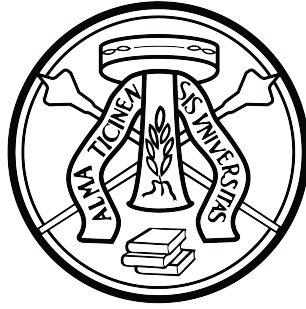




UNIVERSITÀ
DI PAVIA

DIPARTIMENTO DI FISICA

CORSO DI LAUREA MAGISTRALE IN SCIENZE FISICHE

A Phenomenological Study
for $H \rightarrow Za$ Searches

Relazione per la laurea di

Alessandro Codega

Mat. 525792

Supervisore:

Prof.ssa Daniela Rebuzzi

Correlatore:

Dr. Sonia Carrà

Dr. Giovanni Stagnitto

A.A. 2025-2026

*Buttando nel vento
Il lavoro di anni
Perché nemmeno da vecchi si sa
Cosa faremo da grandi*

Lucio Corsi,
"Cosa Faremo da Grandi"

*Douce mémoire, en plaisir consommée;
O siècle heureux que cause tel savoir [...]
Fini le bien, le mal soudain commence.*

Pierre Sandrin,
"Douce mémoire"

And words are futile devices

Sufjan Stevens,
"Futile Devices"

*"E alla fine di un lavoro lungo
ci si siede e si guarda il mare."*

— Cesare Pavese, *Il mestiere di vivere*

I might not end up
amounting to much
but as long as
I treasure the gift
of softness and love
I feel like I might achieve
something for myself

Abstract

The discovery of the Higgs boson by the CERN experiments in 2012 marked a milestone in high-energy physics, confirming the predictive power of the Standard Model but also opening new possibilities for the search of new physics phenomena. This thesis work builds on a previous search carried out by the ATLAS experiment on LHC Run-2 data and a phenomenological study is performed with the aim of improving future searches for new physics at LHC. The theoretical framework used is the 2-Higgs-Doublet Model, where the Higgs sector is extended with an extra scalar doublet and new Higgs boson decay channels. The Beyond-the-Standard-Model Higgs boson decay channel $H \rightarrow Za$ is considered, where the a particle is a light pseudoscalar decaying hadronically into a boosted, two-pronged jet. Two goals are pursued. First, a comparison between Monte Carlo generators, PYTHIA 8 and HERWIG 7, quantifies the nonperturbative modeling uncertainty on jet substructure observables, assessing the degree to which **SoftDrop** grooming can reduce the inter-generator spread. Second, a two-stage neural network pipeline — a multi-layer perceptrons regression for the pseudoscalar mass prediction followed by a classification for signal-versus-background separation — is developed and evaluated across generator and hadronization configurations. The impact of jet grooming technique is also considered. The robustness of the results is verified through a study employing balanced regression training.

Contents

Abstract	vii
Introduction	1
1 The Higgs Boson in the Standard Model	5
1.1 Introduction	5
1.2 Electroweak Theory: Before the Higgs Boson	12
1.3 The Goldstone Model	25
1.4 Electroweak Theory: After SSB	35
1.5 Higgs Boson Phenomenology	44
1.5.1 Higgs Boson Production	49
1.5.2 Higgs Boson Decay	51
1.5.3 Higgs Boson Discovery	54
1.5.4 Higgs Boson Quantum Numbers	56
2 Searches for The Higgs Boson Beyond the Standard Model	59
2.1 Shortcomings of the Standard Model	60
2.1.1 Dark Matter	60
2.1.2 Gravity	61
2.1.3 Matter-Antimatter Asymmetry	61
2.1.4 Neutrino Masses	62
2.1.5 Hierarchy Problem	62
2.2 Beyond the Minimal Higgs Model: Type-1 2HDM	68
2.3 Phenomenology of the Type-1 2HDM	72
3 Tools for The Analysis	77
3.1 The Experimental Setting	78
3.1.1 The Large Hadron Collider	78
3.1.2 The ATLAS Experiment at CERN	79

3.1.3	Charged Particle Tracking and Jet Inputs at ATLAS	82
3.1.4	Analysis Motivation	83
3.1.5	Monte Carlo Simulation	84
3.1.6	Event Generation Example	86
3.2	Looking Inside Jets	89
3.2.1	Quantum Chromodynamics	90
3.2.2	Ultraviolet Divergences and The Running of the Strong Coupling Constant	91
3.2.3	Infrared Divergences in Quantum Chromodynamics	94
3.2.4	The Lund Plane	98
3.2.5	Jet Reconstruction Algorithms	100
	Generalized- k_t Algorithm	102
	k_t Algorithm	102
	Cambridge/Aachen Algorithm	102
	Anti- k_t Algorithm	103
	Experimental Aspects	103
3.2.6	Jet Substructure Variables	103
	Track Multiplicity	106
	Leading Track Observables	106
	Angularity	107
	Energy Correlation Observables	108
	N -subjettiness	109
	Track-based Observables and Non-perturbative Sensitivity	109
3.2.7	Jet Grooming	110
	The Cambridge/Aachen Declustering Tree	111
	SoftDrop	111
	Effect on Substructure Observables	112
3.3	Monte Carlo Event Generators: PYTHIA and HERWIG	115
3.3.1	Why Monte Carlo Simulation Is Necessary	115
3.3.2	Structure of a Simulated Event	115
3.3.3	Parton Showers and Hadronization Modeling: PYTHIA and HERWIG	117
3.3.4	The Underlying Event and Multiple Parton Interactions	120
3.4	Jet Reconstruction and Neural Networks	121
3.5	Analysis Setup and Pipeline	122

4	Results of $H \rightarrow Za$	125
4.1	Previous Analysis of $H \rightarrow Za$	126
4.1.1	Data and Simulated Event Samples	126
4.1.2	Event Selection	128
4.2	New Phenomenological Study	129
4.3	Kinematic Validation	130
4.4	Comparison of Jet Substructure Variables	133
4.4.1	Workflow	133
4.4.2	Jet Substructure Variables Comparison: Ungroomed Case	136
4.4.3	Effect of Hadronization: HAD-on vs HAD-off	140
4.4.4	Groomed Variables	144
4.4.5	Comparison between Ungroomed and Groomed Variables	147
4.5	Neural Network Performance	149
4.5.1	The Two-Stage Pipeline	149
4.5.2	Study A: Delphes Training and Truth-Level Evaluation	151
4.5.3	Study B: Self-Consistent Detector-Level Evaluation	156
4.5.4	Study C: Balanced Regression Training	169
5	Conclusions and Outlook	175
5.1	Summary of the Work	175
5.2	Conclusions	176
5.2.1	Generator Comparison	176
5.2.2	Neural Network Analysis	177
5.2.3	Central Physics Result	177
5.3	Outlook	178
	Bibliography	181

List of Figures

1.1	A contact interaction for β^- decay.	12
1.2	Comparison of weak interactions mediated by W^+ and W^- bosons.	13
1.3	Charged-current interaction vertices.	15
1.4	Neutral current vertex for ν_L	22
1.5	Quartic potentials depending on λ	26
1.6	Mexican hat potential.	28
1.7	Oscillations around the Mexican hat potential's minima.	29
1.8	Feynman diagram for the $e^+e^- \rightarrow f\bar{f}$ process	46
1.9	Parton model proton scattering, two quarks only interact.	46
1.10	Proton-proton collision event generated with Sherpa. Courtesy of [17].	47
1.11	Comparison between PDFs for the proton at two different scales: 10^2GeV (left), 10^4GeV (right). Courtesy of [18].	48
1.12	A 3D rendering of the important variables for a detector experiment, in this case, CMS.	49
1.13	Representative Feynman diagrams for key Higgs boson production mechanisms at hadron colliders. Top row: direct production via gluon fusion (A), vector boson fusion (B), and <i>Higgs-strahlung</i> (C). Bottom row: associated production with top quarks in $t\bar{t}H$ (D), tHq (E), and tHW (F).	50
1.14	Feynman diagrams of Higgs boson decay channels.	53
1.15	(a) The distribution of the four-lepton invariant mass, $m_{4\ell}$, for the selected candidates, compared to the background expectation in the 80–250 GeV mass range. The signal expectation for a SM Higgs with $m_H = 125$ GeV is also shown. Taken from [20].	55
1.15	(b) The distributions of the invariant mass of diphoton candidates after all selections for the combined 7 TeV and 8 TeV data sample. The inclusive sample is shown in (a) and a weighted version of the same sample in (c). Taken from [20].	55

1.15	Distribution of the transverse mass m_T . The expected signal for $m_H = 125$ GeV is shown stacked on top of the background prediction. Taken from [20].	56
2.1	Higgs self-energy diagram with a fermion loop.	63
2.2	Higgs self-energy diagram with a scalar particle S loop.	66
3.1	Scheme of the accelerator machines that feed into the LHC beampipe. Courtesy of [30].	79
3.2	The ATLAS experimental apparatus in its 3D rendition. Courtesy of [31]. .	80
3.3	Illustration of the ATLAS coordinate system(s). Taken from [32].	81
3.4	A slice of ATLAS as well as visualizations of how ATLAS detects different particles. Taken from [33].	82
3.5	The flow of the production of MC events and real data for data analysis. The reconstruction step should work for both MC events (solid) and real data (dashed). Courtesy of [35].	85
3.6	Feynman diagrams of $t\bar{t}$ process in the pp collisions: $t \rightarrow W^+ b \rightarrow u \bar{d} b$, $\bar{t} \rightarrow W^- \bar{b} \rightarrow \bar{\ell} \nu_{\ell} \bar{b}$ with s-channel (left) and t-channel (right)	86
3.7	Proton-proton collision event generated with SHERPA. The different steps of MC simulations are highlighted in different colors. Courtesy of [17]. . . .	87
3.8	Illustration of a event as generated by a Monte Carlo event generator. Quarks and gluons in the initial state (blue) contribute with fractions of the energy of the colliding protons according to PDFs. The hard interaction and additional hard QCD emissions are shown in red. An underlying event interaction is visible in purple. After parton showering and hadronization of intermediate hadrons (light green), the decay into the final state takes place (dark green). Additional photon radiations are added in yellow. Courtesy of [39].	89
3.9	One-loop QCD self-energy and vertex corrections.	92
3.10	Running of QCD and QED coupling constants.	94
3.11	Real and virtual final state radiation in $e^+e^- \rightarrow \gamma^* \rightarrow q\bar{q}(g)$. Momenta $\mathbf{p}_1, \mathbf{p}_2$ denote the incoming e^- and e^+ , $\mathbf{k}_1, \mathbf{k}_2$ the outgoing quark and antiquark, $\mathbf{k}_3$ the emitted or virtual gluon.	95
3.12	The Lund plane.	99
3.13	Correspondence between jet spatial geometry and Lund-plane phase space.	100

3.14	Comparison of the k_t [40], Cambridge/Aachen [41], and anti- k_t [42] jet clustering algorithms. 100-particle toy event, 4 hard seeds, $R = 0.5$. Row 1: collision products. Row 2: distance metric; dashed circles show radius R . Row 3: merge order. Row 4: jet boundaries forming. Row 5: final jets. Soft particles colored by assigned jet in rows 4–5.	104
3.15	Lego plots of the toy collision event for the k_t , Cambridge/Aachen, and anti- k_t jet clustering algorithms. Each bar shows the summed transverse momentum $\sum p_T$ deposited in a $(\Delta\eta \times \Delta\phi)$ simulated calorimeter cell. Bars are colored by their assigned jet; gray bars fall outside all jet cones. The dominant spike in each plot corresponds to the hardest jet (A , $p_T = 220$ GeV).	105
3.16	The SoftDrop boundary visualized in the Lund plane. Anything that is below the boundary is discarded.	113
3.17	Visualization for the Lund hadronization string model adopted by PYTHIA.	119
3.18	Generator and file pipeline.	124
4.1	Feynman diagrams for the observed signal decays for the $H \rightarrow Za$ search.	127
4.2	Invariant mass of the lepton pair plus jet system for data, before fit.	129
4.3	Invariant mass of the lepton pair plus jet system for data, after fit.	129
4.4	The invariant mass $m_{\ell\ell j}$ shows good accord between generators up to statistical fluctuations.	131
4.5	All variables show good accord between generators up to statistical fluctuations. The HERWIG signal is split into the two decay channels, but the plots are normalized.	132
4.6	Workflow for the event generation in PYTHIA/HERWIG.	135
4.7	Kinematical variables from PYTHIA and HERWIG full simulation.	138
4.8	Jet substructure variables reconstructed from PYTHIA and HERWIG full simulation.	139
4.9	Kinematical variables from PYTHIA and HERWIG full simulation, with dashed lines for HAD-off and solid lines for HAD-on	142
4.10	Jet substructure variables reconstructed from PYTHIA and HERWIG full simulation, with dashed lines for HAD-off and solid lines for HAD-on	143
4.11	Groomed kinematical variables from PYTHIA and HERWIG, with dashed lines for HAD-off and solid lines for HAD-on	145

4.12	Groomed jet substructure variables reconstructed from PYTHIA and HERWIG, with dashed lines for HAD-off and solid lines for HAD-on	146
4.13	Schematic of the two-stage neural network pipeline. The seven ghost-track substructure variables are first passed to the regression network, which returns a predicted pseudoscalar mass $\hat{m}_a$. This prediction is appended to the original feature vector, and the augmented eight-feature input is passed to the classification network, which returns a continuous signal-versus-background score.	150
4.14	Study A regression network: loss curves (left) and output distribution on the DELPHES test set (right).	153
4.15	ROC curve for the Study A classification network evaluated on the DELPHES test set at $m_a = 3.00$ GeV. The AUC of 0.831 quantifies the overall separation power of the classifier.	153
4.16	Normalized regression output distributions for the HAD-on (left) and HAD-off (right) truth-level configurations in Study A. All truth-level peaks appear at $\hat{m}_a \approx 2.1$ GeV instead of the true 3.0 GeV, quantifying the ~ 0.9 GeV domain shift introduced by applying the DELPHES-trained model to truth-level inputs.	154
4.17	Effect of SoftDrop grooming on the regression output distribution for PYTHIA (left) and HERWIG (right) with full hadronization. Solid lines show the ungroomed distribution; dashed lines show the groomed distribution. The peak is preserved in both cases, while the ungroomed low-multiplicity artifact at $\hat{m}_a \approx 1.45$ GeV disappears after grooming.	155
4.18	Summary of classification AUC values for the DELPHES baseline and all eight truth-level configurations in Study A. The red dashed line marks the in-domain AUC of 0.831. Groomed configurations (lighter bars) consistently underperform their ungroomed counterparts.	157
4.19	ROC curves for the Study A classifier evaluated on HAD-on truth-level configurations. Ungroomed inputs (left column) yield $\text{AUC} \approx 0.63$, substantially below the DELPHES baseline but still above random. Groomed inputs (right column) yield $\text{AUC} \approx 0.37$, below the random-classifier threshold of 0.5: the network has effectively inverted its classification (see text for interpretation).	158

- 4.20 ROC curves for the Study A classifier evaluated on **HAD-off** truth-level configurations. Ungroomed inputs (left column) yield marginally higher AUC values than their **HAD-on** counterparts, consistent with parton-level features being slightly closer to the DELPHES training distribution. Groomed inputs (right column) remain near or below the random-classifier threshold. 159
- 4.21 Regression output distributions for all four evaluation configurations in Study B. Ratio panels use the PYTHIA **HAD-on** distribution as reference. . . 161
- 4.22 ROC curves for the UNG (AUC = 0.983) and GRO (AUC = 0.618) classifiers evaluated on the DELPHES test set at $m_a = 3.0$ GeV. The gap between the two curves quantifies the discrimination power lost by applying **SoftDrop** grooming before feature extraction. 162
- 4.23 Classifier output score distributions for the GRO (top) and UNG (bottom) classifiers in Study B. *Left panels:* normalized score distributions for the in-domain signal ($m_a = 3$ GeV, PYTHIA **HAD-on**, blue) and the Z +jets background (orange), confirming that both classifiers achieve in-domain signal-background separation consistent with the AUC values of Table 4.6. *Right panels:* signal score distributions across all four evaluation configurations (PYTHIA and HERWIG, **HAD-on** and **HAD-off**), with the Z +jets background shown as a filled histogram for reference. The GRO right panel shows the four evaluation curves nearly superimposed, confirming that the groomed classifier generalizes stably across generator and hadronization configurations. The UNG right panel reveals the generalization collapse: while the in-domain signal peak sits near a score of 0.72, the HERWIG **HAD-on** signal distribution (solid red) collapses onto the background peak, consistent with the near-random AUC of 0.507 reported in Table 4.6. This contrast illustrates that visual inspection of the in-domain score distribution alone is insufficient to assess classifier robustness; evaluation across the full set of generator configurations is required. 163
- 4.24 Track multiplicity N_{tracks} for all four evaluation configurations at $m_a = 3$ GeV, shown for the ungroomed (left) and **SoftDrop**-groomed (right) jet definitions. PYTHIA **HAD-on** (solid blue) and HERWIG **HAD-on** (solid red) differ in both cases; the **HAD-off** distributions (dashed) are shifted to lower multiplicities. Grooming reduces the inter-generator spread but leaves a residual difference at intermediate track counts, which is the principal source of the groomed generator gap $\Delta\text{AUC} = 0.089$ 165

4.25	ROC curves for the HAD-on evaluation configurations in Study B. The UNG AUC values of 0.549 and 0.507 represent a collapse of ≈ 0.44 relative to the in-domain baseline. The GRO AUC values of 0.644 and 0.555 are consistent with the GRO in-domain baseline of 0.618, confirming that the groomed classifier generalizes without collapse.	166
4.26	ROC curves for the HAD-off evaluation configurations in Study B. UNG entries ($\dagger$) are computed from ≈ 50 valid jets and excluded from physics conclusions. The GRO classifier recovers valid features for 84–89% of parton-level jets, achieving AUC values of 0.675 and 0.756.	167
4.27	Summary of classification AUC values for the UNG (dark bars) and GRO (light bars) classifiers across all evaluation configurations in Study B. Vertical dashed lines mark the respective in-domain baselines (UNG = 0.983, GRO = 0.618). The GRO classifier outperforms the UNG classifier on every HAD-on configuration despite its lower in-domain AUC. Entries marked $\dagger$ in the previous plots are excluded from physics conclusions.	168
4.28	Classifier output score distributions for Study C. Left panels: in-domain signal ($m_a = 3$ GeV, PYTHIA HAD-on) vs Z +jets background. Right panels: signal score across all four evaluation configurations. The GRO right panel shows the four evaluation curves nearly superimposed, as in Study B. The UNG right panel shows partial recovery of the generalization collapse: the HERWIG HAD-on signal distribution (solid red) no longer fully collapses onto the background peak, consistent with the improved evaluation AUC of 0.565 reported in Table 4.7.	173

List of Tables

1.1	Degrees of freedom in a locally $U(1)$ -invariant theory. Counting is not accurate.	33
1.2	Degrees of freedom in a locally $U(1)$ -invariant theory. Counting is accurate.	33
1.3	SM Higgs production cross sections and fractional contributions at $\sqrt{s} = 13$ TeV, $m_H = 125$ GeV (ATLAS Run 2 data: [19])	51
1.4	Higgs boson decay branching ratios at $\sqrt{s} = 13$ TeV, $m_H = 125$ GeV.	53
2.1	Gauge and scalar interaction vertices for the Type-1 2HDM, part 1.	73
2.2	Gauge and scalar interaction vertices for the Type-1 2HDM, part 2.	74
3.1	Branching ratios of the Standard Model Higgs boson at $m_H = 125$ GeV, including rare decays into charmonium and light resonances. The last three lines are theoretical estimations. The others are taken from [34].	84
3.2	Summary of the jet substructure variables used in this study. All variables are computed from charged particle tracks with $p_T > 0.5$ GeV, $ \eta < 2.5$, and $\Delta R < R = 0.4$. The IRC safety column refers to the theoretical status of the observable at parton level.	110
4.1	Monte Carlo generators and configurations used for signal and background simulation in the ATLAS search for $h \rightarrow Za$ with $a \rightarrow$ hadrons.	127
4.2	Summary of event and jet selection criteria used in the analysis.	135
4.3	Regression performance on the DELPHES detector-level test set. Signal accuracy is the fraction of predictions within 1 GeV of the true m_a	152
4.4	Classification AUC for the DELPHES-trained network of Study A evaluated on truth-level signal mixed with DELPHES Z +jets background.	156
4.5	Per-class regression test accuracy and MAE for the ungroomed (UNG) and groomed (GRO) studies in Study B, evaluated on the DELPHES test split.	160

4.6	Classification AUC for the self-consistent detector-level evaluation in Study B. Entries marked † are computed from approximately 50 valid jets and are excluded from all physics conclusions.	164
4.7	Classification AUC for Study C compared to Study B. Entries marked † are computed from approximately 50 valid jets and are excluded from physics conclusions.	171

Introduction

*“And you may ask yourself,
well...how did I get here?”*

— Talking Heads, *“Once in a Lifetime”*



THE discovery of a Higgs boson at the Large Hadron Collider (LHC) in 2012 by the ATLAS and CMS collaborations [1, 2] marked a milestone in the history of particle physics, completing the particle content predicted by the Standard Model (SM) and confirming the Brout–Englert–Higgs mechanism of electroweak symmetry breaking. Yet the discovery of the Higgs boson, rather than closing the book on the Higgs sector, has opened a new chapter of precision measurements and exotic searches. The SM Higgs boson is the minimal realization of electroweak symmetry breaking, and its existence does not preclude the presence of additional scalar or pseudoscalar states appearing in extended Higgs sectors. Searching for and characterizing these states is one of the central tasks of the LHC physics program in the coming years.

Beyond the Standard Model theories are necessary to be able to include all of the phenomena that are not well explained by the Standard Model known up until now, such as the dark matter problem, matter/antimatter asymmetry, neutrino masses and the hierarchy problem, and more. Therefore, the Standard Model as we know it requires extensions. Among the most compelling theoretical extensions of the minimal Higgs sector is the Two-Higgs-Doublet Model (2HDM) [3], which introduces a second $SU(2)$ scalar doublet and predicts a richer spectrum of physical states: two CP-even scalars, one CP-odd pseudoscalar, and a charged Higgs pair. In the Type-1 2HDM, one of the neutral scalars plays the role of the SM-like Higgs boson, while the pseudoscalar a can be made light without conflicting with electroweak precision constraints. This opens the possibility of observing the decay $H \rightarrow Za$, where the 125 GeV Higgs boson decays into a Z boson and a light

pseudoscalar. If the pseudoscalar is sufficiently light ($m_a \lesssim \text{few GeV}$), it will be produced with a large Lorentz boost and decay predominantly into a collimated pair of hadrons, resulting in a signature consisting of two opposite-sign same-flavor leptons from the Z decay and a single boosted jet with a distinctive two-pronged internal substructure.

The ATLAS Collaboration has searched for this process using the full Run 2 dataset at $\sqrt{s} = 13$ TeV, obtaining no evidence for new low-mass pseudoscalar particles [4]. A key limiting factor in the sensitivity of this search is the large theoretical systematic uncertainty on the signal yield from the choice of Monte Carlo generator, quoted at approximately 60% from a PYTHIA/HERWIG comparison. This uncertainty directly limits the statistical power that additional luminosity can provide, and motivates a detailed investigation of its origin and potential mitigation.

The jet substructure of the a decay plays a dual role in this context. On one hand, substructure observables — track multiplicity, energy correlations, N -subjettiness — are the natural discriminating variables for separating the two-pronged pseudoscalar signal from the smoothly distributed QCD background. On the other hand, these same observables are known to be sensitive to nonperturbative effects such as hadronization and the underlying event, which are modelled differently by different generators [5]. The generator dependence of jet substructure is therefore not merely a theoretical inconvenience: it is the principal source of the systematic that limits the search. Understanding which observables are robust and which are sensitive to modeling assumptions, and whether jet grooming techniques can reduce this sensitivity, is a necessary step toward a more precise measurement.

Research in machine learning has demonstrated that neural networks trained on jet substructure inputs can significantly outperform traditional cut-based analyses for boosted-object tagging [6]. The question of whether such a network can improve the sensitivity of the $H \rightarrow Za$ search — and, crucially, whether its performance is robust against the generator systematic that limits the existing analysis — is the second core question addressed in this work. The risk of overfitting to generator-specific patterns is particularly acute when the training sample is small and the observable space is dominated by nonperturbative radiation, making a careful study of generalization across generators an essential part of any machine-learning approach to this channel.

This thesis presents a phenomenological study that addresses both of these questions within a unified simulation and analysis framework. The work is organized as follows.

Chapter 1 provides the theoretical foundations of the study. Starting from the gauge-theoretic structure of the Standard Model and the role of the Brout–Englert–Higgs mechanism in generating gauge-boson and fermion masses, the chapter builds toward the phenomenology of the SM Higgs boson: its production mechanisms at hadron colliders, its decay modes, its discovery in 2012, and the measurement of its quantum numbers. The emphasis throughout is on developing the physical intuition required to appreciate why the Higgs sector is simultaneously the best-tested and the most theoretically uncertain part of the SM.

Chapter 2 motivates the search for physics beyond the minimal Higgs sector. After reviewing the principal shortcomings of the Standard Model — the hierarchy problem,

the absence of a dark matter candidate, the matter–antimatter asymmetry, and the puzzle of neutrino masses — the chapter introduces the Type-1 Two-Higgs-Doublet Model as a theoretically well-motivated extension. The physical spectrum of the 2HDM, its gauge and scalar interaction vertices, and the phenomenological constraints on the model parameters are discussed, culminating in the identification of $H \rightarrow Za$ as one of the accessible new-physics signatures at the LHC.

Chapter 3 describes the experimental and theoretical tools used in the analysis. The ATLAS detector and the LHC environment are introduced, followed by a detailed treatment of the theory and phenomenology of jets: the running of the strong coupling, infrared and collinear safety, the Lund plane, jet reconstruction algorithms — in particular the anti- k_t and Cambridge/Aachen algorithms as implemented in **FastJet** — and the suite of jet substructure observables employed in this study. The chapter also describes the Monte Carlo event generators PYTHIA 8 and HERWIG 7, the DELPHES fast-detector simulation framework, and the full analysis pipeline from event generation to neural network input.

Chapter 4 presents the ATLAS Run 2 analysis. After a brief review of the most recent ATLAS $H \rightarrow Za$ search that provides context and motivation, the chapter is organized into three main parts. The first part presents the kinematic validation of the independent simulation pipeline, confirming that PYTHIA 8 and HERWIG 7 reproduce consistent distributions for reconstructed leptons, jets, and the three-body invariant mass $m_{\ell\ell j}$ when seeded from the same POWHEG hard-process input. The second part presents a comprehensive generator comparison of jet substructure observables, including a hadronization study (HAD-on vs. HAD-off) that isolates the nonperturbative contribution to each variable, and a **SoftDrop** grooming analysis that quantifies the degree to which grooming can reduce the inter-generator spread. The third part presents the two-stage neural network pipeline: the regression network for pseudoscalar mass prediction and the classification network for signal-versus-background separation. Three sequential studies are presented. Study A establishes a domain-transfer baseline and identifies the preprocessing inconsistencies that prevent a physically interpretable evaluation. Study B corrects these inconsistencies and produces the first clean measurement of classifier generalization across generators, establishing the collapse of the ungroomed classifier and the stable generalization of the classifier exploiting groomed variables. Study C investigates whether this central result is sensitive to the near-mean-collapse of the regression network identified in Study B, by introducing a balanced class sampler and a per-class normalized loss into the regression training; it confirms that the groomed generator gap is stable under this change.

Chapter 5 collects the conclusions of the work and an outlook for future improvements. The central finding — that **SoftDrop** grooming is a necessary preprocessing step for classifier generalization, and that the groomed generator gap is substantially smaller than the corresponding ungroomed collapse — is placed in the context of the ATLAS analysis and its $\sim 60\%$ theory systematic. The outlook identifies four concrete next steps ordered by expected impact, from adding kinematically motivated features at the hard splitting to applying the pipeline to fully reconstructed simulation.



Chapter 1

The Higgs Boson in the Standard Model

*“Have you ever heard about the Higgs Boson
Blues?”*

*I’m going down to Geneva, baby,
Gonna teach it to you.”*

— Nick Cave and the Bad Seeds, *“Higgs Boson
Blues”*



GOAL of this Chapter is giving an introduction to the fundamental role that the Higgs boson covers in the framework of the Standard Model of particle physics — this is done through a semi-thorough overview of the Standard Model as a gauge theory, and through choosing the electroweak model as a case study. The final goal is to paint a complete picture of why a Higgs boson is pivotal to both the theoretical framework and the experimental realm of physics.

1.1 Introduction

Talking about the Higgs boson is quite a Herculean task, since this particle represents the coronation of a lifelong dream of multiple generations of physicists who had been on its trail, dreaming of a consistent explanation for why most of the universe’s particles possess

mass. It is therefore a huge undertaking to try to put together a consistent and cohesive introduction to present this very special and quirky particle. To understand why the Higgs boson represents the cornerstone of the Standard Model one must, first and foremost, introduce what the Standard Model is and how it works as a gauge theory. It is the goal of this Chapter to now guide the reader step by step towards an understanding of why the Higgs boson is so important for particle physicists, with no pretense of completeness or extreme rigor.

The Standard Model is a gauge-symmetric quantum field theory based on the symmetry group:

$$SU(3)_C \otimes SU(2)_L \otimes U(1)_Y.$$

This denomination, which might sound very generic as well as very confusing, encodes precisely everything that matters now for all intents and purposes about this theory. We will therefore start by breaking down each of the components of its definition. One refers to a quantum field theory (QFT) as a theory that respects a few basic principles:

- **Everything is a field:** fields constitute the basic building blocks of reality. Everything is determined by the interaction of fields and their propagation in spacetime.
- **Special relativity principles hold:** these fields have to satisfy special relativity conditions, mainly that particles cannot travel faster than the speed of light, and that there is no preferred frame of reference. A special consequence of these principles lies in the mathematical formulation of the theory, which must therefore be Lorentz invariant.
- **Quantum mechanics is at the core:** quantum mechanics is built inside the theory, quantum superposition and quantum behaviors are encoded in the formalism, as well as in the theory's locality.

To effectively build a theory satisfying these conditions one must adopt the Lagrangian formalism. Lagrangian densities are Lorentz scalars, and are used to describe fields and their interactions with one another. Depending on the kind of field one would like to describe one uses a different Lagrangian — we will list each useful Lagrangian in the following:

- **Real Scalar Field Lagrangian:** a spin-0 real field is described by

$$\mathcal{L} = \frac{1}{2} \partial_\mu \phi \partial^\mu \phi - \frac{m^2}{2} \phi^2;$$

- **Complex Scalar Field Lagrangian:** a spin-0 complex field is described by

$$\mathcal{L} = \partial_\mu \phi^* \partial^\mu \phi - m^2 |\phi|^2;$$

- **Spinor Field Lagrangian (Dirac Lagrangian):** a spin- $\frac{1}{2}$ field is described by

$$\mathcal{L} = \bar{\psi} (i \not{\partial} - m) \psi;$$

- **Vector Field Lagrangian:** a spin-1 field is described by

$$\mathcal{L} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + \frac{m^2}{2}A_\mu A^\mu,$$

where $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$;

- **Yukawa Lagrangian:** this special Lagrangian describes interactions between spin- $\frac{1}{2}$ fields and scalar fields, such that:

$$\mathcal{L} = g\bar{\psi}\phi\psi,$$

where g is the coupling constant.

Surprisingly enough, these are the only Lagrangians that are required to build the Standard Model. But what does one mean when one mentions that the Standard Model is gauge-symmetric?

Through a more classical approach to the mathematical development of Lagrangian and Hamiltonian formalism one finds that if a Lagrangian possesses a symmetry, *i.e.* it is invariant under a transformation of the fields, there exists a conserved current such that its four-divergence is trivial ($\partial_\mu j^\mu = 0$). If one defines a transformation $U(\alpha)$ dependent on a constant parameter α that transforms the fields through $\phi' = U(\alpha)\phi$, Noether's theorem predicts that the expression for the conserved current should be:

$$j_\mu = \sum_{\text{fields}} \frac{\partial \mathcal{L}}{\partial^\mu \phi} \delta \phi,$$

where the sum is performed over the fields contained in the Lagrangian, and $\delta \phi$ refers to the infinitesimal change of the field operated by $U(\alpha)$, since α is considered to be a small parameter:

$$\delta \phi = \phi' - \phi.$$

If we consider a worked example we can see how this theorem works in practice. For the Dirac Lagrangian we have that $U(\alpha) = e^{i\alpha}$. Therefore:

$$\begin{aligned}\psi' &= U(\alpha)\psi = e^{+i\alpha}\psi \\ \bar{\psi}' &= U^{-1}(\alpha)\bar{\psi} = e^{-i\alpha}\bar{\psi}.\end{aligned}$$

As for the infinitesimal transformations:

$$\begin{aligned}\delta \psi &= \psi' - \psi = (e^{+i\alpha} - 1)\psi \approx +i\alpha\psi \\ \delta \bar{\psi} &= \bar{\psi}' - \bar{\psi} = (e^{-i\alpha} - 1)\bar{\psi} \approx -i\alpha\bar{\psi}.\end{aligned}$$

One can verify that under this transformation the Dirac Lagrangian is invariant, since the $U(\alpha)$ operator commutes with the derivative operator: therefore $\mathcal{L}' = \mathcal{L}$, and Noether's

theorem holds. The Dirac conserved current is then:

$$j^\mu = \frac{\partial \mathcal{L}}{\partial_\mu \psi} \delta \psi + \frac{\partial \mathcal{L}}{\partial_\mu \bar{\psi}} \delta \bar{\psi} = \alpha \bar{\psi} \gamma^\mu \psi.$$

We can prove that its four-divergence is trivial using the Dirac equation.

$$\partial_\mu j^\mu = \alpha \partial_\mu (\bar{\psi} \gamma^\mu \psi) = \alpha (\partial_\mu \bar{\psi} \gamma^\mu \psi + \bar{\psi} \gamma^\mu \partial_\mu \psi) = \alpha (+m \bar{\psi} \psi - m \bar{\psi} \psi) = 0.$$

Returning to our previous question, we can look at what it means for theory to be gauge-symmetric. A gauge transformation is a transformation such as the ones from earlier: we have looked at a special example of a global gauge transformation, *i.e.* a transformation whose parameter does not depend on spacetime coordinates. The Standard Model is a theory that is built on local gauge transformations, ones where the parameters change locally. The issue is that local gauge transformations do not leave Lagrangians invariant as they are, since by definition a derivative term must act on the transformation, yielding extra terms that one cannot get rid of easily.

One then employs the **gauge principle**, a special "recipe" that allows to make Lagrangians invariant under local transformations. What one does is look at the specific local transformation in question, and according to the dimensions of the parameter, also referred to as the generator(s) of the transformation, and add a number of so-called gauge boson fields accordingly. These fields appear in the so-called covariant derivative, whose goal is to be covariant under the transformation. Given a Lagrangian $\mathcal{L}$ invariant under a local transformation $\mathcal{U}(x)$, one can verify that if the introduced covariant derivative D_μ satisfies:

$$(D_\mu \phi)' = D'_\mu \phi' = \mathcal{U}(D_\mu \phi),$$

From this relation one also retrieves what one calls gauge transformations for the gauge bosons. In the case of quantum electrodynamics, by adding to a $U(1)$ local transformation a gauge field A_μ one can retrieve the same formalism as the classically derived quantum electrodynamics¹. Analogously, one can then extend this principle to more complex symmetry transformations. The only piece that would be missing would be a kinetic term for the gauge bosons, which constitutes the last step of the recipe — this is added via the calculation of the following quantity, called a Wilson loop:

$$[D_\mu, D_\nu] f(x).$$

From this one can obtain a gauge field tensor that allows for the construction of a Lorentz-invariant kinetic term.

This procedure is deeply well-justified mathematically, and its explanation is beyond the scope of this Chapter.

¹The gauge principle derivation for quantum electrodynamics is much more general than the classical counterpart since it makes no assumptions around the theory and yields the same results from more general principles.

It is useful to cover a worked example, in order to further cement our understanding. We will consider an $SU(2)_L$ symmetry, whose generators are three 2x2 matrices, the Pauli matrices. A transformation of this kind is applied in the context of the Standard Model to yield a working formalism for electroweak theory, a parity-violating theory. Due to the special properties of this theory, it is important to consider massless fields in a special representation — eigenstates of the chirality operator. A spinor field is then rewritten as:

$$\psi = \begin{pmatrix} \chi_R \\ \chi_L \end{pmatrix},$$

since chirality is a well-defined eigenstate only when the field is massless: a massless particle is such that there exists no frame of reference in which the chirality number could change. We introduce a left-handed field doublet $L = \begin{pmatrix} \nu_L \\ e_L \end{pmatrix}$ whose Lagrangian is:

$$\mathcal{L}_{SU(2)_L} = i\bar{L}\not{\partial}L = i\bar{\nu}_L\not{\partial}\nu_L + i\bar{e}_L\not{\partial}e_L$$

This special doublet is a doublet in an isospin abstract space, whose associated quantum numbers are:

$$\hat{T}\nu_L = +\frac{1}{2}\nu_L \quad \hat{T}e_L = -\frac{1}{2}e_L,$$

and $\hat{T}$ is the isospin operator. We can verify that the Lagrangian is symmetric under a transformation $U(\vec{\alpha})$, where now $\vec{\alpha}$ is a three-dimensional vector of parameters. The introduction of an extra algebraic complexity to this Lagrangian implies that the generators of this symmetry are noncommuting generators. These T generators are such that $T_i = \frac{\sigma_i}{2}$. We can prove that the $SU(2)_L$ symmetry Lagrangian is symmetric under a symmetry of this kind:

$$L' = e^{i\vec{\alpha}\cdot\vec{T}}L,$$

we can look at the transformation for the Hermitian conjugate part of the field:

$$\bar{L}' = \bar{L}e^{-i\vec{\alpha}\cdot\vec{T}}$$

since the Lagrangian is invariant on sight under this transformation, we can move on to the application of the gauge principle. We therefore introduce the covariant derivative for the $SU(2)_L$ Lagrangian:

$$\mathcal{L}_{SU(2)_L} = i\bar{L}\not{D}L \quad D_\mu = \partial_\mu + ig\vec{T} \cdot \vec{W}_\mu.$$

From this we are able to derive the transformation rules for the three gauge fields, where $a = 1, 2, 3$:

$$(D_\mu L)' = D'_\mu L' = D_\mu L'$$

$$W_\mu'^a = W_\mu^a + \frac{1}{g}\partial_\mu \alpha^a - \epsilon^{abc}\alpha^b W_\mu^c,$$

where we recognize the same structure as the previous $U(1)$ symmetry as well as a tensorial component: that is the non-Abelian contribution to the transformation. One can then proceed to build the gauge-invariant field tensor for W_μ^a :

$$F_{\mu\nu}^a = \partial_\mu W_\nu^a - \partial_\nu W_\mu^a + g\epsilon^{abc}W_\mu^b W_\nu^c.$$

It is now possible to build the gauge invariant Lagrangian:

$$\mathcal{L}_{SU(2)_L} = i\bar{L}\not{D}L - \frac{1}{4}F_{\mu\nu}^a F_a^{\mu\nu}.$$

A few remarks about this new Lagrangian are in order:

- This procedure has yielded a gauge invariant theory.
- The gauge principle has modified the phenomenology of this theory, introducing three new gauge fields W_μ^a that interact with the matter fields, and this interaction is encoded inside the covariant derivative.
- Due to the non-Abelian nature of the symmetry we have found that the gauge fields can interact with each other.
- All of the theory is at this point massless, since mass terms are not included in this Lagrangian.

We have been able to lay the foundations in which the Higgs boson operates. One needs a procedure to add field mass terms that does not disrupt gauge symmetry. Adding them by hand is impossible, since a term of the $m^2 W_\mu^a W_a^\mu$ would disrupt gauge symmetry. We can prove:

$$m^2 W_\mu^a W_a^\mu = m^2 (W_\mu^a + \frac{1}{g}\partial_\mu \alpha^a - \epsilon^{abc}\alpha^b W_\mu^c)(W_a^\mu + \frac{1}{g}\partial^\mu \alpha_a - \epsilon_{ade}\alpha_d W_e^\mu),$$

this term evidently generates terms that disrupt the gauge symmetry.

We are now left with a dilemma: we can build a successfully interacting gauge theory with the $SU(3)_C \otimes SU(2)_L \otimes U(1)_Y$ group that is able to describe real fundamental interactions, but the whole model must be massless to keep gauge invariance. This is a difficult conundrum since real particles (both matter fields and gauge fields except the neutrino, the photon, and the gluon) are indeed massive, and they are also observed to be

massive experimentally, so that implies that we require something else. We either choose to scrap this theory, or find a new piece to this mass puzzle. The missing piece turns out to be the Higgs boson and its associated symmetry breaking.

Section Roadmap:**Covered:**

Introduced the Standard Model as a gauge theory built on quantum field principles and local symmetries. Explained how gauge invariance leads to the introduction of force carriers and set up the foundational challenge of giving mass to particles while preserving gauge symmetry.

Coming Up:

Application of gauge symmetries to the electroweak interaction via $SU(2)_L \otimes U(1)_Y$, and the structure of its field content.

References: [7]; [8]; [9].



1.2 Electroweak Theory: Before the Higgs Boson

We now briefly pause the development of the general framework for quantum field theories and the Higgs boson to focus on one of the biggest successes of the Standard Model: Electroweak theory. It all started with Fermi and his theory for the β^- decay. The process turns a neutron into a proton, while also producing an electron and antineutrino as side products. Fermi studied this process from a contact-interaction standpoint, since it seemed to be a process that happened over very short distances.

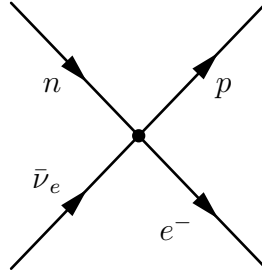


FIGURE 1.1: A contact interaction for β^- decay.

This is a process that depends on the coupling constant named after Fermi, $\frac{G_F}{\sqrt{2}}$, as seen in the Feynman diagram of Figure (1.1). Upon further developments of quantum field theories, it was seen that Fermi theory led to renormalization issues due to the dimensions of the coupling constants, and a bunch of other problems, such as violation of unitarity. All in all, Fermi theory was impracticable for the β decay. This issue was solved by increasing the energy of the process. We can somehow think of zooming in into the process and seeing that the β^- decay is not a contact interaction, but there is an exchange of a quite heavy particle — the heavy mass of the mediator accounted for the very short interaction range. We can assume the existence of a W particle that decays into two leptons. This heavy W particle exchange can happen for different processes in particle physics, as an example:

$$\begin{aligned}\mu^- &\longrightarrow e^- \bar{\nu}_e \nu_\mu \\ \mu^+ &\longrightarrow e^+ \bar{\nu}_\mu \nu_e\end{aligned}$$

It is important to note that these processes all involve an exchange of a charged particle, so the W particle must come in two different charges, W^+ , W^- . These processes are represented in Figure (1.2).

From the theory we can calculate half-lives for processes, and by using this model for weak interactions it is possible to calculate values that perfectly fit experimental data: we have been able to elaborate a theory that is predictive, all done via the application of the gauge principle.

It is now time to try to bridge together what we have obtained from gauge theories with the physics of electroweak interactions. This theory is based on the $SU(2)_L \otimes U(1)_Y$ symmetry group, and it is built with the idea of unifying quantum electrodynamics and weak interactions. If we start from the hypothesis of having massless particles, we can



FIGURE 1.2: Comparison of weak interactions mediated by W^+ and W^- bosons.

classify the electron and its neutrino as a field doublet, using a new quantum number called weak isospin. Note that this formalism can be extended to more than one lepton family, as will be presented in the following Sections. This then becomes:

$$L = \begin{pmatrix} \nu_{eL} \\ e_L \end{pmatrix},$$

in the following we will omit the e subscript for the neutrino field. Since these fields are massless, we will consider the Weyl Lagrangian:

$$\mathcal{L} = \bar{L} i \not{\partial} L = i \bar{\nu}_L \not{\partial} \nu_L + i \bar{e}_L \not{\partial} e_L$$

If we now consider the weak isospin and since it is conserved in weak interactions, it describes a symmetry for the Lagrangian. By defining the isospin operator $T = \frac{\vec{\sigma}}{2}$, we can write the transformations for the field doublets:

$$\begin{aligned} L &\longrightarrow L' = \exp\left(+i\vec{\alpha} \cdot \vec{T}\right) L \\ \bar{L} &\longrightarrow \bar{L}' = \bar{L} \exp\left(-i\vec{\alpha} \cdot \vec{T}\right). \end{aligned}$$

We can then write a Noether current for this $SU(2)_L$ transformation:

$$j_{SU(2)_L, i}^\mu = \bar{L} \gamma^\mu \frac{\sigma_i}{2} L,$$

and by of the gauge principle we covered earlier, we can assume that we need to introduce three fields, W_i :

$$D_\mu = \partial_\mu - ig \frac{\sigma_i}{2} W_{\mu i}.$$

Due to the definition of the σ_i matrices, we have that for $i = 1, 2$ the terms of these fields are off diagonal, while the $i = 3$ term is diagonal. We can write this identity easily:

$$\frac{1}{2} [\sigma_1 W_\mu^1 + \sigma_2 W_\mu^2] = \frac{1}{2} [\sigma_1 W_\mu^1 \pm i\sigma_2 (\mp i W_\mu^2)]$$

This implies that there must exist a combination of Pauli matrices such that:

$$W_\mu^\pm \propto W_\mu^1 \mp i W_\mu^2,$$

and this yields two fields that are complex in nature, representing charged particles. The combination of Pauli matrices such that this holds is:

$$\sigma_\pm := \sigma_1 \pm i\sigma_2,$$

and their explicit expressions are:

$$\sigma_+ = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad \sigma_- = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}.$$

By defining the gauge fields in this way we can see how the Noether current couples to these new fields $W_\mu^\pm$:

$$\begin{aligned} \bar{L}\gamma^\mu\sigma_+L &= (\bar{\nu}_L \quad \bar{e}_L) \gamma^\mu\sigma_+ \begin{pmatrix} \nu_L \\ e_L \end{pmatrix} = (\bar{\nu}_L \quad \bar{e}_L) \gamma^\mu \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \nu_L \\ e_L \end{pmatrix} = \bar{e}_L\gamma^\mu\nu_L \\ \bar{L}\gamma^\mu\sigma_-L &= (\bar{\nu}_L \quad \bar{e}_L) \gamma^\mu\sigma_- \begin{pmatrix} \nu_L \\ e_L \end{pmatrix} = (\bar{\nu}_L \quad \bar{e}_L) \gamma^\mu \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \nu_L \\ e_L \end{pmatrix} = \bar{\nu}_L\gamma^\mu e_L \end{aligned}$$

By introducing the fields, we have two interaction terms that directly relate to the interaction vertices below:

$$\begin{aligned} \mathcal{L}_{int} &= \bar{e}_L\gamma^\mu\nu_L W_\mu^+ \\ \mathcal{L}_{int} &= \bar{\nu}_L\gamma^\mu e_L W_\mu^- \end{aligned}$$

These interaction terms are of the charged-current type since the field particle that is exchanged is charged, and they constitute great pieces of formalism to interpret β_- decay, or the decay of muons. These interaction vertices are the object of Figure (1.3a) and (1.3b).

These terms have a weak coupling constant $\frac{g}{\sqrt{2}}$, but what happens to other component, the one diagonal in the Pauli matrices? We shall calculate its contribution to the Noether current:

$$\bar{L}\gamma^\mu\sigma_3L = (\bar{\nu}_L \quad \bar{e}_L) \gamma^\mu \begin{pmatrix} +1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} \nu_L \\ e_L \end{pmatrix} = \bar{\nu}_L\gamma^\mu\nu_L - \bar{e}_L\gamma^\mu e_L,$$

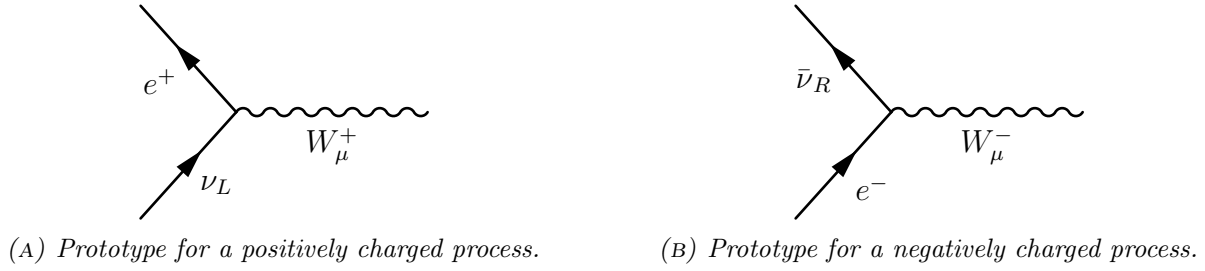


FIGURE 1.3: Charged-current interaction vertices.

and it is easy to see that this couples to the W_μ^3 component of the gauge field with a different coupling, but always dependent on g . The nice thing about this part of the theory is that the charged current components are physically accurate. The problem with this formalism is that it is not able to describe neutral current interactions accurately — if we were to identify the W_μ^3 field with a photon field, we see that the interaction term would not be consistent with reality, since photons do not interact with neutrinos. We will need to unify the electromagnetic and weak interactions to be able to yield terms that are consistent with QED.

We are now ready to unify electroweak interactions in order to obtain a working formalism for this special kind of interaction. We do this by considering part of the symmetry group we started with:

$$SU(2)_L \otimes U(1)_Y.$$

We will introduce this new formalism piece by piece — firstly we look at the global symmetries of the $U(1)$ type, and then we will join them together through two symmetry groups, which will yield two new currents that will enter the interaction Lagrangians. We will be able to do this by considering the right component of the electron field as a singlet under $SU(2)_L$. This amounts to considering:

$$L = \begin{pmatrix} \nu_L \\ e_L \end{pmatrix} \quad , \quad e_R$$

If we now take a look at the weak isospin quantum number we see that we can define it to be:

$$\begin{aligned} T_3(\nu_L) &= +\frac{1}{2} \\ T_3(e_L) &= -\frac{1}{2} \\ T_3(e_R) &= 0, \end{aligned}$$

this implies that under a transformation of the $SU(2)_L$ type, we have that only the doublet changes.

We are then allowed to write out a new Lagrangian term that includes the kinetic propagation of e_R :

$$\mathcal{L} = \bar{L}i\not{\partial}L + \bar{e}_R i\not{\partial}e_R,$$

and this expression is still globally gauge invariant by inspection. We will now analyze all the fields of the theory, namely ν_L , e_R , e_L . We are sticking to one lepton family for simplicity — the same treatment could be given to the other leptonic families. We will group these states in another superfield ψ such that:

$$\psi = \begin{pmatrix} \nu_L \\ e_L \\ e_R \end{pmatrix} \quad \psi'_i = e^{i\alpha} \psi_i,$$

where now all states are rotated under a global $U(1)$ symmetry. The Lagrangian is trivially symmetric under this transformation. Thanks to the Noether theorem, we have that there are different conserved currents. We can write these currents out:

$$j^\mu = \sum_i \bar{\psi}_i \gamma^\mu \psi_i = \bar{\nu}_L \gamma^\mu \nu_L + \bar{e}_L \gamma^\mu e_L + \bar{e}_R \gamma^\mu e_R$$

This is a special current term that cannot directly be coupled to fields in the Standard Model: we say that this kind of global $U(1)$ symmetry is an accidental symmetry, and it is related to the conservation of lepton number.

In order to now build an electroweak theory, we introduce a special relation that is able to join two different symmetry operators. This relation is the following:

$$Q = T_3 + \frac{Y}{2},$$

where Y is a new quantum operator, called the hypercharge operator. We can rewrite this relation in terms of the operators we already know, obtaining:

$$Y = 2(Q - T_3).$$

This is indeed very useful because this encodes both conservation of weak isospin and charge, and hence even hypercharge. We now would like to analyze the quantum numbers of the states we are considering. Taking a look at the doublet, we see:

$$\begin{array}{llll} \nu_L \longrightarrow & Q(\nu_L) = 0 & T_3(\nu_L) = +\frac{1}{2} & \implies Y(\nu_L) = -1 \\ e_L \longrightarrow & Q(e_L) = -1 & T_3(e_L) = -\frac{1}{2} & \implies Y(e_L) = -1, \end{array}$$

we then notice that the hypercharge operator on the doublet looks like a diagonal matrix, proportional to the identity:

$$Y(L)L = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} L$$

We also need to consider the hypercharge for the isospin singlet:

$$e_R \longrightarrow \quad Q(e_R) = -1 \quad T_3(e_R) = 0 \quad \implies \quad Y(e_R) = -2$$

We could also work backwards to reconstruct the fact that the charge operator on the field doublet looks like this:

$$Q(L)L = \left(T_3 + \frac{Y}{2}\right)L = \frac{1}{2}(\tau_3 + Y)L = \begin{pmatrix} 0 & 0 \\ 0 & -1 \end{pmatrix} L$$

We briefly mention the fact that we could have added another singlet to this theory, namely the right-handed neutrino, to fulfill some kind of innate symmetry requirement, but this would have had zero quantum numbers all over: no charge, no weak isospin, and hence no hypercharge. This kind of particle is then said to be sterile, since it does not interact by means of any of the interactions².

All of this hence begs the question: is there a symmetry associated to this special hypercharge operator? We define the field transformations as:

$$\begin{aligned} L' &= \exp(i\alpha Y(L)) L \\ e'_R &= \exp(i\alpha Y(L)) e_R, \end{aligned}$$

where we take particular care in noticing that we have defined a hybrid transformation, which uses the same transformations parameters, due to the fact that it is a unified transformation that deals with the doublet and singlet in different ways. The doublet transformation makes use of 2×2 matrix operators, while the singlet uses a scalar operator. A transformation like this makes the starting Lagrangian invariant due to the diagonal nature of the hypercharge operator. This yields a Noether conserved current that looks like this:

$$j^\mu = \bar{L}\gamma^\mu Y(L)L + \bar{e}_R\gamma^\mu Y(e_R)e_R.$$

This is a hypercharge current, which is more general than the electromagnetic current, but in a way it contains it. We have started from an $SU(2)_L$ symmetry group, and by adding a $U(1)_Y$ symmetry, we have obtained a new symmetry group: $SU(2)_L \otimes U(1)_Y$, which is a symmetry group that contains the electromagnetic symmetry group $U(1)_{em}$. We can verify this fact by applying the gauge protocol to this joint symmetry group. We now start by doing some counting: the $SU(2)_L \otimes U(1)_Y$ group requires three gauge

²The sterile neutrino has been under studies as a probable cause for lots of different inconsistencies in experimental results.

bosons and an extra one, coming from the $U(1)_Y$ group. We consider two different couplings for this theory, g and g' . We saw earlier that applying the gauge protocol to the $SU(2)_L$ yielded a theory with three different fields that could be combined into two charged fields, and one neutral one: this produced two terms for the interaction Lagrangian, a neutral current interaction, and a charged current interaction. We will now extend this procedure to the composite symmetry group. In building the new gauge protocol for this theory we can use the notion of covariant derivative taking care of adding another field B^μ , with a different gauge coupling g' , different from the gauge coupling constant g used for the W_μ^i fields. Since the Lagrangian must be symmetric under the simultaneous application of the two symmetry transformations, we will have to keep track of the corresponding generators and the gauge fields. We give the expression of the full covariant derivative:

$$D^\mu = \partial^\mu - ig \frac{\sigma_i}{2} W_\mu^i - ig' \frac{Y}{2} B^\mu.$$

The gauge protocol imposes that the covariant derivative must transform as the field:

$$\begin{aligned} (D_L^\mu L)' &= U_Y(\beta) U_L(\alpha) (D_L^\mu L) \\ (D_R^\mu e_R)' &= U_Y(\beta) U_L(\alpha) (D_R^\mu e_R), \end{aligned}$$

we must not forget to add the kinetic terms of the gauge field we have added. Then the Lagrangian density of this gauge theory becomes:

$$\mathcal{L} = \bar{L} i \not{D} L + \bar{e}_R i \not{D} e_R - \frac{1}{4} W_{\mu\nu}^i W_i^{\mu\nu} - \frac{1}{4} B_{\mu\nu} B^{\mu\nu},$$

where the kinetic parts for the gauge bosons have been normalized to the standard convention³. Since the gauge symmetry holds, we have that the different interaction terms that arise are:

$$\mathcal{L}_{int} = \mathcal{L}_{NC} + \mathcal{L}_{CC},$$

where the charged current interaction is the same as the one for the pure $SU(2)_L$ group, with the $W_\mu^\pm$ fields. The term that we need to work on is the neutral current interaction, since it is the one that needs to be reworked to be able to include the electromagnetic interaction via A_μ and the interaction with Z_μ . We start from:

$$\mathcal{L}_{NC}^{EW} = \frac{g}{2} (\bar{\nu}_L \gamma^\mu \nu_L + \bar{e}_L \gamma^\mu e_L) W_\mu^3 + \frac{g'}{2} \bar{L} \gamma^\mu Y(L) L B_\mu + \frac{g'}{2} \bar{e}_R \gamma^\mu Y(e_R) e_R B_\mu,$$

where we can distinguish between two different kinds of contributions: the one with the g coupling constant is generated by the non-Abelian group $SU(2)_L$, while the terms with the g' coupling constant is generated by the Abelian group $U(1)_Y$. We will now use the

³The standard convention is such that the normalization for the field tensor for the electromagnetic field yields the right expressions for energy densities, forcing the constant to be $\frac{1}{4}$.

matricial representation for the interaction Lagrangian by using the superfield notation ψ used in the previous sections:

$$\mathcal{L}_{NC}^{EW} = \bar{\psi} \gamma_\mu \left[g T_3 W_3^\mu + g' \frac{Y(\psi)}{2} B^\mu \right] \psi,$$

where a generalization for the matricial form of T_3 , $Y(\psi)$ has been implicitly defined as the following:

$$T_3 = \begin{pmatrix} +\frac{1}{2} & 0 & 0 \\ 0 & -\frac{1}{2} & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad Y(\psi) = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -2 \end{pmatrix}.$$

It is easy to verify that this representation is equivalent to the doublet and singlet representation.

