects Reconstruction 59

3.3.1 Particle-flow algorithm 61

3.3.2 Muons 65

3.3.3 Electrons and Photons 65

3.3.4 Hadrons 67

3.3.5 Jets 67

3.3.6 Wide jets with substructures 70

3.3.7 Jet ρ 73

3.3.8 Jet substructure 74

3.3.9 Missing transverse energy 77

3.3.10 Identification of b-quark jets 78

4 Search for boosted $H \to b\bar{b}$ 85

4.1 Introduction . . . 85

4.1.1 Analysis strategy . . . 86

4.2 Data and simulated samples . . . 87

4.2.1 Data . . . 87

4.2.2 Background simulations . . . 88

4.2.3 Signal simulations . . . 92

4.3 Triggers . . . 98

4.4 Objects reconstruction . . . 102

4.4.1 Jets . . . 104

4.4.2 Jet mass . . . 107

4.4.3 Jet ρ . . . 109

4.4.4 Jet substructure . . . 110

4.4.5 Deep double-b tagger . . . 115

4.4.6 Lepton veto . . . 117

4.4.7 Additional AK4 b jets . . . 118

4.4.8 Missing transverse energy . . . 118

4.5 Event selections . . . 120

4.5.1 Higgs candidate jet selection . . . 121

4.5.2 Substructure and deep double-b optimization . . . 122

4.5.3 Data and simulation comparison . . . 127

4.6 Background estimation . . . 131

4.6.1 V+jets estimate . . . 131
4.6.2 $t\bar{t}$ estimate . . . 133
4.6.3 QCD multijet estimation . . . 143
4.7 Full simultaneous fit . . . 147
4.7.1 Polynomial order and the F-test . . . 152
4.7.2 Fit results with simulated QCD events . . . 157
4.7.3 Signal injection tests based on data best fit . . . 166
4.7.4 Empirical function fit sensitivity . . . 167
4.8 Systematic uncertainties . . . 168
4.8.1 Mass shapes in QCD . . . 168
4.8.2 Mass scale, resolution and substructure scale factors . . . 170
4.8.3 Deep double-b tagging . . . 173
4.8.4 Higgs signal theory uncertainties . . . 175
4.8.5 Theory uncertainties for W/Z+jets . . . 177
4.8.6 JER and JES . . . 178
4.8.7 MC statistics . . . 178
4.8.8 AK4 b-tagging . . . 179
4.8.9 Jet trigger . . . 179
4.8.10 Muon trigger/ID/Isolation efficiency . . . 180
4.8.11 Correlation of systematic across different years . . . 182
4.8.12 Summary of systematic uncertainties . . . 184
4.9 Results . . . 185

4.9.1 Mass spectrum 185
4.9.2 Goodness of fit in data 186
4.9.3 Profile Likelihoods 189
4.9.4 Signal strengths and significance 193
4.9.5 Channel compatibility 195
4.9.6 Fiducial cross section 198
4.10 Summary 204

A List of triggers **205**

B Full m_{SD} distribution **207**

C Full transfer factors **215**

Bibliography **218**

List of Tables

2.1 Coupling strength of fundamental interactions in the SM. 6

3.1 Parameters of instantaneous luminosity equation at the design peak luminosity of $1.0 \times 10^{34}\text{cm}^{-2}\text{s}^{-1}$ [32]. 34

4.1 Datasets used for main analysis. Data corresponds to the ReReco processing unless specified as prompt reco. The integrated luminosity and the run-ranges are shown for each data period. 89

4.2 List of background MC samples. H_{T} denotes the scalar sum of transverse momenta of all jets in the events. Datasets marked with * are LO in QCD and EWK but will have NLO corrections applied. 90

4.3 Uncertainty of different components of NLO k-factors applied to W and Z+jets re-weighting as a function of p_{T} of the vector boson. . . . 92

4.4 List of signal MC samples. In all samples, the Higgs boson is required to decay to $\text{b}\bar{\text{b}}$. Cross sections do not include $\text{H} \to \text{b}\bar{\text{b}}$ branching ratio. All samples are generated with the POWHEG generator except for the HJ-MiNLO gluon fusion sample. 95

4.5 Lepton veto requirements, see text for description of isolation variables. 117

4.6 Relative $S/\sqrt{B}$ values (normalized to the best $S/\sqrt{B}$) for different $N_2^{1,\mathrm{DDT}}$ and deep double-b cuts. 123

4.7 Expected total simulated signal and background yields remaining after each of the jet- and event-level selections applied in this analysis. The jet signal region kinematic selection is $450 \leq p_{\mathrm{T}} < 1200\,\mathrm{GeV}$, $47 \leq m_{\mathrm{SD}} < 201\,\mathrm{GeV}$. Yields correspond to the 2017 data-taking conditions and integrated luminosity of $41.1\,\mathrm{fb}^{-1}$. 125

4.8 Summary of $\mathrm{t\bar{t}}$ scale factors measured in single muon control region for different data-taking years. 136

4.9 Order determined by F-test and corresponding GOF p-value for the empirical function fit in each p_{T} category 167

4.10 Expected results based on an Asimov data set generated from 2017 MC. 168

4.11 Summary of mass scale/resolution and $N_2^{1,\mathrm{DDT}}$ scale factors measured in semileptonic $\mathrm{t\bar{t}}$ control region for different data-taking years. . . . 173

4.12 Summary of measured and applied deep double b scale factors for different data-taking years. 175

4.13 Summary of nuisance parameters with prior constraints specific to this analysis and their affected processes. 184

4.14 Major sources of uncertainty in the measurement of the signal strength μ_{H} based on the HJ-MiNLO prediction, and their observed impact ($\Delta\mu_{\mathrm{H}}$) from a fit to the combined data set, are listed. The total uncertainty is separated into three components: statistical, systematic, and theory. Detailed decompositions of the statistical, systematic, and theory components are specified. The impact of each uncertainty is evaluated considering only that source. The sum in quadrature for each source does not in general equal the total uncertainty of each component because of correlations in the combined fit between nuisance parameters in different sources. 185

4.15 Fitted signal strength, expected and observed significance of the Higgs and Z boson signals. The Higgs boson results are presented with two gluon fusion signal models, one using the nominal HJ-MiNLO sample and the other simulated with the same procedure described in Ref. [1]. The 95% confidence level upper limit (UL) on the Higgs boson signal strength is also listed. In the results for the Higgs boson, the Z boson yield is fixed to the SM prediction value with the corresponding theoretical uncertainties to better constrain *in situ* the data-to-simulation scale factor of the deep double-b tagger. For the expected and observed signal strength of Z boson, the Higgs boson signal strength is freely floating. 196

4.16 Measured and predicted differential fiducial cross section as a function of Higgs boson p_{T}. All cross sections are in units of fb. 201

4.17 Correlation coefficients between the three $p_{\mathrm{T}}^{\mathrm{H}}$ bins of the unfolded Higgs boson differential cross section measurement. 201

A.1 A description of the triggers used in this analysis. 206

List of Figures

2.1 Particle content of the standard model. 4

2.2 Sketch of the Higgs potential $V(\phi)$ [10]. 19

2.3 Main leading order Feynman diagrams contributing to the Higgs boson production in (a) gluon fusion, (b) Vector-boson fusion, (c, d) associated production with W (tree-level, loop-level) and (e) associated production with a pair of top quarks [15]. 24

2.4 Minimal SM Higgs production and decay at the LHC [14]. Left: Production cross sections at $\sqrt{s} = 13\,\text{TeV}$, for $m_{\text{H}} = 120$–$130\,\text{GeV}$. Right: Decay Branching Fractions for $m_{\text{H}} = 120$–$130\,\text{GeV}$. 25

2.5 Coupling strengths of Higgs boson as a function of particle masses, extracted using LHC run 1 data [25]. Note that the uncertainties on fermionic couplings are significantly larger than the ones in W and Z. 26

2.6 Part of the Feynman diagrams involved in predicting Higgs production cross section at different orders. 31

3.1 The CERN accelerator complex [33]. 36

3.2 Distribution of the average number of interactions per crossing (pileup) for *pp* collisions in different data-taking years in the LHC Run II. The overall mean values and the minimum bias cross sections, which is *pp* collision cross section measured with minimum experimental selections, are also shown. 37
3.3 Cumulative luminosity versus day delivered to CMS during stable beams for pp collisions at nominal center-of-mass energy. 38
3.4 Cumulative luminosity versus day delivered to CMS during stable beams for *pp* collisions in LHC Run II. 39
3.5 Cut-out diagram of the CMS detector [34]. 40
3.6 Comparison of the layouts of the Phase-1 pixel detector (top) and the original pixel detector before the year-end technical stop between 2016 and 2017 (bottom) [37]. 43
3.7 Schematic diagram the CMS tracker [34]. 44
3.8 Layout of the CMS electromagnetic calorimeter [37]. 46
3.9 Layout of the CMS hadronic calorimeter [37]. 48
3.10 Response of the HE scintillators tiles as a function of the delivered integrated luminosity for several pseudorapidity values. The blue, red and black dots represent response measurements performed during 2012, 2015, and 2016, respectively. The colored lines are fits to the 2012 measurements for different η values [42]. 49

3.11 Depth segmentation of CMS HCAL before (top) and after (bottom) Phase-1 upgrade [37, 41]. 50
3.12 Three-level hierarchical structure of CMS run control software, see text for the descriptions of each entity. 52
3.13 Map of HCAL (HB, HE and HF only) backend electronics structure, called uHTR crates, in the $\eta-\phi$ plane segmented in HCAL granularity. Each grid corresponds to one read-out channel. The gray boxes covers the crates corresponding to the partition HBHEa, HBHEb and HBHEc, whereas the crates in the large $i\eta$ regions belongs to HF. 53
3.14 State-machine structure of CMS run control software [43]. 55
3.15 Layout of the CMS muon system [37]. 56
3.16 Diagram of basic unit of DT and CSC system, showing the drift cell (left) and cathode stripe chamber (right) [37]. 57
3.17 Level-1 Trigger system architecture [37]. 60
3.18 Diagram of how different particles traverse different detector components in the CMS detector [37]. 61
3.19 Efficiency (left) and misreconstruction rate (right) of global combinatorical track finder at different iterations as a function of track p_{T}[44]. 63
3.20 Muon transverse momentum resolution as a function of muon transverse momentum [37]. 66

3.21 Fit to (left) M_{PDG}/M_{GEN} as a function of jet p_T, where M_{PDG} = 80.4 GeV and M_{GEN} is the fitted mean of the generator level mass and (right) M_{GEN}/M_{RECO}, where M_{RECO} is the reconstructed softdrop mass, as a function of jet p_T in two η bins. [58] 73

3.22 Distribution of ρ (left) and m_{SD} (right), shown for QCD MC for different p_T bins. 74

3.23 Illustration of properties of b-hadrons that can be used to differentiate from light quark/gluon jets [65]. 79

3.24 Receiver operating characteristic (ROC) curve comparing the performance of different b-tagging algorithms in CMS [65]. 80

3.25 The weighted, normalized mass distributions for the "tagged" and "untagged" background and signal before adding the KL divergence penalty terms to the training (top) and after adding the KL divergence penalty terms to the training (bottom). 84

4.1 NLO QCD (left) and EW (right) corrections and the corresponding uncertainties. Ref [79] 93

4.2 Boson p_T distributions at the generator level for W (left) and Z (right) in 2016 (upper) and 2017/18 (lower) generation configurations, showing the uncorrected spectrum, as well as the successive corrections to arrive at NLO QCD+EW. Also included is the target sample, generated as described in the text ("V NLO"), with NLO EW corrections applied. The ratio is shown with respect to the target sample. 94

4.3 Transverse momentum distribution of the Higgs boson at the LHC with $\sqrt{s} = 13\,\text{TeV}$ computed in Ref. [84]. The upper panel shows absolute predictions at LO and NLO in the full SM and in the infinite top-mass approximation (EFT). The lower panel shows respective NLO/LO correction factors. The bands indicates theoretical errors of the full SM result due to scale variation. 97

4.4 Comparison between the LHC Higgs cross section working group prescription [86] for the p_{T}^{H} spectrum and the spectrum found in several CMS datasets. The factorization and renormalization scale uncertainties were evaluated as the envelope of the 7 scale variations stored as weights in the samples. For MINLO, the weights were not available due to technical issues, and for the HIG-17-010 sample they are the same as found in the POWHEG mass effects sample. In the lower panel, the ratio of each described p_{T} spectrum relative to the LHCHXSWG spectrum is shown. See text for further details on the various samples. 99

4.5 Measured trigger efficiency of the "OR" as function of jet p_{T} and jet m_{SD} is shown for 2016. 102

4.6 Measured trigger efficiency of the "OR" as function of jet p_{T} and jet m_{SD} is shown separately for each Run period in 2017. 103

4.7 Measured trigger efficiency of the "OR" as function jet p_{T} and m_{SD} jet for Run 2017B-F and Run 2017C-F. 104

4.8 Measured trigger efficiency of the "OR" as function of jet p_T and jet m_{SD} is shown separately for each Run period in 2018. 105
4.9 Measured trigger efficiency of the "OR" as function jet p_T and m_{SD} jet for Run 2018A-D. 106
4.10 Measured trigger efficiency of the "OR" as function jet p_T and deep double-b score for all 2017 and 2018 runs 106
4.11 AK8 jet p_T distribution. Left, simulated signal events normalized to their cross sections (Table 4.4). Right, background and signal events normalized to their cross section. 108
4.12 Uncorrected soft drop mass distribution for the p_T-leading jet. Left, simulated signal events normalized to their cross sections (Table 4.4). Right, background and signal events normalized to their cross sections. 108
4.13 Corrected soft drop mass distribution for the p_T-leading jet. Left, simulated signal events normalized to their cross sections (Table 4.4). Right, background and signal events normalized to their cross sections. 109
4.14 Left, distribution of ρ for simulated signal events. Right, data and simulation comparison of the ρ spectrum for jets with $p_T > 450$ GeV. At analysis level, only events with $-6.0 < \rho < -2.1$ are required. . . 110
4.15 Comparison between $N_2^{1,DDT}$ and N_2^1 efficiency as a function of m_{SD} with QCD simulation. The cut on N_2^1 corresponds to same overall efficiency of 26% in the analysis phase space. 112

4.16 $N_2^{1,\mathrm{DDT}}$ kinematic dependence as a function of ρ in different p_T bins: p_T:450-500 GeV (top-left), p_T:500-550 GeV (top-right), p_T:550-600 GeV (middle-left), p_T:600-675 GeV (middle-right), p_T:675-800 GeV (bottom-left), p_T:800-1000 GeV (bottom-right). The blue points represent again the 26% quantile value of $N_2^{1,\mathrm{DDT}}$, which by construction is flat as function of ρ. 113

4.17 Transformation maps of $N_2^1 \rightarrow N_2^{1,\mathrm{DDT}}$ for different data-taking years(top left: 2016,top right: 2017, bottom: 2018) corresponding to the N_2^1 quantile value of 26% (i.e. for $\epsilon_\mathrm{QCD} = 26\%$), as a function of ρ and p_T. 114

4.18 $N_2^{1,\mathrm{DDT}}$ distribution for the p_T-leading jet. Left, simulated signal events normalized to their cross sections (Table 4.4). Right, background and signal events normalized to their cross sections. 114

4.19 ROC curve for double-b evaluated on 2017 MC (orange), double-b evaluated on 2016 MC (orange dashed) and deep double-b (decorrelated) taggers, evaulated on 2017 MC. 115

4.20 Mass sculpting for deep double-b without (left) and with (right) mass decorrelation. 116

4.21 Deep double-b tagger distribution for the p_T-leading jet. Left, simulated signal events normalized to their cross sections (Table 4.4). Right, background and signal events normalized to their cross sections. . . . 116

4.22 ROC curves for different $\mathrm{t\bar{t}}$ sensitive variables. 119

4.23 Distribution of the maximum DeepCSV discriminator value for the four leading-p_{T} AK4 jets in the opposite hemisphere from the leading-p_{T} AK8 jet. Note, the medium working point discriminator cut value is 0.4941 for 2017 data. Left, simulated signal events normalized to their cross sections (Table 4.4). Right, background and signal events normalized to their cross sections. 119

4.24 $E_{\mathrm{T}}^{\mathrm{miss}}$ distributions. Left, simulated signal events normalized to their cross sections (Table 4.4). Right, background and signal events normalized to their cross sections. 120

4.25 Fraction of $\mathrm{H} \to \mathrm{b\overline{b}}$ events where the generated Higgs boson was found within $\Delta R < 0.8$ of either the leading reconstructed AK8 jet, the subleading jet, or not matched to any AK8 jet in the event, as a function of generated Higgs boson p_{T}. The fractions are shown separately for gluon fusion H production with the MINLO generator (left) and for VBF H production (right). Gluon fusion is the dominant production mode in the $p_{\mathrm{T}} > 450\,\mathrm{GeV}$ region, followed by VBF production. . . . 122

4.26 Scan of relative $S/\sqrt{B}$ for the deep double-b tagger and $N_2^{1,\mathrm{DDT}}$ substructure cuts, using simulated multijet QCD and ggF $\mathrm{H(b\overline{b})}$ signal events. The white contours show the value of $S/\sqrt{B}$. The maximum occurs near deep double-b values of 0.86 – 0.90 and $N_2^{1,\mathrm{DDT}}$ quantiles of 0.38 – 1 123

4.27 Efficiency of each of the requirement of the kinematic and top rejection selections, for simulated signal and background events. 124

4.28 m_{SD} distribution for simulated signal events after all the selection criteria in both the failing (left) and passing (right) b-tagging categories. Each process is normalized to its cross section as shown in Table 4.4 . 125

4.29 Signal region m_{SD} distribution for simulated signal and background events after all the selection criteria in both the failing (left) and passing (right) deep double-b tagging categories. 126

4.30 Distributions of variables in 2016 data and simulated samples after the kinematic preselection. From top to bottom: number of primary vertices, $E_{\mathrm{T}}^{\mathrm{miss}}$, p_{T}, η, m_{SD}, deep double-b tagger, $N_2^{1,\mathrm{DDT}}$, and the $\mathrm{t\bar{t}}$ rejection variable. The $VV(4q)$ category is mislabeled, and represents instead inclusive diboson production (yellow). The background processes, QCD(blue), W+jets(green), Z+jets(red), $\mathrm{t\bar{t}}$(grey), single-top(brown) and diboson(yellow), are shown as stacked distributions. 128

4.31 Distributions of variables in 2017 data and simulated samples after the kinematic preselection. From top to bottom: number of primary vertices, $E_{\mathrm{T}}^{\mathrm{miss}}$, p_{T}, η, m_{SD}, deep double-b tagger, $N_2^{1,\mathrm{DDT}}$, and the $\mathrm{t\bar{t}}$ rejection variable. The $VV(4q)$ category is mislabeled, and represents instead inclusive diboson production. The background processes, QCD(blue), W+jets(green), Z+jets(red), $\mathrm{t\bar{t}}$(grey), single-top(brown) and diboson(yellow), are shown as stacked distributions. 129

4.32 Distributions of variables in 2018 data and simulated samples after the kinematic preselection. From top to bottom: number of primary vertices, $E_{\mathrm{T}}^{\mathrm{miss}}$, p_{T}, η, m_{SD}, deep double-b tagger, $N_2^{1,\mathrm{DDT}}$, and the $\mathrm{t\bar{t}}$ rejection variable. The $VV(4q)$ category is mislabeled, and represents instead inclusive diboson production. The background processes, QCD(blue), W+jets(green), Z+jets(red), $\mathrm{t\bar{t}}$(grey), single-top(brown) and diboson(yellow), are shown as stacked distributions. 130

4.33 Schematic drawing of the effect of data-MC corrections factors applied on W/Z+jets backgrounds. 132

4.34 Muon p_{T} (left) and η (right) in the single muon control region for $\mathrm{t\bar{t}}$ background in 2016 data (top), 2017 data (middle), and 2018 data (bottom) respectively. QCD is scaled down by a k-factor of 0.78 and $\mathrm{t\bar{t}}$ is scaled to achieve the same overall background integral as data. . 138

4.35 AK8 Puppi jet p_T (left), η (middle), and deep double-b tag (right) in the single muon control region for $t\bar{t}$ background in 2016 data (top), 2017 data (middle) and 2018 data (bottom) respectively. QCD is scaled down by a k-factor of 0.78 and $t\bar{t}$ is scaled to same integral as data. . 139

4.36 Single muon control region for $t\bar{t}$ background in 2016 data. (Top) Inclusive in deep double-b tag. (Bottom left) Failing region with deep double-b tag < 0.89. (Bottom right) Passing region with deep double-b tag > 0.89. QCD is scaled down by a k-factor of 0.78 and $t\bar{t}$ is scaled to same integral as data. 140

4.37 Single muon control region for $t\bar{t}$ background in 2017 data. (Top) Inclusive in deep double-b tag. (Bottom left) Failing region with deep double-b tag < 0.89. (Bottom right) Passing region with deep double-b tag > 0.89. QCD is scaled down by a k-factor of 0.78 and $t\bar{t}$ is scaled to same integral as data. 141

4.38 Single muon control region for $t\bar{t}$ background in 2018 data. (Top) Inclusive in deep double-b tag. (Bottom left) Failing region with deep double-b tag < 0.89. (Bottom right) Passing region with deep double-b tag > 0.89. QCD is scaled down by a k-factor of 0.78 and $t\bar{t}$ is scaled to same integral as data. 142

4.39 m_{SD} distributions of the QCD MC in the pass and fail regions (top panel) and their ratio (bottom panel) in each p_{T} category. This polynomial TF is used to derive the QCD prediction (shown in red) for the pass region (defined as the fail region distribution times the polynomial). The χ^2 reported here refers to the compatibility between shapes in the pass and fail regions. 144

4.40 The F-test for different hypotheses on the MC (with $\mu = 1$), corresponding to the 2016(top),2017(middle),2018(bottom) dataset. 154

4.41 The F-test for different hypotheses on the 2016 dataset. 155

4.42 The F-test for different hypotheses on the 2017 (top) and 2018 (bottom) dataset. 156

4.43 The fitted pass-fail ratio $\epsilon^{\mathrm{QCDMC}}$ as a function of p_{T} and m_{SD} (left) or p_{T} and ρ (right) for QCD MC corresponding to conditions in 2016 (top), 2017 (middle), and 2018 (bottom). 158

4.44 Post-fit m_{SD} distributions in simulated MC events corresponding to 2016 conditions for the pass and fail regions and the first three p_{T} categories (450-500,500-550,550-600 from top to bottom) by using a polynomial 2nd order in ρ and 2nd order in p_{T}. The EW and signal events are also shown separately. 159

4.45 Post-fit m_{SD} distributions in simulated MC events corresponding to 2016 conditions for the pass and fail regions and the second three p_T categories (600-675,675-800,800-1200 from top to bottom) by using a polynomial 2nd order in ρ and 2nd order in p_T The EW and signal events are also shown separately. 160

4.46 Post-fit m_{SD} distributions in simulated MC events corresponding to 2017 conditions for the pass and fail regions and the first three p_T categories (450-500,500-550,550-600 from top to bottom) by using a polynomial 2nd order in ρ and 2nd order in p_T. The EW and signal events are also shown separately. 161

4.47 Post-fit m_{SD} distributions in simulated MC events corresponding to 2017 conditions for the pass and fail regions and the second three p_T categories (600-675,675-800,800-1200 from top to bottom) by using a polynomial 2nd order in ρ and 2nd order in p_T The EW and signal events are also shown separately. 162

4.48 Post-fit m_{SD} distributions in simulated MC events corresponding to 2018 conditions for the pass and fail regions and the first three p_T categories (450-500,500-550,550-600 from top to bottom) by using a polynomial 2nd order in ρ and 2nd order in p_T. The EW and signal events are also shown separately. 163

4.49 Post-fit m_{SD} distributions in simulated MC events corresponding to 2017 conditions for the pass and fail regions and the second three p_T categories (600-675,675-800,800-1200 from top to bottom) by using a polynomial 2nd order in ρ and 2nd order in p_T The EW and signal events are also shown separately. 164

4.50 Post-fit m_{SD} distributions in simulated MC events corresponding to 2016 (top), 2017 (middle), and 2018 (bottom) conditions for the pass and fail regions summing all 6 p_T categories by using a polynomial 2nd order in ρ and 2nd order in p_T The EW and signal events are also shown separately. 165

4.51 The pull distribution when generating and fitting with the nominal function. Plots are shown for different injected $H(b\bar{b})$ signal strengths: $\mu_H = 0, 1, 2, 5$. 166

4.52 Empirical function signal+background fit to the pass category directly with the 2017 MC. 168

4.53 $R_{p/f}$ TF function signal+background fit to the pass category directly with the 2017 MC. 169

4.54 $N_2^{1,DDT}$ Passing and failing region fits in the semileptonic $t\bar{t}$ sample for 2016 (top), 2017 (middle) and 2018 (bottom). Note for 2016, pre-fit and post-fit models are shown, while only post-fit is shown for 2017 and 2018. In addition, the ratio panels for 2017 and 2018 correspond to different quantities. 171

4.55 Effect of varying the jet mass scale by $\pm 10\sigma$ and the jet mass resolution by $\pm 2\sigma$ for W/Z+jets and the signal ggH($b\bar{b}$) in p_T category 1 ($450 < p_T < 500\,\text{GeV}$). 172

4.56 p_T dependence of the mass scale SF measured in the merged-W jet sample. 173

4.57 Distribution of deep double-b(left) and double-b(right) for QCD, gluon splitting and $H \to b\bar{b}$ processes. 175

4.58 The envelope of the relative effect of the 6 combinations (ignoring $0.5, 2$ and $2, 0.5$) about the nominal ggHbb yield, as a function of reconstructed Higgs boson p_T. The yields are as predicted by the HJ-MiNLO generator, with 2017 detector conditions. 176

4.59 The envelope of the relative effect on signal acceptance of the 6 combinations (ignoring $0.5, 2$ and $2, 0.5$) about the nominal ggHbb yield, as a function of reconstructed Higgs boson p_T, for three regions in generated Higgs boson p_T. The yields are as predicted by the HJ-MiNLO generator, with 2017 detector conditions. 177

4.60 Relative effect of the AK4 DeepCSV b-tagging systematic uncertainty on the expected event yield, per jet p_{T} bin in the signal region. The effect is shown separately for events in the AK8 deep double-b passing (bottom) and failing (top) regions. The effect is as measured in a sample of $\mathrm{t\bar{t}}$ production (left) and Z+jets (right) events simulated with 2017 data-taking conditions and corrected by the 2017 BTV-POG scale factors. 179

4.61 Relative effect of the jet trigger systematic uncertainty on the expected event yield, per jet p_{T} bin in the signal region. The effect is shown separately for events in the AK8 deep double-b passing (bottom) and failing (top) regions. The effect is as measured in a sample of Z+jets events simulated with 2017 data-taking conditions and corrected by the 2017 jet trigger efficiency as measured in an independent control sample. 180

4.62 2016 muon trigger efficiency and ID and isolation efficiency SFs (and their uncertainties) measured in data as a function of muon p_{T} and η. 181

4.63 2017 muon trigger efficiency and ID and isolation efficiency SFs (and their uncertainties) measured in data as a function of muon p_{T} and η. 182

4.64 2018 muon trigger efficiency and ID and isolation efficiency SFs (and their uncertainties) measured in data as a function of muon p_{T} and η. 183

4.65 Post-fit m_{SD} distributions in the full 2016, 2017, and 2018 data set for the pass and fail regions and the first three p_{T} categories (450-500,500-550,550-600 from top to bottom) summed over all three years. The EW and signal events are also shown separately. 187

4.66 Post-fit m_{SD} distributions in the full 2016, 2017, and 2018 data set for the pass and fail regions and the second three p_{T} categories (600-675,675-800,800-1200 from top to bottom) summed over all three years. The EW and signal events are also shown separately. 188

4.67 Post-fit m_{SD} distributions in the full Run II data set for the pass and fail regions. The EW and signal events are also shown separately. . . 189

4.68 The fitted data-mc transfer function TF^{data} as a function of p_{T} and m_{SD} (left) or p_{T} and ρ (right) for 2016 (top), 2017 (middle), and 2018 (bottom) data. The polynomial orders of used these functions are 1st order in p_{T} and 1st order in ρ for 2017 and 2018, while the order in p_{T} and ρ are 2nd order and 1st order for 2016 respectively. 190

4.69 The fitted pass-fail ratio $\epsilon^{\mathrm{QCDMC}}$ as a function of p_{T} and m_{SD} (left) or p_{T} and ρ (right) for QCD MC corresponding to conditions in 2016 (top), 2017 (middle), and 2018 (bottom). The polynomial orders of used these functions are 2nd order in p_{T} and 2nd order in ρ. 191

4.70 Results of the goodness-of-fit test for the full (a) 2016 data set , (b) 2017 data set, (c) 2018 data set and (d) full Run 2 data set using the saturated test statistic. The order of polynomial used are second(first) order in ρ and first(first) order in p_{T} for all 2016(2017,2018) data sets. 192

4.71 1D profile test statistic $-2\Delta\log(\mathcal{L})$ scans as a function of the signal strengths μ_{H}, fixing $\mu_{\mathrm{Z}} = 1$ (left) and floating μ_{Z} (right), with Asimov data set for 2016+2017+2018. The solid black curve, labeled "stat+syst," corresponds to the scan when profiling all nuisance parameters and the dotted blue curve, labeled "stat only," corresponds to only profiling the unconstrained nuisance parameters, such as the QCD transfer factor parameters. 193

4.72 1D profile and 2D likelihood test statistic $-2\Delta\log(\mathcal{L})$ scans as a function of the signal strengths μ_{H} and μ_{Z} on data for 2016+2017+2018. In the 1D profile, the solid black curve, labeled "stat+syst," corresponds to the scan when profiling all nuisance parameters and the dotted blue curve, labeled "stat only," corresponds to only profiling the unconstrained nuisance parameters, such as the QCD transfer factor parameters. 194

4.73 The best-fit H signal strength μ_{H} and uncertainty per p_{T} category extracted from a simultaneous fit of all channels, performed on different data taking periods. 197

4.74 The best-fit Z signal strength μ_Z and uncertainty per p_T category extracted from a simultaneous fit of all channels, performed on different data taking periods. The H signal strength μ_H is also floating in these fits. 198

4.75 The best-fit signal strength μ_H and uncertainty per year, extracted from a simultaneous fit of all channels. 199

4.76 The best-fit H signal strength μ_H and uncertainty per p_T category extracted from a simultaneous fit of all channels using full Run 2 dataset. 199

4.77 STXS Stage 1.2 binning. 201

4.78 Measured differential fiducial cross section as a function of Higgs boson p_T, in comparison to the predictions of Ref. [86] and HJ-MiNLO [105]. In the bottom two panels, the dotted line corresponds to a ratio of one. The measured cross section uncertainty in the first bin is larger because of limited acceptance. The cross section measurements in the first and second bins have a mild anti-correlation, primarily due to jet energy scale uncertainties, which inflates the corresponding uncertainties. The observed cross section in the third bin has a smaller relative uncertainty than that in the second bin because of the larger magnitude of the central value in that bin. The relative uncertainties in the predictions of Ref. [86] and HJ-MiNLO are approximately 10% and 20%, respectively. 202

4.79 (a) Response matrix corresponding to the STXS binning of the generated Higgs p_{T}. Reconstructed p_{T} binning is fixed to the p_{T} categories defined in the analysis. The z-axis value quantifies the fraction of events in a given reconstructed p_{T} bin for the given generated p_{T} bin, normalized to the total production rate in the generated p_{T} bin. (b) Correlation between per-p_{T}-bin signal strength parameters corresponding to the STXS binning of the generated Higgs p_{T}. 203

B.1 Post-fit m_{SD} distributions in the full 2016 dataset for the pass and fail regions and the first three p_{T} categories (450-500,500-550,550-600 from top to bottom) by using a polynomial 2nd order in ρ and 1st order in p_{T}. The EW and signal events are also shown separately. 208

B.2 Post-fit m_{SD} distributions in the full 2016 dataset for the pass and fail regions and the second three p_{T} categories (600-675,675-800,800-1200 from top to bottom) by using a polynomial 2nd order in ρ and 1st order in p_{T}. The EW and signal events are also shown separately. . . 209

B.3 Post-fit m_{SD} distributions in the full 2017 dataset for the pass and fail regions and the first three p_{T} categories (450-500,500-550,550-600 from top to bottom) by using a polynomial 1st order in ρ and 1st order in p_{T}. The EW and signal events are also shown separately. 210

B.4 Post-fit m_{SD} distributions in the full 2017 dataset for the pass and fail regions and the second three p_T categories (600-675,675-800,800-1200 from top to bottom) by using a polynomial 1st order in ρ and 1st order in p_T. The EW and signal events are also shown separately. 211

B.5 Post-fit m_{SD} distributions in the full 2018 dataset for the pass and fail regions and the first three p_T categories (450-500,500-550,550-600 from top to bottom) by using a polynomial 1st order in ρ and 1st order in p_T. The EW and signal events are also shown separately. 212

B.6 Post-fit m_{SD} distributions in the full 2018 dataset for the pass and fail regions and the second three p_T categories (600-675,675-800,800-1200 from top to bottom) by using a polynomial 1st order in ρ and 1st order in p_T. The EW and signal events are also shown separately. 213

B.7 Post-fit m_{SD} distributions in the full 2016, 2017, and 2018 dataset for the pass and fail regions. The EW and signal events are also shown separately. 214

C.1 The top(middle/bottom) rows corresponds to quantities detailed below, corresponding to data collected in 2016(2017/2018) condition respectively. (Left) The expected pass-fail ratio ϵ^{QCD} as a function of jet p_{T} and m_{SD} for QCD multijet simulation. (Middle) The correction TF^{data} for residual differences between data and simulation as a function of jet p_{T} and m_{SD} (Right) The fitted pass-fail ratio $R_{\mathrm{p/f}}$ as a function of jet p_{T} and m_{SD}. The ratio relates the QCD multijet event yield in the passing region to that of the failing region. The pass-fail ratio is the product of the previous two components. 216

C.2 The top(middle/bottom) rows corresponds to quantities detailed below, corresponding to data collected in 2016(2017/2018) condition respectively. (Left) The expected pass-fail ratio ϵ^{QCD} as a function of jet p_{T} and ρ for QCD multijet simulation. (Middle) The correction TF^{data} for residual differences between data and simulation as a function of jet p_{T} and ρ (Right) The fitted pass-fail ratio $R_{\mathrm{p/f}}$ as a function of jet p_{T} and ρ. The ratio relates the QCD multijet event yield in the passing region to that of the failing region. The pass-fail ratio is the product of the previous two components. 217

Chapter 1

Introduction

Particle physics tries to answer one of the oldest questions of humanity with scientific method: What is our world made of? From the classical elements era, to the ancient Greek's idea of a single indivisible building block called atom, the scientific method have revealed a more complicated, yet elegant, picture than these wild guesses.

We built tools to propel this endeavor, both in terms of experimental apparatus and mathematical framework. From cloud chambers to particle accelerators, we have refined our tools to probe nature from a table-top size box to a 27 km underground tunnel. It took us the invention of quantum mechanics and special relativity to see just beyond the two small clouds on the horizon, and many more efforts to combine the two into quantum field theory that forms the basis of today's understanding of particle physics, the Standard Model.

The discovery of the Higgs boson in 2012 marks the completion of the Standard

Model, which embodies a century's worth of effort both theoretically and experimentally. CMS pioneered an inclusive search [1] for the Higgs boson decaying into bottom quark pairs with 35.9 fb^{-1} of LHC data at $\sqrt{s} = 13$ TeV. This dissertation extends the search using more than 3 times worth of data, collected in the LHC run 2 and improved identification of bottom quark pairs.

This thesis begins with a brief review of the Standard model in Chapter 2, highlighting the theoretical concepts that are fundamental to this work, which is then followed by a description of the CMS experiment in Chapter 3 to provide the basis for a detailed analysis of boosted Higgs boson decaying into bottom quark pairs in Chapter 4.

Chapter 2

The Standard Model

2.1 Particle content

Before we dive into the theoretical framework of the Standard Model (SM), it is helpful to first have an overview of what are the particle content of the theory. Instead of taking an historical approach to introduce the particles in the order of their discoveries, we will use the benefit of hindsight and present them in logical groups.

After introducing the particle content, we will connect the pieces with the central idea of gauge principle to illustrate the theoretical structure of SM. As an overview, the particles in the SM, together with some of their quantum numbers, is summarized in Fig. 2.1.

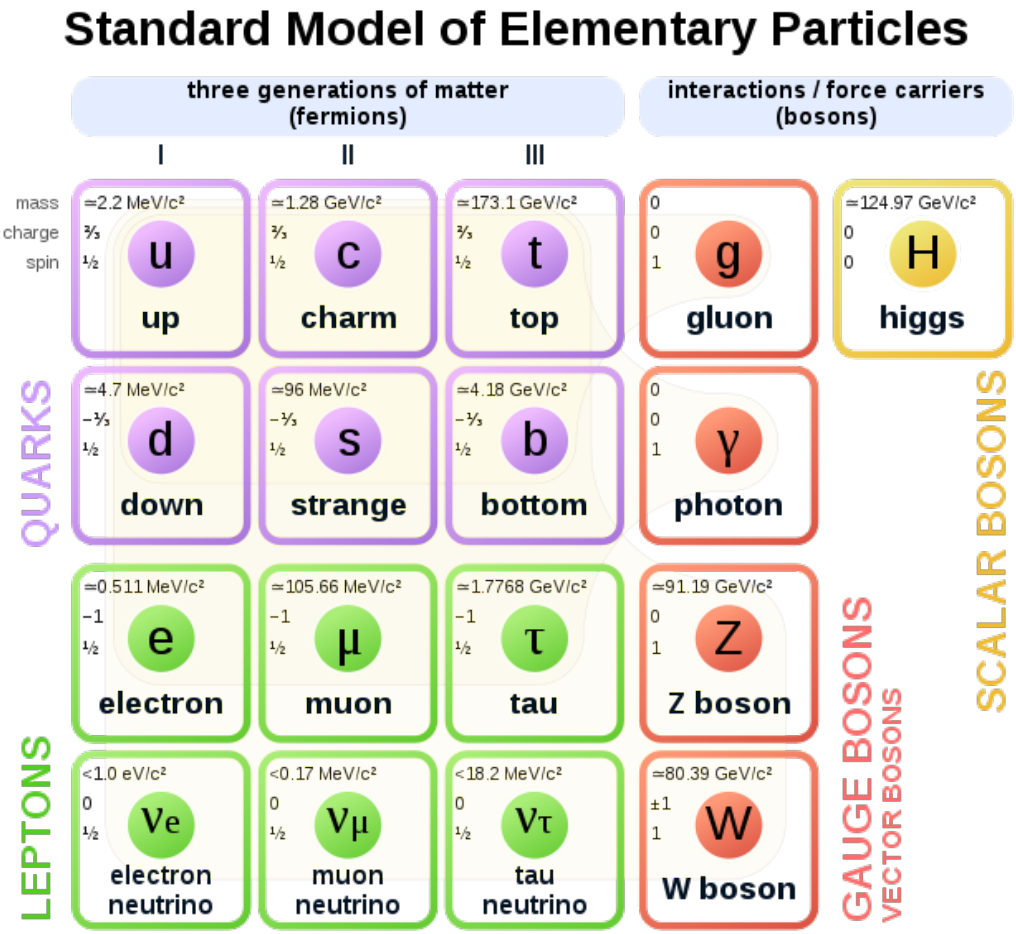


Figure 2.1: Particle content of the standard model.

2.1.1 Bosons

Gauge bosons

Interactions between elementary particles are mediated by particles called the gauge bosons, which are spin-1 particles. As will be described in the next sections, the gauge bosons correspond to the generators of the underlying symmetry group. Their name was given by the fact that the gauge boson arises naturally after applying gauge principle to the symmetry group.

SM describes three of the four fundamental forces of nature, electromagnetic (EM) interaction, weak interaction and strong interaction. Electromagnetism is mediated by the photon, a massless particle mediating a long-ranged interaction without carrying the EM charge. Weak interaction is mediated by the charged $W^{\pm}$ boson and neutral Z boson, which are both massive. Strong interaction is mediate by the gluons, which are self-interacting massless particles, carrying a superposition of colour charges.

The gauge group that generates the 3 interactions is

$$SU_C(3) \otimes SU_L(2) \otimes U_Y(1). \tag{2.1}$$

For various reasons [2, 3, 4], the remaining fundamental interaction, gravity, which originates from the curvature of spacetime as described in the very successful general relativity, is not able to fit under the same theoretical framework, the quantum field theory (QFT), as the SM. Reconciling the quantum nature of spacetime with general

relativity remains to be one of the biggest challenges of theoretical Physics.

The coupling strengths of the forces are summarized in Table 2.1.

Table 2.1: Coupling strength of fundamental interactions in the SM.

Interaction	Mediators	Coupling strength
Strong	Gluons	0.1
Weak[1]	$W^{\pm}$ and Z	10^{-6}
Electromagnetism	Photon	1/137

Higgs boson

The Higgs boson is the only fundamental scalar particle in the SM, which is the last discovered in 2012 by the ATLAS and CMS experiment at the LHC. It differs with the rest of gauge boson in that it does not represent the generator of the SM gauge group, thus not a gauge boson. Yet it plays the central role of giving masses to both the weak gauge bosons, by spontaneously breaking the electroweak symmetry such that

$$SU_L(2) \otimes U_Y(1) \rightarrow U_{em}(1). \tag{2.2}$$

It also gives masses to the fermions via Yukawa interactions. This will be discussed in details in Sec. 2.4.1.

[1] This is the Fermi contact interaction coupling constant, G_F, which, unlike the couplings shown for the strong and EM ones, is a dimensional quantity. G_F is chosen here for illustration, as it sets the relevant strength for weak interactions at low energies. The actual weak coupling constant depends on energy scale [5].

2.1.2 Fermions

Fermions in the SM are spin-1/2 particles that follow Fermi-Dirac statistics. The most striking difference between fermions and gauge bosons in SM is that there are three *generations* of fermions, with each generation being a heavier copy of the previous generation and having the same quantum numbers otherwise. This generation structure is not explained in the SM and the masses of fermions, which spans at least 5 orders of magnitude [2], are treated as fundamental input parameters to the theory.

Fermions are divided into two groups based on their gauge interactions, called quarks and leptons.

Quarks

Quarks are the only fermions with color charge in the SM. Each generation consists of a up-type quark with EM charge of $+2/3$ and a down-type quark with EM charge of $-1/3$. Each quark forms a colour triplet under the $SU_C(3)$ transformation and forms a weak isospin doublet with the other quark in the same generation under the $SU_L(2)$ transformation.

Quarks are observed experimentally in the forms of colourless bound states of two quarks (mesons) or three quarks (hadrons) [3]. This is due to the fact that the coupling constant (α_s) between quarks and gluons decreases as the energy scale increases. The

[2] The heaviest quark is the top quark with 174 GeV and the lightest quark is around 3 MeV. Neutrinos masses are of the order *milli*-eV.

[3] Pentaquarks, bound states made out of 5 (anti)-quarks, have also been observed recently [6], but are much more rare than hadrons and mesons.

renormalization group equation is given by

$$\alpha_s(\mu) = \frac{2\pi}{11 - 2n_f/3} \frac{1}{\ln \frac{\mu}{\Lambda_{QCD}}}, \tag{2.3}$$

where μ is the energy scale, $n_f = 6$ is the number of quark flavors, Λ_{QCD} is the scale of QCD. As a result, strong interaction becomes effectively short-ranged and bounds quarks into colourless bound states. This is called *asymptotic freedom*. Eq. 2.3 also implies that QCD calculations are only perturbative at high-energies where the coupling is sufficiently small for purturbative technique.

In terms of weak interactions, quarks flavour eigenstates do not align exactly with their respective mass eigenstates. This allows flavor-changing decays between generations of quarks. The transformation between flavor eigenstates and mass eigenstates is parametrized as the so called CKM matrix [7, 8], whose parameters are taken as input to the SM.

Leptons

Each generation of lepton consists a charged lepton and a neutral lepton, called neutrino. Without electical charge, the neutrinos only experience weak interactions. Lepton numbers are straightly conserved within each generations in the SM, therefore the lightest charged lepton, electron, is a stable particle.

Unlike other fermions, both the charged leptons and quarks, there are *only* right-handed neutrinos found experimentally. Therefore, only the left-handed leptons in a generation forms a complete weak isospin doublet, whereas the right-handed charged

leptons is a singlet.

2.2 Local gauge invariance

It is one of the most fundamental realization in modern Physics that symmetries governs dynamics. Thus specifying the underlying symmetries already dictates a large part of a theory and study the symmetry groups has become one of the themes of theoretical Physics. The symmetry that links between the symmetry and dynamics is the local gauge symmetry. In this section, we will see how the idea of local gauge symmetry connects all three standard model forces.

2.2.1 The Quantum Electrodynamics

We begin the discussion with Quantum Electrodynamics (QED), which is a abelian gauge theory with symmetry group $U(1)$. The Lagrangian of a free Dirac scalar field $\psi(x)$ is

$$\mathcal{L} = \bar{\psi}(i\partial\!\!\!/ - m)\psi, \tag{2.4}$$

where $\partial\!\!\!/ = \gamma^\nu \partial_\nu$ and γ^ν are the gamma matrices. We require the Lagrangian in Eqn. 2.4 to be invariant, i.e. $\mathcal{L} \to \mathcal{L},$, under a local phase transformation

$$\psi \to e^{-iq\lambda(x)}\psi, \tag{2.5}$$

where $\lambda(x)$ is a function of space and q is some coupling constant. It is clear that the free Lagrangian in Eqn. 2.4 will not satisfy local phase invariance as that gives rise

an extra term $\mathcal{L} \to \mathcal{L} + (q\bar{\psi}\gamma^\mu\psi)\partial_\mu\lambda$.

By inspection, the local phase invariance can be restored by introducing a new gauge field A_μ which transform as

$$A_\mu \to A_\mu + \partial_\mu\lambda. \tag{2.6}$$

with an additional term $\mathcal{L}_{int}$ in the Lagrangian

$$\mathcal{L}_{int} = -(q\bar{\psi}\gamma^\mu\psi)A_\mu, \tag{2.7}$$

because the new term transform as $\mathcal{L}_{int} \to \mathcal{L}_{int} - (q\bar{\psi}\gamma^\mu\psi)\partial_\mu\lambda$.

Given the new gauge field has 4 components, A_μ is a spin-1 vector field, which is described by the Proca Lagrangian,

$$\mathcal{L}_{Proca} = -\frac{1}{4}F^{\mu\nu}F_{\mu\nu} + \frac{1}{2}m_A A^\mu A_\mu, \tag{2.8}$$

where $F^{\mu\nu} = \partial^\mu A^\nu - \partial_\nu A_\mu$ is the anti-symmetric derivative of A_μ and m_A is the mass of the A_μ.

It is easy to see that the $A^\mu A_\mu$ term is *not* invariant under local phase transformation, as it generates terms with product of derivatives of λ. Therefore, m_A has to be *zero* if we require local phase invariance.

Putting the terms together, the local phase invariant Lagrangian becomes

$$\mathcal{L} = \bar{\psi}(i\partial\!\!\!/ - m)\psi - \frac{1}{4}F^{\mu\nu}F_{\mu\nu} - (q\bar{\psi}\gamma^\mu\psi)A_\mu, \tag{2.9}$$

which is nothing but the Lagrangian of QED and the A_μ field is the electromagnetic potential or the photon field.

Hence, we explicitly see how demanding the local phase (gauge) invariance of the Lagrangian generates the interaction term $\mathcal{L}_{int}$ between the photon and matter fields. This relation can also be seen by re-writing Eq. 2.9 using the covariant derivative $\mathcal{D}_\mu$

$$\mathcal{D}_\mu = \partial_\mu + iqA_\mu, \tag{2.10}$$

to be

$$\mathcal{L} = \bar{\psi}(i\gamma^\mu \mathcal{D}_\mu - m)\psi - \frac{1}{4}F^{\mu\nu}F_{\mu\nu} \tag{2.11}$$

Note how Eqn. 2.11 looks like a interaction free theory, but the interaction term between the gauge-field and matter field is hidden in the covariant derivative.

Eq. 2.5 can be treated as multiplying a 1×1 unitary matrix $U = e^{-iq\lambda}$ such that $\psi \rightarrow U\psi$. Since the matrices U commute with each others, we call QED a $U(1)$ abelian gauge theory.

It turns out that the gauge fields being massless in a local gauge theory is a general result, which is the reason why the Higgs mechanism needs to be introduced to explain the mass of W and Z boson.

2.2.2 Yang-Mills Theory

The same idea can be generalized symmetry group of higher dimensions. In 1954, Yang and Mills generalized the local gauge invariance to the $SU(2)$ group [9], which

is to be applied on a Dirac field doublet:

$$\psi \equiv \begin{pmatrix} \psi_1(x) \\ \psi_2(x) \end{pmatrix}. \tag{2.12}$$

The local gauge transformation then becomes

$$\psi \to e^{-iq\boldsymbol{\tau}\cdot\boldsymbol{\lambda}(x)}\psi, \tag{2.13}$$

where $\boldsymbol{\tau} = (\sigma_1, \sigma_2, \sigma_3)$ are the 2×2 Pauli matrices. Instead of a trivial 1×1 matrix, $e^{-iq\boldsymbol{\tau}\cdot\boldsymbol{\lambda}(x)}$ can now represent *any* 2×2 unitary matrices with determinant 1 or $SU(2)$ matrices. The Pauli matrices are called the generators of $SU(2)$ group, since it generates all the elements in the group. Equation 2.13 essentially represents a local rotation in the weak-isospin space and local gauge invariance demands the Lagrangian to be invariant under this transformation.

The same ideas in QED apply here, we need the introduction of massless gauge fields to restore local gauge invariance, except this time we need 3 additional gauge fields

$$\boldsymbol{A}^\nu = (A_1^\nu, A_2^\nu, A_2^\nu)^T \tag{2.14}$$

Then we can arrive at a Lagrangian of the same form as Eq. 2.9,

$$\mathcal{L} = \bar{\psi}(i\partial\!\!\!/ - m)\psi - \frac{1}{4}\boldsymbol{F}^{\mu\nu}\boldsymbol{F}_{\mu\nu} - (q\bar{\psi}\gamma^\mu\boldsymbol{\tau}\psi)\cdot\boldsymbol{A}_\mu, \tag{2.15}$$

where the dual tensor $\boldsymbol{F}^{\mu\nu}$ has gained an extra anti-symmetric term out of the 3 gauge

fields as follows

$$\boldsymbol{F}^{\mu\nu} = \partial^{\mu}\boldsymbol{A}^{\nu} - \partial_{\nu}\boldsymbol{A}_{\mu} - 2q(\boldsymbol{A}^{\mu} \times \boldsymbol{A}^{\nu}). \tag{2.16}$$

We shall see the 3 gauge fields will become the $W^{\pm}$ boson and the Z boson. Since the $SU(2)$ generators do not commute with each others, this model is called non-Abelion gauge theory.

2.3 Quantum Chromodynamics

The next generalization is then "obvious" with our hindsight, we want a Lagrangian that is invariant under local gauge transformation in the *color* space. Therefore, the Dirac doublet in Eqn 2.12 becomes a triplet

$$\psi \equiv \begin{pmatrix} \psi_r(x) \\ \psi_g(x) \\ \psi_b(x) \end{pmatrix}. \tag{2.17}$$

We can now focus on the $SU(3)$ generators T^a,

$$T^a = \frac{1}{2}\lambda^a, \qquad a = 1, \ldots, 8 \tag{2.18}$$

where λ^a are the eight 3×3 Gell-Man matrices. In general, there are 9 degrees of freedom in a 3×3 matrix, but constrain of having determinant of 1 reduces it by one, hence we have 8 distinct Gell-Man matrices. Similar to the Pauli matrices, the

generators obey commutation relations

$$[T^a, T^b] = if^{abc}T^c, \tag{2.19}$$

where f^{abc} are the structure constants of the $SU(3)$ group. Any $SU(3)$ transformation can then be represented by $U = e^{-i\boldsymbol{\lambda}\cdot\boldsymbol{a}}$.

Applying the same local gauge principle, we can recover the Lagrangian for QCD in the same form,

$$\mathcal{L} = \bar{\psi}(i\not{\partial} - m)\psi - \frac{1}{4}\boldsymbol{F}^{\mu\nu}\boldsymbol{F}_{\mu\nu} - (g_s\bar{\psi}\gamma^\mu\boldsymbol{\lambda}\psi)\cdot\boldsymbol{A}_\mu, \tag{2.20}$$

except with replacing the Pauli matrices $\boldsymbol{\tau}$ with the Gell-Mann matrices $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_8)$ and electric charge q replaced by the strong coupling constrant g_s. Similarly, the 8 gauge fields A_μ represents the 8 gluons in QCD.

Including all 6 quark flavours as independent Dirac fields, the full QCD Lagrangian can be written using the covariant derivatives,

$$\mathcal{D}_\mu = \partial_\mu + iqA_\mu, \tag{2.21}$$

as

$$\mathcal{L}_{QCD} = \sum_f \left(\bar{\psi}^f(i\gamma^\mu\mathcal{D}_\mu - m_f)\psi^f\right) - \frac{1}{4}\boldsymbol{F}^{\mu\nu}\boldsymbol{F}_{\mu\nu}, \tag{2.22}$$

where f are the 6 quark flavors.

2.4 Electroweak theory

We now have all the essential theory ingredients to describe the electroweak sector of the SM. The Yang-Mills theory in Sec. 2.2.2 has a fundamental limitation that the gauge field must be massless, which does not align with the massive weak boson observed experimentally. To reconcile this, the masses of the gauge fields are introduced via the spontaneous symmetry breaking by introducing a elementary scalar field, called the Higgs field. This also allows a unified description of weak interaction and electromagnetism at high energies before the gauge symmetry is spontaneously broken. Specifically, the unified electroweak theory is a non-abelion gauge theory with $SU(2)_L \otimes U(1)_Y$ symmetry, which is broken down to a $U(1)_{em}$ symmetry in the presence of the Higgs field. The subscripts denote which space does the symmetry operates, with L indicates the left-handed component of Dirac fermion fields, Y denotes the weak-hypercharge. The direct product between the groups means that these two groups are independent.

The electroweak Lagrangian $\mathcal{L}_{EW}$ consists of 4 parts,

$$\mathcal{L}_{EW} = \mathcal{L}_G + \mathcal{L}_F + \mathcal{L}_H + \mathcal{L}_{Yuk} \tag{2.23}$$

where $\mathcal{L}_G$ describes the gauge fields, $\mathcal{L}_F$ describes the fermion fields, $\mathcal{L}_H$ describes the Higgs field and $\mathcal{L}_{Yuk}$ describes the Yukawa interactions between the Higgs field and fermion fields.

Gauge fields

As we have seen in Sec. 2.2.1 and 2.2.2, the Lagrangian terms for the $SU(2)$ and $U(1)$ gauge fields are

$$\mathcal{L}_G = -\frac{1}{4}W^i_{\mu\nu}W^{\mu\nu i} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu} \tag{2.24}$$

where $i = 1, 2, 3$, the field strength tensors are written in index notations with the Levi-Civita symbol ϵ_{ijk}

$$B_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu \tag{2.25}$$

$$W_{\mu\nu} = \partial_\mu W^i_\nu - \partial_\nu W^i_\mu - g\epsilon_{ijk}W^j_\mu W^k_\nu. \tag{2.26}$$

and g is the $SU(2)$ gauge coupling constant.

We shall see that these four gauge fields, B and W^i will eventually mix with each others, after symmetry breaking, to form the photon, the $\mathrm{W}^\pm$ and the Z bosons.

Fermion-gauge interaction

As seen in Eq. 2.11, the fermion-gauge interaction is best represented in terms of covariant derivatives. The fermion-gauge interaction Lagrangian is given by,

$$\mathcal{L}_F = \sum_{j,f} \bar{\psi}^j_{L,f} i\gamma^\mu \mathcal{D}^L_\mu \psi^j_{L,f} + \sum_{j,f} i\gamma^\mu \mathcal{D}^R_\mu \psi^j_{R,f}, \tag{2.27}$$

where $\psi_{L(R)} \equiv (1 \mp \gamma^5)\psi/2$ are the left (right) chiral projection of the Dirac fermion field. The left and right covariant derivatives are given by,

$$\mathcal{D}_\mu^L = \partial_\mu - ig\frac{1}{2}\sigma_i W_\mu^i + ig'\frac{Y}{2}B_\mu, \tag{2.28}$$

$$\mathcal{D}_\mu^R = \partial_\mu + ig'\frac{Y}{2}B_\mu, \tag{2.29}$$

where g' is the $U(1)_Y$ gauge coupling constant and σ_i are the Pauli matrices.

Since the right-handed fermions do not transform under the $SU_L(2)$ group, there are no interaction terms between the $SU_L(2)$ gauge fields and the right handed fermions. The sum over j in Eq. 2.27 are the 3 generations of fermions and the index f indicates flavor. Writing out the first generation explicitly, we have

$$\psi_{L,q}^1 = \begin{pmatrix} u_L \\ \\ d_L \end{pmatrix} \qquad \psi_{L,l}^1 = \begin{pmatrix} \nu_L^e \\ \\ e_L \end{pmatrix} \tag{2.30}$$

for left-handed fermions of the first generations (up u and down d quark, electron e and electron neutrino ν_e), where we have dropped the ψ inside the spinor. And we have

$$\psi_{R,u} = u_R \quad \psi_{R,d} = d_R \quad \psi_{R,e} = e_R, \tag{2.31}$$

for the right-handed fermions, which do not contain neutrinos.

2.4.1 Higgs Mechanism

The Higgs field a complex scalar field which is a weak isospin doublet

$$\phi(x) = \begin{pmatrix} \phi^+(x) \\ \phi^0(x) \end{pmatrix}, \tag{2.32}$$

where ϕ^+ and ϕ^0 are complex fields.

The Lagrangian for a scalar field consists of a kinetic term and a potential term $V(\phi)$ as follows,

$$\mathcal{L}_H = (\mathcal{D}^\mu \phi)^\dagger \mathcal{D}_\mu \phi - V(\phi) \tag{2.33}$$

where the covariant derivative is given by

$$\mathcal{D}_\mu = \partial_\mu - ig\frac{1}{2}\sigma_i W_\mu^i + ig'\frac{Y}{2}B_\mu \tag{2.34}$$

and the Higgs potential term is

$$V(\phi) = -\mu^2 \phi^\dagger \phi + \frac{\lambda}{4}(\phi^\dagger \phi)^2, \qquad \text{with } \mu^2, \lambda > 0 \tag{2.35}$$

By differentiating $V(\phi)$ against ϕ, it is easy to check that the minimum of the potential occurs at $\phi^\dagger\phi = \frac{2\mu^2}{\lambda}$. Figure 2.2 illustrates the shape of the potential, which exhibit rotational symmetry and it is clear that the minimum location depends only on the radius of the field value. This means the Higgs field will acquire a non-zero vaccum

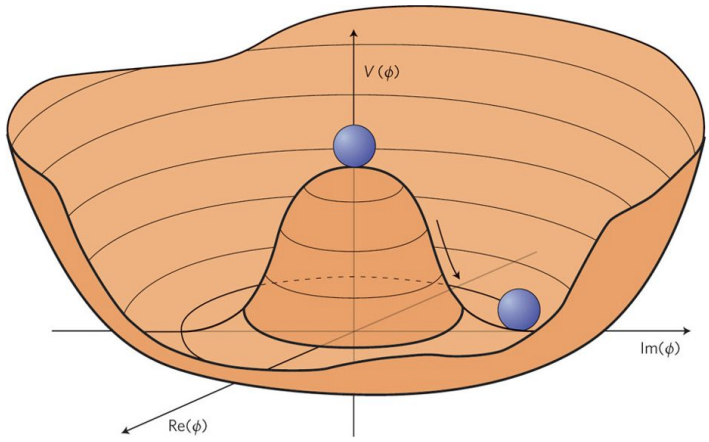


Figure 2.2: Sketch of the Higgs potential $V(\phi)$ [10].

expectation value (VEV) of

$$\langle 0|\phi|0\rangle = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ \nu \end{pmatrix}, \tag{2.36}$$

where $\nu = \frac{2\mu}{\sqrt{\lambda}} = 246\,\text{GeV}$ is a well-measured constant from the mass ratios of W and Z bosons. Since ϕ is a $SU_L(2)$ doublet, the ground state of Eq. 2.36 selecting a particular field value clearly breaks the symmetry.

A natural question is then, what would the Lagrangian $\mathcal{L}_H$ in Eq. 2.33 becomes in this vaccum. In general, we need 4 parameters to describe two complex scalar field, say we use $\theta_a, a = 1, 2, 3$ and h

$$\phi(x) = \frac{1}{2} \begin{pmatrix} \theta_2 + i\theta_1 \\ \nu + h(x) - i\theta_3 \end{pmatrix}, \tag{2.37}$$

If θ_a and h are small, we can write

$$\phi(x) = \frac{\exp(i\theta_a \sigma^a/\nu)}{2} \begin{pmatrix} 0 \\ \nu + h(x) \end{pmatrix}, \tag{2.38}$$

which allows us to recognize $\exp(i\theta_a \sigma^a/\nu)$ is the $SU_L(2)$ gauge transformation. We can choose a gauge transform such that $\phi' = \exp(-i\theta_a \sigma^a/\nu)\phi$, which gives

$$\phi'(x) = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ \nu + h(x) \end{pmatrix}. \tag{2.39}$$

This choice of gauge is called the Unitary gauge. The three scalar degrees of freedom θ_a, which are called Goldstone Boson, has disappeared in this gauge because the symmetry group that generates them is broken spontaneously.

By selecting a particular gauge, we have traded generality with a more transparent physical picture. It is much easier now to evaluate the Lagrangian in the Unitary gauge. First, we look at the kinetic term

$$(\mathcal{D}^\mu \phi)^\dagger \mathcal{D}_\mu \phi = \left| \left(\frac{g}{2} \sigma^i W_\mu^i + \frac{g'}{2} B_\mu \right) \begin{pmatrix} 0 \\ \nu/\sqrt{(2)} \end{pmatrix} \right|^2 + h \text{ terms} \tag{2.40}$$

which can be much simplified if we define

$$W_\mu^\pm = \frac{1}{\sqrt{2}} \left(W_\mu^1 \mp i W_\mu^2 \right), \tag{2.41}$$

$$\begin{pmatrix} Z_\mu & A_\mu \end{pmatrix} = \begin{pmatrix} \cos\theta_W & \sin\theta_W \\ -\sin\theta_W & \cos\theta_W \end{pmatrix} \begin{pmatrix} W^3_\mu \\ B_\mu \end{pmatrix} \tag{2.42}$$

where θ_W is the weak mixing angle.

Eq. 2.40 can then be written as

$$\frac{g^2\nu^2}{4} W^+_\mu W^{-\mu} + \frac{(g^2 + g'^2)\nu^2}{8} Z_\mu Z^\mu, \tag{2.43}$$

from which we can read-off the mass of W and Z boson to be

$$M_W = \frac{g\nu}{2} \tag{2.44}$$

$$M_Z = \frac{\sqrt{(g^2 + g'^2)}\nu}{2} = \frac{M_W}{\cos\theta_W} \tag{2.45}$$

This completes the electroweak symmetry breaking from which the $SU_L(2) \otimes U_Y(1)$ gauge fields acquire masses.

Finally, the Higgs field itself also acquires a mass through self-coupling in the $V(\phi)$ terms, which becomes

$$V(\phi) = \mu^2 h^2 + \frac{\mu}{\nu} h^3 + \frac{\mu}{4\nu^2} h^4, \tag{2.46}$$

where we can read-off the physical Higgs boson mass from the first term as

$$M_H = \sqrt{(2)}\mu \tag{2.47}$$

Yukawa interaction

With an elementary scalar field, the minimal gauge invariant interaction term between the Higgs field and the fermion fields is

$$-y\bar{\psi}_L\phi\psi_R, \tag{2.48}$$

where y is some coupling constant. Therefore, one can constructs terms with left-handed quark (lepton) doublets with their right-handed singlet components

$$\mathcal{L}_{Yuk} = \sum_j \left(-y_l^j \bar{L}_L^j \phi l_R^j - y_d^j \bar{Q}_L^j \phi d_R^j - y_u^j \bar{Q}_L^j \phi u_R^j + h.c.\right) \tag{2.49}$$

where $j = 1, 2, 3$ denotes the fermion generations, y^j are the Yukawa couplings and similar to Eq. 2.30 the j-th generation fermions fields are

$$L_L^j = \begin{pmatrix} \nu_L^j \\ l_L^j \end{pmatrix}, \quad Q_L^j = \begin{pmatrix} u_L^j \\ d_L^j \end{pmatrix}, \quad \nu_R^j, \quad l_R^j, \quad u_R^j, \quad d_R^j, \tag{2.50}$$

In unitary gauge, Eq. 2.49 becomes

$$\mathcal{L}_{Yuk} = \sum_j -y_f \frac{\nu}{\sqrt{2}}\bar{\psi}_f\psi_f - y_f \frac{1}{\sqrt{2}}\bar{\psi}_f\psi_f h, \tag{2.51}$$

where the charged fermions $f = l, u, d$ are now massive with masses

$$m_f = y_f \frac{\nu}{\sqrt{2}} \tag{2.52}$$

The 2nd term in Eq. 2.51 indicates the coupling between the physical Higgs field h and the fermions are proportional to their masses.

Therefore, the Higgs field gives both the gauge bosons and the charged fermions in the SM masses. What is left out are the neutrino masses, which we cannot write down a Yukawa term like Eq. 2.48 due to the lack of right-handed neutrinos. Thus neutrinos in the SM are strictly massless. However, different flavors neutrinos are experimentally observed to be oscillating among themselves [11, 12, 13], which implies there are some non-zero mass differences between the neutrinos generations and that their flavor eigenstates does not align with their mass eigenstates. The origin of neutrino masses remain one of the unsolved puzzle in particle physics.

2.5 Higgs boson at the LHC

2.5.1 Production and decay

The dominant production mode of Higgs boson at the LHC is the gluon-fusion production, in which gluons fuses into a Higgs boson via a virtual fermion loop. Since the Higgs boson couples proportionally with fermion's mass, the heaviest top-quark dominates the loop contribution. Since this process is loop-induced, QCD radiative correction are important and have been studied in details [14], and the inclusive cross section have reached N3LO accuracy. Other production modes involving weak interactions are naturally suppressed since the LHC is a hadron collider. At LHC with $\sqrt{s} = 13\,\text{TeV}$, the total cross section is around 55.1 pb, and gluon-fusion accounts

for $48.6^{+4.6\%}_{-6.7\%}$ [14]. Figure 2.3 shows the leading order Feynman diagrams of different production modes at the LHC.

On the decay side, at the mass of 125 GeV, the largest branching ratio goes to the heaviest fermion less than the Higgs mass, which is the bottom quarks. Although with the largest branching ratio of around 58%, $b\bar{b}$ final states are relatively difficult to detect in a hadron collider.

