$t\bar{t}$ and QCD Multijet Backgrounds in the $hh \rightarrow b\bar{b}b\bar{b}$ Boosted Analysis with ATLAS

Josephine Brewster[1,2], Maximilian Swiatlowski[2], and Marco Valente[2]

[1]University of Victoria, Canada
[2]TRIUMF, Canada

August 23, 2024

Abstract

The goal of the ATLAS $hh \rightarrow b\bar{b}b\bar{b}$ analysis is to constrain the Higgs self-coupling κ_λ. A measurement of this parameter will be the first time the shape of the Higgs potential will be able to be probed experimentally. In the study presented in this report, $t\bar{t}$ backgrounds in the $hh \rightarrow b\bar{b}b\bar{b}$ boosted regime are studied using simulated data. This is a significant background for the boosted analysis though the majority of the backgrounds are QCD multijet. In this study, a transformer neural network (GN2X) $b\bar{b}$ tagger is applied to jets in $t\bar{t}$ events to test how much $t\bar{t}$ events are reduced compared to previous analyses. The composition of $t\bar{t}$ jets making it past the tagger is also studied and compared to simulated multijet backgrounds. A reduction of $t\bar{t}$ events is found compared to previous studies, both in proportion of total background and in total number of events.

Contents

1 Introduction

The Standard Model of particle physics (SM) has been shown to be a remarkable tool for making predictions at the subatomic scale. Many parameters of the SM have been tested experimentally and have agreed very well with the predictions. However, there are still many parameters that have yet to be tested. Something that has not yet been confirmed experimentally is the shape of the Higgs potential.

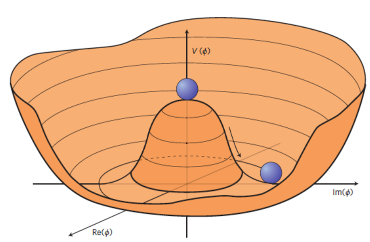


Figure 1: Standard Model Higgs potential. Source [5].

According to SM predictions, the Higgs potential should have the shape shown in Figure 1. However, if the Higgs potential were to take on a different shape, this would have big consequences for electroweak symmetry breaking and many other theoretical implications [6].

The next way that the shape of the Higgs potential will be able to be examined is by making a measurement of the Higgs self-coupling κ_λ [6]. This will give an idea of the shape of the Higgs potential around the minima.

To measure the Higgs self-coupling, events with two Higgs bosons (diHiggs events) are studied. It is possible to access the Higgs self-coupling though diHiggs events from Feynman diagrams such as Figure 2.

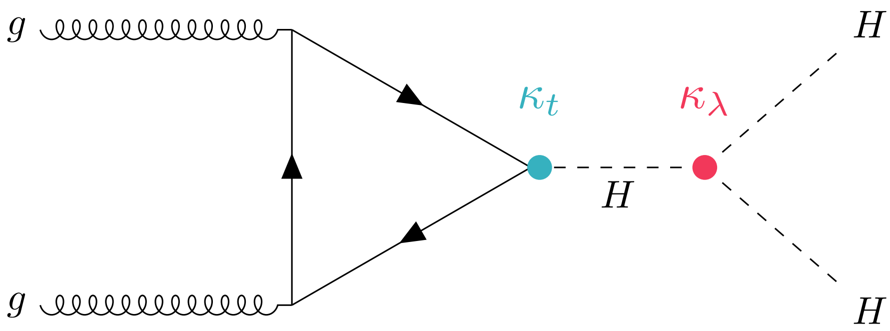


Figure 2: A Higgs pair production method via gluon-gluon fusion (ggF). Source: [4].

By accounting for many Higgs pair production modes (see Figure 10 in appendix for more diagrams) and observing the kinematics of hh events, the value of κ_λ can be extracted.

2 ATLAS $hh \to b\bar{b}b\bar{b}$ Boosted Analysis

The ATLAS $hh \to b\bar{b}b\bar{b}$ analysis is a search for the pair production of Higgs bosons with the aim to constrain the Higgs self-coupling κ_λ.

Several other decays of hh events are also being studied, but the motivation for the $b\bar{b}b\bar{b}$ decay mode is that it is the most common [7]. The backgrounds for $hh \to b\bar{b}b\bar{b}$ are high compared to some decay modes but this decay mode produces the most hh events overall.

This analysis is divided into the boosted and resolved regimes. The following study focuses on the boosted regime. These are events where the two Higgs bosons have large forward momentum. Thus, when each decays into two b quarks, the two jets coming from each b quark are very close together in detector space and can be reconstructed as one large radius (large-R) jet. Most hh events are created at low momentum where each b quark is resolved individually in a small radius jet. However, selecting the boosted events reduces backgrounds.

