Comparing resolved and boosted jet identification algorithms to search for beyond the Standard Model scalar bosons with the ATLAS detector

Aidan Gardner-O'Kearny[1], Abraham Tishelman-Charny[2], and Elizabeth Brost[2]

[1]University Of Oregon

[2]Brookhaven National Lab

August 7th, 2024

Abstract

The Large Hadron Collider (LHC) located along the Swiss-French border near Geneva, Switzerland at CERN accelerates protons to near light speed before colliding them with a center of mass energy of $\sqrt{s} = 13.6$ TeV. The ATLAS detector is one of the general purpose detectors at the LHC. The ATLAS and CMS Experiments' discovery of the Higgs boson in 2012 completed the Standard Model, a theory that agrees with the vast majority of physical evidence. However, several well motivated beyond-the-Standard Model theories, which address a number of theoretical and experimental considerations, predict that there are several more Higgs-like scalar particles. In such a model, a generic scalar particle X could decay to a Higgs boson and another generic scalar, S. Our analysis is looking for the $SH \rightarrow \bar{b}b\gamma\gamma$ decay channel of the X particle, with the S decaying to b-jets and the Higgs decaying to a pair of photons. In cases where the S mass is sufficiently small compared to the X mass, the resulting b-jets are highly boosted, resulting in a very low separation angle between the two. This causes us to lose sensitivity without a way to tag the $S \rightarrow \bar{b}b$ decay. We are working on quantifying the increase in sensitivity to be gained from implementing a boosted $\bar{b}b$ tagger. This can be done by utilizing a score for jets generated by the tagging algorithm. We found that the boosted $\bar{b}b$ tagger score distribution in signal and background indicates a good separation between the two, a promising result. As a result of this project, I am now more capable of utilizing ROOT tools to perform particle physics analysis tasks, as well as understanding, utilizing, and debugging large code bases that I did not create myself.

Acknowledgements

I would like to take this opportunity to thank my mentor, Dr. Abraham Tishelman-Charny for his support with this project and support with my continued professional development. I would also like to thank Dr. Elizabeth Brost for her feedback on numerous matters. This project was supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Science Undergraduate Laboratory Internships Program (SULI).

This project was supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Science Undergraduate Laboratory Internships Program (SULI).

The research described herein is Fundamental as defined in the ITAR (22 CFR 120.34(a)(8)), EAR (15 CFR 734.8), or Part 810 (10 CFR 810.3), as applicable, and as described in the USD (AT & L) memoranda on Fundamental Research, dated May 24, 2010, and on Contracted Fundamental Research, dated June 26, 2008.

1 Introduction And Motivation

The Standard Model of particle physics is one of the most successful physical theories ever created. It describes three of the four fundamental interactions, as well as providing a highly accurate description of the everyday matter we interact with. Particles in the Standard Model are split into two categories: Fermions, with half integer spins, are matter particles, and bosons, with whole integer spins, are force carriers [1]. The Higgs boson, which provides mass to the fundamental particles, was discovered in 2012 by the ATLAS and CMS collaborations, and its discovery completed the Standard Model [2].

Despite this accuracy, the Standard Model is unable to provide satisfactory answers for a number of experimental observations. For instance, the Standard Model provides no explanation for dark matter, as it contains no particle that matches dark matter's observed behavior. Similarly, the Standard Model provides no explanation for how so much more matter than antimatter came into being during Baryogenesis.

To address these deficiencies, there are many proposed extensions to the Standard Model. In many of these extensions, there are predicted to be additional Higgs-like particles that are beyond the Standard Model (BSM). Supersymmetry predicts an extended Higgs sector, and simple Higgs doublet models predict there being five additional Higgs bosons [3][4]. These models are well motivated and consequently of high experimental interest.

2 Methods

2.1 The LHC and ATLAS Detector

Located along the Swiss-French border near lake Geneva, the Large Hadron Collider (LHC), is a circular particle accelerator capable of accelerating protons up to near the speed of light and is the largest particle accelerator every constructed. Protons are grouped in bunches and accelerated, before being collided at $\sqrt{s} = 13.6$ TeV.

In order to reconstruct these collisions, we use a large detector. The ATLAS detector is one of the general purpose detectors at the LHC. The ATLAS detector is constructed of several layers and sub-detectors that are each designed to detect and reconstruct different interactions. There is the inner detector, which reconstructs the trajectories of charged particles, the electromagnetic and hadronic calorimeters, which record the energy deposited by electrons, photons, and hadrons, and the muon spectrometer, which detects muons [5].

