

Search for New Physics in di-Higgs Decays with $H \rightarrow \gamma\gamma$ and without B-jet Pairs

Lucas Ehinger

August 2021

Abstract

I present a summary of my work with UC Berkeley's ATLAS collaboration in the search for beyond standard model (BSM) physics in the di-Higgs to di-photon plus X decay channel. I begin with an overview of the group's broader BSM search effort, before focusing on my role of introducing detector level defects into truth level simulated Monte Carlo (MC) events. I then conclude with a brief discussion of my work with the di-Higgs statistical analysis framework, describing the general techniques used to isolate a di-Higgs signal from background and explaining my role in the sensitivity analysis of the program.

1 Introduction

1.1 Broader Search

Much of the current UC Berkeley ATLAS research effort is currently directed towards the analysis of di-Higgs production and decay. Events with two Higgs bosons are of particular interest to experimentalists as it allows for the measurement of the Higgs self-coupling parameter λ_H which describes the Higgs' interaction with itself. This self-coupling constant is of special interest as it has yet to be precisely measured and could be an indicator to BSM particles and processes. By measuring the di-Higgs decay channels, we are thus able to constrain the Higgs self-coupling parameter.

A previous analysis included six different decay channels and constrained λ_H with an upper limit of 6.9 times that predicted by the standard model with a 95% confidence level.

The $HH \rightarrow \gamma\gamma b\bar{b}$ was one of the main decay channels of the analysis, and is one of the most sensitive to the Higgs self-coupling constant. In this decay channel, one of the Higgs decays to a pair of photons while the other decays into a $b\bar{b}$ pair. While the photons are easily detected with high levels of precision, the bottom quarks quickly hadronize, producing a jet of particles.

In our analysis, we investigate the similar, but orthogonal, decay channel with two photons and fewer than 2 b-jets. A sub-set of this decay channel, the $HH \rightarrow \gamma\gamma WW$, was included in the 6-decay channel analysis, but the broader decay channel in this analysis has yet to be fully studied. Figure 1 shows the different branching probabilities of di-Higgs decays, where one of the Higgs decays into two photons. During the LHC Run 2, about 4309 $gg \rightarrow HH$ events are produced. With a $H \rightarrow \gamma\gamma$ branching ration of 0.23%, approximately 19.8 of these events will include at least one di-photon decay. The $H \rightarrow b\bar{b}$ branching ratio is 58%, so we expect 11.5 of the di-photon di-Higgs decays to decay to a $b\bar{b}$ pair. However, detecting b-jets is very difficult experimentally and not all b-jets are correctly identified. For this analysis, we use the 77% b-jet tagging—that is only 77% true b-jets are correctly identified. Thus, while our analysis focuses on di-photon di-Higgs decays without two b-jets, we expect over 1/3 of the di-Higgs events to be mis-identified $b\bar{b}\gamma\gamma$ events.

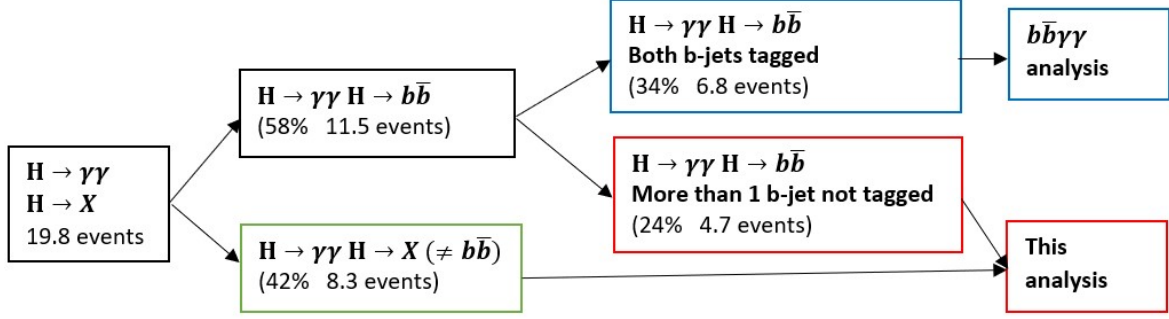


Figure 1: Sea Quark QCDNUM Evolution

Once all events containing a di-photon decay and fewer than two b-jets are collected, the HH sample must then be extracted from the background, consisting of both single Higgs and continuum background events. A Graphical Neural Network classifier is then trained to categorize the selected events into different bins, based on their GNN score. The higher the GNN score, the more likely a given event is to be a HH process. Thus, the GNN allows us to better separate the di-Higgs sample from the background.

