Measuring Higgs Boson Couplings, including to the Top Quark, in the Diphoton Decay Channel with Run 2 Data Collected by the ATLAS Detector

by

Rachel Jordan Hyneman

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Physics)
in the University of Michigan
2020

Doctoral Committee:

Associate Professor Tom Schwarz, Chair
Professor Sara Pozzi
Assistant Professor Joshua Spitz
Professor James Wells
Professor Bing Zhou

Rachel Jordan Hyneman

rhyneman@umich.edu

ORCID ID: 0000-0002-9093-7141

Dedicated to all the women physicists who have come before me.

ACKNOWLEDGEMENTS

I've had the support of so many wonderful people while working towards this dissertation. First of all, thank you to my parents, Roberta and Michael, for supporting me in seeking out a career which doesn't involve teeth. More seriously, they have put in so much effort learning about my research, and they have been supportive of all my decisions to move far away from Arizona. I also owe a deep gratitude to my grandmother, Beki, and my great aunt, Joan, for helping instill a love of science throughout my life. I also owe a huge thanks to my advisor, Tom Schwarz. I've had so much freedom to pursue unique and crazy ideas while a student, which I try not to take for granted. Tom has been incredibly supportive beyond my research, as well. He's always been available if I've had concerns, including personal ones. I hope that if I ever end up managing people, that I can do so as empathetically as Tom has. I've had a fantastic time working with the other members of the ATLAS "Fondue" group - Daniel Marley, Hao Liu, Garrett Merz, Allison McCarn-Deiana, Xueye Hu, Yanlin Liu, Jem Guhit, and Sabrina Corsetti. I wish all of you the best in the future. The other ATLAS professors – Bing Zhou, Junjie Zhu, Dante Amidei, and Jianming Qian – have been great to work with, as well. My professors and advisors from my undergraduate studies also deserve a huge thank you. In particular, I owe my various advisors from William & Mary: Irina Novakova, Patricia Vahle, Gina Hoatson, Joshua Erlich, and Wouter Deconinck. Additionally, my early research advisors from UCLA, Stephanie Wissel and Konstantin Belov, played a huge role in steering me towards graduate school. I can't leave out my first ATLAS advisor, Antonio Sidoti, nor the organizers of the University of Michigan CERN-REU program for giving me a wonderful introduction to LHC physics. I should thank the many friends who have helped me through the last few years, both at Michigan and abroad. Michael Viray, Brandon and Kate Berg,

Shruti Paranjape, Joseph and Alissa Kleinhenz, and many others made Michigan so much better despite the cold and dreary winters. Christine McLean, Riju Dasgupta, Reyer Band, Matthias Weber, Chris Brainerd, Chris not-from-Coldplay Martin, Bing Liu, Paul Lujan, Juliana Froggatt, Jennett Dickinson, Savannah Thais, Juliette Alimena, Alex Morton, and so many others made me feel at home while living far away in Switzer-France. I also want to express my appreciation to Eve Chase, Joseph Stief, and Aaron Epstien for keeping up our friendships from William & Mary. Brittany Creech deserves a special mention, too, for being a great friend for over twenty years. I owe a very special thank you to my high school physics teacher, Jim Archambault, for encouraging me to continue studying physics. Your belief that I would be successful in the field meant a lot in helping me start to go down this long road. Lastly, I can't express how thankful I am to have Callum Jones, the kindest, most patient, and most supportive person I know, in my life.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

LIST OF APPENDICES

APPENDIX

LIST OF ABBREVIATIONS

SM Standard Model

LHC Large Hadron Collider

ID Inner Detector

SCT Silicon Microstrip

IBL Insertable B Layer

MBTS Minimum Bias Trigger Scintillators

TRT Transition Radiation Tracker

EM Electromagnetic

LAr Liquid Argon

HEC Hadronic Endcap Calorimeter

FCAL Forward Calorimeter

PMT Photomultiplier Tube

MS Muon Spectrometer

MDT Monitor Drift Tube

CSC Cathode Strip Chamber

TGC Thin Gap Chamber

RPC Resistive Plate Chamber

HLT High Level Trigger

CTP Central Trigger Processor

RoI Region of Interest

CBM Combined Muon

STM Segment Tagged Muon

CTM Calorimeter Tagged Muon

ME Extrapolated Muon

BSM Beyond Standard Model

STXS Simplified Template Cross Section

BDT Boosted Decision Tree

DSCB Double-Sided Crystal Ball

CB Crystal Ball

GP Gaussian Process

RBF Radial Basis Function

PoI Parameter of Interest

NLL Negative Log Likelihood

QCD Quantum Chromodynamics

PDF Parton Distribution Function

PER Photon Energy Resolution

PES Photon Energy Scale

NP Nuisance Parameter

ABSTRACT

This work presents the measurement of Higgs boson couplings through the Higgs boson to diphoton decay ($H \rightarrow \gamma\gamma$) channel using pp collision data with a center-of-mass energy of $\sqrt{s} = 13$ TeV recorded by the ATLAS detector. In order to probe the coupling of the Higgs boson to the fermions and vector bosons, the inclusive cross sections times the branching ratio of the Higgs boson to two photons are measured for four primary production modes: gluon-gluon fusion (ggF), vector boson fusion (VBF), vector boson associated production (VH), and top quark associated production ($t\bar{t}H$). The measurement of the $t\bar{t}H$ production mode was performed using the full Run 2 dataset taken between 2015 and 2018, amounting to 139 fb^{-1} of pp collision data. Measurements of the remaining production modes were performed with the subset of Run 2 data taken between 2015 and 2017, amounting to 79.8 fb^{-1}. Two significant sources of uncertainty in the analysis are that relating to photon isolation and that from background modeling. A new methodology for calculating corrections to photon isolation in the Electromagnetic calorimeter from a high statistics, low purity data sample is presented in detail. The method utilizes a multistep template fit to remove fake photons from the data sample, which allows for a direct comparison of the isolation shape in data and simulation. The background modeling uncertainty is found to be inflated in many analysis categories due to insufficient simulation statistics. A novel technique using Gaussian Processes to smooth out fluctuations in low statistics simulation samples is explored, along with extensive validation studies. The technique shows significant promise in reducing the inflation of the background modeling uncertainty due to statistical fluctuations in simulation samples. The observed signal strength of the $t\bar{t}H$ production mode is $\mu_{t\bar{t}H} = 1.38\ ^{+0.41}_{-0.36}$, which is compatible with the Standard Model prediction. The observed signal strength of the ggF production mode was found to be $\mu_{ggF} = 0.97^{+0.17}_{-0.15}$, that of the VBF production mode $\mu_{VBF} = 1.4^{+0.44}_{-0.37}$, and that

of the combined VH processes ($W^{\pm}H$, $q\bar{q}ZH$, and $ggZH$) is measured to be $\mu_{VH} = 1.09^{+0.61}_{-0.55}$. These results are all consistent with the Standard Model predictions, and so no direct evidence of new physics is observed.

CHAPTER 1

Introduction

The Higgs boson, discovered by the ATLAS [6] and CMS [32] collaborations in 2012 [33, 34], is the most recently discovered fundamental particle. Since then, its properties have been studied extensively by both collaborations. One such property is how strongly the Higgs boson interacts, or *couples*, to other fundamental particles. In particular, studying the rates at which Higgs bosons are produced through different mechanisms offers an essential channel through which to probe these couplings, as they involve different fundamental particles and often leave unique experimental signatures.

This work probes Higgs boson couplings through the measurement of the total cross section of specific physical processes. The observed cross section of a physical process is proportional to the couplings between the particles involved in that process. Translating between a coupling and a cross section requires a complete knowledge of the theory describing process; some additional, unknown process with identical initial and final states could also contribute to the observed cross section.

The couplings between the Higgs boson and other fundamental particles are defined within the Standard Model (SM), the current framework describing fundamental physics at the level of subatomic particles. Assuming this theory is complete, it therefore offers a framework in which to precisely determine the cross sections of Higgs boson processes. Measuring the inclusive cross sections of various Higgs boson processes offers a pathway through which to experimentally test the accuracy of the SM. If a significant deviation from the SM prediction of a cross section is observed, it may indicate that the coupling strength differs from the SM value, or that some new physics process is also contributing to the total cross section. These hypothetical new physics phenomena could be connected to other grand questions regarding the nature of fundamental physics, such as the nature of dark matter and energy.

One particularly interesting Higgs boson coupling to study is that to the top quark. This coupling can be studied by measuring the cross section of Higgs boson production in association with a top/anti-top quark pair. The top quark and its antiparticle partner leave unique signatures, which

may be exploited identify events resulting from this process. However, this process is also very rare compared to other ways of producing a Higgs boson. The top quark is predicted to interact very strongly with the Higgs boson, however, and so any new physics interaction affecting this process may be amplified.

This work presents the measurement of Higgs boson couplings through the Higgs boson to diphoton decay ($H \to \gamma\gamma$) channel using pp collision data with a center-of-mass energy of $\sqrt{s} = 13$ TeV recorded by the ATLAS detector. Particular emphasis is placed on measuring the coupling of the Higgs boson to the top quark through top-associated Higgs boson production ($t\bar{t}H$). This $t\bar{t}H$ portion of the analysis was completed after the remainder of the coupling measurements. It utilizes the full Run 2 dataset taken between 2015 and 2018, amounting to 139 fb^{-1} of pp collision data. The remaining couplings measurements were performed with the subset of Run 2 data taken between 2015 and 2017, amounting to 79.8 fb^{-1}.

The strategy of the analysis presented is to exploit the fact that the primary background, continuum diphoton production, creates a smooth, falling distribution as a function of the diphoton invariant mass. Given sufficient statistics, this background shape can be sufficiently well-modeled by certain analytic functions (those which roughly resemble a falling exponential distribution). The Higgs boson signal, on the other hand, is a narrow resonance that produces a sharp peak on top of the background invariant mass distribution. The narrow signal resonance can also be well-modeled by an analytic function. In order to determine the number of background events underneath the Higgs boson signal peak, a combined signal and background model (using the chosen analytic functions) is fit to a large window in the invariant mass spectrum of the observed data. The mass window includes "side-band" regions (regions on either side of the signal peak which contain negligible signal contamination), which normalize the background function. The number of events fitted with the signal function is then the number of observed Higgs boson events.

The analysis method has the advantage of using a data-driven background estimation, which eliminates the possibility of bias from mismodeling in simulating background processes. The analytic function modeling the background, however, is chosen using simulated background samples. If these samples contain too few events, determining the optimal functional choice may become challenging, and the uncertainty associated with the choice of background function may be inflated. One possible solution to mitigate the statistical fluctuations in these simulated background samples is to apply a smoothing procedure. Such a technique using Gaussian Processes is discussed in detail in this work.

The analysis includes many orthogonal categories, each of which are simultaneously fit with the combined analytic signal and background model. Dedicated selections are applied to sort Higgs boson-like events into the categories. The selections are designed to reduce the magnitude of the continuum background, so that it does not obscure the Higgs boson signal peak. The selections

also aim to separate different types of Higgs boson events, in order to study the properties of the Higgs boson in finer granularity.

The structure of this work is as follows. First, the theoretical background is discussed in Chapter 2. This chapter will detail the fundamental particles of the Standard Model, as well as the nature of the Standard Model Higgs boson. This chapter will also discuss the main Higgs boson production modes targeted by the analysis. Next, a detailed overview of the ATLAS detector and its operational principles is presented in Chapter 3. The following chapter, Chapter 4, discusses how the signals recorded by the ATLAS detector are reconstructed into more familiar physics objects. This chapter also extensively details one specific quantity of photons, that of isolation. The dataset and simulation samples relevant to the analysis are presented in Chapter 5. Next, the selection criteria used in the analysis is discussed in Chapter 6. Chapter 7 explains the strategies used to model both the Higgs boson signal and the non-resonant diphoton background. An experimental method using Gaussian Processes to reduce the uncertainty associated with the background modeling procedure is detailed in Chapter 8. Next, the statistical model used to process the observed data is explained in Chapter 9. The systematic uncertainties included in the model are detailed in Chapter 10, while the results obtained by the model are presented in Chapter 11. The results of the analysis are discussed in the conclusions, in Chapter 12.

CHAPTER 2

The Higgs Boson in the Standard Model

2.1 The Standard Model

The SM of particle physics describes the fundamental particles which make up the visible universe, as well as how these particles interact with each other [35, 36, 37]. These particles fall into one of two groups: fermions and bosons. The particles of the SM are presented schematically in Figure 2.1. The general structure of these fundamental particles will be presented initially, and a discussion of the underlying forces responsible for said structure will be presented after.

The fermions are defined as having a spin value of $\frac{1}{2}$, and they obey Fermi-Dirac statistics. There exist two families of fermions, quarks and leptons, and within each of these families there are three generations. Fermions from the three generations are very similar, with the main difference being their mass. Those with the largest mass are defined as the third generation, while those with the smallest are defined as the first generation. Within each generation, the quark sector contains two families: the up-type quarks (the up, charm, and top quarks) and the down-type quarks (the down, strange, and bottom quarks). The up-type quarks have an electric charge of $+\frac{2}{3}$, while the down-type quarks have an electric charge of $-\frac{1}{3}$.

Each generation of the lepton sector contains two families. The first family contains the electron, the muon, and the tau particle. The second family consists of the neutrinos. Each neutrino is paired with a member of the first family, and hence they are aptly named the electron neutrino, the muon neutrino, and the tau neutrino. The electron, muon, and tau lepton carry an electric charge of -1, and the neutrinos carry no electric charge. The neutrinos have very small (but non-zero) masses, which have not yet been measured experimentally.

The bosons by definition have integer spin values, and so they obey Bose-Einstein statistics. Two of these bosons, the photon and the gluon, are massless, while the others, the W, Z, and Higgs bosons, are massive. The photon interacts with particles which carry electric charge, and the gluon, the mediator of the strong force, interacts with those which carry color charge. The W and Z bosons (also called the "gauge bosons") mediate the weak force; they couple to all left-handed

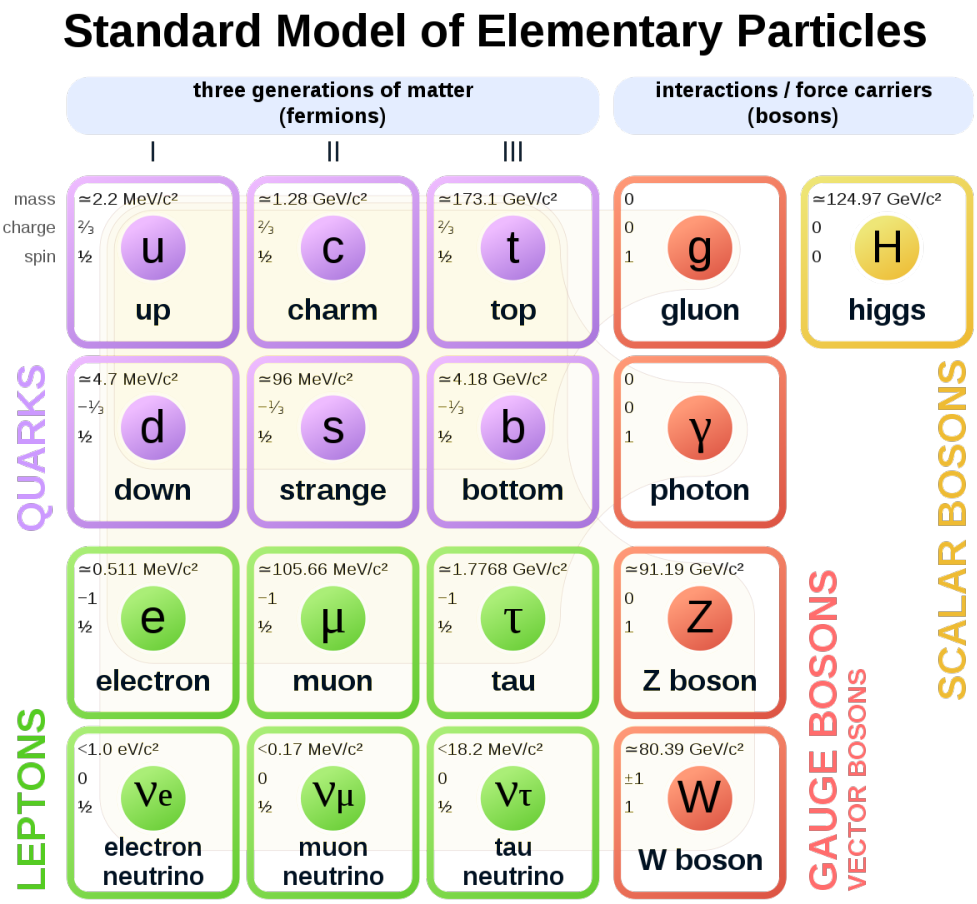


Figure 2.1: A schematic of the fundamental particles described by the Standard Model. Image taken from Ref. [1].

fermions, as well as to each other. The Higgs boson couples to all charged fermions, as well as to the W and Z bosons. This interaction with the Higgs boson is what gives the charged fermions and gauge bosons their mass, as described below.

The Standard Model itself is an example of a quantum field theory. A quantum field theory is a theory describing the behavior of quantized fields, where the fields evolve in time as governed by the Klein-Gordon equation (explained below) while also obeying special relativity. Quantum field theories, as with classical field theories, are often described by a Lagrangian. A Lagrangian equation is often presented as:

$$\mathcal{L} = T - V \tag{2.1}$$

where T represents the kinetic energy of a system and V represents the potential energy. Recalling the Lagrangian of classical mechanics, the kinetic term of a generic object of mass m is generally

given as:

$$T = \frac{1}{2}mv^2 = \frac{1}{2}m\left(\frac{dx}{dt}\right)^2 \tag{2.2}$$

where v is the object's velocity. The potential term depends upon the system being considered, and it may depend not only on the object's mass, but also on its position in spacetime. For example, the classical theory potential for an object (*not* a field) attached to a spring is given as:

$$V(x) = \frac{1}{2}kx^2 \tag{2.3}$$

where k represents the "spring constant" (the strength of the spring) and x represents the objects position (with the un-stretched spring position defined as the origin).

In order to move from a classical Lagrangian for a single particle to a quantum field theory, we will first begin with the relativistic energy relationship:

$$E^2 = p^2c^2 + m^2c^4 \tag{2.4}$$

where E is the total energy of the particle, p is the particle's spatial momentum, m is the particle's mass, and c is the constant speed of light in a vacuum. We may now address the issue of quantization by examining the *quantum operator* forms of E and p:

$$E = -i\hbar\frac{\partial}{\partial t} \qquad\qquad p = -i\hbar\nabla$$

where ∇ is the spatial three-derivative. These operators may be applied to a *quantum state*, which we will define as $\psi(x)$. Substituting the operator forms into Equation 2.4 and applying to the state $\psi(x)$ yields:

$$\begin{aligned}
\left(-i\hbar\frac{\partial}{\partial t}\right)^2\psi(x) &= \left(-i\hbar\nabla\right)^2\psi(x) + m^2c^4\psi(x) \\
-\hbar^2\frac{\partial^2}{\partial t^2}\psi(x) &= -\hbar^2c^2\nabla^2\psi(x) + m^2c^4\psi(x) \\
0 &= \hbar^2c^2\left(\frac{1}{c^2}\frac{\partial^2}{\partial t^2} - \nabla^2\right)\psi(x) + m^2c^4\psi(x)
\end{aligned} \tag{2.5}$$

The above equation may be simplified by defining the spacetime derivatives ∂_μ and ∂^μ:

$$\partial_\mu = \left(+\frac{1}{c}\frac{\partial}{\partial t}, \nabla\right)$$
$$\partial^\mu = \left(-\frac{1}{c}\frac{\partial}{\partial t}, \nabla\right)$$

where μ is the spacetime coordinate. Notable, the two may be multiplied to obtain the *covariant* derivative, ∂^2:

$$\partial^\mu \partial_\mu = \partial^2 = -\nabla^2 + \frac{\partial^2}{\partial t^2}$$

Substituting the covariant derivative into Equation 2.5 gives:

$$\hbar^2 c^2 \partial^2 \psi(x) + m^2 c^4 \psi(x) = 0$$

An additional simplification may be obtained by moving to so-called *natural* units ($c = \hbar = 1$), yielding:

$$\left(\partial^2 + m^2\right) \psi(x) = 0 \tag{2.6}$$

Lastly, we will finally introduce the concept of a field. A field is defined as an output value given some spacetime coordinate. The output may be a single value (a scalar), or it may be a set of values (a vector or matrix). We will simply replace our wave function $\psi(x)$ with a scalar field, $\phi(x)$. This replacement is not trivial, but the details of doing so lie beyond the scope of this work. After the substitution, Equation 2.6 becomes:

$$(\partial^2 + m^2)\phi(x) = 0, \tag{2.7}$$

which is known as the Klein-Gordon equation. The Klein-Gordon equation describes the kinematics of a non-interacting *scalar* field.

As mentioned above, however, the fermions of the SM are not scalars. The fermions are constructed of so-called *Weyl spinor* fields (one form of irreducible vector field which behaves well under Lorentz transforms). A massive, spin-$\frac{1}{2}$ fermion field, denoted as Ψ, can be constructed in two fashions. The first is as two Weyl fields, called a *Dirac* fermion. One of the Weyl fields is denoted the "left-handed" component, while the other is denoted the "right-handed" component. Taking the conjugate transpose of a Dirac fermion yields another distinct Dirac fermion with opposite charge, which is called the "anti-particle" partner of the original Dirac fermion. The second construction is as one neutral Weyl field, called a *Majorana* fermion. Given that there is only one Weyl field, taking the conjugate transpose of a Majorana fermion yields the original fermion; therefore, Majorana fermions are identical to their antiparticles (and hence cannot carry charge). The quarks and charged leptons within the standard model are observed to have oppositely charged antiparticle partners, and therefore they are Dirac fermions. Neutrinos are electrically neutral, but it is currently unknown whether they are of the Dirac or Majorana type.

The kinematics of fermions (both Dirac and Majorana) are described by the Dirac equation:

$$(-i\gamma^\mu \partial_\mu + m)\Psi = (-i\not{\partial} + m)\Psi = 0 \tag{2.8}$$

where γ^μ is a 4×4 matrix (related to the Pauli matrices). The Dirac equation may be manipulated by acting on it with $(i\partial\!\!\!/ + m)$, noting that $\partial\!\!\!/\partial\!\!\!/ = -\partial^2$, giving:

$$0 = (i\partial\!\!\!/ + m)(-i\partial\!\!\!/ + m)\Psi = (\partial\!\!\!/\partial\!\!\!/ + m^2)\Psi = (-\partial^2 + m^2)\Psi. \tag{2.9}$$

This may be recognized as the Klein-Gordon equation from 2.7.

The Lagrangian for a free scalar field can be constructed directly from the Klein-Gordon equation, with a factor of $-\frac{1}{2}$ for convenience:

$$\mathcal{L} = -\frac{1}{2}\partial^\mu\phi\partial_\mu\phi - \frac{1}{2}m^2\phi\phi \tag{2.10}$$

Very loosely comparing the above equation to Equation 2.2 above, one can observe that, again, m represents the mass, and that the terms of Equation 2.10 again represent the kinetic energy of the field. However, free fields are not of particular interest; experimentalists can only measure particles which interact in some way with other particles. To this end, one may consider an example of an additional term in the Lagrangian describing a potential:

$$V = \frac{1}{24}\lambda\phi^4 \tag{2.11}$$

In the above equation, the factor of $\frac{1}{24}$ is simply a combinatoric factor for convenience. Comparing (again, very loosely) to Equation 2.3, it becomes apparent that λ represents the interaction strength described within the term, similar to how k in Equation 2.3 described the strength of a spring. Unlike in the classical analogy, though, the term no longer describes simply the motion of one particle, but rather the interaction between many. Each occurrence of a field in a term indicates its presence in a given point interaction; for example, the ϕ^4 term above indicates that four ϕ particles may interact at a point. These point interactions are called *vertices*, and are often represented pictorially via Feynman diagrams.

2.1.1 Symmetries and Gauge Fields

Consider a quantum field theory with a complex scalar field $\phi(x)$ (and its complex conjugate, $\phi^*(x)$), and a free-field Lagrangian as follows:

$$\mathcal{L} = -\frac{1}{2}\partial^\mu\phi_\mu\partial_\mu\phi^{*\mu} - \frac{1}{2}m^2\phi\phi^* \tag{2.12}$$

Next, we will introduce a constant phase to these two fields such that $\phi(x) \rightarrow e^{i\alpha}\phi(x)$ and $\phi^*(x) \rightarrow e^{-i\alpha}\phi^*(x)$, where α is a constant. The above equation now becomes:

$$\begin{aligned}\mathcal{L} &= -\frac{1}{2}\partial^\mu(e^{-i\alpha}\phi)\partial_\mu(e^{-i\alpha}\phi^*) - \frac{1}{2}m^2(e^{-i\alpha}\phi)(e^{-i\alpha}\phi^*) \\ &= -\frac{1}{2}(e^{i\alpha}e^{-i\alpha})\partial^\mu\phi\partial_\mu\phi^* - \frac{1}{2}m^2(e^{i\alpha}e^{-i\alpha})\phi\phi^* \\ &= -\frac{1}{2}\partial^\mu\phi\partial_\mu\phi^* - \frac{1}{2}m^2\phi\phi^*\end{aligned} \tag{2.13}$$

One can observe that the Lagrangian is *invariant* under this symmetry. This symmetry is a *global* symmetry, meaning that is does not depend on the spacetime coordinate x. The transformation applied ($e^{i\alpha}$) is a single value, and hence a one-dimensional matrix. Its value lies along the unit circle in the Real-Imaginary plane, meaning that this matrix is "unitary." This type of symmetry is defined as a $U(1)$ transformation.

Another type of symmetry, the *special-unitary* or SU group, is especially relevant for describing the Standard Model. The $SU(N)$ group contains $N \times N$ matrices which obey the relation $M^\dagger = M^{-1}$, where the † denotes the conjugate-transpose, and whose determinants are equal to 1 (the "special" part).

Many of the symmetries of the SM are not global symmetries (unlike in the example above) – they are *local* symmetries. To understand the consequences of a local symmetry, one may again consider the example of a complex scalar field $\phi(x)$ with a $U(1)$ symmetry. However, the *local* nature of the symmetry means that the complex phase introduced above will now depend on the spacetime coordinate x: $e^{i\alpha} \rightarrow e^{i\alpha(x)}$. Inserting this transformation into the Lagrangian given in Equation 2.12 yields:

$$\begin{aligned}\mathcal{L} &= -\frac{1}{2}\partial^\mu(e^{i\alpha(x)}\phi)\partial_\mu(e^{-i\alpha(x)}\phi^*) - \frac{1}{2}m(e^{i\alpha(x)}\phi)(e^{-i\alpha(x)}\phi^*) \\ &= -\frac{1}{2}[e^{i\alpha(x)}\partial^\mu\phi + i\phi e^{i\alpha(x)}\partial^\mu\alpha(x)][e^{-i\alpha(x)}\partial_\mu\phi^* - i\phi^* e^{-i\alpha(x)}\partial_\mu\alpha(x)] - \frac{1}{2}m\phi\phi^* \\ &= -\frac{1}{2}e^{i\alpha(x)}e^{-i\alpha(x)}[\partial^\mu + i\partial^\mu\alpha(x)]\phi[\partial_\mu - i\partial_\mu\alpha(x)]\phi^* - \frac{1}{2}m\phi\phi^* \\ &= -\frac{1}{2}[\partial^\mu + i\partial^\mu\alpha(x)]\phi[\partial_\mu - i\partial_\mu\alpha(x)]\phi^* - \frac{1}{2}m\phi\phi^*\end{aligned} \tag{2.14}$$

Notably, the above equation does not match the original Lagrangian. The local $U(1)$ symmetry may be maintained, however, with the introduction of a *gauge field* $A_\mu(x)$. This new field will be incorporated into a new quantity, the *gauge covariant derivative*, as:

$$D_\mu = \partial_\mu + iqA_\mu \tag{2.15}$$

where q is a constant. The $A_\mu(x)$ field is required to transform under the local $U(1)$ symmetry as: $A_\mu(x) \rightarrow A_\mu(x) - \frac{1}{q}\partial^\mu\alpha(x)$. Looking at the kinetic term of Equation 2.14 in terms of the new covariant derivative yields:

$$\begin{aligned}\mathcal{L}_{kin} &= -\frac{1}{2}D^\mu\phi D_\mu\phi^* \\ &= -\frac{1}{2}[\partial^\mu + iqA^\mu]\phi[\partial_\mu - iqA_\mu]\phi^*\end{aligned} \tag{2.16}$$

Now, taking again a local $U(1)$ transformation yields:

$$\begin{aligned}\mathcal{L}_{kin} &= -\frac{1}{2}[\partial^\mu + iq(A^\mu - \frac{1}{q}\partial^\mu\alpha)]e^{i\alpha}\phi[\partial_\mu - iq(A_\mu - \frac{1}{q}\partial_\mu\alpha)]e^{-i\alpha}\phi^* \\ &= -\frac{1}{2}[e^{i\alpha}\partial^\mu\phi + ie^{i\alpha}\phi\partial^\mu\alpha + iqA^\mu e^{i\alpha}\phi - ie^{i\alpha}\phi\partial^\mu\alpha] \\ &\quad \cdot[e^{-i\alpha}\partial_\mu\phi^* - ie^{-i\alpha}\phi^*\partial_\mu\alpha - iqA_\mu e^{-i\alpha}\phi^* + ie^{-i\alpha}\phi^*\partial_\mu\alpha] \\ &= -\frac{1}{2}[\partial^\mu\phi + iqA^\mu\phi][\partial_\mu\phi^* - iqA_\mu\phi^*]e^{i\alpha}e^{-i\alpha} \\ &= -\frac{1}{2}[\partial^\mu + iqA^\mu]\phi[\partial_\mu - iqA_\mu]\phi^* \\ &= -\frac{1}{2}D^\mu\phi D_\mu\phi^*\end{aligned} \tag{2.17}$$

This example demonstrates how the introduction of a gauge field through a gauge covariant derivative may be used to maintain a local symmetry. The kinematics of a gauge field are described using the quantity $F_{\mu\nu}$ (called the "field strength tensor"):

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu \tag{2.18}$$

and the Lorentz-invariant kinetic term in the Lagrangian of such a theory is:

$$\mathcal{L}_{kin} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} \tag{2.19}$$

which can similarly be shown to obey the local $U(1)$ gauge symmetry. Notably, adding a mass term to the gauge boson would violate the symmetry:

$$\begin{aligned}\mathcal{L} &= -\frac{1}{2}m^2(A^\mu - \frac{1}{q}\partial^\mu\alpha)(A_\mu - \frac{1}{q}\partial_\mu\alpha) \\ &= -\frac{1}{2}m^2[A^\mu A_\mu - \frac{2}{q}A^\mu\partial_\mu\alpha + \frac{1}{q^2}\partial^\mu\alpha\partial_\mu\alpha] \\ &\neq -\frac{1}{2}m^2A^\mu A_\mu\end{aligned} \tag{2.20}$$

Therefore, gauge bosons generally cannot have mass.

The SM is described in terms of local symmetry groups as:

$$SU(3) \times SU(2) \times U(1) \tag{2.21}$$

$SU(N)$ groups will result in $n^2 - 1$ generators, and each generator corresponds to a gauge field (boson) of that sector. The $SU(3)$ portion describes the *strong* sector. The strong sector contains eight types of massless gluons, corresponding to different color charge states (permutations of three colors, red, blue, and green, and their anti-colors). The $SU(2) \times U(1)$ portion describes the *electroweak* sector. The three bosons emerging from the $SU(2)$ correspond to the W_μ^a gauge field (where a may be 1, 2, or 3) and the $U(1)$ group gives the B_μ gauge field representing *hypercharge*. Note that the W_μ^a and B_μ fields do not directly correspond to the electroweak bosons described above (the $W^\pm$, Z, and γ).

2.1.2 Symmetry Breaking – An Introductory Example

The existence of the SM Higgs boson is a result of the *spontaneous breaking* of the $SU(2) \times U(1)$ symmetry of the electroweak sector. To better understand the mechanism of spontaneous symmetry breaking, one may return to the example given in Equation 2.13 involving a complex scalar field with a global $U(1)$ symmetry. However, the Lagrangian will now include a non-zero potential term:

$$\mathcal{L} = -\frac{1}{2}\partial^\mu\phi\partial_\mu\phi^* - \frac{1}{2}\varsigma^2\phi\phi^* - \frac{1}{2}\lambda(\phi\phi^*)^2 \tag{2.22}$$

where λ and ς are constants. This potential energy still obeys the global $U(1)$ symmetry. In order to obtain the equations of motion of a particle from the Lagrangian, one may apply the Eurler-Lagrange equation:

$$\frac{\partial\mathcal{L}}{\partial\phi} - \partial_\mu\frac{\partial\mathcal{L}}{\partial(\partial_\mu\phi)} = 0 \tag{2.23}$$

Substituting in the Lagrangian of Eqution 2.22 gives:

$$-\frac{1}{2}\varsigma^2\phi^* - \lambda\phi^*(\phi\phi^*) = -\frac{1}{2}\partial_\mu\partial^\mu\phi^* \tag{2.24}$$

The left side of the above equation is the potential term, and minimizing it gives the "vacuum" solution:

$$\varsigma^2\phi^* + 2\lambda\phi^*(\phi\phi^*) = 0 \tag{2.25}$$

Examining the above equation reveals an obvious solution where $\phi = 0$ and $\phi^* = 0$. However, the above equation also results in the following relation:

$$\sqrt{\phi^*\phi} = \sqrt{-\frac{\varsigma^2}{2\lambda}} \tag{2.26}$$

In the case where $\varsigma^2 < 0$, there exists another minimum in the potential, lower than 0 (the minimum of the $\phi^* = 0$ solution). The field values satisfying the above relation are $\phi = e^{i\alpha}\sqrt{\frac{-\varsigma^2}{2\lambda}}$ and $\phi^* = e^{-i\alpha}\sqrt{\frac{-\varsigma^2}{2\lambda}}$. Notably, the minimum is not a point, but rather a circle in α. The value $\sqrt{-\frac{\varsigma^2}{2\lambda}}$ can be redefined as v, the *vacuum potential*. Although the minimum may be symmetric, there may only be one true vacuum; a value of α must be chosen (for convenience, $\alpha = 0$ will be used here). This choice "spontaneously" breaks the theory's symmetry in α.

Now that a vacuum has been chosen, the fields ϕ and ϕ^* may be expanded about the new vacuum as:

$$\begin{aligned} \phi(x) &= \frac{1}{\sqrt{2}}[v + \rho(x) + i\beta(x)] \\ \phi^*(x) &= \frac{1}{\sqrt{2}}[v + \rho(x) - i\beta(x)] \end{aligned} \tag{2.27}$$

where ρ and β are both real. Expanding the Lagrangian in terms of the above fields yields:

$$\mathcal{L} = -\frac{1}{2}\partial_\mu\rho\partial^\mu\rho - \frac{1}{2}\partial_\mu\beta\partial^\mu\beta - \lambda v^2\rho^2 - \lambda v\rho(\rho^2 + \beta^2) - \frac{1}{4}\lambda(\rho^2 + \beta^2)^2 \tag{2.28}$$

Analyzing the above terms shows that the β field is massless, while the ρ field has a non-zero mass equal to:

$$m_\rho^2 = 2\lambda v^2 = 2\lambda(\frac{-\varsigma^2}{2\lambda}) = -\varsigma^2 \tag{2.29}$$

The β field is called a Goldstone boson – a massless boson arising from the "flat" direction of a spontaneously broken symmetry.

The resulting Lagrangian differs in a number of important ways from the original presented in Equation 2.22. First, there is no longer an explicit $U(1)$ symmetry in the resulting Lagrangian – the symmetry has been broken. Secondly, the original ϕ field its complex conjugate have been replaced by two fields ρ and β, which have distinctly different kinematics. The original theory did not contain any three-particle vertices, which appear in the $\lambda v\rho(\rho^2 + \beta^2)$ term above.

2.1.3 The Standard Model Electroweak Sector

As previously stated, the W^a_μ and B_μ fields (the $SU(2) \times U(1)$ generators), do not directly correspond to the $W^\pm$, Z, and γ bosons. The $W^\pm$, Z, and γ bosons all emerge as a result of the spontaneous breaking of the $SU(2) \times U(1)$ symmetry. Additionally, the existence of the SM Higgs boson is predicted through this mechanism.

Before digging further into this spontaneous symmetry breaking in the SM, an overview regarding the nature of the electroweak sector should be considered. The electrons, muons, and taus are Dirac fermions with left- and right-handed components, which can be respectively denoted as e_L and e_R. The same is true of their antiparticle partners, denoted as $\bar{e}_L$ and $\bar{e}_R$, respectively. Neutrinos, however, have only been observed as left-handed particles, denoted as ν_L. The left-handed (non-neutrino) leptons and the left-handed neutrinos form a doublet L, while the right-handed (non-neutrino) leptons form a singlet R:

$$L = \begin{pmatrix} \nu_L \\ e_L \end{pmatrix} \qquad\qquad R = (e_R) \tag{2.30}$$

As demonstrated in Section 2.1.1, conserving a local $U(1)$ symmetry in a theory requires a gauge field. The gauge field introduced by the SM $U(1)$ symmetry is the B_μ field, which transforms as:

$$B_\mu \to B_\mu + \frac{1}{g'}\partial_\mu \beta \tag{2.31}$$

where g' is the coupling to the *weak hypercharge* (similar to the magnitude of the electric charge, ignoring its sign) and β is the phase applied as part of the symmetry (analogous to $\alpha(x)$ in Section 2.1.1). The leptons will transform under the $U(1)$ phase shift as $(L, R) \to e^{\frac{iY\beta}{2}}(L, R)$. The quantity Y is the weak hypercharge, which takes on a value of -1 for the components of the L doublet and -2 for the component of the R singlet:

$$L = \begin{pmatrix} \nu_L \\ e_L \end{pmatrix} \to e^{\frac{-i\beta}{2}} \begin{pmatrix} \nu_L \\ e_L \end{pmatrix} \qquad\qquad R = (e_R) \to e^{-i\beta}(e_R) \tag{2.32}$$

The conservation of the $SU(2)$ symmetry requires an additional gauge field, W_μ. The transformation of this field is somewhat more complicated than that of the B_μ field:

$$\tau \cdot W_\mu \to \tau \cdot W_\mu + \frac{1}{g}\tau \cdot (\partial_\mu \alpha) - \tau \cdot (\alpha \times W_\mu) \tag{2.33}$$

Here, the phase α is a 3×3 matrix, and τ is a row vector of the three Pauli matrices. Similarly, the fields themselves will transform under the symmetry as $(L, R) \to e^{iI\tau\cdot\alpha}(L, R)$, with I representing

weak isospin. Interestingly, the weak isospin is 0 for the component of R (it is $\frac{1}{2}$ for the components of L). Physically, this means that the $SU(2)$ portion of the SM breaks *chiral symmetry* (it treats left- and right-handed particles differently). Explicitly, L and R transform as:

$$L = \begin{pmatrix} \nu_L \\ e_L \end{pmatrix} \rightarrow e^{\frac{-i}{2}\tau\cdot\alpha} \begin{pmatrix} \nu_L \\ e_L \end{pmatrix} \qquad\qquad R = (e_R) \rightarrow (e_R) \tag{2.34}$$

As in the example from Section 2.1.1, a covariant derivative must be introduced along with the gauge field:

$$\partial_\mu \rightarrow D_\mu = \partial_\mu - \frac{i}{2} g' Y B_\mu - igI\tau \cdot W_\mu \tag{2.35}$$

Applying the covariant derivatives with the appropriate hypercharge and weak isospin values yields the result:

$$\begin{aligned} \mathcal{L} &= \overline{R} i\gamma^\mu (\partial_\mu + ig' B_\mu) R + \overline{L} i\gamma^\mu (\partial_\mu - \frac{i}{2} g\tau \cdot W_\mu + \frac{i}{2} g' B_\mu) L - \frac{1}{4} G^{(W)}_{\mu\nu} \cdot G^{(W)\mu\nu} - \frac{1}{4} F^{(B)}_{\mu\nu} F^{(B)\mu\nu} \\ &= i\overline{R}\gamma^\mu D_\mu R + i\overline{L}\gamma^\mu D_\mu L - \frac{1}{4} G^{(W)}_{\mu\nu} \cdot G^{(W)\mu\nu} - \frac{1}{4} F^{(B)}_{\mu\nu} F^{(B)\mu\nu} \end{aligned} \tag{2.36}$$

where:

$$F^{(B)}_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu \tag{2.37}$$

and

$$G^{(W)}_{\mu\nu} = \partial_\mu W_\nu - \partial_\nu W_\mu + g W_\mu \times W_\nu \tag{2.38}$$

One notable quality of the above Lagrangian is that it does not allow for the fermions to have mass. A mass term would take a form proportional to $m\overline{\Psi}\Psi$, which involves mixing of the left- and right-handed components. For the electron, for example, the mass term expands to $m(\bar{e_L} e_R + \bar{e_R} e_L)$. Such a term does not obey the $SU(2)$ symmetry, because this symmetry would transform the two components differently.

2.1.4 Symmetry Breaking in the Standard Model

The Lagrangian of Equation 2.36 above describes the kinematics of the electroweak sector before the spontaneous breaking of the $SU(2) \times U(1)$ symmetry. The symmetry breaking of the SM is similar to that shown in Section 2.1.2; a complex scalar field will "fall" into a non-zero vacuum configuration. This complex scalar field will be donoted as ϕ, and it has a hypercharge of $Y = +1$ and $I = +\frac{1}{2}$. The potential of this field will be the same as that in Section 2.1.2: $-\frac{1}{2}\varsigma^2 \phi^\dagger \phi - \frac{1}{2}\lambda(\phi^\dagger\phi)^2$, and its kinematic term will be the covariant derivative above (with the appropriate Y

and I), $D^\mu\phi^\dagger D_\mu\phi$. The fermions will also be coupled to this scalar with the interaction form:

$$G_e(\overline{L}\phi R + \overline{R}\phi^\dagger L) \tag{2.39}$$

where G_e is the coupling strength. As in the previous example, the minimum of the potential is not at $\phi = 0$, but rather at $(\phi^\dagger\phi) = \left(-\frac{\varsigma^2}{2\lambda}\right) = v^2$. The ϕ field can be redefined through a gauge transformation as:

$$\phi = \begin{pmatrix} 0 \\ v + \frac{1}{\sqrt{2}}h \end{pmatrix} \tag{2.40}$$

Inserting the new broken symmetry representation of ϕ leads to a number of interesting consequences. The covariant derivative portion can be expanded to:

$$D^\mu\phi^\dagger D_\mu\phi \rightarrow \frac{1}{2}\partial_\mu h\partial^\mu h + \frac{1}{4}g^2v^2(W_\mu^1)^2 + \frac{1}{4}g^2v^2(W_\mu^2)^2 + \frac{1}{4}v^2[(gW_\mu^3 - g'B_\mu)^2] + ... \tag{2.41}$$

A closer look at the above equation reveals that there appear to be two new mass terms for W_μ^1 and W_μ^2 – these are new, massive particles ($m_W = \frac{1}{4}g^2v^2$ for both). These particles are in fact the positively charged and negatively charged W bosons which can be observed in our universe. Before analyzing the $\frac{1}{4}v^2[(gW_\mu^3 - g'B_\mu)^2]$ term, it is useful to define the *Weinberg angle*, $\theta_W = \tan^{-1}(g'/g)$. The two B_μ and W_μ^3 fields may be simplified by defining two new fields, Z_μ and A_μ, which are proportional to B_μ and W_μ^3, as follows:

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

With this redefinition, the last term of Equation 2.41 becomes:

$$\frac{1}{4}v^2[(gW_\mu^3 - g'B_\mu)^2] \rightarrow \frac{g^2v^2}{4\cos^2(\theta_W)}Z_\mu Z^\mu \tag{2.43}$$

The Z_μ field is the Z boson, and its mass is $m_Z = m_W/\cos(\theta_W)$.

Returning to the terms of Equation 2.39, one may observe that their expansion leads to the following:

$$G_e(\overline{L}\phi R + \overline{R}\phi^\dagger L) \rightarrow -G_e(\overline{e}_L v e_R + \overline{e}_R v e_L) - G_e\left(\overline{e}_L\frac{h}{\sqrt{2}}e_R + \overline{e}_R\frac{h}{\sqrt{2}}e_L\right) \tag{2.44}$$

The first term on the right hand side is a mass term – $G_e v$ is a constant, and it appears alongside the field "squared." The charged leptons therefore obtain their masses through the Higgs mechanism. Additionally, there is an interaction term involving the Higgs boson and the charged leptons.

The interaction strength and the particle's mass are both proportional to the same constant (G_e), which indicates that heavier particles couple more strongly with the Higgs boson. Although not shown here explicitly, the quarks also obtain their mass through coupling to the Higgs boson. The proportionality between the mass and coupling strength to the Higgs boson holds for the quarks, too.

The redefinition of the B_μ and W_μ^3 fields in terms of Z_μ and A_μ can be continued into the W_μ, B_μ, and fermion interaction terms as follows:

$$\begin{aligned} i\overline{R}\gamma^\mu D_\mu R + i\overline{L}\gamma^\mu D_\mu L \rightarrow\ & i\overline{e}\gamma^\mu\partial_\mu e + i\overline{\nu}\gamma^\mu\partial_\mu\nu - g\sin(\theta_W)\overline{e}\gamma^\mu e A_\mu \\ & + \frac{g}{\cos(\theta_W)}[\sin^2(\theta_W)\overline{e}_R\gamma^\mu e_R - \frac{1}{2}\cos(2\theta_W)\overline{e}_L\gamma^\mu e_L + \frac{1}{2}\overline{\nu}\gamma^\mu\nu]Z_\mu \\ & + \frac{g}{2}[\overline{\nu}\gamma^\mu e_L(W_\mu^{(1)} - iW_\mu^{(2)}) + \overline{e}_L\gamma^\mu\nu(W_\mu^{(1)} + iW_\mu^{(2)})] \end{aligned} \quad (2.45)$$

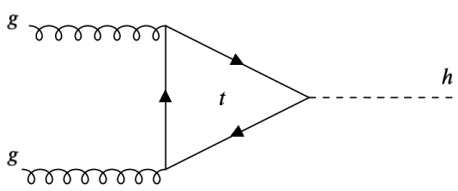
The above terms yield a number of insights. First, the charged leptons interact with the new A_μ field, which has not picked up a mass from the symmetry breaking. This A_μ is the massless photon of electromagnetism. Second, all of the leptons interact with the Z boson. Lastly, the W bosons interact only with the left-handed leptons (as expected).

2.1.5 The Higgs Boson – An Experimentalist's Perspective

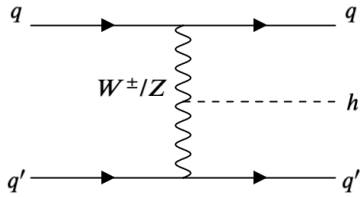
As one may see, the SM gives a detailed picture of the behavior of the fundamental forces and particles which make up our universe. Thus far, the predictions of the SM have been overwhelmingly confirmed by experiment. However, the SM does appear to have its shortfalls. For example, the SM does not explain the observed expansion of the universe (there is no "dark energy" force), and it does not contain any particle consistent with dark matter. These remaining questions suggest that there exists some physics beyond the SM, but how its effects will manifest are still a mystery.

This work focuses on the measurement of the coupling strength of the SM Higgs boson to other SM particles. The processes analyzed do not contain any inherent Beyond Standard Model (BSM) effects. However, as the most recently-discovered fundamental particle, the behavior of the Higgs boson has not been measured with a very high precision (compared to theoretical precision). New physics could affect the Higgs boson couplings, causing their observed values to differ from the SM prediction. Therefore, improving the precision of Higgs boson measurements may either constrain or implicate the existence of new BSM physics.

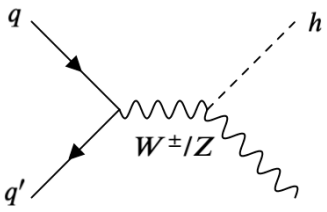
To understand the behavior of the Higgs boson, one may study its couplings to other particles by measuring the cross-sections of different production modes. There exist many ways to produce a Higgs boson; the four ways most accessible to experimentalists are (in order of most common to least common at a proton-proton collider): gluon-gluon fusion (ggF), vector-boson fusion (VBF),



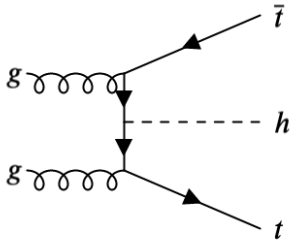
(a) Gluon-Gluon Fusion



(b) Vector-Boson Fusion



(c) Vector-Boson Associated



(d) Top Associated

Figure 2.2: The four main production modes of the Higgs boson at a proton-proton collider: (a) gluon-gluon fusion (ggF), (b) vector-boson fusion (VBF), (c) vector-boson associated production (VH), and (d) top-associated production ($t\bar{t}H$). These Feynman diagrams were generated using the Feynman diagram maker tool, available from Ref. [2].

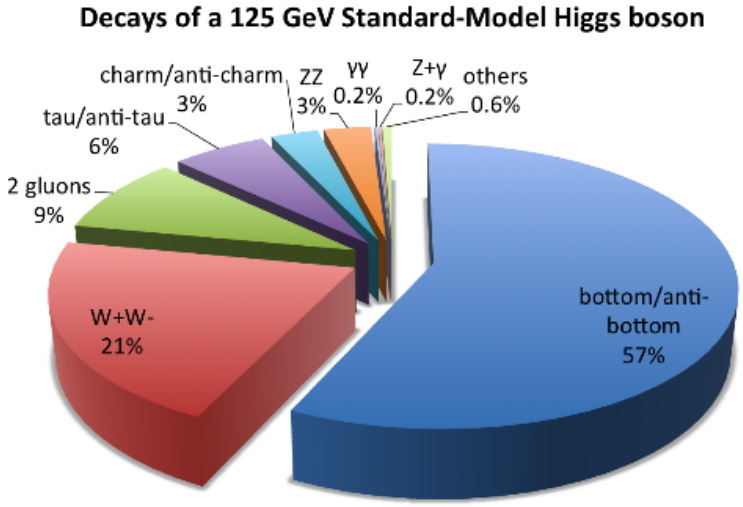
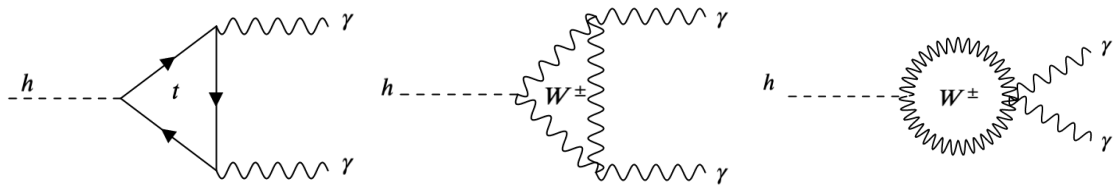


Figure 2.3: A pie chart of the decay channels of the Higgs boson, as well as their branching ratios. Image from Ref. [3].



(a) Top Quark Loop (b) $W^{\pm}W^{\pm}\gamma$ Loop (c) $W^{\pm}W^{\pm}\gamma\gamma$ Loop

Figure 2.4: The three dominant one-loop processes contributing to the diphoton decay mode of the Higgs boson. Diagram (a) shows the Higgs boson decaying via a top quark loop, which radiates two photons. Diagrams (b) and (c) show the Higgs boson decaying through a $W^{\pm}$ loop, emitting two photons in either (b) two $W^{\pm}W^{\pm}\gamma$ vertices or in (c) one $W^{\pm}W^{\pm}\gamma\gamma$ vertex. These Feynman diagrams were generated using the Feynman diagram maker tool, available from Ref. [2].

vector-boson associated production (VH), and top associated production ($t\bar{t}H$). The Feynman diagrams of these four production modes are shown in Figure 2.2.

The ggF process involves a fermion loop. All of the fermions (quarks and massive leptons) contribute to the loop, but due to the significantly larger coupling of the Higgs boson to the top quark, the contribution from the top-quark dominates. The second and third production modes, VBF and VH, involve the vertex with a Higgs boson and two $W^{\pm}$ or Z bosons. In the VBF process, the two vector-bosons are radiated from quarks, and "fuse" to produce the Higgs boson. In VH, the Higgs boson is radiated from a $W^{\pm}$ or Z boson, which also is a part of the final state. The last production mode, $t\bar{t}H$, is similar to the ggF process but with the fermion loop "opened up" to produce a top/anti-top quark pair in the final state. Although the coupling of the Higgs boson to the top quark is strong, this production mode is suppressed due to kinematic constraints; producing a top/anti-top quark pair and a Higgs boson requires a significant amount of energy from the incoming quarks.

The decay modes of the Higgs boson are also of interest, as they define the experimentalists' approach to finding Higgs boson events. The primary decay channels are listed in Figure 2.3, along with their branching ratios (the predicted fraction of Higgs bosons which will decay through each mode). The most common decay mode of the Higgs boson is to a $b\bar{b}$ quark pair (note that the $t\bar{t}$ decay is kinematically disallowed, since $m_H < m_{t\bar{t}}$). Searches for $H \to b\bar{b}$ in a proton-proton collider are complicated by the difficulty in identifying and measuring the b-quark decay products (jets, as discussed in more detail in Section 4.2) with sufficient accuracy and precision. Additionally, many unrelated QCD processes produce $b\bar{b}$ pairs, which may be difficult to disentangle from b-quarks originating from Higgs boson decays.

Two other decay channels of the Higgs boson, $H \to \gamma\gamma$ and $H \to ZZ^* \to 4\ell$ (where ℓ is an electron or muon), may be targeted for precision measurements, despite their small branching ratios. The branching ratio for $H \to ZZ^*$ is approximately 3%; this rate is further reduced due to

the small branching ratio ($< 7\%$) of an individual Z boson to an e^+e^- or $\mu^+\mu^-$ pair [38]. However, the ATLAS detector is capable of measuring with very high precision the energies of muons, through a dedicated Muon Spectrometer, and electrons, through the electromagnetic calorimeter (discussed in detail in Chapters 3 and 4).

The branching ratio of the $H \to \gamma\gamma$ decay is only 0.227% [13]. The Higgs boson is neutral, and so it does not directly couple to photons. This decay instead requires an additional loop of charged particles, primarily the top quark and the $W^{\pm}$ boson, which emit two photons [39]. The dominant loop processes of the diphoton decay mode are shown in Figure 2.4. As in the case of the $H \to ZZ^* \to 4\ell$ mode, the two photons of the diphoton decay process can be identified with high accuracy, and their energies can be measured with a high precision within a calorimeter. More detailed information on the measurement and reconstruction of photons in the ATLAS detector is provided in Chapters 3 and 4.

In addition to the advantage of precision measurement capabilities of the final state decay products in the $H \to \gamma\gamma$ decay mode, this process also benefits from having relatively well-understood backgrounds. The primary background (continuum $\gamma\gamma$ evens produced through strong and electromagnetic interactions) has a smooth, falling diphoton invariant mass spectrum. The Higgs boson has a predicted width of only about 4 MeV [38], and so its decay products are expected to have an invariant mass very close to the Higgs boson mass. This narrow width means that the Higgs decay will appear as a narrow "bump" on top of a smooth, continuous background spectrum.

CHAPTER 3

The ATLAS Detector

The Large Hadron Collider (LHC) [40] at CERN is currently the most powerful particle accelerator in operation. With a circumference of about 27 km, it straddles the Swiss-French border and lies approximately 100 m underground. As of the start of Run 2 in 2015, the LHC has reached a center-of-mass proton-proton collision energy of $\sqrt{s} = 13$ TeV, with collisions occurring every 25 ns. The peak instantaneous luminosity (a measure of the number of expected collision events per unit time per unit area) has varied significantly during Run 2, from $0.55 \times 10^{34}\text{cm}^{-1}\text{s}^{-1}$ in 2015 to 1.9×10^{34} $\text{cm}^{-1}\text{s}^{-1}$ in 2018 (the last year of Run 2) [41]. Notably, the peak instantaneous luminosity reached roughly twice the design instantaneous luminosity of 1.0×10^{34} $\text{cm}^{-1}\text{s}^{-1}$. By of the conclusion of Run 2, the LHC has delivered an *integrated* luminosity of 156 fb^{-1} of collision data to the ATLAS experiment, which in turn has recorded 147 fb^{-1}.

Along the LHC ring sit four main experiments: the ATLAS [6] experiment; the Compact Muon Solenoid (CMS) [32] experiment; A Large Ion Collider Experiment (ALICE) [42]; and the Large Hadron Collider beauty (LHCb) [43] experiment.

The ATLAS detector is a general purpose detector which sits at Point 1 along the ring of the LHC. The term "general purpose" refers to the fact that the ATLAS detector is capable of recording a large variety of physics signatures, and therefore data taken by the detector can be used to study many differing physics processes (such as studying the Higgs boson, or searching for new particles not predicted by the SM). The detector is approximately 25 meters tall and 44 meters long. It sits about 100 meters underground (in order to shield it from cosmic ray backgrounds). The detector can be broken down roughly into four parts, moving from the collision point outwards: the inner detector, which records the tracks of charged particles close to the interaction point with a very fine granularity; the electromagnetic calorimeter, which measures the energy deposited from showering photons and electrons; the hadronic calorimeter, which measures the energy deposited by showering hadrons (composite particles made of two or more quarks); and the muon system, which measures the tracks of muons. In addition, two magnets, one a solenoid and the other a toroid, generate a strong magnetic field in the detector volume which causes the trajectories

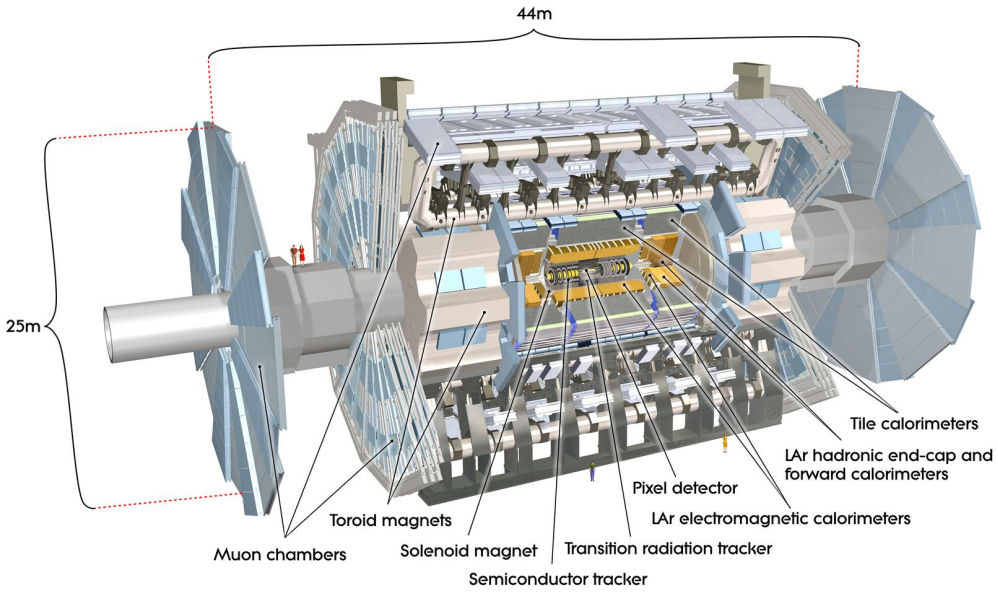


Figure 3.1: A cut-out illustration of the ATLAS detector. The cut out allows the various sections of the detector to be seen.

of charged particles to bend. The magnitude and direction of bending may be used to infer the momenta of charged particles. These sub-detector systems, their relevant technologies, and the magnets are described in more detail the following sections.

The ATLAS detector is designed to be roughly cylindrically symmetric along the z-axis, which corresponds to the line of the LHC beams. The beam collision point is located at the nominal center of the cylinder, which is defined as $z = 0$. The y-axis has been designated the "upward" direction, and the x-axis has been designated as the direction towards the center of the LHC ring. This choice means that the positive z side of the detector is that which sits counterclockwise of the collision point. This side is labeled as the "A" side, and the negative z side is labeled as the "C" side. The detector is nominally forward-backward symmetric. A cut-out view of the detector can be viewed in Figure 3.1.

3.0.1 Inner Detector

The ID consists of a number of tracking detectors which record the tracks (primarily central tracks, with a pseudorapidity range of $|\eta| < 2.5$) of charged particles emitted from particle collisions. A cutout view of the ID is presented in Figure 3.2. It is composed of Silicon Pixel and SCT detectors. Silicon is a semiconductor, which allows it to be *doped*, meaning that impurities may be introduced which hold a significant charge. Positively charged regions are denoted as p regions,

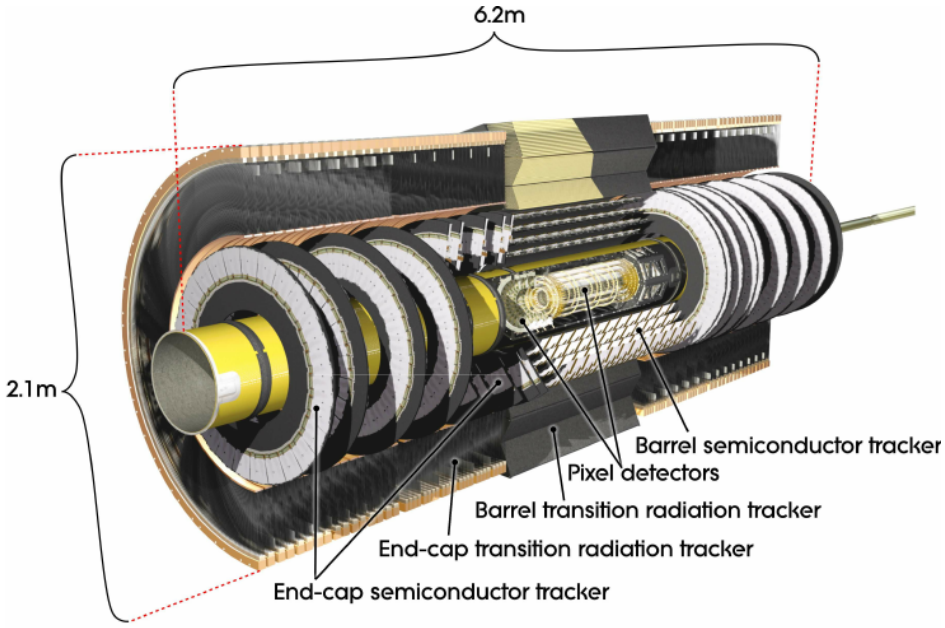


Figure 3.2: A cut-out illustration of the ATLAS Inner Detector (ID). The cut out allows the various sections of the ID to be seen. The "semiconductor tracker" labels refer to the Silicon Microstrip (SCT) detectors.

which negatively charged ones are denoted as *n* regions. A *junction* is defined as the interface between a *p* region and an *n* region. A large voltage bias can be applied across a junction, which will push out any free charge-carriers in the material [38]. High-energy charged particles passing through the material will release a number of charge carriers due to ionization. These free charges can be collected and measured by electronics bonded to the silicon. One advantage of silicon detectors are that they can be manufactured to be very thin, meaning that ionized charges travel to the readout electronics very quickly (often on the order of around 10 ns). The small size also allows for precision measurements of the position of the incoming particle. The ATLAS pixel detectors are formed with small *n+* ("extra" negative regions) implants within an *n*-type bulk [44], while the SCTs are comprised of *p*-type strips on an *n*-type bulk [45].

Both the pixels and strips provide a precise measurement of the position where a charged track has crossed. The pixel detectors offer a finer granularity due to their smaller size (50×400 μm^2 in $R\phi \times z$ at minimum), and as such are placed closest to the expected interaction point. The SCT detectors operate on the same principle as the pixels, but instead of having a square geometry, they are 12 cm long rectangles (in the barrel) or wedges (in the endcaps) in order to cover more area per channel. The longer dimension of the SCTs in the barrel region is close to that parallel to the beamline, while in the endcaps, the longer dimension is close to that of the radial dimension. The SCTs are offset by a stereo angle of 40 mrad, which allows for position measurements along the

z axis. The ID contains three layers of pixels and four layers of SCTs in the barrel region, and in the endcap region, it contains three layers of pixel detectors and nine layers of SCTs .

During the shutdown period between Run 1 and Run 2, significant upgrades were made to the ID [46]. The Insertable B Layer (IBL) is a layer of silicon pixel detectors which sits 33.25 mm from the beamline, providing a significant gain in tracking precision. Additionally, the Minimum Bias Trigger Scintillators (MBTS) were added to the endcaps in order to improve triggering capabilities for forward particles. The MBTS are made of doped plastic scintillating material, and they are arranged in one ring of eight segments in ϕ within each endcap.

The pixel and SCT detectors of the ID are surrounded by the Transition Radiation Tracker (TRT). The TRT measures transition radiation – the radiation emitted when relativistic charged particles traverse across a material boundary in which there is a significant difference in the index of refraction [38]. The amount of energy radiated as transition radiation is dependent upon how relativistically a particle is traveling. Lighter particles (such as electrons) will therefore emit more transition radiation than will heavier ones. The TRT also may discriminate photons from electrons, as photons will not emit transition radiation.

The TRT consists of layers of Polyimide straw tubes, which are 4 mm in diameter, arranged in planes. The 73 planes in the barrel region run parallel to the beam direction. The 160 planes in the endcap regions are divided between two sets of wheels of radially-oriented straws on each side. The inner set contains twelve wheels, and each wheel contains twelve layers of straws with an inter-layer spacing of 8 mm. The outer set of wheels contains eight wheels of eight straw layers, with an inter-layer distance of 15 mm.

3.0.2 Solenoid Magnet

The entirety of the ID is surrounded by a superconducting solenoid magnet, which is designed to provide a magnetic field of 2 T within the volume of the ID [4]. The solenoid magnet sits in front of the ATLAS calorimetry system, and as such is designed to provide a large magnetic field while also being compact in size. The magnet uses an Al-stabilised NbTi conductor in a single coil, with 1154 turns over the coil. The magnetic field provided by the solenoid will cause charged particles in the ID to bend depending on their charge, mass, and momentum, and so the amount and direction of bending may be used as an input for particle reconstruction. Because the magnetic field provided by the solenoid runs along the z axis, the resulting deflection of charged particles will be in the ϕ direction (the sign depending on the particle's charge). A field map of the magnetic field in the ATLAS detector is shown in Figure 3.3

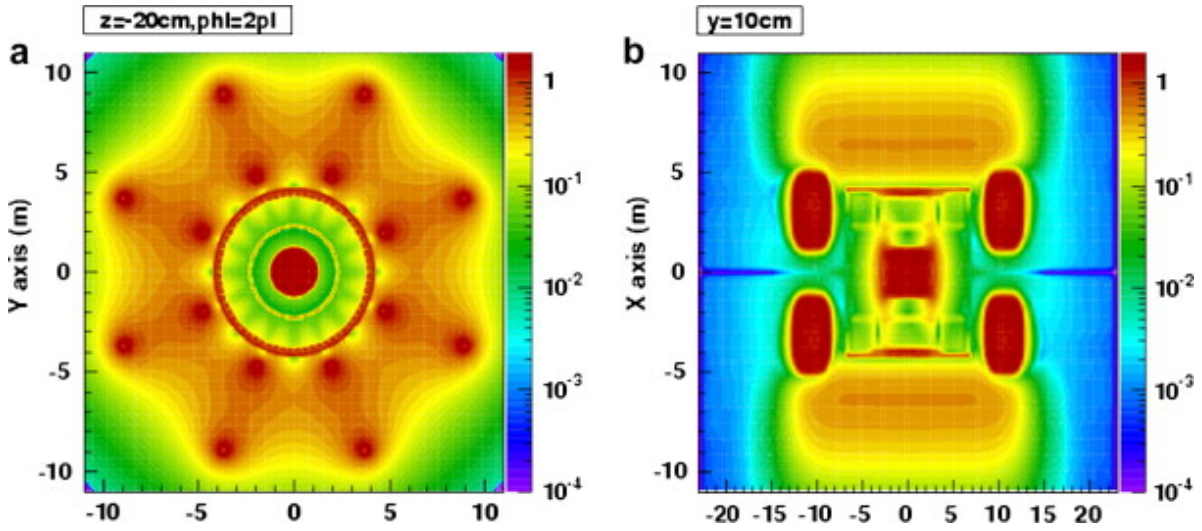


Figure 3.3: A mapping of the magnetic field strength within the ATLAS detector, as presented in Ref. [4]. The left panel shows the field strength within a cross section of the barrel region of the detector in the transverse plane, where the endcap toroid has been removed. The right panel shows a longitudinal section of the detector, including the endcaps. The central area with a large magnetic field is from the solenoid magnet, while the outer four areas of large field are from the endcap toroid coils.