Each set of collisions is called an event, and in each event there is a chance for a hard scatter process that has the potential to produce physics processes that are of interest to us. We search through the large quantities of data produced to search for signs of these new processes, as well as to make precision measurements of Standard Model particles.

2.2 $SH \to \bar{b}b\gamma\gamma$ Analysis

In the ATLAS collaboration, much work is being done on di-Higgs (production of two Standard Model Higgs bosons) and extended Higgs sector physics. Our specific analysis is searching for generic BSM scalar bosons, referred to as X and S. These scalars both couple to the Standard Model Higgs and can be produced through gluon fusion, as seen in Figure 1.

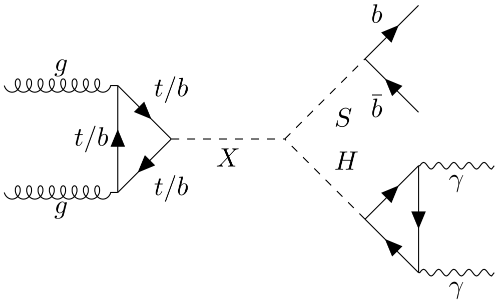

Figure 1: The signal $X \to SH \to \bar{b}b\gamma\gamma$ process produced through gluon fusion

Neither the X particle, the S particle, nor the Higgs are stable, and decay in fractions of a second. Signatures of the S and H decay products are used to reconstruct events, and identify which events may have come from S and H particles. This analysis is searching for the $\bar{b}b\gamma\gamma$ final state, where the Higgs decays to two photons and the S particle decays to two b quarks.

The X and S particles can have a wide range of masses. The range of masses we study is related to the available data set as well as the kinds of events we can reconstruct. The $X \to SH \to \bar{b}b\gamma\gamma$ analysis' studied mass range, as seen from Figure 2, have the X particle's mass ranging from $0.17 - 1$ TeV and the S particle's ranging from $15 - 600$ GeV.

The targeted signal process is not the only way to end up with two boosted b-jets and a pair of photons. In our analysis, one of the major backgrounds is the $\gamma\gamma +$ jets case. This is any case in which we end up with two non-resonant b-jets.

The analysis functions by training a parametrized neural network (PNN) on simulated signal and background to learn how to differentiate signal and background events. This allows real data to be input to the PNN, which outputs a score for each identified event. We can compare these scores versus the expected scores in a background only and background + signal case, with significant discrepancy from the background only hypothesis being a sign that we have observed our signal in data.

This analysis was performed on the complete Run 2 data set, and found a local excess of 3.5σ for the $m_S = 575$ GeV, $m_S = 200$ GeV case [6]. We are in the process of updating the analysis with the partial Run-3 data set to follow up on this excess.

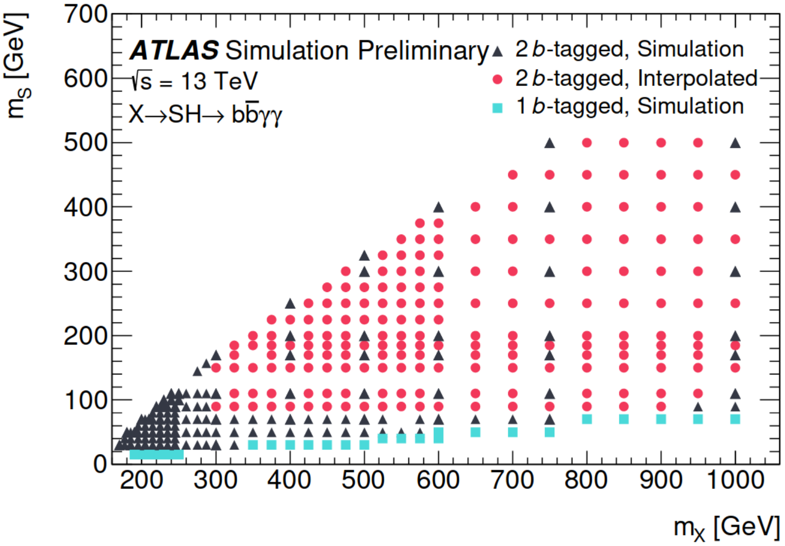

Figure 2: X and S mass values for produced Monte Carlo samples [7]

2.3 Jets

Quarks are governed by a theory known as Quantum Chromodynamics (QCD). In addition to having an electric charge, each quark also has a color charge. Hadrons, particles composed of quarks, must exist in a color neutral state. This property is called confinement, and it means that quarks are never observed by themselves [1]. In particle colliders, we will oftentimes see a process called hadronization occurring because of this. In hadronziation, two or more quarks are produced in a bound state at high energy. These quarks are produced in such a way that they move away from one another. As the quarks move apart, the energy of the bond between the quarks increases, eventually to the point where it is favorable for the bond to break and for a new quark to be formed in its place, maintaining the confined state. In colliders, this process will repeat many times, resulting in a shower of hadrons that interact with the detector. Reconstructing all of these particles individually would be basically impossible, and so they are reconstructed instead as objects called jets by a jet clustering algorithm.