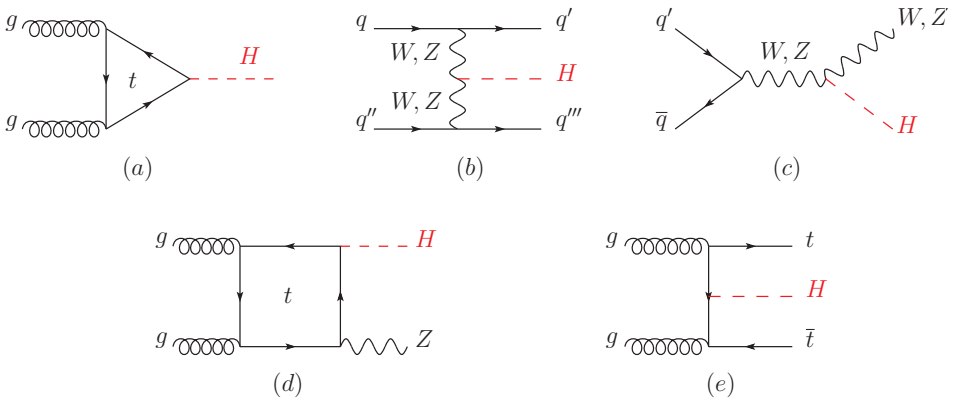


Figure 2.3: Main leading order Feynman diagrams contributing to the Higgs boson production in (a) gluon fusion, (b) Vector-boson fusion, (c, d) associated production with W (tree-level, loop-level) and (e) associated production with a pair of top quarks [15].

2.5.2 Fermionic couplings

As we have seen in Sec. 2.4.1, the Higgs boson generates mass in SM via two inter-related yet distinct mechanisms The electroweak symmetry breaking gives masses to the electroweak vector bosons(W, Z), whereas the fermions in SM obtain the masses through Yukawa couplings with the Higgs boson. It is therefore important to examine both mechanisms in details experimentally.

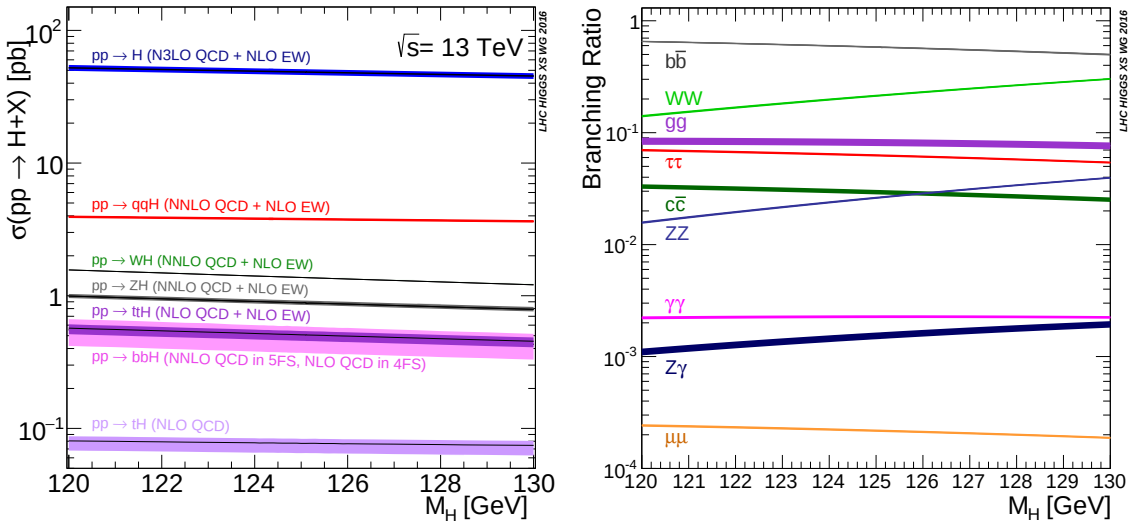


Figure 2.4: Minimal SM Higgs production and decay at the LHC [14]. Left: Production cross sections at $\sqrt{s} = 13\,\text{TeV}$, for $m_\text{H} = 120$–$130\,\text{GeV}$. Right: Decay Branching Fractions for $m_\text{H} = 120$–$130\,\text{GeV}$.

In the initial measurements of the Higgs boson (H) [16, 17, 18] during the LHC Run 1, the coupling between the Higgs boson and the electroweak vector bosons were more established via the associated production channels, whereas the coupling between the Higgs boson and fermions were inferred from the loop-induced di-photon decay channel.

Fermionic couplings were more difficult to establish mainly due to experimental challenges involved in the search channels, such as complicated event topologies and low signal-to-background ratio. Nevertheless, with higher energy and luminosity in the LHC Run 2, the couplings to the charged fermions of the third generations (the bottom quark [19, 20], top quark [21, 22] and the tau lepton [23, 24]) are established.

The event topology challenge were partly overcome with sufficient statistics in finely-binned event categories together with the more matured event classification

techniques using multivariate algorithms such as deep neural network and boosted decision trees.

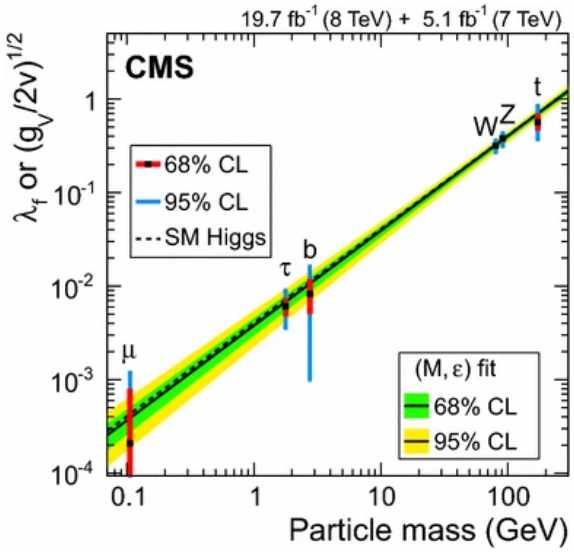


Figure 2.5: Coupling strengths of Higgs boson as a function of particle masses, extracted using LHC run 1 data [25]. Note that the uncertainties on fermionic couplings are significantly larger than the ones in W and Z.

Observing $H \to b\bar{b}$ decay

The observation with the bottom quarks decays were mainly driven by search channels in which the Higgs boson is produced in association with a W or Z boson (VH) decaying to electrons and muons. As will be discussed in Chapter **??**, the presence of energetic electrons and muons in the events are striking experimental signatures in hadron colliders such as LHC, and detectors are designed to measure them with high efficiency and good resolution. Hence, the well-measured electrons and muons can be leveraged to reject majority of SM backgrounds, without losing Higgs signals.

The price to pay in this approach is a significant reduction in cross section. On

the production side, the VH cross section is around 20 times smaller than the most dominant gluon fusion production; furthermore on the decay side, the leptonic decays of vector bosons retain around 20%(WH) and 7%(ZH) of the VH events. This turns out to give reasonable sensitivity, which typically goes as number of signals divided by the square-root of backgrounds.

Probing the $H \to b\bar{b}$ in gluon-gluon fusion (ggF) production, however, is a much more difficult task. As seen in Fig. 2.4, a standalone Higgs boson is produced from two gluons in ggF. Therefore, when the Higgs boson decays to bottom quarks, the final state of the event would be extremely similar to the QCD di-jet process, which has a cross section of the order mb – six orders of magnitudes larger than the ggF Higgs cross section. In the bulk of the kinematic phase space, this was considered as an impossible task.

Another novel approach to search for $H \to b\bar{b}$ decay, is to require the Higgs recoiling a jet with a large boost and achieve a reasonable sensitivity with jet substructure techniques and novel b-tagging inside the merged $b\bar{b}$ jet.

This gives a complementary approach to the VH observation as it leaves *other* production modes unexplored and limits the reach to probe the cross sections at larger transverse momentum (p_{T}) of the Higgs boson because of the reduced cross sections.

As will be discussed in the next section, the cross sections of Higgs bosons produced with high-p_{T} are of particular theoretical interests, but traditional channels have limited sensitivity due to the exponentially falling p_{T} spectrum.

2.5.3 Boosted Higgs boson

Extending SM via EFT

The SM is a remarkably successful theory, predicting cross sections of processes across 9 orders of magnitude and agrees with experimental results. While there are fundamental issues and limitation within the SM that we know SM is not the ultimate theory of elementary particles, it serves as an good approximation to the grand unification theory (GUT) at low energies.

One systematic way to extend the SM is to explore the possible new interaction using the SM fields with an effective coupling constant that are heavily suppressed by a higher energy scale. The is called effective field theory (EFT). In order for the EFT to reduce as SM at low energies and to remain calculable at high energy, the new terms are required to respect the SM gauge symmetries and are renormalizable. The Lagrangian of the EFT [26, 27] can be written as

$$\mathcal{L}_{eff} = \mathcal{L}_{SM} + \sum_i \frac{1}{\Lambda^{d_i-4}} c_i \mathcal{O}_i, \tag{2.53}$$

where $\mathcal{L}_{SM}$ is the SM Lagrangian, Λ is the cut-off scale of the EFT, $\mathcal{O}_i$ are a set of dimension d_i operators that respect $SU_C(3) \otimes SU_L(2) \otimes U_Y(1)$ gauge invariance of $\mathcal{L}_{SM}$ and c_i are the effective couplings. This is similar to how Fermi's theory on beta decay [28] is a good EFT to the underlying electroweak theory, before the experimental energy is comparable to the masses W.

Anomalous Higgs coupling

Being the least explored sector in the SM, it is natural to look for indirect signs of new Physics in the Higgs sector. While under current experimental precision, the newly discovered Higgs agrees with SM predictions in terms of mass, spins and couplings, deviations could be hidden in the differential measurements and could be uncovered as more phase space of the Higgs boson decays are being covered by the experiments.

One such example is the inclusive verses differential cross section of the gluon-gluon fusion, which is the predominant production mode of the Higgs boson, at the LHC. The Higgs low energy theorem [29] ensures that the inclusive production cross section cannot differentiate modifications to the effective top-Higgs coupling, κ_t, and the effective gluon-Higgs coupling, κ_g. Specifically, the modified inclusive cross section $\sigma_{\text{incl}}(\kappa_t, \kappa_g)$ only scales with the *sum* of the two couplings [30],

$$\frac{\sigma_{\text{incl}}(\kappa_t, \kappa_g)}{\sigma_{\text{incl}}{}^{\text{SM}}} \sim (\kappa_t + \kappa_g)^2, \tag{2.54}$$

where $\sigma_{\text{incl}}^{\text{SM}}$ is the SM inclusive Higgs cross section.

In order to break the degeneracy, one could either measure the associated production with top quarks, which is notoriously difficult and only observed recently [21, 22]. Another complementary approach would be to measure the differential cross section of Higgs recoiled with some minimum p_{T} threshold, p_{T}^{min}, with a jet. With the p_{T}^{min}

requirement, Eq. 2.54 becomes

$$\frac{\sigma_{\mathrm{p_T}^{\min}}(\kappa_t, \kappa_g)}{\sigma_{\mathrm{p_T}^{\min \mathrm{SM}}}} = (\kappa_t + \kappa_g)^2 + \delta\kappa_t\kappa_g + \epsilon\kappa_g^2, \tag{2.55}$$

where ϵ and δ are numerical coefficients depending on the p_{T} threshold. With these two complimentary approaches, one could extract valuable top-Higgs coupling modifications information in the SM. Such anomalous couplings can be generated with a additional fermionic top-partners [31], which could make sizable contribution to the cross-section at high-p_{T}.

Besides coupling modification, it is also worth-noting that the differential cross sections at very high boost is also a kind of precision test between theory and experiment, as cross section drops exponentially as a function of p_{T}. As shown in Fig. 2.6, in order to have the same order of accuracy for the boosted Higgs cross-section, extra loop diagrams have to be considered. In other words, the cross section is being probed at a higher energy scale, thus resolving the interactions involved in the production.

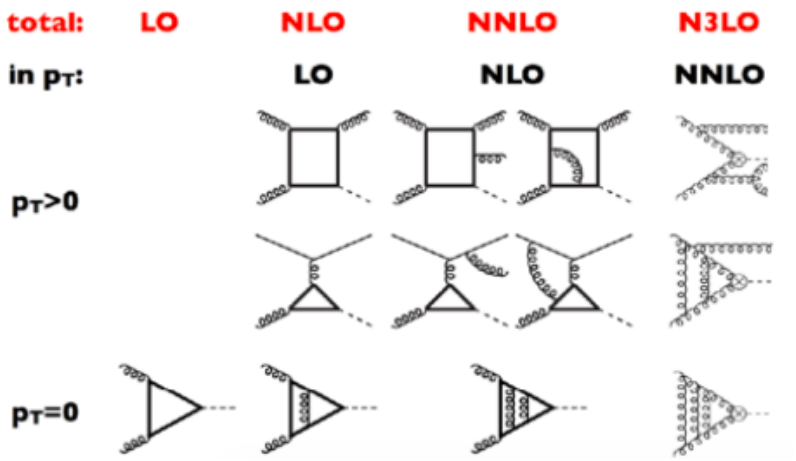


Figure 2.6: Part of the Feynman diagrams involved in predicting Higgs production cross section at different orders.

Chapter 3

The Compact Muon Solenoid Experiment

Since their invention, particle colliders have been a major tool for particle physicists to probe the interactions of the elementary particles. There are two common challenges to all experimental studies of elementary particles: creating the interested particles for study and detecting the particles either directly or indirectly. Particle collider is one of the ways to create an intense source of elementary particles and their decay products, providing a very controlled environment for physics studies. Detectors are built around the interaction point to measure the particle coming out of the collisions.

In this chapter, we will discuss how elementary particles are studied in the largest particle collider program, Large Hadron Collider(LHC), and one of the general purpose particle detector, the Compact Muon Solenoid (CMS), at the LHC.

3.1 The Large Hadron Collider

The Large Hadron Collider is the world's largest particle accelerator built with the highest collision energy. Located in a 26.7 km circular tunnel roughly 100 meters beneath the surface of the France-Switzerland border, the LHC holds the current record for two of the most important parameters of a particle collider: the center-of-mass collision energy($\sqrt{s} = 13\,\text{TeV}$) of proton-proton($pp$) collision and the instantaneous luminosity ($2 \times 10^{34}\text{cm}^{-2}\text{s}^{-1}$). The LHC achieve this by accelerating two opposite circulating proton beams of 6.5 TeV and colliding them at a bunch spacing of 25 ns. The instantaneous luminosity L_{int} delivered by the LHC is given by the following formula

$$L_{\text{int}} = \frac{N_p^2 \cdot n_b \cdot f}{4\pi \cdot \sigma_x \cdot \sigma_y} \times F(\theta_c, \sigma_x, \sigma_y), \tag{3.1}$$

$$F(\theta_c, \sigma_x, \sigma_y) = \left(1 + \left(\frac{\sigma_z}{\sigma_x}\right)^2 \left(\frac{\theta_c}{2}\right)\right)^{-1/2}, \tag{3.2}$$

where N_p is the number of protons in a bunch, n_b is the number of bunches in a beam, f is the frequency of the beam revolution, σ_x, σ_y and σ_z are transverse and longitudinal beam width characterized by the beam optics, F accounts for luminosity reduction due to the finite beam crossing angle θ_c [32]. Values of the above parameter at the LHC design peak luminosity is shown in Table 3.1. The unprecedented luminosity of the LHC is largely a result of completing the engineering challenge of realizing these parameters. The number of observed events N_{obs} from the collisions is proportional

to the integrated luminosity, L, over time,

$$N_{obs} = \epsilon \cdot \sigma \cdot L, \tag{3.3}$$

where ϵ is the efficiency any experimental procedures to observe the events given the production in collision, σ is the cross-section of the underlying process of the events and $L = \int L_{\text{int}} dt$.

Symbol	Description	Value
N_p	Number of protons in a bunch	1.15×10^{11}
n_b	Number of bunches	2808
f	Frequency of revolution	11.245 Hz
σ_x, σ_y	RMS transverse beam size at CMS interaction point	16.7 μm
σ_z	RMS longitudinal bunch length	7.55 cm
F	Geometric luminosity reduction factor	0.836
θ_c	Beam crossing angle at CMS interaction point	285 μ rad

Table 3.1: Parameters of instantaneous luminosity equation at the design peak luminosity of $1.0 \times 10^{34}\text{cm}^{-2}\text{s}^{-1}$ [32].

The LHC acceleration chain

The LHC reaches its collision energy through multiple stages acceleration in the CERN accelerator complex shown in Fig. 3.1. The proton source comes from hydrogen atoms with electrons removed. The acceleration chain starts with a linear accelerator("Linac2") which accelerates the protons to 50 MeV. The protons are injected in the Proton Synchrotron Booster to reach the energy of 25 GeV and then are fed to the Super Proton Synchrotron (SPS), which accerlerates the protons to 450

GeV before injecting them in the LHC. The protons are separated into two counter-rotating breams at the LHC, where the final collision energy of 6.5 TeV is reach after approximately 20 minutes of gradual ramping in beam energy.

Superconducting dipole magnets capable of generate a magnetic field of 8 T are used in the curved sections of the LHC tunnel to guide the beam to follow an approximate circular trajectory. Sixteen radiofrequency (RF) cavities are then used in the straight sections of the LHC tunnel to accelerate the protons with an oscillating electric field. The oscillating field will impart energy to the protons when the they are correctly synchronized time-wise with the electric field accelerating along the direction of motion. Finally, superconducting quadrupole magnets are used near the interaction points (IPs) to focus the beam to a transverse size on the order of millimeter in order to increase the interaction rate at the IPs.

Pileup interactions

When the two bunches of protons cross each others at the interaction point, multiple interactions could take place due to the large number of protons in the bunches and the precise focusing optics around the IP. The interaction which creates the most energetic vertex in the detector, the primary vertex, is often the most interesting event, due to the large energy exchange in the interaction. The remaining vertices from the other interactions are called pileup (PU) vertices. The exact number of PU vertices depends on the number of protons in the proton bunches and other running conditions, therefor it varies between data-taking periods and even within the same

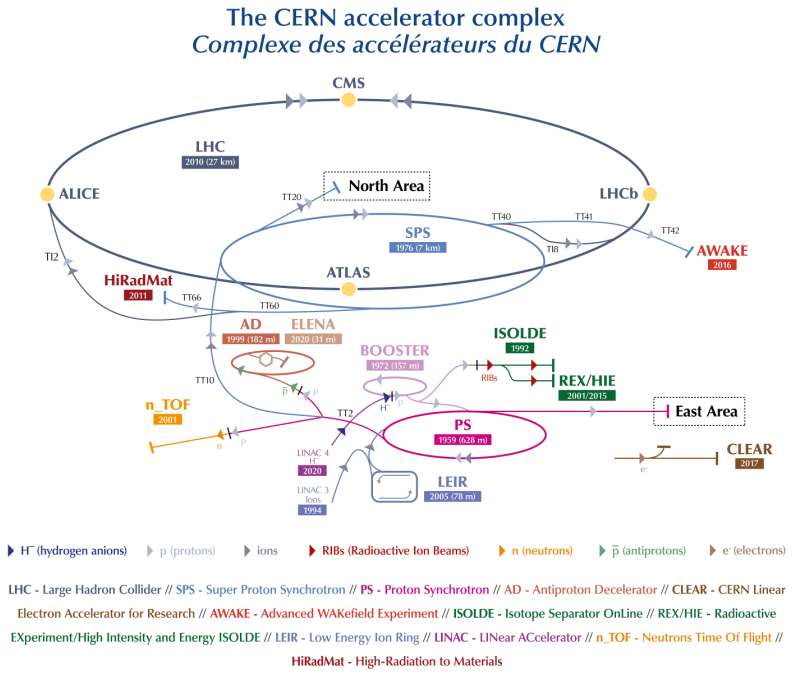


Figure 3.1: The CERN accelerator complex [33].

fill of protons. Figure 3.2 shows the measured average number of interactions per crossing for pp collisions in different data-taking years in the LHC Run II. These measurements are essential inputs to accurately simulate the effect of PU in the event reconstructions.

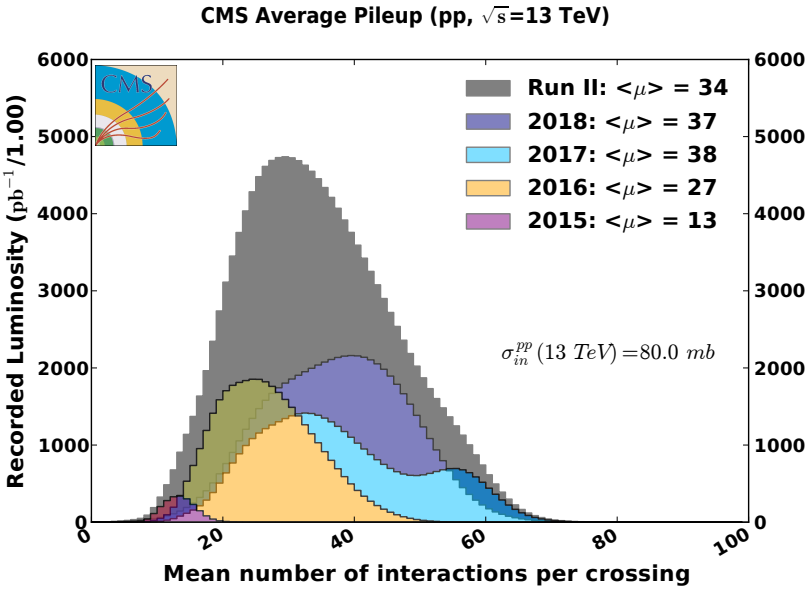


Figure 3.2: Distribution of the average number of interactions per crossing (pileup) for pp collisions in different data-taking years in the LHC Run II. The overall mean values and the minimum bias cross sections, which is pp collision cross section measured with minimum experimental selections, are also shown.

The LHC run II

The LHC started delivered the first physics run at $\sqrt{s} = 7\,\text{TeV}$ in 2010, after delayed by the magnet quench accident when it was brought to its the first operation in 2008. After a very successful Run I between 2010 and 2012, delivering almost $30\,\text{fb}^{-1}$, LHC underwent 2 years of Long Shutdown (LS1) and began its Run II operation in 2015,

which is the first run with $\sqrt{s} = 13\,\text{TeV}$. With the experience of operating the LHC and the respective detectors at $\sqrt{s} = 13\,\text{TeV}$ and bunch spacing of 25 ns in 2015, the LHC ramped up the intensity quickly in 2016 and deliver $> 97\%$ of all Run II data in the following 3 years. The data used in this thesis correspond to a total of 137 fb^{-1} of pp collision data at $\sqrt{s} = 13\,\text{TeV}$ collected between 2016 and 2018 by the CMS experiment.

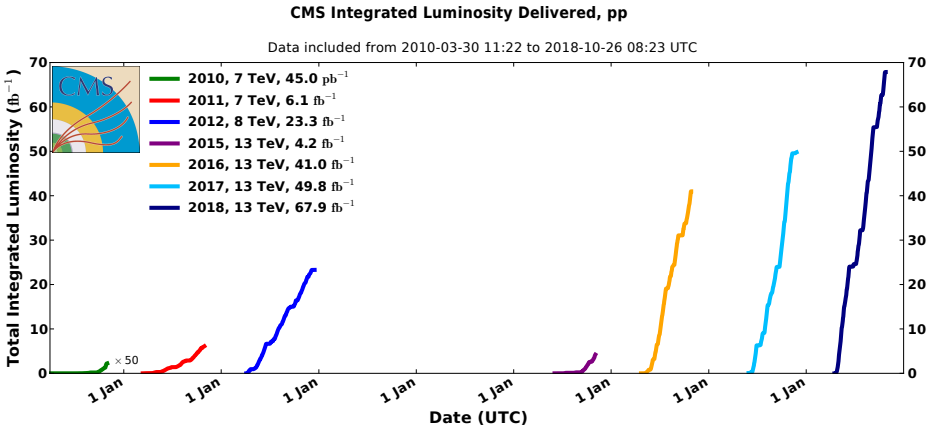


Figure 3.3: Cumulative luminosity versus day delivered to CMS during stable beams for pp collisions at nominal center-of-mass energy.

3.2 The CMS detector

The Compact Muon Solenoid detector (CMS) is a general-purpose particle detector at the LHC. By general-purpose, it means the design goal of the detector is to measure the energy, momentum, charge and other properties of all the particles emerging from the collision point. Therefore, CMS is constructed with a hermetic coverage around the interaction point (IP5) using several layers of sub-detectors, each specialized at

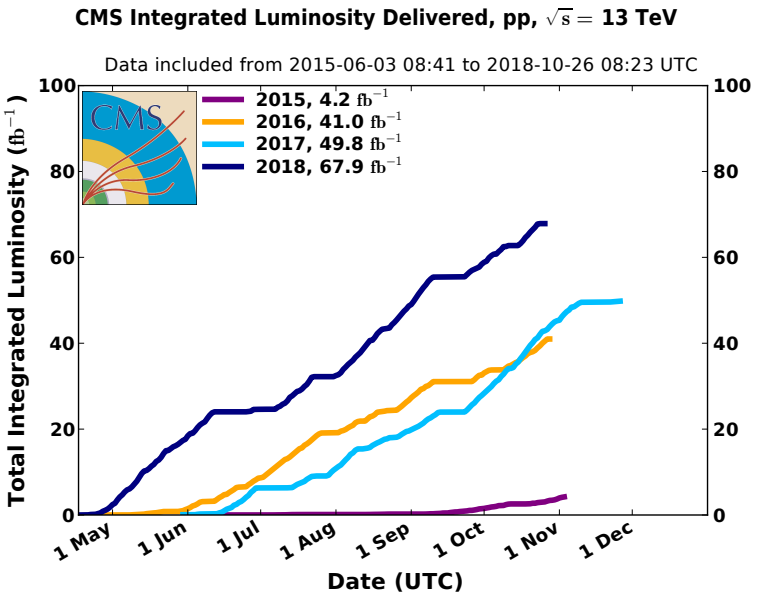


Figure 3.4: Cumulative luminosity versus day delivered to CMS during stable beams for *pp* collisions in LHC Run II.

measuring a class of particles. The coverage is provided in a cylindrical geometry, with each sub-detectors further divided into barrel region and end-cap regions. The central feature of the detector is a superconducting solenoid of 6 m internal radius, providing a 3.8 T nearly homogeneous magnetic field within its volume. The detector weighs about 14,000 tons with a radius of 15 m and a length of 28.7 m.

From the interaction point outwards, the CMS sub-detector consists of a finely granulated silicon tracker system, an electromagnetic calorimeter (ECAL), followed by a hadonic calorimeter (HCAL) and the outermost Muon system. Figure 3.5 shows the layout of the interior of the CMS detector.

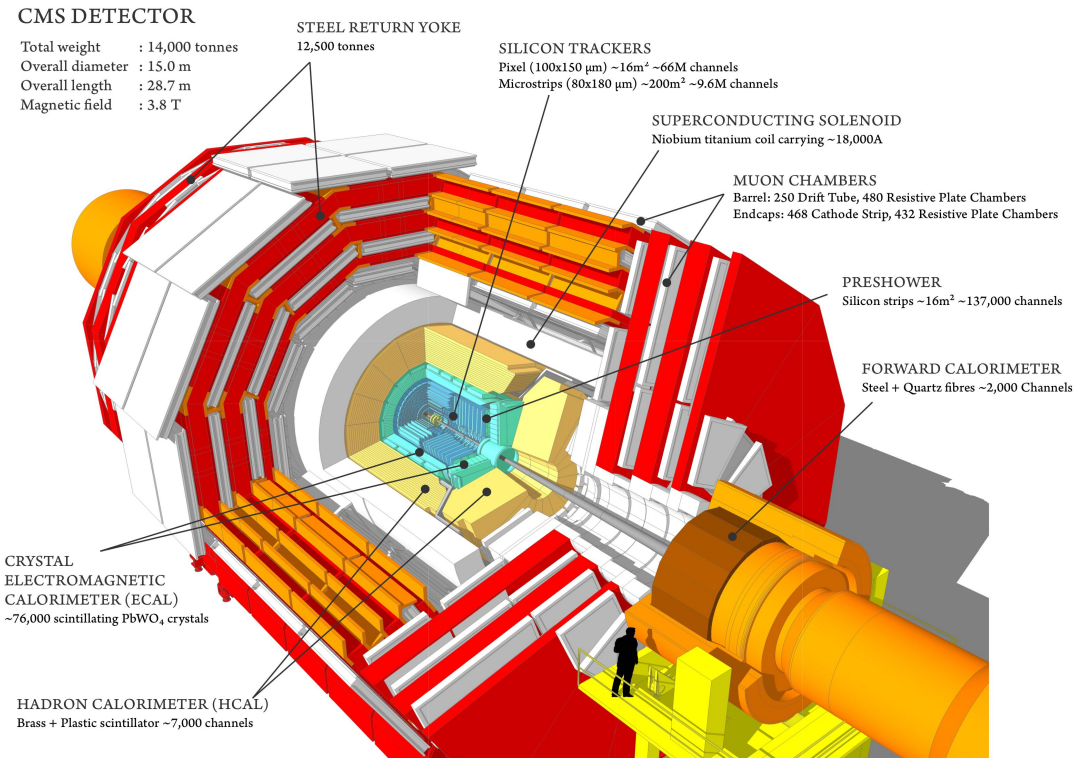


Figure 3.5: Cut-out diagram of the CMS detector [34].

3.2.1 Coordinate System

CMS uses a coordinate right-handed coordinate system where the origin is centered at the nominal *pp* collision point. The positive z-axis points along the beam line going counter-clockwise when looking at the LHC from above, while the positive the y-axis points vertical upward, leaving the positive x-axis pointing radially inwards towards the LHC center. Due to the cylindrical arrangement of the detector components, it is more often to use cylindrical coordinates for detector or event descriptions. The azimuthall angle ϕ is measured from the x-axis and spans in the x-y plane, where the radial coordinate in this plane is denoted by r. The polar angle, θ, is measured from the positive z-axis. It is, however, more common to use pseudorapidity $\eta = -\ln(\tan(\theta/2))$ instead of θ. Pseudorapidity has the advantage that, in the limit of highly relativistic particles, it is invariant under a Lorentz boost along the z-axis.

3.2.2 The Pixel Detector

The pixel detector is the inner-most sub-detector system of CMS, which therefore is crucial for precision tracking of the trajectory of charged particles. It is particularly important for the reconstruction of the secondary vertices and impart parameters of tracks which are some key features for identifying jets originated from a b-quark. The pixel detector measures the three-dimensional coordinates of hits from charged particles with layers of reverse-biased silicon pixel sensors.

The pixel detector is composed of 1440 pixel detector modules with a total of 66 million channels. Each pixel cell is 100×150 μm^2 in size. Three layers of pixel

modules are arranged in a 53-cm-long barrel layers (BPix) at $r = 4.4, 7.3$ and 10.2 cm, while two end-cap layer disks (FPix) of radii ranging from 6 to 15 cm are placed at each end at $|z| = 34$ and $|z| = 46.5$ cm respectively. The pixel detectors covers a pseudorapidity range of $-2.5 < \eta < 2.5$. For a 100 GeV muon, the pixel detector is designed to achieve a 1-2% transverse momentum resolution up to $|\eta| = 1.6$ and around 10 μm transverse impact parameter resolution within $|\eta| < 2.5$ [35].

The pixel detector was designed to withstand the LHC design luminosity of $1 \times 10^{34}\text{cm}^{-2}\text{s}^{-1}$, which was already exceeded in 2016. In the year-end technical stop between 2016 and 2017, the phase-1 upgrade of the pixel detector is installed, which replaced the original pixel modules. The phase-1 pixel detector added an additional fourth layer of pixel modules in the barrel region and a third disk per side in the end cap region, resulting nearly twice as much total pixels at 124 million. The innermost pixel layer was also placed closer to the beam pipe which was installed during the LS1. The layout of the pixel detector before and after the phase-1 upgrade is shown in Fig. 3.6. With these improvements, the pixel detector is able to maintain a hit efficiency $> 97\%$ at $2 \times 10^{34}\text{cm}^{-2}\text{s}^{-1}$ in 2018 data-taking period [36].

3.2.3 Silicon Strip Tracker

As the distance from the interaction point increases, the detector volume increases rapidly and thus become unfeasible to maintain the same detector granularity as the innermost pixel detector, both due to costs and difficulty of reading out channels. Instead of pixels, silicon strips with typical cell size of 10 cm $\times$ 80 μ are used in the

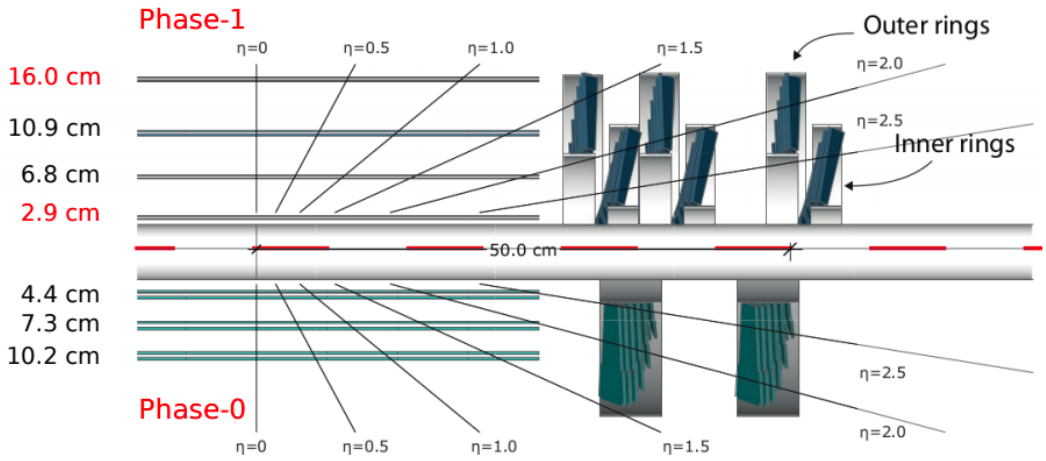


Figure 3.6: Comparison of the layouts of the Phase-1 pixel detector (top) and the original pixel detector before the year-end technical stop between 2016 and 2017 (bottom) [37].

radial region between 20 and 116 cm, covering almost 200 m^2, around 200 times of that of the pixel detector, with 9.3 million strips. Adjacent strips sensors are arranged with a small relative angles between them such that multiple 1D strips measurements can be used to construct a 2D position. The strip modules are divided in 3 different subsystems, namely the Tracker Inner Barrel and Disk (TIB/TID), the Tracker Outer Barrel (TOB) and the Tracker Endcaps (TEC). There are 4 layers of modules in TIB, extending in radius up to 55 cm and supplemented by 3 disks at each ends of TID. The TOB consists of 6 barrel layers, covering radius up to 116 cm, while the TEC is composed of 9 disks, carrying up to 7 rings of strip sensors. The whole tracker system extends in z between $\pm$ 282 cm. Figure 3.7 illustrates the arrangement of tracker subsystems.

The strip trackers are designed to withstand the radiation from integrated luminosity of 500 fb^{-1} corresponding to about 10 years of LHC operation [38].

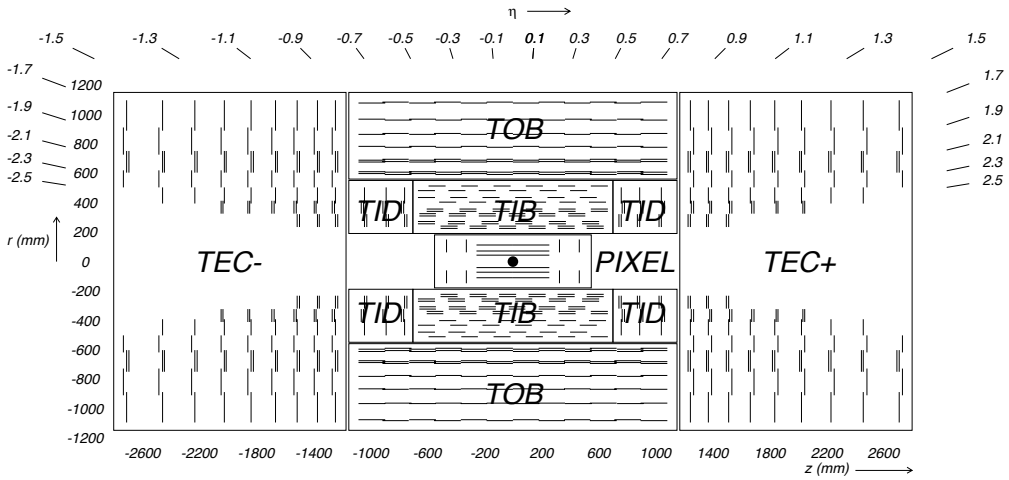


Figure 3.7: Schematic diagram the CMS tracker [34].

3.2.4 Electromagnetic Calorimeter (ECAL)

The electromagnetic calorimeter (ECAL) of CMS is a homogeneous and hermetic calorimeter made of 61,200 lead tungstate ($PbWO_4$) crystals, designed to measure the energy of electrons and photons by triggering the electromagnetic shower in the dense crystals. Lead tungstate crystal is chosen to be the active material due to its high density (8.28 g/cm^3), short radiation length (0.89 cm) and small Moliére radius (2.2 cm). It is also very fast – about 80% the scintillation light in the crystals is emitted within 25 ns, which is the LHC bunch crossing time. The crystals are also radiation-hard, with the ability to maintain a good ECAL performance throughout LHC operations.

In the barrel part of the ECAL (EB), the $PbWO_4$ crystals are approximately $2.2 \times 2.2 \times 23\text{cm}^3$ in size, each covering roughly a square of size 0.0174×0.0174 in the $\eta - \phi$ plane, or 1 degree in ϕ and 1/85 of the full η range of the barrel. These crystals are grouped into 36 supermodules, each spanning 20 degree in ϕ. The crystal length corresponds to 25.8 radiation length X_0. The endcaps (EE) crystals are approximately $2.2 \times 2.2 \times 22\text{cm}^3$ in size, grouped in a 5×5 unit called supercrystals, which are arranged in the regular $x - y$ grid. The EB covers a pseudorapidity range of $|\eta| < 1.479$ whereas the EE covers $1.479 < |\eta| < 3.0$. In front of the EE, the preshower detector, made of layers of lead and silicon strip sensors, is installed to identify neutral pions and improves identification of electrons against minimum ionizing particles. Figure 3.8 shows the layout of different components of ECAL.

The ECAL was designed with the capability to detect the photons from the decay of the Higgs boson as a driving criteria. The homogeneous crystal provides an excellent energy resolution to achieve this purpose. The electromagnetic shower develops inside these crystals as electrons and photons cascade into each others through bremsstrahlung and photon conversion. A typical energy resolution measured in EB is found to be

$$\left(\frac{\sigma}{E}\right)^2 = \left(\frac{2.8\%}{\sqrt{E}}\right)^2 + \left(\frac{0.12}{E}\right)^2 + (0.3\%)^2 \tag{3.4}$$

where E is measured in GeV, the first, second and third term correspond to the stochastic, the noise and the constant term respectively [39]. The homogeneous nature of the ECAL crystals reduces the stochastic term by minimizing the lost of scintillation photons in the absorber. The ECAL is designed to contain more than 98% of the

energy of electrons and photons up to 1 TeV and cause up to two thirds of hadrons passing passing through them to shower [39].

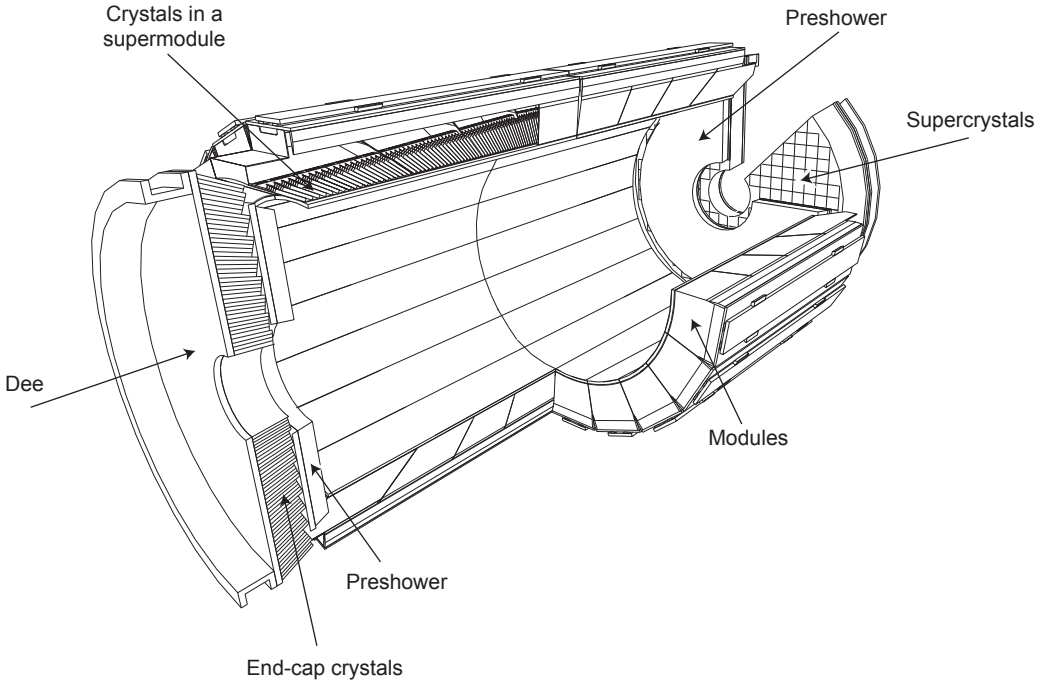


Figure 3.8: Layout of the CMS electromagnetic calorimeter [37].

3.2.5 Hadronic Calorimeter (HCAL)

The hadronic calorimeter (HCAL) of CMS is a sampling calorimeter made of alternating layers of absorbers and active layers, designed to measure the energy of charged and neutral hadrons. Located outside of ECAL, HCAL consists of 4 sub-detectors to cover a wide pseudorapidity range up to $|\eta| = 5.2$, namely the barrel and end-cap (HB,HE), the forward calorimeter (HF) and the outer calorimeter (HO). The

HB consists of 36 identical azimuthal wedges made with bass absobers and plastic scintillators as active materials. The plastic scintillators are segmented into 16 η sectors, resulting a granularity of $\Delta\eta \times \Delta\phi = 0.087 \times 0.087$. A total number of 17 brass-scintillators layers consisting of $\sim$ 5 cm brass and 3.7 mm scintillators give an interaction lengths (λ_I) of 5.82 at $\eta = 0$, which increases up to 10.6 λ_I at $|\eta| = 1.3$. The ECAL in front of HB adds about 1.1 λ_I.

In HE, the same brass-scintillators materials made up 19 sampling layers using thicker (7.9 cm) brass plates. The total interaction length is about 10 $lamdba_I$, including ECAL's contribution. The scintillators are arranged in trays of tiles called megatiles. The granularity of each tile is $\Delta\eta \times \Delta\phi = 0.087 \times 0.087$ for $|\eta| < 1.6$ and $\Delta\eta \times \Delta\phi \approx 0.17 \times 0.17$ for higher $|\eta| \geq 1.6$. Each side of HE consists of 36 megatiles, each of which spans 10 degree in ϕ. The HB and HE cover
$eta| < 1.3$ and $1.3 < |\eta| < 3.0$ respectively. The HO utilizes the solenoid coil as an additional absorber to serve as a tail catcher for late showers to compensate the smaller radiation lengths at low η. HF covers the very forward pseudorapidity range of $3 < |\eta| < 5$ with steel absorbers and Cerenkov radiating quartz fibers as active material to withstand the high radiation in the forward region.

The scintillation lights in the megatiles are guided towards photodetectors, where optical signals are converted into electrical pulses. Both HB, HE use hybrid photodiodes(HPD) as the photodetectors where as HF uses the more radiation resistant photomultiplyer tube (PMT) and HO uses silicon photomultiplier (SiPM).

Since hadronic showers start the developments within the ECAL, the hadronic

energy resolution depends on both ECAL and HCAL. Due to the sampling nature of HCAL, the energy resolution is generally worse than ECAL. The typical resolution measured to be

$$\left(\frac{\sigma}{E}\right)^2 = \left(\frac{0.847}{\sqrt{E}}\right)^2 + (0.074)^2 \tag{3.5}$$

in the barrel region [40]. Note that the stochastic term is larger than then resolution in ECAL. The endcap region have similar resolutions to the barrel.

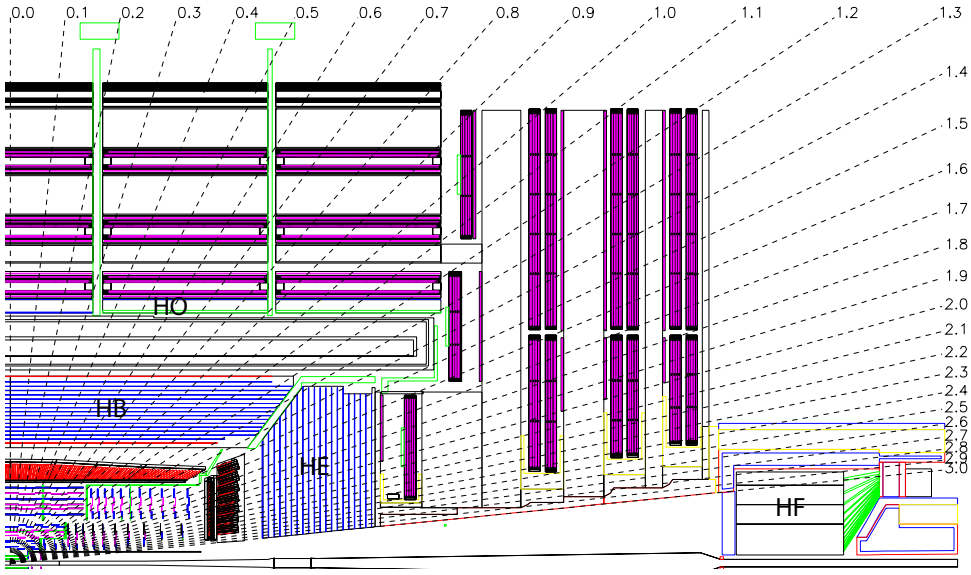


Figure 3.9: Layout of the CMS hadronic calorimeter [37].

HCAL Phase-1 upgrade

While the scintillators in HCAL are chosen to withstand the radiation damage throughout the LHC operation, the HPD shows premature aging due to radiation damage during the LHC Run 2. Figure 3.10 shows the drop of HE scintillators response as

a function of integrated luminosity. SiPMs was chosen to replace the HPDs in the HCAL Phase 1 upgrade. SiPMs have three times the photodetection efficiency and approximately 200 times the gain of HPDs at a much smaller bias voltage of less than 100 V [41]. With a lower signal-to-noise ratio of the SiPMs and the advancement in data link technology, the upgraded HCAL can handle a significant increase in depth segmentation, which allows a more detailed reconstruction of hadronic shower development. Figure 3.11 shows the depth segmentation of HB and HE before and after the Phase-1 upgrade. The backend electronics of HE and HB, which process digital

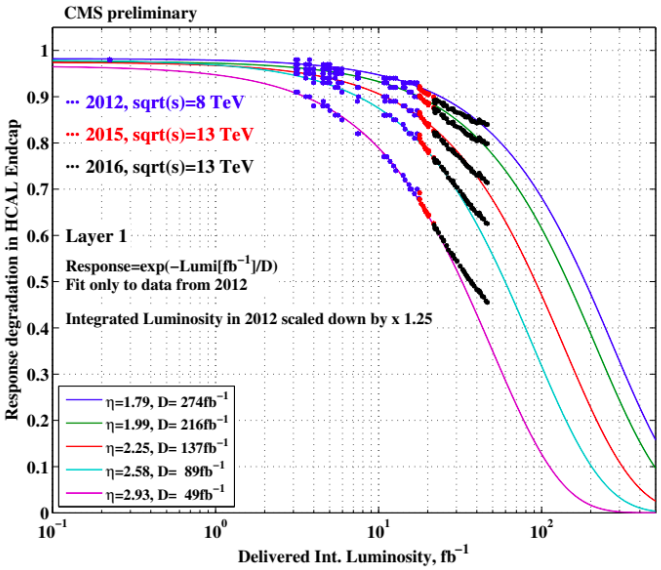


Figure 3.10: Response of the HE scintillators tiles as a function of the delivered integrated luminosity for several pseudorapidity values. The blue, red and black dots represent response measurements performed during 2012, 2015, and 2016, respectively. The colored lines are fits to the 2012 measurements for different η values [42].

signals from the on-detector electronics and propagates them to the data-acquisition (DAQ) path, were first upgraded in the year-end technical stop (YETS) in 2017 and 2018 respectively, in preparation of handling the increased number of channels. The

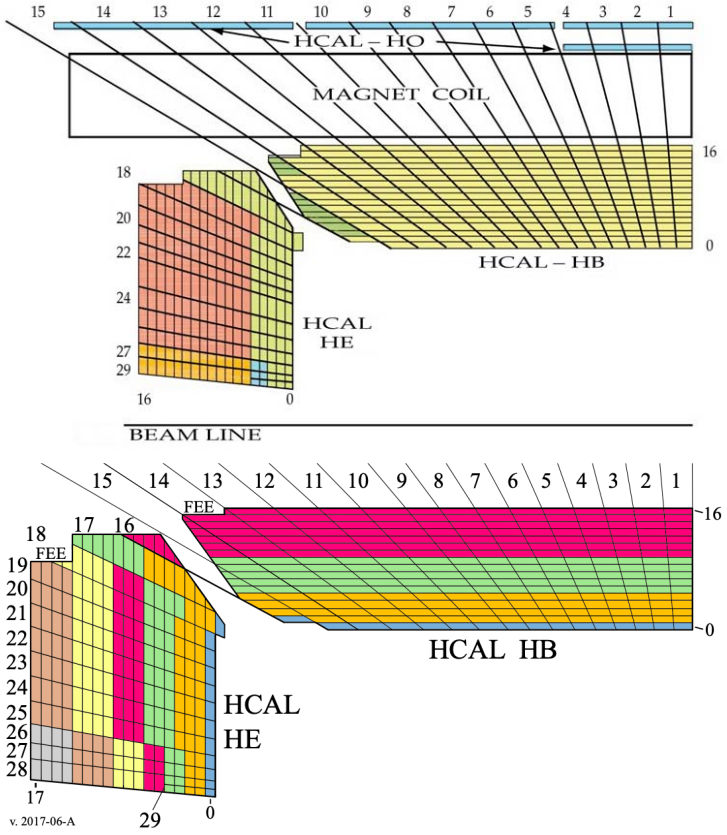


Figure 3.11: Depth segmentation of CMS HCAL before (top) and after (bottom) Phase-1 upgrade [37, 41].

front end electronics, including the SiPMs, of HE and HB were installed in YETS 2018 and 2019 respectively. The PMTs of HF was also upgraded to have multi-anode readout in the 2016 YETS to reduce the anomalous signals due the direct hits of particles on anode readout. The full HCAL phase 1 upgrade was successfully completed in October 2019.

HCAL Run Control

With the evolving set of detector and read-out components, it is important for the corresponding control software to be both flexible for including new components and robust for continuous operations.

To achieve the required flexibility, the run control software develops over the course of LHC Run 2 a one-to-one correspondence between HCAL backend electronic structures and software components. The control software is organized in a 3-level hierarchical structure of software objects called Function Manager (FM) as shown in Fig. 3.12. The top level (LV0) FM is the CMS-wide control structure, which controls the level-1 FMs, each of which corresponds to one CMS sub-system, such as HCAL and ECAL. This gives CMS the flexibility to include or exclude any sub-systems during calibration or testing phases, and since CMS has a static number of sub-systems, this level of flexibility was planned into the run control software framework [43]. The HCAL run control software took the same idea one step further, creating level-2 FMs that each controls a partition of HCAL with common clock signal (a TCDS partition). Figure 3.13 shows the map of these partitions in the $\eta - \phi$ plane segmented in HCAL granularity. Creating an extra software layer of control allowed

commissioning work to happen in parallel between different sectors of HCAL, which was essential during the tight schedule of YETS. Finally, each level-2 FM controls the electronic-specific software applications, called XDAQ. Typical XDAQ applications are readout unit application (`uHTRManager`) and backplane clock distribution unit of a crate (`DTCManager`). The separation XDAQ applications from FMs decouples the changes in front-end electronics from the static control structure, which concerns only the partitioning of HCAL. Thus, this gives stability over the phase-1 upgrade period.

In order to take full advantage of this FM-partition correspondence structure, a fine-grained masking scheme is developed, where any specific component can be removed from a run if requested by the user. Masking is implemented in the partition level (LV2-FMs), in the backend crate level (group of XDAQ applications), and down to individual XDAQ application. Masking greatly simplifies the need to create combinatorics of individual configurations using different parts of subsystems.

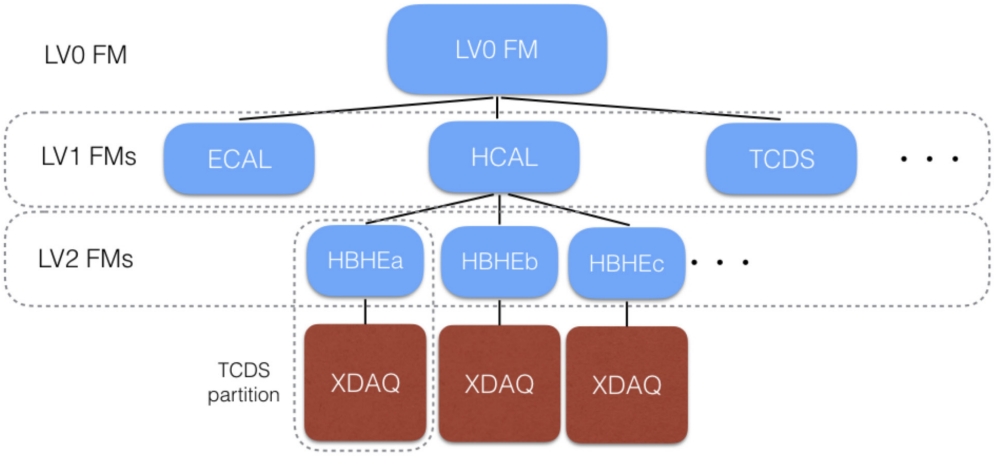


Figure 3.12: Three-level hierarchical structure of CMS run control software, see text for the descriptions of each entity.

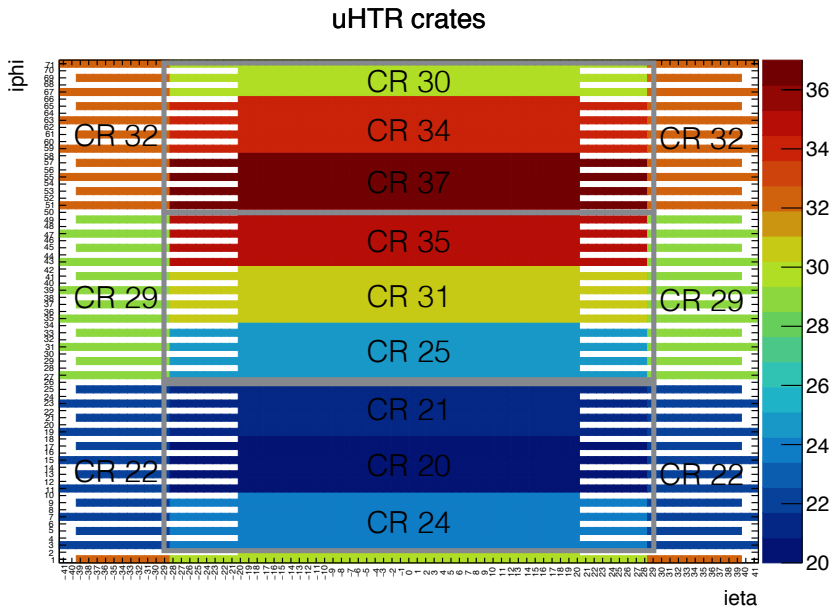


Figure 3.13: Map of HCAL (HB, HE and HF only) backend electronics structure, called uHTR crates, in the $\eta - \phi$ plane segmented in HCAL granularity. Each grid corresponds to one read-out channel. The gray boxes covers the crates corresponding to the partition HBHEa, HBHEb and HBHEc, whereas the crates in the large $i\eta$ regions belongs to HF.

To achieve robustness, the control software implements a common state machine defined in the software framework, as illustrated in Fig. 3.14. The statemachine is implemented via asynchronous notifications. To be specific, the initiation of state transition comes in a top-down direction, causing the down-stream FMs to enter transition states (e.g configuration). Lower level FMs then propagates the transition command down to XDAQ application which executes the command (e.g. configure) on the relevant electronics. Upon completion of the command, XDAQ sends notification back to respective FMs, which updates its own state to determine if the transition should be completed (e.g. go to configured if all controlling XDAQ applications reports configured). State updates are then propagated upwards to complete the full transition of all CMS subsystems. This approach propagates the commands throughout the system efficiently and ensures the top-level state corresponds to an exact set of states of hardware, thus protecting hardware from receiving unwanted commands.

3.2.6 Muon System

Located in the outermost part of CMS, the muon system consists of about 25,000 m^2 of detection planes, therefore the muon system components have to be inexpensive, reliable and robust, which is well satisfied by gaseous detectors. Bearing muon in the experiments name, precise reconstruction and identification of high energy muons is one of the central design concepts. The goal is to achieve $\sim$ 1% di-muon mass resolution for 100 GeV muons and to be able to determine the charge of muons up

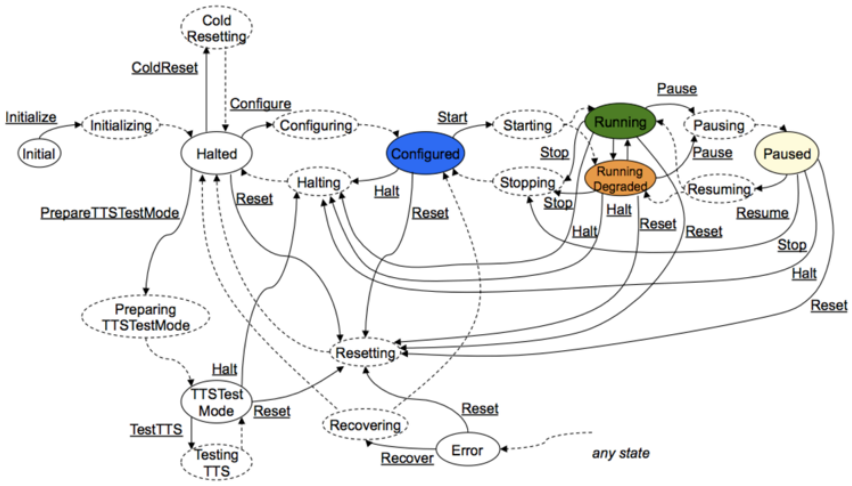


Figure 3.14: State-machine structure of CMS run control software [43].

to 1 TeV.

The muon system consists of 3 gaseous subdetectors, namely the Drift Tubes (DTs), the Cathode Strip Chambers (CSCs) and the Resistive Plate Chambers (RPCs). These subdetector stations are interspersed within the iron return-yoke of the solenoid magnet, which is magnetized steels structure to help maintaining a nearly homogeneous magnetic field throughout the detector volume *outside* of the Solenoid. The DTs and CSCs cover the barrel and endcaps region respectively, providing high position resolution with the wire chambers in the $r-\phi$ plane. RPC provides additional coverage in both the barrel and the endcap regions with fast, independent and highly segmented detectors for triggering. Figure 3.20 shows the layout of the CMS muon system.

The DT system consists of 4 stations, where each station consists of 8 layers of tubes measuring the position in the $r-\phi$ plane and 4 layers in the longitudinal plane

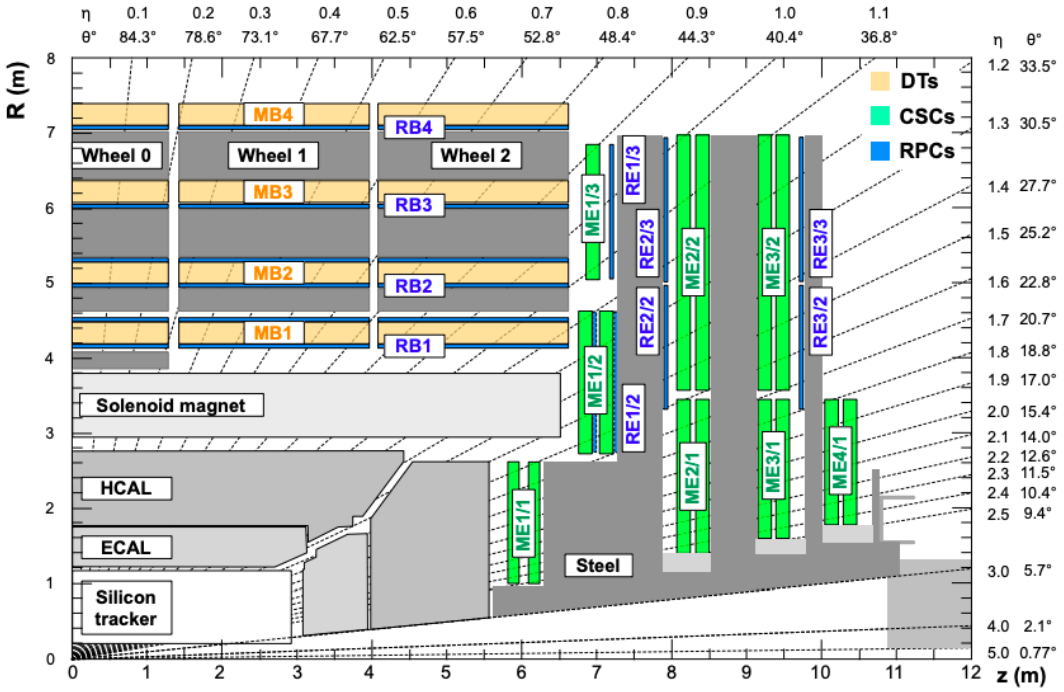


Figure 3.15: Layout of the CMS muon system [37].

(except the outermost layer). A drift cell is the basic unit of the DT system, which is an aluminium cathode cell with a 50-μm-diameter gold-plated anode wire at the center, filled with a gas mixture of Ar (85%) and CO_2 (15%). The CSC system also consists of 4 stations, where each consists of 6 layers. The basic unit of CSC are multiwire proportional chambers, with an array of anode wires arranged at an angle to the cathode strips. The CSC chambers are filled with a gas mixture of 50% CO_2, 40% Ar and 10% CF_4. The DT covers pseudorapidity range of $|\eta| < 1.2$, where as CSC covers the $0.9 < |\eta| < 2.4$ range. The RPCs are fast gaseous parallel-plate detectors that has a timing resolution of about 1 ns, covering $|\eta| < 1.6$ in both the barrel and endcap regions. Figure 3.16 shows the basic units of the DT and CSC system.

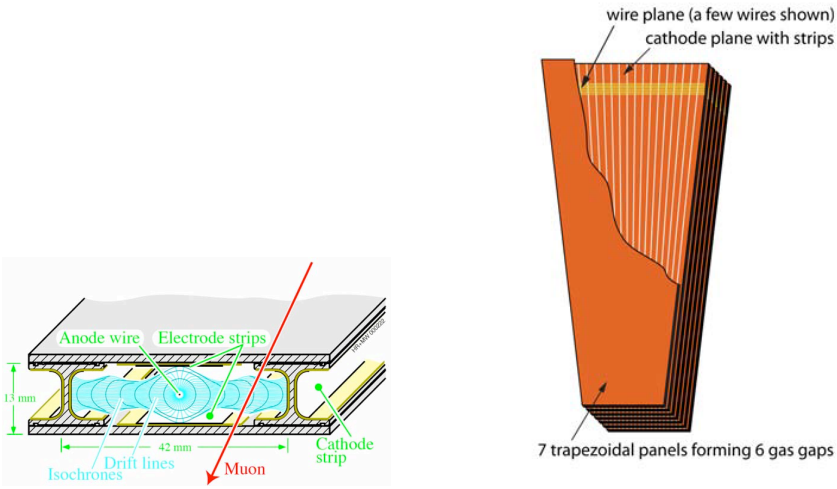


Figure 3.16: Diagram of basic unit of DT and CSC system, showing the drift cell (left) and cathode stripe chamber (right) [37].

3.2.7 Trigger System

The LHC proton bunch crossing interval is 25 ns, corresponding to a crossing frequency of 40 MHz. It is therefore unfeasible to readout millions of CMS channels in between two bunch crossings. Trigger system is designed to reduce this rate to a manageable level, at least a factor of 10^6, for data storage and detailed offline physics analysis. CMS uses a two-tiered trigger system, called the Level-1 (L1) Trigger and High-level Trigger, to reduce this rate to around 1 kHz of total bandwidth for $O(100)$ different physics processes. The L1 trigger consists of custom-designed electronics with the output rate limit of 100 kHz, whereas the HLT is a software based system implemented in a computer farm.

The L1 triggers collect coarsely segmented data from the calorimeter and muon systems to make fast trigger decisions called L1 Accept (L1A), while holding the high-resolution data in the electronics buffers. The HLT has access to the complete read-out data, including those from inner trackers, which allows more complicated trigger algorithms to be computed, such as the ones that involve b-tagging. Given the software-based nature of HLT, the list of HLT triggers, called HLT menu, can be changed relatively easily over time, depending on the physics interests of CMS.

The L1A is constructed from local calorimeter (HF, HCAL and ECAL) and muon (RPC, CSC and DT) trigger information to form stages of global trigger decision. Figure 3.17 illustrates this architecture. On the calorimeter side, trigger towers are formed from the transverse energy (E_T) of 25 ECAL crystals and HCAL read-out towers summing the HB and HE segments in the same η, ϕ area of 0.087×0.087. The

global calorimeter trigger then constructs upto 4 central/tau/forward jets, calculates the scalar sum of jet transverse momenta (H_T) and missing transverse energy (MET). The τ-veto bits, electron/photon candidates, minimum-ionizing particle (MIP) and isolation (ISO) bits are also generated in the global calorimeter trigger.

On the muon side, local track segments are formed with local DT, CSC and RPC measurements, which are then combined in the global muon trigger to select up to 4 muon candidates with the MIP/ISO information from the calorimeter trigger side. The regional information from DT and CSC are track p_{T}, η and ϕ in barrel and endcaps respectively, whereas RPC provides precise bunch crossing information and less precise p_{T} measurement from comparing hit patterns with template patterns.

The maximum latency of L1 trigger to make trigger decision is 3.2 μs.

3.3 Physics Objects Reconstruction

The goal of physics objects reconstruction is to identify the particles types, energy and momentum, given the detector measurements. In practice, there are only a few types of particles with sufficient lifetime before decaying to leave signatures in the detectors. They are muons, electrons, photons and both charged and neutral hadrons. For charged particles, the magnetic fields in the detector bends the trajectories and allow charges and momenta to be measured. For neutral particles, the energies are measured through energy deposits when interacting with detector materials, usually in the form of scintillation lights or ionization. In CMS, electrons and photons, which develop electromagnetic showers when traveling through dense materials, are

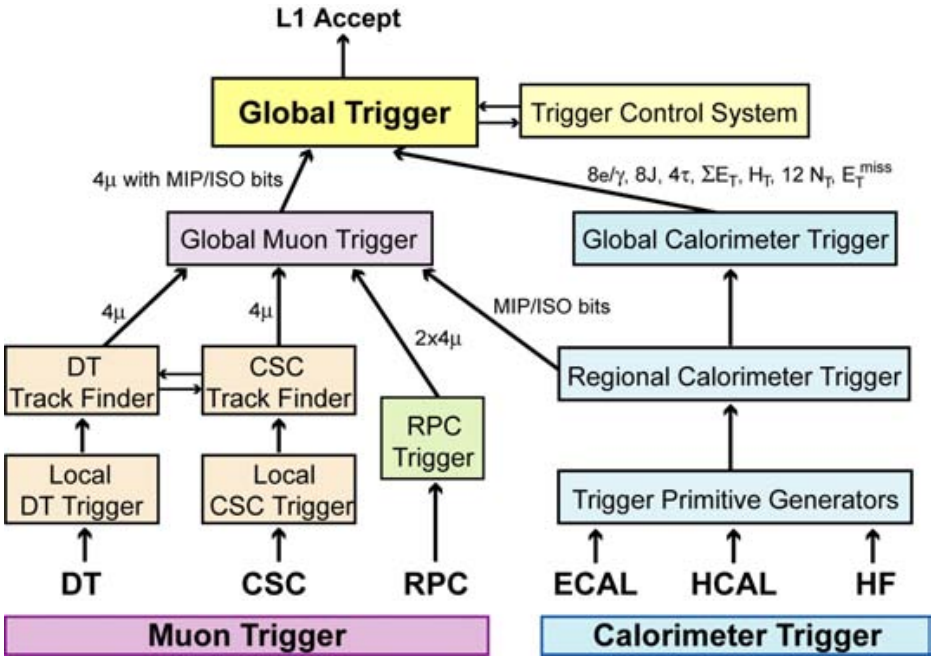


Figure 3.17: Level-1 Trigger system architecture [37].

absorbed in the ECAL. With 200 times of the electron mass, muons leave little or no energies in the calorimeters, but leave hits in both inner trackers and the muon stations outside the magnet. Charged and neutral hadrons may initiate a hadronic shower in ECAL and are subsequently absorbed in the HCAL. Figure 3.18 summarizes how each of type of particle traverse through CMS.

