

UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Search for Anomalous Production of the Higgs Boson at the CMS Experiment and Exascale Computing for the LHC

Permalink

<https://escholarship.org/uc/item/9hf0p5zz>

ISBN

9798193289630

Author

Arora, Aashay

Publication Date

2026-10-12

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA SAN DIEGO

Search for Anomalous Production of the Higgs Boson at the CMS Experiment and
Exascale Computing for the LHC

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Physics

by

Aashay Arora

Committee in charge:

Professor Frank K. Würthwein, Chair
Professor Daniel R. Green
Professor Ramesh R. Rao
Professor Avi Yagil

Copyright

Aashay Arora, 2026

All rights reserved.

The Dissertation of Aashay Arora is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2026

EPIGRAPH

Not only is the universe stranger than we think,
it is stranger than we can think

Werner Heisenberg

Yesterday is history, tomorrow is a mystery, but today is a gift.
That's why it's called the present.

Master Oogway, *Kung Fu Panda*

TABLE OF CONTENTS

Dissertation Approval Page	iii
Epigraph	iv
Table of Contents	v
List of Figures	vii
List of Tables	xiv
Acknowledgements	xvi
Vita	xviii
Abstract of the Dissertation	xix
Introduction	1
Chapter 1 Standard Model of Particle Physics	3
1.1 Classical Field Theory and the Gauge Principle	4
1.2 The Standard Model Lagrangian	7
1.3 The Gauge Sector of the Standard Model	8
1.4 Electroweak Symmetry Breaking and the Higgs Mechanism	10
1.5 The Flavor Sector of the Standard Model	14
1.6 The Free Parameters of the Standard Model	15
1.7 From Amplitudes to Observables	16
1.8 The Higgs Boson	21
1.9 Limitations of the Standard Model and Physics Beyond	25
Chapter 2 The Large Hadron Collider and the CMS Experiment	27
2.1 The Large Hadron Collider (LHC)	27
2.2 The Compact Muon Solenoid (CMS) Detector	35
2.2.1 The Phase-2 Upgrade for the HL-LHC	52
2.3 World Wide LHC Computing Grid	57
Chapter 3 Simulation and Reconstruction	60
3.1 Simulation	60
3.2 Reconstruction	63
3.2.1 Charged-Particle Track Reconstruction at CMS	63
3.2.2 Particle Flow	69
3.2.3 Reconstructed Objects	70
Chapter 4 Anomalous Higgs Boson Production	76
4.1 Vector Boson Scattering (VBS)	76

4.2	Associated VVH Production and the Higgs Couplings	78
4.2.1	Limitations of the κ -framework	81
4.3	Overview of the Search	82
4.4	The Previous Run 2 Search	83
4.4.1	Datasets and Objects	83
4.4.2	Analysis Strategy and Event Selection	83
4.4.3	Background Estimation and Signal Extraction	84
4.4.4	Results	84
4.5	The Run 2 + Run 3 Analysis	85
4.5.1	Data and Simulated Samples	86
4.5.2	Object Reconstruction and Identification	89
4.5.3	Analysis Strategy and Event Selection	93
4.5.4	Background Estimation and Signal Extraction	94
4.5.5	The All-Hadronic Fully-Boosted Channel	99
4.5.6	The Single-Lepton Fully-Boosted Channel	108
4.5.7	The Single-Lepton Semi-Boosted Channel	111
4.5.8	Statistical Methods	116
4.5.9	Signal Region Definition and Fit	121
4.5.10	Systematic Uncertainties	124
4.5.11	Results	125
4.5.12	Summary and Outlook	128
Chapter 5	Tracking Performance and Upgrades for the High-Luminosity LHC	131
5.1	Line Segment Tracking (LST)	131
5.2	Machine Learning Based Track Reconstruction	135
5.3	Summary	147
Chapter 6	Cyberinfrastructure for the LHC	149
6.1	Introduction: The HL-LHC Data Challenge	149
6.2	Saturating the Data Plane: Benchmarking to 400 Gbps	151
6.3	Network-Aware Data Movement: The SENSE-Rucio Architecture	159
6.4	The Data Movement Manager (DMM)	161
6.5	Evaluation on a Multi-Domain Testbed	166
6.6	Summary	169
Appendix A	Supplementary Distributions	171
A.1	All-Hadronic Fully-Boosted Channel	172
A.2	Single-Lepton Fully-Boosted Channel	178
A.3	Single-Lepton Semi-Boosted Channel	186
Bibliography		196

LIST OF FIGURES

Figure 1.1.	The particles of the Standard Model, categorized into fermions (quarks and leptons), gauge bosons (force carriers), and the Higgs boson.	4
Figure 1.2.	The Higgs potential $V(H)$ for $\mu^2 < 0$, with a continuum of degenerate minima at $H^\dagger H = v^2/2$	11
Figure 1.3.	Leading-order Feynman diagrams for the dominant Higgs boson production mechanisms at the LHC.	23
Figure 1.4.	Measured Higgs coupling strengths versus particle mass, compared with the SM mass-proportional prediction [34].	25
Figure 1.5.	The Higgs potential, as an expansion about the electroweak vacuum, indicating regions of stability and metastability.	26
Figure 2.1.	The CERN accelerator complex feeding the LHC, with the four main experiments at their interaction points [49].	28
Figure 2.2.	Beam optics near an LHC interaction point (top) and an example β -function profile defining β^* (bottom).	30
Figure 2.3.	Interactions per bunch crossing in CMS during Phase-1 operation, with the mean $\langle\mu\rangle$ per run [35].	34
Figure 2.4.	Timeline of the LHC runs and long shutdowns, with the energy and integrated luminosity of each, including the HL-LHC [27].	35
Figure 2.5.	Schematic view of the CMS detector, showing the various subdetectors and their arrangement around the interaction point [29].	36
Figure 2.6.	The CMS coordinate system: the x, y, z axes, azimuth ϕ , polar angle θ , and pseudorapidity η [29].	37
Figure 2.7.	Value of $ B $ and the field lines predicted on a longitudinal section of CMS [39]	38
Figure 2.8.	Mean energy loss $\langle -dE/dx \rangle$ versus $\beta\gamma$ for muons, pions, and protons, showing the minimum-ionizing regime and the relativistic rise [71].	40
Figure 2.9.	One quarter of the Phase-1 CMS tracker (r - z view): pixels (green), single- and double-sided strip modules (red, blue) [16].	41
Figure 2.10.	Fractional electron energy loss per radiation length from ionization and bremsstrahlung, crossing at the critical energy E_c [71].	44

Figure 2.11.	Photon mass attenuation versus energy, decomposed into the photoelectric effect, Compton scattering, and pair production [71].	44
Figure 2.12.	Simplified electromagnetic shower cascade, developing over the radiation length X_0 [62].	45
Figure 2.13.	Layout of the CMS electromagnetic calorimeter [29].	46
Figure 2.14.	Schematic view of one quadrant of the CMS hadronic calorimeter [29]. . .	48
Figure 2.15.	Simplified hadronic shower cascade, developing over the nuclear interaction length λ_I [62].	49
Figure 2.16.	Geometry of Cherenkov radiation: a coherent wavefront at half-angle θ_C with $\cos \theta_C = 1/(n\beta)$ [62].	50
Figure 2.17.	The CMS muon system (top): drift tubes, cathode strip chambers, and resistive plate chambers; cross-section of a DT and a CSC (bottom) [29]. .	51
Figure 2.18.	One quarter of the Phase-2 CMS tracker (r - z view): the Outer Tracker is built from PS modules (blue) at smaller radius and 2S modules (red) at larger radius, with the Inner Tracker pixel modules shown in green and orange [1].	52
Figure 2.19.	Tiered organization of the WLCG: data flows from the detector to the Tier-0 at CERN, then out to the Tier-1 and Tier-2 sites [64].	58
Figure 3.1.	Schematic of the simulation chain, data collection and reconstruction in CMS.	61
Figure 3.2.	Parton shower and hadronization: the hard interaction (red), the parton-shower cascade (blue), secondary scatters (purple), and hadronization (green).	62
Figure 3.3.	The five helix track parameters at the point of closest approach: transverse (x - y , left) and longitudinal (r - z , right) views.	64
Figure 3.4.	Track building with the combinatorial Kalman filter, shown in a transverse slice of the tracker	69
Figure 3.5.	Schematic representation of the deposits in the CMS detector and the corresponding reconstructed objects [20]	70
Figure 4.1.	Representative tree-level VVH vector-boson-scattering diagrams showing the dependence on κ_λ (left) and κ_{2V} (right).	79

Figure 4.2.	Summary of production cross sections measured by CMS, from the inelastic proton-proton cross section to the rarest electroweak processes, at center-of-mass energies from 2.76 to 13.6 TeV.	80
Figure 4.3.	Previous Run 2 search: 95% CL upper limit on the VBS VVH cross section versus κ_{2V} (left) and the exclusion contour in the $(\kappa_{2W}, \kappa_{2Z})$ plane (right) [5, 40].	85
Figure 4.4.	LHE-level $\Delta\eta_{jj}$ (left) and scalar p_T sum of the three bosons (right) at several $(\kappa_\lambda, \kappa_{2V})$ benchmark points.	87
Figure 4.5.	Representative tree-level Feynman diagrams for the principal backgrounds of the leptonic channels.	89
Figure 4.6.	Training and inference flow of the VBS jet-pair selection BDT.	92
Figure 4.7.	Schematic of the ABCD method: a signal region A and three background-dominated control regions B, C, D, related by Eq. (4.5).	96
Figure 4.8.	Closure test of the ABCD prediction (Eq. (4.5)) against the observed yield in a control region.	98
Figure 4.9.	Final-state topology of the all-hadronic fully-boosted channel.	101
Figure 4.10.	QCD tagger-score resampling in the all-hadronic channel (Run 3), before (left) and after (right): Higgs candidate (top) and the two V candidates (middle, bottom).	103
Figure 4.11.	Soft-drop mass (top) and τ_{21} (bottom) of the Higgs (left) and the two V candidates, all-hadronic channel, Run 2.	105
Figure 4.12.	Tagging-jet m_{jj} (left) and $ \Delta\eta_{jj} $ (right), all-hadronic channel, Run 2.	105
Figure 4.13.	DNN score (top) and VBS BDT score (bottom), all-hadronic channel, Run 2 (left) and Run 3 (right).	107
Figure 4.14.	Final-state topology of the single-lepton fully-boosted channel.	108
Figure 4.15.	Minimum lepton- b -jet mass, single-lepton fully-boosted channel	110
Figure 4.16.	DNN (top) and VBS BDT (bottom) scores, single-lepton fully-boosted channel, Run 2 (left) and Run 3 (right).	113
Figure 4.17.	Final-state topology of the single-lepton semi-boosted channel.	114

Figure 4.18.	DNN (top) and VBS BDT (bottom) scores, single-lepton semi-boosted channel, Run 2 (left) and Run 3 (right).....	118
Figure 4.19.	Observed and expected 95% CL upper limits on $\mu = \sigma/\sigma(\kappa_{2V} = 1.5)$ for each channel and era and for the combination of all six fits. The dotted vertical line marks $\mu = 1$, the rate of the $\kappa_{2V} = 1.5$ benchmark.	127
Figure 4.20.	Observed and expected 95% CL upper limit on the VBS VVH production cross section as a function of κ_{2V} , from the Run 3 dataset. The blue curve is the theory prediction; the values of κ_{2V} at which the observed limit crosses it delimit the excluded region.	129
Figure 5.1.	Construction of a LST T3 and underlying objects in the Phase-2 Outer Tracker.	132
Figure 5.2.	Learned condensation charge β for true and fake triplets (left) and its ROC as a true/fake classifier (right).	140
Figure 5.3.	Architecture of the Transformer encoder.	141
Figure 5.4.	The ML track-reconstruction pipeline: LST triplets embedded by a Transformer under an object-condensation loss, then clustered with DBSCAN. .	143
Figure 5.5.	Number of Outer Tracker layers N_L spanned by the reconstructed candidates, peaking at six versus the five-layer T5 baseline.	144
Figure 5.6.	Tracking efficiency versus η (left) and p_T (right) for the Transformer + object-condensation candidates and the LST T5 baseline, under the matched $N_L \geq 5$ requirement.	144
Figure 5.7.	Fake rate (top) and duplicate rate (bottom) versus η (left) and p_T (right) for the Transformer + object-condensation candidates and the LST T5 baseline, under the matched $N_L \geq 5$ requirement.	145
Figure 5.8.	Tracking efficiency versus η (left) and p_T (right) for the Transformer + object-condensation candidates with the relaxed $N_L \geq 4$ requirement, compared with the same LST T5 baseline.	145
Figure 5.9.	Fake rate (top) and duplicate rate (bottom) versus η (left) and p_T (right) for the Transformer + object-condensation candidates with the relaxed $N_L \geq 4$ requirement, compared with the same LST T5 baseline.	146
Figure 6.1.	Projected CMS CPU (a) and disk (b) requirements for the HL-LHC, against a flat budget [3].	149

Figure 6.2.	Aggregate XRootD HTTPS-TPC throughput sustained at 400 Gbps across the UCSD-Caltech link.	152
Figure 6.3.	Throughput versus stream count and t_c -emulated latency, for combinations of CPU cores and XRootD origins. Adapted from [10].	155
Figure 6.4.	The variable-latency testbed: two co-located SDSC nodes interconnected through static routes across FABRIC, realizing round-trip times of 5-120 ms.	157
Figure 6.5.	Throughput under real latencies: (a) versus stream count for several round-trip times; (b) progress toward 100 Gbps as the number of XRootD origins grows at fixed total core count [9].	158
Figure 6.6.	Single-server throughput versus stream count at zero added latency, plateauing near 260 Gbps [9].	158
Figure 6.7.	The DMM transfer lifecycle.	163
Figure 6.8.	Example transfer topology for the priority-based bandwidth allocation: nodes are sites (with port capacity), edges are rules (with priority p and assigned bandwidth).	165
Figure 6.9.	The MIST testbed	167
Figure 6.10.	End-to-end performance on the MIST testbed: three prioritized transfers (Rules 1-3, 2:5:3) sharing a 100 Gbps path [67].	169
Figure A.1.	Transverse momentum (left), pseudorapidity (middle), and azimuth (right) of the boosted candidates, all-hadronic fully-boosted channel, Run 2.	172
Figure A.2.	Groomed mass, regressed mass, N -subjettiness ratio τ_{21} , and GLOPART tagger score (left to right) of the boosted candidates, all-hadronic fully-boosted channel, Run 2.	173
Figure A.3.	Tagging-jet system and missing transverse momentum, all-hadronic fully-boosted channel, Run 2: pair mass, rapidity gap and BDT score with p_T^{miss} (top), and the kinematics of the leading and subleading tagging jet (lower rows).	174
Figure A.4.	Transverse momentum (left), pseudorapidity (middle), and azimuth (right) of the boosted candidates, all-hadronic fully-boosted channel, Run 3.	175
Figure A.5.	Groomed mass, regressed mass, N -subjettiness ratio τ_{21} , and GLOPART tagger score (left to right) of the boosted candidates, all-hadronic fully-boosted channel, Run 3.	176

Figure A.6.	Tagging-jet system and missing transverse momentum, all-hadronic fully-boosted channel, Run 3: pair mass, rapidity gap and BDT score with p_T^{miss} (top), and the kinematics of the leading and subleading tagging jet (lower rows).	177
Figure A.7.	Transverse momentum (left), pseudorapidity (middle), and azimuth (right) of the boosted candidates, single-lepton fully-boosted channel, Run 2.	178
Figure A.8.	Groomed mass, regressed mass, N -subjettiness ratio τ_{21} , and GLOPART tagger score (left to right) of the boosted candidates, single-lepton fully-boosted channel, Run 2.	179
Figure A.9.	Tagging-jet system, single-lepton fully-boosted channel, Run 2: pair mass, rapidity gap and BDT score (top), then the kinematics of the leading (middle) and subleading (bottom) tagging jet.	180
Figure A.10.	Missing transverse momentum, $m_{\ell b}$, and lepton kinematics, single-lepton fully-boosted channel, Run 2.	181
Figure A.11.	Transverse momentum (left), pseudorapidity (middle), and azimuth (right) of the boosted candidates, single-lepton fully-boosted channel, Run 3.	182
Figure A.12.	Groomed mass, regressed mass, N -subjettiness ratio τ_{21} , and GLOPART tagger score (left to right) of the boosted candidates, single-lepton fully-boosted channel, Run 3.	183
Figure A.13.	Tagging-jet system, single-lepton fully-boosted channel, Run 3: pair mass, rapidity gap and BDT score (top), then the kinematics of the leading (middle) and subleading (bottom) tagging jet.	184
Figure A.14.	Missing transverse momentum, $m_{\ell b}$, and lepton kinematics, single-lepton fully-boosted channel, Run 3.	185
Figure A.15.	Transverse momentum (left), pseudorapidity (middle), and azimuth (right) of the boosted candidates, single-lepton semi-boosted channel, Run 2.	186
Figure A.16.	Groomed mass, regressed mass, N -subjettiness ratio τ_{21} , and GLOPART tagger score (left to right) of the boosted candidates, single-lepton semi-boosted channel, Run 2.	187
Figure A.17.	Tagging-jet system, single-lepton semi-boosted channel, Run 2: pair mass, rapidity gap and BDT score (top), then the kinematics of the leading (middle) and subleading (bottom) tagging jet.	188
Figure A.18.	Missing transverse momentum, $m_{\ell b}$, and lepton kinematics, single-lepton semi-boosted channel, Run 2.	189

Figure A.19.	Resolved system, single-lepton semi-boosted channel, Run 2: leading dijet mass, angular separation and p_T (top), and the mass residuals against the W , Z and Higgs hypotheses (bottom).	190
Figure A.20.	Transverse momentum (left), pseudorapidity (middle), and azimuth (right) of the boosted candidates, single-lepton semi-boosted channel, Run 3.	191
Figure A.21.	Groomed mass, regressed mass, N -subjettiness ratio τ_{21} , and GLOPART tagger score (left to right) of the boosted candidates, single-lepton semi-boosted channel, Run 3.	192
Figure A.22.	Tagging-jet system, single-lepton semi-boosted channel, Run 3: pair mass, rapidity gap and BDT score (top), then the kinematics of the leading (middle) and subleading (bottom) tagging jet.	193
Figure A.23.	Missing transverse momentum, $m_{\ell b}$, and lepton kinematics, single-lepton semi-boosted channel, Run 3.	194
Figure A.24.	Resolved system, single-lepton semi-boosted channel, Run 3: leading dijet mass, angular separation and p_T (top), and the mass residuals against the W , Z and Higgs hypotheses (bottom).	195

LIST OF TABLES

Table 1.1.	Gauge quantum numbers of the first-generation fermions; the other two generations are identical.	9
Table 1.2.	SM Higgs production cross sections at $\sqrt{s} = 13$ TeV ($m_H = 125$ GeV) and the coupling each mode probes [44].	24
Table 4.1.	Leading-order cross sections (fb) of the four VBS VVH sub-processes at $\sqrt{s} = 13.6$ TeV ($m_{jj} > 100$ GeV).	88
Table 4.2.	Hyperparameters of the VBS jet-pair selection BDT.	92
Table 4.3.	VBS jet-pair matching accuracy of the era-dedicated and combined BDTs, on Run 2 and Run 3 signal.	93
Table 4.4.	Summary of the analysis channels ("FJ" a fat jet, "X" extra AK4 or p_T^{miss} activity); the three detailed in this chapter are in bold.	94
Table 4.5.	Architecture and training hyperparameters of the two discriminants (per-channel ranges where they vary).	100
Table 4.6.	Cutflow for the all-hadronic fully-boosted channel, Run 2.	102
Table 4.7.	Cutflow for the all-hadronic fully-boosted channel, Run 3.	102
Table 4.8.	Signal-region definitions and yields for the all-hadronic fully-boosted channel, Run 2.	106
Table 4.9.	Signal-region definitions and yields for the all-hadronic fully-boosted channel, Run 3.	106
Table 4.10.	Cutflow for the single-lepton fully-boosted channel, Run 2.	110
Table 4.11.	Cutflow for the single-lepton fully-boosted channel, Run 3.	111
Table 4.12.	Signal-region definitions and yields for the single-lepton fully-boosted channel, Run 2.	112
Table 4.13.	Signal-region definitions and yields for the single-lepton fully-boosted channel, Run 3.	112
Table 4.14.	Cutflow for the single-lepton semi-boosted channel, Run 2.	115
Table 4.15.	Cutflow for the single-lepton semi-boosted channel, Run 3.	115

Table 4.16.	Signal-region definitions and yields for the single-lepton semi-boosted channel, Run 2.	117
Table 4.17.	Signal-region definitions and yields for the single-lepton semi-boosted channel, Run 3.	117
Table 4.18.	Observed and expected 95% CL upper limits on the signal strength $\mu = \sigma/\sigma(\kappa_{2V} = 1.5)$, per channel and era and for the combination of all three channels in both eras. The median expected values and the observed combined limit are set in bold.	126

ACKNOWLEDGEMENTS

The work presented in this dissertation is a product of years of guidance, patience, and friendship from more people than I can properly repay here. This is an attempt to do so anyway, and I hope it conveys the depth of my gratitude.

First, and most of all, I want to thank Mario, Leonardo, and Diego, who mentored me through every part of this. Mario taught me very nearly everything I know about tracking and the ins and outs of CMS and how an analysis is built from the ground up; his guidance has been invaluable, and he is the single biggest reason my time at CERN was as productive, and as meaningful, as it was. Leonardo was a constant presence in the analysis work, and I owe him for the countless hours he spent as the wall I bounced my many, mostly stupid, ideas against, always listening patiently and helping me refine them. Diego guided the computing work with that same generosity, and I am grateful for how steadily he showed up for me across this entire journey.

I am deeply grateful to my advisors, Frank for giving me the freedom to wander the vast phase space of CMS physics and computing, to enable all the opportunities I wanted to pursue, and the support to follow every detour I found along the way. And Avi for his continued guidance and attempts to make me a better scientist. Their mentorship shaped not only this research but the way I think about problems, and it carried me through the harder stretches of graduate school.

Thank you to my committee members, Professor Daniel R. Green and Professor Ramesh R. Rao, whose feedback and insight made this a better piece of work than it would otherwise have been. And thank you to Professor Javier Duarte, who acted as a secondary supervisor for the tracking work and was unfailingly generous with his time and expertise.

To the rest of my group at UCSD, Alex, Brunella, Chris, Jade, Jonathan, Manos, Slava, Terrence, Yanxi, and Yifan: thank you for the company, the conversations, and the quiet solidarity of people working through the same thing at the same time. It made all the difference.

This dissertation rests on the efforts of three teams, one behind each part of it: the LST group, whose work underpins Chapter 5; the SENSE-Rucio team, who accommodated my many questions and requests over the years; and the VBS *VVH* analysis group, whose dedication and

shared effort made the search of Chapter 4 possible. I am grateful to all of them.

To my friends and colleagues at UCSD, Elias, Francis, Harry, Jai, Jayke, Owen, Rolando, Simon, Yu-Hsueh, and the entire 2021 cohort: thank you for making the program feel survivable, and often genuinely fun. And to my friends at CERN, Christian, Hayden B., Hayden S., Heewon, Ian, and Nico, who turned that time into something I will carry with me for the rest of my life.

Finally, none of this would exist without my family. To my parents and my brother: you believed in me long before there was much evidence to justify it, and you never once wavered. Thank you for the unconditional love, for the constant encouragement, and for never making me feel I had to justify the path I chose. This dissertation is as much yours as it is mine.

Chapter 4, in part, is a reprint of the material as it appears in "Search for associated production of a Higgs boson and two vector bosons via vector-boson scattering at $\sqrt{s} = 13$ TeV," Physical Review Letters, 2026, A. Hayrapetyan *et al.* (CMS Collaboration) [5, 40]. It also contains material from an analysis of the combined Run 2 and Run 3 datasets that is currently being prepared for publication by the CMS Collaboration.

Chapter 5, in part, is a reprint of the material as it appears in "Using Line Segment Tracking to enable ML-based Track Reconstruction in Phase-2 of the CMS Experiment," CMS Detector Performance Summary CMS-DP-2026-030, 2026, CMS Collaboration [41].

Chapter 6, in part, is a reprint of the material as it appears in "400 Gbps benchmark of XRootD HTTP-TPC," EPJ Web of Conferences, 2024 [10]; "Data Movement Manager (DMM) for the SENSE-Rucio Interoperation Prototype," EPJ Web of Conferences, 2025 [8]; "Benchmarking XRootD-HTTPS on 400 Gbps Links with Variable Latencies," EPJ Web of Conferences, 2025 [9]; and "Using Microsoft Azure for XRootD Network Benchmarking," 2021 [11]. The chapter also contains material from "Integrating End-to-End Exascale SDN into the LHC Data Distribution Cyberinfrastructure," PEARC, 2022 [54]; "Managed Network Services for Exascale Data Movement Across Large Global Scientific Collaborations," XLOOP, 2022 [86]; and "SENSE in Practice: Quantifying the End-to-End Benefits of Intent-Based Bandwidth Reservation for Exascale Science Workflows," SC Workshops, 2025 [67].

VITA

2017-2021 Bachelor of Science, University of California San Diego

2021-2026 Doctor of Philosophy, University of California San Diego

FIELDS OF STUDY

Major Field: Particle Physics

ABSTRACT OF THE DISSERTATION

Search for Anomalous Production of the Higgs Boson at the CMS Experiment and
Exascale Computing for the LHC

by

Aashay Arora

Doctor of Philosophy in Physics

University of California San Diego, 2026

Professor Frank K. Würthwein, Chair

This dissertation presents a search for anomalous production of the Higgs boson with the Compact Muon Solenoid (CMS) experiment, together with two developments in reconstruction and computing for the High-Luminosity LHC (HL-LHC).

The search targets vector boson scattering production of a Higgs boson with two vector bosons. It aims to constrain the quartic VVHH coupling, quoted relative to its Standard Model value: departures from unity grow the cross section quadratically and boost the final-state bosons. The search uses the full Run 2 dataset, 138 inverse femtobarns at a center-of-mass energy of 13 TeV and 283 inverse femtobarns of Run 3 data at 13.6 TeV, in three channels: an all-hadronic

state with three large-radius jets and two single-lepton states. The dominant multijet and top quark backgrounds are estimated from data with an ABCD method built on two decorrelated classifiers, and the signal is extracted from a simultaneous fit to several orthogonal signal regions per channel. The combination excludes a coupling of 1.5, with an observed 95 percent confidence level upper limit of 0.48 on that rate against an expectation of 0.64; the Run 3 data alone constrain the coupling between 0.612 and 1.387, improving on the published Run 2 result.

The second development addresses charged-particle tracking at the HL-LHC using a machine learning approach: the first demonstration in CMS of a transformer-based model trained with object condensation, built on Line Segment Tracking triplets to reconstruct full track candidates. The model reduces the fake rate by roughly a factor of three and all but eliminates duplicates, at an efficiency comparable to the algorithmic baseline.

The third concerns moving the resulting data. The HL-LHC will produce an unprecedented volume of data, so the wide-area network is no longer a best-effort service but a critical resource that must be managed. This work integrates the Rucio data management system with the SENSE software-defined networking service, so that bandwidth is allocated by transfer priority and reservations can be reconfigured mid-transfer without interrupting traffic in flight. It is demonstrated on a multi-domain testbed and in benchmarks at 400 Gbps.

Introduction

At its core, this dissertation is concerned with two of the most fundamental questions in the physical sciences: what are the elementary constituents of matter, and what are the laws that govern their interactions. Over the past century, the pursuit of these questions has produced the Standard Model of Particle Physics, a relativistic quantum field theory that describes the known elementary particles and three of the four fundamental forces with extraordinary precision. Yet the Standard Model is known to be incomplete: it offers no account of gravity, dark matter, or the matter-antimatter asymmetry of the universe, and it leaves the values of its own parameters unexplained.

Establishing where the Standard Model breaks down, and discovering the physics that lies beyond it, is the central goal of contemporary particle physics. The most direct experimental window onto these questions is the Large Hadron Collider (LHC) at CERN, which collides protons at the highest energies ever achieved in a laboratory. Among the discoveries enabled by the LHC, the observation of the Higgs boson in 2012 by the ATLAS and CMS experiments stands out: it confirmed the mechanism of electroweak symmetry breaking and completed the particle content of the Standard Model. The Higgs boson is, however, the most recently discovered and least explored sector of the theory. Because its couplings to the electroweak gauge bosons follow directly from the symmetry breaking mechanism, they are an especially sensitive probe of new physics: even small deviations from the Standard Model predictions would betray the presence of new interactions. This dissertation presents a search for anomalous production of the Higgs boson in vector boson scattering with an associated Higgs boson, a rare electroweak process being probed for the first time at the LHC, using data recorded by the Compact Muon

Solenoid (CMS) experiment. Any modification of the electroweak couplings would manifest as an anomalous enhancement of this process, making it a direct and largely model-independent test of electroweak symmetry breaking. Extracting such a measurement from the LHC is inseparable from the computational enterprise that supports it.

The CMS experiment records and reconstructs data at a scale that places it among the most demanding scientific workflows in the world, and the forthcoming High-Luminosity LHC will increase both the data volume and the event complexity by an order of magnitude. Meeting this challenge requires advances on two fronts that form the second thread of this dissertation. The first is the movement of exascale data volumes across wide-area networks: we present a network-aware data-movement system that couples software-defined networking to the experiment's data-management infrastructure, together with benchmarks of the data plane at the 400 Gbps frontier. The second is the reconstruction of charged-particle trajectories at the extreme particle densities expected after the Phase-2 detector upgrade, where we explore machine-learning approaches that incorporate physics knowledge to improve tracking performance. These contributions are not prerequisites for the physics measurement but parallel efforts that, together, sustain the discovery potential of the experiment at the energy and intensity frontier.

The remainder of this dissertation is organized as follows. We begin by reviewing the Standard Model of Particle Physics, developing the theory from first principles and emphasizing the role of experiment in determining its parameters. We then describe the LHC and the CMS detector, the interactions of particles with matter that underpin its design, and the simulation and reconstruction chain that turns detector signals into physics objects. Building on this foundation, we present the search for anomalous Higgs boson production that motivates the entire enterprise. We then turn to the two supporting efforts, track reconstruction for the Phase-2 upgrade and the exascale computing infrastructure that the experiment depends upon. Each of these three chapters closes with a summary of its results and an outlook on the road ahead.

Chapter 1

Standard Model of Particle Physics

The Standard Model of Particle Physics (SM) is a theoretical framework that describes the known elementary particles and three of the four fundamental interactions between them: the electromagnetic, weak, and strong forces. The theory is built upon a single organizing idea: that the dynamics and interactions of the fundamental fields are dictated by local gauge symmetry, encoded in the gauge group $SU(3)_C \times SU(2)_L \times U(1)_Y$. From this principle, together with the spontaneous breaking of the electroweak symmetry, the entire spectrum of particles shown in Figure 1.1 and their interactions follow. The SM is among the most precisely tested theories in physics, with predictions confirmed by experiment across many orders of magnitude in energy. Its final cornerstone, the Higgs boson, was discovered in 2012 by the ATLAS and CMS experiments at the Large Hadron Collider (LHC) [12, 30].

This chapter develops the theory from first principles. We begin with classical field theory and the gauge principle (Section 1.1) and assemble the compact Standard Model Lagrangian (Section 1.2), whose three pieces organize the rest of the chapter: the gauge sector (Section 1.3), the Higgs mechanism through which the electroweak symmetry is spontaneously broken (Section 1.4), and the flavor sector that generates the fermion masses and the mixing between generations (Section 1.5). We then collect the parameters of the SM that are *not* predicted by the theory and must instead be measured (Section 1.6), emphasizing the central role of experiment. We describe how theoretical predictions are connected to measurable quantities

through scattering amplitudes and cross sections (Section 1.7), examine the Higgs boson itself, its couplings, production, and decay, together with the current experimental picture (Section 1.8), and we close with the known limitations of the SM that motivate the search for new physics (Section 1.9).

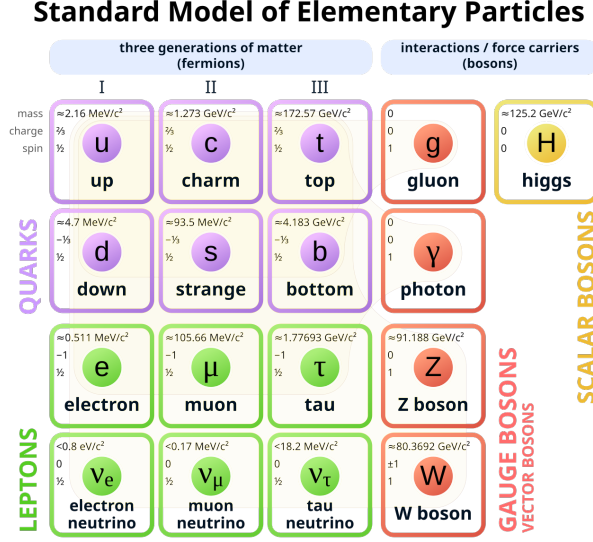


Figure 1.1. The particles of the Standard Model, categorized into fermions (quarks and leptons), gauge bosons (force carriers), and the Higgs boson.

1.1 Classical Field Theory and the Gauge Principle

The Lagrangian Formalism

The dynamics of a classical field theory are encoded in a Lagrangian density $\mathcal{L}$, a local function of a set of fields $\phi_a(x)$ and their spacetime derivatives $\partial_\mu \phi_a(x)$. The action is the spacetime integral of the Lagrangian density,

$$S = \int d^4x \mathcal{L}(\phi_a, \partial_\mu \phi_a). \quad (1.1)$$

The physical field configurations are those that render the action stationary under variations $\phi_a \rightarrow \phi_a + \delta\phi_a$ that vanish on the boundary. Imposing $\delta S = 0$ yields the Euler-Lagrange

equations of motion,

$$\partial_\mu \left(\frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi_a)} \right) - \frac{\partial \mathcal{L}}{\partial \phi_a} = 0. \quad (1.2)$$

This single principle generates the field equations of the entire SM once the appropriate $\mathcal{L}$ is specified. Two building blocks recur throughout. A free real scalar field of mass m is described by the Klein-Gordon Lagrangian,

$$\mathcal{L}_{\text{KG}} = \frac{1}{2} \partial_\mu \phi \partial^\mu \phi - \frac{1}{2} m^2 \phi^2, \quad (1.3)$$

which through Eq. (1.2) gives the Klein-Gordon equation $(\partial_\mu \partial^\mu + m^2)\phi = 0$. A free spin- $\frac{1}{2}$ fermion of mass m is described by the Dirac Lagrangian,

$$\mathcal{L}_{\text{Dirac}} = \bar{\psi} (i\gamma^\mu \partial_\mu - m) \psi, \quad (1.4)$$

where ψ is a four-component spinor, $\bar{\psi} = \psi^\dagger \gamma^0$, and the γ^μ are the Dirac matrices satisfying the Clifford algebra $\{\gamma^\mu, \gamma^\nu\} = 2g^{\mu\nu}$. The equation of motion is the Dirac equation $(i\gamma^\mu \partial_\mu - m)\psi = 0$.

Noether's Theorem

The deep connection between symmetry and conservation laws is established by Noether's theorem [69]: every continuous symmetry of the action corresponds to a conserved current. If the Lagrangian is invariant under an infinitesimal transformation $\phi_a \rightarrow \phi_a + \alpha \Delta \phi_a$, then the current

$$j^\mu = \frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi_a)} \Delta \phi_a \quad (1.5)$$

is conserved, $\partial_\mu j^\mu = 0$, implying a conserved charge $Q = \int d^3x j^0$. Invariance under spacetime translations yields the conserved energy-momentum tensor, while invariance under internal symmetries yields conserved charges such as electric charge. This theorem is the reason

symmetries occupy the central role in the construction of the SM.

From Global to Local Symmetry: The Gauge Principle

Consider the Dirac Lagrangian of Eq. (1.4). It is invariant under the *global* phase rotation $\psi \rightarrow e^{i\alpha} \psi$ with α a constant, whose Noether current $j^\mu = \bar{\psi} \gamma^\mu \psi$ describes a conserved charge. The gauge principle elevates this global $U(1)$ symmetry to a *local* one, demanding invariance when $\alpha = \alpha(x)$ depends on the spacetime point. The ordinary derivative then fails to transform covariantly, because $\partial_\mu (e^{i\alpha(x)} \psi)$ generates an extra term $i(\partial_\mu \alpha) \psi$. Invariance is restored by introducing a gauge field A_μ and replacing the ordinary derivative with the covariant derivative,

$$D_\mu = \partial_\mu + ieA_\mu, \quad A_\mu \rightarrow A_\mu - \frac{1}{e} \partial_\mu \alpha, \quad (1.6)$$

so that $D_\mu \psi \rightarrow e^{i\alpha(x)} D_\mu \psi$ transforms like ψ itself. Promoting $\partial_\mu \rightarrow D_\mu$ in the Dirac Lagrangian and adding a gauge-invariant kinetic term for the new field built from the field strength $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$ yields the Lagrangian of quantum electrodynamics (QED),

$$\mathcal{L}_{\text{QED}} = \bar{\psi} (i\gamma^\mu D_\mu - m) \psi - \frac{1}{4} F_{\mu\nu} F^{\mu\nu} = \bar{\psi} (i\gamma^\mu \partial_\mu - m) \psi - e \bar{\psi} \gamma^\mu \psi A_\mu - \frac{1}{4} F_{\mu\nu} F^{\mu\nu}. \quad (1.7)$$

Remarkably, the interaction between the electron and the photon A_μ , together with electric charge conservation, is *dictated* by the demand of local gauge invariance rather than being inserted by hand. The massless photon arises because a mass term $\frac{1}{2} m_A^2 A_\mu A^\mu$ would violate gauge invariance; this will have important consequences for the weak bosons.

Non-Abelian Gauge Theories

The SM strong and weak sectors generalize this construction to non-Abelian gauge groups, following Yang and Mills. For a gauge group with Hermitian generators T^a obeying the Lie algebra $[T^a, T^b] = if^{abc} T^c$, where f^{abc} are the structure constants, the fermion fields

transform in some representation and the covariant derivative becomes

$$D_\mu = \partial_\mu - ig A_\mu^a T^a, \quad (1.8)$$

with one gauge field A_μ^a for each generator. The non-Abelian field strength acquires a self-interaction term absent in QED,

$$F_{\mu\nu}^a = \partial_\mu A_\nu^a - \partial_\nu A_\mu^a + gf^{abc} A_\mu^b A_\nu^c, \quad (1.9)$$

and the Yang-Mills Lagrangian is $\mathcal{L}_{\text{YM}} = -\frac{1}{4} F_{\mu\nu}^a F^{a\mu\nu}$. The cubic and quartic terms hidden in $F_{\mu\nu}^a$ mean that the gauge bosons of a non-Abelian theory carry charge and interact with one another, gluons couple to gluons, and the weak bosons couple to themselves.

1.2 The Standard Model Lagrangian

Before constructing each piece in detail it is useful to see the whole. Once the gauge group and the particle content are fixed, gauge invariance and renormalizability permit only a small number of terms, and the entire Standard Model is summarized by the compact Lagrangian

$$\mathcal{L}_{\text{SM}} = \underbrace{-\frac{1}{4} F_{\mu\nu}^a F^{a\mu\nu} + i \bar{\psi} \gamma^\mu D_\mu \psi}_{\text{gauge sector}} + \underbrace{(D_\mu H)^\dagger (D^\mu H) - V(H)}_{\text{EWSB sector}} - \underbrace{(\bar{\psi}_L Y \psi_R H + \text{h.c.})}_{\text{flavor sector}}. \quad (1.10)$$

The three braces organize the remainder of this chapter. The *gauge sector* contains the kinetic terms and self-interactions of the gauge fields, $-\frac{1}{4} F_{\mu\nu}^a F^{a\mu\nu}$, together with the fermion kinetic terms whose covariant derivative D_μ dictates how matter couples to the force carriers (Section 1.3). The *electroweak symmetry breaking sector* comprises the Higgs doublet's kinetic term and the scalar potential $V(H)$, which on acquiring a vacuum expectation value gives mass to the weak bosons (Section 1.4). The *flavor sector* is the set of Yukawa couplings Y between the fermions and the Higgs field, which generate the fermion masses and the mixing between

generations (Section 1.5). The generation, gauge, and Lorentz indices and the sums over fermion species are suppressed here for clarity; each term is made explicit in the sections that follow.

1.3 The Gauge Sector of the Standard Model

The SM is the gauge theory built on

$$SU(3)_C \times SU(2)_L \times U(1)_Y, \quad (1.11)$$

where $SU(3)_C$ governs the strong interaction acting on color charge C , $SU(2)_L$ acts on left-handed weak isospin L , and $U(1)_Y$ acts on weak hypercharge Y . The corresponding gauge fields are the eight gluons G_μ^a ($a = 1, \dots, 8$) with coupling g_s , the three weak bosons W_μ^i ($i = 1, 2, 3$) with coupling g , and the hypercharge boson B_μ with coupling g' .

Fermion Content and Representations

The matter content consists of three generations of quarks and leptons. A defining feature of the electroweak sector is that it is *chiral*: left- and right-handed fermions, projected out by $\psi_{L,R} = \frac{1}{2}(1 \mp \gamma^5)\psi$, transform differently. The left-handed fields form $SU(2)_L$ doublets while the right-handed fields are $SU(2)_L$ singlets. For the first generation,

$$Q_L = \begin{pmatrix} u_L \\ d_L \end{pmatrix}, \quad L_L = \begin{pmatrix} \nu_{eL} \\ e_L \end{pmatrix}, \quad u_R, \quad d_R, \quad e_R, \quad (1.12)$$

with the second and third generations (c, s, ν_μ, μ) and (t, b, ν_τ, τ) being exact copies differing only in mass. The quarks additionally carry color and transform as triplets of $SU(3)_C$, while leptons are color singlets. The hypercharge Y of each multiplet is fixed by the Gell-Mann-Nishijima relation $Q = T^3 + Y$, where Q is the electric charge and T^3 is the third component of weak isospin. The quantum-number assignments are summarized in Table 1.1.

Table 1.1. Gauge quantum numbers of the first-generation fermions; the other two generations are identical.

Field	$SU(3)_C$	$SU(2)_L$	Y	Q
$Q_L = (u_L, d_L)$	3	2	$+\frac{1}{6}$	$(+\frac{2}{3}, -\frac{1}{3})$
u_R	3	1	$+\frac{2}{3}$	$+\frac{2}{3}$
d_R	3	1	$-\frac{1}{3}$	$-\frac{1}{3}$
$L_L = (\nu_{eL}, e_L)$	1	2	$-\frac{1}{2}$	$(0, -1)$
e_R	1	1	-1	-1

Quantum Chromodynamics

The strong interaction is described by quantum chromodynamics (QCD), the $SU(3)_C$ Yang-Mills theory coupled to the quarks. Its Lagrangian is

$$\mathcal{L}_{\text{QCD}} = -\frac{1}{4}G_{\mu\nu}^a G^{a\mu\nu} + \sum_q \bar{q} (i\gamma^\mu D_\mu - m_q) q, \quad D_\mu = \partial_\mu - ig_s G_\mu^a T^a, \quad (1.13)$$

where the T^a are the eight generators of $SU(3)$ (proportional to the Gell-Mann matrices) and the gluon field strength is $G_{\mu\nu}^a = \partial_\mu G_\nu^a - \partial_\nu G_\mu^a + g_s f^{abc} G_\mu^b G_\nu^c$. The gluon self-interaction leads to two hallmark features of QCD. Asymptotic freedom causes the coupling g_s to weaken at short distances, so that quarks behave as nearly free particles in high-energy collisions, while color confinement prevents the isolation of colored states, so that quarks and gluons are observed only as color-neutral hadrons and as collimated sprays of particles known as jets.

The Electroweak Lagrangian

The electromagnetic and weak interactions are unified into the $SU(2)_L \times U(1)_Y$ electroweak theory of Glashow, Weinberg, and Salam [52, 85, 77]. The gauge-invariant Lagrangian for the gauge fields and fermions is

$$\mathcal{L}_{\text{EW}} = -\frac{1}{4}W_{\mu\nu}^i W^{i\mu\nu} - \frac{1}{4}B_{\mu\nu} B^{\mu\nu} + \sum_f \bar{\psi}_f i\gamma^\mu D_\mu \psi_f, \quad (1.14)$$

where the sum runs over all left-handed doublets and right-handed singlets, and

$$W_{\mu\nu}^i = \partial_\mu W_\nu^i - \partial_\nu W_\mu^i + g \varepsilon^{ijk} W_\mu^j W_\nu^k, \quad (1.15)$$

$$B_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu. \quad (1.16)$$

The covariant derivative encodes how each field couples to the gauge bosons according to its weak isospin $T^i = \frac{1}{2}\sigma^i$ (the Pauli matrices, acting only on left-handed doublets) and hypercharge Y ,

$$D_\mu = \partial_\mu - ig W_\mu^i T^i - ig' B_\mu Y. \quad (1.17)$$

Crucially, because the theory is chiral, a Dirac mass term $m\bar{\psi}\psi = m(\bar{\psi}_L\psi_R + \bar{\psi}_R\psi_L)$ couples the left-handed doublet to the right-handed singlet and is therefore *not* gauge invariant. The same is true of explicit mass terms for the W and Z bosons. Yet the weak bosons are massive and the fermions clearly have mass. The resolution to this tension is the Higgs mechanism.

1.4 Electroweak Symmetry Breaking and the Higgs Mechanism

The Higgs Potential

The Brout-Englert-Higgs mechanism [47, 57, 55] introduces a complex scalar field that is a doublet of $SU(2)_L$ with hypercharge $Y = +\frac{1}{2}$,

$$H = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix}. \quad (1.18)$$

Its dynamics are governed by the gauge-invariant Lagrangian

$$\mathcal{L}_{\text{Higgs}} = (D_\mu H)^\dagger (D^\mu H) - V(H), \quad V(H) = \mu^2 H^\dagger H + \lambda (H^\dagger H)^2, \quad (1.19)$$

with D_μ given by Eq. (1.17). Stability of the potential requires $\lambda > 0$. The decisive choice is the sign of μ^2 . If $\mu^2 > 0$ the minimum sits at $H = 0$ and the symmetry is unbroken; but for $\mu^2 < 0$ the potential takes the "Mexican hat" shape as shown in Figure 1.2, and the minimum lies on a continuum of degenerate states at

$$H^\dagger H = -\frac{\mu^2}{2\lambda} \equiv \frac{v^2}{2}, \quad v = \sqrt{-\frac{\mu^2}{\lambda}}. \quad (1.20)$$

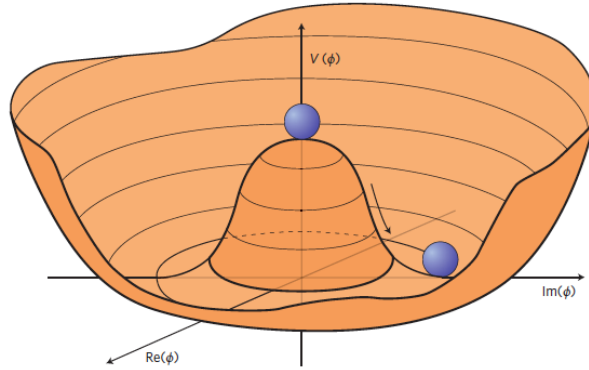


Figure 1.2. The Higgs potential $V(H)$ for $\mu^2 < 0$, with a continuum of degenerate minima at $H^\dagger H = v^2/2$.

Selecting one such vacuum spontaneously breaks the electroweak symmetry. By an $SU(2)_L$ gauge rotation the vacuum can be aligned along the neutral component,

$$\langle H \rangle = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v \end{pmatrix}, \quad (1.21)$$

which is invariant under the combination of generators corresponding to the electric charge $Q = T^3 + Y$. The pattern of breaking is therefore

$$SU(2)_L \times U(1)_Y \longrightarrow U(1)_{\text{EM}}, \quad (1.22)$$

leaving the photon massless while the three remaining generators are broken. The measured value of the vacuum expectation value (vev) is $v \approx 246 \text{ GeV}$, which sets the scale of

electroweak physics.

Gauge Boson Masses

Expanding the scalar kinetic term $(D_\mu H)^\dagger (D^\mu H)$ about the vacuum produces quadratic terms in the gauge fields, that is, mass terms. Working in unitary gauge, the doublet reduces to its neutral component, $\phi^0 = \frac{1}{\sqrt{2}}(v + H)$, where H is the physical Higgs boson, and the gauge-field part yields

$$\frac{1}{2} \frac{(v + H)^2}{4} \left[g^2 (W_\mu^1)^2 + g^2 (W_\mu^2)^2 + (g W_\mu^3 - g' B_\mu)^2 \right]. \quad (1.23)$$

Diagonalizing these terms defines the physical gauge-boson mass eigenstates. The charged weak bosons are $W_\mu^\pm = \frac{1}{\sqrt{2}}(W_\mu^1 \mp i W_\mu^2)$, and the neutral states mix through the weak mixing (Weinberg) angle θ_W , defined by $\tan \theta_W = g'/g$,

$$Z_\mu = \cos \theta_W W_\mu^3 - \sin \theta_W B_\mu, \quad (1.24)$$

$$A_\mu = \sin \theta_W W_\mu^3 + \cos \theta_W B_\mu. \quad (1.25)$$

The resulting masses are

$$m_W = \frac{1}{2} g v, \quad m_Z = \frac{1}{2} v \sqrt{g^2 + g'^2} = \frac{m_W}{\cos \theta_W}, \quad m_\gamma = 0. \quad (1.26)$$

The photon A_μ remains massless, as required for the unbroken $U(1)_{\text{EM}}$, and the relation $m_W = m_Z \cos \theta_W$ is a sharp, testable prediction of the theory. The terms linear in H encode the couplings of the Higgs boson to the weak bosons, proportional to their masses.

Fermion Masses and the Yukawa Sector

The same scalar doublet generates fermion masses through gauge-invariant Yukawa couplings. For a single generation,

$$\mathcal{L}_{\text{Yukawa}} = -y_d \bar{Q}_L H d_R - y_u \bar{Q}_L \tilde{H} u_R - y_e \bar{L}_L H e_R + \text{h.c.}, \quad (1.27)$$

where $\tilde{H} = i\sigma^2 H^*$ is the conjugate doublet (with $Y = -\frac{1}{2}$) needed to give mass to the up-type quarks. When H acquires its vev, each term produces a Dirac mass plus a coupling to the Higgs boson,

$$m_f = \frac{y_f v}{\sqrt{2}}, \quad \mathcal{L} \supset -\frac{m_f}{v} \bar{\psi}_f \psi_f H. \quad (1.28)$$

A profound consequence follows: the strength with which the Higgs boson couples to any fermion is proportional to that fermion's mass. Heavy particles such as the top quark couple strongly, while light particles couple weakly. This mass-proportional coupling structure is a defining and falsifiable prediction of the SM, and measuring it, particularly for the rarer and anomalous production and decay channels, is a central theme of the Higgs physics program and of this dissertation.

For three generations these Yukawa couplings become matrices in generation space, and diagonalizing them to reach the physical mass eigenstates induces mixing between the generations. This structure, which generates the fermion mass hierarchy and the quark and lepton mixing matrices, constitutes the flavor sector of the SM and is developed in Section 1.5.

After spontaneous symmetry breaking, the Higgs field itself acquires a mass by expanding $V(H)$ about the vacuum,

$$m_H = \sqrt{2\lambda} v = \sqrt{-2\mu^2}. \quad (1.29)$$

The particle predicted by this mechanism was observed in 2012 by the ATLAS and CMS collaborations with a mass of approximately 125 GeV [12, 30], completing the experimental verification of the SM.

1.5 The Flavor Sector of the Standard Model

The gauge and Higgs sectors treat the three generations of fermions identically; everything that distinguishes one generation from another, the pattern of masses and the mixing between families, resides in the Yukawa couplings introduced in Section 1.4. Promoting those couplings to three generations, they become complex 3×3 matrices Y_u , Y_d , and Y_e acting in generation space,

$$\mathcal{L}_{\text{Yukawa}} = -\bar{Q}_L^i(Y_d)_{ij}Hd_R^j - \bar{Q}_L^i(Y_u)_{ij}\tilde{H}u_R^j - \bar{L}_L^i(Y_e)_{ij}He_R^j + \text{h.c.}, \quad (1.30)$$

with $i, j = 1, 2, 3$ the generation indices. Gauge symmetry places no constraint on these matrices, so their entries are free; they account for the large majority of the parameters the SM cannot predict (Section 1.6).

Each Yukawa matrix can be brought to a real, positive, diagonal form by a biunitary transformation, $Y_f = U_f \hat{Y}_f W_f^\dagger$, that rotates the left- and right-handed fields independently. The eigenvalues fix the fermion masses, $m_f = y_f v / \sqrt{2}$, and define the *mass basis* of physical particles. The resulting spectrum is strikingly hierarchical, spanning nearly six orders of magnitude from the electron (0.5 MeV) to the top quark (173 GeV); the SM organizes this hierarchy but offers no explanation for it.

The flavor-rotations leave the flavor-universal gauge interactions diagonal, with one exception. In the charged-current weak interaction the up- and down-type quarks are rotated by different matrices, and the residual mismatch survives as a unitary mixing matrix, the Cabibbo-Kobayashi-Maskawa (CKM) matrix,

$$V_{\text{CKM}} = U_u^\dagger U_d = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix}, \quad (1.31)$$

which rotates the down-type mass eigenstates into the basis to which the W boson couples. After

absorbing unphysical phases, a three-generation V_{CKM} retains three mixing angles and a single irreducible complex phase. That phase is the sole source of CP violation in the quark sector of the SM, and is essential to, though quantitatively insufficient for, the observed matter-antimatter asymmetry of the universe. Experimentally V_{CKM} is close to the identity and strongly hierarchical, a structure made manifest by the Wolfenstein parametrization as an expansion in the Cabibbo parameter $\lambda = |V_{us}| \approx 0.225$.

The lepton sector mirrors the quark sector. In the minimal SM the neutrinos are massless and lepton flavor is exactly conserved, but the observation of neutrino oscillations establishes that neutrinos are massive and that the leptonic charged current carries an analogous mixing matrix, the Pontecorvo-Maki-Nakagawa-Sakata (PMNS) matrix. In contrast to the small quark mixing angles, the leptonic angles are large, and whether the neutrinos are Dirac or Majorana particles, the latter introducing additional CP-violating phases, remains unknown. The accommodation of neutrino masses adds the corresponding parameters to those of Section 1.6. Taken together, the unexplained hierarchy of fermion masses and the disparate pattern of quark and lepton mixing constitute the *flavor puzzle*, one of the open questions that motivate the search for new physics (Section 1.9).

1.6 The Free Parameters of the Standard Model

A striking feature of the SM is that, despite its predictive power, it does not determine its own fundamental constants. The structure of the theory, its gauge group, particle content, and interactions, is fixed by symmetry, but the numerical values of its parameters are not predicted and must be supplied by measurement. With massless neutrinos, the SM contains 19 free parameters [71, 72]:

- the three gauge couplings g_s , g , and g'
- the two parameters of the Higgs potential μ^2 and λ , equivalently the vev v and the Higgs mass m_H ;

- the nine charged-fermion masses: six quarks and three charged leptons;
- the four parameters of the CKM matrix, three mixing angles and one CP-violating phase;
- the QCD vacuum angle θ_{QCD} , which is experimentally constrained to be extremely small.

The observation of neutrino oscillations establishes that neutrinos are in fact massive: accommodating this adds three neutrino masses together with the three mixing angles and one CP-violating phase of the PMNS matrix, raising the total to at least 26, and to 28 if the neutrinos are Majorana particles and two further CP-violating phases are present.

That every one of these quantities must be measured rather than derived underscores the indispensable role of experiment. The theory is a scaffold; only experiment fixes the values that the scaffold supports, and only experiment can test whether the relations the theory predicts between them, such as $m_W = m_Z \cos \theta_W$, or the mass-proportionality of Higgs couplings, hold to the precision now attainable. Each measured parameter is simultaneously an input that the SM cannot explain and a probe through which deviations from the SM might first appear. It is precisely in the high-precision and rare-process frontier, where small discrepancies could betray the presence of new physics, that the experimental program of this dissertation operates.

1.7 From Amplitudes to Observables

The bridge between the Lagrangian of the SM and the quantities measured at a collider is the scattering cross section. This section builds the formalism from its classical roots up to the relativistic quantum field theory used for precision predictions, and then to the observable event rates at a hadron collider.

Quantum Field Theory Formulation

In the relativistic regime of high-energy physics, particle numbers are not conserved, particles can be created or annihilated, requiring the framework of quantum field theory (QFT).

The central object is the S -matrix (scattering matrix), which maps initial asymptotic free states $|i\rangle$ at $t \rightarrow -\infty$ to final asymptotic free states $|f\rangle$ at $t \rightarrow +\infty$,

$$|f\rangle_{\text{out}} = S|i\rangle_{\text{in}} \Rightarrow S_{fi} = \langle f|S|i\rangle. \quad (1.32)$$

In the interaction picture, the S -matrix is expressed using Dyson's formula as a time-ordered exponential of the interaction Hamiltonian density V_I ,

$$S = \mathcal{T} \exp \left(-i \int_{t_i}^t dt' V_I(t') \right). \quad (1.33)$$

Expanding this exponential generates the perturbation series; each term, when evaluated using Wick's theorem, corresponds to a set of Feynman diagrams in which internal lines are propagators and vertices are read off directly from the interaction terms of the SM Lagrangian.

For specific processes it is conventional to separate the non-interacting part from the interactions, writing $S = 1 + iT$, where the 1 represents no scattering. The non-trivial part T_{fi} contains a delta function enforcing overall four-momentum conservation,

$$\langle f|iT|i\rangle = (2\pi)^4 \delta^{(4)}(P_f - P_i) i\mathcal{M}_{fi}, \quad (1.34)$$

where P_i and P_f are the total initial and final four-momenta and $\mathcal{M}_{fi}$ is the invariant scattering matrix element, or Feynman amplitude [72]. The amplitude $\mathcal{M}_{fi}$ is computed systematically using Feynman rules derived from the path-integral formulation. The Lehmann-Symanzik-Zimmermann (LSZ) formula relates S -matrix elements to the time-ordered correlation functions of the interacting fields; for a scalar field $\phi(x)$, the scattering of n incoming particles into m outgoing particles is related to the $(n+m)$ -point Green's function. Because the SM is a weakly coupled theory over much of its range, $\mathcal{M}$ is expanded order by order in the couplings, with each successive order suppressed by additional powers of the coupling and represented by diagrams with more loops.

Cross Sections and Decay Rates

Once $\mathcal{M}_{fi}$ is known, observable rates follow from integrating the squared amplitude over the available phase space [72]. The n -body Lorentz-invariant phase space for a final state $1 + 2 + \dots + n$ of total four-momentum P is

$$\int d\Pi_n = \int \prod_{j=1}^n \frac{d^3 p_j}{(2\pi)^3 2E_j} (2\pi)^4 \delta^{(4)}\left(P - \sum_{j=1}^n p_j\right), \quad (1.35)$$

a manifestly Lorentz-invariant volume of final-state momenta constrained by four-momentum conservation. Combined with the squared invariant matrix element, summed over final-state spins and averaged over initial-state spins, it yields two master formulae. For an unstable particle A of mass M_A decaying to a final state f , the partial width is

$$\Gamma(A \rightarrow f) = \frac{1}{2M_A} \int d\Pi_n |\mathcal{M}(A \rightarrow f)|^2. \quad (1.36)$$

Summing the partial widths over all accessible final states yields the total width $\Gamma_A = \sum_f \Gamma(A \rightarrow f)$, whose inverse is the particle's lifetime $\tau_A = 1/\Gamma_A$. The branching ratio $\mathcal{B}_f = \Gamma(A \rightarrow f)/\Gamma_A$ gives the probability of decaying through a particular channel; for the Higgs boson, the set of branching ratios into $b\bar{b}$, WW , ZZ , $\tau\tau$, $\gamma\gamma$, and other final states encodes the mass-proportional couplings derived in Section 1.4 and is a primary target of experimental measurement.

For a collision $A + B \rightarrow f$, the cross section is

$$\sigma(A + B \rightarrow f) = \frac{1}{2E_A 2E_B |v_A - v_B|} \int d\Pi_n |\mathcal{M}(A + B \rightarrow f)|^2, \quad (1.37)$$

where the prefactor $2E_A 2E_B |v_A - v_B|$ is the relativistic flux of the incoming particles, with v_A and v_B their velocities along the beam axis.

Most processes of interest have two particles in the final state, for which the phase space simplifies considerably. Working in the center-of-mass (CM) frame and writing the kinematics

in terms of the Lorentz-invariant Mandelstam variables

$$s = (p_A + p_B)^2, \quad t = (p_A - p_1)^2, \quad u = (p_A - p_2)^2, \quad s + t + u = \sum_i m_i^2, \quad (1.38)$$

with $E_{\text{CM}} = \sqrt{s}$ the total CM energy, the two-body phase space reduces to

$$\int d\Pi_2 = \frac{1}{8\pi} \left(\frac{2|\vec{p}_f|}{E_{\text{CM}}} \right) \int \frac{d\Omega}{4\pi}, \quad (1.39)$$

where $|\vec{p}_f|$ is the momentum of either outgoing particle and reduces to $E_{\text{CM}}/2$ in the massless limit. The differential cross section then takes the compact form

$$\left(\frac{d\sigma}{d\Omega} \right)_{\text{CM}} = \frac{1}{64\pi^2 s} \frac{|\vec{p}_f|}{|\vec{p}_i|} |\mathcal{M}_{fi}|^2, \quad (1.40)$$

which for all four particles effectively massless ($|\vec{p}_f| = |\vec{p}_i|$) simplifies to

$$d\sigma/d\Omega = |\mathcal{M}_{fi}|^2 / 64\pi^2 s \quad (1.41)$$

Observables at a Hadron Collider

At the LHC the colliding objects are protons, which are composite. The short-distance scattering occurs between their constituent partons (quarks and gluons), each carrying a fraction x of the proton momentum. By the factorization theorem, the observable hadronic cross section is obtained by convolving the perturbatively calculable partonic cross section $\hat{\sigma}_{ij}$ with the parton distribution functions (PDFs) $f_i(x, \mu_F)$, which give the probability of finding parton i with momentum fraction x at a factorization scale μ_F [72],

$$\sigma(pp \rightarrow X) = \sum_{i,j} \int dx_1 dx_2 f_i(x_1, \mu_F) f_j(x_2, \mu_F) \hat{\sigma}_{ij}(\hat{s}, \mu_F), \quad (1.42)$$

where $\hat{s} = x_1 x_2 s$ is the partonic center-of-mass energy squared, related to the hadronic invariant s of Eq. (1.38) by the parton momentum fractions.