2.1 Backgrounds

In the boosted analysis, data-driven background estimation is used. The backgrounds from a control region surrounding the signal region are extrapolated into the signal region. The signal region will be some shape surrounding the point (125 GeV, 125 GeV). The exact signal region has yet to be defined for this analysis. This data-driven method is used because the backgrounds are mostly QCD multijet backgrounds, which are typically not well-modelled by simulation. QCD multijet events are events that contain multiple jets being produced from the pp collisions that are not the decay products of something "interesting" such as a heavy resonance.

$t\bar{t}$ is also a significant background. From previous studies, it appears to be about 10-30% of the overall backgrounds [4]. The reason $t\bar{t}$ can appear similar to a boosted hh event is because the most common decays of $t\bar{t}$ are the following:

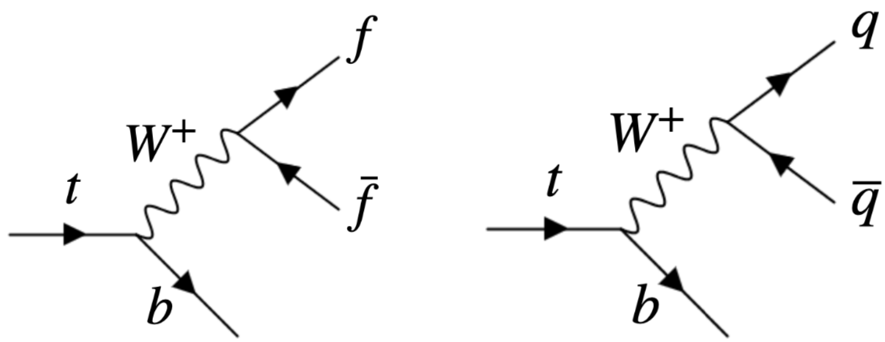


Figure 3: Common decays of a top quark.

where f is a fermion and q is a quark. When $t\bar{t}$ decays, two jets are produced. Each should contain one b quark and the decay products of the W boson. This can look very similar to two jets each containing two b quarks, especially if the W decay products appear similar to a second b quark.

A tagger is employed to try to distinguish between these background events and the signal.

2.2 GN2X Tagger

To select which jets are good candidates to be coming from a Higgs boson decaying to $b\bar{b}$, a tagger is used. For this analysis the tagger is the GN2X Hbb tagger [2]. GN2X is a transformer neural network tagger whose training includes tracks. The Hbb discriminant is defined as

$$D_{Hbb} = \ln\left(\frac{p_{Hbb}}{f_{cc}p_{Hcc} + f_{top}p_{top} + (1 - f_{cc} - f_{top})p_{qcd}}\right),\ f_{cc} = 0.02,\ f_{top} = 0.25$$

where p_{Hbb}, p_{Hcc}, p_{top}, p_{qcd} correspond to the probabilities of the jets coming from $h \to b\bar{b}$, $h \to c\bar{c}$, a top quark decay, and some other QCD jet respectively.

By looking at D_{Hbb} cuts on simulations of the signal ($h \to b\bar{b}$), percent efficiencies can be mapped to different values of D_{Hbb}. These efficiencies represent what percent of the signal is retained. That is, a larger percent value is a looser cut than a smaller percent value.

3 Studying $t\bar{t}$ Backgrounds Using Simulations

The following study focuses on examining the state of the $t\bar{t}$ backgrounds in the boosted regime after applying the GN2X tagger. This is done using Monte Carlo (MC) simulations to be able to study the MC truth information. Simulations are not very accurate, especially when simulating QCD multijet backgrounds. They still do however give a good idea of what is happening, especially with $t\bar{t}$.

Using MC truth information, the following study examines the content of the $t\bar{t}$ jets that are surviving the tagger cuts. This can be used to see what is "faking" the second b quark in the jets that make it past the tagger, which could possibly give some other rejection criteria for these events. Comparisons are made between the compositions of $t\bar{t}$ and multijet backgrounds with the same cuts.

3.1 Selection Criteria

For each aspect studied, three different D_{Hbb} cuts are applied: 85%, 70%, 60%. The current working point of the analysis is 60%. Looking at three different cuts gives a picture of how jets evolve with tighter cuts.

Since simulation events are used, all plots are scaled to 140 fb^{-1}. The region of most interest is the signal region. Since this has yet to be defined exactly for the current analysis, the signal region used is specified on a figure by figure basis.