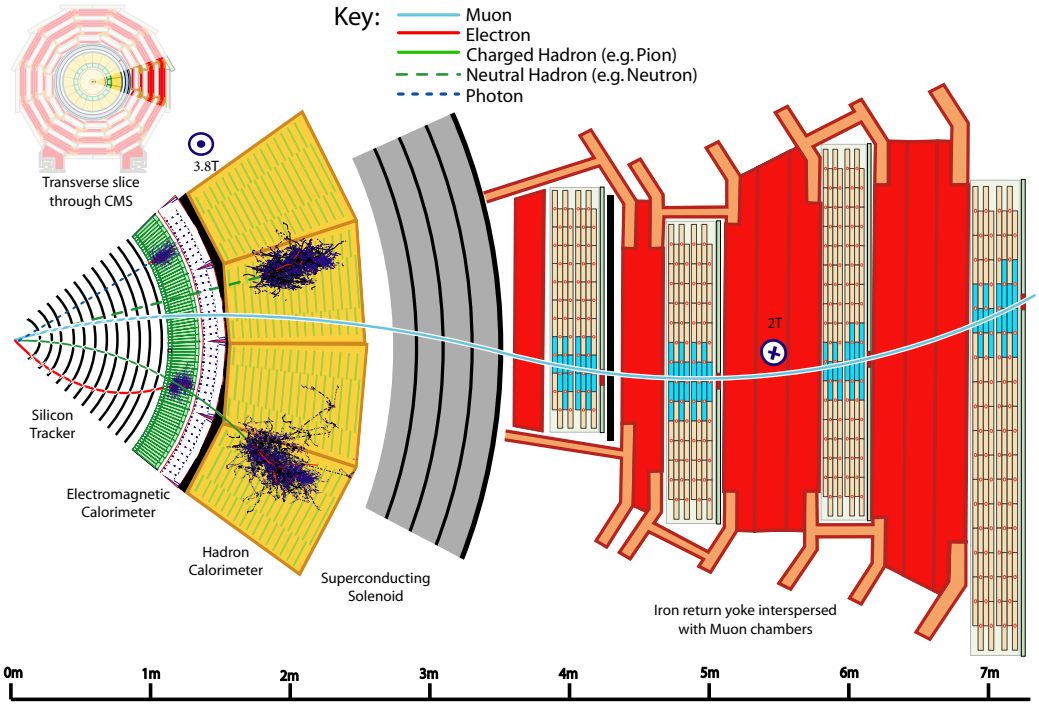


Figure 3.18: Diagram of how different particles traverse different detector components in the CMS detector [37].

3.3.1 Particle-flow algorithm

From Fig. 3.18, it is clear that the best reconstruction performance will make use of information from all subdetectors. For example, tracking information linked to ECAL

clusters could help distinguishing between electrons and photons. The Particle Flow (PF) algorithm [44] is developed in CMS to use full detector information consistently for particle identification and reconstruction to give a global description of the event. For each event, the PF algorithm returns a list of PF candidates for muon, electron, photon, charged hadrons and neutral hadrons, together with the reconstructed 4 momentum vectors.

Algorithm outline

The inputs to the PF reconstruction are called *PF elements*, which are the tracks in the inner tracker, electron tracks, muon tracks, ECAL and HCAL clusters. Since a particle can give rise to several PF elements, the PF elements are linked together by the link algorithm to form *PF blocks*, which are elements associated either by a direct link or by an indirect link through common elements. In each PF block, a sequence of identification and reconstruction is performed, attempting to select muon, electron, photons and hadrons candidates, until all PF blocks have been processed. Finally, the reconstructed event is re-visited to handle cases arising from rare misidentification and misreconstructions.

Tracks

Charged-particle tracks are primarily reconstructed from hits in the silicon inner detector, supplemented with muon systems for muon tracks. A combinatorical track finder based on Kalman Filtering (KF) [45] is used to reconstruct these tracks in three

stages: initial seed generations from finding a few hits compatible with a charged-particle trajectory; trajectory building and final fitting, which determines the origin, transverse momentum and direction of the charged particle. The track finder are applied iteratively, masking hits that already assigned to tracks in previous steps, such that the hits needed to fit are reduced in the next step. The first iteration is set to have the tightest seeding requirements and then progressively loosen in the iteration steps, in which the track quality requirements, such as track χ^2 and number of hits in track, are tightened to keep the misreconstruction rate low. Figure 3.19 illustrates the track efficiency and misreconstruction rate at different iteration stages. KF and iterative tracking are the two key techniques to overcome the combinatorical challenge of hit-to-track reconstruction under high pile-up environment. Electron

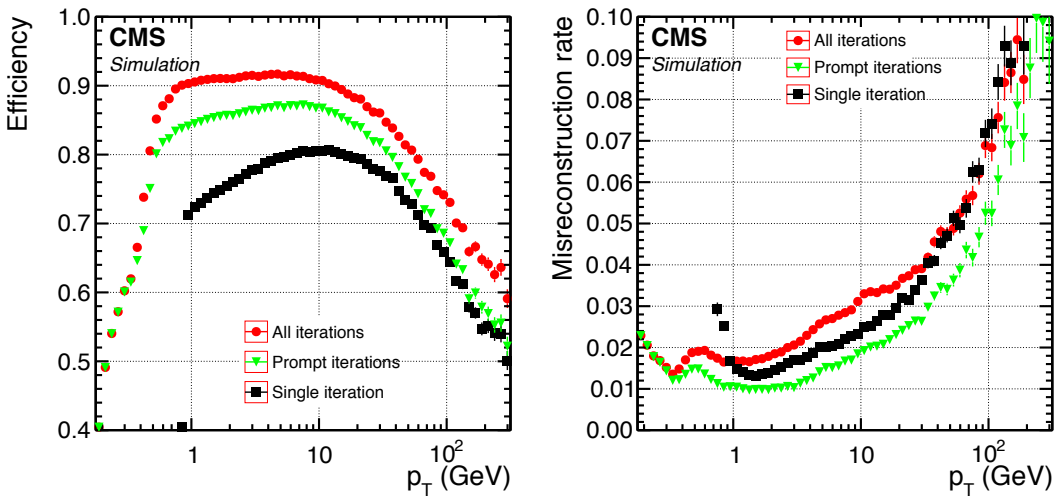


Figure 3.19: Efficiency (left) and misreconstruction rate (right) of global combinatorical track finder at different iterations as a function of track p_{T}[44].

tracks differ slightly from charged hadron tracks due to a sizable probability of losing energy through bremsstrahlung radiation. To find the tracks for radiated electrons,

selected tracks are fitted again with a Gaussian-sum filter (GSF) [46], which better matches the radiated electron track patterns. Muon tracks can be formed with muon systems only, called standalone muon, or with inner tracker, called tracker muon. Each standalone muon track is extrapolated to a inner tracker track and form a global muon track if the combined hits improves momentum resolution with respect to the tracker-only fit.

Calorimeter clusters

Calorimeter clusters are constructed separately in the barrels and endcaps of ECAL and HCAL. Clusters are first formed with *cluster seeds*, which are cells with energies above a certain threshold and have more energies than neighboring cells. Adjacent cluster seeds are summed to form topological clusters, which are then clustered into calorimeter clusters elements.

Link algorithm

The link algorithm is used to establish links between any two PF elements, can be track-to-track, track-to-cluster or cluster-to-cluster. For track-cluster links, the algorithm extrapolates the track to calorimeter cluster and link the two elements if extrapolation lies within boundaries of calorimeter cluster. For cluster-cluster links, a link is formed from the more granular calorimeter cluster, usually ECAL, to the less granular cluster, usually HCAL, if it is within the cluster envelope. A global muon is formed by the track-to-track link between standalone muon and tracker muon track. With the PF elements linked into PF blocks, PF candidate reconstruction will be

performed in each block.

3.3.2 Muons

Muons are first reconstructed from the PF blocks, by a set of selection based on the global and tracker muon properties. Isolated muons are required to have small calorimeter energy deposit and high momentum tracks around them. The sum of p_{T} of tracks and the sum of the E_T of the calorimeter energy, within a radius of $\Delta R = 0.3$ in the muon direction in $\eta - \phi$ plane, is required to be less than 10% of the muon p_{T}. For non-isolated muons, tighter selection on fit quality and number of hits in muon stations are required [47]. The PF elements that are used in the identified muon are masked from the PF blocks in the further processing.

3.3.3 Electrons and Photons

Electron reconstruction uses information from the inner tracker and calorimeters, mainly ECAL. As mentioned in Sec. 3.3.1, electron tracks are reconstructed as GSF tracks in the inner tracker, therefore electron candidates are seeded from a GSF track linked to an ECAL cluster, whereas photon candidates are seeded with an ECAL supercluster with E_T larger than 10 GeV with no link to a GSF track. Other ECAL clusters in the PF block are associated to the electron candidates if they are linked to the candidate's GSF track or ECAL cluster, to provide good energy containment. The electron candidate are then required to pass criteria to distinguish it from a charged hadrons, such as the width of the supercluster in η direction and HCAL/ECAL energy

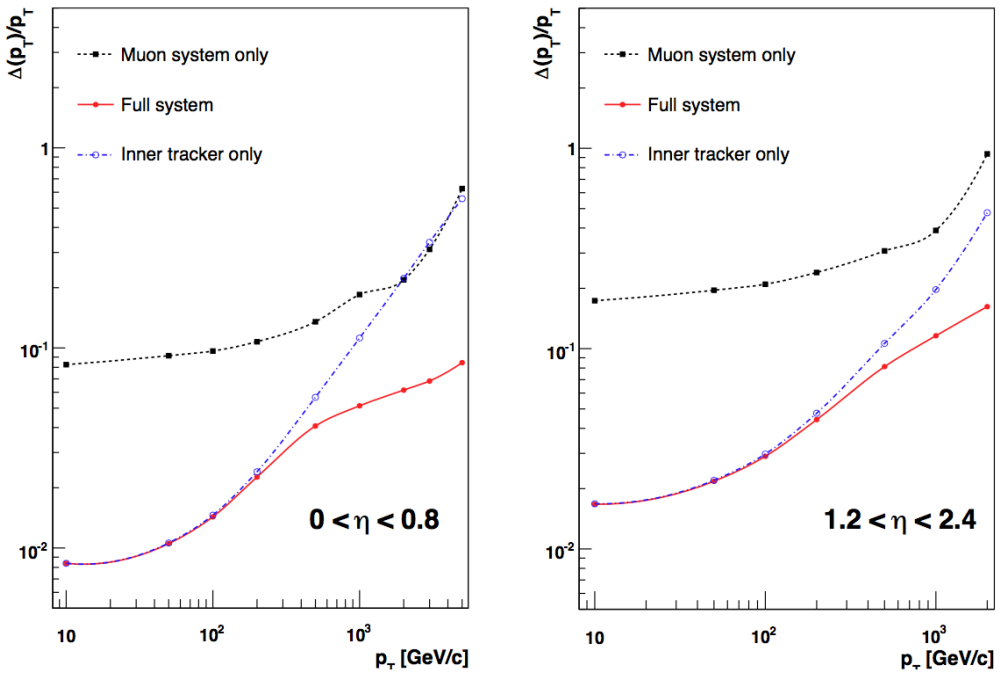


Figure 3.20: Muon transverse momentum resolution as a function of muon transverse momentum [37].

cluster ratio (H/E). Up to 14 such variables are combined in a boosted decision tree (BDT) score to select the final PF electron [48]. Similarly, photon candidates are required to pass requirements on ECAL cell energy distribution and H/E that are compatible with those expected from a photon shower. All tracks and clusters in the PF block used to reconstruct electrons and photons are masked against further processing.

3.3.4 Hadrons

After selecting the muon, electrons and isolated photons from the PF blocks, the remaining particles are charged hadrons ($\pi^{\pm}, \mathrm{K}^{\pm}$ or protons) or neutral hadrons ($\mathrm{K}^{0}_{\mathrm{L}}$ or neutrons). In the remaining PF blocks, the HCAL and ECAL clusters linked to tracks are identified as charged hadrons. For neutral hadrons, within the tracker acceptance of $|\eta| < 2.5$, they are identified as HCAL clusters that are not linked to any track. Beyond the tracker acceptance, charged and neutral hadrons cannot be distinguished unambiguously, therefore neutral hadrons is assumed to be ECAL clusters linked with a HCAL cluster, since a significant fraction of neutral hadrons energy are deposited in ECAL.

3.3.5 Jets

After identifying the five types of PF candidates, composite objects can be constructed. In particular, quarks and gluons manifest themselves as sprays of collimated particles resulting from hadronization and fragmentation. Therefore, it requires an

algorithm to reconstruct the underlying partons that give rise to these sprays of particles called jets. Jets can be clustered using different inputs with various algorithms. In CMS, the jets that are clustered with PF candidates are called PF jets.

The anti-k_{T} algorithm [49] is a very common choice of jet algorithm, favored by both experimentalists and theorists. To experimentalists, the jets resulting from the anti-k_{T} algorithm has a regular circular shape in the $\eta - \phi$ plane, leading to a better control and calibration procedure with jets; whereas to theorists, the algorithm is infar-red and collinear (IRC) safe, meaning the quantity of reconstructed jets is unchanged by an additional soft/collinear gluon emission, so that predictions can of jet quantities can be made.

The anti-k_{T} algorithm is an sequential recombination jet algorithm that successively cluster the closest pairs of particles i and j defined by a custom distance metric d_{ij}. The distance d_{ij} is defined as:

$$d_{ij} = \min(k_{ti}^{-2}, k_{tj}^{-2}) \frac{\Delta_{ij}^2}{R^2}, \tag{3.6}$$

where $\Delta_{ij}^2 = (\eta_i - \eta_j)^2 + (\phi_i - \phi_j)^2$ is the eculidian distance between particle i and j in the $\eta - \phi$ plane, k_{ti} is the transverse momentum of particle i and R is the distance parameter. The momenta of the closest pairs are added together, then d_{ij} is recalculated for all particles. This procedure is repeated until no particles are left. Essentially, the algorithm favours clustering more energetic particles first and sequentially recover the softer particles around it. This can be thought of as the reverse process of how jets arise from quarks/gluon in the semi-classical picture of

iterative $1 \rightarrow 2$ splitting. The resulting jets usually has a circular shape in the $\eta - \phi$ plane with a radius equal to the distance parameter R. Typical choices of R in CMS are 0.4 for regular jets (AK4 jets) and 0.8 for wide jets (AK8 jets) to look for substructures from heavy resonance within the jet, which will be discussed later.

Jet corrections

Jets are complicated objects to calibrate, partly due to the procedure involved in the reconstruction and partly due to non-linear response in CMS calorimeters. CMS uses a multiplicative factorized approach to correct for jet energies. The raw jet transverse momentum (p_{T}) from summing constituents p_{T}, $p_{\mathrm{T}}^{\mathrm{raw}}$, is corrected to the final jet p_{T}, $p_{\mathrm{T}}^{\mathrm{corr}}$ using the following equation [50]:

$$p_{\mathrm{T}}^{\mathrm{corr}} = p_{\mathrm{T}}^{\mathrm{raw}} \times C_{\mathrm{offset}}(p_{\mathrm{T}}^{\mathrm{raw}}) \times C_{\mathrm{MC}}(p'_{\mathrm{T}}, \eta) \times C_{\mathrm{rel}}(\eta) \times C_{\mathrm{abs}}(p''_{\mathrm{T}}), \tag{3.7}$$

where C_{offset} is the offset correction due to pile-up based on the average energy deposited in the jet area, C_{MC} is correction for reconstruction effects in Monte Carlo (MC) simulation to match the reconstructed jet energy to the generator level, C_{rel},C_{abs} are small residual differences between data and MC, derived from data using transverse momentum balance in γ+jets and Z+jets events [51], p'_{T} is the transverse momentum after applying the offset corrections and p''_{T} is the transverse momentum after all previous corrections.

In addition to jet energy scale corrections, the jet resolutions in data is found to be worse than jets in MC. Data-to-MC scale factors for jet resolutions are also derived

from the same γ+jets and Z+jets events.

3.3.6 Wide jets with substructures

As the energy of the LHC reaches the 10 TeV regime, heavy resonances in the SM, with masses of the order 100 GeV, are produced more and more often with large momentum. The large Lorentz boost of these heavy resonances, such as the top quark and the W/Z boson, will result in more colliminated hadronic decay products, which can be clustered within the radius of a single jet. Such jets are different than jets arising from QCD in that two or three energetic centers, or "prong", corresponding to the hadrons decayed from the resonance, can be found within the jet. These differences in substructures of jets can be exploited to differentiate from QCD jets.

In CMS, jet with substructures are clustered with a larger distance parameter of $R = 0.8$. In the following sections, we will discuss the quantities related to these wide jets in order to study the substructures of jets.

Jet mass

Ideally, the mass of a jet, computed as $(\sum_i E^2 - \sum_i p_{\mathrm{T}}^2)^{-1/2}$ from jet constituents, should reflect the mass of the quark/gluon or heavy resonances that the jet originates from. However, QCD can jets can result in a non-zero peak in the plain mass distribution simply by having more gluon radiations inside the jet [52].

In order to recover the bare jet mass, we apply a jet grooming algorithm that removes soft and wide-angle contributions from the jet, in order to mitigate the effects

of contamination from initial state radiation (ISR), underlying event (UE) and pileup.

The soft drop algorithm [53] has become widely used in the community due to its robustness and well-found theoretical foundation. This algorithm controls the soft wide-angle radiation by a soft radiation fraction threshold z_{cut} and an angular exponent parameter β, where $\beta = 0$ corresponds to the (modified) mass-drop procedure (mMDT) detailed in [54]. Given a jet of radius R_0 with only two constituents, the soft drop procedure removes the softer constituent unless:

$$\frac{\min(p_{\text{T}}^1, p_{\text{T}}^2)}{p_{\text{T}}^1 + p_{\text{T}}^2} > z_{\text{cut}} \left(\frac{\Delta R_{12}}{R_0} \right)^{\beta} \tag{3.8}$$

The default parameters used by CMS are $\beta = 0$ and $z_{\text{cut}} = 0.1$, meaning the grooming threshold is independent of subjet separation and soft radiation fraction of less than 0.1 will be removed. The soft drop algorithm has the benefit of performing jet grooming in a theoretically safer way [53, 55] and its behavior is constant across different clustering distance parameters R and p_{T}, which is not true for the pruning technique [56, 57]. Although the soft drop algorithm is primarily aimed at separating boosted W-jets from light quark/gluon-jets, it can fully reject contributions from underlying event and pileup when combined with the PUPPI algorithm.

To ensure the *simulated* soft drop mass reproduces the mass of boosted resonances such as W, Z and H in *data*, we apply three stages of factorized corrections as follows:

- **gen corrections** : a p_{T}-dependent correction to account for a small shift between generated m_{SD} and physical mass

- **reco corrections** : a p_{T}-dependent correction to the reconstructed jet m_{SD}, applied separately for jets in the barrel and endcaps regions

- **Data/MC corrections**: derived from a dedicated semi-leptonic $\mathrm{t\bar{t}}$ decay sample in data, probing the m_{SD} of the merged W opposite to the energetic muon, to be discussed in details in Sec. 4.6.1

The gen and reco correction are derived from a study with boosted W boson in simulation [58]. Generated and reconstructed m_{SD} are the soft-drop mass obtained from using generated particles and reconstructed particles as input to the soft-drop algorithm. The shift in generated softdrop mass at lower p_{T} is of the order of 2-3% while the difference between reconstructed and generated softdrop mass is a 5-10% effect. The mass shift introduced at generator level is corrected by a fit to $M_{\mathrm{PDG}}/M_{\mathrm{GEN}}$ as a function of jet p_{T}, where $M_{\mathrm{PDG}} = 80.4$ GeV and M_{GEN} is the fitted mean of the generator level mass. To correct for the residual shift between generator and reconstruction level, a fit to $M_{\mathrm{GEN}}/M_{\mathrm{RECO}}$, where M_{RECO} is the reconstructed jet mass. The distribution and corresponding fits for the two weights is shown in Fig. 3.21 for the "gen correction" and "reco correction" [58], showing the size of this corrections and the associated uncertainty.

The raw reconstructed m_{SD} is corrected by the product of the two fitted ratios as follows:

$$m_{\mathrm{SD}} = m_{\mathrm{SD,raw}} \times w_{\mathrm{GEN}} \times w_{\mathrm{RECO}} \tag{3.9}$$

where $w_{\mathrm{GEN}} = M_{\mathrm{PDG}}/M_{\mathrm{GEN}}$ and $w_{\mathrm{RECO}} = M_{\mathrm{GEN}}/M_{\mathrm{RECO}}$.

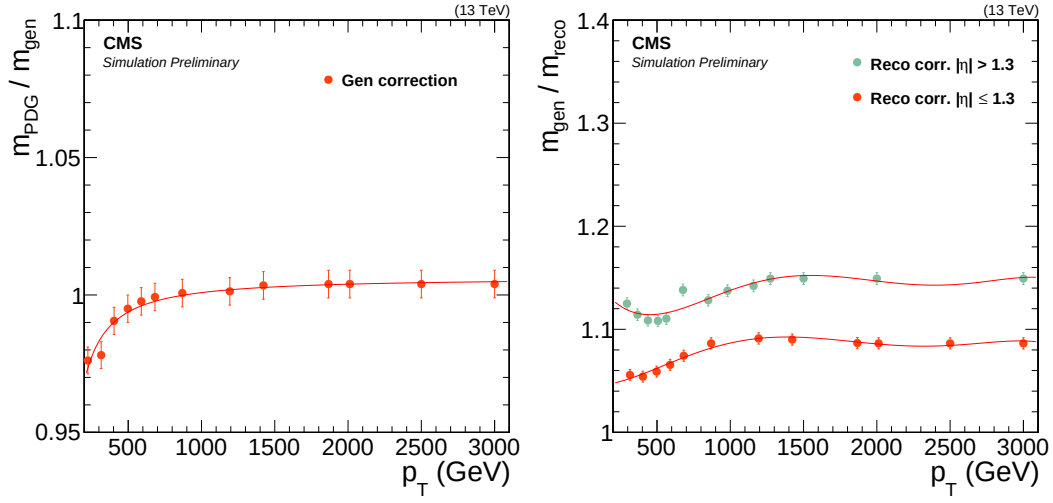


Figure 3.21: Fit to (left) M_{PDG}/M_{GEN} as a function of jet p_T, where $M_{PDG} = 80.4$ GeV and M_{GEN} is the fitted mean of the generator level mass and (right) M_{GEN}/M_{RECO}, where M_{RECO} is the reconstructed softdrop mass, as a function of jet p_T in two η bins. [58]

3.3.7 Jet ρ

Before the discussion of the correlation of jet substructure variables with the jet mass and p_T in the next section, we introduce the scaling variable for QCD jets ρ, defined as $\rho = \log(m_{SD}^2/p_T^2)$, where log is the natural logarithm function. The ρ variable is motivated by the fact that QCD jet mass scales with the jet p_T and is found to have robust scaling properties for substructure variables in certain kinematic regime [59]. Note that the p_T is computed with the jet constituents of the original jet, without grooming applied.

The distribution of ρ and m_{SD} for QCD multijets simulation is shown in Fig 3.22 for different p_T bins.

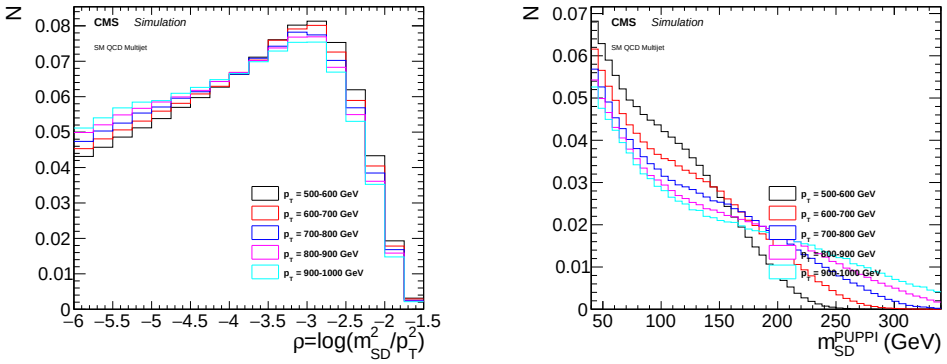


Figure 3.22: Distribution of ρ (left) and m_{SD} (right), shown for QCD MC for different p_{T} bins.

3.3.8 Jet substructure

Jet substructure observables are used to distinguish boosted H bosons that decay hadronically from the hadronization products of single light quarks or gluons. Examples of observables widely used in jet substructure include $N-$ subjettiness ratios [60] and energy correlation functions (ECFs) [61]. These are usually constructed using power counting techniques, from a basis of infrared and collinear (IRC) safe observables that probe $N-$prong substructure within a jet. A IRC safe observable is insensitive to the emission of soft or collinear gluons. In this work, ECF-based observable is chosen over the $N-$ subjettiness ratios because ECF-based observables is more theoretically motivated since they do of not rely on the ambiguous definition of sub-jet axes and often provides better robustness across kinematic regimes.

For a jet with n_J constituents the 2− and 3− point ECFs are defined as follows:

$$e_2^\beta = \sum_{1\leq i<j\leq n_J} z_i z_j \Delta R_{ij}^\beta \tag{3.10}$$

$$e_3^\beta = \sum_{1\leq i<j<k\leq n_J} z_i z_j z_k \Delta R_{ij}^\beta \Delta R_{ik}^\beta \Delta R_{jk}^\beta, \tag{3.11}$$

where z and ΔR represent the energy fraction and angular variable, respectively. From these, a number of powerful 2-prong discriminant have been formed by taking ratios, namely $D_2 = \frac{e_3^\beta}{(e_2^\beta)^3}$ and $C_2 = \frac{e_3^\beta}{(e_2^\beta)^2}$ [62], which are able to identify the extent to which a jet is 1- or 2- prong like. For a 2-prong structure, signal jets have a stronger 2-point correlation than a 3-point correlation.

In this context, a new basis of IRC safe observables was recently proposed in [63]. These observables, named as generalized energy correlation functions (${}_oe_n^\beta$), are defined as follows:

$${}_oe_n^\beta = \sum_{1\leq i_1<i_2<\ldots<i_n\leq n_J} \prod_{1\leq k\leq n} z_{i_k} \times \min\{ \prod_{s<t\in \text{pairs}\{i_1,i_2,\ldots,i_n\}}^{o} \Delta R_{st}^\beta \}, \tag{3.12}$$

where:

- n, denotes the number of particles to be correlated i.e. the order of the correlation function. For an $n-$ pronged jet, $e_n >> e_m$, for $m \geq n$.
- o, denotes the order of the angular factor, and
- β, is related to the angular weighting.

The ${}_oe_n^\beta$ correlate o pairwise angles among n particles, whereas the original correlators were restricted to $o = n = 2$. Therefore, the generalization in Eq. 3.12 introduces more flexibility in angular scaling. Note that similarly to the original ECFs, they do not rely on external axes.

Using these generalized correlators, it is possible to apply power counting to identify new jet substructure observables. For boosted vector boson tagging, ratios between the 2-point and 3-point (with its three variants) energy correlation functions can provide a well motivated tagger:

$$ {}_1e_2^\beta = e_2^\beta = \sum_{1\le i<j\le n_J} z_i z_j \Delta R_{ij}^\beta \tag{3.13} $$

$$ {}_1e_3^\beta = \sum_{1\le i<j<k\le n_J} z_i z_j z_k \min\{\Delta R_{ij}^\beta, \Delta R_{ik}^\beta, \Delta R_{jk}^\beta\} \tag{3.14} $$

$$ {}_2e_3^\beta = \sum_{1\le i<j<k\le n_J} z_i z_j z_k \min\{\Delta R_{ij}^\beta \Delta R_{ik}^\beta, \Delta R_{ij}^\beta \Delta R_{jk}^\beta, \Delta R_{ik}^\beta \Delta R_{jk}^\beta\} \tag{3.15} $$

$$ {}_3e_3^\beta = e_3^\beta = \sum_{1\le i<j<k\le n_J} z_i z_j z_k \Delta R_{ij}^\beta \Delta R_{ik}^\beta \Delta R_{jk}^\beta \ , \tag{3.16} $$

In particular, three dimensionless ratios, corresponding to the three variants of the 3-point correlator, are proposed in [63] for this purpose:

$$ M_2^\beta = \frac{{}_1e_3^\beta}{{}_1e_2^\beta}, \quad N_2^\beta = \frac{{}_2e_3^\beta}{({}_1e_2^\beta)^2}, \quad D_2^{\alpha,\beta} = \frac{{}_3e_3^\alpha}{({}_1e_2^\beta)^{3\alpha/\beta}}, \tag{3.17} $$

Their performance were studied in simulation [64], before and after grooming. Grooming in general improves the signal-background separations in the substructure variables, which confirms the procedure is effective of removing features in QCD background while preserving genuine hard-radiation process. It is also shown that N_2 with $\beta = 1$ is the variable which provides the most discrimination power and shows better discrimination power as the analogous $N-$ subjettiness ratio τ_{21}.

3.3.9 Missing transverse energy

Since the initial momentum in the transverse plane is zero, summing all the transverse momenta of the particles in the final states should be zero. Since CMS is capable of detecting most of the final state particles except neutrinos, summing all of *detected* particles could be used to get indirect information about neutrinos. Using the PF candidates as input, the missing transverse momentum vector $\vec{p}_T^{\text{miss}}$, is defined as the negative vector sum of p_{T} of all PF candidates,

$$\vec{p}_T^{\text{miss}} = -\sum_{i}^{\text{PFcand}} \vec{p}_{T,i}. \tag{3.18}$$

Missing transverse energy (MET) is the magnitude of the missing transverse momentum vector. Since jets could receive substantial energy correction, the type-I corrected MET is then,

$$|\vec{p}_T^{\text{miss}}| = -\sum_{i}^{\text{PFcand}} \vec{p}_{T,i} - \sum_{j}^{\text{PFjets}} (\vec{p}_{T,j}^{\,corr} - \vec{p}_{T,j}), \tag{3.19}$$

where $p_{T,j}^{corr}$ are corrected p_{T} of the PF jets.

3.3.10 Identification of b-quark jets

The jets that originates from a b-quark is called b-jets. There are several properties of b hadrons that allows b-jets to be differentiated from the jets from light quarks or gluon. Most notably, the lifetime of b-hadrons are of the order of 1.5 ps, which allows them to travel a few mm before decaying – a distance well within the resolution of CMS inner tracker, such that secondary vertex can be reconstructed. The relatively large masses of b-hadrons also result in different kinematic features in the decay products, such as larger transverse momenta with respect the jet axis. Low energy (soft) muons are often resulted in the decay chain of b-hadrons, which can also be measured by the muon system in CMS.

Algorithms, often called taggers, can then be built on these variables to exploit the small differences between light jets and b-jets to differentiate between the two. It is important to note that, the presence of muon provides an independent way of verifying the taggers performance in data, which is an handle to define an orthogonal sample to the tracking based variables.

Many algorithms have been developed in CMS for b-tagging, reaching as high as 70% b-jet efficiency at 1% light-jet misidentification rate. Figure 3.24 shows the misidentification rate as a function of b-jet efficiency for different CMS b-tagging algorithms. While CMS has a long history of b-tagging program with AK4 jets, the b-tagging program is further enriched in the context of jet substructure studies, in which two b-quarks can be present in a single wide jet. We discuss one of such tagger in the following section.

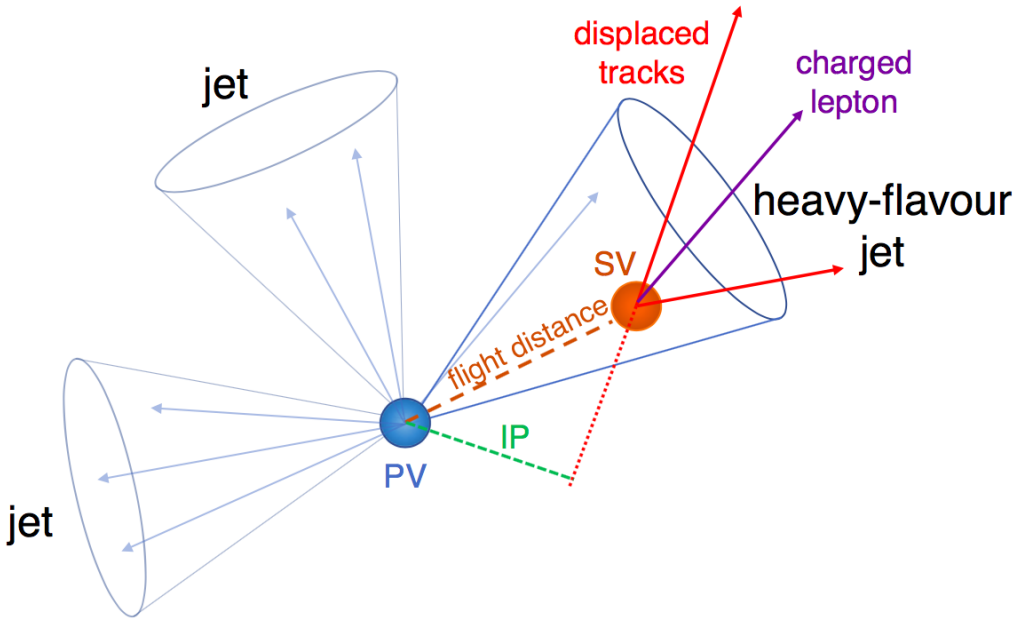


Figure 3.23: Illustration of properties of b-hadrons that can be used to differentiate from light quark/gluon jets [65].

Deep double-b tagger

In order to identify the candidate H jet most likely to contain two b quarks, we use the deep double-b tagger discriminant [66, 67], based on a deep neural network. It is a progression of the original double-b deep tagger based on a boosted decision tree (BDT) [65].

A brief description of the inputs to the deep double-b tagger algorithm is as follows: tracks and secondary vertices associated to a jet are used, as for standard b-tagging. First, the 27 input features used for the BDT of the double-b tagger are still used. These are based on tracking and vertexing information, targeting to resolve the two B-hadron decay chains we expect for a $X \to b\bar{b}$ signal. For instance, N-subjettiness(τ)

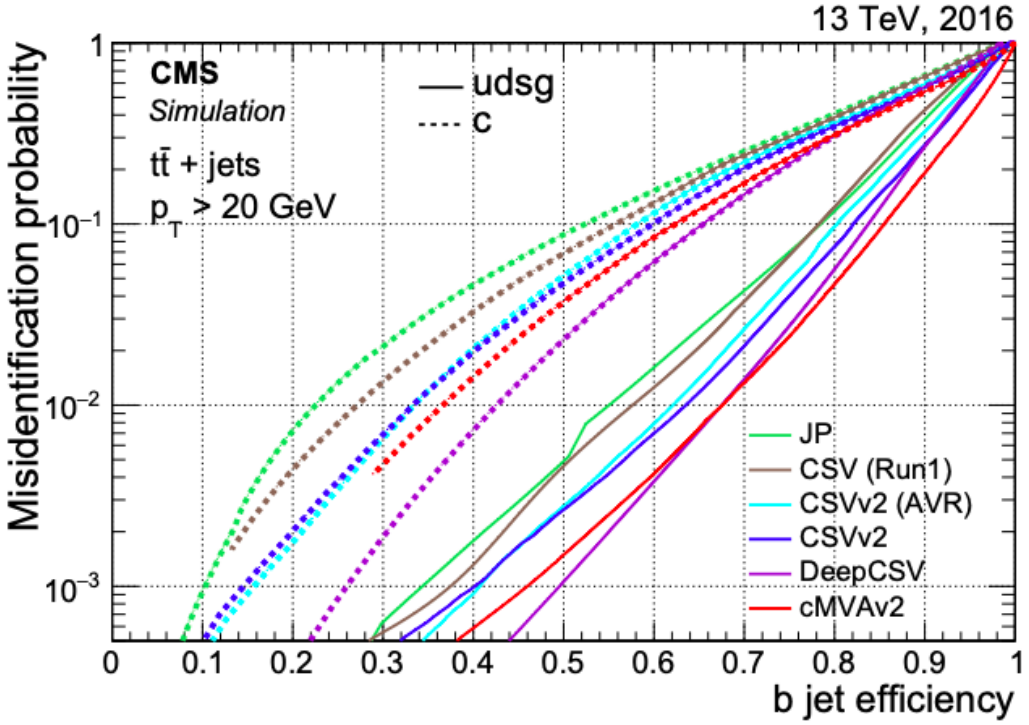


Figure 3.24: Receiver operating characteristic (ROC) curve comparing the performance of different b-tagging algorithms in CMS [65].

axes are calculated from tracks of a jet, and the two τ-axis directions are used to approximate the directions of the two b-quarks in the signal. Secondary vertex (SV) observables are assigned to each axis independently of the subjet components. Then tagging observables are computed using displaced tracks, SVs, and the two-SV system information. Additionally, the following track variables are taken for up to 60 tracks with $p_{\mathrm{T}} > 0.95\,\mathrm{GeV}$ sorted by impact parameter significance

- 2D and 3D impact parameter and impact parameter significance of charged tracks from B-tagged AK4 jet.
- Track transverse momentum, along the jet axis, normalized to its energy.
- Track parallel momentum, along the jet axis, normalized to its energy.
- Track transverse momentum, relative to the jet axis, normalized to its energy.
- Minimum track approach distance to jet axis.

Similarly, the following SV variables are taken for up to 5 SVs sorted by impact parameter significance:

- 3D flight distance.
- 3D flight distance significance.

The algorithm is based on a deep neural network. The track and SV low-level inputs are processed separately by Batch Normalization layers, then two 1D convolution layers, each with 32 filters, a kernel size of 1, and a stride of 1, followed by a Spatial

Dropout layer with 10% dropout, which randomly drops 10% of the filters during training. The output feature sequence (32 features per track or SV) are then fed into a Gated Recurrent Unit layer with 50 hidden units, again followed by a dropout layer with 10% dropout during training, yielding 50 total features for each category of inputs (track or SV). These 50 track and 50 SV features are then concatenated with the pre-computed 27 high-level input features and processed together with a fully connected layer with 100 neurons and 10% dropout. Finally, the output layer of the DNN has either two neurons, with softmax activation, which compute the probability for each class (QCD or $\mathrm{H}(\mathrm{b}\overline{\mathrm{b}})$.

In Fig. 4.19, the performance of the deep double-b tagger is quantified in a ROC curve for QCD and $\mathrm{H}(\mathrm{b}\overline{\mathrm{b}})$ jets with $450 < p_{\mathrm{T}} < 1200\,\mathrm{GeV}$ and $40 < m_{\mathrm{SD}} < 200\,\mathrm{GeV}$.

Mass decorrelation

Empirically, deep networks are able to learn the invariant mass of the jet when provided sufficient low-level kinematic inputs. As such, selecting on the output discriminator value "sculpts" the jet mass distribution of the background to artificially create a "peak" at the signal resonance mass. This is demonstrated in Fig. 4.20 (left). To combat this, the training is performed with a dedicated penalty term in the loss function based on the Kullback-Liebler (KL) divergence between the "tagged" and "untagged" distributions.

The KL divergence is a measure of how one probability distribution is different from the referenced one. Given two distributions represented as binned normalized

histogram vectors, $\vec{w}$ and $\vec{z}$, the KL divergence is given by,

$$D_{KL}(\vec{w},\vec{z}) = \sum_j \log\left(\frac{w_i}{z_i}\right), \tag{3.20}$$

where the two distribution are normalized as $\sum_j w_j = \sum_j z_j = 1$. During the training of the deep double-b tagger, the KL divergence terms are evaluated with the binned normalized histograms of the simulated signal and background process. To form the "tagged" and "untagged" distribution, we compute

$$\vec{h}_{\text{b,untagged}} = \sum_k y_0^k p_0^k \vec{h}^k \tag{3.21}$$

$$\vec{h}_{\text{b,tagged}} = \sum_k y_0^k p_1^k \vec{h}^k \tag{3.22}$$

$$\vec{h}_{\text{s,untagged}} = \sum_k y_1^k p_0^k \vec{h}^k \tag{3.23}$$

$$\vec{h}_{\text{s,tagged}} = \sum_k y_1^k p_1^k \vec{h}^k, \tag{3.24}$$

where the sum is over a training batch of sample, $\vec{h}^k$ is the one-hot encoded mass vector, $y^k_{0(1)}$ is the background (signal) truth label, and $p^k_{0(1)}$ is the background (signal) discriminator probability output. The KL divergence between Eq. 3.22 and Eq. 3.21 is a measure of the background mass sculpting, while the KL divergence between Eq. 3.24 and Eq. 3.23 is the measure of the signal mass sculpting. Therefore, the final

loss function is given by,

$$L = -\sum_{i=0}^{1} y_i^k \log\left(p_i^k\right) + \lambda_s D_{KL}(\vec{h}_{\text{s,tagged}}, \vec{h}_{\text{s,untagged}}) + \lambda_b D_{KL}(\vec{h}_{\text{b,tagged}}, \vec{h}_{\text{b,untagged}}), \tag{3.25}$$

where the first term is the categorical cross-entropy, $\lambda_{s(b)}$ are the regularization parameters for the background and signal mass sculpting. Minimizing Eq. 3.25 with respect to the training sample would mitigate the difference of the mass distribution between the tagged and untagged sample.

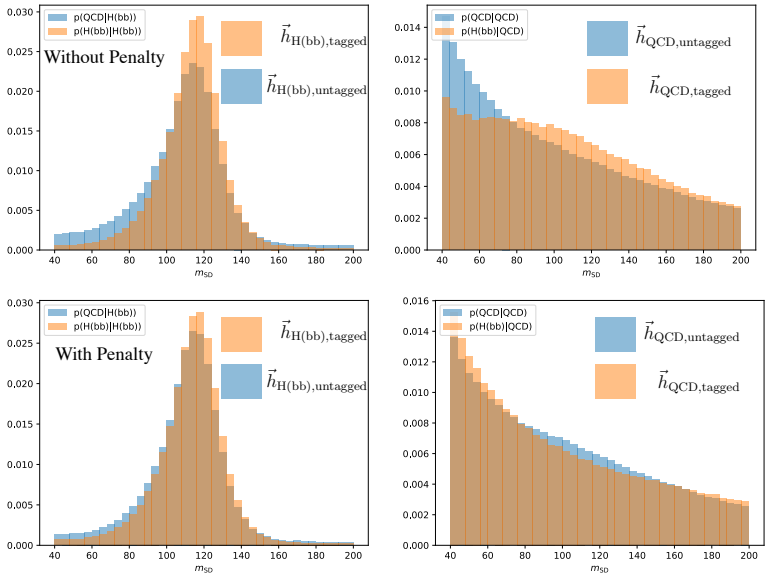


Figure 3.25: The weighted, normalized mass distributions for the "tagged" and "untagged" background and signal before adding the KL divergence penalty terms to the training (top) and after adding the KL divergence penalty terms to the training (bottom).

Chapter 4

Search for a boosted Higgs boson decaying in to $b\bar{b}$

4.1 Introduction

CMS and ATLAS collaborations have performed the first searches for highly boosted Higgs boson decaying to a bottom quark-antiquark pair ($b\bar{b}$) that is inclusive in the production modes of the Higgs boson [1, 68]. Both searches are made possible with the recent advancement in jet substructure techniques and novel b-tagging algorithms, which are the two key handles used to discriminate the over-whelming background events. With around 1/3 of the full LHC Run 2 data, the two searches observed significances of 1.5 and 1.6 standard deviations, respectively, with respect to background-only hypothesis, demonstrating the potential with the search method.

In this section, I will describe an improved CMS search for highly boosted Higgs

boson decaying into $b\bar{b}$ using full LHC Run 2 dataset.

4.1.1 Analysis strategy

One of the main challenges of performing an inclusive search for the Higgs boson decaying to bottom quarks is the overwhelming multi-jet backgrounds from QCD events.

With the jet substructure and b-tagging selections, the signal-to-background ratio is brought to around 0.1% level without requiring any jet mass window selection. It is therefore possible to perform a parametric fit on a smoothly falling jet mass spectrum to look for an excess of Higgs signal above background around the Higgs mass, provided that the selections do not "sculpt" the shape of QCD jet mass spectrum. In practice, the jet substructure and b-tagger selections are often correlated with the jet mass. One of the focus of the analysis would be to decorrelate as much as we can such that we manage to estimate the QCD jet mass shape after the selections.

Motivated by the discussion in Sec. 2.5, we search for a $H \to b\bar{b}$ decay, merged into a single anti-k_T jet with $R = 0.8$ and $p_T > 450\,\text{GeV}$. After selecting the Higgs candidate jet, the QCD background prediction is performed in three main stages:

1. Select events with the jet substructure variable $N_2^{1,\text{DDT}}$, which is decorrelated with respect to p_T and $\rho = \log\left(m_{SD}^2/p_T^2\right)$ in QCD simulations.

2. Use a tight working point cut on the deep double-b tag discriminator value (1% QCD mistagging rate) to define the signal regions and the QCD control regions, in which we analyze the softdrop mass and the p_T of the Higgs candidate jet.

3. Predict the m_{SD} SM background distribution by fitting for a QCD transfer factor pass-fail ratio ($R_{\mathrm{p/f}}$) with a polynomial dependence on p_{T} and $\rho = \log\left(m_{\mathrm{SD}}^2/p_{\mathrm{T}}^2\right)$.

The search then proceeds by looking for a bump on the predicted QCD mass distribution.

In addition to the QCD background, the other SM processes contribute up to around 3% of the total background events. The most notable ones are from $\mathrm{W}(\mathrm{q}\bar{\mathrm{q}})$, $\mathrm{Z}(\mathrm{b}\bar{\mathrm{b}})$ and $\mathrm{t}\bar{\mathrm{t}}$ production. For $\mathrm{W}(\mathrm{q}\bar{\mathrm{q}})$ and $\mathrm{Z}(\mathrm{b}\bar{\mathrm{b}})$ production, which we expect will give us resonant backgrounds on top of the smooth falling QCD background, we model their mass shapes via simulations, corrected by dedicated data-driven scale factors and then simultaneously constrain the uncertainties in the parametric fit to extract QCD background. For $\mathrm{t}\bar{\mathrm{t}}$ production, a separate dedicated control region is defined with enriched semileptonic $\mathrm{t}\bar{\mathrm{t}}$ events, which is used to derive data-to-simulation scale factors.

4.2 Data and simulated samples

4.2.1 Data

The data used for this analysis was collected by the CMS detector over the 3 years of LHC Run II operations from 2016 to 2018, when protons are collided at $\sqrt{s} = 13\,\mathrm{TeV}$ with bunch spacing of 25 ns. These data samples are listed in Table 4.1. All the data are certified by CMS collaboration's Data Quality Monitoring and Data Certification

teams to ensure the quality of the data are suitable for Physics analysis. Validations are performed during data-taking for the nominal operation conditions of the detectors, as well as after data collection where exceptional conditions/incidents are scrutinized by relevant experts. The JetHT primary dataset is used as the main analysis sample. The SingleMuon dataset is used for deriving dedicated control samples and measurement for trigger efficiencies, which will be detailed later in Section 4.3

4.2.2 Background simulations

Standard model backgrounds are simulated with various Monte Carlo(MC) event generators, with the CMS detector response modeled by GEANT4 [69]. The `MadGraph5 amc@nlo` 2.4.2 [70] generator is used for the diboson, W+jets, Z+jets, and QCD multijet samples, at LO accuracy with matching [71] between jets from the matrix element calculation and the parton shower description, while POWHEG 2.0 [72, 73, 74] at NLO precision is used to model the $t\bar{t}$ and single top processes. Since it is important to have accurate description of the resonant backgrounds of W+jets and Z+jets, higher-order QCD and electroweak(EW) corrections are applied to both samples. This will be detailed in the next section.

For parton showering, fragmentation and hadronization, all samples are interfaced with PYTHIA 8.212 [75]. The PYTHIA parameters for the underlying event description are set to the CUETP8M1 tune [76] for samples corresponding to the 2016 running conditions, while the CP5 tune [77] is used for samples produced to be used with the 2017-2018 running conditions. The parton distribution function set NNPDF3.0 [78] is

Table 4.1: Datasets used for main analysis. Data corresponds to the ReReco processing unless specified as prompt reco. The integrated luminosity and the run-ranges are shown for each data period.

Primary Dataset	Era	Campaign	Processed and certified L (fb^{-1})	Run range
JetHT	Run2016B	07Aug17	5.7	272007–275376
JetHT	Run2016C	07Aug17	2.6	275657–276283
JetHT	Run2016D	07Aug17	4.2	276315–276811
JetHT	Run2016E	07Aug17	3.9	276831–277420
JetHT	Run2016F	07Aug17	3.1	277772–278808
JetHT	Run2016G	07Aug17	7.4	278820–280385
JetHT	Run2016H	07Aug17	8.6	280919–284044
JetHT Run2016 Total			**35.5**	
JetHT	Run2017B	17Nov2017-v1	4.8	297046–299329
JetHT	Run2017C	17Nov2017-v1	9.7	299368–302029
JetHT	Run2017D	17Nov2017-v1	4.2	302030–303434
JetHT	Run2017E	17Nov2017-v1	9.3	303824–304797
JetHT	Run2017F	17Nov2017-v1	13.5	305040–306462
JetHT Run2017 Total			**41.5**	
JetHT	Run2018A	17Sep2018	13.7	315252–316995
JetHT	Run2018B	17Sep2018	6.9	317080–319310
JetHT	Run2018C	PromptReco	6.9	319337–320065
JetHT	Run2018D	PromptReco	31.6	320673–325175
JetHT Run2018 Total			**59.2**	
SingleMuon	Run2016B	07Aug17	5.7	272007–275376
SingleMuon	Run2016C	07Aug17	2.6	275657–276283
SingleMuon	Run2016D	07Aug17	4.2	276315–276811
SingleMuon	Run2016E	07Aug17	4.0	276831–277420
SingleMuon	Run2016F	07Aug17	3.1	277772–278808
SingleMuon	Run2016G	07Aug17	7.5	278820–280385
SingleMuon	Run2016H	07Aug17	8.0	280919–284044
SingleMuon Run2016 Total			**35.2**	
SingleMuon	Run2017B	17Nov2017	4.8	297046–299329
SingleMuon	Run2017C	17Nov2017	9.6	299368–302029
SingleMuon	Run2017D	17Nov2017	4.2	302030–303434
SingleMuon	Run2017E	17Nov2017	9.3	303824–304797
SingleMuon	Run2017F	17Nov2017	13.2	305040–306462
SingleMuon Run2017 Total			**41.1**	
SingleMuon	Run2018A	17Sep2018	13.8	315252–316995
SingleMuon	Run2018B	17Sep2018	7.1	317080–319310
SingleMuon	Run2018C	PromptReco	6.9	319337–320065
SingleMuon	Run2018D	PromptReco	31.2	320673–325175
SingleMuon Run2018 Total			**59.0**	

used to produce all simulated samples. Pileup effects are simulated by superimposing minimum bias collisions on respective hard scattering events. Since the data taking conditions differ in each year, samples are generated with an approximate number of inelastic collisions per bunch crossing, which are then re-weighted to reproduce the distribution of the number of primary vertices in each respective data-taking year.

Table 4.2 shows the list of background samples used in this analysis and their respective cross sections.

Table 4.2: List of background MC samples. H_{T} denotes the scalar sum of transverse momenta of all jets in the events. Datasets marked with * are LO in QCD and EWK but will have NLO corrections applied.

Sample description	Generator	σ [pb]
QCD multijets, $500\,\mathrm{GeV} < H_{\mathrm{T}} < 700\,\mathrm{GeV}$	MadGraph5_aMC@NLO	29370
QCD multijets, $700\,\mathrm{GeV} < H_{\mathrm{T}} < 1000\,\mathrm{GeV}$	MadGraph5_aMC@NLO	6524
QCD multijets, $1000\,\mathrm{GeV} < H_{\mathrm{T}} < 1500\,\mathrm{GeV}$	MadGraph5_aMC@NLO	1064
QCD multijets, $1500\,\mathrm{GeV} < H_{\mathrm{T}} < 2000\,\mathrm{GeV}$	MadGraph5_aMC@NLO	121.5
QCD multijets, $2000\,\mathrm{GeV} < H_{\mathrm{T}} < \infty$	MadGraph5_aMC@NLO	25.42
$\mathrm{W}(\mathrm{q\bar{q}})$+jets, $400\,\mathrm{GeV} < H_{\mathrm{T}} < 600\,\mathrm{GeV}$ *	MadGraph5_aMC@NLO	315.6
$\mathrm{W}(\mathrm{q\bar{q}})$+jets, $600\,\mathrm{GeV} < H_{\mathrm{T}} < 800\,\mathrm{GeV}$ *	MadGraph5_aMC@NLO	68.57
$\mathrm{W}(\mathrm{q\bar{q}})$+jets, $800\,\mathrm{GeV} < H_{\mathrm{T}} < \infty$ *	MadGraph5_aMC@NLO	34.9
$\mathrm{Z}(\mathrm{q\bar{q}})$+jets, $400\,\mathrm{GeV} < H_{\mathrm{T}} < 600\,\mathrm{GeV}$ *	MadGraph5_aMC@NLO	145.4
$\mathrm{Z}(\mathrm{q\bar{q}})$+jets, $600\,\mathrm{GeV} < H_{\mathrm{T}} < 800\,\mathrm{GeV}$ *	MadGraph5_aMC@NLO	34.0
$\mathrm{Z}(\mathrm{q\bar{q}})$+jets, $800\,\mathrm{GeV} < H_{\mathrm{T}} < \infty$ *	MadGraph5_aMC@NLO	18.67
Hadronic $\mathrm{t\bar{t}}$	powheg 2.0	377.96
Semileptonic $\mathrm{t\bar{t}}$	powheg 2.0	365.34
Leptonic $\mathrm{t\bar{t}}$	powheg 2.0	88.29
Single-antitop t-channel	powheg 2.0	80.95
Single-antitop tW-channel	powheg 2.0	34.97
Single-top t-channel	powheg 2.0	136.02
Single-top tW-channel	powheg 2.0	34.91
Single-top s-channel	powheg 2.0	11.24
WW	MadGraph5_aMC@NLO	63.21
ZZ	MadGraph5_aMC@NLO	10.32
WZ	MadGraph5_aMC@NLO	22.82

W and Z corrections

The base W+Jets and Z+Jets MC samples are corrected to NLO in QCD and further re-weighted to incorporate the reduction of the cross section at high p_{T} due to higher-order electroweak effects (EW). The correction factors are applied to match the generator-level p_{T} distributions with those predicted by the highest available order in the perturbative expansion.

The higher order QCD and EW corrections are computed in [79, 80, 81, 82] and shown in Fig. 4.1 as a function of p_{T}.

To reweight the base samples to achieve the same accuracy, a dedicated, large generator level NLO Z and W sample was generated using MadGraph5_*a*mc@nlo merged with FxFx merging [83]. The nominal LO samples are then re-weighted to match the NLO samples according to their generated boson p_{T}. Fig. 4.2 shows a comparison of the p_{T} distributions of the base MC and the QCD NLO and QCD+EW NLO corrected simulations. The ratio shows the size of these corrections as function of the boson p_{T}, which are used as the event-by-event weight. Good agreement between the reweighted base sample and NLO samples are observed.

The uncertainties due to these corrections are evaluated as a function of the p_{T} and propagated later as normalization systematic. There are three components in these corrections, namely QCD, EW and EW-pT, where the last component accounts for the different p_{T} dependence of the EW corrections to W and Z. The uncertainty due to QCD higher-order corrections is evaluated by reweighting the QCD NLO samples by event-by-event factors for alternative renormalization and factorization scale choices

(where the up and down variations are scaled by 0.5 and 2 respectively). This gives an overall modification on the W and Z normalization that is 10% and flat across the whole p_{T} range.

For the EW correction uncertainty, we take a conservative approach by applying the full correction from QCD NLO to QCD+EW NLO as the uncertainty. We decorrelate the full EW correction per p_{T} bin. This corresponds to an independent uncertainty of $5-7\%$ of the W and Z combined normalization for the p_{T} bins ranging from $450-1200\,\mathrm{GeV}$ respectively. Because the origins of the electroweak corrections to W and Z processes differ, we decorrelate the W and Z normalization by $2-6\%$ in each p_{T} category.

Table 4.3 summarizes the uncertainty scheme adopted for QCD and EW correction for different p_{T} bins.

Table 4.3: Uncertainty of different components of NLO k-factors applied to W and Z+jets re-weighting as a function of p_{T} of the vector boson.

Process	W,Z		W
p_{T} range (GeV)	QCD	EW	EW-pT
450 - 500	10%	5%	2%
500 - 550	10%	5%	2%
550 - 600	10%	5%	2%
600 - 675	10%	7%	6%
675 - 800	10%	7%	6%
800 - 1200	10%	7%	6%

4.2.3 Signal simulations

Since this search does not have explicit selections to differentiate different production modes, we simulate all the SM Higgs boson production modes at the LHC. All signal

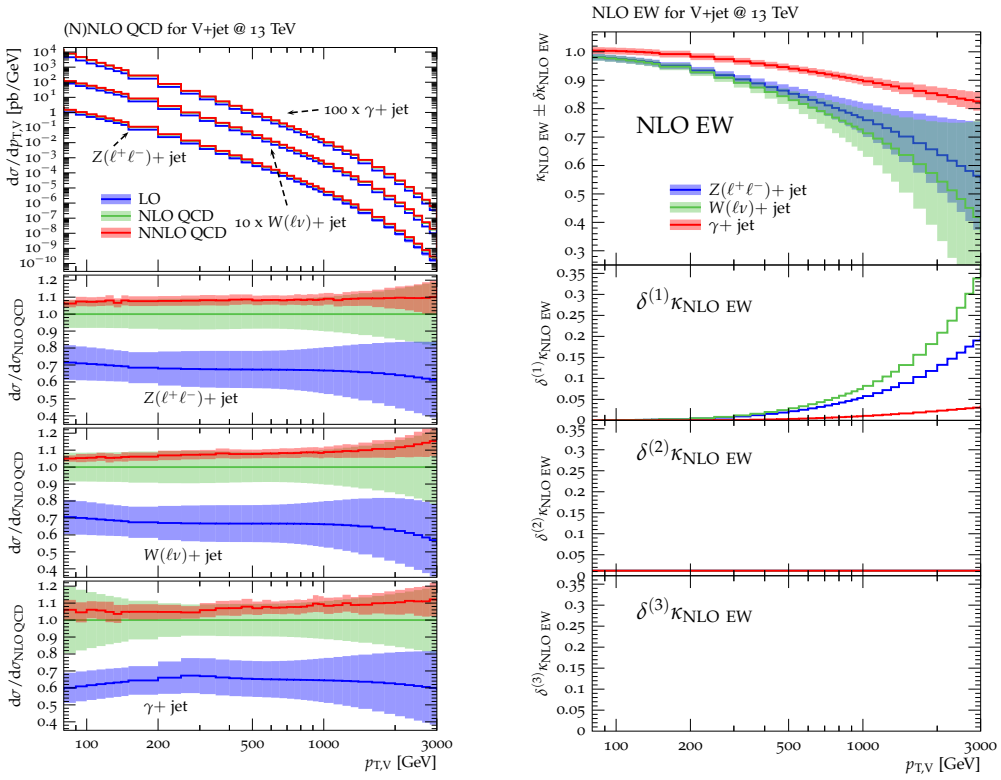


Figure 4.1: NLO QCD (left) and EW (right) corrections and the corresponding uncertainties. Ref [79]

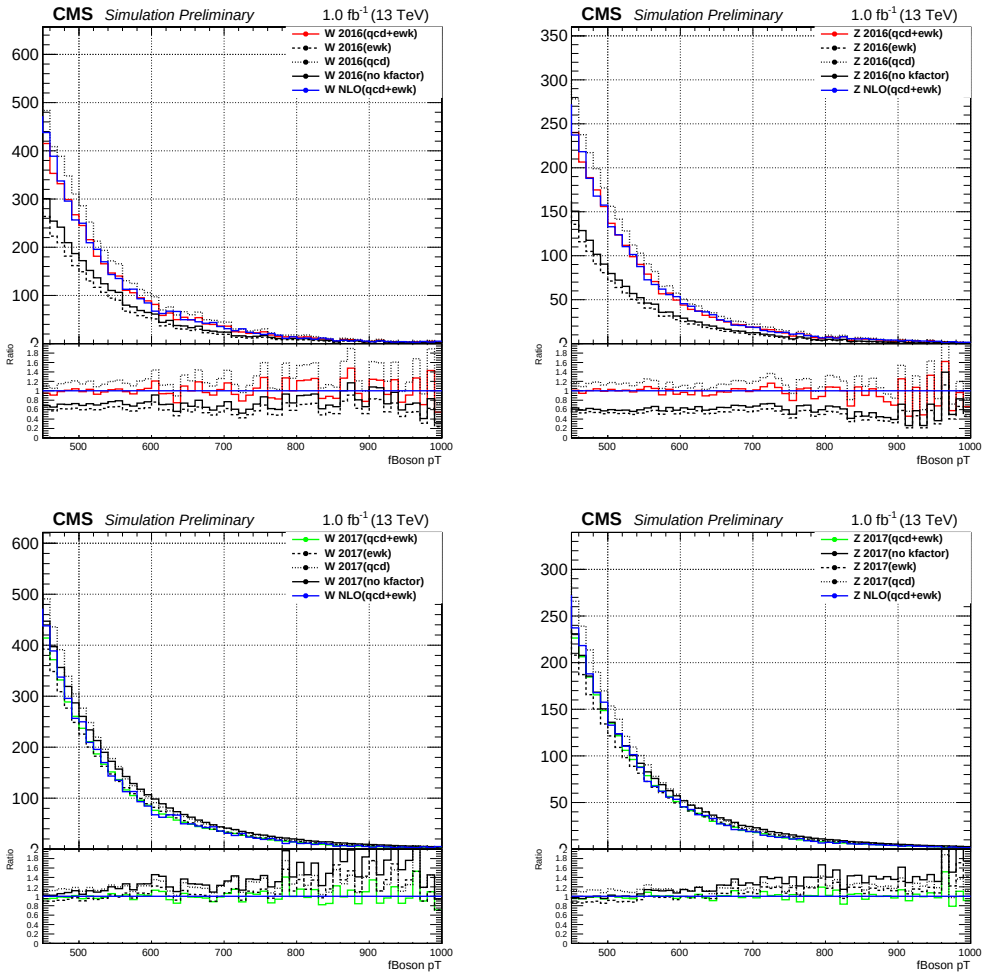


Figure 4.2: Boson p_{T} distributions at the generator level for W (left) and Z (right) in 2016 (upper) and 2017/18 (lower) generation configurations, showing the uncorrected spectrum, as well as the successive corrections to arrive at NLO QCD+EW. Also included is the target sample, generated as described in the text ("V NLO"), with NLO EW corrections applied. The ratio is shown with respect to the target sample.

samples, except for the gluon fusion mode, are generated with the POWHEG generator, interfaced with PYTHIA 8.212 for parton showering, fragmentation and hadronization. Pile-up reweighting to match the data-taking conditions in each year were done with the same procedure as described in Sec. 4.2.2. A list of all signal MC samples and the corresponding cross sections are given in Table 4.4.

Table 4.4: List of signal MC samples. In all samples, the Higgs boson is required to decay to $b\bar{b}$. Cross sections do not include $H \to b\bar{b}$ branching ratio. All samples are generated with the POWHEG generator except for the HJ-MiNLO gluon fusion sample.

Sample description	σ [pb]
Gluon fusion	48.85
Gluon fusion (HJ-MiNLO)	27.80
Vector boson fusion	3.78
ZH	0.7612
gg→ZH	0.1230
W^+H	0.8400
W^-H	0.5330
ttH	0.5071

Higgs p_T spectrum for gluon fusion production

Computing the Higgs boson p_T spectrum for the gluon fusion production mode in the region of interest for the analysis is theoretically challenging for a number of reasons. First, the gluon fusion production is loop induced, which implies the process is very sensitive to higher order corrections. This can be seen from how the cross section converges as higher order effects are included in the calculation. When going from the LO to the next to NLO, the cross section increases by more than a factor of 2 and then increases by an extra 30% going from NLO to the current highest order

of N^{3}LO [26]. Just as any loop induced processes, additional modification must be considered at the energy scale where the loop becomes resolved. For Higgs production, this occurs at twice the mass of the of the particles that yields the dominant loop contribution, which is the top quark. The resolution of top loop occurs roughly at a p_{T} of 350 GeV. While the finite top-mass correction is not very important in the inclusive cross section, it can become relatively large at the large p_{T} region. The Higgs production calculated with the heavy (or infinite) top-mass approximation is referred to as Effective Field Theory (EFT).

Motivated by the previous search with 2016 data [1], the full finite top-mass correction has been calculated recently in Ref. [84, 85]. Based on the above results, the LHC Higgs Cross Section Working Group (LHCHXSWG) has recently put forth a recommendation for high-p_{T} Higgs boson production [86]. The "best prediction" for the differential cross section is currently obtained by taking a fixed order H + 1j calculation in the $m_{\mathrm{t}} \to \infty$ EFT approximation at NNLO in QCD, and then rescaling it by the ratio of the NLO EFT cross section to the NLO cross section computed using fully resolved top loops ("QCD NLO"),

$$\frac{d\sigma^{\mathrm{best}}}{dp_{\mathrm{T}}} = \frac{\frac{d\sigma^{\mathrm{QCD\ NLO}}}{dp_{\mathrm{T}}}}{\frac{d\sigma^{\mathrm{EFT\ NLO}}}{dp_{\mathrm{T}}}} \frac{d\sigma^{\mathrm{EFT\ NNLO}}}{dp_{\mathrm{T}}}. \tag{4.1}$$

It is observed that in both the EFT and the full QCD calculation, the correction from LO to NLO is uniform across p_{T}, whereas including full top mass effect softens the Higgs p_{T} spectrum with respect to the EFT. This can be seen from Fig. 4.3. Assuming the exact NNLO QCD correction follow the same pattern, Eq. 4.1 yields a

very good approximation to the exact NNLO with full top mass effect.

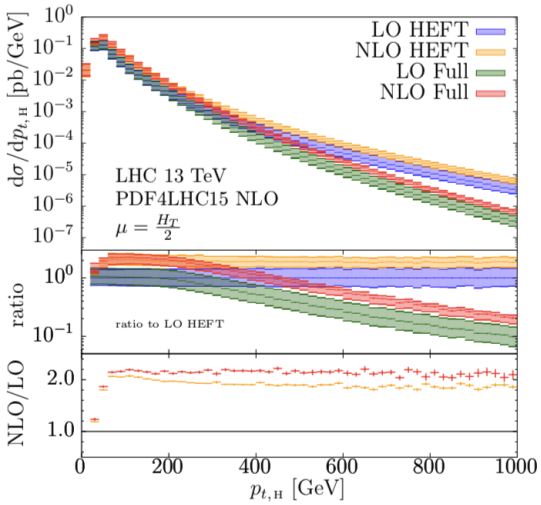


Figure 4.3: Transverse momentum distribution of the Higgs boson at the LHC with $\sqrt{s} = 13\,\text{TeV}$ computed in Ref. [84]. The upper panel shows absolute predictions at LO and NLO in the full SM and in the infinite top-mass approximation (EFT). The lower panel shows respective NLO/LO correction factors. The bands indicates theoretical errors of the full SM result due to scale variation.

