

**PARAMETER INFERENCE AT THE LARGE HADRON
COLLIDER USING NEURAL LIKELIHOOD RATIOS,
AND A MEASUREMENT OF THE HIGGS BOSON
DECAY WIDTH USING THE ATLAS EXPERIMENT**

A Dissertation Presented

by

JAY AJITBHAI SANDESARA

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

February 2025

Department of Physics

© Copyright by Jay Ajitbhai Sandesara 2025
All Rights Reserved

**PARAMETER INFERENCE AT THE LARGE HADRON
COLLIDER USING NEURAL LIKELIHOOD RATIOS,
AND A MEASUREMENT OF THE HIGGS BOSON
DECAY WIDTH USING THE ATLAS EXPERIMENT**

A Dissertation Presented

by

JAY AJITBHAI SANDESARA

Approved as to style and content by:

Rafael Coelho Lopes de Sá, Chair

Stephane Willocq, Member

Benjamin Heidenreich, Member

Verena Martinez Outschoorn, Member

David Rousseau, Member

Anthony Dinsmore, Department Head
Department of Physics

*The same thrill, the same awe and mystery, comes
again and again when we look at any question deeply
enough.*

*With more knowledge comes a deeper, more wonderful
mystery, luring one on to penetrate deeper still.*

*Never concerned that the answer may prove
disappointing, with pleasure and confidence we turn
over each new stone to find unimagined strangeness
leading on to more wonderful questions and mysteries -
certainly a grand adventure!*

- Richard Feynman

Acknowledgments

First and foremost, I would like to express my gratitude to my advisor, Rafael Coelho Lopes de Sá, for his outstanding mentorship throughout my PhD journey. His foresight, dedication, and guidance shaped the development of this work. I am deeply thankful for our many insightful discussions, that helped shape various aspects of the new framework presented in this thesis.

I am also grateful to R.D. Schaffer, who has been working with me nearly full-time since the last two years, and without whom the analysis would not have been completed in the timescale that it did. He stepped in at key moments and took on the work needed to complete the measurement, contributing to key software developments for the setup.

The work has also been greatly influenced and supported by David Rousseau and his team including Aishik Ghosh and Arnaud Maury. I am very thankful for the opportunity of spending a year at IJCLab while being sponsored by the Chateaubriand Fellowship; this time was the founding of this collaborative effort and I have benefited greatly from David's expertise and support since. I am also very grateful for the dedicated efforts on building Neyman Construction tools by Arnaud, and for all the insights and help shared by Aishik throughout the years. During this time I also had the opportunity to collaborate with and benefit from the expertise of Gilles Louppe, who has shared key insights throughout the years and helped develop a robust and practical analysis framework.

I am grateful to Verena Martinez Outschoorn, Stephane Willocq and Carlo Dallapiccola, who have been very supportive throughout the PhD program. I particularly thank Verena for introducing me to the rest of the ATLAS-Google team, particularly to Fernando Barreiro Megino. I cannot thank Fernando enough for his dedicated efforts, responsiveness and support in making the Google cloud-based infrastructure usable for my efforts. The rest of the ATLAS-Google team includes, but is not limited to, Johannes Elmsheuser, Alexei Klimentov and Kaushik De, all of whom I am thankful to.

I would also like to express my gratitude to several people within the collaboration and their key support in pushing the analysis through the various approval stages within the collaboration - Fabio Cerutti, Kyle Cranmer, Lukas Heinrich, Michael Kagan, Wouter Verkerke, Nicolas Berger, Paolo Francavilla, Monica Dunford, Zach Marshall, Walter Hopkins, Aurelio Juste Rozas, and many other ATLAS colleagues. I am thankful to Kyle Cranmer for providing key insights into the theoretical aspects of the new framework and Matthew Feickert, for his help and advice on certain computational aspects of the analysis. I was also lucky to have Max Goblirsch-Kolb and Nora Pettersson as my advisors during the ATLAS qualification task. I am grateful for their depth of knowledge on tracking, as well as all their help and support. Conversations with Samyukta Krishnamurthy and Michiel Jan Veen on the off-shell Higgs boson analysis proved invaluable in building parts of the new framework.

Finally, none of this work would have been possible without the multi-pronged efforts of the entire ATLAS collaboration and CERN in general. I am thankful to each of the current and past collaborators who have contributed in small and big ways to the successful working of the various aspects of the ATLAS experiment.

Abstract

PARAMETER INFERENCE AT THE LARGE HADRON COLLIDER USING NEURAL LIKELIHOOD RATIOS, AND A MEASUREMENT OF THE HIGGS BOSON DECAY WIDTH USING THE ATLAS EXPERIMENT

FEBRUARY 2025

JAY AJITBHAI SANDESARA

B.Sc., GUJARAT UNIVERSITY

M.Sc., DURHAM UNIVERSITY

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Prof. Rafael Coelho Lopes de Sá

A new statistical technique is developed for physics parameter estimation at the Large Hadron Collider (LHC) that uses modern deep-learning tools to realize a more general and fundamental approach to data analysis compared to the ad-hoc techniques commonly used. Coming under the general umbrella of Neural Simulation-Based Inference (NSBI) techniques, the new workflow uses a large number of Neural Networks (NNs) to directly learn event-by-event likelihood ratios and thus handles high-dimensional parameter estimation without the need to bin data into low-dimensional summary histograms. Novel techniques were developed for parameterizing the likelihood ratios as a function of a large number of parameters common in LHC experiments, along with the development of modern computational workflows that make it possible to apply NSBI to a full-scale ATLAS experiment analysis.

A measurement of the Higgs boson in the off-shell phase space is then performed using the new workflow, in the $H^* \rightarrow ZZ \rightarrow 4\ell$ channel. The evidence sensitivity is increased by a factor of 3.1 using the new technique in the $H^* \rightarrow ZZ \rightarrow 4\ell$ channel, compared to the previous measurement published by ATLAS using the same Run-2 data with more standard techniques. A combination of this channel with the off-shell $H^* \rightarrow ZZ \rightarrow 2\ell 2\nu$ channel is done finding evidence for the off-shell Higgs boson with

3.7σ confidence level, superseding the last measurement. The off-shell measurement is then combined with the on-shell measurement for an indirect measurement of the Higgs boson decay width under a few assumptions. The observed (expected) value of the Higgs boson width at 68% CL is $4.3^{+2.7}_{-1.9}$ ($4.1^{+3.5}_{-3.4}$) MeV.

This new method is promising for a wide range of measurements at the LHC, where no single observable may be optimal to scan over the entire theoretical phase space under consideration, or where binning data into histograms could result in a loss of sensitivity. It also offers easy re-interpretability and broader use.

Contents

Acknowledgments	i
Abstract	iii
Introduction	1
1 The ATLAS Experiment	4
1.1 The ATLAS Detector	4
1.1.1 The Coordinate System	6
1.2 The Inner Detector	7
1.2.1 Detector Layout and Technologies	8
1.3 The Calorimeter System	9
1.4 The Muon Spectrometer	9
1.5 Trigger System	11
1.6 Event Reconstruction in the ID	11
1.6.1 Track Reconstruction	13
1.6.2 Vertex Reconstruction	16
1.6.3 Data Post-Processing	16
1.7 ID Tracking Performance Validation	16
1.8 Physics Objects and Reconstructed Data	18
2 Neural Simulation-Based Inference	23
2.1 Review of the Statistical Framework	23
2.1.1 Modelling in the presence of uncertainties	24
2.1.2 Profile Negative Log-Likelihood Test Statistic	25
2.2 Building the Probability Model	26
2.2.1 Measurements within the Fiducial volume	27
2.3 Standard Inference Strategy	28
2.3.1 Summary Observables	28
2.3.2 Limitations of Standard Techniques	30
2.4 New Proposal: Neural Simulation-Based Inference	32
2.4.1 Density Ratios	33
2.4.2 Biases in modelling and estimators	34

2.5	Special Cases - Analytical Parameter Dependence	36
2.5.1	Search-Oriented Mixture Models	37
2.5.2	Signal Strength Measurements	38
2.5.3	Parameterized Optimal Observables	40
3	Uncertainties in the Model	42
3.1	Uncertainty modelling in NSBI	43
3.1.1	Building the α -parametrized density ratios	44
3.1.2	Amortized inference	45
3.2	Calculation of uncertainties and NP impacts	46
3.3	Neyman Construction	47
3.3.1	Poisson Bootstrap technique	48
4	Theory and the Experimental Model	51
4.1	The Higgs boson at the LHC	53
4.1.1	Higgs boson production and decays	54
4.1.2	Higgs boson interactions	55
4.2	The off-shell Higgs boson	56
4.3	The Experimental Model of off-shell Higgs boson	60
4.4	Application of the Search-Oriented Mixture Model	65
4.5	Previous measurement	66
4.6	Systematic Uncertainties in the Off-shell Higgs boson model	67
4.6.1	The full model	69
5	Simulations and Event Selections	70
5.1	Monte Carlo Simulations	70
5.2	Object Reconstruction and Event Selection	72
5.2.1	Object Reconstruction	72
5.2.2	Event Selection	72
5.3	Final State Observables	73
5.4	Signal and control regions	76
6	Neural Likelihood Ratio Estimation	80
6.1	Training of the Neural Networks	80
6.2	Training Ensembles with Cross-validation	83
6.2.1	Training Workflow	83
6.3	Quality of the Density Ratios	85
6.4	Estimating the systematic variations $g_c(\mathbf{x}, \alpha^\pm)$	99
6.5	Using NN Ensembles to estimate MC Statistical Uncertainty	100
6.5.1	Monte Carlo Uncertainty Propagation	104

7	Results	106
7.1	Results in the $ZZ \rightarrow 4\ell$ decay channel	106
7.1.1	Test statistic scans	107
7.1.2	Neyman Construction	108
7.1.3	Data-MC checks - optimal observables	109
7.1.4	Impacts from systematic uncertainties	109
7.2	Combination with the analysis in the $2\ell 2\nu$ decay channel	112
7.3	Reinterpretations in the κ -framework	112
7.4	Combination with on-shell 4ℓ analysis and Higgs boson width interpretation	114
8	Computational Challenges and Alternatives	118
8.1	Computational challenges in NN training	118
8.2	Computational challenges in statistical inference	118
8.3	An Alternative: Using parameterized observables	121
8.3.1	Understanding the Improvements in the New Analysis	121
8.3.2	Application to Real Data	124
9	Conclusion and Outlook	126
A	Histogram Result Comparisons	128
B	Neyman Construction Closures	132
	Bibliography	144

List of Figures

1	The invariant mass m_{ZZ} distribution of the $H \rightarrow ZZ \rightarrow 4\ell$ decay channel shows a large non-resonant Higgs boson signal $gg \rightarrow H^* \rightarrow ZZ \rightarrow 4\ell$ (S) off its mass shell of 125 GeV, along with a large quantum interference (S+I) with the background $gg \rightarrow ZZ \rightarrow 4\ell$ (gg bkg.) in this phase space. The measurement of the off-shell Higgs boson production is measuring the S+I distribution shown in the figure.	2
1.1	The LHC, shown as the dark grey ring, is the final and largest component in a network of particle accelerators. Smaller accelerators in the chain gradually increase particle energy before feeding them into the LHC and supplying beams to various other experiments designed to explore the fundamental nature of the Universe. Figure taken from Ref. [10]	5
1.2	The layout of the full ATLAS detector at the Large Hadron Collider. The Pixel, Semi-conductor and Transition radiation tracker constitute the Inner Detector, which is immersed in a 2T magnetic field parallel to the beam axis produced by the Solenoid magnet. The Inner Detector is followed by electromagnetic and hadron calorimeters, and a Muon spectrometer chambers incorporating three large superconducting Toroid magnets. . .	6
1.3	The ATLAS detector coordinate system. The definitions of the various quantities are given in Sec. 1.1.1. Figure source code taken and modified from Ref. [11].	7
1.4	The ATLAS Inner Detector system. Figure taken from Ref. [12].	8
1.5	A simplified layout of the various sensors and structural elements constituting the ATLAS Inner Detector. IBL stands for the Insertable B-layer introduced before the start of Run-2, followed by the Pixel detectors, the Semi-conductor Trackers (SCT) and the Transition Radiation Trackers (TRT). The red line shows a charged particle track with 10 GeV transverse momentum p_T and $\eta = 0.3$ traversing the Inner Detector. Figure taken from Ref. [14].	10
1.6	The schematic of the ATLAS Inner Detector (ID) system, in the $r - z$ plane. Figure taken from Ref. [14].	11
1.7	A simulation of the various particles interacting with the different sub-detector systems in the ATLAS detector. Figure taken from Ref. [18].	12
1.8	General structure of an Athena application. More details in Ref. [20], from which this figure is taken.	13

1.9	A flowchart of the various stages of ID track reconstruction. A first inside-out primary pass (top) is performed followed by an outside-in secondary backtracking pass (bottom). Figure taken from Ref. [20].	14
1.10	The perigee representation $(d_0, z_0, \phi, \theta, q/p)$ used to parameterize ATLAS ID tracks. Figure taken from Ref. [22].	15
1.11	An example use of the ntuple created from the <code>writeSeedNtuple</code> tool. The figures show the $r - z$ correlation plots of the hits corresponding to the second spacepoint in a seed triplet inside the Pixel layer of the ID. The right figure shows the hits corresponding to spacepoints of the seeds that go on to form tracks, with the left figure corresponding to seeds that do not form tracks. The 3 layers correspond to the L0, L1 and L2 of the Pixel sub-detector. The IBL hits cannot correspond to the second spacepoint and are thus invisible in this plot. Note the increased density of SPs in the 2nd layer of the barrel region for seeds that go on to form track candidates.	18
1.12	An example use of the validation ntuple made using the <code>writeSeedNtuple</code> tool. On the left, is the $x - y$ correlation plot in the barrel layer of the ID, using a stable software release. On the right, is the same plot focused on the Pixel layers, where a problematic configuration in a modified Athena release caused issues with the outermost L2 pixel layer having invisible alternating sensors. The latter resulted in strange behavior in downstream validations, shown in Fig. 1.13, which was further investigated and then resolved using this ntuple tool, showing its versatility in both development and as a debugging tool.	19
1.13	An example of a discrepancy between a development release (20.0.16) of Athena and the Master branch. The left plot shows the number of Pixel hits, and the right plot shows the number of holes as a function of pseudo-rapidity η . The bottom plots are ratio plots between the development release and the Master branch. On average, one pixel hit per track seems to be missing (plot on the left) in the barrel, and instead a hole appears (plot on the right). To investigate the origins of this discrepancy, the new validation <code>writeSeedNtuple</code> tool was used, as is shown in Figs. 1.14 and 1.12. This is just an example issue to showcase the versatility of the tool in general.	19
1.14	An example use of the validation ntuple made using the <code>writeSeedNtuple</code> tool. The $z - \phi$ correlation plots of the hits in the L2 sub-layer (figure on the right) seems to show missing pixels in alternate locations of the L2, as is also clear from the right plot in Fig. 1.12. For comparison the left plot shows the same $z - \phi$ with the third SP in a seed located in the L1 layer, with fewer missing pixels. The black shadows in the figure on the left can be further explained using the new tool, as shown in Fig. 1.15.	20

- 1.15 An example use of the validation ntuple made using the `writeSeedNtuple` tool. A problematic software release was diagnosed by investigating the peculiarities in the $r-\phi$ correlation plot at the SP formation level. The figure on the left corresponds to the third SP in L1, forcing the seed to have IBL+L0+L1 topology. These selections are also done using the information stored in the ntuple written by the tool. The black shadows (blue circles) in the left plot correspond to a lack of hits and potentially indicate dead pixels on the layers below. The figure on the right, with SP hits in the L0 layer was made to confirm this (blue circles surrounding missing pixels) and subsequently used to debug the software release. 20
- 1.16 This table is used to dump all the relevant logistics of the track finding stage. It contains useful statistics extracted mainly from the SiTrackFinder and SiCombinatorialTrackFinder tools. The PPP, PPS, PSS and SSS stand for Pixel-Pixel-Pixel seed, Pixel-Pixel-SCT seed, and so on. DE stands for detector elements and the term road stands for the 3D roads built during the track reconstruction stage. The term *brem model* stands for bremsstrahlung recovery model used in electron reconstruction. The Track/Used row refers to the fraction of seeds that result in track formation compared to the total number of seeds, as a function of the pseudo-rapidity η 21
- 2.1 From left to right: a) An example Feynman diagram of parton-parton hard scattering of interest. One of the many diagrams typically simulated using perturbative QCD, and summed over to produce final cross-sections associated with a given final state b) Sketch of the full hadron-hadron collision as simulated by a Monte-Carlo event generator. The red blob in the center represents the hard collision, generated using the first Feynman diagram for example, surrounded by a tree-like structure representing Bremsstrahlung as simulated by parton showers. The purple blob indicates a secondary hard scattering event. Parton-to-hadron transitions are represented by light green blobs, dark green blobs indicate hadron decays, while yellow lines signal soft photon radiation. Figure taken from Ref. [29]. c) A simulation of the various particles interacting with the different sub-detector systems in the ATLAS detector. Figure taken from Ref. [18]. d) An event display of the final state, also showing reconstructed objects like muons and jets. . . . 27
- 2.2 (a) The distribution of the four-lepton invariant mass, $m_{4\ell}$, used to identify the Higgs boson resonance peak during the first discovery announced in 2012. (b) An example of the standard inference strategy - building observables that optimally discriminate signal from background. 29

- 2.3 An example of a standard inference workflow used for measurements at the LHC. Using perturbative methods, a physics process of interest is simulated using Monte Carlo techniques, and then undergo parton showering, hadronization and detector reconstruction to produce the final detector-level output $\mathbf{x} \sim p(\mathbf{x}|\boldsymbol{\mu})$. Same technique is used to build the full model as a function of some parameter of interest $\boldsymbol{\mu}$. A powerful observable discriminating signal from background is then binned and a Poisson likelihood fit is performed to infer the parameter of interest μ , giving us insight in the physics process of interest. The particular observable and inference shown in the figure is from the measurement of the VH production in Ref. [34]. 31
- 2.4 Left: The probability density function estimated using a combination of analytical approximations and NN output (red line) is shown in comparison with the $m_{4\ell}$ distribution in signal MC simulation (black points) for $m_H = 125$ GeV. Both the probability density function and the MC simulation are normalised to unity. Right: The test statistic values, $-2\ln(\lambda)$, calculated using the unbinned fit with an approximate, but parameterized, probability distribution function (shown on left for $m_H = 125$ GeV). The vertical (red) line on the right plot indicates the combined Run 2 result, and the grey band its total uncertainty. 34
- 2.5 A cartoon representation of a typical NSBI workflow, where NNs are trained to estimate precise and accurate per-event parameterized likelihood ratios, then used for direct statistical inference using the profile negative log-likelihood ratio test statistic defined in Eq. 2.11. This is a generalization of the workflow depicted in Fig. 2.3 using deep NNs acting as surrogates for the likelihood ratios using the likelihood ratio technique shown in Fig. 6.1. The results in this cartoon will be discussed in detail in Chap. 7. 35
- 3.1 The Neyman construction for the off-shell Higgs boson analysis, defined in Chap. 4. Sampled pseudo-experiments are used to build the t_μ distribution, which is then integrated to find the boundaries corresponding to 68.24% and 95.45% pseudo-experiments, defining the 1σ and 2σ confidence belts as a function of $\mu = 0.5, 1.0, 2.5, 3.5$ (top left, top right, bottom left, bottom right). This is repeated for the full scan range of $\mu_{\text{off-shell}}$ as to build confidence belts and find the $\mu_{\text{off-shell}}$ intersection points corresponding to 1σ and 2σ uncertainties on the POI $\mu_{\text{off-shell}}$ as shown in Fig. 7.1. In an asymptotic case, the distributions will be χ^2 -distributed and the vertical lines corresponding to 68.24% and 95.45% pseudo-experiments would be located at $t_\mu \approx 1$ and $t_\mu \approx 4$ respectively. The degenerate solutions of the off-shell Higgs boson probability model breaks the asymptotic approximation. 50

4.1	The Mexican-hat shaped potential of the Higgs boson, described by Eq. 4.3. The Universe starts out with the unstable minimum $V(\phi) = 0$, corresponding to $\text{Re}(\phi) = \text{Im}(\phi) = 0$. About 10^{-12} seconds after the big-bang, the Higgs field acquires a non-zero vacuum expectation value, that does not respect the original $\text{SU}(2)_L \times \text{U}(1)_Y$ symmetry, but acquires a new $\text{U}(1)_{\text{EM}}$ symmetry associated with Maxwell's electromagnetism in the process. Massive vector bosons and fermions are predicted in this broken symmetry Universe through interactions with the Higgs field which now has a non-zero vacuum state. The transverse oscillations of the Higgs field along the radial direction is associated with the new scalar boson called the Higgs boson.	53
4.2	Higgs width prediction in the Standard Model as a function of its mass. Around the measured mass of 125 GeV, the Higgs width is predicted to be ~ 4.1 MeV. Figure taken from Ref. [63].	55
4.3	The relativistic Breit-Wigner distribution of the Higgs boson resonance with a width (Γ_H) of 4.1 MeV. The small value of the width makes it impossible to directly measure using the LHC experiments. Figure courtesy of Michiel Jan Veen.	55
4.4	The various production and decay channels of the Higgs boson at the LHC from Ref. [64]. The Higgs boson is produced via gluon–gluon fusion (a), vector boson fusion (VBF; b) and associated production with vector bosons (c), top or b quark pairs (d), or a single top quark (e). f–i, The Higgs boson decays into a pair of vector bosons (f), a pair of photons or a Z boson and a photon (g), a pair of quarks (h), and a pair of charged leptons (i). Loop-induced Higgs boson interactions with gluons or photons are shown in blue, and processes involving couplings to W or Z bosons in green, to quarks in orange, and to leptons in red. Two different shades of green (orange) are used to separate the VBF and VH (and tH) production processes.	56
4.5	The pie chart showing the probabilities of Higgs boson decays to various possible final states. Figure from Ref. [65].	56
4.6	SM predictions of the Higgs boson decay branching ratios to various particle pairs, as a function of the Higgs boson mass. Figure from Ref. [63].	56
4.7	The Higgs boson coupling strengths to the various fermions and scalar bosons, shown using modifiers measured experimentally. The vertical bar on each point denotes the 68% confidence interval. The lower panel shows the values of the coupling strength modifiers. Data are from ATLAS Run-2. Figure taken from Ref. [66].	57
4.8	The m_{ZZ} distribution comparison between a Breit-Weigner lineshape centered at $m_H = 125$ GeV and the actual lineshape due to threshold effects at $m_{ZZ} \gtrsim 180$ GeV and $m_{ZZ} \gtrsim 350$ GeV. Here the H stands for the Higgs signal process and <i>cont</i> stands for the interfering background shown in Fig. 4.9. ZWA stands for the zero-width approximation that assumes a Breit-Weigner distribution with a narrow decay width. Figure taken from Ref. [68].	59

4.9	Illustrative leading-order Feynman diagrams for $gg \rightarrow ZZ$ production. The diagrams indicate the effective couplings g_g and g_V . Diagram a corresponds to the signal component and diagram b corresponds to the background component. A large destructive interference between the two components is present in the off-shell regime.	60
4.10	Illustrative leading-order Feynman diagrams for the electroweak $qq \rightarrow ZZ+2j$ processes. The diagrams indicate the effective couplings g_V used to define the off-shell Higgs boson production signal strengths. Diagrams a, b, and c correspond to the vector boson fusion, t -channel and ZH signal components, respectively. In diagram c, one of the Z bosons decays into a $q\bar{q}$ pair. Diagram d corresponds to the electroweak background component. A large destructive interference is present in the off-shell regime between the vector boson fusion signal and the background components.	61
4.11	Illustrative leading-order Feynman diagrams for non-interfering background processes. Diagram a shows the leading $q\bar{q}ZZ$ process and diagram b shows the top-quark induced sub-leading VVV process. The leptonic decays of the Z and W bosons are not shown.	62
4.12	The NN discriminant used as a summary statistic in the first Run-2 analysis of the off-shell Higgs boson by the ATLAS experiment. The parameter-independent nature of this observable makes it sub-optimal for parameter values that favor a dominant interference effect over signal, motivating the need for a new optimal analysis technique. Figure from Ref. [3].	67
5.1	Representation of the angular observables used to describe the probability density ratio of each event. The three-momenta of the fermions ($\mathbf{q}_{11}^*$, $\mathbf{q}_{21}^*$) and anti-fermions ($\mathbf{q}_{12}^*$, $\mathbf{q}_{22}^*$) are shown in their parent rest-frames, and the three-momenta of the vector bosons ($\mathbf{q}_1$, $\mathbf{q}_2$) are shown in the quadruplet rest frame.	74
5.2	Comparison between observed and expected distribution of two observables used to describe events: a the four-lepton invariant mass and the b invariant mass of the two leading jets showing good agreement for two of the observables used in this analysis. The expected distributions for $q\bar{q} \rightarrow ZZ$, ggF SBI ($gg \rightarrow (H^* \rightarrow)ZZ$), EW SBI ₁ ($qq \rightarrow (H^* \rightarrow)ZZ + 2j$) and other backgrounds are shown as stacked histograms, and the expected signal (interference) is shown as a red solid (blue dashed) line. The background is estimated under the SM hypothesis (post-fit, $\mu_{\text{off-shell}} = 1$). The lower panels show the ratio of data to expectation. The hashed band shows the total systematic uncertainty in the expected distribution. The last bin contains overflow events.	75
5.3	Distribution of the pre-selection NN score used to select events with $p_{\text{ref}}(\mathbf{x}) \gg 0$, defining the signal region (SR) and the control region (CR) in the analysis. The yellow dashed line shows the division between the CR, where a standard binned Poisson likelihood fit is done, and the SR where an unbinned likelihood fit using NSBI techniques is done. The discriminant score is the value of $D_{\text{pre}}(\mathbf{x})$ as defined in Eq. 5.3.	76

5.4	a Comparison between observed and expected distribution of the preselection discriminant $D_{\text{pre}}(\mathbf{x})$ showing a good description of the acceptance times efficiency in each region. b Comparison between observed and expected distribution of n_{jets} showing a good description of the normalization obtained in each bin with the data-driven background normalization parameters. The expected distributions for $q\bar{q} \rightarrow ZZ$, ggF SBI ($gg \rightarrow (H^* \rightarrow)ZZ$), EW SBI ₁ ($qq \rightarrow (H^* \rightarrow)ZZ + 2j$) and other backgrounds are shown as stacked histograms, and the expected signal (interference) is shown as a solid red (dashed blue) line. The background is estimated under the SM hypothesis (post-fit, $\mu_{\text{off-shell}} = 1$). The lower panels show the ratio of data to expectation. The hashed band shows the total systematic uncertainty in the expected distribution. The last bin of panel b contains overflow events.	78
6.1	A representation of the <i>likelihood-ratio trick</i> or the CARL technique - the optimal minimum $\hat{s}(\mathbf{x})$ of the Binary Cross-Entropy loss function can have a probabilistic interpretations which can be used to regress the probability density ratio between two hypothesis. <Figure needs update>	81
6.2	Four of the simpler diagnostic plots corresponding to a sample NN training. The top left plot shows the increase in accuracy per epoch of training, which eventually converges to a value and stays constant. Top right shows a similar evolution but for the value of binary cross-entropy loss over the epochs. The bottom two plots are used to test overfitting, by comparing the NN prediction distribution between training and testing datasets for each hypothesis.	82
6.3	A cartoon representation of the ensembling technique used in the analysis along with k-fold cross-validation. Each blob in the box on the left represents the 80% of the training data sampled for each NN. The total number of sampled blobs, here only a few shown for representation, are equal to N_{ens} and corresponds to the total number of NNs trained using the train-test resampling technique (see text). The mean of the predictions from these N_{ens} NNs are used as final predictions in the profile likelihood fit. The ensemble training is done for each channel $c = \text{ggF SBI, ggF B, ggF S, etc.}$ independently. . . .	84
6.4	Diagnostic check defined in Eq. 6.6 of the relative probability predicted by the NNs with the relative probability using binned MC events. This plot uses the prediction from the enesmbles of NNs (see Section 6.2). The binning is chosen as narrow as allowed by the size of simulated samples in order to highlight the well-calibrated predictions for very rare signal events on the far left ($\hat{s}_{\text{pred}} \approx 0$) and far right ($\hat{s}_{\text{pred}} \approx 1$).	86
6.5	Diagnostic check defined in Eq. 6.7 of the log of probability density ratio predicted by the NNs with the log of density ratio using binned MC events. This plot uses the prediction from the enesmbles of NNs (see Section 6.2). The binning is chosen as narrow as allowed by the size of simulated samples in order to highlight the well-calibrated predictions for very rare signal events.	86

- 6.6 Comparison between the NN-based and histogram-based estimate of the density ratio $p(\mathbf{x}|\mu_{\text{off-shell}})/p_{\text{ref}}(\mathbf{x})$ for a $\mu_{\text{off-shell}} = 0.3$ and b $\mu_{\text{off-shell}} = 1.7$. The comparison is done as a function of $\log[p(\mathbf{x}|\mu_{\text{off-shell}})/p_{\text{ref}}(\mathbf{x})]$ to separate the comparison for events that are very signal-like and reference-like. The lower panels show the pull, defined as the difference between the NN and MC-based estimate divided by the statistical uncertainty due to the finite number of MC events. The error bars indicate the uncertainty due to the finite number of simulated events in the histogram-based estimate of the density ratio. 87
- 6.7 Reweighting multi-dimensional diagnostic: a NN classifier is trained to try and distinguish the original sample and the reference sample reweighted to the original using Eq. 6.8. The AUC value of 0.5 corresponds to an indistinguishable sample. This provides a multi-dimensional test of the quality of predicted density ratios. Note how training a classifier NN to distinguish reference sample from the original without reweighting results in a large AUC. 88
- 6.8 Reweighting diagnostics: The dijet observable $\Delta\eta_{jj}$. Testing the quality of the predicted density ratio by reweighting 1D observables with the ratios, using Eq. 6.8. The blue curve shows the reference sample and the orange one the sample after reweighting to match the target, shown in green: the ratio plots show the ratio of the reweighted distribution bin to the target distribution bin. All input variables are tested for all NN, showing excellent agreement. 89
- 6.9 Reweighting diagnostics: High-level ME observable calculated using MCFM and not used in the training of the NNs. Testing the quality of the predicted density ratio by reweighting 1D observables with the ratios, using Eq. 6.8. The blue curve shows the reference sample and the orange one the sample after reweighting to match the target, shown in green: the ratio plots show the ratio of the reweighted distribution bin to the target distribution bin. All input variables are tested for all NN, showing excellent agreement. 89
- 6.10 Comparison between the $D_{\text{pre}}(\mathbf{x})$ normalized distribution obtained using an Asimov sample with a $\mu_{\text{off-shell}} = 0.3$ and b $\mu_{\text{off-shell}} = 1.7$, and the same distributions obtained through reweighting of a SM ($\mu_{\text{off-shell}} = 1.0$) Asimov sample. The ratio plots provide a comparison between the two estimates (middle panel) and between the distributions at the given value of $\mu_{\text{off-shell}}$ and the SM expectation (bottom panel). 90
- 6.11 Comparison between the $D_{\text{pre}}(\mathbf{x})$ normalized distribution obtained using an Asimov sample with a $\mu_{\text{off-shell}} = 0.3$ and b $\mu_{\text{off-shell}} = 1.7$, and the same distributions obtained through reweighting of a SM ($\mu_{\text{off-shell}} = 1.0$) Asimov sample. The ratio plots provide a comparison between the two estimates (middle panel) and between the distributions at the given value of $\mu_{\text{off-shell}}$ and the SM expectation (bottom panel). 91

- 6.15 Comparison between data and expectation in the SR ($D_{\text{pre}}(\mathbf{x}) > -0.85$) of the NN scores of the all major samples used to describe the probability density ratio of the reference process and a ggF SBI, b ggF S, c ggF B, d EW SBI₁, e EW SBI₁₀, f EW B, g $q\bar{q} \rightarrow ZZ$, and h VVV . The comparisons show a good modeling of the probability density ratios throughout the SR. The lower panels show the ratio of data to expectation. The expected SM distributions are shown as stacked histograms and the expected signal (interference) is shown as a solid red (dashed blue) line. The background is estimated under the SM hypothesis (post-fit, $\mu_{\text{off-shell}} = 1$). The hatched band shows the total systematic uncertainty in the expected distribution. The last bin contains overflow events. 95
- 6.17 Comparison between data and expectation in the SR ($D_{\text{pre}}(\mathbf{x}) > -0.85$) of the pre-selection NN. In a, events are weighted by $p(\mathbf{x}|\mu = 0.3)/p(\mathbf{x}|\mu = 1)$ to highlight the low- $\mu_{\text{off-shell}}$ region dominated by the interference component. In b, events are weighted by $p(\mathbf{x}|\mu = 1.7)/p(\mathbf{x}|\mu = 1)$ to highlight the high- $\mu_{\text{off-shell}}$ region dominated by the signal component. The solid red line shows the sum of the signal and interference components estimated with the $\mu_{\text{off-shell}}$ hypothesis used to reweight the distributions. The lower panels show the ratio of the data to expectation. The background is estimated under the SM hypothesis (post-fit, $\mu_{\text{off-shell}} = 1$). The hatched area corresponds to the total systematic uncertainty in the expected distributions. The first and last bins contains overflow events. 97
- 6.19 Maximum likelihood estimator $\hat{\mu}_{\text{off-shell}}$ when fitting $\mu_{\text{off-shell}}$ using MC samples with different true values $\mu_{\text{off-shell}}$. The error bar shows the estimate of the uncertainty from the limited number of MC events in the Asimov dataset, which is introduced as a spurious signal uncertainty as described in Sec. 6.5. 99
- 6.20 Example diagnostic plots for systematic uncertainty weight estimation, showing unbiased modelling of the systematic uncertainties. The first ratio plot shows the ratio between the nominal sample reweighted to the systematic variation and original systematic variation. The second ratio plot shows the ratio of systematic variation distribution (both reweighted and original - green and orange curves respectively) to the nominal (blue curve). 101
- 6.21 Ensemble NN score prediction distributions for random example events with NN trained for the $p_{\text{B}}^{\text{ggF}}/p_{\text{ref}}$ density ratio. Larger spreads indicate larger variances on individual NN predictions from the MC statistical uncertainty. 102
- 6.22 Building template ensembles from the trained NN pool to estimate the variance on the $\hat{\mu}$. Each blob in the box on the left represents the 80% of the training data sampled for each NN. These NNs have been trained already for the nominal fit, and it is the predictions from these trained NNs that are sampled N_{ens} times with replacement to build the template model. This is equivalent to sampling and ensembling independently in the limit of large enough N_{ens} . The mean of the predictions from these N_{ens} samples of each template are used to determine the variance on the mean. The template ensemble sampling is done for each channel $c = \text{ggF SBI, ggF B, ggF S, etc.}$ independently. . . . 102

6.23	Ensemble prediction distributions for a randomly selected event for the $p_B^{\text{ggF}}/p_{\text{ref}}$ density ratio. As an example two samplings using the bootstrap technique are done to show the corresponding spread in the ensemble mean prediction (dashed lines).	103
6.24	Ensemble $p_{\text{SBI}}^{\text{ggF}}/p_{\text{ref}}$ prediction distributions for two randomly selected events. A hundred samplings of the density ratio are done using the bootstrap technique described in text, to show the corresponding spread in the ensemble mean prediction (dashed lines). . . .	103
6.25	Distribution of fitted $\hat{\mu}$ from the various bootstrapped ensembles. The dashed blue line at $\mu = 1.002$ represents the value of the nominal fit $\hat{\mu}_{\text{nom}}$ using the nominal density ratio predictions. For each of the subsequent fits, a randomly sampled template ensemble for each of the channels $c = \text{ggF SBI}, \text{ggF B}, \text{ggF S}, \text{etc.}$ is used to calculate the mean of the event-by-event density ratio predictions p_c/p_{ref} which are then used to find the best fit value of $\hat{\mu}$. The resulting distribution $\Delta\hat{\mu}$ represents the uncertainty in the fit at $\mu = 1.0$ due to the limited Monte Carlo statistics used in the training of the NNs. This is repeated using MC samples at different values of $\mu = \mu_{\text{truth}}$ to then build the parameterized $\Delta\hat{\mu}(\mu)$	105
7.1	a Values of the test statistic $t_{\mu_{\text{off-shell}}}$ assuming a single parameter of interest $\mu_{\text{off-shell}}$ obtained with an Asimov dataset (expected, dashed blue) and with data (observed, solid black) in the $H^* \rightarrow ZZ \rightarrow 4\ell$ decay channel. The values from the histogram-based analysis [3], using fixed summary statistic of Eq. 4.29 for the fits, are added in dash-dotted lines for comparison. The dotted gray lines show the 68% and 95% confidence belt, obtained from the Neyman construction. b Same values obtained with data (observed, solid black) and Asimov dataset (expected, dashed blue) compared with the <i>statistics-only</i> case with all NP fixed at their best-fit values $\hat{\alpha}$	107
7.2	Expected distribution of $t_{\mu_{\text{off-shell}}=0}$ estimated with pseudo-experiments for the case of $\mu_{\text{off-shell}}^{\text{truth}} = 0$ (solid green, no off-shell Higgs boson production hypothesis) and $\mu_{\text{off-shell}}^{\text{truth}} = 1$ (dash-dotted red, SM hypothesis). The vertical solid black (dashed blue) line shows the observed (expected) value of $t_{\mu_{\text{off-shell}}=0}$. The dotted vertical green lines at $t_{\mu_{\text{off-shell}}=0} = 1.99$ and 5.18 represent the one-sided 1σ and 2σ significance thresholds, respectively, under the $\mu_{\text{off-shell}}^{\text{truth}} = 0$ hypothesis. These values differ slightly from the ones in the asymptotic approximation, in which the thresholds would be 1.97 and 5.19	108
7.3	Post-fit pulls and pull errors of the all the nuisance parameters representing the systematic uncertainties in the $\mu_{\text{off-shell}}$ fit.	109

- 7.4 Comparison between observed and the expected background distributions of a the optimal observable at $\mu_{\text{off-shell}} = 0$ and b the quadruplet mass $m_{4\ell}$. The solid red lines shows the expected distribution of the best-fit hypothesis $\hat{\mu}$. The lower panel shows a comparison between the distribution of the background-only and best-fit hypotheses. The background is estimated under the SM hypothesis (post-fit, $\mu_{\text{off-shell}} = 1$). A comparison between the two distributions indicate that the optimal observable built with the NSBI method provides better evidence for off-shell Higgs boson production than only the $m_{4\ell}$ distribution. The hatched area corresponds to the total systematic uncertainty in the expected distributions. The first and last bins contain overflow events. 110
- 7.5 a Values of the test statistic $t_{\mu_{\text{off-shell}}}$ assuming a single parameter of interest $t_{\mu_{\text{off-shell}}}$ obtained with an Asimov dataset (expected, dashed blue) and with data (observed, solid black) combining the $H^* \rightarrow ZZ \rightarrow 4\ell$ and $H^* \rightarrow ZZ \rightarrow 2\ell 2\nu$ decay channels. The dash-dotted curves show the *statistics-only* results where all NP are fixed to their best-fit values $\hat{\alpha}$. The dotted gray lines show the 68% and 95% confidence belt, obtained from the Neyman construction. b Expected distribution of $t_{\mu_{\text{off-shell}}=0}$ estimated with pseudo-experiments for the case of $\mu_{\text{off-shell}}^{\text{truth}} = 0$ (solid green, no off-shell Higgs boson production hypothesis) and $\mu_{\text{off-shell}}^{\text{truth}} = 1$ (dashed-dotted red, SM hypothesis). The vertical solid black (dashed blue) line shows the observed (expected) value of $t_{\mu_{\text{off-shell}}=0}$. The vertical dotted green lines represent the 1σ , 2σ , and 3σ significance thresholds under the $\mu_{\text{off-shell}}^{\text{truth}} = 0$ hypothesis. 113
- 7.6 Values of the a test statistic $t_{\kappa_{g,\text{off-shell}}}$ as a function of $\kappa_{g,\text{off-shell}}$ and b the test statistic $t_{\kappa_{V,\text{off-shell}}}$ as a function of $\kappa_{V,\text{off-shell}}$ obtained with an Asimov dataset (expected, dashed blue) and with data (observed, solid black). The dash-dotted curves show the *statistics-only* results where all NP are fixed to their best-fit values $\hat{\alpha}$. The dotted gray lines show the 68% and 95% confidence belt, obtained from the Neyman construction. The κ parameter not shown is profiled in both cases. The abrupt change at $\kappa_{V,\text{off-shell}} = 0$ comes from the strong $\kappa_{V,\text{off-shell}}^4$ dependency in the EW production of off-shell Higgs bosons. 114
- 7.7 Values of the test statistic t_{κ_H} as a function of $\kappa_H = \Gamma_H/\Gamma_H^{\text{SM}}$ obtained with an Asimov dataset (expected, dotted black) and with data (observed, solid black). The dash-dotted curves show the *statistics-only* results where all NP are fixed to their best-fit values $\hat{\alpha}$. The dotted gray lines show the 68% and 95% confidence belt, obtained from the Neyman construction. 115
- 7.8 Values of a the test statistic $t_{R_{gg}}$ as a function of $R_{gg} = \kappa_{g,\text{on-shell}}^2/\kappa_{g,\text{off-shell}}^2$ and of b the test statistic $t_{R_{VV}}$ as a function of $R_{VV} = \kappa_{V,\text{on-shell}}^2/\kappa_{V,\text{off-shell}}^2$ obtained with an Asimov dataset (expected, dashed blue) and with data (observed, solid black). The values obtained when all NP are fixed to their best-fit values $\hat{\alpha}$, corresponding to the *statistics-only* case, are shown in dash-dotted lines for comparison. The dotted gray lines show the 68% and 95% confidence belt, obtained from the Neyman construction. In both figures, the R parameter not shown is profiled and κ_H is fixed to 1. 116

- 8.1 (a) NN architecture of one of the many thousands of NNs used in the analysis, each with millions of trainable parameters. (b) Training time comparison, between a CPU and an NVIDIA T4 GPU, for one of the NNs used in the analysis. Training on GPU offers almost $5\times$ the speed-up compared to the impractically large training time on CPU, with the same $O(1M)$ simulated events used for each training. This time difference becomes especially relevant when optimizing an ensemble of thousands of NNs. 119
- 8.2 A cartoon schematic of the full NSBI workflow developed in this thesis. For brevity only two processes c_1 and c_2 are shown along with two systematics. Each channel uses a large ensemble of NNs to estimate the ratios $p_{c_1}(\mathbf{x})/p_{\text{ref}}(\mathbf{x})$. The Interp function stands for the polynomial-exponential interpolation defined in Chap. 3. 119
- 8.3 Comparison of the time, in seconds, to perform a profile likelihood fit using the NSBI test statistic $t_{\mu_{\text{off-shell}}}$ written as a JIT-compiled JAX function and as a default python function. The fit is performed using event-by-event systematic variations for an Asimov dataset with $O(10M)$ events. While the computation time remains similar between the python function and the JIT-compiled JAX function for a small number of NPs, the latter scales significantly better as the number of NPs in the fit increases. 120
- 8.4 The OO discriminant $u(\mathbf{x}|\mu_{\text{off-shell}} = 0.2)$ with 55 bins, corresponding to a sufficient summary statistic for $\mu_{\text{off-shell}} = 0.2$. Poisson likelihoods are extracted from each bin and are used to do the t_μ scans shown in Fig. 8.6. 122
- 8.5 The plot shows a comparison of expected stat-only sensitivity from various analysis strategies which can potentially be used in this channel. The dashed pink (black) curves show the expected t_μ scans from an analysis which uses $\log p(\mathbf{x}|\mu_{\text{off-shell}} = 1)/p_{\text{ref}}$ as a discriminant with 10 (24) bins. The various dots show the analysis sensitivity for specific values of the POI $\mu_{\text{off-shell}} (= 0.0, 0.1, 0.2, 1.5, 2.0, 2.5 \text{ in this plot})$, where the discriminant $u(\mathbf{x}|\mu_{\text{off-shell}})$ is used as described in Eq. 8.2. Important takeaway from this plot is that as the number of bins in the OO discriminant approach large values, the sensitivity approaches that of the per-event analysis. Also interesting to note is that having an OO even with fewer number of bins still has better sensitivity than the $\log p(\mathbf{x}|\mu_{\text{off-shell}} = 1)/p_{\text{ref}}(\mathbf{x})$ observable optimized only for $\mu_{\text{off-shell}} = 1$ 123
- 8.6 The floating dots in the Fig. 8.5 are actually just part of the full stat-only t_μ curve with the $u(\mathbf{x}|\mu)$ discriminant optimized for that particular value of μ . Here we plot the full NLL curve with the dot to clarify the situation. Note how the OO optimized for $\mu_{\text{scan}} = 0.0$ offers increased sensitivity for $\mu = 0.0$ while showing sub-optimal performance for other values of $\mu_{\text{off-shell}}$, for example $\mu_{\text{off-shell}} = 0.2$ which has increased sensitivity when the OO is optimized for $\mu_{\text{scan}} = 0.2$ 124

- 8.7 A stat-only binned Poisson likelihood fit using real ATLAS experiment data, with the OO discriminant optimized for $\mu_{\text{off-shell}} = 0$. The left plot shows the discriminant $u(\mathbf{x}|\mu_{\text{off-shell}} = 0)$ modeled with 200 bins, and the right plot shows the expected and observed t_μ scans using the binned observable. The right plot also shows comparisons to the result from the NSBI technique, described in Chap. 7. Notably, the locally optimal $u(\mathbf{x}|\mu_{\text{off-shell}} = 0)$ discriminant replicates the sensitivity and data fluctuation in the vicinity of $\mu_{\text{off-shell}} = 0$, while being relatively sub-optimal for $\mu_{\text{off-shell}} \gg 0$. This highlights the importance of parameterized OO in an analysis with degenerate solutions. 125
- A.1 Values of the test statistic $t_{\mu_{\text{off-shell}}}$ assuming a single parameter of interest $\mu_{\text{off-shell}}$ obtained with an Asimov dataset (expected, dashed blue) and with data (observed, solid black) when combining the $H^* \rightarrow ZZ \rightarrow 4\ell$ and $H^* \rightarrow ZZ \rightarrow 2\ell 2\nu$ decay channels. The values from the histogram-based analysis [3] are added with dash-dotted lines for comparison. The dotted gray lines show the 68% and 95% confidence belt, obtained from the Neyman construction, for the NSBI analysis. 129
- A.2 a Values of the test statistic t_{κ_H} as a function of $\kappa_H = \Gamma_H/\Gamma_H^{\text{SM}}$ for the $H \rightarrow ZZ \rightarrow 4\ell$ channel obtained with an Asimov dataset (expected, dashed blue) and with data (observed, solid black). The values from the histogram-based analysis [3] are added with dash-dotted lines for comparison. The dotted gray lines show the 68% and 95% confidence belt, obtained from the Neyman construction, for the NSBI analysis. b Same comparison of the values of the test statistic for $\kappa_H = \Gamma_H/\Gamma_H^{\text{SM}}$ after combination with the $H^* \rightarrow ZZ \rightarrow 2\ell 2\nu$ channel. 130
- A.3 Values of a the test statistic $t_{R_{gg}}$ as a function of R_{gg} , and of b the test statistic $t_{R_{VV}}$ as a function of R_{VV} for the combined $H \rightarrow ZZ \rightarrow 4\ell$ and $H^* \rightarrow ZZ \rightarrow 2\ell 2\nu$ channels obtained with an Asimov dataset (expected, dashed blue) and with data (observed, solid black). The equivalent graphs from the histogram-based analysis [3] are added with dash-dotted lines for comparison. The dotted gray lines show the 68% and 95% confidence belt, obtained from the Neyman construction, for the NSBI analysis. . . . 130
- B.2 The sampling distribution $g(\hat{\mu}|\mu_{\text{truth}})$ for the POI $\mu_{\text{off-shell}}$ of the off-shell Higgs boson analysis, defined in Chap. 4. Sampled pseudo-experiments are used to fit the POI μ using the t_μ test statistic with sampled AOs. The resulting fit distributions $\hat{\mu}$ are shown as a function of $\mu = 0.5, 1.0, 2.5, 3.5$ (top left, top right, bottom left, bottom right). . . 133

List of Tables

4.1	Definition of processes in the probability model used to interpret the data and the simulated samples used to describe them. The multipliers define both the measured off-shell Higgs boson production signal strength and the data-driven normalization of leading backgrounds. The $(H^* \rightarrow)$ notation is used for the SBI process. The multipliers are obtained after substituting the interference terms of Eqs. 4.22 and 4.23 into Eq. 4.19, since interference-only samples are not simulated. For instance, in the case of ggF production: $\mu\nu_S + \sqrt{\mu}\nu_I + \nu_B = (\mu - \sqrt{\mu})\nu_S + \sqrt{\mu}(\nu_S + \nu_I + \nu_B) + (1 - \sqrt{\mu})\nu_B$	64
4.2	Description of the different sources of modeling systematic uncertainties considered for each process and the method used to estimate it.	68
5.1	Matrix element generator, parton shower (PS), and higher-order corrections used to describe the different samples in the measurement. The notation $[Xj]$ indicates that up to X jets are included in the hard-scatter matrix element calculation.	71
5.2	Definition of the observables $\mathbf{x}$ used to describe an event. The observables are defined relative to the ATLAS coordinate system, and are various mappings of the full four-momenta of final state particles. The vectors $\mathbf{q}_1$, $\mathbf{q}_2$, $\mathbf{q}_{11}$, $\mathbf{q}_{21}$, $\mathbf{n}_1$, $\mathbf{n}_2$, and $\mathbf{n}_{sc}$ are defined in Sec. 5.3.	74
7.1	Absolute systematic uncertainties in the measurement of $\mu_{\text{off-shell}}$ in the $H \rightarrow ZZ \rightarrow 4\ell$ decay channel. Two methods of estimation are presented: based on the variation of nuisance parameters and on the variation of auxiliary observables. Uncertainties are given using the auxiliary observables methods since it allows variations to be summed in quadrature. The total uncertainty is independent of the method used to estimate systematic uncertainties.	111

7.2	Summary of the results for $\kappa_{g, \text{ off-shell}}$, $\kappa_{V, \text{ off-shell}}$, and $\mu_{\text{ off-shell}}$ combining the measurement of the off-shell Higgs boson production analysis in the $H \rightarrow ZZ \rightarrow 4\ell$ decay channel presented in this paper with the analysis in the $H \rightarrow ZZ \rightarrow 2\ell 2\nu$ decay channel from Ref. [3]. The results for Γ_H , R_{gg} and R_{VV} are obtained combining the off-shell production measurement with the on-shell analysis from Ref. [126]. All results are presented with their 68% CL and 95% CL intervals. No expected 95% CL interval for $\kappa_{g, \text{ off-shell}}$ is obtained because the Asimov sample is unable to break the degeneracy between $\kappa_{g, \text{ off-shell}}$ and $\kappa_{V, \text{ off-shell}}$ in ggF production. All results use the full Run 2 dataset with 140 fb^{-1} of integrated luminosity.	117
A.1	The impact of most important systematic uncertainties in the observed upper value of $\mu_{\text{ off-shell}}$ for which $t_{\mu_{\text{ off-shell}}} = 4$, obtained by the combined fit. This value corresponds to the two standard deviation upper limit of $\mu_{\text{ off-shell}}$ in the asymptotic approximation. The first column denotes the systematic uncertainty that was excluded from the fit. The last row gives the nominal upper limit, where all uncertainties are included. The importance of an uncertainty its given by its deviation from the last row.	129
A.2	Summary of the results for the NSBI analysis with a comparison to the corresponding results of the histogram analysis of Ref. [3]. The value of the parameter estimates and the 68% and 95% confidence intervals are provided. All results use the full Run 2 dataset with 140 fb^{-1} of integrated luminosity.	131

Introduction

At the Large Hadron Collider (LHC), the ATLAS and CMS experiments widely employ frequentist composite hypothesis testing to detect evidence of fundamental interactions of elementary particles produced during proton-proton collisions and to establish confidence limits on relevant parameter spaces. This hypothesis testing often involves reducing complex, multi-dimensional detector outputs into low-dimensional distributions using specific mapping functions or discriminant variables that optimally separate Monte Carlo (MC) simulated signals from the backgrounds. Histogram templates of these variables are constructed and used to fit the parameters of interest to the observed data, while simultaneously profiling away the nuisance parameter effects associated with various sources of uncertainty in the physics modeling and detector response. The fit is performed bin by bin using a maximum likelihood estimation method, with the assumption that the number of events in each bin follows a Poisson distribution.

These techniques have been optimized for resonance measurements at the LHC, aiming to minimize Type II errors (false negatives) while remaining robust against Type I errors (false positives). As the LHC enters a precision era, traditional resonance searches where the observables scale linearly with the parameter of interest are being complemented or replaced by measurements of non-resonant effects with observables that follow a more complex parameter dependence, such as those explored within the Effective Field Theory (EFT) framework. This increasing complexity and precision required in LHC analyses necessitate the adoption of more general inference techniques that don't rely on ad-hoc assumptions.

Neural Simulation-Based Inference (NSBI), also sometimes referred to as Simulation-Based Inference [1] or Likelihood-Free Inference [2], is a class of deep learning techniques for statistical inference that effectively handles high-dimensional data without relying on low-dimensional summary histograms. They are also by construction optimal throughout the parameter space being scanned over, regardless of the complexity in the parameterization. In this thesis, I develop a type of NSBI analysis to be applied to the ATLAS experiment, that utilizes over 10,000 neural networks to estimate precise and accurate event-by-event parameterized likelihood ratios. These carefully estimated per-event likelihood ratios are then directly used to fit the parameters of interest using the standard negative log-likelihood ratio test statistic. While NSBI techniques in general have been widely employed in Bayesian settings and toy phenomenological models, this work represents the first application of NSBI methods for frequentist parameter inference in a full-scale particle physics experiment. This advancement was made possible through novel developments presented here that enable the parameterization of likelihood ratios as functions of an arbitrary number of uncorrelated parameters, as

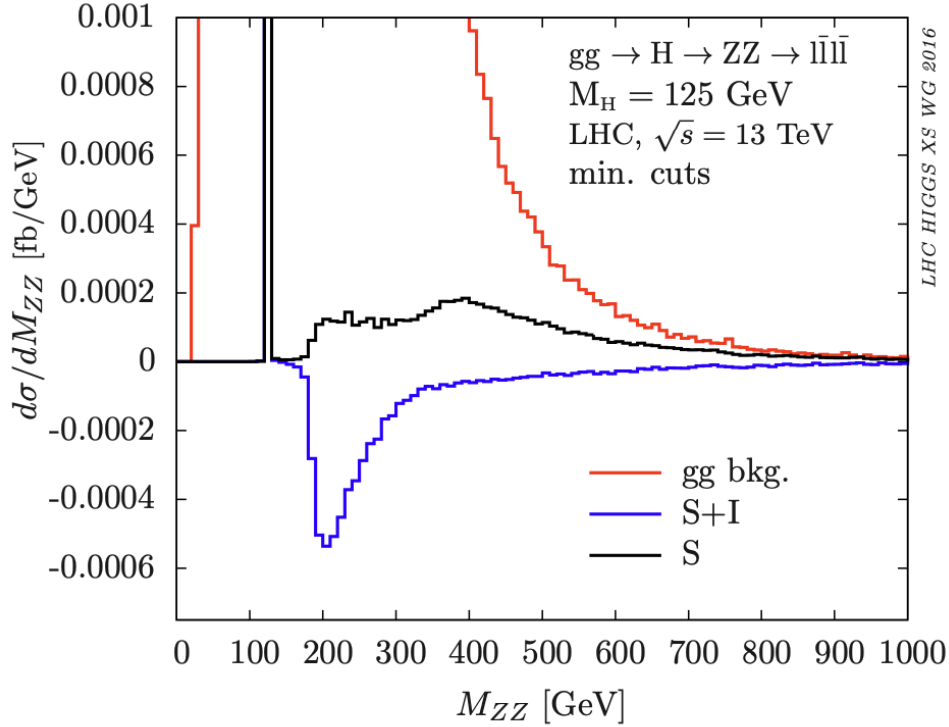


Figure 1: The invariant mass m_{ZZ} distribution of the $H \rightarrow ZZ \rightarrow 4\ell$ decay channel shows a large non-resonant Higgs boson signal $gg \rightarrow H^* \rightarrow ZZ \rightarrow 4\ell$ (S) off its mass shell of 125 GeV, along with a large quantum interference (S+I) with the background $gg \rightarrow ZZ \rightarrow 4\ell$ (gg bkg.) in this phase space. The measurement of the off-shell Higgs boson production is measuring the S+I distribution shown in the figure.

well as the implementation of innovative computing workflows leveraging large-scale GPU and CPU infrastructures.

For a first application of this novel technique, a measurement of the off-shell Higgs boson production in the $H^* \rightarrow ZZ \rightarrow 4\ell$ decay channel is done. This production is an example of non-resonant production with a large signal-background interference that lowers the optimality of standard techniques in minimizing Type II errors. Fig. 1 shows the large signal-background interference distribution compared to the non-resonant signal $gg \rightarrow H^* \rightarrow ZZ \rightarrow 4\ell$ distribution in the phase space regions $m_{ZZ} \gtrsim 180.0$ GeV. The measurement uses the 140 fb^{-1} of integrated luminosity collected by the ATLAS detector during the Run 2 proton-proton collisions of the Large Hadron Collider at $\sqrt{s} = 13$ TeV and supersedes the previous result in this decay channel using the same dataset.

The observed (expected) off-shell Higgs boson production signal strength in the $ZZ \rightarrow 4\ell$ decay channel, for which the NSBI techniques have been used, is $0.87^{+0.75}_{-0.54}$ ($1.00^{+1.04}_{-0.95}$) at 68% CL. The previously published result [3] using the same dataset did not achieve an expected sensitivity to quote a two-sided interval at this CL, which the new measurement does due to improved sensitivity. The expected plus-side uncertainty is reduced by 10%, while the evidence for off-shell Higgs boson production has an observed (expected) significance of 2.5σ (1.3σ) using the $ZZ \rightarrow 4\ell$ decay channel only. The latter represents an observed (expected) improvement of 300% (260%) in excluding the background-only

hypothesis, compared to the previously published analysis using the same data.

When the new measurement in the $ZZ \rightarrow 4\ell$ decay channel is combined with the most recent ATLAS measurement of the off-shell Higgs boson in the $ZZ \rightarrow 2\ell 2\nu$ decay channel, which used standard inference techniques, the evidence for off-shell Higgs boson production has an observed (expected) significance of 3.7σ (2.4σ). The off-shell measurements in the $ZZ \rightarrow 4\ell, 2\ell 2\nu$ channels are combined with the measurement of on-shell Higgs boson production in the $ZZ \rightarrow 4\ell$ decay channel to obtain constraints on the Higgs boson total width, under certain assumptions. The observed (expected) value of the Higgs boson width is found to be $4.29^{+2.66}_{-1.88}$ ($4.10^{+3.51}_{-3.44}$) MeV at 68% CL, in agreement with the SM value.

The work presented in this thesis has resulted in several publications:

- The ATLAS collaboration, *Measurement of off-shell Higgs boson production in the $H^* \rightarrow ZZ \rightarrow 4\ell$ decay channel using a neural simulation-based inference technique in 13 TeV pp collisions with the ATLAS detector* [4].
- The ATLAS collaboration, *An implementation of neural simulation-based inference for parameter estimation in ATLAS* [5].
- Sandesara et al, *ATLAS Data Analysis using a Parallel Workflow on Distributed Cloud-based Services with GPUs* [6].
- Barreiro Megino, et al, *Operational Experience and R&D results using the Google Cloud for High Energy Physics in the ATLAS experiment* [7].
- The ATLAS collaboration, *Evidence of off-shell Higgs boson production from ZZ leptonic decay channels and constraints on its total width with the ATLAS detector* [3].
- The ATLAS collaboration, *Effective field theory interpretation of the measurement of off-shell Higgs boson production from $ZZ \rightarrow 4\ell$ and $ZZ \rightarrow 2\ell 2\nu$ decay channels with the ATLAS detector* [8].

The thesis is arranged in a pedagogical way, with a brief introduction to the ATLAS experiment (Chap. 1) and the subsequent chapters going from the general review of the statistical framework used at the LHC experiments to the introduction of the novel statistical method using NSBI techniques (Chaps. 2, 3), followed by a full-scale analysis of the off-shell Higgs boson and an interpretation of this measurement to constrain the Higgs boson decay width (Chaps. 4, 5, 6, 7), and finally going into the details of the computational challenges involved in this measurement using NSBI techniques with practical solutions for wide-scale use within the LHC (Chap. 8).

Chapter 1

The ATLAS Experiment

A brief description of the ATLAS experimental setup is given here. After the basic introduction to the detector layout, the focus will be on Inner Detector track reconstruction, setting a base for the introduction of some new developments done towards building low-level validation tools for the various stages of the full track reconstruction chain. This chapter is not essential to the development of Neural Simulation-Based Inference techniques presented in the rest of the thesis.

The Large Hadron Collider (LHC) at CERN is the world's most powerful particle accelerator designed to study fundamental particle physics at the high-energy frontier. It has been designed to accelerate and collide proton beams upto an unprecedented maximum centre-of-mass energy of 14 TeV. The full setup at CERN leading up to the two opposing, highly accelerated beams of protons inside the LHC ring is shown in Fig. 1.1. The four main experiments, ATLAS, ALICE, CMS and LHCb are located at the four Interaction Points (IP), where the two opposing proton beams cross each other to produce collisions that are then recorded by the detectors of each of the experiments.

The number of proton-proton collision *events* per second generated in the LHC collisions is given by:

$$N = \sigma \int \mathcal{L} \cdot dt \tag{1.1}$$

where σ is the total cross-section and $\mathcal{L}$ is the instantaneous luminosity. While the LHC was originally designed for instantaneous luminosities of upto $10^{34} \text{ cm}^{-2} \text{ s}^{-1}$, this has been surpassed in the Run-2 and Run-3 data-taking operations of the ATLAS and CMS experiments. A detailed summary on the LHC can be found in Ref. [9].

1.1 The ATLAS Detector

The ATLAS (A Toroidal LHC ApparatuS) detector is one of the two large, general-purpose particle detectors at the LHC, along with CMS (Compact Muon Solenoid). It studies a wide range of physics phenomena by analyzing the particles produced from proton-proton collisions at its IP. The primary physics goals of the ATLAS detector include the precise measurement of Standard Model (SM) processes, such as the production and decay of the Higgs boson, top quark, and electroweak



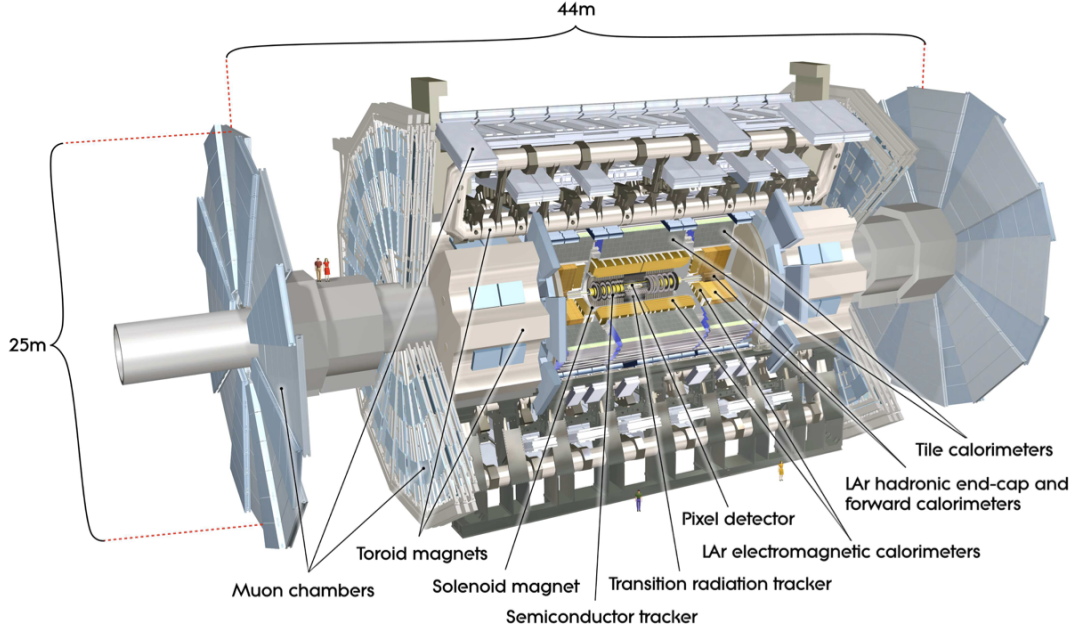


Figure 1.2: The layout of the full ATLAS detector at the Large Hadron Collider. The Pixel, Semiconductor and Transition radiation tracker constitute the Inner Detector, which is immersed in a 2T magnetic field parallel to the beam axis produced by the Solenoid magnet. The Inner Detector is followed by electromagnetic and hadron calorimeters, and a Muon spectrometer chambers incorporating three large superconducting Toroid magnets.

gauge bosons; the search for evidence of physics beyond the Standard Model (BSM), including supersymmetry (SUSY), extra dimensions, and dark matter candidates.

The ATLAS detector consists of an inner tracking detector surrounded by a thin superconducting solenoid, electromagnetic and hadron calorimeters, and a muon spectrometer incorporating three large superconducting air-core toroidal magnets. The full layout can be seen in Fig. 1.2.

1.1.1 The Coordinate System

To record the positions and momenta of the particles produced from proton-proton collisions at the IP located in the center of the detector system, the ATLAS experiment uses a right-handed coordinate system with its origin at the IP and the z -axis along the beam pipe. The x -axis points from the IP to the centre of the LHC ring, and the y -axis points upwards. Cylindrical coordinates (r, ϕ) are used in the transverse plane, ϕ being the azimuthal angle around the z -axis. The pseudorapidity is defined in terms of the polar angle θ as $\eta = -\ln \tan(\theta/2)$. Angular distance is measured in units of $\Delta R \equiv \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$. The geometric layout is shown in Fig. 1.3.

The transverse momentum p_T of a particle reconstructed in the ATLAS detector is given by

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

where p_x and p_y are the components of the momentum vector in the plane perpendicular to the beam axis. The transverse energy E_T is similarly defined as $E_T = E \sin \theta$, where E is the total energy and θ

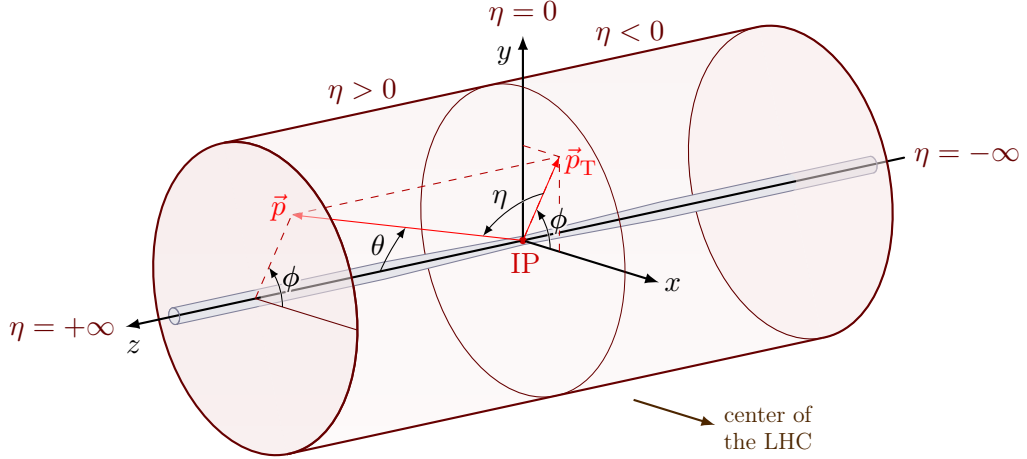


Figure 1.3: The ATLAS detector coordinate system. The definitions of the various quantities are given in Sec. 1.1.1. Figure source code taken and modified from Ref. [11].

the polar angle relative to the beam axis. These transverse definitions are important because the initial momentum along the beam axis is not precisely known, whereas the net initial transverse momentum is always zero. The missing transverse energy E_T^{miss} (or $\vec{E}_T^{\text{miss}}$) is the negative of the vector sum of measured transverse momenta of all the selected and calibrated physics objects in the detector, with a soft term added to account for soft energy that is not associated with any of the selected objects

$$\vec{E}_T^{\text{miss}} = - \sum_i p_{T,i} - \sum p_{T,\text{soft}}.$$

This quantity for example, captures transverse energy contributions from neutrinos or other weakly interacting particles that escape detection, resulting in a non-zero sum in the transverse direction to the beam pipe (or along with $x - y$ plane). The transverse quantities defined here will be used in the off-shell Higgs boson analysis described in Chap. 5, with the E_T^{miss} playing a key role in the $ZZ \rightarrow 2\ell 2\nu$ decay channel with neutrinos in the final state.

1.2 The Inner Detector

The inner-detector system (ID), shown in Fig. 1.4, is the innermost ATLAS sub-detector system. Its entire volume is immersed in a 2T magnetic field along the axis from a solenoid magnet surrounding it, and it is responsible for reconstructing the trajectories of charged particles originating from the the LHC interaction point (IP) in the range $|\eta| < 2.5$. The layout and technology of the ID are hereby described, followed by a brief account of the full chain of track reconstruction, and some corresponding versatile tools that were developed for validations of the reconstruction chain at the low-level stages, before the final track formation.