We can now press on and see for ourselves how we can yield a purely electromagnetic from this neutral current term. We will work under the hypothesis that B_μ and W_μ^3 are not physical fields, but physical fields could be a linear combination of gauge fields, since nothing prevents us from combining these two in some way to yield two new real fields, which we will call Z_μ and A_μ . These are spin-1 neutral fields which will be perfect for describing neutral electroweak interactions. We will use an orthogonal rotation matrix to define the fields:

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

We are then able to invert the orthogonal matrix O by transposition, since $O^{-1} = O^T$.

$$\begin{pmatrix} B_\mu \\ W_\mu^3 \end{pmatrix} = \begin{pmatrix} \cos \theta_W & -\sin \theta_W \\ +\sin \theta_W & \cos \theta_W \end{pmatrix} \begin{pmatrix} A_\mu \\ Z_\mu \end{pmatrix}.$$

This yields definitions for gauge fields in terms of physical fields:

$$\begin{aligned} B_\mu &= A_\mu \cos \theta_W - Z_\mu \sin \theta_W \\ W_\mu^3 &= A_\mu \sin \theta_W + Z_\mu \cos \theta_W, \end{aligned}$$

which we can then substitute inside the interaction Lagrangian:

$$\begin{aligned} \mathcal{L}_{NC}^{EW} &= \bar{\psi} \gamma^\mu \left[g \sin \theta_W T_3 + g' \cos \theta_W \frac{Y}{2} \right] \psi A_\mu \\ &\quad + \bar{\psi} \gamma^\mu \left[g \cos \theta_W T_3 - g' \sin \theta_W \frac{Y}{2} \right] \psi Z_\mu. \end{aligned}$$

Our goal is to now retrieve the electromagnetic interaction Lagrangian, this implies that $\mathcal{L}_{NC} = \mathcal{L}_{QED} + \mathcal{L}_{NWC}$, and since we know the QED interaction Lagrangian to be $\mathcal{L}_{QED} = e \bar{\psi} \gamma^\mu Q \psi A_\mu$, we must then extend the notion of Q by imposing the equality between the

coefficients in the Lagrangian obtained earlier. This yields:

$$eQ = g \sin \theta_W T_3 + g' \cos \theta_W \frac{Y}{2}.$$

Using the matricial representation of the operators we can rewrite:

$$e \begin{pmatrix} 0 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix} = g \sin \theta_W \begin{pmatrix} +\frac{1}{2} & 0 & 0 \\ 0 & -\frac{1}{2} & 0 \\ 0 & 0 & 0 \end{pmatrix} + \frac{g'}{2} \cos \theta_W \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -2 \end{pmatrix}.$$

This representation helps us in writing the electroweak unification relation. If we consider the states, each one corresponding to one of the rows, we can conclude:

$$\begin{aligned} 0 &= \frac{g}{2} \sin \theta_W - \frac{g'}{2} \cos \theta_W \\ -e &= -\frac{g}{2} \sin \theta_W - \frac{g'}{2} \cos \theta_W \\ -e &= -g' \cos \theta_W, \end{aligned}$$

where the third equation is a linear combination of the other two, so it will not be taken into consideration. Solving this system of equations we can write the electroweak unification relation, which relates the couplings of electromagnetic theory to the one of gauge theory: this is an extremely important part of the Standard Model. This implies that electric charge is derived from the gauge coupling of electroweak theories through a mixing angle θ_W , the Weinberg angle, also known as weak angle:

$$\boxed{g \sin \theta_W = e = g' \cos \theta_W}.$$

Now is a good moment to sit back and take in our result: we have proved that the electromagnetic and weak interactions are linked together by this relation, and looking back we also have predicted the existence of a new field with which there can be an interaction vertex similar to QED, but with a different coupling constant that is measurable at collider machines. Its existence was proven at LEP, which carried out precision measurements of the Z^0 particle. It is now our goal to continue and isolate the Z_μ field couplings. Starting from the neutral current weak interaction Lagrangian,

$$\mathcal{L}_{NC}^W = \bar{\psi} \gamma^\mu \left[g \cos \theta_W T_3 - \frac{Y}{2} g' \sin \theta_W \right] \psi Z_\mu,$$

and by using the expression of Y , $\frac{Y}{2} = Q - T_3$, we can write this as:

$$\begin{aligned}\mathcal{L}_{NC}^W &= \bar{\psi}\gamma^\mu [g \cos \theta_W T_3 - g' Q \sin \theta_W + T_3 g' \sin \theta_W] \psi Z_\mu \\ &= \bar{\psi} [g \cos \theta_W + g' \sin \theta_W] T_3 \psi Z_\mu - g' \sin \theta_W \bar{\psi}\gamma^\mu Q \psi Z_\mu = \\ &= gg' \bar{\psi}\gamma^\mu \left[\left(\frac{\cos \theta_W}{g'} + \frac{\sin \theta_W}{g} \right) T_3 \right] \psi Z_\mu - g' \sin \theta_W \bar{\psi}\gamma^\mu Q \psi Z_\mu,\end{aligned}$$

up until now this has been mere algebraic manipulation. If we now make use of the electroweak unification condition:

$$g = \frac{e}{\sin \theta_W} \quad g' = \frac{e}{\cos \theta_W},$$

we can rewrite:

$$\begin{aligned}\mathcal{L}_{NC}^W &= \frac{e^2}{\sin \theta_W \cos \theta_W} \bar{\psi}\gamma^\mu \left[\left(\frac{\cos^2 \theta_W}{e} + \frac{\sin^2 \theta_W}{e} \right) T_3 \right] \psi Z_\mu - g' \sin \theta_W \bar{\psi}\gamma^\mu Q \psi Z_\mu \\ &= \frac{e}{\sin \theta_W \cos \theta_W} \bar{\psi}\gamma^\mu T_3 \psi Z_\mu - g' \sin \theta_W \bar{\psi}\gamma^\mu Q \psi Z_\mu.\end{aligned}$$

By now using the electroweak unification condition on the second term in Q , we can write:

$$\begin{aligned}\mathcal{L}_{NC}^W &= \frac{e}{\sin \theta_W \cos \theta_W} \bar{\psi}\gamma^\mu T_3 \psi Z_\mu - \frac{e}{\cos \theta_W \sin \theta_W} \bar{\psi}\gamma^\mu Q \sin^2 \theta_W \psi Z_\mu = \\ &= \bar{\psi}\gamma^\mu \left[\frac{e}{\sin \theta_W \cos \theta_W} (T_3 - Q \sin^2 \theta_W) \right] \psi Z_\mu.\end{aligned}$$

We can define a new quantity, called Q_z such that:

$$\begin{aligned}\mathcal{L}_{NC}^W &= \bar{\psi}\gamma^\mu Q_z \psi Z_\mu \\ Q_z &\doteq \frac{e}{\sin \theta_W \cos \theta_W} [T_3 - Q \sin^2 \theta_W].\end{aligned}$$

We note that this is a diagonal interaction term since it is the sum of diagonal terms, and it is also interesting to point out that the Z_μ field couples differently to the fields we are considering depending on their charge. If we now examine case by case, we can see two different situations.

If we take our ψ field to be ν_L , we can see that:

$$Q(\nu_L) = 0 \quad T_3(\nu_L) = \frac{1}{2},$$

hence the interaction term has a simple expression:

$$\mathcal{L}_{NC}^W = \frac{e}{\sin \theta_W \cos \theta_W} \bar{\nu}_L \gamma^\mu \nu_L Z_\mu,$$

with associated neutral weak current vertex as in Figure (1.4).

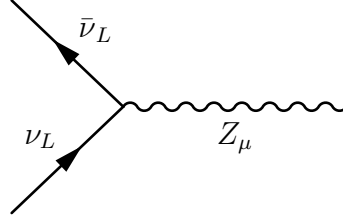


FIGURE 1.4: Neutral current vertex for ν_L .

If we now consider our field to be equal to e_L, e_R , we have:

$$T_3(e_L) = -\frac{1}{2} \quad T_3(e_R) = 0,$$

which implies that we can write the interaction Lagrangian as:

$$\begin{aligned} \mathcal{L}_{NC}^W &= \frac{e}{\sin \theta_W \cos \theta_W} (\bar{e}_L \quad \bar{e}_R) \gamma^\mu Q_z \begin{pmatrix} e_L \\ e_R \end{pmatrix} Z_\mu = \\ &= (\bar{e}_L \gamma^\mu Q_{zL} e_L + \bar{e}_R \gamma^\mu Q_{zR} e_R) Z_\mu, \end{aligned}$$

due to the diagonal nature of Q_z for the fields we are considering, there exist two components for this matrix, Q_{zL} and Q_{zR} . By now using the chiral decomposition for fields we can obtain:

$$\begin{aligned} \bar{e}_L \gamma^\mu e_L &= \bar{e} \gamma^\mu \frac{(1 - \gamma_5)}{2} e \\ \bar{e}_R \gamma^\mu e_R &= \bar{e} \gamma^\mu \frac{(1 + \gamma_5)}{2} e, \end{aligned}$$

and by substituting these expression we obtain:

$$\mathcal{L}_{NC}^W = \left(\bar{e} \gamma^\mu \frac{(Q_{zL} + Q_{zR})}{2} e + \bar{e} \gamma^\mu \gamma_5 \frac{(Q_{zR} - Q_{zL})}{2} \right) Z_\mu.$$

We ask ourselves if we can write expressions for the components Q_{zL}, Q_{zR} . Indeed we can by considering the quantum numbers of the fields in question:

$$Q_{zL} = \frac{e}{\sin \theta_W \cos \theta_W} (T_3(e_L) - Q(e_L) \sin^2 \theta_W) = \frac{e}{\sin \theta_W \cos \theta_W} \left(-\frac{1}{2} + \sin^2 \theta_W \right)$$

$$Q_{zR} = \frac{e}{\sin \theta_W \cos \theta_W} (T_3(e_R) - Q(e_R) \sin^2 \theta_W) = \frac{e}{\sin \theta_W \cos \theta_W} (\sin^2 \theta_W),$$

our linear combinations of components then become:

$$Q_{zR} - Q_{zL} = \frac{e}{2 \sin \theta_W \cos \theta_W}$$

$$Q_{zR} + Q_{zL} = \frac{e}{\sin \theta_W \cos \theta_W} \left[2 \sin^2 \theta_W - \frac{1}{2} \right].$$

By doing this manipulation we have obtained that the final current is written with the full physical field of the electron, and not with its chirality eigenstates. We finally write this interaction term as:

$$\mathcal{L}_{NC}^W = \frac{e}{2 \sin \theta_W \cos \theta_W} \bar{e} \gamma^\mu (v_e - a_e \gamma_5) e Z_\mu,$$

where two new quantities have been introduced, v_e, a_e , such that:

$$a_e \doteq -(Q_{zR} - Q_{zL}) = -\frac{1}{2} = T_3(e_L)$$

$$v_e \doteq Q_{zR} + Q_{zL} = -\frac{1}{2} (1 - 4 \sin^2 \theta_W) = T_3(e_L) (1 + 4Q(e) \sin^2 \theta_W).$$

These quantities are defined via the quantum numbers of the fields we are considering, and it is not a pure V-A current, since the coefficients are not the right ones: this implies a hybrid interaction term that however does not conserve parity.

Section Roadmap:

Covered:

Described the development of electroweak theory from Fermi's β -decay model to a gauge theory based on $SU(2)_L \otimes U(1)_Y$. Introduced weak isospin, hypercharge, and the gauge bosons $W^\pm$, Z^0 , and the photon. Derived the electroweak unification relations and showed how the photon and Z arise from mixing.

Coming Up:

Presentation of the limitations of the unbroken electroweak theory—specifically its inability to explain particle masses—leading into the Goldstone model and symmetry

breaking.

References: [10]; [7]; [11].



1.3 The Goldstone Model

This section deals with the Higgs-Brout-Englert mechanism and spontaneous symmetry breaking. It is the first approach to the Higgs mechanism, and it is simple and yet very profound. Up until now, we have been dealing with electroweak unification: all of the considerations we have done were done starting from this Lagrangian:

$$\mathcal{L} = \bar{L}i\not{D}L + \bar{e}_R i\not{D}e_R - \frac{1}{4}W_{\mu\nu}^i W_i^{\mu\nu} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu},$$

but there is an inherent pressing issue with the fields described by this formalism, *i.e.*, these fields are massless. Since we know things inside our universe have mass, we must find a way to insert masses inside these theories. One could wonder whether it is possible to add a Proca mass term by hand, such that, for a fermion field:

$$\begin{aligned} -m\bar{\psi}\psi &= -m(\bar{\psi}_L + \bar{\psi}_R)(\psi_L + \psi_R) = \\ &= -m(\bar{\psi}P_R + \bar{\psi}P_L)(P_L\psi + P_R\psi) = \\ &= -m(\bar{\psi}_L\psi_R + \bar{\psi}_R\psi_L), \end{aligned}$$

if we specialize this to an electron field:

$$-m_e\bar{e}e = -m_e(\bar{e}_Le_R + \bar{e}_Re_L).$$

By taking a transformation under the $SU(2)_L \otimes U(1)_Y$ group, we have that this term violates gauge invariance, since the right-handed and left-handed components change differently. But the true issue is that we do indeed need to build massive theories, both for the usual fermions, but also for gauge bosons, since the $W^\pm, Z^0$ particles are indeed quite heavy.

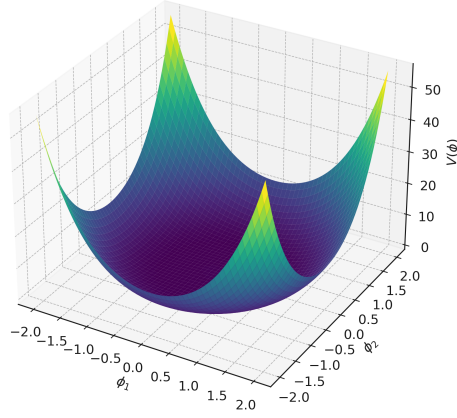
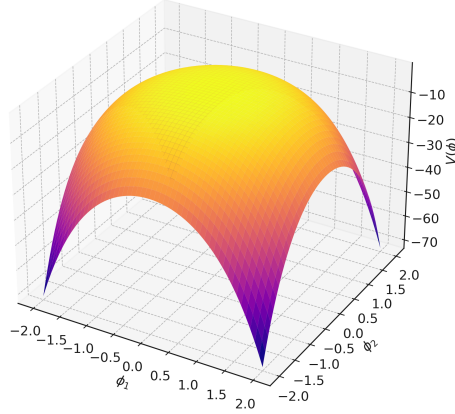
To this avail, we must then begin by introducing the Goldstone model, which will at first seem completely unrelated, but it will be useful in the following.

We start the introduction of this model by considering a complex scalar field ϕ such that it satisfies:

$$\begin{aligned} \phi &= \frac{1}{\sqrt{2}}[\phi_1 + i\phi_2] \\ \phi^* &= \frac{1}{\sqrt{2}}[\phi_1 - i\phi_2] \end{aligned}$$

We add a free propagation term and a potential to this theory such that the Lagrangian density reads:

$$\mathcal{L} = \partial_\mu\phi\partial^\mu\phi^* - V(\phi),$$

(A) Quartic potential with $\lambda > 0$.(B) Quartic potential with $\lambda < 0$.FIGURE 1.5: Quartic potentials depending on λ .

where $V(\phi)$ is a quartic potential of this kind:

$$V(\phi) = m^2 \phi \phi^* + \lambda (\phi \phi^*)^2 = m^2 |\phi|^2 + \lambda |\phi|^4.$$

We will make some assumptions about the values of the parameters at hand. We assume λ to be a positive coupling constant, while the term that looks like a mass term, the one in m^2 , will have any possible sign. We are sure that we can fix the positivity of λ by considering the possible shapes that a potential can assume. In this case, the sign of λ fixes the concavity of the potential, and if we ask that our theory is a physical theory, we will impose the potential to be bounded from below⁴, which corresponds to $\lambda > 0$. The different situations are represented in Figure (1.5).

⁴Requiring the potential to be bounded from below ensures that the Hamiltonian is bounded from below as well, which guarantees the existence of a stable vacuum state with minimal energy.

We would like to study the minima of this potential, we will look at the Hamiltonian density associated to this Lagrangian:

$$\mathcal{H} = \pi_i \dot{\phi}_i - \mathcal{L},$$

where π_i are the conjugate momenta associated to this field. If we do the calculation we obtain the expression of the Hamiltonian density:

$$\mathcal{H} = \partial_0 \phi^* \partial_0 \phi + \nabla \phi^* \nabla \phi + V(\phi),$$

since the terms in the derivatives of the fields are quadratic, the minimum of energy corresponds to the minimum of the potential $V(\phi)$. We stop for a moment and we look at the fact that the potential is invariant upon phase rotations of this kind:

$$\begin{aligned}\phi' &= e^{+i\alpha} \phi \\ \phi^{*'} &= e^{-i\alpha} \phi\end{aligned}$$

One could argue that this potential could have been written in a higher power, say $|\phi|^6$, all while maintaining gauge invariance, but this would have yielded inconsistent results with the renormalization conditions⁵ for the constant λ . The $\lambda|\phi|^4$ option is the most general and most simple expression possible for a potential of this kind.

We move on to studying the vacuum state of this QFT, by calculating the minimum of $V(\phi)$:

$$\frac{\partial}{\partial |\phi|} V(\phi) = 2m^2 |\phi| + 4\lambda |\phi|^3 = 2|\phi| (m^2 + 2\lambda |\phi|^2) = 0,$$

there are two cases for the solution of this equation, and they depend on the sign of the mass: $m^2 > 0$, $m^2 < 0$. We start from the case in which the mass parameter is positive, this implies that the extremum equation is satisfied just for $|\phi| = 0 \iff \phi(x) = 0$. We note that this vacuum state has the same symmetry of the dynamic field. If we transform $\phi(x) = 0$ with the phase transformations, we get $\phi(x) = 0$ identically. A symmetry like this is called an exact symmetry of Wigner-Weyl. If we consider particles to be oscillating states of fields around the vacuum state, we yield:

$$\langle 0 | \phi | 0 \rangle = 0.$$

The other case is a bit more ambitious, $m^2 < 0$. In this case there is another solution for the extremum of potential:

$$|\phi|^2 = -\frac{m^2}{2\lambda},$$

⁵By power counting, a $|\phi|^6$ interaction in 3+1D is non-renormalizable since the coupling has negative mass dimension.

we define the constants:

$$|\phi|^2 \doteq \frac{v^2}{2}, \quad v^2 = \frac{-m^2}{\lambda} = \frac{\mu^2}{\lambda}.$$

We find a new vacuum state. The potential has extrema, in the case of $m^2 < 0$, for $|\phi|=0$ and $|\phi| = \frac{v}{\sqrt{2}} = \phi_0$. We calculate the two extrema of the potential:

$$\begin{aligned} V(|\phi|=0) &= 0 \\ V(\phi_0) &= -\frac{\lambda}{4}v^4 < 0, \end{aligned}$$

this makes ϕ_0 a minimum and $|\phi|=0$ a local maximum. This is the famous Mexican hat potential, which will be extremely useful in the following, as depicted in Figure (1.6).

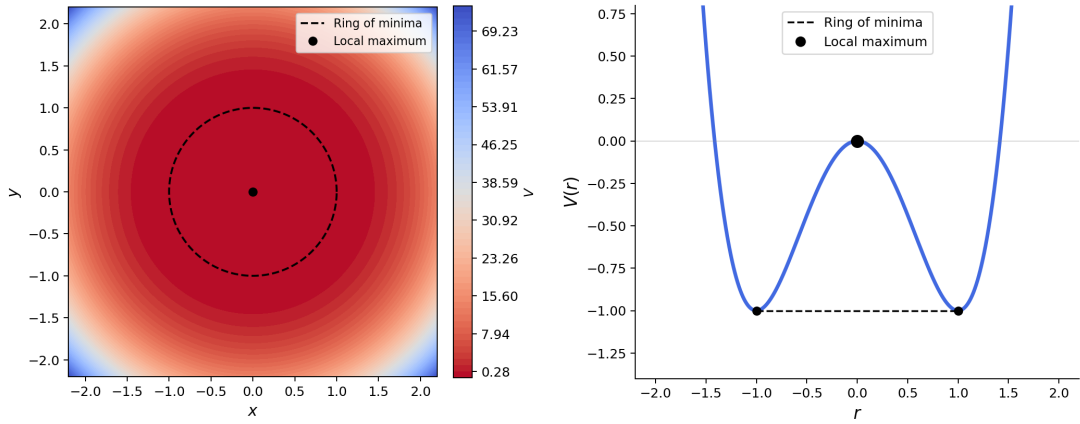


FIGURE 1.6: Mexican hat potential.

This potential has minima for all fields such that:

$$\phi(x) = \frac{v}{\sqrt{2}} e^{i\alpha x},$$

that means that the minima all lie on a circumference with radius $\frac{v}{\sqrt{2}}$. Since all minima around this circumference are equivalent, we can choose just one, and for simplicity we will choose $\phi_0 = \frac{v}{\sqrt{2}}$. We look at the transformation of the minimum: does this state fulfill the same symmetry as the Lagrangian does?

$$\phi_0 \longrightarrow \phi'_0 = e^{i\alpha} \phi_0 \approx (1 + i\alpha + \dots) \frac{v}{\sqrt{2}} = \phi_0 + i\alpha \frac{v}{\sqrt{2}}$$

The vacuum state ϕ_0 is now seen to change under a gauge symmetry transformation. This instance is called the Nambu-Goldstone symmetry, where the vacuum state does not fulfill the same symmetry of the Lagrangian, and this leads to spontaneous symmetry breaking. We say that the generator of symmetry $U(\alpha) = e^{i\alpha}$ is broken due to the fact that the vacuum state does indeed change under its action. Due to the definition of the

constants, it is also interesting to notice that the vacuum state seems to possess some kind of energy: the constant v is also referred to as the characteristic energy scale of spontaneous symmetry breaking⁶. We can now look at expansions around the minima of this vacuum state to get closer to an understanding of the importance of the Higgs boson. What we can do now is study the mass spectrum of the theory we have developed. We will not be studying the situation for which $\phi_0 = 0$ since it is unstable. We will be choosing one particular minimum state and study the oscillations of a field around the extreme. This will be clearer in the following. We will introduce two new fields, $\eta(x)$ and $\xi(x)$ such that they are defined to be oscillations around the vacuum state. The vacuum state field can be rewritten as, then:

$$\phi(x) = \frac{1}{\sqrt{2}} [v + \eta(x) + i\xi(x)]$$

These fields are defined to be different kinds of oscillations — η is a field that oscillates radially around the minimum, while ξ oscillates around the ring of minima, as can be seen in Figure (1.7).

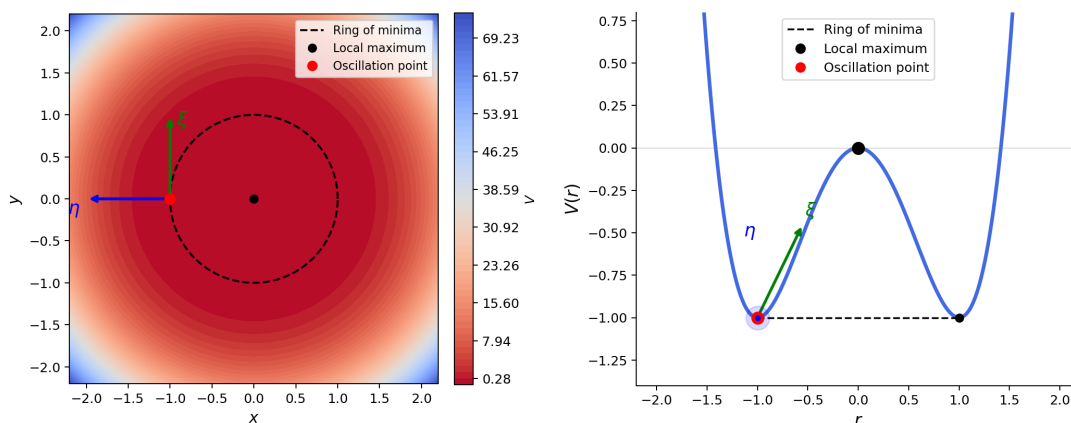


FIGURE 1.7: Oscillations around the Mexican hat potential's minima.

We now ask ourselves a different kind of question: why did we choose a complex scalar field to describe vacuum states? Theoretically we could have started by defining a fermionic field to describe a generic vacuum state, but the real issue is that fermionic fields by themselves do not behave well under Lorentz transformations: the vacuum must be Lorentz invariant, so we choose a scalar field to do this⁷.

⁶The constant v determines the field value at which the potential is minimized, setting the energy scale where electroweak symmetry breaking occurs. This is not the energy of a particle, but the scale at which new physics — such as gauge boson mass generation — becomes relevant. In the Standard Model, $v \approx 246$ GeV, a value extracted from the measured Fermi constant in weak interactions. Importantly, the theory does not predict v ; it must be fixed by experiment.

⁷This is done because a scalar field is the most minimal hypothesis. It is possible to build Lorentz-invariant combinations of fields, but they are more complex.

Having now cleared up this apparent inconsistency, we now move on to rewriting the Lagrangian with the new expression for ϕ .

$$\begin{aligned}\mathcal{L} &= \partial^\mu \phi^* \partial_\mu \phi - m^2 |\phi|^2 - \lambda |\phi|^4 \\ &= \frac{1}{2} \partial^\mu (v + \eta(x) - i\xi(x)) \partial_\mu (v + \eta(x) + i\xi(x)) - m^2 |\phi|^2 - \lambda |\phi|^4,\end{aligned}$$

working out the calculations for the kinetic term yields a Lagrangian for η, ξ :

$$\mathcal{L} = \frac{1}{2} \partial^\mu \eta \partial_\mu \eta + \frac{1}{2} \partial_\mu \xi \partial^\mu \xi - m^2 |\phi|^2 - \lambda |\phi|^4.$$

We now have to endure some algebra that will help us in studying the mass terms for these oscillations: we have obtained that both of these new oscillations propagate like a real Klein-Gordon field, now we need to see what happens to the potential. To this avail, we calculate the moduli of ϕ :

$$\begin{aligned}|\phi|^2 &= \frac{1}{2} [\text{Re } \phi^2 + \text{Im } \psi^2] = \frac{1}{2} [(v + \eta)^2 + \xi^2] = \frac{1}{2} [v^2 + \eta^2 + 2v\eta + \xi^2] \\ |\phi|^4 &= (|\phi|^2)^2 = \frac{1}{4} [v^2 + \eta^2 + 2v\eta + \xi^2]^2 = \\ &= \frac{1}{4} [v^4 + \eta^4 + 6v^2\eta^2 + \xi^4 + 4v^3\eta + 2v^2\xi^2 + 4v\eta^3 + 2\eta^2\xi^2 + 4v\eta\xi^2]\end{aligned}$$

In theory, now we should combine everything and tally up the results. We will get rid of terms that mix fields together and terms that are beyond the second order in the fields for simplicity and clarity:

$$|\phi|^4 = \frac{1}{4} [6v^2\eta^2 + 4v^3\eta + 2v^2\xi^2].$$

We are now ready to put everything into the potential part of the Lagrangian, yielding:

$$V(\phi) = \frac{m^2}{2} [v^2 + \eta^2 + 2v\eta + \xi^2] + \frac{\lambda}{4} [6v^2\eta^2 + 4v^3\eta + 2v^2\xi^2].$$

We now remind ourselves of the relations between the constants v, λ, m^2 :

$$v^2 = \frac{\mu^2}{\lambda} = -\frac{m^2}{\lambda},$$

allowing us to write:

$$\begin{aligned} V(\phi) &= -\frac{v^2\lambda}{2} [v^2 + \eta^2 + 2v\eta + \xi^2] + \frac{\lambda}{4} [6v^2\eta^2 + 4v^3\eta + 2v^2\xi^2] = \\ &= -\frac{v^4\lambda}{2} - \frac{v^2\lambda}{2}\eta^2 - \cancel{v^3\lambda\eta} - \frac{v^2\lambda}{2}\xi^2 + \frac{3\lambda v^2}{2}\eta^2 + \cancel{v^3\lambda\eta} + \frac{v^2\lambda}{2}\xi^2, \end{aligned}$$

where we notice that the uncomfortable linear term in η cancels out. This term is uncomfortable because it does not relate to any quantity we are familiar with in the Lagrangians we already know, and hence it cannot be directly associated to them. We also notice that the quadratic term in ξ cancels out. Since we are looking at mass terms of the Proca kind, *i.e.* $\propto \eta^2, \propto \xi^2$, we are going to identify the mass of the fields with the coefficient in front of the quadratic terms. The final Lagrangian without interaction terms then looks like this:

$$\mathcal{L} = \frac{1}{2}\partial^\mu\eta\partial_\mu\eta + \frac{1}{2}\partial_\mu\xi\partial^\mu\xi + \frac{v^4\lambda}{2} - \lambda v^2\eta^2.$$

We have then obtained that, by inspection, the mass associated to the ξ oscillation is zero, while the mass of the oscillation associated to η is nontrivial and it is equal to $2\lambda v^2$. From this calculation we can however conclude something that is quite surprising in itself: spontaneous symmetry breaking changes the mass spectrum of the theory we are considering. From a heuristic standpoint the fact that the η field has mass is not surprising, its oscillations are defined such that the potential changes upon the oscillations, while the ξ field in its oscillation does not change the potential, which is always at its minimum. The ξ field is also called a Nambu-Goldstone boson.

As we will see, although this result that seems a bit far-fetched at first, this is quite important for the development of our gauge theories: spontaneous symmetry breaking is the process that is able to give masses to the gauge bosons.

We now move on to the application to the gauge principle to this model. The Lagrangian can be seen to be invariant under a $U(1)$ global symmetry, but not under a $U(1)$ local symmetry. We want to look at this last case now, by applying the gauge principle and adding a massless gauge field B_μ . We follow the same procedure as before, defining a covariant derivative and a field tensor:

$$\begin{aligned} D_\mu &= \partial_\mu + igB_\mu \\ B_{\mu\nu} &= \partial_\mu B_\nu - \partial_\nu B_\mu. \end{aligned}$$

We can then write out the gauge invariant Lagrangian:

$$\mathcal{L} = (D^\mu\phi)^*(D_\mu\phi) - m^2|\phi|^2 - \lambda|\phi|^4 - \frac{1}{4}B_{\mu\nu}B^{\mu\nu}.$$

If we now operate the same expansion as before, we find that the η and ξ fields now interact with the gauge field — for this reason we will only focus on the covariant derivative term:

$$\begin{aligned} (D^\mu \phi)^*(D_\mu \phi) &= \frac{1}{2} [(\partial_\mu - igB_\mu)(v + \eta - i\xi)] [(\partial^\mu + igB^\mu)(v + \eta + i\xi)] \\ &= \frac{1}{\sqrt{2}} [\partial_\mu \eta - i\partial_\mu \xi - igvB_\mu - igB_\mu \eta - gB_\mu \xi] \\ &\quad \frac{1}{\sqrt{2}} [\partial^\mu \eta + i\partial^\mu \xi + igvB^\mu + igB^\mu \eta - gB^\mu \xi]. \end{aligned}$$

For this product, we will consider only the purely real terms, since the imaginary ones cancel out⁸. Expanding the product yields a final covariant derivative expression:

$$\begin{aligned} (D_\mu \phi)^*(D^\mu \phi) &= \frac{1}{2}(\partial_\mu \eta)(\partial^\mu \eta) + \frac{1}{2}(\partial_\mu \xi)(\partial^\mu \xi) + gvB^\mu(\eta \partial_\mu \xi - \xi \partial_\mu \eta) \\ &\quad + \frac{1}{2}g^2 B_\mu B^\mu (\eta^2 + \xi^2 + 2v\eta + v^2) + gvB_\mu \partial^\mu \xi. \end{aligned}$$

We notice that the gauge field we have introduced has gained a mass term! We can then see that a spontaneously broken symmetry not only generates mass terms for the oscillation field, but also for any gauge field whose corresponding generator is broken. There are also some odd terms that seem to imply mixing between spin-1 fields (B_μ) and spin-0 fields (ξ): we do not like such terms, since they are unphysical. This corresponds to a Feynman diagram that represents oscillations between spin-1 and spin-0 fields, and this is not possible.

$$\begin{array}{c} \text{---} \quad \text{---} \quad \text{---} \quad \text{---} \quad \text{---} \quad \text{---} \\ \xi \quad B_\mu \end{array} \Rightarrow \begin{array}{c} \text{---} \quad \text{---} \quad \text{---} \quad \text{---} \quad \text{---} \quad \text{---} \\ \xi \quad B_\mu \quad \xi \quad B_\mu \quad \dots \end{array}$$

There is another issue with this piece of theory, and it has to do with the count of degrees of freedom of the system. Before symmetry breaking, we had a complex field (ϕ) with two degrees of freedom and a gauge boson field (B_μ) with two degrees of freedom as well, for a total of four. After symmetry breaking, we find the same two degrees of freedom coming from η and ξ , but the B_μ gains an extra one due to its newfound mass, as seen in Table (1.1). This is inconsistent.

We notice that we can rid ourselves of the ξ oscillation via a gauge transformation. If we look at ξ as:

$$\phi' = \frac{1}{\sqrt{2}} e^{i\xi} (v + \eta) \approx \frac{1}{\sqrt{2}} (v + \eta + i\xi).$$

⁸ $(a + ib)(a - ib) = a^2 + b^2 - iab + iab = a^2 + b^2.$

Field	Before SSB	After SSB
ϕ	2 real d.o.f.	2 massive scalars (η, ξ)
Gauge field B_μ	2 (massless)	3 (massive)
Total d.o.f.	4	5

TABLE 1.1: Degrees of freedom in a locally $U(1)$ -invariant theory. Counting is not accurate.

We can then operate a transformation on the gauge field such that terms dependent on ξ just cancel out:

$$B_\mu(x) \rightarrow B'_\mu(x) = B_\mu(x) + \frac{1}{gv} \partial_\mu \xi(x).$$

The simultaneous application of these choices is called the unitary gauge, and it will be useful from now on. This is an important instance of the Goldstone theorem: We state it:

Theorem 1.3.1 (Goldstone's Theorem). *For every spontaneously broken continuous global symmetry, there exists a massless scalar particle — a Goldstone boson.*

In this case, the Goldstone boson is ξ , the field that is "gauged out", whose degree of freedom is eaten up by the massive vector field B_μ , and it is in this case called a Higgs-Kibble ghost. We can now count the number of degrees of freedom again to find them to be consistent before and after spontaneous symmetry breaking, as seen in Table (1.2).

Field	Before SSB	After SSB
ϕ	2 real d.o.f.	1 massive scalars (η)
Gauge field B_μ	2 (massless)	3 (massive)
Total d.o.f.	4	4

TABLE 1.2: Degrees of freedom in a locally $U(1)$ -invariant theory. Counting is accurate.

We were able to solve all the issues that this procedure yielded. We are now ready to move on to more complex theories, by tackling spontaneous symmetry breaking in the electroweak model, in such a way that the gauge fields that require mass ($W^\pm, Z$) become massive, taking special care in leaving the photon field A_μ unchanged.

Section Roadmap:
Covered:

Introduced spontaneous symmetry breaking via the Goldstone mechanism using a scalar field with a Mexican hat potential. Demonstrated how massless Goldstone bosons emerge and how one can gain mass terms for gauge bosons by “eating” degrees of freedom.

Coming Up:

Extension of spontaneous symmetry breaking to electroweak theory with the Higgs field, completing the Higgs mechanism in the Standard Model.

References: [12]; [13]; [9]; [7].



1.4 Electroweak Theory: After SSB

It is now time to apply this spontaneous symmetry breaking process to the electroweak model, and to do this, we start by looking at the symmetry group associated to this sector and its generators. The group is the well known:

$$SU(2)_L \otimes U(1)_Y,$$

and due to the relations we have already seen, we have that the number of generators we are expecting from this group is $3 + 1$. We have also proved that the electromagnetic $U(1)_Q$ is encoded in this group. This means that we have an extra generator to keep track of:

$$\begin{aligned} SU(2)_L &\Longleftrightarrow T_1, T_2, T_3 \\ U(1)_Y &\Longleftrightarrow Y \\ U(1)_Q &\Longleftrightarrow Q, \end{aligned}$$

where we remind the definitions of the generators:

$$\begin{aligned} T_i &= \frac{\sigma_i}{2} \\ Q &= T_3 + \frac{1}{2}Y. \end{aligned}$$

In the electroweak model there are three fields that need to acquire masses, and one field that does not need a mass. This implies that we need to break symmetry for three generators and keep one unbroken. The starting point is defining a left-handed doublet and a right-handed singlet for one family of leptons:

$$L = \begin{pmatrix} \nu_L \\ e_L \end{pmatrix} \quad e_R.$$

This is a theory that is built on the electroweak theory symmetry group, hence its interacting Lagrangian can be stated as:

$$\mathcal{L} = \bar{L}i\not{D}L + \bar{e}_Ri\not{D}e_R - \frac{1}{4}W_{\mu\nu}^iW^{\mu\nu i} - \frac{1}{4}F_{\mu\nu}F^{\mu\nu},$$

where we can identify different terms:

- $\not{D}$ is the covariant derivative for this theory, making the first two terms kinetic and interaction terms for the field doublet and singlet.
- The other two terms are tensors that describe the free propagation of the W_μ^i and B_μ fields, where $i = 1, 2, 3$.

Taking a look at the covariant derivative expression yields:

$$D_\mu = \partial_\mu - i\frac{g}{2}\sigma_i W_i^\mu - ig'\frac{Y}{2}B_\mu,$$

where two different gauge couplings have been introduced, g and g' . This is important because at first the symmetry groups at first are kept separate, and therefore have different coupling constants, but as we will observe later, they can be related to e when the symmetry is unified. We have to be extra careful in dealing with this expression since the forms of the generators changes depending on which field they act. We see that the fields W_μ^i, B_μ can be mapped into other physical fields with linear transformations to yield the physical (but still massless) fields:

- $W_\mu^\pm$, the field operator describing $W^\pm$ particles,
- Z_μ , the field operator describing the Z_0 particle,
- A_μ , the field operator describing the photon particle.

This linear transformation of fields yields the electroweak unification relation, where θ_W is the Weinberg mixing angle:

$$e = g \sin \theta_W = g' \cos \theta_W$$

This Lagrangian is however insufficient for giving masses to the gauge bosons. We need to add a Mexican hat potential term and a propagation term:

$$\mathcal{L} = \mathcal{L}_{ew} + \mathcal{L}_{SSB},$$

where we already have seen that the spontaneous symmetry breaking Lagrangian is both renormalizable and gauge invariant, so adding it to the Lagrangian should cause no issues. Since the term we have added must satisfy the symmetry group condition the field must be a doublet and it must possess hypercharge and isospin quantum numbers. We operate a somewhat arbitrary choice⁹ for this field, following the works of Weinberg:

$$\Phi = \begin{pmatrix} \phi_+ \\ \phi_0 \end{pmatrix},$$

and since we need four degrees of freedom, three Higgs-Kibble ghosts to generate masses, and a Higgs field mode, we impose that the field doublet is composed of two complex

⁹A choice that is not arbitrary because its form allows to yield spontaneous symmetry breaking for all generators except the charge operator, keeping the photon field massless.

Klein-Gordon fields, each with two degrees of freedom. We define them to be:

$$\begin{aligned}\phi_+ &= \frac{1}{\sqrt{2}} [\phi_1 + i\phi_2] \\ \phi_0 &= \frac{1}{\sqrt{2}} [\phi_3 + i\phi_4]\end{aligned}$$

If we extend the steps we have seen earlier for the Goldstone model, we can find the vacuum minimum state through defining a Lagrangian for this field:

$$\mathcal{L}_{SSB} = (D_\mu \Phi)^\dagger (D^\mu \Phi) - V(\Phi) = (D_\mu \Phi)^\dagger (D^\mu \Phi) - \mu^2 \Phi^\dagger \Phi - \lambda (\Phi^\dagger \Phi)^2,$$

where $\mu^2 < 0$ as before. We calculate the extrema of this function, and the possible extreme vacuum states are such that;

$$\begin{aligned}|\Phi|^2 &= 0 \\ |\Phi|^2 &= -\frac{\mu^2}{2\lambda} = \frac{v^2}{2},\end{aligned}$$

only one combination of these states, however, yields a minimum for the function. This value is called $\tilde{\Phi}_0$, the true vacuum state of the theory:

$$\tilde{\Phi}_0 \doteq \begin{pmatrix} 0 \\ \frac{v}{\sqrt{2}} \end{pmatrix}$$

If we now look at small oscillations around the potential, we have that the field Φ turns into a Higgs field doublet where the other three degrees of freedom have been reabsorbed in the unitary gauge condition¹⁰:

$$\Phi \longrightarrow \Phi = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + H(x) \end{pmatrix}$$

Why is this physically relevant? Nature is sensitive to isospin and hypercharge numbers, and we see that this field has $T_3 = \frac{1}{2}$ for ϕ_+ and $T_3 = -\frac{1}{2}$ for ϕ_0 , while the hypercharge operator on this field doublet is just a two-by-two identity. What happens to the charge operator under these assumptions? From electroweak unification we have:

$$Q = T_3 + \frac{Y}{2} = \frac{\sigma_3}{2} + \frac{Y}{2},$$

¹⁰The unitary gauge choice reabsorbs the Higgs-Kibble ghosts into arbitrary gauge transformations for the gauge fields. They are effectively removed from the theoretical development as unphysical degrees of freedom.

and then we look at its action on Φ :

$$Q\Phi = \left[\begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix} + \frac{1}{2}\mathbb{I}_{2 \times 2} \right] \Phi = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \phi_+ \\ \phi_0 \end{pmatrix} = \phi_+$$

This is a nice fact because we then see that the only component that interacts with the charge operator is ϕ_+ , while ϕ_0 is neutral, even if it is complex. Furthermore, this is great since we would like the Lagrangian to be invariant under some symmetries:

$$U(1)_{em} \longleftrightarrow e^{i\alpha(x)Q}.$$

If we take the vacuum state and transform it, we see that the $U(1)_{em}$ group generators are unbroken:

$$e^{i\alpha(x)Q}\tilde{\Phi}_0 \approx [1 + i\alpha(x)Q + \dots]\tilde{\Phi}_0$$

if we study this transformation, we have:

$$\tilde{\Phi}_0 \longrightarrow \tilde{\Phi}'_0 \approx \tilde{\Phi}_0 + i\alpha(x)Q\tilde{\Phi}_0.$$

By expanding the last term, we see that the electric charge operator is an unbroken generator:

$$Q\tilde{\Phi}_0 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ v \end{pmatrix} = 0.$$

If we now look at the more complex symmetries of $SU(2)_L$ we see that all three generators are broken under this choice of vacuum state.

$$e^{i\vec{\beta} \cdot \vec{T}}\tilde{\Phi}_0 \approx \left(\mathbb{I} + i\vec{\beta} \cdot \vec{T} + \dots \right) \tilde{\Phi}_0,$$

by expanding the $\vec{T}$ terms we have that the vacuum state choice breaks the symmetry, hence the three generators are broken.

$$\begin{aligned} T_1\tilde{\Phi}_0 &= \begin{pmatrix} 0 & +1 \\ +1 & 0 \end{pmatrix} \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v \end{pmatrix} \neq 0 \\ T_2\tilde{\Phi}_0 &= \begin{pmatrix} 0 & +i \\ -i & 0 \end{pmatrix} \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v \end{pmatrix} \neq 0 \\ T_3\tilde{\Phi}_0 &= \begin{pmatrix} +1 & 0 \\ 0 & -1 \end{pmatrix} \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v \end{pmatrix} \neq 0 \end{aligned}$$

We can look at the hypercharge symmetries to see that they are broken as well:

$$e^{i\gamma(x)Y}\tilde{\Phi}_0 \approx [1 + i\gamma(x)Y + \dots],$$

and since the Y operator is an identity, we have that this generator is broken as well. We have finally obtained a model in which the right gauge bosons acquire mass. It is now time to see how this happens by computing mass terms with the Higgs field.

We therefore proceed by adding a gauge invariant spontaneous symmetry breaking Lagrangian to this model:

$$\mathcal{L}_{SSB} = (D_\mu \Phi)^\dagger (D^\mu \Phi) - V(|\Phi|)$$

Since the Lagrangian must satisfy the gauge principle, we rebuild the same $SU(2)_L \otimes U(1)_Y$ covariant derivative:

$$D_\mu = \partial_\mu + ig \frac{\sigma^i}{2} W_\mu^i - ig' \frac{Y}{2} B^\mu = \partial_\mu + \frac{i}{2} \begin{pmatrix} g' B_\mu - g W_\mu & -g(W_\mu^1 - iW_\mu^2) \\ -g(W_\mu^1 + iW_\mu^2) & g' B_\mu + g W_\mu^3 \end{pmatrix}$$

where we can define the $W_\mu^\pm$ fields. We can then calculate the $D_\mu \Phi$ term: from this calculation we will find kinetic and interaction terms for the Higgs field, as well as mass terms for gauge fields:

$$D_\mu \Phi = \left[\partial_\mu + \frac{i}{2} \begin{pmatrix} g' B_\mu - g W_\mu & -\sqrt{2} g W_\mu^+ \\ -\sqrt{2} g W_\mu^- & g' B_\mu + g W_\mu^3 \end{pmatrix} \right] \begin{pmatrix} 0 \\ \frac{v+H(x)}{\sqrt{2}} \end{pmatrix}$$

By finishing this calculation, we yield:

$$D_\mu \Phi = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ \partial_\mu H(x) \end{pmatrix} + \frac{i}{2} \begin{pmatrix} -g W_\mu^+ \\ \frac{g' B_\mu + g W_\mu^3}{\sqrt{2}} \end{pmatrix} (v + H(x)).$$

Thanks to this result we can calculate $(D_\mu \Phi)^\dagger$:

$$(D^\mu \Phi)^\dagger = \frac{1}{\sqrt{2}} (0 \quad \partial^\mu H(x)) - \frac{i}{2} \begin{pmatrix} -g W^{-\mu} & \frac{g' B^\mu + g W^{3\mu}}{\sqrt{2}} \end{pmatrix} (v + H(x)).$$

Putting all of the pieces together we obtain:

$$\begin{aligned} (D^\mu \Phi)^\dagger D_\mu \Phi = & \left[\frac{1}{\sqrt{2}} (0 \quad \partial^\mu H(x)) - \frac{i}{2} \begin{pmatrix} -g W^{-\mu} & \frac{g' B^\mu + g W^{3\mu}}{\sqrt{2}} \end{pmatrix} (v + H(x)) \right] \cdot \\ & \left[\frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ \partial_\mu H(x) \end{pmatrix} + \frac{i}{2} \begin{pmatrix} -g W_\mu^+ \\ \frac{g' B_\mu + g W_\mu^3}{\sqrt{2}} \end{pmatrix} (v + H(x)) \right]. \end{aligned}$$

By expanding the products we obtain:

$$\begin{aligned} & \frac{1}{2} \partial^\mu H(x) \partial_\mu H(x) - \frac{i}{2\sqrt{2}} (v + H(x)) \begin{pmatrix} 0 \\ \partial_\mu H(x) \end{pmatrix} \begin{pmatrix} -gW^{-\mu} & \frac{g'B^\mu + gW^{3\mu}}{\sqrt{2}} \end{pmatrix} \\ & \frac{i}{2\sqrt{2}} (v + H(x)) \begin{pmatrix} -gW_\mu^+ \\ \frac{g'B_\mu + gW_\mu^3}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} 0 & \partial^\mu H(x) \end{pmatrix} \\ & + \frac{1}{4} (v + H(x))^2 \begin{pmatrix} -gW_\mu^+ \\ \frac{g'B_\mu + gW_\mu^3}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} -gW^{-\mu} & \frac{g'B^\mu + gW^{3\mu}}{\sqrt{2}} \end{pmatrix}. \end{aligned}$$

We notice that the mixed terms cancel out because they are real with opposite signs. Therefore we are left with:

$$\begin{aligned} (D^\mu \Phi)^\dagger D_\mu \Phi &= \frac{1}{2} \partial_\mu H(x) \partial^\mu H(x) \\ &+ \frac{1}{4} (v + H(x))^2 \left[g^2 W^{-\mu} W_\mu^+ + \frac{1}{2} (g'B_\mu + W_\mu^3)(g'B^\mu + W^{3\mu}) \right] \end{aligned}$$

Putting the Higgs kinetic term aside, we can look at the other ones. The first term in $W_\mu^+ W^{-\mu}$ yields pure interaction and mass terms when multiplied with $(v + H(x))^2$. It is now our goal to make the second term a pure sum of quadratic terms in A_μ and Z_μ . This procedure is equivalent to the diagonalization of the mass matrix of these fields. We take the term in question and manipulate it algebraically:

$$\frac{1}{2} (g^2 + g'^2) \begin{pmatrix} g'B_\mu + gW_\mu^3 \\ \sqrt{g^2 + g'^2} \end{pmatrix} \begin{pmatrix} g'B^\mu + gW^{3\mu} \\ \sqrt{g^2 + g'^2} \end{pmatrix}.$$

We can define Weinberg mixing angle trigonometric functions via:

$$\sin \theta_W = \frac{g'}{\sqrt{g^2 + g'^2}} \quad \cos \theta_W = \frac{g}{\sqrt{g^2 + g'^2}}.$$

Substituting into the previous equations and reminding ourselves of the mixing relations between A_μ and Z_μ :

$$\frac{1}{2} (g^2 + g'^2) (\sin \theta_W B_\mu + \cos \theta_W W_\mu^3) (\sin \theta_W B^\mu + \cos \theta_W W^{3\mu}) = \frac{1}{2} (g^2 + g'^2) Z_\mu Z^\mu.$$

Thanks to the Higgs mechanism we see that only the Z_μ field acquires a mass term. We can therefore move forwards and calculate the explicit mass terms and couplings:

$$\begin{aligned}
& \frac{1}{4} (v + H(x))^2 g^2 W_\mu^- W^{+\mu} + \frac{1}{4} (v + H(x))^2 (g^2 + g'^2) Z_\mu Z^\mu = \\
& = \frac{1}{4} (v^2 + H(x)^2 + 2vH(x)) g^2 W_\mu^- W^{+\mu} \\
& \quad + \frac{1}{4} (v^2 + H(x)^2 + 2vH(x)) (g^2 + g'^2) Z_\mu Z^\mu = \\
& = \frac{1}{4} v^2 g^2 W_\mu^- W^{+\mu} + \frac{1}{4} v^2 (g^2 + g'^2) Z_\mu Z^\mu + \frac{1}{4} g^2 H(x)^2 W_\mu^- W^{+\mu} + \\
& \quad + \frac{1}{4} (g^2 + g'^2) Z_\mu Z^\mu H(x)^2 + \frac{1}{2} g^2 v H(x) W_\mu^- W^{+\mu} + \frac{1}{2} v (g^2 + g'^2) H(x) Z_\mu Z^\mu.
\end{aligned}$$

We are now face to face with mass terms for gauge fields and interaction terms between gauge fields and the Higgs field:

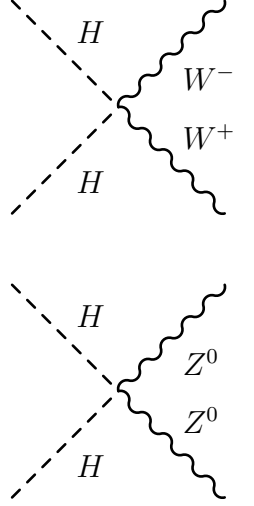
- The first two terms are indeed mass terms for the W and Z fields. Their masses depend on the electroweak couplings as well as on the vacuum expectation value v :

$$\begin{aligned}
M_W^2 &= \frac{1}{4} v^2 g^2 & M_W &= \frac{1}{2} v \frac{e}{\sin \theta_W} \\
M_Z^2 &= \frac{1}{4} v^2 (g^2 + g'^2) & M_Z &= \frac{1}{2} v \sqrt{g^2 + g'^2} = \frac{ev}{2 \cos \theta_W \sin \theta_W}.
\end{aligned}$$

- We have trilinear interaction terms with the Higgs field:

$$\begin{aligned}
& \begin{array}{c} \text{wavy line } W^- \\ \text{wavy line } W^+ \end{array} \text{ --- } \text{dashed line } H \quad \propto \quad \frac{1}{2} g^2 v = \frac{2M_W^2}{v} \\
& \begin{array}{c} \text{wavy line } Z \\ \text{wavy line } Z \end{array} \text{ --- } \text{dashed line } H \quad \propto \quad \frac{1}{2} (g^2 + g'^2) = \frac{2M_Z^2}{v}
\end{aligned}$$

- We have quadrilinear interaction terms with the Higgs field:



$$\propto \frac{1}{4}g^2 = \frac{e^2}{4 \sin^2 \theta_W}$$

$$\propto \frac{1}{4}(g^2 + g'^2) = \frac{e^2}{8 \sin^2 \theta_W \cos^2 \theta_W}$$

We have therefore enriched the physics of the Standard Model in a great manner: we have added gauge-invariant mass terms in a way that does not break symmetry, and we have obtained new interaction terms between the newly introduced Higgs field and the gauge bosons. We still have to add mass terms to the matter fields L and e_R . We shall do just that by employing a Yukawa Lagrangian.

To this avail, we introduce another Higgs field doublet, defined to be the Hermitian conjugate of Φ . This in the unitary gauge is trivial and can be written out to be:

$$\Phi^\dagger = \begin{pmatrix} 0 & \frac{(v+H(x))}{\sqrt{2}} \end{pmatrix}.$$

We can use this new field to build the new Yukawa Lagrangian for the mass of leptons:

$$\mathcal{L}_Y = -h [\bar{L}\Phi e_R + \bar{e}_R \Phi^\dagger L].$$

This Lagrangian is $SU(2)_L \otimes U(1)_Y$ invariant, since Φ is a doublet whose quantum numbers are:

$$Y\Phi = \mathbb{I}_{2 \times 2} \Phi.$$

Under symmetries the Lagrangian is seen to be invariant:

- $SU(2)_L$ transformation:

$$(\bar{L}\Phi e_R)' = \bar{L}'\Phi e_R' = \bar{L}U^\dagger U\Phi e_R = \bar{L}\Phi e_R$$

- $U(1)_Y$ transformation:

$$Y(L) = -\mathbb{I}_{2 \times 2} \quad Y(\Phi) = \mathbb{I}_{2 \times 2} \quad Y(e_R) = -2$$

whose combined transformation is dependent on the sum of these quantum numbers, which is zero, making it invariant.

We can see by expanding this Lagrangian that it yields a useful mass term:

$$-h \left[(\bar{\nu}_L \quad \bar{e}_L) \begin{pmatrix} 0 \\ \frac{v+H(x)}{\sqrt{2}} \end{pmatrix} e_R + \bar{e}_R \begin{pmatrix} 0 \\ \frac{v+H(x)}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} \nu_L \\ e_L \end{pmatrix} \right].$$

By now collapsing all of the algebraic structures:

$$-h \frac{(v + H(x))}{\sqrt{2}} [\bar{e}_L e_R + \bar{e}_R e_L] = -h \frac{(v + H(x))}{\sqrt{2}} \bar{e} e.$$

We have laid the last piece of the puzzle. The electroweak theory is now complete with all the required mass terms. The h terms are Yukawa couplings whose values must be obtained via experiment. We can then write the full electroweak Lagrangian for the first family of leptons:

$$\begin{aligned} \mathcal{L}_{ew} = & \bar{L} i \not{D} L + \bar{e}_R i \not{D} e_R - \frac{1}{4} W_{\mu\nu}^i W^{\mu\nu i} - \frac{1}{4} F_{\mu\nu} F^{\mu\nu} + \\ & + (D_\mu \Phi)^\dagger (D^\mu \Phi) - V(|\Phi|) - h [\bar{L} \Phi e_R + \bar{e}_R \Phi^\dagger L]. \end{aligned}$$

We can extend this to all three lepton families simply by introducing a new index ℓ that will denote which lepton flavor we are referring to. Therefore the complete Lagrangian is:

$$\begin{aligned} \mathcal{L}_{ew} = & \sum_{\ell=e,\mu,\tau} \bar{L}^\ell i \not{D} L^\ell + \bar{\ell}_R i \not{D} \ell_R - \frac{1}{4} W_{\mu\nu}^i W^{\mu\nu i} - \frac{1}{4} F_{\mu\nu} F^{\mu\nu} + \\ & + (D_\mu \Phi)^\dagger (D^\mu \Phi) - V(|\Phi|) - h [\bar{L}^\ell \Phi \ell_R + \bar{\ell}_R \Phi^\dagger L^\ell]. \end{aligned}$$

We have therefore been able to build a massive gauge-invariant theory thanks to spontaneous symmetry breaking and the Higgs field.

Section Roadmap:

Covered:

Integration of the Higgs field into the $SU(2)_L \otimes U(1)_Y$ theory, generation of W and Z boson masses, introduced the Higgs boson as a scalar excitation after symmetry breaking.

Coming Up:

Experimental consequences of the Higgs boson: how it is produced and detected at high-energy colliders.

References: [14]; [15]; [13]; [10]; [16].



1.5 Higgs Boson Phenomenology

In this Section we are interested in giving a better outline to the phenomenology of the Higgs boson. We therefore start from the basics, and retrieve the spontaneous symmetry breaking Lagrangian we used earlier:

$$\mathcal{L}_{SSB} = (D_\mu \Phi)^\dagger (D^\mu \Phi) - V(|\Phi|),$$

where

$$V(|\Phi|) = m^2 \Phi^\dagger \Phi + \lambda (\Phi^\dagger \Phi)^2.$$

We know that the minimum of the Mexican hat potential is related to the vacuum expectation value v such that $V(|\Phi|) = \frac{-\lambda v^4}{4}$. From these elements we can lay the foundations to the physics of the Higgs sector. We can do this by using the unitary gauge representation of the Φ field, the one where the the Goldstone modes of oscillations are "gauged out":

$$\Phi = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + H(x) \end{pmatrix}.$$

The phenomenology for the Higgs field is contained in the spontaneous symmetry breaking potential, and if we expand the $V(|\Phi|)$ term we find:

$$\begin{aligned} m^2 \Phi^\dagger \Phi &= \frac{m^2}{2} (v + H(x))^2 = \frac{m^2}{2} (v^2 + 2vH(x) + H(x)^2) \\ \lambda (\Phi^\dagger \Phi)^2 &= \frac{\lambda}{4} [(v^2 + 2vH(x) + H(x)^2)]^2. \end{aligned}$$

Expanding the term in λ yields:

$$\lambda (\Phi^\dagger \Phi)^2 = \frac{\lambda}{4} [v^4 + 4v^2 H(x)^2 + H(x)^4 + 4v^3 H(x) + 4vH(x)^3 + 2v^2 H(x)^2].$$

Since $\lambda v^2 = -m^2 = \mu^2$ we can use these relations to build the final term:

$$\begin{aligned} m^2 (\Phi^\dagger \Phi) + \lambda (\Phi^\dagger \Phi)^2 &= \\ &= -\lambda v^2 (\Phi^\dagger \Phi) + \lambda (\Phi^\dagger \Phi)^2 = \\ &= -\frac{\lambda v^2}{2} (v^2 + 2vH(x) + H(x)^2) \\ &\quad + \frac{\lambda}{4} [v^4 + 4v^2 H(x)^2 + H(x)^4 + 4v^3 H(x) + 4vH(x)^3 + 2v^2 H(x)^2] = \\ &= -\frac{\lambda v^4}{2} - \lambda v^3 H(x) - \frac{\lambda v^2}{2} H(x)^2 \\ &\quad + \frac{\lambda v^4}{4} + \lambda v^2 H(x)^2 + \frac{\lambda}{4} H(x)^4 + \lambda v^3 H(x) + \lambda v H(x)^3 + \frac{\lambda v^2}{2} H(x)^2. \end{aligned}$$

Some terms cancel out and we find that the Higgs potential determines a specific set of self-interactions for the Higgs field:

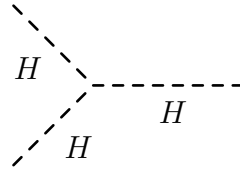
$$\lambda v^2 H(x)^2 + \lambda v H(x)^3 + \frac{\lambda}{4} H(x)^4.$$

We can then analyze each of these terms:

- The term quadratic in $H(x)$ is a mass term, determining the mass of the Higgs boson to be:

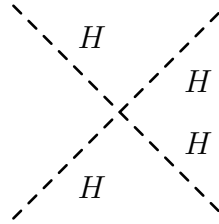
$$\frac{1}{2}M_H^2 \doteq \lambda v^2.$$

- We have a trilinear term in $H(x)$ such that:



$$\propto \lambda v = \frac{M_H^2}{2v}.$$

- We have a quadrilinear term in $H(x)$ such that:



$$\propto \frac{\lambda}{4} = \frac{M_H^2}{8v^2}.$$

As we have seen previously, the coupling of the Higgs depends on the mass of the particle it is interacting with, in this case, it is the Higgs itself. The trilinear and quadrilinear interactions terms cannot yet be observed at LHC due to the low cross section of the process, so we must find the Higgs's mass in another way. This M_H is a parameter of the Standard Model that cannot be found through calculation. It is possible to find bounds on it through fits to experimental data of electroweak processes. Looking at a $e^+e^- \rightarrow f\bar{f}$ process, one can measure its cross section. If we look at NLO processes, these diagrams, as in Figure (1.8) involve loops with other particles.

By fitting with LEP electroweak precision data, it was possible to impose a bound on the Higgs boson mass such that:

$$M_H = 91_{-23}^{+30} \text{ GeV} \quad (\text{standard fit})$$

$$M_H = 120_{-5}^{+12} \text{ GeV} \quad (\text{complete fit})$$

The corresponding upper bounds at 95% (99%) confidence level are:

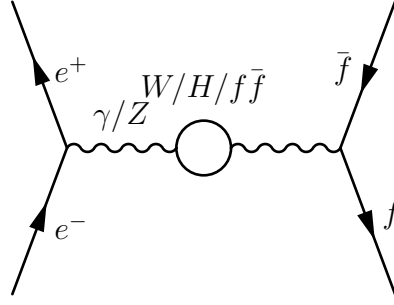


FIGURE 1.8: Feynman diagram for the $e^+e^- \rightarrow f\bar{f}$ process

- Standard fit: $M_H < 163$ (194) GeV
- Complete fit: $M_H < 143$ (149) GeV

These bounds include theoretical uncertainties, estimated to contribute approximately ± 8 GeV to the Higgs mass determination.

We can now move forwards and study how the Higgs boson behaves inside the Standard Model. We will first start with how one can produce a Higgs boson at LHC. The Large Hadron Collider is a collider machine that is able to accelerate protons up to center-of-mass energies of 14 TeV and study the products of these collisions via the four experiments that are lying on the particle beam track: ALICE, ATLAS, CMS, and LHCb. What one does is accelerate protons and smash them together in hope to produce a Higgs boson, then study the decay products of said particle. Since protons are baryons, the smallest particle unit that can interact is a quark: we need to look at the quark parton model. In every pp scattering interaction only one of the partons takes direct part to the interaction following the sample Feynman diagram of Figure (1.9).