In order to train the GNN, data is needed where the HH processes are already identified. Thus, simulated Monte Carlo (MC) data is required for both the di- and single-Higgs samples. For reasons that are beyond the scope of this paper, the continuum background is not created with simulated MC data.

2 Simulation

Simulation of data can generally be broken down into four broad steps; namely, Event Generation, Detector Simulation, Digitization, and Reconstruction.

Event generation uses a monte-carlo generator to simulate the particles created in the pp collisions. All particles created at this step are said to be truth-level, as they represent what was actually created. All following steps introduce the real-world uncertainties of experimental data collection at ATLAS.

The detector simulation step recreates the detector geometry and material properties, simulating how the particles interact (or fail to) with the detector arrays. Digitization converts the truth-level particles into electrical signals in the detector, after which reconstruction then converts the electrical signals back into particles.

The event generation process is fairly straightforward and multiple software packages exist to perform this step with a high level of precision. However, the introduction of detector-level inefficiencies is a significantly more complex process. While multiple third party programs also exist for these steps, their complexity and accuracy vary significantly. ATLAS's simulation program is by far the most

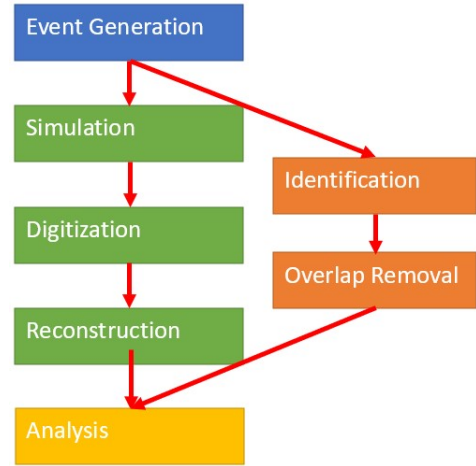


Figure 2: Simulation Workflow

accurate; however, it is run internally and simulation requests require well supported proposals.

Our group is not yet ready to submit a simulation request for HH data to CERN; however, we still need analysis-level data to train the GNN and setup the broader data analysis framework. In the following section, I describe our approach to introducing detector effects into truth-level data, thereby creating analysis-level data to train the GNN.

2.1 Approach

Fully simulating the ATLAS detector and particle interactions within the detector, as is done by ATLAS, is not feasible in the given time frame, nor was it necessary given that fully simulated ATLAS data will ultimately be used in the final GNN training. This, a simpler 2-step approach of identification and overlap removal was chosen. Converting the truth-level particles into electrical signals and then reconstructing them has multiple effects such as smearing the measured energies, momenta, and trajectories, potentially failing to identify the particle, or misidentifying it as another. We chose to directly introduce these inefficiencies in the identification step. During the overlap removal process, when particles had similar trajectories, one was removed. This simulates the effect of two particles hitting a detector very close to one another and being reconstructed as a single particle.

To develop and fine tune our identification and overlap removal scripts, we utilized an existing set of fully ATLAS simulated MC samples for single Higgs processes ¹. By running our two-step chain on truth-level single Higgs samples, we could then compare our analysis-level results (hereafter denoted truth smeared), with the fully simulated ATLAS single Higgs samples. Only after achieving sufficient agreement between truth smeared and fully simulated samples, do we then run our simulation chain on the desired di-Higgs sample.