Because an s -channel process at leading order depends on the momentum fractions only through the product $x_1 x_2$, it is convenient to factor the hadronic cross section into a parton luminosity and a partonic cross section [78]. Defining the parton luminosity for initial-state partons i and j ,

$$\mathcal{L}_{ij}(\hat{s}) = \int dx_1 dx_2 f_i(x_1, \mu_F) f_j(x_2, \mu_F) \delta(x_1 x_2 s - \hat{s}), \quad (1.43)$$

the cross section becomes $\sigma = \sum_{i,j} \int d\hat{s} \mathcal{L}_{ij}(\hat{s}) \hat{\sigma}_{ij}(\hat{s})$. The parton luminosity cleanly separates the collider and PDF dependence from the short-distance dynamics, and comparing luminosities at different $\hat{s}$ quantifies the relative importance of the various initial-state channels that feed a given final state.

The PDFs are non-perturbative and extracted from experimental data, while their evolution with scale is predicted by QCD. Because the strong and electroweak couplings "run" with energy scale according to the renormalization group, the values measured in one process can be evolved to predict another, providing stringent consistency tests of the theory.

Finally, the connection to what is actually recorded in a detector is the expected number of events,

$$N = \sigma \int \mathcal{L} dt = \sigma \mathcal{L}_{\text{int}}, \quad (1.44)$$

where $\mathcal{L}$ is the instantaneous luminosity delivered by the collider and $\mathcal{L}_{\text{int}}$ is the time-integrated luminosity. A cross section computed from the SM Lagrangian thus translates directly into a predicted event yield, which experiment confronts with the observed count, rendering even the rarest processes, where new physics is most likely to hide, quantitatively testable.

Resonances and the Narrow-Width Approximation

Many of the most important processes at a collider proceed through the production and subsequent decay of an unstable intermediate particle, such as $q\bar{q} \rightarrow Z \rightarrow \ell^+ \ell^-$ or the Higgs boson channels studied in this thesis [78]. Near the mass of the intermediate state the amplitude

is dominated by its propagator and takes the relativistic Breit-Wigner form

$$\frac{d\hat{\sigma}}{d\hat{s}} \propto \frac{1}{(\hat{s} - m_R^2)^2 + m_R^2 \Gamma_R^2}, \quad (1.45)$$

peaked at $\hat{s} = m_R^2$ with a spread set by the total width Γ_R of the resonance. Plotting the rate as a function of the invariant mass of the decay products therefore reveals a resonant peak at the mass of the intermediate particle, the basic signature by which the Z , W , and Higgs bosons are identified.

When the width is small compared to the mass, $\Gamma_R \ll m_R$, as holds for the electroweak bosons (for example $\Gamma_Z = 2.5$ GeV against $m_Z = 91$ GeV), the Breit-Wigner distribution collapses toward a delta function at $\hat{s} = m_R^2$, and the cross section factorizes into independent production and decay steps,

$$\sigma(ab \rightarrow R \rightarrow f) = \sigma(ab \rightarrow R) \mathcal{B}(R \rightarrow f), \quad (1.46)$$

with the branching ratio $\mathcal{B}(R \rightarrow f) = \Gamma(R \rightarrow f)/\Gamma_R$ as defined following Eq. (1.36). This narrow-width approximation, which neglects the quantum interference between the production and decay subprocesses, underlies the standard treatment of essentially all resonant signals, including the factorized signal models used in the analysis of this thesis.

1.8 The Higgs Boson

The physical Higgs boson H is the single neutral scalar excitation that remains after electroweak symmetry breaking (Section 1.4), the others having been absorbed as the longitudinal polarizations of the W and Z . It is a spin-zero, CP-even state, $J^{PC} = 0^{++}$, a hypothesis now firmly established by the measured angular distributions of its decays [13, 34]. Because it is the particle corresponding to the field responsible for the masses of every other massive particle, its interactions are not free parameters but are fixed by those masses, making the Higgs boson a uniquely predictive probe of the SM and a sensitive window onto physics beyond it. This section

collects the couplings that follow from the Lagrangians of Section 1.4, describes how the Higgs boson is produced and how it decays at the LHC, and summarizes the current experimental picture.

Couplings of the Higgs Boson

Expanding the electroweak and Yukawa Lagrangians about the vacuum, with the neutral component in unitary gauge $\phi^0 = \frac{1}{\sqrt{2}}(v + H)$, collects all interactions of the physical field H . The gauge couplings arise from the scalar kinetic term $(D_\mu H)^\dagger (D^\mu H)$,

$$\mathcal{L}_{HVV} = \frac{2m_W^2}{v} H W_\mu^+ W^{-\mu} + \frac{m_Z^2}{v} H Z_\mu Z^\mu + \frac{m_W^2}{v^2} H^2 W_\mu^+ W^{-\mu} + \frac{m_Z^2}{2v^2} H^2 Z_\mu Z^\mu, \quad (1.47)$$

comprising both single-Higgs (HVV) and quartic ($HHVV$) vertices. The couplings to fermions come from the Yukawa terms of Section 1.4,

$$\mathcal{L}_{Hff} = -\sum_f \frac{m_f}{v} H \bar{f} f, \quad (1.48)$$

and the Higgs self-interactions follow from the scalar potential $V(H)$,

$$V \supset \frac{m_H^2}{2v} H^3 + \frac{m_H^2}{8v^2} H^4, \quad (1.49)$$

with the trilinear and quartic self-couplings fixed by m_H and v . The unifying feature is that every Higgs coupling is set by the mass of the particle it couples to: linearly in the mass for fermions, and quadratically (through m_V^2) for the weak bosons. This rigid structure is the central prediction tested by Higgs measurements. Deviations are parameterized in the so-called κ -framework by multiplicative coupling modifiers $\kappa_i = g_i/g_i^{\text{SM}}$, defined such that the SM corresponds to $\kappa_i = 1$ for every coupling; the modifiers to the vector-boson and fermion couplings, and in particular the quartic $HHVV$ couplings probed by vector-boson scattering, are

the parameters of interest in the analysis presented in this thesis. Notably, the kappa framework is not free of theoretical limitations, a more detailed discussion of which is given in Section 4.2.1.

Production at the LHC

At the LHC the Higgs boson is produced in four dominant modes, each governed by one of the couplings above and illustrated in Figure 1.3. Gluon-gluon fusion (ggF), proceeding through a virtual top-quark loop, has by far the largest cross section. Vector-boson fusion (VBF), in which two quarks each radiate a weak boson that fuse into a Higgs, leaves two forward "tag" jets. Associated production with a weak boson (VH , i.e. WH and ZH) and with a top-quark pair ($t\bar{t}H$) directly expose the HVV and top-Yukawa couplings, respectively. The predicted cross sections at $\sqrt{s} = 13$ TeV for $m_H = 125$ GeV are summarized in Table 1.2.

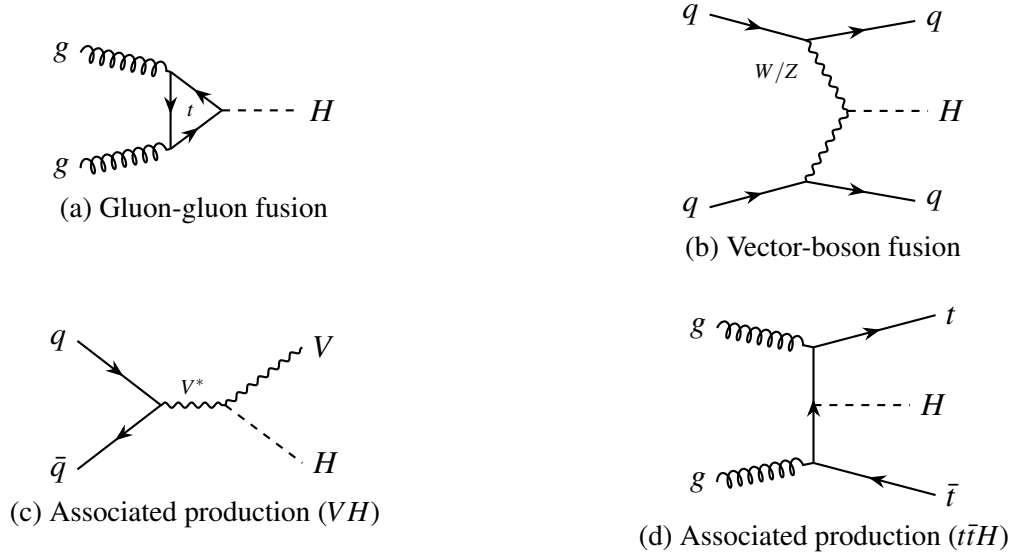


Figure 1.3. Leading-order Feynman diagrams for the dominant Higgs boson production mechanisms at the LHC.

Decay Modes and Branching Ratios

Because its couplings scale with mass, the Higgs boson decays preferentially to the heaviest states that are kinematically accessible. For $m_H = 125$ GeV the largest branching ratio is to $b\bar{b}$, followed by the off-shell WW^* and ZZ^* channels and the $\tau^+\tau^-$ mode; the loop-induced

Table 1.2. SM Higgs production cross sections at $\sqrt{s} = 13$ TeV ($m_H = 125$ GeV) and the coupling each mode probes [44].

Production mode	Coupling probed	σ at 13 TeV [pb]
Gluon-gluon fusion (ggF)	top-quark loop ($Ht\bar{t}$)	48.6
Vector-boson fusion (VBF)	HVV	3.78
Associated WH	HWW	1.37
Associated ZH	HZZ	0.88
Associated $t\bar{t}H$	top Yukawa	0.51

$\gamma\gamma$ decay, though rare, provides one of the cleanest experimental signatures. The total width predicted by the SM is only $\Gamma_H \approx 4.1$ MeV, far smaller than the experimental mass resolution, so that the narrow-width approximation of Eq. (1.46) applies and production and decay factorize cleanly.

Experimental Status

Since the discovery, the properties of the Higgs boson have been measured with steadily increasing precision and found to be remarkably consistent with the SM. Its mass is now known to better than one part in a thousand, $m_H = 125.20 \pm 0.11$ GeV [71], and its spin-parity is established as 0^+ . The overall production and decay rate, expressed through the signal strength $\mu = \sigma/\sigma_{\text{SM}}$, is compatible with unity at the few-percent level in the Run 2 combinations of both experiments [13, 34]. The Higgs boson has been observed in the $\gamma\gamma$, ZZ^* , WW^* , $\tau^+\tau^-$, and $b\bar{b}$ channels, with evidence for the rare $\mu^+\mu^-$ decay, and the coupling modifiers κ_i extracted from a global fit agree with the SM within their uncertainties, of order 10%. Crucially, when the measured couplings are plotted against the corresponding particle masses they fall on the linear (for fermions) and quadratic (for bosons) relations predicted by Eqs. (1.47) and (1.48), spanning three orders of magnitude in mass from the muon to the top quark (Figure 1.4). This agreement is among the most striking confirmations of the SM, even as the couplings most weakly constrained, the self-coupling and the rarer and anomalous production modes, remain open and are precisely where new physics could still hide.

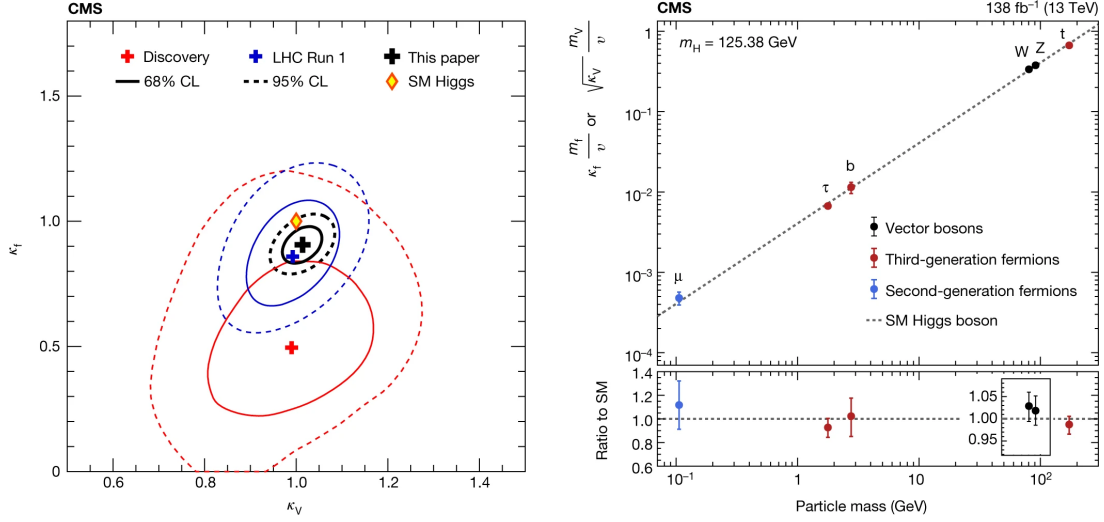


Figure 1.4. Measured Higgs coupling strengths versus particle mass, compared with the SM mass-proportional prediction [34].

1.9 Limitations of the Standard Model and Physics Beyond

Despite its extraordinary success, the SM is known to be incomplete. It does not incorporate gravity, it provides no candidate for the dark matter that astrophysical and cosmological observations indicate constitutes the bulk of the matter in the universe, nor does it explain the dark energy driving cosmic acceleration. The observed matter-antimatter asymmetry of the universe cannot be accounted for by the limited CP violation present in the SM. The discovery of neutrino oscillations demonstrates that neutrinos have mass, which the minimal SM does not accommodate. The theory also leaves several structural questions unanswered: the hierarchy problem asks why the Higgs mass is so much smaller than the Planck scale despite large quantum corrections; the flavor puzzle asks why the fermion masses span many orders of magnitude; and the strong CP problem asks why θ_{QCD} is so unnaturally small. A further, more subtle concern is the stability of the electroweak vacuum itself. The shape of the Higgs potential far from our vacuum is not fixed by low-energy measurements but depends sensitively on the Standard Model parameters, in particular the top quark and Higgs boson masses, whose measured values place the vacuum remarkably close to the boundary between absolute stability and metastability

(Figure 1.5).

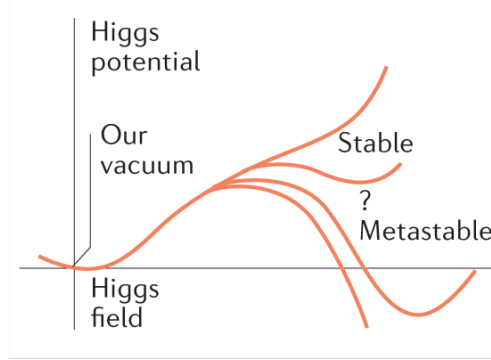


Figure 1.5. The Higgs potential, as an expansion about the electroweak vacuum, indicating regions of stability and metastability.

These shortcomings motivate the search for physics beyond the Standard Model (BSM). A wide range of theoretical frameworks have been proposed to address them, including supersymmetry, additional gauge symmetries and heavy gauge bosons, extended Higgs sectors, extra spatial dimensions, and grand unified theories. Experimentally, BSM physics is pursued through direct searches for new particles at high-energy colliders, precision measurements that could reveal deviations from SM predictions, and astrophysical and cosmological observations. The Higgs boson, as the most recently discovered and least explored sector of the SM, is a particularly sensitive probe: its couplings and production mechanisms could be modified by new physics. Such modifications would manifest as *anomalous Higgs production*: production rates and kinematic distributions that depart from their SM predictions because the underlying couplings differ from their SM values. This anomalous Higgs production, the subject of this dissertation, offers a direct window onto possible deviations from the Standard Model.

Chapter 2

The Large Hadron Collider and the CMS Experiment

2.1 The Large Hadron Collider (LHC)

The Large Hadron Collider (LHC) is the world's largest and highest-energy particle accelerator, located at CERN, the European Organization for Nuclear Research, near Geneva, Switzerland [49]. Housed in a 27-kilometer circular tunnel roughly 100 m underground, the LHC collides two counter-rotating beams of protons (and, in dedicated runs, heavy ions) at four interaction points (IPs), where the ATLAS, CMS, LHCb, and ALICE detectors record the products of the collisions. Designed for a maximum pp center-of-mass energy of $\sqrt{s} = 14$ TeV and a design instantaneous luminosity of $10^{34} \text{ cm}^{-2} \text{ s}^{-1}$, although having operated at lower than design energies and luminosity until the end of Run 3, the LHC provides the energy and collision rate needed both to study the Standard Model with precision, most notably through the 2012 discovery of the Higgs boson by ATLAS and CMS, and to search for physics beyond it. This section describes how the proton beams are produced and accelerated, how they are guided and focused, and how the key figures of merit of the machine, luminosity, the beam spot, and pileup, are defined and optimized.

The CERN Accelerator Complex and Injection Chain

The protons that collide in the LHC are produced and brought to high energy by a chain of successively larger accelerators as depicted in Fig. 2.1, each boosting the beam energy before handing it to the next. Protons are obtained by stripping electrons from hydrogen and are first accelerated as H^- ions in the linear accelerator LINAC4 to an energy of 160 MeV [84]. At injection into the Proton Synchrotron Booster (PSB) the two electrons are stripped, leaving bare protons, which the PSB accelerates to 2 GeV. The beam is then passed to the Proton Synchrotron (PS), which raises the energy to 26 GeV and, crucially, imposes the characteristic 25 ns bunch spacing of the LHC beam. The Super Proton Synchrotron (SPS) accelerates the protons further to 450 GeV before injecting them into the two LHC rings. In the LHC itself, superconducting radio-frequency cavities operating at 400 MHz accelerate the beams to their collision energy of 6.8 TeV per beam, giving $\sqrt{s} = 13.6$ TeV in Run 3.

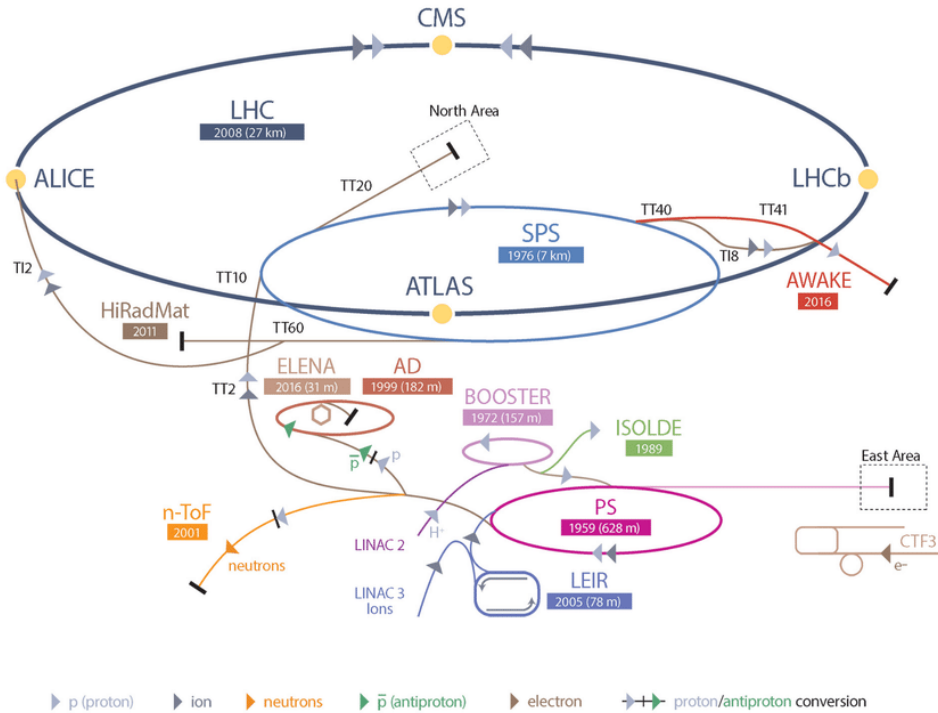


Figure 2.1. The CERN accelerator complex feeding the LHC, with the four main experiments at their interaction points [49].

Each LHC beam is not continuous but is organized into bunches: under nominal operation up to 2808 proton bunches per beam circulate with a 25 ns spacing, each bunch containing on the order of 1.2×10^{11} protons. The beams revolve at a frequency of $f_{\text{rev}} \approx 11.245$ kHz, so that bunch crossings occur at the IPs every 25 ns, a rate of 40 MHz. This time structure sets the fundamental clock of the experiments and drives the design of their trigger and readout systems (Section 2.2).

Magnet System

Bending the high-energy proton beams around the 27 km ring and keeping them tightly focused requires an extensive system of superconducting magnets. The 1232 main dipole magnets provide the bending field; built from niobium-titanium (NbTi) coils cooled to 1.9 K with superfluid helium, they reach magnetic fields of up to 8.3 T at the design beam energy. Because the two counter-rotating beams require opposite bending fields, the dipoles use a twin-bore "two-in-one" design that houses both beam channels in a common cold mass and cryostat, making the magnet system both compact and economical.

Focusing is provided primarily by 392 main quadrupole magnets arranged in a FODO lattice, an alternating sequence of horizontally and vertically focusing quadrupoles that confines the beam transversely while it circulates. Additional higher-order magnets (sextupoles, octupoles, and corrector magnets) compensate for chromatic and nonlinear effects and preserve beam stability over the many hours of a physics fill. Near each interaction point, a set of strong low- β quadrupoles known as the inner triplet performs the final focusing that squeezes the beams to their minimum size at the collision point, as discussed below.

Beam Optics

The transverse motion of the protons about the ideal orbit consists of oscillations, called betatron oscillations, whose local amplitude is governed by the optical β -function $\beta(s)$, where s is the coordinate along the beam. If β is low, the beam is narrower, "squeezed". If β is high, the

beam is wide and straight. The transverse size of the beam at any point is

$$\sigma(s) = \sqrt{\varepsilon \beta(s)}, \quad (2.1)$$

where ε is the geometric emittance, a measure of the area occupied by the beam in transverse position-momentum (phase) space. As the beam is accelerated, the geometric emittance shrinks (adiabatic damping), but the *normalized* emittance

$$\varepsilon_n = \gamma_r \beta_r \varepsilon \approx \gamma_r \varepsilon \quad (2.2)$$

is conserved, where γ_r and β_r are the relativistic Lorentz factor and velocity ratio of the protons. The LHC operates with a normalized emittance of roughly $\varepsilon_n \approx 2\text{-}3.5 \mu\text{m}$, set by the injector chain.

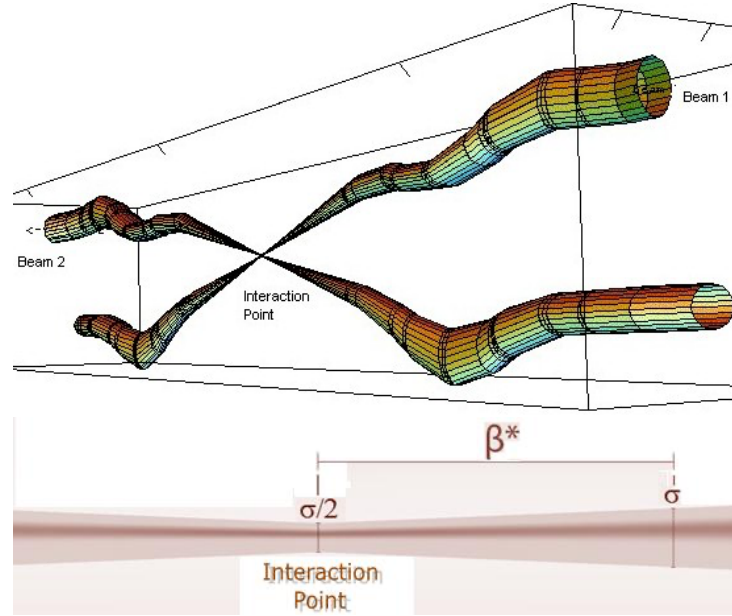


Figure 2.2. Beam optics near an LHC interaction point (top) and an example β -function profile defining β^* (bottom).

At the interaction point, the value of the β -function is denoted β^* . Making β^* as small as possible, "squeezing" the beams with the inner-triplet quadrupoles, minimizes the transverse

beam size at the collision point,

$$\sigma^* = \sqrt{\epsilon \beta^*} = \sqrt{\frac{\epsilon_n \beta^*}{\gamma_r}}, \quad (2.3)$$

and thereby maximizes the density of protons at the crossing and the resulting collision rate. The LHC injects and ramps the beams with a relatively large $\beta^* \approx 11$ m for stability, then squeezes to a small value at top energy before bringing the beams into collision; the design value is $\beta^* = 0.55$ m, while Run 2 routinely achieved $\beta^* \approx 0.30$ m at the ATLAS and CMS interaction points, corresponding to a transverse beam size at the IP of only $\sim 10\text{-}20$ μm .

Because the bunches are squeezed so tightly and follow a common path through the shared focusing magnets near the IP, the two beams cross at a small angle θ_c (a few hundred microradians) rather than head-on, in order to avoid unwanted parasitic collisions away from the nominal crossing point. This crossing angle reduces the geometric overlap of the colliding bunches and hence the luminosity, an effect captured by a geometric reduction factor discussed next.

Luminosity

While the center-of-mass energy sets the kinematic reach of the collider, the instantaneous luminosity $\mathcal{L}$ sets the rate at which collisions, and therefore rare processes, occur. The rate of any process with cross section σ is $dN/dt = \mathcal{L} \sigma$, and for (approximately) Gaussian beams colliding at the LHC the luminosity is

$$\mathcal{L} = \frac{n_b N^2 f_{\text{rev}} \gamma_r}{4\pi \epsilon_n \beta^*} F, \quad (2.4)$$

where n_b is the number of colliding bunches per beam, N the number of protons per bunch, f_{rev} the revolution frequency, and $F \leq 1$ the geometric reduction factor arising from the crossing

angle,

$$F = \left[1 + \left(\frac{\theta_c \sigma_z}{2 \sigma^*} \right)^2 \right]^{-1/2}, \quad (2.5)$$

with σ_z the longitudinal (bunch-length) RMS and σ^* the transverse beam size at the IP. Equation (2.4) makes explicit why β^* and the emittance are central to collider performance: luminosity grows as β^* is squeezed and as the beams are made brighter (small ε_n , large N). The LHC has delivered peak instantaneous luminosities exceeding the design value of $10^{34} \text{ cm}^{-2} \text{ s}^{-1}$, reaching $\sim 2 \times 10^{34} \text{ cm}^{-2} \text{ s}^{-1}$ in Run 2.

The quantity relevant to a physics analysis is the time-integrated luminosity,

$$L_{\text{int}} = \int \mathcal{L} \, dt, \quad (2.6)$$

usually quoted in inverse femtobarns ($1 \text{ fb}^{-1} = 10^{39} \text{ cm}^{-2}$), which together with the cross section determines the expected event yield $N = \sigma L_{\text{int}}$ (Section 1.7). The LHC delivered roughly 160 fb^{-1} to CMS over Run 2. Precise knowledge of L_{int} is itself a significant experimental task, as its uncertainty propagates directly into every cross-section measurement.

The Beam Spot

The region in which the two bunches actually overlap and collide is called the luminous region, or beam spot. Because each bunch is approximately Gaussian, the beam spot is an ellipsoid whose transverse extent is set by σ^* (of order 10-20 μm in x and y) and whose longitudinal extent σ_z is several centimeters, reflecting the bunch length. The beam spot is not perfectly static: its position and size drift slowly during a fill as the beams are adjusted, and it is therefore measured continuously *in situ* from the distribution of reconstructed primary collision vertices.

The beam spot is an essential input to event reconstruction. It provides a prior constraint on where collisions occur, improving the reconstruction of primary vertices and the determination of track impact parameters, which in turn underpins the identification of displaced vertices from

heavy-flavor decays (b-tagging) central to many analyzes, including the one presented in this thesis. Its longitudinal spread also sets the scale over which pileup vertices must be separated, motivating the precision tracking and timing detectors described in Section 2.2.

Pileup

The high luminosity that makes rare processes accessible carries an unavoidable cost: at each bunch crossing, many pp interactions occur simultaneously, a phenomenon known as pileup. The mean number of inelastic interactions per crossing is

$$\langle\mu\rangle = \frac{\mathcal{L} \sigma_{\text{inel}}}{n_b f_{\text{rev}}}, \quad (2.7)$$

where $\sigma_{\text{inel}} \approx 80$ mb is the inelastic pp cross section at LHC energies. During Run 2 the average pileup was typically $\langle\mu\rangle \approx 30$ -50 as shown in Fig. 2.3, meaning that the single hard-scattering event of interest is accompanied by tens of soft, overlapping collisions in the same crossing. In addition to this in-time pileup, the finite response time of some detectors makes them sensitive to collisions from neighboring bunch crossings, known as out-of-time pileup.

Pileup degrades the resolution of reconstructed quantities such as jet energies and missing transverse momentum and complicates the association of particles with the correct primary vertex. Mitigating it requires highly granular detectors, precise tracking to assign charged particles to their vertex of origin, dedicated algorithms such as pileup-per-particle identification, and, in the upgraded detectors, precision timing to resolve overlapping vertices in time as well as space. Managing ever-higher pileup is one of the central experimental challenges driving the detector upgrades discussed in Section 2.2.1 and Chapter 5.

LHC Run Periods and the High-Luminosity Upgrade

The LHC physics program is organized into multi-year run periods separated by long shutdowns (LS) for maintenance and upgrades. Run 1 (2010–2012) operated at $\sqrt{s} = 7$ and 8 TeV and culminated in the discovery of the Higgs boson in 2012. Following LS1, Run 2

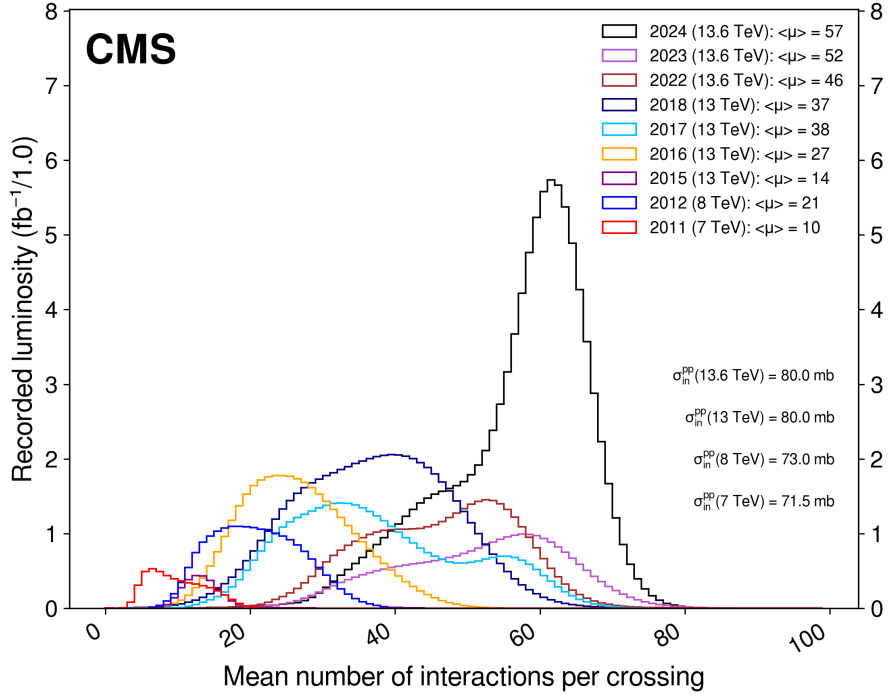


Figure 2.3. Interactions per bunch crossing in CMS during Phase-1 operation, with the mean $\langle\mu\rangle$ per run [35].

(2015–2018) operated at $\sqrt{s} = 13$ TeV and delivered the bulk of the dataset used for precision Higgs measurements and a broad program of searches. After LS2, during which the injector chain was upgraded with LINAC4 and the experiments installed major detector improvements, Run 3 began in 2022 at the increased energy of $\sqrt{s} = 13.6$ TeV and is scheduled to continue through 2026, accumulating a substantially larger dataset.

After Run 3, the accelerator will undergo its most significant transformation in the High-Luminosity LHC (HL-LHC) upgrade [21], scheduled to begin operation toward the end of the decade. The HL-LHC will replace the inner-triplet quadrupoles with new, large-aperture niobium-tin (Nb_3Sn) magnets capable of squeezing the beams to $\beta^* \approx 0.15$ m, and will introduce crab cavities that rotate the bunches to restore head-on overlap despite the crossing angle, recovering the geometric factor of Eq. (2.5). Together with brighter beams, these improvements will raise the instantaneous luminosity to a leveled $5 \times 10^{34} \text{ cm}^{-2} \text{ s}^{-1}$ (with the capability to reach 7.5×10^{34}), at the cost of an average pileup of $\langle\mu\rangle \approx 140\text{--}200$. Over roughly a decade of operation the



Figure 2.4. Timeline of the LHC runs and long shutdowns, with the energy and integrated luminosity of each, including the HL-LHC [27].

HL-LHC aims to deliver an integrated luminosity of 3000 fb^{-1} , an order of magnitude beyond the original program. This unprecedented dataset, and the extreme pileup that accompanies it, is the primary motivation for the Phase-2 detector upgrades and the computing and tracking developments that form a major part of this thesis.

2.2 The Compact Muon Solenoid (CMS) Detector

The Compact Muon Solenoid (CMS) is a general-purpose particle detector designed to investigate a wide range of physics phenomena at the Large Hadron Collider [29]. It is built as a sequence of nested, roughly cylindrical subdetectors arranged like the layers of an onion around the interaction point, so that a particle produced in a collision encounters each measuring layer in turn. Reading outward from the beam: the silicon tracker records the trajectories of charged particles; the electromagnetic calorimeter (ECAL) absorbs electrons and photons; the hadronic calorimeter (HCAL) absorbs hadrons; a large superconducting solenoid provides the magnetic field that bends charged tracks; and gas-based muon detectors embedded in the iron return yoke identify the only charged particles that penetrate that far. No single technology measures energy, momentum, charge, and identity at once, so each layer is built around the

particular interaction its target particles undergo in matter, and the layers are ordered so that each species is measured by the instrument best suited to it before another absorbs it: tracking first, because it must be non-destructive; then the calorimeters, which measure precisely by absorbing; and the muon chambers last, reached only by the particles that survive everything else. The relevant interaction is introduced below alongside the subdetector built to exploit it. Together the layers provide the nearly hermetic, 4π coverage required to reconstruct an entire event and to infer the presence of non-interacting particles from momentum imbalance. The detector measures 21.6 m in length and 15 m in diameter and weighs roughly 14,000 tonnes. Its name reflects its three defining characteristics: it is compact for the field it generates, it is optimized for precise muon measurement, and it is built around a powerful solenoidal magnet.

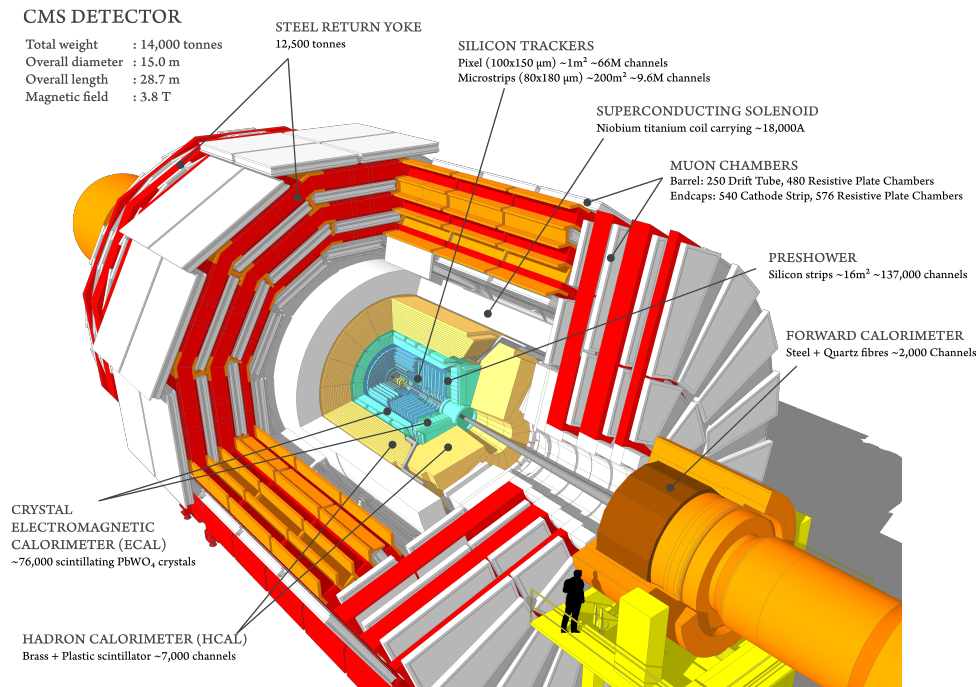


Figure 2.5. Schematic view of the CMS detector, showing the various subdetectors and their arrangement around the interaction point [29].

Coordinate System and Conventions

CMS uses a right-handed coordinate system with its origin at the nominal interaction point. The x -axis points radially inward toward the center of the LHC ring, the y -axis points

vertically upward, and the z -axis runs along the anticlockwise beam direction. Because the detector is cylindrical, it is natural to use the azimuthal angle ϕ , measured from the x -axis in the transverse x - y plane, and the polar angle θ , measured from the z -axis. In place of θ , high-energy physics conventionally uses the pseudorapidity,

$$\eta = -\ln \tan \frac{\theta}{2}, \quad (2.8)$$

which for the relativistic particles typical of LHC collisions approximates the rapidity. Pseudorapidity is the preferred coordinate because particle production is approximately uniform in η , and because differences in η are invariant under Lorentz boosts along the beam axis, a crucial property given that the colliding partons carry unknown longitudinal momentum fractions (Section 1.7). Values $\eta = 0$ correspond to the plane transverse to the beam, while large $|\eta|$ denotes the forward, near-beam region.

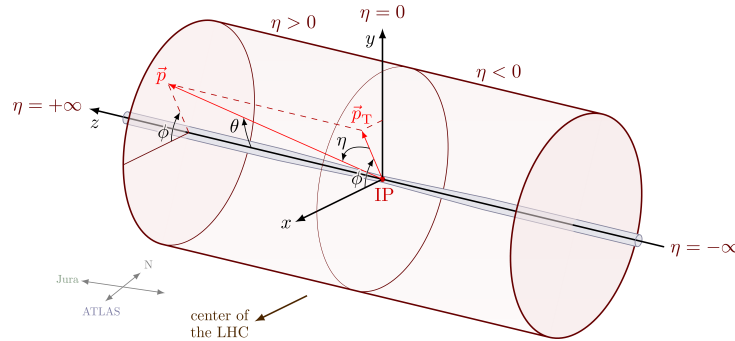


Figure 2.6. The CMS coordinate system: the x, y, z axes, azimuth ϕ , polar angle θ , and pseudorapidity η [29].

Since the initial momentum of the colliding partons along the beam is unknown but their net transverse momentum is negligible, quantities defined in the transverse plane are especially useful. The transverse momentum p_T and transverse energy E_T are the projections onto the x - y plane, and an imbalance in the vectorial sum of transverse momenta, the missing transverse momentum, signals undetected particles such as neutrinos. The angular separation between two objects is measured by $\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$, which is approximately invariant under

longitudinal boosts and is used throughout to quantify isolation and to cluster particles into jets.

The Superconducting Solenoid

The centerpiece of CMS, and the feature around which the whole experiment is designed, is its superconducting solenoid. The magnet is a niobium-titanium coil 12.5 m long with a 6 m inner bore, generating a uniform axial magnetic field of 3.8 T (Figure 2.7) and storing some 2.6 GJ of energy. Its bore is large enough to enclose the entire tracker and both calorimeters, so that charged particles bend within the field over their full measured trajectory. The magnetic flux is returned through a massive iron yoke, which both confines the return field (about 2 T) and is interleaved with the muon chambers, serving simultaneously as a structural skeleton, a hadron absorber, and the medium in which muon momenta are independently measured.

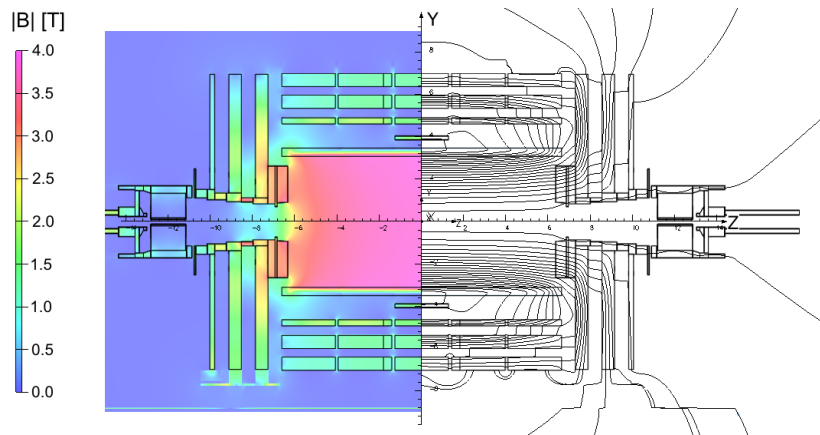


Figure 2.7. Value of $|B|$ and the field lines predicted on a longitudinal section of CMS [39]

Tracker

A precise measurement of charged particle momenta is intrinsic for the physics performance of CMS. In a homogenous magnetic field, the trajectory of a charged particle follows an approximately helical path, with a curvature that is inversely proportional to the particle momentum. The CMS tracker facilitates the measurement of this curvature by providing discrete spatial measurements (“hits”) of charged particles as they traverse the successive layers of the detector.

Those hits are produced by ionization. A charged particle heavier than the electron, a muon, pion, proton, or other hadron, traversing a medium loses energy mainly through its electromagnetic interaction with the atomic electrons, exciting and ionizing the atoms along its path; the liberated ion-electron pairs are the signal a tracking detector collects. The mean rate of energy loss per unit path length is given by the Bethe-Bloch equation [71],

$$\left\langle -\frac{dE}{dx} \right\rangle = K z^2 \frac{Z}{A} \frac{1}{\beta^2} \left[\frac{1}{2} \ln \frac{2m_e c^2 \beta^2 \gamma^2 T_{\max}}{I^2} - \beta^2 - \frac{\delta(\beta\gamma)}{2} \right], \quad (2.9)$$

where $K \approx 0.307 \text{ MeV mol}^{-1} \text{ cm}^2$, z is the charge of the incident particle in units of e , Z/A is the atomic-number-to-mass ratio of the medium, β and γ are the particle's velocity and Lorentz factor, $T_{\max}$ is the largest energy transferable to a single electron in one collision, I is the mean excitation energy, and $\delta(\beta\gamma)$ is the density-effect correction that screens the field at high energies. Expressed as a mass stopping power the result is nearly material-independent, since $Z/A \approx 0.5$ for most materials.

The shape of Eq. (2.9), shown in Figure 2.8, is what makes silicon tracking work. At low velocity the $1/\beta^2$ prefactor dominates and the loss falls steeply; it reaches a broad minimum near $\beta\gamma \approx 3\text{-}4$, where a particle is termed minimum-ionizing (MIP) and deposits at a nearly universal rate of about $2 \text{ MeV cm}^2/\text{g}$, before the logarithmic term produces a slow relativistic rise that the density effect flattens into the Fermi plateau. Essentially every charged particle of interest at the LHC sits at or above this minimum, so each layer sees a comparable, predictable deposit. Equation (2.9) gives only the *mean*: in a thin sensor the deposit fluctuates from one traversal to the next with a long tail from rare large transfers to δ -rays, described by the asymmetric Landau-Vavilov distribution, so the most probable deposit lies below the mean and the calibration must account for the difference. Because at a fixed momentum the loss depends on velocity, and hence on mass, dE/dx also provides particle identification at low momentum.

The tracker is designed to deliver high spatial resolution while strictly minimizing the overall material budget. A low material budget reduces both multiple Coulomb scattering and

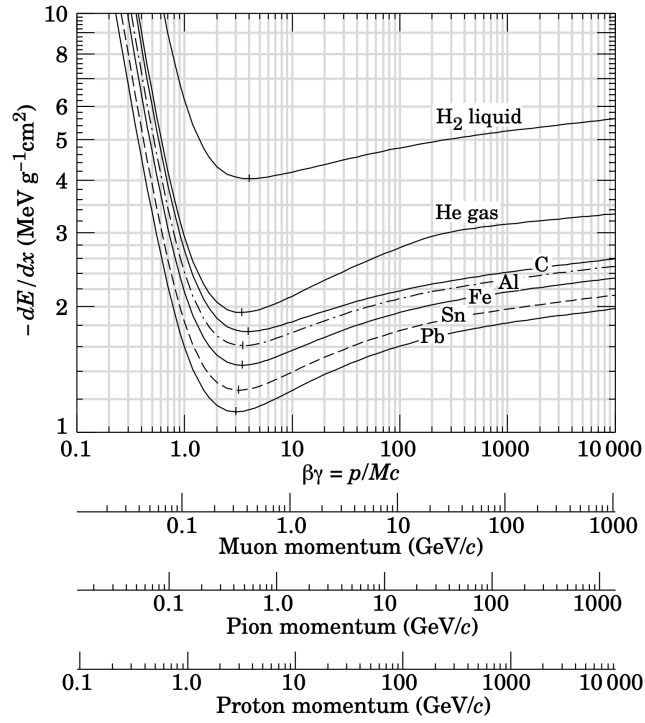


Figure 2.8. Mean energy loss $\langle -dE/dx \rangle$ versus $\beta\gamma$ for muons, pions, and protons, showing the minimum-ionizing regime and the relativistic rise [71].

energy loss, effects that otherwise degrade momentum and vertex resolution, and limits the rate of secondary interactions that complicate pattern recognition. As the innermost CMS subdetector, the tracker operates in an extreme radiation environment and must satisfy stringent requirements regarding radiation tolerance and operational stability.

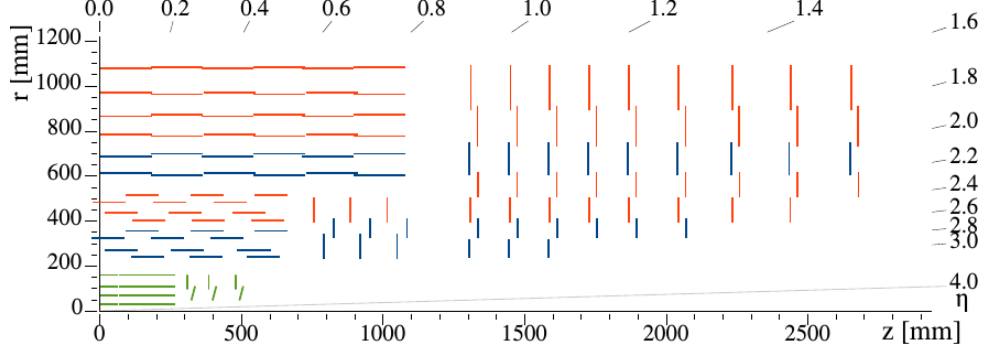


Figure 2.9. One quarter of the Phase-1 CMS tracker (r - z view): pixels (green), single- and double-sided strip modules (red, blue) [16].

To meet these demanding constraints, the CMS tracker is instrumented entirely with silicon sensors. The inner tracker comprises a highly segmented pixel detector, providing maximal granularity and spatial precision near the interaction point. During Phase-1, the pixel detector consists of four cylindrical barrel layers (BPIX) placed at radii of approximately 3, 7, 11, and 16 cm, supplemented by endcap disks (FPIX) on either side. This proximity to the beam pipe is vital for the efficient reconstruction of secondary vertices from short-lived particles, such as b -hadrons and τ -leptons.

The pixel system encompasses approximately 124 million channels. The individual silicon pixels have dimensions of $100 \times 150 \mu\text{m}^2$. As a charged particle traverses the depleted silicon bulk, it creates electron-hole pairs via ionization. The resulting charge drifts in the applied electric field to the readout electrodes, inducing small current pulses. These analogue signals are subsequently collected, amplified, and digitized by custom readout chips (ROCs) bump-bonded to the silicon sensors.

Operating in close proximity to the interaction point exposes the pixel detector to severe particle fluences; the innermost layer experiences hit rates of approximately $600 \text{ MHz}/\text{cm}^2$. The

readout architecture is explicitly designed to disentangle these highly populated environments with minimal dead time and data loss. Furthermore, despite consuming strictly limited power per channel ($\sim 50 \mu\text{W}$), the aggregate power dissipation of the pixel detector is roughly 7.5 kW. To manage this thermal load and mitigate radiation-induced damage, specifically to reduce leakage currents and prevent reverse annealing of the silicon bulk, the detector is actively cooled to -20°C using dedicated cooling loops.

Surrounding the pixel detector is the silicon strip tracker, which provides expansive geometric coverage alongside robust track-finding capabilities. After exiting the pixel layers, particles traverse up to ten layers of silicon strip sensors extending to a radial distance of 130 cm. The strip tracker is subdivided into four main structural components: the Tracker Inner Barrel (TIB) comprising four layers, two Tracker Inner Disks (TID) composed of three small disks each, the Tracker Outer Barrel (TOB) consisting of six concentric layers, and two Tracker Endcaps (TEC) closing off the tracker volume on either side.

In total, the strip tracker contains approximately 15,200 microstrip modules, yielding about 10 million individual readout channels. Similar to the pixel sensors, passing charged particles induce ionization signals in the silicon strips. These signals are read out by Analogue Pipeline Voltage (APV25) front-end ASICs. Four or six APV25 chips are housed on a front-end hybrid per module, which continuously samples the sensor signals at the LHC bunch-crossing frequency, buffers them pending Level-1 trigger decisions, and subsequently multiplexes the analogue data for optical transmission to off-detector electronics.

Like the pixel system, the strip tracker is operated at -20°C to suppress radiation damage effects in the silicon sensors. The original CMS strip tracker was installed in 2007 and has operated reliably since the start of LHC Run 1. The original pixel detector was replaced during the Extended Year-End Technical Stop (EYETS) of 2016–2017 with the upgraded Phase-1 pixel detector, specifically designed to cope with the elevated pileup and higher instantaneous luminosity of LHC Run 2 and Run 3.

For the HL-LHC this tracker no longer suffices, and it will be replaced in its entirety by a

new all-silicon system; that Phase-2 tracker, which forms a major part of this thesis, is described together with the other upgrades in Section 2.2.1.

Electromagnetic Calorimeter

Where the tracker is built to disturb particles as little as possible, a calorimeter is built to stop them completely, and it measures their energy by doing so. For electrons and photons the mechanism is an electromagnetic cascade, driven by two processes that feed one another.

The first is bremsstrahlung. A charged particle deflected in the electric field of a nucleus radiates photons, and the radiated power scales as $1/m^2$, so at a given energy the effect is overwhelmingly larger for the electron than for muons or hadrons. The mean energy of a high-energy electron falls exponentially with the distance traversed,

$$\langle E(x) \rangle = E_0 e^{-x/X_0}, \quad (2.10)$$

which defines the radiation length X_0 , the thickness over which an electron retains only $1/e$ of its energy. It is a property of the material ($X_0 \approx 0.56$ cm in lead) and is the natural unit of thickness for electromagnetic processes. Which loss mechanism dominates is set by the critical energy E_c , at which the radiative and ionization rates are equal (Figure 2.10): below it a particle loses energy mainly by ionization, above it by radiation. For electrons in lead $E_c \approx 7$ MeV. Muons, some 200 times heavier, have a critical energy of hundreds of GeV, which is precisely why they cross the entire detector depositing only ionization energy, the fact the muon system of Section 2.2 is built on.

The second is photon conversion. Photons carry no charge and so do not ionize continuously; they interact discretely, attenuating a beam as $I(x) = I_0 e^{-\mu x}$. Three processes dominate in turn as the energy rises (Figure 2.11): the **photoelectric effect** at the keV scale, where the photon is absorbed and a bound electron ejected, with a cross section rising steeply with atomic number ($\sigma \propto Z^n$, $n \approx 4-5$) and falling as E^{-3} ; **Compton scattering** at intermediate energies,

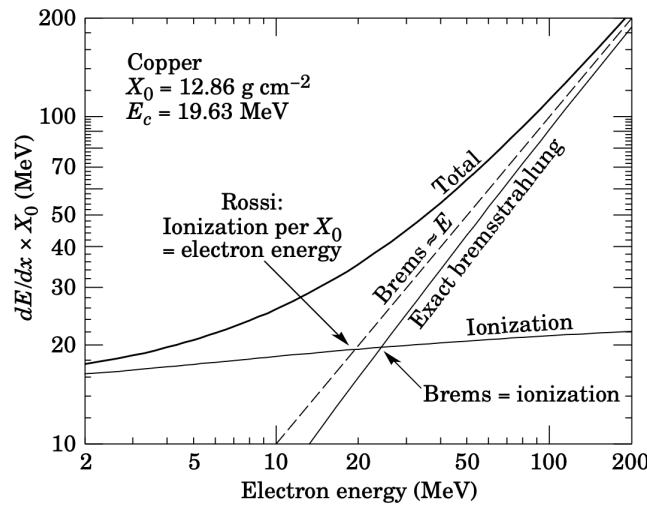


Figure 2.10. Fractional electron energy loss per radiation length from ionization and bremsstrahlung, crossing at the critical energy E_c [71].

falling roughly as $\ln(E)/E$; and **pair production** above $2m_e c^2 = 1.022$ MeV, where the photon converts in the field of a nucleus into an e^+e^- pair. Pair production dominates at high energy and is what makes photon calorimetry possible at all; its mean free path is closely related to the radiation length, $\lambda_{\text{pair}} = \frac{9}{7}X_0$.

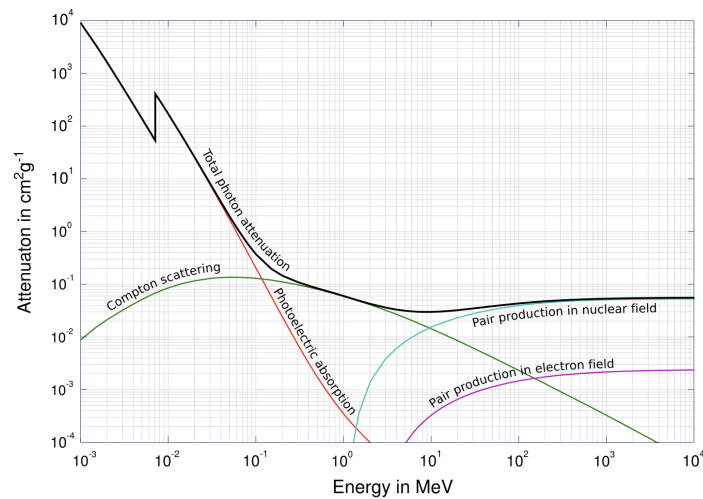


Figure 2.11. Photon mass attenuation versus energy, decomposed into the photoelectric effect, Compton scattering, and pair production [71].

Together these two processes produce the cascade sketched in Figure 2.12: an energetic

electron radiates photons, each of which converts into a pair, which radiate in turn. In the simplest model each step takes one radiation length, so the population roughly doubles every X_0 while the energy per particle halves, giving $\sim 2^N$ particles of energy $E_0/2^N$ after N radiation lengths. Multiplication continues until the per-particle energy drops below E_c , after which ionization and absorption end the cascade. The depth required, $N_c = \ln(E_0/E_c)/\ln 2$, grows only logarithmically with the incident energy, so a calorimeter of modest thickness in units of X_0 contains showers over a wide energy range, while the transverse spread is characterized by the Molière radius. Because the total number of particles produced is proportional to the incident energy, the total deposited signal measures the energy of the original electron or photon.

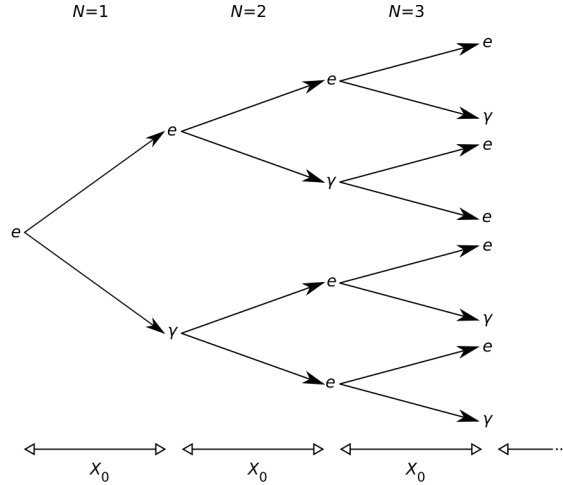


Figure 2.12. Simplified electromagnetic shower cascade, developing over the radiation length X_0 [62].

The CMS electromagnetic calorimeter (ECAL) is built to contain these showers in full. It is a homogeneous calorimeter built out of 75,848 lead tungstate (PbWO_4) crystals, a scintillating material chosen for its high density, short radiation length ($X_0 \approx 0.89$ cm), and small Molière radius (≈ 2.2 cm), which together permit a compact yet finely segmented detector that contains a shower within about $25 X_0$. The scintillation light is fast, about 80% is emitted within the 25 ns bunch spacing, which is essential for operation at the LHC. The ECAL is divided into a barrel region (EB) covering $|\eta| < 1.479$ and two endcap regions (EE) covering $1.479 < |\eta| < 3.0$. Each

crystal is read out by avalanche photodiodes (APDs) in the barrel and radiation-hard vacuum phototriodes (VPTs) in the endcaps, which convert the scintillation light into an electrical signal. A dedicated preshower detector, described below, is mounted in front of the endcaps. The energy resolution of a homogeneous calorimeter takes the form

$$\left(\frac{\sigma_E}{E}\right)^2 = \left(\frac{A}{\sqrt{E}}\right)^2 + \left(\frac{B}{E}\right)^2 + C^2, \quad (2.11)$$

where A is the stochastic term from shower fluctuations, B the noise term from electronics and pileup, and C a constant term from calibration and non-uniformities that dominates at high energy; for the CMS ECAL the measured values are approximately $A \approx 2.8\% \sqrt{\text{GeV}}$, $B \approx 0.12 \text{ GeV}$, and $C \approx 0.3\%$. This excellent resolution was decisive for the discovery of the Higgs boson in the $H \rightarrow \gamma\gamma$ and $H \rightarrow ZZ^* \rightarrow 4\ell$ channels.

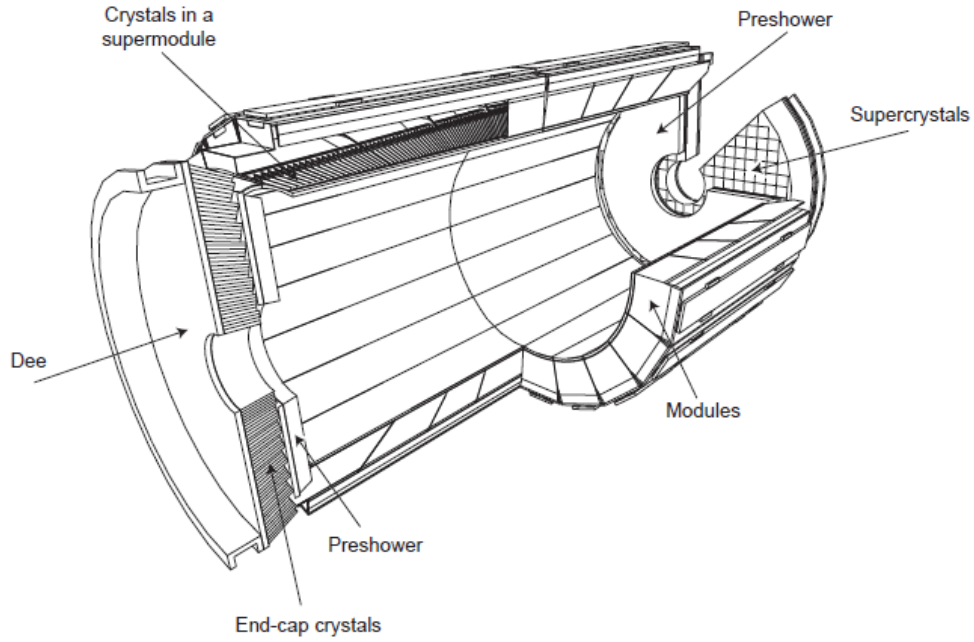


Figure 2.13. Layout of the CMS electromagnetic calorimeter [29].

Preshower Detector

The $H \rightarrow \gamma\gamma$ channel that drove the ECAL design places a premium on distinguishing genuine prompt photons from the dominant background of neutral pions, which decay as $\pi^0 \rightarrow \gamma\gamma$. When the two photons are sufficiently collimated their showers merge in adjacent crystals and mimic a single high-energy photon. This ambiguity is most severe in the endcaps, where the π^0 boost is largest and the opening angle between the photons smallest, so a preshower detector (ES) is mounted in front of each endcap to resolve them (as showcased in Fig 2.13). Its silicon strips are 2 mm wide, more than an order of magnitude finer than the ≈ 3 cm transverse size of the ECAL crystals, which is sufficient to separate the two closely spaced showers.

The preshower is a sampling device comprising two lead radiators, each followed by a plane of silicon strip sensors of the same type used in the tracker (Section 2.2). A photon initiates an electromagnetic shower in the lead, and the resulting e^+e^- pairs are measured in the silicon, sampling the shower energy; the two planes, with their strips oriented orthogonally, together localize the shower in both transverse coordinates. Photon candidates reconstructed in the ECAL are extrapolated back toward the interaction point, the associated preshower deposits are added to the crystal energy, and the lateral profile recorded in the strips discriminates a single photon from a $\pi^0 \rightarrow \gamma\gamma$ pair.

Each endcap is instrumented with roughly 8 m² of silicon, tiled from sensors of $\approx 6.3 \times 6.3$ cm² each divided into 32 strips and arranged to approximate the circular acceptance of the endcap crystals. As in the tracker, the sensors must be held between -10 and -15°C to limit radiation-induced leakage current, whereas the adjacent ECAL crystals must be stabilized to within 0.1°C of their higher operating temperature to preserve the constant term of Eq. (2.11); the preshower is therefore cooled on the inside while kept warm on the outside through combined cooling and heating. The complete assembly forms a disc only about 20 cm thick that nonetheless integrates both lead radiators, both instrumented silicon planes and their electronics, and the thermal control, with a central aperture for the beam pipe.