In CMS, we have evaluated the differential cross section of various samples, each with different schemes of estimating the top-mass effect, and compared them with respect to the LHCXSWG recommendation. As a baseline, the Madgraph sample(MG5 $m_t \to \infty$) is generated at NLO with the EFT, which assumes an infinite top-mass approximation. Next, the "CMS standard" POWHEG sample (POWHEG gg_H_quark-mass-effects) also uses EFT but further approximates the finite top mass correction with the h parameter [87], which is a scale of choice when the Sudakov suppression becomes important. The previous iteration of this analysis ("HIG-17-010 sample") [1] starts from nominal POWHEG rescaled to the NNLO EFT inclusive cross

section, and includes an additional correction to account for top mass effects as described in Ref. [1]. Lastly, the HJ-MiNLO (POWHEG-HJMINLO) is the generator recommended by the LHCHXSWG, which is shown in their study to be compatible with the latest fix-order calculations with full finite-top mass effect.

All distributions are presented in Fig. 4.4, as inclusive cross sections as a function of a minimum $p_{\mathrm{T}}^{\mathrm{H}}$ cut.

Following the LHCHXSWG recommendation, the HJ-MiNLO is chosen as the nominal signal model in the analysis for gluon fusion production, whereas the "HIG-17-010" sample is used as a reference for comparison with previous results in Ref. [1].

4.3 Triggers

Given the relative simple final state objects, the main trigger strategy is to trigger on an energetic jet or sufficient hadronic energy deposition in the calorimeters. As the instantaneous luminosity of the LHC has increased over the data taking periods, the thresholds in the trigger paths are also increased to maintain a manageable rate for the CMS trigger system. Besides raising the threshold, we also developed trigger paths with a small trim mass requirement or very loose b-tagging requirement to maintain a low p_{T} threshold for the jet. Trimming [88] is a jet grooming algorithm in which subjets are clustered with a smaller cone size than the original jet, and are removed if the p_{T} of the subjet is smaller than a chosen fraction of the original jet. Trimming is chosen to be the grooming algorithm at the HLT level due to its simplicity and robustness against pile-up.

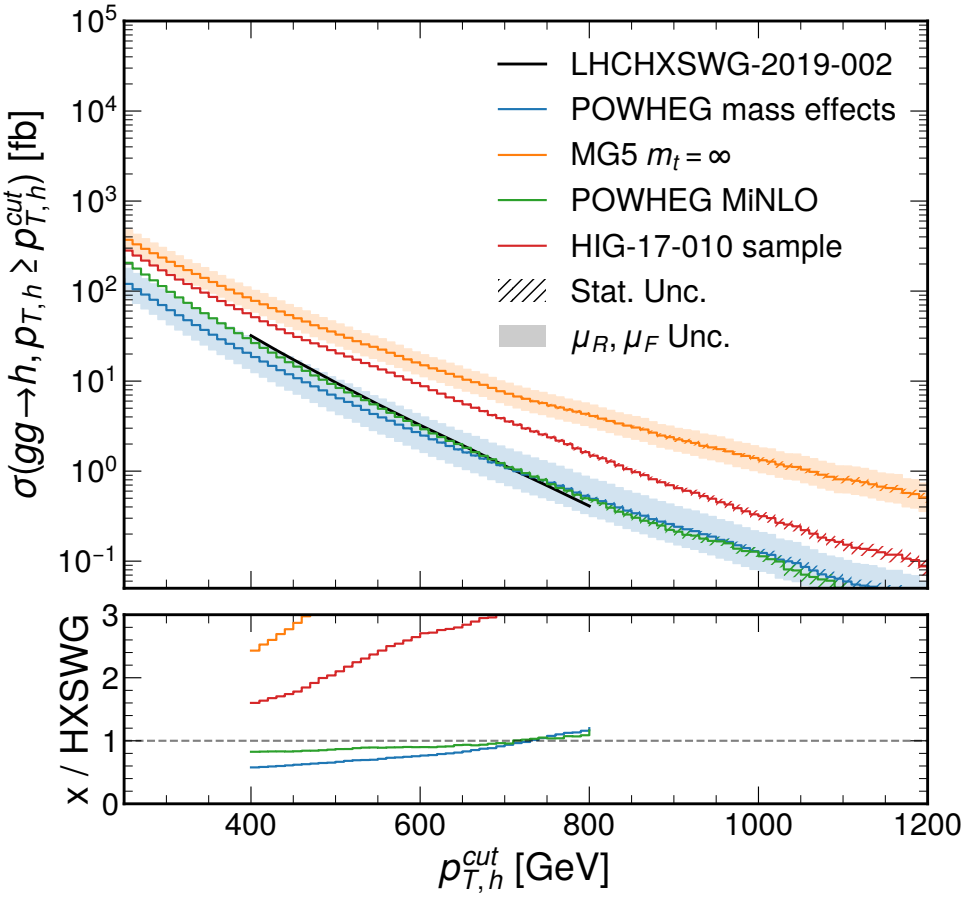


Figure 4.4: Comparison between the LHC Higgs cross section working group prescription [86] for the $p_{\mathrm{T}}^{\mathrm{H}}$ spectrum and the spectrum found in several CMS datasets. The factorization and renormalization scale uncertainties were evaluated as the envelope of the 7 scale variations stored as weights in the samples. For MINLO, the weights were not available due to technical issues, and for the HIG-17-010 sample they are the same as found in the POWHEG mass effects sample. In the lower panel, the ratio of each described p_{T} spectrum relative to the LHCHXSWG spectrum is shown. See text for further details on the various samples.

To maximize the signal efficiency, we use the of logical "OR" of the following un-prescaled triggers, in an approximate decending order of contribution to trigger efficiencies:

- HLT_PFHT[800/900/1050]
- HLT_AK8PFJet[360/400/420]_TrimMass30
- HLT_AK8PFHT800_TrimMass50
- HLT_AK8PFHT700_TrimR0p1PT0p03Mass50
- HLT_PFJet[450/500]
- HLT_AK8PFJet[450/500]
- HLT_PFHT650_WideJetMJJ[900/950]DEtaJJ1p5
- HLT_AK8PFJet330_BTagCSV_p17
- HLT_AK8DiPFJet280_200_TrimMass30_BTagCSV_p20
- HLT_AK8PFJet330_TrimMass30_PFAK8BoostedDoubleB_np2

Not all trigger paths are available through out the data taking periods, due to the updated thresholds. When available, the path with the lowest thresholds are used. A brief description of the triggers used in the analysis is shown in Table A.1.

Trigger Efficiency

The efficiencies of these triggers were measured and studied using the `SingleMuon` dataset and the single muon trigger `Mu50` (described also in Table A.1) is used as reference trigger. Offline selections are chosen to be in a region of phase space where the trigger efficiency is close to 100%. Then events are selected by asking at least one AK8 Puppi jet with $p_{\mathrm{T}} > 180\,\mathrm{GeV}$ and $|\eta| < 2.5$, passing the tight jet ID requirement. Since some of the boosted triggers have two legs (i.e. AK8/AK4 H_{T} or AK8 jet p_{T} and jet mass), the trigger efficiency is measured in a 2D space: leading jet p_{T} vs softdrop mass. The efficiencies are then computed with the following formula:

$$\text{Efficiency} = \frac{\text{Pass Mu50 \&\& (OR of Jet paths) \&\& offline cuts}}{\text{Pass Mu50 \&\& offline cuts}} \tag{4.2}$$

For the 2016 dataset, the measured trigger efficiencies is shown in Fig. 4.5.

In Run 2017B, the `AK8PFJet(400/380/360)_TrimMass30`, `AK8PFHT800_TrimMass50`, and `AK8PFJet330_BTagCSV_p17` triggers were not online and therefore, the "OR" of the triggers does not become fully efficient until $\sim$ 650 GeV, as shown in Fig. 4.6 where the measured efficiency of the triggers logical "OR" is shown as a function of jet p_{T} and jet m_{SD} separately for each Run period. For Runs 2017C-F, each become approximately fully efficient for jet $p_{\mathrm{T}} > 450\,\mathrm{GeV}$. The trigger efficiency "OR" of the triggers for $p_{\mathrm{T}} > 450$ GeV for Run 2017B-F and Run 2017C-F are shown in Fig. 4.7.

Similarly, Fig. 4.8 shows the trigger efficiency of the "OR" of the triggers for 2018 data, split by different run periods. The trigger efficiency for the "OR" of the triggers for Run 2018A-D are shown in Fig. 4.9.

Since some trigger paths involve some loose b-tagging requirements, it is also checked that the trigger efficiency is also close to fully efficient as a function of deep double-b. This is illustrated in Fig. 4.10.

Finally, simulated events are reweighted accordingly to the measured efficiency as function of jet p_{T} and jet m_{SD}. An overall systematic uncertainty of 2% per data-taking year is assigned to this measured efficiency.

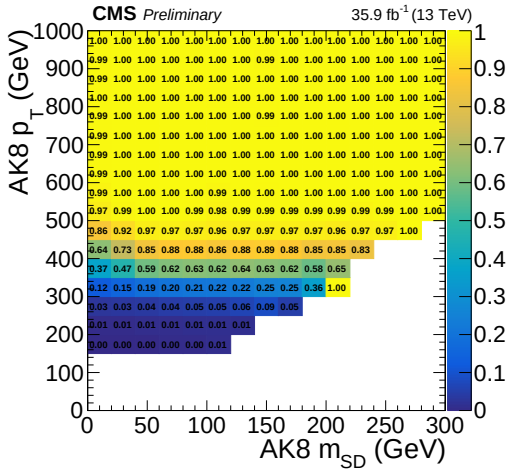


Figure 4.5: Measured trigger efficiency of the "OR" as function of jet p_{T} and jet m_{SD} is shown for 2016.

4.4 Objects reconstruction

In this section, we will describe how each of the physics object used in this analysis are reconstructed. Undoubtedly, the most important object are jets. This analysis hinges on a good understanding of jet mass and jet substructure to reconstruct the resonances $(\mathrm{W}, \mathrm{Z}, \mathrm{H})$ on top of a smoothly falling background. We will discuss how

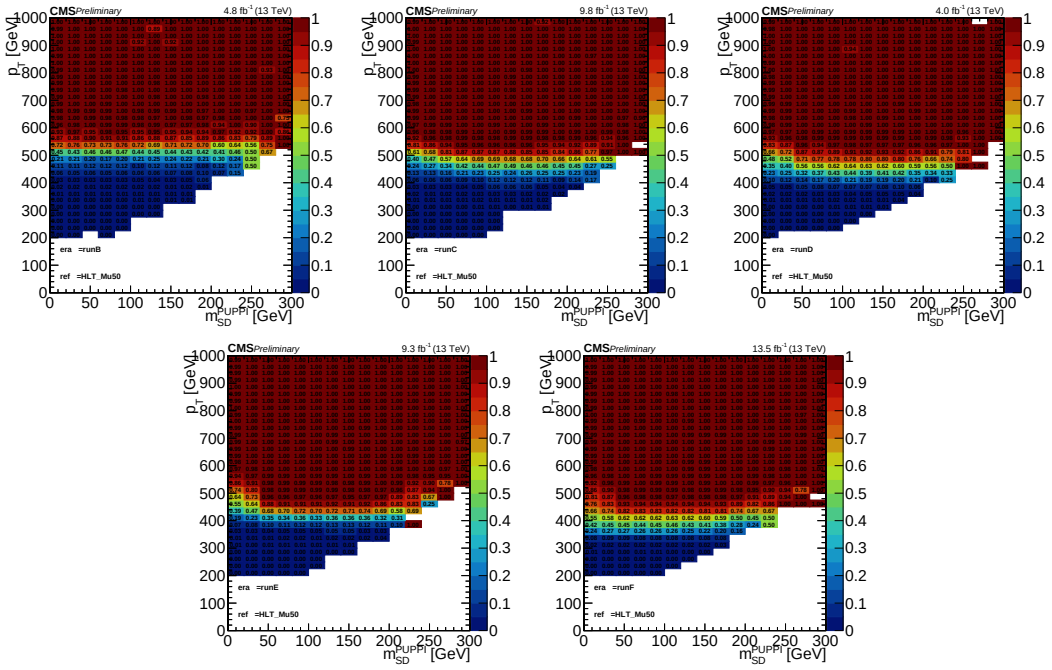


Figure 4.6: Measured trigger efficiency of the "OR" as function of jet p_{T} and jet m_{SD} is shown separately for each Run period in 2017.

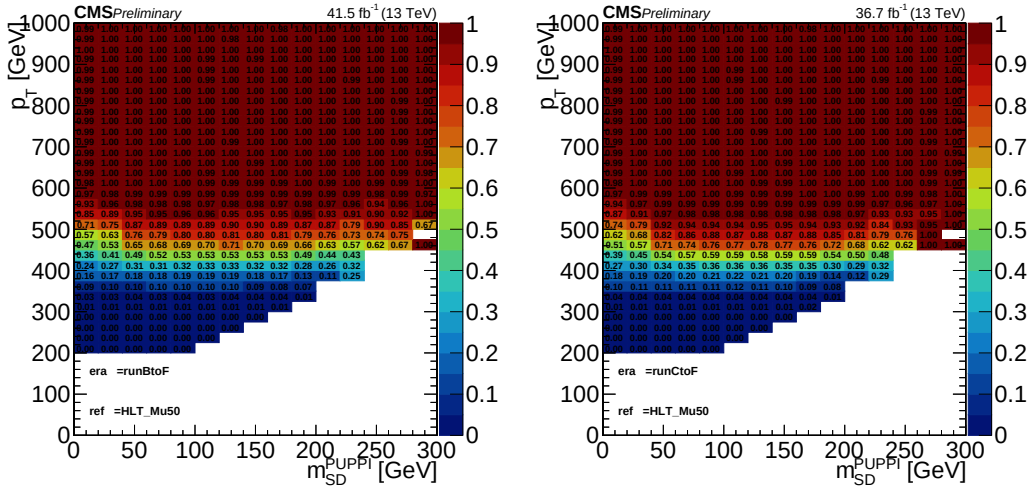


Figure 4.7: Measured trigger efficiency of the "OR" as function jet p_{T} and m_{SD} jet for Run 2017B-F and Run 2017C-F.

.

the substructure selection is explicitly decorrelated with the mass and p_{T} to provide a foundation for the analysis method. We will also discuss the details of the new deep double-b tagger. The rest of the section will focus on selections that aims to reduce conteminations from other SM processes.

4.4.1 Jets

Jets used in the analysis are reconstructed by clustering particle flow candidates in the event using the anti-k_{T} algorithm [49] with a distance parameter of 0.8.

To mitigate the PU effect in AK8 jets, the pileup per particle identification (PUPPI) algorithm [89] is used to weight the particle flow candidates prior jet clustering. This pileup jet discriminator is built based on event pileup properties, local shape and tracking information.

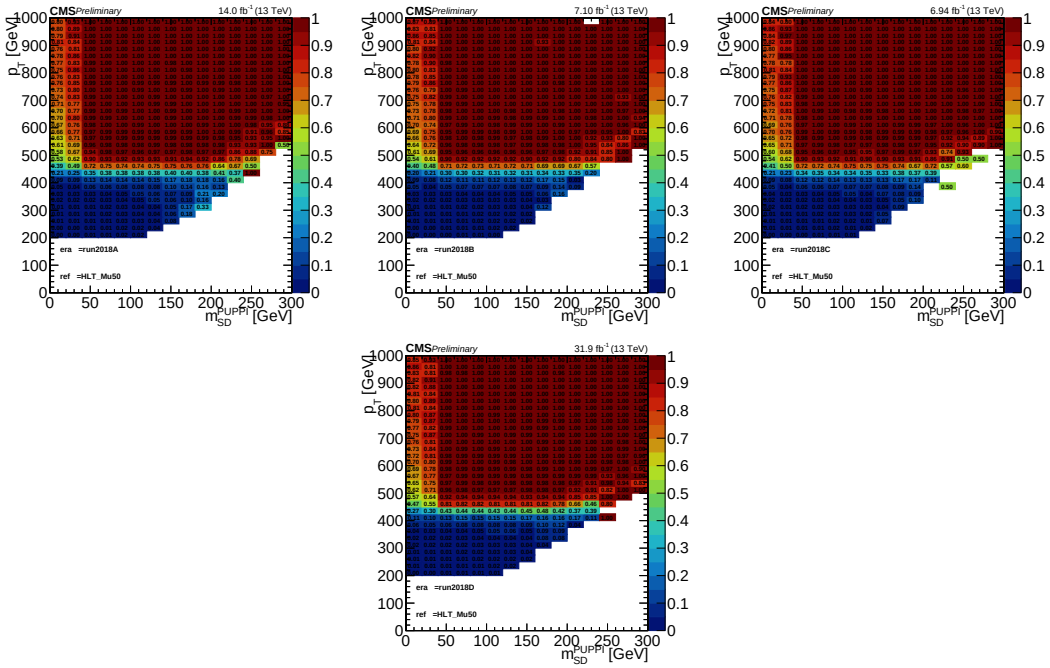


Figure 4.8: Measured trigger efficiency of the "OR" as function of jet p_{T} and jet m_{SD} is shown separately for each Run period in 2018.

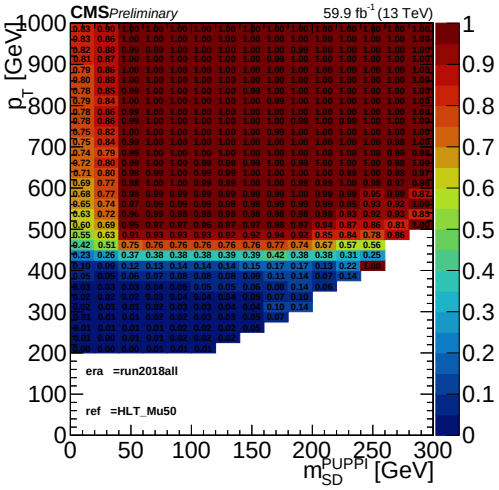


Figure 4.9: Measured trigger efficiency of the "OR" as function jet p_{T} and m_{SD} jet for Run 2018A-D.

.

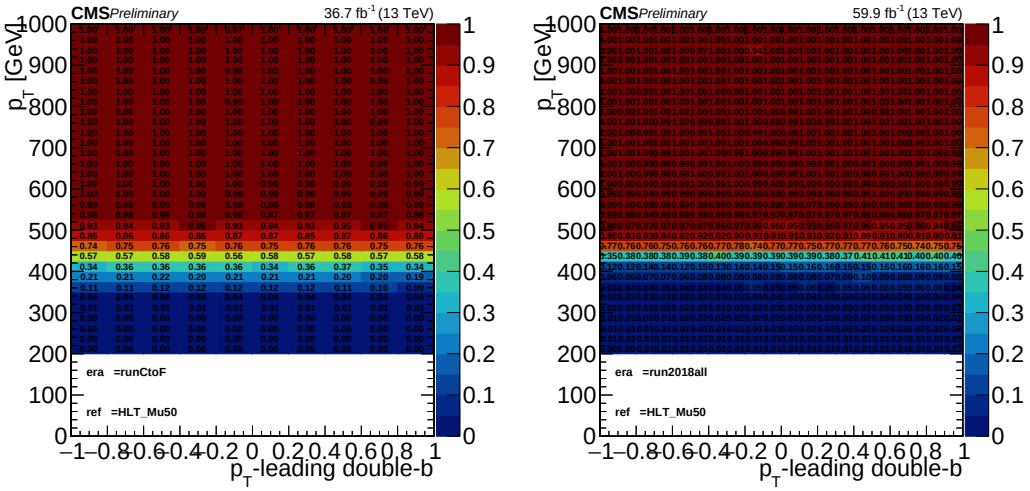


Figure 4.10: Measured trigger efficiency of the "OR" as function jet p_{T} and deep double-b score for all 2017 and 2018 runs

.

The latest jet energy corrections are applied to data (`Summer16_07Aug2017[BCD,EF,GH]_V11_DATA`, `Fall17_17Nov2017[B,C,DE,F]_V32_DATA` and `Autumn18_Run[A,B,C,D]_V8_DATA`) and MC (`Summer16_07Aug2017_V11_MC`, `Fall17_17Nov2017_V32_MC`,

`Autumn18_V8_MC`), following the JME POG recommendations [90]. Since the jet energy resolution (JER) differs in data and MC, in order to get a better agreement an additional smearing is applied to the simulation. The `Summer16_25nsV1`(2016),`Fall17_V3`(2017,2018) tags are currently applied to respective MC samples.

The official JME "tight" PF jet ID [91] is also applied. We reject a jet if an electron, a muon or a photon with sufficient energy and passing good quality requirements is found within a cone size of $\Delta R = 0.8$ around the jet. For an electron or a muon, the p_T thresholds are 40 GeV and passing tight ID [92, 93]; for photon, the p_T threshold is 175 GeV and with medium ID [93].

The AK8 PUPPI jets are required to have $p_T > 450\,\text{GeV}$ to be on the trigger plateau and be within the $|\eta| < 2.5$. As will be discussed later in Sec. 4.5.1, the AK8 PUPPI jets leading in p_T in the event is chosen as the Higgs candidate jet. In Fig. 4.11 the jet p_T distribution is shown for the p_T leading jet for simulated signal (left) and background events overlayed with signal events(right).

4.4.2 Jet mass

In Sec. 3.3.6, we discussed the reconstruction and motivation for the soft drop jet mass (m_{SD}) and we evaluate the variable in the samples in the section. The m_{SD} distribution peaks at the H mass for signal events and pushes the mass of background quark- and gluon-initiated events towards lower values. In Fig. 4.12 the jet m_{SD} distribution is shown for the p_T leading jet for simulated signal (left) and background (right) events.

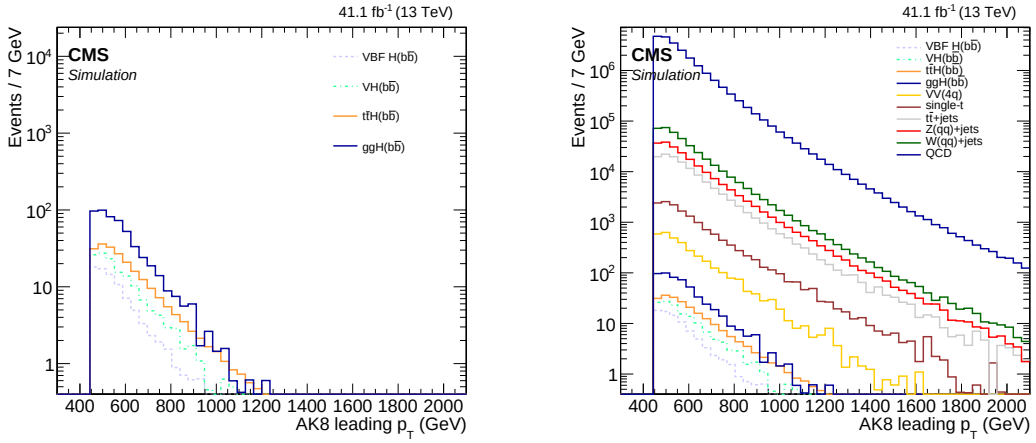


Figure 4.11: AK8 jet p_{T} distribution. Left, simulated signal events normalized to their cross sections (Table 4.4). Right, background and signal events normalized to their cross section.

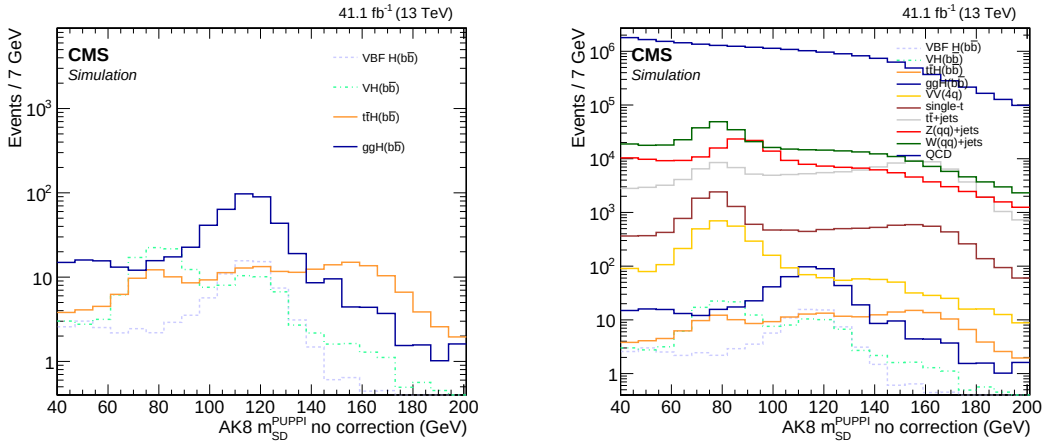


Figure 4.12: Uncorrected soft drop mass distribution for the p_{T}-leading jet. Left, simulated signal events normalized to their cross sections (Table 4.4). Right, background and signal events normalized to their cross sections.

The spectrum of the corrected m_{SD} is shown in Fig. 4.13 for simulated signal (left) and background (right) events.

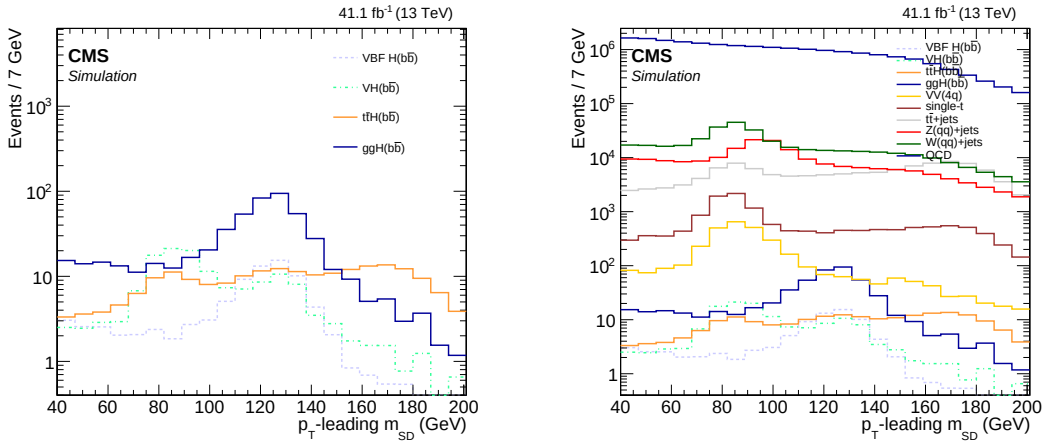


Figure 4.13: Corrected soft drop mass distribution for the p_{T}-leading jet. Left, simulated signal events normalized to their cross sections (Table 4.4). Right, background and signal events normalized to their cross sections.

4.4.3 Jet ρ

To avoid instabilities at the edges of the distribution, due to finite cone effects from the AK8 jet clustering (around $\rho \sim -2$), and to avoid the non-perturbative regime of the soft drop mass calculation (below $\rho \sim -6$), we only consider events in the ρ range $-6.0 < \rho < -2.1$. This requirement is fully efficient for the signal.

The distribution of ρ for data and MC, is shown in Fig. 4.14, for the full dataset after the kinematic selection. The region defined by values of $\rho < -6$, is where the Higgs boson is not fully merged or is merged but the kinematics of the subjets is such that one of them gets groomed away.

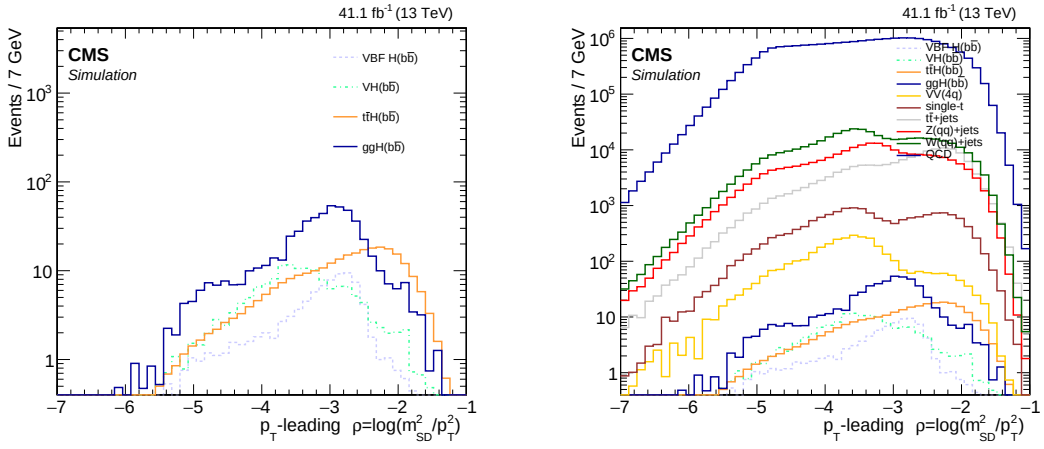


Figure 4.14: Left, distribution of ρ for simulated signal events. Right, data and simulation comparison of the ρ spectrum for jets with $p_{\mathrm{T}} > 450$ GeV. At analysis level, only events with $-6.0 < \rho < -2.1$ are required.

4.4.4 Jet substructure

The definitions and motivation of using N_2^1 as the jet substructure variable is discussed in Sec. 3.3.8. We focus on the deccorelation procedure of applying N_2^1 in this section. Because we want to preserve a smoothly falling jet mass distribution as a function of p_{T}, it is natural to determine a substructure variable's stability as a function of the QCD scaling variable $\rho = \log\left(m_{\mathrm{SD}}^2/p_{\mathrm{T}}^2\right)$. Since the QCD (quark or gluon-initiated) jet mass scales with p_{T}, decorrelating a given substructure variable as a function of ρ and p_{T} is a well-bounded procedure.

The decorrelation procedure starts by at finding the 2D contour map of N_2^1 as a function of ρ and p_{T} at a specific background efficiency, i.e.

$$N_2^1(\rho, p_{\mathrm{T}}) = \text{const},$$

where the constant is chosen by optimizing the expected sensitivity with respect to this substructure threshold and double-b tagging. The optimization procedure is described in details in Sec. 4.5.2. Given this map of the N_2^1 as a function of ρ and p_{T}, we define the "DDT" transformation as

$$N_2^{1,\mathrm{DDT}} = N_2^1 - N_2^1(\text{cut at } 26\%) \tag{4.3}$$

such that $N_2^{1,\mathrm{DDT}} < 0$ has a *constant* background efficency of 26% as a function of ρ and p_{T} by construction.

The decorrelation procedure is performed using the high statistic QCD MC samples. At the moderate background efficency of 26%, the large sample well populates the majority of the phase space, and we apply a mild smoothing procedure on top of the 26% quantile from direct MC events. Separate decorrelation maps are derived corresponding to each data-taking year. For 2016, the smoothing was done with the k-Nearest-Neighbour (kNN) procedure [64], while the 2017 and 2018 maps are smoothed with a gaussian filter of 1 sigma. The 2D maps corresponding to different years are shown in Fig. 4.17 for $\epsilon_{QCD} = 26\%$.

Using this transformation map, we can show now the correlation between $N_2^{1,\mathrm{DDT}}$ and ρ. We see this in Fig. 4.16 where now by definition the background is flat at 26% at a cut value of $N_2^{1,\mathrm{DDT}} = 0$. As an illustration, we compare the $N_2^{1,\mathrm{DDT}}$ efficiency and N_2^1 selection without DDT as a function of m_{SD} in Fig. 4.15

In Fig. 4.18 the jet $N_2^{1,\mathrm{DDT}}$ distribution is shown for the p_{T}-leading jet for simulated signal (left) and background (right) events.

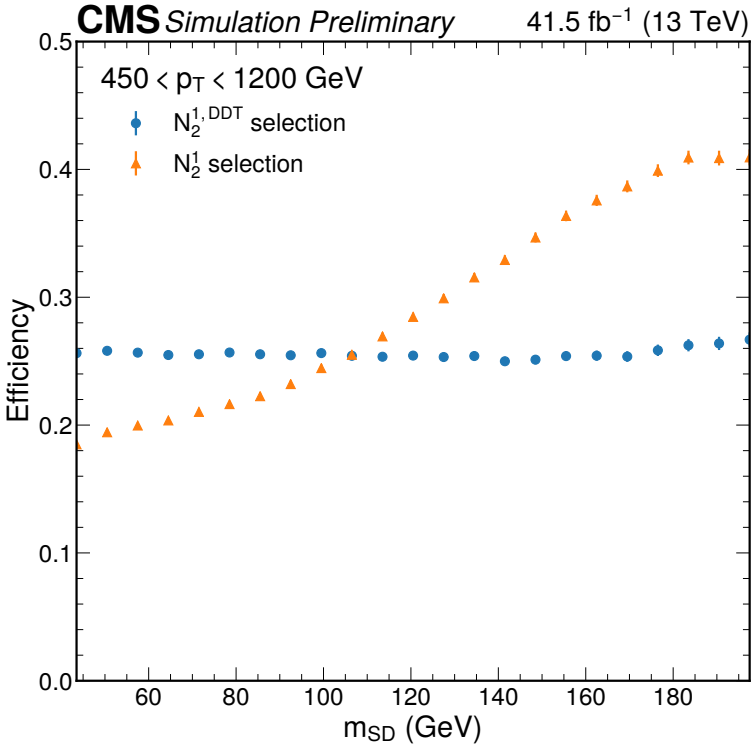


Figure 4.15: Comparison between $N_2^{1,\mathrm{DDT}}$ and N_2^1 efficiency as a function of m_{SD} with QCD simulation. The cut on N_2^1 corresponds to same overall efficiency of 26% in the analysis phase space.

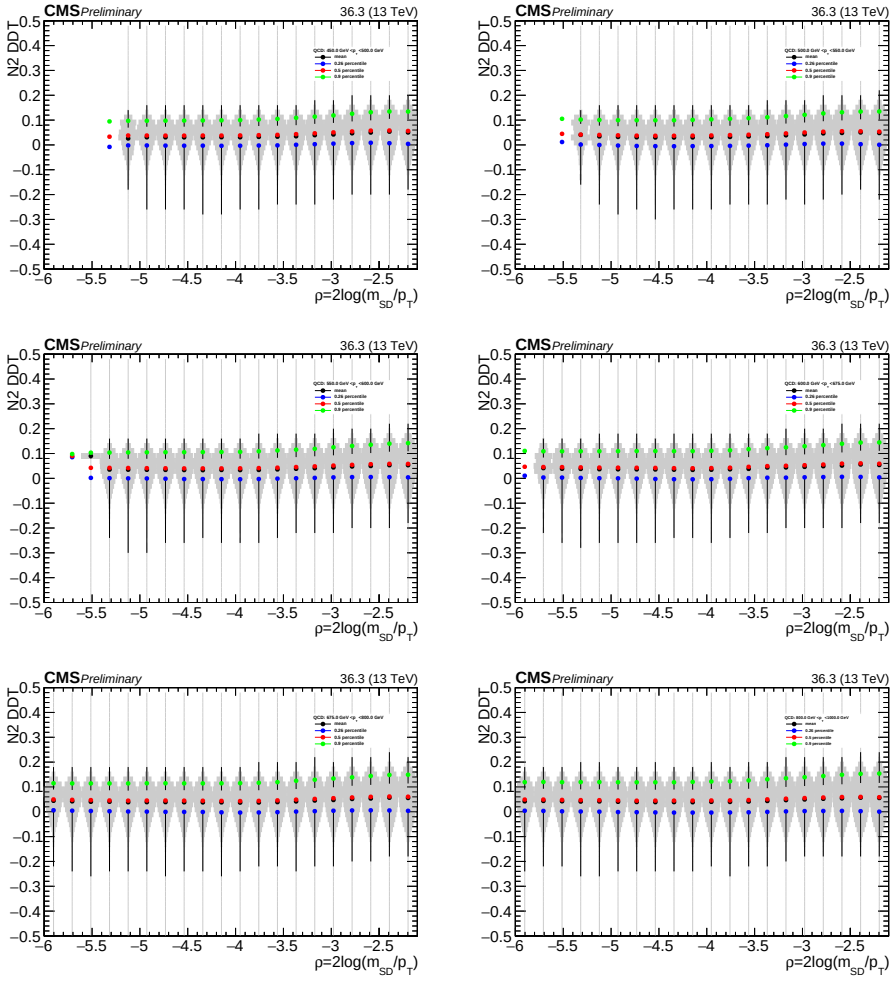


Figure 4.16: $N_2^{1,\mathrm{DDT}}$ kinematic dependence as a function of ρ in different p_T bins: p_T:450-500 GeV (top-left), p_T:500-550 GeV (top-right), p_T:550-600 GeV (middle-left), p_T:600-675 GeV (middle-right), p_T:675-800 GeV (bottom-left), p_T:800-1000 GeV (bottom-right). The blue points represent again the 26% quantile value of $N_2^{1,\mathrm{DDT}}$, which by construction is flat as function of ρ.

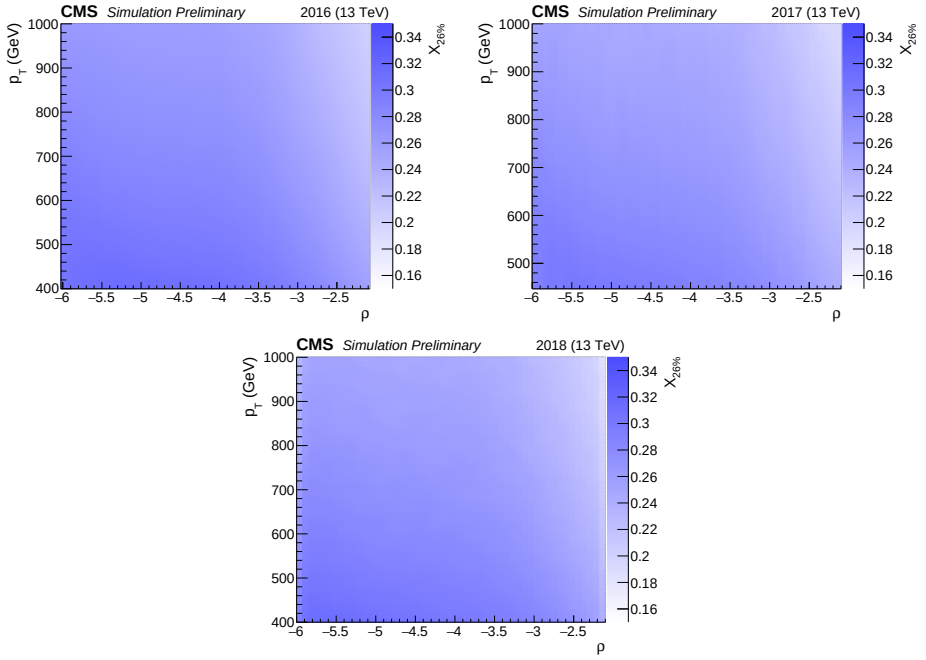


Figure 4.17: Transformation maps of $N_2^1 \to N_2^{1,\mathrm{DDT}}$ for different data-taking years(top left: 2016,top right: 2017, bottom: 2018) corresponding to the N_2^1 quantile value of 26% (i.e. for $\epsilon_{\mathrm{QCD}} = 26\%$), as a function of ρ and p_{T}.

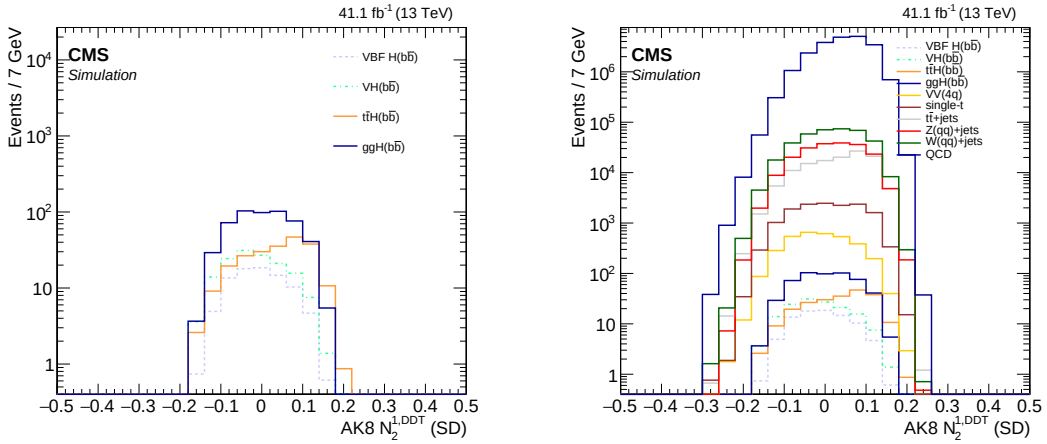


Figure 4.18: $N_2^{1,\mathrm{DDT}}$ distribution for the p_{T}-leading jet. Left, simulated signal events normalized to their cross sections (Table 4.4). Right, background and signal events normalized to their cross sections.

4.4.5 Deep double-b tagger

The details of the deep double-b tagger are discussed in Sec. 3.3.10. We focus here on the performance of the tagger in the simulated samples. With respect to the BDT-based double-b tagger, the decorrelated deep double-b tagger improves the H($b\bar{b}$) tagging efficiency by approximately 50% at the QCD misidentification rate of 1%. Figure 4.19 shows the comparisons of QCD mistagging rate as a function of signal efficiency between double-b tagger and deep double-b tagger. In Fig. 4.21 the deep double-b tagger distribution is shown for the p_{T} leading jet for simulated signal (left) and background (right) events.

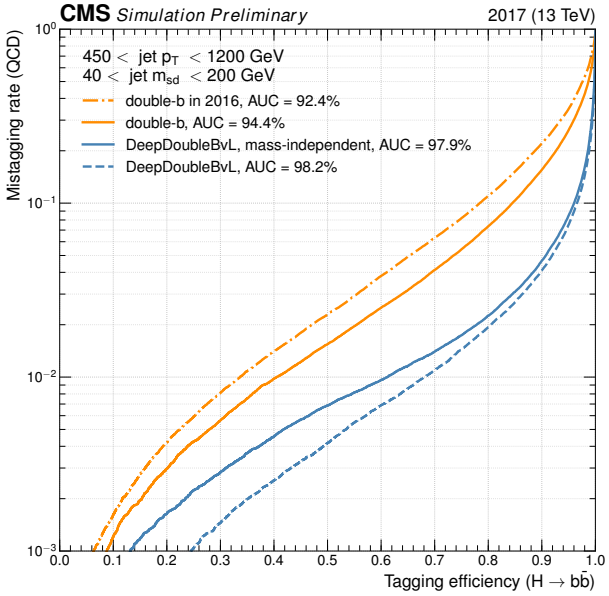


Figure 4.19: ROC curve for double-b evaluated on 2017 MC (orange), double-b evaluated on 2016 MC (orange dashed) and deep double-b (decorrelated) taggers, evaulated on 2017 MC.

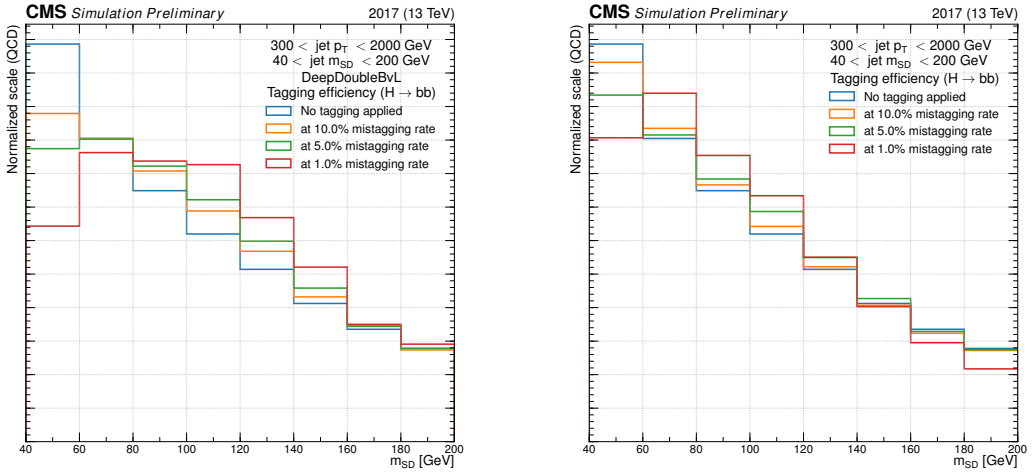


Figure 4.20: Mass sculpting for deep double-b without (left) and with (right) mass decorrelation.

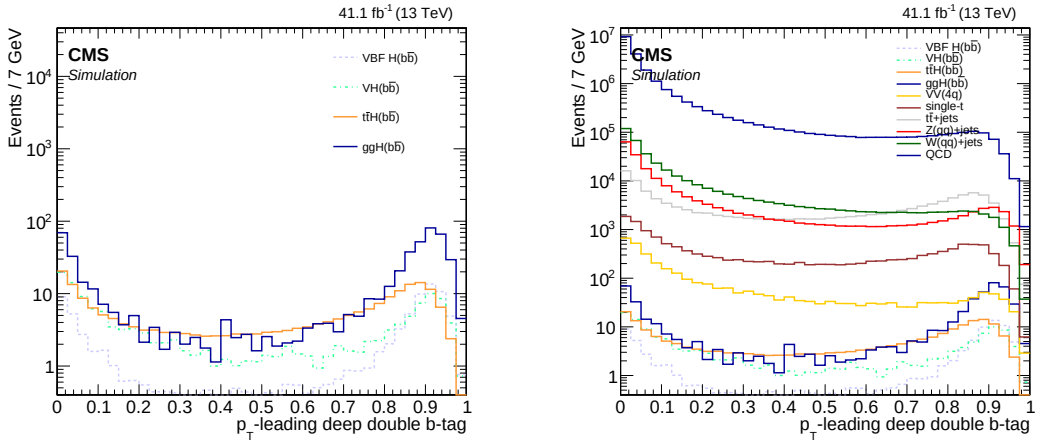


Figure 4.21: Deep double-b tagger distribution for the p_T-leading jet. Left, simulated signal events normalized to their cross sections (Table 4.4). Right, background and signal events normalized to their cross sections.

4.4.6 Lepton veto

To reduce contamination from other EW SM backgrounds, we veto events based on the presence of identified and isolated leptons. Two algorithms are used to reconstruct muons [47]: one proceeds from the inner tracker outwards, while the other starts from tracks measured in the muon chambers and matches them to those reconstructed in the silicon tracker. Muons are required to pass identification and isolation criteria. The muon isolation is computed as the sum of the transverse energy of the particles inside a cone of radius $\Delta R = 0.3$ around the muon direction divided by the muon transverse momentum. A correction for the pileup contribution, called β-correction, to the energy inside the cone is applied. The resulting β-corrected isolation variable is required to be lower than 0.25. Electrons are reconstructed using a Gaussian-sum filter algorithm [94], and each electron candidate must pass identification and β-corrected isolation of less than 0.15. Taus are reconstructed by identifying number of reconstructed charged/neutral pions compatible with tau decays. Isolation for taus are required to be less than 4.5 GeV with more than 3 hits for the charged candidates in the isolation sum.

The following table summarizes the p_{T},η and isolation requirement for leptons: Any events with any number of electrons, muon or taus satisfying the requirements

	e	μ	τ_h
p_{T} (GeV)	> 10	> 10	> 18
$\lvert\eta\rvert$	< 2.5	< 2.4	< 2.3
isolation	$< 15\%$	25%	< 5 GeV

Table 4.5: Lepton veto requirements, see text for description of isolation variables.

in Table 4.5 are vetoed. The ggH signal efficiency for these requirements is 98.6%, 100% and 97.8% respectively.

4.4.7 Additional AK4 b jets

In order to reduce the $t\bar{t}$ contamination we also veto additional AK4 jets which pass b-tagging selection. The DeepCSV algorithm [95] is used to identify AK4 jets that come from b-quarks. We have checked all three CSVv2 operating points and different p_T thresholds for central AK4 jets in order to build this additional veto, as shown in Fig. 4.22. We then ask that there is no medium DeepCSV b-tagged jet in the opposite hemisphere of the leading AK8 jet among the four leading-p_T AK4 jets.

In Fig. 4.23 the maximum DeepCSV discriminator value for the four leading AK4 jets in the opposite hemisphere of the AK8 jet is reported for simulated signal and background events. Note, the medium working point discriminator cut values are 0.6321,0.4941 and 0.4184 for 2016,2017 and 2018 data respectively.

4.4.8 Missing transverse energy

As discussed in Sec. 3.3.9, we compute the Type-I corrected MET in the analysis. Since we do not expect real E_T^{miss} for the dominant signal processes, we veto events with $E_T^{miss} > 140\,\text{GeV}$ in order to further reduce the $t\bar{t}$ background contamination. In Fig. 4.24 the E_T^{miss} is shown for simulated signal and background events.

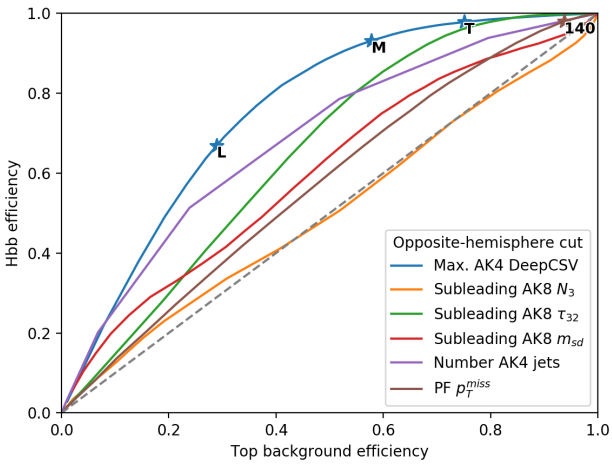


Figure 4.22: ROC curves for different $t\bar{t}$ sensitive variables.

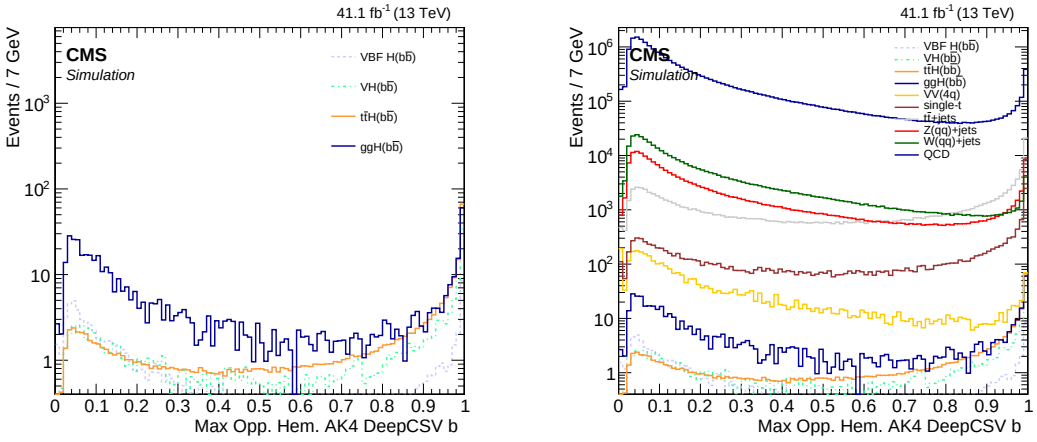


Figure 4.23: Distribution of the maximum DeepCSV discriminator value for the four leading-p_T AK4 jets in the opposite hemisphere from the leading-p_T AK8 jet. Note, the medium working point discriminator cut value is 0.4941 for 2017 data. Left, simulated signal events normalized to their cross sections (Table 4.4). Right, background and signal events normalized to their cross sections.

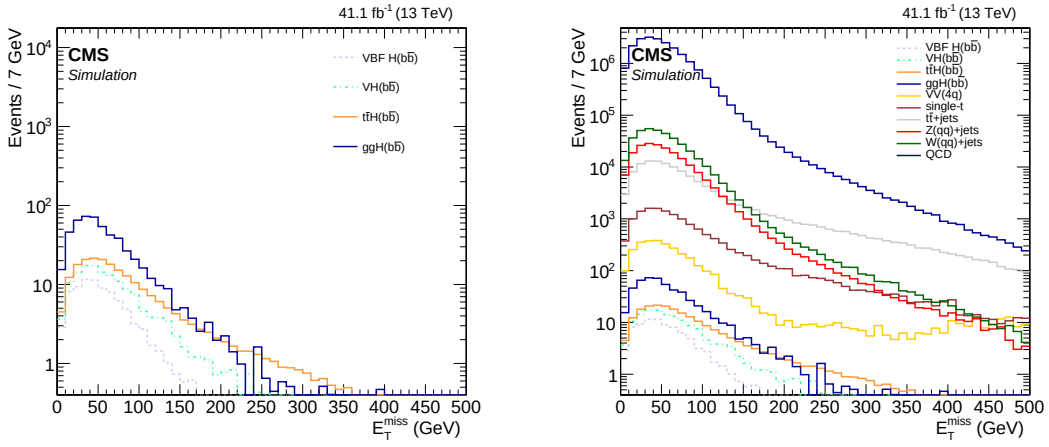


Figure 4.24: $E_{\mathrm{T}}^{\mathrm{miss}}$ distributions. Left, simulated signal events normalized to their cross sections (Table 4.4). Right, background and signal events normalized to their cross sections.

4.5 Event selections

To keep the search as inclusive as possible, the number of selection is kept to a minimal. With a relatively simple signal topology, the event selection can be summarized as follows:

- **Trigger/dataset selection**
- **Kinematic selection (Jets):** Require a high p_{T} AK8 PUPPI jet ($p_{\mathrm{T}} > 450\,\mathrm{GeV}$, $m_{\mathrm{SD}} > 40\,\mathrm{GeV}$, $|\eta| < 2.5$, and $-6.0 < \rho < -2.1$) - we select the leading p_{T} jet.
- **Lepton veto:** Veto the presence of identified "loose" electrons, muons, and hadronic taus.
- **Top rejection:** Require $p_{\mathrm{T}}^{\mathrm{miss}} < 140\,\mathrm{GeV}$ and veto a medium DeepCSV AK4

b-jet opposite the AK8 jet to further reduce the top contamination.

- **Substructure selection:** Require the selected jet to satisfy the 2-prong hypothesis by asking $N_2^{1,\mathrm{DDT}} < 0$ for 26% quantile.

In this section, we discuss the optimization of selection on the signal candidate jets, which will be studied in detail with the background estimation method. We will also look at the data-MC agreement at the pre-selection level, which is defined as the jet p_{T} and m_{SD} selection.

4.5.1 Higgs candidate jet selection

The leading p_{T} jet is taken to be the boson candidate jet. As shown in Fig. 4.25, this selection method correctly identifies the Higgs boson in simulation about 60% of the time. In the previous iteration of this analysis, a study of an alternative boson candidate selection strategy based on the tagger score was performed [1], where no substantial gain in significance was found. A simple calculation corroborates this story: if the leading jet by discriminator score is chosen, this promotes QCD background by offering a given QCD event two opportunities to pass, so $\epsilon_{qcd} = 1 - (1 - \epsilon_{fake})^2 \approx 2\epsilon_{fake}$, where the approximation holds as we use a discriminator cut with 1% fake rate. Meanwhile, we gain at most the other 40% of the Higgs boson signal events where the subleading jet by p_{T} is the true boson jet, leading to an maximum gain in $S/\sqrt{B}$ of factor $(1/0.6)/\sqrt{2} = 1.17$. This small improvement is not deemed worthwhile in the face of possible complications due to discriminator sculpting and cross-trigger efficiency uncertainties.

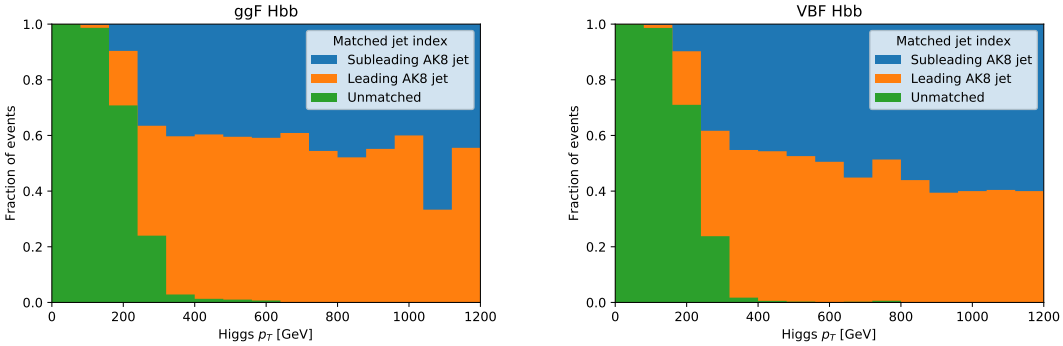


Figure 4.25: Fraction of $H \to b\bar{b}$ events where the generated Higgs boson was found within $\Delta R < 0.8$ of either the leading reconstructed AK8 jet, the subleading jet, or not matched to any AK8 jet in the event, as a function of generated Higgs boson p_T. The fractions are shown separately for gluon fusion H production with the MINLO generator (left) and for VBF H production (right). Gluon fusion is the dominant production mode in the $p_T > 450\,\text{GeV}$ region, followed by VBF production.

4.5.2 Substructure and deep double-b optimization

To ensure the substructure variable is decorrelated from jet p_T and mass, we use $N_2^{1,\text{DDT}}$ instead of N_2 (as defined in Sec. 4.4.4). The substructure selection, based on $N_2^{1,\text{DDT}}$, has been optimized by maximizing $S/\sqrt{B}$ in a simultaneous scan of substructure and deep double-b cut values after kinematic selection. The results of this optimization are shown in Fig. 4.26 and Table 4.6.

We find the optimal substructure ($N_2^{1,\text{DDT}}$) cut to be around 0.38 – 1 and a deep double-b cut value around 0.86 – 0.90. As the optimal $N_2^{1,\text{DDT}}$ quantile does not perform significantly better than the previously used value, we keep a $N_2^{1,\text{DDT}}$ quantile of 0.26 and a deep double-b cut corresponding to the M2 working point of 0.89 to aid in comparison to previous results.

In Fig. 4.27 the efficiency of each of the requirement described above is reported

Table 4.6: Relative $S/\sqrt{B}$ values (normalized to the best $S/\sqrt{B}$) for different $N_2^{1,\mathrm{DDT}}$ and deep double-b cuts.

$N_2^{1,\mathrm{DDT}}$ cut	deep double-b cut	relative $S/\sqrt{B}$
26%	0.89	0.94
38%	0.89	0.97
100%	0.89	0.96

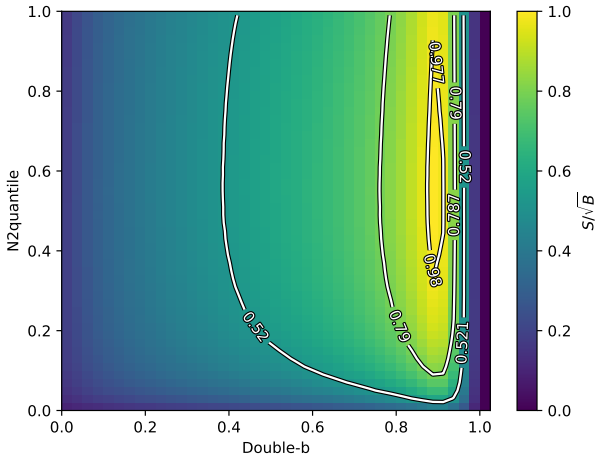


Figure 4.26: Scan of relative $S/\sqrt{B}$ for the deep double-b tagger and $N_2^{1,\mathrm{DDT}}$ substructure cuts, using simulated multijet QCD and ggF H($b\bar{b}$) signal events. The white contours show the value of $S/\sqrt{B}$. The maximum occurs near deep double-b values of 0.86 – 0.90 and $N_2^{1,\mathrm{DDT}}$ quantiles of 0.38 – 1

for simulated signal and background events. The expected total simulated signal and background yields remaining after each requirement is shown for 2017 conditions in Table 4.7.

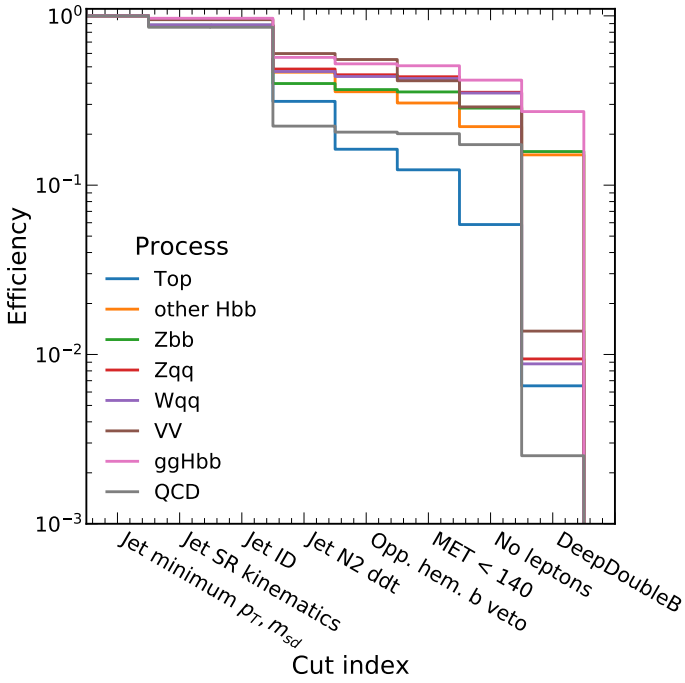


Figure 4.27: Efficiency of each of the requirement of the kinematic and top rejection selections, for simulated signal and background events.

After all the selection criteria have been satisfied we split the events in two categories depending whether the jet has deep double-b tagger> 0.89 or not (passing/failing). The signal extraction will be performed simultaneously in both regions, as described in more detailed in the next section.

In Fig. 4.28 and Fig. 4.29 the m_{SD} distribution is shown for simulated signal and background events after all the selection criteria described in both the passing and failing regions. Although the gluon fusion process dominates, the other Higgs

Table 4.7: Expected total simulated signal and background yields remaining after each of the jet- and event-level selections applied in this analysis. The jet signal region kinematic selection is $450 \leq p_{\mathrm{T}} < 1200\,\mathrm{GeV}$, $47 \leq m_{\mathrm{SD}} < 201\,\mathrm{GeV}$. Yields correspond to the 2017 data-taking conditions and integrated luminosity of $41.1\,\mathrm{fb}^{-1}$.

	ggHbb	other Hbb	Zbb	VV	Top	Zqq	Wqq	QCD
Jet minimum p_{T}, m_{sd}	154	210	11543	3732	250584	64605	155485	24045691
Jet SR kinematics	149	184	10032	3542	221018	56804	137330	20588794
Jet ID	149	184	10032	3541	221006	56799	137316	20587152
Jet N2 ddt	87	98	4598	2235	78316	31347	72950	5368046
Opp. hem. b veto	80	75	4235	2061	40845	28991	68136	4945930
MET $<$ 140 GeV	78	64	4102	1543	30877	28224	66297	4841168
No leptons	64	47	3292	1082	14666	22854	54418	4173588
DeepDoubleB	42	32	1823	51	1636	608	1368	60676

production mechanisms also contribute. In the signal extraction, all contributions are taken into account.

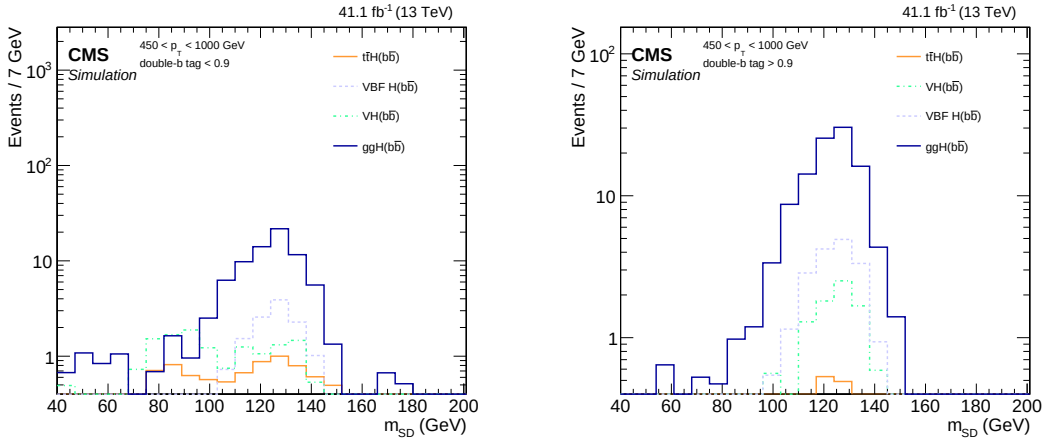


Figure 4.28: m_{SD} distribution for simulated signal events after all the selection criteria in both the failing (left) and passing (right) b-tagging categories. Each process is normalized to its cross section as shown in Table 4.4

.

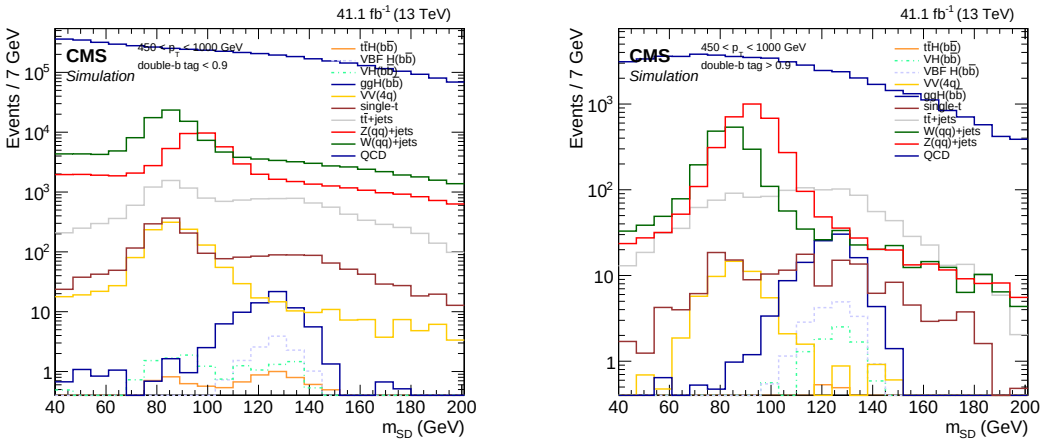


Figure 4.29: Signal region m_{SD} distribution for simulated signal and background events after all the selection criteria in both the failing (left) and passing (right) deep double-b tagging categories.

4.5.3 Data and simulation comparison

In this section in Fig. 4.30 to 4.32, we show distributions of variables in data and simulated samples after the kinematic selection. The main background is composed of multijet production followed by $t\bar{t}$ and other electroweak backgrounds. Note that the QCD background estimation method is data-driven, the method does not rely on exact data-MC agreement. The QCD and EW corrections to the V+jets samples are also included.