A special kind of jet that we are concerned with are b-jets. Whereas most hadrons produced in particle collisions decay immediately and so the jet appears to come from the primary vertex, hadrons containing b quarks actually have sufficient lifetimes to travel meaningful distances into the detector from the primary vertex before a jet forms, as seen in Figure 3 [8].

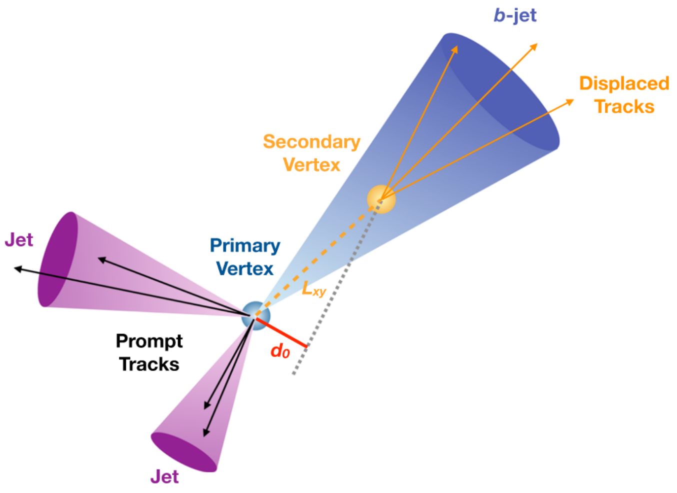

Figure 3: Different kinds of jets [8]

2.3.1 Boosted $\bar{b}b$-Jets

In cases where the X particle's mass is sufficiently large compared to the S particle's mass, the resulting b-jets from the S decay will be highly boosted. We define the threshold between boosted and resolved as $\frac{\mathrm{m}_S}{\mathrm{m}_X} < 0.09$. When this is true, the S particle will have a very high energy, resulting in the jets it decays into being highly boosted and consequently having a very low separation angle between them. As seen in Figure 4, after this threshold, the jets become close enough together that they can no longer be reconstructed individually.

One way to reconstruct these highly boosted events is with a boosted $\bar{b}b$-tagger. This tagger uses ATLAS detector info to determine if two collinear b-jets are present and reconstructs them as a single large-radius jet [10].

The Run 2 version of the analysis did not properly tag highly boosted, collinear, b-jets. This resulted in a loss of sensitivity for the cases where the jets were highly boosted. In order

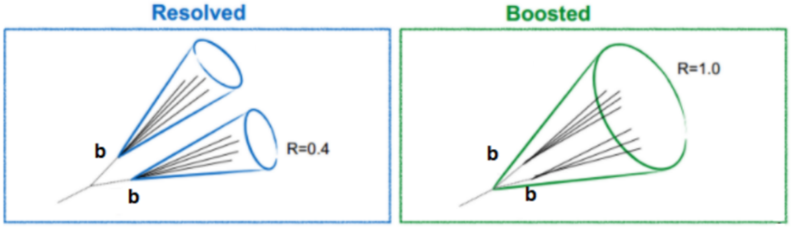

Figure 4: Small radius and large radius jets [9]

to determine the sensitivity to be gained from implementing a specific boosted $\bar{b}b$-tagger, we can investigate the way that implementing one allows us to separate signal from background.

For this study, we choose a single set of mass points for simulated samples, $m_X = 1$ TeV and $m_S = 70$ GeV. These masses are within the range we expect to be sensitive to at the LHC with $\sqrt{s} = 13.6$ TeV proton-proton collisions. It also means that the resulting jets will be able to be considered boosted.

2.4 Selections

2.4.1 Preselections

A framework utilizing C++ classes and Python configuration files was used to process simulated physics processes in the ATLAS detector. This results in the samples undergoing a preliminary selection process, operating on the event level, ensuring that each event meets a set of quality criteria that fit our desired signal.