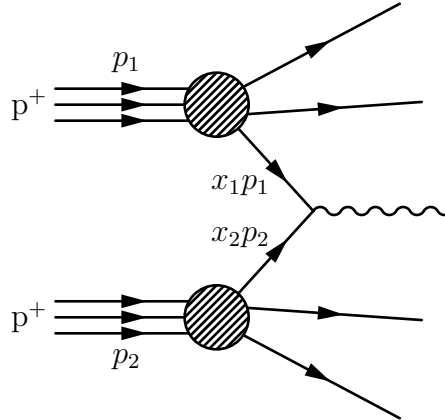


FIGURE 1.9: Parton model proton scattering, two quarks only interact.

What happens in this case is that a quark gets "picked" with a fraction x of the total momentum of the original proton to take part to the interaction: this seems extremely difficult to reconstruct with a detector. Not only is it difficult to find the exact fractions

x_1, x_2 of momentum carried by the partons, but one must also consider the fact that the two other partons continue their motion and interact with the detectors, leaving traces as well. This means that a simple interaction vertex like the one in Figure (1.9), coupled with quantum chromodynamics phenomena like gluon emission becomes much more complex, and it more closely resembles something like in Figure (1.10).

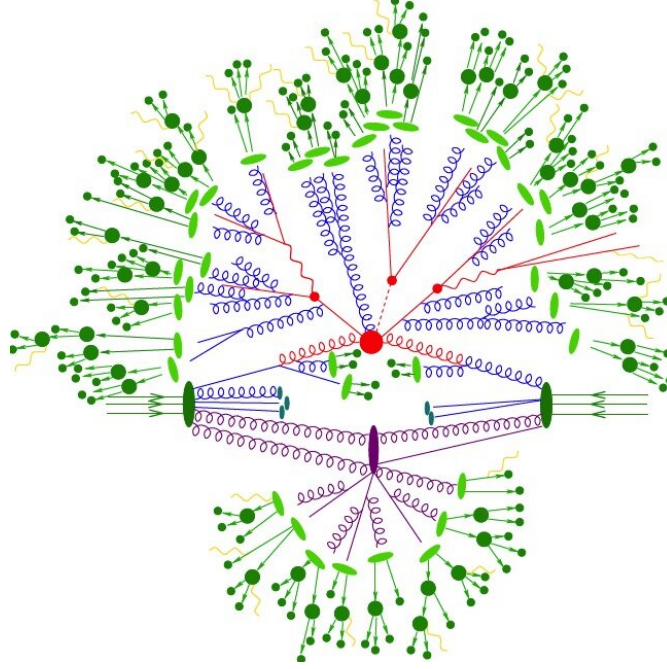


FIGURE 1.10: Proton-proton collision event generated with Sherpa. Courtesy of [17].

This seems just impossible to reconstruct, since there seem to be too many variables. Luckily, from the study of electron scattering we can extract a new tool that is helpful towards getting out of this conundrum. Parton distribution functions, PDFs for short, are non-perturbative tools that help us in figuring out the fractions of momentum carried by the interacting parton. We can use these experimental tools and the factorization theorem of QCD to state that the cross section of a proton-proton collision can be written as:

$$\sigma_{pp \rightarrow X}(S) = \sum_{i,j} \int_0^1 dx_1 \int_0^1 dx_2 f_i(x_1, \mu_F^2) f_j(x_2, \mu_F^2) \hat{\sigma}_{ij \rightarrow X}(\hat{s} = x_1 x_2 S, \mu_F^2, \mu_R^2),$$

where:

- $f_i(x, \mu_F^2)$ is the PDF for parton i carrying momentum fraction x ,
- μ_F is the factorization scale,
- μ_R is the renormalization scale,

- $\hat{\sigma}_{ij \rightarrow X}$ is the perturbative partonic cross section,
- S is the hadronic center-of-mass energy squared,
- $\hat{s} = x_1 x_2 S$ is the partonic center-of-mass energy squared.

This implies that the PDFs $f_i(x, \mu_F^2)$ are functions that describe the probability of extracting a parton i with momentum x from a hadron. These are nonperturbative, so they cannot be calculated, but must be measured experimentally. They are ubiquitous since they apply to any hadron, but they constitute the largest uncertainty for measurements. The sample PDF for the proton used at LHC is as seen in Figure (1.11).

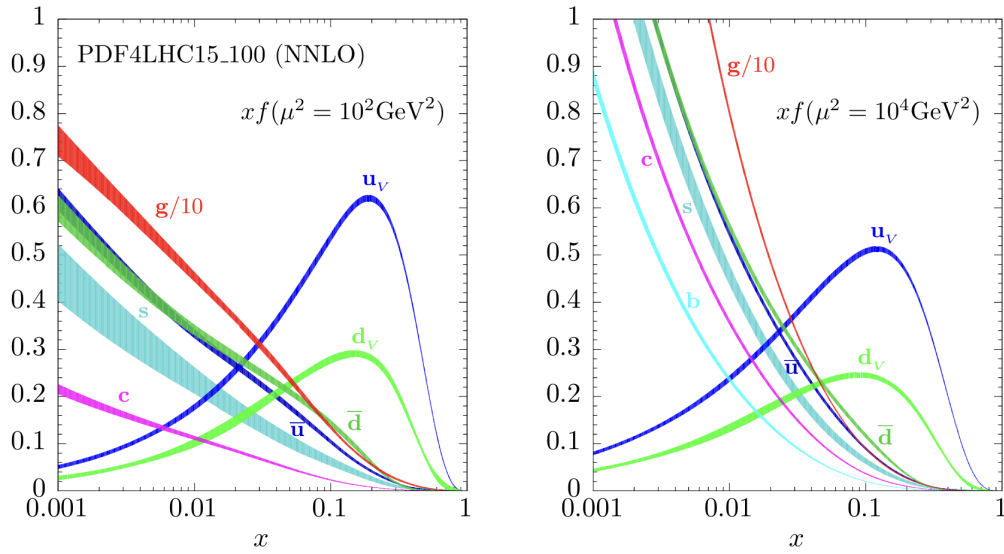


FIGURE 1.11: Comparison between PDFs for the proton at two different scales: 10^2 GeV (left), 10^4 GeV (right). Courtesy of [18].

Since the momentum fractions of the extracted partons are different, one can also assume that the center of mass of the interaction does not coincide with the laboratory's frame of reference: we then assume that the interaction's frame of reference is boosted with respect to the collider's frame. This implies that the kinematical variables of interest will be:

- Transverse momentum p_T : the only component of momentum that is not boosted and, therefore, conserved;
- Azimuthal angle ϕ ;
- Pseudorapidity η , a measurement of the angle of the particle with respect to the beam line:

$$\eta = -\ln \left[\tan \left(\frac{\theta}{2} \right) \right] \quad \eta \in [-\infty, +\infty]$$

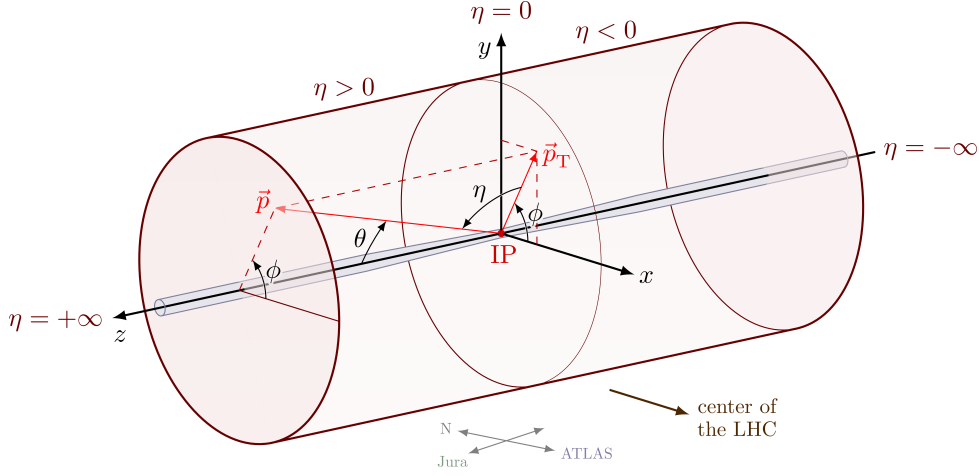


FIGURE 1.12: A 3D rendering of the important variables for a detector experiment, in this case, CMS.

Figure (1.12) perfectly depicts the variables that have just been presented. Now that we have a better understanding of how a measurement works at LHC, which will be useful for the understanding of the rest of this Thesis, we can dive head first into the phenomenology of the Higgs boson, starting from the way it can be produced.

1.5.1 Higgs Boson Production

There are many different ways in which one can produce a Higgs boson at the Large Hadron Collider.

The first and dominant process is gluon-gluon fusion: this process fuses two gluons via a top quark loop to produce a Higgs boson. This process is pictured in the Feynman diagram in Figure (1.13a).

The second most important production mode is the one that involves vector boson fusion (VBF): $qq \rightarrow qq + H$. Two virtual bosons, either W or Z fuse into a Higgs, as seen in Figure (1.13b).

Another process is the *Higgs-strahlung* process, where we have a spontaneous emission of a Higgs boson in the process $q\bar{q} \rightarrow W/Z + H$, and this is depicted in Figure (1.13c).

There are two additional channels, one in which a Higgs boson is produced with a top (or bottom) quark pair ($t\bar{t}H/b\bar{b}H$), and one in which a Higgs boson is produced with a single top (tH) — Figure (1.13d) and Figure (1.13e), respectively. Both of these channels are quite rare and amount to little of the total Higgs production, but they are still important to observe the Higgs properties, namely its interactions and couplings.

Table (1.3) presents the cross section for each of the production modes listed earlier, their percentages, and the precision of the calculation used to obtain the numbers.

The calculations to obtain these numbers span several perturbative orders, and therefore these numbers represent the most precise estimate for cross sections for Higgs production,

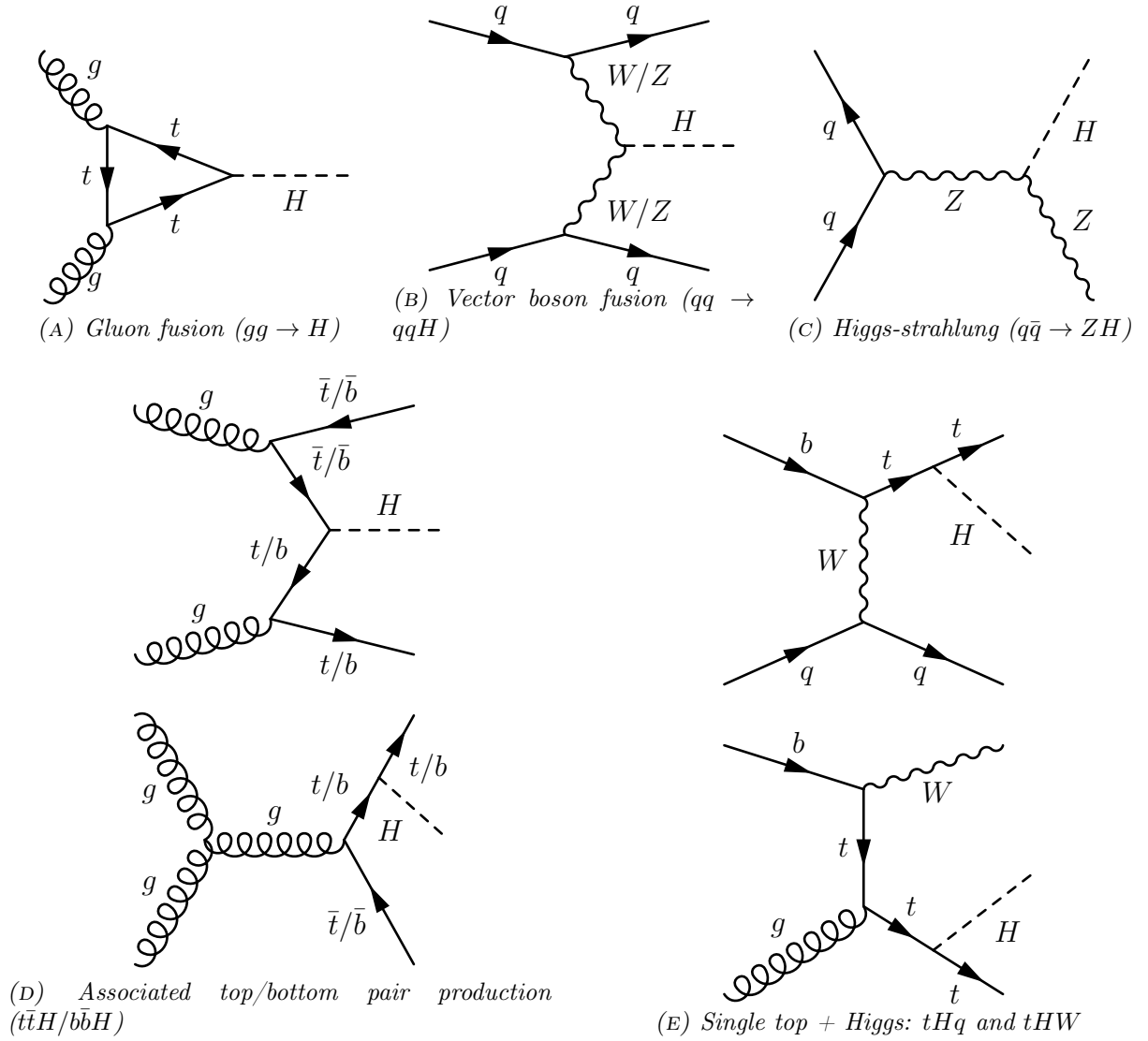


FIGURE 1.13: Representative Feynman diagrams for key Higgs boson production mechanisms at hadron colliders.

Top row: direct production via gluon fusion (A), vector boson fusion (B), and Higgs-strahlung (C).

Bottom row: associated production with top quarks in $t\bar{t}H$ (D), tHq (E), and tHW (F).

TABLE 1.3: *SM Higgs production cross sections and fractional contributions at $\sqrt{s} = 13$ TeV, $m_H = 125$ GeV (ATLAS Run 2 data: [19])*

Mode	σ [pb]	Fraction [%]	Precision (Perturbative Order)
ggF	48.58	87.06	N ³ LO QCD + NLO EW
VBF	3.749	6.72	NNLO QCD + NLO EW
WH	1.373	2.46	NNLO QCD + NLO EW
ZH	0.884	1.58	NNLO QCD + NLO EW
$t\bar{t}H$	0.611	1.09	NLO QCD
$b\bar{b}H$	0.494	0.88	N ³ LO QCD (5FS)
tH	0.103	0.18	NLO QCD
Total	55.83	100	—

have been used for analyses for Run2 and Run3 — they represent the most up-to-date benchmark for the sector.

1.5.2 Higgs Boson Decay

We now move on to the Higgs boson decay modes: we will present the most up to date branching ratios (BR) for the Higgs sector. Since the Higgs boson couples to any particle of the Standard Model, as we have seen throughout this Chapter, we can expect its decay modes to be quite varied. We list them and their numerical values in Table (1.4):

- $H \rightarrow b\bar{b}$: This is the most common decay channel of the Higgs boson, involving its transformation into a pair of bottom quarks. Detecting this decay is challenging because its background is experimentally significantly complex. This channel is depicted in Figure (1.14a).
- $H \rightarrow W^+W^-$ (off-shell): The Higgs can decay into a pair of W bosons, where one or both are often off their mass shell (virtual). This decay mode is experimentally favorable because the following decays of the W bosons to leptons produce clear traces with cleaner backgrounds. This channel is depicted in Figure (1.14c).
- $H \rightarrow gg$: The decay into two gluons is not a tree-level diagram, since gluons are not massive therefore they do not interact with the Higgs. This loop-diagram is primarily mediated by heavy quarks such as the top quark. While it cannot be directly observed due to overwhelming QCD backgrounds, it plays a vital theoretical role, influencing the Higgs production and total decay width. This channel is depicted in Figure (1.14d).
- $H \rightarrow \tau^+\tau^-$: The Higgs boson decays into tau lepton pairs, providing a unique insight into its coupling to leptons. Taus decay in identifiable ways that make this

channel experimentally accessible despite its smaller probability. This channel is depicted in Figure (1.14b).

- $H \rightarrow c\bar{c}$: Decay into charm quarks is difficult to isolate because it mimics the more prevalent bottom quark decays and is buried under large QCD backgrounds. This channel is depicted in Figure (1.14a).
- $H \rightarrow ZZ^*$: The Higgs decays into two Z bosons, one of which is often off-shell due to mass constraints. This channel, especially when both Z bosons subsequently decay into charged leptons, produces a very clean and distinct signature with excellent mass resolution. This channel is depicted in Figure (1.14c).
- $H \rightarrow \gamma\gamma$: The decay into two photons is a rare, loop-induced process. Despite its small branching fraction, this channel is very important experimentally because the photons produce a clean, distinctive final state with high energy resolution. This channel is depicted in Figure (1.14f).
- $H \rightarrow Z\gamma$: Another rare loop-induced decay, the Higgs boson can decay into a Z boson and a photon. Detecting this mode is experimentally challenging. This channel is depicted in Figure (1.14f).
- $H \rightarrow \mu^+\mu^-$: The Higgs boson decay to a pair of muons is extremely rare, but it is key to probing the Higgs coupling to second-generation leptons. This channel is depicted in Figure (1.14b).

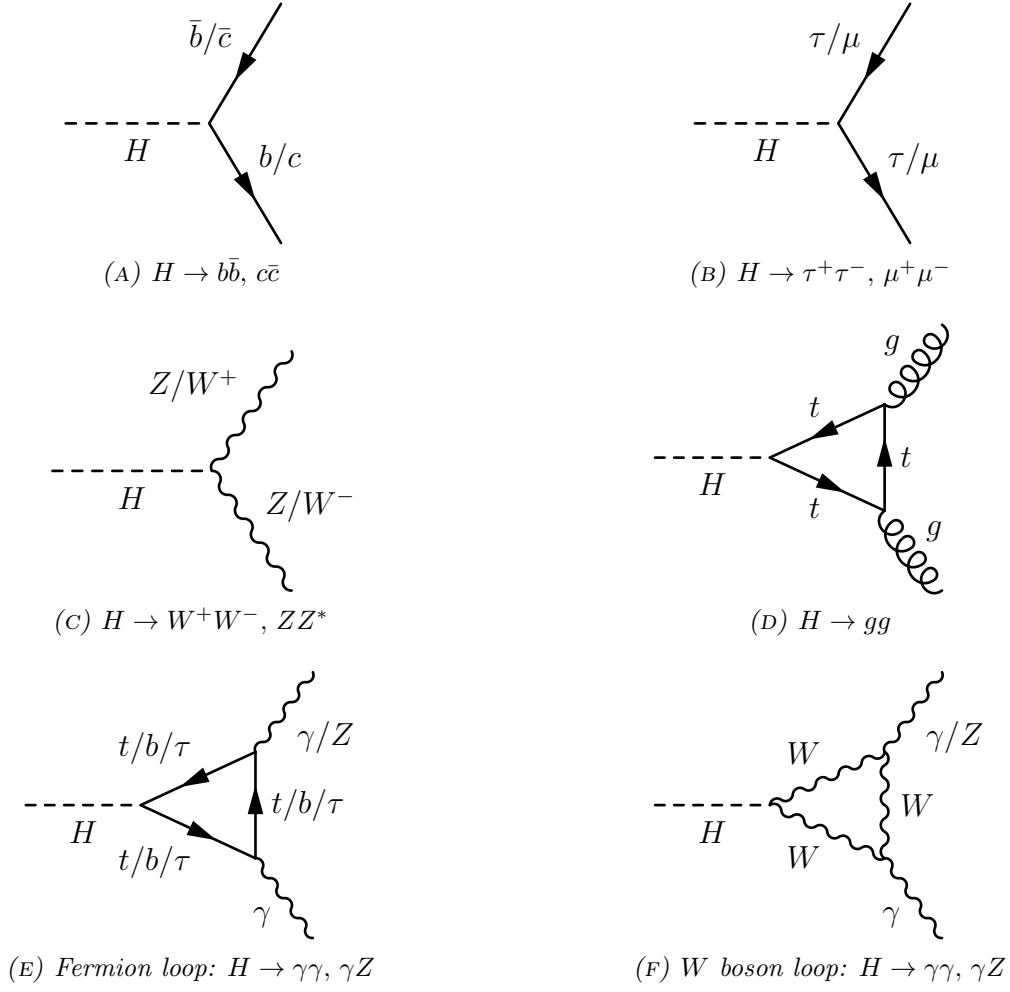


FIGURE 1.14: Feynman diagrams of Higgs boson decay channels.

TABLE 1.4: Higgs boson decay branching ratios at $\sqrt{s} = 13 \text{ TeV}$, $m_H = 125 \text{ GeV}$.

Decay Mode	Branching Ratio [%]
$H \rightarrow b\bar{b}$	58.2
$H \rightarrow W^+W^-$	21.4
$H \rightarrow gg$	8.18
$H \rightarrow \tau^+\tau^-$	6.32
$H \rightarrow c\bar{c}$	2.90
$H \rightarrow ZZ^*$	2.64
$H \rightarrow \gamma\gamma$	0.227
$H \rightarrow Z\gamma$	0.153
$H \rightarrow \mu^+\mu^-$	0.022

1.5.3 Higgs Boson Discovery

On July 4th, 2012, the ATLAS and CMS collaborations at CERN Large Hadron Collider (LHC) announced the observation of a new boson with a mass around 125 GeV, consistent with the long-sought Higgs particle. The latest estimate for the Higgs boson mass from the Particle Data Group (PDG) is:

$$M_H = 125.25 \pm 0.17 \text{ GeV}.$$

The Higgs boson interacts with almost all particles of the Standard Model, and it was first observed in a special subset of decay channels, the ones that gave rise to a simpler background to analyze. The few select decay channels are as follows:

- $H \rightarrow ZZ^* \rightarrow 4\ell$: this channel provides sensitivity over a mass range spanning from 110 GeV to 600 GeV. The signal requires the identification of two opposite-charge pairs of leptons, where the most energetic one, the leading lepton pair, has invariant mass m_{12} close to the one of the Z : $m_{12} \in [50 \text{ GeV}, 106 \text{ GeV}]$, while the other pair has variable invariant mass range. The main background for this channel is made up of ZZ^* and its subsequent decays in jets and leptons. Figure (1.15) depicts the distribution of the four-lepton invariant mass and the expected Higgs signal.
- $H \rightarrow \gamma\gamma$: this search is performed over mass range between 110 – 150 GeV. The main background is constituted of Standard Model photon pair production. The invariant mass of the two photons is evaluated using the photon energies measured in the calorimeter, the azimuthal angle ϕ between the photons as determined from the positions of the photons in the calorimeter, and the values of η calculated from the position of the identified primary vertex and the impact points of the photons in the calorimeter: the invariant mass for these diphotons is in the range $m_{\gamma\gamma} \in [100 \text{ GeV}, 160 \text{ GeV}]$. Figure (1.15) shows the invariant mass distribution of signal diphoton events against the background.

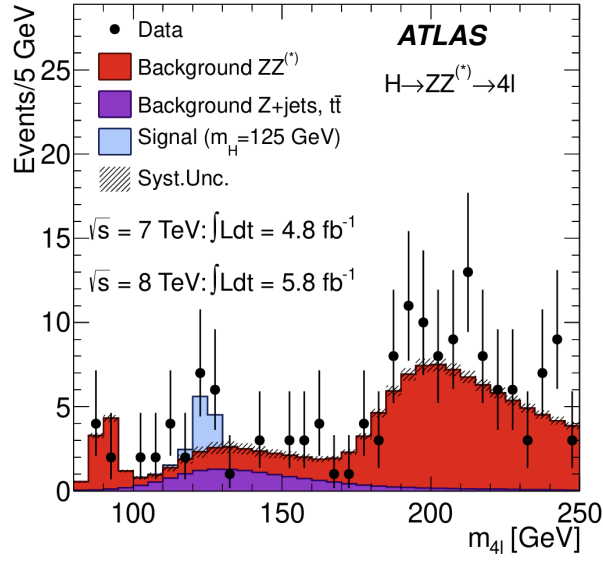


FIGURE 1.15: (a) The distribution of the four-lepton invariant mass, $m_{4\ell}$, for the selected candidates, compared to the background expectation in the 80–250 GeV mass range. The signal expectation for a SM Higgs with $m_H = 125$ GeV is also shown. Taken from [20].

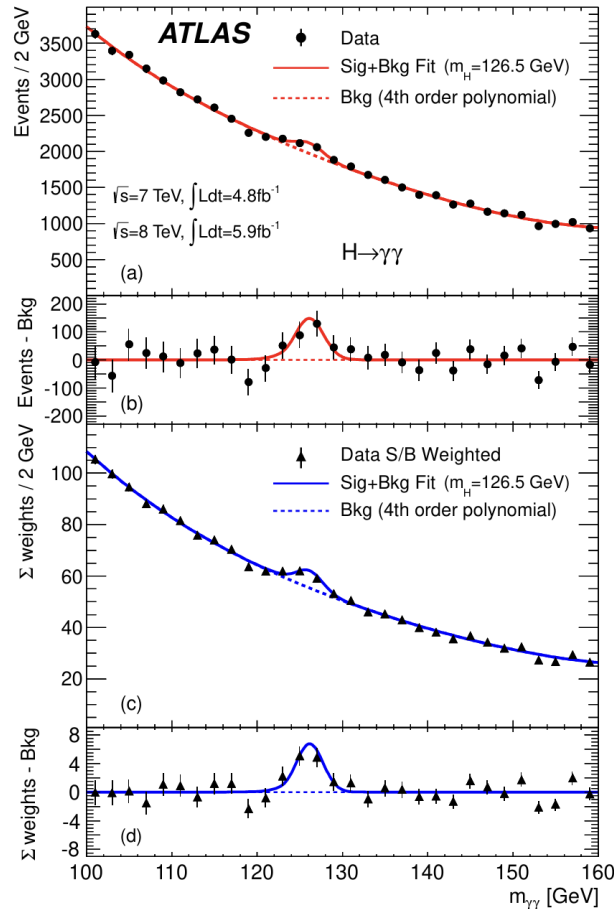


FIGURE 1.15: (b) The distributions of the invariant mass of diphoton candidates after all selections for the combined 7 TeV and 8 TeV data sample. The inclusive sample is shown in (a) and a weighted version of the same sample in (c). Taken from [20].

- $H \rightarrow WW^* \rightarrow e\nu\mu\nu$: this channel focuses on the invariant mass range of $m_H \in [100 \text{ GeV}, 200 \text{ GeV}]$. Its main signal signatures are two opposite-charge leptons with large transverse momentum and a large momentum imbalance in the event due to the escaping neutrinos. The main background to this channel is constituted of WW , $t\bar{t}$, and Wt production, all of which have real W pairs in the final state. Figure (1.15) depicts the transverse mass distribution for this channel, since the invariant mass is not available for reconstruction because of neutrino energy loss.

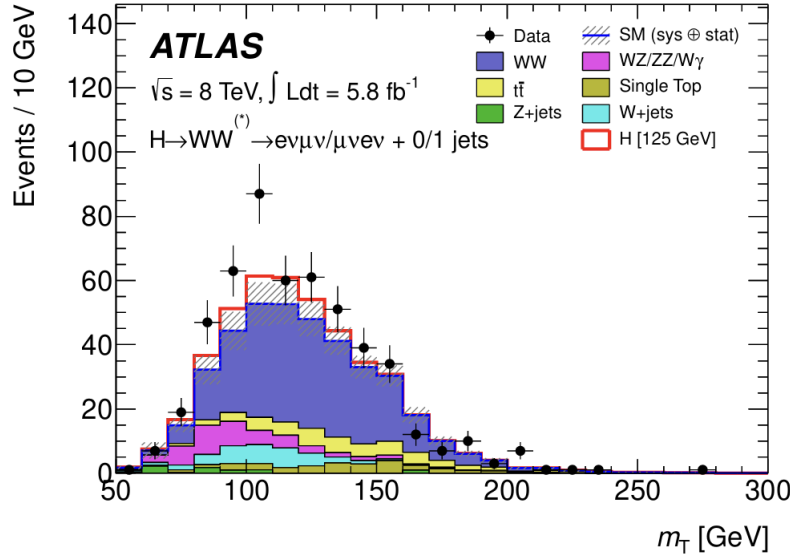


FIGURE 1.15: Distribution of the transverse mass m_T . The expected signal for $m_H = 125 \text{ GeV}$ is shown stacked on top of the background prediction. Taken from [20].

1.5.4 Higgs Boson Quantum Numbers

The Higgs boson is predicted to be a scalar particle with $J^P = 0^+$, with no color or electric charge. This is reflected in the fact that it is described by a real field, so it does not possess any kind of charge. We therefore expect to find that its spin number is zero, its parity to be positive, and its behavior under a CP transformation to be even.

There have been many studies done on the scalar nature of the resonance compatible with the Standard Model Higgs boson, including some that rejected a possible vector-like nature for this particle with a high confidence level. Even one of the main decay channels for the Higgs boson is a strong hint to its scalar nature: $H \rightarrow \gamma\gamma$. The Landau-Yang theorem¹¹ forbids vector particles to decay into pairs of photons. One of the most prolific

¹¹In quantum mechanics, the Landau–Yang theorem is a selection rule for particles that decay into two on-shell photons. The theorem states that a massive particle with spin 1 cannot decay into two photons.

decay channels for the study of spin is the $H \rightarrow ZZ^* \rightarrow 4\ell$ channel, which allows for the complete reconstruction of the kinematics of the process and of the variables that depend on spin. Another notable decay channel for these studies is $H \rightarrow WW^* \rightarrow \ell\nu\ell\nu$, where the distribution of the leptons and the missing energies is used for this study. Multiple hypothesis for J^P have been tested against total data, and the consensus is that with great confidence level the Higgs boson seems to be a $J^P = 0$ kind of particle. The 0^- hypothesis is rejected at 97.8% CL by using the $H \rightarrow ZZ^* \rightarrow 4\ell$ decay alone. The 1^+ and 1^- hypotheses are rejected with a CL of at least 99.7% by combining the $H \rightarrow ZZ^* \rightarrow 4\ell$ and the $H \rightarrow WW^* \rightarrow \ell\nu\ell\nu$ channels. Finally, the $J^P = 2^+$ model is rejected at more than 99.9% CL by combining all three bosonic channels, $H \rightarrow \gamma\gamma$, $H \rightarrow ZZ^* \rightarrow 4\ell$ and $H \rightarrow WW^* \rightarrow \ell\nu\ell\nu$. A strongly favored CP -even state is also analyzed and confirmed.

Section Roadmap:
Covered:

Explored the Higgs field potential and derived self-interaction terms to determine Higgs mass and couplings. Introduced collider phenomenology including Higgs production mechanisms (gluon fusion, vector boson fusion, associated production), Higgs decay channels (e.g., $H \rightarrow \gamma\gamma$, ZZ^* , WW^*), and how these processes are detected and interpreted experimentally. Higgs boson quantum numbers.

References:[14]; [15]; [20]; [19]; [21]; [22]; [23].



Chapter 2

Searches for The Higgs Boson Beyond the Standard Model

*“E ora che ne sarà
del mio viaggio?
Troppo accuratamente l’ho studiato
senza saperne nulla. Un imprevisto
è la sola speranza.”*

— Eugenio Montale, *Prima del viaggio*



DISCOVERING the Higgs boson in 2012 completed the particle content of the Standard Model and confirmed the mechanism of electroweak symmetry breaking. Despite this success, several observations—including dark matter, neutrino oscillations, the baryon asymmetry of the Universe, and the instability of the Higgs mass under radiative corrections—indicate that the SM is not a complete description of Nature. These open questions strongly motivate the exploration of extensions to the Higgs sector, where new scalar degrees of freedom could provide natural explanations for phenomena unresolved within the SM.

The Higgs boson is the only known fundamental scalar particle, and many of its properties, such as its couplings and rare decay modes, have not yet been measured with high precision. Even small deviations from SM expectations can yield powerful evidence of new physics. For this reason, models that expand the scalar sector in a minimal way have

attracted significant attention. Among these, the Two-Higgs- Doublet Model (2HDM) stands out as a simple and well-motivated framework capable of addressing several shortcomings of the SM while remaining compatible with current experimental constraints. By introducing a second scalar $SU(2)_L$ doublet, the 2HDM predicts an expanded Higgs spectrum, including a CP-odd pseudoscalar a , a heavier CP-even scalar H , and charged Higgs bosons $H^\pm$. Depending on the parameters, the pseudoscalar a can naturally be light, opening the possibility of exotic Higgs boson decays such as

$$h \rightarrow Za,$$

which are not expected in the Standard Model. If a has a mass in the range of a few hundred MeV to a few GeV, its decay products become hard to distinguish from the background, making its detection experimentally challenging and motivating dedicated reconstruction techniques. Such signatures can be investigated with the ATLAS detector at the LHC, providing a unique opportunity to probe extended scalar sectors.

The Chapter is structured as follows. We begin with a review of the shortcomings of the Standard Model that motivate the introduction of new physics. We then present the Type-1 2HDM, its scalar potential, and the resulting mass eigenstates. The phenomenological implications of the model are then discussed, including interactions with gauge bosons and possible exotic decay modes. We conclude by describing how these theoretical considerations motivate searches for light resonances in Higgs decays with the ATLAS detector, setting the stage for the analysis performed in the subsequent Chapters.

2.1 Shortcomings of the Standard Model

The Standard Model as it is is not enough to be a sufficient description of reality, and the Higgs boson, despite its essential role in the process of spontaneous symmetry breaking and mass term generation, is not a conclusive find for the physics community either. This particle has been an answer to a subset of questions, but plenty of other questions do not have an answer yet, and even more have arisen because of its discovery. In this Section our focus will be to address questions and problems in modern physics that still have no definitive answer, and effectively prove that the Standard Model as it is is not a complete theory.

2.1.1 Dark Matter

The physics community universally observes that only a small percentage of the whole universe is accessible to our observation¹ ($\approx 5\%$): the rest is composed of something we cannot quite grasp or experimentally analyze — we just know it is out there. This mysterious component of the universe is commonly referred to as dark matter and dark

¹We define the word "observation" in the physical sense: observable through experimental physical means.

energy, and it makes up the better part of the composition of the universe. One of the main culprits for this imbalance could be identified in weakly interactive massive particles: there exists no formal definition of a WIMP, but broadly speaking, it is an elementary particle which interacts via gravity and any other force (or forces) which is as weak as or weaker than the weak nuclear force. The Standard Model as it is does not produce any likely candidate for this kind of particle, so it falls short of providing an exhaustive description of reality on a fundamental level.

2.1.2 Gravity

Gravity is not included in the set of interactions contained in the Standard Model. There are no force mediators that are naturally present in it to represent said interaction. Moreover, gravity as described in the theory of General Relativity is highly different in formalism than any other Quantum Field Theory: the former has a classical formalism, while the latter uses a quantum one. This makes the gravitational interaction impossible to add to the Standard Model formalism, for they are fundamentally incompatible. The Standard Model, however, does not suffer from the lack of gravity, since this interaction is comparatively weak with respect to the other ones, one can neglect it *in toto*. The gravitational coupling constant, in natural units, is inversely proportional to the square of the Planck mass:

$$G_N \sim \frac{1}{M_{Pl}^2}, \quad M_{Pl} \approx 1.22 \times 10^{19} \text{GeV};$$

This implies that quantum gravitational effects are expected to become relevant only near the Planck scale, a regime vastly beyond the reach of current or foreseeable particle accelerators which probe up to the TeV scale.

2.1.3 Matter-Antimatter Asymmetry

It is a known fact that there is an imbalance between the amounts of matter and antimatter in the known universe, with the former strongly overcoming the latter, despite the fact that the Standard Model treats matter and antimatter on the same plane: this implies that there must exist a way to generate this imbalance. In 1967, Andrei Sakharov identified three necessary conditions for generating a baryon asymmetry from an initially symmetric universe:

1. Baryon number violation;
2. C and CP violation;
3. Departure from thermal equilibrium.

The Standard Model technically satisfies all of these three conditions, but the amount of CP violation in the Cabibbo–Kobayashi–Maskawa matrix² is not enough to explain such an imbalance.

2.1.4 Neutrino Masses

In the Standard Model, neutrinos are massless due to the lack of a right component for the field ν_R (it is therefore impossible to build a Yukawa-type Lagrangian), however, several experiments such as Super-Kamiokande, SNO, and KamLAND, have successfully shown that neutrinos do have a mass, and not only do they possess mass, but they come in three different mass eigenstates, which are a quantum superposition of massless flavor states. This is direct proof that the Standard Model is, once again, incomplete: these particles must then acquire their mass from a different mechanism that does not take into account interactions with the Higgs boson. The neutrino sector technically allows for the existence of neutrino particles that interact with none of the current fundamental interactions, calling them "sterile neutrinos". These can be identified as good candidates for dark matter.

2.1.5 Hierarchy Problem

The hierarchy problem is directly related to the role of the Higgs boson in the Standard Model, and will be the major premise to further development of BSM models in this Chapter. The Higgs potential for the electroweak model is:

$$V(H(x)) = m_H^2 |H(x)|^2 + \lambda |H(x)|^4.$$

In order to have spontaneous symmetry breaking in our model, we have required $\lambda > 0$ and $m_H^2 < 0$. These two conditions yield that the vacuum expectation value for the Higgs field is:

$$\langle 0 | H(x) | 0 \rangle = \sqrt{\frac{-m_H^2}{2\lambda}} = \frac{v}{\sqrt{2}},$$

implying that the vacuum state possesses a nonzero energy, *i.e.* the vacuum state lies on the ring of minima of the Mexican hat potential. From experimental discoveries we know that the Higgs mass is approximately 125 GeV and that the values of λ and m_H^2 are determined to be 0.126 and $-(92.9 \text{ GeV})^2$, respectively. A problem of this is that the Higgs mass receives huge corrections from any kind of interaction happening with the Higgs boson, so this implies that further analysis is required. If we consider a process where a Higgs boson propagator receives corrections from a fermion loop, we can consider the Feynman diagram in Figure (2.1).

²The CKM matrix describes quark mixing: it is a 3×3 matrix with complex components, such that $V^\dagger = V$. It inherently violates CP.

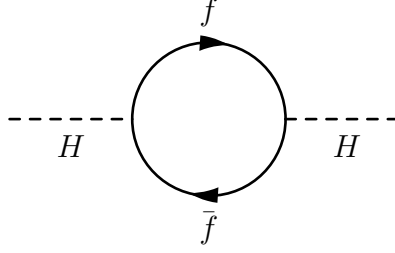


FIGURE 2.1: Higgs self-energy diagram with a fermion loop.

The corresponding mathematical expression, following Feynman rules, is given by integrating over all the possible momenta flowing in the loop. Calculating such an integral yields the following expression:

$$\begin{aligned}
 -i\Pi(p^2) &= (-1)(-i\lambda_f)^2 \int \frac{d^4k}{(2\pi)^4} \text{Tr} \left[\frac{i(\not{k} + m_f)}{k^2 - m_f^2 + i\varepsilon} \cdot \frac{i(\not{k} + \not{p} + m_f)}{(k+p)^2 - m_f^2 + i\varepsilon} \right] \\
 &= i\lambda_f^2 \int \frac{d^4k}{(2\pi)^4} \text{Tr} \left[\frac{(\not{k} + m_f)(\not{k} + \not{p} + m_f)}{(k^2 - m_f^2 + i\varepsilon)((k+p)^2 - m_f^2 + i\varepsilon)} \right]
 \end{aligned}$$

where:

1. λ_f is the Yukawa coupling of the fermion to the Higgs boson,
2. m_f is the fermion mass,
3. p is the external momentum of the Higgs boson.

This kind of diagram has a degree of divergence of two³, and through the renormalization procedure one finds a certain correction Δm_H^2 . Through the iteration of this one-loop diagram we can see the rise of a correction to the Higgs boson propagator. The Green's function has expression:

$$\Delta_H^{(0)}(p^2) = \frac{i}{p^2 - m_H^2 + i\varepsilon}.$$

³One can easily verify this through power counting:

$$i\Pi(p^2) \xrightarrow{k \rightarrow +\Lambda} \int^\Lambda d^4k \frac{k^2}{k^4} = \int^\Lambda dk \, k \sim \Lambda^2,$$

where Λ is the momentum cutoff, making the degree of divergence quadratic.

If we now want to consider all the possible corrections to the propagator that take into account one-loop corrections, we will build a series of Feynman diagrams, as below:

$$\begin{array}{c}
 \text{---} \\
 H
 \end{array}
 +
 \begin{array}{c}
 f \\
 \text{---} \bigcirc \text{---} \\
 \bar{f}
 \end{array}
 H
 +
 \begin{array}{c}
 f \quad f \\
 \text{---} \bigcirc \text{---} \bigcirc \text{---} \\
 \bar{f} \quad \bar{f}
 \end{array}
 H
 + \dots$$

This series of diagrams has the following mathematical expression:

$$\begin{aligned}
 \Delta_H^{(0)}(p^2) + \Delta_H^{(0)}(p^2)[-i\Pi(p^2)]\Delta_H^{(0)}(p^2) \\
 + \Delta_H^{(0)}(p^2)[-i\Pi(p^2)]\Delta_H^{(0)}(p^2)[-i\Pi(p^2)]\Delta_H^{(0)}(p^2) + \dots
 \end{aligned}$$

If we consider infinitely many one-loop corrections, we can write the sum of these as a power series for $\Delta_H^{(1)}(p^2)$:

$$\Delta_H^{(1)}(p^2) = \Delta_H^{(0)}(p^2) \sum_{n=0}^{\infty} \left[(-i\Pi(p^2))\Delta_H^{(0)}(p^2) \right]^n.$$

We can successfully perform the sum of the series⁴ to obtain:

$$\Delta_H^{(1)}(p^2) = \Delta_H^{(0)}(p^2) \sum_{n=0}^{\infty} \left[(-i\Pi(p^2))\Delta_H^{(0)}(p^2) \right]^n = \Delta_H^{(0)}(p^2) \left[\frac{1}{1 - (-i\Pi(p^2))\Delta_H^{(0)}(p^2)} \right].$$

Finally we manipulate this expression algebraically by substituting the propagator's expression, and we obtain:

$$\frac{i}{p^2 - m_H^2 + i\varepsilon} \left[\frac{p^2 - m_H^2 + i\varepsilon}{p^2 - m_H^2 + i\varepsilon - \Pi(p^2)} \right] = \frac{i}{p^2 - m_H^2 - \Pi(p^2) + i\varepsilon}.$$

With this simple calculation we have found that the correction given by the self-energy one-loop diagram moves the pole of propagator, effectively changing the mass⁵ of the Higgs boson. We can expand the self-energy correction around the physical mass of the particle to obtain the most accurate expression for the on-mass-shell propagator. This implies that:

$$\Pi(p^2) = \Pi(m^2) + (p^2 - m^2) \left. \frac{\partial \Pi(p^2)}{\partial p^2} \right|_{p^2=m^2},$$

⁴We do not worry about the fact that the contribution should be less than one because the “resummation” of self-energy insertions is formal: it is a consequence of how diagrams combine, not a statement about a convergent series. Its validity does not depend on the size of any single correction.

⁵In quantum field theory the mass of a particle is defined as the pole of the propagator.

therefore the total propagator becomes:

$$\frac{i}{p^2 - m_H^2 - \Pi(p^2) + i\varepsilon} = \frac{i}{(p^2 - m_H^2 + i\varepsilon)} \left[1 - \frac{\partial \Pi(p^2)}{\partial p^2} \Big|_{p^2=m_H^2} \right]^{-1},$$

where the term in square parentheses gets absorbed in the definition of the field⁶. The self-energy correction is generally a complex number, so it will be made up of a real and imaginary part. One defines the mass of the propagator as follows:

$$m_H^{r2} = m_H^2 - \text{Re}[\Pi(m_H^2)]$$

The imaginary part of the self-energy correction is related, via the optical theorem, to the decay width of the propagator:

$$i\text{Im}[\Pi(m_H^2)] = i\Gamma_H$$

The final expression for the renormalized propagator is then given by:

$$\Delta_H^{r(1)}(p^2) = \frac{i}{p^2 - m_H^{r2} + i\Gamma_H}.$$

We have obtained a correction to the mass of the propagator, and to its width — we can therefore take a look at this correction. After dimensional regularization of the integral, we obtain a term that contains an ultraviolet divergence Λ_{UV} . Highlighting this divergence, we write:

$$\Delta m_H^2 = -\frac{|\lambda_f|^2}{8\pi^2} \Lambda_{UV}^2 + \dots$$

This divergent term Λ_{UV}^2 can be interpreted as an energy-momentum cutoff where we expect the theory to break down and be affected by these corrections. Any fermion in this case can contribute since the mathematical form would be the same up to some factors to account for color, in the case of quarks, and other things. The largest correction is given by a top quark loop, where $\lambda_t = 0.94$. This seems all well, but when one considers that when the energies approach the Planck scale $\Lambda_{UV}^2 \approx M_P^2 \approx 2.4 \times 10^{18} \text{ GeV}^2$, the correction Δm_H^2 becomes thirty orders of magnitude larger than the required mass parameter value $m_H^2 \approx -(92.9)^2 \text{ GeV}^2$. This is a big problem, since all the massive particles inherit mass from the interaction with the Higgs boson: these corrections are enough to hinder the

⁶The derivative term in Z ,

$$Z^{-1} = \left[1 - \frac{\partial \Pi(p^2)}{\partial p^2} \Big|_{p^2=m_H^2} \right],$$

comes from the fact that the self-energy correction $\Pi(p^2)$ depends on momentum. This changes the shape (slope) of the propagator near the mass. The purpose of Z is to rescale the field so that the propagator near the physical mass looks like that of a free particle. Without this correction, the particle would not behave like a proper "one-particle state" in the theory.

accuracy of the Standard Model, so they have to be in some way reduced. This procedure is not possible inside the theoretical framework of the Standard Model, so we must look beyond. In other words, the Standard Model lacks any built-in mechanism to protect the Higgs mass from large radiative corrections. A natural solution would be the existence of additional fields or symmetries that can cancel or compensate for these divergent contributions. This motivates us to extend the scalar sector of the Standard Model in a minimal but physically meaningful way — for example, by introducing additional scalar degrees of freedom.

Suppose that there exists a scalar complex particle S that couples to the Higgs boson and has mass m_S with the following Lagrangian interaction term: $-\lambda_S|H|^2|S|^2$. The addition of this term implies an extra one-loop diagram correction to the Higgs boson propagator, as in Figure (2.2).

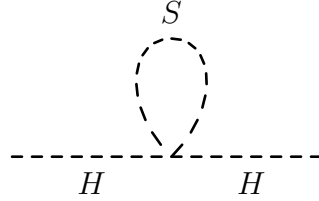


FIGURE 2.2: Higgs self-energy diagram with a scalar particle S loop.

If we repeat the same procedure as before we obtain a correction term that has this expression:

$$\Delta m_H^2 = \frac{\lambda_S}{16\pi^2} \left[\Lambda_{UV}^2 - 2m_S^2 \ln \left(\frac{\Lambda_{UV}}{m_S} \right) + \dots \right]$$

This clearly explains that the mass of the Higgs boson is highly sensitive to the particles it interacts with, no matter the particle. When m_S grows large, its contributions to Standard Model physics still remain relevant. Therefore it is not understood why the m_H parameter must be so small. From this we can start to think that either the Higgs boson does not, in fact, couple with all those particles, which would mean that the Standard Model has to be significantly reworked, or an extra option to save ourselves the trouble of sourcing another mass-acquisition mechanism is to consider that these sensible contributions to mass corrections in some way cancel out. This can be seen through a symmetry: what if there exists a symmetry that is able to fix all of our issues? What if we are able to consider an extra field doublet for which we introduce extra particles to make sure that the hierarchy problem is not one anymore? We will do just this in the following, by introducing another field doublet of scalar fields. The presence of an additional scalar doublet does not only address some of the conceptual limitations of the Standard Model, but also leads to a richer phenomenology. Notably, it modifies the Higgs self-interactions and opens new production and decay channels. Higgs boson pair production, in particular, becomes a key observable to probe the structure of the scalar potential and discriminate between different BSM scenarios.

Section Roadmap:

Covered: Identification of open problems in the SM that motivate extensions — including the lack of dark matter candidates, absence of gravity, baryon asymmetry, neutrino masses and oscillations, and the hierarchy problem related to Higgs mass corrections. Analysis of the quadratic divergences in Higgs mass corrections, their dependence on the cutoff scale, and motivation for BSM solutions such as extended scalar sectors to stabilize the electroweak scale.

Coming up: Deep-dive into the Type-1 2HDM and its foundations.

References: [24]; [25]; [26]; [7].



2.2 Beyond the Minimal Higgs Model: Type-1 2HDM

The Minimal Higgs model, treated up until now, assumed the existence of just one Higgs field doublet: but what if we were to add another one? We will justify our choices in the following, seeing how some issues find resolution in this simple extension.

To this avail, we start from an experimental fact. We can define a parameter that relates the masses of the gauge bosons of electroweak theory, ρ :

$$\rho \doteq \frac{m_W^2}{m_Z^2 \cos^2 \theta_W}.$$

This special quantity in the tree-level Minimal Standard Model is assumed to be precisely one. This is due to how the masses of the gauge bosons relate to the Higgs boson and to the coupling constants:

$$m_W = \frac{1}{2}gv \quad m_Z = \frac{1}{2}\sqrt{g^2 + g'^2}v \quad \cos \theta_W = \frac{g}{\sqrt{g^2 + g'^2}}.$$

Using these relations we can directly obtain:

$$\rho = \frac{m_W^2}{m_Z^2 \cos^2 \theta_W} = \frac{\frac{1}{4}g^2v^2}{\frac{1}{4}(g^2 + g'^2)v^2} \cdot \frac{g^2}{g^2} = 1.$$

Among the simplest and most studied extensions of the Higgs sector is the addition of a second scalar doublet. This choice is not arbitrary: it provides a natural setting to soften the hierarchy problem through new scalar interactions, while also leaving room for richer phenomenology such as new sources of CP violation, modified couplings, and potential dark matter candidates. Importantly, it can be constructed to respect key experimental constraints, such as the electroweak ρ parameter. What one does is measure this constant in experiments, and the result is that this special ratio becomes a strong constraint for our theory: any observable deviation from $\rho = 1$ would signal a departure from the minimal Higgs structure of the Standard Model, indicating the presence of new fields or interactions that modify the relation between the W and Z boson masses. This is important because the value $\rho = 1$ reflects an underlying custodial symmetry⁷ in the Higgs sector — a symmetry that ensures the consistency of electroweak mass generation. Violating this symmetry leads to conflicts with precision measurements. As a result, any extension of the Standard Model must either preserve custodial symmetry or suppress its breaking to a degree compatible with current experimental bounds. One such extension

⁷Custodial symmetry refers to an approximate global $SU(2)$ symmetry that emerges in the Standard Model Higgs sector when the hypercharge gauge coupling is neglected. It ensures that the masses of the W and Z bosons are related in such a way that the parameter ρ equals one at tree level. This symmetry arises when the Higgs potential respects an $SU(2)_L \otimes SU(2)_R$ symmetry that is spontaneously broken to the diagonal subgroup $SU(2)_V$. Extensions of the Higgs sector can break this symmetry explicitly, leading to deviations in ρ , which are strongly constrained by precision electroweak measurements.

is the class of Two Higgs Doublet Models (2HDMs), where a second scalar doublet is added to the theory. Crucially, the presence of an additional doublet does not, by itself, violate custodial symmetry — as long as the vacuum expectation values (VEVs) of the two doublets are aligned appropriately. This leaves room for richer scalar dynamics, additional CP violation, and new phenomena, while still satisfying the $\rho = 1$ constraint at tree level. However, misalignment between the VEVs or the inclusion of higher-dimensional scalar representations would typically break custodial symmetry and thus be tightly constrained by electroweak precision data. Due to the previous reasoning on the mass corrections to the Higgs propagator, we consider a theoretical model called the Type-1 2HDM, which is the model on which the following analysis is based. To consider this class of models, one can limit themselves to keeping into account the most simple of extensions: what we will do in the following is extend the Higgs sector of the Standard Model with an extra field doublet.

To formalize this extension, we introduce two scalar $SU(2)_L$ doublets, Φ_1 and Φ_2 . The dynamics of their interactions are encoded in the scalar potential, which determines the vacuum structure and the physical particle spectrum after electroweak symmetry breaking. The most general Lagrangian for this situation is given by:

$$\mathcal{L} = (D_\mu \Phi_1)^\dagger (D^\mu \Phi_1) + (D_\mu \Phi_2)^\dagger (D^\mu \Phi_2) - V(\Phi_1, \Phi_2),$$

where the V potential is:

$$\begin{aligned} V(\Phi_1, \Phi_2) = & m_{11}^2 \Phi_1^\dagger \Phi_1 + m_{22}^2 \Phi_2^\dagger \Phi_2 - m_{12}^2 (\Phi_1^\dagger \Phi_2 + \Phi_2^\dagger \Phi_1) + \frac{\lambda_1}{2} (\Phi_1^\dagger \Phi_1)^2 + \frac{\lambda_2}{2} (\Phi_2^\dagger \Phi_2)^2 \\ & + \lambda_3 \Phi_1^\dagger \Phi_1 \Phi_2^\dagger \Phi_2 + \lambda_4 \Phi_1^\dagger \Phi_2 \Phi_2^\dagger \Phi_1 + \frac{\lambda_5}{2} [(\Phi_1^\dagger \Phi_1)^2 + (\Phi_2^\dagger \Phi_2)^2]. \end{aligned}$$

One can find the vacuum state for this theory. Since all parameters are real, we find:

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

At this point, it is useful to parametrize the fields around their vacuum expectation values. This allows us to identify the physical degrees of freedom and the Goldstone bosons that give mass to the $W^\pm$ and Z bosons, as well as the remaining physical Higgs states. By expanding two complex doublet fields around the vacuum states we can find eight different fields:

$$\Phi_1 = \begin{pmatrix} \phi_1^+ \\ \frac{1}{\sqrt{2}} (v_1 + \rho_1 + i\eta_1) \end{pmatrix}, \quad \Phi_2 = \begin{pmatrix} \phi_2^+ \\ \frac{1}{\sqrt{2}} (v_2 + \rho_2 + i\eta_2) \end{pmatrix}.$$

As we have covered in Chapter 1, we can see that three of these fields get "gauged" out to yield mass terms for the $Z_0, W^\pm$ bosons, and the remaining five are physical scalar "Higgs" fields. There are two charged scalars ($H^\pm$), two neutral scalars (h, H), and one pseudoscalar (a). With the above minimum, the mass terms for the charged scalars are

given by this mass matrix:

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

The mass terms for the pseudoscalars are given by:

$$\mathcal{L}_\eta = \frac{m_a^2}{v_1^2 + v_2^2} \begin{pmatrix} \eta_1 & \eta_2 \end{pmatrix} \begin{pmatrix} v_2^2 & -v_1v_2 \\ -v_1v_2 & v_1^2 \end{pmatrix} \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix}$$

Finally, the mass terms for the scalars are given by:

$$\mathcal{L}_\rho = - \begin{pmatrix} \rho_1 & \rho_2 \end{pmatrix} \begin{pmatrix} m_{12}^2 \frac{v_1}{v_2} + \lambda_1 v_1^2 & -m_{12}^2 + (\lambda_3 + \lambda_4 + \lambda_5)v_1v_2 \\ -m_{12}^2 + (\lambda_3 + \lambda_4 + \lambda_5)v_1v_2 & m_{12}^2 \frac{v_1}{v_2} + \lambda_2 v_2^2 \end{pmatrix} \begin{pmatrix} \rho_1 \\ \rho_2 \end{pmatrix}.$$

All of these matrices can be diagonalized via linear combinations of the fields, introducing two angles, α and β . If one defines:

$$v_1 = v \cos \beta \quad v_2 = v \sin \beta,$$

where v is the Standard Model vacuum expectation value. From these definitions one finds that:

$$\tan \beta = \frac{v_2}{v_1}$$

one can find that the physical fields are mixed versions of the previous ones. The angle β characterizes the ratio of the vacuum expectation values of the two doublets and controls how electroweak symmetry breaking is shared between them. The angle α mixes the CP-even components of the two doublets, determining how the physical Higgs states align with the Standard Model Higgs boson. The Goldstone boson is G^0 :

$$G^0 = \eta_1 \cos \beta + \eta_2 \sin \beta.$$

The linear combination of η fields is the pseudoscalar particle a :

$$a = \eta_1 \sin \beta - \eta_2 \cos \beta.$$

Furthermore, this model allows the existence of two Higgs scalars, one lighter, h , and one heavier, H :

$$\begin{aligned} h &= \rho_1 \sin \alpha - \rho_2 \cos \alpha \\ H &= -\rho_1 \cos \alpha - \rho_2 \sin \alpha, \end{aligned}$$

The Standard Model Higgs boson is given by:

$$h_{SM} = \sin(\beta - \alpha)h + \cos(\beta - \alpha)H$$

It is important to mention the Higgs alignment limit, where $h = h_{SM}$. This is obtained by considering that $\sin(\beta - \alpha) = 1$, and this limit the following interaction vertices are forbidden: $H \rightarrow WW/ZZ$, $a \rightarrow Zh$, and $H^\pm \rightarrow W^\pm h$.

This extended scalar sector predicts a richer phenomenology compared to the Standard Model. In addition to the light scalar h (which in the alignment limit behaves like the observed 125 GeV Higgs boson), we obtain a heavier neutral scalar H , a pseudoscalar a , and a pair of charged scalars $H^\pm$. These states can lead to distinctive collider signatures, making 2HDM scenarios directly testable at experiments such as the LHC. This model allows us to predict the existence of a theoretical pseudoscalar particle a which will be the subject of the following Sections. The scalars that we have been able to add with this model are effective in helping fix the hierarchy problem mentioned in Section (1.1.5).

Section Roadmap:

Covered: Introduction of the Two Higgs Doublet Model (2HDM) as a minimal scalar-sector extension preserving custodial symmetry and the $\rho = 1$ condition. Derivation of the scalar potential, vacuum expectation values, field expansions, and mass eigenstates ($h, H, H^\pm, A$) with mixing angles α and β .

Coming up: Explanation of the Type-1 2HDM phenomenology.

References: [27]; [3]; [28].



2.3 Phenomenology of the Type-1 2HDM

Having established the structure of the scalar sector, we now turn to its phenomenological implications. The presence of new scalar degrees of freedom modifies existing interactions and introduces new ones, leading to potentially measurable deviations from Standard Model expectations. The introduction of new particles gives rise to new extra interaction vertices. One uses the potential seen in the previous section to apply the gauge principle and spontaneous symmetry breaking as one does for the Standard Model. In this case, however, there are two field doublets whose $SU(2)_L \otimes U(1)_Y$ invariant Lagrangian is:

$$\mathcal{L}_{\text{Higgs}} = \sum_{i=1,2} (D_\mu \Phi_i)^\dagger (D^\mu \Phi_i) - V(\Phi_i),$$

where the covariant derivative is:

$$D_\mu = \partial_\mu - ig_2 \frac{\tau^a}{2} W_\mu^a + ig_1 \frac{Y(\Phi_i)}{2} B_\mu$$

If we now consider the terms coming from the covariant derivatives we find the following interaction Lagrangian between the Type-1 2HDM fields and the gauge bosons $W^\pm$, Z_0 , and A_μ , which give rise to the phenomenology for this theory:

$$\begin{aligned} \mathcal{L}_{\text{int}} = & \left(g_2 m_W W_\mu^+ W^{\mu-} + \frac{1}{2} \frac{g_2}{\cos \theta_W} m_Z Z_\mu Z^\mu \right) \left[\sin(\beta - \alpha) h + \cos(\beta - \alpha) H \right] \\ & + i \frac{g_2}{2} \left\{ \cos(\beta - \alpha) W_\mu^+ (h \partial^\mu H^- - H^- \partial^\mu h) - \sin(\beta - \alpha) W_\mu^+ (H \partial^\mu H^- - H^- \partial^\mu H) \right. \\ & \quad \left. - \cos(\beta - \alpha) W_\mu^- (h \partial^\mu H^+ - H^+ \partial^\mu h) + \sin(\beta - \alpha) W_\mu^- (H \partial^\mu H^+ - H^+ \partial^\mu H) \right\} \\ & - \frac{g_2}{2} \left[W_\mu^+ (H^- \partial^\mu A - A \partial^\mu H^-) + W_\mu^- (H^+ \partial^\mu A - A \partial^\mu H^+) \right] \\ & + i \left[e A_\mu + \frac{g_2}{2 \cos \theta_W} (\sin^2 \theta_W - \cos^2 \theta_W) Z_\mu \right] \left[H^+ \partial^\mu H^- - H^- \partial^\mu H^+ \right] \\ & + \frac{g_2}{2 \cos \theta_W} Z_\mu \left[\cos(\beta - \alpha) (A \partial^\mu h - h \partial^\mu A) - \sin(\beta - \alpha) (A \partial^\mu H - H \partial^\mu A) \right]. \end{aligned}$$

We are interested in looking at the phenomenology, so we will look into the allowed Feynman diagrams for this model — we will break each line down into the single interaction vertices.