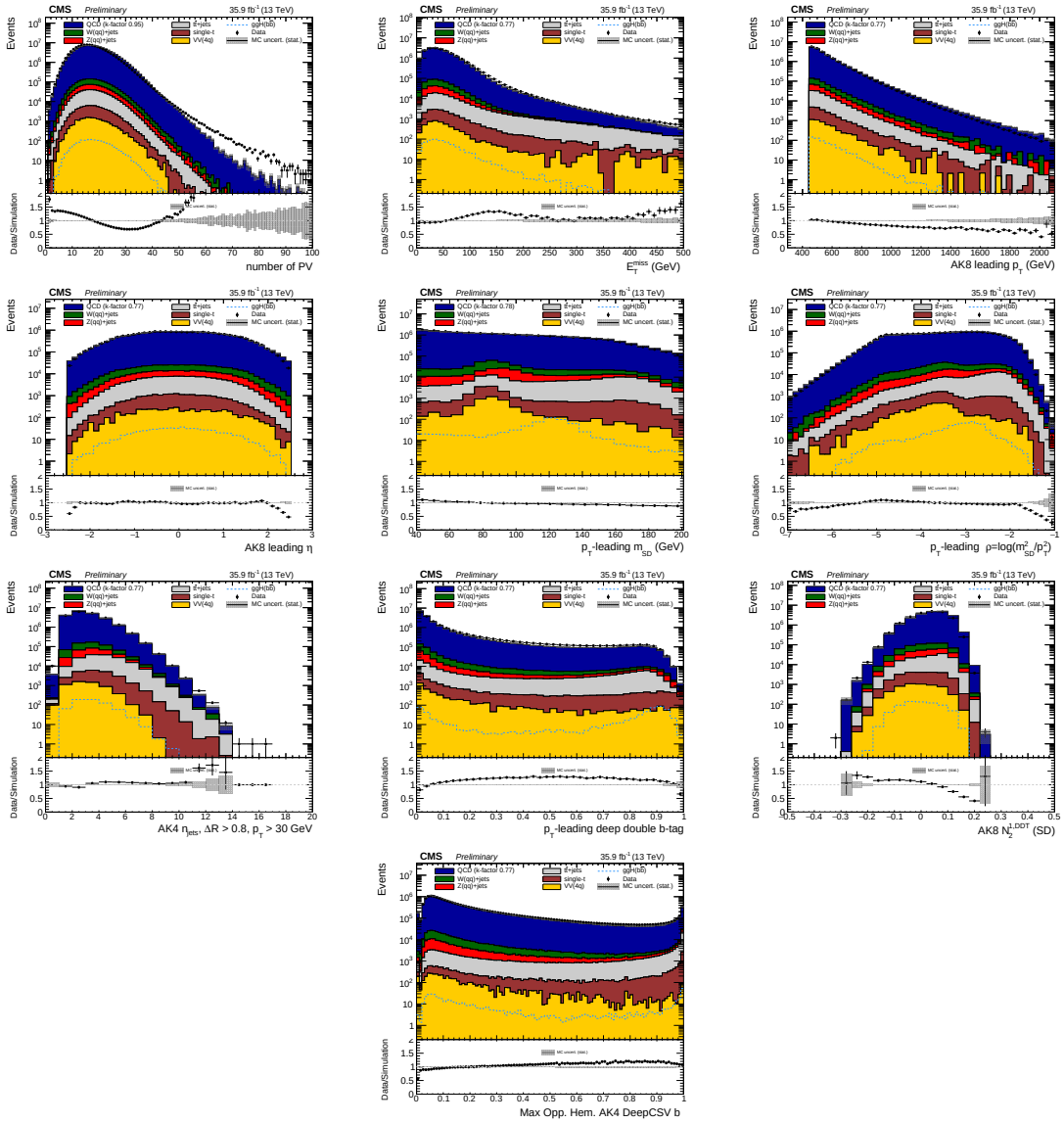


Figure 4.30: Distributions of variables in 2016 data and simulated samples after the kinematic preselection. From top to bottom: number of primary vertices, $E_{\mathrm{T}}^{\mathrm{miss}}$, p_{T}, η, m_{SD}, deep double-b tagger, $N_2^{1,\mathrm{DDT}}$, and the $\mathrm{t\bar{t}}$ rejection variable. The $VV(4q)$ category is mislabeled, and represents instead inclusive diboson production (yellow). The background processes, QCD(blue), W+jets(green), Z+jets(red), $\mathrm{t\bar{t}}$(grey), single-top(brown) and diboson(yellow), are shown as stacked distributions.

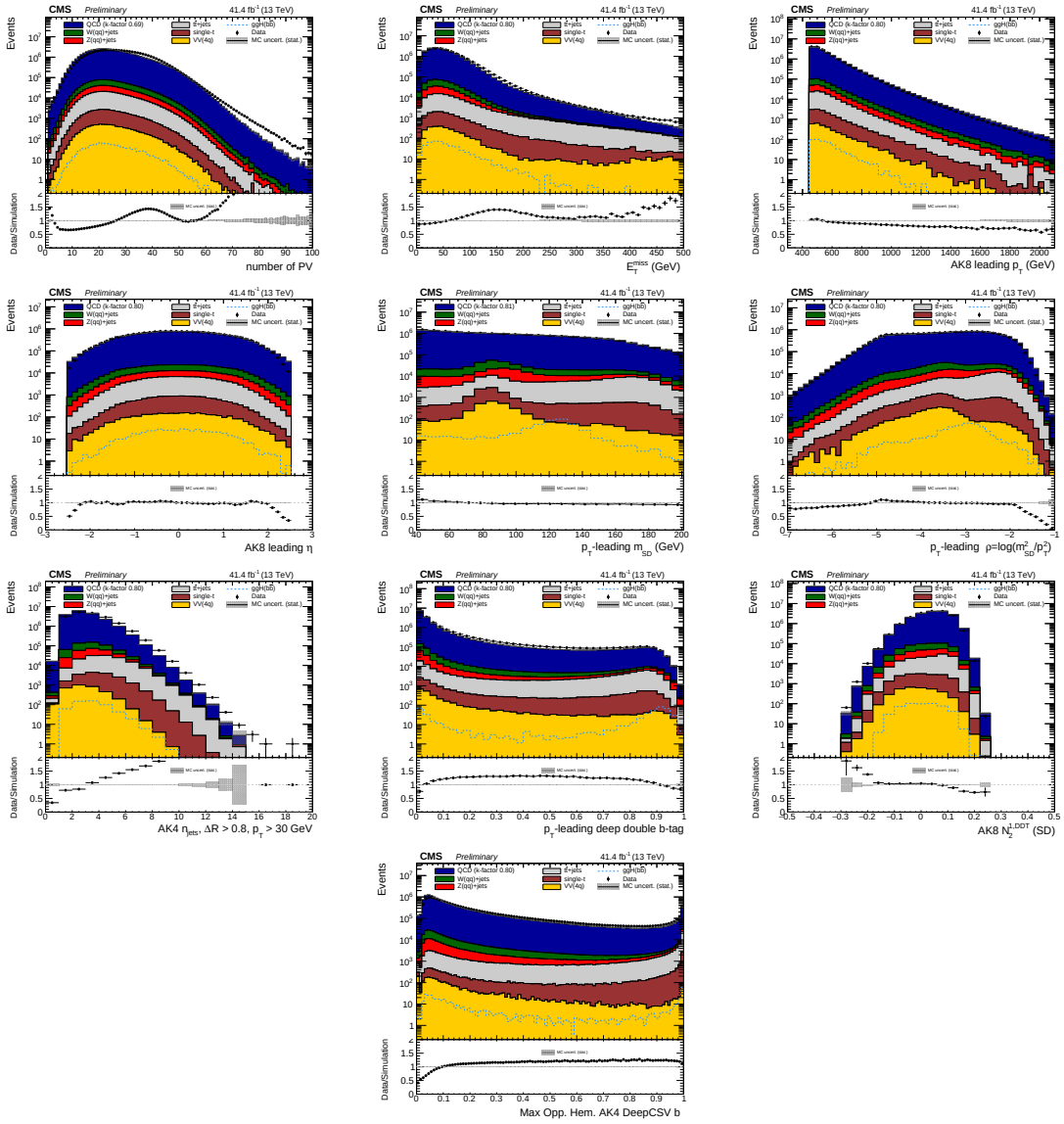


Figure 4.31: Distributions of variables in 2017 data and simulated samples after the kinematic preselection. From top to bottom: number of primary vertices, $E_{\mathrm{T}}^{\mathrm{miss}}$, p_{T}, η, m_{SD}, deep double-b tagger, $N_2^{1,\mathrm{DDT}}$, and the $\mathrm{t\bar{t}}$ rejection variable. The $VV(4q)$ category is mislabeled, and represents instead inclusive diboson production. The background processes, QCD(blue), W+jets(green), Z+jets(red), $\mathrm{t\bar{t}}$(grey), single-top(brown) and diboson(yellow), are shown as stacked distributions.

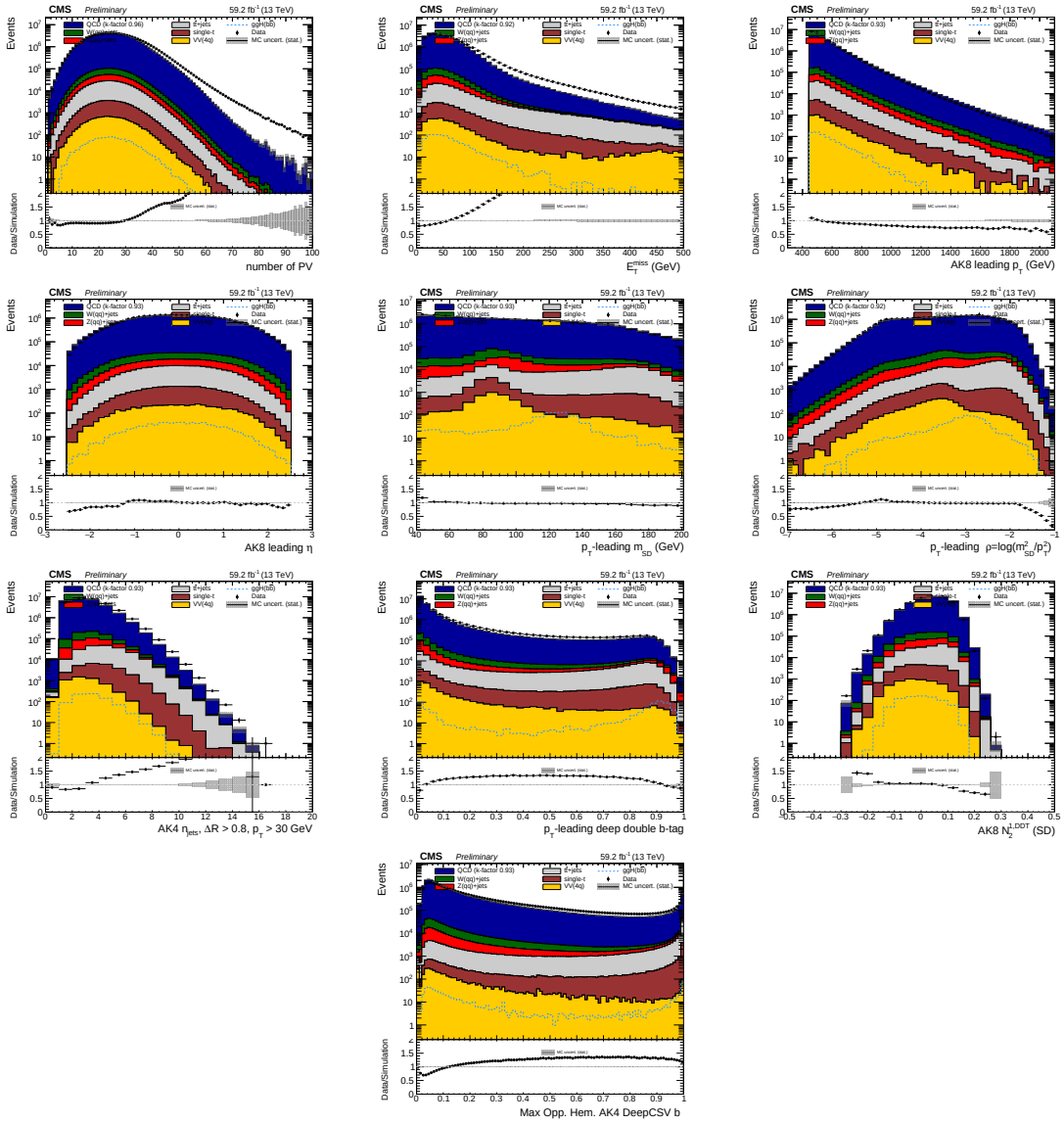


Figure 4.32: Distributions of variables in 2018 data and simulated samples after the kinematic preselection. From top to bottom: number of primary vertices, $E_{\mathrm{T}}^{\mathrm{miss}}$, p_{T}, η, m_{SD}, deep double-b tagger, $N_2^{1,\mathrm{DDT}}$, and the $\mathrm{t\bar{t}}$ rejection variable. The $VV(4q)$ category is mislabeled, and represents instead inclusive diboson production. The background processes, QCD(blue), W+jets(green), Z+jets(red), $\mathrm{t\bar{t}}$(grey), single-top(brown) and diboson(yellow), are shown as stacked distributions.

4.6 Background estimation

In this section, we will describe the methods used to estimate the SM background processes. After the event selections listed in Sec. 4.5, we categorize the events based on the Higgs candidate jets into 6 p_{T} bins ranging from 450 to 1200 GeV and 22 m_{SD} bins between 47 and 201 GeV. The signal region (or the "pass" region) is defined by the events that passes the deep double-b (DDB) tagger threshold of 0.89 – a working point optimized as discussed in Sec.4.5.2. The events that fail DDB threshold but pass the rest of the selections define a control region referred to as the "fail" region. In the ideal case that the tagger is completely decorrelated with p_{T} and m_{SD} for QCD jets, the shape of the QCD m_{SD} spectrum in the "fail" region would be identical to that of the "pass" region. In Sec 4.6.3, we will describe how we model this residual correlation with a transfer factor to predict the QCD m_{SD} spectrum in the pass region. In addition, resonant background estimation (V+jets and $\mathrm{t\bar{t}}$) on top the QCD m_{SD} spectrum will be discussed.

4.6.1 V+jets estimate

The W/Z+jets processes are estimated from simulation that have higher order corrections (NLO QCD and NLO EW) included, as discussed in Sec. 4.2.2. As they are resonance backgrounds on top of the smooth falling QCD m_{SD} spectrum, they are treated similarly as the signal processes. In particular, the m_{SD} mass scale, resolution and $N_2^{1,\mathrm{DDT}}$ need to be validated in data. The m_{SD} correction with respect to simulations has been discussed in Sec. 4.4.2. The final correction to match simulation to

data is derived with a dedicated semileptonic $t\bar{t}$ data sample.

To select a semileptonic $t\bar{t}$ enriched sample in data, events are required to have an energetic muon with $p_T > 100\,\text{GeV}$, $p_T^{\text{miss}} > 80\,\text{GeV}$, a high-$p_T$ AK8 jet with $p_T > 200\,\text{GeV}$, and an additional b-tagged AK4 jet separated from the AK8 jet by $\Delta R > 0.8$ with $p_T > 30\,\text{GeV}$. In this sample, the high-p_T AK8 jet is enriched in merged hadronically decaying W boson. We then divide the sample into events that pass and fail the $N_2^{1,\text{DDT}}$ selection for merged W boson jets in data and simulation. A simultaneous fit to the two samples is performed in order to extract the selection efficiency of a merged W jet in simulation and in data. The W mass peak in data calibrates the m_{SD} scale and resolution in simulation, whereas the ratio between passing and failing the $N_2^{1,\text{DDT}}$ selection gives a data-mc scale factor for the substructure selection. A schematic drawing of the effect of these correction factors on the W and Z mass peak is shown in Fig. 4.33. Typical correction sizes for jet mass scale is a few GeV and $\sim$ up to 10% for mass resolution and $N_2^{1,\text{DDT}}$ SF.

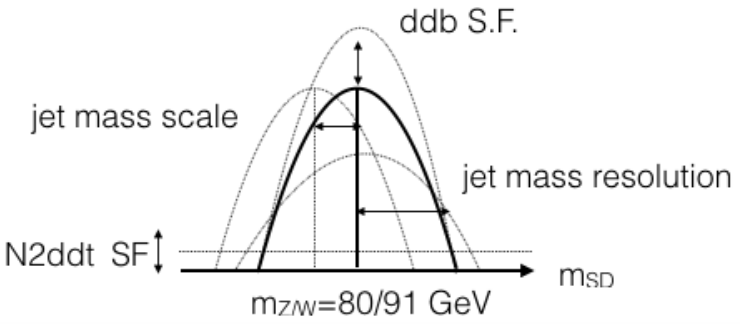


Figure 4.33: Schematic drawing of the effect of data-MC corrections factors applied on W/Z+jets backgrounds.

Uncertainties in these scale factors and corrections are discussed in Sec. 4.8.2.

These uncertainties are treated as nuisance parameters, which are profiled in the final fit to extract the Higgs signal. Due to the presence of W/Z+jets in the final mass distribution, these post-fit uncertainties are further constrained *in situ*.

4.6.2 $t\bar{t}$ estimate

The $t\bar{t}$ background contribution is at the few percent level after all selections. Similar to W and Z+jets backgrounds, it is estimated from MC with corrections based on the data comparisons. To provide a robust estimate of the $t\bar{t}$ background, we use a single muon control region to accomplish two goals:

1. validate the agreement between data and MC simulation of the shape of the $t\bar{t}$ background

2. derive a data-driven estimate of the $t\bar{t}$ normalization in the phase space of the analysis (AK8 jet $p_T > 400\,\text{GeV}$, $m_{SD} > 40\,\text{GeV}$) separately for the passing (deep double-b tag > 0.89) and failing (deep double-b tag < 0.89) regions. Note, we lower the p_T requirement from 450 to 400 GeV to increase the number of events in the single muon control region.

Since the single muon control region aims at probing $t\bar{t}$ contribution in the signal region, the selection is identical to the signal region selection, except for essential changes to enrich $t\bar{t}$ contributions, i.e. the different trigger/dataset selection, the addition of the muon selection, the additional b-tag requirement, and the removal of the top rejection criteria.

Events in the single muon control region are selected as follows:

- **Trigger/dataset selection:** Events from the Single Muon primary dataset (as specified in Sec. 4.2) passing the `HLT_Mu50_v* Or HLT_TkMu50_v*` trigger.
- **Muon selection:** At least one "loose" muon with $p_{\mathrm{T}} > 55\,\mathrm{GeV}$, $|\eta| < 2.1$, and $\Delta\phi(\mu, \mathrm{AK8\ jet}) > 2\pi/3$ (i.e. the muon is required to be in the opposite "hemisphere" from the selected AK8 jet).
- **Additional b-tag selection:** At least one AK4 PUPPI jet satisfying the "medium" CSVv2 b-tag, with $p_{\mathrm{T}} > 50\,\mathrm{GeV}$, $|\eta| < 2.5$, and $\Delta R(\mathrm{AK4\ b-tag, AK8\ jet}) > 0.8$ (i.e. the b-tag is requried to be outside of the cone of the selected AK8 jet).
- **Electron and tau veto:** Veto the presence of identified "loose" electrons and "loose" hadronic taus.
- **Kinematic selection (Jets):** Require a high p_{T} AK8 PUPPI jet ($p_{\mathrm{T}} > 400\,\mathrm{GeV}$, $m_{\mathrm{SD}} > 40\,\mathrm{GeV}$ and $|\eta| < 2.4$) – we select the leading p_{T} jet.
- **Substructure selection :** Require the selected AK8 jet to satisfy the 2-prong hypothesis by requiring $N_2^{1,\mathrm{DDT}} < 0$.

The semi-leptonic $\mathrm{t\bar{t}}$ sample we used to derive m_{SD} and $N_2^{1,\mathrm{DDT}}$ correction factors is very similar to this single-muon control region, except that the AK8 jet p_{T} is lowered to achieve better statistical precision for the scale-factor measurements.

The properties of the muon (p_{T} and η) are shown Fig. 4.34, while the properties of the AK8 PUPPI jet (p_{T}, η, and deep double-b tag) are shown in Fig. 4.35. We find satisfactory agreement between data and MC shapes in all the distributions.

Fig. 4.36- 4.38 shows the distribution of m_{SD} for the leading p_{T} AK8 PUPPI jet in the single muon control region, inclusive in deep double-b tag discriminator value(top), events with deep double-b tag < 0.89 (bottom left) and events with deep double-b tag > 0.89 (bottom right) in 2016,2017 and 2018 data respectively. We can see that the m_{SD} shape of the $\mathrm{t\bar{t}}$ background in data is reasonably reproduced in the simulation.

Since we apply a substructure requirement to the AK8 jet, we remove most of the top contribution coming from the 3-prongs (i.e. cases where the top is fully merged into an AK8 jet) and indeed the peak around the top mass is not seen in Fig. 4.36- 4.38.

We derive two separate data/MC scale factors using the signal muon control region: an overall $\mathrm{t\bar{t}}$ normalization scale factor $\mathrm{SF}^{\mathrm{t\bar{t}}}_{\mathrm{norm}}$, and an efficiency scale factor for the deep double-b tag selection (deep double-b tag > 0.89) $\mathrm{SF}^{\mathrm{top}}_{\mathrm{mis-tag}}$. We compute the normalization scale factor $\mathrm{SF}_{\mathrm{t\bar{t}}}$ by performing a maximum likelihood fit, after which we have

$$\mathrm{SF}^{\mathrm{t\bar{t}}}_{\mathrm{norm}} \approx \frac{N_{\mathrm{data}} - N_{\mathrm{other\ SM}}}{N_{\mathrm{t\bar{t}}}} \ , \tag{4.4}$$

where $N_{\mathrm{t\bar{t}}}$ is the expected yield from the $\mathrm{t\bar{t}}$ MC sample and $N_{\mathrm{other\ SM}}$ is the expected yield from the remaining MC samples in the single muon control region. Similarly, we compute the efficiency scale factor for the deep double-b tag selection with the

Data	$SF^{t\bar{t}}_{norm}$	$SF^{top}_{mis-tag}$
2016	0.73 ± 0.05	1.14 ± 0.15
2017	0.76 ± 0.05	0.80 ± 0.11
2018	0.86 ± 0.09	0.99 ± 0.10

Table 4.8: Summary of $t\bar{t}$ scale factors measured in single muon control region for different data-taking years.

same fit, after which scale factor approximately satisfy

$$\epsilon^{\text{top mis-tag}}_{\text{data}} = \frac{N^{\text{pass}}_{\text{data}} - N^{\text{pass}}_{\text{other SM}}}{N_{\text{data}} - N_{\text{other SM}}} \tag{4.5}$$

$$\epsilon^{\text{top mis-tag}}_{\text{MC}} = \frac{N^{\text{pass}}_{t\bar{t}}}{N_{t\bar{t}}} \tag{4.6}$$

$$SF^{\text{top}}_{\text{mis-tag}} \approx \frac{\epsilon^{\text{top mis-tag}}_{\text{data}}}{\epsilon^{\text{top mis-tag}}_{\text{MC}}} \ , \tag{4.7}$$

where $N^{\text{pass}}_{\text{data}}$, $N^{\text{pass}}_{t\bar{t}}$, and $N^{\text{pass}}_{\text{other SM}}$ are the yields in the pass region (deep double-b tag > 0.89) of the single muon control region.

We also compute the uncertainties in the scale factors by propagating the statistical uncertainties from the data and MC yields (based on sum of squared weights for MC). Table 4.8 summarizes the scale factors and uncertainties measured for all three data taking years.

We treat the single muon control region as a counting experiment, as we found that the MC shapes with low statistics tend to de-stablilze the final fit combined with signal region. Integrating contribution from small processes over m_{SD} provides a easier treatment of the uncertainty from limited MC statistics. The muon control region simultaneously takes into account additional systematic uncertainties related to luminosity and other MC reweighting as well as any possible Higgs signal contamination

in the control region.

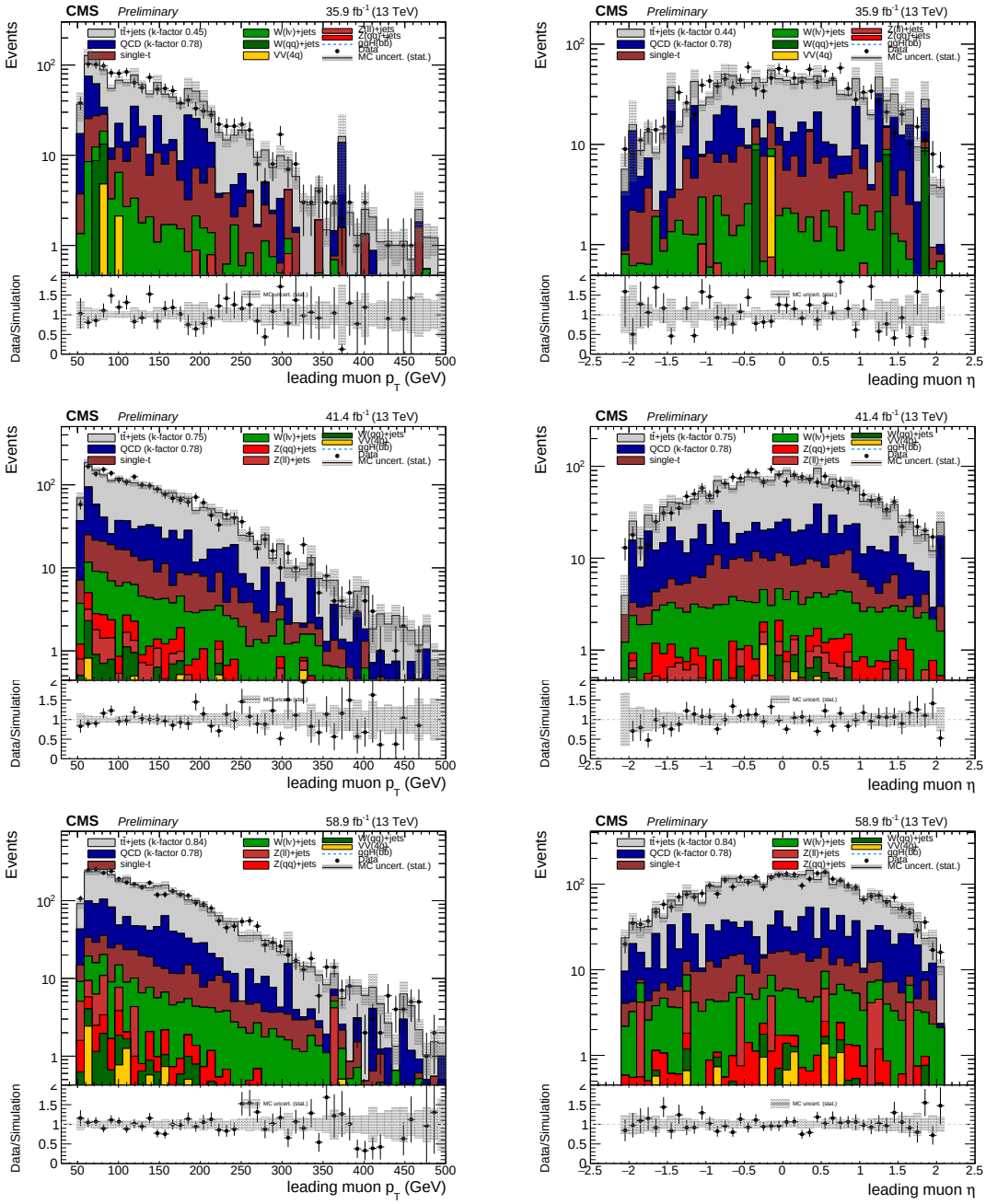


Figure 4.34: Muon p_{T} (left) and η (right) in the single muon control region for $\mathrm{t\bar{t}}$ background in 2016 data (top), 2017 data (middle), and 2018 data (bottom) respectively. QCD is scaled down by a k-factor of 0.78 and $\mathrm{t\bar{t}}$ is scaled to achieve the same overall background integral as data.

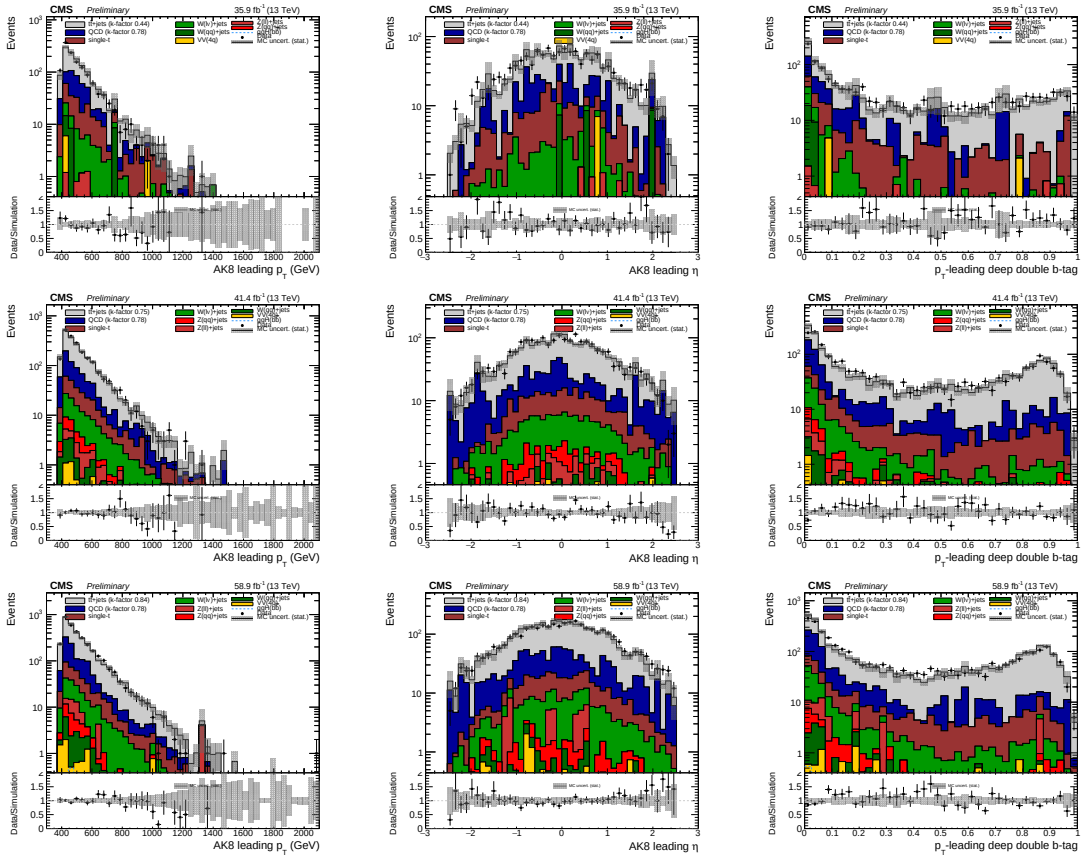


Figure 4.35: AK8 Puppi jet p_{T} (left), η (middle), and deep double-b tag (right) in the single muon control region for $\mathrm{t\bar{t}}$ background in 2016 data (top), 2017 data (middle) and 2018 data (bottom) respectively. QCD is scaled down by a k-factor of 0.78 and $\mathrm{t\bar{t}}$ is scaled to same integral as data.

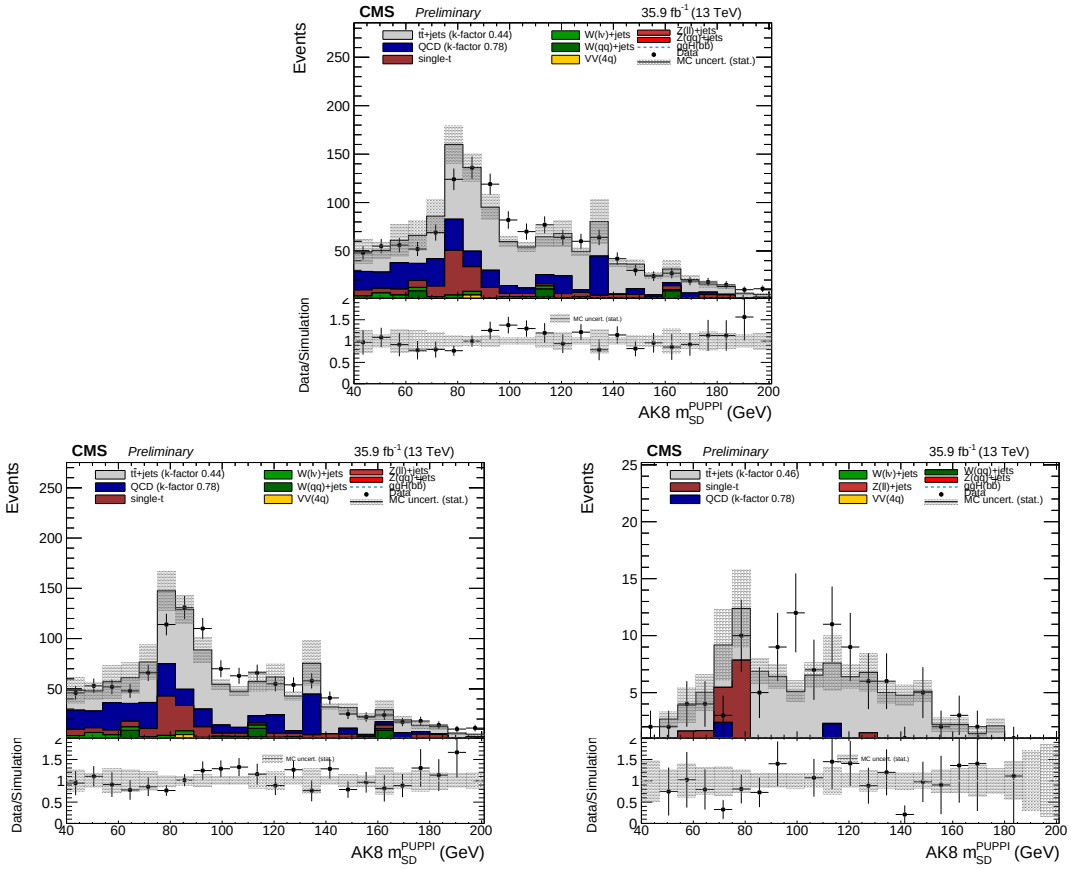


Figure 4.36: Single muon control region for $t\bar{t}$ background in 2016 data. (Top) Inclusive in deep double-b tag. (Bottom left) Failing region with deep double-b tag < 0.89. (Bottom right) Passing region with deep double-b tag > 0.89. QCD is scaled down by a k-factor of 0.78 and $t\bar{t}$ is scaled to same integral as data.

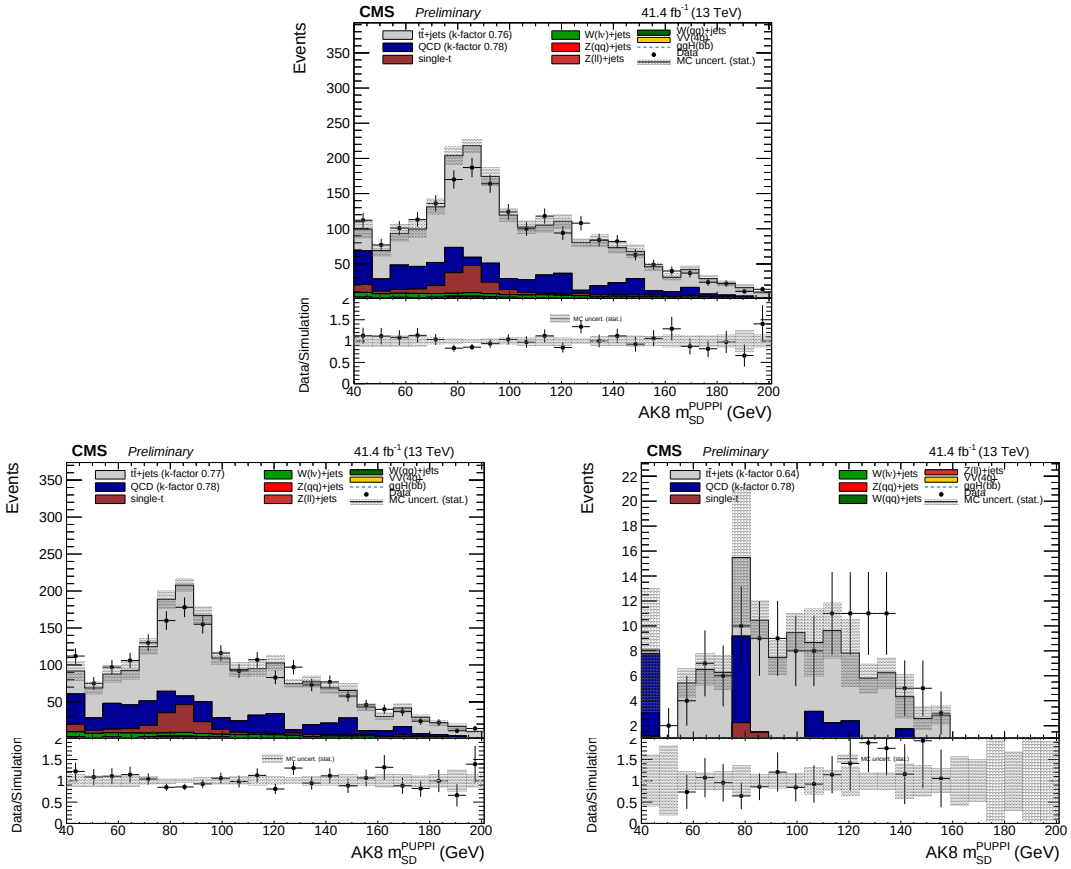


Figure 4.37: Single muon control region for $t\bar{t}$ background in 2017 data. (Top) Inclusive in deep double-b tag. (Bottom left) Failing region with deep double-b tag < 0.89. (Bottom right) Passing region with deep double-b tag > 0.89. QCD is scaled down by a k-factor of 0.78 and $t\bar{t}$ is scaled to same integral as data.

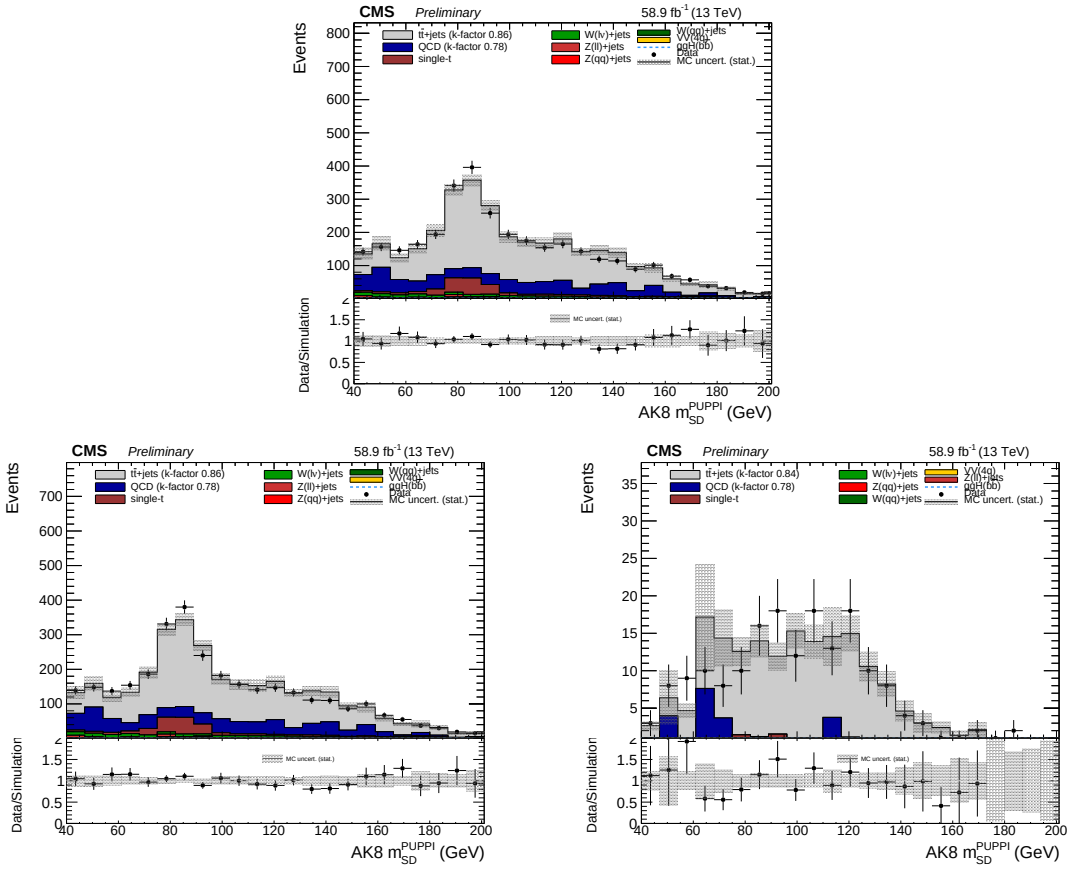


Figure 4.38: Single muon control region for $t\bar{t}$ background in 2018 data. (Top) Inclusive in deep double-b tag. (Bottom left) Failing region with deep double-b tag < 0.89. (Bottom right) Passing region with deep double-b tag > 0.89. QCD is scaled down by a k-factor of 0.78 and $t\bar{t}$ is scaled to same integral as data.

4.6.3 QCD multijet estimation

The main idea of the background estimation technique in this analysis is to predict the QCD background jet mass shape from a signal depleted region defined by failing the deep double-b tag discriminator cut. Specifically, we specify a "pass-to-fail ratio" $R_{\mathrm{p/f}}$, which would be a flat ratio across jet mass and p_{T} for an ideal tagger and perfect simulation. As we show in the QCD MC (see Fig. 4.39), the deep double-b tagger discriminant has residual correlation with jet mass and p_{T}, despite the efforts to decorrelation in the design process. However, due to imperfections in the decorrelation method, there is an anticorrelation at high tagger discriminator values and low jet mass. To account for this, the expected pass-fail ratio as a function of ρ and p_{T}, $\epsilon^{\mathrm{QCD}}(\rho, p_{\mathrm{T}})$, is estimated by fitting QCD multijet simulated events. As discussed in Sec. 4.4.3, the QCD jet mass scales with p_{T}, and thus a more natural variable to consider is the QCD scaling variable ρ, which should have a roughly invariant shape as a function of p_{T}.

The remaining differences are thus due to discrepancies between the data and simulation $R^{\mathrm{data}}_{\mathrm{p/f}}$, which we parametrize using a Berinstein polynomial in ρ and p_{T}. Then, the complete pass-fail ratio in data is given by the product of these two distinct factors.

Explicitly, we define the transfer factor $R_{\mathrm{p/f}}$ to be a function of ρ and p_{T},

$$p^{\mathrm{QCD}}_{\mathrm{pass}}(m_{\mathrm{SD}}, p_{\mathrm{T}}) \propto R_{\mathrm{p/f}}(\rho, p_{\mathrm{T}}) \times p^{\mathrm{QCD}}_{\mathrm{fail}}(m_{\mathrm{SD}}, p_{\mathrm{T}}) \ , \tag{4.8}$$

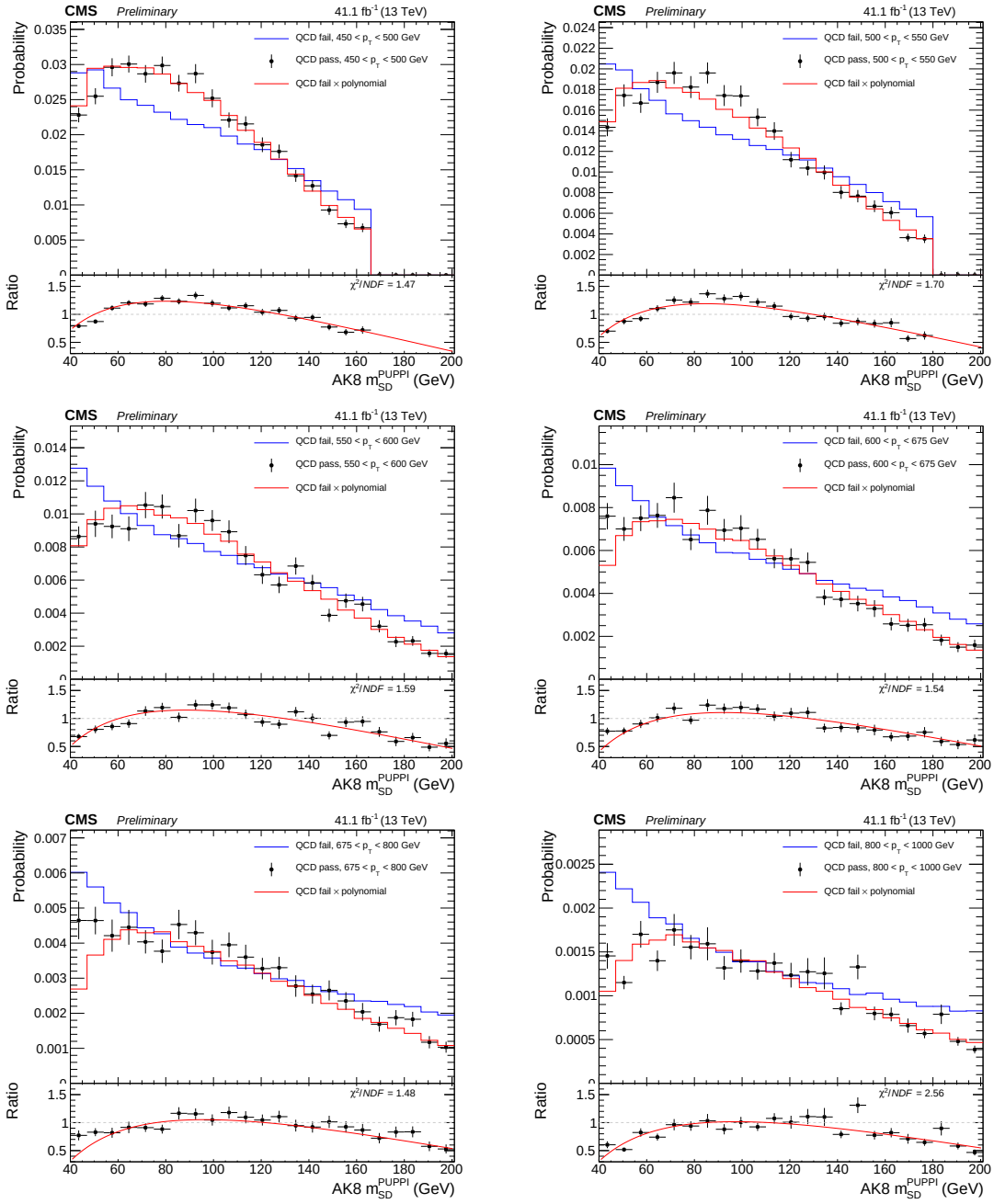


Figure 4.39: m_{SD} distributions of the QCD MC in the pass and fail regions (top panel) and their ratio (bottom panel) in each p_{T} category. This polynomial TF is used to derive the QCD prediction (shown in red) for the pass region (defined as the fail region distribution times the polynomial). The χ^2 reported here refers to the compatibility between shapes in the pass and fail regions.

where $p_{\text{pass}}^{\text{QCD}}(m_{\text{SD}}, p_{\text{T}})$ are the $p_{\text{fail}}^{\text{QCD}}(m_{\text{SD}}, p_{\text{T}})$ QCD probability density functions (modeled as binned two-dimensional templates) normalized to unity. Since the QCD mass shape is not trivial (see Fig. 4.39), the estimation of QCD prediction is performed in two steps:

1. Extract the pass-to-fail ratio in QCD MC events, i.e. $\epsilon^{\text{QCD,MC}}(\rho, p_{\text{T}})$. This ratio describes the expected degree of mass sculpting, which is the difference in the shape of m_{SD} distribution due to deep double-b selection, from QCD MC events.

2. Fit the residual difference of pass-to-fail ratios between MC and data with a smooth transfer factor $R_{\text{p/f}}$ parametrized by a Bernstein polynomial in ρ and p_{T}.

These two functions $\epsilon^{\text{QCD,MC}}(\rho, p_{\text{T}})$ and $R_{\text{p/f}}(\rho, p_{\text{T}})$ are estimated in two separate fits. In both fits, we parametrize $R_{\text{p/f}}$ and $\epsilon^{\text{QCD,MC}}$ as Bernstein polynomials in ρ and p_{T},

$$R_{\text{p/f}}(\rho, p_{\text{T}}) = \epsilon^{\text{QCD,MC}}(\rho, p_{\text{T}}) \sum_{k=0}^{n_\rho} \sum_{\ell=0}^{n_{p_{\text{T}}}} a_{k,\ell} \left[b_{k,n_\rho}(\rho) b_{\ell,n_{p_{\text{T}}}}(p_{\text{T}}) \right], \tag{4.9}$$

$$= \epsilon^{\text{QCD,MC}}(\rho, p_{\text{T}}) \times TF^{\text{data}}(\rho, p_{\text{T}}), \tag{4.10}$$

$$\epsilon^{\text{QCD,MC}}(\rho, p_{\text{T}}) = \overline{\epsilon^{\text{QCD,MC}}} \sum_{k=0}^{n_\rho} \sum_{\ell=0}^{n_{p_{\text{T}}}} c_{k,\ell} \left[b_{k,n_\rho}(\rho) b_{\ell,n_{p_{\text{T}}}}(p_{\text{T}}) \right], \tag{4.11}$$

$$= \overline{\epsilon^{\text{QCD,MC}}} \times TF^{\text{QCD}}(\rho, p_{\text{T}}) \tag{4.12}$$

where n_ρ is the degree of the polynomial in ρ, $n_{p_{\text{T}}}$ is the degree of the polynomial in

p_{T}, $\overline{\epsilon^{\mathrm{QCD,MC}}}$ is the average QCD efficiency in the whole analysis phase space, $a_{k,\ell}, c_{k,\ell}$ are Bernstein coefficients, and

$$b_{\nu,n}(x) = \binom{n}{\nu} x^{\nu} (1-x)^{n-\nu} \tag{4.13}$$

is a Bernstein basis polynomial of degree n. The 2D-Bernstein polynomial functions constructed from the Bernstein coefficients in Eq.4.9 and Eq.4.11 are called transfer functions(TF).

$$TF^{\mathrm{data}}(\rho, p_{\mathrm{T}}) = \sum_{k=0}^{n_\rho} \sum_{\ell=0}^{n_{p_{\mathrm{T}}}} a_{k,\ell} \left[b_{k,n_\rho}(\rho) b_{\ell,n_{p_{\mathrm{T}}}}(p_{\mathrm{T}}) \right], \tag{4.14}$$

$$TF^{\mathrm{QCD}}(\rho, p_{\mathrm{T}}) = \sum_{k=0}^{n_\rho} \sum_{\ell=0}^{n_{p_{\mathrm{T}}}} c_{k,\ell} \left[b_{k,n_\rho}(\rho) b_{\ell,n_{p_{\mathrm{T}}}}(p_{\mathrm{T}}) \right], \tag{4.15}$$

The number of required coefficients in the QCD MC fit(Eq.4.11), as well as in the final data residual fit(Eq.4.9), are determined by a Fisher F-test on data, as discussed further in Section 4.7.1.

Since $\overline{\epsilon^{\mathrm{QCD,MC}}}$ sets the overall scale ($\sim 1\%$), the coefficients $a_{k,\ell}, c_{k,\ell}$are of the order 1. Ideally, if the deep double-b tagger discriminant value were completely uncorrelated from the jet p_{T} and m_{SD}, $\epsilon^{\mathrm{QCD,MC}}(\rho, p_{\mathrm{T}})$ would be a constant and $R_{\mathrm{p/f}}(\rho, p_{\mathrm{T}})$ would be flat.

With $R_{\mathrm{p/f}}(\rho, p_{\mathrm{T}})$ and $\epsilon^{\mathrm{QCD,MC}}(\rho, p_{\mathrm{T}})$, the QCD contribution in the passing region

($N_{\text{pass}}^{\text{QCD}}$) is derived from the QCD contribution in the failing region ($N_{\text{fail}}^{\text{QCD}}$) as follows,

$$N_{\text{pass}}^{\text{QCD}}(m_{\text{SD}}, p_{\text{T}}) = R_{\text{p/f}}(\rho, p_{\text{T}}) \times N_{\text{fail}}^{\text{QCD}}(m_{\text{SD}}, p_{\text{T}}) \ , \tag{4.16}$$

$$= \epsilon^{\text{QCD,MC}}(\rho, p_{\text{T}}) \times TF^{\text{data}}(\rho, p_{\text{T}}) \times N_{\text{fail}}^{\text{QCD}}(m_{\text{SD}}, p_{\text{T}}) \tag{4.17}$$

Note that in the maximum likelihood fit performed on data, $\epsilon^{\text{QCD,MC}}(\rho, p_{\text{T}})$ are treated as constrained parameters with uncertainties assigned as described in section 4.8.1.

After the maximum likelihood fit, the QCD yields in the failing region approximately satisfy

$$N_{\text{fail}}^{\text{QCD}} \approx N_{\text{fail}}^{\text{data}} - N_{\text{fail}}^{V} - N_{\text{fail}}^{\text{t}\bar{\text{t}}} - N_{\text{fail}}^{\text{H}(\text{b}\bar{\text{b}})} \ . \tag{4.18}$$

That is, the QCD contribution in the failing region is essentially the data yield after subtracting the other SM background yields and the $\text{H}(\text{b}\bar{\text{b}})$ signal yields.

The final fit is performed simultaneously in the passing and failing regions. Contributions from Higgs signal processes (ggH, VBF, VH, $\text{t}\bar{\text{t}}$H), $\text{t}\bar{\text{t}}$ and the resonant electroweak background processes (W/Z+jets) are accounted for in both regions.

4.7 Full simultaneous fit

We perform a fit that simultaneously computes the 'pass-to-fail' ratio $R_{\text{p/f}}(\rho, p_{\text{T}})$, the two $\text{t}\bar{\text{t}}$-related scale factors ($\text{SF}_{\text{norm}}^{\text{t}\bar{\text{t}}}, \text{SF}_{\text{mis-tag}}^{\text{top}}$), and performs the signal extraction. We achieve this by

1. fitting the failing and passing signal regions, as well as the failing and the passing

single muon control region simultaneously,

2. binning in the m_{SD} and p_{T} of the jet in the signal region, while doing counting experiment in the single muon control region,

3. and allowing the QCD yield in each m_{SD} bin in the failing category to float with the simultaneous constraint from the passing region.

Explicitly, we bin the m_{SD} distribution in each p_{T} category. Then for a given m_{SD} and p_{T} bin, the QCD yield in the passing and failing regions are tied together through the transfer function $TF(\rho, p_{\mathrm{T}})$. That is, in the i^{th} m_{SD} bin, with central value m^i_{SD} and j^{th} p_{T} bin with value p^j_{T}, corresponding to the midpoint of the bin in logarithmic scale, ρ^{ij} is given by $\rho^{ij} = \log\big((m^i_{\mathrm{SD}})^2/(p^j_{\mathrm{T}})^2\big)$. Then the QCD yield in the i^{th} m_{SD} bin and j^{th} p_{T} bin of the passing region is given by,

$$N^{\mathrm{QCD}}_{\mathrm{pass}}(m^i_{\mathrm{SD}}, p^j_{\mathrm{T}}) = \epsilon^{\mathrm{QCD,MC}}(\rho^{ij}, p^j_{\mathrm{T}}) \times TF^{\mathrm{data}}(\rho^{ij}, p^j_{\mathrm{T}}) \times N^{\mathrm{QCD}}_{\mathrm{fail}}(m^i_{\mathrm{SD}}, p^j_{\mathrm{T}}) \ . \quad (4.19)$$

Eq. 4.19 is the core of the background estimation method and it is crusial to realize that the QCD background shape is tightly constrained by the transfer functions. To see this, it is helpful to count the degree of freedoms in Eq. 4.19. There are 22 m_{SD} bins times 6 p_{T} bins for both the pass and fail region, giving 264 observed data to fit. The parameters in $\epsilon^{\mathrm{QCD,MC}}$ are constrained since they come from fitting QCD MC data. Hence, the only unconstrained nuisance parameters in Eq. 4.19 are the 132 $N^{\mathrm{QCD}}_{\mathrm{fail}}$ in each mass and p_{T} bin and the bernstein coefficients in the transfer function TF, whose total number is typically less than 10 as will be explained in Sec. 4.7.1.

In other words, the 132 observation in the pass(signal) region is fitting with less than 10 unconstrained parameters.

For the signal, $t\bar{t}$, and resonant electroweak backgrounds (W/Z), we add them as binned templates derived from MC to both the failing and passing regions. The advantage of this approach is that it handles the signal, W/Z, and $t\bar{t}$ contamination in the failing region correctly.

Finally, we add the single muon control region as an additional analysis category in the combined fit. As an additional category, the normalizations of the $t\bar{t}$ background contributions in the single muon control region and the hadronic signal region are both tied to the two $t\bar{t}$ scale factors, $\mathrm{SF}^{t\bar{t}}_{\mathrm{norm}}$ and $\mathrm{SF}^{\mathrm{top}}_{\mathrm{mis-tag}}$. Since signal contamination in the single muon control region is minimal, the two scale factors are approximately the only free parameters to fit for the observed data that passes and fails the DDB selection in the single muon control region. Hence, the two $t\bar{t}$ scale factors are mostly constrained by the single muoncontrol region, while handling the contemination between hadronic signal region correctly.

The likelihood for the signal extraction can thus be written as

$$\begin{aligned}\mathcal{L}(\mathrm{data}|\mu,\boldsymbol{\theta}) = &\prod_{i,j} \mathrm{Poisson}\left(N^{\mathrm{data}}_{\mathrm{fail},i,j}|N^{\mathrm{QCD}}_{\mathrm{fail},i,j} + N^{t\bar{t}}_{\mathrm{fail},i,j} + N^{V}_{\mathrm{fail},i,j} + \mu N^{\mathrm{H}(b\bar{b})}_{\mathrm{fail},i,j}\right) \\ &\times \prod_{i,j} \mathrm{Poisson}\left(N^{\mathrm{data}}_{\mathrm{pass},i,j}|N^{\mathrm{QCD}}_{\mathrm{pass},i,j} + N^{t\bar{t}}_{\mathrm{pass},i,j} + N^{V}_{\mathrm{pass},i,j} + \mu N^{\mathrm{H}(b\bar{b})}_{\mathrm{pass},i,j}\right) \\ &\times \mathcal{L}_{1\mu\ \mathrm{CR}}(\mathrm{data}|\mu,\boldsymbol{\theta}) \times \mathrm{Constraint}(\boldsymbol{\theta}|\bar{\boldsymbol{\theta}},\boldsymbol{\delta\theta}) \ , \end{aligned} \tag{4.20}$$

where $N^{\mathrm{data}}_{\mathrm{fail},i,j}$ ($N^{\mathrm{data}}_{\mathrm{pass},i,j}$) is the observed number of events in the i^{th} m_{SD} bin and j^{th}

p_{T} bin of the failing (passing) region, $N^{\text{process}}_{\text{fail},i,j}$ ($N^{\text{process}}_{\text{pass},i,j}$) is the expected number of events for each background or signal process in the i^{th} m_{SD} bin and j^{th} p_{T} bin of the failing (passing) region, μ is the SM Higgs signal strength, $\text{Poisson}(x|y) = y^x e^{-y}/x!$, $\mathcal{L}_{1\mu\ \text{CR}}(\text{data}|\mu, \boldsymbol{\theta})$ is the likelihood for the single muon control region, which takes on a very similar Poisson counting experiment form, and $\text{Constraint}(\boldsymbol{\theta}|\bar{\boldsymbol{\theta}}, \boldsymbol{\delta\theta})$ is the constraint term for the nuisance parameters $\boldsymbol{\theta}$. Note, some of the nuisance parameters do not have an additional constraint in the fit, namely: the QCD yields in the failing category $N^{\text{QCD}}_{\text{fail},i,j}$, the Bernstein coefficients of the data-residual transfer factor $a_{k,\ell}$, and the $\mathrm{t\bar{t}}$-related scale factors. However, other nuisance parameters that model systematic uncertainties, such as luminosity uncertainty, jet energy scale/resolution, etc., are constrained.

Practically, the maximum likelihood fit is performed using the Higgs combine tool [96]. The fit determines the nuisance parameters $\boldsymbol{\theta}$ and the signal strength μ simultaneously via profile likelihood.

Profile likelihood

Given the binned likelihood on data $\mathcal{L}(\text{data}|\mu, \boldsymbol{\theta})$, we define the profile likelihood test statistic following the LHC $\mathrm{CL_s}$ procedure [97],

$$\tilde{q}_\mu = -2 \log \frac{\mathcal{L}(\text{data}|\mu, \hat{\boldsymbol{\theta}}_\mu)}{\mathcal{L}(\text{data}|\hat{\mu}, \hat{\boldsymbol{\theta}})} \ , \quad 0 \leq \hat{\mu} \leq \mu \ , \tag{4.21}$$

where $\hat{\boldsymbol{\theta}}_\mu$ refers to the conditional maximum likelihood estimators of $\boldsymbol{\theta}$ assuming a given value μ, and $\hat{\mu}$ and $\hat{\boldsymbol{\mu}}$ correspond to the global maximum of the likelihood. In

the asymptotic regime, we may derive the observed 95% CL upper limit on the signal strength by computing the value of μ that satisfies,

$$\mathrm{CL_s} \equiv \frac{\mathrm{CL_{s+b}}}{\mathrm{CL_b}} = \frac{1 - \Phi(\sqrt{\tilde{q}_\mu})}{\Phi(\sqrt{\tilde{q}_{\mu,\mathrm{A}}} - \sqrt{\tilde{q}_\mu})} = \alpha \ , \tag{4.22}$$

where $\alpha = 0.05$, $\tilde{q}_{\mu,\mathrm{A}}$[1] is the test statistic evaluated on the Asimov dataset [98] corresponding exactly to the expected background and the nominal nuisance parameters (setting all statistical fluctuations to zero) [99, 97], and $\Phi(x)$ is the cumulative distribution function of the standard normal distribution. A similar expression is used to derive the median expected 95% CL upper limit,

$$\frac{1 - \Phi(\sqrt{\tilde{q}_{\mu,\mathrm{A}}})}{0.5} = \alpha \ . \tag{4.23}$$

Conversely, in the case of a discovery, one tests $\mu = 0$ and measures the "local significance" using a modified test statistic,

$$q_0 = -2 \log \frac{\mathcal{L}(\mathrm{data}|0, \hat{\boldsymbol{\theta}}_0)}{\mathcal{L}(\mathrm{data}|\hat{\mu}, \hat{\boldsymbol{\theta}})} \ , \quad \hat{\mu} \geq 0 \ . \tag{4.24}$$

The observed local significance is then simply $Z = \sqrt{q_0}$.

[1] Note $\tilde{q}_{\mu,\mathrm{A}} = \frac{\mu^2}{\sigma_\mathrm{A}^2}$ where σ_A is an estimator for the variance of μ.

4.7.1 Polynomial order and the *F*-test

We perform the Fisher-test to determine the minimum number of parameters needed to provide a good fit to the two transfer functions, one describing the expected mass sculpting in QCD MC and the other describing the data-MC residual differences.

For two fit models, 1 and 2, where model 1 with p_1 parameters is "nested" within model 2 with p_2 parameters ($p_2 > p_1$), F-test determines whether model 2 gives a *significantly* better fit to the data. For example, a polynomial function of 2nd order is considered nested within the 3rd order polynomial. To compare the goodness-of-fit(GOF) between the two models, we can construct a test statistics from the ratios of GOF test statistics normalized by the number of degree of freedoms in the respective model.

Explicitly, a random variate X will follow the F-distribution with parameters (d_1, d_2)

$$X = \frac{U_1/d_1}{U_2/d_2} \tag{4.25}$$

where U_1 and U_2 are independent and follow chi-squared distributions with d_1 and d_2 degrees of freedom respectively.

For the choice of GOF test statistics, we follow the "saturated model" [100, 101]

$$-2\log\lambda = 2\sum_{j}^{n_{\text{bins}}} = \big(f_j - d_j + d_j \log\big(d_j/f_j\big)\big) \tag{4.26}$$

where d_j is the data in the j^{th} bin and f_j is the prediction in the j^{th} bin.

As noted Ref. [100, 101], this test statistic asymptotically follows a chi-squared

distribution with $n_{\text{bins}} - p_i$ degrees of freedom and is an appropriate metric for the goodness-of-fit of maximum likelihood fits using a Poisson likelihood.

Therefore, we can construct the F-statistic as,

$$F = \frac{(-2\log(\lambda_1/\lambda_2)/(p_2 - p_1))}{(-2\log\lambda_2/(n_{\text{bins}} - p_2))} , \tag{4.27}$$

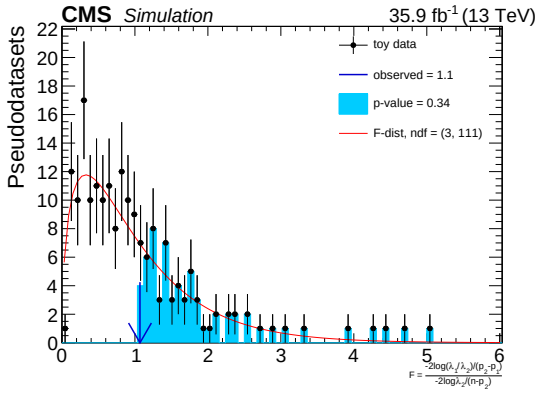
where n_{bins} is the number of bins, $-2\log\lambda_i$ is the "saturated model" goodness-of-fit test statistic for model i, given by Eqn. 4.26.

Under the null hypothesis that model 2 does not provide a significantly better fit than model 1, F will have an F-distribution with $(p_2 - p_1, p_2 - n)$ degrees of freedom. We may reject the null hypothesis if F calculated from data is greater than the critical value of the F-distribution for the desired false rejection probability of $\alpha = 0.05$.

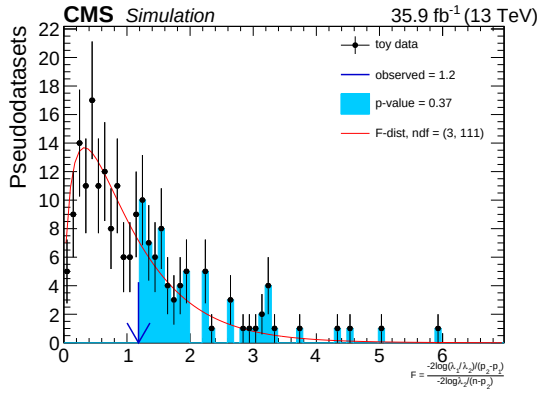
We sequentially test each order of the transfer factor polynomial against the next higher order. We begin with $(n_\rho = 1, n_{p_{\text{T}}} = 1)$ and test against $(n_\rho = 2, n_{p_{\text{T}}} = 1)$ and $(n_\rho = 1, n_{p_{\text{T}}} = 2)$. If one provides a significantly better fit (p-value $< 5\%$), we choose that as the new default. If both provide a significantly better fit, we choose the one with the better goodness-of-fit p-value. Subsequently, we test the new default against $(n_\rho = 2, n_{p_{\text{T}}} = 2)$ and so on. We stop once we evaluate a p-value $> 5\%$ for the higher order polynomials, indicating that the current default has sufficient parameters.

First, we perform the F-test on MC, shown in Fig. 4.40. This F-test shows the $(n_\rho = 2, n_{p_{\text{T}}} = 2)$ polynomial is sufficient to fit the QCD transfer factor in MC.