First, all events must pass one of two trigger chains, each with different working points: a requirement that one photon has a $p_T > 35$ GeV and another has $p_T > 25$ GeV or a requirement that there is one photon in an event with $p_T > 140$ GeV. Events were also required to have at least two photons that were isolated, have a p_T to diphoton mass ratio of > 0.35 in one and > 0.25 in the other, and have a diphoton mass of 105 GeV $\leq m_{\gamma\gamma} <$ 160 GeV. There were required to be no leptons in an event, and each event needed to have at least two b-jets and less than six central jets. This last requirement is meant to cut away the $\bar{t}tH$ background where the Higgs decay to photons and the t pair decay via the $t \to bW$ channel, with the W's then decaying hadronically. These preselections are meant to reject background in ways that shouldn't affect our signal. For example, If the diphoton mass exists outside of the 105 GeV $\leq m_{\gamma\gamma} <$ 160 GeV range, it is highly unlikely that it came from the Standard Model Higgs.

2.4.2 Overlap Removal

It was noticed during an early attempt to plot the $\bar{b}b$-tagger score of signal and background that there was a large spike in the distribution at a consistent value. Further investigation revealed that this value was a default value, and that this behavior had first been noticed by another member of the analysis team. The working understanding of the behavior was that photons were being reconstructed as large-R jets. These jets would come from events with no tracks. In order to confirm this, we compared the separation between the closest photon and the large-R jet. This metric, ΔR, is defined as

$$\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2} \quad \text{where}$$
$$\eta = \ln\left(\cot\frac{\theta}{2}\right) \quad \text{is the pseudorapidity [11]}$$

and ϕ is the azimuthal angle in the detector.

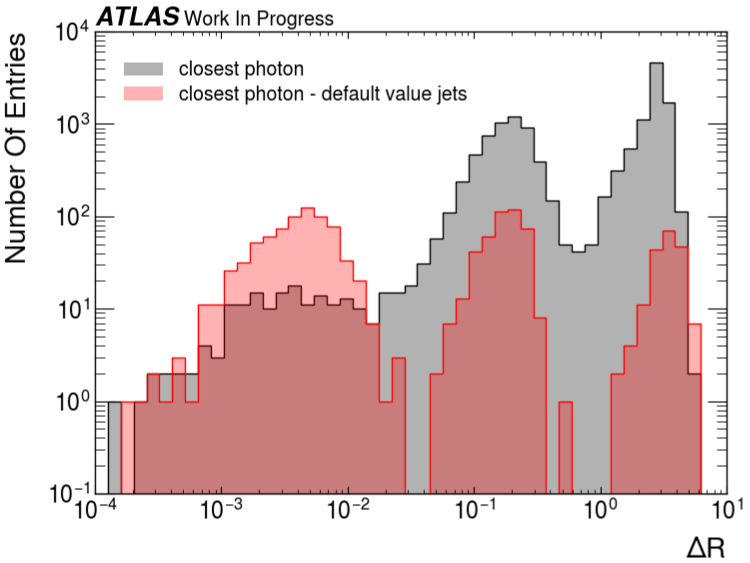

Figure 5: The separation between the closest photon in an event and large-R jet in simulation

This indicates that the large majority of default value jets are coming from photons, rather than the $S \to \bar{b}b$ candidate. In order to account for this behavior, we decided to implement a ΔR selection for all large-R jets. There is currently no standard for overlap removal involving large-R jets in ATLAS. This means that the information used to construct both the large-R jets and the photons. The value for the selection was chosen by calculating the ratio of large-R jets with tracks to default value large-R jets across a number of different threshold values, as seen in Figure 6. In a previous version of this plot, the maximum was at $\Delta R = 0.1$ separation between a large-R jet and the closest photon, and so that value was used for all future work. The latest version of this plot suggests $\Delta R \approx 0.25$ would be a better value, and this value would be used in the future.