Hadronic Calorimeter

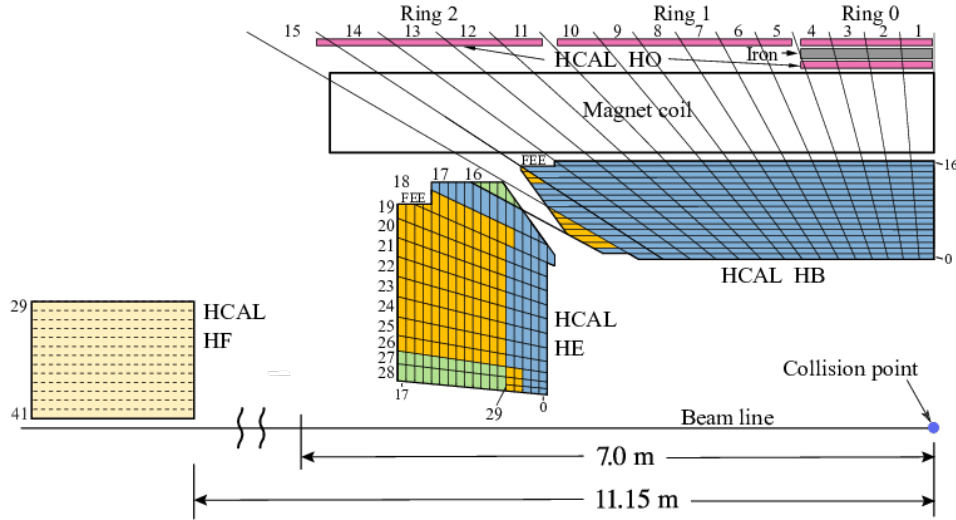


Figure 2.14. Schematic view of one quadrant of the CMS hadronic calorimeter [29].

Hadrons, protons, neutrons, pions, and kaons, interact with matter not only electromagnetically but also through the strong force, undergoing inelastic nuclear collisions that produce secondary hadrons. Those secondaries interact in turn, and the result is a hadronic shower (Figure 2.15) whose development mirrors the electromagnetic cascade but is set by a different scale: the nuclear interaction length λ_I , the mean distance between inelastic collisions, which follows from $\sigma_{\text{inel}} \approx \sigma_0 A^{2/3}$ and is typically much larger than the radiation length ($\lambda_I \approx 17$ cm against $X_0 \approx 0.56$ cm in lead). Hadronic showers therefore begin later and run deeper and wider than electromagnetic ones, and a calorimeter that is to contain them must be many interaction lengths thick. They also carry a fluctuating electromagnetic component, largely from $\pi^0 \rightarrow \gamma\gamma$ decays, and lose a variable fraction of their energy to nuclear binding and to undetected neutrinos and muons, which is why hadronic energy resolution is intrinsically poorer than electromagnetic.

The hadronic calorimeter (HCAL) (Figure 2.14) is built around these constraints. Because full containment demands a great deal of material, the HCAL is a sampling calorimeter built from alternating layers of dense absorber and active plastic scintillator: the scintillation light

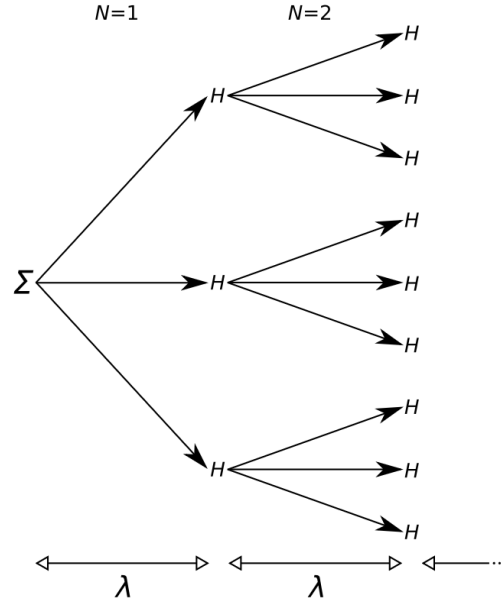


Figure 2.15. Simplified hadronic shower cascade, developing over the nuclear interaction length λ_I [62].

is collected by wavelength-shifting fibers and read out by photodetectors. It is divided into a barrel (HB, $|\eta| < 1.3$) and endcap (HE, $1.3 < |\eta| < 3.0$) using brass absorber, complemented by an outer calorimeter (HO) placed beyond the solenoid in the barrel to catch the tails of late-developing showers, and a forward calorimeter (HF, $3.0 < |\eta| < 5.2$). The HF sits in an intense radiation environment and therefore uses radiation-hard quartz fibers that detect Cherenkov light rather than scintillation, extending the hermetic coverage essential for a reliable missing-transverse-momentum measurement.

Cherenkov radiation arises when a charged particle crosses a medium faster than the phase velocity of light in it, $v > c/n$, whereupon it emits a coherent cone of light at a fixed angle to its direction,

$$\cos \theta_C = \frac{1}{n\beta}, \quad (2.12)$$

so that the emission has a velocity threshold at $\beta = 1/n$ and the opening angle grows with velocity above it (Figure 2.16). Unlike bremsstrahlung, which the particle radiates incoherently by itself, Cherenkov light is emitted coherently by the polarized medium as a whole. In the

HF the quartz responds only to the relativistic charged particles of a shower, which makes the signal fast and largely immune to the radiation damage that would quickly destroy a scintillator at $|\eta| > 3$. Elsewhere the same effect is used for particle identification: measuring the presence or angle of Cherenkov light fixes a particle's velocity, which together with its momentum gives its mass.

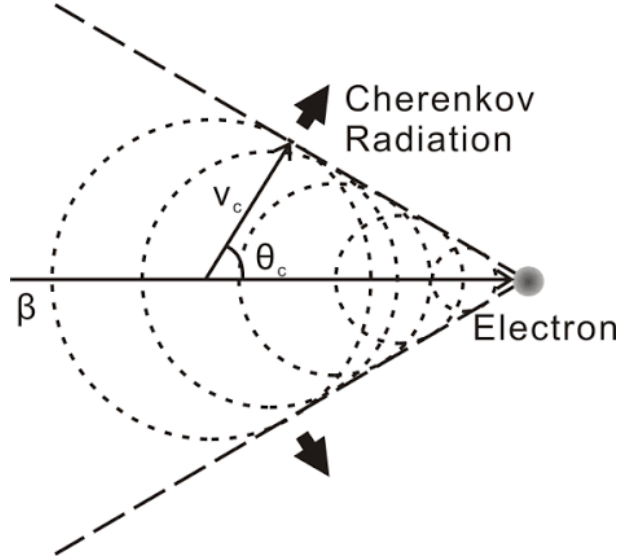


Figure 2.16. Geometry of Cherenkov radiation: a coherent wavefront at half-angle θ_C with $\cos \theta_C = 1/(n\beta)$ [62].

Hadronic showers fluctuate more than electromagnetic ones, so the HCAL energy resolution is intrinsically coarser than the ECAL's; for this reason CMS does not rely on the HCAL alone but combines tracker, ECAL, and HCAL information in the particle-flow algorithm to reconstruct jets and missing transverse momentum with substantially improved resolution.

Muon System

Muons are the only charged particles that traverse the calorimeters without showering, passing through as minimum-ionizing particles by virtue of their large critical energy (Section 2.2); CMS exploits this by placing its muon detectors in the outermost layers, embedded in the iron return yoke where only muons and weakly interacting particles penetrate. The muon system covers $|\eta| < 2.4$ and uses three complementary gaseous detector technologies: drift tubes

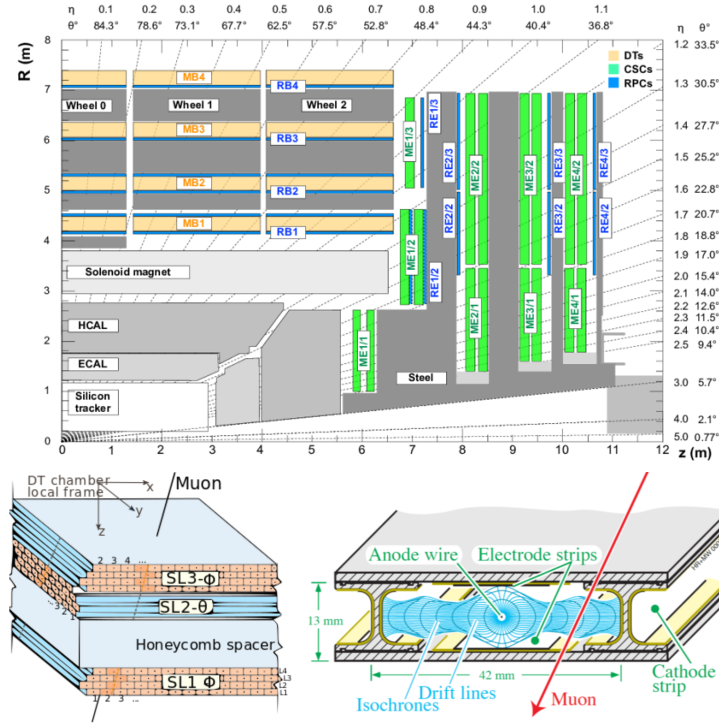


Figure 2.17. The CMS muon system (top): drift tubes, cathode strip chambers, and resistive plate chambers; cross-section of a DT and a CSC (bottom) [29].

(DTs) in the barrel ($|\eta| < 1.2$), where the rate is low and the field uniform; cathode strip chambers (CSCs) in the endcaps ($0.9 < |\eta| < 2.4$), which tolerate the higher rates and non-uniform field of the forward region; and resistive plate chambers (RPCs) covering both regions, which provide fast timing information for triggering. The DTs and CSCs deliver precise position measurements that yield a standalone momentum measurement from the bending in the yoke field, but the best muon momentum resolution is obtained by combining these hits with the inner tracker.

Trigger and Data Acquisition System

The LHC delivers bunch crossings at 40 MHz (Section 2.1), a rate far beyond what can be recorded; the trigger and data acquisition (DAQ) system reduces this to a manageable rate by selecting only the most interesting events in real time. CMS uses a two-tier trigger. The Level-1 (L1) trigger is implemented in custom hardware (ASICs and FPGAs) using coarse information from the calorimeters and muon systems, and reaches a decision within a fixed latency of a few

microseconds, reducing the rate from 40 MHz to about 100 kHz. Events accepted by L1 are read out and passed to the High-Level Trigger (HLT), a software system running a streamlined version of the full offline reconstruction on a large computing farm, which further reduces the rate to roughly 1-2 kHz. The accepted events are written to permanent storage at the CERN Tier-0 computing center, the entry point to the worldwide computing grid where the data is archived, first reconstructed, and distributed for offline analysis (Section 2.3).

2.2.1 The Phase-2 Upgrade for the HL-LHC

The HL-LHC (Section 2.1) will deliver an order of magnitude more integrated luminosity than the original program, at an average pileup of $\langle\mu\rangle \approx 140\text{-}200$ and with radiation doses that the present detector was never designed to withstand. To preserve, and in places extend, its physics performance under these conditions, CMS is undertaking a comprehensive Phase-2 upgrade. The following subsections describe, subsystem by subsystem, how each of the detectors introduced above is either rebuilt or replaced. The upgrade of the tracker, which forms a major part of this thesis, is treated first and in the most detail.

Phase-2 Tracker

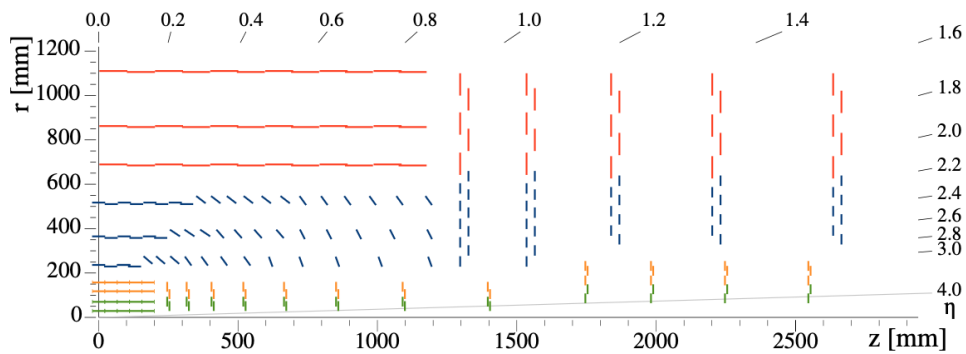


Figure 2.18. One quarter of the Phase-2 CMS tracker (r - z view): the Outer Tracker is built from PS modules (blue) at smaller radius and 2S modules (red) at larger radius, with the Inner Tracker pixel modules shown in green and orange [1].

Before the start of the HL-LHC both the strip tracker and the Phase-1 pixel detector will

have to be replaced to cope with the more demanding operational conditions and due to the significant damage and performance degradation they would suffer during the operation.

Accumulated radiation damage in the pixel sensors reduces the charge collection efficiency as well as the Lorentz angle, leading initially to decreased charge sharing among neighboring pixels and hence to deteriorated spatial resolution, and eventually to reduced hit efficiency. For the present strip tracker, the most prominent changes of detector properties with irradiation are the increase of the sensor depletion voltage and of the leakage current. The latter can be mitigated, up to a certain point, by lowering the operating temperature of the cooling system, while the former cannot. It has been demonstrated that basically all double-sided strip modules cannot be operated anymore at the nominal cooling temperature already after 1000 fb^{-1} .

Studies of the expected performance of the current tracking system as a function of integrated luminosity have shown unacceptable degradation beyond about 1000 fb^{-1} , including the deterioration of tracking and b tagging performance and a worsening of the impact parameter resolution. The physics program would also be affected by limitations in readout bandwidth and trigger latency. The tracker must therefore be entirely replaced for the Phase-2 physics program.

Similar to the Phase-1 configuration, the Phase-2 tracker will consist of an Inner Tracker (IT) based on silicon pixel modules and an Outer Tracker (OT) made from silicon modules with strip and macro-pixel sensors.

The main requirements for the tracker upgrade can be summarized as follows:

1. *Radiation Tolerance*: The upgraded tracker must be fully efficient up to an integrated luminosity of 4500 fb^{-1} (target luminosity of $3000 + 50\%$ tolerance for uncertainties in simulations of radiation exposure and potential of additional delivery should the ultimate luminosity scenario be realized.).
2. *Increased granularity*: In order to ensure efficient tracking performance with a high level of pileup, the channel occupancy must be kept at around or below the per cent level (per

mille level) in the Outer Tracker (Inner Tracker), which requires a high channel density. Target values of 140 and 200 collisions per bunch crossing are used to benchmark the performance of the detector.

3. *Improved two-track separation:* The present tracker has limited track finding performance in highly energetic jets, due to hit merging in the pixel detector. In order to optimally exploit the large amounts of collision data that will be taken during high luminosity operation, two-track separation needs to be improved
4. *Reduced material in the tracking volume.* The exploitation of the high luminosity will greatly benefit from a lighter tracker. The performance of the current tracker is affected by the amount of material, which also influences the performance of the calorimeters and of the overall event reconstruction in CMS.
5. *Robust pattern recognition.* Track finding under high pileup conditions becomes increasingly more difficult and time consuming. The design of the upgraded tracker should enable fast and efficient track finding, notably at the HLT.
6. *Contribution to the level-1 trigger.* The selection of interesting physics events at the first trigger stage becomes extremely challenging at high luminosity, not only because of the rate increase, but also because selection algorithms become inefficient in high pileup conditions. The CMS trigger will operate with substantially increased latency and output rate, and the tracker has to comply with those. In addition, in order to preserve and possibly enhance the performance in a wide spectrum of physics channels, CMS has chosen to use tracking information in the L1 event selection, anticipating part of the reconstruction presently performed in the HLT.
7. *Extended tracking acceptance.* The overall CMS physics capabilities will greatly benefit from an extended acceptance of the tracker and calorimeters in the forward region. The upgraded tracking system will provide efficient tracking up to about $|\eta| \sim 4$.

Phase-2 Calorimeters

For Phase-2 the barrel ECAL retains its lead-tungstate crystals and APDs but receives entirely new front-end and back-end electronics. These deliver single-crystal granularity and precise timing, with a resolution of about 30 ps for high-energy photons and electrons, to the upgraded Level-1 trigger, and the operating temperature is lowered from 18 to about 9°C to suppress the APD dark current that grows with accumulated radiation damage. The barrel HCAL likewise retains its brass-scintillator structure but is read out with silicon photomultipliers in place of the original hybrid photodiodes, an upgrade that increases the light yield and depth segmentation.

The endcap calorimetry is transformed far more radically. The endcap ECAL, the endcap HCAL, and the preshower are all removed and replaced by a single high-granularity calorimeter (HGCAL) covering $1.5 < |\eta| < 3.0$. It is a sampling calorimeter of unprecedented transverse and longitudinal segmentation, with roughly six million channels, conceived to enable particle-flow reconstruction and pileup rejection in the harsh forward environment of the HL-LHC. The electromagnetic section and the high-radiation part of the hadronic section use hexagonal silicon sensor modules as the active material, interleaved with lead, copper, and stainless-steel absorbers, while the lower-radiation hadronic section at larger depth uses plastic scintillator tiles read out by silicon photomultipliers. The fine three-dimensional granularity, complemented by precision timing of a few tens of picoseconds, allows individual showers to be resolved and assigned to the correct vertex, recovering in the endcaps the calorimetric performance that pileup would otherwise overwhelm; its intrinsic two-photon separation also removes the need for a dedicated preshower.

Phase-2 Muon System

The existing muon system is retained and reinforced. Gas electron multiplier (GEM) detectors are added in the forward region to maintain efficient muon triggering and reconstruction under the high rates of the HL-LHC; GEMs amplify the ionization from passing particles in

perforated metal foils and combine high rate capability with good spatial resolution. Three GEM stations are deployed (GE1/1, GE2/1, and ME0), the innermost of which, ME0, extends muon coverage to $|\eta| \approx 2.8$ in combination with the new tracker. The forward region is further instrumented with improved-timing resistive plate chambers (iRPCs), and the existing DT and CSC systems receive new readout electronics to cope with the higher Level-1 accept rate and longer latency.

Phase-2 Trigger

The trigger and data acquisition system is substantially upgraded to cope with the higher rates and pileup of the HL-LHC. The L1 accept rate rises to 750 kHz with a latency of about $12.5 \mu\text{s}$, and, for the first time, tracking information is incorporated at L1 (the track trigger), a capability that the tracker upgrade was explicitly designed to provide and that is essential to preserve sensitivity under extreme pileup.

MIP Timing Detector

At the HL-LHC, the average pileup of $\langle \mu \rangle \approx 140\text{--}200$ (Section 2.1) produces collision vertices so densely packed along the luminous region (Section 2.1) that many overlap spatially and cannot be separated by position alone. The solution is to add a fourth, temporal coordinate. The Phase-2 MIP Timing Detector (MTD) [32] measures the arrival time of charged particles with a precision of about 30–40 ps, enabling four-dimensional reconstruction of vertices in space and time and thereby recovering the ability to associate tracks, jets, and leptons with the correct primary vertex. The MTD comprises a barrel timing layer (BTL), based on cerium-doped lutetium-yttrium orthosilicate (LYSO:Ce) crystals read out by silicon photomultipliers and located between the tracker and the ECAL ($|\eta| < 1.45$), and an endcap timing layer (ETL), based on radiation-hard low-gain avalanche diodes (LGADs) covering $1.6 < |\eta| < 3.0$. By suppressing pileup that would otherwise overwhelm reconstruction, precision timing is one of the key enablers of the Phase-2 physics program and is closely connected to the tracking

developments that form a major part of this thesis (Chapter 5).

2.3 World Wide LHC Computing Grid

The volume of data produced by the LHC experiments far exceeds what any single computing center could store or process. After the trigger and data acquisition system has reduced the event rate (Section 2.2), CMS still records data at a rate of several gigabytes per second during stable beams, accumulating tens of petabytes per year, with comparable volumes again of simulated events. To store, reconstruct, and analyze this data, the LHC experiments share a single globally distributed computing infrastructure, the Worldwide LHC Computing Grid (WLCG) [64].

The WLCG federates the resources of more than 170 computing centers in over 40 countries, pooling on the order of a million processor cores and an exabyte of disk and tape storage into a common system that presents itself to users as a single facility. Rather than move every analyst to the data or every dataset to every analyst, grid middleware schedules computational jobs to run at the sites that already hold, or can readily access, the required data, an arrangement often summarized as "sending the jobs to the data."

The resources are organized hierarchically into tiers, with each tier assigned a distinct role and level of service. The single *Tier-0* center at CERN sits at the apex. It receives the raw data directly from the CMS high-level trigger farm (Section 2.2), archives the first, custodial copy to tape, performs the initial "prompt" reconstruction that converts raw detector signals into physics objects (Section 3.2), and distributes both the raw and reconstructed data outward to the next tier. The Tier-0 performs no end-user analysis; its role is the safe-keeping of the primary copy of the data and the timely first-pass processing of newly recorded events.

Beneath it, roughly a dozen *Tier-1* centers, large national computing facilities such as Fermilab for CMS in the United States, provide custodial tape storage of a second copy of the raw data, the computing capacity for large-scale reprocessing and re-reconstruction as calibrations

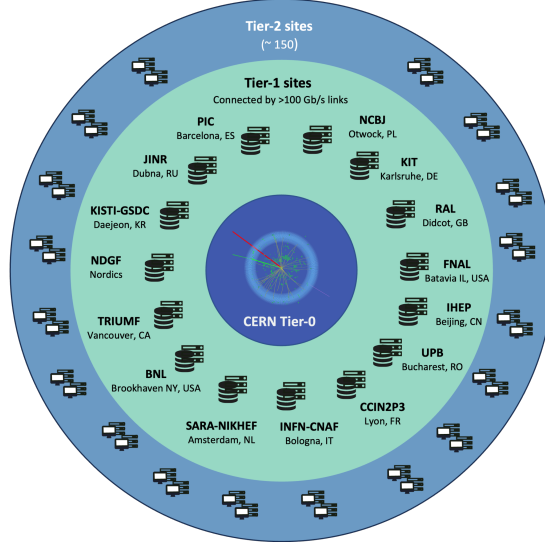


Figure 2.19. Tiered organization of the WLCG: data flows from the detector to the Tier-0 at CERN, then out to the Tier-1 and Tier-2 sites [64].

improve, and high-availability data serving to the sites below them. The Tier-0 and Tier-1 centers are interconnected by a dedicated, high-bandwidth optical-fiber private network, the LHC Optical Private Network (LHCOPN), that guarantees the throughput required for bulk data distribution. The far more numerous *Tier-2* centers, typically hosted at universities and research institutes, supply the bulk of the resources for Monte Carlo simulation production (Section 3.1) and for end-user physics analysis; they hold disk-resident copies of the datasets most in demand but make no long-term custodial commitment.

Finally, *Tier-3* resources are local institutional clusters and interactive machines used opportunistically by individual groups, with no formal obligation to the collaboration. This tiered structure is illustrated in Figure 2.19.

Operating this system requires not only distributed storage and processing but also the reliable, large-scale movement of data between sites. Events are grouped into files of $\mathcal{O}(\text{GB})$ and aggregated into named datasets, whose locations and replicas are tracked in central catalogs so that they can be discovered and addressed independently of the site that physically holds them. Bulk transfers between sites are orchestrated by a data-management system, for which CMS uses Rucio together with the File Transfer Service (FTS) [15], which queues, schedules,

and retries the underlying point-to-point copies. The transfers themselves are carried over grid protocols such as XRootD [45] and HTTPS. The same XRootD protocol additionally underpins a data federation that allows a job running at one site to read a file served from another over the wide-area network, decoupling the location of the computation from the location of the data and improving the overall efficiency of resource use.

The performance and scalability of this data-movement layer becomes a critical constraint in the HL-LHC era, when data volumes grow by roughly an order of magnitude; the resulting challenges, and the work undertaken in this dissertation to address them, are the subject of Chapter 6.

Chapter 3

Simulation and Reconstruction

3.1 Simulation

Comparing the predictions of the Standard Model, or of a proposed extension to it, with the data recorded by CMS requires bridging an enormous gap: theory is expressed in terms of Lagrangians and scattering amplitudes (Chapter 1), while the experiment yields electronic signals in millions of detector channels. Monte Carlo (MC) simulations help cover this gap. It generates collision events using theoretical models, propagates the resulting particles through a detailed model of the detector, emulates digitization and produces output in the same format as real data, so that simulation and data can be subjected to identical reconstruction and analysis. Simulation is indispensable for designing detectors, estimating signal and background yields, correcting for detector effects, and quantifying systematic uncertainties. The simulation proceeds in a sequence of stages, each handled by dedicated software and corresponding to a distinct physical regime, from the perturbative hard scattering down to the detector response. A schematic of the simulation chain is shown in Figure 3.1, and the following sections describe each stage in more detail.

Hard-Process Generation

The chain begins with the hard scattering, the high-momentum-transfer parton interaction of interest, computed from the matrix element $\mathcal{M}$ of the process using perturbative QCD

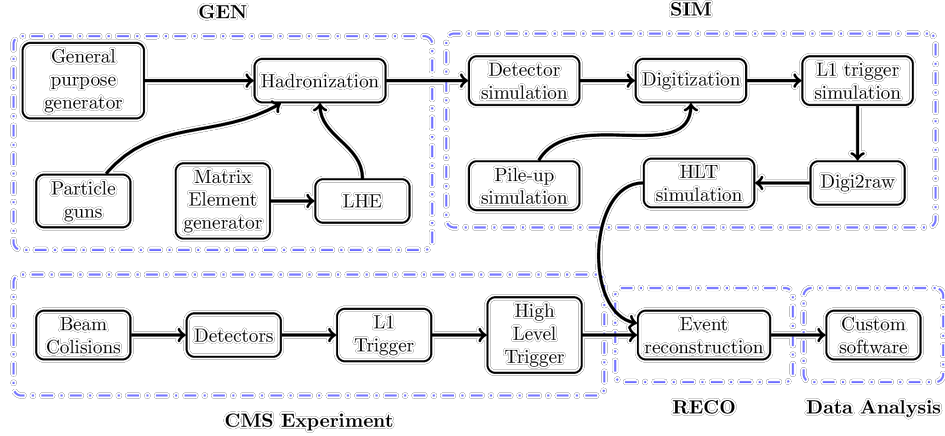


Figure 3.1. Schematic of the simulation chain, data collection and reconstruction in CMS.

(Section 1.7). For pp collisions this partonic cross section is convolved with the parton distribution functions (Section 1.7) to obtain the observable rate, as described by the factorization theorem. Tools such as MadGraph5_aMC@NLO [7] automatically generate the relevant Feynman diagrams and compute the matrix elements at leading or next-to-leading order in the strong coupling, then sample the corresponding phase space to produce a set of hard-scattering events with physically distributed kinematics. The ability to specify an arbitrary Lagrangian makes such generators equally suited to simulating Standard Model processes and hypothetical beyond-the-SM signals, which is essential for the search presented in this thesis.

Parton Showering and Hadronization

The partons emerging from the hard process carry color charge and large virtuality, and they radiate further gluons and quarks. This evolution toward lower momentum scales is modeled by a parton shower, which applies the splitting probabilities to generate successive $q \rightarrow qg$, $g \rightarrow gg$, and $g \rightarrow q\bar{q}$ emissions as both initial- and final-state radiation. Because QCD confines color (Section 1.3), the resulting partons cannot exist freely; at scales around 1 GeV, where the strong coupling becomes large and perturbation theory breaks down, they undergo hadronization into color-neutral hadrons, described by phenomenological models such as the Lund string model. The program Pythia [79] performs both the showering and the hadronization,

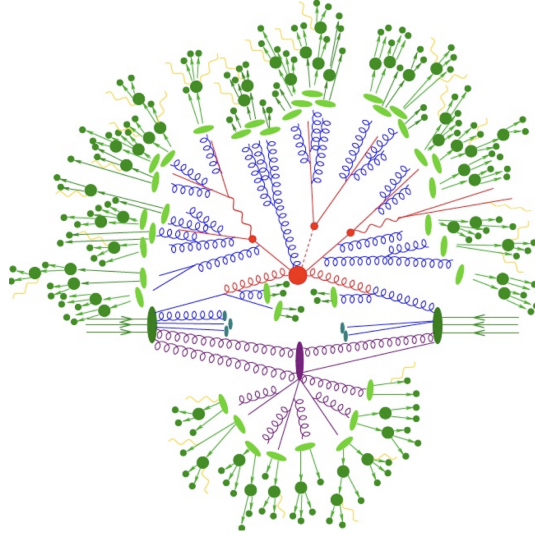


Figure 3.2. Parton shower and hadronization: the hard interaction (red), the parton-shower cascade (blue), secondary scatters (purple), and hadronization (green).

and additionally models the underlying event from secondary parton interactions and the decays of unstable hadrons. Additional inelastic pp interactions are overlaid on the hard scatter to reproduce the pileup conditions of the LHC (Section 2.1). A graphic showcasing the parton shower and hadronization is shown in Figure 3.2

Detector Simulation

The stable particles produced by the generator are then propagated through a detailed model of CMS using Geant4 [4], which transports each particle step by step through the detector geometry and materials, simulating the same electromagnetic and hadronic interactions that the subdetectors of Section 2.2 are built to exploit: ionization, bremsstrahlung, photon conversion, showering, and nuclear interactions. The energy deposited in each active element is recorded and then passed through a digitization step that emulates the response of the sensors and front-end electronics, including noise and the effects of pileup, producing simulated raw data in exactly the same format as the detector output. This full simulation ("FullSim") is highly accurate but computationally expensive, which, given the billions of events required for the HL-LHC, is one of the principal motivations for the computing developments discussed later in this thesis; faster,

parametrized alternatives (“FastSim”) trade some accuracy for speed when large samples are needed.

3.2 Reconstruction

Reconstruction is the process of converting the raw signals recorded by the detector, whether from real collisions or from simulation, into physical objects with measured momenta, energies, and identities. Because simulated events are produced in the same raw format, the identical reconstruction chain is applied to data and simulation, ensuring a like-for-like comparison. Reconstruction proceeds from local to global. Local reconstruction first interprets the signals within each subdetector: charge deposits in adjacent silicon channels are grouped into hits, and energy deposits in neighboring calorimeter cells are grouped into clusters. Global reconstruction then combines this information across subdetectors. Charged-particle trajectories are reconstructed from tracker hits using the combinatorial Kalman filter described in Section 3.2.1, and the reconstructed tracks are extrapolated back to identify primary interaction vertices and the displaced secondary vertices characteristic of heavy-flavor decays. Calorimeter clusters provide energy measurements for both charged and neutral particles. The quality of these reconstructed quantities directly limits the precision of every downstream measurement, which is why the development of fast and accurate reconstruction algorithms is a central experimental effort.

3.2.1 Charged-Particle Track Reconstruction at CMS

Principles of Charged Particle Tracking

In high-energy physics, the primary objective of a tracking system is to precisely measure the kinematics of charged particles emerging from collision events. When a charged particle traverses a uniform magnetic field, such as that provided by the CMS solenoid (Section 2.2), the Lorentz force compels it to follow a helical trajectory. By recording the spatially localized interaction points (hits) left by the particle in the sensitive layers of the silicon tracker (Section 2.2), the underlying trajectory can be reconstructed [42].

Mathematically, a helical track is fully defined by five parameters at a given point of closest approach (PCA) to a reference axis (typically the beamline). In the CMS coordinate system (Section 2.2), these track parameters are conventionally chosen as (Figure 3.3):

- d_0 : The transverse impact parameter, defined as the distance of closest approach to the origin in the transverse (x - y) plane.
- z_0 : The longitudinal impact parameter, defining the z -coordinate at the point of closest transverse approach.
- ϕ : The azimuthal angle of the momentum vector at the PCA.
- θ (or η): The polar angle (or pseudorapidity) of the momentum vector.
- q/p or q/p_T : The signed inverse momentum or transverse momentum, where q dictates the direction of bending in the magnetic field and relates directly to the track curvature.

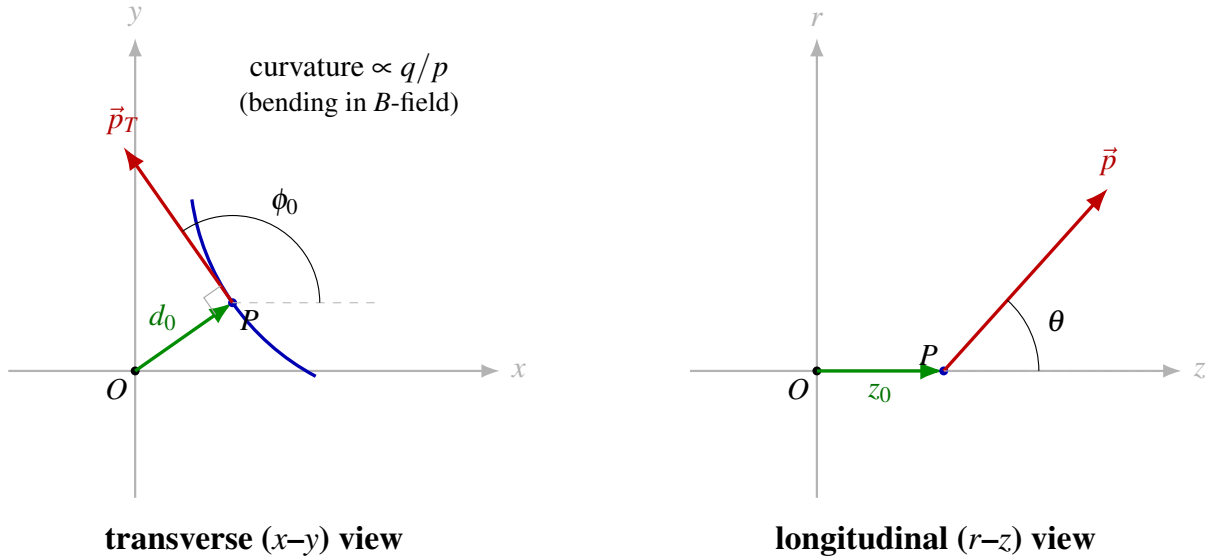


Figure 3.3. The five helix track parameters at the point of closest approach: transverse (x - y , left) and longitudinal (r - z , right) views.

Interactions with Detector Material

As particles traverse the inner detector, their trajectories are subject to physical perturbations due to interactions with the detector material. These effects must be accounted for during track reconstruction:

- **Multiple Coulomb Scattering (MCS):** Successive small-angle deflections off atomic nuclei alter the particle's trajectory. MCS introduces uncertainties in the track parameters and limits the momentum resolution, especially for low- p_T particles.
- **Energy Loss:** Particles continuously lose kinetic energy primarily via ionization (described by the Bethe-Bloch formula, Eq. (2.9)) and, for low-mass particles like electrons, via bremsstrahlung radiation (Section 2.2). Unaccounted energy loss leads to variations in the track curvature along the trajectory.

Overview of the General Tracking Sequence

The reconstruction of charged particle trajectories constitutes a critical element of the event reconstruction paradigm in CMS. To effectively isolate physics signals from complex backgrounds, the standard tracking sequence is conceptually partitioned into four primary stages [42]:

- **Track Seeding:** Initial track candidates, or seeds, are formulated utilizing a minimal subset of hits (typically two or three, often paired with a vertex position) within the innermost layers of the tracker. These seeds establish the initial trajectory parameters and their associated covariance matrices.
- **Track Building:** Originating from the seeds, trajectories are extrapolated outward through the active tracking volume. As the extrapolated trajectory intersects sensitive detector layers, spatially compatible hits from respective layers are iteratively incorporated into the track candidate using a search window predicated on trajectory uncertainties.

- **Track Fitting:** Upon the completion of hit collection, the assembled hit groupings are subjected to a definitive mathematical fit (typically using a Kalman Filter). This procedure optimally integrates the measurements and accounts for material effects to extract the final track parameters.
- **Track Selection:** The reconstructed tracks are evaluated using Multivariate Analysis (MVA) techniques or deep learning models to assess their quality. Tracks are filtered based on variables such as the normalized χ^2 of the fit, the number of missing hits, and kinematic compatibility, culminating in a highPurity designation for physics analysis.

Iterative Tracking

The CMS tracking framework operates as an iterative procedure, necessitating sequential passes over the available hit collections. Each iteration is optimized to reconstruct a distinct topological class of tracks, commencing with those most readily identifiable (e.g., high- p_T prompt tracks originating from the primary interaction vertex) and systematically advancing to more complex topologies (e.g., low- p_T or displaced tracks). Following each iteration, hits unambiguously associated with successfully reconstructed tracks are masked from subsequent iterations, thereby mitigating combinatorial complexity. This methodology ensures overarching tracking efficiency while rigorously regulating combinatorial background, a feature particularly vital in environments characterized by high pileup.

The Current CMS Tracking Sequence

The tracking implementation relies on the aforementioned iterative paradigm to preserve high reconstruction efficiency while keeping fake rates acceptable. Initial iterations prioritize high transverse momentum tracks to rapidly remove their corresponding hits from the available pool. The baseline sequence predominantly employs two primary iterative steps:

- **initialStep:** Uses pixel quadruplets to seed prompt tracks with moderate to high p_T .

- **highPtTripletStep:** Uses hit triplets to recover tracks that may miss pixel-layer measurements, focusing on high- p_T kinematics.

The Combinatorial Kalman Filter

The Combinatorial Kalman Filter (CKF) is the traditional algorithmic foundation for both track building (pattern recognition) and track fitting in CMS. It extends the Kalman filter [51], a recursive, statistically optimal linear estimator, with a combinatorial branching strategy that simultaneously resolves which hits belong to a given track. Its advantage over a global least-squares fit is that it is recursive: rather than inverting a single large matrix relating all measurements at once, it processes the detector layers one at a time, refining the trajectory estimate and incorporating material effects locally as each new measurement is encountered. This makes it natural to interleave the estimation of the trajectory with the decision of which hits to attach to it.

Track model and the Kalman recursion

The trajectory is represented at each layer k by a state vector $\mathbf{x}_k$ of the five local track parameters (Section 3.2.1) and a covariance $\mathbf{C}_k$. A *system model* $\mathbf{x}_k = \mathbf{f}_k(\mathbf{x}_{k-1}) + \mathbf{w}_k$ describes the deterministic propagation $\mathbf{f}_k$ of the helix through the magnetic field between layers, plus a stochastic process-noise term $\mathbf{w}_k$ (covariance $\mathbf{Q}_k$) that encodes material effects (Section 3.2.1): the angular smearing from multiple Coulomb scattering, whose width follows the Highland approximation

$$\theta_0 = \frac{13.6 \text{ MeV}}{\beta c p} z \sqrt{\frac{x}{X_0}} \left[1 + 0.038 \ln \left(\frac{x}{X_0} \right) \right], \quad (3.1)$$

and the curvature change from ionization energy loss (Eq. (2.9)). A *measurement model* $\mathbf{m}_k = \mathbf{H}_k \mathbf{x}_k + \varepsilon_k$ relates the state to the measured coordinates through the projection matrix $\mathbf{H}_k$, with measurement error ε_k of covariance $\mathbf{V}_k$ set by the sensor resolution.

The filter then proceeds recursively. In the **prediction** step the filtered state at layer $k-1$ is extrapolated to layer k using the linearized propagator $\mathbf{F}_k = \partial \mathbf{f}_k / \partial \mathbf{x}$, giving $\mathbf{x}_{k|k-1} = \mathbf{F}_k \mathbf{x}_{k-1|k-1}$

with its covariance inflated by $\mathbf{Q}_k$. In the **filtering** step a candidate measurement is compared with this prediction through the residual $\mathbf{r}_k = \mathbf{m}_k - \mathbf{H}_k \mathbf{x}_{k|k-1}$ and its covariance $\mathbf{S}_k$, and the state is corrected toward the measurement,

$$\mathbf{x}_{k|k} = \mathbf{x}_{k|k-1} + \mathbf{K}_k \mathbf{r}_k, \quad \mathbf{K}_k = \mathbf{C}_{k|k-1} \mathbf{H}_k^\top \mathbf{S}_k^{-1}, \quad (3.2)$$

where the Kalman gain $\mathbf{K}_k$ optimally weights prior knowledge against the new measurement. The compatibility of a hit with the trajectory is quantified by the χ^2 increment,

$$\chi_k^2 = \mathbf{r}_k^\top \mathbf{S}_k^{-1} \mathbf{r}_k, \quad (3.3)$$

which defines the search window used to collect candidate hits and accumulates into the total track χ^2 used for quality selection. After the forward pass, a backward **smoothing** pass combines the forward and backward estimates so that the result at every layer reflects all n measurements; the smoothed parameters at the point of closest approach are the final track parameters reported to analyzes.

Combinatorial track building and modern algorithms

In pattern recognition the assignment of hits to a track is not known in advance, and this is where the *combinatorial* extension enters. Starting from a seed (Section 3.2.1), the filter is propagated outward, and on each layer every hit inside the search window defined by Eq. (3.3) is treated as compatible. When more than one hit qualifies, the candidate bifurcates into competing hypotheses, one per hit, as illustrated in Figure 3.4; an additional no-hit branch is retained so that detector inefficiencies do not prematurely terminate a track. To prevent the branches from growing exponentially, candidates are continually pruned: only a bounded number of the lowest- χ^2 hypotheses survive to the next layer, and branches exceeding χ^2 or missing-hit thresholds are discarded. Finally, an ambiguity-resolution step removes candidates that share too many hits, suppressing duplicate and fake tracks.

The CKF is statistically optimal for Gaussian noise and treats material effects rigorously, which is why it has long served as the CMS workhorse. Its limitations motivate the modernization described earlier: the Gaussian assumption is violated by the heavy radiative tails of electron energy loss (better handled by the Gaussian-Sum Filter), and, more consequentially for the HL-LHC, the combinatorial branching is intrinsically sequential and its cost grows steeply with pileup, making the CKF one of the dominant contributors to reconstruction time. This is precisely the motivation for the vectorized and parallel approaches, mkFit, Patatrack, and LST (Section 5.1), which restructure the same Kalman-filter mathematics to exploit the SIMD units of CPUs and the parallelism of GPUs.

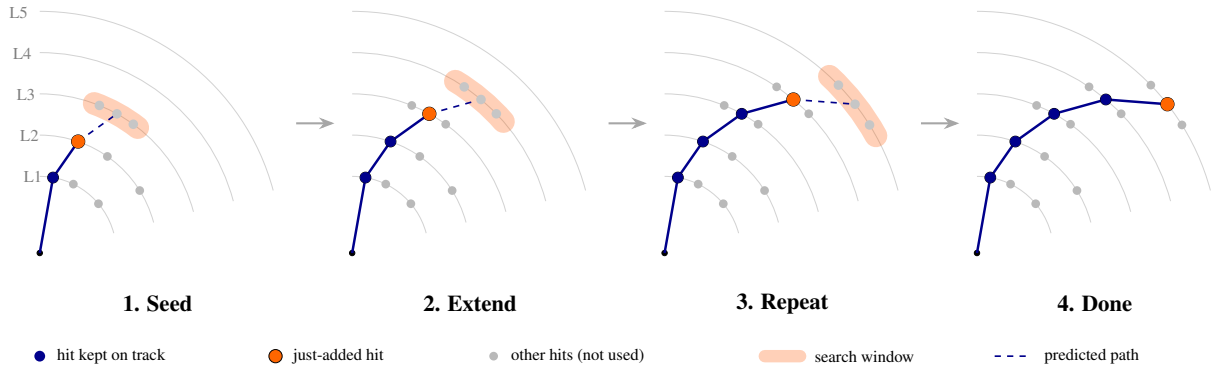


Figure 3.4. Track building with the combinatorial Kalman filter, shown in a transverse slice of the tracker

3.2.2 Particle Flow

The particle-flow (PF) algorithm [31] is the cornerstone of event reconstruction at CMS. Rather than treating each subdetector independently, PF combines the measurements from all of them, tracker (Section 2.2), ECAL (Section 2.2), HCAL (Section 2.2), and muon system (Section 2.2), into a single, coherent list of the individual final-state particles in the event: charged hadrons, neutral hadrons, photons, electrons, and muons, as showcased in Figure 3.5

The strategy is to link the elements observed in different subdetectors that originate from the same particle: a track is associated with the calorimeter clusters it points to, and the type

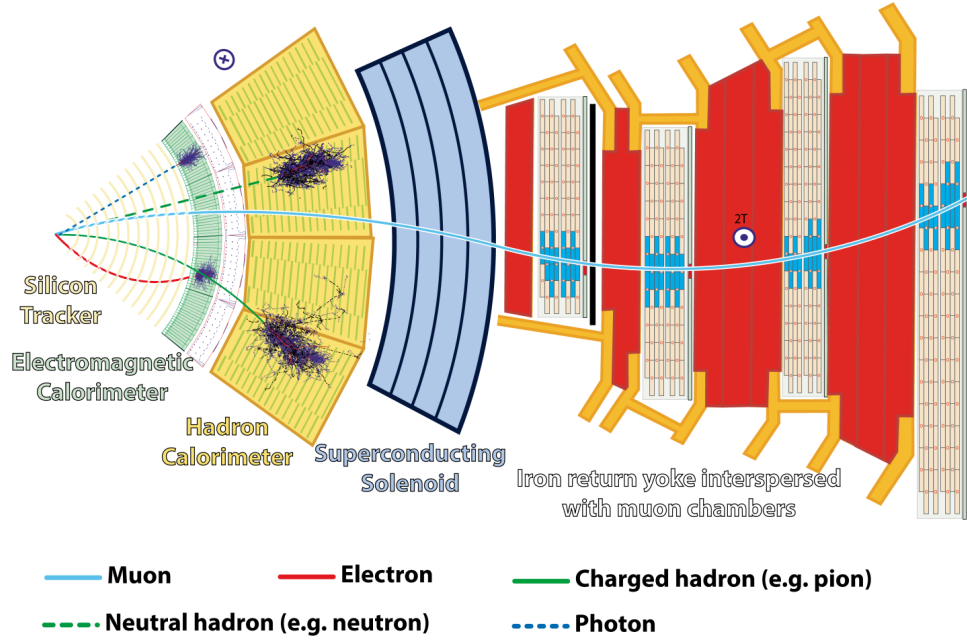


Figure 3.5. Schematic representation of the deposits in the CMS detector and the corresponding reconstructed objects [20]

of particle is then inferred from the combination. A track matched to an ECAL cluster with compatible energy is an electron; a track matched to an HCAL cluster is a charged hadron, with its momentum taken from the precise tracker measurement rather than the coarser calorimeter; a calorimeter cluster with no associated track is a photon (ECAL) or a neutral hadron (HCAL). This approach exploits the strengths of each subdetector and is especially powerful in CMS because the strong solenoidal field (Section 2.2) separates charged from neutral particles and the finely segmented tracker and ECAL allow individual particles to be resolved. The PF candidates form the common input from which all higher-level physics objects are built.

3.2.3 Reconstructed Objects

The particle-flow candidates are combined and interpreted to form the higher-level objects used directly in physics analyzes.

Jets

Quarks and gluons produced in a collision hadronize and appear in the detector as jets, collimated sprays of particles (Section 3.1). Jets are reconstructed by clustering the PF candidates with the anti- k_t algorithm [26], a sequential-recombination algorithm defined by the distance measures

$$d_{ij} = \min\left(p_{T,i}^{-2}, p_{T,j}^{-2}\right) \frac{\Delta R_{ij}^2}{R^2}, \quad d_{iB} = p_{T,i}^{-2}, \quad (3.4)$$

between every pair of candidates i, j (with $\Delta R_{ij}^2 = \Delta\eta_{ij}^2 + \Delta\phi_{ij}^2$) and between a candidate i and the beam B . At each step the smallest distance is found: a d_{ij} merges the two candidates, while a d_{iB} promotes candidate i to a jet and removes it. The negative power of p_T makes the clustering proceed outward from the hardest candidates, so that soft radiation is gathered around hard cores; the algorithm is consequently infrared and collinear safe and produces regular, nearly circular jets in the η - ϕ plane (Section 2.2), with a characteristic size set by the radius parameter R . The measured jet energy is corrected to the true particle-level scale through jet energy scale and resolution calibrations. A significant challenge is pileup: the additional soft pp collisions that overlap the hard-scatter interaction in the same and neighboring bunch crossings deposit extra particles that contaminate the jet and bias its energy, mass, and substructure. CMS suppresses this contamination at the level of the PF candidates, before the anti- k_t clustering, and the algorithms used include Charged-hadron subtraction (CHS) and pileup per particle identification (PUPPI).

Charged-hadron subtraction (CHS).

Charged-hadron subtraction [31] exploits the tracker's ability to associate each charged PF candidate with a reconstructed vertex: charged hadrons that are unambiguously assigned to a pileup vertex, rather than to the primary vertex of the hard scatter, are removed from the PF collection before clustering. CHS therefore eliminates the charged component of pileup exactly, but leaves the neutral component, which carries no tracking information, to be corrected only on average through the jet energy calibration. This residual neutral pileup limits the jet energy and

mass resolution, particularly at high pileup.

Pileup per particle identification (PUPPI).

PUPPI [22] generalizes the subtraction to neutral particles by assigning every PF candidate a weight between 0 and 1 that quantifies the probability that it originates from the primary interaction. The weight is computed from a local shape variable, built from the momenta and angular distances of the surrounding particles, that discriminates the collimated radiation of the hard scatter from the diffuse deposits of pileup, combined with the vertex compatibility of charged particles. Each candidate's four-momentum is rescaled by its weight before clustering, so that pileup particles, charged and neutral alike, are smoothly suppressed rather than removed by a binary decision. PUPPI thereby yields a more stable jet energy scale, mass, and substructure response at high pileup, which is especially important for the boosted large-radius jets on which this analysis relies.

Jets originating from b quarks are identified ("b-tagged") by exploiting the long lifetime of b hadrons, which produces tracks with large impact parameters and displaced secondary vertices measured relative to the beam spot (Section 2.1). Modern taggers are deep neural networks that combine these track and vertex features to distinguish the flavor of a jet: small-radius jets are tagged with dedicated b -jet identification networks such as DEEPJET [24] and the unified particle-transformer (UPART) tagger, while the flavor of boosted large-radius jets is identified with jet-level networks such as PARTICLENET [74] and GLOPART. The specific taggers used in this analysis are detailed in the analysis chapter. Robust b tagging and boosted-jet flavor tagging are essential to the analysis presented in this thesis.

The radius parameter R of the anti- k_t algorithm sets the angular size of the reconstructed jet and is chosen according to the topology being targeted. CMS uses two standard values. **AK4 jets**, clustered with $R = 0.4$, are the default choice for resolving individual quarks and gluons: when the decay products of a heavy particle are well separated, each parton is captured in its own narrow jet, and AK4 jets are the objects that carry the b -tagging information described above.

AK8 jets, clustered with $R = 0.8$ and often called "fatjets," are large-radius jets designed for boosted topologies. When a massive particle such as a W , Z , Higgs, or top is produced with large momentum, its hadronic decay products become collimated into a single wide cone; a fatjet captures the entire decay, so that its mass and internal substructure reflect those of the parent particle.

Two classes of substructure observable characterize a fatjet. The first is the jet mass, which for a genuine boosted decay peaks at the mass of the parent particle rather than falling smoothly as for QCD jets. The raw jet mass is, however, badly distorted by soft and wide-angle radiation and by pileup, so a grooming procedure is applied: the soft-drop algorithm [63] recursively removes soft, wide-angle constituents and retains the hard core of the jet, yielding a groomed "soft-drop mass" m_{SD} that provides a clean, well-resolved mass peak. The second is N -subjettiness [81], a set of jet-shape variables τ_N that measure how consistent a jet is with having N subjets,

$$\tau_N = \frac{1}{\sum_{i=1}^{n_{\text{constituents}}} p_{\text{T},i} R} \sum_{i=1}^{n_{\text{constituents}}} p_{\text{T},i} \min\{\Delta R_{1,i}, \Delta R_{2,i}, \dots, \Delta R_{N,i}\}, \quad (3.5)$$

where the sum runs over the jet constituents, R is the jet radius, and $\Delta R_{k,i}$ is the angular distance between constituent i and the k -th subjet axis. By construction τ_N approaches zero when the constituents are all aligned with N (or fewer) subjet axes, and is large otherwise. A single hard prong therefore gives a small τ_1 , a genuine two-prong decay a small τ_2 , and so on, so it is the ratios that discriminate, in particular $\tau_{21} = \tau_2/\tau_1$, which is small for the two-prong decays of W/Z /Higgs bosons, and $\tau_{32} = \tau_3/\tau_2$, which is small for the three-prong decays of top quarks, whereas single-pronged QCD jets give larger values. These observables, together with the large-radius flavor taggers noted above, identify the origin of a fatjet and separate genuine boosted decays from ordinary QCD jets. The boosted, large-radius jets on which this analysis relies are AK8 fatjets, tagged using the groomed mass and these substructure variables.

Leptons

Charged leptons are reconstructed by combining tracking and calorimeter information.

- **Electrons:** Electrons are identified as a track matched to an electromagnetic cluster; because electrons radiate heavily by bremsstrahlung (Section 2.2), their tracks are fit with the Gaussian-sum filter, which accommodates the non-Gaussian energy loss that the standard Kalman filter (Section 3.2.1) cannot.
- **Muons:** Muons are reconstructed by associating a track in the inner tracker with hits in the outer muon system (Section 2.2), the two combined to yield excellent momentum resolution over a wide range.
- **Taus:** Tau leptons that decay leptonically are reconstructed as electrons or muons, while their hadronic decays are reconstructed from the PF candidates using dedicated algorithms that resolve the characteristic one- and three-prong decay topologies.

In all cases an isolation requirement, a limit on the additional energy within a cone of radius ΔR around the lepton, is used to distinguish prompt leptons from those produced inside jets.

Missing Transverse Energy

Particles that do not interact with the detector, such as neutrinos and many hypothesized dark-matter candidates, escape undetected but reveal themselves through momentum conservation in the transverse plane. Because the colliding partons carry negligible net transverse momentum (Section 2.2), the vector sum of the transverse momenta of all reconstructed particles should be zero; any imbalance defines the missing transverse momentum, whose magnitude is the missing transverse energy,

$$\vec{p}_T^{\text{miss}} = -\sum_i \vec{p}_{T,i}, \quad E_T^{\text{miss}} = \left| \vec{p}_T^{\text{miss}} \right|, \quad (3.6)$$

where the sum runs over all PF candidates in the event. A significant E_T^{miss} can signal the production of invisible particles and is a key observable in many searches for new physics. Because it is computed from every object in the event, its accurate reconstruction depends on the quality of the entire particle-flow description and is particularly sensitive to pileup and to mismeasured jets.

Chapter 4

Anomalous Higgs Boson Production

4.1 Vector Boson Scattering (VBS)

The Higgs mechanism developed in Section 1.4 does more than generate the masses of the weak bosons and fermions: it rests on a specific assumed shape for the scalar potential, and that shape is itself a prediction that remains only partially tested. Expanding the Higgs potential of Eq. (1.19) about the vacuum $\langle\phi^0\rangle = v/\sqrt{2}$, with $v \approx 246$ GeV, in unitary gauge $\phi = \frac{1}{\sqrt{2}}(0, v + H)^T$, generates beyond the quadratic mass term $\frac{1}{2}m_H^2 H^2$, cubic and quartic self-interactions of the physical Higgs boson,

$$V(\phi) \supset \lambda v H^3 + \frac{\lambda}{4} H^4, \quad \lambda = \frac{m_H^2}{2v^2}. \quad (4.1)$$

In the SM the self-coupling λ is one of the two parameters of the Higgs sector (Section 1.6); once the Higgs mass m_H and the vev v are measured, the trilinear coupling λv is no longer free but *predicted*. Measuring the Higgs self-interaction directly therefore overconstrains the theory and provides a sharp test of whether the observed boson is truly the field responsible for electroweak symmetry breaking. The tightest current constraints on λ come from searches for di-Higgs boson production, which proceeds through gluon-gluon fusion and vector boson fusion, but these measurements remain limited by the small cross section and the large irreducible backgrounds.

The covariant derivative of the Higgs field, Eq. (1.17), which as we discussed, generates

the $W^\pm$ and Z masses in Section 1.4, simultaneously fixes the Higgs gauge couplings of the form $g^2 v (HW^+W^-$ and the quartic $HHVV$ vertices). The same non-Abelian $SU(2)_L \times U(1)_Y$ structure that makes the weak bosons self-interacting (Section 1.3) produces the triple and quartic gauge boson vertices; the key distinction from the Higgs-gauge couplings is the appearance of a factor of the vacuum expectation value v relating the two. A direct consequence of the W and Z acquiring mass is that they develop a longitudinal polarization mode, absent for massless gauge bosons. The amplitude for scattering longitudinally polarized weak bosons, computed in the framework of Section 1.7, grows with the center-of-mass energy as

$$\mathcal{M}(V_L V_L \rightarrow V_L V_L) \sim \frac{E^2}{v^2}, \quad (4.2)$$

which in the absence of the Higgs boson would violate perturbative unitarity above the TeV scale. In the SM, diagrams involving s - and t -channel Higgs boson exchange produce cancellations that tame this growth order by order, restoring unitarity and making VBS a direct precision test of the mechanism of electroweak symmetry breaking. Historically, this unitarity argument provided a firm upper bound on the Higgs boson mass and helped motivate the energy reach of the LHC.

Vector boson scattering is, in this sense, a model-independent probe of the electroweak sector: any deviation from the SM that modifies the gauge boson or Higgs interactions would spoil the delicate cancellation in Eq. (4.2), causing the VBS cross section to grow anomalously with energy rather than remaining well-behaved, a signature distinct from any specific new-physics model. By the Goldstone boson equivalence theorem, longitudinal VBS at high energies is equivalent to the scattering of the would-be Goldstone bosons of spontaneous symmetry breaking, giving a direct experimental window onto the scalar sector itself. VBS is simultaneously sensitive to the Higgs self-coupling of Eq. (4.1) and to the triple and quartic gauge couplings, and it forms an important background to a wide range of beyond-the-Standard-Model searches (Section 1.9), so its precise understanding is essential to the broader physics program of the LHC.

VBS has been established experimentally at both CMS and ATLAS in several final states. VBS with the associated production of a Higgs boson is a qualitatively richer process that is being probed for the first time at the LHC. It is sensitive to the HVV couplings that underpin the Higgs mechanism and provides constraints complementary to those from inclusive Higgs and di-Higgs searches, while simultaneously probing the quartic gauge-Higgs vertex that appears at one higher order in the electroweak expansion. As any modification of the electroweak interactions discussed above would manifest as an anomalous enhancement of this rare process, VBS with associated Higgs production is a direct experimental handle on the anomalous Higgs boson production that is the subject of this dissertation.

4.2 Associated VVH Production and the Higgs Couplings

The process studied in this dissertation is the production of a Higgs boson in association with two vector bosons through vector boson scattering, $pp \rightarrow VVHjj$ with $V = W, Z$, illustrated in Figure 4.1. Two quarks each radiate a vector boson; these scatter and, in the diagrams of interest, radiate a Higgs boson, leaving the two spectator quarks as the forward “VBS” jets that tag the process [46, 56]. The crucial feature of this final state is that it provides simultaneous, tree-level access to two of the least constrained couplings of the Higgs sector: the trilinear Higgs self-coupling (through diagrams with s -channel Higgs exchange, Eq. (4.1)) and the quartic $VVHH$ coupling (through the contact vertex). Deviations of these couplings from their SM values are parametrized in the κ -framework by the modifiers κ_λ and κ_{2V} , respectively, defined such that $\kappa = 1$ recovers the SM. The HVV couplings, parametrized by κ_V , also enter and are fixed to their precisely measured SM values.

The SM cross section for VBS VVH production at $\sqrt{s} = 13.6$ TeV is only about 1.68 fb at leading order (Table 4.1), making it one of the rarest electroweak processes accessible at the LHC. Its experimental appeal, however, lies in how strongly it responds to new physics: because the amplitude depends on κ_{2V} , departures from $\kappa_{2V} = 1$ grow the cross section quadratically

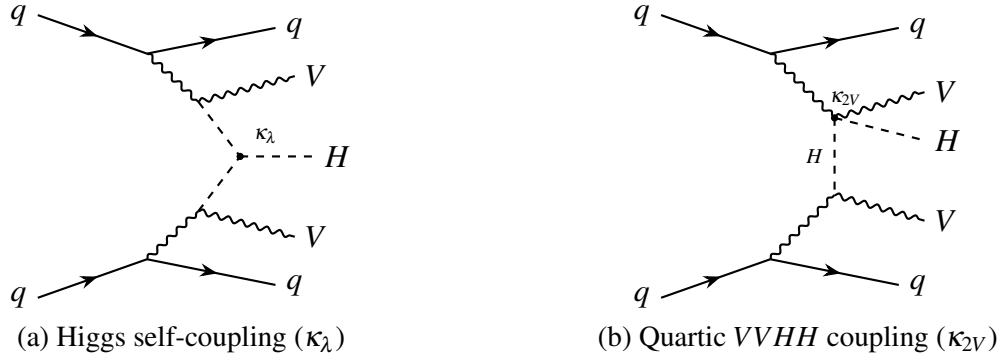


Figure 4.1. Representative tree-level VVH vector-boson-scattering diagrams showing the dependence on κ_λ (left) and κ_{2V} (right).

and substantially boost the final-state objects, sharpening the experimental sensitivity precisely in the regime where new physics would appear. The process is therefore complementary to Higgs pair production, which currently provides the tightest constraints on these couplings: the CMS and ATLAS di-Higgs combinations restrict κ_{2V} to roughly $[0.64, 1.40]$ and $[0.6, 1.5]$, respectively [38, 14], while the individual κ_{2W} and κ_{2Z} modifiers are only loosely bounded by the CMS search for VHH production [37]. Because the change in cross section and kinematics with κ_λ is far milder than with κ_{2V} , the κ_{2V} coupling is the primary target, with the self-coupling explored through the more abundant resolved topologies.

The scale of that number is best appreciated against the full range of processes CMS has measured, shown in Figure 4.2: from the inelastic proton-proton cross section down to the rarest electroweak diboson and multiboson final states, the summary spans some fourteen orders of magnitude, and a few femtobarns places VBS VVH at its very bottom edge, alongside the electroweak same-sign $W^\pm W^\pm$ and triboson processes. For every VBS VVH event the LHC produces on the order of 10^{11} inelastic collisions. That ratio is the essential experimental difficulty of this search, and the reason the analysis of Chapter 4 leans so heavily on boosted-object tagging to isolate the signal and on a data-driven estimate of the background that survives it.