3.0.3 Electromagnetic Calorimeter

The Electromagnetic (EM) calorimeter, shown in Figure 3.4, is designed to measure the energy deposited by electrons and photons through electromagnetic showers. It is a sampling calorimeter, meaning that it has alternating layers of absorber material (designed to induce showering) and active material (where measurements of deposited energy are made) [38]. Sampling calorimeters suffer from larger uncertainties on the total energy deposited compared to calorimeters made of entirely active (measuring) material. However, the absorber materials in sampling calorimeters generally have much shorter radiation lengths (the characteristic length scale of a particle shower) than do active materials. The presence of absorbing materials reduces the overall size of the EM showers being measured, therefore reducing the necessary volume (and cost) of an EM calorimeter. The ATLAS EM calorimeter uses Liquid Argon (LAr) sampling layers and lead absorbing layers. The large number of protons in the lead nucleus means that the atom's electron cloud is well populated, greatly increasing the chances that an incoming high-energy electron or photon will interact via electromagnetism.

The EM calorimeter is broken into six portions, and the total pseudorapidity range of the EM calorimeter is $|\eta| < 3.2$. The barrel consists of two portions, which are separated by a gap of 4 mm at $z = 0$. The two half-barrels each have a pseudorapidity coverage of up to $|\eta| < 1.475$. Each

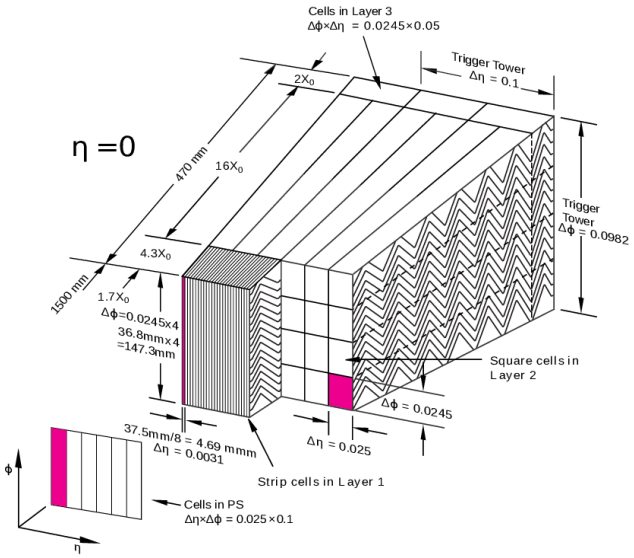


Figure 3.4: An illustration of one segment of the ATLAS EM Calorimeter, as presented in Ref. [5]. The jagged lines propagating outwards radially show the so-called accordion geometry of the detector.

endcap contains an inner wheel and outer wheel portion, which respectively cover the pseudorapidity ranges of $1.375 < |\eta| < 2.5$ and $2.5 < |\eta| < 3.2$. The absorber and sampling layers are arranged in an accordion fashion, with the folds in the barrel region resembling a compression in the ϕ direction. In the endcap region, the folds resemble a compression in the z direction. The accordion geometry allows for signals to be read out at the front or back of the calorimeter, which allows the detector to remain azimuthally symmetric [47].

The barrel region of the EM calorimeter contains three layers in the central range of $|\eta| < 1.35$ and two layers in the slightly more forward range of $1.35 < |\eta| < 1.475$. The endcap region contains three layers in the region of $1.375 < |\eta| < 1.5$ and two layers in the remainder of its pseudorapidity range. A presampling layer is also positioned before the first layer of the EM calorimeter in the range of $|\eta| < 1.8$. The primary goal of the presampler detector is to detect electromagnetic radiation from interactions of electrons and photons which occur upstream of the EM calorimeter, such as in the solenoid magnet.

3.0.4 Hadronic Calorimeter

Hadronic showers result from a combination of strong and electromagnetic interactions between an incoming hadron (a composite state of two or more quarks) and the surrounding material [38]. Hadronic showers tend to have much larger characteristic length scales than do electromagnetic showers, and due to the complex nature of low-energy strong force interactions, they are sig-

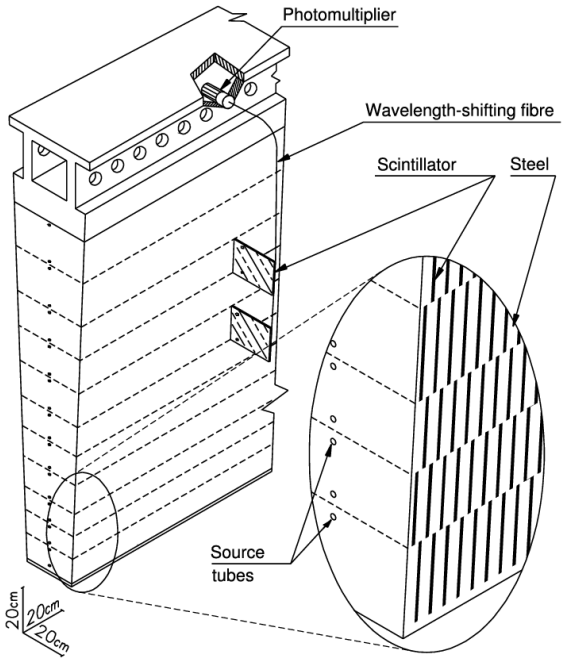


Figure 3.5: An illustration of one wedge of the ATLAS Tile Calorimeter, as presented in Ref. [6]. A closeup of the layout of the scintillating material within the steel showering material is shown in the oval to the right.

nificantly more difficult to model. The energy resolution of a hadronic calorimeter is therefore typically worse than that of an electromagnetic calorimeter, as is the case for the ATLAS detector.

The Hadronic Calorimeter consists of three distinct subsystems: the Tile calorimeter, the LAr Hadronic Endcap Calorimeter (HEC), and the LAr Forward Calorimeter (FCAL). It is designed to measure the energy of showers from hadronic particles (such as protons and neutrons). Similar to the EM calorimeter, all three hadronic calorimeter subsystems are sampling calorimeters.

The sampling layers of the Tile calorimeter are made of a scintillating material, polystyrene doped with PTP and POPOP for their wavelength shifting properties. The showering material is steel, which also provides structural support. It covers the inner region of $|\eta| < 1.7$ and is further subdivided into three subsections, a central barrel and two extended barrels. Each barrel subsection contains 64 modules (or wedges), with each wedge covering 5.625 degrees in ϕ. The scintillator panels in each section are trapezoidal in shape in order to cover a constant range in ϕ across their range in r. They are 3mm thick in the z direction and are placed in a staggered configuration among the steel absorber plates (as shown in Figure 3.5). Two wavelength-shifting fibers are attached to each scintillator tile in order to readout the scintillated photons from hadronic showers. The light carried by each fibre is read out by two Photomultiplier Tubes (PMTs) on opposite sides of the tile in order to provide redundancy in the measurement, as well as to provide an estimation of the

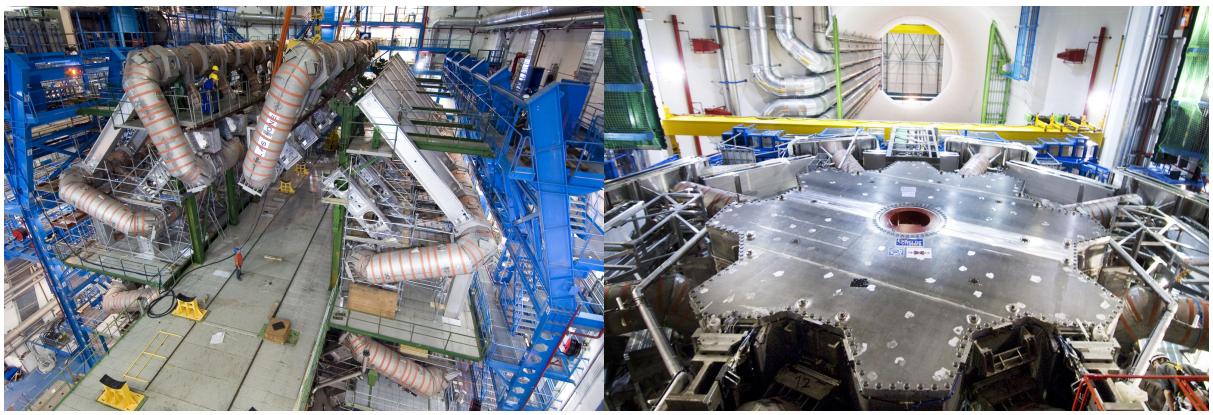


Figure 3.6: Photos of (a) the barrel toroid magnet and (b) one of the endcap toroid magnets for the ATLAS detector. Photos taken from Ref.s [7] and [8], respectively.

impact parameter of the shower.

The HEC uses liquid argon as its active material. The absorber material chosen is copper, and it covers the pseudorapidity range of $1.5 < |\eta| < 3.2$. It consists of two wheels in each endcap (the "front wheel" and the "rear wheel"), and each wheel comprises two longitudinal subsections. Both the front and rear wheels are divided into 32 wedges in the azimuthal direction.

The FCAL provides measurements of very forward electromagnetic and hadronic showers, in the range of $3.1 < |\eta| < 4.9$. The FCAL contains three modules in each endcap, labeled "FCAL1" (the closest to the interaction point), "FCAL2," and "FCAL3" (the furthest from the interaction point). FCAL1 is designed to primarily measure electromagnetic showers, which FCAL2 and FCAL3 are focused for hadronic showers. All three FCAL modules use LAr sampling material. The sampling material is arranged in thin layers perpendicular to the beamline direction. Using thin layers prevents excessive buildup of charged particles which may be induced due to the FCAL's forward position (and hence large particle flux). FCAL1 utilizes copper as an absorber, while FCAL2 and FCAL3 use Tungsten as the primary absorber, along with a copper "shielding plug" mounted behind the modules to reduce any energy leakage into the muon system. Read-out is performed by evenly-spaced electrode rods made of copper (in FCAL1) or tungsten (in FCAL2 and FCAL3) running in the z-direction.

3.0.5 Toroid Magnet

A trio of superconducting toroid magnets, one in each endcap and one in the barrel region, provides an additional magnetic field with strength of approximately 1 T [4] within the volume of the Muon Spectrometer (MS) (described in Section 3.0.6). These magnets sit outside of the calorimeters, but inside the muon subsystem. They are air-core toroids, and each contains eight coils running in the z-direction (the plane made by each coil lies in the $R - z$ plane, and hence the magnetic field lines

run in the ϕ direction). The purpose of the toroid magnets is to bend the tracks of muons, which allows for their momentum to be measured directly through their curvature. Unlike in the case of the solenoid magnet, the deflection of tracks caused by the toroid magnet lies in the $R - z$ plane, perpendicular to the deflection caused by the solenoid magnet in the inner part of the detector. The barrel and encap and toroid magnets can be seen in Firgure 3.6, and a map of the magnetic field in the ATLAS detector is shown in Figure 3.3

3.0.6 Muon Spectrometer

The ATLAS MS sits on the outermost portion of the detector. It consists of four main types of detectors: Monitor Drift Tubes (MDTs), Cathode Strip Chambers (CSCs), Thin Gap Chambers (TGCs), and Resistive Plate Chambers (RPCs)). The first two types are used for precisely measuring muon tracks, while the latter two are used for triggering.

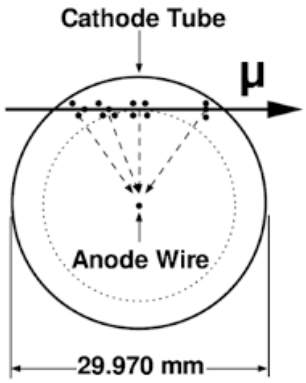


Figure 3.7: A cross section of a MDT, used as part of the ATLAS MS. Image from Ref. [6]

The MDTs are used as the primary precision muon detector; a cross section of an MDT is shown in Figure 3.7. They consist of an aluminum tube 2.997 cm in diameter with a wire running along the center of the tube. A voltage difference of 3080 V is applied between the tube and the wire, and the tube is filled with a gas mixture of 93% Argon and 7% Carbon Dioxide. Muons passing through the tubes will ionize the contained gas, and the freed electrons will collect on the wire. The drift time versus position of the electrons in the gas volume is calculable (and ranges up to a maximum of about 700 ns), and so the time of arrival compared to that of the bunch crossing can be used to determine the minimum distance between the muon track and the central wire.

Because multiple tube hits are needed to precisely reconstruct muon tracks, many MDTs are arranged together in so-called chambers, the exact geometry of which depends on the location within the detector. Almost all chambers consist of two "multilayer's' of MDTs separated by an empty space of 6.5-317 mm, where each multilayer contains either three or four individual layers of tubes. The chambers are arranged in sixteen segments in ϕ, alternating between so-called large

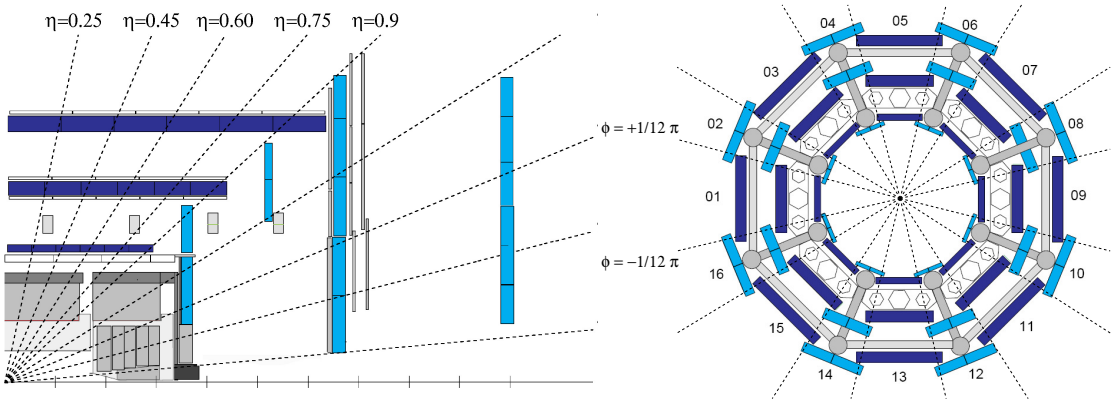


Figure 3.8: The geometry of the MDT chambers (in blue), where (a) shows the segmentation in $|\eta|$ and (b) shows the segmentation in ϕ. Image taken from Ref. [9].

and small chambers. In both the barrel and endcap regions, chambers are placed such that a muon will pass through three separate chambers, labeled the inner, medium, and outer chambers. MDT coverage extends to $|\eta| < 2.7$, although this full range is only covered by the middle endcap layer. A schematic of the MDT chamber placement is shown in Figure 3.8.

Due to the larger expected rate close to the beampipe, CSCs are used in place of MDTs to provide coverage of the range of $2 < |\eta| < 2.7$ in the inner endcap chambers. Similar to the MDT chambers, the CSC chambers are divided into sixteen segments in ϕ, also with alternating large and small sized chambers. The CSCs operate in a manner similar to the MDTs in that they rely on the ionization of an Argon-CO_2 gas. An electric field is applied across the gas volume, causing the ionized electrons drift towards the four radially-oriented anode within each CSC chamber. Each endcap contains two wheels of CSCs, and each wheel contains eight chambers in the azimuthal direction (with evenly-spaced gaps between chambers). The chambers on the wheel closer to the collision point contain "small" CSC chambers, while the outer wheels contain "large" CSC chambers. The chambers overlap such that the small chambers cover the gaps between the large chambers.

Each of the sixteen CSC chambers contain eight layers of cathode strips. Four of the layers are finely segmented along the η direction, with 192 segments per layer, while the other four are finely segmented in the ϕ direction, with 48 segments per layer [48]. Different levels of charge will be induced on the anode strips, depending on their proximity to the muon track, which allows for the precise determination of the muon's track through the detector. In order to improve precision, the detector is set with a pitch angling the face of the detector towards the interaction point. The pitch is 5.31 mm in the large chambers and 5.56 mm in the small chambers.

Two different triggering detectors types are used in ATLAS: RPCs in the barrel ($|\eta| < 1.05$) and TGCs in the endcaps. These two systems provide very fast but coarse track information to be

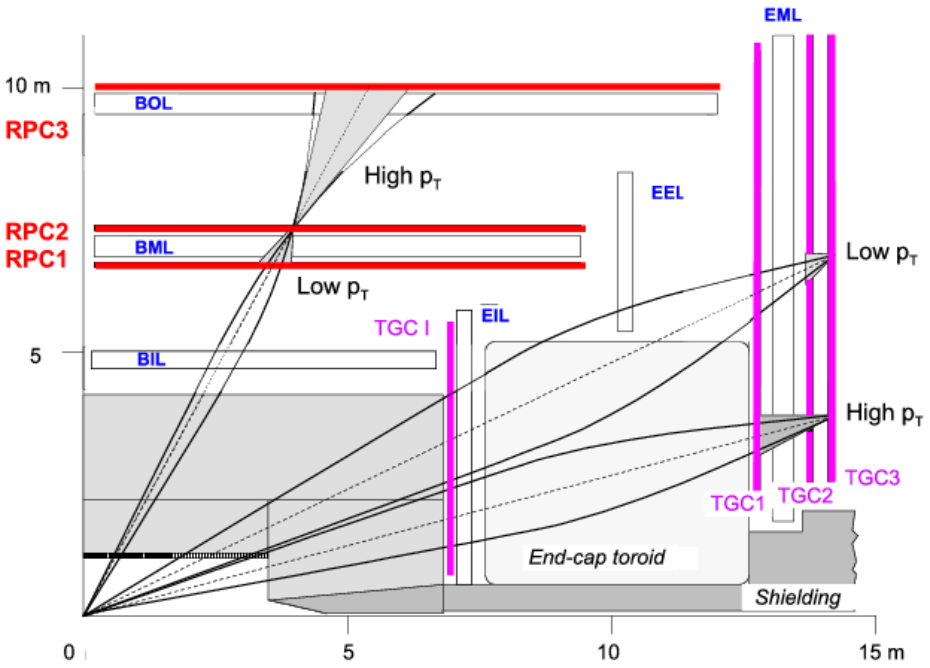


Figure 3.9: The arrangement of the ATLAS MS trigger detectors. Image from Ref. [6]

used in the online trigger system, and as such their segmentation may be much larger than for the MDTs and CSCs. The placement of the TGCs and RPCs is illustrated in Figure 3.9.

The RPCs are gaseous detectors, where the gas is contained within a thin volume (2mm wide) between two sheets of resistive material [49]. The outer surface of each resistive plates is coated in a thin layer of conductive graphite paint, which allows the plates to be held at a large voltage differential. Muons passing through the plates will ionize gas particles in the thin volume, which will induce a charge through capacitive coupling on copper strips arranged on the outside of the resistive plate. The RPCs are segmented in ϕ in the same manner as the MDT chambers, and three layers of RPCs are used for triggering. The first two layers are located on either side of the middle MDT chambers, and the last layer is situated just outside the outer MDT chambers.

For the endcap region ($1.05 < |\eta| < 2.4$), TGCs are used for triggering due to increased rate demands (the rate capability of the RPCs is lower than that of the TGCs, due to the large area covered per plate in the RPCs). The TGCs are also used to better determine the azimuthal coordinate of muon tracks in the endcaps, where the radial coordinate is more precisely measured by the MDTs. Three chambers of TGCs are placed adjacent to the middle MDT chamber in the endcap region to form the "Big Wheel." One sits just inside the chamber, and two just outside, with respect to the interaction point. An additional chamber of TGCs is placed just inside the inner MDT chamber in the endcap. The TGCs are arranged in either a doublet (two layers of gas volume) or triplet (three layers of gas volume). The inner of the Big Wheel TGC chambers is of

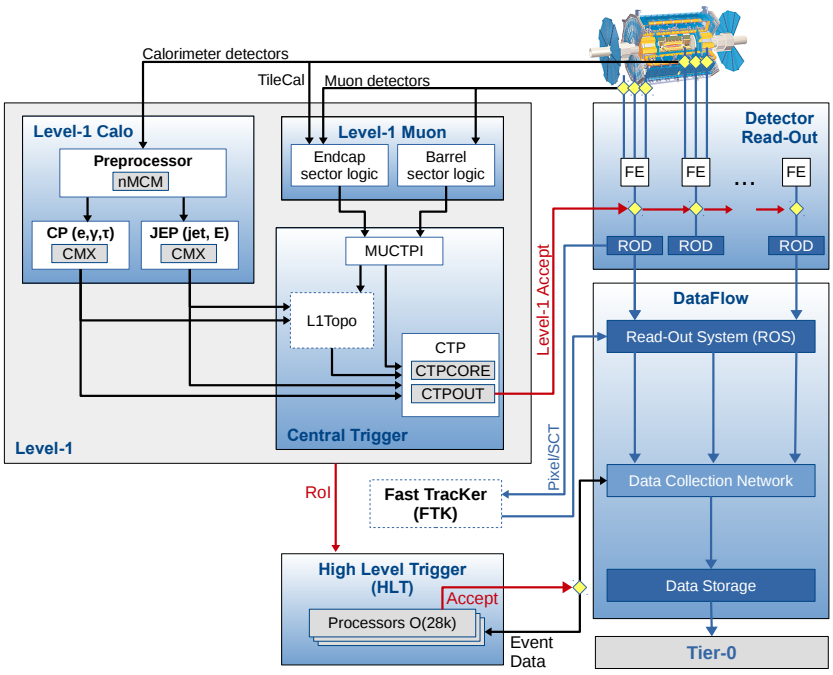


Figure 3.10: A schematic of the ATLAS Trigger System. Image from Ref. [10]

the triplet type, while the others are doublets. Unlike the MDTs, the TGCs are subdivided into twelve segments in ϕ in the Big Wheel.

3.0.7 Trigger System

In Run 2, the LHC collided proton bunches every 25 ns, giving a bunch crossing frequency of 40 MHz. The ATLAS trigger system [50, 10] is designed to buffer and filter events in real time in order to reduce the event recording rate to approximately 200 Hz to conform to data storage and writing limitations. The trigger system is designed as a three-tiered system, with large buffering capabilities to hold events while the trigger algorithms are performed. As the trigger level increases, the algorithms performed become more complex, and hence more costly in terms of time. The lowest level is defined as L1, the second as L2, and the third as the Event Filter. The L1 trigger system is a hardware-based system. Together, L2 and the Event Filter comprise the High Level Trigger (HLT), and they are both software-based systems. The trigger system consists of many individual "options" to indicate an event of interest; these options make up the "trigger menu."

The L1 trigger relies on coarse information from the various subdetector systems to assess whether an event contains an object of interest. Possible high energy muons are identified using the muon system trigger subdetectors discussed in Section 3.0.6, the RPCs and the TGCs. These subsystems and their L1 menu comprise the "L1Muon" triggers. The "L1Calo" trigger uses information from the electromagnetic and hadronic calorimeters, though not at the full granularity of either calorimeter. If either a high energy muon or calorimeter deposit matching the requirements outlined in the L1 trigger menu is found, then the trigger sector of that object is marked as a Region of Interest (RoI). Information from L1Muon and L1Calo is combined with inputs from a number of other subdetector systems (such as the MBTS, described in Section 3.0.1) at the L1 Central Trigger Processor (CTP). The CTP contains another L1 trigger menu, "L1Topo," which analyzes the combined information of the L1Muon and L1Calo triggers in order to identify more global quantities of interest, such as missing transverse energy ($E_{\mathrm{T}}^{\mathrm{miss}}$).

If an event passes the L1 trigger, it is then buffered before being sent to the HLT. In order to avoid overfilling the event buffer, the CTP applies some preventative "dead time" - limits on the number and frequency of L1 acceptances. This dead time takes on two forms: "simple" (an L1 acceptance cannot come within a given time of a previous L1 acceptance) and "complex" (only a set number of L1 acceptances are allowed for a given number of bunch crossings).

In Run 1, the L2 and Event Filter were housed at separate computing farms. Prior to Run 2, the two systems were merged into a single farm, meaning that the HLT functioned more as one integrated unit. The L2 trigger analyzes the RoI marked by the L1 trigger in greater detail, using the full granularity of the subdetector systems within that RoI. If the available information meets one of the criteria of the L2 trigger menu, then the recorded data of the entire event (not just the RoI) is passed to the Event Filter for further analysis. The full event reconstruction at the HLT includes information from the ID, and online track reconstruction algorithms are applied to both tracks in the ID and the MS. Online identification algorithms of physics objects, such as photons, electrons, and b-tagged jets (jets originating from the decay of a high energy bottom quark) are also applied at the HLT stage.

Finally, events which pass one of the options of the HLT menu are recorded. The menus may be changed between runs to account for different data-taking conditions. Some triggers in the menus are pre-scaled, meaning that only a subset of events passing that trigger menu criteria are saved. Pre-scaling allows for the recording of certain event types for which the event rate would otherwise be too high to record.

The triggers of particular relevance to this work are described in further detail in Section 5.1.

CHAPTER 4

Object Reconstruction and Identification in ATLAS

Signals from the ATLAS detector do not generally translate directly into the physics objects, i.e. photons, electrons, muons, and jets, used by physics analyses. They must be reconstructed from the detector signals, usually from a combination of subdetector systems. An overview of object reconstruction is provided below. A diagram summarizing basic particle identification is provided in Figure 4.1.

4.1 Tracks and Vertices

Tracks are reconstructed from energy deposits in the silicone pixel and SCT detectors within the ID [51]. The energy from a charged track is deposited across multiple pixels, and so individual deposits are combined into clusters by grouping neighboring pixels with energy deposited above a set threshold. Because of the alternating nature of the silicone strips, the energy deposits from two layers of strips are combined to obtain the spatial coordinate, called the space-point, of the track. An individual pixel is sufficient to obtain a track's space point. Once a collection of space points within the ID has been determined, three such points are combined to form track seeds. An iterative combinatorial method determines the underlying tracks from the available track seeds with high efficiency and good rejection of random combinatorial tracks.

Vertices are reconstructed by fitting the charged tracks [52]. At least two tracks must be associated with a vertex for it to be considered. Additionally, associated tracks must satisfy the following cirteria: the track transverse momentum (p_{T}) must be greater than 400 MeV; the track's absolute pseudorapidity ($|\eta|$) must be less than 2.5; and the track must have at least 9 (11) hits in the silicone pixels and SCT detectors for $|\eta| < 1.65$ ($|\eta| > 1.65$), as well as at least one hit in the IBL or B-Layer. Additional requirements on the track quality are applied, as well. The so-called "primary vertex" (or vertex of interest) is defined nominally as that with the greatest $\sum(p_{\mathrm{T}})^2$ of the associated tracks. Note that a different definition of the primary vertex is used for diphoton

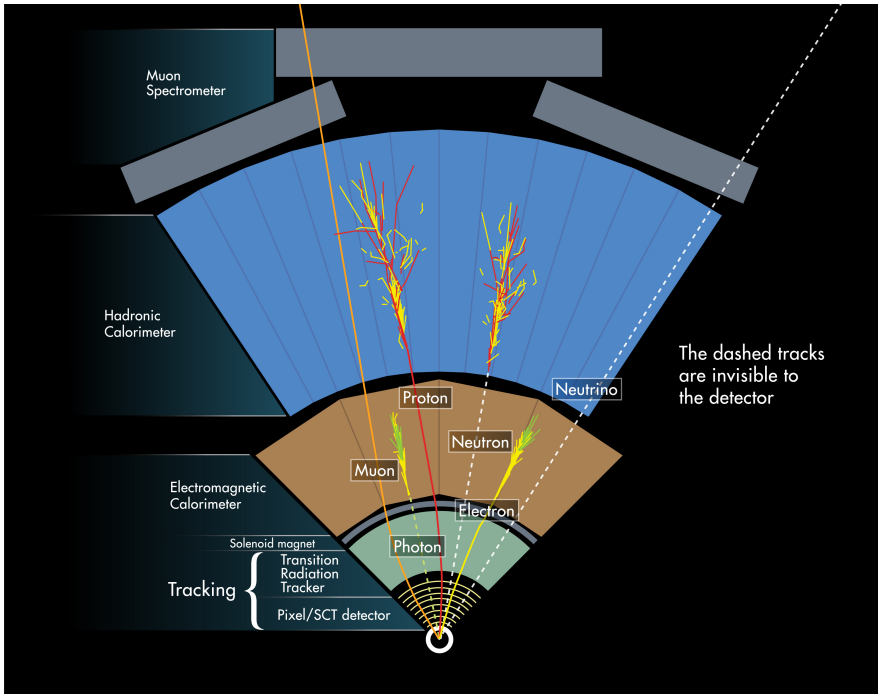


Figure 4.1: The different combinations of detector signatures left by a variety of particles. Figure taken from Ref. [11].

events, discussed in Section 6.1. Tracks originating from vertices other than the primary vertex are considered to be pileup and are not considered for analysis purposes.

4.2 Jets

Jets are reconstructed from *topological clusters* [53] (also called topo-clusters) of energy deposited in both the EM and hadronic calorimeters. Electrons and photons are similarly constructed from topo-clusters; the construction of a topo-cluster from the energy deposited in individual calorimeter cells is detailed extensively in Section 4.4 below. An individual topo-cluster may contain energy deposited by multiple jet constituents, and sub-jet information such as the internal distribution of energy may be relevant for physics analysis. Therefore, topo-clusters with multiple local maxima satisfying certain criteria are split into two topo-clusters. To be split, a local maxima must have at least 500 MeV of energy deposited in that cell, and it must have at least four neighboring cells available for comparison. The separation of local maxima occurs in two steps. In the first step, local maxima from only the second and third layers of the EM calorimeter and the first layer of the forward calorimeter are considered. In the second step, additional maxima from the first layer of the EM calorimeter, the hadronic calorimeter, and the remainder of the forward calorimeter. Once

the topo-clusters have been defined and separated, they are then clustered into jet objects using the anti-k_t algorithm using a radius parameter of 0.4 [54, 55].

4.3 Muons

Muons are reconstructed primarily from information measured by the muon spectrometer by combining segments (small pieces of a muon track measured by multiple hits within a chamber) of hits from the four muon subdetector systems (MDTs, CSCs, RPCs and TGCs) [56]. Segments in the MDTs are obtained by fitting a straight line to the hits recorded in the tubes from the multilayer in each MDT chamber. These segments, along with corresponding hits from the nearby trigger chambers, are then passed through a Hough transform [57], which translates the straight lines in the $x-y$ plane (that orthogonal to the bending induced by the toroid magnet) into points in Hough $r-\phi$ space. Segments in the CSCs are constructed by passing hits from the multilayer to a combinatorial algorithm scanning the $\eta-\phi$ plane (that of the bending induced by that toroid magnet). The hits measured by the trigger systems (the RPCs and TGCs) will give the coordinate in the orthogonal direction to the bending.

Once the track segments from the different subdetector components have been determined, full muon tracks are constructed by combining at least two segments (or one high-quality segment if that segment lies in the barrel-endcap transition region). Initially, only track segments from the middle layer of the MS are used as track seeds for a combinatorial algorithm; track segments from the inner and outer layers are used after. A χ^2 algorithm assigns individual hits to a track. A dedicated overlap removal algorithm determines whether a track segment may be shared by multiple muon tracks, which may occur in the case of close-by muons.

Once muon tracks have been reconstructed in the MS, they are matched to corresponding signals in the other subdetector components (the ID and the calorimeters). Depending on which subdetector systems provide relevant information for the muon, four different muon types are defined: Combined Muon (CBM), Segment Tagged Muon (STM), Calorimeter Tagged Muon (CTM), and Extrapolated Muon (ME). CBMs are those for which the track has been reconstructed in both the ID and the MS. The tracks for one system are extrapolated to match a track in the other system (normally, those from the MS are extrapolated to the ID, but the inverse may be utilized as well). A global fit is performed on the combined track, which may result in the addition or subtraction of MS hits in order to improve the overall track quality. STMs are those constructed from ID tracks which are matched to one or more track segments in the MS, but not to a full MS track. CTMs are constructed by matching ID tracks to calorimeter energy deposits consistent with those of minimum ionizing particles, and hence they do not use inputs from the MS. This type of muon is considered in order to increase muon efficiency in the region of $|\eta| < 0.1$, where the MS is not

fully instrumented, and it is the least pure of the four types. The last type, MEs, are constructed from MS tracks which are not matched to an ID track, but which are loosely compatible with the primary vertex. This type mostly accounts for very forward muons in the range of $2.5 < |\eta| < 2.7$, whose tracks lie outside the acceptance of the ID.

4.4 Photons and Electrons

Photons and electrons are reconstructed from topological clusters of energy deposited in the EM calorimeter [53, 58]. Topological clusters are seeded by calorimeter cells in the second and third layers of the EM Calorimeter in which the energy deposited is greater than four standard deviations above the expected noise level of that cell. Once seed cells are defined, neighboring cells from all EM calorimeter layers (including the presampling layer) in which the energy deposited is greater than two standard deviations above the expected noise level of that cell are iteratively added to the cluster. Lastly, the neighboring cells surrounding the defined topo-clusters are included. Clusters which have two local maxima are considered to be two separate topo-clusters, and all topo-clusters must have at least 400 MeV of deposited energy. The splitting of local maxima into separate topo-clusters is described in detail in Section 4.2.

Once topo-clusters have been defined within the EM calorimeter, they are matched to tracks from the ID, with preference given to tracks in which the track seed falls within the RoI of the topo-cluster. The candidate track is extrapolated to the EM calorimeter to asses its compatibility with the topo-cluster. The extrapolation is done using two possible definitions of the particle's momentum: the first being the measured track momentum, and the second being scaled to match the energy deposited in the topo-cluster. The latter option accounts for cases in which the particle loses significant energy from Bremsstrahlung radiation, which would otherwise lead to a mismatch of the momenta estimated from the track and the calorimeter. Once the track has been extrapolated, it is defined as matched if it satisfies the following criteria:

$$|\Delta\eta| < 0.05$$

and

$$-0.10 < q \times (\phi_{track} - \phi_{cluster}) < 0.05$$

where q is the charge of the candidate charged particle (essentially, the projected track must hit close to the location of the energy deposit).

Prompt electrons and positrons will appear in the detector as an EM calorimeter deposit with one associated track. Unconverted photons carry no charge, and so no track will be matched to the EM calorimeter deposit. Converted photons are photons which have converted into an electron-

positron pair in the detector volume (due to interactions with the detector material). They appear as two clusters, each with an associated track, and generally have a reconstructable conversion vertex. These converted photon candidates are required to have two oppositely-charged tracks, which may be either in the silicon tracker or in the TRT, and the tracks must be compatible with a conversion vertex from a massless particle. Converted photons may also be reconstructed from a single track, where no hits in the innermost layers are present. In both cases, tracks must be determined to be very likely from an electron by the TRT, in order to reduce the contamination of other charged tracks. In cases where multiple conversion vertices appear consistent with the converted photon candidate, those reconstructed from two silicone detector tracks are prioritized over those formed from one or more TRT track or from only a single track. Preference is also given to vertices consistent with a smaller radius of conversion.

In Run 2, the reconstruction of photons and electrons has been improved through the use of superclusters, which allow energy from secondary showers (such as from a Bremsstrahlung photon) to be captured and factored into energy reconstruction. The topo-clusters discussed above are used to seed superclusters. To be defined as a supercluster seed for an electron (photon), the topo-cluster must have at least 1 GeV (1.5 GeV) of energy deposited. Candidate electron topo-clusters must also satisfy basic tracking requirements: the track must have at least four hits in the silicons pixels or strips. Additional topo-clusters are defined as satellites of a supercluster seed if they fall within a window of $\Delta|\eta| \times \Delta\phi = 0.075 \times 0.125$ of the seed. For electron candidates, additional topo-clusters within a wider window of $\Delta|\eta| \times \Delta\phi = 0.125 \times 0.300$ may also be defined as satellites if their matched track is the same as that of the seed supercluster. Converted photon candidates may also be assigned additional satellite topo-clusters if their conversion vertices were formed from tracks in the silicone detectors; the matched track of the satellite topo-cluster must match the seed's conversion vertex. Once all topo-clusters clusters have been assigned, the combined seed and satellites are defined as a supercluster. The superclusters are limited to a maximum size of $\Delta\eta < 0.075$ ($\Delta\eta < 0.125$) in the barrel (endcap) in order to reduce the impact of pileup topo-clusters being assigned to the supercluster.

An additional offline "identification" algorithm is applied to both photons and electrons; the photon identification algorithm will be detailed extensively due to the importance of photons in the analysis. The purpose of the identification algorithm is to reject "fake" photon candidates, which are primarily charged jets mis-reconstructed as photons.

Three primary identification working points are defined: *loose*, *medium*, and *tight*. The *loose* photon identification requirement is based on the leakage from the electromagnetic shower into the hadronic portion of the calorimeter, as well as the shower shape observed in the second layer of the EM calorimeter. The *loose* requirement uses the following shower shape variables:

- R_{had} (the ratio of transverse energy deposited in all layers of the Hadronic calorimeter to that

deposited in the cluster in the EM calorimeter, for clusters in the pseudorapidity range of $0.8 < |\eta| < 1.37$)

- R_{had_1} (the ratio of transverse energy deposited in the first layer of the Hadronic calorimeter to that deposited in the cluster in the EM calorimeter, for clusters outside the pseudorapidity range of $0.8 < |\eta| < 1.37$)

- R_η (the ratio of the energy deposited in a 3×7 ($\eta \times \phi$) cell rectangle to that deposited in a 7×7 cell rectangle, where both rectangles are centered on the cell with the most deposited energy)

- w_{η_2} (the lateral shower width $\sqrt{\frac{\sum_i E_i \eta_i^2}{\sum_i E_i)} - (\frac{\sum_i E_i \eta_i}{\sum_i E_i})^2}$, where E_i is the energy deposited in cell i and η_i is the pseudorapidity of cell i, calculated in a 3×5 ($\eta \times \phi$) cell rectangle).

The *medium* requirement (used by the diphoton trigger, described in Section 5.1) uses the same variables as described above, as well as an additional variable:

- E_{ratio} (the ratio of the difference in energy deposited between the maximum deposit and the second largest deposit within a cluster to the sum of the energy in those two deposits).

Both the *loose* and *medium* working points are optimized in bins of photon pseudorapidity, but neither is optimized separately for converted and unconverted photons.

The *tight* selection includes the requirements of the *loose* working point, and it adds information from the finely segmented strip layer of the calorimeter. The *tight* working point is separately optimized for unconverted and converted photons, to account for the generally broader lateral shower profile of the latter. As for the other two working points, it is also optimized in bins of photon pseudorapidity.

For the portion of the analysis measuring the top-associated Higgs boson production mode, the *tight* identification cut requirement was re-optimized in sub-ranges of the photon's transverse energy (E_{T}), as the photon's electromagnetic shower depends significantly on its energy. Using the E_{T}-dependent *tight* photon identification cuts, photons with $p_{\text{T}} > 25$ GeV are expected to have a reconstruction and identification efficiency of greater than 82%, comparable to that of the non-E_{T}-dependent definition. Overall, the new *tight* selection provides an increase in fake photon rejection, while keeping the same identification efficiency.

Alternative *loose*-like identification criteria are also used within the analysis. These are designated as *Loose'* working points. There are four *Loose'* working points in total, numbered 2-5. These criteria are stricter than the *loose* working point. Events must pass the *loose* cut, and they must fail specific shower shape criteria used in the *tight* working point definition. The primary *Loose'* working point used in this work is the *Loose'4* working point.

4.5 Calorimetric Photon Isolation

Photon isolation is a variable which quantifies the amount of activity in the vicinity of a photon candidate. The purpose of the isolation quantity is to help discriminate prompt photons (those originating from the physics process of interest, at or near the primary vertex) from non-prompt photons (those produced far from the primary vertex). A cartoon illustration of photon isolation being used to discriminate between prompt and non-prompt photons is presented in Figure 4.2. Essentially, non-isolated photons are defined as having large amounts of activity nearby, which is often from a close-by jet. These photons are more likely to have been radiated by charged hadrons in the jet, and they do not originate from the physics interaction being studied.

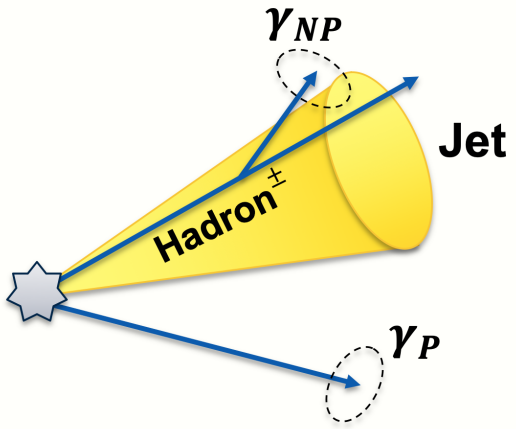
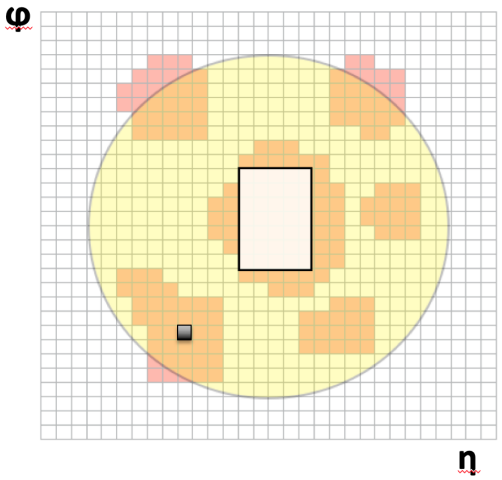
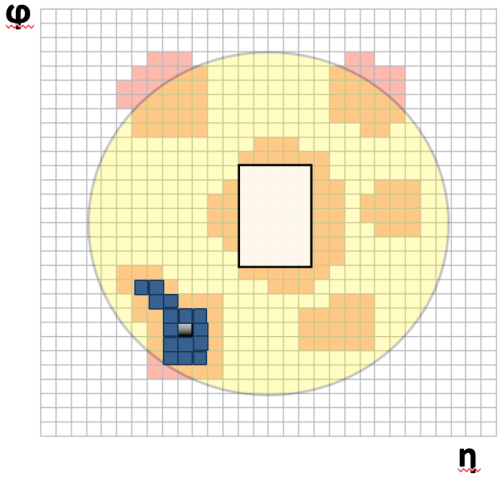


Figure 4.2: A cartoon illustration of photon isolation discriminating prompt and non-prompt photons. The grey star represents the physics interaction of interest at the primary vertex. The yellow cone is a jet, which contains many charged hadrons. One of these charged hadrons has radiated a non-prompt (not originating from the primary vertex) photon, denoted as γ_{NP}. An additional prompt photon is also shown, denoted as γ_P. The dotted ovals surrounding the photons denote the isolation radius; if significant energy is deposited within this radius, the photon is defined as non-isolated. The part of the jet cone overlapping with the photon isolation radius will deposit significant amounts of energy, causing γ_{NP} to be labeled as non-isolated. No objects in the cartoon deposit energy around γ_P, so it is considered isolated. In this way, the isolation requirement may be used to identify non-prompt photons, which are not usually of interest in physics analysis.

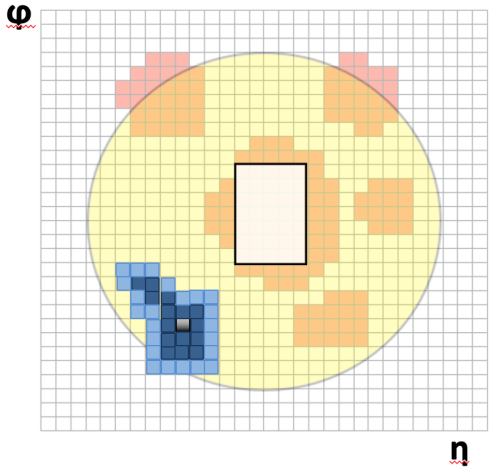
Photon isolation is calculated separately in both the ID (using the charged tracks within a given radius of the projected photon track) and the calorimeter (using energy deposits within the EM calorimeter). The calorimetric isolation variable is calculated by summing the energy from *topological clusters* seeded within a given radius of the photon within the EM calorimeter. The construction of topological clusters (also described in Section 4.4) is illustrated in Figure 4.3. First, seed cells are defined, as shown by the black cell in Figure 4.3a. The energy deposited in the cell must be more than four sigma greater than the expected noise in the cell. Seed cells must fall within



(a) A seed cell, which falls within the defined isolation radius ($\Delta R < 0.2$ or $\Delta R < 0.4$) around the photon candidate. The photon candidate is within the white 5×7 cell "core"– cells within this rectangle are not considered when building topoclusters.



(b) Neighbors of the seed cell in which the energy deposited is more than two sigma above the expected noise. These cells may fall outside the isolation radius, but they may not fall within the core.



(c) The final iteration of neighbors completing the cluster. These cells may fall outside the isolation radius, but they may not fall within the core.

Figure 4.3: The steps to construct a topocluster in the EM calorimeter. The first step (a) is to define seed cells, shown in black, in which the energy deposited is greater than four sigma above the expected noise. The second step (b) is to iteratively add neighboring cells, shown in dark blue, in which the energy deposited is greater than two sigma above the expected noise. The last step (c) is to add the neighboring cells surrounding the existing cluster, as shown by the light blue cells. The central white 5×7 cell is defined as the "core," and it contains most of the energy deposited by the photon candidate.

the defined isolation radius ($\Delta R < 0.2$ or $\Delta R < 0.4$) surrounding the photon candidate, and they may not fall within the central 5×7 cell "core" (shown by the white rectangle in 4.3) containing most of the energy deposited by the photon itself. Next, neighboring cells in which the deposited energy is greater than two standard deviations above the expected noise level of that cell are added to the cluster. This process is repeated iteratively until all connected cells satisfying the energy-to-noise requirement are included in the cluster, as shown by the dark blue cells in Figure 4.3b. These cells may fall outside the isolation radius, but they may not be within the 5×7 cell core. Finally, all cells neighboring the existing cluster are added to the cluster, as shown by the light blue cells in Figure 4.3c. These cells also may be outside the isolation radius, and they must not be within the core.

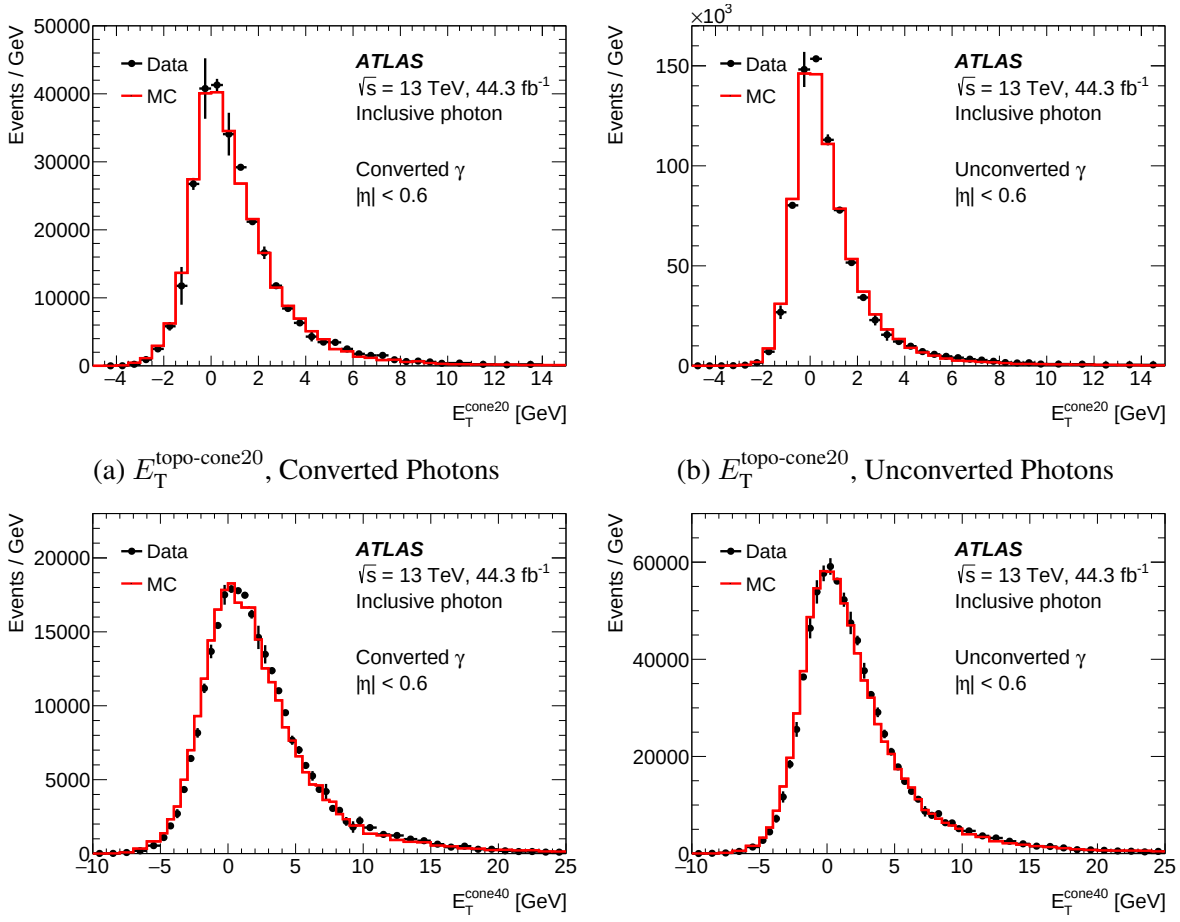
At low energy ($10 < E_{\mathrm{T}}^{\gamma} < 100$ GeV), the isolation quantity is studied using simulated "radiative Z" events, which are those where a Z boson decays to a pair of electrons or muons, with one of the leptons radiating a final state photon ($Z \to \ell\ell\gamma$). This sample contains relatively few events with "fake" photons (usually mis-reconstructed jets). For higher energies, however, the sample lacks sufficient statistics to study the isolation quantity precisely.

For higher energies ($25 < E_{\mathrm{T}}^{\gamma} < 1000$ GeV), an alternative "single photon" sample is used instead. The single photon sample simulated using PYTHIA8, described in Section 5.2, is utilized for the single photon calorimetric isolation studies presented here. The single photon data sample, described in Section 5.1, consists of events passing one of the single photon triggers. Although the data sample contains ample statistics, it also contains a significant fraction of fake photons. The removal of the fake photons from the single photon data set is detailed below in Section 4.5.1. For the studies presented here, all photons are required to pass the *loose* identification working point [58], as described in Section 4.4.

Table 4.1: The definitions of the three standard photon isolation working points.

Working Point	Calorimetric Isolation Requirement	Track Isolation Requirement
`FixedCutLoose`	$E_{\mathrm{T}}^{\text{topo-cone20}} - 0.065 \times E_{\mathrm{T}}^{\gamma} > 0.0$ GeV	$p_{\mathrm{T}}^{\text{cone20}} - 0.05 \times E_{\mathrm{T}}^{\gamma} > 0.0$ GeV
`FixedCutTight`	$E_{\mathrm{T}}^{\text{topo-cone40}} - 0.02 \times E_{\mathrm{T}}^{\gamma} > 2.45$ GeV	$p_{\mathrm{T}}^{\text{cone20}} - 0.05 \times E_{\mathrm{T}}^{\gamma} > 0.0$ GeV
`FixedCutTight(CaloOnly)`	$E_{\mathrm{T}}^{\text{topo-cone40}} - 0.02 \times E_{\mathrm{T}}^{\gamma} > 2.45$ GeV	–

Two calorimetric isolation variables are defined depending on the isolation radius considered: $E_{\mathrm{T}}^{\text{topo-cone20}}$ ($\Delta R = 0.2$) and $E_{\mathrm{T}}^{\text{topo-cone40}}$ ($\Delta R = 0.4$). Example of the calorimetric isolation variables distribution in the single photon samples are presented in Figure 4.4. Using these two variables, along with a track isolation variable $p_{\mathrm{T}}^{\text{cone20}}$ (discussed in Ref. [58]), three standardized isolation working point are defined: `FixedCutLoose`, `FixedCutTight`, and `FixedCutTightCaloOnly`. Their definitions are presented in Table 4.1.



(a) $E_{\mathrm{T}}^{\text{topo-cone20}}$, Converted Photons

(b) $E_{\mathrm{T}}^{\text{topo-cone20}}$, Unconverted Photons

(c) $E_{\mathrm{T}}^{\text{topo-cone40}}$, Converted Photons

(d) $E_{\mathrm{T}}^{\text{topo-cone40}}$, Unconverted Photons

Figure 4.4: Inclusive distributions of the calorimetric isolation variables in the single photon data (shown in black points) and corresponding simulation sample (shown in red) from the 2017 data taking period. The top row shows the distributions of the $E_{\mathrm{T}}^{\text{topo-cone20}}$ variable, while the bottom row shows those of the $E_{\mathrm{T}}^{\text{topo-cone40}}$ variable. The left column contains converted photons, and the right column contains unconverted photons.

4.5.1 Corrections to Calorimetric Photon Isolation in Simulation

The efficiency of the photon calorimetric isolation requirements is obtained by fitting the shape of the calorimetric isolation variable, $E_{\mathrm{T}}^{\text{topo-cone20}}$ or $E_{\mathrm{T}}^{\text{topo-cone40}}$, minus the relevant E_{T} fraction for the targeted working point. From Table 4.1, this translates to fitting the following distributions:

- `FixedCutTight(CaloOnly)`: $E_{\mathrm{T}}^{\text{topo-cone40}} - 0.022 \times E_{\mathrm{T}}^{\gamma}$
- `FixedCutLoose`: $E_{\mathrm{T}}^{\text{topo-cone20}} - 0.065 \times E_{\mathrm{T}}^{\gamma}$

An example distribution of this quantity targeting the `FixedCutTight` working point is shown in Figure 4.5. The advantage of considering the combined $E_{\mathrm{T}}^{\text{topo-cone20}}/E_{\mathrm{T}}^{\text{topo-cone40}}$ and E_{T} quantity is that events passing the working point criteria will fall to the left of a defined value when plotted. In Figure 4.5, for example, events passing the working point criteria fall to the left of 2.45 GeV (shown by a dashed green line in the plot). The efficiency can then be obtained by taking the integral of the isolation shape to the left of the cut value, divided by the total integral of the isolation shape.

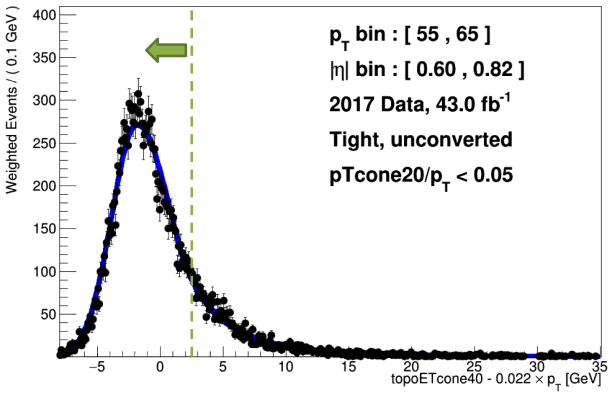


Figure 4.5: An example of the distribution of the $E_{\mathrm{T}}^{\text{topo-cone40}}$ variable minus the relevant fraction of photon E_{T}, shown by the black points (the blue shape is a fit to this distribution). The `FixedCutTight` working point requires that the quantity $E_{\mathrm{T}}^{\text{topo-cone40}} - 0.22 \times E_{\mathrm{T}}$ be *less than* 2.45 GeV. In the plot, this means that all events to the left of the dashed green line will pass the cut, while those to the right fail.

The efficiencies of the isolation cuts on photons from simulation and data have been found to disagree. In order to correct for this discrepancy, a *scale factor* is applied to simulated photons passing the isolation cut in physics analyses. The scale factor is defined as:

$$S = \frac{\epsilon_{\text{Data}}}{\epsilon_{\text{Simu}}}$$

where ϵ_{Data} is the isolation efficiency of real photons in the data sample and ϵ_{Simu} is the efficiency of simulated photons.

Because simulated photons are real photons by definition, performing a proper comparison to data requires that the data sample also contains only real photons. In order to isolate real photons, the scale factor is calculated using photons in data and simulation which pass the *tight* identification working point [58], as described in Section 4.4. Additionally, the cut on the track isolation variable (see table 4.1) is applied at pre-selection level, except when targeting the `FixedCutTight-CaloOnly` working point, in order to reduce additional fake photon contamination.

Still, the primary complication of the efficiency calculation is in determining the shape of only the true photons in the data sample, since the single photon data sample is contaminated by a significant number of fake photons which pass the strict *tight* identification requirements. Subtracting the remaining fake photon shape involves a process of iterative fits, which is described below in detail.

In order to better estimate the isolation shape of fake photons, a fake photon enriched region is defined using the alternative *Loose'4* identification criteria (discussed in Section 4.4). The *Loose'4* is chosen instead of the nominal *loose* region because it has been observed to be somewhat less correlated with photon isolation. Although background-enriched, the *Loose'4* identification region also contains some amount of true photons failing the *tight* identification requirement, defined as "leakage" photons.

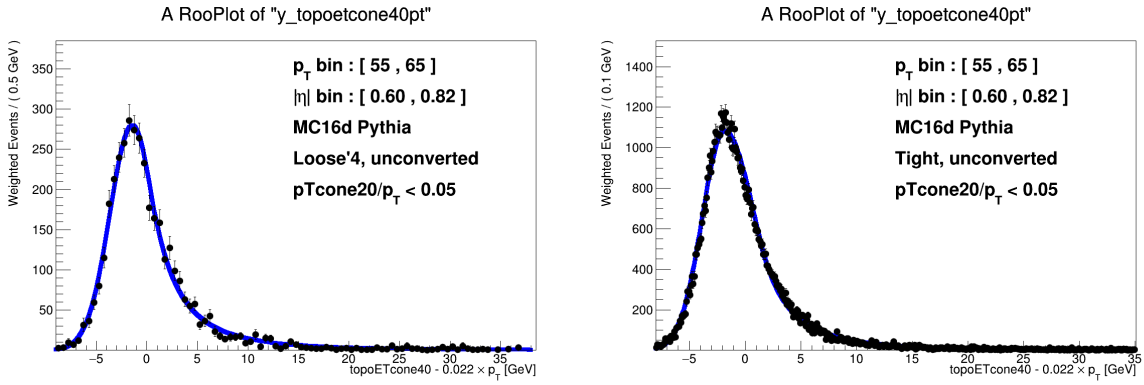
Scale factors are determined for fifteen bins in transverse momentum, which are chosen to correspond with the single photon trigger thresholds. The scale factors are binned in six bins in the absolute-value of pseudo-rapidity. The choice of $|\eta|$ binning is based upon differences in detector material, and it is the same as that used for the photon identification working point definitions. The scale factors are provided separately for converted or unconverted photons, as well as separately for the 2015-2016 and 2017 data-taking periods.

The sequence of fits, performed to remove the fake photon contamination from data, proceeds as follows:

- An initial fit to simulated photons passing the *tight* identification requirement in is performed using an Asymmetric Crystal Ball function [59]. An example of this fit can be seen in Figure 4.6b.

- A fit to simulated photons passing the *Loose'4* requirement is performed using a Crystal Ball function. An example of this fit can be seen in Figure 4.6a.

- A naive two-component fit is performed to photons in data passing the *Loose'4* requirement. An unconstrained Crystal Ball function is used to fit the fake photon shape, and the Crystal Ball fit to simulated photons in the *Loose'4* region is used for the shape of the "leakage" photons (real photons failing the *tight* requirement). An example of this two-component



(a) *Loose'4* , Simulation

(b) *tight*, Simulation

Figure 4.6: The (a) Crystal Ball fit to simulated photons in the *Loose'4* region and the (b) Asymmetric Crystal Ball fit to simulated photons in the *tight* region.

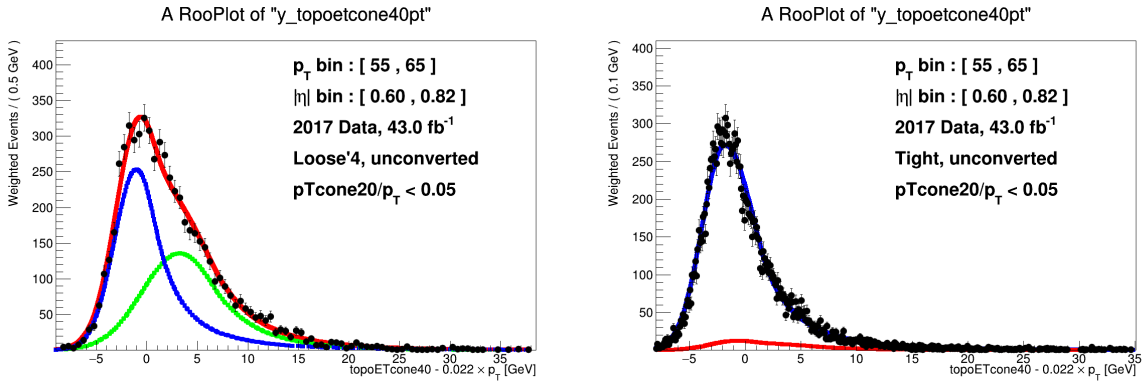
fit can be seen in Figure 4.7a, where the leakage component is shown in blue, the fake component is shown in green, and the total fit is shown in red.

- A naive two-component fit is performed on the photons from the data sample passing the *tight* requirement in order to estimate the number of real photons in this region, $N_{\text{True, } tight \text{ Data}}$. The fake photon component of the sample is fitted with the background shape from the fit to photons from data in the *Loose'4* region (the red curve in Figure 4.7a), and the real photon component is fitted with the shape obtained from the fit to simulated photons in the *tight* region (the curve in Figure 4.6b). An example of this fit is shown in Figure 4.7b.

The relative fractions of real and fake photons in the naive two-component fits to photons in data are not derived from the observed fractions in simulation. They are determined based on the best fit of the real and fake photon shapes to the data in each region. However, a more precise prediction for the number of leakage photons expected in the *Loose'4* region of the data can be obtained from the ratio of simulated *tight* to *Loose'4* photons. This prediction can then be used to constrain the magnitude of the leakage photon shape in the *Loose'4* data region. If the number of true photons in the *tight* region of the data sample is known, then the prediction for the number of leakage photons in the *Loose'4* region of the data is simply given by:

$$N_{\text{Leak, } Loose'4 \text{ Data}} = N_{\text{True, } tight \text{ Data}} \times \frac{N_{Loose'4 \text{ Simulation}}}{N_{tight \text{ Simulation}}} \tag{4.1}$$

where $N_{\text{True, } tight \text{ Data}}$ is the number of true photons in data passing the *tight* criteria, $N_{Loose'4 \text{ Simulation}}$ is the number of simulated photons passing the *Loose'4* criteria, and $N_{tight \text{ Simulation}}$ is the number



(a) *Loose'4* , Data

(b) Tight, Data

Figure 4.7: The initial two-component fit to (a) photons from data in the *Loose'4* region and (b) photons from data in the *tight* region. In the fit to photons from data in the *Loose'4* region (a), the blue line shows the fit to the "leakage" photons (true photons failing the *tight* criteria) component, and the green line shows the fit to the the fake photon component. The red line shows the total fit. In the fit to photons from data in the *tight* region (b), the blue line shows the total fit (to both the real and fake components), while the red line shows the fit to only the fake component. The shape of the fake component is taken from the red curve in (a).

of simulated photons passing the *tight* criteria.

The total number of events in data passing the *tight* criteria (the integral of the blue curve in Figure 4.7b) minus the fitted number of fake photons (the integral of the red curve in Figure 4.7b) gives an estimate of the number of true photons in the *tight* region of the data. Because the total background (red) shape includes real photons from leakage, though, this prediction will underestimate $N_{\text{True, } \textit{tight} \text{ Data}}$.

A second estimate of $N_{\text{True, } \textit{tight} \text{ Data}}$ can be obtained by assuming that the leakage shape fitted to photons in data in the *Loose'4* fit (the blue component in Figure 4.7a) contains entirely true photons. $N_{\text{True, } \textit{tight} \text{ Data}}$ is then estimated by subtracting only the integral of the *fake* component of the background shape (the green component in Figure 4.7a) from the total number of events. Due to the consistent overestimation of the number of leakage events by the naive fit to photons in the *Loose'4* region of data, this second prediction will overestimate the value of $N_{\text{True, } \textit{tight} \text{ Data}}$.

Since the two estimates of $N_{\text{True, } \textit{tight} \text{ Data}}$ are expected to err in opposite directions, an average of the two is taken to give the final prediction of $N_{\text{True, } \textit{tight} \text{ Data}}$. The difference between the final $N_{\text{True, } \textit{tight} \text{ Data}}$ value and the two predictions is taken as the uncertainty. The number of leakage events $N_{\text{Leak, } \textit{Loose'4} \text{ Data}}$ can then be determined using the above formula, and the fitting procedure with this updated leakage prediction continues as follows:

- The two-component fit to photons from data in the *Loose'4* region is performed again, with

the number of leakage events constrained to $N_{\text{Leak, } \textit{Loose'4} \text{ Data}}$ with a Poisson constraint. This second fit tends to show a decrease in the fraction of fitted leakage events, as shown in the blue component of the example fit in Figure 4.8a. The red component shows the fit to fake photon component of the *Loose'4* Data, while the green component shows the total fit.

- The final two-component fit is performed on photons in the *tight* region of the data. The signal shape is again taken from the fit to simulated photons passing the *tight* criteria (from Figure 4.6b). The background shape is taken from the fake component of the second fit to photons from data in the *Loose'4* region (the red component in Figure 4.8a). An example fit can be seen in Figure 4.8b, where the red component shows the predicted fake photon component and the blue shows the total fit.

- The fake component (the red shape in Figure 4.8b) of the two-component fit is subtracted from the shape of the photons in data passing the *tight* criteria. The remainder is the best estimate for the shape of true photons in the *tight* region of the data.

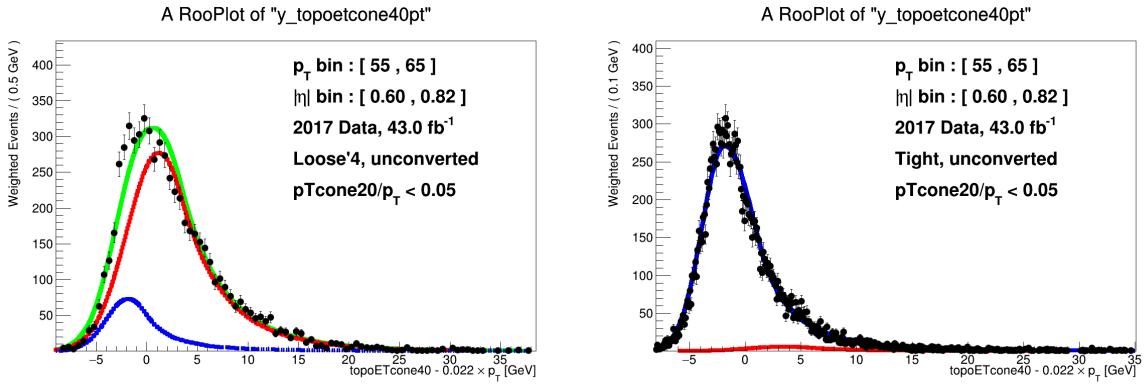
The calorimetric isolation efficiency in the data can finally be obtained by integrating the background-subtracted *tight* photon isolation shape up to the working point cutoff of 2.45 GeV (`FixedCutTight` and `FixedCutTightCaloOnly`) or 0.0 GeV (`FixedCutLoose`), then dividing by the total integral of the isolation shape. The simulation efficiency is similarly obtained by taking the integral of the isolation shape in simulated photons in the *tight* region up to the working point cutoff, then dividing by the total integral.

Three sources of systematic uncertainty are considered: the estimation of the number of leakage photons in the *Loose'4* region of data; the choice of the background-enriched identification region; and errors resulting from mismodeling by the choice of fit function.

To determine the uncertainty from the estimate of the number of leakage events in the *Loose'4* region of data, additional fits are performed using both the overestimate and underestimate of the number of leakage events instead of the average of the two, as discussed above. The maximum deviation between the $N_{overestimate}$ and $N_{underestimate}$ scale factor from the nominal result is taken as the leakage error.