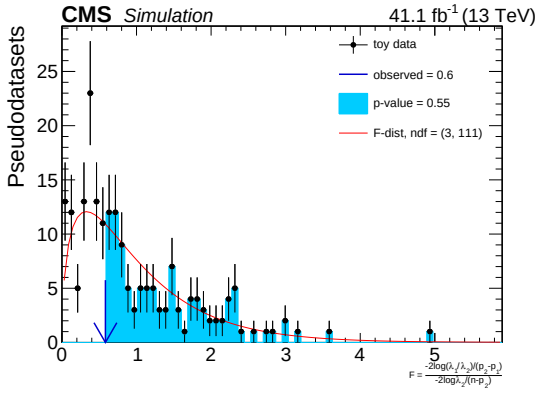
Subsequently, we perform F-tests to the 2016, 2017, and 2018 datasets for the



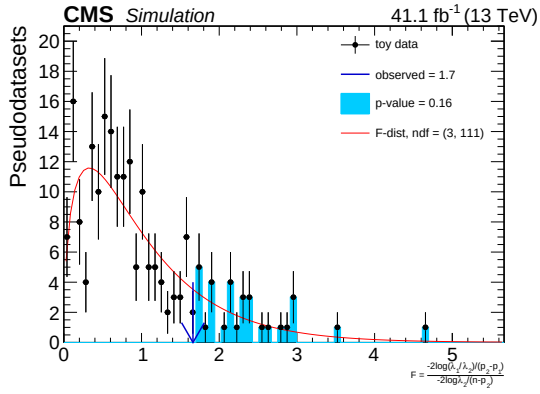
(a) $(n_\rho = 2, n_{p_\mathrm{T}} = 2)$ vs. $(n_\rho = 2, n_{p_\mathrm{T}} = 3)$



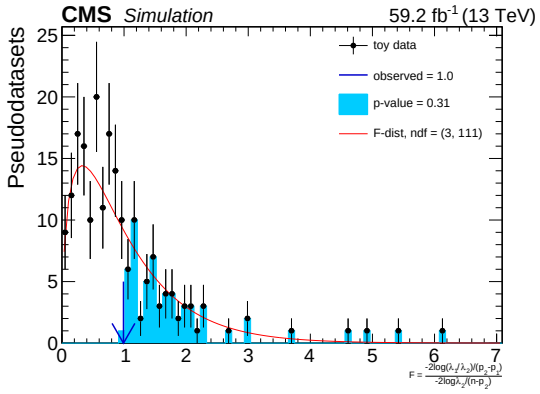
(b) $(n_\rho = 2, n_{p_\mathrm{T}} = 2)$ vs. $(n_\rho = 3, n_{p_\mathrm{T}} = 2)$



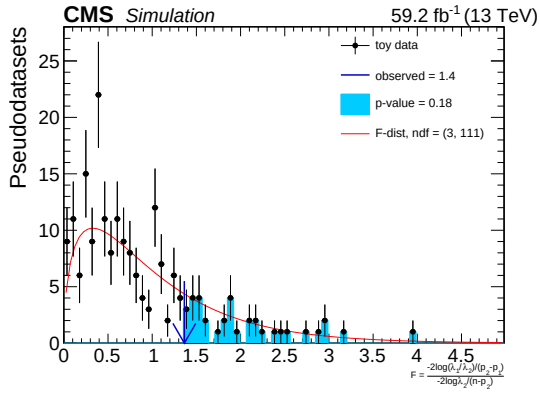
(c) $(n_\rho = 2, n_{p_\mathrm{T}} = 2)$ vs. $(n_\rho = 2, n_{p_\mathrm{T}} = 3)$



(d) $(n_\rho = 2, n_{p_\mathrm{T}} = 2)$ vs. $(n_\rho = 3, n_{p_\mathrm{T}} = 2)$



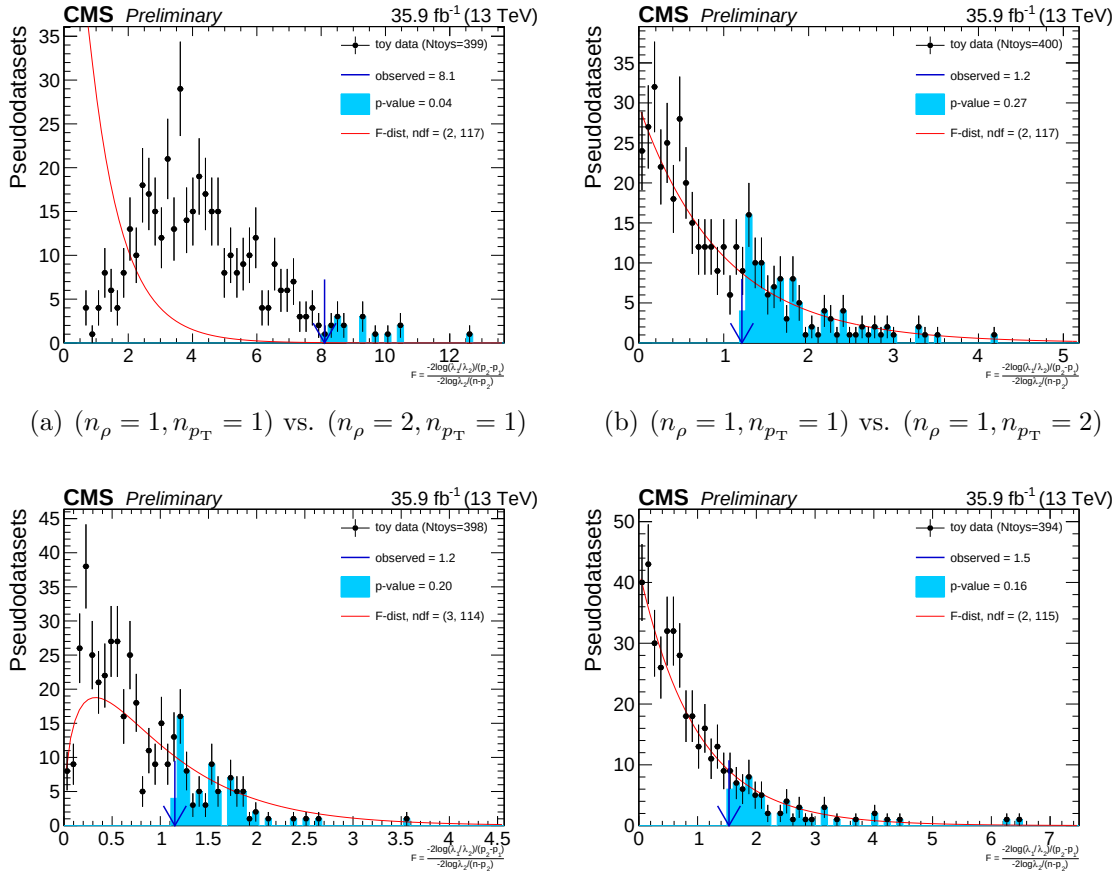
(e) $(n_\rho = 2, n_{p_\mathrm{T}} = 2)$ vs. $(n_\rho = 2, n_{p_\mathrm{T}} = 3)$



(f) $(n_\rho = 2, n_{p_\mathrm{T}} = 2)$ vs. $(n_\rho = 3, n_{p_\mathrm{T}} = 2)$

Figure 4.40: The F-test for different hypotheses on the MC (with $\mu = 1$), corresponding to the 2016(top),2017(middle),2018(bottom) dataset.

data/MC ratio after accounting for the expected QCD pass-fail shape from MC. Figures 4.41-4.42 shows the F-test results when fitting the 2016, 2017 and 2018 datasets. We find the necessary order of the polynomial is $n_\rho = 2, n_{p_T} = 1)$ for 2016 data, and $(n_\rho = 1, n_{p_T} = 1)$ for 2017 and 2018 data.



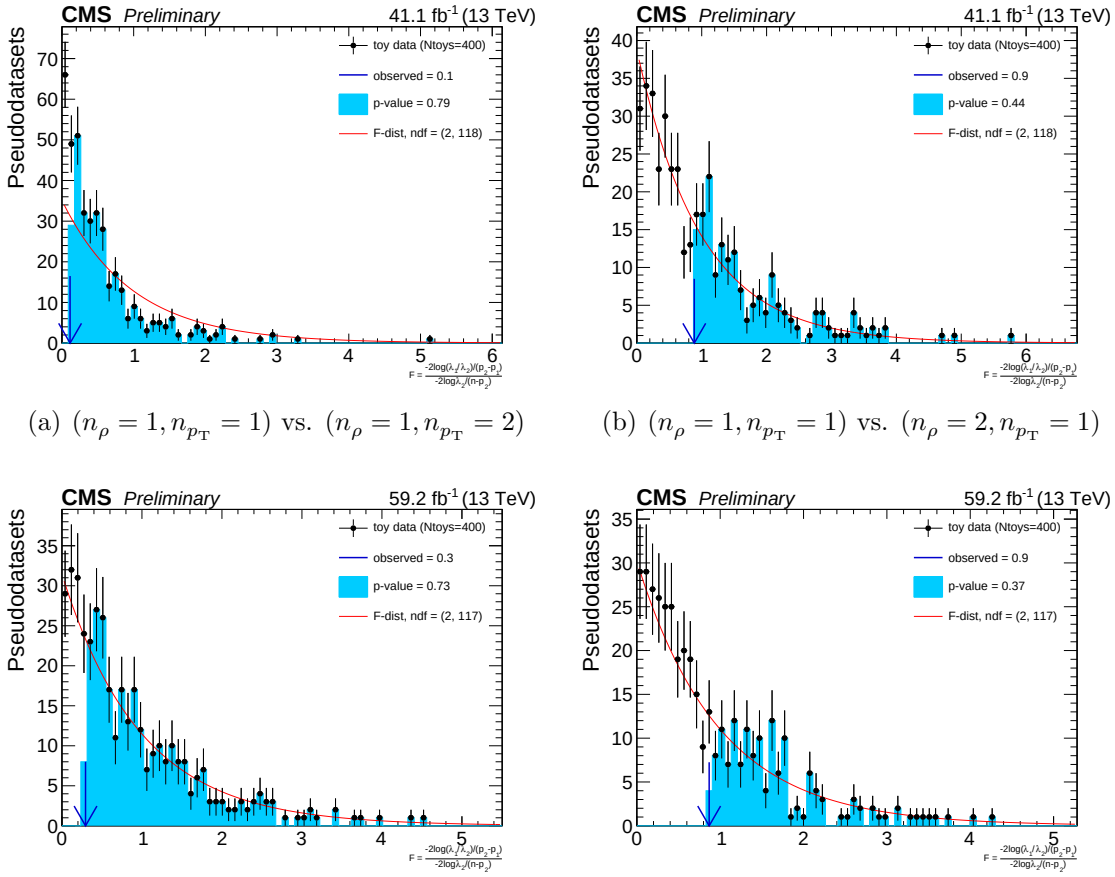
(a) $(n_\rho = 1, n_{p_T} = 1)$ vs. $(n_\rho = 2, n_{p_T} = 1)$

(b) $(n_\rho = 1, n_{p_T} = 1)$ vs. $(n_\rho = 1, n_{p_T} = 2)$

(c) $(n_\rho = 2, n_{p_T} = 1)$ vs. $(n_\rho = 2, n_{p_T} = 2)$

(d) $(n_\rho = 2, n_{p_T} = 1)$ vs. $(n_\rho = 3, n_{p_T} = 1)$

Figure 4.41: The F-test for different hypotheses on the 2016 dataset.



(a) $(n_\rho = 1, n_{p_\mathrm{T}} = 1)$ vs. $(n_\rho = 1, n_{p_\mathrm{T}} = 2)$

(b) $(n_\rho = 1, n_{p_\mathrm{T}} = 1)$ vs. $(n_\rho = 2, n_{p_\mathrm{T}} = 1)$

(c) $(n_\rho = 1, n_{p_\mathrm{T}} = 1)$ vs. $(n_\rho = 1, n_{p_\mathrm{T}} = 2)$

(d) $(n_\rho = 1, n_{p_\mathrm{T}} = 1)$ vs. $(n_\rho = 2, n_{p_\mathrm{T}} = 1)$

Figure 4.42: The F-test for different hypotheses on the 2017 (top) and 2018 (bottom) dataset.

4.7.2 Fit results with simulated QCD events

With the order of Bernstein polynomials determined from the F-test as $n_\rho = 2, n_{p_T} = 2$, Eq. 4.11 becomes

$$\epsilon^{\mathrm{QCD,MC}}(\rho, p_T) = \overline{\epsilon^{\mathrm{QCD,MC}}} \sum_{k=0}^{2} \sum_{\ell=0}^{2} c_{k,\ell} \left[b_{k,2}(\rho) b_{\ell,2}(p_T) \right], \tag{4.28}$$

where $b_{0,2}(x) = (1-x)^2$,$b_{1,2}(x) = 2x(1-x)$,$b_{2,2}(x) = x^2$ are the Bernstein basis functions and $c_{k,\ell}$ are the 9 parameters of the transfer function.

Fig. 4.69 shows the 2D transfer factors for the signal+background fit to the 2016, 2017, and 2018 MC simulation.

As a demonstration of the method, we also performed the full simultaneous fit method using simulations events only. Since there are no data-mc differences in a pure simulation setting, we set the coefficients for $TF^{\mathrm{data}}(\rho, p_T)$ in Eq. 4.9 to be 1. Figures 4.44-4.49 show the signal+background fit to the 2016, 2017, and 2018 MC simulation including the EW and top backgrounds, separately for each p_T category. Similarly, Fig. 4.50 shows the signal+background fit to the 2016, 2017, and 2018 MC simulation summed over the p_T categories.

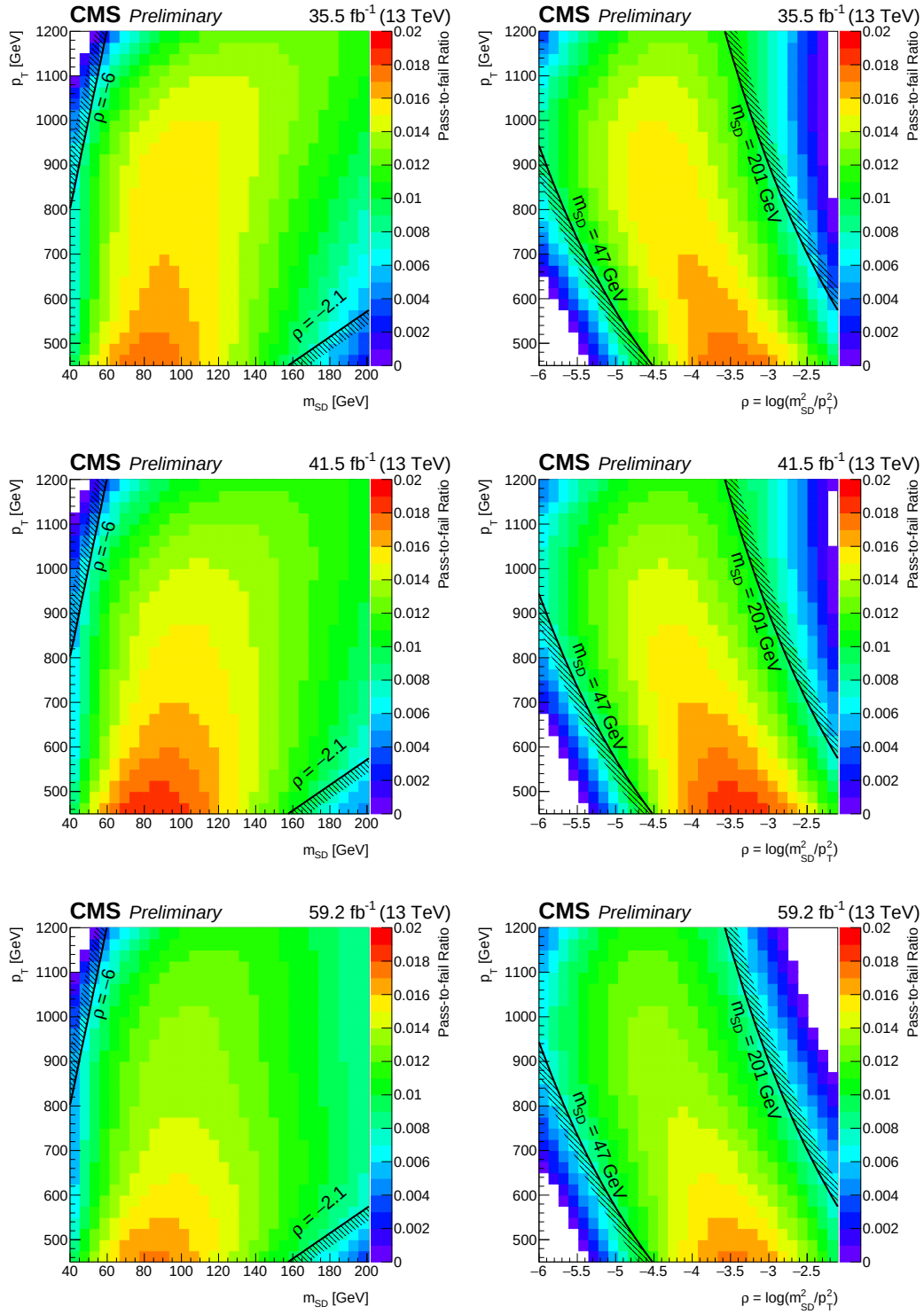


Figure 4.43: The fitted pass-fail ratio $\epsilon^{\mathrm{QCDMC}}$ as a function of p_{T} and m_{SD} (left) or p_{T} and ρ (right) for QCD MC corresponding to conditions in 2016 (top), 2017 (middle), and 2018 (bottom).

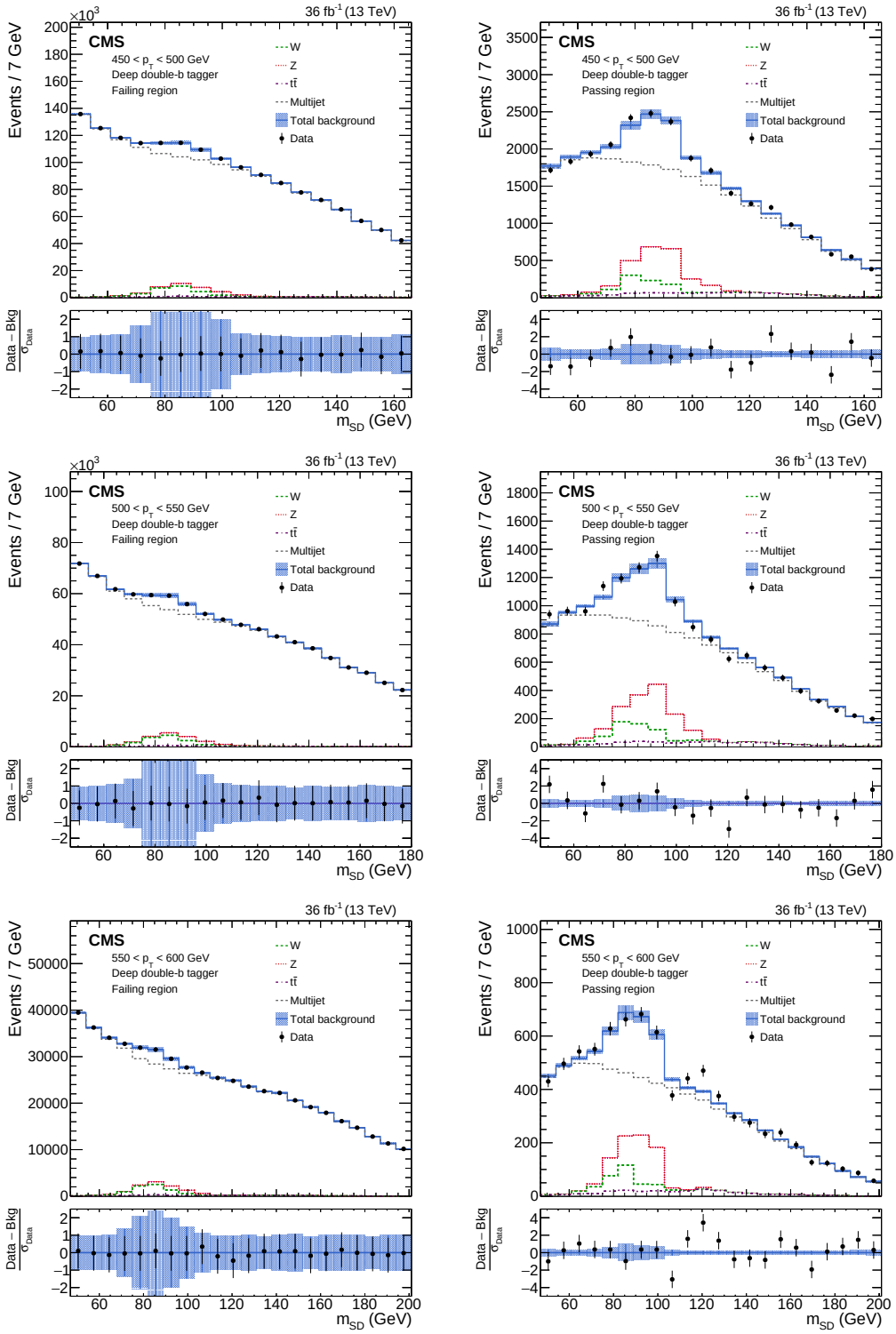


Figure 4.44: Post-fit m_{SD} distributions in simulated MC events corresponding to 2016 conditions for the pass and fail regions and the first three p_{T} categories (450-500,500-550,550-600 from top to bottom) by using a polynomial 2nd order in ρ and 2nd order in p_{T}. The EW and signal events are also shown separately.

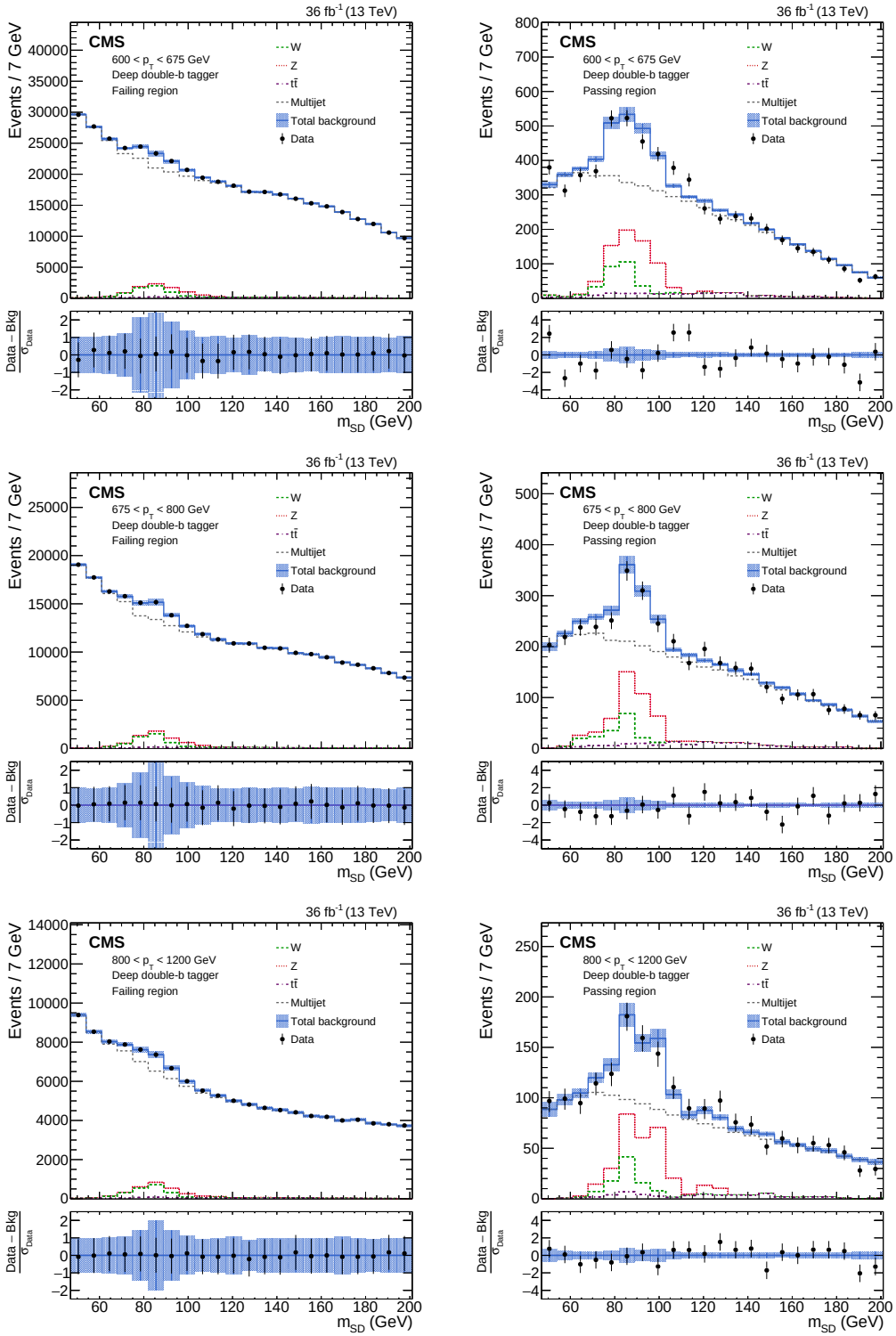


Figure 4.45: Post-fit m_{SD} distributions in simulated MC events corresponding to 2016 conditions for the pass and fail regions and the second three p_{T} categories (600-675,675-800,800-1200 from top to bottom) by using a polynomial 2nd order in ρ and 2nd order in p_{T} The EW and signal events are also shown separately.

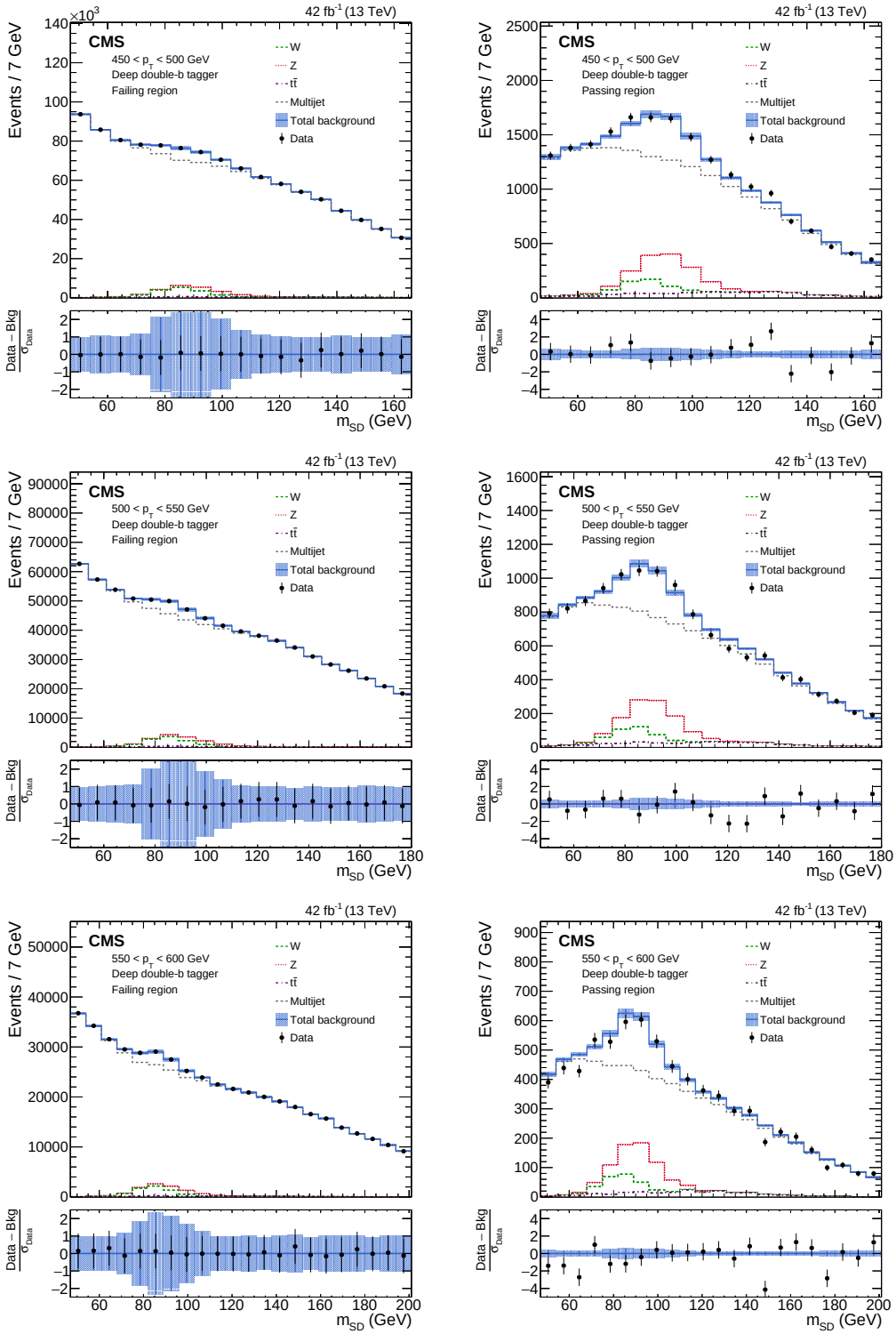


Figure 4.46: Post-fit m_{SD} distributions in simulated MC events corresponding to 2017 conditions for the pass and fail regions and the first three p_{T} categories (450-500,500-550,550-600 from top to bottom) by using a polynomial 2nd order in ρ and 2nd order in p_{T}. The EW and signal events are also shown separately.

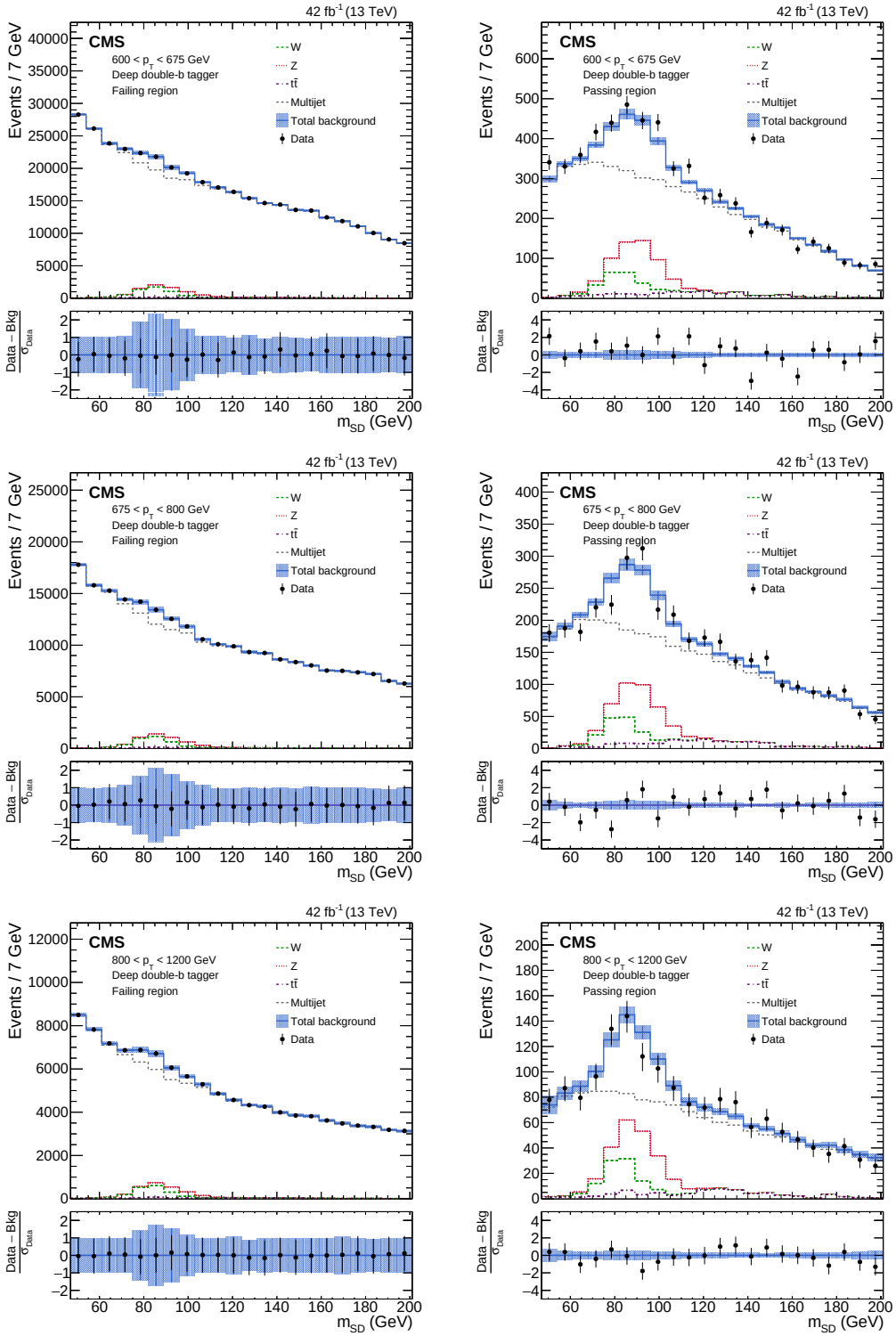


Figure 4.47: Post-fit m_{SD} distributions in simulated MC events corresponding to 2017 conditions for the pass and fail regions and the second three p_{T} categories (600-675,675-800,800-1200 from top to bottom) by using a polynomial 2nd order in ρ and 2nd order in p_{T} The EW and signal events are also shown separately.

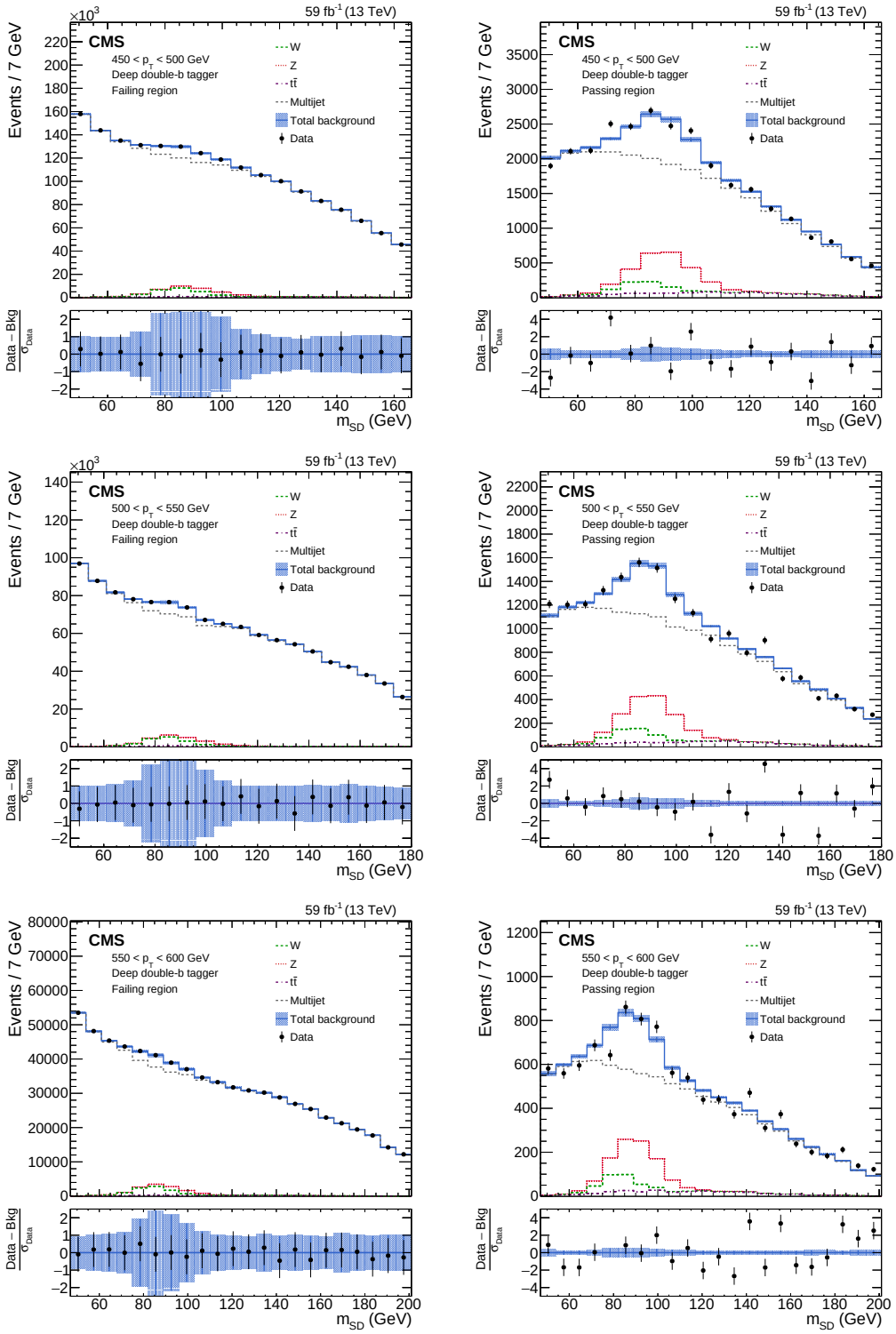


Figure 4.48: Post-fit m_{SD} distributions in simulated MC events corresponding to 2018 conditions for the pass and fail regions and the first three p_{T} categories (450-500,500-550,550-600 from top to bottom) by using a polynomial 2nd order in ρ and 2nd order in p_{T}. The EW and signal events are also shown separately.

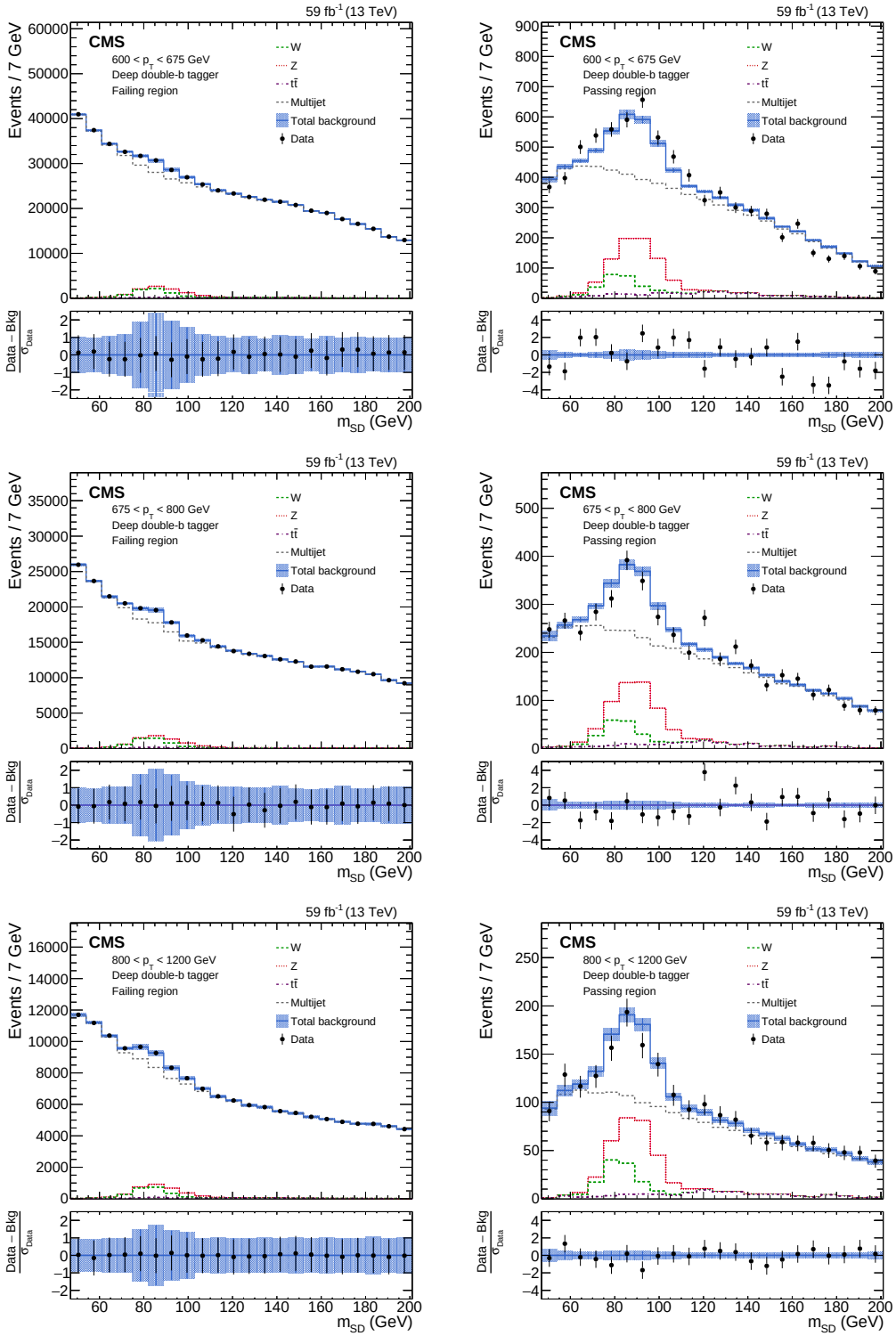


Figure 4.49: Post-fit m_{SD} distributions in simulated MC events corresponding to 2017 conditions for the pass and fail regions and the second three p_{T} categories (600-675,675-800,800-1200 from top to bottom) by using a polynomial 2nd order in ρ and 2nd order in p_{T} The EW and signal events are also shown separately.

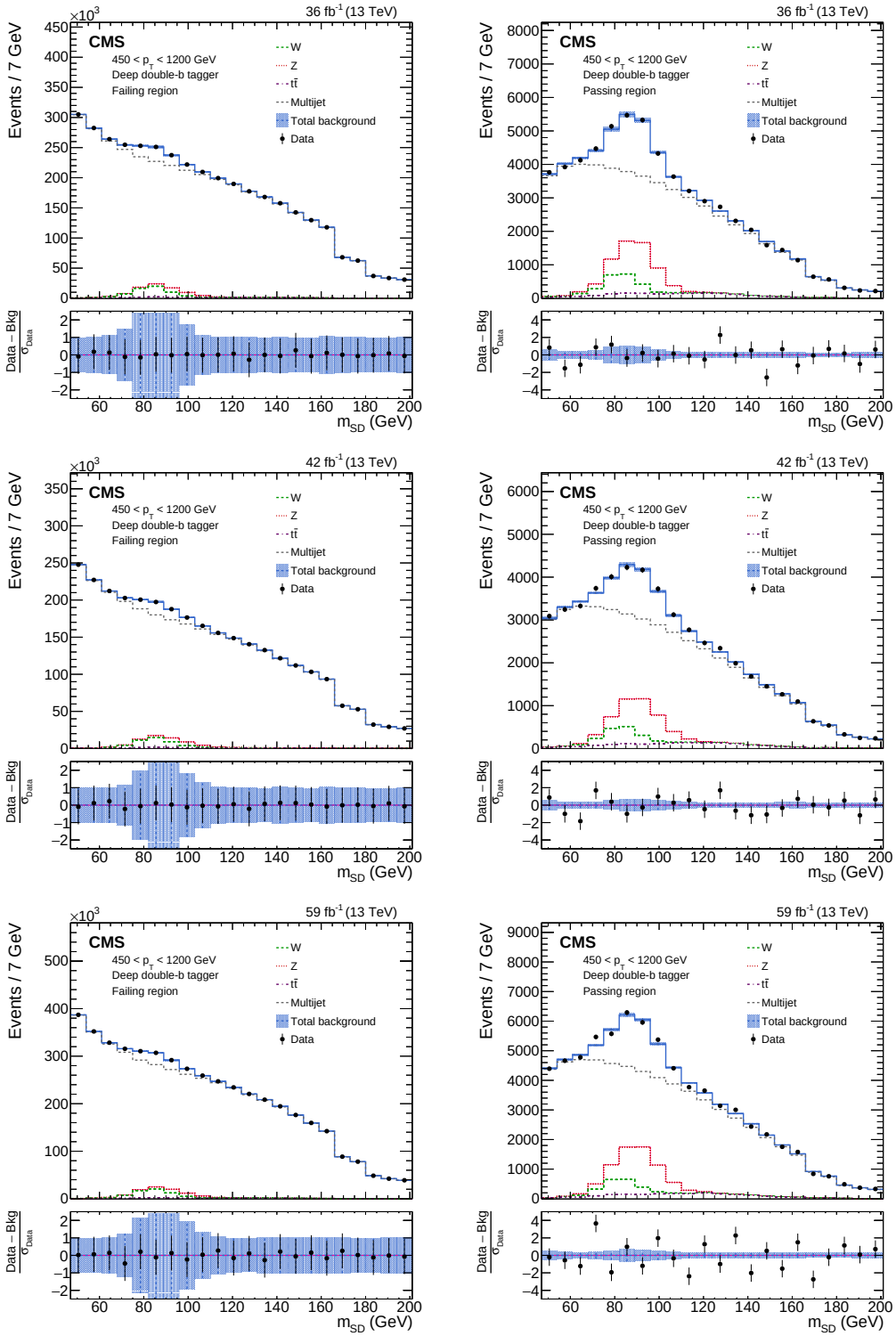


Figure 4.50: Post-fit m_{SD} distributions in simulated MC events corresponding to 2016 (top), 2017 (middle), and 2018 (bottom) conditions for the pass and fail regions summing all 6 p_{T} categories by using a polynomial 2nd order in ρ and 2nd order in p_{T} The EW and signal events are also shown separately.

4.7.3 Signal injection tests based on data best fit

The closure of the fit method is tested using pseudoexperiments drawn from the best fit in data. The result of the closure studies are shown in Fig. 4.51, for different injected $\mathrm{H}(\mathrm{b\bar{b}})$ and signal strengths: $\mu_{\mathrm{H}} = 0, 1, 2, 5$. From the observed fits, the pulls are found to be consistent with unity and we do not observe any significant deviation. All uncertainties related to the QCD estimation are included.

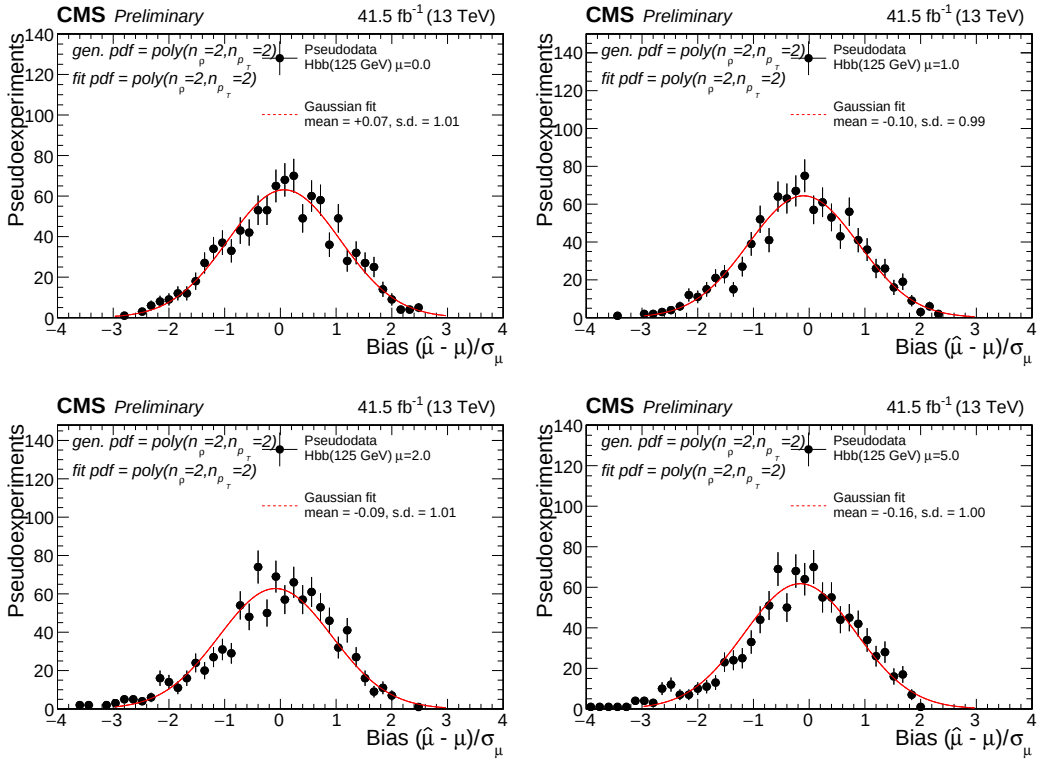


Figure 4.51: The pull distribution when generating and fitting with the nominal function. Plots are shown for different injected $\mathrm{H}(\mathrm{b\bar{b}})$ signal strengths: $\mu_{\mathrm{H}} = 0, 1, 2, 5$.

Table 4.9: Order determined by F-test and corresponding GOF p-value for the empirical function fit in each p_{T} category

p_{T} category [GeV]	$[450, 500]$	$[500, 550]$	$[550, 600]$	$[600, 675]$	$[675, 800]$	$[800, 1200]$
$n_{m_{\mathrm{SD}}}$ order via F-test	4	3	3	3	2	2
GOF p-value	36%	14%	2%	3%	19%	99%

4.7.4 Empirical function fit sensitivity

We also check the sensitivity with an empirical function fitting the QCD mass spectrum in the pass category directly. The parameterization we use is a Bernstein polynomial,

$$f(m_{\mathrm{SD}}) = \sum_{k=0}^{n_{m_{\mathrm{SD}}}} a_k b_{k,n_{m_{\mathrm{SD}}}}(m_{\mathrm{SD}}), \tag{4.29}$$

with independent polynomial coefficient parameters in each p_{T} category. Note that the m_{SD} shape changes non-trivially as a function of p_{T}, therefore without taking advantage of the similarity between the pass and fail category, it is expected to be a more difficult task to model the background shape in pass region.

The minimum order $n_{m_{\mathrm{SD}}}$ needed for the fit in each category is determined via an F-test as shown in Table 4.9. Note that satisfactory goodness-of-fit is not achieved in p_{T} categories possibly indicating the need for a higher-order parameterization. This indicates that, without using the fail category, it takes many more free parameters in the fit function to model the m_{SD} spectrum across the p_{T} range in the signal region.

Fig. 4.52 shows the signal+background fit result in the pass category for Asimov data based on the 2017 MC with the empirical function, while Fig. 4.53 shows the signal+background fit result in the pass category for Asimov data based on 2017 MC with the pass-fail ratio $R_{\mathrm{p/f}}$ transfer factor. The expected sensitivities are shown in

Table 4.10: Expected results based on an Asimov data set generated from 2017 MC.

2017 MC	Empirical function fit	$R_{\mathrm{p/f}}$ TF fit
Expected H significance ($\mu_{\mathrm{H}} = 1$)	0.73σ	1.00σ

Table 4.10. The sensitivity improves by approximately 40% from 0.7σ to 1σ by using the pass-fail ratio $R_{\mathrm{p/f}}$ transfer factor formulation of the background estimate.

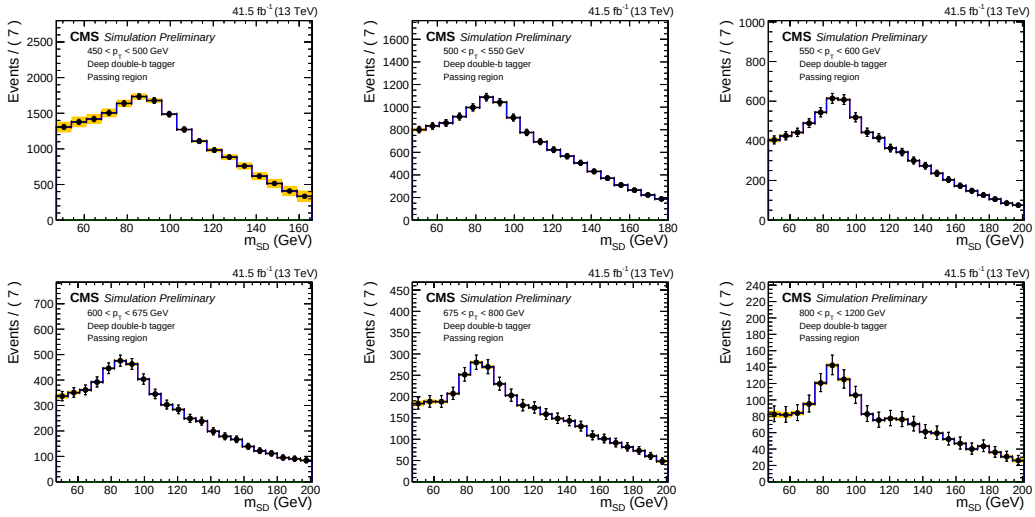


Figure 4.52: Empirical function signal+background fit to the pass category directly with the 2017 MC.

4.8 Systematic uncertainties

4.8.1 Mass shapes in QCD

The background estimation for QCD can be summarized in Eq. 4.19:

$$N^{\mathrm{QCD}}_{\mathrm{pass}}(m^i_{\mathrm{SD}}, p^j_{\mathrm{T}}) = \epsilon^{\mathrm{QCD,MC}}(\rho^{ij}, p^j_{\mathrm{T}}) \times TF^{\mathrm{data}}(\rho^{ij}, p^j_{\mathrm{T}}) \times N^{\mathrm{QCD}}_{\mathrm{fail}}(m^i_{\mathrm{SD}}, p^j_{\mathrm{T}}) \ . \quad (4.30)$$

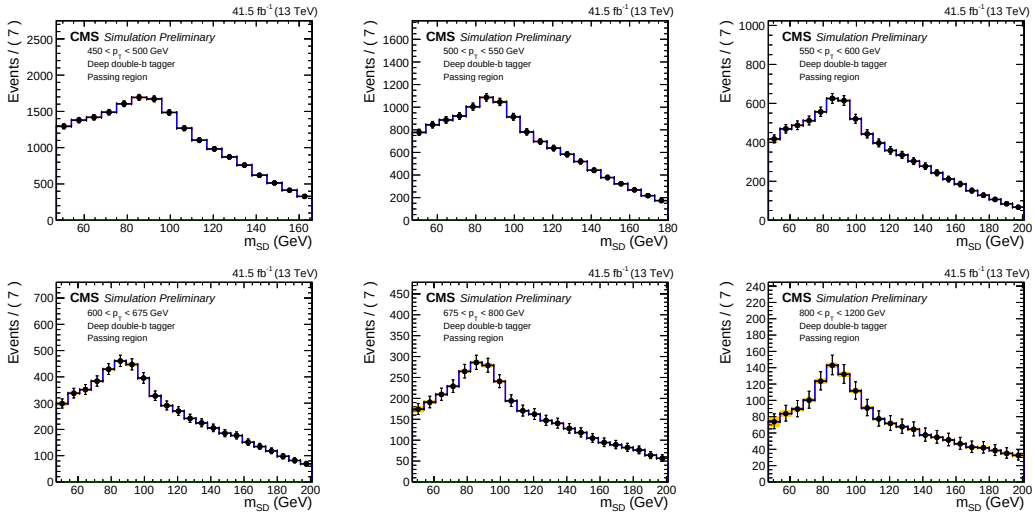


Figure 4.53: $R_{\mathrm{p/f}}$ TF function signal+background fit to the pass category directly with the 2017 MC.

From the above equation, we can see the uncertainties in this method come from three main sources and are accounted bin-by-bin:

- Statistical uncertainty in the statistics of the fail region per bin ($N_{\mathrm{fail}}^{\mathrm{QCD}}$),
- The polynomial coefficient uncertainties in the data-residual fit (TF^{data}),
- The polynomial coefficient uncertainties in the QCD MC fit ($\epsilon^{\mathrm{QCD,MC}}$)

The first two terms are unconstrained parameters, while the last term is treated as constrained parameters whose uncertainty are propagated from the fit to QCD MC as follows.

The separate fit to QCD MC is performed using a maximum likelihood statistic with the `RooFit` program. We enable the `SumW2Error` option to properly embed the significant MC statistical uncertainty into the sculpting parameters. The full covariance matrix of the coefficients $c_{k,\ell}$ from this fit is propagated to the data/MC

"rhalphabet" fit described in Eq. 4.9 by constructing a transform from decorrelated nuisances with unit Gaussian constraints to the properly correlated parameters using the eigenvalues and eigenvectors of the covariance matrix. Typical per-bin uncertainties are between 1-5% depending on p_{T} and ρ.

4.8.2 Mass scale, resolution and substructure scale factors

The efficiency of our V-tag selection using substructure ($N_2^{1,\mathrm{DDT}}$) is measured in a separate semi-leptonic top control region in data, by the JetMET POG.

The semileptonic $\mathrm{t\bar{t}}$ control sample is selected by requiring one lepton and the presence of b jets, as detailed in Sec. 4.6.1. This provides a pure sample of $\mathrm{t\bar{t}}$ events from which we isolate those with the particular topology of boosted W bosons. From this sample we can derive the data/MC scale factor based on the $n_{\mathrm{pass}}/n_{\mathrm{fail}}$ in each of the data and MC.

In addition, we use this sample to derive a nominal value and uncertainty on the jet mass scale and resolution. In particular, we allow the peak position and resolution to vary within the uncertainties derived using shape templates in the analysis signal regions, displayed in Fig. 4.55.

Table 4.11 summarizes the jet mass scale, resolution, $N_2^{1,\mathrm{DDT}}$ and their uncertainties measured for all three data taking years.

As the semileptonic $\mathrm{t\bar{t}}$ sample does not contain a large population of jets with very high p_{T}, an additional systematic uncertainty is included to account for the extrapolation to very high p_{T} jets. This additional uncertainty is estimated to be

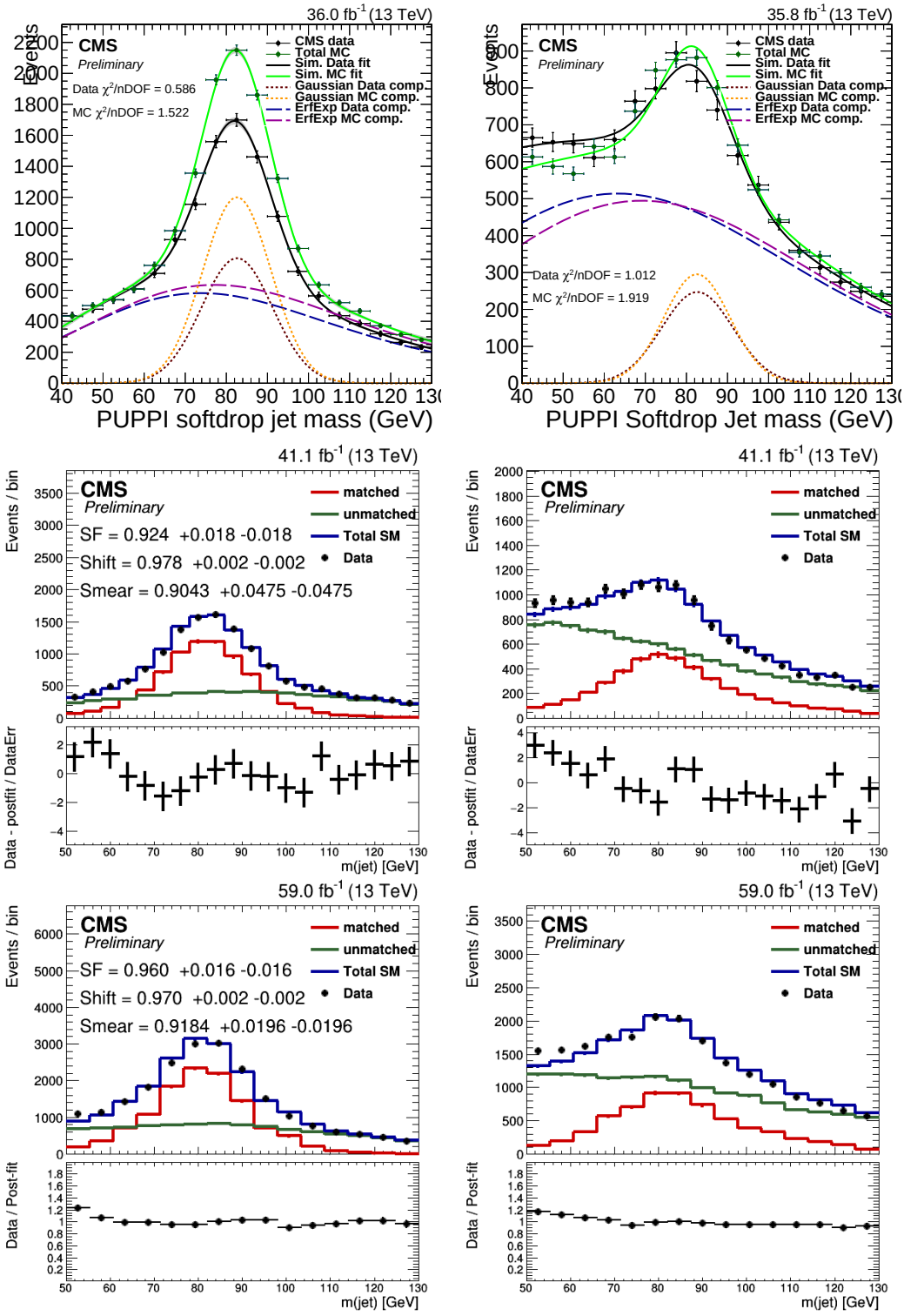
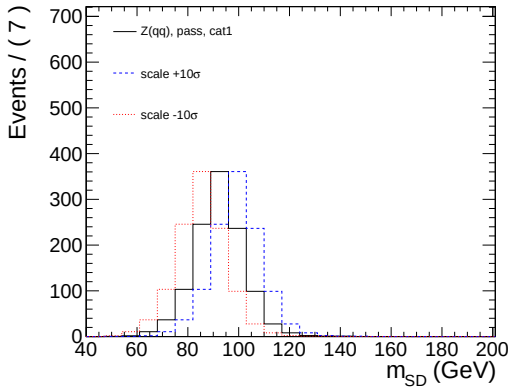
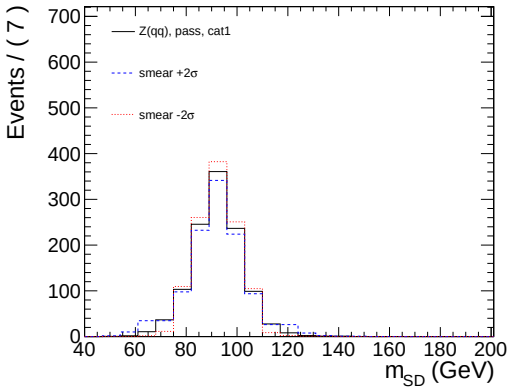


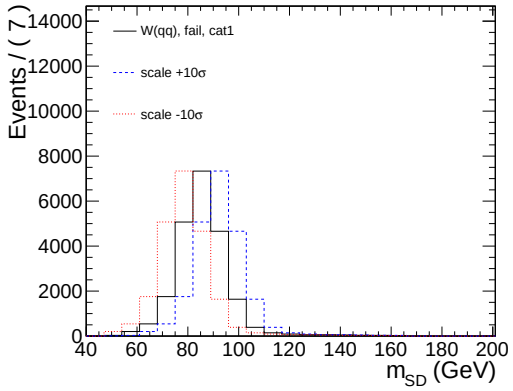
Figure 4.54: $N_2^{1,\mathrm{DDT}}$ Passing and failing region fits in the semileptonic $t\bar{t}$ sample for 2016 (top), 2017 (middle) and 2018 (bottom). Note for 2016, pre-fit and post-fit models are shown, while only post-fit is shown for 2017 and 2018. In addition, the ratio panels for 2017 and 2018 correspond to different quantities.



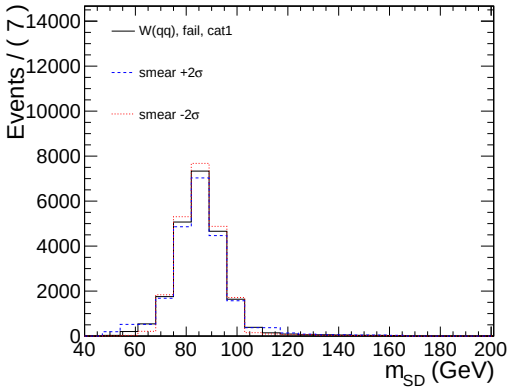
(a) scale for Z+jets



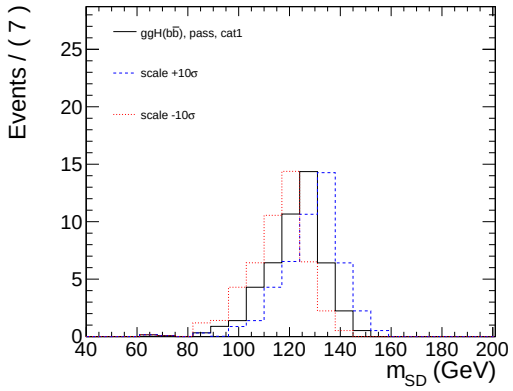
(b) smear for Z+jets



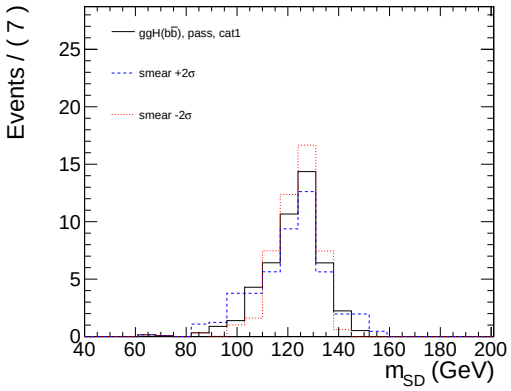
(c) scale for W+jets



(d) smear for W+jets



(e) scale for $\mathrm{ggH(b\bar{b})}$



(f) smear for $\mathrm{ggH(b\bar{b})}$

Figure 4.55: Effect of varying the jet mass scale by $\pm 10\sigma$ and the jet mass resolution by $\pm 2\sigma$ for W/Z+jets and the signal $\mathrm{ggH(b\bar{b})}$ in p_{T} category 1 ($450 < p_{\mathrm{T}} < 500\,\mathrm{GeV}$).

Table 4.11: Summary of mass scale/resolution and $N_2^{1,\mathrm{DDT}}$ scale factors measured in semileptonic $\mathrm{t\bar{t}}$ control region for different data-taking years.

Data	Jet mass scale	Jet mass resolution	$N_2^{1,\mathrm{DDT}}$ selection
2016	1.00 ± 0.012	1.084 ± 0.091	0.993 ± 0.043
2017	0.987 ± 0.012	0.905 ± 0.048	0.924 ± 0.018
2018	0.970 ± 0.012	0.908 ± 0.014	0.953 ± 0.016

0.5% per 100 GeV, based on a study of fitting the m_{SD} distributions of p_{T}-binned samples of merged top quark jets with $p_{\mathrm{T}} > 350$ GeV. Fig. 4.56 shows the fitted p_{T} dependence of the mass scale measured in this sample. Thus, the jet mass scale uncertainty is allowed to vary in the signal extraction differently depending on the jet p_{T}, and ranges from 1.2% at 450 GeV to 2.1% at 800 GeV.

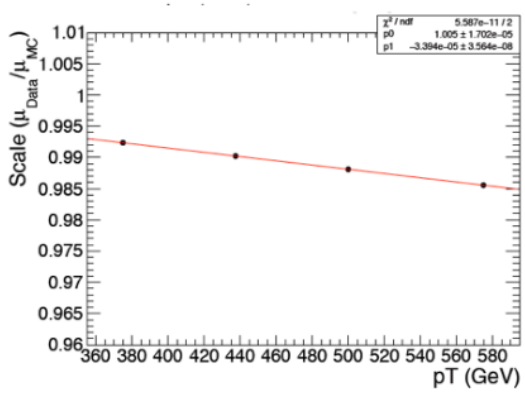


Figure 4.56: p_{T} dependence of the mass scale SF measured in the merged-W jet sample.

4.8.3 Deep double-b tagging

The efficiency of the deep double-b tagger is measured in the data sample, consisting of high p_{T} AK8 jets enriched in $\mathrm{b\bar{b}}$ from gluon splitting. In order to select topologies as

similar as possible to a signal jet, the AK8 jet has $p_{\mathrm{T}} > 300\,\mathrm{GeV}$ and mass $> 50\,\mathrm{GeV}$, it has also to be matched to at least two muons, each with $p_{\mathrm{T}} > 7$ GeV and $|\eta| < 2.4$.