This sensitivity has a clean theoretical origin that motivates the entire measurement. As

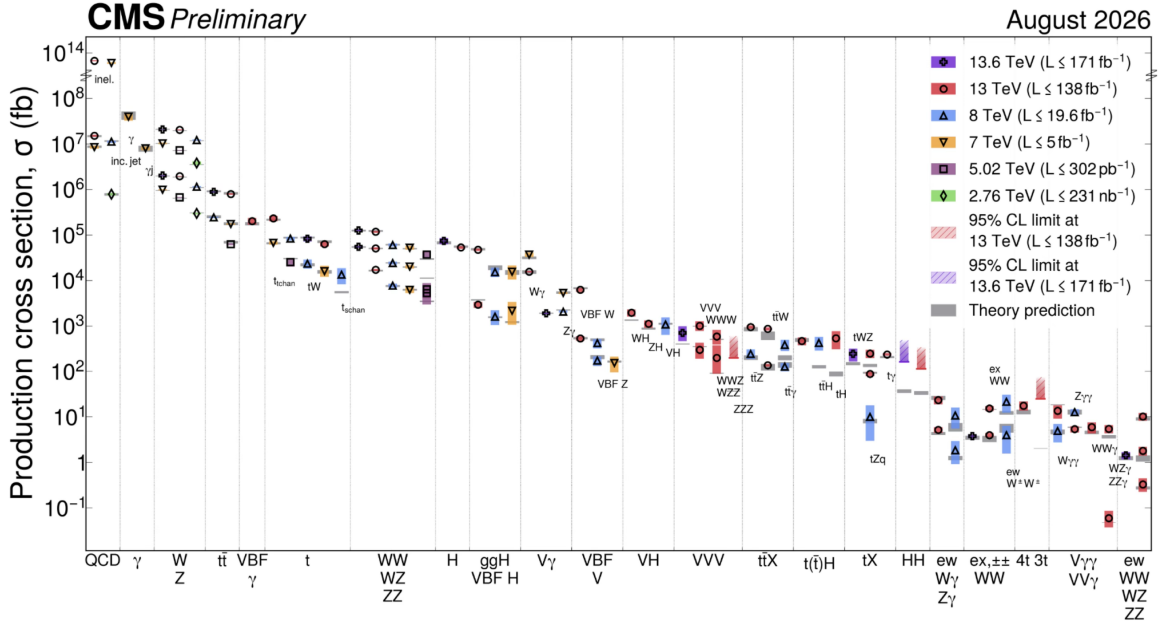


Figure 4.2. Summary of production cross sections measured by CMS, from the inelastic proton-proton cross section to the rarest electroweak processes, at center-of-mass energies from 2.76 to 13.6 TeV.

emphasized by Henning, Lombardo, Rimbau, and Riva [56], modifying the $VVHH$ coupling away from its SM value spoils precisely the cancellation that unitarizes longitudinal vector boson scattering, so that the amplitude no longer remains well behaved but instead grows with the partonic center-of-mass energy as in Eq. (4.2). The deviation therefore imprints itself most strongly not on the total rate but on the high-energy tail of the distribution, where the energy growth accumulates and boosts the final-state bosons and Higgs. In effect, the off-shell, high-energy behavior of the VBS amplitude measures the Higgs-sector coupling κ_{2V} directly: a determination of a Higgs coupling that does not rely on reconstructing an on-shell Higgs resonance, and one that is complementary and at times competitive with on-shell Higgs and di-Higgs measurements [56]. This is the physics rationale for targeting the boosted, high-invariant-mass regime of VBS VVH production, where the sensitivity to κ_{2V} is concentrated and which the merged-jet channels of this analysis are designed to exploit.

4.2.1 Limitations of the κ -framework

The κ -framework used above is a convenient benchmark, but it is not derived from a gauge-invariant Lagrangian; it instead rescales on-shell physical amplitudes by real multiplicative factors at fixed kinematics (m_H). Because the SM Higgs-gauge boson couplings originate from the single covariant-derivative term $|D_\mu H|^2$ (Eq. (1.17)), they are intrinsically correlated with the gauge boson self-couplings and kinetic terms; no local, gauge-invariant Lagrangian reproduces an independently rescaled HVV vertex ($\kappa_V \neq 1$) while leaving the rest of the electroweak sector unchanged. A direct physical consequence is that $\kappa_V \neq 1$ reintroduces the unitarity-violating high-energy growth in longitudinal vector boson scattering that is otherwise cancelled by Higgs exchange in the SM (Eq. (4.2)),

$$\mathcal{M}(W_L W_L \rightarrow W_L W_L) \sim (1 - \kappa_V^2) \frac{s}{v^2}, \quad (4.3)$$

since the cancellation requires exactly $\kappa_V = 1$. The κ -framework can parametrize such a deviation but provides no mechanism for the new physics required to restore unitarity, nor any prediction for the energy scale or kinematic dependence of the associated effects, the very information that VBS measurements such as the one presented here probe directly.

The κ modifiers are nonetheless retained throughout this dissertation: they map directly onto the previous CMS result and onto the di-Higgs and VHH combinations against which these constraints are benchmarked (Section 4.4.4), and the analysis varies a single coupling at a time rather than performing a global operator fit. The κ -framework is thus best understood as a leading-order benchmark for consistency with the SM; where a gauge-invariant interpretation is required, the limits reported here can be reinterpreted as constraints on the corresponding SMEFT Wilson coefficient.

This dissertation builds directly on the first search for this process, CMS HIG-24-003 [5, 40], which established the channel using the full Run 2 dataset. How that search was performed,

and how the analysis developed here extends it, is the subject of the remainder of this chapter.

4.3 Overview of the Search

The preceding sections established the theoretical motivation for why VBS VVH production measurements are a worthwhile effort; the remainder of this chapter turns to how the search is carried out. A search at the LHC is the attempt to establish, or constrain, the existence of a rare process amid an overwhelmingly larger flux of known Standard Model reactions. The process of interest is typically produced only a handful of times in a dataset of billions of collisions, and its experimental signature is mimicked by far more abundant backgrounds. A search is therefore not a single measurement but a carefully constructed chain of inference, designed to enhance the sensitivity to the target process while keeping the background under quantitative control, so that a statistically meaningful statement can be made about whether the process is present and at what rate.

This chain proceeds in three broad stages:

First, the data is collected by the trigger and reconstructed into physics objects, jets, leptons, and missing transverse momentum, using the algorithms detailed in the preceding chapter; simulated samples of the signal and the principal backgrounds are produced in the same format to guide the design of the analysis.

Second, the events are selected according to their topology and kinematics to isolate a signal-rich region: requirements on the reconstructed objects define one or more channels matched to the ways the process can manifest, and the events are further classified, often with multivariate discriminants, to concentrate the signal against the background. Third, the signal is extracted from the remaining background through a statistical analysis: the background in the signal region is estimated, preferably from data where the simulation is unreliable, the systematic uncertainties are assessed, and a likelihood fit across all regions and channels yields the measured rate or, in the absence of a signal, an exclusion limit.

The remainder of this chapter applies this scheme to the search for VBS VVH production. It first summarizes the previous Run 2 result (Section 4.4), and then presents in detail the improved Run 2 + Run 3 analysis developed in this dissertation (Section 4.5), following the three stages above: the data and simulated samples and the objects built from them (Sections 4.5.1-4.5.2), the channel definitions and event selection (Section 4.5.3), and the data-driven background estimation, signal extraction, and resulting sensitivity (Sections 4.5.4-4.5.11).

4.4 The Previous Run 2 Search

The first search for VBS VVH production, CMS HIG-24-003 [5, 40], established the process as an experimental channel using the full Run 2 dataset. Because the analysis developed in this dissertation builds directly upon it, its strategy and result are summarized here before the improved Run 2 + Run 3 analysis is presented in Section 4.5. The new analysis is not merely an update: it reprocesses the data in a newer format with a substantially different set of reconstructed objects, so the two searches are described separately throughout this chapter.

4.4.1 Datasets and Objects

The previous search analyzed the 138 fb^{-1} of pp collisions recorded by CMS in 2016-2018 at $\sqrt{s} = 13 \text{ TeV}$ (Section 2.1), reconstructed with an earlier set of object definitions. Its physics objects differ from those of the present analysis: jets were clustered from charged-hadron-subtracted (CHS) particle-flow candidates, b jets were identified with the DEEPJET tagger [24], and the boosted $H \rightarrow b\bar{b}$ and $V \rightarrow q\bar{q}$ decays, each reconstructed as a single large-radius (AK8) jet, were tagged with PARTICLENET [74].

4.4.2 Analysis Strategy and Event Selection

Events were partitioned into mutually exclusive all-hadronic, single-lepton, and multi-lepton channels according to the number of reconstructed leptons and AK8 jets, covering the hadronic and leptonic decays of the two vector bosons while preserving orthogonality, with

the fully-boosted all-hadronic channel and the fully-boosted single-lepton channels providing the bulk of the sensitivity. Two forward jets with a large pseudorapidity separation $|\Delta\eta_{jj}|$ and invariant mass m_{jj} tagged the VBS production. The coupling dependence was probed through MadGraph matrix-element reweighting, which scanned κ_{2V} continuously within a single simulated sample.

4.4.3 Background Estimation and Signal Extraction

The dominant backgrounds, QCD multijet production in the all-hadronic channel, and $t\bar{t}$ and W +jets production in the leptonic channels, were estimated directly from data using the same data-driven ABCD strategy adopted also by the present analysis (Section 4.5.4). Each channel defined a single signal region at high values of the discriminating tagger and classifier scores, and the couplings were extracted from a simultaneous binned maximum-likelihood fit of the signal and control regions across all channels, performed with the CMS COMBINE tool [36] using the statistical methods described in Section 4.5.8.

4.4.4 Results

The previous Run 2 search established the first observation of the VBS VVH process as an experimental channel and set the strongest constraints to date on the $VVHH$ coupling from a single measurement. Combining all channels, and assuming SM values for the remaining Higgs couplings, the VBS VVH production is excluded at 95% confidence level for values of κ_{2V} outside the observed (expected) interval

$$0.40 < \kappa_{2V} < 1.60 \quad (0.34 < \kappa_{2V} < 1.66), \quad (4.4)$$

as shown in Figure 4.3. The individual quartic modifiers are separately constrained to the observed (expected) ranges $0.17 < \kappa_{2W} < 1.84$ ($0.11 < \kappa_{2W} < 1.89$) and $-0.37 < \kappa_{2Z} < 2.38$ ($-0.54 < \kappa_{2Z} < 2.54$), and a two-dimensional likelihood scan in the κ_{2W} - κ_{2Z} plane substantially

improves on the constraints from the VHH search. These results are complementary to, and competitive with, the di-Higgs determinations of κ_{2V} .

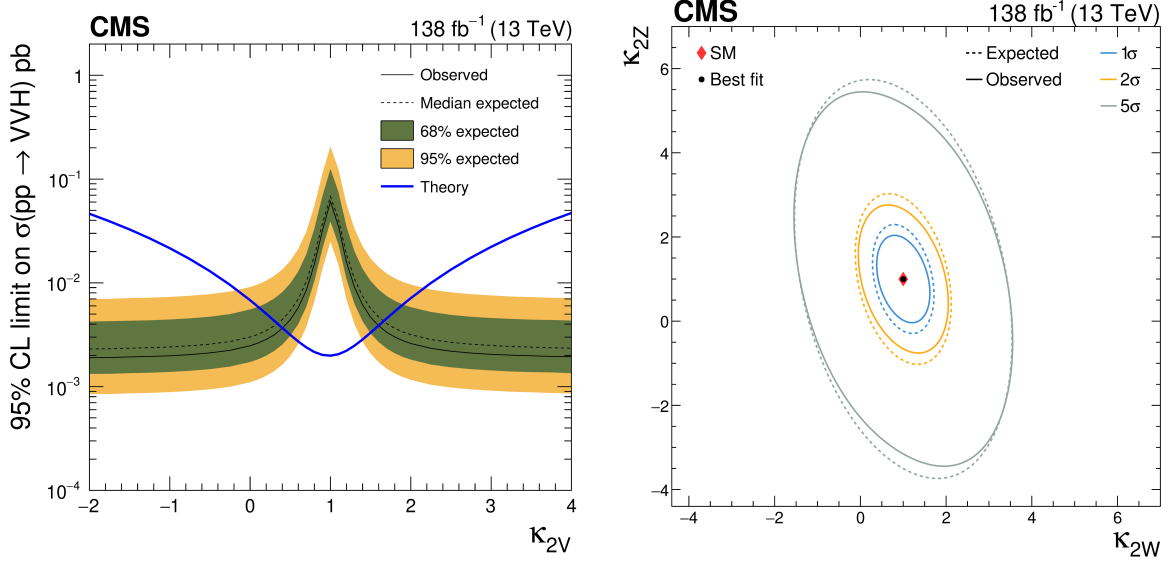


Figure 4.3. Previous Run 2 search: 95% CL upper limit on the VBS VVH cross section versus κ_{2V} (left) and the exclusion contour in the $(\kappa_{2W}, \kappa_{2Z})$ plane (right) [5, 40].

4.5 The Run 2 + Run 3 Analysis

The analysis developed in this dissertation reprocesses the full Run 2 data together with the partial Run 3 sample in a common, newer reconstruction format and analyzes them jointly. It differs from the previous search in almost every reconstructed object, adopting PUPPI jets and the far more performant Particle Transformer taggers in place of the CHS jets, DEEPIET, and PARTICLENET used in the previous iteration of the analysis, while retaining the same data-driven ABCD background strategy, now extended from a single signal region per channel to many. It also broadens the analysis into a larger set of channels, adding semi-resolved topologies that recover less-boosted events, and scans the Higgs self-coupling κ_λ in addition to the primary κ_{2V} target. The remainder of this chapter describes this analysis in detail.

4.5.1 Data and Simulated Samples

The analysis uses pp collision data recorded by CMS, reprocessed with a common, up-to-date reconstruction so that a single set of object definitions and selections applies identically across both data-taking eras. The Run 2 sample, collected in 2016-2018 at $\sqrt{s} = 13$ TeV (Section 2.1), corresponds to an integrated luminosity of 138 fb^{-1} ; the partial Run 3 sample, recorded in 2022-2025 at $\sqrt{s} = 13.6$ TeV, adds 282.79 fb^{-1} , about twice the Run 2 dataset, for a combined sample about three times larger than the previous search.

Signal Samples

The signal process is generated at leading order with MadGraph5_aMC@NLO [7] as four sub-processes distinguished by the pair of vector bosons, opposite-sign $W^{\pm}W^{\mp}H$, same-sign $W^{\pm}W^{\pm}H$, WZH , and ZZH , each produced in association with the two VBS jets. An additional generator-level requirement $m_{jj} > 100 \text{ GeV}$ is imposed on the jet pair in order to suppress diagrams with the same final state but without the VBS topology. The matrix-element events are interfaced to Pythia8 [79] with the CP5 tune [33] for parton showering and hadronization (Section 3.1), using the dipole-recoil scheme, and the NNPDF 3.1 parton distribution functions [68]. The detector response is modeled with Geant4 [4].

The distinct kinematic imprints of the two couplings are already visible at generator (LHE) level in Figure 4.4, which compares the two distributions across the benchmark points. The pseudorapidity separation of the two tagging jets, $\Delta\eta_{jj}$ (left), peaks at large values for every benchmark, the characteristic wide rapidity gap of vector boson scattering, and is essentially independent of the couplings, since the forward-jet topology is a feature of the production mechanism rather than of the Higgs couplings; this is what makes $\Delta\eta_{jj}$ and m_{jj} robust, coupling-agnostic tagging variables. The scalar p_T sum of the three bosons (right) behaves oppositely: raising κ_{2V} above unity shifts the distribution to markedly higher momenta, reflecting the energy growth of the amplitude once the unitarity-restoring cancellation is spoiled (Section 4.2), whereas anomalous κ_{λ} leaves the spectrum close to the SM shape. Anomalous κ_{2V} therefore boosts the

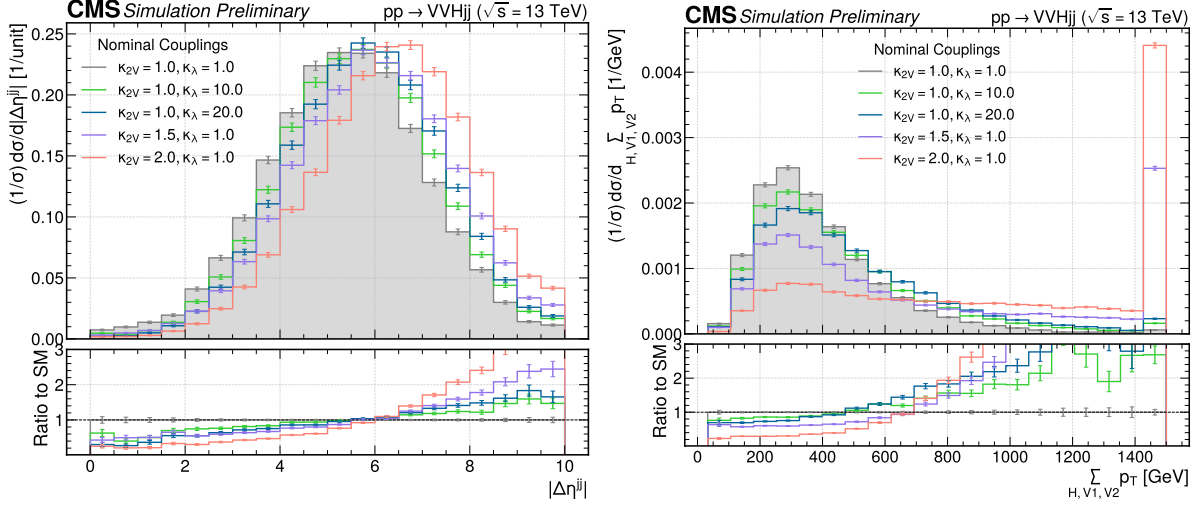


Figure 4.4. LHE-level $\Delta\eta_{jj}$ (left) and scalar p_T sum of the three bosons (right) at several $(\kappa_\lambda, \kappa_{2V})$ benchmark points.

final-state bosons into the merged, large-radius-jet regime, which is precisely why the boosted channels drive the sensitivity to κ_{2V} while κ_λ is better accessed through the more resolved topologies.

The coupling dependence is probed by generating each sub-process at several benchmark points in $(\kappa_{2V}, \kappa_\lambda)$. Because the VBS VVH amplitude grows with energy when the couplings depart from their SM values (Section 4.2), the cross section rises steeply away from the SM point: raising κ_{2V} to 1.5 roughly doubles the rate, while $\kappa_\lambda = 10$ enhances the opposite-sign WWH and ZZH channels by factors of a few, as summarized in Table 4.1. The previous Run 2 result relied on MadGraph matrix-element reweighting to scan κ_{2V} continuously within a single sample, but studies for this dissertation found the reweighting to be unreliable below $\kappa_{2V} \approx 1.5$ with the updated generator settings; the present analysis therefore generates dedicated samples at discrete coupling points, and the selection is optimized near $\kappa_{2V} = 1.5$, a value close to the expected exclusion that enhances the boosted signature.

Table 4.1. Leading-order cross sections (fb) of the four VBS VVH sub-processes at $\sqrt{s} = 13.6$ TeV ($m_{jj} > 100$ GeV).

Sub-process	(1, 1) (SM)	(1.5, 1)	(1, 10)
$W^\pm W^\mp H$ (OS)	0.723	1.277	1.789
$W^\pm W^\pm H$ (SS)	0.359	0.736	0.508
WZH	0.474	0.825	0.711
ZZH	0.125	0.438	0.469

Background Samples

A comprehensive set of simulated background processes, covering every reaction that can mimic the VVH plus forward-jet signature, is used to train and optimize the analysis; as described in Section 4.5.4, the final background yield in the signal regions is then taken from data rather than from this simulation. The samples fall into five categories:

- **QCD multijet**, generated at leading order with MadGraph5_aMC@NLO in bins of H_T and jet p_T . This is the overwhelming background in the all-hadronic boosted channel, where a pair of high- p_T QCD jets can fake the boosted H and V candidates.
- **V+jets**: $W(\rightarrow \ell\nu)$ +jets and $Z/\gamma^*(\rightarrow \ell\ell)$ +jets, together with the hadronic $W \rightarrow q\bar{q}$ and $Z \rightarrow q\bar{q}$ samples, binned in boson p_T and H_T . These are the largest electroweak backgrounds in the leptonic channels.
- **Top**: top quark-antiquark pair ($t\bar{t}$) production in its dileptonic, semileptonic, and all-hadronic decay modes, simulated at NLO with POWHEG [6]; single top quark production in the t -, s -, and tW -channels; and $t\bar{t}$ produced in association with a vector or Higgs boson.
- **Multiboson**: WW , WZ , and ZZ diboson production, the WWW , WWZ , WZZ , and ZZZ tribosons, and electroweak diboson-plus-jets processes including same-sign WW scattering, which share the forward-jet topology of the signal.
- **Higgs**: associated WH , ZH , and $gg \rightarrow ZH$ production and $t\bar{t}H$, which are irreducible where a genuine Higgs boson is produced without the VBS topology, together with the

$gg \rightarrow ZZ$ continuum.

The dominant background is channel-dependent: QCD multijet production overwhelms the all-hadronic channel, while $t\bar{t}$ and W +jets production dominate the single-lepton channels. Representative tree-level diagrams for the principal reducible backgrounds of the leptonic channels are shown in Figure 4.5: each produces genuine leptons, b jets, or hadronic jets that can mimic the VVH final state without the forward VBS topology.

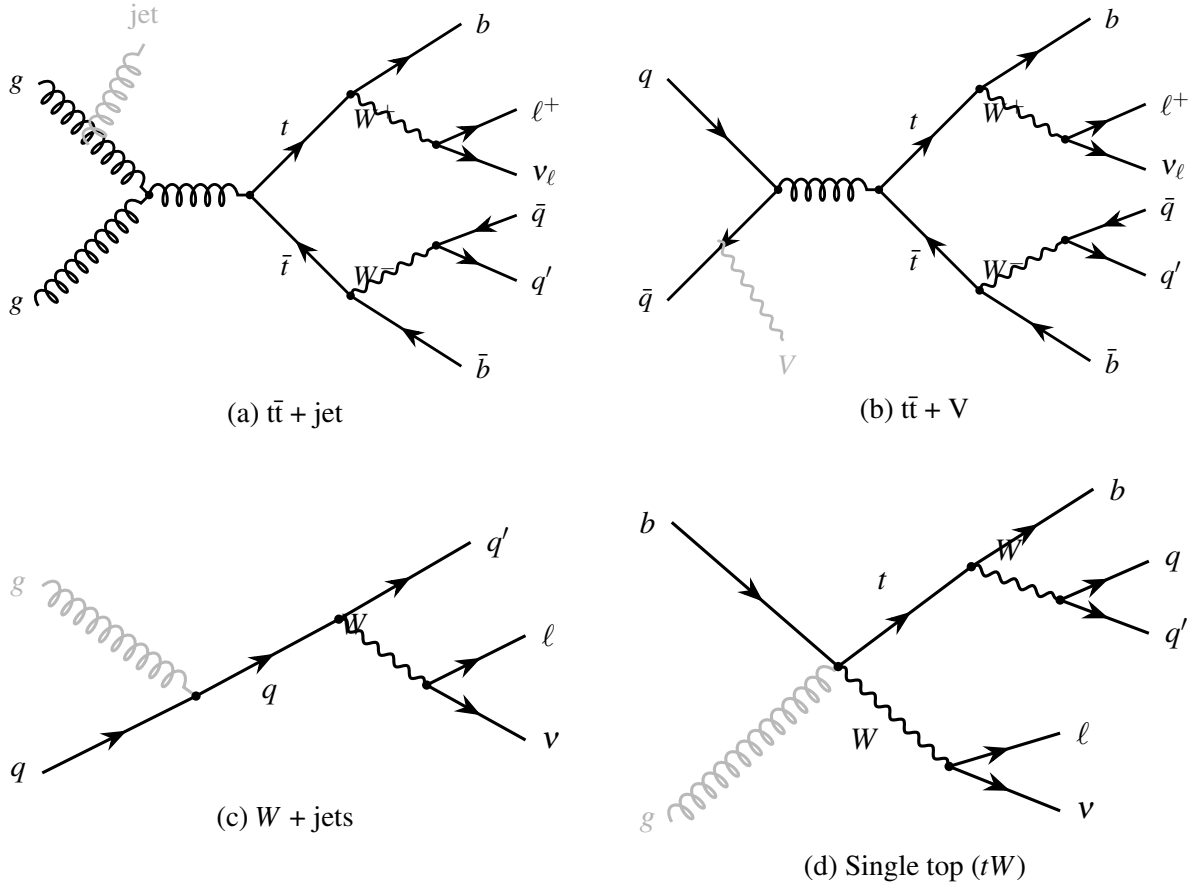


Figure 4.5. Representative tree-level Feynman diagrams for the principal backgrounds of the leptonic channels.

4.5.2 Object Reconstruction and Identification

All physics objects are built from the particle-flow (PF) candidates of Section 3.2.2. The reconstruction of each object type is described in the reconstruction chapter; this section specifies

only the working points, kinematic acceptances, and taggers that are particular to this analysis.

Electrons and Muons

Electrons ($|\eta| < 2.5$) and muons ($|\eta| < 2.4$) are reconstructed as described in Section 3.2.3 and identified at two working points. A *loose* category, with a low p_T threshold (7 GeV for electrons, 5 GeV for muons) is used to veto additional leptonic activity and so enforce orthogonality between channels. A *tight* category, built on tighter identification and isolation requirements and $p_T > 30$ GeV, defines the leptons of the single-lepton signal channels. The isolation requirement that separates prompt leptons from those inside jets (Section 3.2.3) is applied in a cone of $\Delta R \approx 0.3$ -0.4 around the lepton.

Jets

Jets are clustered from the PUPPI-weighted PF candidates with the anti- k_t algorithm (Section 3.2.3). Two radii are used. The small-radius **AK4 jets** ($R = 0.4$) are required to have $p_T > 20$ GeV and are kept out to $|\eta| < 5.0$ so that the forward VBS jets are retained; they are cleaned against selected leptons ($\Delta R > 0.4$) and, in the collection used for the hadronic recoil, against the AK8 jets ($\Delta R > 0.8$).

The large-radius **AK8 jets** ($R = 0.8$), which capture the collimated decay products of a single boosted heavy object, are required to have $p_T > 250$ GeV, $|\eta| < 2.5$, and a soft-drop mass above 40 GeV. The AK8 jet mass is computed both from the soft-drop grooming procedure [63], which recursively removes soft, wide-angle radiation so that the groomed mass peaks at the parent-boson mass for genuine $V \rightarrow q\bar{q}$ and $H \rightarrow b\bar{b}$ decays, and from a dedicated mass regression provided by the boosted tagger.

The two forward VBS jets are selected from the AK4 jets, and the central objects of the analysis are reconstructed as single AK8 jets: a Higgs candidate decaying to a pair of b quarks ($H \rightarrow b\bar{b}$) and, where the topology is boosted, hadronically decaying vector bosons ($V \rightarrow q\bar{q}$).

Small-radius jets originating from b quarks are identified with the UPART tagger (Sec-

tion 3.2.3), a transformer-based network that classifies the flavor of an AK4 jet from its constituents; it supersedes the DEEPJET [24] tagger of the previous analysis. In the single-lepton channels a b -tagged AK4 jet is used to veto the $t\bar{t}$ background.

Jet Flavor Tagging

The boosted Higgs and vector-boson candidates are identified with mass-decorrelated machine-learning taggers that assign each AK8 jet a score for the $X \rightarrow b\bar{b}$ and $X \rightarrow q\bar{q}$ hypotheses against the dominant QCD multijet background. Earlier taggers relied on a handful of physics-motivated substructure observables, but these have been superseded by deep networks that operate directly on the full set of PF constituents of the jet. The previous analysis used the PARTICLENET tagger [74], which represents the jet as an unordered "particle cloud" and applies a graph neural network whose convolutions act on each constituent and its nearest neighbors. This dissertation instead adopts the GLOPART tagger, built on the Particle Transformer architecture [75], in which self-attention layers allow every constituent to interact with every other one, augmented by pairwise kinematic interaction features; the attention mechanism captures longer-range correlations across the jet and yields improved discrimination over the graph-based approach. From the tagger outputs, two discriminants are formed: a *Higgs-tagger* score that separates $H \rightarrow b\bar{b}$ jets from QCD jets, and a *V-tagger* score that separates hadronic $V \rightarrow q\bar{q}$ jets from QCD jets. In both cases the taggers are decorrelated from the jet mass so that a selection on the score does not sculpt an artificial peak in the background mass spectrum.

VBS Jets Tagging

The two forward tagging jets are selected from the AK4 jets that survive overlap removal against the reconstructed Higgs and vector-boson candidates. Rather than a simple kinematic rule such as taking the pair of largest invariant mass, the choice is made by a dedicated gradient-boosted decision tree, trained with XGBOOST [28], that scores how VBS-like each candidate jet pair is; the highest-scoring pair in the event is taken as the tagging pair. The training and

inference flow is summarized in Figure 4.6.

The tree is trained on simulated signal events only. For every event with at least two jets, all jet pairings are formed, and each is labeled as the true VBS pair, matched to the two generator-level scattered partons, or as combinatorial background, the latter arising from initial- and final-state radiation or the underlying event. Because a single true pair competes with many combinatorial ones, the classes are strongly imbalanced; the positive class is reweighted by the background-to-signal ratio $\Sigma(y=0)/\Sigma(y=1)$ so that the two contribute equally to the loss. Each pair is described by twelve kinematic features: the transverse momentum, pseudorapidity, azimuth, and mass of each of the two jets, together with the pair’s invariant mass m_{jj} , transverse momentum $p_{T,jj}$, pseudorapidity separation $|\Delta\eta_{jj}|$, and azimuthal separation $\Delta\phi_{jj}$. The tree minimizes a binary logistic loss, monitored on a held-out validation set through the area under the ROC curve with early stopping; its hyperparameters are collected in Table 4.2.

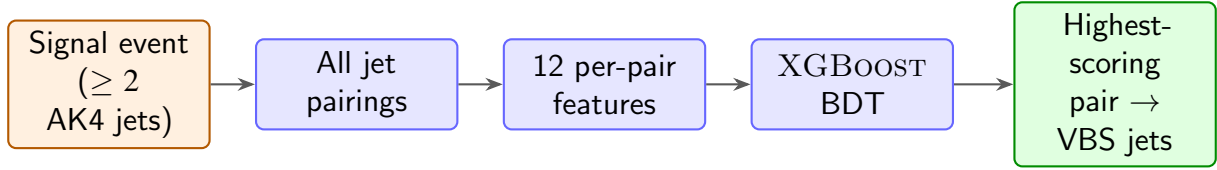


Figure 4.6. Training and inference flow of the VBS jet-pair selection BDT.

Table 4.2. Hyperparameters of the VBS jet-pair selection BDT.

Objective / tree method	binary logistic / hist
Evaluation metric	AUC
Trees (max) / early stopping	500 / 100 rounds
Learning rate	0.03
Maximum depth	4
Minimum child weight	6
Minimum split loss γ	0.5
Row subsampling	0.85
Feature subsampling (tree / level)	0.9 / 0.9
L_2 / L_1 regularization	5.0 / 0.5
Class balancing	<code>scale_pos_weight</code> = $\Sigma(y=0)/\Sigma(y=1)$
Train/validation split	80/20

A single model, trained on the combined Run 2 and Run 3 signal samples, is used

throughout, since models trained per era and on the combination differ in matching performance by only a few tenths of a percent. At inference the trained tree scores every jet pair in the event and the highest-scoring pair is selected. The model identifies the correct pair with an efficiency of roughly 90% (both jets correct) and recovers at least one of the two jets in over 99% of events (Table 4.3).

Table 4.3. VBS jet-pair matching accuracy of the era-dedicated and combined BDTs, on Run 2 and Run 3 signal.

Matching category	Run 2 (%)		Run 3 (%)	
	Dedicated	Combined	Dedicated	Combined
Both jets correct	89.82	89.40	90.01	89.94
Exactly one jet correct	9.78	10.14	9.61	9.66
No jets correct	0.40	0.45	0.38	0.40
At least one jet correct	99.60	99.55	99.62	99.60

This classifier only selects the VBS jet pair; one of the two axes of the data-driven background estimate is instead provided by a *separate* boosted decision tree, trained on the selected tagging-jet system to separate signal from background, described in Section 4.5.4.

Missing Transverse Momentum

The missing transverse momentum $\vec{p}_T^{\text{miss}}$ (Section 3.2.3), which in the leptonic channels captures the momentum carried away by neutrinos, is computed from the PUPPI-weighted PF candidates and recomputed after the jet energy corrections are applied.

4.5.3 Analysis Strategy and Event Selection

The experimental signature combines three elements: two forward VBS jets with a large pseudorapidity separation $|\Delta\eta_{jj}|$ and a large invariant mass m_{jj} ; a boosted $H \rightarrow b\bar{b}$ reconstructed as a single AK8 jet; and the decay products of the two vector bosons, which may be hadronic (further AK8 or AK4 jets) or leptonic (charged leptons and missing transverse momentum). Events are partitioned into mutually exclusive channels according to the number of reconstructed leptons and the number of AK8 jets, so that the range of V decay modes is covered while

orthogonality is preserved. This iteration expands the analysis into a large set of such channels, spanning all-hadronic, single-lepton, and multilepton final states with both fully boosted and semi-resolved hadronic topologies, each analyzed separately in the Run 2 and Run 3 datasets and combined in the final fit. This dissertation describes in detail the all-hadronic fully-boosted channel and the single-lepton channels, which together provide much of the sensitivity to κ_{2V} . The full set of channels is summarized in Table 4.4.

Table 4.4. Summary of the analysis channels ("FJ" a fat jet, "X" extra AK4 or p_T^{miss} activity); the three detailed in this chapter are in bold.

Channel	Leptons	Fat jets	Topology
<i>Fully hadronic</i>			
0 lep + 3 FJ	0	3	Fully boosted
0 lep + 2 FJ + p_T^{miss}	0	2	Boosted + invisible
0 lep + 2 FJ + 2 AK4	0	2	Semi-boosted
0 lep + 1 FJ + p_T^{miss} + 2 AK4	0	1	Mixed
0 lep + 1 FJ + 4 AK4	0	1	Semi-resolved
0 lep + 6 AK4	0	0	Fully resolved
<i>Single lepton</i>			
1 lep + 2 FJ	1	2	Boosted
1 lep + 1 FJ + X	1	1	Semi-boosted
<i>Dilepton</i>			
2 lep (OS OF) + 1 FJ	2	1	Boosted
2 lep (OS SF) + 2 FJ	2	2	Boosted
2 lep (OS SF) + 1 FJ + X	2	1	Semi-boosted
2 lep (SS)	2	—	Inclusive
<i>Multilepton</i>			
3 lep	3	—	Inclusive
4 lep	4	—	Inclusive

4.5.4 Background Estimation and Signal Extraction

Because the principal backgrounds are large and difficult to simulate reliably, they are estimated directly from data using a matrix ("ABCD") method. This analysis retains the ABCD strategy of the previous search; the method and its decorrelation are described first, followed by the signal-region construction, which is extended here from a single region per channel to many and is where the two analyzes differ most.

The ABCD Method

The ABCD method estimates a background directly from data, without relying on its simulation [25]. It uses two discriminating variables, x and y , that are statistically independent for the background process. A threshold on each variable partitions the x - y plane into four mutually exclusive regions, labeled A, B, C, and D (Figure 4.7). Region A is the signal region, placed where the signal concentrates, for instance where both variables pass a tight requirement; the remaining three are background-dominated control (sideband) regions, designed to contain a negligible amount of signal.

When the two variables are independent for the background, its density factorizes, $p(x, y) = p(x)p(y)$, so the four background yields obey $N_A N_D = N_B N_C$. The unknown background in the signal region can then be predicted entirely from the observed counts in the three control regions,

$$N_A^{\text{bkg}} = \frac{N_B N_C}{N_D}, \quad (4.5)$$

with B and C the regions adjacent to A, each sharing one passing requirement, and D the diagonally opposite region in which both requirements fail. Two conditions make the prediction unbiased. The two variables must be genuinely uncorrelated for the background, since any residual correlation breaks the factorization and biases Eq. (4.5), an effect quantified by the *non-closure* of the method; and the control regions must be signal-depleted, or else the signal contamination must be subtracted using the expected signal yields. The first condition is the more delicate and motivates the decorrelation procedure described below.

Decorrelation via Distance Correlation

In this analysis the two ABCD axes are provided by two dedicated multivariate discriminants built from complementary sets of observables. The first is a deep neural network (DNN) that separates signal from background using the kinematics of the central VVH system: the lepton and missing-transverse-momentum kinematics, the momenta of the Higgs and vector-

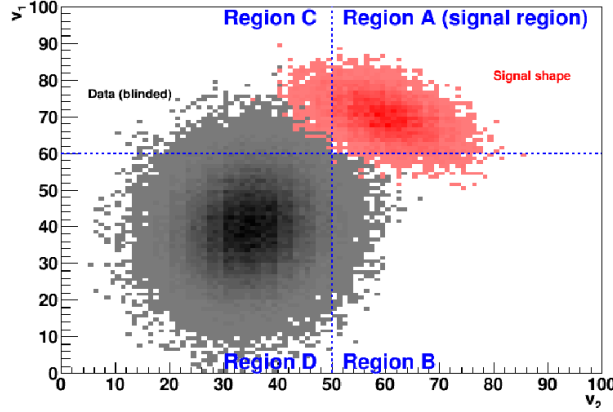


Figure 4.7. Schematic of the ABCD method: a signal region A and three background-dominated control regions B, C, D, related by Eq. (4.5).

boson candidates and their groomed and regressed masses and N -subjettiness ratio τ_{21} , and the top-quark discriminator $m_{\ell b}$. The second is the score of a dedicated boosted decision tree (BDT) trained to separate signal from background using the forward VBS tagging-jet observables alone, distinct from the BDT that selects the tagging pair (Section 4.5.2).

The two discriminants therefore probe orthogonal parts of the event, the central resonance system and the forward VBS topology, which makes them natural candidates for the two independent ABCD variables: the signal region A is defined by high values of both, and the control regions by inverting one or both requirements (Figure 4.7).

An ordinary classifier trained only to separate signal from background will not in general satisfy the independence condition the method requires: the DNN output typically becomes correlated with the BDT score, biasing the prediction of Eq. (4.5). The analysis therefore enforces independence directly during training using the distance-correlation (DisCo) technique of Ref. [60]. Distance correlation is a measure of statistical dependence between two variables that, unlike the linear Pearson coefficient, vanishes *if and only if* the variables are independent, and therefore penalizes nonlinear as well as linear correlations. For a training batch of n events with DNN outputs a_i and second-variable values b_i , it is built from the doubly-centered matrix

of pairwise distances,

$$A_{jk} = d_{jk} - \frac{1}{n} \sum_l d_{lk} - \frac{1}{n} \sum_m d_{jm} + \frac{1}{n^2} \sum_{l,m} d_{lm}, \quad d_{jk} = |a_j - a_k|, \quad (4.6)$$

with an analogous matrix B_{jk} constructed from the pairwise distances $|b_j - b_k|$. The squared distance covariance and the distance correlation then follow as

$$\text{dCov}^2(a, b) = \frac{1}{n^2} \sum_{j,k} A_{jk} B_{jk}, \quad \text{dCor}(a, b) = \frac{\text{dCov}(a, b)}{\sqrt{\text{dCov}(a, a) \text{dCov}(b, b)}} \in [0, 1], \quad (4.7)$$

which equals zero only when a and b are independent and one when they are perfectly dependent.

The DNN is trained to minimize a composite loss that adds this quantity, evaluated over background events only, to the usual classification objective,

$$\mathcal{L} = \mathcal{L}_{\text{BCE}} + \lambda \text{dCor}_{\text{bkg}}^2(a, b), \quad (4.8)$$

where $\mathcal{L}_{\text{BCE}}$ is the binary cross-entropy that drives the signal-versus-background separation, a and b are the DNN and VBS BDT scores, and the second term is the decorrelation penalty.

Restricting the penalty to background events preserves the network's discrimination power, since it never decorrelates the signal, while forcing the DNN output to be statistically independent of the BDT score for background, exactly the condition the ABCD method requires. The hyperparameter λ sets the trade-off: a larger value enforces stronger decorrelation, and hence better closure of the background prediction, at the price of some classification performance; in practice it is set between about 5 and 10 depending on the channel, and its working point is chosen to balance the two.

The class weights are normalized so that signal and background contribute equally to the classification term. A more direct alternative was also investigated, in which the penalty is the differentiable ABCD non-closure itself: soft region yields $N_A, \dots, N_D$ are formed for

background events from smooth threshold functions at randomly sampled cuts and combined into $|N_A N_D - N_B N_C| / (N_A N_D + N_B N_C)$, penalizing violations of the factorization directly. In practice the limited size of the background sample made this estimator too statistically noisy to train against reliably, and the distance-correlation penalty of Eq. (4.8) was retained.

The closure of the method, the agreement between the prediction of Eq. (4.5) and the observed yield, is verified in background-dominated control regions before any signal region is examined (Figure 4.8). The closure test remains blind to region A: it scans internal splits within the control band, using only the sidebands B, C, and D, and forms many independent “mini-ABCD” sub-divisions from which two closure metrics, $2(A - \hat{A}) / (A + \hat{A})$ and $1 - A / \hat{A}$, are computed between the observed and predicted yields of each sub-division. The spread of these metrics across the sub-divisions measures the residual non-closure of the method and provides the empirical basis for the associated systematic uncertainty (Section 4.5.10).

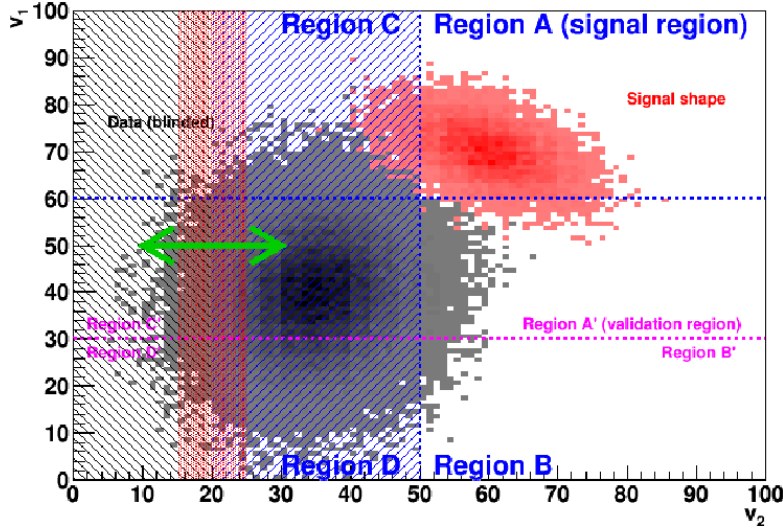


Figure 4.8. Closure test of the ABCD prediction (Eq. (4.5)) against the observed yield in a control region.

Classifier Architectures and Training

Both discriminants share a common design across the channels, with only their inputs and a few per-channel settings changed; their principal hyperparameters are collected in Table 4.5.

The central-system classifier is a fully-connected deep neural network terminating in a single sigmoid output. Its input dimensionality is fixed by the channel, ranging from 17 variables in the all-hadronic fully-boosted channel to 49 in the semi-resolved single-lepton channel, where the several candidate resolved dijet pairs and their mass compatibilities with the H , W , and Z hypotheses enlarge the input set. Continuous inputs are scaled to the unit interval, with a logarithm applied first to those spanning several orders of magnitude, and undefined entries are flagged and mapped to zero. The network minimizes the composite loss of Eq. (4.8) with the AdamW optimizer at a learning rate of 10^{-4} , reduced on plateau, for up to a thousand epochs with early stopping on the validation loss and an 80/20 training-validation split; the signal and background weights are separately normalized to unit sum so that the two classes weigh equally in the classification term. Only the decorrelation strength λ_{DisCo} and the batch size are tuned per channel: λ_{DisCo} lies between 5 and 10, taking the larger value where the residual correlation between the central and forward systems is stronger, and the batch size grows with the surviving event count of the channel.

The second discriminant is a gradient-boosted decision tree that separates signal from background using only the forward tagging-jet system, taking the invariant mass and pseudo-rapidity separation of the tagging pair and the transverse momentum, pseudorapidity, azimuth, and mass of each of its two jets. It is a different classifier from the tree that selects the VBS pair (Section 4.5.2): the pair is chosen first, and this classifier then scores how VBS-like it is. Because a decision-tree ensemble is invariant to monotonic rescalings of its inputs and handles undefined values natively, these inputs are passed unscaled, and the same tree hyperparameters are used in every channel and both eras.

4.5.5 The All-Hadronic Fully-Boosted Channel

This fully boosted channel targets events in which the Higgs boson and both vector bosons decay hadronically, each collimated into a single large-radius (AK8) jet (Figure 4.9). It is the only fully-hadronic channel currently contributing and, together with the single-lepton

Table 4.5. Architecture and training hyperparameters of the two discriminants (per-channel ranges where they vary).

Feed Forward Network (Decorrelated w.r.t. BDT)	
Hidden layers	128, 64, 32, 16
Regularization	batch normalization, dropout 0.2
Activation / output	ReLU / single sigmoid node
Input variables	17–49 (channel dependent)
Optimizer	AdamW, learning rate 10^{-4} , weight decay 0.01
Learning-rate schedule	reduce on plateau (factor 0.5, patience 40)
Epochs	up to 1000, early stopping (patience 40)
Training/validation split	80/20, stratified
Batch size	2048–16384 (channel dependent)
λ_{DisCo}	5–10 (channel dependent)
VBS gradient-boosted decision tree	
Inputs	m_{jj} , $ \Delta\eta_{jj} $, and p_T, η, ϕ, m of each tagging jet
Trees / maximum depth	up to 1000 / 6
Learning rate	0.03
Row / feature subsampling	0.8 / 0.7
L_2 regularization	10
Early stopping	40 rounds
Class balancing	signal reweighted to the background yield

fully-boosted channel, carries much of the sensitivity to κ_{2V} . Events are collected online with triggers on the scalar sum of jet transverse momenta (H_T); offline, they are required to contain no loose leptons, which enforces orthogonality with the leptonic channels, at least three AK8 jets, and at least two AK4 jets to serve as forward tagging-jet candidates.

Object Preselections

The two VBS tagging jets are identified first, by the dedicated BDT (Section 4.5.2); then, among the AK8 jets not overlapping the VBS pair, the one with the highest Higgs-tagger score is taken as the Higgs candidate and the next two, ranked by V -tagger score, as the two vector-boson candidates. Before entering the background-estimation stage, a loose preselection requires the Higgs candidate and both vector-boson candidates to have tagger scores above a low threshold, removing the bulk of the QCD multijet events while retaining essentially all of the signal. The selection is optimized at $\kappa_{2V} = 1.5$ (Section 4.5.1), and the dominant QCD multijet contribution

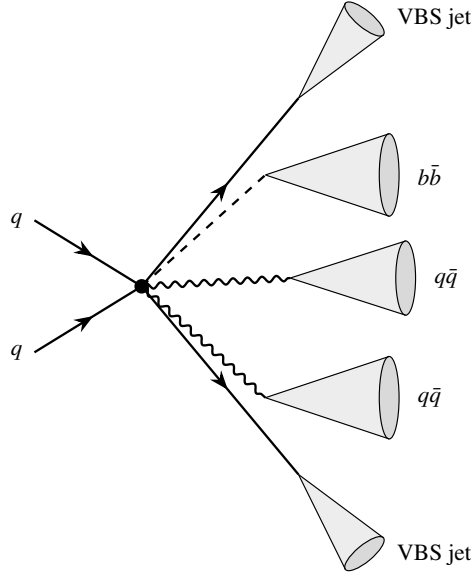


Figure 4.9. Final-state topology of the all-hadronic fully-boosted channel.

in the signal region is estimated from data (Section 4.5.4).

Requiring all three bosons to be boosted selects almost exclusively the high-momentum tail of the signal, which gives this channel a steep coupling dependence: the fraction of generated signal surviving the full selection rises by roughly an order of magnitude between the SM point and $\kappa_{2V} = 1.5$, because anomalous couplings populate the boosted regime far more efficiently (Section 4.2). This channel is therefore primarily a probe of anomalous κ_{2V} rather than of SM-rate VVH production, consistent with its optimization at $\kappa_{2V} = 1.5$.

Cutflow

The signal, background, and observed yields at each stage of the selection are given in Tables 4.6 and 4.7.

QCD Tagger-Score Resampling

After the preselection the channel is overwhelmingly dominated by QCD multijet production, which supplies roughly 93% of the background. The simulation does not reproduce the tails of the boosted-tagger score distributions for QCD jets accurately enough to be relied upon

Table 4.6. Cutflow for the all-hadronic fully-boosted channel, Run 2.

Selection	Signal			Background	Data
	SM	$\kappa_{2V} = 1.5$	$\kappa_{\lambda} = 10$		
All events	194.6	368.0	405.1	15 834 205.5	14 623 923
MET filters	194.3	366.5	404.3	15 801 876.7	13 740 743
Trigger selection	13.3	101.8	30.8	7 118 670.9	8 175 659
0 lepton, ≥ 3 fat jets	0.5	12.2	1.9	1 480 290.9	1 737 193
≥ 2 jets	0.5	12.2	1.9	1 480 276.7	1 737 174
Boosted H candidate	0.5	12.2	1.9	1 480 276.7	1 737 174
Boosted V candidates	0.5	12.2	1.9	1 478 004.6	1 734 259
VBS pair candidate found	0.3	5.8	1.1	879 439.4	1 074 806
After SFs and corrections	0.3	5.8	1.1	881 392.3	1 041 075

Table 4.7. Cutflow for the all-hadronic fully-boosted channel, Run 3.

Selection	Signal					Background	Data
	SM	$\kappa_{2V} = 1.5$	$\kappa_{\lambda} = 10$	$\kappa_{2V} = 0.25$	$\kappa_{2V} = 0.75$		
All events	475.7	926.9	983.6	1393.8	570.6	28 365 412.2	28 059 909
MET filters	474.7	923.5	981.6	1386.8	568.9	28 311 082.4	27 816 775
Trigger selection	18.8	208.1	44.3	451.1	65.2	9 946 980.2	11 623 677
0 lepton, ≥ 3 fat jets	0.9	25.6	3.1	57.9	6.9	2 350 837.4	2 744 977
≥ 2 jets	0.9	25.6	3.1	57.9	6.9	2 350 837.4	2 744 977
Boosted H candidate	0.9	25.6	3.1	57.9	6.9	2 350 837.4	2 744 977
Boosted V candidates	0.9	25.6	3.1	57.9	6.9	2 347 172.9	2 740 191
VBS pair candidate found	0.6	14.7	2.0	33.1	4.1	1 383 343.8	1 638 403
After SFs and corrections	0.6	14.7	2.0	33.6	4.1	1 468 175.1	1 592 991

in the high-score region where the signal concentrates and the signal regions are defined. To remedy this without discarding the simulation, the per-jet tagger scores of the simulated QCD sample, and only that sample, with data and all other processes left untouched, are replaced by values drawn from template distributions measured in data, binned in jet transverse momentum and pseudorapidity. The V -tagger score is drawn conditionally on the Higgs-tagger score already assigned to the same jet, so that the correlation between the two taggers observed in data is preserved rather than destroyed by independent sampling, and the replacement is applied before the candidate assignment, so that the choice of which jet becomes the Higgs candidate is itself made with the corrected scores. As Figure 4.10 shows, the procedure markedly improves the agreement between data and simulation across all three tagger scores, and in particular in the high-score tails from which the signal regions are carved.

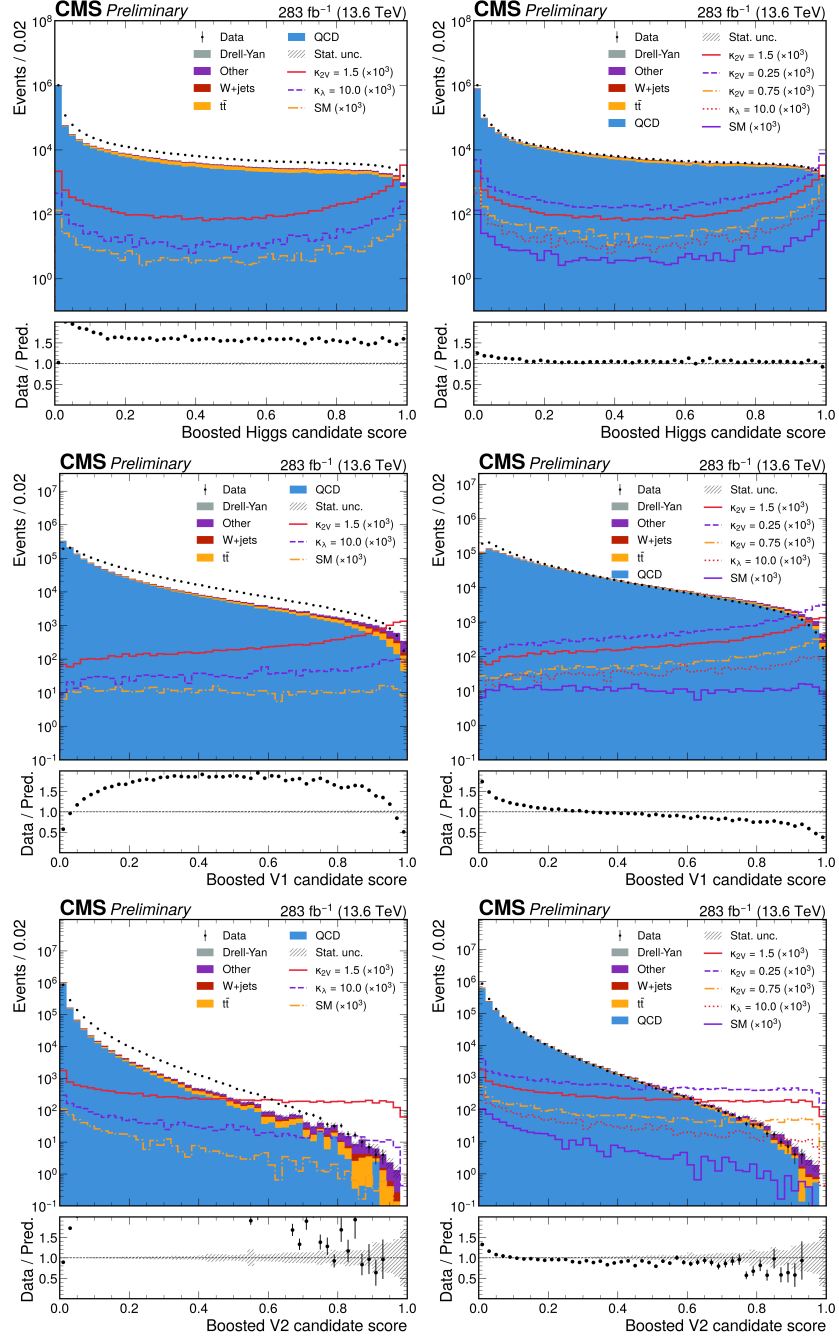


Figure 4.10. QCD tagger-score resampling in the all-hadronic channel (Run 3), before (left) and after (right): Higgs candidate (top) and the two V candidates (middle, bottom).

Kinematic Distributions

The observables that feed the two multivariate discriminants of Section 4.5.4 are shown in Figures 4.11–4.13. In each, the simulated backgrounds are stacked, the signal is overlaid and scaled by the factor given in the legend, and the lower panel shows the data-to-prediction ratio; the QCD simulation has had its tagger scores resampled as described above. The two that carry the discrimination are shown here. The groomed (soft-drop) mass and N -subjettiness ratio τ_{21} of the three boosted candidates (Figure 4.11) each show a distinct peak at the boson mass and a low- τ_{21} two-prong structure for genuine boson decays, against the broad, single-prong QCD background; the forward tagging-jet pair (Figure 4.12) populates the high-mass, large-rapidity-gap region characteristic of vector boson scattering. The candidate transverse momenta, pseudorapidities and azimuths, the missing transverse momentum, and the corresponding Run 3 distributions are collected in Appendix A.

Multivariate Analysis

From these inputs the two decorrelated discriminants of Section 4.5.4 are formed; their output scores, which define the two axes of the data-driven background estimate, are shown in Figure 4.13. Because this channel carries three tagged bosons, the signal-region scan is five-dimensional, over the three tagger scores and the two discriminant thresholds, and yields five orthogonal signal regions per era. The most significant reaches an expected Asimov significance of $Z = 5.04$ in Run 3 with about 1.7 expected signal events; the tight tagger requirements that give the channel its sensitivity also leave the control regions sparsely populated, especially in Run 2. In Run 3 the channel reaches an observed (median expected) 95% CL upper limit of $\mu < 1.31$ (1.43) on the signal strength (Table 4.18), close to the rate of the $\kappa_{2V} = 1.5$ benchmark and, despite the overwhelming QCD background, the second most sensitive of the three channels. The signal-region definitions and yields from the scan are listed in Tables 4.8 and 4.9.

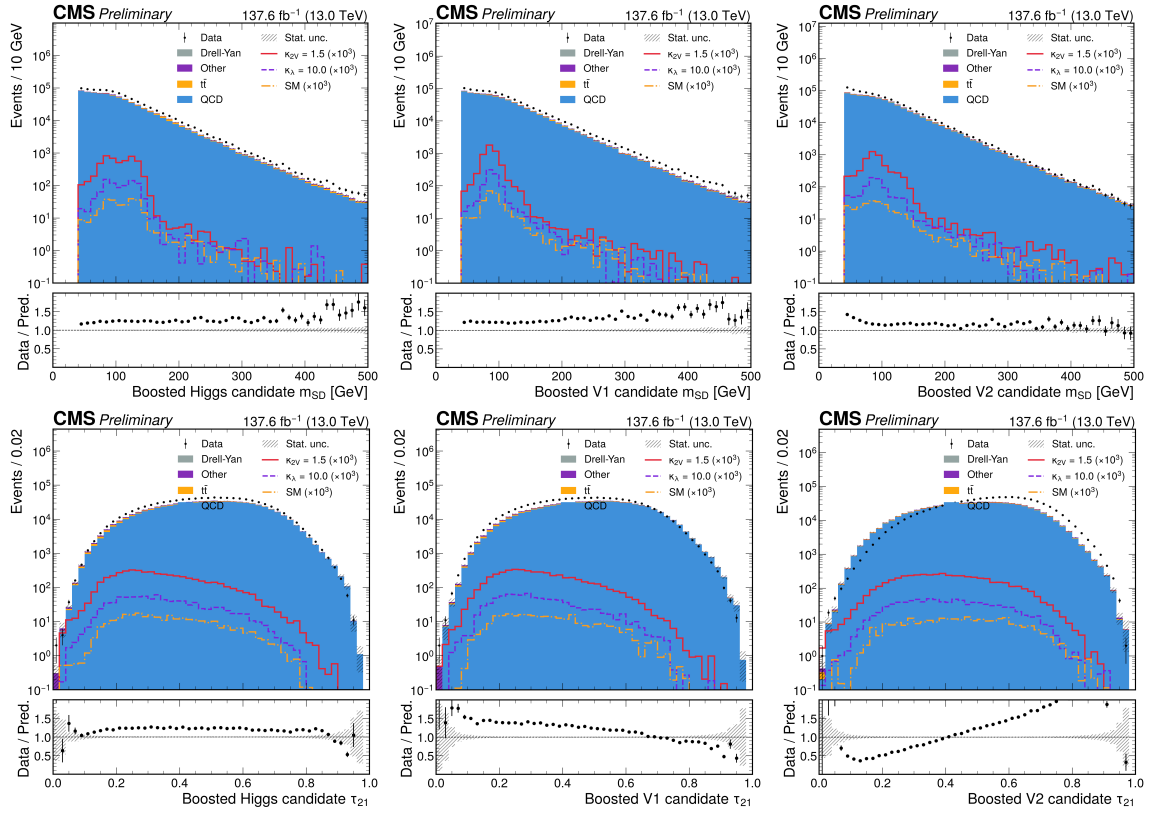


Figure 4.11. Soft-drop mass (top) and τ_{21} (bottom) of the Higgs (left) and the two V candidates, all-hadronic channel, Run 2.

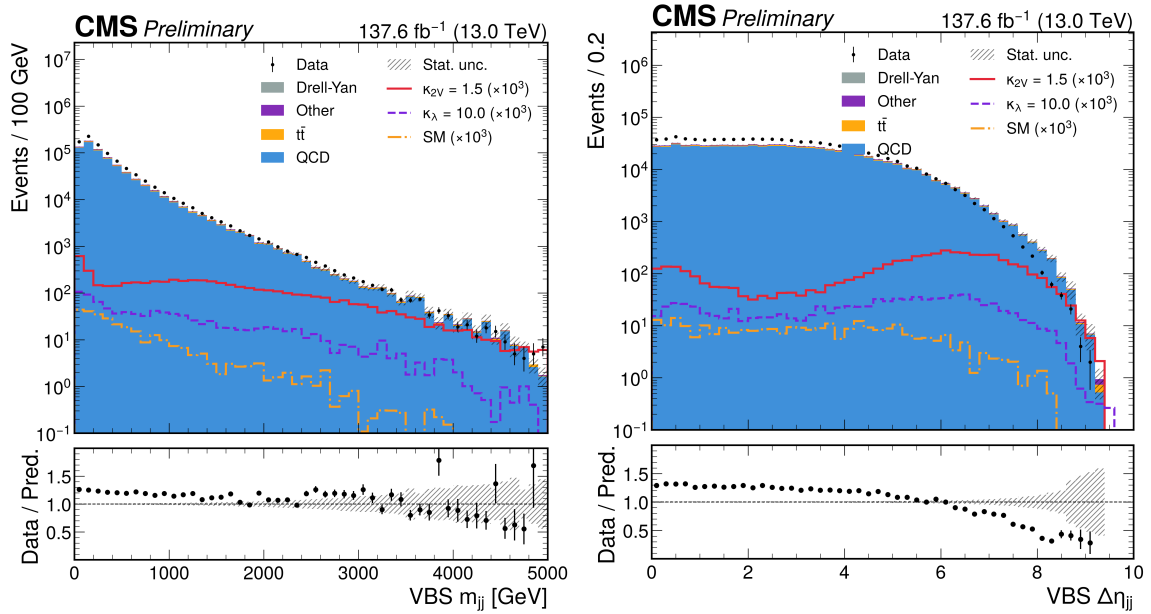


Figure 4.12. Tagging-jet m_{jj} (left) and $|\Delta\eta_{jj}|$ (right), all-hadronic channel, Run 2.

Table 4.8. Signal-region definitions and yields for the all-hadronic fully-boosted channel, Run 2.

