

SEARCH FOR THE DECAY OF STANDARD MODEL HIGGS BOSONS TO A CHARM QUARK-ANTIQUARK PAIR USING 138 FB^{-1} OF CMS PROTON-PROTON COLLISION DATA AT $\sqrt{s} = 13 \text{ TEV}$

by

Bjorn Burkle

M.Sc. in Physics, Brown University, 2018

B.Sc. in Physics, University of California Santa Barbara, 2016

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR

THE DEGREE OF DOCTOR OF PHILOSOPHY

IN THE DEPARTMENT OF PHYSICS AT BROWN UNIVERSITY

Providence, Rhode Island

May 2022



© Copyright 2022 by Bjorn Burkle

This dissertation by Bjorn Burkle is accepted in its present form by
the Department of Physics as satisfying the
dissertation requirement for the degree of
Doctor of Philosophy

Date.....
Ulrich Heintz, Advisor

Recommended to the Graduate Council

Date.....
Greg Landsberg, Reader

Date.....
Rick Gaitskell, Reader

Approved by the Graduate Council

Date.....
Andrew G. Campbell
Dean of the Graduate School

Bjorn Burkle

Curriculum Vitae

Physics Department

Barus and Holley Building
Brown University
184 Hope Street
Providence, RI 02906

Email:

bjorn_burkle@brown.edu
bjorn.burkle@cern.ch
burkle.bjorn@gmail.com

Home Address:

298 Wickenden Street,
Apt. 7
Providence, RI 02903
858-945-1477 (cell)

Education

- **Brown University**, Graduate Studies in Physics, Providence, RI
 - Candidate for Doctor of Philosophy in Physics, Aug 2016 – present
 - * Thesis: Search for the decay of Standard Model Higgs bosons to a charm quark-antiquark pair using 138 fb^{-1} of CMS proton-proton collision data at $\sqrt{s} = 13 \text{ TeV}$
 - * Advisor: Prof. Ulrich Heintz
 - Master of Science in Physics, May 2018
- **University of California Santa Barbara**, College of Letters and Sciences, Santa Barbara, CA
 - Bachelor of Science in Physics, May 2016

Experience

Research

Physics Research Assistantship, Brown U., September 2017 – present

- **CMS Experiment**, September 2017 – present
 - Full author, January 2019 – present
 - Upgrade Work: **Quality control of sensors used for production of phase II CMS outer tracker upgrade**, March 2017 – October 2021
 - * Developed procedure and performed full characterization of 2S and PS-s sensors that will be used to build the upgraded version of the outer tracker for the Phase II upgrade of the CMS detector.
 - * Developed procedures and performed IV measurements on selected 2S modules between each module assembly step to check if the silicon sensors are damaged during assembly procedures.
 - Upgrade Work: **Characterization of nuclear reactor neutron energy spectrum for used irradiation testing of the CMS high granularity calorimeter**, October 2016 – present
 - * I planned and carried out a study to use the spectrum unfolding via foil activation method to characterize a functional form of the neutron radiation spectrum of the Rhode Island Nuclear Science Center (RINSC) fast neutron reactor. The RINSC reactor is one of the primary reactors used to perform neutron irradiation on silicon sensors that will be used to build the High Granularity Calorimeter (HGCAL) for the Phase II upgrade of the CMS detector. Portions of the HGCAL are expected to receive an unprecedented $10^{16} n_{eq}/cm^2$ dose of radiation, and studies must be performed to verify the detector integrity at said radiation doses. Results of this study are used to actively monitor the radiation doses received by silicon sensors during these irradiation tests.
 - * Developed a series of analysis tools to measure the 1 MeV neutron equivalent radiation dose received by monitor foils when performing silicon irradiation studies¹.
 - Data Quality Management: **Monitoring of the CMS outer tracker zero-suppression algorithm during data taking using spy channel data taking**, April 2018 – present
 - * During data taking, the bandwidth required to readout the raw data of the CMS outer tracker is orders of magnitude too large given the detector readout rate. To account for this, a zero-suppression algorithm is run on the detector readout hardware

¹https://github.com/bjornburkle/rabbit_normalization

to greatly reduce the size and complexity of the outer tracker data. The spy channel is a separate readout channel allowing periodic full detector readouts. This allowing us to verify that the assumptions of the zero-suppression algorithms continue to be true across long data taking periods.

- * Developed a code base which looks at one of the intermediary steps of the zero-suppression algorithm where a flat offset is subtracted from the analog-to-digital conversion (ADC) counts from the module readouts. The algorithm assumes that the shape of the offset is flat, and the spy data analysis allows us to see if any deformations in this offset occur during data taking.
- * Maintained a study to monitor the average noise values in the outer tracker and observe how the value changes during data taking.
- * Developed a python code base which automatically converts raw spy-data to to the EDM format used for CMS analysis and transfers the converted data to a common EOS storage space accessible for spy data analysis².

– Ph.D. Thesis: Search for the decay of Standard Model Higgs bosons to a charm quark-antiquark pair using 138 fb^{-1} of CMS proton-proton collision data at $\sqrt{s} = 13 \text{ TeV}$

- * Worked with a small team of researches to perform one of two analyses performed by the CMS collaboration to look for the rare $VH(H \rightarrow c\bar{c})$ Standard Model process using the Full Run 2 dataset. The analysis framework was designed from studies based on physics events produced via Monte Carlo simulations and later applied to real physics events collected during the 2016–2018 data taking years. The analysis makes use of the DeepJet DNN algorithm to identify jets originating from the hadronization of charm quarks, as well as isolate primary background processes based on the flavor of jets produced during the events. Furthermore, a boosted decision tree is trained and applied to improve separation of the $VH(H \rightarrow c\bar{c})$ signal from standard model background processes.
- * The analysis is tuned for three separate channels based on the number of leptons and neutrinos present in the final state
- * A final negative log-likelihood fit is made using a frequentist approach to measure the $VH(H \rightarrow c\bar{c})$ signal strength while free floating a normalization on the overall

²<https://github.com/trackerpro/SiStripCronJob>

cross section of dominant standard model backgrounds.

- * Through combination of results with the second $VH(H \rightarrow c\bar{c})$ analysis, produced the strongest limits on the $VH(H \rightarrow c\bar{c})$ process production rate and the charm Yukawa coupling to date.
- * An additional validation analysis was performed to measure the signal strength of the $VZ(Z \rightarrow c\bar{c})$ standard model process. After combination of results with the second $VH(H \rightarrow c\bar{c})$ analysis, we became the first study to measure the $Z \rightarrow c\bar{c}$ process in a hadron collider.

- **Independent Machine Learning Studies using CMS Opendata**, September 2018 – present

- Analysis: **End-to-End Jet Classification of Boosted Top Quarks with the CMS OpenData**, April 2021 – present

- * We use a novel application of the end-to-end deep learning technique on the task of discriminating top quark-initiated jets from those originating from the hadronization of a light quark or a gluon. In this study, we map calorimeter energy measurements and track-based quantities in the vicinity of a high momentum jet to two-dimensional images. A neural network using the residual network architecture is then trained on these image and used to determine whether the jet was performed from the hadronization of top quarks or the hadronization of other strongly interacting particles.
- * This is the first study to ever make use of the spatial location of low level reconstructed hits in CMS tracker as inputs to a jet tagging algorithm.
- * Worked with members of Google to train and evaluate the network using their newest TPU hardware and benchmark the TPU training speeds against leading end GPUs.

Teaching

- **Physics Teaching Assistantship**, Brown U., September 2016 – May 2018

- TA for undergraduate introductory physics labs for non physics-including concentrators (Phys 0030, Fall 2016; Phys 0040, Spring 2018)
 - * Obtained Award for Excellence as Graduate Teaching Assistance for work as Phys 0040 TA

- TA for undergraduate physics lab on modern quantum physics experiments (Phys 0560, Spring 2017)

List of Publications

- Andrews et al, “End-to-End Jet Classification of Boosted Top Quarks with the CMS Open-Data”, Phys. Rev. D **105**, 052008 – Published 22 March 2022.
- The CMS Collaboration, “Direct search for the standard model Higgs boson decaying to a charm quark-antiquark pair”, Phys. Rev. Lett. – targetting.

List of Presentations

- vCHEP2021, "End-to-End Jet Classification of Boosted Top Quarks with CMS Open Data", May 2021.
- PS April 2021: American Physical Society April Meeting 2021, "Search for a Higgs Boson Decaying to Two Charm Quarks with Resolved Jet Topology using Run 2 CMS Data", April 2021.
- Higgs 2020 Conference, "Higgs to cc (ATLAS+CMS)", October 2020.
- Technologies for Discovery IV: The CPAD Instrumentation Frontier Workshop 2018, "End-to-End Boosted Jet ID for Top/QCD Discrimination in a 200 PU Environment in the Phase II CMS Detector", December 2018.
- 2018 US Users Association Meeting, "Neutron Irradiation of Silicon Sensors for the Phase 2 Upgrade", October 2018.

Preface and Acknowledgments

This thesis represents a compendium of projects worked on by myself, Bjorn Burkle, while pursuing my doctorate in philosophy at Brown University. Throughout my degree, I was able to explore work involving radiation dosimetry, the testing of silicon based tracking sensors, data quality management, computer-vision approaches to particle decay identification, and as the name of this thesis implies: an analysis of particle physics data to look for evidence of a Higgs boson directly decaying to second generation charm quark-antiquark pairs. From this, I have gained a vast array of knowledge and experiences that I will proudly carry with me for the rest of my life. As such, there is a plethora of people who I would like to thank and recognize for the roles they played in my journey through grad school.

First and foremost, I would like to send thanks to all of my friends and family who were there to support me morally and emotionally through the all of trials and tribulations posed by graduate school. I would like to thank my advisor Ulrich Heintz as well as professor Meenakshi Narain for the guidance they have given during my career as a physics researcher. I would then like to thank the two additional members of my thesis committee: Greg Landsberg and Rick Gaitskell for taking the time to read my dissertation and (hopefully) deem me deserving of a Ph.D. in particle physics.

I would then like to thank all of those I have worked with in the lab at Brown. I would like to specifically thank Nick Hinton and Eric Spenser for the countless hours of guidance and help equipment debugging that they had provided me. I would like to thank Justin Voelker for measuring irradiated diodes to help understand the neutron spectrum of the RINSC reactor. I would like to thank Farrah Simpson and Nick Pervan for undertaking the continuation of sensor quality control measurements after I leave. And finally, I would then like to give an extra special shout-out to Soumya Ghosh, Isaac Fugate, and Ginger Cheng for all of the support they gave me and fun times we had in the early years of my graduate student career.

Next, I would like to thank those whom I worked with in exploring the end-to-end approaches to jet image based top tagging. In particular, I would like to thank Emanuele Usai who developed

the way to add reconstructed tracker hits into the AOD dataframe and produce the CMS opendata files used to construct the top jet images. I would like to thank Davide Di Croche for furthering our ability to test the training of our network on a larger variety of GPU architectures as well as his hospitality that he showed me when I was briefly at CERN. I would like to thank Sergei Gleyzer and Michael Andrews in their development of the original jet identification framework from which the study was derived. Finally, I would like to thank Sergei one more time for all of his help in editing the final top tagging paper into the form that was eventually published.

I would then like to thank all of those who I worked with in the $VH(H \rightarrow c\bar{c})$ analysis. In particular, I would like to thank my collaborators Xavier Coubez, Andrey Pozdnyakov, Spandon Mondal, and Luca Mastrolorenzo in the resolved jet analysis team. We all worked hard to improve the results of the 2016 analysis well beyond what would have been expected from an increase in luminosity alone. I would like to further thank Luca for the hospitality that he showed me during my time at CERN. I would also like to thank my collaborators in the merged jet analysis team, in particular Huilin Qu and Loukas Gouskos who were able to develop an absolutely mind-blowing jet tagging and mass regression algorithm for identification of a high p_T Higgs boson decaying to charm quarks. I would then like to additionally congratulate all of my collaborators on both of the analysis teams for the phenomenal results produced from this analysis. In my (very biased) opinion, this is one of the most impressive results to come out of Run 2 of the LHC.

I would like to once again thank all of the current and former high energy physics Ph.D. students that I have had the joy of sharing a workspace with when here at Brown. In particular, Daniel Li, Mary Hadley, Nick Pervan, Farrah Simpson, Wenyu Zhang, Jess Wong, Michael Lukasik, Rizki Syarif, Eric Scotti, and Jangbae Lee. Last and not least, I would like to thank and recognize all of the other friends I have made through the Brown University Physics Department.

Contents

Curriculum Vitae	iv
Preface and Acknowledgments	x
Contents	xi
List of Figures	xvii
List of Tables	xxvii
1 Introduction	1
2 The Standard Model	4
2.1 The Fundamental Particles	4
2.1.1 Bosons	5
2.1.2 Fermions	6
2.2 Particle Physics as a Quantum Field Theory	8
2.2.1 Quantum Chromo Dynamics	9
2.2.2 Electroweak Interactions	12
2.2.2.1 Spontaneous Symmetry Breaking and the Higgs Boson	13
2.2.2.2 The Higgs-Yukawa Couplings to Fermionic Fields	17
2.3 High Energy Proton Proton Scattering	18
2.3.1 Higgs Production at Proton-Proton Colliders	19
3 The Compact Muon Solenoid Experiment and the Large Hadron Collider	22
3.1 The Large Hadron Collider	24
3.1.1 Beam Parameters	24
3.1.2 The Pre-Accelerator Complex	26

3.1.2.1	Linac2	27
3.1.2.2	The Proton Synchrotron Booster	27
3.1.2.3	The Proton Synchrotron	28
3.1.2.4	The Super Proton Synchrotron	28
3.1.3	Acceleration in the LHC Ring	29
3.1.4	The Stable Proton Beam	32
3.1.5	The LHC Beam Dump	34
3.2	The Compact Muon Solenoid	36
3.2.1	The CMS Subsystems	37
3.2.1.1	The Silicon Tracker	37
3.2.1.2	The Pixel Detector	39
3.2.1.3	The Silicon Strip Detector	42
3.2.1.4	The Electromagnetic Calorimeter	43
3.2.1.5	The Hadronic Calorimeter	46
3.2.1.6	The Magnet	48
3.2.1.7	The Muon System	48
3.2.2	Event Triggering	51
3.2.3	Physics Object Reconstruction	52
3.2.3.1	Particle Flow	53
3.2.3.2	Tracks	53
3.2.3.3	Vertices	54
3.2.3.4	Electrons and Photons	54
3.2.3.5	Muons	55
3.2.3.6	Jets	56
3.2.3.7	Missing Transverse Energy (MET)	58
4	Quality Control Measurements for the CMS Phase II Outer Tracker Upgrade	60
4.1	The Phase II Tracker	61
4.1.1	The Phase II Outer Tracker	62
4.1.2	Design Specifications of 2S and PS-s Strip Sensors	63
4.2	Sensor Quality Control	64
4.2.1	Visual Inspection	66
4.2.2	CV and IV Characterization	66

4.2.3 Sensor Full Characterization	69
4.2.4 Long Term Stability Testing	74
5 Jets Originating from the Hadronization of Charm Quarks	76
5.1 Production of charm quarks at the LHC	77
5.2 Key Features of Charmed Meson Decays	77
5.2.1 Key Features for Identifying Charm Jets	81
5.3 Jet Tagging at the CMS Detector	82
5.3.1 DeepJet	85
5.3.2 Charm Tagging with DeepJet	87
5.4 Calibration of c-tagging scale factors	87
5.4.1 Phase Space Selections	90
5.4.1.1 c-enriched jets	90
5.4.1.2 b-enriched jets	91
5.4.1.3 light-enriched jets	92
5.4.2 Iterative Fit	92
5.4.3 Scale factor application	94
6 The search for $VH(H \rightarrow c\bar{c})$ at the CMS Detector with Full Run 2 Data	96
6.1 Analysis Overview	97
6.2 Simulations and Datasets	99
6.2.1 Data	99
6.2.2 Monte Carlo Samples	102
6.2.2.1 Signal Samples	102
6.2.2.2 Background Samples	103
6.2.3 Corrections to Differential Cross Sections of Simulated Processes	104
6.2.3.1 NLO in Electroweak Corrections to VH Differential Cross Section	104
6.2.3.2 NNLO in QCD and NLO in Electroweak Corrections to Diboson Differential Cross Section	108
6.3 Physics Object and Event Reconstruction	109
6.3.1 Primary Vertex	109
6.3.2 Lepton Identification	110
6.3.2.1 Electron Identification	110
6.3.2.2 Muon Identification	111

6.3.2.3 Jet Identification	111
6.3.2.4 Identification of Charm Jets	112
6.3.2.5 Charm Jet Energy Regression	113
6.3.3 Missing Transverse Energy	113
6.3.4 Data Taking Year Specific Corrections	113
6.3.4.1 ECAL Prefiring in 2016 and 2017	114
6.3.4.2 HEM15/16 Issue in 2018	114
6.4 Vector Boson Reconstruction	116
6.4.1 Vector Boson Transverse Momentum Re-Weighting in Z+jets Events	117
6.4.2 Jet Separation Re-weighting in Z+jets Events	118
6.4.2.1 Scale Factor Derivation	120
6.5 Higgs Boson Reconstruction	122
6.5.1 FSR Recovery	123
6.5.2 Application of a Charm Jet Energy Regression	124
6.5.3 Application of a Kinematic Fit in the 2-Lepton Channel	126
6.6 Event Selection	127
6.7 BDT Training	129
6.7.1 Intrinsic Performance of BDTs	132
6.7.2 BDT Binning and Reshaping	133
6.8 Background Modeling	134
6.8.1 Construction of Control Region Shapes for Signal Extraction	136
6.9 VZ Validation	136
6.10 Combination with Merged Jet Analysis	137
6.10.1 VH($H \rightarrow c\bar{c}$) Merged Jet Analysis Overview	138
6.10.2 Combination of Resolved and Merged Jet Analysis	138
6.11 Systematic Uncertainties	139
6.12 Results	144
6.12.1 Standalone Results	146
6.12.1.1 VH($H \rightarrow c\bar{c}$) Results	147
6.12.1.2 VZ($Z \rightarrow c\bar{c}$) Results	151
6.12.2 Results After Combination With Merged Analysis	153
6.12.2.1 VH($H \rightarrow c\bar{c}$) Results	153
6.12.2.2 VZ($Z \rightarrow c\bar{c}$) Results	154

6.12.3 Constraints on the Charm-Higgs Yukawa Coupling	154
7 Summary and Conclusion	158
7.1 The Future of $VH(H \rightarrow c\bar{c})$	159
A Categorization of the neutron irradiation spectrum at the Rhode Island Nuclear Science Center	161
A.1 The Interactions of Free Neutrons with Atomic Nuclei	162
A.2 The Need for Radiation Hardness in Particle Physics	164
A.2.1 Radiation Damage in Silicon	165
A.3 Spectrum Unfolding Using Foil Activation	166
A.3.1 Monitoring of Neutron Fluence via D0 Diodes	168
A.3.2 Measuring the Neutron Spectrum of the RINSC Reactor	169
A.3.3 Measurements of the Pneumatic Rabbit System Flux	177
A.3.4 Continual Monitoring of Neutron Fluence using Iron Foils	180
B Monitoring Silicon Strip Readouts Using Spy Data Taking	182
B.1 The Outer Tracker Readout System	183
B.1.1 Tracker Zero Suppression Algorithm	184
B.2 Spy Data Taking	186
B.2.1 Automated Processing and Transferring of Spy Data	187
B.3 Monitoring of Common Mode Slope Progression Using Spy Data	188
C Top Tagging Using End-to-End Images	192
C.1 The Identification of Top Quarks	193
C.2 Top and QCD Jet Datasets	194
C.2.1 CMS Opendata Samples	194
C.2.2 Jet Selection	195
C.3 The CMS Detector as an Image	195
C.3.1 Sub-ECAL Granularity Images	199
C.3.2 Network Architecture	200
C.4 Jet Identification Results	203
C.4.1 Interpretation of Results	204
C.5 Computing Architecture Training Speed Comparisons	206

List of Figures

2.1.1	Diagram showing the fundamental particles of the standard model. Leptons are grouped as quarks or fermions, and ordered in terms of generation. Semi-transparent clusters are drawn to show which particles interact with each vector boson. Image obtained from [1].	5
2.2.1	Example of Feynman diagrams showing fermion anti-fermion scattering at first order (left) and second order (right)	8
2.2.2	The estimated values of α_s at various energies based on eq 2.2.5 with a reference scale of $\alpha_s(m_Z)$ compared to data from multiple particle physics experiments. Figure from [2].	11
2.3.1	Proton parton distribution function at $\sqrt{s}$ energies of $\sqrt{10}$ GeV (left) and 100 GeV (right). The x-axis represents the fraction of proton momentum a colliding parton of said flavor contains, and the y-axis represents the probability of the parton having said momentum fraction multiplied by the momentum fraction itself. Distributions are made such that the integral of a single distribution represents the the total fraction of the proton momentum that will be carried by the corresponding parton. Figure obtained from [3].	19
2.3.2	The dominant production modes of the Higgs boson for proton-proton collisions. The production modes are gluon-gluon fusion (top left), vector boson fusion (top right), associated production with gauge bosons (bottom left), and associated production with top quarks (bottom right). Figure obtained from [4].	20
2.3.3	Cross sections of the various Higgs boson production modes for proton-proton collision with $\sqrt{s}$ in the TeV energy range. Figure obtained from [4].	21

3.0.1	Total luminosity delivered by the LHC and collected by the CMS detector collected during the 2015–2018 run 2 period. The 3.86 fb^{-1} of data collected during 2015 was not used for the $\text{VH}(\text{H} \rightarrow c\bar{c})$ search.	23
3.1.1	Full schematic of the CERN accelerator complex. In addition to the LHC, position of the four primary detectors, the pre-accelerator chain, and auxiliary experiments. Figure taken from [5].	26
3.1.2	Diagram outlining the 8 octants of the LHC along with their primary functionality. The stars indicate where the beams cross to produce collisions. Diagram obtained from [6].	30
3.1.3	Schematics of a cryomodule used to cool the RF cavities inside of the LHC. Diagram taken from [7].	31
3.1.4	Cross section of the dipole magnets used to bend the two proton beams inside of the arcs of the LHC. Diagram taken from [8].	32
3.2.1	Distribution of the number of simultaneous PU collisions per bunch crossing for the 2016 (left), 2017 (middle) and 2018 (right) data taking years.	37
3.2.2	Diagram showing the CMS detector along with its various subsystems. Diagram taken from [9].	38
3.2.3	A diagram of a quarter of the CMS tracker in the r – z plane showing both the outer (red and blue) and inner (green) subsystems. For the outer tracer, the layers/disks in blue represent the modules built in the stereo configuration. The inner tracker shown is for the phase 1 tracker used in the 2017 and 2018 data taking years.	40
3.2.4	Comparison of the geometries of the Phase 0 pixel detector used in the 2016 data taking year and the phase 1 pixel detector used in the 2017 and 2018 data taking years. Image taken from [10].	41
3.2.5	Diagram of the CMS detector in the transverse planer. The spatial locations of each detector components can be observed in the r and η coordinates as well as the distance along the beamline z	43
3.2.6	Simulated magnetic flux lines in the CMS detector. Image taken from [11].	49
3.2.7	Schematics of the L1 trigger system employed by the CMS detector. Information from the calorimeter and muon systems are combined to make decisions about whether or not an event should be sent for further processing within $4\text{ }\mu\text{s}$ of the collisions. Diagram taken from [12].	52

3.2.8	Diagram showing results of jet clustering when using the anti- k_T algorithm. Image taken from [13]	57
4.1.1	A diagram of the layout of one quarter of the phase II tracker in the r - z plane. The inner tracker is composed of pixels with two readout chips (green) and four readout chips (yellow). The outer tracker is composed of PS modules (red) and 2S modules (blue). Sensors located near the current TID are tilted towards the impact region to reduce the amount of active volume particle will travel through while also improving the probability that particles will pass through both sensors in a single PS module. Figure obtained from [14].	62
4.1.2	A diagram depicting the way in which a 2S module can differentiate between high and low p_T tracks. Detected high p_T tracks are identified as stubs and able to be used for L1 triggering. Diagram obtained from [14].	63
4.2.1	A computer-aided design drawing of a PS-s wafer. From a single wafer, two PS-s sensors (central left and central right) are fabricated. Additional test structures are built on the peripherals of the wafer to allow for PQC and irradiation testing.	65
4.2.2	Diagram depicting the circuitry configuration for CV and IV measurements. The decoupler consists to a pair of capacitors used to decouple the DC voltages applied from the high voltage supply from the terminals of the LCR meter.	67
4.2.3	A set of IV (left) and CV (right) curves for 2S sensors batch 39480. To ease human readability, applied voltage and measured currents are plotted as their absolute values. The double bump structure in the IV measurements is an observed feature of the sensors, the phenomenon is not fully understood but it is potentially due to the depletion of a secondary depletion zone forming inside of the sensors.	68
4.2.4	The configuration of probe needles when used for single strip measurements. The needle used to bias the ground needle is stationary with respect to the vacuum chuck, and as such will not move during the measurement. Between individual strip measurements, the chuck will lower, move to the next strip position, and raise such that the needles probing the strips will once again make contact.	70

4.2.5	The strip leakage current (top left), polysilicon resistance (top right), coupling capacitance (middle left), dielectric current (middle right), interstrip resistance (bottom left), and interstrip capacitance (bottom right) measured for a subset of strips on select sensors from PS-s batch 36697 at Brown University. Black dashed lines are used to represent the design specifications listed in Table 4.2 for a PS-s sensor.	73
4.2.6	The sensor holder used to perform long term stability testing (left) and the enclosure used to keep multiple sensors in stable conditions during measurements (right). Up to 10 sensors can be loaded into the enclosure and kept at high voltages simultaneously. The sensors inside of the enclosure are kept at a stable temperature of 20° C and dry air is pumped into the enclosure to keep the relative humidity below 10% during measurements.	74
4.2.7	The reverse bias leakage current for select sensors from PS-s batches 36696 and 36697 taken at a backplane potential of -600 V over a Z day period. Individual current measurements were performed every 60 s.	75
5.2.1	Feynmann diagrams of the Drell-Yan and three gluon decay modes of the J/ψ (top left and top right), semileptonic decay mode of the D^+ (bottom left), and W decay channel D_s (bottom right).	78
5.3.1	Diagram showing how a jet may contain a secondary vertex (SV) and it's relationship to the impact parameter (IP) variable commonly used for jet tagging algorithm. Figure taken from [15].	82
5.3.2	Comparison of the value of the IP significance of the track with the largest IP value for b -jets, c -jets, and uds -jets. Figure taken from [15].	83
5.3.3	Diagram showing types of inputs and network architecture used by the DeepCSV jet tagging algorithm. Track features were fed into the network for the six tracks with the highest IP values, as well as a collection of features related to the SV of the jet and overall jet properties.	84
5.3.4	Comparison of the c -tagging performance as a function of b -jet and uds -jet mistag rate for early Run 2 jet taggers. Values closer to the bottom right are considered better.	85

5.3.5	Diagram showing the network architecture and types of input features used by the DeepJet algorithm. The network can take features associated with up to 25 charged PF candidates, 25 neutral PF candidates, and four potential secondary vertices as inputs. The network then makes use of convolution, recursive, and fully connected network layers to produce multiple jet flavor discriminants.	86
5.3.6	Comparison of the DeepCSV and DeepJet CvsL (left) and CvsB (right) distributions for b -, c -, and uds -jets with $p_T > 20$ GeV and $ \eta < 2.5$. Jets generated using 2017 $t\bar{t}$ MC samples.	88
5.3.7	Receiver operator characteristics (ROC) showing the c -tagging mis-tag rate as a function of tagging efficiency for CvsL (left) and CvsB (right) discriminators.	88
5.4.1	Distributions of CvsL (left) and CvsB (right) scores for soft muon containing jets obtained from the same-sign – opposite-sign $W+c$ event selection in both data and simulation. Generator level information from the simulated samples shows that the selection leads to a phase space heavily enriched in c -jets. Figures taken from [16].	91
5.4.2	Full 2D profiles in CvsL and CvsB of DeepJet scale factors for 2017 data taking year. A separate set of scale factors are derived and applied for light-jets (top left), c -jets (top right), and b -jets (bottom).	95
6.1.1	The Feynmann diagrams for the three dominant $VH(H \rightarrow c\bar{c})$ production modes in proton-proton collisions.	98
6.2.1	Event weights applied to the WH (left) and ZH (right) signal samples as a function of generated $p_T(V)$ to correct for the NLO in electroweak corrections to the differential cross section.	108
6.2.2	Differential NNLO in QCD and NLO in EW k-factors applied as a function of the generated $p_T(V)$ of the subleading boson in WW (left), WZ (middle), and ZZ (right) processes. Correction factors were derived and distributions were produced by the $VH(H \rightarrow c\bar{c})$ merged jet analysis team.	109
6.3.1	Distribution of $\phi(\text{MET})$ in the 0L channel for the 2018 data taking year before (left) and after (right) application of the $\phi(\text{MET})$ veto in data and weighting in MC. A residual deficit in data yield can still be observed in the opposite side of the ϕ spectrum.	115
6.4.1	The distribution Drell-Yan NLO 2018 MC events before and after corrections (right).	118

6.4.2 Inclusive distribution of the $p_T(V)$ variable for combined high and low $p_T(V)$ channels in the $Z(\ell\ell)H$ control region used to derive the correction factors. Distributions are plotted before (top) and after (bottom) the correction factors are applied for all three data taking years.	119
6.4.3 Distribution of the $\Delta R(jj)$ variable in the light flavored control regions of the 2-lepton, 1-lepton, and 0-lepton analysis channels in the 2017 data taking year. Residual mismodeling in simulation is observed for events with $\Delta R(jj) < 1$	119
6.4.4 Shape of $\Delta R(jj) < 1.0$ distribution of W+Jet (left) and Z+Jet (right) events for various jet flavor compositions.	120
6.4.5 Shape of $\Delta R(jj)$ distribution of the 2-lepton (left), 1-lepton (middle) and 0-lepton (right) channels in the light-flavor control regions for the 2016 data taking year after application of the $\Delta R(jj)$ scale factors.	121
6.4.6 Shape of $\Delta R(jj)$ distribution of the 2-lepton (left), 1-lepton (middle) and 0-lepton (right) channels in the light-flavor control regions for the 2017 data taking year after application of the $\Delta R(jj)$ scale factors.	121
6.4.7 Shape of $\Delta R(jj)$ distribution of the 2-lepton (left), 1-lepton (middle) and 0-lepton (right) channels in the light-flavor control regions for the 2018 data taking year after application of the $\Delta R(jj)$ scale factors. For the 0 lepton channel, the fit was performed in such a way that the normalization offset between data and MC was preserved for low and high $\Delta R(jj)$ events.	122
6.5.1 Reconstructed Higgs mass of 2016 $Z(\ell\ell)H(c\bar{c})$ signal events with (blue) and without (red) the application of the FSR recovery procedure for the 2-lepton low (left) and high (right) $p_T(V)$ analysis channels binned inclusively for electron and muon Z-boson decays. The mass peaks are fitted with a double crystal ball function to obtain the dashed lines.	124
6.5.2 Distribution of the reconstructed Higgs mass with and without the application of the c -jet energy regression for $Z(\ell\ell)H(c\bar{c})$ high $p_T(V)$ (left) and $W(\ell\nu)H(c\bar{c})$ (right) events. Central value and standard deviation of the distributions are computed with a double crystal ball function. For the 2-lepton analysis channel, the c -jet regression is also used with a kinematic fit to further improve the calculation of the Higgs boson mass.	125
6.7.1 Overtraining curves for the BDTs train to perform $VH(H \rightarrow c\bar{c})$ signal separation in the 2018 data taking year.	132

6.7.2 Receiver operator curves for the BDTs trained to perform $VH(H \rightarrow c\bar{c})$ signal separation in the 2018 data taking year.	132
6.7.3 Overtraining curves for the BDTs train to perform $VZ(Z \rightarrow c\bar{c})$ signal separation in the 2018 data taking year.	132
6.7.4 Receiver operator curves for the BDTs trained to perform $VZ(Z \rightarrow c\bar{c})$ signal separation in the 2018 data taking year.	133
6.10.1 Reconstruction efficiency of different ways to reconstruct a Higgs boson. At p_T greater than 300 GeV, the reconstruction efficiency for a H-candidate reconstructed from the highest p_T AK15 jet in an event approaches unity.	139
6.12.1 Control regions used to constrain light-jet dominated (far left), b -jet dominated (middle left), and c -jet dominated (middle right) Z +jet events and $t\bar{t}$ dominated (far right) events in the high- (top two) and low- $p_T(V)$ (bottom two) 2-lepton analysis channels for the 2018 data taking year. MC event yields and uncertainties are set to their best fit values obtained from the fit to the full Run 2 dataset.	149
6.12.2 Control regions used to constrain light-jet dominated (left) and c -jet dominated (middle) W +jet events and $t\bar{t}$ dominated (right) events in the 1-lepton analysis channel (top two) and the light-jet dominated (far left), b -jet dominated (middle left), and c -jet dominated (middle right) Z +jet events and $t\bar{t}$ dominated (far right) events in the 0-lepton for the 2018 data taking year. MC event yields and uncertainties are set to their best fit values obtained from the fit to the full Run 2 dataset.	150
6.12.3 Signal regions used to extract the best fit $VH(H \rightarrow c\bar{c})$ signal strength for the 2018 data taking year. Nuisance parameters and signal strength applied to MC event yields and uncertainty bands were obtained from the full Run 2 fit to all analysis channels. Because the best fit signal strength was measured to be $\hat{\mu} = -9.50^{+9.94}_{-9.54}$, the signal yield is set to 0.	151
6.12.4 Signal regions used to extract the best fit $VZ(Z \rightarrow c\bar{c})$ signal strength for the 2018 data taking year. Nuisance parameters and signal strength applied to MC event yields and uncertainty bands were obtained from the full Run 2 $VZ(Z \rightarrow c\bar{c})$ fit to all analysis channels.	153
6.12.5 Observed and expected 95% CL upper limit on the $VH(H \rightarrow c\bar{c})$ signal strength for the various channels, topologies, and the combined fit to all channels, topologies, and data taking years.	155

6.12.6	Distribution of signal region bin counts weighted as a function of $\mathcal{S}/\mathcal{B}$, where $\mathcal{S}$ and $\mathcal{B}$ are the expected signal and event yields in each of the $VH(H \rightarrow c\bar{c})$ signal region bins of either of the two topologies.	155
6.12.7	Distribution of signal region bin counts weighted as a function of $\mathcal{S}/\mathcal{B}$, where $\mathcal{S}$ and $\mathcal{B}$ are the expected signal and event yields in each of the $VZ(Z \rightarrow c\bar{c})$ signal region bins of the resolved jet topology (left) and combined resolved and merged jet topologies (right).	156
6.12.8	Parameter scan over $ \kappa_c $ with and the corresponding test statistic calculated with the combined topology fit results using eq 6.12.2.	157
A.2.1	The measured post-irradiation leakage current at the depletion voltage in n-type and p-type silicon as a function of equivalent neutron dose. Increases in leakage current were found to have a linear relationship with radiation fluence, and were found to be independent of the type of radiation after conversion of the effective dose. Image taken from [14].	165
A.3.1	An image of the foil holder used to put foils into the 6"a beamport. The foils were placed in small holes bowed into a milled wooden disc and secured using a polyethylene disc attached to the face of the wood.	170
A.3.2	Neutron activation cross sections as a function of incident neutron energy for the 6 absorption processes considered when fitting the neutron energy spectrum in each of the beam ports.	171
A.3.3	The derived neutron energy spectrum and fit normalizations for the neutron energy spectrum found for the 6"a (left), 6"b (middle), and 8" (right) beamports at the RINSC reactor.	172
B.1.1	The data stream readout from a pair of APV25 chips with a minimum ionizing particle (MIP) signal injected. The digital header is used to inform the FED that input data is coming as well as signify the corresponding address of the APV25 pair. The tails are then used to inform the FED that the data stream has ended, with the second tail existing for redundancy.	183

B.1.2	The various forms that the strip data takes after being re-ordered (top left), having the pedestal subtracted (top right), having the common mode subtracted (bottom left), and after the full zero-suppression algorithm (bottom right). The post common mode subtracted distribution is plotted beside the assumed strip noise (red). In the zero-suppressed form, only information about strips with detected hits are sent to the DAQ.	185
B.3.1	The average (left) and root mean squared (right) of the common mode slope distributions fit for every APV25 chip in the outer tracker during spy run 321779 in bins of 500 spy events. In all partitions, a residual linear trend that increased as the runs went on was observed. However, the overall size of the trends are too small to have an effect on the zero suppression algorithm.	188
B.3.2	The distribution average ADC counts for the post pedestal subtracted ADC count of each APV25 chip in bins of 500 events (left) and for events 12000–1276 (left) for spy run 321779. A slight increase in the ADC count is observed at the end of the run, which we believe is due to the tracker changing configurations to perform cosmic data taking.	189
B.3.3	The RMS of the common mode slope distribution per event for events 6000–6500 of run 321779 (left) and the common mode slope distribution for the different layers of the TOB in event 6262 (right) which was found to have a large RMS value.	190
C.3.1	An image representation of a CMS detector image for a $t\bar{t}$ event. All image layers described in this section are shown, with the exception of the track $d0$ and dz layers which are covered by the track p_T layer.	196
C.3.2	The set of BPix layer 1 (top left), BPix layer 2 (top middle) BPix layer 3 (top right), track p_T (bottom left), ECAL (bottom middle), and HCAL (bottom right) for one of the jets shown in Figure C.3.1. The bottom portions of the BPix layer 2 and 3 image layers have no BPix due to the lower η acceptance of the outer layers of the BPix.	198
C.3.3	Comparison of jet images for a simulated top jet at the nominal ECAL resolution (left) and after subdividing the ECAL cells by a factor of 3 (middle) and 5 (right).	200

C.3.4 A pictorial diagram of a ResBlock (left) and the ResNet-15 architecture used for the network (right). Filter size in ResBlocks are modifiable, but for this study all ResBlocks used a filter size of 3×3 . ReLU activation functions exist after the initial convolutional layer, and the fully connected layer is passed through a sigmoid activation function before being passed to the output node.	201
---	-----

List of Tables

2.1	The eigenvalues of weak isospin (I_3) and hypercharge (Y) for the different fermion flavors of the standard model. The left- and right-handed chiral projections of the fermionic spinors transform separately under $SU(2)_L \times U(1)_Y$ giving rise to two sets of values. No values are listed for right-handed neutrinos since no experimental confirms their existence.	13
2.2	Masses of the SM massive gauge bosons as a function of free parameters in the SM, as well as the experimentally measured values. Numbers quoted are averages of many experimental measurements aggregated by [4].	17
3.1	Designed LHC Beam parameters according to [7] and peak values reached during the run 2 data taking period.	25
3.2	Parameters related to the different accelerators used in the LHC proton pre-accelerator complex. The number of pulses for the SPS to fill the LHC ring is dependent on the filling scheme used during the various data taking years.	29
4.1	Various dimensional quantities for 2S, PS-s, and PS-p sensors. Active refers to the dimensions of the p-type silicon bulk, and outer refers to the dimensions of the entire sensors. All distance quantities are given in units of μm	64
4.2	Design specification for strip and interstrip properties when the sensors are biased at a -600 V backplane voltage. Normalized measurements are normalized per cm, with the exception being the interstrip resistance which is normalized in units of $\text{G}\Omega\text{cm}$. 2S and PS-s specifications are set by approximating the strip lengths as 5 cm and 2 cm respectively.	69

5.1 Mass, lifetime, decay distance, and dominant branching fractions of the ground state charmed mesons. Exact values obtained from [4] and approximate values obtained based on available phase space for decay modes. Uncertainties dropped from mass, lifetime, and $c\tau$ measurements and values truncated to two significant figures for readability.	79
6.1 2016 datasets used for the analysis and HLT paths used for their corresponding event triggers.	100
6.2 2017 and 2018 datasets used for the analysis and HLT paths used for their corresponding event triggers.	101
6.3 CMS datasets names used to obtain the $VH(H \rightarrow c\bar{c})$ simulated events for all three data taking years. Cross sections obtained from [17].	103
6.4 CMS datasets names used to obtain the $VH(H \rightarrow b\bar{b})$ simulated events for all three data taking years. Cross sections obtained from [17].	103
6.5 Processes used to model standard model background for 2016 data taking year. Events in the WZTo1L1Nu2Q, ZZTo2L2Q, and ZZTo2Q2Nu processes in which generator level information showed that a Z-boson decayed to a pair of c -jets was used to model the signal in the VZ($Z \rightarrow c\bar{c}$) validation analysis.	105
6.6 Processes used to model standard model background for 2017 data taking year. Events in the WZTo1L1Nu2Q, ZZTo2L2Q, and ZZTo2Q2Nu processes in which generator level information showed that a Z-boson decayed to a pair of c -jets was used to model the signal in the VZ($Z \rightarrow c\bar{c}$) validation analysis.	106
6.7 Processes used to model standard model background for 2018 data taking year. Events in the WZTo1L1Nu2Q, ZZTo2L2Q, and ZZTo2Q2Nu processes in which generator level information showed that a Z-boson decayed to a pair of c -jets was used to model the signal in the VZ($Z \rightarrow c\bar{c}$) validation analysis.	107
6.8 Correction factor parameters extracted from fit for 2016, 2017, and 2018 data taking years.	117
6.9 The percentage of events in which the two jets used to reconstruct the H-candidate properly reconstruct the Higgs boson in $VH(H \rightarrow c\bar{c})$ signal events for various selection methods. We found that with the exception of $Z(\nu\nu)H$ events, in which reconstruction efficiency had a small dependence on selection methodology, selecting jets based on their CvsL score had the highest probability of properly selecting the two c -jets. . .	123

6.10	The mean and width of the reconstructed H-candidate mass obtained from a double crystal-ball fit with (bottom) and without (top) use of regressed c -jets. In the bottom table, the relative improvement after applying the kinematic fit is lower, because it is being compared to the already improved upon m_H resolution obtained from the energy regression algorithm.	127
6.11	Baseline selection used to construct the signal region in each of the four analysis categories. For each channel, events are first pre-selected based on whether the corresponding V-boson decay being reconstructed ($Z \rightarrow \nu\nu$, $W \rightarrow \ell\nu$, $Z \rightarrow \ell\ell$) and having the H-candidate reconstructed from two AK4 jets.	128
6.12	Primary hyper parameters used to train the BDTs. For the 2- and 0-lepton channels, there were no differences in parameters used between the $VZ(Z \rightarrow c\bar{c})$ and $VH(H \rightarrow c\bar{c})$ BDT trainings. However, a smaller number of trees were used for the $WZ(cc)$ BDT due to lower number of MC signal sample events.	130
6.13	Input variables for the BDTs trained in each of the analysis channels. The 1-lepton channel uses an additional set of variables to help discriminate against $t\bar{t}$ events, and the 2-lepton channel uses the variables extracted from the kinematic fit described in Section 6.5.3.	131
6.14	The parameters and their ranges which were scanned over when performing the BDT optimization procedure. x and y refer to the free parameters in eq 6.7.1, and with the exception of the final bin all other bins remain the same width.	134
6.15	Additional selections used to construct the four control regions in each of the analysis channels. For all variables not mentioned, the same selections used to construct the signal regions are used.	135
6.16	Overview of the systematic uncertainties considered in the analysis, how they are correlated across the topologies, analysis channels, physics processes, and data taking years, as well as whether we considered the uncertainty to effect the overall shape of the distribution or only present itself in an overall change in event normalization (rate).140	
6.17	Expected and observed 95% CL limits for the VH production cross section times $H \rightarrow c\bar{c}$ branching ratio signal strength, for the various analysis channels, data taking years, and the full Run 2 fit.	148
6.18	Background normalization rate parameters derived from the $VH(H \rightarrow c\bar{c})$ full Run 2 fit to all analysis channels.	148

6.19	Relative contribution to the uncertainties in the $\hat{\mu}_{\text{VH}(\text{H} \rightarrow c\bar{c})}$ signal strength from the standalone resolved jet analysis.	151
6.20	Expected and observed significance of the best fit $\text{VZ}(\text{Z} \rightarrow c\bar{c})$ signal strength in the various analysis channels and data taking years.. . . .	152
6.21	Background normalization rate parameters derived from the $\text{VZ}(\text{Z} \rightarrow c\bar{c})$ full Run 2 fit to all analysis channels.	152
6.22	Expected and observed 95% CL upper limits on the $\text{VH}(\text{H} \rightarrow c\bar{c})$ signal strength after combination of the resolved and merged jet topologies for all three data taking year. The quoted $\hat{\mu}_{\text{VZ}(\text{Z} \rightarrow c\bar{c})}$ is the best-fit value obtained for the fit to all channels and topologies in the corresponding row.	154
6.23	Expected and observed $\text{VZ}(\text{Z} \rightarrow c\bar{c})$ significance and the best-fit combined channel signal strength after combination of the resolved and merged jet topologies for all three data taking year. After combination of all channels, topologies, and data taking years we observe the $\text{VZ}(\text{Z} \rightarrow c\bar{c})$ signal strength at over 5σ	156
A.1	Comparison of the measured saturated activity to the values calculated from the parameterized fit for each of the three beam port measurements in units of decays/nuclei/s. The calculated values were obtained using the neutron energy spectra shown in Figure A.3.3.	173
A.2	The 1 MeV equivalent neutron flux for the three beam ports as measured via spectrum unfolding from foil activation.	173
A.3	Measurements associated with the 6"a beam port irradiation. Measured activity from gamma spectroscopy (A^γ), the back calculated activity at the end of the irradiation A^{irrad} , and the associated saturated activity A^∞ are given for each irradiated sample. In the case of iron, Fe54, Fe56, and Fe58 measurements were derived from the same set of foils.	174
A.4	Measurements associated with the 6"b beam port irradiation. Measured activity from gamma spectroscopy (A^γ), the back calculated activity at the end of the irradiation A^{irrad} , and the associated saturated activity A^∞ are given for each irradiated sample. In the case of iron, Fe54, Fe56, and Fe58 measurements were derived from the same set of foils.	175

A.5	Measurements associated with the 8" beam port irradiation. Measured activity from gamma spectroscopy (A^γ), the back calculated activity at the end of the irradiation A^{irrad} , and the associated saturated activity A^∞ are given for each irradiated sample. In the case of iron, Fe54, Fe56, and Fe58 measurements were derived from the same set of foils.	176
A.6	The 2017 saturated activity values measured and obtained through normalized beam port fit of the saturated activities (in decays/nuclei/s) used to measure the rabbit system flux. Because the foils were encased in cadmium, the Fe58(n, γ)Fe59 activation processes was not used to calculate the 6"b beam port normalization.	178
A.7	The 2018–2019 saturated activity values measured and obtained through normalized beam port fit of the saturated activities (in decays/nuclei/s) used to measure the rabbit system flux. [†] Mn56 activity measurement appeared orders of magnitude higher than expected, and did not agree with activity Mn54 and Fe59 activities measured on the same foil. As such, those activity values were not used in the fit and are omitted from the average values quoted in the table. ^{††} Foils encased in cadmium were not used when computing the average activity values for these foils	179
A.8	2017 rabbit system flux and fluence measurements from the 6"b and 8" beamport re-normalizations, D0 diode measurements, and the desired values. The Fe58(n, γ)Fe59 activation measurement were not used when computing the flux based on the 6"b beamport measurement. The flux is shown in units of $n_{\text{eq}}/\text{cm}^2/\text{s}$ and the fluence is in units of $n_{\text{eq}}/\text{cm}^2$. Measurements are in the same order as Table A.6.	180
A.9	2018–2019 rabbit system flux and fluence measurements from the 6"b and 8" beamport re-normalizations, D0 diode measurements, and the desired values. The flux is calculated in units of $n_{\text{eq}}/\text{cm}^2/\text{s}$ and the fluence is in units of $n_{\text{eq}}/\text{cm}^2$. Measurements are in the same order as Table A.7. [†] foils encased in cadmium not used when computing 6"b beamport fluence estimates. ^{††} Irradiation time needed to reach the desired fluence assumed a lower $n_{\text{eq}}/\text{cm}^2$ flux.	180
C.1	List of the CMS Open Data simulated samples used and total number of events considered for each of the datasets.	194

C.2	Number of jets used for training and validating the top quark and non-top quark jet categories after the jet p_T re-sampling of the QCD jets. When training the network on subsets of layers, only 198k jets were used for validation.	195
C.3	Performance of the classifier trained up to 20 epochs. Results are shown for different combinations of tracking and calorimeter layers and evaluated on a sample of 198k jets.	203
C.4	Comparison of architecture specification for the NVIDIA Tesla P100 [18], NVIDIA Tesla V100 [19], and Google TPUv3-8 [20] architectures. The floating point operations per second speed (FLOPS) are quoted based on the data type and architecture setup used during training.	206
C.5	Comparison of I/O and training time for different computing architectures. A larger batch size was used when training on the Tesla V100 and TPUv3-8. Training times were found to vary by approximately 10% between epochs.	208

Chapter 1

Introduction

On July 4th, 2012 the CMS and ATLAS collaborations announced the joint discovery of the Higgs boson [21, 22]. The Higgs had been predicted as a fundamental particle since 1964 [23], with its primary feature being that it generates the masses of the fundamental particles. Since then, the continuation of the study of the Higgs and measuring its couplings to the other fundamental particles has been a primary goal of both collaborations, and is a major factor for the planning for future collider experiments [24, 25, 26]. As it currently stands, physicists have now been able to directly observe the Higgs boson couplings to the following particles, and found that the observations agree with theoretical predictions: the W [27, 28], the Z [29, 30], top quarks [31, 32], bottom quarks [33, 34], and taus [35, 36]. Additionally, there are now measurements of the Higgs boson's coupling to muons [37, 38], with evidence of the process reported by the CMS collaboration. In this thesis, I present the continuation of our quest to understand the Higgs boson via the latest CMS search for a Higgs boson decaying to charm quarks, giving us a direct method for measuring the charm quark Yukawa-coupling to the Higgs field. This is a continuation of a study performed on a set of CMS data collected in 2016 [39], with the new iteration now being performed on the full CMS Run 2 data set consisting of 138 fb^{-1} of data and introducing a plethora of improvements to the overall analysis techniques.

The thesis is written with the narrative focusing on the $H \rightarrow c\bar{c}$ analysis, with discussions of additional projects I worked on primarily present in the appendix. As such, the thesis is formatted as follows: Chapter 2 contains an introduction to the main theoretical concepts needed to understand this thesis. It discusses the Standard Model and the mechanisms of the Higgs boson. It then ends with an additional section on proton-proton collisions with a primary focus on the parton distribution

function, its implications for the Large Hadron Collider, and how Higgs bosons can be produced from proton-proton collisions.

Chapter 3 then gives an overview of the Large Hadron Collider (LHC) and the Compact Muon Solenoid (CMS) detector which was used to collect the data for the primary study presented in this thesis. In discussing the LHC, I walk the reader through the steps taken to go from stable H_2 gas to creating co-rotating bunches of protons traveling at 99.9999991% the speed of light. When discussing the CMS experiment, I introduce the primary partitions of the detector, their purpose, and the primary design principles under which they operate. I then describe the ways in which we transform the raw detector output into a data format usable by physicists to perform a physics analysis.

Chapter 4 is a slight detour from the primary narrative of the thesis to briefly talk about the upgrade currently being developed for the CMS outer tracker, and one of the ways in which Brown University is involved in its development. I have spent a large part of my time at Brown actively working in our lab and helping to measure the quality of sensors which will be used to build the next generation of the outer tracker. As such, I could spend hours discussing the details of this subdetector and the years worth of experiences that I have had in our lab. But unfortunately, time is short, my thesis is long, and as such I am only able to give brief details on the portions of our lab duties related to sensor quality control with a primary focus on the measurements I have had a direct influence on.

Chapter 5 returns to the primary narrative and gives an overview of charm quarks and their properties. In particular, I discuss the types of matter charm quarks manifest in at the LHC and how the signatures associated with the production of charm matter looks through the lens of the CMS detector. I then go over the ways in which the CMS collaboration aims to identify jets, which are the type of particle showers produced from the decay of quark matter. In doing so, I provide an extra focus on the DeepJet algorithm [40] which is used as part of the $H \rightarrow c\bar{c}$ search as well as the way in which DeepJet is calibrated to perform identically on real LHC data and simulated data produced from Monte Carlo simulation.

Chapter 6 is the chapter describing the full analysis techniques used to perform the low p_T portion of the $H \rightarrow c\bar{c}$ search. For this search, we make use of the VH production mode in which the Higgs boson is produced in association with a W or Z vector gauge boson which we heavily make use to aid in our ability to identify $H \rightarrow c\bar{c}$ events. I then briefly introduce a second $H \rightarrow c\bar{c}$ analysis aimed at identifying high p_T Higgs bosons and how its results are combined with the low p_T analysis in which I worked on. I present the new set of limits we are able to place on the

VH($H \rightarrow c\bar{c}$) process as well as the charm quark Yukawa coupling, both of which are currently the strongest limits to date. Additionally, I discuss the results of a validation analysis in which the VH($H \rightarrow c\bar{c}$) analysis framework is re-cast to look for events where a Z-boson which decays to a charm quark-antiquark pair is produced with an additional Z- or W-boson. Through the combined results, the CMS collaboration is now the first experiment to observe the $Z \rightarrow c\bar{c}$ process above a 5σ level at a hadron collider.

The appendix of this thesis then presents three additional projects which I worked on during my time as a graduate student at Brown University. Appendix A is an overview of a project aimed at measuring the neutron energy spectrum of the RINSC nuclear reactor. RINSC is now actively being used for irradiation testing of sensors that will be used to construct the next upgrade of the CMS detector. Appendix B describes spy data, a special type of data in which the outer tracker of the CMS detector is able to be read out in its raw format during data taking. I then discuss a data quality management project which I was able to use this special type of data for. Appendix C gives an overview of a project in which computer vision techniques were used to identify jets originating from the decay of top quarks. Additionally, this is the first jet identification algorithm which makes use of the location of particles as they pass through the tracker as a jet discriminator input.

Chapter 2

The Standard Model

The standard model (SM) of particle physics is the theory in which we describe the fundamental particles of our universe and the ways in which they interact. Despite being proposed more than 50 years ago, it has proven robust in its ability to describe the fundamental nature of the universe. Even in instances where new physics not yet described by the SM had been observed, they could later be naturally incorporated into the theory like a missing piece being added to a jigsaw puzzle. With the Large Hadron Collider (LHC) we are now once again testing the limits of the SM by measuring its predictions with unprecedented precision and searching for new phenomenon beyond the standard model (BSM).

At its core, the SM is a gauge theory governed by the symmetries of the non-Abelian $SU(3)_C \times SU(2)_L \times U(1)_Y$ symmetry group. Each of the three symmetry groups corresponds to a fundamental force, where $SU(3)_C$ corresponds with the strong interaction and $SU(2)_L \times U(1)_Y$ corresponds with the combined electroweak interaction. At lower energies, this then further breaks into the $SU(2)_L$ weak interaction and the $U(1)_{EM}$ electromagnetic interaction. In this chapter, I will give a brief overview of the fundamental particles of the SM and the quantum field theory framework in which they are defined. I will then briefly discuss the fundamental features of proton-proton collisions, and how they can lead to the production of a Higgs boson at the LHC.

2.1 The Fundamental Particles

The SM predicts that the most basic building blocks of matter and the forces in which they interact exist as a series of fundamental particles which cannot be further subdivided. Each fundamental particle can be described by a series of intrinsic properties primarily describing the ways in which

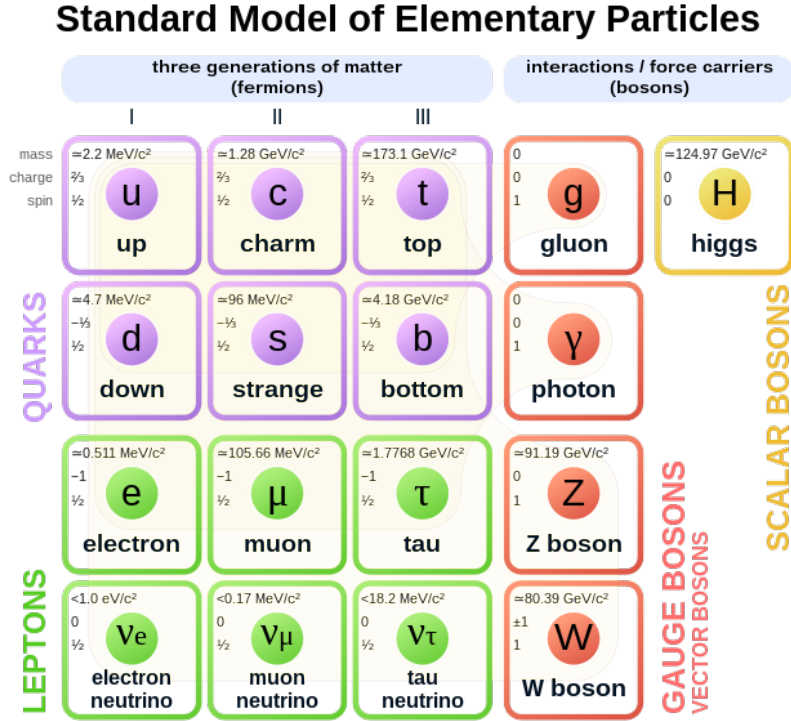


Figure 2.1.1: Diagram showing the fundamental particles of the standard model. Leptons are grouped as quarks or fermions, and ordered in terms of generation. Semi-transparent clusters are drawn to show which particles interact with each vector boson. Image obtained from [1].

they interact. From one of these properties, fundamental particles can be broken into two separate families: Fermions which make up the building blocks of matter, and bosons which carry the fundamental forces/interactions between particles. Beyond a naming convention, these family of particles are separated based on their intrinsic properties of spin. Fermions have half-integer spin causing them to follow Fermi-Dirac statistics, while boson have an integer spin causing them to follow Bose-Einstein statistics [41]. A full diagram showing all fundamental particles along with their mass, electromagnetic charge, and spin is shown in Figure 2.1.1.

2.1.1 Bosons

In addition to all having integer spin, the bosons include the family of particles which mediate the various fundamental forces described by the standard model. There are a total of 12 force carrying vector (spin 1) gauge bosons: The photon (γ), the Z, two Ws (W^+ and W^-), and eight gluons (g). Additionally, there is a single scalar (spin 0) particle that is the Higgs boson (H). The photon is the massless and chargeless particles which mediates the electromagnetic interactions. As such,

photons only interact with electromagnetically charged particles. Similar to the photon is the gluon, it is also massless and has no electromagnetic charge. However, it carries the nuclear strong force which acts through the exchange of *color* charge. Additionally, gluons contain color and anti-color charge, allowing them to self-interact while also creating a total of eight different gluons based on the possible combinations of color and anti-color that they can carry. The W- and Z-bosons are then the massive mediators of the weak force. In particular, the $W^\pm$ boson mediates nuclear decay, and can induce flavor change. Furthermore, due to the charged nature of the W (the + or - is sometimes dropped when referring to the particle) it is able to interact with photons. The Z on the other hand interacts in the same instances as a photon, with the added property that, in addition with the W-boson, it is able to interact with electrically neutral particles, including itself.

2.1.2 Fermions

Fundamental fermions can exist as one of 12 unique flavors. The 12 fermionic flavors can then be further split into 6 leptons and 6 quarks, which can both be further subdivided into three unique pairs referred to as generations. It is according to these generations that the columns of fermions are split in Figure 2.1.1. For both leptons and quarks, the pairs of particles exhibit very similar behavior across generation. The primary difference being that the masses of particles increase as you go up in generation. Due to the W-boson's ability to mediate particle decay, this leads to higher generation fermions rapidly decaying to lighter particles within their families, until you are left with only first generation particles. For this reason, the first generation contains the up (u) and down (d) quarks which make up protons and neutrons, as well as electron (e) which makes up the outer portion of atomic matter.

In addition to the electron, the electron-neutrino (ν_e) is the second first generation lepton. The ν_e is an electrically neutral almost-massless particle which always accompanies the electron upon interactions with a W-boson. As we go up in generation, we then have the muon (μ) and tau (τ) which behave as the more massive partners to the electron. As such, they both have an electric charge of -1. They also both contain an electrically neutral and near-massless ν_μ and ν_τ in their corresponding generations. As fermions, the primary features of leptons is that they do not interact via the strong interaction, they have integer charge, and that the neutrinos are weakly interacting particles. Of further interest for neutrinos is that we know that they possess unique masses due to the observation of neutrino oscillations [42], but so far have not been able to make any direct measurements of their mass. Additionally, the muon occupies an interesting space in particle physics

due to its mass being approximately 200 times larger than an electron's as well as having a lifetime of approximately $2.2 \mu\text{s}$ [4]. These two features allow muons to travel large distances through matter before decaying, and also exhibit a large radius of curvature when being bent via an external force. Both of these features can be exploited for particle detection.

The primary features of quarks is that they interact through the strong force and that they have fractional electric charge. The fractional charge is a result of their relationship with the electroweak interaction, and causes the up quarks to have a charge of $+2/3$ and down quarks a charge of $-1/3$. As such, the second and third generations of quarks each contain an up-like quark with charge of $+2/3$ and down-like quark of charge $-1/3$. These are respectively the charm (c) and top (t) quarks for the up-like quarks and the strange (s) and bottom (b) for the down-like quarks. Of the higher generation particles, the most unique is the top quark which has a mass of approximately 172.9 GeV [4] making it the most massive fermion by a factor of 40. Because quarks interact with the strong force, each quark carries color charge. Similarly to the gluon, this causes each flavor of quark to be able to exist in one of three separate color states. As such, there are a total of 18 unique quarks constituting the 6 different flavors and their 3 different color states. Due to the additional color states, we often find that the electroweak bosons decay to specific quark flavors at roughly 3 times the rate they would decay to specific lepton flavors.

The fundamental fermions described by the standard model are all spin $1/2$. As such, they all behave as spinors according to the Dirac Equation:

$$(i\hbar\gamma^\mu\partial_\mu - mc)\psi = 0 \tag{2.1.1}$$

This produces a few additional interesting features common to all quarks and leptons. The first is that each fermion can be decomposed into a left-handed (ψ_L) and right-handed (ψ_R) chiral state. As will be discussed in Section 2.2.2 this leads to unique behaviors via the electroweak interactions. The second is that for every fundamental fermion, there must be an anti-fermion ($\bar{\psi}$) which exists as a solution to the adjoint Dirac Equation. Anti-fermions have been observed to exhibit all of the same behaviors as normal counterparts, except that all charge/eigenvalues associated with the anti-fermions have opposite sign.

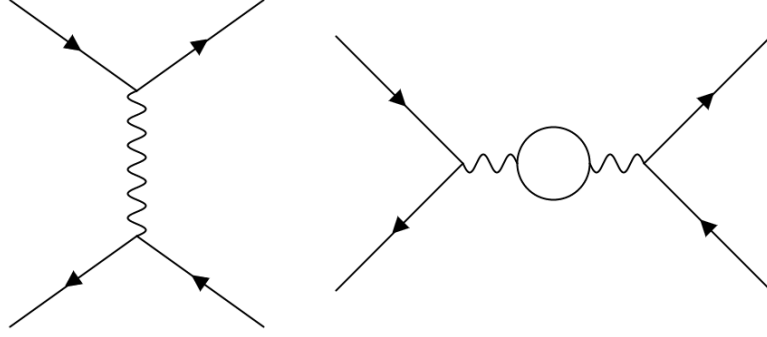


Figure 2.2.1: Example of Feynman diagrams showing fermion anti-fermion scattering at first order (left) and second order (right)

2.2 Particle Physics as a Quantum Field Theory

Under quantum field theory (QFT) particles are represented by excitations in relativistic quantum fields which span the universe. As such, a fermionic excitation can be represented by the creation operator $\bar{\psi}\psi$ and bosonic excitation can be represented by the creation operator $\phi^\dagger\phi$ or $A_\mu A^\mu$ depending on whether it is an excitation of a scalar or vector field. The interactions between fields is then represented by a Lagrangian density ($\mathcal{L}$). For example, a non-interacting fermionic field in the QFT formalism can be described from the Lagrangian representation of eq 2.1.1:

$$\mathcal{L} = \bar{\psi}_f(i\hbar c\gamma^\mu\partial_\mu - mc^2)\psi_f \quad (2.2.1)$$

Under this QFT formalism, the multiplication of field operators represents the interactions of their respective particles species and the corresponding constants accompanying their multiplication represents the coupling strength of the interaction.

From understanding the structure of these Lagrangian densities for the various SM interactions, we can deduce a large variety of predictions about the behavior of particles. For a given SM process, we can represent it pictorially through a Feynman diagram showing the potential ways in which particles would interact, be produced, and be destroyed to go from the process's initial to final state. For example, the Feynman diagrams in Figure 2.2.1 shows two pictorial representations in which a fermion and its corresponding anti-fermion could scatter. The terms of the SM Lagrangian density representing the propagation and interactions of each of the particles can then be used to construct a path integral [43] describing the transfer of energy from the initial to final state particles. This integral can be solved to obtain an amplitude. From this amplitude, information about the rate at which the interaction occurs can be directly obtained.

The one caveat of using Feynman diagrams is that if we want to describe the ways in which a physics process goes from an initial to a final state, we can always draw more diagrams in which additional internal interactions are taking place. By solving the corresponding path integrals, we would then find that there is a different amplitude associated with each diagram. As such, a single diagram only represents one term in an infinite perturbative expansion in the ways in which a SM process can unfold. Phenomenological calculations in QFT are then often described by an order indicating how many total internal interactions were considered when incorporating Feynman diagram amplitudes into a calculation. In general, the probability decreases as the number of internal interaction increases because the coupling strength of the interacting fields will be less than one. As such, a leading-order calculation is said to account for all diagrams involving the minimum number of interactions. For each additional pair of internal interactions that are accounted for in a set of calculations, an additional next-to- is appended to the description. As an example, the left diagram in Figure 2.2.1 is a leading-order interaction and the right diagram is a next-to-leading-order diagram. For next-to-leading-order diagrams and beyond, it is possible for associated amplitudes to be negative and cause destructive interference with other interaction methods, leading to some diagrams decreasing the overall probability of an interaction and others increasing the probability.

2.2.1 Quantum Chromo Dynamics

Quantum chromo dynamics (QCD) is the formal QFT theory behind the strong interaction. It is encapsulated by the $SU(3)_C$ portion of the SM gauge symmetry group, making it a theory involving a set of three distinct color charges: red (r), blue (b), and green (g). Similarly to how positive and negative electromagnetic charges can be considered the "anti-charge" of one another, there are additionally a $\bar{r}$, $\bar{b}$, and $\bar{g}$ anti-color carried by gluons and anti-quarks. As mentioned previously, the SM consists of 8 distinct gluons, these correspond to the different ways in which they are able to carry a color current. A single gluon carries a color and anti-color as described by the Gell-Mann matrices, which constitute the Lie algebra of $SU(3)$, giving us the following color combinations:

$$r\bar{b}, r\bar{g}, b\bar{g}, b\bar{r}, g\bar{r}, g\bar{b}, \frac{1}{\sqrt{2}}(r\bar{r} - b\bar{b}), \frac{1}{\sqrt{6}}(r\bar{r} + b\bar{b} - 2g\bar{g})$$

The QCD portion of the SM Lagrangian density thus constitutes the dynamics and interactions of the quark and gluon fields. Because the quarks are fermions, they are still described by the Dirac-Equation in eq 2.1.1. However, the derivative ∂_μ must now be replaced with the QCD covariant

derivative

$$(D_\mu^{\text{QCD}})_{ij} = \partial_\mu \delta_{ij} - ig(T_a)_{ij} A_\mu^a \quad (2.2.2)$$

to ensure that the quark interactions are invariant under an SU(3) gauge transformation. This can then be combined with the gluon field strength tensor

$$G_{\mu\nu}^a = \partial_\mu A_\nu^a - \partial_\nu A_\mu^a + gf^{abc} A_\mu^b A_\nu^c \quad (2.2.3)$$

to produce the full QCD Lagrangian density

$$\mathcal{L}_{\text{QCD}} = \sum_{\text{quarks}} (\bar{\psi}_i (i\gamma^\mu (D_\mu^{\text{QCD}})_{ij} - m\delta_{ij}) \psi_j) - \frac{1}{4} G_{\mu\nu}^a G_a^{\mu\nu} \quad (2.2.4)$$

In eq 2.2.2–2.2.4, we find that the ψ terms represent the various quark flavors with mass m and indices i, j running over the three color states. T_a represents the Gell-Mann matrices divided by a numeric factor of 2 and A_μ^a represents the gluon fields (analogous to the potential of an electromagnetic field). g is then the strong coupling constant corresponding to the coupling strength between gluons and color-charged particles and f^{abc} is the structure constant of SU(3) representing the 8 different ways in which the gluons can be multiplied together. A full expansion of eq 2.2.4 would lead us to find that the quark portion of the Lagrangian contains the standard Dirac equation as well as an additional term showing the coupling to the gluon field proportional to the strong coupling strength. The $G_{\mu\nu}^a$ products then create a term that is quadratic in A_μ^a representing the kinematic energy of gluons, as well as a cubic and quartic term in A_μ^a respectively proportional to g and g^2 representing the possible 3-point and 4-point self-couplings that are possible by gluons.

When looking directly at the strong-interaction coupling strength, the convention is often to divide the coupling constant which appears in $\mathcal{L}$ by the constant 4π to explore the features of the more fundamental constant α_s . Of particular interest about the strong coupling constant is that its strength depends strongly on the energy being transferred via the gluons [44] as approximately

$$\alpha_s(q^2) = \frac{\alpha_s(\mu^s)}{1 + \frac{7}{4\pi} \alpha_s(\mu^2) \ln(\frac{q^2}{\mu^2})} \quad (2.2.5)$$

where q is the energy being transferred via the gluon and $\alpha_s(\mu^2)$ refers to a measured strong coupling strength at pre-determined energy scale μ . Because this requires α_s at a known value, it is common to refer to the coupling with respect to the Z mass pole $\alpha_s(m_Z) = 0.1181 \pm 0.0011$ [4]. The predicted

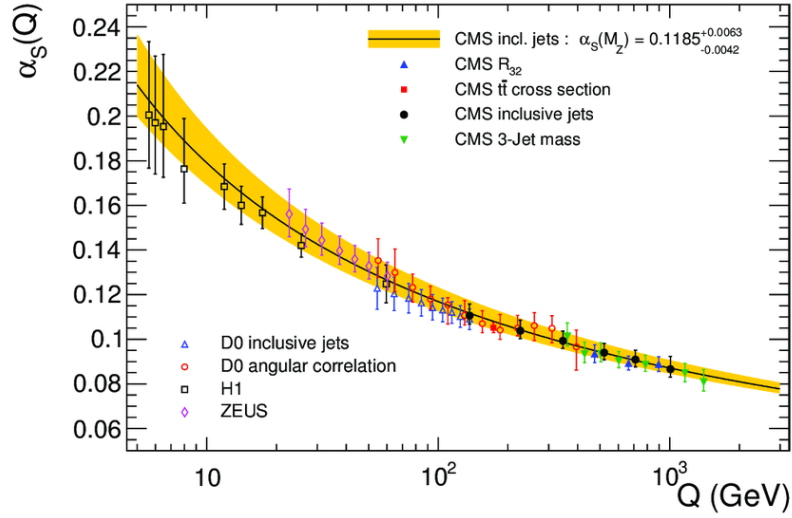


Figure 2.2.2: The estimated values of α_s at various energies based on eq 2.2.5 with a reference scale of $\alpha_s(m_Z)$ compared to data from multiple particle physics experiments. Figure from [2].

and measured values of α_s at the GeV energy scale can be seen in Figure 2.2.2. Of particular note about the energy dependence of α_s is that in comparison to the electromagnetic coupling strength $\alpha_e \approx 1/137$ we find that $\alpha_s(m_Z)$ is more than an order of magnitude larger. Furthermore, we find that at very low energies, the term $\ln(q^2/\mu^2)$ is negative and causes the coupling strength to grow rapidly as the energy scale approaches 0. Additionally, increasing the energy will cause the value of α_s to asymptotically reach 0. This is the underlying process defining quark confinement, in which lone quarks are not found in nature. A lone quark will have an α_s coupling approaching infinity, causing it to produce new strongly interacting particles out of the vacuum with which it can interact. Alternatively, the energy required to remove a quark from a strong bound state is larger than the energy scale required to pair-produce new quarks out of the vacuum. It is also for this reason that the mass of protons and neutrons are much larger than the mass of the up and down quarks as listed in Figure 2.1.1. Unlike electromagnetic bound states, strong bound states exhibit a positive binding energy because the bound states are still at lower energy state than the near infinite energy required to have a free quark.

As briefly mentioned in the start of this chapter, QFT calculations are commonly performed as perturbative expansions about the number of internal interactions in a process. This means that the coefficient about which the expansion is being performed is the coupling constant of the interactions taking place. This leads to low energy QCD expansions to converge slowly, making phenomenological predictions for low energy strong interactions exceedingly challenging. Alternatively, α_s becomes

a suitable expansion coefficient at higher energy scales allowing for high energy QCD interactions to be predicted via perturbation theory. We find that most of the new physics processes being explored by the LHC exist at high energies, allowing the use of perturbation theory when modeling these theories. However, the large α_s present at low energy scales makes it not currently feasible to accurately model low energy proton scatterings, which is by far the most dominant process during proton collision.

2.2.2 Electroweak Interactions

The electroweak (EWK) interaction is the formal QFT theory describing the combined electromagnetic and the weak interactions. At high energy scales, it is found that the electromagnetic and weak interactions of the standard model are unified. As such, the EWK interaction is described by the $SU(2)_L \times U(1)_Y$ portion of the SM gauge symmetry group. This also means that the EWK interaction interacts via two unique charges: Weak isospin associated with $SU(2)_L$ and hypercharge associated with $U(1)_Y$. It is found that from the EWK force, we obtain 4 massless vector gauge bosons each associated with a quantum field $W_\mu^{(1)}$, $W_\mu^{(2)}$, and $W_\mu^{(3)}$ to carry interactions via isospin and B_μ to carry hypercharge. It is from these two electroweak charges that the electric charge (Q) associated with the electromagnetic interaction is obtained. From the respective EWK interactions, Q can be calculated as

$$Q = I_3 + \frac{1}{2}Y \quad (2.2.6)$$

where I_3 is a projection of the isospin upon an abstract choice of axis, analogous to the s_z spin projection used in the $SU(2)$ description of spin 1/2 particles, and Y is the hypercharge. The values of I_3 and Y for the various flavors of SM fermions can be found in Table 2.1. Of note is how like-particles of the same family but different generations share the same isospin and hypercharge values.

Similar to QCD, the fermions originally described by the dirac equation must have ∂_μ formulated into a new covariant derivative to ensure that the particles are invariant under a $SU(2)_L \times U(1)_Y$ gauge transformation. This leads to the EWK covariant derivative to be defined as

$$D_\mu^{\text{EWK}} = \partial_\mu - i\frac{g'}{2}YB_\mu - i\frac{g}{2}I_jW_\mu^j \quad (2.2.7)$$

where g' and g are the corresponding coupling strengths associated with the interactions via hyper-

family	flavor	Left-Handed		Right-Handed	
		I_3	Y	I_3	Y
leptons	e^-, μ^-, τ^-	$-1/2$	-1	0	-2
	ν_e, ν_μ, ν_τ	$1/2$	-1		—
quarks	u, c, t	$1/2$	$1/3$	0	$4/3$
	d, s, b	$-1/2$	$1/3$	0	$-2/3$

Table 2.1: The eigenvalues of weak isospin (I_3) and hypercharge (Y) for the different fermion flavors of the standard model. The left- and right-handed chiral projections of the fermionic spinors transform separately under $SU(2)_L \times U(1)_Y$ giving rise to two sets of values. No values are listed for right-handed neutrinos since no experimental confirms their existence.

charge (Y) and weak isospin (I_j), which take the form of the identity (Y) and pauli spin matrices (I_j). The corresponding $W^{(i)}$, and B field tensors can be described by:

$$W_a^{\mu\nu} = \partial^\mu W_a^\nu - \partial^\nu W_a^\mu - ig\epsilon^{abc}W_b^\mu W_c^\nu \quad (2.2.8)$$

$$B^{\mu\nu} = \partial^\mu B^\nu - \partial^\nu B^\mu \quad (2.2.9)$$

The interactions of the fermions, $W^{(i)}$, and B fields, can then be used to build the EWK Lagrangian density

$$\mathcal{L}_{\text{EWK}} = \sum_{\text{fermions}} (\bar{\psi}(i\gamma^\mu D_\mu^{\text{EWK}} - m)\psi) - \frac{1}{4}W_a^{\mu\nu}W_{\mu\nu}^a - \frac{1}{4}B^{\mu\nu}B_{\mu\nu} \quad (2.2.10)$$

where the fermion portion now describes their kinematics, mass, and interactions with the massless EWK vector gauge bosons, the $W_\mu^{(i)}$ portion describes the quadratic kinematics and the cubic and quartic interactions of its respective fields, and the B_μ portion only has a quadratic kinematic term.

2.2.2.1 Spontaneous Symmetry Breaking and the Higgs Boson

The EWK Lagrangian density shown in eq 2.2.10 exhibits multiple contradictions with respect to the observed SM. The first is that the gauges fields described by the Lagrangian are all massless fields. However, experiments have shown us that the $W^\pm$ and Z boson of the weak interaction should all have masses in the GeV scale. The second is that the fermionic portion of the Lagrangian is gauge invariant under an $SU(2) \times U(1)_Y$ gauge transformation, but not $SU(2)_L \times U(1)_Y$. That is because the $SU(2)_L$ gauge symmetry only acts on the left-handed chiral state of the fermions (as shown in Table 2.1). In other words, the left- and right-handed decomposition of a fermion can be written as

the bispinor:

$$\psi = \begin{pmatrix} \psi_R \\ \psi_L \end{pmatrix} \quad (2.2.11)$$

This means that under our EWK gauge symmetry, the left- and right-handed components will transform separately as

$$\psi'_L = e^{i\alpha(x)I + i\beta(x)Y} \psi_L \quad (2.2.12)$$

$$\psi'_R = e^{i\beta(x)Y} \psi_R \quad (2.2.13)$$

which will lead to a broken gauge symmetry in eq 2.2.10 due to the two decomposed mass terms in

$$-\bar{\psi}m\psi = -(\bar{\psi}_R m \psi_L + \bar{\psi}_L m \psi_R) \quad (2.2.14)$$

transforming differently after the gauge transformation (because a right-handed anti-fermion is a left-handed chiral state).

It ends up that introducing a complex scalar field with $I_3 = -1/2$, $Y = 1$, and a non-zero vacuum excitation value (VEV) is able to spontaneously break the $SU(2)_L \times U(1)_Y$ in a way that naturally gives rise to the $W^\pm$, Z , and γ force carriers observed at our energy levels, and also introduce fermion masses in a way that does not break the $SU(2)_L \times U(1)_Y$ gauge symmetry. In the SM, this spontaneous symmetry breaking can be carried out through the existence a self-interacting complex scalar field with the form

$$\phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} \phi_1 - i\phi_2 \\ \phi_3 - i\phi_4 \end{pmatrix} \quad (2.2.15)$$

as a doublet under $SU(2)_L \times U(1)_Y$. This is introduced as the Higgs field with a Lagrangian density of

$$\mathcal{L}_H = (D^\mu \phi)(D_\mu \phi)^\dagger + \mu^2(\phi^\dagger \phi) - \lambda(\phi^\dagger \phi)^2 \quad (2.2.16)$$

where D_μ is the EWK covariant derivative introduced in eq 2.2.7, μ is an effective mass term for the field with the opposite sign as a standard mass term, and λ is the quartic self-coupling of the scalar

field. Via the standard Lagrangian formalism, we find that the potential energy of this field is

$$V(\phi) = -\mu^2(\phi^\dagger\phi) + \lambda(\phi^\dagger\phi)^2 \quad (2.2.17)$$

which can have two potential minima based on the value of μ^2 . If $\mu^2 < 0$ we find that the potential energy is minimized at $\phi = 0$. However, if $\mu^2 > 0$ we find it results in a non-zero VEV which can be formulated as:

$$\phi_\circ = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v \end{pmatrix}, \text{ with } v = \sqrt{\frac{\mu^2}{\lambda}} \quad (2.2.18)$$

It is believed that cooling in the early universe caused the average temperature to drop below v , causing the Higgs field to undergo a Landau-style phase transition leading to the non-zero VEV. It can be seen that at this new minimum, $\phi_\circ$ is no longer an SU(2) invariant doublet, and now only remains invariant via the U(1) portion of the electroweak gauge symmetry. This spontaneous electroweak symmetry breaking as a result of a cooling early universe is what causes the breaking of $SU(2)_L \times U(1)_Y$ into the separate weak and electromagnetic interactions at low energy.

Through the Higgs field, we can now consider its full interactions with the standard model by considering quantum fluctuations about the VEV. This fluctuation constitutes excitations in the field, giving rise to the interactions between the SM particles and the Higgs and it's non-zero VEV. We can do this by representing the post-symmetry breaking field as:

$$H = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + h \end{pmatrix} \quad (2.2.19)$$

To better understand the properties of the real Higgs, we can plug H into eq 2.2.17 to obtain

$$V(h) = \frac{\mu^4}{4\lambda} + \mu^2 h^2 + \mu\sqrt{\lambda}h^3 + \frac{1}{4}\lambda h^4 \quad (2.2.20)$$

giving rise to a constant energy offset associated with the non-zero VEV, the physical Higgs mass being equal to $m_H = \sqrt{2}\mu$, as well as the cubic and quartic self-coupling terms associated with λ which continues to exist as an unconstrained parameter in the Lagrangian.