The efficiency of the deep double-b tagger is measured in data and MC for four different operating points by the BTV-POG [102]. The measurement relies on the Jet Probability (JP) discriminant, for which the expected simulated distributions ("templates") are different for the various jet flavors. The fraction of b (from gluon splitting) jets is estimated by fitting the data distribution of the JP variable with the templates. This so-called Lifetime Tagging (LT) method [103] is also used to perform the measurement of the b jet identification efficiency scale factors(SF) for the standard anti-k_{T} $R = 0.4$ (AK4) jets [95]. No significant bias due to the di-muon requirement in the efficiency evaluation has been found.

Unlike the double-b tagger, we found that for deep double-b tagger, the $\mathrm{b\overline{b}}$ from gluon splitting process in QCD events has a significantly different distribution than signal $\mathrm{H} \to \mathrm{b\overline{b}}$ process in simulation, as shown in Fig. 4.57. This could also be related to the fact that the S.F. are measured to be systematically smaller than 1 for different working points [102]. As a result, we take a more agnostic approach to the S.F. measured with gluon splitting as proxy in data: we take the central value of 1.0 with a prior uncertainty that covers the measured S.F. in data.

Table 4.12 summarizes the deep double b efficiency scale factors and uncertainties measured and applied for all three data taking years.

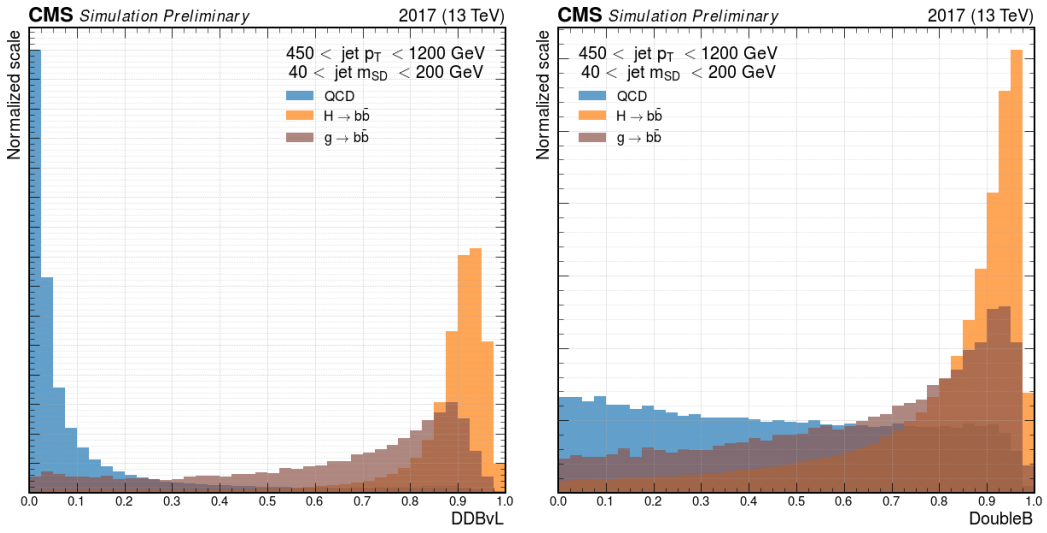


Figure 4.57: Distribution of deep double-b(left) and double-b(right) for QCD, gluon splitting and H $\to$ b$\bar{\text{b}}$ processes.

Table 4.12: Summary of measured and applied deep double b scale factors for different data-taking years.

Data	Working point	Measured DDB eff. S.F.	Applied DDB eff. S.F.
2016	M2 (0.89)	0.77 ± 0.07	1.0 ± 0.23
2017	M2 (0.89)	0.68 ± 0.06	1.0 ± 0.30
2018	M2 (0.89)	0.70 ± 0.07	1.0 ± 0.30

4.8.4 Higgs signal theory uncertainties

Following the recommendation of Ref. [86], the renormalization (μ_R) and factorization (μ_F) scale uncertainties are evaluated in the HJ-MiNLO simulation sample by adjusting them by factors of 0.5 and 2, finding the new event weight, and propagating the sum of weights to the signal region yield. The envelope of the relative effect of the 6 combinations (ignoring $0.5, 2$ and $2, 0.5$) about the nominal yield is computed, and shown in Fig. 4.58 as a function of reconstructed Higgs boson p_{T}. Based on this result, a flat 20% uncertainty is assigned to the signal simulation yield to account for

these effects.

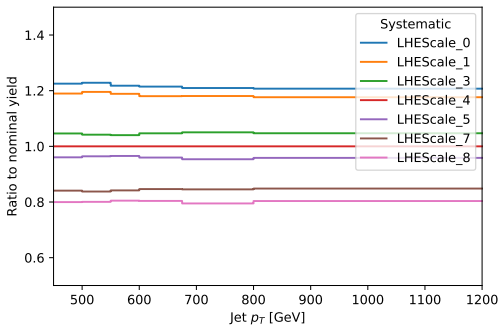


Figure 4.58: The envelope of the relative effect of the 6 combinations (ignoring $0.5, 2$ and $2, 0.5$) about the nominal ggHbb yield, as a function of reconstructed Higgs boson p_{T}. The yields are as predicted by the HJ-MiNLO generator, with 2017 detector conditions.

For the purposes of calculating the fiducial cross section (see Section 4.9.6), only the component of this uncertainty that affects the acceptance of the signal into the nominal signal region need be considered. To isolate this component, the total yields are found for each of the 6 scale variations (along with the nominal yield), as a function of generated Higgs boson p_{T}^{gen} and reconstructed jet p_{T}, for both the signal selection (n_{sig}) and a fully inclusive selection of events (n_{tot}). The acceptance times efficiency is then defined for a given bin in p_{T}^{gen} as $A\epsilon = n_{sig}/n_{tot}$, also a function of reconstructed jet p_{T}. The envelope of the relative variation in $A\epsilon$ is the acceptance uncertainty due to renormalization and factorization scale, and is shown in Fig. 4.59 for 3 p_{T}^{gen} bins.

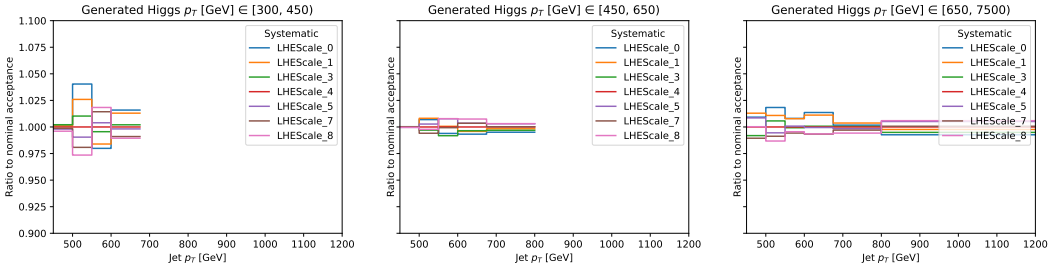


Figure 4.59: The envelope of the relative effect on signal acceptance of the 6 combinations (ignoring $0.5, 2$ and $2, 0.5$) about the nominal ggHbb yield, as a function of reconstructed Higgs boson p_{T}, for three regions in generated Higgs boson p_{T}. The yields are as predicted by the HJ-MiNLO generator, with 2017 detector conditions.

4.8.5 Theory uncertainties for W/Z+jets

Other sub-dominant backgrounds come from inclusive W and Z production. There are corrections to these cross-sections for these processes from NNLO k-factors and also from reweighting the LO p_{T} distribution to NLO. Uncertainties in the W/Z cross-sections are p_{T} dependent due to the uncertainties growing at higher p_{T} from higher order effects. These are correlated per p_{T} bin to account for some p_{T} dependent deviations from the higher order corrections. An additional systematic of 5% is included to account for potential differences between the W and Z higher order corrections.

The re-weighting procedure is discussed in detail in Sec. 4.2.2, and the associated uncertainties in the re-weighting procedures are summarized in Table 4.3.

When considering the process where a vector boson decays to jets $V \to \mathrm{q\bar{q}}$ an ambiguity arises in the selection of the jet. The chosen jet can either be from hadrdonically decaying vector boson or it can arise from one of the jets produced in the sample. Since vector boson events with a $p_{\mathrm{T}} > X$ will have a recoiling jet with a p_{T}

that is roughly equivalent it is not unusual for the selected jet to *not* be from the vector boson. These jets, the unmatched component, consist of QCD jets and should behave more like QCD background jets than tagged boson jets. We let this be treated as QCD background by the fit.

4.8.6 JER and JES

We check the effect of varying the JER and JES by $\pm 1\sigma$ for W/Z+jets, $t\bar{t}$, and the $H(b\bar{b})$ signal processes in each p_T category and apply a log-normal normalization uncertainty equal to the relative size of the effect.

We also include a p_T dependent mass scale uncertainty, according to the variations at high p_T observed in the soft-drop mass corrections, as can be seen in Fig. 3.21, that corresponds to a mass scale shift of $0.5\%/100$ GeV in p_T.

4.8.7 MC statistics

For each process, W/Z+jets, $t\bar{t}$, and the $H(b\bar{b})$ signal processes, we vary each distribution by $\pm 1\sigma$ of the statistical uncertainty (based on the square root of the sum of the squared weights) and apply a log-normal normalization uncertainty equal to the relative size of the effect.

4.8.8 AK4 b-tagging

The efficiency and misidentification rate of the AK4 DeepCSV b-tagging algorithm is measured in data by the BTV-POG [102]. We apply the efficiency and misidentification scale factors as event weights to the MC and generate up/down shape templates by varying the scale factors by $\pm 1\sigma$ of their uncertainties. The relative effect of this source of uncertainty is shown for a representative set of processes in Fig. 4.60

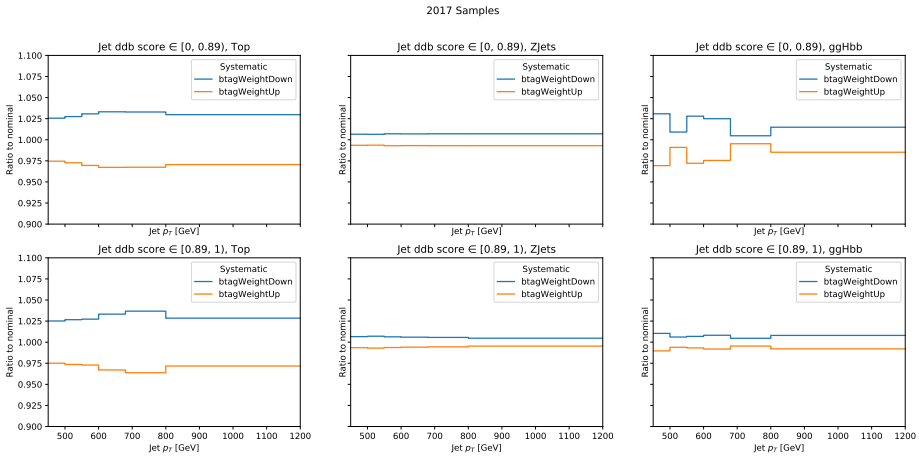


Figure 4.60: Relative effect of the AK4 DeepCSV b-tagging systematic uncertainty on the expected event yield, per jet p_{T} bin in the signal region. The effect is shown separately for events in the AK8 deep double-b passing (bottom) and failing (top) regions. The effect is as measured in a sample of $\mathrm{t\bar{t}}$ production (left) and Z+jets (right) events simulated with 2017 data-taking conditions and corrected by the 2017 BTV-POG scale factors.

4.8.9 Jet trigger

The measured trigger efficiency in data, shown in Sec. 4.3, is applied to the simulation as a function of jet p_{T} and m_{SD}. The statistical uncertainty in the measurement of

the trigger efficiency is propagated into the analysis region, and the relative effect of this uncertainty on the signal region phase space is shown in Fig. 4.61. Based on this result, an overall 2% systematic uncertainty per data-taking year is assigned.

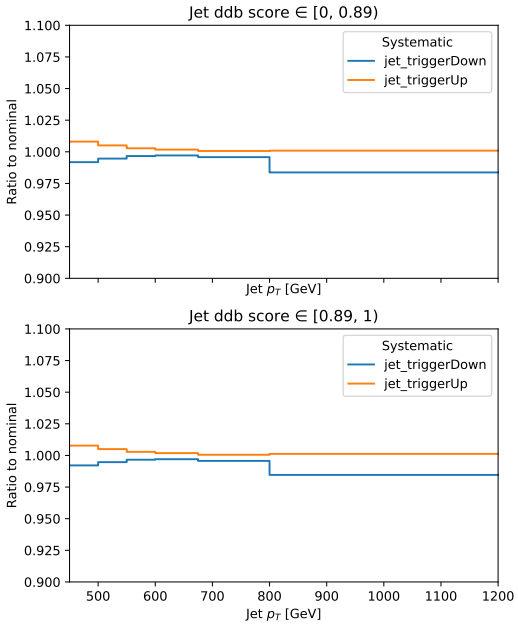
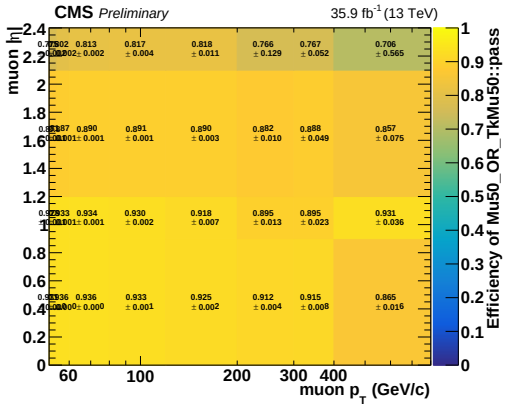


Figure 4.61: Relative effect of the jet trigger systematic uncertainty on the expected event yield, per jet p_{T} bin in the signal region. The effect is shown separately for events in the AK8 deep double-b passing (bottom) and failing (top) regions. The effect is as measured in a sample of Z+jets events simulated with 2017 data-taking conditions and corrected by the 2017 jet trigger efficiency as measured in an independent control sample.

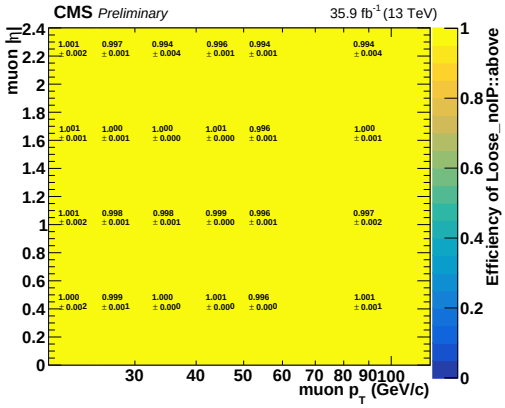
4.8.10 Muon trigger/ID/Isolation efficiency

The muon trigger, loose ID, and loose isolation efficiencies are measured directly in data based on a tag-and-probe method as a function of muon p_{T} and η as indicated

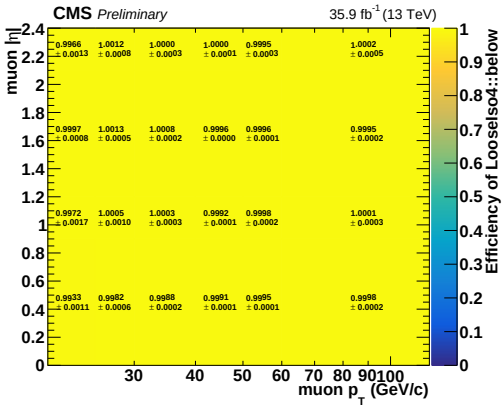
in Ref. [104]. Then, we apply an event-by-event weight (and uncertainty) to the MC samples in the single muon control region depending on the p_{T} and η of the selected muon. The muon trigger efficiency measured in data is applied directly to the MC, whereas the muon loose ID and loose isolation efficiency scale factors are applied to the MC. Figures 4.62-4.64 shows these muon-related efficiencies and uncertainties for 2016, 2017, and 2018, separately.



(a) Muon trigger efficiency in 2016 data

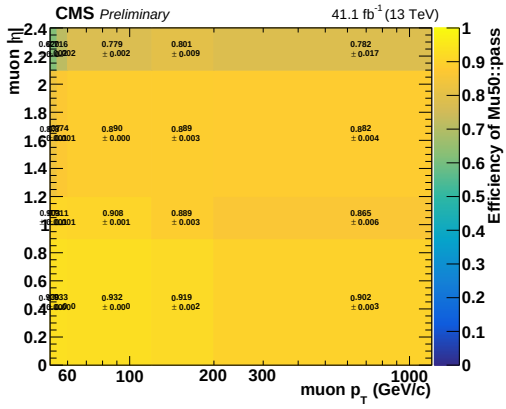


(b) Muon ID efficiency 2016 SF

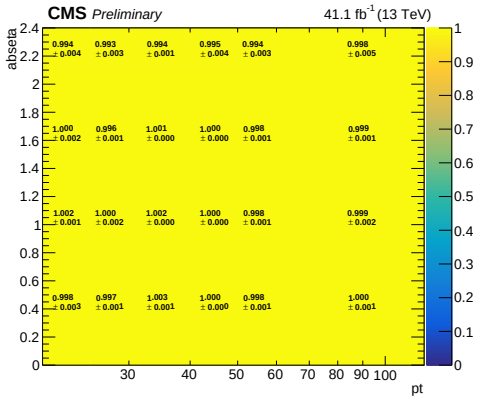


(c) Muon isolation efficiency 2016 SF

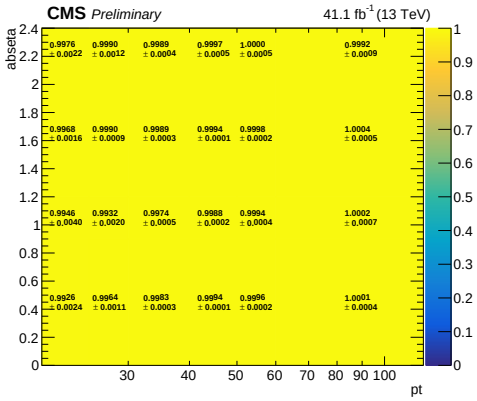
Figure 4.62: 2016 muon trigger efficiency and ID and isolation efficiency SFs (and their uncertainties) measured in data as a function of muon p_{T} and η.



(a) Muon trigger efficiency in 2017 data



(b) Muon ID efficiency 2017 SF

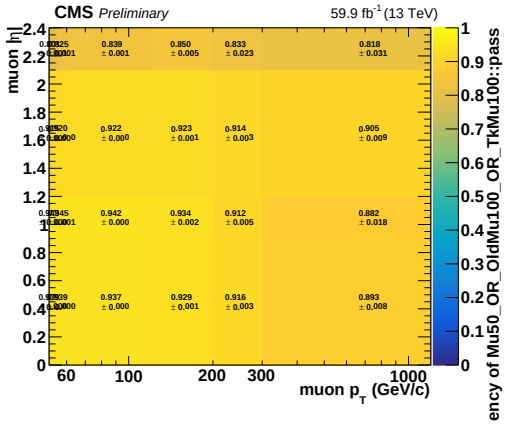


(c) Muon isolation efficiency 2017 SF

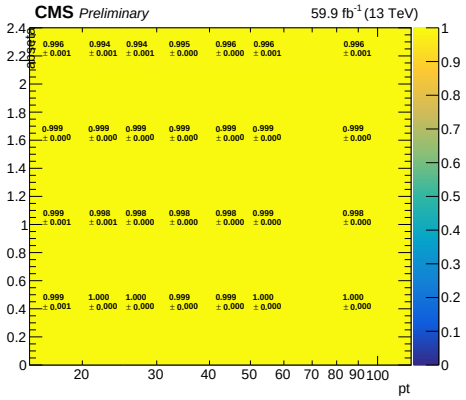
Figure 4.63: 2017 muon trigger efficiency and ID and isolation efficiency SFs (and their uncertainties) measured in data as a function of muon p_{T} and η.

4.8.11 Correlation of systematic across different years

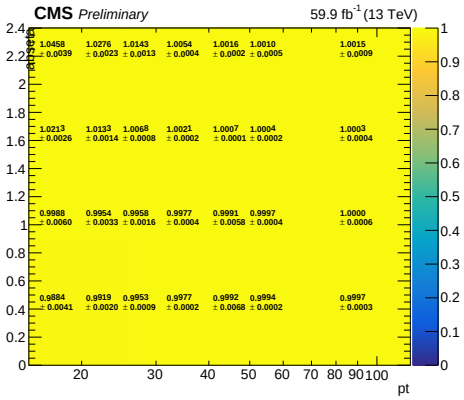
We take a conservative approach towards the correlation of experimental systematic uncertainties across different years. Most of them are un-correlated between different years, such as luminosity, JER, JES, as well as the scale factors measured in-situ and



(a) Muon trigger efficiency in 2018 data



(b) Muon ID efficiency 2018 SF



(c) Muon isolation efficiency 2018 SF

Figure 4.64: 2018 muon trigger efficiency and ID and isolation efficiency SFs (and their uncertainties) measured in data as a function of muon p_{T} and η.

in dedicated control regions. QCD transfer factors are un-correlated between years as well.

Systematic related to theoretical uncertainties are correlated between different years, including the uncertainty k-factors for NLO EWK/QCD corrections for W+jets and Z+jets, as well as those used for Higgs p_{T}-reweighting in gluon fusion production mode.

4.8.12 Summary of systematic uncertainties

A summary of the systematic uncertainties with prior constraints specific to this analysis and their affected processes is tabulated in Table 4.13. A summary of the systematic uncertainties considered in this analysis and their relative size is tabulated in Table 4.14.

Table 4.13: Summary of nuisance parameters with prior constraints specific to this analysis and their affected processes.

Description	Affected process	Correlated between years?	Remarks
Normalization of the ggH signal, uniform in p_{T}	ggHbb only	Yes	30% for HIG-17-010, 20% for MiNLO
p_{T} dependence of the ggH signal	ggHbb only	Yes	30% for HIG-17-010, none for MiNLO
NLO QCD k-factor for W and Z	W,Z	Yes	10% [79]
NLO EW k-factor for W and Z	W,Z	Yes	5-7%, depending on p_{T} [79]
Extra corrections on EW k-factor for W	W	Yes	2-6%, depending on p_{T} [79]
Soft-drop mass scale	W,Z, all H signals	No	1.2% proportional to resonance mass
p_{T} dependence of soft-drop mass scale	W,Z, all H signals	No	0.5% per 100 GeV in p_{T}
Soft-drop mass resolution	W,Z, all H signals	No	
$N_2^{1,\mathrm{DDT}}$ scale factor uncertainty	W,Z, all H signals	No	
deep double-b scale factor uncertainty	Z, all H signals	No	

Table 4.14: Major sources of uncertainty in the measurement of the signal strength μ_H based on the HJ-MiNLO prediction, and their observed impact ($\Delta\mu_H$) from a fit to the combined data set, are listed. The total uncertainty is separated into three components: statistical, systematic, and theory. Detailed decompositions of the statistical, systematic, and theory components are specified. The impact of each uncertainty is evaluated considering only that source. The sum in quadrature for each source does not in general equal the total uncertainty of each component because of correlations in the combined fit between nuisance parameters in different sources.

Uncertainty source	$\Delta\mu_H$	
Statistical	+1.20	−1.20
QCD pass-fail ratio (data correction)	+0.71	−0.70
$t\bar{t}$ normalization and misidentification	+0.36	−0.36
Systematic	+0.63	−0.66
QCD pass-fail ratio (simulation)	+0.48	−0.52
Deep double-b tag efficiency	+0.32	−0.41
Jet mass scale and resolution	+0.11	−0.11
Jet energy scale and resolution	+0.32	−0.29
Simulation sample size	+0.14	−0.15
Other experimental uncertainties	+0.07	−0.09
Theory	+0.81	−0.46
V+jets modeling	+0.62	−0.38
H modeling	+0.54	−0.27
Total	+1.58	−1.45

4.9 Results

4.9.1 Mass spectrum

The full simultaneous fits are performed with the respective polynomial orders chosen in Sec. 4.7.1 on data set collected in different years. There are 21 different categories (6 p_T categories and a muon control region for each year, each with a pass and a fail region), in the full simultaneous fit.

To visualize the fit, we show the m_{SD} distribution summed across different data-taking years and/or p_T categories. Figures 4.65 and 4.66 show the same post-fit

distributions combining the three data-taking years, but keeping the p_{T} categories separate. Figure 4.67 shows the combination of all years and p_{T} categories. Figure 4.68 shows the 2D fitted transfer factor for each year, whereas fig. 4.69 shows the QCD mass sculpting shapes used in the corresponding fits for each year. The full post-fit m_{SD} distributions for 2016, 2017 and 2018 separately are shown in Appendix B.

As will be discussed in detail in Sec. 4.7, the normalization for the $\mathrm{Z}(\mathrm{b\bar{b}})$ process in the fit is fixed at the SM value, with associated uncertainties, to provide a better constraint on the Higgs signal strength. This choice is motivated by the fact that we have a relatively loose prior constraint on deep double-b scale factor uncertainty, and that we have sufficient $\mathrm{Z}(\mathrm{b\bar{b}})$ events in the signal region to constrain the scale factor in-situ.

4.9.2 Goodness of fit in data

In Fig. 4.70, the results of goodness-of-fit (GOF) tests in data are reported. As in the F-test, we use the "saturated model" GOF test statistic [100, 101], which asymptotically follows a chi-squared distribution with $n_{\mathrm{bins}} - p_i$ degrees of freedom. In the presence of a signal strength considerably larger than 1, we find the overall GOF is slightly less than 5%, indicating a small tension between the data and the fit result.

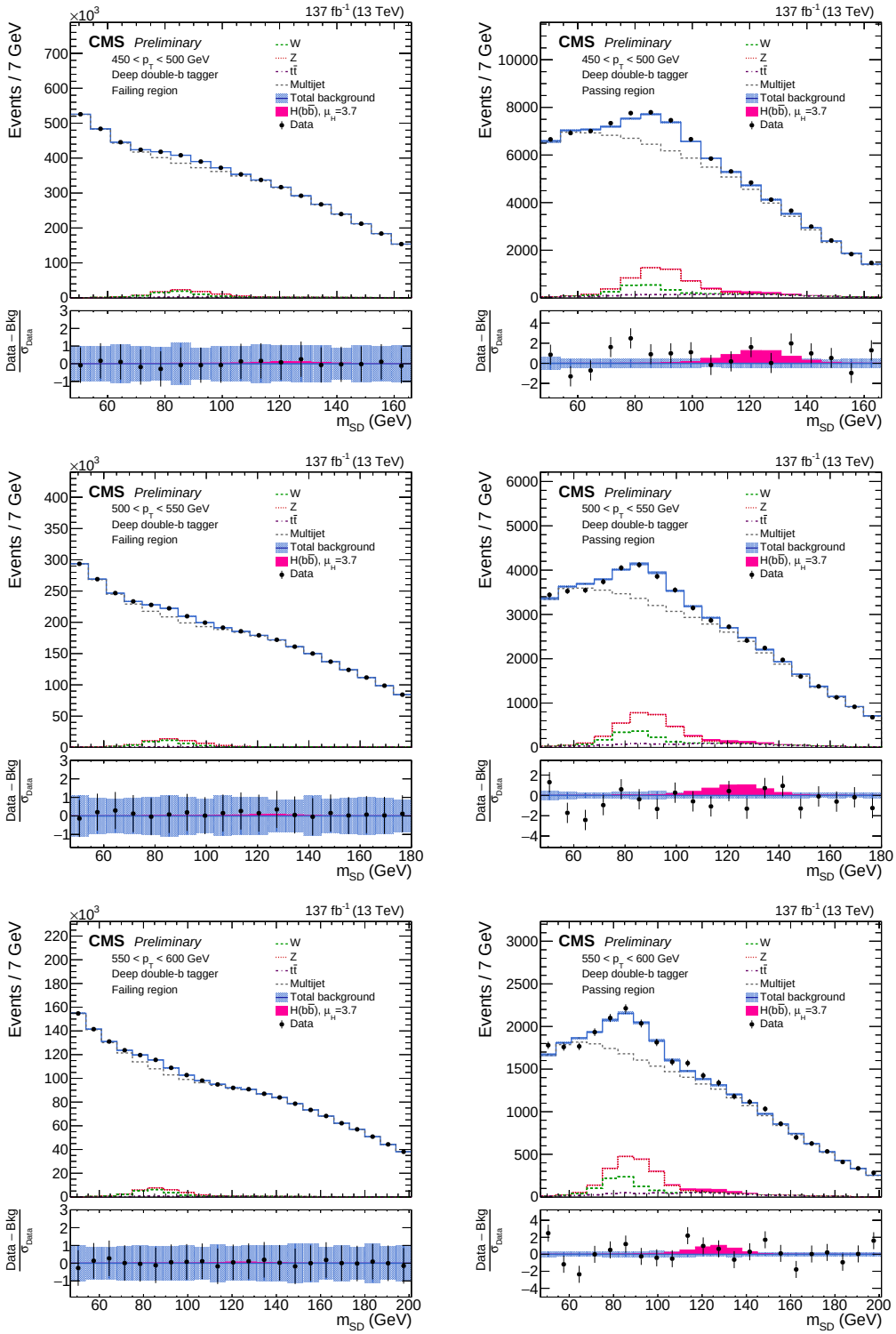


Figure 4.65: Post-fit m_{SD} distributions in the full 2016, 2017, and 2018 data set for the pass and fail regions and the first three p_{T} categories (450-500,500-550,550-600 from top to bottom) summed over all three years. The EW and signal events are also shown separately.

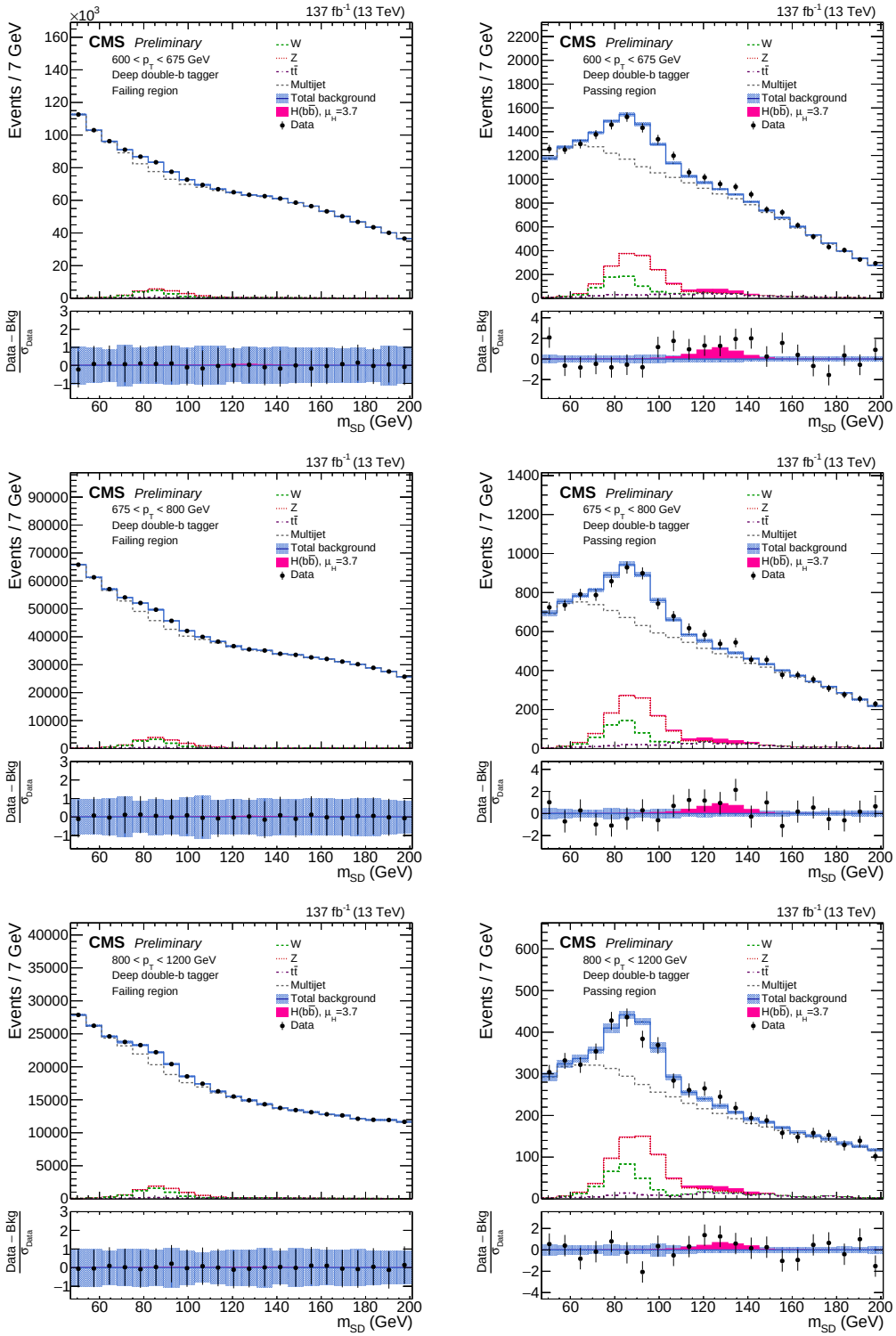


Figure 4.66: Post-fit m_{SD} distributions in the full 2016, 2017, and 2018 data set for the pass and fail regions and the second three p_{T} categories (600-675,675-800,800-1200 from top to bottom) summed over all three years. The EW and signal events are also shown separately.

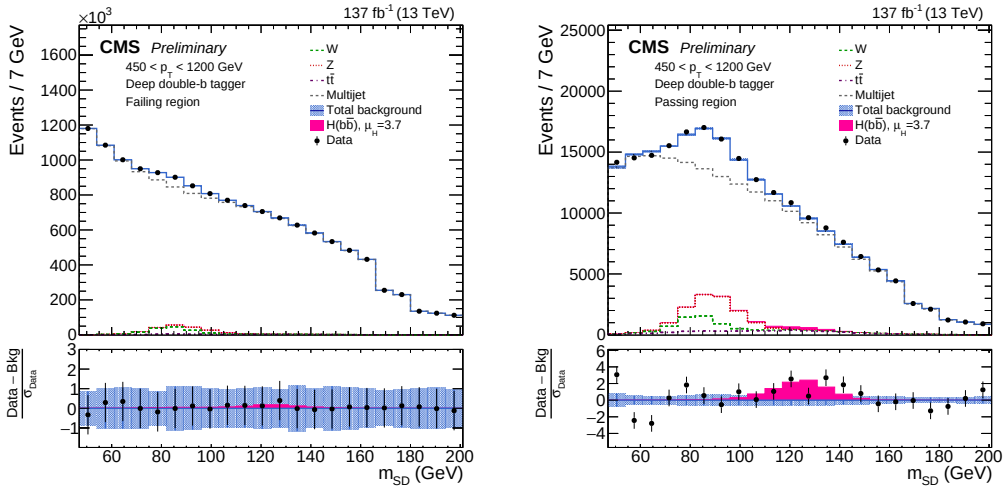


Figure 4.67: Post-fit m_{SD} distributions in the full Run II data set for the pass and fail regions. The EW and signal events are also shown separately.

4.9.3 Profile Likelihoods

As discussed in Sec. 4.7, profile likelihood is used to extract parameter of interests and corresponding uncertainties from the likelihood

First, we determine if signal strength of $\mathrm{Z}(\mathrm{b\bar{b}})$ could improve the sensitivity of the H signal strength. Fig. 4.71 shows the 1D profile test statistics scan fixing $\mu_{\mathrm{Z}} = 1$ (left) and floating μ_{Z} (right) with Asimov data set corresponding to full Run 2 luminosity. As expected, fixing $\mu_{\mathrm{Z}} = 1$ indirectly constrained the μ_{H} better due to the reduced uncertainty in the deep double-b scale factors.

Figure 4.72 shows the 1D and 2D profile likelihood test statistic scans as a function of μ_{H} and μ_{Z} on data. The 1D scan for μ_{H} shows the nominal result with μ_{Z} fixed at 1 and from the scan, it is clear that the uncertainty on μ_{H} is statistically dominated. The 1D scan for μ_{Z} and 2D $\mu_{\mathrm{H}}, \mu_{\mathrm{Z}}$ scan serve as cross check and robustness of the minimum of the nominal result. With μ_{Z} minimizing the likelihood at 1, this gives

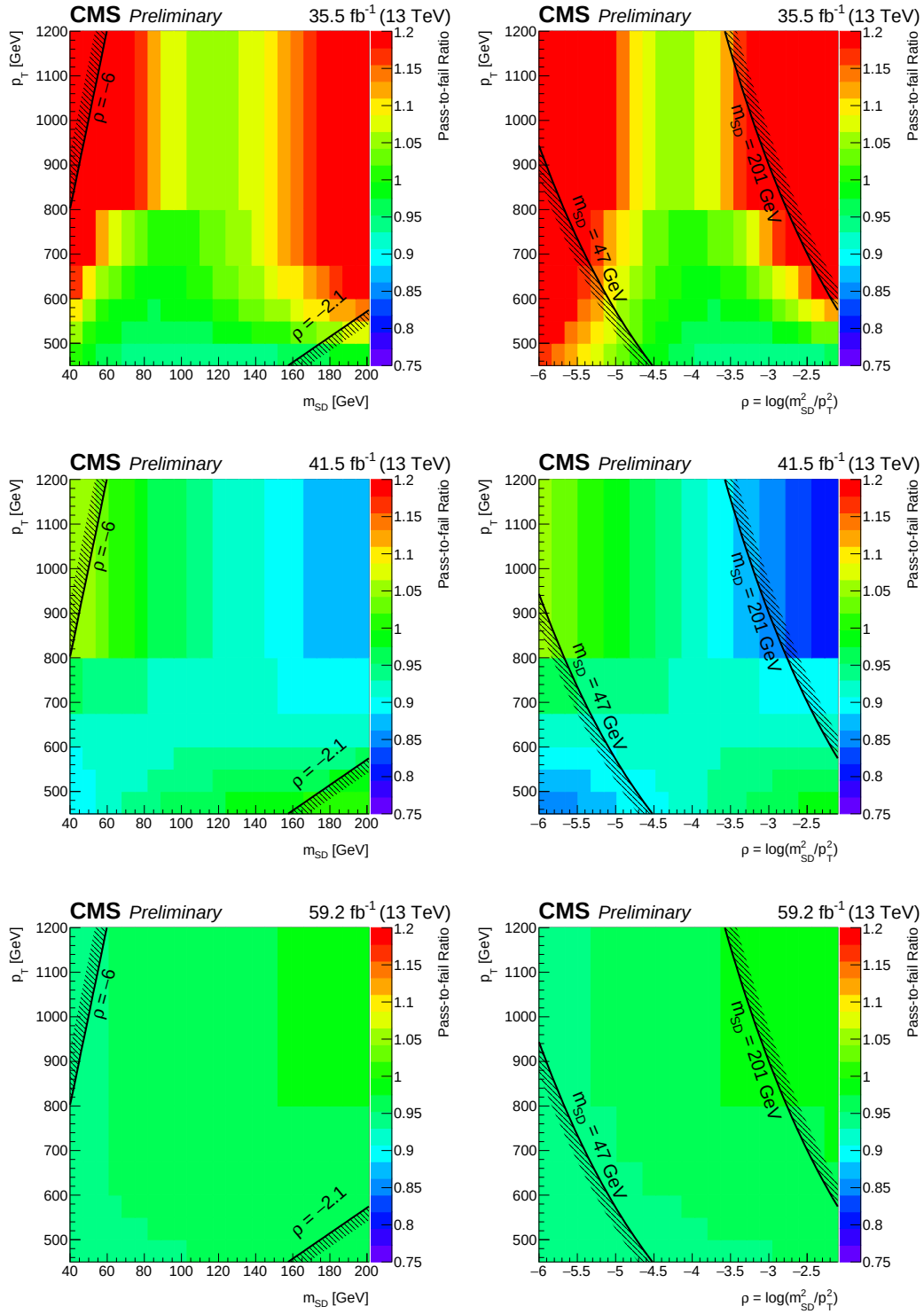


Figure 4.68: The fitted data-mc transfer function TF^{data} as a function of p_{T} and m_{SD} (left) or p_{T} and ρ (right) for 2016 (top), 2017 (middle), and 2018 (bottom) data. The polynomial orders of used these functions are 1st order in p_{T} and 1st order in ρ for 2017 and 2018, while the order in p_{T} and ρ are 2nd order and 1st order for 2016 respectively.

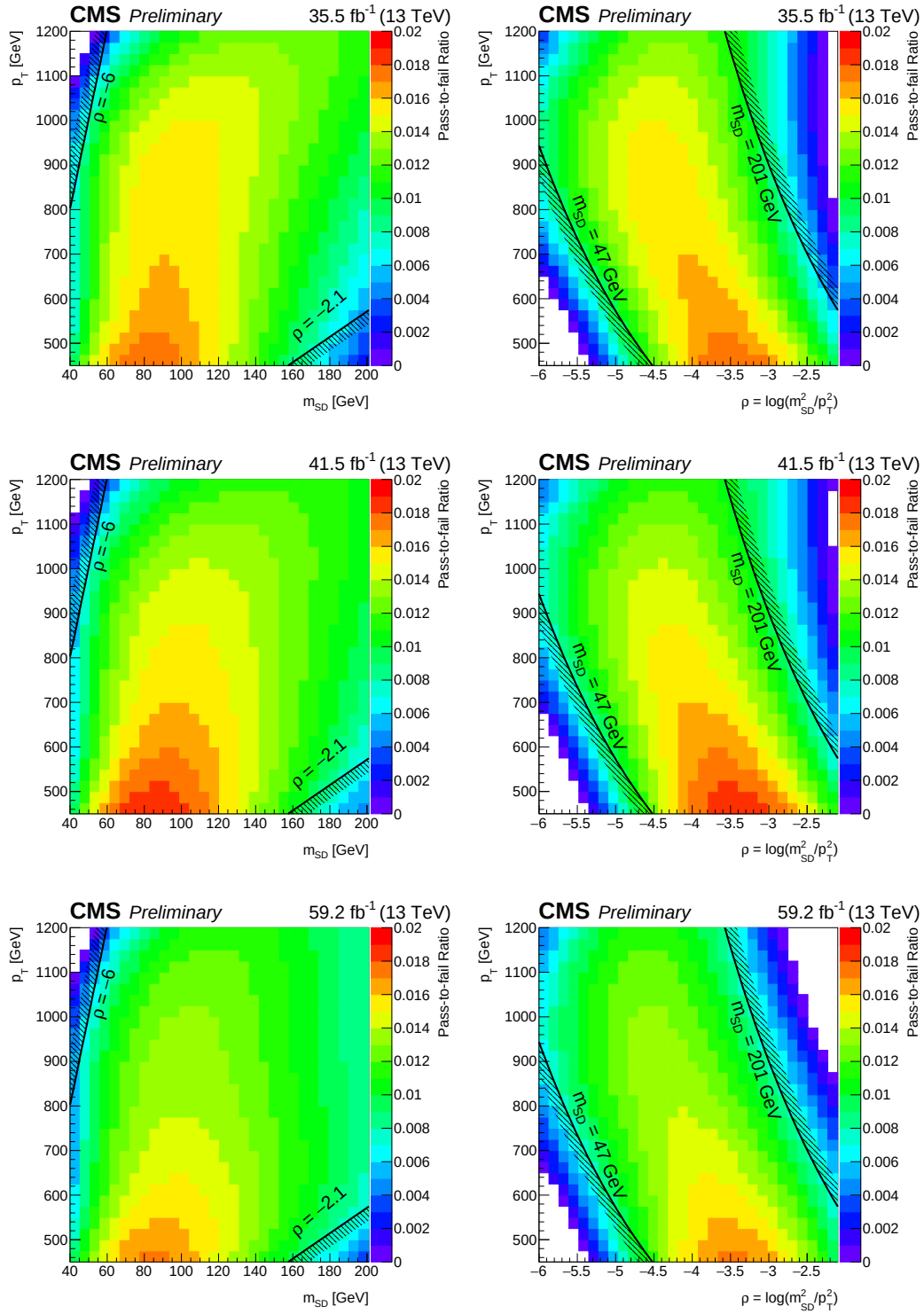
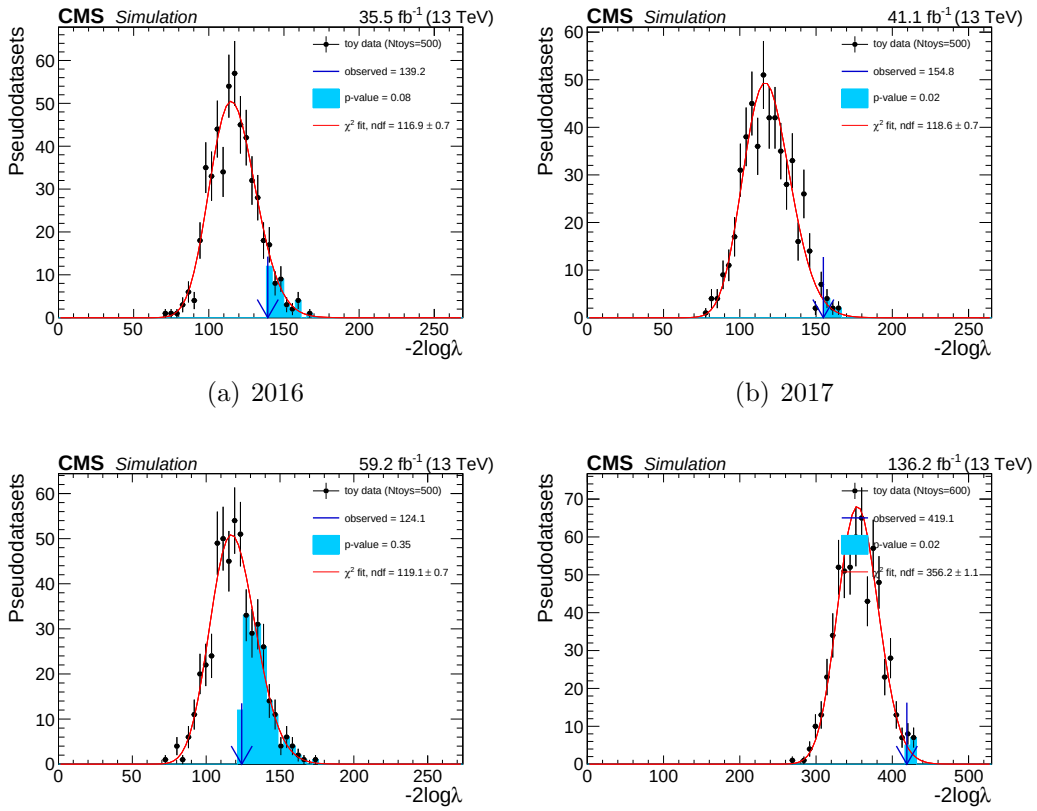


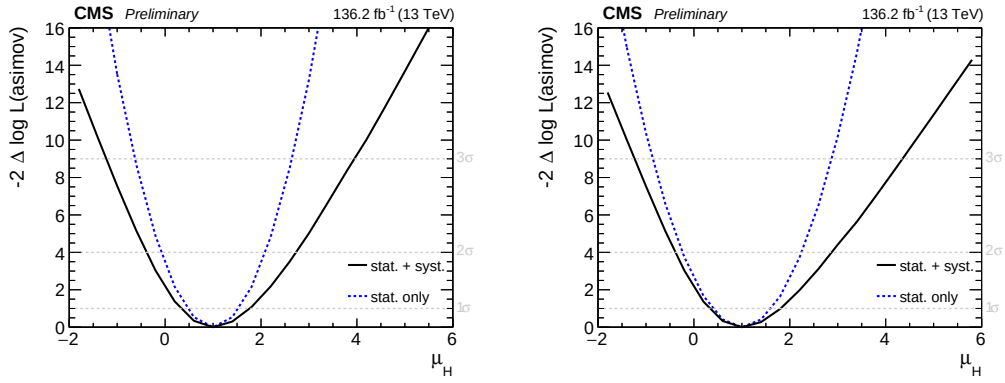
Figure 4.69: The fitted pass-fail ratio $\epsilon^{\mathrm{QCDMC}}$ as a function of p_{T} and m_{SD} (left) or p_{T} and ρ (right) for QCD MC corresponding to conditions in 2016 (top), 2017 (middle), and 2018 (bottom). The polynomial orders of used these functions are 2nd order in p_{T} and 2nd order in ρ.



(a) 2016 (b) 2017

(c) 2018 (d) 2016+2017+2018

Figure 4.70: Results of the goodness-of-fit test for the full (a) 2016 data set , (b) 2017 data set, (c) 2018 data set and (d) full Run 2 data set using the saturated test statistic. The order of polynomial used are second(first) order in ρ and first(first) order in p_{T} for all 2016(2017,2018) data sets.



(a) μ_{H} scan, $\mu_{\mathrm{Z}} = 1$ (b) μ_{H} scan, μ_{Z} =float

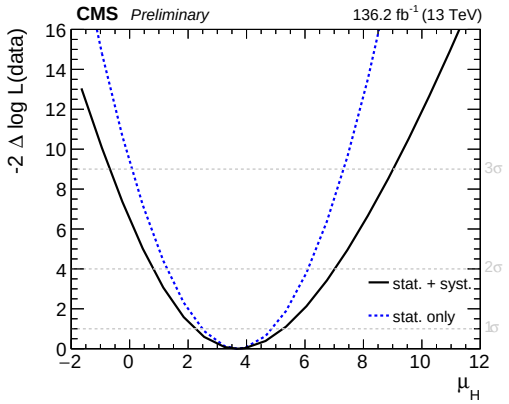
Figure 4.71: 1D profile test statistic $-2\Delta\log(\mathcal{L})$ scans as a function of the signal strengths μ_{H}, fixing $\mu_{\mathrm{Z}} = 1$ (left) and floating μ_{Z} (right), with Asimov data set for 2016+2017+2018. The solid black curve, labeled "stat+syst," corresponds to the scan when profiling all nuisance parameters and the dotted blue curve, labeled "stat only," corresponds to only profiling the unconstrained nuisance parameters, such as the QCD transfer factor parameters.

a good indication that the $\mathrm{Z}(\mathrm{b\overline{b}})$ process in data is well modeled by the analysis method and it remains to be so when we are profiling the μ_{H} simultaneously.

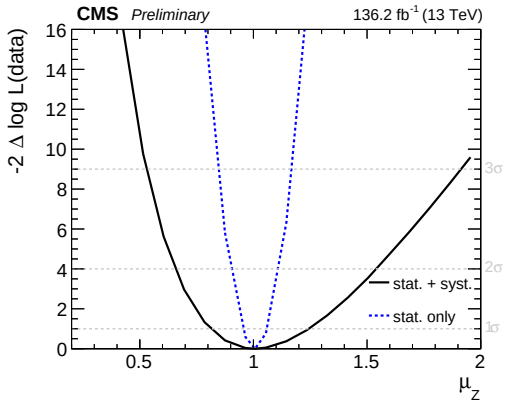
4.9.4 Signal strengths and significance

To validate the substructure and b tagging techniques employed in this search, a maximum likelihood fit is performed using a model where the Z and H signal strengths are left unconstrained. The measured Z boson signal strength is $\mu_{\mathrm{Z}} = 1.01 \pm 0.05\,(\mathrm{stat})^{+0.20}_{-0.15}\,(\mathrm{syst})^{+0.13}_{-0.09}\,(\mathrm{theo})$. This demonstrates that the Z boson is clearly separable from the background.

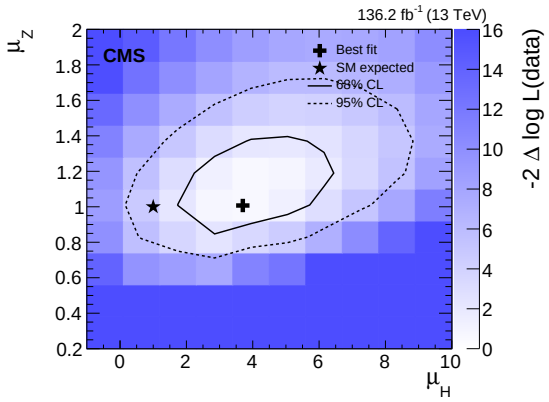
We then proceed with the extraction of Higgs signal strength. The measured Higgs boson signal strength is $\mu_{\mathrm{H}} = 3.68 \pm 1.20\,(\mathrm{stat})^{+0.63}_{-0.66}\,(\mathrm{syst})^{+0.81}_{-0.46}\,(\mathrm{theo})$ using the



(a) μ_{H}, 2016+2017+2018



(b) μ_{Z}, 2016+2017+2018



(c) $\mu_{\mathrm{Z}}, \mu_{\mathrm{H}}$2016+2017+2018

Figure 4.72: 1D profile and 2D likelihood test statistic $-2\Delta\log(\mathcal{L})$ scans as a function of the signal strengths μ_{H} and μ_{Z} on data for 2016+2017+2018. In the 1D profile, the solid black curve, labeled "stat+syst," corresponds to the scan when profiling all nuisance parameters and the dotted blue curve, labeled "stat only," corresponds to only profiling the unconstrained nuisance parameters, such as the QCD transfer factor parameters.

HJ-MiNLO sample as the gluon fusion signal model. The corresponding observed and expected upper limits on the Higgs boson signal strength at the 95% confidence level are 3.70 and 2.90, respectively, while the observed and expected significance with respect to the background-only hypothesis are 2.54σ and 0.71σ. The measurement exhibits a clear excess over the SM expectation ($\mu_{\mathrm{H}} = 1$), with a significance of 1.85σ.

Table 4.15 summarizes the measured signal strengths and significance for the Higgs and Z boson processes. As a cross-check, the results are also presented with the ggF signal p_{T} spectrum used in the previous version of this analysis. As discussed in Sec. 4.2.3 and shown in Fig. 4.4, the cross sections predicted by the previous p_{T} spectrum is approximately 50% of the latest HJ-MiNLO prediction, as a results, the signal strengths and other results scales accordingly.

4.9.5 Channel compatibility

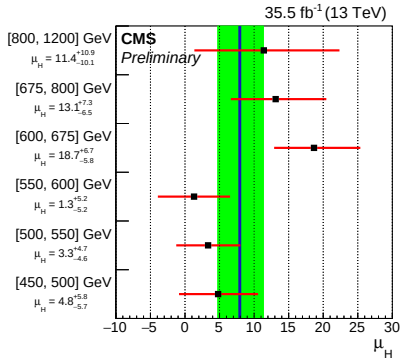
To assess the compatibility between the observed signal strengths in the different jet p_{T} categories, an alternative fit to data is performed. In this fit, an independent H signal strength is assigned to each of the six reconstructed jet p_{T} bins. These signal strengths are unconstrained in the fit and are varied simultaneously. All other parameters are profiled, as in the nominal fit. Figure 4.73 illustrates the compatibility in the best fit signal strength between the different p_{T} categories in different data taking periods. With the full Run 2 dataset, we see a small trend for the signal strength to increase with the p_{T} of the categories and this trend is similarly consistent in different data taking periods. The combined result using full Run 2 dataset is

Table 4.15: Fitted signal strength, expected and observed significance of the Higgs and Z boson signals. The Higgs boson results are presented with two gluon fusion signal models, one using the nominal HJ-MiNLO sample and the other simulated with the same procedure described in Ref. [1]. The 95% confidence level upper limit (UL) on the Higgs boson signal strength is also listed. In the results for the Higgs boson, the Z boson yield is fixed to the SM prediction value with the corresponding theoretical uncertainties to better constrain *in situ* the data-to-simulation scale factor of the deep double-b tagger. For the expected and observed signal strength of Z boson, the Higgs boson signal strength is freely floating.

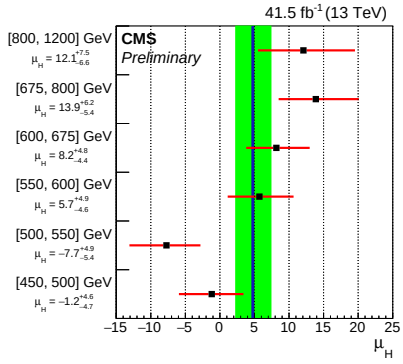
	2016	2017	2018	Combined
Expected signal strength μ_Z	$1.00^{+0.38}_{-0.28}$	$1.00^{+0.42}_{-0.29}$	$1.00^{+0.43}_{-0.29}$	$1.00^{+0.23}_{-0.19}$
Observed signal strength μ_Z	$0.86^{+0.32}_{-0.24}$	$1.11^{+0.48}_{-0.33}$	$0.91^{+0.37}_{-0.26}$	$1.01^{+0.24}_{-0.20}$
HJ-MiNLO [105, 86]				
Expected signal strength μ_H	$1.00^{+3.34}_{-3.54}$	$1.00^{+2.46}_{-2.50}$	$1.00^{+2.30}_{-2.37}$	$1.00^{+1.42}_{-1.40}$
Observed signal strength μ_H	$7.92^{+3.42}_{-3.19}$	$4.76^{+2.64}_{-2.47}$	$1.70^{+2.31}_{-2.33}$	$3.68^{+1.58}_{-1.46}$
Expected H significance ($\mu_H = 1$)	0.29σ	0.41σ	0.43σ	0.71σ
Observed H significance	2.42σ	1.90σ	0.73σ	2.54σ
Expected UL signal strength ($\mu_H = 0$)	< 6.84	< 5.02	< 4.67	< 2.90
Observed UL signal strength	< 7.97	< 4.78	< 1.70	< 3.70
Ref. [1] H p_T spectrum				
Expected signal strength μ_H	$1.00^{+1.52}_{-1.50}$	$1.00^{+1.14}_{-1.06}$	$1.00^{+1.12}_{-1.05}$	$1.00^{+0.68}_{-0.58}$
Observed signal strength μ_H	$3.97^{+1.89}_{-1.55}$	$2.18^{+1.37}_{-1.19}$	$1.13^{+1.12}_{-1.06}$	$1.90^{+0.86}_{-0.70}$
Expected H significance ($\mu_H = 1$)	0.68σ	0.94σ	0.95σ	1.73σ
Observed H significance	2.62σ	1.84σ	1.06σ	2.85σ
Expected UL signal strength ($\mu_H = 0$)	< 3.41	< 2.42	< 2.29	< 1.42
Observed UL signal strength	< 4.00	< 2.18	< 1.13	< 1.90

shown in Fig. 4.76. We perform a similar compatibility checks for Z signal strengths on different p_{T} categories on Figure 4.74 and did not find similar trend to be present with μ_{Z}.

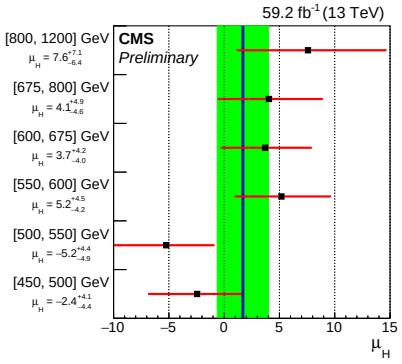
Similarly, we also checked the compatibility across different data taking periods with an independent signal strength scaling with different data sets. Figure 4.75 illustrates the compatibility in the best fit signal strength between the different data taking periods.



(a) 2016

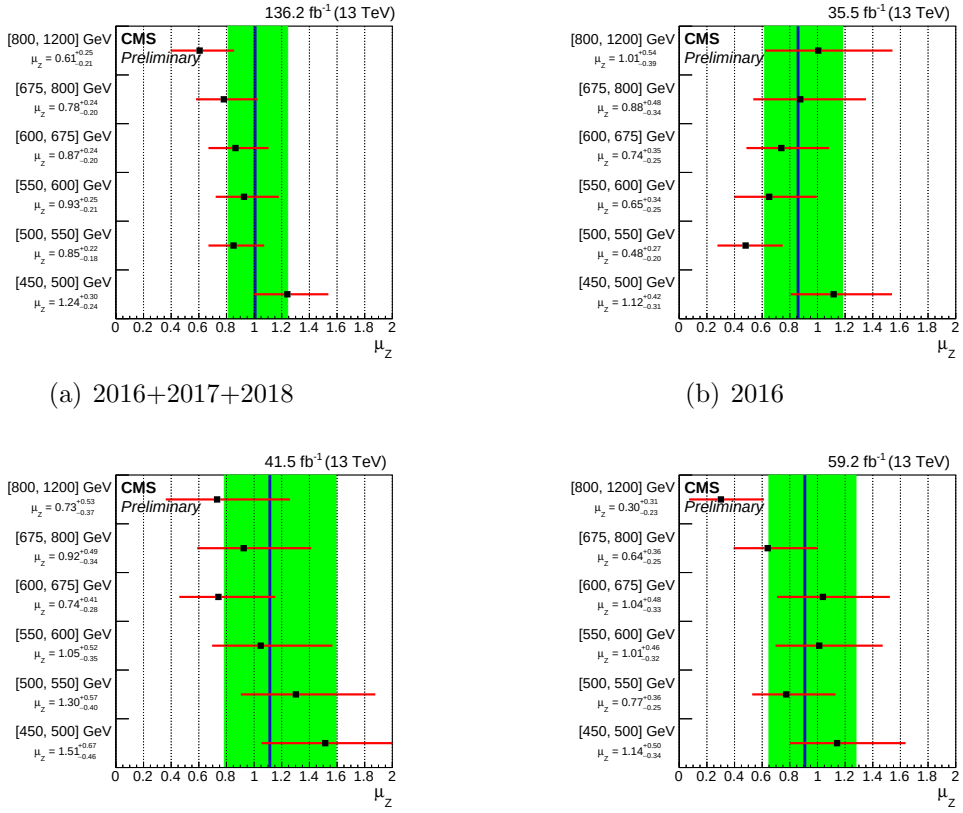


(b) 2017



(c) 2018

Figure 4.73: The best-fit H signal strength μ_{H} and uncertainty per p_{T} category extracted from a simultaneous fit of all channels, performed on different data taking periods.



(a) 2016+2017+2018 (b) 2016

(c) 2017 (d) 2018

Figure 4.74: The best-fit Z signal strength μ_{Z} and uncertainty per p_{T} category extracted from a simultaneous fit of all channels, performed on different data taking periods. The H signal strength μ_{H} is also floating in these fits.

4.9.6 Fiducial cross section

To facilitate comparisons with theoretical predictions, we isolate and remove the effects of limited detector acceptance and response to the ggF Higgs boson production cross section using a maximum-likelihood unfolding technique. In our treatment, the remaining Higgs boson production modes are assumed to occur at SM rates, as predicted by the HJ-MiNLO simulation. The ggF Higgs boson signal is split into several

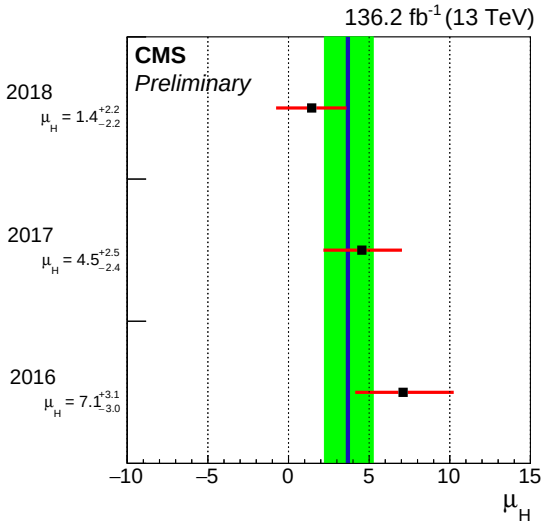
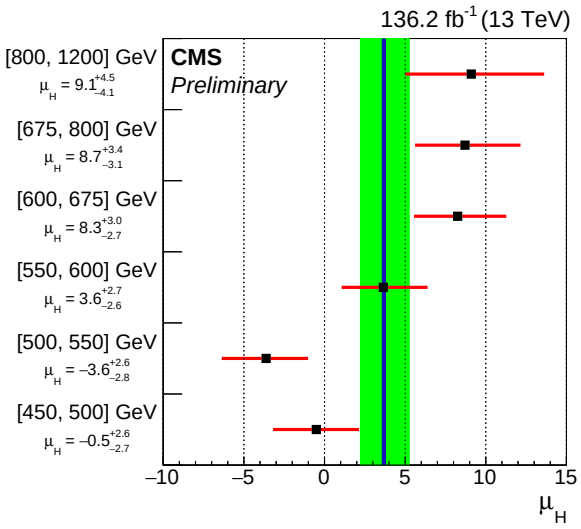


Figure 4.75: The best-fit signal strength μ_{H} and uncertainty per year, extracted from a simultaneous fit of all channels.



(a) 2016+2017+2018

Figure 4.76: The best-fit H signal strength μ_{H} and uncertainty per p_{T} category extracted from a simultaneous fit of all channels using full Run 2 dataset.

bins according to the generated Higgs boson p_{T} ($p_{\mathrm{T}}^{\mathrm{H}}$), and each $p_{\mathrm{T}}^{\mathrm{H}}$ bin is considered as a separate process with a freely floating signal strength parameter in the likelihood

model. The respective $p_{\mathrm{T}}^{\mathrm{H}}$ bins are 300–450, 450–650, and $> 600\,\mathrm{GeV}$. This binning choice follows the simplified template cross section (STXS) recommendation [106]. As the minimum reconstructed jet p_{T} is 450 GeV, a negligible signal contribution is expected from events with $p_{\mathrm{T}}^{\mathrm{H}} < 300\,\mathrm{GeV}$. The fiducial cross section in each $p_{\mathrm{T}}^{\mathrm{H}}$ bin is then extracted by scaling the cross section found in simulation, imposing no selection requirements other than those on $p_{\mathrm{T}}^{\mathrm{H}}$, by the corresponding signal strength parameter. The uncertainty in this value is taken from the correspondingly scaled signal strength uncertainty. For the theoretical uncertainties, only those that affect the acceptance of signal events into the reconstructed selection are taken into account. Based on the envelope of acceptance values from varying the renormalization and factorization scales by factors of two, this theoretical acceptance uncertainty is estimated to be 2%.

The result of this unfolding procedure is shown in Fig. 4.78 and Table 4.16, along with the predicted cross sections from Ref. [86] and the predictions of the signal event generators described in Section 4.2. The correlation coefficients among the three $p_{\mathrm{T}}^{\mathrm{H}}$ bins are shown in Table 4.17. The measured cross section uncertainty in the first $p_{\mathrm{T}}^{\mathrm{H}}$ bin is larger because of limited acceptance. The first and second $p_{\mathrm{T}}^{\mathrm{H}}$ bins have a mild anti-correlation, primarily due to jet energy scale uncertainties, which inflates the corresponding per-bin uncertainties in the unfolded cross section. The observed cross section in the third $p_{\mathrm{T}}^{\mathrm{H}}$ bin has a smaller relative uncertainty than that in the second bin because of the larger magnitude of the central value in that bin. With respect to the SM, the upward deviation of the cross section in the third $p_{\mathrm{T}}^{\mathrm{H}}$ bin, when profiling

the other two, corresponds to a significance of 2.64σ. Instead, when considering all three cross section parameters of interest simultaneously, the total deviation from the SM corresponds to a significance of 1.91σ.

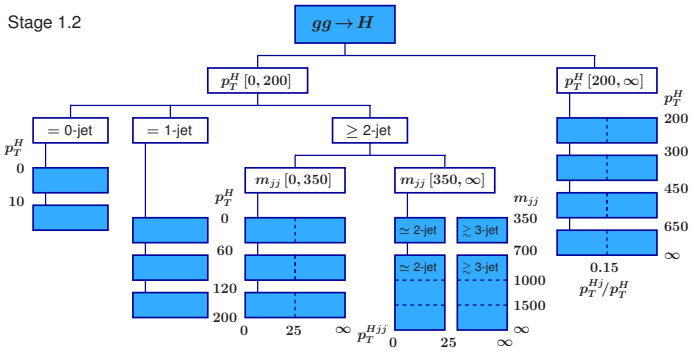


Figure 4.77: STXS Stage 1.2 binning.