Region	Definition	Z	Bin	Run 2		
				Signal	Background (MC)	Data
1	$H > 0.68, V_1 > 0.84, V_2 > 0.24, \text{BDT} > 0.94, \text{DNN} > 0.98$	2.12	A	0.3654	0.0003	0 ($\tilde{N}_A = 0.11$)
			B	0.5614	7.7252	13
			C	0.0601	0.8696	1
			D	0.0888	78.5213	121
2	$H > 0.96, V_1 > 0.10, V_2 > 0.14, \text{BDT} > 0.88, \text{DNN} > 0.98$	1.26	A	0.1704	0.0006	0 ($\tilde{N}_A = 0.10$)
			B	0.1282	2.2281	4
			C	0.0593	5.7600	6
			D	0.0572	185.8865	249
3	$H > 0.10, V_1 > 0.66, V_2 > 0.58, \text{BDT} > 0.96, \text{DNN} > 0.98$	0.66	A	0.0671	0.0011	0 ($\tilde{N}_A = 0.00$)
			B	0.1452	11.5017	18
			C	0.0219	0.9334	0
			D	0.0627	212.8037	242
4	$H > 0.90, V_1 > 0.44, V_2 > 0.12, \text{BDT} > 0.92, \text{DNN} > 0.86$	0.55	A	0.0637	0.0025	0 ($\tilde{N}_A = 0.81$)
			B	0.0914	16.2228	42
			C	0.0085	0.2661	5
			D	0.0172	225.1981	260
5	$H > 0.98, V_1 > 0.16, V_2 > 0.10, \text{BDT} > 0.58, \text{DNN} > 0.98$	0.37	A	0.0178	0.0002	1 ($\tilde{N}_A = 0.00$)
			B	0.0026	0.8298	0
			C	0.0093	1.0389	2
			D	0.0058	18.9341	38

Table 4.9. Signal-region definitions and yields for the all-hadronic fully-boosted channel, Run 3.

Region	Definition	Z	Bin	Run 3		
				Signal	Background (MC)	Data
1	$H > 0.82, V_1 > 0.64, V_2 > 0.28, \text{BDT} > 0.92, \text{DNN} > 0.96$	5.04	A	1.6906	0.0003	0 ($\tilde{N}_A = 0.31$)
			B	1.5794	15.6948	21
			C	0.1456	6.4377	4
			D	0.1892	223.5603	272
2	$H > 0.72, V_1 > 0.24, V_2 > 0.10, \text{BDT} > 0.98, \text{DNN} > 0.96$	1.86	A	0.2932	0.0003	1 ($\tilde{N}_A = 0.28$)
			B	1.3966	93.5034	131
			C	0.0634	4.0511	10
			D	0.4725	4224.4954	4693
3	$H > 0.10, V_1 > 0.80, V_2 > 0.70, \text{BDT} > 0.92, \text{DNN} > 0.98$	1.50	A	0.2017	0.0003	0 ($\tilde{N}_A = 0.09$)
			B	0.2117	17.2664	4
			C	0.0315	1.7344	2
			D	0.0302	128.9281	93
4	$H > 0.38, V_1 > 0.84, V_2 > 0.10, \text{BDT} > 0.96, \text{DNN} > 0.96$	0.90	A	0.0944	0.0005	0 ($\tilde{N}_A = 0.17$)
			B	0.1713	53.8857	38
			C	0.0134	7.1203	2
			D	0.0363	492.4253	440
5	$H > 0.10, V_1 > 0.84, V_2 > 0.10, \text{BDT} > 0.98, \text{DNN} > 0.98$	0.56	A	0.0418	0.0004	0 ($\tilde{N}_A = 0.03$)
			B	0.2144	50.5135	25
			C	0.0111	2.4172	1
			D	0.0746	932.2725	716

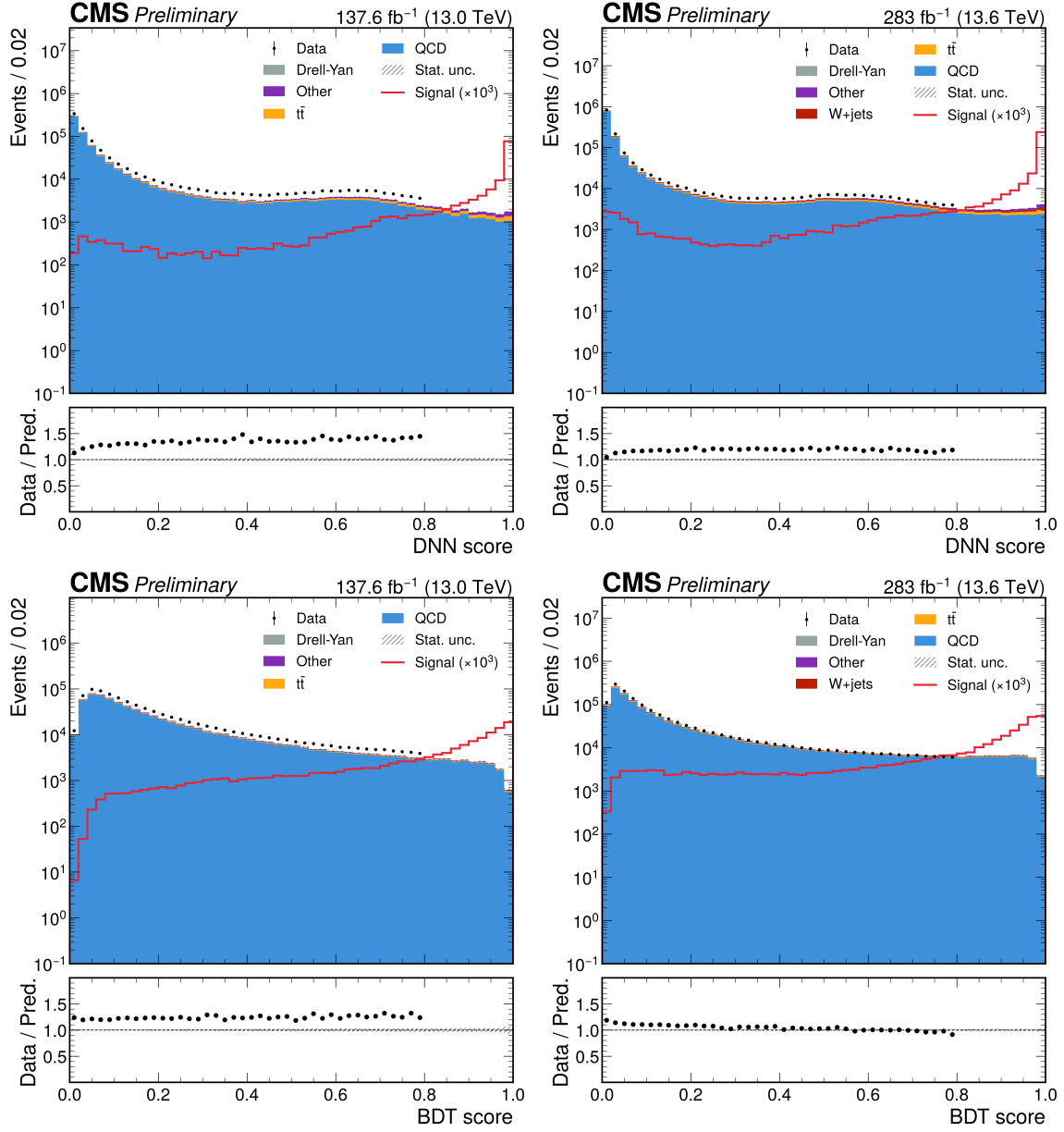


Figure 4.13. DNN score (top) and VBS BDT score (bottom), all-hadronic channel, Run 2 (left) and Run 3 (right).

4.5.6 The Single-Lepton Fully-Boosted Channel

This channel targets the fully boosted single-lepton topology, in which one W boson decays leptonically and both the Higgs boson and the remaining vector boson are each reconstructed as a single AK8 jet (Figure 4.14). It is the most sensitive of the three channels described here. Events are collected with single-lepton triggers and required to contain exactly one tight lepton with $p_T > 40$ GeV and no additional loose leptons, at least two AK8 jets, and at least two AK4 tagging-jet candidates. The AK8 jet with the highest Higgs-tagger score is taken as the Higgs candidate; among the remaining AK8 jets separated from it by $\Delta R > 0.8$, the one with the highest V -tagger score is the boosted vector boson. Both are required to pass the loose tagger preselection, and the forward tagging pair is selected by the dedicated BDT (Section 4.5.2).

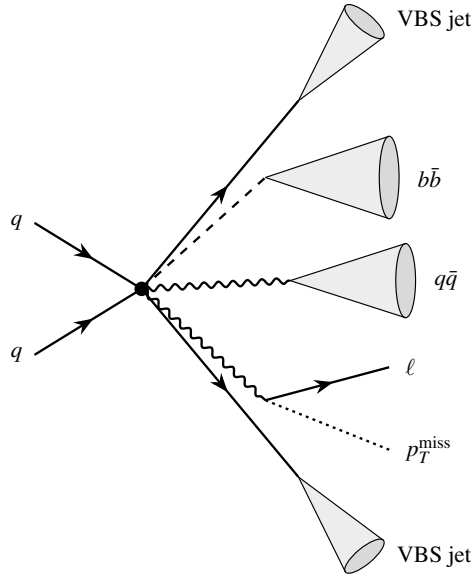


Figure 4.14. Final-state topology of the single-lepton fully-boosted channel.

Object Preselection

The requirement of two boosted bosons is the single most powerful background rejector in this channel; in Run 3 it retains roughly 15% of the $\kappa_{2V} = 1.5$ signal but only about 1.5% of the background. The dominant background is nonetheless top-quark pair ($t\bar{t}$) production, followed by W +jets: a semileptonic $t\bar{t}$ event naturally supplies an isolated lepton, missing

transverse momentum, and several jets, two of which can fake the two boosted bosons. It is further suppressed with a variable that exploits the correlation between the lepton and the b jets rather than their individual kinematics: the minimum invariant mass $m_{\ell b}$ of the lepton and a b -tagged AK4 jet. In a top-quark decay $t \rightarrow bW \rightarrow b\ell\nu$ the lepton and the b quark descend from the same parent, so their invariant mass is kinematically bounded; neglecting the lepton and b -quark masses,

$$m_{\ell b} \leq \sqrt{m_t^2 - m_W^2} \approx 153 \text{ GeV}, \quad (4.9)$$

for $m_t = 172.5 \text{ GeV}$ and $m_W = 80.4 \text{ GeV}$, so the $t\bar{t}$ distribution rises from threshold and terminates in a sharp edge at this value (Figure 4.15). For the signal no such structure exists: the lepton comes from a vector boson produced in the scattering, while any b jets come from the $H \rightarrow b\bar{b}$ decay, so the two do not share a parent and $m_{\ell b}$ is unbounded. Taking the smallest $m_{\ell b}$ over the available b -tagged jets sharpens the edge, and the variable is supplied to the classifier as an input rather than applied as a hard cut, so its discrimination is exploited without the accompanying loss of efficiency.

Cutflow

The signal, background, and observed yields at each stage of the selection are given in Tables 4.10 and 4.11.

Kinematic Distributions

The observables entering the two discriminants follow the same pattern as in the all-hadronic channel. The two boosted candidates are described by their tagger score, regressed mass, and τ_{21} : the signal peaks cleanly at the Higgs and W/Z masses with low τ_{21} , while the background is broad. The lepton and missing-momentum kinematics and $m_{\ell b}$ complete the central system, and the forward tagging pair again populates the high- m_{jj} , large- $|\Delta\eta_{jj}|$ vector-boson-scattering region. These distributions, for both eras, are collected in Appendix A.

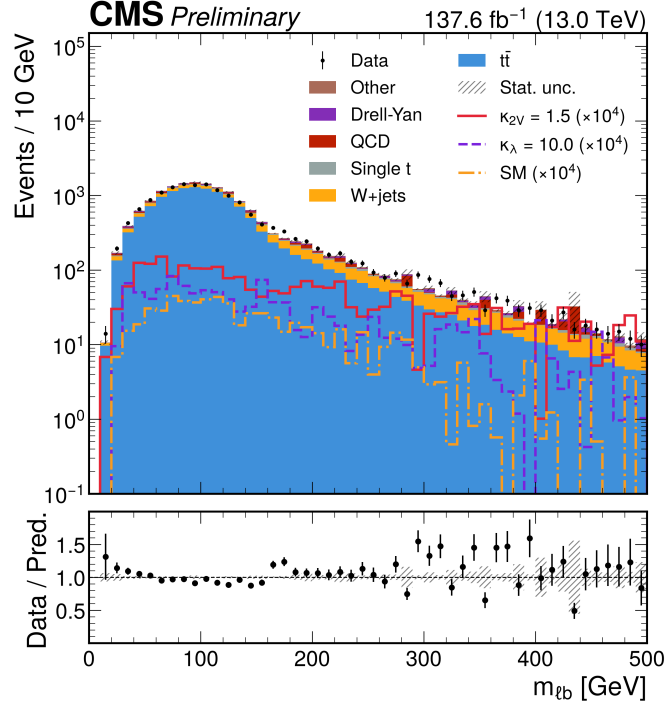


Figure 4.15. Minimum lepton- b -jet mass $m_{\ell b}$, single-lepton fully-boosted channel, Run 2; the $t\bar{t}$ edge near 153 GeV is visible (Eq. (4.9)).

Table 4.10. Cutflow for the single-lepton fully-boosted channel, Run 2.

Selection	Signal			Background	Data
	SM	$\kappa_{2V} = 1.5$	$\kappa_{\lambda} = 10$		
All events	194.6	368.0	405.1	20 054 920.4	16 229 414
MET filters	194.3	366.5	404.3	20 033 542.3	16 202 852
Trigger selection	47.1	98.3	104.9	8 542 747.0	9 565 469
1-lepton selection	24.8	56.2	56.5	5 706 329.6	6 385 959
≥ 2 fat jets	0.8	8.1	2.1	111 329.7	111 247
≥ 2 jets	0.8	8.1	2.1	111 190.2	111 138
Boosted Higgs candidate	0.8	8.1	2.1	111 190.2	111 138
Boosted vector boson candidate	0.8	8.1	2.1	111 174.7	111 117
VBS pair candidate found	0.5	4.2	1.3	68 908.1	68 792
Final VBS jet selection	0.5	4.2	1.3	68 908.1	68 792
After SFs and corrections	0.5	4.1	1.3	67 064.3	66 649

Table 4.11. Cutflow for the single-lepton fully-boosted channel, Run 3.

Selection	Signal					Background	Data
	SM	$\kappa_{2V} = 1.5$	$\kappa_\lambda = 10$	$\kappa_{2V} = 0.25$	$\kappa_{2V} = 0.75$		
All events	475.7	926.9	983.6	1393.8	570.6	66 458 607.4	55 138 503
MET filters	474.7	923.5	981.6	1386.8	568.9	66 371 065.4	54 841 105
Trigger selection	100.3	218.4	221.5	347.1	126.4	22 467 958.9	20 103 915
1-lepton selection	51.6	123.0	116.4	204.0	67.3	15 871 351.2	12 960 875
≥ 2 fat jets	1.4	18.1	4.2	39.5	5.5	251 667.8	208 773
≥ 2 jets	1.4	18.1	4.2	39.5	5.5	251 666.0	208 772
Boosted Higgs candidate	1.4	18.1	4.2	39.5	5.5	251 666.0	208 772
Boosted vector boson candidate	1.4	18.1	4.2	39.5	5.5	251 628.4	208 749
VBS pair candidate found	1.0	11.1	2.9	23.9	3.5	143 190.2	124 035
Final VBS jet selection	1.0	11.1	2.9	23.9	3.5	143 190.2	124 035
After SFs and corrections	1.0	10.9	2.8	23.3	3.4	141 153.6	120 535

Multivariate Analysis

The output scores of the two classifiers are shown in Figure 4.16. They define a four-dimensional signal-region scan, over the two tagger scores and the two discriminant thresholds, giving five orthogonal regions per era; the leading region reaches an expected significance of $Z = 5.05$ in Run 3 with about 1.6 expected signal events. In Run 3 the channel reaches an observed (median expected) 95% CL upper limit of $\mu < 1.03$ (1.24) (Table 4.18), the strongest of the three channels; Run 3 alone is about three times stronger than Run 2, reflecting the larger dataset, the larger signal cross sections at $\sqrt{s} = 13.6$ TeV, and the improved taggers. The signal-region definitions and yields from the scan are listed in Tables 4.12 and 4.13.

4.5.7 The Single-Lepton Semi-Boosted Channel

This channel targets the semi-resolved single-lepton topology, in which one W boson decays leptonically, one boson is reconstructed as a single AK8 jet, and the other is reconstructed from a pair of resolved AK4 jets (Figure 4.17). It recovers signal events in which only one of the two hadronic bosons is boosted enough to be captured by a large-radius jet, extending the search into the lower-momentum regime that the fully boosted channels miss. The lepton and trigger requirements match the fully-boosted channel; exactly one AK8 jet is required, which ensures orthogonality with it, and is classified as the Higgs or the vector-boson candidate according to

Table 4.12. Signal-region definitions and yields for the single-lepton fully-boosted channel, Run 2.

Region	Definition	Z	Bin	Run 2		
				Signal	Background (MC)	Data
1	$H > 0.10, V > 0.48, \text{BDT} > 0.92, \text{DNN} > 0.98$	2.24	A	0.3803	0.0002	0 ($\hat{N}_A = 0.01$)
			B	0.3990	9.9815	2
			C	0.2715	20.9751	17
			D	0.4244	2141.4220	2316
2	$H > 0.86, V > 0.10, \text{BDT} > 0.78, \text{DNN} > 0.96$	0.79	A	0.1224	0.0042	0 ($\hat{N}_A = 0.05$)
			B	0.0444	1.9257	1
			C	0.0841	16.5913	26
			D	0.0702	431.7903	543
3	$H > 0.10, V > 0.22, \text{BDT} > 0.94, \text{DNN} > 0.98$	0.40	A	0.0222	0.0002	0 ($\hat{N}_A = 0.02$)
			B	0.0264	4.2555	4
			C	0.0270	16.9123	19
			D	0.0914	2710.2530	3288
4	$H > 0.36, V > 0.10, \text{BDT} > 0.94, \text{DNN} > 0.96$	0.16	A	0.0092	0.0012	0 ($\hat{N}_A = 0.08$)
			B	0.0109	5.2106	8
			C	0.0093	9.7459	18
			D	0.0400	1494.7351	1844
5	$H > 0.16, V > 0.10, \text{BDT} > 0.98, \text{DNN} > 0.62$	0.09	A	0.0019	0.0001	0 ($\hat{N}_A = 0.14$)
			B	0.0168	65.9811	94
			C	0.0003	0.7253	2
			D	0.0078	1109.5473	1319

Table 4.13. Signal-region definitions and yields for the single-lepton fully-boosted channel, Run 3.

Region	Definition	Z	Bin	Run 3		
				Signal	Background (MC)	Data
1	$H > 0.88, V > 0.38, \text{BDT} > 0.82, \text{DNN} > 0.96$	5.05	A	1.6085	0.0002	0 ($\hat{N}_A = 0.21$)
			B	0.5724	5.2052	5
			C	0.4536	23.0524	15
			D	0.2881	351.4403	354
2	$H > 0.26, V > 0.22, \text{BDT} > 0.92, \text{DNN} > 0.98$	2.43	A	0.5322	0.0008	0 ($\hat{N}_A = 0.18$)
			B	0.5118	31.6511	12
			C	0.4089	115.0418	107
			D	0.6289	7147.3154	7231
3	$H > 0.96, V > 0.10, \text{BDT} > 0.54, \text{DNN} > 0.92$	1.05	A	0.0967	0.0001	0 ($\hat{N}_A = 0.00$)
			B	0.0157	2.3835	0
			C	0.0474	7.6018	8
			D	0.0280	80.8647	77
4	$H > 0.10, V > 0.22, \text{BDT} > 0.96, \text{DNN} > 0.92$	1.00	A	0.1122	0.0005	0 ($\hat{N}_A = 0.19$)
			B	0.2175	44.3685	36
			C	0.0435	22.9366	20
			D	0.1616	4439.7083	3778
5	$H > 0.90, V > 0.10, \text{BDT} > 0.28, \text{DNN} > 0.96$	0.53	A	0.0387	0.0004	1 ($\hat{N}_A = 0.00$)
			B	0.0030	1.3169	0
			C	0.0416	51.8182	46
			D	0.0138	189.3759	204

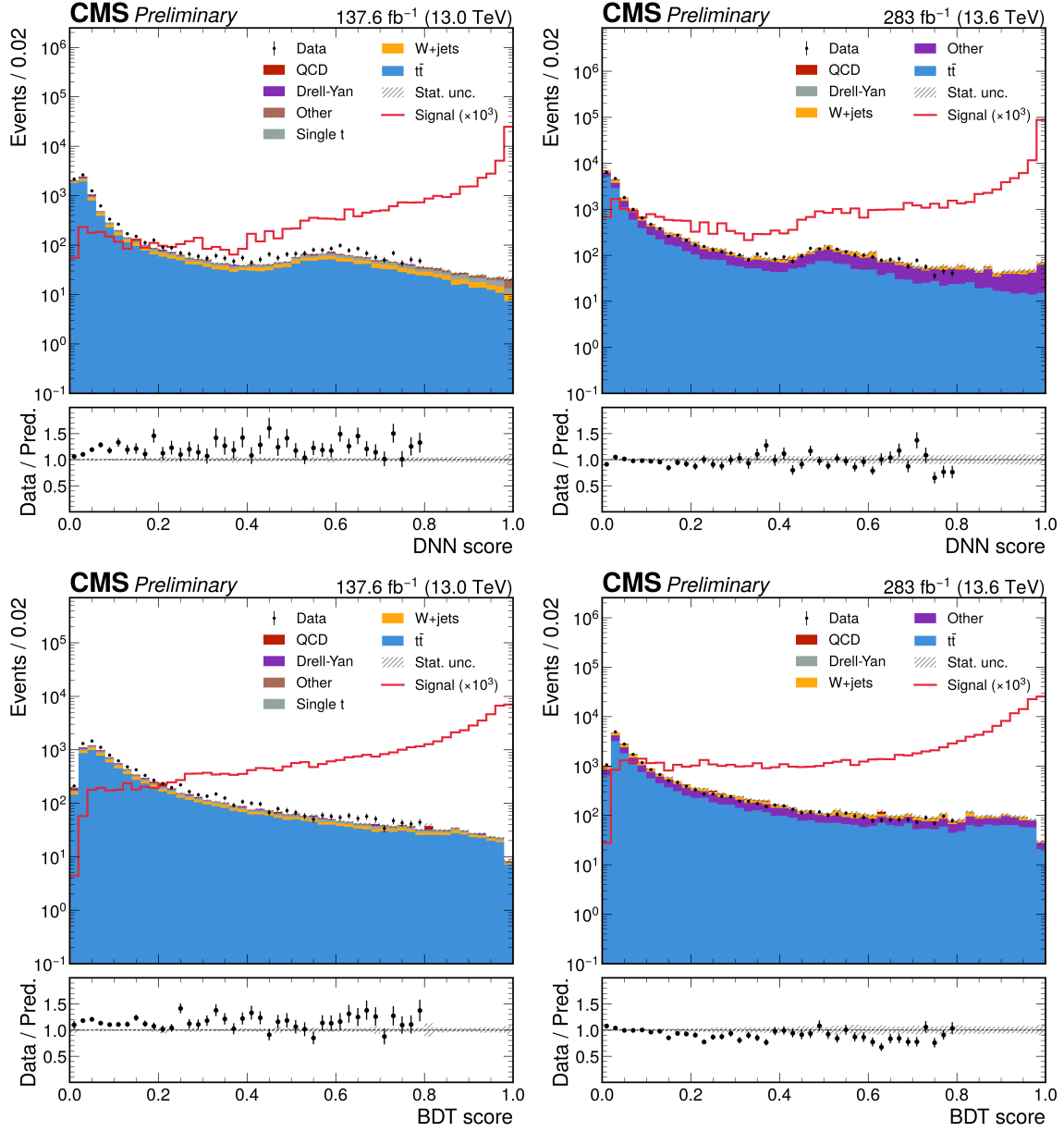


Figure 4.16. DNN (top) and VBS BDT (bottom) scores, single-lepton fully-boosted channel, Run 2 (left) and Run 3 (right).

which of its two tagger scores is larger. At least four AK4 jets are required—two to form the forward tagging pair and two to reconstruct the resolved boson.

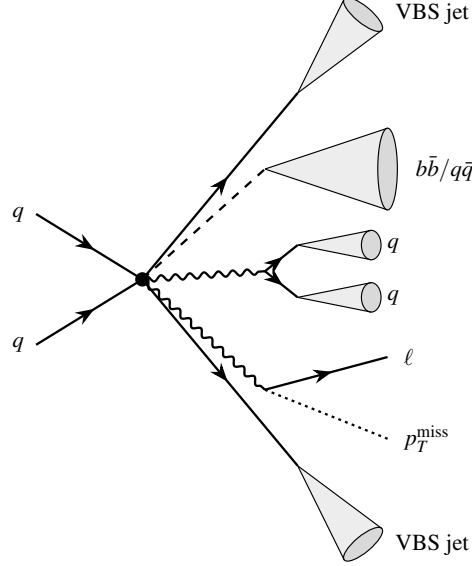


Figure 4.17. Final-state topology of the single-lepton semi-boosted channel.

Object Preselections

To reconstruct the second boson, the AK4 jets that are neither part of the tagging pair nor within $\Delta R < 0.8$ of the AK8 jet are paired. The pair with the smallest angular separation is taken as the nominal candidate, but rather than committing to a single choice, the invariant mass, transverse momentum, and ΔR of the five leading pairs, together with the absolute residual of each dijet mass from the H , W , and Z hypotheses, are all supplied to the classifier, so that the combinatorial ambiguity is resolved by the network rather than by a hard upstream choice. With the lepton, missing-momentum, boosted-candidate, and $m_{\ell b}$ variables added, this channel has by far the largest classifier input set of the three.

Top-quark pair production dominates the background by a wider margin than in the fully boosted channel: one AK8 jet plus four AK4 jets is close to the natural jet content of a semileptonic $t\bar{t}$ event, so the topological requirements alone provide little rejection, and $m_{\ell b}$ is correspondingly more important. In compensation, requiring only one boosted boson makes

the channel far less sensitive to the hardness of the signal momentum spectrum, so it recovers a much larger fraction of the SM-like signal, its acceptance at the SM point is several times higher than that of the fully boosted channels.

Cutflow

The signal, background, and observed yields at each stage of the selection are given in Tables 4.14 and 4.15.

Table 4.14. Cutflow for the single-lepton semi-boosted channel, Run 2.

Selection	Signal			Background	Data
	SM	$\kappa_{2V} = 1.5$	$\kappa_{\lambda} = 10$		
All events	194.6	368.0	405.1	20 054 920.4	16 229 414
MET filters	194.3	366.5	404.3	20 033 542.3	16 202 852
Trigger selection	47.1	98.3	104.9	8 542 747.0	9 565 469
1-lepton selection	24.8	56.2	56.5	5 706 329.6	6 385 959
Exactly 1 fat jet	4.2	19.2	11.7	2 072 990.3	2 076 387
≥ 4 jets	2.1	6.9	5.3	567 514.7	540 935
Boosted candidate selection	2.1	6.9	5.3	567 514.7	540 935
VBS candidate selection	2.1	6.9	5.3	567 514.7	540 935
Resolved candidate selection	1.8	5.7	4.6	523 422.8	506 275
After SFs and corrections	1.8	5.5	4.4	509 922.4	490 267

Table 4.15. Cutflow for the single-lepton semi-boosted channel, Run 3.

Selection	Signal					Background	Data
	SM	$\kappa_{2V} = 1.5$	$\kappa_{\lambda} = 10$	$\kappa_{2V} = 0.25$	$\kappa_{2V} = 0.75$		
All events	475.7	926.9	983.6	1393.8	570.6	66 458 607.4	55 138 503
MET filters	474.7	923.5	981.6	1386.8	568.9	66 371 065.4	54 841 105
Trigger selection	100.3	218.4	221.5	347.1	126.4	22 467 958.9	20 103 915
1-lepton selection	51.6	123.0	116.4	204.0	67.3	15 871 351.2	12 960 875
Exactly 1 fat jet	8.2	42.7	23.0	85.5	16.4	5 354 917.3	4 120 670
≥ 4 jets	4.3	16.8	11.0	31.9	7.2	1 165 441.0	951 872
Boosted candidate selection	4.3	16.8	11.0	31.9	7.2	1 165 441.0	951 872
VBS candidate selection	4.3	16.8	11.0	31.9	7.2	1 165 441.0	951 872
Resolved candidate selection	3.7	14.1	9.7	26.8	6.2	1 094 942.3	907 917
After SFs and corrections	3.6	13.9	9.4	26.4	6.1	1 085 799.5	881 878

Kinematic Distributions

The principal observables are the tagger scores and regressed mass of the single boosted candidate, the invariant mass and angular separation of the leading resolved dijet, the minimum lepton- b -jet mass $m_{\ell b}$, and the forward tagging-jet mass and rapidity gap; they are collected, for both eras, in Appendix A.

Multivariate Analysis

The classifier scores are shown in Figure 4.18. Because the single AK8 jet is classified as either the Higgs or the vector boson but never both, the two tagger requirements are combined with a logical OR, rather than an AND, in the signal-region scan. The per-region significances are lower than in the boosted channels, with only one boosted boson and the rest resolved, the discrimination available to the scan is weaker, and the channel reaches an observed (median expected) 95% CL upper limit of $\mu < 9.82$ (7.13) in Run 3 (Table 4.18), roughly five to six times weaker than the fully-boosted channels. It nonetheless covers a region of phase space inaccessible to the fully boosted channels and, recovering the largest fraction of SM-like signal, contributes in the combination of all channels. The signal-region definitions and yields from the scan are listed in Tables 4.16 and 4.17.

4.5.8 Statistical Methods

Once an event sample has been reconstructed and the events of interest counted, the final step is to infer the underlying physics: to estimate a parameter, to quantify the significance of a possible signal, or to set a limit when no signal is observed. This is done within a frequentist framework built on the likelihood function, and the same machinery underlies both the expected-sensitivity figures of merit used to optimize an analysis and the final results extracted from data. This section sets out the formalism used to extract the results of this search.

Table 4.16. Signal-region definitions and yields for the single-lepton semi-boosted channel, Run 2.

Region	Definition	Z	Bin	Run 2		
				Signal	Background (MC)	Data
1	$H > 1.00$ OR $V > 0.96$, BDT > 0.88 , DNN > 0.98	0.28	A	0.0717	0.0456	0 ($\hat{N}_A = 0.05$)
			B	0.0651	30.5608	2
			C	0.0711	55.4490	38
			D	0.1194	2554.5027	1430
2	$H > 0.84$ OR $V > 0.40$, BDT > 0.84 , DNN > 0.98	0.15	A	0.3252	4.6651	0 ($\hat{N}_A = 1.66$)
			B	0.2563	116.8276	53
			C	0.7545	2229.8968	2146
			D	1.2465	73714.7818	68560
3	$H > 0.66$ OR $V > 0.38$, BDT > 0.96 , DNN > 0.98	0.17	A	0.0063	0.0003	0 ($\hat{N}_A = 0.03$)
			B	0.0245	16.9576	19
			C	0.0093	35.0372	30
			D	0.1710	17731.6827	17315
4	$H > 0.66$ OR $V > 0.18$, BDT > 0.96 , DNN > 0.98	0.09	A	0.0059	0.0024	1 ($\hat{N}_A = 0.06$)
			B	0.0561	41.0793	40
			C	0.0316	65.5015	81
			D	0.5234	50741.7299	57554
5	$H > 0.66$ OR $V > 0.12$, BDT > 0.96 , DNN > 0.94	0.07	A	0.0070	0.0074	0 ($\hat{N}_A = 0.11$)
			B	0.0305	82.8747	84
			C	0.0091	29.7587	40
			D	0.1702	25689.8454	29389

Table 4.17. Signal-region definitions and yields for the single-lepton semi-boosted channel, Run 3.

Region	Definition	Z	Bin	Run 3		
				Signal	Background (MC)	Data
1	$H > 0.94$ OR $V > 0.96$, BDT > 0.96 , DNN > 0.96	0.95	A	0.2202	0.0131	0 ($\hat{N}_A = 0.31$)
			B	1.1865	163.7005	78
			C	0.0769	39.8628	26
			D	0.9837	10317.3808	6599
2	$H > 0.64$ OR $V > 0.12$, BDT > 0.94 , DNN > 0.98	0.28	A	0.6096	4.6095	6 ($\hat{N}_A = 3.06$)
			B	1.9039	879.0806	540
			C	0.9347	2953.9565	2035
			D	5.4828	471289.5263	359484
3	$H > 0.52$ OR $V > 0.12$, BDT > 0.96 , DNN > 0.96	0.10	A	0.0045	0.0009	0 ($\hat{N}_A = 0.15$)
			B	0.0391	94.0928	57
			C	0.0061	82.5079	59
			D	0.1067	25122.1550	21707
4	$H > 0.50$ OR $V > 0.10$, BDT > 0.98 , DNN > 0.52	0.07	A	0.0013	0.0001	0 ($\hat{N}_A = 0.12$)
			B	0.1535	4505.5385	3219
			C	0.0000	3.4934	1
			D	0.0844	36432.4722	26908
5	$H > 0.10$ OR $V > 0.10$, BDT > 0.96 , DNN > 0.96	0.06	A	0.0088	0.0225	0 ($\hat{N}_A = 0.49$)
			B	0.0612	320.2235	190
			C	0.0271	529.8623	298
			D	0.4362	167239.2140	116515

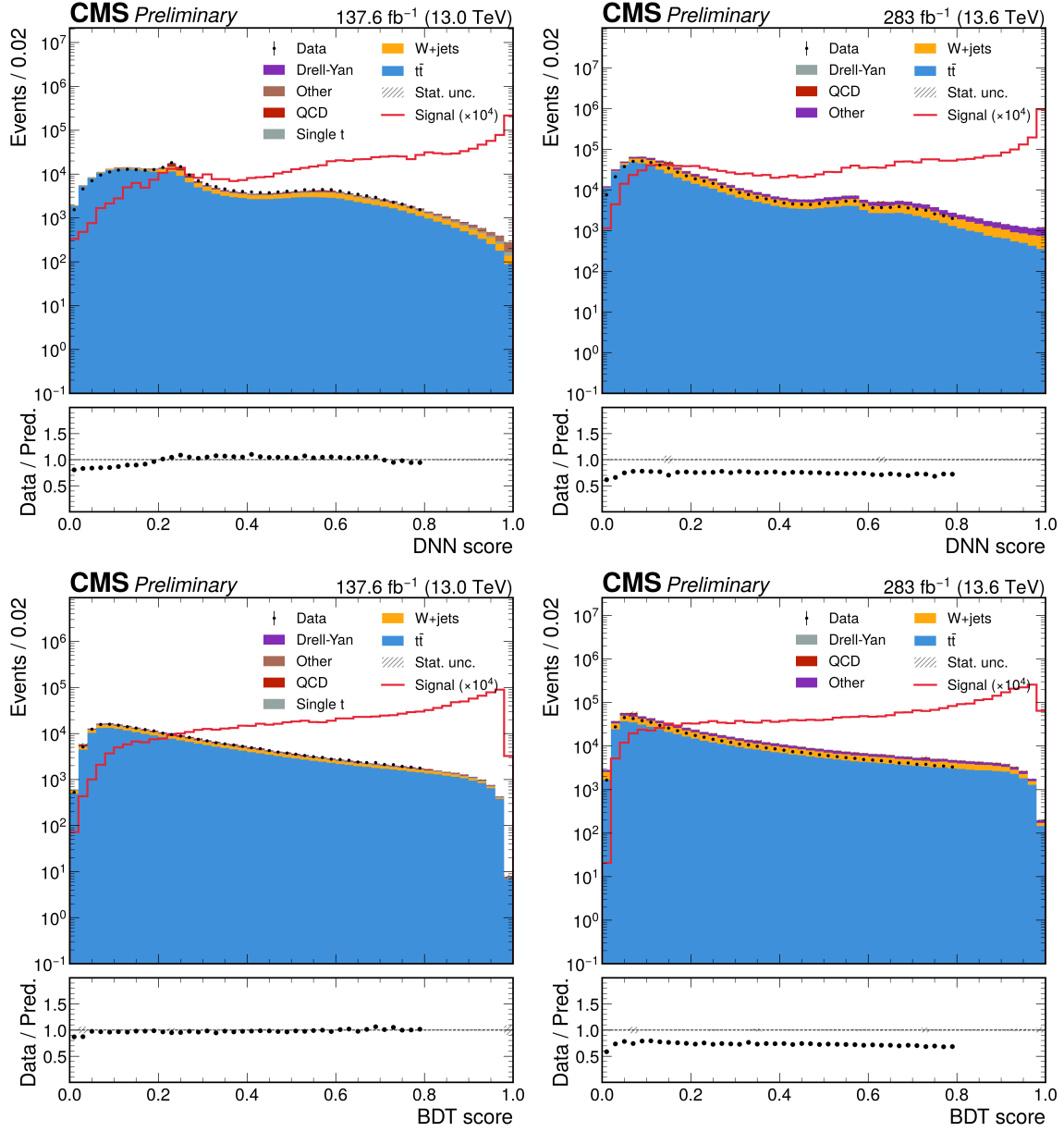


Figure 4.18. DNN (top) and VBS BDT (bottom) scores, single-lepton semi-boosted channel, Run 2 (left) and Run 3 (right).

The Likelihood Function

A modern search is typically binned: the selected events are distributed across a set of bins or regions indexed by i (for example the signal and control regions of Section 4.5.4). In each bin the observed count n_i is treated as a Poisson variable whose mean is the sum of the expected signal and background yields. The signal is scaled by a parameter of interest, the signal strength μ , or equivalently a coupling modifier such as κ_{2V} , while the systematic uncertainties (Section 4.5.10) are represented by a set of nuisance parameters θ that shift the predicted yields and are themselves constrained by auxiliary measurements. The likelihood is the product over all bins of the Poisson probabilities, multiplied by the constraint terms $\pi(\theta_j)$,

$$L(\mu, \theta) = \prod_i \frac{[\mu s_i(\theta) + b_i(\theta)]^{n_i} e^{-[\mu s_i(\theta) + b_i(\theta)]}}{n_i!} \prod_j \pi(\theta_j), \quad (4.10)$$

where s_i and b_i are the expected signal and background yields in bin i . The method of maximum likelihood estimates the parameters by the values $\hat{\mu}, \hat{\theta}$ that maximize L ; these are the best-fit signal strength and nuisance parameters given the data.

The form of each constraint term $\pi(\theta_j)$ reflects the nature of the uncertainty it encodes. Normalization uncertainties that must remain positive, such as the integrated luminosity and the 35% ABCD non-closure of Section 4.5.10, are given log-normal constraints; uncertainties that shift a yield through a calibration parameter, such as the jet energy scale and resolution, are given Gaussian constraints on the underlying parameter; and uncertainties arising from the finite size of a simulated sample are described by gamma distributions, which correctly reproduce the Poisson statistics of the generated events. A nuisance parameter may also be left *unconstrained*, with no term $\pi(\theta_j)$ at all, in which case it is determined purely by the data. The control-region yields of the background estimate described next are of exactly this kind.

Where a background is estimated from data rather than from simulation, the yields b_i are themselves parameters of the fit rather than fixed inputs. The search in this thesis is of that kind, and the resulting likelihood, in which each signal region is one corner of a data-driven ABCD

construction and its background rate is pinned to the observed control-region counts, is set out in Section 4.5.9 below.

Test Statistic and Profiling

Hypotheses about μ are tested with the profile likelihood ratio,

$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\theta}}(\mu))}{L(\hat{\mu}, \hat{\theta})}, \quad (4.11)$$

in which the numerator is maximized over the nuisance parameters at a fixed value of μ (the “profiled” values $\hat{\hat{\theta}}(\mu)$) while the denominator is the global maximum. Profiling is how the systematic uncertainties enter the result: allowing the nuisance parameters to adjust to the data broadens the likelihood and degrades the sensitivity in proportion to the size of the uncertainties, while also letting the data constrain those uncertainties in situ. The corresponding test statistic is $q_\mu = -2 \ln \lambda(\mu)$, large values of which indicate incompatibility between the data and the tested value of μ .

Expected Sensitivity and the Asimov Dataset

In the asymptotic regime of sufficiently large samples, the distribution of q_μ is known analytically, so the expected sensitivity of an analysis can be computed without costly ensembles of pseudo-experiments [43]. The central tool is the Asimov dataset: a single artificial dataset in which every observed count is set exactly equal to its expected value, so that the maximum-likelihood estimators return the true parameters. Evaluating the test statistic on the Asimov dataset yields the median expected significance and the expected limits directly. The discovery significance of a putative signal is $Z = \sqrt{q_0}$, the value obtained when testing the background-only hypothesis $\mu = 0$; it is precisely this Asimov significance that is maximized when the signal regions of the analysis are optimized (Section 4.5.4).

The closed form this significance takes for a single counting bin is given with the region-defining scan that uses it in Section 4.5.9, and the definition of the $\pm 1\sigma$ and $\pm 2\sigma$ bands with the

limits they accompany in Section 4.5.11.

Limit Setting and the CL_s Criterion

When no significant excess is observed, an upper limit is placed on the signal strength, and hence on the parameter of interest. To avoid excluding signal hypotheses to which the analysis has no genuine sensitivity, a pathology of a pure CL_{s+b} limit when the observed yield fluctuates below the background expectation, the limits are set using the modified-frequentist CL_s criterion [59, 76],

$$CL_s = \frac{CL_{s+b}}{CL_b}, \quad (4.12)$$

where CL_{s+b} and CL_b are the p -values of the signal-plus-background and background-only hypotheses. A value of μ is excluded at 95% confidence level when $CL_s < 0.05$. The coupling intervals quoted as the results of this thesis (Section 4.5.11) are obtained in exactly this way, by scanning the parameter of interest and identifying the range excluded at 95% CL.

Implementation

In practice, the construction of the likelihood of Eq. (4.10), the profiling of nuisance parameters, and the evaluation of the test statistics, expected significances, and CL_s limits are all carried out with the CMS statistical analysis and combination tool, COMBINE [36], built on the ROOFIT and ROOSTATS frameworks. The limits quoted in this thesis use its asymptotic method. COMBINE also enables the simultaneous fit of many regions and channels and their statistical combination, which is essential for an analysis whose sensitivity is distributed across the many final states described in Section 4.5.3.

4.5.9 Signal Region Definition and Fit

Where the previous search used a single signal region per channel, this analysis defines several orthogonal signal regions per channel. Within a kinematic phase space fixed by cuts on the Higgs and vector-boson tagger scores, the signal region is the corner A of the DNN–

BDT plane defined by a threshold on each score; the background there is predicted from the sidebands through Eq. (4.5). The region boundaries are optimized by a multidimensional scan over the tagger-score cuts and the two ABCD thresholds that maximizes the Asimov discovery significance,

$$Z = \sqrt{2 \left[(S+B) \ln \left(1 + \frac{S}{B} \right) - S \right]}, \quad (4.13)$$

where S and B are the expected signal and background yields in the trial signal region, both evaluated from simulation. This is the Asimov significance of Section 4.5.8; it reduces to the familiar $S/\sqrt{B}$ only in the limit $S \ll B$, which is emphatically not the regime here, where the expected yields in a signal region are of order one. The scan is performed iteratively: after the most significant region is found, its events are removed from the sample and the scan is repeated on the remainder, up to five times, yielding up to five mutually orthogonal signal regions per channel ordered by signal purity.

The final result is obtained from a binned maximum-likelihood fit of the signal and background predictions to the data across all regions and channels simultaneously, performed with the CMS COMBINE tool [36]. The general form of the likelihood, the profile likelihood ratio used to test hypotheses about μ , and the CL_s criterion used to set limits are those of Section 4.5.8; what is particular to this search is how the data-driven background enters, and that is described here.

Because the background is taken from data rather than from simulation, the expected yields of Eq. (4.10) are not supplied by Monte Carlo. Each channel–scan combination instead contributes a four-bin block corresponding to the four regions of the ABCD construction of Section 4.5.4. Within a block the yields of the three control regions B, C, and D are free parameters, unconstrained apart from the Poisson statistics of their own observed counts, while the background rate in the signal region A is not an independent parameter at all: it is fixed to the product $N_B N_C / N_D$ of Eq. (4.5) through a constrained parameter in the fit. The background prediction where the signal lives is therefore determined entirely by data in regions where it does

not, and its uncertainty follows automatically from the statistical precision of the control regions, propagated through that relation by the fit; the residual bias of the factorization assumption is covered separately by the non-closure nuisance parameter of Section 4.5.10. All such blocks, across every channel, scan, and era, are maximized in a single likelihood, so that nuisance parameters shared between them, the jet energy scale or the integrated luminosity for instance, are correlated across the whole combination.

One consequence of this construction shapes how the results below should be read. Because the control-region parameters are unconstrained and are driven by counts that are themselves small, they are among the parameters with the largest influence on the fitted signal strength: the search is limited by the statistical precision of its own control regions rather than by any externally supplied uncertainty. The data in the signal regions were kept blinded throughout the development of the analysis: the region definitions, the tagger and classifier thresholds, the ABCD scan, and the non-closure uncertainty were all fixed on simulation and on the control regions alone, and region A was read only once that procedure was frozen. The results presented in Section 4.5.11 are those of the subsequent unblinded fit.

Because the sensitivity of this search is concentrated at large κ_{2V} , where the cross section grows rapidly and the signal is boosted into the topologies selected here, the signal strength is quoted relative to the $\kappa_{2V} = 1.5$ benchmark rather than to the SM point,

$$\mu = \frac{\sigma}{\sigma(\kappa_{2V} = 1.5)}, \quad (4.14)$$

with all other couplings held at their SM values. This is the benchmark on which the channel selections were tuned (Section 4.5.3), and $\mu < 1$ therefore corresponds to excluding the $\kappa_{2V} = 1.5$ hypothesis. The constraint on κ_{2V} itself is obtained separately, from a scan of the upper limit on the VBS VVH cross section as a function of κ_{2V} against the theory prediction.

4.5.10 Systematic Uncertainties

The systematic uncertainties are incorporated as nuisance parameters in the likelihood. Because the background in every signal region is fixed to its control regions through the ABCD constraint of Section 4.5.9, the data-driven background estimate is largely self-protecting, and the systematic uncertainties act predominantly on the signal and on any residual simulation-derived component. Since each region enters the fit as a single counting bin, the uncertainties affect only the expected yields and not the shape of any distribution: each source is propagated either as a multiplicative event weight or, when the variation acts on the reconstructed objects themselves, by recomputing the full selection under the varied condition and taking the resulting change in acceptance.

The statistical model carries a small, well-defined set of dominant nuisances. The largest is the **ABCD non-closure**, a 35% log-normal uncertainty on the background prediction in each signal region, whose size is informed by the spread of the closure metrics measured in the control-region sub-divisions (Section 4.5.4); this is the leading background systematic. The **integrated luminosity** contributes a normalization uncertainty on the signal, and the finite size of the simulated signal samples enters as a per-region **signal Monte Carlo statistical** uncertainty. The control-region yields themselves float in the fit within Poisson bounds derived from their observed counts, so their statistical precision propagates directly into the background estimate.

Among the experimental sources acting on the signal, the corrections applied as event weights, the pileup reweighting, the L1 trigger pre-firing correction in Run 2, and the electron and muon reconstruction, identification, isolation, and trigger efficiency scale factors, are varied by one standard deviation and the resulting change in yield is taken as the uncertainty. The jet-related sources instead act on the object kinematics and are propagated through the selection. The jet energy scale is treated through the reduced set of eleven sources provided by the CMS Jet/MET group, of which six describe effects common to all data-taking periods and are correlated across eras and five describe period-specific effects and enter as one independent nuisance parameter per

era; each is varied in both directions, giving twenty-two variations, applied to both the AK4 and the AK8 jets. The scale shift is evaluated after the energy corrections but before the resolution smearing, with the smearing re-evaluated at the shifted momentum and seeded so that the same random number is drawn for the nominal and for every variation, so that the templates differ only by the calibration effect being probed. The jet energy resolution uncertainty is obtained by recomputing the smearing with the resolution scale factor varied within its uncertainty, and is decorrelated across eras since the scale factors are measured separately for each. All jet energy scale and resolution variations are propagated to p_T^{miss} by repeating the Type-1 rebuild with the varied jet momenta, and the unclustered remainder of p_T^{miss} carries its own scale uncertainty. Among the theoretical sources, the renormalization and factorization scale variations and the parton-shower initial- and final-state radiation weights are included.

Several sources remain to be added. The boosted GLOPART H -tagging and V -tagging efficiency scale factors, the associated AK8 jet mass scale and resolution, the UPART b -tagging scale factors, the hadronic H_T trigger scale factors, and the PDF uncertainty are not yet propagated into the limits reported below; for the PDF the per-event Hessian weight vector is stored in the signal samples but the corresponding variation has not yet been evaluated. The tagger scale factors are the most consequential of these, since the signal regions sit in the extreme tail of the tagger score distributions where the corrections are least well constrained. The limits of Section 4.5.11 should therefore be read as somewhat optimistic. Even so, given the very low signal and background yields in the signal regions, the search is limited primarily by the statistical uncertainty, with the ABCD non-closure the dominant systematic.

4.5.11 Results

With the region definitions and the background model frozen as described above, the signal regions were unblinded and the full likelihood fit was performed on the data. The results are quoted in two forms: an upper limit on the signal strength μ of Eq. (4.14), that is, on the rate of the $\kappa_{2V} = 1.5$ benchmark on which the selections were tuned, and a constraint on κ_{2V} itself

obtained from a scan of the cross-section limit against the theory prediction.

Limits on the signal strength

The 95% confidence-level upper limits on μ , computed with the asymptotic CL_s method, are given in Table 4.18 and shown in Figure 4.19, separately for each channel and era and for the full combination. The median expected limit quoted alongside each observed value is the limit obtained on the Asimov dataset, and the $\pm 1\sigma$ and $\pm 2\sigma$ bands are the corresponding quantiles of the distribution of the upper limit under the background-only hypothesis: the range within which the limit would fall in 68% and 95% of repetitions of the experiment were no signal present. Every channel–era fit is consistent with the background-only hypothesis: each observed limit falls within the $\pm 1\sigma$ band of its expectation, with a mild deficit in the two fully-boosted Run 3 fits and a mild excess in the single-lepton semi-boosted Run 3 fit.

Table 4.18. Observed and expected 95% CL upper limits on the signal strength $\mu = \sigma/\sigma(\kappa_{2V} = 1.5)$, per channel and era and for the combination of all three channels in both eras. The median expected values and the observed combined limit are set in bold.

Channel	Dataset	Observed	-2σ	-1σ	Median	$+1\sigma$	$+2\sigma$
All-hadronic, 3 FJ	Run 2	4.60	1.42	2.28	4.03	7.50	12.22
	Run 3	1.31	0.55	0.85	1.43	2.56	4.33
Single-lepton, 1 FJ	Run 2	7.22	4.23	6.23	10.03	16.99	28.29
	Run 3	9.82	3.09	4.48	7.13	11.95	19.84
Single-lepton, 2 FJ	Run 2	3.95	1.38	2.32	4.31	7.99	13.07
	Run 3	1.03	0.47	0.73	1.24	2.24	3.77
Combined		0.48	0.26	0.39	0.64	1.10	1.85

The combined fit gives an observed limit

$$\mu_{\text{obs}}^{95\%} = 0.48 \quad \left(\mu_{\text{exp}}^{95\%} = 0.64_{-0.25}^{+0.46} \right), \quad (4.15)$$

so that the $\kappa_{2V} = 1.5$ benchmark is excluded at 95% confidence level, with the observed limit about a factor of two below the benchmark rate. The exclusion was expected: the median expected limit and the whole of the -1σ to $+1\sigma$ band, $[0.39, 1.10]$, sit at or below unity. The

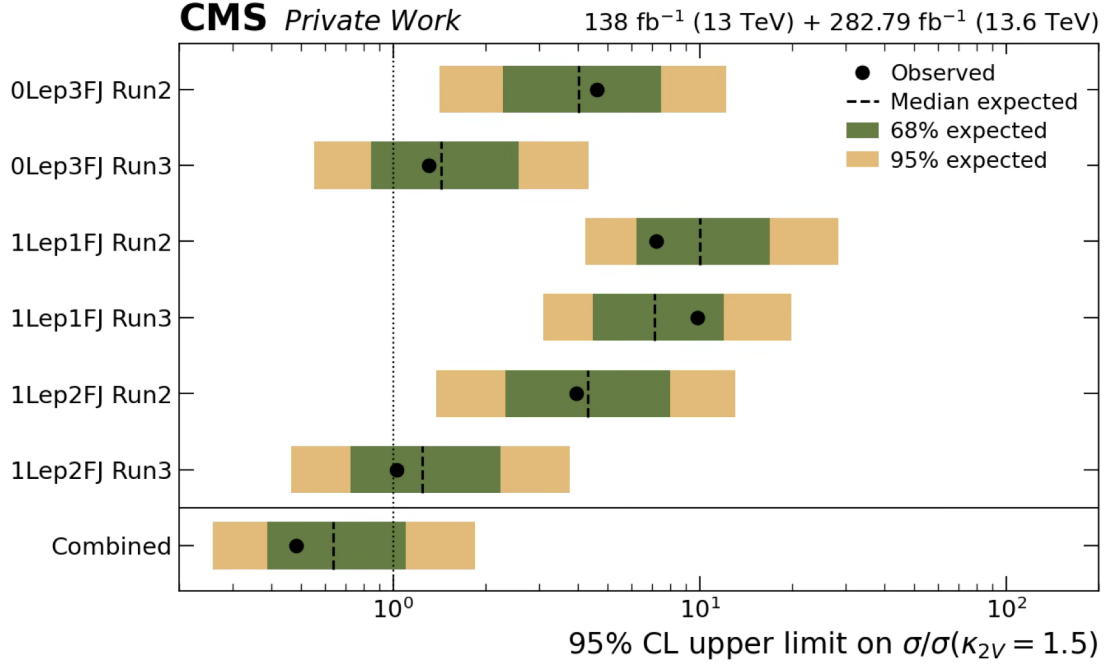


Figure 4.19. Observed and expected 95% CL upper limits on $\mu = \sigma/\sigma(\kappa_{2V} = 1.5)$ for each channel and era and for the combination of all six fits. The dotted vertical line marks $\mu = 1$, the rate of the $\kappa_{2V} = 1.5$ benchmark.

observed limit is slightly stronger than the median expectation, reflecting the small deficit of events in the most sensitive signal regions, where the leading regions of the two fully-boosted channels observe no data event against a background prediction of a few tenths of an event.

The combination is considerably stronger than any of its parts, no single channel–era fit reaching $\mu < 1$ on its own, because the six fits are statistically independent and each contributes several near-empty, high-purity signal regions. Among the individual channels, the two fully-boosted topologies dominate: the single-lepton two-fat-jet channel is the most sensitive, with a median expected limit of 1.24 in Run 3, followed closely by the all-hadronic three-fat-jet channel at 1.43, the only fully-hadronic channel currently contributing. The single-lepton semi-boosted channel is roughly a factor of five weaker, a direct consequence of the much larger $t\bar{t}$ background admitted when only one boosted boson is required, but it covers a region of phase space inaccessible to the fully-boosted channels and recovers a substantially larger fraction of the SM-like signal, so it still contributes in the combination. In every channel the expected

sensitivity of the Run 3 dataset alone surpasses that of the full Run 2 sample, reflecting the roughly twice larger integrated luminosity, the approximately 10% larger signal cross sections at $\sqrt{s} = 13.6$ TeV, and the improved GLOPART taggers.

Constraint on κ_{2V}

Scanning the upper limit on the VBS VVH cross section as a function of κ_{2V} , with all other couplings held at their SM values, and intersecting it with the theory prediction gives the exclusion shown in Figure 4.20. From the Run 3 dataset alone, VBS VVH production is excluded at 95% confidence level for values of κ_{2V} outside the observed (expected) interval

$$0.612 < \kappa_{2V} < 1.387 \quad (0.573 < \kappa_{2V} < 1.428). \quad (4.16)$$

This is a substantial improvement on the published Run 2 result of Section 4.4.4, $0.40 < \kappa_{2V} < 1.60$ observed, narrowing the allowed interval by about a third with the Run 3 data alone. The improvement follows from the larger dataset, the improved GLOPART and UPART taggers, the additional analysis channels, and the iterative multi-region signal extraction of Section 4.5.4, and it is obtained before the Run 2 sample is added to the κ_{2V} scan. As in the signal-strength fit, the observed and expected intervals agree closely, the observed being slightly the narrower of the two on both sides, consistent with the mild deficit seen in the combined signal-strength limit. In addition to the primary κ_{2V} target, the same fits are used to scan benchmark points in the Higgs self-coupling κ_λ , extending the interpretation of these datasets to that coupling.

4.5.12 Summary and Outlook

This chapter has presented a search for VBS VVH production in the full Run 2 dataset and in 282.79 fb^{-1} of Run 3 data, reprocessed in a common format and analyzed with a single set of object definitions. Three channels were developed in detail: the all-hadronic fully-boosted topology, the single-lepton fully-boosted topology, and the single-lepton semi-boosted topology. In each, the two decorrelated classifiers of Section 4.5.4 define an ABCD plane in which the

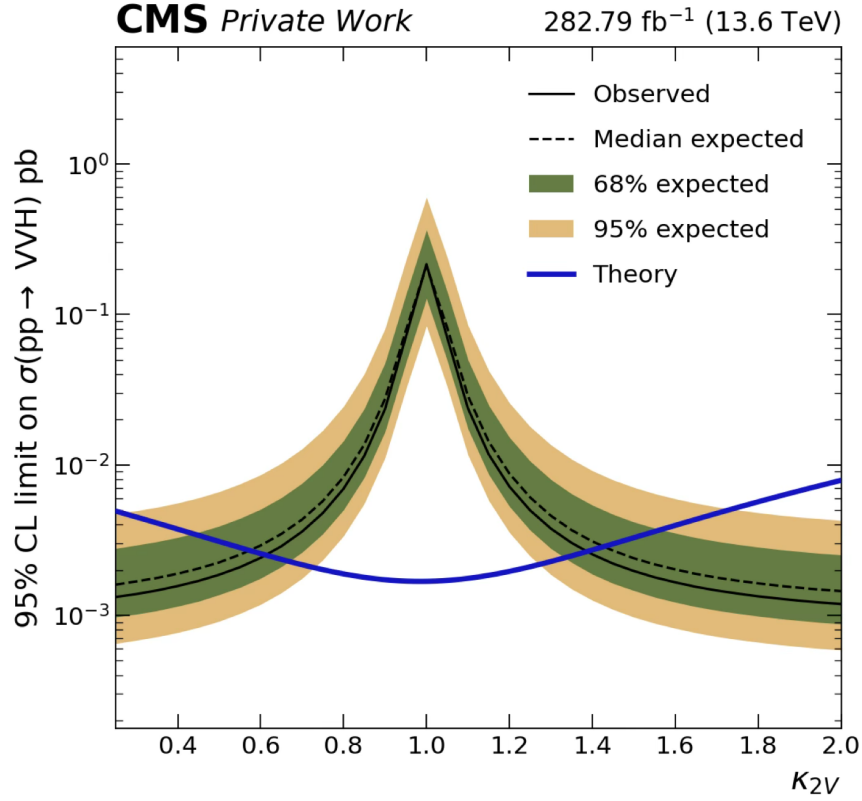


Figure 4.20. Observed and expected 95% CL upper limit on the VBS VVH production cross section as a function of κ_{2V} , from the Run 3 dataset. The blue curve is the theory prediction; the values of κ_{2V} at which the observed limit crosses it delimit the excluded region.

background is predicted from the data, and an iterative scan carves out up to five orthogonal signal regions per channel and era, all of which are fit simultaneously. Unblinding the signal regions, the combination of the three channels and both eras sets an observed 95% CL upper limit of 0.48 on the rate of the $\kappa_{2V} = 1.5$ benchmark, against an expectation of 0.64, and the Run 3 data alone constrain $0.612 < \kappa_{2V} < 1.387$, a marked improvement on the published Run 2 interval.

Several avenues remain open. The most immediate is the completion of the systematic model: the GLOPART and UPART tagger scale factors, the AK8 jet mass scale and resolution, and the PDF uncertainty are yet to be propagated, and their inclusion will loosen the limits quoted here by an amount that is expected to be modest given how strongly the search is dominated by statistical uncertainty, but that must be quantified before the result is final. Beyond that, the

remaining channels of Table 4.4, in particular the semi-resolved all-hadronic and the dilepton and multilepton topologies, are still to be brought into the combination, and the κ_{2V} scan is still to be extended to the Run 2 sample. The κ_λ interpretation of the same fits is likewise being developed. Taken together, these are expected to sharpen the constraint further, and the resulting combination will constitute the final result of this search.

Acknowledgements

Chapter 4, in part, is a reprint of the material as it appears in "Search for associated production of a Higgs boson and two vector bosons via vector-boson scattering at $\sqrt{s} = 13$ TeV," Physical Review Letters, 2026, A. Hayrapetyan *et al.* (CMS Collaboration) [5, 40]. It also contains material from an analysis of the combined Run 2 and Run 3 datasets that is currently being prepared for publication by the CMS Collaboration.

Chapter 5

Tracking Performance and Upgrades for the High-Luminosity LHC

5.1 Line Segment Tracking (LST)

Line Segment Tracking (LST) [83] is a bottom-up, intrinsically parallelizable pattern-recognition algorithm developed to augment or supersede the Combinatorial Kalman Filter (Section 3.2.1) in the Phase-2 Outer Tracker (OT). As discussed in Section 3.2.1, the combinatorial track building of the CKF is fundamentally sequential and its cost scales steeply with hit density, making it one of the dominant contributors to reconstruction time at the HL-LHC. LST attacks this bottleneck by restructuring track finding into a hierarchy of purely local construction steps, each of which can be executed concurrently across the entire detector. The algorithm is written against the Alpaka portability layer so that the same source runs, without modification, on both CPUs and GPUs, and it has been integrated and validated within the CMS software as an external package for the Phase-2 High Level Trigger [2].