The interaction terms are obtained by multiplying terms and are listed in Table (2.1) and Table (2.2).

One notices that the phenomenology of this model is much richer than before, and it is testable at colliders. As a concrete illustration of how extended scalar sectors — such as

Interaction term	Feynman diagram	Coupling
$g_2 m_W \sin(\beta - \alpha) W_\mu^+ W_-^\mu h$		$g_2 m_W \sin(\beta - \alpha)$
$g_2 m_W \cos(\beta - \alpha) W_\mu^+ W_-^\mu H$		$g_2 m_W \cos(\beta - \alpha)$
$\frac{g_2}{2} m_Z \sin(\beta - \alpha) Z_\mu Z^\mu h$		$\frac{g_2}{2} m_Z \sin(\beta - \alpha)$
$\frac{g_2}{2} m_Z \cos(\beta - \alpha) Z_\mu Z^\mu H$		$\frac{g_2}{2} m_Z \cos(\beta - \alpha)$
$i \frac{g_2}{2} \cos(\beta - \alpha) W_\mu^+ \times (h \partial^\mu H^- - H^- \partial^\mu h)$		$i \frac{g_2}{2} \cos(\beta - \alpha)$
$-i \frac{g_2}{2} \sin(\beta - \alpha) W_\mu^+ \times (H \partial^\mu H^- - H^- \partial^\mu H)$		$-i \frac{g_2}{2} \sin(\beta - \alpha)$
$-i \frac{g_2}{2} \cos(\beta - \alpha) W_\mu^- \times (h \partial^\mu H^+ - H^+ \partial^\mu h)$		$-i \frac{g_2}{2} \cos(\beta - \alpha)$

TABLE 2.1: Gauge and scalar interaction vertices for the Type-1 2HDM, part 1.

Interaction term	Feynman diagram	Coupling
$i\frac{g_2}{2}\sin(\beta-\alpha)W_\mu^- \times (H\partial^\mu H^+ - H^+\partial^\mu H)$	A vertex where a dashed line labeled H^+ enters from the top left, a dashed line labeled H exits to the top right, and a wavy line labeled W^- exits to the bottom left.	$i\frac{g_2}{2}\sin(\beta-\alpha)$
$-\frac{g_2}{2}W_\mu^+(H^-\partial^\mu A - A\partial^\mu H^-)$	A vertex where a dashed line labeled H^- enters from the top left, a dashed line labeled a exits to the top right, and a wavy line labeled W^+ exits to the bottom left.	$-\frac{g_2}{2}$
$-\frac{g_2}{2}W_\mu^-(H^+\partial^\mu A - A\partial^\mu H^+)$	A vertex where a dashed line labeled H^+ enters from the top left, a dashed line labeled H exits to the top right, and a wavy line labeled W^- exits to the bottom left.	$-\frac{g_2}{2}$
$ieA_\mu(H^+\partial^\mu H^- - H^-\partial^\mu H^+)$	A vertex where a dashed line labeled H^- enters from the top left, a dashed line labeled H^+ exits to the bottom left, and a wavy line labeled γ exits to the top right.	ie
$i\frac{g_2}{2\cos\theta_W}(\sin^2\theta_W - \cos^2\theta_W)Z_\mu \times (H^+\partial^\mu H^- - H^-\partial^\mu H^+)$	A vertex where a dashed line labeled H^- enters from the top left, a dashed line labeled H^+ exits to the bottom left, and a wavy line labeled Z exits to the top right.	$\frac{ig_2}{2\cos\theta_W} \times (\sin^2\theta_W - \cos^2\theta_W)$
$\frac{g_2\cos(\beta-\alpha)}{2\cos\theta_W}Z_\mu \times (A\partial^\mu h - h\partial^\mu A)$	A vertex where a dashed line labeled h enters from the top left, a dashed line labeled a exits to the bottom left, and a wavy line labeled Z exits to the top right.	$\frac{g_2\cos(\beta-\alpha)}{2\cos\theta_W}$
$\frac{g_2\sin(\beta-\alpha)}{2\cos\theta_W}Z_\mu \times (A\partial^\mu H - H\partial^\mu A)$	A vertex where a dashed line labeled a enters from the top left, a dashed line labeled H exits to the bottom left, and a wavy line labeled Z exits to the top right.	$\frac{g_2\sin(\beta-\alpha)}{2\cos\theta_W}$

TABLE 2.2: Gauge and scalar interaction vertices for the Type-1 2HDM, part 2.

the Type-1 2HDM discussed above — may manifest in exotic Higgs decays, one might consider the following decay channels:

$$h \rightarrow Za \quad \text{and} \quad h \rightarrow Z\eta_c$$

where h is the discovered ~ 125 GeV Higgs boson, a is a light pseudoscalar, and η_c is a quarkonium-like resonance. In the following Section we will therefore look into how these searches have been performed experimentally from a data analysis standpoint.

Section Roadmap:

Covered: Exploration of the interaction structure and phenomenological implications of the 2HDM, including couplings among new scalar states and gauge bosons. Tables of interaction vertices illustrate the extended Higgs–gauge dynamics.

Coming up: Experimental methods to probe the model.

References: [27]; [3]; [28]; [29].



Chapter 3

Tools for The Analysis

*“All the king’s horses and all the king’s men /
Couldn’t put Humpty together again.”*

— Anon., *Nursery Rhyme*



JETS are among the most abundant and information-rich objects produced in high-energy collisions at the LHC, and they consist in the physical manifestation of particles inside the detector — this makes them the target of the physics’ community observation. Arising from the fragmentation and hadronization of quarks and gluons produced in the hard scattering, they encode the dynamics of Quantum Chromodynamics (QCD) across a wide range of energy scales. Despite their complexity, a precise understanding of jet formation — from the perturbative emission of radiation to the non-perturbative transition into stable hadrons — remains an active area of both theoretical and experimental investigation, and it is of special interest for this work.

The internal structure of a jet carries a wealth of information about the underlying hard process, decay, and the subsequent parton shower. Jet substructure variables, constructed to be sensitive to the angular and momentum distributions of constituents within a jet, have emerged as powerful discriminants between jets originating from quarks, gluons, or massive boosted particles. Their proper theoretical description, however, requires accurate modeling of both perturbative QCD effects and the poorly constrained non-perturbative corrections that dominate at low energy scales. Two of the Monte Carlo tools used in this work, PYTHIA and HERWIG, implement distinct approaches to parton showering and

hadronization, leading to systematic differences in their predictions that must be carefully understood.

Recent advances in machine learning have opened new avenues for jet classification and reconstruction. Neural network architectures trained directly on jet constituent information are capable of exploiting correlations among substructure observables that are inaccessible to traditional cut-based approaches, offering improved discrimination power in a variety of physics analyses.

This Chapter provides the theoretical and technical foundations for the study presented in subsequent Chapter. This begins with a description of the experimental apparatus used, and continues on by motivating the central role of jet physics in collider phenomenology. Jet reconstruction algorithms are presented, followed by the key jet substructure variables of interest in this case study and the neural network approaches used, too. Monte Carlo generators PYTHIA and HERWIG are then described, with particular attention to their differing treatments of the parton shower and hadronization. The Chapter ends with a description of the analysis setup and computational pipeline for the studies carried out in this work.

3.1 The Experimental Setting

The search for $H \rightarrow Za$ decays is ultimately an experimental endeavor: however well-motivated theoretically, the signal must be extracted from real collision data recorded by a physical detector, and the simulation chain used to model it must faithfully reproduce the conditions under which that data is collected. Before turning to the theoretical and algorithmic tools of the analysis, it is therefore necessary to describe the experimental apparatus and the methods through which a particle physics measurement is carried out in practice.

This section introduces the Large Hadron Collider and the ATLAS detector, the two experimental devices at the heart of this work. It then outlines the general strategy of a particle physics analysis — from hard-scatter simulation to reconstructed physics objects — and closes with a concrete example of Monte Carlo event generation that will serve as a reference throughout the following sections.

3.1.1 The Large Hadron Collider

The LHC is the final stage of a complex accelerator chain at CERN, where protons¹ are gradually accelerated to the extreme energies needed for high-energy collisions. They begin in a linear accelerator, then pass through a series of booster synchrotrons —including the Proton Synchrotron (PS) and the Super Proton Synchrotron (SPS) — each progressively increasing their energy and intensity. By the time the protons enter the 27 km LHC ring, shown schematically in Figure (3.1), they are moving at nearly the speed

¹Protons are not the only particles that are accelerated in the ring, but they are the ones of interest in this case. LHC accelerates lead nuclei, too, for example.

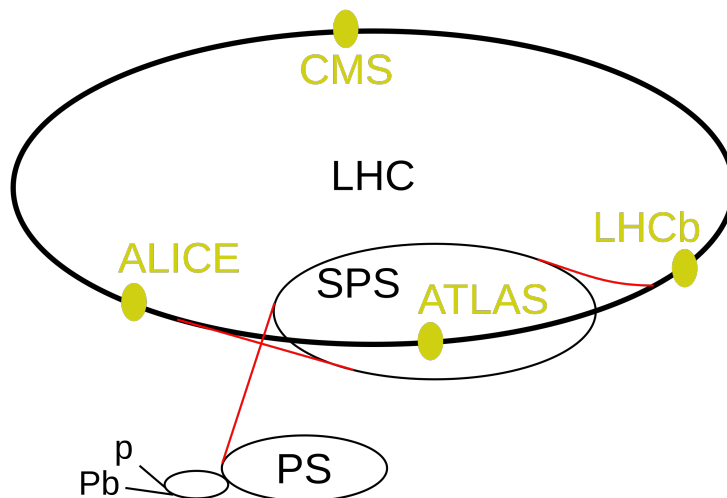


FIGURE 3.1: Scheme of the accelerator machines that feed into the LHC beampipe. Courtesy of [30].

of light, ready to collide at the four main experimental points. This staged approach ensures stable, high-energy collisions that allow experiments like ATLAS to probe rare processes, such as exotic Higgs boson decays, with unprecedented precision. We will focus our attention on giving out a couple of technical details about the ATLAS experiment and its reconstruction methods, as well as the methods of a physics analysis at a particle physics experiment.

3.1.2 The ATLAS Experiment at CERN

The ATLAS experiment at CERN, depicted in Figure (3.2), is a multipurpose particle detector with a forward-backward symmetrical cylindrical geometry and a near 4π coverage in solid angle: this implies that it is capable of detecting and classifying almost all of the different particles that pass through it. Its design allows physicists to reconstruct the four-momenta and properties of particles produced in high-energy collisions with great precision.

The ATLAS experiment is one of four experiments that lie along the LHC beampipe, along with ALICE, CMS, and LHCb. Together, these experiments form a complementary suite of detectors that explore different aspects of high-energy physics phenomena. The joint goal of collider and detector experiments is to study high energy physics interactions: LHC is the main collider experiment, accelerating proton packets through a chain of increasingly larger machines up to 13.6 TeV and leading them to collide head-on at the crossing points, where the aforementioned experiments are located.

ATLAS consists of an inner tracker surrounded by a solenoid capable of delivering a magnetic field of 2 T, electromagnetic and hadronic calorimeters, muon chambers, and a toroidal magnetic field. These layers work together to identify and measure the energy and

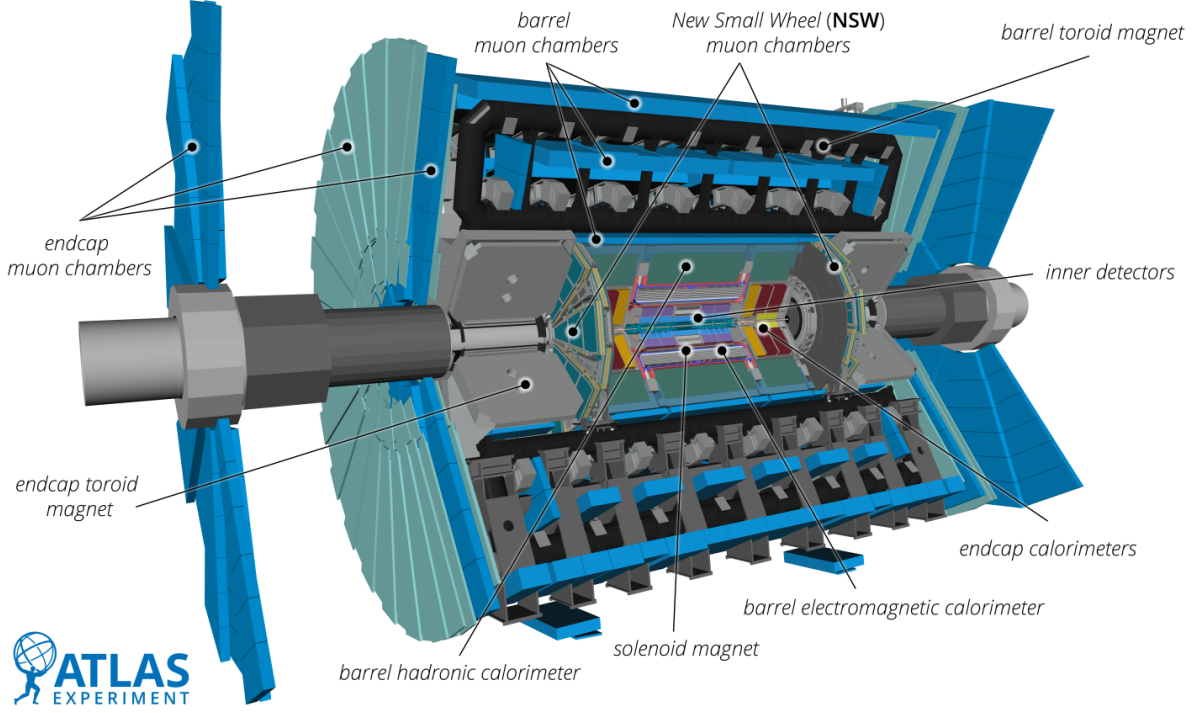


FIGURE 3.2: The ATLAS experimental apparatus in its 3D rendition. Courtesy of [31].

momentum of particles produced in collisions, from the innermost charged tracks to the outermost muons. It uses a right-handed coordinate system with the origin at the nominal interaction point at the center of the detector and with the z -axis along the beam pipe. The x -axis points to the center of the LHC ring, while the y -axis points upwards. One can also use polar coordinates for the plane transverse to the beam pipe: this special set of coordinates allows for an easier representation of the important kinematical variables in the experiment, such as pseudorapidity η and angular distance ΔR :

$$\eta = -\ln\left(\tan\frac{\theta}{2}\right) \quad \Delta R = \sqrt{(\Delta y)^2 + (\Delta\phi)^2}$$

These quantities are particularly useful when comparing angular separations and rapidity distributions of particles produced in high-energy collisions. Figure (3.3) shows a graphical representation of the coordinate system used at the ATLAS experiment.

The ATLAS detector relies on sophisticated reconstruction algorithms to turn raw collision data into meaningful physics objects. Charged particle tracks are first reconstructed in the inner detector using hits from the silicon pixel and strip detectors, as well as the transition radiation tracker, which together allow precise measurement of particle trajectories and momenta. Energy deposits in the electromagnetic and hadronic calorimeters

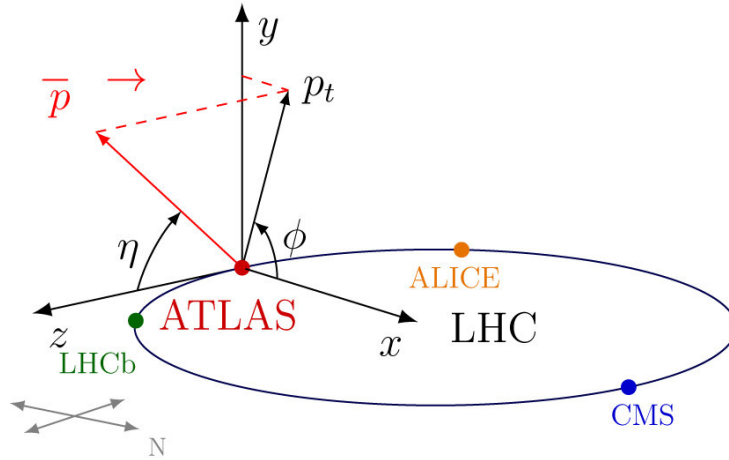


FIGURE 3.3: Illustration of the ATLAS coordinate system(s). Taken from [32].

are clustered to identify electrons, photons, and jets, while sophisticated algorithms separate overlapping signals from nearby particles. Muons are reconstructed by combining tracks in the inner detector with tracks in the outer muon spectrometer, which operates in the toroidal magnetic field, allowing precise momentum measurements even at high energies. The reconstruction also includes vertex identification, crucial for distinguishing the primary collision from multiple simultaneous interactions, known as pileup², and the calculation of missing transverse energy, which provides information about undetected particles such as neutrinos. By combining all these elements, ATLAS is able to reconstruct complete collision events with high accuracy, forming the foundation for subsequent physics analyses.

The physics objects relevant to this analysis — electrons, jets, and muons — each leave a characteristic signature across the detector layers, as seen in Figure (3.4). Electrons are identified by a charged track in the inner detector matched to a compact, localized energy deposit in the electromagnetic calorimeter, with the shower shape used to discriminate them from photons and hadronic activity. Jets are reconstructed by clustering energy deposits in both the electromagnetic and hadronic calorimeters, with inner detector tracks matched to the cluster to provide additional information on the charged component of the jet and to associate it with a primary vertex. Muons, being minimum-ionizing particles, deposit very little energy in the calorimeters; they are instead identified by combining a charged track in the inner detector with a matching track in the muon spectrometer, which operates in the toroidal magnetic field and allows precise momentum measurements up to the TeV scale. The interplay between these subsystems underpins the full reconstruction of the $H \rightarrow Za$ decay topology studied in this work.

²Pileup refers to additional proton-proton interactions occurring in the same or nearby bunch crossings as the primary event of interest. These extra collisions can produce additional tracks and energy deposits, complicating the reconstruction of the main event.

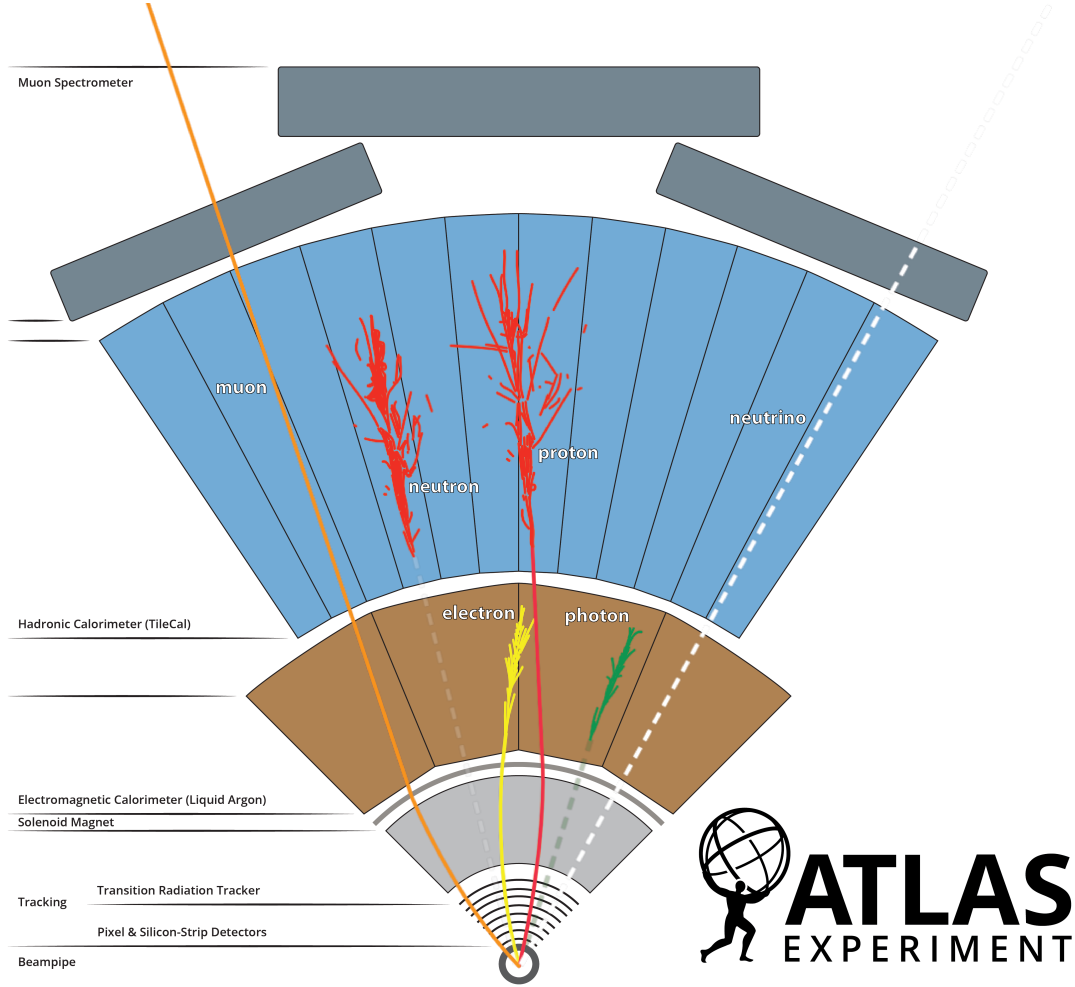


FIGURE 3.4: A slice of ATLAS as well as visualizations of how ATLAS detects different particles. Taken from [33].

3.1.3 Charged Particle Tracking and Jet Inputs at ATLAS

The reconstruction of charged particle trajectories is the first step in building the physics objects relevant to this analysis. The ATLAS inner detector covers $|\eta| < 2.5$ and provides sufficient precision hits per track to determine the five helix parameters (d_0 , z_0 , ϕ_0 , θ , q/p), where d_0 and z_0 are the transverse and longitudinal impact parameters with respect to the primary vertex, and q/p encodes the signed inverse momentum in the solenoidal field of 2 T. The hard-scatter primary vertex is identified as the reconstructed vertex with the largest $\sum p_T^2$ of associated tracks, and only tracks consistent with it are retained for jet substructure calculations, providing inherent robustness against pileup contamination. For jet reconstruction, the primary calorimeter inputs are topo-clusters, three-dimensional

energy deposits grown from seeded calorimeter cells. However, for the substructure measurements central to this analysis, the calorimeter angular granularity is too coarse to resolve the closely-spaced decay products of a highly boosted light resonance, and charged particle tracks ghost-associated to the jet become the primary source of information. The theoretical and algorithmic content of this statement — what ghost association is, why it is infrared safe, and what substructure variables are built from these tracks — is developed in detail in the sections that follow. What those tracks are ultimately used to find is the subject of the next section.

3.1.4 Analysis Motivation

Ever since its discovery, the Higgs boson and its sector have been extensively studied via the aforementioned experiments, and it has been found to be consistent with Standard Model predictions. The limited precision on its decay width of 4.1 GeV against a mass of approximately 125 GeV allow for significant contributions from BSM decays. Since the decay width is several orders of magnitude lower than its mass, it is possible to have a small coupling to a non-SM particle and new decay modes: one cannot rule out decays of the Higgs boson in light resonances, such as unflavored mesons and quarkonium states, namely J/ψ and the η_c . Higgs boson decays into elementary light resonances are predicted by many extensions of the Standard Model — one of those is the one that has been covered in Section (1.2) and (1.3). If we consider the Higgs boson to be close to the alignment limit, so that $\sin(\alpha - \beta) \approx 1$, we can comfortably accommodate a light CP-odd scalar (a pseudoscalar) particle with a mass of a few GeV. Both ATLAS and CMS have published several results, most of them focusing on mass ranges of $\mathcal{O}(1 - 10 \text{ GeV})$, pairs of a -resonances, and leptonic or photonic decays of said a -resonance. A light scalar resonance originating from a Higgs boson can also be interpreted within the framework of the Standard Model: light quark meson resonances composed of $c\bar{c}$ constitute background processes for the search of the q particle. These processes are hard to study since the branching ratios for the exclusive Higgs boson decays are of $\mathcal{O}(10^{-5})$ or less, therefore the expected yield of event counts is heavily reduced. Some of these ratios are presented in Table (3.1).

Rare decays of the Higgs boson into light elementary or composite resonances are essential towards the betterment of precision measurements of properties of the boson, as well as searches for new physics. In the following analysis hadronic decays of the light scalar particle are going to be studied since they constitute more than 50% of the branching ratio³. This boost-induced collimation is quantified by the characteristic angular separation of the decay products:

$$\Delta R \sim \frac{2m_a}{p_T^a},$$

which for $m_a \lesssim 4 \text{ GeV}$ and $p_T^a \sim 100 \text{ GeV}$ yields $\Delta R \lesssim 0.08$. This is smaller than the calorimeter tower spacing in the hadronic calorimeter, so no angular separation between

³As an example, $\mathcal{BR}(\eta_c \rightarrow \text{hadrons}) \approx 52\%$.

Decay Mode	Branching Ratio $\mathcal{BR} \doteq \frac{\Gamma_i}{\Gamma_{tot}}$
$H \rightarrow b\bar{b}$	0.577
$H \rightarrow \tau^+\tau^-$	0.062
$H \rightarrow W^+W^-$	0.215
$H \rightarrow ZZ$	0.027
$H \rightarrow \gamma\gamma$	0.00227
$H \rightarrow gg$	0.085
$H \rightarrow \mu^+\mu^-$	0.000022
$H \rightarrow Z\eta_c$	$\simeq 1.4 \times 10^{-5}$
$H \rightarrow \rho^0\gamma$	$\simeq 1.68 \times 10^{-5}$
$H \rightarrow J/\psi\gamma$	$\simeq 2.95 \times 10^{-6}$

TABLE 3.1: Branching ratios of the Standard Model Higgs boson at $m_H = 125$ GeV, including rare decays into charmonium and light resonances. The last three lines are theoretical estimations. The others are taken from [34].

the a decay products is resolvable in the calorimeter: they merge into a single topo-cluster and are reconstructed as one anti- k_T jet of radius $R = 0.4$. The inner detector, with its $\mathcal{O}(10 \mu\text{m})$ hit resolution, can in principle resolve the individual tracks from the two prongs, but at $m_a \sim 1$ GeV and the relevant boosts even this becomes marginal. The practical consequence is that the analysis is forced to work entirely with the internal structure of this single merged jet, using the p_T -weighted angular moments and prong-counting variables of the following sections, computed from ghost-associated tracks, rather than with the multi-jet kinematics that would be accessible for heavier, less boosted resonances.

3.1.5 Monte Carlo Simulation

A clear explanation of the general methods for a particle physics analysis should be provided. The search for $H \rightarrow Za$ decays goes through a few steps of simulation. The objective of Monte Carlo simulations is to emulate the real sets of data obtained in the experiment and its response, providing a testable and reproducible benchmark. The simulation of a particle physics experiment data are obtained through a few steps of simulation, constituting a workflow. The first step in this workflow is the simulation and computation of hard scattering between the two colliding protons: for this, the partonic

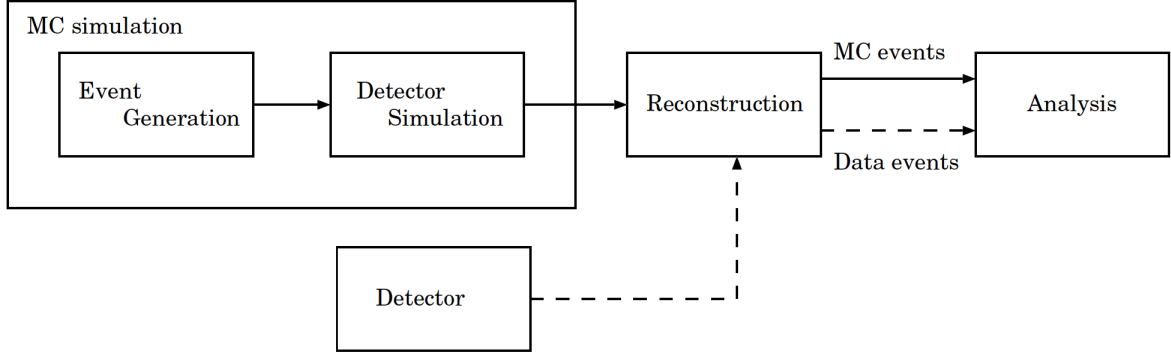


FIGURE 3.5: The flow of the production of MC events and real data for data analysis. The reconstruction step should work for both MC events (solid) and real data (dashed). Courtesy of [35].

cross section we have mentioned in Chapter 1 is employed:

$$\begin{aligned}
 \sigma_{pp \rightarrow X}(S) &= \sum_{i,j} \int_0^1 dx_1 \int_0^1 dx_2 f_i(x_1, \mu_F^2) f_j(x_2, \mu_F^2) \hat{\sigma}_{ij \rightarrow X}(\hat{s} = x_1 x_2 S, \mu_F^2, \mu_R^2), \\
 &= \sum_{i,j} \int_0^1 dx_1 \int_0^1 dx_2 f_i(x_1, \mu_F^2) f_j(x_2, \mu_F^2) \sum_{k=0}^{\infty} \int d\Phi_{X+k} \left| \sum_{\ell=0}^{\infty} \mathcal{M}_{X+k}^{(\ell)} \right|^2.
 \end{aligned}$$

One has that the sum over k spans over all the initial state or final state quark/gluon emission. We call these Initial State Radiation (ISR) or Final State Radiation (FSR), respectively. The ℓ index spans all the possible loop corrections to the process, instead. Therefore, $\mathcal{M}_{X+k}^{(\ell)}$ describes the matrix element of $X + k$ particle production at the ℓ -th loop order with a renormalization scale and a PDF choice. Since the simulated events are one-on-one parton interactions, the spectator partons can also interact, generating what are referred to as underlying events. The next step in the process is the hadronization step, where the products from hard scattering simulations hadronize in primary hadrons, which can furthermore decay. The code simulates the decay of these unstable particles into stable particles. The interaction of the stable particles with the detector then is mediated by another code, Geant4⁴. After the interaction of stable particles with the detector, the workflow simulates the detector reconstruction step. What one obtains is a set of data in the same format as the output of a detector experiment. It is possible to use the data just before detector interaction to be able to do analysis: these samples are called "truth-level" or "particle level" simulations, and this is useful for some studies, but having a full set of data simulating the detector is fundamental for real data comparison. Our study, for ease of computation will use truth-level particles.

⁴Toolkit for the simulation of the passage of particles through matter. Its areas of application include high energy, nuclear and accelerator physics, as well as studies in medical and space science.

Figure (3.5) describes a simplified simulation workflow. The Monte Carlo simulation steps are used to fabricate real data coming from the detector and are fed to the reconstruction algorithm in order to do an analysis. Real detector data follow the same workflow: the result is the only test for new theories and particles, since the Monte Carlo part is something we can easily reproduce and modify based on what we are interested in testing. The sets coming from Monte Carlo simulations are used in later steps of the analysis: a detailed comparison of the two generators used — PYTHIA 8 and HERWIG 7 — is deferred to Section 3.3, where their differing physical approximations and their consequences for jet observables are discussed.

3.1.6 Event Generation Example

We now give a brief example to the process of Monte Carlo event generation in particle physics experiments. The goal of this procedure is to produce a set of data that is in the same format as the one coming out of the actual detector, in order to be able to match and simulation and real data coming from physics. We will consider a specific example to outline the main steps and ideas behind the procedure, considering the production of $t\bar{t}$ in pp collisions. Production of $t\bar{t}$ is mediated by gluon-gluon processes. The t and $\bar{t}$ will follow subsequent decays in:

$$t \rightarrow W^+ b \rightarrow u \bar{d} b \quad \bar{t} \rightarrow W^- \bar{b} \rightarrow \bar{\ell} \nu_{\ell} \bar{b}$$

All of the aforementioned decays are depicted in Figure (3.6).

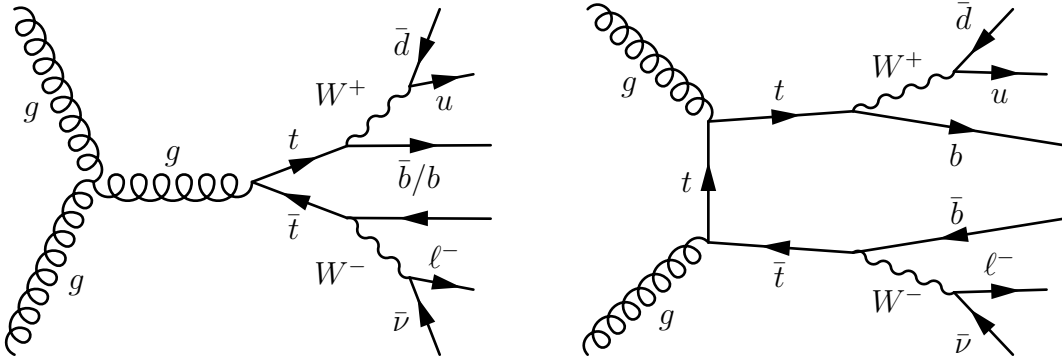


FIGURE 3.6: Feynman diagrams of $t\bar{t}$ process in the pp collisions: $t \rightarrow W^+ b \rightarrow u \bar{d} b$, $\bar{t} \rightarrow W^- \bar{b} \rightarrow \bar{\ell} \nu_{\ell} \bar{b}$ with s -channel (left) and t -channel (right)

We can simulate the hard-part of the Feynman diagrams via matrix element generators (ME generators). ME generators also perform the simulation of a gluon (g) from a proton in the proton-proton collision, considering the relative PDFs. However, the production of a hard-process using ME generators is not the end of the process but a starting point of the event generation: several extra steps are needed to obtain the full dataset for the

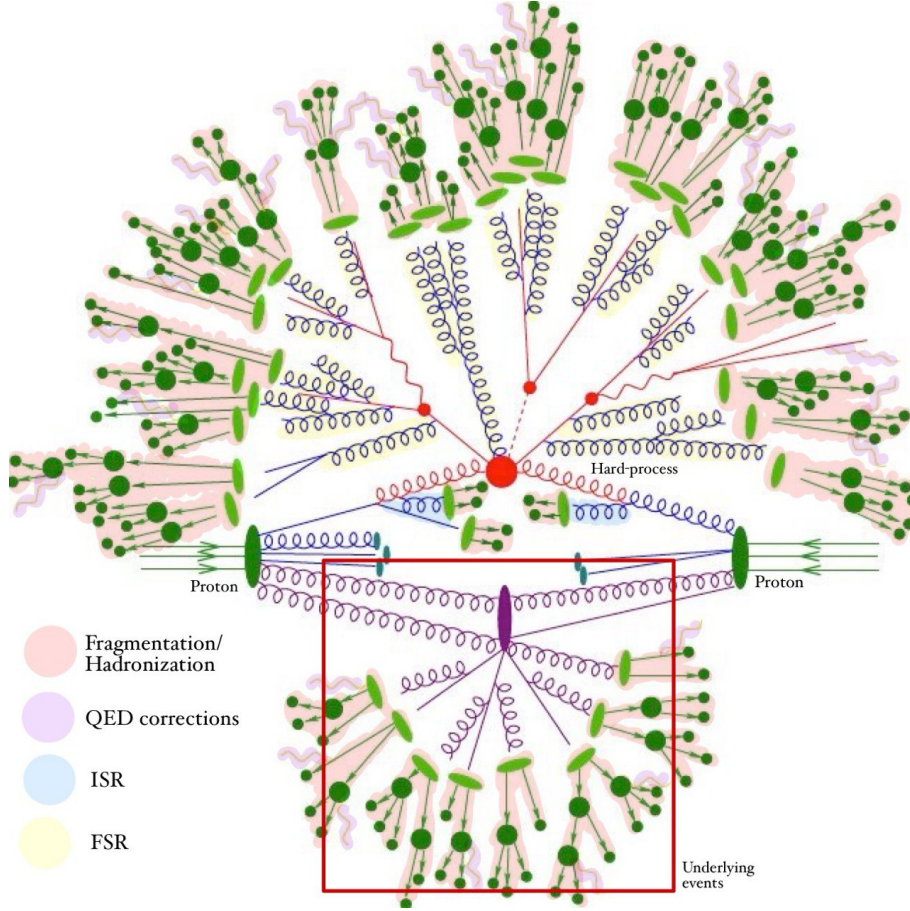


FIGURE 3.7: Proton-proton collision event generated with SHERPA. The different steps of MC simulations are highlighted in different colors. Courtesy of [17].

collision. One of these is the parton shower step: lots of quarks and gluons with relatively small p_T are emitted via soft and collinear emissions, making what is called a parton shower. There are two different types of showers for such emissions: initial state radiation (ISR) and final state radiation (FSR); the theoretical idea behind them is similar⁵, the use of ISR and FSR depends on when the emissions happen: ISR is for radiations emitted before the hard-process, while FSR is for after the hard-process. ISR gluons and quarks are radiated from the initial gluons of $gg \rightarrow t\bar{t}$ but FSR gluons are radiated from quarks (u and/or $\bar{d}$) of $W^+ \rightarrow u\bar{d}$. From QCD we know, that however, due to color confinement,

⁵Gluons can emit other gluons, and the probability of this happening strongly depends on the energy of the gluon: the more energetic the gluon, the less likely emission is. In perturbative quantum field theory, the amplitudes for ISR and FSR can be described using analogous Feynman rules, with the distinction that the former originates from incoming particles before the hard scattering, while the latter originates from outgoing particles after the scattering. Both processes are governed by the same gauge couplings and share similar soft and collinear singularity structures, which leads to comparable resummation techniques in calculations.

it is impossible to observe free gluons or quarks: they need to be combined to compose hadrons by a process called hadronization. New additional quarks and gluons are created from the existing quarks, gluons or vacuums, which are gluon fields in this case. This step is called fragmentation but is sometimes included in the step of hadronization. Photons can be radiated from charged leptons and quarks and this constitutes the QED radiative corrections step. The remaining parts of protons not used in the hard-process are called "underlying events" and must be dealt with properly with parton showers, hadronization and fragmentation steps themselves. Finally, the decay of mesons, baryons and leptons is performed until particles produced from decays become "stable". Kinematics (energy and momentum) of all the particles are determined under conservation rules (energy, momentum, spin/polarisation, etc.) with the MC method. Figure (3.7) shows the scheme of an event, with all the steps highlighted in different colors.

Section Roadmap:

Covered: Experimental methods of observation — experimental apparatus (ATLAS), its coordinate systems, and the reconstruction of charged particle tracks as the primary inputs to jet substructure measurements; the motivation for studying boosted $H \rightarrow Za$ decays and why the merged-jet topology forces a track-based substructure analysis; the general methods and steps of a particle physics analysis, with an example of Monte Carlo event generation.

Coming up: The QCD theoretical framework that explains why the charged tracks inside a merged jet carry the information needed to identify a boosted $a \rightarrow$ hadrons decay — from infrared singularity structure and the Lund plane to the specific substructure variables computed from ghost-associated tracks.

References: [36]; [37]; [4]; [38].

3.2 Looking Inside Jets

The experimental tools described in the previous Section — the solenoidal tracker, the calorimeters, the vertex reconstruction, and the track-to-jet association — define the objects with which any LHC analysis must work. Understanding what those objects actually measure, however, requires a theoretical account of how quarks and gluons produced in the hard scattering evolve into the collimated showers of hadrons and charged tracks that the detector records. The next Sections provide this account, starting from the QCD Lagrangian and the infrared singularity structure that shapes the parton shower, and ending with the specific track-based substructure variables whose theoretical properties — IRC safety, non-perturbative sensitivity, and Lund-plane geometry — determine how faithfully they can be predicted and compared between generators.

Jets — collimated sprays of hadrons ubiquitous in collider experiments — are the signature of the production of particles carrying color charge. When protons collide inelastically with a large energy transfer, one can isolate the hard process at the core of the collision from all accompanying activity, as seen in Figure (3.8).

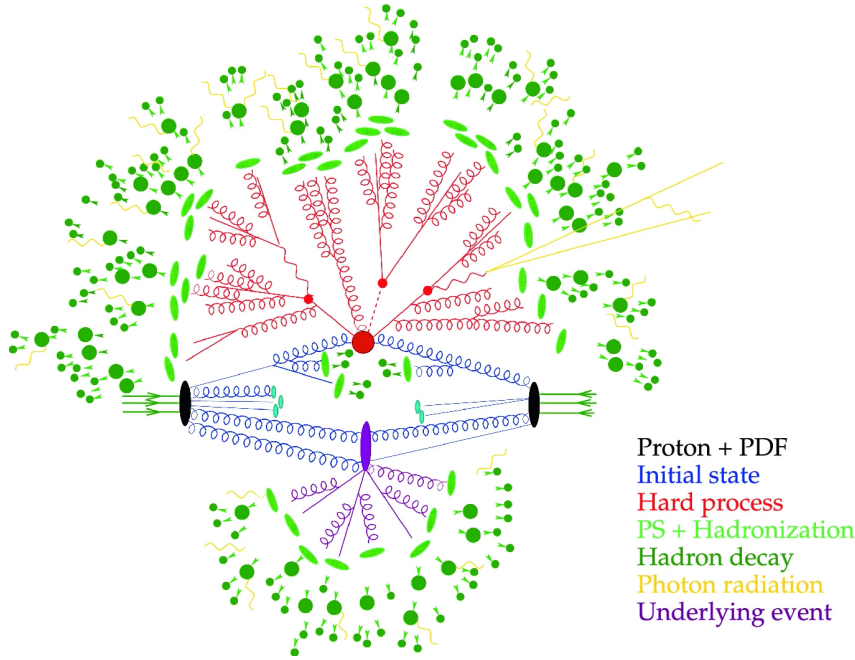


FIGURE 3.8: Illustration of a event as generated by a Monte Carlo event generator. Quarks and gluons in the initial state (blue) contribute with fractions of the energy of the colliding protons according to PDFs. The hard interaction and additional hard QCD emissions are shown in red. An underlying event interaction is visible in purple. After parton showering and hadronization of intermediate hadrons (light green), the decay into the final state takes place (dark green). Additional photon radiations are added in yellow. Courtesy of [39].

Since the energy of the hard process itself is large, typically in the range of a few hundred GeV and a few TeV, this leaves a large gap between the energies for the incoming protons and the hard process, leaving a large phase space for the decay products to populate. To complicate things further, the LHC collides protons in bunches of order $\mathcal{O}(10^{11})$. The interaction of the extra protons generates what is referred to as the interaction pileup. Learning more about jets is of our interest because hard process products we might be interested in, such as BSM hadronically decaying particles like the a we introduced in Chapter 2. Jet analysis is important since light BSM particles can decay in highly boosted jets that are hard to distinguish from the normal QCD background.

3.2.1 Quantum Chromodynamics

Jet physics is QCD physics. An understanding of such objects implies the knowledge of Quantum Chromodynamics, as described by the Standard Model of particle physics. We will follow a theoretical approach to explain how color charge works, and why jets come to be the way they are. Quantum Chromodynamics is a non-Abelian gauge theory based on the symmetry group $SU(3)_C$. The theory describes the interaction between fermionic and bosonic fields associated to quarks and gluons, as described in the Lagrangian, where $A = 1, \dots, 8$:

$$\mathcal{L} = -\frac{1}{4}F_{\mu\nu}^A F_A^{\mu\nu} + \sum_{\text{flavors}} \bar{\psi}_a (i\not{D} - m)_{ab} \psi_b,$$

where $F_{\mu\nu}^A$ is the gluon field strength, and D_μ is the covariant derivative, both defined as:

$$\begin{aligned} F_{\mu\nu}^A &= \partial_\mu A_\nu^A - \partial_\nu A_\mu^A + g_s f^{ABC} A_\mu^B A_\nu^C \\ (D_\mu)_{ab} &= \partial_\mu \delta_{ab} - ig_s A_\mu^A t_{ab}^A. \end{aligned}$$

Note that in the following definitions we have introduced f^{ABC} , which is the totally anti-symmetric tensor that defines the structure constants for the color number algebra, and t_{ab}^A are the 3×3 matrices, named after Gell-Mann, that generate the $SU(3)_C$ transformations. A defining feature of QCD is the running of the coupling constant $\alpha_s = \frac{g_s^2}{4\pi}$, that scales with the energy of the process we are considering. One can use this theory and its associated Feynman rules to calculate cross sections for observables. We have previously observed such a cross section formula, as shown below:

$$\sigma_{pp \rightarrow X}(S) = \sum_{i,j} \int_0^1 dx_1 \int_0^1 dx_2 f_i(x_1, \mu_F^2) f_j(x_2, \mu_F^2) \hat{\sigma}_{ij \rightarrow X}(\hat{s} = x_1 x_2 S, \mu_F^2, \mu_R^2),$$

where the parton-level cross section is given by the square of the matrix element:

$$\sigma_{pp \rightarrow X}(S) = \sum_{i,j} \int_0^1 dx_1 \int_0^1 dx_2 \int d\Phi_n f_i(x_1, \mu_F^2) f_j(x_2, \mu_F^2) \frac{1}{2\hat{s}} |\mathcal{M}_{ij \rightarrow X}(\Phi_n, \mu_F^2, \mu_R^2)|^2.$$

In the above result, the parton-level cross section has been replaced with the squared amplitude of the process obtained using Feynman rules, where $d\Phi_n$ denotes the differential phase space element over the n final states:

$$d\Phi_n = \prod_{i=1}^n \frac{d^3 p_i}{(2\pi)^2 2E_i} (2\pi)^4 \delta^{(4)}(p_i + p_j - \sum_{i=1}^n p_i).$$

This formula holds for any perturbative order, and the full calculation is usually approximated via a power series expansion in the strong coupling constant α_s . Therefore if v is a generic observable,

$$\sigma(v) = \sigma_0 + \alpha_s \sigma_1 + \alpha_s^2 \sigma_2 + \alpha_s^3 \sigma_3 + \mathcal{O}(\alpha_s^4).$$

Each of increasing power expansion term of α_s takes into consideration a higher-order correction to the cross section. Some caution however must be taken in order to be able to proceed: QCD, as any other renormalizable theory where a massless gauge boson is present, is plagued by divergences that increase the difficulty of the calculations. These divergences are of two different types, ultraviolet (UV) and infrared (IR), where the former is related to high momentum exchanges in loop integrals, while the latter is related to the low momentum behavior of a massless propagator. In the following Sections we will take a look at what these divergences do, and what their existence implies for our theory.

3.2.2 Ultraviolet Divergences and The Running of the Strong Coupling Constant

A central feature of QCD is that its coupling constant $\alpha_s = g_s^2/4\pi$ is not a fixed parameter but depends on the energy scale μ at which it is probed. This scale dependence, known as the running of the coupling, is a direct consequence of the renormalization procedure required to render the perturbative expansion of QCD finite.

In any quantum field theory, loop corrections to Green functions introduce ultraviolet divergences that must be regulated and absorbed into a redefinition of the bare parameters of the Lagrangian.

At one loop, these diagrams involve a single internal momentum k^μ that is not fixed by external kinematics and must be integrated over all values $\int d^4 k$. For large $|k| \rightarrow \infty$, the integrand generically falls off too slowly for the integral to converge, producing ultraviolet

(UV) divergences. It is easy to figure out which diagrams will then be divergent: power counting⁶ yields this formula:

$$D = 4 - \frac{3}{2}N_q - N_g,$$

which states that the degree of divergence of a loop diagram, noted with D , is determined by the number of external quark propagator legs (N_q) and by the number of external gluon legs (N_g). At one loop, we then can figure out exactly which diagrams will be problematic for our purposes, since any value for $D \geq 0$ will yield a divergence⁷. In QCD, the relevant one-loop diagrams that renormalize the coupling g_s fall into three classes. The first consists of quark self-energy diagrams, in which a quark of momentum p emits and reabsorbs a virtual gluon; the loop integral over the gluon momentum k diverges as $k \rightarrow \infty$ and yields a $1/\epsilon$ pole that renormalizes the quark propagator. The second class comprises gluon self-energy (vacuum polarization) diagrams, which include a quark–antiquark loop, a loop of virtual gluons, and a ghost loop; the latter two are purely non-Abelian contributions with no analogue in QED⁸. Finally, vertex correction diagrams, in which a virtual gluon connects two legs of the quark–gluon or three-gluon vertex, renormalize the coupling directly. It is the relative weight of these contributions (depicted in Figure (3.9)) that determines the behavior of the theory at large energies.

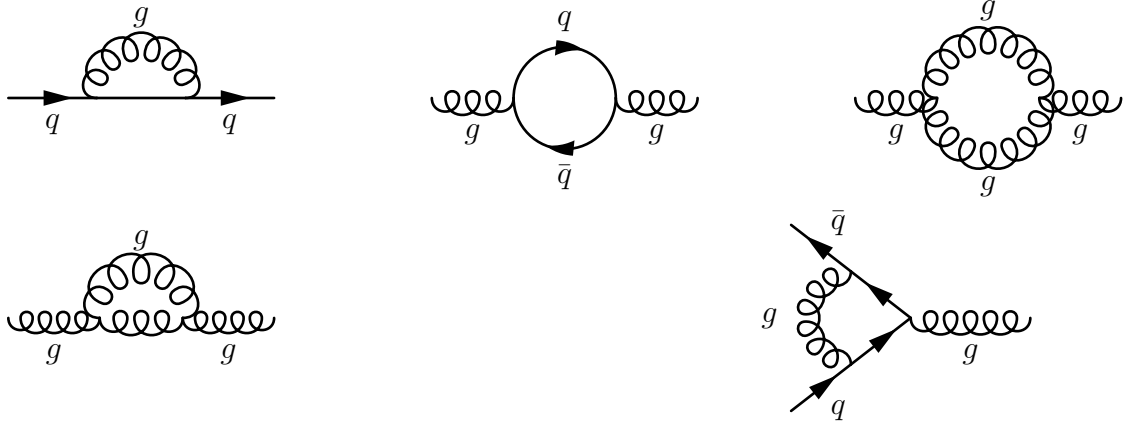


FIGURE 3.9: One-loop QCD self-energy and vertex corrections.

In dimensional regularization, working in $d = 4 - 2\epsilon$ spacetime dimensions, the bare coupling $g_s^{(0)}$ is related to the renormalized coupling $g_s(\mu)$ through

$$g_s^{(0)} = \mu^\epsilon Z_g g_s(\mu),$$

⁶Power counting is a crude procedure to determine whether a loop diagram is divergent or not, and it hinges itself on counting the dimension of energy inside the loop integral.

⁷The divergences can have different "intensities": if $D = 0$, we have a logarithmic divergence, if $D = 1$ we have a linear divergence, if $D = 2$ we have a quadratic divergence, and so on.

⁸We are not considering the ghost-divergent diagram due to the Fadeev-Popov procedure, since it is beyond the scope of this text.

where μ is the renormalization scale and Z_g is the coupling renormalization constant, which absorbs the $1/\epsilon$ poles arising from loop integrals⁹. The requirement that physical observables be independent of the arbitrary scale μ leads to the Callan–Symanzik equation,

$$\mu \frac{d\alpha_s}{d\mu} = \beta(\alpha_s),$$

where the beta function $\beta(\alpha_s)$ encodes the response of the coupling to a change in scale. At one loop, the beta function takes the form

$$\beta(\alpha_s) = -2\alpha_s \left(\beta_0 \frac{\alpha_s}{4\pi} + \mathcal{O}(\alpha_s^2) \right),$$

with the one-loop coefficient

$$\beta_0 = 11 - \frac{2}{3}n_f,$$

where n_f is the number of active quark flavors. The two contributions to β_0 have distinct physical origins: the term 11 arises from gluon self-interactions and ghost loops, a consequence of the non-Abelian $SU(3)_C$ structure of QCD, while the term $-2n_f/3$ reflects the screening effect of quark loops, analogous to vacuum polarization in QED.

For $n_f \leq 16$, one has $\beta_0 > 0$, and therefore $\beta(\alpha_s) < 0$: the coupling decreases as the energy scale increases. Integrating the one-loop Callan–Symanzik equation yields the explicit running

$$\alpha_s(\mu^2) = \frac{\alpha_s(\mu_0^2)}{1 + \frac{\beta_0}{2\pi} \alpha_s(\mu_0^2) \ln(\mu^2/\mu_0^2)},$$

or equivalently, introducing the non-perturbative scale Λ_{QCD} at which α_s formally diverges,

$$\alpha_s(\mu^2) = \frac{2\pi}{\beta_0 \ln(\mu/\Lambda_{\text{QCD}})}.$$

Experimentally, $\Lambda_{\text{QCD}} \approx 200\text{--}300$ MeV, and the coupling evaluated at the Z boson mass is $\alpha_s(M_Z^2) \approx 0.118$.

This asymptotic freedom is one of the defining properties of QCD. At large momentum transfers $\mu \gg \Lambda_{\text{QCD}}$, quarks and gluons interact weakly and perturbative calculations are reliable (this is called the pQCD regime). Conversely, as $\mu \rightarrow \Lambda_{\text{QCD}}$, the coupling grows large and the perturbative expansion breaks down, giving rise to the confinement of quarks into hadrons and the onset of non-perturbative phenomena. It is precisely in this regime that the descriptions of jet formation, hadronization, and the underlying event encoded in Monte Carlo generators as no analytic first-principles calculation is currently tractable, but it relies on model fitting and experimental data.

Whereas in QED the behavior is opposite: the one-loop beta function coefficient is $\beta_0^{\text{QED}} = -\frac{4}{3}n_f < 0$, so that $\beta(\alpha) > 0$ and the electromagnetic coupling α_{em} increases with energy.

⁹The contents of Z_g also depend on the renormalization scheme, but they always at least contain the divergent part.

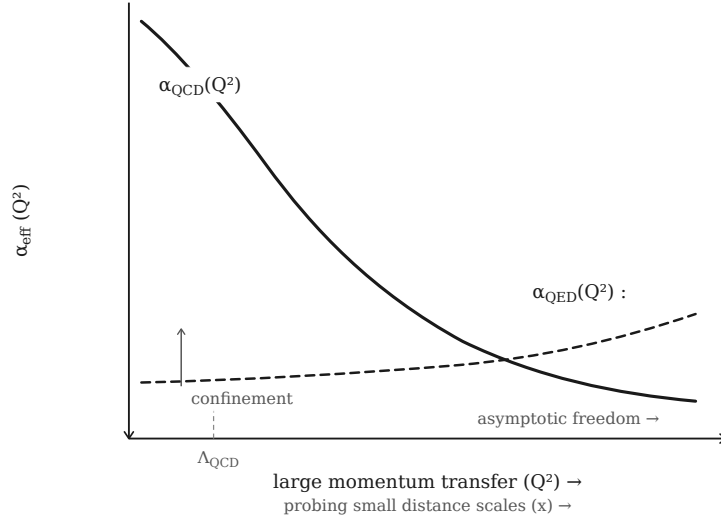


FIGURE 3.10: Running of QCD and QED coupling constants.

This is because QED has no self-interacting gauge bosons — the photon carries no electric charge — and the only contribution to the running comes from charged fermion loops, which screen the bare charge. The well-known value $\alpha_{\text{em}} \approx 1/137$ is therefore a low-energy quantity; at the scale of the Z boson mass one finds $\alpha_{\text{em}}(M_Z^2) \approx 1/128$. The contrast with QCD is stark: it is precisely the non-Abelian gluon self-coupling, absent in QED, that reverses the sign of β_0 and drives asymptotic freedom. In this sense, the $SU(3)$ color structure of QCD is not merely a technical detail but the physical origin of the entire phenomenology of jets, confinement, and the hard-soft factorization that underpins collider physics. This behavior can be seen in Figure (3.10).

3.2.3 Infrared Divergences in Quantum Chromodynamics

Infrared divergences are characteristic of theories in which the interaction's mediator is massless, so its Green function, expressed as:

$$iG_{\mu\nu}(p^2) = \frac{-ig_{\mu\nu}}{p^2 + i\epsilon},$$

shows problematic behavior for $p_\mu \rightarrow 0$. In order to highlight the structure of IR singularities in matrix elements we consider the NLO corrections to this process:

$$e^+e^- \rightarrow \gamma^* \rightarrow q\bar{q}.$$

This is a simplified example since it does not take the initial-state radiation into consideration, but it is nevertheless useful for explaining how these divergences behave and arise. The relevant Feynman diagrams are shown in Figure (3.11).

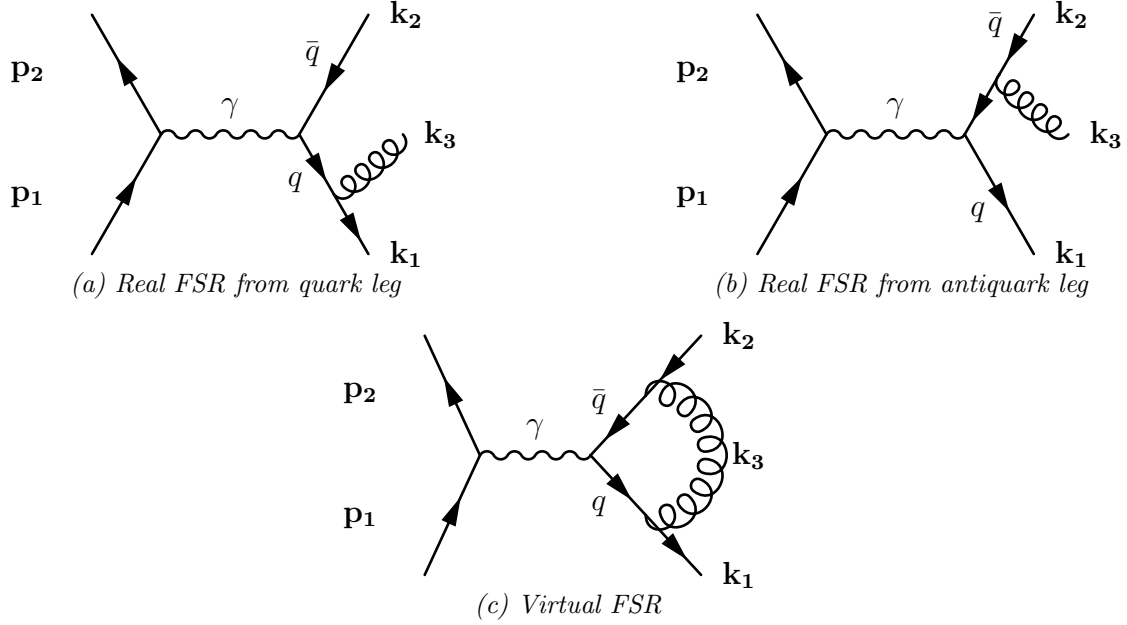


FIGURE 3.11: Real and virtual final state radiation in $e^+e^- \rightarrow \gamma^* \rightarrow q\bar{q}(g)$. Momenta $\mathbf{p}_1$, $\mathbf{p}_2$ denote the incoming e^- and e^+ , $\mathbf{k}_1$, $\mathbf{k}_2$ the outgoing quark and antiquark, $\mathbf{k}_3$ the emitted or virtual gluon.