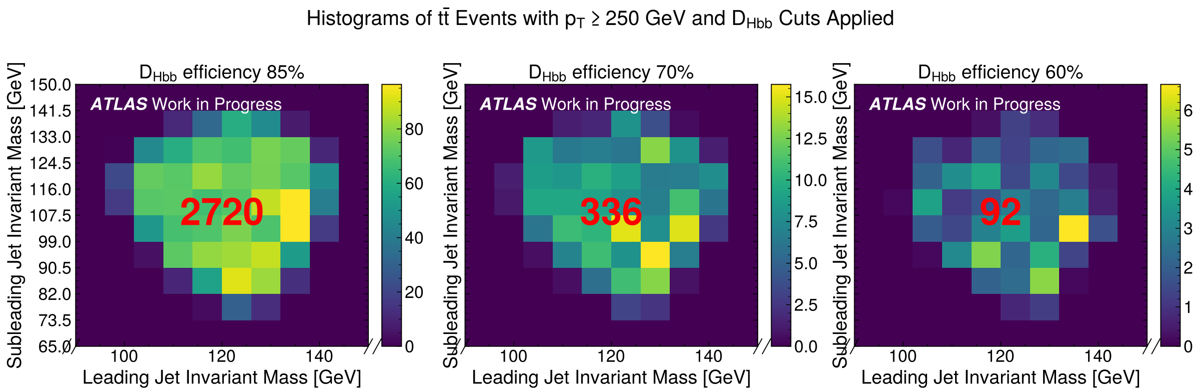
3.2 Initial Results with Tagger Cuts

Initially, a 2D mass plane histogram of both $t\bar{t}$ and multijet backgrounds is produced. This gives a general overview of the state of the backgrounds and also shows if there is any mass sculpting by the tagger. The overall number of events in the signal region is also shown.

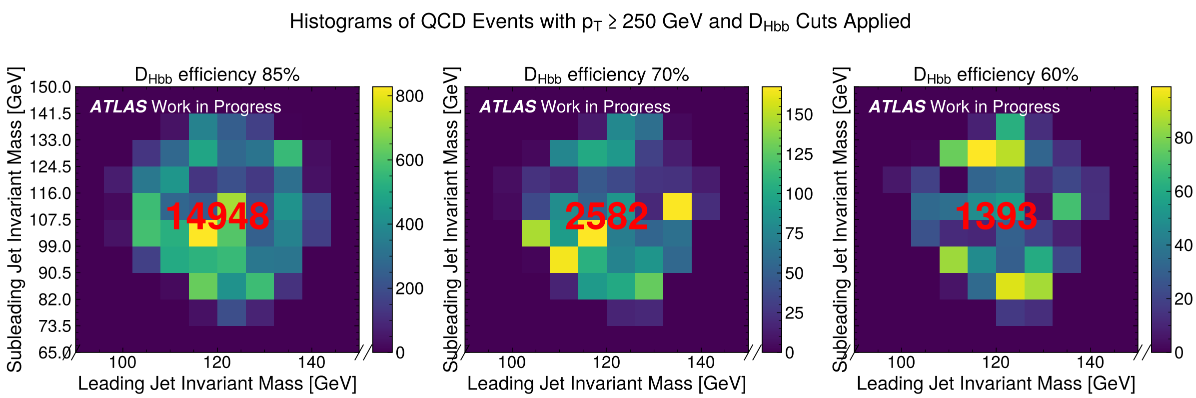
From the plots shown in Figure 4, the percent of backgrounds made up by $t\bar{t}$ is extracted. $t\bar{t}$ makes up 15%, 12%, 6% of the total backgrounds at 85%, 70%, and 60% tagger efficiencies respectively. Since the analysis working point is 60%, $t\bar{t}$ is a smaller proportion of the background than the 10-30% in the previous study [4]. GN2X seems to be handling $t\bar{t}$ very well proportionally. The next step is to compare absolutely.

3.3 Comparison to Previous Results

If the p_T cut on the jets is made at 450 GeV instead of 250 GeV, a comparison of the total number of remaining $t\bar{t}$ events can be made to a previous study [3]. As is shown in Figure 5, after tagger cuts of 85%, 70%, and 60%, the number of $t\bar{t}$ events remaining are on the order of 10^2, 10^1, and 10^0 respectively. From the previous study, at three similar cuts with their equivalent tagger, the number of $t\bar{t}$ events were on the order of 10^3, 10^2, and 10^0 respectively [3] (See Figure 11 in appendix). Since the signal regions are defined differently, only an order of magnitude comparison can be made.