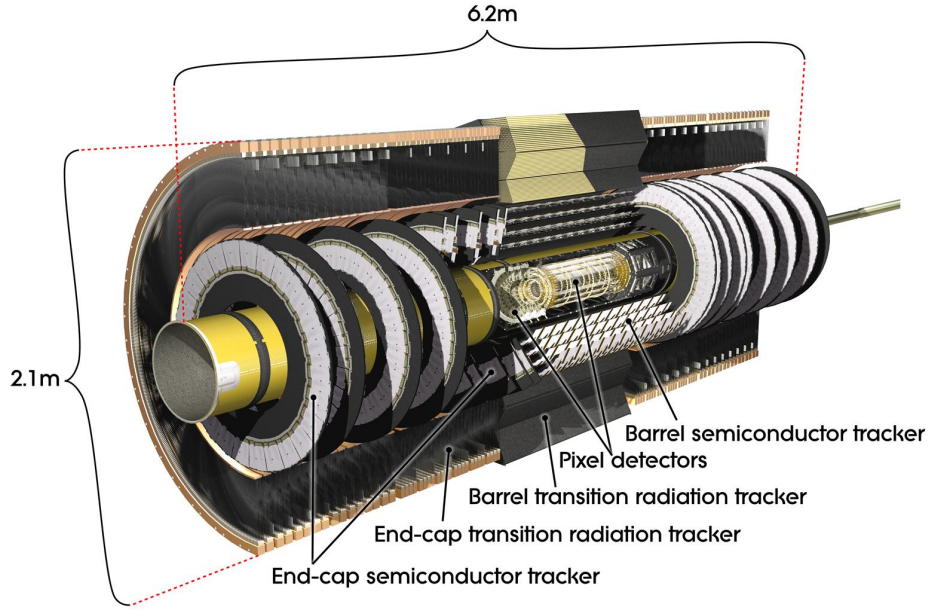


Figure 1.4: The ATLAS Inner Detector system. Figure taken from Ref. [12].

1.2.1 Detector Layout and Technologies

The ID is organized in concentric cylindrical layers around the beam line, as shown in Fig. 1.5, extending into forward regions with disk-like endcaps, shown in Fig. 1.4. The main detector technologies are the Pixel Detector, the Semiconductor Tracker (SCT), and the Transition Radiation Tracker (TRT). Together, they provide measurements of a charged particle’s position over multiple radii, enabling robust momentum determination and precise vertex reconstruction.

Pixel Detector

The Pixel Detector is the innermost subsystem of the ID and is located closest to the beam pipe. It consists of silicon pixel sensors with a very fine pitch. The pixel modules are arranged in cylindrical “barrel” layers and “endcap” disks, as shown in Fig. 1.6, ensuring good acceptance in pseudorapidity up to about $|\eta| \approx 2.5$. The barrel region consists of four concentric layers, and together with the endcap disks comprises of roughly 92 million pixels, with an equal number of readout channels, providing excellent spatial resolution.

The first layer is labeled Insertable B-Layer (IBL), a new tracking layer added before the start of Run 2 and which is the closest to the beam pipe and designed to improve the precision and robustness of ID track reconstruction. The remaining layers are labeled Layer 0 (L0), Layer 1 (L1) and Layer 2 (L2) in order of succession from the beam pipe. Each pixel in the detector sub-layers, achieves a spatial resolution of about $10\mu\text{m}$ in the transverse plane and $115\mu\text{m}$ along the beam pipe axis. In the IBL, the longitudinal resolutions along the beam pipe axis are further reduced to $60\mu\text{m}$ [13]. The four layers of the Pixel Detector can provide up to four hits for a single particle, one per layer, each with enough information to be converted into three-dimensional spacepoints as will be discussed in Sec. 1.6.

Semiconductor Tracker (SCT)

Surrounding the Pixel Detector is the Semiconductor Tracker (SCT), which employs silicon micro-strip detectors (as opposed to individual pixels in the Pixel detector). Each module consists of two pairs of back-to-back single-sided silicon strip sensors with a 40 mrad angle between them. This geometry helps to resolve the two-dimensional position of particle hits, complementing the fine granularity of the pixel measurements. Like the pixel system, the SCT detector system is arranged in four barrel layers and two endcaps with nine disks each, as shown in Figs. 1.4 and 1.5. The coverage is also the same as Pixel detector system, covering particle trajectories within $|\eta| < 2.5$. Each of the four SCT layers offer a precision of $580\mu m$ along the beam pipe and $17\mu m$ in the transverse direction, considerably larger than the precision in the Pixel detector. As such, fewer readout channels are required adding up to over 6 million for the full sub-detector system.

Transition Radiation Tracker (TRT)

The TRT is the outermost part of the ID, composed of thin gas-filled straw tubes interleaved with radiator material. Charged particles crossing the boundaries of materials generate transition radiation, which is more intense for lighter particles such as electrons compared to heavier charged particles. This feature helps with electron identification at intermediate and high momenta. The straw tubes also measure drift times to provide a position measurement along the azimuthal coordinate, adding further confirmation of the track and improving overall tracking efficiency and robustness.

1.3 The Calorimeter System

The calorimeter system covers the pseudorapidity range $|\eta| < 4.9$. Within the region $|\eta| < 3.2$, electromagnetic calorimetry is provided by barrel and endcap high-granularity lead/liquid-argon (LAr) calorimeters, with an additional thin LAr presampler covering $|\eta| < 1.8$ to correct for energy loss in material upstream of the calorimeters. Hadron calorimetry is provided by the steel/scintillator-tile calorimeter, segmented into three barrel structures within $|\eta| < 1.7$, and two copper/LAr hadron endcap calorimeters. The solid angle coverage is completed with forward copper/LAr and tungsten/LAr calorimeter modules optimised for electromagnetic and hadronic energy measurements respectively.

1.4 The Muon Spectrometer

The muon spectrometer (MS) comprises separate trigger and high-precision tracking chambers measuring the deflection of muons in a magnetic field generated by the superconducting air-core toroidal magnets. The field integral of the toroids ranges between 2.0 and 6.0 Tm across most of the detector. Three layers of precision chambers, each consisting of layers of monitored drift tubes, cover the region $|\eta| < 2.7$, complemented by cathode-strip chambers in the forward region, where the background is highest. The muon trigger system covers the range $|\eta| < 2.4$ with resistive-plate chambers in the barrel, and thin-gap chambers in the endcap regions.

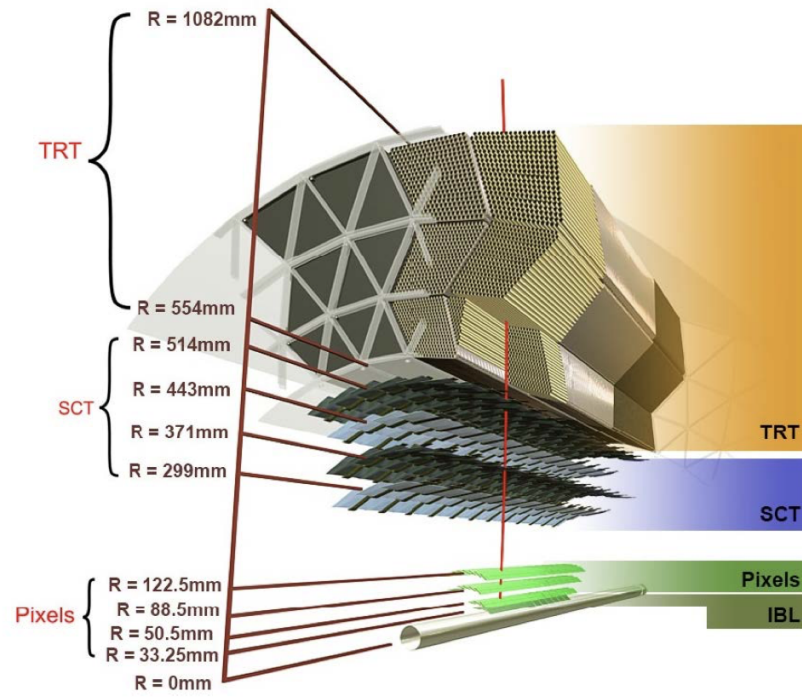


Figure 1.5: A simplified layout of the various sensors and structural elements constituting the ATLAS Inner Detector. IBL stands for the Insertable B-layer introduced before the start of Run-2, followed by the Pixel detectors, the Semi-conductor Trackers (SCT) and the Transition Radiation Trackers (TRT). The red line shows a charged particle track with 10 GeV transverse momentum p_T and $\eta = 0.3$ traversing the Inner Detector. Figure taken from Ref. [14].

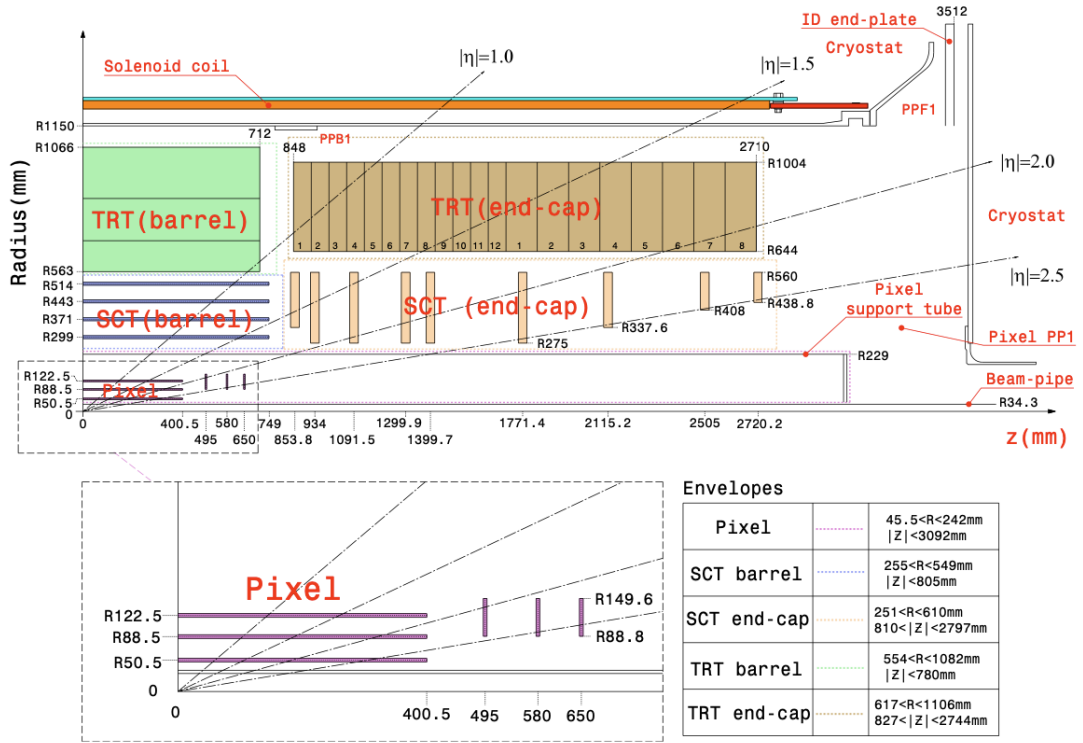


Figure 1.6: The schematic of the ATLAS Inner Detector (ID) system, in the $r - z$ plane. Figure taken from Ref. [14].

1.5 Trigger System

Interesting events are selected by the first-level trigger system implemented in custom hardware, followed by selections made by algorithms implemented in software in the high-level trigger [15]. The first-level trigger accepts events from the 40 MHz bunch crossings at a rate below 100 kHz, which the high-level trigger further reduces in order to record events to disk at about 1 kHz.

1.6 Event Reconstruction in the ID

When a charged particle passes through the silicon or gas-based subdetectors of the ID, it leaves localized energy deposits (hits). These hits are read out by the front-end electronics, digitized, and sent to the data-acquisition (DAQ) system, if the event leaving the hits is accepted by the Level-1 trigger [16]. Tracks and vertices of charged particles are reconstructed using the hits in the ID, and the corresponding information is used to reconstruct *physics objects* such as the electrons, muons or hadronic jets. The simulations of the signatures left by these objects, both in the ID and other parts of the detector, are shown in Fig. 1.7. Detailed algorithms for track and vertex reconstruction are described in Ref. [17], but parts of it will be discussed here as a basis for the introduction of some novel validation tools in Sec. 1.7.

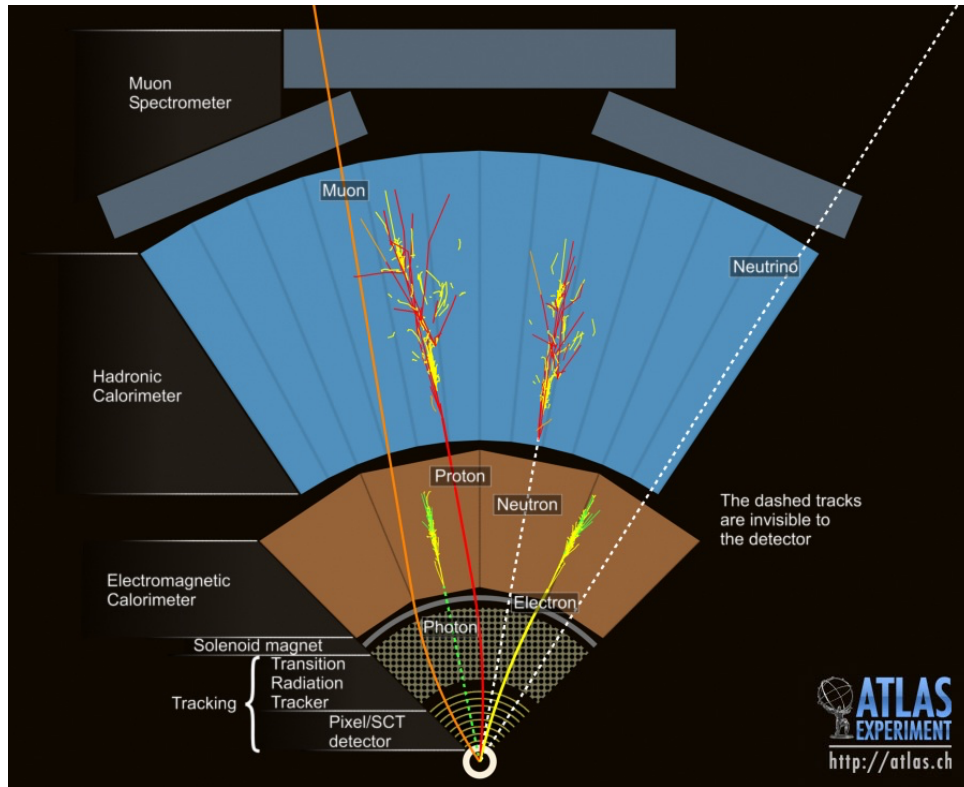


Figure 1.7: A simulation of the various particles interacting with the different sub-detector systems in the ATLAS detector. Figure taken from Ref. [18].

The Athena framework

The entire codebase used for event reconstruction inside the ATLAS experiment is contained in the open-source software framework called Athena [19]. An Athena application consists of dynamically-loadable components, which implement the concepts of Algorithms, Services, and Tools, described below:

1. Algorithms are used to process the data accessible from a shared event store. They read objects from the shared store and write new objects back to the store for use in subsequent algorithms.
2. Services are objects used by other components. For example the event store itself is a service, and so is error logging, and random number generation used in the reconstruction. The Service class is designed to provide dedicated functionality during the entire program execution, e.g. the central data storages (transient event store or the detector store) are realised as ATHENA Service objects. Service instances are handled by a central ExtSvc manager, that regulates initialisation and finalisation.
3. Tools serve as helpers for other components in Athena. They may be uniquely owned by Algorithms, Services, or even other Tools.

A typical reconstruction job in Athena thus makes use of hundreds of these Algorithms, Services, and Tools, all of which must be correctly configured in order for the job to run. A schematic of a typical

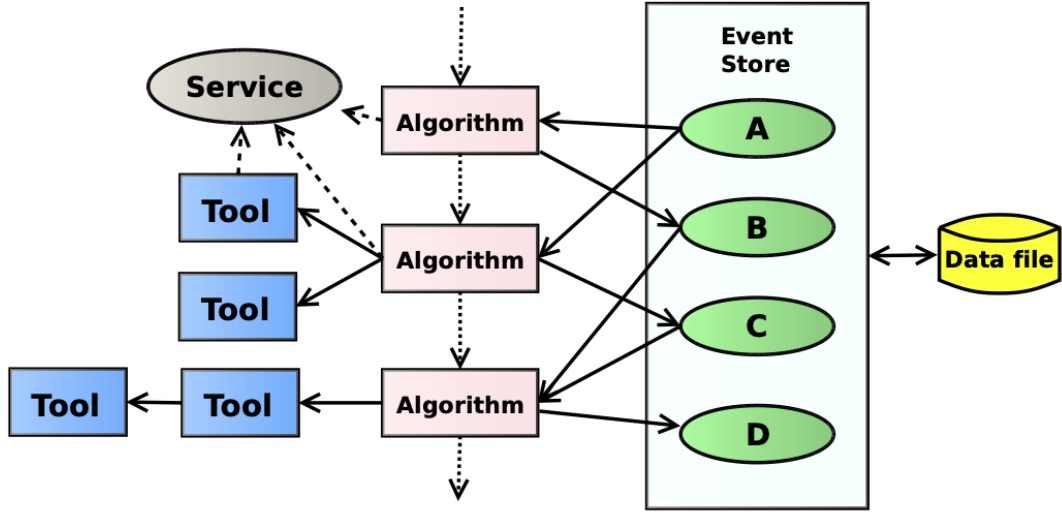


Figure 1.8: General structure of an Athena application. More details in Ref. [20], from which this figure is taken.

Athena application is shown in Fig. 1.8. More details can be found in Ref. [20].

1.6.1 Track Reconstruction

The first step in reconstructing a collision event inside the ATLAS detector is the track reconstruction of charged particles in the ID. The track reconstruction algorithm is described in Ref. [17], and additional details on the track reconstruction can be found in the dedicated paper [21]. The thousands of hits produced in the silicon detectors of the ID, from each hard-scattering event stored for offline processing, are converted into high-quality tracks through a successive execution of algorithms. As shown in Fig. 1.9, this is divided into two passes - first, an inside-out approach starts from reconstruction using the hits in the silicon layers (pixels and SCTs) of the ID, with outward extrapolation by searching for hits in the TRT that are compatible with the silicon layer hits. This baseline approach for track reconstruction is also known as the Primary Tracking Pass, and is followed by an *outside-in* Back-Tracking pass, where track segments reconstructed from the TRT hits are extrapolated inwards by searching for compatible hits in the silicon layers. The focus of this section will be on the inside-out primary tracking pass. The main algorithm within the Athena framework that encompasses this inside-out stage of the track reconstruction is **SiSpSeededTrackFinder**¹.

Each ID track is parameterized by five parameters and a reference point,

$$(d_0, z_0, \phi, \theta, q/p) \quad (1.2)$$

using a perigee representation depicted in Fig. 1.10. The first two parameters are the transverse and longitudinal impact parameters d_0 and z_0 , defined as the transverse and longitudinal distances of the

¹The code corresponding to the algorithms and tools listed in this chapter can be found in the Athena framework [19]. The algorithm names and descriptions mentioned in the thesis correspond to those in the Athena framework at the time of writing.

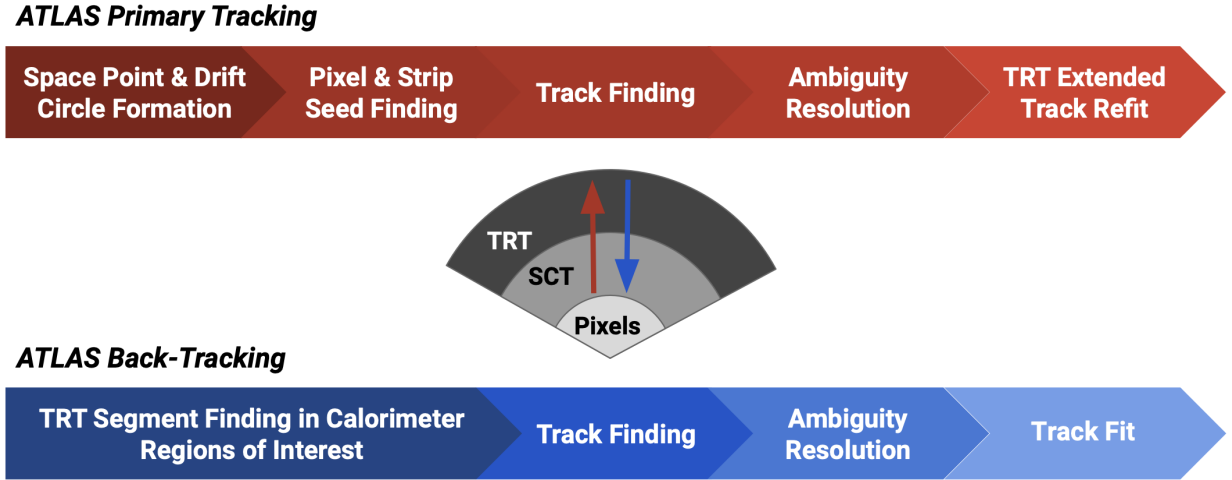


Figure 1.9: A flowchart of the various stages of ID track reconstruction. A first inside-out primary pass (top) is performed followed by an outside-in secondary backtracking pass (bottom). Figure taken from Ref. [20].

single point of closest approach transverse to the reference point. The other three parameters are the azimuthal angle ϕ and the polar angle θ of the track momentum at the reference point and the ratio q/p of the charge of the reconstructed track divided by the magnitude of its momentum. The reference point used is the average measured position of the proton–proton interactions, also referred to as the beamspot position.

Clustering and Spacepoint formation

One of the first steps in track reconstruction is the formation of clusters and spacepoints from raw digitized signals. This stage is crucial for measuring the positions of charged-particle hits in three dimensions, enabling subsequent pattern recognition and track fitting. Clusters are formed using signals from adjacent channels of pixels and SCT sub-detectors, interpreted as deposits left by individual traversing charged particles. The clusters from the single pixel-layer and from the pairs of one-dimensional SCT clusters on either side of each sensor module, defined earlier, are then further converted into three-dimensional *space-points*. These space-points are the beginning of the track reconstruction process, with the top-level algorithm `SiTrackerSpacePointFinder`. The clusterization is done using the `PixelClusterization` and `SCT_Clusterization` algorithms, and the spacepoint formation proceeds through the `SiTrackerSpacePointFinder` algorithm.

Track Seeding

The primary track reconstruction pass starts by forming track seeds, which are a combination of three consecutive space-points (SP), in either the Pixel or SCT subdetectors. The seeding stage is motivated to maximize the number of possible SP combinations that can potentially result in a good quality track matched to a charged particle. This step of the reconstruction is handled by the `SiSPSeededTrackFinder` algorithm within Athena. To start, all possible combinatorial triplets are

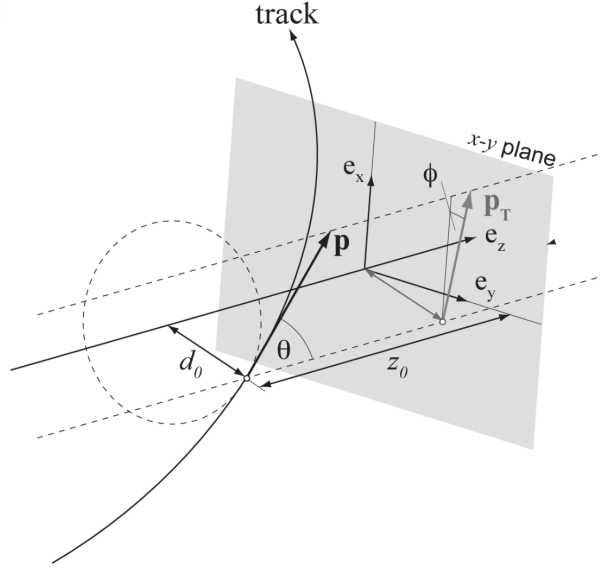


Figure 1.10: The perigee representation $(d_0, z_0, \phi, \theta, q/p)$ used to parameterize ATLAS ID tracks. Figure taken from Ref. [22].

tested as seeds, with their momentum and impact parameters estimated coarsely by assuming a perfect helical extrapolated trajectory. These track seeds then have some loose selection criteria, based for example on the impact parameter and momentum estimates, that are applied to abort reconstruction of seeds that are guaranteed to produce low-quality tracks. Put another way, the seeds are checked to be consistent with the expected helical (or nearly helical) trajectory of a charged particle in a 2 T magnetic field. If these seed patterns are confirmed to be physically plausible (for instance, by checking the compatibility of the curvature and directions with a real particle passing through the detector), the algorithm attempts to attach additional hits from the surrounding layers or the TRT.

Track Extension and Ambiguity Solving

Once an initial track seed has been found, three-dimensional search roads are built through the remaining detector based on the estimated seed trajectory, searching for compatible hits in the next layers. Compatible hits are those whose positions and errors (in conjunction with the estimated track parameters) satisfy statistical criteria, typically a cut on the χ^2 increment when incorporating a new hit into the track fit. This iterative addition of hits continues outward, culminating in a complete track candidate that spans multiple layers.

Due to the high occupancy in certain regions and the possibility that multiple seeds could converge on the same physical trajectory, *track ambiguity resolution* steps are crucial. Algorithms score track candidates based on the number of shared hits, the overall χ^2 , and the number of hits per track, preferring tracks that exhibit the highest probability of representing a real trajectory while discarding duplicate or spurious fits.

Track Fitting and Momentum Determination

Once a set of hits is assigned to a given track candidate, a dedicated track-fitting algorithm refines the trajectory and determines the particle's momentum. Commonly used fitting techniques include the Kalman filter, which updates track parameters (position, slope, curvature, etc.) as successive hits are added. The Kalman filter can also propagate the covariance matrix of the track parameters through the inhomogeneous magnetic field and material layers, ensuring robust error estimates.

In ATLAS, material effects such as multiple scattering and energy loss are carefully modeled. Multiple scattering in the ID layers can degrade the track resolution, while energy loss (mainly by ionization) slightly alters the particle's momentum as it passes through the material. The fitting algorithm corrects for these effects to provide an unbiased momentum estimate. A final global quality assessment is performed, including a check of the final fit χ^2 and the number of degrees of freedom.

1.6.2 Vertex Reconstruction

In parallel with or subsequent to track fitting, reconstructed tracks are used for vertex reconstruction. Since the ID typically observes a high track multiplicity from each proton–proton collision, vertex-finding algorithms must disentangle multiple interaction vertices in a single event (the so-called *pileup*). The primary vertex (or vertices) is identified by maximizing the probability that the observed ensemble of tracks originates from a common point. Secondary vertices, such as those from long-lived heavy-flavor hadrons, are searched for in a narrower subset of tracks that exhibit significant impact-parameter displacements from the primary vertex. This vertexing information is of particular importance for flavor tagging in analyses that aim to identify *b*-jets and other processes involving displaced decays.

1.6.3 Data Post-Processing

After the track-fitting and vertexing procedures, the reconstructed tracks are stored in data structures accessible by the event analysis framework. Additional algorithms perform data-quality checks to evaluate detector performance, including hit efficiencies, local alignment checks, and momentum and impact-parameter resolution studies. These validation steps are critical to ensuring that the ID performs according to design specifications and that subtle effects (such as misalignments or sensor degradation) are properly corrected in the reconstruction.

1.7 ID Tracking Performance Validation

The Inner Detector track reconstruction must balance competing demands: it should achieve high efficiency in finding genuine tracks (to avoid losing signal particles), maintain a low fake rate (so as not to degrade analyses with spurious tracks), and operate within a reasonable time budget given the enormous number of collisions recorded by ATLAS. Advances in algorithms, improved calibration, and refined alignment procedures continue to push the performance envelope, allowing the ID to operate effectively in increasingly challenging conditions, including higher instantaneous luminosities and more complex event topologies. The detailed reconstruction chain presented in the previous chapter, from hit formation and calibration, through pattern recognition and track fitting, up to vertex finding and

quality validation, lies at the heart of the ID's capability to provide high-quality momentum, vertex, and identification information for charged particles.

Identifying key regions that can be made more efficient in future upgrades, especially towards the High-Luminosity LHC era, remains an essential goal. Moreover, novel developments in the ID reconstruction algorithm have the possibility to introduce unwanted effects or drops in efficiencies in the reconstruction procedure and thus need detailed validations. To address these two goals, a new set of low-level diagnostic tools have been developed for validation accross the various stages of the track reconstruction.

Seed Maker Validation Ntuple

In the seed making step of the reconstruction chain, seeds are being produced from combinations of three spacepoints in the pixel and SCT detector modules, as described in Sec. 1.6. A new tool, called `writeSeedNtuple`, has been developed for the `SiSPSeededTrackFinder` algorithm in Athena, that writes an ntuple² with seed information like its (d_0, z_0, p_T, η) prediction as well as the (x, y, z, r) coordinates of each of the three spacepoints used to make the seeds, and other useful quantities like the event number corresponding to the seed and the dz/dr variable corresponding to top-central and central-bottom space point pairs. In order to distinguish whether a seed in the ntuple is formed from Pixel modules or SCT modules, a binary `seedType` variable is also stored in the ntuple. Moreover the written ntuple contains the p_T and η predictions of track candidates formed by each of these seeds, along with a binary `givesTrack` variable indicating if a seed is eventually used to form track candidates or is discarded. The idea behind storing this low-level information corresponding to the seed-forming stage of the ID reconstruction is to allow for detailed validations of novel and old algorithms as well as motivate new developments towards more efficient reconstruction algorithm.

The stored (r, z) coordinates of the individual SPs, for example, allows one to make correlation plots that help identify regions of hits that primarily result in track formation by using the `givesTrack` entries in the stored ntuple. Along with the seed type information stored in a `seedType` binary variable, this was used to make the Pixel detector hit plots in Fig. 1.11 using a sample dataset. They show how the distribution of SPs in the different Pixel layers influence the efficiency of the corresponding seeds to form viable preliminary track candidates. As can be seen from the two figures, majority of good quality seeds for track formation have the second SP hitting the L1 or the endcap layers. This type of plots can help identify potential cuts and selections for improved efficiency of the reconstruction algorithm particularly in seed-filtering stages, while also providing a robust debugging tool.

Another application of this tool concerns debugging of novel software releases still under development that can potentially produce unwanted effects due to faulty algorithms or incorrectly used configurations. As an example, consider the missing Pixel hits (and correspondingly more Pixel holes) observed in a development release compared to the stable Master version shown in Fig. 1.13. The new tool can be used to create an ntuple using the developmental release, with low-level information stored for a detailed debugging run. Plotting the $x - y$ correlation plots immediately identified the problem to be coming from the outermost Pixel layer, shown on the right plot in Fig. 1.12. A faulty

²An ntuple or NTuple, in the context of particle physics, is a simplified data structure that allows for straightforward access to the parameters in an event.

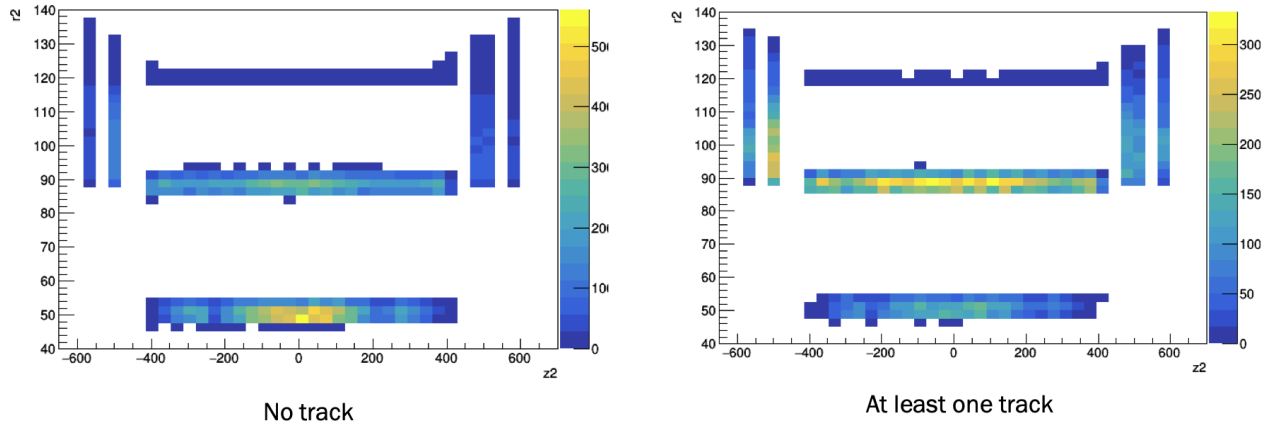


Figure 1.11: An example use of the `ntuple` created from the `writeSeedNtuple` tool. The figures show the $r - z$ correlation plots of the hits corresponding to the second spacepoint in a seed triplet inside the Pixel layer of the ID. The right figure shows the hits corresponding to spacepoints of the seeds that go on to form tracks, with the left figure corresponding to seeds that do not form tracks. The 3 layers correspond to the L0, L1 and L2 of the Pixel sub-detector. The IBL hits cannot correspond to the second spacepoint and are thus invisible in this plot. Note the increased density of SPs in the 2nd layer of the barrel region for seeds that go on to form track candidates.

configuration of the L2 layer is evident from the peculiar alternating trend in the missing pixels. This was further investigated using $z - \phi$ correlation plots with SP hits in the different Pixel layers - with the L2 layer showing the locations of the missing pixels, on the right plot of Fig. 1.14

The missing Pixel hits from a development Athena release shown in Fig. 1.13 were found to be coming from missing Pixels in the L2 layer of the Pixel sub-detector system due to a faulty configuration.

Track Finder Summary Table

A summary table that dumps all the important summary logistics of the `SiTrackMaker` tool, which is called by the `SiSPSeededTrackFinder` algorithm among others, has been ported to the newer release of Athena with support for multi-threaded offline reconstruction runs. This summary table was aimed to be helpful for tuning efforts in the ID track reconstruction algorithm, and an example of this is shown in Fig. 1.16. All the counters shown in the table have been made thread safe with the use of mutexes.

1.8 Physics Objects and Reconstructed Data

The detailed event reconstruction algorithms, briefly presented here in the case of the ID sub-system, are used to reconstruct physics objects and their corresponding four-momenta inside the ATLAS detector for subsequent use in physics analysis. Following are some details of the main physics objects used in the off-shell Higgs boson analysis in the $ZZ \rightarrow 4\ell$ decay channel, with a more analysis-specific description and corresponding pre-selections described in Sec. 5.2.

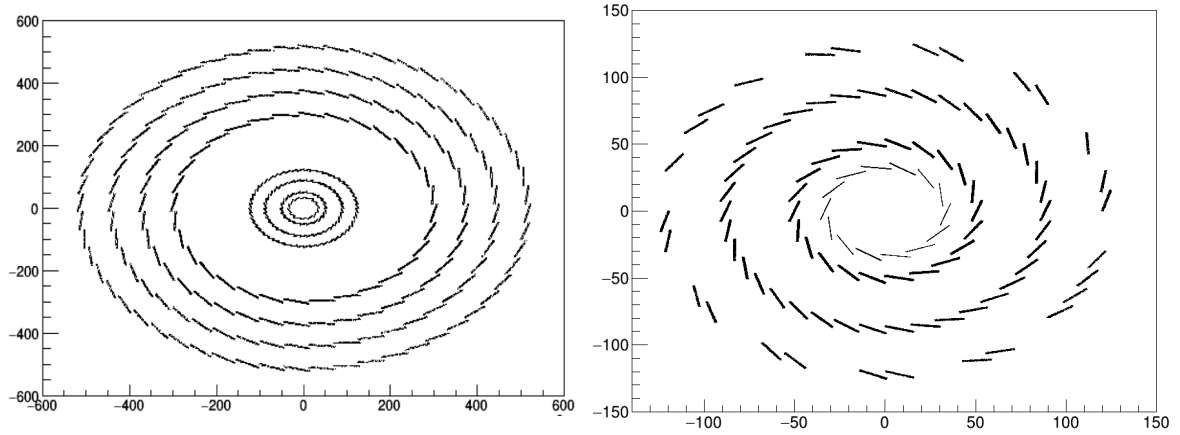


Figure 1.12: An example use of the validation ntuple made using the `writeSeedNtuple` tool. On the left, is the $x - y$ correlation plot in the barrel layer of the ID, using a stable software release. On the right, is the same plot focused on the Pixel layers, where a problematic configuration in a modified Athena release caused issues with the outermost L2 pixel layer having invisible alternating sensors. The latter resulted in strange behavior in downstream validations, shown in Fig. 1.13, which was further investigated and then resolved using this ntuple tool, showing its versatility in both development and as a debugging tool.

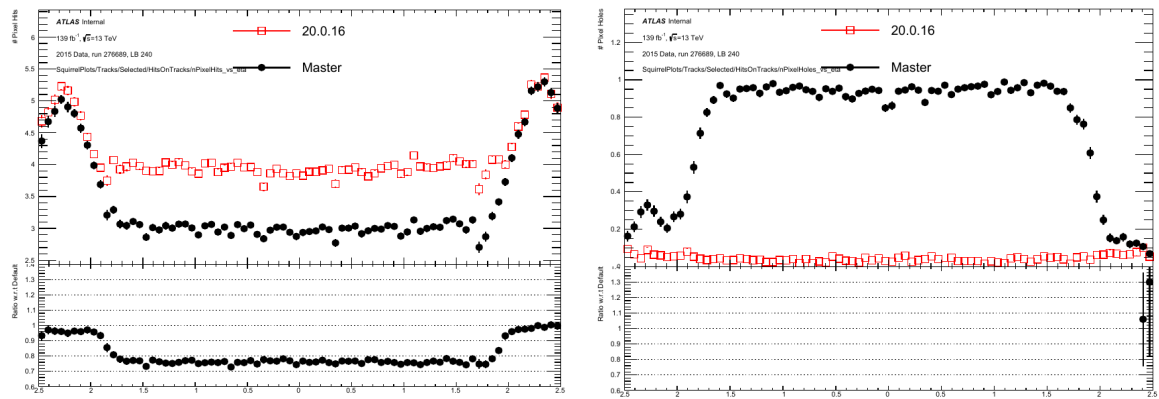


Figure 1.13: An example of a discrepancy between a development release (20.0.16) of Athena and the Master branch. The left plot shows the number of Pixel hits, and the right plot shows the number of holes as a function of pseudo-rapidity η . The bottom plots are ratio plots between the development release and the Master branch. On average, one pixel hit per track seems to be missing (plot on the left) in the barrel, and instead a hole appears (plot on the right). To investigate the origins of this discrepancy, the new validation `writeSeedNtuple` tool was used, as is shown in Figs. 1.14 and 1.12. This is just an example issue to showcase the versatility of the tool in general.

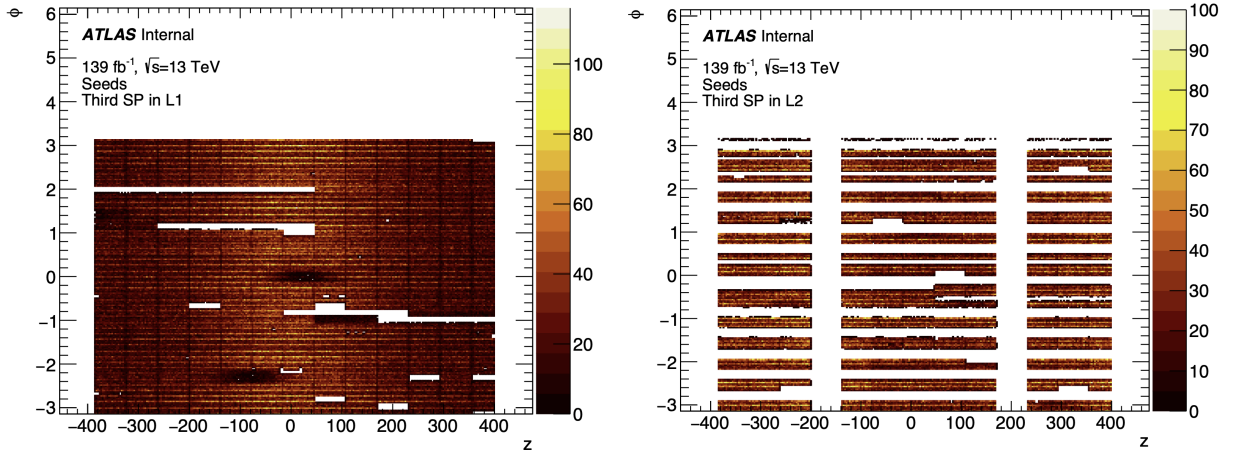


Figure 1.14: An example use of the validation ntuple made using the `writeSeedNtuple` tool. The $z-\phi$ correlation plots of the hits in the L2 sub-layer (figure on the right) seems to show missing pixels in alternate locations of the L2, as is also clear from the right plot in Fig. 1.12. For comparison the left plot shows the same $z-\phi$ with the third SP in a seed located in the L1 layer, with fewer missing pixels. The black shadows in the figure on the left can be further explained using the new tool, as shown in Fig. 1.15.

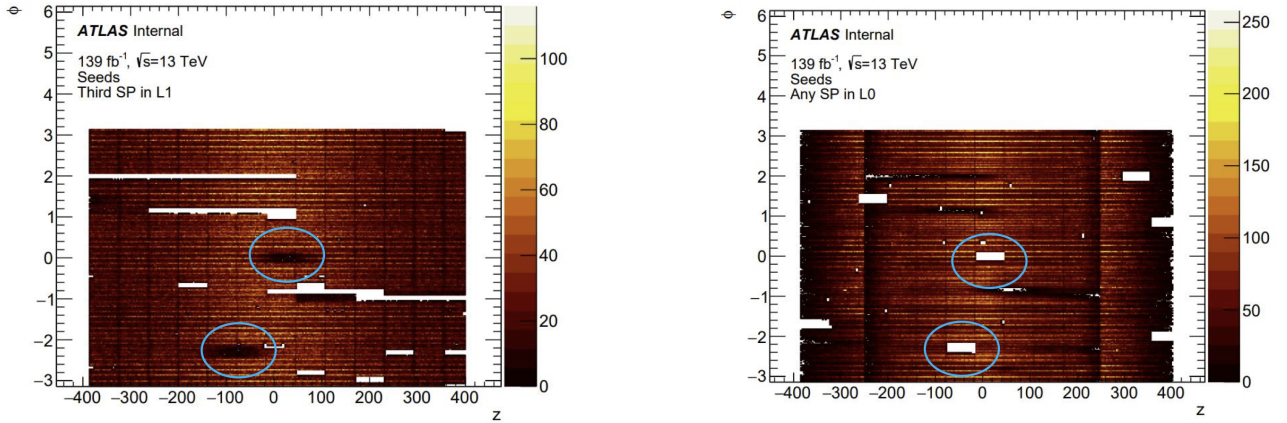


Figure 1.15: An example use of the validation ntuple made using the `writeSeedNtuple` tool. A problematic software release was diagnosed by investigating the peculiarities in the $r-\phi$ correlation plot at the SP formation level. The figure on the left corresponds to the third SP in L1, forcing the seed to have IBL+L0+L1 topology. These selections are also done using the information stored in the ntuple written by the tool. The black shadows (blue circles) in the left plot correspond to a lack of hits and potentially indicate dead pixels on the layers below. The figure on the right, with SP hits in the L0 layer was made to confirm this (blue circles surrounding missing pixels) and subsequently used to debug the software release.

Kind of seed	PPP	PPS	PSS	SSS	ALL	
Input seeds	79449	0	0	103834	183283	
No track parameters	949	0	0	5187	6136	
Used seeds	64129	0	0	71892	136021	
Used seeds brem	2411	0	0	2623	5034	
Det elements in road	45.11	0	0	44.042	44.545	
Two clusters on DE	0	0	0	502	502	
Wrong DE road	0	0	0	21	21	
Wrong initialization	800	0	0	2148	2948	
Can not find track	55989	0	0	50596	106585	
It is not new track	1844	0	0	1792	3636	
Attempts brem model	1477	0	0	0	1477	
Output tracks	5820	0	0	17556	23376	
Output extra tracks	592	0	0	1501	2093	
Output tracks brem	316	0	0	0	316	
Seeds with track	5228	0	0	16055	21283	Number seeds
Track/Used 0.0-0.5	0.06204	0	0	0.23	0.1744	3382
0.5-1.0	0.06427	0	0	0.2198	0.1637	3758
1.0-1.5	0.08226	0	0	0.2325	0.1671	4481
1.5-2.0	0.06111	0	0	0.2279	0.1384	4164
2.0-2.5	0.1166	0	0	0.2054	0.1555	5245
2.5-3.0	0.0533	0	0	0.296	0.08325	253
3.0-3.5	0	0	0	0	0	0
3.5-4.0	0	0	0	0	0	0

Figure 1.16: This table is used to dump all the relevant logistics of the track finding stage. It contains useful statistics extracted mainly from the SiTrackFinder and SiCombinatorialTrackFinder tools. The PPP, PPS, PSS and SSS stand for Pixel-Pixel-Pixel seed, Pixel-Pixel-SCT seed, and so on. DE stands for detector elements and the term road stands for the 3D roads built during the track reconstruction stage. The term *brem model* stands for bremsstrahlung recovery model used in electron reconstruction. The Track/Used row refers to the fraction of seeds that result in track formation compared to the total number of seeds, as a function of the pseudo-rapidity η .

- **Electrons/Photons:** Charged particles like electrons leave the first hits in the ID, subsequently being absorbed by the electromagnetic calorimeter. The reconstructed tracks in the ID, using the algorithms described in the sections above, are matched to the energy deposits in the calorimeter for a robust electron object reconstruction. Photons do not play a direct role in the off-shell Higgs boson analysis, but can sometimes convert into an electron-positron pair via interactions with the detector material - leaving tracks in the ID.
- **Muons:** Muons are also charged particles that first interact with and leave tracks in the ID. Muons, owing to their heavier masses, are not stopped inside the electromagnetic calorimeter and propagate the entire length of the detector to the outermost MS detector sub-system. They are primarily identified by tracks or segments reconstructed in the MS and matched to the tracks reconstructed in the ID.
- **Jets:** Jets are collimated sprays of particles formed as a result of hadronization of color-charged or hadronic particles in the final state as they traverse through the detector. Jets used in the off-shell Higgs boson analysis are reconstructed from particle-flow objects [23] using the anti- k_t algorithm [24, 25] with radius parameter $R = 0.4$. The jet-energy scale is calibrated using simulation and further corrected with in situ methods [26].

Chapter 2

Neural Simulation-Based Inference

This chapter starts with a review of the general statistical framework for parameter measurements at the LHC for completeness, followed by a discussion on general limitations of the traditional inference techniques commonly used. A new inference framework using Neural Simulation-Based Inference techniques is then developed, followed by a cross-section measurement example. The outlook of the new formalism, including some simplifications, will also be discussed.

2.1 Review of the Statistical Framework

The probability density for an inelastic process $pp \rightarrow A$ in a particle collision experiment is characterized by the n -dimensional inclusive differential cross-section $d\sigma_{pp \rightarrow A}/d\mathbf{x}$ of the observed final state of A as a function of the reconstructed observables $\mathbf{x} = (x_1, x_2, \dots, x_n)$ measured by the ATLAS detector, constituting the physics objects and their corresponding track kinematics described briefly in Chap. 1:

$$p(\mathbf{x}|\boldsymbol{\theta}) = \frac{1}{\sigma_{pp \rightarrow A}(\boldsymbol{\theta})} \frac{d\sigma_{pp \rightarrow A}}{d\mathbf{x}}(\mathbf{x}|\boldsymbol{\theta}) \quad (2.1)$$

where $\boldsymbol{\theta}$ is a parameter which represents an underlying physics hypothesis. This probability density in Eq. 2.1 is normalized using the total cross-section of the particular final state A :

$$\int \frac{d\sigma_{pp \rightarrow A}}{d\mathbf{x}}(\mathbf{x}|\boldsymbol{\theta}) d\mathbf{x} = \sigma_{pp \rightarrow A}(\boldsymbol{\theta}) \quad (2.2)$$

The aim of a typical analysis is to fit the parameter $\boldsymbol{\theta}$ values given the observed set of independent and identically distributed (i.i.d) data $\mathcal{D}_{obs} = \{\mathbf{x}_i\}$ at the ATLAS detector. Given a dataset with $\mathcal{N}_{data}$ events, the *joint* probability model is then given by:

$$\begin{aligned} p(\mathcal{D}|\boldsymbol{\theta}) &= \text{Pois}(\mathcal{N}_{data}|\nu_{pp \rightarrow A}(\boldsymbol{\theta})) \prod_i^{\mathcal{N}_{data}} p(\mathbf{x}_i|\boldsymbol{\theta}) \\ &= \text{Pois}(\mathcal{N}_{data}|\nu_{pp \rightarrow A}(\boldsymbol{\theta})) \prod_i^{\mathcal{N}_{data}} \frac{1}{\sigma_{pp \rightarrow A}} \frac{d\sigma_{pp \rightarrow A}}{d\mathbf{x}}(\mathbf{x}_i|\boldsymbol{\theta}) \end{aligned} \quad (2.3)$$

where $\nu_{pp \rightarrow A}(\boldsymbol{\theta}) = \sigma_{pp \rightarrow A}(\boldsymbol{\theta}) \cdot L$ is the expected number of inelastic processes $pp \rightarrow A$ written as a function of the integrated luminosity $L = \int \mathcal{L} \cdot dt$ and Poisson refers to the Poisson distribution function. The Poisson term in Eq. 2.3 is used to account for the variability in the total number of observed events which itself depends on the parameters $\boldsymbol{\theta}$.

In principle, all the collision events $\mathbf{x}_i$ recorded and processed by the experiment could be used for inference in Eq. 2.3. In practice, the number of recorded collisions can exceed 10^{10} , and only a small subset carry significant information about a given process. The events that are effectively used for parameter inference as said to be *pre-selected*. The details of the pre-selection used in the analysis of the off-shell Higgs boson presented in this thesis will be given in Sec. 5.2.

The total cross-section $\sigma_{pp \rightarrow A}$ can be broken down as a sum of cross-sections from the various signal and background processes that contribute to the final state of interest:

$$\sigma_{pp \rightarrow A} = \sum_s \sigma_{pp \rightarrow A}^s \quad (2.4)$$

This gives us the general probability model used in LHC analysis:

$$p(\mathcal{D}|\boldsymbol{\theta}) = \text{Pois} \left(\mathcal{N}_{\text{data}} \middle| \sum_s \nu_{pp \rightarrow A}^s(\boldsymbol{\theta}) \right) \times \prod_i^{\mathcal{N}_{\text{data}}} \left[\sum_s \frac{1}{\sigma_{pp \rightarrow A}^s(\boldsymbol{\theta})} \frac{d\sigma_{pp \rightarrow A}^s}{d\mathbf{x}}(\mathbf{x}_i|\boldsymbol{\theta}) \right] \quad (2.5)$$

The probability density Eq. 2.5 gives us a *joint* model for the observation of i.i.d. events $\mathcal{D} = \{\mathbf{x}\}$ given some underlying hypothesis $\boldsymbol{\theta}$, which can then be used to perform a composite hypothesis test on the parameters $\boldsymbol{\theta}$. The probability density function in Eq. 2.5 is the same as the likelihood function to be used for statistical inference, if the input data $\mathcal{D}$ is fixed to be either the Asimov (expected) or the observed dataset, and float the $\boldsymbol{\theta}$ parameters. The terms likelihood and probability density will thus be used interchangeably in this thesis.

2.1.1 Modelling in the presence of uncertainties

The parameter vector $\boldsymbol{\theta}$ consists of not just the parameters to be measured but also *nuisance parameters* (NPs) that affect the model but are not used in the hypothesis testing. As such, the full parameter vector $\boldsymbol{\theta}$ is generally broken down into two components:

$$\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\alpha}) \quad (2.6)$$

where $\boldsymbol{\mu}$ refers to the measured parameters of interest (POI) whereas $\boldsymbol{\alpha} = (\alpha_m)$ is a vector with NPs. Here $\boldsymbol{\mu}$ can either represent a single parameter or a vector of parameters, and the formalism accommodates both cases. The NPs are introduced to quantify the uncertainty in the simulation model of Eq. 2.5. Each part of the simulation chain shown in Fig. 2.1 contributes to the full uncertainty model parameterized using $\boldsymbol{\alpha}$. For example, the hard-scattering and parton-shower calculations have theoretical modelling uncertainties due to the missing higher orders in the perturbative QCD calculations. There are also a range of experimental uncertainties associated with the detector limitations such as the limited physics object reconstruction efficiencies, uncertainty in the measured energy scale

and resolution as well as the uncertainty in the overall recorded luminosity during the LHC operation. Some of the uncertainties relevant to the off-shell Higgs measurements will be discussed in Sec. 4.6.

Each parameter α_m of the vector $\boldsymbol{\alpha}$ is associated with a particular uncertainty. For experimental uncertainties, dedicated calibration measurements are performed using independent datasets to constrain the possible values of the NPs, enabling a precise measurement of the POIs $\boldsymbol{\mu}$ using the dataset of interest. In the case of theoretical uncertainties, where dedicated measurements are not possible, prior assumptions are used to constrain the corresponding NPs. In principle, the full model from the auxiliary measurements can be fit alongside the nominal analysis model. Although, in practice only the maximum likelihood estimate a_m , also known as the *global observable* or in this work the *auxiliary observable* (AO), and a standard error δ_m is extracted from the auxiliary measurements. This will be discussed in more details in Chap. 3. These quantities are then used to build a simplified probability model of the auxiliary measurement approximated as a Gaussian,

$$p_{\text{aux}}(a_m|\alpha_m) = \frac{1}{\sqrt{2\pi}} \exp \left[-\frac{(\alpha_m - a_m)^2}{2\delta_m^2} \right]. \quad (2.7)$$

This auxiliary model often undergoes further simplification by a standard change of variables $(\alpha_m - a_m)/\delta_m \rightarrow \alpha_m$, to convert to a unit Gaussian. Assuming that there are N_{syst} constrained NPs affecting the measurement, and that each source of uncertainty is chosen to be independent of the rest, the constraint term in the likelihood is written simply as,

$$L_{\text{aux}}(\boldsymbol{\alpha}|\mathcal{G}) = \prod_{p=1}^{N_{\text{syst}}} \frac{1}{\sqrt{2\pi}} \exp \left[-\frac{\alpha_m^2}{2} \right] \quad (2.8)$$

where $\mathcal{G} = \{a_m\}$ is the dataset of global observables, consisting of one auxiliary measurement per element α_m of the full NP vector $\boldsymbol{\alpha}$.

2.1.2 Profile Negative Log-Likelihood Test Statistic

The Neyman-Pearson lemma [27] provides a framework for *simple* hypothesis testing that minimizes the probability of Type II error (false negative) for a fixed probability of Type I error (false positive). It states that the most powerful test for a given significance level is based on the likelihood ratio. In practice, for the typical *composite* hypothesis tests on continuous parameters employed in LHC experiments, the profile likelihood ratio formalism is used:

$$\frac{L_{\text{phys}}(\boldsymbol{\mu}, \hat{\boldsymbol{\alpha}}|\mathcal{D}, \mathcal{G})}{L_{\text{phys}}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\alpha}}|\mathcal{D}, \mathcal{G})} = \frac{e^{-\nu(\boldsymbol{\mu}, \hat{\boldsymbol{\alpha}})} \nu(\boldsymbol{\mu}, \hat{\boldsymbol{\alpha}})^{\mathcal{N}_{\text{data}}}}{e^{-\nu(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\alpha}})} \nu(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\alpha}})^{\mathcal{N}_{\text{data}}}} \prod_{i=1}^{\mathcal{N}_{\text{data}}} \frac{p(\mathbf{x}_i|\boldsymbol{\mu}, \hat{\boldsymbol{\alpha}})}{p(\mathbf{x}_i|\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\alpha}})} \quad (2.9)$$

where $\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\alpha}}$ are the values of $\boldsymbol{\mu}, \boldsymbol{\alpha}$ which maximizes the likelihood function $L(\boldsymbol{\mu}, \boldsymbol{\alpha})$, and $\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\alpha}}(\boldsymbol{\mu})$ is the value of NPs $\boldsymbol{\alpha}$ that locally maximizes the likelihood function $L(\boldsymbol{\mu}, \boldsymbol{\alpha})$ for a fixed value of $\boldsymbol{\mu}$. This profiling of NPs increases the uncertainty on the POIs $\boldsymbol{\mu}$ in the test, reflecting loss of information due to the uncertain parameters. This uncertainty is reduced by introducing L_{aux} that constrains the NPs $\boldsymbol{\alpha}$ using auxiliary calibration measurements, giving the full likelihood model,

$$L(\boldsymbol{\mu}, \boldsymbol{\alpha}) = L_{\text{phys}}(\boldsymbol{\mu}, \boldsymbol{\alpha} | \mathcal{D}, \mathcal{G}) \cdot L_{\text{aux}}(\boldsymbol{\alpha} | \mathcal{G}). \quad (2.10)$$

Maximizing the likelihood ratio is mathematically equivalent to minimizing the negative of the log of the likelihood ratio. The test statistic used for hypothesis testing is given by

$$\begin{aligned} t_{\boldsymbol{\mu}} &= -2 \log \left(\frac{L_{\text{phys}}(\boldsymbol{\mu}, \hat{\boldsymbol{\alpha}} | \mathcal{D}, \mathcal{G})}{L_{\text{phys}}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\alpha}} | \mathcal{D}, \mathcal{G})} \cdot \frac{L_{\text{aux}}(\hat{\boldsymbol{\alpha}} | \mathcal{G})}{L_{\text{aux}}(\hat{\boldsymbol{\alpha}} | \mathcal{G})} \right) \\ &= -2 \mathcal{N}_{\text{data}} \left(\log \frac{\nu(\boldsymbol{\mu}, \hat{\boldsymbol{\alpha}})}{\nu(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\alpha}})} \right) + 2 \left(\nu(\boldsymbol{\mu}, \hat{\boldsymbol{\alpha}}) - \nu(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\alpha}}) \right) - 2 \sum_i^{\mathcal{N}_{\text{data}}} \log \frac{p(\mathbf{x}_i | \boldsymbol{\mu}, \hat{\boldsymbol{\alpha}})}{p(\mathbf{x}_i | \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\alpha}})} + \sum_p^{N_{\text{syst}}} \left(\hat{\boldsymbol{\alpha}}^2 - \hat{\boldsymbol{\alpha}}^2 \right) \end{aligned} \quad (2.11)$$

While the Neyman-Pearson lemma does not strictly apply to composite hypothesis tests, the profile negative log-likelihood test statistic saturates the Cramér-Rao bound, giving the minimum possible variance on the POIs $\boldsymbol{\mu}$ asymptotically [28]. As such, this is the most common test statistic used at the LHC for parameter inference. The full procedure of building confidence intervals using this test statistic is defined in Sec. 3.3. Estimating the test statistic exactly is not straightforward and several ad-hoc techniques have been developed over the years to this end. This will be discussed in Sec. 2.3. The focus of this thesis is building Eq. 2.11 as optimally and accurately as possible, using modern deep-learning techniques as described in Sec. 2.4.

2.2 Building the Probability Model

If the full analytical model $p(\mathbf{x} | \boldsymbol{\theta})$ of Eq. 2.5 were available the optimal test statistic shown in Eq. 2.11 can be easily built from first principles. However, it is impossible to analytically calculate the detector-level cross sections of a process $pp \rightarrow A$ in the LHC experiments. What can be calculated is the partonic, or the hard-scattering, cross section $d\sigma_{ab \rightarrow A}/d\mathbf{k}$ as a function of Born-level variables, call it $\mathbf{k}$. These partonic final state objects $\{\mathbf{k}\}$ radiate additional partons before undergoing hadronization to give final observable objects inside the detector. This cascading process is driven by non-perturbative (or the so-called soft-) Quantum Chromodynamics (QCD).

These processes, shown using a cartoon in Fig. 2.1, can only be modelled using Markov chain Monte Carlo techniques using a combination of phenomenological and data-driven inputs. The resulting final state objects then undergo simulations to model the reaction with the detector electronics producing a $O(100M)$ dimensional vector corresponding to the total number of sensors - adding to the complexity of the final state which can now only be modelled using Monte Carlo integration. The full probability model used for simulations can be schematically expressed as:

$$p(\mathbf{x} | \boldsymbol{\theta}) = \frac{1}{\sigma(\boldsymbol{\theta})} \int f_a(X_a) f_b(X_b) W(\mathbf{x} | \mathbf{y}) H(\mathbf{y} | \mathbf{k}) \frac{d\sigma_{ab \rightarrow A}}{d\mathbf{k}}(\mathbf{k} | \boldsymbol{\theta}) dX_a dX_b d\mathbf{y} d\mathbf{k}, \quad (2.12)$$

where f_a and f_b are parton density functions of the initial state partons, H encodes all shower and non-perturbative effects simulated by PYTHIA [30] or similar Monte Carlo software, and W describes

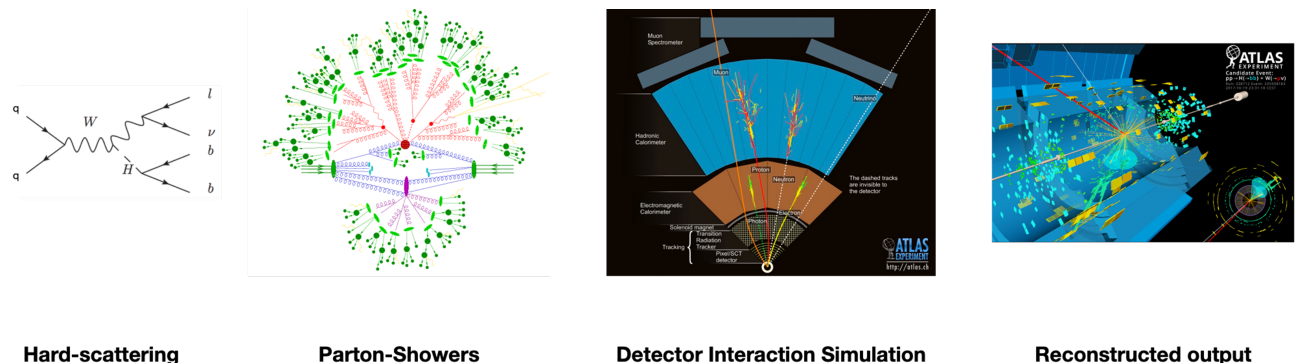


Figure 2.1: From left to right: a) An example Feynman diagram of parton-parton hard scattering of interest. One of the many diagrams typically simulated using perturbative QCD, and summed over to produce final cross-sections associated with a given final state b) Sketch of the full hadron-hadron collision as simulated by a Monte-Carlo event generator. The red blob in the center represents the hard collision, generated using the first Feynman diagram for example, surrounded by a tree-like structure representing Bremsstrahlung as simulated by parton showers. The purple blob indicates a secondary hard scattering event. Parton-to-hadron transitions are represented by light green blobs, dark green blobs indicate hadron decays, while yellow lines signal soft photon radiation. Figure taken from Ref. [29]. c) A simulation of the various particles interacting with the different sub-detector systems in the ATLAS detector. Figure taken from Ref. [18]. d) An event display of the final state, also showing reconstructed objects like muons and jets.

the detector response simulated by the detector simulation software GEANT [31]. A condensed version of Eq. 2.12 can be written as:

$$p(\mathbf{x}|\boldsymbol{\theta}) = \frac{1}{\sigma(\boldsymbol{\theta})} \int \frac{d\sigma_{pp \rightarrow A}}{d\mathbf{x} d\mathbf{z}}(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) d\mathbf{z}, \quad (2.13)$$

where the variables $\mathbf{z}$ are marginalized over to produce the observed final state. These are often called *latent variables*. In complex simulations like the one used in ATLAS, the latent space can have millions of variables which are integrated over: some are physical degrees of freedom (such as intermediate particles like individual hadrons in jets, or the digitized detector readout) while others are present just for stochastic calculation purposes. While this Monte Carlo model can be used to do forward simulations of the detector output, it cannot take a value $\mathbf{x}$ and then analytically calculate $p(\mathbf{x}|\boldsymbol{\theta})$. As a result, the inference on $\boldsymbol{\theta}$ from the detector output $\mathbf{x}$ is a difficult task.

2.2.1 Measurements within the Fiducial volume

The detectors at LHC experiments are not designed to cover the full solid angle surrounding the particle collision and thus have a limited *fiducial* volume. Moreover, the detector components do not have perfect resolutions and efficiencies in detecting outgoing particles. Taking into account the efficiency and acceptance of the ATLAS detector (also referred to as the fiducial volume), only the differential event rate can be directly measured in an analysis:

$$p(\mathbf{x}|\boldsymbol{\theta}) = \frac{1}{\nu(\boldsymbol{\theta})} \frac{d\nu_{pp \rightarrow A}}{d\mathbf{x}}(\mathbf{x}|\boldsymbol{\theta}), \quad (2.14)$$

where $\nu_{pp \rightarrow A} = \sigma_{pp \rightarrow A} \cdot \mathcal{L} \cdot \mathcal{A} \cdot \varepsilon$ is the number of expected events within the ATLAS detector volume taking into account the detector efficiency ε and the acceptance $\mathcal{A}$ in the fiducial volume of the ATLAS detector. From here on, $p(\mathbf{x}|\boldsymbol{\theta})$ will refer to the differential rate of events in the fiducial phase space as defined in Eq. 2.14 unless otherwise specified.

Goal of an analysis:

Estimate the probability model $p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\alpha})$ to be the best approximation of Eq. 2.14, to be then used in the test statistic function of Eq. 2.11 for a composite frequentist hypothesis test.

In the sections below, the traditional strategy to estimate $p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\alpha})$ will be reviewed along with its limitations, followed by the new strategy using Neural Simulation-Based Inference technique. The latter will be followed with a specific example of signal-strength measurements.

2.3 Standard Inference Strategy

Ideally, having a simulation model is sufficient to build a proxy for the probability distribution $p(\mathbf{x}|\boldsymbol{\theta})$ by sampling enough events $\mathbf{x} \sim p(\mathbf{x}|\boldsymbol{\theta})$ and binning them in the multi-dimensional space of $\mathbf{x}$. In practice this is unfeasible in LHC experiments due to the so-called *curse of dimensionality*, i.e. the number of events needed to be sampled to build a good histogram representation of $p(\mathbf{x}|\boldsymbol{\theta})$ grows exponentially with each additional dimension.

Both the parameter vector $\boldsymbol{\theta}$ and the final state observables $\mathbf{x}$ in Eq. 2.14 are vectors with a large number of dimensions. The dimension of the vector $\mathbf{x}$ can be as large as $\sim O(100M)$ - roughly the number of independent sensor elements in the ATLAS detector. But the collaboration has developed reconstruction algorithms [32] that are used to convert the raw sensor output from the ATLAS detector to a set of physics objects and their corresponding kinematics, charges and flavours. The resulting dimensionality, however reduced relative to the original output, is still large enough to make the direct modelling of $p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\alpha})$ using histograms unfeasible.

2.3.1 Summary Observables

The standard procedure then is to create a one-dimensional (1D) summary observable $\mathbf{x} \rightarrow u(\mathbf{x})$ from the simulated multi-dimensional output $\mathbf{x} \sim p(\mathbf{x}|\boldsymbol{\theta})$, that optimally discriminates signal processes of interest from the large backgrounds. Take as an example, the reconstructed invariant mass $m_{4\ell}(\mathbf{x})$ observable from the 4-lepton final state in the $H \rightarrow ZZ$ channel that was used by ATLAS to discover the Higgs boson in 2012.

This is an example of physics-motivated observables where domain knowledge is used to project the multi-dimensional detector output $\mathbf{x}$ to a 1D summary observable $m_{4\ell}(\mathbf{x})$ containing relevant physics information that optimally searches for the signal - in this case the Higgs boson resonance. This type of physics-motivated observables have the added advantage of offering direct physical insight; the resonance peak observed in Fig. 2.2a offers direct insight to the peak mass value at which the new resonance appears.

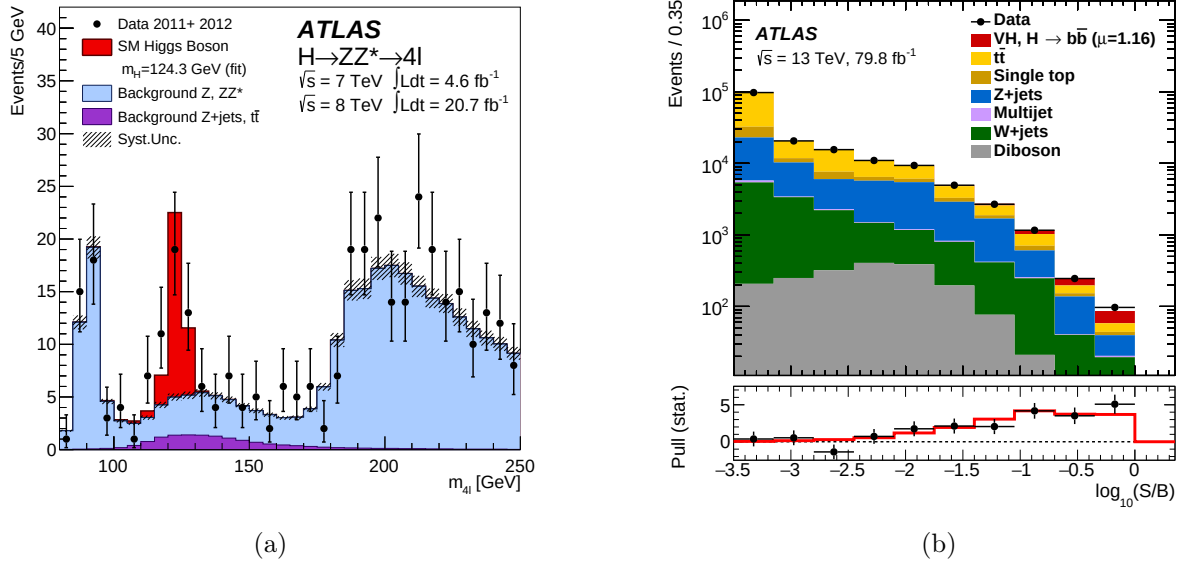


Figure 2.2: (a) The distribution of the four-lepton invariant mass, $m_{4\ell}$, used to identify the Higgs boson resonance peak during the first discovery announced in 2012. (b) An example of the standard inference strategy - building observables that optimally discriminate signal from background.