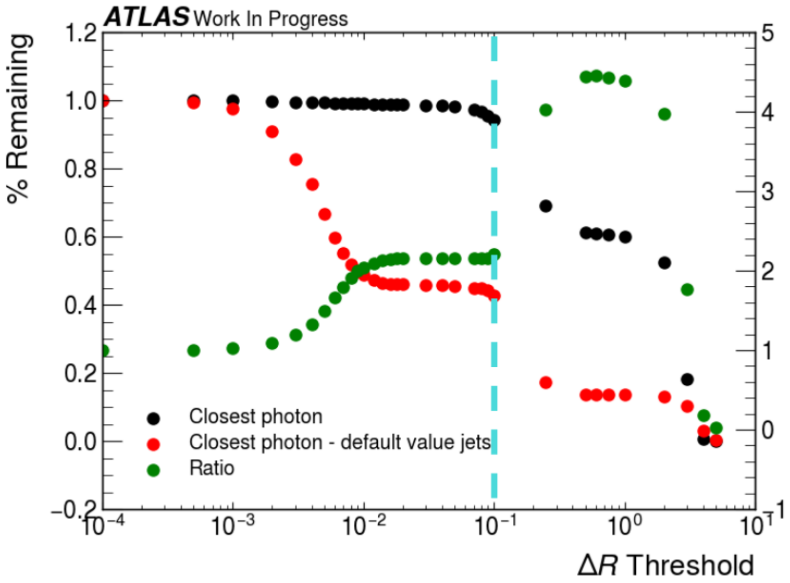

Figure 6: Ratio (shown in green) of surviving real and default value jets in simulation

2.4.3 Kinematic Selections

It was recommended that we implement kinetic selections on the large-R jets in a given event. These selections include > 250 GeV p_T requirement, a > 50 GeV mass requirement, and a < 2 pseudorapidity requirement.

These selections, as seen in Figures 7 and 9, cut out a large portion of large-R jets that would be unwanted, due to having very low masses and momenta.

2.5 Large-R Jets

One of the first important steps was to understand the way the jets in the samples behaved. Plotting the number of large-R jets per event in each sample revealed some unexpected behavior that led to a greater understanding of how to work with the samples.

The bulk of events in the background sample had no large-R jets in them, which was expected. As shown in Figure 8 the vast majority of SH events had two or more large-R jets being reconstructed, which was unexpected behavior. It was hypothesized that the extra

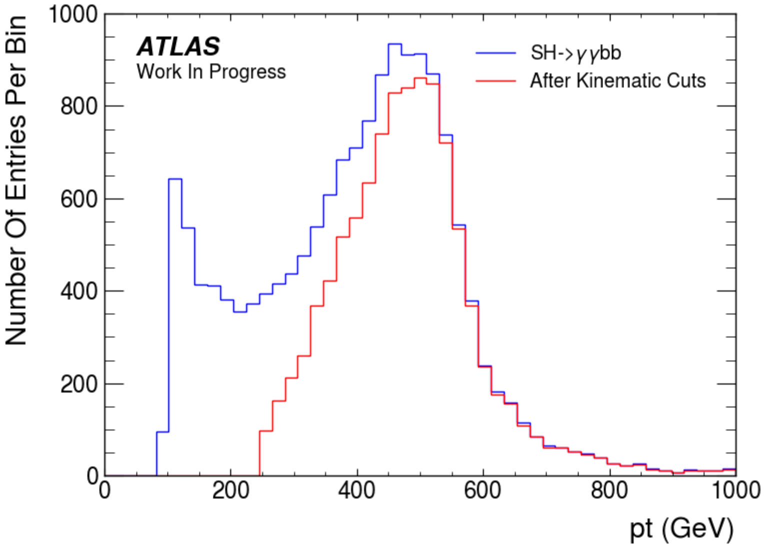

Figure 7: The p_T distribution of large-R jets in signal before and after kinematic selections in simulation. $m_X = 1$ TeV, $m_S = 70$ GeV

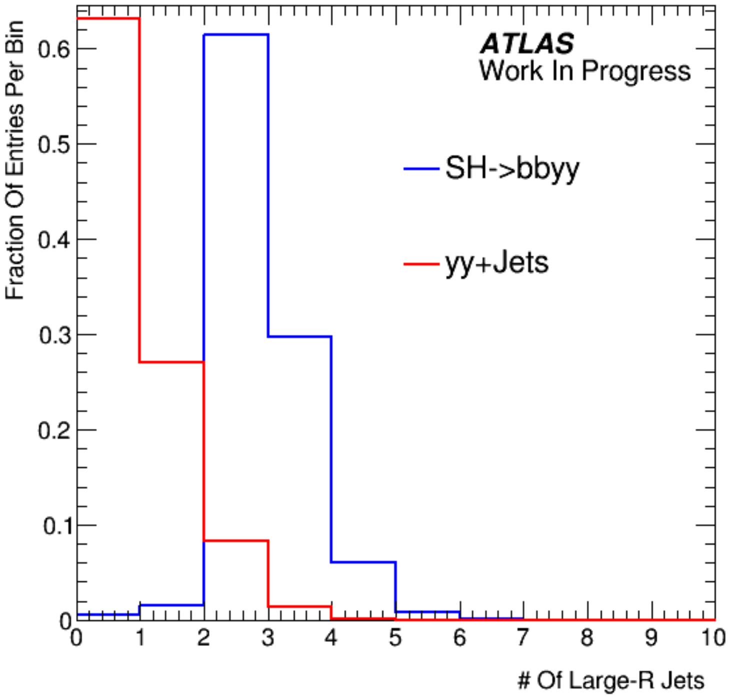

Figure 8: Number of large-R jets per event in both signal and background in simulation

large-R jet in an event would be coming from a $H \to \gamma\gamma$ decay being reconstructed as a large-R jet. Plotting the distribution of the signal large-R jets' mass provided further motivation to investigate that hypothesis.