The Phase-2 Outer Tracker and stubs

LST is designed around a distinctive feature of the Phase-2 OT geometry. Each OT layer is instrumented with p_T modules, built from two closely spaced silicon sensors, either two strip sensors (2S modules) at larger radius or a macro-pixel paired with a strip sensor (PS modules) at smaller radius. A charged particle traversing the module deposits a hit in each sensor,

and because the trajectory is bent in the 3.8 T solenoidal field, the displacement between the two hits within a single module encodes the local bending, and hence a coarse estimate of the track's transverse momentum. The on-detector electronics exploit this to form a "stub" only when the two hits are compatible with a track above roughly 2 GeV, rejecting the overwhelming low- p_T background already at the front end and supplying stubs to the L1 trigger. The crucial consequence for pattern recognition is that every OT layer provides not just a position but also a built-in directional and momentum constraint; LST is built bottom-up precisely to exploit this local physics prior at every stage.

Hierarchical track building

LST reconstructs tracks by assembling a hierarchy of objects of increasing complexity, each formed from objects of the preceding level using only *local* geometric information. The construction is illustrated in Figure 5.1.

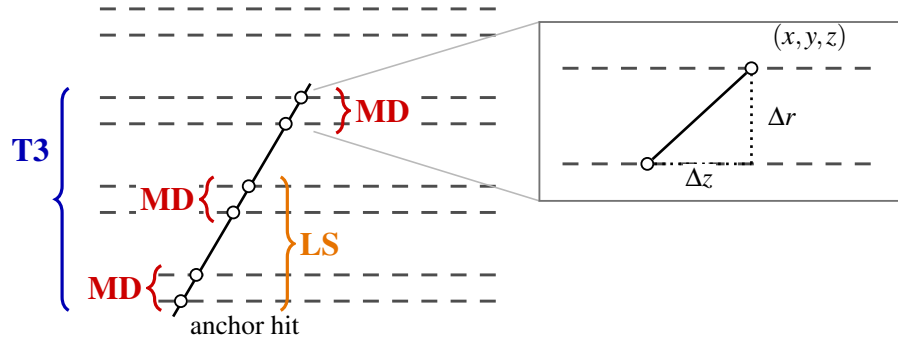


Figure 5.1. Construction of a LST T3 and underlying objects in the Phase-2 Outer Tracker.

Mini-doublets (MD). The two hits recorded by the paired sensors of a single p_T module are linked into a mini-doublet whenever their relative displacement is consistent with a track above a p_T threshold of 0.8 GeV, the offline counterpart of the on-detector stub. A mini-doublet is the elementary building block of LST: it fixes a space point and, through the intra-module displacement, an associated direction.

Line segments (LS). Two mini-doublets in neighboring layers are joined into a line segment

when their positions and directions are mutually compatible with a single charged-particle trajectory. The allowed pairings are restricted in advance through precomputed module connection maps, so that only geometrically sensible layer-to-layer links are ever considered, and the compatibility selections encode the trajectory constraints that physical tracks must obey.

Triplets (T3). Two line segments that share a common middle mini-doublet are merged into a triplet, a three-mini-doublet object spanning three layers. The shared mini-doublet enforces continuity, and the curvature implied by the three points provides a refined local p_T estimate.

Quintuplets (T5). Two triplets sharing a middle mini-doublet are combined into a quintuplet, spanning five layers. Quintuplets are the principal pure-OT track candidates: by requiring five consistent layers they achieve high purity without any input from the inner tracker.

Because each object is built solely from the local properties of its constituents, there are no sequential data dependencies between objects of the same type: every mini-doublet, every line segment, and every triplet in the event can be constructed simultaneously. This is what makes LST naturally suited to the massively parallel execution model of GPUs, in contrast to the branch-by-branch track building of the CKF.

Pixel seeding and final track candidates

The Inner Tracker (IT) pixel detector is incorporated through pixel line segments (pLS), seeds reconstructed from pixel hits in the spirit of the Patatrack pixel-tracking workflow [23]. Matching a pixel line segment to a triplet or a quintuplet in the OT produces a pixel-matched object, a pT3 or a pT5 respectively, that benefits from the excellent impact-parameter resolution of the pixel detector while retaining the long lever arm of the OT.

The final list of track candidates is assembled by combining these object types in a fixed priority order, pT5, then pT3, then pure-OT T5, and finally pixel line segments with at least

four hits, where an object later in the order is admitted only if its hits are not already used by an earlier, higher-priority candidate, which simultaneously enforces the priority and suppresses duplicates. The priority ordering reflects the relative purity of each category: pixel-matched objects supply the bulk of the prompt-track efficiency, while the pure-OT quintuplets, which do not require a pixel seed, extend acceptance to displaced trajectories whose production vertices lie centimeters from the beamline and which would be missed by pixel-seeded reconstruction. Unlinked pixel line segments recover efficiency in the forward region ($|\eta| \gtrsim 2.5$) where OT coverage runs out.

Parallelization and hardware portability

A central design goal of LST is to be hardware-agnostic. All performance-critical kernels are written once against the Alpaka abstraction layer and compiled to run either on NVIDIA GPUs or on multi-core CPUs, so that the algorithm can adapt to whatever accelerators are deployed at the HLT. The static parts of the combinatorics, in particular the module connection maps that define which layer pairs may form line segments, are precomputed and reused, leaving the per-event work dominated by the embarrassingly parallel construction of mini-doublets, line segments, triplets, and quintuplets.

LST Performance

In simulated $t\bar{t}$ events at $\langle\mu\rangle = 200$, configurations that build tracks with LST reach efficiencies comparable to, and at low momentum ($p_T \lesssim 5$ GeV) about 2% higher than, the baseline CKF, while delivering a substantially lower fake rate: up to roughly 60% lower for $p_T < 10$ GeV and up to about 80% lower in the range $0.5 < |\eta| < 2.5$, where combinatorial backgrounds are most severe [83, 2]. The pure-OT quintuplet candidates additionally provide acceptance for displaced tracks, exceeding the pixel-seeded baseline for production radii beyond about 2 cm and extending out to several tens of centimeters, with the efficiency falling off only once fewer than four OT layers remain. On GPUs the per-event throughput matches or exceeds

that of the CKF, up to about 35% faster in the quintuplet-based configuration, while on CPUs the algorithm incurs at most a $\sim 30\%$ slowdown; in both cases LST adds displaced-tracking reach without the steep CKF cost scaling, which is precisely what motivates its adoption for the Phase-2 HLT.

It is the triplet objects (T3s) introduced above, each already carrying a position, a direction, and a local p_T estimate, that serve as the physics-rich tokens for the machine-learning track reconstruction model developed in the following sections.

5.2 Machine Learning Based Track Reconstruction

In recent years, machine learning (ML) has emerged as a highly successful approach across many domains of high-energy physics, particularly within the event reconstruction chain. Techniques such as deep neural networks, graph neural networks (GNNs), and convolutional neural networks have demonstrated unprecedented performance in tasks ranging from calorimeter clustering and particle identification to energy regression. Building upon these remarkable successes, the application of ML methodologies to track reconstruction has garnered significant attention.

While ML-based sequences have proven highly effective for other reconstruction processes, this success has yet to be fully demonstrated for tracking within the CMS experiment. Several contender algorithms based on architectures like GNNs and Transformers are currently being explored, as they are inherently capable of learning the contextual relationships between inputs that are essential for accurate tracking. Although these models have shown considerable promise when applied to datasets with simplified geometries, translating this success to the complex, realistic geometry of CMS remains an ongoing challenge.

A primary difficulty with many of these ML tracking models is that they attempt to build tracks directly from individual detector hits without incorporating any prior physics information. This forces the model to learn all underlying physics and structural relationships from scratch

during training.

An alternative and promising approach is to tokenize the hits into larger, physically meaningful segments. By exploiting existing physics knowledge, such as the effects of the magnetic field and the expected constraints of the detector geometry, these segments provide the model with context-rich inputs. This tokenization simplifies the relationships the model must learn, which can in turn yield better physics performance. One such approach, detailed in the following sections, builds upon the line segment tracking algorithm described in Section 5.1. By utilizing the triplet objects created by LST as the tokenized input to a Transformer model, this method bridges the gap between traditional physics-driven reconstruction and modern machine learning capabilities. The remainder of this section first reviews the two neural-network architectures most relevant to tracking, graph neural networks and Transformers (Section 5.2), then describes the object condensation framework used to turn network outputs into reconstructed objects (Section 5.2), and finally presents a concrete realization of this strategy that clusters LST triplets with a Transformer trained under an object condensation objective (Sections 5.2 and 5.2) [41].

Neural Network Architectures for Tracking

The central difficulty of track reconstruction is one of pattern recognition: given an unordered collection of $\mathcal{O}(10^5)$ detector hits in a high-pileup event, the algorithm must decide which hits belong to the same particle trajectory. This is naturally a problem about *relationships* between inputs rather than about the inputs in isolation, and the two architectures that have proven best suited to it, graph neural networks and Transformers, are precisely those built to learn such relationships. Both also respect a property essential for tracking: the output must be invariant under any reordering of the input hits, since the labelling of hits carries no physical meaning.

A graph neural network (GNN) operates on data structured as a graph of nodes connected by edges [58]. For tracking, the detector hits (or higher-level objects) are taken as the nodes, and

edges are drawn between pairs of hits that could plausibly belong to the same track, typically based on geometric proximity or compatibility with a trajectory. Each node and edge carries a feature vector. The network then performs a sequence of *message-passing* steps: at each step every node aggregates information from its neighbors through the connecting edges, updating its own feature vector with a learned function. After several iterations each node’s representation encodes information about an extended neighborhood of the graph, and a final classifier acts on the edges, labelling each as ”true” (the two hits belong to the same particle) or ”false.” The set of edges surviving this classification partitions the graph into connected components, each of which is interpreted as a track candidate. Because message passing is defined locally and identically at every node, a GNN is permutation invariant by construction and scales to the large, sparse graphs characteristic of tracker data.

A Transformer dispenses with the explicit graph and instead learns the relationships between inputs directly through the *self-attention* mechanism [82]. Each input element, or token, is mapped to three learned vectors: a query $\mathbf{q}$, a key $\mathbf{k}$, and a value $\mathbf{v}$. The compatibility of token i with token j is measured by the scaled dot product of their query and key, and these scores, normalized across all tokens, weight a sum over the value vectors,

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V, \quad (5.1)$$

where Q , K , and V collect the query, key, and value vectors of all tokens and d_k is their dimension. Each token thus builds an updated representation as a learned, data-dependent average over all other tokens, allowing the model to capture long-range correlations in a single layer. In practice several such operations run in parallel as independent attention ”heads,” and the resulting layers are stacked to form a deep encoder. Like the GNN, the Transformer is permutation invariant, since no ordering of the tokens is assumed; its principal cost is that the full attention of Eq. (5.1) scales quadratically with the number of tokens, a limitation addressed below by restricting attention to geometrically nearby tokens.

Object Condensation

Both architectures produce, for each input, a learned representation; what remains is to convert these into a definite list of reconstructed objects whose number is not known in advance and varies event by event. Object condensation [61, 65] is a loss formulation designed for exactly this clustering-free, variable-multiplicity setting. Each input point is mapped by the network into a low-dimensional latent space and, in addition, is assigned a *condensation charge* $\beta_i \in [0, 1]$ that reflects how strongly that point should act as the representative, or condensation point, of an object. The charge enters the loss through

$$q_i = \operatorname{arctanh}^2 \beta_i + q_{\min}, \quad (5.2)$$

which grows sharply as $\beta_i \rightarrow 1$, so that a small number of high-charge points dominate.

To make the loss concrete, consider an event with N input points indexed by i and K true objects indexed by k , and let $M_{ik} = 1$ if point i belongs to object k and 0 otherwise. Within each object the algorithm selects the point of highest charge as that object’s *condensation point* α_k ,

$$\alpha_k = \arg \max_{i: M_{ik}=1} q_i, \quad (5.3)$$

with latent coordinate $\mathbf{x}_{\alpha_k}$ and charge q_{α_k} . The condensation point acts as the anchor toward which every other constituent of the object is drawn and away from which all foreign points are pushed.

The geometry of the latent space is shaped by two potentials defined relative to each condensation point. An *attractive* potential grows quadratically with distance, so that points of the same object are pulled into a tight cluster, while a *repulsive* potential is a hinge that acts only within unit distance and vanishes beyond it, so that points of different objects are separated but

distant clusters exert no force,

$$\check{V}_k(\mathbf{x}_i) = q_{\alpha_k} \|\mathbf{x}_i - \mathbf{x}_{\alpha_k}\|^2, \quad \hat{V}_k(\mathbf{x}_i) = q_{\alpha_k} \max(0, 1 - \|\mathbf{x}_i - \mathbf{x}_{\alpha_k}\|). \quad (5.4)$$

Both potentials are weighted by the charge q_{α_k} of the condensation point, so that high-confidence objects exert proportionally stronger forces. Summing the attractive contribution over an object's own points and the repulsive contribution over all foreign points, and weighting each term by the charge q_i of the point being acted upon, gives the potential loss

$$L_V = \frac{1}{N} \sum_{i=1}^N q_i \sum_{k=1}^K [M_{ik} \check{V}_k(\mathbf{x}_i) + (1 - M_{ik}) \hat{V}_k(\mathbf{x}_i)]. \quad (5.5)$$

The double charge weighting, by both q_i and q_{α_k} , is what makes the scheme self-organizing: the network is free to raise β only for the points it can place reliably, and those high-charge points then dominate the geometry of the latent space.

A second term governs the charges themselves. It pushes each object to nominate a clear representative by rewarding a high charge on its condensation point, and simultaneously suppresses the charge of *noise* points, those belonging to no true object, such as fake or spurious inputs, so that they neither anchor clusters nor survive the inference-time charge threshold,

$$L_\beta = \frac{1}{K} \sum_{k=1}^K (1 - \beta_{\alpha_k}) + s_B \frac{1}{N_B} \sum_{i=1}^N (1 - M_i) \beta_i, \quad (5.6)$$

where $M_i = \max_k M_{ik}$ indicates whether point i belongs to any object, $N_B = \sum_i (1 - M_i)$ is the number of noise points, and s_B is a tunable background-suppression strength. The full objective is the sum of the two contributions,

$$L = L_V + L_\beta. \quad (5.7)$$

Minimizing L organizes a chaotic cloud of inputs into well-separated clusters in latent space, each anchored by a condensation point that summarizes the object's properties, while

simultaneously learning a per-point charge that serves as a built-in noise filter. Crucially, the number of objects need never be specified: it emerges from the data as the number of distinct condensates, which makes the method well matched to track reconstruction, where the number of particles is itself one of the quantities to be determined.

LST Triplets as Tokens: Model and Training

The realization described here applies these ideas to the CMS Phase-2 Outer Tracker, taking as input the line segment tracking triplets (T3s) introduced in Section 5.1 rather than raw hits [41]. Using LST objects as tokens supplies the model with a strong inductive bias: each T3 already encodes geometric and physics-motivated quantities, most notably a local transverse-momentum estimate derived from the correlated hit pairs in the p_T modules, so that latent physics information is folded directly into the input basis. Each T3 is represented by a feature vector built from the spatial coordinates of the constituent hits, the angular and longitudinal differences within each mini-doublet and line segment, and the p_T , η , and ϕ of the triplet. In a $\langle\mu\rangle = 200$ event roughly 10^4 T3s are constructed, of which about half are fake, defined as triplets in which fewer than 75% of the constituent mini-doublets originate from a single simulated particle; this large fake fraction is precisely what the condensation noise term is designed to suppress.

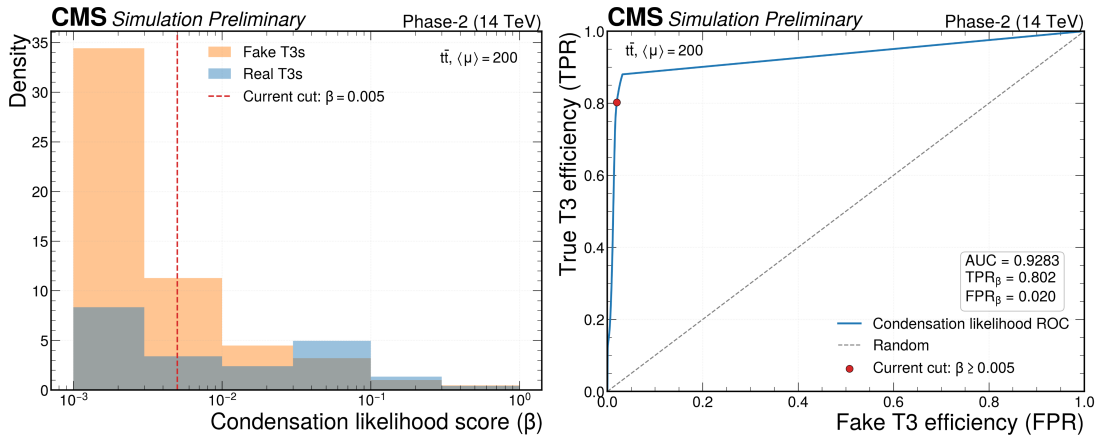


Figure 5.2. Learned condensation charge β for true and fake triplets (left) and its ROC as a true/fake classifier (right).

The model is a Transformer encoder that projects the input features to a hidden dimension

of 256 and maps them through six self-attention layers, each with eight attention heads and a 512-dimensional feed-forward sublayer, down to a final 128-dimensional latent space; these values were fixed by a systematic hyperparameter scan. A schematic of the model architecture is shown in Figure 5.3.

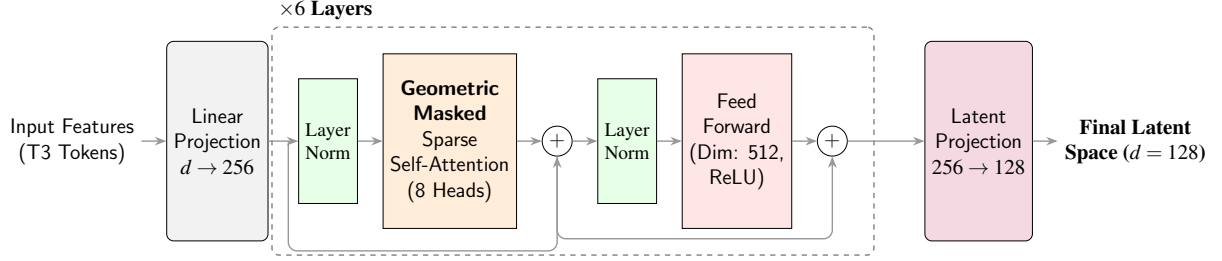


Figure 5.3. Architecture of the Transformer encoder.

To keep the quadratic attention of Eq. (5.1) tractable at the scale of 10^4 tokens, the self-attention is sparsified by a geometrically motivated mask: a given T3 attends only to other T3s whose nearest hits lie within $\Delta R < 0.1$ in η - ϕ space. This locality constraint reflects the physical expectation that triplets from the same track are geometrically close, and reduces the attention cost to a sparse operation without discarding relevant context.

Concretely, for each token i a neighbor set $\mathcal{N}(i)$ is defined from the angular separation of the triplets,

$$\mathcal{N}(i) = \{ j : \Delta R_{ij} \leq \Delta R_{\max}, e_j = e_i \}, \quad \Delta R_{ij} = \sqrt{(\eta_i - \eta_j)^2 + \Delta \phi_{ij}^2}, \quad (5.8)$$

where $\Delta \phi_{ij} = \text{atan2}(\sin(\phi_i - \phi_j), \cos(\phi_i - \phi_j))$ wraps the azimuthal difference into $(-\pi, \pi]$, and the constraint $e_j = e_i$ restricts attention to tokens of the same event. To bound the memory footprint the set is truncated to the k nearest neighbors in ΔR ; a token with no neighbor inside $\Delta R_{\max}$ attends only to itself, so that $\mathcal{N}(i)$ is never empty.

The attention of Eq. (5.1) is then evaluated only over these sets. For attention head h , with per-token query, key, and value vectors $\mathbf{q}_i^{(h)}, \mathbf{k}_i^{(h)}, \mathbf{v}_i^{(h)} \in \mathbb{R}^{d_h}$ of dimension $d_h = d/H$, the

normalization runs over $\mathcal{N}(i)$ rather than all tokens,

$$\alpha_{ij}^{(h)} = \frac{\exp(\mathbf{q}_i^{(h)} \cdot \mathbf{k}_j^{(h)} / \sqrt{d_h})}{\sum_{j' \in \mathcal{N}(i)} \exp(\mathbf{q}_i^{(h)} \cdot \mathbf{k}_{j'}^{(h)} / \sqrt{d_h})}, \quad j \in \mathcal{N}(i), \quad (5.9)$$

and the updated representation of token i concatenates the H heads and applies a learned output projection W_O ,

$$\mathbf{o}_i = W_O \bigoplus_{h=1}^H \sum_{j \in \mathcal{N}(i)} \alpha_{ij}^{(h)} \mathbf{v}_j^{(h)}, \quad (5.10)$$

where $\bigoplus$ denotes concatenation over heads. Because each token attends to at most $k \ll N$ others, the cost of Eqs. (5.9) and (5.10) scales as $\mathcal{O}(Nk)$ rather than the $\mathcal{O}(N^2)$ of dense attention, recovering the full-attention result of Eq. (5.1) in the limit $\Delta R_{\max} \rightarrow \infty$. Each sparse-attention block is wrapped in the standard pre-normalization residual structure, with a layer normalization preceding the attention and a two-layer feed-forward network, and the blocks are stacked to form the encoder.

The network is trained under the object condensation objective of Section 5.2, with the attractive, repulsive, and noise terms jointly pulling together T3s from the same particle, separating those from different particles, and rejecting fakes. Training uses roughly 15,000 simulated $t\bar{t}$ events at $\langle \mu \rangle = 200$ with a 60/40 training-validation split, run for about 100 epochs over approximately three days on a single NVIDIA A100 GPU, with all loss components converging smoothly to a stable plateau.

At inference the learned embeddings are clustered into track candidates with the density-based algorithm DBSCAN [48], which requires no prior specification of the number of clusters and so adapts naturally to the varying track multiplicity of each event; its neighborhood radius and minimum cluster size are optimized to $\varepsilon = 0.05$ and $N_{\min} = 2$. Prior to clustering, a threshold on the condensation charge removes low-confidence triplets as likely fakes, providing a first-pass reduction of the combinatorial background; at the chosen operating point this retains roughly 80%

of true T3s while rejecting all but about 2% of fakes. Each resulting candidate is characterized by the number of unique Outer Tracker layers it spans, N_L . Requiring $N_L \geq 5$ permits a direct comparison with LST quintuplets (T5s), which by construction span five layers, while relaxing the requirement to $N_L \geq 4$ probes the efficiency-versus-purity tradeoff and the recovery of shorter trajectories. A schematic of the full pipeline, from LST triplet construction through Transformer embedding and object condensation to final track candidates, is shown in Figure 5.4.

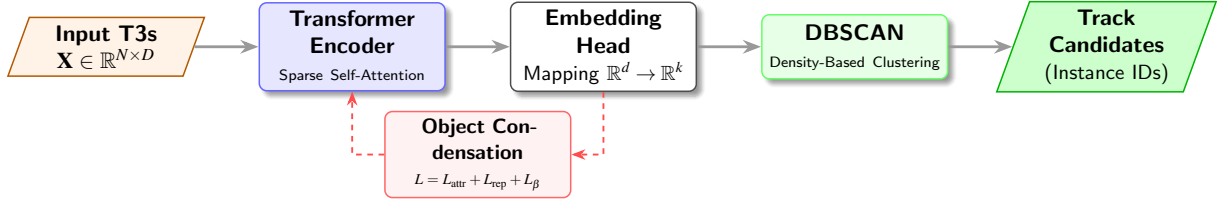


Figure 5.4. The ML track-reconstruction pipeline: LST triplets embedded by a Transformer under an object-condensation loss, then clustered with DBSCAN.

Model Performance

Performance is evaluated on simulated $t\bar{t}$ events at $\langle\mu\rangle = 200$ in the Phase-2 detector geometry, with reconstructed candidates associated to simulated particles through their constituent mini-doublets: a candidate is matched if more than 75% of its mini-doublets originate from a single simulated particle, and is otherwise counted as fake. Simulated particles from the hard-scattering vertex with $p_T > 0.8$ GeV and $|\eta| < 2.4$ define the denominator of the efficiency. Three metrics are reported as functions of η and p_T : the tracking efficiency, the fraction of simulated particles matched to at least one candidate; the fake rate, the fraction of candidates not matched to any particle; and the duplicate rate, the fraction of candidates that redundantly match a particle already reconstructed.

The distribution of the number of layers spanned by the reconstructed candidates peaks at six layers (Figure 5.5), exceeding the five-layer span of the baseline T5s and indicating improved track continuity.

The matched comparison is made at $N_L \geq 5$, the requirement that reproduces the five-

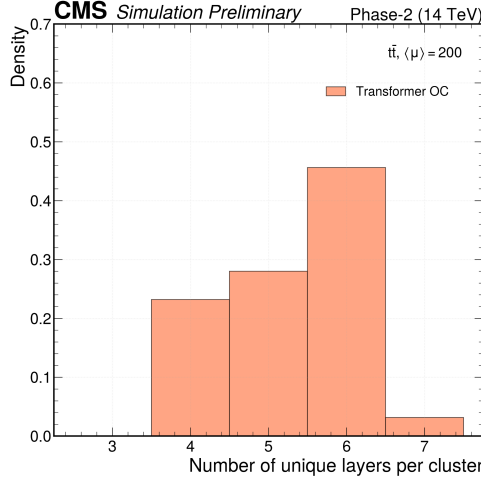


Figure 5.5. Number of Outer Tracker layers N_L spanned by the reconstructed candidates, peaking at six versus the five-layer T5 baseline.

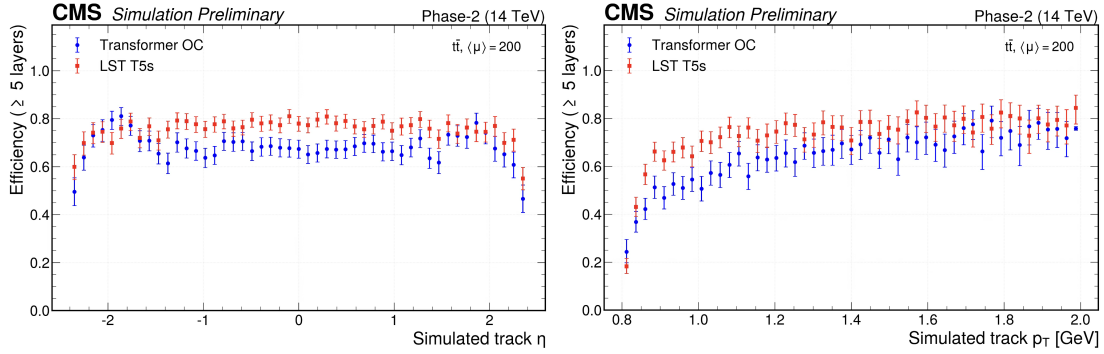


Figure 5.6. Tracking efficiency versus η (left) and p_T (right) for the Transformer + object-condensation candidates and the LST T5 baseline, under the matched $N_L \geq 5$ requirement.

layer span of a T5 by construction and therefore places the two algorithms on the same footing (Figures 5.6 and 5.7). At this working point the model trades efficiency for purity. Its efficiency sits systematically below the T5 baseline over the core of the acceptance, by roughly ten percentage points, about 0.67 against 0.78 for $|\eta| \lesssim 1.5$, with the gap closing towards $|\eta| \approx 2$; in p_T the deficit is largest at the low end and the two converge to within their uncertainties above roughly 1.4 GeV. The purity, on the other hand, improves markedly. The fake rate falls steadily with p_T , from a value comparable to the baseline near 0.8 GeV to about 0.05 at 2 GeV, while the T5 fake rate stays flat at 0.18–0.19 throughout, so the model is roughly three times cleaner over most of the range. The duplicate rate is the most striking difference: it is consistent with

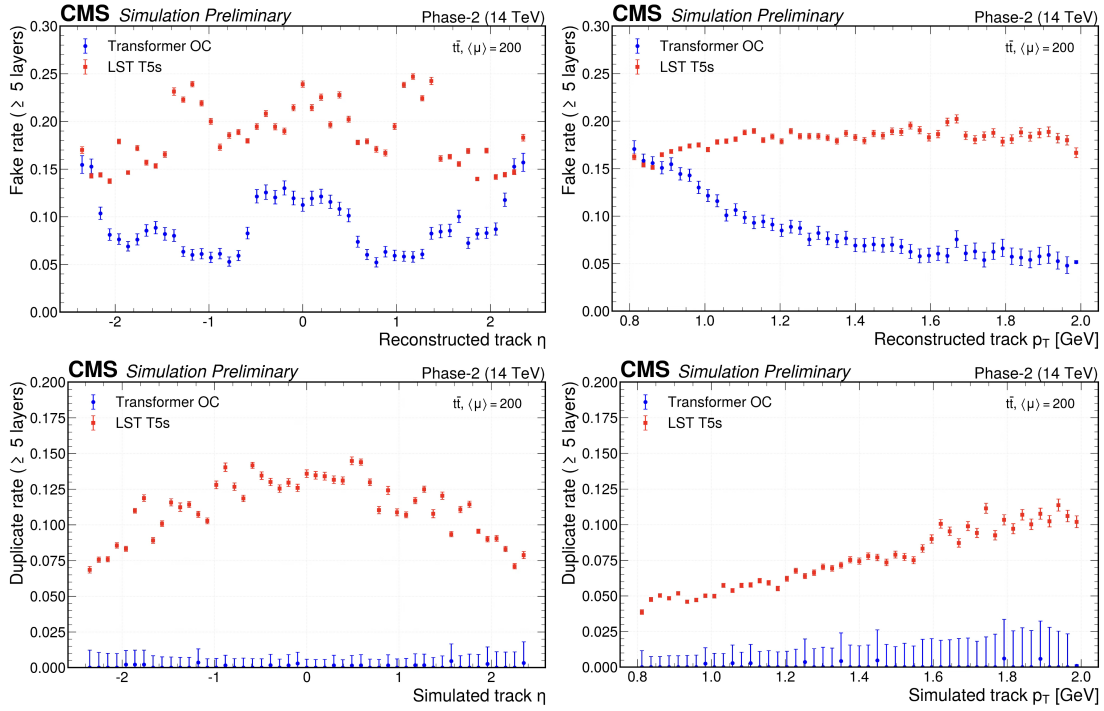


Figure 5.7. Fake rate (top) and duplicate rate (bottom) versus η (left) and p_T (right) for the Transformer + object-condensation candidates and the LST T5 baseline, under the matched $N_L \geq 5$ requirement.

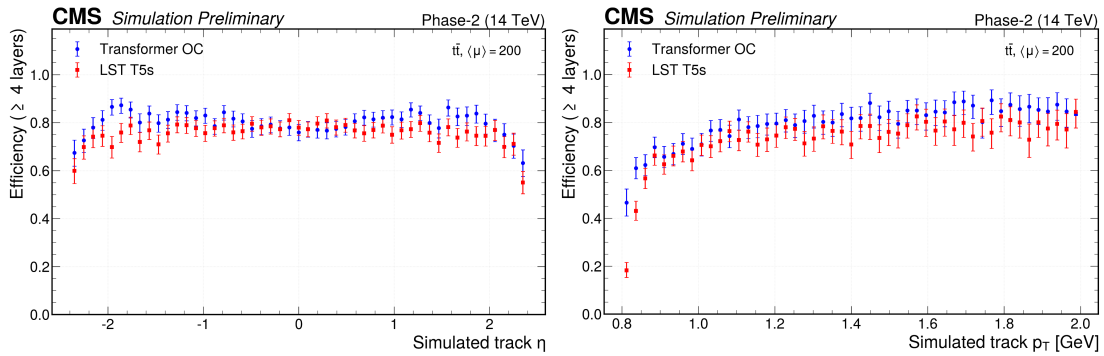


Figure 5.8. Tracking efficiency versus η (left) and p_T (right) for the Transformer + object-condensation candidates with the relaxed $N_L \geq 4$ requirement, compared with the same LST T5 baseline.

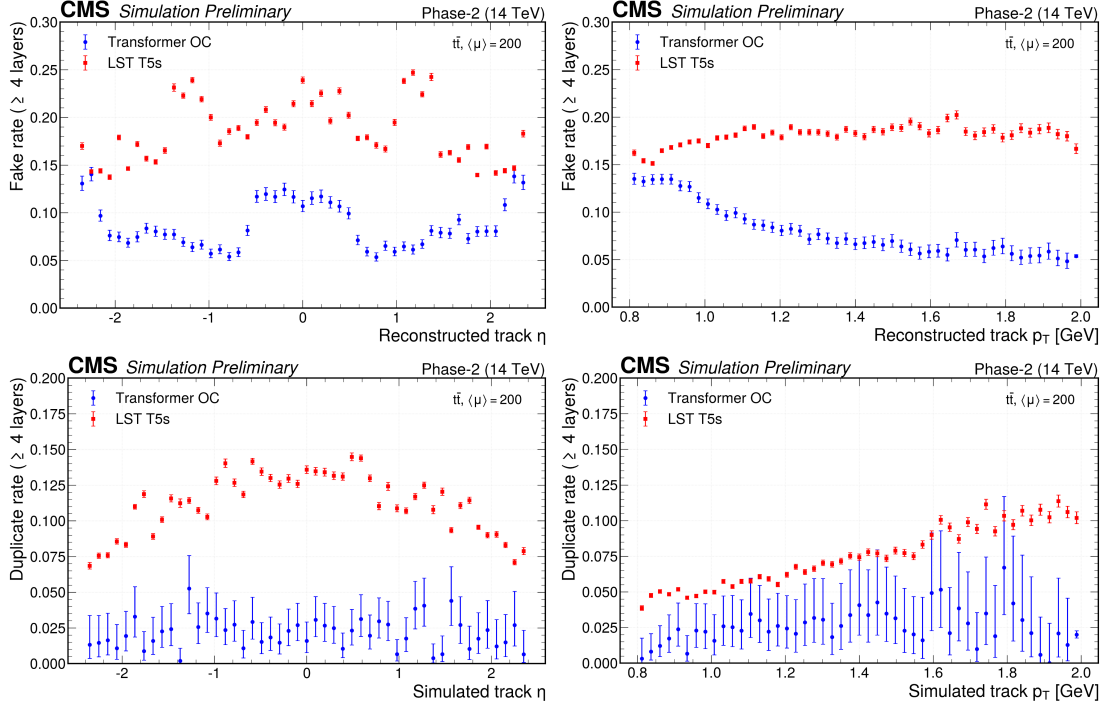


Figure 5.9. Fake rate (top) and duplicate rate (bottom) versus η (left) and p_T (right) for the Transformer + object-condensation candidates with the relaxed $N_L \geq 4$ requirement, compared with the same LST T5 baseline.

zero everywhere, whereas the T5 duplicate rate climbs from 0.04 to above 0.10 with increasing p_T . Object condensation assigns each particle a single condensation point by construction, so redundant candidates are suppressed by the design of the loss rather than by a downstream cleaning step.

Relaxing the requirement to $N_L \geq 4$ recovers the shorter trajectories that the matched selection discards, and with them the efficiency (Figures 5.8 and 5.9). The model then matches or exceeds the T5 baseline across the full range, reaching 0.80–0.87 against 0.70–0.80; the gain is concentrated at $1 \lesssim |\eta| \lesssim 2$, while in the central barrel the two are comparable. The fake and duplicate rates remain well below the five-layer baseline. This is the more representative operating point for a learned algorithm whose candidates are not constrained to a fixed layer span, and it indicates that the efficiency deficit seen at $N_L \geq 5$ reflects the layer-count requirement rather than a failure of the embedding to group hits correctly.

Taken together, these results constitute the first CMS demonstration of a machine-learning track reconstruction model built on Transformers and object condensation, and establish LST objects as an effective, physics-rich tokenization for learned tracking.

5.3 Summary

This chapter addressed the reconstruction of charged-particle trajectories at the extreme pileup of the HL-LHC, where the sequential, combinatorial track building of the Kalman filter (Section 3.2.1) becomes one of the dominant costs of event reconstruction. Line Segment Tracking reformulates pattern recognition in the Phase-2 Outer Tracker as a hierarchy of purely local construction steps, mini-doublets, line segments, triplets, and quintuplets, each built concurrently and each exploiting the built-in momentum constraint of the p_T -module stubs. Written once against the Alpaka portability layer, LST runs unchanged on CPUs and GPUs; it reaches the tracking efficiency of the Combinatorial Kalman Filter with a substantially lower fake rate, adds acceptance for displaced trajectories, and runs at or above CKF throughput on GPUs, which is why it has been integrated into the CMS software for the Phase-2 High-Level Trigger.

Building directly on that algorithm, the second half of the chapter presented the first CMS demonstration of a machine-learning track-reconstruction model founded on Transformers and object condensation. Rather than learning from raw hits, the model takes LST triplets as tokens, so that the geometry and local momentum already encoded in each triplet enter the network as a strong physics prior. A geometrically masked, sparse self-attention encoder embeds the triplets into a latent space under an object-condensation objective that pulls together triplets of the same particle, separates those of different particles, and suppresses fakes, after which DBSCAN forms the final track candidates without any prior on their number. Evaluated on $t\bar{t}$ events at $\langle\mu\rangle = 200$, the model matches the efficiency of the LST quintuplet baseline while achieving markedly lower fake and duplicate rates, and, when shorter trajectories are admitted ($N_L \geq 4$), recovers additional

efficiency in the barrel.

Together these results establish LST objects as an effective, physics-rich tokenization for learned tracking, and show that a physics-informed machine-learning model can match, and in places surpass, a state-of-the-art combinatorial algorithm under the harshest conditions foreseen at the HL-LHC. The natural next steps are to integrate the learned model into the CMS reconstruction and trigger software alongside LST, to measure its computational cost and throughput against the strict latency budget of the Phase-2 High-Level Trigger, and to extend the approach from the Outer Tracker toward the full tracking volume and a broader range of physics signatures.

Acknowledgements

Chapter 5, in part, is a reprint of the material as it appears in "Using Line Segment Tracking to enable ML-based Track Reconstruction in Phase-2 of the CMS Experiment," CMS Detector Performance Summary CMS-DP-2026-030, 2026, CMS Collaboration [41].

Chapter 6

Cyberinfrastructure for the LHC

6.1 Introduction: The HL-LHC Data Challenge

The physics program of the High-Luminosity LHC (HL-LHC, Section 2.1) rests not only on the detector and accelerator upgrades described in earlier chapters but also on the computing infrastructure that must store, move, and process the resulting data. The scale of the challenge is large: the CMS experiment alone is projected to record on the order of 0.5 exabytes of new data per year, and the associated processing and storage requirements (Figure 6.1) outpace what flat computing budgets can supply. Much of this data is replicated between the sites of the Worldwide LHC Computing Grid (Section 2.3), so the wide-area network (WAN) that connects them becomes as critical a resource as the processors and disks it serves.

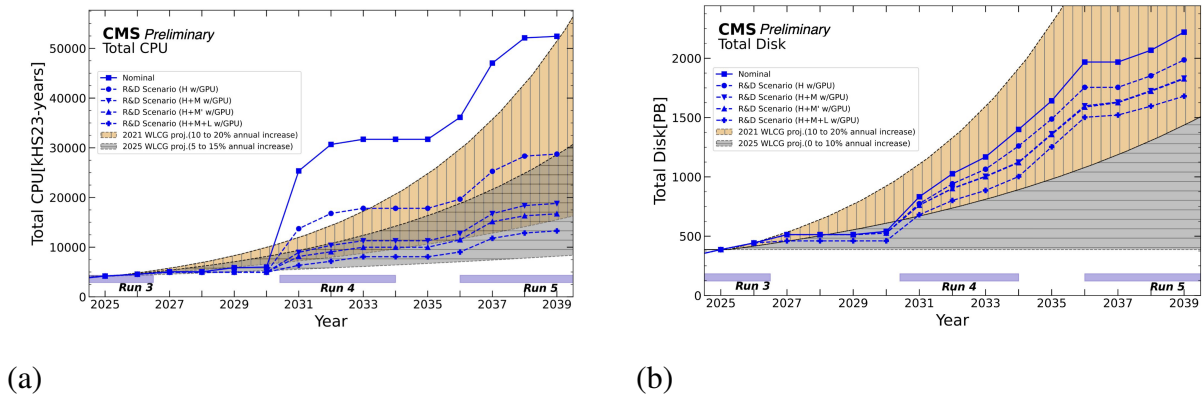


Figure 6.1. Projected CMS CPU (a) and disk (b) requirements for the HL-LHC, against a flat budget [3].

Yet while grid resource management agents schedule processors and manage storage explicitly, the network has historically been treated as an opaque, best-effort resource: every transfer competes for bandwidth on equal terms, with no notion of priority or guarantee [54, 86]. Bulk production transfers carried over National Research and Education Networks (NRENs) therefore experience throughput that oscillates unpredictably between excellent and poor. The consequences propagate through the system: when the WAN congests, remote processors sit idle waiting for input, and sites are forced to over-provision local disk to buffer workflows against the network’s variability. At present these consequences are tolerable: current data volumes leave enough slack that idle processors and generously sized buffers can absorb the network’s variability without threatening the physics program. In the HL-LHC era this slack disappears. The sheer volume of data to be moved leaves no room for processors to sit idle waiting on a congested link, and no realistic amount of local disk can buffer against sustained variability at exabyte scale. What is today an inefficiency becomes a bottleneck, and the best-effort model that has served the grid so far ceases to be adequate.

Treating the network instead as a *schedulable* resource, one that can offer a guaranteed bandwidth between two sites for a bounded time, would let high-value transfers proceed deterministically and reduce the storage overhead built to hide network jitter. This transformation is made possible by software-defined networking (SDN) [66].

This chapter presents a first effort towards that goal. It first establishes, through dedicated benchmarking, that the existing file-transfer software stack can physically saturate next-generation 400 Gbps links (Section 6.2). It then describes the SENSE-Rucio architecture that elevates the network to a managed resource, and the Data Movement Manager (DMM) as the component that bridges grid data management and SDN orchestration (Sections 6.3 and 6.4). Finally, it reports the evaluation of the integrated system on a multi-domain testbed (Section 6.5).

6.2 Saturating the Data Plane: Benchmarking to 400 Gbps

Before the network can be orchestrated, it is necessary to confirm that the data plane, the file-transfer software and the storage servers that drive it, can physically fill a next-generation link. In the HL-LHC era the Tier-2 sites in the United States are expected to require burst capabilities of 400 Gbps for hours at a time, with a steady-state consumption of roughly 100 Gbps to their storage systems [70]. The projected production tool for this exabyte-scale movement is XRootD [45] performing third-party copy (TPC) over HTTPS, in which a controller instructs a source server to transfer directly to a destination server without the bytes passing through the controller. The goal of this study was to identify the limits of that stack in reaching the target rate. It built on a prior 100 Gbps comparison on the National Research Platform, which had shown XRootD HTTPS-TPC to perform on par with or slightly better than the legacy GridFTP protocol [50], and on a subsequent demonstration of 1 Tbps over low-latency links in a commercial cloud [11].

Testbed and Deployment

The testbed consisted of two XRootD clusters, one at the University of California, San Diego (UCSD) and one at the California Institute of Technology (Caltech), located roughly 120 miles apart and connected by a mostly dedicated link with an end-to-end capacity of 400 Gbps and a round-trip latency of about 3 ms. The Caltech cluster comprised thirteen data servers and the UCSD cluster two high-core-count servers (dual AMD EPYC 7763, 128 cores each); within each cluster the data servers register with a redirector that acts as a load balancer. All servers were deployed and managed with Kubernetes, and the test files (1 GB each) were served from an in-memory `tmpfs` volume to remove the local filesystem as a bottleneck. The XRootD configuration mirrored the CMS production systems, bootstrapping authentication with X.509 credentials and then performing the TPC using Macaroon tokens. Transfers were driven by a `gfal-copy`-based orchestrator running on a dedicated host, since the client itself

is CPU-intensive, with each client launching single-stream transfers concurrently through the redirector.

Saturating the 400 Gbps Link

By empirical tuning it was found that 40 streams per destination server, that is, $40 \times 13 = 520$ concurrent streams out of UCSD, were required to saturate the link, sustaining an aggregate throughput of 400 Gbps stably for about an hour (Figure 6.2). This established that a containerized XRootD deployment can reach the next-generation target rate, but only when driven at a concurrency far beyond that of any single transfer.

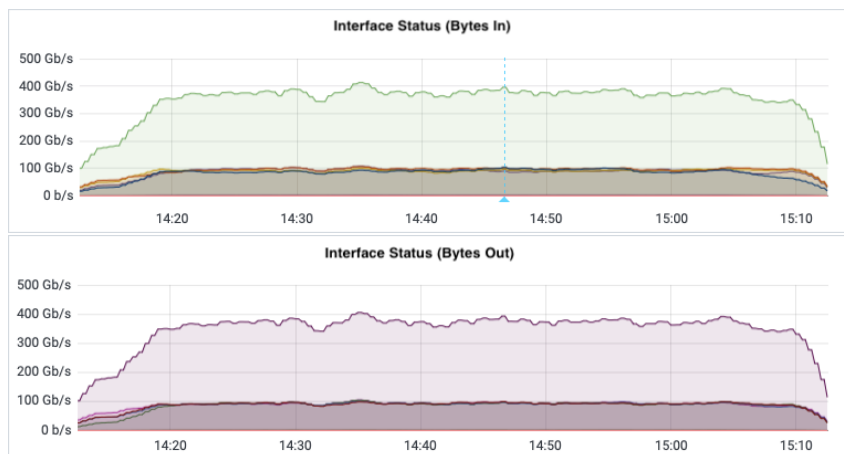


Figure 6.2. Aggregate XRootD HTTPS-TPC throughput sustained at 400 Gbps across the UCSD-Caltech link.

TCP Windows and the Bandwidth-Delay Product

That 520 concurrent streams were needed to fill a 400 Gbps link, rather than one fast stream, is not an accident of XRootD but a direct consequence of how TCP, the transport protocol underlying HTTPS, behaves on high-bandwidth paths [73]; understanding why makes sense of the parameterization that follows. TCP is a windowed, acknowledgement-based protocol: at any instant a sender may have at most one *window* of unacknowledged bytes in flight, after which it must pause until an acknowledgement returns from the receiver. Because each acknowledgement

takes one round-trip time (RTT) to arrive, the throughput of a single TCP stream is bounded by

$$\text{Throughput} \leq \frac{W}{\text{RTT}}, \quad (6.1)$$

where W is the effective window, the smaller of the sender’s congestion window and the receiver’s advertised window, the latter itself limited by the socket buffer allocated to the connection. The quantity $\text{BDP} = \text{Throughput} \times \text{RTT}$ is the *bandwidth-delay product*: the amount of data that must be simultaneously in flight to keep the pipe full [73]. Networks that combine high bandwidth with long delay, so-called “long fat networks,” have a correspondingly large BDP, and Eq. (6.1) shows that a single stream can fill such a link only if its window is at least as large as the BDP.

This is precisely the regime of wide-area LHC transfers, and it predicts a severe sensitivity to latency. Sustaining 100 Gbps across a transcontinental path of $\text{RTT} = 120$ ms requires $\text{BDP} = 100 \text{ Gbps} \times 120 \text{ ms} \approx 1.5 \text{ GB}$ in flight at once; against the modest socket buffers of a default kernel configuration, of order a few megabytes, Eq. (6.1) caps a single stream to well under a gigabit per second, orders of magnitude short of the link rate. The same window that saturates the 3 ms UCSD-Caltech hop (Section 6.2) is therefore useless across a CERN-scale delay. The measurements that follow bear this out, with the throughput of a fixed configuration falling roughly five-fold between 0.1 and 10 ms and bottoming out near 1 Gbps at 120 ms.

Two complementary remedies follow directly from Eq. (6.1), and both are exercised in the benchmarks below. The first is to enlarge the window itself: the TCP window-scaling option (RFC 1323) lifts the 64 kB ceiling of the original protocol, and provisioning socket buffers comparable to the BDP then allows a single stream to fill a long-latency pipe; the variable-latency study of Section 6.2 configures 1 GB socket buffers and 9 kB jumbo frames for exactly this reason, sizing the window to the largest BDP under test while reducing per-packet processing overhead. The second remedy is to parallelize: N concurrent streams contribute N independent windows, so the aggregate throughput scales as $\approx NW/\text{RTT}$, recovering the link rate even when each individual window is modest, which is why hundreds of streams, and not one, were needed

above. This is the mechanism behind the central trend seen in the sections that follow, that the number of streams required to saturate a link grows with its latency, and behind the observation that at long RTT the throughput rises only gradually with each stream added. Parallel streams carry a further advantage under packet loss: because loss-based congestion control reduces the congestion window sharply on each detected loss and then rebuilds it over many round trips, a single stream on a high-RTT path recovers slowly from even rare losses, whereas the aggregate of many streams is far less sensitive to any one of them.

There remains a ceiling that no amount of windowing or parallelism can raise, because it lies above the transport layer: once enough streams are in flight to fill the window, the binding constraint becomes the host's ability to process packets, set by CPU cores and memory bandwidth rather than by TCP. This will appear below as a single-server plateau near 260 Gbps even at zero added latency (Section 6.2) and as diminishing returns from additional streams once a core saturates. The practical consequence for production is that a transfer service cannot apply a single tuning everywhere; it must size both the socket buffers and the stream count to the bandwidth-delay product of each individual site pair, using few streams on short domestic hops and large buffers with high concurrency on transcontinental and intercontinental paths. It is exactly this per-circuit tuning of the FTS stream count that the Data Movement Manager automates, in proportion to each circuit's allocated bandwidth (Section 6.4).

Parameterizing the Throughput

The saturation test establishes that the target rate is reachable, but says nothing about the minimum resources needed to reach a given throughput. To expose those limiting factors, the achievable throughput was parameterized as a function of four quantities: the inter-site latency, the number of concurrent streams, the number of CPU cores allocated to each server, and the number of XRootD origins in the cluster. The aim was, in the ideal case, to obtain the minimum value of each parameter required to attain a specified throughput. Each parameter was varied systematically, with transfers run between the two identical high-core-count UCSD servers of

Section 6.2 (the dual AMD EPYC 7763 hosts) until the throughput stabilized and then stopped. The hosts' raw round-trip latency of 0.1 ms served as the baseline, and larger latencies were emulated with Linux Traffic Control (tc), which introduces an artificial delay by holding packets in a buffer before releasing them; CPU cores were allocated through Kubernetes. Because each server's network interface is rated at 100 Gbps, adding origins raises the aggregate bandwidth capacity proportionally, while CPU cores are counted across all origins and the requested streams are divided evenly among them, so that 100 streams across four origins is 25 streams per origin. One practical wrinkle was that tc did not cleanly relinquish the interface: fully reversing an emulated delay required rebooting the affected machine.

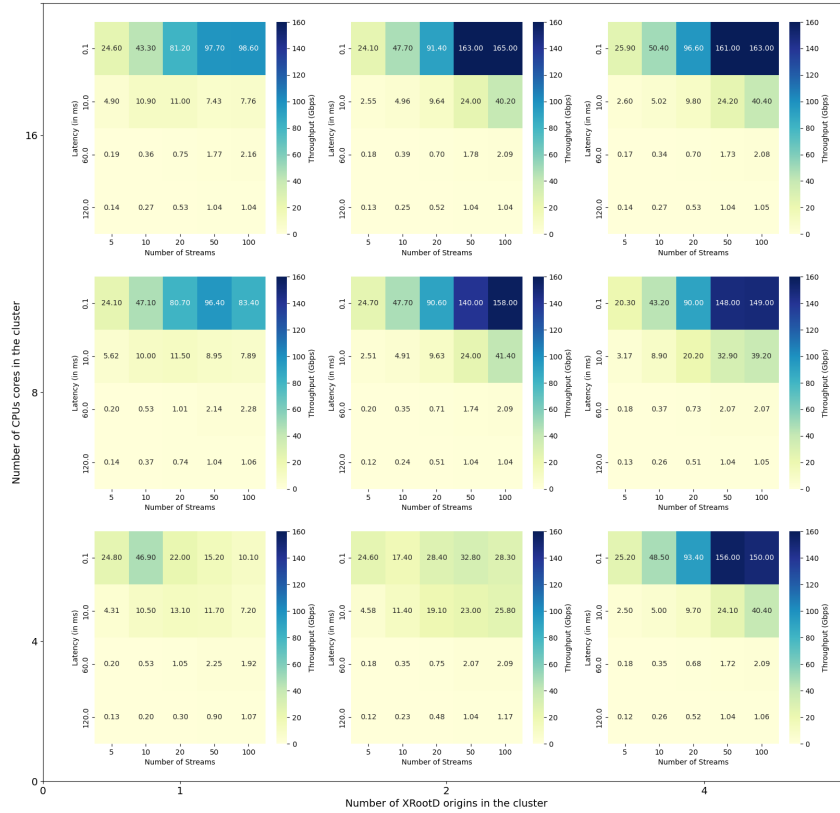


Figure 6.3. Throughput versus stream count and tc-emulated latency, for combinations of CPU cores and XRootD origins. Adapted from [10].

Three trends emerged (Figure 6.3). First, for a fixed number of CPU cores the throughput rises with the number of streams up to a point and then falls as CPU usage saturates, indicating a

maximum throughput per core set by packet-processing contention. Second, adding more origins while holding the core count fixed improves performance, attributable to better kernel-level load balancing across more processes. Third, throughput degrades severely with latency: increasing the round-trip latency from 0.1 to 10 ms reduces the data transferred roughly five-fold, and at 120 ms the throughput remains limited to about 1 Gbps even with two origins, full 128-core access, and up to 500 streams. Taken together, these results confirm that XRootD HTTPS-TPC comfortably sustains the throughput demanded by the HL-LHC era at low latency, and the trends suggest that with the right combination of streams, cores, and origins the target could also be met at higher latencies, though establishing this with confidence would require further study. They also expose a limitation of the method itself: the magnitudes produced by `tc` emulation proved unreliable for high-bandwidth, high-latency tests, which motivated repeating the characterization with real latencies (Section 6.2).

Extending to Variable, Real-World Latencies

The parameterization above emulated inter-site delay with Linux `tc`, and while the resulting trends were qualitatively correct their magnitude was not: the artificial delay produced throughput far below what real links of the same latency deliver. This matters because the round-trip time between any pair of US Tier-2 sites, and from them to CERN, spans roughly 5 to 120 ms, so confirming that XRootD scales across that entire range, and not merely on the short UCSD-Caltech hop, is a prerequisite for production. A follow-up study therefore repeated the characterization using *real* latencies rather than emulated ones [9]. Two identical data-transfer nodes (dual 32-core Intel Xeon Gold 6430, 2 TB of DDR5, and a 400 Gbps ConnectX-7 NIC, tuned for high-latency throughput with 1 GB socket buffers and 9 kB jumbo frames) were placed side by side at SDSC and then interconnected through a set of static routes that loop across the FABRIC testbed (Figure 6.4), using SENSE’s layer-2 and routing capabilities to realize a range of genuine round-trip times between 5 and 120 ms. SENSE’s quality-of-service guarantees reserved bandwidth on each route to isolate the measurements from competing testbed traffic.

As before, XRootD origins ran on a Kubernetes cluster serving 4 GB files from `tmpfs`, with transfers driven by parallel `gfal-copy` clients, and the study swept the number of streams, the CPU cores per origin, and the number of origins against each real latency.



Figure 6.4. The variable-latency testbed: two co-located SDSC nodes interconnected through static routes across FABRIC, realizing round-trip times of 5-120 ms.

The results sharpened the earlier findings and exposed hard limits. The throughput-versus-streams curve was found to depend strongly on latency: at short round-trip times throughput climbed steeply with added streams but collapsed just as sharply once a threshold was passed, whereas at long latencies both the rise and the fall were gentler (Figure 6.5(a)). A single XRootD origin could not reach 100 Gbps at any configuration, appearing unable to scale past roughly 16 cores; two origins reached 100 Gbps at low latency but not at high, and only with four origins across 64 total cores was 100 Gbps attainable even at the longest latencies, and then with fewer than 100 streams, confirming that distributing cores across more origins pays substantial dividends (Figure 6.5(b)). Proximity to the bandwidth ceiling was itself found to depress throughput before the ceiling is reached: a 200 Gbps target that was unattainable under a 200 Gbps allocation became attainable once the allocation was raised to 400 Gbps, all else equal. Finally, a single physical server was found to saturate at about 260 Gbps even at zero added latency and with hundreds of streams (Figure 6.6), establishing a per-host ceiling that any

production deployment must plan around. These measurements provide the empirical basis for the stream- and origin-count tuning that managed-transfer services such as FTS [15] and the DMM described later must perform to sustain throughput across a heterogeneous, high-latency grid.

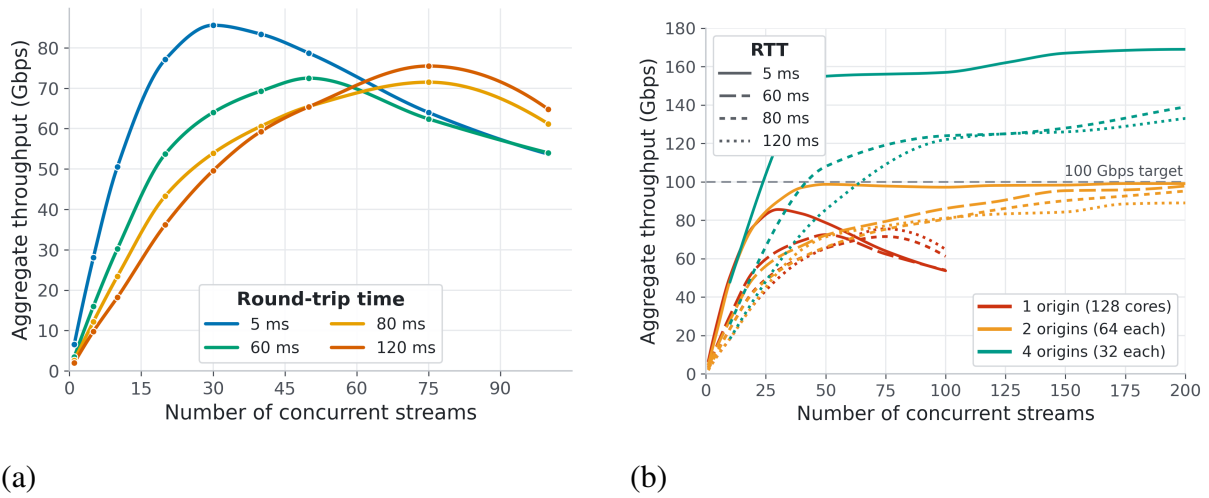


Figure 6.5. Throughput under real latencies: (a) versus stream count for several round-trip times; (b) progress toward 100 Gbps as the number of XRootD origins grows at fixed total core count [9].

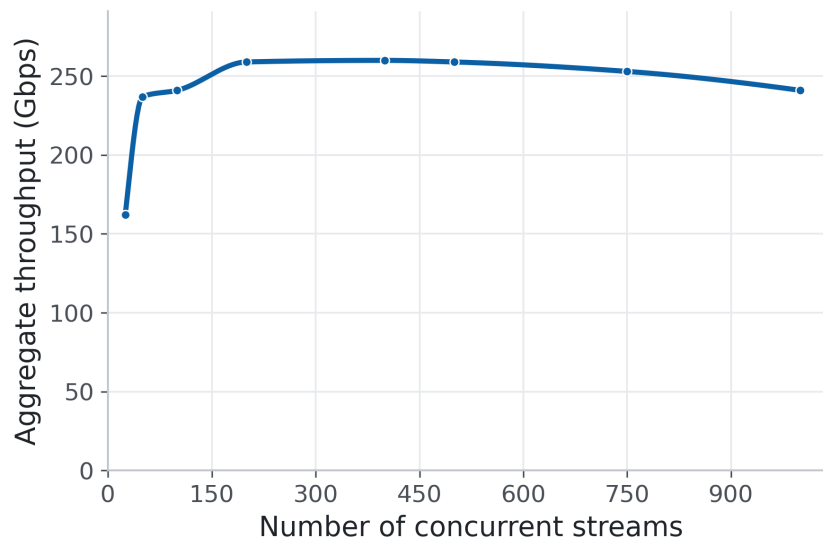


Figure 6.6. Single-server throughput versus stream count at zero added latency, plateauing near 260 Gbps [9].

Operational Lessons

Two practical findings inform the production design. Using the redirector for load balancing, rather than hand-pairing source and destination servers, introduced negligible overhead even at high concurrency, confirming that the redirector-based architecture is viable at scale. The choice of client tool mattered modestly: `gfal-copy` outperformed `curl` by about 15%, implying that production tools such as FTS [15], which use the `gfal` client, should achieve comparable performance. Taken together, these benchmarks confirm that the data plane can reach next-generation targets at low latency, but that doing so in an unmanaged fashion, with hundreds of competing streams and a steep sensitivity to latency, risks systemic instability. This motivates the orchestration architecture described next.

6.3 Network-Aware Data Movement: The SENSE-Rucio Architecture

The purpose of the SENSE-Rucio project is to let selected, high-value data transfers ride dedicated, bandwidth-guaranteed network circuits instead of the shared best-effort paths used by ordinary grid traffic [54]. Doing so requires connecting two production systems that, until now, knew nothing of one another: the experiment’s data-management system, which decides *what* data to replicate *where*, and an SDN control plane, which can build a network circuit of a requested capacity between two sites.

The control plane is provided by SENSE, the Software-Defined Network for End-to-end Networked Science at the Exascale [66], developed by the Energy Sciences Network. SENSE follows an *intent-based* model: an application declares *what* it needs, for example a guaranteed-bandwidth link between two sites for a fixed duration, and the SENSE Orchestrator (SENSE-O) translates that intent into the concrete, multi-domain network configuration that realizes it, provisioning a point-to-point virtual circuit across the backbone and configuring the routers and switches at each site. This elevates the network to a first-class schedulable resource, managed

alongside compute and storage. The operational model reserves a baseline fraction of each link’s capacity for ordinary best-effort traffic and dedicates the remainder to scheduled, high-priority circuits governed by explicit agreements, so that prioritization never starves the general-purpose grid [86]. What matters here is its external contract, an API through which a circuit of a given bandwidth can be requested, modified, and torn down; the remainder of this section sketches how SENSE realizes that contract.

How SENSE Works

SENSE distributes its function across two roles: a central *Orchestrator* (SENSE-O), one per science collaboration, and a set of domain-specific *Resource Managers* (RMs). Each RM owns the resources of a single administrative domain, a wide-area network such as ESnet or an end site with its data-transfer nodes, and is the authoritative source of truth for that domain’s topology and state. Every RM publishes a uniform *model* of its resources, and the Orchestrator periodically pulls these models and stitches them into a single end-to-end graph spanning every domain between source and destination. Because all information travels as models, the RM-facing API reduces to two operations, pull a model and push a change, that need not evolve as the underlying resources do.