The corresponding observable is then binned, with the likelihood of an event in a bin $I(\mathbf{x}|u)$ given simply by [33]:

$$p(\mathbf{x}_i|\boldsymbol{\theta}) = \frac{1}{\nu(\boldsymbol{\theta})} \frac{\nu_{I(\mathbf{x}_i|u)}(\boldsymbol{\theta})}{\Delta_{I(\mathbf{x}_i|u)}}. \quad (2.15)$$

where $I(\mathbf{x}|u)$ is the function that returns the bin index of the binned $u(\mathbf{x})$ observable associated with an event $\mathbf{x}$, $\Delta_{I(\mathbf{x}_i|u)}$ corresponds to the width of the bin and $\nu_{I(\mathbf{x}_i|u)}$ is the yield in the bin associated with index $I(\mathbf{x}|u)$. This gives us a simplified, or binned, version of the probability density function defined in Eq.2.3:

$$p(\mathcal{D}|\boldsymbol{\theta}) = \text{Pois}(\mathcal{N}_{\text{data}}|\nu(\boldsymbol{\theta})) \prod_i^{\mathcal{N}_{\text{data}}} \frac{1}{\nu(\boldsymbol{\theta})} \frac{\nu_{I(\mathbf{x}_i|u)}(\boldsymbol{\theta})}{\Delta_{I(\mathbf{x}_i|u)}} \quad (2.16)$$

where $\mathcal{N}_{\text{data}}$ is the observed number of events used to fit the parameters $\boldsymbol{\theta}$. This *simplified* model can be used to build the test statistic of Eq. 2.11. The unbinned expression of Eq. 2.16 is equivalent to a product of Poisson probability ratios for each bin, up to a constant offset [33]:

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{\text{bin}}^{\mathcal{N}_{\text{bins}}} \frac{e^{-\nu_{\text{bin}}} \cdot \nu_{\text{bin}}(\boldsymbol{\theta})^{\mathcal{N}_{\text{bin}}}}{\mathcal{N}_{\text{bin}}!} \quad (2.17)$$

where $\mathcal{N}_{\text{bin}}$ corresponds to the observed number of events in the bin of the summary statistic $u(\mathbf{x})$ and $\nu_{\text{bin}}(\boldsymbol{\theta})$ the expected number of events in the bin corresponding to a hypothesis $\boldsymbol{\theta}$. This is the familiar expression for the joint probability density model used in LHC analysis. A typical workflow using this simplified model with a summary observable $u(\mathbf{x})$ is shown in Fig. 2.3, using a randomly

chosen measurement using the ATLAS experiment, which will not be discussed further in this thesis.

The test statistic from the probability model approximation of Eq. 2.17 will have an unbiased estimator $\hat{\boldsymbol{\theta}}$, despite being not an exact description. Note that the probability model relies on simulations for an estimate of the parameterized yields $\nu_{\text{bin}}(\boldsymbol{\theta})$, even during the inference step. This is a feature of the standard inference strategy with Poisson likelihoods that is distinct from a more direct inference strategy where simulations are used to train the full likelihood (ratio) model $p(\mathcal{D}|\boldsymbol{\theta})$, which is then used as a direct surrogate to perform inference on new data. The latter is referred to as amortized [1] approach to inference and will be discussed more in Sec. 3.1.2.

2.3.2 Limitations of Standard Techniques

There are three possible limitations to the simplification of the full model of Eq. 2.14 to the approximation of Eq. 2.16 using binned summary observable:

1. The dimensional reduction of the full final-state information to a single-dimensional summary observable $\mathbf{x} \rightarrow u(\mathbf{x})$ can result in a loss of information.
2. The summary statistic $u(\mathbf{x})$ is constant in Eq. 2.16 for the full scan across the parameter space. Even if the observable is carefully designed to be locally optimal at certain parameter values, it might lose optimality for parameter values that change the physics significantly.
3. All the events in a bin have the same probability. Information from rare possibly signal-like events can be lost in a binned approximation.

While in most cases these limitations can be addressed using a careful choice of the summary observable and a binning optimization, in several cases the limitations cannot be resolved without a restructuring of the analysis techniques.

For example, in a simple signal versus background hypothesis testing, training a NN to optimally discriminate signals from backgrounds and using its output to estimate the probability ratio results in an optimal summary statistic, which minimizes Type II errors (false negatives) for a fixed Type I error rate (false positives). This is motivated by the Neyman-Pearson lemma, which establishes that the likelihood ratio,

$$u(\mathbf{x}) = \frac{p_S(\mathbf{x})}{p_B(\mathbf{x})},$$

is the most powerful test statistic for binary hypothesis testing, and hence can also be considered an optimal summary observable. In general, ATLAS analysis do *composite* hypothesis testing to scan over a range of continuous parameter values. This makes the optimal discriminant dependent on the parameter being scanned:

$$u_{\text{optimal}}(\mathbf{x}|\boldsymbol{\mu}) \approx \frac{p(\mathbf{x}|\boldsymbol{\mu})}{p(\mathbf{x}|\boldsymbol{\mu}')} \quad (2.18)$$

where $\boldsymbol{\mu}'$ is the nominal hypothesis. The Neyman-Pearson lemma applies to simple hypothesis testing and thus its interpretation in the context of composite hypothesis testing, particularly in the presence of nuisance parameters (which are suppressed here), is not as robust. **Asymptotically, however,**

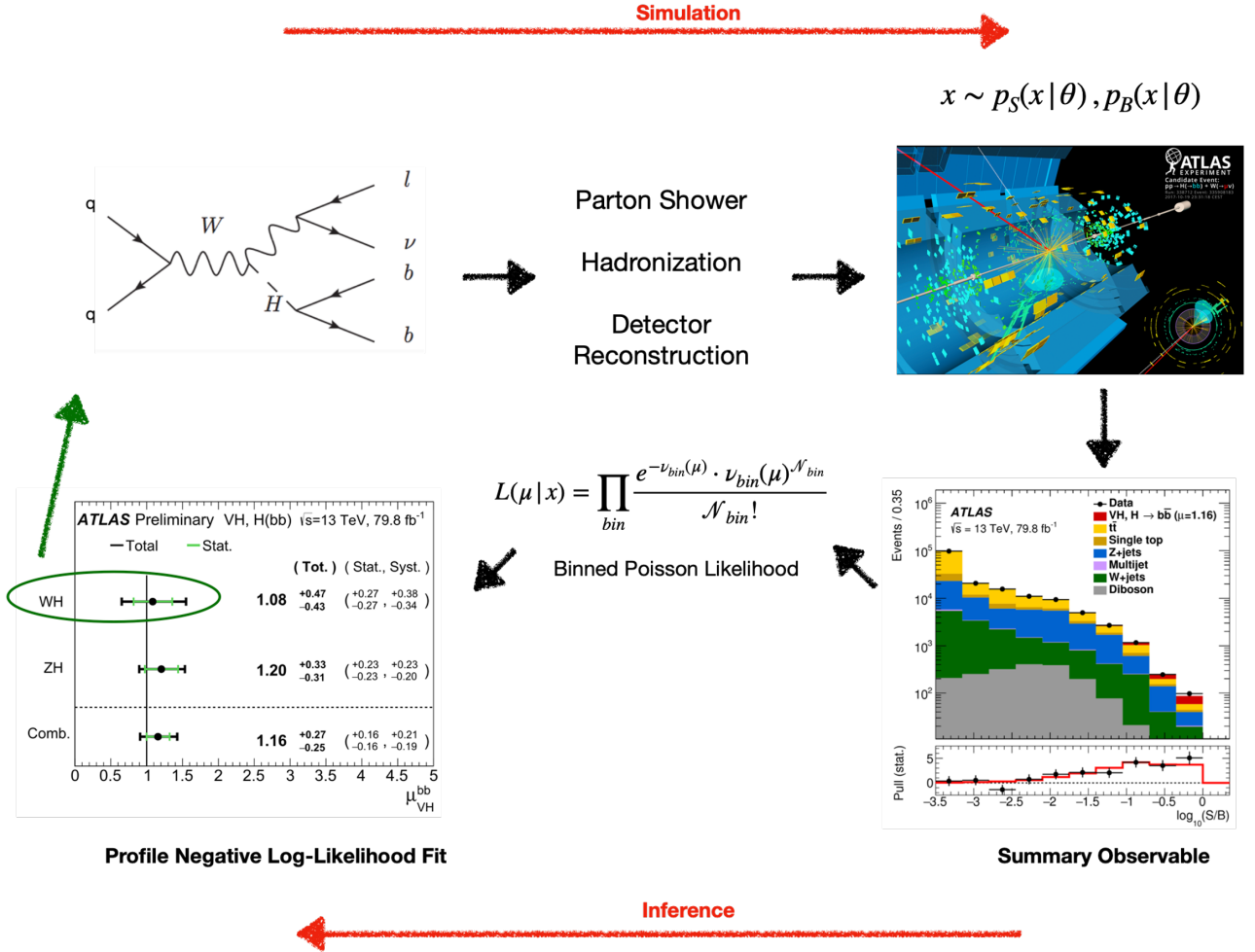


Figure 2.3: An example of a standard inference workflow used for measurements at the LHC. Using perturbative methods, a physics process of interest is simulated using Monte Carlo techniques, and then undergo parton showering, hadronization and detector reconstruction to produce the final detector-level output $\mathbf{x} \sim p(\mathbf{x}|\mu)$. Same technique is used to build the full model as a function of some parameter of interest μ . A powerful observable discriminating signal from background is then binned and a Poisson likelihood fit is performed to infer the parameter of interest μ , giving us insight in the physics process of interest. The particular observable and inference shown in the figure is from the measurement of the VH production in Ref. [34].

the likelihood ratio saturates the Cramér-Rao minimum variance bound and is thus considered in this thesis as an optimal observable. In certain cases like the measurement of the cross-section of the signal process, the parameterized density ratio often reduces to:

$$u_{\text{optimal}}(\mathbf{x}|\boldsymbol{\mu}) = \frac{\sum_S f_S(\boldsymbol{\mu}) \cdot p_S(\mathbf{x}) + \sum_B p_B(\mathbf{x})}{\sum_B p_B(\mathbf{x})} \quad (2.19)$$

This is further simplified if the several signal probabilities have a parameter dependence that scales the signal probabilities uniformly:

$$u_{\text{optimal}}(\mathbf{x}|\boldsymbol{\mu}) = f(\boldsymbol{\mu}) \cdot \frac{\sum_S p_S(\mathbf{x})}{\sum_B p_B(\mathbf{x})} + \frac{\sum_B p_B(\mathbf{x})}{\sum_B p_B(\mathbf{x})} \quad (2.20)$$

In cases where this simplification is possible, a *fixed* observable

$$u_{\text{optimal}}(\mathbf{x}|\boldsymbol{\mu}) \sim u(\mathbf{x}) = \frac{\sum_S p_S(\mathbf{x})}{\sum_B p_B(\mathbf{x})}, \quad (2.21)$$

is an asymptotically optimal summary statistic, provided the probability densities $p_c(\mathbf{x})$ are well modeled using the full multi-dimensional phase space, and can thus be used to optimally scan over the entirety of the parameter space. An example of this is cross-section measurements where the parameter of interest is a signal strength measurement, as will be discussed in Sec. 2.5.

In measurements of complex, non-linear parameterizations, where the simplification of Eq. 2.20 cannot be done, the commonly used parameter-independent observable $u(\mathbf{x})$ in Eq. 2.21 is no longer globally optimal, even asymptotically. Examples of such analyses include the off-shell Higgs boson production measurements and the various effective field theory (EFT) measurements, where quantum interference introduces a quadratic dependence of the probability model on the parameter of interest.

A simple way to think about this is that if a probability model allows for degenerate solutions, a fixed summary statistic optimized for a local parameter value is not sufficient for the scan over the entire parameter space.

2.4 New Proposal: Neural Simulation-Based Inference

With the rapid developments in the field of deep learning, there has been a rise in literature advocating the use of deep learning for direct estimation of probability densities $p(\mathbf{x}|\boldsymbol{\mu})$ or probability density ratios $p(\mathbf{x}|\boldsymbol{\mu})/p(\mathbf{x}|\boldsymbol{\mu}_0)$ without the need for dimensional reduction and binned approximations. This approach is broadly named Likelihood-Free Inference or Simulation-Based Inference [35]. In this thesis, the approach is labeled Neural Simulation-Based Inference (NSBI), inspired by some of the recent papers [36, 37]. The preferred choice of nomenclature is driven mainly due to the fact that traditional analysis can also be termed Simulation-Based and the Likelihood-Free term can be an ambiguous term as NN-estimated likelihoods are used for inference in the new technique despite being implicit in the original model.

Estimating the true probability densities or density ratios makes building an *unbinned* test statistic using Eq. 2.11 possible, which resolves the three limitations (see Sec. 2.3.2) of a standard analysis strategy. Instead of the simplified approximation of the underlying probability model from a 1D summary statistic as in Eq. 2.15, the NNs can be directly used to estimate the probability model:

$$p(\mathbf{x}|\boldsymbol{\theta}) = \frac{1}{\nu(\boldsymbol{\theta})} \frac{d\nu_{pp \rightarrow A}}{d\mathbf{x}}(\mathbf{x}|\boldsymbol{\theta}) \quad (2.22)$$

The idea behind this approach is the same as the so-called Matrix Element Method (MEM) [38], which has been used in several analyses like the early measurement (with less than 100 candidate events) of the top-quark mass by D0 [39]. These approaches estimate the probability density for each event and use it to build an unbinned test statistic.

This is exactly the approach needed to perform high precision measurements for rare processes like the off-shell Higgs boson production. It can extract the maximum possible information from each event and is maximally sensitive to interference effects due to its $\boldsymbol{\theta}$ -parametrized nature, addressing limitations 1 and 2 defined in Sec. 2.3.2. The per-event nature of the probability densities ensures that limitation 3 is resolved as well.

Analytical approximations in resonant measurements

In the case of resonant measurements like that of the on-shell Higgs boson, the probability model of Eq. 2.22 can be approximated as a function of the invariant mass resonance lineshape of Fig. 2.2a with an approximate analytical function. Such a technique was used to perform the most precise measurement to date of the Higgs boson mass using the $H \rightarrow ZZ \rightarrow 4\ell$ channel [40], and is also typically used for $H \rightarrow \gamma\gamma$ analysis [41]. The results of the An analytical function approximation is not practical for non-resonant measurements, where there is no simple dependence on a single observable.

2.4.1 Density Ratios

In order to do hypothesis testing and inference on a parameter of interest μ using the test statistic defined in Eq. 2.11, estimating the probability density ratios is sufficient. Estimating the probability density $p(\mathbf{x}|\mu)$ using NNs is possible, but is currently a technically difficult problem for inputs with large dimensions. This thesis focuses on a relatively easier problem of estimating the ratio of probability densities $p(\mathbf{x}|\mu)/p_{\text{ref}}(\mathbf{x})$ directly using well-trained classification NNs:

$$\frac{p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\alpha})}{p_{\text{ref}}(\mathbf{x})} = \frac{\frac{1}{\nu(\boldsymbol{\mu}, \boldsymbol{\alpha})} \frac{d\nu_{pp \rightarrow A}}{d\mathbf{x}}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\alpha})}{\frac{1}{\nu^{\text{ref}}} \frac{d\nu_{pp \rightarrow A}^{\text{ref}}}{d\mathbf{x}}(\mathbf{x})} \quad (2.23)$$

where $p_{\text{ref}}(\mathbf{x})$ can be any $\boldsymbol{\mu}, \boldsymbol{\alpha}$ -independent probability density so it cancels out during the profile likelihood fit, giving us the usual test statistic (2.11):

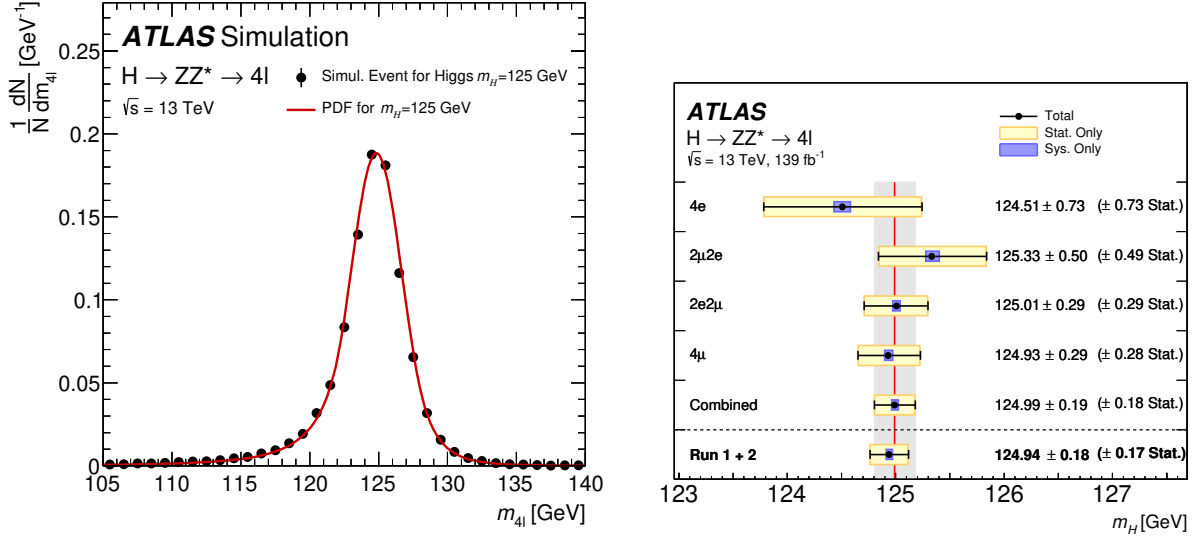


Figure 2.4: Left: The probability density function estimated using a combination of analytical approximations and NN output (red line) is shown in comparison with the $m_{4\ell}$ distribution in signal MC simulation (black points) for $m_H = 125$ GeV. Both the probability density function and the MC simulation are normalised to unity. Right: The test statistic values, $-2\ln(\lambda)$, calculated using the unbinned fit with an approximate, but parameterized, probability distribution function (shown on left for $m_H = 125$ GeV). The vertical (red) line on the right plot indicates the combined Run 2 result, and the grey band its total uncertainty.

$$\begin{aligned}
 t_{\boldsymbol{\mu}} = & -2 \mathcal{N}_{\text{data}} \left(\log \frac{\nu(\boldsymbol{\mu}, \hat{\boldsymbol{\alpha}})}{\nu(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\alpha}})} \right) + 2 \left(\nu(\boldsymbol{\mu}, \hat{\boldsymbol{\alpha}}) - \nu(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\alpha}}) \right) - 2 \sum_i^{\mathcal{N}_{\text{data}}} \log \frac{p(\mathbf{x}_i | \boldsymbol{\mu}, \hat{\boldsymbol{\alpha}}) / p_{\text{ref}}(\mathbf{x})}{p(\mathbf{x}_i | \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\alpha}}) / p_{\text{ref}}(\mathbf{x})} \\
 & + \sum_p^{N_{\text{syst}}} \left(\hat{\boldsymbol{\alpha}}^2 - \hat{\boldsymbol{\alpha}}^2 \right)
 \end{aligned} \tag{2.24}$$

The new NSBI technique reproduces Eq. 2.11 with no loss of information, and with no ad-hoc approximations like the one used in Eq. 2.16, discussed in Sec. 2.3.

The likelihood ratios can be estimated using supervised learning techniques. Of the numerous supervised learning techniques summarized in [2] for the estimation of probability density ratios, this work picks a variation of the simplest one - called the CARL method [42], without the additional calibration layer. In general this is also known as the *likelihood-ratio trick*. Note that estimating these exact $\boldsymbol{\mu}$ -parameterized density ratios directly achieves optimality at each POI scan point, even in cases where the parameterization allows for degenerate solutions, as will be further explored in Sec. 2.5. A cartoon workflow of the new analysis technique is shown in Fig. 2.5.

2.4.2 Biases in modelling and estimators

Standard techniques use Poisson PDFs to model the spread of events in each bin of the summary statistic $u(\mathbf{x})$ as described in Sec. 2.3. Maximum likelihood estimators $\hat{\boldsymbol{\theta}}$ derived from the test statistic

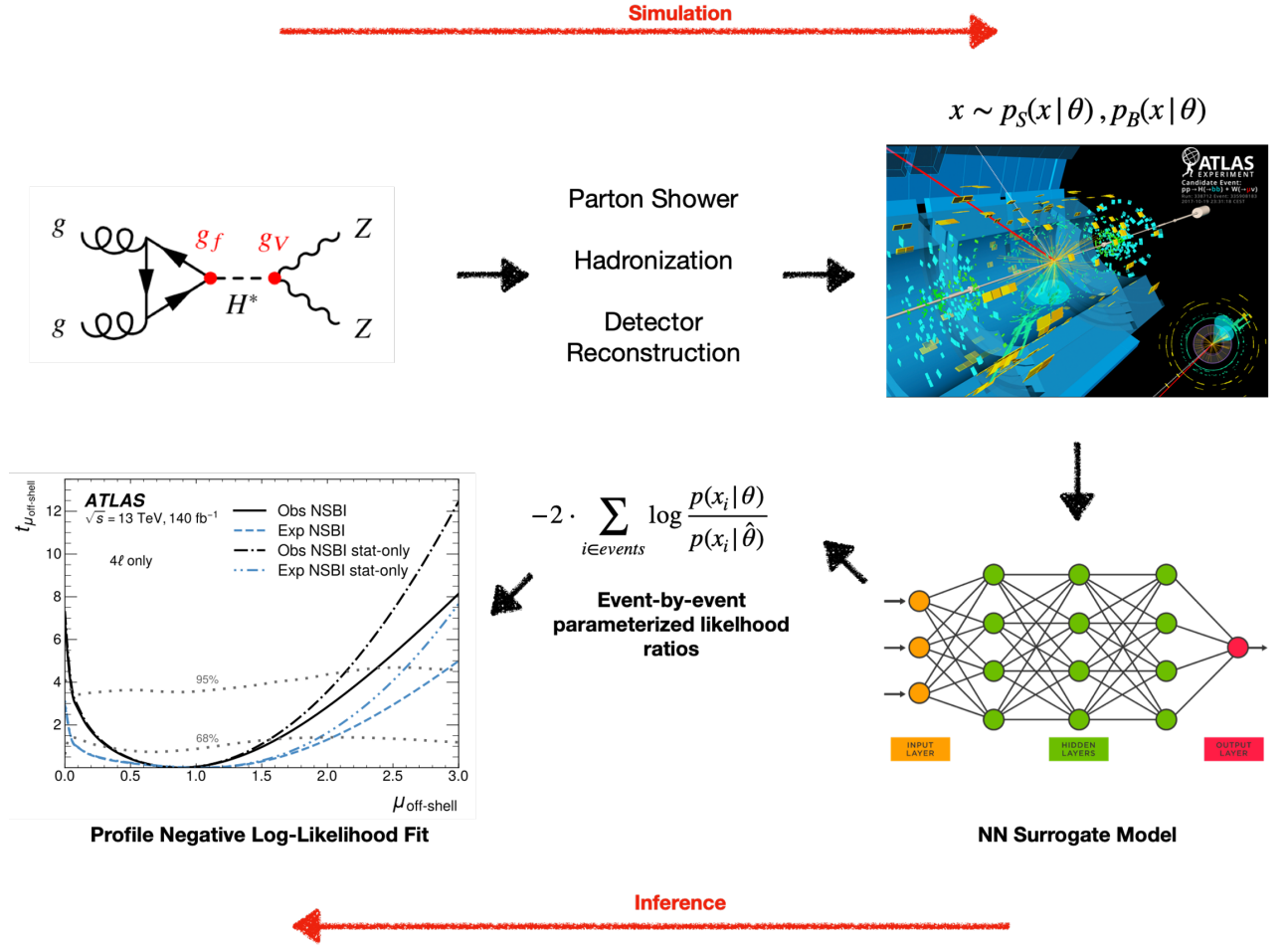


Figure 2.5: A cartoon representation of a typical NSBI workflow, where NNs are trained to estimate precise and accurate per-event parameterized likelihood ratios, then used for direct statistical inference using the profile negative log-likelihood ratio test statistic defined in Eq. 2.11. This is a generalization of the workflow depicted in Fig. 2.3 using deep NNs acting as surrogates for the likelihood ratios using the likelihood ratio technique shown in Fig. 6.1. The results in this cartoon will be discussed in detail in Chap. 7.

using approximate PDFs of Eq. 2.15 are asymptotically unbiased, i.e. for a large number of observations $\hat{\boldsymbol{\theta}}$ converges to the true parameter values $\boldsymbol{\theta}$. This is true even when the summary statistic $u(\mathbf{x})$ itself is calculated using a biased model that does not capture the full parameter dependence. This is because the spread of events is always approximated using Poisson probabilities which have well defined statistical properties. The focus in traditional techniques is given to building a low-variance function $u(\mathbf{x})$ - or in other words a function that does not over-fit on the simulation and is thus flexible.

Unlike this standard framework, the estimator $\hat{\boldsymbol{\theta}}$ from the test statistic using the unbinned approximation of Eq. 2.23 is not guaranteed to be asymptotically un-biased estimator of the true $\boldsymbol{\theta}$. To ensure this, rigorous diagnostic checks for bias in the modelling of probability density ratios of Eq. 2.23 needs to be done along with checks ensuring robust probability calibration. Low-bias and low-variance modelling is a major challenge in NSBI techniques, where it is easy to produce incorrect inference results if any of the biases in the modelling or the estimator are not accounted for.

A comprehensive set of diagnostics will be developed that ensure that the density ratio estimates of Eq. 2.23 are true representations of the underlying MC model, and that the estimator $\hat{\boldsymbol{\mu}}$ in the test statistic of Eq. 2.24 is an unbiased estimator of $\boldsymbol{\mu}$. This will be fully developed in Chaps. 6 and 3.

Moreover, the variance of the estimator in the standard analysis framework can be asymptotically approximated using the inverse of the Fisher information matrix in most cases. This is only guaranteed if the asymptotic limit [43] holds. For analysis that have a large signal-background interference the assumptions that enforce the asymptotic limit break down and confidence intervals need to be constructed by sampling pseudo-experiments from the MC samples. A robust technique for this Neyman Construction of the frequentist confidence intervals in an unbinned NSBI analysis will also be developed in Sec. 3.3.

To account for any residual biases due to limited MC statistics used in the NN modelling of Eq. 2.23, an ensemble of NNs will be used and the effects from the resulting uncertainty will be introduced as nuisance parameters in the test statistic. This is done using an approach very similar to the typical *spurious-signal* approach of unbinned $H \rightarrow \gamma\gamma$ analysis of the ATLAS experiment. This will be discussed in more details in Sec. 6.5.

2.5 Special Cases - Analytical Parameter Dependence

In most cases, the physics processes contributing to the final state can be broken down into components that scale with the parameter of interest and those that do not. This was briefly touched upon in Sec. 2.3.2. Without loss of generality, the POI-dependent components will be referred to as signal (S) processes and the others as background (B) processes, giving a generalization of Eq. 2.5:

$$p(\mathcal{D}|\boldsymbol{\mu}, \boldsymbol{\alpha}) = \text{Pois} \left(\mathcal{N}_{\text{data}} \left[\sum_S \nu_{pp \rightarrow A}^S(\boldsymbol{\mu}, \boldsymbol{\alpha}) + \sum_B \nu_{pp \rightarrow A}^B(\boldsymbol{\alpha}) \right] \right) \frac{1}{\sum_S \nu_{pp \rightarrow A}^S(\boldsymbol{\mu}, \boldsymbol{\alpha}) + \sum_B \nu_{pp \rightarrow A}^B(\boldsymbol{\alpha})} \prod_i^{\mathcal{N}_{\text{data}}} \left[\sum_S \frac{d\nu_{pp \rightarrow A}^S}{d\mathbf{x}}(\mathbf{x}_i|\boldsymbol{\mu}, \boldsymbol{\alpha}) + \sum_B \frac{d\nu_{pp \rightarrow A}^B}{d\mathbf{x}}(\mathbf{x}_i|\boldsymbol{\alpha}) \right] \quad (2.25)$$

Furthermore, the parameter of interest dependence can often be factorized out from the individual components, if the analytical dependence of the probability scaling is known:

$$\nu_{pp \rightarrow A}^S(\boldsymbol{\mu}, \boldsymbol{\alpha}) = f_S(\boldsymbol{\mu}) \cdot \nu_{pp \rightarrow A}^S(\boldsymbol{\alpha}) \quad (2.26)$$

$$\frac{d\nu_{pp \rightarrow A}^S(\mathbf{x}_k | \boldsymbol{\mu}, \boldsymbol{\alpha})}{d\mathbf{x}} = f_S(\boldsymbol{\mu}) \cdot \frac{d\nu_{pp \rightarrow A}^S(\mathbf{x}_k | \boldsymbol{\alpha})}{d\mathbf{x}} \quad (2.27)$$

where $f_S(\boldsymbol{\mu})$ is known based on first principles. This allows us to factorize the parameter-dependence in the density ratio estimation problem:

$$\begin{aligned} \frac{p(\mathcal{D} | \boldsymbol{\mu})}{\prod_k^{\mathcal{N}_{\text{data}}} p_{\text{ref}}(\mathbf{x}_k)} &= \text{Pois} \left(\mathcal{N}_{\text{data}} \left| \sum_S f_S(\boldsymbol{\mu}) \cdot \nu_S + \sum_B \nu_B \right| \right) \times \prod_k^{\mathcal{N}_{\text{data}}} \frac{p(\mathbf{x}_k | \boldsymbol{\mu})}{p_{\text{ref}}(\mathbf{x}_k)} \\ &= \text{Pois} \left(\mathcal{N}_{\text{data}} \left| \sum_S f_S(\boldsymbol{\mu}) \cdot \nu_S + \sum_B \nu_B \right| \right) \times \\ &\quad \frac{1}{\sum_S f_S(\boldsymbol{\mu}) \cdot \nu_S + \sum_B \nu_B} \prod_k^{\mathcal{N}_{\text{data}}} \left[\sum_S f_S(\boldsymbol{\mu}) \cdot \nu_S \cdot \frac{p_S(\mathbf{x}_k)}{p_{\text{ref}}(\mathbf{x}_k)} + \sum_B \nu_B \cdot \frac{p_B(\mathbf{x}_k)}{p_{\text{ref}}(\mathbf{x}_k)} \right], \end{aligned} \quad (2.28)$$

where

$$p_c(\mathbf{x}) = \frac{1}{\nu_c} \frac{d\nu_c}{d\mathbf{x}}(\mathbf{x}) \quad (2.29)$$

where the dependence on the NPs $\boldsymbol{\alpha}$ has been suppressed for the moment. These parameters need special treatment and will thus be re-introduced in Chap. 3 with the relevant formalism.

A single *parametrized network* may be trained to learn a conditional probability density ratio of Eq. 2.23 that varies with the hypothesis $\boldsymbol{\mu}$ under consideration [42, 44]. If the parametric dependence of a test statistic can be expressed analytically in terms of the POIs, as in Eq. 2.28, then **estimating a finite number of parameter-independent density ratios** p_c/p_{ref} can be used to build the full model of Eq. 2.23, without the need for a NN to learn the parametric dependence.

This analytical decomposition reduces the burden of validating the test statistic from testing its interpolation performance over the full theory parameter space to validating the performance of a small number of classifiers with fixed hypothesis. If every classifier is well-trained and well-calibrated, then their combination may be expected to remain well-behaved, although this must be explicitly verified.

2.5.1 Search-Oriented Mixture Models

While there is freedom to make any choice for the reference hypothesis $p_{\text{ref}}(\mathbf{x})$ in Eq. 2.23, this work defines a *search-oriented* reference process, built as a combination of all signal processes,

$$p_{\text{ref}}(\mathbf{x}_i) = \frac{1}{\sum_S \nu_S} \sum_S \nu_S \cdot p_S(\mathbf{x}_i), \quad (2.30)$$

This definition ensures that the denominators in Eq. 2.28 is larger than zero over the entire signal region of an analysis, which is the region that is sensitive to the signal processes. Here, p_{ref} is defined to be independent of $\boldsymbol{\mu}$, which allows the construction of the final profile likelihood ratio that is independent of p_{ref} as discussed in Sec. 2.4. The term p_{ref} contributes to a parameter-independent offset, which can be ignored in the maximization of the likelihood.

The search-oriented mixture model overcomes issues that may arise in alternative formulations of the reference sample. If probability density ratios $p_c(\mathbf{x})/p_{\text{ref}}(\mathbf{x})$ are used to estimate the likelihood ratio in regions of phase-space with $p_{\text{ref}}(x) \approx 0$, the final estimate will depend on a fine-tuned cancellation of large numbers, which is numerically unstable. A pre-selected signal region for the analysis must be defined that ensures $p_{\text{ref}}(x_i) > 0$ throughout the region. This definition of p_{ref} ensures that all signal-sensitive regions of the phase space can remain in the pre-selection region. This choice of p_{ref} also aids in the sample-efficient training of the individual classifiers. The type of pre-selection needed to filter out the events $\mathbf{x}$ with $p_{\text{ref}}(\mathbf{x}) \sim 0$ is discussed in Sec. 5.4.

Finally, it may be convenient to define p_{ref} such that it can be represented using simulated samples with only positive weighted events. This simplifies the procedure to construct confidence intervals, as sampling from negative weights is an ill-defined task. This will be discussed in Sec. 3.3.

2.5.2 Signal Strength Measurements

Signal strength is a dimensionless parameter that quantifies the ratio of the observed rate of a specific process to the rate predicted by the SM. In the case of a hypothesis test for a signal model defined in the SM, this is parameterized using a single POI $\boldsymbol{\mu} = \{\mu\}$ defined as:

$$\mu = \frac{\nu_{\text{observed}}^S}{\nu_{\text{SM}}^S} \quad (2.31)$$

where the ν_{observed}^S refers to the observed rate of events and ν_{SM}^S refers to the expected rate of signal events given a model. In this type of measurements, the signal processes can have fairly low expected yields compared to the backgrounds with the same final state. This can make the statistically significant observation of these signals difficult. This results in a need for optimal inference techniques that do not lose valuable information for signal identification. In other words, the test statistic of Eq. 2.11 needs to be calculated using the most robust estimate of the probability model $p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\alpha})$, for each value of vmu being scanned.

Linear parameterization

In the simplest case, the POI scales all signal hypothesis in the model uniformly. This can be written as:

$$\nu_{pp \rightarrow A}^S(\mu) = f(\mu) \cdot \nu_{pp \rightarrow A}^S \quad (2.32)$$

$$\frac{d\nu_{pp \rightarrow A}^S(\mathbf{x}_k|\mu)}{d\mathbf{x}} = f(\mu) \cdot \frac{d\nu_{pp \rightarrow A}^S(\mathbf{x}_k|\boldsymbol{\alpha})}{d\mathbf{x}} \quad (2.33)$$

While in principle the process-independent $f(\mu)$ can be any function of μ , upto a diffeomorphism it can be defined without loss of generality as $f(\mu) = \mu$. In this thesis, the case where the simplifications of Eqs. 2.32 and 2.33 is possible will be referred to as the "linear" cases since the probability model then can be rewritten as a linear function of the POI μ , and has a single non-degenerate solution:

$$p(\mathbf{x}|\mu) \sim \mu \cdot \sum_S \nu_S \cdot p_S(\mathbf{x}) + \sum_B \nu_B \cdot p_B(\mathbf{x}) \quad (2.34)$$

Note that were $f_{S_1}(\mu) \neq f_{S_2}(\mu)$ for any two signal hypothesis S_1, S_2 used in the model, the above simplification would not hold.

In this type of signal-strength measurements, $\mu = 1$ corresponds to the nominal signal hypothesis, while $\mu = 0$ corresponds to the background-only hypothesis.

Using simulated events from the probability models $p_S(\mathbf{x})$ and $p_B(\mathbf{x})$, the *summary statistic* $u(\mathbf{x})$ is built that is essentially a ratio of signal probability to background probability:

$$u(\mathbf{x}) \sim \frac{\sum_S p_S(\mathbf{x})}{\sum_B p_B(\mathbf{x})} \quad (2.35)$$

As discussed in Sec. 2.3.2, this is asymptotically optimal when the simplification of Eq. 2.20 is possible, and when the probabilities $p_c(\mathbf{x})$ are modeled accurately. Multivariate techniques are used to create the summary statistic $u(\mathbf{x})$ by training a Machine Learning algorithm such as a NN or a Boosted Decision Tree (BDT), to get $u(\mathbf{x})$ of Eq. 2.35 to be optimal as a function of the full final-state phase space. The corresponding observable $u(\mathbf{x})$ is then used to get an estimate of the probability model as defined in Eq. 2.16. There is inevitably loss of information since the binned probability model of Eq. 2.16 assumes the same likelihood for all events grouped together in a single bin, which may not be a good approximation when there are rare signal events.

Take for example the $VH(H \rightarrow b\bar{b})$ measurement that uses multiple BDTs optimized to discriminate signal from background in different phase space regions to measure the corresponding signal strength parameter. The resulting signal-background separation from a combination of these phase space regions is shown in Fig. 2.2b, and a representation of the standard workflow used is shown in Fig. 2.3.

Non-linear parameterization

The fixed summary statistic $u(\mathbf{x})$ might not capture the full dependence on the parameter of interest μ , unless the probability model scales linearly as a function of μ as in Eq. 2.34. This is not the case for example, when the signal-background interference is non-negligible:

$$p(\mathbf{x}|\mu) \sim \mu \cdot \sum_S \nu_S \cdot p_S(\mathbf{x}) + \sqrt{\mu} \cdot \sum_I \nu_I \cdot p_I(\mathbf{x}) + \sum_B \nu_B \cdot p_B(\mathbf{x}) \quad (2.36)$$

This type of model allows degenerate solutions for the POI μ , and the probability model can no longer be written as a linear function - hence the term *non-linear* parameterization in this thesis. It is easy to see how the observable of Eq. 2.35 is no longer an optimal summary statistic, since $f_S(\mu) \neq f_I(\mu)$ and thus no single signal-over-background ratio, like in Eq. 2.35, can be used as a proxy of the full density ratio across the parameter space. The local degenerate solution will appear in parameter space regions with very large interference effects $\sqrt{\mu} \cdot p_I(\mathbf{x}) \gg \mu \cdot p_S(\mathbf{x})$, where the fixed summary statistic will be especially sub-optimal. The interference is just one example - in principle **any model with parameter dependence that allows degenerate solutions faces the same sub-optimality from standard inference techniques using fixed summary statistics**. This is even when the summary statistic is built to be asymptotically optimal in some local parameter space.

In these more general cases, using the full probability model of Eq. 2.28 that uses *parameterized* density ratios for hypothesis testing ensures asymptotic optimality across the parameter space. In these cases, an NSBI analysis can thus bring significant improvement in the measurement sensitivity compared to using more traditional techniques.

2.5.3 Parameterized Optimal Observables

Instead of building a full unbinned analysis using the test statistic, the parameterized density ratios can themselves be used as a binned summary statistic. But unlike a traditional *fixed* summary, this will be a parameter-dependent observable $u(\mathbf{x}|\mu)$ that is optimal for each value of μ being scanned over:

$$u(\mathbf{x}|\mu) = \frac{p(\mathbf{x}|\mu)}{p(\mathbf{x}|\mu_0)} \quad (2.37)$$

where μ_0 is the nominal hypothesis. This type of parameter-dependent observable will be referred to in this thesis as a parameterized *Optimal Observable*. Note that these observables are now **asymptotically optimal for the scan over the entire parameter space, not just locally at some reference parameter value**. The observable of Eq. 2.37 can be used to infer the probability density using binned approximations as in Eq. 2.16:

$$p(\mathcal{D}|\mu) = \text{Pois}(\mathcal{N}_{\text{data}}|\nu(\mu)) \prod_i^{\mathcal{N}_{\text{data}}} \frac{1}{\nu(\mu)} \frac{\nu_{I(\mathbf{x}_i|u(\mu))}(\mu)}{\Delta_{I(\mathbf{x}_i|u(\mu))}} \quad (2.38)$$

Note the μ -dependence in the observable $u(\mathbf{x}|\mu)$ in contrast to Eq. 2.15. As discussed in Sec. 2.3.2, this resolves two of the limitations of a more traditional binned summary but can be less optimal than an unbinned analysis using parameterized density ratios. Chap. 8 will show how the gradual unbinned of these optimal observables results in a sensitivity gain that is comparable to a full unbinned NSBI analysis, using the probability model defined in Eq. 2.38.

The term optimal observables is based on Refs. [45, 46, 47], and refers to an observable that is only optimal in a local parameter phase space. Traditionally these optimal observables are calculated as the ratio of Matrix Elements (ME) between two *fixed* hypothesis. The ME calculations are also only done at the parton-level, and thus do not account for parton-showering and detector effects. Ref. [2] motivates the use of *score* $t(x|\theta_{\text{ref}}) = \nabla_{\theta} p(x|\theta)|_{\theta_{\text{ref}}}$ as the optimal observable that saturates

the Cramer-Rao bound locally in the vicinity of some chosen value $\boldsymbol{\theta} = \boldsymbol{\theta}_{\text{ref}}$. Both of these approaches can be optimal in most cases of parameter estimation but end up losing sensitivity for a scan over the entire parameter space, in the presence of degenerate solutions.

Chapter 3

Uncertainties in the Model

This chapter introduces a computationally and statistically efficient treatment of estimating density ratios parameterized on the full systematic uncertainty model of the ATLAS experiment, as well as proposes a novel and practical technique for performing Neyman Construction in an NSBI analysis that only estimates the probability density ratios.

The challenge of *model misspecification* in NSBI, where the simulations have systematic differences from real data, is addressed by the introduction of NPs $\boldsymbol{\alpha}$ representing systematic uncertainties, and by testing the modeling of these systematic uncertainties themselves, in a similar way to how it is done for traditional analysis techniques in particle physics [33].

Most of the systematic uncertainty NPs in a typical LHC experiment analysis are constrained by a complimentary auxiliary measurement which provides a maximum likelihood estimate a_m , also referred to as the Auxiliary Observable (AO), and a corresponding uncertainty δ_m on the nominal value, as was briefly described in Sec. 2.1.1. These values are used to build the auxiliary likelihood in Eq. 2.7.

Several proposals can be found in the simulation-based inference literature for how to build the $\boldsymbol{\alpha}$ -parametrized density ratios $p_c(\mathbf{x}, \boldsymbol{\alpha})/p_{\text{ref}}(\mathbf{x})$. The most common solution [48] involves training a parametrized NN for density ratio estimation with the NPs as new inputs. This would involve expanding the NN input $\mathbf{x}$ to also include randomly assigned $\boldsymbol{\alpha}$ values, with correspondingly varied events for the loss function. The output can then be used model the joint probability density ratio model of Eq. 2.28 in an NSBI analysis:

$$\frac{p(\mathcal{D}|\boldsymbol{\mu}, \boldsymbol{\alpha})}{\prod_k^{\mathcal{N}_{\text{data}}} p_{\text{ref}}(\mathbf{x}_k)} = \text{Pois} \left(\mathcal{N}_{\text{data}} \left| \sum_S f_S(\boldsymbol{\mu}) \cdot \nu_S(\boldsymbol{\alpha}) + \sum_B \nu_B(\boldsymbol{\alpha}) \right| \right) \times \frac{1}{\sum_S f_S(\boldsymbol{\mu}) \cdot \nu_S(\boldsymbol{\alpha}) + \sum_B \nu_B(\boldsymbol{\alpha})} \prod_k^{\mathcal{N}_{\text{data}}} \left[\sum_S f_S(\boldsymbol{\mu}) \cdot \nu_S(\boldsymbol{\alpha}) \cdot \frac{p_S(\mathbf{x}_k|\boldsymbol{\alpha})}{p_{\text{ref}}(\mathbf{x}_k)} + \sum_B \nu_B(\boldsymbol{\alpha}) \cdot \frac{p_B(\mathbf{x}_k|\boldsymbol{\alpha})}{p_{\text{ref}}(\mathbf{x}_k)} \right], \quad (3.1)$$

where

$$p_c(\mathbf{x}|\boldsymbol{\alpha}) = \frac{1}{\nu_c(\boldsymbol{\alpha})} \frac{d\nu_c}{d\mathbf{x}}(\mathbf{x}|\boldsymbol{\alpha}) \quad (3.2)$$

In practice, the direct parameterized modelling of the various $p_c(\mathbf{x}|\boldsymbol{\alpha})/p_{\text{ref}}(\mathbf{x})$ is not feasible due to three main limitations. First, simulations corresponding to $\boldsymbol{\alpha}$ -variations are typically generated in ATLAS and CMS experiments by varying a single NP $\alpha_m \in \boldsymbol{\alpha}$ at a time, with no simulation model available where multiple parameters $\alpha_m, \alpha_n \in \boldsymbol{\alpha}$ vary simultaneously. Second, only three sets of samples are typically available per NP: $\alpha_m = \{0, \pm 1\}$, which before the standard change of variables $(\alpha_m - a_m)/\delta_m \rightarrow \alpha_m$ correspond to the best-fit and $\pm 1\sigma$ uncertainties from the auxiliary measurements $\alpha_m = \{a_m, a_m \pm \delta_m\}$. And finally, the dimension of the NP vector $\boldsymbol{\alpha}$ is high-dimensional - making the parameterized density ratio estimation computationally and simulation-heavy. To summarize, the input simulations available for a typical ATLAS experiment analysis are insufficient for a NN to learn the full parametric dependence. Moreover, validating the interpolation capabilities of a classifier across the high-dimensional space of NPs is a challenging task.

Instead, this thesis extends the systematic uncertainty framework used in histogram-based analyses, based on HistFactory [49], to an unbinned multi-dimensional setting. This assumes that the impact of systematic uncertainties from independent sources is treated independently, following the standard practice at the LHC. This assumption is well-motivated due to the design of the systematics model in LHC experiments.

3.1 Uncertainty modelling in NSBI

Parametrizing $\nu_c(\boldsymbol{\alpha})$ is a common problem in LHC statistical data analysis and the standard solution will be reviewed in this section. The total rate in the cases where factorizing the POI dependence is possible can be written as:

$$\begin{aligned} \nu(\boldsymbol{\mu}, \boldsymbol{\alpha}) &= \sum_S f_S(\boldsymbol{\mu}) \cdot \nu_S(\boldsymbol{\alpha}) + \sum_B \nu_B(\boldsymbol{\alpha}) \\ &= \sum_S f_S(\boldsymbol{\mu}) \cdot G_S(\boldsymbol{\alpha}) \cdot \nu_S + \sum_B G_B(\boldsymbol{\alpha}) \cdot \nu_B \end{aligned} \quad (3.3)$$

where $\nu_c = \nu_c(\boldsymbol{\alpha} = 0)$ due to the rescaling of the NPs that turns the auxiliary likelihood to a unit Gaussian as in Eq. 2.8. One then assumes that the ratio $G_c(\boldsymbol{\alpha}) = \nu_c(\boldsymbol{\alpha})/\nu_c$ factorizes, which is quite well-motivated in a typical systematics model used at the LHC experiments:

$$G_c(\boldsymbol{\alpha}) = \prod_{m=1}^{N_{\text{syst}}} G_c(\alpha_m) = \prod_{m=1}^{N_{\text{syst}}} \frac{\nu_c(\alpha_m)}{\nu_c} \quad (3.4)$$

There are dedicated interpolation methods described in [50] which can be used to build $G_c(\alpha_m)$. Following the prescription used in most recent ATLAS analyses, the polynomial-exponential interpolation is used:

$$\frac{\nu_X(\alpha_m)}{\nu_X} = \begin{cases} \left(\frac{\nu_X^+}{\nu_X}\right)^{\alpha_m} & \alpha_m > 1 \\ 1 + \sum_{i=1}^6 a_i \alpha_m^i & -1 \leq \alpha_m \leq 1 \\ \left(\frac{\nu_X^-}{\nu_X}\right)^{-\alpha_m} & \alpha_m < -1 \end{cases} \quad (3.5)$$

where a_i are chosen to ensure the continuity of the function as well as its derivatives at $\alpha = \pm 1$. The values of ν_c^+ and ν_c^- are the standard $+1\nu$ and -1ν variations in the yield. As discussed above, this corresponds to the values $\alpha = \pm 1$.

3.1.1 Building the α -parametrized density ratios

In this section, a solution to the per-event uncertainty parameterization problem is proposed which not only avoids the problem of training NNs with unfeasible large-dimensional input vectors, but also reduces the computational cost by factorizing out the estimation of NP parameterization of the estimation of α -independent density ratios $p_c(\mathbf{x})/p_{\text{ref}}(\mathbf{x})$.

The method begins with the same hypothesis used to parametrize the rate. The α -dependence is factorized out as in the yields:

$$\frac{p_c(\mathbf{x}|\alpha)}{p_{\text{ref}}(\mathbf{x})} = g_c(\mathbf{x}; \alpha) \cdot \frac{p_c(\mathbf{x})}{p_{\text{ref}}(\mathbf{x})} \quad (3.6)$$

The first hypothesis is that the α - parameterized per-event density ratio:

$$g_c(\mathbf{x}; \alpha) = \frac{\frac{1}{\nu_c(\alpha)} \frac{d\nu_c}{d\mathbf{x}}(\mathbf{x}, \alpha)}{\frac{1}{\nu_c} \frac{d\nu_c}{d\mathbf{x}}(\mathbf{x})} \quad (3.7)$$

factorizes as:

$$g_c(\mathbf{x}, \alpha) = \prod_{m=1}^{N_{\text{syst}}} g_c(\mathbf{x}, \alpha_m) \quad (3.8)$$

This is the same hypothesis used for every analysis in ATLAS and holds true since different NPs represent different *sources* of systematic variation and should not depend on each other. Then, the same polynomial-exponential interpolation used for the rates can be used, but now directly for the per-event weights:

$$g_c(\mathbf{x}, \alpha_m) = \begin{cases} (g_c(\mathbf{x}, \alpha_m^+))^{\alpha_m} & \alpha_m > 1 \\ 1 + \sum_{i=1}^6 a_i \alpha_m^i & -1 \leq \alpha_m \leq 1, \\ (g_c(\mathbf{x}, \alpha_m^-))^{-\alpha_m} & \alpha_m < -1 \end{cases} \quad (3.9)$$

with the a_i terms are calculated to keep the function and its derivative continuous at $\alpha_m^\pm = \pm 1$. The estimation of $g_c(\mathbf{x}, \alpha_m^\pm)$ is described in the next sub-section. Inserting these parameterized ratios into

the the full α -dependant joint probability density ratio of Eq. 3.1:

$$\frac{p(\mathcal{D}|\boldsymbol{\mu}, \boldsymbol{\alpha})}{\prod_k^{\mathcal{N}_{\text{data}}} p_{\text{ref}}(\mathbf{x}_k)} = \text{Pois} \left(\mathcal{N}_{\text{data}} \left| \sum_c f_c(\boldsymbol{\mu}) \cdot \left[\prod_{m=1}^{N_{\text{syst}}} G_c(\alpha_m) \right] \nu_c \right. \right) \times \frac{1}{\sum_c f_c(\boldsymbol{\mu}) \cdot \left[\prod_{m=1}^{N_{\text{syst}}} G_c(\alpha_m) \right] \nu_c} \prod_k^{\mathcal{N}_{\text{data}}} \left[\sum_c f_c(\boldsymbol{\mu}) \cdot \left[\prod_{m=1}^{N_{\text{syst}}} G_c(\alpha_m) \right] \nu_c \cdot \left[\prod_{m=1}^{N_{\text{syst}}} g_c(\mathbf{x}|\alpha_m) \right] \frac{p_c(\mathbf{x}_k)}{p_{\text{ref}}(\mathbf{x}_k)} \right], \quad (3.10)$$

where c corresponds to the various S and B channels in the summations of Eq. 3.1, and $f_B(\boldsymbol{\mu}) = 1$ for all B . The $p_c(\mathbf{x})/p_{\text{ref}}(\mathbf{x})$ terms are the α -independent probability density ratios that can be evaluated without any information of the full systematic model. Each source of systematic uncertainty NP $\alpha_m \in \boldsymbol{\alpha}$ can then be individually evaluated using two NN trained ratios $p_c(\mathbf{x}|\alpha_m^\pm)/p_c(\mathbf{x})$ per NP α_m . The details on estimation and validations of these ratios will be discussed in Chap. 6. The full likelihood model can be written as:

$$\frac{L(\boldsymbol{\mu}, \boldsymbol{\alpha}|\mathcal{D}, \mathcal{G})}{\prod_k^{\mathcal{N}_{\text{data}}} p_{\text{ref}}(\mathbf{x}_k)} = \frac{p(\mathcal{D}|\boldsymbol{\mu}, \boldsymbol{\alpha})}{\prod_k^{\mathcal{N}_{\text{data}}} p_{\text{ref}}(\mathbf{x}_k)} \cdot \prod_{m=1}^{N_{\text{syst}}} \frac{1}{\sqrt{2\pi}} \exp \left[-\frac{(\alpha_m - a_m)^2}{2\delta_m^2} \right] \quad (3.11)$$

where the $\prod_k^{\mathcal{N}_{\text{data}}} p_{\text{ref}}(\mathbf{x}_k)$ term contributes as a constant offset in the log-likelihood fit and can thus be neglected in a hypothesis test using the test statistic of Eq. 2.24.

The formalism developed here factorizes the estimation of the nominal probability density ratio $p_c(\mathbf{x})/p_{\text{ref}}(\mathbf{x})$ from the estimation of the effect from each of systematic uncertainties, considerably simplifying the estimation and validation of the full likelihood model. This makes simulation-based inferences practical for any physics analysis that allows factorization of POI dependence and has a uncorrelated systematic uncertainty model.

3.1.2 Amortized inference

A distinct feature of NSBI technique developed in this thesis is that once the per-event parameterized probability model of Eq. 3.1 is estimated using MC simulations, the model is *amortized* for any subsequent inference [1]. This is unlike a traditional binned summary statistic measurement, where the MC predictions need to be used for inference even when fitting observed detector data. This is because the Poisson probability model looks like:

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{\text{bin}}^{N_{\text{bins}}} \frac{e^{-\nu_{\text{bin}}} \cdot \nu_{\text{bin}}^{\mathcal{N}_{\text{bin}}}}{\mathcal{N}_{\text{bin}}!} \quad (3.12)$$

where the observed data $\mathcal{N}_{\text{bin}}$ is compared to the expected yields, estimated using simulated samples parameterized on $\boldsymbol{\theta}$. On the other hand, when building a log-likelihood model for NSBI:

$$-2 \log \frac{L(\boldsymbol{\mu}, \boldsymbol{\alpha})}{\prod_k^{\mathcal{N}_{\text{obs}}} p_{\text{ref}}(\mathbf{x}_k)} = -2 \sum_k^{\mathcal{N}_{\text{obs}}} \log \frac{p(\mathbf{x}_k | \boldsymbol{\mu}, \boldsymbol{\alpha})}{p_{\text{ref}}(\mathbf{x}_k)} \quad (3.13)$$

there is no distinction between data and MC. This means that the $\boldsymbol{\mu}, \boldsymbol{\alpha}$ -parameterized density ratio model (equivalent of the summary observable in a traditional analysis) is evaluated on data, independently of the simulations used to train the model, to perform the final fit. A practical consequence of this in the context of LHC analysis is that systematic uncertainty parameterization is done on both the MC and the observed data, unlike the standard approach where only the MC is parameterized. Chap. 8 discusses the full workflow for training and inference in more detail.

3.2 Calculation of uncertainties and NP impacts

The HESSE [51, 52, 53] covariance matrix is estimated using the inverse of the Hessian matrix at the maximum likelihood estimate $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\alpha}})$, which is an approximation of the inverse of the Fisher information matrix,

$$I_{nm}(\boldsymbol{\theta}) = \mathbb{E} \left[- \frac{\partial^2 \log p(\mathbf{x} | \boldsymbol{\theta})}{\partial \theta_n \partial \theta_m} \middle| \boldsymbol{\theta} \right] \quad (3.14)$$

where $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\alpha})$ and the $\mathbb{E}[\cdot]$ denotes the expectation value, which involves integration over the entire phase space of $\mathbf{x}$. Since this integral would be impractical to compute, in the HESSE approximation, the covariance matrix is instead calculated using the log-likelihood ratio model $\lambda(\boldsymbol{\mu}, \boldsymbol{\alpha}) = -2 \ln \left(L(\boldsymbol{\mu}, \boldsymbol{\alpha} | \mathcal{D}, \mathcal{G}) / \prod_k^{\mathcal{N}_{\text{data}}} p_{\text{ref}}(\mathbf{x}_k) \right)$ for the observed data $(\mathcal{D}, \mathcal{G})$:

$$V_{nm} = \left[\frac{1}{2} \frac{\partial^2 \lambda}{\partial \theta_n \partial \theta_m}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\alpha}}) \right]^{-1}, \quad (3.15)$$

where the parameter-independent nature of $p_{\text{ref}}(\mathbf{x})$ ensures that it does not play a role in the gradient. The likelihood ratio model $\lambda(\boldsymbol{\mu}, \boldsymbol{\alpha})$ of Eq. 3.11 is a differentiable function. Its dependence on the POIs $\boldsymbol{\mu}$ and NPs $\boldsymbol{\alpha}$ is introduced through differentiable functions, either through smooth functions or through NNs that are themselves differentiable. It is thus natural to leverage auto-differentiation techniques [54] to calculate the Hessian matrix of the likelihood ratio model. Auto-differentiation tools are useful since they can be leveraged to perform the profile fitting using exact gradients. Moreover, the calculation of the Hessian matrices can be parallelized on computing clusters [6], by calculating Hessian vector products instead of materializing the full Hessian matrix. This also reduces the memory usage in the gradient computation as described in Chap. 8. In the HESSE approximation, the $\pm 1\sigma$ uncertainty on the maximum likelihood estimator $\hat{\mu}$ of a POI $\mu \in \boldsymbol{\mu}$ is given by:

$$\sigma_{\hat{\mu}} = \pm \sqrt{\left[\frac{1}{2} \frac{\partial^2 \lambda}{\partial \mu^2}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\alpha}}) \right]^{-1}} \quad (3.16)$$

and similarly the variance on the estimators $\hat{\alpha}$ can be extracted from the covariance matrix. The pull of a NP $\alpha_m \in \boldsymbol{\alpha}$ is defined as

$$\hat{\alpha}_m \rightarrow \frac{\hat{\alpha}_m - a_m}{\sqrt{V_{mm}}}. \quad (3.17)$$

where a_m is the pre-fit (or nominal) value from the auxiliary measurement. The LHS of the equation corresponds to the NP α_m after applying the rescaling described in Sec. 2.1.1. The impact of a systematic uncertainty is an estimate of the propagation of the uncertainty to the POIs $\boldsymbol{\mu}$. Like in case of estimation of pulls, the impact of a systematic uncertainty can be approximated by a local expression based solely on the Hessian matrices or by a more accurate, but computationally more expensive, expression based on the maximum likelihood estimate calculated at different parameter values.

The impact of a systematic uncertainty on a POI $\mu \in \boldsymbol{\mu}$ is traditionally estimated by the difference $\Gamma_k^{\text{NP},(\pm)} = \hat{\mu}(\hat{\alpha}_m \pm \sqrt{V_{mm}}) - \hat{\mu}$ where $\hat{\mu}(\hat{\alpha}_m \pm \sqrt{V_{mm}})$ is the conditional maximum likelihood estimate obtained by keeping the NP associated with the systematic uncertainty fixed at $\alpha_m = \hat{\alpha}_m \pm \sqrt{V_{mm}}$. Recently [55], an alternative definition has been proposed where the impact is estimated with the difference $\Gamma_m^{\text{AO},(\pm)} = \hat{\mu}(\hat{\alpha}_m \mp \delta_m/\sqrt{V_{mm}}) - \hat{\mu}$ of the maximum likelihood estimates when the value of the auxiliary observable (AO) a_m representing the auxiliary measurement is shifted by the externally provided uncertainty and when it is kept at its nominal value. The latter definition has been shown to allow for a consistent decomposition of systematic and statistical uncertainties, as will be discussed in detail in Sec. 7.1.4.

Both definitions of the impact of a systematic uncertainty have local definitions which can be calculated efficiently with auto-differentiation. In the case of a single POI, a local estimate of the impact Γ_m^{NP} is given by

$$\Gamma_m^{\text{NP}} = \frac{\partial \hat{\mu}}{\partial \alpha_m} \times \sqrt{V_{mm}} = \left[\frac{\partial^2 \lambda}{\partial \mu^2}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\alpha}}) \right]^{-1} \frac{\partial^2 \lambda}{\partial \mu \partial \alpha_m}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\alpha}}) \times \sqrt{V_{mm}}, \quad (3.18)$$

considerably simplifying the analysis of the profile likelihood ratio. The effect of pulls on the impact can be evaluated by calculating the so-called *pre-fit* impact, which is obtained by replacing $\sqrt{V_{mm}} \rightarrow \delta_m$ in Eq. 3.18. A local estimate for the alternative definition of impact is given by $\Gamma_m^{\text{AO}} = V_{0m}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\alpha}})$ [55], where the covariance matrix is defined in Eq. 3.15. Further details about these calculations for NSBI using auto-differentiation techniques are described in Ref. [6].

3.3 Neyman Construction

Well known cases when the test statistic $t_{\boldsymbol{\mu}}$ is not chi-squared distributed include analysis with very low number N of pre-selected data events (low-background searches) and with probability models that allow degenerate solutions for the POI. In the case of NSBI analyses, any residual bias in the probability density ratios will yield a non chi-square distributed test statistics, even if there is a large number of data events pre-selected and if the probability model is linear so as to allow a single solution. When the distribution is not asymptotic, Monte Carlo methods are required to define frequentist confidence intervals. The standard procedure for LHC analyses is called Neyman construction.

Neyman construction for histogram-based analyses are a standard procedure since the spread of events in each bin is described using a Poisson PDF, which is then used to sample the pseudo-data for each bin I for a given set of parameters θ . In the NSBI analysis presented in this thesis, the probability densities $p(\mathbf{x}|\mu, \alpha)$ are intractable and only the density ratios between two hypothesis $p(\mathbf{x}|\mu, \alpha)/p_{\text{ref}}(\mathbf{x})$ is estimated using NNs. This leaves out the possibility of directly sampling events using a PDF.

3.3.1 Poisson Bootstrap technique

In this work, we instead propose sampling using a Poisson bootstrapping technique [56, 57]. The Poisson bootstrap is a resampling method that generates new datasets by assigning each observation a random number of replications (or integer weights) drawn from a Poisson distribution, with the frequency determined by its associated weight w_i .

$$w_i \rightarrow w_i^{\text{pseudo-exp}} = \text{Poisson}(w_i) \quad (3.19)$$

The basic idea is to treat each event's Monte Carlo weight as the expected value (mean) of a Poisson distribution. We then use a random number generator to sample Poisson distributed integers corresponding to each event, and use these sampled integers as the weight for the corresponding events to build the pseudo-experiment $\mathcal{D}_{\text{pseudo-exp}}$. This converts the event weights from real numbers to a non-negative integer distributed according to a Poisson PDF with a mean equal to its Monte Carlo weight, effectively un-weighting the pseudo-experiment.

The Poisson bootstrapping method presented here cannot always be directly used on the Asimov MC dataset. This is true for example if the dataset contains events with negative weights, because the Poisson distribution can only have a positive mean value. A solution to this problem is to use trained probability density ratio estimates to reweight a dataset with all positive weighted events to an unbiased (as a function of the same observable $\mathbf{x}$ used in the analysis) alternative Asimov dataset. Since both the density ratios and the event weights are positive, these weights are positive-definite and the Poisson sampling of Eq. 3.19 can proceed on this reweighted dataset.

In order to obtain pseudo experiments at different values of μ and α , one can also just use the probability density ratios trained to build the test statistic, if the reference sample p_{ref} is chosen to have only positive weighted events as suggested in the Search-Oriented Mixture model proposed Sec. 2.5.1.

$$w_i^{\text{ref}} \rightarrow w_i^{\text{Asimov}}(\mu, \alpha) = \frac{\nu(\mu, \alpha)}{\nu_{\text{ref}}} \cdot \frac{p(\mathbf{x}|\mu, \alpha)}{p_{\text{ref}}(\mathbf{x})} \cdot w_i^{\text{ref}} \quad (3.20)$$

Using Eq. 3.20, the Asimov dataset can be generated from the p_{ref} sample at any value of μ, α , followed by a Poisson bootstrap using Eq. 3.19 to generate a large number of pseudo-experiments $\mathcal{D}_{\text{pseudo-exp}} = \{w_i^{\text{pseudo-exp}}(\mu, \alpha)\}$, using a different seed in the random number generator for each pseudo-experiment.

$$w_i^{\text{Asimov}}(\mu, \alpha) \rightarrow w_i^{\text{pseudo-exp}}(\mu, \alpha) = \text{Poisson}(w_i^{\text{Asimov}}(\mu, \alpha)) \quad (3.21)$$

In order for this method to work, it is necessary that the Asimov datasets resulting from Eq. 3.20 are un-biased representations of the Monte Carlo dataset at any value of μ, α . This is demonstrated by

the reweighted diagnostic checks in Sec. 6.3. Moreover, the asymptotic properties of the re-weighted dataset are verified with that of the original Monte Carlo dataset by comparing the t_μ curves in Appendix B.

Confidence Interval Construction

Now that formalism for generation of pseudo-experiments has been established, the Neyman construction proceeds by scanning over the full parameter range of each of the POIs $\mu \in \boldsymbol{\mu}$ to build confidence intervals. For each pseudo-experiment, the value of the AO a_m associated with each of the constrained NP α_m in the fit is sampled from a Gaussian probability density $\mathcal{G}_{\text{pseudo-exp}} \sim \{\text{Gaus}(a_m|\hat{\alpha}_m^{\text{obs}}, 1)\}$ [50]. The values of the POIs $\boldsymbol{\mu}$ being scanned are used together with the NPs fixed to the best fit value from an initial global fit $\hat{\boldsymbol{\alpha}}^{\text{obs}}$ using the full observed dataset $\mathcal{D}$, to sample pseudo-experiments which are then used to fit the likelihood with sampled values of a_m per pseudo-experiment. This is the same general procedure highlighted in Ref. [50], but with the likelihood ratio model given by:

$$\frac{L(\boldsymbol{\mu}, \boldsymbol{\alpha}|\mathcal{D}_{\text{pseudo-exp}}, \mathcal{G}_{\text{pseudo-exp}})}{\prod_k^{\mathcal{N}_{\text{pseudo-exp}}} p_{\text{ref}}(\mathbf{x}_k)} = \frac{p(\mathcal{D}_{\text{pseudo-exp}}|\boldsymbol{\mu}, \boldsymbol{\alpha})}{\prod_k^{\mathcal{N}_{\text{pseudo-exp}}} p_{\text{ref}}(\mathbf{x}_k)} \cdot \prod_{m=1}^{N_{\text{syst}}} \frac{1}{\sqrt{2\pi}} \exp \left[-\frac{(\alpha_m - a_m)^2}{2\delta_m^2} \right] \quad (3.22)$$

where $\mathcal{D}_{\text{pseudo-exp}} = \{w_i^{\text{pseudo-exp}}(\boldsymbol{\mu}, \hat{\boldsymbol{\alpha}}^{\text{obs}})\}$ is the sampled pseudo-experiment corresponding to the Asimov truth $(\boldsymbol{\mu}, \hat{\boldsymbol{\alpha}}^{\text{obs}})$ values and $\mathcal{N}_{\text{pseudo-exp}}$ the corresponding number of events with non-zero weights. Note that for clarity the standard change of variables has been reverted in this formula.

The construction of confidence intervals then takes place by estimating the distributions of t_μ built using the likelihood model of Eq. 3.22, from the pseudo-experiments sampled for a range of truth $\mu \in \boldsymbol{\mu}$ values keeping the truth $\boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}}^{\text{obs}}$ fixed. To build the 1σ and 2σ confidence belts, the t_μ distribution for each $\mu \in \boldsymbol{\mu}$ being scanned is then integrated over to find the boundaries corresponding to 68.24% and 95.45% of the pseudo-experiments. This is shown in Fig. 7.1 for the case of the off-shell Higgs boson analysis (described in the next chapter). These correspond to the 1σ and 2σ boundaries for the particular values of POI $\mu_{\text{off-shell}}$, defined in Eq. 4.16, which is scanned over in finite intervals for the entire parameter range to build the $\mu_{\text{off-shell}}$ -parameterized confidence belts, shown later in Fig. 7.1. It is the projections along the $\mu_{\text{off-shell}}$ scan of the intersection of these confidence belts and the test statistic scan (shown in Fig. 7.1 for the off-shell Higgs boson analysis) that determines the 1σ and 2σ uncertainties on the POI μ being scanned over. These uncertainties are quoted in Chap. 7 for the measurement of POI $\mu_{\text{off-shell}}$ of the off-shell Higgs boson analysis. The best-fit values $\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\alpha}}$ for each pseudo-experiment is also stored in memory during the Neyman Construction procedure and used to build the $\hat{\mu}, \hat{\alpha}$ distributions $g(\hat{\mu}|\mu_{\text{truth}})$ and $g(\hat{\alpha}_m|\hat{\alpha}_m^{\text{obs}})$, also sometimes referred to as sampling distributions. This is shown in Fig. B.2 in Appendix B for the off-shell Higgs boson analysis.