2.2 Overlap Removal

Although overlap removal is the second step in this process, it is the simpler one, and will thus be discussed before identification. Generally, overlap removal is a post-reconstruction level step which is designed to avoid double-counting particles when multiple analysis-level particles are reconstructed from the same signal. While our technique does not involve a reconstruction step, we still apply the procedure. To start, if two particles are close enough, they will be detected as a single electrical signal; thus, instead of one of the particles being lost in the digitization phase, we lose it in the overlap removal phase. Second, in the event that two nearby particles are accurately detected and reconstructed, they may still be removed by the overlap removal procedure.

Table 1 gives the overlap removal procedure used. The non-jet events model two nearby particles being detected as a single particle. The four jet removals model a portion of the ATLAS post-reconstruction overlap removal procedure[1].

Kept	Lost	$\Delta R <$
Photon	Photon	0.1
Electron	Electron	0.1
Muon	Muon	0.1
photon	electron	0.4
photon	muon	0.4
photon	jet	0.4
electron	jet	0.2
jet	electron	0.4
jet	muon	0.4

Table 1: Overlap Removal Procedure

2.3 Identification

We began our identification procedure by applying simple p_t based identification probabilities to photons, electrons, and muons. That is, we adjusted the probability that a given particle would be identified given the particle type and transverse momenta. Note that the probability of detection is

¹For our analysis, we used the MC16a/d/e samples at 13TeV, the combination of which represents the all Run 2 data

not simply the ratio of truth-level events to fully simulated events, due to additional particle removals from the overlap removal procedure and a series of post analysis cuts. Thus, to accurately tune the momentum-based identification probabilities, we developed an iterative process of running our simulation script and then proportionally adjusting the identification probability (r) for the next iteration ($i+1$), as given in equation 1.

$$r_{i+1} = r_i + \frac{N_{full\ sim} - N_{truth\ smeared}}{N_{full\ sim}} 0.5 \quad (1)$$

The truth smeared distributions usually converged to the full simulation within ten iterations.

This transverse momentum based identification was tuned to the single Higgs ttH process. However, it did not perform as well for other single Higgs processes, such as ggH , as shown in figure 3c. The primary cause of this was due to the difference in event multiplicity; that is, there were far fewer particles in ggH events, and thus the identification probability was larger. Thus, to account for this, the photon, electron, and muon identification probabilities were adjusted using the scale factor given in equation 2, where 11.3 is the average number of particles for ttH events, for which the unscaled probabilities were tuned, and 100 is a constant adjusted to best fit the ggH process. As shown in figure 3b, this multiplicity correction has little effect on the previously well-tuned ttH process.

$$\alpha = 1 + \frac{11.3 - N_{particles}}{100} \quad (2)$$

ttH and ggH processes have the maximum and minimum particle multiplicities, with averages of 11.3 and 2.7, respectively; thus, all tuning was done with respect to the two processes. The proportional scaling between these two extremes also fit the mid-multiplicity processes well, and thus no further adjustments were made for those processes.