To determine the uncertainty from the choice of the *Loose'4* region as the background-enriched region, the fitting process is re-performed using photons in the so-called *Loose'3* and *Loose'5* regions. The *Loose'3* criteria is slightly stricter than *Loose'4*, while *Loose'5* is closer to the *loose* requirement. The maximum discrepancy between the *Loose'3* and *Loose'5* scale factors from the nominal *Loose'4* result is taken as the *Loose'* error.

A binomial statistical error on the scale factors is also calculated. The contribution from the fit quality is included as a scaling on the statistical error, and it is obtained from the $\sqrt{\chi^2/(N_{bins}-1)}$



(a) *Loose'4* , Data (b) Tight, Data

Figure 4.8: The final two-component fit to (a) photons in the *Loose'4* region of data and (b) photons in the *tight* region of data. In the fit to *Loose'4* photons (a), the blue line shows the fit to the leakage photons (true photons failing the *tight* criteria), and the red line shows the fit to the the fake photon component. The green line shows the total fit. In the fit to photons in the *tight* region of data (b), the red line shows the fitted fake photon shape, taken from the red curve in (a), while the blue line shows the total fit.

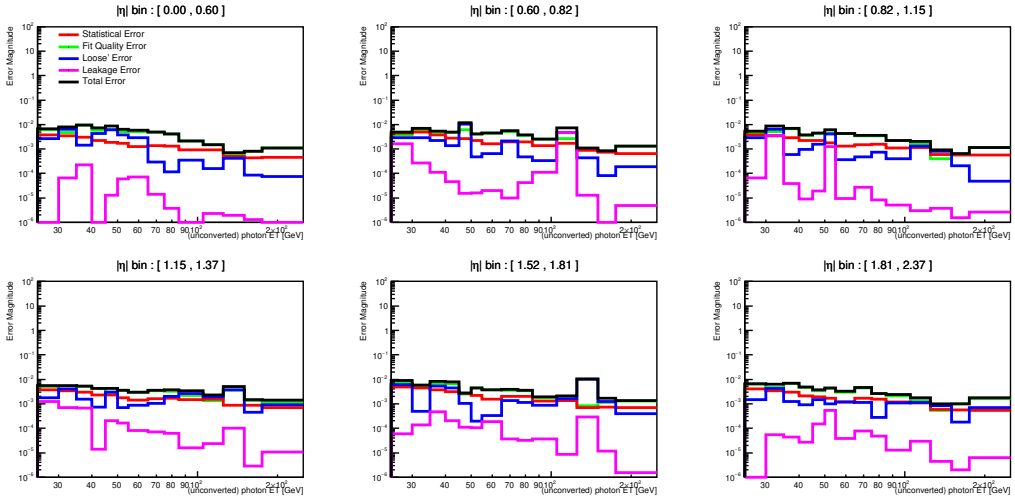


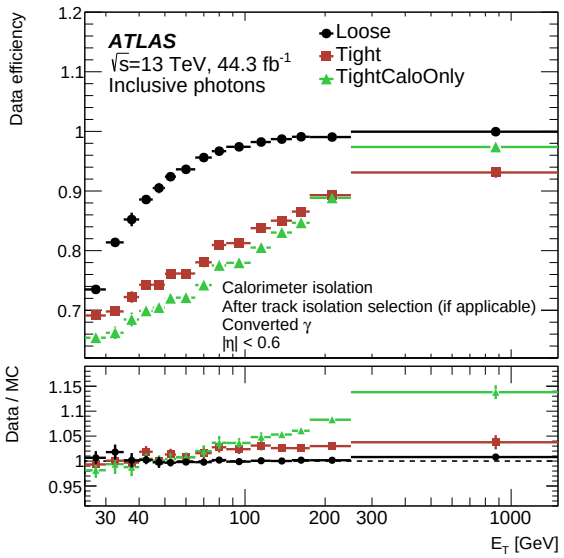
Figure 4.9: The breakdown of the systematic errors for the `FixedCutLoose` scale factors for unconverted photons, obtained using the 2017 subset of single photon data and corresponding simulation sample. The error values versus photon E_T are plotted separately for each bin in $|\eta|$. The leakage error is shown in magenta, and the $Loose'$ error is shown in blue. The raw statistical error is shown in red, and the the scaled statistical error (to account for the fit error) is shown in green. The black line shows the total error.

value of the fits. To obtain the total uncertainty, the two systematic (Loose'-ID choice and leakage estimate) components are added in quadrature with the scaled statistical component. An example breakdown of the systematic uncertainty components can be found in Figures 4.9 for the `FixedCutLoose` working point. Typically, the fit quality systematic is the dominant contribution.

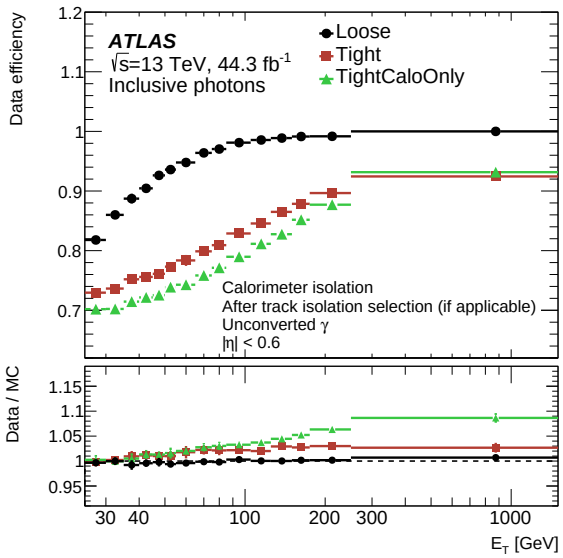
The calorimetric isolation data efficiencies and scale factors from the single photon data set in the innermost ($|\eta| < 0.6$) and outermost ($1.81 < |\eta| < 2.37$) pseudorapidity bins are shown in Figure 4.10. The final photon isolation efficiency for each working point is determined by combining the calorimetric isolation measurements from single photons (presented here) with track isolation measurements from single photons and combined isolation measurements from the $Z \to \ell\ell\gamma$ sample.

In addition to the calorimetric isolation scale factors, another correction is also derived using the single photon samples. This correction, known as a "data-driven shift," addresses the fact that the mean of the calorimetric isolation distribution in simulation is observed to be displaced from that seen in data. The data-driven shift is then simply a constant value added to the simulated isolation variable. The shifts are calculated using the same fitting procedure described above, but with the fits performed directly to the isolation variables $E_{\mathrm{T}}^{\text{topo-cone20}}$ and $E_{\mathrm{T}}^{\text{topo-cone40}}$ (no fraction of E_{T}^{γ} is subtracted). The shifts are extracted by taking the difference between the mean of the final fitted shape of real photons in data passing the *tight* criteria and that of the fit to photons in the *tight* region of the simulation sample.

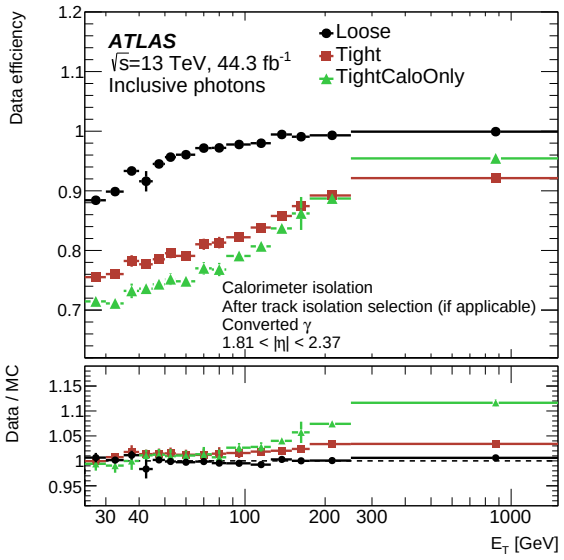
Examples of the data-driven shifts measured using the 2017 subset of data and corresponding simulation sample are presented in Figure 4.11. Shifts for both the $E_{\mathrm{T}}^{\text{topo-cone20}}$ and $E_{\mathrm{T}}^{\text{topo-cone40}}$ variables are shown, as well as for both converted and unconverted photons. The shifts are calculated in the same six pseudorapidity bins as for the scale factors; the figure shows those in the innermost ($|\eta| < 0.6$) and outermost ($1.81 < |\eta| < 2.37$) bins.



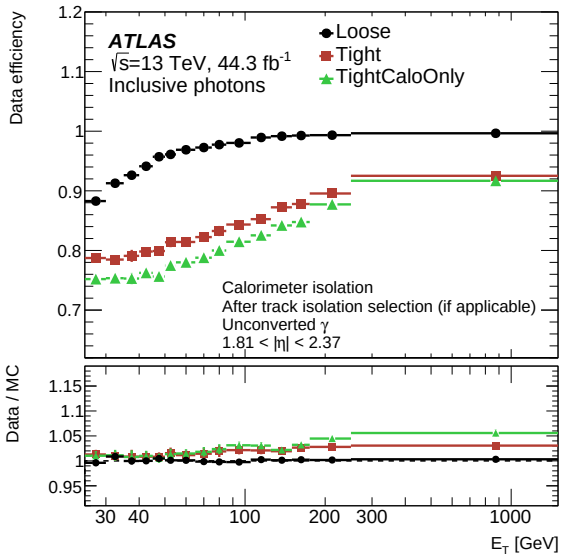
(a) Converted, $|\eta| < 0.6$



(b) Unconverted, $|\eta| < 0.6$

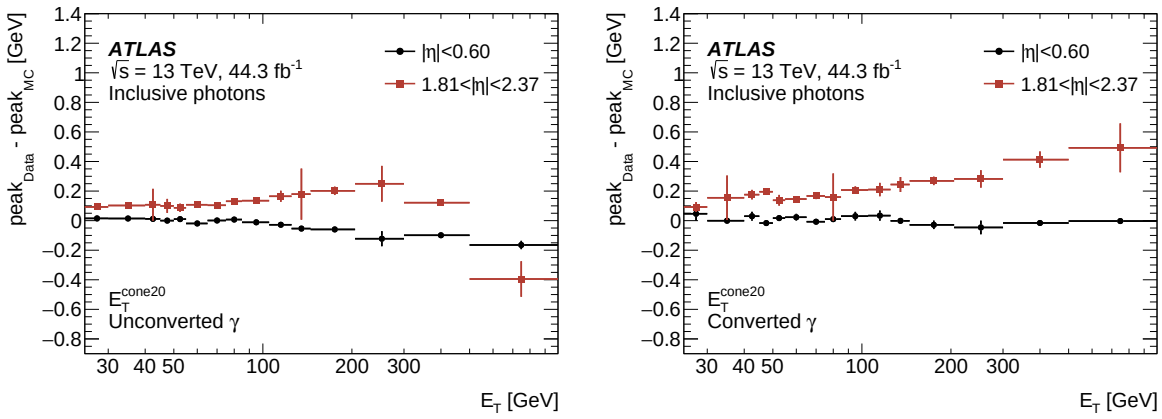


(c) Converted, $1.81 < |\eta| < 2.37$



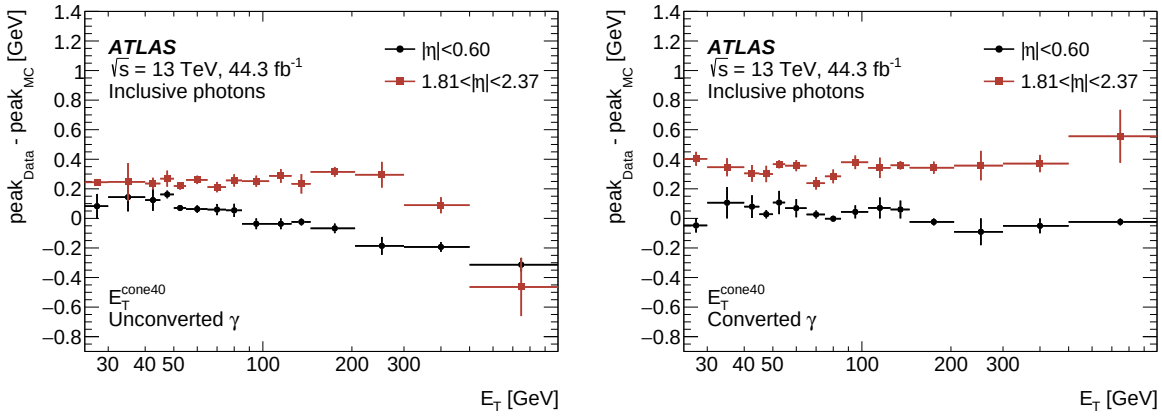
(d) Unconverted, $1.81 < |\eta| < 2.37$

Figure 4.10: Measurement using single photons, calo-only: the efficiencies as a function of E_{T} for the three working points, `FixedCutLoose` (black), `FixedCutTight` (red), and `FixedCutTightCaloOnly` (green). The bottom panels of each subplot show the scale factors. The left plots are the results using converted photons, while the right plots are those using unconverted photons. The top row shows the results in the inner pseudorapidity range of $|\eta| < 0.6$, while the lower plots show the pseudorapidity range of $1.81 < |\eta| < 2.37$.



(a) $E_{\mathrm{T}}^{\text{topo-cone20}}$, Unconverted Photons

(b) $E_{\mathrm{T}}^{\text{topo-cone20}}$, Converted Photons



(c) $E_{\mathrm{T}}^{\text{topo-cone40}}$ Unconverted Photons

(d) $E_{\mathrm{T}}^{\text{topo-cone40}}$, Converted Photons

Figure 4.11: The data-driven shifts calculated for the (top) $E_{\mathrm{T}}^{\text{topo-cone20}}$ and (bottom) $E_{\mathrm{T}}^{\text{topo-cone40}}$ variables, using the 2017 data taking period and corresponding simulated single photon samples. The left plots show the shifts for unconverted photons, while the right show those for converted photons. The black points show the shifts for the most central pseudorapidity bin ($|\eta^{\gamma}| < 0.6$), while the maroon points show the shifts for the most forward pseudorapidity bin ($1.81 < |\eta^{\gamma}| < 2.37$).

CHAPTER 5

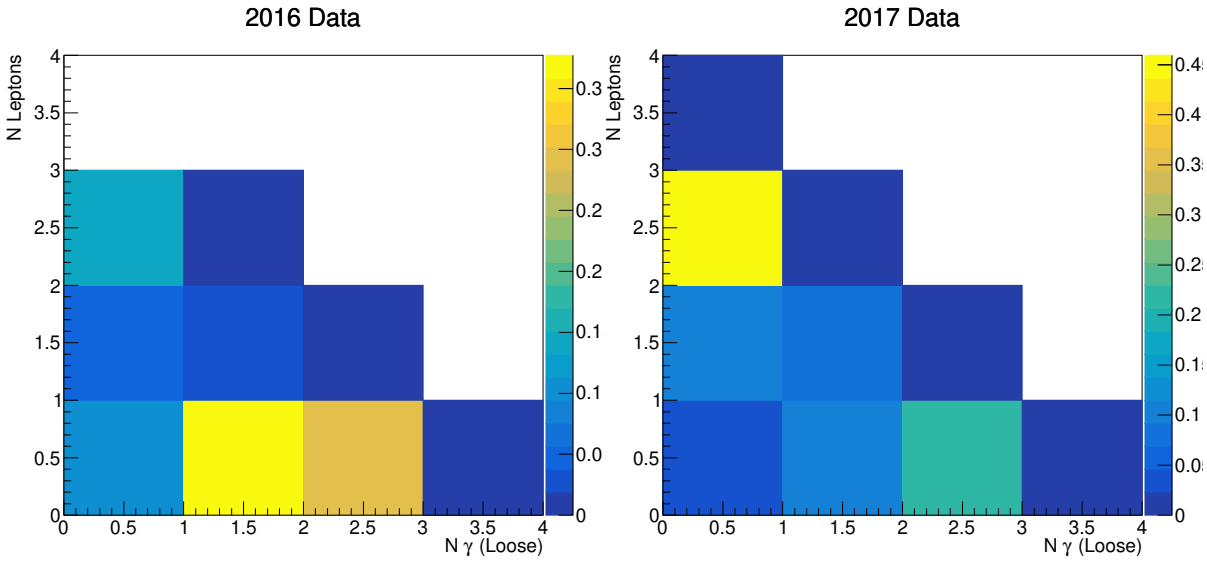
Data and Simulation Samples

This chapter discusses the details of the dataset used by the analysis. The relevant samples of simulated events, used to model specific physics processes, are discussed as well. The dataset was collected by the ATLAS detector, as discussed in Chapter 3. The simulation of physics processes includes the detector response.

5.1 Data Sample

The analysis presented includes proton-proton (pp) collision data taken at a center-of-mass energy $\sqrt{s} = 13$ TeV during Run 2 (2015-2018) of the LHC. The $H(\gamma\gamma)$ Couplings analysis uses the subset of data taken between 2015 and 2017, amounting to a total integrated luminosity of 79.8 fb^{-1}. The portion of the analysis concerning $t\bar{t}H$ production contains the full Run 2 pp dataset, which amounts to a total integrated luminosity of 139 fb^{-1}. The mean number of interactions $< \mu >$ per bunch crossing was 34 over the full Run 2 period. The average μ per bunch crossing was 23 during the 2015-2016 data-taking period, and it increased to 37 during the 2017-2018 data-taking period. Data events are required to pass a set of data-quality criteria, which requires that all subdetector components are functioning properly.

The ATLAS trigger system is described in detail in Section 3.0.7. The analysis uses data passing the following triggers: `HLT_g35_loose_g25_loose` (in the 2015-2016 data-taking period) and `HLT_g35_medium_g25_medium` (in the 2017-2018 data-taking period). The diphoton triggers are constructed from two primary single photon triggers, with one requiring a photon candidate with a transverse energy E_{T} of at least 35 GeV and the other requiring a candidate with a E_{T} of at least 25 GeV. The *loose* and *medium* labels within the trigger names denote the photon identification requirement used; these requirements are described in detail in Section 4.4. The trigger identification requirement was tightened from *loose* to *medium* due to the increases in pileup and instantaneous luminosity between 2016 and 2017 in order to preserve a manageable trigger rate [58]. The efficiency of each of the primary single photon "legs" of the diphoton triggers has



(a) 2016 Data Subset

(b) 2017 Data Subset

Figure 5.1: Normalized distributions of the number of leptons (y-axis) and the number of *loose*-identified photons (x-axis) for a sample of the 2016 data (left) and a sample of the 2017 data (right).

been measured to be about 95% or higher when the photons have a transverse energy at least 5 GeV above the trigger thresholds (25 and 35 GeV) [60].

The tightening of the online photon identification requirement in the 2017 diphoton trigger was expected to increase the fraction of recorded data events which would pass the requirement of having two photons satisfying the offline *loose* identification criteria. However, the opposite was observed; the efficiency of the "two loose photons" cut decreased in the 2017 data set compared to the efficiency in the 2015-2016 data set. One theory to explain the decrease in efficiency was that there was a larger contamination of electrons passing the 2017 trigger.

In order to test this theory, small subset of data were produced without the requirement of two loose photons. A portion of run 301915 (2016) and run 330166 (2017) were analyzed, which corresponded to 2735 events from 2016 and 1385 events from 2017. The relative fractions of leptons and *loose*-identified photons were plotted in 2D histograms, as presented in Figure 5.1. One can see that the 2017 data contains a relatively large fraction of events which have two leptons, but no *loose* photons. The 2016 data, on the other hand, does not show this behavior, and most events appear to have two *loose* photons. This discrepancy supports the theory that the 2017-2018 trigger selects a comparably higher fraction of electrons than the 2015-2016 trigger, despite the nominal tightening of the identification requirement.

The photon isolation studies, described in Section 4.5.1, utilize the so-called "single photon"

data set. This data set includes all proton-proton collision events passing one of the single photon triggers (the lowest threshold of which is 25 GeV) and also passing the basic data-quality requirements described above. The single photon triggers use the *loose* online identification requirement. Due to their high rates, they are pre-scaled (meaning only a fraction of events passing a given trigger are recorded). The isolation studies only consider data taken between 2015 and 2017.

5.2 Simulation Samples

The analysis uses simulated samples of Higgs boson events decaying to two photons. The simulated production modes include: gluon-guon fusion (ggF) [61]; vector boson fusion (VBF) [62]; vector-boson associated production (VH) [63, 64], split into $W^{\pm}$ and Z associated production ($W^{+}H$, $W^{-}H$, and ZH); top-associated production ($t\bar{t}H$); and single-top production (tH), split into $tHjb$ and tWH. The most prevalent of these modes are discussed in Section 2.1.5. Additionally, the rare production modes $ggZH$ and $b\bar{b}H$ are simulated, as well. Most of the signal samples are simulated using the POWHEG generator [65, 66, 67], using the PDF4LHC15 [68] Parton Distribution Function (PDF) set. Subsequent parton showering and hadronization are modeled with PYTHIA8 [69] using either the AZNLO parameter set [70] tuned to data, or, for the case of $t\bar{t}H$ and $b\bar{b}H$, the A14 parameter set [71]. The exceptional signal samples are $tHjb$ and tWH. The $tHjb$ sample is generated with MADGRAPH [72] using the CT10 PDF set [73], interfaced to PYTHIA8 with the A14 tune for showering and hadronization modeling. The tWH sample is generated with MADGRAPH5_aMC@NLO using the CT10 PDF set, interfaced to Herwig++ [74, 75, 76] for showering and hadronization using the the UEEE5 parameter set.

The simulated Higgs boson samples are normalized to their expected SM cross sections (as reported in Refs. [13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31]) times the expected SM branching ratio to two photons (as reported in Refs. [13, 77, 78, 79, 80, 81]) at a center-of-mass energy of 13 TeV and with a Higgs boson mass of 125.09 GeV. The inclusive cross section values per Higgs boson production mode are presented in Table 5.1. The inclusive ggF cross section is calculated at the next-next-next-leading-order (N3LO) for QCD processes and next-leading-order (NLO) for Electroweak processes. The inclusive VBF, $W^{\pm}H$, and ZH *QCD* cross sections are calculated at the NNLO for QCD processes and NLO for Electroweak processes. The inclusive $t\bar{t}H$ cross section is calculated at the NLO for QCD and NLO for Electroweak processes. The inclusive $tHjb$ and tWH cross sections are calculated using the 5FS at NLO QCD accuracy (no NLO Electroweak corrections are included). The inclusive $b\bar{b}H$ cross section is calculated using a combination of the five-flavor-scheme (5FS) NNLO and four-favor-scheme (4FS) NLO for QCD processes; no NLO Electroweak corrections are included.

The analysis uses simulated events from the continuum $\gamma\gamma$, $t\bar{t}\gamma\gamma$, and $V\gamma\gamma$ background pro-

Table 5.1: The predicted SM cross sections of the Higgs boson production modes, as reported in Refs. [13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31].

Production Mode	Cross Section [pb]
ggF	4.852×10^{1}
VBF	3.779
$W^{\pm}H$	1.369
ZH	8.824×10^{-1}
$t\bar{t}H$	5.065×10^{-1}
$tHjb$	7.426×10^{-2}
tWH	1.517×10^{-2}
$b\bar{b}H$	4.863×10^{-1}

cesses. The $\gamma\gamma$ and $V\gamma\gamma$ samples are generated with SHERPA 2.2.4 [82] using the CT10 PDF set. SHERPA parton showering and hadronization is also used, with the ME+PS@NLO prescription [83, 84] and dedicated parton showering tune developed by the authors of SHERPA [85]. The $t\bar{t}\gamma\gamma$ sample is generated with MADGRAPH5_aMC@NLO using the PDF4LHC15 PDF set and is interfaced to PYTHIA8 with the A14 parameter set. Additionally, a small set of events containing one prompt photon and one prompt jet (γj) are produced for studying backgrounds in which a jet is mis-reconstructed as a photon. These samples are produced, showered, and hadronized with SHERPA, using the CT10 PDF set and the same prescription as for the $\gamma\gamma$ samples.

The photon isolation (detailed in Section 4.5.1) studies utilize dedicated single photon simulation samples generated using PYTHIA8; showering and hadronization in these samples is also modeled using PYTHIA8. These samples contain one prompt photon and one prompt jet (additional radiated photons and jets may be present in the events) resulting from the $q\bar{q} \to \gamma g$ and $qg \to q\gamma$ processes. The A14 parameter set is used, along with the PDF4LHC15 PDF set.

All generated Higgs boson events are passed through a full simulation of the ATLAS detector response [86] using GEANT4 [87]. The continuum $\gamma\gamma$ and $t\bar{t}\gamma\gamma$ background samples are processed with a fast GEANT4 simulation, which uses a parameterization of the calorimeter response in place of full simulation of the calorimeter [88]. The single photon samples are produced using both the full and fast detector simulations in order to study isolation mismodeling introduced through the use of the parameterized calorimeter response. The generation of the simulated event samples includes the effect of pileup (multiple pp interactions per bunch crossing), as well as the effect on the detector response due to interactions from bunch crossings before or after the one containing the hard interaction. Due to the significant difference in the pileup profile seen across Run 2, three separate subsamples of simulated events were generated. One set ("mc16a") contains a pileup profile similar to that in the 2015-2016 data subset. The remaining two, "mc16d" and "mc16e," contain pileup profiles similar to those of the 2017 and 2018 data subsets, respectively.

CHAPTER 6

Analysis Selection

The analysis selection is designed to select the targeted signal events, while the rejecting background events from other physics processes. A general *diphoton* preselection is applied to isolate events consistent with a Higgs boson decay to two photons. A finer selection is then used to separate events into one of 29 analysis categories. Broadly, these categories target specific production modes of the Higgs boson. On a finer level, they may target different areas of phase space within a production mode. Particular attention is paid to the selection of events consistent with originating from the $t\bar{t}H$ process.

The physics objects utilized in the analysis presented here are reconstructed from detector signatures as described in Chapter 4.

6.1 Diphoton Preselection

The photons used in the analysis are reconstructed from topological clusters of energy deposits in the electromagnetic calorimeter, as detailed in Chapter 4. Photon candidates are required to pass a *loose* identification requirement [58], designed to reduce the number of energy deposits from jets misidentified as photons. The photon identification requirement is based on properties of the energy shower observed in the EM and Hadronic calorimeters, as detailed in Section 4.4

Events are required to pass the diphoton trigger (the trigger requirement of the analysis is discussed in Section 5.1). Within each triggered event, the leading photon candidate is required to have a E_{T} of at least 35 GeV, while the subleading candidate's E_{T} must be greater than 25 GeV. The photon candidates must fall within the pseudorapidity range of $|\eta| < 2.37$; additionally, they must not fall within the "crack" region (the transition between the barrel and endcap regions of the calorimeter), corresponding to $1.37 < |\eta| < 1.52$.

Once all of the *loose* photon candidates have been identified, the two with the greatest transverse momentum (p_{T}) are selected as the potential decay products of the Higgs boson. The pointing

direction (the z position most compatible with the shower shapes in the calorimeter) of these photon candidates is fed into a neural network [89]. The neural net is also given the scalar sum of the transverse momentum ($\sum p_{\mathrm{T}}$) and the sum of the squared transverse momentum ($\sum p_{\mathrm{T}}^2$) of the tracks originating from all reconstructed vertices in the event, as well as the azimuthal angle between the vector sum of the p_{T} from the available vertices and the p_{T} of the diphoton system ($\Delta\phi(\text{tracks}, \gamma\gamma)$). The neural network is trained using simulated $ggF(H \rightarrow \gamma\gamma)$ events to select the vertex corresponding to the hard scatter event in which the Higgs boson is produced. When tested using simulated ggF events in Run 1, the network was found to choose a vertex within 0.3 mm of the true interaction point with around an 85% accuracy [89]. The use of a dedicated diphoton vertex algorithm is motivated by the fact that a significant fraction of Higgs boson interactions (especially those from the ggF production mode) do not result from the hardest vertex (the default vertex choice) in an event.

After the diphoton primary vertex has been identified, the stricter *tight* identification requirement is applied. The tight selection, also described in detail in Section 4.4, includes the requirements of the *loose* working point, as well as additional information to further reduce the contamination of fake photons.

In addition to identification requirements on the photon EM calorimeter showers, isolation requirements are imposed on photon candidates in order to suppress jets misidentified as photons. The isolation requirement can be broken into two parts: the first part is a condition on the EM calorimeter information and the second part is a condition on information from the inner tracker. The EM calorimeter condition demands that the transverse energy deposited in clusters within a cone of $\Delta R < 0.2$ (with ΔR defined as $\sqrt{\Delta\phi^2 + \Delta\eta^2}$) surrounding the photon candidate in the EM calorimeter does not exceed 6.5% of the photon's transverse energy – this working point is defined as `FixedCutLoose`. Note that the deposited transverse energy is corrected for the average pileup transverse energy, and the transverse energy from the photon candidate is subtracted. The tracking-related condition demands that the scalar sum of the transverse momenta of tracks within a cone of $\Delta R < 0.2$ surrounding the photon be less than 5% of the photon candidate's p_{T}. Only tracks with $p_{\mathrm{T}} > 1$ GeV and consistent with the reconstructed primary vertex are considered. The use of the photon isolation cut helps reduce the contribution from "non-prompt" photons (those not originating from the inetraction of interest). The systematic uncertainties associated with photon isolation were recalculated for the $t\bar{t}H$ portion of the analysis. Extensive detail on photon isolation is presented in Section 4.5.1.

After the application of the *tight* identification and `FixedCutLoose` isolation requirements, kinematic cuts are imposed such that the p_{T} of the (sub)leading photon is >0.35 (0.25) times the the diphoton invariant mass, $m_{\gamma\gamma}$. Assuming the Higgs boson is a spin-0 particle (as predicted by the SM), the two decay photons will have an isotropic distribution in the reference frame of

the Higgs boson. Background diphoton processes, on the other hand, are more likely to produce forward photons, which are preferentially cut by the p_{T} requirements.

Lastly, the diphoton invariant mass is required to fall within a window of $105 < m_{\gamma\gamma} < 160$ GeV. This window contains the SM Higgs mass, as well as a sufficient range both above and below the Higgs mass used for validating the background model (discussed in Section 7.3). Events which satisfy these criteria (two *tight*-identified, isolated photons with satisfactory kinematic properties) are delineated as those passing the so-called *diphoton selection*.

6.2 Other Object Selection

The physics objects (electrons, muons, and jets) utilized in the analysis are reconstructed as detailed in Chapter 4. Reconstructed jets are required to have a transverse momentum $p_{\mathrm{T}} > 25$ GeV to be considered in the $t\bar{t}H$ categories, and $p_{\mathrm{T}} > 30$ GeV to be considered for the remaining analysis categories. In all categories, jets must have a pseudorapidity $|\eta| < 4.4$. Of these jets, those with $|\eta| < 2.5$ and containing b-hadrons are identified using the MV2c10 b-tagging algorithm [90, 91, 92] with the b-tagging average efficiency of 77%, corresponding to a light flavor mistagging rate of approximately 1%. In addition, jets with $p_{\mathrm{T}} < 120$ GeV and $|\eta| < 2.4$ which originate from pileup collisions are identified and removed via a jet vertex tagger multivariate discriminant [93].

Electron candidates are identified using a likelihood disciminant, which takes into account the candidate's shower shape in the EM calorimeter and the matched track information. The *Medium LH* identification requirement (as outlined in Ref. [94]) is used in this analysis. Electron candidates in the analysis are required to have $p_{\mathrm{T}} > 15$ GeV and fall within the region of $|\eta| < 2.47$ (as well as fall outside of the transition region between the central and endcap EM calorimeters), and have $|z_0 \sin\theta| < 0.5$ mm (where z_0 is the longitudinal impact parameter along the beamline) in order to ensure consistency with the diphoton vertex. The electron candidate's transverse impact parameter d_0 (the transverse distance from the beamline) divided by its uncertainty σ_{d_0} must be less than 5, also to ensure track and diphoton vertex compatibility. As with photons, isolation criteria derived from track and calorimeter information are applied to electron candidates in order to reject fake electron candidates. The *Fix (Loose)* (as outlined in Ref. [94]) criteria is chosen, for which the efficiency of real electrons considered in the analysis is greater than 95% for the range of $p_{\mathrm{T}} > 15$ GeV and $|\eta| < 2.47$.

Muon candidates in the analysis are required to have both $p_{\mathrm{T}} > 15$ GeV and $|\eta| < 2.7$. In addition, an identification requirement with the *medium* defined working point [95] is used to select real muons. As with electrons, additional constraints on the longitudinal and transverse impact parameters ($|z_0 \sin\theta| < 0.5$ mm and $|d_0|/\sigma_{d_0} < 3$) are imposed in order to ensure consistency

between the muon candidate's track and the reconstructed diphoton vertex. Also as with electrons, isolation criteria derived from track and calorimeter information are imposed. These criteria were chosen such that the efficiency of real muons is greater than 95% for the range of $p_{\mathrm{T}} > 15$ GeV and $|\eta| < 2.7$.

The missing transverse momentum, $E_{\mathrm{T}}^{\mathrm{miss}}$, is defined as the negative vector sum of the transverse momenta of all objects associated with the reconstructed primary diphoton vertex. This includes photons, electrons, muons, jets, and any additional unidentified low-p_{T} tracks [96]. In order to remove double-counted objects, overlap removal criteria based on the distance apart ΔR of two objects in the detector is utilized, with cuts applied in the following sequence:

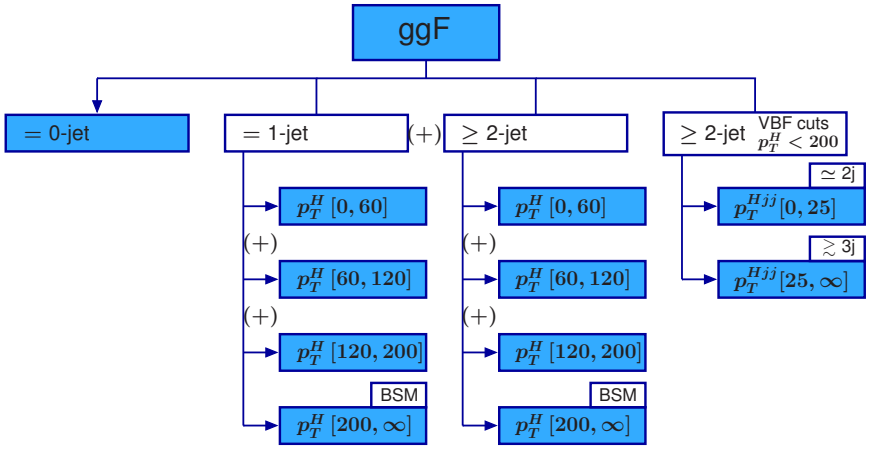
1. Remove electrons and jets overlapping ($\Delta R < 0.4$) with the two selected photons
2. Remove jets overlapping ($\Delta R < 0.2$) with the remaining electrons
3. Remove electrons overlapping ($\Delta R < 0.4$) with the remaining jets
4. Remove muons overlapping ($\Delta R < 0.4$) with the two selected photons
5. Remove muons overlapping ($\Delta R < 0.4$) with the remaining jets

6.3 Analysis Categorization

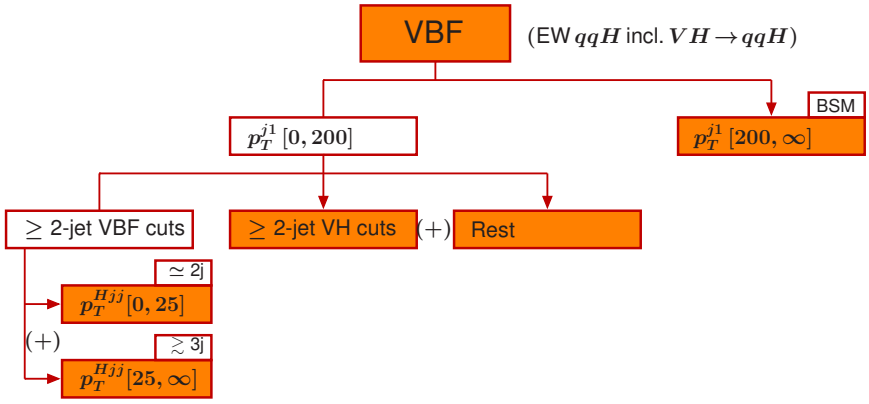
The $H(\gamma\gamma)$ Couplings Analysis is designed to measure the cross-section of the four leading production modes (ggF, VBF, VH and $t\bar{t}H$) of the Higgs Boson at the LHC, using the diphoton decay channel. These production modes, as well as the diphoton decay channel of the Higgs Boson, are described in Section 2.1.5.

The analysis utilizes 29 orthogonal categories in order to improve final sensitivity and to fit into the Simplified Template Cross Section (STXS) [13] framework. The STXS framework is designed to separate Higgs boson events into so-called "truth bins," which represent different portions of kinematic phase space within the production modes. By separating into the truth bins, the analysis results may be easier to compare with theoretical BSM predictions. The STXS scheme in the different production modes is presented in Figure 6.1. Due to statistical limitations, the analysis categories do not capture the full granularity of the STXS truth bins. Categories with insufficient statistics are merged together, indicated by a "+" sign in Figure 6.1. Additionally, events from the $t\bar{t}H$ and tH production modes are merged into the single "Top" category in the STXS scheme.

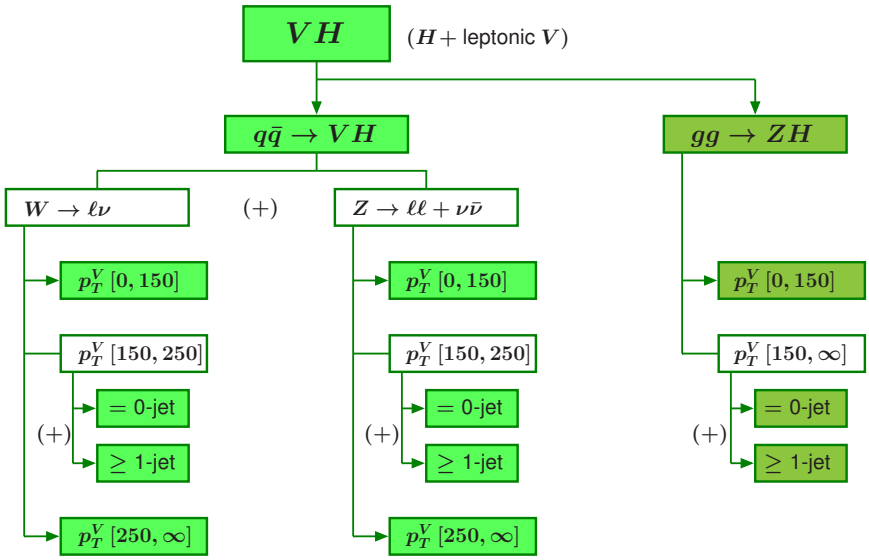
The analysis categorization is mostly unchanged with respect to the previous iteration, using 36 fb^{-1} and presented in Ref. [97]. The qqH BSM category now includes an additional requirement that at least two jets are present in the events in order to reduce contamination from ggF



(a) ggF STXS Bins



(b) VBF STXS Bins



(c) VH STXS Bins

Figure 6.1: The STXS truth bin scheme in the (a) ggF, (b) VBF, and (c) VH production modes. The "+" symbols mark where truth bins have been merged into coarser analysis categories, due to limited statistics.

events. Additionally, the $t\bar{t}H$ categorization was altered significantly. Two dedicated Boosted Decision Trees (BDTs) are used, resulting in seven orthogonal analysis categories. The dedicated $t\bar{t}H$ selection is detailed below in Section 6.4.

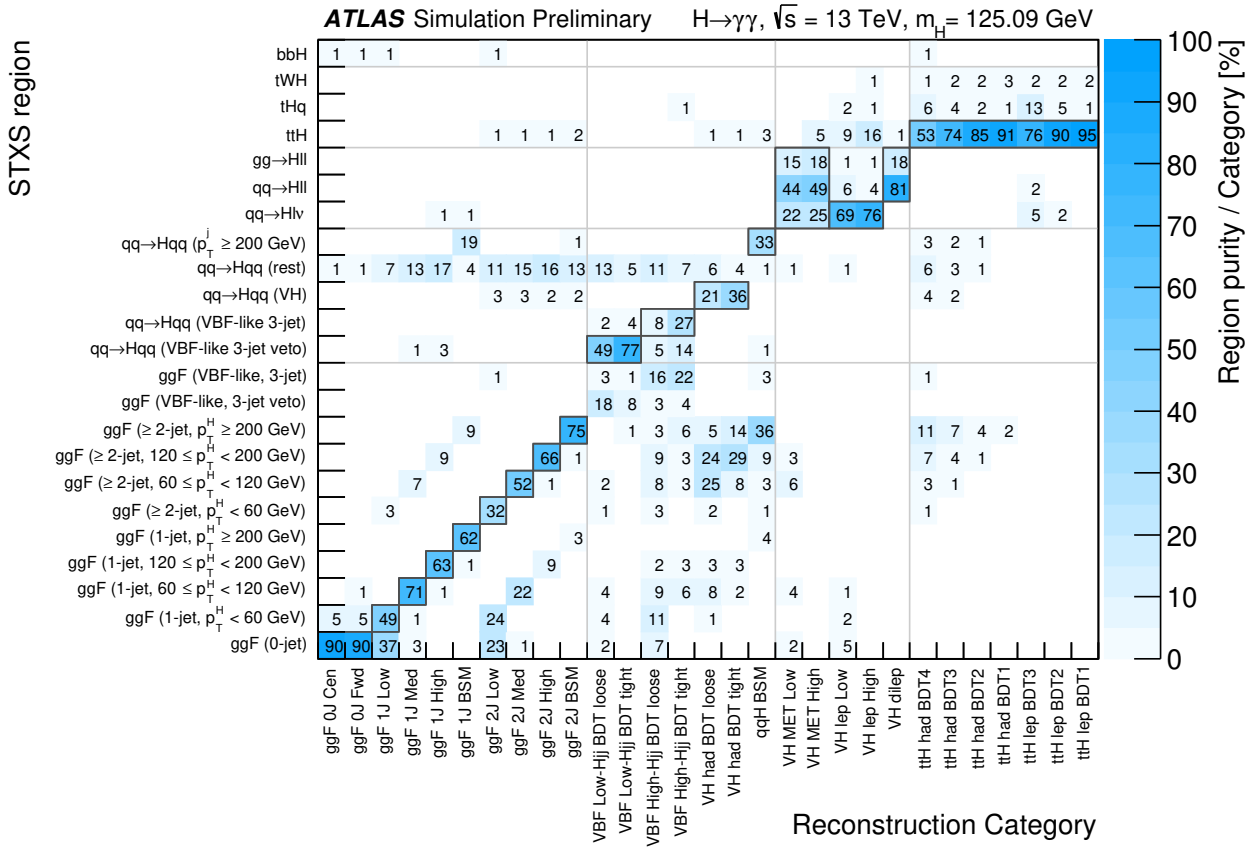


Figure 6.2: The fraction of Higgs boson events in each of the 29 analysis categories, broken down by production mode. The fractions reflect the prediction of the Higgs boson simulation samples.

The ggF and VH categories are entirely cut-based. A dedicated BDT is used to select VBF events. The input variables used by this BDT are:

- The invariant mass of the leading two jets, m_{jj}
- The separation in pseudorapidity of the leading two jets, $\Delta\eta(jj)$
- The azimuthal separation between the dijet system and diphoton system, $\Delta\phi(jj,\gamma\gamma)$
- The transverse momentum of the diphoton system, projected perpendicular to the diphoton thrust axis, $p_{\mathrm{T}t}$
- The minimum ΔR between one of the two leading photons and one of the two leading jets, $\Delta R^{\mathrm{min}}_{\gamma,j}$

- The Zeppenfeld pseudorapidity [98], defined as $\eta^{\text{Zeppenfeld}} = \eta_{\gamma\gamma} - 0.5(\eta_{j1} + \eta_{j2})$

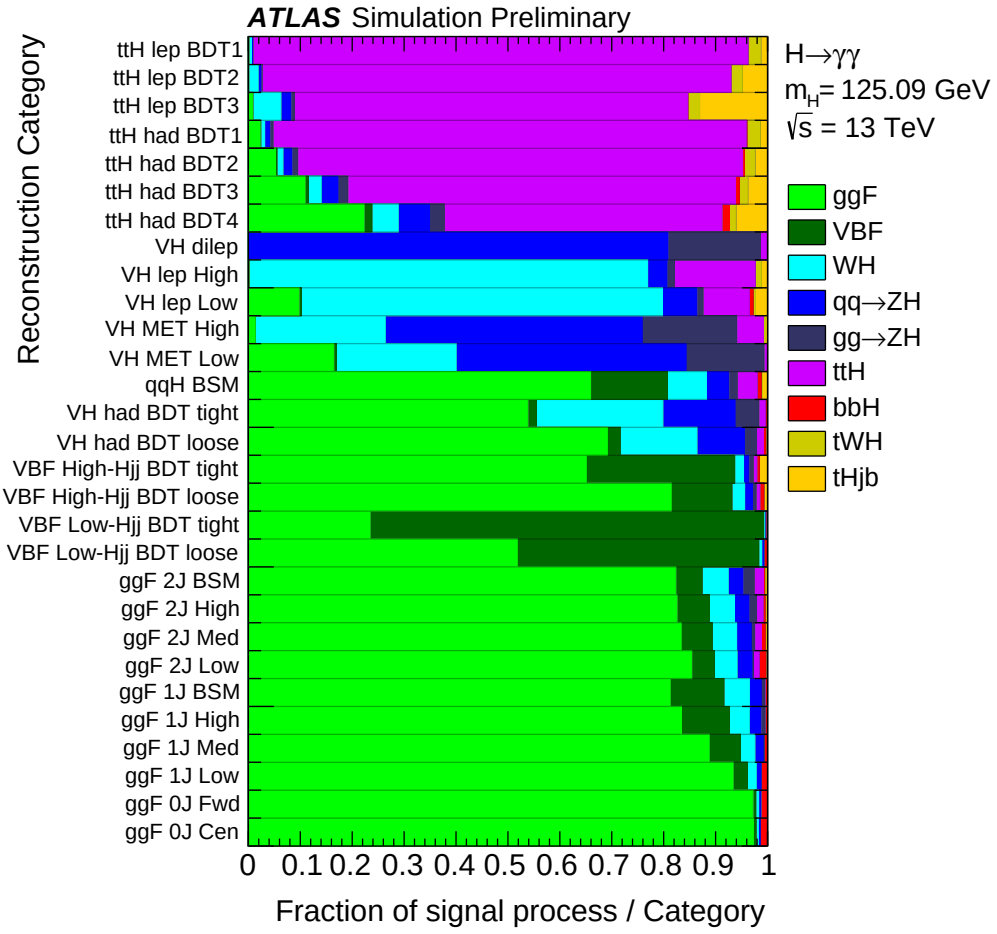


Figure 6.3: The fraction of Higgs boson events in each of the 29 analysis categories, broken down by production mode. The fractions reflect the prediction of the Higgs boson simulation samples.

The full list of categories, along with their selections, is provided in Table 6.1. Additionally, a breakdown of fraction of Higgs boson events from each production mode populating the analysis categories is presented in Figure 6.3, and a breakdown of the fraction of events in each analysis category originating from the STXS truth bins is provided in Figure 6.2. Note that events are sorted into the categories corresponding to the rarest production modes first ($t\bar{t}H$, then VH, then VBF, and finally ggF). In both Table 6.1 and Figure 6.3, the categories are listed in this corresponding order of priority. In order to preserve the orthogonality of the categories, events which are selected for one category will not be included in any other categories.

Index	Category	Selection
33	$t\bar{t}H$ Lep 1	$N_{\text{lep}} \geq 1$, $N_{b\text{-jets}} \geq 1$ ($p_{\text{T}}^{\text{jet}} > 25$ GeV, 77% WP), BDT > 0.987
32	$t\bar{t}H$ Lep 2	$N_{\text{lep}} \geq 1$, $N_{b\text{-jets}} \geq 1$ ($p_{\text{T}}^{\text{jet}} > 25$ GeV, 77% WP), BDT > 0.942
31	$t\bar{t}H$ Lep 3	$N_{\text{lep}} \geq 1$, $N_{b\text{-jets}} \geq 1$ ($p_{\text{T}}^{\text{jet}} > 25$ GeV, 77% WP), BDT > 0.705
30	$t\bar{t}H$ Had 1	$N_{\text{lep}} = 0$, $N_{\text{jets}} \geq 3$, $N_{b\text{-jets}} \geq 1$ ($p_{\text{T}}^{\text{jet}} > 25$ GeV, 77% WP), BDT > 0.996
29	$t\bar{t}H$ Had 1	$N_{\text{lep}} = 0$, $N_{\text{jets}} \geq 3$, $N_{b\text{-jets}} \geq 1$ ($p_{\text{T}}^{\text{jet}} > 25$ GeV, 77% WP), BDT > 0.991
28	$t\bar{t}H$ Had 1	$N_{\text{lep}} = 0$, $N_{\text{jets}} \geq 3$, $N_{b\text{-jets}} \geq 1$ ($p_{\text{T}}^{\text{jet}} > 25$ GeV, 77% WP), BDT > 0.971
27	$t\bar{t}H$ Had 1	$N_{\text{lep}} = 0$, $N_{\text{jets}} \geq 3$, $N_{b\text{-jets}} \geq 1$ ($p_{\text{T}}^{\text{jet}} > 25$ GeV, 77% WP), BDT > 0.911
24	VH Dilep	$N_{\text{lep}} \geq 2$, $70 \leq m_{\ell\ell} \leq 110$ GeV
23	VH Lep High	$N_{\text{lep}} = 1$, $\lvert m_{e\gamma} - 89\rvert > 5$ GeV, $p_{\text{T}}^{\ell+E_{\text{T}}^{\text{miss}}} > 150$ GeV
22	VH Lep Low	$N_{\text{lep}} = 1$, $\lvert m_{e\gamma} - 89\rvert > 5$ GeV, $p_{\text{T}}^{\ell+E_{\text{T}}^{\text{miss}}} < 150$ GeV, $E_{\text{T}}^{\text{miss}}$significance > 1
21	VH MET High	$150 < E_{\text{T}}^{\text{miss}} < 250$ GeV $E_{\text{T}}^{\text{miss}}$significance > 9 or $E_{\text{T}}^{\text{miss}} > 250$ GeV
19	VH MET Low	$80 < E_{\text{T}}^{\text{miss}} < 150$ GeV $E_{\text{T}}^{\text{miss}}$significance > 8
18	qqH BSM	$N_{\text{jets}} \geq 2$ ($p_{\text{T}}^{\text{jet}} > 30$ GeV), $p_{\text{T}}^{\text{jet1}} > 200$ GeV
16	VH Had Tight	$60 < m_{jj} < 120$ GeV, BDT Score > 0.78
15	VH Had Loose	$60 < m_{jj} < 120$ GeV, $0.35 <$ BDT Score < 0.78
14	VBF Tight, High p_{T}^{Hjj}	$\Delta\eta_{jj} > 2$, $\eta^{\text{Zeppenfeld}} < 5$, $p_{\text{T}}^{Hjj} > 25$ GeV, BDT > 0.48
13	VBF Loose, High p_{T}^{Hjj}	$\Delta\eta_{jj} > 2$, $\eta^{\text{Zeppenfeld}} < 5$, $p_{\text{T}}^{Hjj} > 25$ GeV, $-0.5 <$ BDT < 0.48
12	VBF Tight, Low p_{T}^{Hjj}	$\Delta\eta_{jj} > 2$, $\eta^{\text{Zeppenfeld}} < 5$, $p_{\text{T}}^{Hjj} < 25$ GeV, BDT > 0.87
11	VBF Loose, Low p_{T}^{Hjj}	$\Delta\eta_{jj} > 2$, $\eta^{\text{Zeppenfeld}} < 5$, $p_{\text{T}}^{Hjj} < 25$ GeV, $-0.39 <$ BDT < 0.87
10	ggF 2J BSM	$N_{\text{jets}} \geq 2$ ($p_{\text{T}}^{\text{jet}} > 30$ GeV), $p_{\text{T}}^{\gamma\gamma} \geq 200$ GeV
9	ggF 2J HIGH	$N_{\text{jets}} \geq 2$ ($p_{\text{T}}^{\text{jet}} > 30$ GeV), $120 < p_{\text{T}}^{\gamma\gamma} < 200]$ GeV
8	ggF 2J MED	$N_{\text{jets}} \geq 2$ ($p_{\text{T}}^{\text{jet}} > 30$ GeV), $60 < pT^{\gamma\gamma} < 120$ GeV
7	ggF 2J LOW	$N_{\text{jets}} \geq 2$ ($p_{\text{T}}^{\text{jet}} > 30$ GeV), $p_{\text{T}}^{\gamma\gamma} < 60$ GeV
6	ggF 1J BSM	$N_{\text{jets}} = 1$ ($p_{\text{T}}^{\text{jet}} > 30$ GeV), $p_{\text{T}}^{\gamma\gamma} \geq 200$ GeV
5	ggF 1J HIGH	$N_{\text{jets}} = 1$ ($p_{\text{T}}^{\text{jet}} > 30$ GeV), $120 < p_{\text{T}}^{\gamma\gamma} < 200]$ GeV
4	ggF 1J MED	$N_{\text{jets}} = 1$ ($p_{\text{T}}^{\text{jet}} > 30$ GeV), $60 < pT^{\gamma\gamma} < 120$ GeV
3	ggF 1J LOW	$N_{\text{jets}} = 1$ ($p_{\text{T}}^{\text{jet}} > 30$ GeV),$p_{\text{T}}^{\gamma\gamma} < 60$ GeV
2	ggF 0J FWD	$N_{\text{jets}} = 0$ ($p_{\text{T}}^{\text{jet}} > 30$ GeV), One Photon with $\lvert\eta\rvert > 0.95$
1	ggF 0J CEN	$N_{\text{jets}} = 0$ ($p_{\text{T}}^{\text{jet}} > 30$ GeV), Both Photons with $\lvert\eta\rvert \leq 0.95$

Table 6.1: Summary of analysis category definitions. The skipped numbers in the category index column reflect merged categories (17 and 20) and the old tH categories (25-26) which were removed from this iteration of the analysis.

6.4 Dedicated $t\bar{t}H$ Selection

Due to the rarity of the $t\bar{t}H$ process, a dedicated categorization must be carefully constructed in order to observe the process with any significance. The categorization may target the decay products of the top quark pair, as these particles are not a direct decay product of any of the other main Higgs boson production modes. The top quark is the heaviest fundamental particle, and as such it rapidly decays before hadronization (unlike the light quarks). Its decay is governed by the Cabibbo-Kobayashi-Maskawa matrix [99, 100]. This matrix (although not strictly a probability a matrix) implies that the top quark decays to a bottom quark through the weak interaction the vast majority of the time [38].

The dedicated $t\bar{t}H$ selection targets the b quarks from the top quark decays by requiring a b-tagged jet. Although two b-jets are produced in each $t\bar{t}H$ event, the efficiency with which a jet may be identified as coming from a b decay is non-trivially far from 100%. Additionally, jets may be lost due to falling outside of the active detector region. Therefore, requiring both b-jets would result in too low an efficiency for $t\bar{t}H$ signal events.

The decay of the top quark into a b quark also produces a W boson, which then itself decays into either multiple high energy quarks (resulting in jets) or into a lepton and a neutrino. The $t\bar{t}H$ signal region is then divided into two separate preselection regions, each targeting one of these cases. The first is the so-called "hadronic" region, which targets events in which the W boson decays to jets. This region also targets most events where the W decays into a τ lepton and τ neutrino, since the τ particle decays into quarks approximately 65% of the time [38]. The $t\bar{t}H$ hadronic region is defined as events containing at least three jets, where at least one of the jets is tagged as a b-jet.

The second region is the "leptonic" region, which targets events in which the W boson decays into either an electron or muon, along with a neutrino. The W boson only decays in these modes about 20% of the time [38], so this region is expected to have fewer events than the hadronic region. However, the precision afforded by the EM Calorimeter and MS of the ATLAS detector, along with the relatively high accuracy in identifying electrons and muons, mean that this channel may be comparatively pure. The $t\bar{t}H$ leptonic region is defined as events which contain at least one electron or muon, along with at least one b-tagged jet.

Events in both the hadronic and leptonic preselection regions must also pass the diphoton preselection criteria, defined in Section 6.1. Notably, the transverse momentum requirement of the jets considered for the $t\bar{t}H$ selection is 25 GeV (jets from events in non-$t\bar{t}H$ categories must have a p_{T} of at least 30 GeV).

After the definition of the two preselection regions, the analysis selection is further refined through the use of a dedicated BDT in each region. Both BDTs were constructed using the XG-

Boost package [101], and they use mostly "low-level" quantities (primarily object four-vector information) as inputs. Both BDTs are trained and evaluated using simulated $t\bar{t}H$ events as signal and "non-tight, non-isolated" (NTI) data events (those in which one or both photons fails the identification and/or isolation requirements) as a background sample. The details of each BDT are described below in Sections 6.4.2 (hadronic) and 6.4.3 (leptonic).

The strategy behind the BDT-based selection is to sort events by the continuous output score of the BDT. Here, events deemed "signal-like" will have a score closer to 1, while events deemed more "background-like" will have a score closer to 0. A cut may be placed on the BDT score of an event in order to preferentially select signal-like events. The optimal value of the BDT score cut is based on two competing phenomena. First, the tighter (closer to 1) the BDT score cut, the more pure the category. However, for an imperfect BDT, a tighter cut will also lead to fewer signal statistics. Additionally, multiple BDT-based categories may be defined. Because each category is orthogonal (an event has one BDT score, so it may only fall within one BDT-score category), the significances of each category may be added in quadrature. Therefore, defining multiple BDT categories may increase the sensitivity of the analysis. The gains of adding additional BDT categories diminish as the BDT score decreases, however, and so using only a few categories in each preselection region may achieve near-optimal sensitivity while avoiding any over-complication of the analysis.

Table 6.2: The breakdown of the signal (simulated $t\bar{t}H$ events) and background (non-tight, non-isolated data events) subsamples used for training, validation (hyper-parameter/variable optimization and categorization), and testing of the BDTs. The approximate number of events in each subsample is given in parentheses after the percent value. The non-tight, non-isolated data is denoted as "NTI Data."

$t\bar{t}H$ **Hadronic Sample-Breakdown**				
Sample	Training	Hyper-Parameter/ Variable Optimization	Categorization	Testing
$t\bar{t}H$ Signal	60% (605k)	20% (202k)		20% (202k)
NTI Data	60% (56k)	20% (19k)		20% (19k)
$t\bar{t}H$ **Leptonic Sample-Breakdown**				
Sample	Training	Hyper-Parameter/ Variable Optimization	Categorization	Testing
$t\bar{t}H$ Signal	60% (271K)	20% (90k)		20% (90k)
0 b-jet NTI Data	75% (91k)	25% (3.0k)	–	–
$\geq$1 b-jet NTI Data	–	–	50% (0.6k)	50% (0.6k)

$t\bar{t}H$ **Hadronic Preselection**						
Sample	p_{T}^{γ} Cuts	Mass Window [GeV]	N_{jets}	$p_{\mathrm{T}}^{\text{jet}}$ [GeV]	$N_{b-\text{jets}}$	b-tag WP
Training	Flat	$105 < m_{\gamma\gamma} < 160$	≥ 3	> 25	≥ 1	77%
Hyper-Par/ Var Opt						
Categorization	Relative					
Testing						
$t\bar{t}H$ **Leptonic Preselection**						
Sample	Relative p_{T}^{γ} Cuts	Mass Window [GeV]	N_{jets}	$p_{\mathrm{T}}^{\text{jet}}$ [GeV]	$N_{b-\text{jets}}$	b-tag WP
Training	Flat	$80 < m_{\gamma\gamma} < 250$	≥ 1	> 25	$= 0$ (NTI Data)	77%
Hyper-Par/ Var Opt					≥ 1 (MC)	
Categorization	Relative	$105 < m_{\gamma\gamma} < 160$	≥ 1		≥ 1	
Testing						

Table 6.3: A summary of the different preselections applied for the different subsamples used in the BDT construction (training, hyperparameter/variable optimization, categorization, and testing). The selection listed under the "Testing" category reflects that used for the final analysis event selection.

6.4.1 Sensitivity Metrics Used to Evaluate BDT Performance

Two different metrics of sensitivity evaluation were used in order optimize the $t\bar{t}H$ BDTs. These metrics are estimations of the final analysis sensitivity, but they are somewhat simpler and faster than the full analysis framework (the unbinned fits to the invariant mass distribution in all analysis categories, as described in Chapter 9). These estimations were used in order to more efficiently optimize the BDTs. The metrics use two different data samples for various purposes:

- "TI" data events (events in which both photons pass both the isolation and identification criteria)
- "NTI" data events (events in which one or both photons fail either the isolation and/or identification criteria)

The two metrics are:

- Number-counting of TI data side-band events (where the side-bands are defined as $105 < m_{\gamma\gamma} < 120$ GeV and $130 < m_{\gamma\gamma} < 160$ GeV)
- Number-counting of NTI data events

In both methods of significance estimation, the number of "signal" events N_s is equal to the number of TI events from the simulated $t\bar{t}H$ signal sample within the signal window (defined as the diphoton invariant mass window of $123 < m_{\gamma\gamma} < 127$ GeV). However, the method of calculating the number of background events differs. Generally, backgrounds in the signal region can be classified into either non-$t\bar{t}H$ Higgs events (Higgs events coming from production modes other than $t\bar{t}H$ or tH), or continuum background events (all other events). The predicted non-$t\bar{t}H$ Higgs background component is equal to the number of TI events in the signal window, obtained from the simulation samples. The expected number of continuum background events can be determined using one of the two methods listed above.

The method of TI number-counting estimates the number of continuum background events in the signal window based on the number of TI events in the side-bands. First, the number of TI data side-band events is counted, then the number is scaled by the ratio of the number of NTI events in the signal window to the number of NTI events in the side-bands:

$$N_{\text{TI, sw}} = N_{\text{TI, sb}} \times \frac{N_{\text{NTI, sw}}}{N_{\text{NTI, sb}}}$$

where "sb" denotes events in the side-bands and "sw" denotes events in the signal window.

The method of NTI number-counting on the other hand, begins by counting the number of NTI events in the side-bands. This yield is then scaled by the ratio of NTI events in the signal window to that in the side-bands. Due to limited statistics, this scaling factor relating the signal window to side-band yield is taken using only the diphoton preselection (and not specific $t\bar{t}H$ category) yields. In equation form, this can be expressed as:

$$N_{\text{NTI, sw}} = N_{\text{NTI, sb}} \times \frac{N^{\text{presel}}_{\text{NTI, sw}}}{N^{\text{presel}}_{\text{NTI, sb}}}$$

where, again, "sb" denotes side-band events and "sw" denotes signal-window events. The "presel" label indicates that the yield was calculated using only the preselection cuts, not the full $t\bar{t}H$ category cuts.

Once the number of continuum background events, $N_{\text{TI, sw}}$ or $N_{\text{NTI, sw}}$ has been determined, it must be added to the number of TI, non-$t\bar{t}H$ Higgs boson events in the signal window in order to obtain the total background estimation, N_b. The significance is then calculated using the following approximation:

$$Z = \sqrt{2\left[(N_s + N_b)\ln\left(1 + \frac{N_s}{N_b}\right) - N_s\right]}$$

6.4.2 Hadronic $t\bar{t}H$ BDT

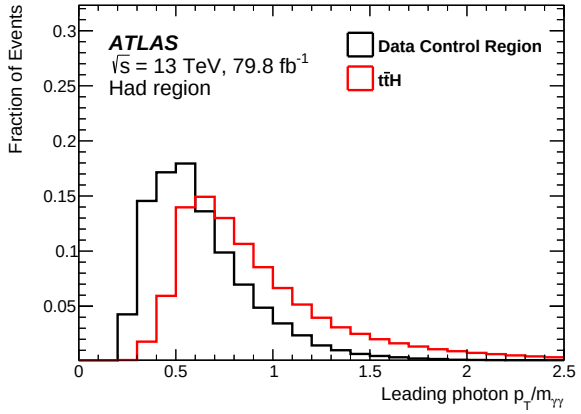
The hadronic BDT is used to analyzes events falling within the so-called "hadronic" $t\bar{t}H$ preselection region. Events in this region are required to pass the diphoton preselection, contain no reconstructed electrons or muons, and contain at least three jets, at least one of which is tagged as originating from a b-quark. The full list of input variables utilized in the BDT is as follows:

- The four-vector information ($p_{\mathrm{T}}/m_{\gamma\gamma}$, η, ϕ, $E/m_{\gamma\gamma}$) of the leading and subleading photons, with p_{T} and energy scaled by the event $m_{\gamma\gamma}$
- The four-vector information (p_{T}, η, ϕ, E) of the first six jets, ordered by jet p_{T}
- The b-tagging status (using the 77% b-tagging working point) of the first six jets, ordered by jet p_{T}
- The magnitude and ϕ of the $E_{\mathrm{T}}^{\mathrm{miss}}$

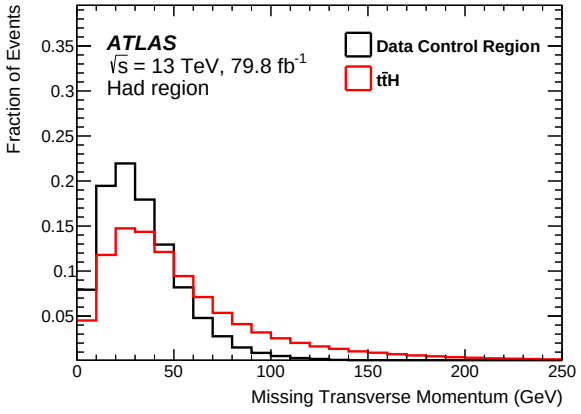
The p_{T} and energy of the photons are scaled by $m_{\gamma\gamma}$ in order to prevent the BDT from learning the invariant mass. If the BDT were to learn the mass, it would preferentially select events, including from background processes, in the signal window. This would cause the continuum background distribution to no longer be smooth, and the interpolation of the background form the side-bands to the signal window would no longer be valid. Examples of some of the input variable distributions used in the $t\bar{t}H$ Hadronic BDT are presented in Figure 6.4.

The granularity at which a BDT can select on different input variables is dependent upon the statistics of the training sample. Ideally, the BDT will be trained on as many events as possible. However, subsets of the signal (simulated $t\bar{t}H$ events) and background (NTI data events) samples must be set aside to test and optimize the BDT. A subset consisting of 20% of the events in each sample are used to optimize the hyper-parameters of the BDT (such as the number of trees and the tree depth) and optimize the choice of input variables. This subset is also used to define the BDT-score boundaries of the analysis categories. Another 20% of the samples is used for testing (to prevent overtraining the BDT) and assessing the expected sensitivity. This leaves 60% of the samples remaining for training the BDT. The breakdown of input samples used for the hadronic BDT is summarized in Table 6.2

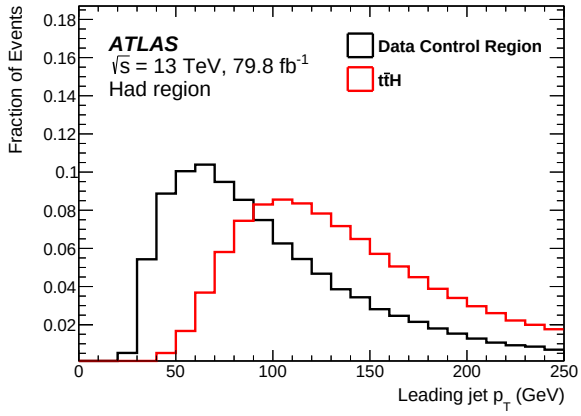
In order to increase the available training statistics, the relative kinematic cuts on the leading and subleading photon are replaced with flat p_{T} cuts of 35 GeV and 25 GeV, respectively. These loosened kinematic cuts are also used within the training and hyper-parameter optimization subsamples. The preselections of the different subsamples used within the hadronic BDT are summarized in Table 6.3.