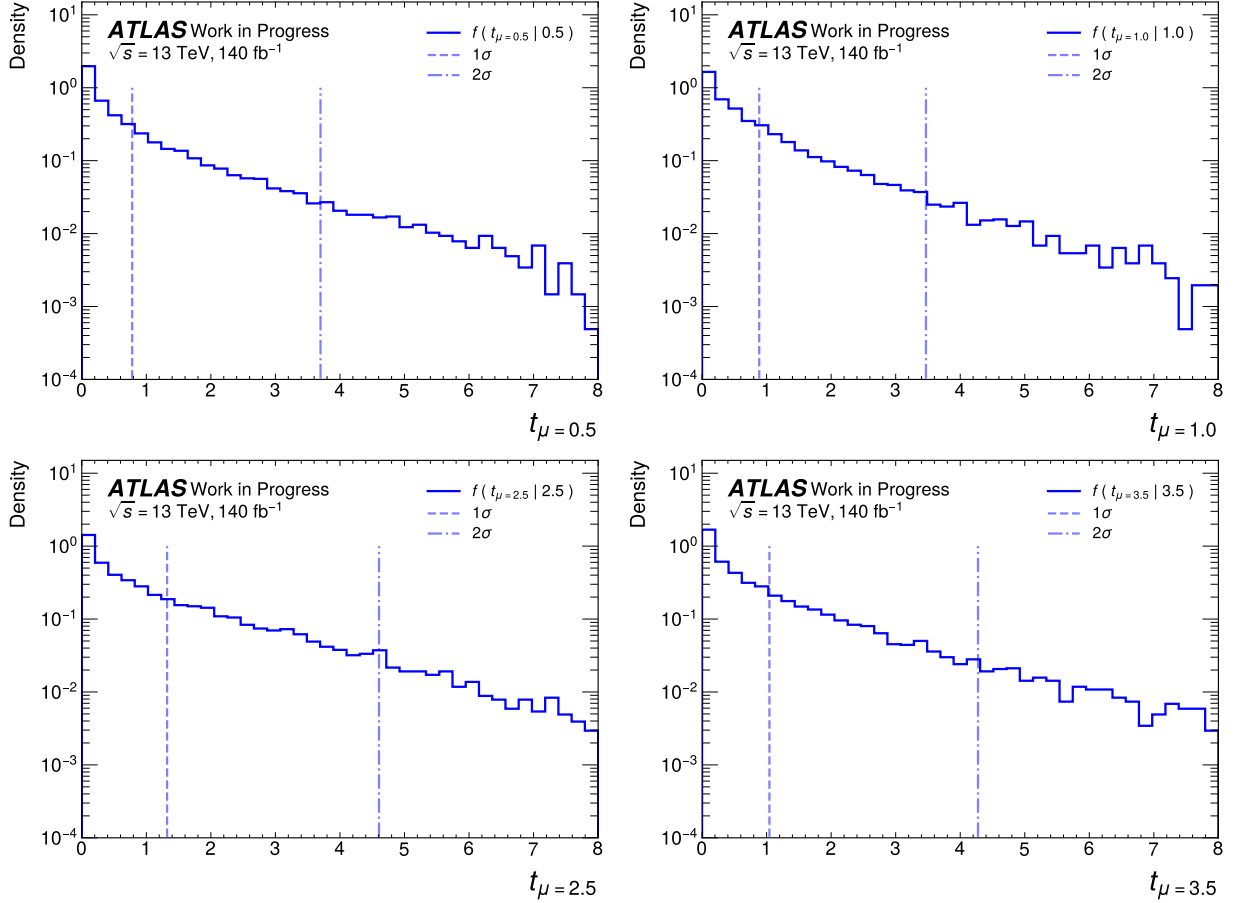


Figure 3.1: The Neyman construction for the off-shell Higgs boson analysis, defined in Chap. 4. Sampled pseudo-experiments are used to build the t_μ distribution, which is then integrated to find the boundaries corresponding to 68.24% and 95.45% pseudo-experiments, defining the 1σ and 2σ confidence belts as a function of $\mu = 0.5, 1.0, 2.5, 3.5$ (top left, top right, bottom left, bottom right). This is repeated for the full scan range of $\mu_{\text{off-shell}}$ as to build confidence belts and find the $\mu_{\text{off-shell}}$ intersection points corresponding to 1σ and 2σ uncertainties on the POI $\mu_{\text{off-shell}}$ as shown in Fig. 7.1. In an asymptotic case, the distributions will be χ^2 -distributed and the vertical lines corresponding to 68.24% and 95.45% pseudo-experiments would be located at $t_\mu \approx 1$ and $t_\mu \approx 4$ respectively. The degenerate solutions of the off-shell Higgs boson probability model breaks the asymptotic approximation.

Chapter 4

Theory and the Experimental Model

In this chapter, the physics behind the off-shell Higgs boson is briefly reviewed along with the implications of precise measurement of its cross-section at the LHC. The basics of the Standard Model will not be reviewed in this thesis, with the exception of a brief introduction to the Higgs mechanism for completeness. The experimental model of the off-shell Higgs boson is then setup, along with a motivation for the use of NSBI techniques in the corresponding signal-strength analysis.

The Standard Model (SM) of particle physics is one of the most successful models in physics in terms of experimental precision, encompassing the theory for strong and electroweak fundamental interactions. It is, however, not complete and requires as input nineteen independent parameters - the masses of the three generations of leptons and those of the six quarks spanning three generations, coupling constants of the three electroweak gauge bosons, three quark flavour mixing angles and one complex phase, mass of the Higgs boson and its quartic coupling constant and the QCD vacuum angle. By precisely measuring these input parameters using experiments, and subsequently checking the internal consistency of the SM by measuring the predicted values, the experiments at the LHC aim to find a breaking point for this incomplete model that can motivate a more general and complete theory of particle interactions that can, among other things, predict the values of the input parameters to the SM.

The Brout-Englert-Higgs mechanism in the Standard Model

The SM of particle physics describes electroweak interactions using a $SU(2)_L \times U(1)_Y$ invariant gauge field theory. The general form of the Lagrangian respecting this symmetry group does not allow for the corresponding gauge bosons to have masses. The Brout-Englert-Higgs (BEH) mechanism provides a mechanism for the breaking of the electroweak $SU(2)_L \times U(1)_Y$ symmetry to a $U(1)_{EM}$ symmetry of electromagnetism, and in the process introduces massive gauge bosons as force carriers for the extremely short-ranged weak interactions, while keeping the gauge boson (also called the photon γ) associated with the remaining $U(1)_{EM}$ symmetry massless to avoid violation of local gauge invariance. This electroweak symmetry breaking (EWSB) mechanism postulates a new massive scalar field permeating the Universe, called the Higgs field, which is a self-interacting complex doublet:

$$\phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} \phi_1 + i\phi_2 \\ \phi_3 + i\phi_4 \end{pmatrix} \quad (4.1)$$

with a weak hypercharge $Y = 1$ and weak isospin $I = 1/2$. This Higgs field contributes with the Lagrangian:

$$\mathcal{L}_{\text{Higgs}} = D_\mu \phi D^\mu \phi - V(\phi) \quad (4.2)$$

where the D_μ is the covariant derivative, essential for ensuring local gauge invariance under the SM symmetry group $\text{SU}(3)_C \times \text{SU}(2)_L \times \text{U}(1)_Y$ and the potential associated with the field is given by:

$$V(\phi) = -\frac{m_h^2}{2} \phi^\dagger \phi + \lambda (\phi^\dagger \phi)^2 \quad (4.3)$$

where m_h corresponds to the mass of the Higgs boson and λ the self-coupling strength. The negative sign on the mass term of the potential results in a shape represented as the so-called Mexican-hat potential shown in Fig. 4.1. This has the consequence that the ground state of this new field is not zero, and thus the Higgs field is unstable to symmetry breaking. The Higgs field can spontaneously break the original $\text{SU}(2)_L \times \text{U}(1)_Y$ symmetry at which it existed with $\langle \phi \rangle = 0$ and take on a unique state with non-zero vacuum expectation value (vev) v , $\langle \phi \rangle = v (= m_h/\sqrt{\lambda})$ from the infinitely many (but gauge redundant) possible values in the valley of the potential shown in Fig. 4.1. This new state is still invariant under local gauge transformations associated with the $\text{U}(1)_{\text{EM}}$ symmetry group. In this process three of the degrees of freedom in the Higgs field 4.1 are absorbed to form the three massive gauge bosons $W^\pm, Z$, and the remaining degree of freedom is associated with a massive scalar boson, called the Higgs boson H .

$$m_W^2 = \frac{g^2 v^2}{4}, \quad m_Z^2 = \frac{(g'^2 + g^2) v^2}{4} \quad (4.4)$$

where g and g' are the coupling strengths which are inputs to the SM to be determined experimentally. Using the experimental measurements of the mass of the $W^\pm$ and Z bosons, the vev associated with the Higgs field is predicted to be $v = 246$ GeV. The oscillations of this field in the new broken symmetry ground state and along the radial direction shown in Fig. 4.1 correspond to the Higgs boson, which in unitary gauge can be written simply as

$$\phi = \begin{pmatrix} 0 \\ v + \frac{h(x)}{\sqrt{2}} \end{pmatrix} \quad (4.5)$$

where $h(x)$ represents the quantum fluctuation corresponding to the physical Higgs boson. Similarly, the fermions in the SM acquire mass in this theory through a new type of interactions with the Higgs boson referred to as the Yukawa interactions.

$$m_f = \frac{h_f v}{\sqrt{2}} \quad (4.6)$$

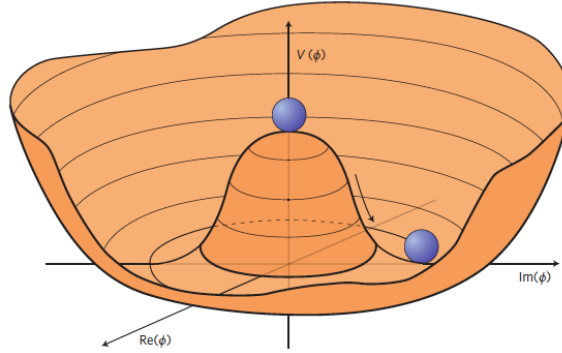


Figure 4.1: The Mexican-hat shaped potential of the Higgs boson, described by Eq. 4.3. The Universe starts out with the unstable minimum $V(\phi) = 0$, corresponding to $\text{Re}(\phi) = \text{Im}(\phi) = 0$. About 10^{-12} seconds after the big-bang, the Higgs field acquires a non-zero vacuum expectation value, that does not respect the original $\text{SU}(2)_L \times \text{U}(1)_Y$ symmetry, but acquires a new $\text{U}(1)_{\text{EM}}$ symmetry associated with Maxwell's electromagnetism in the process. Massive vector bosons and fermions are predicted in this broken symmetry Universe through interactions with the Higgs field which now has a non-zero vacuum state. The transverse oscillations of the Higgs field along the radial direction is associated with the new scalar boson called the Higgs boson.

where h_f is the Yukawa coupling strength associated with a fermion f . Note that the SM offers no explanation on the hierarchy of masses observed in reality due to different interaction strengths h_f of the various fermions, which must be determined experimentally.

If this theory describes reality, the charged leptons in the SM take on different mass values, the neutrinos remain massless and a massless gauge boson, called the photon γ , corresponding to the newly emerged $\text{U}(1)_{\text{EM}}$ symmetry comes into existence. The massive $W^\pm, Z$ bosons predicted by this so-called Weinberg-Salam model of electroweak interactions in the presence of BEH mechanism have since been discovered, hinting at the validity of this model of electroweak interactions. The prediction of massive charged leptons with a hierarchy of masses and a massless photon γ lends further validity to this model. Neutrinos, due to the observed oscillations [58], are known to be massive (with a mass close to zero) and thus do not fit the originally proposed Weinberg-Salam model. However, massless neutrinos were a postulate due to the lack of observational data at the time and it is possible to extend the SM to include neutrino masses in several ways, some of which predict the existence of new physics whereas others are straightforward corrections to the SM theory [59, 60, 61]. The true masses of the neutrinos and thus the nature of the model that predicts them remains to be experimentally determined.

4.1 The Higgs boson at the LHC

While the discoveries of the $W^\pm$ and Z bosons already pointed to the evidence of a broken symmetry Universe with a non-zero Higgs field vev, the final degree of freedom in the field - the Higgs boson - remained elusive and its properties undetermined until recently. In 2012, the ATLAS and CMS experiments announced the discovery of a scalar particle with properties consistent to that of the Higgs boson predicted by the SM using data from the first run, Run-1, of the LHC. This marked a

monumental milestone, confirming the existence of the last missing particle predicted by the SM and providing strong evidence of the BEH mechanism of electroweak symmetry breaking, along with the final inputs needed to complete the SM description. Since then, the ATLAS experiment during the Run-2 of the LHC - lasting between 2015 to 2018 - recorded over 30 times more Higgs bosons, allowing a precise investigation of the properties and interactions of this new scalar particle to confirm if it is indeed the Higgs boson as predicted by the SM.

However, in the 12 years since the discovery of the new particle, not all of its properties have been precisely measured. The shape of the Higgs potential of Eq. 4.3, for example, is fully determined using a measurement of both the Higgs boson mass and its self-interaction. And while the mass of the Higgs boson has been measured precisely to be 125.11 ± 0.11 GeV [62], the direct measurement of the self-interactions $H \rightarrow HH$ remains elusive, due to a low cross-section at the LHC further complicated by a large and destructive quantum interference. So a very important property of the Higgs field, that can be modified in reality due to effects from unknown Beyond the SM (BSM) physics, remains to be measured using the experimental program at the LHC.

Another property of the Higgs boson that is difficult to measure at the LHC is its total decay width Γ_H . The observed mass of the Higgs boson at 125 GeV results in an expected value of $\Gamma_H \sim 4.1$ MeV (See Fig. 4.2). This relatively small value of Γ_H makes it impossible to directly measure at the LHC due to limited resolution of its detectors, which is of the order $O(1 \text{ GeV})$, or about three orders of magnitude larger than the width.

The current uncertainties on these two fundamental properties of the newly discovered scalar field allow ample room for new physics effects that might reveal that the Higgs boson, or its couplings with other particles, do not behave according to the SM predictions. Precision measurements of theoretical physics parameters will be a major element of the scientific program at the LHC experiments over the coming years, potentially leading to a subtle deviation from the SM predictions or confirming its predictions to a high degree of precision. The focus of this thesis will be the precision measurement of the Higgs boson decay width Γ_H using the off-shell Higgs boson measurement described below.

4.1.1 Higgs boson production and decays

The Higgs boson is produced during the proton-proton collisions as a virtual particle at the LHC. The protons are made up of quarks and gluons, also referred to as partons in general, that interact during these high-energy collisions to produce the Higgs boson, which subsequently decays into other particles. The various production and decay interactions of the Higgs boson with other SM particles at the LHC are shown in Fig. 4.4. The main production channel, accounting for 87% of the total Higgs bosons produced during the Run-2 of the LHC is the gluon-gluon fusion process (ggF) mediated by a heavy quark ($t/b/c$) loop. This is followed by Vector-Boson Fusion (VBF) production where two W or Z bosons fuse to produce a Higgs boson, accounting for about 7% of the total production. This type of production is also referred to as Electroweak (EW) production mechanism, along with associated production of the Higgs boson with a vector boson that accounts for 4% of the total Higgs boson production. The other production channels in association with a single top quark or a top quark pair accounts for about 2% of the total production. The remaining interactions are much too subdued at

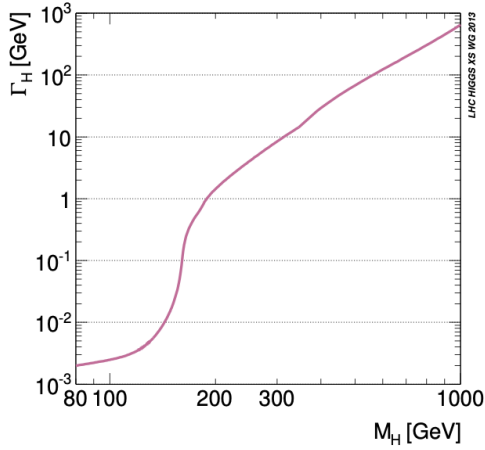


Figure 4.2: Higgs width prediction in the Standard Model as a function of its mass. Around the measured mass of 125 GeV, the Higgs width is predicted to be ~ 4.1 MeV. Figure taken from Ref. [63].

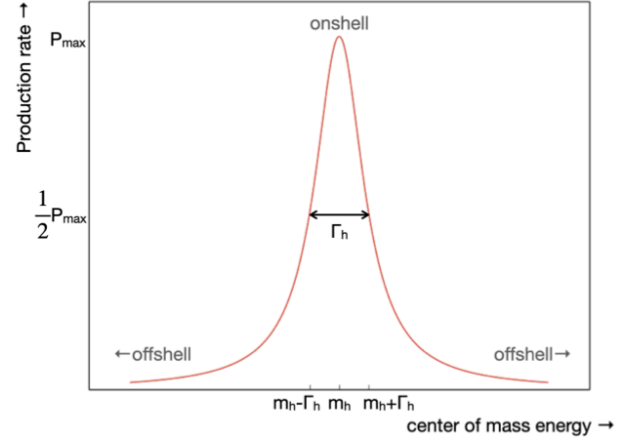


Figure 4.3: The relativistic Breit-Wigner distribution of the Higgs boson resonance with a width (Γ_H) of 4.1 MeV. The small value of the width makes it impossible to directly measure using the LHC experiments. Figure courtesy of Michiel Jan Veen.

the LHC and will not be discussed here.

The Higgs boson is an unstable particle and exists only as a virtual particle in LHC production. The SM predicts that a Higgs boson with a mass of $m_H = 125$ GeV decays pre-dominantly using the $H \rightarrow b\bar{b}$ channel, followed by $H \rightarrow W\bar{W}$ channel. The SM predictions of the Higgs boson Branching Ratios is shown in Fig. 4.6 as a function of the Higgs boson mass, where the branching ratio is $BR = \Gamma_{H \rightarrow XX} / \Gamma_H$, where $\Gamma_{H \rightarrow XX}$ is the decay cross-section of the $H \rightarrow XX$ channel and Γ_H is the total *decay width* of the Higgs boson (See Fig. 4.3). The decay probabilities into various final states for a Higgs boson with $m_H = 125$ GeV are shown in Fig. 4.5. While decays like $H \rightarrow b\bar{b}/W\bar{W}/g\bar{g}$ present large probabilities, they are experimentally difficult to isolate and measure precisely due to hadronization in the case of quarks and gluons and W -boson decays into invisible neutrinos. As such, while the $H \rightarrow Z\bar{Z}$ and $H \rightarrow \gamma\gamma$ decay channels have much lower probabilities, they were used to first discover the Higgs boson, due to the clean final state from these decays.

4.1.2 Higgs boson interactions

A key prediction of the EWSB theory is that the Higgs boson couplings with the other particles scale with their masses, and the coupling strengths can be predicted precisely once the mass of the particles are known experimentally. By comparing these predictions based on the observed masses of the particles, an independent consistency check can be performed on the SM. The precise measurements of these couplings moreover can be used to put stringent constraints on theories beyond the SM. Fig. 4.7 shows the experimental measurements of the coupling strengths of the SM particles as a function of their mass, showing compatibility within experimental precision to the SM prediction shown by the diagonal line.

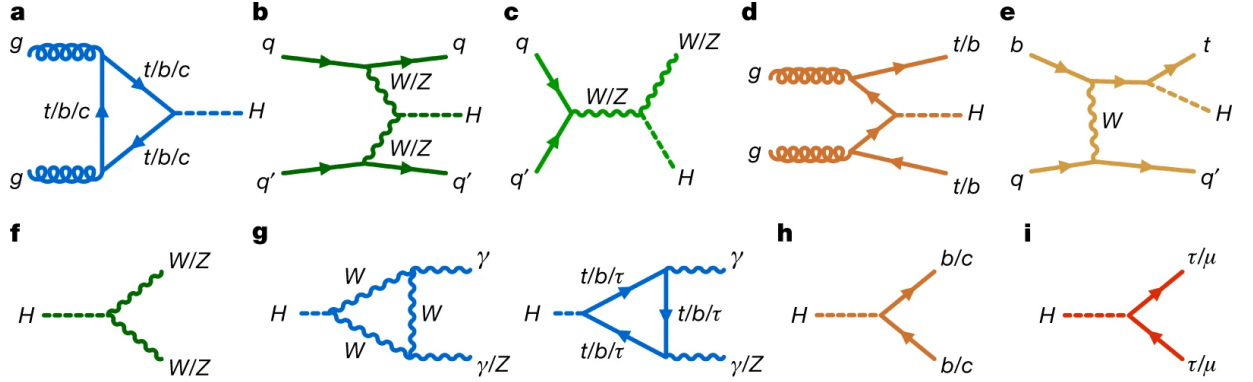


Figure 4.4: The various production and decay channels of the Higgs boson at the LHC from Ref. [64]. The Higgs boson is produced via gluon–gluon fusion (a), vector boson fusion (VBF; b) and associated production with vector bosons (c), top or b quark pairs (d), or a single top quark (e). f–i, The Higgs boson decays into a pair of vector bosons (f), a pair of photons or a Z boson and a photon (g), a pair of quarks (h), and a pair of charged leptons (i). Loop-induced Higgs boson interactions with gluons or photons are shown in blue, and processes involving couplings to W or Z bosons in green, to quarks in orange, and to leptons in red. Two different shades of green (orange) are used to separate the VBF and VH (and tH) production processes.

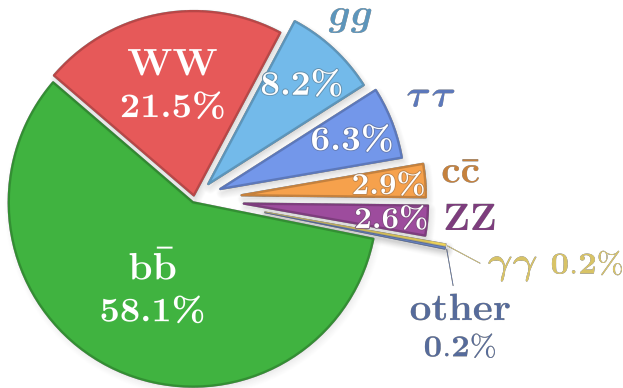


Figure 4.5: The pie chart showing the probabilities of Higgs boson decays to various possible final states. Figure from Ref. [65].

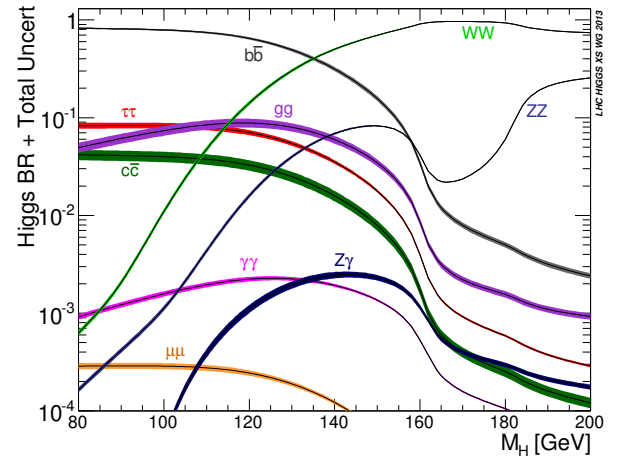


Figure 4.6: SM predictions of the Higgs boson decay branching ratios to various particle pairs, as a function of the Higgs boson mass. Figure from Ref. [63].

4.2 The off-shell Higgs boson

The Higgs boson is produced at the LHC as a virtual particle, with various production and decay channels described in Sec. 4.1.1, and shown in Fig. 4.4. The dynamic term of the Higgs Lagrangian of Eq. 4.2 gives the Higgs propagator:

$$D_H(\hat{s}) = \frac{1}{(\hat{s}^2 - m_H^2)^2 + m_H^2 \Gamma_H^2} \quad (4.7)$$

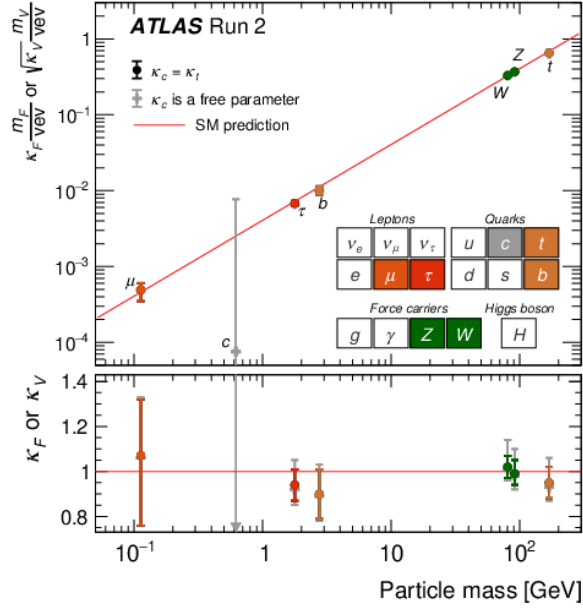


Figure 4.7: The Higgs boson coupling strengths to the various fermions and scalar bosons, shown using modifiers measured experimentally. The vertical bar on each point denotes the 68% confidence interval. The lower panel shows the values of the coupling strength modifiers. Data are from ATLAS Run-2. Figure taken from Ref. [66].

This corresponds to a resonance effect shown in Fig. 4.3 approximated by the so-called Breit-Wigner distribution. For the Higgs boson decaying into a pair of vector bosons $H \rightarrow VV$ (where $V = W, Z$ bosons), the differential cross-section scales as:

$$\frac{d\sigma^{H \rightarrow VV}}{dm_{VV}^2} \propto \frac{g_{\text{prod}}^2(\hat{s})g_{\text{decay}}^2(\hat{s})}{(m_{VV}^2 - m_H^2)^2 + m_H^2\Gamma_H^2} \quad (4.8)$$

where $\hat{s} = m_{VV}^2$ is the reconstructed mass of the Higgs boson, using the invariant mass of the VV final state, giving an estimate of its virtuality as well as the partonic center-of-mass energy, and $g_{\text{prod}}(\hat{s})$ and $g_{\text{decay}}(\hat{s})$ are, respectively, the effective couplings associated with the Higgs boson production and decay as a function of the Higgs boson virtuality $\hat{s}$.

The Breit-Wigner distribution of $D_H(\hat{s})$ drops off quickly while moving away from the peak at $m_H = 125$ GeV, with the relative contribution of the tails more than n Higgs widths ($n \cdot \Gamma_H$) from $m_H = 125$ GeV can be estimated as $1/(n\pi)$, since [67]:

$$\int_{(m_H - n\Gamma_H)^2}^{(m_H + n\Gamma_H)^2} \frac{dm_{VV}^2}{2\pi} \frac{1}{(m_{VV}^2 - m_H^2)^2 + m_H^2\Gamma_H^2} \approx \frac{1}{2m_H\Gamma_H} \left(1 - \frac{1}{n\pi}\right) \quad (4.9)$$

As such, one does not expect the Higgs boson to be produced with significant cross-sections at masses $m_H \gg 125$ GeV. However, the Breit-Wigner distribution only describes the scaling of the Higgs propagator. In principle, $\hat{s}$ dependence from the production and decay couplings ($g_{\text{prod}}(\hat{s})$, $g_{\text{decay}}(\hat{s})$) can affect the total production cross-section differently at different energies. This is true in the $H \rightarrow VV$ decays, if the Higgs boson is produced using the ggF, VBF or VH production channels shown in Fig. 4.4.

In this analysis, focus is on the $H \rightarrow ZZ$ decay channel.

As shown in Fig. 4.8, in the $H \rightarrow ZZ$ channel the off-shell Higgs boson production (solid black curve) is significantly boosted in regions $m_{ZZ} \geq 180.0$ GeV, compared to predictions from a simple Breit-Wigner distribution (dashed black curve). This is due to the fact that while the Higgs propagator scales down as $\sqrt{s} \gg 125$ GeV, the coupling terms corresponding to HZZ and $t\bar{t}H$ (in the gluon-fusion ggF production as shown in Fig (a) of. 4.4) interactions receive a boost, due to the t -quark and the Z -boson going on-shell in this kinematic phase space, that compensates for the decrease in amplitude due to the propagator.

$$g_{HZZ}(\hat{s}) \sim \hat{s}^2 \quad \text{for } \hat{s} \gtrsim 2m_Z \quad (4.10)$$

$$g_{t\bar{t}H}(\hat{s}) \sim m_t \hat{s} \quad \text{for } \hat{s} \gtrsim 2m_t \quad (4.11)$$

where m_Z and m_t denote the pole masses of the Z -boson and the t -quark respectively. The boost from these two threshold effects at $m_{ZZ} \sim 2m_Z, 2m_t$ cancel out the $\sim 1/(m_{ZZ}^2)^2$ down-scaling of the Higgs propagator in Eq. 4.8 in certain production channels - the ggF production channel due to the t -quark loop and the VBF and VH production channels due to the HZZ production mechanism. This enhancement in Higgs boson production rate in the off-shell region compared to a simple Breit-Wigner resonance associated with the Higgs propagator is shown in Fig. 4.8. Note that in this off-shell phase space $m_{VV} \gtrsim 180.0$:

$$D_H(m_{ZZ}) \approx \frac{1}{(m_{ZZ}^2 - m_H^2)^2} \quad (4.12)$$

giving

$$\frac{d\sigma_{\text{off-shell}}^{H \rightarrow VV}}{dm_{VV}^2} \propto \frac{g_{\text{prod}}^2(\hat{s})g_{\text{decay}}^2(\hat{s})}{(m_{VV}^2 - m_H^2)^2} \quad (4.13)$$

where it is interesting to note that the cross-section loses dependence on the decay width Γ_H . While the same behaviour described here can be applied for $H \rightarrow WW$ decay channel, this thesis focuses on the $H \rightarrow ZZ$ production channel - in particular to the experimentally clean $ZZ \rightarrow 4\ell$ final state.

Constraining Higgs width using the off-shell Higgs boson

As mentioned above, the Higgs width is an important and as-yet poorly measured piece of the SM puzzle. This is because it is too narrow $\Gamma_H \sim 4.1$ MeV to be measured directly using the resonance lineshape at the LHC experiments due to their large resolution of $O(1 \text{ GeV})$. To visualize this, note the magnitudes of the width in the theoretical resonance of Fig. 4.3 with the width of the experimentally observed $m_{4\ell}$ lineshape in Fig. 2.2a. This motivates the need for indirect measurements of the Higgs width using theoretical model assumptions. An important thing to note is that while the cross-section associated with the on-shell Higgs boson is dependent on the decay width

$$\sigma_{\text{on-shell}}^{H \rightarrow VV} \propto \frac{g_{\text{prod}}^2(m_H)g_{\text{decay}}^2(m_H)}{\Gamma_H}, \quad (4.14)$$

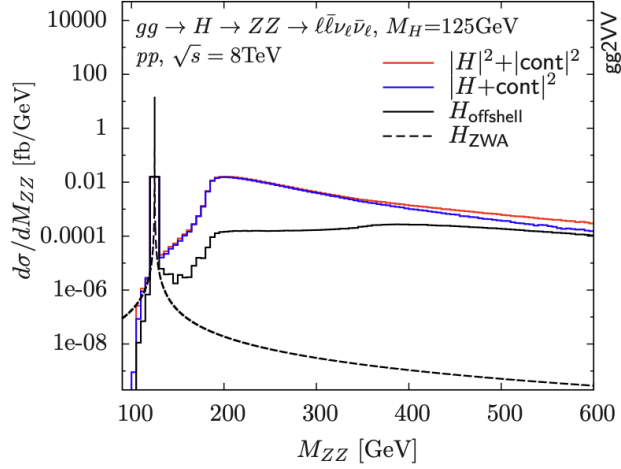


Figure 4.8: The m_{ZZ} distribution comparison between a Breit-Weigner lineshape centered at $m_H = 125$ GeV and the actual lineshape due to threshold effects at $m_{ZZ} \gtrsim 180$ GeV and $m_{ZZ} \gtrsim 350$ GeV. Here the H stands for the Higgs signal process and $cont$ stands for the interfering background shown in Fig. 4.9. ZWA stands for the zero-width approximation that assumes a Breit-Weigner distribution with a narrow decay width. Figure taken from Ref. [68].

there is degeneracy between the effective couplings $g(m_H)$ and the Higgs boson decay width Γ_H that does not allow for inference on the Γ_H from an experimental measurement of this cross-section. Note, however, that the off-shell cross-section in Eq. 4.13 is independent of the Higgs boson decay width Γ_H , giving the integrated cross-section in the $m_{VV} \gtrsim 180.0$ GeV phase space:

$$\sigma_{\text{off-shell}}^{H^* \rightarrow VV} \propto g_{\text{prod}}^2(\hat{s}) g_{\text{decay}}^2(\hat{s}) \quad (4.15)$$

As such, a precise measurement of the off-shell and on-shell signal strengths:

$$\mu_{\text{off-shell}} = \frac{\sigma_{\text{off-shell,observed}}^{H^* \rightarrow VV}}{\sigma_{\text{off-shell,SM}}^{H^* \rightarrow VV}}, \quad \mu_{\text{on-shell}} = \frac{\sigma_{\text{on-shell,observed}}^{H \rightarrow VV}}{\sigma_{\text{on-shell,SM}}^{H \rightarrow VV}} \quad (4.16)$$

gives

$$\frac{\mu_{\text{off-shell}}}{\mu_{\text{on-shell}}} = \frac{\Gamma_H}{\Gamma_{H,\text{SM}}} \quad (4.17)$$

under the assumption that both the on-shell and off-shell effective coupling modifiers $\left[(g_{\text{prod}}^2 g_{\text{decay}}^2) / (g_{\text{prod,SM}}^2 g_{\text{decay,SM}}^2) \right]_{\text{off-shell}} \approx \left[(g_{\text{prod}}^2 g_{\text{decay}}^2) / (g_{\text{prod,SM}}^2 g_{\text{decay,SM}}^2) \right]_{\text{on-shell}}$, which roughly translates to the assumption that no new physics affects the effective coupling strengths in the off-shell and on-shell phase spaces differently (i.e. has a strong scale $\hat{s}$ -dependence). This has been proposed in [69, 70, 71], as a minimally model-dependent and powerful way to perform inference on the Higgs boson decay width Γ_H using the LHC experiments. Other approaches, with different model dependencies, include precisely measuring the on-shell Higgs boson interference that modify the resonance lineshape [72] in $H \rightarrow Z\gamma, \gamma\gamma$ final states, and measuring the $t\bar{t}H^*$ production using the $H^* \rightarrow t\bar{t}$ decay channel in the off-shell phase space [73]. The approach presented in this analysis with the off-shell



Figure 4.9: Illustrative leading-order Feynman diagrams for $gg \rightarrow ZZ$ production. The diagrams indicate the effective couplings g_g and g_V . Diagram a corresponds to the signal component and diagram b corresponds to the background component. A large destructive interference between the two components is present in the off-shell regime.

$H^* \rightarrow ZZ$ production is the most precise determination of Γ_H among these proposed methods.

4.3 The Experimental Model of off-shell Higgs boson

The measurement of the off-shell Higgs boson production in the $H^* \rightarrow ZZ \rightarrow 4\ell$ channel is presented here as a signal strength measurement as described in Sec. 2.5.2, and defined in Eq. 4.16. The measurement is also interpreted using a model in which the effective couplings of the Higgs boson to gluons and to electroweak vector bosons in the SM can have independent anomalous scalar modifications. This latter interpretation, known as the κ -framework [63] can be understood as a sector of a larger Higgs Effective Field Theory (HEFT) [74].

The gluon–gluon fusion (ggF) off-shell Higgs boson production and subsequent decay into a ZZ pair, $gg \rightarrow ZZ$, can be described by using the Higgs boson effective coupling constants $g_g(\hat{s})$ and $g_V(\hat{s})$ to gluons and vector bosons, shown in Fig. 4.9a. The effective couplings $g_g(=g_{\text{prod}})$ and $g_V(=g_{\text{decay}})$ depend on the Higgs boson virtuality $\hat{s}$, but the notation is suppressed hereafter for simplicity. The Higgs boson is represented by H^* to denote explicitly that its virtuality is well above the pole mass $m_H = 125 \text{ GeV}$ [75].

In $gg \rightarrow ZZ$ production, the *signal* (S) component is defined at leading order (LO) in perturbation theory by the absolute value squared of the amplitude of the diagram in Fig. 4.9a. The signal contribution scales as $g_g^2 g_V^2$. The *background* (B) component is defined at LO in perturbation theory by the absolute value squared of the amplitude of the diagram in Fig. 4.9b. The background component is independent of g_g and g_V . The *interference* (I) between the two diagrams scales as $g_g g_V$. The integrated interference between the two diagrams is negative, as required by perturbative unitarity conservation [76].

The effective coupling between gluons and the Higgs boson g_g is represented by a blob in Fig. 4.9a since in the SM, at LO in perturbation theory, it can be resolved to a fermion triangle loop dominated by the top-quark contribution. In several BSM scenarios, the contribution of new heavy particles cannot be resolved at the scales probed by this measurement, but can modify the Higgs boson effective couplings as well as the background process.

The same concepts can be applied to the electroweak (EW) production of $q\bar{q} \rightarrow ZZ + 2j \rightarrow 4\ell + 2j$. In this case, the *signal* scales as g_V^4 , the *interference* scales as g_V^2 , and the *background* component is

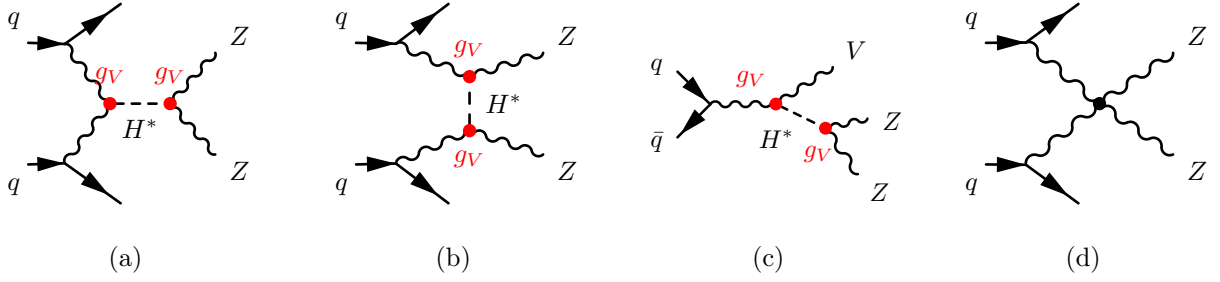


Figure 4.10: Illustrative leading-order Feynman diagrams for the electroweak $qq \rightarrow ZZ + 2j$ processes. The diagrams indicate the effective couplings g_V used to define the off-shell Higgs boson production signal strengths. Diagrams a, b, and c correspond to the vector boson fusion, t -channel and ZH signal components, respectively. In diagram c, one of the Z bosons decays into a $q\bar{q}$ pair. Diagram d corresponds to the electroweak background component. A large destructive interference is present in the off-shell regime between the vector boson fusion signal and the background components.

independent of g_V . The LO Feynman diagrams for the production of EW $q\bar{q} \rightarrow ZZ + 2j \rightarrow 4\ell + 2j$ are shown in Fig. 4.10.

The scaling of the cross-sections with the effective couplings g_g and g_V defines uniquely each component (signal, interference, and background) in the ggF and EW production of off-shell Higgs bosons. The probability density model used to measure the off-shell Higgs boson production is defined as function of the coupling modifiers $\kappa_g = g_g/g_{g,\text{SM}}$ and $\kappa_V = g_V/g_{V,\text{SM}}$, independently of the Higgs boson virtuality, which are used to define the signal strengths:

$$\mu_{\text{off-shell}}^{\text{ggF}} = \kappa_{g,\text{off-shell}}^2 \kappa_{V,\text{off-shell}}^2, \quad \mu_{\text{off-shell}}^{\text{EW}} = \kappa_{V,\text{off-shell}}^4, \quad (4.18)$$

where the subscript *off-shell* indicates that the modifier only affects processes with virtuality sufficiently above the pole mass. These signal strengths are used in the definition of a probability density model that is used to interpret the collected data:

$$p(\mathbf{x} | \mu_{\text{off-shell}}^{\text{ggF}}, \mu_{\text{off-shell}}^{\text{EW}}) = \frac{1}{\nu(\mu_{\text{off-shell}}^{\text{ggF}}, \mu_{\text{off-shell}}^{\text{EW}})} \times \left[\mu_{\text{off-shell}}^{\text{ggF}} \nu_{\text{S}}^{\text{ggF}} p_{\text{S}}^{\text{ggF}}(\mathbf{x}) + \sqrt{\mu_{\text{off-shell}}^{\text{ggF}}} \nu_{\text{I}}^{\text{ggF}} p_{\text{I}}^{\text{ggF}}(\mathbf{x}) + \nu_{\text{B}}^{\text{ggF}} p_{\text{B}}^{\text{ggF}}(\mathbf{x}) + \mu_{\text{off-shell}}^{\text{EW}} \nu_{\text{S}}^{\text{EW}} p_{\text{S}}^{\text{EW}}(\mathbf{x}) + \sqrt{\mu_{\text{off-shell}}^{\text{EW}}} \nu_{\text{I}}^{\text{EW}} p_{\text{I}}^{\text{EW}}(\mathbf{x}) + \nu_{\text{B}}^{\text{EW}} p_{\text{B}}^{\text{EW}}(\mathbf{x}) + \nu_{\text{NIPNI}}(\mathbf{x}) \right], \quad (4.19)$$

where $\mathbf{x}$ is a vector of reconstructed observables, which are defined in Sec. 5.3, and where $p_c(x)$ and ν_c are the probability density function and the expected yield for process c , respectively. Eq. 4.19 is the general model based on the κ -framework. The main focus of this analysis will be the measurement of the signal strength parameter $\mu_{\text{off-shell}} = \mu_{\text{off-shell}}^{\text{ggF}} = \mu_{\text{off-shell}}^{\text{EW}}$, defined in Eq. 4.16, for which the model is written as:

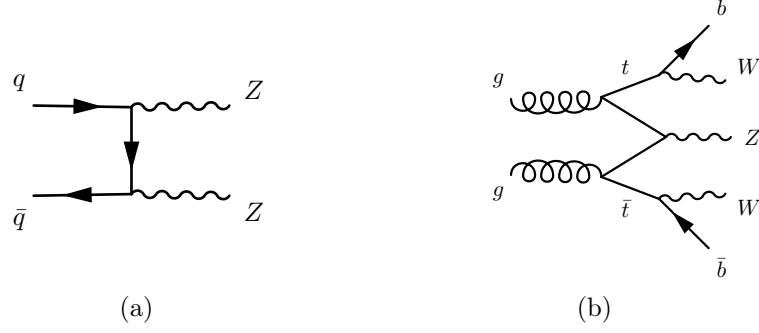


Figure 4.11: Illustrative leading-order Feynman diagrams for non-interfering background processes. Diagram a shows the leading $q\bar{q}ZZ$ process and diagram b shows the top-quark induced sub-leading VVV process. The leptonic decays of the Z and W bosons are not shown.

$$\begin{aligned}
 p(\mathbf{x}|\mu_{\text{off-shell}}) = & \frac{1}{\nu(\mu_{\text{off-shell}})} \times \\
 & \left[\mu_{\text{off-shell}} \nu_{\text{S}}^{\text{ggF}} p_{\text{S}}^{\text{ggF}}(\mathbf{x}) + \sqrt{\mu_{\text{off-shell}}} \nu_{\text{I}}^{\text{ggF}} p_{\text{I}}^{\text{ggF}}(\mathbf{x}) + \nu_{\text{B}}^{\text{ggF}} p_{\text{B}}^{\text{ggF}}(\mathbf{x}) + \right. \\
 & \left. \mu_{\text{off-shell}} \nu_{\text{S}}^{\text{EW}} p_{\text{S}}^{\text{EW}}(\mathbf{x}) + \sqrt{\mu_{\text{off-shell}}} \nu_{\text{I}}^{\text{EW}} p_{\text{I}}^{\text{EW}}(\mathbf{x}) + \nu_{\text{B}}^{\text{EW}} p_{\text{B}}^{\text{EW}}(\mathbf{x}) + \nu_{\text{NI}} p_{\text{NI}}(\mathbf{x}) \right], \quad (4.20)
 \end{aligned}$$

This is the more familiar signal-strength model that was defined in Sec. 2.5.2. Here, the total yield can also be written as a function of $\mu_{\text{off-shell}}$:

$$\begin{aligned}
 \nu(\mu_{\text{off-shell}}) = & f_{\text{S}}(\mu_{\text{off-shell}}) \nu_{\text{S}}^{\text{ggF}} + f_{\text{I}}(\mu_{\text{off-shell}}) \nu_{\text{I}}^{\text{ggF}} + f_{\text{B}}(\mu_{\text{off-shell}}) \nu_{\text{B}}^{\text{ggF}} \\
 & + f_{\text{S}}(\mu_{\text{off-shell}}) \nu_{\text{S}}^{\text{EW}} + f_{\text{I}}(\mu_{\text{off-shell}}) \nu_{\text{I}}^{\text{EW}} + f_{\text{B}}(\mu_{\text{off-shell}}) \nu_{\text{B}}^{\text{EW}} + f_{\text{NI}}(\mu_{\text{off-shell}}) \nu_{\text{NI}} \quad (4.21)
 \end{aligned}$$

where $f_{\text{S}}(\mu) = \mu$, $f_{\text{I}}(\mu) = \sqrt{\mu}$ and $f_{\text{B}}(\mu) = f_{\text{NI}}(\mu) = 1$. The general forms of these multipliers for each process used in the off-shell Higgs boson analysis are defined in Tab. 4.1.

Since the multipliers $f_c(\mu)$ are process dependent, the simplification of Eq. 2.33 is not possible in the off-shell Higgs boson modelling and hence a *fixed* summary statistic $u(\mathbf{x})$ is not a sufficient summary statistic for the full parameter $\mu_{\text{off-shell}}$ range being scanned. This motivates the use of parameterized optimal observables or NSBI techniques for this measurement. For more discussion, refer to Sec. 2.5.

The term $p_{\text{NI}}(\mathbf{x})$ represents the probability density for processes that do not interfere with the ggF and EW processes described above. The leading non-interfering process is $q\bar{q} \rightarrow ZZ \rightarrow 4\ell$ production via a t -channel exchange at LO. Triboson VVV processes, including those from top-quark decays $t\bar{t}V$, are subleading processes, but are also included in the analysis. Fig. 4.11a shows the LO Feynman diagram of the leading non-interfering $q\bar{q} \rightarrow ZZ$ process ($q\bar{q}ZZ$) and Fig. 4.11b shows the corresponding LO Feynman diagram of the subleading non-interfering top-quark-induced VVV process.

Monte Carlo (MC) simulated samples are used to model the probability densities $p_c(\mathbf{x})$ in Eq. 4.19.

Due to the technical challenges associated with the efficient production of interference-only MC simulations [77], the interference terms in Eq. 4.19 are not generated separately, but inferred from samples generated with signal, interference, and background terms (SBI sample). For ggF production, a single SBI sample is generated and the interference term $\nu_I^{\text{ggF}} p_I^{\text{ggF}}(\mathbf{x})$ is calculated as:

$$\nu_I^{\text{ggF}} p_I^{\text{ggF}}(\mathbf{x}) = \nu_{\text{SBI}}^{\text{ggF}} \cdot p_{\text{SBI}}^{\text{ggF}}(\mathbf{x}) - \nu_S^{\text{ggF}} p_S^{\text{ggF}}(\mathbf{x}) - \nu_B^{\text{ggF}} p_B^{\text{ggF}}(\mathbf{x}). \quad (4.22)$$

In the simulation of EW production, it is impossible to generate an off-shell signal-only sample. Due to the diagrams shown in Fig. 4.10b and Fig. 4.10c, where an s -channel Higgs boson propagator is absent, there is always contamination of on-shell events, which are formally part of the EW B process. Instead of generating pure signal and interference samples, two linear combinations (EWSBI₁ and EWSBI₁₀) are used to model the EW processes:

$$\begin{aligned} \nu_{\text{SBI}_1}^{\text{EW}} p_{\text{SBI}_1}^{\text{EW}}(\mathbf{x}) &= \nu_S^{\text{EW}} p_S^{\text{EW}}(\mathbf{x}) + \nu_I^{\text{EW}} p_I^{\text{EW}}(\mathbf{x}) + \nu_B^{\text{EW}} p_B^{\text{EW}}(\mathbf{x}), \\ \nu_{\text{SBI}_{10}}^{\text{EW}} p_{\text{SBI}_{10}}^{\text{EW}}(\mathbf{x}) &= 10\nu_S^{\text{EW}} p_S^{\text{EW}}(\mathbf{x}) + \sqrt{10}\nu_I^{\text{EW}} p_I^{\text{EW}}(\mathbf{x}) + \nu_B^{\text{EW}} p_B^{\text{EW}}(\mathbf{x}). \end{aligned} \quad (4.23)$$

These samples contain the signal, interference, and background components, including any possible on-shell contamination, albeit with different fractions. An additional sample is generated exclusively modeling the diagram in Fig. 4.10d to describe the EW B process $\nu_B^{\text{EW}} p_B^{\text{EW}}(\mathbf{x})$. While this sample does not contain the on-shell contamination from diagrams Fig. 4.10b and Fig. 4.10c discussed above, it was found to be a good approximation for the missing on-shell parts as well.

The EW SBI₁₀ sample is simulated by choosing the effective coupling between the Higgs boson and the EW vector bosons to achieve an EW signal cross-section ten times larger than the SM value ($\kappa_V^4 = 10$). The change of the effective coupling scales the interference component by a factor $\sqrt{10}$, yielding the expression in Eq. 4.23. When simulating the EW SBI₁₀ sample, the Higgs boson width is modified to keep the on-shell effective coupling unchanged at g_V^4/Γ_H . Linear combinations of the EW B, EW SBI₁, and EW SBI₁₀ samples can be inverted to obtain the EW signal, interference and background processes separately, as shown in Tab. 4.1.

As described earlier, the non-interfering background process $\nu_{\text{NIPNI}}(\mathbf{x})$ in Eq. 4.19 contains both the leading $q\bar{q}ZZ$ and the subleading VVV processes:

$$\nu_{\text{NIPNI}}(\mathbf{x}) = \nu_{q\bar{q}ZZ} p_{q\bar{q}ZZ}(\mathbf{x}) + \nu_{VVV} p_{VVV}(\mathbf{x}). \quad (4.24)$$

All VVV processes are considered separately as a single non-interfering background process, whereas the $ZZZ \rightarrow 4\ell + \text{jets}$ sample is modeled as part of the EW background process.

A data-driven normalization is introduced for the leading $q\bar{q} \rightarrow ZZ$ process as a function of the number of reconstructed jets, which is one of the observables in the vector $\mathbf{x}$ defined in Sec. 5.3:

Table 4.1: Definition of processes in the probability model used to interpret the data and the simulated samples used to describe them. The multipliers define both the measured off-shell Higgs boson production signal strength and the data-driven normalization of leading backgrounds. The $(H^* \rightarrow)$ notation is used for the SBI process. The multipliers are obtained after substituting the interference terms of Eqs. 4.22 and 4.23 into Eq. 4.19, since interference-only samples are not simulated. For instance, in the case of ggF production: $\mu\nu_S + \sqrt{\mu}\nu_I + \nu_B = (\mu - \sqrt{\mu})\nu_S + \sqrt{\mu}(\nu_S + \nu_I + \nu_B) + (1 - \sqrt{\mu})\nu_B$.

Process	Multipliers $f_c(\mu, \theta_{qq})$	Samples
ggF S	$\mu_{\text{off-shell}}^{\text{ggF}} - \sqrt{\mu_{\text{off-shell}}^{\text{ggF}}}$	$gg \rightarrow H^* \rightarrow ZZ \rightarrow 4\ell$
ggF SBI	$\sqrt{\mu_{\text{off-shell}}^{\text{ggF}}}$	$gg \rightarrow (H^* \rightarrow) ZZ \rightarrow 4\ell$ ($\kappa_V^2 = 1$)
ggF B	$1 - \sqrt{\mu_{\text{off-shell}}^{\text{ggF}}}$	$gg \rightarrow ZZ \rightarrow 4\ell$ ($\kappa_V^2 = 0$)
EW B	$\frac{(1 - \sqrt{10})\mu_{\text{off-shell}}^{\text{EW}} + 9\sqrt{\mu_{\text{off-shell}}^{\text{EW}}} - 10 + \sqrt{10}}{-10 + \sqrt{10}}$	$EW q\bar{q} \rightarrow ZZ + 2j \rightarrow 4\ell + 2j$ ($\kappa_V^4 = 0$) $ZZZ \rightarrow 4\ell + 2j$
EW SBI ₁	$\frac{\sqrt{10}\mu_{\text{off-shell}}^{\text{EW}} - 10\sqrt{\mu_{\text{off-shell}}^{\text{EW}}}}{-10 + \sqrt{10}}$	$EW q\bar{q} \rightarrow (H^* \rightarrow) ZZ + 2j \rightarrow 4\ell + 2j$ ($\kappa_V^4 = 1$) $ZZZ \rightarrow 4\ell + 2j$
EW SBI ₁₀	$\frac{-\mu_{\text{off-shell}}^{\text{EW}} + \sqrt{\mu_{\text{off-shell}}^{\text{EW}}}}{-10 + \sqrt{10}}$	$EW q\bar{q} \rightarrow (H^* \rightarrow) ZZ + 2j \rightarrow 4\ell + 2j$ ($\kappa_V^4 = 10$) $ZZZ \rightarrow 4\ell + 2j$
$q\bar{q}ZZ$ $n_{\text{jets}} = 0$	$\theta_{q\bar{q}ZZ}^{0j}$	$q\bar{q} \rightarrow ZZ \rightarrow 4\ell$
$q\bar{q}ZZ$ $n_{\text{jets}} = 1$	$\theta_{q\bar{q}ZZ}^{0j} \theta_{q\bar{q}ZZ}^{1j}$	$q\bar{q} \rightarrow ZZ \rightarrow 4\ell$
$q\bar{q}ZZ$ $n_{\text{jets}} \geq 2$	$\theta_{q\bar{q}ZZ}^{0j} \theta_{q\bar{q}ZZ}^{1j} \theta_{q\bar{q}ZZ}^{2j}$	$q\bar{q} \rightarrow ZZ \rightarrow 4\ell$
VVV	—	$WWZ \rightarrow 4\ell$
		$WZZ \rightarrow 4\ell$
		$t\bar{t}Z \rightarrow 4\ell$

$$\begin{aligned}
 \nu_{\text{NIPNI}}(\mathbf{x}) = & \theta_{q\bar{q}ZZ}^{0j} \nu_{q\bar{q}ZZ}^{0j} p_{q\bar{q}ZZ}^{0j}(\mathbf{x}) + \theta_{q\bar{q}ZZ}^{0j} \theta_{q\bar{q}ZZ}^{1j} \nu_{q\bar{q}ZZ}^{1j} p_{q\bar{q}ZZ}^{1j}(\mathbf{x}) \\
 & + \theta_{q\bar{q}ZZ}^{0j} \theta_{q\bar{q}ZZ}^{1j} \theta_{q\bar{q}ZZ}^{2j} \nu_{q\bar{q}ZZ}^{2j} p_{q\bar{q}ZZ}^{2j}(\mathbf{x}) + \nu_{VVV} p_{VVV}(\mathbf{x}).
 \end{aligned} \tag{4.25}$$

The parameter $\theta_{q\bar{q}ZZ}^{0j}$ provides a data-driven normalization for the total $q\bar{q} \rightarrow ZZ \rightarrow 4\ell$ observed yield. The parameter $\theta_{q\bar{q}ZZ}^{1j}$ ($\theta_{q\bar{q}ZZ}^{2j}$) provides a data-driven normalization for the ratio of the observed yield of $q\bar{q} \rightarrow ZZ \rightarrow 4\ell$ events with one (two) reconstructed jet and that with zero (one) reconstructed jets. The process referred to as $\nu_{q\bar{q}ZZ}^{2j} p_{q\bar{q}ZZ}^{2j}(\mathbf{x})$ includes all events with at least two reconstructed jets.

Tab. 4.1 summarizes the processes that are used in the model after all transformations, as well as the simulated samples used to describe them.

4.4 Application of the Search-Oriented Mixture Model

The probability density ratio model for the off-shell Higgs boson is now ready to be built:

$$\begin{aligned}
\frac{p(\mathbf{x}|\mu_{\text{off-shell}}^{\text{ggF}}, \mu_{\text{off-shell}}^{\text{EW}})}{p_{\text{ref}}(\mathbf{x})} &= \frac{1}{\nu(\mu_{\text{off-shell}}^{\text{ggF}}, \mu_{\text{off-shell}}^{\text{EW}})} \times \\
&\left[\left(\mu_{\text{off-shell}}^{\text{ggF}} - \sqrt{\mu_{\text{off-shell}}^{\text{ggF}}} \right) \nu_{\text{S}}^{\text{ggF}} \frac{p_{\text{S}}^{\text{ggF}}}{p_{\text{ref}}}(\mathbf{x}) + \sqrt{\mu_{\text{off-shell}}^{\text{ggF}}} \nu_{\text{SBI}}^{\text{ggF}} \frac{p_{\text{SBI}}^{\text{ggF}}}{p_{\text{ref}}}(\mathbf{x}) \right. \\
&+ \left(1 - \sqrt{\mu_{\text{off-shell}}^{\text{ggF}}} \right) \nu_{\text{B}}^{\text{ggF}} \frac{p_{\text{B}}^{\text{ggF}}}{p_{\text{ref}}}(\mathbf{x}) + \left(\mu_{\text{off-shell}}^{\text{EW}} - \sqrt{\mu_{\text{off-shell}}^{\text{EW}}} \right) \nu_{\text{S}}^{\text{EW}} \frac{p_{\text{S}}^{\text{EW}}}{p_{\text{ref}}}(\mathbf{x}) \\
&+ \sqrt{\mu_{\text{off-shell}}^{\text{EW}}} \nu_{\text{SBI}}^{\text{EW}} \frac{p_{\text{SBI}}^{\text{EW}}}{p_{\text{ref}}}(\mathbf{x}) + \left(1 - \sqrt{\mu_{\text{off-shell}}^{\text{EW}}} \right) \nu_{\text{B}}^{\text{EW}} \frac{p_{\text{B}}^{\text{EW}}}{p_{\text{ref}}}(\mathbf{x}) \\
&\left. + f(\theta_{qq}) \nu_{q\bar{q}ZZ} \frac{p_{q\bar{q}ZZ}}{p_{\text{ref}}}(\mathbf{x}) + \nu_{VVV} \frac{p_{VVV}}{p_{\text{ref}}}(\mathbf{x}) \right], \tag{4.26}
\end{aligned}$$

where $f(\theta_{qq})$ can be inferred from Eq. 4.25 and where a single density ratio is trained for $p_{q\bar{q}ZZ}$ to model all three $q\bar{q}ZZ$ processes (with $n_{\text{jets}} = 0, 1$, and ≥ 2). This is possible without loss of information because the observable that distinguishes the three processes (n_{jets}) is part of the vector $\mathbf{x}$ that describes the event, and therefore:

$$\nu_{q\bar{q}ZZ}^{n_{\text{jets}}=i} \frac{p_{q\bar{q}ZZ}^{n_{\text{jets}}=i}(\mathbf{x})}{p_{\text{ref}}(\mathbf{x})} = \delta_{n_{\text{jets}}, i} \nu_{q\bar{q}ZZ} \frac{p_{q\bar{q}ZZ}(\mathbf{x})}{p_{\text{ref}}(\mathbf{x})}, \tag{4.27}$$

where $\delta_{n_{\text{jets}}, i}$ is the Kronecker delta and the corresponding θ_{qq} dependence $f(\theta_{qq})$ can be introduced here using the formulae in Tab. 4.1. Using the model of Eq. 4.26, the task of estimating a fully μ -parameterized density ratio $p(\mathbf{x}|\mu_{\text{off-shell}}^{\text{ggF}}, \mu_{\text{off-shell}}^{\text{EW}})/p_{\text{ref}}(\mathbf{x})$ can be split into estimating eight parameter μ -independent density ratios $p_c(\mathbf{x})/p_{\text{ref}}(\mathbf{x})$ in Eq. 4.26.

The choice of p_{ref} is important as one needs to remove phase space regions $\mathbf{x}$ with $p_{\text{ref}}(\mathbf{x}) \sim 0$ in order for the density ratios p_c/p_{ref} in Eq. 4.26 to be numerically stable. Choosing p_{ref} to be a signal-rich probability density following the search-oriented mixture model procedure defined in Sec. 2.5.1:

$$\begin{aligned}
p_{\text{ref}}(\mathbf{x}) &= \frac{1}{\nu_{\text{S}}^{\text{ggF}} + \nu_{\text{SBI}_{10}}^{\text{EW}}} \left[\frac{d\nu_{\text{S}}^{\text{ggF}}}{dx}(\mathbf{x}) + \frac{d\nu_{\text{SBI}_{10}}^{\text{EW}}}{dx}(\mathbf{x}) \right] \\
&= \frac{1}{\nu_{\text{S}}^{\text{ggF}} + \nu_{\text{SBI}_{10}}^{\text{EW}}} \left[\nu_{\text{S}}^{\text{ggF}} \cdot p_{\text{S}}^{\text{ggF}}(\mathbf{x}) + \nu_{\text{SBI}_{10}}^{\text{EW}} \cdot p_{\text{SBI}_{10}}^{\text{EW}}(\mathbf{x}) \right] \tag{4.28}
\end{aligned}$$

The Monte Carlo sample corresponding to the chosen $p_{\text{ref}}(\mathbf{x})$ hypothesis is a combination of $p_{\text{S}}^{\text{ggF}}$ and $p_{\text{SBI}_{10}}^{\text{EW}}$ samples randomly shuffled together, with their combined weights normalized. As discussed above, for the EW production of the off-shell Higgs boson, it is impossible to formally define a signal sample, and hence the reference sample chooses the signal-rich hypothesis $p_{\text{SBI}_{10}}^{\text{EW}}$. The choice of

p_{ref} is motivated by the observation that, in any well-designed search, the most valuable events will be in the support of the signal process, but not necessarily in the support of background processes.

The key advantage in the choice of p_{ref} in Eq. 4.28 is that only events which do not fall within the support of the signal-rich $p_{\text{ref}}(\mathbf{x})$ hypothesis need to be removed. The corresponding pre-selection of signal-rich events will be discussed in the section 5.4.

Note that Eq. 4.26 is analogous to Eq. 2.23, both featuring a ratio of probability densities where the denominator is independent of any of the fitted parameters. During a profile likelihood ratio fit, the common denominator p_{ref} cancels out giving the usual profile log-likelihood ratio test statistic 2.24.

4.5 Previous measurement

The ATLAS experiment found the first evidence for the off-shell Higgs boson in 2022 [3] using the same Run-2 data used in this analysis, but using a more traditional analysis technique. The previous analysis made use of a *fixed* summary statistic built using a multi-classifying NN to discriminate the signal from the two types of background:

$$O_{NN}(\mathbf{x}) = \log \frac{p_S^{\text{ggF}}(\mathbf{x})}{p_B^{\text{ggF}}(\mathbf{x}) + 0.1 \cdot p_{q\bar{q}ZZ}} \quad (4.29)$$

where the p_c in this equation corresponds to the outputs of a softmax activation function used in the output layer of the trained NN discriminator. This is the observable used in two of the signal regions rich in ggF production, as defined in Ref. [78], and an analogous NN was trained for the EW signal region. The discriminant distribution for the various signals and backgrounds in the off-shell Higgs, for the 0 jet ggF signal region is shown in Fig. 4.12.

As discussed in Sec. 2.3.2, the summary statistic of Eq. 4.29 is only locally optimal around the $\mu_{\text{off-shell}} = 1$ values, due to the non-linear probability model of the off-shell Higgs boson. The summary statistic of Eq. 4.29 is independent of the POI $\mu_{\text{off-shell}}$ and is only trained to discriminate signal hypothesis from the background hypothesis using MC simulated at $\mu_{\text{off-shell}} = 1$. This makes the summary statistic sub-optimal for parameter regions with a large effect from quantum interference, which corresponds to the region with $\sqrt{\mu_{\text{off-shell}}} p_I \gg \mu_{\text{off-shell}} p_S$, which is the region with $\mu_{\text{off-shell}} < 1$ since $p_I \approx p_S$. This will become apparent in the comparisons of the sensitivity between this old analysis and the new NSBI analysis presented in Chap. 7.

In the NSBI analysis, the probability model of Eq. 4.26 aims to model not just the signal and background effects but also the interference effects $p_I(\mathbf{x})(\sim p_{\text{SBI}}(\mathbf{x}) - p_B(\mathbf{x}) - p_S(\mathbf{x}))/p_{\text{ref}}(\mathbf{x})$ using dedicated NNs, unlike the simple signal versus background observable used in the previous analysis and defined in Eq. 4.29.

The NSBI model of Eq. 4.26 ensures that an optimal summary statistic is used, based on the Neyman Pearson lemma, across the parameter range being scanned over. This combined with the effect from un-binned nature of the technique results in optimal sensitivity to $\mu_{\text{off-shell}}$ even in the interference-rich parameter regions $\sqrt{\mu_{\text{off-shell}}} p_I \gg \mu_{\text{off-shell}} p_S$. This will be explicitly shown using binned optimal observable technique in Chap 8.

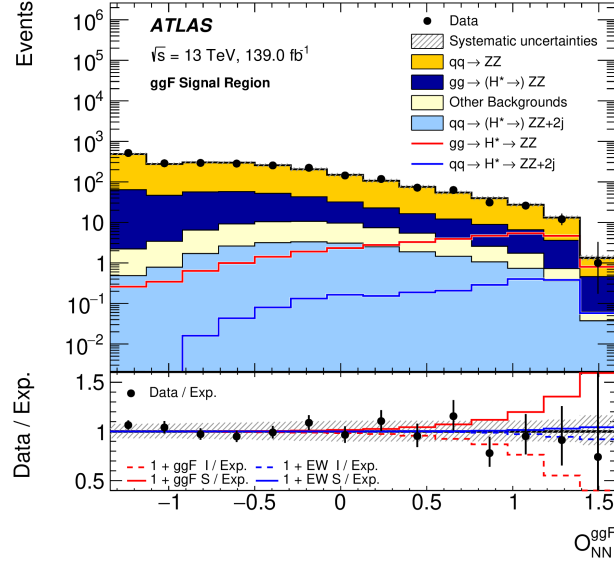


Figure 4.12: The NN discriminant used as a summary statistic in the first Run-2 analysis of the off-shell Higgs boson by the ATLAS experiment. The parameter-independent nature of this observable makes it sub-optimal for parameter values that favor a dominant interference effect over signal, motivating the need for a new optimal analysis technique. Figure from Ref. [3].

4.6 Systematic Uncertainties in the Off-shell Higgs boson model

Systematic uncertainties include experimental uncertainties in the reconstructed objects and modeling uncertainties in the simulated samples. The systematic uncertainty model follows closely that reported in Ref. [3] and only a summary is presented here. Modeling uncertainties and uncertainties in the jet energy scale and resolution dominate the systematic uncertainty of the measurement. Jet-related observables used to describe the events provide important information for discrimination between hypotheses, especially for EW off-shell Higgs boson production. However, they also make the result sensitive to modeling of jet energy scale and resolution.

The uncertainty in the integrated luminosity for the full Run 2 dataset is 0.83% [79], obtained using the LUCID-2 detector [80] for the primary luminosity measurements. Experimental uncertainties include momentum scale and resolution uncertainties for muons, electrons, and jets. These uncertainties are estimated by using calibrations performed for each individual object and by comparing them with different simulation models. Uncertainties in the reconstruction, identification and trigger efficiency of electrons [81, 82] and muons [83, 84] are determined from tag-and-probe efficiency measurements using $Z \rightarrow \ell^+ \ell^-$ and $J/\psi \rightarrow \ell^+ \ell^-$ events. uncertainties in the jet energy scale are derived by combining information from test-beam data, LHC collision data and simulation [26]. Uncertainties in the jet energy resolution are estimated as a function of jet p_T and rapidity using dijet events, based on a similar method as in Ref. [26].

Modeling uncertainties arise from the choice of PDF, missing higher-order corrections in both QCD and EW perturbative calculations, the merging of additional partons to the hard-scatter ME, and the description of the parton shower. The methods used to estimate modeling uncertainties are summarized in Tab. 4.2.

Table 4.2: Description of the different sources of modeling systematic uncertainties considered for each process and the method used to estimate it.