Applications interact with the Orchestrator through an *intent-based* interface: they declare the outcome they want, a guaranteed-bandwidth circuit between two endpoints, without specifying the underlying network configuration, and may first *negotiate*, asking for instance what maximum bandwidth a given path can support. To satisfy an intent, the Orchestrator computes the change each domain must make and applies it in two phases: a fast *propagate*, in which each RM inspects and may revise the request, followed by an asynchronous *commit*, in which each RM provisions its local resources. The result is an end-to-end virtual circuit with quality-of-service guarantees that coexists with ordinary best-effort traffic over the same physical infrastructure. Steering the right packets onto that circuit is handled at each site border, where the RM configures the router to route a designated range of globally routable IPv6 addresses onto

the circuit via BGP; only traffic to and from those addresses follows the guaranteed path, while everything else takes the default best-effort route [17]. End to end, this cycle of computation, propagation, and commitment completes in roughly one to two minutes [66].

On the data-management side, CMS uses Rucio [19], which expresses the data it must move as *replication rules*, each carrying a priority, and dispatches the underlying file transfers through the File Transfer Service (FTS) [15] to XRootD storage servers [45]. The two worlds meet in the *SENSE-Rucio* interoperation environment, whose components span five layers: Rucio (what to move), FTS and XRootD (the engine and servers that move it), SENSE (the network control plane), and, bridging Rucio and SENSE, the **Data Movement Manager (DMM)**. DMM is the interface that watches Rucio for transfers that should be accelerated, asks SENSE for an appropriately sized circuit, decides how the available bandwidth should be divided among competing transfers, redirects the actual data flow onto the circuit, and retires the circuit when the transfer completes.

6.4 The Data Movement Manager (DMM)

DMM is the orchestration layer that closes the gap between Rucio and SENSE [8]. Its task is conceptually simple, to keep a set of network circuits sized and routed in agreement with a continually changing set of prioritized transfers, but it must do so against two independent, distributed, and occasionally unreliable production systems. The design meets this challenge through a single organizing decision, described next, from which its resilience and observability follow.

DMM Architecture

DMM is built as a state machine: each transfer it manages, one per Rucio replication rule, advances through a fixed sequence of states, and that state is recorded in a database rather than held in memory. The work of moving transfers from one state to the next is divided among a handful of small, independent daemons, each responsible for a single step, such as provisioning

a circuit or beginning a teardown. The daemons never call one another; they coordinate only through the shared record, each picking up the transfers that have reached the state it handles and advancing them.

This design tries to mirror Rucio’s architecture. Because every transfer’s progress lives in the database and not in memory, DMM can be stopped and restarted at any time and each transfer resumes exactly where it left off; new behavior can be added by introducing a new state and a daemon to handle it; and because the whole system’s status can be read off at any instant, it is easy to monitor. Equally important, DMM never assumes that a request to SENSE has taken effect: it continually compares its own view of each transfer against the circuit state that SENSE reports and reconciles the two, so that circuits which fail, complete late, or are torn down without warning are detected and handled rather than silently lost. This reconciliation is what keeps DMM consistent with a network control plane it does not itself control.

The Transfer Lifecycle

The progression of a single transfer through DMM is summarized in Figure 6.7. A transfer enters when an ingestion daemon discovers a new Rucio rule. Rules are classified as SENSE-eligible by their activity label, and only eligible rules are tracked further; all others are recorded and ignored, so that non-SENSE production is never disturbed.

For an eligible rule, DMM first **allocates** the network endpoints it will use, claiming an unused, globally routable IPv6 subnet at each of the source and destination sites from SENSE’s address pools. It then **stages** a SENSE circuit: it submits an intent binding those two endpoints and reads back, from SENSE, both the bandwidth available on the link and the concrete circuit identifier. With the available bandwidth known for every competing transfer, a decision daemon **optimizes** the division of each link’s capacity among them (Section 6.4) and assigns each transfer a bandwidth. The circuit is then **provisioned** at that bandwidth, committing the SENSE configuration, after which DMM substitutes the circuit’s endpoints into the transfer so that the data flows over it (Section 6.4).

While a transfer runs, DMM tracks changes on both sides. If the Rucio rule's priority changes, the optimizer re-runs and the circuit is resized in place through SENSE's modify operation, without interrupting the transfer in flight, a property essential for long-running, petabyte-scale workflows. When the transfer finally completes, DMM does not immediately destroy the circuit. It first throttles it down to a minimal bandwidth to release capacity to others, and keeps it alive for a short, configurable window; if another transfer on the same site pair arrives within that window, it inherits the live circuit wholesale and skips straight to the provisioned state. This *circuit reuse* amortizes the comparatively slow and expensive provision/teardown cycle across back-to-back transfers. Only after the keep-alive window expires, and provided no other transfer still references it, is the circuit cancelled and its endpoints freed.

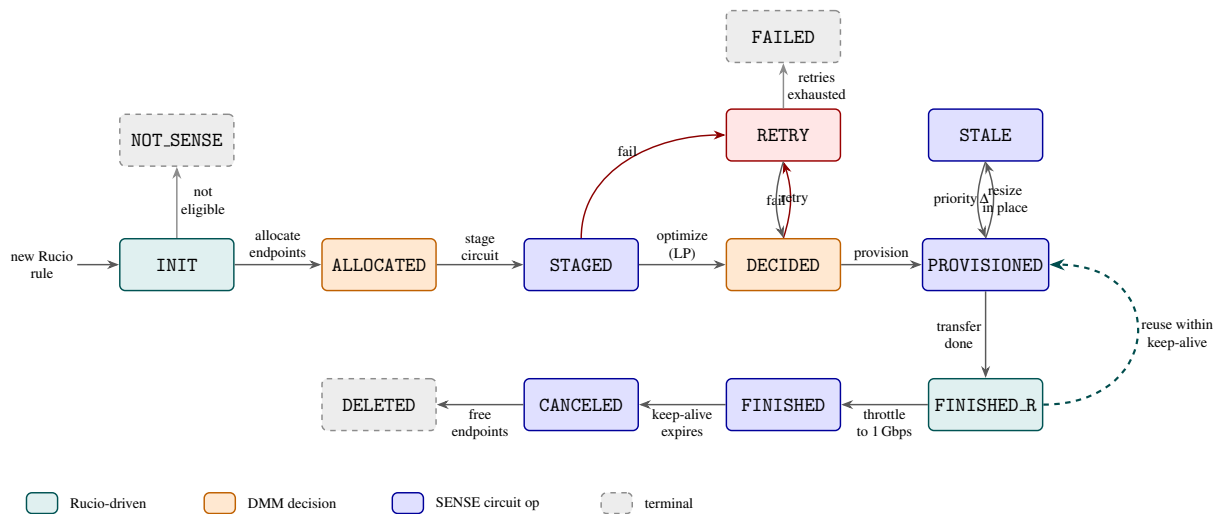


Figure 6.7. The DMM transfer lifecycle.

The same lifecycle absorbs the failure modes of a distributed system. Transient errors during staging or provisioning route a transfer to a retry state, from which it is re-attempted a bounded number of times before being marked permanently failed. A gateway timeout on a synchronous provisioning call is treated as "in flight, outcome unknown" and reconciled by the status poller rather than assumed to have failed. A circuit that SENSE has silently reaped is detected through the circuit-state machine and rebuilt from the already-allocated endpoints. And because a partial allocation could otherwise leak network addresses, every allocation failure

unwinds its own partial work, freeing both the database reservations and the SENSE-side address allocations, in the manner of a compensating transaction.

Priority-Based Bandwidth Allocation

At the analytical core of DMM is the decision of how to divide each link's capacity among the transfers competing for it. When DMM is the sole tenant of the SENSE-enabled paths and the binding constraint is the port capacity of each site, this is naturally posed as a linear program [8]. The set of active transfers is represented as a directed multigraph with the sites as nodes N and the replication rules as edges E ; several rules sharing a site pair appear as parallel edges, which are first merged into a single edge whose priority is the sum of theirs. Writing A for the $N \times E$ node-edge incidence matrix, b for the per-node port-capacity vector, and c for the merged edge-priority vector, DMM solves

$$\text{maximize } c^\top x \quad \text{subject to } Ax \leq b, \quad x \geq l \geq 0, \quad (6.2)$$

where x is the per-edge bandwidth allocation. The objective maximizes the priority-weighted throughput, while the lower bound l guarantees even the lowest-priority transfer a usable share. Rather than fix l arbitrarily, DMM binary-searches for the largest uniform lower bound that keeps the program feasible, which drives the solution toward *max-min fairness*: it raises the minimum guaranteed allocation as high as possible before maximizing weighted throughput. The solved bandwidth on each merged edge is then split among its constituent rules in proportion to their priorities and rounded down to a coarse granularity for stability.

As a concrete example, consider the four-site topology of Figure 6.8, spanning T2_US_UCSD, T2_US_Caltech, T1_US_FNL, and T2_US_Nebraska, with two rules of priority 5 and 3 running between UCSD and Caltech. Merging those parallel rules into a single edge of combined priority 8, the linear-programming system is

$$A = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad b = \begin{bmatrix} 400 \\ 400 \\ 200 \\ 100 \end{bmatrix}, \quad c = \begin{bmatrix} 8 \\ 3 \\ 4 \\ 2 \end{bmatrix}, \quad (6.3)$$

whose optimal solution is the merged-edge allocation $x = [335, 65, 65, 70]$ Gbps. The 335 units on the UCSD-Caltech edge are then redistributed between its two rules in proportion to their priorities, 210 to the priority-5 rule and 125 to the priority-3 rule, yielding the final per-rule allocations. The optimization is re-run whenever the set of transfers or their priorities changes, and where no feasible solution exists, the lower bounds are relaxed to recover a sub-optimal but valid allocation.

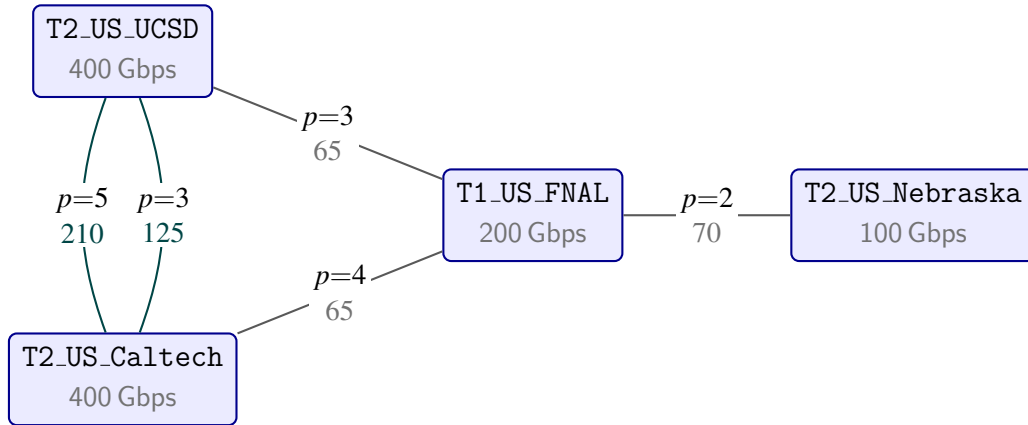


Figure 6.8. Example transfer topology for the priority-based bandwidth allocation: nodes are sites (with port capacity), edges are rules (with priority p and assigned bandwidth).

Redirecting the Data Flow

Deciding and provisioning a circuit is only half the problem; the actual bytes must be made to travel over it. DMM does not move data itself, FTS and XRootD do, so it must steer their transfers onto the provisioned endpoints. This is achieved through a small patch to Rucio's FTS transfer tool. When Rucio builds a transfer job for a SENSE-eligible rule, the patch queries DMM for that rule and receives, in return, the hostnames of the provisioned IPv6 endpoints at

the source and destination. It then rewrites the source and destination of the transfer URLs to those endpoints, so that FTS drives the data over the SENSE circuit rather than the default best-effort path; this application-layer steering is what makes the transfer use the SENSE-allocated IPv6 addresses, which the site-border BGP configuration of Section 6.3 in turn routes onto the guaranteed circuit. The query is retried briefly if DMM has not yet finished provisioning, smoothing the race between Rucio submitting a transfer and DMM building its circuit, and rules that are not SENSE-eligible are passed through untouched. This closes the control loop: Rucio decides what to move, DMM provisions a circuit and reports its endpoints, FTS moves the data over that circuit, and DMM monitors the flow and retires the circuit when it is done.

Two further couplings keep the circuit fully utilized. Because the throughput attainable on a link grows with the number of concurrent transfers in flight (Section 6.2), DMM adjusts the FTS concurrency limit for each circuit in proportion to its allocated bandwidth, dividing each site pair’s stream budget among its transfers by priority. And to confirm that a circuit is actually carrying the traffic it was sized for, a monitoring daemon polls per-interface byte counters from the sites’ telemetry, computes the achieved throughput, and flags a circuit as healthy only when it sustains at least 80% of its allocated bandwidth, exposing the result for operators and automated diagnostics.

6.5 Evaluation on a Multi-Domain Testbed

Quantifying the benefits of this paradigm requires testing under conditions that reflect real production networks [67]. To this end the Multi-Infrastructure SENSE Testbed (MIST) was constructed, federating compute, storage, and network resources across three distinct global research infrastructures, FABRIC [18], the National Research Platform [80], and the GNA-G AutoGOLE testbed [53], interconnected through distinct autonomous systems into a single virtual overlay. This heterogeneous, multi-domain environment forces SENSE to coordinate different control planes, performance limits, and trust boundaries that cannot be reproduced in a

single-domain laboratory. The simplified topology of the MIST testbed is shown in Figure 6.9.

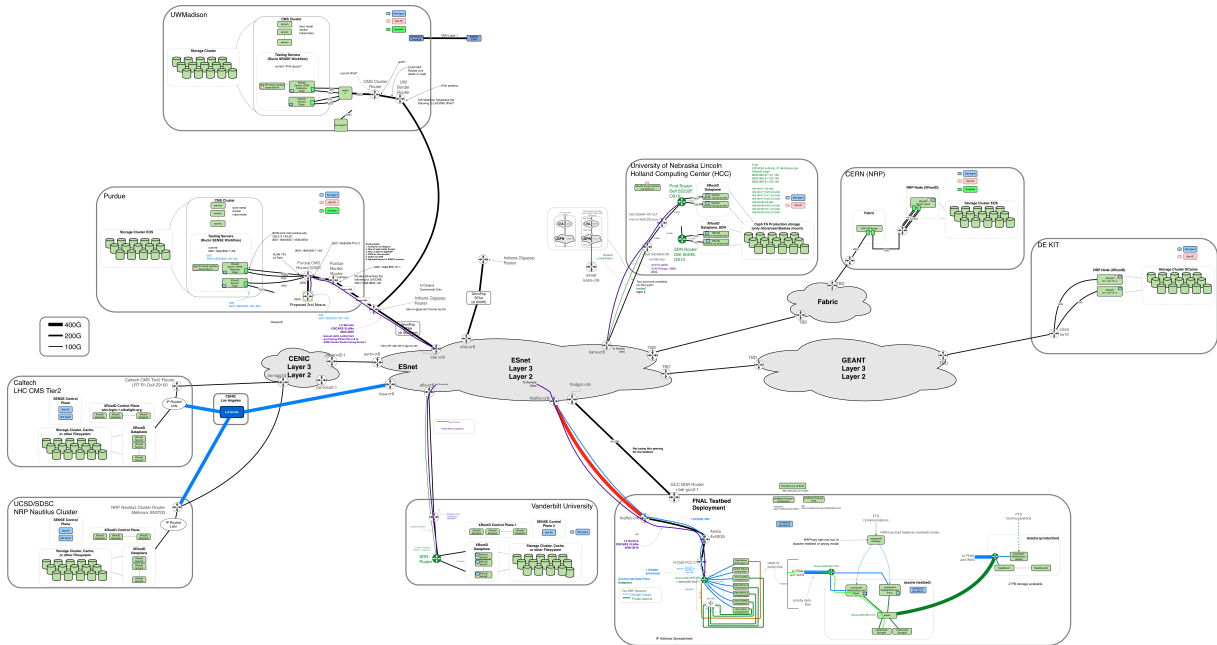


Figure 6.9. The MIST testbed

End-to-End Application Performance

The end-to-end benefit was measured with a set of four experiments that transfer three 12 TB datasets between high-performance sites (SDSC and Caltech) over a shared 100 Gbps path, each dataset mapped to a SENSE flow, or *Rule*, of a distinct priority in the ratio 2:5:3 on its own VLAN [67]. The three Rules and their transfers are started sequentially, one every ten minutes, so that the link passes through phases of one, two, and three simultaneously active flows. The four experiments cross the transfer tool, the Fast Data Transfer (FDT) tool used to baseline the raw network against the full Rucio/FTS/DMM application stack, with the network policy, best-effort against SENSE-guaranteed quality of service (QoS); the resulting traffic is shown in Figure 6.10.

The first two experiments use FDT to characterize the network itself, driving the transfers directly while the SENSE Rules are configured by hand, first under best-effort and then under guaranteed QoS (Figure 6.10a,b). Three features stand out. A production host left uncontended

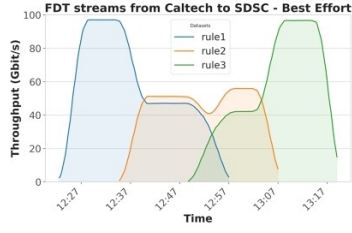
occupies almost the entire 100 Gbps link; when all three flows are active under best-effort they share the link equally and fairly, as on any traditional network; and under QoS the peak rates are lower, because SENSE caps each flow at a guaranteed bandwidth set by the priorities, so that each flow receives a share proportional to its priority, 2:5:3 when all three are active and the corresponding proportions when only two remain. That the measured shares track the configured priorities so closely confirms that the infrastructure responds accurately to SENSE settings, and establishes the baseline against which the full-stack experiments are read.

The third experiment exercises the complete data-management stack, Rucio, FTS, and DMM, moving data over the SENSE paths but without requesting any QoS priority from the network (Figure 6.10c,d). The three flows share the 100 Gbps link equally and fairly until the first completes, after which the remaining two split the freed capacity, and finally the last flow occupies the entire link on its own. This reproduces the best-effort behavior of the first experiment, now through the full end-to-end application, confirming that inserting Rucio, FTS, and DMM into the path introduces no penalty of its own.

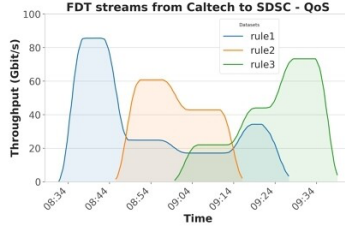
The fourth experiment is the one that matters for production: the same stack now requests priorities and guaranteed bandwidth caps through SENSE, and DMM uses the SENSE *modify* operation to adjust the guarantees of the existing flows as new ones start and old ones finish (Figure 6.10e,f). The highest-priority flow (Rule 2) is granted the largest share and finishes first; once it completes and only the lower-priority Rule 3 and Rule 1 remain, Rule 3 receives only slightly more bandwidth than Rule 1, as their nearer priorities dictate. Crucially, every one of these SENSE reconfigurations is applied without interrupting or stopping the transfers already in flight, a property that is essential for the long-running, petabyte-scale workflows of the HL-LHC.

Taken together, the four experiments provide direct empirical evidence that high-level application intent can be translated into enforceable, non-disruptive network policy even at the demanding scale of LHC data movement: the best-effort cases are indistinguishable from a traditional shared link, while under SENSE-guaranteed service bandwidth is allocated in proportion to priority and rebalanced on the fly without disturbing the transfers already under

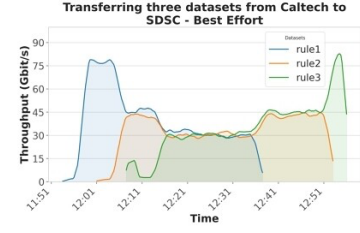
way.



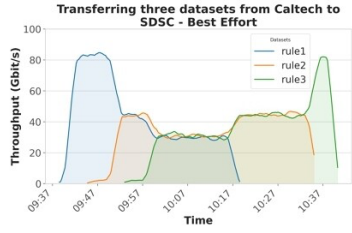
(a) Experiment 1: FDT with No QoS defined in the network



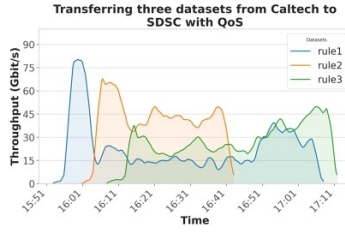
(b) Experiment 2: FDT with QoS and guaranteed bandwidth caps provisioned by SENSE



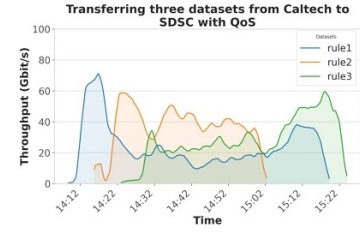
(c) Experiment 3: Full stack with no QoS defined in the network – Run 1



(d) Experiment 3: Full stack with no QoS defined in the network – Run 2



(e) Experiment 4: Full stack with QoS defined in the network – Run 1



(f) Experiment 4: Full stack with QoS defined in the network – Run 2

Figure 6.10. End-to-end performance on the MIST testbed: three prioritized transfers (Rules 1-3, 2:5:3) sharing a 100 Gbps path [67].

6.6 Summary

This chapter has developed the case for treating the wide-area network as a managed, schedulable resource for HL-LHC data movement. Dedicated benchmarking established that a containerized XRootD deployment can saturate a 400 Gbps link, confirming that the data plane can meet next-generation demands but is unstable when driven blindly [10, 9]. The Data Movement Manager [8] then closed the loop between grid data management and SDN orchestration: built as a database-backed state machine of independent daemons, it translates transfer priorities into guaranteed network service through a linear-programming optimizer with a max-min-fair floor, redirects the real data flow onto provisioned circuits, and reconciles its state against an unreliable control plane while reusing circuits to amortize their cost. Evaluation on the multi-domain MIST testbed [67] showed that application-driven prioritization delivers proportionally faster completion of high-priority transfers, with bandwidth allocated in proportion

to priority and rebalanced on the fly without disrupting ongoing flows. Together these results demonstrate a practical path to transforming the network from a passive, best-effort medium into a deterministic, schedulable resource capable of sustaining the exascale scientific workflows of the next decade.

Acknowledgements

Chapter 6, in part, is a reprint of the material as it appears in "400 Gbps benchmark of XRootD HTTP-TPC," EPJ Web of Conferences, 2024 [10]; "Data Movement Manager (DMM) for the SENSE-Rucio Interoperation Prototype," EPJ Web of Conferences, 2025 [8]; "Benchmarking XRootD-HTTPS on 400 Gbps Links with Variable Latencies," EPJ Web of Conferences, 2025 [9]; and "Using Microsoft Azure for XRootD Network Benchmarking," 2021 [11]. The chapter also contains material from "Integrating End-to-End Exascale SDN into the LHC Data Distribution Cyberinfrastructure," PEARC, 2022 [54]; "Managed Network Services for Exascale Data Movement Across Large Global Scientific Collaborations," XLOOP, 2022 [86]; and "SENSE in Practice: Quantifying the End-to-End Benefits of Intent-Based Bandwidth Reservation for Exascale Science Workflows," SC Workshops, 2025 [67].

Appendix A

Supplementary Distributions

Chapter 4 shows, for each channel, the distributions that most directly motivate the selection and the discriminants. This appendix collects the remaining input distributions for all three channels in both data-taking eras. In every figure the simulated backgrounds are stacked, the signal is overlaid and scaled by the factor given in the legend, and the lower panel shows the data-to-prediction ratio; in the all-hadronic channel the QCD simulation has had its tagger scores resampled as described in Section 4.5.5. All distributions are shown after the full preselection of the corresponding channel and on a logarithmic vertical scale.

A.1 All-Hadronic Fully-Boosted Channel

Run 2

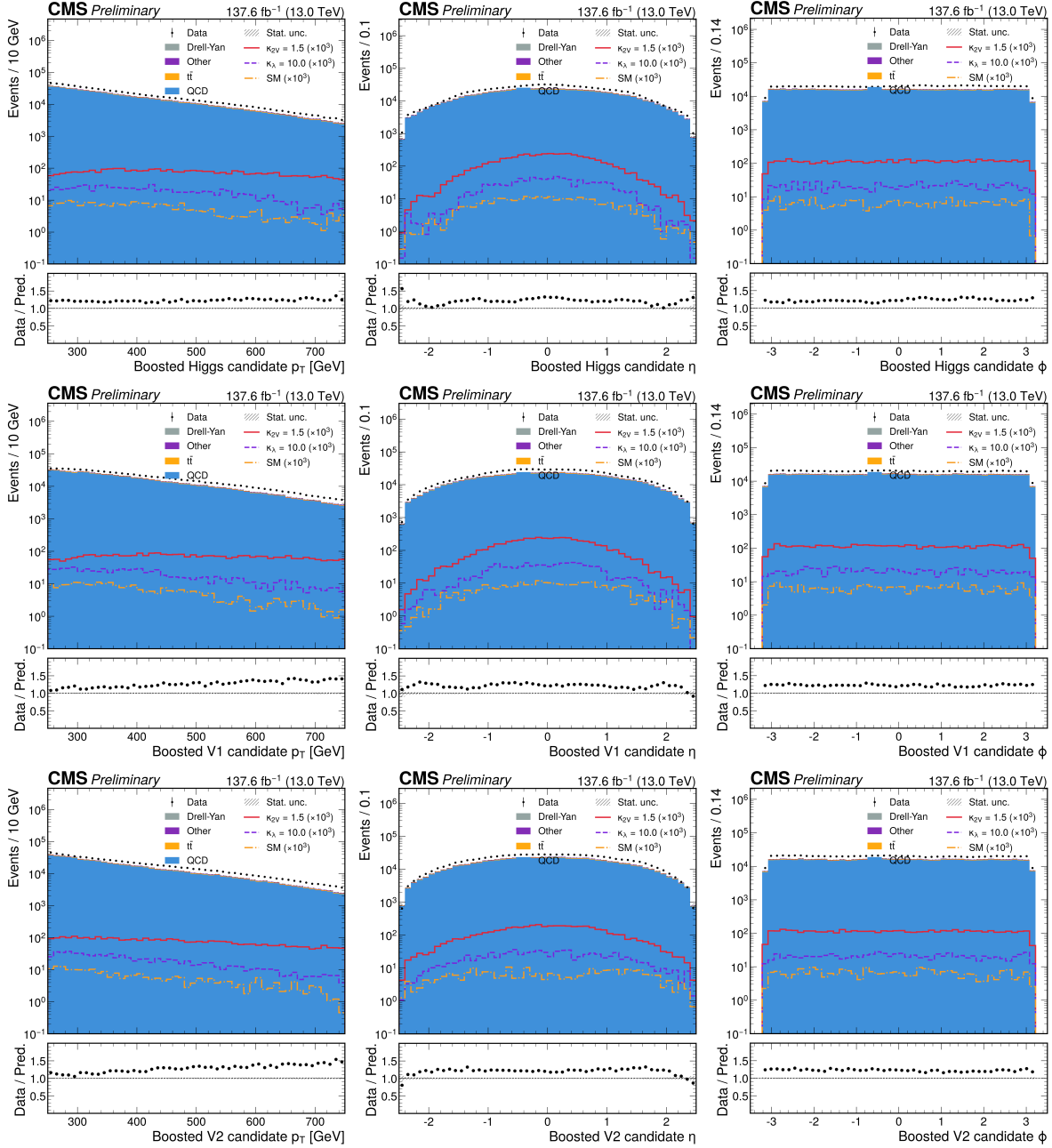


Figure A.1. Transverse momentum (left), pseudorapidity (middle), and azimuth (right) of the boosted candidates, all-hadronic fully-boosted channel, Run 2.

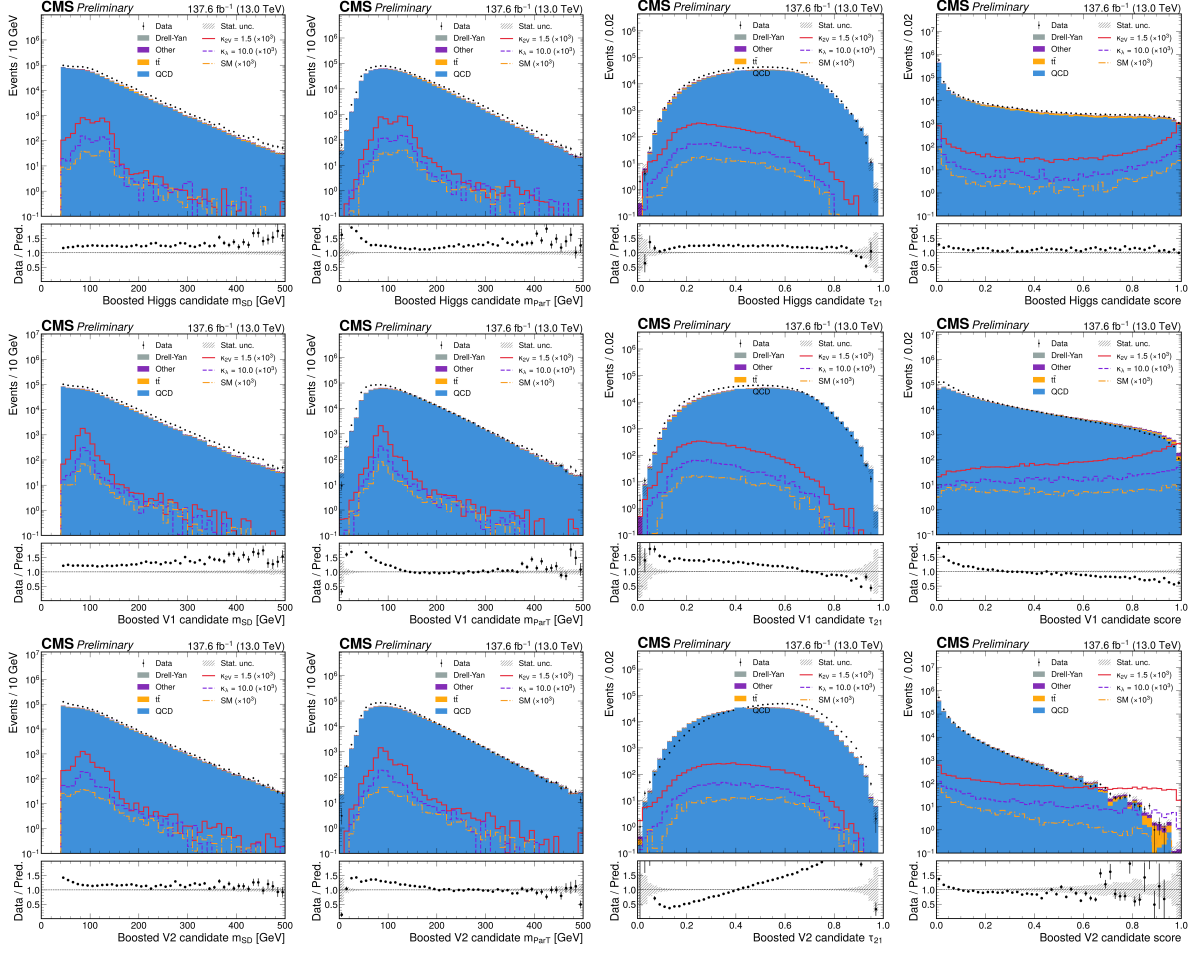


Figure A.2. Groomed mass, regressed mass, N -subjettiness ratio τ_{21} , and GLOPART tagger score (left to right) of the boosted candidates, all-hadronic fully-boosted channel, Run 2.

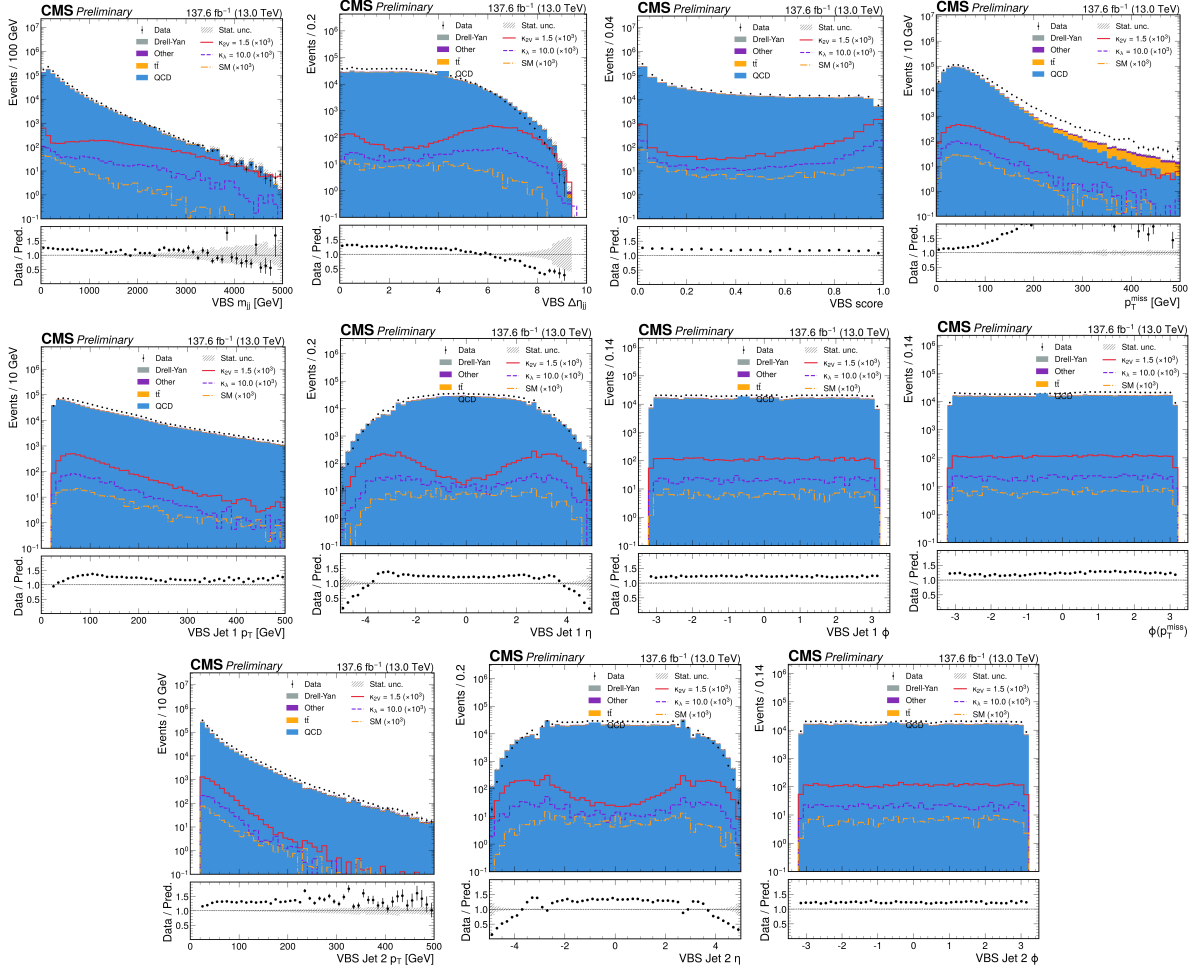


Figure A.3. Tagging-jet system and missing transverse momentum, all-hadronic fully-boosted channel, Run 2: pair mass, rapidity gap and BDT score with p_T^{miss} (top), and the kinematics of the leading and subleading tagging jet (lower rows).

Run 3

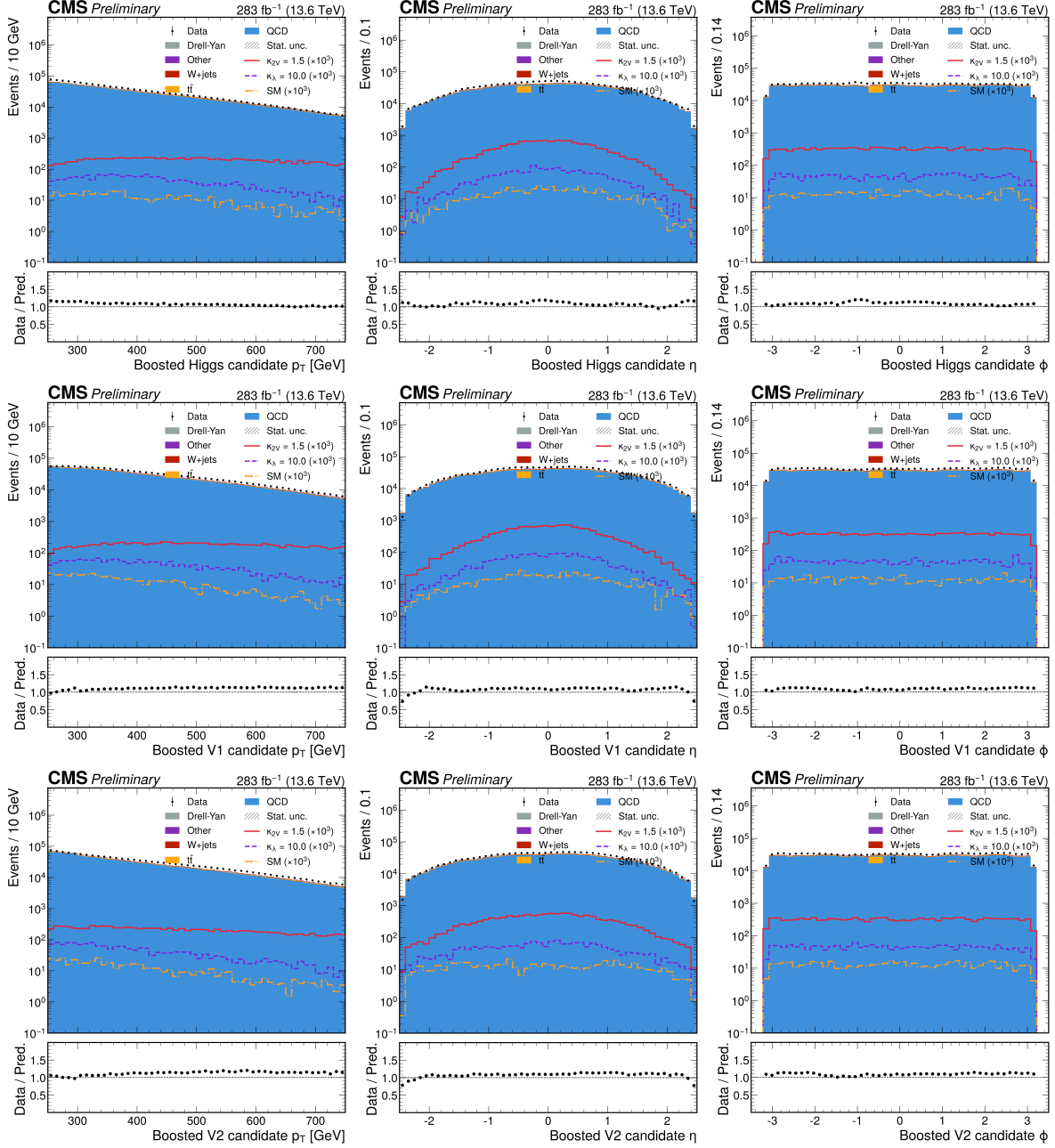


Figure A.4. Transverse momentum (left), pseudorapidity (middle), and azimuth (right) of the boosted candidates, all-hadronic fully-boosted channel, Run 3.

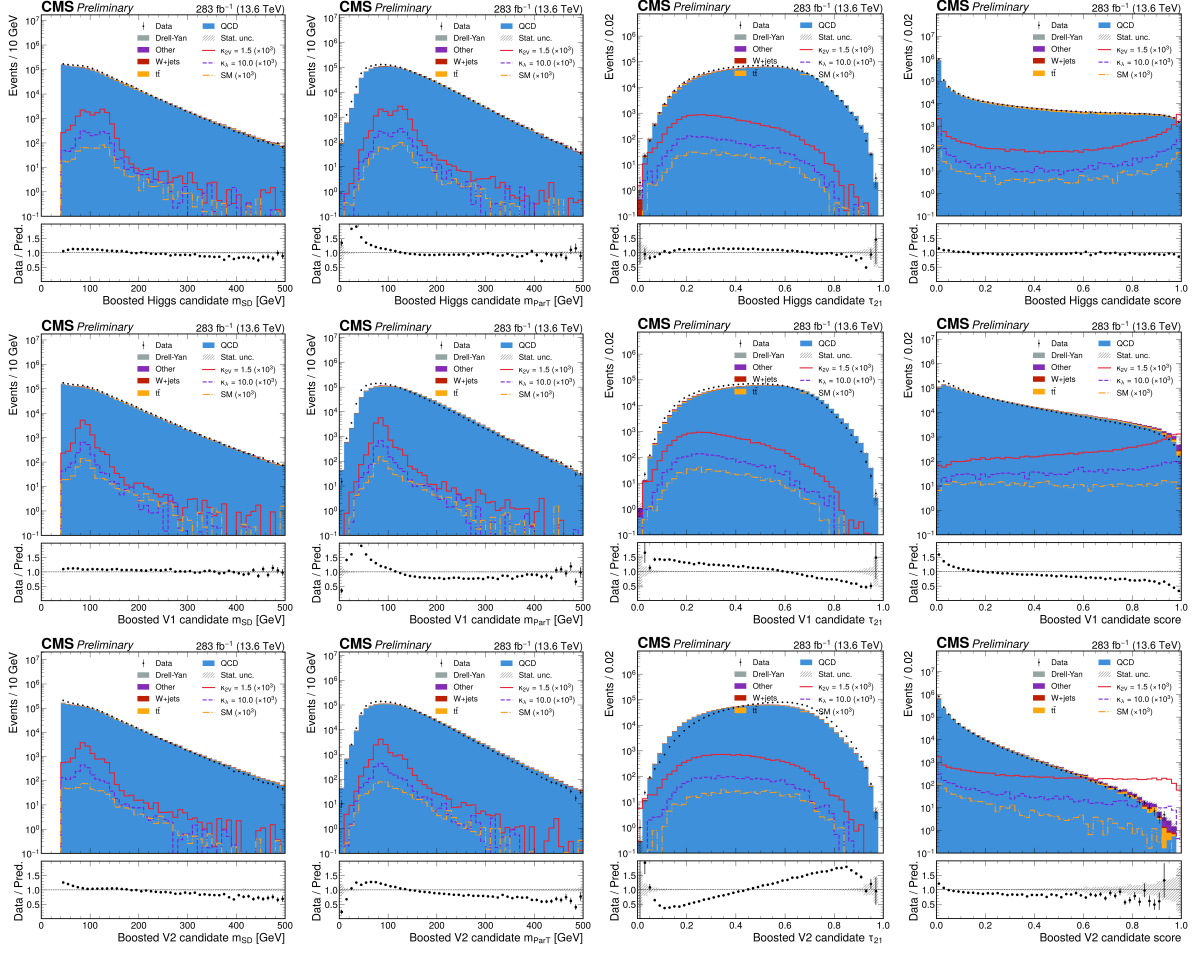


Figure A.5. Groomed mass, regressed mass, N -subjettiness ratio τ_{21} , and GLOPART tagger score (left to right) of the boosted candidates, all-hadronic fully-boosted channel, Run 3.

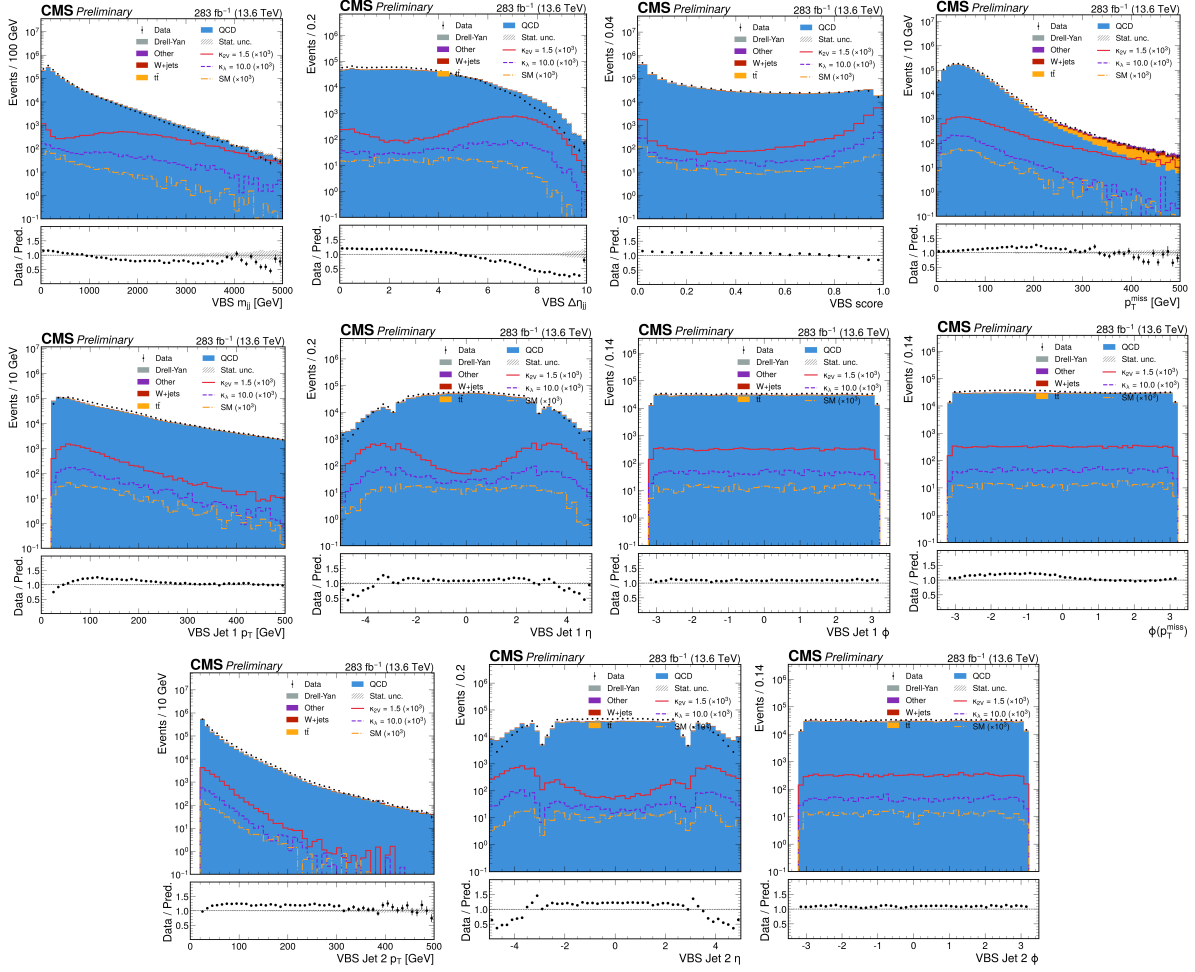


Figure A.6. Tagging-jet system and missing transverse momentum, all-hadronic fully-boosted channel, Run 3: pair mass, rapidity gap and BDT score with p_T^{miss} (top), and the kinematics of the leading and subleading tagging jet (lower rows).

A.2 Single-Lepton Fully-Boosted Channel

Run 2

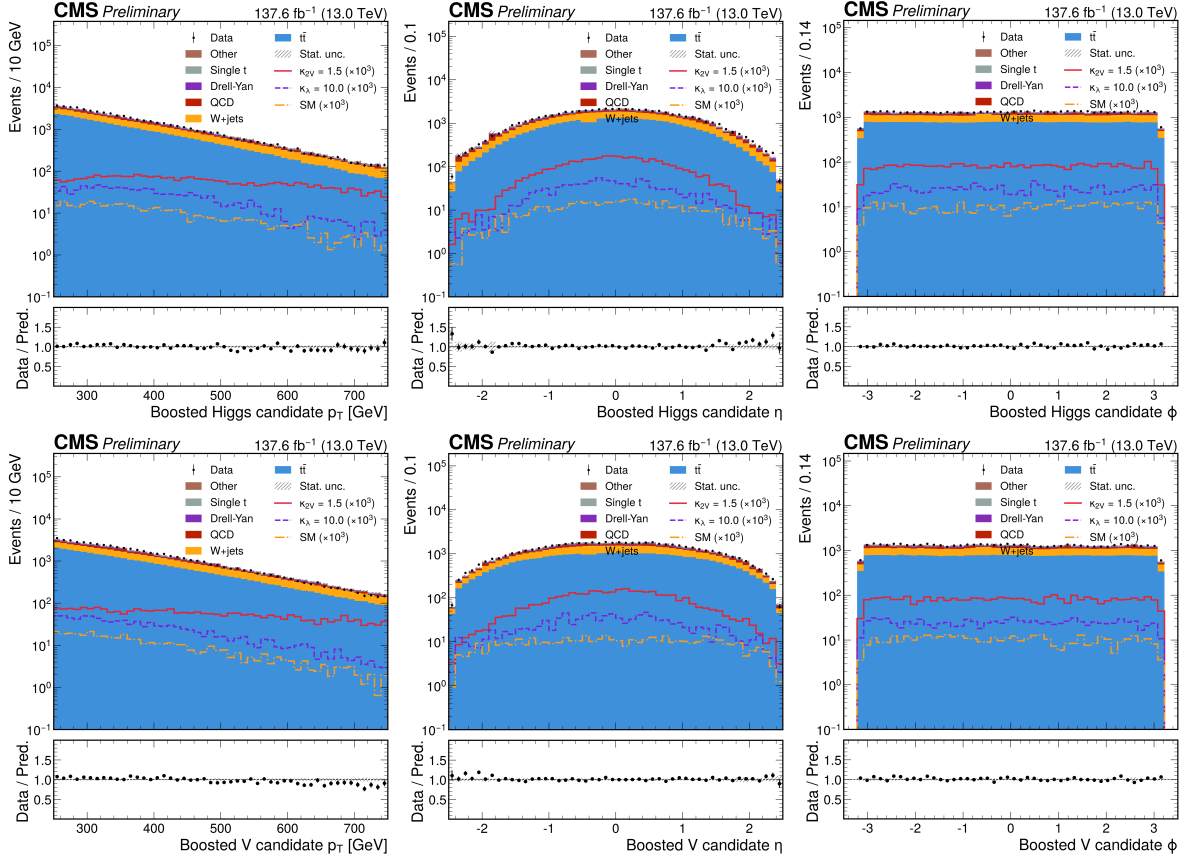


Figure A.7. Transverse momentum (left), pseudorapidity (middle), and azimuth (right) of the boosted candidates, single-lepton fully-boosted channel, Run 2.

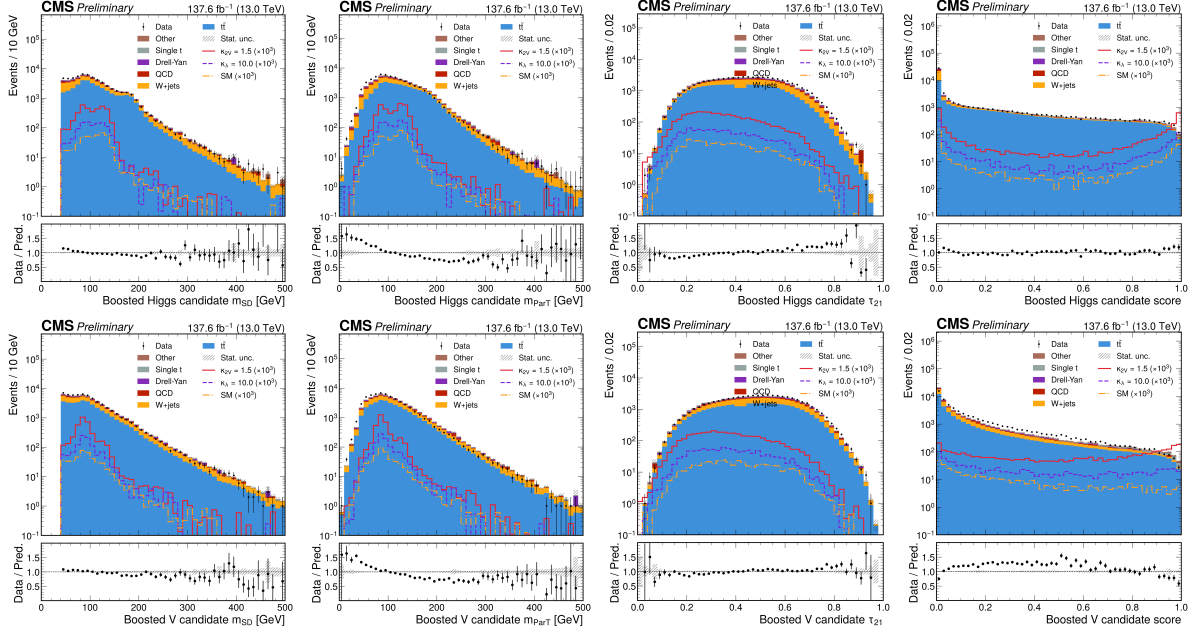


Figure A.8. Groomed mass, regressed mass, N -subjettiness ratio τ_{21} , and GLOPART tagger score (left to right) of the boosted candidates, single-lepton fully-boosted channel, Run 2.

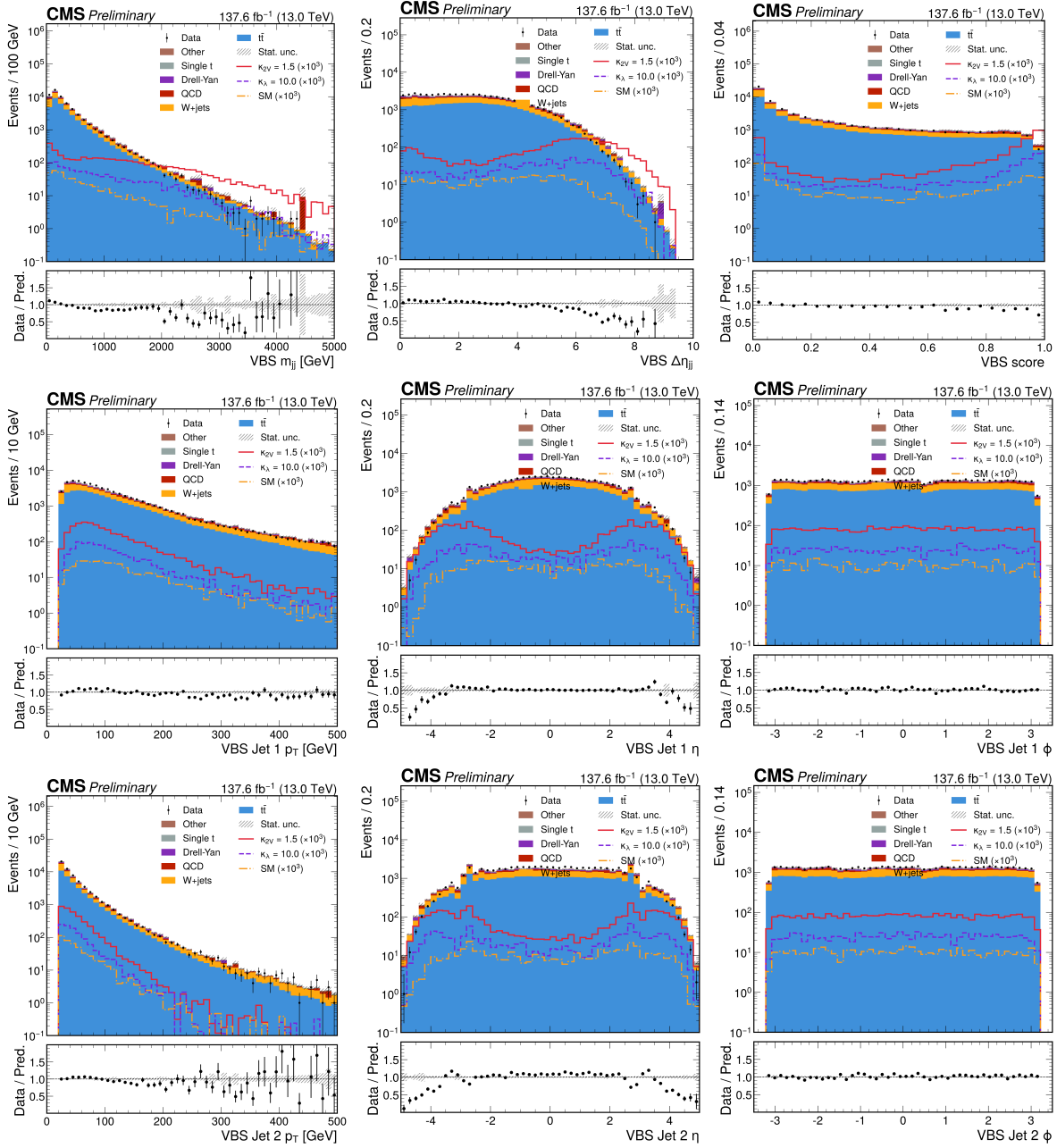


Figure A.9. Tagging-jet system, single-lepton fully-boosted channel, Run 2: pair mass, rapidity gap and BDT score (top), then the kinematics of the leading (middle) and subleading (bottom) tagging jet.

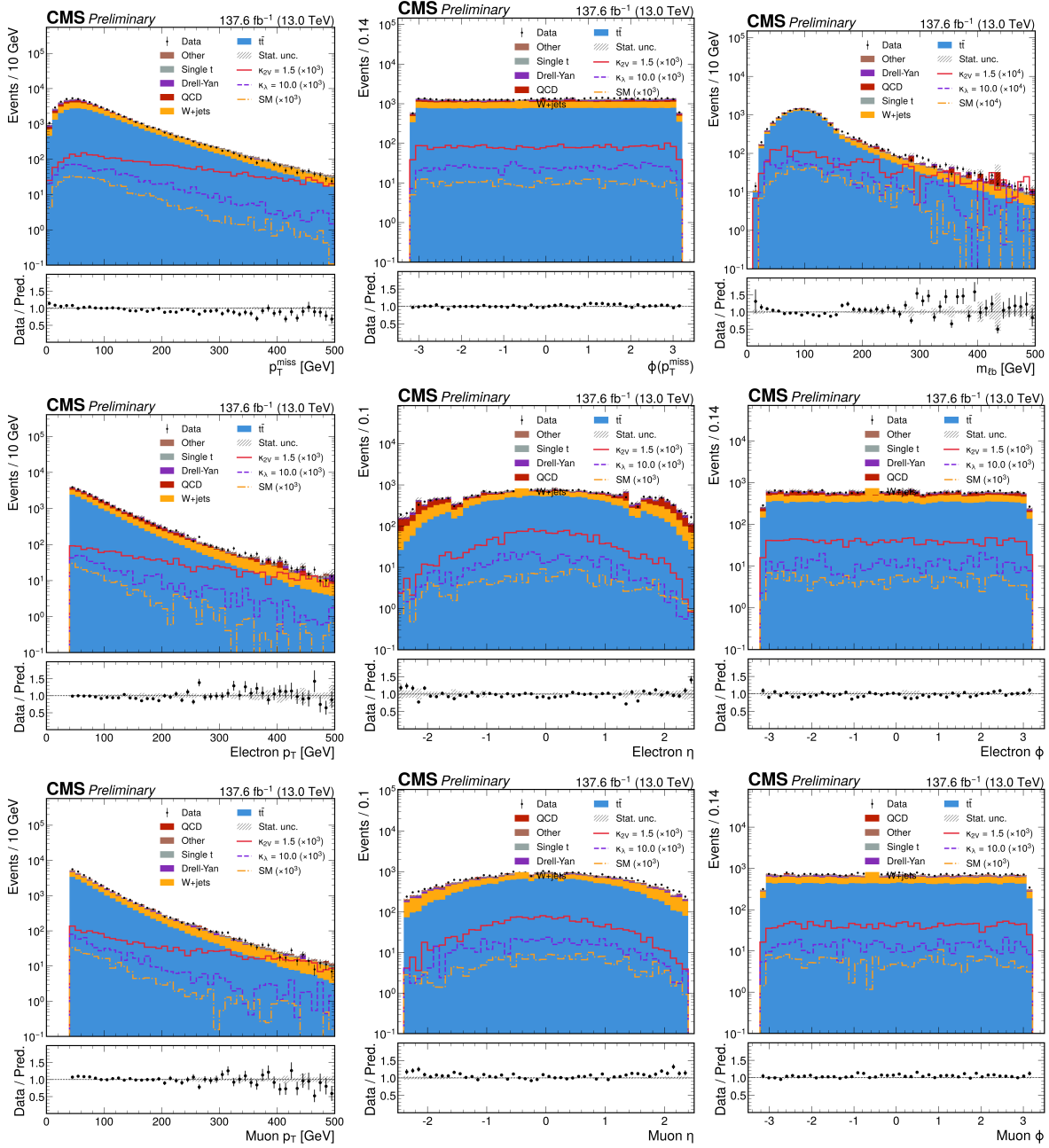


Figure A.10. Missing transverse momentum, m_{lb} , and lepton kinematics, single-lepton fully-boasted channel, Run 2.