From the kinetic term in eq 2.2.16 we can explore the coupling of the physical Higgs boson and the VEV to the electroweak fields. By specifically looking at their interaction of the covariant

derivative with the Higgs VEV, we find:

$$D_\mu \phi = -\frac{iv}{\sqrt{8}} \begin{pmatrix} g(W^{(1)} - iW^{(2)}) \\ -gW^{(3)} + g'Y_{\phi_0}B_\mu \end{pmatrix} \quad (2.2.21)$$

Considering the full kinematic term of the Higgs VEV we then find

$$(D^\mu v)^\dagger (D_\mu v) = \frac{v^2}{8} \left(g^2(W^{(1)2} + W^{(2)2}) - (gW^{(3)} + g'B)^2 \right) \quad (2.2.22)$$

producing an effective mass for the quadratic terms of our electroweak fields in the SM Lagrangian dependent on the Higgs VEV. Due to the influence of the Higgs VEV breaking the $SU(2)_L \times U(1)_Y$ symmetry we can replace our electroweak gauge fields with the real gauge bosons associated with the weak and electromagnetic interactions. For the charged currents, the $W^{(1)}$ and $W^{(2)}$ fields mix as

$$W_\mu^\pm = \frac{1}{\sqrt{2}} (W_\mu^{(1)} \mp iW_\mu^{(2)}) \quad (2.2.23)$$

to produce the physical $W^\pm$ bosons. Similarly, the $W^{(3)}$ and B fields mix as

$$\begin{pmatrix} Z_\mu & A_\mu \end{pmatrix} = \begin{pmatrix} \cos \theta_W & \sin \theta_W \\ -\sin \theta_W & \cos \theta_W \end{pmatrix} \begin{pmatrix} W_\mu^{(3)} \\ B_\mu \end{pmatrix} \quad (2.2.24)$$

to produce the associated neutral current carrying fields Z and A . Here, the term θ_W is the Weinberg angle representing the mixing between the interaction couplings of the $W^{(i)}$ and B_μ gauge fields as $\tan \theta_W = g'/g$. By plugging eq 2.2.23 and 2.2.24 into eq 2.2.22 we arrive at

$$\mathcal{L}_{VV} = \frac{g^2 v^2}{8} (W^{+\mu} W_\mu^- + W^{-\mu} W_\mu^+) + \frac{(g^2 + g'^2) v^2}{8} Z^\mu Z_\mu \quad (2.2.25)$$

producing the Lagrangian mass terms for the $W^\pm$ and Z weak bosons while leaving no remaining term for the A_μ gauge field associated with the massless photon. From this, we can also find the relationship between the coupling strength of the Higgs VEV and EWK fields and the masses of the vector gauge bosons given in Table 2.2 in which the masses are dependent on various terms appearing in the SM Lagrangian. Due to the dependence of the Higgs VEV of the masses of the W and Z bosons and their coupling strength of the EWK interaction, the value of $v \approx 246$ GeV had been known for years before the eventually discovery of the Higgs boson at the LHC by the ATLAS

mass	constants	experimental value
m_H	$\sqrt{2\lambda}v$	$125.25 \pm 0.17 \text{ GeV}$
m_W	$\frac{1}{2}gv$	$80.379 \pm 0.021 \text{ GeV}$
m_Z	$\frac{1}{2}\sqrt{g^2 + g'^2}v$	$91.1876 \pm 0.0021 \text{ GeV}$

Table 2.2: Masses of the SM massive gauge bosons as a function of free parameters in the SM, as well as the experimentally measured values. Numbers quoted are averages of many experimental measurements aggregated by [4].

and CMS collaborations [21, 22]. However, we still have no measurements of the Higgs self-coupling leading to no direct measurements of λ .

2.2.2.2 The Higgs-Yukawa Couplings to Fermionic Fields

As shown in eq 2.2.14 the existence of fermion masses in the SM Lagrangian breaks its invariance under the $SU(2)_L \times SU(1)_Y$ gauge symmetry. However, this can be remedied through reinterpretation of the mass through a Yukawa-coupling to the Higgs field shown in eq 2.2.19. By dropping higher order terms (in particular those associated with flavor mixing) we can obtain the following Yukawa term associated with the Lagrangian

$$\mathcal{L}_y = \sum^{\text{gens}} (-\lambda_\ell \bar{L}_L H \ell_R - \lambda_d \bar{Q}_L H d_R - \lambda_u \bar{Q}_L H u_R) \quad (2.2.26)$$

where the sum is over all three generations, and L_L and Q_L represents left-handed particles in the lepton and quark doublets, and ℓ_R , d_R , and u_R represents the charged right-handed lepton, down-like quarks, and up-like quarks. λ_f is the free parameter associated with each fermion's coupling to the Higgs field. By expanding H and repackaging like terms, we can represent the Yukawa terms of the SM Lagrangian as:

$$\mathcal{L}_y = - \sum_f \left(\frac{\lambda_f v}{\sqrt{2}} \bar{\psi}_f \psi_f + \frac{\lambda_f}{\sqrt{2}} \bar{\psi}_f \psi_f h \right) \quad (2.2.27)$$

From this we find that the mass of each fermion is directly proportional to the Higgs VEV multiplied by its individual coupling to the Higgs field. This gives physicists two distinct ways to measure the Higgs-Yukawa coupling with the SM fermions. The first is through the direct measurement of the fermion masses, which is something that has been established for many decades. The second way is by directly measuring each fermion's coupling to a Higgs boson particle produced at a particle accelerator. This second method is just now a possibility due to the LHC, and has already been performed for multiple particles [31, 32, 33, 34, 35, 36, 37, 38].

2.3 High Energy Proton Proton Scattering

Hadron colliders like the LHC produce new particles through the collisions of strongly-interacting bound states, with the most common hadron candidate being protons due to their stable nature. Because protons are composite particles, they exhibit a very non-trivial scattering scheme. During low momentum collisions, incident particles will scatter off a proton as if it is a single non-point particle. But as the momentum of the incident particle increases, the composite nature of the proton emerges. When scattering at high energies, incident particles will scatter with the individual partons which make up the proton. Because these partons are co-interacting when inside of the proton, the fraction of proton momentum contained in a single parton will not be a set value, and instead must be described probabilistically through a parton distribution function (PDF). These parton-parton interactions further complicate matters, as the gluons exchanged between partons will often pair-produce virtual quark-antiquark pairs. As momentum of the incident particle continues to increase, scatters directly off the virtual particles produced by gluons. At scattering energies above $2m_c$ and $2m_b$ it additionally becomes possible for incident particles to scatter off c and b quarks produced from gluon pair-production, leading to the PDF having a large dependence on scattering energies. The distributions of the proton PDF at a scattering energies of $\sqrt{10}$ GeV and 100 GeV can be found in Figure 2.3.1. Here the x-axis represents the fraction of momentum carried by a species of parton, and the y-axis represents the probability to find said particle carrying the corresponding momentum fraction multiplied by the momentum fraction. Because α_s is suppressed at high energies, the probability that virtual quarks and anti-quarks contain a large fraction of the proton momentum is low. As such, the majority of the proton momentum is most likely carried by the up and down valence quarks and the gluons.

At the LHC, the high energy scale of the collisions means that protons collide via their PDFs. This means that for a given proton-proton interaction, the real collision that takes place is between a set of partons in each of the protons. The collision energy is thus determined by the fraction of the protons energies' contained in the two colliding partons, which will always be lower than the quoted $\sqrt{s}$ center of mass energy of the colliding proton beams. The general consequence of this is two fold: The vast majority of proton-proton collisions will be between partons holding a low fraction of the proton momentum, causing the vast majority of collisions to be "uninteresting" for the purposes of a collider whose purpose is to push new energy scales. However, this leads to second consequence which is that the LHC to be able to scan a large range of scattering processes. Due to the properties of the proton PDF, the LHC is able to produce nominal collisions between g, u, d, s, c, and b (and

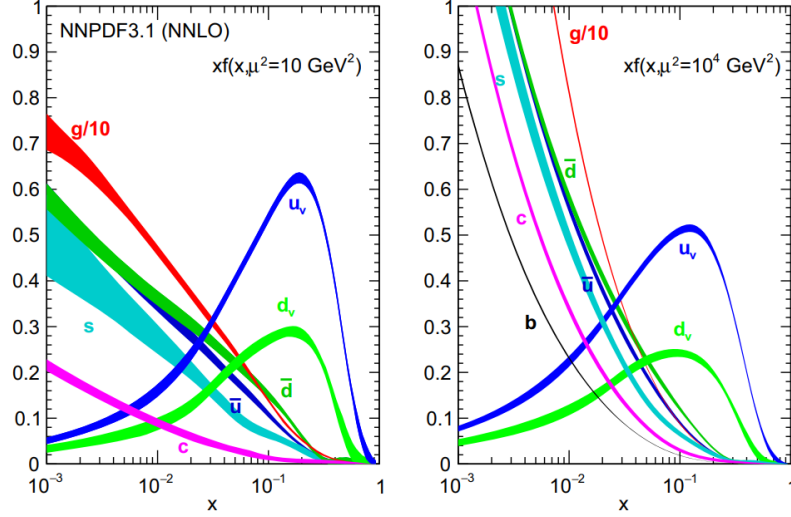


Figure 2.3.1: Proton parton distribution function at $\sqrt{s}$ energies of $\sqrt{10}$ GeV (left) and 100 GeV (right). The x-axis represents the fraction of proton momentum a colliding parton of said flavor contains, and the y-axis represents the probability of the parton having said momentum fraction multiplied by the momentum fraction itself. Distributions are made such that the integral of a single distribution represents the the total fraction of the proton momentum that will be carried by the corresponding parton. Figure obtained from [3].

the antiquark partners) from low energies all the way up to energies in the TeV scale.

2.3.1 Higgs Production at Proton-Proton Colliders

There are multiple ways in which a Higgs boson can be produced as a result of proton-proton collisions. But as described in Section 2.2.2.1 and 2.2.2.2 the coupling strength of a Higgs boson to massive particles is directly proportional to the mass of the particle. This leads to Higgs boson production being greatly suppressed when it is being radiated from or produced from the fusion of the lighter fermions. As such, the dominant production modes of the Higgs boson involves processes in which the Higgs boson is interacting with the weak gauge bosons or t quarks.

The dominant Higgs boson production modes of proton-proton collisions are thus gluon-gluon fusion, vector boson fusion, associated production with a gauge boson, and associated production with top quarks. The leading order feynman diagram for each of the production modes can be found in Figure 2.3.2 and their descriptions are as follows:

gluon-gluon fusion (ggF) – The vastly dominant production mode by an order of magnitude.

Two scattering gluons produce a virtual pair of top and anti-top quarks which fuse into a Higgs boson. Cross section for a $m_H = 125$ GeV at $\sqrt{s} = 13$ TeV proton-proton collisions is $48.6^{+5\%}_{-5\%}$ pb.

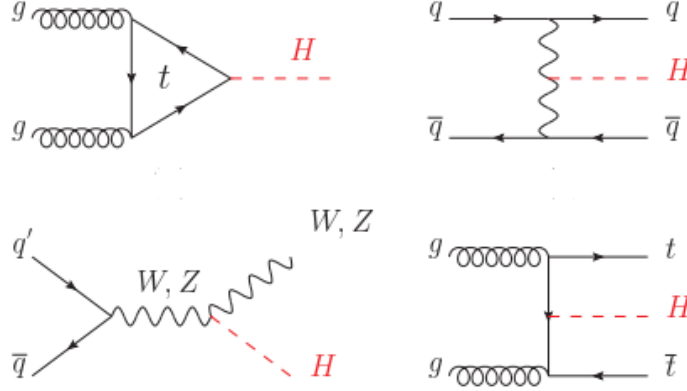


Figure 2.3.2: The dominant production modes of the Higgs boson for proton-proton collisions. The production modes are gluon-gluon fusion (top left), vector boson fusion (top right), associated production with gauge bosons (bottom left), and associated production with top quarks (bottom right). Figure obtained from [4].

Vector Boson Fusion (VBF) – Two quarks scatter via the exchange of a Z or W boson. The gauge boson will in turn radiate a Higgs boson while being exchanged between the quarks. This process generally leads to the two quarks having low scattering angles with respect to the angle of the incident colliding protons, and the Higgs primarily being produced orthogonally to the axis of the incident protons. Cross section for a $m_H = 125 \text{ GeV}$ at $\sqrt{s} = 13 \text{ TeV}$ proton-proton collisions is $3.78^{+5\%}_{-5\%} \text{ pb}$.

Associated production with a gauge boson (WH/ZH) – A quark and anti-quark undergo s -channel scattering and fuse into a virtual W or Z boson. The gauge boson then splits into a real W/Z and a Higgs boson. Cross section for a $m_H = 125 \text{ GeV}$ at $\sqrt{s} = 13 \text{ TeV}$ proton-proton collisions is $1.37^{+5\%}_{-5\%} \text{ pb}$ for WH and $0.88^{+5\%}_{-5\%} \text{ pb}$ for ZH.

Associated production with top quarks (ttH) – Two gluons scatter to produce a $t\bar{t}$ pair. A Higgs boson is then radiated from one of the t quarks. Cross section for a $m_H = 125 \text{ GeV}$ at $\sqrt{s} = 13 \text{ TeV}$ proton-proton collisions is $0.50^{+5\%}_{-5\%} \text{ pb}$.

The cross sections listed for each of the production modes was obtained from [4]. A comparison of the cross sections for a range of TeV collision energies for the various Higgs production modes can be found in Figure 2.3.3.

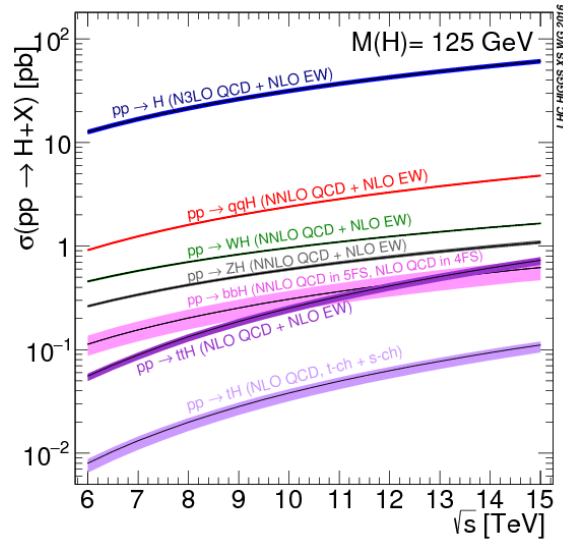


Figure 2.3.3: Cross sections of the various Higgs boson production modes for proton-proton collision with $\sqrt{s}$ in the TeV energy range. Figure obtained from [4].

Chapter 3

The Compact Muon Solenoid Experiment and the Large Hadron Collider

The $VH(H \rightarrow c\bar{c})$ analysis was performed on 138 fb^{-1} of proton-proton collision data collected using the Compact Muon Solenoid (CMS) detector from 2016-2018 during the "Run 2" data taking period. The collisions took place at $\sqrt{s} = 13 \text{ TeV}$ using high energy proton beams produced at the Large Hadron Collider (LHC). An overview of the amount of data delivered by the LHC and recorded by the CMS detector over the course of the Run 2 data taking period can be found in Figure 3.0.1. In this chapter, I will give a formal introduction to the of LHC and CMS detector given the state they were in when producing the data used in the $VH(H \rightarrow c\bar{c})$ analysis. As such, I will not be discussing all components of the LHC and CMS detector in their current states, which have since undergone upgrades during the long shutdown 2 period.

In discussing the LHC, I will start with a discussion of the purpose of the detector followed by an overview of how the accelerator works. I will describe how protons go from a stable proton gas to being in bunches of $1.15 \cdot 10^{11}$ protons at 450 GeV that are injected into the LHC from the pre-accelerator complex. I will then describe how the LHC is able to accelerate the protons to an energy of 6.5 TeV and maintain a stable proton beam at these energies for up to 25 hours at a time. After my discussion of the LHC, I will give a brief overview of the CMS detector and its various subsystems. I will then describe the trigger system used by the detector to rapidly filter out events

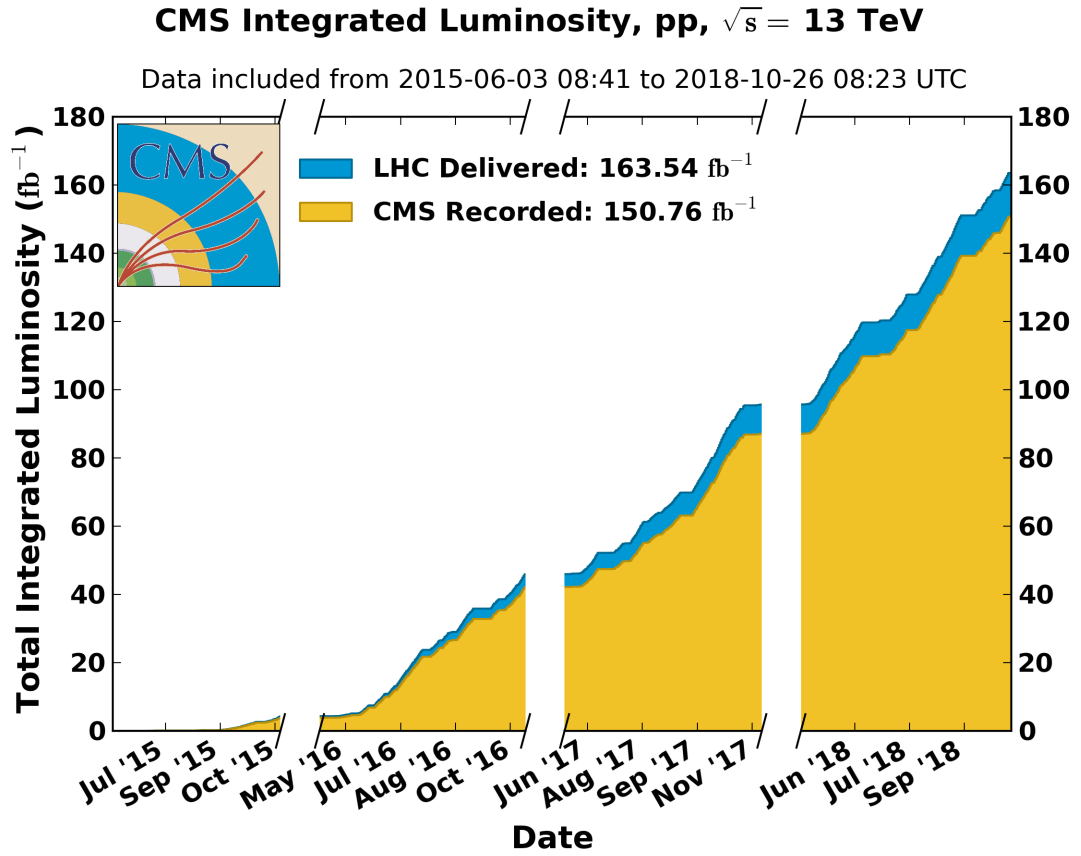


Figure 3.0.1: Total luminosity delivered by the LHC and collected by the CMS detector collected during the 2015–2018 run 2 period. The 3.86 fb^{-1} of data collected during 2015 was not used for the $VH(H \rightarrow c\bar{c})$ search.

and select which collisions to record. Finally, I will go over the types of high level physics objects used to carry out the analysis, as well as how these physics objects are produced using the low level detector readout information.

3.1 The Large Hadron Collider

The Large Hadron Collider (LHC) is a 26.7 km storage ring collider operated by the Conseil Européen pour la Recherche Nucléaire (CERN) and located approximately 100 meters beneath Geneva, Switzerland and the surrounding French cities. The LHC, along with a series of previous CERN accelerators, are designed to accelerate two counter rotating beams of protons and/or lead ions to a maximum beam energy of 7 TeV for protons and 2.76 TeV/nucleon for lead ions [7]. However, full energy has not yet been reached for proton beams leading to beam energies of $E_{\text{beam}} = 3.5$ TeV in 2011, $E_{\text{beam}} = 4$ TeV in 2012, $E_{\text{beam}} = 6.5$ TeV in 2015–2018, and planned energies of $E_{\text{beam}} = 6.8$ TeV in the upcoming Run 3. At 6.5 TeV of energy, protons fly around the LHC at approximately 99.9999991% the speed of light leading to a circular frequency of 11.2 kHz, and collide at one of four interaction points each located inside one of the primary detectors: ATLAS [45], CMS [46], ALICE [47], and LHCb [48].

3.1.1 Beam Parameters

To obtain the collision rates required to observe rare processes like Higgs boson production and potential BSM particles, the LHC was designed to collide proton beams at a peak luminosity of $10^{34}/\text{cm}^2/\text{s}$. However, the probability of proton-proton scattering falls off as $\sqrt{s}$ increases. At the $\sqrt{s} = 13$ TeV energies of the LHC, this means that the probability of two individual protons colliding head on is less than 1 in 10^{10} . To reach the desired luminosity, the LHC is designed to pack protons together into tightly packed bunches and collide two bunches every 25 ns. Assuming that in each bunch protons are distributed about a central point following a Gaussian distribution, the luminosity of head on collisions can be expressed as

$$\mathcal{L} = \frac{N_1 N_2 f n_b}{4\pi\sigma_x\sigma_y} \quad (3.1.1)$$

where N_1 and N_2 are the number of protons in each bunch, f is the revolution frequency, n_b are the number of bunches, and σ_x and σ_y are the standard deviation of the proton density in the x and y direction. Given the design specified beam parameters shown in Table 3.1 combined with the fact

that protons moving at speed c would revolve around the LHC with a frequency of ~ 11 kHz, we find that eq 3.1.1 gives a luminosity of $1.2 \cdot 10^{34}/\text{cm}^2/\text{s}$.

Parameter	Design	Peak Run 2
Energy	7 TeV	6.5 TeV
Peak Luminosity	$10^{34}/\text{cm}^2/\text{s}$	$2 \cdot 10^{34}/\text{cm}^2/\text{s}$
Number of Bunches	2808	2556
Bunches per Train	288	144
Bunch Separation	3.7 m	7.5 m
Protons Per Bunch	$1.15 \cdot 10^{11}$	$1.25 \cdot 10^{11}$
Bunch Length RMS	7.55 cm	7.55 cm
Bunch Width RMS	16.7 μm	9.5 μm
Crossing Angle at CMS	285 μrad	260 μrad

Table 3.1: Designed LHC Beam parameters according to [7] and peak values reached during the run 2 data taking period.

There is one additional complication when colliding the proton bunches in that it is unfeasible to have head-on collisions. To have head-on collisions would require the two circulating beams to overlap, meaning that they would be colliding all throughout the LHC ring. Instead what happens is that in each of the interaction points, the two beams cross at an angle such that only a single pair of oppositely circulating bunches collide at a time. In this crossing scheme, we effectively reduce the overlap of the bunches during collision and must introduce a geometric reduction factor S to the right hand side of eq 3.1.1. The geometric reduction factor can be approximated to

$$S \approx \frac{1}{\sqrt{1 + \left(\frac{\sigma_s}{\sigma_x} \frac{\phi}{2}\right)^2}} \quad (3.1.2)$$

where σ_s is the standard deviation of the bunch length, σ_x is still the standard deviation of the bunch radius, and ϕ is the crossing angle at which the two bunches intersect. Giving the design parameters specified in Table 3.1 we find a geometric reduction factor of $S = 0.84$. When multiplied by the head-on luminosity of $\mathcal{L} = 1.2 \cdot 10^{34}/\text{cm}^2/\text{s}$ obtained from eq 3.1.1, we then arrive at the design goal luminosity of $\mathcal{L} = 10^{34}/\text{cm}^2/\text{s}$.

Procuring a beam of this specification and continually circulating and colliding bunches is nothing short of one of the greatest feats of modern engineering. It first requires a complex pre-acceleration system required to deliver stable bunches with proton energies of $E_{\text{beam}} = 450$ GeV to the LHC ring. The LHC must then increase the beam energy to the desired $E_{\text{beam}} = 6.5$ TeV energies used during Run 2 and continuously circulate the colliding bunches of protons for hours at a time. During this circulation, extra care must also be taken to ensure that the quality of the beam does not degrade beyond what is unavoidable during collisions.

3.1.2 The Pre-Accelerator Complex

The LHC is the last in a series of many particle accelerators used to accelerate protons to their final energies. Protons start as a H_2 gas which has the electrons stripped away. It then goes through a chain of particle accelerators consisting of the Linear Accelerator 2 complex (Linac2), Proton Synchrotron booster (PS booster), Proton Synchrotron (PS), and Super Proton Synchrotron (SPS) accelerators. For ring based accelerators, the maximum number of particles that can circulate scales according to $\beta^2\gamma^3$ where both β and γ are the standard relativistic factors for the beams upon entering the ring. Thus, by using a pre-accelerator system before injection into the LHC, the maximum luminosity of the LHC beam can be vastly increased. Additionally, it allows each ring to be tuned to only bend and accelerate protons of specific energy ranges. This aids in reducing the complexity of the earlier stages of acceleration. Especially in terms of the caliber of magnets used to turn the beams, where super conducting magnets are not needed in any of the accelerators in the pre-acceleration complex. The schematics of the full CERN complex showing the LHC, proton pre-accelerator chain, and the four primary detectors can be seen in Figure 3.1.1.

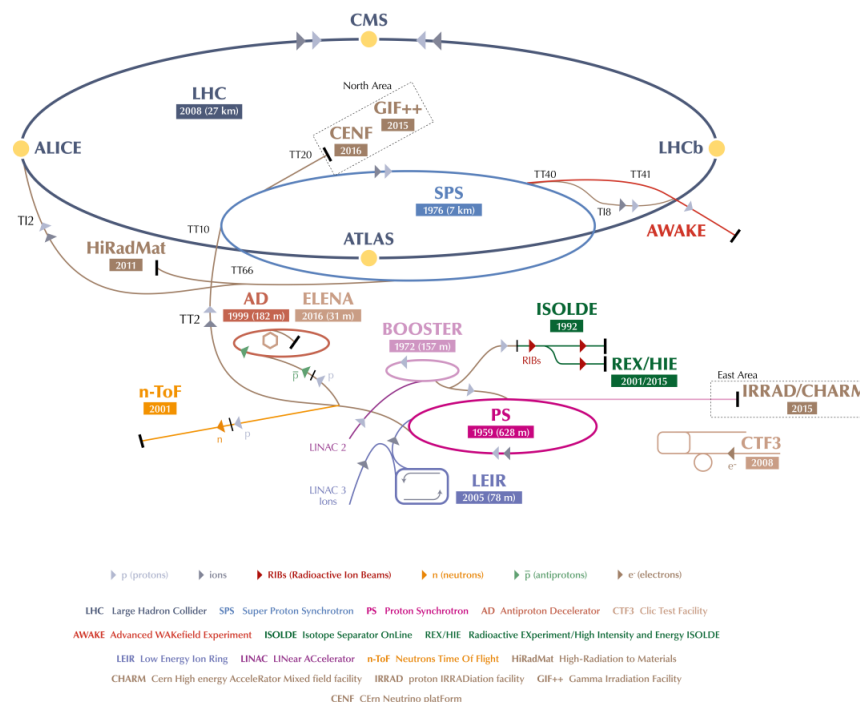


Figure 3.1.1: Full schematic of the CERN accelerator complex. In addition to the LHC, position of the four primary detectors, the pre-accelerator chain, and auxiliary experiments. Figure taken from [5].

3.1.2.1 Linac2

The protons which eventually move into the LHC initially start as stable H_2 gas sitting in a canister that will eventually feed into the Linac2 linear accelerator complex. The hydrogen gas is first pulsed out of its canister and into a strong electric field which will strip away the electrons. The protons then move through an approximately one meter long radio frequency quadrupole (RFQ) [49] to squeeze the protons into a beam. The RFQ consists of a long cylinder with four rods along the inside. The rods continue down the full length of the cylinder and are all spaced 90° apart along the end of the cylinder. A modulating RF polarity is then run through each rod such that they are all in phase with the rod across from them and 180° out of phase with the rods next to them. When the protons pass through the RFQ, they are forced to oscillate between being squeezed horizontally and longitudinally. When the protons leave the RFQ, they are squished together in a beam that is undergoing quadrupole oscillations, which aids in keeping the beam together despite the proton's electric repulsion from each other.

After having passed through the RFQ, the protons move into Linac2. Linac2 contains the first of many Radio Frequency (RF) cavities which will accelerate the protons. RF cavities consist of large cylindrical conductors which will switch between positive and negative polarity 180° out of phase with the conductor in front and behind it. As the protons pass through the RF cavity, they oscillate such that the positively charged conductors are pushing the protons forward with the help of the negatively charged conductors pulling them. At the end of Linac2, the protons have been accelerated to an energy of around 50 MeV. In total, it takes approximately 1.2 seconds for the protons to go from stable H_2 gas to the end of Linac2

3.1.2.2 The Proton Synchrotron Booster

The Linac2 then sends the proton to the PS booster, which is a set of four super imposed synchrotron rings each with a circumference of 157 m. Because there are four separate rings, the protons coming from Linac2 are fed into a distributor which pass the protons into one of the four rings. Linac2 continues to feed protons into the PS booster rings via the pulsed hydrogen gas until each ring holds approximately 10^{13} protons. When in the PS booster, the protons will pass through a sequence of RF cavities to accelerate the beam, and magnets to circulate them. As the protons travel through the PS booster, the RF frequency increases from 0.6 to 1.7 MHz along with the magnetic fields continuing to ramp up until the protons reach an energy of 1.4 GeV. Like Linac2, it takes 1.2 seconds for the PS booster to go from being empty to having all four rings full and at maximum energy.

Once ready, the booster ejects all four rings such that they form four distinct bunches, each spaced 327 ns apart. By design, the sum of the circumferences of the four PS booster rings is exactly equal to the 628 m PS circumference. This was originally to ensure that proton bunch injected by the PS booster could occupy a quarter of the PS ring. However, for use with the LHC the PS and PS booster were upgraded to use a two batch filling scheme. As the proton bunches leave the PS booster, a new set of protons is added to fill a set of two more booster rings. The second batch of bunches are then accelerated and injected into the PS exactly 1.2 seconds after injection of the first four batches. This lets the PS go from being empty to having six 1.4 GeV proton bunches circulating inside the ring in a time-frame of approximately 2.4 seconds.

3.1.2.3 The Proton Synchrotron

Once filled, the PS accelerates the protons to an energy of 26 GeV. To obtain these energies, the RF cavities increase to a frequency of 80 MHz. In reaching this RF frequency, the oscillation wave lengths become finer than the proton bunch length of the input beam. This allows the RF cavities to split the proton beam into groups such that they fit inside of the RF waveform. It is at this time that the beam will be split into the proton bunches of $\sim 1.1 \times 10^{11}$ separated in time by 25 ns.

The total number of bunches the PS splits the injected beam into (referred to as a batch) varied throughout the 2016–2018 data taking years. The pre-accelerator design specs are for the PS to first split the initial PS booster bunches into thirds. The bunches would then go through two additional sets of splitting where each bunch is split into two pieces, leading to a total of 72 bunches circulating inside the PS. However, this splitting scheme was not achieved until partway through 2017. In 2016, the PS booster bunches were split in half a total of three times, leading to only 48 bunches circulating in the PS. In all three years, the batch is produced with a 300 ns gap between the final and first bunch, to produce a large enough gap for the deflector (kicker) magnets to ramp up. After kicker magnet ramp up, the batch is injected into the SPS. Once the first fill begins, it takes 3.6 s for the PS to go from being empty to having finished production of a train.

3.1.2.4 The Super Proton Synchrotron

The SPS is a 6.9 km synchrotron accelerator, and its relationship with the PS is much like the relationship between the PS and the PS booster. In addition to the 300 ns gap in a single batch, the PS will produce and send batches into the SPS with an additional separation of 225 ns leading to a total gap of 525 ns between adjacent batches. This procedure is continued until the SPS is considered to be full.

Given the ratio of circumferences between the PS and SPS, a total of 11 batches could fit into the SPS. But the LHC pre-accelerator complex was planned such that the SPS accelerates up to four batches at a time. However, the SPS has so far been limited to two batches. Once filled, the protons inside the SPS are accelerated to 450 GeV using a series of RF cavities that reach a maximum frequency of 200 MHz. In total, it takes 10.8 s for the SPS to go from empty to fully accelerating a pair of batches. Once fully accelerated, the SPS is ready to inject the batches into the LHC ring. Because the LHC consists of two beams circulating in opposite directions, there are two separate transfer lines used to transfer protons from the SPS into the LHC. TI2 which injects the train into the clockwise rotating beam and TI8 which transfers protons into the anti-clockwise rotating beam, both of which can be seen in Figure 3.1.1.

A breakdown of the pre-accelerators and some of their basic parameters can be found in Table 3.2. From this, it can be seen that during Run 2 it took only 16.8 s for a proton to go from stable hydrogen gas to being injected into the LHC. For a run consisting of 40 batches, it would take approximately 5¹/₂ minutes to completely fill the LHC.

	Linac2	PS Booster	PS	SPS
Final Energy	50 MeV	1.4 GeV	26 GeV	450 GeV
Length	30 m	157 m $\times$ 4	628 m	6.9 km
Repetition time	1.2 s	1.2 s	3.6 s	10.8 s
# of pulse to fill downstream accelerator	1	2	2	20–26

Table 3.2: Parameters related to the different accelerators used in the LHC proton pre-accelerator complex. The number of pulses for the SPS to fill the LHC ring is dependent on the filling scheme used during the various data taking years.

3.1.3 Acceleration in the LHC Ring

After their journey through the pre-accelerator complex, protons are finally injected into the LHC beamline. They are injected to the LHC from the SPS in batches, where the full set of batches in a single beam is referred to as a train. Each train consists of two batches of 48, 64, or 72 proton bunches (based on the data taking year) with the final bunch of the first batch and first bunch of the second batch separated by at least a 3 μ s gap. Inside of each batch the proton bunches are separated by 7.5 m, corresponding to a flight time of 25 ns, consist of $\sim 1.1 \cdot 10^{11}$ protons traveling with an energy of 450 GeV. Because the LHC houses two co-rotating beams, each beam is encased in a separate beam pipe which lie adjacent to each other in the horizontal plane. The use of two beam pipes is required for both acceleration and turning of two same-charged co-rotating particle beams. For most portions of the LHC, the center of the two beams will be separated by approximately 195

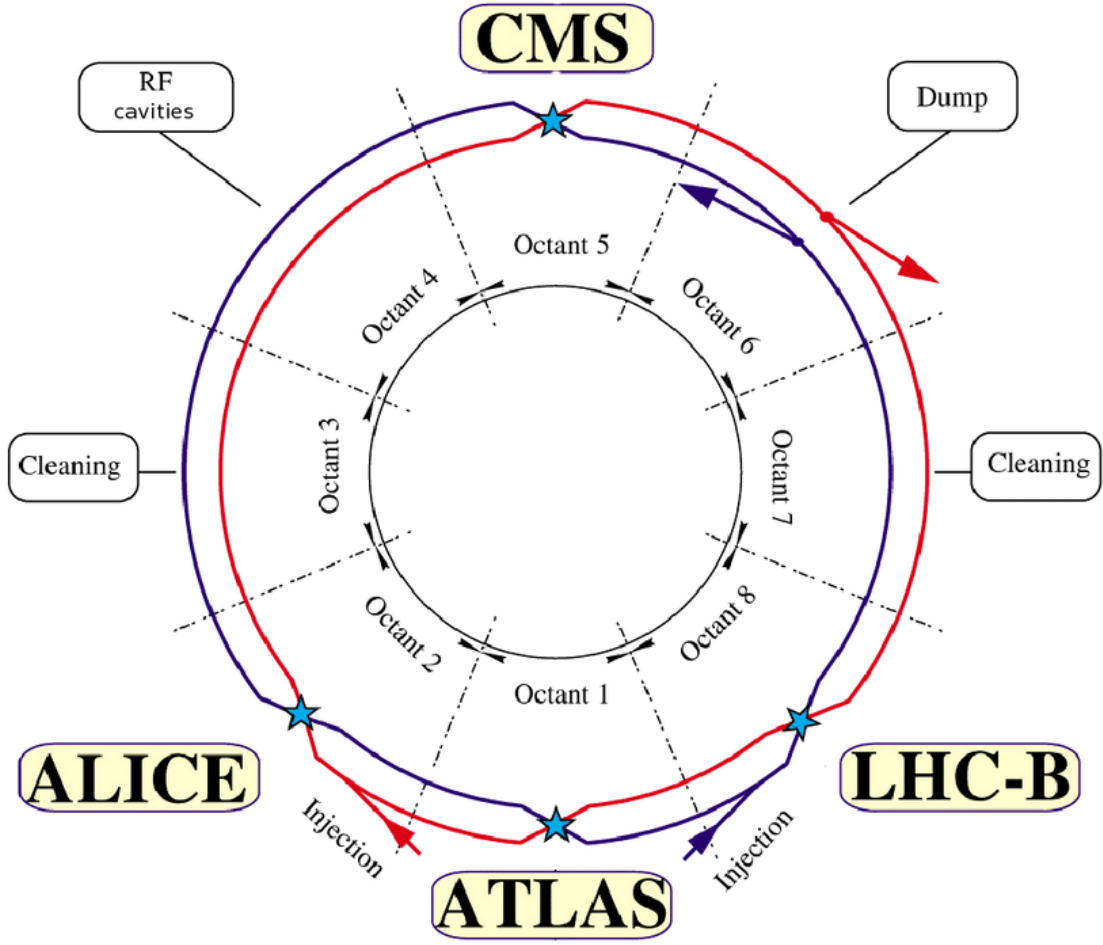


Figure 3.1.2: Diagram outlining the 8 octants of the LHC along with their primary functionality. The stars indicate where the beams cross to produce collisions. Diagram obtained from [6].

mm. However, this distance changes based on the functionality of the section of the LHC ring.

As shown in Figure 3.1.2, the LHC can be split into eight separate octants, each with their own unique functionality. Each octant starts and stops halfway through a large 2.45 km arc consisting of dipole magnets used to bend the beam as it moves through the ring. At the ends of each arc is a dispersion suppression region (DSR) consisting of quadrupole magnets used to maintain the beam's stable quadrupole oscillations. In the center of the arcs are 528 m long straight regions, where the functionality of each straight region depends on the octant. Octants 1, 2, 5, and 8 house the ATLAS, ALICE, CMS, and LHCb detectors respectively. It is within the center of each segment where the detectors are housed and collisions occur. Octants 3 and 7 are used for beam cleaning, which helps maintain long fill periods. Octant 6 is used to facilitate a beam dump, which will be covered in subsection Section 3.1.5. Octant 4 contains the RF cavities used to accelerate the beams to their

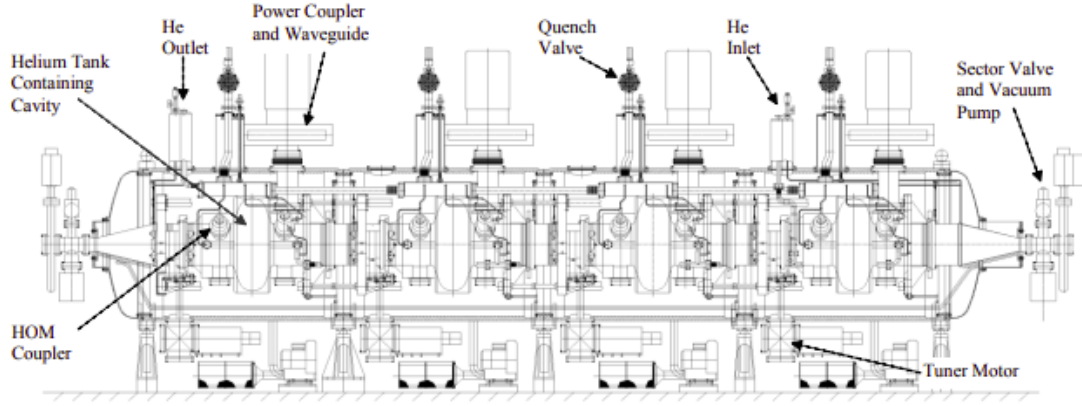


Figure 3.1.3: Schematics of a cryomodule used to cool the RF cavities inside of the LHC. Diagram taken from [7].

final energy.

Octant 4 contains a total of 16 RF cavities used to accelerate the proton beams. Unlike, the pre-accelerator complex the RF cavities in the LHC need to be operated at cryogenic temperatures and make use of superconducting components. To ease in the maintenance of cryogenic temperatures, groups of 4 RF cavities are housed in cryomodules which are cooled using liquid helium. A schematic of a cryomodule can be found in Figure 3.1.3. Inside the cryomodules, the RF cavities are designed to accelerate protons moving in a single direction. As such, two of the cryomodule–RF cavity complexes will be used to accelerate the clockwise rotating beam and two will be used to accelerate the anti-clockwise rotating beam. To facilitate the active material which must surround the accelerated beam, the two beams are moved to a separation of 420 mm inside of the straight region of octant 4. During normal operations, the RF cavities in the LHC oscillate at a frequency of 400 MHz. At this oscillation frequency the proton bunches reach their final energy. In total, it takes approximately 20 minutes for the 450 GeV proton beams injected into the LHC to reach their final energy of 6.5 TeV.

To maintain circulation around the LHC ring, the proton bunches must be turned within the arc regions. Because the LHC consists of two equal-charge counter circulating beams, the beams must be guided using opposite polarity magnetic fields in order to stay within the same ring. This is performed using a total of 1232 dipole magnets placed within the arc sections combining each of the 8 octants. Large constraints are placed on the design specification due to the high energy of the proton beam, the large mass of the proton, as well as size constraints imposed by the LHC cavern. To meet these specifications the dipole magnet is produced using 14.3 meter long Nb-Ti cables which carry currents of up to 11.9 kA along the beam pipes. At these currents, the dipole magnets are

LHC DIPOLE : STANDARD CROSS-SECTION

CERN AC/DI/MM - HE107 - 30 04 1999

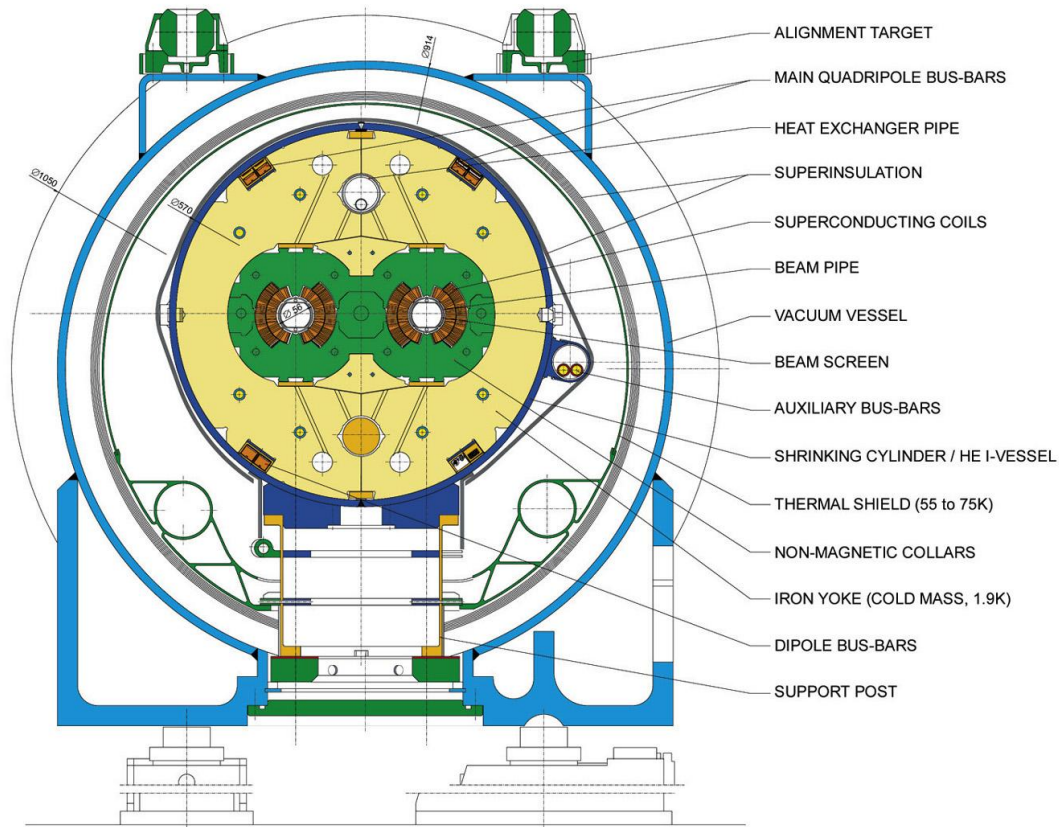


Figure 3.1.4: Cross section of the dipole magnets used to bend the two proton beams inside of the arcs of the LHC. Diagram taken from [8].

able to reach magnetic field strengths of 8.3 T. This is only possible because the Nb-Ti cables are cooled to approximately 1.9° K using superfluid helium. At this temperature, the material becomes superconducting allowing them to push large amounts of charge through the cabling unimpeded. A diagrammatic cross section of a dipole magnet can be seen in Figure 3.1.4.

3.1.4 The Stable Proton Beam

Granted there are no problems introduced during acceleration, the fully accelerated proton bunches will be declared a stable beam. During stable beams, they are allowed to collide within the impact regions of the four detectors for data taking. At this point, it is the LHC's goal to take as much collision data as possible. This is primarily done through two main mechanisms: Keeping the beam emittance low and continuing to collide with the current fill for as long as possible. The emittance is a measure of proton density spread, and in turn goes hand in hand with the LHC's ability to take data over a long fill. Once a fill has ended the LHC needs to prepare a new stable beam, but

that takes time. As such, if the emittance has risen such that the time required to produce a new stable beam will be made up for by its reset emittance, the fill would be cut short. There are many mechanical ways in which improperly tuned LHC operations can cause the emittance to increase. However, under normal operating conditions the primary cause of emittance increase comes from the internal repulsion of protons within a bunch. The primary way in which the LHC can mitigate this effect is through the quadrupole oscillations which the proton bunches have been forced into since their first pre-acceleration step. However, even in optimal conditions the maximum fill time for the LHC is approximately 25 hours, at which point the unavoidable natural emittance increase requires a planned beam dump.

When maintaining stable beams, there are two other primary ways in which an early beam dump can be triggered. The first is from an inability to properly cool the dipole magnets used to bend the proton beams. As mentioned previously, the Nb-Ti cables are cooled to $\sim 1.9^\circ$ K to maintain superconducting properties. If the temperature of the cables rise too much, the magnetic fields inside the material can grow too great causing them to quench and instantly lose superconductive properties. In this instance, the large current forced through the cables causes them to rapidly heat due to Joule heating. Without a beam dump, the continued heating would cause the material to rapidly expand and inevitably leading to the material exploding, greatly damaging the LHC. The primary method in which uncontrolled heating can occur is through the production of an electron gas inside of the LHC beam pipe (often referred to as the electron cloud). The electron cloud is created from the synchrotron radiation emitted by the turning protons colliding with the wall of the beam pipe and exciting free electrons. These electrons are only produced with a energy on the order of eV. However, as the run goes on they will continue to build up, and can also be accelerated by the proton bunches rotating through the LHC. The motion of the electron cloud in response to the electromagnetic field is actually one the largest engineering challenges for the LHC. That is because the motion of the electrons inside of the cloud takes place within the magnetic field of the dipole magnets, which via Newton's third law in turn applies forces on the Nb-Ti cable producing additional heat. As the clouds continue to build, the heat they add to the cables will continue to increase. There are many mechanisms in place to decrease the electron cloud buildup, but a description of them is beyond the scope of this thesis. However, the build-up of the electron cloud is currently one of the primary factors limiting the magnetic field strength of the LHC [7].

The other primary mechanism which can lead to an early beam dump comes from the LHC losing control of the beams. This can refer to any mechanism in which a previously stable beam moves beyond its standard oscillation pattern. Losing control can be caused by a problem occurring

in the operation of the LHC, such as a problem with an RF cavity or a kicker pushing bunches out of their standard orbit. However, it most commonly occurs as a result of long range beam-beam interactions from the counter rotating bunches. When the bunches approach an impact region, they will leave the beam pipes and continue to move closer to each other. Despite the 25 ns spacing ensuring that only a single pair of counter rotating bunches directly collide, the electromagnetic fields of the counter rotating bunches are still able to interact. This leads to non-linear interactions between the bunches. In general these are damped because the direction of forces felt before a bunch crossing will be in the opposite direction of the forces felt by bunches after the bunch crossing a bunch crossing. However, for the first and last bunches in a train, they receive an un-even number of long-range interactions before and after each collisions. In some instances, the build-up of these interactions can lead to the bunches entering various orbital patterns about the center of their RF bucket. As these grow, they can cause the sides of the bunches to collide directly with the beam pipe or the inside of the detector. This type of uncontrolled beam can cause lasting damage to the LHC and its experiments, requiring an immediate beam dump.

3.1.5 The LHC Beam Dump

When a fill is deemed over, the proton bunches need to be evacuated from the LHC beam pipe in a process referred to as a beam dump. However, the high energies of the LHC beam makes beam dumping a difficult engineering feat. During Run 2 a nominal proton beam contained energies of 280–315 MJ, which is comparable to the kinetic energy held in a landing passenger aircraft. To start the beam dump, a series of kicker magnets in octant 6 are ramped up in order to steer the beam out of the LHC ring. The magnet ramp up takes approximately $3\ \mu\text{s}$, and takes place while the abort gap, which is a minimum $3\ \mu\text{s}$ gap between the last and first train, is moving past octant 6. Due to the possibility that the LHC loses control of the beam in such a way that bunches are moving through the kicker magnets while ramping up, dedicated dilution devices are placed in front of the local equipment to prevent damage.

Once protons have been deflected out of the LHC ring, they then go through a 600 m tunnel where they will eventually collide head on with a cylindrical graphite absorber that is approximately 8 m long and 700 mm in diameter. Before colliding with the absorber the proton beam passes through a series of kicker magnets to spread the bunches into a ring-like pattern. This spreads the beam out more evenly over the graphite. But despite the bunch dispersion, the impulse caused by the proton energies being absorbed into the graphite causes it to reach temperatures of 1000° K in

approximately $100\ \mu\text{s}$. To aid in cooling as well as reinforce the structure of the graphite absorber, it is encased in a 12 mm thick stainless steel which is cooled using water and nitrogen gas. The steel casing is then further encased in approximately 7 tonnes of iron for radiation shielding.

3.2 The Compact Muon Solenoid

The Compact Muon Solenoid (CMS) detector [46] is one of the four primary particle detectors located along the LHC ring. CMS is a general purpose detector designed to discover unobserved particles (at the time of its inception) such as the Higgs boson, search for new beyond standard model (BSM) physics, perform new measurements on SM phenomenon, and improve the precision and accuracy of existing measurements of SM processes. The primary features of the CMS detector which allow it to reach these goals with unprecedented performance are its fully silicon tracker, a hermetic calorimeter, and a compact design allowing everything except for the muon systems to sit within the uniform field of a superconducting solenoid magnet.

Inside of the CMS detector, proton bunches cross in the horizontal plane and collide. The detector is designed to measure the properties of the particles emanating from the collisions, as well as the ensuing decay products, through direct interaction with the particles as they travel through the detector. Due to the large number of protons in a single bunch it is expected that there will be multiple simultaneous collisions, referred to as pile-up (PU), per single bunch crossing. The detector is thus designed to differentiate particles based on their point of collision, allowing for PU discrimination. The distribution of PU collisions per bunch crossing for the 2016–2018 data taking years can be seen in Figure 3.2.1

The CMS detector is approximately two stories tall, and utilizes a cylindrical geometry where collisions take place within the geometric center of the cylinder. To best facilitate the shape of the detector, CMS utilizes a special set of coordinates borrowing from both cylindrical and spherical coordinates as well as relativistic quantities. These coordinates are: z , r , η , and ϕ . The origin of the coordinate system is at the center of the detector. The coordinates r and ϕ constitute the standard radial distance and azimuthal angle about the central axis in a cylindrical coordinate system, which is defined by the beam line on which protons enter the detector. The z coordinate is the distance along the beam line with $z = 0$ referring to the geometric center of the detector. However, z is primarily only used when referring to the location of objects (such as collision vertices) close to the origin or when specifying the location of detector components. When referring to the direction of outgoing particles, it is often preferred to refer to them in terms of a polar angle coordinates. The CMS coordinate system makes use of the pseudorapidity η due to its relevance in relativistic kinematics. The pseudorapidity is defined as

$$\eta = -\ln \left(\tan \frac{\theta}{2} \right) \tag{3.2.1}$$

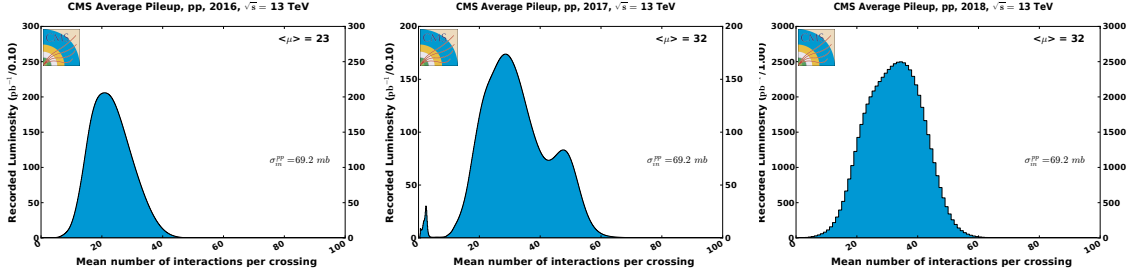


Figure 3.2.1: Distribution of the number of simultaneous PU collisions per bunch crossing for the 2016 (left), 2017 (middle) and 2018 (right) data taking years.

where θ is the standard polar angle with respect to the z -axis in a spherical coordinate system. For massless particles, η is a Lorentz invariant quantity of angular separation. Due to the high particle energies at the LHC, all particles which reach the detector can be kinematically considered massless allowing for us to treat $\Delta\eta$ as being a Lorentz invariant unit of separation for all measured particles. In other words, if a particle undergoes a two prong decay the separation in η between the daughters will be highly correlated with the transverse momentum of the mother particle, but be completely decoupled from the z component of the momentum.

3.2.1 The CMS Subsystems

The CMS detector is composed of multiple layers of subsystems, each with the goal of measuring specific properties of subatomic particles. Through combinations of these measurements, the CMS detector is able to understand intrinsic properties of the particles which directly interact with the detector, as well as the properties of particles produced from collisions that have decayed mid-flight before reaching the detector. These subsystems are layered sequentially and can extend from the barrel region into the endcap and forward regions of the detector. In total, the CMS detector is made of four different subsystems: The silicon tracker [50, 51], the calorimeter [52, 53], the magnet [11], and the muon systems [54]. Furthermore, the tracker can be further split into the pixel detector and the silicon strip detectors, and the calorimeter can be split into the electromagnetic and the hadronic calorimeters. A diagram of the CMS detector with its various subsystems labeled can be seen in Figure 3.2.2 and 3.2.5.

3.2.1.1 The Silicon Tracker

The fully silicon tracker is one of the most important features of the CMS detector. It is able to track the location of particles as they move through the innermost part of the detector with

CMS DETECTOR

Total weight : 14,000 tonnes
Overall diameter : 15.0 m
Overall length : 28.7 m
Magnetic field : 3.8 T

STEEL RETURN YOKE
12,500 tonnes

SILICON TRACKERS
Pixel ($100 \times 150 \mu\text{m}^2$) $\sim 1 \text{ m}^2 \sim 66\text{M}$ channels
Microstrips ($80\text{--}180 \mu\text{m}$) $\sim 200 \text{ m}^2 \sim 9.6\text{M}$ channels

SUPERCONDUCTING SOLENOID
Niobium titanium coil carrying $\sim 18,000 \text{ A}$

MUON CHAMBERS
Barrel: 250 Drift Tube, 480 Resistive Plate Chambers
Endcaps: 540 Cathode Strip, 576 Resistive Plate Chambers

PRESHOWER
Silicon strips $\sim 16 \text{ m}^2 \sim 137,000$ channels

FORWARD CALORIMETER
Steel + Quartz fibres $\sim 2,000$ Channels

CRYSTAL
ELECTROMAGNETIC
CALORIMETER (ECAL)
 $\sim 76,000$ scintillating PbWO_4 crystals

HADRON CALORIMETER (HCAL)
Brass + Plastic scintillator $\sim 7,000$ channels

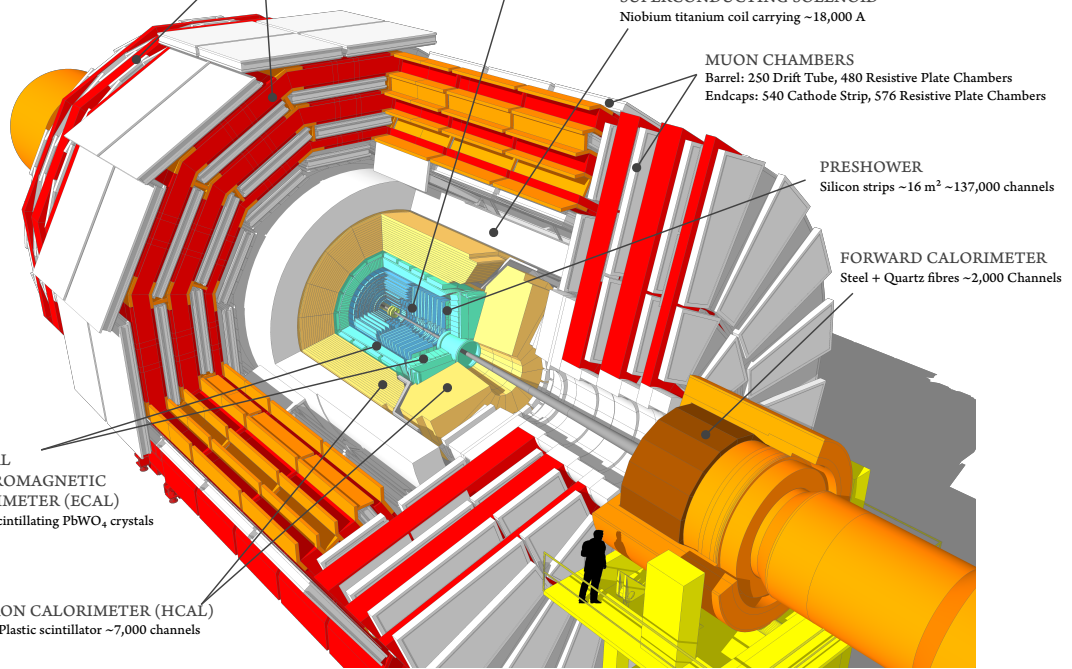


Figure 3.2.2: Diagram showing the CMS detector along with its various subsystems. Diagram taking from [9].

unprecedented precision. With the tracker, the CMS detector is able to reconstruct the trajectory of charged particles allowing it to accurately measure particle momentum and determine the location of the proton-proton collision which produced the particle. Furthermore, it is able to do this while fully resolving the tracks of almost all particles produced during a bunch crossing in the PU environments described in Figure 3.2.1. Because of its ability to resolve the originating location of particles, the tracker is also able to identify particles originating from displaced vertices caused by the decay of long-lived SM (and potential BSM) particles.

To perform these measurements, the tracker is split into an inner and outer tracker. The inner tracker being the pixel detector and the outer tracker being the silicon strip detector, both of which can be seen in the innermost regions of the detector in Figure 3.2.2. Both subsystems make use of sensors largely composed of an N-type silicon bulk sitting on-top of a N-doped silicon backplane. P-type strips or pixels (depending on the sensor type) are implanted along the surface of the sensors. When energetic charged particles pass through the depletion zone of the P-N junction, a localized charge will be produced in the vicinity of the particle. The location of the charge clusters can then be measured by measuring the induced currents in the P-type implants. These sensors are then put in concentric layers, as can be seen in Figure 3.2.3, allowing charged particles to produce a trail of charge deposits as they pass through the inner and outer tracker. The CMS detector can then use this information to reconstruct the full trajectory of charged particles as they pass through the tracker, as well as extrapolate their trajectory through the outer subsystems. In the iterations of the CMS detector used during the $VH(H \rightarrow c\bar{c})$ analysis, the tracker extended out to a distance of $\eta = \pm 2.5$ with respect to the center of the detector, and was able to properly reconstruct the trajectory of particles with $|\eta| < 2.4$ with respect to the collision location.

3.2.1.2 The Pixel Detector

The pixel detector is the first part of the CMS tracker as well as the innermost part of the CMS detector. It provides 3D vertexing by having a high spatial resolution in both η and ϕ . This is achieved by implanting n- (2016) or p-type (2017–2018) pixels on the surface of the silicon sensors. The pixel implants are then bump-bonded to readout chips allowing for a simultaneous measurement of a charged particle’s η and ϕ hit location based on the locations of the closest pixels. A single pixel has a spacial resolution of $100 \times 150 \mu\text{m}^2$ in the η – ϕ plane. However, a single sensor can have a 1D spacial resolution of 10 – $30 \mu\text{m}$ due to the sharing of deposited charge among nearby pixels. Sensors are placed in either the barrel (BPix) or forward (FPix) regions of the pixel detector. The BPix is composed of multiple cylindrical layers of modules which extend to approximately ± 300 mm in the

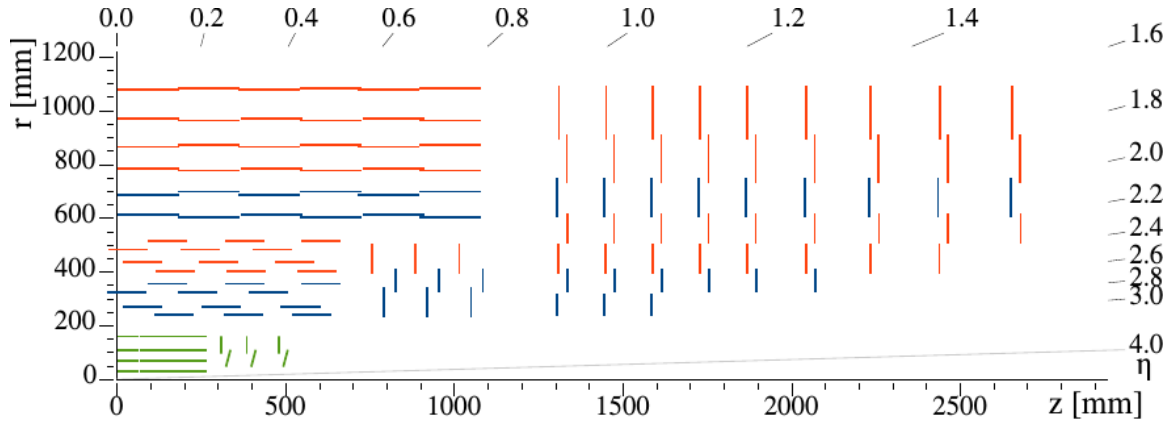


Figure 3.2.3: A diagram of a quarter of the CMS tracker in the r - z plane showing both the outer (red and blue) and inner (green) subsystems. For the outer tracer, the layers/disks in blue represent the modules built in the stereo configuration. The inner tracker shown is for the phase 1 tracker used in the 2017 and 2018 data taking years.

z direction. The FPix modules start at $z \approx \pm 400$ mm and are mounted orthogonally with respect to the beamline to better coincide with the incoming particles angle of incidence when $|\eta| > 1.5$.

During the 2016–2018 data taking periods, two separate iterations of the pixel detector were used. During 2016 the phase 0 pixel detector was used, and for 2017 and 2018 the phase 1 pixel detector was used. Between these two iterations multiple improvements were made to the upgrade of the pixel including: increased number of layers in the BPix and FPix, increased spacial resolution of the pixel sensors, and improvements in readout electronic bandwidth. The differences in geometries between the phase 0 and phase 1 pixel detectors can be found in Figure 3.2.4.

In 2016, the CMS detector made use of the phase 0 pixel detector. The phase 0 pixel detector consisted of three layers of pixel modules in the BPix, with r distances of 43 mm, 73 mm, and 102 mm from the beamline. The FPix consisted of two layers of modules on each edge of the BPix situated at a z distance of 32.5 and 46.5 mm from the center of the detector. In total, there are approximately 66 million pixels in the phase 0 pixel detector. As can be seen in Figure 3.2.4 the BPix and FPix are situated such that any particle passing through the pixel will pass through at least three layers. With the 3-hit design of the phase 0 pixel, a track resolution of $10.4 \mu\text{m}$ in the transverse direction and $20\text{--}45 \mu\text{m}$ in the longitudinal direction could be reached. Such resolution allows for the detector to not just isolate the location of a specific collision in the PU environments of Run 2 of the LHC, but also isolate whether or not a group of particles were produced from the secondary decay of weakly decaying particles like B - or D -mesons produced during a collision. One drawback of the 3-hit design is that pixel tracking resolution decreases if the detector failed to detect a hit in any layer that a particle passed through. This could either be due to hardware spontaneously

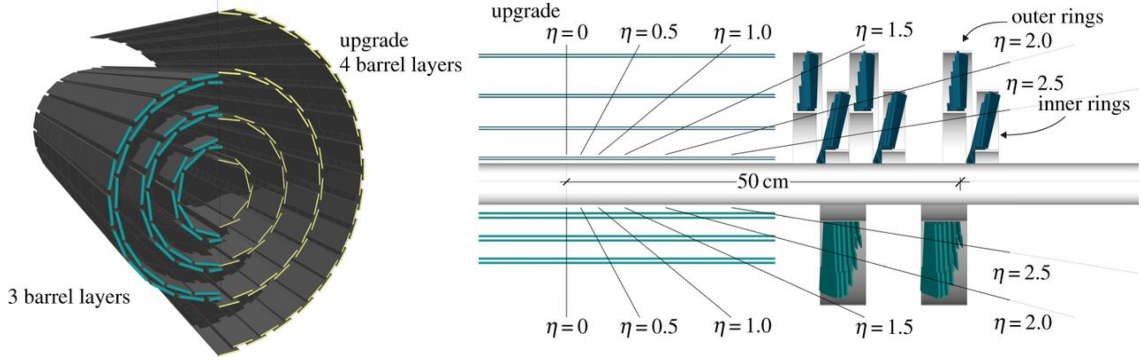


Figure 3.2.4: Comparison of the geometries of the Phase 0 pixel detector used in the 2016 data taking year and the phase 1 pixel detector used in the 2017 and 2018 data taking years. Image taken from [10].

going down during normal data taking operations, the maximum readout bandwidth of the inner most layers of the pixel detector being reached during high PU bunch crossings, or a particle passing through a dead zone between adjacent sensors in a layer. Particularly in the case of spontaneous hardware failure, this becomes more likely as the detector components age and radiation induced damage in the material builds up.

Between the 2016 and 2017 data taking periods, the phase 1 upgrade of the pixel detector was installed in the CMS detector. The phase 1 pixel brought multiple improvements to the previous iteration of the subdetector. Due to the shrinking of the CMS beam pipe to a smaller diameter, the distance from the innermost BPix layer to the beam line was decreased from 43 mm to 29 mm. Furthermore, the pixel detector was able to add additional pixel layers allowing it to move from a 3-hit to a 4-hit design. This was achieved by placing the BPix layers at r distances of 29 mm, 68 mm, 109 mm, and 160 mm from the beam line. The FPix was improved to have three layers incorporating two-ring design allowing for a greater sensor coverage per layer. The FPix layers were then placed at a z distance of 309–338 mm, 384–413 mm, and 479–508 mm from the center of the detector. To further improve hit efficiency the bandwidth of the readout electronics was increased to better handle the higher PU rates achieved in Run 2, and the placement of sensors at $\phi = 0$ and $\phi = 180^\circ$ was modified to remove a small gap that existed in the phase 0 pixel.

Early into the 2017 data taking year a problem was noticed where the readout chips in the phase 1 tracker would occasionally send out large volumes of noise rather than hit data. It was found that this issue could be remedied by power cycling the low-voltage supplies which powered the readout chips. However, this led to an additional fault in the system where DC-DC converters used to increase the low-voltage signal designed to power the phase 0 pixel to higher voltages required to

power the greater number of readout chips would occasionally break upon being power-cycled. These DC-DC converter failures lead to a slow build-up in pixel sensors which were unable to be readout during the 2017 data taking period. This problem was especially prominent in the first layer of the pixel detector where the higher radiation exposure lead to more readout chips requiring power cycles. Overall, this problem lead to lower hit reconstruction efficiencies than intended for the phase 1 pixel detector which worsened for runs later in the data taking period. However, between the 2017 and 2018 data taking periods the DC-DC converters were replaced with new chips that did not have this flaw, allowing the phase 1 pixel to operate as expected during the 2018 data taking period.

3.2.1.3 The Silicon Strip Detector

The outer section of the tracker is composed of the silicon strip detector. The primary purpose of the outer tracker is to extrapolate tracks from the inner tracker and measure of the momentum of charged particles. Silicon strip sensors are designed by placing multiple parallel (or near parallel in the case of the endcaps) p-type implant strips along the surface of the silicon bulk. When a charged particle passes through the sensor, the charge is then collected along the strips and is able to be read off via read-out electronics positioned near the base of the sensor. Unlike pixel sensors, strip sensors only possess a fine resolution in the ϕ coordinate, and the resolution in the η coordinate is roughly the length of the sensor. The magnetic field of the large magnet containing the inner portions of the CMS detector points in the direction of the beam line, meaning that charged particles will curve in the r - ϕ plane as they move through the detector. The choice to use strip sensors allows the outer tracker to accurately measure the momentum of charged particles by measuring this curvature without any loss in performance due to coarser resolution when measuring the track location in η .

The outer tracker can be split into four separate partitions: The tracker inner barrel (TIB), tracker outer barrel (TOB), tracker inner disk (TID), and the tracker end cap (TEC). Sensors can have either 512 or 768 strips and a strip separation of 80–205 μm based on their location in the detector. Both the TIB and the TOB are part of the barrel region of the detector and encapsulate a region of $|\eta| \lesssim 1$. Like the BPix, the TIB and TOB are fitted with sensors lying in the η - ϕ plane and facing the beam line. The TID and TEC encapsulate the end cap portion of the outer tracker, and like the FPix have sensors facing near orthogonally with respect to the beamline. The TID consists of three rings on each side of the TIB in the z direction and extends out to the maximum length of the TOB. The TEC consists of nine additional rings on each side of the TOB and TID. The location of all four partitions of the tracker can be seen as the red and blue sensors in Figure 3.2.3. The sensors in blue represent those built in a stereo configuration, where two sensors are placed

back-to-back with one sensor rotated 100 mrad with respect to the other. Modules in the stereo configuration are able to measure the location of charged particles along two separate axis giving the modules moderate 3D tracking capabilities.

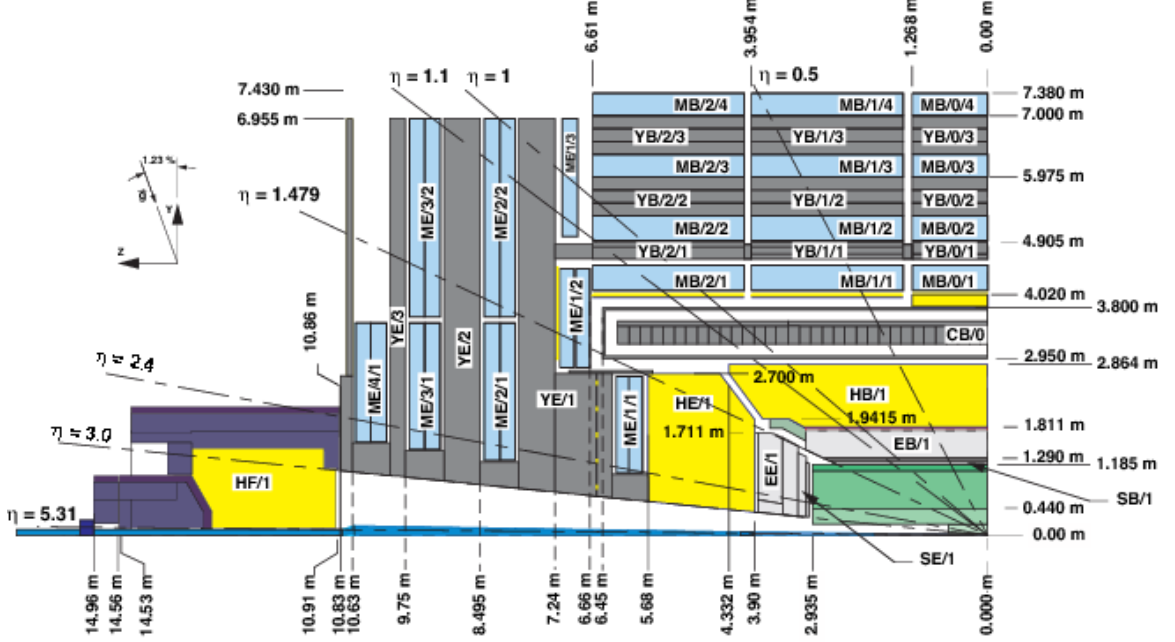


Figure 3.2.5: Diagram of the CMS detector in the transverse planer. The spatial locations of each detector components can be observed in the r and η coordinates as well as the distance along the beamline z .

3.2.1.4 The Electromagnetic Calorimeter

After passing through the outer tracker, particles produced from collisions makes their way to the electromagnetic calorimeter (ECAL). The ECAL is made up of cells primarily composed of a Lead Tungstate (PbWO_4) crystal, and has the primary goal of measuring the energy from photons and electrons that pass through its portion of the detector. Other electromagnetically interacting particles like charged-mesons and muons will still have a portion of their energy measured by the ECAL. However, due to their higher masses they normally pass through the ECAL without depositing all of their energy into the subdetector. The ECAL is broken into a barrel (EB) and end-cap (EE) region, with an additional preshower component installed in front of the EE. Between the EB and the EE, there is a small gap in the region of $1.479 < |\eta| < 1.653$ in which the ECAL is unable to properly reconstruct the energy of particles. The locations of the ECAL components with respect to the rest of the detector can be seen in Figure 3.2.5.

Electrons and positrons with $E \gtrsim 20$ MeV primarily undergo bremsstrahlung inside of the ECAL,

a process where the $e^\pm$ interacts with the nucleus of an atom and loses roughly half of its energy through the emission of a photon. Similarly, photons with $E \gtrsim 20$ MeV will pair produce into an electron and positron of equal energies after scattering off a nuclei. This leads to a cascade of bremsstrahlung and pair production interactions to occur when high energy electrons, positrons, and photons pass through the ECAL. This cascading effect leads to a rapid deposition of the incident particle energy into the ECAL through a process called an electromagnetic shower.

Because each interaction causes the incident particle to produce two out-going particles of roughly equal energy, the energy of the originating electron, positron, or photon can be inferred based on the size of the electromagnetic shower. For this to work, it is imperative that the full shower is contained within the ECAL. This can be achieved by ensuring that the length of a single ECAL crystal is approximately $25\times$ longer than the average interaction length for bremsstrahlung or pair production to occur. Furthermore, to maximize the ECAL's ability to differentiate between the showers produced from separate incident particles, the width of an ECAL crystal needs to be comparable to the average interaction length. For PbWO_4 , the average interaction length is approximately 0.9 cm. This leads to a design length of each ECAL crystal of 23 cm. The radially-facing ends of the crystals are then produced to be equal lengths of 0.0175×0.0175 in η and ϕ . At a radial distance of 129 cm between the front of the EB and the beamline, this gives an $r - \phi$ length of 2.25 cm in the EB. The EE is designed with a 1D length of 3 cm on the front-facing sides of the crystal. For particles like muons and charged mesons, some level of electromagnetic showering will still occur. However, the energy lost per unit distance is inversely proportional to the mass of the particle; leading to a larger interaction length for the incident particle. In this instance the particles only lose a fraction of their energy in the ECAL, and the remainder must be determined from other detector subsystems.

Despite being 2–3 times longer than the targeted interaction length, the granularity of the ECAL was found to be adequate to resolve showering from individual prompt particles, including cases where a Higgs boson decays to a pair of photons. However, neutral pions which decay to two photons in the forward region of the detector will often produce photons which are too closely-separated to be resolved by the $3 \times 3 \text{ cm}^2$ resolution of the EE. In such a case, neutral pions would often be misidentified as being a single high energy photon. To address this scenario, the ECAL preshower is installed in front of the EE. The preshower makes use of two thin lead layers, each of which are followed by a silicon strip sensor similar to those used in the outer tracker. The preshower strip sensors make use of a 2 mm strip pitch, allowing it to fully resolve the tracks from $e^\pm$ pairs produced via pair-production when incident photons collide with the lead layers of the preshower. Forward pions would then leave a signature of two closely-separated charge clusters in the preshower,

followed by one large merged shower in the EE. This gives them a unique signature to be exploited during particle reconstruction

The decision to use PbWO_4 crystals for the ECAL is because in addition to producing electromagnetic showers, they also scintillate as a result of the showering inside of the material. As showering occurs, electrons in the crystal will be excited. When the electrons fall back to their original energy levels, they will release additional photons in the visible range of 360–570 nm, far below the energy required to pair produce. It is expected that this electron excitation will only occur during bremsstrahlung and pair production, making the total number of visible range photons produced proportional to the energy of the incident particle. In the ECAL, the visible scintillation photons are collected at the base of the crystal, and have their yield multiplied by a factor of 50 through avalanche photodiodes tuned to their frequency range. The output of the photodiodes then undergo an analogue to digital conversion in the readout electronics which gets converted to an energy during event reconstruction. For the end cap region of the ECAL, energy measurements obtained from the preshower are added to the energy measured by the EE.

Two additional features about the shape of electromagnetic showers were considered in the ECAL design. The first is due to electrons, positrons, and photons interactions with material in the tracker. As these particles pass through the tracker, they are able to scatter off the nuclei in the sensors or supporting structure. On average, the tracker has an effective interaction length of 0.9, causing electromagnetic showers to start before the particles reach the ECAL. The second is due to the large magnetic field inside the CMS detector causing electrons and positrons produced during the showers to drift. In both instances, this leads to electromagnetic showers often spreading into multiple nearby ECAL crystals. Because of this, ECAL cells are installed in 5×5 groups referred to as superclusters which can be readout all at once for rapid energy estimation. This allows the CMS detector to activate its L1 trigger (Section 3.2.2) based on the energy output of broad electromagnetic showers.

A final prominent feature of the ECAL comes from the fact that during normal data taking operations, radiation will cause short term damage to the PbWO_4 crystal by causing the material to become more opaque. This leads to lower production rates of visible spectrum photons, which in turn leads to underestimation in the energy of ECAL showers. To address this, a 440 nm laser periodically shines light via fiber optic cabling into the ECAL crystals. This calibration is performed periodically during proton beam abort gap, allowing the ECAL to measure the intensity of the laser without interference from collisions. The laser energy measurements are recorded and used offline to calibrate the ECAL allowing for accurate energy measurements despite the buildup of radiation effects. Between proton fills, the temperature of the ECAL is then temporarily raised to reverse the

radiation effects which caused the increased opaqueness of the material.

3.2.1.5 The Hadronic Calorimeter

The HCAL is situated outside of the ECAL, as well as in the very forward region of the detector approximately 10 m from the impact region in the z direction. In total, it is made up of 4 separate partitions: The hadron calorimeter barrel (HB), outer calorimeter (HO), hadron calorimeter endcap (HE), and the forward calorimeter (HF). Unlike the ECAL, the components of the HCAL are situated such that no gaps exist in the detector where particles produced from the impact region could pass through a crack in the HCAL and not have their energy recorded. This can be seen in Figure 3.2.5, in which no straight line in η will fail to intersect a full portion of the HCAL. This makes the HCAL a hermetic calorimeter, allowing it to fully measure the energy of all particles which reach the subsystem by stopping them and absorbing their energy. The two exceptions are muons, which continue to pass through the material in the HCAL with little energy loss, and rarely interacting particles like neutrinos which effectively pass through the CMS detector without interacting with any components.

The HB and HE partitions of the HCAL are situated just outside of the EB and EE of the ECAL. Along with the HO, which is installed just outside of the central region of the magnet in r , they compose what is referred to as the central HCAL. The central HCAL is designed to fully capture the energy of all particles with $|\eta| < 3.0$, where the HB has coverage of $|\eta| < 1.3$ and the HE has coverage of $1.3 < |\eta| < 3$. In order to perform jet identification and maintain a resolution capable of resolving single jets, and in some instances single particles, the HCAL makes use of sampling calorimeters. Unlike the ECAL where a single bulk material is used to induce showering and scintillating, the HB and HE regions of the HCAL make use of repeating layers of a dense absorber material which cause incident particles shower, followed by a plastic scintillator which measures the energy of the particles produced in the showers. The scintillating layers are segmented to have a segmentation of 0.875×0.875 in η and ϕ matching the size of an ECAL supercluster. In the HB, the absorber plates are installed in long planes called wedges. Each wedge has a width of about 20° in ϕ and extends the length of the HB in z . Scintillators are stacked between absorber wedges in what is referred to as towers, in such a way that the 0.875×0.875 scintillator segments will form a column that is inline with an ECAL supercluster. In the HB, the towers contain a total of 17 layers of absorbers and scintillators, giving an effective interaction depth of 5.8 at $\eta = 0$. In the HE, the absorber plates are installed like slices of a disc to better match the endcap geometry. The HE towers then constitute 19 layers of absorbers and scintillators, giving an effective interaction depth of 10 at $|\eta| = 3$.

Due to the the shorter travel distance in the low η regions of the HB, the HO is installed to ensure that the energy of any particles which leak through the HB is measured. To perform this task, two additional layers of scintillators are placed between the CMS magnet and the barrel region of the muon detector. The scintillators are designed to have the same 0.875×0.875 segmentation as the HB and HE, and also sit radially in-line with the HB towers. Due to the large amount of material in the magnet, it is expected that any remaining particles which leak through the HB will have fully showered by the time they have passed through the magnet. The HO is then able to fully measure the energy of the late-showering particles.

Finally, the HF exists in the very forward region of the CMS detector. In order for the CMS detector to detect and measure the energy of non-interacting particles, it must first properly measure the energy and out-going direction of all interacting particles produced from a collision. It can then use conservation laws to find mis-matches in momentum, and infer that it must be due to the presence of an unmeasured particle. To achieve full energy measurements, the CMS detector employs the HF in the $3.0 < |\eta| < 5.0$ region of the detector. The HF starts approximately ± 10 m from the impact region in the z direction and has no detector material sitting in front of it. As such, the HF must be able to fully contain the energy of all forward particles, have good energy resolution in its measurements, and withstand the high radiation doses intrinsic of its distance to the beamline. To achieve this, the HF is made from a large steel absorber embedded with quartz fibers. The charged particles in the showers produce Cherenkov radiation inside of the quartz which is read out to measure the energy of the showers. Quartz fibers with a length of 165 cm (approximately 10 interaction lengths) are used to measure the full energy of the showers. Additional 22 cm long fibers are also embedded into the steel and readout separately to measure the ratio of energy deposited at the two different showering depth. This allows the HF to differentiate between electrons and photons, which deposit their energy much faster, from hadrons, which take the full length of the HF to lose their energy. Because the HF's primary goal is energy measurement rather than object resolution and discrimination, it has approximately half of the granularity as the HCAL with an η - ϕ segmentation of 0.175×0.175 .

During the 2018 data taking year, wedges 15 and 16 in the minus hadronic endcap (HEM) lost readout functionality approximately $1/3$ of the way through the data taking period. This resulted in an inability for the HCAL to measure energy in the region of $-3.0 < \eta < -1.3$ and $280^\circ < \phi < 320^\circ$ corresponding to approximately 3% of the HB+HE coverage. This was observed to have a negative effect on some physics analyses, primarily those heavily relying on distributions of transverse momentum during a collision. However, by vetoing events with missing transverse

momentum pointing in the vicinity of the HEM15/16 modules, missing transverse momentum based analysis strategies can be employed with minimal loss in data.

3.2.1.6 The Magnet

The magnet is one of the defining features of the CMS detector. It exists as a large 12000 tonne superconducting solenoid magnet situated outside of the central region of the HCAL. As such, it is able to immerse the tracking and calorimeter components of the CMS detector inside a uniform 3.8 T magnetic field. It is because of this that the CMS detector is able to reach unprecedented tracking and object reconstruction capabilities. To reach the magnetic field strengths, the CMS detector pumps approximately 18.2 kA of current through a $20.63 \times 2.44 \text{ mm}^2$ superconducting cable. The cable is then reinforced through a casing made of pure aluminum and aluminum based alloys. To produce a solenoid which surrounds the tracker and calorimeter subsystems, approximately 45 km of cabling is used to make a total of 2179 turns around the detector.

To support the magnetic coil, multiple iron return yokes are used in the barrel and endcap regions of the detector. The coil sits within the innermost barrel yoke which houses cooling pipes that pump liquid He across the surface of the aluminum casing to keep the coil at an operating temperature of approximately 4.2° K. The outer return yokes are then used to help direct the magnetic field lines back to the detector. This greatly reduces the magnetic field strength far away from the detector, preventing it from interfering with electronics situated within the CMS cavern. A simulated flux map of the CMS detector demonstrating the use of the return yokes can be found in Figure 3.2.6.