(a) $t\bar{t}$ jets



(b) QCD jets

Figure 4: Histogram of events with two jets that pass the noted tagger cut, and each have $p_T \geq 250$ GeV, in the signal region. For this figure, the signal region used is the one defined in [4]. Sum of events in the signal region (scaled to 140 fb^{-1}) is in red in the centre of each plot.

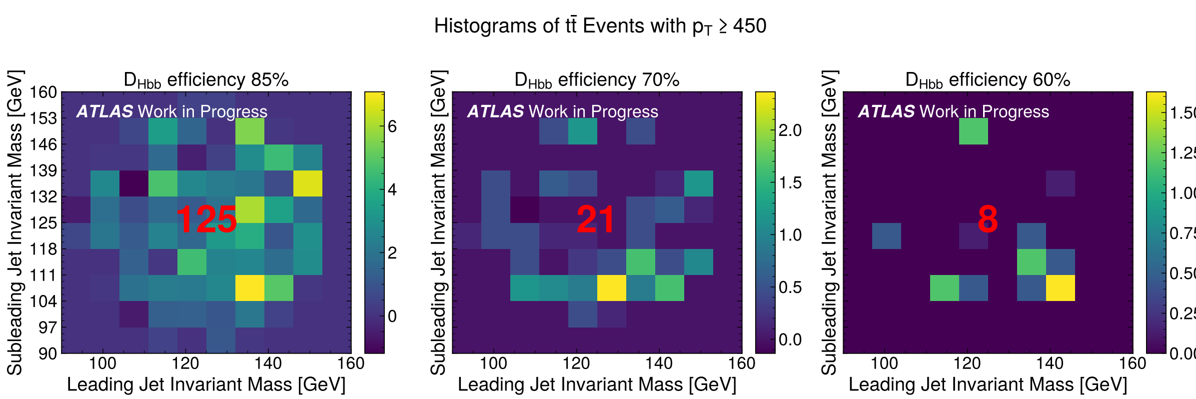


Figure 5: Histogram of events with two jets that pass the noted tagger cut, and each have $p_T \geq 250$ GeV, in the signal region. For this figure, the signal region used is a 30 GeV radius circle around (125 GeV, 125 GeV). Sum of events in the signal region (scaled to 140 fb^{-1}) is in red in the centre of each plot.

3.4 Truth Content of Jets

The above has given a picture of what the backgrounds look like after tagger cuts. Next, the composition of jets that are still making it past the tagger cuts will be studied.

3.4.1 τ Content of Jets

Using the MC truth information, the lepton content of the jets is examined. It is possible that τs from W decays could be getting $t\bar{t}$ jets by the tagger by acting as a second b. To test this, jets are divided into three categories: jets with a τ, jets with a lepton but no τ and jets with no leptons. They are also divided into the three different $t\bar{t}$ decay modes: full hadronic (both Ws decay into quarks), semi-leptonic (one W decays into quarks and the other decays into fermions), and leptonic (both Ws decay into fermions)[1]. These are labelled as $0l$, $1l$, $2l$ respectively in the following plots. Note that $0l$ can still contain leptons from a semi-leptonic b quark decay.