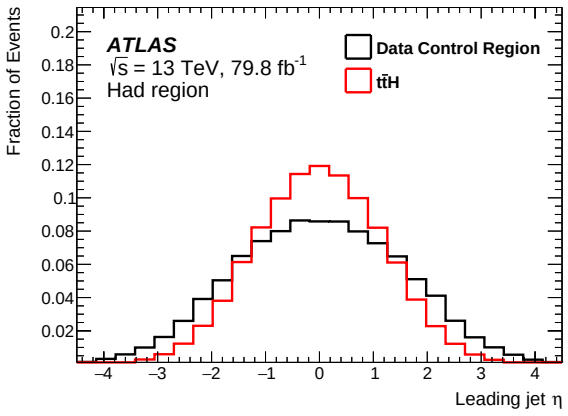
(a) $p_{\mathrm{T}}^{\gamma_1}/m_{\gamma\gamma}$ ($t\bar{t}H$ Hadronic BDT)



(b) $E_{\mathrm{T}}^{\mathrm{miss}}$ ($t\bar{t}H$ Hadronic BDT)



(c) $p_{\mathrm{T}}^{j_1}$ ($t\bar{t}H$ Hadronic BDT)



(d) η^{j_1} ($t\bar{t}H$ Hadronic BDT)

Figure 6.4: Example distributions of the simulated $t\bar{t}H$ signal (red) and NTI data background (black) for some of the hadronic BDT input variables: (a) the p_{T} of the leading photon (scaled by $m_{\gamma\gamma}$), (b) the magnitude of the $E_{\mathrm{T}}^{\mathrm{miss}}$, (c) the p_{T} of the leading jet, and (d) the η of the leading jet.

Table 6.4: The BDT score boundaries of the four hadronic $t\bar{t}H$ categories.

Category	BDT Score Boundaries
Had BDT1	> 0.996
Had BDT2	> 0.991
Had BDT3	> 0.971
Had BDT4	> 0.911

Using the validation sample, four orthogonal categories are defined using cuts on the BDT score. Defining more than four categories was found to have a negligible effect on the projected analysis sensitivity. The determination of the category boundaries was performed by iteratively scanning the categorization subsample for the four BDT score regions with the largest NTI number-counting significance. In the scan, all BDT categories were required to have at least 0.8 TI continuum background events in the signal mass window ($123 < m_{\gamma\gamma} < 127$ GeV) in order to ensure that the side-bands contain enough events to perform a successful background fit. The final BDT score boundaries are listed in Table 6.4.

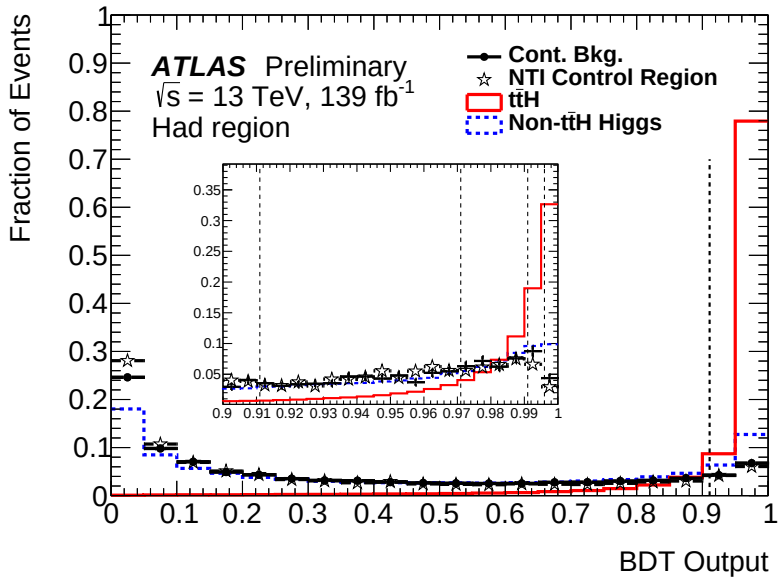
After the categorization is determined, the NTI number counting significance of the test sample is calculated to cross-check against over-training. The hadronic BDT response for the simulated $t\bar{t}H$ signal sample, NTI data, and TI side-band data are shown in Figure 6.5.

6.4.3 Leptonic $t\bar{t}H$ BDT

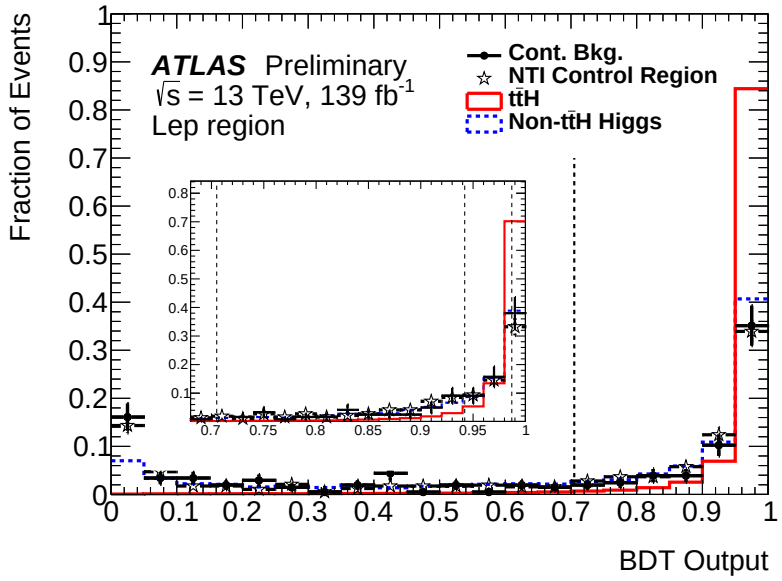
The leptonic BDT is used to analyze events falling within the so-called "leptonic" $t\bar{t}H$ preselection region. Events in this region are required to contain at least one electron or muon, pass the diphoton preselection, and contain at least one jet tagged as originating from a b-quark. The full list of input variables utilized in the BDT is as follows:

- The four-vector information ($p_{\mathrm{T}}/m_{\gamma\gamma}$, η, ϕ, $E/m_{\gamma\gamma}$) of the leading and subleading photons, with p_{T} and energy scaled by the event $m_{\gamma\gamma}$
- The four-vector information (p_{T}, η, ϕ, E) of the first four jets, ordered by jet p_{T}
- The four-vector information (p_{T}, η, ϕ, E) of the first two leptons, ordered by lepton p_{T}
- The magnitude and ϕ of the $E_{\mathrm{T}}^{\mathrm{miss}}$

Again, the p_{T} and energy of the photons are scaled by $m_{\gamma\gamma}$ in order to prevent the BDT from learning the invariant mass. Examples of some of the input variable distributions used in the $t\bar{t}H$ Leptonic BDT are presented in Figure 6.6.

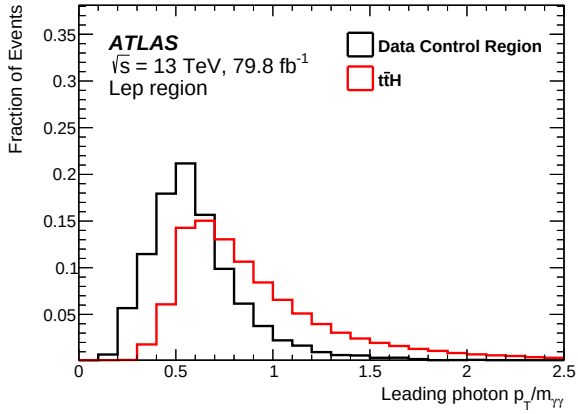


(a) "Had" BDT Response

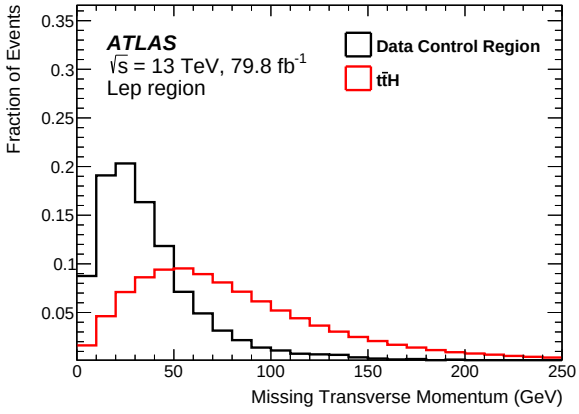


(b) "Lep" BDT Response

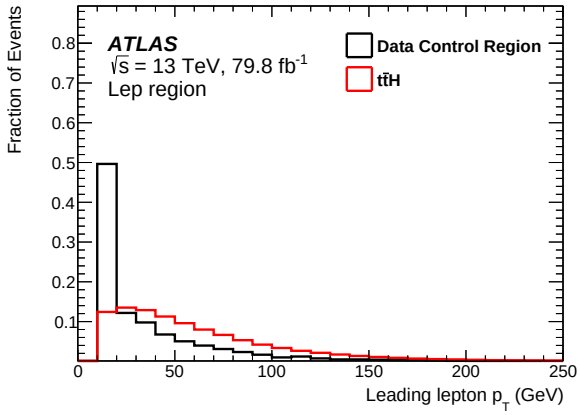
Figure 6.5: The normalized fraction of events in bins of BDT score in the (a) "Hadronic" and (b) "Leptonic" regions of: simulated $t\bar{t}H$ signal events (red); simulated non-$t\bar{t}H$ Higgs boson events (blue); "Not Tight/Isolated" (NTI) data events used as the background sample in testing the BDTs (open stars); and data side-band events (filled black circles). The "Not Tight/Isolated" data events shown are those used in testing the BDTs, and, as such, they are required to pass all cuts in the diphoton and $t\bar{t}H$ preselections, other than the identification and isolation criteria. The dashed lines on the x-axis denote the BDT-score cut of the loosest category in each region. The insets provide a zoomed-in picture of the BDT score distribution for events which are selected for the BDT categories. The dashed lines in the insets denote the category boundaries.



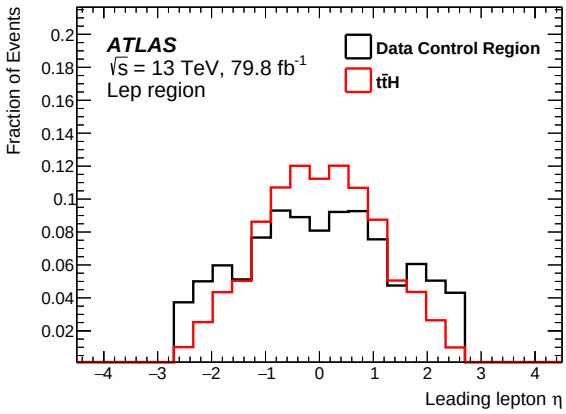
(a) $p_{\mathrm{T}}^{\gamma_1}/m_{\gamma\gamma}$ ($t\bar{t}H$ Leptonic BDT)



(b) $E_{\mathrm{T}}^{\mathrm{miss}}$ ($t\bar{t}H$ Leptonic BDT)



(c) $p_{\mathrm{T}}^{\ell_1}$ ($t\bar{t}H$ Leptonic BDT)



(d) η^{ℓ_1} ($t\bar{t}H$ Leptonic BDT)

Figure 6.6: Example distributions of the simulated $t\bar{t}H$ signal (red) and NTI data background (black) for some of the leptonic BDT input variables: (a) the p_{T} of the leading photon (scaled by $m_{\gamma\gamma}$), (b) the magnitude of the $E_{\mathrm{T}}^{\mathrm{miss}}$, (c) the p_{T} of the leading lepton, and (d) the η of the leading lepton.

Table 6.5: The BDT score boundaries of the three leptonic $t\bar{t}H$ categories.

Category	BDT Score Boundary
Lep BDT1	> 0.987
Lep BDT2	> 0.942
Lep BDT3	> 0.705

Orthogonal subsamples of the signal (simulated $t\bar{t}H$ events) and background (NTI data events) samples are defined for the purposes of training; hyper-parameter and variable optimization; categorization; and testing. In order to increase the available statistics for the training and hyper-paramater/variable optimization, certain cuts are loosened for these two subsamples. First, the requirement on the diphoton invariant mass is loosened to $80 < m_{\gamma\gamma} < 250$ GeV. Again, in as for the hadronic BDT, the flat 35 (25) GeV p_{T} cut on the (sub)leading photon is used to replace the tighter relative p_{T} cuts. Lastly, events in these two subsamples are required to have no b-tagged jets, but have at least one jet. The preselections of the different subsamples used within the leptonic BDT is summarized in Table 6.3.

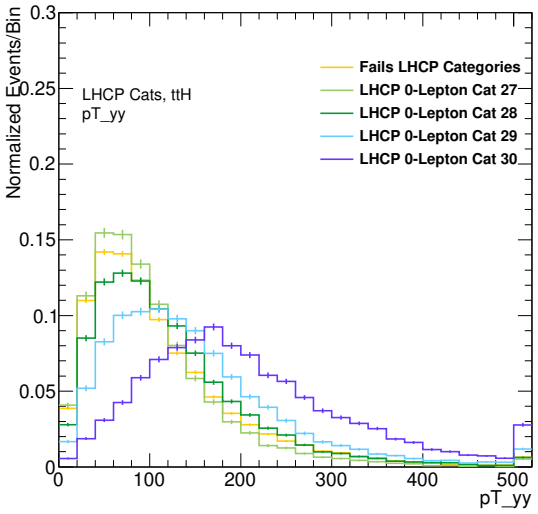
A subset containing 75% of the "zero b-tagged" NTI sample is used for training, and 25% is used for hyper-parameter and variable optimization. Of the "nominal" NTI events (containing at least one b-tagged jet), 50% are used for categorization, and the remaining 50% are used for testing. The breakdown of training, validation, and testing events is shown in Table 6.2.

Using the validation sample, three orthogonal categories are defined scanning the possible BDT score cuts for the maximum significance, as calculated using the NTI number counting significance of the categorization subsample. The BDT score boundaries of these categories are listed in Table 6.5. The leptonic BDT response for the simulated $t\bar{t}H$ sample, the NTI data, and the TI side-band data are shown in Figure 6.5.

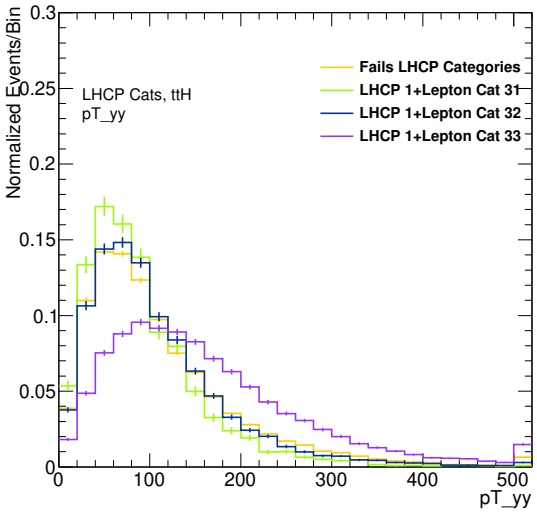
6.4.4 Performance Studies of the $t\bar{t}H$ BDT

The dedicated $t\bar{t}H$ BDTs utilize primarily "low-level" kinematic variables as inputs. The distributions of some of these variables in both the simulated $t\bar{t}H$ signal sample and the NTI data background sample are shown in Figures 6.4 (for the hadronic BDT) and 6.6 (for the leptonic BDT). Significant differences may be observed between the signal and background distributions of many of these variables, which contribute to the BDTs' ability to separate signal and background events effectively.

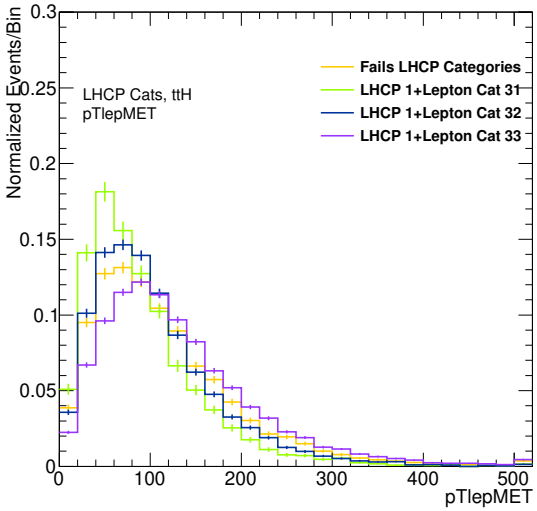
However, some of the BDT performance could be explained by the BDT learning so called "high-level" variables, which may have additional discrimination power between signal and back-



(a) $p_{\mathrm{T}}^{\gamma\gamma}$ ($t\bar{t}H$ Hadronic BDT Categories)



(b) $p_{\mathrm{T}}^{\gamma\gamma}$ ($t\bar{t}H$ Leptonic BDT Categories)



(c) $p_{\mathrm{T}}^{\ell+\mathrm{MET}}$ ($t\bar{t}H$ Leptonic BDT Categories)

Figure 6.7: The distributions of two of the "high-level" variables in the BDT categories: (a) $p_{\mathrm{T}}^{\gamma\gamma}$ in the hadronic BDT categories, (b) $p_{\mathrm{T}}^{\gamma\gamma}$ in the leptonic BDT categories, and (c) $p_{\mathrm{T}}^{\ell+\mathrm{MET}}$ in the leptonic BDT categories. The sculpting of these variables between the BDT categories indicates that the BDT is at least partially selecting on that variable. The category numbers in the legends correspond with those provided in Table 6.1. Category 33 corresponds to the tightest leptonic BDT category, and category 30 corresponds to the tightest hadronic BDT category.

ground events. In order to determine what high-level variables the BDTs were learning, the distributions of variables were compared across analysis categories. Distributions which appear strongly sculpted by the BDT cuts indicate that the variable is being "learned" in some way by the BDT.

One of the most striking distributions is that of $p_{\mathrm{T}}^{\gamma\gamma}$ (the transverse momentum of the diphoton system, equivalent to the Higgs boson p_{T} in signal events), shown in Figures 6.7a (in the hadronic BDT categories) and 6.7b (in the leptonic BDT categories). The importance of $p_{\mathrm{T}}^{\gamma\gamma}$ is reasonable, given that Higgs boson events tend to be more central than background events. Additionally, the sculpting of this variable indicates the separation power obtained through the use of the photon kinematic information in the BDT.

Another highly-sculpted distribution is that of $p_{\mathrm{T}}^{\ell+\mathrm{MET}}$ (the transverse momentum of the lepton and $E_{\mathrm{T}}^{\mathrm{miss}}$ system), shown in Figure 6.7c (note that this variable is only relevant to the leptonic BDT categories). The sculpting of this distribution may help explain some of the performance increase obtained by using the lepton four-vector and $E_{\mathrm{T}}^{\mathrm{miss}}$ information. The lepton and $E_{\mathrm{T}}^{\mathrm{miss}}$ system effectively reflects the kinematics of the W boson in $t\bar{t}H$ decays. Therefore, it is reasonable that the BDT would be able to use these variables to roughly discriminate between events where the $E_{\mathrm{T}}^{\mathrm{miss}}$ and lepton originated from a W boson or from some other source.

CHAPTER 7

Signal and Background Modeling

Both the signal and background shapes in the analysis are modeled using analytic functions. The width of the Higgs boson decay is approximately 4 MeV, leading to a very narrow resonance in the diphoton invariant mass. Due to limitations of the EM calorimeter resolution, the observed signal shape is significantly wider (on the order of about 1 GeV), and its features are dominated by the detector response. The analytic Double-Sided Crystal Ball (DSCB) [59, 102] shape has been found to model the Higgs boson simulation shape with minimum bias. The primary advantage of using an analytic parameterization of the signal shape is in the calculation of systematic uncertainties. These uncertainties may be accounted for by varying specific parameters of the analytic signal function within the signal and background fits (described in Chapter 10). For example, uncertainties in the Higgs boson mass may be measured by varying the mean of the DSCB function.

The dominant background process, continuum diphoton production, produces a smooth, falling spectrum in the diphoton invariant mass. The raw spectrum is a falling exponential distribution, but effects such as trigger thresholds induce additional features in the spectrum. Because the background spectrum is smooth, including underneath the narrow Higgs boson peak, the background contribution directly underneath the signal peak can be interpolated from the regions on either side of the peak in the invariant mass, named the "side-bands." The advantage of interpolating the background contribution from the sidebands is that the background contribution is derived from data instead of simulation. Additionally, the background shape itself then lacks narrow statistical fluctuations.

7.1 Signal Model

The Higgs boson signal shape is modeled with an analytic DSCB [59, 102] shape in each analysis category. The DSCB function models the central portion of the signal as a Gaussian peak and the

outer tails of the signal as power-law curves, with the following functional form:

$$f_{\text{DSCB}}(m_{\gamma\gamma}) = N \times \begin{cases} e^{-t^2/2} & \text{if } -\alpha_{low} \leq t \leq \alpha_{high} \\ \frac{e^{-\frac{1}{2}\alpha_{low}^2}}{\left[\frac{1}{R_{low}}(R_{low}-\alpha_{low}-t)\right]^{n_{low}}} & \text{if } t < -\alpha_{low} \\ \frac{e^{-\frac{1}{2}\alpha_{high}^2}}{\left[\frac{1}{R_{high}}\left(R_{high}-\alpha_{high}+t\right)\right]^{n_{high}}} & \text{if } t > \alpha_{high} \end{cases} \tag{7.1}$$

where μ_{CB} defines the peak position of the Gaussian portion, σ_{CB} describes the Gaussian width, $\alpha_{low/high}$ describes the transition between the Gaussian center and power law tails, and $n_{low/high}$ describe the power of the tails.

In past $H(\gamma\gamma)$ couplings measurements, multiple analytic functions, including the Crystal Ball (CB) function [59], the DSCB function [102], and a CB function with an additional Gaussian core, were tested to fit the simulated signal shape. The DSCB function was found to best model the signal shape.

The analytic function is fit to the combined shape of all simulated Higgs boson samples described in Section 5.2 in a given category, including processes which are not targeted specifically by the category. The yields of the different processes are taken from the simulation prediction. The expected signal resolutions (defined as the width in GeV of the smallest window containing either 68% or 90% of the inclusive Higgs boson signal events) in each of the analysis categories are presented in Table 7.1. Examples of the DSCB fits to two of the $t\bar{t}H$ analysis categories are presented in Figure 7.1. The DSCB fits to the simulated signal samples in all of the analysis categories are presented in Appendix A. Overall, the agreement between the fitted shape and the simulated events appears to be good.

7.2 Background Decomposition

In order to better understand and model the analysis backgrounds, effort was taken to *decompose* the background events into different sub-processes. The first form of decomposition is to separate background events by whether they contain real or "fake" photons. Fake photons are primarily jets which deposit a significant portion of energy into the EM calorimeter, which are then misidentified as photons by both the event reconstruction and offline photon identification algorithms. Events are categorized into three groups: those containing two real photons ($\gamma\gamma$), those containing one real photon and one jet misidentified as a photon (γj and $j\gamma$, depending on whether the jet is misidentified as the leading or subleading photon, respectively), and those containing two jets misidentified as photons (jj). The fraction of $\gamma\gamma$ events in the background sample is defined as the *purity*. This decomposition will be discussed in Section 7.2.1.

Table 7.1: The Higgs boson signal resolution, defined as the width in GeV of the smallest window containing 68% (σ_{68}) and 90% (σ_{90}) of the inclusive Higgs boson signal events, in each of the analysis categories. The $t\bar{t}H$ categories are ordered such that the category with the highest signal purity in each of the "Hadronic" and "Leptonic" regions is labeled as category 1, while that with the lowest signal purity is labeled with the largest number. Note that these are the expected values extracted from the simulation samples at $m_H = 125$ GeV. The $t\bar{t}H$ shapes are normalized to an expected 139 fb^{-1} of data, while the non-$t\bar{t}H$ shapes are normalized to an expected 79.8 fb^{-1} of data.

Index	Category	σ_{68} (GeV)	σ_{90} (GeV)
33	$t\bar{t}H$ "Lep" 1	1.56	2.80
32	$t\bar{t}H$ 'Lep" 2	1.75	3.13
31	$t\bar{t}H$ "Lep" 3	1.85	3.30
30	$t\bar{t}H$ "Had" 1	1.39	2.48
29	$t\bar{t}H$ "Had" 2	1.58	2.84
28	$t\bar{t}H$ "Had" 3	1.65	2.96
27	$t\bar{t}H$ "Had" 4	1.67	3.00
24	VH Dilep	1.7	3.2
23	VH Lep High	1.5	3.0
22	VH Lep Low	1.9	3.4
21	VH MET High	1.6	3.0
19	VH MET Low	1.9	3.6
18	$q\bar{q}H$ BSM	1.5	2.8
16	VH had tight	1.6	2.9
15	VH had loose	1.8	3.2
14	VBF Tight, High p_{T}^{Hjj}	1.7	3.2
13	VBF Loose, High p_{T}^{Hjj}	1.8	3.5
12	VBF Tight, Low p_{T}^{Hjj}	1.7	3.1
11	VBF Loose, Low p_{T}^{Hjj}	1.9	3.5
10	ggF 2J BSM	1.5	2.8
9	ggF 2J HIGH	1.7	3.1
8	ggF 2J MED	1.9	3.4
7	ggF 2J LOW	2.0	3.6
6	ggF 1J BSM	1.4	2.7
5	ggF 1J HIGH	1.7	3.1
4	ggF 1J MED	1.8	3.4
3	ggF 1J LOW	1.9	3.6
2	ggF 0J FWD	2.1	3.8
1	ggF 0J CEN	1.7	3.0

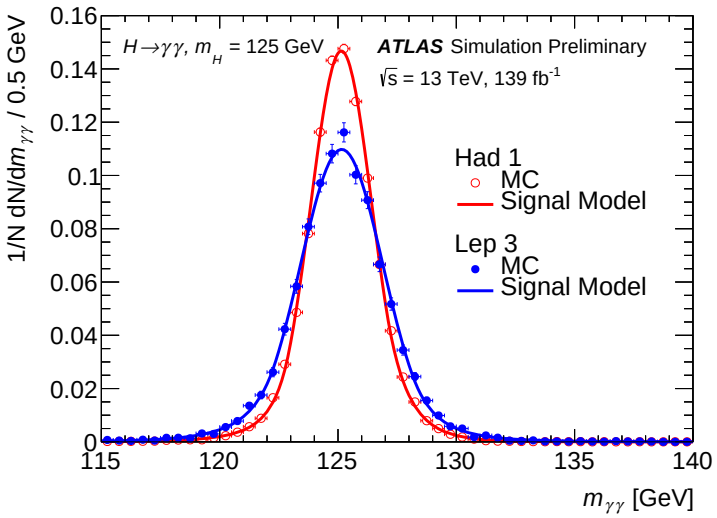


Figure 7.1: Examples of the DSCB fit (solid lines) to the simulated Higgs boson signal shapes (points) in two of the $t\bar{t}H$ analysis categories: $t\bar{t}H$ Had 1 (the tightest $t\bar{t}H$ hadronic category) in red, and $t\bar{t}H$ Lep 3 (the loosest $t\bar{t}H$ leptonic category) in blue. The $t\bar{t}H$ Had 1 category has the narrowest resolution of the $t\bar{t}H$ categories, while $t\bar{t}H$ Lep 3 has the widest.

The decomposition in terms of $\gamma\gamma$, $\gamma j/j\gamma$, and jj events is motivated by the difference in the invariant mass distribution observed between the subsamples. Effectively, the mass distribution of real diphoton events differs from that of events containing one or more fake photons. The discrepancy is due to effects such as the energy dependence of photon identification and isolation efficiencies.

The second decomposition specifically targets the $t\bar{t}H$ categories: separating background events which contain a top quark from those which do not. Measuring this fraction may motivate the treatment of the $t\bar{t}H$ background modeling in terms of relevant simulation samples. Additionally, calculating a significant percentage of events containing top quarks helps validate that the event selection successfully targets the top quark decay products. This decomposition will be discussed in Section 7.2.2.

7.2.1 Background Purity

The purity of the data set is defined as the fraction of background events which contain two *real* photons. The purity within the analysis categories was determined using a so-called 2×2D ABCD side-band method [103, 33]. The method utilizes two (mostly) uncorrelated variables, photon identification and isolation. An "ABCD" grid of regions is constructed based on these variables: (A) the signal region, containing events which pass the *tight* identification and the isolation requirement; (B) a control region containing events which pass the *tight* identification but fail the isolation

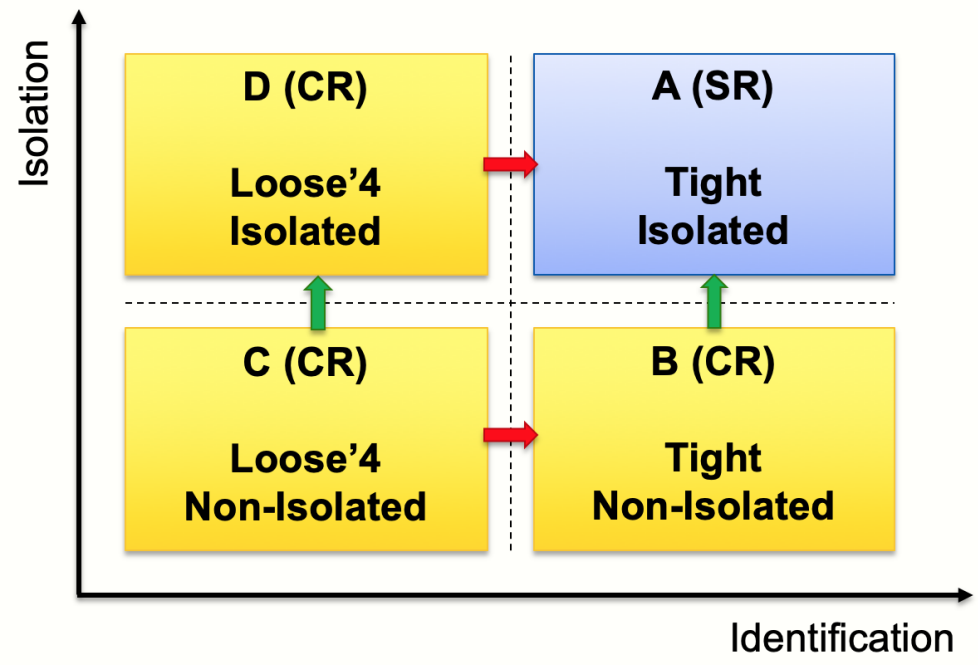


Figure 7.2: An illustration of the ABCD grid, based on the photon isolation and identification variables. The blue box contains the signal region, in which events pass both the photon identification and isolation criteria. The yellow boxes contain orthogonal control regions, in which events fail either the identification, isolation, or both criteria. The green arrows represent the efficiency of the isolation cut, while the red arrows represent the efficiency of the identification cut. In the case where photon isolation and identification are completely uncorrelated, the efficiencies of each color arrow are the same.

requirement; (C) a control region containing events which fail the *tight* identification and fail the isolation requirement; and (D) a control region containing events which fail the *tight* identification but pass the isolation requirement. The *Loose'4* identification criteria, described in Section 4.4, is applied to photons failing the *tight* criteria. Importantly, the control regions are assumed to contain a very small fraction of *real* photon events. A cartoon representation of the 2×2 grid is presented in Figure 7.2.

If the two variables are indeed completely uncorrelated, then the efficiency of the *tight* identification cut will be the same for both isolated and non-isolated photon candidates. Similarly, the efficiency of the isolation cut will be the same for both *tight* and *Loose'4* photon candidates. If the control regions contain only fake photon events, then the relationships between the control regions and the signal regions can be manipulated to obtain the number of *fake* photons in the signal region, A. Therefore, the number of fake photon events in the signal region, $N_{A,\,\text{fake}}$, is equal to:

$$N_{A,\,\text{fake}} = \epsilon_T N_D = \epsilon_I N_B = \epsilon_T \epsilon_I N_C$$

where N_B, N_C, and N_D are the number of events in control regions B, C, and D, respectively; ϵ_T is the *tight* identification efficiency; and ϵ_I is the isolation efficiency.

The above ABCD grid applies to a single photon in an event. If the behavior of the two photons in a diphoton event is completely uncorrelated, then a second, independent ABCD grid may be constructed for the second photon (hence the "2×2D" portion of the method name). However, for this analysis, there exist non-trivial correlations between the leading and subleading photons. Additionally, there also exist non-negligible correlations between the photon identification and isolation variables. As a result, the full 2×2D ABCD method involves sixteen input equations with nineteen variables. Of these variables, four are fixed based on the simulated diphoton sample:

- (ϵ_{T2}) ϵ_{T1}, the tight-identification efficiency of the (sub)leading photon
- (ϵ_{I2}) ϵ_{I1}, the tight-isolation efficiency of the (sub)leading photon

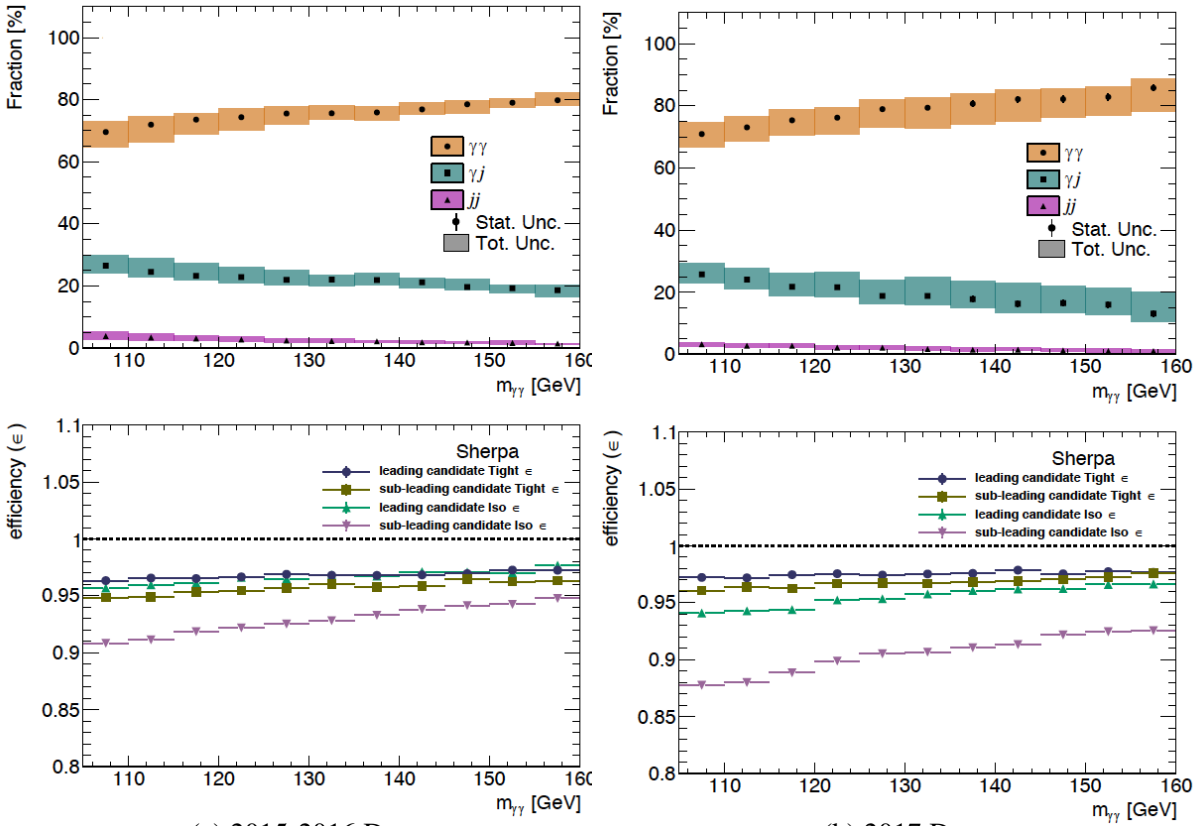
In addition, two variables are fixed to 1.0, then allowed to vary in the range of ± 0.01 in order to quantify systematic impacts:

- (ξ_{j2}) ξ_{j1}, the correlation between the identification and isolation fake rate for the (sub)leading jet

The remaining variables are allowed to float in a fit to the best describe the number of measured events in each of the ABCD regions of the leading and subleading photon grids.

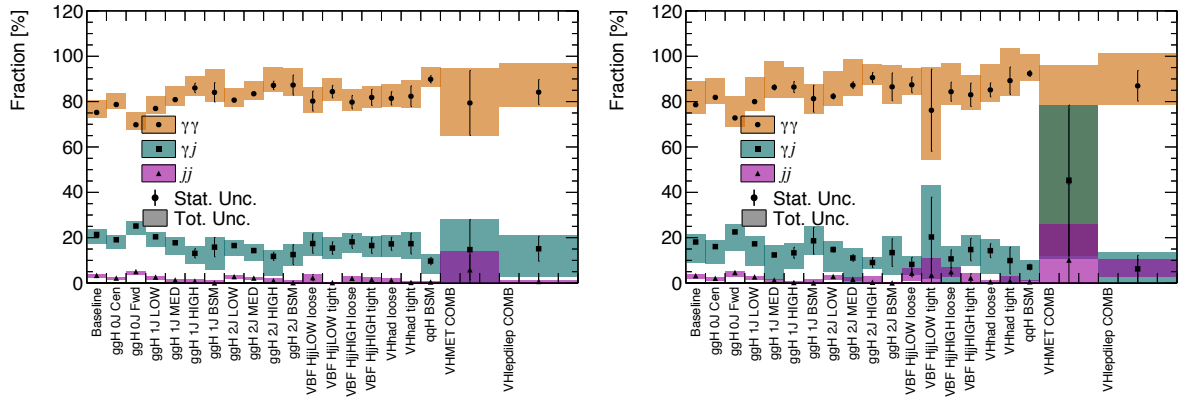
The inclusive purities for both the 2015-2016 and 2017 data taking periods are presented in Figure 7.3, along with the extracted identification and isolation efficiencies for the leading and subleading photons.The changes in the trigger used and the pileup conditions of the data were expected to have a non-negligible impact on the data purity, which can be observed in the figure. The purity was also calculated within most of the individual analysis categories, again separately for the 2015-2016 and 2017 data taking periods. The purities of the non-$t\bar{t}H$ categories are presented in Figure 7.4. The individual VH MET and VH Lep categories are combined when calculating the purity, again due to limited simulation statistics.

In addition, the relative $\gamma\gamma$, γj, and jj fractions in the 2015-2016 and 2017 data taking periods were estimated using the 2×2D ABCD method within the $t\bar{t}H$ analysis preselection regions, described in Section 6.4. The 2018 data taking period is assumed to be similar to 2017 period (both periods use the same trigger and have similar pileup profiles), although this compatibility was not explicitly tested here. The relative purities in the $t\bar{t}H$ regions are shown in Figures 7.5. The simulated diphoton sample contains few events passing the $t\bar{t}H$ hadronic or leptonic preselection requirements. Therefore, the purity estimations contain large statistical errors and should only be considered as a rough estimation.



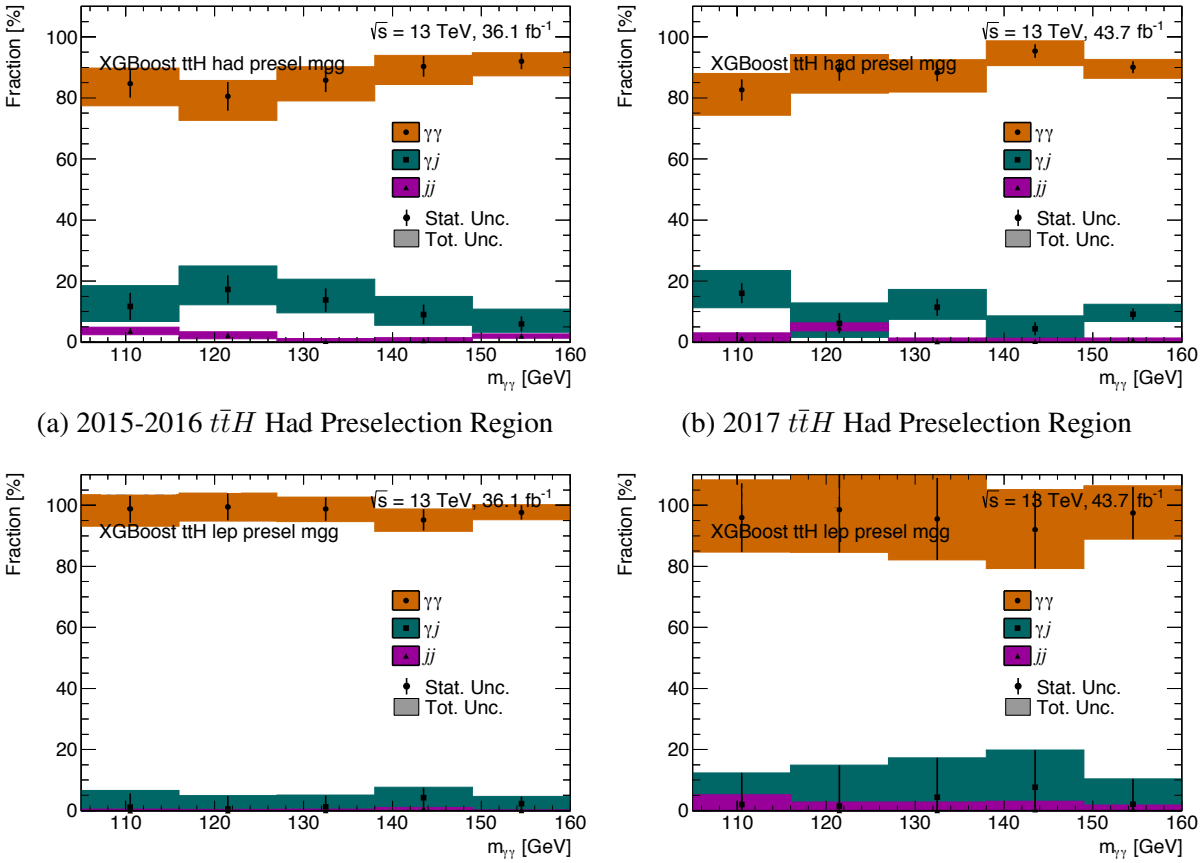
(a) 2015-2016 Data (b) 2017 Data

Figure 7.3: The inclusive $\gamma\gamma$, γj, and jj purities in bins of $m_{\gamma\gamma}$ for the (left) 2015-2016 and (right) 2017 data taking periods, as obtained by the 2×2D ABDC Sideband method. Below are presented the identification and isolation efficiencies of both the leading and subleading photons as obtained from simulated diphoton events.



(a) 2015-2016 Data (b) 2017 Data

Figure 7.4: The $\gamma\gamma$, γj, and jj purities within a selection of the analysis categories for the (a) 2015-2016 and (b) 2017 data taking periods, as obtained by the 2×2D ABDC Sideband method.



(a) 2015-2016 $t\bar{t}H$ Had Preselection Region

(b) 2017 $t\bar{t}H$ Had Preselection Region

(c) 2015-2016 $t\bar{t}H$ Lep Preselection Region

(d) 2017 $t\bar{t}H$ Lep Preselection Region

Figure 7.5: The $\gamma\gamma$, γj, and jj purities in bins of $m_{\gamma\gamma}$ for the 2015-2016 and 2017 data taking periods within the $t\bar{t}H$ preselection regions, as obtained by the 2x2D ABCD side-band Sideband method. The left plots show the 2015-2016 data taking period, while the right plots show the 2017 data taking period. The top row shows the hadronic $t\bar{t}H$ preselection region, while the bottom row shows the leptonic preselection region.

7.2.2 Events Containing Top Quarks in the Hadronic $t\bar{t}H$-Enriched Region

A method for estimating the background composition, called the "cluster template method," was explored for the $t\bar{t}H$ portion of the analysis. This method uses a Machine Learning Clustering technique to obtain a simplified template based on a multidimensional space and then performs a template fit. This Cluster Template method was used to roughly estimate the relative fraction of so-called "real-top" ($t\bar{t}\gamma\gamma$ events) and "non-top" backgrounds (continuum diphoton events) in the $t\bar{t}H$ hadronic preselection region.

This method relies on a machine learning Agglomerative Clustering algorithm in the Python Scikit-Learn library [104]. It is a hierarchical clustering algorithm which begins by defining each event as a unique cluster. Pairs of clusters are then merged in such a way that the total variance within each cluster is minimized. This merging is performed until the desired final number of clusters is reached.

The cluster template method was tested using the data subsample taken in 2016, along with the subset of simulated continuum diphoton ($\gamma\gamma$), single photon (γj), $t\bar{t}\gamma\gamma$, $t\bar{t}\gamma j$ ($t\bar{t}\gamma\gamma$ with a jet faking a photon), and $t\bar{t}H$ Higgs boson events. The following input variables are used by the clustering algorithm:

- m_{T}, the transverse mass of the event
- $p_{\mathrm{T}}^{\ell,\mathrm{MET}}$, the transverse momentum of the lepton and $E_{\mathrm{T}}^{\mathrm{miss}}$ system
- N_{jet}, the jet multiplicity (for jets with $p_{\mathrm{T}}^{\mathrm{jet}} > 25$ GeV)
- $N_{\mathrm{jet,\,c}}$, the central jet multiplicity ("central" defined as $-2.5 < \eta^{\mathrm{jet}} < 2.5$) (for jets with $p_{\mathrm{T}}^{\mathrm{jet}} > 25$ GeV)
- $N_{b-\mathrm{jet}}$, the b-tagged jet multiplicity (for b-tagged jets with $p_T^{\mathrm{jet}} > 25$ GeV, using the 77% working point)
- The b-tagging status of the leading two jets, ordered by p_T^{jet}

The input variables are normalized such that the range of all input values for each variable is equal to one. This prevents the scale of each variable from having an effect on how much it contributes to the total variance within the clusters. A loose pre-selection requirement of $N_{\mathrm{jets}} > 1$ is applied on the data and simulation samples.

The agglomerative clustering algorithm is first performed using the data events in order to define clusters. Once the clusters are defined, the Monte Carlo events are assigned a cluster label using the Scikit-Learn NearestNeighbors algorithm [104]. This algorithm assigns an event to a cluster by taking a vote of the nearest $N_{\mathrm{neighbors}}$ data points in the multidimensional normalized

input space. The value of $N_{\text{neighbors}}$ is set to 25, but it generally does not affect performance strongly so long as $N_{\text{neighbors}} \ll N_{\text{events}}$. The clustering of the TI data (events where both photons pass the identification and isolation criteria); the NTI data (events where both photons fail the identification and isolation criteria); and the $\gamma\gamma$ and $t\bar{t}\gamma$ and $t\bar{t}\gamma\gamma$ simulation samples can be seen in Figure 7.6.

Once the data and simulated events have been assigned into clusters, the relative fraction of events from each sample within each cluster is calculated. This relative distribution within the clusters is defined as the cluster template. Statistical errors are assigned for each sample as $1/\sqrt{N_i}$ for a cluster i with N_i events.

Assuming all relevant background types are included in the simulation cluster template, the data cluster template is by definition equal to a weighted sum of the simulation sample cluster templates. The weights then represent the fraction of each background type present within the data. This relationship can be expressed in equation form as:

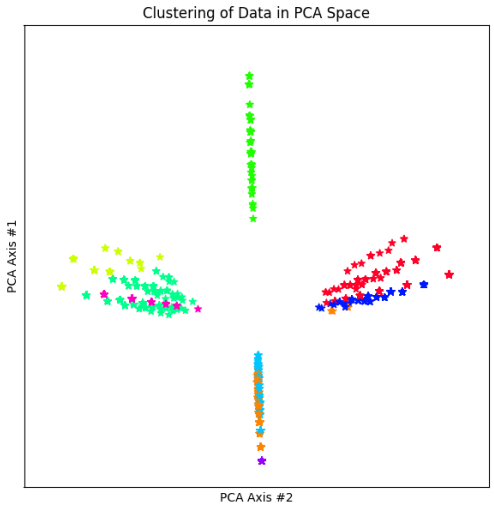
$$C_{\text{data}} = \Sigma_{i_{\text{sample}}=1}^{N_{\text{simulation samples}}} w_{i_{\text{sample}}} \times C_{i_{\text{sample}}}$$

where C denotes the vector (of length N_{Clusters}) of relative composition in each cluster for either data or a simulation sample. A binned fit is then performed where the data and simulation cluster templates, C_{data} and the $C_{\text{simulation samples}}$, are fixed, and the weights $w_{i_{\text{sample}}}$ are left free. The fitted values of $w_{i_{\text{sample}}}$ then give the final estimate of the amount of each background sample found in the data.

The resulting cluster templates are shown in Figure 7.7. Here, nine clusters are used; the value of each bin represents the fraction of events in a sample which belong to that cluster. The method predicts that in the hadronic $t\bar{t}H$ preselection region, the non-tight, non-isolated data contains $91.8\% \pm 12.3\%$ non-top ($\gamma\gamma$ and γj) backgrounds and $8.2\% \pm 1.8\%$ top ($t\bar{t}\gamma\gamma$ and $t\bar{t}\gamma j$) backgrounds. The tight, isolated data is predicted to contain $90.9\% \pm 20.4\%$ non-top backgrounds and $9.0\% \pm 16.1\%$ top backgrounds. The limited statistics of the tight, isolated data lead to large statistical errors on the results. However, the background composition estimates for the non-tight, non-isolated and the tight, isolated data are compatible.

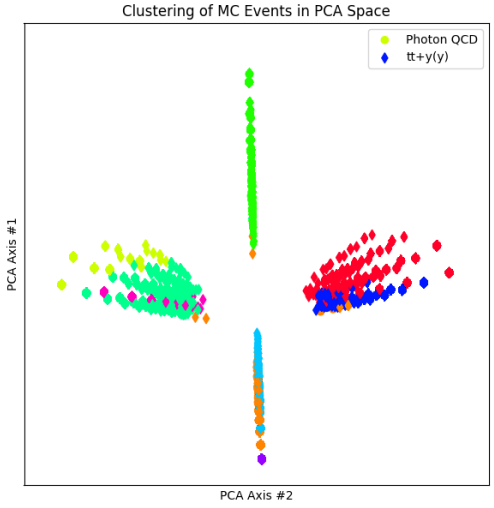
7.3 Background Modeling

Because the continuum diphoton background produces a smooth, falling shape in in the diphoton invariant mass, an analytic function is used in order to describe the background shape. The Higgs boson signal, on the other hand, is narrow, and so signal events are confined to a small window about the Higgs boson mass in the invariant mass spectrum. The continuum background is smooth over a large mass range, including underneath the signal peak. Therefore, the number of



(a) NTI Data Clusters

(b) TI Data Clusters



(c) Simulation Clusters

Figure 7.6: The clusters, in principle component analysis space, of the (a) non-tight, non-isolated ("NTI") data, (b) tight, isolated ("TI") data (top right), and simulated background samples (bottom). Each color represents a different cluster, and each marker shape in the simulation cluster plot represents a different simulated sample. The axes of all three plots are equivalent, and are derived from a Principle Component Analysis (PCA). Note that the PCA axes are used solely for plotting purposes.

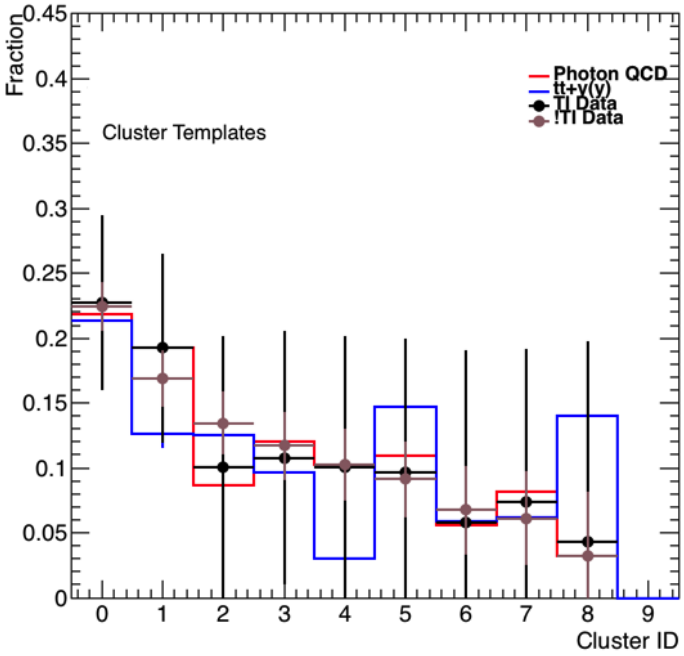


Figure 7.7: The Cluster Templates of the simulated $\gamma\gamma$ and γj samples (red), $tt\gamma j$ and $tt\gamma\gamma$ MC (blue), the tight, isolated ("TI") data (black), and non-tight, non-isolated ("NTI") data (brown).

background events underneath the Higgs boson signal peak can be estimated by interpolating the smooth background from the background-pure side-band regions.

Continuum diphoton events are expected to have a falling exponential shape. However, experimental effects, such as trigger thresholds and the energy-dependence of the photon identification efficiency, induce additional features in the spectrum. Therefore, a number of "exponential-like" functions are tested to model the background. The choices for the analytic function that have been considered are:

- Exponential Function: $f(m_{\gamma\gamma}) = e^{c \cdot m_{\gamma\gamma}}$
- Exponential Function of 2^{nd} Order Polynomial: $f(m_{\gamma\gamma}) = e^{c_1 \cdot m_{\gamma\gamma}^2 + c_2 \cdot m_{\gamma\gamma}}$
- Exponential Function of 3^{nd} Order Polynomial: $f(m_{\gamma\gamma}) = e^{c_1 \cdot m_{\gamma\gamma}^3 + c_2 \cdot m_{\gamma\gamma}^2 + c_3 \cdot m_{\gamma\gamma}}$
- Bernstein polynomial of order N ($3 \leq N \leq 5$): $B_N(m_{\gamma\gamma}) = \sum_{i=0}^{N} c_i \cdot b_{i,N}$ with $b_{i,N} = \binom{N}{i} m_{\gamma\gamma}^i (1 - m_{\gamma\gamma})^{N-i}$
- First-Order Power Law Function: $f(m_{\gamma\gamma}) = m_{\gamma\gamma}^c$

Typically, the experimental features in the spectrum are relatively small, and they may be hidden by statistical uncertainties. Therefore, categories with small statistics are usually sufficiently modeled

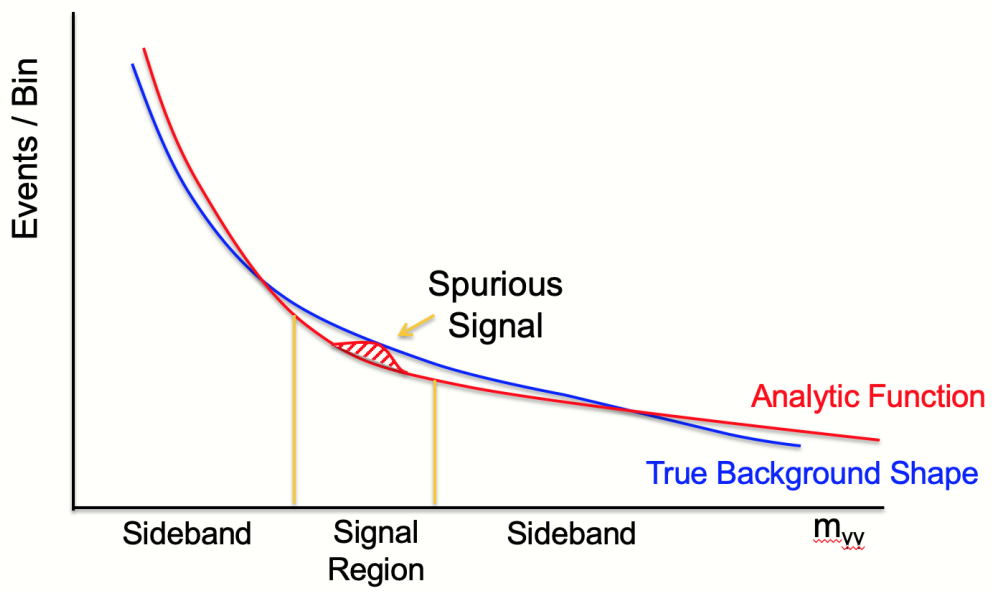


Figure 7.8: A cartoon illustration of the spurious signal test. The "true" shape of the continuum background is shown in blue. An analytic signal and background model, shown in red, is fit to the observed background. Due to the mismodeling of the real shape by the analytic background function, a non-zero signal is fit as well, in order to improve the overall fit quality. In this manner, the background mismodeling is parameterized in terms of a signal yield.

using analytic functions with fewer parameters. In categories with very high statistics, these effects may become more pronounced compared to the statistical uncertainties, and so more degrees of freedom are needed to adequately describe the background spectrum.

The optimal functional form for each category is chosen using the *spurious signal test*. This test in essence determines the analytic function which best models the continuum background spectrum, and it parameterizes the mismodeling in terms of a number of fake signal events induced by that mismodeling. To perform the spurious signal test, the full analytic signal plus background model is fitted to a background-only template in each individual analysis category. Because the sample contains only background events by definition, the fitted signal is actually just a relic of the discrepancy between the analytic function and the real background shape (hence the use of the term "spurious" signal). The number of fitted spurious signal events as a function of the Higgs mass is scanned in intervals of 1 GeV within the range of $121 \leq m_{\gamma\gamma} \leq 129$ GeV. The number of spurious signal events N_{sp} is then defined as the maximum of the absolute value of the fitted number of signal events within the signal mass sub-range (note that N_{sp} is allowed to be negative). A cartoon illustration of the test at a single mass point is presented in Figure 7.8.

The choice of analytic function is made such that the function satisfies at least one of two specific criteria relating to the real Higgs boson signal in each category. The first criteria requires that the fitted spurious signal must be less than 10% of the expected number of *real* signal events

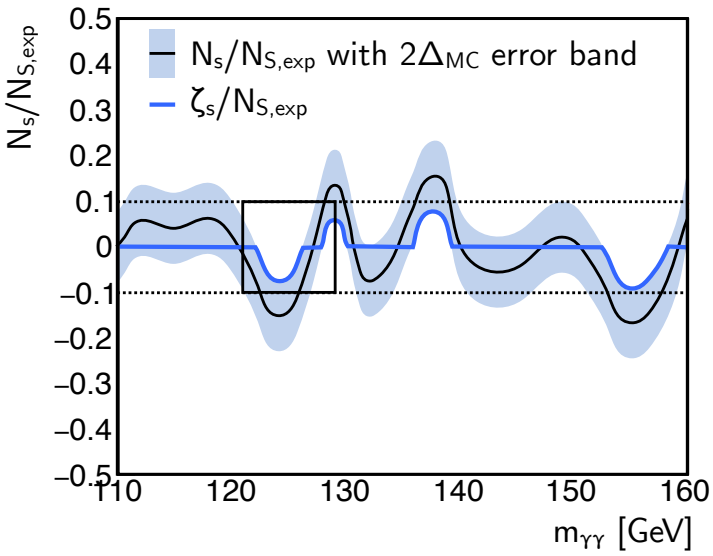


Figure 7.9: A cartoon illustration of the *relaxed* spurious signal quantity, ζ_{sp}.

in a category. The second criteria requires that it must be less than 20% of the statistical error on the expected number of *real* signal events in a category.

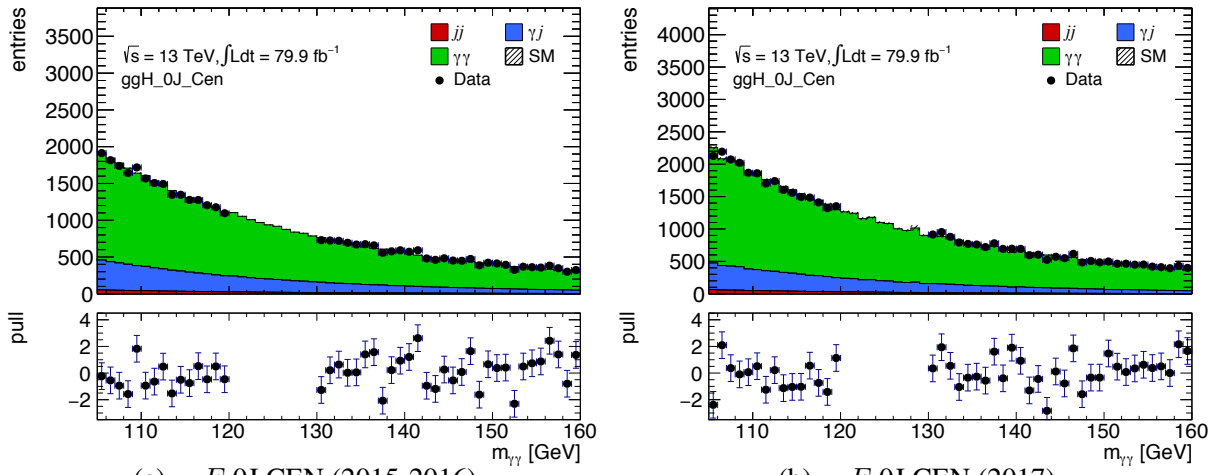
The background-only simulation samples of some of the categories (especially those measuring the VH and $t\bar{t}H$ processes) are statistically limited. The resulting large statistical fluctuations may cause all tested functions to fail the two criteria of the spurious signal test. In order to address these cases, a *relaxed* spurious signal criteria is used, using the variable ζ_{sp}:

$$\zeta_{sp} = \begin{cases} N_{sp} + 2\Delta_{Bkg} & \text{when} N_{sp} + 2\Delta_{Bkg} < 0 \\ N_{sp} - 2\Delta_{Bkg} & \text{when} N_{sp} - 2\Delta_{Bkg} > 0 \\ 0 & \text{otherwise} \end{cases}$$

where the Δ_{Bkg} variable represents the statistical uncertainty on the background-only template. The behavior of ζ_{sp} is demonstrated in Figure 7.9. The new ζ_{sp} quantity is then required to pass one of the two criteria outlined above, instead of the raw N_{sp} value.

It is possible that multiple functions may pass the spurious signal test. In this case, the function with the fewest parameters is preferred. If two functions with the same number of parameters pass the test, then the one with the least resulting spurious signal is chosen. Once the analytic background function has been chosen for a category, the value of N_{sp} taken as a systematic uncertainty on the measured signal yield of that category.

For the ggF, VBF, and VH Had categories, the background-only templates used for the spurious signal test are obtained from the simulated continuum diphoton sample, described in



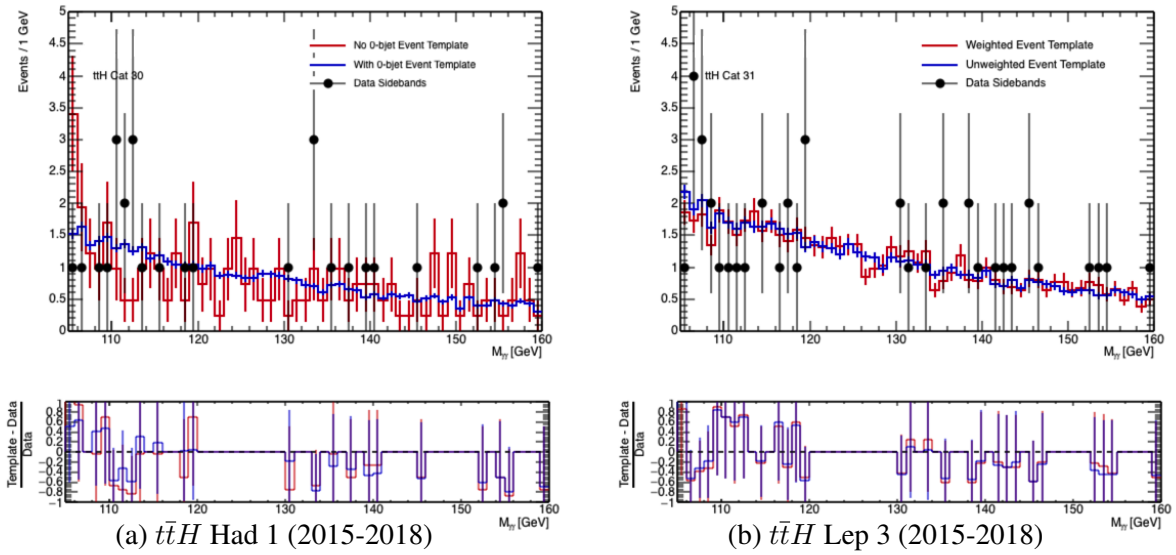
(a) ggF 0J CEN (2015-2016) (b) ggF 0J CEN (2017)

Figure 7.10: The 2015-2016 (left) and 2017 (right) background templates, along with the data side-band events in black, for the ggF 0J Cen category. The green shows the $\gamma\gamma$ events, the blue shows the γj events, and the red shows the jj events. The bottom panel of each plot shows the pull between the template and the data sideband events (the difference between the two, quantified in terms of the statistical error of the side-bands).

Section 5.2 (the number of events from this sample in the VH MET, VH Lep, and $t\bar{t}H$ categories is insufficient to successfully perform the test). The relative fractions of $\gamma\gamma$, γj (where one photon is faked by a jet), and jj (where both photons are faked by a jet) events are obtained using the 2×2D ABCD method, described in detail in Section 7.2.1. The approximate shape of the γj component is estimated by looking at the data shape in the case where one of the photons fails the *tight* identification requirement, but passes the *Loose'4* photon identification criteria (as described in Section 4.4). Similarly, the jj shape is estimated from events where both photons fail the *tight* identification requirement, but pass the *Loose'4* requirement. The simulated diphoton events passing the category selection are then re-weighted by a linear factor to correct the simulated shape to that obtained from the γj or jj data shape. The $\gamma\gamma$, γj, and jj shapes are then added together, with each component weighted by the relative fraction estimated in the *tight*, isolated data region.

Examples of the background templates obtained for the 2015-2016 and 2017 periods in the ggF 0J CEN category are presented in Figure 7.10. The remaining ggF, VBF, and VH Had background templates are presented in Appendix B. The agreement between the templates and the data side-bands is satisfactory, and as such it validates the use of the templates in describing the background shape in the region underneath the Higgs boson signal peak.

Due to limited statistics in the simulated diphoton samples, the background templates in the VH MET and VH Lep categories were constructed from the simulated $V + \gamma\gamma$ events. No break-down of the $\gamma\gamma$, γj, and jj sub-fractions was attempted. The background templates for the VH



(a) $t\bar{t}H$ Had 1 (2015-2018) (b) $t\bar{t}H$ Lep 3 (2015-2018)

Figure 7.11: The full Run 2 background-only templates (using the data control region events failing the identification and/or isolation requirement, without the b-jet requirement) for the (a) $t\bar{t}H$ Had 1 and (b) $t\bar{t}H$ Lep 3 categories, shown in blue. The data side-band events are shown as black dots. The red shape shows an alternative template obtained using only the data control region events which do pass the b-jet requirement applied to the nominal analysis categories. The apparent compatibility between the nominal (blue) and tighter alternative (red) templates confirms the validity of using the loosened sample. The bottom panel shows the fractional difference between the two templates and the data side-bands.

MET and Lep categories are presented in Appendix B. As for the ggF, VBF, and VH Had categories, the agreement between the background templates and the data side-bands appears acceptable, especially given the limited data statistics of the VH MET and VH Lep categories.

The $t\bar{t}H$ categories contain also too few simulated continuum diphoton events to provide useful background templates for the spurious signal test. The background templates for the $t\bar{t}H$ Had categories are instead constructed using the a data control region in which one or both both photons fail the tight identification or isolation requirement. The $t\bar{t}H$ Had preselection is loosened for the background template, as well – events are not required to contain a b-tagged jet. In the $t\bar{t}H$ Lep categories, simulated $t\bar{t}\gamma\gamma$ events are used. In order to gain statistics in the background templates, the simulated $t\bar{t}\gamma\gamma$ events are not required to pass the photon identification or isolation criteria, and the event weights are not applied. Due to the limited number of data events in the $t\bar{t}H$ categories, only the exponential and power law functions were tested to model the continuum background.

Examples of the background templates used for two of the $t\bar{t}H$ categories ($t\bar{t}H$ Had 1 and $t\bar{t}H$ Lep 3) are presented in Figure 7.11. The figure also includes "alternative" background templates used to validate the loosened constraints (applied in order to increase statistics) discussed above. The alternative background templates in the $t\bar{t}H$ hadronic categories were obtained by imposing the b-jet cut ($N_{b-\text{jets}} \geq 1$) on the data control region events. The alternative background templates in the $t\bar{t}H$ leptonic categories were obtained by using the event weights provided in the simulated

$t\bar{t}\gamma\gamma$ sample. The agreement between the nominal and alternative background templates appears satisfactory. Combined with the reduced statistical power observed in the alternative templates, the agreement helps motivate the use of the looser, higher-statistics template options. Both the nominal and alternative background templates constructed in all of the $t\bar{t}H$ categories are presented in Appendix B. The nominal and alternative background templates appear to acceptably agree with the data side-bands in all categories, although the limited statistics of many $t\bar{t}H$ categories make a direct comparison challenging.