The matrix elements for the real emission diagrams, denoted with $\mathcal{M}_{FSR}^{(a)/(b)}$, will be:

$$\mathcal{M}_{FSR}^{(a)} = \bar{v}(\mathbf{p}_2)(ie\gamma^\mu)u(\mathbf{p}_1) \left[\frac{-g_{\mu\nu}}{(\mathbf{p}_1 + \mathbf{p}_2)^2 + i\epsilon} \right] \bar{u}(\mathbf{k}_1)(ie\gamma^\nu) \left[\frac{\not{\mathbf{k}}_2 + \not{\mathbf{k}}_3}{(\mathbf{k}_2 + \mathbf{k}_3)^2 + i\epsilon} \right] (t_1^a g_s \gamma^\alpha) \epsilon_\alpha(\mathbf{k}_3) v(\mathbf{k}_2)$$

$$\mathcal{M}_{FSR}^{(b)} = \bar{v}(\mathbf{p}_2)(ie\gamma^\mu)u(\mathbf{p}_1) \left[\frac{-g_{\mu\nu}}{(\mathbf{p}_1 + \mathbf{p}_2)^2 + i\epsilon} \right] \bar{u}(\mathbf{k}_1)(ie\gamma^\nu) \left[\frac{\not{\mathbf{k}}_1 + \not{\mathbf{k}}_3}{(\mathbf{k}_1 + \mathbf{k}_3)^2 + i\epsilon} \right] (t_1^a g_s \gamma^\alpha) \epsilon_\alpha(\mathbf{k}_3) v(\mathbf{k}_2),$$

where $t_{1/2}^a$ are the color matrices associated to gluons in the adjoint representation. We can apply the soft limit to these expressions, considering soft final state radiation, $\mathbf{k}_3 \rightarrow 0$. This will turn the terms into:

$$\mathcal{M}_{FSR}^{(a)} = \bar{v}(\mathbf{p}_2)(ie\gamma^\mu)u(\mathbf{p}_1) \left[\frac{-g_{\mu\nu}}{(\mathbf{p}_1 + \mathbf{p}_2)^2 + i\epsilon} \right] \bar{u}(\mathbf{k}_1)(ie\gamma^\nu) \left[\frac{\not{\mathbf{k}}_2}{2\mathbf{k}_2 \cdot \mathbf{k}_3 + i\epsilon} \right] (t_1^a g_s \gamma^\alpha) \epsilon_\alpha(\mathbf{k}_3) v(\mathbf{k}_2)$$

$$\mathcal{M}_{FSR}^{(b)} = \bar{v}(\mathbf{p}_2)(ie\gamma^\mu)u(\mathbf{p}_1) \left[\frac{-g_{\mu\nu}}{(\mathbf{p}_1 + \mathbf{p}_2)^2 + i\epsilon} \right] \bar{u}(\mathbf{k}_1)(ie\gamma^\nu) \left[\frac{\not{\mathbf{k}}_1}{2\mathbf{k}_1 \cdot \mathbf{k}_3 + i\epsilon} \right] (t_1^a g_s \gamma^\alpha) \epsilon_\alpha(\mathbf{k}_3) v(\mathbf{k}_2),$$

We can now exploit gamma matrix anticommutation rules and the massless Dirac equations for fermions, dropping the bold face for 4-vectors:

$$\begin{aligned}\bar{u}(k_1) \gamma^\mu \not{k}_2 &= \bar{u}(k_1) (2k_2^\mu - \not{k}_2 \gamma^\mu) = 2k_2^\mu \bar{u}(k_1) \\ \bar{u}(k_1) \gamma^\mu \not{k}_1 &= \bar{u}(k_1) (2k_1^\mu - \not{k}_1 \gamma^\mu) = 2k_1^\mu \bar{u}(k_1)\end{aligned}$$

The result of this calculation is that, in the soft limit, the matrix element for final state real radiation correction factorizes into two terms: one term is related to the LO process $e^+e^- \rightarrow \gamma \rightarrow q\bar{q}$ multiplied by a special factor, called the eikonal factor. We see it in the following:

$$\mathcal{M}_{FSR}^{(a)} + \mathcal{M}_{FSR}^{(b)} \rightarrow g_s J^\mu(k_3) \epsilon_\mu(k_3) \mathcal{M}_{LO},$$

where the eikonal factor $J^\mu(k)$ is:

$$J^\mu(k) = \sum_{i=1}^2 t_i^a \frac{k_i^\mu}{k \cdot k_i} = t_1 \frac{k_1^\mu}{k \cdot k_1} + t_2 \frac{k_2^\mu}{k \cdot k_2}.$$

It is important to report that the eikonal factor's structure does not depend on the specific structure of the process we consider. From the physical point of view this reflects the fact that soft radiation is not able to resolve the details of the interactions.

We can now square the amplitude and obtain:

$$|\mathcal{M}_{FSR}|^2 \rightarrow |\mathcal{M}_{LO}|^2 g_s^2 J^\mu(k_3) J_\mu(k_3),$$

which in turn yields:

$$|\mathcal{M}_{LO}|^2 g_s^2 \left[- \sum_{i,j} t_i^a t_j^a \frac{k_i \cdot k_j}{(k_i \cdot k_3)(k_j \cdot k_3)} \right] = |\mathcal{M}_{LO}|^2 g_s^2 C_{12} \frac{k_1 \cdot k_2}{(k_1 \cdot k_3)(k_2 \cdot k_3)},$$

where C_{12} has been introduced as the effective color charge.

We can apply the soft limit to virtual corrections, too. The one-loop virtual correction is:

$$I = g_s^2 (-i) C_F \int \frac{d^4 k_3}{(2\pi)^d} \frac{\bar{u}(\mathbf{k}_1) \gamma^\mu (\not{\mathbf{k}}_1 + \not{\mathbf{k}}_3) \gamma^\rho (\not{\mathbf{k}}_3 - \not{\mathbf{k}}_2) \gamma_\mu v(\mathbf{k}_2)}{[(\mathbf{k}_3 + \mathbf{k}_1)^2 + i\epsilon][(\mathbf{k}_3 - \mathbf{k}_2)^2 + i\epsilon][\mathbf{k}_3^2 + i\epsilon]}.$$

We can then apply the soft approximation to obtain:

$$g_s^2 C_F (-i) \int \frac{d^4 k_3}{(2\pi)^4} \frac{(k_1 \cdot k_2) [\bar{u}(k_1) \gamma^\rho v(k_2)]}{[k_3 \cdot k_1 + i\epsilon] [-k_3 \cdot k_2 + i\epsilon] [k_3^2 + i\epsilon]}.$$

Having done this, we need to evaluate the integral by using a special parametrization for the four-momenta:

$$\begin{aligned} k_1^\mu &= E_1(+1, 0, 0, +1) \\ k_2^\mu &= E_2(+1, 0, 0, -1) \\ k_3^\mu &= (k_3^0, \vec{k}_{3\perp}, k_3^z). \end{aligned}$$

The integral then becomes:

$$I = g_s^2 C_F (-i) \int \frac{d^3 k_3}{(2\pi)^4} \frac{2 dk_3^0 [\bar{u}(k_1) \gamma^\rho v(k_2)]}{(k_3^0 - k_3^z + i\epsilon) (-k_3^0 - k_3^z + i\epsilon) (k_3^{02} - k_3^{z2} - k_{3\perp}^2 + i\epsilon)}$$

Skipping the full contour integral evaluation, we give the final integral result:

$$I = g_s^2 C_F [\bar{u}(k_1) \gamma^\rho v(k_2)] \int \frac{d^3 k_3}{(2\pi)^3} \left[\frac{-(k_1 \cdot k_2)}{2|\vec{k}_3|(k_1 \cdot k_3)(k_2 \cdot k_3)} \right]$$

We can collect real and virtual contributions together to find the NLO distribution of a generic variable v by introducing a specific measurement function $V_n(\{k_i\})$, which describes the value of v for a set of final state particles k_i ¹⁰. With these results under our belt, we can write out the cross-section for v in NLO accuracy:

$$\begin{aligned} \sigma(v) &= \frac{1}{2s} \int d\Phi_2 |\mathcal{M}_{LO}|^2 V_2(k_1, k_2) \\ &+ \frac{1}{2s} \int d\Phi_2 |\mathcal{M}_{LO}|^2 \int \frac{d^3 k_3}{(2\pi)^3 2|\vec{k}_3|} 2g_s^2 C_F \frac{(k_1 \cdot k_2)}{(k_1 \cdot k_3)(k_2 \cdot k_3)} [V_3(k_1, k_2, k_3) - V_2(k_1, k_2)] \end{aligned}$$

We can see that the virtual and real corrections sum up, with the only real difference being the fact that the measurement function V lives in a three-particle space for the real terms.

We have therefore derived that any observable v in the limit where the radiated parton is soft exhibits a singular behavior encoded in the eikonal factor. Taking a look at the denominator terms:

$$k_i \cdot k_3 = E_i E_3 (1 - \cos \theta_{i3}).$$

We note a divergent behavior in the case of soft emitted gluons ($k_i \rightarrow 0$), but also in the case of collinear particles ($\theta_{i3} \rightarrow 0$), giving rise to the so-called infrared and collinear divergences.

¹⁰This special function depends on how many particles are present in the final state and can contain a series of Dirac deltas, or Heaviside functions, it is related to the kinematics of the process.

In order to consider finite contributions to our observables, we must achieve a complete cancellation of the divergent parts. This is given by requiring:

$$\begin{aligned} \text{Collinear safety} & \quad \Leftrightarrow \quad V_{m+1}(\dots k_i, k_j \dots) \longrightarrow V_m(\dots, k_i + k_j, \dots) & \quad \text{if } k_i \parallel k_j \\ \text{Infrared safety} & \quad \Leftrightarrow \quad V_{m+1}(\dots, k_i, \dots) \longrightarrow V_m(\dots, k_{i-1}, k_{i+1}, \dots) & \quad \text{if } k_i \rightarrow 0. \end{aligned}$$

Whenever a parton is split into two collinear partons, or whenever a soft parton is added, the value of the observable must remain unchanged in order to guarantee a correct cancellation of the divergence. If we consider exclusive measurements¹¹, the singularities in this case cancel, but the kinematical dependence of the observable can cause an imbalance between real and virtual contributions. These contributions then spoil the perturbative expansion and must be resummed. This is the theoretical background of the problem of really soft radiation, and the justification to the usage of grooming algorithms for jets, about which we will discuss in later sections.

3.2.4 The Lund Plane

The infrared and collinear singularities derived in the previous section are not merely algebraic accidents: they reflect a genuine geometric structure in the phase space of QCD radiation. It is useful to make this structure explicit before turning to jet algorithms and substructure variables, since the same diagram will be useful in the following for describing grooming, substructure, and non-perturbative effects.

Consider a parton — a quark or gluon — that emits a secondary gluon with transverse momentum k_t relative to the emitting parton's direction, at an emission angle Δ . One defines the two logarithmic coordinates

$$\ell \equiv \ln \frac{1}{\Delta}, \quad \kappa \equiv \ln \frac{k_t}{p_T},$$

where p_T is the transverse momentum of the jet. The two-dimensional plane spanned by (ℓ, κ) is called the Lund plane. Each point in this plane corresponds to a single primary emission from the hard parton, and the full radiation pattern of a jet can be thought of as a set of points scattered across it.

The Lund plane has a natural physical geography. The soft singularity of the eikonal factor, $k_t \rightarrow 0$, corresponds to the lower half of the plane ($\kappa \rightarrow -\infty$), while the collinear singularity, $\Delta \rightarrow 0$, corresponds to the right half of the plane ($\ell \rightarrow +\infty$). The two limits together define the lower-right corner as the most divergent region, where emissions are simultaneously soft and collinear. The non-perturbative regime, where $k_t \lesssim \Lambda_{\text{QCD}}$,

¹¹An exclusive observable (or measurement) in particle physics refers to a measurement in which all final-state particles produced in a collision are identified, reconstructed, and accounted for. This is in contrast to an inclusive measurement, where only a subset of the final-state particles is observed, and the rest are summed or integrated over. Exclusive measurements provide more detailed information about the underlying hard-scattering process and are sensitive to the full kinematic configuration of the event, but are experimentally challenging due to the requirement of complete final-state reconstruction.

occupies a horizontal strip at the very bottom of the plane, below which no perturbative description is valid and the hadronization regime takes over.

At leading logarithmic accuracy, the density of primary emissions is approximately uniform per unit area of the Lund plane,

$$\frac{d^2 P}{d\ell d\kappa} \approx \frac{\alpha_s C_i}{\pi},$$

where $C_i = C_F$ for a quark-initiated jet and $C_i = C_A$ for a gluon-initiated jet. This uniformity is the logarithmic structure of the QCD parton shower made manifest: it is a direct consequence of the double-logarithmic enhancement of soft-collinear radiation visible in the denominator $k_i \cdot k_3 = E_i E_3 (1 - \cos \theta_{i3})$ of the previous eikonal calculation. The fact that $C_A > C_F$ means that gluon jets populate the Lund plane more densely at every point, which is the underlying reason why gluon jets are broader, more populated, and harder to distinguish from signal than quark jets.

The IRC safety conditions derived at the end of the previous section have a clean geometric interpretation in this language. Infrared safety requires that the observable be insensitive to emissions in the $\kappa \rightarrow -\infty$ strip, and collinear safety requires insensitivity to the $\ell \rightarrow +\infty$ edge. An IRC-safe observable is one that does not resolve the lower-right corner of the Lund plane, depicted in Figure (3.12).

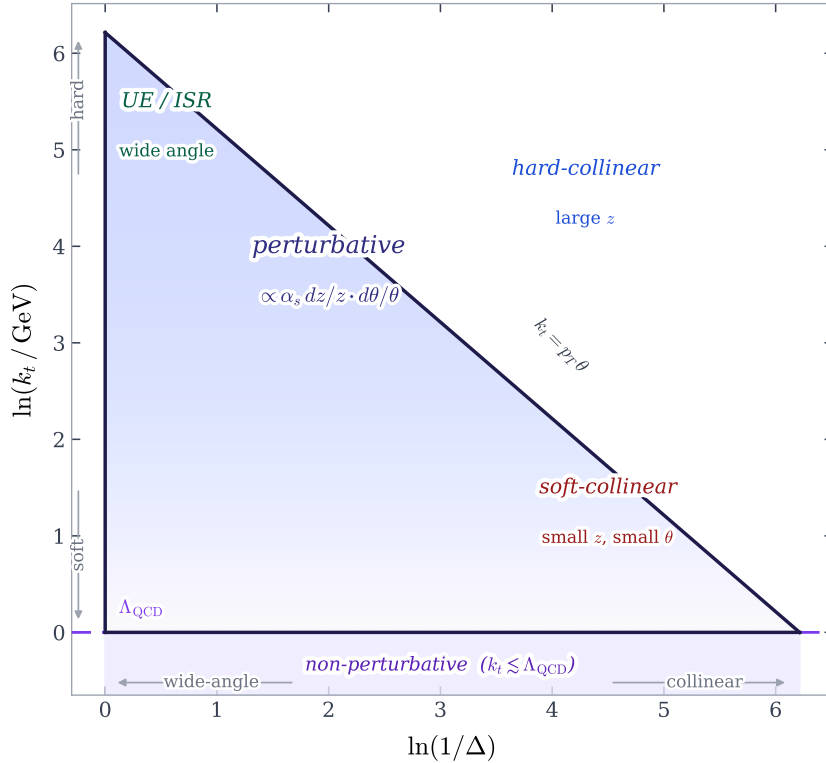


FIGURE 3.12: The Lund plane.

Jet algorithms must collect radiation from a well-defined region of the Lund plane without being destabilized by its soft-collinear corner: this is the geometric content of the IRC-safety requirements discussed in the previous paragraphs. Once a jet is defined, a substructure observable assigns a numerical value to it by summing a contribution from every emission inside the jet. The contribution of each emission depends on where it sits in the Lund plane — that is, on its k_t and Δ . Different observables use different functions of these two coordinates as weights, so they end up being more or less sensitive to different parts of the plane.

For example, an angularity $\lambda_\beta^\kappa \sim \sum_i k_{t,i}^\kappa \Delta_i^\beta$ with large β is dominated by emissions at large angles (large Δ), because the factor Δ^β grows fast and makes wide-angle emissions contribute more. Conversely, a small β suppresses that growth and shifts the sensitivity toward the collinear region. And **SoftDrop** grooming, discussed in the following, acts by removing all emissions below a straight boundary in the (ℓ, κ) plane, cleanly cutting off the non-perturbative strip and the soft wide-angle region while leaving the hard collinear core intact. The Lund plane is therefore not merely a visualization tool but the natural coordinate system for the entire phenomenology of jet substructures and reconstruction algorithms. Figure (3.13) depicts the intuitive relationship between jet geometry and the Lund plane.

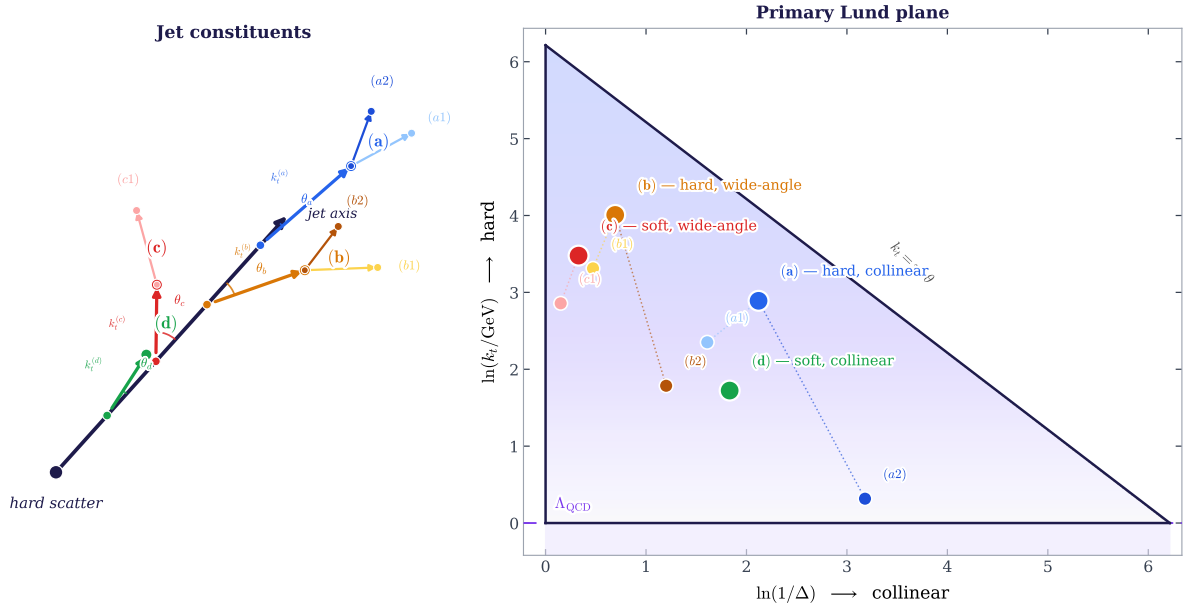


FIGURE 3.13: Correspondence between jet spatial geometry and Lund-plane phase space.

3.2.5 Jet Reconstruction Algorithms

Final state partons are the daily occurrence in events generated at LHC. Since we collide protons, after the hard scattering, a QCD parton can radiate from the initial state, the

incoming partons. Moreover, some of the particles that are produced in the hard event can decay to partons, and to reconstruct these decays one needs to look at decay chains involving quarks and gluons. The high energy quarks and gluons however are never directly observed, since they tend to fragment and undergo decay chains involving progressively less energetic particles, producing a series of collimated quarks and gluons, which can be seen to interact with the detector. A typical shower starts from a quark or a gluon with high virtuality and progresses until the particles produced reach the energy of the hadronization scale, which is of order of 1 GeV: at these stage will the partons produce hadrons in a process called hadronization. Jets can be defined to be collimated flows of hadrons and they can be seen as the fingerprint of the production of high-energy quarks and gluons in a collision. The simple concept of what a jet is meant to represent, however, is not sufficient to practically identify one. To this avail, one relies on a jet definition: a well defined procedure that allows for jet reconstruction from the set of recorded hadrons in a collision. This definition is made up of different blocks:

1. The jet algorithm: the actual computational procedure to reconstruct such objects;
2. The parameters of the algorithm;
3. A recombination scheme: it specifies how to reconstruct the kinematic properties of the jet from its constituents.

There are a few guidelines that have been followed for the development of the following jet definition algorithms: these are listed in the following box.

Important properties that a jet definition should satisfy:

- Simple to implement in the experimental analysis;
- Simple to implement in theoretical calculations;
- Defined at any perturbative order;
- Yields finite cross sections at any perturbative order;
- Yields a cross section that is relatively insensitive to hadronization.

Whereas the first few points are rather self-explanatory, the last one is a bit more complex. The description of a particle collision events relies on different blocks, and since jets are supposed to capture the hard partons in an event, one would expect that the jets coming out of different simulation steps would be in good agreement: this implies that observables built from jet reconstructed quantities should be as little sensitive as possible to nonperturbative effects like hadronization.

We are now ready to cover the sequential recombination algorithms used for jet definition. From the perturbative QCD viewpoint jets are the products of successive branchings of

partons. These algorithms aim to reverse the process combining two particles into one, trying to parse the chain in reverse. These algorithms come into different shapes and sizes, and are presented in the following.

Generalized- k_t Algorithm

Most of the algorithms for jet recombination belong to this specific family, which reconstructs jets with the following procedure:

1. It takes the list of our particle events as input.
2. From the object list, two kind of distances are built: the inter-particle distance, defined as:

$$d_{ij} = \min(p_{t,i}^{2p}, p_{t,j}^{2p}) \Delta R_{ij},$$

where the p at the exponent is a free parameter and ΔR_{ij} is the rapidity-azimuthal angle plane. One also defines a beam distance:

$$d_{iB} = p_{t,i}^{2p} R^2,$$

where R is a free parameter usually referred to as the jet radius.

3. Iteratively the code should find the smallest distance between all of the d_{ij} and d_{iB} :
 - If the smaller distance is d_{ij} , then the objects i and j are removed from the list of items and merged into one new object k which is added to the list.
 - If the smaller distance is a d_{iB} object, the i object is classified as a jet and the removed from the list.

One then iteratively starts again from step 2 until the objects in the list have all been parsed.

k_t Algorithm

The best known algorithm belonging to the previous family is the k_t -algorithm, corresponding to a choice of $p = 1$. A soft emission, with small p_t would be associated with a small distance and therefore would be recombined early in the process. This algorithm is therefore sensitive to soft emissions, but this becomes an issue as it tends to make jets more sensitive to extra soft radiation in the event, coming from pileup or the underlying event.

Cambridge/Aachen Algorithm

Another specific case of the generalized- k_t algorithm is the Cambridge/Aachen algorithm, obtained by choosing $p = 0$. In this case the jet distance becomes geometrical and suffers less from soft contamination than the previous algorithm does.

Anti- k_t Algorithm

In LHC physics, the anti- k_t algorithm has become the most popular and in use for jet reconstruction, obtained choosing $p = -1$. The advantage of such a choice is that it favors hard particles: the more energetic a particle is, the earlier it will be clustered. A hard jet will grow by successively aggregating soft particles around it until it has reached a radius R . Hard jets will therefore be insensitive to soft radiation and will have conical shape. This resistance to soft radiation highly facilitates the algorithm's calibration and usage by all the LHC experiments. A simulation of these reconstruction algorithms, step by step, is shown in Figure (3.14): the three columns describe three different reconstruction algorithms starting from the same set of simulated data. The jets are reconstructed according to each of the algorithm instructions, and the final objects differ. It is possible then to observe the peculiarities of each algorithm used.

Experimental Aspects

From the interaction of the particle with the detector, the experiments are able to rebuild special plots, called lego plots, that describe the energy deposits inside the calorimeter as a function of the pseudo-rapidity η and the azimuthal angle ϕ . In this case the distances between energy deposits is calculated via:

$$\Delta R_{ij}^{(\text{detector})} = \sqrt{(\phi_i - \phi_j)^2 + (\eta_i - \eta_j)^2}$$

These plots are a great tool for visualization of the energy deposits in the detector parts. A depiction of the lego plots from the toy model in Figure (3.14) is shown in Figure (3.15). The same simulated data are used to build a lego plot for visualization.

3.2.6 Jet Substructure Variables

Once a jet has been reconstructed by a clustering algorithm, its total four-momentum and cone radius R capture only the less definite features of the underlying hard process. The jet clustering algorithms described in the previous section are deliberately insensitive to the jet's internal structure: the anti- k_t algorithm produces conical jets regardless of how radiation is distributed inside the cone — it is intentional for the following reasons. To access the information encoded in the jet interior, one must turn to a different class of observables, collectively known as jet substructure variables.

These are observables constructed from the momenta and angles of the jet's individual constituents, built to quantify the pattern of radiation inside the cone. They are the natural toolset for distinguishing jets of different origin, for identifying boosted massive particles decaying hadronically, and for probing the transition from the perturbative parton shower to non-perturbative hadronization. A central constraint on any substructure observable that is to receive a finite perturbative prediction is that it must be infrared and collinear safe. Infrared safety requires that the value of the observable is unchanged

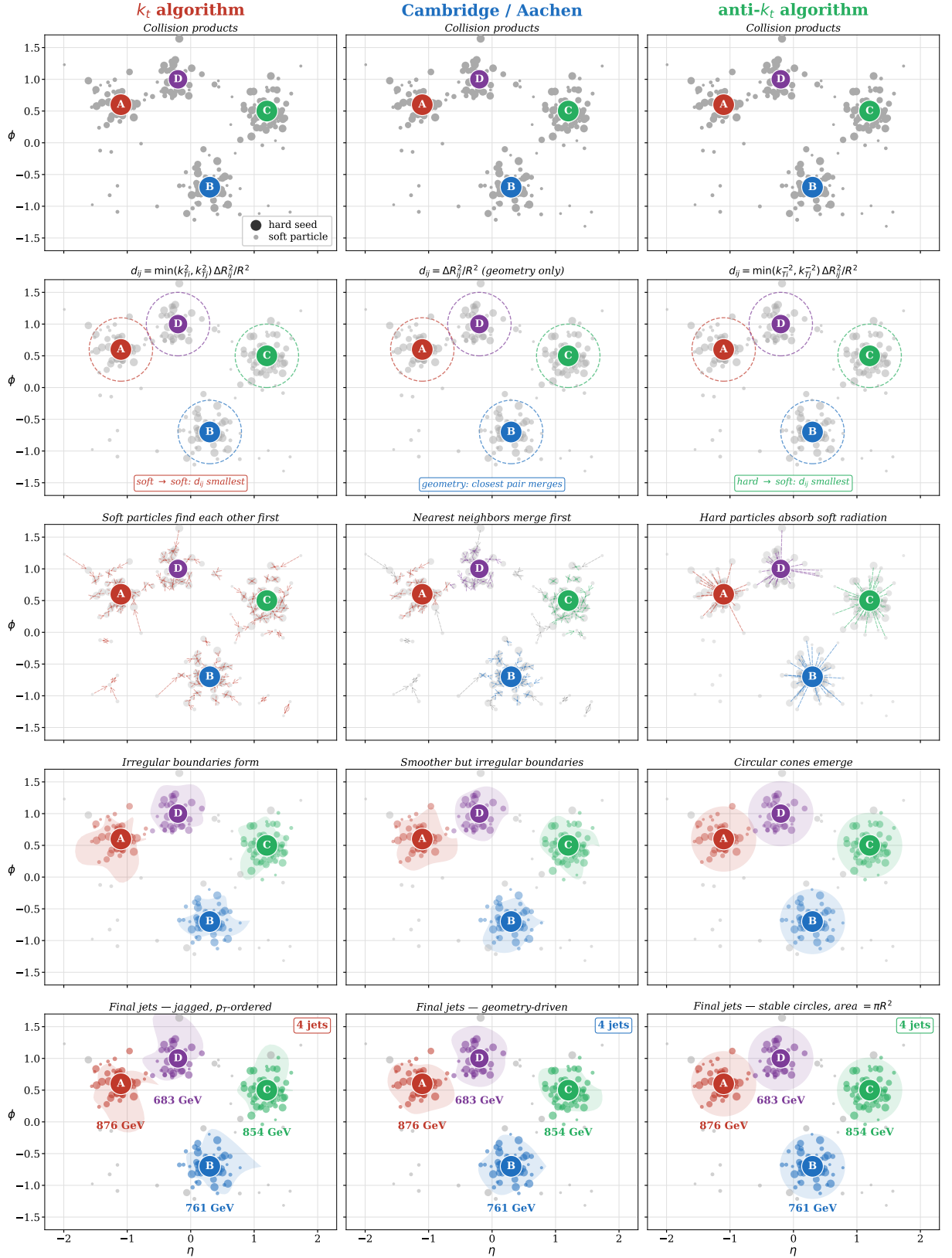


FIGURE 3.14: Comparison of the k_t [40], Cambridge/Aachen [41], and anti- k_t [42] jet clustering algorithms. 100-particle toy event, 4 hard seeds, $R = 0.5$. Row 1: collision products. Row 2: distance metric; dashed circles show radius R . Row 3: merge order. Row 4: jet boundaries forming. Row 5: final jets. Soft particles colored by assigned jet in rows 4–5.

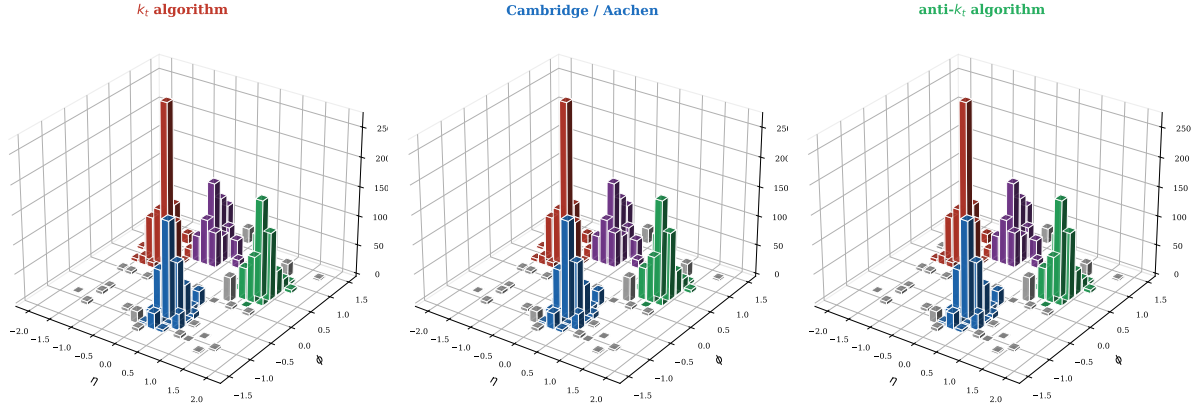


FIGURE 3.15: *Lego plots of the toy collision event for the k_t , Cambridge/Aachen, and anti- k_t jet clustering algorithms. Each bar shows the summed transverse momentum $\sum p_T$ deposited in a $(\Delta\eta \times \Delta\phi)$ simulated calorimeter cell. Bars are colored by their assigned jet; gray bars fall outside all jet cones. The dominant spike in each plot corresponds to the hardest jet (A , $p_T = 220$ GeV).*

when an arbitrarily soft particle is added to the final state. Collinear safety requires that the value is unchanged when a constituent is replaced by two collinear daughters carrying the same total momentum. Observables that fail either of these conditions receive divergent real-emission corrections that are not cancelled by virtual contributions, exactly as in the eikonal structure of the NLO calculation discussed earlier in this chapter. The perturbative expansion then breaks down and the observable can only be predicted through resummation or Monte Carlo simulation. In the Lund plane view from earlier, we have that the IRC safe jet substructure variables do not resolve the lower right corner of the plane. With the exception of the track multiplicity discussed below, all variables introduced in this section satisfy the IRC safety requirement by construction.

The variables employed in this analysis fall into two broad classes. The first consists of fragmentation-sensitive observables, which characterize how the jet's total momentum is distributed among individual charged particles. The second consists of radiation-pattern observables, which measure the p_T -weighted angular spread of constituents inside the jet cone and have closer connections to the analytic structure of the QCD parton shower. All variables are computed exclusively from charged particle tracks, for reasons discussed at the end of this section.

Track selection applied throughout:

- Charged final-state particles only (neutral hadrons and invisible particles excluded);
- $p_T > 0.5$ GeV and $|\eta| < 2.5$;

- Angular distance from the jet axis: $\Delta R < R = 0.4$.

The jet transverse momentum p_T appearing in normalization factors includes all jet constituents, both charged and neutral.

Track Multiplicity

The simplest quantity one can construct from the jet interior is the number of charged particles it contains. The track multiplicity N_{tracks} counts all charged final-state particles passing the selection above, and is the most direct experimental probe of the jet fragmentation.

As noted above, N_{tracks} is not infrared safe: adding an arbitrarily soft charged particle to the event increments the count by one, and no cancellation with virtual corrections occurs. Fixed-order perturbative predictions for the multiplicity distribution are therefore meaningless, and reliable predictions require either all-orders resummation or Monte Carlo simulation.

Despite this theoretical limitation, N_{tracks} is among the most powerful quark/gluon discriminants available. The reason is rooted in the color structure of QCD. Consider a parton shower initiated by a quark or a gluon. At each branching, the probability of emitting an additional gluon is proportional to the color charge of the radiating parton: $C_A = 3$ for a gluon and $C_F = 4/3$ for a quark. Because $C_A > C_F$, a gluon-initiated shower produces more secondary emissions at every step of the cascade, and the final hadron multiplicity is correspondingly higher. At leading logarithmic accuracy, the ratio of the mean multiplicities of gluon and quark jets is

$$\frac{\langle N_{\text{tracks}}^{\text{gluon}} \rangle}{\langle N_{\text{tracks}}^{\text{quark}} \rangle} \longrightarrow \frac{C_A}{C_F} = \frac{9}{4},$$

though in practice this ratio is reduced by subleading corrections and is typically observed to be closer to 1.5–1.7 at LHC energies. Predictions for N_{tracks} are consequently highly sensitive to the details of both the parton shower and the hadronization model, making it an especially revealing observable for the Monte Carlo comparison carried out in this study.

Leading Track Observables

Two variables characterize the hardest charged particle inside the jet. Both probe the high- z end of the fragmentation function, where z denotes the fraction of the jet momentum carried by a single hadron.

The first is the angular separation between the leading track and the jet axis,

$$\Delta R(\text{lead trk, jet}) = \sqrt{(\eta_{\text{lead}} - \eta_{\text{jet}})^2 + (\phi_{\text{lead}} - \phi_{\text{jet}})^2}.$$

In a hard, well-collimated jet initiated by a high-transverse momentum quark, the leading fragment tends to carry a large fraction of the jet momentum and to travel close to the reconstructed jet axis, giving a small value of ΔR . In a broader gluon jet, where the shower develops more secondary emissions and no single fragment dominates, the leading track is more likely to be displaced from the axis. The distribution of this variable therefore carries information both about the collimation of the jet and about the balance between hard and soft fragmentation.

The second leading-track variable is the momentum fraction

$$\frac{p_T^{\text{lead}}}{p_T^{\text{jet}}},$$

which measures what fraction of the total jet transverse momentum is carried by the single hardest charged particle. The denominator includes all jet constituents, charged and neutral, while the numerator is a single charged track. The ratio is therefore sensitive not only to the fragmentation function but also to the charged-to-neutral momentum balance of the jet, which is a non-perturbative quantity that varies between generators at the level of their hadronization models. A value close to unity signals a jet dominated by a single hard fragment, while small values indicate that the momentum is shared among many constituents. Neither leading-track variable is IRC safe.

Angularity

Moving beyond the properties of a single constituent, to characterize the global momentum-weighted shape of the jet, one introduces the class of observables known as angularities. These are p_T -weighted moments of the angular distribution of constituents inside the jet, parametrized by an angular exponent β that controls the sensitivity to wide-angle versus collinear radiation. For a general exponent β , the angularity is

$$\lambda^{(\beta)} = \frac{\sum_{i \in \text{tracks}} p_{T,i} \Delta R_i^\beta}{\sum_{i \in \text{tracks}} p_{T,i}},$$

where ΔR_i is the angular distance of constituent i from the jet axis. The p_T normalization ensures collinear safety: if one constituent splits into two collinear daughters at angle $\Delta R \rightarrow 0$, their combined contribution to the numerator approaches that of the parent, and the sum is unchanged. Infrared safety follows from the p_T weighting, which suppresses the contribution of soft particles so that adding a gluon with $p_T \rightarrow 0$ leaves the observable unchanged. For $\beta \geq 1$ the observable is therefore IRC safe.

The choice of β interpolates between different physical regimes. For small β the power of ΔR in the numerator is small, so the observable is relatively insensitive to the angular

position of constituents and probes primarily the p_T distribution. As β increases, wider-angle radiation receives more weight and the observable becomes increasingly sensitive to the global breadth of the jet.

Common choices for the parameters are $\beta = 2$ or $\beta = 1$,

$$\lambda^{(2)} = \frac{\sum_{i \in \text{tracks}} p_{T,i} \Delta R_i^2}{\sum_{i \in \text{tracks}} p_{T,i}}.$$

The quadratic angular weight suppresses collinear radiation and gives relatively more weight to wide-angle constituents. A broad jet with radiation spread throughout the cone will yield a large value of $\lambda^{(2)}$, while a narrow jet concentrated near the axis will yield a small value. Since gluon jets are broader on average than quark jets at the same p_T , $\lambda^{(2)}$ is an effective quark/gluon discriminant. Its sensitivity to wide-angle radiation also makes it particularly susceptible to non-perturbative corrections from the underlying event and from the hadronization of soft secondary partons.

Energy Correlation Observables

A natural generalization of the angularity framework is to restrict the p_T -weighted angular moments to specific angular regions inside the jet. This allows one to probe the radiation pattern at different scales and to distinguish the structure of the jet core from the wide-angle periphery, isolating regions where non-perturbative effects are expected to be most pronounced.

The first such variable is

$$U_1 = \frac{\sum_{i \in \text{tracks}}^{\Delta R_i < 0.7} p_{T,i} \Delta R_i}{\sum_{i \in \text{tracks}} p_{T,i}},$$

a first angular moment restricted to tracks within $\Delta R < 0.7$ and normalized to the total track p_T . It is sensitive primarily to the overall broadening of the jet and to the balance between central and peripheral radiation.

The second variable probes the innermost region of the jet,

$$M_2 = \frac{\sum_{i \in \text{tracks}}^{\Delta R_i < 0.3} p_{T,i} \Delta R_i^2}{\sum_{i \in \text{tracks}}^{\Delta R_i < 0.3} p_{T,i}},$$

a second angular moment restricted to $\Delta R < 0.3$, corresponding to the inner three-quarters of the jet cone by area. Both the numerator and the denominator are restricted to the same region, so the normalization is local to the core. This means that M_2 measures the intrinsic angular spread of the core independently of how much total momentum falls inside $\Delta R = 0.3$. By restricting to small angles, where the strong coupling α_s is large and perturbative control is limited, M_2 is particularly sensitive to hadronization and to the details of the soft collinear shower, making it one of the observables most susceptible to differences between Monte Carlo generators.

***N*-subjettiness**

The angularity-based variables introduced above measure the global shape of a jet but do not directly address its prong structure: whether the radiation is organized into one, two, or more distinct hard cores. This question is central both to the identification of boosted hadronic decays and to quark/gluon discrimination, and it is answered by *N*-subjettiness. The construction begins by choosing *N* candidate subjet axes inside the jet. One then assigns each constituent to its nearest axis and sums the p_T -weighted angular distances,

$$\tau_N = \frac{1}{p_{T\text{jet}} R} \sum_{i \in \text{constituents}} p_{T,i} \min_{k=1,\dots,N} \Delta R_{i,k},$$

where $\Delta R_{i,k}$ is the distance of constituent *i* from axis *k* and the prefactor $p_{T\text{jet}} R$ normalizes the result to the unit interval. The observable is small when the jet's radiation is well described by *N* prongs, since each constituent is then close to its nearest axis, and large when more prongs are required, since some constituents will be far from all axes. *N*-subjettiness is IRC safe: adding a soft particle or splitting a collinear one changes the sum by a negligible amount.

The discrimination power of *N*-subjettiness is maximized by choosing the subjet axes to minimize τ_N over all possible axis configurations. In this study τ_2 is used, making the observable sensitive to two-pronged structure, due to the structure of the decay of the $H \rightarrow Za$ process, in which *a* decays into two highly-boosted partons. A quark jet, which typically develops a single dominant core, will have small τ_2 . A gluon jet, which undergoes harder wide-angle splittings and is more likely to develop a second distinct core, tends to yield larger values.

Track-based Observables and Non-perturbative Sensitivity

All variables in this section are computed from charged particle tracks rather than from the full constituent list of the jet. This choice is motivated by two separate considerations. From the experimental side, silicon tracking detectors provide superior angular resolution compared to calorimeters, particularly for highly collimated jets where adjacent calorimeter towers begin to merge. Track-based substructure variables are better resolved and less affected by pileup, since tracks can be associated to the primary interaction vertex through impact parameter requirements in a way that calorimeter deposits cannot.

From the theoretical side, restricting to charged particles introduces an additional source of non-perturbative sensitivity beyond the usual hadronization corrections to the parton-level observable. The charged-to-neutral momentum ratio of a jet is not a perturbative quantity: it depends on the detailed composition of the hadron spectrum produced during fragmentation, and in particular on which neutral hadrons are produced and what fraction of the jet energy they carry. Even two generators that agree on the parton-level jet shape can disagree on the track-based observable simply because they predict different charged-to-neutral fractions.

The comparison between hadronization-on and hadronization-off samples presented in the following chapters therefore measures the combined sensitivity of these track-based variables to all non-perturbative effects simultaneously: hadronization, the underlying event, and the charged-to-neutral decomposition of the fragmentation function. Disentangling these contributions using the variables alone is in general not possible, and differences between the two generators should be interpreted as the total modeling uncertainty of this non-perturbative sector.

The seven variables are summarized in Table (3.2).

Variable	Definition	Angular weight	IRC safe
N_{tracks}	charged multiplicity	—	no
$\Delta R(\text{lead trk, jet})$	leading track offset	—	no
$p_T^{\text{lead}}/p_T^{\text{jet}}$	leading track momentum fraction	—	no
$\lambda^{(2)}$	angularity, $\beta = 2$	ΔR^2	yes
U_1	first angular moment	ΔR ($\Delta R < 0.7$)	yes
M_2	core second angular moment	ΔR^2 ($\Delta R < 0.3$)	yes
τ_2	N-subjettiness, $N = 2$	$\min_k \Delta R$	yes

TABLE 3.2: Summary of the jet substructure variables used in this study. All variables are computed from charged particle tracks with $p_T > 0.5$ GeV, $|\eta| < 2.5$, and $\Delta R < R = 0.4$. The IRC safety column refers to the theoretical status of the observable at parton level.

3.2.7 Jet Grooming

The substructure variables introduced in the previous section are powerful probes of the internal radiation pattern of a jet. However, they share a common limitation when computed on the raw output of a clustering algorithm: they are all sensitive, to varying degrees, to soft wide-angle radiation that has nothing to do with the hard process of interest. This contamination has two main sources. The first is the underlying event, the collection of soft interactions among the spectator partons in the colliding protons that accompanies every hard collision and deposits low- p_T hadrons throughout the detector. The second is pileup, the additional inelastic proton–proton collisions that occur in the same bunch crossing as the hard scatter and contribute soft particles that can fall inside the jet cone. Both effects inflate the apparent mass and breadth of a jet and introduce

non-perturbative corrections that are difficult to compute analytically and that differ significantly between Monte Carlo generators.

Grooming algorithms address this problem by systematically removing soft wide-angle radiation from the jet before computing substructure variables, recovering a cleaner perturbative object whose properties are more theoretically tractable and whose modeling uncertainties are reduced. The key idea is that the hard collinear structure of the jet, which carries the physically relevant information, is well separated in phase space from the soft wide-angle contamination, and that this separation can be exploited by an appropriate procedure operating on the jet's internal branching history.

The Cambridge/Aachen Declustering Tree

All modern grooming algorithms operate on the Cambridge/Aachen clustering history of the jet rather than directly on its constituent list. The reason is that the Cambridge/Aachen algorithm, which uses a purely geometric distance metric $d_{ij} = \Delta R_{ij}^2/R^2$ independent of transverse momentum, produces an angle-ordered branching tree: when the clustering history is run in reverse to decluster the jet, each successive splitting corresponds to a wider-angle emission than the previous one. The declustering tree therefore provides a natural angularly-ordered decomposition of the jet's radiation, from the hard collinear core at small angles outward to the soft wide-angle periphery.

This is precisely the structure that grooming exploits. Starting from the root of the tree, which corresponds to the widest-angle splitting in the jet, the grooming procedure walks inward along the harder branch, evaluating at each node whether the emission is sufficiently hard to be considered part of the genuine jet. Soft wide-angle emissions are encountered first and removed, and the procedure terminates as soon as it reaches a hard enough splitting, leaving the inner hard core intact.

SoftDrop

The grooming algorithm used in this analysis is **SoftDrop**, a generalization of the earlier modified mass drop tagger. The jet constituents are first reclustered with the Cambridge/Aachen algorithm to obtain the angle-ordered declustering tree. The tree is then walked from the root outward. At each splitting into two branches with transverse momenta $p_{T,1} \geq p_{T,2}$ separated by an angle ΔR , the **SoftDrop** condition

$$z \equiv \frac{p_{T,2}}{p_{T,1} + p_{T,2}} \geq z_{\text{cut}} \left(\frac{\Delta R}{R} \right)^\beta$$

is evaluated. If the condition is satisfied, the splitting is considered sufficiently hard and the grooming terminates, returning the current jet as the groomed result. If the condition fails, the softer branch is discarded as soft contamination and the procedure continues recursively on the harder branch.

The two free parameters of **SoftDrop** have distinct physical interpretations. The parameter z_{cut} sets the overall momentum asymmetry threshold under which a splitting is

considered soft: a larger value results in more aggressive grooming, removing a larger fraction of the jet's constituents. The parameter β controls the angular dependence of this threshold. For $\beta = 0$, known as the modified mass drop tagger limit, the threshold is flat and equal to z_{cut} at all angles: splittings with $z < z_{\text{cut}}$ are always removed, regardless of their angle. For $\beta > 0$, the threshold scales as $(\Delta R/R)^\beta$ and therefore decreases as ΔR decreases. This means the groomer becomes progressively more permissive at small angles, removing wide-angle soft emissions while retaining narrow-angle ones even if they carry a small momentum fraction. The choice $\beta = 1$ used in this study represents a moderate grooming that is more aggressive at large angles and more lenient toward the collinear core of the jet.

The physical picture is most transparent when viewed in the Lund plane. In this representation, the **SoftDrop** condition defines a straight boundary in the $(\ln 1/\theta, \ln k_t)$ plane: emissions above this line are retained and those below are removed. We can see this if we remind ourselves of the axes of the Lund plane:

$$x = \ln(1/\Delta R), \quad y = \ln z \Delta R p_{T\text{jet}} = \ln k_t,$$

we substitute the **SoftDrop** boundary, $z = z_{\text{cut}}(\Delta R/R)^\beta$, into $\ln k_t$:

$$\begin{aligned} \ln k_t &= \ln z + \ln \Delta R + \ln p_{T\text{jet}} = \\ &= \ln z_{\text{cut}} + \beta \ln \Delta R + \ln \Delta R + \ln p_{T\text{jet}} = \\ &= \ln z_{\text{cut}} - (1 + \beta) \ln(1/\Delta R) + \ln p_{T\text{jet}} \end{aligned}$$

The **SoftDrop** condition turns into:

$$y > \ln z_{\text{cut}} - (1 + \beta) \ln(1/\Delta R) + \ln p_{T\text{jet}}.$$

Anything below the condition is then discarded, while the rest is kept, as in Figure (3.16). The groomed jet corresponds to the primary Lund-plane emission that first satisfies the condition, together with all radiation collinear to it. This geometric picture makes the theoretical properties of **SoftDrop** particularly transparent: the groomed jet mass, for instance, receives contributions only from emissions in a well-defined region of the Lund plane and can be computed analytically at next-to-leading logarithmic accuracy.

Effect on Substructure Observables

The primary motivation for applying grooming in this analysis is to disentangle the perturbative and non-perturbative contributions to the substructure variables. A given variable computed on the raw jet receives contributions from three sources: the hard perturbative shower, the soft non-perturbative hadronization, and the underlying event. When the same variable is computed on the groomed jet, the underlying event contribution is largely removed, since it consists precisely of the soft wide-angle radiation that **SoftDrop** discards. The comparison between the groomed and ungroomed distributions therefore

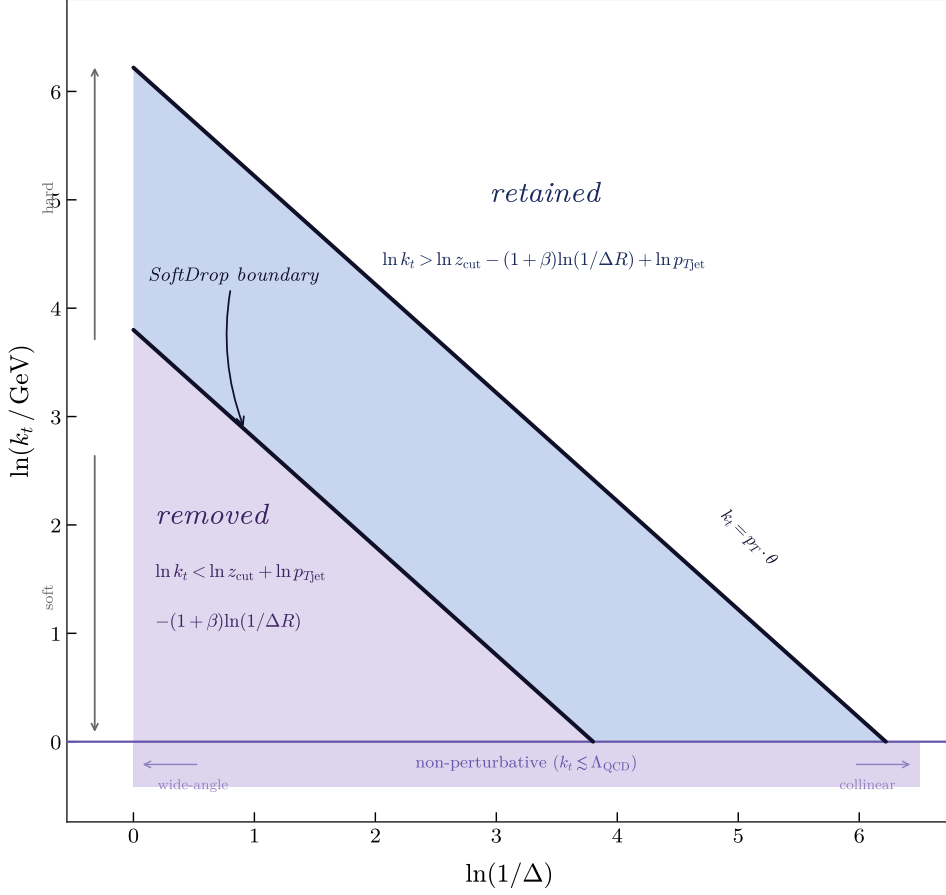


FIGURE 3.16: The *SoftDrop* boundary visualized in the Lund plane. Anything that is below the boundary is discarded.

isolates the contribution of the underlying event, and comparing their hadronization-on and hadronization-off ratios measures the degree to which the ungroomed observable is polluted by this contamination.

The availability of both groomed and ungroomed versions of the substructure variables in this analysis allows one to separate the question of whether two generators agree on the hard collinear structure of the jet from the question of whether they agree on the soft wide-angle environment. A disagreement on the ungroomed variables but agreement on the groomed ones would point specifically to the modeling of the underlying event, while persistent disagreement after grooming would indicate genuine differences in the perturbative shower or in the collinear hadronization. This separation is essential for interpreting the PYTHIA–HERWIG comparison presented in the following Chapter.

Section Roadmap:**Covered:**

Jet formation from parton-level dynamics to hadronization; soft and collinear divergences in Quantum Chromodynamics and their role in shaping jet evolution; jet clustering algorithms and the definition of reconstructed jets; jet substructure techniques as tools for probing the internal structure of jets; non-perturbative effects relevant for jet observables.

References:[43]; [5]; [44]; [40]; [41]; [42] [45]; [46]; [47]; [48]; [49]; [50]; [6]; [51]; [52]; [6].



3.3 Monte Carlo Event Generators: *PYTHIA* and *HERWIG*

3.3.1 Why Monte Carlo Simulation Is Necessary

The theoretical description of a proton–proton collision at the LHC spans an enormous range of energy scales, from the hard scattering at hundreds of GeV down to the non-perturbative regime at ~ 1 GeV where quarks and gluons are confined into hadrons, and as we know, it is a very complex endeavor to reconstruct it. No single analytic framework is capable of describing this entire chain from theoretical formalisms. At the highest scales, perturbative QCD (pQCD) provides systematic predictions via the factorization theorem, but the perturbative expansion breaks down as the strong coupling α_s grows larger at low scales. At those scales, non-perturbative effects — hadronization, confinement, the underlying event — must be modeled phenomenologically due to the non-perturbativity of the objects that are involved in such calculations, PDFs. Furthermore, even where pQCD is applicable, the final-state phase space has dimension $3n - 4$ for an n -particle event; at the LHC, n can easily reach $\mathcal{O}(100)$, making direct numerical integration over the full phase space impossible.

Monte Carlo event generators address all three of these challenges simultaneously. They exploit the factorized structure of QCD to simulate each regime of the collision separately and then stitch them together into a fully exclusive event. The large dimensionality of phase space is handled naturally by Monte Carlo integration, whose statistical error scales as $1/\sqrt{N}$ regardless of dimension, making it the most efficient way of carrying out such a calculation. The non-perturbative sector is encoded in physically motivated models with tunable parameters, fitted to experimental data. The result is a program that takes the Lagrangian of the Standard Model as figurative input and produces, as output, a list of stable four-momenta indistinguishable in format from what a detector records.

3.3.2 Structure of a Simulated Event

A general-purpose MC event generator simulates a proton–proton collision through a sequence of well-defined stages, ordered by decreasing momentum scale. Each stage is treated with the appropriate theoretical tool, and they are combined consistently through the factorization theorem. The stages are the following.

1. Hard subprocess. The simulation begins with the selection of a hard scattering process — for example, the production of a $q\bar{q}$ or gg pair at large transverse momentum. The user specifies the process type; the generator then extracts the initial-state parton momenta from the PDFs of the incoming protons, and generates the partonic kinematics according to the relevant tree-level (or NLO) matrix element squared, $|\mathcal{M}_{ij \rightarrow X}|^2$. This is the only step of the simulation that is derived directly from the Feynman rules of QCD

without further approximation. In practice, the hard matrix element is commonly evaluated using MADGRAPH5_AMC@NLO, or POWHEG, which automates the generation of Feynman diagrams and the computation of $|\mathcal{M}_{ij \rightarrow X}|^2$ for arbitrary processes at LO and NLO in QCD. The user specifies the process in a simple syntax, and the generator produces weighted partonic events in the Les Houches Event (LHE) format, which are subsequently passed to a parton-shower program.

2. Parton shower. The outgoing colored partons produced in the hard process are not free: they carry color charge and radiate gluons, which can in turn split into further quark–antiquark pairs or radiate more gluons. This cascade of emissions — the parton shower — continues until the typical scale of the process falls to the non-perturbative threshold $Q_0 \sim 1$ GeV, where the shower is terminated.

In the leading-logarithmic (LL) approximation, the probability that a parton a of type a splits into two partons b and c during the shower evolution is governed by the DGLAP splitting functions¹² $P_{a \rightarrow bc}(z)$, where z is the longitudinal momentum fraction carried by parton b . The three relevant splitting functions at leading order are:

$$\begin{aligned} P_{q \rightarrow qg}(z) &= C_F \frac{1+z^2}{1-z}, \\ P_{g \rightarrow gg}(z) &= C_A \frac{(1-z(1-z))^2}{z(1-z)}, \\ P_{g \rightarrow q\bar{q}}(z) &= T_R [z^2 + (1-z)^2], \end{aligned}$$

where $C_F = 4/3$, $C_A = 3$, and $T_R = 1/2$ are the QCD color factors. The larger value of C_A relative to C_F means that gluon-initiated showers are more numerous than quark-initiated ones — a fact that directly determines why gluon jets are broader and more populated than quark jets, as we covered in the previous sections.

The shower evolution is monitored by the Sudakov form factor $\Delta_a(t_1, t_2)$, which calculates the probability that parton a evolves from scale t_1 to a lower scale t_2 without emitting a new parton. It is defined as:

$$\Delta_a(t_1, t_2) = \exp \left(- \int_{t_2}^{t_1} \frac{dt}{t} \frac{\alpha_s(t)}{2\pi} \sum_{b,c} \int_{z_-}^{z_+} dz P_{a \rightarrow bc}(z) \right),$$

where the sum runs over all allowed splittings of parton a . The Sudakov factor is an the all-orders probability of no emission between the two scales. This resummation is

¹²DGLAP stands for Dokshitzer–Gribov–Lipatov–Altarelli–Parisi. The DGLAP evolution equations describe how parton distribution functions $f_a(x, \mu^2)$ evolve with the factorization scale μ^2 :

$$\mu^2 \frac{\partial f_a(x, \mu^2)}{\partial \mu^2} = \frac{\alpha_s(\mu^2)}{2\pi} \sum_b \int_x^1 \frac{dz}{z} P_{a \leftarrow b}(z) f_b\left(\frac{x}{z}, \mu^2\right),$$

where $P_{a \leftarrow b}(z)$ is the probability that parton b produces parton a carrying a fraction z of b 's momentum.

the key feature that makes parton showers accurate in the collinear and soft limits, even though no single fixed-order calculation is used they are therefore IRC safe. In practice, the shower is generated by repeated application of the Sudakov veto algorithm: a trial emission scale is drawn from Δ_a , and accepted or rejected according to the full splitting probability; if rejected, the evolution continues from the trial scale downward until one reaches hadronization scale.

3. Underlying event. In a proton–proton collision, the two hard-scattering partons are not the only constituents of the incoming protons. The remaining spectator partons — those not involved in the primary hard scatter — can undergo additional softer interactions with each other. These secondary scatterings give rise to the so-called underlying event (UE): a distribution of soft particles spread throughout the detector that accompanies every hard event, and acts as background noise. The UE is modeled through multiple parton interactions (MPI), which will be discussed in later sections.

4. Hadronization. Once the shower reaches the non-perturbative scale Q_0 , the system consists of a set of quarks and gluons connected by a specific color flow — the color topology¹³ inherited from the hard process and shower. At this point, confinement must set in: the colored partons must be rearranged into color-singlet hadrons. This cannot be computed from pQCD, and is handled by phenomenological hadronization models, implemented in the code. The two most widely used models — the Lund string model (*PYTHIA*) and the cluster model (*HERWIG*) — are described in upcoming sections.

5. Hadron decays. Many of the hadrons produced during hadronization are unstable resonances, such as π^0 , K_S^0 , ρ , B mesons, and so on. These are subsequently decayed to the lighter, long-lived particles that reach the detector. Sophisticated decay tables and spin-correlation treatments are employed to ensure accurate final-state topologies. A crucial property of this framework is its approximate universality: the hadronization models, once tuned to data from one collider environment (for example, e^+e^- annihilation at LEP), are expected to be valid at the LHC as well, since hadronization is a property of color-connected parton systems and is largely independent of how those systems were produced.

3.3.3 Parton Showers and Hadronization Modeling: *PYTHIA* and *HERWIG*

The two generators employed in this analysis, *PYTHIA* 8 and *HERWIG* 7, implement distinct approaches to both parton showering and hadronization. These differences are not merely technical choices: they reflect genuinely different physical approximations,

¹³The color topology is simply the successive application of color conservation for each of the splitting vertices.

and lead to systematic differences in jet-level predictions that are central to the PYTHIA–HERWIG comparison carried out in the following Chapter.

PYTHIA and the $p_{\perp}$ -ordered shower. PYTHIA evolves the parton shower using the transverse momentum $p_{\perp}$ of successive emissions as the ordering variable. Starting from the hard scale $p_{\perp,\text{hard}}$, the shower steps downward in $p_{\perp}$, generating emissions at progressively lower transverse momenta. This choice of ordering variable naturally incorporates the dominant soft–collinear singularity structure of QCD, and the shower correctly resums leading and next-to-leading logarithms of $p_{\perp}$ via the Sudakov form factor.

For hadronization, PYTHIA employs the Lund string model. The physical picture is rooted in the linear confinement of QCD: the color-electric field between a separating $q\bar{q}$ pair does not spread outward as in electromagnetism, but is instead compressed into a narrow flux tube, or string, with an elastic potential of approximately constant energy density $\kappa \approx 1 \text{ GeV/fm}$. As the quark and antiquark move apart, the string stretches and stores potential energy. When the stored energy is sufficient, the string breaks via the quantum-mechanical tunnelling of a new $q\bar{q}$ pair from the vacuum, splitting the original string into two shorter strings. This process repeats iteratively until the remaining string pieces carry too little invariant mass to break further, at which point they are identified as the final hadrons. The model predicts a characteristically steeply falling fragmentation function, in good agreement with data. In multi-parton events, gluons produced in the shower appear as kinks on the string, and the string connects all color-charged partons in the event in a topology determined by the large- N_c color flow. This process is depicted in Figure (3.17).

HERWIG and the angular-ordered shower. HERWIG uses the emission angle θ as the ordering variable for the parton shower, rather than $p_{\perp}$. Angular ordering emerges naturally from the coherent treatment of soft gluon emission: in QCD, soft gluons emitted at large angles from a color-connected quark–gluon system interfere destructively due to color coherence, and the net effect is that successive emissions must be ordered in decreasing angle. Without this constraint, the shower would over-count soft wide-angle emissions and predict a particle multiplicity that grows too rapidly with energy. Angular ordering resums these coherence effects at leading logarithmic accuracy.

For hadronization, HERWIG employs the cluster model. At the end of the angular-ordered shower, all gluons that have not yet branched are split non-perturbatively into $q\bar{q}$ pairs. The resulting quarks and antiquarks are then paired according to their color connections into color-singlet clusters — pre-hadronic objects with a distribution of invariant masses M_{cluster} that is sharply peaked near 1 GeV due to the angular-ordered shower structure. Each cluster then decays isotropically (in its rest frame) into a pair of hadrons, chosen according to the available phase space and the quantum numbers of the cluster. The model has fewer free parameters than the Lund string, because the cluster mass distribution is largely fixed by the perturbative shower; however, it is less flexible in describing the details of the fragmentation function at high z .