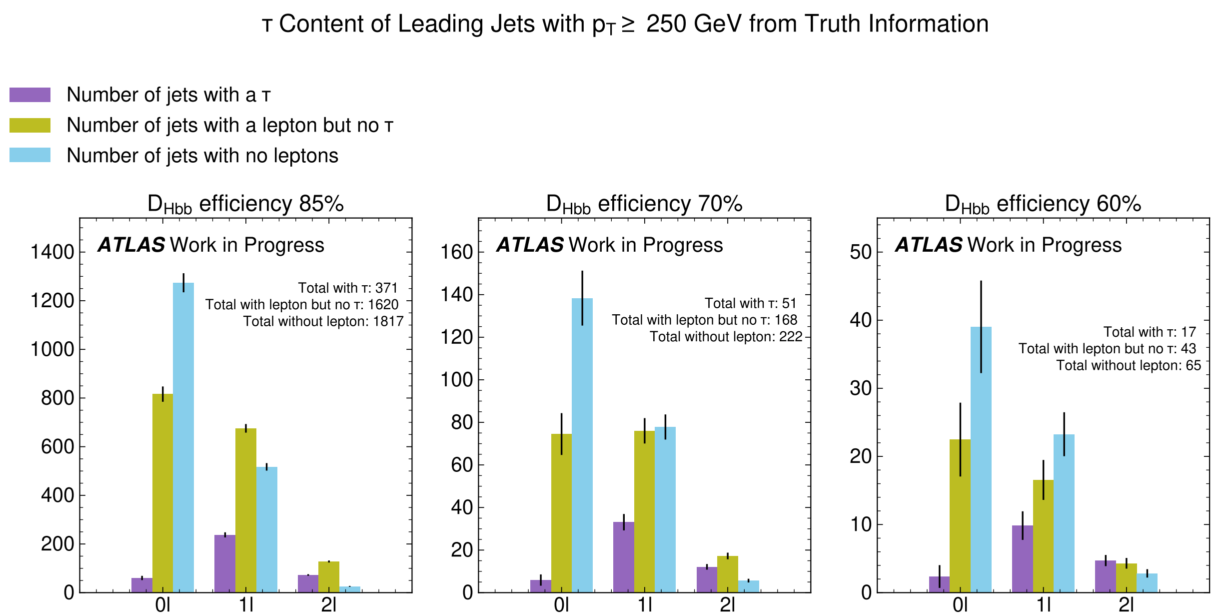


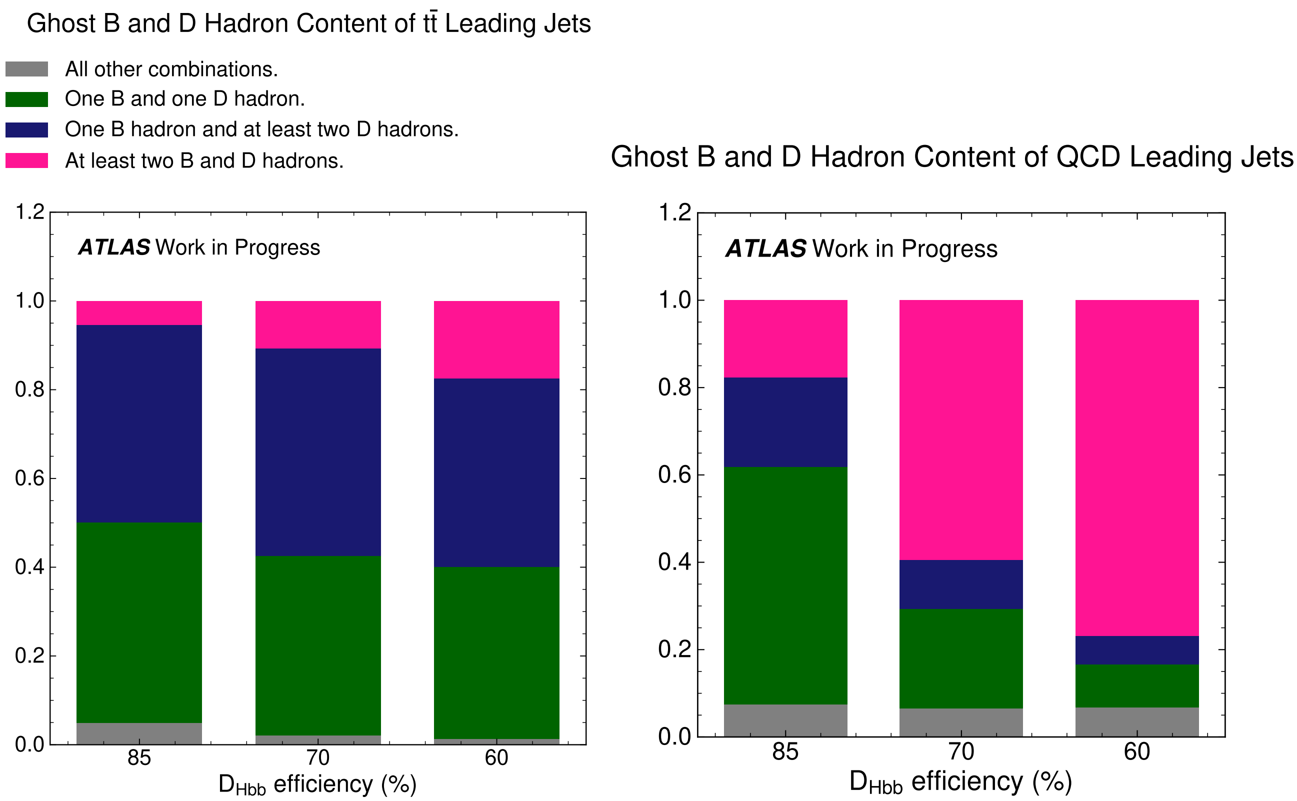
Figure 6: Lepton and τ content of leading jets at three different GN2X efficiencies. Jets are only taken from events in the signal region (30 GeV radius circle around (125 GeV, 125 GeV) in the mass plane). $0l$, $1l$, and $2l$ are full hadronic, semi-leptonic, and leptonic $t\bar{t}$ decays respectively.