Run 3

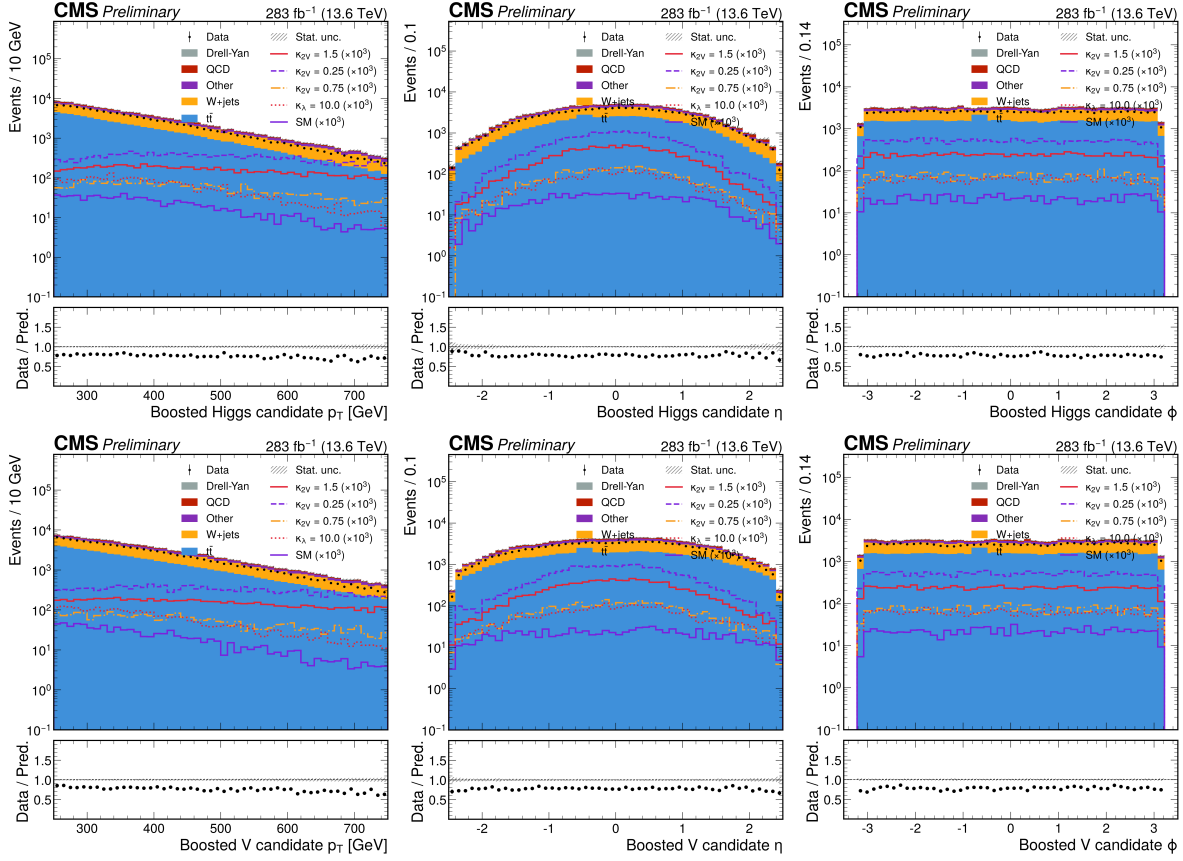


Figure A.11. Transverse momentum (left), pseudorapidity (middle), and azimuth (right) of the boosted candidates, single-lepton fully-boosted channel, Run 3.

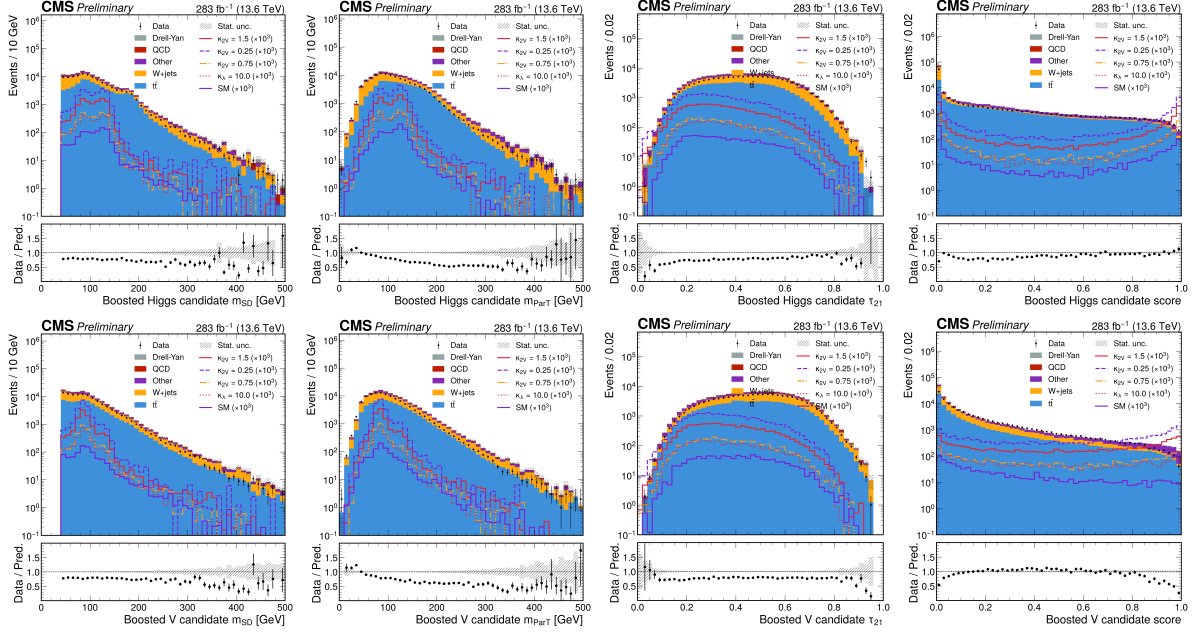


Figure A.12. Groomed mass, regressed mass, N -subjettiness ratio τ_{21} , and GLOPART tagger score (left to right) of the boosted candidates, single-lepton fully-boosted channel, Run 3.

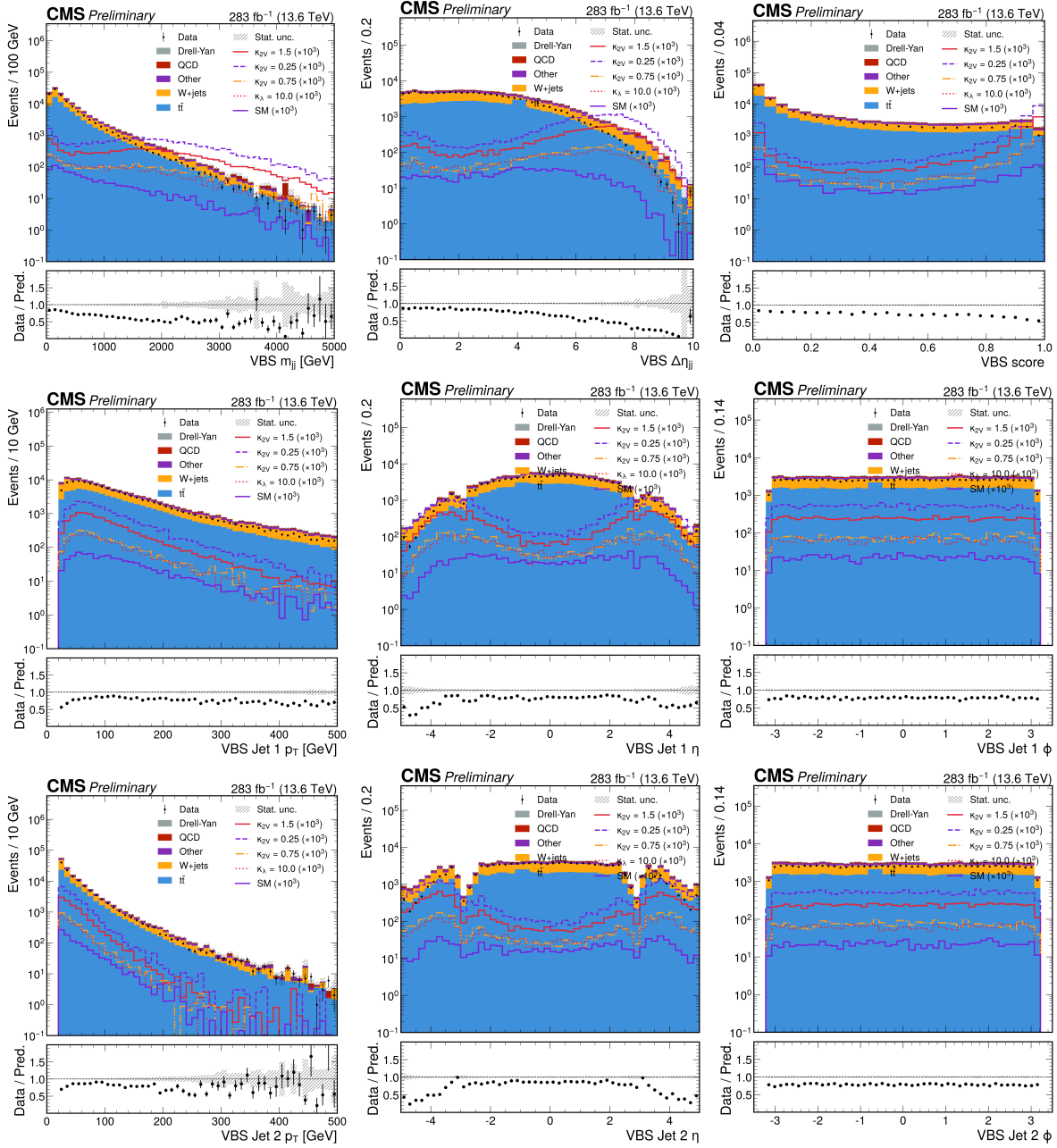


Figure A.13. Tagging-jet system, single-lepton fully-boosted channel, Run 3: pair mass, rapidity gap and BDT score (top), then the kinematics of the leading (middle) and subleading (bottom) tagging jet.

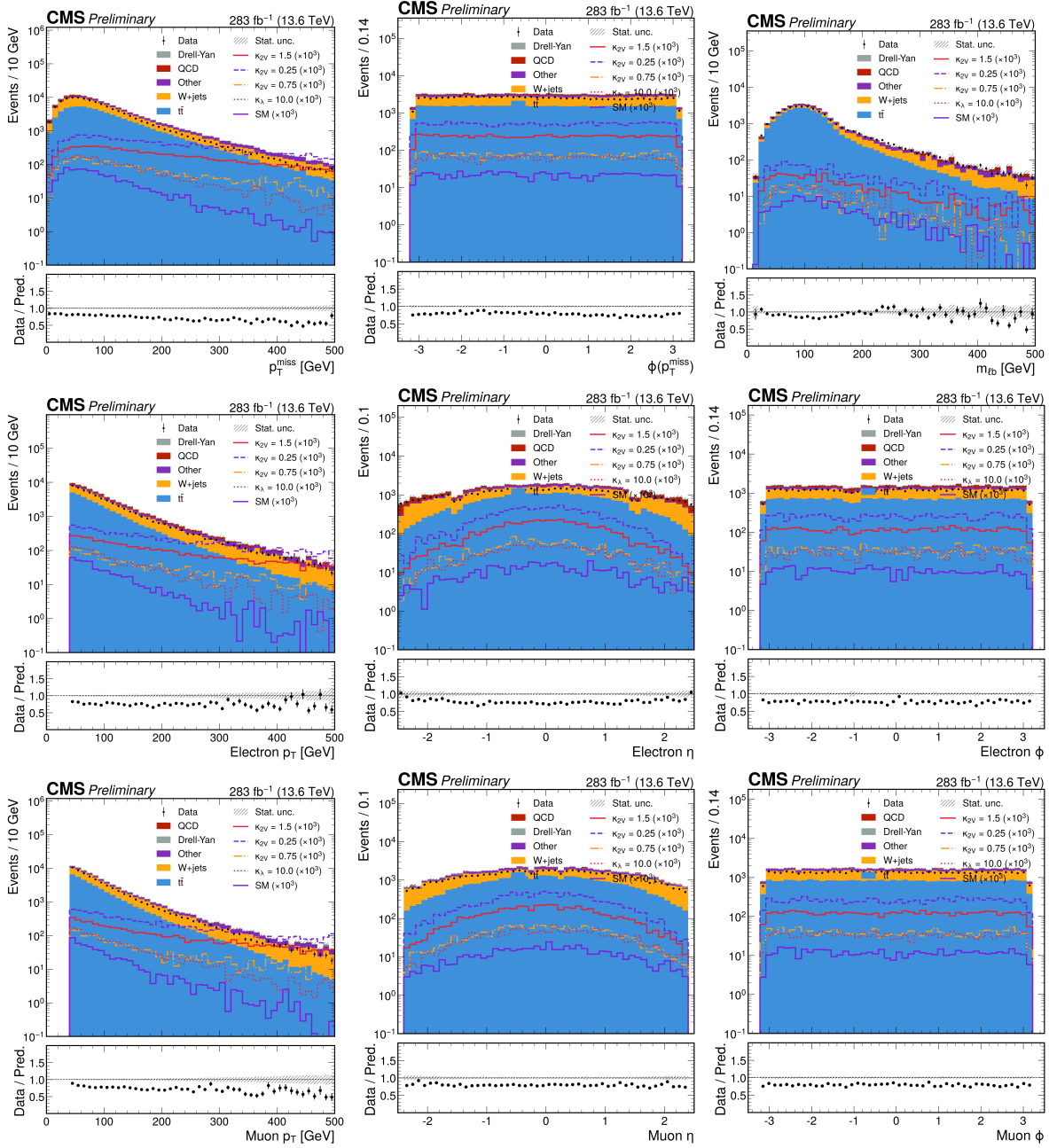


Figure A.14. Missing transverse momentum, m_{lb} , and lepton kinematics, single-lepton fully-boasted channel, Run 3.

A.3 Single-Lepton Semi-Boosted Channel

Run 2

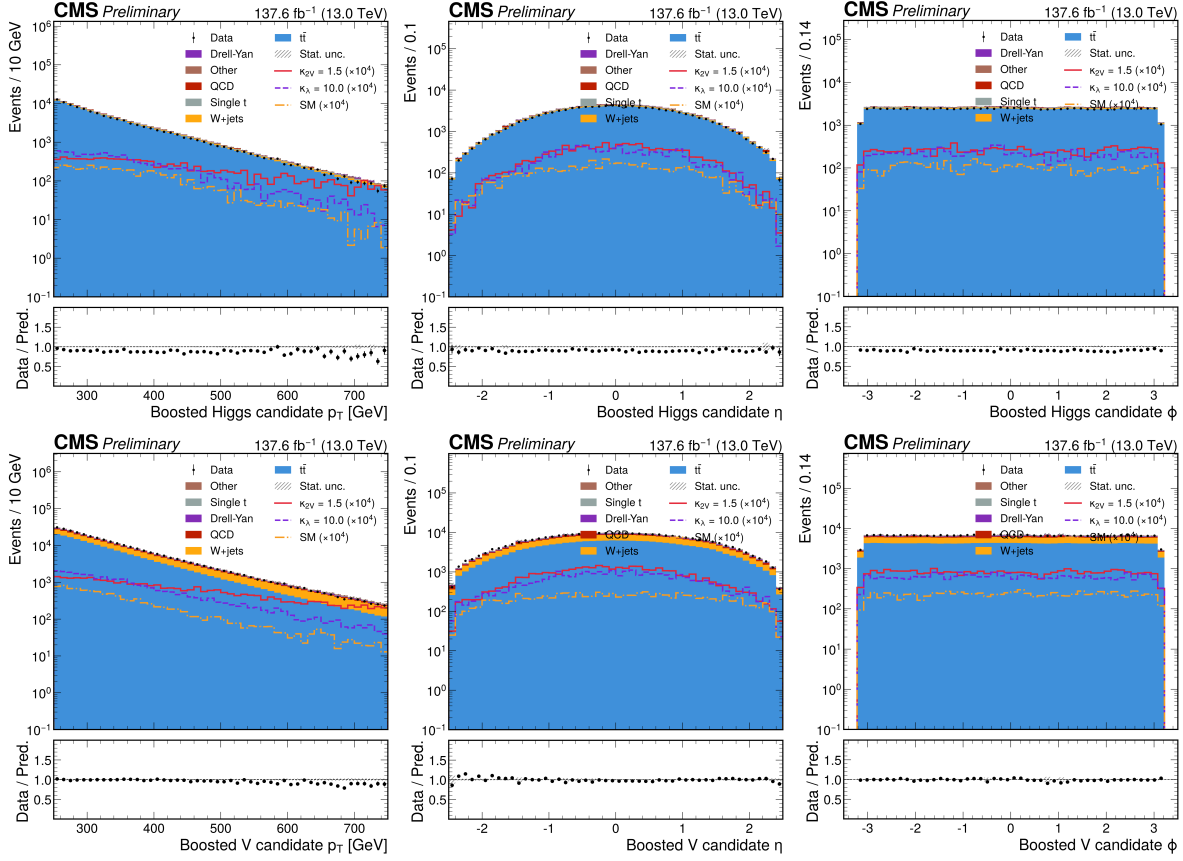


Figure A.15. Transverse momentum (left), pseudorapidity (middle), and azimuth (right) of the boosted candidates, single-lepton semi-boosted channel, Run 2.

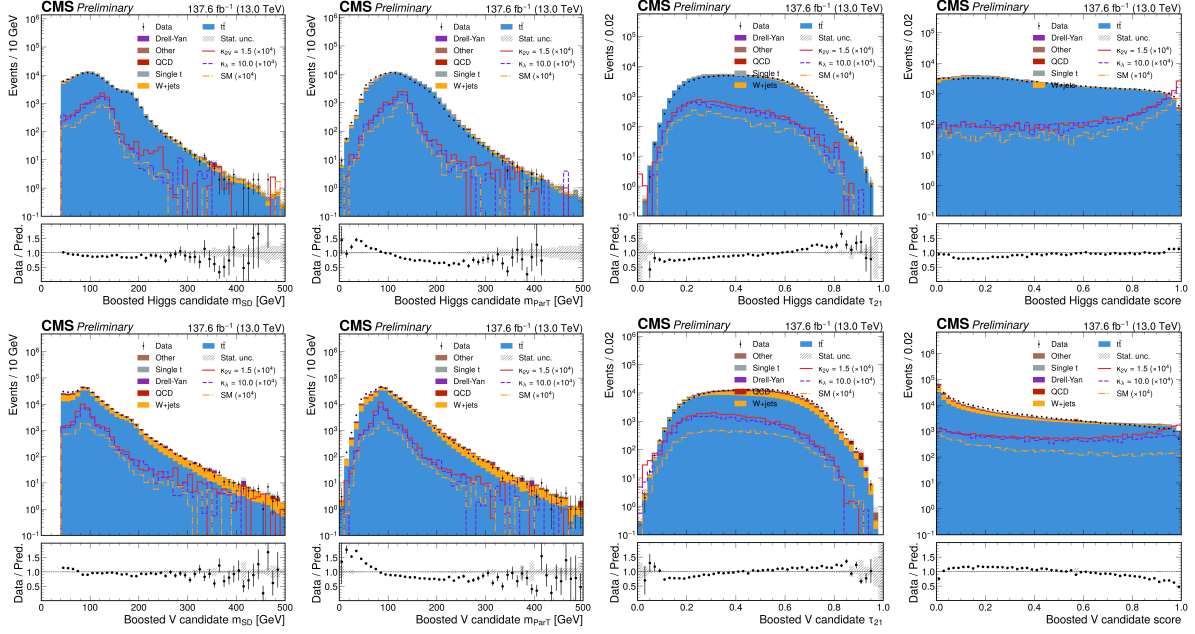


Figure A.16. Groomed mass, regressed mass, N -subjettiness ratio τ_{21} , and GLOPART tagger score (left to right) of the boosted candidates, single-lepton semi-boosted channel, Run 2.

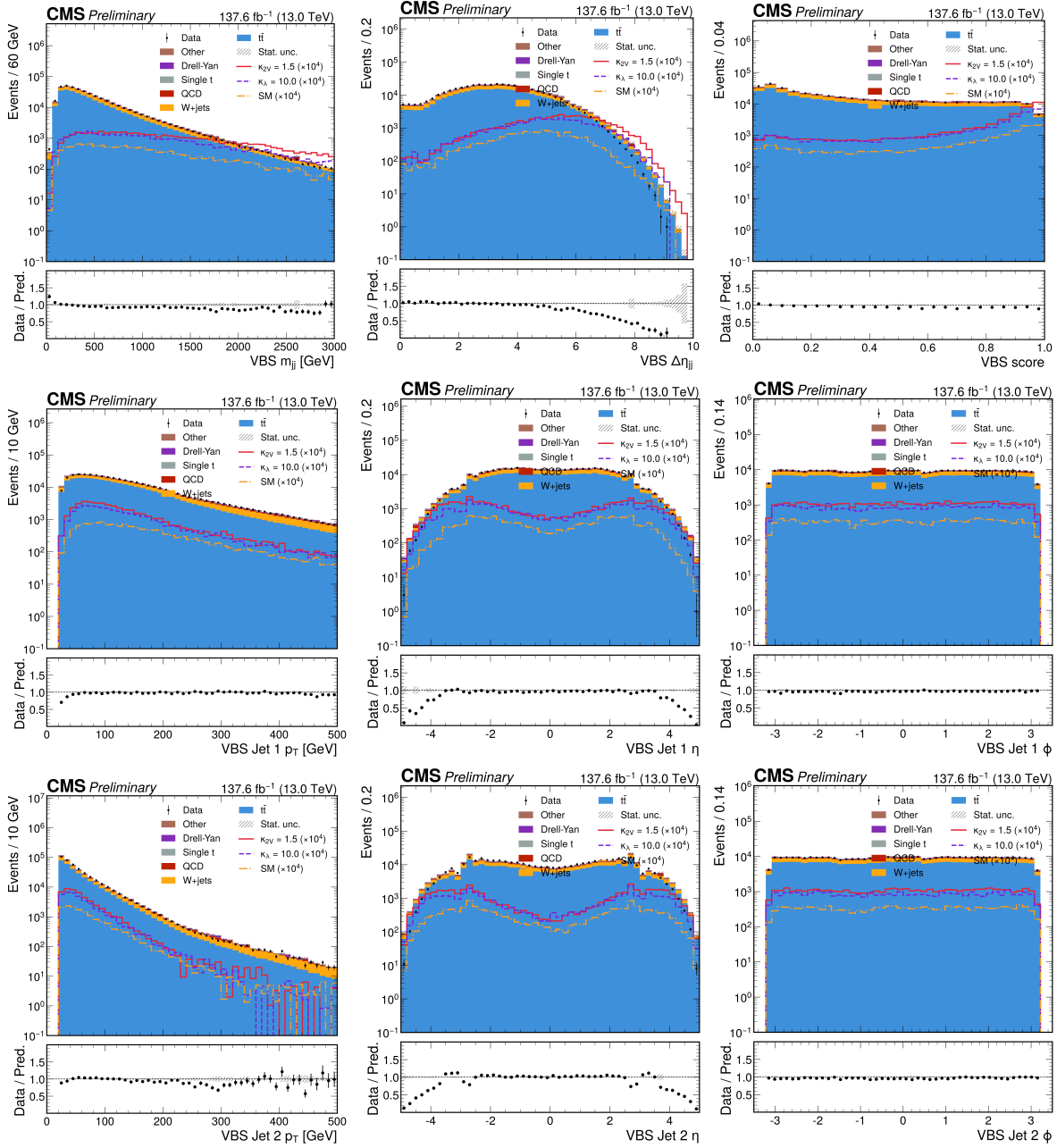


Figure A.17. Tagging-jet system, single-lepton semi-boosted channel, Run 2: pair mass, rapidity gap and BDT score (top), then the kinematics of the leading (middle) and subleading (bottom) tagging jet.

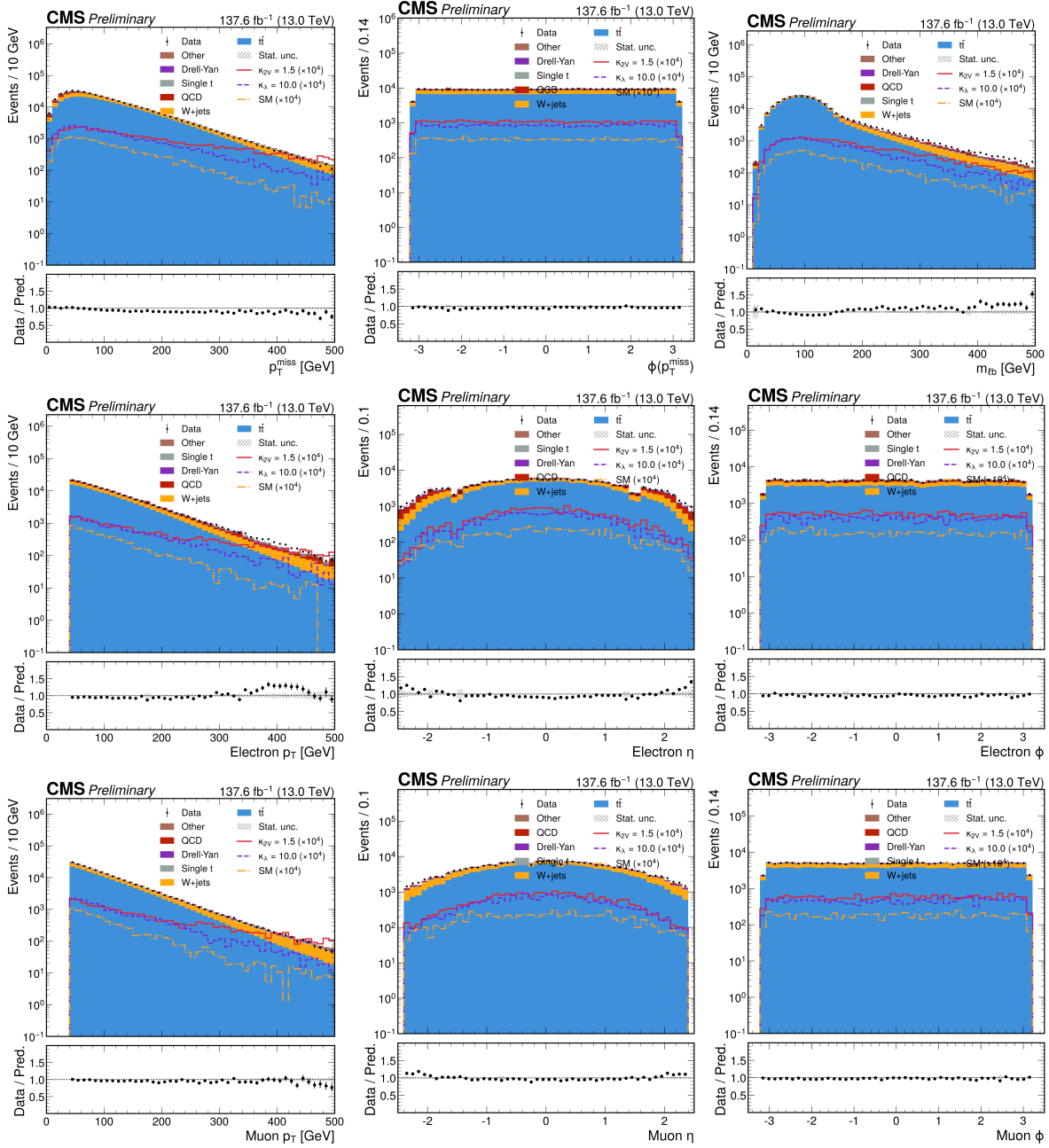


Figure A.18. Missing transverse momentum, m_{lb} , and lepton kinematics, single-lepton semi-boostered channel, Run 2.

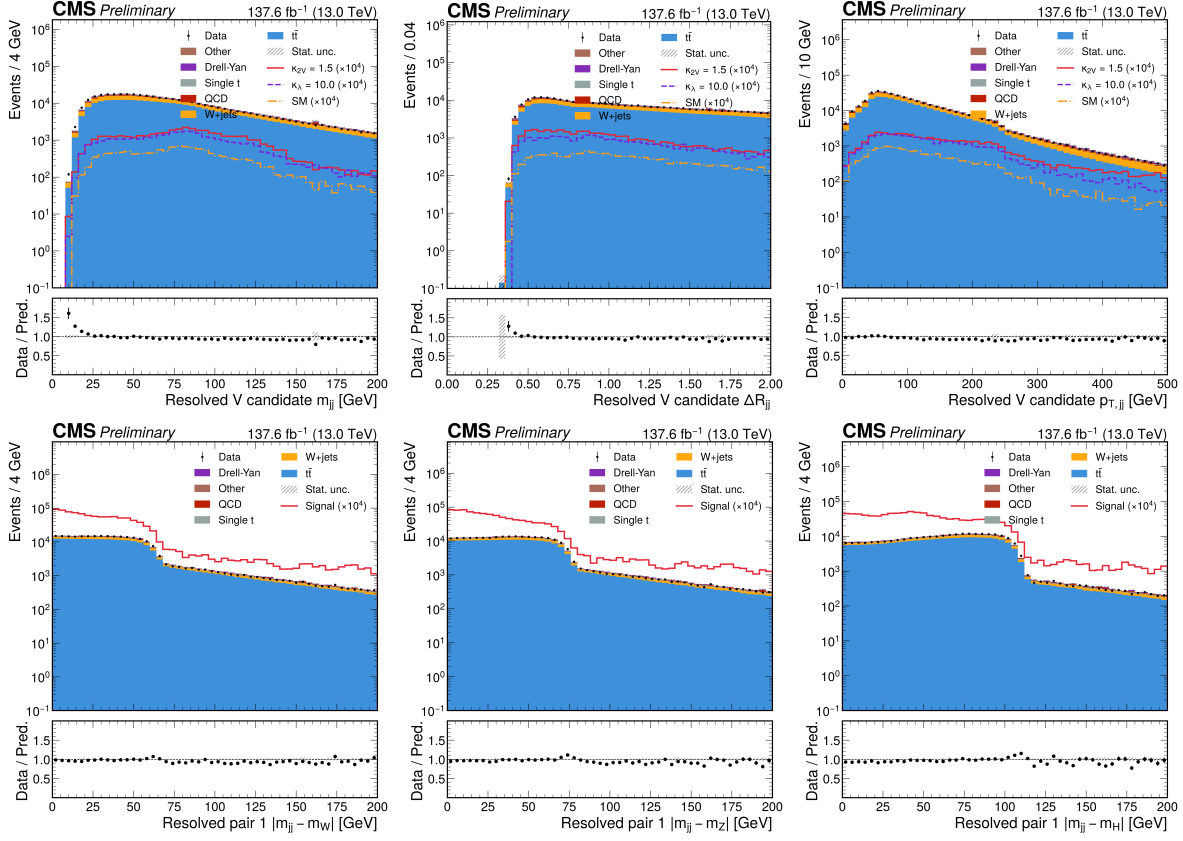


Figure A.19. Resolved system, single-lepton semi-boosted channel, Run 2: leading dijet mass, angular separation and p_T (top), and the mass residuals against the W , Z and Higgs hypotheses (bottom).

Run 3

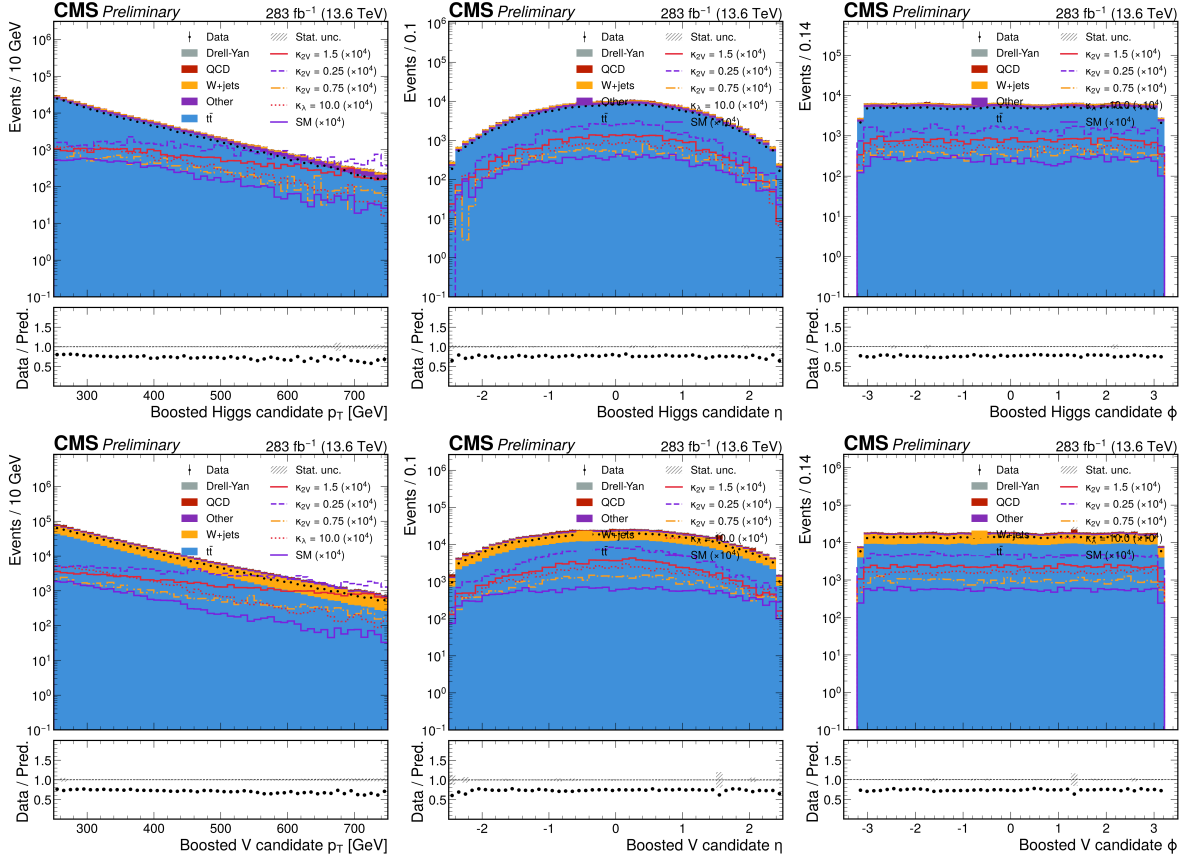


Figure A.20. Transverse momentum (left), pseudorapidity (middle), and azimuth (right) of the boosted candidates, single-lepton semi-boosted channel, Run 3.

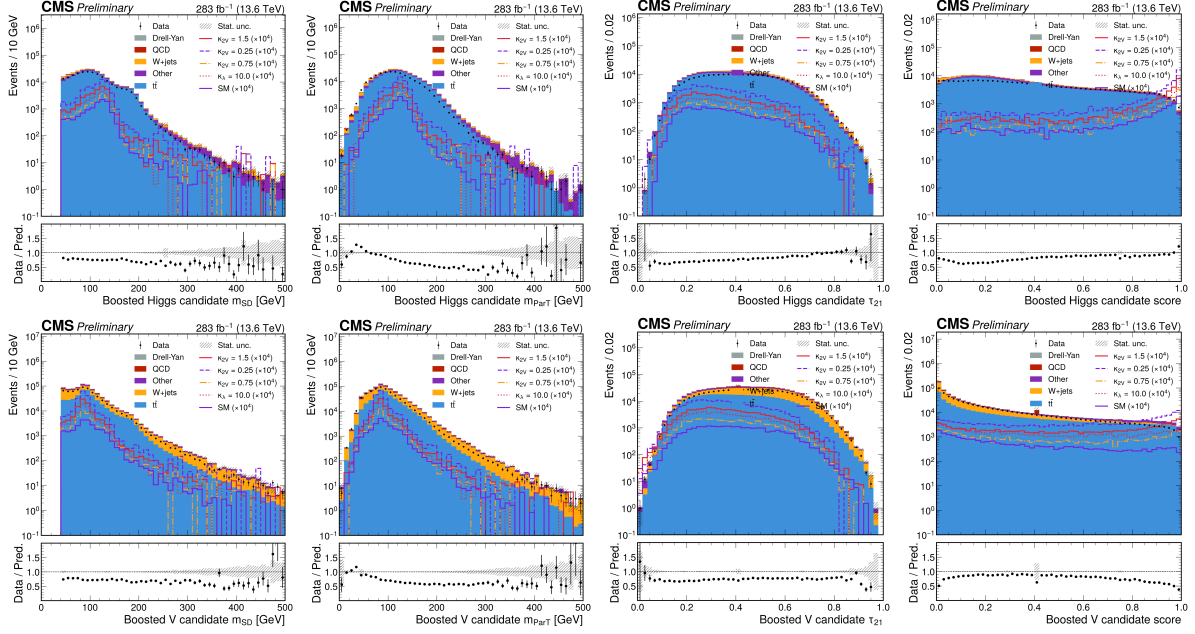


Figure A.21. Groomed mass, regressed mass, N -subjettiness ratio τ_{21} , and GLOPART tagger score (left to right) of the boosted candidates, single-lepton semi-boosted channel, Run 3.

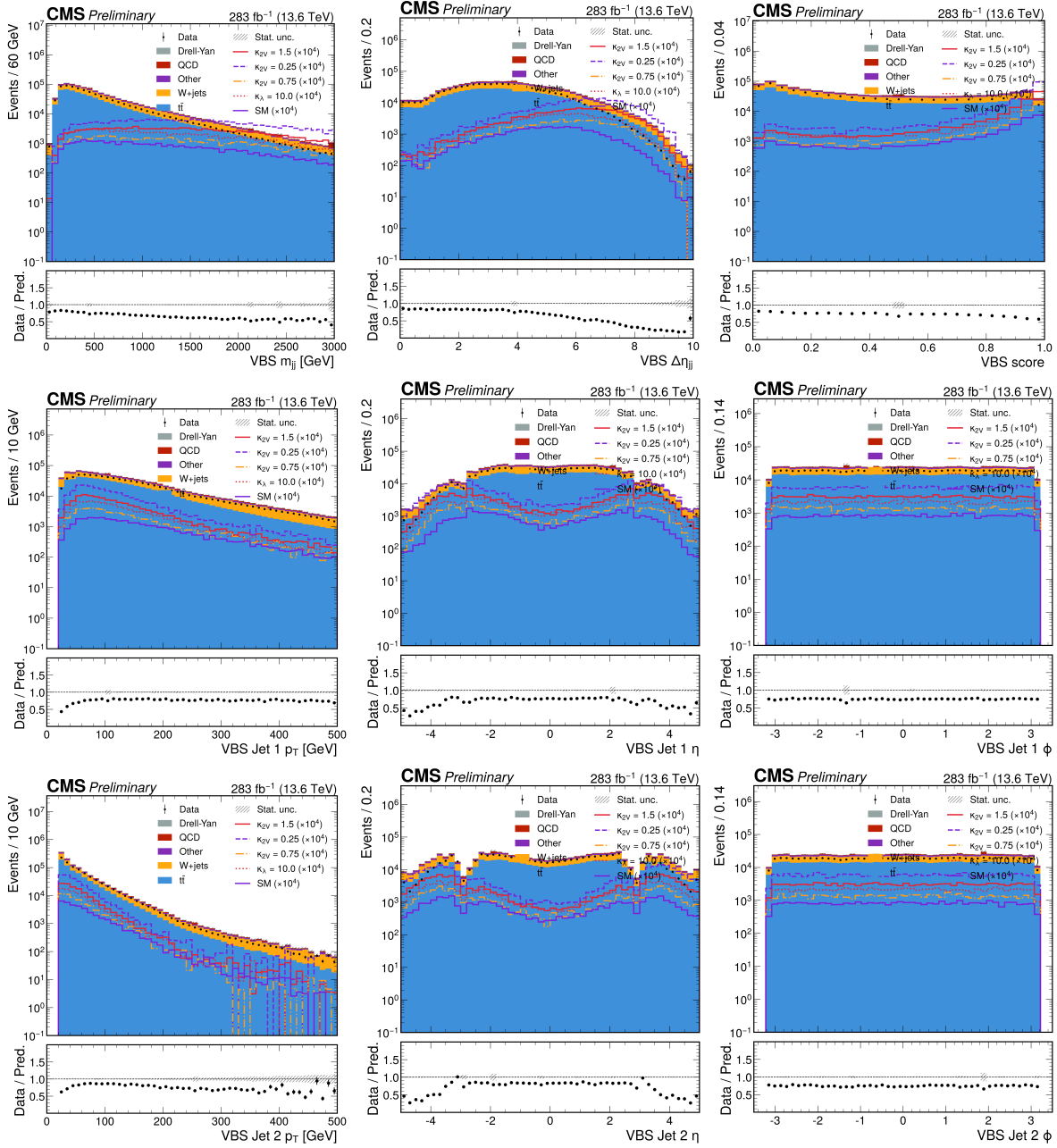


Figure A.22. Tagging-jet system, single-lepton semi-boosted channel, Run 3: pair mass, rapidity gap and BDT score (top), then the kinematics of the leading (middle) and subleading (bottom) tagging jet.

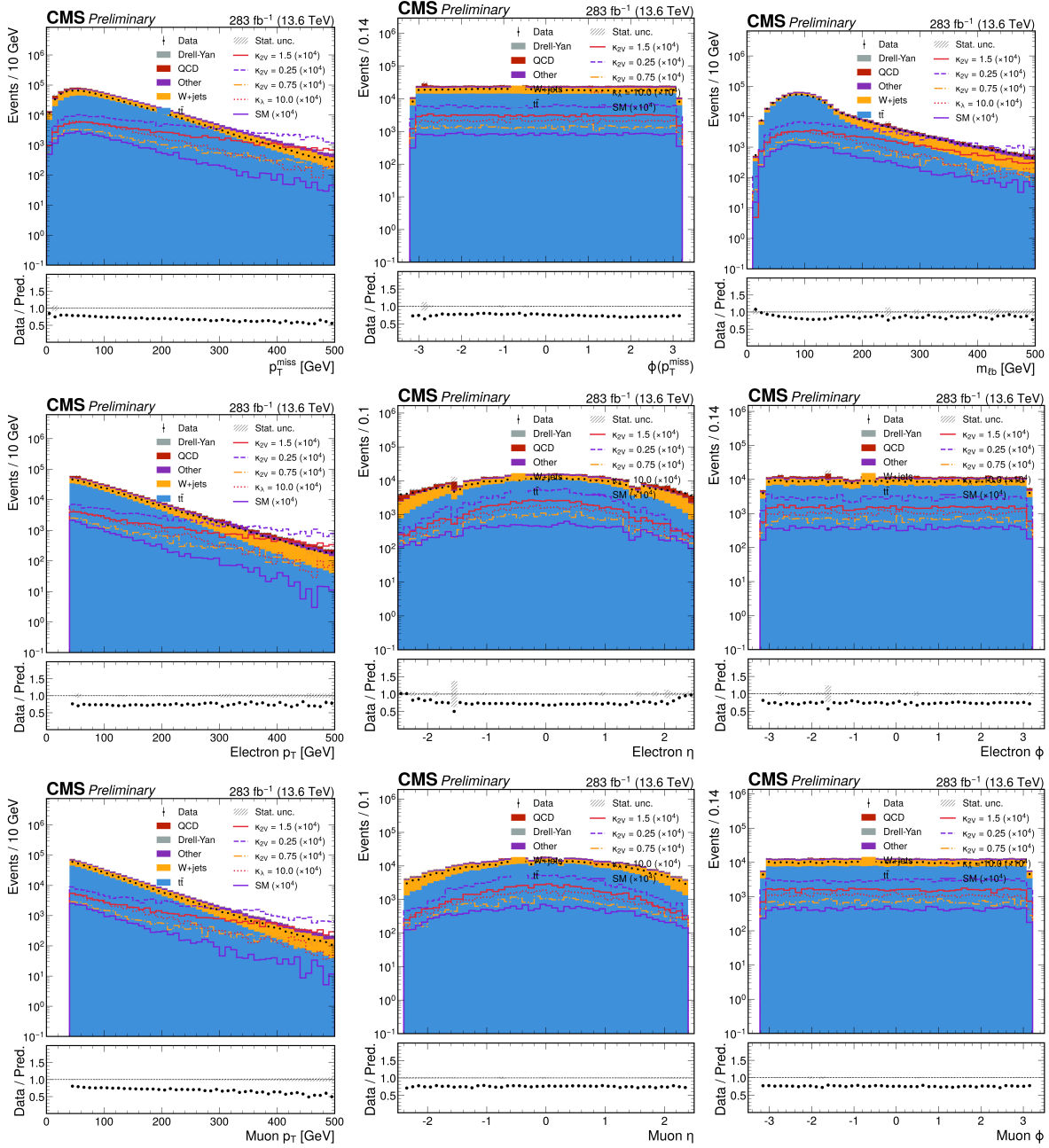


Figure A.23. Missing transverse momentum, m_{lb} , and lepton kinematics, single-lepton semi-boostered channel, Run 3.

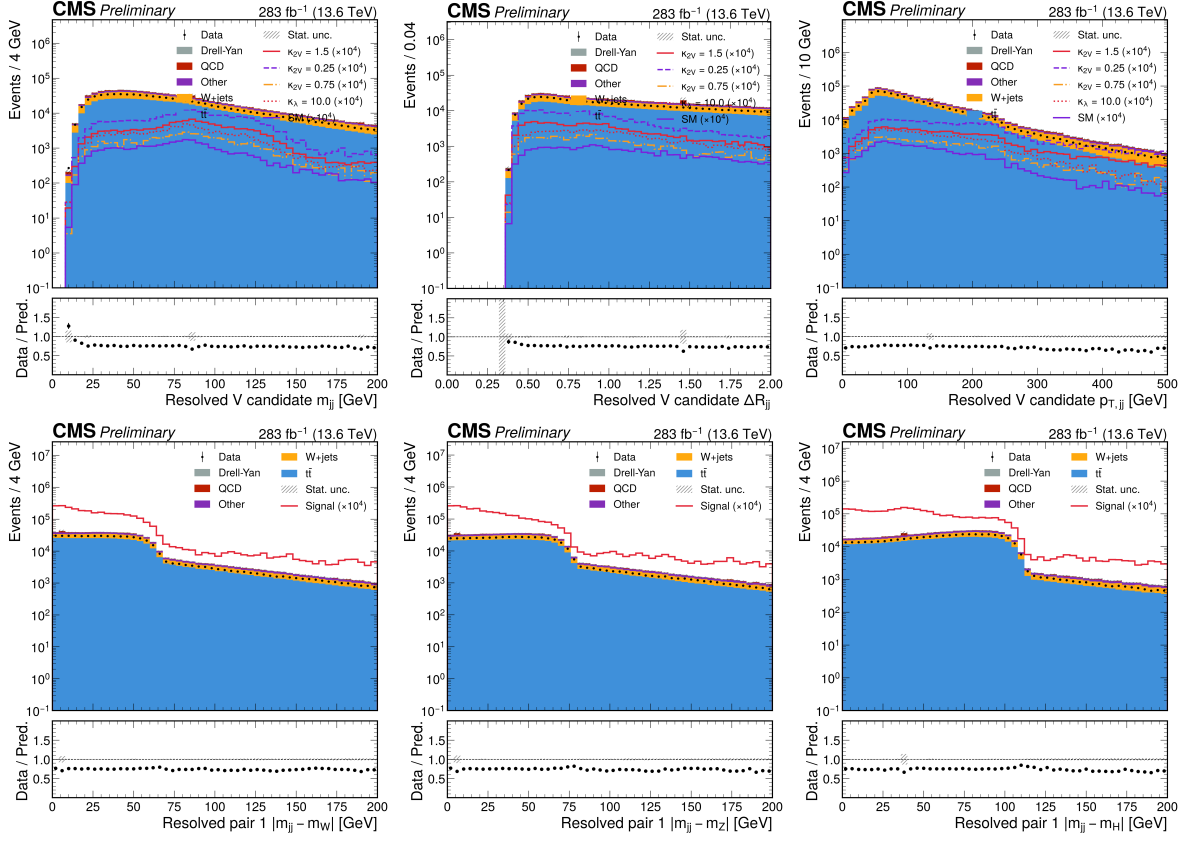


Figure A.24. Resolved system, single-lepton semi-boostered channel, Run 3: leading dijet mass, angular separation and p_T (top), and the mass residuals against the W , Z and Higgs hypotheses (bottom).

Bibliography

- [1] The Phase-2 Upgrade of the CMS Tracker. Technical report, CERN, Geneva, 2017.
- [2] Performance of mkFit and LST algorithms in the CMS Phase-2 High Level Trigger Tracking. Technical Report CMS-DP-2025-051, 2025.
- [3] CMS Resource Projection Study for HL-LHC. 2026.
- [4] S. Agostinelli et al. GEANT4—a simulation toolkit. *Nucl. Instrum. Meth. A*, 506:250–303, 2003.
- [5] A. Hayrapetyan et al. Search for associated production of a higgs boson and two vector bosons via vector-boson scattering at $\sqrt{s} = 13$ tev. *Phys. Rev. Lett.*, Jun 2026.
- [6] Simone Alioli, Paolo Nason, Carlo Oleari, and Emanuele Re. A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX. *JHEP*, 06:043, 2010.
- [7] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H.-S. Shao, T. Stelzer, P. Torrielli, and M. Zaro. The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations. *JHEP*, 07:079, 2014.
- [8] Aashay Arora, Diego Davila, Jonathan Guiang, Frank Würthwein, Harvey Newman, Justas Balcas, Tom Lehman, and Xi Yang. Data movement manager (dmm) for the sense-rucio interoperability prototype. In *EPJ Web of Conferences*, volume 337, page 01208. EDP Sciences, 2025.
- [9] Aashay Arora, Diego Davila, Frank Würthwein, John Graham, Dima Mishin, Justas Balcas, Tom Lehman, Xi Yang, Chin Guok, and Harvey Newman. Benchmarking xrootd-https on 400gbps links with variable latencies. In *EPJ Web of Conferences*, volume 337, page 01212. EDP Sciences, 2025.
- [10] Aashay Arora, Jonathan Guiang, Diego Davila, Frank Würthwein, Justas Balcas, and Harvey Newman. 400gbps benchmark of xrootd http-tpc. In *EPJ Web of Conferences*, volume 295, page 01001. EDP Sciences, 2024.
- [11] Aashay Arora, Igor Sfiligoi, Diego Davila, and Frank Würthwein. Using microsoft azure for XRootD network benchmarking, 2021.

- [12] ATLAS Collaboration. Observation of a new particle in the search for the standard model higgs boson with the ATLAS detector at the LHC. *Phys. Lett. B*, 716:1–29, 2012.
- [13] ATLAS Collaboration. A detailed map of Higgs boson interactions by the ATLAS experiment ten years after the discovery. *Nature*, 607(7917):52–59, 2022.
- [14] ATLAS Collaboration. Combination of searches for higgs boson pair production in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. *Phys. Rev. Lett.*, 133:101801, 2024.
- [15] A. A. Ayllon, M. Salichos, M. K. Simon, and O. Keeble. FTS3: New data movement service for WLCG. *J. Phys.: Conf. Ser.*, 513(3):032081, 2014.
- [16] Paolo Azzurri. The cms silicon strip tracker. *Journal of Physics: Conference Series*, 41:127–134, May 2006.
- [17] Justas Balcas, Harvey Newman, Preeti P Bhat, Frank Würthwein, Jonathan Guiang, Aashay Arora, Diego Davila, John Graham, Thomas Hutton, Tom Lehman, et al. Automated network services for exascale data movement. In *EPJ Web of Conferences*, volume 295, page 01009. EDP Sciences, 2024.
- [18] Ilya Baldin, Anita Nikolich, James Griffioen, Indermohan Inder S. Monga, Kuang-Ching Wang, Tom Lehman, Paul Ruth, and Ewa Deelman. FABRIC: A national-scale programmable experimental network infrastructure. *IEEE Internet Computing*, 23(6):38–47, 2019.
- [19] Martin Barisits, Thomas Beermann, Frank Berghaus, Brian Bockelman, et al. Rucio: Scientific data management. *Comput. Softw. Big Sci.*, 3(1):11, 2019.
- [20] David Barney. CMS detector slice. <https://cds.cern.ch/record/2120661>, 2016. CMS Collaboration, CERN Document Server record 2120661.
- [21] I. Béjar Alonso, O. Brüning, P. Fessia, M. Lamont, L. Rossi, L. Taviani, and M. Zerlauth. High-luminosity large hadron collider (HL-LHC): Technical design report. Technical Report CERN-2020-010, CERN, Geneva, 2020.
- [22] Daniele Bertolini, Philip Harris, Matthew Low, and Nhan Tran. Pileup per particle identification. *JHEP*, 10:059, 2014.
- [23] Andrea Bocci, Vincenzo Innocente, Matti Kortelainen, Felice Pantaleo, and Marco Rovere. Heterogeneous reconstruction of tracks and primary vertices with the CMS pixel tracker. *Front. Big Data*, 3:601728, 2020.
- [24] Emil Bols et al. Jet flavour classification using DeepJet. *JINST*, 15:P12012, 2020.
- [25] William Buttinger. Background estimation with the abcd method featuring the troofit toolkit. 2018.
- [26] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. The anti- k_t jet clustering algorithm. *JHEP*, 04:063, 2008.

- [27] CERN. High-luminosity LHC project schedule. <https://project-hl-lhc-industry.web.cern.ch/content/project-schedule>, 2025. Accessed 2026-07-21.
- [28] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.
- [29] CMS Collaboration. The CMS experiment at the CERN LHC. *JINST*, 3:S08004, 2008.
- [30] CMS Collaboration. Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Phys. Lett. B*, 716:30–61, 2012.
- [31] CMS Collaboration. Particle-flow reconstruction and global event description with the CMS detector. *JINST*, 12:P10003, 2017.
- [32] CMS Collaboration. A MIP timing detector for the CMS phase-2 upgrade. Technical Report CERN-LHCC-2019-003, CMS-TDR-020, CERN, Geneva, 2019.
- [33] CMS Collaboration. Extraction and validation of a new set of CMS PYTHIA8 tunes from underlying-event measurements. *Eur. Phys. J. C*, 80:4, 2020.
- [34] CMS Collaboration. A portrait of the Higgs boson by the CMS experiment ten years after the discovery. *Nature*, 607(7917):60–68, 2022.
- [35] CMS Collaboration. CMS luminosity – public results. <https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults>, 2024. Accessed 2026-07-21.
- [36] CMS Collaboration. The CMS statistical analysis and combination tool: Combine. *Comput. Softw. Big Sci.*, 8:19, 2024.
- [37] CMS Collaboration. Search for higgs boson pair production with one associated vector boson in proton-proton collisions at $\sqrt{s} = 13$ TeV. *JHEP*, 10:061, 2024.
- [38] CMS Collaboration. Combination of searches for nonresonant higgs boson pair production in proton-proton collisions at $\sqrt{s} = 13$ TeV. *arXiv:2510.07527*, 2025.
- [39] CMS Collaboration. Precise mapping of the magnetic field in the cms barrel yoke using cosmic rays. *Journal of Instrumentation*, 5(03):T03021–T03021, March 2010.
- [40] CMS Collaboration. Search for associated production of a higgs boson and two vector bosons via vector boson scattering at $\sqrt{s} = 13$ tev, 2026.
- [41] CMS Collaboration. Using Line Segment Tracking to enable ML-based Track Reconstruction in Phase-2 of the CMS Experiment. 2026.
- [42] The CMS Collaboration. Description and performance of track and primary-vertex reconstruction with the cms tracker. *Journal of Instrumentation*, 9(10):P10009–P10009, October 2014.

- [43] Glen Cowan, Kyle Cranmer, Eilam Gross, and Ofer Vitells. Asymptotic formulae for likelihood-based tests of new physics. *Eur. Phys. J. C*, 71:1554, 2011. Erratum: *Eur. Phys. J. C* 73 (2013) 2501.
- [44] D. de Florian et al. Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector. Technical report, 2017.
- [45] Alvisio Dorigo, Peter Elmer, Fabrizio Furano, and Andrew Hanushevsky. XROOTD—a highly scalable architecture for data access. *WSEAS Transactions on Computers*, 4(3):348–353, 2005.
- [46] Christoph Englert et al. VBS $w^\pm w^\pm h$ production at the HL-LHC and a 100 TeV pp -collider. *Int. J. Mod. Phys. A*, 32(18):1750106, 2017.
- [47] F. Englert and R. Brout. Broken symmetry and the mass of gauge vector mesons. *Phys. Rev. Lett.*, 13:321–323, 1964.
- [48] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pages 226–231. AAAI Press, 1996.
- [49] Lyndon Evans and Philip Bryant. Lhc machine. *Journal of Instrumentation*, 3(08):S08001–S08001, August 2008.
- [50] Edgar Fajardo, Aashay Arora, Diego Davila, Richard Gao, Frank Würthwein, and Brian Bockelman. Systematic benchmarking of https third party copy on 100gbps links using xrootd. In *EPJ Web of Conferences*, volume 251, page 02001. EDP Sciences, 2021.
- [51] R. Frühwirth. Application of kalman filtering to track and vertex fitting. *Nucl. Instrum. Meth. A*, 262:444–450, 1987.
- [52] Sheldon L. Glashow. Partial-symmetries of weak interactions. *Nucl. Phys.*, 22:579–588, 1961.
- [53] Global Network Advancement Group (GNA-G). AutoGOLE/SENSE: Automated multi-domain network service provisioning. <https://www.gna-g.net/join-working-group/autogole-sense/>, 2024. Accessed 2026-07-16.
- [54] Jonathan Guiang, Aashay Arora, Diego Davila, John Graham, Dima Mishin, Igor Sfiligoi, Frank Wuerthwein, Tom Lehman, Xi Yang, Chin Guok, et al. Integrating end-to-end exascale sdn into the lhc data distribution cyberinfrastructure. In *Practice and Experience in Advanced Research Computing 2022: Revolutionary: Computing, Connections, You*, pages 1–4. 2022.
- [55] G. S. Guralnik, C. R. Hagen, and T. W. B. Kibble. Global conservation laws and massless particles. *Phys. Rev. Lett.*, 13:585–587, 1964.

- [56] Brian Henning, Davide Lombardo, Marc Riembau, and Francesco Riva. Measuring higgs couplings without higgs bosons. *Phys. Rev. Lett.*, 123(18):181801, 2019.
- [57] Peter W. Higgs. Broken symmetries and the masses of gauge bosons. *Phys. Rev. Lett.*, 13:508–509, 1964.
- [58] Xiangyang Ju et al. Performance of a geometric deep learning pipeline for hl-lhc particle tracking. *The European Physical Journal C*, 81(10):876, 2021.
- [59] Thomas Junk. Confidence level computation for combining searches with small statistics. *Nucl. Instrum. Meth. A*, 434:435–443, 1999.
- [60] Gregor Kasieczka, Benjamin Nachman, Matthew D. Schwartz, and David Shih. Automating the abcd method with machine learning. *Physical Review D*, 103(3), February 2021.
- [61] Jan Kieseler. Object condensation: one-stage grid-free multi-object reconstruction in physics detectors, graph, and image data. *The European Physical Journal C*, 80(9):886, 2020.
- [62] H. Kolanoski and N. Wermes. *Particle Detectors: Fundamentals and Applications*. Oxford University Press, 2020.
- [63] Andrew J. Larkoski, Simone Marzani, Gregory Soyez, and Jesse Thaler. Soft drop. *JHEP*, 05:146, 2014.
- [64] LCG Collaboration. LHC computing grid: Technical design report. Technical Report CERN-LHCC-2005-024, LCG-TDR-001, CERN, 2005.
- [65] Kilian Lieret and Gage DeZoort. An object condensation pipeline for charged particle tracking at the high luminosity lhc. *EPJ Web of Conferences*, 295:09004, 2024.
- [66] Inder Monga, Chin Guok, John MacAuley, Alex Sim, Harvey Newman, Justas Balcas, Phil DeMar, Linda Winkler, Tom Lehman, and Xi Yang. Software-defined network for end-to-end networked science at the exascale. *Future Gener. Comput. Syst.*, 110:181–201, 2020.
- [67] Inder Monga, Mazahir Hussain, Justas Balcas, Aashay Arora, Diego Davila, Buseung Cho, Xi Yang, and Cees De Laat. Sense in practice: Quantifying the end-to-end benefits of intent-based bandwidth reservation for exascale science workflows. In *Proceedings of the SC '25 Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC Workshops '25. Association for Computing Machinery, 2025.
- [68] NNPDF Collaboration. Parton distributions from high-precision collider data. *Eur. Phys. J. C*, 77:663, 2017.
- [69] Emmy Noether. Invariante variationsprobleme. *Nachr. d. König. Gesellsch. d. Wiss. zu Göttingen, Math-phys. Klasse*, pages 235–257, 1918.

- [70] None None. *Basic Energy Sciences Network Requirements Review (Final Report)*. November 2022.
- [71] Particle Data Group. Review of particle physics. *Phys. Rev. D*, 110(3):030001, 2024.
- [72] Michael E. Peskin. *Concepts of Elementary Particle Physics*. Oxford Master Series in Physics. Oxford University Press, 9 2019.
- [73] Larry L. Peterson and Bruce S. Davie. *Computer Networks: A Systems Approach*. Morgan Kaufmann, San Francisco, CA, 5th edition, 2011.
- [74] Huilin Qu and Loukas Gouskos. ParticleNet: Jet tagging via particle clouds. *Phys. Rev. D*, 101(5):056019, 2020.
- [75] Huilin Qu, Congqiao Li, and Sitian Qian. Particle transformer for jet tagging. *Proceedings of the 39th International Conference on Machine Learning (PMLR)*, 162:18281–18292, 2022.
- [76] Alexander L. Read. Presentation of search results: the CL_s technique. *J. Phys. G*, 28:2693–2704, 2002.
- [77] Abdus Salam. Weak and electromagnetic interactions. In N. Svartholm, editor, *Elementary Particle Theory: Relativistic Groups and Analyticity (Nobel Symposium No. 8)*, pages 367–377, Stockholm, 1968. Almqvist & Wiksell.
- [78] Matthew D. Schwartz. Tasi lectures on collider physics, 2017.
- [79] Torbjörn Sjöstrand, Stefan Ask, Jesper R. Christiansen, Richard Corke, Nishita Desai, Philip Ilten, Stephen Mrenna, Stefan Prestel, Christine O. Rasmussen, and Peter Z. Skands. An introduction to PYTHIA 8.2. *Comput. Phys. Commun.*, 191:159–177, 2015.
- [80] Larry Smarr, Thomas DeFanti, John Graham, et al. The national research platform: Stretched, multi-tenant, scientific Kubernetes cluster. In *Practice and Experience in Advanced Research Computing (PEARC '25)*. Association for Computing Machinery, 2025.
- [81] Jesse Thaler and Ken Van Tilburg. Identifying boosted objects with N-subjettiness. *JHEP*, 03:015, 2011.
- [82] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.
- [83] Emmanouil Vourliotis, Philip Chang, Peter Elmer, Yanxi Gu, Jonathan Guiang, Vyacheslav Krutelyov, Balaji Venkat Sathia Narayanan, Gavin Niendorf, Michael Reid, Mayra Silva, Andres Rios Tascon, Matevž Tadel, Peter Wittich, and Avraham Yagil. Line segment tracking: Improving the phase 2 cms high level trigger tracking with a novel, hardware-agnostic pattern recognition algorithm, 2024.

- [84] M. Vretenar et al. Linac4 design report. *CERN Yellow Reports: Monographs*, 6(CERN-2020-006), 2020.
- [85] Steven Weinberg. A model of leptons. *Phys. Rev. Lett.*, 19:1264–1266, 1967.
- [86] Frank Würthwein, Jonathan Guiang, Aashay Arora, Diego Davila, John Graham, Dima Mishin, Thomas Hutton, Igor Sfiligoi, Harvey Newman, Justas Balcas, et al. Managed network services for exascale data movement across large global scientific collaborations. In *2022 4th Annual Workshop on Extreme-scale Experiment-in-the-Loop Computing (XLOOP)*, pages 16–19. IEEE, 2022.