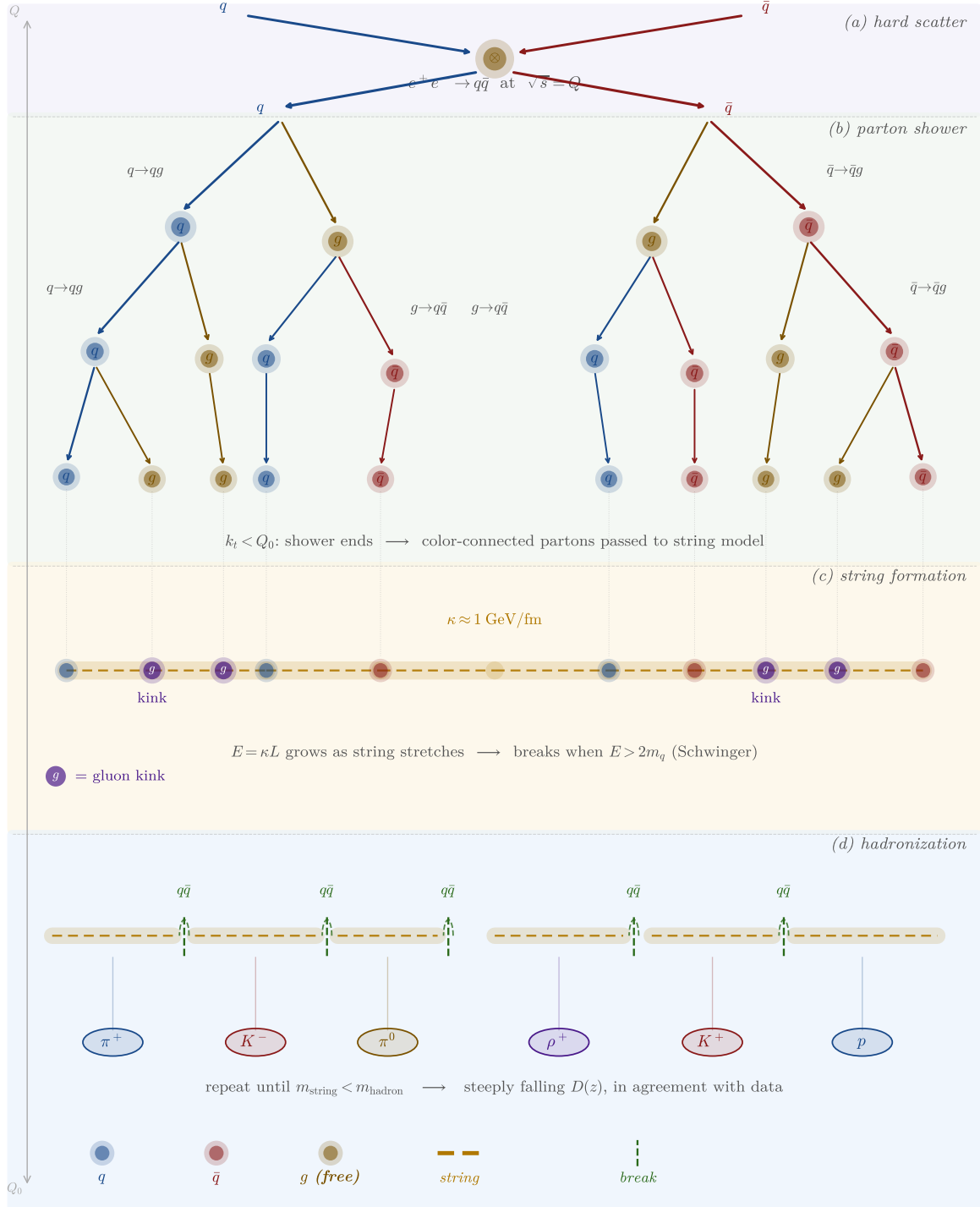


FIGURE 3.17: Visualization for the Lund hadronization string model adopted by PYTHIA.

Comparison of the two approaches. The two generators therefore differ at every stage of the simulation beyond the hard matrix element. The shower ordering variable ($p_\perp$ vs. θ), the treatment of color coherence, and the hadronization model (string vs. cluster) are all distinct. These differences are not independent: the angular-ordered shower in HERWIG naturally produces clusters of narrow invariant mass, which is precisely the prerequisite for the cluster model to work well. Conversely, the $p_\perp$ -ordered shower in PYTHIA spans a broader range of color topologies, and the string model's global treatment of color flow is better suited to accommodate this. The systematic comparison of predictions from both generators is therefore a powerful probe of the sensitivity of observables to the non-perturbative sector of QCD, and their analysis is fundamental for the following sections.

3.3.4 The Underlying Event and Multiple Parton Interactions

At the LHC, each colliding proton is a composite object: it contains not only the two or three valence quarks participating in the hard subprocess, but also a sea of gluons and virtual $q\bar{q}$ pairs. In a given collision, only one pair of partons from the two protons undergoes the primary hard scattering; the remaining partons — the spectators — continue moving through each other and can themselves interact softly. These secondary interactions collectively constitute the underlying event.

The UE is modeled through multiple parton interactions (MPI). Because these additional interactions are softer than the primary hard process, their cross sections are enhanced by the large value of α_s at low scales, and the average number of MPI per event is of order a few at LHC energies. Each additional interaction deposits soft particles throughout the event, adding to the jet multiplicity, broadening the jet energy profile, and increasing the apparent jet mass.

In PYTHIA, the MPI are generated in the same $p_\perp$ -ordered downward evolution as the parton shower, and are intertwined with it: at each step, the generator chooses randomly between a shower emission and an additional MPI, with the relative rates governed by their respective cross sections. The PDFs for each successive interaction are explicitly reweighted to account for the momentum already removed by earlier interactions, ensuring exact energy-momentum conservation.

In HERWIG, the expected number of MPI per event is computed from an eikonal model of the proton-proton overlap function, which depends on the assumed transverse density profiles of the two incoming protons. The interactions are generated sequentially and terminated once the available center-of-mass energy is exhausted.

The UE contributes directly to the jet substructure observables studied in this thesis. Since MPI deposit soft, wide-angle particles inside the jet cone, they inflate the values of observables such as the angularity $\lambda^{(2)}$ and the jet multiplicity N_{tracks} , in particular for ungroomed jets. The **SoftDrop** grooming algorithm discussed in the previous Section removes this contamination by discarding soft wide-angle emissions below a momentum asymmetry threshold, allowing the perturbative jet core to be studied independently of the UE. The comparison of groomed and ungroomed distributions between PYTHIA and

HERWIG therefore provides direct sensitivity to the differences in the MPI models of the two generators and is a good estimator for systematic uncertainties.

3.4 Jet Reconstruction and Neural Networks

Signal events that are of interest for this work are characterized by a highly boosted light pseudoscalar resonance a , whose hadronic decay products are collimated into a single narrow jet. Due to the low mass of the resonance ($m_a \lesssim 4$ GeV) and its large Lorentz boost, the decay products cannot be resolved as separate jets, and the resonance is instead reconstructed as a single anti- k_T jet with radius parameter $R = 0.4$, clustered with **FastJet**. Jet reconstruction is performed on the final-state stable particles from the shower output of both PYTHIA 8 and HERWIG 7, allowing systematic comparisons between the two parton-shower models. Jets are required to satisfy $p_T > 20$ GeV and $|\eta| < 2.5$. In events containing multiple jets, the highest- p_T jet is taken as the resonance candidate, provided that the three-body invariant mass $m_{\ell\ell j} < 250$ GeV.

From the particles associated to the leading jet, a suite of substructure discriminants is computed and used as input to the neural networks. These include the number of associated particles (N_{tracks}), the ratio of the leading-particle p_T to the scalar sum of all associated particle p_T values (**leadTrackPtRatio**), the angular separation between the leading particle and the jet axis (ΔR_{lead}), N -subjettiness τ_2 (evaluated with the exclusive k_T sub-jet algorithm to measure compatibility with a two-pronged decay), the modified energy-correlation functions $U_1(0.7)$ and $M_2(0.3)$, and the angularity variable $\text{angularity}_{(2)}$.

Because the calorimeter resolution is insufficient to reconstruct the jet mass at $\mathcal{O}(1)$ GeV, a two-stage neural network approach is employed. A regression network is first trained on signal samples spanning $m_a \in \{0.5, 0.75, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0\}$ GeV to predict the mass of the resonance directly from the substructure variables. The network is a feed-forward multi-layer perceptron implemented in **PyTorch** [53], trained using a Huber loss function [54] minimized with the ADAM optimizer [55]. Hyperparameters are optimized through Bayesian optimization using the Optuna framework [56], which achieves convergence significantly faster than conventional grid or random search strategies.

The predicted mass is then passed, together with the original substructure variables, to a classification network that discriminates signal jets from the Z +jets background. By receiving the regressor's mass estimate as an additional input, the classifier effectively operates as a parametric neural network, implicitly conditioning its output on the resonance mass hypothesis without requiring separate trainings for each mass point. The classifier uses a sigmoid output and binary cross-entropy loss, and is also optimized with ADAM and Optuna-based hyperparameter tuning, with the AUC used as the figure of merit.

Systematic uncertainties associated with the parton-shower and hadronization model are evaluated by comparing the nominal PYTHIA 8 signal samples with alternative samples generated with HERWIG 7, as described in previous sections.

3.5 Analysis Setup and Pipeline

The phenomenological analysis is organized as a sequential pipeline that transforms parton-level hard-scatter events into fully showered, jet-reconstructed, and neural-network-analyzed samples. The chain is steered by a series of shell scripts and Python utilities collected in the `pheno_h2zllahad` repository.

Hard-scatter generation. Signal events for the process $gg \rightarrow H \rightarrow Z(\ell\ell)a$ are generated at NLO in QCD using POWHEG-BOX v2, with the $H \rightarrow Za$ decay handled internally. Background Z +jets events at parton level are produced as Les Houches Event (LHE) files using MADGRAPH5_AMC@NLO.

Parton shower and hadronization. LHE¹⁴ files from both POWHEG and MADGRAPH are passed to PYTHIA 8, which handles the parton shower, hadronization, and UE modeling, using the NNPDF 2.3 LO PDF set. Signal samples are generated for mass hypotheses $m_a \in \{0.5, 0.75, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 8.0\}$ GeV, with separate programs handling the $H \rightarrow Z\eta_c$, $H \rightarrow ZJ/\psi$, and $H \rightarrow Za$ decay channels. Alternative signal samples showered with HERWIG 7 are produced to assess the impact of parton-shower modeling on the neural-network selection efficiency.

Event conversion and jet reconstruction. The shower output is stored in HepMC format and read by a dedicated Python script, compatible with both HepMC 2 and HepMC 3. The script applies a truth-level preselection requiring at least one same-flavor opposite-sign lepton pair with both leptons satisfying $p_T > 18$ GeV, a leading lepton $p_T > 27$ GeV, a dilepton invariant mass in the Z -peak window $81 < m_{\ell\ell} < 101$ GeV, and at least one anti- k_T $R = 0.4$ jet with $p_T > 20$ GeV, $|\eta| < 2.5$, and $m_{\ell\ell j} < 250$ GeV. For each passing event, jet clustering is performed with FastJet and the substructure variables described above are computed and stored alongside the jet and lepton kinematics in ROOT ntuples, which are subsequently converted to HDF5 format for efficient tensor-based processing.

Neural network training. HDF5 files for all signal mass hypotheses and the Z +jets background are supplied to the PyTorch-based training scripts. The regression network is trained first; its learned mass predictions are then appended to each event's feature vector before the classification network is trained on the augmented feature set. Feature standardization uses a robust scaler (stored alongside the model weights) to mitigate the impact of outliers. Automated hyperparameter optimization is performed with Optuna (40 trials per network by default), sampling network width and depth, learning rate, dropout, batch size, and loss choice; the best configuration is identified from the study database and used for a final training run that produces all diagnostic plots and the saved

¹⁴Les Houches Event File (LHEF) is a standard ASCII format for exchanging parton-level events between matrix element generators and parton shower programs, encoding particle momenta, flavors, color flow, and event weights.

model checkpoint. An overview of this generator pipeline is summed up in Figure (3.18).

Section Roadmap:

Covered:

The necessity of Monte Carlo simulation for multi-scale proton–proton collisions and the sequential structure of a simulated event; the DGLAP splitting functions and Sudakov form factor as the theoretical foundation of parton showers; the contrasting approaches of PYTHIA 8 ($p_\perp$ -ordered shower, Lund string hadronization) and HERWIG 7 (angular-ordered shower, cluster hadronization); modeling of the underlying event via multiple parton interactions in both generators; signal jet topology and anti- k_T reconstruction for the boosted $H \rightarrow Za$ decay; the seven track-based substructure discriminants and their use as inputs to a two-stage neural network comprising a mass regression network and a signal-versus-background classification network; the full analysis pipeline from NLO hard-scatter generation through parton showering, HepMC conversion, FastJet jet reconstruction, ROOT-to-HDF5 output, and Optuna-based hyperparameter optimization.

References: [44]; [57]; [58]; [59]; [60]; [61]



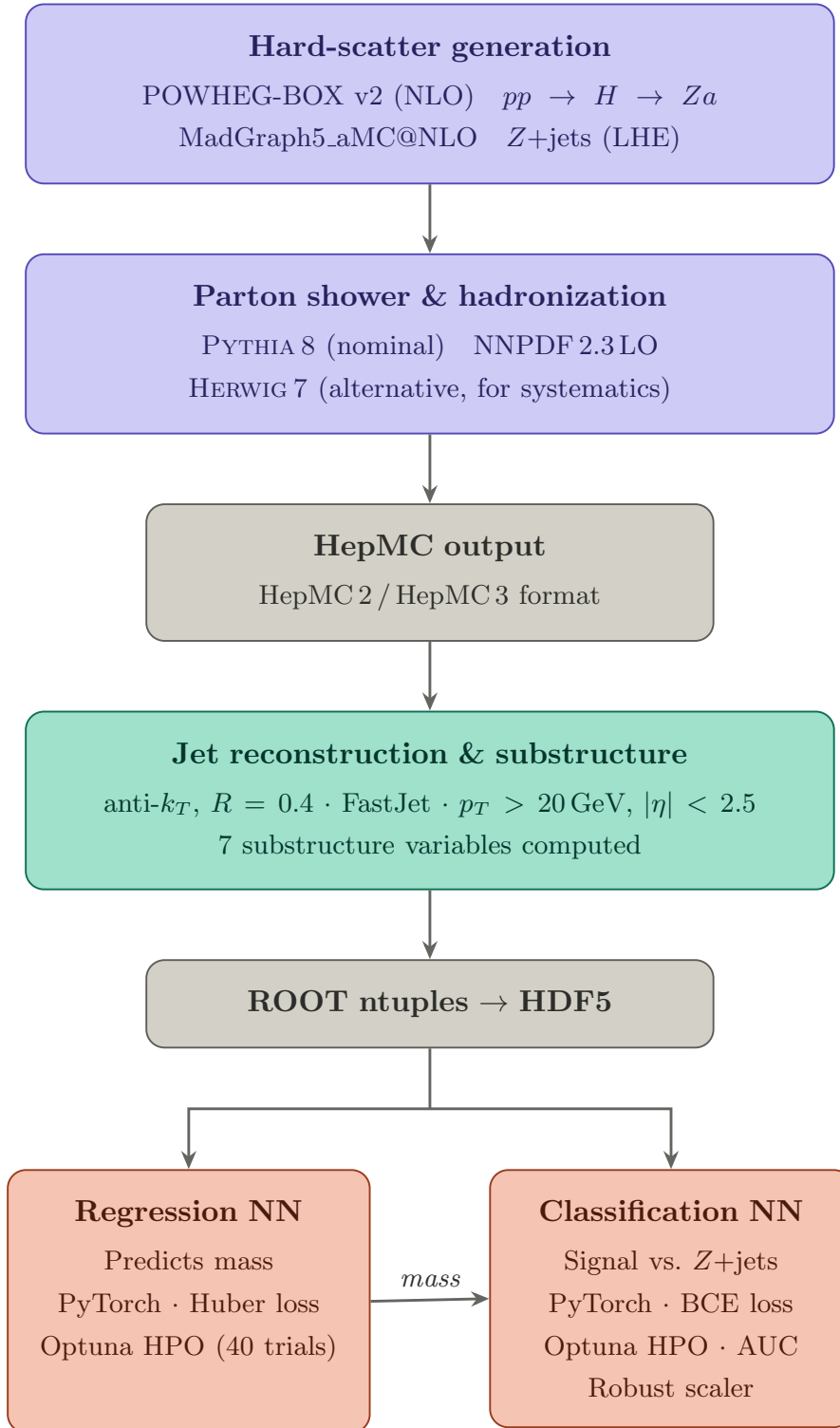


FIGURE 3.18: Generator and file pipeline.

Chapter 4

Results of $H \rightarrow Za$

“I don’t know, Mitch. With work like this, we may have to get you back here for grad school.”

— Morrie Schwartz, in Mitch Albom’s *Tuesdays with Morrie*



THE preceding three Chapters have built, layer by layer, the conceptual and technical scaffolding required to carry out a meaningful phenomenological study of the $H \rightarrow Za$ decay. Chapter 1 established the role of the Higgs boson within the Standard Model and motivated the search for deviations from it. Chapter 2 introduced the theoretical framework of the Type-1 Two-Higgs-Doublet Model, showed how a light pseudoscalar $a \rightarrow q\bar{q}/gg$ arises naturally in this context, and placed the decay $H \rightarrow Za$ within the broader landscape of exotic Higgs searches at the LHC. Chapter 3 assembled the practical tools: the experimental apparatus of the ATLAS detector, the jet substructure observables sensitive to the internal structure of a boosted a decay, the Monte Carlo generators PYTHIA 8 and HERWIG 7, and the analysis pipeline that connects parton-level events to reconstructed jet observables.

This Chapter brings all of these elements together. The goal is twofold: first, to characterize the theoretical uncertainty on the signal modeling by comparing the two generators across the set of jet substructure observables; and second, to develop and evaluate a machine-learning pipeline for signal-versus-background separation, and to assess how robust it is between different generators — a question that has emerged as central in the

subsequent ATLAS analysis, which independently pursued a neural-network approach to the same decay channel.

The Chapter is organized as follows. We begin with a review of the most recent ATLAS result for the search of the $H \rightarrow Za \rightarrow \ell\ell + \text{hadrons}$ process, which used the full Run 2 dataset of 140 fb^{-1} at $\sqrt{s} = 13 \text{ TeV}$: this serves as the experimental benchmark and the primary motivation for the improvements pursued in this work. We then present the kinematic validation of our independent simulation pipeline, establishing that the PYTHIA and HERWIG samples, when seeded by the same POWHEG hard process, reproduce consistent distributions for all reconstructed objects before any generator-specific differences are examined. The main generator comparison follows, organized into three parts: the full simulation with hadronization enabled, the **HAD-on** vs **HAD-off** study that isolates the nonperturbative contribution to each observable, and the groomed variables that assess the resilience of the observables to soft wide-angle radiation. The Chapter closes with the performance of the two-stage neural network pipeline — a regression network for pseudoscalar mass prediction and a classification network for signal-versus-background separation — organized into three studies: Study A (domain-transfer baseline), Study B (self-consistent evaluation), and Study C (balanced regression training), together a reduction in systematics and an improvement in the neural network performance.

4.1 Previous Analysis of $H \rightarrow Za$

In the current Section the previous work by the ATLAS Collaboration will be presented ([38]). The latest research on an analysis for the BSM 2HDM $H \rightarrow Za$ decay, authored by the ATLAS collaboration, has been performed using the full 140 fb^{-1} dataset of 13 TeV proton-proton collisions from Run 2.

4.1.1 Data and Simulated Event Samples

This analysis has been performed using the entire recorded dataset from LHC’s Run 2 of integrated luminosity of 140 fb^{-1} at pp center-of-mass energy of $\sqrt{s} = 13 \text{ TeV}$. For the Run 2 analysis leptonic decays of the Z boson and hadronic decays of the light resonance are considered.

This simulation is able to obtain NNLO accuracy for arbitrary inclusive $gg \rightarrow H$ observables. The successive steps, particle decays, hadronizations, parton showers and underlying events are modelled using PYTHIA 8.212 ([62]) and EVTGEN 1.6.0 ([63]) with the A14 tuning and the NNPDF23LO PDF set ([64]; [65]). The branching ratios for Higgs decay are determined using the PYTHIA 8 MSSM scenario with two Higgs doublets, and it predicts that the a ’s decay into gluons is favored until $a \rightarrow c\bar{c}$ is kinematically accessible. The signal Monte Carlo samples span the region of 0.5–4.0 GeV for the a mass. The Z is required to decay into pairs of leptons, and the alternative signal samples made up of η_c or J/ψ decays are also simulated using PYTHIA 8. HERWIG 7.1 ([66]) is also used to model the parton shower and hadronization in order to estimate uncertainties on the decay of

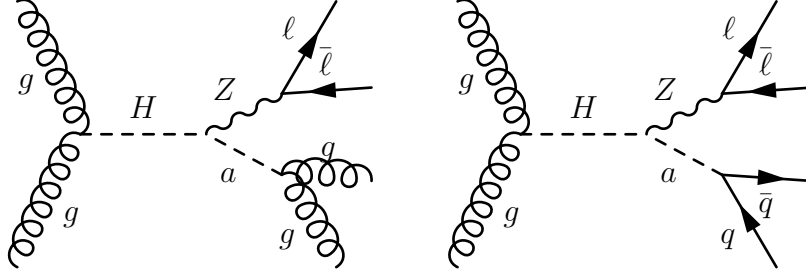


FIGURE 4.1: Feynman diagrams for the observed signal decays for the $H \rightarrow Za$ search.

the a particle into hadrons. Moreover, the simulated events ($gg \rightarrow H \rightarrow Za \rightarrow \ell\ell gg$ or $\ell\ell q\bar{q}$, as depicted in Figure 4.1) are simulated using POWHEGBOX v2 ([67]) and the PDF4LHC15NNLO ([64]; [65]) set of PDFs.

The background of this analysis is dominated mainly by Z +jet events. They are modeled in the same way as the signal samples, using the same MC generators. Only samples with decays $Z \rightarrow ee$ and $Z \rightarrow \mu\mu$ are considered. The processes ZZ and ZW constitute a small part of the background in this analysis, amounting to $\approx 1\%$ — they are modeled using SHERPA 2.2.1 ([68]). The $t\bar{t}$ decay also produces a small part of the background ($\approx 1\%$) and is modeled using POWHEG ([67]) and PYTHIA 8+EVTGEN. The MC sample also includes a full simulation of the ATLAS detector and its response, as well as other interactions between pp in bunch crossings. Table 4.1 summarizes the generators employed for each step of the simulation.

Process	Matrix element gen.	Shower+HAD+PDF+tuning
Signal		
$H \rightarrow Za$ ($m_a = 0.5$ – 4 GeV)	POWHEG-BOX v2 with reweighting to HNNLO; alternative samples with HERWIG 7.1.	PYTHIA 8.212+EVTGEN 1.6.0; PDF4LHC15NNLO / NNPDF23LO; A14 tuning.
Backgrounds		
$Z(\ell\ell)$ +jets (nominal)	POWHEG (NLO QCD).	PYTHIA 8+EVTGEN; CTEQ6L1 / AZNLO tuning.
$Z(\ell\ell)$ +jets (alternative)	SHERPA 2.2.1.	SHERPA internal; NNPDF3.0NNLO.
ZZ, ZW	SHERPA 2.2.1.	SHERPA internal; NNPDF3.0NNLO.
$t\bar{t}$	POWHEG.	PYTHIA 8+EVTGEN; standard ATLAS setup.

TABLE 4.1: Monte Carlo generators and configurations used for signal and background simulation in the ATLAS search for $h \rightarrow Za$ with $a \rightarrow \text{hadrons}$.

4.1.2 Event Selection

The kinematic selections used to separate signal from background are now described. Low-mass signal resonances are highly boosted due to the mass difference between the Higgs boson and the Z boson. The corresponding jet is therefore reconstructed with a small radius $R = 0.4$. In order not to be rejected as pile-up¹ the jets are required to pass a set of conditions interaction events: $p_T > 60$ GeV and $|\eta| > 2.4$. Electron candidates are reconstructed by matching tracks in the inner detector and energy clusters in the electromagnetic calorimeter, while muons are reconstructed from inner detector tracks matching tracks in the muon spectrometer. The pairs of leptons must satisfy the following: both should have $p_T > 18$ GeV and one of them must have $p_T > 27$ GeV. Electrons must be reconstructed within $|\eta| < 2.47$ and muons within $|\eta| < 2.7$. It is also paramount for the detected leptons to have an opposite-charge counterpart and that they have a combined dilepton invariant mass compatible with the Z boson mass: $81 < m_{\ell\ell} < 101$ GeV. Higgs boson candidates are reconstructed from jet-dilepton systems compatible with its invariant mass, with a loose pre-selection requirement of $m_{\ell\ell j} < 250$ GeV. Figure 4.2 shows the invariant mass of the lepton-pair-plus-jet system for data, background and its systematic uncertainties, and a signal hypothesis.

The analysis found that for signals of invariant mass of ≈ 130 GeV, meaning that the mass of the a particle is around $m_a = 0.5\text{--}2.5$ GeV, a bump in the data does not accommodate the existence of an a particle compatible with the 2HDM. The largest significance is 1.5σ , observed for $m_a = 0.5$ GeV in the $a \rightarrow gg$ scenario. Conversely, the other signals become compatible with the background-only hypothesis, as seen in Figure 4.3, which depicts the invariant mass after fitting to a signal of $m_a = 0.5$ GeV. The background shape and normalization come from the fit to the data. The signal includes only gg decays, its shape comes from the fit, and it is normalized to $\mathcal{B}(h \rightarrow Za) = 100\%$.



¹Pile-up interactions result in events that are recorded alongside the signal events, but originate from additional underlying pp interactions within the same bunch crossing.

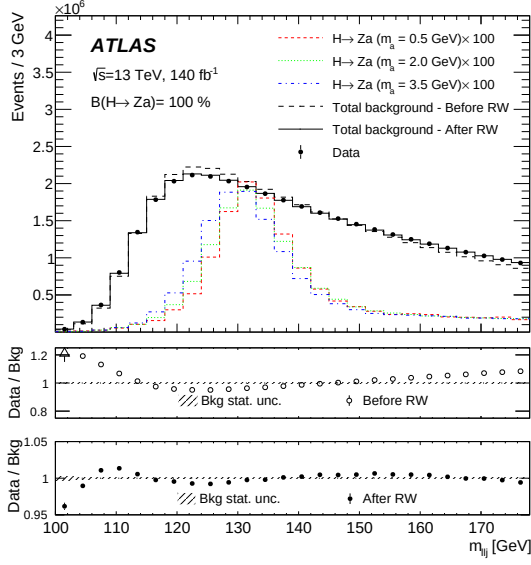


FIGURE 4.2: Invariant mass of the lepton pair plus jet system for data, before fit.

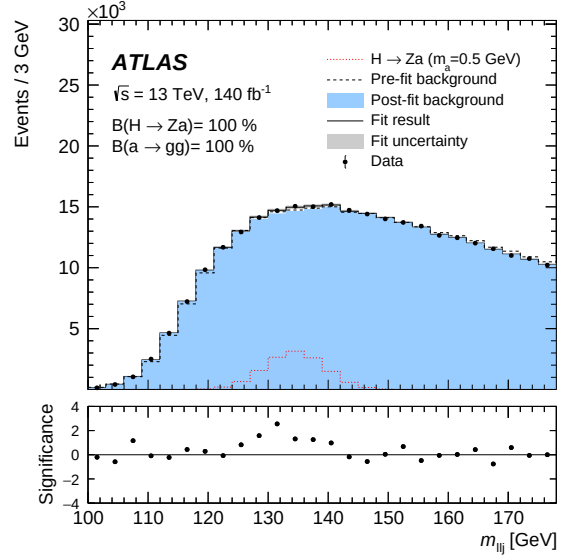


FIGURE 4.3: Invariant mass of the lepton pair plus jet system for data, after fit.

4.2 New Phenomenological Study

The ATLAS Run 2 search has targeted hadronically decaying BSM scalar particles compatible with the ones from the 2HDM using data from the ATLAS experiment at CERN, and it yielded no significant excess above the Standard Model background prediction. The objective of this Thesis is to develop an independent ML-based pipeline for this decay channel, characterize the theoretical uncertainties associated with the Monte Carlo generator choice, and assess the robustness of machine-learning discriminants under changes of simulation configuration. These questions have since been addressed in an updated ATLAS analysis using the full Run 2 dataset; the present work constitutes an independent and complementary study that reached consistent conclusions regarding the role of jet grooming and generator systematics. While the ATLAS search relies on a full detector simulation and a dedicated reconstruction chain, the present study employs a lighter framework — POWHEG for the hard process, interfaced with PYTHIA 8 and HERWIG 7 for parton shower and hadronization, **FastJet** for jet reconstruction directly on final-state particles, and DELPHES for detector-effect simulation. Before any meaningful comparison between the two generators can be made, it is essential to verify that this pipeline reproduces the expected kinematic behavior of the signal process. The validation presented in this Section is therefore not a repetition of the ATLAS result, but rather a sanity check on our own tools: we confirm that both PYTHIA and HERWIG, when fed the same hard-process input, yield consistent distributions for the reconstructed leptons, jets, and the full $m_{\ell\ell j}$ invariant mass. Only once this agreement is established does the generator

comparison become interpretable — any differences observed there can then be attributed to genuine physical modeling choices rather than artifacts of the simulation chain.

4.3 Kinematic Validation

A first step in this study is to validate the samples and the generators we are using, since the generator comparison of jet substructure variables in the following Section is only meaningful if both generators produce kinematically consistent events. In this Section PYTHIA and HERWIG samples will be compared.

The signal $H \rightarrow Za \rightarrow \ell\ell + q\bar{q}/g\bar{g}$ is simulated via both Monte Carlo generators and the samples are analyzed using kinematical variables: p_T , transverse momentum; η , pseudo-rapidity; and ϕ , azimuthal angle. All plots are normalized to unit area.

The reconstructed objects we are analyzing are called `lep1`, `lep2`², and `jet`, identifying the three objects that we could observe, and they are shown in Figure 4.5. The distributions are mutually consistent between generators, confirming that both pipelines reproduce the expected kinematics of the signal process. The ratio panels do not fluctuate wildly; the large statistical fluctuations visible at high p_T are due to the limited number of events populating the tails, and can be reduced by increasing the number of simulated events. The validation step therefore confirms that the pipeline is working correctly and serves as a sanity check before proceeding. Another useful cross-check is the $m_{\ell\ell j}$ plot, shown in Figure 4.4. One expects to see a prominent peak near 125 GeV, corresponding to the Higgs boson mass, since $m_{\ell\ell j}$ reconstructs the full three-body system. The Z mass manifests separately in the dilepton invariant mass $m_{\ell\ell}$, not in this distribution. The a peak is not separately visible in $m_{\ell\ell j}$, since the three-body invariant mass resolves the full system rather than the individual decay products, and the low mass of a causes its contribution to merge into the Z peak region. This plot shows good accord between generators and validates the previous results.

²The leptons are ordered by p_T , where `lep1` has a larger momentum than `lep2`.

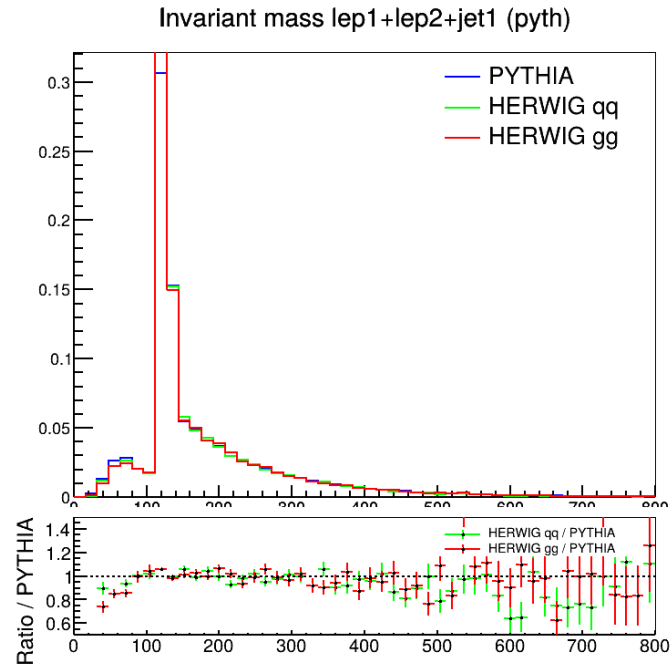


FIGURE 4.4: The invariant mass $m_{\ell\ell j}$ shows good accord between generators up to statistical fluctuations.



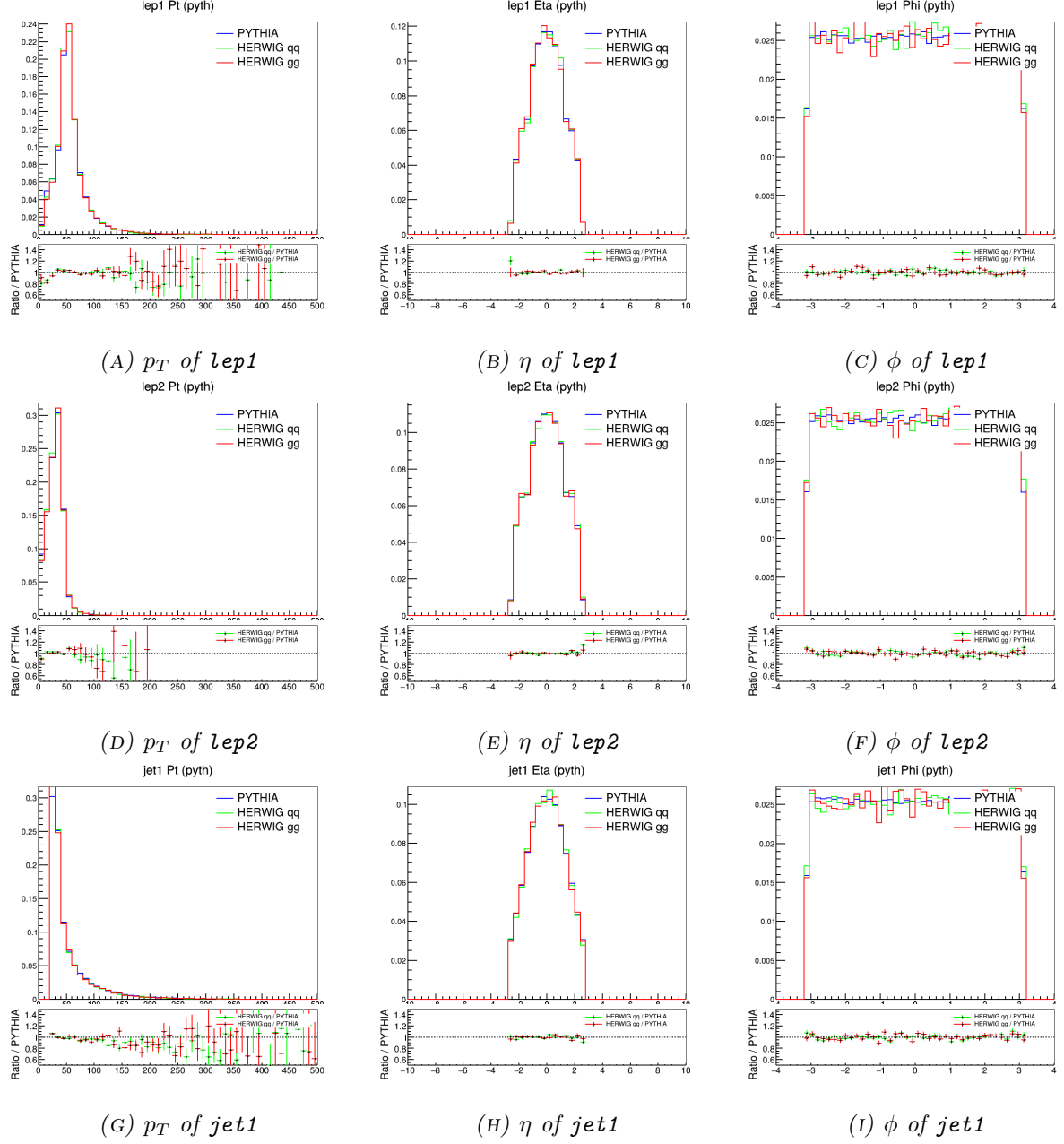


FIGURE 4.5: All variables show good accord between generators up to statistical fluctuations. The HERWIG signal is split into the two decay channels, but the plots are normalized.

4.4 Comparison of Jet Substructure Variables

This Section holds the results of a comparison between the two generators we are considering, PYTHIA and HERWIG. The comparison is motivated by the fact that the two generators implement fundamentally different models for the nonperturbative stages of the simulation. PYTHIA employs the Lund string model for hadronization and a p_T -ordered dipole shower, while HERWIG uses a cluster fragmentation model and an angular-ordered parton shower. These differences are expected to leave distinct imprints on jet substructure observables, which are sensitive precisely to the soft and collinear structure of the final state. Any systematic difference observed between the two generators in the following plots can therefore be attributed to genuine physical modeling choices rather than numerical artifacts, and serves as an estimate of the theoretical uncertainty associated with the choice of Monte Carlo generator. We shall first look at the variables of interest in the full simulation with hadronization enabled. After that, we look at how hadronization impacts the results by comparing the HAD-on and HAD-off outputs for both generators. Lastly, we examine how the variables are affected by applying a grooming algorithm to suppress soft radiation that might disturb signal identification. The simulation is run for $m_a = 3.00$ GeV and $N_{\text{events}} = 10^6$. These choices are made in order to satisfy constraints imposed by HERWIG³ and to achieve sufficient suppression of statistical fluctuations.

4.4.1 Workflow

A few words about the workflow are important in order to understand the following material. POWHEG-BOX v2 is used to produce Higgs bosons from pp collisions at $\sqrt{s} = 8$ TeV⁴, and an input LHEF file is produced. PYTHIA then takes this as input and handles the decays of the BSM Higgs and a particle, producing another LHE file containing the hard process $H \rightarrow Za \rightarrow \ell\ell + \bar{q}q/\bar{g}g$. This is due to the fact that in HERWIG it is rather difficult to define a BSM particle that is not pre-existing in the runcards. PYTHIA is instead able to add the BSM particle easily and calculate its branching ratios from first principles. The result is an LHE file containing the Higgs boson production and subsequent decay, ready for the hadronization step. The LHE file is fed to both PYTHIA and HERWIG in order to produce showers and hadronization, introducing generator-specific effects. The result of this process is two HepMC⁵ files, which need to be analyzed and processed further.

³For masses lower than the one considered here, HERWIG appears to become numerically unstable and unable to produce valid events for such a light particle. Further investigation of this limitation is needed.

⁴The simulation runs at $\sqrt{s} = 8$ TeV (4 TeV per beam), compared with the 13 TeV center-of-mass energy of Run 2. This difference affects the absolute production cross-section and the p_T spectrum of the Higgs, but since the generator comparison is performed at fixed $m_a = 3.00$ GeV and the jet substructure observables are evaluated after the standard kinematic selection, the kinematic differences are largely absorbed into the event selection. The qualitative conclusions of the generator comparison are not expected to be significantly affected; a dedicated simulation at $\sqrt{s} = 13$ TeV would be required to quantify any residual effect.

⁵HepMC is a standard C++ event record library and file format for storing fully hadronized Monte Carlo events, including the complete particle history, decay tree, and vertex information.

An `hepmc2root` script is used to convert the PYTHIA/HERWIG output into a ROOT file, and jet substructure variables are calculated using **FastJet**⁶ ([44]), a C++ library for jet reconstruction (anti- k_t in our case) applied to final-state particles. This code is also able to apply grooming to reconstructed objects using **SoftDrop**⁷ ([69]). The selection criteria applied to jets are summarized in Table 4.2. From the ROOT files produced it is possible to build the histograms shown in the following. A summary of the workflow is shown in Figure 4.6.

⁶**FastJet** is a widely used C++ package for jet finding and jet substructure analysis, providing efficient implementations of sequential clustering algorithms such as anti- k_t , k_t , and Cambridge/Aachen.

⁷**SoftDrop** is a jet grooming procedure that recursively declusters a jet and removes soft wide-angle branches failing the condition $z > z_{\text{cut}} (\Delta R_{12}/R)^\beta$, reducing sensitivity to non-perturbative effects and pile-up.

Category	Variable	Value
Leptons	Subleading lepton p_T	18.0 GeV
	Leading lepton p_T	27.0 GeV
	Maximum electron $ \eta $	2.47
	Maximum muon $ \eta $	2.70
Z boson	$m_{\ell\ell}$ window	[81.0, 101.0] GeV
	Maximum $m_{\ell\ell j}$	250.0 GeV
Jets	Minimum jet p_T	20.0 GeV
	Maximum jet $ \eta $	2.5
	Jet radius R	0.4
Tracks	Minimum track p_T	0.5 GeV
	Maximum track $ \eta $	2.5
Grooming	z-cut z_{cut}	0.1
	β	1

TABLE 4.2: Summary of event and jet selection criteria used in the analysis.

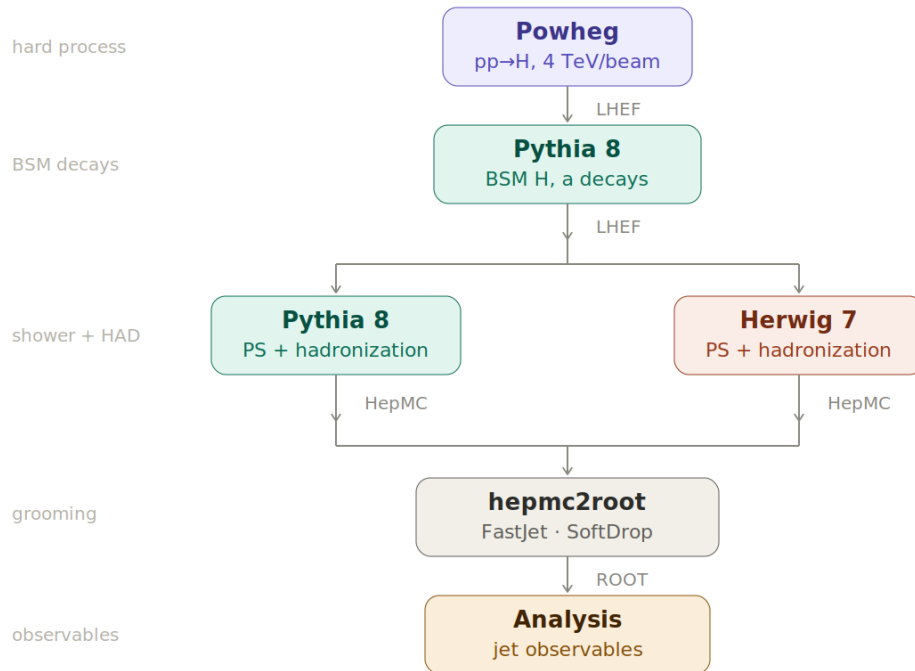


FIGURE 4.6: Workflow for the event generation in PYTHIA/HERWIG.

4.4.2 Jet Substructure Variables Comparison: Ungroomed Case

We are now ready to present and compare the results coming from the full simulation of hadronization effects from PYTHIA and HERWIG. In these plots, an extra line is added to represent an approximation — it corresponds to the hard process $H \rightarrow Za \rightarrow \ell\ell + \bar{q}q/\bar{g}g$, which is not technically the LO process, but serves as a useful guideline.

to the tree-level process, without NLO corrections. The ratio panels between the generators are shown in order to keep fluctuations in check.

We start by looking at the kinematical variables for the different elements of the decay: we will consider two leptons and a jet, as shown in Figure 4.7. Jet-related variables seem to be consistent between generators, as indicated by plots 4.7a, 4.7b, and 4.7c. HERWIG seems to have better modeling for highly energetic regimes, while it substantially agrees with PYTHIA for lower energy spectra. Invariant masses are well reproduced, as shown in plots 4.7d and 4.7e. The structure observed in the $m_{\ell\ell j}$ distribution around ~ 150 GeV is not necessarily associated with a true resonance, but can instead arise from the incomplete reconstruction of the $H \rightarrow Za \rightarrow \ell\ell + jj$ final state in the boosted regime. In particular, the large boost of the a boson causes the two jets from the $a \rightarrow jj$ decay to become highly collimated, such that only part of the hadronic system may be captured by the jet reconstruction algorithm. As a consequence, the observable $m_{\ell\ell j}$ does not correspond exactly to the invariant mass of the parent Higgs boson, but rather to a partially reconstructed system whose shape is strongly affected by jet clustering, out-of-cone radiation, and parton showering effects. From the analysis of these kinematical variables one concludes that PYTHIA has a tendency to overestimate the high-energy behavior of the objects we are considering, with the PYTHIA/HERWIG ratio in the p_T tails reaching values of order 1.2–1.3 as visible in the ratio panels of Figure 4.7.

We can now take a look at the jet substructure variables depicted in Figure 4.8. Since the a particle is produced with high boost, its decay products are expected to be collimated within the jet cone, resulting in distributions peaked at low values for angularity, U_1 , and M_2 , and a comparatively low track multiplicity with respect to a QCD jet. These plots are the product of the **FastJet** reconstruction algorithm, and the variables shown have been presented and detailed in Chapter 3. The first two plots, 4.8a and 4.8b, represent the number of reconstructed charged tracks and the ΔR angular separation between the leading track and the jet axis respectively. Even in this case PYTHIA tends to overestimate the number of events satisfying the conditions for the variables. Since the hard event (black dashed line) does not involve any hadronization effect, it makes sense that it would show many fewer counts. The same behavior is observed in the following plots: 4.8c, 4.8d, 4.8e. The angularity and U_1 show approximately the same behavior: there is a stronger peak near zero, consistent with the highly-boosted topology. This behavior is a direct consequence of the boost: for a highly boosted two-body decay $a \rightarrow \bar{q}q/\bar{g}g$, the decay products are strongly collimated within the jet cone, so the bulk of the jet momentum is carried by two narrow cores and the soft transverse activity measured by angularity and U_1 is kinematically suppressed. We see the same behavior in M_2 , consistent with a highly boosted object. The N-subjettiness variable τ_2 quantifies the residual radiation relative to a two-subjet hypothesis: smaller values indicate better compatibility with a two-pronged

structure. The bulk of the distribution lying at low τ_2 is therefore consistent with the expected topology of $a \rightarrow \bar{q}q/\bar{g}g$, where the jet genuinely contains two hard cores.

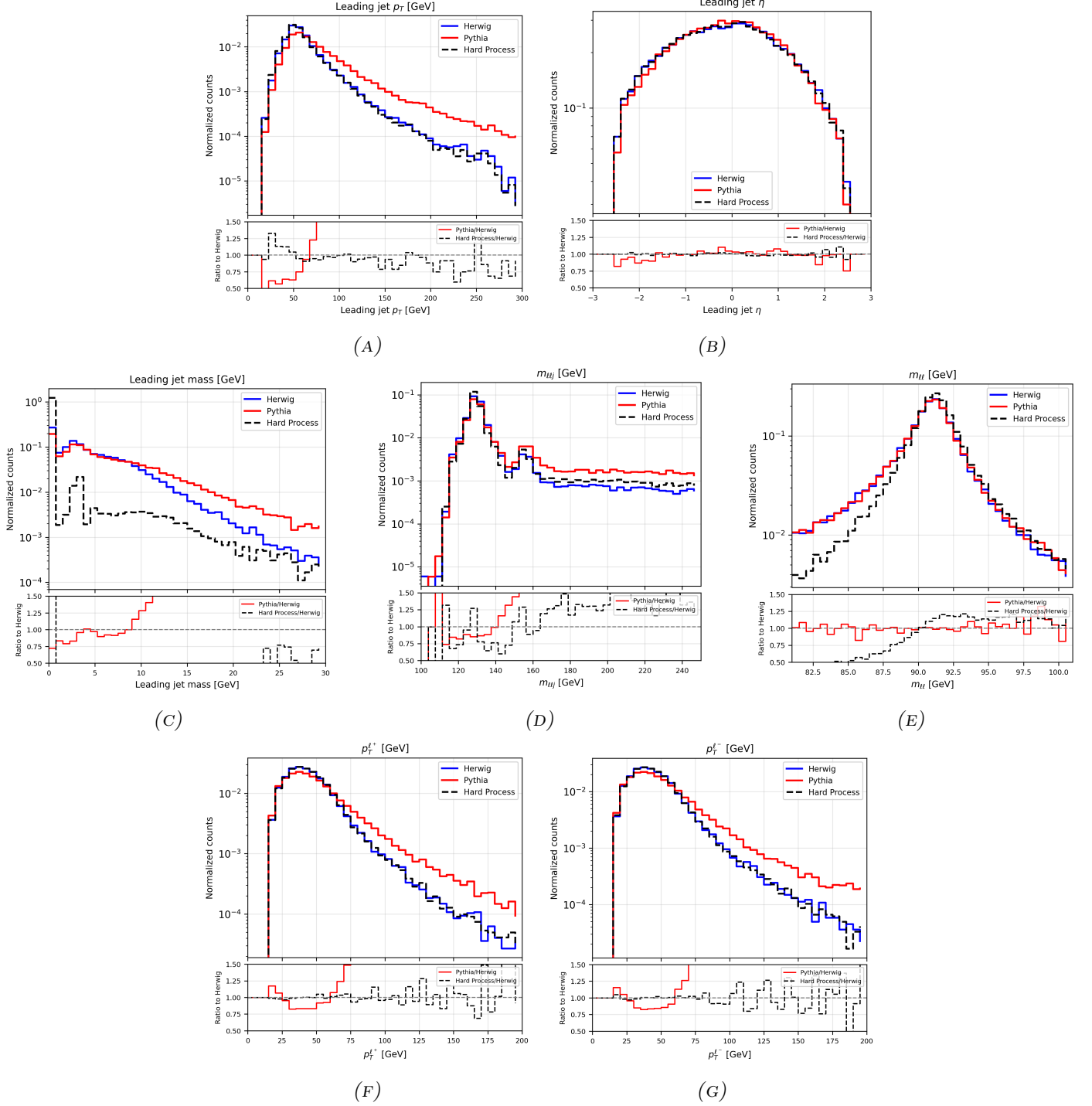


FIGURE 4.7: Kinematical variables from PYTHIA and HERWIG full simulation.

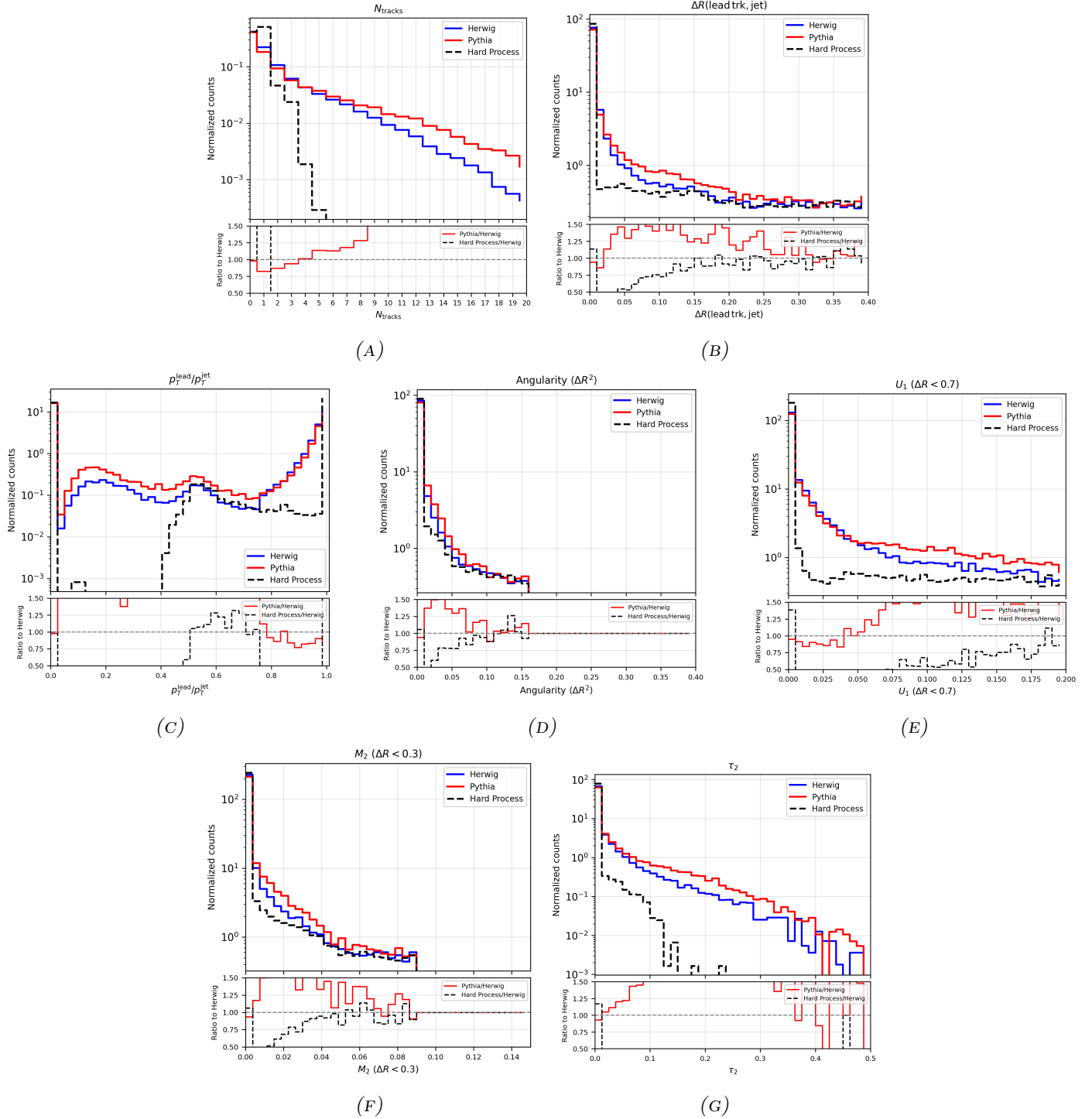


FIGURE 4.8: Jet substructure variables reconstructed from PYTHIA and HERWIG full simulation.

4.4.3 Effect of Hadronization: HAD-on vs HAD-off

The following plots investigate the impact of hadronization on both kinematical and jet substructure variables by comparing, for each generator, the fully hadronized (HAD-on) run against the parton-level (HAD-off) run in which hadronization is disabled. In all figures, solid lines denote HAD-on and dashed lines denote HAD-off, with PYTHIA shown in red and HERWIG in blue. The aim of this comparison is to isolate the nonperturbative contribution to each observable by contrasting, for each generator separately, the fully hadronized output against the parton-level output, thereby providing a direct estimate of the systematic uncertainty associated with the choice of hadronization model.

The kinematical variables are shown in Figure 4.9 as a sanity check. The dilepton invariant mass $m_{\ell\ell}$ (4.9e) and the full three-body invariant mass $m_{\ell\ell j}$ (4.9d) likewise show very little sensitivity to hadronization: the peak position is stable, and only minor shape differences appear in the tails, attributable to statistics. The jet η distribution (4.9b) is also stable, confirming that the angular geometry of the reconstructed jet axis is not altered significantly by the nonperturbative step. The most visible kinematical effect of hadronization appears in the jet p_T (4.9a) and, more markedly, in the jet mass (4.9c): turning hadronization on systematically shifts the jet mass distribution towards higher values, broadening the distribution and depleting the near-zero region. This is a direct consequence of the soft particles added during the hadronization step, which increase the invariant mass of the jet constituents. The effect is qualitatively the same for both generators, though PYTHIA produces a somewhat larger shift than HERWIG, consistent with the broader soft-particle activity already observed in the full simulation comparison. The effect of hadronization is considerably more pronounced for the jet substructure variables, as shown in Figure 4.10. The track multiplicity N_{tracks} (4.10a) provides the most striking illustration: at parton level, the jet contains only a small number of hard perturbative partons, and the distribution peaks at very low multiplicities. Once hadronization is enabled, the distribution is dramatically broadened and its peak shifts to higher values, as each parton fragments into multiple color-neutral hadrons. PYTHIA consistently produces more tracks than HERWIG in the HAD-on case, reflecting the difference between the Lund string and cluster fragmentation models. The angular separation ΔR between the leading track and the jet axis (4.10b) is similarly sensitive: hadronization populates the outer region of the jet cone more densely, broadening the distribution relative to the parton-level result. The p_T -ratio variable (4.10c) shows a redistribution of momentum fraction between the HAD-on and HAD-off cases: at parton level, the leading constituent tends to carry a larger fraction of the total jet p_T , while hadronization smears this fraction downward by redistributing energy among more particles.

The energy-correlation observables angularity (4.10d), U_1 (4.10e), and M_2 (4.10f) all show the same qualitative behavior: the HAD-off distributions are more sharply peaked at low values, consistent with the highly collimated two-parton topology of the hard process, while HAD-on populates higher values due to the soft, wide-angle radiation introduced by fragmentation. The effect is larger in PYTHIA than in HERWIG for all three variables, attributable to the greater soft-particle activity of the Lund string model. The

N-subjettiness ratio τ_2 (4.10g) likewise shifts to higher values when hadronization is enabled: at parton level the jet is dominated by two clean hard cores and τ_2 is suppressed, while hadronization adds soft radiation between and around the cores, increasing the residual two-subjet incompatibility.

Overall, the **HAD-on/HAD-off** comparison confirms that the leptonic sector and gross kinematic quantities are robust against the choice of hadronization model, while jet substructure observables are sensitive nonperturbative probes. The magnitude of the hadronization correction and the size of the PYTHIA–HERWIG spread are consistent across all variables, and the differences observed between generators in this comparison can be taken as a direct estimate of the nonperturbative modeling uncertainty.

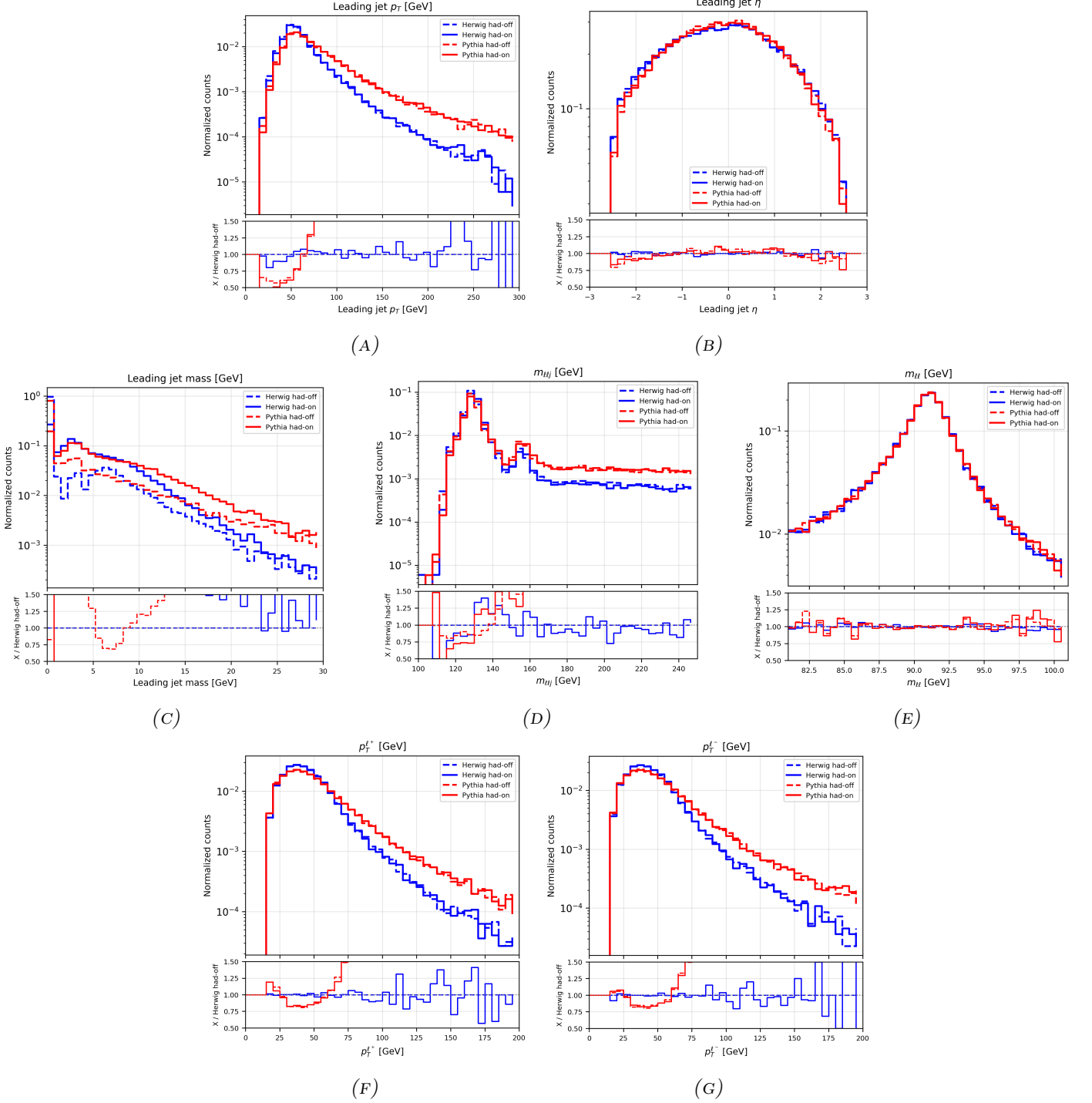


FIGURE 4.9: Kinematical variables from PYTHIA and HERWIG full simulation, with dashed lines for HAD-off and solid lines for HAD-on.