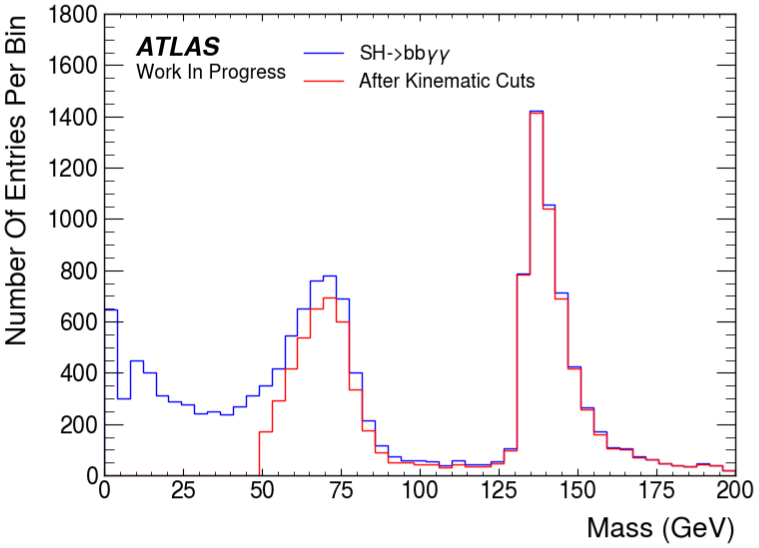

Figure 9: The mass distribution of large-R jets in simulated signal before and after kinematic selections, a peak can be seen at ≈ 70 GeV and another at ≈ 130 GeV. $m_X = 1$ TeV, $m_S =$ 70 GeV

In order to test this suspicion, we plotted the separation between the large-R jet and the position of the simulated Higgs boson (or "truth" position) versus the mass of the large-R jet. As seen in Figure 10, there are two major peaks, the 70 GeV peak occurs at a ΔR of around 3.14, indicating very high separation. The other peak occurred at the Higgs mass and a ΔR of close to 0. This indicated that the large-R jets with the Higgs mass were almost entirely coming from the $H \to \gamma\gamma$ decay.

3 Results

3.1 Candidate Selection

Because there was on average more than one large-R jet per event in the signal sample, we needed to identify a method to select the large-R jet that came from the $S \to \bar{b}b$ decay and not from the $H \to \gamma\gamma$ decay or another source of large-R jets, such as non-resonant di-jets. There were two proposed methods; selecting the large-R jet with the highest p_T, and selecting the large-R jet with the mass closest to the S particle's mass.

The separation between the selected candidate jet and the truth S particle position is a good way to measure the accuracy of the candidate selection method. A perfect method