3.2.1.7 The Muon System

The CMS muon system is the outermost part of the detector, and sits within the iron return yokes of the CMS magnet. Muons are both long lived particles and able to pass through matter without depositing much of their energy. This allows most muons to freely pass through the entire CMS tracker, calorimeter, and magnet without losing a large fraction of their energy. As such, muons are detected inside the dedicated muon systems through the use of three different types of gas-based subdetectors: Drift tube chambers (DT), cathode strip chambers (CSC), and resistive plate chambers (RPC). All three types of subdetectors work via a similar design philosophy in which muons ionize a gas as they pass through the active material in the detectors and induce a localized current. The location of the currents are recorded and layers of subdetectors can be used to reconstruct the tracks of muons passing through the muon system. Because the muon system lives within the return yokes of the magnet, muons experience an $\sim 2\text{T}$ magnetic field, but in the opposite direction as the field

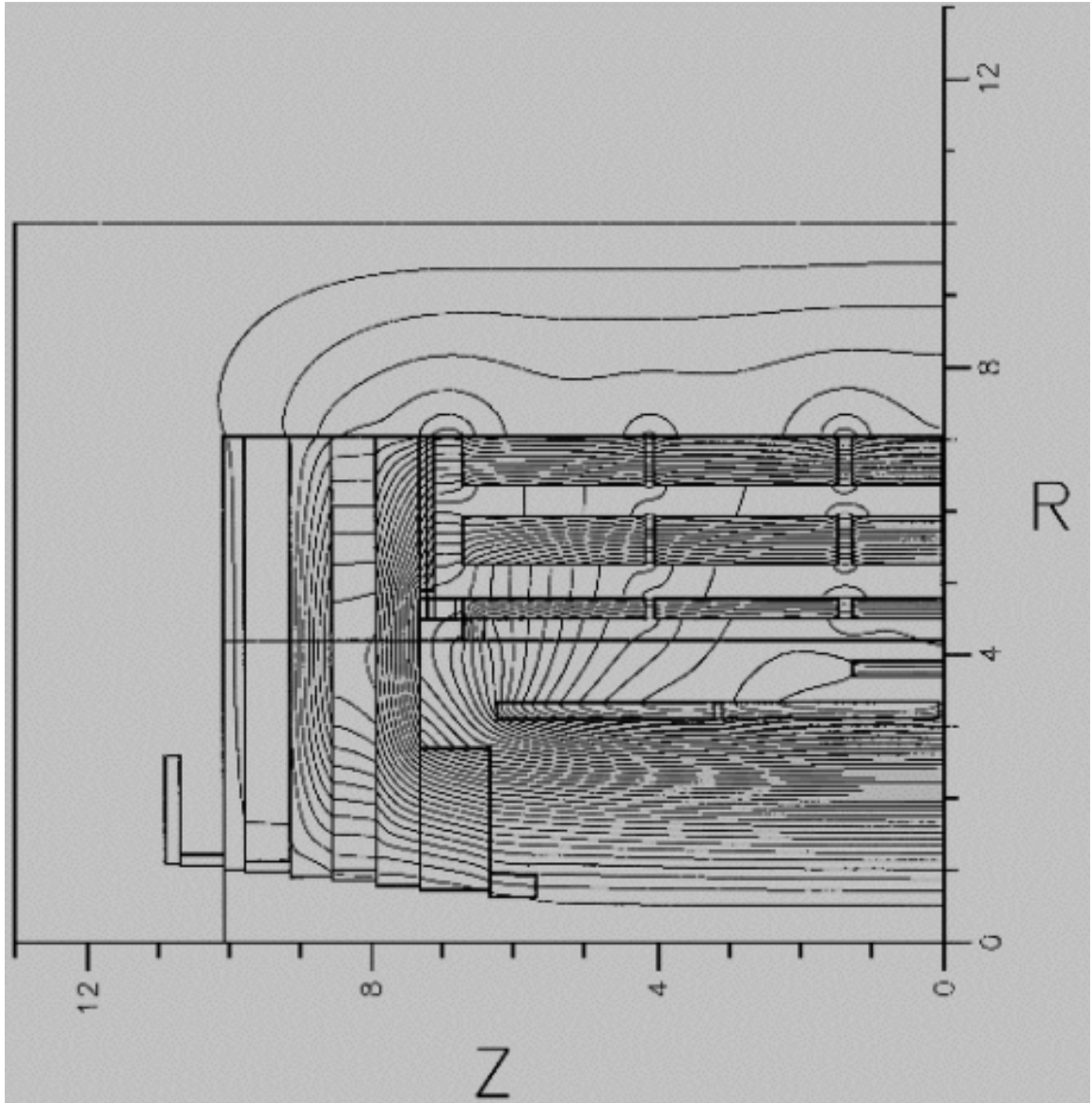


Figure 3.2.6: Simulated magnetic flux lines in the CMS detector. Image taken from [11].

inside the magnet, while traveling through the muon system. This allows the muon system to look for curved tracks enabling strong muon reconstruction efficiencies, especially after being combined with information from the CMS tracking system.

There are two separate partitions in the muon system: The barrel region which is composed of DTs and RPCs situated in the $|\eta| < 1.2$ region of the detector, and the end cap region which is composed of CSCs and RPCs situated in the $1.2 < |\eta| < 2.4$ region of the detector. Because of its far distance from the beam port, the radiation levels experienced by the barrel region will be lower than in other parts of the detector. As such, DTs are used to detect muons in the barrel region of the muon systems. The drift tubes use an 85/15% mixture of Ar and CO₂ gas contained in drift cells with cathodes capable of detecting the ionization of the ArCO₂ mixture. A series of four layers are stacked on top of each other to produce a super layer which resides between each layer of the magnet's return yoke, with an additional layer situated outside the return yoke, giving a total of 4 DT layers. The DTs are able to reconstruct the track of muons passing through the muon system with a precision of 100 μm in $r\phi$ and 150 μm in z and an independent momentum resolution of 15%.

Instead of DTs, CSCs were chosen for use in the end cap region of the detector due to their increased robustness to radiation exposure. The cathodes consist of long trapezoidal chambers with a total width of 10° or 20° in ϕ depending on the layer. The chambers are filled with a 50/40/10% mixture of CO₂, Ar, and CF₄ gases, which produce electrons and positively charged gas ions when muons pass through a chamber. Series of perpendicularly crossing anode and cathode wires measure the position of muons passing through the chambers by detecting currents generated by the ionized gas. A single endcap layer is composed of six overlapping CSC chambers. Like the barrel region of the muon systems, there are a total of four separate layers situated within the end cap regions of the iron return yokes of the magnet. The CSCs are able to reconstruct muons within the end cap region of the detector with a comparable spatial resolution of the barrel region, and with an independent momentum resolution of 25%.

In addition to the DT and CSC components, the RPC is located within the barrel and end cap regions of the muon systems. In total, there are six RPC layers in the barrel region of the detector and four inside the end cap region, extending to a distance of $|\eta| < 2.1$. The RPC is designed to work independently from the DTs and CSCs for the sole purposes of flagging the L1 trigger system. It does this through the use of a 95.2/4.5/0.3% mixture of C₂H₂F₄, i-C₄H₁₀, and SF₆ gases which produce an avalanche of electrons when energetic muons pass through the chambers. This type of ionization gives a much poorer spatial resolution than the DT and CSC designs. However, it makes the RPC excellent at only detecting muons above specific energy thresholds, and is able to do so

with a timing resolution of 3 ns. This makes it excellent at triggering events based on high energy muons produced during collisions.

3.2.2 Event Triggering

There is a slightly paradoxical feature about data taking at the LHC and the CMS detector. The goal is to maximize the amount of data collected in order to search for exceedingly rare physical processes. However, at the 40 MHz bunch crossing frequency it is logistically impossible to readout information from the detector for every set of collisions. This limitation is due to a combination of it being too computationally intensive to convert the detector readouts to a data format that can be saved to computer hardware, there would be too much data to feasibly save, and furthermore the bandwidth required to read out the detector every 25 ns is not feasible with current technology. However, CMS is able to address this issue by exploiting the fact that the frequency at which *interesting* events occur at the LHC is much lower than the bunch crossing rate. To exploit this feature of proton-proton collisions, the CMS detector makes use of a two tiered trigger system which allows the detector to only read-out and save physics events which are the most relevant to analyze.

The first stage of the triggering is done via the level 1 (L1) trigger system. The L1 trigger makes use of components inside of the various pieces of the CMS readout electronics to make rapid decisions about whether or not the detector should be read out. To do this, the L1 trigger system starts by producing trigger primitives from the calorimeter and muon subsystems which look for the presence of muons, electrons/photons, high energy jets, tau leptons, or missing energy. The combination of trigger primitives is then sent to a global trigger system which is able to combine information from all of the trigger primitives to decide whether or not the detector information should be read-out and sent to the next tier of the trigger system. A diagram showing the various ways detector and trigger information are combined for the L1 trigger system can be found in Figure 3.2.7. In total, it takes 4 μ s for the L1 trigger to perform the necessary amount of low level event processing and decide whether or not an event should be triggered. Through the L1 trigger system, the read-out frequency is lowered from the 40 MHz bunch crossing frequency to an average frequency 100 kHz.

After being selected by the L1 trigger system, events are sent to the high level trigger (HLT) system. Due to the decrease in event frequency from 40 MHz to 100 kHz, the HLT is able to perform a higher level of event reconstruction to make improved decisions about whether or not it should be saved and used for analysis. To do this, the detector information is sent to a computer farm which first combines various event fragments from all of the different components of the detector into a

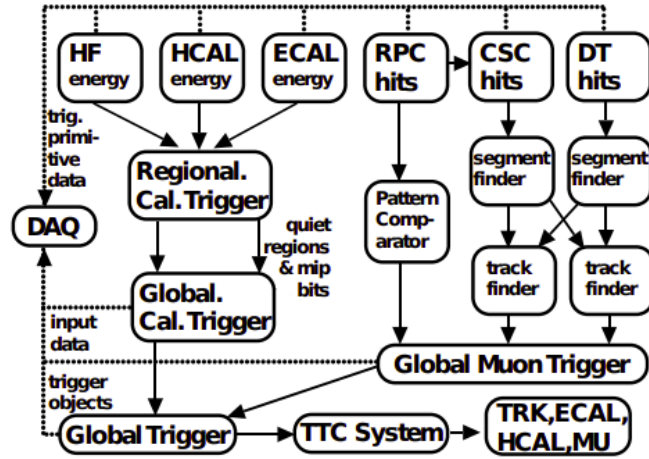


Figure 3.2.7: Schematics of the L1 trigger system employed by the CMS detector. Information from the calorimeter and muon systems are combined to make decisions about whether or not an event should be sent for further processing within $4 \mu\text{s}$ of the collisions. Diagram taken from [12].

single global event. Each full event is then run through an HLT path, where each path represents a sequence of object reconstruction steps followed by a series of selections made on those objects. The path is designed to perform selections in increasing order of computational intensity. This allows the algorithms to decrease average CPU time required to run an HLT path, because if selections which require it to pass simpler algorithms are not met, the HLT path can be aborted without running the more CPU-intensive algorithms. During Run 2, there were over 400 different HLT paths which events would run through before being either selected or discarded. Despite the large computational task required of the HLT system it would only take 50–250 ms on average, with larger times for larger instantaneous luminosities, to run a single event through all of the HLT paths. Based on the HLT, the rate of saved events is decreased to a frequency on the order of 100 Hz. Through the trigger system, the CMS detector is able to decrease the data acquisition rate by 99.999% while managing to save analysis suitable events at high efficiency.

3.2.3 Physics Object Reconstruction

The raw data collected by the CMS detector is in too abstract of a state to be readily used for a physics analysis. To address this issue, detector information is combined to create high level physics objects allowing physicists to use our understanding of particle phenomenology to intuitively analyze the detector data. These objects are a representations of singular or groups of particles which created the various signatures left in the CMS detector during a saved bunch crossing (also called an event). Unlike the actual particles produced during collisions, physics objects are fundamentally

statistical objects. No reconstructed object is produced with 100% accuracy or precision, and as such reconstruction efficiencies and their associated systematic uncertainties are accounted for when analyzing CMS data.

3.2.3.1 Particle Flow

Due to the hermetic nature of the HCAL, the CMS detector is capable of completely reconstructing all interacting particles in an event. Due to this property as well as the high granularity of the ECAL and tracker, particle flow (PF) is the primary algorithm used for event and particle reconstruction with the CMS detector [55]. PF is able to combine information from the tracker, calorimeter, and muon systems of the detector to construct a comprehensive list of final state particles produced during a collision, referred to as PF candidates, to give a global event description. It is from these PF candidates that almost all physics analyses by the CMS collaboration are built on.

3.2.3.2 Tracks

Tracks are a low-level physics object which are traditionally not used for a high level physics analysis. However, they are heavily used in the identification and reconstruction of more high level physics objects. Tracks are reconstructed by combining hits in both the inner and outer regions of the tracker. This is done in an iterative way, where the easiest to identify tracks are reconstructed first to reduce the number of unmatched hits when fitting more difficult to reconstruct tracks. During the reconstruction process, tracks are first seeded by assuming a set of initial constraints and finding hits in the 3–4 inner tracker layers that fall within a certain transverse distance of each other and reconstruct to pass a minimum p_T threshold. The seeds expand one layer at a time into the outer tracker, where after each layer the momentum and other track parameters are re-calculated and used to find a search region for the next hit. After there are no more tracker layers to search through and the full collection of hits are found, the original constraints imposed during track seeding are dropped and the set of track parameters are re-calculated using the full collection of track hits. After all tracks have been produced, a quality selection is made in which all tracks which were found to be missing hits from too many tracking layers are dropped. It has been found that these tracks are often *fakes*, in which the reconstructed track is not associated with a trajectory of a charged particle. A more in depth description of track reconstruction can be found in Section 4 of [56]

3.2.3.3 Vertices

In the CMS detector, a vertex corresponds to the location from which particles originated. In general, there are two types of vertices: primary vertices and secondary vertices. A primary vertex (PV) refers to a location along the beam-line where a proton-proton collision took place. Primary vertices are identified by first extrapolating all tracks built from the tracker back to their origin and extrapolating their location of closest approach to the beam line. Vertices are then formed where the locations of closest approach cluster, and tracks are associated with a given vertex. Additionally, particles originating from the primary vertex are referred to as prompt particles. When referring to an event, the term primary vertex is generally used to refer to the single vertex associated with the location of the proton-proton interaction resulting in the collision being studied. This PV is identified by finding the vertex with the largest sum of transverse momentum of the associated tracks in an event. All other collision vertices are considered to be originating from PU interactions during the bunch crossing, and their associated particles are generally discarded from the global event description. A full description of primary vertex reconstruction can be found in Section 6 of [56].

In addition to primary vertices, we are often interested in the secondary vertices of a collision. A secondary vertex (SV) often corresponds to locations where heavy particles produced from the PV traveled a distance measurable by the CMS detector and then decayed into multiple lighter particles. SVs are identified in a similar way to the PV, where tracks associated with PV particles are traced back to their origin, and locations where a subset of tracks converge can be reconstructed as an SV. In general, two types of algorithms are used to reconstruct the SV, adaptive vertex reconstruction [57] and the inclusive vertex finder [58].

3.2.3.4 Electrons and Photons

Electrons and photons produced in the CMS experiment can have very similar signatures. Both are expected to produce electromagnetic showers that are fully contained within the ECAL. However, electrons leave a signature in the tracker while photons do not. Additionally, electrons still bend as they shower through the ECAL leading to slightly different electromagnetic shower shapes than photons. However, often both electrons and photons will start to shower when passing through the tracker which can obfuscate the differences in their signatures. To identify these objects, ECAL crystal seeds are first produced based on the presence of energy measured above a threshold of 80 MeV in the EB or 300 MeV in the EE. Clusters from nearby crystals are then combined into a

single super cluster representing the energy of a single shower. A special tracking algorithm is then used which combines pixel seeds, the general tracks as described in Section 3.2.3.2, potential tracks associated with showering in the tracker, and super cluster position to produce a set of GSF tracks. A boosted decision tree is used to try to match the GSF tracks with the super cluster, and if it identifies a matching track the super cluster is identified as being an electron or a photon. In the case where no GSF track was found, the object is identified as a photon.

Regardless of the presence of the GSF track, all the components used for reconstruction are fed into PF so that it can be reconstructed as a single EGamma object. The reason for this is because at this point, it can still be difficult to differentiate between the object being an electron, photon, or hadron. This is because high energy electrons and photons can have electromagnetic showers which are not fully contained in the ECAL and could produce hadron-like signatures. Additionally, it is possible that the electron or photon is not actually a prompt particle, but was produced as a secondary particle during the hadronization process. To discriminate between true EGamma objects and hadrons, additional minimum selections as described in [55] are applied to determine whether the EGamma object is appropriately identified or is actually the signature of a hadron. In the case that it is found to be a true EGamma object, a set of loose isolation requirements are also applied to ensure that the particle was not produced during hadronization.

Finally, an additional check is performed to determine whether the EGamma object should be reconstructed as an electron or a photon. Two single discriminate boosted decision trees are applied to the object, one to determine the probability of the object being an electron and the other to determine probability of the object being a photon. The final selection is then based on whether the particle passes a given electron or photon working point. In the situation where it passes the selected working points for both the electron and photon boosted decision trees, the particle is identified as an electron if it has an associated GSF track with a hit in the first layer of the pixel detector. Otherwise it is identified as a photon. More information about electron reconstruction can be found at [59].

3.2.3.5 Muons

The CMS detector is able to reconstruct muons through a combination of the muon systems and the outer tracker. In total there are three levels of muon reconstruction: Local (level-1) reconstruction, stand-alone (level-2) reconstruction, and global (level-3) muon reconstruction. Local reconstruction is the first step of muon reconstruction, and only uses information from the individual DT or CSC sections of the muon detector. In local reconstruction, muon signatures left in multiple super layers

in the DT or 6-cell layers in the CSC are combined to find potential tracks segments inside the muon system. These track segments are then sent into the stand-alone reconstruction step, in which the muon system reconstructs the muon tracks independently of the information from any of the other CMS subsystems. Once all potential tracks through the muon system have been reconstructed, they are fed into the global reconstruction step which looks to match tracks in the muon system with tracks reconstructed by the outer tracker. Two types of track matching are then performed to identify muons: **tracker muons** where tracks from within the tracker are extrapolated into the muon system and then matched with muon tracks (inside-out), and **global muons** where tracks from the muon system are extrapolated back to the outer tracker and matched with general tracks (outside-in). Both inside-out and outside-in reconstruction algorithms are run for redundancy, giving the CMS detector the ability to reconstruct muons with 99% reconstruction efficiency. More information about muon reconstruction can be found at [60].

3.2.3.6 Jets

Jets are the most complex physics object reconstructed in the CMS detector. Due to the high coupling constant of the strong force, quarks and gluons produced from collisions will quickly hadronize, producing more strongly interacting particles from the vacuum in order to maintain a colorless bound state. This process starts to occur in the order of 10^{-22} s after collisions, making it so many hadronic final states involve a cluster of strongly interacting particles. Additionally, heavier hadrons can go through multiple strong and electroweak decay steps which can add leptons, photons, and even more quark matter into the final-state collections. The best way to study these collections is to cluster them together into what is referred to as a jet, with the underlying concept being that the jet encapsulates all of the final state particles produced from the fragmentation of a prompt quark or gluon. In a physics analysis, jets can then be used as a proxy object for quarks or gluons produced during collisions or in the decay of a massive particle.

Because multiple types of particles can end up being classified as a jet, many different algorithms are used to both cluster particles into jets and identify the flavor of the jet progenitor. In the CMS detector, most analysis make use of sequential clustering algorithms in which nearby particles are clustered based on a distance parameter

$$d_{ij} = \min\left(k_{Ti}^{2p}, k_{Tj}^{2p}\right) \frac{\Delta_{ij}}{D} \quad (3.2.2)$$

where k_T is the transverse component of the particle energy, p is an algorithm specific parameter,

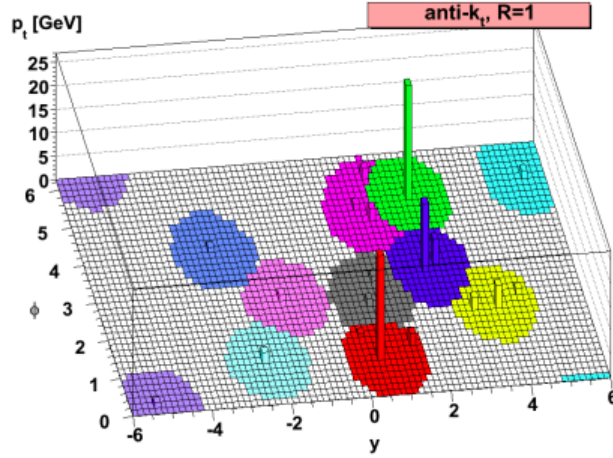


Figure 3.2.8: Diagram showing results of jet clustering when using the anti- k_T algorithm. Image taken from [13]

$\Delta_{ij} = (\eta_i - \eta_j)^2 + (\phi_i - \phi_j)^2$ is the separation in the η - ϕ plane, and D is the cone or jet size which is the maximum separation in the η - ϕ plane in which a clustered particle can have with respect to the center of the jet. The most energetic particle is chosen, and particles are clustered with it in ascending d_{ij} . Once there are no more particles within the distance D to cluster, the most energetic un-clustered particle is selected and used to seed another jet. This procedure is repeated until no more particles are left to cluster. The exact type of algorithm is set by choice in the parameter p where $p = -1$ is the anti- k_T algorithm [13], $p = 0$ is the Cambridge/Aachen algorithm [61], and $p = 1$ is the k_T algorithm [62]. Within CMS anti- k_T is the chosen algorithm for jet clustering. The VH($H \rightarrow c\bar{c}$) resolved jet analysis described in Chapter 6 makes use of anti- k_T jets with a radius of 0.4 (AK4 jets). A simulation showing the clustering of the anti- k_T algorithm with $D=4$ can be found in Figure 3.2.8.

Unfortunately, the nonlinear relationship between the true energy of a particle and the energies measured by the ECAL and HCAL leads to imperfections in properly measuring jet kinematics. To account for this, the CMS collaboration employs a factorized approach in which jet energy corrections (JEC) are applied to all jets after they have been clustered. The JECs start with an L1 correction in which energy clusters from the jet found to be associated with pileup collisions are subtracted from the jet. In addition to ensuring that the jet is only associated with the primary vertex, this allows all further JEC steps to be agnostic of the instantaneous luminosity of the LHC. After the L1 corrections, L2 Relative and L3 Absolute corrections are applied in which the jet's measured energy is adjusted based on its measured p_T and η values. These L2 and L3 corrections are obtained

through a simulation study in which QCD dijet events produce neutrino-less jets that are measured by a GEANT4 [63] simulation of the CMS detector. The difference between true energy of the simulated jet and the simulated detector response is then extracted and used to construct the L2 and L3 corrections. A final L2L3Residuals correction can additionally be applied which account for the residual differences (on the order of a few %) between the detectors response to jets in data and simulation. The L2L3 corrections are made to account for slight mis-modelings of the detector response in simulation which are propagated into the L2 and L3 corrections when applying them to data. An in depth description of the JEC procedures employed by the CMS collaboration can be found at [64].

Beyond measuring jets kinematic properties, large scales efforts exist within CMS to properly identify the type of particle that produced the jet. This process, known as flavor tagging, is vital for the success of analyses in which the production of specific quark flavors is a primary property of the process being studied. Thanks to the advent of machine learning, there are now many competing algorithms which are able to tag jets originating from W bosons, H bosons, top quarks, bottom quarks, charm quarks, or lighter quarks. These jet tagging algorithms have been developed for various different cone sizes, allowing physicists in the CMS collaboration to have a large selection in choosing the algorithm that best suits their needs. A more in depth description of jet tagging at the CMS detector can be found in Section 5.3.

3.2.3.7 Missing Transverse Energy (MET)

Because the counter circulating proton beams are only moving along the beam-line before collision, conservation laws tells us that the sum of the transverse momentum of all final state particles produced in a collision must be 0. Furthermore, because the HCAL is hermetic up to $|\eta| = 5$ we can expect the CMS detector to be able to measure and reconstruct the energy of all final state particles with significant transverse momentum produced in the vast majority of events. The only exception to this is from the creation of non-interacting particles which would pass through the detector without leaving any signatures. This allows the CMS detector to reconstruct a single missing transverse momentum (MET) object in each event to represent the presence of any unmeasured particles. The MET is reconstructed as the transverse momentum vector

$$p_{\text{MET}} = - \sum_i^N \vec{p}_T \quad (3.2.3)$$

where p_i is the transverse momentum of the i -th particle reconstructed in the event. For a purely SM process, the MET would then be equal to the sum of the transverse of all neutrinos produced during the collisions convoluted with any energy mis-measurements. However, there are many BSM models being explored at the LHC, such as the production of dark matter during collisions, which the CMS detector would also be able to reconstruct as MET.

Chapter 4

Quality Control Measurements for the CMS Phase II Outer Tracker Upgrade

During the high-luminosity run of the Large Hadron Collider (HL-LHC) the particle detectors will be faced with the unprecedented task of collecting 3000 fb^{-1} of proton-proton collision data. To do this, the LHC is expected to produce a 200 PU environment for at least 10 years of data taking. Prior to the start of the HL-LHC, the LHC and the associated accelerator complex will under go the three year Long Shutdown 3 (LS3). At this time, the CMS detector will undergo its Phase II upgrade to replace and upgrade components in preparation for the challenges posed by the HL-LHC.

To aid the development of the Phase II CMS detector, Brown University is taking part in the construction of the phase II upgrade of the outer tracker. Silicon sensors, associated test structures, module components, and readout electronics will be sent to Brown University. A series of quality control measurements will be performed on the sensors and accompanying test structures to ensure the manufacturer has produced them to the desired specifications. Groups of sensors which have been certified as good will then be assembled into modules which will eventually be used in the Phase II CMS detector. During my time at Brown, I had a central role in the development and performance of quality control measurements aimed at verifying the electrical properties of the silicon sensors which will go into modules. In particular, those involving the full characterization of strip sensor properties through single and inter strip measurements. In this chapter, I will briefly discuss

the Phase II tracker upgrade of the CMS detector. I will then describe the sensor quality control procedure, with emphasis on global and full sensor characterization, for the strip sensors which are currently being performed at Brown University.

4.1 The Phase II Tracker

The phase II tracker must be designed to withstand the expected radiation doses imposed by 3000 fb^{-1} of proton-proton collision data, while also being able to resolve particle tracks in the expected 200 PU environment of the HL-LHC. It is expected that sensors will receive radiation doses of 10^{14} – $10^{15} \text{ n}_{\text{eq}}/\text{cm}^2$ in the outer tracker and 10^{15} – $10^{16} \text{ n}_{\text{eq}}/\text{cm}^2$ in the inner tracker [14]. To fully withstand these conditions, the phase II tracker is being designed with a target tolerance that is 50% greater than the expected doses. As with the current tracker design, the phase II CMS detector will use a fully silicon tracker design due to its ability to perform tracking with high resolution and natural radiation hardness. Unlike the current tracker, the strip sensors used in the outer tracker will now be produced with a p-doped silicon bulk with additional improvements made during the silicon fabrication to further enhance radiation hardness. As mentioned in Section 3.2.1.1 the total strip/pixel occupancy for a given collision must be below 1% to properly resolve particle tracks. This requires that in addition to making a more radiation hard detector, the overall granularity of the silicon sensors used to build the phase II tracker must be even finer than the $100 \times 150 \mu\text{m}^2$ in $\eta \times \phi$ pixel separation and 80–205 μm in ϕ strip separation of the current tracker.

In addition to addressing new challenges posed by HL-LHC, the phase II tracker is also being designed to enhance current features of the current CMS tracker. To improve its ability to reconstruct forward objects the phase II tracker will extend its range in the CMS tracker from $|\eta| = 2.5$ to $|\eta| = 4$. The phase II tracker will also be attached to the CMS detector’s L1 trigger system through a new low-latency way to detect high p_{T} tracks using tracker modules with two closely separated sensors. This allows for a new type of L1 trigger primitive to be used which rejects low p_{T} particles, and is the first time such a trigger is implemented at a hadron collider. Lastly, the phase II tracker is designed to meet all previously listed design requirements while also decreasing the total tracking volume. As mentioned in Section 3.2.1, electrons and photons pass through the equivalence of approximately 0.9 interaction lengths before reaching the ECAL. This leads to a large number of electrons and photons starting to shower within the tracker, which decreases the reconstruction efficiency of the object. To decrease the material particles have to traverse in the tracker, the phase II tracker will use thinner sensors, less dense structural material in the modules, tilt some outer tracker modules

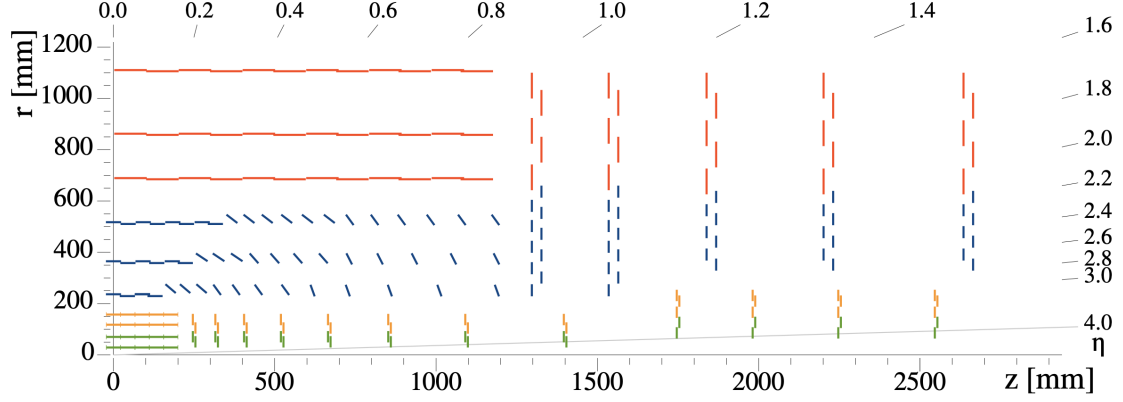


Figure 4.1.1: A diagram of the layout of one quarter of the phase II tracker in the r - z plane. The inner tracker is composed of pixels with two readout chips (green) and four readout chips (yellow). The outer tracker is composed of PS modules (red) and 2S modules (blue). Sensors located near the current TID are tilted towards the impact region to reduce the amount of active volume particle will travel through while also improving the probability that particles will pass through both sensors in a single PS module. Figure obtained from [14].

to be closer to perpendicular to the direction of particles from the interaction point, and re-direct associated wiring to be less densely packed. A diagram depicting the layout of the phase II tracker in the r - z plane can be found in Figure 4.1.1. The extended η range and tilt in the orientation of sensors in the former TID region of the outer tracker can be observed. Each tracker module in the outer tracker consists of two sensors, leading to the apparent decreased number of outer tracker layers with respect to the current tracker drawn in the diagram.

4.1.1 The Phase II Outer Tracker

The phase II outer tracker will be constructed out of new p_T modules. These modules are built out of two closely separated sensors with strips that are parallel in ϕ . Because the trajectory of charged particles passing through the 3.8 T magnetic field of the CMS detector will bend in ϕ with a radius of curvature inversely proportional to the p_T of the particle, the tracker modules can quickly discriminate between high and low p_T tracks based on the distance in ϕ between hits in the two sensors. Outer tracker modules in which both sensors have a detected hit with similar ϕ values (referred to as stubs) can then send the information to the L1 trigger system as a new type of trigger primitive. A diagram depicting the stub detection method for two strip sensors can be seen in Figure 4.1.2. Because both strip sensors in the modules must be parallel for use in a p_T module, the stereo design used in the current CMS outer tracker to perform 3D vertexing can no longer be implemented in the phase II tracker. Instead, the phase II tracker makes use of p_T modules

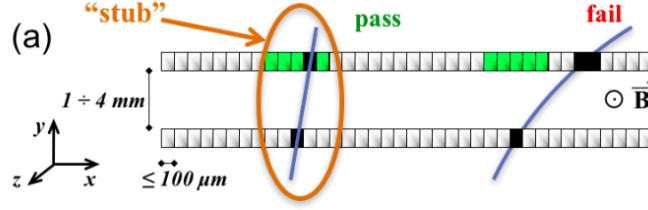


Figure 4.1.2: A diagram depicting the way in which a 2S module can differentiate between high and low p_T tracks. Detected high p_T tracks are identified as stubs and able to be used for L1 triggering. Diagram obtained from [14].

consisting of a strip sensor and a pixel sensor (called PS modules) for the inner layers of the outer tracker and modules consisting of two strip sensors (called 2S modules) for the outer layers. The phase II tracker can then be broken into three separate partitions: The Tracker Barrel with PS modules (TBPS) composed of three layers of PS modules covering the volume with $200 \text{ mm} < r < 600 \text{ mm}$ and $|z| < 1200 \text{ mm}$, the Tracker Barrel with 2S modules (TB2S) composed of 3 layers of 2S modules covering the volume defined by $600 \text{ mm} < r < 1200 \text{ mm}$ and $|z| < 1200 \text{ mm}$, and the Tracker Endcap Double-Discs (TEDD) composed of 5 disks of PS and 2S sensors based on the distance from the beamline, and covering its volume with $1200 \text{ mm} < |z| < 2700 \text{ mm}$.

4.1.2 Design Specifications of 2S and PS-s Strip Sensors

Both 2S and PS modules will be constructed of similar types of strip sensors. They are both composed of $290 \text{ } \mu\text{m}$ thick float zone silicon sensors. The silicon bulk is p-doped and surrounded on its bottom face and all edges by highly p-doped silicon. Along the surface of the active area are n-doped strips implanted (so called n-in-p sensors) to collect charges produced from MIPs passing through the bulk. To improve resolution in η , each strip implant is then split along the center of the sensor such that there are two rows of equally spaced strips which bisect the sensors. To ensure that strips are suitably isolated, especially after the buildup of radiation damage, each strip is independently surrounded by a thin p-stop implant [65]. The active area of the p-type bulk is then surrounded by an n-doped bias ring implant which distributes the ground voltage to each of the strip implants via a polysilicon resistor. As with the strips, an additional p-stop implant ring is placed between the bias ring and the active area occupied by the strips. Finally, a floating guard ring sits outside of the bias ring to ensure high voltage stability. An insulating oxide layer then sits on-top of the n-in-p silicon substrate. Above each strip implant then sits an aluminum readout strip which maintains an AC electric coupling to the strip implants. For data taking, pads at the end of the aluminum readout strips are wire bonded to the read-out chips to measure strip currents.

Type	Outer			Active			Strip/Pixel Cell		Quantity
	width	length	depth	width	length	depth	pitch	length	
2S	94 183	102 700	320	91 440	100 548	290	90	50 274	15360
PS-s	98 140	49 160	320	96 075	47 183	290	100	23 479	5616
PS-p	98 740	49 160	320	96 075	47 183	290	100	1 467	5616

Table 4.1: Various dimensional quantities for 2S, PS-s, and PS-p sensors. Active refers to the dimensions of the p-type silicon bulk, and outer refers to the dimensions of the entire sensors. All distance quantities are given in units of μm .

Similarly, aluminum is layered over the bias ring, ground ring, the outer highly p-doped sensor edge, and on a set of DC pads attached to where the polysilicon resistors contact the strip implants.

The primary differences between the 2S and PS-s sensors stems from differences in their geometry, number of strips, and strip spacing. These changes are primarily motivated by the smaller size of the PS-p pixel sensors which the PS-s sensors must match in size. As such, 2S sensors are approximately $10 \times 10 \text{ cm}^2$ in size and contain 2×1016 strips of length $\sim 5 \text{ cm}$ and separation of $90 \mu\text{m}$. PS-s sensors are approximately $5 \times 10 \text{ cm}^2$ in size and contain 2×960 strips of length $\sim 2.3 \text{ cm}$ and separation of $100 \mu\text{m}$. The full set of differing geometric quantities between 2S and PS-s sensors are listed in Table 4.1.

4.2 Sensor Quality Control

Due to the high demands of the HL-LHC and the importance of tracking to the success of the CMS detector, it is vital to ensure that every sensor which goes into the detector will be able to run within operating specifications for the lifetime of the detector. To ensure this, sensors must pass a rigorous set of qualification criteria before they will be used to build a p_T module. All 2S and PS-s sensors used to construct the phase II outer tracker will be manufactured by the third party vendor Hamamatsu Photonics K. K. (HPK). The sensors will be produced on 6" wafers, a diagram of which is shown in Figure 4.2.1, which are sliced from a larger silicon ingot. Each wafer consists of either a single 2S sensor or two PS-s sensors in the center region as well as associated test structures along the periphery. Sets of diced wafers will then be sent to CERN in batches consisting of approximately 40–60 wafers. The sensors and test structures associated with a single batch will then be redistributed to universities designated as sensor quality control (SQC) and process quality control (PQC) centers. Brown University is designated as both an SQC and PQC center and will individually perform SQC measurements on received sensors and PQC measurements on received test sensors. Once a batch has passed its full set of SQC and PQC tests, the associated sensors will

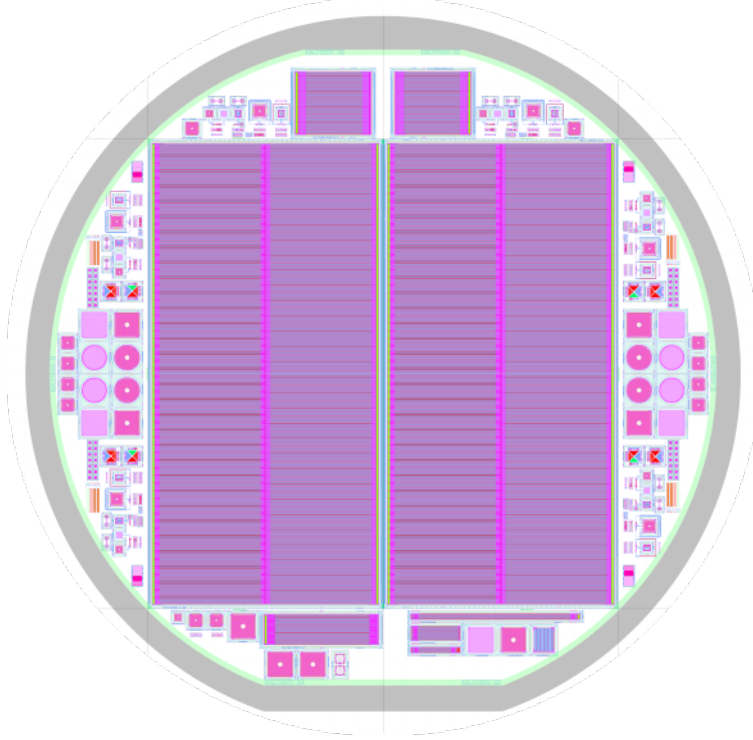


Figure 4.2.1: A computer-aided design drawing of a PS-s wafer. From a single wafer, two PS-s sensors (central left and central right) are fabricated. Additional test structures are built on the peripherals of the wafer to allow for PQC and irradiation testing.

be shipped to one of 10 assembly centers to build p_T modules. In the situation where a batch fails any portion of the SQC or PQC tests, the entire batch of sensors will be rejected and will not be used for construction of the phase II tracker.

In performing SQC measurements, sensors will undergo a series of tests to ensure that they were produced within a set of electrical design specifications described further in this chapter. This includes a set of visual inspections to ensure that there is no visible physical damage to the sensors, a set of current-voltage (IV) and capacitance-voltage (CV) measurements to determine global electrical properties of the sensors, a set of full characterization measurements where individual measurements are performed with respect to the individual strip implants to measure the full set of sensor qualifications, and a set of long term stability measurements where sensors sit at high voltage for multiple days to ensure that their operating currents do not change. However, due to the vast quantity of sensors received per batch it is not feasible to measure every sensor, especially at the expected rate of receiving multiple batches of sensors a month. Instead, a visual inspection, CV, and IV measurements are performed on 10 (15) 2S (PS-s) sensors per batch. These 10 (15) sensors are selected based on a set of initial IV measurements that HPK performs on all sensors,

with preference towards those that displayed worse quality. Of those which underwent the initial measurements, a smaller subset of sensors are then set aside for strip measurements, with preference going towards those exhibiting the worst performance from the previous measurements taken at Brown and by HPK. Finally, 3 (4) 2S (PS-s) sensors selected based on previous IV measurements are kept at high voltage for multiple days to ensure that the leakage current remains stable. The primary motivation behind this approach is that any large defect in the silicon would manifest itself in the poor performance of multiple sensors within a batch. Similarly, any large scale problem arising during production of the strip implants will likely have influence on the set of global characterization measurements.

4.2.1 Visual Inspection

Upon receiving a batch of sensors, the sensors designated for SQC testing will first undergo an optical inspection in which the full front side of each sensor will be imaged. At Brown University, this is performed using an OGP smartscope [66]. Sensors are first placed side-by-side and strip-side up onto the glass optical table of the OGP. The OGP will then automatically scan the full area of sensors taking and saving photos after every motion step. For the visual inspection, automated defect detection software is run to look for any chipping/damage along the edges of the sensors or scratching along the surface of the sensor which could have occurred during the dicing procedure or shipping. Photos of the sensors are additionally used in the instance that poor electrical performance of the sensor is found during testing to see if it could be the result of previously un-noticed markings on the sensor.

4.2.2 CV and IV Characterization

After the visual inspection steps, individual CV and IV measurement are performed on each of the sensors. At Brown University, these measurements are performed with a Tesla200 [67] probe station, which provides a temperature controlled and light-tight enclosure. Sensors are first placed backplane-down on top of the probe station's actively cooled vacuum chuck. Once they are determined to be placed in the correct location, the vacuum is then turned on to secure the sensor in place and the chuck is loaded into the volume of the probe station. A probe needle is then lowered onto one of the bias ring passivation windows with the aid of a digital microscope. Once the probe needle has made proper electrical contact with the bias ring, the enclosure is shut and dry air is pumped into the enclosure.

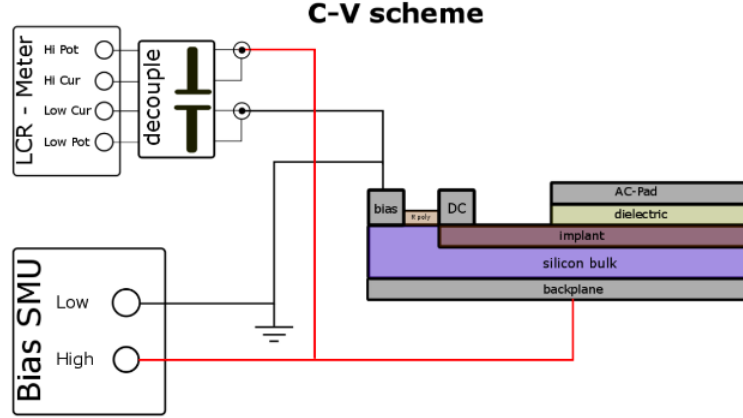


Figure 4.2.2: Diagram depicting the circuitry configuration for CV and IV measurements. The decoupler consists to a pair of capacitors used to decouple the DC voltages applied from the high voltage supply from the terminals of the LCR meter.

For performing the measurements, the vacuum chuck is attached to the high end of an external high voltage supply which will be used to apply a reverse bias current to the backplane of the sensor. The probe needle is attached to the low end of the high voltage supply and grounded. At Brown University, this is performed using a Keithly 6517B [68] which additionally contains an electrometer capable of measuring the current to the bias ring. To perform the CV measurements, the high and low ends of the circuit are additionally attached to the high and low ends of an LCR meter, which is performed by an Agilent 4284A [69] at Brown University. To decouple the applied DC voltage from the leads of the LCR meter, the wires are passed through a capacitor bank before being sent to the LCR meter. A diagram depicting the circuitry scheme used for the CV and IV global characterization measurements can be found in Figure 4.2.2.

At Brown University, the IV and CV measurements are performed in two separate steps. For both sets of measurements, the relative humidity and temperature of the air inside the probe station enclosure as well as the temperature of the chuck are measured through the use of Resistance Temperature Detectors (RTD). To perform the measurements, a negative bias voltage is applied to the backplane of the sensor in 5 V increments from 0 to -400 V, followed by 10 V steps from -400 V to -1000 V. After each voltage step, the circuit is left stationary for approximately half a second to allow the internal currents of the sensor to stabilize. When performing IV measurements, the current flowing through the sensor is recorded through the electrometer housed inside the Keithly 6517B. If it is found that the sensor exhibits sharp breakdown behavior at anytime during an IV measurement, the measurement is stopped and the voltage is safely ramped down. The maximum

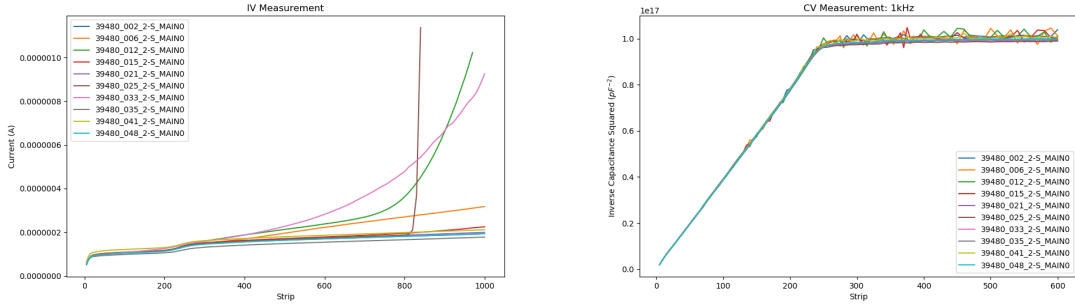


Figure 4.2.3: A set of IV (left) and CV (right) curves for 2S sensors batch 39480. To ease human readability, applied voltage and measured currents are plotted as their absolute values. The double bump structure in the IV measurements is an observed feature of the sensors, the phenomenon is not fully understood but it is potentially due to the depletion of a secondary depletion zone forming inside of the sensors.

CV current is then lowered such that it does not exceed the breakdown current of the sensor. When performing CV measurements, at each voltage step the capacitance of the sensor is measured by the LCR meter applying a 0.5 V and 1 kHz AC potential to the backplane of the sensor, and measuring the AC current response from the grounded bias ring. Additionally, the capacitance of the sensor is measured before any voltage is supplied and subtracted from the $V < 0$ measurements to remove any offset capacitance associated with the circuit.

After the CV and IV measurements have been finished, the results are plotted, as shown in Figure 4.2.3, and inspected to ensure that they are within sensor specifications. All sensors coming from the manufacturer need to be fully depleted at a voltage below $|V| < 350$ V. This can be found by plotting $1/C^2$ as a function of applied voltage, extrapolating the two linear portions of the distribution, and finding the voltage at which they intersect. As can be seen in the righthand plot in Figure 4.2.3, depletion occurs at approximately -250 V for all sensors in 2S batch 39480.

The IV distributions are then checked to ensure that the sensor does not break down until $|V| > 600$ V, the current at -600 V is less than 2.5 nA/mm^3 ($7.25 \text{ }\mu\text{A}$ for 2S and $3.625 \text{ }\mu\text{A}$ for PS-s), and that the ratio of currents between the -600 V and -400 V applied potentials is less than 2.5. Additionally, a more stringent set of criteria are set for sensors to not breakdown until $|V| > 800$ V and for the ratio of currents at -800 V and -600 V to be less than a factor of 2.5 as well. Sensors which pass both sets of criteria are then marked as Grade A, and sensors which pass the first set but not the second are marked as Grade B. Sensors which fail the initial set of criteria are further investigated to determine whether the problem originated from the manufacturer. In case that three or more sensors are fail inspection and the cause is determined to originate from their manufacturing, it is grounds for rejecting the entire batch.

	Single Strip				Interstrip	
	I_{leak} (nA)	R_{poly} (M Ω)	C_{coup} (pF)	I_{diel} (pA)	R_{inter} (G Ω)	C_{inter} (pF)
Normalized (cm ⁻¹)	< 10	—	> 1.2 μm	—	> 10	< 0.5
2S	< 50	1.5 $\pm$ 0.5	> 132	< 100	> 2	< 2.5
PS-s	< 25	1.5 $\pm$ 0.5	> 75	< 100	> 4	< 1.0

Table 4.2: Design specification for strip and interstrip properties when the sensors are biased at a -600 V backplane voltage. Normalized measurements are normalized per cm, with the exception being the interstrip resistance which is normalized in units of G Ω cm. 2S and PS-s specifications are set by approximating the strip lengths as 5 cm and 2 cm respectively.

From the full set of CV and IV measurements, we select which sensors will be sent for full characterization through a series of strip and interstrip measurements. As can be seen in the IV distributions in Figure 4.2.3, all of the sensors show Grade A IV characteristics. However, among the sensors we observe one with a sharp breakdown, two with potentially slow breakdown behavior, and one with a slightly elevated current. As such, all four of those sensors would be selected for full characterization testing.

4.2.3 Sensor Full Characterization

Full sensor characterization corresponds with a set of measurements to measure the electrical properties associated with single strips, and the p-stop homogeneity/isolation between neighboring strips. The single strip measurements aim to measure the following strip properties: The leakage current through the n-type strip implants, the resistance of the polysilicon resistor grounding each strip implant to the bias ring, the capacitance between the AC coupled aluminum strips from the corresponding strip implants, and the ability for DC current to leak through the dielectric layer separating the the aluminum strips from their corresponding strip implants. The interstrip measurements then aim to measure the strip isolation through the effective resistance and the capacitance of the silicon bulk + p-stop implant volume between neighboring strips. The full set of strip and interstrip specification for 2S and PS-s sensors can be found in Table 4.2. For a sensor to fail full characterization, more than 1% of all strips or 2 strips within 5 strips of each other must have single/inter strip characteristics outside of the specifications listed.

At Brown University, single strip and interstrip measurements are also performed using a Tesla200 probe station. As with the global characterization measurements, the backplane of the sensors are secured onto the chuck, and a probe needle which stays stationary with respect to the chuck makes contact with the passivation window in the bias ring to ground the sensor. Two additional probe needles are then used to contact measurement pads along the strips, with the exact configuration of the needles dependent on whether single or interstrip measurements are being performed. The

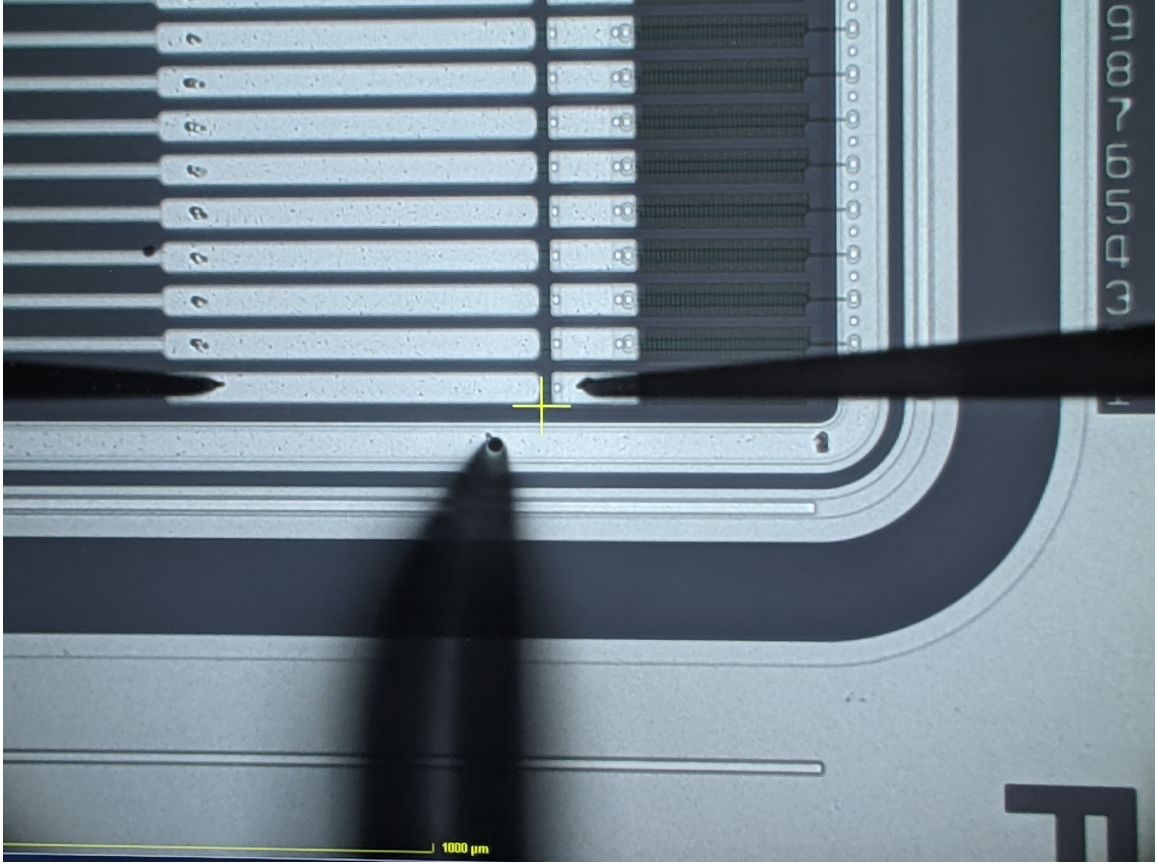


Figure 4.2.4: The configuration of probe needles when used for single strip measurements. The needle used to bias the ground needle is stationary with respect to the vacuum chuck, and as such will not move during the measurement. Between individual strip measurements, the chuck will lower, move to the next strip position, and raise such that the needles probing the strips will once again make contact.

needle configuration for single strip measurements can be seen in Figure 4.2.4, where the bottom needle is used to ground the bias ring and the needles coming in from the left and right are used to make contact with probe pads on the strips. The left set of pads are the AC pads allowing electrical contact with the aluminum strips. The right set of smaller pads constitute the DC pads which are each in electrical contact with the end of an n-type strip implant. When switching to the interstrip needle configuration, the left side needle is manually repositioned to make contact with the DC pad of the strip above the one being contacted by the right side needle. The vacuum chuck is able to move in 3-dimensions, allowing it automatically lower and raise the chuck based on whether or not we need to move the location of the sensor relative to the probe needles, or if we want the probe needles to contact the strip pads.

To measure the various single and interstrip properties of the sensor, the vacuum chuck is first actively cooled to approximately 20°C and dry air is pumped into the probe station enclosure.

Once the environment is stable, the chuck is lowered such that the probe needles used for strip measurements are not making contact with the sensor. A -600 V potential is then applied to the backplane of the sensor in 10 V steps by a Keithly 6517B. Once the backplane has reached -600 V, a set of open capacitance measurements are taken by an LCR meter at an applied 1 V AC potential at a 1 kHz and 100 kHz frequency. The chuck then raises the sensor such that the probe needles used for strip measurements make contact with their respective strip pads. The two probe needles are attached to an Agilent 34970A switching matrix [70] which is able to automatically redirect inputs to various electrical apparatuses based on the strip quantity being measured. The following quantities are then measured as listed, based on whether the needles are set into the single strip or interstrip configuration:

Leakage Current (I_{leak}) – The DC pad connection is attached to the high terminal of a picoammeter, which in turn has the low terminal attached to the common ground. The AC pad connection is then left floating. The current passing through the picoammeter is then read off as the I_{leak} measurement.

Polysilicon Resistance (R_{poly}) – The DC pad connection is attached to the high terminal of a low voltage supply, which in turn has the low terminal attached to the common ground. The AC pad connection is left floating. The low voltage supply applies a -5 V potential to the DC pad, and the -5 V potential is then divided by the corresponding supplied current to obtain the R_{poly} measurement.

Coupling Capacitance (C_{coup}) – The DC pad connection is attached to the low terminals of an LCR meter. The AC pad connection in turn is attached to the high end of the LCR connections. As with the CV measurements, both connections are passed through a DC decoupling unit before connecting to the terminals of the LCR meter. The LCR meter then applies a 1 V AC potential at a 1 kHz frequency to the AC pad, and the resulting AC response current from the DC pad is readoff and used by the LCR meter to compute the capacitance. The 1kHz open capacitance is subtracted from the value measured by the LCR meter to obtain the C_{coup} measurement.

Dielectric Current (I_{diel}) – The DC pad connection is attached to the low terminal of a low voltage supply. The AC pad connection in turn is attached to the high end of the low voltage supply. A 10 V potential is applied to the AC pad, and the resulting supplied current is read off as the I_{diel} current.

Interstrip Resistance (R_{inter}) – The DC pad connection of one strip is attached to the high terminal of a low voltage supply. The DC pad connection of the neighboring strip is in turn attached to the low terminal of the low voltage supply. A 5 V potential is applied to one of the DC pad connections in 1 V steps, and the corresponding change in current passing through the neighboring strip is read off. A linear fit is then made to the distribution of current vs applied voltage, and the voltage/current slope is taken as the R_{inter} measurement.

Interstrip Capacitance (C_{inter}) – The DC pad connection of one strip is attached to the high terminals of an LCR meter. The DC pad connection of the neighboring strip is in turn attached to the low terminals of the LCR meter. As with the CV measurements, both connections are passed through a DC decoupling unit before connecting to the terminal of the LCR meter. The LCR meter then applies a 1 V AC potential at a 100 kHz frequency to the DC pad, and the resulting AC response current from the neighboring DC pad is read off and used by the LCR meter to compute the capacitance. The 100 kHz open capacitance measurement is subtracted from the value measured by the LCR meter to obtain the C_{inter} measurement.

At Brown University, the following measurement apparatuses are used: A Keithly 6485 [71] for the picoammeter, A Keithly 2400 [72] for the low voltage supply, and an Agilent 4284A for the LCR meter. Additionally, during each set of strip measurements the relative humidity and temperature of the air inside of the probe station is measured using a pair of RTDs and the temperature of the chuck is directly measured by the Tesla200 probe station. The current coming from the grounding needle is additionally measured before every set of strip measurements to ensure that it maintains electrical contact before performing the strip measurements. After a single strip's worth of measurements, the chuck will lower the needles and then move 4 strips over for single strip measurements or 15 strips over for interstrip measurements¹. Because there are strips on both sides of the sensor, the type of pads being probed by the respective probe needles will swap halfway through performing single strip measurements on a sensor. To account for this, the switching matrix is set to switch the associated circuits DC and AC circuit connections after it has moved to the other side of the sensor.

A full set of strip measurements taken at Brown University for PS-s batch 36697 can be found in Figure 4.2.5 with the required specifications listed in Table 4.2 shown as the horizontal dashed lines. Two different markings are used to represent sensors coming from the left and right halves of the wafers. As such, a 180° rotational phase can be observed with respect the coupling capacitance trends. Overall, it has been found that sensors made by HPK have electrical characteristics that are

¹A decision by the outer tracker SQC group was made to decrease the number of strips measured per sensor in order to increase the number of unique sensors we can measure per batch

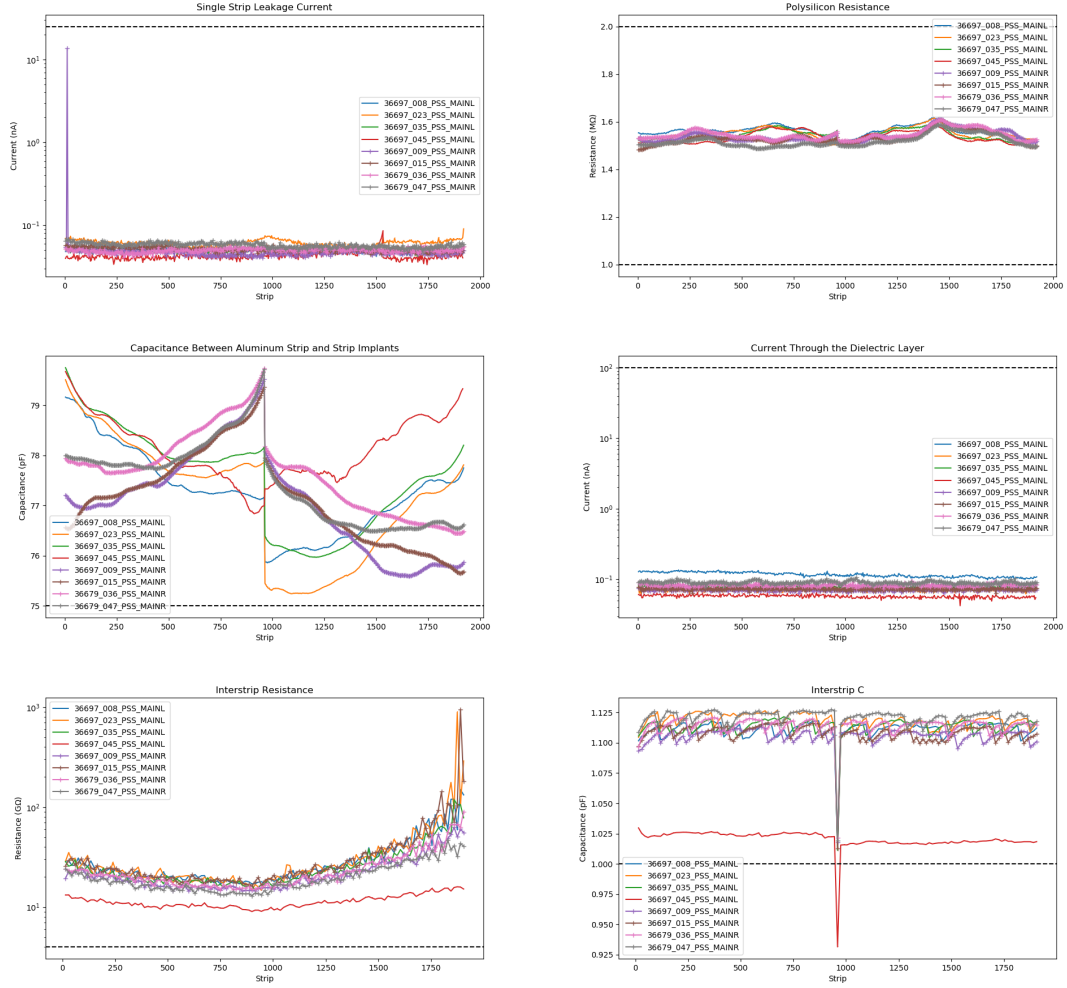


Figure 4.2.5: The strip leakage current (top left), polysilicon resistance (top right), coupling capacitance (middle left), dielectric current (middle right), interstrip resistance (bottom left), and interstrip capacitance (bottom right) measured for a subset of strips on select sensors from PS-s batch 36697 at Brown University. Black dashed lines are used to represent the design specifications listed in Table 4.2 for a PS-s sensor.

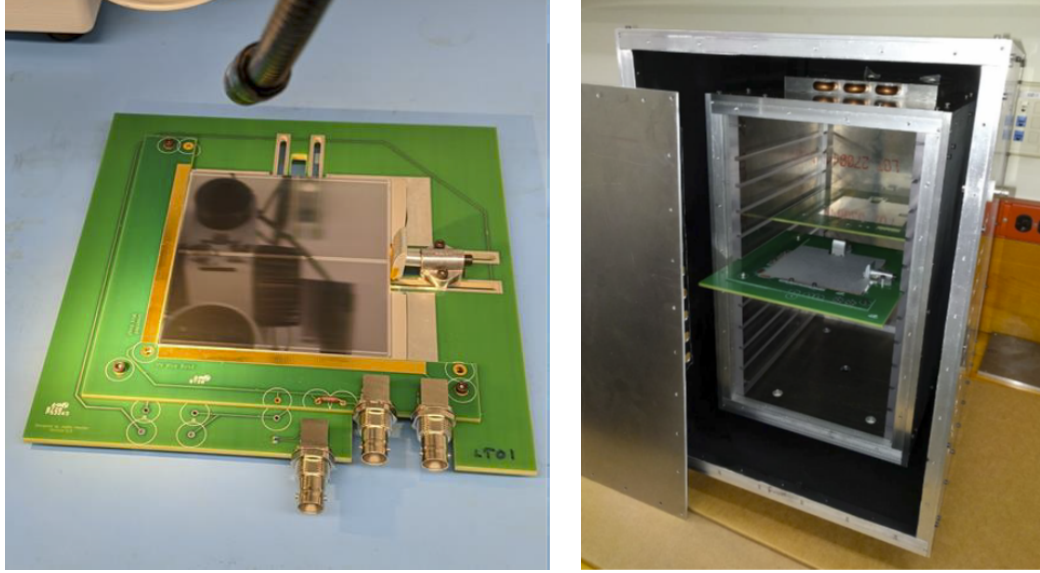


Figure 4.2.6: The sensor holder used to perform long term stability testing (left) and the enclosure used to keep multiple sensors in stable conditions during measurements (right). Up to 10 sensors can be loaded into the enclosure and kept at high voltages simultaneously. The sensors inside of the enclosure are kept at a stable temperature of 20° C and dry air is pumped into the enclosure to keep the relative humidity below 10% during measurements.

well within the required specifications. For example, in sensor 36697_009_PSS_MAINR we observe a leaky strip which is normally indicative of defects in the silicon crystal in the active region below the strip. But despite the strip exhibiting defect behavior, we still find that is within the desired specification. When measuring the interstrip capacitance, we find when measuring the capacitance between the final two strips on a half of the sensor the capacitance drops due to the decreased amount of active silicon surrounding the final strip. As such, we consider dips below the minimum interstrip capacitance value for said pairs of strips, as seen in Figure 4.2.5, to be acceptable behavior.

4.2.4 Long Term Stability Testing

In addition to the previous global and full characterization tests, three of the test sensors will additionally undergo a long term testing procedure. During the long term testing, the sensors are reverse biased at their design operating voltage of -600 V and have their reverse bias leakage currents through the bias ring monitored for at least 2 consecutive days. For testing, sensors are secured onto a sensor holder shown in Figure 4.2.6 and have their bias ring connected to the ground terminal through a wire bond. Up to 10 holders are then loaded into the environment controlled enclosure stabilized to approximately 20° C and 10% relative humidity and have their currents regularly measured. A full set of long term stability measurements for PS-s batches 36696 and 36697 can

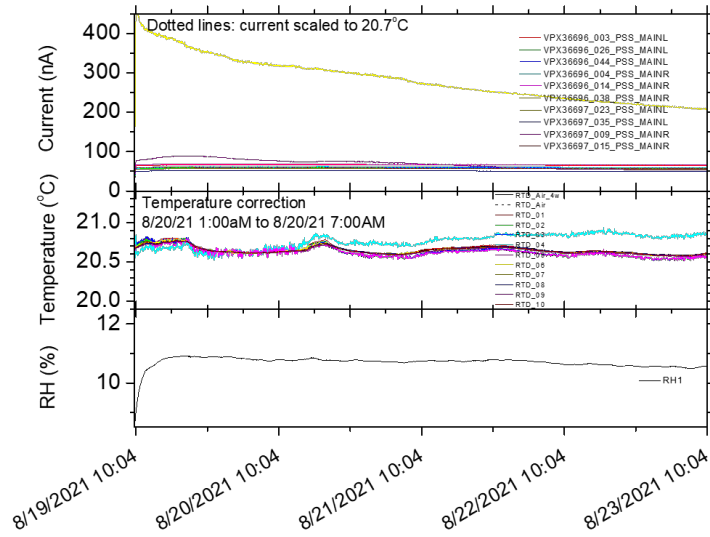


Figure 4.2.7: The reverse bias leakage current for select sensors from PS-s batches 36696 and 36697 taken at a backplane potential of -600 V over a Z day period. Individual current measurements were performed every 60 s.

be seen in Figure 4.2.7. The measurements were performed over a total of 92 hours and 9 of the 10 sensors measured showed stable behavior. Sensor 36696_38R_PSS showed an elevated sensor current during initial global characterization measurements. This appears to be a common behavior with some sensors, where elevated currents will slowly dissipate over long periods of being biased.

Chapter 5

Jets Originating from the Hadronization of Charm Quarks

Identifying jets originating from the hadronization of charm quarks has been a difficult problem for the CMS detector, as well as the ALICE and ATLAS detectors, at the LHC. In general, jets are categorized based on the heaviest quark which contained during the hadronization process. For the $VH(H \rightarrow c\bar{c})$ analysis, we are primarily interested in jets originating from the hadronization of bottom mesons (b -jet), charmed mesons (c -jet), or from the hadronization processes containing only up, down, or strange quarks ($udsg$ -jet or light-jet). These can then be further broken into the following definitions:

b -jet – A jet which at some point in the ancestry chain of the final state particles contained a bottomed-meson.

c -jet – A jet which at some point in the ancestry chain of the final state particles contained a charmed-meson but never contained a bottomed-meson.

$udsg$ -jet – A jet which never contained a bottomed- or charmed-meson in the decay chain of any of its final state particles.

The signatures of c -jets lie in an intermediate phase space between that of b -jets and $udsg$ -jets which can make c -tagging a difficult task at the LHC. Furthermore, the identification of b -jets at hadron colliders had been heavily studied at the Tevatron for the discovery of the top-quark, and the CMS detector was designed with b -tagging in mind. However, high efficiency c -tagging had not been explored for hadron colliders until recent years at the LHC.

In this chapter, I will start by going over the most common ways that charm quarks can be produced at the LHC and some of the key features of c -jets. I will then give a brief overview of jet tagging, also known as heavy flavor tagging, at the CMS detector, with a special importance given to the DeepJet algorithm used to perform c -tagging for the resolved jet VH($H \rightarrow c\bar{c}$) analysis. Afterwards, I will give a description of the methods used to calibrate the DeepJet tagger to ensure that it performs the same way on data as it does on the MC simulation used to train it.

5.1 Production of charm quarks at the LHC

At the LHC, prompt charm quarks are either produced directly from collisions inside the proton via the parton distribution function, from a d or s quark flavor changing after emitting a W boson during a collision, from the splitting of a virtual gluon during collisions as $g \rightarrow c\bar{c}$, from the splitting of virtual photons as $\gamma \rightarrow c\bar{c}$, from the decay of massive vector bosons like $Z \rightarrow c\bar{c}$, $W^+ \rightarrow c\bar{s}$, or $W^- \rightarrow \bar{c}s$, or from the Higgs boson decay of $H \rightarrow c\bar{c}$. Additionally, charm quarks can be produced as secondary particles resulting from the decay of particles that have traveled a measurable distance from the primary vertex before decaying. In general, this is the result of weakly decaying b -mesons and would thus be a c quark produced during the production of a b -jet.

After production, charm quarks quickly hadronize forming a charmed mesons in conjunction with additional hadronic matter. These charmed mesons can be charmonium containing $c\bar{c}$, a D -meson containing a c and u or d quark, or a D_s -meson containing a c and an s quark. There can exist many possible charmonium, D -, and D_s -mesons states due to an abundance of excitation energy accessible to a quark-antiquark bound state. However, the excited states will nearly instantaneously radiate photons and/or light-mesons and fall to their lowest long-lived energy state. For charmonium two lowest states exist based on the spin state of the particle: The J/ψ with a spin of $J = 1$ or the η_c with a spin of $J = 0$. All of the lowest energy states for the D - and D_s -mesons have a spin of $J = 0$ and are defined by their quark content as $D^+(c\bar{d})$, $D^0(c\bar{u})$, $\bar{D}^0(\bar{c}u)$, $D^-(\bar{c}d)$, $D_s^+(c\bar{s})$, and $D_s^-(\bar{c}s)$.

5.2 Key Features of Charmed Meson Decays

The primary features of c -jets are the results of the mass, lifetime, and common decay modes of the charmed mesons. As such, it is important to understand the varying decay features of charmed mesons. These features can lead to c -jets producing unique topologies inside of the CMS detector which can potentially be exploited for c -tagging. Despite similar mass ranges of the various flavors

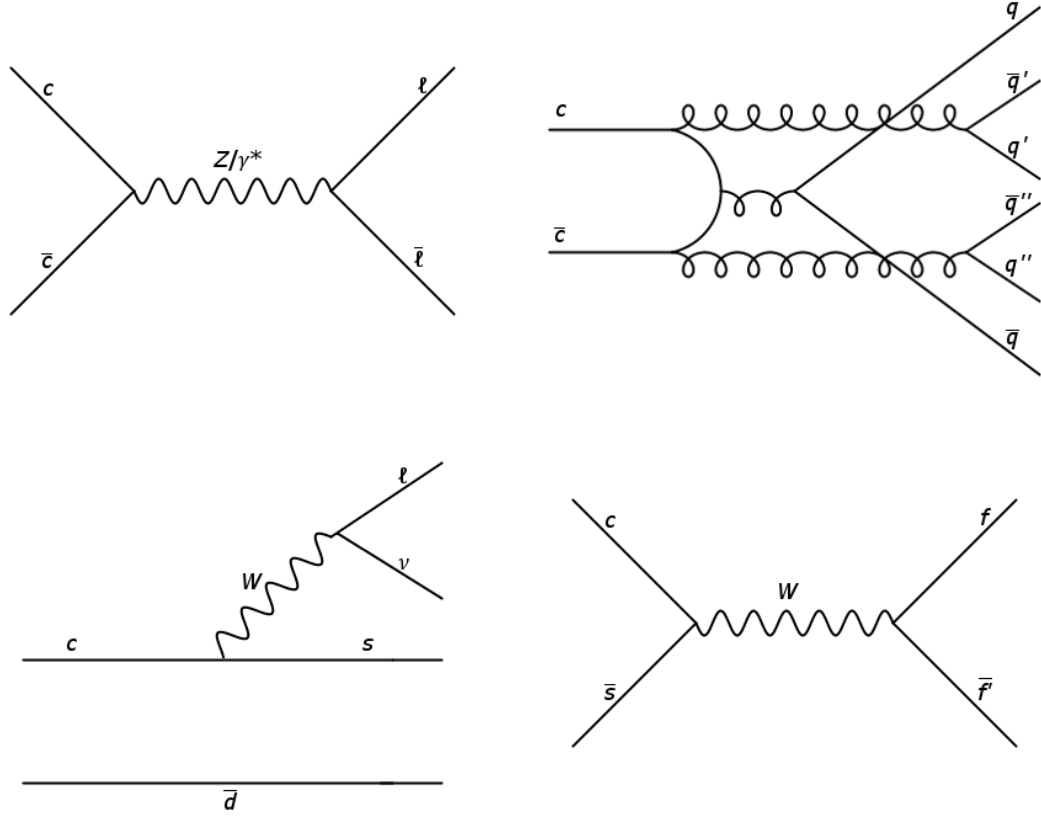


Figure 5.2.1: Feynmann diagrams of the Drell-Yan and three gluon decay modes of the J/ψ (top left and top right), semileptonic decay mode of the D^+ (bottom left), and W decay channel D_s (bottom right).

of charmed mesons, decay parameters can greatly vary leading to some flavors of c -jets being easier to identify than others.

Of all the charmed mesons, the η_c is the most difficult to identify at the LHC [73]. The η_c has a mass of 2983.9 ± 0.5 MeV with an average lifetime of approximately 7.1×10^{-21} s [4]. This gives an average $c\tau$ of 2.1 pm which is too short of a distance to be discerned by the CMS tracker, even after consideration of relativistic effects increasing the decay length. Furthermore, leptonic decay modes via Drell-Yan (DY) are kinematically forbidden because of the $J = 0$ spin state of the η_c . This leads to all decay modes leading to light mesons, and in very rare cases baryon anti-baryon pairs. The rare $\eta_c \rightarrow p\bar{p}$ decay has been used to measure η_c production at LHCb [74] via a specialized analysis. However, the branching fraction of this process is too small to produce a viable signal for c -taggers to reliably identify c -jets containing η_c decays, and the most dominant decay modes all create signals that are not distinguishable from types of light-jets hadronization.

Similar to the η_c , the J/ψ has a mass of 3096.916 ± 0.011 MeV, an average lifetime of 7.2×10^{-21}

Meson	Mass (GeV)	Lifetime (s)	$c\tau$ (m)	Inclusive (semi)leptonic decays	Notable hadronic decays
η_c	3.0	7.1×10^{-21}	2.1×10^{-12}	Forbidden	$\mathcal{B}(\eta_c \rightarrow p\bar{p}) = 0.01\%$ (incl.)
J/ψ	3.1	7.2×10^{-21}	2.2×10^{-12}	$\mathcal{B}(J/\psi \rightarrow e^+e^-) = (5.94 \pm 0.06)\%$ $\mathcal{B}(J/\psi \rightarrow \mu^+\mu^-) = (5.93 \pm 0.06)\%$	$\mathcal{B}(J/\psi \rightarrow ggg) = (64.1 \pm 1.0)\%$ $\mathcal{B}(J/\psi \rightarrow \gamma gg) = (8.8 \pm 0.5)\%$
$D^\pm$	1.9	1.0×10^{-12}	3.1×10^{-4}	$\mathcal{B}(D^\pm \rightarrow e^\pm \nu_e) = (16.07 \pm 0.30)\%$ $\mathcal{B}(D^\pm \rightarrow \mu^\pm \nu_\mu) = (17.6 \pm 3.2)\%$	$\mathcal{B}(D^\pm \rightarrow K + \geq 2 \text{ mesons}) \approx 61\%$
D^0	1.9	4.1×10^{-13}	1.2×10^{-4}	$\mathcal{B}(D^0 \rightarrow e^\pm \nu_e) = (6.49 \pm 0.11)\%$ $\mathcal{B}(D^0 \rightarrow \mu^\pm \nu_\mu) = (6.7 \pm 0.6)\%$	$\mathcal{B}(D^0 \rightarrow K + \geq 2 \text{ mesons}) \approx 78\%$ $\mathcal{B}(D^\pm \rightarrow K\pi) = (6.08 \pm 0.09)\%$
D_s	2.0	5.0×10^{-13}	1.5×10^{-4}	$\mathcal{B}(D_s \rightarrow \tau^\pm \nu_\tau) = (5.48 \pm 0.23)\%$ $\mathcal{B}(D_s \rightarrow e^\pm \nu_e) = (6.5 \pm 0.4)\%$ $\mathcal{B}(D_s \rightarrow \mu^\pm \nu_\mu) \approx 7\%$	$Br(D_s \rightarrow K + \text{mesons}) \approx 48\%$ $Br(D_s \rightarrow \text{non-}s \text{ mesons}) \approx 33\%$

Table 5.1: Mass, lifetime, decay distance, and dominant branching fractions of the ground state charmed mesons. Exact values obtained from [4] and approximate values obtained based on available phase space for decay modes. Uncertainties dropped from mass, lifetime, and $c\tau$ measurements and values truncated to two significant figures for readability.

s, and a $c\tau$ flight distance of 2.2 pm [4]. However, due to its $J = 1$ angular momentum state the leptonic e^+e^- and $\mu^+\mu^-$ decay modes are possible via DY. Additionally, the strong decay modes of the J/ψ are heavily suppressed via OZI suppression leading to large enhancements to the leptonic decay modes. This leads to leptonic branching fractions of $\mathcal{B}(J/\psi \rightarrow e^+e^-) = 5.94 \pm 0.06\%$ and $\mathcal{B}(J/\psi \rightarrow \mu^+\mu^-) = 5.93 \pm 0.06\%$. The remainder of the phase space is dominated by hadronic decay modes, with the two most dominant modes being ggg or γgg with branching fractions of $\mathcal{B}(J/\psi \rightarrow ggg) = 64.1 \pm 1.0\%$ and $\mathcal{B}(J/\psi \rightarrow \gamma gg) = 8.8 \pm 0.5\%$. The gluons will hadronize to variable numbers of pions and kaons, with π^0 s that quickly decay into photons often being produced in the ggg decay mode. In both cases, these can lead to unique topologies, especially in the case of the leptonic decays where the mass of the J/ψ can be reconstructed from 4-momentum of the two leptons.

The presence of only one charm quark greatly reduces the mass of D -mesons to be 1869.65 ± 0.05 MeV for the $D^\pm$, 1864.83 ± 0.05 MeV for the D^0 , and 1968.43 ± 0.07 MeV for the D_s . Similarly, the lack of $q\bar{q}$ annihilation makes it so the only SM decays of the D -meson ground state are through weak decay via a W-boson, greatly increasing their lifetimes. This leads to a mean lifetime and a $c\tau$ flight distance of 1.04×10^{-12} s and $311.8 \mu\text{m}$ for $D^\pm$, 4.10×10^{-13} s and $122.9 \mu\text{m}$ for the D^0 , and 5.04×10^{-13} s and $151.2 \mu\text{m}$ for the D_s . The real flight distance within the CMS detector is then even longer due to the relativistic effects of the high energy collisions, for example a 70 GeV $D^\pm$ will have an average $\beta\gamma c\tau$ flight distance of approximately 3 mm. At these flight distances, the 3D vertexing of the inner tracker is able to resolve the secondary vertices arising from D -meson decays. For their decay modes, all three D -mesons can decay hadronically or semileptonically. $D^\pm$ -mesons will decay semileptonically roughly $1/3$ of the time with the $e\nu_e$ mode happening 16.07% of the time and the $\mu\nu_\mu$ mode happening 17.6% of the time.

Of the hadronic decay modes the vast majority of the final states include one or more kaons with additional pions. The branching fraction of a $D^\pm$ decaying to a kaon(s) and at least two additional mesons approximately is 61%. Despite similar quark content to the $D^\pm$, the D^0 -mesons have roughly half the average lifetime and have their semileptonic decay rates suppressed by roughly a factor of two due to the additional hadronic decay phase space accessible from D^0 - $\bar{D}^0$ mixing. As such, we see that D^0 -mesons inclusively decays to an $e^\pm$ 6.49% of the time and to a $\mu^\pm$ 6.7% of the time. The hadronic decay sector is largely comparable to that of the $D^\pm$ with almost all decays involving one or more kaons. Like the D^0 , D_s -mesons have a shortened lifetime and suppressed semileptonic decay rate compared to the $D^\pm$ -mesons. However, this is due to the c and s quarks being in the same generation allowing for an s-channel decay via a W-boson, as shown in Figure 5.2.1, which is not Cabibbo suppressed. This new decay mechanisms allows the D_s mesons to decay to τ leptons 5.48% of the time. Exclusive of any secondary τ decays, D_s -mesons then semileptonically decay to $e^\pm$ final states 6.5% of the time. Exact values for the branching fraction of μ semileptonic decay modes are not readily available, but can be assumed to be comparable to the e branching fractions [4]. Like the other D -meson flavors, the D_s undergoes similar hadronic decay modes to kaons with additional pions. However, because of the extra s-channel decay mode the D_s can decay to hadronic final states with only π and ρ mesons approximately $1/3$ of the time.

In summary, charmonium states will almost instantaneously decay, with the J/ψ decaying to a pair of opposite charged leptons via a DY processing approximately 12% of the time. Otherwise, it decays via the strong interaction and primarily hadronizes to a large number of mesons, often times with photons present in the final state. D -mesons will often travel a few mm before decaying, with the mesons decaying semileptonically 12–23% of the time depending on the flavor of the non- c quark. In the hadronic decay modes, there are normally three or more mesons being produced, with at least one of them being a kaon. Additionally, charmed mesons can be produced in excited states which will instantaneously decay to one of the ground state charmed mesons through the radiation of photons or mesons. The signatures of excited charmed meson decays would then look similar to that of the ground state versions, but with the addition of one or two extra final state particles. Table 5.1 shows a summary of the various decay features of the ground state charmed mesons discussed in this section. Furthermore, Feynman diagrams showing the most notable dominant decay modes of the J/ψ and D -mesons are shown in Figure 5.2.1.