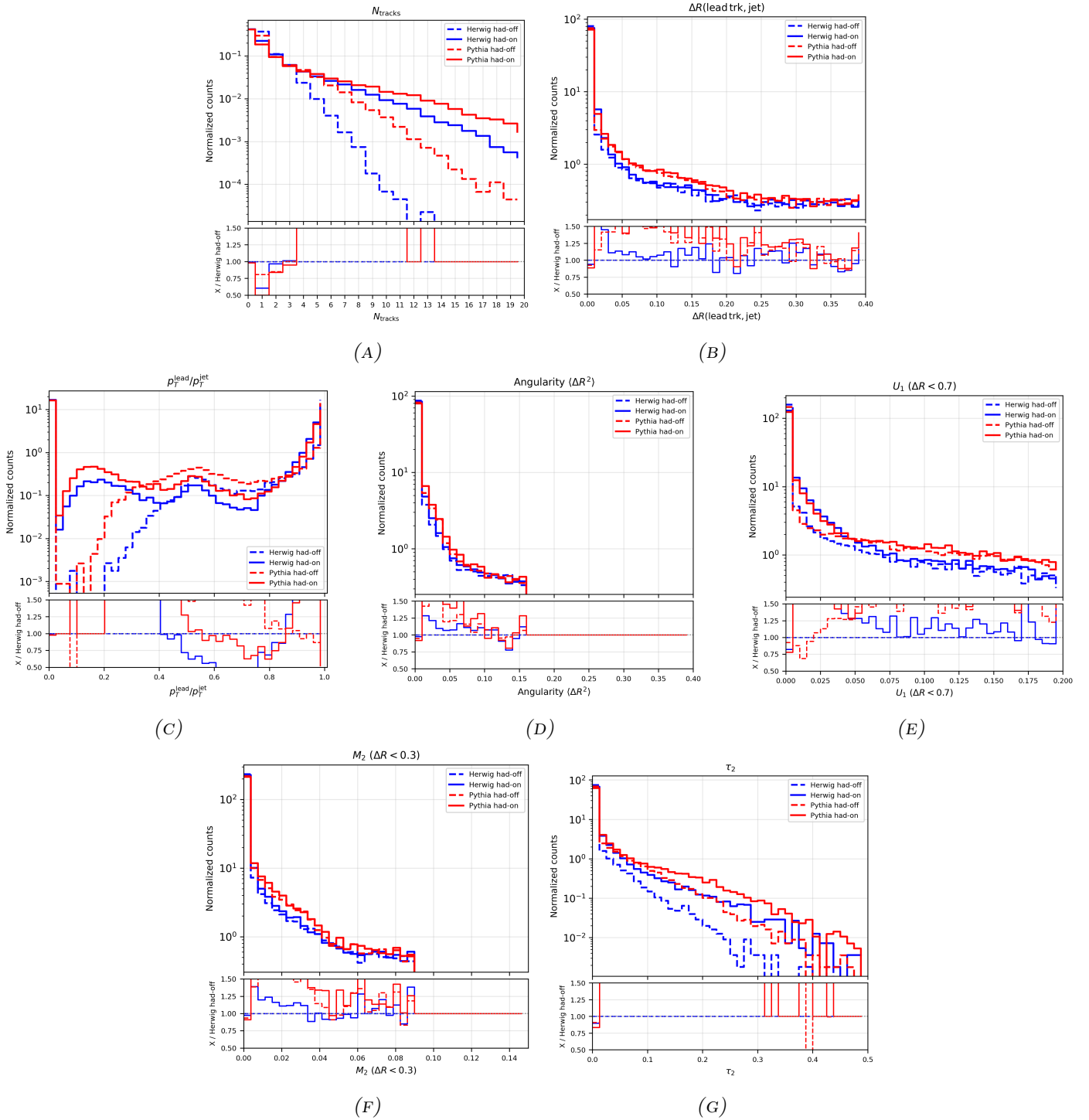


FIGURE 4.10: Jet substructure variables reconstructed from PYTHIA and HERWIG full simulation, with dashed lines for HAD-off and solid lines for HAD-on.

4.4.4 Groomed Variables

The **HAD-on/HAD-off** comparison presented in the previous Section has made explicit what was anticipated on theoretical grounds: the jet substructure observables carry a substantial nonperturbative component, and the size of this component differs meaningfully between PYTHIA and HERWIG. This generator dependence is not merely a theoretical inconvenience — in a real analysis it directly inflates the systematic uncertainty on the signal model, potentially limiting the sensitivity of the search. Jet grooming offers a practical handle on this problem. By recursively declustering the jet and discarding soft, wide-angle branches that fail the **SoftDrop** condition $z > z_{\text{cut}} (\Delta R_{12}/R)^\beta$, grooming suppresses precisely the nonperturbative radiation that drives the inter-generator differences, while leaving the hard two-pronged core of the $a \rightarrow \bar{q}q/\bar{g}g$ decay intact. The groomed observables are therefore expected to be more infrared-safe and less sensitive to hadronization model assumptions — all properties that make them preferable as inputs to a signal extraction procedure. In the following, we apply **SoftDrop** with parameters $z_{\text{cut}} = 0.1$ and $\beta = 1$, and compare the resulting distributions between PYTHIA and HERWIG to assess whether grooming succeeds in reducing the theoretical uncertainty estimated in the previous Sections.

The groomed jet kinematic and substructure distributions are shown in Figures 4.11 and 4.12 respectively as a sanity check. For the kinematic variables the effect of grooming is mild, as expected: the leading-jet p_T and η distributions are essentially unchanged with respect to the ungroomed case, and the ratio panels remain flat at unity to within statistical fluctuations. The jet mass, however, is visibly affected: the groomed distribution is shifted towards lower values relative to the ungroomed **HAD-on** result, and the **HAD-on/HAD-off** ratio panel is flatter, indicating that the soft-particle contribution responsible for inflating the jet mass in the ungroomed case has been successfully suppressed by the **SoftDrop** procedure.

Turning to the substructure variables, the observables — angularity $\langle \Delta R^2 \rangle$, U_1 , and M_2 — show the most pronounced response to grooming: both the PYTHIA and HERWIG distributions are sharpened and pulled back towards lower values, with the **HAD-on** and **HAD-off** curves converging more closely than in the ungroomed case. The **HAD-on/HAD-off** ratio panels for these variables remain within $\pm 15\%$ of unity across most of the distribution, and the residual shape differences are confined largely to the tails. The N_{tracks} distribution, by contrast, retains a **HAD-on/HAD-off** separation even after grooming: the **HAD-on** peak shifts to multiplicities of order 5–8, while the **HAD-off** distribution remains concentrated at $N_{\text{tracks}} \lesssim 4$, reflecting the fact that grooming removes soft wide-angle branches from the clustering tree but does not suppress the fragmentation of the hard constituents within the surviving groomed core. The ΔR and p_T -ratio variables show an intermediate behavior, with grooming reducing but not eliminating the generator and hadronization sensitivity. The groomed τ_2 distribution is more sharply peaked at low values than its ungroomed counterpart for both generators, consistent with the two hard cores of the $a \rightarrow \bar{q}q/\bar{g}g$ decay becoming more resolved once soft radiation is removed; the **HAD-on/HAD-off** ratio remains within $\pm 20\%$ across the bulk of the distribution.

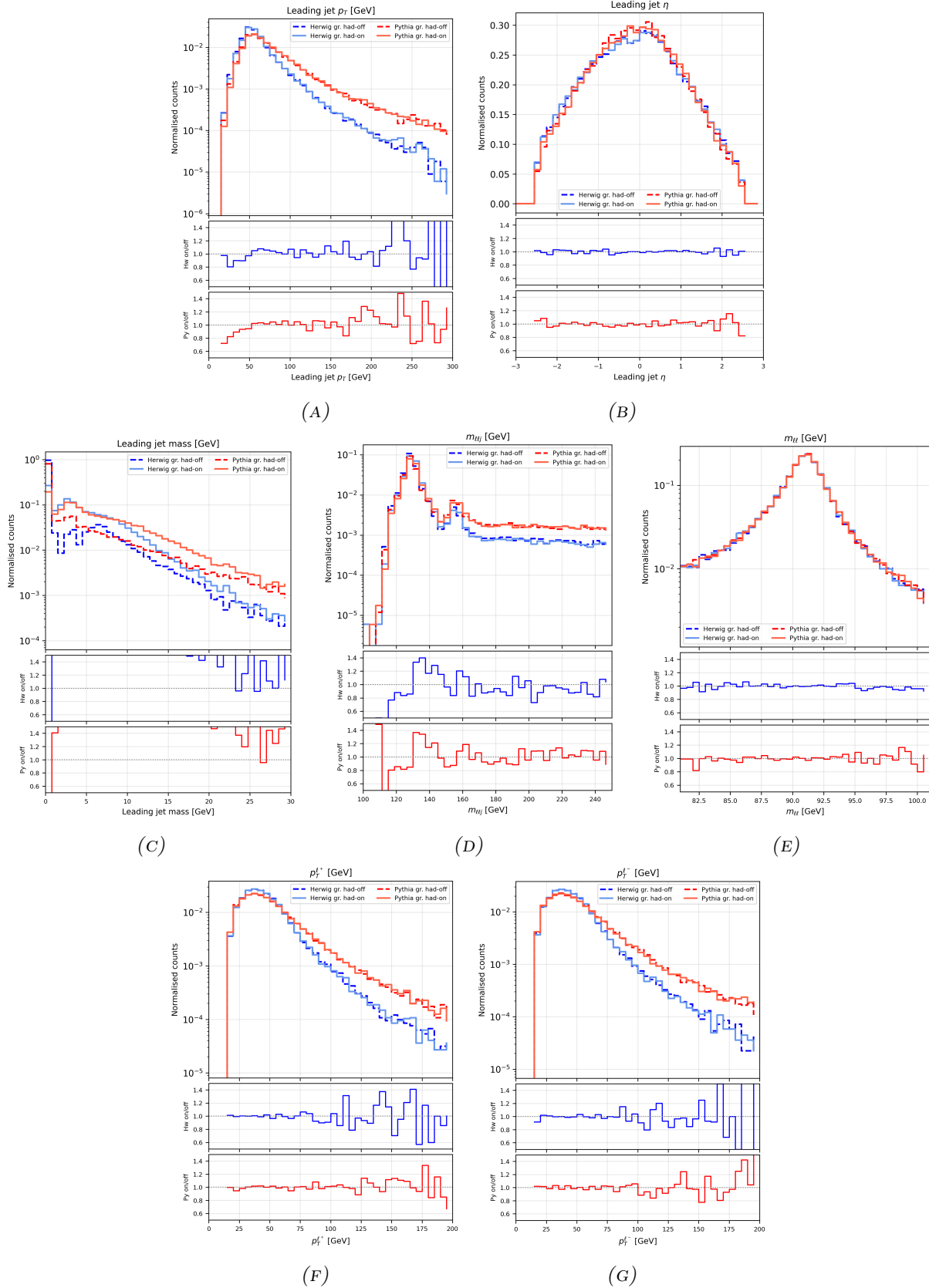


FIGURE 4.11: Groomed kinematical variables from PYTHIA and HERWIG, with dashed lines for HAD-off and solid lines for HAD-on.

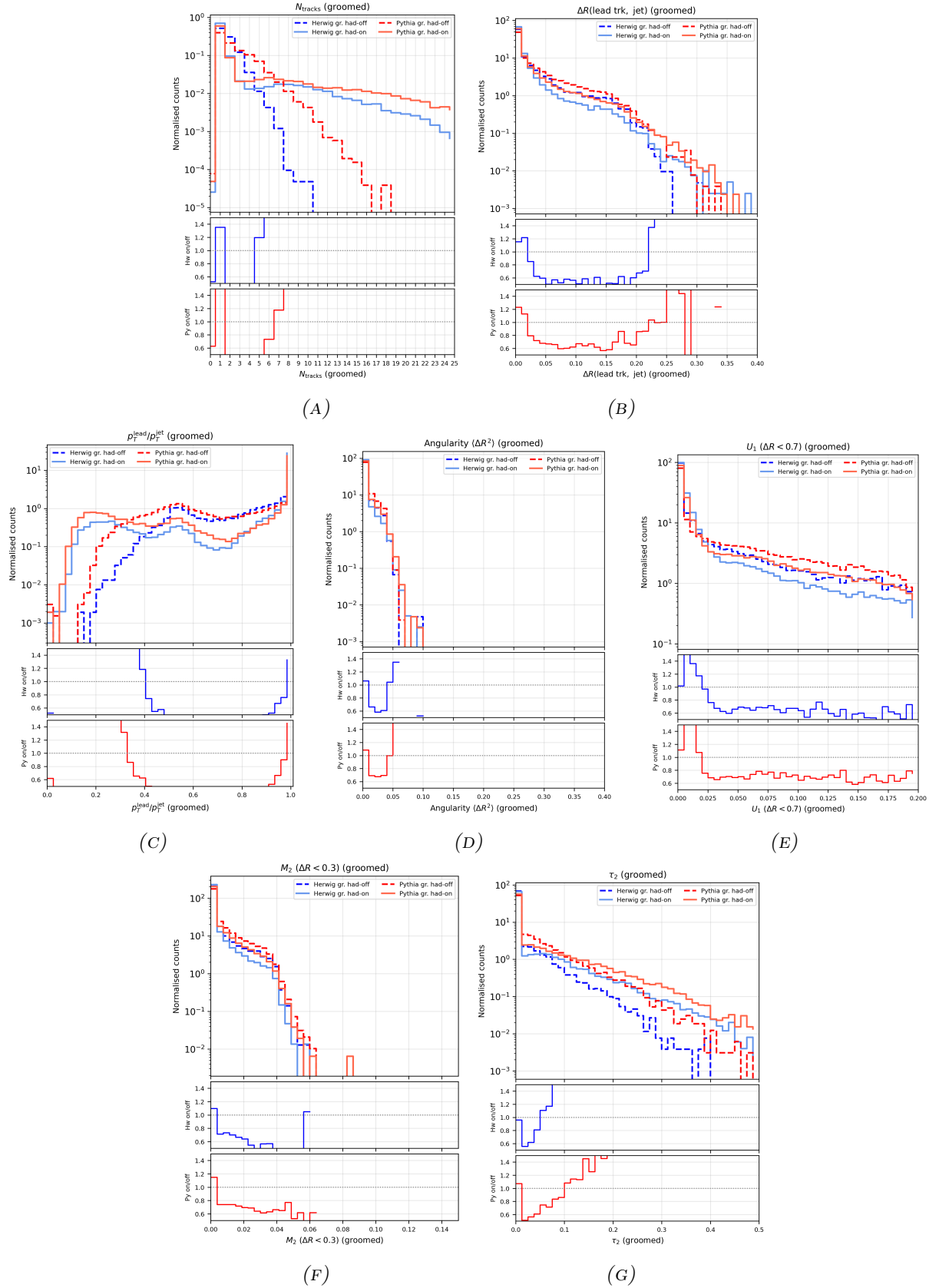


FIGURE 4.12: Groomed jet substructure variables reconstructed from PYTHIA and HERWIG, with dashed lines for *HAD-off* and solid lines for *HAD-on*.

4.4.5 Comparison between Ungroomed and Groomed Variables

Comparing the groomed results of Figures 4.11 and 4.12 directly against their ungroomed counterparts in Figures 4.9 and 4.10 reveals a clear but variable improvement. For the energy-correlation observables the improvement is most evident: in the ungroomed case the **HAD-on**/**HAD-off** ratio panels for angularity, U_1 , and M_2 show systematic deviations from unity of order 20–40% across the bulk of the distribution, with both generators exhibiting a consistent upward shift of the **HAD-on** curve relative to **HAD-off**. After grooming these deviations are reduced to the 10–20% level, and the two generators respond to hadronization in a more similar way, suggesting that the dominant source of nonperturbative uncertainty in these observables is indeed the soft wide-angle radiation that **SoftDrop** is designed to remove. A complementary reduction is observed for the jet mass, where the **HAD-on**/**HAD-off** shift of roughly 2–4 GeV visible in the ungroomed distribution is substantially suppressed after grooming, with the ratio panels converging to within $\pm 10\%$ of unity over most of the spectrum. The improvement for τ_2 is more modest: the ungroomed ratio panels deviate by up to $\sim 30\%$ from unity, while the groomed ratio panels reduce this to roughly $\pm 20\%$, indicating that a residual nonperturbative component survives inside the groomed jet core. For N_{tracks} and ΔR , however, the comparison yields a sobering result: the **HAD-on**/**HAD-off** separation in the groomed distributions is comparable in magnitude to the ungroomed case, with ratio panels that depart from unity by factors of order 1.3–1.5 at the extremes of both distributions. This confirms that grooming does not reduce the hadronization sensitivity of multiplicity-based observables, since those observables probe the fragmentation structure of the hard core rather than the soft peripheral radiation that **SoftDrop** removes. Taken together, the comparison suggests that grooming is most beneficial as a pre-processing step for the energy-correlation family of variables, and that any machine-learning pipeline relying on N_{tracks} as an input should treat its associated nonperturbative uncertainty as a significant and largely unmitigated systematic.

Section Roadmap:

Covered: Kinematic validation of the independent simulation pipeline, confirming that PYTHIA 8 and HERWIG 7 reproduce consistent distributions for reconstructed leptons, jets, and the three-body invariant mass $m_{\ell\ell j}$ when seeded from the same POWHEG hard-process input; generator comparison of jet substructure observables, organized into three parts: the full simulation with hadronization enabled, the **HAD-on** versus **HAD-off** study isolating the nonperturbative contribution to each observable, and the **SoftDrop**-groomed variables assessing the resilience of the observables to soft wide-angle radiation; quantification of the inter-generator spread as a direct estimate of the nonperturbative modeling uncertainty; direct comparison between ungroomed and groomed distributions, establishing that grooming reduces the hadronization sensitivity of the energy-correlation family — angularity, U_1 , and M_2 — from the 20–40% level to the 10–20% level, while the track-based observables N_{tracks} and ΔR retain a

large nonperturbative uncertainty that `SoftDrop` cannot mitigate.

References: [67]; [62]; [66]; [44]; [69].



4.5 Neural Network Performance

The preceding generator comparison has established that the jet substructure observables carry a significant dependence on the choice of Monte Carlo generator and on the treatment of nonperturbative effects. This Section assesses whether a machine-learning approach can exploit these same observables to improve the sensitivity of the $H \rightarrow Za$ search, and investigates how performance changes when the trained model is confronted with samples that differ from its training distribution. The study is organized into three phases. Study A trains and evaluates the full pipeline using DELPHES detector-level samples for training and truth-level samples for evaluation. Study B corrects the three sources of inconsistency identified in Study A and carries out a fully self-consistent evaluation within the same simulation framework. Study C repeats the Study B pipeline with a modified regression training procedure designed to address the near-mean-collapse identified in Study B, and verifies that the central conclusion regarding grooming and generator robustness is stable under this change. The three studies are presented in Sections 4.5.2, 4.5.3, and 4.5.4 respectively.

4.5.1 The Two-Stage Pipeline

The neural network pipeline consists of two stages operating in sequence, illustrated schematically in Figure 4.13. The first stage is a regression network that maps the seven jet substructure variables directly to a predicted pseudoscalar mass $\hat{m}_a$. The second stage is a classification network that receives the same seven input variables augmented by the regression output $\hat{m}_a$ as an eighth feature, and returns a continuous score separating signal jets from background. The motivation for this augmentation is that a correct mass prediction is informative for signal identification: a jet whose regression output clusters near a specific mass value is more likely to originate from a pseudoscalar decay of that mass, whereas background jets are less likely to produce a consistent or well-clustered mass prediction. By giving the classifier access to this mass estimate, the network can use both the low-level ghost-track information and the reconstructed mass simultaneously to improve signal identification.

Both networks are multi-layer perceptrons (MLPs) with three hidden layers of widths 128, 64, and 32 neurons respectively, trained with the Adam optimizer, batch normalization after each hidden layer, and early stopping on the validation loss.

A multi-layer perceptron is the simplest realisation of a deep neural network: the input features are passed through a sequence of layers, each of which multiplies the incoming values by a matrix of learnable weights, adds a bias, and then applies a nonlinear function — typically a rectified linear unit, $\text{ReLU}(x) = \max(0, x)$ — to the result. The nonlinearity is essential: without it, stacking multiple layers would be equivalent to a single linear transformation, and the network would be unable to capture the complex correlations present in jet substructure data. With three hidden layers of widths 128, 64, and 32 neurons, the architecture progressively compresses the input into a lower-dimensional

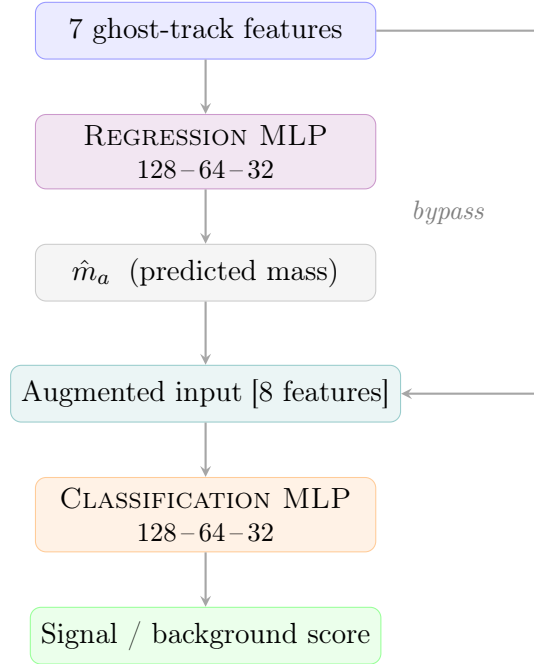


FIGURE 4.13: Schematic of the two-stage neural network pipeline. The seven ghost-track substructure variables are first passed to the regression network, which returns a predicted pseudoscalar mass $\hat{m}_a$. This prediction is appended to the original feature vector, and the augmented eight-feature input is passed to the classification network, which returns a continuous signal-versus-background score.

representation, retaining only the combinations of features most useful for the task at hand.

Training a neural network means finding the set of weights that minimizes a chosen loss function, measured on a labeled training sample. This is done iteratively: at each step, a small random subset of the training data — a mini-batch — is passed through the network, the loss is computed, and the weights are nudged in the direction that reduces it. The size of each nudge is controlled by the learning rate, and the Adam optimizer determines this automatically for each weight individually, by keeping track of how the gradient has been behaving over recent steps; weights whose gradients are large and consistent receive smaller updates, while weights whose gradients are noisy receive larger corrections, leading to faster and more stable convergence than a fixed learning rate would achieve.

Two additional techniques are used to make training more robust. Batch normalization rescales the output of each hidden layer so that, within every mini-batch, the values have zero mean and unit variance before being passed to the next layer; this prevents any single layer from producing activations so large or so small that the gradients propagating back through the network become numerically unstable. Early stopping addresses the risk of overfitting: once the loss measured on a held-out validation set stops improving for a fixed number of consecutive epochs, training is halted and the model weights are restored

to the checkpoint that achieved the best validation performance. This is particularly important in the present study, where the number of available signal jets is limited and the network could otherwise memorize the training sample rather than learning features that generalize to unseen data.

Regression training setup. The regression network was trained on DELPHES detector-level signal samples spanning nine pseudoscalar mass hypotheses: $m_a = 0.50, 0.75, 1.00, 1.50, 2.00, 2.50, 3.00, 3.50, 4.00$ GeV, together with the Z +jets background sample. Each signal file was produced by the same DELPHES fast detector simulation described in Chapter 3, ensuring that track impact parameters, momentum resolution, and reconstruction inefficiencies are consistently modeled. The regression target for each signal jet is the generator-level pseudoscalar mass m_a associated with its source sample; for background jets no meaningful mass target exists, and these entries are filtered out before computing the loss. The Huber loss function[54]

$$\mathcal{L}_\delta(\hat{m}, m) = \begin{cases} \frac{1}{2}(\hat{m} - m)^2 & |\hat{m} - m| \leq \delta, \\ \delta(|\hat{m} - m| - \frac{1}{2}\delta) & \text{otherwise,} \end{cases} \quad (4.1)$$

with $\delta = 1$ GeV was chosen in preference to mean-squared error because it is less sensitive to the outlier mass predictions that arise from jets with very few reconstructed tracks. Input features were normalized with a robust scaler (subtracting the median and dividing by the interquartile range) fitted on the training set only.

Classification training setup. The classification network was trained at $m_a = 3.00$ GeV on a balanced mixture of signal and background jets. It receives as input the same seven ghost-track variables used for regression, augmented by the scalar regression output $\hat{m}_a$ as an eighth feature. Both networks use the Adam optimizer with batch size 256 and early stopping with patience 25 on the validation loss.

4.5.2 Study A: Delphes Training and Truth-Level Evaluation

Study A trains and evaluates the full two-stage pipeline using DELPHES detector-level samples for training. Performance is first assessed within the training domain (in-domain), then the trained models are applied without retraining to truth-level configurations to assess how the pipeline degrades under a domain shift.

In-domain regression performance. Table 4.3 summarizes the regression performance on the DELPHES test set. The overall test MAE⁸ of 1.93 GeV, evaluated across all

⁸The Mean Absolute Error (MAE) quantifies the bin-by-bin agreement between two normalized distributions. Given two histograms f and g with N bins, it is defined as $\text{MAE} = \frac{1}{N} \sum_{i=1}^N |f_i - g_i|$. A value of 0 indicates perfect agreement; larger values reflect systematic shape differences.

signal jets irrespective of their true mass, reflects the intrinsic difficulty of regressing a continuous parameter from variables that are, individually, only weakly mass-discriminating. The more informative quantity is the signal accuracy: the fraction of signal jets for which the network prediction falls within 1 GeV of the true mass. This reaches 68% on the test set, rising above 88% for the most favorable mass points. The coefficient of determination R^2 ⁹ is slightly negative across all training runs, which is expected in the multi-mass setting: a negative R^2 does not indicate that the model performs worse than chance on any individual mass hypothesis, but rather that the overall variance of targets spanning 0.5 to 4 GeV exceeds the variance explained by the network. The signal accuracy is therefore adopted as the primary figure of merit for the regression stage.

Mass point	Signal accuracy (%)	MAE (GeV)
$m_a = 0.50$ GeV	59.5	1.08
$m_a = 1.00$ GeV	69.3	0.75
$m_a = 1.50$ GeV	88.3	0.48
$m_a = 2.00$ GeV	72.2	0.73
$m_a = 2.50$ GeV	64.5	0.93
$m_a = 3.00$ GeV	84.1	0.54
$m_a = 3.50$ GeV	88.9	0.52
$m_a = 4.00$ GeV	56.0	1.01
All mass points	68.3	1.93

TABLE 4.3: Regression performance on the DELPHES detector-level test set. Signal accuracy is the fraction of predictions within 1 GeV of the true m_a .

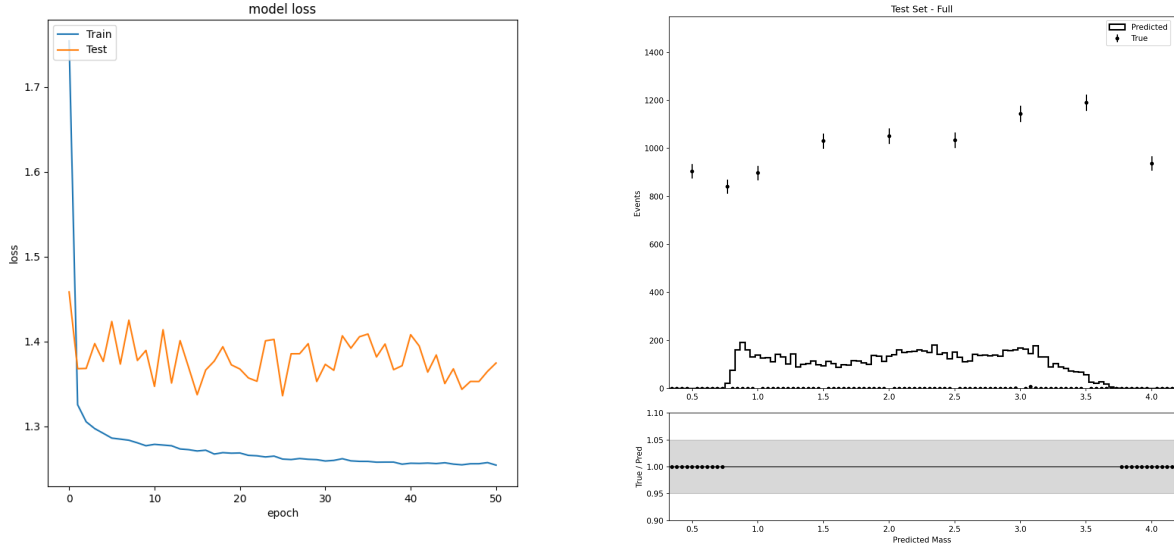
Figure 4.14 shows the training and validation loss curves and the regression output distribution on the DELPHES test set.

In-domain classification performance. The classification network achieves a test AUC¹⁰ of 0.831 and a validation AUC of 0.838. Figure 4.15 shows the corresponding receiver operating characteristic (ROC) curve, which plots the signal efficiency (true positive rate) against the background efficiency (false positive rate); a perfect classifier hugs the top-left corner, while a random classifier lies on the diagonal.

Domain-transfer study. A central question for the robustness of the analysis is how the pipeline degrades when applied to samples from a different simulation domain. To

⁹The coefficient of determination R^2 measures the fraction of variance in one distribution explained by another. Given histograms f and g with mean $\bar{f} = \frac{1}{N} \sum_i f_i$, it is defined as $R^2 = 1 - \frac{\sum_i (f_i - g_i)^2}{\sum_i (f_i - \bar{f})^2}$. A value of $R^2 = 1$ indicates perfect agreement; $R^2 = 0$ means the comparison histogram explains none of the variance of the reference.

¹⁰Area Under the ROC Curve: a threshold-independent summary of classifier performance. An AUC of 1 indicates perfect separation; an AUC of 0.5 corresponds to a random classifier.



(A) Training and validation loss curves. Early stopping triggers after 25 consecutive epochs without improvement; the best model is recovered at the minimum of the validation curve.

(B) Distribution of the regression network output on the DELPHES test set: predicted mass $\hat{m}_a$ for signal jets alongside the true mass distribution. The agreement confirms that the network has learned a reasonable approximation of the multi-mass target distribution.

FIGURE 4.14: Study A regression network: loss curves (left) and output distribution on the DELPHES test set (right).

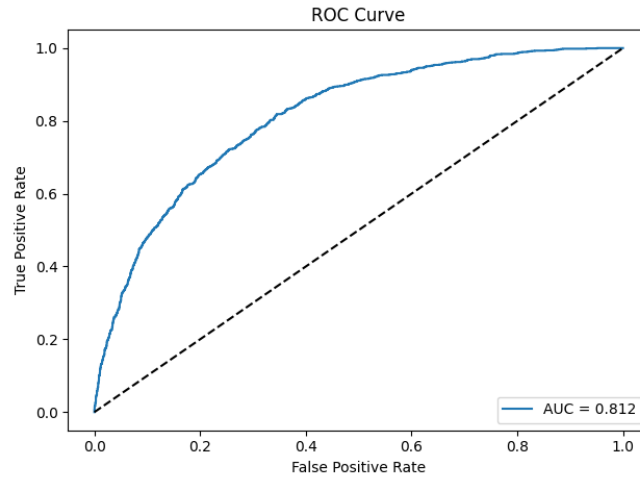
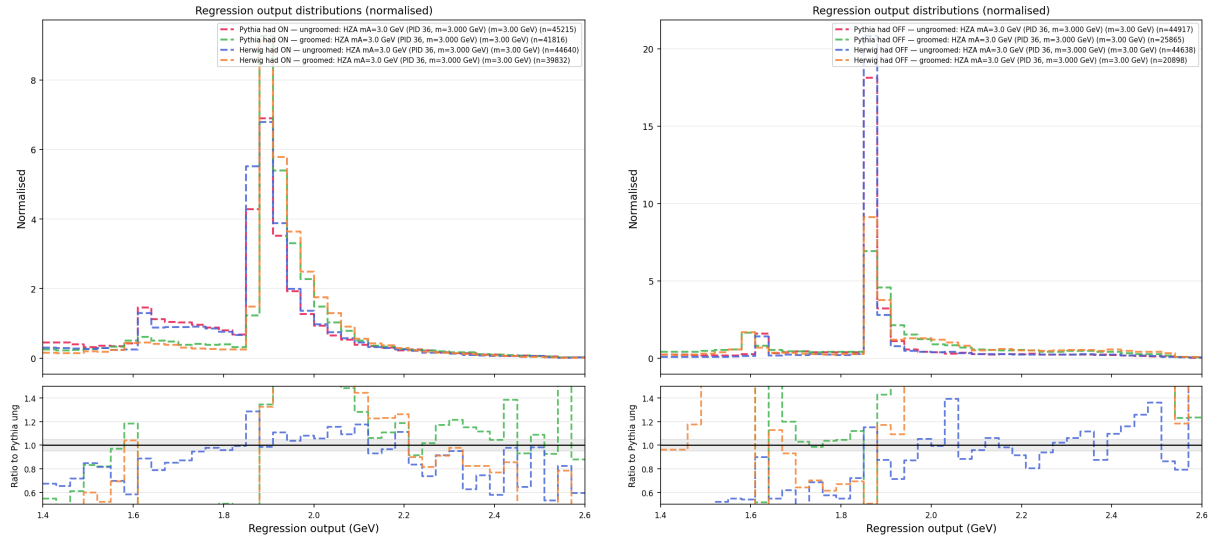


FIGURE 4.15: ROC curve for the Study A classification network evaluated on the DELPHES test set at $m_a = 3.00$ GeV. The AUC of 0.831 quantifies the overall separation power of the classifier.

assess this, the DELPHES-trained models were applied without retraining to four truth-level configurations: PYTHIA and HERWIG, each with hadronization enabled (**HAD-on**) and disabled (**HAD-off**). For each configuration, inference was performed twice — once using the ungroomed ghost-track variables and once using the **SoftDrop**-groomed variables. All truth-level configurations yield a prominent regression peak shifted to $\hat{m}_a \approx 2.1$ GeV, approximately 0.9 GeV below the true value of 3.0 GeV. This systematic offset is a direct consequence of the domain shift: the robust scaler was fitted on DELPHES feature distributions, which differ from the truth-level distributions in mean, width, and shape because of detector smearing, tracking inefficiency, and the different particle content of EFlow objects versus generator-level final-state particles. When truth-level features are passed through the DELPHES scaler, they are mapped to the wrong region of the input space, and the model responds with a systematically lower mass prediction. The resulting regression output distributions are shown in Figure 4.16.



(A) **HAD-on** configurations: PYTHIA and HERWIG, ungroomed and groomed. The ratio panel uses the ungroomed PYTHIA distribution as reference.

(B) **HAD-off** configurations: PYTHIA and HERWIG, ungroomed and groomed. The ratio panel uses the ungroomed PYTHIA distribution as reference.

FIGURE 4.16: Normalized regression output distributions for the **HAD-on** (left) and **HAD-off** (right) truth-level configurations in Study A. All truth-level peaks appear at $\hat{m}_a \approx 2.1$ GeV instead of the true 3.0 GeV, quantifying the ~ 0.9 GeV domain shift introduced by applying the DELPHES-trained model to truth-level inputs.

Several features of Figure 4.16 merit discussion. The ungroomed PYTHIA and HERWIG distributions are nearly identical for both **HAD-on** and **HAD-off**, confirming that the domain shift is not sensitive to the generator choice. Within each generator, the **HAD-on** and **HAD-off** ungroomed distributions differ by only a few percent in peak position, so the regression output cannot discriminate between hadronization configurations. The groomed distributions behave differently: for **HAD-on**, the groomed peak is slightly sharper, suggesting that **SoftDrop** removes some of the soft radiation that contributes to the spread

of predictions. For **HAD-off**, the groomed distribution is strikingly narrow, reflecting the clean two-parton topology at parton level. An ungroomed spike at $\hat{m}_a \approx 1.45$ GeV, arising from jets with very few reconstructed tracks for which the energy-correlation functions and τ_2 take degenerate values near zero, is entirely absent from the groomed distributions. This is one of the clearest practical benefits of grooming as a preprocessing step: it suppresses not only nonperturbative radiation but also the pathological low-track configurations that cause the regression network to produce uninformative predictions. The effect is shown explicitly for both generators in Figure 4.17.

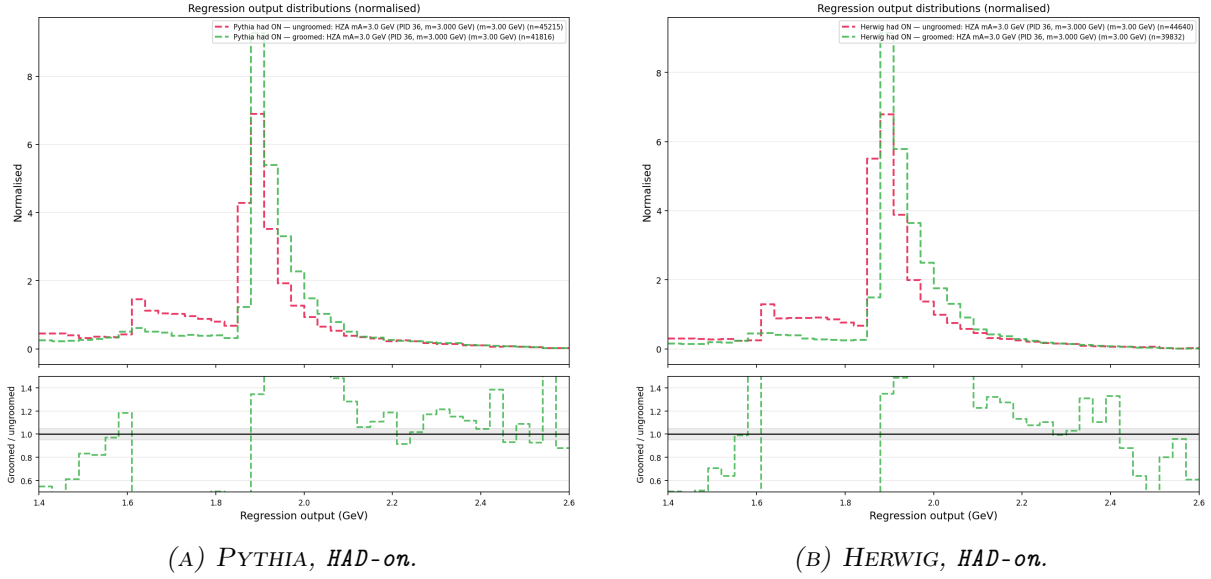


FIGURE 4.17: Effect of *SoftDrop* grooming on the regression output distribution for PYTHIA (left) and HERWIG (right) with full hadronization. Solid lines show the ungroomed distribution; dashed lines show the groomed distribution. The peak is preserved in both cases, while the ungroomed low-multiplicity artifact at $\hat{m}_a \approx 1.45$ GeV disappears after grooming.

Since the DELPHES-trained regression output is the eighth feature fed to the classifier, the ~ 0.9 GeV domain shift has direct implications for classification: the biased mass estimate may mislead the decision boundary, and the nature of this degradation depends on whether grooming has been applied. Table 4.4 reports the AUC for all eight truth-level configurations; the corresponding ROC curves are shown in Figures 4.19 and 4.20, and a summary comparison against the DELPHES baseline is shown in Figure 4.18. The degraded performance of the ungroomed classifier ($\text{AUC} \approx 0.63\text{--}0.66$) relative to the DELPHES baseline reflects the domain shift introduced by applying a scaler fitted on DELPHES distributions to truth-level inputs; the inversion of the groomed classifier below the random-classifier threshold ($\text{AUC} \approx 0.37\text{--}0.51$) arises because **SOFTDROP** grooming removes the soft-track structure that the network learned to associate with signal, causing it to systematically score groomed signal jets as background.

Three distinct regimes emerge from Table 4.4. All four ungroomed configurations yield AUC values in the range 0.62–0.66, a reduction of approximately 0.17–0.21 relative to the

Generator	HAD	Grooming	AUC
DELPHES (baseline)	ON	ungroomed	0.831
PYTHIA	ON	ungroomed	0.629
PYTHIA	ON	groomed	0.369
PYTHIA	OFF	ungroomed	0.657
PYTHIA	OFF	groomed	0.507
HERWIG	ON	ungroomed	0.625
HERWIG	ON	groomed	0.375
HERWIG	OFF	ungroomed	0.643
HERWIG	OFF	groomed	0.454

TABLE 4.4: Classification AUC for the DELPHES-trained network of Study A evaluated on truth-level signal mixed with DELPHES Z +jets background.

DELPHES baseline. The classifier continues to perform above chance, indicating that the ghost-track variables retain discriminating power even under the domain shift. Crucially, the generator (PYTHIA vs. HERWIG) and hadronization (HAD-on vs. HAD-off) make negligible differences: the four ungroomed AUC values are mutually consistent within ± 0.02 . Three of the four groomed configurations yield $\text{AUC} < 0.5$, with values as low as 0.37. An AUC below 0.5 does not indicate random performance; rather, it indicates that the classifier systematically assigns higher signal scores to background jets than to signal jets. **SoftDrop** grooming removes soft tracks from the truth-level jets, making signal jets look sparser and simpler — more characteristic of the background in the feature space the model has learned. The classification score is consequently inverted. This is a diagnostic result rather than a failure: it confirms that the network has extracted genuine discriminating information, but the preprocessing misalignment swaps the signal and background labels in the groomed feature space, and it directly motivates the need for a dedicated classifier trained on groomed features. The groomed HAD-off configurations show partial recovery (AUC values of 0.507 and 0.454), since the clean two-parton core at parton level is in some respects more stable once **SoftDrop** has removed the soft branches.

4.5.3 Study B: Self-Consistent Detector-Level Evaluation

The domain-transfer study of Study A has exposed a structural limitation of evaluating a DELPHES-trained model on truth-level samples. Three sources of inconsistency compound one another. The first is the feature normalization: truth-level features, when passed through a scaler fitted on DELPHES distributions, are mapped to the wrong region of the input space, explaining the uniform ~ 0.9 GeV shift observed across all eight evaluation configurations. The second is the asymmetric background treatment: the classifier was trained to separate DELPHES signal from DELPHES background, but evaluation signal came from truth level, misaligning the decision boundary in a way that can invert the classifier entirely, as the groomed rows of Table 4.4 demonstrate. The third follows

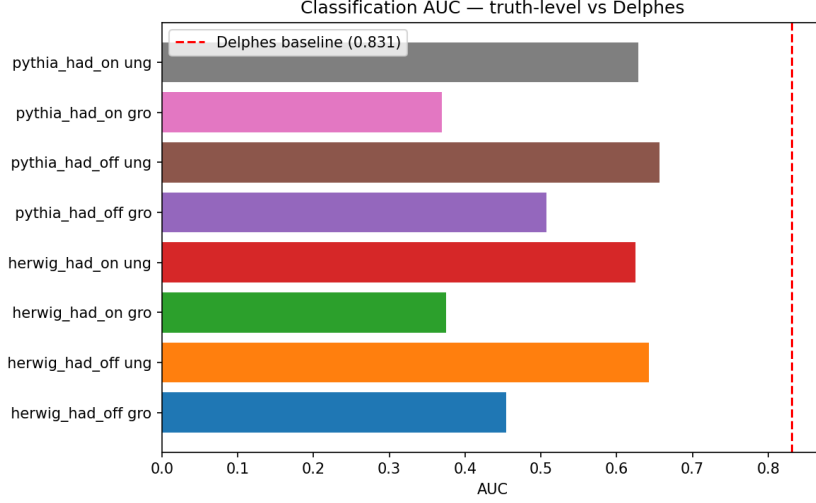


FIGURE 4.18: Summary of classification AUC values for the DELPHES baseline and all eight truth-level configurations in Study A. The red dashed line marks the in-domain AUC of 0.831. Groomed configurations (lighter bars) consistently underperform their ungroomed counterparts.

from the first two: the AUC spread across the eight configurations reflects preprocessing misalignment, not genuine generator differences. Study B corrects all three: signal and background are both drawn from the DELPHES simulation and the robust feature scaler is refitted on DELPHES training data only. Any residual AUC difference between evaluation configurations then constitutes a direct, uncontaminated estimate of the theoretical uncertainty due to the choice of Monte Carlo generator at detector level.

Setup. Using Study A’s setup, two classifiers are trained in parallel: one using the seven ungroomed ghost-track variables augmented by the ungroomed regression output (hereafter UNG), and one using the same variables recomputed after **SoftDrop** grooming augmented by the groomed regression output (hereafter GRO). Both are three-hidden-layer MLPs with widths 128, 64, and 32 neurons, trained with the Adam optimizer, batch size 256, and early stopping with patience 25. The regression networks are trained first, separately for each study, on all nine signal mass hypotheses spanning $m_a = 0.50\text{--}4.00$ GeV together with the DELPHES Z +jets background; their scalar outputs are appended as eighth features before the classifiers are trained at $m_a = 3.0$ GeV.

Regression performance. The regression trained on ungroomed variables achieves a test MAE of 0.960 GeV and the groomed regression achieves 0.931 GeV, with signal accuracies of 52.5% and 55.0% respectively. Both networks exhibit a near-mean-collapse in which the predicted mass concentrates near the training mean of approximately 2.2 GeV regardless of the true mass hypothesis. The per-class breakdown is shown in Table 4.5: accuracy peaks at 99.8% for $m_a = 2.0$ GeV and drops to 0% for $m_a = 4.0$ GeV in both studies. The collapse is a consequence of the limited discriminating power of the seven

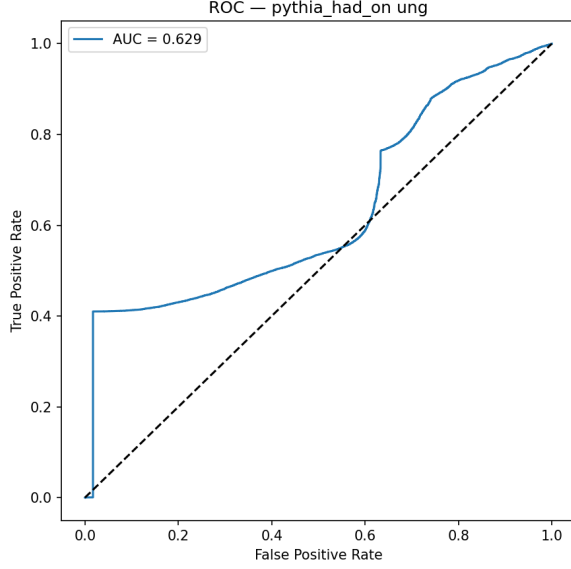
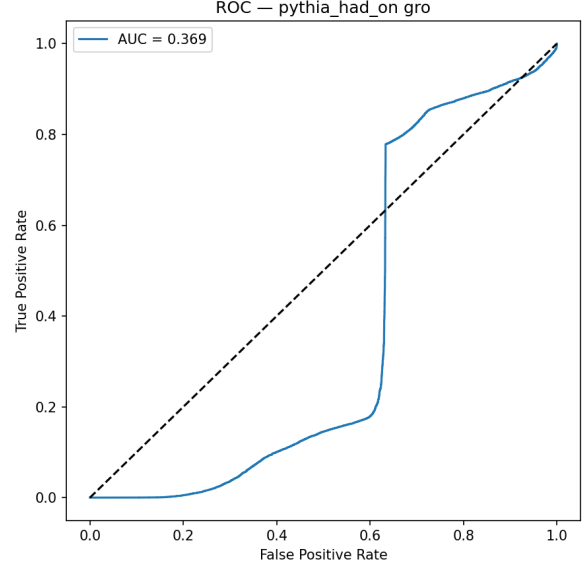
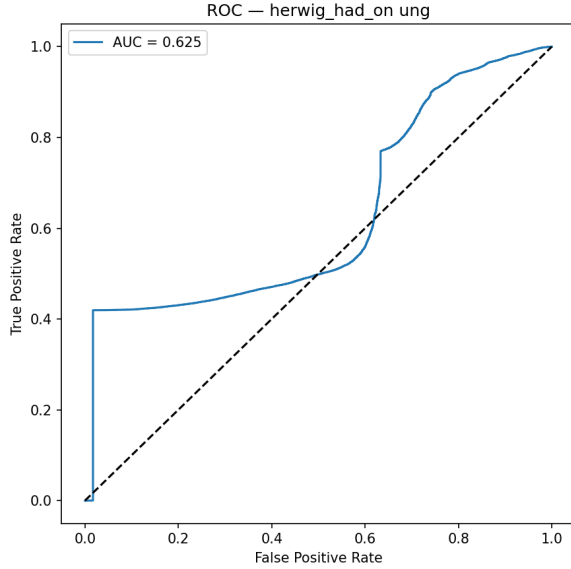
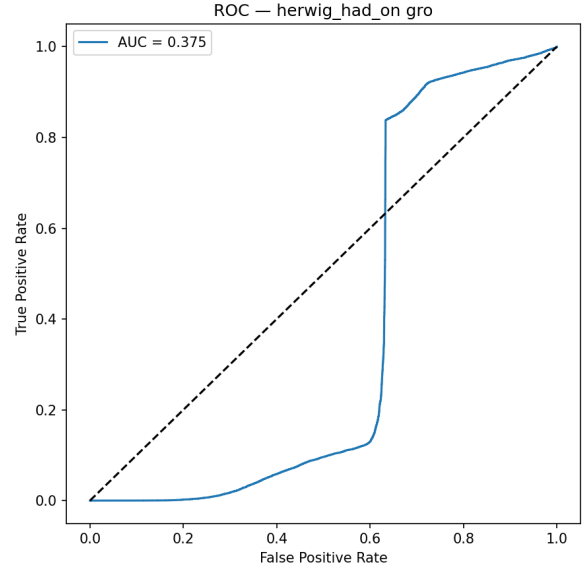
(A) PYTHIA *HAD-on*, ungroomed. $AUC = 0.629$.(B) PYTHIA *HAD-on*, groomed. $AUC = 0.369$.(C) HERWIG *HAD-on*, ungroomed. $AUC = 0.625$.(D) HERWIG *HAD-on*, groomed. $AUC = 0.375$.

FIGURE 4.19: ROC curves for the Study A classifier evaluated on *HAD-on* truth-level configurations. Ungroomed inputs (left column) yield $AUC \approx 0.63$, substantially below the DELPHES baseline but still above random. Groomed inputs (right column) yield $AUC \approx 0.37$, below the random-classifier threshold of 0.5: the network has effectively inverted its classification (see text for interpretation).

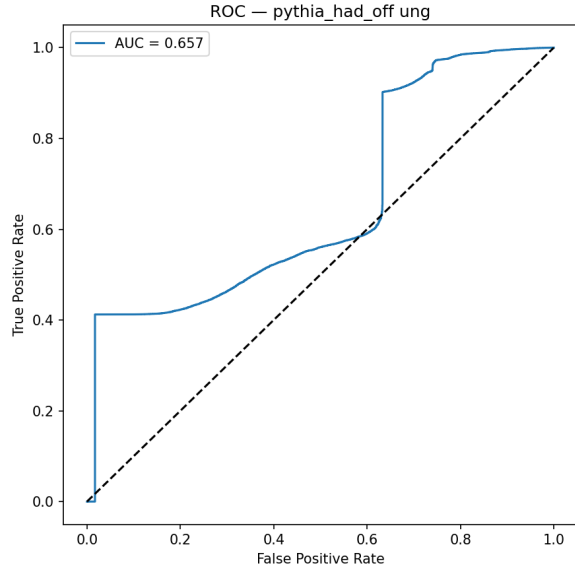
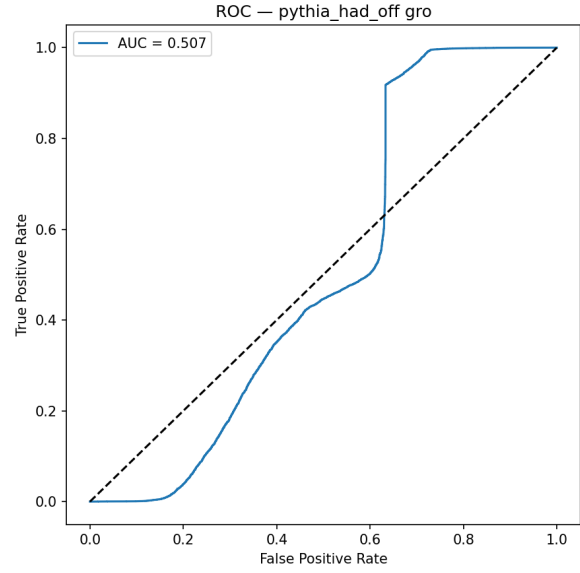
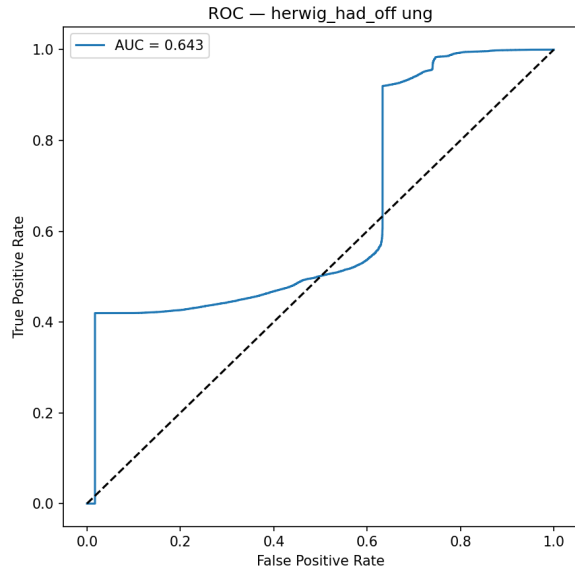
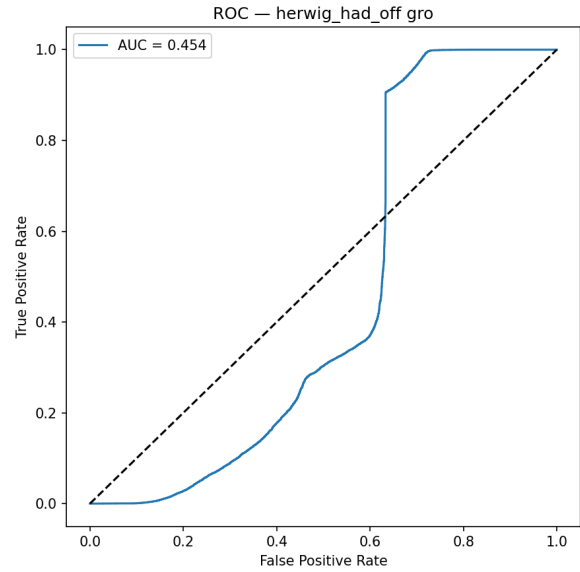
(A) PYTHIA *HAD-off*, ungroomed. $AUC = 0.657$.(B) PYTHIA *HAD-off*, groomed. $AUC = 0.507$.(C) HERWIG *HAD-off*, ungroomed. $AUC = 0.643$.(D) HERWIG *HAD-off*, groomed. $AUC = 0.454$.

FIGURE 4.20: ROC curves for the Study A classifier evaluated on *HAD-off* truth-level configurations. Ungroomed inputs (left column) yield marginally higher AUC values than their *HAD-on* counterparts, consistent with parton-level features being slightly closer to the DELPHES training distribution. Groomed inputs (right column) remain near or below the random-classifier threshold.

track-based variables across the full 0.5–4.0 GeV mass range, and is partially but not fully mitigated by grooming. The groomed regression also produces ten times fewer jets with invalid features (98 versus 1 142), since the **SoftDrop** hard core has more numerically stable kinematics than the full ungroomed jet. A targeted remedy for this collapse is investigated in Study C (Section 4.5.4).

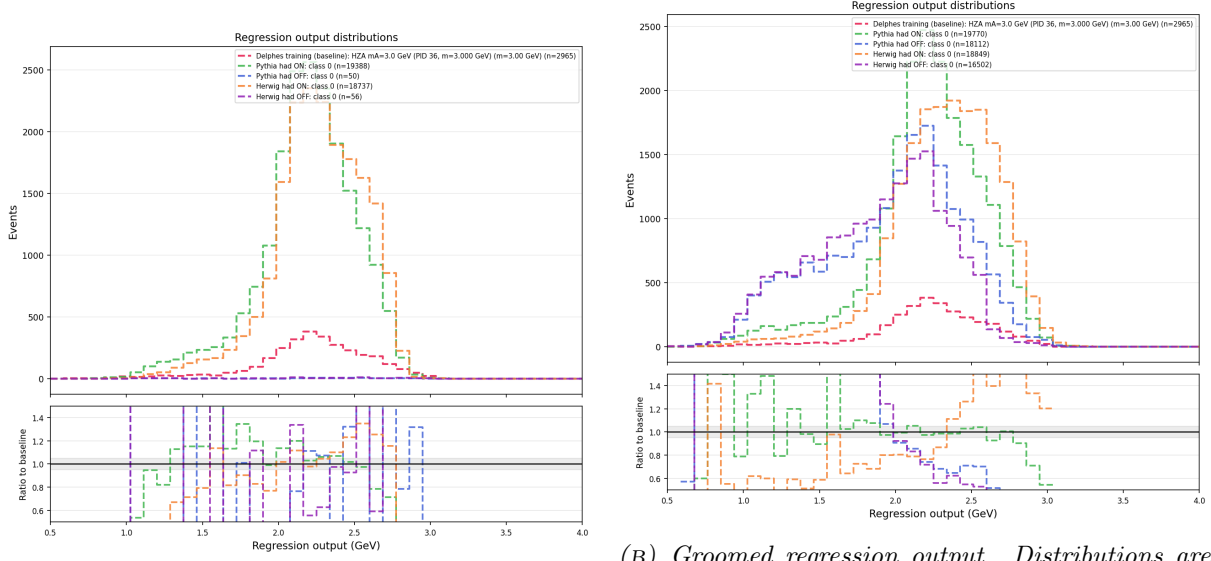
m_a	UNG acc.	UNG MAE	GRO acc.	GRO MAE
0.50 GeV	20.3%	1.473 GeV	32.2%	1.293 GeV
0.75 GeV	32.3%	1.152 GeV	33.7%	1.093 GeV
1.00 GeV	37.9%	0.972 GeV	43.4%	0.917 GeV
1.50 GeV	91.7%	0.565 GeV	91.6%	0.597 GeV
2.00 GeV	99.8%	0.290 GeV	98.1%	0.350 GeV
2.50 GeV	94.4%	0.381 GeV	89.2%	0.432 GeV
3.00 GeV	77.8%	0.809 GeV	80.1%	0.768 GeV
3.50 GeV	16.0%	1.277 GeV	21.6%	1.265 GeV
4.00 GeV	0.0%	1.778 GeV	0.0%	1.744 GeV

TABLE 4.5: Per-class regression test accuracy and MAE for the ungroomed (UNG) and groomed (GRO) studies in Study B, evaluated on the DELPHES test split.

Figure 4.21 shows the regression output distributions for all four evaluation configurations in Study B. In both the ungroomed and groomed cases the distributions are broadly consistent across generator and hadronization settings, confirming that the near-mean-collapse is a property of the feature space rather than a generator-specific artifact. The distributions made by groomed variables are somewhat narrower than their ungroomed counterparts, consistent with the removal of soft radiation that broadens the energy-correlation inputs.

In-domain classification performance. The UNG classifier achieves a test AUC of 0.983 and the GRO classifier achieves 0.618. Figure 4.22 shows the two ROC curves on the same axes. The ungroomed in-domain AUC is substantially higher than the Study A baseline of 0.831, confirming that removing the preprocessing inconsistency reveals the true in-sample discrimination power of the ghost-track variables. The GRO classifier reaches its optimal validation loss at epoch 6 and triggers early stopping at epoch 47; the UNG classifier continues improving for the full 100 epochs. This difference in convergence reflects the information content of the two feature sets: the ungroomed variables carry enough soft-radiation structure to permit near-perfect separation of training signal from background, while the groomed variables exhaust their discriminating power rapidly once that soft structure is removed.