The chosen analytic background function, as well as the resulting value of the spurious signal systematic uncertainty, for each analysis category is presented in Table 7.2.

Category	Function	N_{sp}
ggF 0J CEN	ExpPoly2	-48.3
ggF 0J FWD	ExpPoly2	-228
ggF 1J LOW	ExpPoly2	45.9
ggF 1J MED	ExpPoly2	40.7
ggF 1J HIGH	PowerLaw1	-9.79
ggF 1J BSM	Exponential	-2.63
ggF 2J LOW	ExpPoly2	-38.1
ggF 2J MED	ExpPoly2	21.7
ggF 2J HIGH	PowerLaw1	7.42
ggF 2J BSM	Exponential	0.426
VBF Loose, Low p_{T}^{Hjj}	PowerLaw1	10.3
VBF Tight, Low p_{T}^{Hjj}	Exponential	5.16
VBF Loose, High p_{T}^{Hjj}	PowerLaw1	-12.2
VBF Tight, High p_{T}^{Hjj}	PowerLaw1	-17.4
VH Had Loose	PowerLaw1	6.45
VH Had Tight	Exponential	-1.81
qqH BSM	Exponential	-11.7
VH MET Low	Exponential	2.45
VH MET High	PowerLaw1	1.94
VH Lep Low	Exponential	-4.67
VH Lep High	Exponential	-0.598
VH Dilep	PowerLaw1	1
$t\bar{t}H$ Had 4	PowerLaw	1.07
$t\bar{t}H$ Had 3	PowerLaw	0.658
$t\bar{t}H$ Had 2	PowerLaw	0.756
$t\bar{t}H$ Had 1	Exponential	0.316
$t\bar{t}H$ Lep 3	Exponential	0.165
$t\bar{t}H$ Lep 2	Exponential	0.430
$t\bar{t}H$ Lep 1	Exponential	0.241

Table 7.2: Background modeling functions selected for all of the analysis categories, as well as the value of the spurious signal systematic uncertainty.

CHAPTER 8

Reducing Background Modeling Systematics Using Gaussian Processes

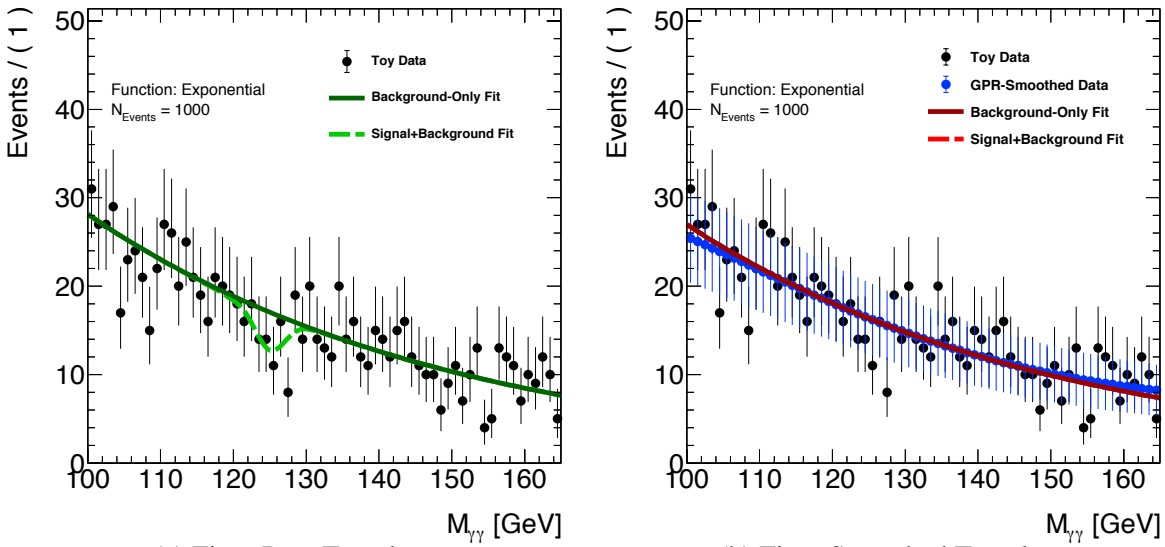
The signal and background shapes used in the fit to data in each of the analysis categories are derived from analytic functions, as described in Sections 7.1 and 7.3. However, the true distribution of the continuum background events may not be perfectly modeled by the choice of analytic function. In this case, background mismodeling may induce a so-called "spurious" signal, or background events fitted as signal incorrectly in order to improve the quality of the signal and background fit. The spurious signal uncertainty is described in detail in Section 7.3.

Because of the limited statistics of the simulation samples used in the calculation, the value of the spurious signal systematic uncertainty is subject to significant statistical fluctuations within many of the analysis categories. Statistical fluctuations in the background-only sample may cause signal-like bumps, which are then fit as spurious signal events. These statistical fluctuations do not capture the shape mismodeling from the analytic function, and they often drastically inflate the value of the systematic. An illustration of a statistical fluctuation inflating the value of the spurious signal uncertainty is presented in Figure 8.1a.

Although simply producing additional simulation samples would alleviate the issue of statistical fluctuations fitted as spurious signal, producing more simulated events is computationally expensive. Additionally, producing events which fall into specific phase spaces (such as diphoton events with a large amount of missing transverse energy) is often highly inefficient. Therefore, an alternative solution using the available simulation samples is preferred. Given that the spurious signal systematic uncertainty is one of the largest sources of experimental error in the analysis, a technique to reduce these fluctuations may significantly improve the precision of the analysis.

8.1 Gaussian Processes

A Gaussian Process (GP) is defined as a set of random processes, where all finite subsets of these processes have a multivariate normal distribution [105]. A multivariate distribution can be de-



(a) Fit to Raw Template (b) Fit to Smoothed Template

Figure 8.1: Example signal and background fits to a toy background template generated from an exponential probability density function. The black points show the toy events in each bin. The left plot shows the fit in green to the raw template. (The dark green shows the background only fit, while the light green shows the complete signal and background fit). The right plot shows the fit in red to the smoothed background template, shown by the blue points. (The dark red shows the background only fit, while the light red shows the complete signal and background fit). The background function used in the fits is an exponential function, the same as used to generate the background templates. Therefore, the analytic function can perfectly model the background template, aside from statistical fluctuations, and no real shape mismodeling is present. The expected value of the spurious signal uncertainty is then zero.

scribed using two quantities: the *mean* and the *covariance*.

An example of a two-dimensional multivariate normal distribution is presented in Figure 8.2, along with the two one-dimensional Gaussian components (one labeled the x-component, in blue, and the other the y-component, in red). In the given example, the mean is $\left[\begin{smallmatrix}0\\0\end{smallmatrix}\right]$. Both the x- and y-component Gaussian distributions are centered at 0, and so the center of the 2D multivariate distribution is simply the origin. The covariance matrix is $\left[\begin{smallmatrix}1 & 3/5\\ 3/5 & 2\end{smallmatrix}\right]$. The width of the 1D Gaussian along the x-axis (1) is given by the first diagonal element of the covariance matrix, while the width of that along the y-axis (2) is given by the second diagonal element. The off-diagonal elements indicate the correlation between the two 1D Gaussian distributions. The positive value $(3/5)$ leads to a diagonal skew in the 2D distribution. Essentially, events which are positive in x are more likely to also be positive in y.

Given a finite dataset – such as the bin contents of a smooth histogram – with corresponding mean and covariance matrices, a Gaussian Process may be defined. The "correct" mean and the covariance, however, are not necessarily well defined, as they encode specific assumptions about

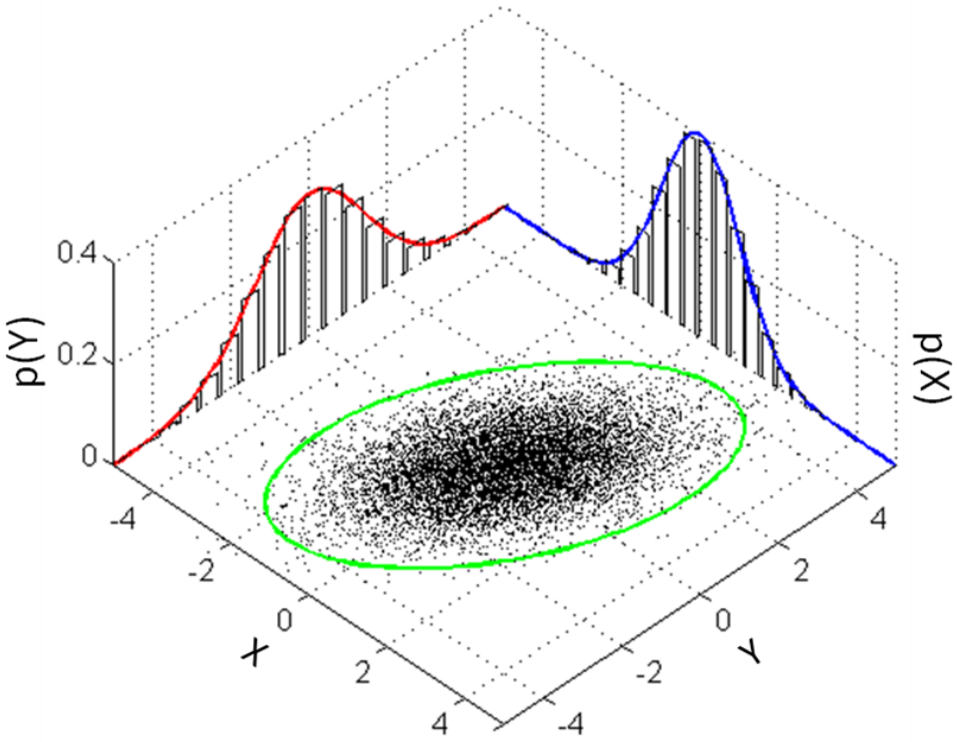


Figure 8.2: An example of a two-dimensional multivariate normal distribution, shown via a density in the 2D plane. The corresponding one-dimensional Gaussian distributions are shown by the blue and red curves. The green ellipse shows the 3 sigma band of the 2D multivariate normal distribution. The mean of the 2D multivariate normal distribution shown is $[0, 0]$ – the distribution is centered at the origin. The width of the 1D Gaussian along the x-axis (1) is given by the first diagonal element of the covariance matrix, while the width of that along the y-axis (2) is given by the second diagonal element. The off-diagonal elements indicate the correlation between the two 1D Gaussian distributions. The positive value $(3/5)$ leads to a diagonal skew in the 2D distribution. Figure taken from Ref. [12].

the underlying dataset. In practice, the two quantities are fit to a finite dataset using a minimization algorithm. In the case of a one-dimensional histogram with a finite number of bins, the mean can be interpreted as a "rough" description of the underlying shape. The diagonal elements of the matrix represent the error of each bin. The off-diagonal elements of the covariance matrix specify how "similar" the bin content of two different bins should be.

The covariance matrix can be simplified through the introduction of a *kernel*, which analytically determines the level of correlation between two distinct points. One commonly used kernel is the Radial Basis Function (RBF) kernel. The RBF kernel is defined as follows:

$$k(x_i, x_j) = e^{-\frac{1}{2}d(x_i/l, x_j/l)^2}$$

where d is the Euclidian distance and l is a hyper-parameter defining the characteristic length scale [104]. In the case of a one-dimensional histogram, the quantities x_i and x_j may be interpreted as the center of each bin. If the distance between the bin centers is much larger than the length scale l, then the content of the two bins will be uncorrelated. If the bin centers are of order l or closer together, then the contents will be highly correlated. The value of l can be interpreted loosely as a measure of the minimum feature size of the distribution. A useful kernel for the case of smoothly falling distributions is the *Gibbs kernel* [106, 107]. The Gibbs kernel is similar to the RBF kernel, except that the length scale l is allowed to increase linearly as a function of x.

As stated above, the uncertainties on data points may be accounted for by a term in the diagonal of the covariance matrix. The diagonal elements may be passed as fixed values, or they may be included as a part of the kernel. One approach to including uncertainties on points within the kernel itself is to add a *white kernel*, which is simply an $N \times N$ identity matrix (where N is the number of data points) multiplied by a constant value. The constant value is a hyper-parameter, which is fit during minimization.

For the case of a histogram with a smoothly falling distribution, the bins at a higher value of x will contain fewer events. Because the error on the content of each bin follows Poisson statistics (for the case of unweighted counts), bins at higher x values will have smaller absolute errors associated with their content. To approximate this effect, a custom *linear error kernel* is used. The linear error kernel is similar to the white noise kernel in that it is a diagonal matrix. However, the diagonal elements decrease linearly as a function of x. The initial value (the magnitude of the uncertainty on the left-most bin) and the slope are both hyper-parameters, which are constrained by the data points.

For this work, the process of fitting a GP to a finite data set is performed using the Scikit-Learn [104] machine learning package.

8.2 Smoothing Background Templates with Gaussian Processes

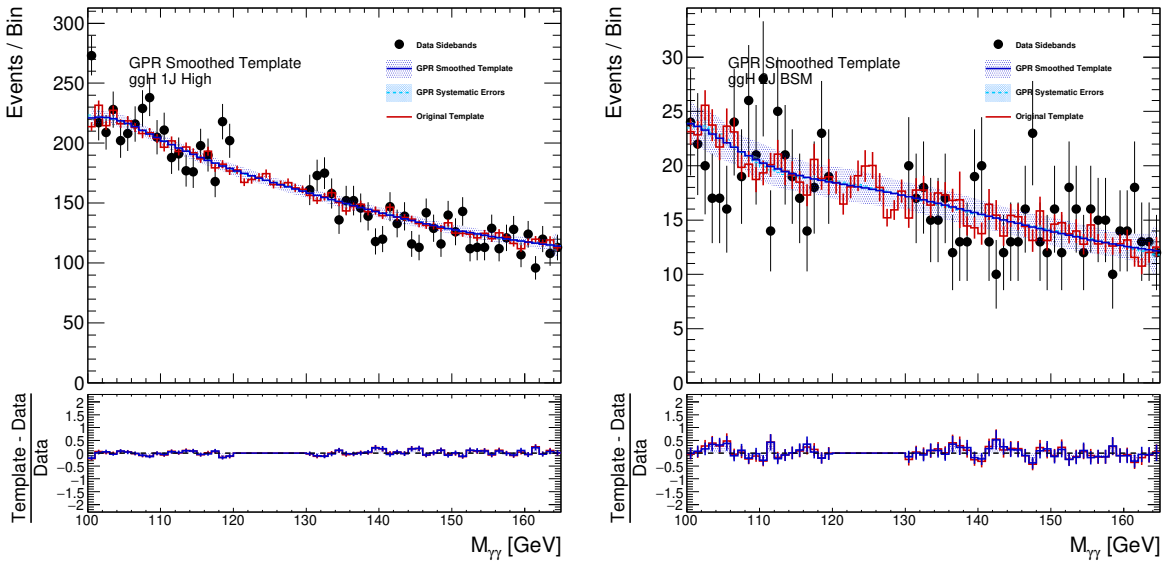
The background templates used in the spurious signal test for the analysis categories are all smooth, roughly exponentially falling distributions with statistical fluctuations. Fitting a background template using Gaussian Process Regression (using the Gibbs kernel with the linear error component discussed above) offers a consistent method of estimating the underlying smooth shape of the template, without the problematic fluctuations. The power of such smoothing is illustrated in Figure 8.1: after smoothing, the measured spurious signal obtained from the toy template is reduced from 28.1 (in Figure 8.1a) to 0.6 (in Figure 8.1b). Notably, the GP smoothing technique makes no assumption on the underlying distribution other than that it is smooth and falling, hence the choice of functional form from the spurious signal test will not be biased.

In order to obtain a reasonable understanding of the underlying shape of the template, the hyper-parameters of the fit must be chosen carefully. A dedicated script was written to optimize the hyper-parameters of the Gibbs kernel. The script first fits the provided template with a range of analytic functions (the same as those used in the spurious signal test, listed in Section 7.3) and takes that where the χ^2 value is closest to 1. Another testing template is produced based on this analytic function, which, unlike the original template, will not contain any statistical fluctuations. Two different narrow features are added to the smooth template shape, then fit individually. The first is a narrow CB shape, where the width is defined as half the bin width of the input template. This narrow feature is designed to replicate a statistical fluctuation. The second is designed to be a signal-like feature; a CB shape with width equal to the expected signal width is used.

The two testing templates (one with the narrow feature and one with a wider feature) are then fit using GPs with a wide range of the two Gibbs kernel hyper-parameters, the length scale and the slope. Ideally, the narrow feature will be mostly smoothed out (since this feature is more similar to the statistical noise the method is designed to mitigate). The Gibbs length scale will therefore be greater than the width of this narrow feature. The wider feature is designed to mimic the size of a real feature in the background template, and so it should not be smoothed out in the fit. Therefore, the length scale should not be too much larger than this width. A set of hyper-parameters is deemed acceptable if the magnitude of the narrow feature is reduced by at least 67% at its mean value, and if the magnitude of the wider feature is reduced by less than 75% at its mean value. Once the scan is complete, a rectangle of good hyper-parameter values is constructed. In cases where the good hyper-parameter space is not rectangular, a rectangular subset is chosen such that the range of the length scale parameter is maximized.

Once the optimal hyper-parameter ranges have been set, a GP is fit to the original (noisy) background template in each category. The GP mean in the fits is defined as an exponential function, the parameters of which are obtained by a fit to the original background template. The exponential shape has been observed to be a sufficiently close guess for the categories used by the analysis. However, in cases where the input template has very few statistics (less than about ten events per bin on average), the resulting GP fit may be nearly identical to the mean exponential shape. This issue occurs when the statistical uncertainties of the original template are so large that the template is fully compatible with the preliminary exponential shape. Although the exponential shape is technically an adequate descriptor of the template shape, the choice of the exponential mean does bias the functional choice of the spurious signal test in this case. Therefore, a check has been added to re-perform the GP fit using a flat mean in cases where the resulting GP shape and the mean exponential shape differ by less than 0.01%.

The resulting smoothed shape obtained from the GP fit is then saved as a new histogram. This smoothed histogram is passed as the background template to the spurious signal test, which then



(a) ggF 1J HIGH (b) ggF 2J BSM

Figure 8.3: Example full Run 2 background templates for two of the analysis categories, (a) ggF 1J HIGH, which contains a medium-level of statistics, and (b) ggF 2J BSM, which contains few statistics. The red shape shows the original background template, the blue shape shows the smoothed background template, and the black points show the data sidebands (for reference). The bottom panel shows the fractional difference between the smoothed- and un-smoothed templates and the data sidebands.

determines the background functional form and spurious signal systematic uncertainty as described in Section 7.3.

The technique was tested using the non-$t\bar{t}H$ analysis categories (the spurious signal systematic uncertainty does not affect the $t\bar{t}H$ sensitivity as much as that of the other production modes). Examples of the smoothed templates are presented in Figure 8.3 for a category with a medium level of statistics (Figure 8.3a) and for a category with a very low level of statistics (Figure 8.3b). The original templates are shown as well, for comparison. The data sidebands are also shown for validation, although the GP smoothing technique does not take into account the data sidebands. The remaining templates are presented in Appendix C. Note that the presented templates include the full Run 2 data set (2015-2018).

A comparison of the measured spurious signal systematic uncertainty in the analysis categories (other than the $t\bar{t}H$-enriched categories) is presented in Table 8.1. After performing the smoothing, a significant reduction in the magnitude of the spurious signal uncertainty was observed for most of the analysis categories. The largest reductions are seen in the categories with lower statistics (as expected). The few categories which see an increase in the magnitude of the systematic (ggF 0J FWD, ggF 2J LOW, and ggF 2J MED) are all well-populated categories, and so no significant

Index	Category	Spurious Signal	Functional Form	Spurious Signal (GP-Smoothed)	Functional Form (GP-Smoothed)
1	ggF 0J CEN	ExpPoly2	-97.1	ExpPoly2	-88.3
2	ggF 0J FWD	Bern3	-592	ExpPoly2	-653
3	ggF 1J LOW	ExpPoly2	-112	ExpPoly2	-70.7
4	ggF 1J MED	ExpPoly2	-72.9	ExpPoly2	-63.4
5	ggF 1J HIGH	Power Law	-13.3	Power Law	3.55
6	ggF 1J BSM	Exponential	-3.23	Exponential	-0.724
7	ggF 2J LOW	ExpPoly2	-81.6	ExpPoly2	-85.3
8	ggF 2J MED	ExpPoly2	-30.1	ExpPoly2	-33.2
9	ggF 2J HIGH	Exponential	-22.4	Exponential	-13.4
10	ggF 2J BSM	Exponential	-5.84	Exponential	-1.82
11	VBF Loose, Low p_{T}^{Hjj}	Exponential	-8.19	Exponential	-5.01
12	VBF Tight, Low p_{T}^{Hjj}	Exponential	4.37	Exponential	2.9
13	VBF Loose, High p_{T}^{Hjj}	Exponential	-19.5	Exponential	-8.25
14	VBF Tight, High p_{T}^{Hjj}	Exponential	-4.67	Exponential	0.276
15	VH Had Loose	Exponential	-14.9	Exponential	-4.68
16	VH Had Tight	Power Law	-7.7	Exponential	0.406
18	qqH BSM	Exponential	-17.1	Exponential	-13.2
19	VH MET Low	Exponential	2.7	Exponential	0.731
21	VH MET High	Exponential	1.87	Exponential	0.221
22	VH Lep Low	Exponential	-6.84	Exponential	-2.04
22	VH Lep High	Power Law	-1.13	Power Law	-0.647
22	VH Dilep	Exponential	1.27	Power Law	-0.302

Table 8.1: Summary of the magnitude of the spurious signal systematic and the associated functional form chosen by the spurious signal test, both before and after the GP smoothing technique is applied.

impact from statistical fluctuations in the value of the systematic is expected.

Extensive validation tests were performed with the GP smoothing technique in order to ensure that the smoothing itself does not introduce a bias. These tests primarily use "toy" templates – randomly-generated background templates constructed from either simulated diphoton events or from the probability distribution function of a known analytic function. These studies are presented in detail in Appendix D. No significant bias was found in the quality of the GP fits to the toy templates. Alternative choices for the GP mean (a decreasing linear function and a flat line) were tested; no significant difference was found in the fit quality as a result of the mean. Additionally, a spurious signal test was performed on many toys (both from analytic functions and from simulated events) to determine whether the GP smoothing affected the spurious signal measurement. No significant bias in the measured spurious signal was found, both in cases where the background function can successfully describe the background and where there exists real mismodeling by the analytic background function.

The technique has not yet been utilized for a public result of the analysis presented here. However, the technique has been used in the analysis of the Higgs boson differential cross sections in the diphoton decay channel, presented in Ref. [108]. In the differential cross sections anay-

sis, the use of the GP smoothing technique reduced the spurious signal uncertainty by an average of approximately 20% across the analysis categories. As expected, the greatest gains came from categories with low statistics. The total impact of the spurious signal systematic uncertainty was calculated to be 6.4% in the full Run 2 iteration of the analysis. This represents an absolute gain of almost one percent over the previous iteration of the analysis (with a total spurious signal systematic impact of 7.3%), which did not use the smoothing technique. Notable, the true impact of the smoothing technique is greater than the absolute reduction of the systematic shown here, since the spurious signal systematic would have increased with the addition of the 2018 dataset.

The GP smoothing approach represents a novel and simple method to mitigating the impact of statistical fluctuations in calculating the spurious signal uncertainty. Although the results presented in this work do not take advantage of this technique, it is a promising approach to be utilized for the full Run 2 measurement of Higgs boson couplings in the diphoton decay channel.

CHAPTER 9

Statistical Method

The statistical model used by the analysis is that of a fit to the diphoton invariant mass spectrum in data, using analytic signal and background shapes. The determination of the analytic signal shape is presented in Section 7.1. The determination of the background shape, along with the systematic uncertainty associated with the choice of analytic function, is described in Section 7.3.

The analysis aims to provide a measurement of the inclusive cross section σ times the Higgs to diphoton branching ratio ($BR_{\gamma\gamma}$) for the four targeted production modes, ggF, VBF, VH, and $t\bar{t}H$. Additionally, a measurement is made of the cross section times branching ratio for the STXS bins which are statistically accessible. Due to the low statistics in the $t\bar{t}H$ categories of the analysis, no attempt is made to separate the $t\bar{t}H$ and tH events into further STXS bins.

9.1 Signal Model

The analysis categories are broadly designed to target a specific production mode, ggF, VBF, VH, or $t\bar{t}H$. However, no categorization can be perfect – each category includes Higgs boson events from all Higgs boson production modes (the predicted fractions of events from each production mode within each analysis category can be found in Figure 6.3). This includes rarer production modes not targeted by the analysis, such as $b\bar{b}H$ production. As discussed in Section 7.1, the analytic signal shape in each category includes the contribution from *all* SM Higgs boson production modes, not just the targeted mode.

The finer level of categorization within the production modes (such as the ten orthogonal categories targeting ggF production) is designed to target the STXS framework, as described in Section 6.3. However, events produced from one STXS bin, defined as the "truth bin," may migrate into another bin during reconstruction, defined as the "reconstruction bin." For example, a ggF event produced with two associated jets will ideally fall within one of the ggF 2J bins. If one of the jets is mis-reconstructed as another object, however, it may migrate into one of the ggF 1J

bins. The estimated fractions of Higgs boson events from each STXS truth bin populating each reconstruction bin are presented in Figure 6.2.

When measuring the STXS $\sigma \times BR_{\gamma\gamma}$, the expected signal yield in a given reconstructed analysis category is given by the following equation:

$$N_r = \Sigma_t \sigma_t^{\gamma\gamma} \epsilon_{tr} \mathcal{L} \tag{9.1}$$

where: r is the index of the reconstructed analysis category; t is the index of the STXS truth bin; $\sigma_t^{\gamma\gamma}$ is the cross-section times branching ration $\sigma_t \times BR_{\gamma\gamma}$ for a given STXS truth bin; ϵ_{tr} is the probability (efficiency) of an event from STXS truth bin t being reconstructed into the analysis category r; and $\mathcal{L}$ is the integrated luminosity of the analyzed dataset. The efficiency ϵ_{tr} is obtained using the simulated Higgs boson samples.

When measuring the $\sigma \times BR_{\gamma\gamma}$ for a given production mode instead of an individual STXS bin, the expected number of events for a given production mode N_p is defined similarly, but with the index t being replaces by an index p denoting each mode.

Additionally, the measurement of the production mode cross sections may be provided in terms of a signal strength μ, rather than as a $\sigma \times BR_{\gamma\gamma}$. The signal strength of a given production mode p is defined as:

$$\mu_p = \frac{\sigma_p^{\gamma\gamma}}{\sigma_{p,SM}^{\gamma\gamma}} \tag{9.2}$$

where $\sigma_{p,SM}^{\gamma\gamma}$ is the SM prediction of the cross section times diphoton branching ratio. This description has the advantage of cancelling out a number of systematics.

Note that for the $t\bar{t}H$ production mode measurement, the $\sigma \times BR_{\gamma\gamma}$ values for the other production modes were fixed the their SM predictions.

9.2 Statistical Framework

The parameters of interest (μ and $\sigma \times BR_{\gamma\gamma}$, for the STXS truth bins and production modes) are extracted through a simultaneous, unbinned fit performed to all the relevant analysis categories. The non-$t\bar{t}H$ Parameters of Interest (PoIs) are extracted from a fit to the 2015-2017 subset of data, and the $t\bar{t}H$ categories are included in the fit. The $t\bar{t}H$ PoIs are extracted from a fit performed only on the $t\bar{t}H$ analysis categories; the other production modes are kept fixed to the SM prediction. The $t\bar{t}H$ fit also includes the 2018 data.

The fitting procedure is discussed in detail in Ref.s [109, 110]. The procedure is based upon the notion of the *likelihood* [38], simply the probability of the observed data $\vec{x}$, given some Nuisance

Parameter (NP) set $\vec{\theta}$:

$$\mathcal{L} = P(\vec{x}|\vec{\theta})$$

In the case of *independent and identically distributed* values [38] of $\vec{x}$ obeying Poisson statistics (as is a particle physics experiment), the likelihood may be expressed as a factorization:

$$\mathcal{L} = \frac{m^n}{n!} e^{-\mu} \Pi_{i=1}^{n} f(\vec{x_i}; \vec{\theta})$$

where m represents the mean of the Poisson distribution. The above function is referred to as the *extended likelihood*.

In the analysis presented here, the extended likelihood is constructed in each analysis category considered using the diphoton invariant mass, $m_{\gamma\gamma}$, in the range of $105 < m_{\gamma\gamma|} < 160$ GeV. The likelihood includes the contribution of the systematic uncertainties discussed in Chapter 10. Each systematic uncertainty corresponds to one or more NPs, which are modeled as either Gaussian or Log-Normal NPs.

This likelihood can be maximized for a fixed value of the PoI, denoted as μ, and floating NP vector $\vec{\theta}$. This gives the conditional maximum likelihood, denoted as $L(\mu, \hat{\hat{\vec{\theta}}})$. Here, the $\hat{\hat{\vec{\theta}}}$ vector contains the values of the NP which maximize the likelihood for some given value of μ. The likelihood may also be maximized when floating both the PoI μ and $\vec{\theta}$. This gives the true maximum likelihood, denoted as $L(\hat{\mu}, \hat{\hat{\vec{\theta}}})$. Here, the statistic $\hat{\mu}$ and the NPs in the $\hat{\hat{\vec{\theta}}}$ vector are all at the values which maximize the likelihood. The maximum likelihood ratio is then:

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

This ratio can be calculated for a PoI of $\mu = 0$ or $\sigma^{\gamma\gamma} = 0$ (the background only hypothesis) and for $\mu = 1$ or $\sigma^{\gamma\gamma} = \sigma^{\gamma\gamma}_{SM}$ (the nominal SM signal hypothesis). A value of $\lambda(\mu/\sigma^{\gamma\gamma})$ closer to one for a given PoI ($\mu/\sigma^{\gamma\gamma}$) indicates greater agreement of the corresponding hypothesis with the observed data. The main benefit of the maximum likelihood statistical approach is that the maximum likelihood ratio is independent of the NPs [111].

An important related quantity is the Negative Log Likelihood (NLL); it is defined as $-2\ln\lambda$ [111]. The NLL takes a parabola-like shape, and, due to the negative sign, it can be minimized numerically with relative ease to find the maximum of λ and hence extract the best-fit PoI value.

In practice, determining the compatibility between a given value of a PoI and a hypothesis is difficult. For the analysis presented here, the Asymptotics approximation [112] is used to determine the level of agreement. The Asymptotics approximation makes use of a representative "Asimov" dataset. Two different constructions of the Asimov may be used. The first is the "pre-

fit" Asimov. This construction involves a fit to the data side-bands in each analysis category using the analytic function defined in Section 7.3. The fit to the side-bands is used to interpolate the background shape to the signal region, $120 < m_{\gamma\gamma} < 130$ GeV. The fitted DSCB signal shape, defined in Section 7.1, in each category is added to the background function, and its magnitude is set to the SM prediction. The second contruction is the "post-fit" Asimov, which instead involves a fit to the observed data in the full mass region in each category, again using the analytic signal and background shapes chosen in Sections 7.1 and 7.3, respectively. The PoI is profiled, and all NPs are allowed to float. This fit will set the post-fit nuisance parameter values, as well as the normalization of the background. As for the pre-fit construction, the fitted DSCB signal shape, normalized to the SM prediction, is added to the fitted background shape. Some of the results presented in this work may be compared to so-called "expected" values – these are calculated using the post-fit Asimov dataset.

Because of the fact that the $t\bar{t}H$ portion of the analysis uses the full Run 2 data set, while the remainder of the analysis uses only the 2015-2017 portion of the Run 2 data set, separate extended likelihoods are used for the $t\bar{t}H$ and non-$t\bar{t}H$ fits.

CHAPTER 10

Systematic Uncertainties

The sources of uncertainty in the analysis come from both experimental and theoretical sources. The experimental sources include those relating to the detection, reconstruction, and identification of the physics objects used in the analysis, as well as the measurement of their kinematic properties. Additionally experimental uncertainties are those associated with the analysis techniques, such as that due to the choice of analytic function used to model the continuum diphoton background. The theoretical sources include those relating to the simulation of different physics processes, as well as uncertainties on the calculations of the cross sections to which the signal samples are normalized.

The theoretical source of uncertainty relevant to the analysis are discussed in Section 10.1, while the experimental sources are discussed in Section 10.2. The impact of each the various uncertainties is evaluated in the final likelihood fit (discussed in Chapter 9), and so their effects on the measurement are presented along with the fit results in Chapter 11.

10.1 Theoretical Sources of Systematic Uncertainty

Theoretical sources of uncertainty considered in the analysis include uncertainties on: Higgs boson branching ratio into the diphoton decay channel ($BR_{\gamma\gamma}$); the Quantum Chromodynamics (QCD) coupling constant α_s; the PDF; the parton showering, underlying event, and hadronization model (*UEPS*); and lack of knowledge of higher-order QCD corrections from perturbative calculations (*QCD*). An additional uncertainty on the rates of gluons splitting to heavy-flavor jets and heavy-flavor jets radiated in Higgs boson events from ggF, VH, and VBF production (*Heavy Flavor*) is included in the $t\bar{t}H$ portion of the analysis, as well.

The *QCD* uncertainties in the VBF, VH, and $t\bar{t}H$ signals are estimated by independently varying both the renormalization and factorization scales (μ_R and μ_F) by factors of $\frac{1}{2}$ and 2. The total *QCD* uncertainty in the ggF signal is calculated in a more complex manor, as the uncertainty may become larger than the value calculated from the μ_R/μ_F variations in certain kinematic regions. The ggF *QCD* uncertainty is broken into nine individual nuisance parameters:

- Four parameters covering the uncertainty in modeling the jet multiplicities of ggF events [13, 113, 114, 115]
- Three parameters covering the uncertainty in modeling the Higgs boson p_{T} spectrum
- Two parameters to account for the uncertainty in the fraction of ggF events which are reconstructed in the VBF analysis categories [116, 117].

The uncertainty of the Higgs boson branching ratio is calculated using HDECAY [77, 118, 78, 119] and PROPHECY4F [79, 80, 81] (as described in Refs. [77, 78, 79, 80, 81, 13]), and the value of the uncertainty is found to be $+2.90\%/-2.84\%$. The total $BR_{\gamma\gamma}$ uncertainty includes a "theoretical" portion, which is primarily due to missing higher order corrections. It also accounts for uncertainties on the top, bottom, and charm quark masses, as well as on the value of α_S, since these values serve as input parameters in calculating the branching ratio [13].

The effects of uncertainties on α_s and the PDF are estimated based on the recommendations in Ref. [120]. The central value of the α_S parameter is set to $\alpha_S(m_Z) = 0.118 \pm 0.0015$ (where m_Z indicates the reference mass scale from renormalization, in this case that of the Z boson). The percentage values of the cross section uncertainty for the Higgs boson production modes due to the combined PDF and α_S uncertainties are presented in Table 10.2, while those due to the QCD uncertainty are presented in Table 10.1. The QCD, PDF, and α_S uncertainties are combined for the $b\bar{b}H$ process; the error on the cross section is $+20.1\%$, -23.9%.

In order to evaluate the effects of incomplete knowledge of parton showering, hadronization processes, and the underlying event, the signal yields and kinematic distributions from alternative simulated signal samples with the same generator and different parton showering algorithms (HERWIG++ and PYTHIA8) are compared. For the ggF signal, an additional uncertainty is derived by combining the effects resulting from different AZNLO eigenvector tunes.

Lastly, in the $t\bar{t}H$ categories, the yields of Higgs boson events from the ggF, VH, and VBF production modes with extra heavy-flavor jets are assigned an additional conservative 100% uncertainty (*Heavy Flavor*), which is correlated between the seven analysis categories.

Uncertainties on the $t\bar{t}H$ cross section and the Higgs boson to diphoton branching ratio are included in the signal strength ($\mu_{t\bar{t}H}$) measurement, as $\mu_{t\bar{t}H}$ is scaled by the SM cross section prediction. However, the $t\bar{t}H$ cross section and the Higgs boson to diphoton branching ratio uncertainties are not included in the $t\bar{t}H$ cross section times branching ratio ($\sigma_{t\bar{t}H} \times BR_{\gamma\gamma}$) measurement.

Uncertainties on the inclusive Higgs boson event yield from non-$t\bar{t}H$ processes are included in the $t\bar{t}H$ cross section times branching ratio ($\sigma_{t\bar{t}H} \times BR_{\gamma\gamma}$) and the signal strength ($\mu_{t\bar{t}H}$) measurements, as all other Higgs boson production processes are assumed to obey their SM predictions. Both the $\mu_{t\bar{t}H}$ and $\sigma_{t\bar{t}H} \times BR_{\gamma\gamma}$ measurements take into account uncertainties which may cause

event migration between (or out of) BDT categories. These include missing higher-order QCD effects and UEPS uncertainties.

Table 10.1: The theoretical uncertainties on the Higgs boson production mode (other than $b\bar{b}H$) cross sections due to the QCD uncertainty. The 3.9% value associated with the ggF production mode is that obtained from the variation of the renormalization and factorization scales, and hence it is smaller than that obtained by the 9 nuisance parameter method. This value is derived at the N3LO in QCD and NLO in Electroweak, as described in Ref. [15].

Production Mode	QCD Uncertainty	
	$+$ [%]	$-$ [%]
ggF	3.9	3.9
VBF	0.4	0.3
$W^{\pm}H$	0.5	0.7
ZH	3.8	3.0
$t\bar{t}H$	5.8	9.2
$tHjb$	6.5	14.7
tWH	4.9	6.7

Table 10.2: The theoretical uncertainties on the Higgs boson production mode cross sections (other than $b\bar{b}H$) due to the combined PDF and α_S uncertainties.

Production Mode	PDF and α_S Uncertainty
	$\pm$[%]
ggF	3.2
VBF	2.1
$W^{\pm}H$	1.9
ZH	1.6
$t\bar{t}H$	3.6
$tHjb$	3.7
tWH	6.3

10.2 Experimental Sources of Systematic Uncertainty

The sources of experimental uncertainty in the analysis can be roughly separated into two groups: those which primarily impact the signal acceptance or yield and those which primarily impact the signal mass shape.

The systematic uncertainties affecting the signal mass shape are Photon Energy Resolution (PER) and Photon Energy Scale (PES) [5] . These two systematics are implemented in the fit as changes to the mean and width (resolution) of the DSCB signal shape. The magnitude of the PES

uncertainty is calculated as a ratio of the mean m of the DSCB shape when varying the PES by one sigma, as follows:

$$\delta m_{DSCB}^{\pm 1\sigma} = \frac{\left\langle m_{\gamma\gamma}^{\pm 1\sigma} \right\rangle}{\left\langle m_{\gamma\gamma}^{\text{Nominal}} \right\rangle} - 1$$

The $\pm 1\sigma$ variations above are taken from Ref.s [5] (for the $t\bar{t}H$ portion of the analysis) and [121] (for the remainder of the analysis). It was observed that $\delta m_{DSCB}^{+1\sigma} \approx \delta m_{DSCB}^{-1\sigma}$, and so the corresponding nuisance parameter in the fit is implemented with a Gaussian constraint, with width of $\delta m_{DSCB}^{+1\sigma}$. The PER uncertainty is treated similarly, but as a ratio of the width. The width is quantified as the inter-quantile Q, defined as: $Q = CDF^{-1}(0.75) - CDF^{-1}(0.25)$. The CDF function is the *cumulative distribution function*, which determines the upper value of $m_{\gamma\gamma}$ for which some given percent of the DSCB shape is contained. The variation of the nuisance parameter is then calculated as:

$$\delta \sigma_{DSCB}^{\pm 1\sigma} = \frac{\left\langle \sigma^{\pm 1\sigma} \right\rangle}{\left\langle \sigma^{\text{Nominal}} \right\rangle} - 1$$

The $\pm 1\sigma$ variations in the resolution are taken from the same sources as provided for the PES uncertainty. Unlike the PES uncertainty, the variations of the PER uncertainty are not symmetric; the product of two asymmetric constraints is used to model the nuisance parameter in the fit. The experimental error on the Higgs boson mass, assigned as 0.24 GeV based on the Run 1 Higgs boson mass measurement [122], also provides an uncertainty on the location of the signal peak.

Dominant systematic uncertainties affecting mainly the signal yield include: the integrated luminosity; the efficiency of the diphoton triggers; and pileup modeling. The uncertainty in the combined 2015–2017 integrated luminosity, used by the non-$t\bar{t}H$ portion of the analysis, is 2.0%, while in the combined 2015–2018 integrated luminosity, used by the $t\bar{t}H$ portion of the analysis, it is 1.7%. It is derived following a methodology similar to that detailed in Ref. [123], using the LUCID-2 detector for the baseline luminosity measurements [124], from calibration of the luminosity scale using x-y beam-separation scans. The uncertainty associated with the diphoton trigger contributes to a 0.4% uncertainty on the yield, as described in Ref. [10].

Other systematic uncertainties that have an impact on the signal acceptance are described as follows. Photon-related systematic uncertainties include photon identification efficiency and isolation efficiency uncertainties [5, 58]. The photon isolation systematic is described in more detail in Section 4.5.1. Jet related systematic uncertainties include the efficiency of the jet-vertex tagger [125] and the jet energy scale and resolution [126]. In addition, systematic uncertainties arising from the reconstruction and identification efficiency of the b-tagging algorithm [127, 92] (*Flavor tagging*) used are included, as well. Both electrons [5, 58] and muons [95] contribute with uncertainties associated to the reconstruction, identification efficiency, isolation efficiency, and their respective energy/momentum scales and resolutions. The uncertainty on $E_{\text{T}}^{\text{miss}}$ [128] from charged tracks

resulting from objects other than the aforementioned high-p_{T} objects is taken into account, as well.

Lastly, the uncertainty on the signal yield related to potential background mis-modeling due to the choice of analytic function is accounted for with the spurious signal, calculated for each analysis category, as described in Section 7.3. Since no systematic trends were observed between categories in spurious signal studies, the spurious signal uncertainty is left uncorrelated between the analysis categories.

CHAPTER 11

Results

The results for the $t\bar{t}H$ portion of the analysis are presented using the full Run 2 data set of pp collisions, corresponding to 139 fb^{-1}, collected at $\sqrt{s} = 13$ TeV with the ATLAS detector. The results of the remaining portion of the analysis are presented using the 2015-2017 subset of the $\sqrt{s} = 13$ TeV Run 2 pp data set, amounting to 79.8 fb^{-1} of data, collected by the ATLAS detector.

11.1 $t\bar{t}H$ Results

The combined signal and background fit to all of the $t\bar{t}H$ analysis categories is shown in Figure 11.1; for illustration purposes, events are weighted by $\ln(1 + S_{90}/B_{90})$, where S_{90} (B_{90}) for each BDT bin is the expected signal (background) in the smallest $m_{\gamma\gamma}$ window containing 90% of the expected signal. The results of the combined fit within each of the individual $t\bar{t}H$ analysis categories are shown in Figures 11.2 ($t\bar{t}H$ hadronic categories) and 11.3 ($t\bar{t}H$ leptonic categories). The fitted yields of the $t\bar{t}H$ signal, non-$t\bar{t}H$ Higgs boson, and continuum background events are presented in Table 11.1.

The top associated Higgs boson production process in the diphoton decay channel is observed with a significance of 4.9σ; the expected significance is 4.2σ. The expected significance is evaluated using the post-fit construction of the Asimov dataset, as described in Chapter 9. The signal is normalized to the SM prediction, while the values of the nuisance parameters and the normalization of the background are taken from the nominal fit to the data. Within the post-fit Asimov, the values of the energy scale and resolution nuisance parameters were found to be somewhat pulled. When fixing these two parameters to their nominal values, the expected significance increases slightly to 4.5σ, closer to the observed significance.

The observed signal strength of the $t\bar{t}H$ process is

$$\mu_{t\bar{t}H} = 1.38\ ^{+0.41}_{-0.36} = 1.38\ ^{+0.33}_{-0.31}\ (\text{stat.})\ ^{+0.13}_{-0.11}\ (\text{exp.})\ ^{+0.22}_{-0.14}\ (\text{theo.}).$$

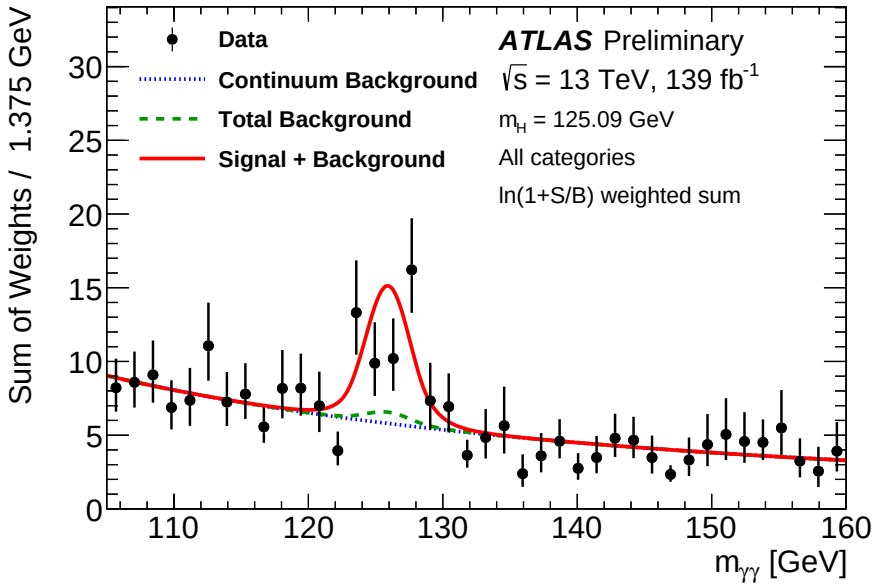


Figure 11.1: Weighted diphoton invariant mass spectrum for the sum of all seven $t\bar{t}H$ BDT categories observed in 139 fb^{-1} of 13 TeV data. Events are weighted by $\ln(1 + S_{90}/B_{90})$, where S_{90} (B_{90}) for each BDT category is the expected signal (background) in the smallest $m_{\gamma\gamma}$ window containing 90% of the expected signal. The error bars represent 68% confidence intervals of the weighted sums. The solid red curve shows the fitted signal-plus-background model with the Higgs boson mass constrained to 125.09 ± 0.24 GeV. The non-resonant and total background components of the fit are shown with the dotted blue curve and dashed green curve. Both the signal-plus-background and background-only curves shown here are obtained from the weighted sum of the individual curves in each BDT category.

The statistical, experimental, and theoretical uncertainties are determined from the difference (in quadrature) of the uncertainty obtained from the unconditional fit and that obtained when the relevant nuisance parameters are fixed to their best fit values.

The observed cross section times branching ratio $\sigma_{t\bar{t}H} \times BR_{\gamma\gamma}$ is measured using the same formalism as above, except that the theoretical uncertainties from the total $t\bar{t}H \to \gamma\gamma$ production cross section and the $H \to \gamma\gamma$ branching ratio are omitted. The effects of these uncertainties on the kinematic distributions are included, as these may lead to the migration of events between the $t\bar{t}H$ BDT categories or a change in the efficiency of the $t\bar{t}H$ selection.

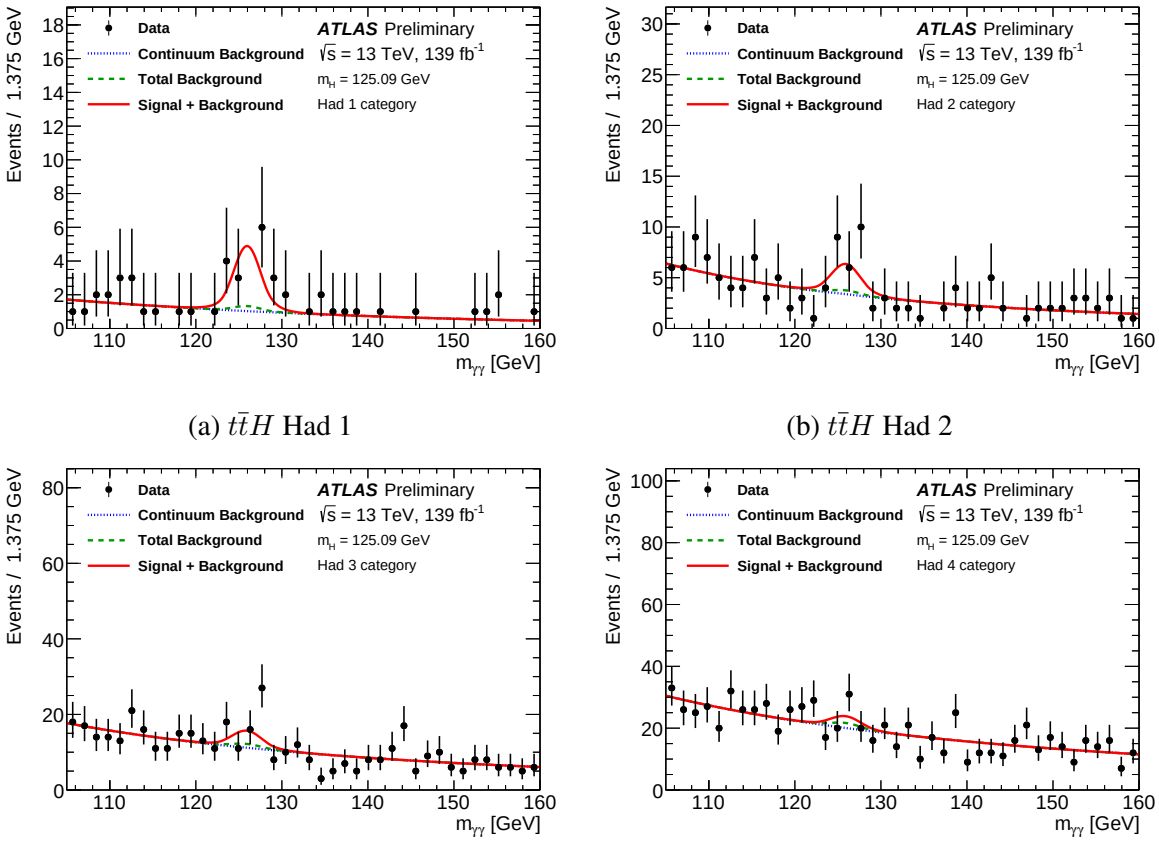
The measured cross section times branching ratio is:

$$\sigma_{t\bar{t}H} \times BR_{\gamma\gamma} = 1.59^{+0.43}_{-0.39}\ \text{fb} = 1.59^{+0.38}_{-0.36}\ (\text{stat.})\ ^{+0.15}_{-0.12}\ (\text{exp.})\ ^{+0.15}_{-0.11}\ (\text{theo.})\ \text{fb}.$$

The above $t\bar{t}H$ results were calculated with the Higgs boson mass fixed to 125.09 ± 0.24 GeV, the value measured by the ATLAS and CMS Collaborations [122]. As a cross check, the signif-

icance, μ, and $\sigma_{t\bar{t}H} \times B_{\gamma\gamma}$ were remeasured with the Higgs boson mass and the signal resolution allowed to be determined directly from data; the results were found to be compatible with the nominal fit.

For reference, the SM prediction of the $\sigma_{t\bar{t}H} \times BR_{\gamma\gamma}$ for the $t\bar{t}H$ process in the diphoton channel is $1.15^{+0.09}_{-0.12}$ fb. The measured $\sigma_{t\bar{t}H} \times BR_{\gamma\gamma}$ value is higher than the SM prediction, but the tension is not great enough to strongly suggest the existence of some BSM physics. Therefore, the results presented here do not indicate any significant deviation from the SM at this time.



(a) $t\bar{t}H$ Had 1

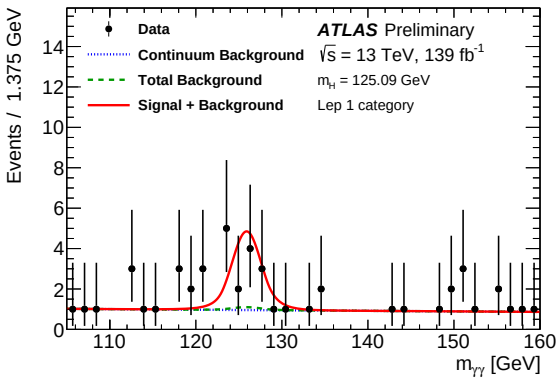
(b) $t\bar{t}H$ Had 2

(c) $t\bar{t}H$ Had 3

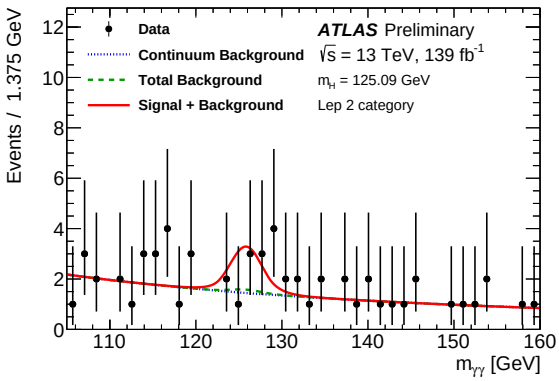
(d) $t\bar{t}H$ Had 4

Figure 11.2: The fitted signal and background shapes for each of the four "Had" region BDT categories. The category with the greatest purity is shown in (a), the second-greatest purity in (b), the third-greatest purity in (c), and the lowest purity in (d). The red line shows the result of the signal plus background unbinned fit to the data, while the data points are shown as black dots. The dotted blue line shows the continuum background component of the fit, and the dashed green line shows the total background (including non-$t\bar{t}H$ Higgs events). The fit results are derived from the combined signal plus background fit to all seven BDT categories.

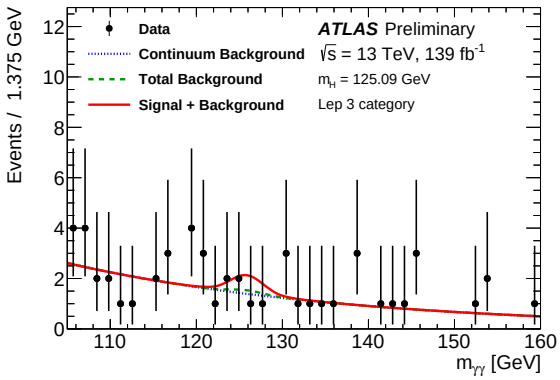
The uncertainties associated with the measurement are significantly larger than those associated with the SM prediction. The breakdown of the contributions to the measurement from each group



(a) $t\bar{t}H$ Lep 1



(b) $t\bar{t}H$ Lep 2



(c) $t\bar{t}H$ Lep 3

Figure 11.3: The fitted signal and background shapes for each of the three "Lep" region BDT categories, with (a) showing the category with the greatest signal purity and (c) showing that with the lowest. The category with intermediate purity is shown in (b). The red line shows the result of the signal plus background unbinned fit to the data, while the data points are shown as black dots. The dotted blue line shows the continuum background component of the fit, and the dashed green line shows the total background (including non-$t\bar{t}H$ Higgs events). The fit results are derived from the combined signal plus background fit to all seven BDT categories.

of uncertainties (as described in Chapter 10) is shown in Table 11.2. The dominant uncertainty in the measurement is currently the statistical uncertainty. The upcoming Run 3 of the LHC is expected to approximately double the existing pp dataset; this additional data will help significantly reduce the statistical errors.

The impacts from the experimental and theoretical uncertainties are comparable. On the theoretical side, the uncertainties due to the underlying event and parton showering are dominant. Reducing these uncertainties will require dedicated improvements to current simulation and parton showering algorithms. The modeling of heavy flavor jets in non-$t\bar{t}H$ Higgs boson (*Heavy Flavor*) processes also contributes significantly to the total uncertainty. This *Heavy Flavor* uncertainty is currently conservatively estimated by the analysis. Dedicated efforts to constrain this uncertainty through differential measurements of heavy flavor jets in Higgs boson events may better constrain this uncertainty in the future. However, such measurements are challenging, and comparisons to simulation are complicated by the large discrepancies between the predictions of different algorithms.

Table 11.1: Observed number of events in the different categories for the cross section times branching ratio measurement, using 13 TeV data corresponding to an integrated luminosity of 139 fb^{-1} ("Data"). The observed yields are compared with the sum of expected $t\bar{t}H$ signal, background from non-$t\bar{t}H$ Higgs boson production, and other background sources. The numbers are counted in the smallest $m_{\gamma\gamma}$ window containing 90% of the expected signal. The background yield is extracted from the fit with freely floating signal. The BDT bins are labeled such that the category with the highest signal purity in each of the "Had" and "Lep" regions is labeled as category 1, while that with the lowest signal purity is labeled with the largest number.

Category	$t\bar{t}H$ Signal	non-$t\bar{t}H$ Higgs	Continuum Background	Total (Expected)	Data
$t\bar{t}H$ "Lep" Category 1	7.9 ± 1.5	0.42 ± 0.12	4.6 ± 0.9	12.9 ± 1.8	15
$t\bar{t}H$ "Lep" Category 2	3.9 ± 0.6	0.43 ± 0.15	7.5 ± 1.2	11.8 ± 1.3	11
$t\bar{t}H$ "Lep" Category 3	1.45 ± 0.24	0.49 ± 0.19	7.5 ± 1.2	9.5 ± 1.2	6
$t\bar{t}H$ "Had" Category 1	6.9 ± 1.6	0.8 ± 0.5	4.5 ± 0.9	12.2 ± 1.9	15
$t\bar{t}H$ "Had" Category 2	5.6 ± 1.0	1.1 ± 0.8	16.5 ± 1.7	23.2 ± 2.3	31
$t\bar{t}H$ "Had" Category 3	7.7 ± 1.3	3.1 ± 2.2	56.0 ± 3.0	67 ± 4	82
$t\bar{t}H$ "Had" Category 4	4.9 ± 0.8	5 ± 4	101 ± 4	111 ± 6	105

On the experimental side, the dominant systematic uncertainty is that from the PER. The PES contributes significantly, as well. These two uncertainties are pulled somewhat in the fit (as mentioned above); this is likely due to the large number of events with a diphoton invariant mass slightly above the expected Higgs mass of 125 GeV. Pulling the PER and PES nuisance parameters allows the signal shape of the fit to encompass these events, since the mass of the Higgs boson is fixed to 125.09 GeV. Barring significant changes to the EM calorimeter, reducing the PER and PES uncertainties will require novel improvements to the existing photon calibration process. The jet related systematic uncertainties also contribute significantly to the total uncertainty, which is

reasonable considering the large number of jets included in the $t\bar{t}H$ selection.

Table 11.2: The contribution of groups of systematic uncertainties to the total error on the observed cross section times branching ratio of the $t\bar{t}H$ process. This is shown as the uncertainty due to each group of systematic uncertainties ($\Delta\sigma$), as a fraction of the total observed cross section (σ). For each group of uncertainties, asymmetric errors are assigned. Here $\Delta\sigma_{\text{high}}$ ($\Delta\sigma_{\text{low}}$) shows the effect of systematic variations that increase (decrease) σ.

Uncertainty source	$\Delta\sigma_{\text{low}}/\sigma$ [%]	$\Delta\sigma_{\text{high}}/\sigma$ [%]
Theory uncertainties	6.6	9.7
Underlying Event and Parton Shower (UEPS)	5.0	7.2
Modeling of Heavy Flavor Jets in non-$t\bar{t}H$ Processes	4.0	3.4
Higher-Order QCD Terms (QCD)	3.3	4.7
Parton Distribution Function and α_S Scale (PDF+α_S)	0.3	0.5
Non-$t\bar{t}H$ Cross Section and Branching Ratio to $\gamma\gamma$ (BR)	0.4	0.3
Experimental uncertainties	7.8	9.1
Photon Energy Resolution (PER)	5.5	6.2
Photon Energy Scale (PES)	2.8	2.7
Jet/$E_{\text{T}}^{\text{miss}}$	2.3	2.7
Photon Efficiency	1.9	2.7
Background Modeling	2.1	2.0
Flavor Tagging	0.9	1.1
Leptons	0.4	0.6
Pileup	1.0	1.5
Luminosity and Trigger	1.6	2.3
Higgs Boson Mass	1.6	1.5

11.2 Remaining Results

As stated in Chapter 9, the non-$t\bar{t}H$ portion of the analysis uses a separate extended likelihood and fit than used for the above $t\bar{t}H$ results. However, the extended likelihood used for the following results does include the $t\bar{t}H$ and tH processes through the inclusion of the seven $t\bar{t}H$ categories. Only the 2015-2017 portion of the data in these categories is used for this fit.

The result of the combined signal and background fit to all 29 analysis categories using the 2015-2017 subset of Run 2 data is presented in Figure 11.4. The results of the fit for the combined categories targeting the ggF, VH, and VBF production modes are presented in Figure 11.5. The same $\ln(1 + S_{90}/B_{90})$ weighting prescription detailed in the previous section is used for these figures.

A measurement of the global Higgs boson signal strength is derived, under the assumption that the ratios of the different Higgs boson production modes agree with the SM predictions (within

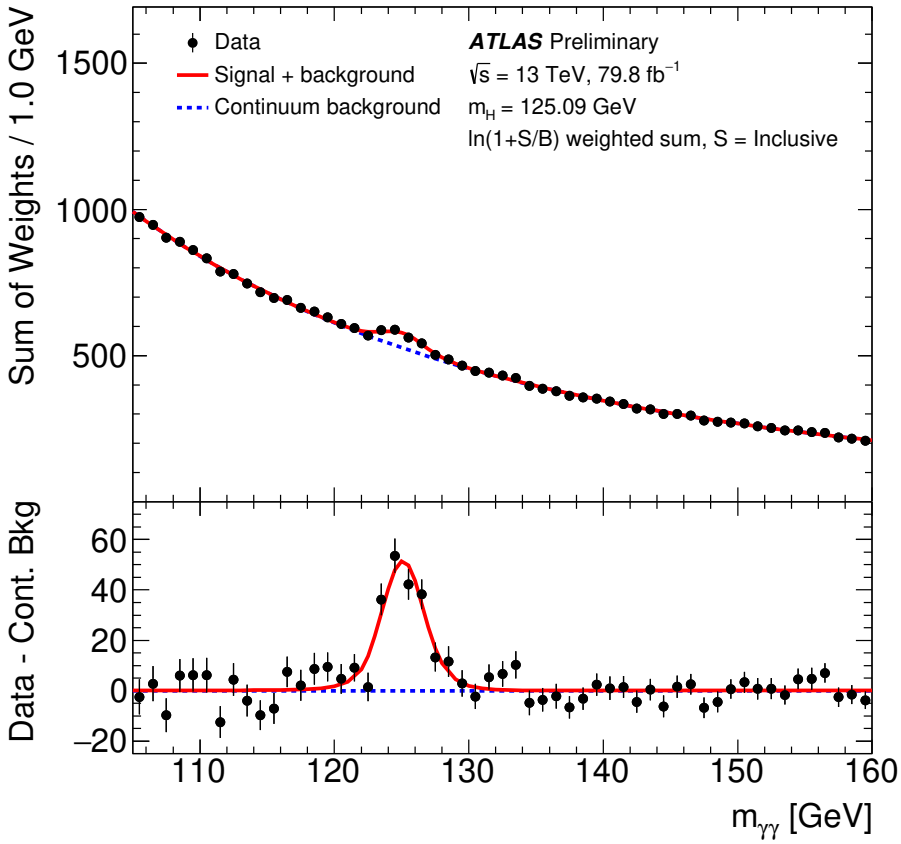
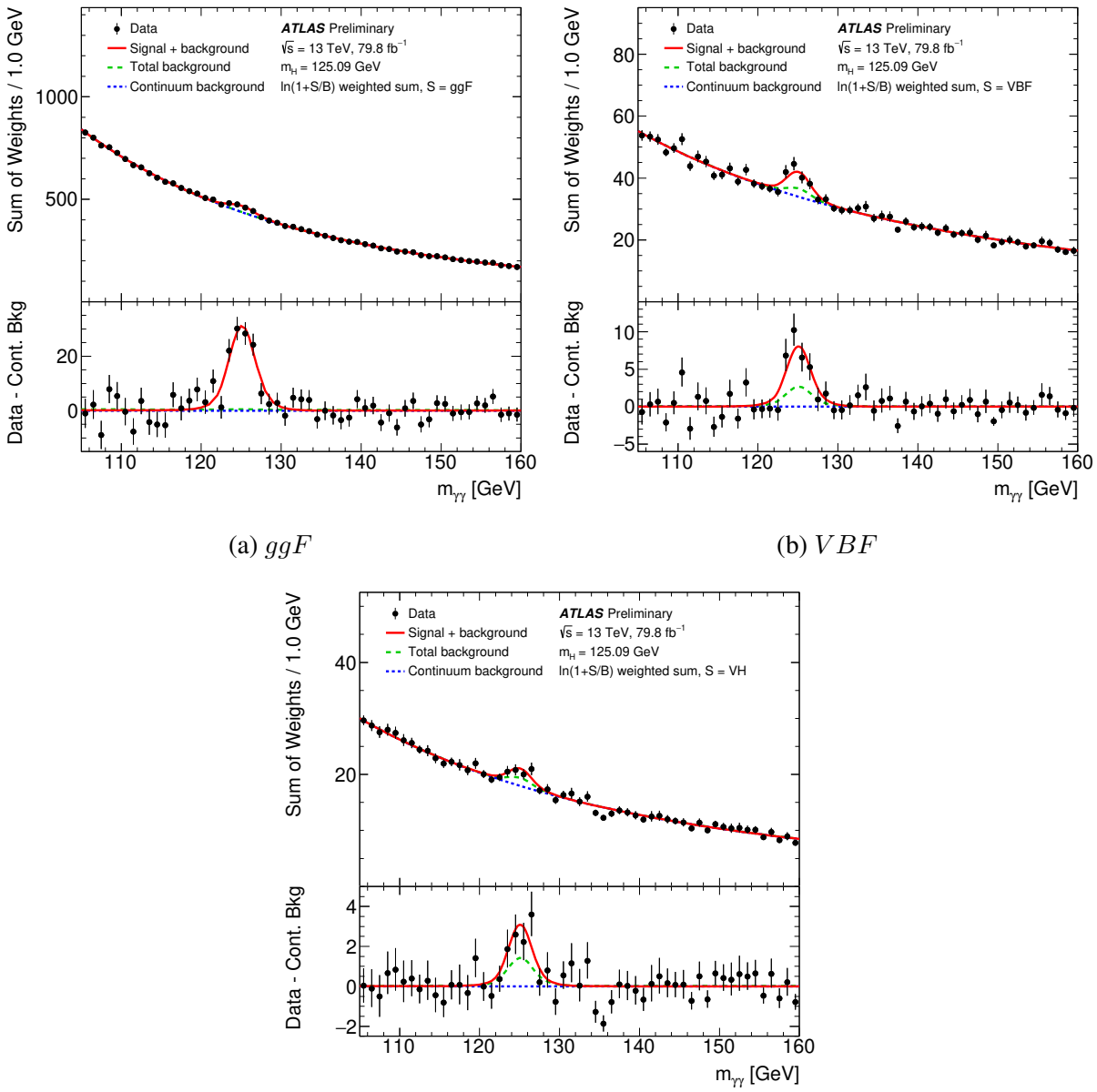


Figure 11.4: The weighted diphoton invariant mass spectrum of all the analysis categories (only the 2015-2017 subset of $t\bar{t}H$ events is included). Events are weighted by $\ln(1 + S_{90}/B_{90})$, where S_{90} (B_{90}) for each category is the expected signal (background) in the smallest $m_{\gamma\gamma}$ window containing 90% of the expected signal. The error bars represent 68% confidence intervals of the weighted sums. The solid red curve shows the fitted signal-plus-background model with the Higgs boson mass constrained to 125.09 ± 0.24 GeV. The dotted blue curve shows the continuum background, while the dashed green curve shows the combined continuum and non-targeted Higgs boson backgrounds. The lower panel shows the difference between the observed data and the fitted continuum background.

theoretical uncertainties). This measured value is:

$$\mu = 1.06\ ^{+0.14}_{-0.12} = 1.06\ \pm 0.08\,(\text{stat.})\ ^{+0.08}_{-0.07}\,(\text{exp.})\ ^{+0.07}_{-0.06}\,(\text{theo.})$$

This result is compatible with the SM prediction.