Process	Uncertainty	Method
$q\bar{q} \rightarrow ZZ$	Missing higher-order QCD	Renormalization and factorization scales
$q\bar{q} \rightarrow ZZ$	Soft-gluon resummation	QSF resummation scale
$q\bar{q} \rightarrow ZZ$	Jet merging	CKKW merging scale
$q\bar{q} \rightarrow ZZ$	Parton shower	SHERPA showering scheme
$q\bar{q} \rightarrow ZZ$	Missing higher-order EW	NLO/LO K -factor in regions of high recoil
$q\bar{q} \rightarrow ZZ$	PDF	NNPDF MC replicas
EW $qq \rightarrow ZZ + 2j$	Missing higher-order QCD	Renormalization and factorization scales
EW $qq \rightarrow ZZ + 2j$	Parton Shower	A14 tune parameters, ISR and FSR scales
EW $qq \rightarrow ZZ + 2j$	PDF	NNPDF MC replicas
$gg \rightarrow ZZ$	Missing higher-order QCD	Renormalization and factorization scales
$gg \rightarrow ZZ$	Soft-gluon resummation	QSF resummation scale
$gg \rightarrow ZZ$	Jet merging	CKKW merging scale
$gg \rightarrow ZZ$	Parton shower	SHERPA showering scheme
$gg \rightarrow ZZ$	PDF	NNPDF MC replicas

The PDF uncertainties are evaluated using the NNPDF prescription with 100 replicas from the NNPDF3.0 set [85]. The uncertainties due to missing higher-order QCD corrections are estimated by varying the renormalization and factorization scales independently, by factors one-half and two.

For the $gg \rightarrow ZZ \rightarrow 4\ell$ processes, the renormalization and factorization scales variations are evaluated on the NLO/LO K -factors and propagated to the measurement as a function of m_{ZZ} only [86]. The $gg \rightarrow ZZ \rightarrow 4\ell$ NLO scale variation is found to be approximately uniform throughout the phase space, with a 10% relative magnitude, and independent of the process (ggF S, ggF B, ggF SBI). The NLO K -factors and scale variations used for the $gg \rightarrow ZZ \rightarrow 4\ell$ processes do not contain the complete top-quark mass dependency, which has only been recently calculated [87]. Therefore, the $gg \rightarrow ZZ \rightarrow 4\ell$ uncertainties due to scale variations are increased by 50% for m_{ZZ} in the $t\bar{t}$ threshold region, and doubled in the phase space containing a jet with $p_T > 150$ GeV [3].

For all other processes, the two variations of the renormalization and factorization scales with largest impact in the expected value ν_c were taken as representative of this uncertainty. An additional uncertainty in missing higher-order QCD corrections is estimated for the $gg \rightarrow ZZ \rightarrow 4\ell$ and $q\bar{q} \rightarrow ZZ \rightarrow 4\ell$ processes by varying the soft-gluon resummation scale (QSF) in SHERPA by half and twice the nominal value used in the simulation. The QSF scale variation is evaluated separately for each $gg \rightarrow ZZ \rightarrow 4\ell$ component (ggF S, ggF B, and ggF SBI) and can induce variations as large as 40% on the expected yield [88].

Jet merging uncertainties are evaluated by varying the matching scale (based on the Catani-Krauss-Kuhn-Webber CKKW prescription [89]) for the processes simulated with the SHERPA generator. Parton-shower uncertainties are evaluated by varying the SHERPA showering scheme [88]. For

those processes simulated with the PYTHIA shower program, the uncertainty is assessed by varying the PYTHIA configurations, such as the parameter values of the A14 tune, the multi-parton models, and the initial and final-state radiation scales.

The uncertainties due to missing higher-order EW corrections are considered for the main $q\bar{q} \rightarrow ZZ \rightarrow 4\ell$ process. The largest missing higher-order component comprises mixed NLO QCD + NLO EW corrections, which are relevant in regions of phase space with large NLO QCD corrections. Following the procedure from Ref. [90], the full magnitude of the NLO EW correction is taken as uncertainty for events in which the quadruplet has large recoil. The same prescription was used in histogram-based analysis and further discussion can be found in Ref. [3]. These uncertainties are subleading when compared to the dominant modeling uncertainties related to missing higher-order corrections in $gg \rightarrow ZZ$, soft-gluon resummation in $gg \rightarrow ZZ$, and jet matching in $q\bar{q} \rightarrow ZZ$.

For the theoretical modeling uncertainties of $gg \rightarrow ZZ \rightarrow 4\ell$ production, common NPs are introduced for the ggF S, ggF B, and ggF SBI processes. The same is done for the theoretical modeling uncertainties of EW $q\bar{q} \rightarrow ZZ + 2j \rightarrow 4\ell + 2j$ production, where common NPs are used for EW SBI₁, EW SBI₁₀, and EW B. In the description of theoretical uncertainties in $gg \rightarrow ZZ \rightarrow 4\ell$, the normalization and shape components are described with different NPs. Since separate normalization parameters are introduced for the $q\bar{q} \rightarrow ZZ \rightarrow 4\ell$ processes with $n_{\text{jets}} = 0, 1, \geq 2$, separate NPs are also used to describe the corresponding modeling of uncertainties for the three different processes.

4.6.1 The full model

Systematic uncertainties are introduced in the measurement model using nuisance parameters (NP) α_m , following the prescription in Chap. 3. The joint probability model for the off-shell Higgs boson is then given by:

$$\frac{p(\mathcal{D}|\boldsymbol{\mu}, \boldsymbol{\alpha})}{\prod_k^{\mathcal{N}_{\text{data}}} p_{\text{ref}}(\mathbf{x}_k)} = \text{Pois} \left(\mathcal{N}_{\text{data}} \left| \sum_c f_c(\boldsymbol{\mu}) \cdot \left[\prod_{m=1}^{N_{\text{syst}}} G_c(\alpha_m) \right] \nu_c \right| \right) \times \frac{1}{\sum_c f_c(\boldsymbol{\mu}) \cdot \left[\prod_{m=1}^{N_{\text{syst}}} G_c(\alpha_m) \right] \nu_c} \prod_k^{\mathcal{N}_{\text{data}}} \left[\sum_c f_c(\boldsymbol{\mu}) \cdot \left[\prod_{m=1}^{N_{\text{syst}}} G_c(\alpha_m) \right] \nu_c \cdot \left[\prod_{m=1}^{N_{\text{syst}}} g_c(\mathbf{x}|\alpha_m) \right] \frac{p_c(\mathbf{x}_k)}{p_{\text{ref}}} \right], \quad (4.30)$$

where $\boldsymbol{\mu} = (\mu_{\text{off-shell}}^{\text{ggF}}, \mu_{\text{off-shell}}^{\text{EW}})$ and the channels c correspond to the eight relevant channels in the analysis, fully elaborated in Eq. 4.26. The techniques for the estimation of the functions $G_c(\alpha_m)$ and $g_c(\mathbf{x}|\alpha_m)$ is detailed in Chap. 3, and will be discussed further in Chap. 6.

Chapter 5

Simulations and Event Selections

This chapter reviews the simulation model used in the analysis, followed by the physics object and event selections. The split into a signal region and a control region, necessary when using a search-oriented mixture model as described in the last section, will also be reviewed specifically for the new analysis.

The off-shell Higgs boson analysis described in the thesis uses ATLAS detector collision data collected during the Run-2 of the LHC, at a center-of-mass energy of $\sqrt{s} = 13$ TeV. The observed data corresponds to a total integrated luminosity of 140 fb^{-1} [79] after data-quality requirements [91]. For more details on the triggers used during this run and their corresponding efficiencies, see Refs. [81, 83, 92]. The following section presents the details of the MC simulations used in the analysis, followed by a section on analysis-specific selections.

5.1 Monte Carlo Simulations

The $gg \rightarrow ZZ \rightarrow 4\ell$ samples (ggF S, ggF B, and ggF SBI) are generated with SHERPA v2.2.2 [93] and OPENLOOPS [94, 95, 96] at LO accuracy in quantum chromodynamics (QCD), with up to one additional parton in the final state, using the NNPDF3.0 NNLO parton distribution function (PDF) set [85]. The merging with the parton shower was performed using the MEPS@NLO prescription [97, 88] and the SHERPA built-in algorithm was used for parton showering and hadronization. Next-to-leading-order (NLO) QCD corrections are included as a function of the invariant mass of the two Z bosons, m_{ZZ} , separately for the ggF B, S, and SBI processes [86]. Fully differential next-to-next-to-leading-order (NNLO) corrections to the $gg \rightarrow H \rightarrow ZZ$ signal process are known [98, 99, 100], but not for the interference and background components. A common, average NNLO/NLO correction of 1.2 is applied to the signal, interference, and background components of the $gg \rightarrow ZZ$ process. Inclusive next-to-next-to-next-to-leading-order (N³LO) corrections to the $gg \rightarrow H \rightarrow ZZ$ signal process are known [101], dominated by the on-shell Higgs boson contribution. Currently, N³LO corrections in the off-shell region and for the interference and background components are not available. An average, common N³LO/NNLO correction is extrapolated to the off-shell region and applied to the signal, interference, and background components of the $gg \rightarrow ZZ$ process [102].

Table 5.1: Matrix element generator, parton shower (PS), and higher-order corrections used to describe the different samples in the measurement. The notation $[Xj]$ indicates that up to X jets are included in the hard-scatter matrix element calculation.

Sample	ME generator	PS	Higher-order correction
$gg \rightarrow ZZ$	SHERPA 2.2.2 (LO [1j])	SHERPA 2.2.2	NLO QCD (m_{ZZ} dependent) [86] Approx. N3LO QCD (global) [102]
EW $q\bar{q} \rightarrow ZZ + 2j$	MG5_AMC@NLO 2.3.3	PYTHIA 8.244	-
$q\bar{q} \rightarrow ZZ$	SHERPA 2.2.2 (NLO [1j], LO [3j])	SHERPA 2.2.2	NLO EW (m_{ZZ} dependent) [107, 108]
WWZ, WZZ, ZZZ	SHERPA 2.2.2	SHERPA 2.2.2	-
$t\bar{t}Z$	MG5_AMC@NLO 2.3.3	PYTHIA 8.210	NLO QCD + NLO EW (global) [109]

The EW $q\bar{q} \rightarrow ZZ + 2j \rightarrow 4\ell + 2j$ samples (EW B, EW SBI₁, and EW SBI₁₀) are generated with MADGRAPH5_AMC@NLO [103] at LO QCD and LO EW accuracy using the NNPDF3.0 NLO PDF set [85]. The PYTHIA 8.2 [104] program was used for parton showering and hadronization with the A14 set of tuned parameters (A14 tune) for the underlying event [105] and NNPDF2.3 LO PDF set [106].

The $q\bar{q} \rightarrow ZZ$ sample is generated with SHERPA v2.2.2 and OPENLOOPS using the NNPDF3.0 NNLO PDF set. The matrix elements (ME) are calculated to NLO accuracy in QCD for 0- and 1-jet final states, and to LO accuracy for 2- and 3-jet final states. The merging with the SHERPA parton shower was performed using the MEPS@NLO prescription. The NLO EW corrections are included as a function of m_{ZZ} [107, 108].

The triboson samples ZZZ , WZZ , and WWZ with fully leptonic decays were modeled with SHERPA v2.2.2 at NLO QCD accuracy. The $ZZZ \rightarrow 4\ell + 2j$ process is included in the EW $q\bar{q} \rightarrow ZZ + 2j$ sample described above. The simulation of $t\bar{t}V$ production with at least one of the top quarks decaying leptonically and the vector boson decaying inclusively into either quarks or leptons was performed with MADGRAPH5_AMC@NLO interfaced to PYTHIA 8.2 for parton showering and hadronization with the A14 tune. The total cross-sections for the $t\bar{t}V$ backgrounds were normalized to the NLO QCD and EW predictions from Ref. [109].

Tab. 5.1 summarizes the order in perturbation theory with which each simulated sample was generated and the source of higher-order correction (K -factors) used to improve the modeling. All simulated samples are processed with the ATLAS detector simulation [110] based on GEANT4 [31]. The effects of multiple inelastic interactions in the same and neighboring bunch crossings (pileup) were modeled by overlaying each simulated hard-scattering event with inelastic pp events generated with PYTHIA 8.186 [111] using the NNPDF2.3 LO PDF set and the A3 tune [112]. Simulated events are reweighted to match the pileup conditions observed in the full Run 2 dataset. Simulated events are reconstructed with the same algorithms and analysis chain as the data.

5.2 Object Reconstruction and Event Selection

The measurement of the off-shell Higgs boson production is performed in the $H^* \rightarrow ZZ \rightarrow 4\ell$ decay channel, where ℓ is either an electron or a muon. The object and event selections aim to identify opposite-charged electron or muon pairs consistent with the decay of a Z boson. A quadruplet is formed from two pairs with a common production vertex and invariant mass above the ZZ threshold. The selection criteria are optimized to increase the acceptance of $H^* \rightarrow ZZ \rightarrow 4\ell$ events while maintaining negligible levels of non-prompt background events. The object and event reconstruction used in this analysis is exactly the same as in the previous result [3] and only a short summary is provided here.

5.2.1 Object Reconstruction

Muons are identified by tracks or segments reconstructed in the MS and matched to tracks reconstructed in the ID, with exceptions in areas where the MS lacks coverage. In the region $2.5 < |\eta| < 2.7$, muons can also be identified by tracks from the MS alone. In the central gap region ($|\eta| < 0.1$) of the MS, muons can be identified by a track from the ID associated with a compatible calorimeter energy deposit (calorimeter-tagged muons). Candidate muons are required to have $p_T > 5$ GeV and $|\eta| < 2.7$, except calorimeter-tagged muons for which the p_T threshold is raised to 15 GeV. Muons must satisfy the *loose* identification criterion [84] with at most one standalone or calorimeter-tagged muon allowed per Higgs boson candidate. Electrons are reconstructed from energy deposits in the electromagnetic calorimeter matched to a track in the ID. Candidate electrons must have $p_T > 7$ GeV and $|\eta| < 2.47$, and satisfy the *loose* identification criteria [82].

All electrons and muons used in both channels must be isolated, satisfying a *loose* isolation criteria [84, 82]. Furthermore, electrons (muons) are required to have associated tracks satisfying $|d_0/\sigma_{d_0}| < 5$ (3) and $|z_0 \sin \theta| < 0.5$ mm, where d_0 is the transverse impact parameter relative to the beam line, σ_{d_0} is its uncertainty, and z_0 is the z coordinate of the r - ϕ impact point, defined relative to the primary vertex. The event is rejected if the minimum angular separation between two leptons is $dR_{\ell\ell} < 0.1$, where $dR_{\ell\ell} = \sqrt{(\Delta\phi_{\ell\ell})^2 + (\Delta\eta_{\ell\ell})^2}$.

Jets are reconstructed from particle-flow objects [23] using the anti- k_t algorithm [24, 25] with radius parameter $R = 0.4$. The jet-energy scale is calibrated using simulation and further corrected with in situ methods [26]. Reconstructed jets are required to have $p_T > 30$ GeV and $|\eta| < 4.5$. A jet-vertex tagger [113] is applied to jets with $p_T < 60$ GeV and $|\eta| < 2.4$ to suppress jets that originate from pileup. In the forward region, for jets with $p_T < 50$ GeV and $2.5 < |\eta| < 4.5$, another tagger based on jet shapes and topological jet correlations [114] is used to suppress pileup jets.

5.2.2 Event Selection

The selection of candidate events follows that described in Ref [3]. Events with at least four leptons (electrons or muons) are used in this analysis. The p_T thresholds for the three leading leptons are 20, 15 and 10 GeV, respectively. The four-lepton invariant mass is required to be above the on-shell ZZ production threshold, $180 < m_{4\ell} < 2000$ GeV. Candidate lepton quadruplets are formed by selecting

two opposite-charge, same-flavor dilepton pairs in each event. In the $4e$ and 4μ channels, in which there are two possible pairings, the one that includes the lepton pair with mass closest to that of the Z boson mass is chosen. In each quadruplet, the lepton pair with mass closest to the Z boson mass, m_{Z_1} , is referred to as the leading pair and required to have $50 < m_{Z_1} < 106$ GeV. The sub-leading pair mass, m_{Z_2} , must satisfy $50 < m_{Z_2} < 115$ GeV when $m_{4\ell} > 190$ GeV. Due to the increased probability of one Z boson being off-shell at lower values of $m_{4\ell}$, the lower threshold for m_{Z_2} decreases linearly from 50 GeV at $m_{4\ell} = 190$ GeV to 45 GeV at $m_{4\ell} = 180$ GeV.

5.3 Final State Observables

Events are described in the analysis by 14 observables $\mathbf{x} = (m_{4\ell}, m_{Z_1}, \dots, \Delta\phi_{jj})$, summarized in Tab. 5.2. The 14 observables, corresponding to various mappings of the full four-momentum of final state particles, provide a complete description of the reconstructed final state phase space [115]. While in principle the density ratios $p_c(\mathbf{x})/p_{\text{ref}}(\mathbf{x})$ can be modeled using the four-momenta information of the various leptons and jets in the $ZZ \rightarrow 4\ell$ final state, using physics-motivated *high-level* observables as in Tab. 5.2 that capture all the relevant physics makes the training of low-bias and low-variance density ratio estimators was found to converge faster and with fewer MC training events.

The three-momentum of the fermion (anti-fermion) in the Z_1 decay is defined as $\mathbf{q}_{11}$ ($\mathbf{q}_{12}$). Similarly, the three-momentum of the fermion (anti-fermion) in the Z_2 decay is defined as $\mathbf{q}_{21}$ ($\mathbf{q}_{22}$). The three-momentum of Z_1 (Z_2) is defined as $\mathbf{q}_1$ ($\mathbf{q}_2$). All three-momenta are defined in the rest frame of the quadruplet. Jets are ordered in p_T and their momenta are defined in the laboratory reference frame.

The observables in Tab. 5.2 are the components of the vector $\mathbf{x}$ in Eq. 4.19. The observables m_{jj} , η_{jj} , and ϕ_{jj} related to the leading dijet system, *i.e.*, the two jets with highest p_T in the event, are only well-defined for events with at least two jets. For events with fewer jets, the value of these observables are chosen as the median of the corresponding distribution for events with at least two jets. The observable n_{jets} is used for classification of the non-interfering background in Eq. 4.25, where all events with more than two jets are described by $n_{\text{jets}} = 2$.

The normal vectors $\mathbf{n}_1$ and $\mathbf{n}_2$ to the Z_1 and Z_2 decay planes and the normal vector $\mathbf{n}_{\text{sc}}$ to the Higgs boson decay plane are defined as:

$$\mathbf{n}_1 = \frac{\mathbf{q}_{11} \times \mathbf{q}_{12}}{|\mathbf{q}_{11} \times \mathbf{q}_{12}|}, \quad \mathbf{n}_2 = \frac{\mathbf{q}_{21} \times \mathbf{q}_{22}}{|\mathbf{q}_{21} \times \mathbf{q}_{22}|}, \quad \mathbf{n}_{\text{sc}} = \frac{\mathbf{n}_z \times \mathbf{q}_1}{|\mathbf{n}_z \times \mathbf{q}_1|}, \quad (5.1)$$

where $\mathbf{n}_z$ is the unit vector in the z direction. These vectors are used to build angles that are sensitive to the spin and parity of the quadruplet. Their geometrical visualization [115] is given in Fig. 5.1, where the three-momenta are shown in the rest frame of each particle's parent for clarity. Figure 5.2 shows the comparison between observed and expected background distributions for two of the observables, $m_{4\ell}$ and m_{jj} , showing good agreement for two of the observables used in the analysis.

Table 5.2: Definition of the observables $\mathbf{x}$ used to describe an event. The observables are defined relative to the ATLAS coordinate system, and are various mappings of the full four-momenta of final state particles. The vectors $\mathbf{q}_1$, $\mathbf{q}_2$, $\mathbf{q}_{11}$, $\mathbf{q}_{21}$, $\mathbf{n}_1$, $\mathbf{n}_2$, and $\mathbf{n}_{sc}$ are defined in Sec. 5.3.

Variable	Definition
$m_{4\ell}$	quadruplet mass
m_{Z_1}	Z_1 mass
m_{Z_2}	Z_2 mass
$\cos \theta^*$	cosine of the Higgs boson decay angle $[\mathbf{q}_1 \cdot \mathbf{n}_z / \mathbf{q}_1]$
$\cos \theta_1$	cosine of the Z_1 decay angle $[-(\mathbf{q}_2) \cdot \mathbf{q}_{11} / (\mathbf{q}_2 \cdot \mathbf{q}_{11})]$
$\cos \theta_2$	cosine of the Z_2 decay angle $[-(\mathbf{q}_1) \cdot \mathbf{q}_{21} / (\mathbf{q}_1 \cdot \mathbf{q}_{21})]$
Φ_1	Z_1 decay plane angle $[\cos^{-1}(\mathbf{n}_1 \cdot \mathbf{n}_{sc}) (\mathbf{q}_1 \cdot (\mathbf{n}_1 \times \mathbf{n}_{sc}) / (\mathbf{q}_1 \cdot \mathbf{n}_1 \times \mathbf{n}_{sc}))]$
Φ	angle between Z_1, Z_2 decay planes $[\cos^{-1}(\mathbf{n}_1 \cdot \mathbf{n}_2) (\mathbf{q}_1 \cdot (\mathbf{n}_1 \times \mathbf{n}_2) / (\mathbf{q}_1 \cdot \mathbf{n}_1 \times \mathbf{n}_2))]$
$p_T^{4\ell}$	quadruplet transverse momentum
$y^{4\ell}$	quadruplet rapidity
n_{jets}	number of jets in the event
m_{jj}	leading dijet system mass
$\Delta \eta_{jj}$	leading dijet system pseudorapidity
$\Delta \phi_{jj}$	leading dijet system azimuthal angle difference

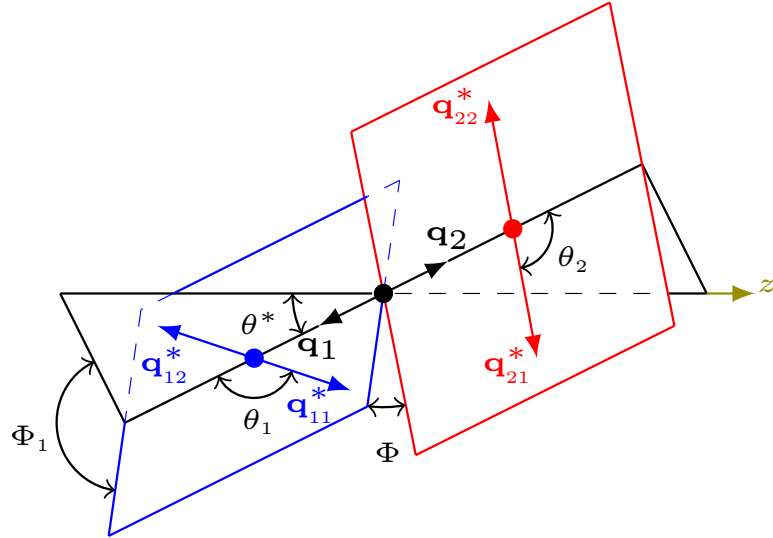


Figure 5.1: Representation of the angular observables used to describe the probability density ratio of each event. The three-momenta of the fermions ($\mathbf{q}_{11}^*$, $\mathbf{q}_{21}^*$) and anti-fermions ($\mathbf{q}_{12}^*$, $\mathbf{q}_{22}^*$) are shown in their parent rest-frames, and the three-momenta of the vector bosons ($\mathbf{q}_1$, $\mathbf{q}_2$) are shown in the quadruplet rest frame.

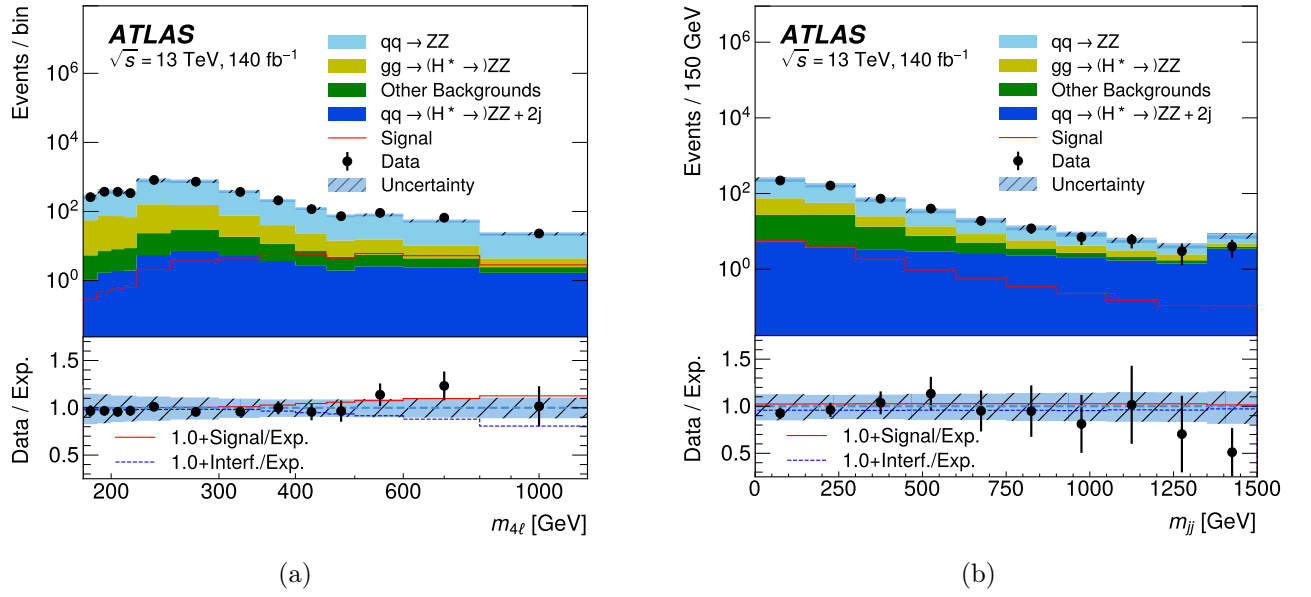


Figure 5.2: Comparison between observed and expected distribution of two observables used to describe events: a the four-lepton invariant mass and the b invariant mass of the two leading jets showing good agreement for two of the observables used in this analysis. The expected distributions for $q\bar{q} \rightarrow ZZ$, ggF SBI ($gg \rightarrow (H^* \rightarrow)ZZ$), EW SBI₁ ($qq \rightarrow (H^* \rightarrow)ZZ + 2j$) and other backgrounds are shown as stacked histograms, and the expected signal (interference) is shown as a red solid (blue dashed) line. The background is estimated under the SM hypothesis (post-fit, $\mu_{\text{off-shell}} = 1$). The lower panels show the ratio of data to expectation. The hashed band shows the total systematic uncertainty in the expected distribution. The last bin contains overflow events.

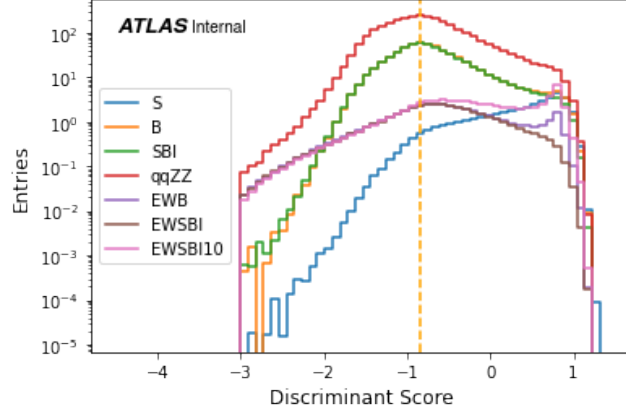


Figure 5.3: Distribution of the pre-selection NN score used to select events with $p_{\text{ref}}(\mathbf{x}) \gg 0$, defining the signal region (SR) and the control region (CR) in the analysis. The yellow dashed line shows the division between the CR, where a standard binned Poisson likelihood fit is done, and the SR where an unbinned likelihood fit using NSBI techniques is done. The discriminant score is the value of $D_{\text{pre}}(\mathbf{x})$ as defined in Eq. 5.3.

5.4 Signal and control regions

In Sec. 4.4, the choice of a reference hypothesis is made for the off-shell Higgs boson analysis based on the Search-Oriented Mixture Model formalism:

$$p_{\text{ref}}(\mathbf{x}) = \frac{1}{\nu_S^{\text{ggF}} + \nu_{\text{SBI}_{10}}^{\text{EW}}} \left[\nu_S^{\text{ggF}} \cdot p_S^{\text{ggF}}(\mathbf{x}) + \nu_{\text{SBI}_{10}}^{\text{EW}} \cdot p_{\text{SBI}_{10}}^{\text{EW}}(\mathbf{x}) \right] \quad (5.2)$$

The density ratios $p_c(\mathbf{x})/p_{\text{ref}}(\mathbf{x})$ in Eq. 4.26 can only reliably be estimated when $p_{\text{ref}}(\mathbf{x}) \gg 0$, which in the context of a Search-Oriented Mixture Model corresponds to a signal-rich phase space. The task then is to identify signal-poor phase space regions and exclude them from the NSBI analysis.

For this purpose, a multi-class classification NN is trained to split the events into signal and control regions using the observables defined in Tab. 5.2. This NN has five hidden layers each with 1 000 neurons and a *swish* activation function [116]. The output layer has five neurons with a *softmax* $[e^{-x_i} / \sum_j e^{-x_j}]$ activation function. The NN is trained with a multi-class cross-entropy loss corresponding to the five processes used in the training: ggF S, ggF B, VBF, EW B, and $q\bar{q} \rightarrow ZZ$. The vector boson fusion (VBF) process is obtained from a dedicated MC sample that uses only the EW VBF diagram in Fig. 4.10a. While this process does not provide a full description of the EW S production of off-shell Higgs bosons, it provides a sufficiently good approximation to define the control region (CR) with $p_{\text{ref}} \sim 0$ and signal region (SR) with $p_{\text{ref}} \gg 0$. The preselection discriminant $D_{\text{pre}}(\mathbf{x})$ is defined as:

$$D_{\text{pre}}(\mathbf{x}) = \log \frac{s_{\text{pre}}^{\text{ggF S}}(\mathbf{x}) + s_{\text{pre}}^{\text{VBF}}(\mathbf{x})}{s_{\text{pre}}^{\text{ggF B}}(\mathbf{x}) + s_{\text{pre}}^{\text{EW B}}(\mathbf{x}) + s_{\text{pre}}^{\text{qqZZ}}(\mathbf{x})}, \quad (5.3)$$

where $s(\mathbf{x})$ denotes the score function of the NN. By pre-selecting events $\mathbf{x}$ with large values of $D_{\text{pre}}(\mathbf{x})$, events with $p_{\text{ref}}(\mathbf{x}) \sim 0$ can be excluded. A distribution of the preselection score for various signals and backgrounds is shown in Fig. 5.3.

As is apparent in the distribution of Fig. 5.3, the number of signal events fall off almost linearly (in the log plot) below a certain threshold. In essence, the selection threshold (yellow line) is a fine-tuning knob that can be adjusted to minimize the trade-off between:

- Including as large a phase space as possible without the numerical stability of density ratio estimation p_c/p_{ref} breaking down.
- Ensuring that the number of simulated events is enough to train a NN that is able to accurately and precisely model the density ratios in the SR (see Chap. 6).

The power of a multi-dimensional unbinned analysis comes from extracting the likelihood ratio information using as many signal events as possible. Even the signal events in background-like phase space can contribute non-negligibly towards the final sensitivity in this type of analysis. However, training overly complex NN models that have low-bias across a large phase-space can introduce a large variance in the prediction if the statistical power of the simulated events is not enough. In order for the unbinned analysis to be successful, the pre-selection threshold of $D_{\text{pre}}(\mathbf{x})$ has to be optimized to allow the training of models that produce low-bias and low-variance estimates for the density ratios p_c/p_{ref} .

Using a careful trial-and-error approach then, different values of $D_{\text{pre}}(\mathbf{x})$ were used to split the events into SR and CR and NNs were trained to model accurate density ratios p_c/p_{ref} in each of the SR, which were subsequently tested using the diagnostics defined in Sec. 6.3. The SR with the optimal trade-off that allows both a large phase space without overfitting on the training data was found to satisfy $D_{\text{pre}}(\mathbf{x}) > -0.85$. Events failing this condition ($D_{\text{pre}}(\mathbf{x}) \leq -0.85$) define the CR in this analysis.

The CR is largely dominated by $q\bar{q} \rightarrow ZZ$ events and can thus be used to constrain the parameters $\theta_{q\bar{q}ZZ}^{0j,1j,2j}$ in Eq. 4.25, and also to validate the NN modelling using real data without *unblinding* the data in the signal-rich phase space. Fig. 5.4a shows a comparison between the observed and expected distribution of the observable $D_{\text{pre}}(\mathbf{x})$ used to define the SR and the CR indicating a good description of the acceptance times efficiency in each region. Fig. 5.4b depicts a comparison of the observed and expected n_{jets} distribution, showing a good description of the normalization obtained in each bin with the data-driven background normalization parameters.

In the CR, only the n_{jets} observable is used as the summary statistic to describe the probability model using the approximation of Eq. 2.16, while in the SR all 14 observables listed in Tab. 5.2 are used to create an NSBI model given by Eq. 2.23. The CR events were fitted using real data before the full unblinding phase of the analysis and also proved valuable for validating the trained density ratio NN outputs, as described in Chap. 6.

After the split of events that pass the pre-selection cuts, into a SR ($D_{\text{pre}} \geq 0.85$) and a CR

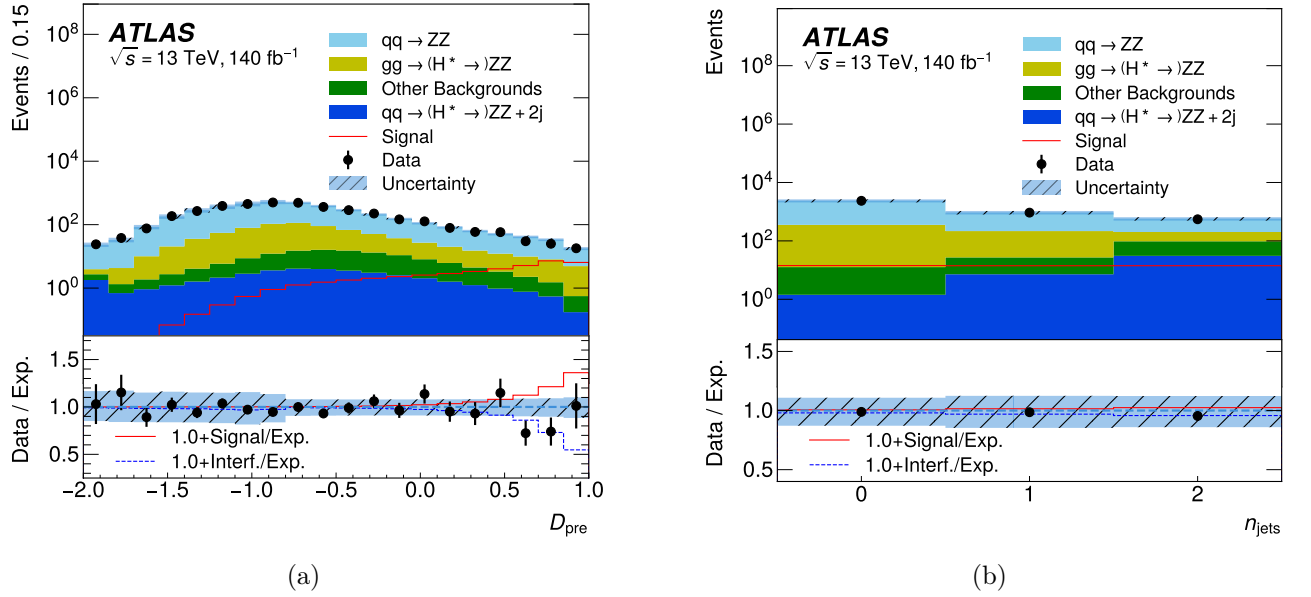


Figure 5.4: a Comparison between observed and expected distribution of the preselection discriminant $D_{\text{pre}}(\mathbf{x})$ showing a good description of the acceptance times efficiency in each region. b Comparison between observed and expected distribution of n_{jets} showing a good description of the normalization obtained in each bin with the data-driven background normalization parameters. The expected distributions for $q\bar{q} \rightarrow ZZ$, ggF SBI ($g\bar{g} \rightarrow (H^* \rightarrow)ZZ$), EW SBI₁ ($q\bar{q} \rightarrow (H^* \rightarrow)ZZ + 2j$) and other backgrounds are shown as stacked histograms, and the expected signal (interference) is shown as a solid red (dashed blue) line. The background is estimated under the SM hypothesis (post-fit, $\mu_{\text{off-shell}} = 1$). The lower panels show the ratio of data to expectation. The hashed band shows the total systematic uncertainty in the expected distribution. The last bin of panel b contains overflow events.

($D_{\text{pre}} < 0.85$), the full negative log-likelihood ratio model can be written as:

$$\begin{aligned}
 -2 \log \frac{L(\mu_{\text{off-shell}}, \theta_{qq}, \boldsymbol{\alpha})}{\prod_k^{\mathcal{N}_{\text{obs}}} p_{\text{ref}}(\mathbf{x})} = & \sum_{\text{bin} \in n_{\text{jets}}} \left[-2 \mathcal{N}_{\text{CR}}^{\text{bin}} \cdot \log \nu_{\text{CR}}^{\text{bin}}(\mu_{\text{off-shell}}, \theta_{qq}, \boldsymbol{\alpha}) + 2 \nu_{\text{CR}}^{\text{bin}}(\mu_{\text{off-shell}}, \theta_{qq}, \boldsymbol{\alpha}) \right] \\
 & - 2 \mathcal{N}_{\text{data}}^{\text{SR}} \cdot \log \nu_{\text{SR}}(\mu_{\text{off-shell}}, \theta_{qq}, \boldsymbol{\alpha}) + 2 \nu_{\text{SR}}(\mu_{\text{off-shell}}, \theta_{qq}, \boldsymbol{\alpha}) \\
 & - 2 \sum_i^{\mathcal{N}_{\text{data}}} \log \frac{p(\mathbf{x}_i | \mu_{\text{off-shell}}, \theta_{qq}, \boldsymbol{\alpha})}{p_{\text{ref}}(\mathbf{x})} + \sum_p^{N_{\text{syst}}} \alpha_p^2
 \end{aligned} \tag{5.4}$$

where the $\mathcal{N}_{\text{data}}^{\text{SR}}$ corresponds to the total events observed in the SR and $\mathcal{N}_{\text{CR}}^{\text{bin}}$ the number of events observed in each bin of the CR. Similarly $\nu_{\text{CR}}^{\text{bin}}$ corresponds to the expected yield in each bin of the CR, and ν_{SR} to the total expected yield in the SR. The final test statistic function for the full SR+CR fit is given by:

$$\begin{aligned}
 t_{\mu_{\text{off-shell}}} = & \sum_{\text{bin} \in n_{\text{jets}}} \left[-2 \mathcal{N}_{\text{CR}}^{\text{bin}} \left(\log \frac{\nu_{\text{CR}}^{\text{bin}}(\mu_{\text{off-shell}}, \hat{\theta}_{qq}, \hat{\boldsymbol{\alpha}})}{\nu_{\text{CR}}^{\text{bin}}(\hat{\mu}_{\text{off-shell}}, \hat{\theta}_{qq}, \hat{\boldsymbol{\alpha}})} \right) + 2 \left(\nu_{\text{CR}}^{\text{bin}}(\mu_{\text{off-shell}}, \hat{\theta}_{qq}, \hat{\boldsymbol{\alpha}}) - \nu_{\text{CR}}^{\text{bin}}(\hat{\mu}_{\text{off-shell}}, \hat{\theta}_{qq}, \hat{\boldsymbol{\alpha}}) \right) \right] \\
 & - 2 \mathcal{N}_{\text{data}}^{\text{SR}} \left(\log \frac{\nu_{\text{SR}}(\mu_{\text{off-shell}}, \hat{\theta}_{qq}, \hat{\boldsymbol{\alpha}})}{\nu_{\text{SR}}(\hat{\mu}_{\text{off-shell}}, \hat{\theta}_{qq}, \hat{\boldsymbol{\alpha}})} \right) + 2 \left(\nu_{\text{SR}}(\mu_{\text{off-shell}}, \hat{\theta}_{qq}, \hat{\boldsymbol{\alpha}}) - \nu_{\text{SR}}(\hat{\mu}_{\text{off-shell}}, \hat{\theta}_{qq}, \hat{\boldsymbol{\alpha}}) \right) \\
 & - 2 \sum_i^{\mathcal{N}_{\text{data}}} \log \frac{p(\mathbf{x}_i | \mu_{\text{off-shell}}, \hat{\theta}_{qq}, \hat{\boldsymbol{\alpha}}) / p_{\text{ref}}(\mathbf{x})}{p(\mathbf{x}_i | \hat{\mu}_{\text{off-shell}}, \hat{\theta}_{qq}, \hat{\boldsymbol{\alpha}}) / p_{\text{ref}}(\mathbf{x})} + \sum_p^{N_{\text{syst}}} (\hat{\alpha}^2 - \alpha^2)
 \end{aligned} \tag{5.5}$$

where the first sum over bins is the contribution from the CR summary statistic. Similar to this, the likelihood model from any binned or unbinned analysis can be combined with the NSBI model to fit a parameter of interest, by just adding on the negative log-likelihood terms from the model. The test statistic of Eq. 5.5 is used to perform fits on the off-shell Higgs boson parameters in Chap. 7.

Chapter 6

Neural Likelihood Ratio Estimation

This chapter provides a detailed description on the estimation of density ratios using large ensembles of NNs, followed by extensive diagnostics for testing the bias in the corresponding estimates. This is followed by detailed data-MC checks in the signal-free CR to test the validity of the simulation model and the NN mapping function. More data-MC checks are performed in to highlight the phase-space regions with significant data-MC differences. This is followed by the description of a novel method used to propagate the MC statistical uncertainty to the final parameter fitting.

In Sec. 4.4 the search-oriented mixture model for the off-shell Higgs boson measurement was presented. Referring to Eq. 4.26, the task of estimating a parametrized density ratio $p(\mathbf{x}|\mu_{\text{off-shell}})/p_{\text{ref}}(\mathbf{x})$ can be reduced to the task of estimating eight $\mu_{\text{off-shell}}$ -independent density ratios p_c/p_{ref} . Neural networks, widely perceived as ideal solutions for the estimation of probability density ratios [42, 117], can be trained for this purpose.

In order to train a NN for a given probability density ratio $p_c(\mathbf{x})/p_{\text{ref}}(\mathbf{x})$, simulated MC events are sampled from the respective hypotheses $\mathbf{x} \sim p_c(\mathbf{x}), p_{\text{ref}}(\mathbf{x})$ in the SR, where $p_{\text{ref}}(\mathbf{x}) > 0$ as defined in the Sec. 5.4. The NNs are trained as functions of the 14 observables in Tab. 5.2, selected to describe the full final-state information in the $H^* \rightarrow ZZ \rightarrow 4\ell$ channel. For each ratio, an ensemble of NNs is trained using gradient descent optimization methods to minimize an appropriate cost function. These trained neural networks can estimate all of the eight density ratios $h_\omega : \mathbb{R}^{14} \rightarrow \mathbb{R}$, $h_\omega(\mathbf{x}) = p_c/p_{\text{ref}}(\mathbf{x})$, where ω is the set of free parameters in the NN which can be tuned to estimate the appropriate density ratios. In the next section, the cost function and training techniques used for this work are reviewed, along with the diagnostics on the quality of the estimated density ratios.

6.1 Training of the Neural Networks

The NNs used are multi-layer perceptrons (MLP) with 5000 neurons arranged in five hidden layers of 1000 neurons each with swish [116] activation functions. The final layer is a single neuron with a sigmoid activation function. The swish function was used as the activation function for the hidden layers since, despite having the gradient squishing property, it still allows for the training of very large NN. **Despite the high computational cost of training very wide and deep NNs each with**

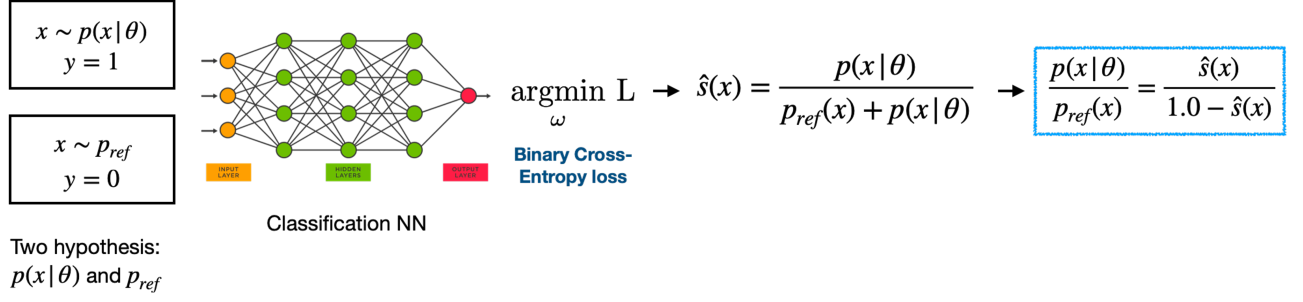


Figure 6.1: A representation of the *likelihood-ratio trick* or the CARL technique - the optimal minimum $\hat{s}(\mathbf{x})$ of the Binary Cross-Entropy loss function can have a probabilistic interpretations which can be used to regress the probability density ratio between two hypothesis. <Figure needs update>

over a million trainable parameters, it was found to be necessary as smaller NNs with fewer trainable parameters could not learn good quality density ratios that satisfied the diagnostics defined in section 6.3.

For the loss function, we use a *weighted* binary cross entropy loss, since the MC simulated events in a typical LHC analysis are weighted:

$$\mathcal{L}[s] = -\frac{1}{(\sum_i w_i)} \sum_i w_i \cdot [y_i \log \hat{s}(\mathbf{x}_i) + (1 - y_i) \log(1 - \hat{s}(\mathbf{x}_i))], \quad (6.1)$$

where $\hat{s}(\mathbf{x})$ is the output of the NN and w_i are the MC event weights. The optimization of weights is performed with the Nadam optimizer [118]. The y_i are defined as truth labels in the BCE loss function of Eq. 6.1. The value of $y_i = 1$ is chosen for events coming from the numerator hypothesis $\mathbf{x} \sim p_c(\mathbf{x})$ and $y_i = 0$ for events coming from the denominator hypothesis $\mathbf{x} \sim p_{ref}(\mathbf{x})$. The cost function is implemented in Keras [119] and TensorFlow [120], which are used for the training.

The optimized $\hat{s}(\mathbf{x}) = \hat{s}(\mathbf{x})_{pred}$ which minimizes the cost function:

$$\hat{s}_{pred}(\mathbf{x}) = \argmin_{\hat{s}(\mathbf{x})} \mathcal{L}[s] \quad (6.2)$$

corresponds to the relative probability of the underlying hypothesis, if the probability of the two hypotheses during training are normalized to the same value [42, 121]. For example, if the NN is trained with one of the samples p_c as the numerator hypothesis and the reference sample p_{ref} as the denominator hypothesis, setting the sum of weights of both samples to be equal, then assuming an *optimal* NN training:

$$\hat{s}_{pred}(\mathbf{x}) \rightarrow s(\mathbf{x}) = \frac{p_c(\mathbf{x})}{p_c(\mathbf{x}) + p_{ref}(\mathbf{x})} \quad (6.3)$$

The optimality of the NN training is verified in section 6.3. The prediction of Eq. 6.3 can then be used to build the probability density ratios (commonly referred to as the likelihood ratio trick):

$$\frac{p_c(\mathbf{x})}{p_{ref}(\mathbf{x})} = \frac{\hat{s}_{pred}(\mathbf{x})}{1.0 - \hat{s}_{pred}(\mathbf{x})} \quad (6.4)$$

This technique of estimating the probabilities is widely used, including in the ATLAS experiment for

$HH \rightarrow 4b$ background estimation [3], systematic uncertainty estimation in the previous off-shell Higgs boson publication [78] as well as in an unfolding setting using the Omnifold technique [122]. This is also referred to as the CARL technique, particularly when an additional calibration layer is applied to ensure the probabilistic behavior of the NN output as defined in Eq. 6.3.

The input features listed in 5.3 are standardized in a pre-processing step by removing the mean and scaling each feature to unit variance, with the exception of the integer and binary encoded observables. Overfitting tests are also done using the train-test split in Fig. 6.2.

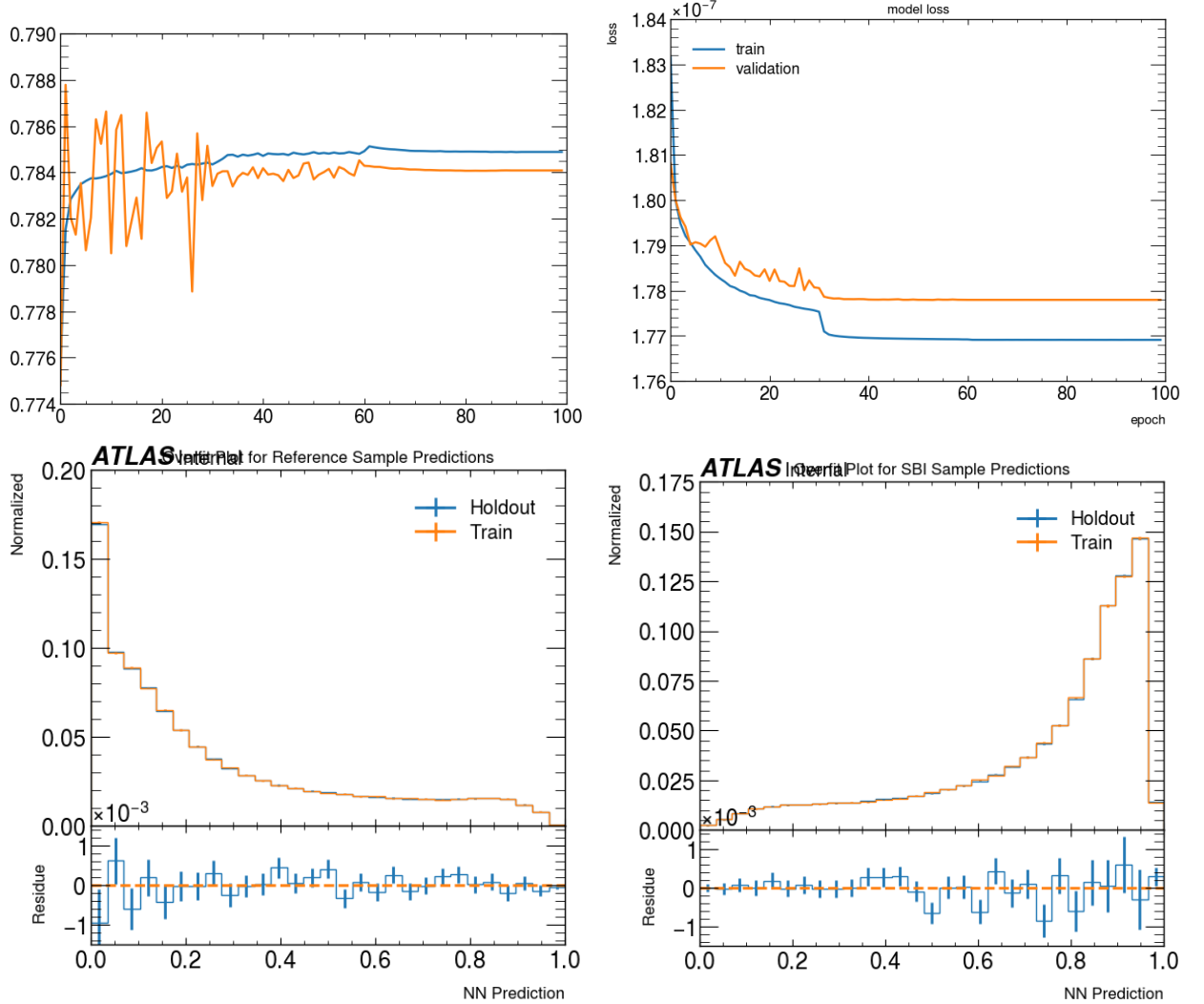


Figure 6.2: Four of the simpler diagnostic plots corresponding to a sample NN training. The top left plot shows the increase in accuracy per epoch of training, which eventually converges to a value and stays constant. Top right shows a similar evolution but for the value of binary cross-entropy loss over the epochs. The bottom two plots are used to test overfitting, by comparing the NN prediction distribution between training and testing datasets for each hypothesis.

In order to build an unbinned test statistics using the density ratios $p_c/p_{\text{ref}}(\mathbf{x})$, the density ratios predicted using the NN need to be accurate representations. In section 6.3, a comprehensive set of diagnostics will be developed to prove that the NN is unbiased and well-calibrated. Several metrics to

optimize the hyperparameters associated with the NN will also be discussed.

6.2 Training Ensembles with Cross-validation

Neural networks have an underlying variance associated with their learning algorithm. They may find slightly different sets of optimized weights ω during each instance of training. This can vary for a number of reasons. For example, how the training dataset is split in each batch or how the full dataset is split between training and validation. Even if the training dataset is kept exactly the same, there is still some variance coming from the random initialization of weights in the neurons before the training starts. This results in a variance associated with the output of the NNs, which can be a disadvantage in an analysis where the aim is to predict precise density ratios.

A commonly used approach to reduce the variance associated with the NNs is to train an *ensemble* of NNs and combine the predictions from these models. There are several different prescriptions for performing the combinations, ranging from taking the mean or median of the estimates, to training another NN which weighs the output of each of the NNs in the ensemble to give maximum accuracy. **Training an ensemble of NNs not only reduces the variance in the output but also improves the predictions compared to training a single model. For the kind of analysis described in this note, training large ensembles is necessary to get an unbiased estimator.** The ensemble predictions can be used in multiple ways, but here we use the mean of the ensemble as the final prediction for every event:

$$\frac{p_c}{p_{\text{ref}}}(\mathbf{x}) = \frac{1}{N_{\text{ens}}} \sum_a^{N_{\text{ens}}} \left[\frac{p_c}{p_{\text{ref}}}(\mathbf{x}) \right]_a \quad (6.5)$$

where $[p_c(\mathbf{x})/p_{\text{ref}}(\mathbf{x})]_a$ corresponds to the prediction from a *single* NN labelled a and N_{ens} corresponds to the total number of NNs used in the ensemble. In this analysis, we use $N_{\text{ens}} = 500$ for each channel c . All the instances of $p_c(\mathbf{x})/p_{\text{ref}}(\mathbf{x})$ from here on refer to the ensemble averaged prediction of Eq. 6.5.

6.2.1 Training Workflow

To train the ensemble of NNs without overfitting, a k-fold cross validation technique ¹ is used, where the input dataset is split into $k = 10$ subsets or folds and each NN in the ensemble is trained on 9 folds, also called the training dataset, and the trained NNs are used to evaluate the likelihood ratios on the 1 fold, also called the holdout dataset, not used in the training. This is then repeated until all the folds are used once for holdout. The density ratios used in the final fitting are the ones estimated on the holdout dataset of each fold. This avoids the NN to overfit on training data as the holdout data used in evaluation always remains statistically independent.

For training the ensemble of NNs, for each iteration of the k-fold technique, the holdout fold is kept fixed and from the remaining k-1 folds, **a random 80 – 20 train-test split is done per ensemble member**. The 80% is sampled randomly and without replacement from the pool of events in the k-1 training folds, and is then used directly to train a single NN ensemble member. The remaining 20%

¹Wikipedia article

is used to validate the single trained NN using the diagnostics defined in Sec. 6.3. A train-validation split of 90 – 10 is also performed during training time on the 80% used as the training data. The 10% validation data from this split is used to estimate the loss and accuracy metrics in each epoch of the training done on the remaining 90%, as shown in Fig. 6.2. The 80–20 split of the k -1 folds is randomized each time a new NN in the ensemble is trained **keeping the holdout dataset untouched**. Hence, each NN in the ensemble is trained on a different subset of the data in the k -1 folds, and the final prediction on the events in the *holdout* dataset is taken to be the mean of the predictions from the full ensemble as in Eq. 6.5.

The evaluated density ratios on the independent holdout dataset is also used for the diagnostic plots that will be presented in Sec. 6.3. A simplified representation of this procedure is shown in Fig. 6.3. The resulting spread of density ratio predictions due to this sampling procedure can also be used to estimate the impact of Monte Carlo statistical uncertainties as described in Sec. 6.5.

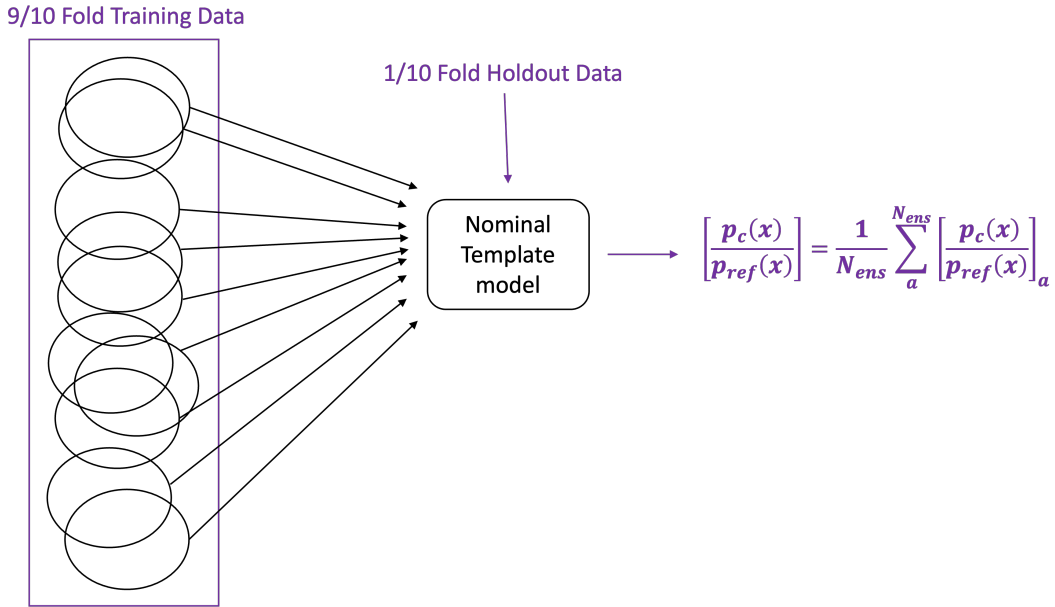


Figure 6.3: A cartoon representation of the ensembling technique used in the analysis along with k -fold cross-validation. Each blob in the box on the left represents the 80% of the training data sampled for each NN. The total number of sampled blobs, here only a few shown for representation, are equal to N_{ens} and corresponds to the total number of NNs trained using the train-test resampling technique (see text). The mean of the predictions from these N_{ens} NNs are used as final predictions in the profile likelihood fit. The ensemble training is done for each channel $c = \text{ggF SBI, ggF B, ggF S, etc.}$ independently.

The choice of $k = 10$ folds was motivated by practical reasons, with a balance between computational cost and having a large enough training dataset. The approach presented here was chosen over having just a single holdout dataset for inference to avoid any stochastic limitations during evaluation of the final test statistic and thus making the maximum use of the limited MC events.

6.3 Quality of the Density Ratios

As mentioned in Sec. 2.4.2, the fitting of parameters using the test statistic of Eq. 2.24 calculated with the NSBI approximated density ratios is not guaranteed to follow the familiar statistical properties of a binned Poisson likelihood fit. To ensure these properties, robust diagnostic checks need to be performed, which will be described in this section.

Calibration Curves

In order to perform direct statistical inference using NN predicted per-event density ratios, the most important check is to make sure that the NN predictions have a correct probabilistic interpretation. To verify this, the NN output probability ratios are compared to the *true* frequencies observed in MC events. This is done by plotting frequency histograms as a function of the predicted score $\hat{s}_{\text{pred}}$ defined in Eq. 6.3, to test the following equality:

$$\left[\frac{p_c(\mathbf{x})}{p_c(\mathbf{x}) + p_{\text{ref}}(\mathbf{x})} \right]_{\text{NN}} \sim \left[\frac{\mathcal{N}_c^{I(\mathbf{x}|\hat{s}_{\text{pred}})}}{\mathcal{N}_c^{I(\mathbf{x}|\hat{s}_{\text{pred}})} + \mathcal{N}_{\text{ref}}^{I(\mathbf{x}|\hat{s}_{\text{pred}})}} \right]_{\text{MC}} \quad (6.6)$$

where we bin the events from p_c and p_{ref} MC samples as a function of the NN predicted score $\hat{s}_{\text{pred}}$ of Eq. 6.3. The binned events are denoted by $\mathcal{N}_c^{I(\mathbf{x}|\hat{s}_{\text{pred}})}$ and $\mathcal{N}_{\text{ref}}^{I(\mathbf{x}|\hat{s}_{\text{pred}})}$ for the p_c and p_{ref} MC samples respectively, where $I(\mathbf{x}|\hat{s}_{\text{pred}})$ returns the index of the $\hat{s}_{\text{pred}}$ bin in which an event $\mathbf{x}$ falls. The $\mathcal{N}_c^{I(\mathbf{x}|\hat{s}_{\text{pred}})}$ histogram is divided by the $\mathcal{N}_c^{I(\mathbf{x}|\hat{s}_{\text{pred}})} + \mathcal{N}_{\text{ref}}^{I(\mathbf{x}|\hat{s}_{\text{pred}})}$ to calculate the relative probability or frequency of the hypothesis c and diagnose its agreement with the predicted score in the LHC of Eq. 6.6. This is shown in Fig. 6.4. If the NN density ratio estimates do indeed match the true MC event frequencies $\mathcal{N}_c^{I(\mathbf{x}|\hat{s}_{\text{pred}})} / (\mathcal{N}_c^{I(\mathbf{x}|\hat{s}_{\text{pred}})} + \mathcal{N}_{\text{ref}}^{I(\mathbf{x}|\hat{s}_{\text{pred}})})$ as obtained from a randomly selected sample of MC events not used in the training (the holdout sample defined in earlier sections), the two curves in Fig. 6.4 are expected to match. They show excellent agreement even for very rare events, as least within the precision of the simulated sample available.

The calibration plot presented in Fig. 6.4 is only used for validation in the method presented in this thesis, but in the case of imperfect agreements in out-of-box NN predictions can also be used as an additional calibration layer. This would imply getting the probability ratio values from the binned estimates in Fig. 6.4, as was done in Ref. [123] for an NSBI analysis to infer Dark Matter substructure. This corresponds to the so-called CARL (Calibrated Approximate Ratio of Likelihoods) method [42], where even if the NN estimates are not perfect approximations of the *true* (or more concretely, unbiased) density ratios, they can be used for final hypothesis test when using a calibration layer that ensures probabilistic interpretation of the final output. Such a technique would result in a sub-optimal Type-II error, but would still have robust statistical properties in the hypothesis test.

Fig. 6.4 show the calibration of the score function $\hat{s}_{\text{pred}}$. For a detailed diagnosis of very signal-rich and signal-deployed regions of phase space, the logarithms of the density ratios are compared to the MC predictions in Fig. 6.5 checking the following equality:

$$\left[\log \frac{p_c(\mathbf{x})}{p_{\text{ref}}(\mathbf{x})} \right]_{\text{NN}} \sim \left[\log \frac{\mathcal{N}_c^{I(\mathbf{x}|\hat{s}_{\text{pred}})}}{\mathcal{N}_{\text{ref}}^{I(\mathbf{x}|\hat{s}_{\text{pred}})}} \right]_{\text{MC}} \quad (6.7)$$

Given that the MC samples have a much larger effective luminosity (or statistics) compared to the observed data, this test provides a very robust validation of the accuracy of the NNs.

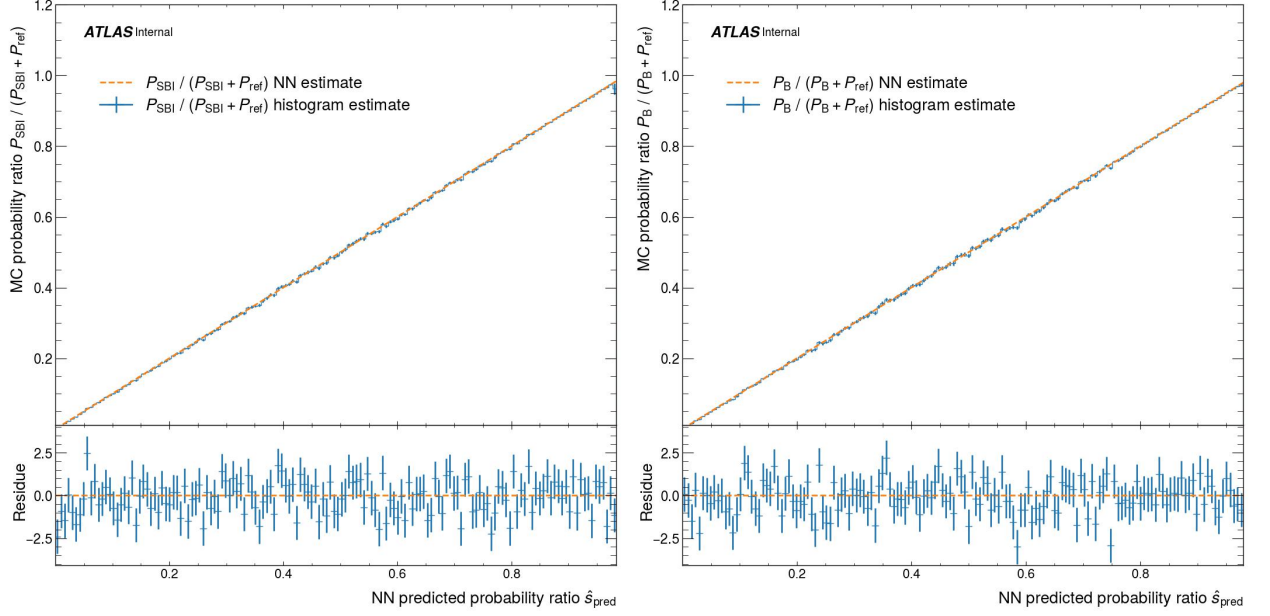


Figure 6.4: Diagnostic check defined in Eq. 6.6 of the relative probability predicted by the NNs with the relative probability using binned MC events. This plot uses the prediction from the ensemble of NNs (see Section 6.2). The binning is chosen as narrow as allowed by the size of simulated samples in order to highlight the well-calibrated predictions for very rare signal events on the far left ($\hat{s}_{\text{pred}} \approx 0$) and far right ($\hat{s}_{\text{pred}} \approx 1$).

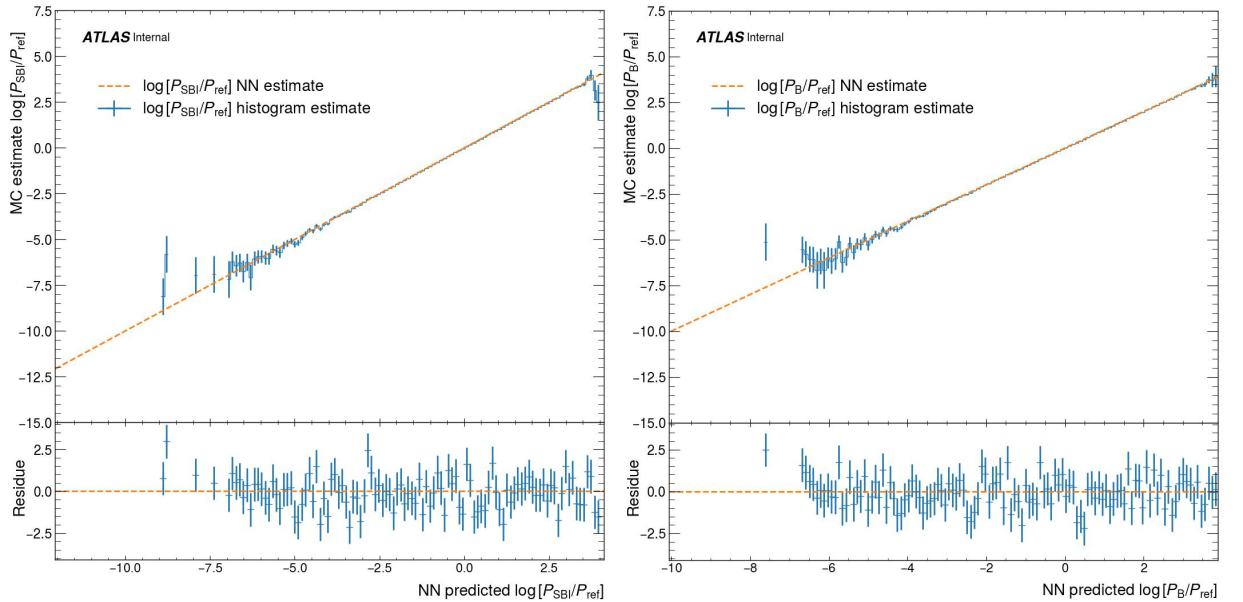


Figure 6.5: Diagnostic check defined in Eq. 6.7 of the log of probability density ratio predicted by the NNs with the log of density ratio using binned MC events. This plot uses the prediction from the ensemble of NNs (see Section 6.2). The binning is chosen as narrow as allowed by the size of simulated samples in order to highlight the well-calibrated predictions for very rare signal events.