Figure 6 shows that there is not an excess of τs in jets that are making it past the tagger cuts. There also does not seem to be any large variation in τ proportion between cuts. It appears that the GN2X Hbb tagger is doing very well with τs. It also does not seem like τs are responsible for backgrounds making it past cuts.

[1]See Figure 3 for the Feynman diagrams.

Thus, contamination may be coming instead from hadronic W decays. This is examined in the next section.

3.4.2 B and D Hadron Content of Jets in Signal Region



(a) $t\bar{t}$ jets

(b) QCD jets

Figure 8: B and D hadron content of leading jets at three different GN2X Hbb efficiency cuts. Jets are only taken from events in the signal region (30 GeV radius circle around (125 GeV, 125 GeV) in the mass plane).

The B and D hadron content is another aspect of jets that can be studied with MC truth information. B hadrons are hadrons that contain a B quark and D hadrons are hadrons that contain a c quark. In the following figures, combinations of numbers of Ghost[2] B and D hadrons are counted and compared.

For $t\bar{t}$, it is expected that most jets will have one B hadron. The most common decay of the b quark involves a c quark (see Figure 7) and so it is also expected that most $t\bar{t}$ jets will have at least one D hadron as well.

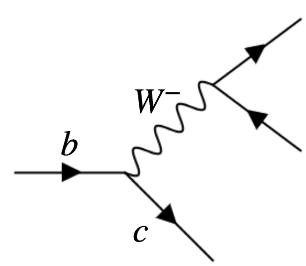
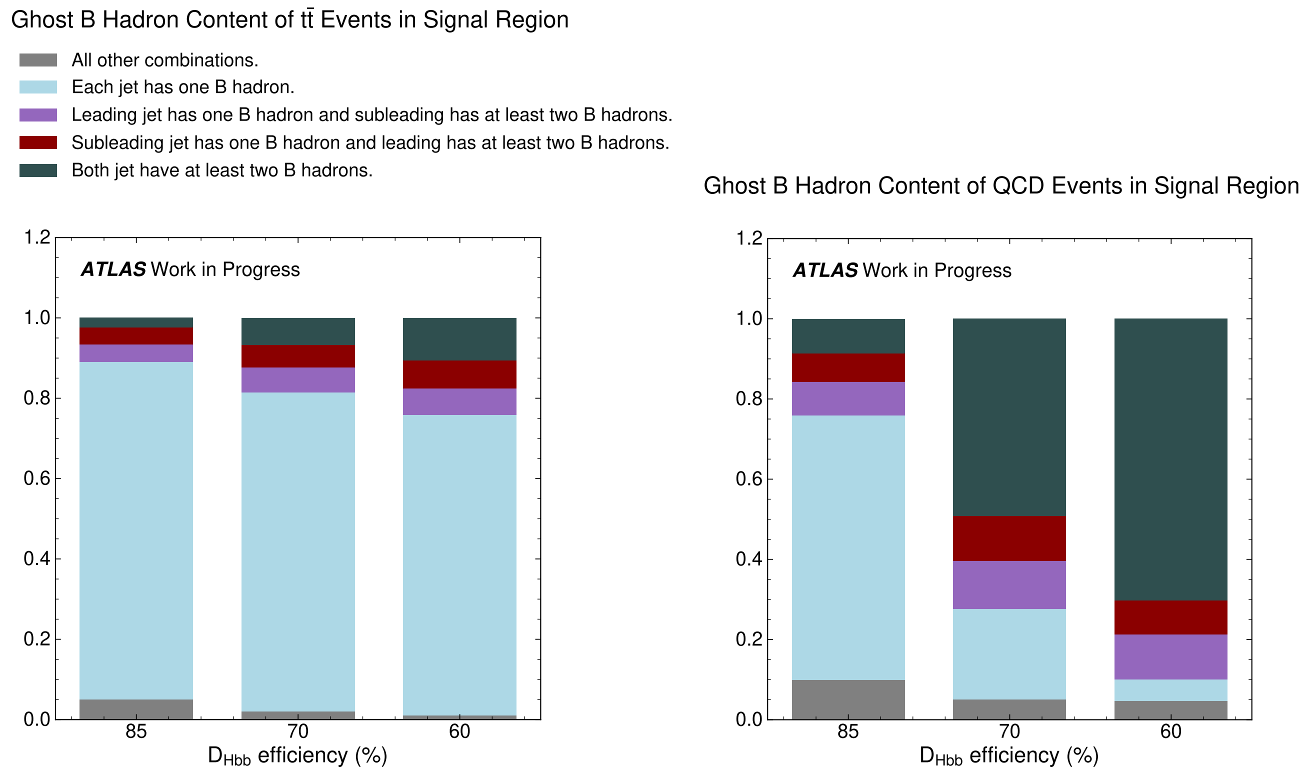


Figure 7

[2] See page 3 of [1] for a description of the ghost-association technique.