(a) ggF

(b) VBF

(c) VH

Figure 11.5: The weighted diphoton invariant mass spectrum in the analysis categories, grouped by the targeted production mode: (a) ggF, (b) VBF, and (c) VH. Events are weighted by $\ln(1 + S_{90}/B_{90})$, where S_{90} (B_{90}) for each category is the expected signal (background) in the smallest $m_{\gamma\gamma}$ window containing 90% of the expected signal. The error bars represent 68% confidence intervals of the weighted sums. The solid red curve shows the fitted signal-plus-background model with the Higgs boson mass constrained to 125.09 ± 0.24 GeV. The dotted blue curve shows the continuum background, while the dashed green curve shows the combined continuum and non-targeted Higgs boson backgrounds. The lower panel shows the difference between the observed data and the fitted continuum background.

The signal strengths of the primary non-$t\bar{t}H$ production modes were found to be:

$$\mu_{ggF} = 0.97^{+0.17}_{-0.15} = 0.97^{+0.11}_{-0.11}\ \text{(stat.)}\ ^{+0.09}_{-0.08}\ \text{(exp.)}\ ^{+0.08}_{-0.06}\ \text{(theo.)}$$
$$\mu_{VBF} = 1.4^{+0.44}_{-0.37} = 1.4^{+0.32}_{-0.31}\ \text{(stat.)}\ ^{+0.21}_{-0.16}\ \text{(exp.)}\ ^{+0.21}_{-0.14}\ \text{(theo.)}$$
$$\mu_{VH} = 1.09^{+0.61}_{-0.55} = 1.09^{+0.53}_{-0.50}\ \text{(stat.)}\ ^{+0.23}_{-0.23}\ \text{(exp.)}\ ^{+0.210}_{-0.12}\ \text{(theo.)}$$

These signal strengths individually are compatible with the SM prediction.

The observed significance of the ggF production mode is $7.6\ \sigma$, compared to an expected significance of 7.9σ. This production mode is beyond the "discovery" stage, and so future work on the analysis should focus on improving the precision of the measurement. The observed significance of the VBF production mode is 4.7σ, compared to an expected significance of 3.6σ. Although this larger observed significance does represent an upward fluctuation compared to the SM prediction, it is smaller than that observed in the previous iteration of the analysis (performed using approximately 36 fb^{-1} of data). The VBF significance may hit the single-channel "discovery" threshold of 5σ with the addition of more data in the future. Lastly, the significance of the VH process is observed at 2.0σ, compared to an expectation of 1.9σ.

The measured cross sections times the diphoton branching ratio of the primary non-$t\bar{t}H$ production modes are presented in Table 11.3. The SM predictions (taken from Ref. [13]) are presented as well for comparison. The inclusive VH cross section includes the $W^{\pm}$ and $q\bar{q}ZH$ (previously referred to simply as "ZH") processes, as well as the much rarer $ggZH$ process. The ratios of these different processes are assumed to be those predicted by the SM. Additionally, the inclusive ggF cross section includes a small contribution from the $b\bar{b}H$ process, which is also assumed to contribute at the ratio predicted by the SM.

The results do not show any significant deviations from the SM predictions. Additionally, the measured cross sections times the diphoton branching ratio for the strong-merging scheme of the STXS framework are presented in Figure 11.6. The strong-merging scheme represents the merging of low-statistics categories. The $t\bar{t}H$ and tH processes are merged into the "top" bin; the 2018 data subset is not included in this result. Again, no significant deviations from the SM prediction are measured.

The theoretical and experimental uncertainties now outweigh the statistical uncertainties associated with the ggF and VBF measurements. This means that future measurements of these two channels will not strongly benefit from additional data. Significant work in reducing the experimental uncertainties will be critical to any near-term improvements in the analysis. The uncertainty on the VH measurements, on the other hand, is still dominated by statistical error.

The main systematics affecting the non-$t\bar{t}H$ measurements are those relating to the background modeling and the experimental measurement of photons. The background modeling (spurious

Table 11.3: Best-fit values and uncertainties of the non-$t\bar{t}H$ production mode cross sections times the Higgs to diphoton branching ratio. The SM predictions (taken from Ref. [13]) are shown for each production process. The central values and uncertainties are rounded.

Process	Result	Uncertainty				SM prediction
	[fb]	Total [fb]	Stat. [fb]	Exp. [fb]	Theo. [fb]	[fb]
ggF	98	$^{+15}_{-14}$	± 11	$^{+9}_{-8}$	$^{+4}_{-3}$	102^{+5}_{-7}
VBF	11.2	$^{+3.4}_{-3.0}$	$^{+2.6}_{-2.4}$	$^{+1.3}_{-1.2}$	$^{+1.9}_{-1.1}$	8.0 ± 0.2
VH	4.9	$^{+2.7}_{-2.5}$	$^{+2.4}_{-2.2}$	$^{+1.0}_{-0.9}$	$^{+0.6}_{-0.5}$	4.5 ± 0.2

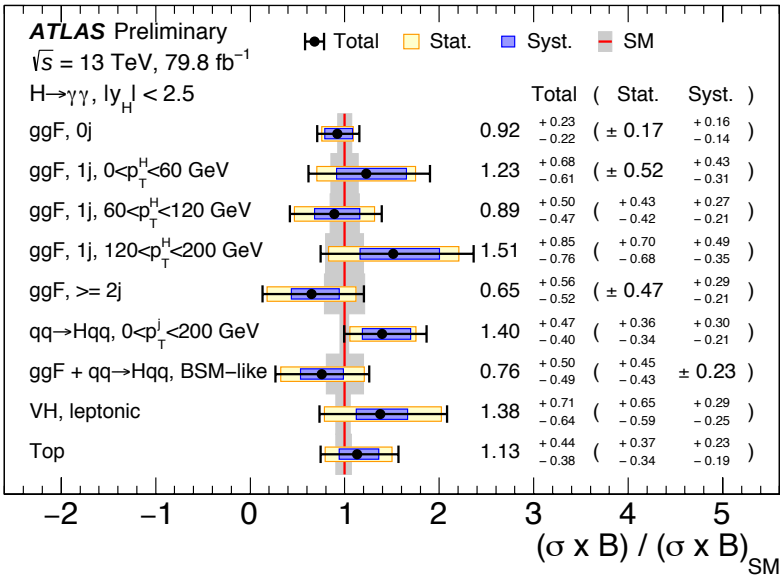


Figure 11.6: Summary plot of the measured cross sections times the Higgs to diphoton branching ratio in the strong merging scheme of the STXS framework. For illustration purposes the central values and uncertainties have been divided by their SM expectations. The uncertainties in the predicted SM cross sections are shown in gray bands in the plot. The "Top" contribution does not include the 2018 subset of data.

signal) systematic uncertainty relates to the choice of analytic background function used to model the continuum diphoton distribution; it is described in detail in Section 7.3. This uncertainty may be mitigated in the future through the use of the Gaussian Process smoothing technique, discussed in Chapter 8. The tests presented demonstrate the potential of the method to significantly reduce the impact of this uncertainty in future iterations of the analysis.

The uncertainty on the photon isolation efficiency also contributes significantly to the uncertainty on the measurement. The calorimetric isolation measurements from single photons (detailed in Section 4.5.1) were initially observed to disagree significantly with separate measurements from a sample of $Z \to \ell\ell\gamma$ events. Therefore, an additional uncertainty on the total photon isolation efficiency was added to cover the discrepancy (totaling about 5%). Later studies have pinpointed the main source of the disagreement to differences in truth-level requirements imposed on the simulated photons (for example, the $Z \to \ell\ell\gamma$ studies excluded photons form fragmentation, which were included in the single photon studies). The additional uncertainty component accounting for the single photon/$Z \to \ell\ell\gamma$ disagreement has since been removed, significantly reducing the photon isolation uncertainty. Therefore, future iterations of the analysis will no longer be as strongly impacted by the photon isolation uncertainty.

CHAPTER 12

Conclusions

This work presents the analysis of the couplings of the Higgs boson through the measurement of the $t\bar{t}H$, ggF, VBF, and inclusive VH production mode cross sections and signal strengths in the diphoton decay channel. The results are also presented in the context of the STXS framework. The $t\bar{t}H$ portion of the analysis uses the full Run 2 data set of proton-proton collisions collected by the ATLAS detector between 2015 and 2018, amounting to a total integrated luminosity of 139 fb^{-1}. The remainder of the analysis uses the 2015-2017 portion of the Run 2 data set, amounting to 78.9 fb^{-1}.

The measurement has very nearly achieved single channel observation (an observed 5σ significance in one decay channel alone) of the $t\bar{t}H$ and VBF processes within the diphoton decay mode. These achievements reflect the culmination of years of dedicated study of the Higgs boson and unlock new fronts through which to probe Higgs boson properties. Future studies of these production modes in the diphoton channel will shift from "discovery"-based analyses to more precision-focused analyses. As of now, however, no significant disagreement has been observed with the predictions of the SM in any of the presented results. Therefore, the measurements do not implicate any BSM physics.

Recalling the inclusive cross section measurements of the ggF production mode, which has the least associated uncertainty, indicates that the precision of the diphoton channel measurements is roughly 15%. Although this level of precision may disfavor BSM models with order-of-magnitude effects on the ggF cross section, there still exists ample room for some smaller BSM effects. These effects may be expected to scale very roughly as $1/\Lambda$, where Λ refers to the energy scale of the new physics. Therefore, the precision of the analysis is not yet strong enough to probe BSM effects multiple orders of magnitude above the Electroweak scale.

The measurements of the VH and $t\bar{t}H$ processes at this point suffer from large statistical uncertainties. Therefore, these channels will benefit from the additional Run 3 data due to be collected by the ATLAS detector in the coming years. However, the theoretical uncertainties associated with these channels are significant. Extensive theoretical work may be needed in order to improve the

precision of these measurements in the future. This work would likely focus on improving the modeling of the VH and $t\bar{t}H$ signal processes themselves. The $t\bar{t}H$ channel in particular may possibly benefit from the addition of higher-order corrections in the calculation of the inclusive cross-section, given that it currently includes only next-leading-order terms.

The $t\bar{t}H$ measurement might also benefit especially from a dedicated measurement of the rate of heavy flavor jet production in association with ggF events. This measurement would likely be best performed in the diphoton or four lepton decay channels of the Higgs boson, due to the fact that they do not contain additional b-jets from the Higgs boson decay. Further theoretical work would also allow for a dedicated comparison of the experimental measurements.

With additional statistics, the $t\bar{t}H$ region specifically could be exploited in order to search for BSM effects with a finer granularity. One option would be to perform a differential measurement of the Higgs boson p_{T} in these categories, since this spectrum could be sensitive to BSM effects. However, the current BDT-based selection would likely need to be reconsidered, since the output score of the BDTs are heavily correlated with the Higgs boson p_{T}. Another potential avenue of study would be to probe the CP (Charge-Parity) nature of the Higgs boson coupling to the top quark. This coupling is predicted to be CP-even by the SM, but some BSM models indicate the possibility of a CP-odd component to the coupling. A high-precision measurement of this coupling could be achieved by selecting $t\bar{t}H$ events which decay dileptonically (both $W^{\pm}$ bosons decay to leptons and neutrinos) and measuring the angular distributions constructed from the two leptons and the diphoton system. This channel has the advantage that there are no ambiguities between the decay products of the top quarks and the Higgs boson, unlike in other channels (such as $t\bar{t}H \to ZZ^* \to 4\ell$, where both the Higgs boson and top quarks produce leptons as part of the decay chain). However, the current analysis has only observed a handful (of order 10) dileptonic events, and so such an analysis will likely not be feasible until late into Run 3 of the LHC. Additionally, the BDT-based selection used by the current analysis may need to be reconsidered for such a measurement, since the leptonic $t\bar{t}H$ BDT cuts on the angular variables of the leptons in an event.

Significantly improving the ggF and VBF measurements will require immediate action, given that these results are not dominated by statistical uncertainties. As mentioned previously, one of the dominant experimental uncertainties on these measurements is that of the background modeling. For the low- and medium-statistics categories targeting this production mode, the GP smoothing technique may offer a significant reduction in the spurious signal and hence drastically reduce the uncertainty. However, the smoothing technique (by design) will not eliminate real features in the background spectrum that may be difficult to fit using an analytic function. Therefore, it may be worth questioning the analysis strategy. One option would be to attempt to model the background spectrum directly using a GP. A GP is not constricted to any particular functional form, and it is flexible enough to fit the many features of the background spectrum. Work would be needed to

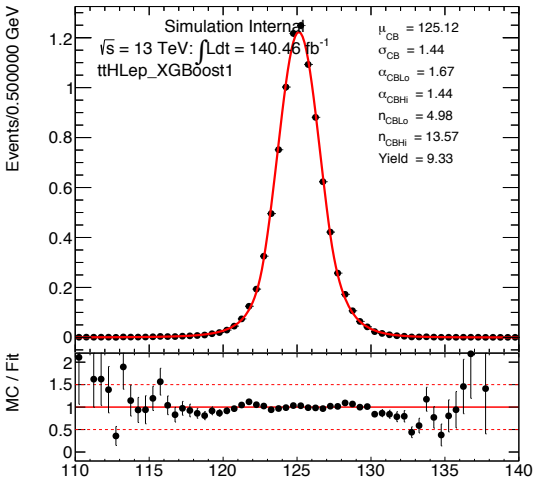
determine a procedure to extract the signal from the GP fit to the data.

Moving forward, the diphoton channel will continue to be an appealing experimental channel through which to study the properties of the Higgs boson. As more data is collected, sufficient statistics of this rare decay mode will allow for finer granularity probes of the SM. However, extensive theoretical and experimental work will be needed moving forward in order to extract extremely high precision measurements of Higgs boson couplings in the future.

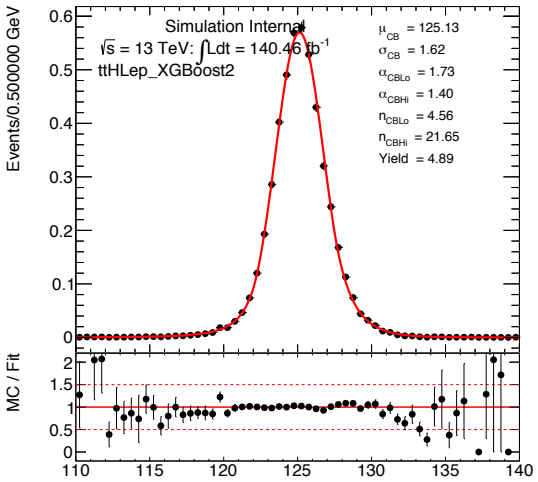
APPENDIX A

Signal Parameterizations

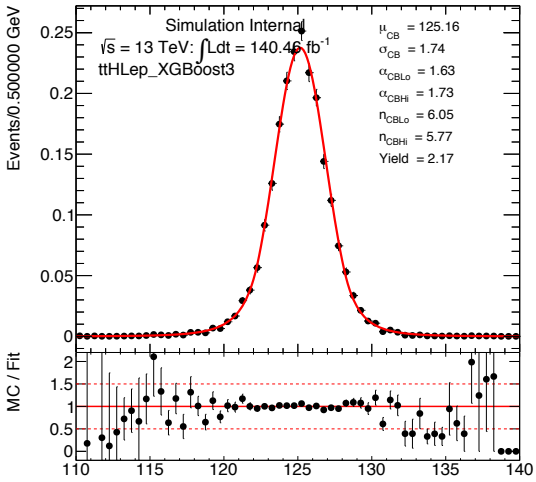
The Double-Sided Crystal Ball (DSCB) fits to the simulated signal samples in each of the $t\bar{t}H$ categories are presented in Figures A.1-A.2, while the signal shapes of the non-$t\bar{t}H$ categories are presented in Figures A.3-A.9. The construction of the signal models is discussed in Section 7.1.



(a) $t\bar{t}H$ Lep 1

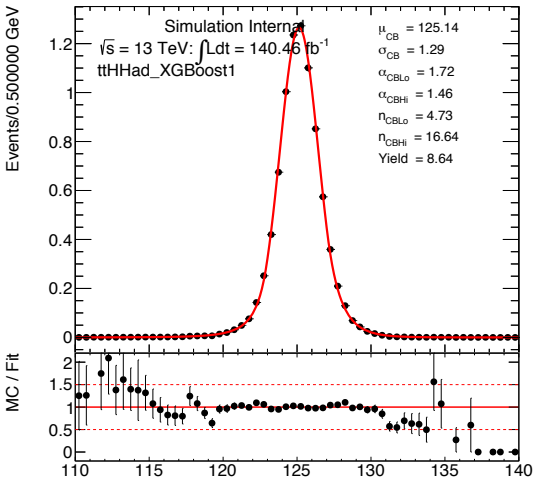


(b) $t\bar{t}H$ Lep 2

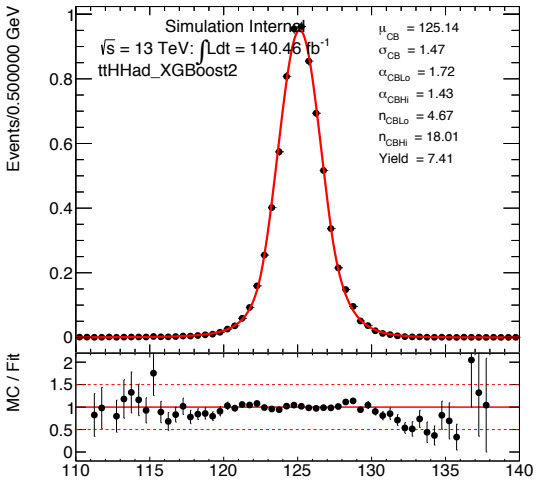


(c) $t\bar{t}H$ Lep 3

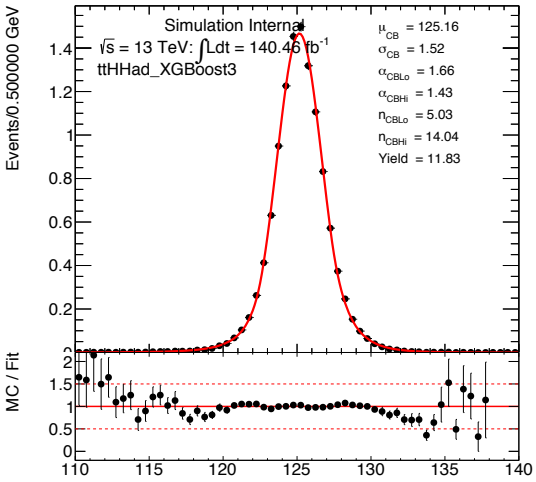
Figure A.1: The signal parameterizations, using simulates signal events, of the $t\bar{t}H$ leptonic categories: (a) $t\bar{t}H$ Lep 1, (b) $t\bar{t}H$ Lep 2, and (c) $t\bar{t}H$ Lep 3.



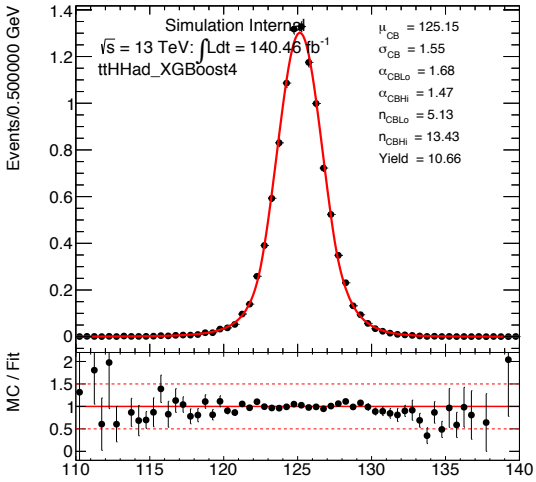
(a) $t\bar{t}H$ Had 1



(b) $t\bar{t}H$ Had 2

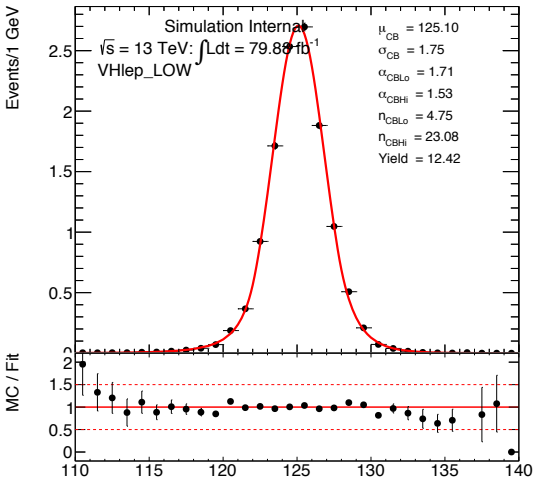


(c) $t\bar{t}H$ Had 3

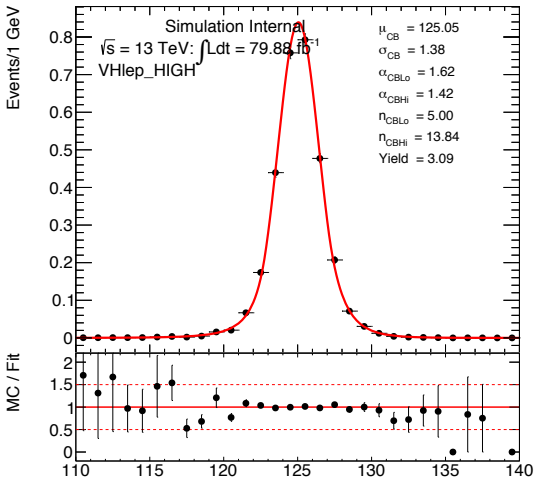


(d) $t\bar{t}H$ Had 4

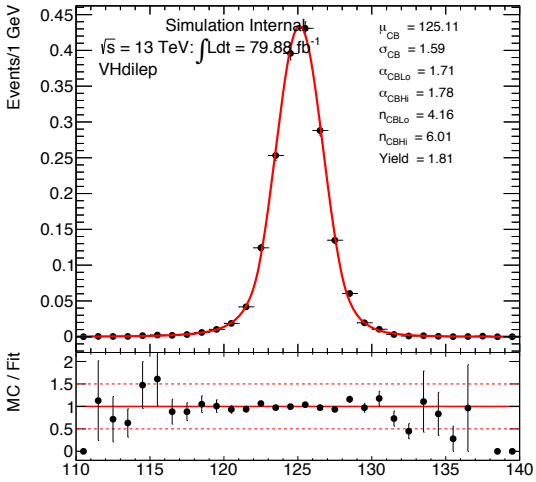
Figure A.2: The signal parameterizations, using simulates signal events, of the $t\bar{t}H$ hadronic categories: (a) $t\bar{t}H$ Had 1, (b) $t\bar{t}H$ Had 2, (c) $t\bar{t}H$ Had 3, and (d) $t\bar{t}H$ Had 4.



(a) VH Lep Low

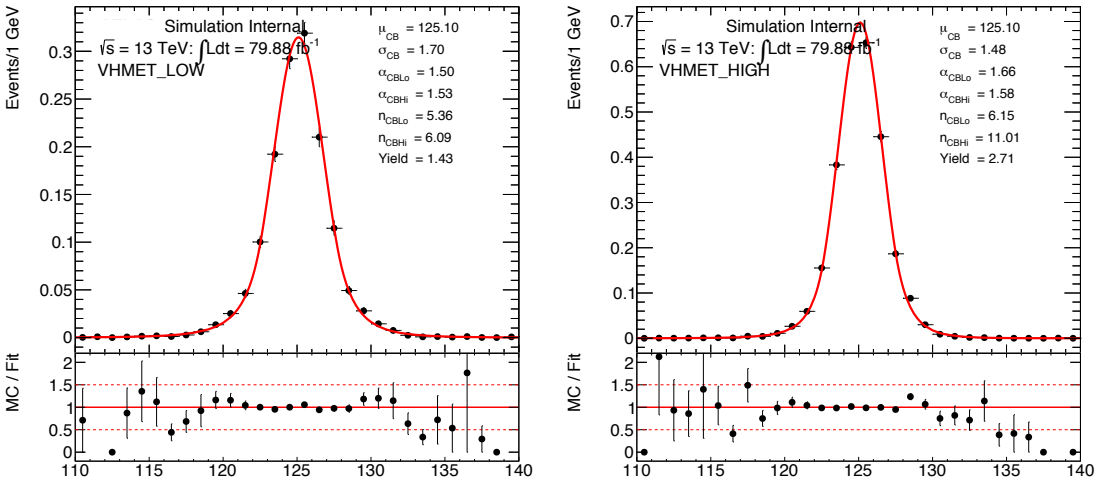


(b) VH Lep High



(c) VH Dilep

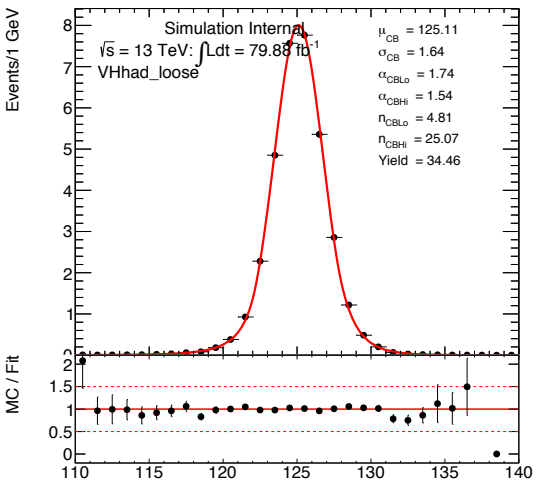
Figure A.3: The signal parameterizations, using simulates signal events, of the VH leptonic categories: (a) VH Lep Low, (b) VH Lep High, and (c) VH Dilep.



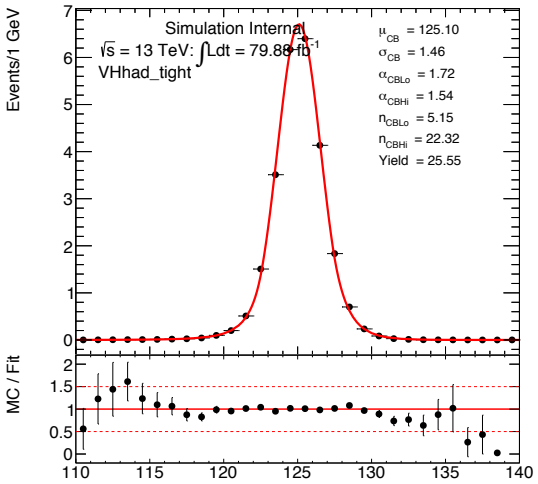
(a) VH MET Low

(b) VH MET High

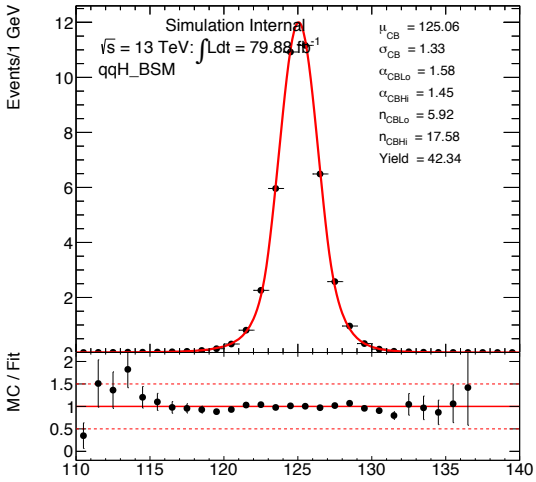
Figure A.4: The signal parameterizations, using simulates signal events, of the VH MET categories: (a) VH MET Low and (b) VH MET High.



(a) VH Had Loose

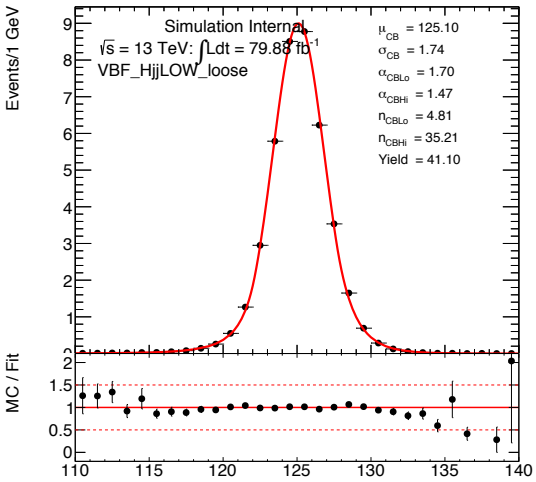


(b) VH Had Tight

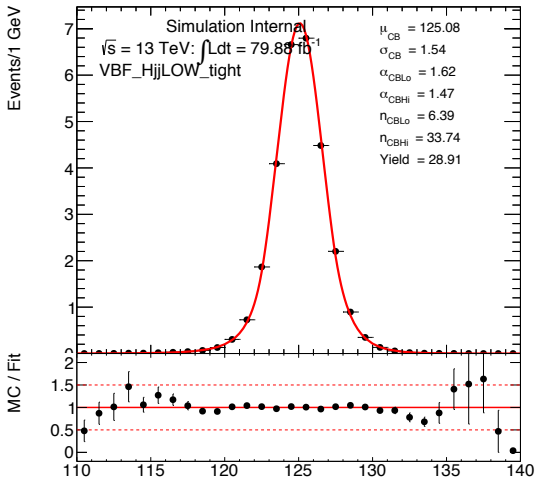


(c) qqH BSM

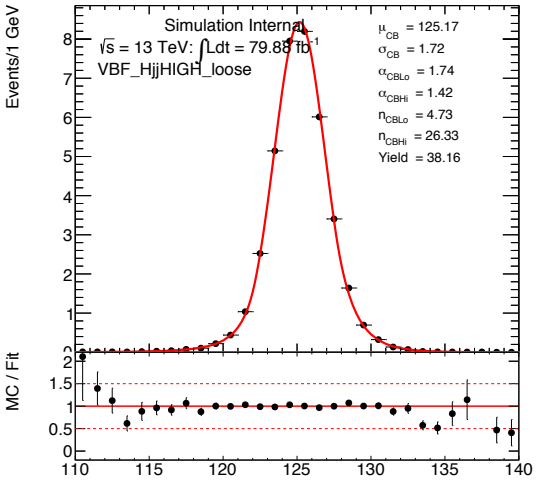
Figure A.5: The signal parameterizations, using simulates signal events, of the VH leptonic categories: (a) VH Had Loose, (b) VH Had Tight, and (c) qqH BSM.



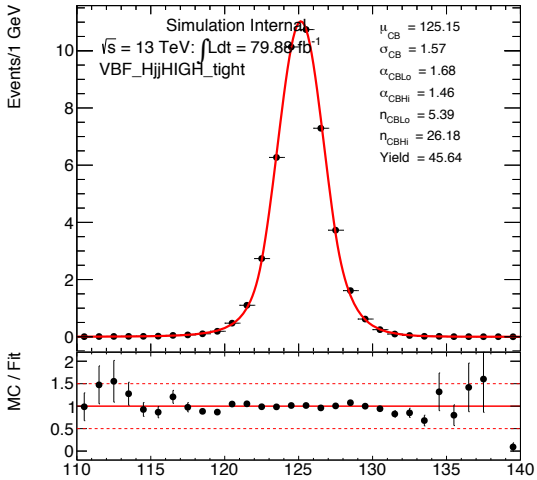
(a) VBF Loose, Low p_{T}^{Hjj}



(b) VBF Tight, Low p_{T}^{Hjj}

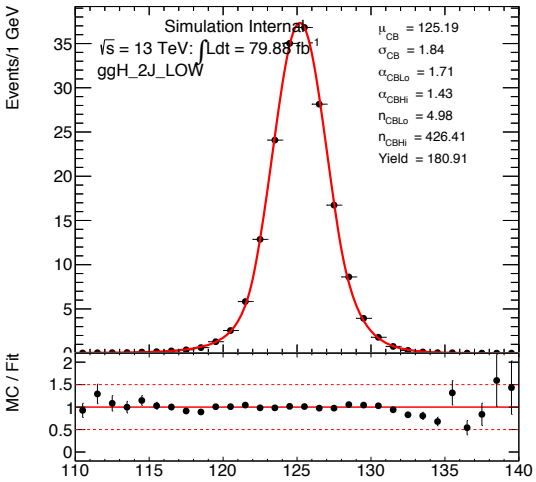


(c) VBF Loose, High p_{T}^{Hjj}

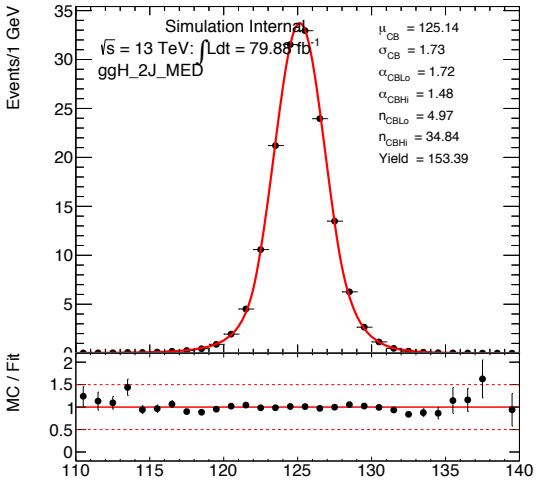


(d) VBF Tight, High p_{T}^{Hjj}

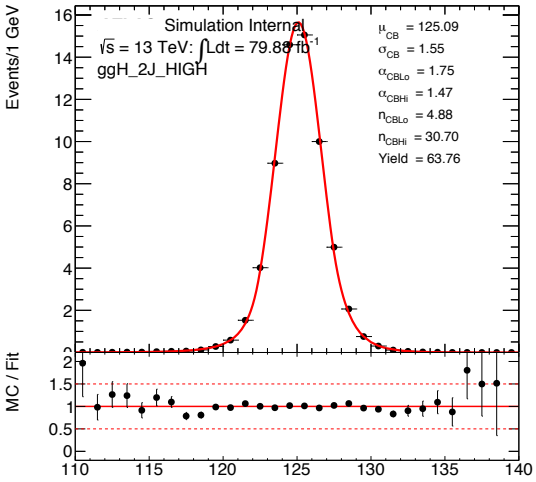
Figure A.6: The signal parameterizations, using simulates signal events, of the VBF categories: (a) VBF Loose, Low p_{T}^{Hjj}, (b) VBF Tight, Low p_{T}^{Hjj}, (c) VBF Loose, High p_{T}^{Hjj}, and (d) VBF Tight, High p_{T}^{Hjj}.



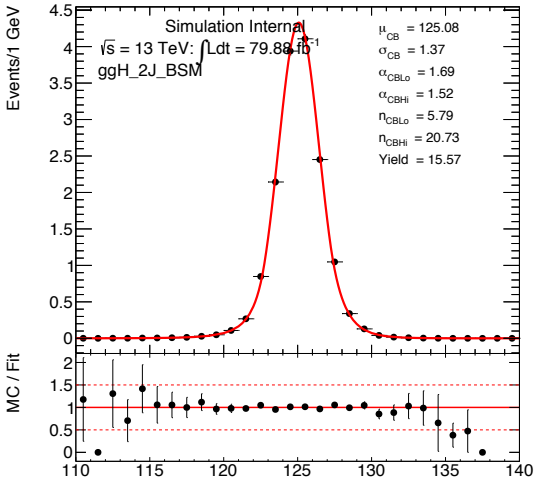
(a) ggH 2J Low



(b) ggH 2J Med

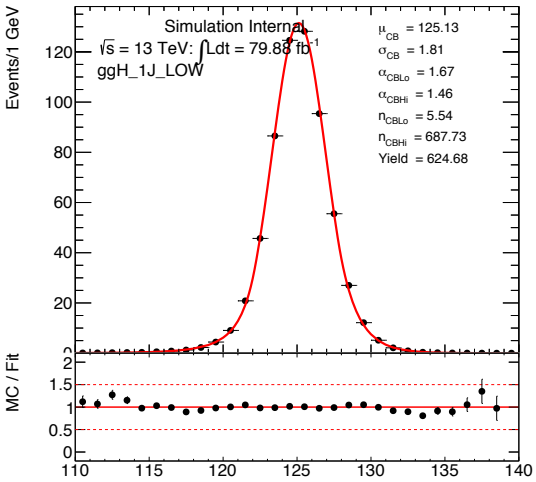


(c) ggH 2J High

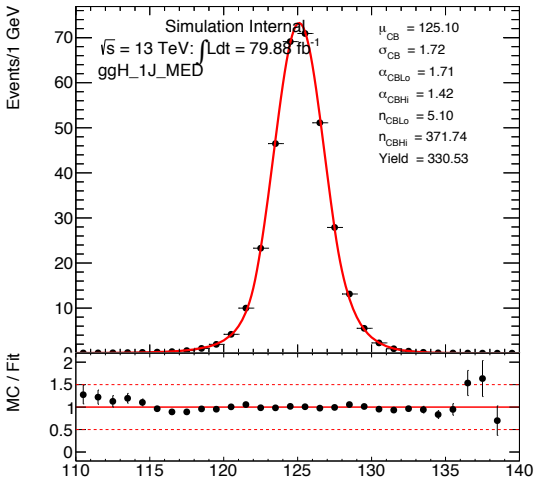


(d) ggH 2J BSM

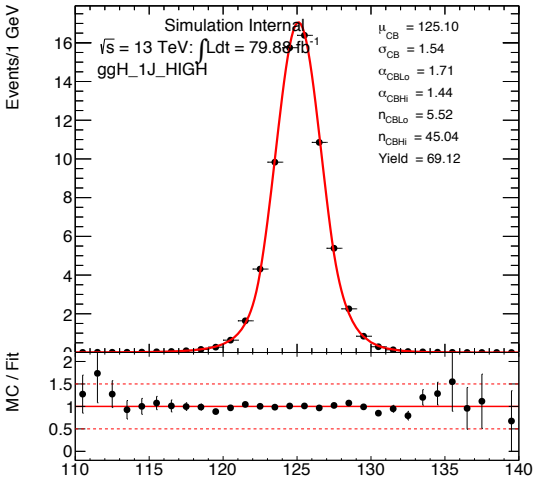
Figure A.7: The signal parameterizations, using simulates signal events, of the ggF 2J categories: (a) ggH 2J Low, (b) ggH 2J Med, (c) ggH 2J High, and (d) ggH 2J BSM.



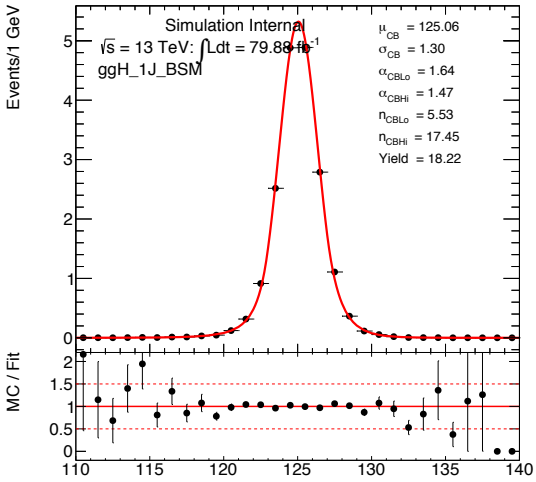
(a) ggH 1J Low



(b) ggH 1J Med

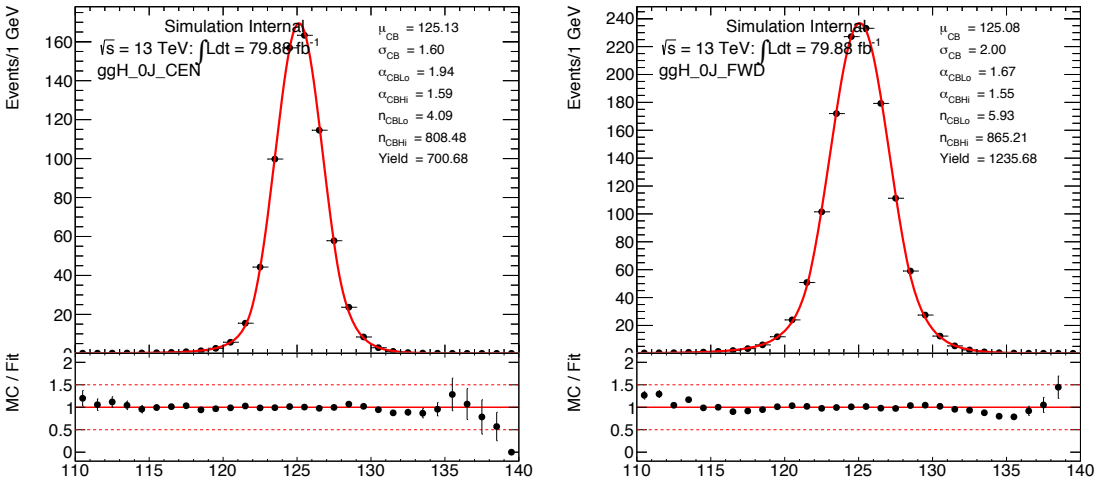


(c) ggH 1J High



(d) ggH 1J BSM

Figure A.8: The signal parameterizations, using simulates signal events, of the ggF 1J categories: (a) ggH 1J Low, (b) ggH 1J Med, (c) ggH 1J High, and (d) ggH 1J BSM.



(a) ggH 0J CEN (b) ggH 0J FWD

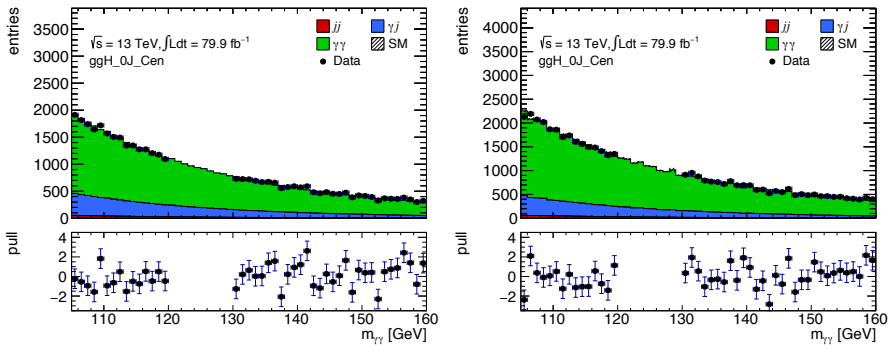
Figure A.9: The signal parameterizations, using simulates signal events, of the ggF 0J categories: (a) ggH 0J CEN and (b) ggH 0J FWD.

APPENDIX B

Background Templates

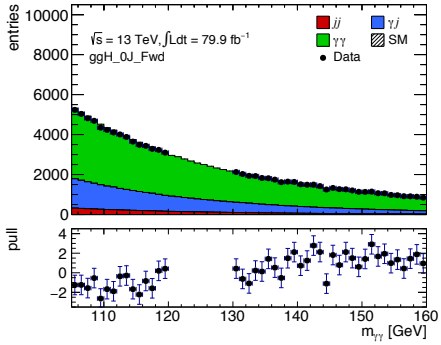
The background templates for the non-$t\bar{t}H$ analysis categories are presented in Figures B.1-B.7. Due to the distinction of the simulation samples modeling the pileup profile of the 2015-2016 data taking period and that of the 2017 data-taking period, the templates for the two periods are presented side-by-side. The exception is for the presentation of the VBF templates, where only the combined 2015-2017 templates are presented due to a statistical fluctuation in the 2017 subsample of simulated continuum diphoton events. The data events in the sideband region (defined as $m_{\gamma\gamma} < 120$ or $m_{\gamma\gamma} > 130$ GeV) are presented as well. The construction of the background templates is presented in Section 7.3.

The background templates used for the $t\bar{t}H$ categories are shown in Figures B.8-B.9. The data sideband events are also shown in order to evaluate the compatibility of the background-only template and the true background shape. Additionally, the "alternative" background templates, constructed to validate the loosened criteria used to improve the template statistics, are presented. The construction of the nominal and alternative $t\bar{t}H$ background templates is presented in Section 7.3.

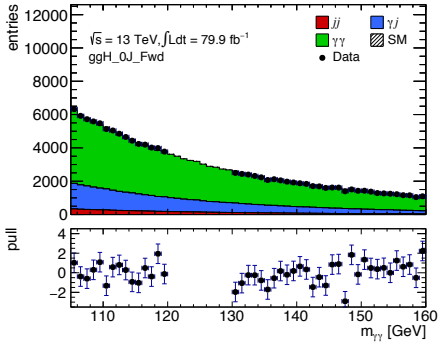


(a) ggF 0J CEN (2015-2016)

(b) ggF 0J CEN (2017)

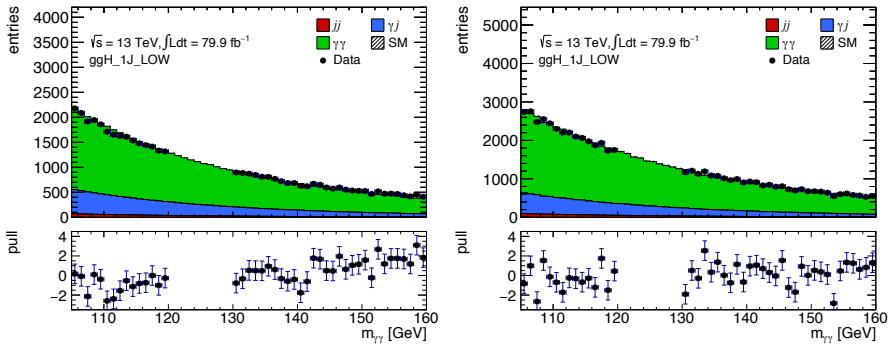


(c) ggF 0J FWD (2015-2016)



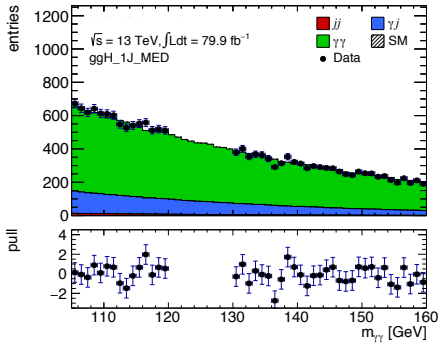
(d) ggF 0J FWD (2017)

Figure B.1: The 2015-2016 (left) and 2017 (right) background templates, along with the data side-band events in black, for the (top row) ggF 0J Cen and (bottom row) ggF 0J Fwd categories. The green shows the $\gamma\gamma$ events, the blue shows the γj events, and the red shows the jj events.

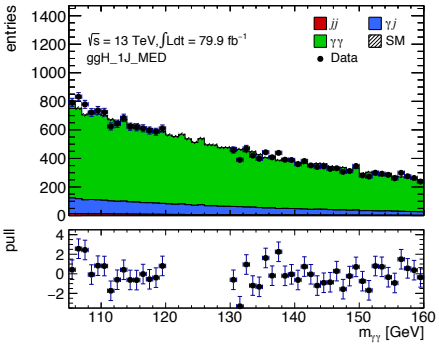


(a) ggF 1J Low (2015-2016)

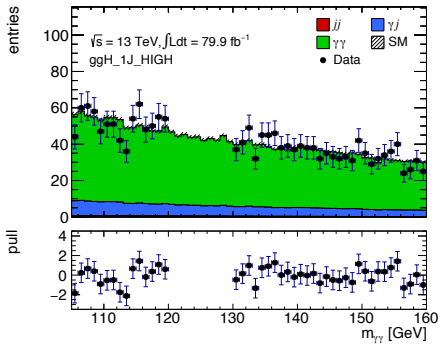
(b) ggF 1J Low (2017)



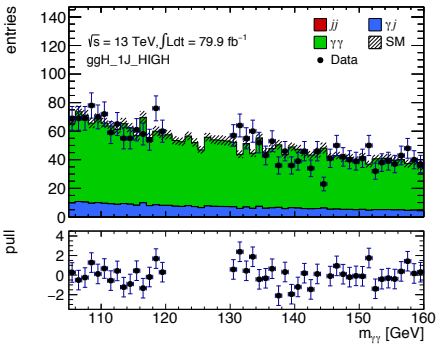
(c) ggF 1J Med (2015-2016)



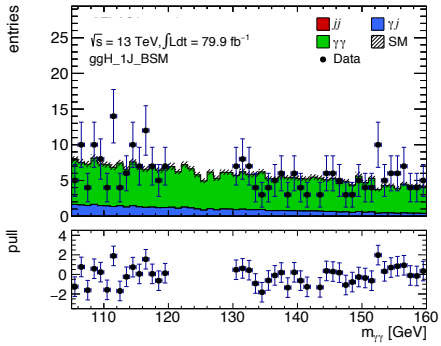
(d) ggF 1J Med (2017)



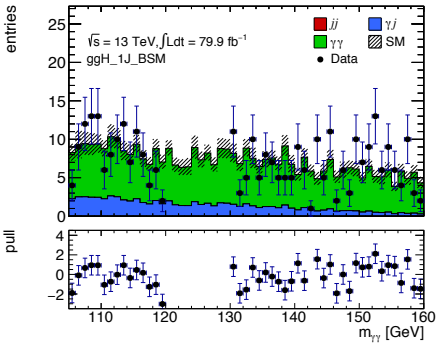
(e) ggF 1J High (2015-2016)



(f) ggF 1J High (2017)

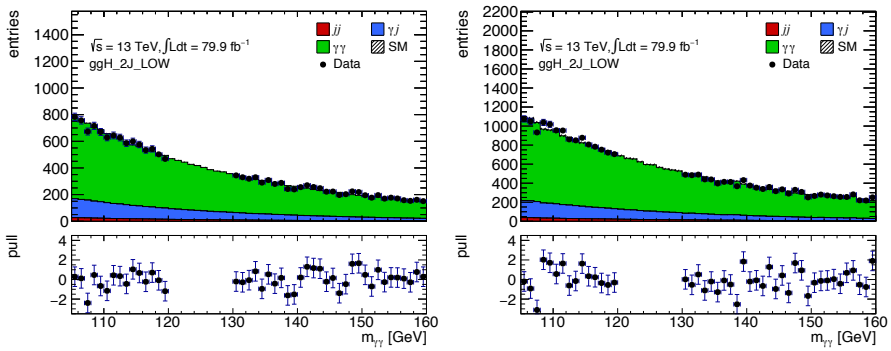


(g) ggF 1J BSM (2015-2016)

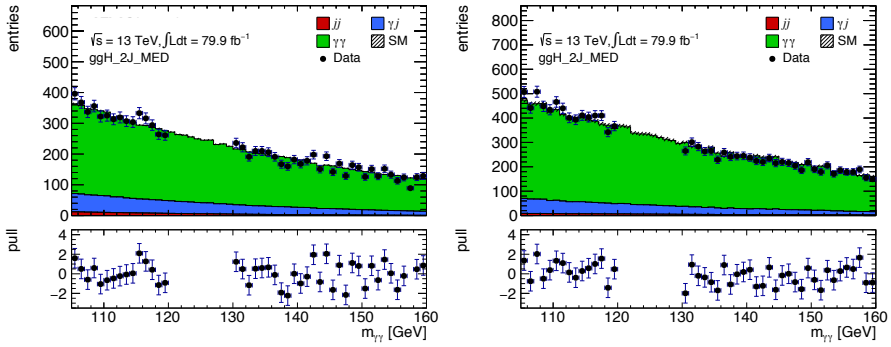


(h) ggF 1J BSM (2017)

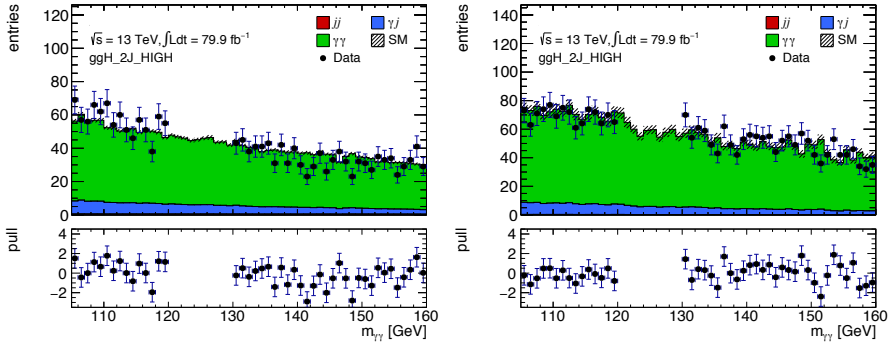
Figure B.2: The 2015-2016 (left) and 2017 (right) background templates, along with the data side-band events in black, for the (top row) ggF 1J LOW, (second row) ggF 1J Med, (third row) ggF 1J High, and (bottom row) ggF 1J BSM categories. The green shows the $\gamma\gamma$ events, the blue shows the γj events, and the red shows the jj events.



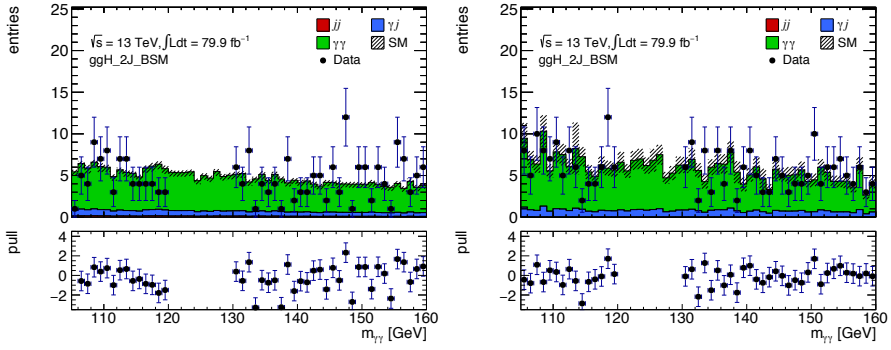
(a) ggF 2J Low (2015-2016) (b) ggF 2J Low (2017)



(c) ggF 2J Med (2015-2016) (d) ggF 2J Med (2017)

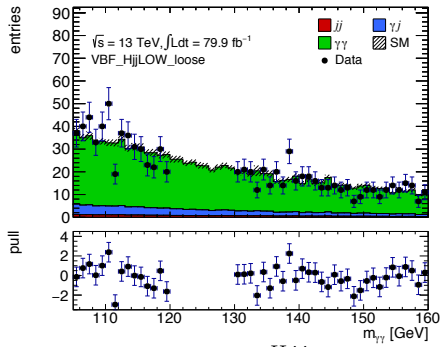


(e) ggF 2J High (2015-2016) (f) ggF 2J High (2017)

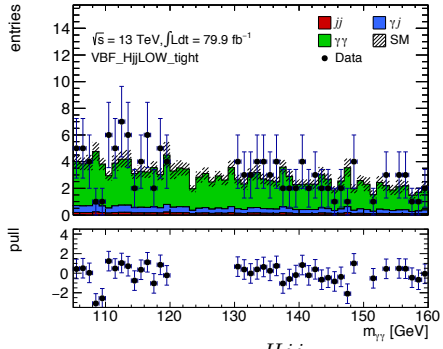


(g) ggF 2J BSM (2015-2016) (h) ggF 2J BSM (2017)

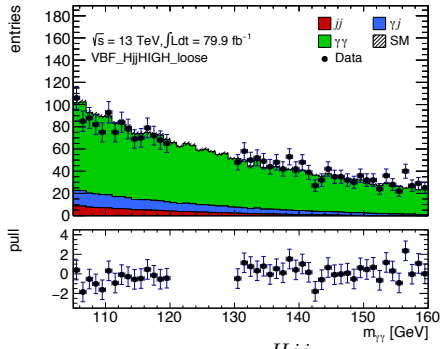
Figure B.3: The 2015-2016 (left) and 2017 (right) background templates, along with the data side-band events in black, for the (top row) ggF 2J LOW, (second row) ggF 2J Med, (third row) ggF 2J High, and (bottom row) ggF 2J BSM categories. The green shows the $\gamma\gamma$ events, the blue shows the γj events, and the red shows the jj events.



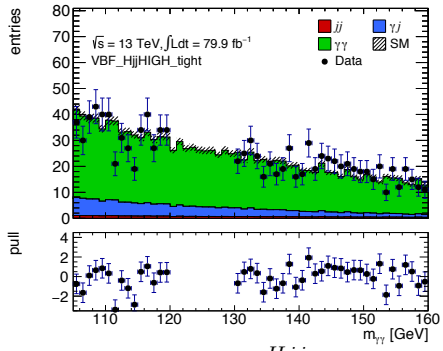
(a) VBF Loose, Low p_{T}^{Hjj} (2015-2017)



(b) VBF Tight, Low p_{T}^{Hjj} (2015-2017)

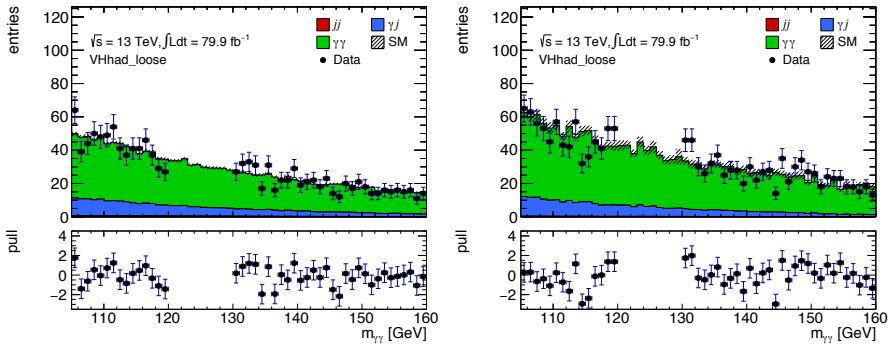


(c) VBF Loose, High p_{T}^{Hjj} (2015-2017)

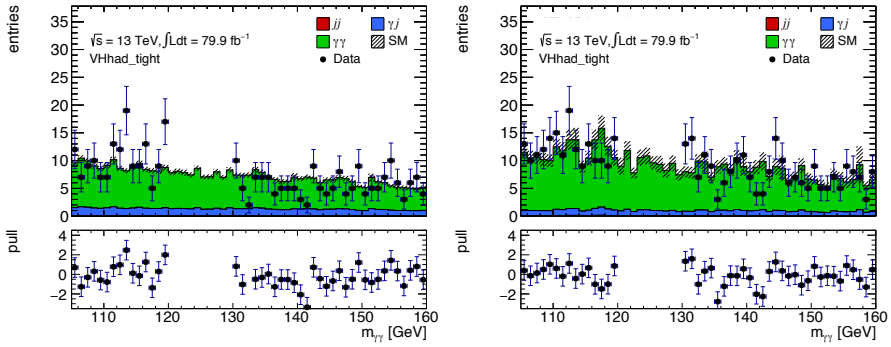


(d) VBF Tight, High p_{T}^{Hjj} (2015-2017)

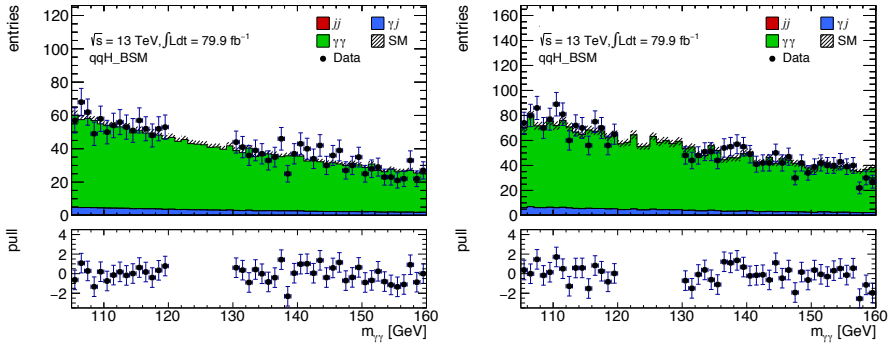
Figure B.4: The 2015-2017 background templates, along with the data side-band events in black, for the (top row) VBF Loose, Low p_{T}^{Hjj}, (second row) VBF Tight, Low p_{T}^{Hjj}, (third row) VBF Loose, High p_{T}^{Hjj}, and (bottom row) VBF Tight, High p_{T}^{Hjj} categories. The green shows the $\gamma\gamma$ events, the blue shows the γj events, and the red shows the jj events.



(a) VH Had Loose (2015-2016) (b) VH Had Loose (2017)

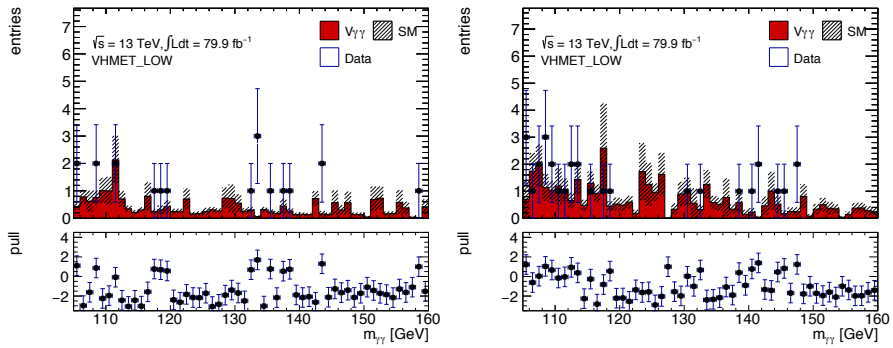


(c) VH Had Tight (2015-2016) (d) VH Had Tight (2017)



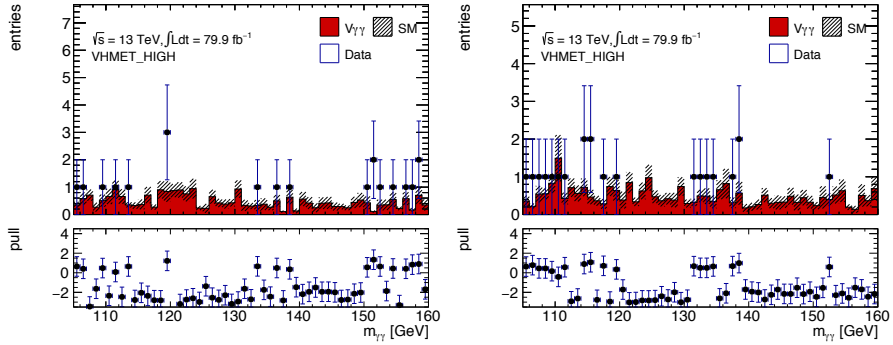
(e) qqH BSM (2015-2016) (f) qqH BSM (2017)

Figure B.5: The 2015-2016 (left) and 2017 (right) background templates, along with the data side-band events in black, for the (top row) VH Had Loose, (middle row) VH Had Tight, and (bottom row) qqH BSM categories. The green shows the $\gamma\gamma$ events, the blue shows the γj events, and the red shows the jj events.



(a) VH MET Low (2015-2016)

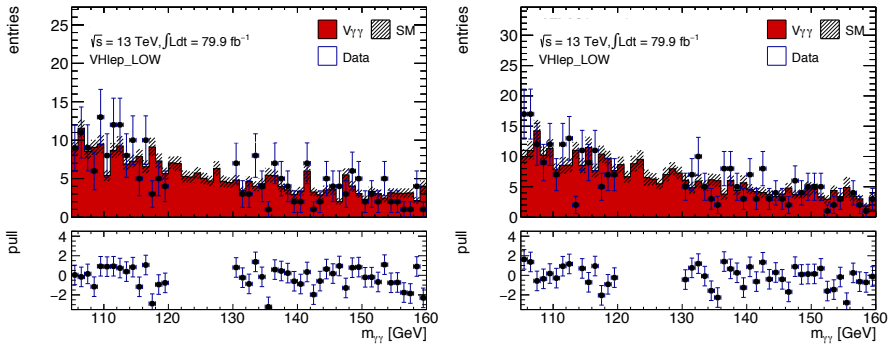
(b) VH MET Low(2017)



(c) VH MET High (2015-2016)

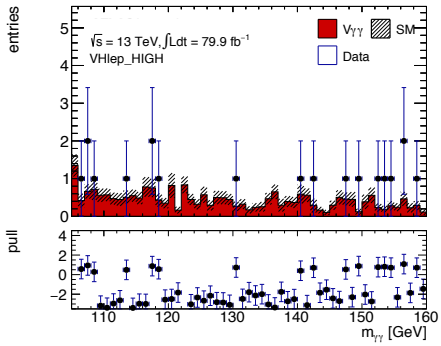
(d) VH MET High (2017)

Figure B.6: The 2015-2016 (left) and 2017 (right) background templates, along with the data sideband events in black, for the (top row) VH MET Low and (bottom row) VH MET High categories. Only the $V\gamma\gamma$ simulated sample is used (shown in red).

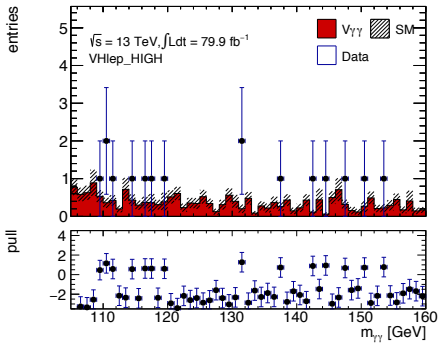


(a) VH Lep Low (2015-2016)

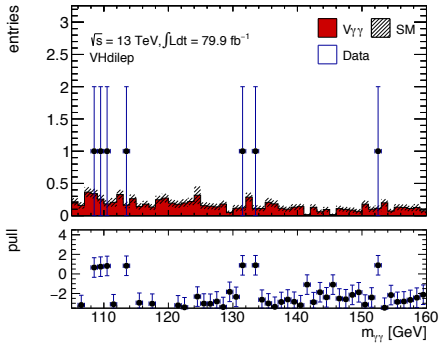
(b) VH Lep Low(2017)



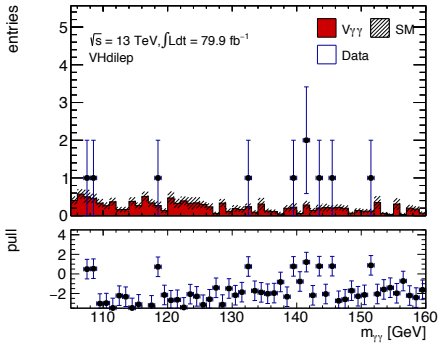
(c) VH Lep High (2015-2016)



(d) VH Lep High(2017)

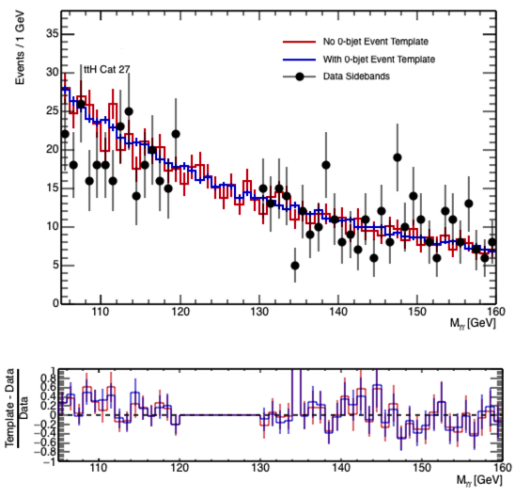


(e) VH Dilep (2015-2016)

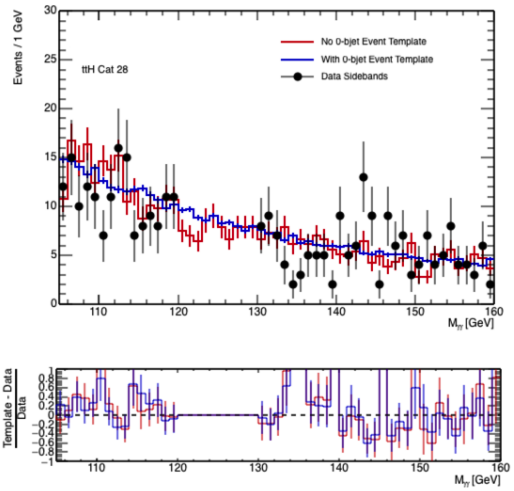


(f) VH Dilep (2017)

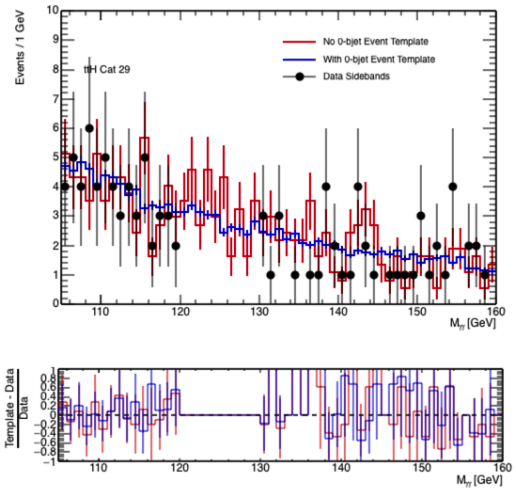
Figure B.7: The 2015-2016 (left) and 2017 (right) background templates, along with the data side-band events in black, for the (top row) VH MET Low and (bottom row) VH MET High categories. Only the $V\gamma\gamma$ simulated sample is used (shown in red).