5.2.1 Key Features for Identifying Charm Jets

Based on the description of the various charm meson decays described in Section 5.2 we can infer information about some of the more notable features of c -jets. For all cases, the meson will decay to at least two measurable particles, meaning that c -jets will on average have more jet constituents than $udsg$ -jets. However, further identification of notable c -jet features depends on whether the jet is originating from a D -meson or charmonium.

One of the most notable features is the case of D -mesons where the mesons will often times travel a distance measurable by the CMS inner tracker of the before decaying to lighter particles. This leads to the production of a reconstructable secondary vertex inside of the jet cone which would not be present from the production of $udsg$ -jets. For fully hadronic decays, the mass of the secondary vertex can be fully reconstructed to a value close to 1.9 GeV. In the case of semileptonic D -meson decays, we expect the leptons to carry a large fraction of the jet momentum. This leads to roughly 20% of c -jets originating from D -mesons to contain a high momentum lepton originating from their secondary vertex. However, these c -jets will also have a large portion of the jet energy and momentum in a neutrino leading to poor reconstruction of the c -jet energy and mass.

Unlike D -mesons, the decay length of charmonium is far too small to be reconstructed by the CMS inner tracker. This makes c -jets originating from charmonium much more difficult to differentiate from the large number of $udsg$ -jets produced at the LHC. In particular, the vast majority η_c final states are indistinguishable from the standard content of $udsg$ -jets making the identification of jets containing the decay products of an η_c infeasible outside of a dedicated analysis. However, the regular presence of photons and leptons in c -jets involving a decaying J/ψ give them some identifiable features. We expect these jets to contain a $\mu^-\mu^+$ or e^-e^+ which reconstruct to an invariant mass of 3.1 GeV approximately 12% of the time. Alternatively, we expect almost all of the fully hadronic decay modes to contain a photon (or closely separated pair of photons) containing $1/3$ or more of the J/ψ energy and momentum.

Given the key features described above, one of the biggest difficulties with successfully identifying c -jets is the degeneracy of said features with b -jets. Despite being more massive particles, weakly decaying B -mesons must decay outside of their generation leading them to have a longer lifetime than D -mesons. As such, the average $c\tau$ flight distance of a B -meson is approximately 450 μm compared to the range of 120–310 μm for D -mesons [4]. Additionally, weakly decaying b quarks will almost always decay to c quarks. This means that final state particle topologies unique charmed-meson decays will be contained inside of b -jets. This adds an extra layer of complexity when identifying

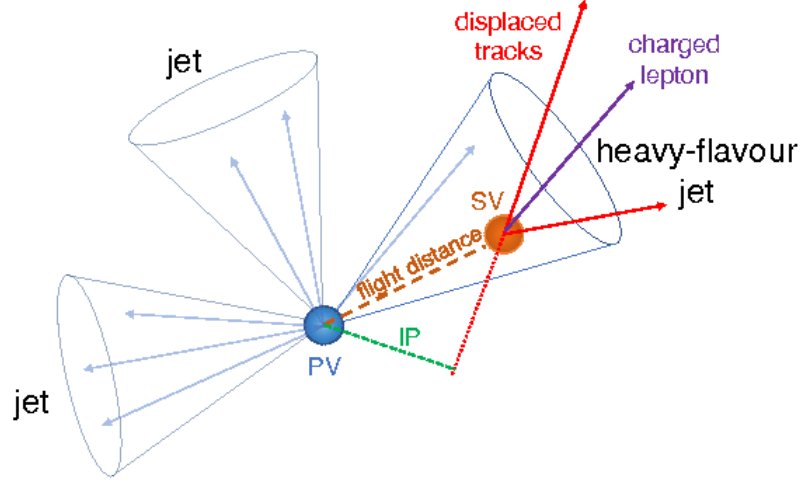


Figure 5.3.1: Diagram showing how a jet may contain a secondary vertex (SV) and its relationship to the impact parameter (IP) variable commonly used for jet tagging algorithm. Figure taken from [15].

c -jets because all of the c -like features of a jet must be weighted against the possibility that it is actually coming from b -jets, which have many of the same unique features but more strongly displayed.

5.3 Jet Tagging at the CMS Detector

High efficiency jet tagging was a design goal of the CMS detector. This was partially motivated by the dominant decay mode of the Higgs boson being the $H \rightarrow b\bar{b}$ decay with a branching fraction of 58% [75]. Furthermore, many BSM processes being studied at the LHC involve the production of TeV scale particles decaying to t quarks, which subsequently decay to hadronizing b quarks. This as well as other factors lead to the ability to perform high efficiency jet tagging to be of the primary design goals of the CMS detector. This was a primary motivation behind the design of the inner tracker which is able to measure the location of secondary and primary vertices with a precision of up to 10–30 μm [50]. Furthermore, the high granularity of the tracker and ECAL allows the CMS detector to fully reconstruct all of the final state particles produced inside of a jet. This allows the CMS detector to use algorithms capable of exploiting a low-level description of a jet and its constituent particles to make decisions about the jet flavor.

Because the Higgs boson is much more massive than the mass of b and lighter quarks, the momentum carried by each of the two quarks in an $H \rightarrow q\bar{q}$ decay is enough that the two quarks will hadronize into two distinct hadrons. This eases the goals of a tagger meant to tag jets originating

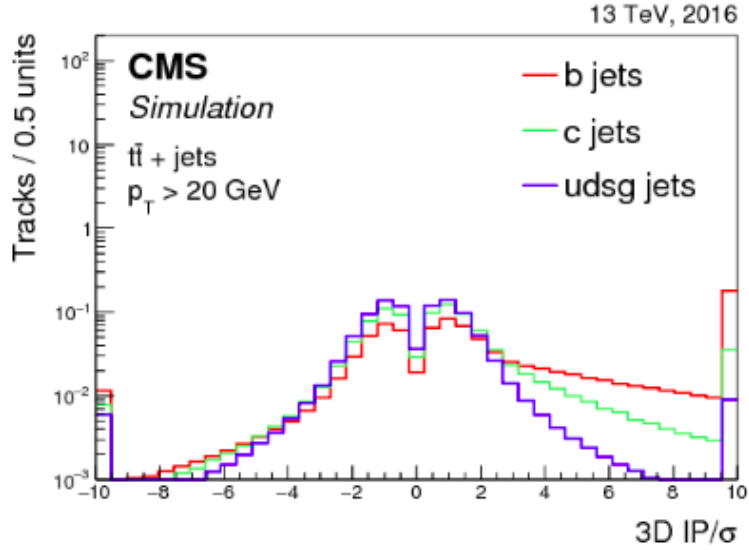


Figure 5.3.2: Comparison of the value of the IP significance of the track with the largest IP value for b -jets, c -jets, and $udsg$ -jets. Figure taken from [15].

from the decay of a Higgs boson, because the vast majority of $H \rightarrow b\bar{b}$ and $H \rightarrow c\bar{c}$ decays will produce jets where the b and c quarks will hadronize into the long lived B - and D -mesons. Because of this, the earliest heavy flavor tagging algorithms at the CMS detector are primarily built off of information about the existence of a secondary vertex within the jet cone. They then calculate a single discriminator value which a selection could be made on to determine whether or not the jet is a b -jet. Often times SV based algorithms make use of the impact parameter (IP) of the tracks within a jet, which is found by calculating a track back to its origin and finding the distance of closest approach to the primary vertex; a diagram of the relationship between the IP and SV can be found in Figure 5.3.1 and a comparison of IP values between jet flavors can be found in Figure 5.3.2. The IP is signed indicating whether the track is curving towards or away from the PV at its location of closest approach. Additionally, an IP significance is often used in which the IP is divided by its relative uncertainty.

The primary heavy flavor tagging algorithms used by CMS during Run 1 of the LHC were: Track Counting, Jet Probability, Soft Lepton, Simple Secondary Vertex, and Combined Secondary vertex [76]. A description of the algorithms follows:

- **Tracking Counting** is a very basic algorithm in which the IP significance of all tracks is calculated, and a b -tagging discriminator is returned as either the second or third largest IP significance of all the tracks.

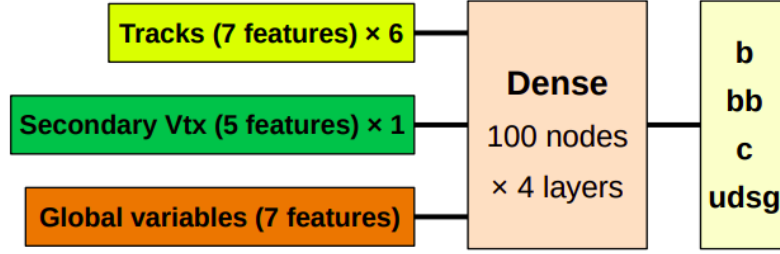


Figure 5.3.3: Diagram showing types of inputs and network architecture used by the DeepCSV jet tagging algorithm. Track features were fed into the network for the six tracks with the highest IP values, as well as a collection of features related to the SV of the jet and overall jet properties.

- **Jet Probability** returns a discriminator equal to the negative log likelihood of the confidence level that all of the jets originated from the PV.
- **Soft Lepton** looks for the presence of an electron or muon inside of the jet. If found, it feeds information about these leptons into a shallow neural network which produces a discriminator determining whether or not the jet is a *b*-jet.
- **Simple Secondary Vertex** uses the adaptive vertex reconstruction (AVR) algorithm [57] to construct a SV inside of the jet cone and uses variables such as the decay length significance to compute a single *b*-tagging discriminator.
- **Combined Secondary Vertex (CSV)** combines information about a SV reconstructed using the AVR algorithm and other kinematic and topological information about the jets to perform a likelihood scan on distributions of the input variables for different jet flavors. The result of the likelihood scan are then used to calculate three discriminator scores describing the likelihood that the jet is a *b*-jet, *c*-jet, or *udsg*-jet.

Moving towards Run 2 the design concepts of the CSV algorithm, in which SV information is combined with global information about the jet, motivated the next steps in heavy flavor tagging. The first widespread jet tagging algorithm was CSVv2, in which the AVR algorithm used to construct the SV was replaced with the higher efficiency Inclusive Vertex Finder (IVF) algorithm [58]. The next was the cMVAv2 algorithm, in which input variables from the CSVv2 algorithm were combined with potential information about soft leptons within the jet and fed into a BDT to obtain a jet-tagging discriminator.

During the shift from analyzing Run 1 to Run 2 data, the use of modern machine learning algorithms like deep neural networks started to take hold at the LHC. At this time, the DeepCSV algorithm was introduced by CMS. DeepCSV was the first DNN based jet tagging algorithm with

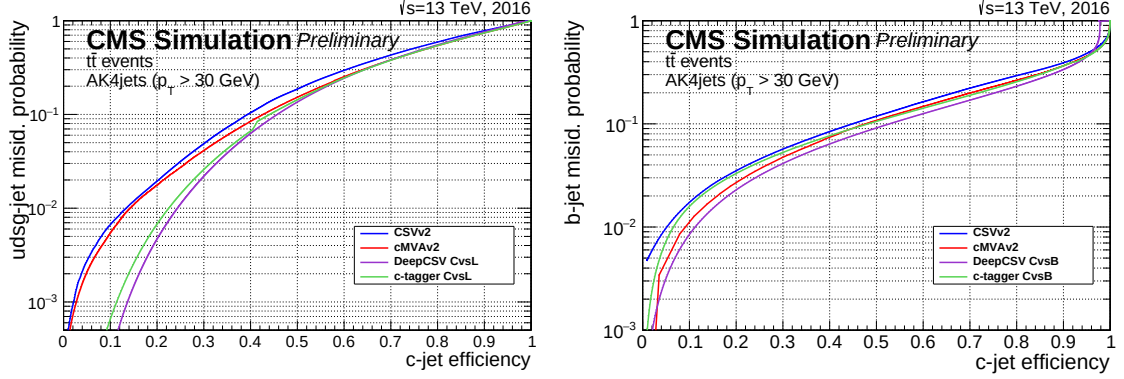


Figure 5.3.4: Comparison of the c -tagging performance as a function of b -jet and $udsg$ -jet mistag rate for early Run 2 jet taggers. Values closer to the bottom right are considered better.

widespread use from the CMS collaboration. It made use of all of the CSVv2 variables including a few more lower level IP associated variables for a total of 66 input features. They were then fed into a DNN with four hidden layers which was trained for 500 epochs on a sample of 50 million jets. Additionally, DeepCSV was produced to perform b -tagging, bb -tagging, c -tagging, and $udsg$ -tagging. A diagram showing the structure of DeepCSV and a comparison of DeepCSV's c -tagging performance to other early Run 2 jet taggers with c -tagging capabilities can be seen in Figures 5.3.3 and 5.3.4. The use of deep learning in CMS lead to even more DNN based taggers for boosted jet tagging such as DeepAK8, BEST, ImageTop, HOTVR, the DeepDoubleB tagger, and many others [77].

As progress in jet tagging capabilities continued to move forward in Run 2, newer algorithms have continued to improve performance through the use of more and more low level jet variables. The DeepJet algorithm, which is described in more detail later in this section, is able to make discriminator decisions using the features of the individual PF candidates that get clustered into the jet, as well as global properties of the jet and potential SVs. There are also state of the art graph neural network (GNN) based algorithms such as ParticleNet [78], which was used in the merged topology channel of the $VH(H \rightarrow c\bar{c})$ analysis with full Run 2 data, which are capable of fully correlating the spatial information of the jet constituents to make the strongest large cone jet tagging decisions to date.

5.3.1 DeepJet

The DeepJet algorithm is a DNN based jet multiclassifier used by the CMS collaboration. It is able to take the low-level features associated with a single AK4 jet as an input and assigns discriminator

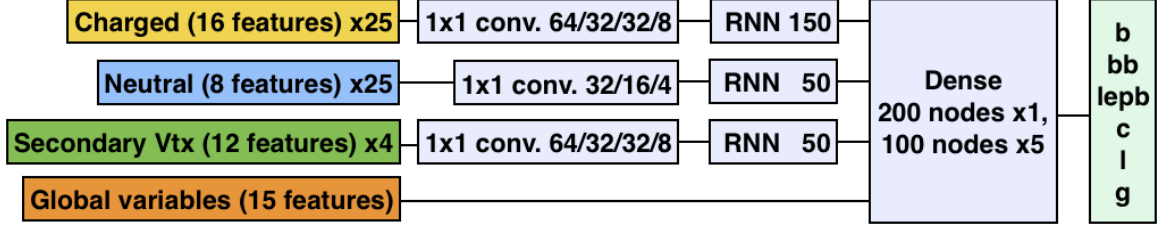


Figure 5.3.5: Diagram showing the network architecture and types of input features used by the DeepJet algorithm. The network can take features associated with up to 25 charged PF candidates, 25 neutral PF candidates, and four potential secondary vertices as inputs. The network then makes use of convolution, recursive, and fully connected network layers to produce multiple jet flavor discriminants.

values to the associated jet giving the probability for the jet to come from the hadronization of: a single bottomed-meson (b), a pair of bottom-mesons (bb), a bottom-meson that decayed leptonically (b_ℓ), a charmed-meson (c), an uds quark (l or uds), or the hadronization of a gluon (g). Additionally, it is the algorithm used to perform c -tagging in the $VH(H \rightarrow c\bar{c})$ resolved jet analysis described in Chapter 6.

DeepJet makes use of up to 663 input features per jet split into four separate categories: charged PF candidate features, neutral PF candidate features, secondary vertex features, and global features. In total these consist of 16 kinematic and track features (8 each) associated with the 25 hardest charged PF candidates, 8 kinematic features associated with the 25 hardest neutral PF candidates, 12 kinematic and distance features associated with the 4 most likely secondary vertices candidates, and 15 global features associated with the jet. The PF candidate and SV features are first fed through separate sets of convolutional layers in order to perform automatic feature-engineering. This is done by passing a series of 1×1 filters over the feature arrays, slowly transforming them to a new set of 8 (or 4 for the neutral candidates) features. The underlying concept is that during the training, the network learned to transform the input features into their most useful representation. The sets of engineered features are then fed into three separate LSTM [79] layers to properly handle the situations where jets with less than the maximum number of PF candidates and potential SVs are fed into the network. The outputs of the three LSTM nodes along with the array of global variables are then fed into a series of fully connected layers, which passed their outputs through a softmax function to create the discriminator. A diagram showing the network architecture and how the different input features are treated by the network can be found in Figure 5.3.5.

For training, the network was trained using supervised learning on a total of 100 million different jets produced from a mix of QCD multijet events and fully hadronic $t\bar{t}$ events. Jets were labelled based on the presence of b - or c -hadrons in the ancestry of the final state particles. In the event that

the jet contained both b - and c -hadrons, it is labelled as a b -jet. Training was performed using the Adam optimizer with a learning rate of 3×10^{-4} for 65 epochs using a categorical cross entropy loss function. Additionally, if the loss on the validation sample was found to stagnate for 10 epochs, the learning rate was cut in half.

5.3.2 Charm Tagging with DeepJet

When using DeepJet, c -tagging efficiencies greatly increases over previous jet tagging algorithms. In particular, DeepCSV which was used for in a previous iteration of the $VH(H \rightarrow c\bar{c})$ resolved jet analysis performed on data taken by the CMS detector in 2016. However, even with stronger discrimination powers it can often time be disadvantageous to tag jets based purely on the raw c -tagging discriminator values. For multiclassifiers, it is often best to judge a discriminator value with respect to the discriminator values assigned to other jet flavors. This leads to a consistent normalization of the newly calculated discriminator value as well as the ability to tag a jet based on the relative performance of subsets of discriminator outputs. For c -tagging the closeness of c -jets to both b -jets and uds -jets makes it advantageous to identify c -jets based on two separate relative discriminator ratios. In particular, the $VH(H \rightarrow c\bar{c})$ resolved analysis performs c -tagging based on the values:

$$\text{CvsL} = \frac{P(c)}{P(c) + P(uds)} \qquad \text{CvsB} = \frac{P(c)}{P(c) + P(b)} \qquad (5.3.1)$$

where $P(c)$ is the raw c -tagging discriminator output, $P(uds)$ is the sum of all light and gluon initiated jet discriminators ($l+g$ for DeepJet), and $P(b)$ is the sum of the probability of all b -tagging discriminators ($b+bb+b_\ell$ for DeepJet). This allows one to make c -tagging decisions on based on a jet's b -ness, c -ness, and light-ness. The distributions of the DeepJet and DeepCSV CvsL and CvsB scores for b -, c -, and uds -jets produced in simulated $t\bar{t}$ events can be found in Figure 5.3.6. Receiver operator characteristics (ROC) in which the mis-tag rate as a function of c -jet tagging efficiency for the CvsL and CvsB discriminators are shown in Figure 5.3.7.

5.4 Calibration of c -tagging scale factors

Due to the requirement to have labels when performing supervised learning, the DeepJet algorithm is trained using simulated data in which we know the true flavor of every particle in the ancestry of the final state particles. However, due to imperfections in our ability to simulate physics events

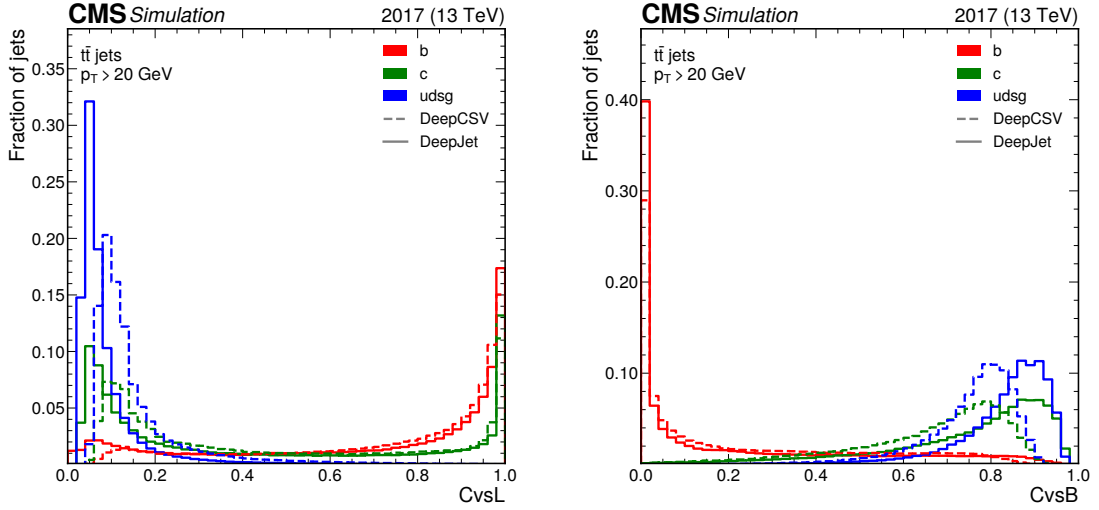


Figure 5.3.6: Comparison of the DeepCSV and DeepJet CvsL (left) and CvsB (right) distributions for b -, c -, and $udsg$ -jets with $p_T > 20$ GeV and $|\eta| < 2.5$. Jets generated using 2017 $t\bar{t}$ MC samples.

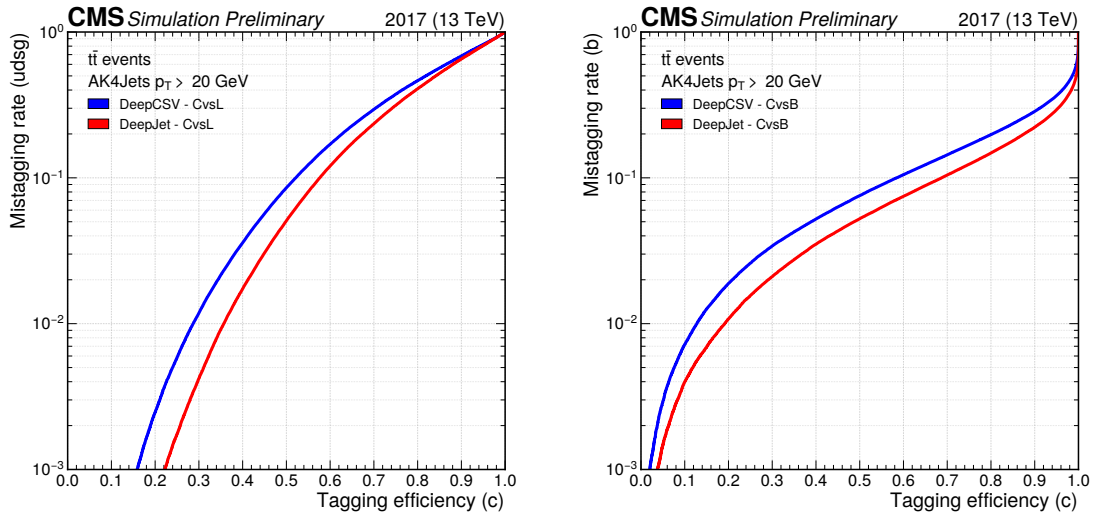


Figure 5.3.7: Receiver operator characteristics (ROC) showing the c -tagging mis-tag rate as a function of tagging efficiency for CvsL (left) and CvsB (right) discriminators.

systematic differences exist between the simulated physics events used to train the networks and the real physics data collected by the CMS detector. These differences can lead DeepJet, as well as many other tagging algorithms, to perform inconsistently when applied on data rather than simulated MC events. This generally leads to the algorithms over performing when applied on simulation with respect to their performance on data. If not accounted for, these changes in performance can make DeepJet unsuitable for a high precision statistical analysis. This is especially true for the $VH(H \rightarrow c\bar{c})$ resolved jet analysis strategy, explained in detail in Chapter 6, where the CvsL and CvsB scores are used as inputs for an multivariate analysis (MVA).

To account for the differences in DeepJet’s response when applied to simulation versus data, a set of corrections are generally applied to the simulated events when the tagger is used for analysis. The c -tagging corrections, often called scale factors, are designed to encapsulate the relative differences in how DeepJet scores the various jets flavors in data and in simulation. After applying the scale factors to a full set of simulation events, it is expected that the full distributions of CvsL and CvsB scores should agree between the simulated and data jets. For the $VH(H \rightarrow c\bar{c})$ resolved jet analysis, corrections to the simulation are applied as per-event weights which are obtained by first deriving a set of per-jet scale factors for c -, b -, and $udsg$ -jets. These scale factors represent the differences in tagger performance on the given jet flavor when applied on simulated events versus real data. For a given simulation event, the product of the individual per-jet scale factors is calculated and applied as an overall event weight. To obtain the per-jet scale factors, the following method is applied:

1. Create three distinct regions of phase space, each enriched in either b -, c -, or $udsg$ -jets.
2. In each flavor enriched phase space, construct a 2D histogram of the total number jets in bins of CvsL and CvsB. Produce such histograms for both data and simulated events.
3. Perform an iterative fit of the simulated events to the data in each of the constructed histogram bins for all the three phase space selections.
4. Extract three 2D profiles representing the fit values obtained for each CvsL and CvsB bin for the three different jet flavors as the scale factors.

Due to differences in the detector conditions and simulated samples used for each of the three data taking years, this procedure is carried out independently for each data taking year and a separate set of scale factors are derived. For all events considered, the derivation of scale factors are performed using AK4 jets with $p_T > 20$ GeV and $|\eta| < 2.5$. Additionally, a **tightLepVeto** (see Chapter 6 for more information) jet selection is applied to ensure that the jets are not misidentified high energy

lepton object which punched-through the ECAL. A full description of the procedures used to derive and apply the scale factors for the DeepCSV and DeepJet taggers can be found in [16].

5.4.1 Phase Space Selections

To properly calculate the scale factors, it is vital that we are able to construct phase spaces which are suitably enriched in a single flavor of jet. However, these phase spaces must be constructed without the use of jet-tagging algorithms and minimal use of jet-tagging related variables when selecting on jets and events. Otherwise, the scale factors may be biased by the same simulation mis-modeling in which they are made to correct for. As such, event based selections are used to isolate standard model processes known to have good modeling in their simulations as well as a large fraction of either b -, c -, or $udsg$ -jets. For this study, the following processes were used to obtain our jet enriched phase spaces: $t\bar{t}$ for b -jet enriched events, $W + c$ events for c -jet enriched events, and $DY+jet$ events for $udsg$ -jet enriched events. Once an event has been selected, a single jet will then be selected based on criteria designed to maximize the probability that the jet is of the desired flavor. If there are no jets in a selected event passing the jet selection criteria, the event is instead discarded.

5.4.1.1 c -enriched jets

The c -jet enriched phase space is obtained using jets from a selection of $W+jet$ events. This is because $W+jet$ events are produced from a quark which hard scatters during a collision radiating a W -boson. If this quark is a d or s quark, the W emission can cause it to flavor change into a c quark with high probability. These events are obtained by selecting on events in which a single leptonically decaying W boson is reconstructed. Additionally there must be 1–3 jets in the event, with the maximum cut of 3 placed to prevent contamination from $t\bar{t}$ events.

To select c -jets from the $W+jet$ events, only AK4 jets in which a reconstructed soft muon with opposite charge as the lepton used to reconstruct the W are selected. This is because of the 17.6% branching ratio of the $D^\pm \rightarrow \mu^\pm \nu$ decay mode as well as the high μ reconstruction efficiency of the CMS detector. The opposite charge selection is made because we expect that the c quark produced in conjunction with the W from $W+c$ events will always have the opposite charge as the W boson. The majority of the remaining background comes from jets produced from a radiated gluon in which the gluon decays as $g \rightarrow b\bar{b}/c\bar{c}$ and one of the c or b quarks decays semileptonically to a μ . However, the percentage of $g \rightarrow b\bar{b}/c\bar{c}$ jets in which the soft muon has the same sign as the W boson is 50%. That allows us to subtract out the distribution of jets in which the soft muon has the same sign as

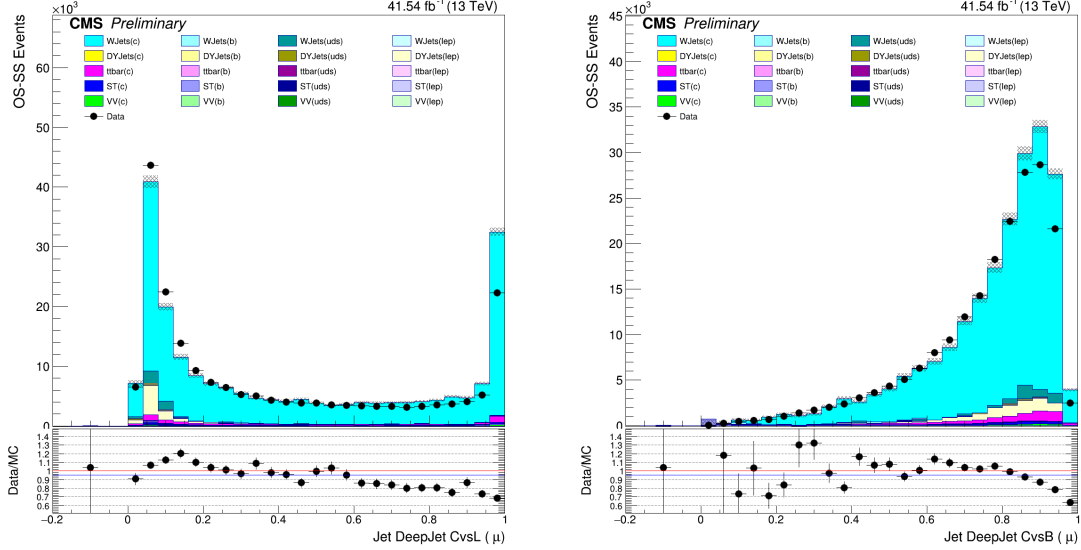


Figure 5.4.1: Distributions of CvsL (left) and CvsB (right) scores for soft muon containing jets obtained from the same-sign – opposite-sign $W+c$ event selection in both data and simulation. Generator level information from the simulated samples shows that the selection leads to a phase space heavily enriched in c -jets. Figures taken from [16].

the W from the jets in which the soft muon has the opposite sign as the W . As seen in Figure 5.4.1, the subtraction procedure leaves a high purity sample of c -jets originating from the hadronization of semileptonically decaying $D^\pm$ mesons. Despite the fact that we are only selecting a single type of charmed-meson topology, [16] shows that the derived scale factors can be extrapolated to other c -jet (and b -jet) topologies to obtain good agreement between simulation and data.

5.4.1.2 b -enriched jets

The b -jet enriched phase space is obtained from jets selected from $t\bar{t}$ events in which one or both of the tops decay semileptonically. $t\bar{t}$ events with a single semileptonically decaying top are selected by requiring the events have a reconstructed leptonically decaying W boson and at least 4 jets. $t\bar{t}$ events with both top quarks decaying semileptonically are selected by requiring that the event reconstructs two prompt high p_T leptons and has at least two jets. If the two leptons have the same flavor, a veto around the Z peak is placed on the mass of the dilepton system and a MET cut are placed to ensure that the event is not originating from a DY process.

Because B -mesons will also decay weakly, the same soft muon requirement used to select c -jets in the W +jet events can be made to differentiate b -jets from the other jets produced in the $t\bar{t}$ events. The highest p_T jet containing a soft muon is then selected from each event and used for scale factor derivation. Unlike the c -jet selection, no same-sign – opposite-sign subtraction is performed.

This is because the near 100% probability of a top quark decaying to a b -hadron and high $t\bar{t}$ event selection efficiency already leads to the the contamination of soft-muon containing jets coming from the $g \rightarrow b\bar{b}/c\bar{c}$ background being highly suppressed.

5.4.1.3 light-enriched jets

The $udsg$ -jet enriched phase space is composed of jets coming from DY events. DY events are identified by selecting on events with two oppositely signed and same flavor leptons. The invariant mass of the lepton pair must then be within the range of the Z peak. Once the DY events have been selected, the jet with the highest p_T is selected and used for scale factor derivation.

5.4.2 Iterative Fit

To calculate the scale factors, an iterative fit procedure is applied to all three flavors of jets with the aim of simultaneously fitting the distributions of the simulated jets to data jets in all three phase space regions. This is done by first creating a 2D binning in CvsL and CvsB that is common to all three phase spaces. The bin width is kept to a uniform value of 0.1 in CvsB for all bins. CvsL uses a variable bin width in the range $[0.02, 0.10]$ such that it is the minimum bin width in which the relative statistical uncertainty in the number of jets is 2% or less. In the instance in which the statistical uncertainty cannot be less than 2%, a bin width of 0.1 in CvsL is used. Three separate versions of the bins, one for each of the three phase space regions, are then filled with simulated and data jets as described in Section 5.4.1 and binned according to their DeepJet CvsL and CvsB scores. For the simulated jets, the distinction between c -, b -, and $udsg$ -jets are maintained so they can be used in the iterative fit procedure.

Once the bins have been filled a unique iterative fit is performed on each range of CvsB, such that the fit procedure is shared between all CvsL bins in all three phase spaces in said bin of CvsB. Before performing the iterative fit, a separate normalization is applied to the simulated jets in of the three phase spaces such that the total number of simulated jets for said phase space is equal to the number of jets found in data. The fit procedure is then applied to each bin in the following manner:

1. A set of three fit parameters, one for each jet flavor and representing a modification to the normalization in the number of jets of said flavor, are initialized to a value of 1 in each bin of CvsL.
2. The phase space with the highest percentage of signal jets (i.e. c -jets for W^+ jet, b -jets for $t\bar{t}$, and $udsg$ -jets for DY) in the given CvsB range is identified.

3. The fit parameter for the corresponding signal jets is modified (by a maximum of 0.005) separately for each bin of CvsL such that the ratio of the total number of simulated jets to the total number of data jets in said phase space and CvsL bin approaches unity.
4. The phase space with the second highest percentage of signal jets is identified, and the updated fit parameters are applied to each jet flavor. The fit parameter for the signal jet of the new phase space is then modified using the same methodology as step 3.
5. The phase space with the third highest percentage of signal jets is identified, and the updated fit parameters are applied to each jet flavor. The fit parameter for the signal jet of the new phase space is then modified using the same methodology as steps 3 and 4.
6. The fit parameters in all three phase spaces are updated and a χ^2 value of

$$\chi_{\text{PS}}^2 = \sum_i^n \frac{(N_{\text{sim},i} - N_{\text{data},i})^2}{\sigma_{\text{sim},i}^2 + \sigma_{\text{data},i}^2}$$

where n is the number of CvsL bins, $N_{\text{sim},i}$ is the total number of simulated jets in the i -th bin after fit parameter application, $N_{\text{data},i}$ is the number of data jets in i -th CvsL bin, $\sigma_{\text{sim},i}$ is the statistical uncertainty of the number of simulated jets in the i -th CvsL bin, and $\sigma_{\text{data},i}$ is the statistical uncertainty of the number of data jets in the i -th CvsL bin, is calculated individually for each phase space. The χ_{PS}^2 value of the three phase spaces are then summed together.

7. Steps 2–6 are repeated with the latest set of fit parameters until the total χ^2 value stops increasing, in which case the fit parameters of the previous iteration of fit parameter updates are extracted as the set of scale factors.

To account for systematic uncertainties affecting the simulated samples, the fit procedure is re-run multiple times with various uncertainties propagated to the simulated jet parameters. This is done one at a time for the up and down variation associated with each theoretical and detector mis-modeling based uncertainty and applied before jets are tagged or selected. A separate set of scale factors are then extracted for each uncertainty propagation, and saved separately from the nominal value to be used by researchers in physics analyses.

5.4.3 Scale factor application

After the fit procedures described in Section 5.4.2 have been concluded, the final set of fit parameters for each jet flavor and bin in CvsL and CvsB are extracted. Three 2D profiles of scale factors, one for each jet flavor, are then produced from the fit parameters in each bin of CvsB and CvsL. The 2D profiles of DeepJet scale factors derived for the 2017 data taking year can be found in Figure 5.4.2.

From the scale factors, event weights can be calculated by first taking each AK4 jet in an event passing the baseline selections and obtaining the scale factor for its corresponding jet flavor, CvsL score, and CvsB score. The c -tagging weight is then calculated as the product of all scale factors:

$$w_{c\text{-tag}} = \prod_{i=1}^{N_j} SF_i(\text{hadron flavor, CvsL, CvsB}) \quad (5.4.1)$$

where i is the i -th jet in the event, N_j is the total number of jets, and SF_i is the scale factor of the i -th jet in the event. Additionally, a separate set of weights can be calculated for each set of scale factors calculated after propagating the systematic uncertainties through the iterative fit procedure. When used for a physics analysis, a shape uncertainty can be obtained for each systematic by taking the ratio of the systematic-associated up and down variation c -tagging weights with respect to the nominal c -tagging event weight.

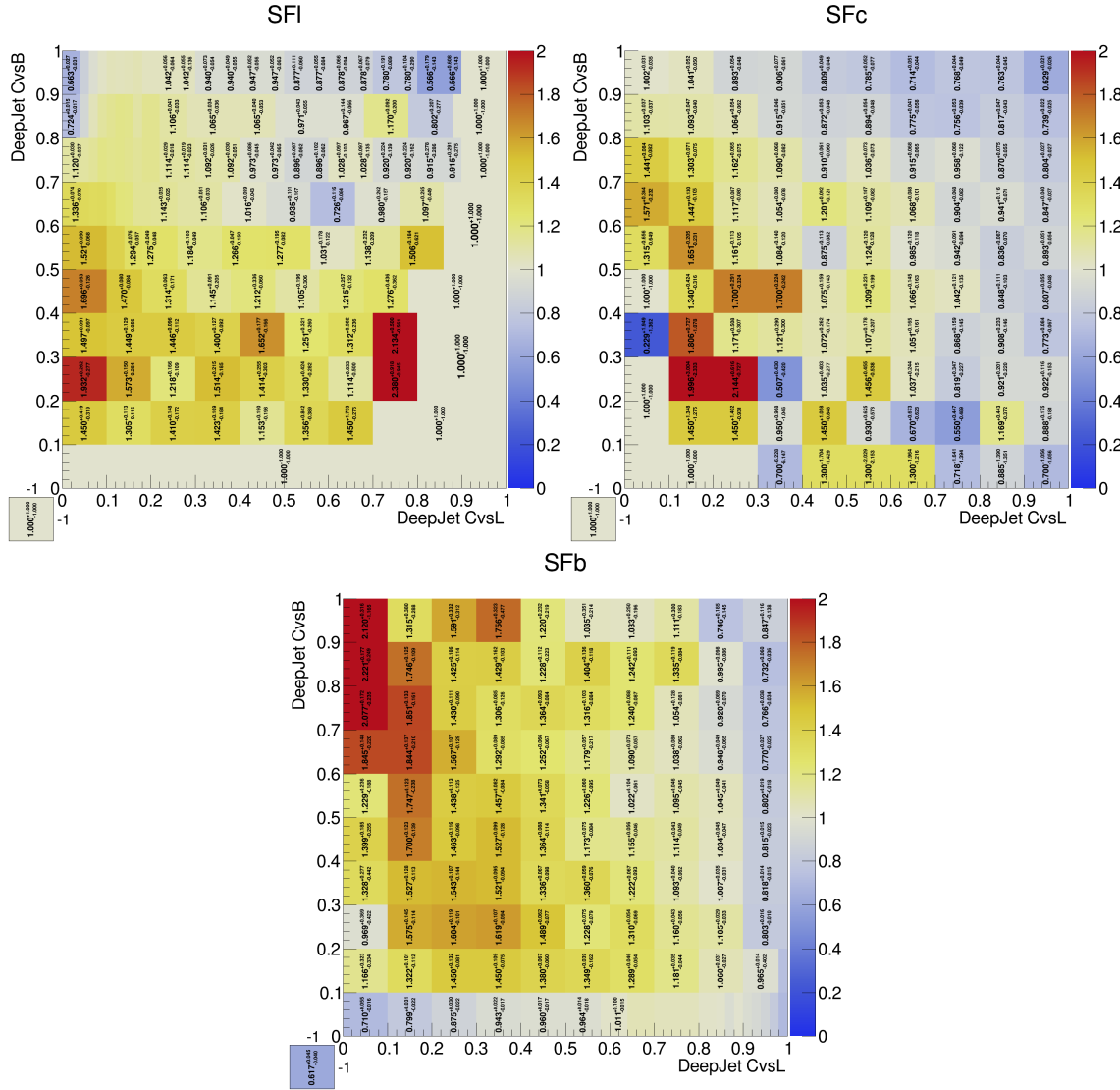


Figure 5.4.2: Full 2D profiles in CvsL and CvsB of DeepJet scale factors for 2017 data taking year. A separate set of scale factors are derived and applied for light-jets (top left), c -jets (top right), and b -jets (bottom).

Chapter 6

The search for $VH(H \rightarrow c\bar{c})$ at the CMS Detector with Full Run 2 Data

The $VH(H \rightarrow c\bar{c})$ analysis presented in this chapter is an extension of the resolved jet analysis outlined in [39] to the full Run 2 CMS dataset. Like [39], a search for the $VZ(Z \rightarrow c\bar{c})$ process is also performed as a means to validate the analysis strategy. In addition to extending the amount of data analyzed from 35.9 fb^{-1} to approximately 138 fb^{-1} , new additions to the analysis strategy are implemented including improvements in the simulation of dominant standard model backgrounds, moving from DeepCSV to DeepJet to perform c -tagging, the inclusion of a kinematic fit for events with a Z decaying to two leptons, and the inclusion of a c -jet energy regression algorithm when reconstructing the Higgs boson. The analysis was primarily carried out by a team of four researchers. For completeness, description of all studies which were part of the overall analysis are presented in this thesis. However, it is explicitly stated in their description when I did not take part in performing specific studies.

As in [39], the $VH(H \rightarrow c\bar{c})$ and $VZ(Z \rightarrow c\bar{c})$ resolved jet analyses are combined with an orthogonal analysis tuned to a high $p_T(H)$ merged jet topology, which also includes many improvements to its previous analysis strategy. The final combination of datacards, extraction of the combined analysis results, and calculation of the charm-Yukawa coupling were performed by the merged jet analysis team. But for completeness I will be presenting the results of the combination on their behalf in this thesis. Compared to the 2016 analysis, the orthogonality cut has been further optimized to reduce the chance that signal events are unintentionally discarded by both topologies. Additionally, object identification and scale factor application has been made uniform across the two topologies wherever

possible to maximize the number of systematic uncertainties which can be correlated during signal extraction. The combined $VH(H \rightarrow c\bar{c})$ analysis leads to the strongest limits on the $VH(H \rightarrow c\bar{c})$ signal strength and charm-Higgs Yukawa coupling to date. Additionally, the combined results of the $VZ(Z \rightarrow c\bar{c})$ analysis leads to the first observation of the $Z \rightarrow c\bar{c}$ decay at a hadron collider.

6.1 Analysis Overview

To exploit the trigger capabilities of the CMS detector and add phase space elements which can be exploited for improved signal separation, the $VH(H \rightarrow c\bar{c})$ analysis makes use of the VH production mode, shown in Figure 6.1.1, to search for Higgs boson production at the LHC. With the VH production mode, we can utilize the $Z \rightarrow \ell\ell$, $W \rightarrow \ell\nu$, and $Z \rightarrow \nu\nu$ leptonic decay modes of the V-bosons to exploit CMS's high efficiency electron, muon, and MET based triggers. Based on the set of triggers passed, we then attempt to reconstruct the Z- or W-boson (V-candidate) using the triggered and other physics objects in the event (Section 6.4). We then construct a multichannel analysis by categorizing and splitting events based on the V-candidate decay type, with the naming convention based on the number of charged leptons: 2-leptons ($Z(\ell\ell)H$), 1-lepton ($W(\ell\nu)H$) or 0-leptons ($Z(\nu\nu)H$). Events in the 2-lepton channel are further split into two categories based on the momentum of the Z-candidate. Additionally, distributions studied in the 2- and 1-lepton events are further split based on the lepton flavor, but the general treatments of analysis channels and applications of non-lepton based efficiencies are performed identically to both lepton flavors.

After identification of the V-candidate, at least two AK4 jets in the event are required for identification of a potential Higgs boson. We reconstruct a potential Higgs boson (H-candidate) by summing the 4-momentum of the two jets with the largest DeepJet CvsL scores (Section 6.5). We then apply a dedicated c -jet energy regression algorithm to the two jets used to reconstruct the H-candidate (also referred to as Higgs jets) to recover the energy of particles which either failed to be clustered into the jet or were lost from neutrinos being produced in potential D -meson decay chains. Additionally, a final state radiation (FSR) recovery procedure is performed to include the 4-momentum of soft jets coming from gluons radiated at high angles during the H-candidate decay chains such that they were not clustered in their AK4 jet cone. For the 2-lepton channel, a kinematic fit is also applied in which we use the invariant mass of the Z-pole and the conservation of p_T in the collision are used to better constrain the kinematics of the H- and V-candidates.

To measure the $VH(H \rightarrow c\bar{c})$ cross section, Monte Carlo (MC) simulations are used to simulate proton-proton collisions taking place inside of the CMS detector (Section 6.3). The real physics

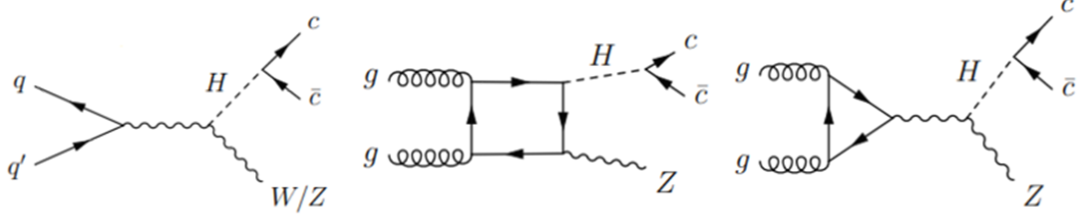


Figure 6.1.1: The Feynman diagrams for the three dominant $VH(H \rightarrow c\bar{c})$ production modes in proton-proton collisions.

data is then compared to the expected standard model event yields obtained from the simulated MC events. To maximize the search potential of the analysis, a dedicated boosted decision tree (BDT) is used to identify regions of an abstract variable-space which are expected to be highly dominated by signal-like events (Section 6.7). BDTs are trained to discriminate the $VH(H \rightarrow c\bar{c})$ processes from SM backgrounds using a mix of c -tagging, H-candidate, V-candidate, and event based variables. By separating our analysis into channels based on the number of leptons, and in the 2-lepton channel $p_T(V)$, we are able to train independent BDTs for each channel using the set of variables best suited for separating the $VH(H \rightarrow c\bar{c})$ signal from said channels respective background processes.

To ensure that we are properly characterizing any mis-modeling in our background simulation, we construct multiple control regions to constrain $t\bar{t}$ background events as well as V +jet events based on the flavor of the jets used to reconstruct the H-candidate (Section 6.8). The $VH(H \rightarrow c\bar{c})$ signal strength is then extracted using a binned maximum likelihood fit made to a reshaped distribution of the BDT discriminator in $VH(H \rightarrow c\bar{c})$ enriched signal regions (Section 6.12). During the fit, the normalization of the $t\bar{t}$ background and flavor split V +jet backgrounds are left floating as rate parameters fit to the DeepJet CvsB distributions of the subleading Higgs jets. Due to the rarity of the $VH(H \rightarrow c\bar{c})$ process, a validation of the analysis procedure is also run in which $VH(H \rightarrow c\bar{c})$ events are considered a background and $VZ(Z \rightarrow c\bar{c})$ events are treated as signal (Section 6.9). The $VZ(Z \rightarrow c\bar{c})$ process is identified in the same way as the $VH(H \rightarrow c\bar{c})$ process, except that the two c -tagged jets are now used to reconstruct a Z-candidate decaying to charm quarks. When measuring the $VZ(Z \rightarrow c\bar{c})$ production rate, the $VH(H \rightarrow c\bar{c})$ process is then treated as a background and the $VZ(Z \rightarrow c\bar{c})$ signal strength is made to fit to its best-fit value. This is performed using a separate set of BDTs trained to discriminate $VZ(Z \rightarrow c\bar{c})$ events and SM background sources.

To maximize the CMS collaboration's ability to search for $VH(H \rightarrow c\bar{c})$ (and $VZ(Z \rightarrow c\bar{c})$) events, the resolved jet $VH(H \rightarrow c\bar{c})$ analysis is combined with a separate $VH(H \rightarrow c\bar{c})$ analysis tuned for identifying events where a high p_T Higgs boson decays to two c -jets which merge to form a

single large cone jet (Section 6.10). For combination with the merged jet analysis, an orthogonality condition is first imposed by requiring that all events containing a jet reconstructed using the anti- k_T algorithm with a radius parameter of 1.5 (AK15 jet) with $p_T > 300$ GeV are selected for the merged analysis, and all others are selected for the resolved jet analysis. Signal extraction is then performed via a simultaneous frequentist fit to both the resolved and merged analysis signal and control regions with all experimental and theoretical systematics correlated between the two analyses.

6.2 Simulations and Datasets

This analysis is performed using a specialized version of the NanoAODv7 [80] files centrally produced by the CMS collaboration. In these files, all information about the DeepJet c -taggers and the AK15 jets are added. These specialized NanoAODv7 files are then run through the post-processing steps recommended by the CMS collaboration [80]. The processing of physics events relating to the 2016 data taking year as well as a portion of events related to the 2017 data taking year were performed by collaborators.

6.2.1 Data

The analysis is performed on data taking in the 2016, 2017, and 2018 data taking periods with the conditions described in Chapter 3 and total integrated luminosity of approximately 138 fb^{-1} . Events were certified against an internal database [81, 82, 83] to filter out any potentially bad events from the NanoAOD files. For each of the analysis channels a different the following datasets/HLT paths were used:

- **2-lepton** – Data in the $Z(ee)$ and $Z(\mu\mu)$ 2-lepton channels are selected using DoubleMuon and DoubleEG triggers where two high p_T muons or EGamma objects are reconstructed.
- **1-lepton** – Data in the $W(e\nu)$ and $W(\mu\nu)$ 1-lepton channels are selected using SingleMuon and SingleElectron triggers where a single high p_T muon or Egamma object is reconstructed.
- **0-lepton** – Data in the $Z(\nu\nu)$ 0-lepton channel is selected using MET and MHT triggers where a large amount of missing p_T is found from the ECAL+HCAL or HCAL only.

The full list of CMS datasets and HLT paths used for each of the three data taking years can be found in Table 6.1 and Table 6.2, where the HLT paths contain the selection criteria passed by the triggered objects.

Primary dataset	HLT path
2016	
/MET/Run2016B-17Jul2018_ver2-v1/MINIAOD /MET/Run2016C-17Jul2018-v1/MINIAOD /MET/Run2016D-17Jul2018-v1/MINIAOD /MET/Run2016E-17Jul2018-v1/MINIAOD /MET/Run2016F-17Jul2018-v1/MINIAOD /MET/Run2016G-17Jul2018-v1/MINIAOD /MET/Run2016H-17Jul2018-v2/MINIAOD	HLT_PFMET110_PFMHT110_IDTight HLT_PFMET170_NoiseCleaned HLT_PFMET170_BeamHaloCleaned HLT_PFMET170_HBHECleaned
/SingleElectron/Run2016B-17Jul2018_ver2-v1/MINIAOD /SingleElectron/Run2016C-17Jul2018-v1/MINIAOD /SingleElectron/Run2016D-17Jul2018-v1/MINIAOD /SingleElectron/Run2016E-17Jul2018-v1/MINIAOD /SingleElectron/Run2016F-17Jul2018-v1/MINIAOD /SingleElectron/Run2016G-17Jul2018-v1/MINIAOD /SingleElectron/Run2016H-17Jul2018-v2/MINIAOD	HLT_Ele27_eta2p1_WPTight_Gsf
/SingleMuon/Run2016B-17Jul2018_ver2-v1/MINIAOD /SingleMuon/Run2016C-17Jul2018-v1/MINIAOD /SingleMuon/Run2016D-17Jul2018-v1/MINIAOD /SingleMuon/Run2016E-17Jul2018-v1/MINIAOD /SingleMuon/Run2016F-17Jul2018-v1/MINIAOD /SingleMuon/Run2016G-17Jul2018-v1/MINIAOD /SingleMuon/Run2016H-17Jul2018-v2/MINIAOD	HLT_IsoMu24 HLT_IsoTkMu24
/DoubleElectron/Run2016B-17Jul2018_ver2-v1/MINIAOD /DoubleElectron/Run2016C-17Jul2018-v1/MINIAOD /DoubleElectron/Run2016D-17Jul2018-v1/MINIAOD /DoubleElectron/Run2016E-17Jul2018-v1/MINIAOD /DoubleElectron/Run2016F-17Jul2018-v1/MINIAOD /DoubleElectron/Run2016G-17Jul2018-v1/MINIAOD /DoubleElectron/Run2016H-17Jul2018-v2/MINIAOD	HLT_Ele23_Ele12_CaloIdL_TrackIdL_IsoVL_DZ
/DoubleMuon/Run2016B-17Jul2018_ver2-v1/MINIAOD /DoubleMuon/Run2016C-17Jul2018-v1/MINIAOD /DoubleMuon/Run2016D-17Jul2018-v1/MINIAOD /DoubleMuon/Run2016E-17Jul2018-v1/MINIAOD /DoubleMuon/Run2016F-17Jul2018-v1/MINIAOD /DoubleMuon/Run2016G-17Jul2018-v1/MINIAOD /DoubleMuon/Run2016H-17Jul2018-v2/MINIAOD	HLT_Mu17_TrkIsoVVL_Mu8_TrkIsoVVL HLT_Mu17_TrkIsoVVL_Mu8_TrkIsoVVL_DZ HLT_Mu17_TrkIsoVVL_TkMu8_TrkIsoVVL HLT_Mu17_TrkIsoVVL_TkMu8_TrkIsoVVL_DZ

Table 6.1: 2016 datasets used for the analysis and HLT paths used for their corresponding event triggers.

Primary dataset	HLT path
2017	
/MET/Run2017B-31Mar2018-v1/MINIAOD /MET/Run2017C-31Mar2018-v1/MINIAOD /MET/Run2017D-31Mar2018-v1/MINIAOD /MET/Run2017E-31Mar2018-v1/MINIAOD /MET/Run2017F-31Mar2018-v1/MINIAOD	HLT_PFMET110_PFMHT110_IDTight HLT_PFMET120_PFMHT120_IDTight HLT_PFMET120_PFMHT120_IDTight_PFHT60 HLT_PFMETTypeOne120_PFMHT120_IDTight
/SingleElectron/Run2017B-31Mar2018-v1/MINIAOD /SingleElectron/Run2017C-31Mar2018-v1/MINIAOD /SingleElectron/Run2017D-31Mar2018-v1/MINIAOD /SingleElectron/Run2017E-31Mar2018-v1/MINIAOD /SingleElectron/Run2017F-31Mar2018-v1/MINIAOD	HLT_Ele32_WPTight_Gsf_L1DoubleEG HLT_Ele32_WPTight_Gsf
/SingleMuon/Run2017B-31Mar2018-v1/MINIAOD /SingleMuon/Run2017C-31Mar2018-v1/MINIAOD /SingleMuon/Run2017D-31Mar2018-v1/MINIAOD /SingleMuon/Run2017E-31Mar2018-v1/MINIAOD /SingleMuon/Run2017F-31Mar2018-v1/MINIAOD	HLT_IsoMu24 HLT_IsoMu27
/DoubleEG/Run2017B-31Mar2018-v1/MINIAOD /DoubleEG/Run2017C-31Mar2018-v1/MINIAOD /DoubleEG/Run2017D-31Mar2018-v1/MINIAOD /DoubleEG/Run2017E-31Mar2018-v1/MINIAOD /DoubleEG/Run2017F-31Mar2018-v1/MINIAOD	HLT_Ele23_Ele12_CaloIdL_TrackIdL_IsoVL HLT_Ele23_Ele12_CaloIdL_TrackIdL_IsoVL_DZ
/DoubleMuon/Run2017B-31Mar2018-v1/MINIAOD /DoubleMuon/Run2017C-31Mar2018-v1/MINIAOD /DoubleMuon/Run2017D-31Mar2018-v1/MINIAOD /DoubleMuon/Run2017E-31Mar2018-v1/MINIAOD /DoubleMuon/Run2017F-31Mar2018-v1/MINIAOD	HLT_Mu17_TrkIsoVVL_Mu8_TrkIsoVVL HLT_Mu17_TrkIsoVVL_Mu8_TrkIsoVVL_DZ HLT_Mu17_TrkIsoVVL_Mu8_TrkIsoVVL_DZ_Mass3p8 HLT_Mu17_TrkIsoVVL_Mu8_TrkIsoVVL_DZ_Mass8
2018	
/MET/Run2018A-17Sep2018-v1/MINIAOD /MET/Run2018B-17Sep2018-v1/MINIAOD /MET/Run2018C-17Sep2018-v1/MINIAOD /MET/Run2018D-PromptReco-v2/MINIAOD	HLT_PFMET110_PFMHT110_IDTight HLT_PFMET120_PFMHT120_IDTight HLT_PFMET120_PFMHT120_IDTight_PFHT60
/EGamma/Run2018A-17Sep2018-v2/MINIAOD /EGamma/Run2018B-17Sep2018-v1/MINIAOD /EGamma/Run2018C-17Sep2018-v1/MINIAOD /EGamma/Run2018D-22Jan2019-v2/MINIAOD	W(ev)H: HLT_Ele32_WPTight_Gsf_L1DoubleEG HLT_Ele32_WPTight_Gsf Z(ee)H: HLT_Ele23_Ele12_CaloIdL_TrackIdL_IsoVL
/SingleMuon/Run2018A-17Sep2018-v2/MINIAOD /SingleMuon/Run2018B-17Sep2018-v1/MINIAOD /SingleMuon/Run2018C-17Sep2018-v1/MINIAOD /SingleMuon/Run2018D-22Jan2019-v2/MINIAOD	HLT_IsoMu24 HLT_IsoMu27
/DoubleMuon/Run2018A-17Sep2018-v2/MINIAOD /DoubleMuon/Run2018B-17Sep2018-v1/MINIAOD /DoubleMuon/Run2018C-17Sep2018-v1/MINIAOD /DoubleMuon/Run2018D-22Jan2019-v2/MINIAOD	HLT_Mu17_TrkIsoVVL_Mu8_TrkIsoVVL HLT_Mu17_TrkIsoVVL_Mu8_TrkIsoVVL_DZ HLT_Mu17_TrkIsoVVL_Mu8_TrkIsoVVL_DZ_Mass3p8 HLT_Mu17_TrkIsoVVL_Mu8_TrkIsoVVL_DZ_Mass8

Table 6.2: 2017 and 2018 datasets used for the analysis and HLT paths used for their corresponding event triggers.

6.2.2 Monte Carlo Samples

Monte Carlo (MC) simulation is used to model proton-proton collisions in the conditions of the 2016, 2017, and 2018 data taking periods, as well as the response of the CMS detector during each of the years. The simulated events are each assigned a weight, such that the weighted event distribution of an infinite number of simulated events would match the true phase space predicted by phenomenological calculation. The weighted event counts of each process are then normalized with respect to an externally supplied cross section and the luminosity of the corresponding data taking year, making the weighted event count equal to the number of expected events for the given physics process, cross section, and luminosity.

The specific generator used when simulating physics processes varies depending on the process. However, all simulated events use GEANT4 [63] to simulate the full response of the CMS detector for the data taking year they are made to simulate. When modeling the 2016 (2017–2018) data taking year, the NLO NNPDF3.0 [84] (NNLO NNPDF3.1 [3]) parton distribution functions (PDF) are used for simulating collisions. Additionally, showering and hadronization is simulated using PYTHIA v8.230 [85] with the CUETP8M1 [86] (CP5 [87]) event tunes. The matching of partons produced via the generator to jets produced from PYTHIA is performed using the FxFx [88] scheme for samples calculated at next-to-leading order (NLO), and MLM [89] scheme for samples calculated at leading order (LO). To model PU, additional collisions are simulated and added to the events. The events are then re-weighted such that the distribution of PU events matches the PU distributions observed in data for the corresponding data taking year.

To properly model the effects of the triggering on data, simulated events are run through the same HLT trigger paths as those described in Table 6.1 and Table 6.2 before being used for analysis. Additional weights (some times called scale factors) are then applied to the MC events as a function(s) of $p_T(e)$, $p_T(\mu)$, or MET (based on the HLT path) to match the trigger efficiency observed in data. The set of trigger scale factors used from this study were obtained from a set of studies performed by the VH($H \rightarrow b\bar{b}$) analysis team [90], in which the same set of triggers, lepton Ids, and baseline event selections were used.

6.2.2.1 Signal Samples

The analysis uses signal samples in which a Higgs boson is produced in conjunction with a Z- or W-boson. Events in which the V-boson is produced from quark fusion are generated at NLO using the POWHEG v2 [91, 92, 93] event generator and improved using the MiNLO procedure [94, 95].

Signal events in which the ZH process is produced via gluon-gluon interaction (ggZH) are produced at LO using POWHEG v2. The $VH(H \rightarrow c\bar{c})$ and $VH(H \rightarrow b\bar{b})$ (used as a background) CMS MC samples used for the analysis as well as the cross sections [17] used for event count normalization can be found in Table 6.3 and 6.4. Beyond what is listed in the table, additional corrections to signal event cross sections are performed as described in Section 6.2.3.1.

In the case of the $VZ(Z \rightarrow c\bar{c})$ validation analysis, the signal is modeled using simulated diboson events in which a Z-boson decaying to a pair of c -jets is found using generator level information. The second boson is then required to decay as $Z \rightarrow \ell\ell$, $W \rightarrow \ell\nu$, or $Z \rightarrow \nu\nu$ depending on the analysis channel. The list of simulated diboson process can be found in Table 6.5–6.7.

Dataset	$\sigma \times \mathcal{B}$ [pb]
/WplusH_HTtoCC_WToLNu_M125_13TeV_powheg_pythia8/ proc /MINIAODSIM	0.008540
/WminusH_HTtoCC_WToLNu_M125_13TeV_powheg_pythia8/ proc /MINIAODSIM	0.005411
/ZH_HTtoCC_ZToLL_M125_13TeV_powheg_pythia8/ proc /MINIAODSIM	0.002342
/ggZH_HTtoCC_ZToLL_M125_13TeV_powheg_pythia8/ proc /MINIAODSIM	0.000357
/ZH_HTtoCC_ZToNuNu_M125_13TeV_powheg_pythia8/ proc /MINIAODSIM	0.004628
/ggZH_HTtoCC_ZToNuNu_M125_13TeV_powheg_pythia8/ proc /MINIAODSIM	0.000713

Table 6.3: CMS datasets names used to obtain the $VH(H \rightarrow c\bar{c})$ simulated events for all three data taking years. Cross sections obtained from [17].

Dataset	$\sigma \times \mathcal{B}$ [pb]
/WplusH_HTtoBB_WToLNu_M125_13TeV_powheg_pythia8/ proc /MINIAODSIM	0.17202
/WminusH_HTtoBB_WToLNu_M125_13TeV_powheg_pythia8/ proc /MINIAODSIM	0.10899
/ZH_HTtoBB_ZToLL_M125_13TeV_powheg_pythia8/ proc /MINIAODSIM	0.04718
/ggZH_HTtoBB_ZToLL_M125_13TeV_powheg_pythia8/ proc /MINIAODSIM	0.00720
/ZH_HTtoBB_ZToNuNu_M125_13TeV_powheg_pythia8/ proc /MINIAODSIM	0.09322
/ggZH_HTtoBB_ZToNuNu_M125_13TeV_powheg_pythia8/ proc /MINIAODSIM	0.01437

Table 6.4: CMS datasets names used to obtain the $VH(H \rightarrow b\bar{b})$ simulated events for all three data taking years. Cross sections obtained from [17].

6.2.2.2 Background Samples

We expect standard model background events to emulate the leptonic decays of the V-bosons (with the number of leptons and neutrinos depending on the channel) as well as the jet topology of a dijet Higgs decay. The most common standard model processes which can display these signatures are events in which a vector boson is produced with additional jets (V +jets), a top quark-antiquark pair are produced ($t\bar{t}$), a single top quark is produced in association with lighter quarks (Single Top), or two V-bosons are produced (diboson). Furthermore, VH events in which the Higgs boson decays to b -jets ($VH(H \rightarrow b\bar{b})$) arise as a source of background events due to the possibility of wrongly

identifying b -jets as c -jets and the ambiguity in Higgs decay mode when the jets hadronize to form charmed-bottom mesons. MC simulations of these processes are generated and used to measure the contamination of standard model backgrounds in $VH(H \rightarrow c\bar{c})$ enriched phase space.

V+jet samples are generated at NLO using MADGRAPH5_aMC@NLO v2.6.0 [96] where we consider events in which the V-boson is produced with one or two partons. Generation was done in groups based on flavor of the V-boson, its decay mode, the generated p_T of the V-boson, and the number of partons produced during collision (2017 and 2018 only). Generation of the Diboson samples were performed in groups based on the flavor of the two bosons as well as their final decay states. The WW samples and ZZ samples where the Zs decay to charged leptons or neutrinos were generated at NLO using POWHEG v2. WZ as well as ZZ samples where one Z decays to charged leptons or neutrinos and the other decays to quarks were generated at NLO using MADGRAPH5_aMC@NLO v2.6.0. For both the V+jet and Diboson samples, additional corrections to the differential cross sections of the samples were applied, as described in detail in Section 6.2.3.2. The $t\bar{t}$ and Single Top processes are both generated at NLO using POWHEG v2. However, the cross sections of the $t\bar{t}$ samples are scaled to next-to-next-to-leading-log (NNLL) using TOP++ v2 [97], and the differential cross section as a function of top quark p_T is corrected to next-to-next-to-leading order (NNLO) in QCD and NLO in EWK predictions based on [98]. The full list of CMS MC samples used to simulate the SM background of the analysis as well as the cross sections used for normalization can be found in Table 6.5–6.7.

6.2.3 Corrections to Differential Cross Sections of Simulated Processes

For some samples, phenomenological calculations exist at higher precision than what was used for MC generation. In the simplest cases, these can be applied by adjusting the cross section (σ) used to normalize the process’s weighted event count. However, for some samples corrections to the differential cross sections are applied as a function of generator level event kinematics. These corrections are then applied as an additional per-event weight to each event in a corresponding processes.

6.2.3.1 NLO in Electroweak Corrections to VH Differential Cross Section

As described in Section 6.2.2.1 the MC samples used to simulate the VH production via quark fusion production mode are produced at NLO and improved upon using the MiNLO procedure. A secondary correction is additionally applied, in which the production cross section is scaled to be in

Process	Dataset	σ [pb]
$W \rightarrow \ell\nu + \text{jets}$	/WJetsToLNu_Pt-50To100_TuneCUETP8M1_13TeV-amcatnloFXFX-pythia8/ proc /MINIAODSIM	3298.37
	/WJetsToLNu_Pt-100To250_TuneCUETP8M1_13TeV-amcatnloFXFX-pythia8/ proc /MINIAODSIM	689.75
	/WJetsToLNu_Pt-250To400_TuneCUETP8M1_13TeV-amcatnloFXFX-pythia8/ proc /MINIAODSIM	24.507
	/WJetsToLNu_Pt-400To600_TuneCUETP8M1_13TeV-amcatnloFXFX-pythia8/ proc /MINIAODSIM	3.1101
	/WJetsToLNu_Pt-600ToInf_TuneCUETP8M1_13TeV-amcatnloFXFX-pythia8/ proc /MINIAODSIM	0.46832
$Z \rightarrow \ell\ell + \text{jets}$	/DYJetsToLL_M-50_TuneCUETP8M1_13TeV-amcatnloFXFX-pythia8/ proc /MINIAODSIM	6508.0
	/DYJetsToLL_Pt-50To100_TuneCUETP8M1_13TeV-amcatnloFXFX-pythia8/ proc /MINIAODSIM	363.81
	/DYJetsToLL_Pt-100To250_TuneCUETP8M1_13TeV-amcatnloFXFX-pythia8/ proc /MINIAODSIM	84.015
	/DYJetsToLL_Pt-250To400_TuneCUETP8M1_13TeV-amcatnloFXFX-pythia8/ proc /MINIAODSIM	3.2283
	/DYJetsToLL_Pt-400To650_TuneCUETP8M1_13TeV-amcatnloFXFX-pythia8/ proc /MINIAODSIM	0.43604
$Z \rightarrow \nu\nu + \text{jets}$	/DYJetsToNuNu_PtZ-50To100_TuneCUETP8M1_13TeV-amcatnloFXFX-pythia8/ proc /MINIAODSIM	688.92
	/DYJetsToNuNu_PtZ-100To250_TuneCUETP8M1_13TeV-amcatnloFXFX-pythia8/ proc /MINIAODSIM	159.95
	/DYJetsToNuNu_PtZ-250To400_TuneCUETP8M1_13TeV-amcatnloFXFX-pythia8/ proc /MINIAODSIM	6.0386
	/DYJetsToNuNu_PtZ-400To650_TuneCUETP8M1_13TeV-amcatnloFXFX-pythia8/ proc /MINIAODSIM	0.81127
	/DYJetsToNuNu_PtZ-650ToInf_TuneCUETP8M1_13TeV-amcatnloFXFX-pythia8/ proc /MINIAODSIM	0.075855
$t\bar{t}, 2\ell$	/TTTo2L2Nu_TuneCP5_PSweights_13TeV-powheg-pythia8/ proc /MINIAODSIM	88.29
$t\bar{t}, 1\ell$	/TTToSemiLeptonic_TuneCP5_PSweights_13TeV-powheg-pythia8/ proc /MINIAODSIM	365.34
Single top	/ST_tW_top_5f_inclusiveDecays_TuneCP5_PSweights_13TeV-powheg-pythia8/ proc /MINIAODSIM	35.85
	/ST_tW_antitop_5f_inclusiveDecays_TuneCP5_PSweights_13TeV-powheg-pythia8/ proc /MINIAODSIM	35.85
	/ST_t-channel_top_4f_InclusiveDecays_TuneCP5_PSweights_13TeV-powheg-pythia8/ proc /MINIAODSIM	136.02
	/ST_t-channel_antitop_4f_InclusiveDecays_TuneCP5_PSweights_13TeV-powheg-pythia8/ proc /MINIAODSIM	80.95
	/ST_s-channel_4f_leptonDecays_TuneCP5_PSweights_13TeV-amcatnlo-pythia8/ proc /MINIAODSIM	3.36
Diboson	/WWTo2L2Nu_13TeV-powheg/ proc /MINIAODSIM	12.178
	/WWToLNuQQ_13TeV-powheg/ proc /MINIAODSIM	49.997
	/WZTo1L1Nu2Q(_TuneCUETP8M1) _13TeV _amcatnloFXFX _madspin _pythia8/ proc /MINIAODSIM	10.71
	/WZTo1L3Nu_13TeV _amcatnloFXFX _madspin _pythia8/ proc /MINIAODSIM	3.06
	/WZTo2L2Q_13TeV _amcatnloFXFX _madspin _pythia8/ proc /MINIAODSIM	5.595
	/WZTo2Q2Nu_13TeV _amcatnloFXFX _madspin _pythia8/ proc /MINIAODSIM	6.317
	/WZTo3LNU _TuneCUETP8M1_13TeV-amcatnloFXFX-pythia8/ proc /MINIAODSIM	4.692
	/ZZTo2L2Nu_13TeV _powheg _pythia8/ proc /MINIAODSIM	0.564
	/ZZTo2L2Q(_TuneCUETP8M1) _13TeV _amcatnloFXFX _madspin _pythia8/ proc /MINIAODSIM	3.22
	/ZZTo2Q2Nu_13TeV _amcatnloFXFX _madspin _pythia8/ proc /MINIAODSIM	4.04
	/ZZTo4L_13TeV _powheg _pythia8/ proc /MINIAODSIM	1.256

Table 6.5: Processes used to model standard model background for 2016 data taking year. Events in the WZTo1L1Nu2Q, ZZTo2L2Q, and ZZTo2Q2Nu processes in which generator level information showed that a Z-boson decayed to a pair of c-jets was used to model the signal in the VZ($Z \rightarrow c\bar{c}$) validation analysis.

Process	Dataset	σ [pb]
$W (\rightarrow \ell \nu) + \text{jets}$	/W1JetsToLNu_LHEWpT_50-150_TuneCP5_13TeV-amcatnloFXFX-pythia8/proc/MINIAODSIM	2661.0
	/W1JetsToLNu_LHEWpT_100-150_TuneCP5_13TeV-amcatnloFXFX-pythia8/proc/MINIAODSIM	286.1
	/W1JetsToLNu_LHEWpT_150-250_TuneCP5_13TeV-amcatnloFXFX-pythia8/proc/MINIAODSIM	71.9
	/W1JetsToLNu_LHEWpT_250-400_TuneCP5_13TeV-amcatnloFXFX-pythia8/proc/MINIAODSIM	8.05
	/W1JetsToLNu_LHEWpT_400-inf_TuneCP5_13TeV-amcatnloFXFX-pythia8/proc/MINIAODSIM	0.885
	/W2JetsToLNu_LHEWpT_50-150_TuneCP5_13TeV-amcatnloFXFX-pythia8/proc/MINIAODSIM	1331.0
	/W2JetsToLNu_LHEWpT_100-150_TuneCP5_13TeV-amcatnloFXFX-pythia8/proc/MINIAODSIM	277.7
	/W2JetsToLNu_LHEWpT_150-250_TuneCP5_13TeV-amcatnloFXFX-pythia8/proc/MINIAODSIM	105.9
	/W2JetsToLNu_LHEWpT_250-400_TuneCP5_13TeV-amcatnloFXFX-pythia8/proc/MINIAODSIM	18.67
	/W2JetsToLNu_LHEWpT_400-inf_TuneCP5_13TeV-amcatnloFXFX-pythia8/proc/MINIAODSIM	3.037
$Z (\rightarrow \ell \ell) + \text{jets}$	/DY1JetsToLL_M-50_TuneCP5_13TeV-amcatnloFXFX-pythia8/proc/MINIAODSIM	6508.0
	/DY1JetsToLL_M-50_LHEZpT_50-150_TuneCP5_13TeV-amcatnloFXFX-pythia8/proc/MINIAODSIM	316.6
	/DY1JetsToLL_M-50_LHEZpT_150-250_TuneCP5_13TeV-amcatnloFXFX-pythia8/proc/MINIAODSIM	9.543
	/DY1JetsToLL_M-50_LHEZpT_250-400_TuneCP5_13TeV-amcatnloFXFX-pythia8/proc/MINIAODSIM	1.098
	/DY1JetsToLL_M-50_LHEZpT_400-inf_TuneCP5_13TeV-amcatnloFXFX-pythia8/proc/MINIAODSIM	0.1193
	/DY2JetsToLL_M-50_LHEZpT_50-150_TuneCP5_13TeV-amcatnloFXFX-pythia8/proc/MINIAODSIM	169.6
	/DY2JetsToLL_M-50_LHEZpT_150-250_TuneCP5_13TeV-amcatnloFXFX-pythia8/proc/MINIAODSIM	15.65
	/DY2JetsToLL_M-50_LHEZpT_250-400_TuneCP5_13TeV-amcatnloFXFX-pythia8/proc/MINIAODSIM	2.737
	/DY2JetsToLL_M-50_LHEZpT_400-inf_TuneCP5_13TeV-amcatnloFXFX-pythia8/proc/MINIAODSIM	0.447
$Z (\rightarrow \nu \nu) + \text{jets}$	/Z1JetsToNuNu_M-50_LHEZpT_50-150_TuneCP5_13TeV-amcatnloFXFX-pythia8/proc/MINIAODSIM	596.3
	/Z1JetsToNuNu_M-50_LHEZpT_150-250_TuneCP5_13TeV-amcatnloFXFX-pythia8/proc/MINIAODSIM	17.98
	/Z1JetsToNuNu_M-50_LHEZpT_250-400_TuneCP5_13TeV-amcatnloFXFX-pythia8/proc/MINIAODSIM	2.057
	/Z1JetsToNuNu_M-50_LHEZpT_400-inf_TuneCP5_13TeV-amcatnloFXFX-pythia8/proc/MINIAODSIM	0.224
	/Z2JetsToNuNu_M-50_LHEZpT_50-150_TuneCP5_13TeV-amcatnloFXFX-pythia8/proc/MINIAODSIM	325.7
	/Z2JetsToNuNu_M-50_LHEZpT_150-250_TuneCP5_13TeV-amcatnloFXFX-pythia8/proc/MINIAODSIM	29.76
	/Z2JetsToNuNu_M-50_LHEZpT_250-400_TuneCP5_13TeV-amcatnloFXFX-pythia8/proc/MINIAODSIM	5.166
	/Z2JetsToNuNu_M-50_LHEZpT_400-inf_TuneCP5_13TeV-amcatnloFXFX-pythia8/proc/MINIAODSIM	0.8457
$t\bar{t}, 2\ell$	/TTTo2L2Nu_TuneCP5(_PSweights)_13TeV-powheg-pythia8/proc/MINIAODSIM	88.29
$t\bar{t}, 1\ell$	/TTToSemileptonic_TuneCP5(_PSweights)_13TeV-powheg-pythia8/proc/MINIAODSIM	365.34
Single top	/ST_tW_top_5f_inclusiveDecays_TuneCP5(_PSweights)_13TeV-powheg-pythia8/proc/MINIAODSIM	35.85
	/ST_tW_antitop_5f_inclusiveDecays_TuneCP5(_PSweights)_13TeV-powheg-pythia8/proc/MINIAODSIM	35.85
	/ST_t-channel_top_4f_InclusiveDecays_TuneCP5_PSweights_13TeV-powheg-pythia8/proc/MINIAODSIM	136.02
	/ST_t-channel_antitop_4f_InclusiveDecays_TuneCP5_PSweights_13TeV-powheg-pythia8/proc/MINIAODSIM	80.95
	/ST_s-channel_top_leptonDecays_13TeV-PSweights_powheg-pythia8/proc/MINIAODSIM	2.0688
	/ST_s-channel_antitop_leptonDecays_13TeV-PSweights_powheg-pythia8/proc/MINIAODSIM	1.2934
Diboson	/WWTo2L2Nu_NNPDF31_TuneCP5_PSweights_13TeV-powheg-pythia8/proc/MINIAODSIM	11.08
	/WWToLNuQQ_NNPDF31_TuneCP5_PSweights_13TeV-powheg-pythia8/proc/MINIAODSIM	45.99
	/WZTo1L1Nu2Q(_TuneCP5)_13TeV_amcatnloFXFX_madspin_pythia8/proc/MINIAODSIM	11.66
	/WZTo1L3Nu_13TeV_amcatnloFXFX_madspin_pythia8/proc/MINIAODSIM	3.294
	/WZTo2L2Q_13TeV_amcatnloFXFX_madspin_pythia8/proc/MINIAODSIM	6.351
	/WZTo2Q2Nu_TuneCP5_13TeV_amcatnloFXFX_madspin_pythia8/proc/MINIAODSIM	6.317
	/WZTo3LNu_TuneCP5_13TeV-amcatnloFXFX-pythia8/proc/MINIAODSIM	5.062
	/ZZTo2L2Nu_13TeV_powheg_pythia8/proc/MINIAODSIM	0.6008
	/ZZTo2L2Q(_TuneCP5)_13TeV_amcatnloFXFX_madspin_pythia8/proc/MINIAODSIM	3.693
	/ZZTo2Q2Nu_TuneCP5_13TeV_amcatnloFXFX_madspin_pythia8/proc/MINIAODSIM	4.554
	/ZZTo4L_13TeV_powheg_pythia8/proc/MINIAODSIM	1.325

Table 6.6: Processes used to model standard model background for 2017 data taking year. Events in the WZTo1L1Nu2Q, ZZTo2L2Q, and ZZTo2Q2Nu processes in which generator level information showed that a Z-boson decayed to a pair of c -jets was used to model the signal in the VZ($Z \rightarrow c\bar{c}$) validation analysis.

Process	Dataset	σ [pb]
$W \rightarrow \ell\nu + \text{jets}$	/WJetsToLNu_Pt-50To100_TuneCP5_13TeV-amcatnloFXFX-pythia8/proc/MINIAODSIM	3570.0
	/WJetsToLNu_Pt-100To250_TuneCP5_13TeV-amcatnloFXFX-pythia8/proc/MINIAODSIM	770.8
	/WJetsToLNu_Pt-250To400_TuneCP5_13TeV-amcatnloFXFX-pythia8/proc/MINIAODSIM	28.06
	/WJetsToLNu_Pt-400To600_TuneCP5_13TeV-amcatnloFXFX-pythia8/proc/MINIAODSIM	3.591
	/WJetsToLNu_Pt-600ToInf_TuneCP5_13TeV-amcatnloFXFX-pythia8/proc/MINIAODSIM	0.5491
$Z \rightarrow \ell\ell + \text{jets}$	/DYJetsToLL_M-50_TuneCP5_13TeV-amcatnloFXFX-pythia8/proc/MINIAODSIM	6508.0
	/DY1JetsToLL_M-50_LHEZpT_50-150_TuneCP5_13TeV-amcatnloFXFX-pythia8/proc/MINIAODSIM	316.6
	/DY1JetsToLL_M-50_LHEZpT_150-250_TuneCP5_13TeV-amcatnloFXFX-pythia8/proc/MINIAODSIM	9.543
	/DY1JetsToLL_M-50_LHEZpT_250-400_TuneCP5_13TeV-amcatnloFXFX-pythia8/proc/MINIAODSIM	1.098
	/DY1JetsToLL_M-50_LHEZpT_400-inf_TuneCP5_13TeV-amcatnloFXFX-pythia8/proc/MINIAODSIM	0.1193
	/DY2JetsToLL_M-50_LHEZpT_50-150_TuneCP5_13TeV-amcatnloFXFX-pythia8/proc/MINIAODSIM	169.6
	/DY2JetsToLL_M-50_LHEZpT_150-250_TuneCP5_13TeV-amcatnloFXFX-pythia8/proc/MINIAODSIM	15.65
	/DY2JetsToLL_M-50_LHEZpT_250-400_TuneCP5_13TeV-amcatnloFXFX-pythia8/proc/MINIAODSIM	2.737
	/DY2JetsToLL_M-50_LHEZpT_400-inf_TuneCP5_13TeV-amcatnloFXFX-pythia8/proc/MINIAODSIM	0.4477
$Z \rightarrow \nu\nu + \text{jets}$	/Z1JetsToNuNu_M-50_LHEZpT_50-150_TuneCP5_13TeV-amcatnloFXFX-pythia8/proc/MINIAODSIM	596.3
	/Z1JetsToNuNu_M-50_LHEZpT_150-250_TuneCP5_13TeV-amcatnloFXFX-pythia8/proc/MINIAODSIM	17.98
	/Z1JetsToNuNu_M-50_LHEZpT_250-400_TuneCP5_13TeV-amcatnloFXFX-pythia8/proc/MINIAODSIM	2.057
	/Z1JetsToNuNu_M-50_LHEZpT_400-inf_TuneCP5_13TeV-amcatnloFXFX-pythia8/proc/MINIAODSIM	0.224
	/Z2JetsToNuNu_M-50_LHEZpT_50-150_TuneCP5_13TeV-amcatnloFXFX-pythia8/proc/MINIAODSIM	325.7
	/Z2JetsToNuNu_M-50_LHEZpT_150-250_TuneCP5_13TeV-amcatnloFXFX-pythia8/proc/MINIAODSIM	29.76
	/Z2JetsToNuNu_M-50_LHEZpT_250-400_TuneCP5_13TeV-amcatnloFXFX-pythia8/proc/MINIAODSIM	5.166
	/Z2JetsToNuNu_M-50_LHEZpT_400-inf_TuneCP5_13TeV-amcatnloFXFX-pythia8/proc/MINIAODSIM	0.8457
$t\bar{t}, 2\ell$ $t\bar{t}, 1\ell$	/TTTo2L2Nu_TuneCP5_13TeV-powheg-pythia8/proc/MINIAODSIM	88.29
	/TTToSemileptonic_TuneCP5_13TeV-powheg-pythia8/proc/MINIAODSIM	365.34
Single top	/ST_tW_top_5f_inclusiveDecays_TuneCP5(_PSweights)_13TeV-powheg-pythia8/proc/MINIAODSIM	35.85
	/ST_tW_antitop_5f_inclusiveDecays_TuneCP5(_PSweights)_13TeV-powheg-pythia8/proc/MINIAODSIM	35.85
	/ST_t-channel_top_4f_InclusiveDecays_TuneCP5_PSweights_13TeV-powheg-pythia8/proc/MINIAODSIM	136.02
	/ST_t-channel_antitop_4f_InclusiveDecays_TuneCP5_PSweights_13TeV-powheg-pythia8/proc/MINIAODSIM	80.95
	/ST_s-channel_top_leptonDecays_13TeV-PSweights_powheg-pythia8/proc/MINIAODSIM	2.0688
Diboson	/ST_s-channel_antitop_leptonDecays_13TeV-PSweights_powheg-pythia8/proc/MINIAODSIM	1.2934
	/WWTo2L2Nu_NNPDF31_TuneCP5_13TeV-powheg-pythia8/proc/MINIAODSIM	11.08
	/WWToLNUQQ_NNPDF31_TuneCP5_13TeV-powheg-pythia8/proc/MINIAODSIM	45.99
	/WZTo1L1Nu2Q_13TeV(_TuneCP5)_amcatnloFXFX_madspin_pythia8/proc/MINIAODSIM	11.66
	/WZTo1L3Nu_13TeV_amcatnloFXFX_madspin_pythia8/proc/MINIAODSIM	3.294
	/WZTo2L2Q_13TeV_amcatnloFXFX_madspin_pythia8/proc/MINIAODSIM	6.351
	/WZTo2Q2Nu_TuneCP5_13TeV_amcatnloFXFX_madspin_pythia8/proc/MINIAODSIM	6.317
	/WZTo3LNU_TuneCP5_13TeV-amcatnloFXFX-pythia8/proc/MINIAODSIM	5.062
	/ZZTo2L2Nu_TuneCP5_13TeV_powheg_pythia8/proc/MINIAODSIM	0.6008
	/ZZTo2L2Q_13TeV(_TuneCP5)_amcatnloFXFX_madspin_pythia8/proc/MINIAODSIM	3.693
	/ZZTo2Q2Nu_TuneCP5_13TeV_amcatnloFXFX_madspin_pythia8/proc/MINIAODSIM	4.554
	/ZZTo4L_TuneCP5_13TeV_powheg_pythia8/proc/MINIAODSIM	1.325

Table 6.7: Processes used to model standard model background for 2018 data taking year. Events in the WZTo1L1Nu2Q, ZZTo2L2Q, and ZZTo2Q2Nu processes in which generator level information showed that a Z-boson decayed to a pair of c -jets was used to model the signal in the VZ($Z \rightarrow c\bar{c}$) validation analysis.