(a) $t\bar{t}$ events (b) Multijet events

Figure 9: B hadron content of events at three different GN2X Hbb efficiency cuts. Jets are only taken from events in the signal region (30 GeV radius circle around (125 GeV, 125 GeV) in the mass plane)

This is something that can be compared to QCD jets. With QCD jets there is less of a constraint on the number of B and D hadrons in each jet. Thus QCD jets are potentially able to have two B hadrons.

Figure 8a shows that the composition is close to what is expected, with one interesting feature. There are two regions that are most abundant: jets with one B and one D hadron, and jets with one B and at least two D hadrons. This is reasonably to be expected. The interesting part about this figure is the region counting jets with at least two B and D hadrons. If there are at least two B hadrons, then it makes sense that there will be at least two D hadrons as explained above. However, as the cuts get tighter, the higher presence of B hadrons is intriguing. The fact that events making it past tighter tagger cuts would be more likely to have two B hadrons makes sense. However, the origin of these extra B hadrons is not known. Potential sources could be W decays and final state radiation being picked up in the $t\bar{t}$ jets. This is a topic for future study.

This is also noticeable in Figure 8b to a much more extreme effect. By the 60% efficiency cut, most of the jets fall into the region with two D and B hadrons. QCD jets are able to have two real B hadrons, so it seems that this is what is enabling jets to make it through the tightest of cuts. For $t\bar{t}$ it seems more likely that a D hadron is helping the jets make it past tagger cuts, though there are some jets that do contain two B hadrons.

This gives a picture of the B and D hadron content at jet level. The D hadron content seems very reasonable so it is informative to look at how often both jets have two B hadrons at event level.

In Figure 9a, the region where both jets have one B hadron each dominates as expected. The probability of jets to have an extra B hadron definitely increases as tagger cuts get tighter, as was seen in Figure 8a.

Similarly to Figure 9a, in Figure 9b, the likelihood of both jets to have two B hadrons grows noticeably with tighter tagger cuts.

4 Conclusion

The main conclusion is that the GN2X Hbb tagger is doing a very good job of reducing $t\bar{t}$ backgrounds. It is doing a better job than the tagger from the previous analysis in this respect, and significantly better in looser cuts.

τ leptons are not present in a significant proportion of the $t\bar{t}$ jets that are making it past cuts. This indicates that the GN2X Hbb tagger is handling τs well. It also means that τs are likely not what is resembling a second b quark in the $t\bar{t}$ jets.

For $t\bar{t}$ jets, the presence of a D hadron is likely what is appearing like a the second b quark to the tagger. An interesting effect is apparent too. As tagger cuts become tighter, the small proportion of $t\bar{t}$ jets that contain two B hadrons does grow. The origin of the second B hadrons is not known. They could be coming from W decay or from final state radiation. This would be interesting for future study.

For QCD jets, it is likely that jets making it past tight tagger cuts do have two B hadrons each.

Acknowledgments

A large portion of this work was completed on the traditional, ancestral, and unceded territory of the xʷməθkʷəy̓əm (Musqueam) people. I am incredibly grateful to have been a guest on this land.

Thank you so much to Max Swiatlowski and Marco Valente for all of the guidance and help throughout this project as well as Dilia Portillo Quintero, Russell Bate, and Sébastien Rettie for the support while at CERN.

Thank you also to the Institute of Particle Physics in Canada and to TRIUMF for the funding and support. Finally, thank you to the CERN Summer Student program for a fantastic summer.

References

[1] Flavor Tagging with Track Jets in Boosted Topologies with the ATLAS Detector. Technical report, CERN, Geneva, 2014. All figures including auxiliary figures are available at https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PUBNOTES/ATL-PHYS-PUB-2014-013.

[2] Transformer Neural Networks for Identifying Boosted Higgs Bosons decaying into $b\bar{b}$ and $c\bar{c}$ in ATLAS. Technical report, CERN, Geneva, 2023. All figures including auxiliary figures are available at https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PUBNOTES/ATL-PHYS-PUB-2023-021.