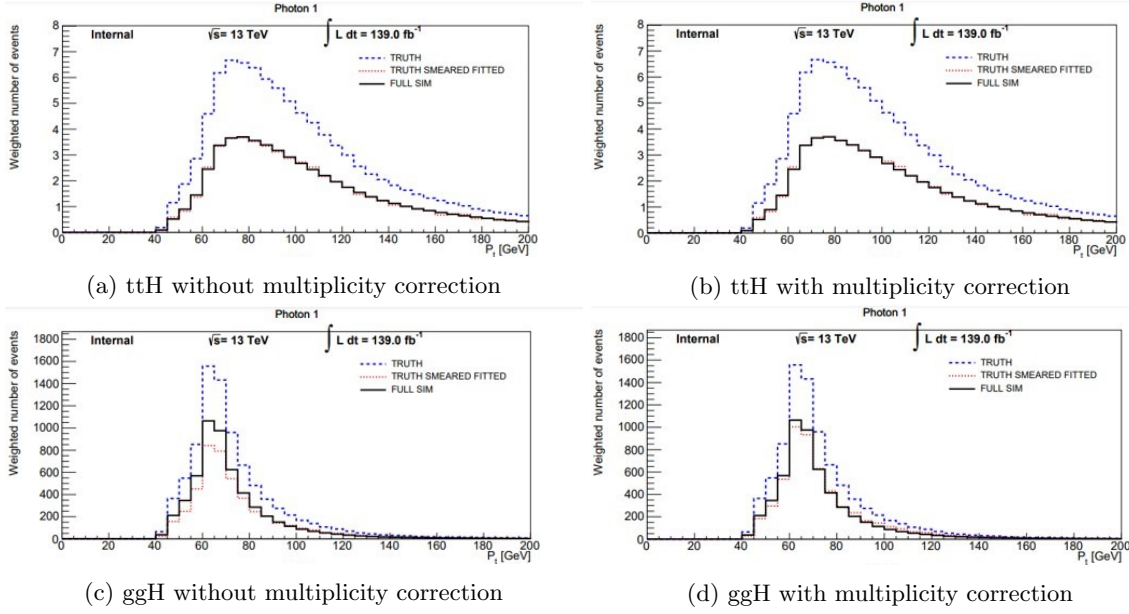


Figure 3: Effect of multiplicity correction on leading photon P_t .

While energy and momentum smearing occur with all particles detected, we chose to only smear the photon energy. As our analysis requires at least one $H \rightarrow \gamma\gamma$ decay, we expect a resonance at 125 GeV.

As seen in figure 4a, without out energy smearing, this peak is too narrow. Thus, we scaled all photon transverse momenta and energies. While not a particle, our GNN treats the MET as one, so we also scaled the missing x and y transverse energy (and thereby the total MET) in a similar manner, as shown in figure 4b. As we do not expect resonances from other particles, we chose not to smear their energies and momenta.

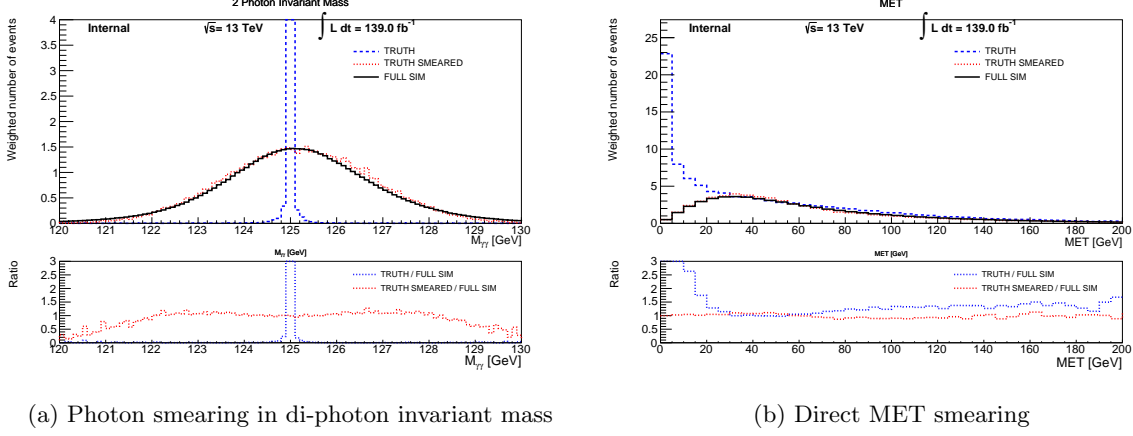


Figure 4: Photon energy and pt, and MET smearing.

Finally, we implemented particle mis-identification for the tau lepton. As shown in figure 5, there are significantly fewer tau events in the truth-level than in the full simulation. Thus, we assumed that roughly 2% of jets were mis-identified as Taus, a value which roughly agrees with experimental analyses (citation). All tau momenta were then scaled down by 40% to better match the full simulation distribution. Finally, the multiplicity based scaling for identification probabilities was also applied to the jet-to-tau mis-identification. To further complicate the identification procedure, adjusting jet distributions affected previously tuned distributions, especially the muons, through the overlap removal phase. Thus, other fits often had to be re-fitted after jet and tau adjustments.