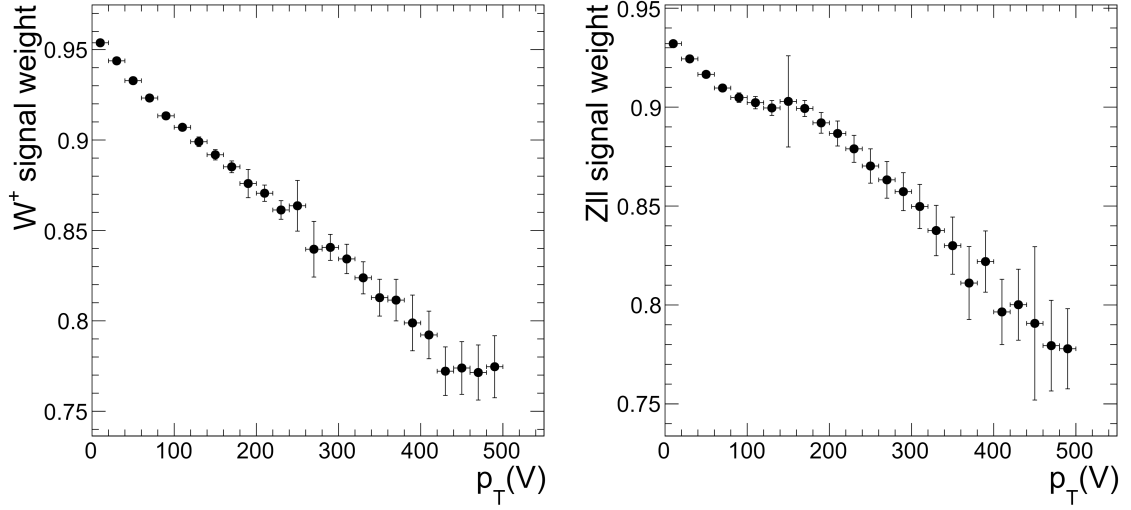


Figure 6.2.1: Event weights applied to the WH (left) and ZH (right) signal samples as a function of generated $p_T(V)$ to correct for the NLO in electroweak corrections to the differential cross section.

agreement with NNLO calculation in QCD. The total production cross section for the ZH and WH processes can be calculated as [17]

$$\sigma^{\text{WH}} = \sigma_{\text{NNLO,QCD}}^{\text{qqWH}}(1 + \gamma_{EW}) + \sigma_{\text{t-loop}} + \sigma_{\gamma'} \quad (6.2.1)$$

$$\sigma^{\text{ZH}} = \sigma_{\text{NNLO,QCD}}^{\text{qqZH}}(1 + \gamma_{EW}) + \sigma_{\text{t-loop}} + \sigma_{\gamma'} + \sigma^{\text{ggZH}} \quad (6.2.2)$$

where $\sigma_{\text{t-loop}}$ and $\sigma_{\gamma'}$ refers to subleading production modes in which the vector boson is created by a top quark loop or interactions of an off-shell photon. The term γ_{EW} represents the NNLO electroweak corrections to cross sections, which in our case are calculated at NLO. At NLO, we are able to apply these electroweak corrections as a function of $p_T(V)$. The signal cross section is then weighted by $(1 + \gamma_{EW})$ as a function of $p_T(V)$ as shown in Figure 6.2.1. This leads to a $\sim 10\%$ decrease in the expected number of VH events produced by the LHC.

6.2.3.2 NNLO in QCD and NLO in Electroweak Corrections to Diboson Differential Cross Section

Due to the need to accurately estimate the expected $VZ(Z \rightarrow c\bar{c})$ event yield in the validation analysis, the best available phenomenological calculation [99] were used to correct for the differential cross sections of the diboson samples used in this analysis. This is done through the application of NNLO in QCD and NLO in EW corrections applied first as an inclusive k-factor, which is a cross section correction factor accounting for higher precision calculations than which the sample was

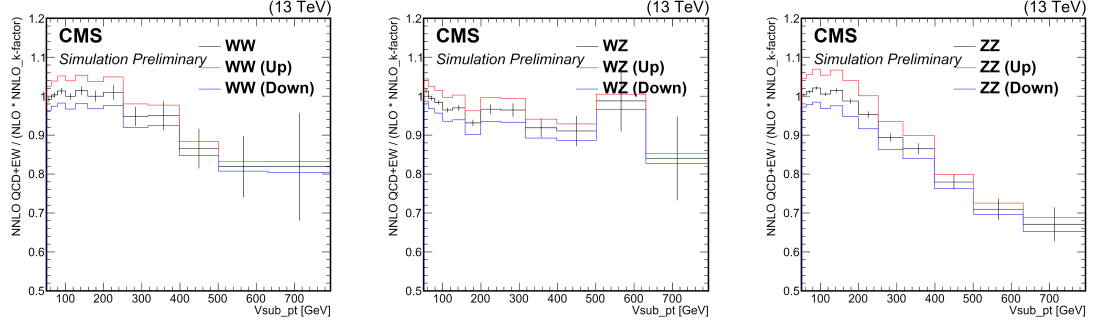


Figure 6.2.2: Differential NNLO in QCD and NLO in EW k-factors applied as a function of the generated $p_T(V)$ of the subleading boson in WW (left), WZ (middle), and ZZ (right) processes. Correction factors were derived and distributions were produced by the VH($H \rightarrow c\bar{c}$) merged jet analysis team.

generated at, to each of the WW, WZ, and ZZ processes. This uniform k-factor application has already been included in the σ column in Table 6.5–6.7. A second set of differential k-factors are then applied by comparing the NNLO in QCD and NLO in EW cross sections calculated in [99] to the NLO cross sections used to generate the samples as a function of the p_T of the subleading vector boson. As with the inclusive k-factors, a separate set of corrections are derived for the WW, WZ, and ZZ processes. Both sets of k-factors were derived for use in the VH($H \rightarrow c\bar{c}$) analysis by the merged jet analysis team. The distributions for the k-factors derived for each of the three diboson processes are shown in Figure 6.2.2.

6.3 Physics Object and Event Reconstruction

The physics objects used for this analysis closely follow the definitions laid out in Section 3.2.3. As such, all objects are reconstructed from measurements made by the CMS detector using the particle flow algorithm, and the physics objects used in the analysis are then first obtained using labels and parameters reconstructed from PF. Due to the importance of electron, muons, and jets to the analysis procedure, a tighter set of definitions are applied on top of the corresponding PF objects. This is to ensure that the analysis is only performed on physics objects with a high reconstruction efficiency.

6.3.1 Primary Vertex

As described in Section 3.2.3.3, the primary vertex is defined as the collision vertex with the largest sum of the transverse momentum associated with the tracks originating from it. For this study, we

only consider a single PV per event, and filter out any physics objects coming from PU collisions from our overall event description.

6.3.2 Lepton Identification

The analysis makes use of two different set of identification requirements for both electrons or muons. These are broken into a *tight* and *loose* classification in which all tight leptons are loose leptons that pass additional more stringent selection criteria. The reconstruction efficiency for loose leptons must be higher than for tight leptons (if defined as reconstructed leptons divided by all leptons). But the purify of the sample is higher for tight leptons than for loose leptons. For the $VH(H \rightarrow c\bar{c})$ analysis, tight leptons are used to reconstruct the W-candidate in 1-lepton channel events. Loose leptons are used for event selection and Z-candidate reconstruction in the 2-lepton channel, as well as vetoing events in the 0- and 1-lepton channel which are found to have too many leptons.

To improve our ability to reconstruct prompt leptons, an isolation requirement is applied in part of our selections to discern whether a lepton was produced as a prompt particle from the primary vertex of the collision or if it was produced during the production of a jet. The lepton isolation is calculated using PF candidates within a radius of 0.3 (0.4) in η and ϕ around and electron (muon), and is defined as:

$$I_\ell = \frac{1}{p_T^\ell} \left(\sum p_T^{\text{charged}} + \max \left[0, \sum p_T^{\text{neutral}} + \sum p_T^\gamma - p_T^{\text{PU}}(\ell) \right] \right) \quad (6.3.1)$$

where the p_T^ℓ is the p_T of the lepton in question, $\sum p_T^{\text{charged}}$ and $\sum p_T^{\text{neutral}}$ are the sum of the p_T of nearby charged and neutral hadrons, $\sum p_T^\gamma$ is the sum of the p_T of photons within isolation radius, and $p_T^{\text{PU}}(\ell)$ refers to the p_T of particles coming from PU interactions. From this, the isolation gives us a measure of what fraction of energy in the cluster of particles surrounding the lepton is contained within the lepton itself, with an extra factor subtracted out attributed to particles produced from PU.

6.3.2.1 Electron Identification

For all electrons, a preselection requirement is first made to select electrons with $p_T > 20 \text{ GeV}$, $|\eta| < 2.5$, and $I_e < 0.25$. Further electron selections are then made based on the MVA2 identification method [100] produced by the CMS EGamma physics object group (POG). MVA2 produces a single discriminator value obtained from a multivariate analysis to improve our ability to identify electrons with high efficiency. Loose electrons are defined as electrons which pass the preselection requirements

and the `Electron_mvaFall17V2Iso_WP90` discriminant cut, which has a 90% efficiency for electrons. Tight electrons are then defined as loose electrons with $p_T > 30 \text{ GeV}$ that additionally pass the `Electron_mvaFall17V2Iso_WP80` discriminant cut, which has an 80% efficiency for electrons.

For leptons used to reconstruct a vector bosons, one additional check is performed in which the electron is vetoed if it was reconstructed in the range $1.44 < |\eta| < 1.56$. This is to ensure that the electron was not reconstructed inside the gap between the EE and EB portions of the ECAL, in which electron and photon reconstruction efficiencies are low.

6.3.2.2 Muon Identification

For all muons, a preselection requirement is first made to select muons with $p_T > 20 \text{ GeV}$, $|\eta| < 2.4$, and $I_\mu < 0.25$. Loose muons are then required to meet the recommended definition from the muon POG, in which the muon object must be reconstructed as a muon from particle flow and that the CMS muon reconstruction algorithm reconstructed the object as either a Global Muon or a Tracker Muon (defined in Section 3.2.3.5). Tight muons are then defined based on an additional set of reconstruction based selections recommended by the muon POG:

- The muon is reconstructed as a global muon
- The muon is reconstructed as a particle flow muon
- The χ^2/N_{dof} of the global-muon track fit is less than 10
- There is at least 1 muon chamber hit in the global-muon track fit
- The muon is reconstructed as passing through at least 2 section of the muon system
- There is at least 1 pixel hit in the global-muon track fit
- There are at least 5 outer tracker hits in the global-muon track fit

Additionally, a tight muon must have $I_\mu < 0.06$, a transverse IP of $|d_{xy}| < 0.05 \text{ cm}$, a longitudinal IP of $|d_z| < 0.2 \text{ cm}$, and a $p_T > 25 \text{ GeV}$.

6.3.2.3 Jet Identification

As outlined in Section 3.2.3.6, jets used for the analysis are clustered using the anti- k_T algorithm [13] with a radius of $R = 0.4$ (AK4 jets). Additionally, the charged hadron subtraction (CHS) algorithm [101], in which energy clusters inside of the jet which are associated with PU vertices via track matching are removed from the jet cluster. The L1, L2, and L3 corrections described in

Section 3.2.3.6 are applied after the CHS procedure in both simulated and data jets. Additionally, the L2L3 residual corrections are applied on top of the data jets. After the CHS and JEC procedures have been applied, we then only select jets with $p_T > 25 \text{ GeV}$ and $|\eta| < 2.4$ for use in the analysis. A TIGHTLEPVETO [102] is then applied to reject any jets likely to be a high energy lepton which was incorrectly reconstructed as a jet. Additionally, jets with $p_T < 50 \text{ GeV}$ must pass a *tight* discriminant cut on BDT [103] produced by the JetMET group to veto jets likely to originate from PU.

After applying JECs, an additional smearing procedure is applied to the energies of simulated jets to account for the fact that the resolution of measured jet energy is sharper in simulation than in data. To apply the smearing, two potential applications are used. If the reconstructed jet is able to be matched to its generator level progenitor particle, the **scaling method** is used to derive a correction factor

$$c_{\text{JER}} = 1 + (s_{\text{JER}} - 1) \frac{p_T - p_T^{\text{ptcl}}}{p_T} \quad (6.3.2)$$

where p_T is the transverse momentum of the reconstructed jet, p_T^{ptcl} is the true momentum of the generated progenitor particle, and s_{JER} is a condition dependent factor. If we are unable to match the jet, a **stochastic smearing** technique is applied in which the correction factor is calculated as

$$c_{\text{JER}} = 1 + \mathcal{N}(0, \sigma_{\text{JER}}) \sqrt{\max(s_{\text{JER}}^2 - 1, 0)} \quad (6.3.3)$$

where $\mathcal{N}(0, \sigma_{\text{JER}})$ is a random number with mean 0 and variance σ_{JER}^2 , which is also a detector conditions dependent factor. For both cases, the correction factor c_{JER} is set to have a minimum value of 0 and is multiplied with the post-JEC simulated jet energy to obtain the smeared jet energy used for analysis.

6.3.2.4 Identification of Charm Jets

As described in Chapter 5 the DeepJet [40] discriminator is used to assign CvsL and CvsB scores to each of the jets in the event. Additionally, a set of c -tagging scale factors are applied on a per-jet basis in the MC samples to construct a per-event c -tagging event weights. The distribution of MC events are then weighted by the c -tagging weights.

6.3.2.5 Charm Jet Energy Regression

As described in Section 5.2 the sizable semileptonic branching fraction of D -mesons will lead to a significant fraction of c -jets having a non-trivial amount of their energy carried away in a neutrino. This will lead to an under measurement of the jet energy, which will in turn lead to decreased resolution in the reconstruction of the Higgs boson 4-momentum. Additionally, the L2 and L3 JECs applied on jets used in this analysis are derived based on flavor mixed QCD dijet events. However, the true calorimeter energy response is expected to be dependent on the particle flavor, and the flavor phase space of a jet will vary based on the flavor of the progenitor particle. This leads to the possibility to improve the JEC if we know the jet flavor. To address both of these factors, a dedicated energy regression algorithm is trained for this analysis to improve our ability to reconstruct the true energy of c -jets. Further information about the energy regression algorithm and its application is described in Section 6.5.2.

6.3.3 Missing Transverse Energy

The missing transverse energy (MET) follows the same definition as given in Section 3.2.3.7. However, to account for the effects of JEC applications, so-called TYPE-1 MET corrections are applied in which the MET is re-calculated after application of the JECs (and jet smearing in simulation). During the calculation of the corrected MET, an additional set of filters are applied to account for known noisy components inside of the CMS detector. In addition to the standard MET, the following MET based variables are used in the 0- and 1-lepton channels: The MET significance (METsig), which is defined as the MET divided by the sum of the transverse momentum of all PF candidates, and the tracker MET (trkMET), which is the vector sum of p_T from the reconstructed tracks in the event.

6.3.4 Data Taking Year Specific Corrections

Due to the ever changing conditions of the CMS detector, multiple changes can occur in the detector leading to changes in reconstruction efficiencies across the three different data taking periods. In general, these changes are accounted for in the GEANT4 detector simulation or absorbed into standard sets of scale factors used to correct for differences in reconstruction efficiencies observed in data and simulation. However, two separate detector level issues arose during the Run 2 data taking period which required specialized corrections for only a subset of the data taking years.

6.3.4.1 ECAL Prefiring in 2016 and 2017

Due to the gradual build-up of radiation effects in the ECAL scintillators, a natural shift occurs in the readout timings over the course of a data taking year. However, this shift was not properly propagated to the trigger primitives in the 2016 and 2017 data taking years. This resulted in portions of the ECAL showers in the most forward regions of the EE ($2.0 < |\eta| < 3.0$) to be incorrectly attributed to the previous bunch crossing as out of time pileup (OOP). Due to the latency associated with reading out the detector, two consecutive bunch crossings are not allowed to fire the L1 trigger system. This lead to some events with forward EGamma objects to trigger on the wrong bunch crossing and be lost.

The L1 trigger prefiring effect was not accounted for when simulating the MC samples used in this analysis. A correction is instead applied as an event level weight to simulated samples in the 2016 and 2017 data taking periods. The weight is calculated by considering the physics objects and HLT paths of the simulated events and computing the probability that it would have triggered the L1 prefire error. An additional systematic uncertainty associated with the ECAL prefiring corrections is applied to 2016 and 2017 simulated samples as well (Section 6.11).

6.3.4.2 HEM15/16 Issue in 2018

During run 319077 of the 2018 data taking period, a power supply failure lead to the complete loss of two sectors of the HCAL Endcap Minus (HEM15 and HEM16) for the remainder of the data taking period. This resulted in a decreased jet energy response as well as an increased probability to improperly reconstruct a jet as an electron in the region of the detector with $-1.57 < \phi < -0.87$ and $-3.0 < \eta < 1.3$. To account for this issue, minor adjustments were applied to the 2018 analysis strategy to veto data events which were potentially impacted by the HEM15/16 failures. Additionally, the effects of the HEM15/16 failures were not included in GEANT4 when simulating the 2018 version of the MC samples, requiring the addition of extra efficiencies and systematic uncertainties to be applied in post-processing to the simulated samples.

To account for reduced object reconstruction efficiencies in the failure region, any event in which one of the two jets used to reconstruct the Higgs boson is found to overlap with the effected η and ϕ region of the detector are vetoed. Additionally, any events in the 1L channel in which the electron used to reconstruct the W-boson falls within the effected region of the detector is also vetoed. Due to the fact that most high energy objects produced at the LHC tend to be boosted towards the low $|\eta|$ region of the detector, we found that the HEM15/16 had an overall minimal

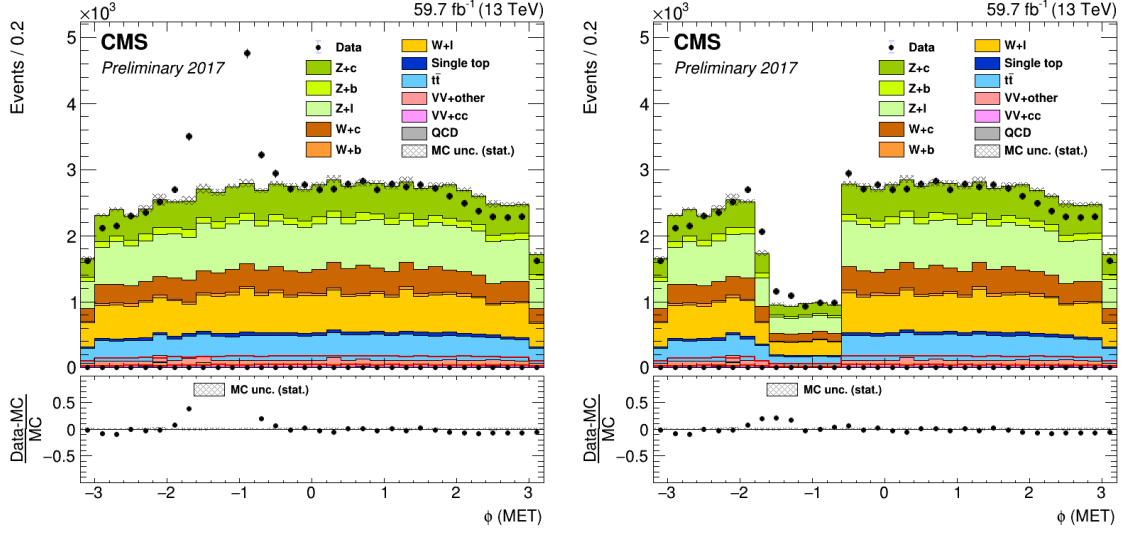


Figure 6.3.1: Distribution of $\phi(\text{MET})$ in the 0L channel for the 2018 data taking year before (left) and after (right) application of the $\phi(\text{MET})$ veto in data and weighting in MC. A residual deficit in data yield can still be observed in the opposite side of the ϕ spectrum.

effect on the 1- and 2-lepton analysis channels. However, the effects of the issue was found to be much greater in the 0-lepton channel, where in data the HEM15/16 was found to produce a large amount of fake MET pointing in the direction of the effected region of the detector. To account for this, a veto was placed on all data events in the 0-lepton channel with $\text{run} \geq 319077$ and $-1.7 < \phi(\text{MET}) < -0.6$. To properly model this veto in MC, all simulated 2018 events in the 0L channel with $-1.7 < \phi(\text{MET}) < -0.6$ were weighted down by a factor of 0.3529 corresponding to the percentage of the 2018 integrated luminosity not affected by the HEM15/16 issue. In total, this leads to $\sim 10\%$ loss in effective luminosity for the 0L channel during 2018. A comparison of the HEM15/16 effects to the 0-lepton channel before and after application of our veto and weighting procedure is shown in Figure 6.3.1.

In addition to measuring events with large fake MET in the direction of the HEM failure region, the HEM issue creates a second-order effect where MET is under measured in events with true MET pointing opposite direction of the failure region. This leads to a small over-estimation of MC events in the 0-lepton channel with $0.9 < \phi(\text{MET})$ or $\phi(\text{MET}) < -2.2$. This over estimation was found to have little correlation with other event kinematics, and predominantly presents itself as a normalization offset between MC and data for Z+jets where the Z undergoes the $Z \rightarrow \nu\nu$ decay mode. Because this normalization of this parameter is allowed to freely float as a rate parameter during out signal extraction, no further action is carried out to correct for the MC to data agreement in this phase space.

6.4 Vector Boson Reconstruction

The proper reconstruction of the V-boson is vital for the identification of $VH(H \rightarrow c\bar{c})$ events. This is done through exploitation of the various leptonic decay modes of the Z- and W-bosons, which have high reconstruction efficiency while simultaneously allowing us to separate our events from the large amount of QCD multijet background. To ensure that we do not double count events by placing them in multiple channels, we only reconstruct a single type of V-candidate per event. As such, we search for the V-candidate in the following order, giving preference to the decay modes with higher reconstruction efficiency: $Z \rightarrow \ell\ell$, $W \rightarrow \ell\nu$, and then $Z \rightarrow \nu\nu$. The tagging and reconstruction is then performed as follows:

Z $\rightarrow \ell\ell$ – Candidates are obtained from events containing a pair of oppositely charged same flavor leptons passing the loose lepton requirement. The Z-boson candidate is then reconstructed by adding the 4-momentum of the pair of opposite charged same flavor leptons with the highest p_T .

W $\rightarrow \ell\nu$ – Candidates are obtained from events containing a lepton passing the tight requirement, and have no additional tight or loose leptons of the same flavor and opposite sign. The W-boson candidate is then reconstructed by summing the 4-momentum of the highest p_T tight lepton and the MET. As such we compute the p_T of the W-boson candidate as

$$p_T(W) = \sqrt{(\text{MET}_x + p_x^\ell)^2 + (\text{MET}_y + p_y^\ell)^2}$$

which is used elsewhere in the analysis.

Z $\rightarrow \nu\nu$ – Candidates are reconstructed from events with $\text{MET} > 170 \text{ GeV}$ and no leptons passing the loose identification requirements. The Z-candidate is then assumed to be the MET object.

As mentioned in Section 6.3.2, any electron which is reconstructed in the range $1.44 < |\eta| < 1.56$ will not be used to reconstruct Z- or W- boson. However, such electrons will still be used to deselect events from the 0- and 1-lepton channels. This is to veto any electrons reconstructed within the gap between the EE and EB partitions of the ECAL, which has poor EGamma identification and energy reconstruction efficiencies.

6.4.1 Vector Boson Transverse Momentum Re-Weighting in Z+jets Events

The p_T spectra of the reconstructed Z-bosons in data are observed to be non-perfectly described (mainly in the low $p_T(V)$ region) by the NLO amcatnloFxFx MC samples. It was found that the $p_T(V)$ of the Z-boson was systematically higher in MC than in data, primarily when $p_T(V) \lesssim 150$ GeV. This disagreement was found to follow a linear trend, as such a $p_T(V)$ dependent correction factor was derived to account for this disagreement. Because of the low Z+jets contamination in the 1-lepton channel and the large MET required for reconstructing the Z-candidate in the 0-lepton channel, we found that corrections were only required for events the 2-lepton channel.

The correction factor was derived by fitting the reconstructed $p_T(Z)$ in MC to data in the low $p_T(V)$ 2-lepton control region used to isolate the Z+light-flavor jets (full description of this phase space in section Section 6.8) inclusively to events where the Z decays to e^+e^- or $\mu^+\mu^-$. Events in the light-flavor control region were chosen due to the high purity of Drell-Yan (DY) events allowing the fit to not be influenced by other background processes. Furthermore, we expect the p_T of the Z to have small dependence on the flavor of the additional partons produced during hard scattering allowing us to extrapolate the fit to Z+jet events with heavier jet flavors. To disentangle this correction from other MC scale factors, the derivation is performed without the c-tagging efficiency nor $\Delta R(jj)$ scale factors (described in Section 6.4.2) applied. Due to slight variations in parameters used to generate MC and lepton efficiencies, correction factors were derived independently for each data taking year.

year	Slope	Intercept
2016	$(1.693 \pm 0.091) \cdot 10^{-3}$	0.8035 ± 0.0088
2017	$(0.886 \pm 0.137) \cdot 10^{-3}$	0.8018 ± 0.0132
2018	$(0.443 \pm 0.112) \cdot 10^{-3}$	0.8115 ± 0.0105

Table 6.8: Correction factor parameters extracted from fit for 2016, 2017, and 2018 data taking years.

The derived fit parameters can be found in Table 6.8. Because we observed improved data modeling at high $p_T(V)$, the correction factor is only applied if it has a value less than one. Otherwise, it is set to unity. Fit uncertainties are treated the same way, where the shape uncertainty variation is set to unity if it were to have a value greater than one. A comparison of the $p_T(V)$ spectrum in data and simulation with and without the application of the scale factors can be found in Figure 6.4.1. The full distributions of the $p_T(V)$ variable before and after application of the correction factor can be seen in Figure 6.4.2. Residual disagreements between data and MC in the distributions are still observed after applying the correction factor due to the presence of the c -tagging scale factors which

were not applied during the fitting procedure.

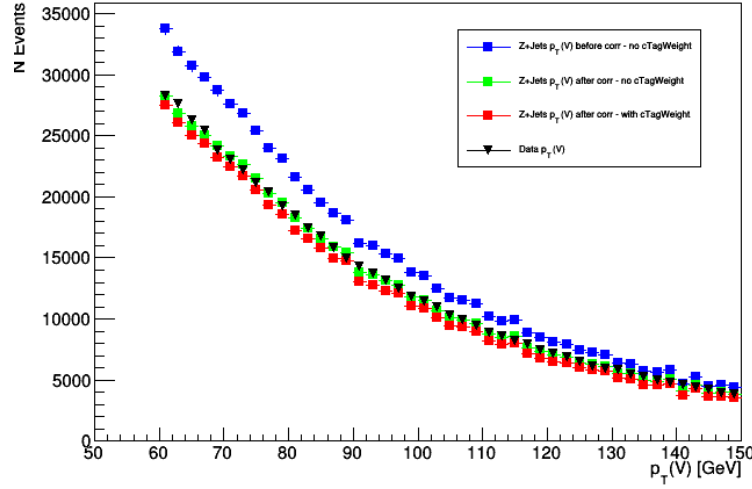


Figure 6.4.1: The distribtution Drell-Yan NLO 2018 MC events before and after corrections (right).

6.4.2 Jet Separation Re-weighting in Z+jets Events

A residual data/MC disagreement was found in events with low $\Delta R(jj)$ in the control and signal regions used to carry out the analysis. Further investigation found an improper modeling of jet separation in the NLO amcatnloFxFx V+Jet MC samples leading to an underestimation of events with $\Delta R(jj) < 1$, as shown in Figure 6.4.3.

The reason for mis-modelling in MC has not yet been determined. Therefore it was decided to apply dedicated scale factors to MC events to match the $\Delta R(jj)$ distributions in data. Despite not knowing the exact cause, one explanation for the disagreement in low $\Delta R(jj)$ could be due to a mismodeling of the gluon splitting processes in MC. Furthermore, gluon splitting is known to have different branching fractions to c-, b-, and light-quark final states, allowing for a potential flavor dependence on the $\Delta R(jj)$ distribution in suitably merged jets. Because of this possibility, extra care was taken to ensure that applying the scale factors does not introduce any new mismodeling in the jet flavor composition in MC. In general, we expect that the flavor dependence due to gluon splitting primarily exists for events with $\Delta R(jj) < 0.4$, which are excluded from our event selection due to criteria of having two fully resolved AK4 jets (cone radius of 0.4). Further studies into the flavor dependence on $\Delta R(jj)$ found the dependence to be negligible for events with $0.55 < \Delta R(jj) < 1.0$, and small for events with $0.4 < \Delta R(jj) < 0.55$, as shown in Figure 6.4.4. Thus, we concluded to derive

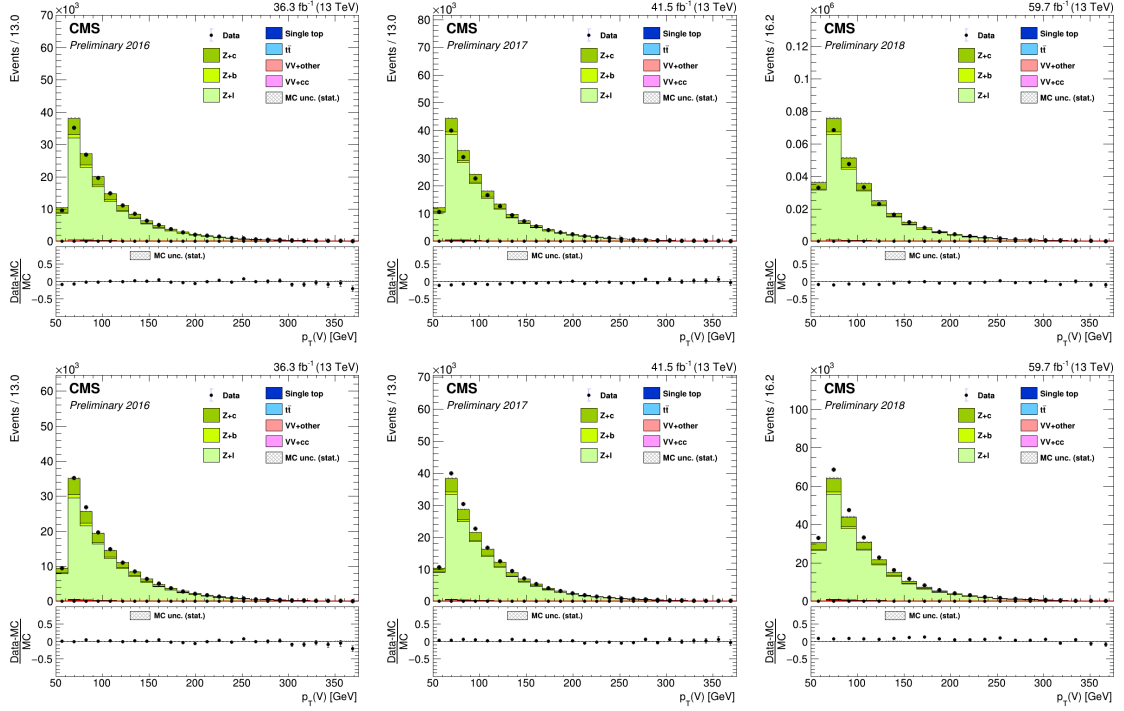


Figure 6.4.2: Inclusive distribution of the $p_T(V)$ variable for combined high and low $p_T(V)$ channels in the $Z(\ell\ell)H$ control region used to derive the correction factors. Distributions are plotted before (top) and after (bottom) the correction factors are applied for all three data taking years.

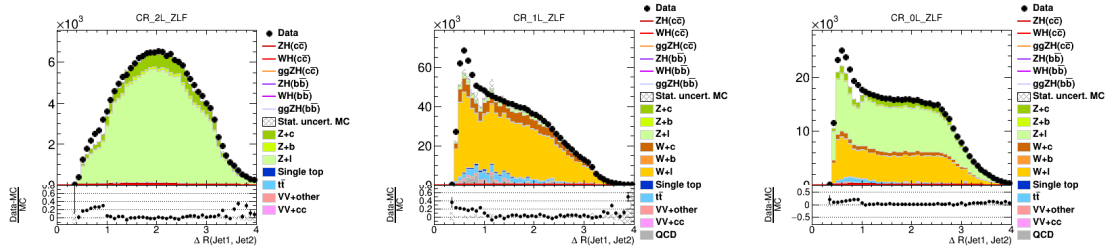


Figure 6.4.3: Distribution of the $\Delta R(jj)$ variable in the light flavored control regions of the 2-lepton, 1-lepton, and 0-lepton analysis channels in the 2017 data taking year. Residual mismodeling in simulation is observed for events with $\Delta R(jj) < 1$.

flavor inclusive $\Delta R(jj)$ scale factors to the different V+Jet processes with additional uncertainties applied to account for the $\Delta R(jj)$ dependence on jet flavor.

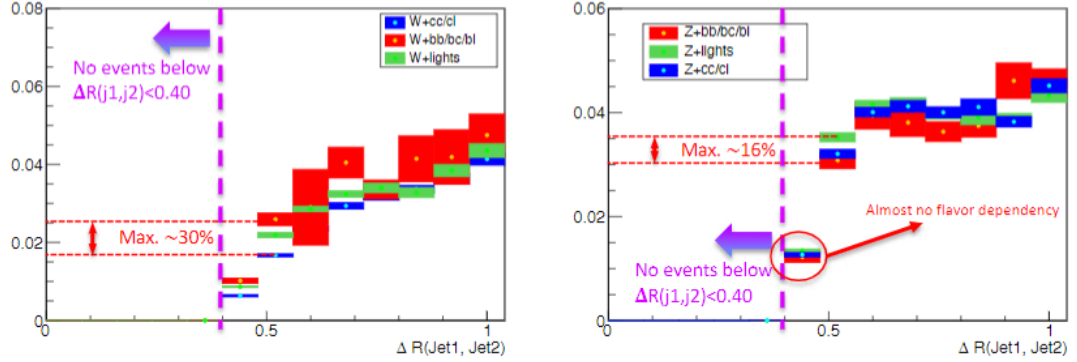


Figure 6.4.4: Shape of $\Delta R(jj) < 1.0$ distribution of W+Jet (left) and Z+Jet (right) events for various jet flavor compositions.

6.4.2.1 Scale Factor Derivation

To derive the scale factors, $\Delta R(jj)$ distributions in MC are sequentially fit to data in the region $\Delta R(jj) < 1.0$, in the light-flavor control region of three different channels: 2-lepton, 1-lepton, and 0-lepton. The light-flavor control region was chosen due to its ability to primarily select V+Jet events with low contamination from the other primary background sources. Scale factor derivation in the 2- and 1-lepton channels was performed inclusively to electron and muon tagged events, and the 2-lepton channel derivation is additionally performed inclusively to high- and low- $p_T(V)$ events. The scale factors were then derived in 3 steps:

1. The scale factors for $Z(\ell\ell)$ +jets events was derived by fitting MC to data with a 1-degree polynomial in the 2-lepton light-flavor control region.
2. The scale factors for $Z(\ell\ell)$ +jets events was then applied to events in the 1-lepton light-flavor control region. After scale factor application, the non- $W(\ell\nu)$ +jets MC event distribution was subtracted from the data distribution in the control region. A 1-degree polynomial fit to $W(\ell\nu)$ +jets events was then applied to the residual "data – MC" distribution.
3. The scale factors for $Z(\ell\ell)$ +jets and $W(\ell\nu)$ +jets were applied to events in the 0-lepton light-flavor control regions. After scale factor application the non- $Z(\nu\nu)$ +jets MC event distribution was subtracted from the data distribution in the control region. A 3-degree polynomial fit to $Z(\nu\nu)$ +jets events was then applied to the residual "data – MC" distribution.

For 2018, one additional step was taken to account for the HEM15/16 issue that was experienced during data taken. After applying the veto to mismeasured 0-lepton events outlined in section 6.3.4.2, there is still the residual over-estimation in the normalization of $Z(\nu\nu)$ +jets MC events which is not corrected until the application of the rate parameters. To ensure that we do not unintentionally correlate the residual HEM15/16 normalization offset with $\Delta R(jj)$, we first find the ratio of $(Z(\nu\nu)$ +jets)/(data - 'non- $Z(\nu\nu)$ +jets MC yield') with $\Delta R(jj) > 1.0$ for the 0-lepton light-flavor control region. We then scale the (data - 'non- $Z(\nu\nu)$ +jets MC yield') distribution used for the scale factor derivation by this ratio before performing the 3-degree polynomial fit.

Distributions of $\Delta R(jj)$ in the light-flavor control regions after the application of the $\Delta R(jj)$ dependent scale factors for each of the data taking years can be found in Figures 6.4.5, 6.4.6, and 6.4.7. In all cases, the shape mismodeling has been rectified through the scale factor application.

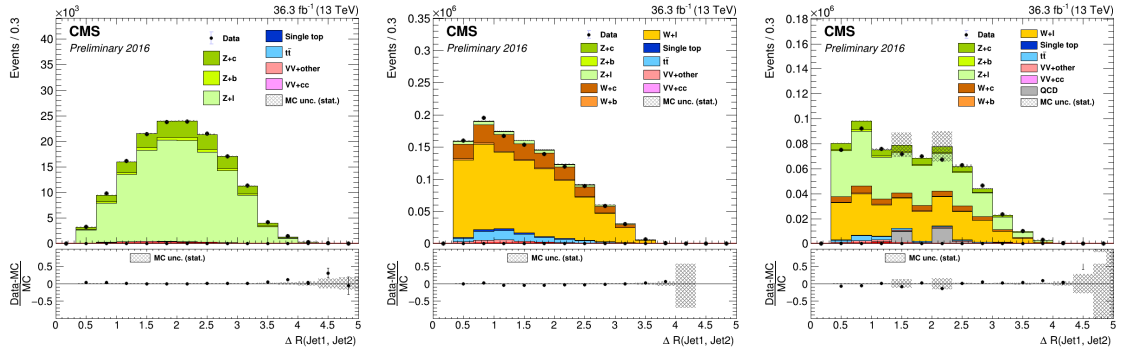


Figure 6.4.5: Shape of $\Delta R(jj)$ distribution of the 2-lepton (left), 1-lepton (middle) and 0-lepton (right) channels in the light-flavor control regions for the 2016 data taking year after application of the $\Delta R(jj)$ scale factors.

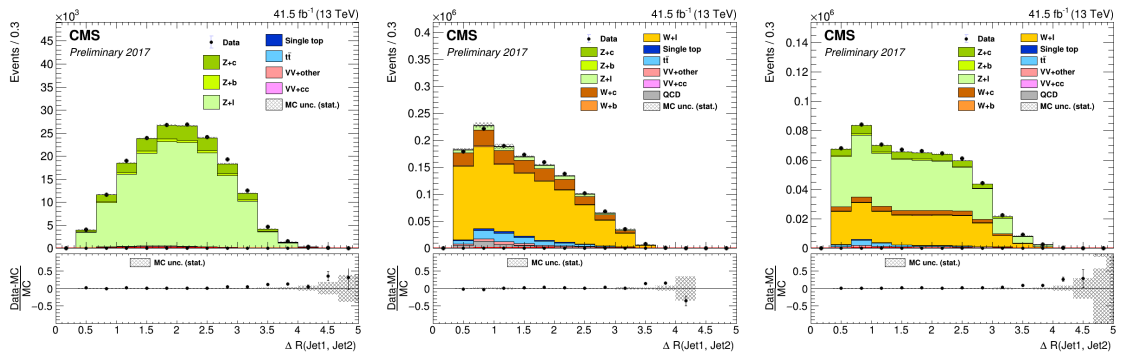


Figure 6.4.6: Shape of $\Delta R(jj)$ distribution of the 2-lepton (left), 1-lepton (middle) and 0-lepton (right) channels in the light-flavor control regions for the 2017 data taking year after application of the $\Delta R(jj)$ scale factors.

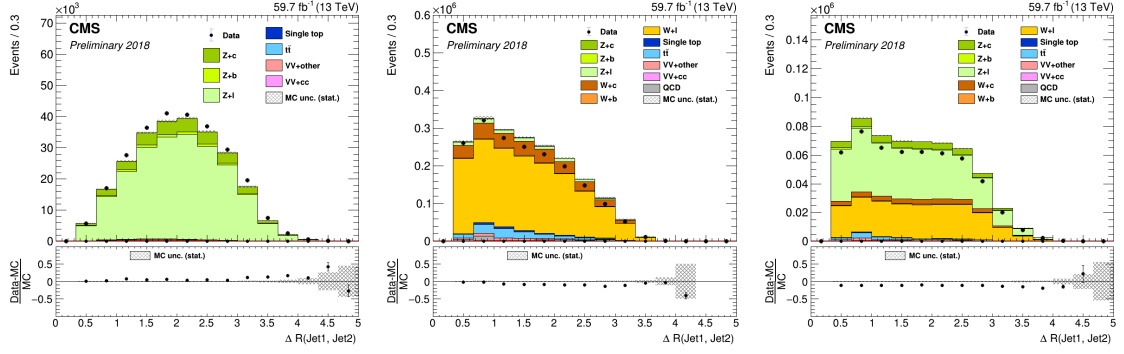


Figure 6.4.7: Shape of $\Delta R(jj)$ distribution of the 2-lepton (left), 1-lepton (middle) and 0-lepton (right) channels in the light-flavor control regions for the 2018 data taking year after application of the $\Delta R(jj)$ scale factors. For the 0 lepton channel, the fit was performed in such a way that the normalization offset between data and MC was preserved for low and high $\Delta R(jj)$ events.

6.5 Higgs Boson Reconstruction

The $VH(H \rightarrow c\bar{c})$ resolved jet topology focuses on events in which the two charm quarks hadronize into two distinct and identifiable AK4 jets. As such, we attempt to reconstruct the H-candidate in an event by summing the 4-momenta of the two jets which it decayed to. However, even for VH events we often expect more than two jets per event. This is primarily due to the presence of initial state radiation (ISR) in which gluons radiated during the scattering process hadronize to produce additional jets in an event. Standard practice is to select the two jets with the largest p_T , since we expect ISR jets to have softer p_T than the Higgs boson. However, because we expect all of the Higgs jets in the $H \rightarrow c\bar{c}$ decay mode to be involve charm quarks which hadronize to D -mesons, we also investigated the possibility to select the Higgs jets based on the c -tagging discriminators of the DeepJet algorithm [40] (Section 5.3.2).

As shown in Table 6.9 we investigated three ways to select the pair of jets to reconstruct the H-candidate, and found that using the two jets with the largest DeepJet CvsL score has a higher reconstruction efficiency in $VH(H \rightarrow c\bar{c})$ MC events than selecting the jets based on jet p_T or CvsB score. As such, we reconstruct our H-candidate using the two jets with the highest DeepJet CvsL scores in the event, referring to the jet with the highest score as the leading jet and the jet with the second highest score as the subleading jet. An additional study from a collaborator was also performed in which a BDT was trained to select jets based on their p_T , CvsB, and CvsL values. A slight improvement in selecting the Higgs jets was found in MC when selecting jets based on the BDT performance. However, we found that such a method introduced mismodeling when comparing the H-candidate based distribution of MC and data events. As such, we decided to not use a BDT

channel		p_T	CvsL	CvsB
0L	Z($\nu\nu$)	73.47%	73.20%	71.43%
1L	W($\mu\nu$)	59.29%	66.91%	51.93%
	W($e\nu$)	58.57%	66.47%	51.18%
2L	Z($\mu\mu$)	59.84%	66.57%	50.88%
	Z(ee)	59.66%	66.46%	50.76%

Table 6.9: The percentage of events in which the two jets used to reconstruct the H-candidate properly reconstruct the Higgs boson in VH($H \rightarrow c\bar{c}$) signal events for various selection methods. We found that with the exception of Z($\nu\nu$)H events, in which reconstruction efficiency had a small dependence on selection methodology, selecting jets based on their CvsL score had the highest probability of properly selecting the two c -jets.

based selection and continued selecting jets based on CvsL performance.

After reconstructing the H-candidate, additional techniques are performed to improve the estimations of the H-candidate’s kinematic properties. In particular, we aim to improve our ability to accurately calculate the p_T and invariant mass of the H-candidate motivated by the variables’ importance in identifying Higgs boson events. For this analysis, both variables, as well as kinematic variables of the jets used to reconstruct the H-candidate, are used as inputs for a BDT used to discriminate between signal and background events. This allows us to improve the potential sensitivity of the analysis by improving our accuracy when calculating these values.

6.5.1 FSR Recovery

Final state radiation (FSR) refers to the phenomenon in which a hadronizing jets emit gluons as part of the hadronization process. In some instances, these FSR gluons can be emitted at a wide enough angle, that their hadronization products fall outside of the cone of the originating jet. In such instances, the FSR will lead to the clustering of a separate soft jet which lies close to the originating jet in η and ϕ . When this process occurs for one of the Higgs jet, the jet selection scheme will fail to attribute the FSR jet as part of the H-candidate decay chain leading to an undermeasurement of the H-candidate’s energy and p_T . To account for this, an FSR recovery scheme is used in which the 4-momentum of all soft jets with $p_T > 20 \text{ GeV}$ and $|\eta| < 3.0$ that are within a distance of $\Delta R < 0.8$ of one of the two Higgs jets are included in the 4-momentum of the H-candidate. After application of the FSR procedure, we observe improvements of 1–8% in the resolution and 0.3–3.6% in the accuracy of the reconstructed Higgs mass in simulation. Distributions of the reconstructed Higgs mass before and after application of the FSR recovery for the 2-lepton low and high $p_T(V)$ analysis channels can be found in Figure 6.5.1. In both instances, none of the other techniques used to improve the calculation of the Higgs boson kinematics described in this section are applied.

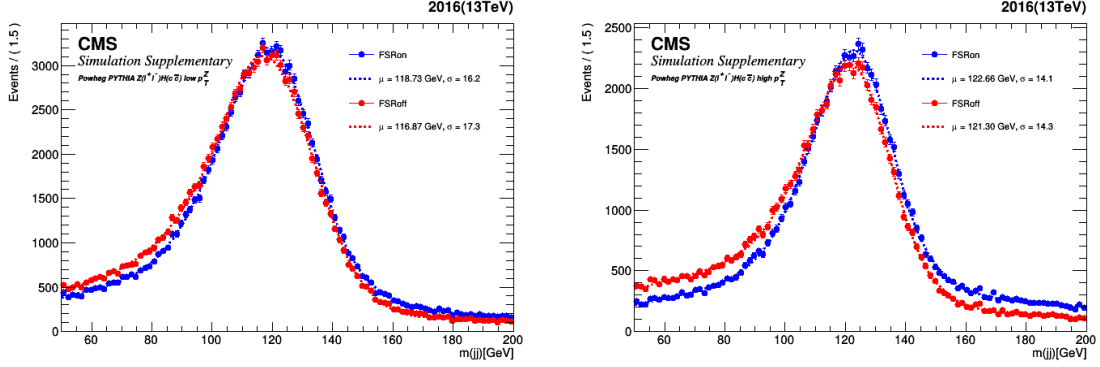


Figure 6.5.1: Reconstructed Higgs mass of 2016 $Z(\ell\ell)H(c\bar{c})$ signal events with (blue) and without (red) the application of the FSR recovery procedure for the 2-lepton low (left) and high (right) $p_T(V)$ analysis channels binned inclusively for electron and muon Z-boson decays. The mass peaks are fitted with a double crystal ball function to obtain the dashed lines.

6.5.2 Application of a Charm Jet Energy Regression

As described in Section 5.2 a sizable fraction of c -jets can contain soft leptons originating from semileptonic decays of the charm mesons. In such instances, the leptons charged will be produced in conjunction with a neutrino that is not directly measurable by the CMS detector. Additionally, because the neutrino is a direct decay product of the charm quark, it can contain a large fraction of the overall jet energy leading to large underestimations of the c -jet energy. For $H \rightarrow c\bar{c}$ decays, the presence of neutrinos in c -jets leads to a broadening of the reconstructed Higgs mass resolution, potentially worsening the sensitivity of the analysis.

To recover the lost energy associated with neutrinos produced in the c -hadron decay chain, a dedicated energy regression algorithm is applied to the c -jets. This is done through the use of a DNN architecture following the setup used in [104] for the $VH(H \rightarrow b\bar{b})$ analysis, but with the algorithm trained using a sample of c -jets rather than b -jets. The training and validation of the regression network was then performed by one of my collaborators on the $VH(H \rightarrow c\bar{c})$ resolved jet analysis team. The DNN consists of 6 fully connected hidden layers and is fed an input of 41 features including jet kinematics, PU information, features of leptons reconstructed as part of the jet, features of the secondary vertex of the jet, and features of the jet constituents. These features are fed into the network after application of JECs, but without any smearing performed on the jet energy. The regression network is trained to calculate three output values: a multiplicative correction factor to the jet energy, the 25% quantile to the correction factor, and the 75% quantile to the correction. Here, the outputs represent the points on a probability distribution for the regressed energy estimate, where the correction factor represents the central value of a gaussian distribution

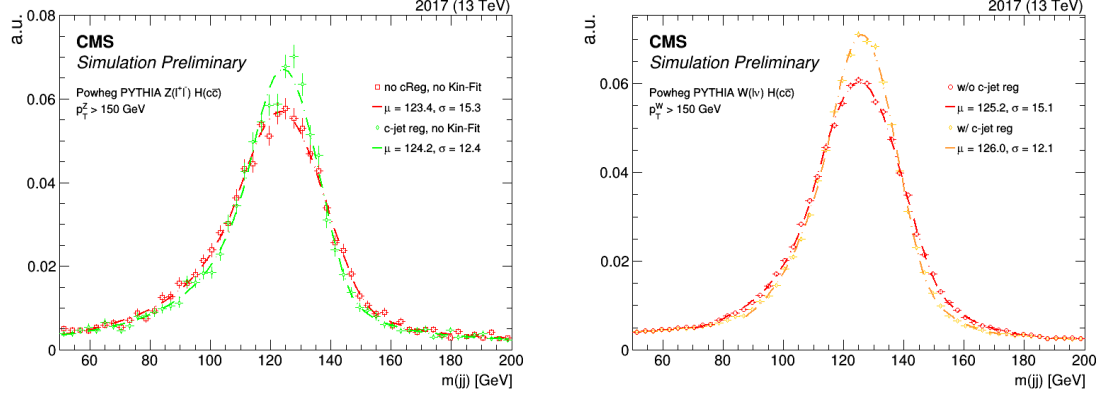


Figure 6.5.2: Distribution of the reconstructed Higgs mass with and without the application of the c -jet energy regression for $Z(\ell\ell)H(c\bar{c})$ high $p_T(V)$ (left) and $W(\ell\nu)H(c\bar{c})$ (right) events. Central value and standard deviation of the distributions are computed with a double crystal ball function. For the 2-lepton analysis channel, the c -jet regression is also used with a kinematic fit to further improve the calculation of the Higgs boson mass.

and the 25% (75%) quantile represents the symmetric range about the central value such that the true correction factor will be within said range 25% (75%) of the time.

During the analysis, we apply the c -jet energy regression only to the two jets selected to reconstruct the H-candidate at the time of their selection. Thus, the jet Id selections described in Section 6.3.2.3, DeepJet discriminator values applied to each of the jets, and the calculation of the MET are fully agnostic to the regressed c -jet energy. However, the 4-momentum of the H-candidate, the event selections defined in Section 6.6, and BDT input variables described in Section 6.7 are computed after application of the c -jet energy regression. Figure 6.5.2 shows the distribution of the reconstructed Higgs mass in 2017 $Z(\ell\ell)H(c\bar{c})$ MC events with and without the application of the c -jet energy regression in the 2-lepton high $p_T(V)$ and 1-lepton analysis channels. After application of the energy regression, we find an improvement of $\approx 20\%$ in the mass resolution.

As described in Section 6.3.2.3 the energy of MC jets are smeared to match the energy resolution of jets reconstructed by the CMS detector. Unfortunately, the standard smearing procedures would not be correct if applied on top of the regressed jet, and instead a dedicated smearing factor is applied. The smearing factors were obtained by identifying events in which a Z-boson which decays to charged leptons is produced with two jets, one of which has low p_T . A fit is then performed to the value $p_T(j_1)/p_T(\ell\ell + j_2)$ in bins of $p_T(j_1)/p_T(\ell\ell)$ and extrapolated to the values in which $p_T(j_2) = 0$. The ratio of $p_T(j_1)/p_T(\ell\ell)$ at $p_T(j_2) = 0$ is then taken as s_{JER} . Because we use the same regression architecture studied by the $VH(H \rightarrow b\bar{b})$ analysis group, we found that the smearing values obtained for their study (none for 2016, 5% for 2017, and 15% for 2018) were suitable to

provide good agreement between regressed energy of jets in data and MC.

6.5.3 Application of a Kinematic Fit in the 2-Lepton Channel

To further improve estimations of the H-candidates kinematics, a kinematic fit is used in the 2-lepton channel of the analysis. The kinematic fit makes use of the fact that, with the exception of neutrinos produced inside of hadronizing jets, there are no unmeasurable particles in the $Z(\ell\ell)H(c\bar{c})$ final state. We can then use the invariant mass of the Z-pole as a constraint to improve the accuracy of the lepton kinematics. We can additionally use the fact that the V- and H-bosons will be produced back to back to place additional constraints on the p_T of the Higgs jets and leptons.

To perform the kinematic fit, we use the p_T , η , and ϕ values of the following physics objects as inputs:

The Higgs candidate jets – The two c -jets selected to reconstruct the H-candidate. The jets are fed as inputs to the fit procedure after the application of the c -jet energy regression. The 4-momenta of these jets do not include the 4-momenta of any soft jets included in the FSR recovery procedure.

The recovered FSR jets – Any jets selected as part of the FSR recovery procedure fed into the kinematic fit as independent 4-vectors.

The two leptons – The 4-momentum of the two opposite charged e^+e^- or $\mu^+\mu^-$ pairs used to reconstruct the Z-candidate.

A Recoil Vector – Jets initiated as ISR are not considered part of the H-candidate, and thus are not used to reconstruct the Higgs. However, these jets are still important during the fit to ensure that we properly account for the conservation of p_T within the event. The 4-momentum of all found ISR jets in the event are then summed together and fed into the kinematic fit as a single 4-momentum. The ISR jet selection criterion is the same jet Id listed in Section 6.3.2.3 but with the η acceptance increased to $|\eta| < 5$.

The 4-momenta of each of the aforementioned inputs are then allowed to float within their experimental uncertainties such that the following fit constraints are met:

Z-boson mass – the p_T of the two lepton inputs are allowed to float within their uncertainty ranges to fit the invariant mass of the SM Z-peak. As such, the target invariant mass value is set to a gaussian distribution centered about 91 GeV with a standard deviation of 5 GeV.

p_T of all input 4-momenta – The values of the p_T , η , and ϕ of each of the kinematic fit inputs are adjusted such that the vectoral sum of the p_T is 0.

Because we are working with a large number of floating variables but only two constraints, there will exist a large number of potential 4-momentum which can satisfy the constraints. As such, the set of 4-momentum variables which minimizes the function $\chi^2 = \sum_i^N (x_i^{\text{reco}} - x_i^{\text{fit}})^2 / \sigma_i$ are selected and extracted to be used as an additional set of BDT variables for the 2-lepton analysis channel (Section 6.7). The study performed to test the effectiveness for the kinematic fit was performed by one of my collaborators on the VH($H \rightarrow c\bar{c}$) resolved jet analysis team, and the comparisons of the central value, standard deviation, and resolution of the Higgs mass as computed from a double crystal ball fit as obtained from their study can be found in Table 6.10. In the final analysis procedure, the kinematic fit is performed using the regressed c -jets.

No c -jet regression							
category	wo/ kinfit			w/ kinfit			improv. res. %
	μ	σ	σ/μ	μ	σ	σ/μ	
2L, low $p_T(V)$	118.7	18.5	0.156	119.6	15.9	0.133	14.0
2L, high $p_T(V)$	123.4	15.3	0.124	124.0	9.8	0.079	36.3

With c -jet regression							
category	μ	σ	σ/μ	μ	σ	σ/μ	improv. res. %
	μ	σ	σ/μ	μ	σ	σ/μ	
2L, low $p_T(V)$	121.2	15.1	0.125	121.2	14.7	0.121	3.2
2L, high $p_T(V)$	124.2	12.4	0.100	124.5	9.1	0.073	27.0

Table 6.10: The mean and width of the reconstructed H-candidate mass obtained from a double crystal-ball fit with (bottom) and without (top) use of regressed c -jets. In the bottom table, the relative improvement after applying the kinematic fit is lower, because it is being compared to the already improved upon m_H resolution obtained from the energy regression algorithm.

6.6 Event Selection

Events are selected as shown in Table 6.11 in order to minimize contamination from the dominant background processes in a VH($H \rightarrow c\bar{c}$) signal rich phase space. As mentioned in Section 6.1 and Section 6.4 the VH($H \rightarrow c\bar{c}$) analysis makes use of a multichannel analysis strategy in which events are separated depending on the identified decay mode of the V-candidate. As such, the type of standard model background process and their corresponding decay topologies in each of these channels differ, leading to a separate set of event selections to best isolate the VH signal from each channel's dominant backgrounds. A description of each of the variables used to make selections are as follows:

Variable	Z(vv)H	W($\ell\nu$)H	Z($\ell\ell$)H low- p_T (V)	Z($\ell\ell$)H high- p_T (V)
$p_T(\text{V})$	> 170	> 100	[60 – 150]	> 150
$m_{\ell\ell}$	–	–	[75 – 105]	[75 – 105]
p_T^ℓ	–	(> 25(μ), > 30(e))	> 20	> 20
$p_T(j_1)$	> 60	> 25	> 20	> 20
$p_T(j_2)$	> 35	> 25	> 20	> 20
$p_T(\text{H})$	> 120	> 100	–	–
m_{H}	[50 – 200]	[50 – 200]	[50 – 200]	[50 – 200]
CvsL $_{j_1}$	> 0.225	> 0.225	> 0.225	> 0.225
CvsB $_{j_1}$	> 0.40	> 0.40	> 0.40	> 0.40
N_{aj}	–	< 2	–	–
N_{al}	= 0	= 0	–	–
MET	> 170	–	–	–
METsig	–	> 4.0	–	–
Anti-QCD	Yes	–	–	–
$\Delta\phi(\text{V}, \text{H})$	> 2.0	> 2.5	2.5	2.5
$\Delta\phi(\text{MET}, \text{trkMET})$	< 0.5	–	–	–
$\Delta\phi(\text{MET}, \text{lep})$	–	< 2.0	–	–

Table 6.11: Baseline selection used to construct the signal region in each of the four analysis categories. For each channel, events are first pre-selected based on whether the corresponding V-boson decay being reconstructed ($Z \rightarrow \nu\nu$, $W \rightarrow \ell\nu$, $Z \rightarrow \ell\ell$) and having the H-candidate reconstructed from two AK4 jets.

$p_T(\text{V})$ – The transverse momentum of the V-boson reconstructed as described in Section 6.4.

$m_{\ell\ell}$ – The invariant mass of the two lepton system used to reconstruct the V-boson in the 2-lepton analysis channel.

p_T^ℓ – The transverse momentum of the one (two) lepton(s) used to reconstruct the V-boson in the 1- (2-)lepton analysis channel. p_T requirement consistent with the lepton definitions for the given channel as described in Section 6.3.2.

$p_T(j_1)$ – The transverse momentum of the jet with the highest CvsL score in the event.

$p_T(j_2)$ – The transverse momentum of the jet with the second highest CvsL score in the event.

$p_T(\text{H})$ – Transverse momentum of the reconstructed H-candidate. Includes recovered 4-momentum from the FSR recovery procedure and c -jet energy regression, but is not calculated using variables extracted from the kinematic fit.

m_{H} – Invariant mass of the reconstructed H-candidate. Includes recovered 4-momentum from the FSR recovery procedure and c -jet energy regression, but is not calculated using variables extracted from the kinematic fit.

CvsL $_{j_1}$ – CvsL score of the jet with the highest CvsL score in the event.

CvsB_{j₁} – CvsB score of the jet with the highest CvsL score in the event.

N_{aj} – Number of additional jets in an event, not counting the two jets selected to reconstruct the H-candidate or jets selected during the FSR recovery procedure.

N_{al} – Number of additional leptons passing the loose requirement beyond those used to reconstruct the V-candidate. For the 1-lepton channel, this only refers to leptons with the same flavor and opposite sign as the lepton used to reconstruct the W-candidate.

MET – The magnitude of the missing transverse momentum as described in Section 6.3.3.

METsig – The MET as described in Section 6.3.3 divided by the number of PF candidates reconstructed in the event.

Anti-QCD – A requirement that the central axis of all jets satisfying the jet Ids in Section 6.3.2.3 have a greater ϕ separation with respect to the MET 4-vector than 0.5.

$\Delta\phi(\mathbf{V}, \mathbf{H})$ – The angular separation in ϕ between the V- and H-candidate 4-momentum.

$\Delta\phi(\text{MET}, \text{trkMET})$ – The angular separation in ϕ between the MET in the event and a separate type of missing transverse momentum calculated using only track objects.

$\Delta\phi(\text{MET}, \text{lep})$ – The angular separation in ϕ between the MET and the lepton used to reconstruct the W-candidate.

As will be discussed in Section 6.8, alternative sets of selections are made to choose non-overlapping event samples similar to the $\text{VH}(\text{H} \rightarrow c\bar{c})$ enriched signal region. These are constructed to individually constrain the most dominant background sources in the $\text{VH}(\text{H} \rightarrow c\bar{c})$ resolved jet analysis as a way to study residual disagreements between simulated and data background events.

6.7 BDT Training

To improve our ability to differentiate between signal and background events, boosted decision trees (BDTs) [105] are trained to discriminate between signal and background events in the signal regions of each analysis channel. A BDT is a multivariate analysis technique in which a large number of *decisions trees* (sometimes called *regression trees*) [106] are built and used to perform classification. A single decision tree contains a variable number of iterative boolean checks based on a given set of input features. At the end of each decision pathway, a single discriminating value is given to the

Analysis Channel	Number of Trees	Maximum Depth
2-lepton (all)	750	4
1-lepton (VH($H \rightarrow c\bar{c}$))	1000	3
1-lepton (VZ($Z \rightarrow c\bar{c}$))	300	3
0-lepton (all)	500	3

Table 6.12: Primary hyper parameters used to train the BDTs. For the 2- and 0-lepton channels, there were no differences in parameters used between the VZ($Z \rightarrow c\bar{c}$) and VH($H \rightarrow c\bar{c}$) BDT trainings. However, a smaller number of trees were used for the WZ(cc) BDT due to lower number of MC signal sample events.

event representing the tree’s confidence that an event following those selections would either be a signal (closer to +1) or background (closer to -1) event. Boosting is a technique in which a multitude of trees are iteratively produced and assigned a weight relative to their respective accuracy. For a given event, it is sequentially passed through each of the trees, with the discriminator output of each tree weighted by their respective tree weight and summed. For this analysis the *gradient boost* [107] machine learning technique is used to build the trees and assign their weights. The set of hyperparameters used for each of the BDTs trained can be found in Table 6.12.

In this analysis, an independent BDT is trained for each of the three lepton-based analysis channels and for each of the three data taking years, with the BDTs for the 2016 and 2017 data taking years were performed by my collaborators. Additionally, two separate BDTs are trained for the high- and low- $p_T(V)$ regions of the 2-lepton channel. The BDTs are trained using half of the weighted MC events (events with odd event numbers assigned) that pass the signal region selections for their corresponding analysis channel. The BDTs are then trained with the VH($H \rightarrow c\bar{c}$) MC samples set as the signal and all other simulated MC samples set as an inclusive background.

In the case of the VZ($Z \rightarrow c\bar{c}$) validation analysis (Section 6.9), a separate set of BDTs are trained where the signal samples were replaced with diboson events in which generator information is used to determine that a Z-boson decayed to two charm quarks and the other vector boson decayed as $Z \rightarrow \nu\nu$, $W \rightarrow \ell\nu$, or $Z \rightarrow \ell\ell$. Additionally, due to lower statistics in the number of VZ($Z \rightarrow c\bar{c}$) MC events, only a single BDT is trained and used in the 2-lepton high- and low- $p_T(V)$ analysis channels for use in the VZ validation analysis.

The set of inputs variables used as inputs to the BDT inputs and their description can be found in Table 6.13. The variables used depend on the analysis channel, but can be broken down into the following categories:

- Kinematic properties of the H- and V-candidates
- DeepJet c -tagging discriminant for the two jets used to reconstruct H-candidate

- Kinematics of the jets, leptons, and MET used to reconstruct the V- and H-candidates
- Variables related to reconstructed top quarks – to better discern rare $t\bar{t}$ topologies faking as $W(\ell\nu)H(c\bar{c})$ events
- Variables extracted from the kinematic fit in the 2-lepton analysis channels.

BDT inputs were selected based on two criteria. The first was based on the variables ability to show different distribution shapes for $VH(H \rightarrow c\bar{c})$ signal and background events. The second was to only choose variables which showed good modeling in MC to ensure that the BDT performance will be consistent between MC simulation and data.

Variable	Description	0L	1L	2L
$m(H)$	H-candidate mass	✓	✓	✓
$p_T(H)$	H-candidate transverse momentum	—	✓	✓
$p_T(V)$	V-candidate transverse momentum	—	✓	✓
$m_T(V)$	V-candidate transverse mass	—	✓	—
p_T^{miss}	missing transverse momentum	✓	✓	—
$p_T(V)/p_T(H)$	ratio between V-candidate and H-candidate transverse momenta	✓	✓	✓
$\Delta\phi(V, H)$	azimuthal angle between vector boson and H	✓	✓	✓
CvsL_{j_1}	CvsL value of the leading Higgs jet	✓	✓	✓
CvsB_{j_1}	CvsB value of the leading Higgs jet	✓	✓	✓
CvsL_{j_2}	CvsL value of the subleading Higgs jet	✓	✓	✓
CvsB_{j_2}	CvsB value of the subleading Higgs jet	✓	✓	✓
p_{Tj_1}	p_T of the leading Higgs jet	✓	✓	✓
p_{Tj_2}	p_T of the subleading Higgs jet	✓	✓	✓
$\sigma_{cReg}(j_1)$	leading p_T jet resolution from c-jet energy regression	✓	✓	✓
$\sigma_{cReg}(j_2)$	subleading p_T jet resolution from c-jet energy regression	✓	✓	✓
$\Delta R(j_1, j_2)$	ΔR between leading and subleading Higgs jets	—	✓	✓
$\Delta\phi(j_1, j_2)$	azimuthal angle between leading and subleading Higgs jets	✓	✓	—
$\Delta\eta(j_1, j_2)$	difference in pseudorapidity between leading and subleading Higgs jets	✓	✓	✓
$\Delta\phi(\ell_1, \ell_2)$	azimuthal angle between leading and subleading p_T leptons	—	—	✓
$\Delta\eta(\ell_1, \ell_2)$	difference in pseudorapidity between leading and subleading p_T leptons	—	—	✓
$\Delta\phi(\ell_1, j_1)$	azimuthal angle between leading p_T lepton and leading Higgs jet	—	✓	—
$\Delta\phi(\ell_2, j_1)$	azimuthal angle between subleading p_T lepton and leading Higgs jet	—	—	✓
$\Delta\phi(\ell_2, j_2)$	azimuthal angle between subleading p_T lepton and subleading Higgs jet	—	—	✓
$\Delta\phi(\ell_1, p_T^{\text{miss}})$	azimuthal angle between leading p_T lepton and missing transverse momentum	—	✓	—
$\Delta\eta(\ell_1, t)$	difference in pseudorapidity between leading p_T lepton and b-tagged jet from top quark decay	—	✓	—
$\Delta\phi(\ell_1, t)$	azimuthal angle between leading p_T lepton and b-tagged jet from top quark decay	—	✓	—
$\Delta R(\ell_1, t)$	ΔR between leading p_T lepton and b-tagged jet from top quark decay	—	✓	—
CvsL_t	CvsL value of the b-tagged jet from top quark decay	—	✓	—
CvsB_t	CvsB value of the b-tagged jet from top quark decay	—	✓	—
$P(b+bb)_t$	DeepJet prob(b+bb) value of the b-tagged jet from top quark decay	—	✓	—
$m(t)$	Reconstructed top quark mass	—	✓	—
$N_{\text{small-}R}^{\text{aj}}$	Number of small- R additional jets after the FSR subtraction	—	✓	—
$\Delta\eta(V, H) _{\text{kinfit}}$	difference in pseudorapidity between vector boson and H, after kinematic-fit	—	—	✓
$\Delta\phi(V, H) _{\text{kinfit}}$	azimuthal angle between vector boson and H, after kinematic-fit	—	—	✓
$m(H) _{\text{kinfit}}$	H mass after kinematic-fit	—	—	✓
$p_T(H) _{\text{kinfit}}$	H transverse momentum after kinematic-fit	—	—	✓
$p_{T\text{max}} _{\text{kinfit}}$	p_T of the leading Higgs jet after kinematic-fit	—	—	✓
$p_{T\text{min}} _{\text{kinfit}}$	p_T of the subleading Higgs jet after kinematic-fit	—	—	✓
$p_T(V)/p_T(H) _{\text{kinfit}}$	ratio between vector boson and H transverse momenta after kinematic-fit	—	—	✓
$\sigma(H) _{\text{kinfit}}$	H invariant mass resolution from kinematic fit	—	—	✓

Table 6.13: Input variables for the BDTs trained in each of the analysis channels. The 1-lepton channel uses an additional set of variables to help discriminate against $t\bar{t}$ events, and the 2-lepton channel uses the variables extracted from the kinematic fit described in Section 6.5.3.

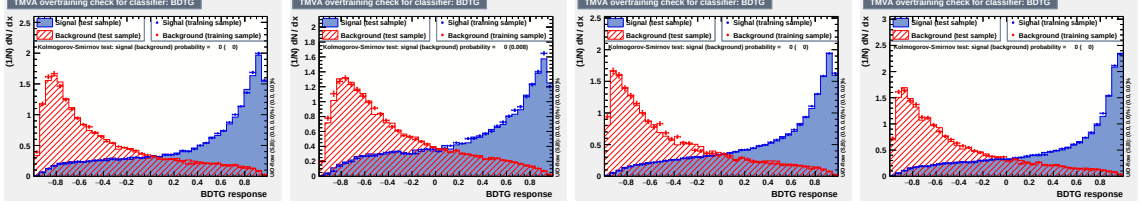


Figure 6.7.1: Overtraining curves for the BDTs train to perform $VH(H \rightarrow c\bar{c})$ signal separation in the 2018 data taking year.

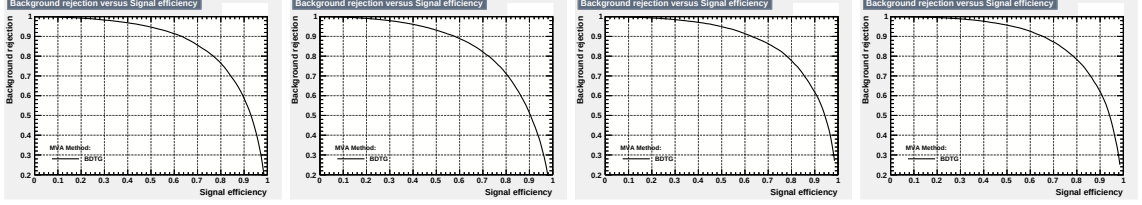


Figure 6.7.2: Receiver operator curves for the BDTs trained to perform $VH(H \rightarrow c\bar{c})$ signal separation in the 2018 data taking year.

6.7.1 Intrinsic Performance of BDTs

To verify the performance of the trained BDTs, two different methods were used. The first is an overtraining check, in which we confirm that the BDT performance is consistent between the half of the MC events used to train the BDT and the other half of the MC events. This is done by running all MC events through the BDT, and independently comparing the BDT output distributions for both signal and background test events for the two halves of the events. The second check is by looking at receiver operator characteristics (ROC). ROCs are produced by scanning a large set of potential discriminator values and determining the network's ability to correctly reject background events and correctly select signal events at each value. This produces a distribution of background rejection for a given signal efficiency value which are plotted to show the BDTs ability to simultaneously perform both discriminating tasks. The distributions of the overtraining and ROC curves for the BDTs used to perform $VH(H \rightarrow c\bar{c})$ and $VZ(Z \rightarrow c\bar{c})$ signal extraction in the 2018 data taking period can be found in Figure 6.7.1–6.7.4

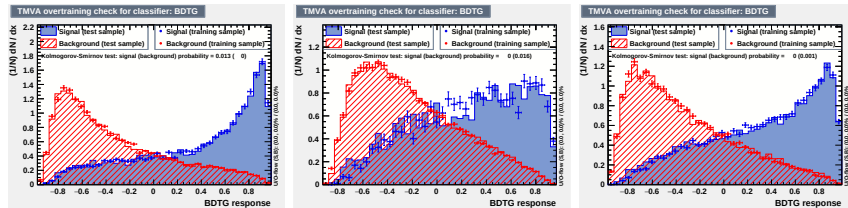


Figure 6.7.3: Overtraining curves for the BDTs train to perform $VZ(Z \rightarrow c\bar{c})$ signal separation in the 2018 data taking year.

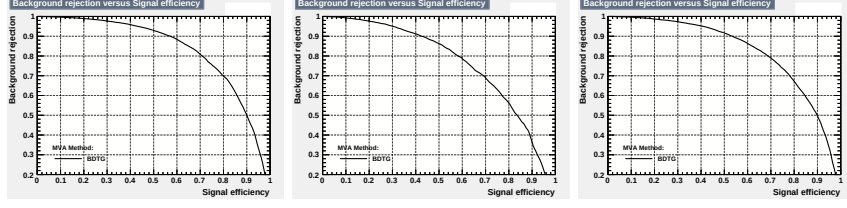


Figure 6.7.4: Receiver operator curves for the BDTs trained to perform $VZ(Z \rightarrow c\bar{c})$ signal separation in the 2018 data taking year.

6.7.2 BDT Binning and Reshaping

The gradient boosting technique of the boosted decision tree will naturally form a discriminator distributions optimized for its chosen loss functions. However, this will not necessarily be the most optimal distribution to be fed into a statistical analysis. During signal extraction, we split the distribution into bins to perform a binned likelihood fit. For each bin, the ratio of signal to background events, the relative number of signal and background events, and the relative uncertainties of the expected event yields are all important for search performance. As such, we developed a method to simultaneously reshape and rebin the distributions of our BDTs to maximize their search effectiveness.

The raw BDT outputs are first reshaped according to the function

$$\text{BDT}_{\text{new}} = \frac{1}{2} \sqrt{\left(\frac{1 + \text{BDT}}{2}\right)^x + \left(\frac{1 + \text{BDT}}{2}\right)^y} \quad (6.7.1)$$

where x and y are free parameters in the reshaping function. A parameter scan is then independently carried out for each of the trained BDTs according to the parameters in Table 6.14, in which the weighted distribution of the training events, weighted by an additional factor of 2 to account the fact that we are using half of the events, are used to calculate the expected statistical significance of each set of parameters. The significance is calculated using the negative log likelihood ratio

$$-\Gamma_{(s+b)/b} = \sum_i^{\text{bins}} \log \left(\frac{e^{-(s_i+b_i)} (s_i+b_i)^{\text{int}(s_i+b_i)}}{\text{int}(s_i+b_i)!} \right) - \sum_i^{\text{bins}} \log \left(\frac{e^{-b_i} b_i^{\text{int}(b_i)}}{\text{int}(b_i)!} \right) \quad (6.7.2)$$

in which the sums are performed over the bins are s are the number of expected signal events and b are the expected number of background events in each bin. Because the analysis is only expected to produce upper limits on the $VH(H \rightarrow c\bar{c})$ production rate, the expected number of signal events are weighted up in each year to find a set of parameters best suited for the expected analysis sensitivity. Due to the increased amount of data in the later data taking periods, $VH(H \rightarrow c\bar{c})$

Parameter	Range	Scan Width
x	[0.1, 1.1]	0.1
y	[2, 25]	1
N_{bins}	[8, 26]	1
Final Bin Width	[0.06, 0.24]	0.02

Table 6.14: The parameters and their ranges which were scanned over when performing the BDT optimization procedure. x and y refer to the free parameters in eq 6.7.1, and with the exception of the final bin all other bins remain the same width.

signal events were weighted up by a factor of 25, 20, and 15 for the 2016, 2017, and 2018 data taking years respectively. These numbers were chosen based on the expected signal strength limits that were obtained without performing the rebinning procedure. For the BDTs trained to perform $VZ(Z \rightarrow c\bar{c})$ identification, no such signal weighting was performed. Furthermore, to account for the effect of statistical uncertainties on our expected limits the selected set of parameters must additionally satisfy the following requirements: We discard any set of parameters which leaves the final two bins with a statistical uncertainty of $VH(H \rightarrow c\bar{c})$ ($VZ(Z \rightarrow c\bar{c})$) signal events greater than 10% (30%), or if any bin has an inclusive statistical uncertainty of background events greater than 30%. The set of parameters which passes the statistical uncertainty check and minimizes eq 6.7.2 are then selected and used for signal extraction.

6.8 Background Modeling

As seen in the over training plots in Figure 6.7.1 and 6.7.3, there is still background contamination present in the most sensitive bins of the BDT distributions. Furthermore, due to how rare the $VH(H \rightarrow c\bar{c})$ signal process is, improper estimations of background process event yields can propagate into large changes in our final measurement of the $VH(H \rightarrow c\bar{c})$ (and $VZ(Z \rightarrow c\bar{c})$) signal strength. To study the accuracy of our simulated MC backgrounds, we construct multiple control regions, each designed to isolate a specific type of background process. These control regions are designed to be orthogonal to our signal region, but with a large enough overlap in its selection criteria to ensure that any statements about their agreement with data can be extrapolated to the signal region.

Of the background sources mentioned in Section 6.2.2.2, the two largest sets of irreducible backgrounds are $t\bar{t}$ events in which one W-boson decays leptonically and the other decays hadronically with a high enough p_T that it is reconstructed as a single c -tagged jet, and V +jet events in which a leptonically decaying V-boson is produced with two c -tagged jets. Previous studies performed during

Background CR	CvsL _{j₁}	CvsB _{j₁}	Z($\ell\ell$)H cuts	W($\ell\nu$)H cuts	Z($\nu\nu$)H cuts
V+light	< 0.225	> 0.4	—	—	$N_{\text{aj}} < 2$
V+c	> 0.225	> 0.4	!(50 < $m_{\text{H}} < 200$)	!(50 < $m_{\text{H}} < 200$)	!(50 < $m_{\text{H}} < 200$) && $N_{\text{aj}} == 0$
V+b	> 0.225	< 0.4	80 < $m_{\text{V}} < 97$	—	$N_{\text{aj}} == 0$
t $\bar{\text{t}}$	> 0.225	< 0.4	!(75 < $m_{\text{V}} < 120$)	$N_{\text{aj}} \geq 2$ && MET < 170	$N_{\text{aj}} \geq 2$ && $\min(\Delta\phi(\text{MET}, \text{j})) < \pi/2$

Table 6.15: Additional selections used to construct the four control regions in each of the analysis channels. For all variables not mentioned, the same selections used to construct the signal regions are used.

the 2016 incarnation of the analysis found that the distribution of hadron flavor in simulated V+jet events does not perfectly model what was observed in data. To account for this flavor mismodeling, we further split the V+jet background processes into three categories:

V+b – V+jet events in which at least one of the jets are a b -jet.

V+c – V+jet events in which at least one of the jets are a c -jet, and none of the jets are b -jets.

V+light – V+jet events in which there are only light flavor jets.

For each of the three V+jet categories, simulated events are tagged as one of the three aforementioned categories using generated level information in the event about all jets passing the jet Id requirements listed in Section 6.3.2.3.