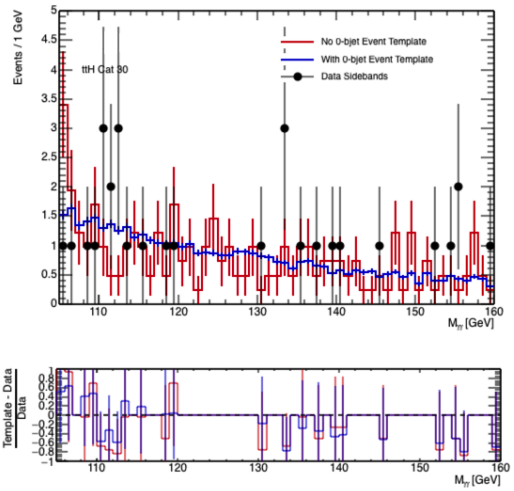
(a) $t\bar{t}H$ Had 4 (2015-2018)



(b) $t\bar{t}H$ Had 3 (2015-2018)

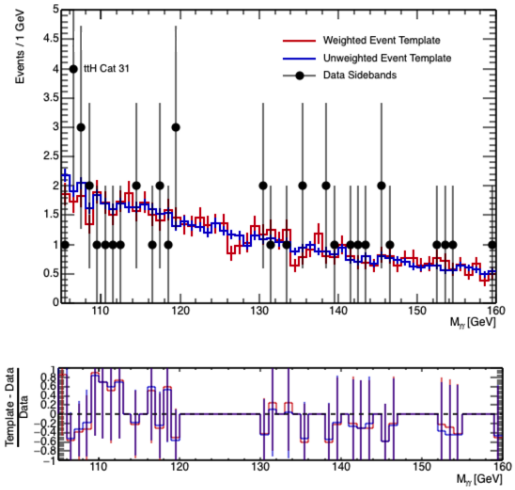


(c) $t\bar{t}H$ Had 2 (2015-2018)

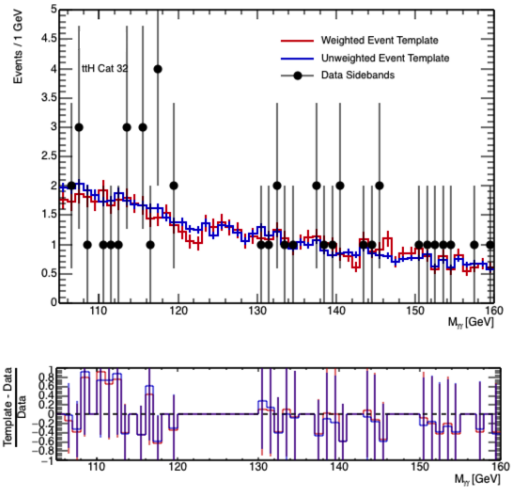


(d) $t\bar{t}H$ Had 1 (2015-2018)

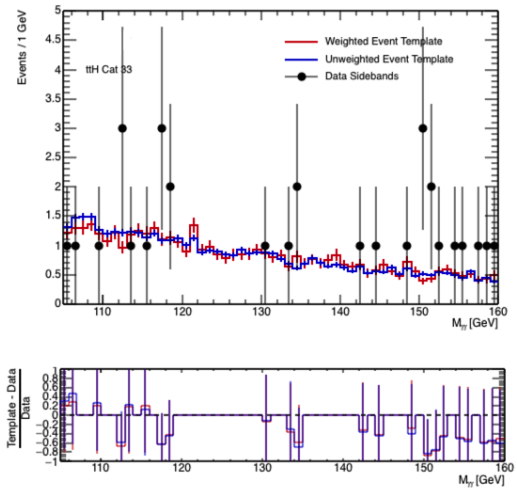
Figure B.8: The full Run 2 background-only templates (using the NTI data control region events, without the b-jet requirement) for the $t\bar{t}H$ hadronic categories, shown in blue. The data sideband events are shown as black dots. The red shape shows an alternative template obtained using only the NTI data control region events which pass the b-jet requirement applied to the nominal analysis categories. The apparent compatibility between the nominal (blue) and tighter alternative (red) templates confirms the validity of using the loosened sample.



(a) $t\bar{t}H$ Lep 3 (2015-2018)



(b) $t\bar{t}H$ Lep 2 (2015-2018)



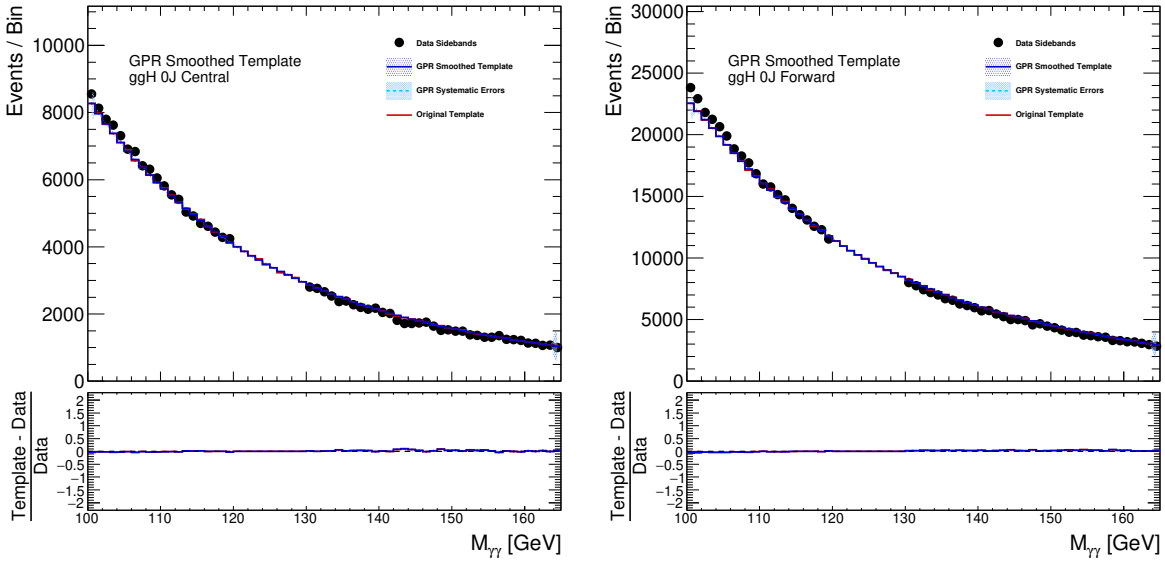
(c) $t\bar{t}H$ Lep 1 (2015-2018)

Figure B.9: The full Run 2 background-only templates (using the simulated $t\bar{t}\gamma\gamma$ sample, without the event weights applied) for the $t\bar{t}H$ leptonic categories, shown in blue. The data sideband events are shown as black dots. The red shape shows an alternative template obtained using by applying the event weights to the simulated $t\bar{t}\gamma\gamma$ events. The apparent compatibility between the nominal (blue) and tighter alternative (red) templates confirms the validity of using the loosened sample.

APPENDIX C

Gaussian Process Smoothed Templates

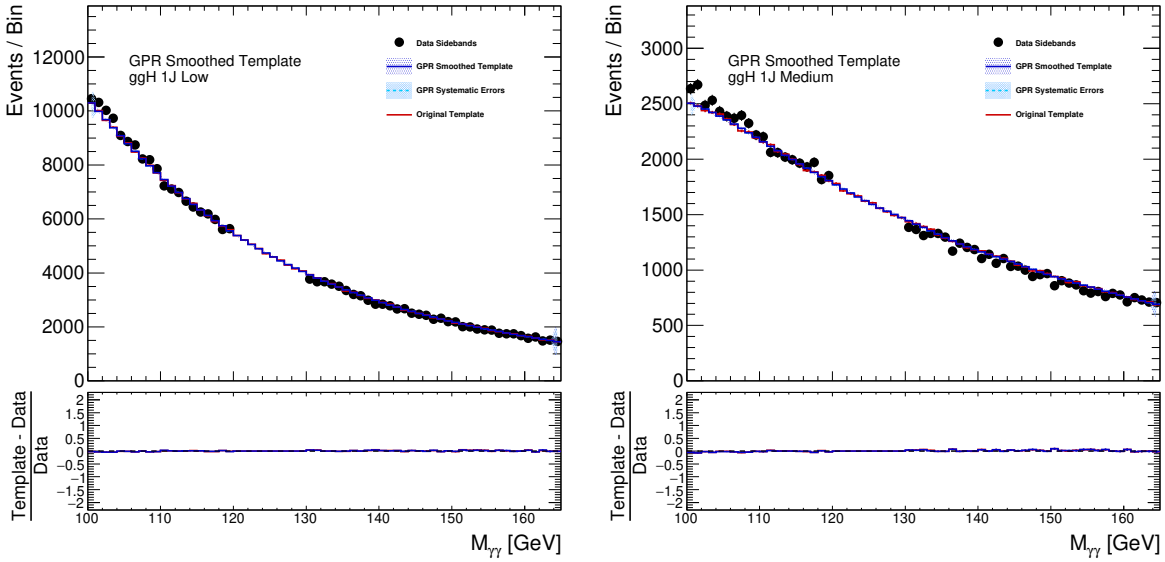
The background templates of all of the (non-$t\bar{t}H$) analysis categories, both before and after the Gaussian Process (GP) smoothing, are presented in Figures C.1-C.7. The data sidebands are shown for comparison, although the GP smoothing technique does not take into account the data sidebands. The $t\bar{t}H$ categories were not tested using the technique, since the spurious signal systematic uncertainty does not affect the $t\bar{t}H$ sensitivity as much as that of the other production modes. Note that the presented templates include the full Run 2 data set (2015-2018).



(a) ggF 0J CEN

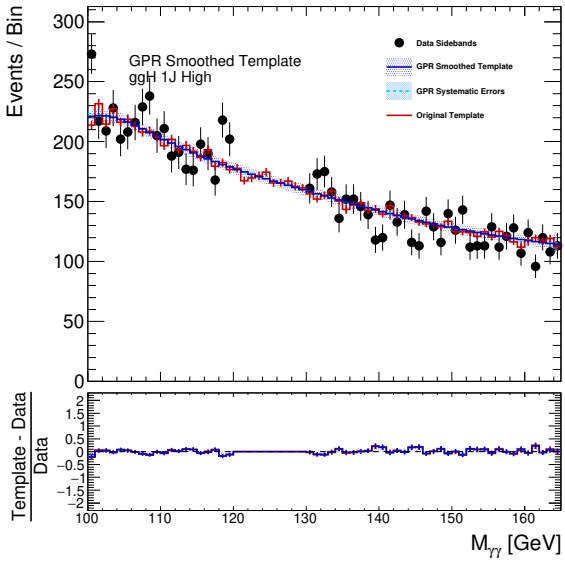
(b) ggF 0J FWD

Figure C.1: The full Run 2 background templates of the ggF 0J analysis categories, (a) ggF 0J CEN and (b) ggF 0J FWD. The red shape shows the original background template, the blue shape shows the smoothed background template, and the black points show the data sidebands (for reference). The bottom panel shows the fractional difference between the smoothed- and unsmoothed templates and the data sidebands.

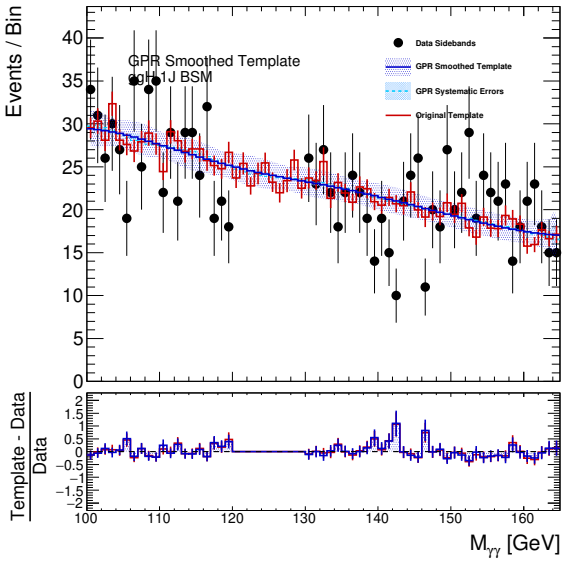


(a) ggF 1J LOW

(b) ggF 1J MED

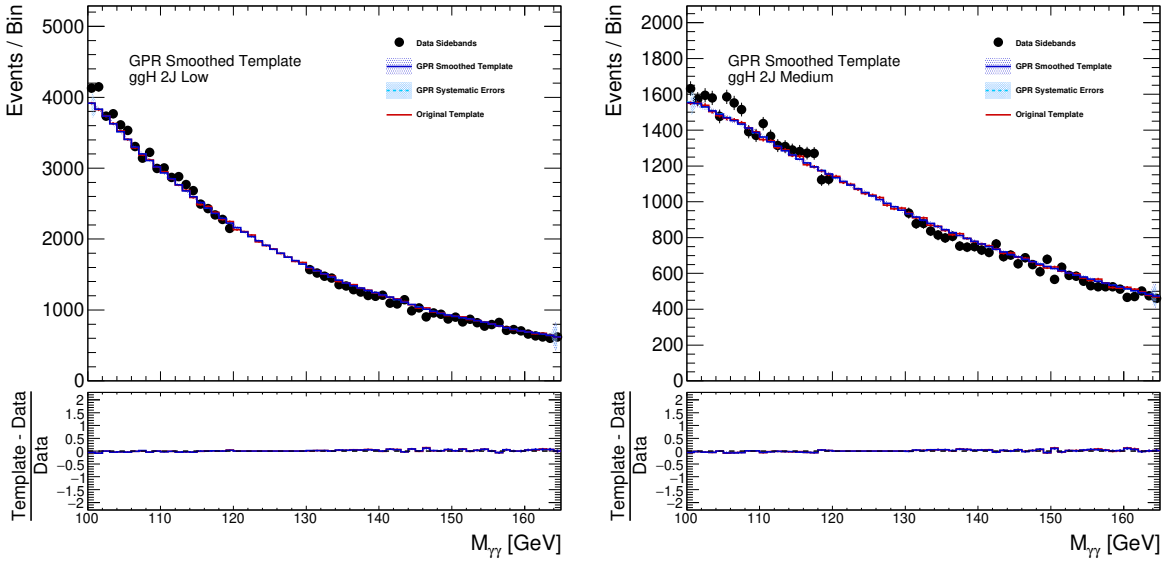


(c) ggF 1J HIGH



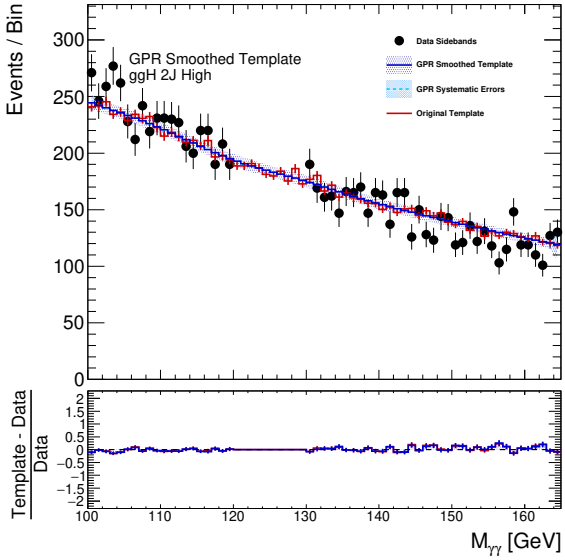
(d) ggF 1J BSM

Figure C.2: The full Run 2 background templates of the ggF 1J analysis categories, (a) ggF 1J LOW, (b) ggF 1J MED, (c) ggF 1J HIGH, and (d) ggF 1J BSM. The red shape shows the original background template, the blue shape shows the smoothed background template, and the black points show the data sidebands (for reference). The bottom panel shows the fractional difference between the smoothed- and un-smoothed templates and the data sidebands.

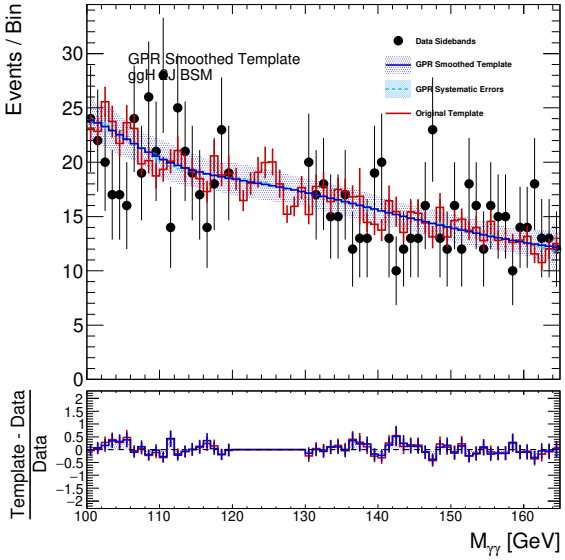


(a) ggF 2J LOW

(b) ggF 2J MED

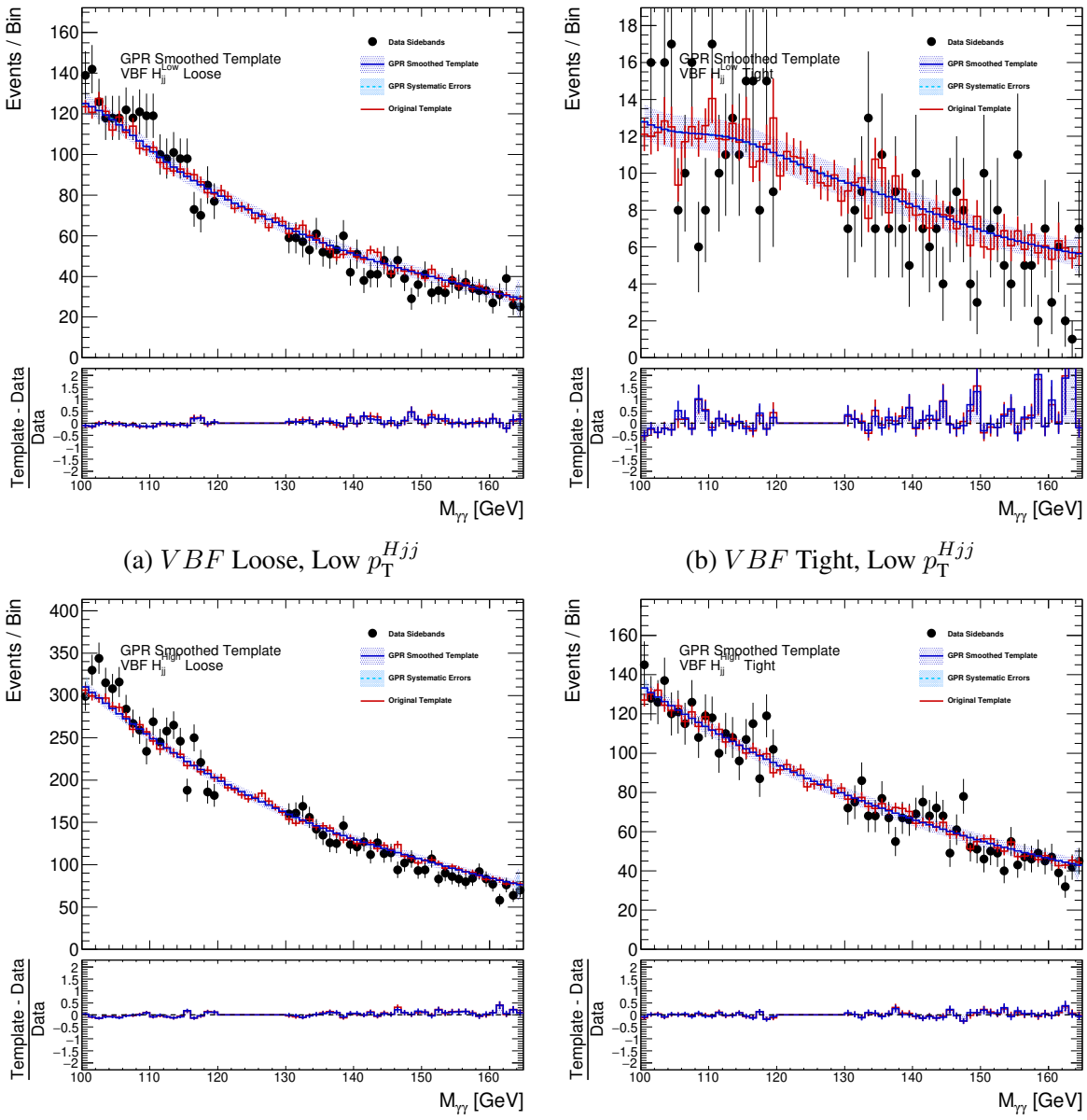


(c) ggF 2J HIGH



(d) ggF 2J BSM

Figure C.3: The full Run 2 background templates of the ggF 2J analysis categories, (a) ggF 2J LOW, (b) ggF 2J MED, (c) ggF 2J HIGH, and (d) ggF 2J BSM. The red shape shows the original background template, the blue shape shows the smoothed background template, and the black points show the data sidebands (for reference). The bottom panel shows the fractional difference between the smoothed- and un-smoothed templates and the data sidebands.



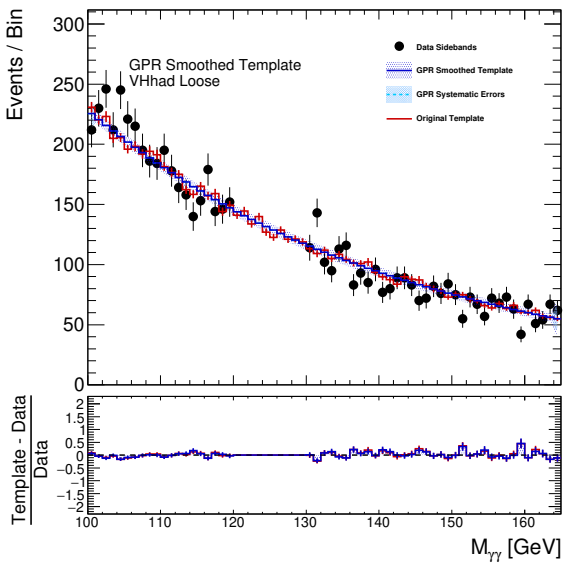
(a) VBF Loose, Low p_{T}^{Hjj}

(b) VBF Tight, Low p_{T}^{Hjj}

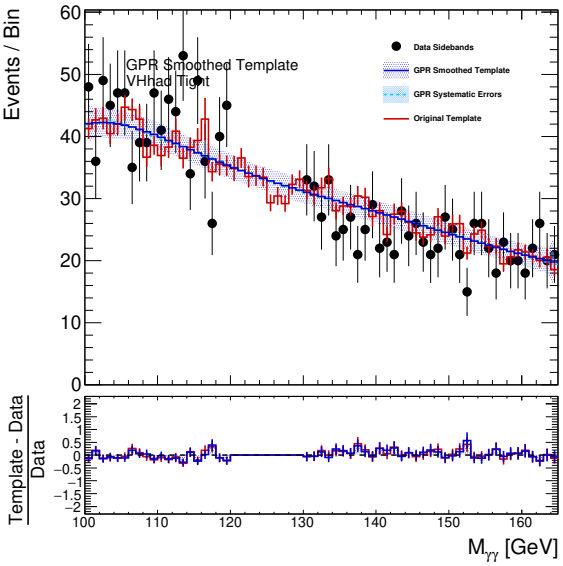
(c) VBF Loose, High p_{T}^{Hjj}

(d) VBF Tight, High p_{T}^{Hjj}

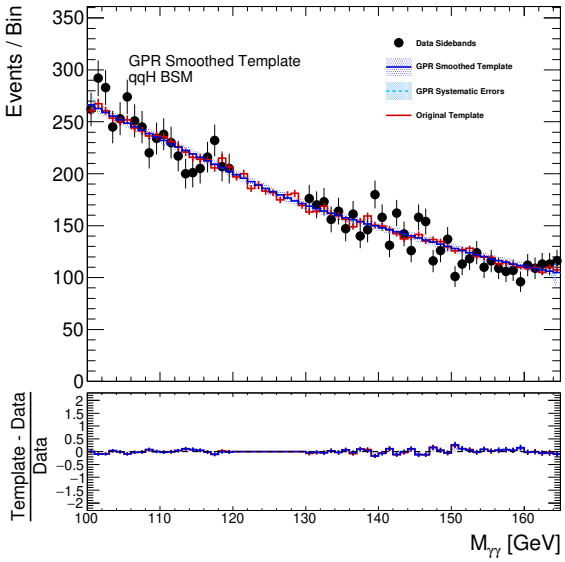
Figure C.4: The full Run 2 background templates of the VBF analysis categories, (a) VBF Loose, Low p_{T}^{Hjj}, (b) VBF Tight, Low p_{T}^{Hjj}, (c) VBF Loose, High p_{T}^{Hjj}, and (d) VBF Tight, High p_{T}^{Hjj}. The red shape shows the original background template, the blue shape shows the smoothed background template, and the black points show the data sidebands (for reference). The bottom panel shows the fractional difference between the smoothed- and un-smoothed templates and the data sidebands.



(a) VH Had Loose

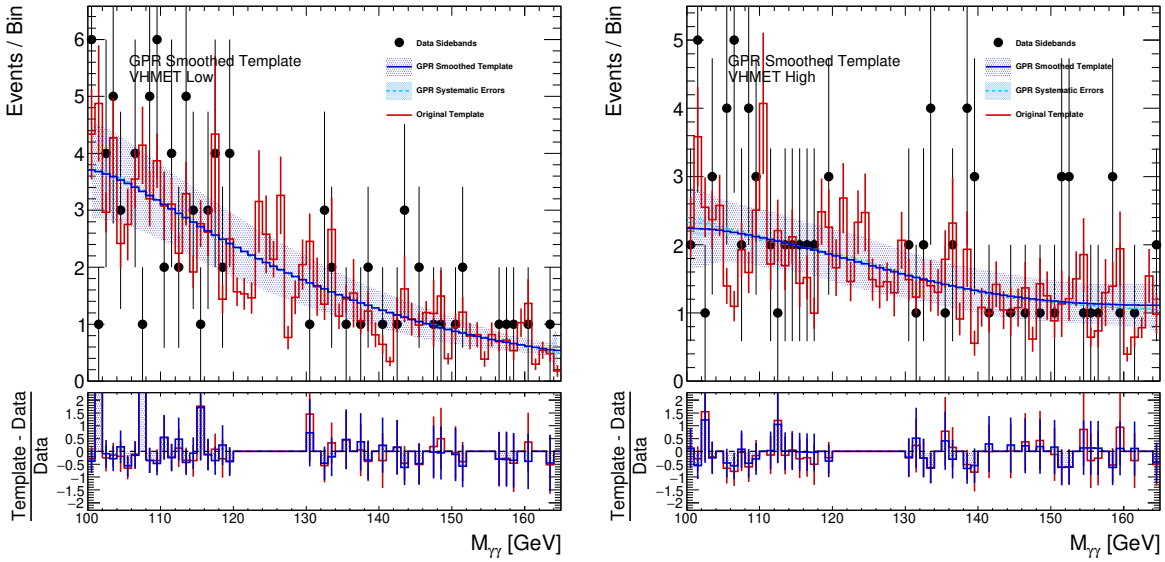


(b) VH Had Tight



(c) qqH BSM

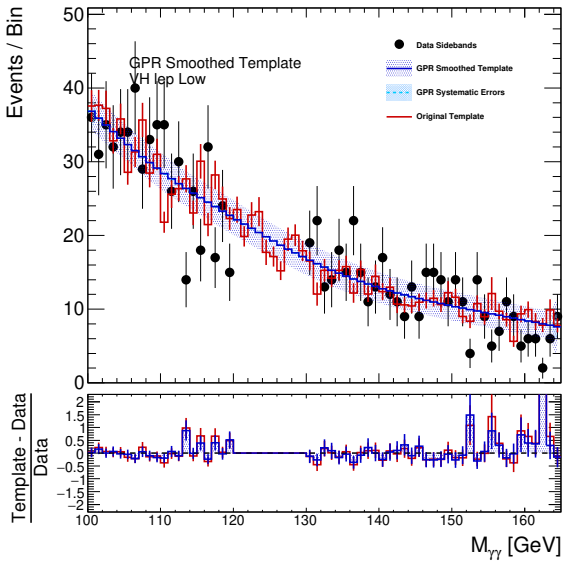
Figure C.5: The full Run 2 background templates of the hadronic VH and qqH BSM analysis categories, (a) VH Had Loose, (b) VH Had Tight, and (c) qqH BSM. The red shape shows the original background template, the blue shape shows the smoothed background template, and the black points show the data sidebands (for reference). The bottom panel shows the fractional difference between the smoothed- and un-smoothed templates and the data sidebands.



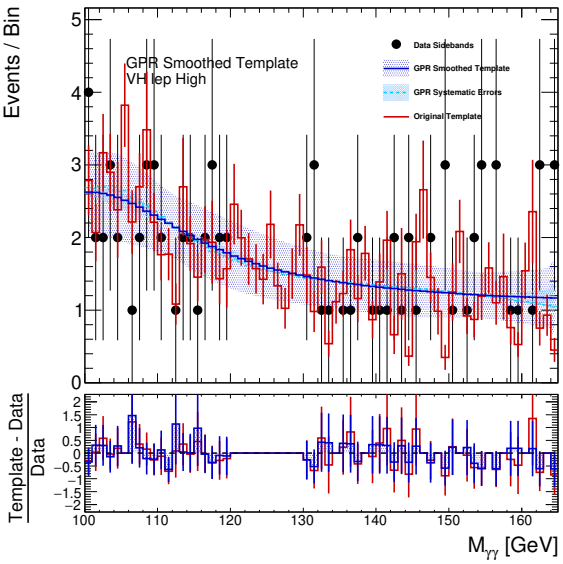
(a) VH MET Low

(b) VH MET High

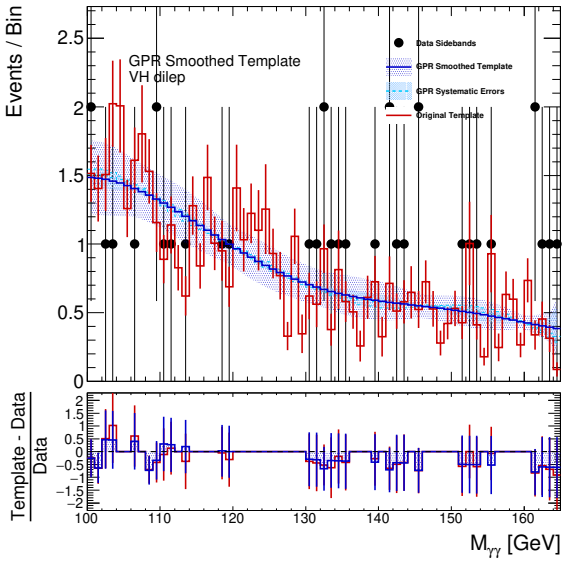
Figure C.6: The full Run 2 background templates of the hadronic VH $E_{\mathrm{T}}^{\mathrm{miss}}$ analysis categories, (a) VH MET Low and (b) VH MET High. The red shape shows the original background template, the blue shape shows the smoothed background template, and the black points show the data sidebands (for reference). The bottom panel shows the fractional difference between the smoothed- and unsmoothed templates and the data sidebands.



(a) VH Lep Low



(b) VH Lep High



(c) VH Dilep

Figure C.7: The full Run 2 background templates of the hadronic VH leptonic analysis categories, (a) VH Lep Low, (b) VH Lep High, and (c) VH Dilep. The red shape shows the original background template, the blue shape shows the smoothed background template, and the black points show the data sidebands (for reference). The bottom panel shows the fractional difference between the smoothed- and un-smoothed templates and the data sidebands.

APPENDIX D

Validation of Gaussian Process Smoothing

Extensive testing was performed of the GP smoothing technique (detailed in Chapter 8) to ensure that no bias was being introduced through the smoothing process. The tests are designed to mimic the conditions of the analysis presented in this work. Hence, the background-like analytic functions used in the tests are the same as those used in the spurious signal test (see Section 7.3), and the mass range of the templates is the same as that of the analysis ($105 < m_{\gamma\gamma} < 160$ GeV).

The first of these tests involves calculating the average bias of many GP fits to "toy" templates. The toy templates were constructed in two ways. The first is from the probability distributions of known analytic functions, but containing random fluctuations. The percent differences between the "true" shape (the analytic function) and the average of the many GP fits are presented in Figures D.1-D.7. Each figure shows the results using one of the analytic functions. The fits were also performed with different statistics, ranging from 1000 events per template to one million. The results consistently demonstrate that the fitting bias is less than about 2% for toy templates with statistics greater than 10,000 events per template. For lower-statistics templates, the bias is also confined to less than about 2% when ignoring within 5 GeV of the edge of the template. Notably, a bias at the edges of the fit range is expected given the nature of the GP fit. The points less than about one characteristic length scale away from the edge are not as highly constrained by the given data points as those in the middle of the template, and so the choice of GP mean (discussed below) begins to affect the fit in this range.

The second method of constructing toy templates for the average bias tests is using the simulated diphoton sample. A subsample consisting of randomly selected events was filled into a histogram, and a GP fit was performed to each template. The average of the many GP fits was compared to the full statistics simulation shape, which was defined as the "true" shape. The number of events ranged between 1000 and one million, again in order to check for any bias in different statistics regimes. The simulation weights were applied to each event, and the events were required to pass the diphoton preselection (defined in Section 6.1). As for the fit bias tests using the analytic function based toys, the average fits to the simulation templates do not appear to show signifi-

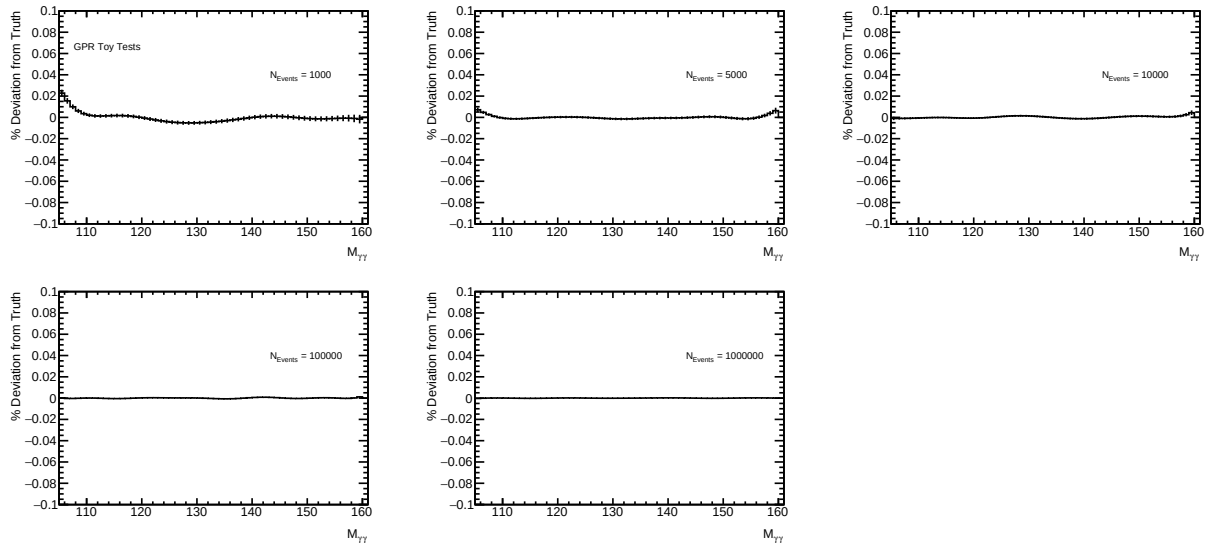


Figure D.1: The average bias of the GP fits to toy templates constructed from an analytic Power Law function. The top left subplot shows the results using templates containing 1000 events, the top middle shows the results using templates containing 5000 events, and the top right shows the results using templates containing 10,000 events. The bottom left subplot shows the results using templates containing 100,000 events, while the bottom middle shows the results using templates containing one million events.

cant biases, other than at the template edges. Some fluctuations on the order of about 1% appear for some individual bins in the templates. However, these features are consistent across the different levels of statistics tested. Therefore, these appear to be fluctuations already present in the full-statistics simulation sample, and not an artifact of the GP fit.

The second group of tests was to evaluate the difference in GP fit bias when different functions were used as the GP mean. The Exponential function, Linear function, and a flat line were all tested. These tests were performed using toy templates constructed from Power Law (Fig. D.9), ExpPoly2 (Fig. D.10), and Bernstein 5 (Fig. D.11) functions. Again, different levels of statistics in the toy templates were tested. Overall, the choice of GP mean does not seem to affect the GP fit bias significantly, other than for templates with less than 1000 events. In the lower statistics templates, most of the differences in fitting bias are again confined to the extremities of the template range, as expected. Therefore, the choice of mean does not appear to introduce a significant bias in the fit result.

The third set of tests performed evaluated the impact of the GP smoothing technique on the measured spurious signal. The tests again involve performing the GP smoothing on a large number of toy templates. The spurious signal test (a signal plus background fit) is performed on both the un-smoothed and smoothed templates. For these tests, a simplified Crystal Ball shape with

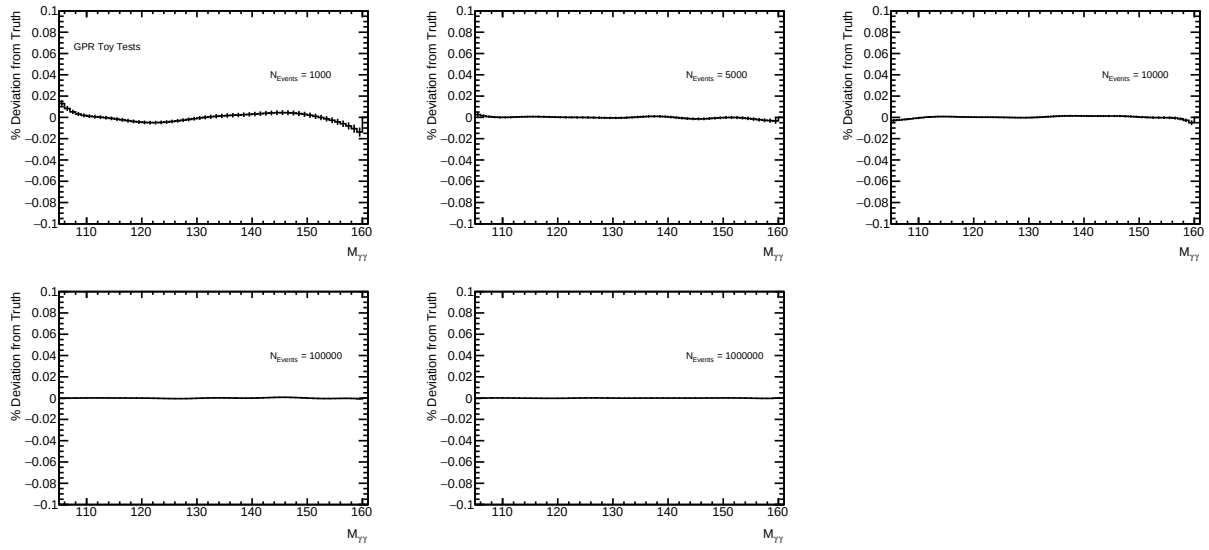


Figure D.2: The average bias of the GP fits to toy templates constructed from an analytic Exponential function. The top left subplot shows the results using templates containing 1000 events, the top middle shows the results using templates containing 5000 events, and the top right shows the results using templates containing 10,000 events. The bottom left subplot shows the results using templates containing 100,000 events, while the bottom middle shows the results using templates containing one million events.

a width of 2 GeV is used as the signal shape, and the full set of analytic functions described in Section 7.3 (Power Law, Exponential, ExpPoly2, ExpPoly3, Bern3, Bern4, and Bern5) are tested as background functions. The many spurious signal measurements are compiled into distributions, such as those shown in Figures D.12-D.17. The distributions are expected to be roughly Gaussian in nature.

The results shown in Figures D.12-D.17 are obtained using toy templates constructed from simulated diphoton events. A set number of events passing the diphoton selection was randomly chosen to construct each template. Each figure shows the results for a different number of events in each template, ranging from 500 to one million. Each subplot within the figures shows the results using the different choices of background functional forms when performing the signal and background fit. The results may be compared between the smoothed (blue) and un-smoothed (red) distributions. For the smoothed templates, the distribution of measured spurious signal values is narrower, especially for those templates with fewer events. This narrowing reflects the reduction of statistical fluctuations by the smoothing process. Additionally, the mean values of both the smoothed and the un-smoothed distributions are compatible. The average value of the un-smoothed distributions may be taken as the "true" spurious signal for a given choice of analytic background function. The compatibility of the smoothed mean values indicates that the GP smoothing does

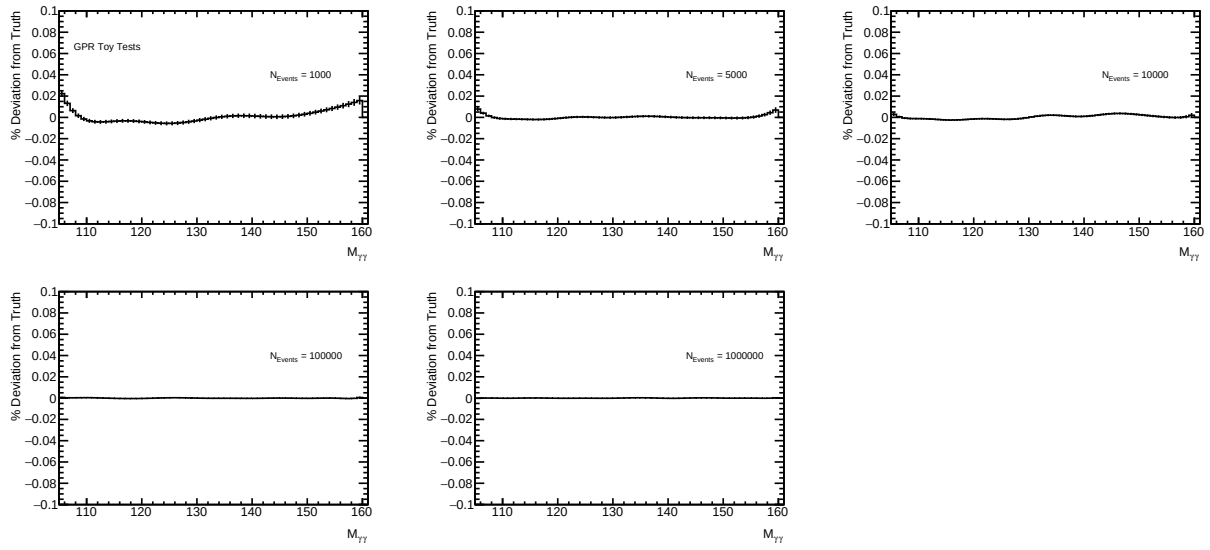


Figure D.3: The average bias of the GP fits to toy templates constructed from an analytic ExpPoly2 function. The top left subplot shows the results using templates containing 1000 events, the top middle shows the results using templates containing 5000 events, and the top right shows the results using templates containing 10,000 events. The bottom left subplot shows the results using templates containing 100,000 events, while the bottom middle shows the results using templates containing one million events.

not appear to strongly bias the spurious signal measurement; it simply increases the measurement precision. Additionally, the compatibility may be seen both in cases where the analytic background function chosen can adequately model the background (cases where the average spurious signal is zero) and in cases where it does produce nontrivial mismodeling (cases where the average spurious signal is non-zero). The high-statistics simulation shape contains features which are not easily modeled by the Power Law and Exponential functions, and so these functions in particular are expected to result in a significant spurious signal.

The toy tests measuring the spurious signal are more compactly summarized in Figure D.18. The results across the range of statistics are shown in a single plot by effectively "rotating" the distributions on their side. The mean values of the scaled spurious signal distributions (based on Gaussian fits the distributions) from both the raw and smoothed templates are plotted on the y-axis as a function of template statistics in the x-axis. The shaded areas indicate the widths of the distributions (defined as the width σ of the Gaussian fit to the distribution). Again, the blue shape shows the results from the smoothed templates, while the red shows those from the un-smoothed templates. The spurious signal has been observed to scale linearly with the number of background events. Therefore, by scaling the spurious signal distributions by the number of background events, the mean values of the distributions will converge to a constant value. The

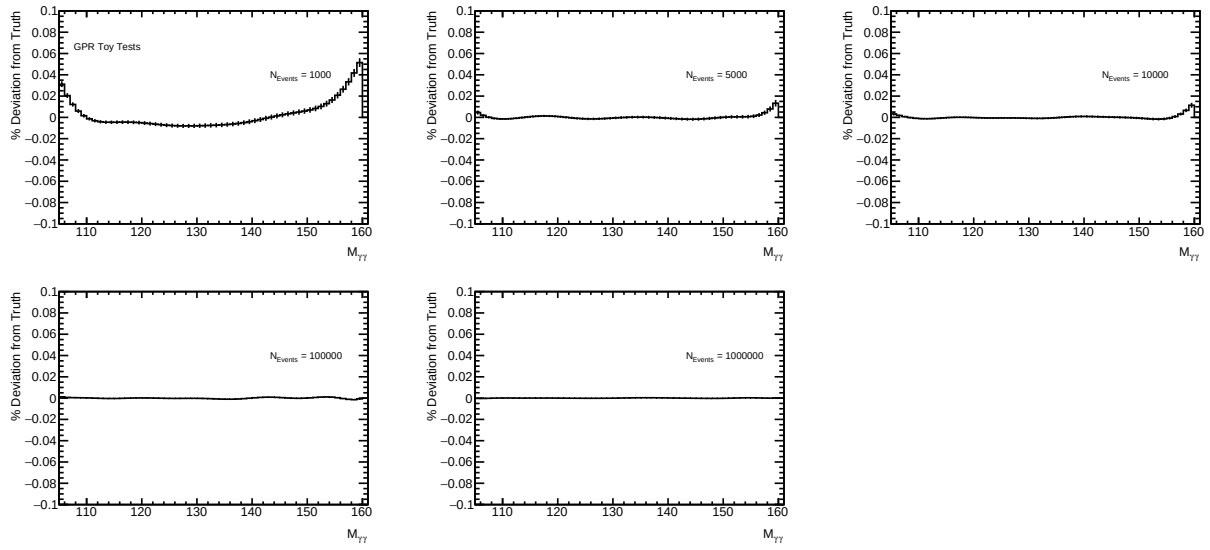


Figure D.4: The average bias of the GP fits to toy templates constructed from an analytic ExpPoly3 function. The top left subplot shows the results using templates containing 1000 events, the top middle shows the results using templates containing 5000 events, and the top right shows the results using templates containing 10,000 events. The bottom left subplot shows the results using templates containing 100,000 events, while the bottom middle shows the results using templates containing one million events.

level of "flatness" of the subplots in Figure D.18 then indicates whether any significant bias in the spurious signal measurement is present (a flat line indicates no bias).

An examination of Figure D.18 indicates that the GP smoothing results in the same spurious signal on average as when measured using the raw templates. The average spurious signal, scaled by the number of background events, measured using both the smoothed and raw templates is a constant value across the range of template statistics considered. Overall, no significant bias is observed as a result of the smoothing. Additionally, the shaded area of the smoothed templates is narrower than that of the raw templates, indicating that the distribution of spurious signal measured is narrower when using the smoothing technique.

These spurious signal distribution tests were also be conducted using toy templates constructed from analytic functions. The individual distributions of spurious signal measurements are not shown; instead, only the summary plots are presented in Figures D.20-D.25. These plots are similar to Figure D.18, in that they present the scaled spurious signal in the y-axis as a function of the template statistics in the x-axis. As for the tests using simulated events, the average spurious signal values measured using the smoothed and un-smoothed templates are compatible. The widths of the smoothed distributions are also narrower than those of the un-smoothed, except for very high statistics templates. Again, the GP smoothing does not appear to significantly shift the mean

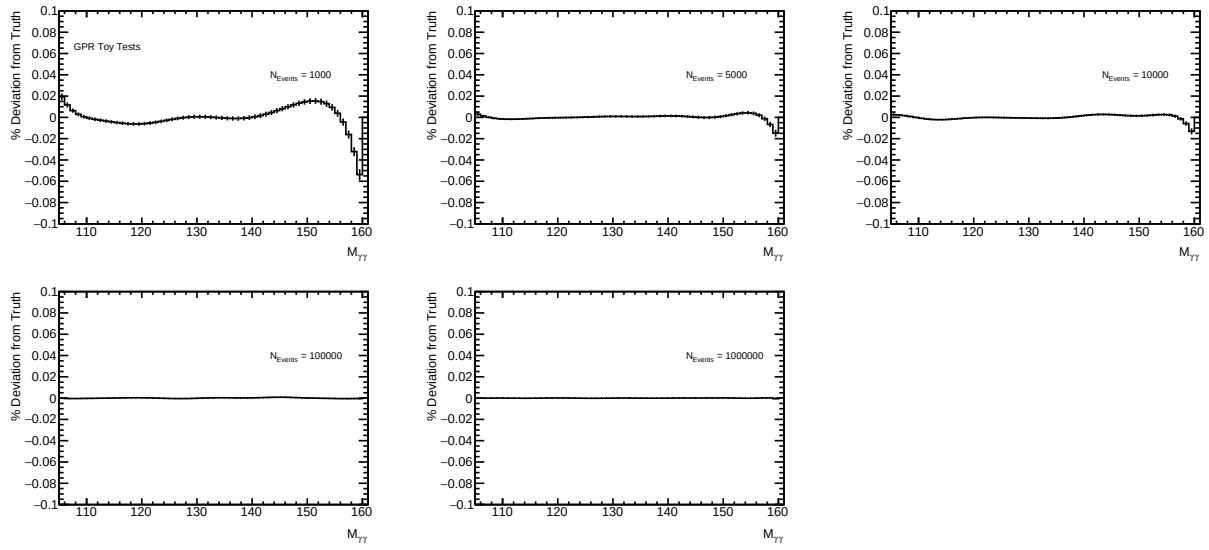


Figure D.5: The average bias of the GP fits to toy templates constructed from an analytic Bernstein 3 function. The top left subplot shows the results using templates containing 1000 events, the top middle shows the results using templates containing 5000 events, and the top right shows the results using templates containing 10,000 events. The bottom left subplot shows the results using templates containing 100,000 events, while the bottom middle shows the results using templates containing one million events.

spurious signal, even if it is non-zero. This implies that the GP smoothing successfully reduces statistical fluctuations, without affecting the real underlying mismodeling which the spurious signal test is designed to probe.

The combination of these validation tests, along with the measured impact to the spurious signal uncertainty in the analysis presented here in Chapter 8 and that presented in Ref. [108], demonstrate the power of using the Gaussian Process approach. By reducing problematic statistical fluctuations, the spurious signal systematic may be reduced enough to significantly increase the precision of the analysis in the future.

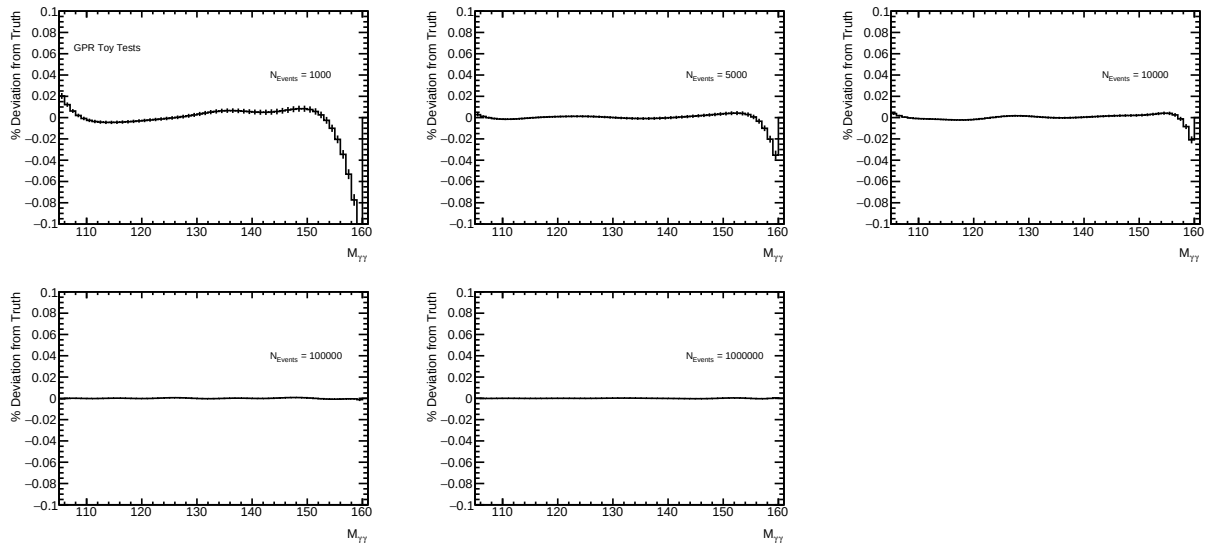


Figure D.6: The average bias of the GP fits to toy templates constructed from an analytic Bernstein 4 function. The top left subplot shows the results using templates containing 1000 events, the top middle shows the results using templates containing 5000 events, and the top right shows the results using templates containing 10,000 events. The bottom left subplot shows the results using templates containing 100,000 events, while the bottom middle shows the results using templates containing one million events.

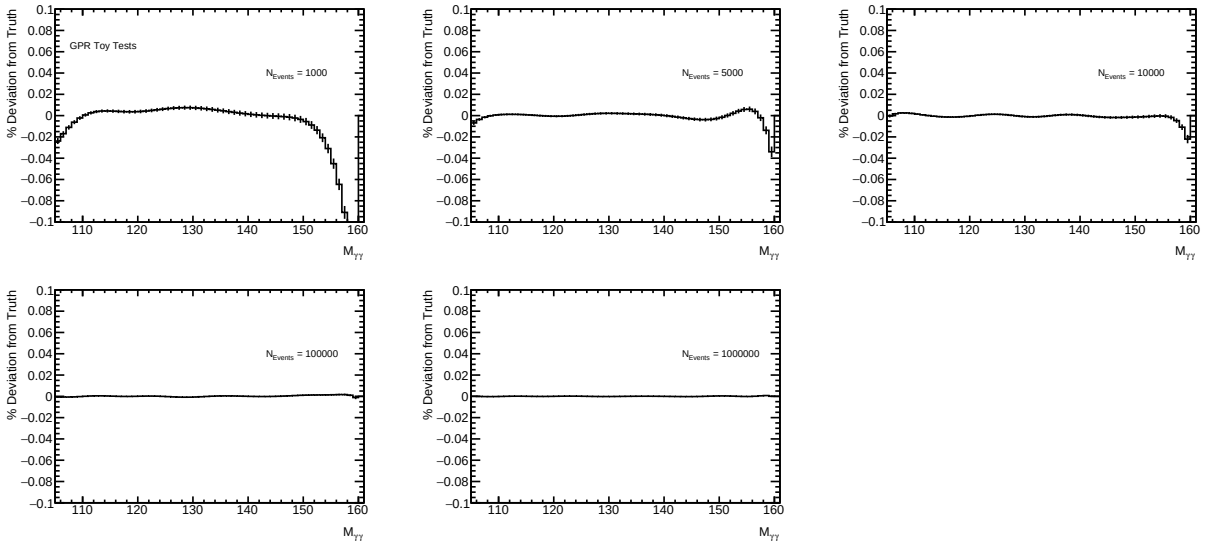


Figure D.7: The average bias of the GP fits to toy templates constructed from an analytic Bernstein 5 function. The top left subplot shows the results using templates containing 1000 events, the top middle shows the results using templates containing 5000 events, and the top right shows the results using templates containing 10,000 events. The bottom left subplot shows the results using templates containing 100,000 events, while the bottom middle shows the results using templates containing one million events.

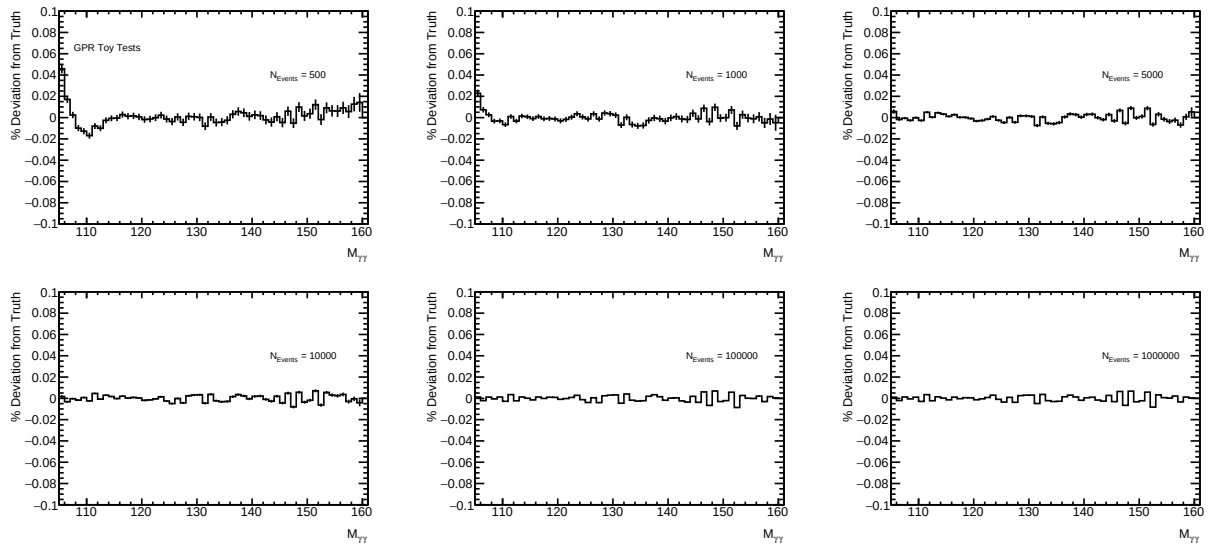


Figure D.8: The average bias of the GP fits to toy templates constructed from randomly selected simulated diphoton events. The top left subplot shows the results using templates containing 500 events, the top middle shows the results using templates containing 1000 events, and the top right shows the results using templates containing 5000 events. The bottom left subplot shows the results using templates containing 10,000 events, the bottom middle shows the results using 100,000 events, and the bottom right shows the results using templates containing one million events.

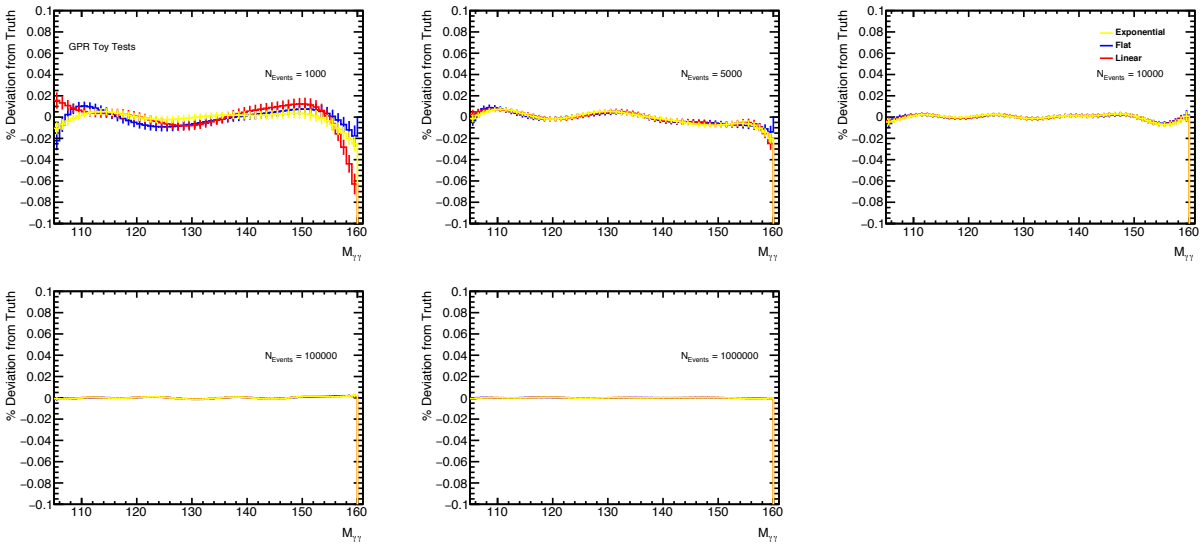


Figure D.9: Comparisons of the average bias induced by the choice of GP mean when fitting toy templates constructed from the analytic Power Law function. The yellow shape shows the results using the default Exponential mean, the blue shape shows the result using a flat line as the mean, and the red shape shows the result using a linear fit as the mean. The top left subplot shows the results using templates containing 1000 events, the top middle shows the results using templates containing 5000 events, and the top right shows the results using templates containing 10,000 events. The bottom left subplot shows the results using templates containing 100,000 events, while the bottom middle shows the results using templates containing one million events.

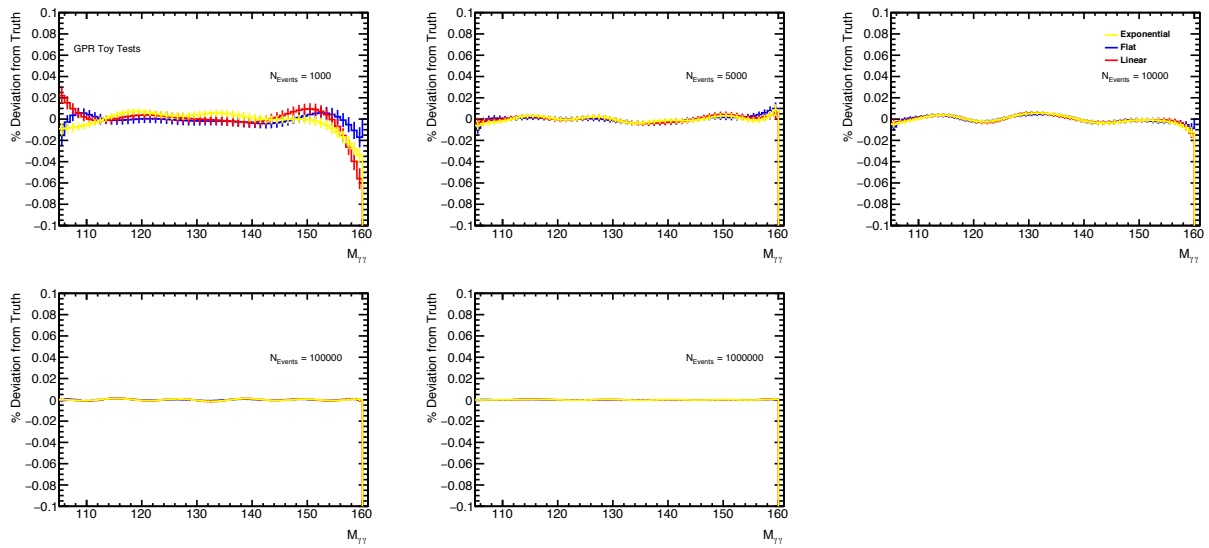


Figure D.10: Comparisons of the average bias induced by the choice of GP mean when fitting toy templates constructed from the analytic ExpPoly2 function. The yellow shape shows the results using the default Exponential mean, the blue shape shows the result using a flat line as the mean, and the red shape shows the result using a linear fit as the mean. The top left subplot shows the results using templates containing 1000 events, the top middle shows the results using templates containing 5000 events, and the top right shows the results using templates containing 10,000 events. The bottom left subplot shows the results using templates containing 100,000 events, while the bottom middle shows the results using templates containing one million events.

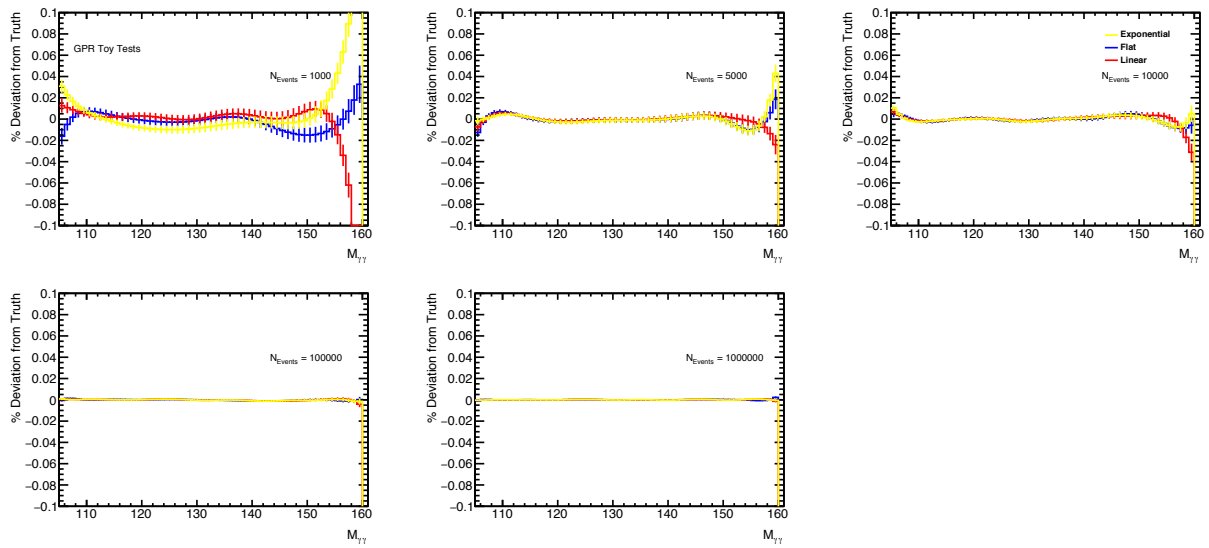


Figure D.11: Comparisons of the average bias induced by the choice of GP mean when fitting toy templates constructed from the analytic Bernstein 5 function. The yellow shape shows the results using the default Exponential mean, the blue shape shows the result using a flat line as the mean, and the red shape shows the result using a linear fit as the mean. The top left subplot shows the results using templates containing 1000 events, the top middle shows the results using templates containing 5000 events, and the top right shows the results using templates containing 10,000 events. The bottom left subplot shows the results using templates containing 100,000 events, while the bottom middle shows the results using templates containing one million events.

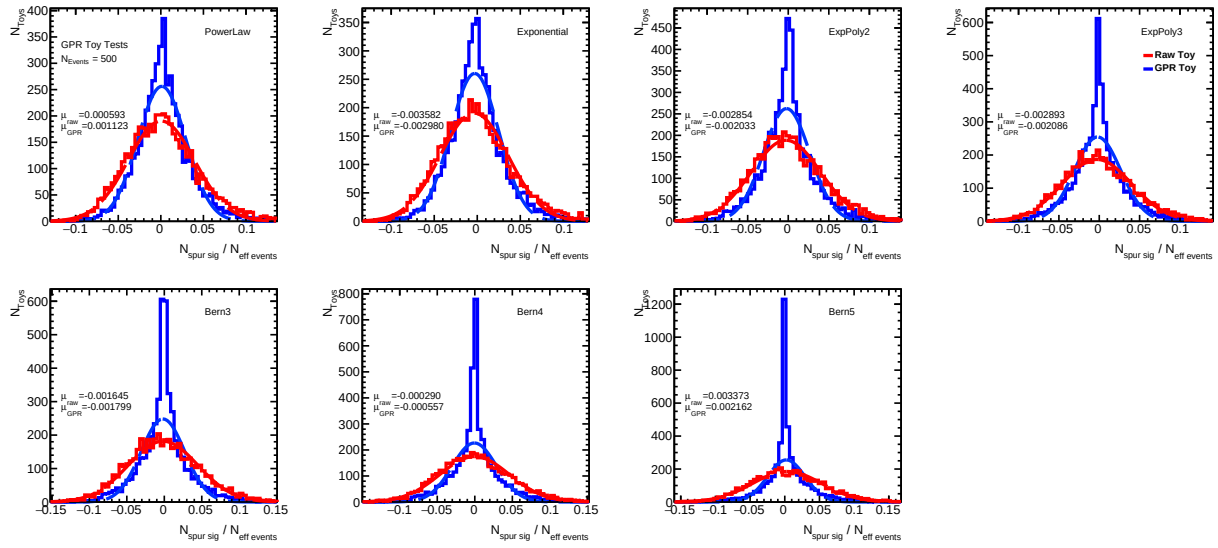


Figure D.12: The distribution of spurious signal values obtained when performing the spurious signal test on toy templates constructed from the continuum diphoton simulation sample. Each toy contains 500 randomly-chosen events, which are required to pass the diphoton selection, detailed in Chapter 6. The x-axis is normalized to the number of events contained in the template. The subplots in the top row show the results when using the (left to right) Power Law, Exponential, ExpPoly2, and ExpPoly3 functions as the background function in the spurious signal test fit. The top row shows the results obtained using the (left to right) Bernstein 3, 4, and 5 functions in the fit. The blue shape shows the distribution of spurious signal measurements when testing the GP smoothed toy templates, while the red shows the measurements from the un-smoothed toys.