Evaluation across generator configurations. Table 4.6 reports the AUC for both classifiers evaluated on all four generator and hadronization configurations.



(A) Ungroomed regression output. Distributions cluster near the training mean of ≈ 2.2 GeV, reflecting the near-mean-collapse of Table 4.5.

(B) Groomed regression output. Distributions are narrower and the inter-generator spread is reduced, consistent with the suppression of soft-radiation contributions to the energy-correlation inputs.

FIGURE 4.21: Regression output distributions for all four evaluation configurations in Study B. Ratio panels use the PYTHIA HAD-on distribution as reference.

Classifier score distributions. The ROC curves of Figure 4.22 summarize the discrimination power of each classifier as a single threshold-integrated number, but they do not reveal how the two classifiers distribute their output scores across signal and background jets. Figure 4.23 shows the classifier output score distributions for both the GRO and UNG classifiers in Study B. The left panels compare the in-domain signal ($m_a = 3$ GeV, PYTHIA HAD-on) against the Z +jets background, while the right panels overlay the signal score distributions across all four evaluation configurations (PYTHIA and HERWIG, HAD-on and HAD-off), with the background shown as a filled histogram for reference.

For the GRO classifier (top row), the signal and background distributions are visibly separated, with the signal peak sitting above the background across the full range of evaluation configurations. Crucially, the four signal score curves are nearly superimposed in the right panel, confirming that the groomed classifier assigns consistent scores to signal jets regardless of the generator or hadronization setting — the robustness already quantified by the AUC values of Table 4.6 is here made visually manifest.

The UNG classifier (bottom row) tells a more instructive story. While the in-domain signal peak sits near a score of 0.72, well separated from the background peak at ≈ 0.48 , the right panel reveals the generalization collapse: the HERWIG HAD-on signal distribution collapses entirely onto the background, consistent with the near-random AUC of 0.507 reported in Table 4.6. This contrast illustrates a general diagnostic point: a high in-domain score separation is not sufficient evidence of classifier robustness, and evaluation across the full set of generator configurations is necessary to expose generator-specific

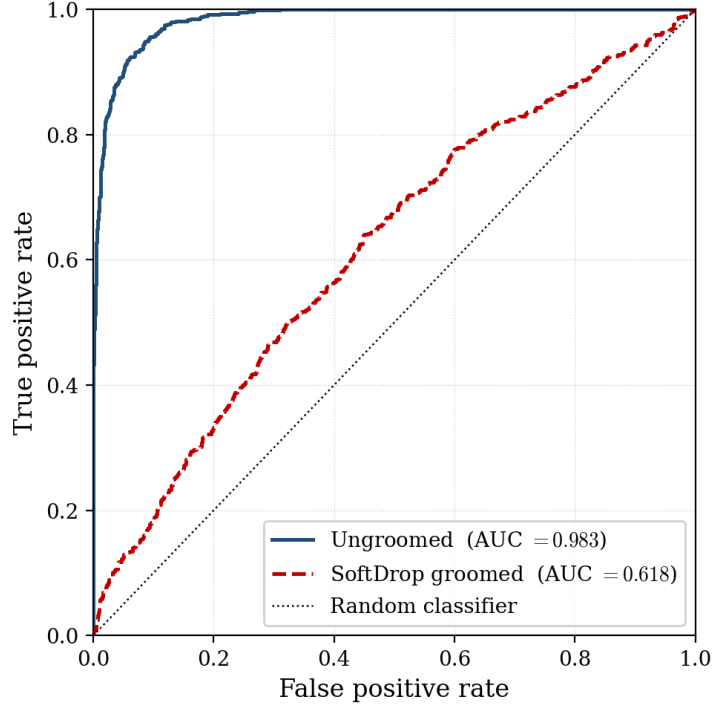


FIGURE 4.22: ROC curves for the UNG ($AUC = 0.983$) and GRO ($AUC = 0.618$) classifiers evaluated on the DELPHES test set at $m_a = 3.0$ GeV. The gap between the two curves quantifies the discrimination power lost by applying *SoftDrop* grooming before feature extraction.

overfitting.

A contributing factor to the degraded ROC performance is the limited number of simulated signal events available for training and evaluation. The classification network was trained on approximately 3 000 signal jets at $m_a = 3.0$ GeV, a sample size small enough for the network to memorize generator-specific features rather than learning robust physical discriminants. The AUC values reported across evaluation configurations should therefore be interpreted with caution: statistical fluctuations in the score distributions are non-negligible at this event count, and the generator gap $\Delta AUC = 0.089$ carries an uncertainty that is difficult to quantify without a dedicated ensemble study. Increasing the number of simulated events — particularly for the mass hypotheses at the extremes of the training range, where the per-class accuracy is lowest — would be the most direct way to disentangle genuine generator systematics from statistical noise.

Generalization collapse of the UNG classifier. When evaluated on PYTHIA HAD-on and HERWIG HAD-on DELPHES configurations, the UNG classifier achieves AUC values of 0.549 and 0.507 respectively, a drop of approximately 0.44 relative to the in-domain baseline of 0.983. This collapse to near-random performance occurs despite the evaluation signal coming from the same detector simulation as the training data: the only difference

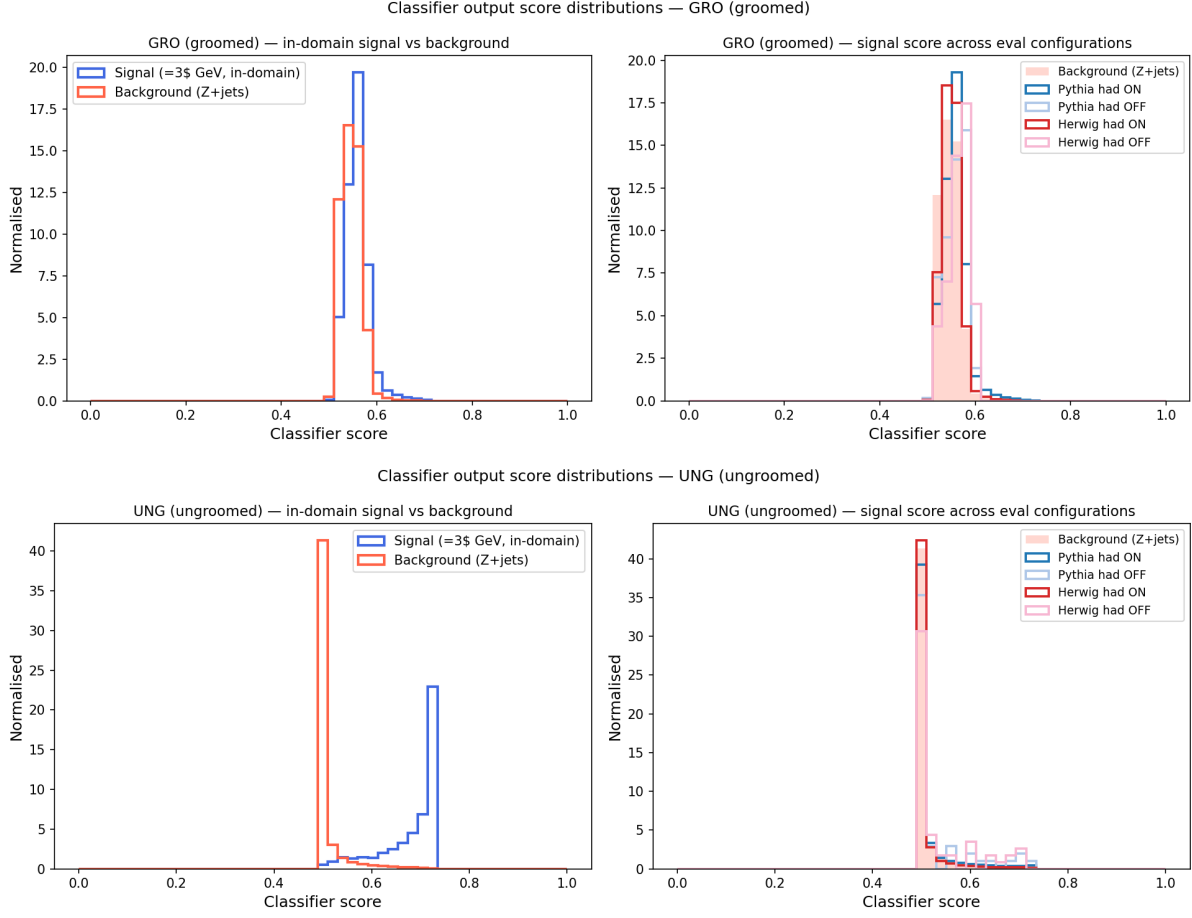


FIGURE 4.23: Classifier output score distributions for the GRO (top) and UNG (bottom) classifiers in Study B. Left panels: normalized score distributions for the in-domain signal ($m_a = 3$ GeV, PYTHIA HAD-on, blue) and the Z+jets background (orange), confirming that both classifiers achieve in-domain signal-background separation consistent with the AUC values of Table 4.6. Right panels: signal score distributions across all four evaluation configurations (PYTHIA and HERWIG, HAD-on and HAD-off), with the Z+jets background shown as a filled histogram for reference. The GRO right panel shows the four evaluation curves nearly superimposed, confirming that the groomed classifier generalizes stably across generator and hadronization configurations. The UNG right panel reveals the generalization collapse: while the in-domain signal peak sits near a score of 0.72, the HERWIG HAD-on signal distribution (solid red) collapses onto the background peak, consistent with the near-random AUC of 0.507 reported in Table 4.6. This contrast illustrates that visual inspection of the in-domain score distribution alone is insufficient to assess classifier robustness; evaluation across the full set of generator configurations is required.

Generator	HAD	UNG AUC	GRO AUC
DELPHES (baseline)	ON	0.983	0.618
PYTHIA	ON	0.549	0.644
PYTHIA	OFF	0.606 [†]	0.675
HERWIG	ON	0.507	0.555
HERWIG	OFF	0.758 [†]	0.756
Generator gap (PYTHIA ON – HERWIG ON)		+0.042	+0.089

TABLE 4.6: Classification AUC for the self-consistent detector-level evaluation in Study B. Entries marked [†] are computed from approximately 50 valid jets and are excluded from all physics conclusions.

is that the evaluation events were generated in a separate run. The UNG classifier has not learned physically generalizable features; it has memorized the specific soft-radiation patterns of the $\approx 3\,000$ training signal jets and fails to transfer to any other configuration of the same signal process.

Grooming as a condition for generalization. The GRO classifier achieves AUC values of 0.644 and 0.555 on the PYTHIA and HERWIG HAD-on configurations respectively, outperforming the UNG classifier on every evaluation setting in Table 4.6. Compared with the GRO in-domain baseline of 0.618, the evaluation performance is consistent or higher, confirming that the groomed classifier generalizes without collapse. The cost of this robustness is a reduction of 0.365 AUC points relative to the UNG in-domain performance: **SoftDrop** grooming removes the soft-radiation features that made the ungroomed signal easily separable from background within the training domain. A classifier that achieves AUC = 0.983 in-sample but 0.549 on evaluation is not useful as a physics discriminant; a classifier that achieves AUC = 0.618 in-sample and 0.644 on evaluation is. **SoftDrop** grooming is therefore a necessary preprocessing step for the classifier to produce predictions that are stable across simulation configurations.

Generator gap and systematic uncertainty. The groomed generator gap $\Delta\text{AUC} = 0.089$ is the first physically interpretable estimate of the PYTHIA/HERWIG systematic uncertainty in this consistent evaluation framework. Both GRO classifiers operate well above random, making the comparison meaningful in a way that the UNG gap is not. The magnitude of this gap is consistent with the generator dependence observed in the jet substructure variable distributions of Section 4.4: **SoftDrop** grooming suppresses but does not eliminate the inter-generator spread in N_{tracks} , and the groomed classifier’s residual generator dependence is a direct reflection of this unmitigated multiplicity sensitivity. Figure 4.24 makes this connection explicit by showing the N_{tracks} distributions for all four evaluation configurations, comparing the ungroomed and groomed cases side by side. The ROC curves for the HAD-on and HAD-off configurations are shown in Figures 4.25 and 4.26. A summary of all AUC values is shown in Figure 4.27.

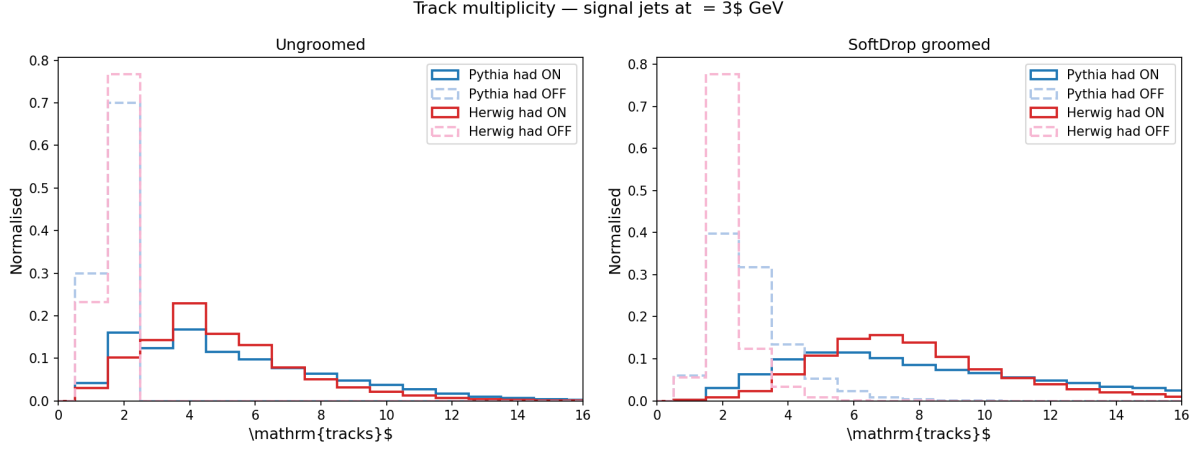


FIGURE 4.24: Track multiplicity N_{tracks} for all four evaluation configurations at $m_a = 3$ GeV, shown for the ungroomed (left) and *SoftDrop*-groomed (right) jet definitions. *PYTHIA HAD-on* (solid blue) and *HERWIG HAD-on* (solid red) differ in both cases; the *HAD-off* distributions (dashed) are shifted to lower multiplicities. Grooming reduces the inter-generator spread but leaves a residual difference at intermediate track counts, which is the principal source of the groomed generator gap $\Delta AUC = 0.089$.

Hadronization study and the ungroomed HAD-off limitation. A structural limitation prevents a meaningful ungroomed hadronization comparison. Without hadronization, the DELPHES simulation produces no EFlowTrack objects for approximately 99.8% of parton-level jets, because there are no final-state hadrons to reconstruct. The seven track-based substructure variables are undefined for these jets, and the UNG HAD-off AUC values are computed from approximately 50 valid jets out of roughly 20 000. They are retained in Table 4.6 for completeness but are excluded from all physics conclusions. The GRO study partially recovers the HAD-off evaluation: *SoftDrop* operates at the jet-clustering level rather than on final-state hadrons, restoring valid features for 84–89% of parton-level jets. The GRO HAD-off AUC values of 0.675 and 0.756 exceed their HAD-on counterparts, indicating that parton-level $a \rightarrow q\bar{q}$ decays present a cleaner two-prong structure that is more distinct from the fully hadronized Z +jets background.

Section Roadmap:

Covered: Two-stage neural network pipeline for $H \rightarrow Za$: the regression network for pseudoscalar mass prediction and the classification network for signal-versus-background separation, with their shared MLP architecture described in Section 4.5.1; Study A establishing the in-domain regression accuracy (68% signal accuracy across nine mass hypotheses) and classification AUC (0.831), and the domain-transfer degradation (ungroomed AUC ≈ 0.63 – 0.66 ; groomed classifier inversion to AUC ≈ 0.37 – 0.51); Study B correcting the three sources of preprocessing inconsistency and revealing the true generalization behavior: UNG collapse from AUC = 0.983 to ≈ 0.53 ,

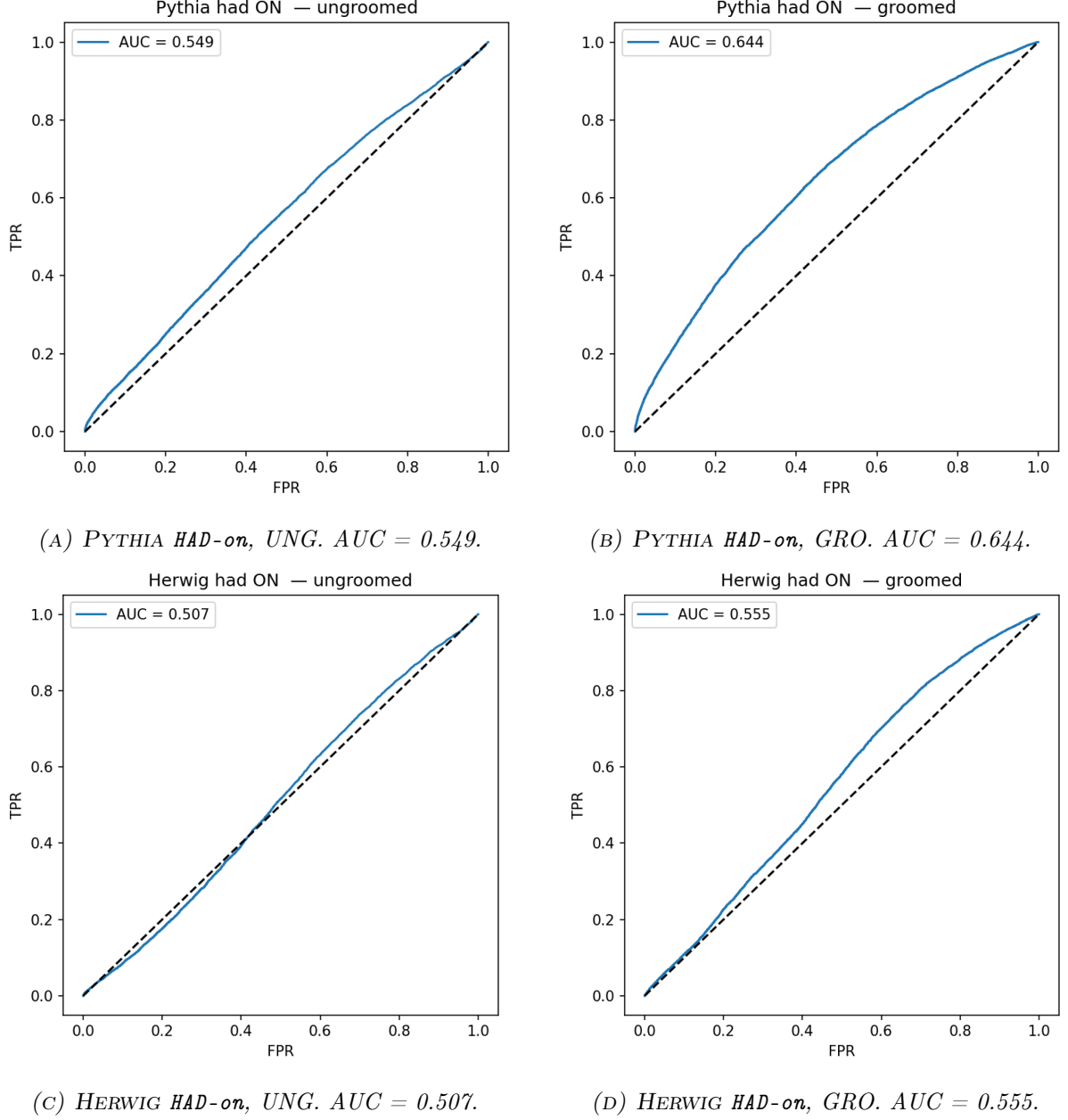


FIGURE 4.25: ROC curves for the *HAD-on* evaluation configurations in Study B. The UNG AUC values of 0.549 and 0.507 represent a collapse of ≈ 0.44 relative to the in-domain baseline. The GRO AUC values of 0.644 and 0.555 are consistent with the GRO in-domain baseline of 0.618, confirming that the groomed classifier generalizes without collapse.

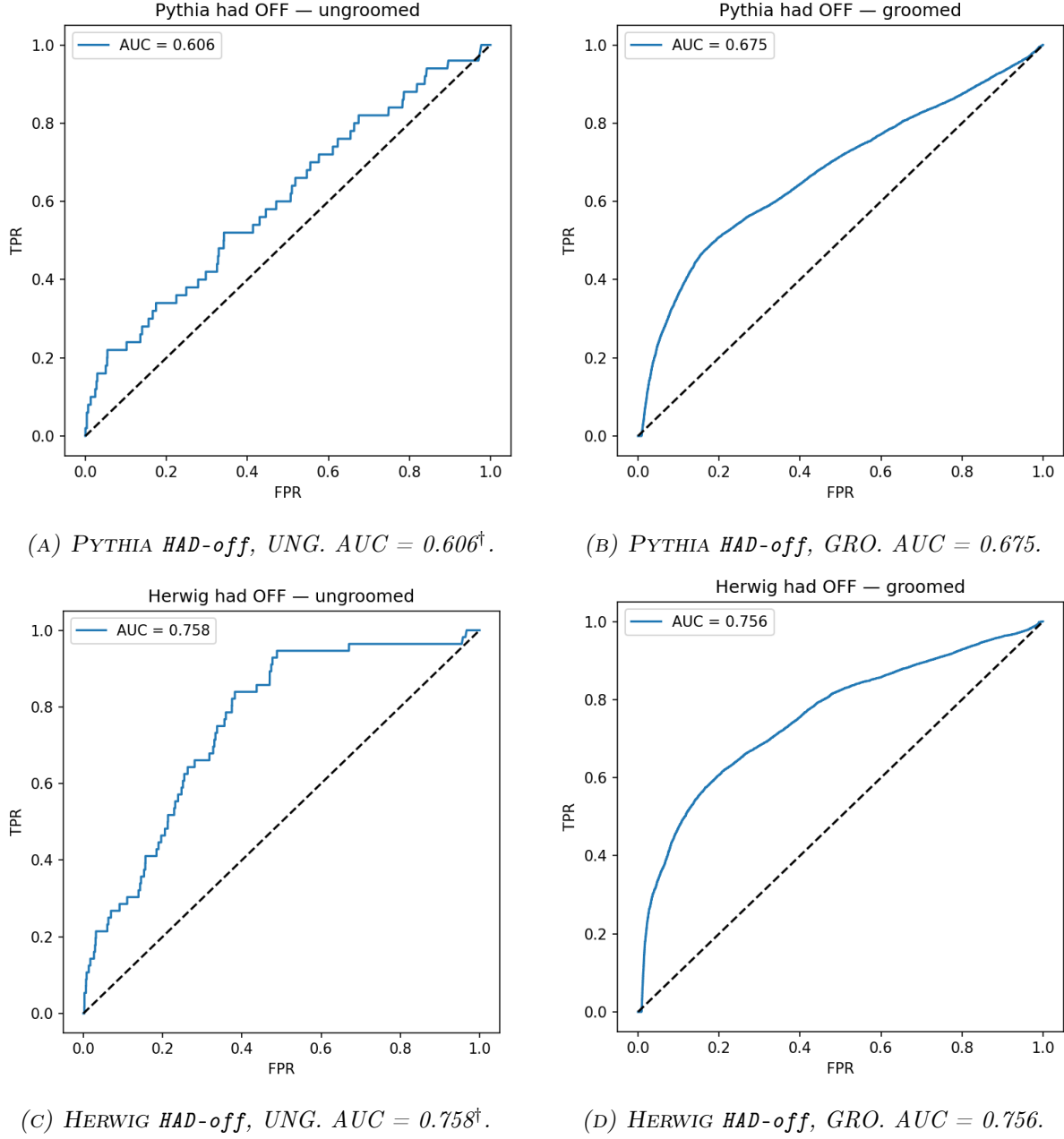


FIGURE 4.26: ROC curves for the *HAD-off* evaluation configurations in Study B. *UNG* entries ([†]) are computed from ≈ 50 valid jets and excluded from physics conclusions. The *GRO* classifier recovers valid features for 84–89% of parton-level jets, achieving AUC values of 0.675 and 0.756.

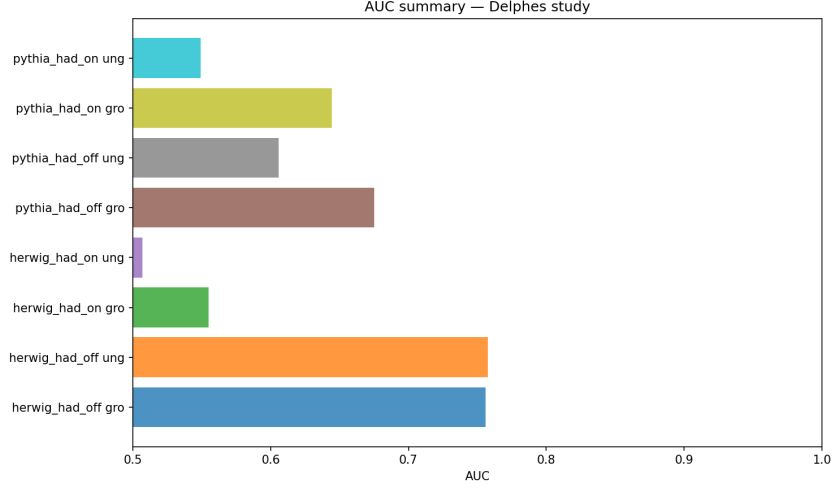


FIGURE 4.27: Summary of classification AUC values for the UNG (dark bars) and GRO (light bars) classifiers across all evaluation configurations in Study B. Vertical dashed lines mark the respective in-domain baselines (UNG = 0.983, GRO = 0.618). The GRO classifier outperforms the UNG classifier on every *HAD-on* configuration despite its lower in-domain AUC. Entries marked † in the previous plots are excluded from physics conclusions.

GRO stable generalization (AUC = 0.618 in-domain, 0.644 on PYTHIA evaluation); the groomed generator gap $\Delta\text{AUC} = 0.089$ as the first physically interpretable estimate of the PYTHIA/HERWIG systematic uncertainty on classifier performance; Study C (Section 4.5.4) verifying the robustness of the GRO generator gap under balanced regression training.

References: [38]; [67]; [64]; [65]; [62]; [63]; [66]; [68]; [70].



4.5.4 Study C: Balanced Regression Training

Study B identified a near-mean-collapse in the regression network: the predicted mass concentrates near the training mean of ≈ 2.2 GeV regardless of the true mass hypothesis, with accuracy dropping to 0% at $m_a = 4.0$ GeV. This collapse arises because the standard training procedure draws mini-batches by uniform random sampling from the concatenated training pool, allowing mass classes with more easily predictable features to dominate the gradient signal. Study C addresses this with two complementary modifications to the regression training, leaving the network architecture, the classifier, and all downstream steps unchanged.

The first modification replaces the default `DataLoader` with a `WeightedRandomSampler`. To understand why this is necessary, consider what happens during standard training. The nine signal mass files contribute approximately 1,600–1,800 jets each to the training pool, and background jets are discarded before computing the regression loss because they carry no mass target. When mini-batches are drawn by uniform random sampling, a jet from $m_a = 2.0$ GeV is drawn with the same probability as a jet from $m_a = 4.0$ GeV. This sounds fair, but it is not: the network finds it far easier to predict the mid-range masses, where the seven variables carry the most discriminating power, than the extremes. The optimizer therefore converges quickly on the easy classes and effectively stops learning from the hard ones, since the loss contribution from a class where predictions are already accurate is small regardless of how often those jets appear in a batch. The net result is that the training mean of ≈ 2.2 GeV acts as an attractor: predicting that value for every jet incurs a lower average loss than learning the full mass dependence, and the network collapses toward it.

Study C assigns each signal jet in class c a sampling weight

$$w_i = \frac{1}{N_c}, \quad i \in \text{class } c,$$

where N_c is the number of training jets in that class. The `WeightedRandomSampler` then draws mini-batches such that the expected number of jets from each mass hypothesis is equal per epoch, regardless of the absolute class sizes. Concretely, a jet from $m_a = 4.0$ GeV is now drawn approximately $1,694/1,600 \approx 1.06$ times more often than a jet from $m_a = 2.0$ GeV — a mild numerical correction, but one that prevents the easy mid-range classes from accumulating gradient steps at the expense of the hard ones. This is the training analog of reweighting a Monte Carlo sample to a flat prior in a control variable: just as one reweights events to remove a known kinematic bias from an efficiency measurement, here the training draws are reweighted to remove the accidental frequency bias from the dataset assembly.

The second modification normalizes the loss within each mini-batch by mass class. Even with a balanced sampler, the Huber loss averaged over all jets in a batch is still dominated by the classes where the per-jet loss is largest — which are precisely the classes the network finds hardest. This sounds beneficial, but it creates a different problem: once the network has reduced the loss on a mid-range class to near zero, that class contributes almost

nothing to the gradient, even though the hard classes (whose loss remains large) would benefit from the network spending more capacity on them. After computing the per-jet Huber loss, the jets in the batch are therefore grouped by their rounded target mass, the mean loss is computed separately within each group, and the batch loss used for backpropagation is the simple average of the K per-class losses:

$$\mathcal{L}_{\text{batch}} = \frac{1}{K} \sum_{k=1}^K \bar{\ell}_k.$$

Under this scheme, a class where the network already predicts accurately contributes exactly zero gradient, while a class where predictions are still poor contributes the same gradient magnitude as any other class — regardless of how many jets from that class appear in the batch. The two modifications therefore act at different levels: the sampler equalizes how often each class is seen across epochs, and the per-class loss equalizes how much each class contributes to the gradient within each batch. Both are needed simultaneously, because using only the sampler leaves the gradient dominated by the high-loss classes within each batch, and using only the per-class loss is ineffective when some classes appear in too few batches to accumulate meaningful gradient signal.

Regression performance. The Study C regression achieves overall test MAE of 0.968 GeV (UNG) and 0.936 GeV (GRO), essentially unchanged from Study B. The per-class pattern, however, shifts significantly: the mid-range classes that were near-perfect in Study B ($m_a = 1.5\text{--}2.5$ GeV, accuracy $> 91\%$) are deliberately de-emphasized, while gains at the mass extremes are modest — $m_a = 0.5$ GeV improves from 20.3% to 21.9% (UNG) and from 32.2% to 33.0% (GRO); $m_a = 4.0$ GeV reaches 4.3% (UNG) and 0.6% (GRO), compared to 0% in Study B. The limited improvement at the extremes reflects a genuine feature-space limitation: the seven track-based variables carry qualitatively similar information for jets at $m_a = 3.5$ and 4.0 GeV, and no reweighting scheme can extract discriminating power that is not present in the inputs. The eval-set regression mean remains at ≈ 2.2 GeV, confirming that the near-mean-collapse on the evaluation configurations is driven by the feature-space limitation rather than by the training imbalance.

Classification performance. The Study C UNG classifier achieves a test AUC of 0.629 in-domain, compared to 0.983 in Study B. This reduction is the direct consequence of the balanced training: by preventing the regression from concentrating on the mid-range mass classes, the classifier’s eighth feature no longer encodes the generator-specific soft-radiation patterns that made in-domain separation near-perfect. The penalty is substantial but expected and physically well-motivated. The GRO in-domain AUC is 0.623, essentially unchanged from Study B’s 0.618.

The key physics result is the evaluation performance. Table 4.7 reports the full AUC comparison between Study B and Study C.

Configuration	Study B		Study C		Δ (C – B)	
	UNG	GRO	UNG	GRO	Δ UNG	Δ GRO
DELPHES (baseline)	0.983	0.618	0.629	0.623	−0.354	+0.005
PYTHIA ON	0.549	0.644	0.645	0.643	+0.096	−0.001
PYTHIA OFF	0.606 [†]	0.675	0.748 [†]	0.619	—	−0.056
HERWIG ON	0.507	0.555	0.565	0.557	+0.058	+0.002
HERWIG OFF	0.758 [†]	0.756	0.907 [†]	0.695	—	−0.062
Generator gap (PYTHIA ON – HERWIG ON)			UNG: +0.080		GRO: +0.086	

TABLE 4.7: Classification AUC for Study C compared to Study B. Entries marked [†] are computed from approximately 50 valid jets and are excluded from physics conclusions.

The UNG evaluation AUC on PYTHIA HAD-on increases from 0.549 to 0.645, and on HERWIG HAD-on from 0.507 to 0.565: the balanced regression partially cures the UNG generalization collapse identified in Study B. The GRO results are unchanged to within statistical fluctuations ($0.644 \rightarrow 0.643$ for PYTHIA HAD-on, $0.555 \rightarrow 0.557$ for HERWIG HAD-on), confirming that the groomed classifier’s generalization behavior is insensitive to the quality of the eighth feature.

The central physics result of Study C is the GRO generator gap: $\Delta_{\text{gen}}^{\text{GRO}} = +0.086$, compared to +0.089 in Study B, a difference of -0.003 . This stability across two regression training procedures with qualitatively different per-class accuracy profiles confirms that the groomed generator gap is a physical property of the jet substructure variables — specifically, the residual inter-generator spread in N_{tracks} visible in Figure 4.24 — rather than a training artifact. The Study B conclusion is therefore robust.

Classifier score distributions. Figure 4.28 shows the classifier output score distributions for the UNG and GRO classifiers in Study C, organized in the same format as Figure 4.23. The left panels show the in-domain signal ($m_a = 3$ GeV, PYTHIA HAD-on) against the Z +jets background; the right panels overlay the signal score distributions across all four evaluation configurations.

For the GRO classifier (bottom panel), the picture is essentially unchanged relative to Study B: the four evaluation curves remain nearly superimposed in the right panel, confirming that the groomed classifier’s stable generalization across generator and hadronization configurations is insensitive to the quality of the regression eighth feature. The in-domain signal–background separation is also consistent with Study B, as expected from the near-identical GRO AUC values of 0.623 and 0.618.

The UNG classifier (top panel) shows the most significant change with respect to Study B. The in-domain separation is reduced compared to Study B, a direct consequence of the balanced regression training preventing the network from concentrating on the mid-range mass classes whose generator-specific soft-radiation patterns made in-domain separation

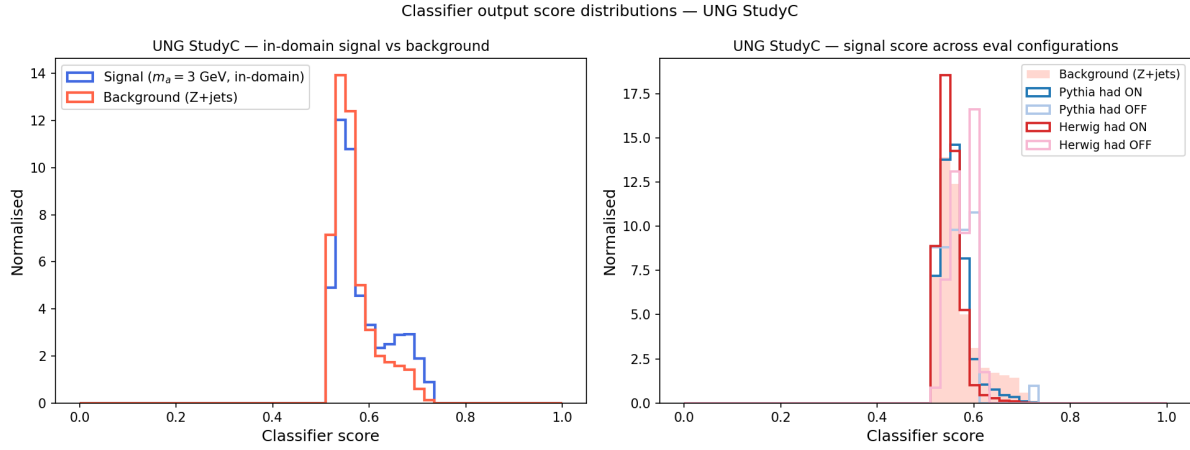
near-perfect. More importantly, the right panel shows a partial recovery of the generalization collapse: the HERWIG HAD-on signal distribution no longer fully collapses onto the background peak, consistent with the improved evaluation AUC of 0.565 reported in Table 4.7. This partial recovery confirms that the Study B collapse was driven in part by the regression memorizing generator-specific features at the mid-range mass classes, and that balanced training partially decouples the classifier from those artifacts — though the residual generator gap $\Delta\text{AUC} = 0.080$ indicates that a generator-dependent component survives in the ungroomed feature space even after rebalancing.

Section Roadmap:

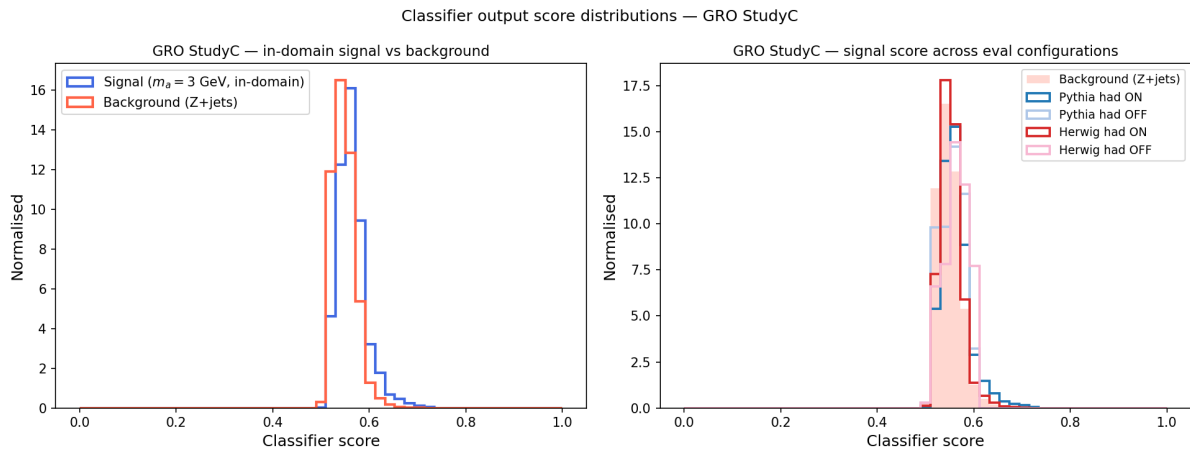
Covered: Summary of generator comparison results from Sections 4.3–4.4; summary of neural network Study A, Study B, and Study C findings from Section 4.5; physics implications for the ATLAS systematic uncertainty, including the confirmed robustness of the groomed generator gap.

References: [38].





(A) UNG classifier score distributions.



(B) GRO classifier score distributions.

FIGURE 4.28: Classifier output score distributions for Study C. Left panels: in-domain signal ($m_a = 3$ GeV, *PYTHIA HAD-on*) vs *Z+jets* background. Right panels: signal score across all four evaluation configurations. The GRO right panel shows the four evaluation curves nearly superimposed, as in Study B. The UNG right panel shows partial recovery of the generalization collapse: the *HERWIG HAD-on* signal distribution (solid red) no longer fully collapses onto the background peak, consistent with the improved evaluation AUC of 0.565 reported in Table 4.7.

Chapter 5

Conclusions and Outlook

*“It’s a magical world, Hobbes ol’ buddy. . .
Let’s go exploring!”*

— Bill Watterson, *Calvin and Hobbes* (1995)



THIS work set out to address two questions about the decay $H \rightarrow Za \rightarrow \ell\ell + \text{hadrons}$: how large is the theoretical uncertainty on the signal model introduced by the choice of Monte Carlo generator, and can a new machine-learning pipeline improve upon the sensitivity of the existing ATLAS search while remaining robust to that uncertainty? The preceding four Chapters built the tools to answer both questions and then applied them. This final Chapter collects the conclusions and identifies the most promising directions for future work.

5.1 Summary of the Work

The starting point was the ATLAS Run 2 search for $H \rightarrow Za$ with $a \rightarrow \text{hadrons}$ using the full 140 fb^{-1} dataset at $\sqrt{s} = 13 \text{ TeV}$, which found no significant excess above the Standard Model background, with a largest deviation of 1.5σ at $m_a = 0.5 \text{ GeV}$. The dominant theoretical uncertainty quoted in that analysis — approximately 60% on the signal yield from the PYTHIA/HERWIG comparison — provided the primary motivation for this study.

An independent simulation pipeline was constructed using POWHEG for the hard process, PYTHIA 8 and HERWIG 7 for parton shower and hadronization, FastJet for jet reconstruction directly on final-state particles, and DELPHES for detector-effect simulation. Before any physics comparison could be made, the pipeline was validated kinematically: both generators, when seeded from the same POWHEG hard-process input, reproduce consistent distributions for the reconstructed leptons, jets, and the three-body invariant mass $m_{\ell\ell j}$, confirming that observed differences arise from genuine modeling choices rather than simulation artifacts.

The generator comparison was then carried out at $m_a = 3.0$ GeV across three configurations: full hadronization enabled (**HAD-on**), hadronization disabled (**HAD-off**), and after SoftDrop grooming ($z_{\text{cut}} = 0.1$, $\beta = 1$). The seven jet substructure variables — N_{tracks} , ΔR , leading p_T ratio, angularity, U_1 , M_2 , and τ_2 — were examined in each configuration. The **HAD-on**/**HAD-off** comparison isolated the nonperturbative contribution to each variable, and grooming was assessed as a tool to reduce it.

On the machine-learning side, a two-stage neural network pipeline was developed. A regression network maps the seven substructure variables to a predicted pseudoscalar mass $\hat{m}_a$, and this prediction is appended as an eighth feature to a classification network that separates signal jets from Z +jets background. Three studies were carried out in sequence. Study A established a baseline using DELPHES-trained models evaluated on truth-level samples, and exposed a systematic domain-shift failure. Study B corrected this by keeping signal, background, and evaluation within the same DELPHES framework, producing the first physically interpretable AUC measurements. Study C then investigated whether the near-mean-collapse of the Study B regression — in which the predicted mass concentrated near 2.2 GeV regardless of the true hypothesis — affected the central classification conclusions, by introducing balanced class sampling and per-class loss normalization to the regression training.

5.2 Conclusions

5.2.1 Generator Comparison

The jet kinematic variables show mild hadronization dependence, most visibly in the jet mass distribution, which shifts toward higher values when hadronization is enabled as soft particles increase the invariant mass of the jet constituents. The jet substructure observables are considerably more sensitive: N_{tracks} , ΔR , leading p_T ratio, angularity, U_1 , M_2 , and τ_2 all shift systematically when hadronization is switched on, with PYTHIA consistently producing more soft activity than HERWIG across all variables. This reflects the fundamental difference between the Lund string model and the cluster fragmentation model, and the inter-generator spread across all observables constitutes a direct estimate of the nonperturbative modeling uncertainty.

SoftDrop grooming reduces the hadronization sensitivity of the energy-correlation family — angularity, U_1 , and M_2 — from the 20–40% level to the 10–20% level, and brings the

PYTHIA and HERWIG distributions into closer agreement. The track multiplicity N_{tracks} and angular separation ΔR , however, retain a hadronization sensitivity of order 30–50% even after grooming. These variables probe the fragmentation structure of the hard core rather than the soft peripheral radiation that **SoftDrop** removes. Any analysis relying on them must treat their associated nonperturbative uncertainty as a significant and largely unmitigated systematic.

5.2.2 Neural Network Analysis

Study A demonstrated the importance of preprocessing consistency: applying a scaler fitted on DELPHES distributions to truth-level inputs introduces a systematic ~ 0.9 GeV offset in the regression output, degrades the ungroomed classifier to $\text{AUC} \approx 0.63\text{--}0.66$, and inverts the groomed classifier entirely to $\text{AUC} \approx 0.37\text{--}0.51$. The inversion is a diagnostic result: the preprocessing misalignment swaps signal and background labels in the groomed feature space, directly motivating a self-consistent evaluation framework.

Study B provided the physically interpretable results. The ungroomed classifier achieves an in-domain AUC of 0.983, but collapses to 0.549 and 0.507 on PYTHIA and HERWIG **HAD-on** evaluation samples — a drop of ≈ 0.44 AUC points. The classifier has memorized generator-specific soft-radiation patterns present in the $\approx 3\,000$ training signal jets and fails to transfer to any other configuration of the same process. The groomed classifier, by contrast, sacrifices in-domain performance ($\text{AUC} = 0.618$) but generalizes consistently, achieving 0.644 and 0.555 on PYTHIA and HERWIG **HAD-on** respectively. **SoftDrop** grooming is therefore a necessary preprocessing step for the classifier to produce predictions that are stable across simulation configurations.

The groomed generator gap $\Delta\text{AUC} = 0.089$ (Study B) is the first physically interpretable estimate of the PYTHIA/HERWIG systematic uncertainty on classifier performance in this framework. Its magnitude is directly traceable to the residual inter-generator spread in N_{tracks} that grooming cannot suppress.

Study C confirmed the robustness of this conclusion. Replacing the standard sampler with a class-balanced weighted sampler and a per-class normalized loss moves the GRO generator gap by only -0.003 (from 0.089 to 0.086) while partially recovering the UNG generalization collapse (**HAD-on** evaluation AUC increases from 0.507–0.549 to 0.565–0.645). The persistent failure at $m_a = 4.0$ GeV — accuracy below 5% in all configurations — is a feature-space information limit: the seven track-based variables carry qualitatively similar information for jets at $m_a = 3.5$ and 4.0 GeV, and no reweighting scheme can extract discriminating power that is not present in the inputs.

5.2.3 Central Physics Result

The central result of this work is twofold. First, a systematic comparison between PYTHIA 8 and HERWIG 7 generators reveals that **SoftDrop** grooming substantially reduces the generator dependence of both the substructure observables and the downstream classifier response. Second, a regression network trained on seven jet substructure variables

achieves good performance in predicting the pseudoscalar mass $\hat{m}_a$, demonstrating that jet substructure carries meaningful information about the $a \rightarrow q\bar{q}/gg$ decay kinematics. The binary classifier, trained to separate boosted $a \rightarrow q\bar{q}/gg$ signal jets from Z +jets background using the same seven variables, achieves $\text{AUC} = 0.618\text{--}0.644$ in the groomed configuration—a modest but stable separation. This performance is achieved only when **SoftDrop** grooming is applied as a preprocessing step: the ungroomed classifier achieves higher in-domain AUC (0.983) but collapses to near-random performance on any evaluation configuration other than the training sample, exposing a severe generator dependence. The groomed generator gap of $\approx 0.087\text{--}0.089$ AUC points, stable across two regression training procedures, is a quantitative and robust estimate of the generator systematic on classifier performance, and is consistent with the $\sim 60\%$ signal yield uncertainty quoted by the ATLAS analysis. Improving the classification performance itself—for instance by incorporating additional high-level variables, pile-up mitigation, or a more expressive network architecture—is a natural next step beyond the scope of this proof-of-concept study.

The present work provides an independent, lightweight proof-of-concept that reaches the same qualitative conclusion as the updated ATLAS analysis: **SoftDrop** grooming is necessary for classifier generalization, the groomed generator gap is substantially smaller than the corresponding ungroomed collapse, and this conclusion is robust to the choice of regression training procedure.

A direct quantitative comparison with the updated ATLAS result would require applying this pipeline to fully reconstructed, pile-up-inclusive simulation at $\sqrt{s} = 13$ TeV.

5.3 Outlook

Several concrete improvements to the pipeline follow directly from the results of this work, ordered by expected impact.

Adding kinematic features at the hard splitting. The regression failure at $m_a = 4.0$ GeV is a feature-space limitation, not a training artifact: Study C demonstrated that reweighting the training procedure does not recover accuracy at the mass extremes. The most direct remedy is to add jet-level features sensitive to the hard splitting. The **SoftDrop** momentum fraction z_g and the groomed jet mass m_{gro} are the most physically motivated additions: m_{gro} is the most direct proxy for m_a accessible from jet substructure, and z_g captures the hard-splitting kinematics of the $a \rightarrow q\bar{q}$ decay. Both variables are already computed in the analysis pipeline and would require only a modification to the feature vector.

Training the classifier across mass hypotheses. The classifier was trained at $m_a = 3.0$ GeV only. Training simultaneously across all nine mass hypotheses would produce

a more general discriminant whose learned features are not specific to one kinematic point, and would make the generator gap mass-dependent rather than a single number — providing a more complete picture of the systematic uncertainty across the signal region.

Increasing the training sample size. The current $\approx 3\,000$ signal jets per mass hypothesis is a significant limiting factor at the classification stage. Increasing this to $\sim 10\,000$ jets would directly improve the classifier AUC and reduce the variance in the validation-test gap observed for the GRO classifier.

Application to fully reconstructed simulation. The most important next step from a physics perspective is the application of this pipeline to fully reconstructed simulation, including a realistic detector response, pile-up, and the full ATLAS event reconstruction at $\sqrt{s} = 13.6$ TeV. The present study provides the proof of concept and the quantitative motivation, and the full Run3 dataset collected by ATLAS in 2022-2026 offers a great opportunity for a new analysis.



Bibliography

- [1] The ATLAS Collaboration. “Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC”. In: *Physics Letters B* 716.1 (2012), pp. 1–29.
- [2] CMS Collaboration. “Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC”. In: *Physics Letters B* 716.1 (2012), pp. 30–61.
- [3] G. C. Branco et al. “Theory and phenomenology of two-Higgs-doublet models”. In: *Physics Reports* 516.1–2 (July 2012), pp. 1–102.
- [4] The ATLAS Collaboration. “Search for Higgs Boson Decays into a Z Boson and a Light Hadronically Decaying Resonance Using 13 TeV pp Collision Data from the ATLAS Detector”. In: *Physical Review Letters* 125.22 (Nov. 2020).
- [5] S. Marzani, G. Soyez, and M. Spannowsky. *Looking Inside Jets: An Introduction to Jet Substructure and Boosted-Object Phenomenology*. Vol. 958. Lecture Notes in Physics. Springer, 2019.
- [6] A. J. Larkoski, I. Moult, and B. Nachman. “Jet Substructure at the Large Hadron Collider: Experimental Review”. In: *Phys. Rept.* 841 (2020), pp. 1–63.
- [7] M. E. Peskin and D. V. Schroeder. *An Introduction to Quantum Field Theory*. Reading, MA: Addison-Wesley, 1995.
- [8] D. J. Griffiths. *Introduction to Elementary Particles*. 2nd. Wiley-VCH, 2008.
- [9] S. Weinberg. *The Quantum Theory of Fields, Volume II: Modern Applications*. Cambridge University Press, 1996.
- [10] S. Weinberg. “A Model of Leptons”. In: *Physical Review Letters* 19.21 (1967), pp. 1264–1266.
- [11] Particle Data Group. “Electroweak Model and Constraints on New Physics”. In: *Progress of Theoretical and Experimental Physics* 2020.8 (2020), p. 083C01.
- [12] J. Goldstone, A. Salam, and S. Weinberg. “Broken Symmetries”. In: *Physical Review* 127 (1962), pp. 965–970.

- [13] G. S. Guralnik, C. R. Hagen, and T. W. B. Kibble. “Global Conservation Laws and Massless Particles”. In: *Physical Review Letters* 13.20 (1964), pp. 585–587.
- [14] P. W. Higgs. “Broken Symmetries and the Masses of Gauge Bosons”. In: *Physical Review Letters* 13.16 (1964), pp. 508–509.
- [15] P. W. Higgs. “Broken Symmetries, Massless Particles and Gauge Fields”. In: *Physics Letters* 12.2 (1964), pp. 132–133.
- [16] A. Djouadi. “The Anatomy of Electro-Weak Symmetry Breaking. I: The Higgs boson in the Standard Model”. In: *Physics Reports* 457 (2008), pp. 1–216.
- [17] S. Habib et al. “ASCR/HEP Exascale Requirements Review Report”. In: (2016).
- [18] J. Gao, L. Harland-Lang, and J. Rojo. “The structure of the proton in the LHC precision era”. In: *Physics Reports* 742 (May 2018), pp. 1–121.
- [19] The ATLAS Collaboration. “Characterising the Higgs boson with ATLAS data from Run 2 of the LHC”. In: *Phys. Rep.* 11 (2024), p. 001.
- [20] The ATLAS Collaboration. “Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC”. In: *Physics Letters B* 716.1 (Sept. 2012), pp. 1–29.
- [21] CERN. “CERN Yellow Reports: Monographs, Vol. 2 (2017): Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector”. In: (2017).
- [22] The ATLAS Collaboration. “Evidence for the spin-0 nature of the Higgs boson using ATLAS data”. In: *Physics Letters B* 726.1–3 (Oct. 2013), pp. 120–144.
- [23] The ATLAS Collaboration. “Determination of spin and parity of the Higgs boson in the $WW^* \rightarrow e\nu\mu\nu$ decay channel with the ATLAS detector”. In: *European Physical Journal C* 75.5 (May 2015), pp. 1–40.
- [24] A. D. Sakharov. “Violation of CP invariance, C asymmetry, and baryon asymmetry of the universe”. In: *Phys. Usp.* 34.5 (1991), pp. 392–393.
- [25] Y. Fukuda et al. “Evidence for Oscillation of Atmospheric Neutrinos”. In: *Physical Review Letters* 81.8 (Aug. 1998), pp. 1562–1567.
- [26] S. P. Martin. “A Supersymmetry Primer”. In: *Perspectives on Supersymmetry*. World Scientific, July 1998, pp. 1–98.
- [27] J. F. Gunion et al. *The Higgs Hunter’s Guide*. Vol. 80. 2000.
- [28] A. Djouadi. “The anatomy of electroweak symmetry breaking Tome II: The Higgs bosons in the Minimal Supersymmetric Model”. In: *Physics Reports* 459.1–6 (Apr. 2008), pp. 1–241.

- [29] K. Cheung et al. “Comprehensive study of the light charged Higgs boson in the type-I two-Higgs-doublet model”. In: *Physical Review D* 105.9 (2022), p. 095044.
- [30] Wikipedia contributors. *Large Hadron Collider*. Online; accessed 23 October 2025. 2025.
- [31] The ATLAS Collaboration. *The ATLAS Experiment Implements Heterogeneous Particle Reconstruction with Intel® oneAPI Tools*. Online; accessed 23 October 2025. 2023.
- [32] G. Strong. “On the impact of selected modern deep-learning techniques to the performance and celerity of classification models in an experimental high-energy physics use case”. In: *Machine Learning: Science and Technology* 1 (Sept. 2020).
- [33] S. Mehlhase. *ATLAS Detector Schematic (with particle visualisations)*. CERN Document Server. Work based on earlier versions by R. Pitt and J. Pequenaio. <https://cds.cern.ch/record/2770815>. May 2021.
- [34] M. Carena et al. “Status of Higgs Boson Physics”. In: *Review of Particle Physics* (2023).
- [35] K. Hanagaki et al. *Experimental Techniques in Modern High-Energy Physics: A Beginner’s Guide*. 2022.
- [36] A. S. Chisholm et al. “Measuring rare and exclusive Higgs boson decays into light resonances”. In: *The European Physical Journal C* 76.9 (Sept. 2016).
- [37] S. Carrà. “Search for electroweak production of supersymmetric particles at the LHC Run 2 with the ATLAS detector”. PhD thesis. Milan, Italy: Università degli Studi di Milano, Dipartimento di Fisica, 2019.
- [38] The ATLAS Collaboration. “Search for Higgs boson decays into a Z boson and a light hadronically decaying resonance in pp collisions at 140 fb^{-1} of 13 TeV with the ATLAS detector”. In: *Physics Letters B* 868 (Sept. 2025), p. 139671.
- [39] M. Rieger. “The $t\bar{t}H$ process in the Standard Model of Particle Physics”. In: *Search for $t\bar{t}H$ Production in the $H \rightarrow b\bar{b}$ Decay Channel: Using Deep Learning Techniques with the CMS Experiment*. Cham: Springer International Publishing, 2021, pp. 5–39.
- [40] S. Catani et al. “Longitudinally invariant k_t clustering algorithms”. In: *Nucl. Phys. B* 406 (1993), p. 187.
- [41] Y. L. Dokshitzer et al. “Better jet clustering algorithms”. In: *JHEP* 08 (1997), p. 001.
- [42] M. Cacciari, G. P. Salam, and G. Soyez. “The anti- k_t jet clustering algorithm”. In: *JHEP* 04 (2008), p. 063.

- [43] G. Salam. *Basics of QCD, Lecture 2: higher orders, divergences*. Lecture notes, ICTP–SAIFR school on QCD and LHC physics, São Paulo, Brazil. July 2015.
- [44] M. Cacciari, G. P. Salam, and G. Soyez. “FastJet user manual (for version 3.0.2)”. In: *The European Physical Journal C* 72.3 (Mar. 2012).
- [45] C. F. Berger, T. Kucs, and G. Sterman. “Event shape / energy flow correlations”. In: *Phys. Rev. D* 68 (2003), p. 014012.
- [46] A. J. Larkoski, G. P. Salam, and J. Thaler. “Energy Correlation Functions for Jet Substructure”. In: *JHEP* 06 (2013), p. 108.
- [47] A. J. Larkoski, I. Mout, and B. Nachman. “New Angles on Energy Correlation Functions”. In: *JHEP* 09 (2018), p. 012.
- [48] J. Thaler and K. Van Tilburg. “Identifying Boosted Objects with N -subjettiness”. In: *JHEP* 03 (2011), p. 015.
- [49] J. Thaler and K. Van Tilburg. “Maximizing Boosted Top Identification by Minimizing N -subjettiness”. In: *JHEP* 02 (2012), p. 093.
- [50] J. Gallicchio and M. D. Schwartz. “Quark and Gluon Jet Substructure”. In: *JHEP* 04 (2013), p. 090.
- [51] A. J. Larkoski et al. “Soft Drop”. In: *JHEP* 05 (2014), p. 146.
- [52] M. Dasgupta et al. “Towards an understanding of jet substructure”. In: *JHEP* 09 (2013), p. 029.
- [53] A. Paszke et al. “PyTorch: An Imperative Style, High-Performance Deep Learning Library”. In: *Advances in Neural Information Processing Systems* 32 (2019).
- [54] P. J. Huber. “Robust Estimation of a Location Parameter”. In: *The Annals of Mathematical Statistics* 35.1 (1964), pp. 73–101.
- [55] D. P. Kingma and J. Ba. “Adam: A Method for Stochastic Optimization”. In: (2014).
- [56] T. Akiba et al. “Optuna: A Next-generation Hyperparameter Optimization Framework”. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (2019), pp. 2623–2631.
- [57] J. Alwall et al. “The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton showers”. In: *JHEP* 07 (2014), p. 079.
- [58] J. Alwall et al. “A standard format for Les Houches event files”. In: *Comput. Phys. Commun.* 176 (2007), pp. 300–304.

- [59] A. Buckley et al. “General-purpose event generators for LHC physics”. In: *Phys. Rept.* 504 (2011), pp. 145–233.
- [60] C. Bierlich et al. “A comprehensive guide to the physics and usage of PYTHIA 8.3”. In: *SciPost Phys. Codebases* 8 (2022).
- [61] J. Bellm et al. “Herwig 7.0 / Herwig++ 3.0 release note”. In: *Eur. Phys. J. C* 76.4 (2016), p. 196.
- [62] T. Sjöstrand et al. “An introduction to PYTHIA 8.2”. In: *Computer Physics Communications* 191 (2015), pp. 159–177.
- [63] D. J. Lange. “The EvtGen particle decay simulation package”. In: *Nuclear Instruments and Methods in Physics Research Section A* 462.1 (2001), pp. 152–155.
- [64] J. Butterworth et al. “PDF4LHC recommendations for LHC Run II”. In: *Journal of Physics G: Nuclear and Particle Physics* 43.2 (Jan. 2016), p. 023001.
- [65] R. D. Ball et al. “Parton distributions with LHC data”. In: *Nuclear Physics B* 867.2 (2013), pp. 244–289.
- [66] J. Bellm et al. “HERWIG 7.0/HERWIG++ 3.0 release note”. In: *The European Physical Journal C* 76.4 (2016), p. 196.
- [67] S. Alioli et al. “A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX”. In: *Journal of High Energy Physics* 2010.6 (2010), p. 043.
- [68] E. Bothmann et al. “Event generation with Sherpa 2.2”. In: *SciPost Phys.* 7 (2019), p. 034.
- [69] A. J. Larkoski et al. “Soft Drop”. In: *Journal of High Energy Physics* 2014.5 (2014), p. 146.
- [70] J. de Favereau et al. “DELPHES 3: a modular framework for fast simulation of a generic detector”. In: *Journal of High Energy Physics* 2014.2 (2014), p. 057.

*“Il sistema tolemaico è un’eccellente
approssimazione
del sentimento provato nel lasciarsi
indietro i paradigmi della scienza imparati,
solo per portarne avanti altri.”*