To independently study each of these four background sources, four control regions were constructed in each of the analysis channels for all of the data taking years. The general strategy is to invert the c -jet discriminator cuts to specify the desired jet flavor in the type of event we are trying to isolate, while minimizing the number of additional kinematic selections which need to be changed. The set of modified event selections used to construct the control regions can be found in Table 6.15, where any selections on variables used to construct the signal regions which were not mentioned in the table are unchanged. The motivation for the selections are the following: V+light events can be selected by inverting the CvsL cut to select events where the H-candidate is reconstructed with two light jets. V+c events can be selected by keeping the same c -tagging selections, but vetoing events where the H-candidate invariant mass is close to the Higgs peak. V+b and t $\bar{\text{t}}$ events can then be selected by inverting the CvsB cut to select events in which the H-candidate is being reconstructed by one or more b -jets. The V+b and t $\bar{\text{t}}$ events are then further separated based on the mass of the V-candidate in the 2-lepton channel, or by the number of additional jets present in the 1- and 0-lepton analysis channels

6.8.1 Construction of Control Region Shapes for Signal Extraction

In addition to using the control regions to study the modeling of the simulated MC backgrounds and construct scale factors (Section 6.4.1 and 6.4.2), the distribution of the sub-leading jet's CvsB score (CvsB_{j_2}) from each of the control regions are used to constrain our dominant background sources during signal extraction. This is done by keeping the normalization of $t\bar{t}$, $V+\text{light}$, $V+c$, and $V+b$ processes floating as rate parameters while fitting the $VH(H \rightarrow c\bar{c})$ or $VZ(Z \rightarrow c\bar{c})$ signal strengths (Section 6.12). By doing so, we can improve our ability to accurately model the background event yields in the most sensitive parts of our signal region. In particular by leaving the flavor-split $V+\text{jet}$ normalizations floating, we are able to exploit the c -tagger values in our BDTs without our results showing a systematic bias due to the mismodeling in $V+\text{jet}$ hadron flavor present in the MC samples.

During signal extraction, we then constrain the floating normalizations as follows: The normalization of $Z(\ell\ell)+\text{jet}$ events are extracted from the control region distributions of the 2-lepton channel, with a separate set of normalizations used for the high- and low- $p_T(V)$ events. The normalization of $W(\ell\nu)+\text{jet}$ events are extracted from the control regions distributions of the 1-lepton channel. Due to the similarity of the CvsB_{j_2} distributions of single top and $W+b$ events and the low contamination of $W+b$ events in the most sensitive bins of the 1-lepton channel BDT, the control region designed to isolate $W+b$ events is dropped from the fit and the $W+b$ process share their normalization with the $W+c$ rate parameter. The normalization of $Z(\nu\nu)+\text{jet}$ events are extracted from the control region distributions of the 0-lepton channel. Additionally, the $W+\text{light}$, $W+b$, and $W+c$ normalizations extracted from the 1-lepton channel are applied to $W+\text{jet}$ events in the 0-lepton channel. A unique set of normalizations are used in each of the three data taking years. This is because of changes in the simulations, reconstruction efficiencies, and detector conditions which vary between the 2016, 2017, and 2018.

6.9 VZ Validation

To ensure confidence in our results, a validation analysis is performed in which we carry out the same analysis procedure with the aim of measuring the $VZ(Z \rightarrow c\bar{c})$ signal strength. The $VZ(Z \rightarrow c\bar{c})$ and $VH(H \rightarrow c\bar{c})$ processes both have very similar kinematics and topologies, with the primary differences stemming from the mass difference between the Higgs and Z-bosons. Furthermore, the production cross section of the ZZ and WZ processes at the LHC are well modeled [99], and the $Z \rightarrow c\bar{c}$ branching ratio has been well measured in past e^+e^- collider experiments [108, 109, 110, 111, 112].

Thus, we can inspire confidence in our analysis framework by measuring the rate of the $VZ(Z \rightarrow c\bar{c})$ standard model process, and obtaining a result that is consistent with the SM.

The $VZ(Z \rightarrow c\bar{c})$ validation procedure closely emulates the $VH(H \rightarrow c\bar{c})$ framework, in which multiple analysis channels are produced to target the leptonic final states of the second V-boson produced with the Z decaying to charm quarks. There are however two main changes between the two analysis. The first, due to kinematic differences between the $VZ(Z \rightarrow c\bar{c})$ and $VH(H \rightarrow c\bar{c})$ processes, is that the BDTs used to perform signal separation in the signal regions are retrained with MC $VZ(Z \rightarrow c\bar{c})$ events used as the signal. Due to decreases in the number of simulated events, the 2-lepton $VZ(Z \rightarrow c\bar{c})$ analysis channels share a common BDT which is trained using events from the high- and low- $p_T(V)$ analysis channels. The second difference is that despite using the same set of events in the signal and control regions, the final frequentist fit used to perform the $VZ(Z \rightarrow c\bar{c})$ signal extraction is performed independent of the $VH(H \rightarrow c\bar{c})$ fit. This means that the final set of nuisance and rate parameters obtained in the VZ validation will vary slightly compared to the $VH(H \rightarrow c\bar{c})$ analysis.

6.10 Combination with Merged Jet Analysis

We expect approximately 95% of $VH(H \rightarrow c\bar{c})$ events at the LHC to produce a Higgs boson with a low enough p_T such that the two charm quarks hadronize into two distinct jets. The analysis strategy described in this thesis is designed to exploit the unique features of this two resolved jet topology. The other 5% of signal events instead produce a topology in which the two jets are at least suitably merged, to the point that they can be clustered together to form a single large-radius jet. This topology exhibits a few highly exploitable features: Primarily a much lower relative contamination of background events which emulate the topology, increased probability of reconstructing a Higgs boson (since there are no combinatorics associated with correctly selecting one jet but failing to select the other), and the ability to perform mass regression algorithms directly on the singular H-candidate jet. Because of these unique features, collaborators within CMS performed a second $VH(H \rightarrow c\bar{c})$ analysis constructed to specifically focus on these high $p_T(H)$ events. To maximize the collaborations ability to measure the $VH(H \rightarrow c\bar{c})$ (and $VZ(Z \rightarrow c\bar{c})$) signal strength, the two analyses are then combined during signal extraction.

6.10.1 VH($H \rightarrow c\bar{c}$) Merged Jet Analysis Overview

The VH($H \rightarrow c\bar{c}$) merged jet analysis is based on the idea that the two hadronizing charm quarks will produce two overlapping or very nearby particle showers. The analysis then attempts to cluster both of the showers into a single AK15 jet (anti-kt jet with a radius parameter of 1.5), and then reconstructs the H-candidate as the highest p_T AK15 jet in the event. A special version of the ParticleNet tagger [78] trained on AK15 jets is used to perform cc -tagging on the AK15 jet, and events are categorized as either high, medium, or low purity based on the cc -tagging discriminator value of the H-candidate. Events passing the baseline selection are then passed through a kinematic BDT trained on a set of flavor independent variables (no variables intrinsic to jet tagging or jet mass) to discriminate between signal-like from background-like events. A cut is then placed on the BDT discriminator to split selected events into a signal region and a V+jets dominated control region. An additional control region is then constructed by selecting events with a large number of additional AK4 jets to isolate $t\bar{t}$ events. Finally, a mass regression algorithm using the same graph neural network architecture as ParticleNet is applied to the H-candidate in each event, and the regressed mass distribution in the control and signal regions for each of the three cc -jet flavor purity categories are used for signal extraction.

6.10.2 Combination of Resolved and Merged Jet Analysis

When constructing the full Run 2 iterations of the two analyses, we aimed to keep as many of their aspects identical as possible. As such, all MC samples and datasets (Section 6.2), triggers and efficiencies (Section 6.2), object Id requirements (Section 6.3), vector boson reconstruction procedures (Section 6.4), and signal region event selections (Section 6.6) are kept constant between the two analysis. The primary exceptions are the use of $CvsL_{j_1}$ and $CvsB_{j_1}$ to build the signal region, which is not used for the merged jet analysis, the triggers for the 2-lepton channel, in which the merged jet analysis uses the same single lepton triggers used in the 1-lepton channel, and the additional scale factors applied based on the V-boson kinematics (Section 6.4.2 and 6.4.1) which were found to not be needed for the high momentum regime of the merged analysis. This matching of analysis procedures is to ensure that we are able to maximally correlate systematic uncertainties between the two analysis, which is described further in Section 6.11.

To ensure that there are no events which are included in both analyses, effectively leading to a double counting of events which could underestimate the statistical precision, an additional orthogonality requirement is placed on each of the analyses. The choice of orthogonality is as follows:

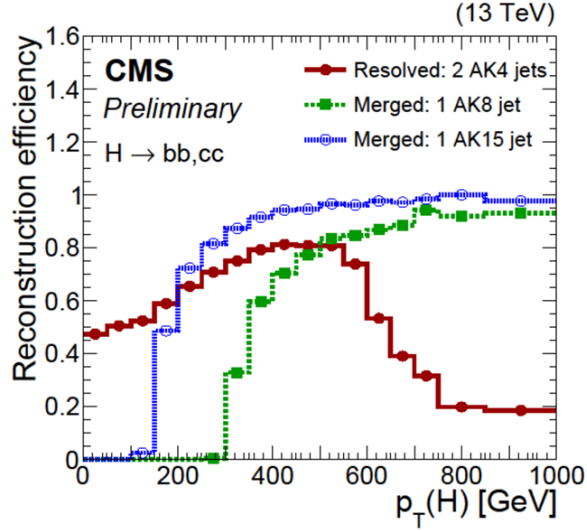


Figure 6.10.1: Reconstruction efficiency of different ways to reconstruct a Higgs boson. At p_T greater than 300 GeV, the reconstruction efficiency for a H-candidate reconstructed from the highest p_T AK15 jet in an event approaches unity.

All events with an AK15 jet with $p_T > 300$ GeV and $|\eta| < 2.4$ are selected for use in the merged jet analysis. All events with no AK15 jets, or an AK15 jet with $p_T < 300$ GeV or $|\eta| > 2.4$ are selected for use in the resolved jet analysis. The decision to base the orthogonality condition on a reconstructed AK15 jet ensures that there are no instances in which a high $p_T(H)$ event is lost. This would otherwise happen if there were a high enough momentum Higgs boson to match the description of a merged Higgs event, but for some reason there are no AK15 jets reconstructed in an event. The choice to set the p_T selection at 300 GeV was based on a scan of expected limits, where it was found that the combined analysis sensitivity was maximized by cuts placed in the vicinity of 300 GeV, while simultaneously maintaining a low negative impact on the standalone performance of either of the two analysis (due to the loss of phase space). This is partially exemplified in Figure 6.10.1 showing that at $p_T(H)$ around 300 GeV the H-candidate reconstruction efficiency becomes much larger using a single AK15 jet.

6.11 Systematic Uncertainties

In performing the analysis, we make a great effort to account for as many systematic uncertainties as possible. The systematics are primarily categorized into those attributed to theoretical uncertainties affecting our MC samples' ability to properly model our data, experimental uncertainties related to our ability to model the performance of the CMS detector as well as our analysis techniques, and

Systematic		Correlation				
Type	Label	Topology	Channel	Processes	Year	Type
Theoretical	VH Cross Sections	✓	✓	—	✓	rate
	VV and Single Top Cross Sections	✓	✓	—	✓	rate
	Higgs Branching Fractions	✓	✓	—	✓	rate
	PDF Set	✓	✓	—	Partially	rate
	Renormalization Scale	✓	✓	—	Partially	shape
	Factorization Scale	✓	✓	—	Partially	shape
	VH p_T Reweighting	✓	✓	—	✓	shape
	Drell-Yan p_T Reweighting	—	✓	—	—	shape
Experimental	V+jet $\Delta R(jj)$ Reweighting	—	✓	—	—	shape
	Luminosity	✓	✓	✓	—	rate
	Jet Energy Resoltuion	✓	✓	✓	—	shape
	Jet Energy Scale	✓	✓	✓	Partially	shape
	Lepton Efficiencies	✓	—	✓	—	shape
	MET Trigger Efficiencies	✓	—	✓	—	shape
	ECAL Prefiring	✓	✓	✓	—	shape
	HEM15/16 Failure	✓	✓	✓	—	shape
c -tagging	Pileup Modeling	✓	✓	✓	—	shape
	c -tagging scale factors	—	✓	✓	Partially	shape
	Post c -tag Renormalization	—	✓	—	—	shape
	c -jet Regression	—	✓	✓	—	shape

Table 6.16: Overview of the systematic uncertainties considered in the analysis, how they are correlated across the topologies, analysis channels, physics processes, and data taking years, as well as whether we considered the uncertainty to effect the overall shape of the distribution or only present itself in an overall change in event normalization (rate).

those attributed to c -tagging (or cc -tagging in the case of the merged jet analysis) and the application of the energy/mass regression techniques. All systematic uncertainties are associated with our ability to properly simulate the physics taking place at the LHC, and thus are represented as ways in which the expected MC event distributions in our control and signal regions may have changed as a result of the potential systematics. We thus represent systematics in two forms: *rate* uncertainties represented by an overall change in the event yield normalization for a given process (uniformly across a set of processes), and *shape* uncertainties represented by a change in the overall distribution of expected event yields in our control and signal regions. In the case of shape uncertainties leading to differences in the reconstructed event or object kinematics, the BDT is re-run to produce a new signal region distribution. Because having larger numbers of unique systematics will naturally lead to a greater uncertainty of the overall analysis, we attempt to correlate systematics across topology, analysis channel, physics process being modeled, or data taking year wherever possible. The types of systematics used in this analysis, as well as the ways in which they are correlated are shown in Table 6.16. A more in depth description of each of the sources of systematic uncertainties considered in the analysis are listed as follows:

VH Cross Sections – The effect of PDF, renormalization scale, and factorization scale are con-

sidered indepdently for the VH process than for the rest of the samples. This is modeled by assuming a normalization uncertainty of $^{+5.5\%}_{-2\%}$ in the VH signal event yield

VV and Single Top Cross Sections – The cross sections of diboson and single top processes are obtained from measurements performed at the LHC. As such, a normalization uncertainty of 15% in each of their overall expected event yields are assumed to account for the uncertainty in these measurements.

Higgs Branching Fractions – A 5% normalization uncertainty is assumed for the $\mathcal{B}(H \rightarrow bb)$ and $\mathcal{B}(H \rightarrow cc)$ branching fractions.

PDF Sets – A 1–3% normalization uncertainty is assumed based on the PDF set used when simulating various standard model processes. Uncertainty is considered to be correlated between identical processes in 2017 and 2018 because of the shared PDF sets used when simulating events.

Renormalization and Factorisation Scales – A shape uncertainty is used to account for systematics arising from the imperfect renormalization and factorization scales used when simulating events. Changes in distributions are obtained by adjusting factors associated with the normalization and factorization scales and adjusting the event weights accordingly. This is correlated in 2017 and 2018 where simulated samples were obtained using the same assumed renormalization and factorization scales.

VH p_T Reweighting – Uncertainties on the function used to calculate the VH scale factors are obtained by varying the parameters in the function by their respective uncertainties and observing the degree to which they change their corresponding event weight. The new event weights are then applied to produce a new set of event distributions.

Drell-Yan p_T Reweighting – Uncertainties associated with the function used to correct the mis-modeled $p_T(Z)$ distribution in the 2-lepton channel are obtained by varying the fit parameters by their respective uncertainty and observing how they vary their corresponding scale factor. The new scale are then applied to produce a new set of event distributions.

V+jet $\Delta R(jj)$ Reweighting – An uncertainty on the functions used to correct the $\Delta R(jj)$ mis-modeling in V+jet samples are obtained by varying the fit parameters by their respective uncertainties and observing how the the corresponding scale factors vary. The new scale are

then applied to produce a new set of event distributions. An additional set of up/down variation in the yield of V+c and V+b events with $\Delta R(jj) < 0.6$ are applied to account for the relationship between $\Delta R(jj)$ and jet flavor for highly merged jets originating from gluon splitting. The uncertainties assigned are: $\pm 20\%$ for Z + c events, $\pm 35\%$ for Z + b events, $\pm 30\%$ for W + c events, and $\pm 30\%$ for W + b events.

Luminosity – A 2.5%, 2.3%, and 2.5% uncertainty in the event normalization is assigned to events simulated for the 2016, 2017, and 2018 data taking years to account for the uncertainty in the total luminosity received by the CMS detector.

Jet Energy Resolution – Uncertainties in the way we smear jets to properly model their energy resolution in the CMS detector are accounted for by adjusting the smearing parameters within their calculated uncertainties. The analysis procedures are then re-run with the new set of jet smearing applied, and the difference in bin content across all distributions are used to encapsulate the effect of the systematic.

Jet Energy Scale – Uncertainties in the jet energy corrections are found by independently varying 1 of 11 sources of jet energy identified by the JetMET group of the CMS collaboration and re-running the analysis with the new set of jet energies. The analysis procedures are then re-run with the new set of jet energies, and the difference in bin content across all distributions are used to encapsulate the effect of the systematic. The following JEC sources are considered to be uncorrelated across years: Absolute energy scale, BBEC1, EC2, HF, and RelativeSample. The uncertainties associated with the following JEC sources were then treated as correlated accross years: Absolute energy scale, BBEC1, EC2, QCD Flavor, HF, and RelativeBal. In the case of sources listed twice, they were found to be partially correlated, and the sysyetmatics were deconstructed into correlated and uncorrelated components.

Lepton Efficiencies – An uncertainty is applied to encapsulate the potential systematic differences in trigger, reconstruction, identification, and isolation efficiencies in data and in simulation. Due to differences in triggers and identification requirements between analysis channels, as well as changes in the tracker between data taking years, lepton efficiencies are set to be uncorrelated between channel and year.

MET Triggers – An uncertainty in the normalization of the expected event yields in the 0-lepton channel at the rate of 2% is applied to account for potential systematic uncertainties of the

MET trigger efficiencies. Due to changes in trigger and detector conditions in each of the data taking years, MET efficiencies are set to be uncorrelated between channel and year.

ECAL Prefiring – An additional uncertainty is assigned to the 2016 and 2017 years to account for the L1 prefiring issues described in Section 6.3. Uncertainty was obtained by varying the pre-fire probability for a given event by their corresponding uncertainties and propagating the value to the event weight. Despite the problem existing in both years, any systematic differences between the assumed and true pre-firing probabilities are assumed to be uncorrelated between years.

HEM15/16 Issue – In addition to the method described in Section 6.3.4.2 to veto events with fake MET originating from the HEM hole in the 2018 0-lepton channel, an additional jet energy uncertainty is applied across all 2018 analysis channels. The energy of jets with their central axis being reconstructed in the following ranges of the detector have their energies scaled down by the following amounts: 20% for jets within $-1.57 < \phi < -0.87$ and $-2.5 < \eta < -1.3$ and 35% for jets within $-1.57 < \phi < -0.87$ and $-3.0 < \eta < -2.5$. The analysis is then re-run with the new set of jet energies, and the difference in event yields in each of the control and signal region bins are extracted as the effect of the uncertainty.

Pileup Modeling – Two separate sets of uncertainties are used to account for differences in our estimations of pileup and their true values. An uncertainty in our ability to properly model the number of pileup interactions is obtained by varying the pileup cross section by 4.6% and propagating the change to the pileup reweighting procedure. The second is through the variation in a set of scale factors associated with a pileup jet identification method used to veto pileup jets. In both cases, the modified weights and scale factors are independently applied to the MC events and the new distributions are taken to be the effect of the uncertainties.

c -tagging Scale Factors – As described in Section 5.4.3 additional sets of c -tagging scale factors are calculated by re-running the iterative fit procedure after independently propagating theoretical (factorization and renormalization scales), experimental (pileup modeling, jet energy scale, jet energy resolution), and statistical uncertainties to the event yields in each of the CvsL×CvsB bins. The effect of these uncertainties on the scale factor derivation are then propagated to the final analysis by re-calculating the event level c -tagging weights with the alternative sets of scale factors, and finding the changes in control region and signal region distributions. A separate set of cc -tagging uncertainties are applied to the distribution in the

merged jet analysis based on the uncertainties in their scale factor derivation.

Post c -tagging Renormalization – The application of c -tagging scale factors are designed to have no effect on the overall event yield of MC events. However, due to the phase space in which scale factors were derived being different from that of the signal and control regions used in the analysis, a residual change in event yields are expected in the analysis. The event yields before and after application of the c -tagging weights are calculated in each of the signal and control regions as a function of number of jets and pileup interactions, and used to quantify how the overall distribution of events would change after restoring the normalization.

c -jet Regression – As described in Section 6.5.2 a dedicated smearing has to be applied to the regressed c -jets after application of the regression algorithm. A separate jet energy resolution uncertainty following the same procedure described above is then calculated for the regressed c -jets. In the case of the merged jet analysis, a separate uncertainty is applied related to the efficiency of the mass regressor applied to the merged H-candidate.

During signal extraction, the systematic uncertainties described in this section produce a set of uncertainty bands around the expected MC yields in our distributions. They are treated under the assumption that each band represents a $\pm 1\sigma$ standard deviation about the nominal expected event yield attributed to the presence of the uncertainty source. The uncertainty bands are then treated as nuisance parameters in the final frequentist fit, where expected yields are able to fluctuate in accordance with their various sources of systematic uncertainties. Additionally, statistical uncertainties for each signal and background process are calculated on bin-by-bin basis as $\pm \sqrt{\sum w_i^2}$ with respect to the nominal event yield, where w_i is the event weight associated with the i -th event for a given process in a given bin. Expected event yields are able to vary within these statistical uncertainty bands in the same way as the modeled systematic uncertainties, and they are treated as being purely uncorrelated between bins.

6.12 Results

Measurements on the $VH(H \rightarrow c\bar{c})$ and $VZ(Z \rightarrow c\bar{c})$ cross sections are based on the calculation of likelihood values originating from a frequentist binned counting experiment. The binned likelihood can be calculated as

$$\mathcal{L}(\text{data}|\mu, \theta) = \prod_i^{\text{bins}} (\text{Poisson}(N_i|\mu \cdot s_i(\theta) + b_i(\theta))) \cdot p(\tilde{\theta}|\theta) \quad (6.12.1)$$

where N_i is the data yield in the i -th bin, s_i and b_i are the expected signal and background yields obtained from our MC simulation in the i -th bin, $\mu = \sigma/\sigma_{\text{SM}}$ is the signal strength used to modify the overall signal yield, and θ is the package of nuisance parameters corresponding to the systematic and statistical uncertainties of our expected event yields which have each settled to some value. The function $p(\tilde{\theta}|\theta)$ then represents the probability associated with the set of nuisance parameters fitting to a set of values θ given their respective uncertainty bands. Additionally, the values of the floating rate parameters are encoded in $b_i(\theta)$ and the probabilities associated with the fit rate parameters are contained within $p(\tilde{\theta}|\theta)$. eq 6.12.1 then tells how well our MC simulation models what we see in data given a chosen signal strength μ and set of nuisance parameters θ . From this, we define the set of signal strength and nuisance parameters $\hat{\mu}$ and $\hat{\theta}$ representing the set of best fit parameter which will maximize the likelihood function.

From our likelihood function and maximum likelihood function, we can scan over various fixed values of μ and obtain the profile likelihood test statistic

$$\bar{q}_\mu = -2 \ln \frac{\mathcal{L}(\text{data}|\mu, \hat{\theta}_\mu)}{\mathcal{L}(\text{data}|\hat{\mu}, \hat{\theta})} \quad (6.12.2)$$

where $\hat{\theta}_\mu$ is the set of nuisance parameters which maximize the likelihood for a chosen signal strength μ . In essence, $\bar{q}_\mu$ will be a positive value which becomes larger the further μ is from the corresponding best fit value. From $\bar{q}_\mu$ we can construct two separate values in order to obtain the finalized results of our analysis. In the event that the analysis is not yet sensitive enough to measure the signal, we scan over various values of $\bar{q}_\mu$ and for each of them calculate the confidence level

$$\text{CL}(\mu) = \frac{\bar{q}_\mu}{\bar{q}_0} \quad (6.12.3)$$

where $\bar{q}_0$ is the test statistic in the assumption that the signal does not exist ($\mu = 0$). The confidence limit thus represents the probability that we would have successfully measured an excess of data over our expected background distribution if the signal were to exist with the corresponding signal strength. During analysis we quote an upper limit at the 95% confidence level, representing the signal strength at which we measure a confidence level of $\text{CL}(\mu) = 0.05$.

In the event that our analysis is sensitive enough to measure $\hat{\mu}$ with a high confidence level, we can treat $\bar{q}_0$ as the likelihood ratio between our signal existing with signal strength $\hat{\mu}$ and the signal not existing. In the assumption that the signal does not exist, it implies any excess of data events beyond expected background amounts are due to fluctuations within the statistical and systematic

uncertainties. From $\bar{q}_0$ we can then directly compute a p -value

$$p = \int_{\bar{q}_0}^{\infty} p(\bar{q}_\mu|0) d\bar{q}_\mu \quad (6.12.4)$$

where $p(\bar{q}_\mu|0)$ has the value

$$p(\bar{q}_\mu|0) = \frac{1}{2} \frac{e^{-\bar{q}_\mu/2}}{\sqrt{2\bar{q}_\mu}} \quad (6.12.5)$$

for values of $\bar{q}_0 > 0$, which is always true for a best fit signal strength of $\hat{\mu} > 0$, the p -value represents the probability in which one would measured the best-fit signal strength $\hat{\mu}$ under a background only hypothesis. In particular, it gives a $p\%$ probability for our measurement to be due to a null signal hypothesis experiencing natural statistical fluctuations. In particle physics, a p -value of 0.003 is the requirement to claim that one has measured evidence for a process, and a p -value of 3×10^{-7} is the requirement to claim observation of a process. There is a one to one mapping between the p -value, the distance between a measured signal strength and the null hypothesis on a Gaussian distribution. and describing how far away the signal strength must be from the mean of a Gaussian centered at 0. Because of this relationship, we often refer to the p -value in terms of the numbers of standard deviations (σ) from $\hat{\mu}$ that the $\mu = 0$ hypothesis must live to have not been excluded from the analysis. For a claim of evidence or observation, the $\mu = 0$ signal strength must respectively live past 3σ and 5σ .

6.12.1 Standalone Results

Before combining with the merged analysis, the frequentist fit approach described above were independently run on the resolved jet analysis. For the $\text{VH}(\text{H} \rightarrow c\bar{c})$ analysis, they were run on the CvsB_{j_2} distributions in the control regions and the reshaped BDT distributions trained with the $\text{VH}(\text{H} \rightarrow c\bar{c})$ signal events in the signal regions. The $\text{VZ}(\text{Z} \rightarrow c\bar{c})$ analysis was then run over the same set of control regions, but with a separate set of reshaped BDTs trained with $\text{VZ}(\text{Z} \rightarrow c\bar{c})$ events as signal in the signal regions. To remove any possible biases from running our likelihood fit over the events used to train our BDT, we only fill the signal region distributions with odd numbered MC events and increase their event weights by a factor of 2.

6.12.1.1 VH($H \rightarrow c\bar{c}$) Results

Due to the small cross sections of the $VH(H \rightarrow c\bar{c})$ signal, the analysis was carried out to construct upper limits at a 95% confidence level on the $VH(H \rightarrow c\bar{c})$ signal strength using the prescriptions described earlier in this chapter. To study the strengths of the individual analysis channels and data taking years, the fit procedure was performed both individually and simultaneously to each of the analysis channels and data taking years. When only performing the fit to the 2-lepton channel, the fit is still performed simultaneously to the low- and high- $p_T(V)$ channels. The full set of 2018 control region distributions using the best fit nuisance and rate parameters obtained from the full Run 2 fit to all analysis channels and data taking years can be found in Figure 6.12.1 and 6.12.2. The set of rate parameters used to normalize the most dominant backgrounds can be found in Table 6.18. From the fit, we see agreement between the data and expected event yields are close to unity in all of our control region bins.

The full set of expected and observed limits for each combination of analysis channels and data taking years can be found in Table 6.17. Figure 6.12.3 then shows the full set of 2018 signal region distributions with the best fit nuisance parameters and signal strength from the full Run 2 fit to all analysis channels applied. The $VH(H \rightarrow c\bar{c})$ resolved jet full Run 2 analysis measures an observed (expected) 95% CL upper limit on the signal strength of 14 (19) times the standard model value with an observed (expected) best fit signal strength of $\hat{\mu}_{VH(H \rightarrow c\bar{c})} = -9.50^{+9.94}_{-9.54}$ ($1.00^{+9.53}_{-9.73}$). This corresponds to a factor of 5 (2) improvement in the observed (expected) limits with respect to the combined resolved+merged performance of the 2016 analysis [39]. Additionally, our observed (expected) upper limits show a factor of 1.9 (1.6) improvement over ATLAS's full Run 2 $VH(H \rightarrow c\bar{c})$ analysis [113]. In both instances, comparisons have been made after cutting away a portion of our signal from the application of the AK15 $p_T > 300$ GeV orthogonality cut, which is expected to decrease the sensitivity of the standalone resolved jet analysis by approximately 35%.

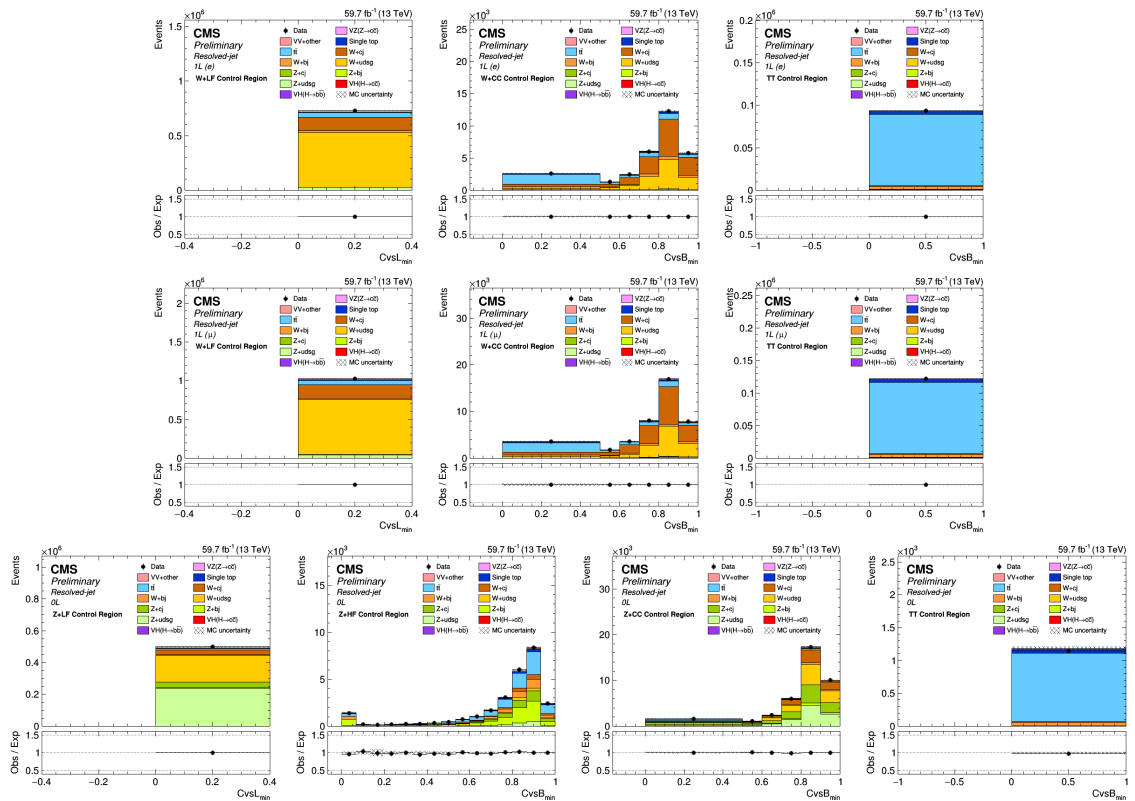
Table 6.19 shows the relative contribution that our largest systematics have on the overall uncertainty of our best fit $VH(H \rightarrow c\bar{c})$ signal strength. Overall, we find that the two largest sources of uncertainties comes from the statistical uncertainty associated with the number of data events in each bin, as well as the uncertainties associated with the finite size of the simulated MC samples. Ideally, both of these sources will be naturally improved in potential Run 3 and HL-LHC iterations of the analysis. The maximum search sensitivity is limited by luminosity, which is expected to be increased by a factor of 2 for Run 3 and 20 for the HL-LHC. Because of the increased statistics in data, we can also expect a correlated increase the MC event generation. Hopefully, the always in-

		0-lepton	1-lepton	2-lepton	combined
2016	SR+CR fit (Exp.)	72	87	41	34
	SR+CR fit (Obs.)	58	63	46	28
2017	SR+CR fit (Exp.)	63	73	40	33
	SR+CR fit (Obs.)	39	71	40	24
2018	SR+CR fit (Exp.)	65	67	43	33
	SR+CR fit (Obs.)	56	50	46	27
Full Run 2	SR+CR fit (Exp.)	38	43	23	19
	SR+CR fit (Obs.)	25	33	23	14

Table 6.17: Expected and observed 95% CL limits for the VH production cross section times $H \rightarrow c\bar{c}$ branching ratio signal strength, for the various analysis channels, data taking years, and the full Run 2 fit.

	Year	Z+c	Z+b	Z+light	W+c	W+b	W+light	$t\bar{t}$
2-lepton high- $p_T(V)$	2016	0.88 ± 0.07	0.96 ± 0.06	1.02 ± 0.05	—	—	—	0.91 ± 0.05
	2017	1.12 ± 0.08	1.20 ± 0.08	0.92 ± 0.05	—	—	—	0.95 ± 0.06
	2018	1.20 ± 0.10	1.22 ± 0.10	0.94 ± 0.06	—	—	—	0.98 ± 0.06
2-lepton low- $p_T(V)$	2016	0.79 ± 0.06	0.83 ± 0.05	1.03 ± 0.05	—	—	—	0.87 ± 0.04
	2017	0.97 ± 0.07	0.94 ± 0.06	0.97 ± 0.05	—	—	—	0.90 ± 0.05
	2018	1.19 ± 0.14	1.02 ± 0.08	0.98 ± 0.09	—	—	—	0.93 ± 0.05
1-lepton	2016	—	—	—	0.93 ± 0.06	—	1.03 ± 0.05	0.93 ± 0.05
	2017	—	—	—	1.03 ± 0.07	—	1.03 ± 0.06	1.09 ± 0.05
	2018	—	—	—	1.22 ± 0.08	—	0.91 ± 0.05	1.05 ± 0.05
0-lepton	2016	1.00 ± 0.13	1.12 ± 0.19	1.27 ± 0.09	0.93 ± 0.06	—	1.03 ± 0.05	0.83 ± 0.05
	2017	1.13 ± 0.24	1.07 ± 0.20	0.75 ± 0.12	1.03 ± 0.07	—	1.03 ± 0.06	1.04 ± 0.06
	2018	1.03 ± 0.20	0.75 ± 0.15	0.76 ± 0.11	1.22 ± 0.08	—	0.91 ± 0.05	0.99 ± 0.06

Table 6.18: Background normalization rate parameters derived from the $VH(H \rightarrow c\bar{c})$ full Run 2 fit to all analysis channels.



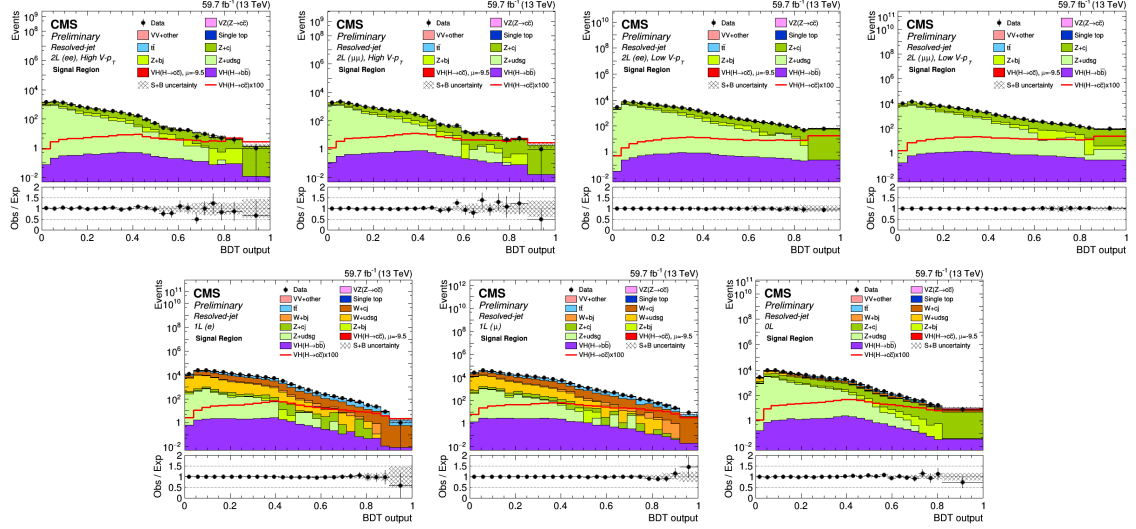


Figure 6.12.3: Signal regions used to extract the best fit $VH(H \rightarrow c\bar{c})$ signal strength for the 2018 data taking year. Nuisance parameters and signal strength applied to MC event yields and uncertainty bands were obtained from the full Run 2 fit to all analysis channels. Because the best fit signal strength was measured to be $\hat{\mu} = -9.50^{+9.94}_{-9.54}$, the signal yield is set to 0.

creasing performance of leading edge computational hardware will continue to improve the feasibility generating higher and higher numbers of MC events.

6.12.1.2 $VZ(Z \rightarrow c\bar{c})$ Results

Measurement of the $VZ(Z \rightarrow c\bar{c})$ process is expected to be within the sensitivity of our analysis, as such the p -value test described previously in this chapter is used to calculate measurement significance in units of σ . Because we used the same set of control regions as the $VH(H \rightarrow c\bar{c})$ analysis, we expect the fit to settle on a similar set of rate parameters. This can be seen in Table 6.18 and

Uncertainty Source	$\Delta\mu/(\Delta\mu)_{\text{tot}}$
Statistical	66%
Background normalizations	28%
Experimental	72%
Sizes of the simulated samples	59%
Jet energy scale and resolution	17%
Charm identification efficiencies	27%
Simulation modeling	20%
Luminosity	13%
Lepton identification efficiencies	10%
Theory	22%
Backgrounds	21%
Signal	7%

Table 6.19: Relative contribution to the uncertainties in the $\hat{\mu}_{VH(H \rightarrow c\bar{c})}$ signal strength from the standalone resolved jet analysis.

6.21 where the differences in rate parameters between the two fits all exist within their respective 1σ uncertainty ranges. The standalone performance of the resolved jet $VZ(Z \rightarrow c\bar{c})$ analysis can be seen in Table 6.20 for different combinations of analysis channels and data taking years as well as the full Run 2 fit. The 2018 $VZ(Z \rightarrow c\bar{c})$ signal region distributions with the respective signal strength and nuisance parameters extracted from the full Run 2 fit to all analysis channels can be found in Figure 6.12.3.

The full Run 2 $VZ(Z \rightarrow c\bar{c})$ analysis measured an observed (expected) best-fit $VZ(Z \rightarrow c\bar{c})$ signal strength of $\hat{\mu}_{VZ(Z \rightarrow c\bar{c})} = 0.96_{-0.32}^{+0.34}$ ($1.00_{-0.32}^{+0.34}$) at a significance of 3.1σ (3.3σ). This brings us to evidence of the $Z \rightarrow c\bar{c}$ decay mode at a hadron collider at a production rate that is in excellent agreement with the standard model. In addition to a strong result, the close to unity observed signal strength demonstrates that the analysis technique performs as expected. The validation of the analysis techniques through measuring a $\hat{\mu}_{VZ(Z \rightarrow c\bar{c})}$ value consistent with unity with similar nuisance parameters to those obtained from the $VH(H \rightarrow c\bar{c})$ fit indicates that our measurement of the $VH(H \rightarrow c\bar{c})$ signal strength is correct, given the statistical and systematic uncertainties. This is further exemplified with the fact that $\hat{\mu}_{VH(H \rightarrow c\bar{c})}$ is in agreement with the standard model expectations within 1.1σ of the measured value.

		0-lepton	1-lepton	2-lepton	combined
2016	SR+CR fit (Exp.)	0.7σ	0.5σ	1.6σ	1.7σ
	SR+CR fit (Obs.)	2.0σ	0.0σ	1.0σ	1.8σ
2017	SR+CR fit (Exp.)	1.2σ	0.7σ	1.8σ	2.2σ
	SR+CR fit (Obs.)	0.6σ	0.2σ	1.5σ	1.7σ
2018	SR+CR fit (Exp.)	0.8σ	0.6σ	1.3σ	1.7σ
	SR+CR fit (Obs.)	0.0σ	1.6σ	0.6σ	1.9σ
Full Run2	SR+CR fit (Exp.)	1.6σ	1.1σ	2.7σ	3.3σ
	SR+CR fit (Obs.)	1.3σ	1.4σ	1.8σ	3.1σ

Table 6.20: Expected and observed significance of the best fit $VZ(Z \rightarrow c\bar{c})$ signal strength in the various analysis channels and data taking years..

	Year	Z+c	Z+b	Z+light	W+c	W+b	W+light	t $\bar{t}$
2-lepton high- p_T (V)	2016	0.88 ± 0.07	0.96 ± 0.06	1.02 ± 0.04	—	—	—	0.91 ± 0.05
	2017	1.11 ± 0.08	1.20 ± 0.08	0.92 ± 0.05	—	—	—	0.95 ± 0.06
	2018	1.20 ± 0.09	1.22 ± 0.09	0.94 ± 0.06	—	—	—	0.98 ± 0.06
2-lepton low- p_T (V)	2016	0.79 ± 0.06	0.83 ± 0.05	1.03 ± 0.05	—	—	—	0.87 ± 0.04
	2017	0.97 ± 0.07	0.94 ± 0.06	0.97 ± 0.06	—	—	—	0.90 ± 0.05
	2018	1.19 ± 0.14	1.02 ± 0.08	0.94 ± 0.05	—	—	—	0.93 ± 0.05
1-lepton	2016	—	—	—	0.95 ± 0.06	—	1.02 ± 0.04	0.93 ± 0.05
	2017	—	—	—	1.03 ± 0.07	—	1.02 ± 0.06	1.09 ± 0.05
	2018	—	—	—	1.21 ± 0.08	—	0.91 ± 0.05	1.04 ± 0.05
0-lepton	2016	1.00 ± 0.13	1.11 ± 0.19	1.27 ± 0.09	0.95 ± 0.06	—	1.02 ± 0.04	0.83 ± 0.05
	2017	1.33 ± 0.16	1.28 ± 0.22	1.02 ± 0.08	1.03 ± 0.07	—	1.02 ± 0.06	1.14 ± 0.07
	2018	1.20 ± 0.13	0.90 ± 0.20	0.89 ± 0.06	1.21 ± 0.08	—	0.91 ± 0.05	1.01 ± 0.07

Table 6.21: Background normalization rate parameters derived from the $VZ(Z \rightarrow c\bar{c})$ full Run 2 fit to all analysis channels.

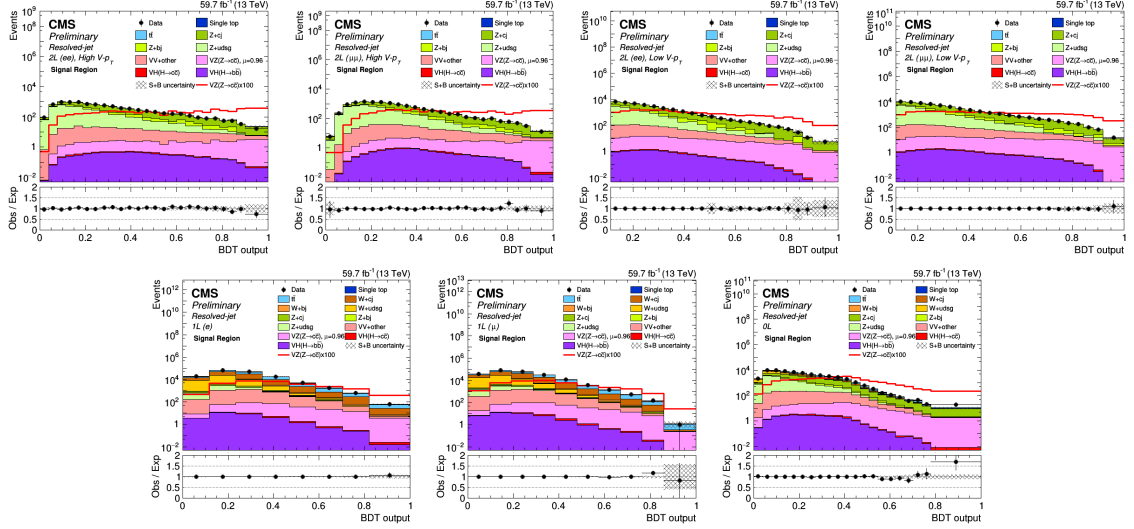


Figure 6.12.4: Signal regions used to extract the best fit $VZ(Z \rightarrow c\bar{c})$ signal strength for the 2018 data taking year. Nuisance parameters and signal strength applied to MC event yields and uncertainty bands were obtained from the full Run 2 $VZ(Z \rightarrow c\bar{c})$ fit to all analysis channels.

6.12.2 Results After Combination With Merged Analysis

When combining the the results of both the merged and resolved jet analyses, the full frequentist fit is run simultaneously in the control and signal regions in both topologies. The systematics are correlated as described in Section 6.11 while the background rate parameters are assumed to be fully uncorrelated between the topologies. A single signal strength is then extracted from the simultaneous fit to the distributions of all signal regions.

6.12.2.1 $VH(H \rightarrow c\bar{c})$ Results

As with the standalone resolved jet analysis, measuring a standard model $VH(H \rightarrow c\bar{c})$ signal strength with a significant p -value is still beyond the expected sensitivities of the combined analysis. As such, the fit is performed to obtain 95% CL upper limits on the signal strength. The set of upper limits and the best-fit signal strength for each of the data taking years of the merged+resolved combined fit can be found in Table 6.22. Figure 6.12.6 shows the distributions of data and MC bin yields as a function of expected signal. The observed and expected 95% CL upper limits for individual analysis channels, topologies, and the fully combined analysis are shown in Figure 6.12.5. From the combination of the two analysis strategies, we measure an observed (expected) best-fit signal strength of $\hat{\mu}_{VH(H \rightarrow c\bar{c})} = 7.7_{-3.5}^{+3.8}$ ($1.0_{-3.4}^{+3.7}$) which is within 2σ of unity. We additionally measure an observed (expected) 95% CL upper limit on $\hat{\mu}_{VH(H \rightarrow c\bar{c})}$ of 14 (7.6). In addition to being the best constraints on the $VH(H \rightarrow c\bar{c})$ production rate to date, the expected sensitivity

	resolved jet	merged jet	combination				
	$(p_T(H) < 300 \text{ GeV})$	$(p_T(H) > 300 \text{ GeV})$	0L	1L	2L	all	$\hat{\mu}_{VZ(Z \rightarrow c\bar{c})}$
2016 (Exp.)	34	19	26	25	30	16	$1.0^{+7.6}_{-7.1}$
2016 (Obs.)	28	20	21	43	21	18	$3.2^{+7.4}_{-7.3}$
2017 (Exp.)	32	17	24	25	26	15	$1.0^{+6.9}_{-6.1}$
2017 (Obs.)	24	24	29	26	31	18	$4.3^{+6.9}_{-6.2}$
2018 (Exp.)	33	14	20	18	24	12	$1.0^{+5.6}_{-4.9}$
2018 (Obs.)	26	26	34	18	42	21	$9.9^{+6.1}_{-5.1}$
Full Run 2 (Exp.)	19	8.8	13	12	14	7.6	$1.0^{+3.7}_{-3.4}$
Full Run 2 (Obs.)	14	17	18	19	20	14	$7.7^{+3.5}_{-3.8}$

Table 6.22: Expected and observed 95% CL upper limits on the $VH(H \rightarrow c\bar{c})$ signal strength after combination of the resolved and merged jet topologies for all three data taking year. The quoted $\hat{\mu}_{VZ(Z \rightarrow c\bar{c})}$ is the best-fit value obtained for the fit to all channels and topologies in the corresponding row.

is comparable to the $\mu_{VH(H \rightarrow c\bar{c})} < 6.4$ expected limit that ATLAS had presented on a potential 3000fb^{-1} $VH(H \rightarrow c\bar{c})$ search at the HL-LHC in October of 2020 [114].

6.12.2.2 $VZ(Z \rightarrow c\bar{c})$ Results

As with the stand alone resolved jet analysis, the $VZ(Z \rightarrow c\bar{c})$ analysis is performed to extract a signal strength encapsulating the inclusive VZ cross section convoluted with the $Z \rightarrow c\bar{c}$ branching ratio, and compute a statistical significance associated with the best fit signal strength. Unlike the resolved jet analysis, the merged jet analysis uses the same regressed mass distribution in their signal region to perform the signal extraction. Measured signal strength and the corresponding significance can be found for the combined analysis in Table 6.23. After combination of both analyses, we measured an observed (expected) signal strength of $\hat{\mu}_{VZ(Z \rightarrow c\bar{c})} = 1.01^{+0.23}_{-0.21}$ ($1.00^{+0.22}_{-0.22}$) with a significance of 5.7σ (5.9σ). This makes CMS the first experiment to observe the $Z \rightarrow c\bar{c}$ decay mode at a hadron collider! The distribution of $VZ(Z \rightarrow c\bar{c})$ SR bin yields for data and MC as a function of expected signal separation can be found in Figure 6.12.7. Despite being measured at previous e^+e^- colliders [108, 109, 110, 111, 112], the unique challenges of disentangling the $Z \rightarrow c\bar{c}$ decay process from the overwhelming QCD background had proven too difficult to fully study prior to the implementation of the various tagging and regression tools utilized by both of the analyses.

6.12.3 Constraints on the Charm-Higgs Yukawa Coupling

Because the $H \rightarrow c\bar{c}$ decay is directly related to the Higgs coupling to charm quarks, we can use the measured $VH(H \rightarrow c\bar{c})$ signal strength to obtain direct limits on the charm quark’s Yukawa coupling to the Higgs field. The interaction can be parameterized by the variable κ_c which is the ratio of

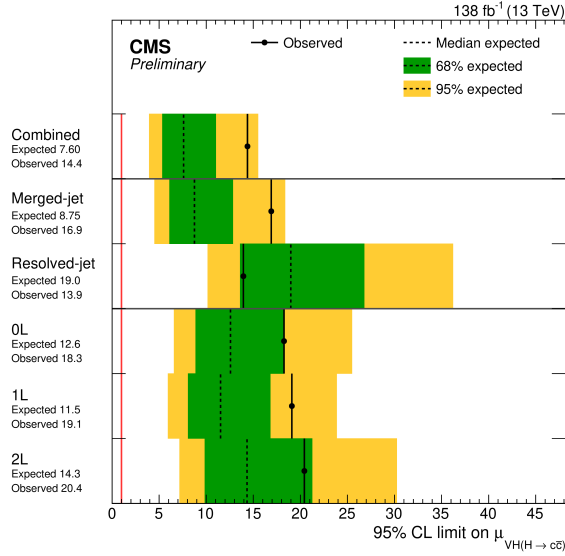


Figure 6.12.5: Observed and expected 95% CL upper limit on the $VH(H \rightarrow c\bar{c})$ signal strength for the various channels, topologies, and the combined fit to all channels, topologies, and data taking years.

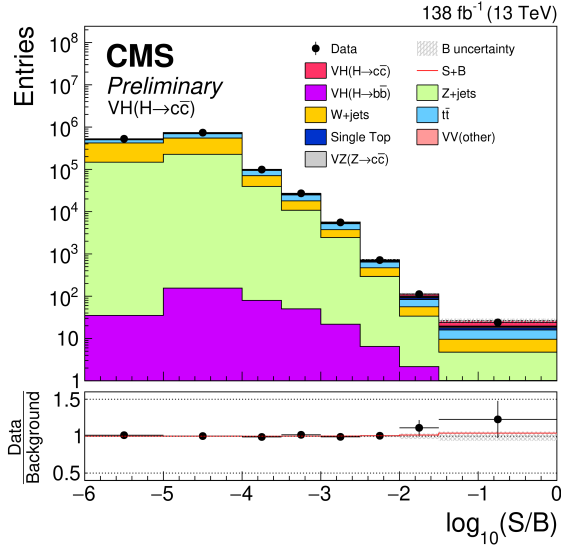


Figure 6.12.6: Distribution of signal region bin counts weighted as a function of S/B , where S and B are the expected signal and event yields in each of the $VH(H \rightarrow c\bar{c})$ signal region bins of either of the two topologies.

	resolved jet ($p_T(H) < 300$ GeV)	merged jet ($p_T(H) > 300$ GeV)	0L	1L	2L	all	$\hat{\mu}_{VZ(Z \rightarrow c\bar{c})}$
2016 (Exp.)	1.7σ	2.4σ	2.5σ	1.0σ	1.9σ	3.0σ	$1.00^{+0.42}_{-0.36}$
2016 (Obs.)	1.8σ	2.0σ	3.6σ	1.0σ	0.4σ	2.9σ	$1.02^{+0.41}_{-0.36}$
2017 (Exp.)	2.2σ	2.9σ	2.9σ	1.3σ	2.6σ	3.7σ	$1.00^{+0.35}_{-0.30}$
2017 (Obs.)	1.7σ	2.6σ	1.3σ	0.0σ	4.0σ	3.2σ	$0.84^{+0.32}_{-0.28}$
2018 (Exp.)	1.7σ	3.0σ	3.1σ	1.4σ	2.0σ	3.6σ	$1.00^{+0.39}_{-0.32}$
2018 (Obs.)	1.9σ	3.1σ	4.0σ	1.4σ	1.2σ	3.8σ	$1.23^{+0.45}_{-0.37}$
Full Run 2 (Exp.)	3.3σ	4.7σ	4.8σ	2.1σ	3.7σ	5.9σ	$1.00^{+0.22}_{-0.20}$
Full Run 2 (Obs.)	3.1σ	4.4σ	4.9σ	1.3σ	3.3σ	5.7σ	$1.01^{+0.23}_{-0.21}$

Table 6.23: Expected and observed $VZ(Z \rightarrow c\bar{c})$ significance and the best-fit combined channel signal strength after combination of the resolved and merged jet topologies for all three data taking year. After combination of all channels, topologies, and data taking years we observe the $VZ(Z \rightarrow c\bar{c})$ signal strength at over 5σ .

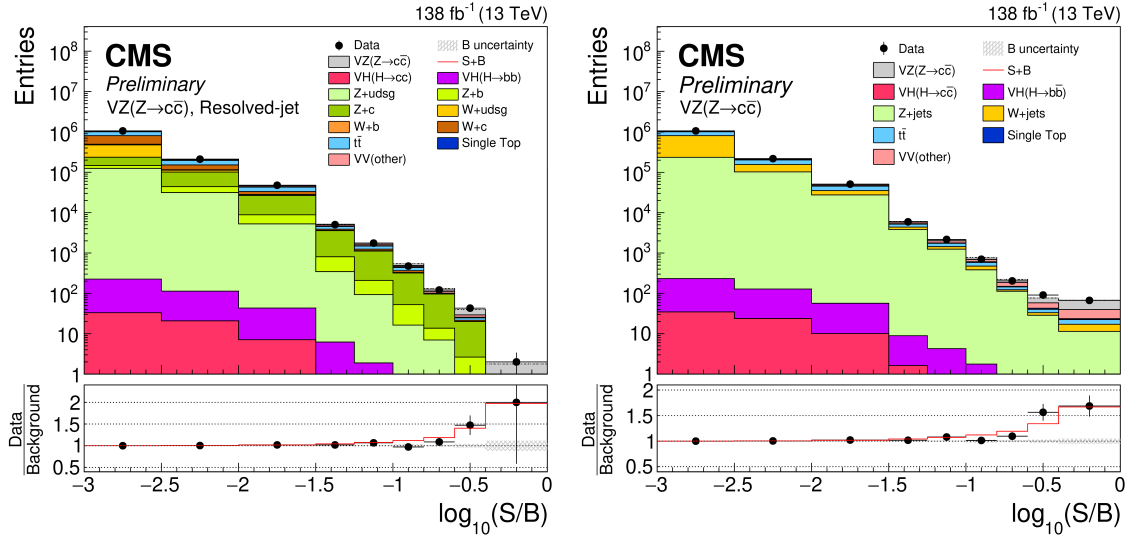


Figure 6.12.7: Distribution of signal region bin counts weighted as a function of S/B , where S and B are the expected signal and event yields in each of the $VZ(Z \rightarrow c\bar{c})$ signal region bins of the resolved jet topology (left) and combined resolved and merged jet topologies (right).

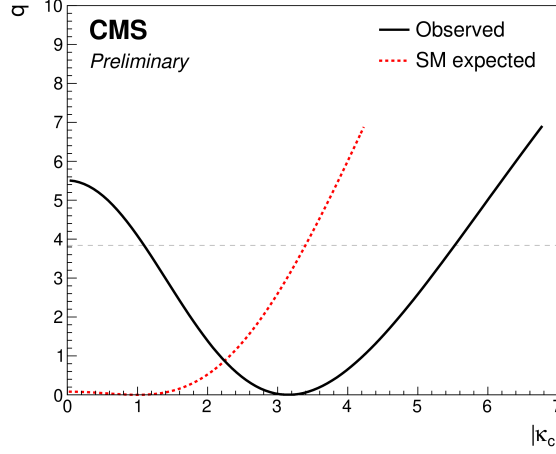


Figure 6.12.8: Parameter scan over $|\kappa_c|$ with and the corresponding test statistic calculated with the combined topology fit results using eq 6.12.2.

assumed coupling to the expected standard model value. As such a κ_c value of unity corresponds to the expected standard model value [115, 116]. Assuming that our signal strength corresponds to a standard model measurement of the VH production mode and that any non-unity value is purely corresponding to the $H \rightarrow c\bar{c}$ portion of the process, the relationship between the measured $VH(H \rightarrow c\bar{c})$ signal strength and κ_c can be approximated as:

$$\mu_{VH(H \rightarrow c\bar{c})} = \frac{\kappa_c^2}{1 + \mathcal{B}_{SM}(H \rightarrow c\bar{c}) \times (\kappa_c^2 - 1)} \quad (6.12.6)$$

A parameter scan over κ_c was performed with the corresponding $\mu_{VH(H \rightarrow c\bar{c})}$ values being extrapolated to their corresponding test statistic $\bar{q}_\mu$ originating from the combined topology $VH(H \rightarrow c\bar{c})$ signal extraction, the results can be seen in Figure 6.12.8. From this, we obtain an observed (expected) 95% CL band of $1.1 < |\kappa_c| < 5.5$ ($|\kappa_c| < 3.4$). Both expected and observed values were found to be stronger than the previous best limit of $|\kappa_c| < 6.2$ obtained from a global fit to all measured and un-measured Higgs decay signal strengths [117].

Chapter 7

Summary and Conclusion

This thesis presents the analysis in which a Higgs boson which decays to a charm quark-antiquark pair is produced in association with a W- or Z-boson. To aid in our ability to identify the signal process, we search for events in which the weak gauge boson decays to leptons and/or neutrinos. The analysis was performed via two separate analyses, a low p_T analysis which reconstructs Higgs bosons with a p_T below 300 GeV into two fully resolved AK4 jets, and a high p_T analysis which reconstructs Higgs bosons with a p_T above 300 GeV into a singled merged AK15 jet. Through a use of leading edge machine learning techniques and the highest precision simulations to date, the combined analyses were able to produce an observed (expected) 95% CL upper limit on the $VH(H \rightarrow c\bar{c})$ signal strength of 14 (7.6), This corresponds to an observed (expected) measured best fit signal strength of $\hat{\mu}_{VH(H \rightarrow c\bar{c})} = 7.7^{+3.8}_{-3.5}$ ($1.0^{+3.7}_{-3.4}$) which is within 2σ of the expected standard model value. Under the assumption that our upper limit signal strength is purely based on the $H \rightarrow c\bar{c}$ portion of the process, we can compute an observed (expected) 95% CL on the charm-Yukawa coupling of $1.1 < |\kappa_c| < 5.5$ ($|\kappa_c| < 3.4$) which is stronger than any previous direct or indirect limit.

The sensitivity of the combined analysis is approaching sensitivities originally presented by the ATLAS collaboration for a $VH(H \rightarrow c\bar{c})$ analysis at the HL-LHC [114]. In order to better instil confidence in the results of the analysis, an additional validation is performed in which the analysis is recast with a $VZ(Z \rightarrow c\bar{c})$ process set as the signal process. From the $VZ(Z \rightarrow c\bar{c})$ validation, we were able to measure an observed (expected) best fit signal strength of $\hat{\mu}_{VZ(Z \rightarrow c\bar{c})} = 1.01^{+0.23}_{-0.21}$ ($1.00^{+0.22}_{-0.22}$) at with a significance of 5.7σ (5.9σ). This result makes CMS the first experiment to have directly observed the $Z \rightarrow c\bar{c}$ decay at a hadron collider. Furthermore, the agreement between

our best fit signal strength and SM expectations demonstrates the analysis techniques' ability to properly extract and measure a $VX(X \rightarrow c\bar{c})$ signal process.

7.1 The Future of $VH(H \rightarrow c\bar{c})$

In the current iteration of the $VH(H \rightarrow c\bar{c})$ analysis, the sensitivity is heavily limited by the available statistics of the Run 2 data set. Given the 1.373 pb and 0.8839 pb cross sections of the WH and ZH production modes at $\sqrt{s} = 13$ TeV and an assumed $H \rightarrow c\bar{c}$ branching ratio of 2.88% [116], we have standard model predictions of the $\sigma(\text{WH})\mathcal{B}(H \rightarrow c\bar{c})$ and $\sigma(\text{ZH})\mathcal{B}(H \rightarrow c\bar{c})$ production rates as 39.5 fb and 29.1 fb respectively. Assuming the accepted $\mathcal{B}(Z \rightarrow \nu\nu) = 20\%$, $\mathcal{B}(W \rightarrow \ell\nu) = 21.34\%$, and $\mathcal{B}(Z \rightarrow \ell^-\ell^+) = 6.72\%$ branching fractions [4], this leads to an expected number of 803 $(Z \rightarrow \nu\bar{\nu})(H \rightarrow c\bar{c})$ events, 1163 $(W \rightarrow \ell\nu)(H \rightarrow c\bar{c})$ events, and $(Z \rightarrow \ell\ell)(H \rightarrow c\bar{c})$ 269 events collected by the CMS detector over the course of the full Run 2 data taking period. However, after considering the efficiencies of the trigger, lepton reconstruction, c -tagging, jet selection, and event selections, the expected number of $VH(H \rightarrow c\bar{c})$ events in each of the signal regions used to extract $\mu_{VH(H \rightarrow c\bar{c})}$ in this analysis are reduced to be in the single digits.

In the next phase of the LHC, we are expected to receive 300 fb⁻¹ of data over the course of Run 3 and 3000 fb⁻¹ of data over the lifetime of the HL-LHC. In addition, Run 3 is expected to begin with an increased energy of $\sqrt{s} = 13.6$ TeV and the HL-LHC aims to reach the LHC's design energy of $\sqrt{s} = 14$ TeV. It is expected that the sensitivity of an analysis will scale with the square root of the number of signal events ($\sqrt{N_S}$). Given the expected upper limit of the combined analysis being 7.6 times the standard model cross section for the full Run 2 iteration of the analysis, it is not expected that we will be able to directly measure the $H \rightarrow c\bar{c}$ process without large improvements to the analysis framework in Run 3 alone. However, for the HL-LHC the outlook may be more hopeful. At 14 TeV we expect the production cross sections of the WH and ZH processes to increase by approximately 10% and 11% respectively leading to increases in the number of expected $VH(H \rightarrow c\bar{c})$ collected by the CMS detector to be increased by a factor of ~ 24 . Assuming no changes in the systematic uncertainties and the relative size of the MC samples scales linearly with luminosity, this would bring the expected 95% CL upper limits down to ~ 1.5 times the standard model value. Preliminary projections have additionally been made by the merged jet analysis team in which theoretical uncertainties are reduced by a factor of 2, experimental uncertainties are scaled to 3000 fb⁻¹, and systematic uncertainties are adjusted according to [118]. For these projections the cut on the Higgs boson p_T was reduced to 200 GeV, and a simultaneous fit was performed to the

VH($H \rightarrow b\bar{b}$) and VH($H \rightarrow c\bar{c}$) signal strength. From this projection, they obtained an expected sensitivity of $\hat{\mu}_{\text{VH}(H \rightarrow c\bar{c})} = 1.0 \pm 1.2$ and $\hat{\mu}_{\text{VH}(H \rightarrow b\bar{b})} = 1.000 \pm 0.075$.

Appendix A

Categorization of the neutron irradiation spectrum at the Rhode Island Nuclear Science Center

Radiation is defined by the prevalence of free particles moving through space. These free particles are able to interact with stable matter on both a molecular and nuclear level, allowing them to cause damage to both organic and inorganic atomic structure. In modern high energy particle physics experiments, the goal is often synonymous to producing and then measuring the properties of high energy radiation. As such, extra care must then be taken to mitigate the damages and risks posed by radiation.

During the run of the High Luminosity LHC (HL-LHC) the particle detectors located within its caverns are expected to receive an unprecedented dose of radiation. For the CMS detector, the highest dose will be experienced in the silicon tracker and the inner layers of the forward calorimeter. The functionality of the strip and pixel detectors in the tracker and the High Granularity Calorimeter (HGCAL) in the forward calorimeter depends heavily on the performance of p-doped silicon semiconductor sensors. Because prolonged radiation exposure degrades semiconductor performance, ample testing will be performed to insure that materials used in the phase II upgrade of the CMS detector will continue to perform within operational specifications after receiving the expected radiation doses of the HL-LHC.

Sensors will undergo neutron and proton irradiation. Afterwards, the sensor's electrical properties

will be measured to determine post irradiation performance. One irradiation site will be at the Rhode Island Nuclear Science Center (RINSC), a neutron reactor located in Narragansett, Rhode Island. It is currently planned that all neutron irradiations happening in the United States will take place at RINSC. However, before starting the irradiation campaign the known neutron flux at the reactor was only known to be approximately $3 \cdot 10^{11} \text{ n}_{\text{eq}}/\text{cm}^2/\text{s}$. This value was obtained by irradiating CoAl monitoring wire and measuring its resulting nuclear activation, and was most likely a factor of 2 estimation at best. Neither was there any understanding of the energy spectrum of the neutrons. Instead, the flux was only given as a single number in units of 1-MeV neutron equivalents. This value gives the number of 1 MeV neutrons that are expected to leave the observed amount of energy in the irradiated sample. It was determined that before irradiations at RINSC could take place, a study would be performed to fully characterize the neutron energy spectrum of the reactor. For this study, the method of spectrum unfolding via foil activation was used to obtain the neutron energy flux as a function of energy in three separate neutron exposure beam ports. Additional foil activation measurements as well as measurements of irradiated diodes were then used to measure the neutron flux in the pneumatic rabbit system, which is used to expose small samples to the reactor core.

A.1 The Interactions of Free Neutrons with Atomic Nuclei

Because neutrons are electrically neutral, they rarely interact directly with an atom's electron cloud. This leaves the nuclei as the only piece of the atom that neutrons are still able to interact with. There are three primary ways in which the neutron can interact: Hard scattering, fission, neutron absorption. In all cases, the probability of interaction occurring is described by a cross section $\sigma(E)$. The cross section tells us the probability that nuclear interaction will occur given the energy of the incident neutron. The cross section is normally an intrinsic property of the nuclei based on the number of neutrons and protons inside, as well as the type of interaction happening. This can lead to a single element to having multiple different neutron energy response profiles.

In hard scattering, the neutron will directly strike the nucleus and elastically bounce off. In this interaction the neutron will transfer some of its kinetic energy to the nucleus, but the nucleus will not change as a result of the collision. Given a target composed of a isomorphous material, the number of hard scatter neutrons per unit for a generic neutron spectrum $\phi(E)$ is given by

$$N = \int_0^{\infty} \phi(E) \sigma(E) dE \quad (\text{A.1.1})$$

where $\phi(E)$ describes the rate per unit area of incident neutrons of energy E that traveled through the material. The scattering cross section $\sigma(E)$ is non-zero for all neutron energies, but is normally larger at lower neutron energies. For very heavy nuclei like U235 and Pu239 the energy transferred to the nucleus when being struck by a neutron can cause the nucleus to split into multiple smaller parts. This splitting of the nucleus into smaller parts as a result of nuclear scattering is known as nuclear fission. Fission is the driving mechanism of nuclear reactors, which primarily consist of a core of highly fissile material as well as moderators like boron, graphite, and water to control the energy flux of neutrons that will scatter off of the core.

Neutron absorption is a different type of interaction than those previously described, and it is often referred to as activation because it causes the nucleus to become radioactive. Rather than having the neutron scatter off a nuclei, the neutron is instead absorbed into the nucleus. This absorption causes the nucleus to become unstable and quickly eject one or more particle to reach a more stable state. Based on the proton and neutron composition of the nucleus, there are four types of particles that can be ejected after neutron absorption: a photon (γ), a proton (p), an alpha particle (α), or two neutrons (2n). The type of absorption and ejection is often written in the format of $X(n,r)Y$, where X and Y refers to the initial and final isotope, n is referring to the neutron being absorbed, and r describes the type of particle ejected from the nucleus. For example $Al_{13}^{27}(n,\alpha)Na_{11}^{24}$ refers to an Al_{13}^{27} isotope absorbing a neutron and then ejecting an alpha particle to become an Na_{11}^{24} isotope. The rate at which neutron absorption happens for an incident neutron flux $\phi(E)$ is similar to eq A.1.1, except that we now must consider the fact that as neutron exposure time increases, the number of un-activated nuclei will also be changing. The number of activated nuclei N_a after exposing N atoms with activation cross section $\sigma(E)$ exposed to a uniform neutron flux $\phi_n(E)$ for time t_i can then be given by the formula:

$$N_a = \int_0^{t_i} \int_0^{\infty} \phi_n(E) \sigma(E) N(t) dE dt \quad (A.1.2)$$

The exact form of $N(t)$ in eq A.1.2 is somewhat complicated, because in addition to the number of un-activated atoms slowly being depleted during irradiation, it is also possible for the activated nuclei to decay back into the original un-activated state. Activated isotopes will decay at a rate of

$$A = \lambda N_a e^{-\lambda t} \quad (A.1.3)$$

where A is the activity and tells us the rate at which the activated nuclei are decaying, and λ is

the decay constant which tells us how much time must pass for an e -fold of decay to take place. From eq A.1.2 and A.1.3 we can infer that as the percentage of activated nuclei grows, the rate of activation will decrease while the rate of decay increases. For a material experiencing a constant and uniform neutron flux, there will exist an equilibrium where the rate of activation and rate of decay become equal. This is referred to as activity at saturation, and the rate at which activated nuclei decay at this saturation point is referred to as the saturated activity (A^∞). If a radiation exposure of time t_i causes a material to become activated with activity A^{irrad} , the saturated activity can then be calculated as

$$A^\infty = \frac{A^{\text{irrad}}}{1 - e^{-\lambda t_i}} \quad (\text{A.1.4})$$

The saturated activity is a useful quantity to work with, as the percentage of nuclei that will be activated during saturation is only dependent on the neutron flux and the activation cross section of the nuclei. As such, the saturated activity can be normalized in such a way that it is independent of the size of the activated sample. In particular, the saturated activity in units of decays/nuclei/s can be given by

$$A_{\text{norm}}^\infty = \int_0^\infty \phi_n(E) \sigma(E) dE \quad (\text{A.1.5})$$

giving us a direct relationship between the saturated activity, activation cross section, and neutron energy flux that is independent of the number of nuclei that is being exposed.

A.2 The Need for Radiation Hardness in Particle Physics

Particle colliders are constantly producing high energy radiation which are interacting with the detectors. Often, this radiation comes in the form of the final state particles which the particle detectors are designed to measure. However, in colliders like this LHC, radiation can also come in the form of protons existing in the beam halo colliding with the inner most layers of the detectors. This creates an extra degree of difficulty when it comes to detector design, because materials must be chosen such that they can not only operate in high radiation environments, but must also be able to continue to operate within specifications after years of radiation exposure.

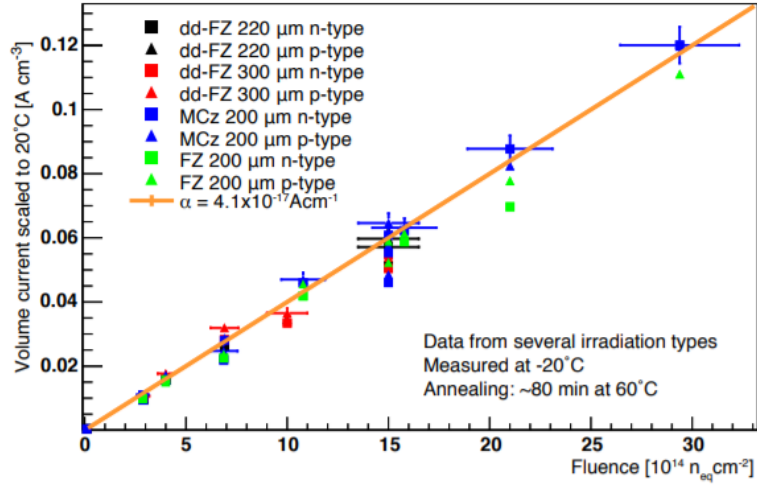


Figure A.2.1: The measured post-irradiation leakage current at the depletion voltage in n-type and p-type silicon as a function of equivalent neutron dose. Increases in leakage current were found to have a linear relationship with radiation fluence, and were found to be independent of the type of radiation after conversion of the effective dose. Image taken from [14].

A.2.1 Radiation Damage in Silicon

Silicon is a preferred detector material because of its natural radiation hardness. When ionizing radiation passes through a silicon bulk, it will produce localized charge clusters through production of electron-hole pairs. However, these charge clusters will quickly decay allowing the silicon behavior to reset before the next set of collisions take place. Because the electron structure of the silicon resets quickly and autonomously, there is no intrinsic build-up of radiation damage over long periods of time. This makes silicon an incredibly robust material against ionizing radiation.

However, silicon is still able to experience long term damage as a result of particles scattering off the nuclei in the silicon lattice. When heavy particles collide with the atoms in a silicon lattice, either the silicon itself or a dopant, it can push the atom out of its lattice site. As these lattice defects slowly build up, the electric performance of silicon based sensors will degrade. In general, this will manifest as an increase in the reverse bias leakage current, as can be seen in Figure A.2.1, and a decrease in the rate at which currents can move through the silicon as a result of charge carriers getting trapped in the vacant lattice sites. In the case of n-doped silicon high enough radiation doses can even cause type inversion to occur, where enough electrons become trapped in the empty lattice sites that their resulting holes become the dominant charge carriers in the material.

It has been found that some of the effects of the radiation damage can be healed through annealing, where the temperature of the silicon is raised such that the thermal vibrations of the nuclei can

help move them back into their lattice sites. However, it has also been found that for long enough annealing times, the thermal vibrations will cause an overall worsening of the silicon performance due to a general displacement of nuclei with respect to the nominal lattice sites. The slow build-up of the annealing induced damage leads to a maximum amount of time that radiation damaged silicon can experience beneficial annealing. This leads to a need for ample testing when building silicon based particle detector components to ensure that they can continue to perform after experiencing their full expected radiation doses.

A.3 Spectrum Unfolding Using Foil Activation

Determining the energy spectrum $\phi(E)$ of particles producing from a fissioning reactor core is a non-trivial problem in nuclear physics. The neutron spectrum of a fissioning reactor core is well modeled [119], but the emergent energy spectrum of the neutrons inside the reactor is heavily dependent on the geometry of the reactor. Additionally, the rate of neutrons production from a super critical U235 core is far too great for modern calorimetry techniques to directly measure the energy of the individual neutrons [120, 121]. Instead a spectrum unfolding technique is used, in which objects with differing radiation response profiles are placed in a reactor and have their resulting radiation response measured. By having suitably different energy and particle type responses, the radiation energy spectrum can be reproduced using a parametric fit to an assumed functional form. For neutron reactors, the two most common ways to obtain radiation response measurements is by using Bonner spheres [122] or foil activation.

Bonner spheres consist of varying diameter moderating spheres each containing identical thermal neutron measuring centers. Because the number of scattering nuclei per unit length can be given by eq A.1.1, varying the diameter and type of moderating material can make each sphere sensitive to a specific portion of the neutron energy spectrum. This makes Bonner spheres a very standardized method for neutron spectrum unfolding. However, the neutron penetration depth can be quite long due to their electrically neutral nature. This requires the variation in Bonner sphere size to be quite large, with the larger spheres having diameters larger than a foot. This makes the Bonner sphere method for neutron spectrum unfolding unusable when trying to measure the energy spectrum inside of a small location.

For cases where the Bonner sphere method will not work, neutron spectrum unfolding via foil activation becomes a preferred method of neutron measurement. The foil activation method uses the phenomenon of neutron absorption described by eq A.1.2. By exposing different elements to

a uniform neutron flux, one can use their varying activation cross sections to study the full form of the neutron energy spectrum. As such, the foil activation method of neutron unfolding aims to irradiate different types of ultra thin foils composed of pure elements to reconstruct the neutron energy spectrum. The choice to use thin foils is made to work under the assumption that any self-shielding of the material, in which the scattering and absorption of neutrons by the foil will cause the neutron flux to have a depth dependence, is negligible.

After irradiating a set of foils, they are first allowed to cool to allow safe handling by reactor personnel. The activity of all radioactive isotopes in the foil can then be measured using gamma spectroscopy [123]. From the activity measured from gamma spectroscopy, the activity at the end of the irradiation time can be calculated using

$$A^{\text{irrad}} = A^{\gamma} e^{\lambda t_w} \quad (\text{A.3.1})$$

where A^{γ} is the activity measured via gamma spectroscopy and t_w is the time between the end of the neutron exposure period and when the activity of the sample was measured via gamma spectroscopy. Given an activity measured via gamma spectroscopy in units of decays per second eq A.1.4, A.1.5, and A.3.1 can be combined to produce the equation:

$$A_{\text{norm}}^{\infty} = \frac{A^{\gamma} e^{\lambda t_w}}{1 - e^{-\lambda t_i}} \frac{N_A m_m f}{m \rho} = \int_0^{\infty} \phi_n(E) \sigma(E) dE \quad (\text{A.3.2})$$

where N_A is Avogadro's number, m_m is the molar mass of the material, f is the fraction of nuclei of the material (by mass) which are the isotope of interest, m is the mass of the foil, and ρ is the density of the foil. By constructing the foil out of elements in their natural occurring form, one can use established values of f and m_m making all values in the central portion of the equality to either be known constants or qualities of the material which can be directly measured. This then allows one to directly measure A_{norm}^{∞} for all isotopes of interest irradiated via the applied neutron spectrum.

Given a collection of normalized saturated activity values, one can then compare the measured values to an expected value calculated using the known activation cross section for the material and an assumed neutron energy spectrum. In doing so, the assumed spectrum which most accurately predicts the measured saturated activity can be taken as the true neutron energy spectrum. The difficulty then comes from choosing how to model the energy spectrum. For our study, we model

the neutron energy spectrum using three distinct sections:

$$\text{Thermal Neutrons} \quad \phi_{\text{Th}}(E) = \sqrt{\frac{E}{(k_B T)^3}} e^{-E/k_B T} \quad (\text{A.3.3})$$

$$\text{Intermediate Neutrons} \quad \phi_{\text{I}}(E) = \frac{1}{E} \quad (\text{A.3.4})$$

$$\text{Fission Neutrons} \quad \phi_{\text{F}}(E) = e^{-E/0.965} \sinh \sqrt{2.29E} \quad (\text{A.3.5})$$

Where the low energy thermal neutrons are modeled via a Maxwell-Boltzmann distribution with an assumed temperature T , the intermediate energy neutrons are assumed to have a falling $1/E$ spectrum, and the high energy fission neutrons are assumed to be produced according to the Watt distribution. For the Watt distribution, the numerical constants are all in units of MeV^{-1} . We then apply a floating normalization to each of the three regions and assume that the spectrum changes distributions at the energy in which the two distributions intersect. In the scenario in which no intersection occurs between the intermediate and fission distributions (due to a large mis-match in normalizations), the distribution change is set to occur no lower than 0.001 MeV and no higher than 0.5 MeV .

A.3.1 Monitoring of Neutron Fluence via D0 Diodes

As described in Section A.2.1, radiation damage in silicon leads to an increase in the leakage current that is proportional to the experienced radiation fluence. This creates an additional metric for determining the reactor flux, in which diodes of known quality are irradiated and have their resulting leakage current at depletion voltage measured. For this study, we perform silicon based fluence measurements by first irradiating leftover diodes from the D0 experiment with a nuclear reactor for a set amount of time. After irradiation, the exposed diodes are stored in a freezer to prevent any un-intended annealing. They are then annealed in an oven at 60° C for 80 minutes, and then have their depletion voltage measured from Capacitance-Voltage measurements. The current at depletion voltage is then measured, and the increase in current (ΔI) as a result of the irradiation is proportional to the fluence as governed by

$$\Delta I = \alpha \Phi_{\text{eq}} V \quad (\text{A.3.6})$$

where α is a factor which accounts for the effects of annealing, Φ_{eq} is the 1 MeV equivalent neutron flux for silicon, and V is the volume of the active silicon bulk. From this relationship, we then

obtain an effective radiation fluence. Because leakage current measurements have a temperature dependence, we attempt to perform all measurements in a temperature controlled environment kept at 20° C. However, in the case of high radiation doses, the diodes often need to be cooled to prevent their currents from reaching the compliance of our equipment and have their measured current scaled to an expected measurement at 20° C. After the effective radiation dose for the diode has been determined, the fluence is divided by the radiation exposure time to obtain the neutron flux. Unlike the previously described measurement methods, this is unable to recreate a full neutron energy spectrum. Instead, it is only able to give a single normalized neutron flux.

A.3.2 Measuring the Neutron Spectrum of the RINSC Reactor

The RINSC reactor has two different methods for exposing samples to the reactor core. The first is by placing samples into a beamport aimed towards the core. At RINSC, there are two beam ports with a 6" diameter (referred to as 6"a and 6"b) and one beamport with an 8" diameter. The second exposure method is through the use of a pneumatic rabbit system in which samples are loaded into cylinders which are then sent to and retrieved from the reactor core automatically. A full set of spectrum unfolding measurements was performed when measuring the neutron spectrum inside the beamports. However, due to size limitations of the rabbit system, an alternative method for characterizing the neutron flux was used.

To probe the full range of neutron energies, multiple sets of foils were used in each irradiation. They were chosen such that we could have multiple data points constraining the high and low neutron energy distributions, while also having a short enough half-life such that we only have to wait a reasonable amount of time between irradiating the foils and being able to safely measure their activity via gamma spectroscopy. To understand the uniformity of the neutron flux across the beamport cross section, we included duplicate sets of foils for each irradiation. Additionally, the 6"a and 8" beamport irradiations included some foils placed in a 0.0254" cadmium case to shield the foil from thermal neutrons. Foils were then placed into a wooden holder that was inserted into the beamport, as can be seen in Figure A.3.1. For all beamport irradiations, the reactor core was set to full power for one hour and then turned off. After the conditions were safe, the holders were then retrieved and the foils were set in lead containers until they could be safely handled. Once their activity was low enough, they then had their resulting activity measured via gamma spectroscopy.