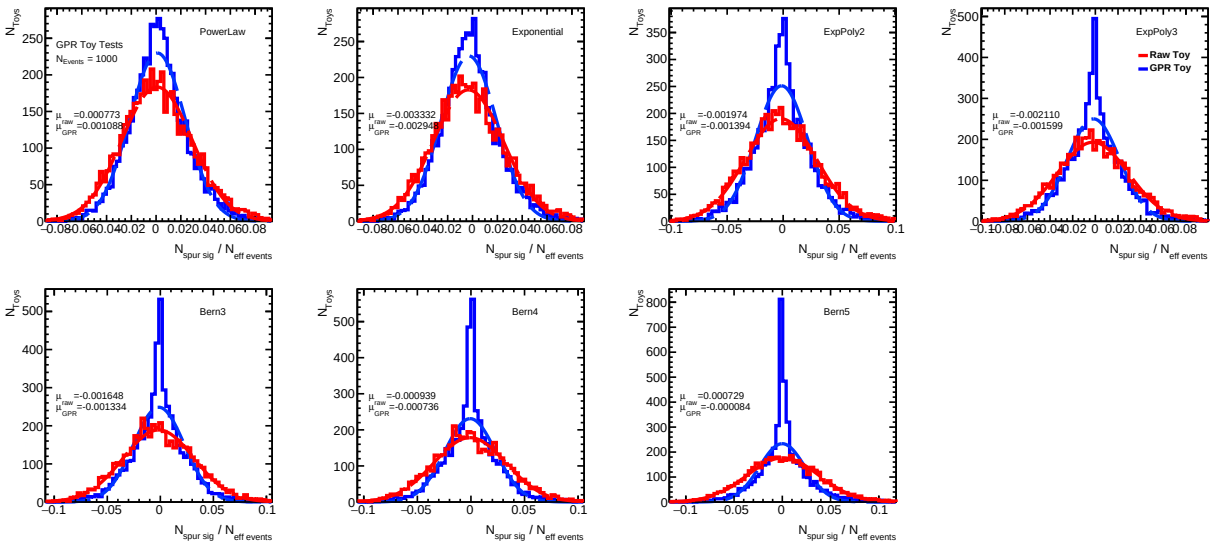


Figure D.13: The distribution of spurious signal values obtained when performing the spurious signal test on toy templates constructed from the continuum diphoton simulation sample. Each toy contains 1000 randomly-chosen events, which are required to pass the diphoton selection, detailed in Chapter 6. The x-axis is normalized to the number of events contained in the template. The subplots in the top row show the results when using the (left to right) Power Law, Exponential, ExpPoly2, and ExpPoly3 functions as the background function in the spurious signal test fit. The top row shows the results obtained using the (left to right) Bernstein 3, 4, and 5 functions in the fit. The blue shape shows the distribution of spurious signal measurements when testing the GP smoothed toy templates, while ths red shows the measurements from the un-smoothed toys.

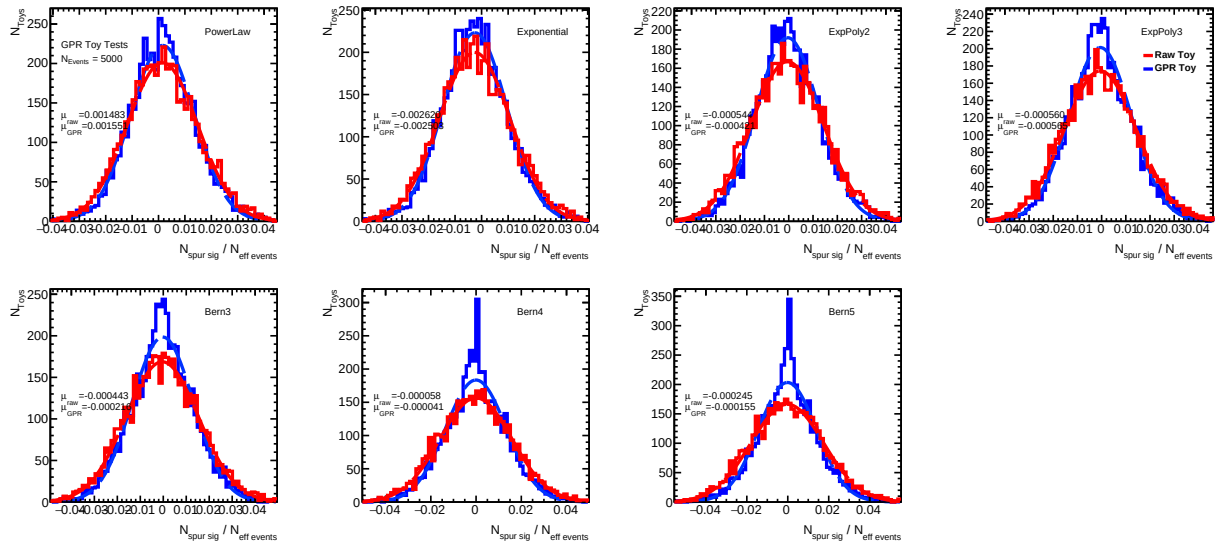


Figure D.14: The distribution of spurious signal values obtained when performing the spurious signal test on toy templates constructed from the continuum diphoton simulation sample. Each toy contains 5000 randomly-chosen events, which are required to pass the diphoton selection, detailed in Chapter 6. The x-axis is normalized to the number of events contained in the template. The subplots in the top row show the results when using the (left to right) Power Law, Exponential, ExpPoly2, and ExpPoly3 functions as the background function in the spurious signal test fit. The top row shows the results obtained using the (left to right) Bernstein 3, 4, and 5 functions in the fit. The blue shape shows the distribution of spurious signal measurements when testing the GP smoothed toy templates, while ths red shows the measurements from the un-smoothed toys.

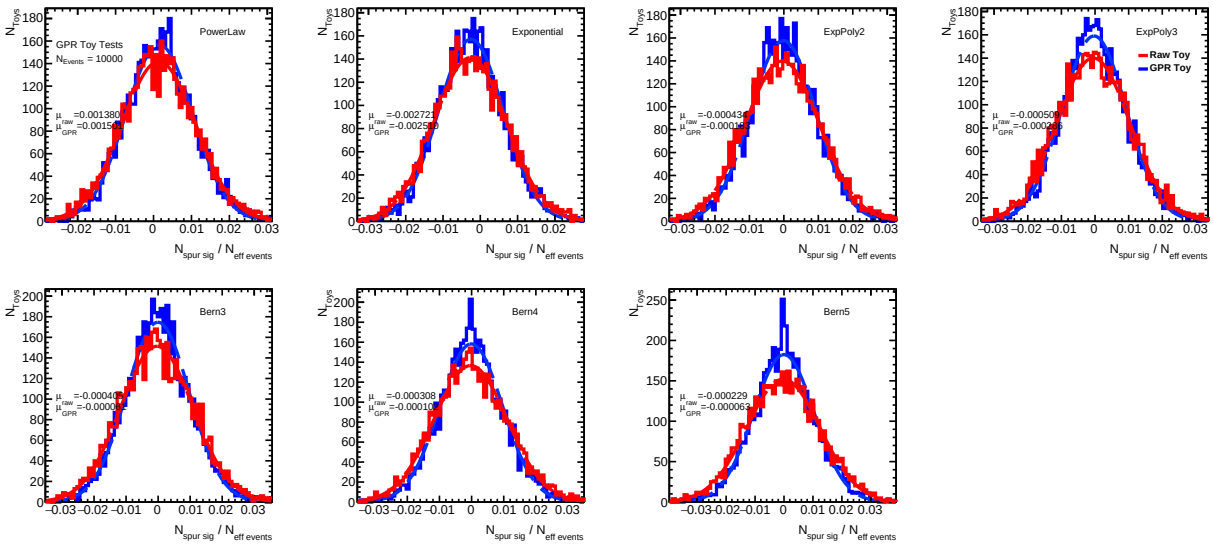


Figure D.15: The distribution of spurious signal values obtained when performing the spurious signal test on toy templates constructed from the continuum diphoton simulation sample. Each toy contains 10,000 randomly-chosen events, which are required to pass the diphoton selection, detailed in Chapter 6. The x-axis is normalized to the number of events contained in the template. The subplots in the top row show the results when using the (left to right) Power Law, Exponential, ExpPoly2, and ExpPoly3 functions as the background function in the spurious signal test fit. The top row shows the results obtained using the (left to right) Bernstein 3, 4, and 5 functions in the fit. The blue shape shows the distribution of spurious signal measurements when testing the GP smoothed toy templates, while ths red shows the measurements from the un-smoothed toys.

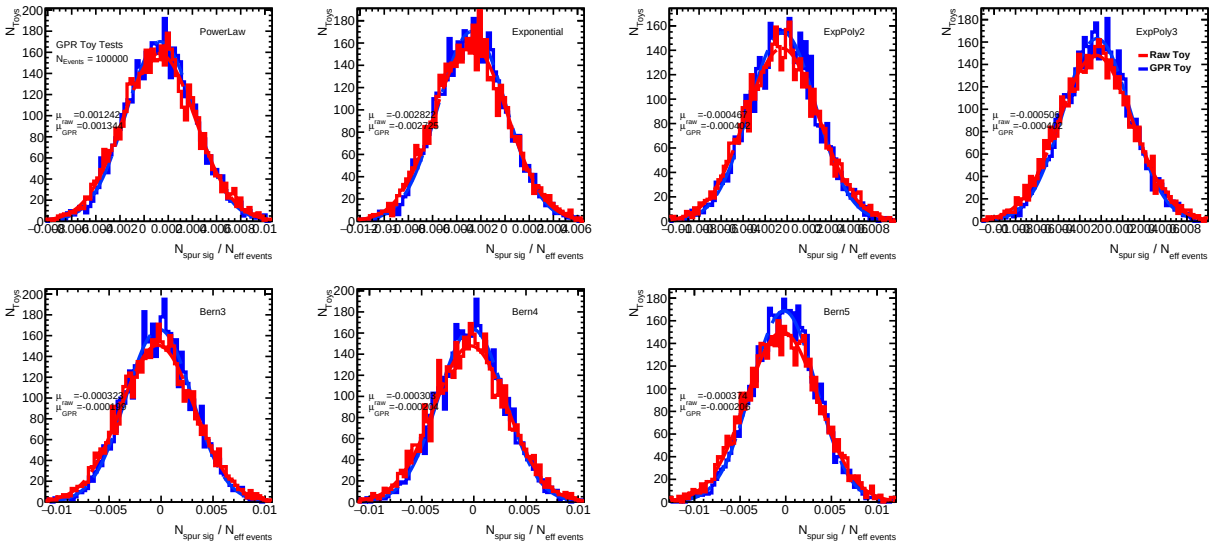


Figure D.16: The distribution of spurious signal values obtained when performing the spurious signal test on toy templates constructed from the continuum diphoton simulation sample. Each toy contains 100,000 randomly-chosen events, which are required to pass the diphoton selection, detailed in Chapter 6. The x-axis is normalized to the number of events contained in the template. The subplots in the top row show the results when using the (left to right) Power Law, Exponential, ExpPoly2, and ExpPoly3 functions as the background function in the spurious signal test fit. The top row shows the results obtained using the (left to right) Bernstein 3, 4, and 5 functions in the fit. The blue shape shows the distribution of spurious signal measurements when testing the GP smoothed toy templates, while ths red shows the measurements from the un-smoothed toys.

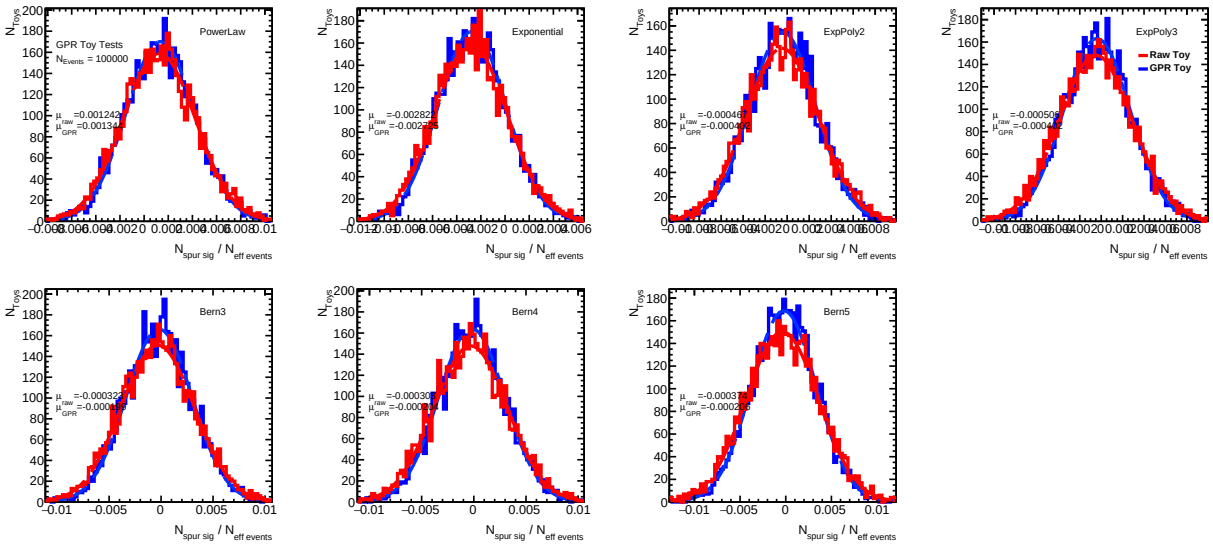


Figure D.17: The distribution of spurious signal values obtained when performing the spurious signal test on toy templates constructed from the continuum diphoton simulation sample. Each toy contains one million randomly-chosen events, which are required to pass the diphoton selection, detailed in Chapter 6. The x-axis is normalized to the number of events contained in the template. The subplots in the top row show the results when using the (left to right) Power Law, Exponential, ExpPoly2, and ExpPoly3 functions as the background function in the spurious signal test fit. The top row shows the results obtained using the (left to right) Bernstein 3, 4, and 5 functions in the fit. The blue shape shows the distribution of spurious signal measurements when testing the GP smoothed toy templates, while ths red shows the measurements from the un-smoothed toys.

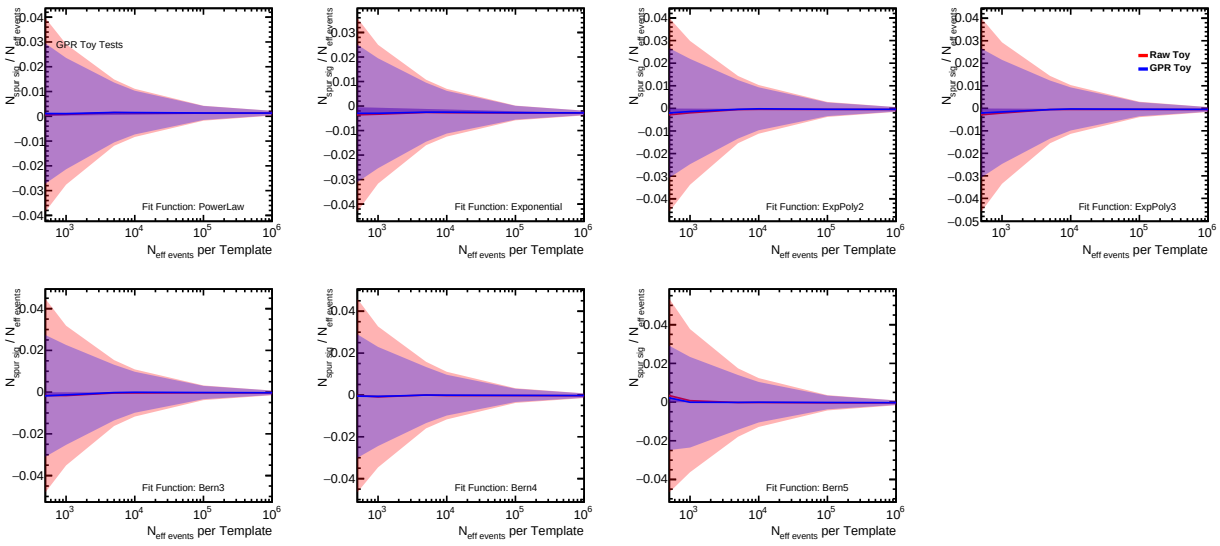


Figure D.18: Summary of the mean value and width of the spurious signal distributions measured from toy templates constructed with randomly selected simulated diphoton events. The x-axis indicates the number of events in the background template, while the y-axis indicates the measured spurious signal, scaled by the number of events in the background template. The top row shows the results obtained when performing the signal and background fit using the (left to right) Power Law, Exponential, ExpPoly2, and ExpPoly3 functions as the background shape in the fit. The bottom row shows the results when using the (left to right) Bernstein 3, 4 and 5 functions as the background shape. The blue shape shows the results obtained using the GP smoothing technique, while the red shape shows the results obtained without the technique.

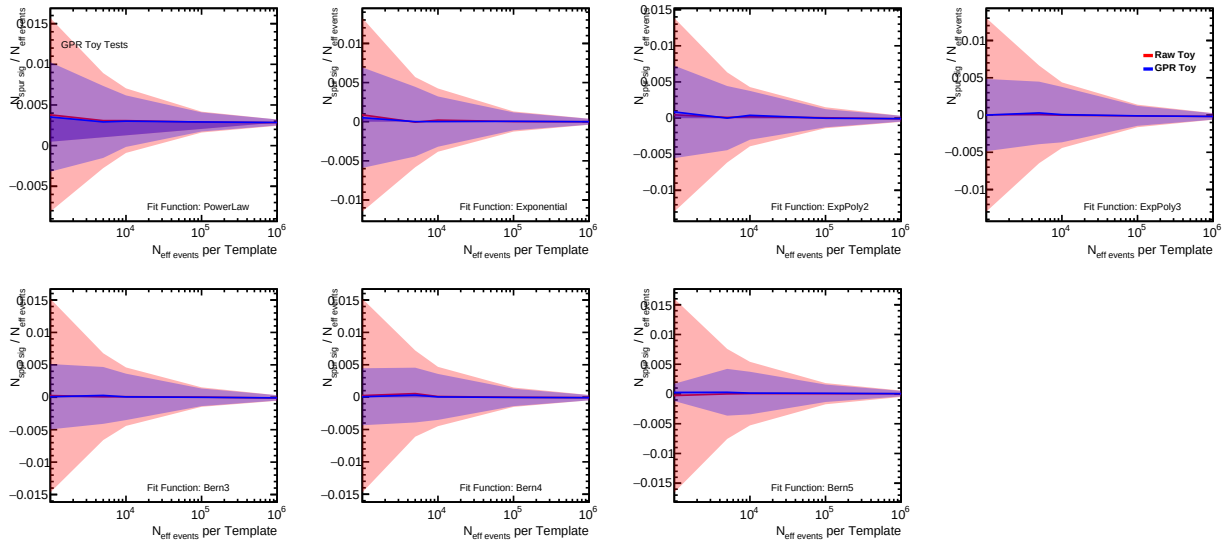


Figure D.19: Summary of the mean value and width of the spurious signal distributions measured from toy templates constructed with the analytic Exponential function. The x-axis indicates the number of events in the background template, while the y-axis indicates the measured spurious signal, scaled by the number of events in the background template. The top row shows the results obtained when performing the signal and background fit using the (left to right) Power Law, Exponential, ExpPoly2, and ExpPoly3 functions as the background shape in the fit. The bottom row shows the results when using the (left to right) Bernstein 3, 4 and 5 functions as the background shape. The blue shape shows the results obtained using the GP smoothing technique, while the red shape shows the results obtained without the technique.

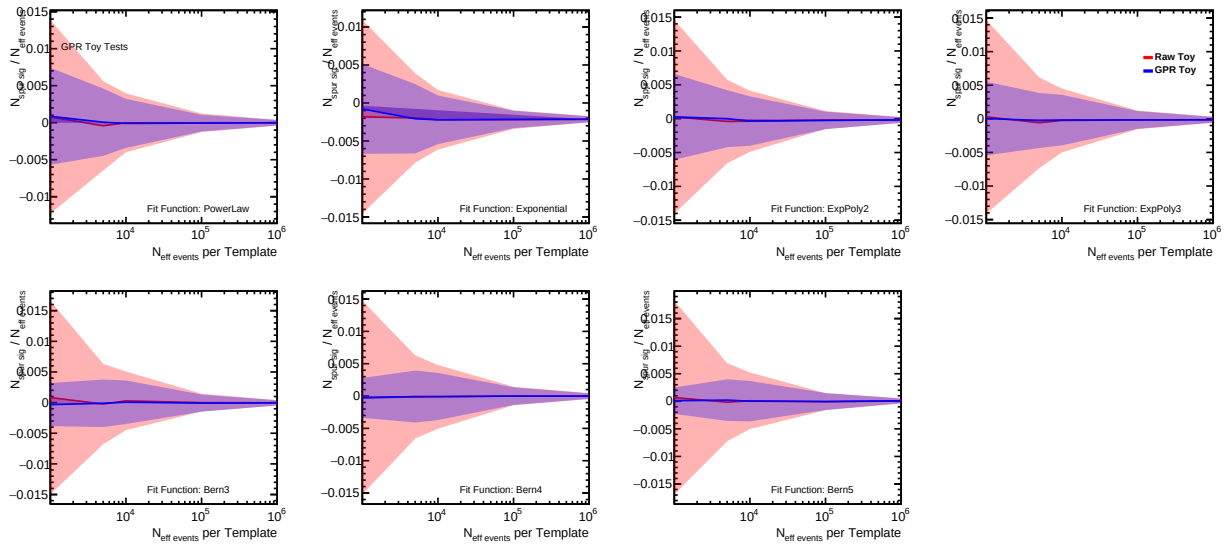


Figure D.20: Summary of the mean value and width of the spurious signal distributions measured from toy templates constructed with the analytic Power Law function. The x-axis indicates the number of events in the background template, while the y-axis indicates the measured spurious signal, scaled by the number of events in the background template. The top row shows the results obtained when performing the signal and background fit using the (left to right) Power Law, Exponential, ExpPoly2, and ExpPoly3 functions as the background shape in the fit. The bottom row shows the results when using the (left to right) Bernstein 3, 4 and 5 functions as the background shape. The blue shape shows the results obtained using the GP smoothing technique, while the red shape shows the results obtained without the technique.

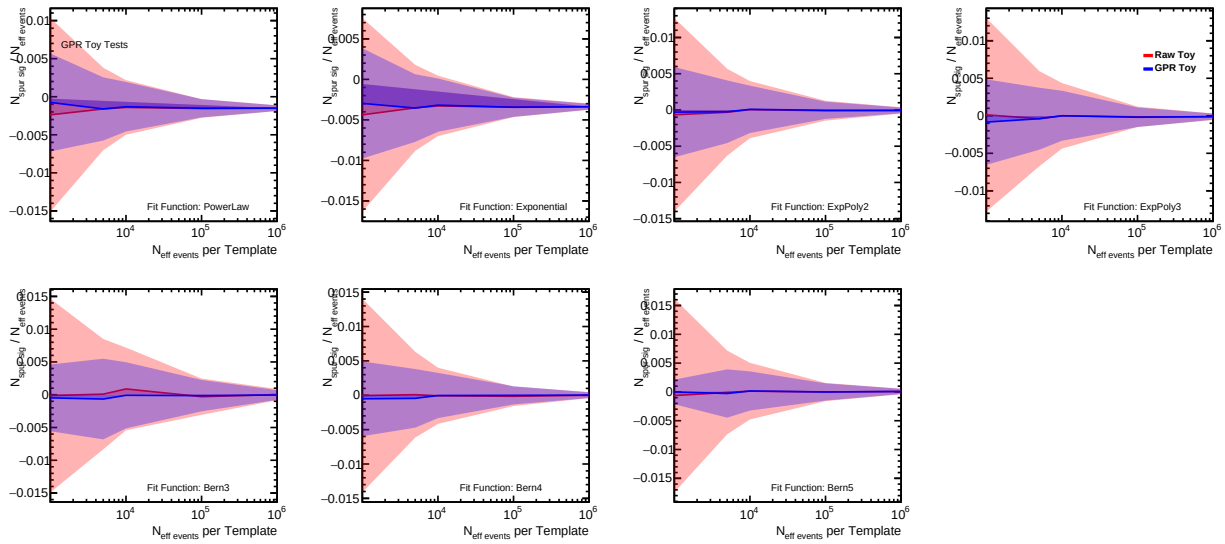


Figure D.21: Summary of the mean value and width of the spurious signal distributions measured from toy templates constructed with the analytic ExpPoly2 function. The x-axis indicates the number of events in the background template, while the y-axis indicates the measured spurious signal, scaled by the number of events in the background template. The top row shows the results obtained when performing the signal and background fit using the (left to right) Power Law, Exponential, ExpPoly2, and ExpPoly3 functions as the background shape in the fit. The bottom row shows the results when using the (left to right) Bernstein 3, 4 and 5 functions as the background shape. The blue shape shows the results obtained using the GP smoothing technique, while the red shape shows the results obtained without the technique.

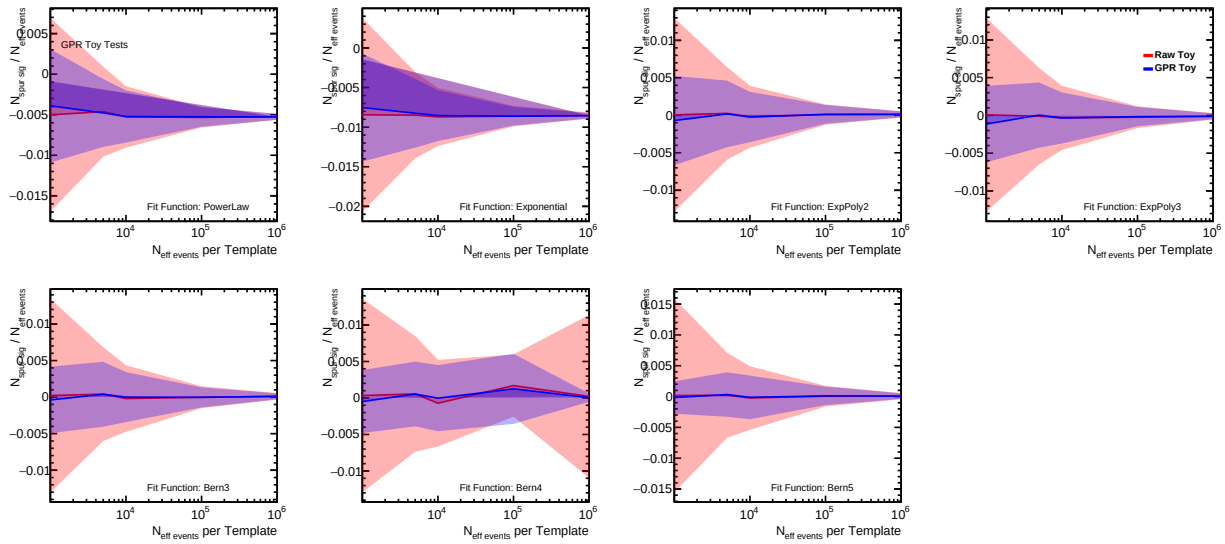


Figure D.22: Summary of the mean value and width of the spurious signal distributions measured from toy templates constructed with the analytic ExpPoly3 function. The x-axis indicates the number of events in the background template, while the y-axis indicates the measured spurious signal, scaled by the number of events in the background template. The top row shows the results obtained when performing the signal and background fit using the (left to right) Power Law, Exponential, ExpPoly2, and ExpPoly3 functions as the background shape in the fit. The bottom row shows the results when using the (left to right) Bernstein 3, 4 and 5 functions as the background shape. The blue shape shows the results obtained using the GP smoothing technique, while the red shape shows the results obtained without the technique.

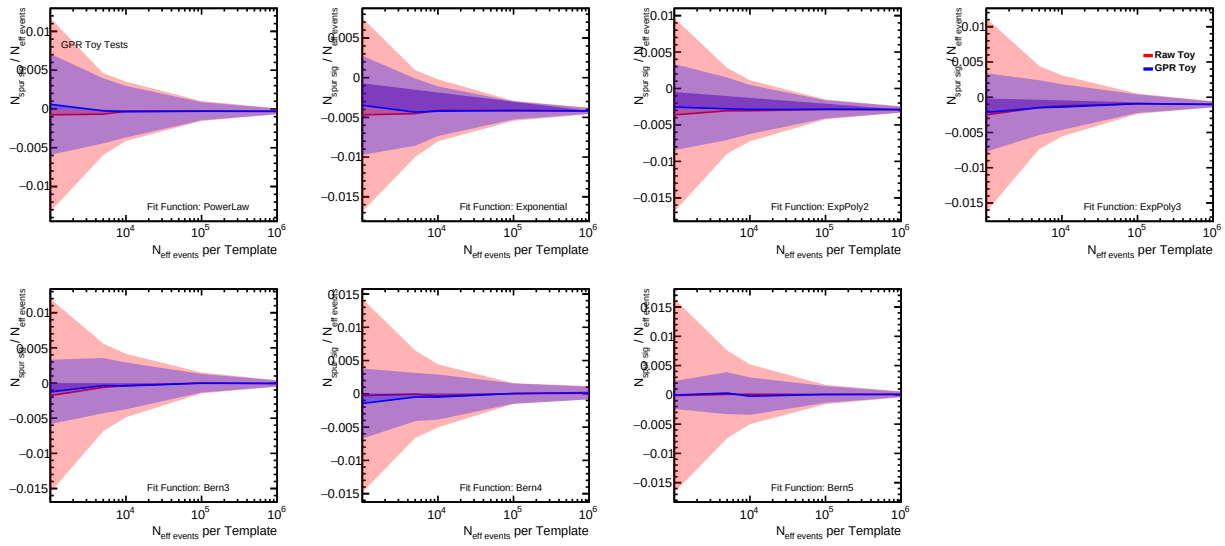


Figure D.23: Summary of the mean value and width of the spurious signal distributions measured from toy templates constructed with the analytic Bernstein 3 function. The x-axis indicates the number of events in the background template, while the y-axis indicates the measured spurious signal, scaled by the number of events in the background template. The top row shows the results obtained when performing the signal and background fit using the (left to right) Power Law, Exponential, ExpPoly2, and ExpPoly3 functions as the background shape in the fit. The bottom row shows the results when using the (left to right) Bernstein 3, 4 and 5 functions as the background shape. The blue shape shows the results obtained using the GP smoothing technique, while the red shape shows the results obtained without the technique.

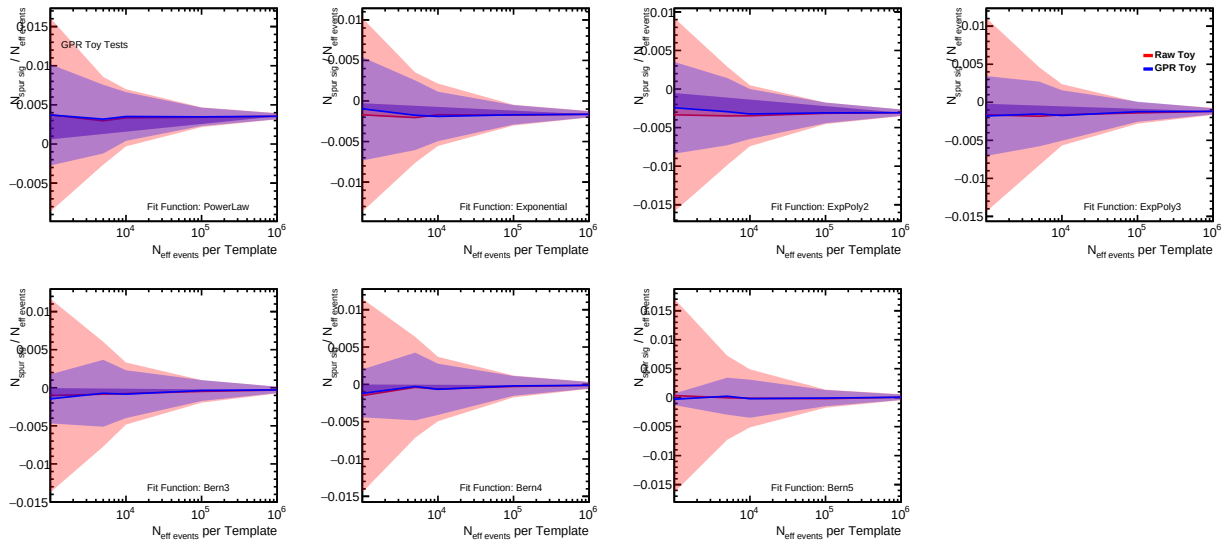


Figure D.24: Summary of the mean value and width of the spurious signal distributions measured from toy templates constructed with the analytic Bernstein 4 function. The x-axis indicates the number of events in the background template, while the y-axis indicates the measured spurious signal, scaled by the number of events in the background template. The top row shows the results obtained when performing the signal and background fit using the (left to right) Power Law, Exponential, ExpPoly2, and ExpPoly3 functions as the background shape in the fit. The bottom row shows the results when using the (left to right) Bernstein 3, 4 and 5 functions as the background shape. The blue shape shows the results obtained using the GP smoothing technique, while the red shape shows the results obtained without the technique.

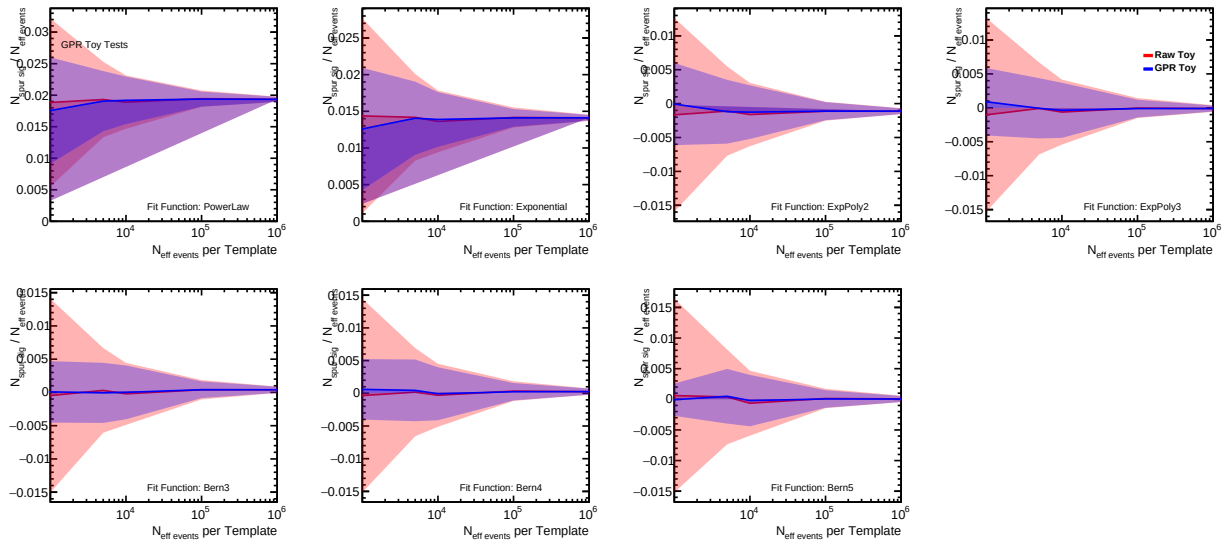


Figure D.25: Summary of the mean value and width of the spurious signal distributions measured from toy templates constructed with the analytic Bernstein 5 function. The x-axis indicates the number of events in the background template, while the y-axis indicates the measured spurious signal, scaled by the number of events in the background template. The top row shows the results obtained when performing the signal and background fit using the (left to right) Power Law, Exponential, ExpPoly2, and ExpPoly3 functions as the background shape in the fit. The bottom row shows the results when using the (left to right) Bernstein 3, 4 and 5 functions as the background shape. The blue shape shows the results obtained using the GP smoothing technique, while the red shape shows the results obtained without the technique.

APPENDIX E

Nuisance Parameter Rankings

Some insight into impact of various systematic uncertainties on the measured result can be obtained by examining the Nuisance Parameter (NP) rankings of the analysis. The rankings of the largest NPs affecting the observed results in the non-$t\bar{t}H$ production modes are presented in Figure E.1.

One important point relating to the nuisance parameter plots is that one systematic uncertainty source does not necessarily correspond directly to one nuisance parameter. Many systematic uncertainties are comprised of multiple nuisance parameters, and so estimating the "total" effect of one systematic uncertainty source is non-trivial from the NP ranking plots. For example, the background modeling ("spurious signal") uncertainty (discussed in Section 7.3) encompasses one nuisance parameter per analysis category.

Additionally, the NP plots provide information about the nuisance parameters both "pre-fit" and "post-fit." Typically, the initial values of the nuisance parameters are calculated independently of the analysis fitting framework; these methods are described in Chapter 10. These are the "pre-fit" values. The "post-fit" values are those obtained in the post-fit Asimov. The NPs are allowed to float in order to increase the agreement with the observed data. Details on the construction of the post-fit Asimov and resulting NPs are provided in Chapter 9.

The NP ranking plots also provide a high density of information for a single plot. The upper x-axis shows the relative impact of varying a single NP on the observed signal strength. The relative impact is shown as the difference in the observed signal strength from the nominal signal strength, $\Delta\hat{\mu}$, divided by the nominal signal strength, $\hat{\mu}$. The impacts of both the variations of the pre-fit (the red shaded areas) and of the post-fit (the blue shaded areas) values of the NPs are shown.

The lower x-axis shows the pulls on the NPs. The points show the difference between the central value of the post-fit NP and the initial (pre-fit) value of the NP, scaled by the error on the NP. The red lines show the magnitude of the pre-fit errors on the NPs, while the black lines show the magnitude of the post-fit errors. None of the NPs are very strongly pulled, which could possibly indicate some modeling issue of the data. None of the NPs appear to be excessively constrained by

the data. The JET_JER NP, corresponding to the jet energy resolution, is the most strongly pulled of the presented NPs. This pull is likely due to the excess observed in the VBF categories.

The dominant nuisance parameters affecting the observed results in the different production modes are shown in Figure E.1.

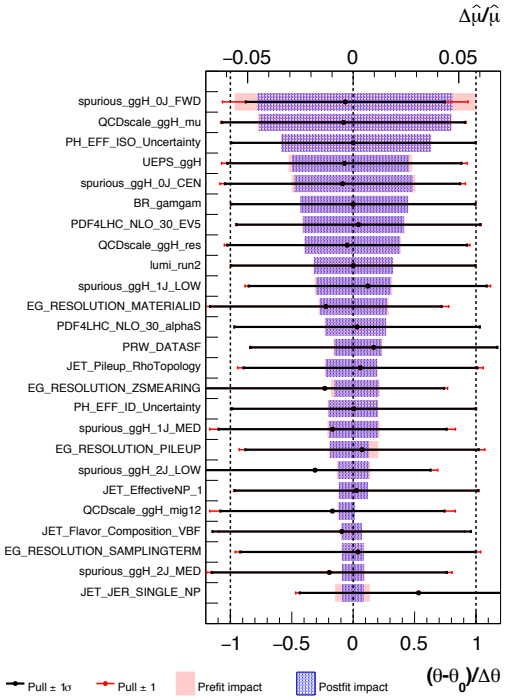
The NPs with "spurious" in the name indicate those which result from the background modeling (spurious signal) uncertainty. This uncertainty is one of the largest in the non-$t\bar{t}H$ portion of the analysis.

The PHOTON_EFF_ISO_Uncertainty NP represented the uncertainty due to the photon isolation efficiency. The EG_RESOLUTION NPs relate to the PER.

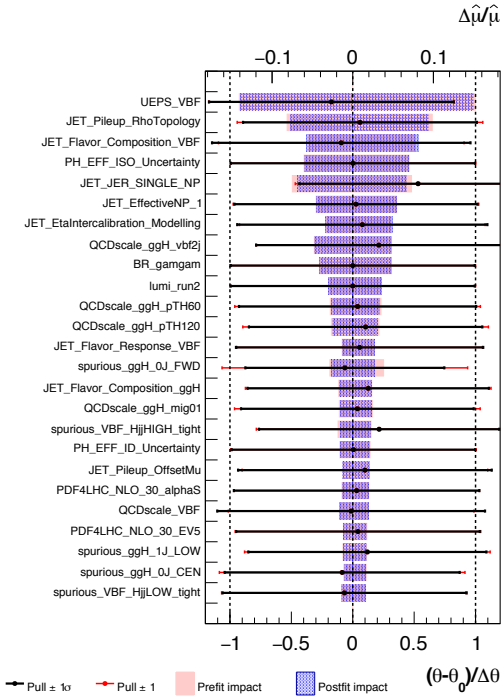
The BR_gamgam NP refers to the theoretical uncertainty on the Higgs boson to diphoton branching ratio.

The QCDscale_ggH NPs are part of the nine-NP scheme describing the higher-order *QCD* uncertainty relating to the ggF process.

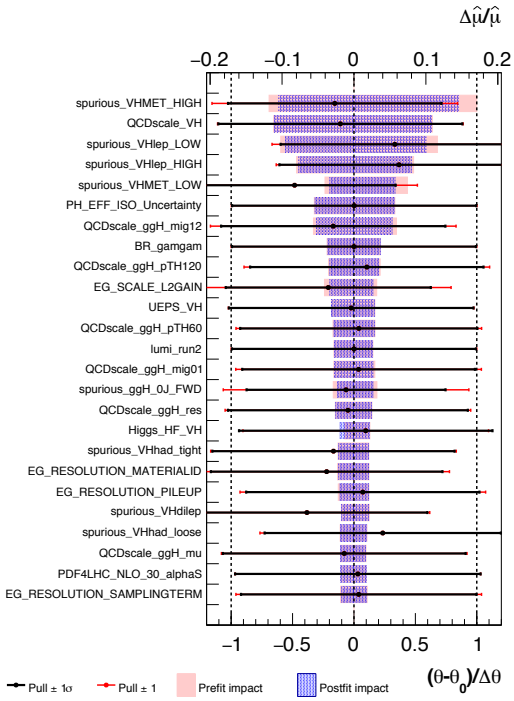
The lumi_run2 NP refers to the uncertainty in the total luminosity collected by the ATLAS detector in Run 2.



(a) ggF



(b) VBF



(c) VH

Figure E.1: The ranking of the largest nuisance parameters affecting the observed results in the main non-$t\bar{t}H$ production modes: (a) ggF, (b) VBF, and (c) VH.

BIBLIOGRAPHY

[1] "Standard Model," https://en.wikipedia.org/wiki/Standard_model, Accessed: 2020-01-16.

[2] "Feynman Diagram Maker Tool," https://www.aidansean.com/feynman/, Accessed: 2020-01-18.

[3] "Decay of a 125 GeV Standard-Model Higgs Boson," https://atlas.physicsmasterclasses.org/zpath_files/img/higgs_torte_DE.png, Accessed: 2020-01-18.

[4] H.H.J. ten Kate, "The ATLAS superconducting magnet system at the Large Hadron Collider," *Physica C: Superconductivity*, Vol. 468, No. 15, 2008, pp. 2137 – 2142, Proceedings of the 20th International Symposium on Superconductivity (ISS 2007).

[5] Aaboud, M., Aad, G., Abbott, B., Abdinov, O., Abeloos, B., Abhayasinghe, D. K., Abidi, S. H., AbouZeid, O. S., Abraham, N. L., and et al., "Measurement of the photon identification efficiencies with the ATLAS detector using LHC Run 2 data collected in 2015 and 2016," *The European Physical Journal C*, Vol. 79, No. 3, Mar 2019.

[6] ATLAS Collaboration, "The ATLAS Experiment at the CERN Large Hadron Collider," *Journal of Instrumentation*, Vol. 3, No. 08, aug 2008, pp. S08003–S08003.

[7] ATLAS Collaboration, "Barrel-toroid: BT-8," https://cds.cern.ch/record/892484, Accessed: 2020-01-28.

[8] ATLAS Collaboration, "Magnet System," https://atlas.cern/discover/detector/magnet-system, Accessed: 2020-01-28.

[9] ATLAS Collaboration, "Illustration of the chosen ϕ and η-binning of the muon spectrometer." https://twiki.cern.ch/twiki/bin/view/AtlasPublic/MuonPerformancePublicPlotsSimulation, Accessed: 2020-01-28.

[10] ATLAS Collaboration, "Performance of the ATLAS Trigger System in 2015," *Eur. Phys. J.*, Vol. C77, No. 5, 2017, pp. 317.

[11] "How ATLAS detects particles: diagram of particle paths in the detector," https://cds.cern.ch/record/1505342, Accessed: 2020-01-16.

[12] "Multivariate normal distribution," https://en.wikipedia.org/wiki/Multivariate_normal_distribution, Accessed: 2020-01-16.

[13] LHC Higgs Cross Section Working Group (D. de Florian et al.), "Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector," 2016.

[14] Anastasiou, C., Duhr, C., Dulat, F., Herzog, F., and Mistlberger, B., "Higgs Boson Gluon-Fusion Production in QCD at Three Loops," *Phys. Rev. Lett.*, Vol. 114, 2015, pp. 212001.

[15] Anastasiou, C., Duhr, C., Dulat, F., Furlan, E., Gehrmann, T., Herzog, F., Lazopoulos, A., and Mistlberger, B., "High precision determination of the gluon fusion Higgs boson cross-section at the LHC," *JHEP*, Vol. 05, 2016, pp. 058.

[16] Actis, S., Passarino, G., Sturm, C., and Uccirati, S., "NLO electroweak corrections to Higgs boson production at hadron colliders," *Phys. Lett. B*, Vol. 670, 2008, pp. 12–17.

[17] Anastasiou, C., Boughezal, R., and Petriello, F., "Mixed QCD-electroweak corrections to Higgs boson production in gluon fusion," *JHEP*, Vol. 04, 2009, pp. 003.

[18] Ciccolini, M., Denner, A., and Dittmaier, S., "Strong and electroweak corrections to the production of Higgs + 2 jets via weak interactions at the LHC," *Phys. Rev. Lett.*, Vol. 99, 2007, pp. 161803.

[19] Ciccolini, M., Denner, A., and Dittmaier, S., "Electroweak and QCD corrections to Higgs production via vector-boson fusion at the LHC," *Phys. Rev. D*, Vol. 77, 2008, pp. 013002.

[20] Bolzoni, P., Maltoni, F., Moch, S.-O., and Zaro, M., "Higgs production via vector-boson fusion at NNLO in QCD," *Phys. Rev. Lett.*, Vol. 105, 2010, pp. 011801.

[21] Brein, O., Djouadi, A., and Harlander, R., "NNLO QCD corrections to the Higgs-strahlung processes at hadron colliders," *Phys. Lett. B*, Vol. 579, 2004, pp. 149–156.

[22] Altenkamp, L., Dittmaier, S., Harlander, R. V., Rzehak, H., and Zirke, T. J. E., "Gluon-induced Higgs-strahlung at next-to-leading order QCD," *JHEP*, Vol. 02, 2013, pp. 078.

[23] Denner, A., Dittmaier, S., Kallweit, S., and Mück, A., "Electroweak corrections to Higgs-strahlung off W/Z bosons at the Tevatron and the LHC with HAWK," *JHEP*, Vol. 03, 2012, pp. 075.

[24] Beenakker, W. et al., "NLO QCD corrections to $t\bar{t}H$ production in hadron collisions," *Nucl. Phys. B*, Vol. 653, 2003, pp. 151–203.

[25] Dawson, S., Jackson, C., Orr, L., Reina, L., and Wackeroth, D., "Associated Higgs production with top quarks at the large hadron collider: NLO QCD corrections," *Phys. Rev. D*, Vol. 68, 2003, pp. 034022.

[26] Zhang, Y., Ma, W.-G., Zhang, R.-Y., Chen, C., and Guo, L., "QCD NLO and EW NLO corrections to $t\bar{t}H$ production with top quark decays at hadron collider," *Phys. Lett. B*, Vol. 738, 2014, pp. 1–5.

[27] Frixione, S., Hirschi, V., Pagani, D., Shao, H.-S., and Zaro, M., "Electroweak and QCD corrections to top-pair hadroproduction in association with heavy bosons," *JHEP*, Vol. 06, 2015, pp. 184.

[28] Dawson, S., Jackson, C., Reina, L., and Wackeroth, D., "Exclusive Higgs boson production with bottom quarks at hadron colliders," *Phys. Rev. D*, Vol. 69, 2004, pp. 074027.

[29] Dittmaier, S., Krämer, M., and Spira, M., "Higgs radiation off bottom quarks at the Tevatron and the CERN LHC," *Phys. Rev. D*, Vol. 70, 2004, pp. 074010.

[30] Harlander, R., Kramer, M., and Schumacher, M., "Bottom-quark associated Higgs-boson production: reconciling the four- and five-flavour scheme approach," 2011.

[31] Demartin, F., Maltoni, F., Mawatari, K., and Zaro, M., "Higgs production in association with a single top quark at the LHC," *Eur. Phys. J. C*, Vol. 75, 2015, pp. 267.

[32] "CMS, the Compact Muon Solenoid: Technical proposal," 1994.

[33] ATLAS Collaboration, "Measurement of the isolated diphoton cross section in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector," *Physical Review D*, Vol. 85, No. 1, Jan 2012.

[34] CMS Collaboration, "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC," *Physics Letters B*, Vol. 716, No. 1, Sep 2012, pp. 30–61.

[35] Lancaster, T. and Blundell, S. J., *Quantum Field Theory for the Gifted Amateur*, Oxford University Press, 2014.

[36] Srednicki, M., *Quantum Field Theory*, Cambridge University Press, 2007.

[37] Robinson, M., *Symmetry and the Standard Model*, Springer, 2011.

[38] Tanabashi, M. et al., "Review of Particle Physics," *Phys. Rev. D*, Vol. 98, Aug 2018, pp. 030001.

[39] Actis, S., Passarino, G., Sturm, C., and Uccirati, S., "NNLO computational techniques: The cases and," *Nuclear Physics B*, Vol. 811, No. 1-2, Apr 2009, pp. 182–273.

[40] Brüning, O. S., Collier, P., Lebrun, P., Myers, S., Ostojic, R., Poole, J., and Proudlock, P., *LHC Design Report*, CERN Yellow Reports: Monographs, CERN, Geneva, 2004.

[41] "Luminosity determination in pp collisions at $\sqrt{s} = 13$ TeV using the ATLAS detector at the LHC," Tech. Rep. ATLAS-CONF-2019-021, CERN, Geneva, Jun 2019.

[42] "ALICE: Technical proposal for a large ion collider experiment at the CERN LHC," 1995.

[43] Amato, S. et al., "LHCb technical proposal," 1998.

[44] ATLAS Collaboration, "The ATLAS silicon pixel sensors," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, Vol. 456, No. 3, 2001, pp. 217 – 232.

[45] ATLAS Collaboration, "Silicon microstrip detectors for the ATLAS SCT," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, Vol. 485, No. 1, 2002, pp. 84 – 88, Proceedings of the 5th International Conference on Large Scale Ap plications and Radiation Hardness of Semiconductor Detectors.

[46] ATLAS Collaboration, "Early Inner Detector Tracking Performance in the 2015 data at $\sqrt{s}$ = 13 TeV," Tech. Rep. ATL-PHYS-PUB-2015-051, CERN, Geneva, Dec 2015.

[47] Aubert, B. et al., "Liquid argon calorimetry with LHC-performance specifications," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, Vol. 315, No. 1, 1992, pp. 285 – 293.

[48] Bartoldus, R. et al., "A new ATLAS muon CSC readout system with system on chip technology on ATCA platform," *Journal of Instrumentation*, Vol. 11, No. 01, jan 2016, pp. C01059–C01059.

[49] Giordano Cattani, "Performance of the ATLAS resistive plate chambers," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, Vol. 661, 2012, pp. S6 – S9, X. Workshop on Resistive Plate Chambers and Related Detectors (RPC 2010).

[50] ATLAS Collaboration, "Performance of the ATLAS Trigger System in 2010," *Eur. Phys. J.*, Vol. C72, 2012, pp. 1849.

[51] ATLAS Collaboration, "Performance of the ATLAS track reconstruction algorithms in dense environments in LHC Run 2," *The European Physical Journal C*, Vol. 77, No. 10, Oct 2017, pp. 673.

[52] ATLAS Collaboration, "Vertex Reconstruction Performance of the ATLAS Detector at $\sqrt{s} = 13$ TeV," ATL-PHYS-PUB-2015-026, 2015.

[53] ATLAS Collaboration, "Topological cell clustering in the ATLAS calorimeters and its performance in LHC Run 1," *Eur. Phys. J. C*, Vol. 77, 2017, pp. 490.

[54] Cacciari, M., Salam, G. P., and Soyez, G., "The anti-k_t jet clustering algorithm," *JHEP*, Vol. 04, 2008, pp. 063.

[55] Cacciari, M., Salam, G. P., and Soyez, G., "FastJet user manual," *Eur. Phys. J. C*, Vol. 72, 2012, pp. 1896.

[56] ATLAS Collaboration, "Muon reconstruction performance of the ATLAS detector in proton–proton collision data at $\sqrt{s} = 13$ TeV," *Eur. Phys. J. C*, Vol. 76, 2016, pp. 292.

[57] Illingworth, J. and Kittler, J., "A survey of the hough transform," *Computer Vision, Graphics, and Image Processing*, Vol. 44, No. 1, 1988, pp. 87 – 116.

[58] ATLAS Collaboration, "Electron and photon performance measurements with the ATLAS detector using the 2015–2017 LHC proton-proton collision data," *Journal of Instrumentation*, Vol. 14, No. 12, Dec 2019, pp. P12006–P12006.

[59] Oreglia, M., *A Study of the Reactions* $\psi' \to \gamma\gamma\psi$, Ph.D. thesis, SLAC, 1980.

[60] Collaboration, A., "Performance of electron and photon triggers in ATLAS during LHC Run 2," 2019.

[61] Hamilton, K., Nason, P., Re, E., and Zanderighi, G., "NNLOPS simulation of Higgs boson production," *JHEP*, Vol. 10, 2013, pp. 222.

[62] Nason, P. and Oleari, C., "NLO Higgs boson production via vector-boson fusion matched with shower in POWHEG," *JHEP*, Vol. 02, 2010, pp. 037.

[63] Mimasu, K., Sanz, V., and Williams, C., "Higher order QCD predictions for associated Higgs production with anomalous couplings to gauge bosons," *JHEP*, Vol. 08, 2016, pp. 039.

[64] Luisoni, G., Nason, P., Oleari, C., and Tramontano, F., "$HW^{\pm}$/HZ + 0 and 1 jet at NLO with the POWHEG BOX interfaced to GoSam and their merging within MiNLO," *JHEP*, Vol. 10, 2013, pp. 083.

[65] Nason, P., "A New method for combining NLO QCD with shower Monte Carlo algorithms," *JHEP*, Vol. 11, 2004, pp. 040.

[66] Frixione, S., Nason, P., and Oleari, C., "Matching NLO QCD computations with Parton Shower simulations: the POWHEG method," *JHEP*, Vol. 11, 2007, pp. 070.

[67] Alioli, S., Nason, P., Oleari, C., and Re, E., "A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX," *JHEP*, Vol. 06, 2010, pp. 043.

[68] Butterworth, J. et al., "PDF4LHC recommendations for LHC Run II," *J. Phys. G*, Vol. 43, 2016, pp. 023001.

[69] Sjöstrand, T., Ask, S., Christiansen, J. R., Corke, R., Desai, N., Ilten, P., Mrenna, S., Prestel, S., Rasmussen, C. O., and Skands, P. Z., "An introduction to PYTHIA 8.2," *Comput. Phys. Commun.*, Vol. 191, 2015, pp. 159.

[70] ATLAS Collaboration, "Measurement of the Z/γ * boson transverse momentum distribution in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector," *Journal of High Energy Physics*, Vol. 2014, No. 9, Sep 2014.

[71] "ATLAS Run 1 Pythia8 tunes," Tech. Rep. ATL-PHYS-PUB-2014-021, CERN, Geneva, Nov 2014.

[72] Alwall, J., Frederix, R., Frixione, S., Hirschi, V., Maltoni, F., et al., "The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations," *JHEP*, Vol. 07, 2014, pp. 079.

[73] Lai, H.-L., Guzzi, M., Huston, J., Li, Z., Nadolsky, P. M., Pumplin, J., and Yuan, C.-P., "New parton distributions for collider physics," *Physical Review D*, Vol. 82, No. 7, Oct 2010.

[74] Corcella, G. et al., "HERWIG 6: An Event generator for hadron emission reactions with interfering gluons (including supersymmetric processes)," *JHEP*, Vol. 01, 2001, pp. 010.

[75] Bähr, M. and others, "Herwig++ physics and manual," *Eur. Phys. J. C*, Vol. 58, 2008, pp. 639.

[76] Bellm, J. et al., "Herwig 7.0/Herwig++ 3.0 release note," *Eur. Phys. J. C*, Vol. 76, No. 4, 2016, pp. 196.

[77] Djouadi, A., Kalinowski, J., and Spira, M., "HDECAY: a program for Higgs boson decays in the Standard Model and its supersymmetric extension," *Comput. Phys. Commun.*, Vol. 108, 1998, pp. 56–74.

[78] Djouadi, A., Mühlleitner, M. M., and Spira, M., "Decays of supersymmetric particles: The Program SUSY-HIT (SUspect-SdecaY-Hdecay-InTerface)," *Acta Phys. Polon.*, Vol. B38, 2007, pp. 635–644.

[79] Bredenstein, A., Denner, A., Dittmaier, S., and Weber, M. M., "Radiative corrections to the semileptonic and hadronic Higgs-boson decays $H \to WW/ZZ \to 4$ fermions," *JHEP*, Vol. 02, 2007, pp. 080.

[80] Bredenstein, A., Denner, A., Dittmaier, S., and Weber, M. M., "Precise predictions for the Higgs-boson decay $H \to WW/ZZ \to 4$ leptons," *Phys. Rev. D*, Vol. 74, 2006, pp. 013004.

[81] Bredenstein, A., Denner, A., Dittmaier, S., and Weber, M. M., "Precision calculations for the Higgs decays $H \to ZZ/WW \to 4$ leptons," *Nucl. Phys. Proc. Suppl.*, Vol. 160, 2006, pp. 131–135.

[82] Gleisberg, T. and Höche, Stefan. and Krauss, F. and Schönherr, M. and Schumann, S. and others, "Event generation with SHERPA 1.1," *JHEP*, Vol. 02, 2009, pp. 007.

[83] Höche, Stefan and Krauss, Frank and Schumann, Steffen and Siegert, Frank, "QCD matrix elements and truncated showers," *JHEP*, Vol. 05, 2009, pp. 053.

[84] Gleisberg, Tanju and Höche, Stefan, "Comix, a new matrix element generator," *JHEP*, Vol. 12, 2008, pp. 039.

[85] Schumann, S. and Krauss, F., "A Parton shower algorithm based on Catani-Seymour dipole factorisation," *JHEP*, Vol. 03, 2008, pp. 038.

[86] ATLAS Collaboration, "The ATLAS Simulation Infrastructure," *The European Physical Journal C*, Vol. 70, No. 3, Sep 2010, pp. 823–874.

[87] Agostinelli, S. et al., "GEANT4—a simulation toolkit," *Nucl. Instrum. Meth. A*, Vol. 506, 2003, pp. 250.

[88] ATLAS, C., Beckingham, M., Duehrssen, M., Schmidt, E., Shapiro, M., Venturi, M., Virzi, J., Vivarelli, I., Werner, M., Yamamoto, S., and Yamanaka, T., "The simulation principle and performance of the ATLAS fast calorimeter simulation FastCaloSim," Tech. Rep. ATL-PHYS-PUB-2010-013, CERN, Geneva, Oct 2010.

[89] ATLAS Collaboration, "Measurement of Higgs boson production in the diphoton decay channel in pp collisions at center-of-mass energies of 7 and 8 TeV with the ATLAS detector," *Physical Review D*, Vol. 90, No. 11, Dec 2014.

[90] ATLAS Collaboration, "Expected performance of the ATLAS b-tagging algorithms in Run-2," ATL-PHYS-PUB-2015-022, 2015.

[91] ATLAS Collaboration, "Measurements of b-jet tagging efficiency with the ATLAS detector using $t\bar{t}$ events at $\sqrt{s} = 13$ TeV," *JHEP*, Vol. 08, 2018, pp. 089.

[92] ATLAS Collaboration, "Measurement of b-tagging efficiency of c-jets in $t\bar{t}$ events using a likelihood approach with the ATLAS detector," ATLAS-CONF-2018-001, 2018.

[93] 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*, Vol. 76, 2016, pp. 581.

[94] ATLAS Collaboration, "Electron reconstruction and identification in the ATLAS experiment using the 2015 and 2016 LHC proton-proton collision data at $\sqrt{s}$ = 13 TeV," *Submitted to: Eur. Phys. J.*, 2019.

[95] ATLAS Collaboration, "Muon reconstruction performance of the ATLAS detector in proton–proton collision data at $\sqrt{s} = 13$ TeV," *Eur. Phys. J. C*, Vol. 76, 2016, pp. 292.

[96] ATLAS Collaboration, "Performance of missing transverse momentum reconstruction with the ATLAS detector using proton–proton collisions at $\sqrt{s} = 13$ TeV," *Eur. Phys. J. C*, Vol. 78, 2018, pp. 903.

[97] ATLAS Collaboraion, "Measurements of Higgs boson properties in the diphoton decay channel with 36 fb^{-1} of pp collision data at $\sqrt{s} = 13$ TeV with the ATLAS detector," *Physical Review D*, Vol. 98, No. 5, Sep 2018.

[98] Rainwater, D., Szalapski, R., and Zeppenfeld, D., "Probing color-singlet exchange in Z+2-jet events at the CERN LHC," *Physical Review D*, Vol. 54, No. 11, Dec 1996, pp. 6680–6689.

[99] Cabibbo, N., "Unitary Symmetry and Leptonic Decays," *Phys. Rev. Lett.*, Vol. 10, Jun 1963, pp. 531–533.

[100] Kobayashi, M. and Maskawa, T., "CP-Violation in the Renormalizable Theory of Weak Interaction," *Progress of Theoretical Physics*, Vol. 49, No. 2, 02 1973, pp. 652–657.

[101] Chen, T. and Guestrin, C., "XGBoost: A Scalable Tree Boosting System," .

[102] ATLAS Collaboration, "Search for Scalar Diphoton Resonances in the Mass Range 65–600 GeV with the ATLAS Detector in pp Collision Data at $\sqrt{s} = 8$ TeV," *Phys. Rev. Lett.*, Vol. 113, Oct 2014, pp. 171801.

[103] ATLAS Collaboration, "Measurement of the inclusive isolated prompt photon cross section in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector," *Phys. Rev. D*, Vol. 83, Mar 2011, pp. 052005.

[104] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, Vol. 12, 2011, pp. 2825–2830.

[105] Ebden, M., "Gaussian Processes: A Quick Introduction," 2015.

[106] Rasmussen, C. E. and Williams, C. K. I., *Gaussian Processes for Machine Learning*, MIT Press, 2006.

[107] Gibbs, M. N., *Bayesian Gaussian Processes for Regression and Classification*, Ph.D. thesis, University of Cambridge, 1997.

[108] "Measurements and interpretations of Higgs-boson fiducial cross sections in the diphoton decay channel using 139 fb^{-1} of $\sqrt{s}$ =13 TeV with the ATLAS detector," Tech. Rep. ATLAS-CONF-2019-029, CERN, Geneva, Jul 2019.

[109] "Procedure for the LHC Higgs boson search combination in Summer 2011," Tech. Rep. CMS-NOTE-2011-005. ATL-PHYS-PUB-2011-11, CERN, Geneva, Aug 2011.

[110] The ATLAS Collaboration, "Measurements of Higgs boson properties in the diphoton decay channel with 36 fb^{-1} of pp collision data at $\sqrt{s} = 13$ TeV with the ATLAS detector," *Physical Review D*, Vol. 98, No. 5, Sep 2018.

[111] Ranucci, G., "The profile likelihood ratio and the look elsewhere effect in high energy physics," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, Vol. 661, No. 1, 2012, pp. 77 – 85.

[112] Cowan, G., Cranmer, K., Gross, E., and Vitells, O., "Asymptotic formulae for likelihood-based tests of new physics," *European Physical Journal C*, Vol. 71, No. 2, 2011.

[113] Liu, X. and Petriello, F., "Reducing theoretical uncertainties for exclusive Higgs-boson plus one-jet production at the LHC," *Phys. Rev. D*, Vol. 87, No. 9, 2013, pp. 094027.

[114] Stewart, I. W., Tackmann, F. J., Walsh, J. R., and Zuberi, S., "Jet p_T Resummation in Higgs Production at $NNLL' + NNLO$," *Phys. Rev. D*, Vol. 89, No. 5, 2014, pp. 054001.

[115] Boughezal, R., Liu, X., Petriello, F., Tackmann, F. J., and Walsh, J. R., "Combining resummed Higgs predictions across jet bins," *Phys. Rev. D*, Vol. 89, 2014, pp. 074044.

[116] Stewart, I. W. and Tackmann, F. J., "Theory uncertainties for Higgs and other searches using jet bins," *Phys. Rev. D*, Vol. 85, 2012, pp. 034011.

[117] Gangal, S. and Tackmann, F. J., "Next-to-leading-order uncertainties in Higgs+2 jets from gluon fusion," *Phys. Rev. D*, Vol. 87, No. 9, 2013, pp. 093008.

[118] Spira, M., "QCD Effects in Higgs physics," *Fortsch. Phys.*, Vol. 46, 1998, pp. 203–284.

[119] Djouadi, A., Kalinowski, J., Mühlleitner, M., and Spira, M., "HDECAY: Twenty$_{++}$ years after," *Comput. Phys. Commun.*, Vol. 238, 2019, pp. 214–231.

[120] Butterworth, J. et al., "PDF4LHC recommendations for LHC Run II," *J. Phys. G*, Vol. 43, 2016, pp. 023001.

[121] "Electron and photon energy calibration with the ATLAS detector using data collected in 2015 at $\sqrt{s} = 13$ TeV," Tech. Rep. ATL-PHYS-PUB-2016-015, CERN, Geneva, Aug 2016.

[122] ATLAS and CMS Collaborations, "Combined Measurement of the Higgs Boson Mass in pp Collisions at $\sqrt{s} = 7$ and 8 TeV with the ATLAS and CMS Experiments," *Phys. Rev. Lett.*, Vol. 114, 2015, pp. 191803.

[123] Aaboud, M., Aad, G., Abbott, B., Abdallah, J., Abdinov, O., Abeloos, B., Aben, R., AbouZeid, O. S., Abraham, N. L., and et al., "Luminosity determination in pp collisions at $\sqrt{s}$ =8 TeV using the ATLAS detector at the LHC," *The European Physical Journal C*, Vol. 76, No. 12, Nov 2016.

[124] Avoni, G. et al., "The new LUCID-2 detector for luminosity measurement and monitoring in ATLAS," *JINST*, Vol. 13, No. 07, 2018, pp. P07017.

[125] ATLAS Collaboration, "Vertex Reconstruction Performance of the ATLAS Detector at $\sqrt{s} = 13$ TeV," ATL-PHYS-PUB-2015-026, 2015.

[126] ATLAS Collaboration, "Jet energy scale measurements and their systematic uncertainties in proton–proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector," *Phys. Rev. D*, Vol. 96, 2017, pp. 072002.

[127] ATLAS Collaboration, "Performance of b-jet identification in the ATLAS experiment," *JINST*, Vol. 11, 2016, pp. P04008.

[128] ATLAS Collaboration, "$E_{\mathrm{T}}^{\mathrm{miss}}$ performance in the ATLAS detector using 2015–2016 LHC pp collisions," ATLAS-CONF-2018-023, 2018.