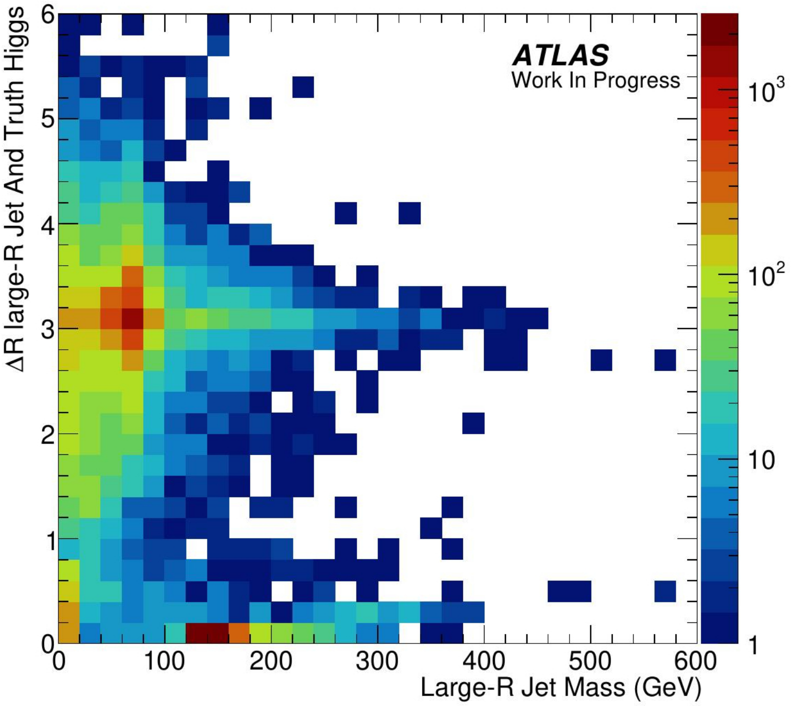

Figure 10: Separation between large-R jet and truth Higgs versus large-R jet mass in simulation

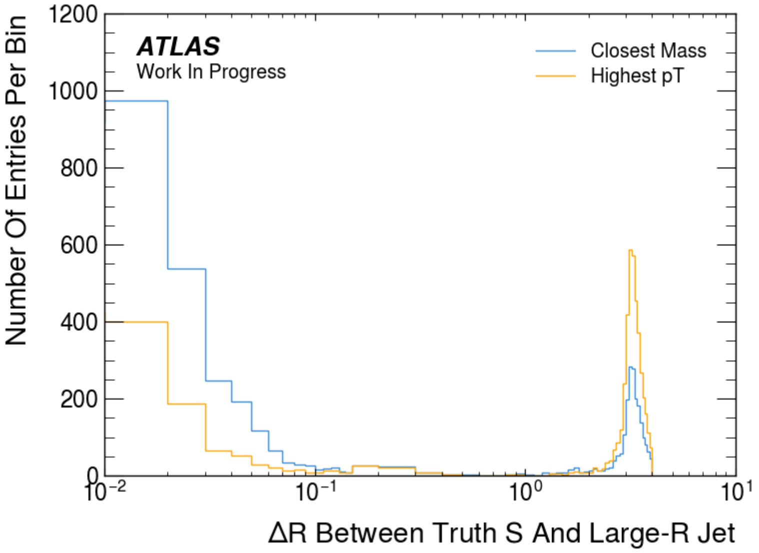

Figure 11: Separation between candidate large-R jet and truth S particle position in simulation

would result in there being nearly no separation between the two. In this case, selecting the large-R jet with the mass closest to the S particle mass proved to be the most accurate method of selecting a candidate, as can be seen in Figure 11 that resulted in the many more large-R jets having very low separations.

3.2 GN2X Scores

Processing the ntuples with our framework assigns a boosted $\bar{b}b$-tagger score to each large-R jet in an event. This score, called the GN2X score, ranges from 0 to 1 where the lowest score is not boosted $\bar{b}b$-like, and a 1 is very boosted $\bar{b}b$-like. The specific GN2X score we are using, pHbb, is used to determine the likelihood a jet came from an $H \to \bar{b}b$ decay, but we are using it for the $S \to \bar{b}b$, and the signature is expected to be similar [10].

After applying the preselections, the $\Delta R > 0.1$ requirement to reduce photon-jet overlap, as well as a slightly more restricting diphoton mass selection (120 GeV $\leq m_{\gamma\gamma} <$ 130 GeV), we were finally able to analyze the distribution of GN2X scores in both signal and background candidate large-R jets.

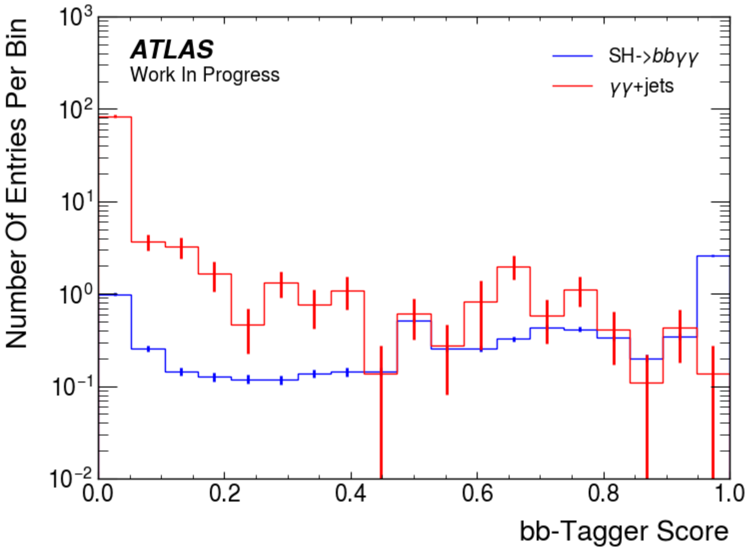

Figure 12: GN2X score distribution in both signal and background with statistical uncertainties in simulation

As seen in Figure 12 there is good separation between signal and background at GN2X scores around 1. As expected, the background distribution skews heavily towards GN2X scores of around 0, with many fewer large-R jets at GN2X scores around 1. The signal distribution trends towards GN2X scores nearer 1, although there is a large fraction of large-R jets with a score near 0, as well as a dip at a GN2X of ≈ 0.8.