The resulting set of measured activities for the 6"a, 6"b, and 8" beamports can be found in Table A.3–A.5. With the exception of the half-life which is listed in lieu of the decay constant, all values



Figure A.3.1: An image of the foil holder used to put foils into the 6" beamport. The foils were placed in small holes bowed into a milled wooden disc and secured using a polyethylene disc attached to the face of the wood.

listed refer to their corresponding measured values in eq A.3.1 and A.3.2. From our measurements of each of the beamports, we found that there was a $\sim 20\%$ variation in the measured saturated activity across foils of the same type. To account for this, when performing the spectrum unfolding procedure we averaged saturated activity measurements associated with the same activation process and beamport together and assigned a 20% uncertainty to the measured values. Additionally, to add uniformity to the unfolding procedure across all three beam ports, the spectrum derivation was only performed with the following activation processes which were common to the foil selection of all three beam ports: $\text{Al}^{27}(\text{n},\alpha)\text{Na}^{24}$, $\text{Fe}^{54}(\text{n},\text{p})\text{Mn}^{54}$, $\text{Fe}^{56}(\text{n},\text{p})\text{Mn}^{56}$, $\text{Fe}^{58}(\text{n},\gamma)\text{Fe}^{59}$, $\text{Zr}^{90}(\text{n},2\text{n})\text{Zr}^{89}$, and $\text{Cu}^{63}(\text{n},\gamma)\text{Cu}^{64}$. Distribution of the normalized activation cross section for each of these processes can be found in Figure A.3.2. Cross sections were obtained from the IRDF database [124] and binned using the standard ENDF-6 640 energy groups.

When performing spectrum unfolding the methodology described in the previous section is used, in which the described thermal, intermediate, and fission neutron energy distributions are each given a floating normalization parameter. For the thermal neutron spectrum, it is assumed that the thermal neutrons reach an equilibrium temperature with the water in which the core is submerged. As such, a temperature of 70°F (converted to Kelvin) is used for the Maxwell-Boltzmann distribution. In order to find the best set of normalization parameters, initial values were set and the resulting energy spectrum was used to calculate an assumed saturated activity based on the set of normalization

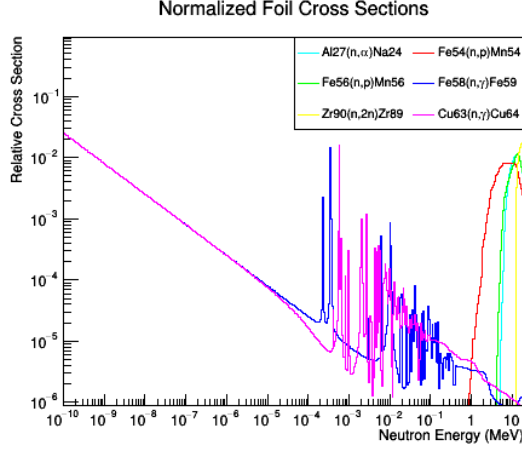


Figure A.3.2: Neutron activation cross sections as a function of incident neutron energy for the 6 absorption processes considered when fitting the neutron energy spectrum in each of the beam ports.

parameters. From saturated activity, we calculate the χ^2 value

$$\chi^2 = \sum_{i,j}^{\text{foils}} A_i^{\infty, \text{meas}} (E_{i,j}^{-1})^2 A_j^{\infty, \text{calc}} \quad (\text{A.3.7})$$

where $A_i^{\infty, \text{meas}}$ and $A_j^{\infty, \text{calc}}$ are the saturated activity values calculated based on the activity measured from the foils and calculated using the assumed spectrum, and $E_{i,j}^{-1}$ is the error matrix capable of accounting for correlations when measuring multiple isotopes from the same foil. The MINUIT software package [125] is then used to adjust the normalization parameters according to the MIDGRAD algorithm such that the resulting χ^2 value is decreased. This procedure is then repeated until a minimum χ^2 value is reached. To ensure that the algorithm did not settle to a local minimum, we performed the fit multiple times after scanning across a range of initial normalization values, their minimum and maximum values, and the distance with which they can move during each minimization step. The resulting neutron energy spectrum for each of the beam ports as well as a comparison of the measured and calculated saturated activity values can be found in Figure A.3.3 and Table A.1. In performing the fitting procedure, we found that the calculated value of the $\text{Zr90}(n,2n)\text{Zr89}$ saturated activity was consistently lower than the measured value. Further investigation lead us to finding that the Watt spectrum used to model the fission neutrons slightly underestimates the distribution of very high energy neutrons. To address this, we looked into convoluting the fission spectrum with an exponentially decaying tail for neutrons with $E > 10$ MeV. However, we found that doing so only modified the effective neutron flux by $\sim 1\%$ while introducing large instabilities in the parameterized fit. As such, we dropped the inclusion of the tail from the

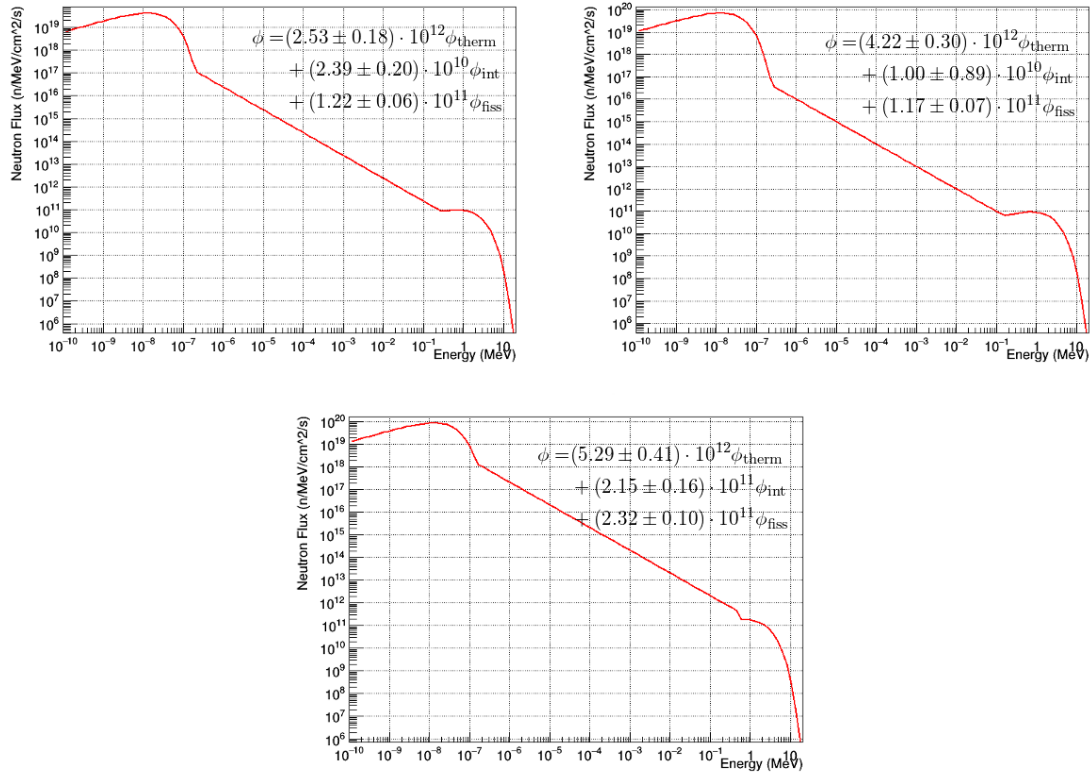


Figure A.3.3: The derived neutron energy spectrum and fit normalizations for the neutron energy spectrum found for the 6"a (left), 6"b (middle), and 8" (right) beamports at the RINSC reactor.

final fitting procedure.

To model the effect that the derived neutron energy spectra would have on silicon sensors, the distributions are normalized with respect to a 1 MeV equivalent dose for silicon. This 1 MeV equivalent dose represents the effective neutron flux if instead of a spectrum of neutron energies, they all had an energy of 1 MeV. This value is obtained by convoluting the energy spectrums with a silicon damage function [126] representing the ratio of damage caused by a neutron of energy E divided by that caused by a 1 MeV neutron. This convoluted energy distribution is then integrated over the energy to obtain the normalized result. The 1 MeV equivalent doses calculated from the spectrum unfolding via foil activation technique for each of the three beamports can be found in Table A.2. When planning to irradiate samples, the desired doses are divided by the 1 MeV equivalent flux to determine the required neutron exposure time. In the case of the 6" beamports, we find that the 1 MeV equivalent neutron dose is in agreement with the pre-established estimation from the CoAl wire measurements.

Beamport	Process	A_{meas}^{∞}	A_{calc}^{∞}	$A_{\text{meas}}^{\infty}/A_{\text{calc}}^{\infty}$
6" a	Al27(n, α)Na24	$1.76 \cdot 10^{-16}$	$1.51 \cdot 10^{-16}$	1.11
	Al27(n, α)Na24 cd	$1.82 \cdot 10^{-16}$	$1.49 \cdot 10^{-16}$	1.05
	Fe54(n,p)Mn54	$2.24 \cdot 10^{-14}$	$2.15 \cdot 10^{-14}$	1.04
	Fe54(n,p)Mn54 cd	$1.48 \cdot 10^{-14}$	$2.13 \cdot 10^{-14}$	0.69
	Fe58(n, γ)Mn59	$3.80 \cdot 10^{-12}$	$3.31 \cdot 10^{-12}$	1.15
	Fe58(n, γ)Mn59 cd	$4.72 \cdot 10^{-14}$	$4.49 \cdot 10^{-14}$	1.05
	Zr90(n,2n)Zr89	$4.99 \cdot 10^{-17}$	$1.71 \cdot 10^{-17}$	2.92
	Zr90(n,2n)Zr89 cd	$4.98 \cdot 10^{-17}$	$1.69 \cdot 10^{-17}$	2.94
	Cu63(n, γ)Cu64	$1.05 \cdot 10^{-11}$	$1.16 \cdot 10^{-11}$	0.91
	Cu63(n, γ)Cu64 cd	$1.38 \cdot 10^{-13}$	$1.44 \cdot 10^{-13}$	0.95
6" a	Al27(n, α)Na24	$1.59 \cdot 10^{-16}$	$1.68 \cdot 10^{-16}$	0.95
	Fe54(n,p)Mn54	$1.79 \cdot 10^{-14}$	$2.06 \cdot 10^{-14}$	0.87
	Fe58(n, γ)Mn59	$5.58 \cdot 10^{-12}$	$5.46 \cdot 10^{-12}$	1.02
	Zr90(n,2n)Zr89	$4.48 \cdot 10^{-17}$	$1.64 \cdot 10^{-17}$	2.74
	Cu63(n, γ)Cu64	$1.87 \cdot 10^{-11}$	$1.91 \cdot 10^{-11}$	0.98
8"	Al27(n, α)Na24	$4.28 \cdot 10^{-16}$	$3.33 \cdot 10^{-16}$	1.28
	Al27(n, α)Na24 cd	$3.86 \cdot 10^{-16}$	$3.30 \cdot 10^{-16}$	1.17
	Fe54(n,p)Mn54	$2.97 \cdot 10^{-14}$	$4.08 \cdot 10^{-14}$	0.73
	Fe54(n,p)Mn54 cd	$3.32 \cdot 10^{-14}$	$4.04 \cdot 10^{-14}$	0.83
	Fe58(n, γ)Mn59	$6.14 \cdot 10^{-12}$	$7.26 \cdot 10^{-12}$	0.85
	Fe58(n, γ)Mn59 cd	$3.18 \cdot 10^{-13}$	$3.62 \cdot 10^{-13}$	0.88
	Zr90(n,2n)Zr89	$8.22 \cdot 10^{-17}$	$3.24 \cdot 10^{-17}$	2.53
	Zr90(n,2n)Zr89 cd	$7.97 \cdot 10^{-17}$	$3.21 \cdot 10^{-17}$	2.48
	Cu63(n, γ)Cu64	$3.67 \cdot 10^{-11}$	$2.53 \cdot 10^{-11}$	1.45
	Cu63(n, γ)Cu64 cd	$1.44 \cdot 10^{-12}$	$1.15 \cdot 10^{-12}$	1.25

Table A.1: Comparison of the measured saturated activity to the values calculated from the parameterized fit for each of the three beam port measurements in units of decays/nuclei/s. The calculated values were obtained using the neutron energy spectra shown in Figure A.3.3.

Beam Port	6" a	6" b	8"
Flux ($n_{\text{eq}}/\text{cm}^2/\text{s}$)	$(2.96 \pm 0.15) \cdot 10^{11}$	$(2.81 \pm 0.17) \cdot 10^{11}$	$(7.03 \pm 0.30) \cdot 10^{11}$

Table A.2: The 1 MeV equivalent neutron flux for the three beam ports as measured via spectrum unfolding from foil activation.

Activation	Half-Life (h)	A^γ (μCi)	t_w (h)	A^{irrad} (μCi)	A^∞ (decays/N/s)
Al27(n, α)Na24	14.959	$3.25 \cdot 10^{-1}$	29.1	1.25	$1.96 \cdot 10^{-16}$
Al27(n, α)Na24 cd		$3.34 \cdot 10^{-1}$	28.7	1.27	$1.82 \cdot 10^{-16}$
Fe54(n,p)Mn54	7489.44	$5.19 \cdot 10^{-3}$	24.0	$5.20 \cdot 10^{-3}$	$2.38 \cdot 10^{-14}$
Fe54(n,p)Mn54		$3.27 \cdot 10^{-3}$	27.2	$3.28 \cdot 10^{-3}$	$1.53 \cdot 10^{-14}$
Fe54(n,p)Mn54		$4.58 \cdot 10^{-3}$	28.4	$4.59 \cdot 10^{-3}$	$2.09 \cdot 10^{-14}$
Fe54(n,p)Mn54 cd		$3.24 \cdot 10^{-3}$	27.1	$3.25 \cdot 10^{-3}$	$1.48 \cdot 10^{-14}$
Fe56(n,p)Mn56	2.5789	$3.86 \cdot 10^{-3}$	24.4	2.72	$3.24 \cdot 10^{-16}$
Fe56(n,p)Mn56		$1.36 \cdot 10^{-3}$	27.2	2.01	$2.44 \cdot 10^{-16}$
Fe56(n,p)Mn56		$1.27 \cdot 10^{-3}$	28.4	2.64	$3.12 \cdot 10^{-16}$
Fe56(n,p)Mn56 cd		$1.32 \cdot 10^{-3}$	27.1	1.90	$2.25 \cdot 10^{-16}$
Fe58(n, γ)Fe59	1067.88	$2.61 \cdot 10^{-1}$	27.2	$2.65 \cdot 10^{-1}$	$3.88 \cdot 10^{-12}$
Fe58(n, γ)Fe59		$1.61 \cdot 10^{-1}$	28.4	$1.64 \cdot 10^{-1}$	$2.45 \cdot 10^{-12}$
Fe58(n, γ)Fe59		$2.51 \cdot 10^{-1}$	27.1	$2.55 \cdot 10^{-1}$	$3.73 \cdot 10^{-12}$
Fe58(n, γ)Fe59 cd		$3.18 \cdot 10^{-3}$	28.4	$3.23 \cdot 10^{-3}$	$4.72 \cdot 10^{-14}$
Ni58(n,p)Co58	1700.64	$4.02 \cdot 10^{-1}$	25.8	$4.06 \cdot 10^{-1}$	$1.86 \cdot 10^{-14}$
Ni58(n,p)Co58 cd		$4.15 \cdot 10^{-1}$	23.1	$4.19 \cdot 10^{-1}$	$1.91 \cdot 10^{-14}$
Ti48(n,p)Sc48	43.67	$2.14 \cdot 10^{-2}$	24.5	$3.21 \cdot 10^{-2}$	$5.68 \cdot 10^{-17}$
Ti48(n,p)Sc48		$2.24 \cdot 10^{-2}$	25.1	$3.34 \cdot 10^{-2}$	$5.92 \cdot 10^{-17}$
V48(n,p)Sc48	43.67	$7.44 \cdot 10^{-4}$	24.3	$1.10 \cdot 10^{-3}$	$4.69 \cdot 10^{-18}$
V48(n,p)Sc48		$5.96 \cdot 10^{-4}$	25.5	$8.68 \cdot 10^{-4}$	$4.20 \cdot 10^{-18}$
Zr90(n,2n)Zr89	78.41	$3.61 \cdot 10^{-3}$	26.6	$4.57 \cdot 10^{-3}$	$4.99 \cdot 10^{-17}$
Zr90(n,2n)Zr89		$3.61 \cdot 10^{-3}$	26.1	$4.55 \cdot 10^{-3}$	$4.98 \cdot 10^{-17}$
In113(n, γ)In114 cd	1188.24	2.44	71.2	2.55	$6.64 \cdot 10^{-12}$
Cu63(n, γ)Cu64	12.7	$3.50 \cdot 10^{-1}$	194.3	$1.41 \cdot 10^4$	$1.05 \cdot 10^{-11}$
Cu63(n, γ)Cu64 cd		3.71	71.6	$1.85 \cdot 10^2$	$1.38 \cdot 10^{-13}$
Au197(n, γ)Au198	64.6121	1.37	911.0	$2.40 \cdot 10^4$	$2.32 \cdot 10^{-10}$
Au197(n, γ)Au198		1.76	912.0	$3.13 \cdot 10^4$	$2.96 \cdot 10^{-10}$
Au197(n, γ)Au198		1.07	914.6	$1.96 \cdot 10^4$	$1.92 \cdot 10^{-10}$
Au197(n, γ)Au198 cd		1.34	670.9	$1.79 \cdot 10^3$	$1.72 \cdot 10^{-11}$

Table A.3: Measurements associated with the 6" beam port irradiation. Measured activity from gamma spectroscopy (A^γ), the back calculated activity at the end of the irradiation A^{irrad} , and the associated saturated activity A^∞ are given for each irradiated sample. In the case of iron, Fe54, Fe56, and Fe58 measurements were derived from the same set of foils.

Activation	Half-Life (h)	A^γ (μCi)	t_w (h)	A^{irrad} (μCi)	A^∞ (decays/N/s)
Al27(n, α)Na24	14.959	$3.61 \cdot 10^{-1}$	24.4	1.12	$1.54 \cdot 10^{-16}$
Al27(n, α)Na24		$2.82 \cdot 10^{-1}$	24.8	$8.89 \cdot 10^{-1}$	$1.41 \cdot 10^{-16}$
Al27(n, α)Na24		$3.70 \cdot 10^{-1}$	25.1	1.19	$1.64 \cdot 10^{-16}$
Fe54(n,p)Mn54	7489.44	$4.42 \cdot 10^{-3}$	22.9	$4.43 \cdot 10^{-3}$	$1.98 \cdot 10^{-14}$
Fe54(n,p)Mn54		$1.03 \cdot 10^{-2}$	23.9	$1.03 \cdot 10^{-2}$	$4.59 \cdot 10^{-14}$
Fe54(n,p)Mn54		$3.92 \cdot 10^{-3}$	23.2	$3.93 \cdot 10^{-3}$	$1.60 \cdot 10^{-14}$
Fe56(n,p)Mn56	2.5789	$4.96 \cdot 10^{-3}$	22.9	2.35	$2.73 \cdot 10^{-16}$
Fe56(n,p)Mn56		$5.19 \cdot 10^{-3}$	23.9	3.18	$3.67 \cdot 10^{-16}$
Fe56(n,p)Mn56		$3.98 \cdot 10^{-3}$	23.2	2.05	$2.18 \cdot 10^{-16}$
Fe58(n, γ)Fe59	1067.88	$4.32 \cdot 10^{-1}$	22.9	$4.38 \cdot 10^{-1}$	$6.27 \cdot 10^{-12}$
Fe58(n, γ)Fe59		$4.82 \cdot 10^{-1}$	23.9	$4.90 \cdot 10^{-1}$	$6.95 \cdot 10^{-12}$
Fe58(n, γ)Fe59		$3.69 \cdot 10^{-1}$	23.2	$3.75 \cdot 10^{-1}$	$4.89 \cdot 10^{-12}$
Ni58(n,p)Co58	1700.64	$3.53 \cdot 10^{-1}$	24.9	$3.57 \cdot 10^{-1}$	$1.57 \cdot 10^{-14}$
Ni58(n,p)Co58		$2.77 \cdot 10^{-1}$	26.1	$2.80 \cdot 10^{-1}$	$1.25 \cdot 10^{-14}$
Ni58(n,p)Co58		$3.02 \cdot 10^{-1}$	26.7	$3.05 \cdot 10^{-1}$	$1.34 \cdot 10^{-14}$
Zr90(n,2n)Zr89	78.41	$2.71 \cdot 10^{-3}$	49.4	$4.19 \cdot 10^{-3}$	$4.50 \cdot 10^{-17}$
Zr90(n,2n)Zr89		$2.67 \cdot 10^{-3}$	50.3	$4.16 \cdot 10^{-3}$	$4.46 \cdot 10^{-17}$
Zr90(n,2n)Zr89		$1.92 \cdot 10^{-3}$	50.7	$3.01 \cdot 10^{-3}$	$3.23 \cdot 10^{-17}$
Cu63(n, γ)Cu64	12.7	2.69	173.9	$3.56 \cdot 10^4$	$2.42 \cdot 10^{-11}$
Cu63(n, γ)Cu64		2.02	173.6	$2.64 \cdot 10^4$	$1.89 \cdot 10^{-11}$
Cu63(n, γ)Cu64		2.02	173.3	$2.60 \cdot 10^4$	$1.85 \cdot 10^{-11}$
W186(n, γ)W187	23.72	$9.81 \cdot 10^{-1}$	359.2	$3.55 \cdot 10^4$	$1.28 \cdot 10^{-10}$
W186(n, γ)W187		$7.63 \cdot 10^{-1}$	362.1	$3.01 \cdot 10^4$	$1.16 \cdot 10^{-10}$
W186(n, γ)W187		$6.97 \cdot 10^{-1}$	360.1	$2.59 \cdot 10^4$	$9.72 \cdot 10^{-11}$

Table A.4: Measurements associated with the 6"b beam port irradiation. Measured activity from gamma spectroscopy (A^γ), the back calculated activity at the end of the irradiation A^{irrad} , and the associated saturated activity A^∞ are given for each irradiated sample. In the case of iron, Fe54, Fe56, and Fe58 measurements were derived from the same set of foils.

Activation	Half-Life (h)	A^γ (μCi)	t_w (h)	A^{irrad} (μCi)	A^∞ (decays/N/s)
Al27(n, α)Na24 cd	14.959	$7.91 \cdot 10^{-1}$	27.2	2.78	$4.01 \cdot 10^{-16}$
Al27(n, α)Na24 cd		$7.45 \cdot 10^{-1}$	25.6	2.53	$3.71 \cdot 10^{-16}$
Al27(n, α)Na24		$9.22 \cdot 10^{-1}$	23.8	2.44	$3.59 \cdot 10^{-16}$
Al27(n, α)Na24		$8.10 \cdot 10^{-1}$	24.5	3.08	$4.49 \cdot 10^{-16}$
Al27(n, α)Na24		$9.39 \cdot 10^{-1}$	25.6	2.80	$4.07 \cdot 10^{-16}$
Fe54(n,p)Mn54 cd	7489.44	$7.26 \cdot 10^{-3}$	45.8	$7.29 \cdot 10^{-3}$	$3.36 \cdot 10^{-14}$
Fe54(n,p)Mn54 cd		$5.50 \cdot 10^{-3}$	51.5	$5.53 \cdot 10^{-3}$	$2.58 \cdot 10^{-14}$
Fe54(n,p)Mn54 cd		$7.14 \cdot 10^{-3}$	47.1	$7.17 \cdot 10^{-3}$	$3.28 \cdot 10^{-14}$
Fe54(n,p)Mn54		$6.65 \cdot 10^{-3}$	51.2	$6.68 \cdot 10^{-3}$	$3.05 \cdot 10^{-14}$
Fe54(n,p)Mn54		$6.29 \cdot 10^{-3}$	49.7	$6.32 \cdot 10^{-3}$	$2.89 \cdot 10^{-14}$
Fe58(n, γ)Fe59 cd	1067.88	$2.12 \cdot 10^{-2}$	45.8	$2.19 \cdot 10^{-2}$	$3.23 \cdot 10^{-13}$
Fe58(n, γ)Fe59 cd		$4.59 \cdot 10^{-2}$	51.5	$4.75 \cdot 10^{-2}$	$7.09 \cdot 10^{-13}$
Fe58(n, γ)Fe59 cd		$2.07 \cdot 10^{-2}$	47.2	$2.14 \cdot 10^{-2}$	$3.13 \cdot 10^{-13}$
Fe58(n, γ)Fe59		$4.16 \cdot 10^{-1}$	51.2	$4.30 \cdot 10^{-1}$	$6.29 \cdot 10^{-12}$
Fe58(n, γ)Fe59		$3.97 \cdot 10^{-1}$	49.7	$4.10 \cdot 10^{-1}$	$5.99 \cdot 10^{-12}$
Zr90(n,2n)Zr89 cd	78.41	$3.14 \cdot 10^{-3}$	97.3	$7.42 \cdot 10^{-3}$	$8.36 \cdot 10^{-17}$
Zr90(n,2n)Zr89 cd		$2.79 \cdot 10^{-3}$	98.7	$6.67 \cdot 10^{-3}$	$7.59 \cdot 10^{-17}$
Zr90(n,2n)Zr89		$2.36 \cdot 10^{-3}$	98.4	$6.67 \cdot 10^{-3}$	$7.59 \cdot 10^{-17}$
Zr90(n,2n)Zr89		$3.09 \cdot 10^{-3}$	97.6	$7.33 \cdot 10^{-3}$	$8.33 \cdot 10^{-17}$
Zr90(n,2n)Zr89		$3.02 \cdot 10^{-3}$	98.0	$7.18 \cdot 10^{-3}$	$8.11 \cdot 10^{-17}$
Cu63(n, γ)Cu64 cd	12.7	8.18	98.9	$1.81 \cdot 10^3$	$1.38 \cdot 10^{-12}$
Cu63(n, γ)Cu64 cd		5.28	99.3	$1.19 \cdot 10^3$	$9.15 \cdot 10^{-13}$
Cu63(n, γ)Cu64 cd		$2.14 \cdot 10^{-1}$	167.2	$1.97 \cdot 10^3$	$1.51 \cdot 10^{-12}$
Cu63(n, γ)Cu64		4.70	167.6	$4.41 \cdot 10^4$	$3.33 \cdot 10^{-11}$
Cu63(n, γ)Cu64		5.50	168.0	$5.27 \cdot 10^4$	$4.01 \cdot 10^{-11}$

Table A.5: Measurements associated with the 8" beam port irradiation. Measured activity from gamma spectroscopy (A^γ), the back calculated activity at the end of the irradiation A^{irrad} , and the associated saturated activity A^∞ are given for each irradiated sample. In the case of iron, Fe54, Fe56, and Fe58 measurements were derived from the same set of foils.

A.3.3 Measurements of the Pneumatic Rabbit System Flux

Due to the small size of the pneumatic rabbit system, we are unable to perform a full set of simultaneous foil measurements to unfold the neutron energy spectrum. Additionally, the containers in which we place samples when sending them to/from the reactor core are free to rotate inside of the pneumatic tube. This can lead to systematic differences in measurements between exposure, especially if the foils are rotated in such a way that the face of the foils are rotated 90° with respect to the incident fission neutron flux. In such instances, we can expect that some amount of self-shielding can take place within the foil, leading to lower measured neutron fluences. To address these issues, a modified approach was used to determine the energy spectrum of the rabbit system where instead of measuring a large variety of foils, a large number of irradiations each consisting of a single iron foil and a D0 diode were performed. Because naturally occurring iron can undergo the $\text{Fe54}(n,p)\text{Mn54}$, $\text{Fe56}(n,p)\text{Mn56}$, and $\text{Fe58}(n,\gamma)\text{Fe59}$ activation processes, a single foil gives us three data points that can be used to determine the neutron fluence. Additionally, the $\text{Fe58}(n,\gamma)\text{Fe59}$ activation cross section is dominated by thermal neutrons, while the $\text{Fe54}(n,p)\text{Mn54}$ and $\text{Fe56}(n,p)\text{Mn56}$ activation cross sections are dominated by fission neutrons. This allows a single iron foil to give us independent information about the thermal and fission neutron fluences. The addition of the D0 diode then allows us to have a second independent method to measure the neutron fluence which we can compare the iron foil measurement against.

Two sets of rabbit system flux measurements were performed, the first during 2017 and the second during the years of 2018–2019. In both instances, the irradiation time was spanned over a large array of values to determine the consistency of flux measurements across varying exposure times. Three separate neutron flux measurements were then determined for each irradiation. From the foil measurements, two separate flux measurements were obtained based on the measured flux of the 6"b and 8" beamports. The flux is measured by re-fitting the derived beamport measurements to the Fe54 , Fe56 , and Fe58 activation measurements. This is performed by first creating a new shared normalization to the thermal and intermediate spectrums obtained from the beamport fit. The normalization is then adjusted such that the ratio of measured to calculated saturated activity for the $\text{Fe58}(n,\gamma)\text{Fe59}$ activation process is equal to unity. The normalization of the fission spectrum is then independently modified such that the average ratio of measured to calculated saturated activity for the $\text{Fe54}(n,p)\text{Mn54}$ and $\text{Fe56}(n,p)\text{Mn56}$ activation processes is equal to unity. This procedure is carried out independently for the 6"b and 8" beamports, and the two potential energy spectra produced from each fit are taken as two separate foil based energy spectrum hypotheses, and are

Rad Time (min)	Rad Date	Activation Process	A_{meas}^{∞}	6"b A_{calc}^{∞}	8" A_{calc}^{∞}
5.8	8/17/17	Fe54(n,p)Mn54	$1.58 \cdot 10^{-14}$	$1.65 \cdot 10^{-14}$	$1.65 \cdot 10^{-14}$
		Fe56(n,p)Mn56	$2.05 \cdot 10^{-16}$	$1.97 \cdot 10^{-16}$	$1.97 \cdot 10^{-16}$
19.2	8/17/17	Fe54(n,p)Mn54	$1.22 \cdot 10^{-14}$	$1.44 \cdot 10^{-14}$	$1.49 \cdot 10^{-14}$
		Fe56(n,p)Mn56	$1.99 \cdot 10^{-16}$	$1.72 \cdot 10^{-16}$	$1.78 \cdot 10^{-16}$
		Fe58(n, γ)Fe59	$1.36 \cdot 10^{-13}$		$1.27 \cdot 10^{-13}$
28.8	8/17/17	Fe54(n,p)Mn54	$1.41 \cdot 10^{-14}$	$1.35 \cdot 10^{-14}$	$1.51 \cdot 10^{-14}$
		Fe56(n,p)Mn56	$2.14 \cdot 10^{-16}$	$1.62 \cdot 10^{-16}$	$1.81 \cdot 10^{-16}$
		Fe58(n, γ)Fe59	$1.56 \cdot 10^{-13}$		$1.29 \cdot 10^{-13}$
57.5	8/17/17	Fe54(n,p)Mn54	$1.08 \cdot 10^{-14}$	$1.25 \cdot 10^{-14}$	$1.61 \cdot 10^{-14}$
		Fe56(n,p)Mn56	$1.70 \cdot 10^{-16}$	$1.50 \cdot 10^{-16}$	$1.92 \cdot 10^{-16}$
		Fe58(n, γ)Fe59	$1.98 \cdot 10^{-13}$		$1.37 \cdot 10^{-13}$
145.7	8/17/17	Fe54(n,p)Mn54	$1.20 \cdot 10^{-14}$	$1.38 \cdot 10^{-14}$	$1.50 \cdot 10^{-14}$
		Fe56(n,p)Mn56	$1.85 \cdot 10^{-16}$	$1.64 \cdot 10^{-16}$	$1.79 \cdot 10^{-16}$
		Fe58(n, γ)Fe59	$1.49 \cdot 10^{-13}$		$1.28 \cdot 10^{-13}$
28.8	9/20/17	Fe54(n,p)Mn54	$1.05 \cdot 10^{-14}$	$1.31 \cdot 10^{-14}$	$1.99 \cdot 10^{-14}$
		Fe56(n,p)Mn56	$1.79 \cdot 10^{-16}$	$1.56 \cdot 10^{-16}$	$2.38 \cdot 10^{-16}$
		Fe58(n, γ)Fe59	$2.87 \cdot 10^{-13}$		$1.70 \cdot 10^{-13}$
57.5	10/16/17	Fe54(n,p)Mn54	$1.24 \cdot 10^{-14}$	$1.27 \cdot 10^{-14}$	$1.50 \cdot 10^{-14}$
		Fe58(n, γ)Fe59	$1.47 \cdot 10^{-13}$		$1.28 \cdot 10^{-13}$
145.7	10/16/17	Fe54(n,p)Mn54	$1.59 \cdot 10^{-14}$	$8.69 \cdot 10^{-15}$	$1.15 \cdot 10^{-14}$
		Fe56(n,p)Mn56	$1.84 \cdot 10^{-17}$	$1.04 \cdot 10^{-16}$	$1.38 \cdot 10^{-16}$
		Fe58(n, γ)Fe59	$1.47 \cdot 10^{-13}$		$9.84 \cdot 10^{-14}$
Average	N/A	Fe54(n,p)Mn54	$(1.30 \pm 0.21) \cdot 10^{-14}$	$(1.31 \pm 0.22) \cdot 10^{-14}$	$(1.31 \pm 0.22) \cdot 10^{-14}$
		Fe56(n,p)Mn56	$(1.61 \pm 0.72) \cdot 10^{-16}$	$(1.51 \pm 0.25) \cdot 10^{-16}$	$(1.51 \pm 0.25) \cdot 10^{-16}$
		Fe58(n, γ)Fe59	$(1.74 \pm 0.54) \cdot 10^{-13}$		$(1.32 \pm 0.23) \cdot 10^{-13}$

Table A.6: The 2017 saturated activity values measured and obtained through normalized beam port fit of the saturated activities (in decays/nuclei/s) used to measure the rabbit system flux. Because the foils were encased in cadmium, the Fe58(n, γ)Fe59 activation processes was not used to calculate the 6"b beam port normalization.

both used to calculate an effective 1 MeV equivalent neutron fluence and flux. The D0 diode then gives an independent measurement of the 1 MeV equivalent neutron fluence and flux based on the methods described in Section A.3.1.

The sets of foil measurements from the rabbit system irradiations can be found in Table A.6 and A.7. For the 2017 irradiations the D0 and foils were encased in cadmium. We found that because there were no foils encased in cadmium used when unfolding the 6"b beamport spectrum, this lead to a systematic disagreement in the normalization of the 6"b beamport's thermal/intermediate spectrum relative to the rabbit system data. The full set of flux and fluence measurements from the iron foils and D0 diodes as well as the desired fluence for the 2017 and 2018–2019 rabbit system irradiations can be found in Table A.8 and A.9. From the initial set of 2017 measurements, we found that the foil measurements systematically measured fluences that were about half the levels of the diode measurements. However, from the 2018–2019 measurements we find good agreement between the beamport and diode based measurements, especially when comparing the flux obtained from re-normalizing the 8" beam port measurement. As such, we attribute the disagreement observed in 2017 to be a result of the cadmium shielding causing unequal responses from the foil and diode

Rad Time (min)	Rad Date	Activation Process	A_{meas}^{∞}	6"b A_{calc}^{∞}	8" A_{calc}^{∞}
4.5	2/1/2019	Fe56(n,p)Mn56 Fe58(n, γ)Fe59	$5.91 \cdot 10^{-17}$ $5.24 \cdot 10^{-12}$	$1.48 \cdot 10^{-16}$ $3.28 \cdot 10^{-12}$	$2.05 \cdot 10^{-16}$ $3.06 \cdot 10^{-12}$
4.5	3/7/2019	Fe56(n,p)Mn56 Fe58(n, γ)Fe59	$2.19 \cdot 10^{-16}$ $4.00 \cdot 10^{-12}$	$2.0 \cdot 10^{-16}$ $4.43 \cdot 10^{-12}$	$2.00 \cdot 10^{-16}$ $4.00 \cdot 10^{-12}$
13.5	2/13/2019	Fe56(n,p)Mn56 Fe58(n, γ)Fe59	$5.03 \cdot 10^{-16}$ $5.31 \cdot 10^{-12}$	$3.71 \cdot 10^{-16}$ $8.23 \cdot 10^{-12}$	$4.29 \cdot 10^{-16}$ $5.07 \cdot 10^{-12}$
19.03	12/19/2018	Fe54(n,p)Mn54 cd Fe56(n,p)Mn56 cd Fe58(n, γ)Fe59 cd	$1.44 \cdot 10^{-14}$ $2.24 \cdot 10^{-16}$ $2.81 \cdot 10^{-13}$	$1.14 \cdot 10^{-13}$ $1.35 \cdot 10^{-15}$ $1.04 \cdot 10^{-13}$	$2.24 \cdot 10^{-14}$ $2.29 \cdot 10^{-16}$ $1.94 \cdot 10^{-13}$
19.03	12/19/2018	Fe54(n,p)Mn54 Fe56(n,p)Mn56 Fe58(n, γ)Fe59	$1.16 \cdot 10^{-14}$ $2.78 \cdot 10^{-16}$ $4.49 \cdot 10^{-12}$	$1.82 \cdot 10^{-14}$ $2.06 \cdot 10^{-16}$ $4.59 \cdot 10^{-12}$	$2.07 \cdot 10^{-14}$ $2.39 \cdot 10^{-16}$ $3.68 \cdot 10^{-12}$
27	2/13/2019	Fe54(n,p)Mn54 Fe56(n,p)Mn56 Fe58(n, γ)Fe59	$2.50 \cdot 10^{-14}$ $1.20 \cdot 10^{-16}$ $5.35 \cdot 10^{-12}$	$1.84 \cdot 10^{-14}$ $2.20 \cdot 10^{-16}$ $4.88 \cdot 10^{-12}$	$2.17 \cdot 10^{-14}$ $2.59 \cdot 10^{-16}$ $3.86 \cdot 10^{-12}$
38.05	12/19/2018	Fe54(n,p)Mn54 cd Fe56(n,p)Mn56 cd Fe58(n, γ)Fe59 cd	$1.73 \cdot 10^{-14}$ $5.50 \cdot 10^{-15\dagger}$ $1.96 \cdot 10^{-13}$	$1.14 \cdot 10^{-13}$ N/A $1.05 \cdot 10^{-13}$	$1.99 \cdot 10^{-14}$ N/A $1.72 \cdot 10^{-13}$
38.05	12/19/2018	Fe54(n,p)Mn54 Fe56(n,p)Mn56 Fe58(n, γ)Fe59	$1.94 \cdot 10^{-14}$ $2.66 \cdot 10^{-16}$ $4.05 \cdot 10^{-12}$	$1.72 \cdot 10^{-14}$ $2.25 \cdot 10^{-16}$ $4.56 \cdot 10^{-12}$	$2.09 \cdot 10^{-14}$ $2.55 \cdot 10^{-16}$ $3.71 \cdot 10^{-12}$
38.05	11/07/2018	Fe54(n,p)Mn54 Fe56(n,p)Mn56 Fe58(n, γ)Fe59	$3.14 \cdot 10^{-14}$ $4.29 \cdot 10^{-16}$ $6.90 \cdot 10^{-12}$	$3.15 \cdot 10^{-14}$ $3.77 \cdot 10^{-16}$ $7.95 \cdot 10^{-12}$	$3.53 \cdot 10^{-14}$ $4.23 \cdot 10^{-16}$ $6.29 \cdot 10^{-12}$
38.05	11/07/2018	Fe54(n,p)Mn54 Fe56(n,p)Mn56 Fe58(n, γ)Fe59	$1.98 \cdot 10^{-14}$ $2.91 \cdot 10^{-16}$ $3.99 \cdot 10^{-12}$	$2.00 \cdot 10^{-14}$ $2.39 \cdot 10^{-16}$ $5.04 \cdot 10^{-12}$	$2.22 \cdot 10^{-14}$ $2.65 \cdot 10^{-16}$ $3.95 \cdot 10^{-12}$
60	3/8/2019	Fe54(n,p)Mn54 Fe56(n,p)Mn56 Fe58(m, γ)Mn59	$2.32 \cdot 10^{-14}$ $1.26 \cdot 10^{-16}$ $4.24 \cdot 10^{-12}$	$1.49 \cdot 10^{-14}$ $1.79 \cdot 10^{-16}$ $3.97 \cdot 10^{-12}$	$1.80 \cdot 10^{-14}$ $2.16 \cdot 10^{-16}$ $3.21 \cdot 10^{-12}$
76.10	11/07/2018	Fe54(n,p)Mn54 Fe56(n,p)Mn56 Fe58(n, γ)Fe59	$2.88 \cdot 10^{-14}$ $4.04 \cdot 10^{-16}$ $6.46 \cdot 10^{-12}$	$2.94 \cdot 10^{-14}$ $3.52 \cdot 10^{-16}$ $7.41 \cdot 10^{-12}$	$3.30 \cdot 10^{-14}$ $3.94 \cdot 10^{-16}$ $5.87 \cdot 10^{-12}$
76.10	11/07/2018	Fe54(n,p)Mn54 Fe56(n,p)Mn56 Fe58(n, γ)Fe59	$2.23 \cdot 10^{-14}$ $3.72 \cdot 10^{-16}$ $4.96 \cdot 10^{-12}$	$2.44 \cdot 10^{-14}$ $2.91 \cdot 10^{-16}$ $6.14 \cdot 10^{-12}$	$2.71 \cdot 10^{-14}$ $3.24 \cdot 10^{-16}$ $4.82 \cdot 10^{-12}$
90	3/8/2019	Fe54(n,p)Mn54 Fe56(n,p)mn56 Fe58(n, γ)Fe59	$1.76 \cdot 10^{-14}$ $2.65 \cdot 10^{-16}$ $4.55 \cdot 10^{-12}$	$1.6 \cdot 10^{-14}$ $1.98 \cdot 10^{-16}$ $4.39 \cdot 10^{-12}$	$1.97 \cdot 10^{-14}$ $2.35 \cdot 10^{-16}$ $3.50 \cdot 10^{-12}$
Average ^{††}	N/A	Fe54(n,p)Mn54 Fe56(n,p)Mn56 Fe58(n, γ)Fe59	$(2.21 \pm 0.60) \cdot 10^{-14}$ $(2.78 \pm 1.34) \cdot 10^{-16}$ $(5.04 \pm 0.95) \cdot 10^{-12}$	$(2.12 \pm 0.59) \cdot 10^{-14}$ $(2.51 \pm 0.78) \cdot 10^{-16}$ $(5.51 \pm 1.70) \cdot 10^{-12}$	$(2.43 \pm 0.61) \cdot 10^{-14}$ $(2.87 \pm 0.84) \cdot 10^{-16}$ $(3.98 \pm 1.64) \cdot 10^{-12}$

Table A.7: The 2018–2019 saturated activity values measured and obtained through normalized beam port fit of the saturated activities (in decays/nuclei/s) used to measure the rabbit system flux. [†]Mn56 activity measurement appeared orders of magnitude higher than expected, and did not agree with activity Mn54 and Fe59 activities measured on the same foil. As such, those activity values were not used in the fit and are omitted from the average values quoted in the table.

^{††}Foils encased in cadmium were not used when computing the average activity values for these foils

Rad Time (min)	Rad Date	6"b Flux	8" Flux	Diode Flux	6"b Fluence	8" Fluence	Diode Fluence	Desired Fluence
5.8	8/17/17	2.25-10 ¹¹	2.84-10 ¹¹	4.48-10 ¹¹	7.83-10 ¹³	9.89-10 ¹³	1.56-10 ¹⁴	1.50-10 ¹⁴
19.2	8/17/17	1.97-10 ¹¹	2.48-10 ¹¹	3.96-10 ¹¹	2.26-10 ¹⁴	2.86-10 ¹⁴	4.56-10 ¹⁴	5.00-10 ¹⁴
28.8	8/17/17	1.85-10 ¹¹	2.33-10 ¹¹	4.78-10 ¹¹	3.19-10 ¹⁴	4.03-10 ¹⁴	8.26-10 ¹⁴	7.50-10 ¹⁴
57.5	8/17/17	1.71-10 ¹¹	2.16-10 ¹¹	3.48-10 ¹¹	5.89-10 ¹⁴	7.44-10 ¹⁴	1.20-10 ¹⁵	1.50-10 ¹⁵
145.7	8/17/17	1.88-10 ¹¹	2.37-10 ¹¹	3.87-10 ¹¹	1.64-10 ¹⁵	2.07-10 ¹⁵	3.38-10 ¹⁵	3.80-10 ¹⁵
28.8	9/20/17	1.78-10 ¹¹	2.25-10 ¹¹	3.48-10 ¹¹	3.08-10 ¹⁴	3.89-10 ¹⁴	6.02-10 ¹⁵	7.50-10 ¹⁴
57.5	10/16/17	1.74-10 ¹¹	2.19-10 ¹¹	4.20-10 ¹¹	5.99-10 ¹⁴	7.57-10 ¹⁴	1.45-10 ¹⁵	1.50-10 ¹⁵
145.7	10/16/17	1.19-10 ¹¹	1.50-10 ¹¹	4.71-10 ¹¹	1.04-10 ¹⁵	1.31-10 ¹⁵	4.12-10 ¹⁵	3.80-10 ¹⁵

Table A.8: 2017 rabbit system flux and fluence measurements from the 6"b and 8" beamport re-normalizations, D0 diode measurements, and the desired values. The Fe58(n, γ)Fe59 activation measurement were not used when computing the flux based on the 6"b beamport measurement. The flux is shown in units of $n_{eq}/cm^2/s$ and the fluence is in units of n_{eq}/cm^2 . Measurements are in the same order as Table A.6.

Rad Time (min)	Rad Date	6"b Flux	8" Flux	Diode Flux	6"b Fluence	8" Fluence	Diode Fluence	Desired Fluence
4.5	2/1/2019	1.68-10 ¹¹	2.96-10 ¹¹	3.16-10 ¹¹	4.55-10 ¹³	8.00-10 ¹³	8.54-10 ¹³	1.00-10 ¹⁴
4.5	3/7/2019	2.28-10 ¹¹	3.51-10 ¹¹	2.84-10 ¹¹	6.15-10 ¹³	9.49-10 ¹³	7.66-10 ¹³	1.00-10 ¹⁴
13.5	2/13/2019	4.23-10 ¹¹	6.19-10 ¹¹	2.64-10 ¹¹	3.42-10 ¹⁴	5.01-10 ¹⁴	2.14-10 ¹⁴	3.00-10 ¹⁴
19.03	12/19/2018	1.56-10 ^{12†}	3.82-10 ¹¹	3.31-10 ¹¹	1.78-10 ^{15†}	4.36-10 ¹⁴	3.78-10 ¹⁴	2.50-10 ¹⁴
19.03	12/19/2018	2.35-10 ¹¹	3.57-10 ¹¹	3.57-10 ¹¹	2.69-10 ¹⁴	4.07-10 ¹⁴	4.08-10 ¹⁴	2.50-10 ¹⁴
27	2/13/2019	2.51-10 ¹¹	3.74-10 ¹¹	2.40-10 ¹¹	4.07-10 ¹⁴	6.06-10 ¹⁴	3.88-10 ¹⁴	6.00-10 ¹⁴
38.05	12/19/2018	1.58-10 ^{12†}	3.47-10 ¹¹	3.33-10 ¹¹	3.61-10 ^{15†}	7.92-10 ¹⁴	7.61-10 ¹⁴	5.00-10 ¹⁴
38.05	12/19/2018	2.57-10 ¹¹	3.67-10 ¹¹	3.29-10 ¹¹	5.87-10 ¹⁴	8.39-10 ¹⁴	7.51-10 ¹⁴	5.00-10 ¹⁴
38.05	11/07/2018	4.24-10 ¹¹	6.09-10 ¹¹	4.94-10 ¹¹	9.68-10 ¹⁴	1.39-10 ¹⁵	1.13-10 ¹⁵	5.00-10 ¹⁴
38.05	11/07/2018	2.69-10 ¹¹	3.82-10 ¹¹	Diode Broke	6.14-10 ¹⁴	8.73-10 ¹⁴	Diode Broke	5.00-10 ¹⁴
60††	3/8/2019	2.04-10 ¹¹	3.11-10 ¹¹	3.33-10 ¹¹	7.34-10 ¹⁴	1.12-10 ¹⁵	1.20-10 ¹⁵	1.00-10 ¹⁵
76.10	11/07/2018	4.01-10 ¹¹	5.68-10 ¹¹	3.61-10 ¹¹	1.83-10 ¹⁵	2.59-10 ¹⁵	1.65-10 ¹⁵	1.00-10 ¹⁵
76.10	11/07/2018	3.28-10 ¹¹	4.67-10 ¹¹	3.81-10 ¹¹	1.50-10 ¹⁵	2.13-10 ¹⁵	1.74-10 ¹⁵	1.00-10 ¹⁵
90††	3/8/2019	2.25-10 ¹¹	3.39-10 ¹¹	3.35-10 ¹¹	1.22-10 ¹⁵	1.83-10 ¹⁵	1.81-10 ¹⁵	1.50-10 ¹⁵

Table A.9: 2018–2019 rabbit system flux and fluence measurements from the 6"b and 8" beamport re-normalizations, D0 diode measurements, and the desired values. The flux is calculated in units of $n_{eq}/cm^2/s$ and the fluence is in units of n_{eq}/cm^2 . Measurements are in the same order as Table A.7.

†foils encased in cadmium not used when computing 6"b beamport fluence estimates.

††Irradiation time needed to reach the desired fluence assumed a lower n_{eq}/cm^2 flux.

based measurements. One additional observation made from the rabbit system measurements is how the accuracy of the foil based measurements breaks down during low fluence measurements. In particular, the foil based measurements were found to be inconsistent for neutron exposures shorter than ~ 15 minutes.

A.3.4 Continual Monitoring of Neutron Fluence using Iron Foils

Based on the results of the initial neutron spectrum measurements, multiple irradiation measurements have taken place at the RINSC reactor. These include 8" beamport irradiations of HGCal sensors and rabbit system irradiations of test sensors produced on the wafers from which tracker upgrade components are manufactured. In all of these irradiations, fluence is continually monitored by including D0 diodes and iron foils with the irradiated samples. In addition to ensuring that the conditions of the reactor have not changed across irradiations, this is also of importance for

beamport irradiations in which multiple sensors are placed in the beamport. We find that sensors in the front of the holders receive a higher fluence than further back sensors. For such irradiations, multiple iron foils are included to monitor the fluence at different distances from the front of the sensor holder. When determining the fluence measured from iron foils, we have opted to use the re-normalization method designed for measuring the rabbit system flux. However, due to the better agreement between diode and foil measurements obtained from the 8" beamport normalization, we have dropped the flux measurements based on the 6"b beamport energy spectrum. Current plans are to continue monitoring the fluence in this way throughout the remainder of the RINSC irradiations planned for testing of the Phase II CMS detector components.

Appendix B

Monitoring Silicon Strip Readouts Using Spy Data Taking

The CMS outer tracker system produces an insurmountable amount of data. There are 24244 silicon sensors in its current iteration, and with each sensor having 512 or 768 strips this leads to approximately 10,000,000 individual strips in the outer tracker. If the current of a single strip is stored as a byte, then strip currents alone leads to over 10 MB of strip data corresponding to a single bunch crossing. At an L1 trigger rate of 10 kHz, this would require a transfer rate 100 GB/s to send all of the strip readouts through the CMS data acquisition system (DAQ). The true amount of data would then be even larger due to additional data related to the sensors corresponding with the various strip currents. In reality, the CMS detector makes use of an advanced zero-suppression algorithm integrated into the front-end driver (FED) [127] used as part of the outer tracker data readout system. From this algorithm, the FED is able determine which strip currents were produced from a minimum ionizing particle (MIP), and only send the readouts related to those strips to the DAQ. After zero-suppression, the total transfer rate of the outer tracker is reduced to 25–150 MB/s depending on the PU conditions.

From the zero-suppression algorithm, there is an unavoidable loss of data. However, a second system is in place to read data from the FEDs in a *scope* mode in which a full readout of the raw tracker information can be sent to the DAQ and saved. The *spy channel* exists to periodically read out the outer tracker in scope mode during normal data taking. This *spy data* can then be analyzed offline to better understand properties of the raw strip data. In particular, spy data gives a way to perform an offline verification of the assumptions about raw strip data and its defects that the

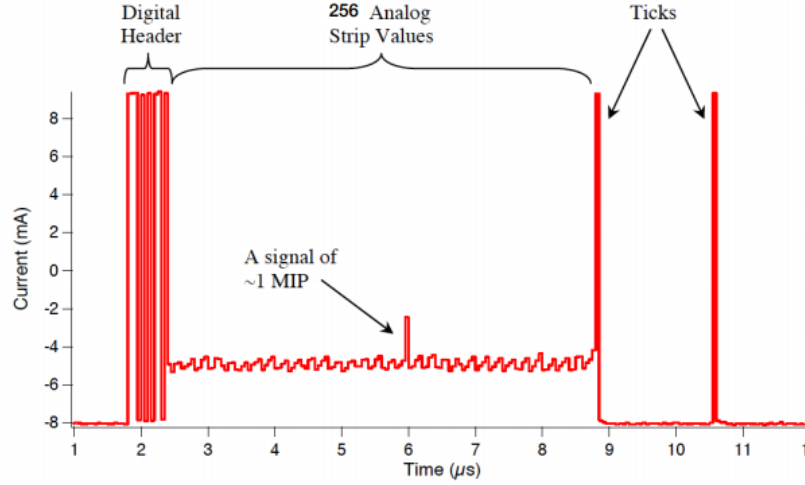


Figure B.1.1: The data stream readout from a pair of APV25 chips with a minimum ionizing particle (MIP) signal injected. The digital header is used to inform the FED that input data is coming as well as signify the corresponding address of the APV25 pair. The tails are then used to inform the FED that the data stream has ended, with the second tail existing for redundancy.

zero-suppression algorithm is otherwise forced to blindly assume.

B.1 The Outer Tracker Readout System

As discussed in Section 3.2.1.1 the outer tracker system makes use of silicon strip sensors. As particles pass through a sensor, they will produce localized currents that will be collected along the strips and sent to one of multiple APV25 readout chips [128] located along the edge of the sensor. Each APV25 chip is connected to 128 consecutive strips, amplifies the currents at a sampling rate of 40 MHz, and stores each strip current in a buffer capable of holding strip readouts for the previous 32 events. Upon receiving a trigger signal, the corresponding event will be read out from pairs of APV25 chip buffers and multiplexed together to produce a 256 strip data stream, an example of which is shown in Figure B.1.1. In the data stream, a header signifying the start of the data stream and address of the pipeline is first produced followed by 256 25 ns analogue samples corresponding to various strip currents for the triggered event. The strip signals produce a 7 μ s long analogue signal that is then followed by two additional ticks signifying the end of the data stream. A linear laser drive (LLD) collects the 2-3 APV25 data streams from its respective sensor and sends the information down separate channels to the FEDs for the next stage of data readout.

Each FED contains 8 optical receiver units each collecting data streams from 12 separate LLD channels. The optical receiver transfers the data inputs to copper wiring via an optical receiver

module [129] and then to analogue to digital converters (ADCs), each of which is controlled by a series of three VIRTEX-II XC2V4 delay FPGAs. After the data streams have been digitized, they are each sent to one of eight VIRTEX-II XC2V1500 FPGAs (front-end FPGAs) to perform the zero-suppression algorithm. The zero-suppressed data is sent to a final back-end VIRTEX-II XC2V2000 FPGA which constructs an event header, combines it with the package of strip data, and stores it in a 2 MB storage buffer. A VME is then used to control the read out of the storage buffer to the CMS DAQ, and waits for a virtual handshake signifying that the DAQ is ready to receive the strip data. After receiving the handshake, the FED will send the full data package to the CMS DAQ via a Front-end Readout Link [130] which are able to run at a maximum 400 MB/s data transfer rate. Once received by the DAQ, the individual event fragments from each of the FEDs will be combined to obtain the full set of outer tracker information for the event and sent to the HLT.

B.1.1 Tracker Zero Suppression Algorithm

In order to perform zero-suppression, the front-end FPGAs must decide whether each of the strip currents are originating from a MIP passing through the sensor or the result of the naturally occurring electrical noise which can originate from the strip sensor or the readout system. However, the signal which arrives at the FPGAs undergoes various deformations that occurs during the readout steps. As such, the zero-suppression algorithm has two primary stages: To first restore the strip sensor signal to be as close as possible to its original form, and to then determine whether the various strip currents are coming from a MIP or electrical noise. The strip information arrives at front-end FPGAs as an array of bytes where each strip is assigned a value of 0-255 as a digitized representation of the strip current. The first step in the zero-suppression algorithm is to reorder the digitized strip currents, such that they are arranged according to strip number. This is because the strip ordering of the analogue data stream produced from the multiplexed APV25 chips is not in sequential order, but an order optimized to maximize the read speed from the memory buffer. After reordering, it is found that for each of the 128 strips pertaining to the strip data read from a single APV25 chip, the ADC count has a linear dependence on the strip number. This slope in the ADC count (the digitized unit of strip current) is the product of the chip design, where the low voltage required to power the chip creates a small potential difference across the strip channels. This potential difference manifests itself as a pedestal offset, that gets individually subtracted off from the strip ADC count distributions for each of the APV25 chips. Once the pedestal has been subtracted off, there still exists a flat offset in the ADC count distribution referred to as the common mode. This common

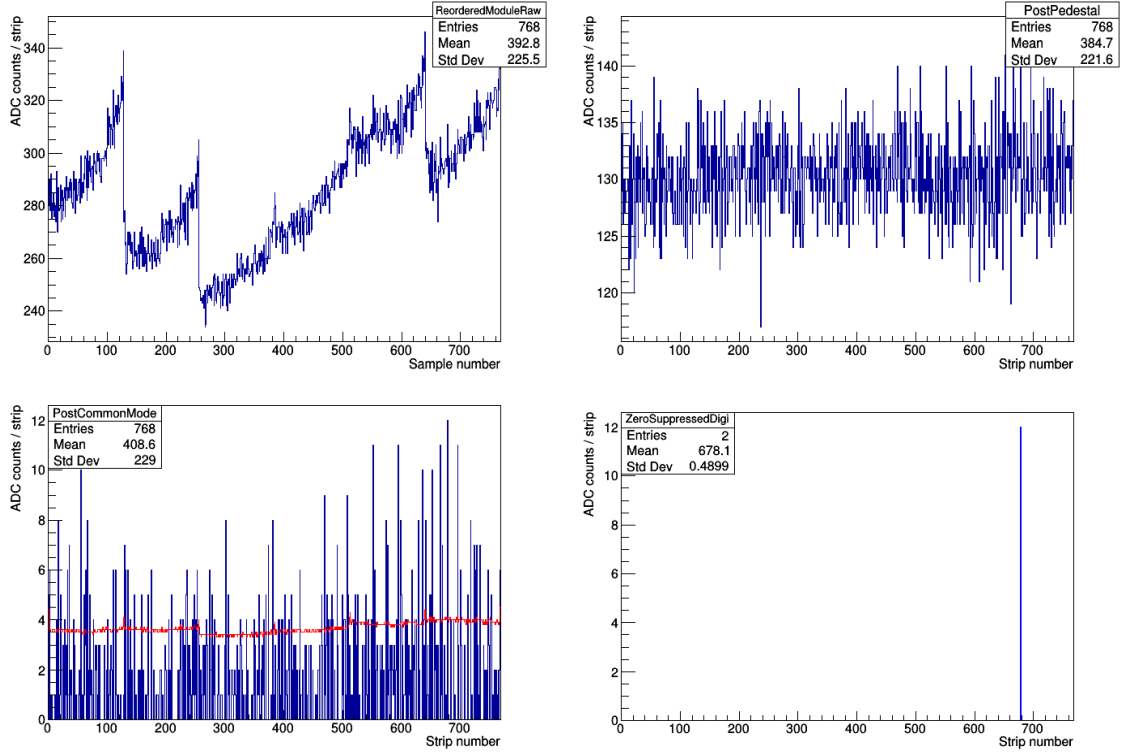


Figure B.1.2: The various forms that the strip data takes after being re-ordered (top left), having the pedestal subtracted (top right), having the common mode subtracted (bottom left), and after the full zero-suppression algorithm (bottom right). The post common mode subtracted distribution is plotted beside the assumed strip noise (red). In the zero-suppressed form, only information about strips with detected hits are sent to the DAQ.

mode offset is removed by subtracting the average ADC count for each multiplexed pair of APV25 chips from the ADC count of each of their corresponding strips. For the common mode subtraction procedure, negative ADC counts are truncated at zero.

From the final common mode subtracted ADC count distribution, the restored strip currents are obtained. To determine whether the ADC count of a strip corresponds with a MIP, they are compared to an existing distribution of known strip noise. In the situation that the ADC count is 3σ above the noise, or two neighboring strips both have an ADC count 2σ above the noise, the ADC count is extracted as a zero-suppressed tracker hit. The collection of zero-suppressed tracker hits is what is then sent to the back-end FPGAs. The distribution of ADC counts after the reordering, pedestal subtraction, common mode subtraction, and the full zero-suppression algorithm is shown in Figure B.1.2. For the post common mode subtracted ADC count, the measured noise distribution for the corresponding sensor is additionally plotted in red. To obtain information about the strip noise, commissioning runs in which the outer tracker is read out during no colliding beams is regularly

performed in between data taking runs. The noise is computed as the root mean squared of the common mode subtracted strip ADC counts for each of the strips.

B.2 Spy Data Taking

After zero-suppression, the rate at which data is sent from the FEDs to the central DAQ becomes well within the acceptable throughput of the Front-end Readout Link. As such, the entire tracker can be periodically reading out at a low rate such that it would have no impact on the CMS detectors ability to transfer the zero-suppressed data from events passing the L1 trigger system. The spy channel exists as separate data taking channel inside of the FEDs to preserve the raw data stream without going through ADC conversion or zero-suppression.

The spy channel is kicked off by a special configuration in the three delay FPGA chips used to control each of the ADCs. In this configuration, they wait for a specified signal meant to go off at specific bunch crossing in the train known as a BGo signal. With the currently taken spy channel data, the BGo signal has been set such that an arming signal occurs at a frequency of approximately 3.9 Hz. After receiving the arming signal, the FPGA chips are told to additionally copy the data stream from each of the LLD channels in scope mode, which does not receive the analogue to digital conversion or undergo any processing, on the front-end FPGAs. The scope mode data instead gets directly stored in a special partition of the 2 MB storage buffer in the back-end FPGA referred to as the spy buffer. The VME can then be sent a separate signal signifying that the next event in the spy buffer should be read out to a separate spy DAQ which constructs the events from the FED fragments and copies them to a specialized virtual machine (VM) storage area used to house outer tracker commissioning data.

The actions of the delay FPGAs to arm based on specific BGo signals and save scope mode data into the spy buffer is something that is now fully automated, as its functionality does not impede with standard data taking. However, the spy DAQ will not send the VME signals to read the spy buffer unless a specialized spy client has been activated by a user. All scope mode events collected in this way are internally labeled as spy data and packed together into a single spy run on the commissioning VM. As long as BGo and L1 triggers are being sent to the outer tracker, the spy run will continue to grow until the spy client is manually turned off. In this way, a single spy run can often consist of multiple different global runs, run types, and even proton fills.

Spy data has one primary draw back originating from the distance between the source of the arming signal and the location of the crates which hold the various FEDs. Due to the rapid rate

of the L1 trigger system, the differences in the cables lengths used to carry the BGo signal to the FEDs as well as the speeds of the delay FPGAs can cause a single arming signal to reach the FEDs at different times. This can lead to a single arm signal causing different FEDs to copy data in scope mode for different L1 trigger signals. When the spy DAQ builds the spy event out of the corresponding FED fragments, it is often the case that sets of fragments will correspond with the tracker output for different sets of global events. This is found to have little effect on the usage of spy data for detector commissioning purposes. However, the mix-match of global events in spy runs makes the data type unusable for physics analysis in its current form.

B.2.1 Automated Processing and Transferring of Spy Data

After being transferred to the VMs, the spy data exists in a raw data format not yet suitable for use in the standard CMS software framework. Instead, a separate set of software located on the VM nodes needs to be run to repackage the raw files into the event data model (EDM) framework [131]. Additionally, the computing machine which houses the VM is not accessible for the general CMS community, requiring a need to transfer the spy data to a separate file system before being analyzed. Previously, the repackaging software had to be manually run for each spy run. A unique request made through a Jira support ticketing system then had to be made for each spy run in order to transfer the EDM format file from the machine housing the VM to an EOS storage system accessible from a general use CERN computing center. Due to general inefficiencies in the process required to go from taking spy data to using it for analysis, I was commissioned to write a new code base to automate the prepackaging and transferring of spy data.

Under the new frame work, a python script is run as a systemd-service which activates every 30 minutes to scan the folder used to hold commissioning data. In the instance that it finds a run it has not yet processed, it will first query a SQL database to determine whether or not the run is still on going. For finished runs, it will check to make sure no one has manually transferred the file to EDM format, and if not it will run a secondary script to manually run the conversion. After the conversion has finished, it will run a series of checks to determine that the output file was completed without errors and that there was no loss of spy events during the conversion. In the instance that any of the checks fail, it will attempt to re-run the conversion one additional time. Upon a second failure it will write the reason to a specialized log associated with failed conversions, and append the run to the list of already processed files. Finally, the script will lookup the size and metadata associated with the processed EDM file and write it to a JSON file used to initiate an automated

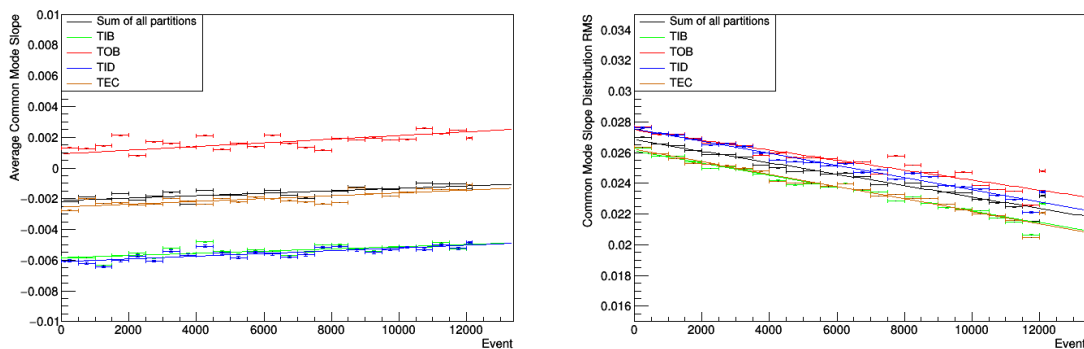


Figure B.3.1: The average (left) and root mean squared (right) of the common mode slope distributions fit for every APV25 chip in the outer tracker during spy run 321779 in bins of 500 spy events. In all partitions, a residual linear trend that increased as the runs went on was observed. However, the overall size of the trends are too small to have an effect on the zero suppression algorithm.

personal data transfer request. The processed file and associated JSON will then automatically be copied to a directory in an EOS file system located in the general CERN computing cluster.

B.3 Monitoring of Common Mode Slope Progression Using Spy Data

One of the many advantages of spy data is that it can be used to get a picture of the CMS outer tracker during live data taking. This lets us better explore the facets of the zero suppression algorithm, which are otherwise based on observations in the ADC count distributions made when outside of standard run conditions. For example, the common mode is expected to exist as a flat offset in the ADC count after the subtraction of the pedestal. However, it is possible that over the course of long global runs defects in the common mode could arise, most likely as the result of the high radiation environment of the LHC. Spy data taking gives us a way to directly verify the validity of these assumptions, and further investigate whether any deformations that do occur are small enough such that they will not impact the performance of the zero-suppression algorithm.

In order to monitor the spy data common mode, an offline emulation of the FED zero-suppression algorithm is run on the spy data. In doing this offline emulation, we are able to save a copy of the ADC count distribution after each step of the zero-suppression algorithm. After the pedestal has been subtracted from the reordered ADC counts, a linear fit is made to the 128 strips of each APV25 chip and the slope of the fit is saved. Additional information such as the size of the common mode (average ADC count) and root mean squared (RMS) of the ADC count distribution for each

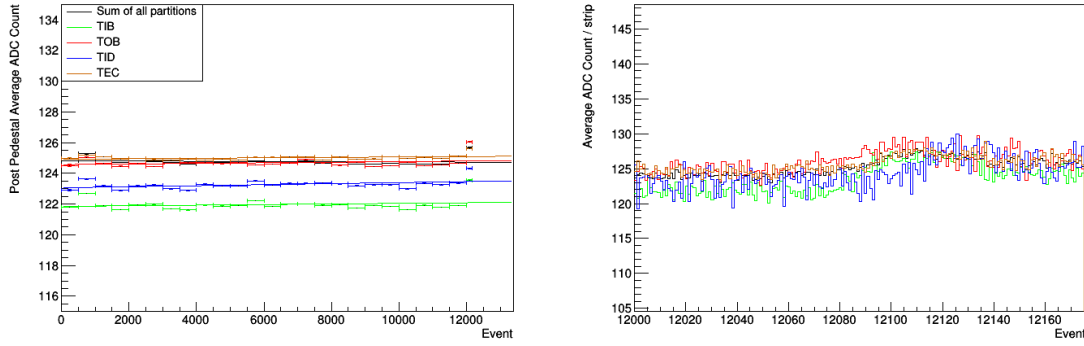


Figure B.3.2: The distribution average ADC counts for the post pedestal subtracted ADC count of each APV25 chip in bins of 500 events (left) and for events 12000–1276 (left) for spy run 321779. A slight increase in the ADC count is observed at the end of the run, which we believe is due to the tracker changing configurations to perform cosmic data taking.

APV25 chip is also recorded. To ensure that this information is not dependent on the instantaneous luminosity of the event, any charge clusters which pass the full offline zero-suppression algorithms are omitted from the post pedestal subtracted ADC count distribution when analyzing the common mode offset. Due to the large number of APV25 chips in the CMS outer tracker, trends in the common mode are analyzed by extracting the average and standard deviation of the slope, average ADC count, and ADC count RMS for every event. The trend in how these parameters changed over the course of a run is then investigated to draw conclusions about the progression of the common mode.

The trends in the slope of the common mode offset for spy run 321779 can be found in Figure B.3.1. Run 321779 consisted of 12179 spy events taken over the course of approximately 9 hours. To better observe overall trends in the data, events were combined into bins of 500 and fit with a straight line. For all partitions of the outer tracker we found that there was a small residual strip dependence on the common mode ADC count. Also, regardless of the magnitude or sign of the initial relationship, it slowly increased to gain a more positive slope (higher strip numbers had higher ADC counts) at an consistent rate in all partitions. Additionally, trends in the RMS show that the distributions of the common mode slope becomes slightly more peaked as the run time continues. In both cases, the removal of strips with zero-suppressed charge clusters leads us to believe that this behavior is not a result of the hit occupancy decreasing as a result of the beam luminosity decreasing at later run times. Despite these trends we find that they would not have an overall impact on the zero-suppression algorithm given the flat common mode hypothesis. Given that a single APV25 chip communicates with 128 strips, an average common mode slope that is less than 0.01 with an RMS of

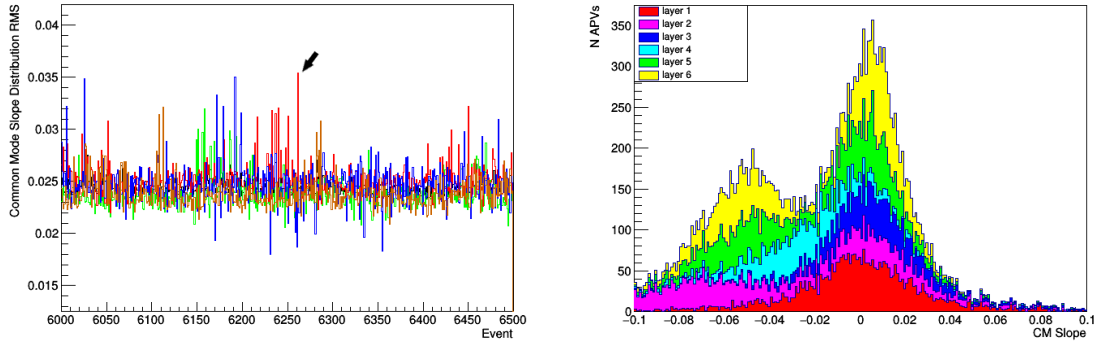


Figure B.3.3: The RMS of the common mode slope distribution per event for events 6000–6500 of run 321779 (left) and the common mode slope distribution for the different layers of the TOB in event 6262 (right) which was found to have a large RMS value.

0.026–0.028 leads to 95% of common modes having a strip dependence that causes the ADC count to vary by less than 2 over the course of 128 strips. When considering the changes in the slope of the common mode, we observe an average difference of approximately 0.001 over the course of the run which is still too small to produce a strip dependence in the ADC count larger than 1 ADC count.

Figure B.3.2 shows the trends in the average ADC counts of the pedestal subtracted digitized strip readout distribution for each of the APV25 chips in spy run 321779. The left plot shows the average ADC count distribution in bins of 500 events. After subtraction of strip with zero-suppressed charge clusters we find that it is relatively unchanged over course of the data taking period. At the very end of the spy run period, we observe a slight increase in the overall ADC counts. Due to the fact that spy runs can take place over multiple global runs, we found that run 321779 continued past the end of a fill and included data from a set of cosmic runs in which tracks produced from cosmic ray muons passing through the tracker are used for module alignment. We believe that this increase in ADC count is due to this change in run configurations used for cosmic data taking. In particular, it's possible that different gains used to amplify the strip currents leads to the increased pedestal subtracted ADC counts.

One last interesting feature that was found in the common mode slope trends came from the investigation of the event-by-event distributions of the RMS in the common mode slope distributions. It was found that periodic spikes in the RMS would be present in a single partition of the tracker, but none of the others. Closer investigation of the events found that some of the layers of the effected partitions would have large offsets in the common mode slope distribution that were centered 2σ or more from the standard mean value. An example of this behavior can be seen in Figure B.3.3 in which the TOB (shown in red) shows a large jump in the RMS of the common mode slope distribution

for event 6262. Further investigation found that the 2nd, 5th, and 6th layers of the TOB showed a double peak structure in which approximately half of the APV25 chips have slopes centered about -0.05 ADC counts/strip. At the time of writing this thesis the full cause of this behavior is not fully understood. It is possibly a byproduct of the asynchronous arming of the FEDs, or the result of an arming signal being received by the FPGAs when they are in the middle of processing a run, and the FPGAs only sending a portion of the unprocessed APV25 readout to the back-end FPGAs. However, further studies in Run 3 will need to be performed to better understand this behavior.

Appendix C

Top Tagging Using End-to-End Images

As mentioned in Section 5.3 jet identification has been a long standing area of study in accelerator physics. Traditional tagging algorithms are based on the properties of reconstructed physics objects, with algorithms in recent years using the features of PF candidates as inputs to machine learning algorithms. However, there is an invariable loss of information associated with compressing detector-level information into PF objects, and some physics objects may be reconstructed as fakes [55]. As such, it can be useful to look at end-to-end identification techniques that apply machine learning algorithms directly to low-level data representations of the detector.

This appendix chapter describes a study in which an end-to-end approach is applied to top-tagging with the CMS detector via a neural network which minimizes the number of reconstructed features used as inputs. This is done through the construction of jet images in which calorimeter and tracking information are mapped to an image based representation using the ECAL granularity and coordinate system. Computer vision techniques are then applied to the jet images and used to discriminate between $p_T > 400$ GeV jets originating from the hadronic decays of top quarks and hadronization of light quarks or gluons. The methods of this study are a natural extensions of two separate studies [132, 133]. In [132] images were first produced from ECAL energy deposits to identify events in which a $H \rightarrow \gamma\gamma$ decay took place. In [133] the study was extended to include HCAL energy deposits and the location of reconstructed tracks weighted by their p_T to discriminate between jets originating from the hadronization of uds quarks from jets produced from the hadronization of gluons.

Due to the traditional use of the secondary vertex for identifying b -jet decays, this iteration of the study introduces vertex information in the form of two new sets of image information. The first is by adding two new layers consisting of tracks weighted by their respective longitudinal and transverse impact parameters. The second is by introducing the locations where charged particles passed through the barrel region of the pixel detector as input features. As described in Section 3.2.1.1 the inner tracker is responsible for the CMS detectors 3D vertexing capabilities, motivating the use of the pixel detector to produce input features for the network.

C.1 The Identification of Top Quarks

Top quarks feature a very consistent decay pattern. Due to the near unity value of the $|V_{tb}|$ term of the CKM matrix [4] we expect that all top quarks will decay to produce a bottom quark and a W -boson. The bottom quark will then hadronize and decay to produce a b -jet. The W will decay leptonically to an electron, μ , or τ and the corresponding neutrino 33.5% of the time or hadronically to a quark and antiquark 66.5% of the time. In this case of the leptonically decaying W , the presence of a b -jet, high p_T lepton, and nearby MET can be used to identify the top quark. For fully hadronic top quark decays the primary feature is the production of three separate jet clusters. For lower p_T top quarks, this can result in three distinct jets. Common prescriptions within the CMS collaboration is to attempt to cluster all jets associated with a top quark into a single large radius jet (standard procedure is to use AK8 jets) and use substructure variables like the N -subjettiness [134] to parameterize information about the the existence of subjets of the unique jet clusters corresponding to the prompt decay products of the top. However, at a high Lorentz-boost the jet clusters of the hadronic decay products can overlap producing a unique yet challenging topology for identifying top quark decays where we can no longer resolve the three subjets. For these high p_T hadronic top decays, the most common background to discriminate against becomes high p_T $udsq$ -jets which can be produced in high enough quantities at the LHC that they can produce a statistically large sample of jets which fake the top decay signatures. This has lead the identification of high p_T top quarks to become a new and exciting way to benchmark the robustness of machine learning based jet tagging algorithms.

Sample	Description	Number of events
TTJets_HadronicMGDecays	$t\bar{t}$, p_T (top) > 400 GeV, all-jet top decays	2969109
QCD_Pt-300to600_TuneZ2star_Flat	QCD, flat $300 < \hat{p}_T < 600$ GeV	1498800
QCD_Pt-400to600_TuneZ2star_Flat	QCD, flat $400 < \hat{p}_T < 600$ GeV	1989000
QCD_Pt-600to3000_TuneZ2star_Flat	QCD, flat $600 < \hat{p}_T < 3000$ GeV	2975400

Table C.1: List of the CMS Open Data simulated samples used and total number of events considered for each of the datasets.

C.2 Top and QCD Jet Datasets

Networks used to perform jet discrimination are often trained using supervised learning in which labelled data is used to ensure that the network is properly classifying the inputs. As such, jets produced from MC simulated events are used to train and test the network due to the ability to use generator level information to label the jets. Events are then split based on two separate labels: 1 to signify a jet originating from the hadronic decay of a simulated top quark, and 0 to signify a jet originating from the decay of a *udsg*-jet. In order to produce a network in which other physicists outside of the CMS collaboration can benchmark their network, we elected to use samples from the CMS opendata portal [135]. This is a CMS project in which MC simulations, physics data, and the software needed to analyze them are periodically released open-source and free for public use.

C.2.1 CMS Opendata Samples

Top jets used to train and test the network were obtained from a simulated SM $t\bar{t}$ sample [136] in which the tops were generated to have a $p_T > 400$ GeV. The *udsg*-jets were then obtained from three separate QCD multijet samples in which jets were produced to have three different ranges of hard-scatter transverse momentum ($\hat{p}_T$): $300 < \hat{p}_T < 600$ GeV [137], $400 < \hat{p}_T < 600$ GeV [138], and $600 < \hat{p}_T < 3000$ GeV [139]. A description of the different CMS opendata samples used obtain the jets for this study as well as the number of simulated events can be found in Table C.1. The samples were generated using the CMS Run 1 data taking conditions during the 2012 data taking year. As such, the collisions were simulated at 8 TeV using MADGRAPH5_AMC@NLO v2.6.6 [140] and used PYTHIA6 [141] with the Z2Star tune for parton showering. An average of 10 additional PU collisions were added to the simulated hard-scattered event. The GEANT4 [63] full detector simulation was used to simulate the response of the CMS detector using CMSSW release 5_3_32 [142] to properly encapsulate the response of the phase 0 CMS detector during the 2012 data taking.

C.2.2 Jet Selection

Jets for use in the study are obtained from their respective MC samples. In each of the events, we look for jets clustered using the anti- k_T algorithm with a radius of $R = 0.8$ and have a $p_T > 400$ GeV and $|\eta| < 1.37$ after application of the L1, L2, and L3 corrections. The p_T requirement is placed to ensure that our top jets are highly boosted such that their subjets and constituents will be suitably merged. The η requirement is to ensure that the jet images (construction defined in following section) will be entirely contained within the $|\eta| < 2.4$ acceptance of the CMS tracker. For top jets, an additional check using generator level information is performed to ensure that the W daughter of the top quark decays hadronically. Additionally, the 4-momentum of the top quark, bottom quark daughter, and the two daughters of the W are checked to ensure that they point within the AK8 cone of the jet. Once one or more jets have been found in an event, the event is saved for image generation and the location of the AK8 jet is saved to seed the jet images in later steps.

Early test showed that our network was sensitive to differences in the p_T distribution between the $t\bar{t}$ and QCD jet samples. In particular, it would over attribute jets existing in the tail of the p_T distribution to belong to the sample with the harder p_T spectrum. To address this issue QCD jets were initially produced such that there were more QCD than top jets across their full p_T range. QCD jets were then pseudorandomly dropped from our QCD jet dataset until we were left with nearly identical p_T distributions in both of our jet samples. The total number of jets used for training and testing the network architecture after the QCD jet re-sampling procedure can be found in Table C.2.

Category	Top quark jets	QCD jets	Total Jets
Train	1280830	1279170	2560000
Validation	319819	320181	640000

Table C.2: Number of jets used for training and validating the top quark and non-top quark jet categories after the jet p_T re-sampling of the QCD jets. When training the network on subsets of layers, only 198k jets were used for validation.