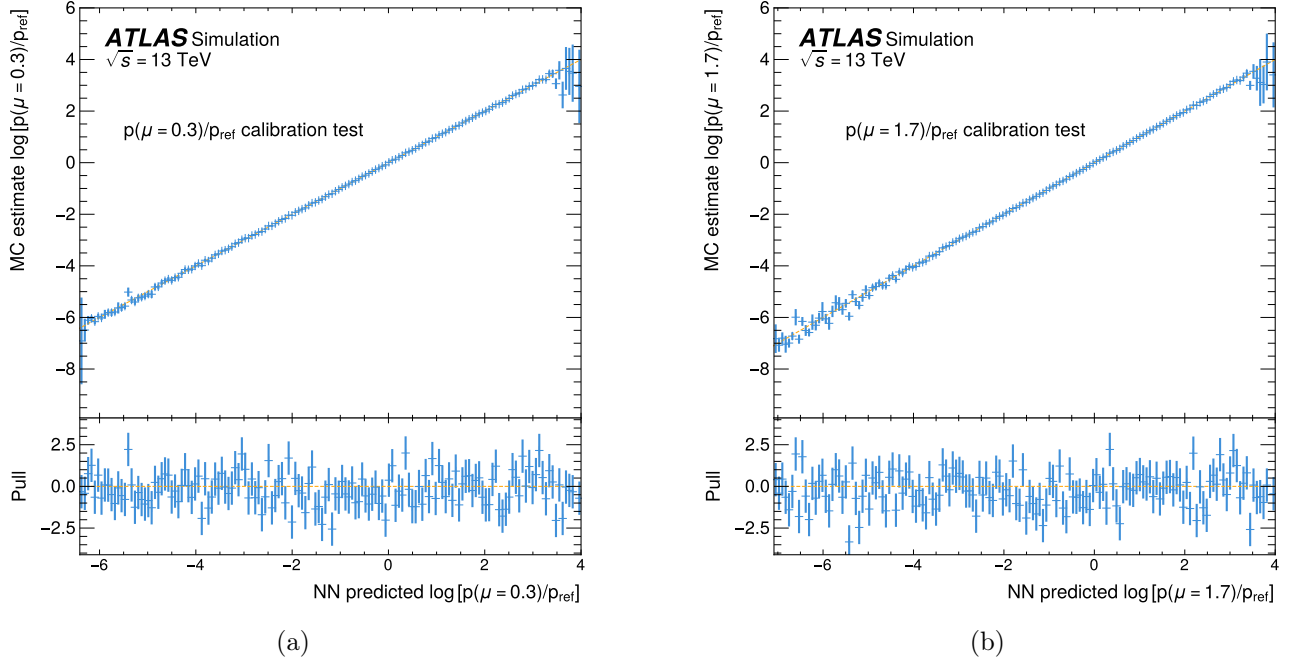


Figure 6.6: Comparison between the NN-based and histogram-based estimate of the density ratio $p(\mathbf{x}|\mu_{\text{off-shell}})/p_{\text{ref}}(\mathbf{x})$ for a $\mu_{\text{off-shell}} = 0.3$ and b $\mu_{\text{off-shell}} = 1.7$. The comparison is done as a function of $\log[p(\mathbf{x}|\mu_{\text{off-shell}})/p_{\text{ref}}(\mathbf{x})]$ to separate the comparison for events that are very signal-like and reference-like. The lower panels show the pull, defined as the difference between the NN and MC-based estimate divided by the statistical uncertainty due to the finite number of MC events. The error bars indicate the uncertainty due to the finite number of simulated events in the histogram-based estimate of the density ratio.

An additional strict test can be performed by injecting different signal values in the truth MC sample used for evaluation. The quality of the full likelihood ratio (Eq. 4.26) at different values of the parameter $\mu_{\text{off-shell}}$ can be tested using similar calibration plots. The calibration plots between the reference hypothesis $p_{\text{ref}}(\mathbf{x})$, and the $\mu = 0.3$ (interference-rich), and $\mu = 1.7$ (signal-rich) hypotheses are shown in Fig. 6.6. The μ -morphing defined in Eq. 4.20 is used to generate truth MC samples at the various values of μ from the available set of simulations.

Density Ratio Reweighting

In principle, if the NNs estimate the *exact* or unbiased density ratios $\hat{r}_c(\mathbf{x}) \approx (p_c/p_{\text{ref}})(\mathbf{x})$, then the following relation must hold:

$$p_c(\mathbf{x}) \approx \hat{r}_c(\mathbf{x}) \cdot p_{\text{ref}}(\mathbf{x}) \quad (6.8)$$

There are two ways to check this equality:

1. By training another NN as a classifier between the reweighted reference sample and the original sample and checking the AUC. In ideal case this should be ≈ 0.5 , i.e. the classifier would be unable to tell the two samples apart. The sample ROC curve and AUC for two of the density ratio checks are shown in Fig. 6.7. This provides a **multi-dimensional** verification of the quality

of the density ratio. The classifiers trained in Fig. 6.7 are deep NNs with the same structure and loss function as the ones used in the density ratio estimations. Moreover, they use the same input features $\mathbf{x}$ listed in Sec. 5.3. They use as inputs events from the test data of the p_c and p_{ref} hypotheses from the train-test split to perform the training.

2. By making 1D projections of the full observable space and checking if the reweighting works for individual observables. In Figs. 6.8 and 6.9 this is shown for two example kinematic distributions for three of the channels in the mixture model, but has been verified individually for all different inputs $\mathbf{x}$ and density ratios $p_c(\mathbf{x})/p_{\text{ref}}(\mathbf{x})$. While this test does not provide multi-dimensional tests of the the probability distributions, it provides a direct test in relevant part of the final-state phase space.

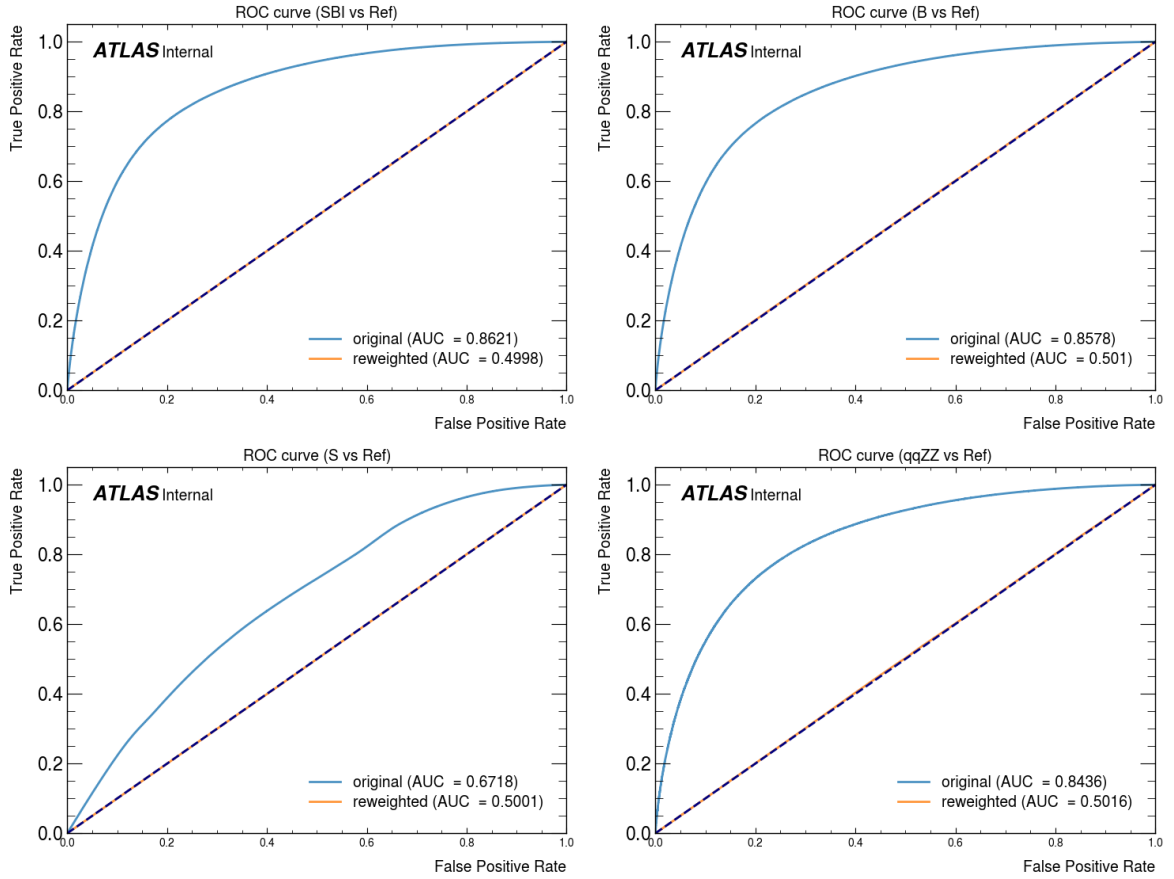


Figure 6.7: Reweighting multi-dimensional diagnostic: a NN classifier is trained to try and distinguish the original sample and the reference sample reweighted to the original using Eq. 6.8. The AUC value of 0.5 corresponds to an indistinguishable sample. This provides a multi-dimensional test of the quality of predicted density ratios. Note how training a classifier NN to distinguish reference sample from the original without reweighting results in a large AUC.

These diagnostics are used to test the quality of the NNs estimated ratios p_c/p_{ref} . The quality of the full likelihood ratio (see Eq. 4.26) at different values of the parameter $\mu_{\text{off-shell}}$ can be tested with a similar reweighting technique.

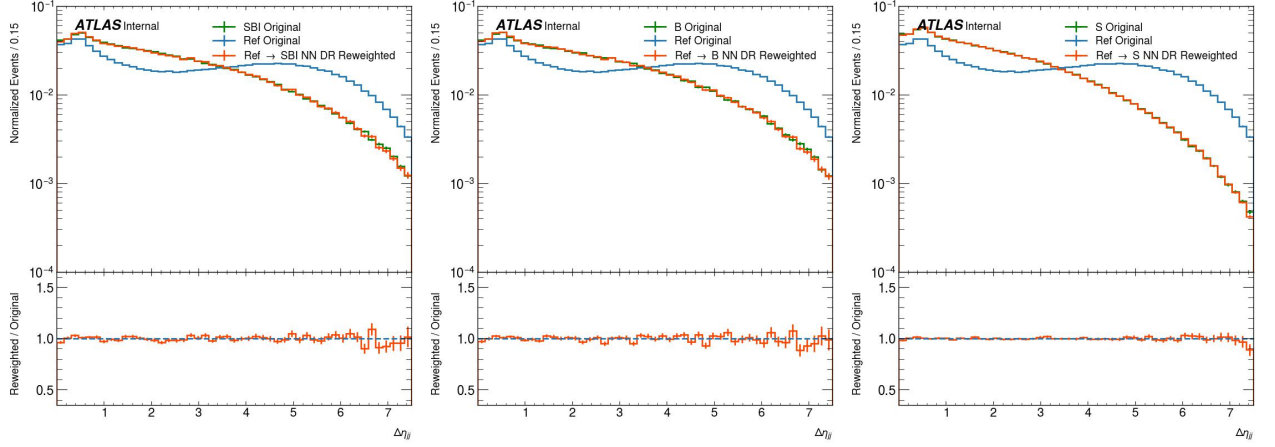


Figure 6.8: Reweighting diagnostics: The dijet observable $\Delta\eta_{jj}$. Testing the quality of the predicted density ratio by reweighting 1D observables with the ratios, using Eq. 6.8. The blue curve shows the reference sample and the orange one the sample after reweighting to match the target, shown in green: the ratio plots show the ratio of the reweighted distribution bin to the target distribution bin. All input variables are tested for all NN, showing excellent agreement.

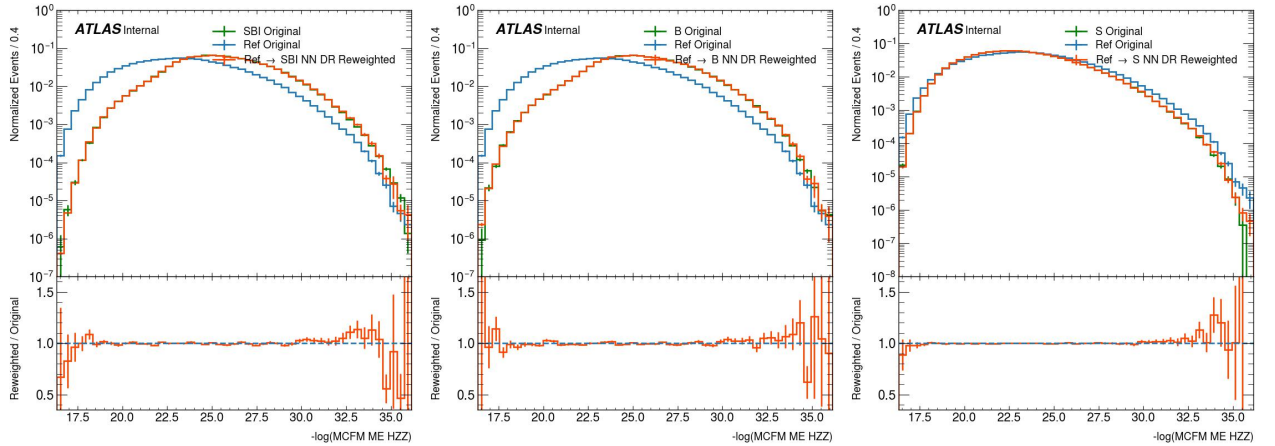


Figure 6.9: Reweighting diagnostics: High-level ME observable calculated using MCFM and not used in the training of the NNs. Testing the quality of the predicted density ratio by reweighting 1D observables with the ratios, using Eq. 6.8. The blue curve shows the reference sample and the orange one the sample after reweighting to match the target, shown in green: the ratio plots show the ratio of the reweighted distribution bin to the target distribution bin. All input variables are tested for all NN, showing excellent agreement.

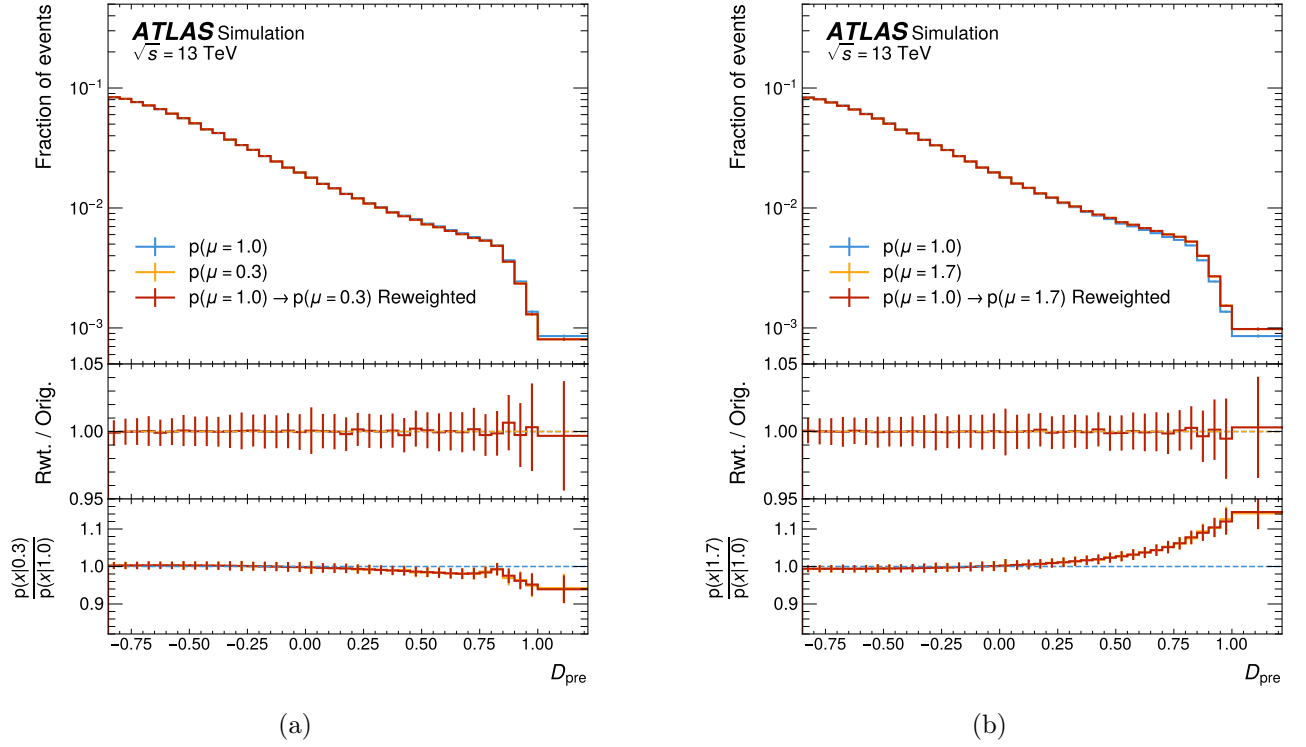


Figure 6.10: Comparison between the $D_{\text{pre}}(\mathbf{x})$ normalized distribution obtained using an Asimov sample with a $\mu_{\text{off-shell}} = 0.3$ and b $\mu_{\text{off-shell}} = 1.7$, and the same distributions obtained through reweighting of a SM ($\mu_{\text{off-shell}} = 1.0$) Asimov sample. The ratio plots provide a comparison between the two estimates (middle panel) and between the distributions at the given value of $\mu_{\text{off-shell}}$ and the SM expectation (bottom panel).

$$p(\mathbf{x}|\mu_{\text{off-shell}}) \approx \frac{p(\mathbf{x}|\mu_{\text{off-shell}})}{p_{\text{ref}}(\mathbf{x})} \cdot p_{\text{ref}}(\mathbf{x}) \quad (6.9)$$

Figs. 6.11 and 6.10 show the reweighting diagnostics of two observables for different $\mu_{\text{off-shell}}$ hypotheses. As before, the $\mu_{\text{off-shell}}$ -morphing defined in Eq. 4.20 is used to generate Asimov samples at the various values of $\mu_{\text{off-shell}}$ from the available set of simulations.

Data-MC Validations in the control region

In a typical ATLAS experiment, the real data is kept unblinded until the full analysis strategy has been formed and validated extensively. While the previous sections explored validations using MC simulations, this section highlights validations that can be performed using real data from a signal-deficient phase space, before using the framework on the full dataset. For the pre-unblinding validation of the MC modeling and the SR-CR split defined in Sec. 5.4 in the off-shell Higgs boson analysis, data events from the CR of the analysis $D_{\text{pre}} < -0.85$ are used. This is one of the primary motivations to not throw away events in the CR, but rather use them for validations and data-driven fits as discussed in Sec. 2.5.1.

The data-MC comparison for several kinematic distributions in the CR as well as the preselection NN score is shown in Fig. 6.13. **Post-fit distributions refer to the plots where the MC pre-**

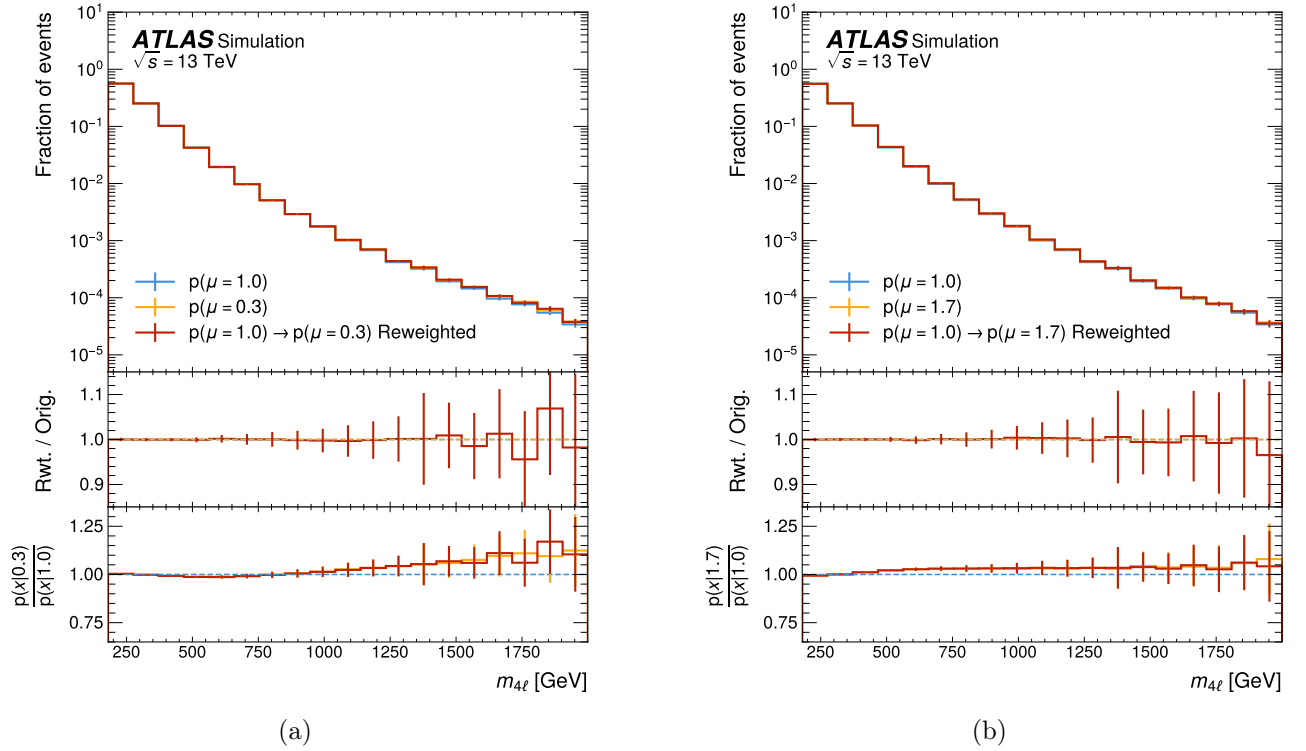


Figure 6.11: Comparison between the $D_{\text{pre}}(\mathbf{x})$ normalized distribution obtained using an Asimov sample with a $\mu_{\text{off-shell}} = 0.3$ and b $\mu_{\text{off-shell}} = 1.7$, and the same distributions obtained through reweighting of a SM ($\mu_{\text{off-shell}} = 1.0$) Asimov sample. The ratio plots provide a comparison between the two estimates (middle panel) and between the distributions at the given value of $\mu_{\text{off-shell}}$ and the SM expectation (bottom panel).

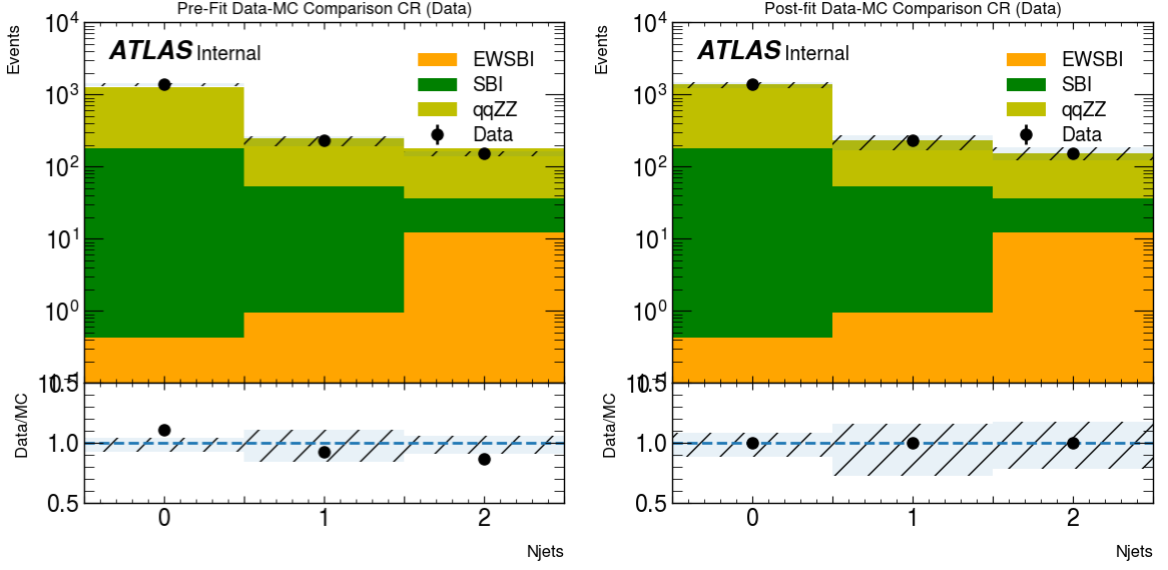


Figure 6.12: **Blinded Checks:** Pre-fit and post-fit data-MC comparisons of the n_{jets} summary statistic used to fit the CR using data. This fit has the data in the SR blinded. The larger uncertainty on the post-fit distribution is due to the introduction of unconstrained nuisance parameters defined in Eq. 4.25, for which there is no pre-fit analog.

dictions have been modified according to the data fitted NP values, while keeping the POI value fixed to the nominal value ($\mu_{\text{off-shell}} = 1$). Similarly, pre-fit refers to the MC predictions where NP values are set to nominal values. For the pre-unblinding post-fit distributions, a fit is performed using only the CR summary statistic of n_{jets} as shown in Fig. 6.12, with the test statistic:

$$t_{\mu_{\text{off-shell}}}^{\text{CR}} = \sum_{\text{bin} \in n_{\text{jets}}} \left[-2 \mathcal{N}_{\text{CR}}^{\text{bin}} \left(\log \frac{\nu_{\text{CR}}^{\text{bin}}(\mu_{\text{off-shell}}, \hat{\theta}_{qq}, \hat{\alpha})}{\nu_{\text{CR}}^{\text{bin}}(\hat{\mu}_{\text{off-shell}}, \hat{\theta}_{qq}, \hat{\alpha})} \right) + 2 \left(\nu_{\text{CR}}^{\text{bin}}(\mu_{\text{off-shell}}, \hat{\theta}_{qq}, \hat{\alpha}) - \nu_{\text{CR}}^{\text{bin}}(\hat{\mu}_{\text{off-shell}}, \hat{\theta}_{qq}, \hat{\alpha}) \right) \right] \quad (6.10)$$

The uncertainty bands in post-fit plots are also estimated using post-fit errors on the various NPs. The total uncertainty is calculated while preserving post-fit correlations between the various NPs, by using the covariance matrix of Eq. 3.15.

The probability density ratios p_c/p_{ref} can also be validated in the CR. Note that the resulting ratios are not estimates of the density ratio for these events, since the NN outputs are only optimized to be density ratio estimates in the SR phase space. Nevertheless, performing these checks in the CR before the unblinding of the full dataset helps validate the mapping function associated with the trained feed-forward NN, along with all the validations already performed using the input kinematic features $\mathbf{x}$ (See Fig. 6.13). The validations of the NN outputs in the CR are shown in Fig. 6.14, for a few of the NN output scores $\hat{s}_{\text{pred}}$.

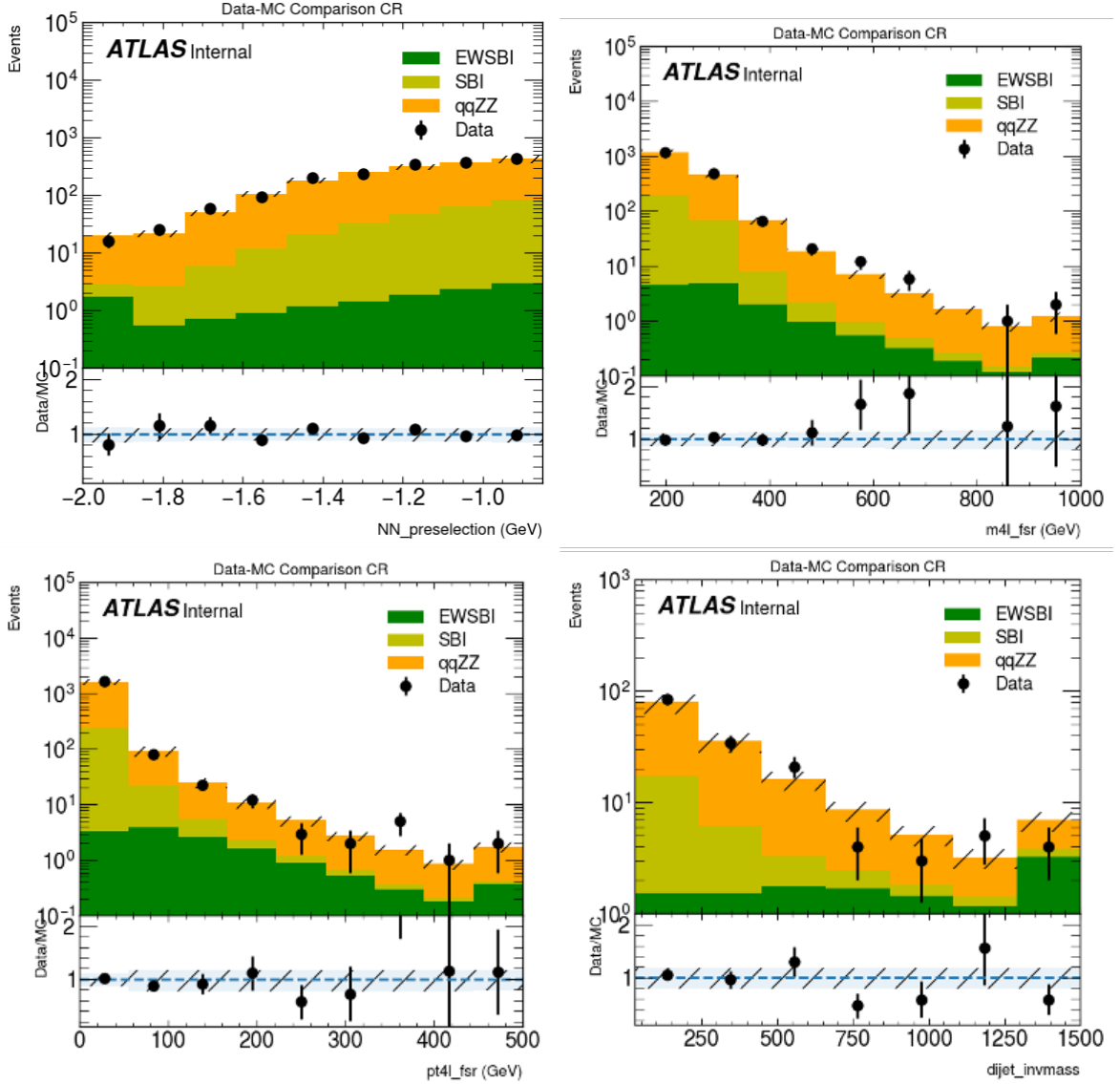


Figure 6.13: **Blinded Checks:** Post fit data-MC validation using data events exclusively from the $D_{pre} < -0.85$ control region. The plots show various kinematic distributions, as well as the preselection NN. The events in these plots are not used to estimate the per-event likelihood ratios.

Data-MC Validations in the full phase space

Post unblinding of the full dataset, data-MC validations can be performed in the full SR+CR phase space. The post-fit plots now refer to a fit done in the full phase space using the test statistic of Eq. 5.5. Just like in the blinded CR plots above, the NPs are fitted to their best fit value in the MC estimates shown in the plots. Histograms of the scores $s_c(x)$ used to estimate the density ratios in the SR are shown in Fig. 6.15 for all processes in this analysis indicating a good modeling of the probability density ratios throughout the SR.

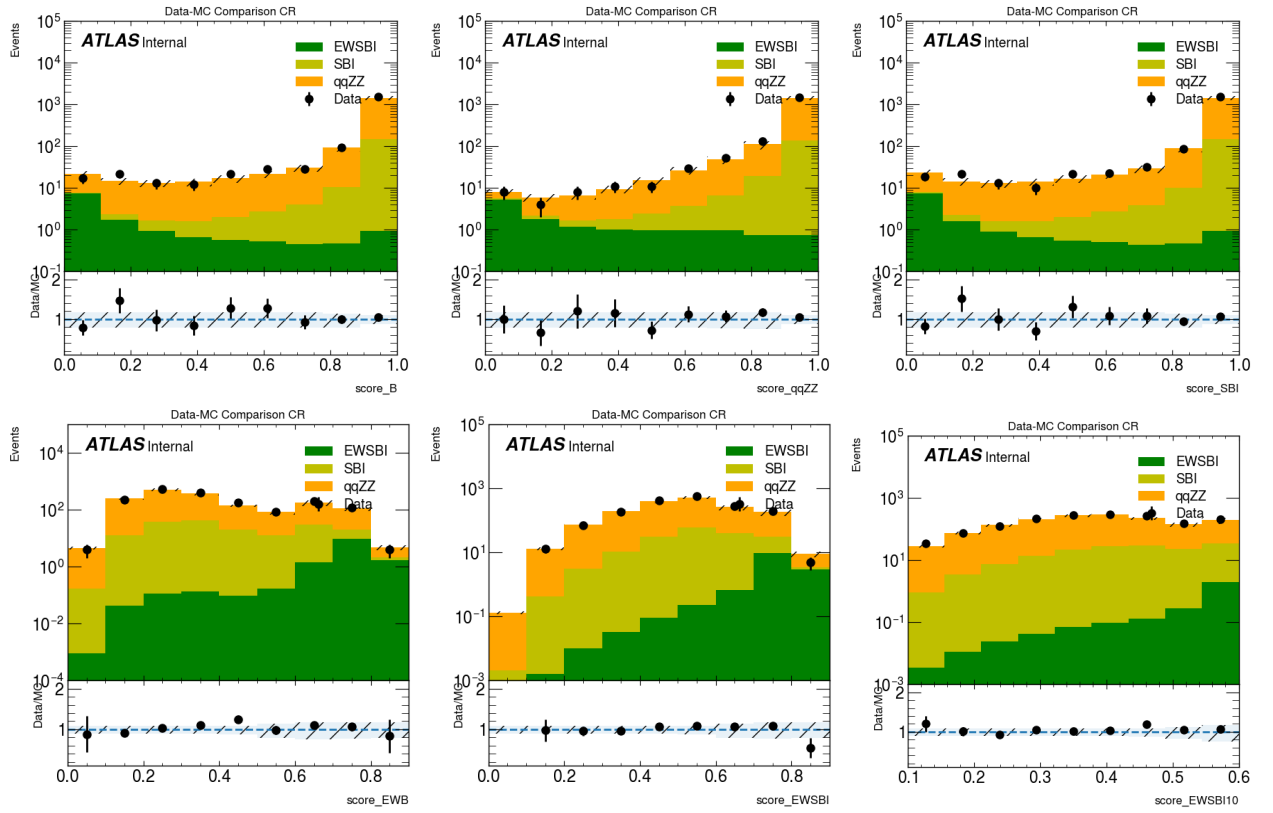


Figure 6.14: **Blinded Checks:** Post fit data-MC validation using estimated density ratio NN scores for data events exclusively from the $D_{pre} < -0.85$ control region. The events in these plots are not used to estimate the per-event likelihood ratios.

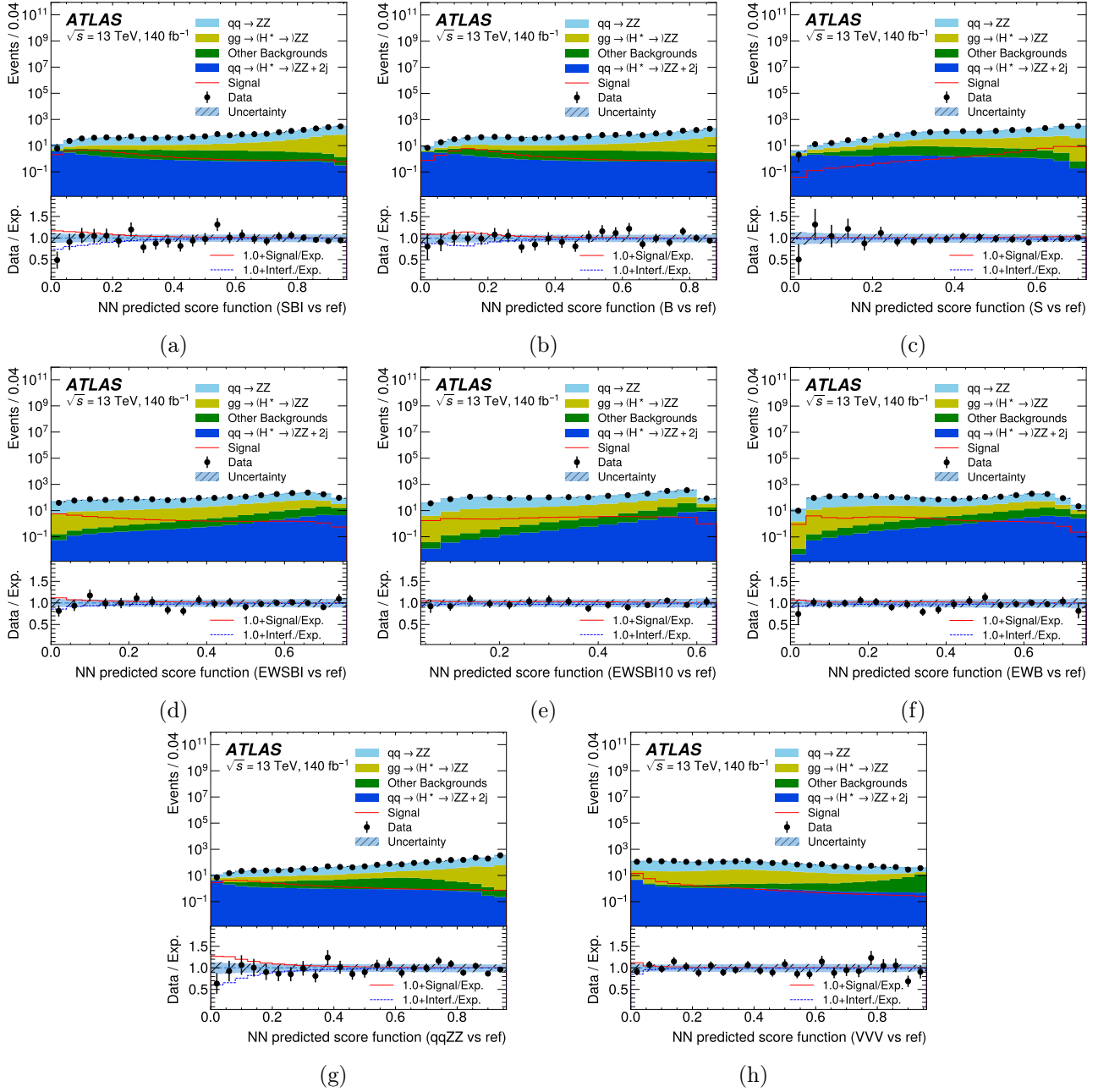


Figure 6.15: Comparison between data and expectation in the SR ($D_{\text{pre}}(\mathbf{x}) > -0.85$) of the NN scores of the all major samples used to describe the probability density ratio of the reference process and a ggF SBI, b ggF S, c ggF B, d EW SBI₁, e EW SBI₁₀, f EW B, g $q\bar{q} \rightarrow ZZ$, and h VVV . The comparisons show a good modeling of the probability density ratios throughout the SR. The lower panels show the ratio of data to expectation. The expected SM distributions are shown as stacked histograms and the expected signal (interference) is shown as a solid red (dashed blue) line. The background is estimated under the SM hypothesis (post-fit, $\mu_{\text{off-shell}} = 1$). The hatched band shows the total systematic uncertainty in the expected distribution. The last bin contains overflow events.

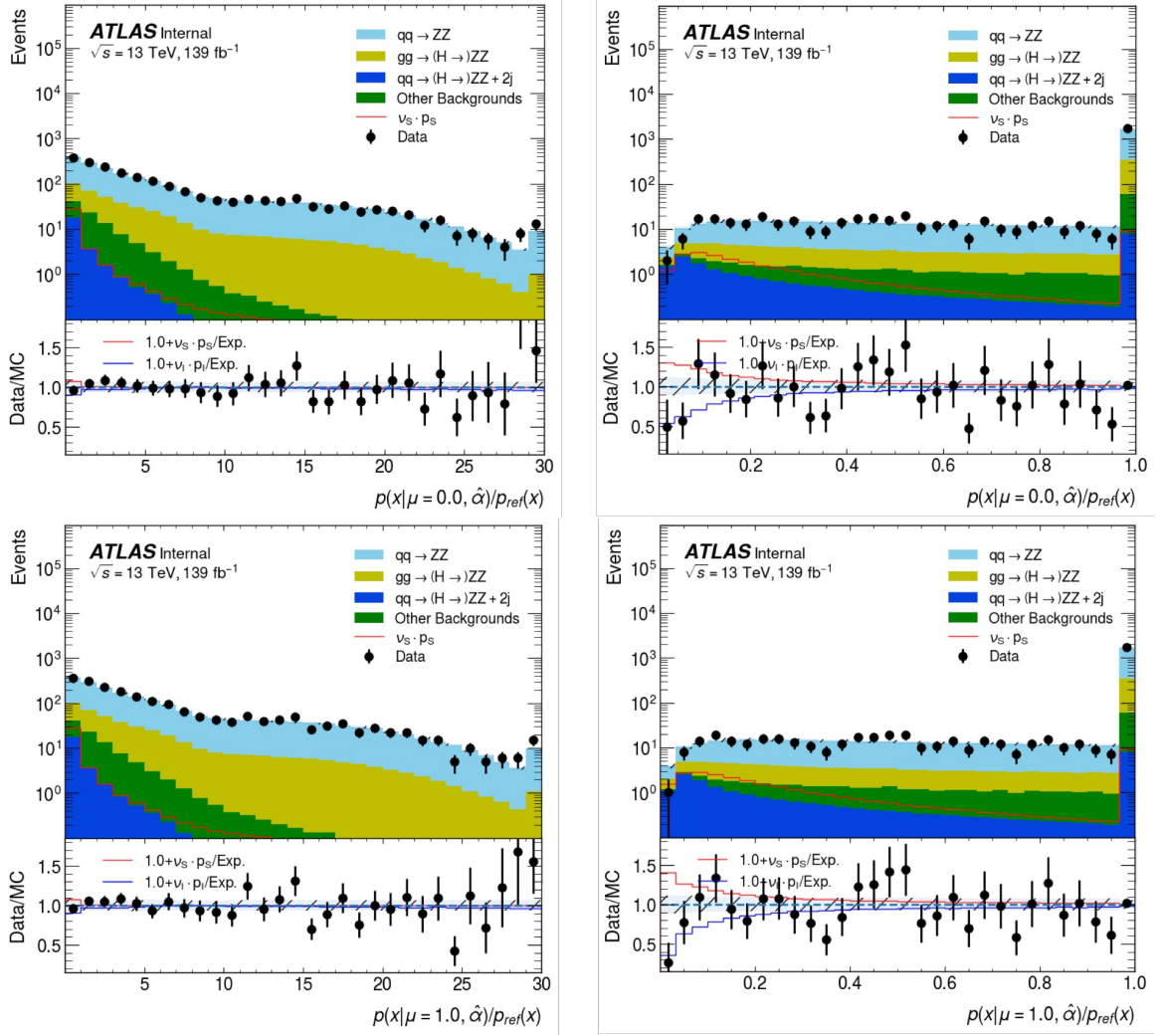


Figure 6.16: Post fit data-MC validation using data events from the SR phase space ($D_{\text{pre}}(\mathbf{x}) > -0.85$). The plots show various NN predicted $p(\mathbf{x}|\mu)/p_{\text{ref}}(\mathbf{x})$ distributions used directly in the profile likelihood fits for $\mu = 0.0$ (bkg-only hypothesis) and $\mu = 1.0$ (SM hypothesis). The left plots show the full range, while the right plots focus on the signal-rich $p_{\text{ref}} \gg p(\mu = 1.0)$ phase space regions. The data-MC agreement is good throughout the phase space.

Interpretations using Data-MC checks

Extensive data-MC checks can be performed as a function of several one-dimensional kinematic distributions to highlight where the data-MC differences in the final fit (if any) arise from. This stage helps with interpreting any excesses, or other anomalies, seen in the final fits using real data. Four of these checks for some of the kinematic observables used to train the analysis NNs were shown in Figs. 5.2 and 5.4.

Here, a new type of data-MC validations for deeper interpretability are presented. In an NSBI analysis of the type presented in this thesis, the MC and data events can be reweighted using the NN predicted density ratios:

$$w_{\text{new}} = R(\mathbf{x}|\mu, \alpha = \hat{\alpha}) \times w(\mathbf{x}|\mu = 1, \alpha = \hat{\alpha}) \quad (6.11)$$

where the weights correspond to the weighted events of the MC dataset generated at the corresponding μ value using the morphing formula of Eq. 4.26. For data events, the weights are replaced by 1.0. The density ratio $R(\mathbf{x})$ is the one that is already estimated using the NNs multiplied by the normalization factor:

$$R(\mathbf{x}|\mu, \alpha = \hat{\alpha}) = \frac{\nu(\mu, \alpha = \hat{\alpha})}{\nu(\mu = 1, \alpha = \hat{\alpha})} \frac{p(\mathbf{x}|\mu, \alpha = \hat{\alpha})}{p(\mathbf{x}|\mu = 1, \alpha = \hat{\alpha})} \quad (6.12)$$

The data-MC comparisons using the reweighted MC and real datasets can help highlight regions of one-dimensional kinematic phase space where data-MC difference are the most significant for different values of the scanned POI μ . **This enables a qualitative assessment of where the differences in data and MC scans (if any) come from, as a function of the scanned POI μ .** Fig.6.17 and Fig.6.18 show the post-fit plots for the preselection NN D_{pre} and $m_{4\ell}$ distributions, for various values of $\mu_{\text{off-shell}}$.

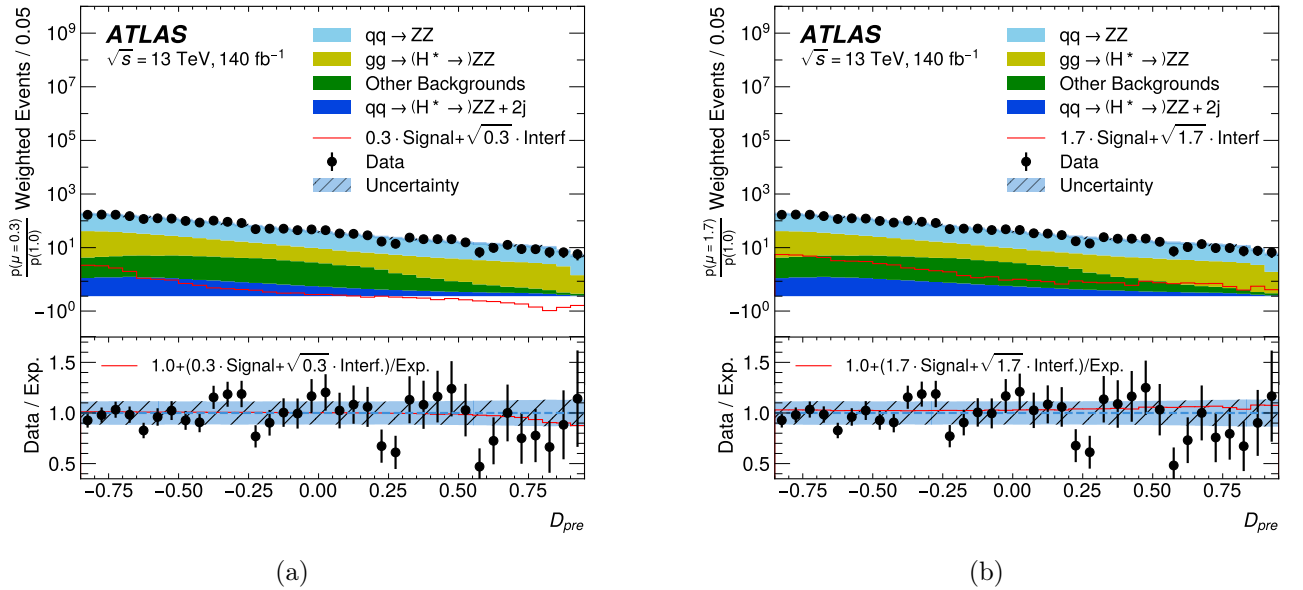


Figure 6.17: Comparison between data and expectation in the SR ($D_{\text{pre}}(\mathbf{x}) > -0.85$) of the preselection NN. In a, events are weighted by $p(\mathbf{x}|\mu = 0.3)/p(\mathbf{x}|\mu = 1)$ to highlight the low- $\mu_{\text{off-shell}}$ region dominated by the interference component. In b, events are weighted by $p(\mathbf{x}|\mu = 1.7)/p(\mathbf{x}|\mu = 1)$ to highlight the high- $\mu_{\text{off-shell}}$ region dominated by the signal component. The solid red line shows the sum of the signal and interference components estimated with the $\mu_{\text{off-shell}}$ hypothesis used to reweight the distributions. The lower panels show the ratio of the data to expectation. The background is estimated under the SM hypothesis (post-fit, $\mu_{\text{off-shell}} = 1$). The hatched area corresponds to the total systematic uncertainty in the expected distributions. The first and last bins contains overflow events.

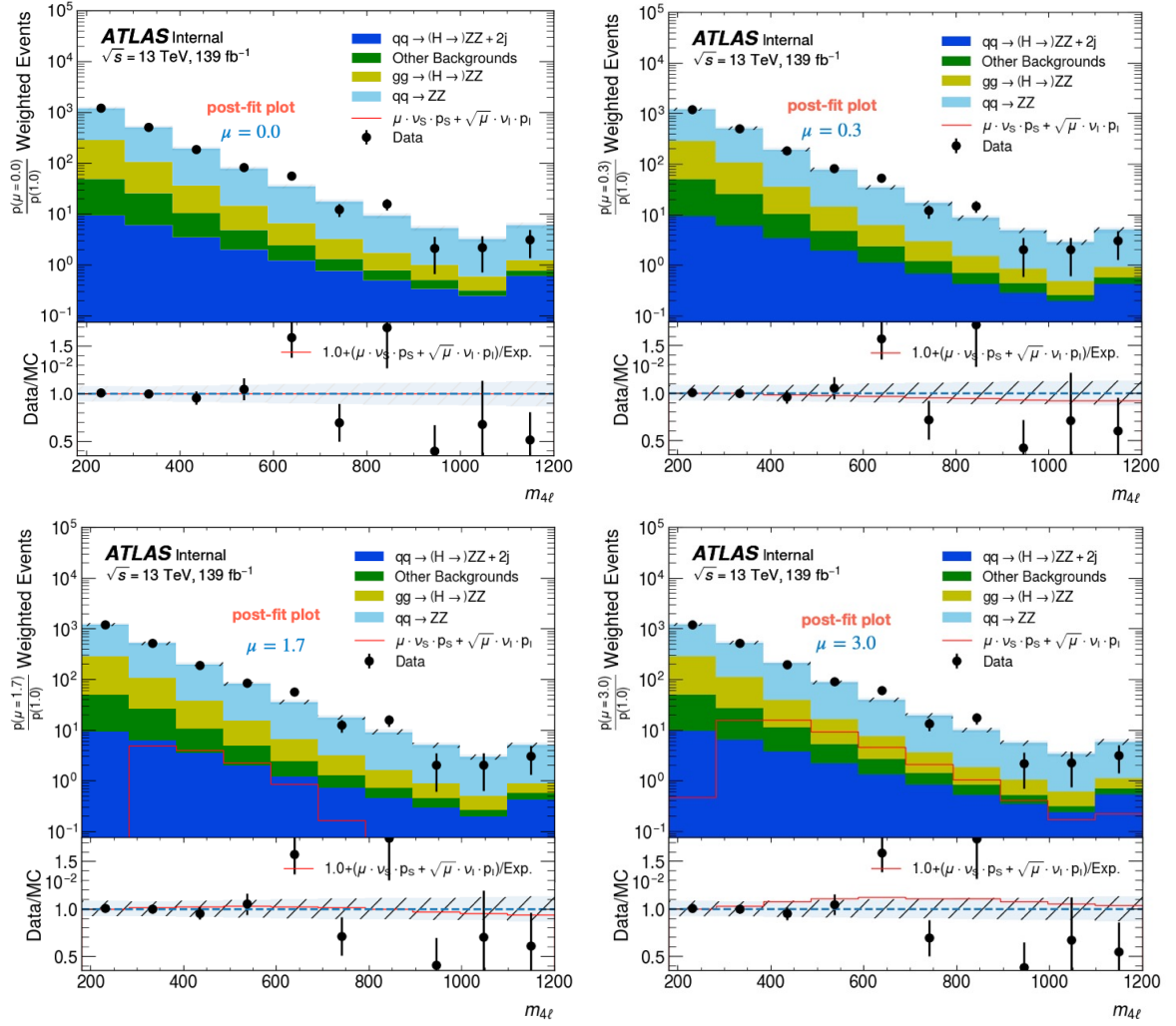


Figure 6.18: Post fit data-MC validation with data events exclusively from the $D_{pre} > -0.85$ signal region. From left to right, starting from the top: $p(\mathbf{x}|\mu = 0.0)/p(\mathbf{x}|\mu = 1.0)$ weighted events, $p(\mathbf{x}|\mu = 0.3)/p(\mathbf{x}|\mu = 1.0)$ weighted events, $p(\mathbf{x}|\mu = 1.7)/p(\mathbf{x}|\mu = 1.0)$ weighted events, $p(\mathbf{x}|\mu = 3.0)/p(\mathbf{x}|\mu = 1.0)$ weighted events.

Checking for Unbiased Estimator

Fig. 6.19 shows the results of an unbiased estimator closure test: the value of the estimates $\hat{\mu}_{\text{off-shell}}$ for Asimov samples with true values of $\mu_{\text{off-shell}}$ for a wide range of $\mu_{\text{off-shell}}$ values. Closure is observed for all values of $\mu_{\text{off-shell}}$, within the MC statistical uncertainty of the simulation samples used. An additional test was performed replacing the $q\bar{q} \rightarrow ZZ$ SHERPA sample (see Sec. 5.1) in the truth MC data by an alternative simulation done with the POWHEG simulator. The hard scattering ME in both samples is calculated at the same perturbative order in QCD, but the samples have different parton shower matching and simulation. Satisfactory closure of the $\mu_{\text{off-shell}}$ MLE is also observed with the alternative MC sample used for evaluation, despite the different physics in final state - highlighting the robustness of the density ratios.

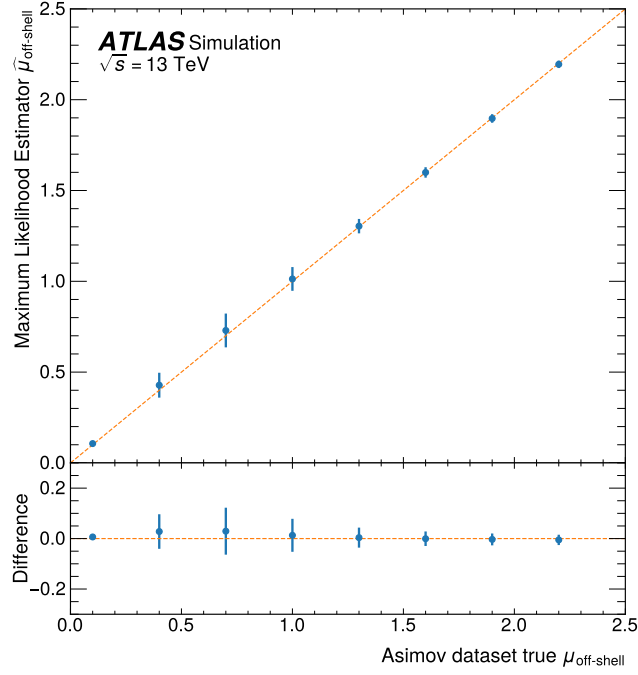


Figure 6.19: Maximum likelihood estimator $\hat{\mu}_{\text{off-shell}}$ when fitting $\mu_{\text{off-shell}}$ using MC samples with different true values $\mu_{\text{off-shell}}$. The error bar shows the estimate of the uncertainty from the limited number of MC events in the Asimov dataset, which is introduced as a spurious signal uncertainty as described in Sec. 6.5.

6.4 Estimating the systematic variations $g_c(\mathbf{x}, \alpha^\pm)$

Accurate estimates of $g_c(\mathbf{x}, \alpha^\pm)$ are needed in order to use the interpolation formula of Eq. 3.9. The $\pm 1\sigma$ simulation model is available in a typical ATLAS analysis as either dedicated per-event weights in the nominal MC sample, namely those coming from efficiency systematic uncertainties and from theoretical uncertainties with analytical expressions, or they are available as dedicated MC simulations altogether, namely those related to scale and resolution systematic uncertainties and theoretical uncertainties obtained from comparing different models. These alternate MC simulations are built by varying the sampling of latent variables in simulators. In the unbinned formalism developed in this note, the latent variables are integrated out and the systematic weight estimates need to be a function of the observable $\mathbf{x}$ only, so that it can be directly estimated from data.

In order to properly describe the systematic variation on a per-event basis, the systematic weights $g_c(\mathbf{x}, \alpha^\pm)$ need to be determined as a function of the input $\mathbf{x}$ used for the density ratio prediction. Due to the uncorrelated nature of the systematics model in the ATLAS experiment, as described in Chap. 3, the per-event weights $g_c(\mathbf{x}, \alpha_m^\pm)$ can be estimated systematic-by-systematic using dedicated NNs. The weights correspond to the probability density ratios between the varied sample and the nominal sample.

$$\frac{p_c(\mathbf{x}|\alpha_m^\pm)}{p_c(\mathbf{x}|\alpha_m=0)} = g_X(\mathbf{x}, \alpha_m^\pm) = \left. \frac{\hat{s}_{\text{pred}}(\mathbf{x})}{1 - \hat{s}_{\text{pred}}(\mathbf{x})} \right|_{\alpha_m^\pm} \quad (6.13)$$

where $\alpha_m = 0$ corresponds to the nominal probability model.

Using the same likelihood-ratio trick (or CARL technique) as was described in Sec. 6.1 for the estimation of the density ratios $p_c(\mathbf{x})/p_{\text{ref}}(\mathbf{x})$, the NN can then convert the description of systematic variations as a function of latent variables (modeled using MC techniques) into weight-based systematic uncertainties based only on reconstructed observables $\mathbf{x}$. To test the validity of the trained NNs, similar diagnostics to the ones described in Sec. 6.3 are used. Some examples are shown in Fig. 6.20.

For each systematic uncertainty nuisance parameter α_m , we train two NNs to estimate the variation weights corresponding to the $\alpha_m^\pm$ variations. This is repeated for the variation of each process c , to be used for the corresponding density ratio $p_c(\mathbf{x}|\alpha_m^\pm)/p_c(\mathbf{x}|\alpha_m^0 = 0)$. The total number of NNs used to estimate the full systematic uncertainty model ends up being equal to the number of nuisance parameters times the number of processes times two, somewhere in the order of $\sim O(10^3)$ for this analysis.

6.5 Using NN Ensembles to estimate MC Statistical Uncertainty

As described in Sec 6.2, an ensemble of NNs is trained to predict density ratios for every event (see Eq. 6.5). The observed distribution of the ensemble predictions can be used as an estimate of the statistical uncertainty coming from the limited MC statistics used for training the NNs. Each ensemble is trained using a subset of the original training data randomly sampled using a different train-test split, as described in Sec. 6.2.1, to mimic different datasets for each training and is an alternative to the standard bootstrap technique which involves randomly sampling from the training pool with replacement.

In the mixture models approach, the sample-specific density ratios p_c/p_{ref} are estimated using a separate ensemble of NNs. Fig. 6.21 shows the distribution of the $p_B^{\text{ggF}}/p_{\text{ref}}$ NN ensemble predictions for a set of randomly selected ggF signal (left) and ggF background (right) events. As expected, the signal events in the background-rich phase space $p_B^{\text{ggF}}/(p_{\text{ref}} + p_B^{\text{ggF}}) \rightarrow 1.0$ tend to have larger variances compared to the signal events in the signal-rich phase space $p_B^{\text{ggF}}/(p_{\text{ref}} + p_B^{\text{ggF}}) \ll 1.0$ and vice versa for background events.

The several diagnostic tools described in Sec. 6.3 show that the mean of these ensembles describes an unbiased density ratio. In order to provide an estimate of the uncertainty coming from the limited number of MC events, while keeping the density ratio unbiased, NN predictions $[p_c(\mathbf{x})/p_{\text{ref}}(\mathbf{x})]_a$ are bootstrapped from the pool of the NNs in the ensemble. Using the bootstrapped predictions, ensemble templates are formed as shown in Fig. 6.22, and used to evaluate the template mean prediction. This gives an estimate of the **variance on the mean** of the ensemble prediction by the definition of bootstrapping.

A distribution of the spread of ensemble means due to this template building procedure can be seen in the Fig. 6.23. Note that this procedure is identical to training multiple ensembles using unique samplings each time in the limit of $N_{\text{ens}} \rightarrow \infty$. The variance on the mean calculated from this procedure is equal to the standard formula for calculating the variance on the mean:

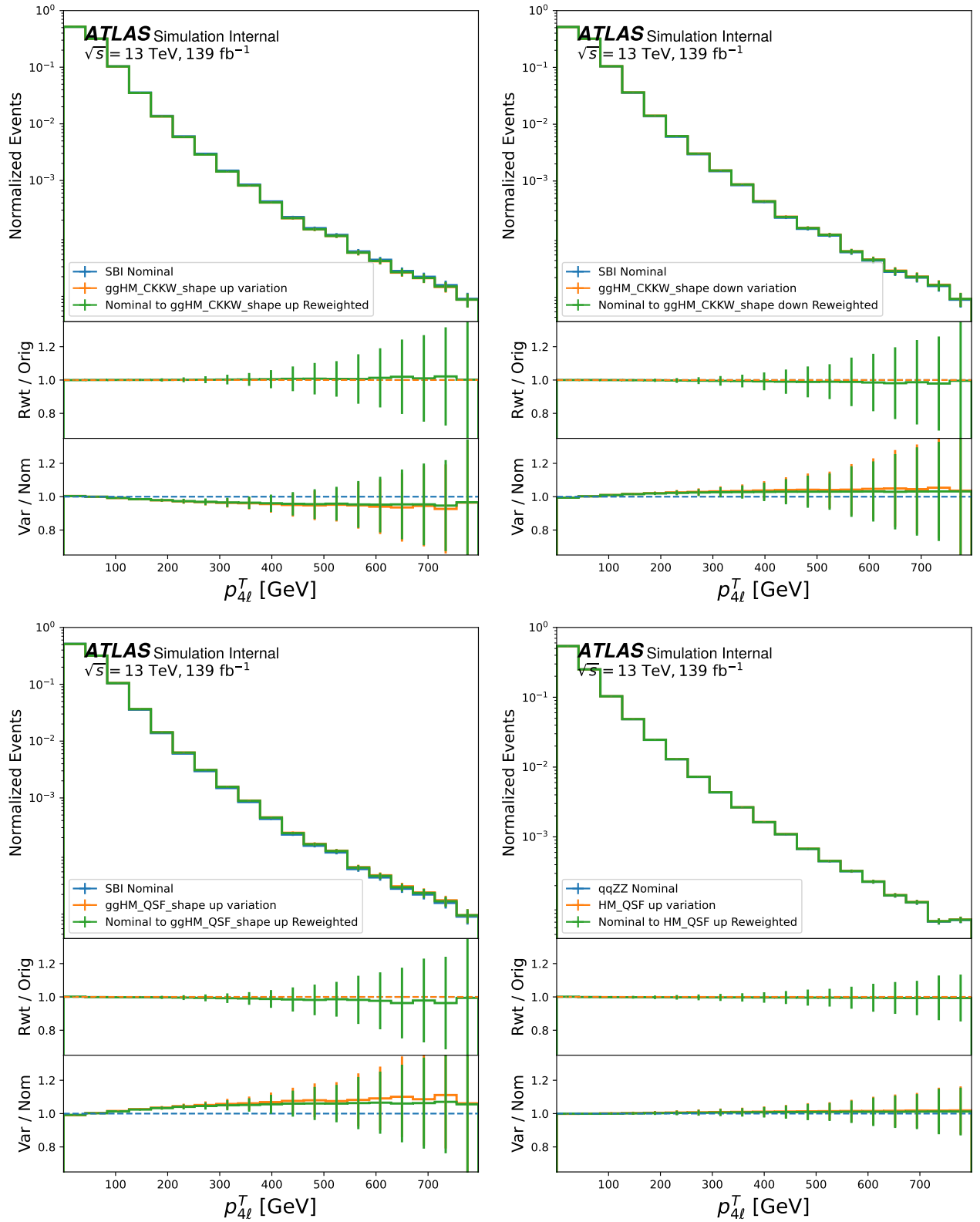


Figure 6.20: Example diagnostic plots for systematic uncertainty weight estimation, showing un-biased modelling of the systematic uncertainties. The first ratio plot shows the ratio between the nominal sample reweighted to the systematic variation and original systematic variation. The second ratio plot shows the ratio of systematic variation distribution (both reweighted and original - green and orange curves respectively) to the nominal (blue curve).

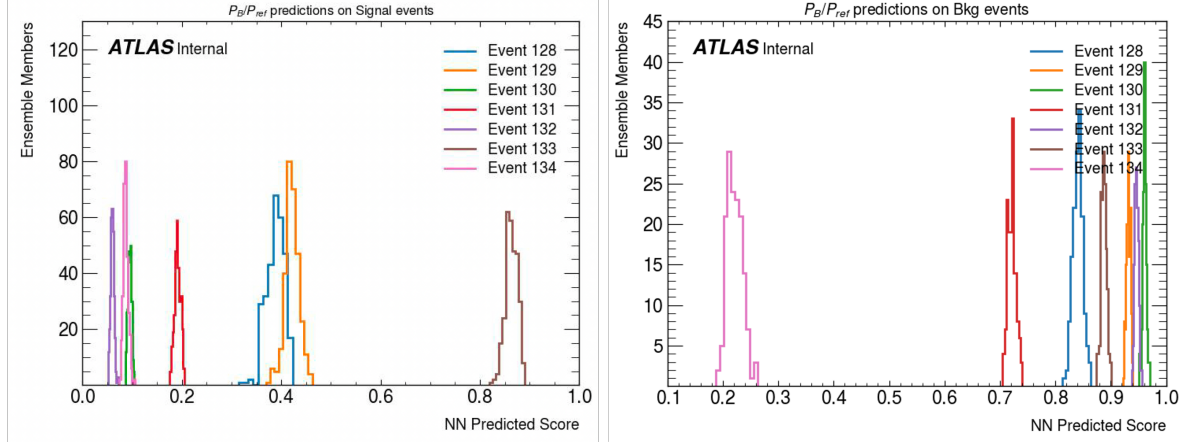


Figure 6.21: Ensemble NN score prediction distributions for random example events with NN trained for the $p_B^{\text{ggF}}/p_{\text{ref}}$ density ratio. Larger spreads indicate larger variances on individual NN predictions from the MC statistical uncertainty.

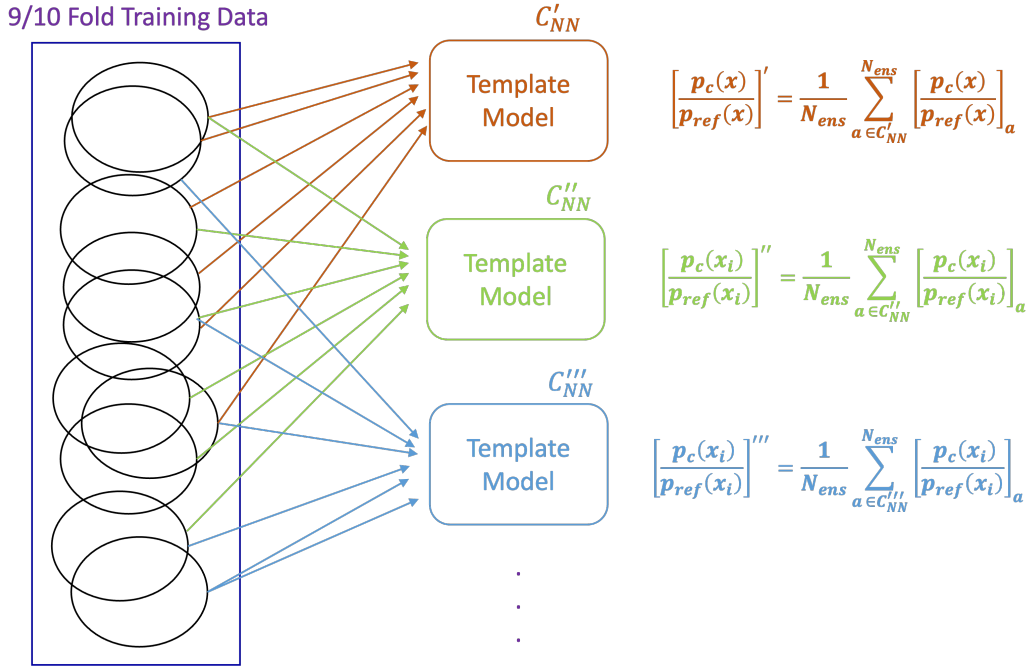


Figure 6.22: Building template ensembles from the trained NN pool to estimate the variance on the $\hat{\mu}$. Each blob in the box on the left represents the 80% of the training data sampled for each NN. These NNs have been trained already for the nominal fit, and it is the predictions from these trained NNs that are sampled N_{ens} times with replacement to build the template model. This is equivalent to sampling and ensembling independently in the limit of large enough N_{ens} . The mean of the predictions from these N_{ens} samples of each template are used to determine the variance on the mean. The template ensemble sampling is done for each channel $c = \text{ggF SBI, ggF B, ggF S, etc.}$ independently.

$$V \left[\frac{p_c}{p_{\text{ref}}} \right] = \frac{\sigma_{\text{ens}}}{\sqrt{N_{\text{ens}}}} \quad (6.14)$$

where σ_{ens} is the variance on the per-event ensemble predictions shown in Fig. 6.21. But the challenge

is propagating this variance to the profile likelihood fit described in Chap. 3, since it is important to preserve the event-by-event correlations in these MC statistical fluctuations.