Even with these multiple adjustments to the tau, there is still a significant discrepancy between the full simulation and truth smeared distributions, demonstrating the complexity of tau (mis-)identification processes. While further adjustments could have been applied to improve the fit, we felt that this would have been over-fitting our data with too many parameters. Given that tau leptons are not used in the current GNN analysis and that we ultimately plan on using ATLAS simulations, we deemed the identification process to be adequate.

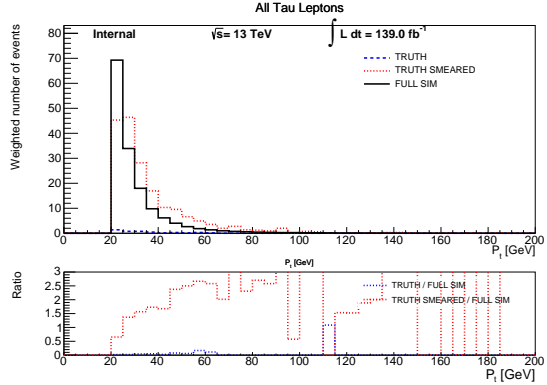


Figure 5: ggH Tau Lepton Pt distribution

3 GNN Mechanics

The Graphical Neural Network consists of a fully-connected graph, where each node represents a particle, jet or MET in the event and each edge represents the relation between two particles, such as ΔR . After training, the GNN then assigns all events a GNN score, and the data are binned according to GNN score. The higher GNN score bins have a higher HH signal-to-background ratio, effectively amplifying the HH signal.

4 Signal Sensitivity Study

Once the GNN categorizes the simulated events into bins, I then performed a sensitivity study on the statistical technique used to identify a di-Higgs resonance. The ultimate goal of this analysis is to produce a 95% confidence level upper bound on the di-Higgs cross-section. [2].

For this analysis, we investigate the leading di-photon invariant mass as our parameter of interest. We expect data to include not only this signal (a peak at 125GeV), but also single Higgs signal (also a peak at 125GeV) and a continuum background. We begin by determining the functional model, or shape, of each expected signal. Since no ATLAS full simulation data is available for the HH processes, it was assumed that the HH resonance would model the shape of the single Higgs processes, with a different normalization. Figure 6a shows the single Higgs fit using a double sided crystal ball function.

The continuum background was then modeled using power, exponential, and exponential polynomial fits, with the goal of selecting the best fit while also minimizing the number of parameters used. The exponential fit is shown in figure 6b.