C.3 The CMS Detector as an Image

Jet images make use of low-level detector information where CMS subdetectors are projected onto one or more layers of an image with each pixel corresponding with the location of an ECAL cell. As such, the images are in dimensions of $i\phi \times i\eta$ where each unit of $i\eta$ represents a distance of 0.1 in η and each unit of $i\phi$ represents a distance of 1° in ϕ . Jet images are produced by first forming an image of the entire barrel+endcap regions of the detector, where the endcap and barrel regions

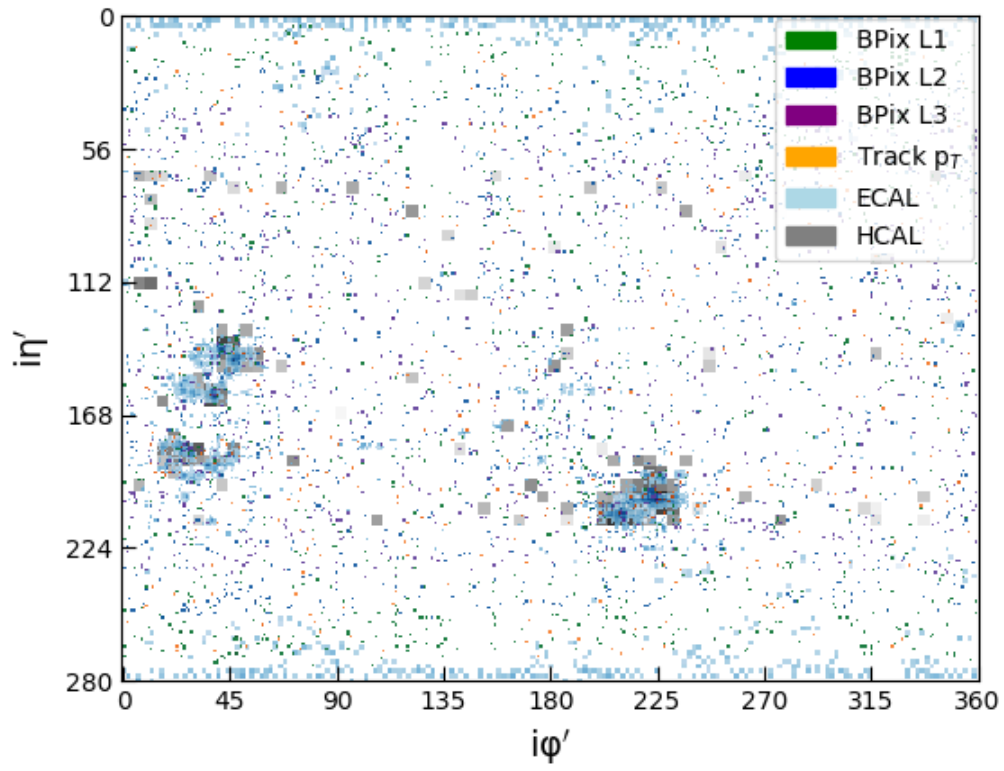


Figure C.3.1: An image representation of a CMS detector image for a $t\bar{t}$ event. All image layers described in this section are shown, with the exception of the track $d0$ and dz layers which are covered by the track p_T layer.

are stitched together in $i\eta$ at their boundary to form a detector image. These images are produced using the following subdetector information, where each subdetector or feature is represented by a separate image channel: ECAL energy deposits, HCAL energy deposits, tracks weighted by p_T , $d0$, and dz , and the reconstructed hits (RecHits) of tracks passing through the barrel of the pixel detector. A description of each of the image layers and how they are formed are as follows:

ECAL Energy Deposits – The energy deposited in each of the ECAL crystals is saved as is into a single layer of the image.

HCAL Energy Deposits – The energy deposited in each of the HCAL crystals is first saved to a temporary image layer. Due to the decreased granularity of the HCAL with respect to the ECAL, the HCAL image is upsampled by a factor of 5, with the energy deposited into an HCAL cell equally distributed between the 25 image pixels matching the corresponding ECAL supercluster.

Track Weighted Quantities – The reconstructed tracks through the general reconstruction algorithm are obtained, and their locations in η and ϕ are propagated to the surface of the ECAL. The track is then binned into the image pixel corresponding to the $i\eta$ and $i\phi$ coverage of the corresponding ECAL cell. A total of three separate images layers are produced by individually weighting each of the track pixels by the p_T , transverse impact parameter ($d0$) significance, and longitudinal impact parameter (dz) significance of the track. The IP significance is defined as the relative $d0$ or dz value divided by their corresponding uncertainty. Additionally, tracks with superfluously large impact parameters, found to be associated with fakes or photon conversion in the tracker, were found to vastly degrade training performance. As such, all $d0$ (dz) weighted tracks with $|d0| > 10$ ($|dz| > 20$) had their $d0$ (dz) values set to 0. Because each layer are produced from the same track collection, the spatial distribution of the tracks in the three layers are all the same, however the weighted values will differ between each of the layers.

BPix Detector RecHits – RecHits corresponding to the locations in the pixel detector in which charged particle tracks passed through the sensors are used to construct an additional set of jet image layers. RecHits are produced from the charge clusters of nearby pixels inside of the sensors which pass an adjustable charge threshold. A straight line then connects the center of the charge cluster to the center of the beam, and the location of the line on the front of the sensor adjusted for Lorentz-drift is taken as the RecHit location. As with the tracks, the η and ϕ locations are projected to the surface of the ECAL and image pixel associated with the

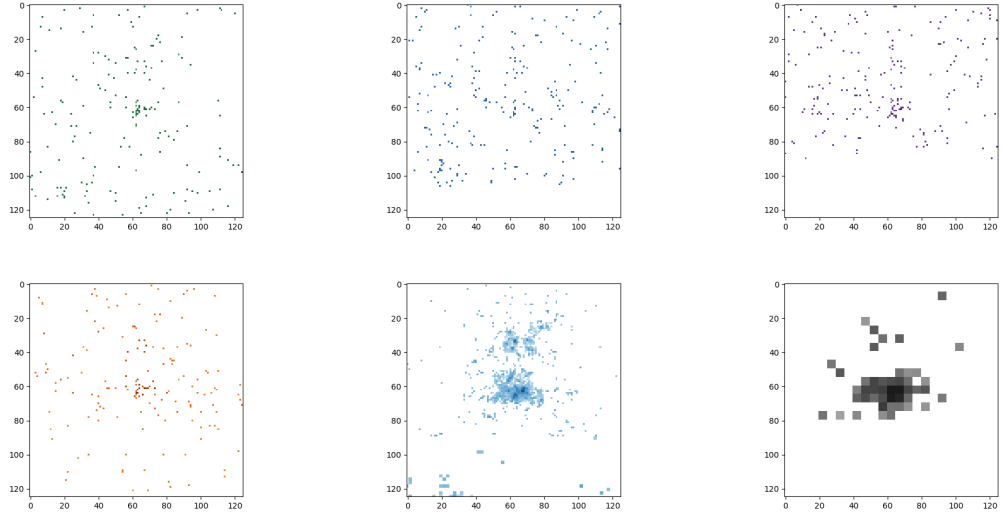


Figure C.3.2: The set of BPix layer 1 (top left), BPix layer 2 (top middle) BPix layer 3 (top right), track p_T (bottom left), ECAL (bottom middle), and HCAL (bottom right) for one of the jets shown in Figure C.3.1. The bottom portions of the BPix layer 2 and 3 image layers have no BPix due to the lower η acceptance of the outer layers of the BPix.

$i\eta$ and $i\phi$ value of the ECAL cell with the closest center to the RecHit is filled. In total, three separate RecHit layers were produced corresponding to the RecHits in each of the layers of the phase 0 BPix. The value in each image pixel is saved as an unsigned integer representing the number RecHits projected to the corresponding ECAL cell. The reason for omitting the FPix is due to the disparity in flight distance between the BPix and FPix as well as the overlapping of the outer rings of the FPix with the BPix producing regions in $i\eta$ where the RecHits would not be best suited for an image processing problem. As such, the more homogenous BPix was used, with RecHits from the FPix as well as the silicon strip detector left to be explored in future studies.

For the track and BPix layers, an additional modification was made in which the η and ϕ location of the track and RecHits are projected to the surface of the ECAL with respect to the location of the primary vertex rather than the center of beam pipe. The reason is due to the fact that straight lines originating from a location along the beam pipe that is not at $z = 0$ will pass through the layers of the CMS detector at different values in η with respect to the origin of the detector. This produced an apparent curvature in η for any RecHits and tracks that are not produced from the center of the axis, and as such the projection with respect to the PV is made to remove this effect. A full image of the CMS detector built from the image layers described above can be found in Figure C.3.1.

Once the full detector images have been produced, the η and ϕ location of the selected AK8 jets

are used to find the relevant top or *udsg*-jets in each detector image. Given the η and ϕ location of the jet axis, the nearby image pixels are searched to find the 5×5 HCAL pixel cluster with the highest energy deposit. A 125×125 window in $i\eta$ and $i\phi$ is then cropped around the center of the HCAL pixel cluster and used to produce a jet image that is fed into the network. The distribution of image pixels for the BPix, track p_T , and calorimeter layers for a cropped jet can be found in Figure C.3.2. Darker pixels correspond to higher p_T tracks and larger calorimeter energy deposits. Tracks weighted by the $d0$ and dz values are omitted due to the lack of spatial information added to the image.

Before feeding the images into the network, two last pre-processing steps are performed on the images. The first is to perform a light zero-suppression to remove track and calorimeter deposits likely to be attributed to noise. Each track with or calorimeter cell with p_T or deposited energy less than 1 MeV are removed from the image. For the tracks that are zero-suppressed based on their p_T , the corresponding $d0$ and dz layers have the corresponding tracks' parameters removed as well. Additionally, it has been found that computer vision based networks are most stable when the pixel intensity of all image channels are normalized to about unity. As such, the second pre-processing step is for the track p_T , $d0$, dz , ECAL, and HCAL layers to have their image pixel intensities scaled down to best facilitate training. Each layer was scaled down by the following amount: Track p_T was scaled down by a factor of 100, track $d0$ was scaled down by a factor of 5, track dz was scaled down by a factor of 8, ECAL energy deposits were scaled down by a factor of 5, and HCAL energy deposits were scaled down by a factor of 25.

C.3.1 Sub-ECAL Granularity Images

Traditionally, the RecHit locations of the inner tracking are used to perform the 3D vertexing required to resolve the secondary vertices of tracks. However, the granularity of the ECAL is too coarse to properly compute information about the impact parameter of a track. Back of the envelope calculations show that the resolution of the inner layer of the BPix is 8-9 times finer η and ϕ than that of the ECAL, meaning that the granularity of our jet images will need to be much finer if we expect the RecHit layers to provide our network with relevant information regarding track displacement. As such, early studies were performed to construct additional images using sub-ECAL granularity. For these images, an adjustable parameter was used to subdivide each ECAL cell into $N \times N$ equal sized slices in η and ϕ . When producing the track and RecHit based images, the tracks and RecHits were placed into the image pixel corresponding with the ECAL sub-cell with the shortest separation

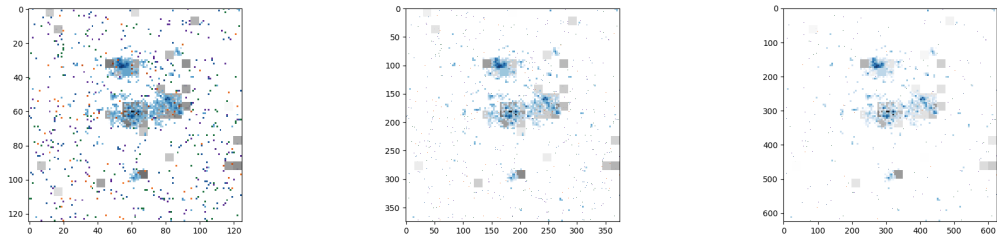


Figure C.3.3: Comparison of jet images for a simulated top jet at the nominal ECAL resolution (left) and after subdividing the ECAL cells by a factor of 3 (middle) and 5 (right).

from the projected track/RecHit in η and ϕ .

A comparison of sets of images produced at $3\times$ and $5\times$ the granularity of the ECAL with respect to a standard jet image can be found in Figure C.3.3. As with the upsampling of the HCAL for the nominal images, the ECAL and HCAL images had their original pixel intensity equally spread over the new 3×3 or 5×5 pixel cluster after upsampling. Despite producing images, we were unable to investigate too far into the use of these high granularity images when training. The reasons were two fold: When training on the $3\times$ granularity images we found a slight decrease in performance due to the images becoming too sparse for nearby track related features to be fully contained within the initial convolution filters. Additionally, the large file sizes made training and the transfer/storage of files cumbersome. This was especially prominent for the $5\times$ granularity images which had a total uncompressed image size of 12 MB and total uncompressed sample size of approximately 36 TB. Ultimately, performing a full hyper parameter scan to see if we could improve the network performance on high granularity images would have required a large amount of power consumption for a potentially limited improvement in performance. Furthermore, the use of high granularity images would be impractical for any form of a deployed end-to-end top tagger. As such, the exploration of high granularity images were dropped partway through the study.

C.3.2 Network Architecture

Due to the image-based representation of our data, this study follows the prescription used in the previous two end-to-end studies [132, 132] in which a ResNet-15 [143] architecture is used to construct the network. A ResNet is a type of convolutional neural network consisting of multiple residual blocks (ResBlocks). Each ResBlock consists of two consecutive convolutional layers separated by a ReLU activation function [144], with the input of the residual block then added back to the output. It has been found that much deeper convolutional networks can be constructed by stringing together

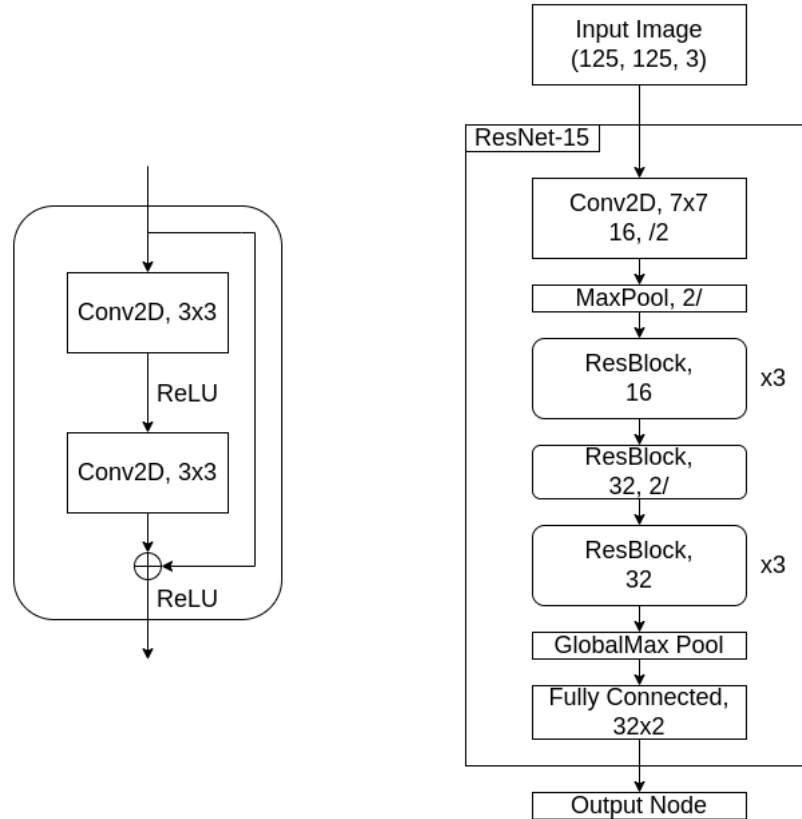


Figure C.3.4: A pictorial diagram of a ResBlock (left) and the ResNet-15 architecture used for the network (right). Filter size in ResBlocks are modifiable, but for this study all ResBlocks used a filter size of 3×3 . ReLU activation functions exist after the initial convolutional layer, and the fully connected layer is passed through a sigmoid activation function before being passed to the output node.

concentric ResBlocks rather than just using concentric convolutions. A diagram showing the overall structure of the network architecture, including a pictorial diagram of a ResBlock, can be seen in Figure C.3.4. A full explanation of the network infrastructure is as follows: The images pass through a 2D convolution consisting of 16 7×7 filters with a stride of 2. The outputs are passed through a ReLU activation function followed by a max pooling layer [145] which compresses each 2×2 cluster of pixels in each output channel, such that only the pixel with the largest value of the four is kept. The images are then passed through 3 ResBlocks each consisting of 16 3×3 filters with a stride of 1. A ReLU activation function is used in between each of the ResBlocks. The network then passes the images through a single ResBlock which makes use of 32 3×3 filters with a stride of 2 followed by another ReLU activation. There is then a final series of 3 more ResBlocks which use 32 3×3 filters with a stride of 1, and as before a ReLU activation function is used after each of the ResBlocks. After the final ResBlock, a global max pooling is applied to the image in which each of the 32 output channels of the final ResBlock layer are compressed into a single floating point number equal to the highest valued image pixel. This produces an array of length 32 which is passed through a fully connected layer to an output node with top and QCD that is then sent through a final sigmoid activation function [146].

During training, images are fed into the network with batch size of 64. A binary cross entropy loss function is used to determine the network's ability to properly discriminate between the two types of jets. The weights are optimized using the ADAM [147] algorithm with an initial learning rate of 5×10^{-4} that drops by half every 10 epochs. Due to the large sample size of input jets we found that 20 epochs was a suitable number of epochs in order for the network to converge, and as such trained the network for 20 epochs when training on subsets of jet image layers. However, when training over all layers we opted to train for a total of 40 epochs.

In order to compare our networks performance against a more traditional approach to jet identification, we additionally trained a fully-connected network consisting of 7 hidden layers and a total of 843,602 trainable parameters. Once again a batch size of 64, binary cross entropy loss function, and the ADAM optimizer were used. The output of the network is passed through a softmax function to obtain the top and QCD jet discriminator scores. The fully-connected network is fed an array of length 12 as an input containing the following features of the AK8 jets used for seeding the jet images: softdrop mass [148], the ratio of 2-subjettiness to 3-subjettiness [134], and the d_0 and dz impact parameter values of the 5 tracks with the largest transverse impact parameter. These features were chosen based on their ability to identify jets produced from the hadronization of bottom quarks [15], as well as their use in other top tagging algorithms [149].

Layer Combinations	ROC-AUC	Sig-eff at 1%
Track p_T (baseline)	0.955 ± 0.002	39.6%
Track p_T + ECAL + HCAL (nominal)	0.967 ± 0.002	49.5%
Track p_T + $d0$ + dz	0.972 ± 0.002	57.2%
Track p_T + $d0$ + dz + ECAL + HCAL	0.981 ± 0.002	64.4%
BPix	0.947 ± 0.002	38.3%
BPix + Track p_T	0.965 ± 0.002	47.6%
BPix + ECAL + HCAL	0.975 ± 0.002	56.9%
BPix + Track $p_T + d0 + dz$	0.977 ± 0.002	61.3%
BPix + Track $p_T + d0 + dz$ + ECAL + HCAL	0.9824 ± 0.0013	66.41%
Fully-Connected Network (groomed mass, substructure, track IPs)	0.9258 ± 0.0013	43.10%

Table C.3: Performance of the classifier trained up to 20 epochs. Results are shown for different combinations of tracking and calorimeter layers and evaluated on a sample of 198k jets.

C.4 Jet Identification Results

Table C.3 shows the area under the receiver operator curve (AUC) and signal efficiency at 1% background misidentification rate for the end-to-end classifier using different combinations of track and calorimeter layers. When training on a subset of layers the networks were trained for 20 epochs and evaluated on a 198k jet subset of the jets used for validation as described in Table C.2. When training on the full combination of eight image layers and for the fully connected-network, the networks were trained for 40 epochs and evaluated on the full 640k jet validation sample. A statistical uncertainty is obtained on the AUC score by inverting the square-root of the number of jets used to evaluate the network, and the signal efficiency is given for the AUC score central value.

Our previous end-to-end deep learning results showed that the Track p_T layer gave the best single layer performance for jet discrimination [133]. Therefore, we choose track p_T layer performance as a baseline for our models' performance. From Table C.3, we observe that the largest single-subdetector performance increase comes with the inclusion of the $d0$ and dz track information, leading to an AUC score improvement of 0.014–0.017. Comparing rows 2 and 3, the combination of track p_T , $d0$, and dz outperforms the nominal combination layers despite the fact that the $p_T + d0 + dz$ images are agnostic to neutral particles. This is in agreement with [133] where the tracks were observed as the most important feature for jet discrimination, and other jet tagging approaches that require the presence of a b-tagged subjet tagged using IP variables [77, 149].

Models trained only on BPix produced a weaker performance than those trained using the track p_T information. However, we observe multiple improvements in network performance after combining BPix with other layers. When training the network on BPix1–3, ECAL, and HCAL layers we find that it outperforms the nominal baseline in [133], improving the AUC score by 0.008 and the signal efficiency from 49.5% to 56.9%. Comparing rows 3 and 8 of Table C.3 shows that adding the BPix

to the track $p_T + d0 + dz$ images improves the AUC by 0.005 and the signal efficiency from 57.2% to 61.3%.

The bottom two rows of Table C.3 show the performance of network trained on all 8 channels and the fully-connected network. The network trained on all 8 channels attains an AUC score of 0.9824 ± 0.0013 and a signal efficiency of 66.41% at 1% misidentification. Comparing this to a fully-connected deep learning based top tagger, we find that all of the combination of layers explored outperform the fully-connected network's AUC score metric of 0.9258 ± 0.0013 .

C.4.1 Interpretation of Results

An in depth look at the performance of neural networks trained on different layer combinations provides insight into which features the network is learning. The observation that the strongest single subdetector performance comes from the reconstructed tracks weighted by their p_T and IP variables is in agreement with expectations based on the current understanding of high momentum top jets. We expect a large number of high p_T tracks, due to the jet containing three merged subjects, and a small subset of tracks having large IP values, attributed to a decaying B -meson. What is particularly interesting is that the network is able to successfully extract this IP information from the addition of the $d0$ and dz layers to the track p_T image layer. The track-only images are composed of a set of perfectly overlapping sparse layers, and our initial intuition was that the 2D convolutional layers would have difficulty extracting new information due to its lack of new spatial structure. We instead observe that this is not the case, and that these track-only images achieved an AUC of 0.972 ± 0.002 . This shows a large performance boost over our baseline comparison, and outperforms the nominal combination of layers that instead added new non-sparse spatial information to the baseline track p_T images.

Further insight into the utility of IP variables can be observed by comparing a networks AUC performance to it's signal efficiency. When comparing performance between networks, we sometimes observe that the network with a higher AUC score does not always possess a higher signal efficiency as well. As mentioned in Section C.4, we observe that the fully-connected network has a lower AUC score than any of the end-to-end networks, but that its signal efficiency is slightly larger than the networks trained only on track p_T or BPix1-3. We also observe this when comparing the network trained on the BPix1-3, ECAL, and HCAL layers to the network trained on track $p_T + d0 + dz$. The BPix1-3, ECAL, and HCAL network has a larger AUC score, but lower signal efficiency. In both of these cases, we find that the networks with direct access to IP information have larger signal

efficiencies at 1% background rejection, but still have a lower AUC value. Further investigation into the Receiver Operator Curves showed that the addition of IP variables primarily improves the networks background rejection at discriminator values with very high signal efficiency greater than 90%. However, this has a smaller impact on the networks background rejection at discriminator values with lower signal efficiency.

The second insight comes from the performance of the BPix RecHits. As mentioned in Section C.4, the BPix do not show a strong standalone single sub-detector performance. However, this is to be expected for multiple reasons. The pixel detector has an η and ϕ resolution of $10 \mu\text{m}$, giving the inner most layers a 1D spatial resolution that is almost eight times finer than the ECAL [50, 52]. Furthermore, we only considered the barrel region of the pixel detector, and do not include any RecHits from the forward region of the pixel detector. Any jets that border the η acceptance of this study will be missing RecHits from portions of the BPix layers. Finally, our network is agnostic to each layer’s distance from the beamline, giving the network incomplete information about the RecHit global positioning. For example, the RecHits will drift in ϕ as the charged particle bends in the CMS detector’s magnetic field. But unless more layers are added to the image, the network does not have enough information to know the order of each hit nor the direction in ϕ the particle is moving. But despite the shortcomings of our current RecHit implementation, we find remarkable results. With the exception of the final layer combination, where BPix RecHits are added to images composed of track $p_T + d0 + dz + \text{ECAL} + \text{HCAL}$ information, we note that adding the BPix RecHits gives a statistically significant increase in network performance. The most notable are cases where BPix RecHits are added on top of the tracking variables (1), and the case where BPix RecHits are used in lieu of the derived tracking information (2).

In the first case (1), we see that the network is able to use the BPix to exploit jet features which it could not parse from the derived track quantities alone. One possible feature is the track charge, where motion through ϕ can be combined with the final location of the track to determine its direction of curvature of the track. However, more abstract features may also exist in these images. In the case of (2), the network does not use any reconstructed variables for its inputs. We see that despite the lack of derived variables, the network outperforms the track $p_T + d0 + dz$ images, and only performs marginally worse than the final performance on the full images. The overall success of our network’s ability to learn from BPix RecHits paves the foundation for future studies of end-to-end taggers where no derived variables are used.

Processor	Manufacturer	Year Released	HBM Memory	Performance
Tesla P100	NVIDIA	2016	16 GB	9.3 Single-Precision TeraFLOPS
Tesla V100	NVIDIA	2017	32 GB / 16 GB	125 Mixed-Precision TeraFLOPS
TPUv3-8	Google	2018	128 GB	420 Mixed-Precision TeraFLOPS

Table C.4: Comparison of architecture specification for the NVIDIA Tesla P100 [18], NVIDIA Tesla V100 [19], and Google TPUv3-8 [20] architectures. The floating point operations per second speed (FLOPS) are quoted based on the data type and architecture setup used during training.

C.5 Computing Architecture Training Speed Comparisons

With respect to the previous end-to-end study [133] the number of training events increased by a factor of 3.2 and the number of image layers has increased by a factor of 2.7. For a single image at a 125×125 resolution with 8 channels we find that it has an uncompressed size of 488 kB. These factors have lead to significant computation requirements for training and testing the performance of the network. However, this also makes training out network an excellent metric for testing the computing capabilities of accelerated computing hardware architectures. In particular, a series of leading end graphical processing units (GPUs) and a tensor processing unit (TPU) manufactured by Google. The three different accelerated hardware architectures used in this study as well as how they were interfaced with are described below:

NVIDIA Tesla P100 – The NVIDIA Tesla P100 is a GPU that utilizes the Pascal architecture [18]. The Tesla P100 GPU was accessed on a shared cluster at the Fermilab National Accelerator Laboratory LHC Physics Center via a dedicated GPU worker node. The node accessed the GPU through a 12 GB/s PCIe connection using the CUDA Toolkit v9.1 drivers. During training, data was stored and read from an HGST 1W10002 hard disc drive [150] located on the GPU machine. Images were uncompressed, pre-processed, and sent to the GPU using a single Intel(R) Xeon(R) Silver 4110 8-core CPU [151].

NVIDIA Tesla V100 – The NVIDIA Tesla V100 uses the Volta architecture [19], incorporating eight tensor cores and an all-around higher performance than the Pascal architecture. Unlike the P100, the V100 is able to make use of mixed precision operations, which were utilized for this comparison, to drastically increase the number of floating point operations per second (FLOPS). During training, images were read from a solid state drives with better random I/O operations per second than traditional disc drives. The computing node ran Cuda v11.0.2 drivers and used four Intel(R) Xeon(R) Gold 5118 12-core CPUs [152] to perform data loading and pre-processing tasks, which were parallelized using the NVIDIA DGX-1 architecture [153]. Of the 48 available CPU cores, 20 were used for data loading.

TPUv3-8 - The TPU is a type of AI-accelerated architecture designed by Google with the purpose of training and running inference on machine learning models [154], and can be accessed through the Google Cloud platform [155]. TPUs boast a high number of FLOPS, made possible by dedicated matrix multiplication units which make use of the bfloat16 data type [156]. Unlike GPU based architectures, the CPU cores used to fetch and pre-process batches all live on the TPU board creating a low latency I/O pipeline which does not have to compete for CPU resources. For this work, a TPUv3-8 [20] running TPU software v1.14 was run using a type n1-standard-1 Google Cloud virtual machine node. Data used to train the networks was stored in Google Cloud buckets located in the same region as the virtual machine and TPU nodes to decrease time associated with transferring data during the training. This gives the cloud storage buckets comparable performance to persistent HDs in the local VM storage area [157].

A comparison of the primary hardware specifications of the P100, V100, and TPUv3-8 can be found in Table C.4. Single-precision refers to the use of operations involving float32 data types. For mixed-precision operations, weights in the layers of the network are stored as float32 data types and the intermediate outputs of each layer are stored as float16 or bfloat16 data types.

Table C.5 provides a timing comparison for training machine learning models using the three different architectures. To better decouple the performance of the TPU and GPUs from bottlenecks in the overall I/O infrastructure of the computing nodes they were accessed from, the time related to I/O time associated with producing the batches was extracted from the total time required for the full training cycle (I/O + computation). The total training performed on the P100 were drastically longer, partially stemming from low I/O speeds caused by the inefficient random reads associated with disk HDs [158] as well the weaker CPU used for building batches and sending them to the GPU. Additionally, the nodes used to access the GPUs had less available RAM, leading to a batch size of 32 to be used when training. When utilizing the Tesla P100, we observe that improvements to the computing clusters I/O infrastructure could decrease the required time to train on a single epoch by up to 2.6 hours.

The Tesla V100 and TPUv3-8 give much stronger performance and were accessed utilizing I/O pipelines with comparable performance. However, in both cases we see that over half of the time associated with training a batch is spent on I/O. By subtracting the single batch I/O time from the single batch train time, we obtain an approximate computation time. From this, we see that the TPUv3-8 spent approximately 2.6 ms to perform forwards and backwards propagation calculations, which is a factor of four faster than the 11 ms required by the Tesla V100. In the end, this lead to the

Category	Tesla P100	Tesla V100	TPUv3-8
Training Batch Size	32 jets	64 jets	64 jets
Batches Per Epoch	80k	40k	40k
I/O Time (batch)	0.119 s	0.0180 s	0.0176 s
Train Time (batch)	0.481 s	0.0290 s	0.0202 s
Train Time (epoch)	321 min	19 min	14 min

Table C.5: Comparison of I/O and training time for different computing architectures. A larger batch size was used when training on the Tesla V100 and TPUv3-8. Training times were found to vary by approximately 10% between epochs.

TPUv3-8 having the overall best performance for performing forward and backwards propogation. However, in all of our cases we find that large improvements in the training speeds can be gained through better tuning of the I/O infrastructure used to train the network.

Bibliography

- [1] M. Cush, “Standard Model of Elementary Particles.” September, 2019.
- [2] D. d’Enterria, P. Skands, S. Alekhin, A. Banfi, S. Bethke, J. Bluemlein, K. Chetyrkin, D. d’Enterria, G. Dissertori, X. Tormo, A. Hoang, M. Klasen, T. Klijnsma, S. Kluth, J. Kneur, B. Kniesl, D. Kolodrubetz, J. Kühn, P. Mackenzie, and I. Stewart, “High-precision α_s measurements from LHC to FCC-ee,”.
- [3] **NNPDF** Collaboration, R. D. Ball *et al.*, “Parton distributions from high-precision collider data,” *Eur. Phys. J. C* **77** (2017) 663. [arXiv:1706.00428 \[hep-ph\]](#).
- [4] **Particle Data Group** Collaboration, M. e. a. Tanabashi, “Review of Particle Physics,” *Phys. Rev. D* **98** (Aug, 2018) 030001. [\[link\]](#).
- [5] E. A. Mobs, “The CERN Accelerator Complex.”.
- [6] J. Bauer and T. Muller, “Prospects for the Observation of Electroweak Top Quark Production with the CMS Experiment,”.
- [7] O. S. Brüning, P. Collier, P. Lebrun, S. Myers, R. Ostojic, J. Poole, and P. Proudlock, *LHC Design Report*. CERN Yellow Reports: Monographs. CERN, Geneva, 2004. [\[link\]](#).
- [8] A. Team, “Diagram of an LHC dipole magnet. Schéma d’un aimant dipôle du LHC.” Jun, 1999.
- [9] **CMS Collaboration** Collaboration, T. Sakuma, “Cutaway diagrams of CMS detector,”. [\[link\]](#).
- [10] M. Ressegotti and O. behalf of the CMS Collaboration, “Overview of the CMS Detector Performance at LHC Run 2,” *Universe* **5** (2019) no. 1, . [\[link\]](#).

- [11] **CMS Collaboration** Collaboration, *The CMS magnet project: Technical Design Report*. Technical design report. CMS. CERN, Geneva, 1997. [\[link\]](#).
- [12] V. Khachatryan, A. Sirunyan, A. Tumasyan, W. Adam, E. Asilar, T. Bergauer, J. Brandstetter, E. Brondolin, M. Dragicevic, J. Erö, M. Flechl, M. Friedl, R. Frühwirth, and et al., “The CMS trigger system,” *Journal of Instrumentation* **12** (Jan, 2017) P01020–P01020. [\[link\]](#).
- [13] M. Cacciari, G. P. Salam, and G. Soyez, “The anti-ktjet clustering algorithm,” *Journal of High Energy Physics* **2008** (apr, 2008) 063–063. [\[link\]](#).
- [14] C. G. T. L. experiments Committee, “The Phase-2 Upgrade of the CMS Tracker,” Report CERN-LHCC-2017-009 ; CMS-TDR-014, Collaboration, CMS, 2017.
- [15] A. Sirunyan, A. Tumasyan, W. Adam, F. Ambroggi, E. Asilar, T. Bergauer, J. Brandstetter, E. Brondolin, M. Dragicevic, J. Erö, A. E. D. Valle, M. Flechl, M. Friedl, and et al., “Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV,” *Journal of Instrumentation* **13** (May, 2018) P05011–P05011. [\[link\]](#).
- [16] **CMS Collaboration** Collaboration, “Calibration of charm jet identification algorithms using proton-proton collision events at $\sqrt{s} = 13$ TeV,” tech. rep., CERN, Geneva, 2021. [\[link\]](#).
- [17] **LHC Higgs Cross Section Working Group** Collaboration, LHC Higgs Cross Section Working Group, C. Anastasiou, D. de Florian, C. Grojean, F. Maltoni, C. Mariotti, A. Nikitenko, M. Schumacher, and R. Tanaka (Eds.), “Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector,” [arXiv:1610.07922](#) [[hep-ph](#)].
- [18] *NVIDIA Tesla P100: The Most Advanced Data Center Accelerator*. [\[link\]](#). Accessed: 28 August 2020.
- [19] *NVIDIA V100 | NVIDIA*. [\[link\]](#). Accessed: 16 September 2020.
- [20] *Cloud Tensor Processing Units (TPUs)*. [\[link\]](#). Accessed: 30 August 2020.
- [21] C. Collaboration, “Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC,” *Physics Letters B* **716** (2012) no. 1, 30–61. [\[link\]](#).

- [22] A. Collaboration, “Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC,” *Physics Letters B* **716** (2012) no. 1, 1–29. [\[link\]](#).
- [23] P. W. Higgs, “Broken Symmetries and the Masses of Gauge Bosons,” *Phys. Rev. Lett.* **13** (Oct, 1964) 508–509. [\[link\]](#).
- [24] H. Baer, T. Barklow, K. Fujii, Y. Gao, A. Hoang, S. Kanemura, J. List, H. E. Logan, A. Nomerotski, M. Perelstein, M. E. Peskin, R. Pöschl, J. Reuter, S. Riemann, A. Savoy-Navarro, G. Servant, T. M. P. Tait, and J. Yu, “The International Linear Collider Technical Design Report - Volume 2: Physics,” tech. rep., 2013. [\[link\]](#).
- [25] CERN, “CERN Yellow Reports: Monographs, Vol 2 (2018): The Compact Linear e^+e^- Collider (CLIC) : 2018 Summary Report,” tech. rep., 2018. [\[link\]](#).
- [26] A. Abada, M. Abbrescia, S. S. AbdusSalam, I. Abdyukhanov, J. A. Fernandez, A. Abramov, M. Aburaia, A. O. Acar, P. R. Adzic, P. Agrawal, J. A. Aguilar-Saavedra, J. J. Aguilera-Verdugo, M. Aiba, I. Aichinger, G. Aielli, A. Akay, A. Akhundov, H. Aksakal, J. L. Albacete, S. Albergo, A. Alekou, M. Aleksa, R. Aleksan, R. M. A. Fernandez, Y. Alexahin, R. G. Alía, S. Alioli, N. A. Tehrani, B. C. Allanach, P. P. Allport, M. Altinli, W. Altmannshofer, G. Ambrosio, D. Amorim, O. Amstutz, L. Anderlini, A. Andreazza, M. Andreini, A. Andriatis, C. Andris, A. Andronic, M. Angelucci, F. Antinori, S. A. Antipov, M. Antonelli, M. Antonello, P. Antonioli, S. Antusch, F. Anulli, L. Apolinário, G. Apollinari, A. Apollonio, D. Appelö, R. B. Appleby, A. Apyan, A. Apyan, A. Arbey, A. Arbuzov, G. Arduini, V. Ari, S. Arias, N. Armesto, R. Arnaldi, S. A. Arsenyev, M. Arzeo, S. Asai, E. Aslanides, R. W. Aßmann, D. Astapovych, M. Atanasov, S. Atieh, D. Attié, B. Auchmann, A. Audurier, S. Aull, S. Aumon, S. Aune, F. Avino, G. Avrillaud, G. Aydın, A. Azatov, G. Azuelos, P. Azzi, O. Azzolini, P. Azzurri, N. Bacchetta, E. Bacchiocchi, H. Bachacou, Y. W. Baek, V. Baglin, Y. Bai, S. Baird, M. J. Baker, M. J. Baldwin, A. H. Ball, A. Ballarino, S. Banerjee, D. P. Barber, D. Barducci, P. Barjhoux, D. Barna, G. G. Barnaföldi, M. J. Barnes, A. Barr, J. B. García, J. B. G. da Costa, W. Bartmann, V. Baryshevsky, E. Barzi, S. A. Bass, A. Bastianin, B. Baudouy, F. Bauer, M. Bauer, T. Baumgartner, I. Bautista-Guzmán, C. Bayındır, F. Beaudette, F. Bedeschi, M. Béguin, I. Bellafont, L. Bellagamba, N. Bellegarde, E. Belli, E. Bellingeri, F. Bellini, G. Bellomo, S. Belomestnykh, G. Bencivenni, M. Benedikt, G. Bernardi, J. Bernardi, C. Bernet, J. M.

Bernhardt, C. Bernini, C. Berriaud, A. Bertarelli, S. Bertolucci, M. I. Besana, M. Besançon, O. Beznosov, P. Bhat, C. Bhat, M. E. Biagini, J.-L. Biarrotte, A. B. Chevalier, E. R. Bielert, M. Biglietti, G. M. Bilei, B. Bilki, C. Biscari, F. Bishara, O. R. Blanco-García, F. R. Blánquez, F. Blekman, A. Blondel, J. Blümlein, T. Boccali, R. Boels, S. A. Bogacz, A. Bogomyagkov, O. Boine-Frankenheim, M. J. Boland, S. Bologna, O. Bolukbasi, M. Bomben, S. Bondarenko, M. Bonvini, E. Boos, B. Bordini, F. Bordry, G. Borghello, L. Borgonovi, S. Borowka, D. Bortoletto, D. Boscherini, M. Boscolo, S. Boselli, R. R. Bosley, F. Bossu, C. Botta, L. Bottura, R. Boughezal, D. Boutin, G. Bovone, I. B. Jelisavić, A. Bozbey, C. Bozzi, D. Bozzini, V. Braccini, S. Braibant-Giacomelli, J. Bramante, P. Braun-Munzinger, J. A. Briffa, D. Britzger, S. J. Brodsky, J. J. Brooke, R. Bruce, P. B. De Renstrom, E. Bruna, O. Brüning, O. Brunner, K. Brunner, P. Bruzzone, X. Buffat, E. Bulyak, F. Burkart, H. Burkhardt, J.-P. Burnet, F. Butin, D. Buttazzo, A. Butterworth, M. Caccia, Y. Cai, B. Caiffi, V. Cairo, O. Cakir, R. Calaga, S. Calatroni, G. Calderini, G. Calderola, A. Caliskan, D. Calvet, M. Calviani, J. M. Camalich, P. Camarri, M. Campanelli, T. Camporesi, A. C. Canbay, A. Canepa, E. Cantergiani, D. Cantore-Cavalli, M. Capeans, R. Cardarelli, U. Cardella, A. Cardini, C. M. C. Calame, F. Carra, S. Carra, A. Carvalho, S. Casalbuoni, J. Casas, M. Cascella, P. Castelnovo, G. Castorina, G. Catalano, V. Cavasinni, E. Cazzato, E. Cennini, A. Cerri, F. Cerutti, J. Cervantes, I. Chaikovska, J. Chakraborty, M. Chala, M. Chamizo-Llatas, H. Chanal, D. Chanal, S. Chance, A. Chancé, P. Charitos, J. Charles, T. K. Charles, S. Chattopadhyay, R. Chehab, S. V. Chekanov, N. Chen, A. Chernoded, V. Chetvertkova, L. Chevalier, G. Chiarelli, G. Chiarello, M. Chiesa, P. Chiggiato, J. T. Childers, A. Chmieleńska, A. Cholakian, P. Chomaz, M. Chorowski, W. Chou, M. Chrzaszcz, E. Chyhyrnyets, G. Cibinetto, A. K. Ciftci, R. Ciftci, R. Cimino, M. Ciuchini, P. J. Clark, Y. Coadou, M. Cobal, A. Coccaro, J. Cogan, E. Cogneras, F. Collamati, C. Coldelram, P. Collier, J. Collot, R. Contino, F. Conventi, and C. T. A. Cook, “FCC Physics Opportunities,” Tech. Rep. 6, Jun, 2019. [\[link\]](#).

- [27] C. Collaboration, “Measurements of properties of the Higgs boson decaying to a W boson pair in pp collisions at $\sqrt{s} = 13$ TeV,” *Physics Letters B* **791** (2019) 96–129. [\[link\]](#).
- [28] A. Collaboration, “Observation and measurement of Higgs boson decays to WW^* with the ATLAS detector,” *Phys. Rev. D* **92** (Jul, 2015) 012006. [\[link\]](#).
- [29] C. Collaboration, “Observation of Z decays to four leptons with the CMS detector at the LHC,” *Journal of High Energy Physics* **2012** (dec, 2012) . [\[link\]](#).

- [30] A. Collaboration, “Measurement of the Higgs boson mass in the $H \rightarrow ZZ^* \rightarrow 4\ell$ and $H \rightarrow \gamma\gamma$ channels with $\sqrt{s} = 13$ TeV pp collisions using the ATLAS detector,” *Physics Letters B* **784** (2018) 345–366. [[link](#)].
- [31] C. Collaboration, “Observation of $t\bar{t}H$ Production,” *Phys. Rev. Lett.* **120** (Jun, 2018) 231801. [[link](#)].
- [32] ATLAS, “Observation of Higgs boson production in association with a top quark pair at the LHC with the ATLAS detector,” *Physics Letters B* **784** (2018) 173–191. [[link](#)].
- [33] C. Collaboration, “Observation of Higgs Boson Decay to Bottom Quarks,” *Phys. Rev. Lett.* **121** (Sep, 2018) 121801. [[link](#)].
- [34] A. Collaboration, “Observation of $H \rightarrow b\bar{b}$ decays and VH production with the ATLAS detector,” *Physics Letters B* **786** (2018) 59–86. [[link](#)].
- [35] C. Collaboration, “Observation of the Higgs boson decay to a pair of τ leptons with the CMS detector,” *Physics Letters B* **779** (2018) 283–316. [[link](#)].
- [36] A. Collaboration, “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. D* **99** (Apr, 2019) 072001. [[link](#)].
- [37] CMS, “Evidence for Higgs boson decay to a pair of muons,” *Journal of High Energy Physics* **2021** (Jan, 2021) . [[link](#)].
- [38] A. collaboration, “A search for the dimuon decay of the Standard Model Higgs boson with the ATLAS detector,” *Physics Letters B* **812** (2021) 135980. [[link](#)].
- [39] C. Collaboration, “A search for the standard model Higgs boson decaying to charm quarks,” *Journal of High Energy Physics* **2020** (Mar, 2020) . [[link](#)].
- [40] E. Bols, J. Kieseler, M. Verzett, M. Stoye, and A. Stakia, “Jet Flavour Classification Using DeepJet,” *JINST* **15** (Aug, 2020) P12012. 14 p. [arXiv:2008.10519](#). 14 pages, 9 figures, accepted for publication in JINST.
- [41] M. Kardar, *Statistical Physics Of Particles*. Cambridge University Press, June, 2007.
- [42] **Super-Kamiokande Collaboration** Collaboration, Y. Fukuda, T. Hayakawa, E. Ichihara, K. Inoue, K. Ishihara, H. Ishino, Y. Itow, T. Kajita, J. Kameda, S. Kasuga, K. Kobayashi,

- Y. Kobayashi, Y. Koshio, M. Miura, M. Nakahata, S. Nakayama, A. Okada, K. Okumura, N. Sakurai, M. Shiozawa, Y. Suzuki, Y. Takeuchi, Y. Totsuka, S. Yamada, M. Earl, A. Habig, E. Kearns, M. D. Messier, K. Scholberg, J. L. Stone, L. R. Sulak, C. W. Walter, M. Goldhaber, T. Barszczak, D. Casper, W. Gajewski, P. G. Halverson, J. Hsu, W. R. Kropp, L. R. Price, F. Reines, M. Smy, H. W. Sobel, M. R. Vagins, K. S. Ganezer, W. E. Keig, R. W. Ellsworth, S. Tasaka, J. W. Flanagan, A. Kibayashi, J. G. Learned, S. Matsuno, V. J. Stenger, D. Takemori, T. Ishii, J. Kanzaki, T. Kobayashi, S. Mine, K. Nakamura, K. Nishikawa, Y. Oyama, A. Sakai, M. Sakuda, O. Sasaki, S. Echigo, M. Kohama, A. T. Suzuki, T. J. Haines, E. Blaufuss, B. K. Kim, R. Sanford, R. Svoboda, M. L. Chen, Z. Conner, J. A. Goodman, G. W. Sullivan, J. Hill, C. K. Jung, K. Martens, C. Mauger, C. McGrew, E. Sharkey, B. Viren, C. Yanagisawa, W. Doki, K. Miyano, H. Okazawa, C. Saji, M. Takahata, Y. Nagashima, M. Takita, T. Yamaguchi, M. Yoshida, S. B. Kim, M. Etoh, K. Fujita, A. Hasegawa, T. Hasegawa, S. Hatakeyama, T. Iwamoto, M. Koga, T. Maruyama, H. Ogawa, J. Shirai, A. Suzuki, F. Tsushima, M. Koshihara, M. Nemoto, K. Nishijima, T. Futagami, Y. Hayato, Y. Kanaya, K. Kaneyuki, Y. Watanabe, D. Kielczewska, R. A. Doyle, J. S. George, A. L. Stachyra, L. L. Wai, R. J. Wilkes, and K. K. Young, “Evidence for Oscillation of Atmospheric Neutrinos,” *Phys. Rev. Lett.* **81** (Aug, 1998) 1562–1567. [[link](#)].
- [43] R. P. Feynman, “Space-Time Approach to Non-Relativistic Quantum Mechanics,” *Rev. Mod. Phys.* **20** (Apr, 1948) 367–387. [[link](#)].
- [44] G. Altarelli and G. Parisi, “Asymptotic freedom in parton language,” *Nuclear Physics B* **126** (1977) no. 2, 298–318. [[link](#)].
- [45] **ATLAS Collaboration** Collaboration, *ATLAS detector and physics performance: Technical Design Report, 1*. Technical design report. ATLAS. CERN, Geneva, 1999. [[link](#)].
- [46] **CMS Collaboration** Collaboration, *CMS Physics: Technical Design Report Volume 1: Detector Performance and Software*. Technical design report. CMS. CERN, Geneva, 2006. [[link](#)]. There is an error on cover due to a technical problem for some items.
- [47] **ALICE Collaboration** Collaboration, “The ALICE experiment at the CERN LHC. A Large Ion Collider Experiment,” *JINST* **3** (2008) S08002. 259 p. [[link](#)]. Also published by CERN Geneva in 2010.
- [48] **LHCb Collaboration** Collaboration, “LHCb: letter of intent,” tech. rep., CERN, Geneva, 1995. [[link](#)].

- [49] A. M. Lombardi, “The radio frequency quadrupole (RFQ),”. [\[link\]](#).
- [50] **CMS Collaboration** Collaboration, Karimäki, V and Mannelli, M and Siegrist, P and Breuker, H and Caner, A and Castaldi, R and Freudenreich, K and Hall, G and Horisberger, R and Huhtinen, M and Cattai, A, *The CMS tracker system project: Technical Design Report*. Technical design report. CMS. CERN, Geneva, 1997. [\[link\]](#).
- [51] **CMS Collaboration** Collaboration, *The CMS tracker: addendum to the Technical Design Report*. Technical design report. CMS. CERN, Geneva, 2000. [\[link\]](#).
- [52] **CMS Collaboration** Collaboration, *The CMS electromagnetic calorimeter project: Technical Design Report*. Technical design report. CMS. CERN, Geneva, 1997. [\[link\]](#).
- [53] **CMS Collaboration** Collaboration, *The CMS hadron calorimeter project: Technical Design Report*. Technical design report. CMS. CERN, Geneva, 1997. [\[link\]](#).
- [54] **CMS Collaboration** Collaboration, J. G. Layter, *The CMS muon project: Technical Design Report*. Technical design report. CMS. CERN, Geneva, 1997. [\[link\]](#).
- [55] C. Collaboration, “Particle-flow reconstruction and global event description with the CMS detector,” *Journal of Instrumentation* **12** (oct, 2017) P10003–P10003. [\[link\]](#).
- [56] T. C. Collaboration, “Description and performance of track and primary-vertex reconstruction with the CMS tracker,” *Journal of Instrumentation* **9** (Oct, 2014) P10009–P10009. [\[link\]](#).
- [57] W. Waltenberger, “Adaptive Vertex Reconstruction,” tech. rep., CERN, Geneva, Jul, 2008. [\[link\]](#).
- [58] G. Piacquadio and C. Weiser, “A new inclusive secondary vertex algorithm for b-jet tagging in ATLAS,” *Journal of Physics: Conference Series* **119** (07, 2008) 032032. [\[doi\]](#).
- [59] A. Sirunyan, A. Tumasyan, W. Adam, T. Bergauer, M. Dragicevic, A. Escalante Del Valle, R. Frühwirth, M. Jeitler, N. Krammer, L. Lechner, D. Liko, I. Mikulec, F. Pitters, and et al., “Electron and photon reconstruction and identification with the CMS experiment at the CERN LHC,” *Journal of Instrumentation* **16** (May, 2021) P05014. [\[link\]](#).
- [60] A. Sirunyan, A. Tumasyan, W. Adam, F. Ambroggi, E. Asilar, T. Bergauer, J. Brandstetter, E. Brondolin, M. Dragicevic, J. ErŒ, A. E. D. Valle, M. Flechl, M. Friedl, and et al.,

- “Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at $\sqrt{s}=13$ TeV,” *Journal of Instrumentation* **13** (Jun, 2018) P06015–P06015. [link].
- [61] CMS Collaboration, “A Cambridge-Aachen (C-A) based Jet Algorithm for boosted top-jet tagging,”. [link].
- [62] M. Cacciari and G. P. Salam, “Dispelling the N^3 myth for the Kt jet-finder,” *Physics Letters B* **641** (Sep, 2006) 57–61. [link].
- [63] GEANT4 Collaboration, S. Agostinelli *et al.*, “GEANT4 — A simulation toolkit,” *Nucl. Instrum. Meth. A* **506** (2003) 250. [doi].
- [64] T. C. collaboration, “Determination of jet energy calibration and transverse momentum resolution in CMS,” *Journal of Instrumentation* **6** (nov, 2011) P11002–P11002. [link].
- [65] M. Printz, “P-stop isolation study of irradiated n-in-p type silicon strip sensors for harsh radiation environments,” *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **831** (2016) 38–43. [link].
 Proceedings of the 10th International Hiroshima Symposium on the Development and Application of Semiconductor Tracking Detectors.
- [66] O. G. Products, “SmartScope Flash 500.”.
- [67] FormFactor, “Cascarde TESLA200.”.
- [68] Tektronix, “Keithly 6517B Electrometer/High Resistance Meter.”.
- [69] Agilent, “4284A Precision LCR Meter.”.
- [70] Agilent, “34970A Data Acquisition / Data Logger Switch Unit.”.
- [71] Textronix, “Keithly 6485 Picoammeter.”.
- [72] Textronix, “Keithley 2400 Standard Series SMU.”.
- [73] Y. Feng, J. He, J.-P. Lansberg, H.-S. Shao, A. Usachov, and H.-F. Zhang, “Phenomenological NLO analysis of η production at the LHC in the collider and fixed-target modes,” *Nuclear Physics B* **945** (Aug, 2019) 114662. [link].

- [74] R. Aaij, C. A. Beteta, B. Adeva, M. Adinolfi, A. Affolder, Z. Ajaltouni, S. Akar, J. Albrecht, F. Alessio, M. Alexander, S. Ali, G. Alkhazov, P. Alvarez Cartelle, and et al., “Measurement of the $\eta_c(1S)$ production cross-section in proton-proton collisions via the decay $\eta_c(1S) \rightarrow p\bar{p}$,” *The European Physical Journal C* **75** (Jul, 2015) . [[link](#)].
- [75] **ATLAS Collaboration** Collaboration, C. Vittori, “Observation of Higgs in $b\bar{b}$ channel from ATLAS and CMS,” tech. rep., CERN, Geneva, Mar, 2019. [[link](#)].
- [76] C. Weiser, “A Combined Secondary Vertex Based B-Tagging Algorithm in CMS,” tech. rep., CERN, Geneva, Jan, 2006. [[link](#)].
- [77] A. Sirunyan, A. Tumasyan, W. Adam, F. Ambroggi, T. Bergauer, M. Dragicevic, J. Erö, A. E. D. Valle, M. Flechl, R. Frühwirth, M. Jeitler, N. Krammer, I. Krätschmer, and et al., “Identification of heavy, energetic, hadronically decaying particles using machine-learning techniques,” *Journal of Instrumentation* **15** (Jun, 2020) P06005–P06005. [[link](#)].
- [78] H. Qu and L. Gouskos, “Jet tagging via particle clouds,” *Physical Review D* **101** (Mar, 2020) . [[link](#)].
- [79] S. Hochreiter and J. Schmidhuber, “Long Short-term Memory,” *Neural computation* **9** (12, 1997) 1735–80. [[doi](#)].
- [80] “NanoAODv7.”
<https://gitlab.cern.ch/cms-nanoAOD/nanoaod-doc/-/wikis/Releases/NanoAODv7>.
- [81] “Golden JSON 2016.” https://cms-service-dqm.web.cern.ch/cms-service-dqm/CAF/certification/Collisions16/13TeV/ReReco/Final/Cert_271036-284044_13TeV_23Sep2016ReReco_Collisions16_JSON.txt.
- [82] “Golden JSON 2017.” https://cms-service-dqm.web.cern.ch/cms-service-dqm/CAF/certification/Collisions17/13TeV/ReReco/Cert_294927-306462_13TeV_EOY2017ReReco_Collisions17_JSON_v1.txt.
- [83] “Golden JSON 2018.” https://cms-service-dqm.web.cern.ch/cms-service-dqm/CAF/certification/Collisions18/13TeV/ReReco/Cert_314472-325175_13TeV_17SeptEarlyReReco2018ABC_PromptEraD_Collisions18_JSON.txt.
- [84] **NNPDF** Collaboration, R. D. Ball, V. Bertone, S. Carrazza, C. S. Deans, L. Del Debbio, S. Forte, A. Guffanti, N. P. Hartland, J. I. Latorre, J. Rojo, and M. Ubiali, “Parton distributions for the LHC Run II,” *JHEP* **04** (2015) 040. [arXiv:1410.8849](https://arxiv.org/abs/1410.8849) [[hep-ph](#)].

- [85] T. Sjostrand, S. Mrenna, and P. Z. Skands, “A Brief Introduction to PYTHIA 8.1,” *Comput. Phys. Commun.* **178** (2008) 852–867. [arXiv:0710.3820 \[hep-ph\]](#).
- [86] CMS Collaboration, V. Khachatryan *et al.*, “Event generator tunes obtained from underlying event and multiparton scattering measurements,” *Eur. Phys. J. C* **76** (2016) 155. [arXiv:1512.00815 \[hep-ex\]](#).
- [87] CMS Collaboration, A. M. Sirunyan *et al.*, “Extraction and validation of a new set of CMS PYTHIA8 tunes from underlying-event measurements,” *Eur. Phys. J. C* **80** (2020) no. 1, 4. [arXiv:1903.12179 \[hep-ex\]](#).
- [88] R. Frederix and S. Frixione, “Merging meets matching in MC@NLO,” *JHEP* **12** (2012) 061. [arXiv:1209.6215 \[hep-ph\]](#).
- [89] J. Alwall, S. Höche, F. Krauss, N. Lavesson, L. Lönnblad, 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** (2008) 473. [arXiv:0706.2569 \[hep-ph\]](#).
- [90] D. Abercrombie , P. Berger , V. Brigljevic , A. Calandri , B. Chitroda , D. Ferencek , K. Gedia , G. Gomez-Ceballos , C. Grab , N. Haubrich , H. Kaveh , J. Konigsberg , D. Kovalskiy , B. Maier , R. Mankel , S. Mishra , A. Nigamova , J. Olsen , C. Palmer , C. Paus , G. Perrin , L. Perrozzi , C. Reissel , A. Rizzi , A. Schmidt , T. Susa , A. de Wit, “Simplified template cross section measurements in VH, H to bb with full Run 2 data set,” *CMS Physics Analysis Note* **AN-19-229** (2019) . [\[link\]](#).
- [91] P. Nason, “A New method for combining NLO QCD with shower Monte Carlo algorithms,” *JHEP* **0411** (2004) 040. [arXiv:hep-ph/0409146 \[hep-ph\]](#).
- [92] S. Frixione, P. Nason, and C. Oleari, “Matching NLO QCD computations with parton shower simulations: the POWHEG method,” *JHEP* **11** (2007) 070. [arXiv:0709.2092 \[hep-ph\]](#).
- [93] 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** (2010) 043. [arXiv:1002.2581 \[hep-ph\]](#).
- [94] K. Hamilton, P. Nason, and G. Zanderighi, “MINLO: Multi-Scale Improved NLO,” *JHEP* **10** (2012) 155. [arXiv:1206.3572 \[hep-ph\]](#).

- [95] G. Luisoni, P. Nason, C. Oleari, and F. Tramontano, “ $\text{HW}^\pm/\text{HZ} + 0$ and 1 jet at NLO with the POWHEG BOX interfaced to GoSam and their merging within MiNLO,” *JHEP* **10** (2013) 083. [arXiv:1306.2542 \[hep-ph\]](#).
- [96] J. Alwall, S. Höche, F. Krauss, N. Lavesson, L. Lönnblad, 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** (2008) 473. [arXiv:0706.2569 \[hep-ph\]](#).
- [97] M. Czakon and A. Mitov, “Top++: A program for the calculation of the top-pair cross-section at hadron colliders,” *Comput. Phys. Commun.* **185** (2014) 2930. [arXiv:1112.5675 \[hep-ph\]](#).
- [98] M. Czakon, D. Heymes, A. Mitov, D. Pagani, I. Tsinikos, and M. Zaro, “Top-pair production at the LHC through NNLO QCD and NLO EW,” *JHEP* **10** (2017) 186. [arXiv:1705.04105 \[hep-ph\]](#).
- [99] M. Grazzini, S. Kallweit, J. M. Lindert, S. Pozzorini, and M. Wiesemann, “NNLO QCD + NLO EW with Matrix+OpenLoops: precise predictions for vector-boson pair production,” *JHEP* **02** (2020) 087. [arXiv:1912.00068 \[hep-ph\]](#).
- [100] “MVA electron ID.” <https://twiki.cern.ch/twiki/bin/viewauth/CMS/MultivariateElectronIdentificationRun2>.
- [101] CMS Collaboration, C. Collaboration, “Pileup Removal Algorithms,”.
- [102] “Jet Identification.” https://twiki.cern.ch/twiki/bin/view/CMS/JetID#nanoAOD_Flags,r94.
- [103] “Jet identification in high pile-up environment (PileupJetID).” [https://twiki.cern.ch/twiki/bin/viewauth/CMS/PileupJetID, r64](https://twiki.cern.ch/twiki/bin/viewauth/CMS/PileupJetID,r64).
- [104] CMS Collaboration, CMS, “A deep neural network for simultaneous estimation of b jet energy and resolution,” *Comput. Softw. Big Sci.* **4** (Dec, 2019) 10. 20 p. [[link](#)]. Submitted to Computing and Software for Big Science. All figures and tables can be found at <http://cms-results.web.cern.ch/cms-results/public-results/publications/HIG-18-027> (CMS Public Pages).

- [105] A. Hoecker, P. Speckmayer, J. Stelzer, J. Therhaag, E. von Toerne, H. Voss, M. Backes, T. Carli, O. Cohen, and A. Christov, “TMVA - Toolkit for Multivariate Data Analysis,” *arXiv e-prints* (Mar, 2007) physics/0703039. [arXiv:physics/0703039](#) [[physics.data-an](#)].
- [106] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees*. Routledge, 2017.
- [107] J. H. Friedman, “Greedy Function Approximation: A Gradient Boosting Machine,” *The Annals of Statistics* **29** (2001) no. 5, 1189–1232. [[link](#)].
- [108] **SLD** Collaboration, K. Abe *et al.*, “Measurement of the branching ratio of the Z0 into heavy quarks,” *Phys. Rev. D* **71** (2005) 112004. [arXiv:hep-ex/0503005](#).
- [109] **DELPHI** Collaboration, P. Abreu *et al.*, “Measurements of the Z partial decay width into c anti-c and multiplicity of charm quarks per b decay,” *Eur. Phys. J. C* **12** (2000) 225–241. [[doi](#)].
- [110] **ALEPH** Collaboration, R. Barate *et al.*, “Study of charm production in Z decays,” *Eur. Phys. J. C* **16** (2000) 597–611. [arXiv:hep-ex/9909032](#).
- [111] **OPAL** Collaboration, K. Ackerstaff *et al.*, “Measurement of $f(c \rightarrow D^* + X)$, $f(b \rightarrow D^* + X)$ and $\Gamma(c \text{ anti-}c) / \Gamma(\text{hadronic})$ using $D^{*+/-}$ mesons,” *Eur. Phys. J. C* **1** (1998) 439–459. [arXiv:hep-ex/9708021](#).
- [112] **OPAL** Collaboration, G. Alexander *et al.*, “A Study of charm hadron production in $Z0 \rightarrow c \text{ anti-}c$ and $Z0 \rightarrow b \text{ anti-}b$ decays at LEP,” *Z. Phys. C* **72** (1996) 1–16. [[doi](#)].
- [113] **ATLAS Collaboration** Collaboration, “Direct constraint on the Higgs-charm coupling from a search for Higgs boson decays to charm quarks with the ATLAS detector,” tech. rep., CERN, Geneva, Jun, 2021. [[link](#)]. All figures including auxiliary figures are available at <https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/CONFNOTES/ATLAS-CONF-2021-021>.
- [114] A. Collaboration, “Extrapolation of ATLAS sensitivity to $H \rightarrow b\bar{b}$ and $H \rightarrow c\bar{c}$ decays in VH production at the HL-LHC,” tech. rep., CERN, Geneva, Oct, 2021. [[link](#)]. All figures including auxiliary figures are available at <https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PUBNOTES/ATL-PHYS-PUB-2021-039>.

- [115] **LHC Higgs Cross Section Working Group** Collaboration, S. Heinemeyer, C. Mariotti, G. Passarino, R. Tanaka, J. R. Andersen, P. Artoisenet, E. A. Bagnaschi, A. Banfi, T. Becher, F. U. Bernlochner, S. Bolognesi, P. Bolzoni, R. Boughezal, D. Buarque, J. Campbell, F. Caola, M. Carena, F. Cascioli, N. Chanon, T. Cheng, S. Y. Choi, A. David, P. de Aquino, G. Degrand, D. Del Re, A. Denner, H. van Deurzen, S. Diglio, B. Di Micco, R. Di Nardo, S. Dittmaier, M. Dührssen, R. K. Ellis, G. Ferrera, N. Fianza, M. Flechl, D. de Florian, S. Forte, R. Frederix, S. Frixione, S. Gangal, Y. Gao, M. V. Garzelli, D. Gillberg, P. Govoni, M. Grazzini, N. Greiner, J. Griffiths, A. V. Gritsan, C. Grojean, D. C. Hall, C. Hays, R. Harlander, R. Hernandez-Pinto, S. Höche, J. Huston, T. Jubb, M. Kadastik, S. Kallweit, A. Kardos, L. Kashif, N. Kauer, H. Kim, R. Klees, M. Krämer, F. Krauss, A. Laureys, S. Laurila, S. Lehti, Q. Li, S. Liebler, X. Liu, E. Logan, G. Luisoni, M. Malberti, F. Maltoni, K. Mawatari, F. Maierhofer, H. Mantler, S. Martin, P. Mastrolia, O. Mattelaer, J. Mazzitelli, B. Mellado, K. Melnikov, P. Meridiani, D. J. Miller, E. Mirabella, S. O. Moch, P. Monni, N. Moretti, A. Mück, M. Mühlleitner, P. Musella, P. Nason, C. Neu, M. Neubert, C. Oleari, J. Olsen, G. Ossola, T. Peraro, K. Peters, F. Petriello, G. Piacquadio, C. T. Potter, S. Pozzorini, K. Prokofiev, I. Puljak, M. Rauch, D. Rebuzzi, L. Reina, R. Rietkerk, A. Rizzi, Y. Rotstein-Habarnau, G. P. Salam, G. Sborlini, F. Schissler, M. Schönherr, M. Schulze, M. Schumacher, F. Siegert, P. Slavich, J. M. Smillie, O. Stål, J. F. von Soden-Fraunhofen, M. Spira, I. W. Stewart, F. J. Tackmann, P. T. E. Taylor, D. Tommasini, J. Thompson, R. S. Thorne, P. Torrielli, F. Tramontano, N. V. Tran, Z. Trócsányi, M. Ubiali, M. Vazquez Acosta, T. Vique, A. Vicini, W. J. Waalewijn, D. Wackerth, C. Wagner, J. R. Walsh, J. Wang, G. Weiglein, A. Whitbeck, C. Williams, J. Yu, G. Zanderighi, M. Zanetti, M. Zaro, P. M. Zerwas, C. Zhang, T. J. E. Zirke, and S. Zuberi, *Handbook of LHC Higgs Cross Sections: 3. Higgs Properties: Report of the LHC Higgs Cross Section Working Group*. CERN Yellow Reports: Monographs. Jul, 2013. [link]. Comments: 404 pages, 139 figures, to be submitted to CERN Report. Working Group web page: <https://twiki.cern.ch/twiki/bin/view/LHCPhysics/CrossSections>.
- [116] D. Florian, C. Grojean, F. Maltoni, C. Mariotti, A. Nikitenko, M. Pieri, P. Savard, M. Schumacher, R. Tanaka, R. Aggleton, M. Ahmad, B. Allanach, C. Anastasiou, W. Astill, S. Badger, M. Badziak, J. Baglio, E. Bagnaschi, A. Ballestrero, A. Banfi, D. Barducci, M. Beckingham, C. Becot, G. Bélanger, J. Bellm, N. Belyaev, F. Bernlochner, C. Beskidt, A. Biekötter, F. Bishara, W. Bizon, N. Bomark, M. Bonvini, S. Borowka, V. Bortolotto,

S. Boselli, F. Botella, R. Boughezal, G. Branco, J. Brehmer, L. Brenner, S. Bressler,
 I. Brivio, A. Broggio, V. Brun, G. Buchalla, C. Burgard, A. Calandri, L. Caminada,
 R. Armadans, F. Campanario, J. Campbell, F. Caola, C. Calame, S. Carrazza, A. Carvalho,
 M. Casolino, O. Cata, A. Celis, F. Cerutti, N. Chanon, M. Chen, X. Chen, B. Nejad,
 N. Christensen, M. Ciuchini, R. Contino, T. Corbett, S. Costa, D. Curtin, M. Dall’Osso,
 A. David, S. Dawson, J. Blas, W. Boer, P. Manzano, C. Degrande, L. Rueda-Delgado,
 F. Demartin, A. Denner, B. Micco, R. Nardo, S. Dittmaier, A. Dobado, T. Dorigo, F. Dreyer,
 M. Dührssen, C. Duhr, F. Dulat, K. Ecker, K. Ellis, U. Ellwanger, C. Englert, D. Espriu,
 A. Falkowski, L. Fayard, R. Feger, G. Ferrera, A. Ferrogli, N. Fianza, T. Figy, M. Flechl,
 J. Fontes, S. Forte, P. Francavilla, E. Franco, R. Frederix, A. Freitas, F. Freitas, F. Frensch,
 S. Frixione, B. Fuks, E. Furlan, S. Gadatsch, J. Gao, Y. Gao, M. Garzelli, T. Gehrmann,
 R. Gerosa, M. Ghezzi, D. Ghosh, S. Gieseke, D. Gillberg, G. Giudice, E. Glover, F. Goertz,
 D. Gonçalves, J. Gonzalez-Fraile, M. Gorbahn, S. Gori, C. Gottardo, M. Gouzevitch,
 P. Govoni, D. Gray, M. Grazzini, N. Greiner, A. Greljo, J. Grigo, A. Gritsan, R. Gröber,
 S. Guindon, H. Haber, C. Han, T. Han, R. Harlander, M. Harrendorf, H. Hartanto, C. Hays,
 S. Heinemeyer, G. Heinrich, M. Herrero, F. Herzog, B. Hespel, V. Hirschi, S. Hoeche,
 S. Honeywell, S. Huber, C. Hugonie, J. Huston, A. Ilnicka, G. Isidori, B. Jäger, M. Jaquier,
 S. Jones, A. Juste, S. Kallweit, A. Kaluza, A. Kardos, A. Karlberg, Z. Kassabov, N. Kauer,
 D. Kazakov, M. Kerner, W. Kilian, M. Kling, K. Köneke, R. Kogler, R. Konoplich,
 S. Kortner, S. Kraml, C. Krause, F. Krauss, P. Krawczyk, A. Kulesza, S. Kuttimalai,
 R. Lane, A. Lazopoulos, W. Lee, P. Lenzi, I. Lewis, Y. Li, S. Liebler, E. van Lindert, X. Liu,
 Z. Liu, F. Llanes-Estrada, H. Logan, D. Lopez-Val, I. Low, G. Luisoni, P. Maierhöfer,
 E. Maina, B. Mansoulié, H. Mantler, M. Mantoani, A. Marini, V. Outschoorn, S. Marzani,
 D. Marzocca, A. Massironi, K. Mawatari, J. Mazzitelli, A. McCarn, B. Mellado, K. Melnikov,
 S. Menari, L. Merlo, C. Meyer, P. Milenovic, K. Mimasu, S. Mishima, B. Mistlberger,
 S. Moch, A. Mohammadi, P. Monni, G. Montagna, M. Llácer, N. Moretti, S. Moretti,
 L. Motyka, A. Mück, M. Mühlleitner, S. Munir, P. Musella, P. Nadolsky, D. Napoletano,
 M. Nebot, C. Neu, R. Neubert, R. Nevzorov, O. Nicosini, J. Nielsen, K. Nikolopoulos, J. No,
 C. O’Brien, T. Ohl, C. Oleari, T. Orimoto, D. Pagani, C. Pandini, A. Papaefstathiou,
 A. Papanastasiou, G. Passarino, B. Pecjak, M. Pelliccioni, J. Perez, L. Perrozzi, F. Petriello,
 G. Petrucciani, E. Pianori, F. Piccinini, M. Pierini, A. Pilkington, S. Plätzer, T. Plehn,
 R. Podskubka, C. Potter, S. Pozzorini, K. Prokofiev, A. Pukhov, I. Puljak,
 M. Queitsch-Maitland, J. Quevillon, D. Rathlev, S. Rauch, E. Re, M. Rebelo, D. Rebuzzi,

- L. Reina, C. Reuschle, J. Reuter, M. Riembau, F. Riva, A. Rizzi, T. Robens, R. Röntsch, J. Rojo, J. Romão, N. Rompotis, J. Roskes, R. Roth, G. Salam, R. Salerno, M. Sampaio, R. Santos, V. Sanz-Nebot, J. Sanz-Cillero, H. Sargsyan, U. Sarica, P. Schichtel, J. Schlenk, T. Schmidt, C. Schmitt, M. Schönherr, U. Schubert, M. Schulze, S. Sekula, M. Sekulla, E. Shabalina, H. Shao, J. Shelton, C. Shepherd-Themistocleous, S. Shim, F. Siegert, A. Signer, J. Da Silva, L. Silvestrini, M. Sjodahl, P. Slavich, M. Slawinska, L. Soffi, M. Spannowsky, C. Speckner, D. Sperka, M. Spira, O. Stål, F. Staub, T. Stebel, T. Stefaniak, M. Steinhauser, I. Stewart, M. Strassler, J. Streicher, D. Strom, S. Su, X. Sun, F. Tackmann, K. Tackmann, A. Teixeira, R. Lima, V. Theeuwes, R. Thorne, D. Tommasini, P. Torrielli, M. Tosi, F. Tramontano, Z. Trócsányi, M. Trott, I. Tsirikos, M. Ubiali, P. Vanlaer, W. Verkerke, A. Vicini, L. Viliani, E. Vryonidou, D. Wackerroth, C. Wagner, J. Wang, S. Wayand, G. Weiglein, C. Weiss, M. Wiesenmann, C. Williams, J. Winter, D. Winterbottom, R. Wolf, S. Xiao, L. Yang, R. Yohay, S. Yuen, G. Zanderighi, M. Zaro, D. Zeppenfeld, R. Ziegler, T. Zirke, and J. Zupan, “Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector,” *arXiv.org* (Oct., 2016) . [doi]. 869 pages, 295 figures, 248 tables and 1645 citations. Working Group web page: <https://twiki.cern.ch/twiki/bin/view/LHCPhysics/LHCHXSWG>.
- [117] G. Perez, Y. Soreq, E. Stamou, and K. Tobioka, “Constraining the charm Yukawa and Higgs-quark coupling universality,” *Physical Review D* **92** (Aug, 2015) . [link].
- [118] A. Dainese, M. Managno, A. B. Meyer, A. Nisati, G. Salam, and M. . Vesterinen, *Report on the Physics of the HL=LHC, and perspectives at the HE-LHC*. 2019. [link].
- [119] L. Cranberg, G. Frye, N. Nereson, and L. Rosen, “Fission Neutron Spectrum of U^{235} ,” *Phys. Rev.* **103** (Aug, 1956) 662–670. [link].
- [120] C. Collaboration, “Performance of CMS hadron calorimeter timing and synchronization using test beam, cosmic ray, and LHC beam data,” *Journal of Instrumentation* **5** (mar, 2010) T03013–T03013. [link].
- [121] S. Bertolucci, M. Cordelli, B. Esposito, M. Curatolo, P. Giromini, S. Miscetti, A. Sansoni, G. Apollinari, F. Bedeschi, S. Belforte, G. Bellettini, N. Bonavita, F. Cervelli, G. Chiarelli, R. Del Fabbro, M. Dell’Orso, E. Focardi, P. Giannetti, A. Menzione, R. Paoletti, G. Punzi, L. Ristori, A. Scribano, P. Sestini, A. Stefanini, G. Tonelli, F. Zetti, V. Barnes, A. Di Virgilio, A. Garfinkel, S. Kuhlmann, A. Laasanen, H. Jensen, and H. Kautzky, “The CDF

- central and endwall hadron calorimeter,” *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **267** (1988) no. 2, 301–314. [[link](#)].
- [122] D. Thomas and A. Alevra, “Bonner sphere spectrometers – a critical review,” *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **476** (2002) no. 1, 12–20. [[link](#)]. Int. Workshop on Neutron Field Spectrometry in Science, Technology and Radiation Protection.
- [123] H. Morinaga and T. Yamazaki, *In-beam gamma-ray spectroscopy*. 1976.
- [124] P. Kocherov, N.P.; McLaughlin, “The international reactor dosimetry file. (IRDF-90 Version 2),” Report IAEA-NDS-141(REV.3), International Atomic Energy Agency, Vienna (Austria). Nuclear Data Section, 1996.
- [125] F. James, “MINUIT Function Minimization and Error Analysis: Reference Manual Version 94.1,”.
- [126] D. M. P. Lazo, M.S.; Woodall, “Silicon and Silicon dioxide neutron damage functions,” Conference SAND-87-0098-VOL.1, 1987.
- [127] C. Foudas, R. Bainbridge, D. Ballard, I. Church, E. Corrin, J. Coughlan, C. Day, E. Freeman, J. Fulcher, W. Gannon, G. Hall, R. Halsall, G. Iles, J. Jones, J. Leaver, M. Noy, M. Pearson, M. Raymond, I. Reid, G. Rogers, J. Salisbury, S. Taghavi, I. Tomalin, and O. Zorba, “The CMS tracker readout front end driver,” *IEEE Transactions on Nuclear Science* **52** (2005) no. 6, 2836–2840. [[doi](#)].
- [128] M. Raymond, G. Cervelli, M. French, J. R. Fulcher, G. Hall, L. Jones, L. K. Lim, G. Marseguerra, P. Moreira, Q. R. Morrissey, A. Neviani, and E. Noah, “The CMS tracker APV25 0.25 μ m CMOS readout chip,”. [[link](#)].
- [129] J. Troska, G. Cervelli, F. Faccio, K. Gill, R. Grabit, A. M. Sandvik, F. Vasey, and A. Zanet, “Optical Readout and Control Systems for the CMS Tracker,”. [[link](#)].
- [130] **CMS Collaboration** Collaboration, S. Cittolin, A. Rácz, and P. Sphicas, *CMS The TriDAS Project: Technical Design Report, Volume 2: Data Acquisition and High-Level Trigger*. CMS trigger and data-acquisition project. Technical design report. CMS. CERN, Geneva, 2002. [[link](#)].

- [131] J. K. M. P. W. T. Christopher Jones, Elizabeth Sexton-Kennedy, “The New CMS Event Data Model and Framework.” CHEP 2006, 2006.
- [132] M. Andrews, M. Paulini, S. Gleyzer, and B. Poczoz, “End-to-End Physics Event Classification with CMS Open Data: Applying Image-Based Deep Learning to Detector Data for the Direct Classification of Collision Events at the LHC,” 2018. [arXiv:1807.11916 \[hep-ex\]](#).
- [133] M. Andrews, J. Alison, S. An, B. Burkle, S. Gleyzer, M. Narain, M. Paulini, B. Poczoz, and E. Usai, “End-to-end jet classification of quarks and gluons with the CMS Open Data,” *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **977** (Oct, 2020) 164304. [\[link\]](#).
- [134] J. Thaler and K. Van Tilburg, “Identifying boosted objects with N-subjettiness,” *Journal of High Energy Physics* **2011** (Mar, 2011) . [\[link\]](#).
- [135] CMS Collaboration, “CMS data preservation, re-use and open access policy,” 2014. [\[link\]](#).
- [136] CMS Collaboration, “Tracker-hit-enriched TTJets_HadronicMGDecays_8TeV-madgraph,” 2019. [\[link\]](#).
- [137] CMS Collaboration, “Tracker-hit-enriched 300 to 600 bin of QCD_Pt-15to3000_TuneZ2star_Flat_8TeV_pythia6,” 2019. [\[link\]](#).
- [138] CMS Collaboration, “Tracker-hit-enriched 400 to 600 bin of QCD_Pt-15to3000_TuneZ2star_Flat_8TeV_pythia6,” 2019. [\[link\]](#).
- [139] CMS Collaboration, “Tracker-hit-enriched 600 to 3000 bin of QCD_Pt-15to3000_TuneZ2star_Flat_8TeV_pythia6,” 2019. [\[link\]](#).
- [140] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H.-S. Shao, T. Stelzer, P. Torrielli, and M. Zaro, “The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations,” *Journal of High Energy Physics* **2014** (Jul, 2014) . [\[link\]](#).
- [141] T. Sjöstrand, S. Mrenna, and P. Skands, “PYTHIA 6.4 physics and manual,” *Journal of High Energy Physics* **2006** (May, 2006) 026–026. [\[link\]](#).
- [142] CMS Collaboration, “CMS Software Version 5_3_32 (CMSSW_5_3_32),” 2016. [\[link\]](#).
- [143] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” 2015. [arXiv:1512.03385 \[cs.CV\]](#).

- [144] A. F. Agarap, “Deep learning using rectified linear units (relu),” *arXiv preprint arXiv:1803.08375* (2018) .
- [145] J. Nagi, F. Ducatelle, G. A. Di Caro, D. Ciresan, U. Meier, A. Giusti, F. Nagi, J. Schmidhuber, and L. M. Gambardella, “Max-pooling convolutional neural networks for vision-based hand gesture recognition,” in *2011 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, pp. 342–347. 2011. [doi].
- [146] S. Narayan, “The generalized sigmoid activation function: Competitive supervised learning,” *Information Sciences* **99** (1997) no. 1, 69–82. [link].
- [147] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization.” 2014.
- [148] A. J. Larkoski, S. Marzani, G. Soyez, and J. Thaler, “Soft Drop,” *JHEP* **05** (2014) 146. [arXiv:1402.2657 \[hep-ph\]](#).
- [149] **CMS** Collaboration, A. M. Sirunyan *et al.*, “Identification of heavy, energetic, hadronically decaying particles using machine-learning techniques,” *JINST* **15** (2020) no. 06, P06005. [arXiv:2004.08262 \[hep-ex\]](#).
- [150] *Western Digital DC HA210 Datasheet, 3.5 Inch Data Center Hard Drives.* [link]. Accessed: 30 August 2020.
- [151] *Intel(R) Xeon(R) Silver 4110 Processor (11M Cache, 2.10 GHz) Product Specifications.* [link]. Accessed: 16 September 2020.
- [152] *Intel(R) Xeon(R) Gold 5118 Processor (16.5M Cache, 2.30 GHz) Product Specifications.* [link]. Accessed: 16 September 2020.
- [153] *NVIDIA DGX-1: Deep Learning Server for AI Research.* [link]. Accessed: 16 September 2020.
- [154] N. P. Jouppi, C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa, S. Bates, S. Bhatia, N. Boden, A. Borchers, R. Boyle, P. luc Cantin, C. Chao, C. Clark, J. Coriell, M. Daley, M. Dau, J. Dean, B. Gelb, T. V. Ghaemmaghami, R. Gottipati, W. Gulland, R. Hagmann, C. R. Ho, D. Hogberg, J. Hu, R. Hundt, D. Hurt, J. Ibarz, A. Jaffey, A. Jaworski, A. Kaplan, H. Khaitan, A. Koch, N. Kumar, S. Lacy, J. Laudon, J. Law, D. Le, C. Leary, Z. Liu, K. Lucke, A. Lundin, G. MacKean, A. Maggiore, M. Mahony, K. Miller, R. Nagarajan, R. Narayanaswami, R. Ni, K. Nix, T. Norrie, M. Omernick, N. Penukonda, A. Phelps,

- J. Ross, M. Ross, A. Salek, E. Samadiani, C. Severn, G. Sizikov, M. Snelham, J. Souter, D. Steinberg, A. Swing, M. Tan, G. Thorson, B. Tian, H. Toma, E. Tuttle, V. Vasudevan, R. Walter, W. Wang, E. Wilcox, and D. H. Yoon, “In-Datacenter Performance Analysis of a Tensor Processing Unit,”. [arXiv:1704.04760 \[cs.AR\]](#).
- [155] *Google Cloud Computing Services*. [\[link\]](#). Accessed: 30 August 2020.
- [156] *BFloat16: The secret to high performance on Cloud TPUs* | *Google Cloud Blog*. [\[link\]](#). Accessed: 18 September 2020.
- [157] *Storage classes* | *Google Cloud*. [\[link\]](#). Accessed: 29 September 2020.
- [158] S.-W. Lee, B. Moon, and C. Park, “Advances in Flash Memory SSD Technology for Enterprise Database Applications,” in *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data*, SIGMOD ’09, pp. 863–870. Association for Computing Machinery, New York, NY, USA, 2009. [\[link\]](#).