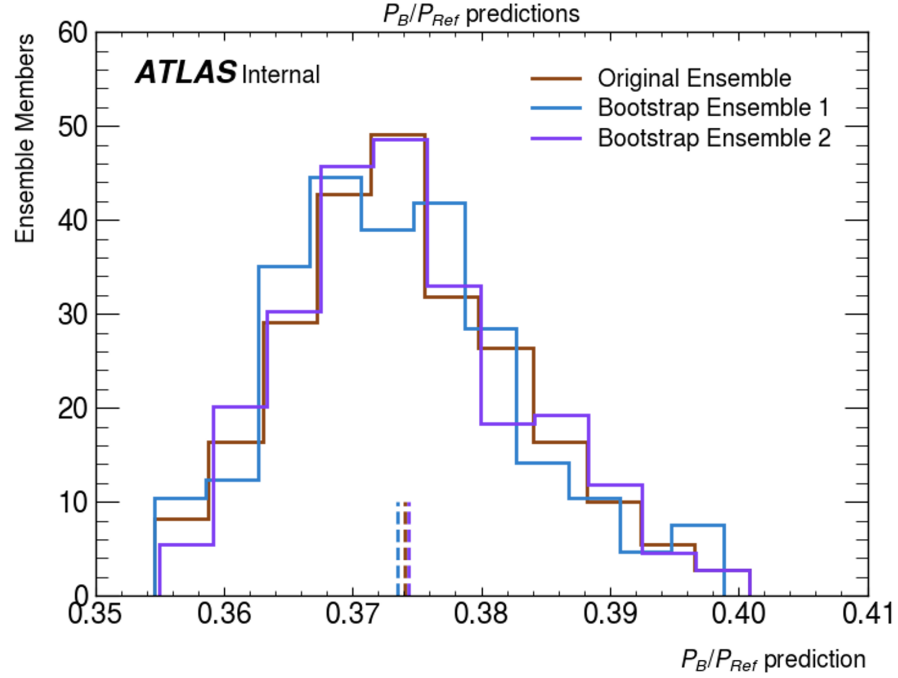


Figure 6.23: Ensemble prediction distributions for a randomly selected event for the $p_B^{\text{ggF}}/p_{\text{ref}}$ density ratio. As an example two samplings using the bootstrap technique are done to show the corresponding spread in the ensemble mean prediction (dashed lines).

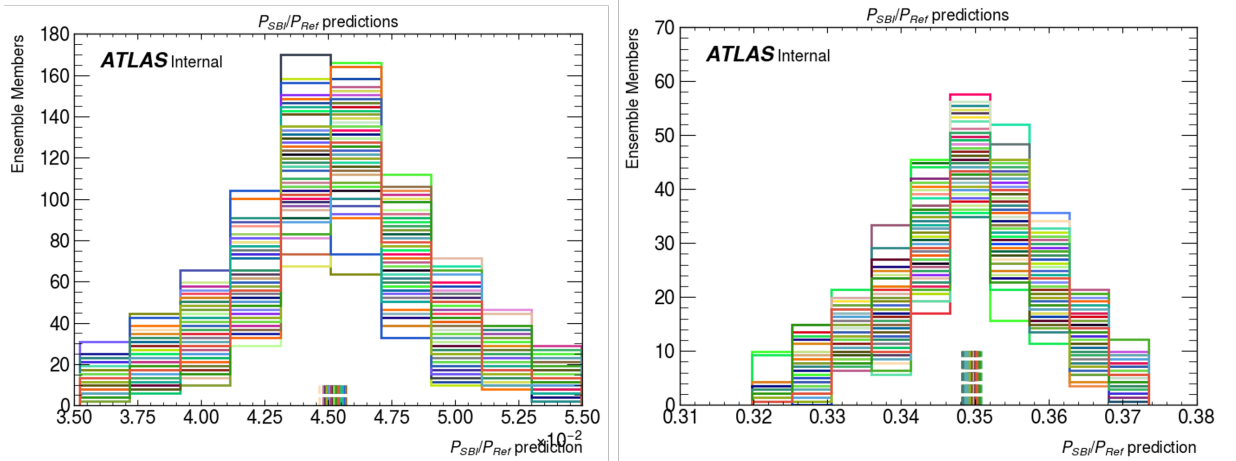


Figure 6.24: Ensemble $p_{\text{SBI}}^{\text{ggF}}/p_{\text{ref}}$ prediction distributions for two randomly selected events. A hundred samplings of the density ratio are done using the bootstrap technique described in text, to show the corresponding spread in the ensemble mean prediction (dashed lines).

6.5.1 Monte Carlo Uncertainty Propagation

While the method discussed above gives an estimate of the variance on the mean as a function of the event $\mathbf{x}$, one needs to propagate this uncertainty in the profile log-likelihood ratio test statistic, while preserving all the correlations between each event. For this a similar strategy as the spurious signal uncertainty [57] used in unbinned analysis like $H\gamma\gamma$ [41], has been adopted for the NSBI analysis.

Using bootstrapped ensembles for predictions of all the channels $p_c(\mathbf{x})/p_{\text{ref}}(\mathbf{x})$, the POI μ is fitted using the MC dataset, with various truth values $\mu = \mu_{\text{truth}}$, per-ensemble template - and the resulting spread of the fitted values $\hat{\mu}$ is used as an uncertainty on the POI μ from the limited MC statistics. For each fit, a randomly sampled ensemble for each of the channels is used to calculate the mean of the event-by-event density ratio predictions p_c/p_{ref} as in Eq. 6.5, then used to calculate the test statistic function of Eq. 5.5 to find the best fit value $\hat{\mu}$. The resulting distribution $\Delta\hat{\mu}$ represents the uncertainty in the fit at $\mu = \mu_{\text{truth}}$ value of the MC dataset used in the fit, due to the limited Monte Carlo statistics used in the training of the NNs. This can be repeated using MC samples at different values of $\mu = \mu_{\text{truth}}$ to then build the parameterized $\Delta\hat{\mu}(\mu)$ for a more robust estimate of the uncertainty across the parameter space. The distribution for $\mu = 1.0$, shown in Fig. 6.25, is centered on the nominal fit value (dashed blue line) and covers the truth value $\mu = 1.0$ within $\sim 1\sigma$ uncertainty. For different values of μ , this uncertainties and corresponding best fit values can be seen in Fig. 6.19.

Drawing inspiration the spurious signal method, a single nuisance parameter α_{stat} is introduced in the μ parameterization of the full probability model:

$$\sum_c f_c(\mu) \rightarrow \sum_c f_c(\mu + \alpha_{\text{stat}} \cdot \Delta\hat{\mu}(\mu)) \quad (6.15)$$

Inserting Eq. 6.15 into the full off-shell Higgs boson likelihood model of Eq. 4.30 and then profiling the α_{stat} dependence in a test statistic fit accounts for the MC statistical uncertainty in the final fitted parameter values. For the off-shell Higgs boson analysis presented in this thesis, the contribution from this uncertainty was found to be negligible.

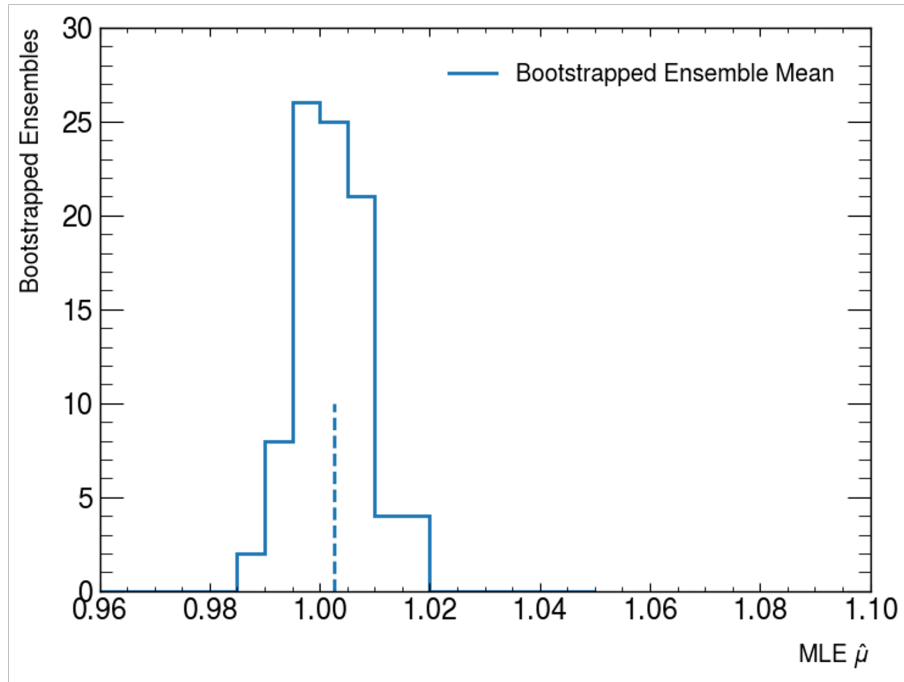


Figure 6.25: Distribution of fitted $\hat{\mu}$ from the various bootstrapped ensembles. The dashed blue line at $\mu = 1.002$ represents the value of the nominal fit $\hat{\mu}_{\text{nom}}$ using the nominal density ratio predictions. For each of the subsequent fits, a randomly sampled template ensemble for each of the channels $c = \text{ggF SBI, ggF B, ggF S, etc.}$ is used to calculate the mean of the event-by-event density ratio predictions p_c/p_{ref} which are then used to find the best fit value of $\hat{\mu}$. The resulting distribution $\Delta\hat{\mu}$ represents the uncertainty in the fit at $\mu = 1.0$ due to the limited Monte Carlo statistics used in the training of the NNs. This is repeated using MC samples at different values of $\mu = \mu_{\text{truth}}$ to then build the parameterized $\Delta\hat{\mu}(\mu)$.

Chapter 7

Results

This chapter presents the statistical inference results using NSBI technique on the parameters of interest in the off-shell Higgs boson analysis. The new results are also compared with the previously published results using the same Run-2 data to highlight the gains in sensitivity from using the new technique.

7.1 Results in the $ZZ \rightarrow 4\ell$ decay channel

To get the final results, a combined fit in the SR and the CR defined in Sec. 5.4 is performed using both the Asimov MC dataset and the real detector data. The events of the SR use the per-event likelihood ratio model of Eq. 4.30:

$$\begin{aligned} \frac{p(\mathcal{D}|\mu_{\text{off-shell}}, \theta_{qq}, \boldsymbol{\alpha})}{\prod_k^{\mathcal{N}_{\text{data}}} p_{\text{ref}}(\mathbf{x}_k)} &= \text{Pois} \left(\mathcal{N}_{\text{data}} \middle| \sum_c f_c(\mu_{\text{off-shell}}, \theta_{qq}) \cdot \left[\prod_{m=1}^{N_{\text{syst}}} G_c(\alpha_m) \right] \nu_c \right) \times \\ &\quad \frac{1}{\sum_c f_c(\mu_{\text{off-shell}}, \theta_{qq}) \cdot \left[\prod_{m=1}^{N_{\text{syst}}} G_c(\alpha_m) \right] \nu_c} \times \\ &\quad \prod_k^{\mathcal{N}_{\text{data}}} \left[\sum_c f_c(\mu_{\text{off-shell}}, \theta_{qq}) \cdot \left[\prod_{m=1}^{N_{\text{syst}}} G_c(\alpha_m) \right] \nu_c \cdot \left[\prod_{m=1}^{N_{\text{syst}}} g_c(\mathbf{x}|\alpha_m) \right] \frac{p_c}{p_{\text{ref}}}(\mathbf{x}_k) \right], \end{aligned} \quad (7.1)$$

To this likelihood model, the Poisson likelihoods of the CR (defined by $D_{\text{pre}}(\mathbf{x}) \leq -0.85$) bins are combined. The CR uses a fixed summary statistic, which is the n_{jets} bins: 0, 1, and ≥ 2 to provide enough constraining power for the $\theta_{qq} = \left\{ \theta_{q\bar{q}ZZ}^{0j}, \theta_{q\bar{q}ZZ}^{1j}, \theta_{q\bar{q}ZZ}^{2j} \right\}$ background normalization parameters, which are treated as unconstrained NPs, i.e., they have no corresponding auxiliary measurements that constrain them. This is discussed in Sec. 5.4. The full negative log-likelihood model is then given by Eq. 5.4

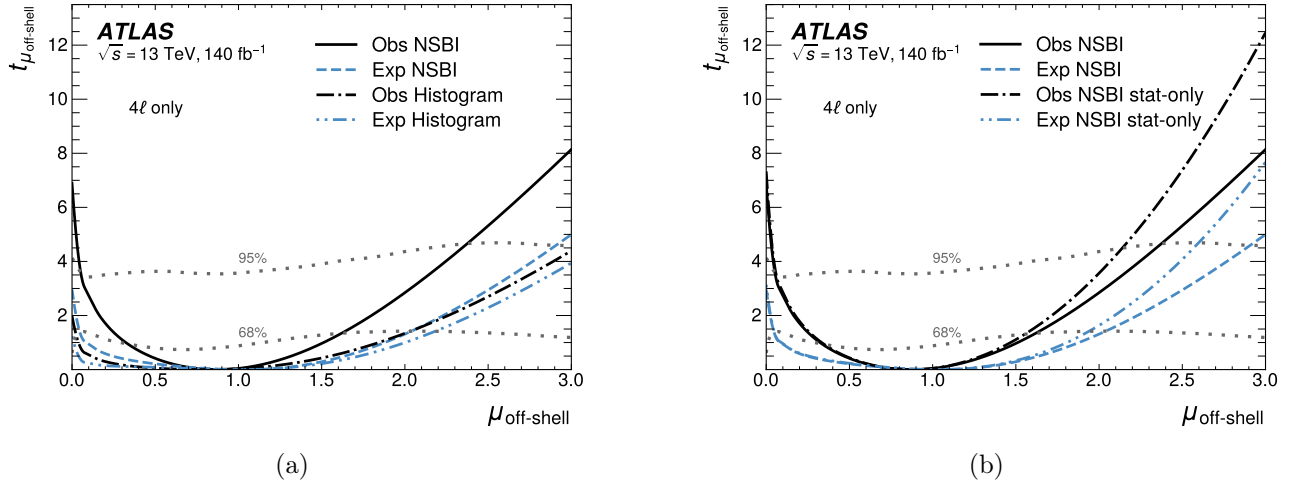


Figure 7.1: a Values of the test statistic $t_{\mu_{\text{off-shell}}}$ assuming a single parameter of interest $\mu_{\text{off-shell}}$ obtained with an Asimov dataset (expected, dashed blue) and with data (observed, solid black) in the $H^* \rightarrow ZZ \rightarrow 4\ell$ decay channel. The values from the histogram-based analysis [3], using fixed summary statistic of Eq. 4.29 for the fits, are added in dash-dotted lines for comparison. The dotted gray lines show the 68% and 95% confidence belt, obtained from the Neyman construction. b Same values obtained with data (observed, solid black) and Asimov dataset (expected, dashed blue) compared with the *statistics-only* case with all NP fixed at their best-fit values $\hat{\alpha}$.

$$\begin{aligned}
 -2 \log \frac{L(\mu_{\text{off-shell}}, \theta_{qq}, \boldsymbol{\alpha})}{\prod_k^{\mathcal{N}_{\text{obs}}} p_{\text{ref}}(\mathbf{x})} &= \sum_{\text{bin} \in n_{\text{jets}}} \left[-2 \mathcal{N}_{\text{CR}}^{\text{bin}} \cdot \log \nu_{\text{CR}}^{\text{bin}}(\mu_{\text{off-shell}}, \theta_{qq}, \boldsymbol{\alpha}) + 2 \nu_{\text{CR}}^{\text{bin}}(\mu_{\text{off-shell}}, \theta_{qq}, \boldsymbol{\alpha}) \right] \\
 &\quad - 2 \mathcal{N}_{\text{data}}^{\text{SR}} \cdot \log \nu_{\text{SR}}(\mu_{\text{off-shell}}, \theta_{qq}, \boldsymbol{\alpha}) + 2 \nu_{\text{SR}}(\mu_{\text{off-shell}}, \theta_{qq}, \boldsymbol{\alpha}) \\
 &\quad - 2 \sum_i^{\mathcal{N}_{\text{data}}} \log \frac{p(\mathbf{x}_i | \mu_{\text{off-shell}}, \theta_{qq}, \boldsymbol{\alpha})}{p_{\text{ref}}(\mathbf{x})} + \sum_m^{\mathcal{N}_{\text{syst}}} \alpha_m^2
 \end{aligned} \tag{7.2}$$

which is used to fit the parameters using the test statistic function of Eq. 5.5.

7.1.1 Test statistic scans

The observed and expected values of the profile likelihood ratios $t_{\mu_{\text{off-shell}}}$ are shown in Fig. 7.1 as functions of $\mu_{\text{off-shell}}$ for a variety of scenarios, assuming $\mu_{\text{off-shell}} = \kappa_{g,\text{off-shell}}^2 \kappa_{V,\text{off-shell}}^2 = \kappa_{V,\text{off-shell}}^4$. Fig. 7.1a compares the profile likelihood ratio for the histogram-based [3] and the NSBI-based analyses, and shows the improved constraints on $\mu_{\text{off-shell}}$ obtained with the latter. The histogram-based analysis uses the fixed summary statistic defined in Eq. 4.29. Additional comparisons with the histogram-based analysis are shown in Appendix A. Fig. 7.1b compares the profile likelihood ratio for the NSBI-based analysis to a variant where the NPs are fixed to the best-fit value $\hat{\alpha}$, reflecting only the statistical uncertainty on the data. The comparison indicates that systematic uncertainties are more important for tests of signal-dominated hypotheses ($\mu_{\text{off-shell}} > 1$) than for tests of interference-dominated hypotheses ($\mu_{\text{off-shell}} < 1$).

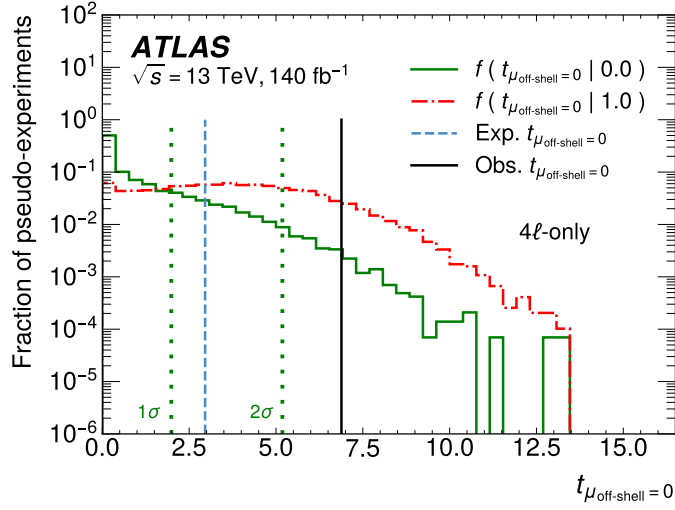


Figure 7.2: Expected distribution of $t_{\mu_{\text{off-shell}}=0}$ estimated with pseudo-experiments for the case of $\mu_{\text{off-shell}}^{\text{truth}} = 0$ (solid green, no off-shell Higgs boson production hypothesis) and $\mu_{\text{off-shell}}^{\text{truth}} = 1$ (dash-dotted red, SM hypothesis). The vertical solid black (dashed blue) line shows the observed (expected) value of $t_{\mu_{\text{off-shell}}=0}$. The dotted vertical green lines at $t_{\mu_{\text{off-shell}}=0} = 1.99$ and 5.18 represent the one-sided 1σ and 2σ significance thresholds, respectively, under the $\mu_{\text{off-shell}}^{\text{truth}} = 0$ hypothesis. These values differ slightly from the ones in the asymptotic approximation, in which the thresholds would be 1.97 and 5.19 .

7.1.2 Neyman Construction

As is the case for the histogram-based analysis, the test statistic is not distributed as a χ^2 probability density, due to the double minima created by the interference terms and due to the constraint $\mu_{\text{off-shell}} > 0$ imposed by the model of Eq. 4.19. Confidence intervals are built using the Neyman construction [124] (NC) instead of relying on the asymptotic approximation. In the NC, pseudo-experiments are built by performing a Poisson bootstrapping on the high-statistic reference sample as described in Sec. 3.3.

Fig. 7.2 shows the expected distribution of $t_{\mu_{\text{off-shell}}=0}$ for the SM hypothesis ($\mu_{\text{off-shell}}^{\text{truth}} = 1$) and the *no off-shell Higgs boson hypothesis* ($\mu_{\text{off-shell}}^{\text{truth}} = 0$). The black and dashed blue vertical lines indicate the observed and expected values of $t_{\mu_{\text{off-shell}}=0}$, respectively. The red dash-dotted curve in Fig. 7.2 shows the expected distribution of $t_{\mu_{\text{off-shell}}=0}$ assuming the SM hypothesis ($\mu_{\text{off-shell}}^{\text{truth}} = 1$). The p -value of the observed value of $t_{\mu_{\text{off-shell}}=0}$ under the SM hypothesis is 0.11 , corresponding to one-sided significance of 1.2σ . The green solid curve shows the expected distribution of $t_{\mu_{\text{off-shell}}=0}$ assuming $\mu_{\text{off-shell}}^{\text{truth}} = 0$ (no off-shell Higgs boson hypothesis). The green dotted lines show the p -value thresholds corresponding to the one-sided significance of 1σ and 2σ under this hypothesis. The evidence for off-shell Higgs boson production has an observed (expected) significance of 2.5σ (1.3σ) using only the $H^* \rightarrow ZZ \rightarrow 4\ell$ decay channel. The evidence for off-shell Higgs boson hypothesis has a larger significance than the one observed (expected) in the previous histogram-based analysis [3] of the same dataset, which had a value of 0.8σ (0.5σ).

Nuisance Parameter Pulls

A total of 127 NPs describing the full systematics model of Sec. 4.6 make up α in the off-shell Higgs boson measurement, and their fitted values (pulls) and pull errors are shown in Fig. 7.3. No significant differences between the MLE estimates $\hat{\alpha}_m$ from fits to real data and the corresponding auxiliary values $\alpha_m^{(0)} = 0$ are observed.

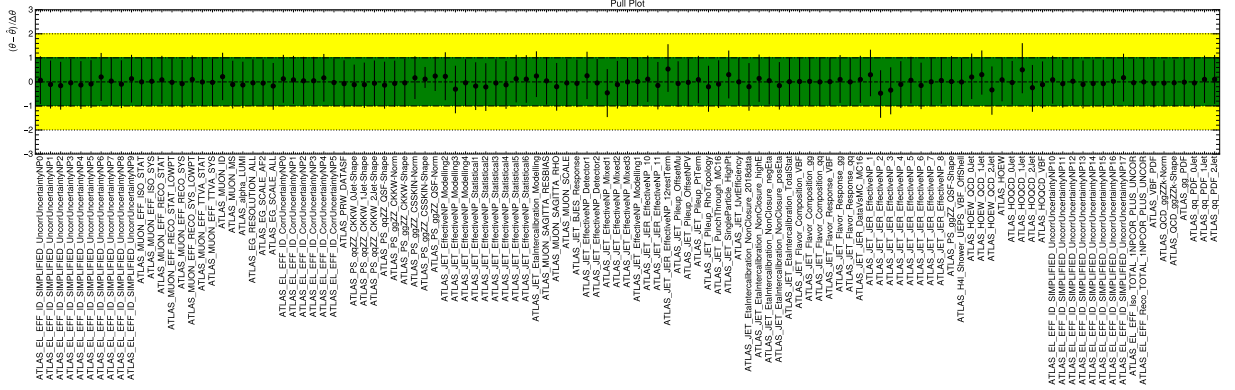


Figure 7.3: Post-fit pulls and pull errors of the all the nuisance parameters representing the systematic uncertainties in the $\mu_{\text{off-shell}}$ fit.

7.1.3 Data-MC checks - optimal observables

Fig. 7.4a shows the distribution of the probability density ratio $p(\mathbf{x}|\mu_{\text{off-shell}} = 0, \hat{\alpha})/p(\mathbf{x}|\mu_{\text{off-shell}} = 1, \hat{\alpha})$, which is a near-optimal observable for $\mu_{\text{off-shell}} = 0$. The lower panel shows a comparison with the distribution from the best-fit hypothesis depicting the data behavior that leads to the observed exclusion of the no off-shell Higgs boson hypothesis. The difference between the observed and expected values of $t_{\mu_{\text{off-shell}}=0}$ indicates that there are regions of phase-space where the data are more interference-like than signal-like, as can be seen by the rightmost bins in Fig. 7.4a where the observed deficit of events is larger than expect. Given the small significance of the difference between expected and observed values of $t_{\mu_{\text{off-shell}}=0}$, it is difficult to isolate a specific region of phase-space with this behavior.

Fig. 7.4b shows the distribution of the quadruplet mass $m_{4\ell}$ and, in the lower panel, the comparison with the best-fit hypothesis indicating that only the quadruplet mass information would not be enough to obtain evidence of off-shell Higgs boson production in this channels and the importance of the ME-based analysis performed with the NSBI method. Further descriptions of the optimal observables can be found in Sec. 2.5.3.

7.1.4 Impacts from systematic uncertainties

Two methods are used to estimate the sensitivity of the measurement to different systematic uncertainties. The two methods differ in what is varied: either the NPs α_m , or of the AOs a_m associated with these NPs [55]. The specifics of the two methods are described in Sec. 3.2. The results of both methods are summarized in Tab. 7.1. The propagated uncertainties are summed in quadrature within

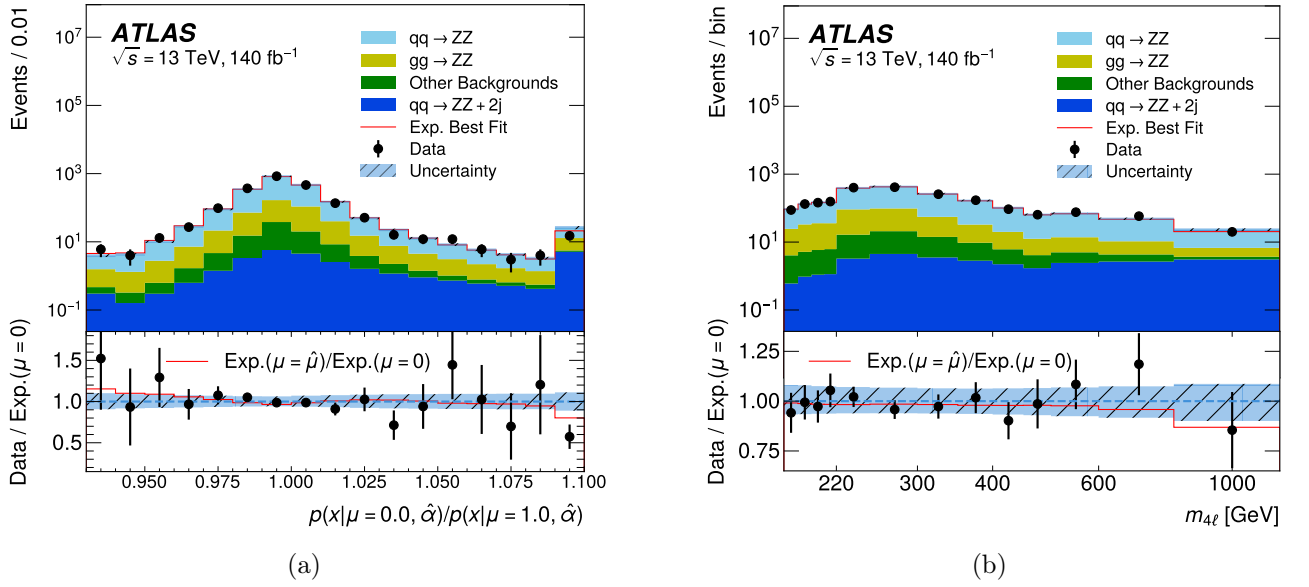


Figure 7.4: Comparison between observed and the expected background distributions of a the optimal observable at $\mu_{\text{off-shell}} = 0$ and b the quadruplet mass $m_{4\ell}$. The solid red lines shows the expected distribution of the best-fit hypothesis $\hat{\mu}$. The lower panel shows a comparison between the distribution of the background-only and best-fit hypotheses. The background is estimated under the SM hypothesis (post-fit, $\mu_{\text{off-shell}} = 1$). A comparison between the two distributions indicate that the optimal observable built with the NSBI method provides better evidence for off-shell Higgs boson production than only the $m_{4\ell}$ distribution. The hatched area corresponds to the total systematic uncertainty in the expected distributions. The first and last bins contain overflow events.

Table 7.1: Absolute systematic uncertainties in the measurement of $\mu_{\text{off-shell}}$ in the $H \rightarrow ZZ \rightarrow 4\ell$ decay channel. Two methods of estimation are presented: based on the variation of nuisance parameters and on the variation of auxiliary observables. Uncertainties are given using the auxiliary observables methods since it allows variations to be summed in quadrature. The total uncertainty is independent of the method used to estimate systematic uncertainties.

Uncertainty source	Absolute impact on $\mu_{\text{off-shell}}$	
	Nuisance Parameter	Auxiliary Observable
Electron uncertainties	(−0.05, +0.06)	(−0.05, +0.06)
Muon uncertainties	(−0.03, +0.03)	(−0.02, +0.03)
Jet uncertainties	(−0.10, +0.10)	(−0.09, +0.11)
Luminosity	(−0.01, +0.01)	(−0.01, +0.01)
Total experimental	(−0.12, +0.13)	(−0.11, +0.12)
$q\bar{q} \rightarrow ZZ$ modeling	(−0.06, +0.07)	(−0.06, +0.07)
$gg \rightarrow ZZ$ modeling	(−0.08, +0.13)	(−0.07, +0.09)
EW $q\bar{q} \rightarrow ZZ + 2j$ modeling	(−0.01, +0.01)	(−0.01, +0.01)
Total modeling	(−0.10, +0.15)	(−0.09, +0.12)
Systematic uncertainty	(−0.16, +0.19)	(−0.14, +0.17)
Statistical uncertainty	(−0.49, +0.72)	(−0.50, +0.73)
Total uncertainty	(−0.54, +0.75)	

each group. The sum in quadrature does not take into account correlations between the different parameters $\hat{\alpha}_m$ when using the method based on the variation of NPs. The same problem does not exist when varying AOs: the sum in quadrature can be made without loss of information and it can be used to provide a consistent decomposition of the total uncertainty into statistical and systematic uncertainties. In both methods, the statistical uncertainty is obtained by the square root of the difference of the total uncertainty squared and the systematic uncertainty squared.

The largest contributions to the measurement uncertainty are the statistical uncertainty on the data, the theoretical modeling uncertainties, and uncertainties on the jet energy scale and resolution. The contribution of MC statistical uncertainty to the total uncertainty is less than 0.01. The observed (expected) value of $\mu_{\text{off-shell}}$ at 68% CL is:

$$\mu_{\text{off-shell}} = 0.87^{+0.75}_{-0.54} \quad (1.00^{+1.04}_{-0.95}).$$

The fitted values of $\theta_{q\bar{q}ZZ}^{0j}$, $\theta_{q\bar{q}ZZ}^{1j}$, and $\theta_{q\bar{q}ZZ}^{2j}$ are 1.12 ± 0.04 , 0.85 ± 0.05 , and 0.90 ± 0.07 , respectively. The result presented here has a reduced uncertainty when compared with the previous histogram-based analysis observed value at 68% CL of $0.79^{+1.21}_{-0.77}$ (expected $\mu_{\text{off-shell}} < 1.14$).

7.2 Combination with the analysis in the $2\ell 2\nu$ decay channel

The new result presented in this paper is combined with the most recent ATLAS off-shell Higgs boson production measurement in the $H^* \rightarrow ZZ \rightarrow 2\ell 2\nu$ decay channel [3]. The histogram-based $H^* \rightarrow ZZ \rightarrow 2\ell 2\nu$ analysis uses the binned transverse mass m_{T}^{ZZ} distribution in three different SRs, in addition to the yield of four CRs enriched in $Z + \text{jets}$, non-resonant $e\mu$ events, and $q\bar{q} \rightarrow WZ$ events. The test statistic used for the combination is built from the log-likelihood ratio in Eq. 7.2, where now an additional Poisson likelihood sum (like in the CR) is included for the several m_{T}^{ZZ} bins and CRs of the analysis in the $H^* \rightarrow ZZ \rightarrow 2\ell 2\nu$ decay channel.

The systematic uncertainty model used here is expanded to include constrained and unconstrained NPs exclusive to the analysis of the $H^* \rightarrow ZZ \rightarrow 2\ell 2\nu$ decay channels. Experimental and common theory uncertainties, i.e., pertaining to $gg \rightarrow ZZ$ and $q\bar{q} \rightarrow ZZ$ modeling, are modeled with common NPs that modify both likelihood components. Uncertainties related to missing higher-order EW corrections in the $q\bar{q} \rightarrow ZZ$ process, which were estimated with different methods in both analyses, are modeled with separate NPs for the $H^* \rightarrow ZZ \rightarrow 4\ell$ and $H^* \rightarrow ZZ \rightarrow 2\ell 2\nu$ analyses. However, the measured off-shell Higgs boson production is largely insensitive to this modeling choice. Common $q\bar{q} \rightarrow ZZ$ data-driven normalization parameters $\theta_{q\bar{q}ZZ}^{0j}$, $\theta_{q\bar{q}ZZ}^{1j}$, and $\theta_{q\bar{q}ZZ}^{2j}$ are used in the two channels since the phase space probed is similar within each n_{jets} bin.

Fig. 7.5a shows the test statistic values as a function of $\mu_{\text{off-shell}}$, for a joint likelihood model with a single common parameter of interest $\mu_{\text{off-shell}}$ ($\kappa_{g,\text{off-shell}}^2 \kappa_{V,\text{off-shell}}^2 = \kappa_{V,\text{off-shell}}^4$). Similar to the result using only the $ZZ \rightarrow 4\ell$ channel, the comparison indicates that systematic uncertainties are more important for tests of signal-dominated hypotheses ($\mu_{\text{off-shell}} \gg 1$) than for tests of interference-dominated hypotheses ($\mu_{\text{off-shell}} \ll 1$). Fig. 7.5b shows the expected distribution of $t_{\mu_{\text{off-shell}}}$ for the SM hypothesis ($\mu_{\text{off-shell}}^{\text{truth}} = 1$) and the no off-shell Higgs boson hypothesis ($\mu_{\text{off-shell}}^{\text{truth}} = 0$) using the bootstrapping technique described above. The observed (expected) value of $\mu_{\text{off-shell}}$ at 68% CL is:

$$\mu_{\text{off-shell}} = 1.06_{-0.45}^{+0.62} \quad (1.00_{-0.83}^{+0.83}).$$

The result using uncertainties at 95% CL can be found in Tab. 7.2. The evidence for off-shell Higgs boson production has an observed (expected) significance of 3.7σ (2.4σ).

7.3 Reinterpretations in the κ -framework

The probability density ratio model of Eq. 4.26 can be rewritten in the κ -framework by substituting:

$$\mu_{\text{off-shell}}^{\text{ggF}} = \kappa_{g,\text{off-shell}}^2 \kappa_{V,\text{off-shell}}^2, \quad \mu_{\text{off-shell}}^{\text{EW}} = \kappa_{V,\text{off-shell}}^4, \quad (7.3)$$

Rewriting the full negative log-likelihood model in terms of these parameters then allows for a reinterpretation of the off-shell Higgs boson measurement in terms of the κ -framework. Note the advantage of using a mixture model approach, which trains only non-parameterized density ratios p_c/p_{ref} that can be re-used without re-training in a different model:

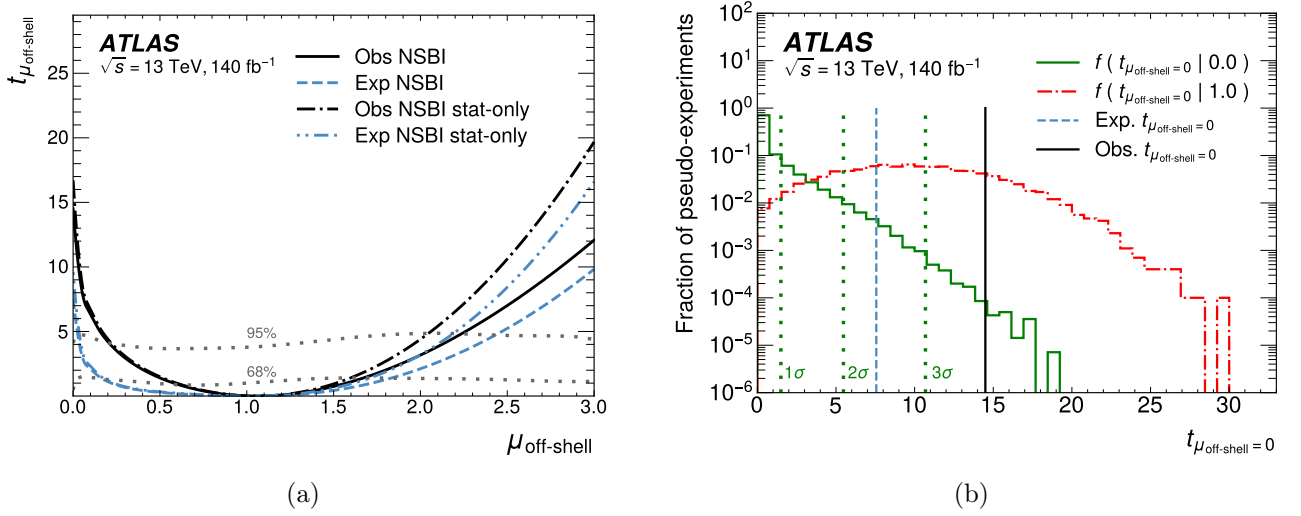


Figure 7.5: a Values of the test statistic $t_{\mu_{\text{off-shell}}}$ assuming a single parameter of interest $t_{\mu_{\text{off-shell}}}$ obtained with an Asimov dataset (expected, dashed blue) and with data (observed, solid black) combining the $H^* \rightarrow ZZ \rightarrow 4\ell$ and $H^* \rightarrow ZZ \rightarrow 2\ell 2\nu$ decay channels. The dash-dotted curves show the *statistics-only* results where all NP are fixed to their best-fit values $\hat{\alpha}$. The dotted gray lines show the 68% and 95% confidence belt, obtained from the Neyman construction. b Expected distribution of $t_{\mu_{\text{off-shell}}=0}$ estimated with pseudo-experiments for the case of $\mu_{\text{off-shell}}^{\text{truth}} = 0$ (solid green, no off-shell Higgs boson production hypothesis) and $\mu_{\text{off-shell}}^{\text{truth}} = 1$ (dashed-dotted red, SM hypothesis). The vertical solid black (dashed blue) line shows the observed (expected) value of $t_{\mu_{\text{off-shell}}=0}$. The vertical dotted green lines represent the 1σ , 2σ , and 3σ significance thresholds under the $\mu_{\text{off-shell}}^{\text{truth}} = 0$ hypothesis.

$$\begin{aligned}
 -2 \log \frac{L(\kappa_g, \kappa_V, \theta_{qq}, \boldsymbol{\alpha})}{\prod_k^{\mathcal{N}_{\text{obs}}} p_{\text{ref}}(\mathbf{x})} &= \sum_{\text{bin} \in n_{\text{jets}}} \left[-2 \mathcal{N}_{\text{CR}}^{\text{bin}} \cdot \log \nu_{\text{CR}}^{\text{bin}}(\kappa_g, \kappa_V, \theta_{qq}, \boldsymbol{\alpha}) + 2 \nu_{\text{CR}}^{\text{bin}}(\kappa_g, \kappa_V, \theta_{qq}, \boldsymbol{\alpha}) \right] \\
 &\quad - 2 \mathcal{N}_{\text{data}}^{\text{SR}} \cdot \log \nu_{\text{SR}}(\kappa_g, \kappa_V, \theta_{qq}, \boldsymbol{\alpha}) + 2 \nu_{\text{SR}}(\kappa_g, \kappa_V, \theta_{qq}, \boldsymbol{\alpha}) \\
 &\quad - 2 \sum_i^{\mathcal{N}_{\text{data}}} \log \frac{p(\mathbf{x}_i | \kappa_g, \kappa_V, \theta_{qq}, \boldsymbol{\alpha})}{p_{\text{ref}}(\mathbf{x})} + \sum_m^{\mathcal{N}_{\text{syst}}} \alpha_m^2
 \end{aligned} \tag{7.4}$$

where $\kappa_{g/V} = \kappa_{g/V, \text{off-shell}}$. Fig. 7.6 shows the values of the test statistic $t_{\kappa_{g, \text{off-shell}}}$ and $t_{\kappa_{V, \text{off-shell}}}$ as a function of $\kappa_{g, \text{off-shell}}$ and $\kappa_{V, \text{off-shell}}$, respectively. In both cases, the κ parameter not shown is profiled. The resulting confidence intervals in $\kappa_{g, \text{off-shell}}$ and $\kappa_{V, \text{off-shell}}$ provide a measurement of the Higgs boson couplings to gluons and vector bosons without any assumption on the Higgs boson total width [125]. The observed (expected) values at 68% CL are:

$$\kappa_{g, \text{off-shell}} = 1.09^{+0.39}_{-0.35} \quad (1.00^{+0.76}_{-0.89}), \quad \kappa_{V, \text{off-shell}} = 0.99^{+0.16}_{-0.19} \quad (1.00^{+0.29}_{-0.45}).$$

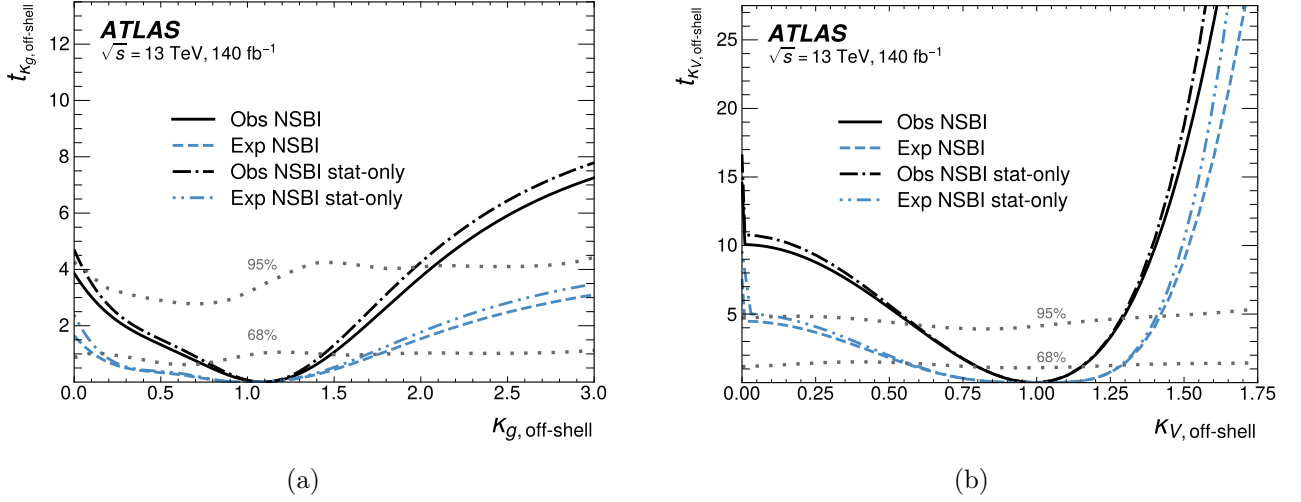


Figure 7.6: Values of the a test statistic $t_{\kappa_{g,\text{off-shell}}}$ as a function of $\kappa_{g,\text{off-shell}}$ and b the test statistic $t_{\kappa_{V,\text{off-shell}}}$ as a function of $\kappa_{V,\text{off-shell}}$ obtained with an Asimov dataset (expected, dashed blue) and with data (observed, solid black). The dash-dotted curves show the *statistics-only* results where all NP are fixed to their best-fit values $\hat{\alpha}$. The dotted gray lines show the 68% and 95% confidence belt, obtained from the Neyman construction. The κ parameter not shown is profiled in both cases. The abrupt change at $\kappa_{V,\text{off-shell}} = 0$ comes from the strong $\kappa_{V,\text{off-shell}}^4$ dependency in the EW production of off-shell Higgs bosons.

7.4 Combination with on-shell 4ℓ analysis and Higgs boson width interpretation

Measurements of the on-shell Higgs boson production [126] provide constraints on the signal strengths $\mu_{\text{on-shell}}^{\text{ggF}} = \kappa_{g,\text{on-shell}}^2 \kappa_{V,\text{on-shell}}^2 / \kappa_H$ and $\mu_{\text{on-shell}}^{\text{EW}} = \kappa_{V,\text{on-shell}}^4 / \kappa_H$, where $\kappa_H = \Gamma_H / \Gamma_H^{\text{SM}}$ is the Higgs boson width normalized to the SM expectation. Recall that

$$\frac{\Gamma_H}{\Gamma_H^{\text{SM}}} = \frac{\mu_{\text{off-shell}}}{\mu_{\text{on-shell}}} \quad (7.5)$$

assuming that $\kappa_{g,\text{on-shell}}^2 \kappa_{V,\text{on-shell}}^2 = \kappa_{V,\text{on-shell}}^4 = \kappa_{g,\text{off-shell}}^2 \kappa_{V,\text{off-shell}}^2 = \kappa_{V,\text{off-shell}}^4$, the off-shell Higgs boson production measurement can be combined with the on-shell $H \rightarrow ZZ \rightarrow 4\ell$ production measurement [126] to provide a measurement of the Higgs boson total width. The joint likelihood model for this measurement extends Eq. 7.2 with a common $H \rightarrow ZZ$ coupling modifier θ_{HZZ} and the modifier κ_H to the Higgs boson width:

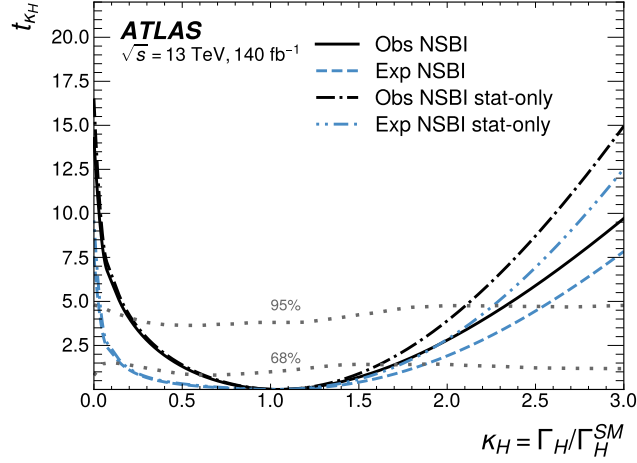


Figure 7.7: Values of the test statistic t_{κ_H} as a function of $\kappa_H = \Gamma_H / \Gamma_H^{SM}$ obtained with an Asimov dataset (expected, dotted black) and with data (observed, solid black). The dash-dotted curves show the *statistics-only* results where all NP are fixed to their best-fit values $\hat{\alpha}$. The dotted gray lines show the 68% and 95% confidence belt, obtained from the Neyman construction.

$$\begin{aligned}
 -2 \log \frac{L(\kappa_H, \theta_{HZZ}, \theta_{qq}, \boldsymbol{\alpha})}{\prod_k^{\mathcal{N}_{\text{data}}} p_{\text{ref}}(\mathbf{x})} &= -2 \sum_{\text{on-shell bins}} \left[\mathcal{N}_{\text{on-shell}}^{\text{bin}} \cdot \log \nu_{\text{on-shell}}^{\text{bin}}(\theta_{HZZ} / \kappa_H, \boldsymbol{\alpha}) + 2 \nu_{\text{on-shell}}^{\text{bin}}(\theta_{HZZ} / \kappa_H, \boldsymbol{\alpha}) \right] \\
 &\quad -2 \sum_{\text{off-shell bins}} \left[\mathcal{N}_{\text{off-shell}}^{\text{bin}} \cdot \log \nu_{\text{off-shell}}^{\text{bin}}(\theta_{HZZ}, \theta_{qq}, \boldsymbol{\alpha}) + 2 \nu_{\text{off-shell}}^{\text{bin}}(\theta_{HZZ}, \theta_{qq}, \boldsymbol{\alpha}) \right] \\
 &\quad -2 \sum_i^{\mathcal{N}_{\text{data}}} \log \frac{p(\mathbf{x}_i | \theta_{HZZ}, \theta_{qq}, \boldsymbol{\alpha})}{p_{\text{ref}}(\mathbf{x})} + \sum_m^{N_{\text{syst}}} \alpha_m^2
 \end{aligned} \tag{7.6}$$

where $\theta_{HZZ} = \kappa_{g,\text{on-shell}}^2 \kappa_{V,\text{on-shell}}^2 = \kappa_{V,\text{on-shell}}^4 = \kappa_{g,\text{off-shell}}^2 \kappa_{V,\text{off-shell}}^2 = \kappa_{V,\text{off-shell}}^4$ and the corresponding off-shell Higgs boson model is defined in Eq. 4.20. The off-shell bins in Eq. 7.6 includes both the $ZZ \rightarrow 4\ell$ and $ZZ \rightarrow 2\ell 2\nu$ SR and CR bins. The on-shell model uses the Standard Template Cross-Section (STXS) prescription, defined in Ref. [126] and will not be reviewed here. The NP model for common experimental uncertainties follows the approach described in Sec. 7.2. Theoretical modeling uncertainties in the on-shell and off-shell measurements are modeled with separate parameters, given the distinctness of the phase-space regions, but the measured Higgs boson width is largely insensitive to this modeling choice. Background normalization factors are also modeled separately.

Fig. 7.7 shows the test statistic values as a function of κ_H when profiling θ_{HZZ} . The observed (expected) value of κ_H and Γ_H at 68% CL are:

$$\kappa_H = 1.05_{-0.46}^{+0.65} \quad (1.00_{-0.84}^{+0.86}), \quad \Gamma_H = 4.3_{-1.9}^{+2.7} \quad (4.1_{-3.4}^{+3.5}),$$

The result using uncertainties at 95% CL can be found in Tab. 7.2.

A similar combined measurement strategy can be used to constrain $R_{gg} = \kappa_{g,\text{on-shell}}^2 / \kappa_{g,\text{off-shell}}^2$ and

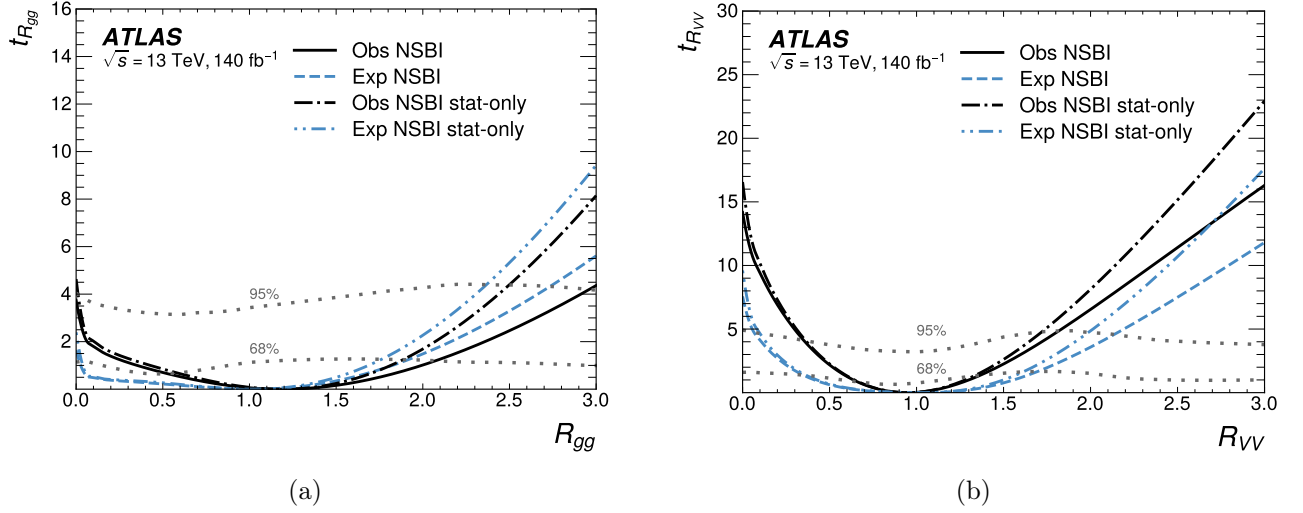


Figure 7.8: Values of a the test statistic $t_{R_{gg}}$ as a function of $R_{gg} = \kappa_{g,\text{on-shell}}^2 / \kappa_{g,\text{off-shell}}^2$ and of b the test statistic $t_{R_{VV}}$ as a function of $R_{VV} = \kappa_{V,\text{on-shell}}^2 / \kappa_{V,\text{off-shell}}^2$ obtained with an Asimov dataset (expected, dashed blue) and with data (observed, solid black). The values obtained when all NP are fixed to their best-fit values $\hat{\alpha}$, corresponding to the *statistics-only* case, are shown in dash-dotted lines for comparison. The dotted gray lines show the 68% and 95% confidence belt, obtained from the Neyman construction. In both figures, the R parameter not shown is profiled and κ_H is fixed to 1.

$R_{VV} = \kappa_{V,\text{on-shell}}^2 / \kappa_{V,\text{off-shell}}^2$. The value of the two test statistics as a function of R_{gg} and R_{VV} are shown in Fig. 7.8. The observed (expected) values at 68% CL are:

$$R_{gg} = 1.19^{+0.89}_{-0.66} \quad (1.00^{+0.92}_{-0.98}), \quad R_{VV} = 0.95^{+0.44}_{-0.35} \quad (1.00^{+0.70}_{-0.69}),$$

profiling the other R parameter and setting κ_H to unity. Tab. 7.2 summarizes all the results presented in this paper.

Table 7.2: Summary of the results for $\kappa_{g, \text{ off-shell}}$, $\kappa_{V, \text{ off-shell}}$, and $\mu_{\text{ off-shell}}$ combining the measurement of the off-shell Higgs boson production analysis in the $H \rightarrow ZZ \rightarrow 4\ell$ decay channel presented in this paper with the analysis in the $H \rightarrow ZZ \rightarrow 2\ell 2\nu$ decay channel from Ref. [3]. The results for Γ_H , R_{gg} and R_{VV} are obtained combining the off-shell production measurement with the on-shell analysis from Ref. [126]. All results are presented with their 68% CL and 95% CL intervals. No expected 95% CL interval for $\kappa_{g, \text{ off-shell}}$ is obtained because the Asimov sample is unable to break the degeneracy between $\kappa_{g, \text{ off-shell}}$ and $\kappa_{V, \text{ off-shell}}$ in ggF production. All results use the full Run 2 dataset with 140 fb^{-1} of integrated luminosity.

Parameter	Value	68% CL interval		95% CL interval	
		Observed	Expected	Observed	Expected
$\mu_{\text{ off-shell}}$	1.06	[0.61, 1.67]	[0.17, 1.83]	[0.21, 2.24]	[0.01, 2.42]
$\kappa_{g, \text{ off-shell}}$	1.09	[0.74, 1.48]	[0.11, 1.76]	< 2.08	-
$\kappa_{V, \text{ off-shell}}$	0.99	[0.80, 1.15]	[0.55, 1.29]	[0.58, 1.30]	[0.01, 1.42]
Γ_H [MeV]	4.29	[2.41, 6.95]	[0.66, 7.61]	[0.76, 9.66]	[0.12, 10.50]
R_{gg}	1.19	[0.53, 2.07]	[0.02, 1.92]	< 2.96	< 2.73
R_{VV}	0.95	[0.61, 1.39]	[0.31, 1.70]	[0.30, 1.86]	[0.06, 2.14]

Chapter 8

Computational Challenges and Alternatives

This chapter highlights the computational challenges associated with a full NSBI analysis and demonstrates the workflow developed and used in the off-shell Higgs boson analysis. Later, an alternative of using parameterized optimal observables in lieu of a full unbinned analysis is presented, particularly useful when computational challenges are a limitation. The sensitivity gain from the new analysis is demonstrated to be replicated using the parameterized optimal observable techniques in certain cases.

The prediction of high quality estimates of the likelihood ratios using real experimental data requires an ensemble of deep and wide NNs as discussed in Chap. 6. Fig. 8.1a shows an example of the NN architecture used in the off-shell Higgs boson analysis to model density ratios with minimal bias and variance.

8.1 Computational challenges in NN training

Training these deep and wide NNs using CPUs can be inefficient. As shown in Fig. 8.1b, GPUs offer a significant speed-up compared to CPU for training the same NN using $O(1M)$ simulated events. On top of the ensemble of NNs predicting nominal density ratios p_c/p_{ref} for each of the channels c , additional NNs are trained to accurately estimate the variations associated with the $O(100)$ nuisance parameters needed to model the full set of systematic uncertainties in a typical ATLAS analysis. A schematic of the full workflow is shown in Fig. 8.2. The total number of NNs needed to estimate the fully parameterized likelihood ratios with sufficient accuracy and precision is on the order of many thousands, **making large scale GPU infrastructure essential for an SBI analysis for an NSBI analysis in a real experimental setting.**

8.2 Computational challenges in statistical inference

Once the density ratios $p(\mathbf{x}|\mu_{\text{off-shell}}, \boldsymbol{\alpha})/p_{\text{ref}}(\mathbf{x})$ have been calculated using the thousands of NNs, the final step of an analysis is to compute the best-fit values and uncertainties of model parame-

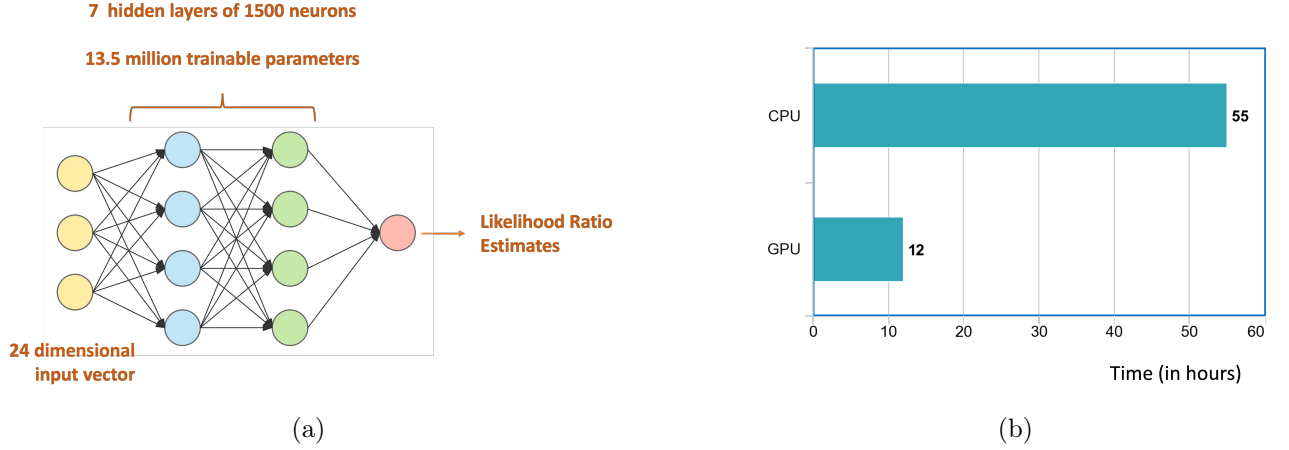


Figure 8.1: (a) NN architecture of one of the many thousands of NNs used in the analysis, each with millions of trainable parameters. (b) Training time comparison, between a CPU and an NVIDIA T4 GPU, for one of the NNs used in the analysis. Training on GPU offers almost 5 $\times$ the speed-up compared to the impractically large training time on CPU, with the same $O(1M)$ simulated events used for each training. This time difference becomes especially relevant when optimizing an ensemble of thousands of NNs.

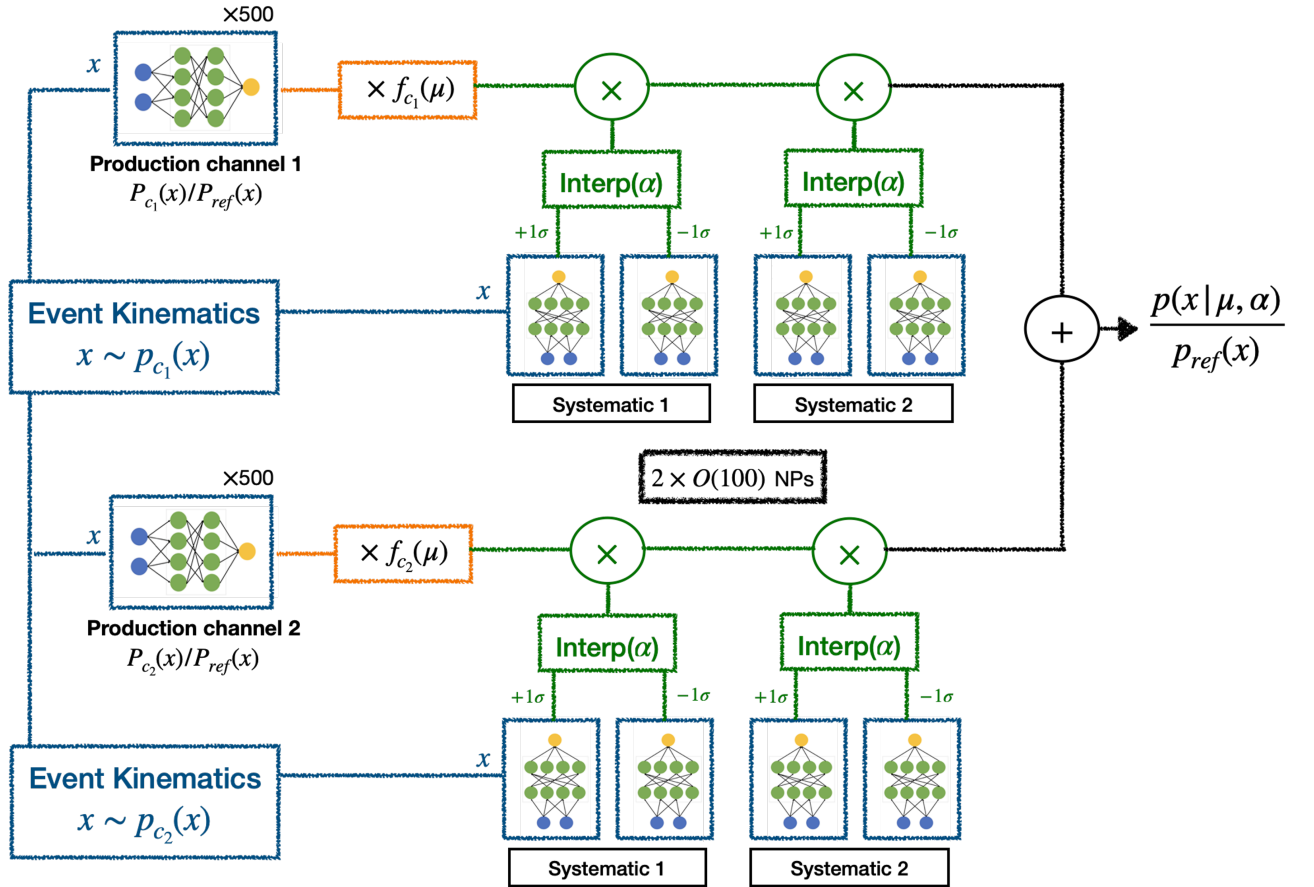


Figure 8.2: A cartoon schematic of the full NSBI workflow developed in this thesis. For brevity only two processes c_1 and c_2 are shown along with two systematics. Each channel uses a large ensemble of NNs to estimate the ratios $p_{c_1}(\mathbf{x})/p_{ref}(\mathbf{x})$. The $Interp$ function stands for the polynomial-exponential interpolation defined in Chap. 3.

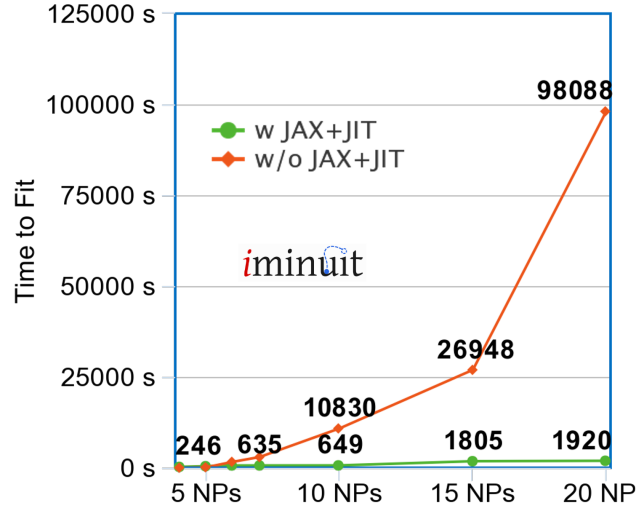


Figure 8.3: Comparison of the time, in seconds, to perform a profile likelihood fit using the NSBI test statistic $t_{\mu_{\text{off-shell}}}$ written as a JIT-compiled JAX function and as a default python function. The fit is performed using event-by-event systematic variations for an Asimov dataset with $O(10M)$ events. While the computation time remains similar between the python function and the JIT-compiled JAX function for a small number of NPs, the latter scales significantly better as the number of NPs in the fit increases.

ters. This is done for example in the off-shell Higgs boson analysis using the log-likelihood model $\lambda(\mu_{\text{off-shell}}, \theta_{qq}, \alpha) = -2 \log \left[L(\mu_{\text{off-shell}}, \theta_{qq}, \alpha) / \prod_k^{\mathcal{N}_{\text{obs}}} p_{\text{ref}}(\mathbf{x}_k) \right]$ defined in Eq. 7.2. With the NSBI analysis, the computation time to perform parameter fitting using $\lambda(\mu_{\text{off-shell}}, \theta_{qq}, \alpha)$ with the typical $O(100)$ NPs of the ATLAS experiment is significantly higher than that of a binned fit. This is because unlike a binned fit with parameterized Poisson probability ratios for number of bins ranging from a few 10s to 100s, in the NSBI analysis there are event-by-event parameterized density ratio predictions for events ranging from 1M to 10M in number. Subsequently, the new analysis makes use of modern tools like auto-differentiation and just-in-time (JIT) compilation using the JAX library [127], which offers the potential to reduce computation time by several orders of magnitude. A comparison of the time to find the best fit values of parameters using the Migrad algorithm of iminuit [128, 129] library is shown in Figure 8.3 for two Python functions - one using JIT compilation with corresponding JAX's lax control flow operations and one without any compiled function.

The new analysis makes use of the auto-differentiation tool offered by JAX to infer both the uncertainties on the NPs and the propagation of these uncertainties to the best fit value of the parameter of interest, by calculating the exact second derivative Hessian matrix of the likelihood model of Eq. 7.2, as discussed in Sec. 3.2. Materializing the full Hessian matrix in the NSBI analysis using JAX is unfeasible, since storing weights during the forward and/or reverse pass calculation of the second derivative $\nabla^2 \lambda : \mathbb{R}^{O(100)} \rightarrow \mathbb{R}^{O(100) \times O(100)}$ requires enormous memory resources owing to the event-by-event nature of the analysis and the $O(10M)$ MC events in a typical application. There exists a memory-efficient solution, which can then be executed in a parallel workflow using distributed cloud-based computing infrastructure.

To make the computation more memory-efficient, instead of directly materializing the full Hessian

matrix, the Hessian vector product (HVP) is calculated making use of the following identity:

$$\nabla^2 \lambda(\mathbf{x}) v = \nabla [\nabla \lambda(x) \cdot v] \quad (8.1)$$

The first step then is to calculate the Jacobian $\nabla \lambda : \mathbb{R}^{O(100)} \rightarrow \mathbb{R}^{O(100)}$, and take its dot product with some vector v . This step transforms the Jacobian into a scalar-valued function $\nabla \lambda \cdot v : \mathbb{R}^{O(100)} \rightarrow \mathbb{R}$. This reduces the required memory resources for the subsequent computation of the gradient of this scalar valued function $\nabla [\nabla \lambda \cdot v] : \mathbb{R}^{O(100)} \rightarrow \mathbb{R}^{O(100)}$ giving us $\nabla^2 \lambda(\mathbf{x}) v$.

Repeating the computations of $\nabla^2 \lambda(\mathbf{x}) v$ with the vector v chosen to be one of the $O(100)$ -dimensional orthogonal unit vectors each time, the Hessian matrix $\nabla^2 \lambda(\mathbf{x})$ is extracted one row at a time. By only computing the gradients of scalar valued functions, the memory requirements are significantly reduced. For NSBI analysis, this new solution makes the calculation of the Hessian matrix possible to perform in practice. However, the memory requirement can still add up to several hundred gigabytes, necessitating the use of specialized hardware with large memory resources. This is developed in Ref. [6].

8.3 An Alternative: Using parameterized observables

The impressive gain in sensitivity using NSBI techniques comes at a large computational cost as discussed extensively in this thesis. An unbinned analysis needs low-bias and low-variance estimates of probability density ratios, only possible using large ensembles of very deep and wide NNs. The training of these large ensembles needs large-scale GPU and high memory CPU infrastructure as discussed in Sec. 8.1 and Sec. 8.2. Moreover, training low-variance large NNs needs large MC statistics which might not be readily available. There are also analysis where the simulations are known to be not a perfect representation of the true data and hence training unbiased density ratios might be a challenge. Using binned parameterized optimal observables (OO) defined in Sec. 2.5.3 with Poisson probability models resolves a lot of these challenges while still gaining significantly in sensitivity compared to a more traditional approach. In order to demonstrate what role the OO play in an NSBI analysis, the next section compares sensitivities from a full unbinned NSBI analysis with a toy analysis using binned parameterized OO.