Table 4.16: Measured and predicted differential fiducial cross section as a function of Higgs boson p_{T}. All cross sections are in units of fb.

	300–450 GeV	450–650 GeV	> 650 GeV
Measured	600^{+800}_{-800}	5^{+42}_{-43}	29^{+12}_{-11}
	$\pm 700\,(\mathrm{stat}) \pm 350\,(\mathrm{syst})$	$\pm 37\,(\mathrm{stat}) \pm 22\,(\mathrm{syst})$	$\pm 9\,(\mathrm{stat}) \pm 7\,(\mathrm{syst})$
LHCHXSWG [86]	—	$15.4^{+1.4}_{-1.8}$	$1.9^{+0.2}_{-0.2}$
HJ-MiNLO [105]	89^{+20}_{-18}	$13.5^{+3.0}_{-2.7}$	$1.9^{+0.4}_{-0.4}$

Table 4.17: Correlation coefficients between the three $p_{\mathrm{T}}^{\mathrm{H}}$ bins of the unfolded Higgs boson differential cross section measurement.

	300–450 GeV	450–650 GeV	> 650 GeV
300–450 GeV	1.0	−0.18	−0.002
450–650 GeV	−0.18	1.0	0.06
> 650 GeV	−0.002	0.06	1.0

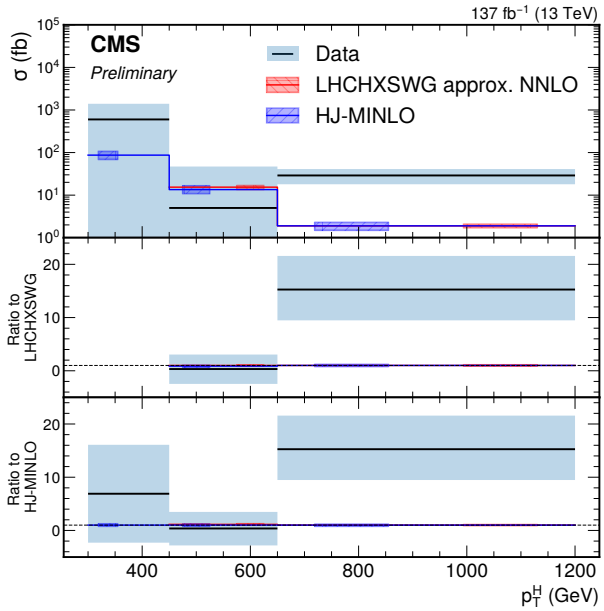
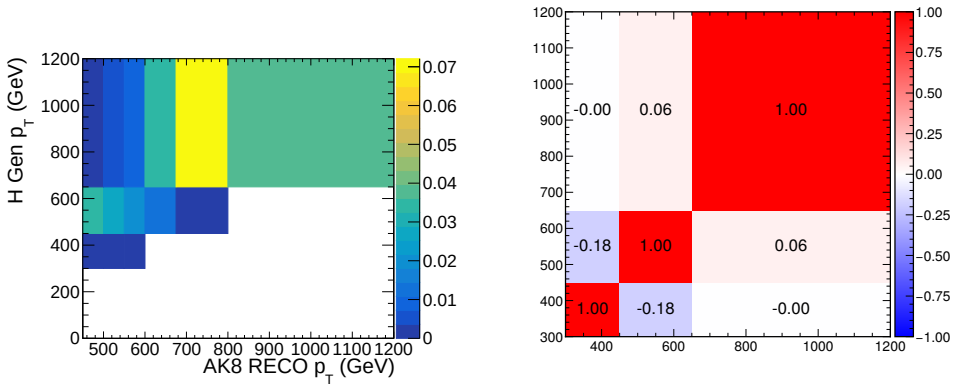


Figure 4.78: Measured differential fiducial cross section as a function of Higgs boson p_{T}, in comparison to the predictions of Ref. [86] and HJ-MiNLO [105]. In the bottom two panels, the dotted line corresponds to a ratio of one. The measured cross section uncertainty in the first bin is larger because of limited acceptance. The cross section measurements in the first and second bins have a mild anti-correlation, primarily due to jet energy scale uncertainties, which inflates the corresponding uncertainties. The observed cross section in the third bin has a smaller relative uncertainty than that in the second bin because of the larger magnitude of the central value in that bin. The relative uncertainties in the predictions of Ref. [86] and HJ-MiNLO are approximately 10% and 20%, respectively.



(a) Response matrix (b) Correlation matrix

Figure 4.79: (a) Response matrix corresponding to the STXS binning of the generated Higgs p_{T}. Reconstructed p_{T} binning is fixed to the p_{T} categories defined in the analysis. The z-axis value quantifies the fraction of events in a given reconstructed p_{T} bin for the given generated p_{T} bin, normalized to the total production rate in the generated p_{T} bin. (b) Correlation between per-p_{T}-bin signal strength parameters corresponding to the STXS binning of the generated Higgs p_{T}.

4.10 Summary

We presented an inclusive search for the standard model Higgs boson decaying to bottom quark-antiquark pairs and reconstructed as a single, large-radius jet with $p_{\mathrm{T}} > 450\,\mathrm{GeV}$. The search uses the LHC full Run II dataset of proton-proton collisions at $\sqrt{s} = 13\,\mathrm{TeV}$ corresponding to an integrated luminosity of $137\,\mathrm{fb}^{-1}$. We validated the analysis method with Z+jets process, which is measured to be consistent with the standard model prediction.

With the latest boosted Higgs p_{T} spectrum prediction from the HJ-MiNLO generator, which is compatible with theoretical calulation with fullfinite top-mass correction at NLO, we measured the Higgs boson signal strength to be $\mu_{\mathrm{H}} = 3.68 \pm 1.20\,(\mathrm{stat})^{+0.63}_{-0.66}\,(\mathrm{syst})^{+0.81}_{-0.46}\,(\mathrm{theo}) = 3.68^{+1.58}_{-1.46}$. The measured signal strength corresponds to an observed significance of 2.54σ with respect to the background-only hypothesis, where the expected significance of the standard model signal is 0.71σ. The significance of the observed excess with respect to the standard model expectation is 1.85σ.

With respect to the previous CMS result, the relative precision of the Higgs boson signal strength measurement improves by approximately a factor of two because of the increased integrated luminosity, improved b tagging based on a deep neural network, and smaller theoretical uncertainties. Finally, the differential cross section for the Higgs boson transverse momentum in the phase space regions recommended by the LHC simplified template cross section framework has also been presented.

Appendix A

List of triggers

Table A.1: A description of the triggers used in this analysis.

Trigger Name	Definition
HLT_PFHT[800/900/1050]	AK4PFJets with $p_{\mathrm{T}} > 30\,\mathrm{GeV}$ and $\lvert\eta\rvert < 2.5$ $H_{\mathrm{T}} > 800/900/1050\,\mathrm{GeV}$
HLT_PFHT650_ _WideJetMJJ[900/950]DEtaJJ1p5	AK4PFJets with $p_{\mathrm{T}} > 150\,\mathrm{GeV}$ and $\lvert\eta\rvert < 2.5$ $AK8HT > 650$ GeV two jets with mass $> 900/950\,\mathrm{GeV}$ and $\Delta\eta < 1.5$
HLT_AK8PFJet[360/400/420] _TrimMass30	AK8PFJets with one jet $p_{\mathrm{T}} > 360/400/420\,\mathrm{GeV}$ and $\lvert\eta\rvert < 2.5$ one jet with trimmed mass $> 30\,\mathrm{GeV}$ trimming parameters: $r = 0.1$ and $p_{\mathrm{T\ frac}} = 0.3$
HLT_AK8PFHT700_ _TrimR0p1PT0p3Mass50_v*	AK8PFJets with $p_{\mathrm{T}} > 150\,\mathrm{GeV}$ and $\lvert\eta\rvert < 2.5$ $H_{\mathrm{T}} > 700$[1] GeV one jet with trimmed mass $> 50\,\mathrm{GeV}$ trimming parameters: $r = 0.1$ and $p_{\mathrm{Tfrac}} = 0.3$
HLT_AK8PFHT800_TrimMass50	AK8PFJets with $p_{\mathrm{T}} > 200\,\mathrm{GeV}$ and $\lvert\eta\rvert < 2.5$ and trimmed mass $> 50\,\mathrm{GeV}$ trimming parameters: $r = 0.1$ and $p_{\mathrm{Tfrac}} = 0.3$ $H_{\mathrm{T}} > 800\,\mathrm{GeV}$
HLT_PFJet50	AK4PFJets with one jet $p_{\mathrm{T}} > 500\,\mathrm{GeV}$ and $\lvert\eta\rvert < 5.0$
HLT_AK8PFJet[450/500]	AK8PFJets with one jet $p_{\mathrm{T}} > 450/500\,\mathrm{GeV}$ and $\lvert\eta\rvert < 5.0$
HLT_AK8PFJet330 _BTagCSV_p17	AK8PFJets with one jet $p_{\mathrm{T}} > 330\,\mathrm{GeV}$ and $\lvert\eta\rvert < 2.5$ one jet with trimmed mass $> 30\,\mathrm{GeV}$ trimming parameters: $r = 0.1$ and $p_{\mathrm{Tfrac}} = 0.3$ one jet with CSV> 0.3
HLT_AK8PFJet330 _TrimMass30 _PFAK8BoostedDoubleB_np2	AK8PFJets with one jet $p_{\mathrm{T}} > 330\,\mathrm{GeV}$ and $\lvert\eta\rvert < 2.5$ one jet with trimmed mass $> 30\,\mathrm{GeV}$ trimming parameters: $r = 0.1$ and $p_{\mathrm{Tfrac}} = 0.3$ one jet with double-b > -0.2
HLT_AK8DiPFJet280_ 200_TrimMass30_BTagCSVM_p20	AK8PFJets with$p_{\mathrm{T}1} > 280\,\mathrm{GeV}$, $p_{\mathrm{T}2} > 200\,\mathrm{GeV}$ at least one jet with trimmed mass $> 30\,\mathrm{GeV}$ and one calo btag jet with CSVM
HLT_Mu50	Muon with $p_{\mathrm{T}} > 50\,\mathrm{GeV}$ and $\lvert\eta\rvert < 2.5$
HLT_TkMu50	Tracker Muon with $p_{\mathrm{T}} > 50\,\mathrm{GeV}$ and $\lvert\eta\rvert < 2.5$

Appendix B

Full m_{SD} distribution

In this section, we show the post-fit m_{SD} distribution for both pass and fail category in 2016,2017,2018 datasets.

In Figs. B.1-B.6, the post-fit distributions of m_{SD} in data are shown for the passing and failing categories for all 6 p_{T} categories and their combination for 2016, 2017, and 2018. Figure B.7 shows the same post-fit distributions combining the p_{T} categories, but keeping the three data-taking years separate.

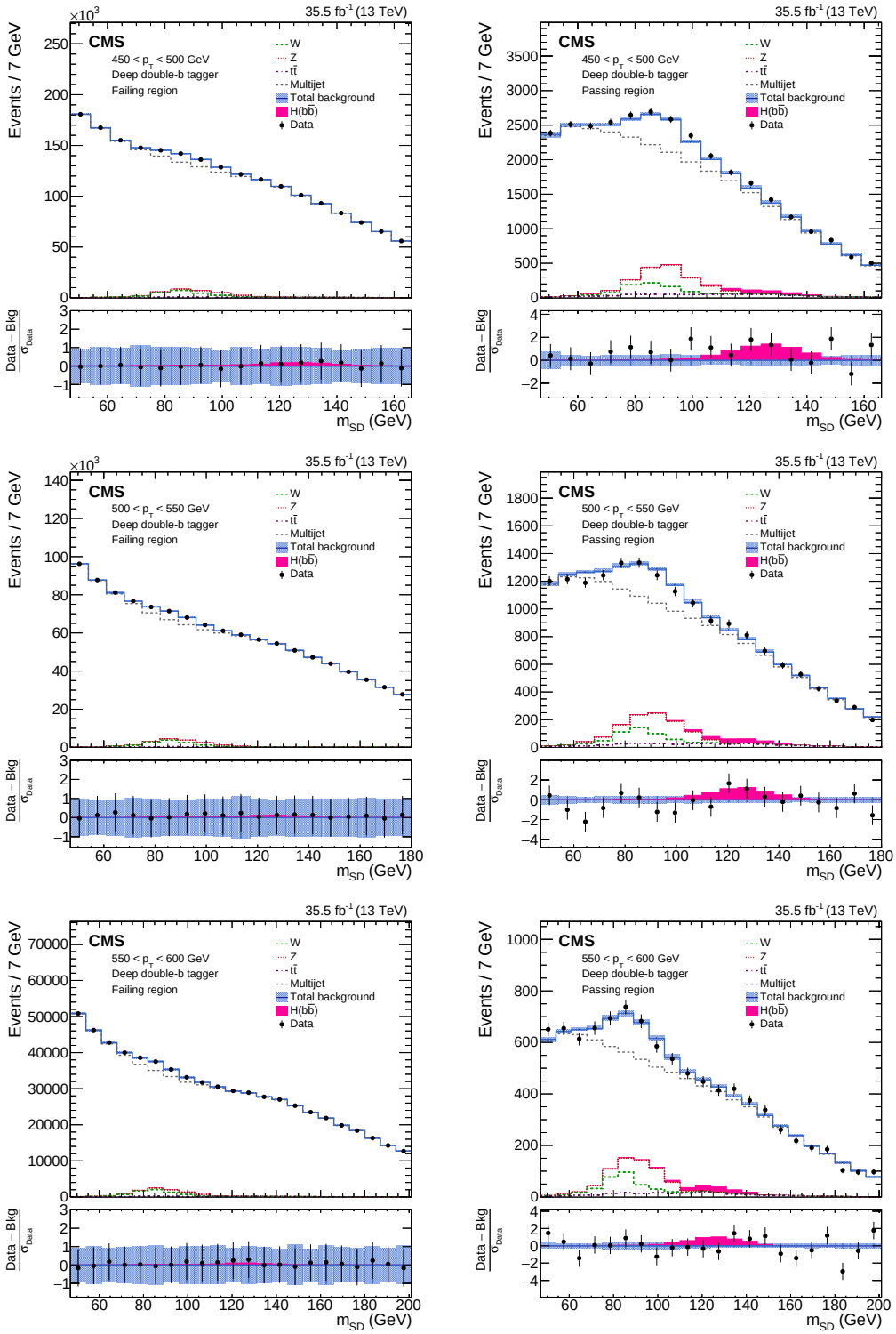


Figure B.1: Post-fit m_{SD} distributions in the full 2016 dataset for the pass and fail regions and the first three p_{T} categories (450-500,500-550,550-600 from top to bottom) by using a polynomial 2nd order in ρ and 1st order in p_{T}. The EW and signal events are also shown separately.

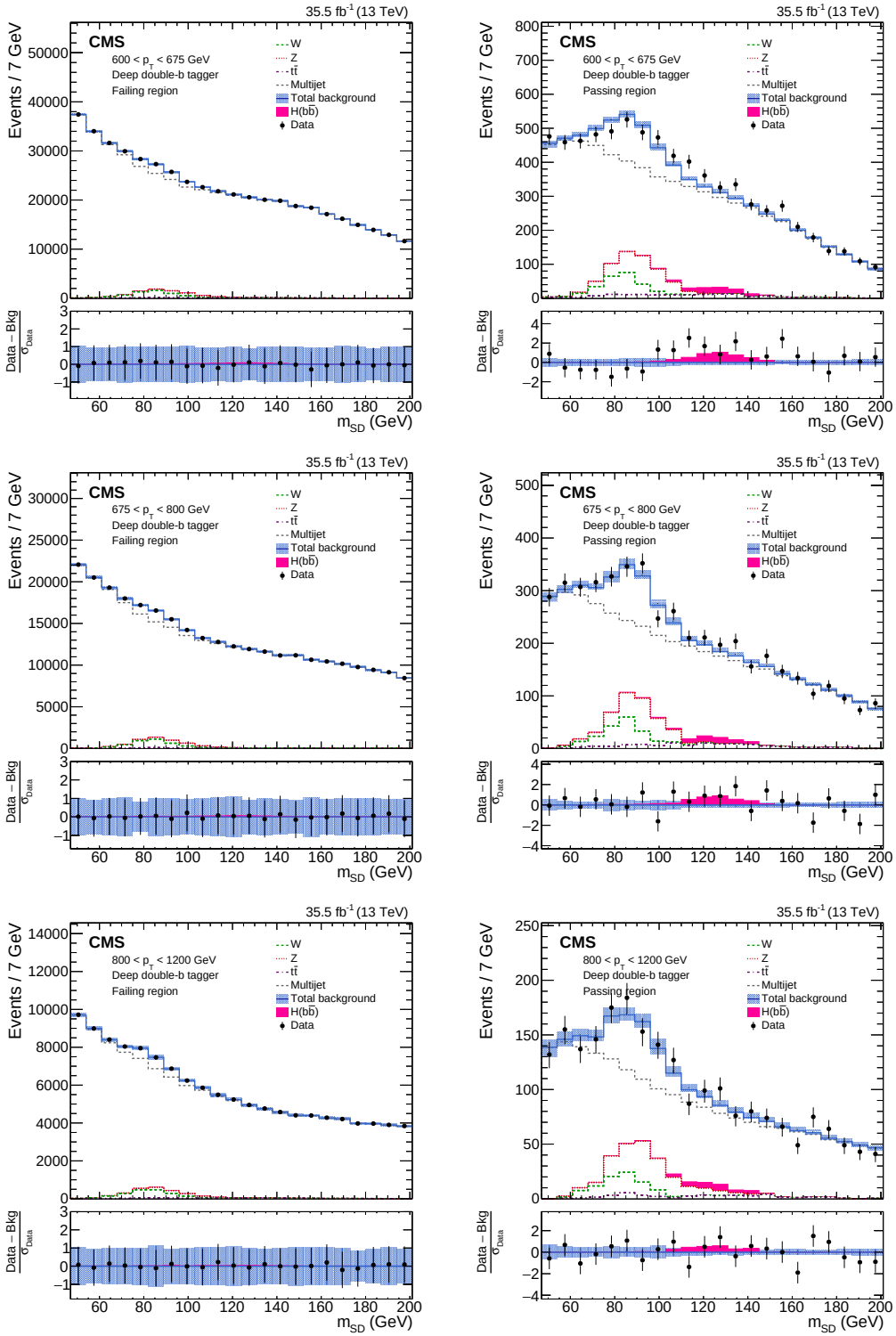


Figure B.2: Post-fit m_{SD} distributions in the full 2016 dataset for the pass and fail regions and the second three p_{T} categories (600-675,675-800,800-1200 from top to bottom) by using a polynomial 2nd order in ρ and 1st order in p_{T}. The EW and signal events are also shown separately.

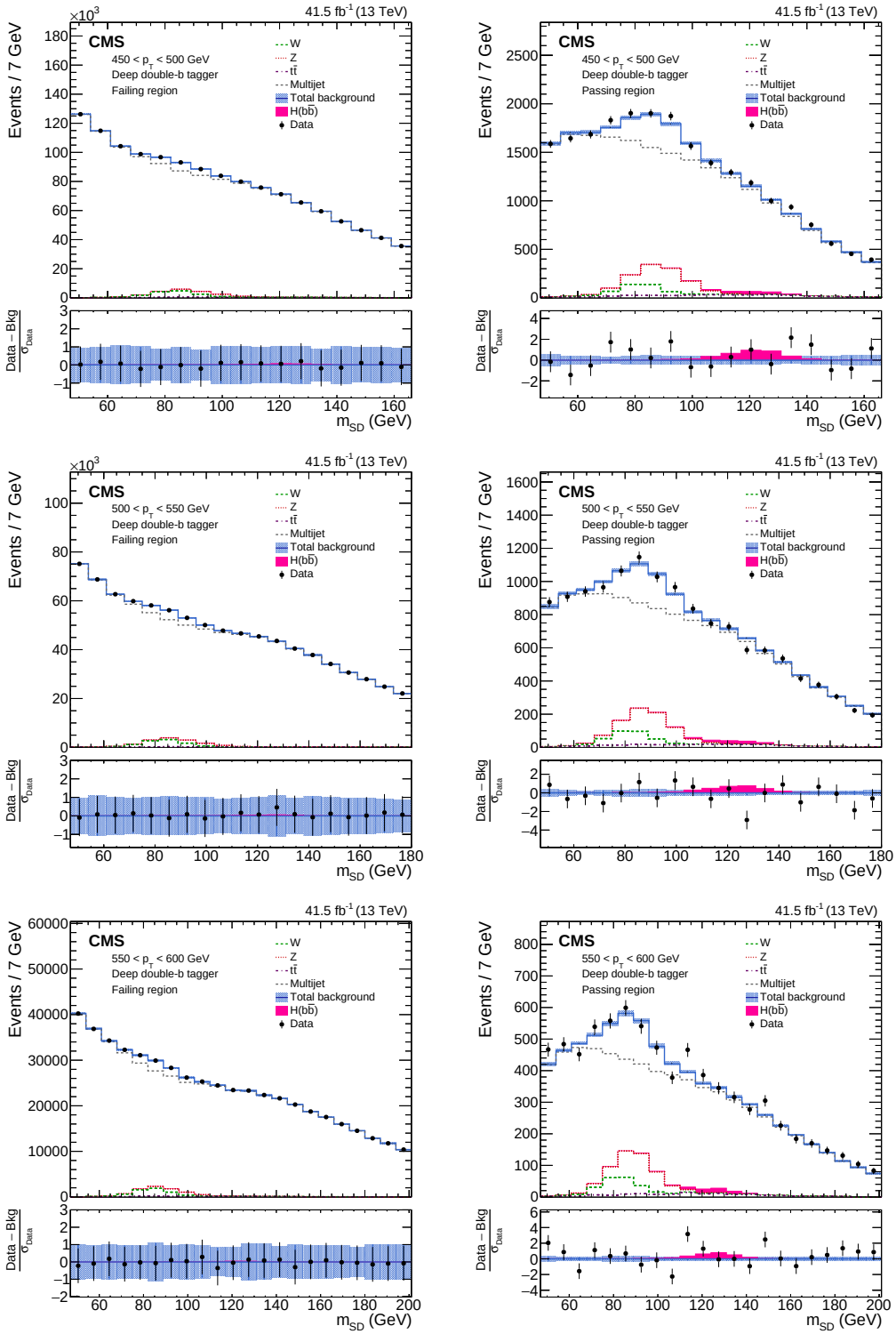


Figure B.3: Post-fit m_{SD} distributions in the full 2017 dataset for the pass and fail regions and the first three p_{T} categories (450-500,500-550,550-600 from top to bottom) by using a polynomial 1st order in ρ and 1st order in p_{T}. The EW and signal events are also shown separately.

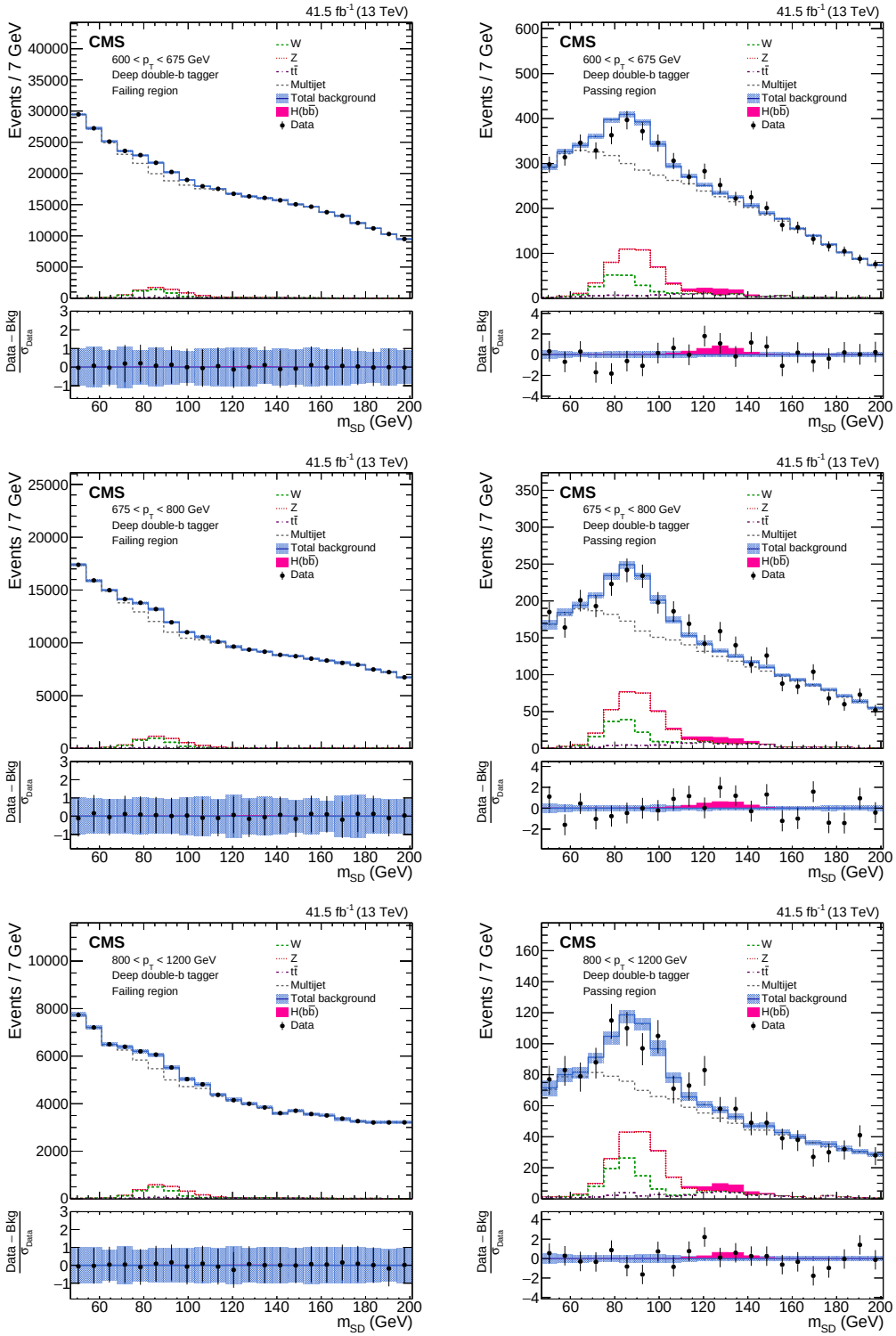


Figure B.4: Post-fit m_{SD} distributions in the full 2017 dataset for the pass and fail regions and the second three p_T categories (600-675,675-800,800-1200 from top to bottom) by using a polynomial 1st order in ρ and 1st order in p_T. The EW and signal events are also shown separately.

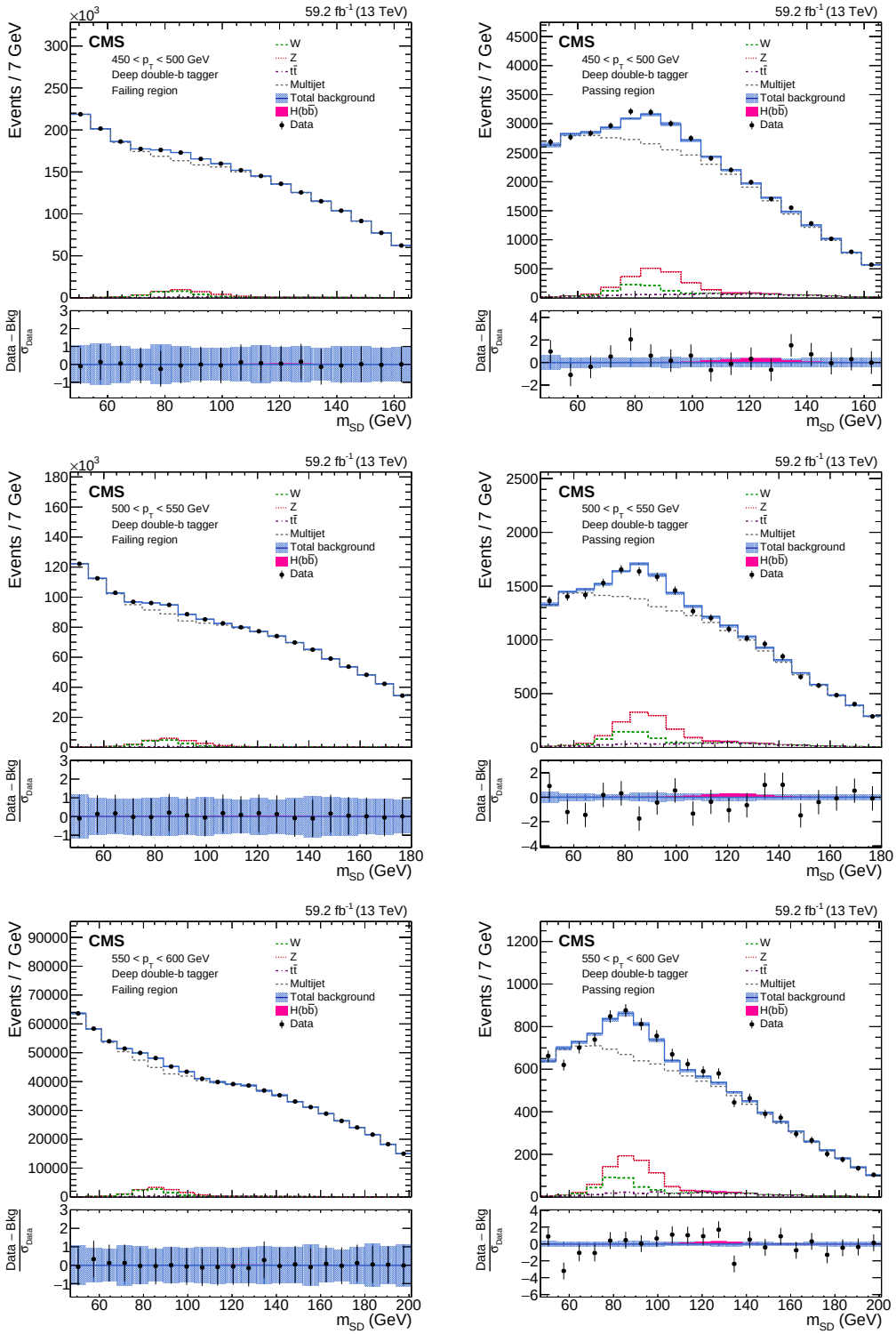


Figure B.5: Post-fit m_{SD} distributions in the full 2018 dataset for the pass and fail regions and the first three p_T categories (450-500,500-550,550-600 from top to bottom) by using a polynomial 1st order in ρ and 1st order in p_T. The EW and signal events are also shown separately.

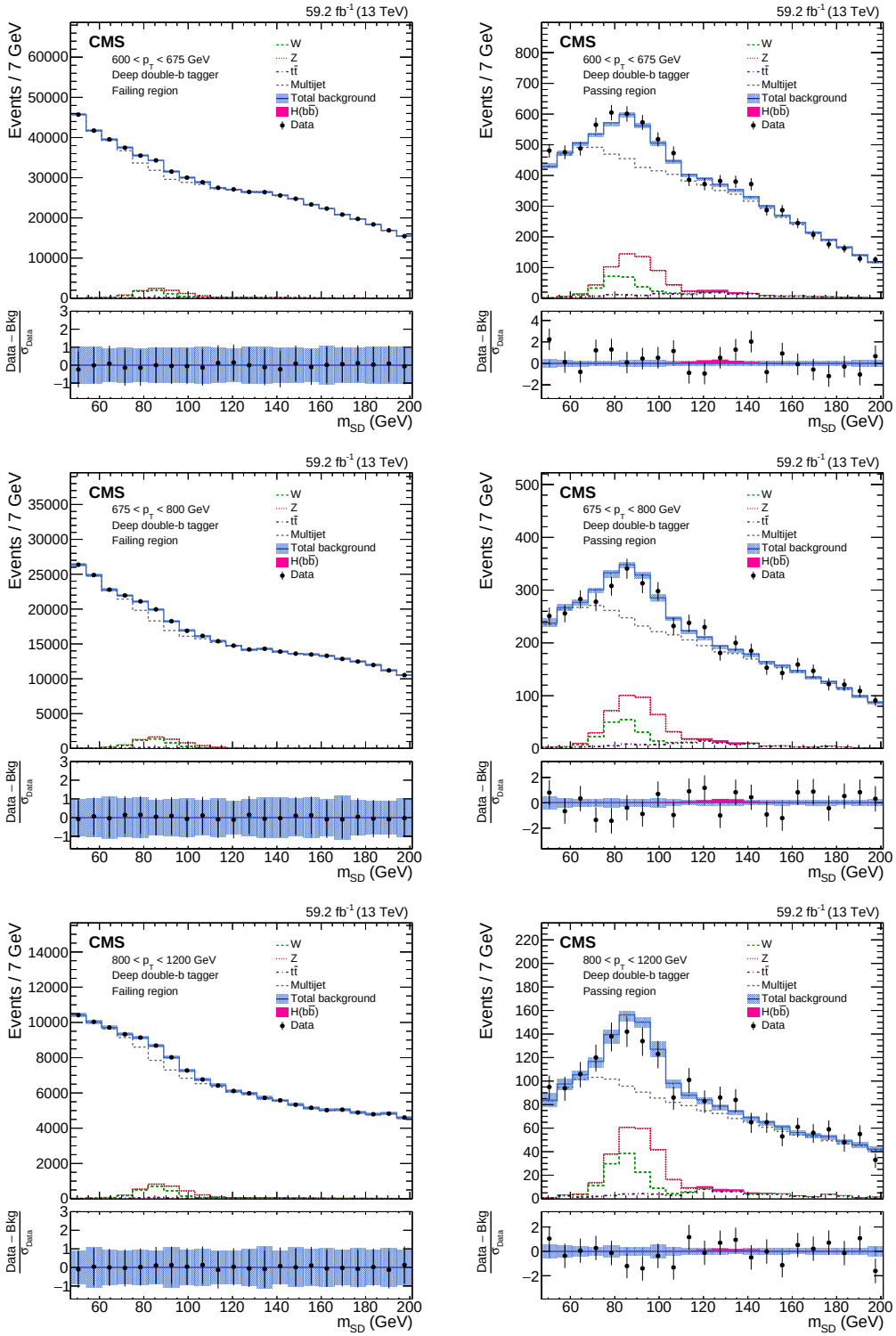


Figure B.6: Post-fit m_{SD} distributions in the full 2018 dataset for the pass and fail regions and the second three p_{T} categories (600-675,675-800,800-1200 from top to bottom) by using a polynomial 1st order in ρ and 1st order in p_{T}. The EW and signal events are also shown separately.

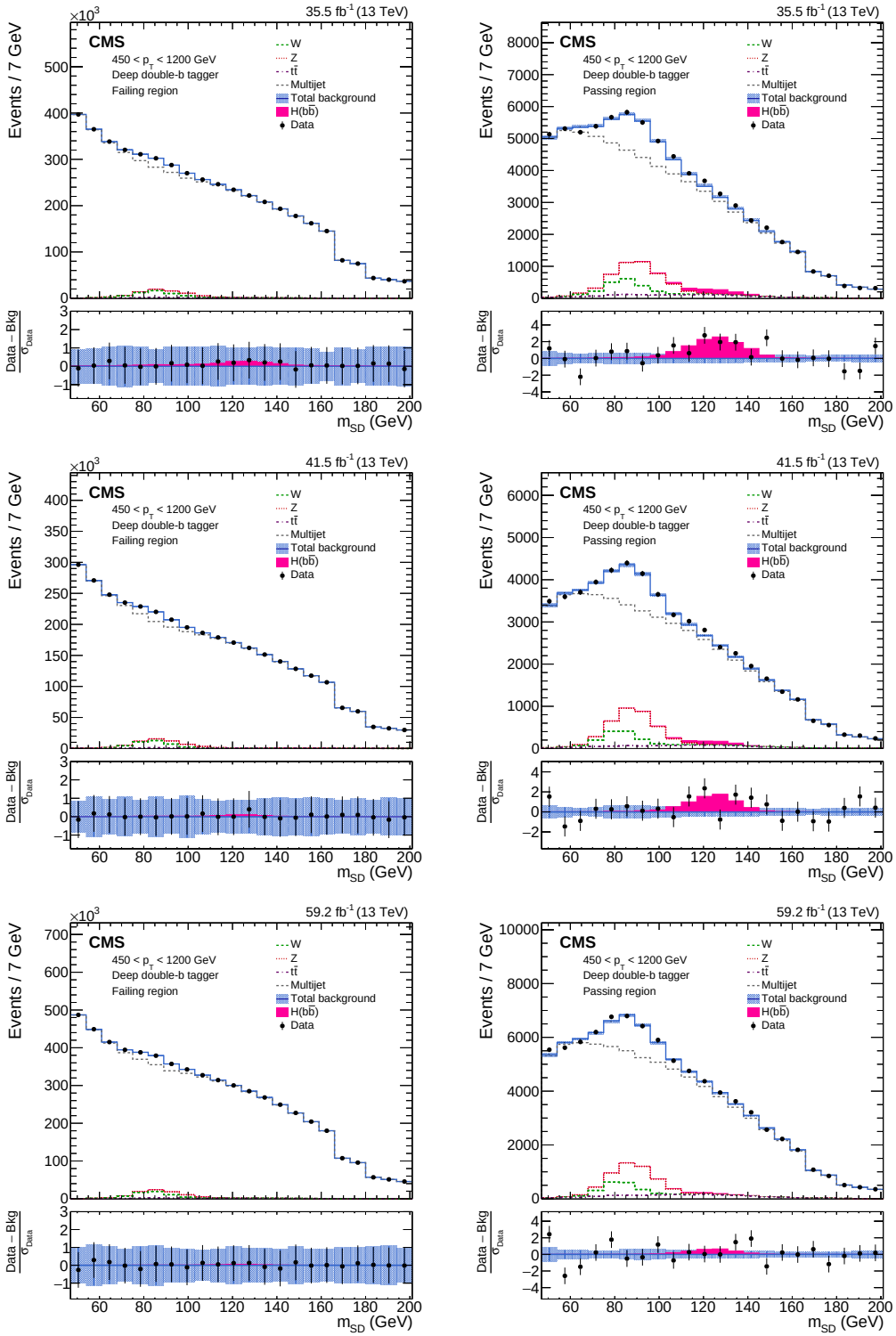
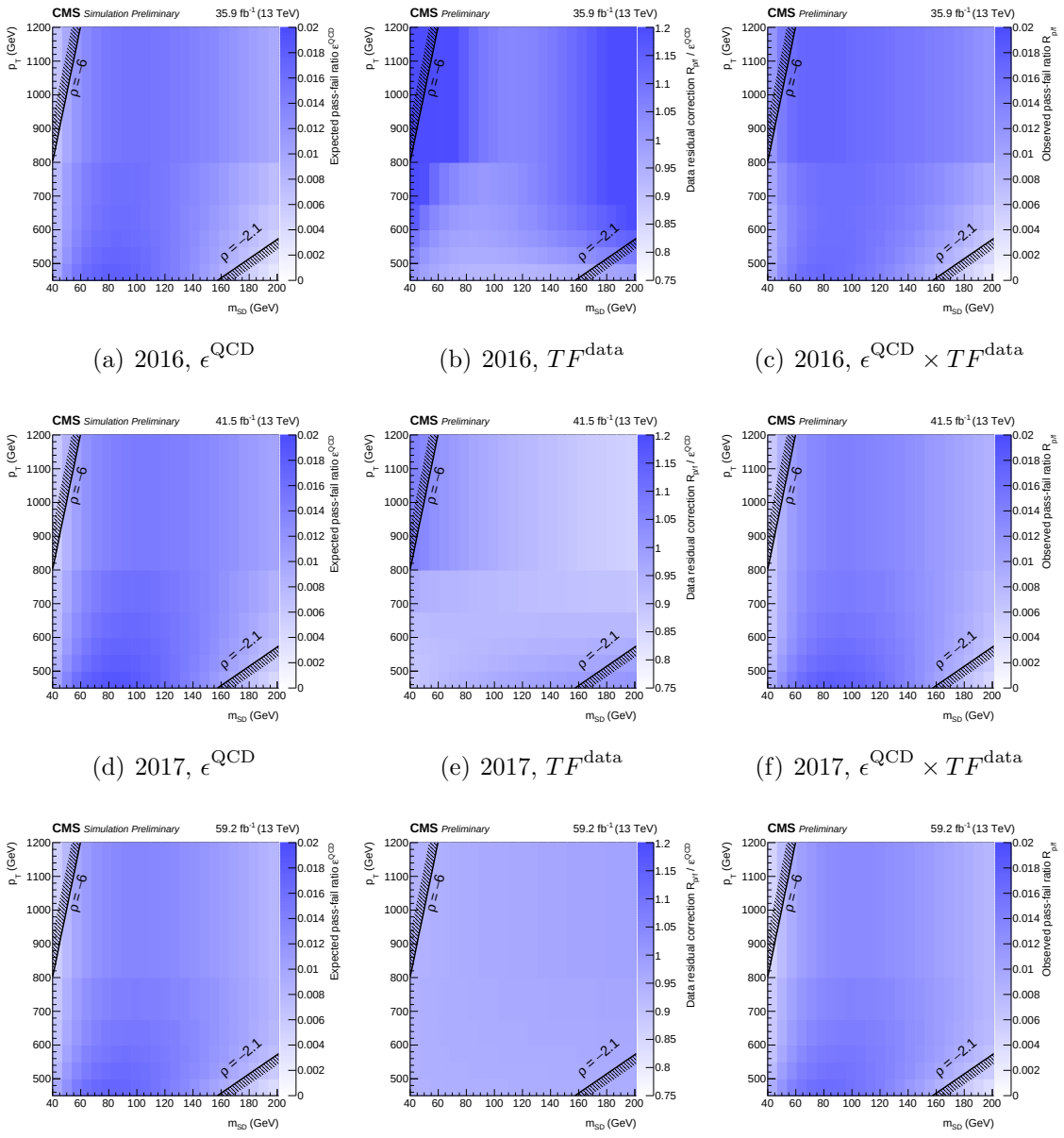


Figure B.7: Post-fit m_{SD} distributions in the full 2016, 2017, and 2018 dataset for the pass and fail regions. The EW and signal events are also shown separately.

Appendix C

Full transfer factors

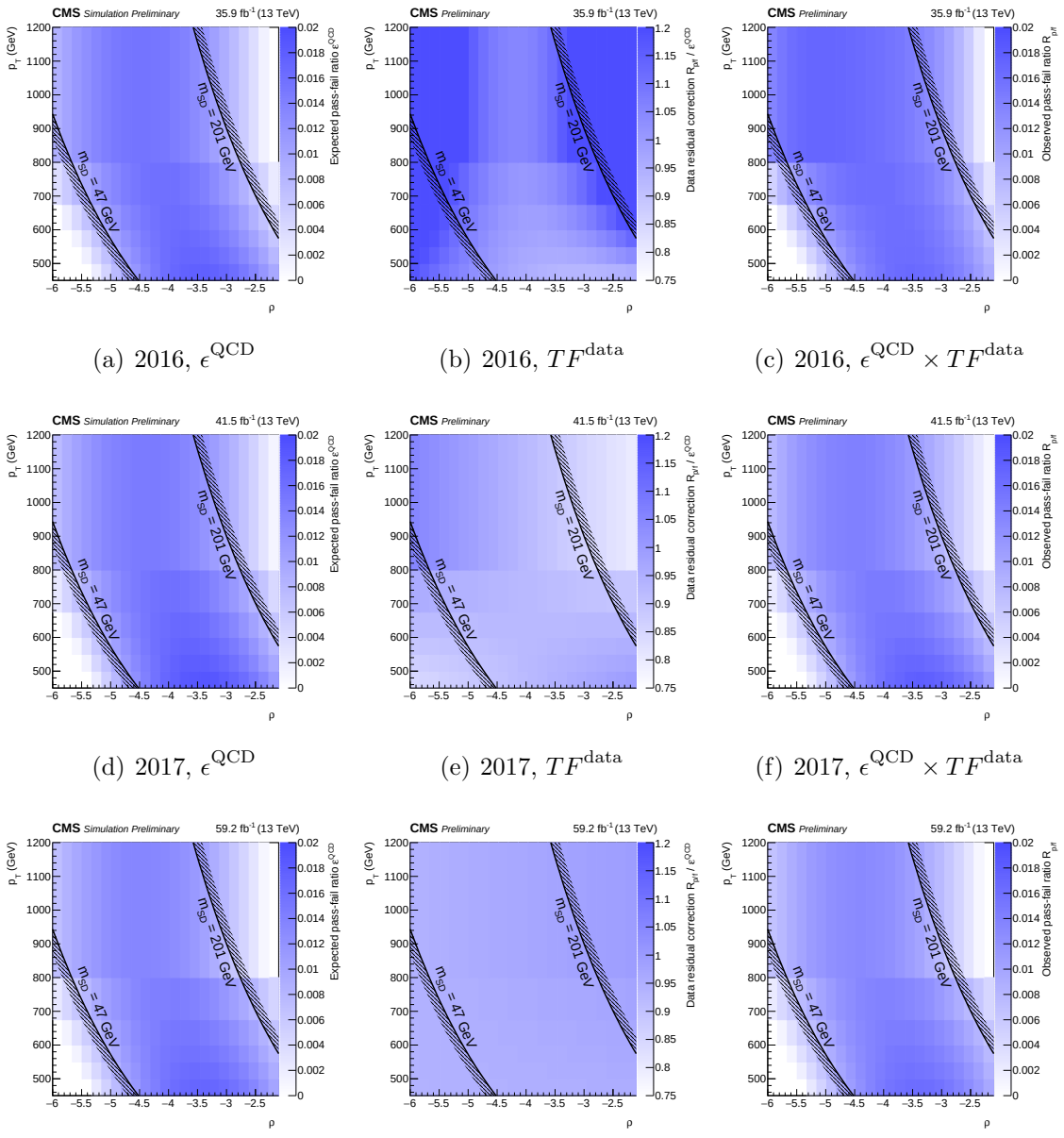


(a) 2016, ϵ^{QCD} (b) 2016, TF^{data} (c) 2016, $\epsilon^{\mathrm{QCD}} \times TF^{\mathrm{data}}$

(d) 2017, ϵ^{QCD} (e) 2017, TF^{data} (f) 2017, $\epsilon^{\mathrm{QCD}} \times TF^{\mathrm{data}}$

(g) 2018, ϵ^{QCD} (h) 2018, TF^{data} (i) 2018, $\epsilon^{\mathrm{QCD}} \times TF^{\mathrm{data}}$

Figure C.1: The top(middle/bottom) rows corresponds to quantities detailed below, corresponding to data collected in 2016(2017/2018) condition respectively. (Left) The expected pass-fail ratio ϵ^{QCD} as a function of jet p_{T} and m_{SD} for QCD multijet simulation. (Middle) The correction TF^{data} for residual differences between data and simulation as a function of jet p_{T} and m_{SD} (Right) The fitted pass-fail ratio $R_{\mathrm{p/f}}$ as a function of jet p_{T} and m_{SD}. The ratio relates the QCD multijet event yield in the passing region to that of the failing region. The pass-fail ratio is the product of the previous two components.



(a) 2016, ϵ^{QCD} (b) 2016, TF^{data} (c) 2016, $\epsilon^{\mathrm{QCD}} \times TF^{\mathrm{data}}$

(d) 2017, ϵ^{QCD} (e) 2017, TF^{data} (f) 2017, $\epsilon^{\mathrm{QCD}} \times TF^{\mathrm{data}}$

(g) 2018, ϵ^{QCD} (h) 2018, TF^{data} (i) 2018, $\epsilon^{\mathrm{QCD}} \times TF^{\mathrm{data}}$

Figure C.2: The top(middle/bottom) rows corresponds to quantities detailed below, corresponding to data collected in 2016(2017/2018) condition respectively. (Left) The expected pass-fail ratio ϵ^{QCD} as a function of jet p_{T} and ρ for QCD multijet simulation. (Middle) The correction TF^{data} for residual differences between data and simulation as a function of jet p_{T} and ρ (Right) The fitted pass-fail ratio $R_{\mathrm{p/f}}$ as a function of jet p_{T} and ρ. The ratio relates the QCD multijet event yield in the passing region to that of the failing region. The pass-fail ratio is the product of the previous two components.

Bibliography

[1] Albert M Sirunyan et al. Inclusive search for a highly boosted Higgs boson decaying to a bottom quark-antiquark pair. *Phys. Rev. Lett.*, 120(7):071802, 2018.

[2] Pablo Antonio Moreno Casares. A review on Loop Quantum Gravity. 2018.

[3] B. L. Hu. Gravity and Nonequilibrium Thermodynamics of Classical Matter. *Int. J. Mod. Phys.*, D20:697–716, 2011.

[4] R. P. Woodard. How Far Are We from the Quantum Theory of Gravity? *Rept. Prog. Phys.*, 72:126002, 2009.

[5] D.I. Kazakov. Beyond the standard model: In search of supersymmetry. In *2000 European School of High-Energy Physics*, pages 125–199, 8 2000.

[6] Roel Aaij et al. Observation of $J/\psi p$ Resonances Consistent with Pentaquark States in $\Lambda_b^0 \to J/\psi K^- p$ Decays. *Phys. Rev. Lett.*, 115:072001, 2015.

[7] Nicola Cabibbo. Unitary symmetry and leptonic decays. *Phys. Rev. Lett.*, 10:531–533, Jun 1963.

[8] Makoto Kobayashi and Toshihide Maskawa. CP Violation in the Renormalizable Theory of Weak Interaction. *Prog. Theor. Phys.*, 49:652–657, 1973.

[9] C. N. Yang and R. L. Mills. Conservation of isotopic spin and isotopic gauge invariance. *Phys. Rev.*, 96:191–195, Oct 1954.

[10] John Ellis, Mary K. Gaillard, and Dimitri V. Nanopoulos. A Historical Profile of the Higgs Boson. 2012.

[11] K. Eguchi et al. First results from KamLAND: Evidence for reactor anti-neutrino disappearance. *Phys. Rev. Lett.*, 90:021802, 2003.

[12] Q. R. Ahmad et al. Direct evidence for neutrino flavor transformation from neutral current interactions in the Sudbury Neutrino Observatory. *Phys. Rev. Lett.*, 89:011301, 2002.

[13] Y. Fukuda et al. Evidence for oscillation of atmospheric neutrinos. *Phys. Rev. Lett.*, 81:1562–1567, 1998.

[14] D. de Florian et al. Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector. 2016.

[15] M. Tanabashi et al. Review of particle physics. *Phys. Rev. D*, 98:030001, Aug 2018.

[16] Georges Aad et al. Observation of a new particle in the search for the standard model higgs boson with the atlas detector at the lhc. *Phys.Lett.*, B716:1–29, 2012.

[17] Serguei Chatrchyan et al. Observation of a new boson at a mass of 125 gev with the cms experiment at the lhc. *Phys.Lett.*, B716:30–61, 2012.

[18] Serguei Chatrchyan et al. Observation of a new boson with mass near 125 GeV in pp collisions at $\sqrt{s} = 7$ and 8 TeV. *JHEP*, 06:081, 2013.

[19] A. M. Sirunyan et al. Observation of Higgs boson decay to bottom quarks. *Phys. Rev. Lett.*, 121:121801, 2018.

[20] Morad Aaboud et al. Observation of $H \to \mathrm{b\bar{b}}$ decays and VH production with the ATLAS detector. *Phys. Lett. B*, 786:59, 2018.

[21] Albert M Sirunyan et al. Observation of $\mathrm{t\bar{t}H}$ production. *Phys. Rev. Lett.*, 120(23):231801, 2018.

[22] M. Aaboud et al. Observation of Higgs boson production in association with a top quark pair at the LHC with the ATLAS detector. *Phys. Lett.*, B784:173–191, 2018.

[23] Albert M Sirunyan et al. Observation of the Higgs boson decay to a pair of τ leptons with the CMS detector. *Phys. Lett.*, B779:283–316, 2018.

[24] Morad Aaboud et al. Cross-section measurements of the Higgs boson decaying into a pair of τ-leptons in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. *Phys. Rev.*, D99:072001, 2019.

[25] Vardan Khachatryan et al. Precise determination of the mass of the Higgs boson and tests of compatibility of its couplings with the standard model predictions using proton collisions at 7 and 8 TeV. *Eur. Phys. J.*, C75(5):212, 2015.

[26] D. de Florian et al. Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector. 2016.

[27] Brian Henning, Xiaochuan Lu, and Hitoshi Murayama. How to use the Standard Model effective field theory. *JHEP*, 01:023, 2016.

[28] G. Rajasekaran. Fermi and the Theory of Weak Interactions. *Resonance J. Sci. Educ.*, 19(1):18–44, 2014.

[29] Mikhail A. Shifman, A.I. Vainshtein, M.B. Voloshin, and Valentin I. Zakharov. Low-Energy Theorems for Higgs Boson Couplings to Photons. *Sov. J. Nucl. Phys.*, 30:711–716, 1979.

[30] Christophe Grojean, Ennio Salvioni, Matthias Schlaffer, and Andreas Weiler. Very boosted Higgs in gluon fusion. *JHEP*, 05:022, 2014.

[31] Andrea Banfi, Adam Martin, and Veronica Sanz. Probing top-partners in Higgs+jets. *JHEP*, 08:053, 2014.

[32] Oliver Sim Bruning, Paul Collier, P Lebrun, Stephen Myers, Ranko Ostojic, John Poole, and Paul Proudlock. *LHC Design Report*. CERN Yellow Reports: Monographs. CERN, Geneva, 2004.

[33] Esma Mobs. The CERN accelerator complex - 2019. Complexe des accélérateurs du CERN - 2019. Jul 2019. General Photo.

[34] Tai Sakuma and Thomas McCauley. Detector and Event Visualization with SketchUp at the CMS Experiment. *J. Phys. Conf. Ser.*, 513:022032, 2014.

[35] S. Chatrchyan et al. The CMS experiment at the CERN LHC. *JINST*, 3:S08004, 2008.

[36] Benedikt Roland Vormwald. The CMS Phase-1 pixel detector – experience and lessons learned from two years of operation. Technical Report CMS-CR-2019-019. 07, CERN, Geneva, Feb 2019.

[37] Martin Lipinski. The Phase-1 Upgrade of the CMS Pixel Detector. Technical Report CMS-CR-2017-135. 06, CERN, Geneva, May 2017.

[38] V Karimäki, M Mannelli, P Siegrist, H Breuker, A Caner, R Castaldi, K Freudenreich, G Hall, R Horisberger, M Huhtinen, and A Cattai. *The CMS tracker system project: Technical Design Report*. Technical Design Report CMS. CERN, Geneva, 1997.

[39] *The CMS electromagnetic calorimeter project: Technical Design Report*. Technical Design Report CMS. CERN, Geneva, 1997.

[40] CMS collaboration et al. Performance of the cms hadron calorimeter with cosmic ray muons and lhc beam data. *Journal of Instrumentation*, 5(03):T03012, 2010.

[41] J Mans, J Anderson, B Dahmes, P de Barbaro, J Freeman, T Grassi, E Hazen, J Mans, R Ruchti, I Schimdt, T Shaw, C Tully, J Whitmore, and T Yetkin. CMS Technical Design Report for the Phase 1 Upgrade of the Hadron Calorimeter. Technical Report CERN-LHCC-2012-015. CMS-TDR-10, Sep 2012. Additional contact persons: Jeffrey Spalding, Fermilab, spalding@cern.ch, Didier Contardo, Universite Claude Bernard-Lyon I, contardo@cern.ch.

[42] N Strobbe. The upgrade of the CMS hadron calorimeter with silicon photomultipliers. The upgrade of the CMS hadron calorimeter with silicon 5 photomultipliers. *JINST*, 12(FERMILAB-CONF-16-501-CMS-E. 01):C01080. 8 p, 2017.

[43] V Brigljevic, N Toniolo, S Ventura, C Schwick, G Maron, P Sphicas, F Meijers, A Csilling, R Gomez-Reino, H Larsen, et al. Run control and monitor system for the cms experiment. *eConf*, 303241(cs/0306110):THGT002, 2003.

[44] A. M. Sirunyan et al. Particle-flow reconstruction and global event description with the CMS detector. *JINST*, 12(10):P10003, 2017.

[45] Wolfgang Adam, Boris Mangano, Thomas Speer, and Teddy Todorov. Track Reconstruction in the CMS tracker. Technical Report CMS-NOTE-2006-041, CERN, Geneva, Dec 2006.

[46] Wolfgang Adam, R Frühwirth, Are Strandlie, and T Todorov. Reconstruction of electrons with the gaussian-sum filter in the cms tracker at the lhc. *Journal of Physics G: Nuclear and Particle Physics*, 31(9):N9, 2005.

[47] Serguei Chatrchyan et al. Performance of CMS muon reconstruction in *pp* collision events at $\sqrt{s} = 7$ TeV. *JINST*, 7:P10002, 2012.

[48] Stephanie Baffioni, Claude Charlot, Federico Ferri, David Futyan, Paolo Meridiani, Ivica Puljak, Chiara Rovelli, Roberto Salerno, and Yves Sirois. Electron reconstruction in CMS. Technical Report CMS-NOTE-2006-040, CERN, Geneva, Feb 2006.

[49] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. The anti-k_t jet clustering algorithm. *JHEP*, 04:063, 2008.

[50] Syarif Rizki. *Search for Vector-Like Heavy Quark Partners Decaying to Leptons and Jets in pp Collisions at $\sqrt{s}$ = 13 TeV by the CMS Experiment.* PhD thesis, Brown University, 2018.

[51] The CMS collaboration. Determination of jet energy calibration and transverse momentum resolution in cms. *Journal of Instrumentation*, 6(11):P11002, 2011.

[52] Mrinal Dasgupta. Jet masses for high p_T QCD jets at all orders. *PoS*, DIS2010:104, 2010.

[53] Andrew J. Larkoski, Simone Marzani, Gregory Soyez, and Jesse Thaler. Soft Drop. *JHEP*, 05:146, 2014.

[54] Mrinal Dasgupta, Alessandro Fregoso, Simone Marzani, and Gavin P. Salam. Towards an understanding of jet substructure. *JHEP*, 09:029, 2013.

[55] Andrew J. Larkoski and Jesse Thaler. Unsafe but Calculable: Ratios of Angularities in Perturbative QCD. *JHEP*, 09:137, 2013.

[56] Stephen D. Ellis, Christopher K. Vermilion, and Jonathan R. Walsh. Techniques for improved heavy particle searches with jet substructure. *Phys. Rev.*, D80:051501, 2009.

[57] Stephen D. Ellis, Christopher K. Vermilion, and Jonathan R. Walsh. Recombination Algorithms and Jet Substructure: Pruning as a Tool for Heavy Particle Searches. *Phys. Rev.*, D81:094023, 2010.

[58] T. K. Aarrestad et al. Identification and calibration of boosted hadronic W/Z bosons at 13 TeV. Technical report, AN-2016/215, 2016.

[59] James Dolen, Philip Harris, Simone Marzani, Salvatore Rappoccio, and Nhan Tran. Thinking outside the ROCs: Designing Decorrelated Taggers (DDT) for jet substructure. 2016.

[60] Jesse Thaler and Ken Van Tilburg. Identifying Boosted Objects with N-subjettiness. *JHEP*, 03:015, 2011.

[61] Andrew J. Larkoski, Gavin P. Salam, and Jesse Thaler. Energy Correlation Functions for Jet Substructure. *JHEP*, 06:108, 2013.

[62] Andrew J. Larkoski, Ian Moult, and Duff Neill. Power Counting to Better Jet Observables. *JHEP*, 12:009, 2014.

[63] Ian Moult, Lina Necib, and Jesse Thaler. New Angles on Energy Correlation Functions. 2016.

[64] Nhan Tran et al. Search for boosted light dijet resonances with an associated jet with 2016 data. Technical report, AN-2016/365, 2016.

[65] Albert M Sirunyan et al. Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV. *JINST*, 13:P05011, 2018.

[66] Performance of deep tagging algorithms for boosted double quark jet topology in proton-proton collisions at 13 TeV with the Phase-0 CMS detector. Technical Report CMS-DP-2018-046, 2018.

[67] J. Duarte et al. . Technical report, AN-2018/045, 2018.

[68] The ATLAS collaboration. Search for boosted resonances decaying to two b-quarks and produced in association with a jet at $\sqrt{s} = 13$ TeV with the ATLAS detector. 2018.

[69] S. Agostinelli et al. GEANT4 — a simulation toolkit. *Nucl. Instrum. Meth. A*, 506:250, 2003.

[70] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H. S. Shao, T. Stelzer, P. Torrielli, and M. Zaro. The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations. *JHEP*, 07:079, 2014.

[71] J. Alwall, S. S. Hoeche, F. Krauss, N. Lavesson, L. Lonnblad, F. Maltoni, M. L. Mangano, M. Moretti, C. G. Papadopoulos, F. Piccinini, S. Schumann, M. Treccani, J. Winter, and M. Worek. Comparative study of various algorithms for the merging of parton showers and matrix elements in hadronic collisions. *Eur. Phys. J. C*, 53:473, 2008.

[72] Paolo Nason. A new method for combining nlo qcd with shower monte carlo algorithms. *JHEP*, 0411:040, 2004.

[73] S. Frixione, P. Nason, and C. Oleari. Matching NLO QCD computations with parton shower simulations: the POWHEG method. *JHEP*, 11:070, 2007.

[74] 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*, 06:043, 2010.

[75] T. Sjöstrand, S. Ask, J. R. Christiansen, R. Corke, N. Desai, P. Ilten, S. Mrenna, S. Prestel, C. O. Rasmussen, and P. Z. Skands. An introduction to PYTHIA 8.2. *Comput. Phys. Commun.*, 191:159, 2015.

[76] Vardan Khachatryan et al. Event generator tunes obtained from underlying event and multiparton scattering measurements. 2015.

[77] Albert M Sirunyan et al. Extraction and validation of a new set of CMS PYTHIA8 tunes from underlying-event measurements. *Eur. Phys. J. C*, 80(1):4, 2020.

[78] Richard D. Ball et al. Parton distributions for the LHC Run II. *JHEP*, 04:040, 2015.

[79] J. M. Lindert et al. Precise predictions for $V+$ jets dark matter backgrounds. *Eur. Phys. J.*, C77(12):829, 2017.

[80] Stefan Kallweit, Jonas M. Lindert, Philipp Maierhöfer, Stefano Pozzorini, and Marek Schönherr. NLO electroweak automation and precise predictions for W+multijet production at the LHC. *JHEP*, 04:012, 2015.

[81] Stefan Kallweit, Jonas M. Lindert, Philipp Maierhofer, Stefano Pozzorini, and Marek Schönherr. NLO QCD+EW predictions for V + jets including off-shell vector-boson decays and multijet merging. *JHEP*, 04:021, 2016.

[82] Stefan Kallweit, Jonas M. Lindert, Stefano Pozzorini, Marek Schönherr, and Philipp Maierhöfer. NLO QCD+EW automation and precise predictions for V+multijet production. In *Proceedings, 50th Rencontres de Moriond, QCD and high energy interactions*, pages 121–124, 2015.

[83] Rikkert Frederix and Stefano Frixione. Merging meets matching in MC@NLO. *JHEP*, 12:061, 2012.

[84] S. P. Jones, M. Kerner, and G. Luisoni. Next-to-Leading-Order QCD Corrections to Higgs Boson Plus Jet Production with Full Top-Quark Mass Dependence. *Phys. Rev. Lett.*, 120(16):162001, 2018.

[85] Jonas M. Lindert, Kirill Kudashkin, Kirill Melnikov, and Christopher Wever. Higgs bosons with large transverse momentum at the LHC. *Phys. Lett.*, B782:210–214, 2018.

[86] Kathrin Becker, Fabrizio Caola, Andrea Massironi, Bernhard Mistlberger, Pier Monni, Xuan Chen, Stefano Frixione, Thomas Kurt Gehrmann, Nigel Glover, Keith Murray Hamilton, Alexander Yohei Huss, Stephen Philip Jones, Alexander Karlberg, Matthias Kerner, Kirill Kudashkin, Jonas Lindert, Gionata Luisoni, Michelangelo Mangano, Stefano Augusto Pozzorini, Emanuele Re, Gavin Salam, Eleni Vryonidou, and Christopher Wever. Recommended predictions for the boosted-Higgs cross section. Technical Report LHCHXSWG-2019-002, CERN, Geneva, Mar 2019.

[87] Emanuele Bagnaschi and Alessandro Vicini. The Higgs transverse momentum distribution in gluon fusion as a multiscale problem. *JHEP*, 01:056, 2016.

[88] David Krohn, Jesse Thaler, and Lian-Tao Wang. Jet Trimming. *JHEP*, 02:084, 2010.

[89] Daniele Bertolini, Philip Harris, Matthew Low, and Nhan Tran. Pileup Per Particle Identification. *JHEP*, 10:059, 2014.

[90] CMS JME POG. Recommended jet energy corrections and uncertainties for data and mc. https://twiki.cern.ch/twiki/bin/view/CMS/JECDataMC, 2017.

[91] CMS JME POG. Recommended jet id. https://twiki.cern.ch/twiki/bin/viewauth/CMS/JetID, 2017.

[92] CMS MUO POG. CMS MUO POG recommendations. Technical report, 2015.

[93] CMS EGM POG. CMS EGM POG recommendations. Technical report, 2015.

[94] Electron reconstruction and identification at sqrt(s) = 7 TeV. 2010.

[95] CMS Collaboration. Identification of b-quark Jets at the CMS Experiment in the LHC Run2 Startup. CMS Physics Analysis Summary CMS-PAS-BTV-15-001, CERN, 2016.

[96] Georges Aad et al. Measurements of the Higgs boson production and decay rates and constraints on its couplings from a combined ATLAS and CMS analysis of the LHC pp collision data at $\sqrt{s} = 7$ and 8 TeV. *JHEP*, 08:045, 2016.

[97] ATLAS and CMS Collaborations. Procedure for the LHC Higgs boson search combination in Summer 2011. CMS Note/ATLAS Pub CMS-NOTE-2011-005, ATL-PHYS-PUB-2011-11, 2011.

[98] Isaac Asimov. *Isaac Asimov: The Complete Stories, Vol. 1.* Broadway Books, New York City, USA, 1990.

[99] Glen Cowan, Kyle Cranmer, Eilam Gross, and Ofer Vitells. Asymptotic formulae for likelihood-based tests of new physics. *Eur. Phys. J.*, C71:1554, 2011. [Erratum: Eur. Phys. J.C73,2501(2013)].

[100] Steve Baker and Robert D. Cousins. Clarification of the Use of Chi Square and Likelihood Functions in Fits to Histograms. *Nucl. Instrum. Meth.*, 221:437–442, 1984.

[101] C. Patrignani et al. Review of Particle Physics. *Chin. Phys.*, C40(10):100001, 2016.

[102] CMS BTV Twiki. Cms btv twiki. Technical report, 2017.

[103] Serguei Chatrchyan et al. Identification of b-quark jets with the CMS experiment. *JINST*, 8:P04013, 2013.

[104] CMS MUO POG. CMS MUO POG SF recommendations. Technical report, 2017.

[105] Keith Hamilton, Paolo Nason, Carlo Oleari, and Giulia Zanderighi. Merging H/W/Z + 0 and 1 jet at NLO with no merging scale: a path to parton shower + NNLO matching. *JHEP*, 05:082, 2013.

[106] Nicolas Berger et al. Simplified Template Cross Sections - Stage 1.1. 2019.