4 Discussion

4.1 Summary

The separation of signal and background at high GN2X is a promising sign that the addition of a boosted $\bar{b}b$-tagger would be a useful addition to the analysis, potentially gaining sensitivity in the boosted regime. The $\gamma\gamma$ + jets background is heavily suppressed after selection, a good sign for sensitivity.

We have also found an efficient and accurate $S \to \bar{b}b$ candidate selection method. Selecting the large-R jet with the mass closest to the S particle's mass proved to be much better at selecting the jet that resulted from the targeted decay than selecting the highest p_T large-R jet.

We also identified the reconstruction of the $H \to \gamma\gamma$ decay as a large-R jet in the signal sample.

4.2 Next Steps

As the analysis moves forwards, there are a number of areas that we would like to understand further. First, we are interested in understanding the number of $H \to \gamma\gamma$ large-R jets are left after the implementation of the ΔR requirement. It is expected that that selection reduces the number of those large-R jets, but more work is needed to confirm this.

The GN2X score distribution indicated that the $\gamma\gamma$ + jets background was highly suppressed, but we need to add other backgrounds to see how they interact with the selections.

While the GN2X scores indicate that the implementation of a boosted $\bar{b}b$-tagger would be a gain in sensitivity, we still need to calculate the $\frac{\text{signal}}{\sqrt{\text{background}}}$ ratio to quantify the gain in sensitivity from implementation.

References

[1] D Griffiths. *Introduction To Elementary Particles.* Wiley-VCH, 2008.

[2] ATLAS Collaboration. Observation of a new particle in the search for the standard model higgs boson with the atlas detector at the lhc. *Physics Letters B*, 2012.

[3] Ilya F. Ginzburg, Maria Krawczyk, and Per Osland. Two-higgs-doublet models with cp violation, 2002.

[4] John Ellis, Mary K. Gaillard, and Dimitri V. Nanopoulos. An Updated Historical Profile of the Higgs Boson. volume 26, pages 255–274. October 2016. arXiv:1504.07217 [hep-ex, physics:hep-ph, physics:hep-th].

[5] ATLAS Collaboration. The atlas experiment at the cern large hadron collider: A description of the detector configuration for run 3. 2023.

[6] ATLAS Collaboration. Search for a resonance decaying into a scalar particle and a Higgs boson in the final state with two bottom quarks and two photons in proton-proton collisions at a center of mass energy of 13 TeV with the ATLAS detector, April 2024. arXiv:2404.12915 [hep-ex].

[7] ATLAS Collaboration. X $\rightarrow SH \rightarrow b\bar{b}$:summary of the full run 2 analysis. 2024.

[8] ATLAS Collaboration. Configuration and performance of the ATLAS b-jet triggers in Run 2. *The European Physical Journal C*, 81(12):1087, dec 2021.

[9] Galetsky Vladlen. Machine learning methods to improve boosted Higgs boson tagging at ATLAS.

[10] ATLAS Collaboration. Transformer neural networks for identifying boosted higgs bosons decaying into $\bar{b}$b and $\bar{c}c$ in atlas. *ATLAS Pub Note*, 2023.

[11] Matthew D. Schwartz. TASI Lectures on Collider Physics, September 2017. arXiv:1709.04533 [hep-ph].

Appendix

Participants

Name	Affiliation	Role
Abraham Tishelman-Charny	Brookhaven National Lab	Project mentor
Elizabeth Brost	Brookhaven National Lab	Provided feedback on numerous occasions

Table 1: Participants

Scientific Facilities

There were no scientific user facilities used in the course of this research.

Notable Outcomes

There were several internal presentations to members of the ATLAS Collaboration. Work in progress updates were provided to BNL's Omega Group meetings. A work in progress presentation was given to the $SH \to \bar{b}b\gamma\gamma$ analysis group as well as to the $\bar{b}b/\bar{c}c$ tagger team.