8.3.1 Understanding the Improvements in the New Analysis

As discussed in Chap. 2, the power of the new analysis comes from the parameterized nature of the density ratio estimation. This can be tested in a more traditional setting where the fit is done using binned Poisson likelihoods from a parameterized version of the usual summary statistic, here referred to as the parameterized OO technique and defined in Sec. 2.5.3. The idea is to build a powerful parameter-dependent summary statistic for each value of the scanned POI μ .

The per-event NSBI analysis as described in this thesis naturally learns the density ratios as a function of the observable $\mu_{\text{off-shell}}$. The result is that we have essentially a $\mu_{\text{off-shell}}$ -parameterized OO¹. Moreover, it is by construction an unbinned analysis - the NNs learn the density ratios as a

¹Note that we don't really use an observable per-se in the NSBI approach presented here. We directly learn the density

smooth differentiable function across the phase space of the input.

In this section, toy studies are performed and it is shown how a combination of the $\mu_{\text{off-shell}}$ -parameterized nature of the density ratios combined with the unbinned effects are responsible for the significant improvements in the sensitivity of the analysis. The OO for a given value of the off-shell signal strength $\mu_{\text{off-shell}} = \mu$ is given by:

$$u(\mathbf{x}|\mu_{\text{off-shell}}) = \frac{p(\mathbf{x}|\mu_{\text{off-shell}})}{p(\mathbf{x}|\mu_{\text{off-shell}} = 1)} \quad (8.2)$$

Note the $\mu_{\text{off-shell}}$ -dependence of the summary statistic, which is not commonly used in LHC analyses. The density ratios on the RHS are already estimated using the NSBI formalism. The summary statistic can now be binned and likelihood extracted using Poisson probability model as in Eq. 2.38.

$$p(\mathcal{D}|\mu_{\text{off-shell}}) = \text{Pois}(\mathcal{N}_{\text{data}}|\nu(\mu_{\text{off-shell}})) \prod_i \frac{1}{\nu(\mu_{\text{off-shell}})} \frac{\nu_{I(\mathbf{x}_i|u(\mu_{\text{off-shell}}))}(\mu_{\text{off-shell}})}{\Delta_{I(\mathbf{x}_i|u(\mu_{\text{off-shell}}))}} \quad (8.3)$$

Moreover, to approach the sensitivity of the unbinned per-event analysis, a sequential un-binning of $u(\mathbf{x}|\mu_{\text{off-shell}})$ is performed. An example OO discriminant for $\mu = 0.2$ with a large number of bins is shown in Fig. 8.4.

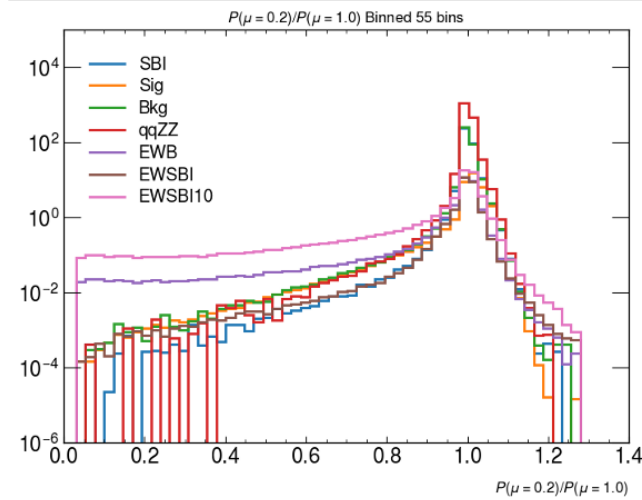


Figure 8.4: The OO discriminant $u(\mathbf{x}|\mu_{\text{off-shell}} = 0.2)$ with 55 bins, corresponding to a sufficient summary statistic for $\mu_{\text{off-shell}} = 0.2$. Poisson likelihoods are extracted from each bin and are used to do the t_μ scans shown in Fig. 8.6.

A comparison between the sensitivity from the new per-event analysis and the sensitivity from these sequential un-binning of the μ -parametrized toy experiments is shown in Fig. 8.5, highlighting how the increased sensitivity in the new analysis is a result of its $\mu_{\text{off-shell}}$ -parametrized nature combined with its un-binned nature.

It might not be clear to the reader how the floating dots fit into Fig. 8.5. To clarify this, Fig. 8.6 shows how these dots are just the value of the t_μ for the value of $\mu_{\text{off-shell}}$ used in the OO. Also interesting ratios from the input. But it serves as a good analogy.

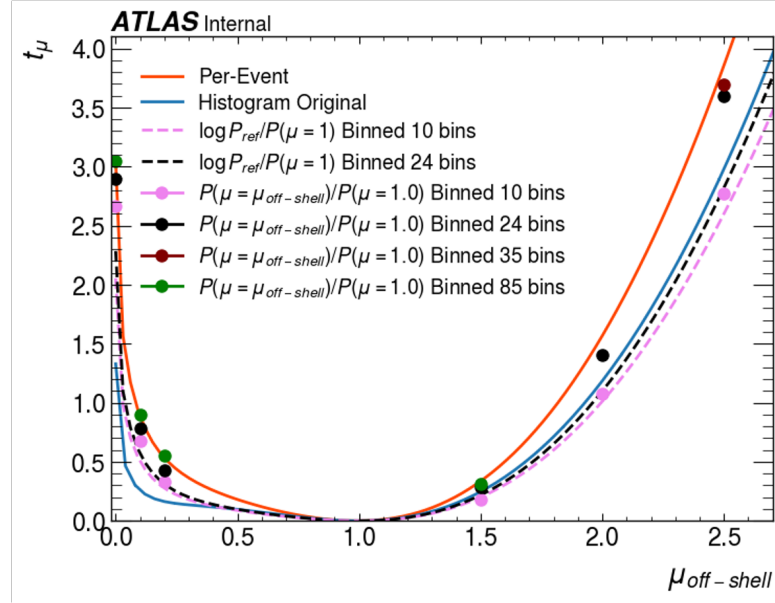


Figure 8.5: The plot shows a comparison of expected stat-only sensitivity from various analysis strategies which can potentially be used in this channel. The dashed pink (black) curves show the expected t_μ scans from an analysis which uses $\log p(\mathbf{x}|\mu_{\text{off-shell}} = 1)/p_{\text{ref}}$ as a discriminant with 10 (24) bins. The various dots show the analysis sensitivity for specific values of the POI $\mu_{\text{off-shell}}$ ($= 0.0, 0.1, 0.2, 1.5, 2.0, 2.5$ in this plot), where the discriminant $u(\mathbf{x}|\mu_{\text{off-shell}})$ is used as described in Eq. 8.2. Important takeaway from this plot is that as the number of bins in the OO discriminant approach large values, the sensitivity approaches that of the per-event analysis. Also interesting to note is that having an OO even with fewer number of bins still has better sensitivity than the $\log p(\mathbf{x}|\mu_{\text{off-shell}} = 1)/p_{\text{ref}}(\mathbf{x})$ observable optimized only for $\mu_{\text{off-shell}} = 1$.

to note in Fig. 8.6 is how the OO optimized for a given value $u(\mathbf{x}|\mu'_{\text{off-shell}})$ offers increased sensitivity for that particular value of $\mu_{\text{off-shell}} = \mu'_{\text{off-shell}}$ while showing sub-optimal performance for other values $\mu_{\text{off-shell}} \neq \mu'_{\text{off-shell}}$ that $u(\mathbf{x}|\mu'_{\text{off-shell}})$ is not optimized for. This is similar to the argument mentioned above, about why having a single summary statistic observable, built using a fixed value of $\mu_{\text{off-shell}} = 1$ for example, can be sub-optimal for other values of $\mu_{\text{off-shell}} \neq 1$, especially when interference effects are non-negligible.

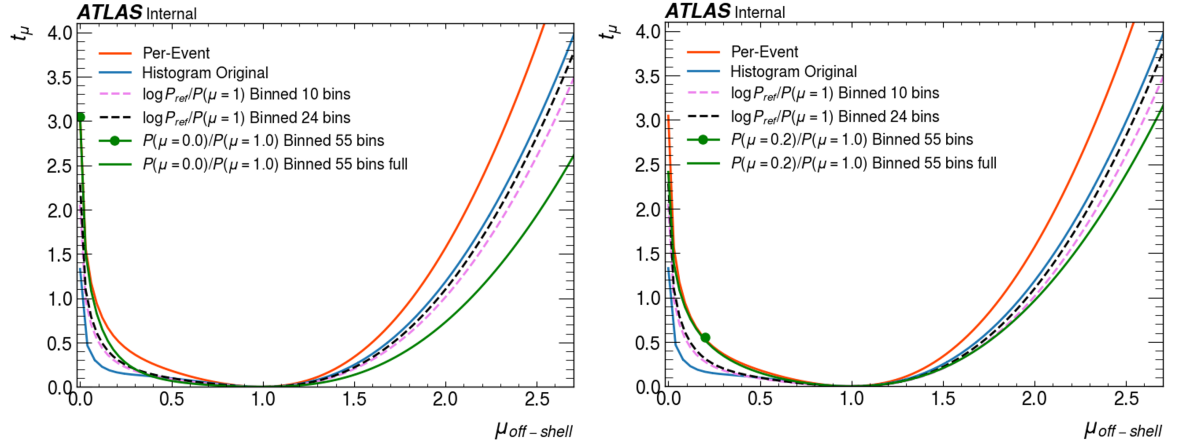


Figure 8.6: The floating dots in the Fig. 8.5 are actually just part of the full stat-only t_μ curve with the $u(\mathbf{x}|\mu)$ discriminant optimized for that particular value of μ . Here we plot the full NLL curve with the dot to clarify the situation. Note how the OO optimized for $\mu_{\text{scan}} = 0.0$ offers increased sensitivity for $\mu = 0.0$ while showing sub-optimal performance for other values of $\mu_{\text{off-shell}}$, for example $\mu_{\text{off-shell}} = 0.2$ which has increased sensitivity when the OO is optimized for $\mu_{\text{scan}} = 0.2$.

The parameterized density ratios trained for an NSBI analysis, when used as a parameter-dependent binned summary statistic $u(\mathbf{x}|\mu)$, can still be asymptotically optimal for the full parameter range being scanned. While not as sensitive as a fully unbinned analysis, this can be a viable alternative when training unbiased density ratios is a challenging problem - since the summary statistic is to be used for a Poisson likelihood fit, which has well-defined statistical properties by construction. The so-called parameterized Optimal Observables method, described in Sec. 2.5.3 can provide significant improvement in sensitivity compared to a more traditional fixed summary statistic.

8.3.2 Application to Real Data

The previous section demonstrates the use of parameterized observables using expected t_μ scans. To demonstrate the effectiveness of this approach in a real analysis, Fig. 8.7 shows both the observed and expected t_μ scans, using a binned Poisson likelihood fit done with the OO optimized at $\mu_{\text{off-shell}} = 0$. It is easy to see how the OO sensitivity approaches that of the NSBI result in the vicinity of $\mu_{\text{off-shell}} = 0$, for both the expected and observed scans. Moving away from $\mu_{\text{off-shell}} = 0$ sees a drop in sensitivity, highlighting the importance of parameterized OO for cases with degenerate solutions. This figure also helps demystify the data fluctuations in the unbinned NSBI fit results, highlighting how a Poisson likelihood fit using the $u(\mathbf{x}|0)$ observable replicates the same fluctuation effect near $\mu_{\text{off-shell}} = 0$.

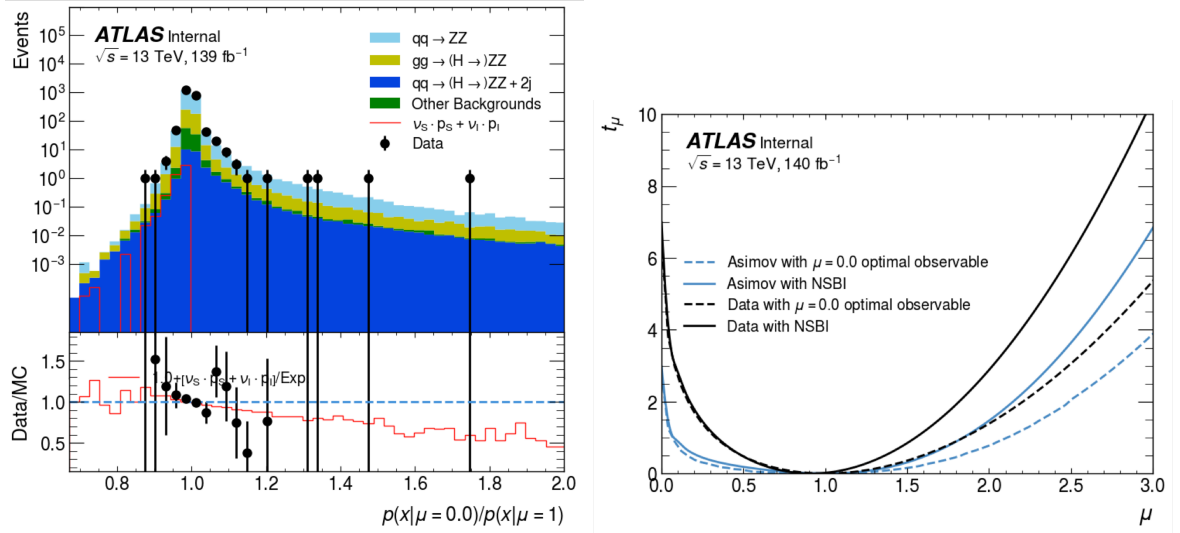


Figure 8.7: A stat-only binned Poisson likelihood fit using real ATLAS experiment data, with the OO discriminant optimized for $\mu_{\text{off-shell}} = 0$. The left plot shows the discriminant $u(\mathbf{x}|\mu_{\text{off-shell}} = 0)$ modeled with 200 bins, and the right plot shows the expected and observed t_μ scans using the binned observable. The right plot also shows comparisons to the result from the NSBI technique, described in Chap. 7. Notably, the locally optimal $u(\mathbf{x}|\mu_{\text{off-shell}} = 0)$ discriminant replicates the sensitivity and data fluctuation in the vicinity of $\mu_{\text{off-shell}} = 0$, while being relatively sub-optimal for $\mu_{\text{off-shell}} \gg 0$. This highlights the importance of parameterized OO in an analysis with degenerate solutions.

Chapter 9

Conclusion and Outlook

A practical formalism for parameter inference using Neural Simulation-Based Inference (NSBI) techniques has been developed for the ATLAS experiment, with a consistent and computationally effective prescription for the propagation of a large systematic uncertainty model - enabling easy application to LHC experiments. A novel technique for reporting robust frequentist confidence intervals using Neyman Construction in the new NSBI formalism has also been developed. The full formalism presented in this thesis can in principle be applied to every parameter inference analysis at the LHC, provided that an analytical form of the parametric model is available as discussed in Chap. 2.

Analyses at the LHC with non-linear probability models such as those with large quantum interference (off-shell Higgs boson, top-quark pair production, Effective Field Theory operators) and with non-analytically known latent-variable dependence such as those measuring fundamental particle masses (top quark mass) can gain significant sensitivity from using NSBI techniques due to the parameterized nature of the observables used in hypothesis testing. Searches for rare signals subdued by large background coming from many different sources should also see large improvements due to the unbinned nature of the analysis. The formalism can also be applied to analyses where simulations are challenging and have known biases, but will need additional R&D efforts to integrate.

A new measurement of the off-shell Higgs boson production in the $H^* \rightarrow ZZ \rightarrow 4\ell$ decay channel using NSBI techniques is also presented. The measurement uses 140 fb^{-1} of integrated luminosity collected at $\sqrt{s} = 13 \text{ TeV}$ during the Run 2 of the LHC by the ATLAS detector. This new result is also combined with the most recent published measurement of the off-shell Higgs boson production in the $H^* \rightarrow ZZ \rightarrow 2\ell 2\nu$ decay channel to set powerful constraints on the off-shell signal strength using Run 2 data. The observed (expected) value of the off-shell Higgs boson signal strength is found to be $1.06^{+0.62}_{-0.45}$ ($1.00^{+0.83}_{-0.83}$) at 68% CL. The evidence for off-shell Higgs boson production has an observed (expected) significance of 3.7σ (2.4σ). The off-shell Higgs boson production measurement in the $H^* \rightarrow ZZ$ decay channel is combined with the on-shell Higgs boson production measurement in the same channel to obtain a constraint on the Higgs boson total width. The observed (expected) value of the Higgs boson total width is $\Gamma_H = 4.3^{+2.7}_{-1.9}(4.1^{+3.5}_{-3.4}) \text{ MeV}$ at 68% CL, improved relative to the ATLAS previous result of $\Gamma_H = 4.4^{+3.1}_{-2.3}(4.1^{+3.8}_{-3.8}) \text{ MeV}$ using the same dataset.

The training and optimization of the $O(10^3)$ NNs in this type of analysis can be computationally expensive. This work was performed as part of the ATLAS-Google collaboration using the Google

Cloud Platform (GCP) as discussed in [6]. Large parts of this workflow have also been done using a high-performance computing (HPC) system with large-scale GPU infrastructure at Southern Methodist University. In the absence of large scale computing infrastructures or theoretically accurate and high-statistics Monte Carlo simulations, a computationally cheaper but powerful technique of parameterized Optimal Observables is suggested in Chap. 8 that retains a lot of the optimality of a full NSBI analysis without the constraints of training strictly low-bias and low-variance density ratios.

Appendix A

Histogram Result Comparisons

This appendix provides a comparison of the results presented in this paper with the results in Ref. [3], which uses a histogram-based analysis using the fixed summary statistic of Eq. 4.29. The results of Ref. [3] used in this comparison have been updated to use the most recent measurement of integrated luminosity of 140 fb^{-1} [79].

Fig. 7.1a provides a comparison of the values of the test statistic $t_{\mu_{\text{off-shell}}}$ assuming a single parameter of interest $\mu_{\text{off-shell}}$ between the NSBI result and the histogram-based hypothesis using only the $H^* \rightarrow ZZ \rightarrow 4\ell$ decay channel. Fig. A.1 shows the same comparison after combining the result in the $H^* \rightarrow ZZ \rightarrow 2\ell 2\nu$ channel. Tab. A.1 compares the impact of different groups of systematic uncertainties in the combined result by using the method described in [3]. The impact of a group of systematic uncertainties is assessed by fixing the associated group of NPs and re-estimating the value of $\mu_{\text{off-shell}}$ for which the value of the test statistic is equal to four, corresponding to 95.5% CL in the asymptotic approximation. This method highlights the importance of different sources of uncertainties in the signal-dominated region. The importance of an uncertainty is given by its deviation from the last row, which has the uncertainty for the full result. The most important sources of systematic uncertainty are common for both analyses and their impacts on $\mu_{\text{off-shell}}$ are similar.

Fig. A.2a shows a comparison between the two analyses for the values of the test statistic for the combination of the off-shell and on-shell Higgs boson production measurements for the $H \rightarrow ZZ \rightarrow 4\ell$ decay channel alone as a function of κ_H while profiling θ_{HZZ} . Fig. A.2b shows the same comparison after combination with the result in the $H^* \rightarrow ZZ \rightarrow 2\ell 2\nu$ decay channel.

Fig. A.3a and A.3b compares the values of the test statistic for R_{gg} and R_{vv} with the results from the histogram-based analysis. Tab. A.2 summarizes the 68% and 95% CL results presented in this appendix, including $\mu_{\text{off-shell}}$ (4ℓ only). These are compared with the results of the previous measurement Ref. [3].

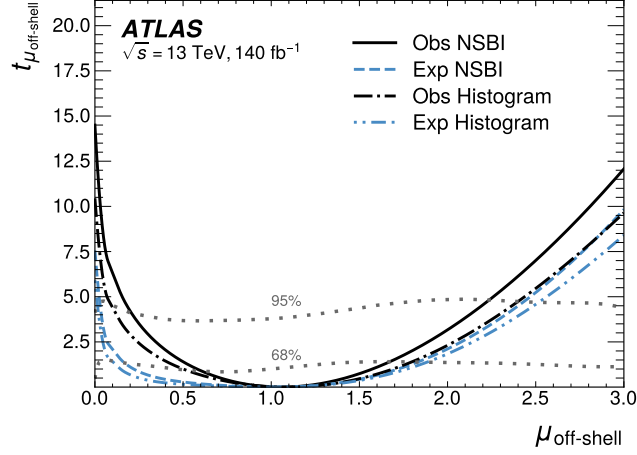


Figure A.1: Values of the test statistic $t_{\mu_{\text{off-shell}}}$ assuming a single parameter of interest $\mu_{\text{off-shell}}$ obtained with an Asimov dataset (expected, dashed blue) and with data (observed, solid black) when combining the $H^* \rightarrow ZZ \rightarrow 4\ell$ and $H^* \rightarrow ZZ \rightarrow 2\ell 2\nu$ decay channels. The values from the histogram-based analysis [3] are added with dash-dotted lines for comparison. The dotted gray lines show the 68% and 95% confidence belt, obtained from the Neyman construction, for the NSBI analysis.

Table A.1: The impact of most important systematic uncertainties in the observed upper value of $\mu_{\text{off-shell}}$ for which $t_{\mu_{\text{off-shell}}} = 4$, obtained by the combined fit. This value corresponds to the two standard deviation upper limit of $\mu_{\text{off-shell}}$ in the asymptotic approximation. The first column denotes the systematic uncertainty that was excluded from the fit. The last row gives the nominal upper limit, where all uncertainties are included. The importance of an uncertainty is given by its deviation from the last row.

Systematic uncertainty fixed	$\mu_{\text{off-shell}}$ value at which $t_{\mu_{\text{off-shell}}} = 4$	
	NSBI analysis	Histogram-based
All (stat-only)	1.96	2.13
Parton shower uncertainty for $gg \rightarrow ZZ$ (normalization)	2.07	2.26
Parton shower uncertainty for $gg \rightarrow ZZ$ (shape)	2.12	2.29
NLO EW uncertainty for $q\bar{q} \rightarrow ZZ$	2.10	2.27
NLO QCD uncertainty for $gg \rightarrow ZZ$	2.09	2.29
Parton shower uncertainty for $q\bar{q} \rightarrow ZZ$ (shape)	2.12	2.29
Jet energy scale and resolution uncertainty	2.11	2.26
None (full result)	2.12	2.30

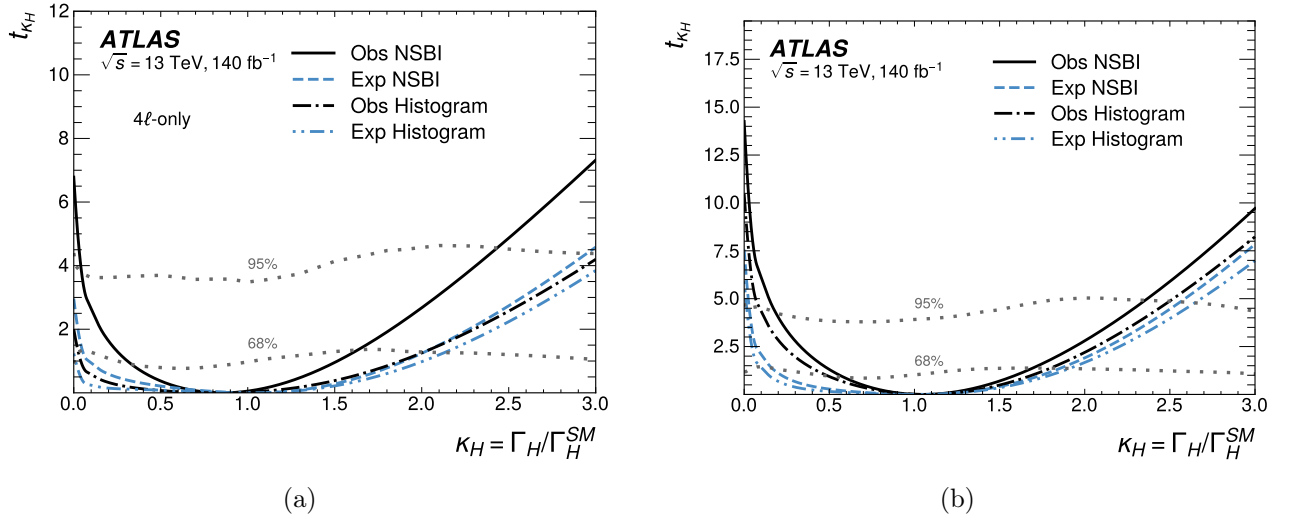


Figure A.2: a Values of the test statistic t_{κ_H} as a function of $\kappa_H = \Gamma_H / \Gamma_H^{SM}$ for the $H \rightarrow ZZ \rightarrow 4\ell$ channel obtained with an Asimov dataset (expected, dashed blue) and with data (observed, solid black). The values from the histogram-based analysis [3] are added with dash-dotted lines for comparison. The dotted gray lines show the 68% and 95% confidence belt, obtained from the Neyman construction, for the NSBI analysis. b Same comparison of the values of the test statistic for $\kappa_H = \Gamma_H / \Gamma_H^{SM}$ after combination with the $H^* \rightarrow ZZ \rightarrow 2\ell 2\nu$ channel.

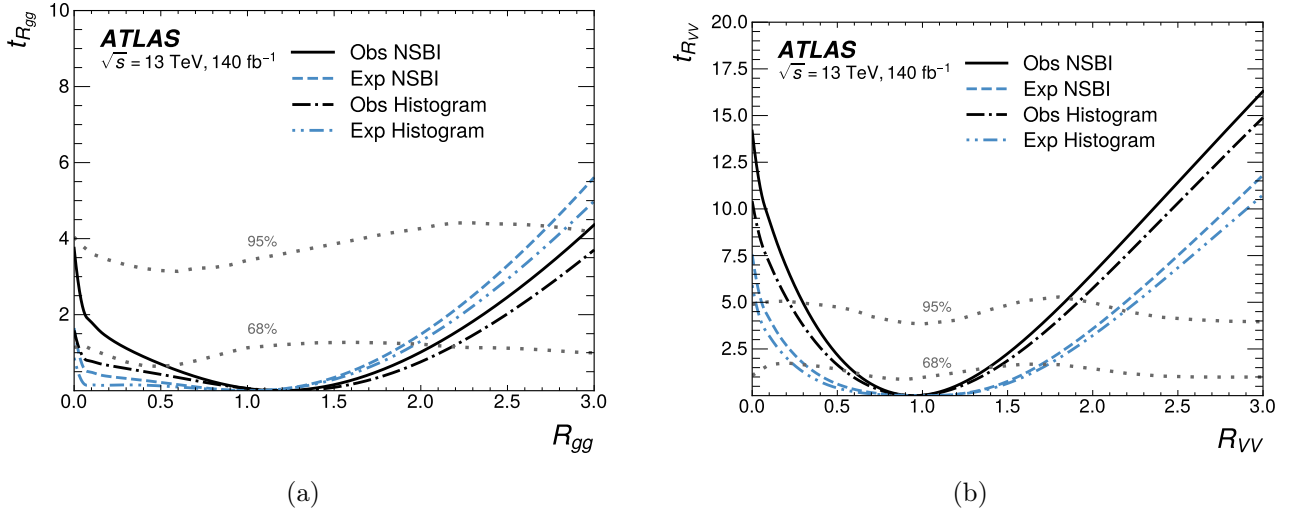


Figure A.3: Values of a the test statistic $t_{R_{gg}}$ as a function of R_{gg} , and of b the test statistic $t_{R_{VV}}$ as a function of R_{VV} for the combined $H \rightarrow ZZ \rightarrow 4\ell$ and $H^* \rightarrow ZZ \rightarrow 2\ell 2\nu$ channels obtained with an Asimov dataset (expected, dashed blue) and with data (observed, solid black). The equivalent graphs from the histogram-based analysis [3] are added with dash-dotted lines for comparison. The dotted gray lines show the 68% and 95% confidence belt, obtained from the Neyman construction, for the NSBI analysis.

Table A.2: Summary of the results for the NSBI analysis with a comparison to the corresponding results of the histogram analysis of Ref. [3]. The value of the parameter estimates and the 68% and 95% confidence intervals are provided. All results use the full Run 2 dataset with 140 fb^{-1} of integrated luminosity.

Parameter	Value	68% CL interval		95% CL interval	
		Observed	Expected	Observed	Expected
NSBI analysis					
$\mu_{\text{off-shell}} \text{ (}4\ell \text{ only)}$	0.87	[0.33, 1.62]	[0.05, 2.04]	[0.05, 2.38]	< 2.38
$\mu_{\text{off-shell}}$	1.06	[0.61, 1.67]	[0.17, 1.83]	[0.21, 2.24]	[0.01, 2.42]
$\Gamma_H \text{ [MeV] (}4\ell \text{ only)}$	3.43	[1.37, 6.71]	[0.20, 8.25]	[0.18, 9.98]	< 12.09
$\Gamma_H \text{ [MeV]}$	4.29	[2.41, 6.95]	[0.66, 7.61]	[0.76, 9.66]	[0.12, 10.50]
R_{gg}	1.19	[0.53, 2.07]	[0.02, 1.92]	< 2.96	< 2.73
R_{VV}	0.95	[0.61, 1.39]	[0.31, 1.70]	[0.30, 1.86]	[0.06, 2.14]
Histogram-based analysis					
$\mu_{\text{off-shell}} \text{ (}4\ell \text{ only)}$	0.79	[0.02, 2.00]	< 2.14	< 2.97	< 3.10
$\mu_{\text{off-shell}}$	1.09	[0.54, 1.81]	[0.08, 1.90]	[0.10, 2.41]	[0.01, 2.52]
$\Gamma_H \text{ [MeV] (}4\ell \text{ only)}$	3.43	[0.10, 8.42]	< 8.89	< 12.48	< 12.89
$\Gamma_H \text{ [MeV]}$	4.37	[2.13, 7.43]	[0.35, 7.94]	[0.39, 10.14]	< 10.79
R_{gg}	1.23	[0.00, 2.20]	< 1.98	< 3.15	< 2.84
R_{VV}	0.95	[0.60, 1.43]	[0.27, 1.74]	[0.26, 1.90]	[0.02, 2.18]

Appendix B

Neyman Construction Closures

In Sec. 3.3, a reweighting method was introduced to produce positive weighted Monte Carlo samples for performing bootstrap sampling using Poisson probability. In order for this method to work, it is necessary that the Asimov datasets resulting from Eq. 3.20 are un-biased representations of the Monte Carlo dataset at any value of μ . This is demonstrated by the reweighted diagnostic checks in Fig. 6.10. These verify that the re-weighted distributions of Eq. 3.20 match the original Monte Carlo sample for any given value of parameter μ . In Fig. B.1, the asymptotic properties of the re-weighted dataset are verified with that of the original Monte Carlo dataset by comparing the test statistic scans for various value of truth $\mu_{\text{off-shell}}$.

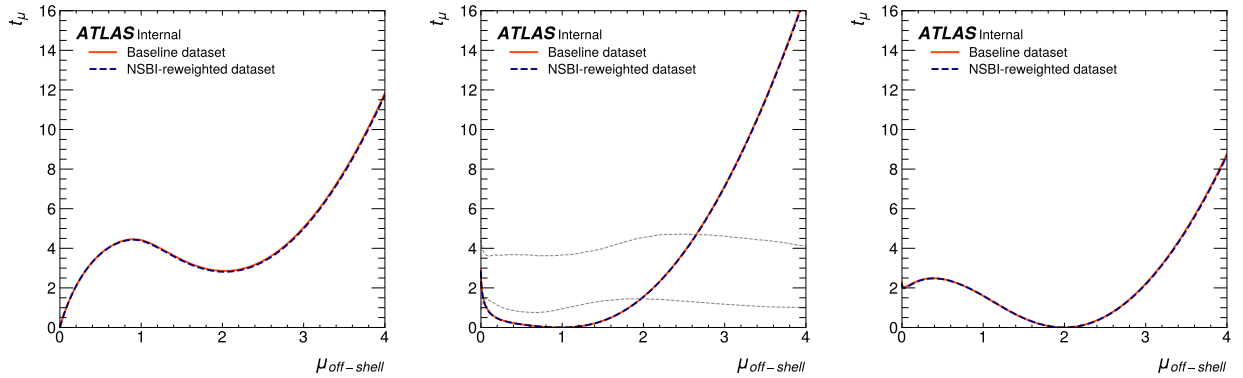


Figure B.1: Bootstrap Validation: the asymptotic properties of the re-weighted dataset are verified with that of the original Monte Carlo dataset by comparing the test statistic scans for $\mu_{\text{off-shell}} = 0, 1, 2$.

Fig. B.2 shows the distributions of $\hat{\mu}_{\text{off-shell}}$ from fitting many pseudo-experiments sampled using Poisson bootstrapping described in Sec. 3.3. Interesting to note that the estimators are in fact biased $\mu_{\text{truth}} \neq E[\hat{\mu}]$. This is not an artefact of the new NSBI analysis but that of the off-shell Higgs boson analysis due to the non-linear probability model combined with $\hat{\mu}_{\text{off-shell}} \geq 0$ constraint.

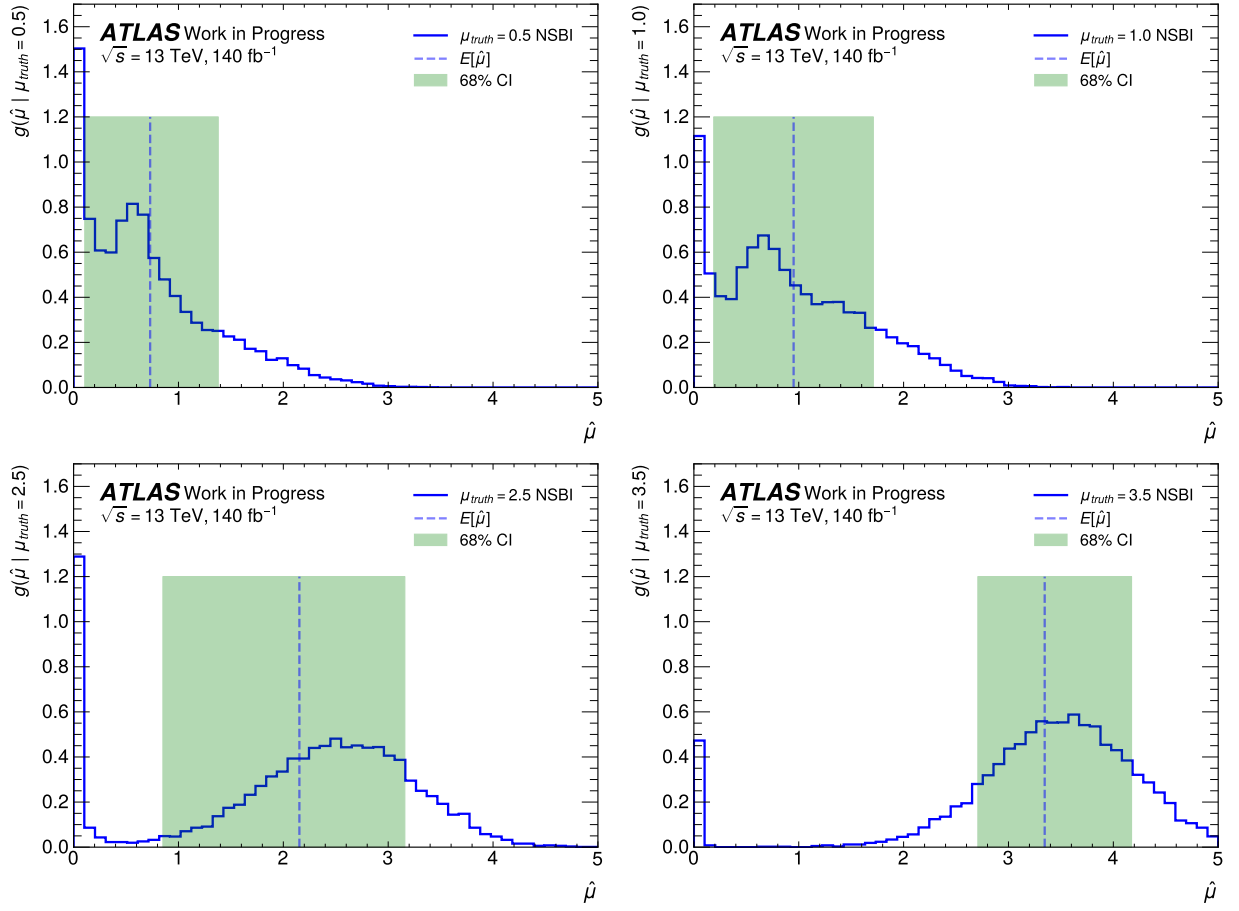


Figure B.2: The sampling distribution $g(\hat{\mu} | \mu_{\text{truth}})$ for the POI $\mu_{\text{off-shell}}$ of the off-shell Higgs boson analysis, defined in Chap. 4. Sampled pseudo-experiments are used to fit the POI μ using the t_{μ} test statistic with sampled AOs. The resulting fit distributions $\hat{\mu}$ are shown as a function of $\mu = 0.5, 1.0, 2.5, 3.5$ (top left, top right, bottom left, bottom right).

Bibliography

- [1] Kyle Cranmer, Johann Brehmer, and Gilles Louppe,
The frontier of simulation-based inference, Proc. Nat. Acad. Sci. **117** (2020) 30055,
arXiv: 1911.01429 [hep-ph] (cit. on pp. 1, 30, 45).
- [2] Johann Brehmer, Kyle Cranmer, Gilles Louppe, and Juan Pavez,
A guide to constraining effective field theories with machine learning,
Phys. Rev. D **98** (2018) 052004, arXiv: 1805.00020 [hep-ph] (cit. on pp. 1, 34, 40).
- [3] ATLAS Collaboration, *Evidence of off-shell Higgs boson production from ZZ leptonic decay channels and constraints on its total width with the ATLAS detector*,
Phys. Lett. B **846** (2023) 138223, arXiv: 2304.01532 [hep-ex]
(cit. on pp. 2, 3, 66–69, 72, 82, 107, 108, 112, 117, 128–131),
Erratum: Phys. Lett. B **854** (2024) 138734.
- [4] Georges Aad et al.,
Measurement of off-shell Higgs boson production in the $H^ \rightarrow ZZ \rightarrow 4\ell$ decay channel using a neural simulation-based inference technique in 13 TeV pp collisions with the ATLAS detector*,
(2024), arXiv: 2412.01548 [hep-ex] (cit. on p. 3).
- [5] Georges Aad et al.,
An implementation of neural simulation-based inference for parameter estimation in ATLAS,
(2024), arXiv: 2412.01600 [hep-ex] (cit. on p. 3).
- [6] Jay Sandesara et al., *ATLAS Data Analysis using a Parallel Workflow on Distributed Cloud-based Services with GPUs*, EPJ Web Conf. **295** (2024) 04007
(cit. on pp. 3, 46, 47, 121, 127).
- [7] Fernando Barreiro Megino et al., *Operational experience and R&D results using the Google Cloud for High-Energy Physics in the ATLAS experiment*,
Int. J. Mod. Phys. A **39** (2024) 2450054, arXiv: 2403.15873 [hep-ex] (cit. on p. 3).
- [8] *Effective field theory interpretation of the measurement of off-shell Higgs boson production from $ZZ \rightarrow 4\ell$ and $ZZ \rightarrow 2\ell 2\nu$ decay channels with the ATLAS detector*, tech. rep.,
All figures including auxiliary figures are available at
<https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PUBNOTES/ATL-PHYS-PUB-2023-012>: CERN, 2023, URL: <https://cds.cern.ch/record/2860128> (cit. on p. 3).

- [9] Lyndon Evans and Philip Bryant, *LHC Machine*, Journal of Instrumentation **3** (2008) S08001, URL: <https://dx.doi.org/10.1088/1748-0221/3/08/S08001> (cit. on p. 4).
- [10] Julie Haffner, *The CERN accelerator complex. Complexe des accélérateurs du CERN*, (2013), General Photo, URL: <https://cds.cern.ch/record/1621894> (cit. on p. 5).
- [11] Izaak Neutelings, *CMS coordinate system*, URL: https://tikz.net/axis3d_cms/ (visited on 12/29/2024) (cit. on p. 7).
- [12] Joao Pequenaio, “Computer generated image of the ATLAS inner detector”, 2008, URL: <https://cds.cern.ch/record/1095926> (cit. on p. 8).
- [13] Georges Aad et al., *Alignment of the ATLAS Inner Detector in Run-2*, Eur. Phys. J. C **80** (2020) 1194, arXiv: 2007.07624 [hep-ex] (cit. on p. 8).
- [14] G Aad et al., *The ATLAS Experiment at the CERN Large Hadron Collider*, JINST **3** (2008) S08003, Also published by CERN Geneva in 2010, URL: <https://cds.cern.ch/record/1129811> (cit. on pp. 10, 11).
- [15] ATLAS Collaboration, *Performance of the ATLAS trigger system in 2015*, Eur. Phys. J. C **77** (2017) 317, arXiv: 1611.09661 [hep-ex] (cit. on p. 11).
- [16] The ATLAS collaboration, *Operation of the ATLAS trigger system in Run 2*, Journal of Instrumentation **15** (2020) P10004, URL: <https://dx.doi.org/10.1088/1748-0221/15/10/P10004> (cit. on p. 11).
- [17] T Cornelissen et al., *Concepts, Design and Implementation of the ATLAS New Tracking (NEWT)*, tech. rep., All figures including auxiliary figures are available at <https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PUBNOTES/ATL-SOFT-PUB-2007-007>: CERN, 2007, URL: <https://cds.cern.ch/record/1020106> (cit. on pp. 11, 13).
- [18] Joao Pequenaio, “Event Cross Section in a computer generated image of the ATLAS detector.”, 2008, URL: <https://cds.cern.ch/record/1096081> (cit. on pp. 12, 27).
- [19] ATLAS Collaboration, *Athena*, version 21.0.127, 2021, URL: <https://doi.org/10.5281/zenodo.4772550> (cit. on pp. 12, 13).
- [20] Georges Aad et al., *Software and computing for Run 3 of the ATLAS experiment at the LHC*, (2024), arXiv: 2404.06335 [hep-ex] (cit. on pp. 13, 14).
- [21] *Performance of the ATLAS Inner Detector Track and Vertex Reconstruction in the High Pile-Up LHC Environment*, tech. rep., All figures including auxiliary figures are available at <https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/CONFNOTES/ATLAS-CONF-2012-042>: CERN, 2012, URL: <https://cds.cern.ch/record/1435196> (cit. on p. 13).
- [22] T G Cornelissen et al., *Updates of the ATLAS Tracking Event Data Model (Release 13)*, tech. rep., All figures including auxiliary figures are available at <https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PUBNOTES/ATL-SOFT-PUB-2007-003>: CERN, 2007, URL: <https://cds.cern.ch/record/1038095> (cit. on p. 15).

- [23] ATLAS Collaboration, *Jet reconstruction and performance using particle flow with the ATLAS Detector*, Eur. Phys. J. C **77** (2017) 466, arXiv: 1703.10485 [hep-ex] (cit. on pp. 22, 72).
- [24] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez, *The anti- k_t jet clustering algorithm*, JHEP **04** (2008) 063, arXiv: 0802.1189 [hep-ph] (cit. on pp. 22, 72).
- [25] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez, *FastJet user manual*, Eur. Phys. J. C **72** (2012) 1896, arXiv: 1111.6097 [hep-ph] (cit. on pp. 22, 72).
- [26] ATLAS Collaboration, *Jet energy scale and resolution measured in proton–proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, Eur. Phys. J. C **81** (2020) 689, arXiv: 2007.02645 [hep-ex] (cit. on pp. 22, 67, 72).
- [27] J. Neyman and E. S. Pearson, *On the Problem of the Most Efficient Tests of Statistical Hypotheses*, Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character **231** (1933) 289, ISSN: 02643952, URL: <http://www.jstor.org/stable/91247> (cit. on p. 25).
- [28] Glen Cowan, Kyle Cranmer, Eilam Gross, and Ofer Vitells, *Asymptotic formulae for likelihood-based tests of new physics*, Eur. Phys. J. C **71** (2011) (cit. on p. 26).
- [29] Stefan Hoche, “Introduction to parton-shower event generators”, 2014, URL: <https://api.semanticscholar.org/CorpusID:119291985> (cit. on p. 27).
- [30] Christian Bierlich et al., *A comprehensive guide to the physics and usage of PYTHIA 8.3*, (2022), arXiv: 2203.11601 [hep-ph] (cit. on p. 26).
- [31] GEANT4 Collaboration, S. Agostinelli, et al., *GEANT4 – a simulation toolkit*, Nucl. Instrum. Meth. A **506** (2003) 250 (cit. on pp. 27, 71).
- [32] ATLAS Collaboration, *The ATLAS Collaboration Software and Firmware*, ATL-SOFT-PUB-2021-001, 2021, URL: <https://cds.cern.ch/record/2767187> (cit. on p. 28).
- [33] Kyle Cranmer, George Lewis, Lorenzo Moneta, Akira Shibata, and Wouter Verkerke, *HistFactory: A tool for creating statistical models for use with RooFit and RooStats*, tech. rep., New York U., 2012, URL: <https://cds.cern.ch/record/1456844> (cit. on pp. 29, 42).
- [34] Morad Aaboud et al., *Observation of $H \rightarrow b\bar{b}$ decays and VH production with the ATLAS detector*, Phys. Lett. B **786** (2018) 59, arXiv: 1808.08238 [hep-ex] (cit. on p. 31).
- [35] Kyle Cranmer, Johann Brehmer, and Gilles Louppe, *The frontier of simulation-based inference*, Proceedings of the National Academy of Sciences **117** (2020) 30055, eprint: <https://www.pnas.org/doi/pdf/10.1073/pnas.1912789117>, URL: <https://www.pnas.org/doi/abs/10.1073/pnas.1912789117> (cit. on p. 32).

- [36] Siddharth Mishra-Sharma and Kyle Cranmer,
Neural simulation-based inference approach for characterizing the Galactic Center γ -ray excess,
Phys. Rev. D **105** (2022) 063017, arXiv: 2110.06931 [astro-ph.HE] (cit. on p. 32).
- [37] Lukas Heinrich, Siddharth Mishra-Sharma, Chris Pollard, and Philipp Windischhofer,
Hierarchical Neural Simulation-Based Inference Over Event Ensembles, 2024,
arXiv: 2306.12584 [stat.ML], URL: <https://arxiv.org/abs/2306.12584> (cit. on p. 32).
- [38] K. Kondo, *Dynamical Likelihood Method for Reconstruction of Events With Missing Momentum. 1: Method and Toy Models*, J. Phys. Soc. Jap. **57** (1988) 4126 (cit. on p. 33).
- [39] V. M. Abazov et al., *A precision measurement of the mass of the top quark*,
Nature **429** (2004) 638, arXiv: hep-ex/0406031 (cit. on p. 33).
- [40] Georges Aad et al.,
Measurement of the Higgs boson mass in the $H \rightarrow ZZ^ \rightarrow 4\ell$ decay channel using 139 fb^{-1} of $\sqrt{s} = 13\text{ TeV}$ pp collisions recorded by the ATLAS detector at the LHC*,
Phys. Lett. B **843** (2023) 137880, arXiv: 2207.00320 [hep-ex] (cit. on p. 33).
- [41] Georges Aad et al., *Measurement of the Higgs boson mass with $H \rightarrow \gamma\gamma$ decays in 140 fb^{-1} of $s=13\text{ TeV}$ pp collisions with the ATLAS detector*, Phys. Lett. B **847** (2023) 138315,
arXiv: 2308.07216 [hep-ex] (cit. on pp. 33, 104).
- [42] Kyle Cranmer, Juan Pavez, and Gilles Louppe,
Approximating Likelihood Ratios with Calibrated Discriminative Classifiers, (2015),
arXiv: 1506.02169 [stat.AP] (cit. on pp. 34, 37, 80, 81, 85).
- [43] Glen Cowan, Kyle Cranmer, Eilam Gross, and Ofer Vitells,
Asymptotic formulae for likelihood-based tests of new physics,
Eur. Phys. J. C **71** (2011) 1554, [Erratum: Eur.Phys.J.C 73, 2501 (2013)],
arXiv: 1007.1727 [physics.data-an] (cit. on p. 36).
- [44] Pierre Baldi, Kyle Cranmer, Taylor Faucett, Peter Sadowski, and Daniel Whiteson,
Parameterized neural networks for high-energy physics, Eur. Phys. J. C **76** (2016) 235,
arXiv: 1601.07913 [hep-ex] (cit. on p. 37).
- [45] Nachtmann Diehl M., *O. Optimal observables for the measurement of three gauge boson couplings in $e^+e^- \rightarrow W^+W^-$* , Z. Phys. C - Particles and Fields **62** (1994) 397,
URL: <https://doi.org/10.1007/BF01555899> (cit. on p. 40).
- [46] Davison E. Soper and Michael Spannowsky,
Finding physics signals with shower deconstruction, Phys. Rev. D **84** (2011) 074002,
arXiv: 1102.3480 [hep-ph] (cit. on p. 40).
- [47] M. Davier, L. Duflot, F. Le Diberder, and A. Roug ,
The optimal method for the measurement of tau polarization,
Physics Letters B **306** (1993) 411, ISSN: 0370-2693,
URL: <https://www.sciencedirect.com/science/article/pii/037026939390101M>
(cit. on p. 40).

- [48] Johann Brehmer, Gilles Louppe, Juan Pavez, and Kyle Cranmer, *Mining gold from implicit models to improve likelihood-free inference*, Proc. Nat. Acad. Sci. **117** (2020) 5242, arXiv: 1805.12244 [stat.ML] (cit. on p. 42).
- [49] Kyle Cranmer, George Lewis, Lorenzo Moneta, Akira Shibata, and Wouter Verkerke, *HistFactory: A tool for creating statistical models for use with RooFit and RooStats*, (2012), URL: <https://cds.cern.ch/record/1456844> (cit. on p. 43).
- [50] Kyle Cranmer, *Practical Statistics for the LHC*, (2014), arXiv: 1503.07622 [physics.data-an] (cit. on pp. 43, 49).
- [51] F James, *MINUIT: Function Minimization and Error Analysis Reference Manual*, (1998), CERN Program Library Long Writeups, URL: <https://cds.cern.ch/record/2296388> (cit. on p. 46).
- [52] Wouter Verkerke and David Kirkby, *The RooFit toolkit for data modeling*, 2003, arXiv: physics/0306116 [physics.data-an] (cit. on p. 46).
- [53] Lorenzo Moneta et al., *The RooStats Project*, 2011, arXiv: 1009.1003 [physics.data-an] (cit. on p. 46).
- [54] Atilim Gunes Baydin, Barak A. Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind, *Automatic differentiation in machine learning: a survey*, J. Mach. Learn. Res. **18** (2018) 1, arXiv: 1502.05767 [cs.SC] (cit. on p. 46).
- [55] Andrés Pinto et al., *Uncertainty components in profile likelihood fits*, Eur. Phys. J. C **84** (2024) 593, arXiv: 2307.04007 [physics.data-an] (cit. on pp. 47, 109).
- [56] Jens Praestgaard and Jon A. Wellner, *Exchangeably Weighted Bootstraps of the General Empirical Process*, Ann. Probab. **21** (1993) 2053 (cit. on p. 48).
- [57] ATLAS Collaboration, *Recommendations for the Modeling of Smooth Backgrounds*, ATL-PHYS-PUB-2020-028, 2020, URL: <https://cds.cern.ch/record/2743717> (cit. on pp. 48, 104).
- [58] M. H. Ahn et al., *Indications of neutrino oscillation in a 250 km long baseline experiment*, Phys. Rev. Lett. **90** (2003) 041801, arXiv: hep-ex/0212007 (cit. on p. 53).
- [59] J. Schechter and J. W. F. Valle, *Neutrino Masses in $SU(2) \times U(1)$ Theories*, Phys. Rev. D **22** (1980) 2227 (cit. on p. 53).
- [60] Clay Cordova, Sungwoo Hong, Seth Koren, and Kantaro Ohmori, *Neutrino Masses from Generalized Symmetry Breaking*, Phys. Rev. X **14** (2024) 031033, arXiv: 2211.07639 [hep-ph] (cit. on p. 53).
- [61] Andre de Gouvea, “TASI lectures on neutrino physics”, *Theoretical Advanced Study Institute in Elementary Particle Physics: Physics in $D \geq 4$* , 2004 197, arXiv: hep-ph/0411274 (cit. on p. 53).

- [62] Georges Aad et al.,
Combined Measurement of the Higgs Boson Mass from the $H \rightarrow \gamma\gamma$ and $H \rightarrow ZZ^ \rightarrow 4\ell$ Decay Channels with the ATLAS Detector Using $s=7, 8$, and 13 TeV pp Collision Data*,
Phys. Rev. Lett. **131** (2023) 251802, arXiv: 2308.04775 [hep-ex] (cit. on p. 54).
- [63] J R Andersen et al., *Handbook of LHC Higgs Cross Sections: 3. Higgs Properties*,
(2013), ed. by S Heinemeyer, C Mariotti, G Passarino, and R Tanaka,
arXiv: 1307.1347 [hep-ph] (cit. on pp. 55, 56, 60).
- [64] Georges Aad et al., *A detailed map of Higgs boson interactions by the ATLAS experiment ten years after the discovery*, Nature **607** (2022) 52, [Erratum: Nature 612, E24 (2022)],
arXiv: 2207.00092 [hep-ex] (cit. on p. 56).
- [65] Izaak Neutelings, *SM decays*,
URL: https://tikz.net/sm_decay_piechart/ (visited on 12/29/2024) (cit. on p. 56).
- [66] Georges Aad et al., *A detailed map of Higgs boson interactions by the ATLAS experiment ten years after the discovery*, Nature **607** (2022) 52, [Erratum: Nature 612, E24 (2022)],
arXiv: 2207.00092 [hep-ex] (cit. on p. 57).
- [67] Nikolas Kauer and Giampiero Passarino,
Inadequacy of zero-width approximation for a light Higgs boson signal, JHEP **08** (2012) 116,
arXiv: 1206.4803 [hep-ph] (cit. on p. 57).
- [68] Nikolas Kauer, *Inadequacy of zero-width approximation for a light Higgs boson signal*,
Mod. Phys. Lett. A **28** (2013) 1330015, arXiv: 1305.2092 [hep-ph] (cit. on p. 59).
- [69] John M. Campbell, R. Keith Ellis, and Ciaran Williams,
Bounding the Higgs width at the LHC using full analytic results for $gg \rightarrow e^-e^+\mu^-\mu^+$,
JHEP **04** (2014) 060, arXiv: 1311.3589 [hep-ph] (cit. on p. 59).
- [70] John M. Campbell, R. Keith Ellis, and Ciaran Williams,
Bounding the Higgs width at the LHC: Complementary results from $H \rightarrow WW$,
Phys. Rev. D **89** (2014) 053011, arXiv: 1312.1628 [hep-ph] (cit. on p. 59).
- [71] Fabrizio Caola and Kirill Melnikov,
Constraining the Higgs boson width with ZZ production at the LHC,
Phys. Rev. D **88** (2013) 054024, arXiv: 1307.4935 [hep-ph] (cit. on p. 59).
- [72] Federico Buccioni et al., *Interference effects in $gg \rightarrow H \rightarrow Z\gamma$ beyond leading order*,
Phys. Lett. B **851** (2024) 138596, arXiv: 2312.12384 [hep-ph] (cit. on p. 59).
- [73] Georges Aad et al., *Constraint on the total width of the Higgs boson from Higgs boson and four-top-quark measurements in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*,
(2024), arXiv: 2407.10631 [hep-ex] (cit. on p. 59).
- [74] Margherita Ghezzi, Raquel Gomez-Ambrosio, Giampiero Passarino, and Sandro Uccirati,
NLO Higgs effective field theory and κ -framework, JHEP **07** (2015) 175,
arXiv: 1505.03706 [hep-ph] (cit. on p. 60).

- [75] ATLAS Collaboration, *Measurement of the Higgs boson mass in the $H \rightarrow ZZ^* \rightarrow 4\ell$ decay channel using 139 fb^{-1} of $\sqrt{s} = 13\text{ TeV}$ pp collisions recorded by the ATLAS detector at the LHC*, (2022), arXiv: 2207.00320 [hep-ex] (cit. on p. 60).
- [76] Giampiero Passarino, *Higgs interference effects in $gg \rightarrow ZZ$ and their uncertainty*, JHEP **08** (2012) 146, arXiv: 1206.3824 [hep-ph] (cit. on p. 60).
- [77] O. Mattelaer and K. Ostrolenk, *Speeding up MadGraph5_aMC@NLO*, Eur. Phys. J. C **81** (2021) 435, arXiv: 2102.00773 [hep-ph] (cit. on p. 63).
- [78] Georges Aad et al., *Evidence of off-shell Higgs boson production from ZZ leptonic decay channels and constraints on its total width with the ATLAS detector*, Phys. Lett. B **846** (2023) 138223, arXiv: 2304.01532 [hep-ex] (cit. on pp. 66, 82).
- [79] ATLAS Collaboration, *Luminosity determination in pp collisions at $\sqrt{s} = 13\text{ TeV}$ using the ATLAS detector at the LHC*, (2022), arXiv: 2212.09379 [hep-ex] (cit. on pp. 67, 70, 128).
- [80] G. Avoni et al., *The new LUCID-2 detector for luminosity measurement and monitoring in ATLAS*, JINST **13** (2018) P07017 (cit. on p. 67).
- [81] ATLAS Collaboration, *Performance of electron and photon triggers in ATLAS during LHC Run 2*, Eur. Phys. J. C **80** (2020) 47, arXiv: 1909.00761 [hep-ex] (cit. on pp. 67, 70).
- [82] ATLAS Collaboration, *Electron and photon efficiencies in LHC Run 2 with the ATLAS experiment*, JHEP **05** (2024) 162, arXiv: 2308.13362 [hep-ex] (cit. on pp. 67, 72).
- [83] ATLAS Collaboration, *Performance of the ATLAS muon triggers in Run 2*, JINST **15** (2020) P09015, arXiv: 2004.13447 [hep-ex] (cit. on pp. 67, 70).
- [84] ATLAS Collaboration, *Muon reconstruction and identification efficiency in ATLAS using the full Run 2 pp collision data set at $\sqrt{s} = 13\text{ TeV}$* , Eur. Phys. J. C **81** (2021) 578, arXiv: 2012.00578 [hep-ex] (cit. on pp. 67, 72).
- [85] NNPDF Collaboration, Richard D. Ball, et al., *Parton distributions for the LHC run II*, JHEP **04** (2015) 040, arXiv: 1410.8849 [hep-ph] (cit. on pp. 68, 70, 71).
- [86] Fabrizio Caola, Kirill Melnikov, Raoul Röntsch, and Lorenzo Tancredi, *QCD corrections to ZZ production in gluon fusion at the LHC*, Phys. Rev. D **92** (2015) 094028, arXiv: 1509.06734 [hep-ph] (cit. on pp. 68, 70, 71).
- [87] Bakul Agarwal, Stephen Jones, Matthias Kerner, and Andreas von Manteuffel, *Complete NLO QCD Corrections to ZZ Production in Gluon Fusion*, (2024), arXiv: 2404.05684 [hep-ph] (cit. on p. 68).
- [88] F. Cascioli et al., *Precise Higgs-background predictions: merging NLO QCD and squared quark-loop corrections to four-lepton + 0,1 jet production*, JHEP **01** (2014) 046, arXiv: 1309.0500 [hep-ph] (cit. on pp. 68, 70).

- [89] S. Catani, F. Krauss, B. R. Webber, and R. Kuhn, *QCD Matrix Elements + Parton Showers*, JHEP **11** (2001) 063, arXiv: [hep-ph/0109231](#) (cit. on p. 68).
- [90] Stefan Gieseke, Tobias Kasprzik, and Johann H. Kühn, *Vector-boson pair production and electroweak corrections in HERWIG++*, Eur. Phys. J. C **74** (2014) 2988, arXiv: [1401.3964 \[hep-ph\]](#) (cit. on p. 69).
- [91] ATLAS Collaboration, *ATLAS data quality operations and performance for 2015–2018 data-taking*, JINST **15** (2020) P04003, arXiv: [1911.04632 \[physics.ins-det\]](#) (cit. on p. 70).
- [92] ATLAS Collaboration, *The ATLAS Inner Detector Trigger performance in pp collisions at 13 TeV during LHC Run 2*, Eur. Phys. J. C **82** (2021) 206, arXiv: [2107.02485 \[hep-ex\]](#) (cit. on p. 70).
- [93] Enrico Bothmann et al., *Event generation with Sherpa 2.2*, SciPost Phys. **7** (2019) 034, arXiv: [1905.09127 \[hep-ph\]](#) (cit. on p. 70).
- [94] Federico Buccioni et al., *OpenLoops 2*, Eur. Phys. J. C **79** (2019) 866, arXiv: [1907.13071 \[hep-ph\]](#) (cit. on p. 70).
- [95] Fabio Cascioli, Philipp Maierhöfer, and Stefano Pozzorini, *Scattering Amplitudes with Open Loops*, Phys. Rev. Lett. **108** (2012) 111601, arXiv: [1111.5206 \[hep-ph\]](#) (cit. on p. 70).
- [96] Ansgar Denner, Stefan Dittmaier, and Lars Hofer, *COLLIER: A fortran-based complex one-loop library in extended regularizations*, Comput. Phys. Commun. **212** (2017) 220, arXiv: [1604.06792 \[hep-ph\]](#) (cit. on p. 70).
- [97] Stefan Höche, Frank Krauss, Marek Schönherr, and Frank Siegert, *QCD matrix elements + parton showers. The NLO case*, JHEP **04** (2013) 027, arXiv: [1207.5030 \[hep-ph\]](#) (cit. on p. 70).
- [98] Charalampos Anastasiou, Kirill Melnikov, and Frank Petriello, *Fully differential Higgs boson production and the di-photon signal through next-to-next-to-leading order*, Nucl. Phys. B **724** (2005) 197, arXiv: [hep-ph/0501130](#) (cit. on p. 70).
- [99] Stefano Catani and Massimiliano Grazzini, *Next-to-Next-to-Leading-Order Subtraction Formalism in Hadron Collisions and its Application to Higgs-Boson Production at the Large Hadron Collider*, Phys. Rev. Lett. **98** (2007) 222002, arXiv: [hep-ph/0703012 \[hep-ph\]](#) (cit. on p. 70).
- [100] Massimiliano Grazzini, Stefan Kallweit, and Marius Wiesemann, *Fully differential NNLO computations with MATRIX*, Eur. Phys. J. C **78** (2018) 537, arXiv: [1711.06631 \[hep-ph\]](#) (cit. on p. 70).
- [101] Charalampos Anastasiou, Claude Duhr, Falko Dulat, Franz Herzog, and Bernhard Mistlberger, *Higgs Boson Gluon-Fusion Production in QCD at Three Loops*, Phys. Rev. Lett. **114** (2015) 212001, arXiv: [1503.06056 \[hep-ph\]](#) (cit. on p. 70).

- [102] Giampiero Passarino, *Higgs CAT*, Eur. Phys. J. C **74** (2014) 2866, arXiv: 1312.2397 [hep-ph] (cit. on pp. 70, 71).
- [103] J. Alwall et al., *The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations*, JHEP **07** (2014) 079, arXiv: 1405.0301 [hep-ph] (cit. on p. 71).
- [104] Torbjörn Sjöstrand et al., *An introduction to PYTHIA 8.2*, Comput. Phys. Commun. **191** (2015) 159, arXiv: 1410.3012 [hep-ph] (cit. on p. 71).
- [105] ATLAS Collaboration, *ATLAS Pythia 8 tunes to 7 TeV data*, ATL-PHYS-PUB-2014-021, 2014, URL: <https://cds.cern.ch/record/1966419> (cit. on p. 71).
- [106] NNPDF Collaboration, Richard D. Ball, et al., *Parton distributions with LHC data*, Nucl. Phys. B **867** (2013) 244, arXiv: 1207.1303 [hep-ph] (cit. on p. 71).
- [107] B. Biedermann, A. Denner, S. Dittmaier, L. Hofer, and B. Jäger, *Electroweak Corrections to $pp \rightarrow \mu^+ \mu^- e^+ e^- + X$ at the LHC: A Higgs Boson Background Study*, Phys. Rev. Lett. **116** (2016) 161803, arXiv: 1601.07787 [hep-ph] (cit. on p. 71).
- [108] Benedikt Biedermann, Ansgar Denner, Stefan Dittmaier, Lars Hofer, and Barbara Jäger, *Next-to-leading-order electroweak corrections to the production of four charged leptons at the LHC*, JHEP **01** (2017) 033, arXiv: 1611.05338 [hep-ph] (cit. on p. 71).
- [109] S. Frixione, V. Hirschi, D. Pagani, H-S. Shao, and M. Zaro, *Electroweak and QCD corrections to top-pair hadroproduction in association with heavy bosons*, JHEP **06** (2015) 184, arXiv: 1504.03446 [hep-ph] (cit. on p. 71).
- [110] ATLAS Collaboration, *The ATLAS Simulation Infrastructure*, Eur. Phys. J. C **70** (2010) 823, arXiv: 1005.4568 [physics.ins-det] (cit. on p. 71).
- [111] T. Sjöstrand, S. Mrenna, and P. Skands, *A brief introduction to PYTHIA 8.1*, Comput. Phys. Commun. **178** (2008) 852, arXiv: 0710.3820 [hep-ph] (cit. on p. 71).
- [112] ATLAS Collaboration, *The Pythia 8 A3 tune description of ATLAS minimum bias and inelastic measurements incorporating the Donnachie–Landshoff diffractive model*, ATL-PHYS-PUB-2016-017, 2016, URL: <https://cds.cern.ch/record/2206965> (cit. on p. 71).
- [113] ATLAS Collaboration, *Performance of pile-up mitigation techniques for jets in pp collisions at $\sqrt{s} = 8$ TeV using the ATLAS detector*, Eur. Phys. J. C **76** (2016) 581, arXiv: 1510.03823 [hep-ex] (cit. on p. 72).
- [114] ATLAS Collaboration, *Forward jet vertex tagging using the particle flow algorithm*, ATL-PHYS-PUB-2019-026, 2019, URL: <https://cds.cern.ch/record/2683100> (cit. on p. 72).
- [115] Sara Bolognesi et al., *Spin and parity of a single-produced resonance at the LHC*, Phys. Rev. D **86** (2012) 095031, arXiv: 1208.4018 [hep-ph] (cit. on p. 73).
- [116] Prajit Ramachandran, Barret Zoph, and Quoc V. Le, *Searching for Activation Functions*, (2017), arXiv: 1710.05941 [cs.NE] (cit. on pp. 76, 80).

- [117] George V. Moustakides and Kalliopi Basioti, *Training Neural Networks for Likelihood/Density Ratio Estimation*, 2019, URL: <https://arxiv.org/abs/1911.00405> (cit. on p. 80).
- [118] Sumu Li, Dequan Li, and Yuheng Zhang, “Incorporating Nesterov’s Momentum into Distributed Adaptive Gradient Method for Online Optimization”, *2021 China Automation Congress (CAC)*, 2021 7338 (cit. on p. 81).
- [119] François Chollet et al., *Keras*, 2015, URL: <https://keras.io> (cit. on p. 81).
- [120] TensorFlow Developers, *TensorFlow*, version v2.9.0-rc2, 2022, URL: <https://doi.org/10.5281/zenodo.6519082> (cit. on p. 81).
- [121] Lluís Garrido and Aurelio Juste, *On the determination of probability density functions by using Neural Networks*, *Comput. Phys. Commun.* **115** (1998) 25, arXiv: physics/9807018 [physics.data-an] (cit. on p. 81).
- [122] Georges Aad et al., *A simultaneous unbinned differential cross section measurement of twenty-four Z +jets kinematic observables with the ATLAS detector*, (2024), arXiv: 2405.20041 [hep-ex] (cit. on p. 82).
- [123] Johann Brehmer, Siddharth Mishra-Sharma, Joeri Hermans, Gilles Louppe, and Kyle Cranmer, *Mining for Dark Matter Substructure: Inferring subhalo population properties from strong lenses with machine learning*, *Astrophys. J.* **886** (2019) 49, arXiv: 1909.02005 [astro-ph.CO] (cit. on p. 85).
- [124] J. Neyman, *Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability*, *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences* **236** (1937) 333, ISSN: 00804614 (cit. on p. 108).
- [125] ATLAS Collaboration, *A detailed map of Higgs boson interactions by the ATLAS experiment ten years after the discovery*, *Nature* **607** (2022) 52, arXiv: 2207.00092 [hep-ex] (cit. on p. 113).
- [126] ATLAS Collaboration, *Higgs boson production cross-section measurements and their EFT interpretation in the 4ℓ decay channel at $\sqrt{s} = 13$ TeV with the ATLAS detector*, *Eur. Phys. J. C* **80** (2020) 957, arXiv: 2004.03447 [hep-ex] (cit. on pp. 114, 115, 117), Erratum: *Eur. Phys. J. C* **81** (2021) 29.
- [127] James Bradbury et al., *JAX: composable transformations of Python+NumPy programs*, version 0.3.13, 2018, URL: <http://github.com/google/jax> (cit. on p. 120).
- [128] Hans Dembinski et al., *scikit-hep/iminuit*, 2020, URL: <https://doi.org/10.5281/zenodo.3949207> (visited on 07/14/2023) (cit. on p. 120).

- [129] F. James and M. Roos, *Minuit - a system for function minimization and analysis of the parameter errors and correlations*, Computer Physics Communications **10** (1975) 343, ISSN: 0010-4655,
URL: <https://www.sciencedirect.com/science/article/pii/0010465575900399>
(cit. on p. 120).