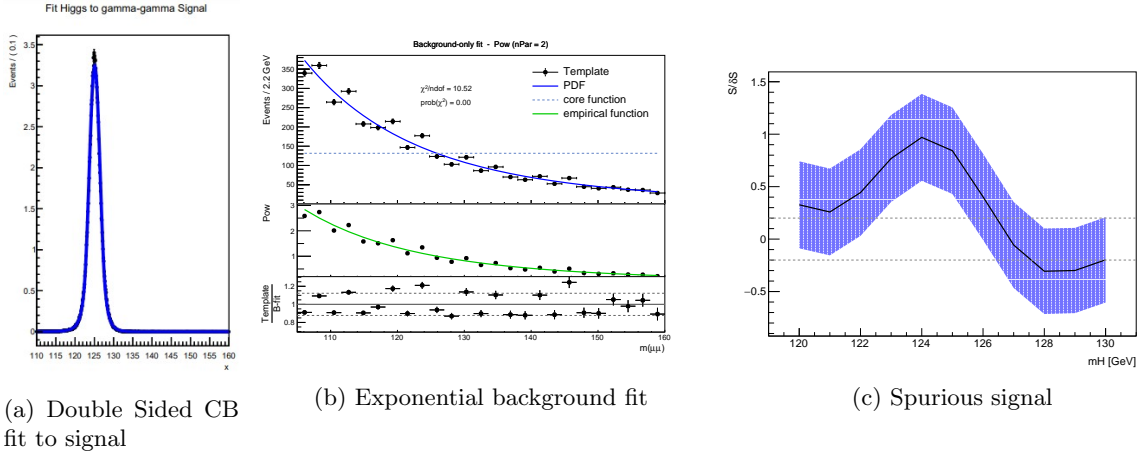


Figure 6: Signal and Background Models

In addition to the continuum background model, a spurious signal analysis is also performed. In this analysis, a signal + background fit is performed on only the background. The signal amplitude of this fit is defined as the spurious signal, as it is a signal component which appears as a result of an imperfect model of the background distribution. Figure 6c shows the spurious signal for the exponential background model, where the x-axis represents the center of the signal peak and the y-axis represents the spurious signal scaled by the statistical uncertainty.

These four signals (Higgs, HH, background and spurious signal) were then used to determine the upper limit on the HH cross section. For this analysis, we assume the single Higgs cross-section is that predicted by the standard, using the HH cross-section as the principal parameter of interest. The limit analysis was then also performed on a background sample (containing no Higgs or HH signal), using 100% uncertainty in the single Higgs signal. The results of both analyses are shown in table 2, with the expected 95% confidence level upper bound in the first row and the statistical uncertainty in the value shown in the rows below.

	No Uncertainty	100% Uncertainty
Expected	5.74	5.77
+1 σ	12.56	12.67
+2 σ	8.45	8.51
-1 σ	4.13	4.16
-2 σ	3.08	3.10

Table 2: 95% Confidence level upper bound on HH cross section. Values represent ratio to the SM prediction.

This analysis was primarily a proof of concept. The background sample used was relatively small; hence the large upper bound on the HH cross-section in a sample where none is expected. The uncertainty in the single Higgs model has a negligible effect on the upper limits, suggesting that data quantity and signal strength will have a more significant effect than the uncertainties.

As a proof of concept, this test principally demonstrates that we now have framework in place to determine an upper limit on the HH cross-section once the ATLAS-level fully simulated di-Higgs samples and our GNN are trained.

5 Conclusion

The majority of my contribution to this project was with the Identification process. Progress was very non-linear, as the first weeks were spend becoming familiarized with Linux and root, and creating simple analysis scripts. The majority of my progress came in the final weeks when I could use my analysis scripts to quickly adjust parameters; focusing on the physics more so than the coding.

My work on the limit analysis modeling in the final weeks was equally enjoyable. The majority of the analysis framework was already setup for this analysis; and thus my work consisted principally of creating configuration files to input our current samples into the analysis framework. Much of my time here was also spent understanding the statistical theory behind the limit analysis.

6 Acknowledgements

I would like to extend my gratitude to Professor Haichen Wang of UC Berkeley and Lawrence Berkeley National Lab for all his help and mentorship over this past summer. I would also like to thank Hongtao Yang for all his help with setting up and understanding the analysis framework, and to Zhicai Zhang for his help in explaining the broader background of the HH search and Linux framework. Thanks to Mary Kate, Junjie Zhu, and Myron Campbell for all their hard work in organizing the REU, including invaluable coding and GRFP sessions. Finally, I am grateful to the CERN Summer Student Program for organizing this summer’s series of high energy physics lectures. This material is based upon work supported by the National Science Foundation under Grant No. PHYS-1949923.

References

- [1] G. Aad et al. “Search for squarks and gluinos in events with hadronically decaying tau leptons, jets and missing transverse momentum in proton–proton collisions at $\sqrt{s} = 13$ TeV recorded with the ATLAS detector”. In: *The European Physical Journal C* 76.12 (Dec. 2016). ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-016-4481-2. URL: <http://dx.doi.org/10.1140/epjc/s10052-016-4481-2>.
- [2] G. Aad et al. “Combination of searches for Higgs boson pairs in pp collisions at s=13TeV with the ATLAS detector”. In: *Physics Letters B* 800 (2020), p. 135103. ISSN: 0370-2693. DOI: <https://doi.org/10.1016/j.physletb.2019.135103>. URL: <https://www.sciencedirect.com/science/article/pii/S0370269319308251>.