[3] ATLAS Collaboration. Search for resonant pair production of higgs bosons in the $b\bar{b}b\bar{b}$ final state using pp collisions at $\sqrt{s} = 13$ tev with the atlas detector. *Physical Review D*, 105(9), May 2022.

[4] ATLAS Collaboration. Search for pair production of boosted higgs bosons via vector-boson fusion in the $b\bar{b}b\bar{b}$ final state using pp collisions at $\sqrt{s} = 13$ tev with the atlas detector, 2024.

[5] John Ellis, Mary K. Gaillard, and Dimitri V. Nanopoulos. A Historical Profile of the Higgs Boson. An Updated Historical Profile of the Higgs Boson. pages 255–274, 2016. 22 pages, 5 figures, update of arXiv:1201.6045, to be published in the volume.

[6] Biagio Di Micco, Maxime Gouzevitch, Javier Mazzitelli, and Caterina Vernieri. Higgs boson potential at colliders: Status and perspectives. *Reviews in Physics*, 5:100045, November 2020.

[7] S. Navas et al. Review of particle physics. *Phys. Rev. D*, 110(3):030001, 2024.

5 Appendix

5.1 Feynman Diagrams for diHiggs Events

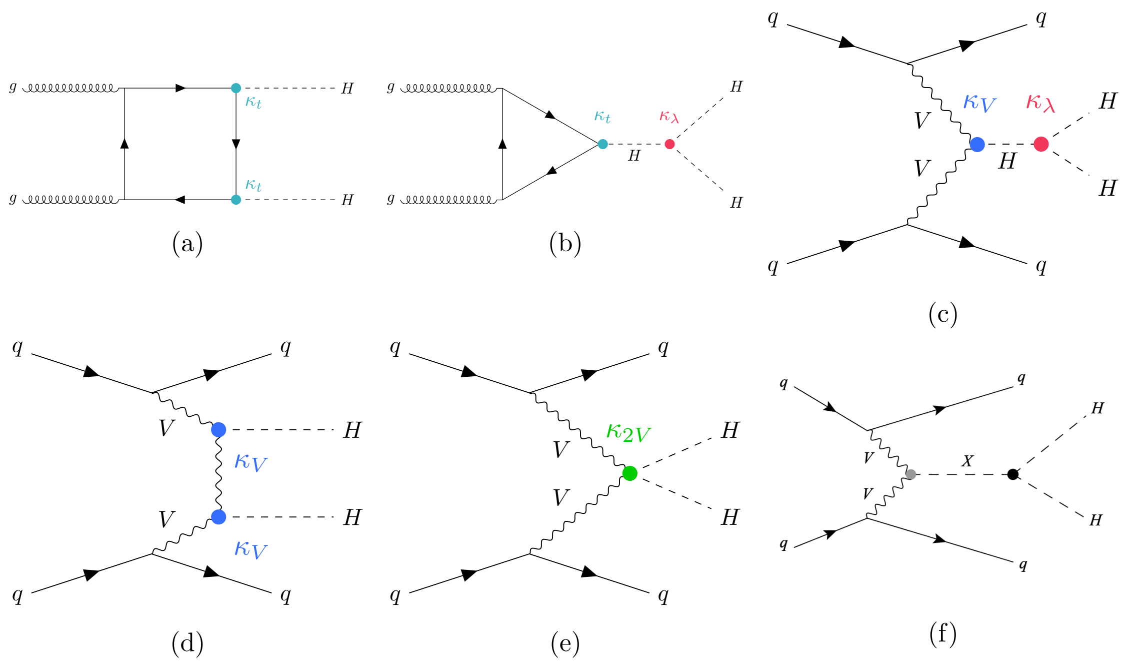
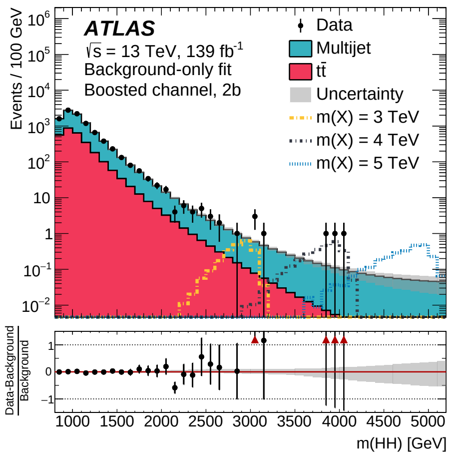
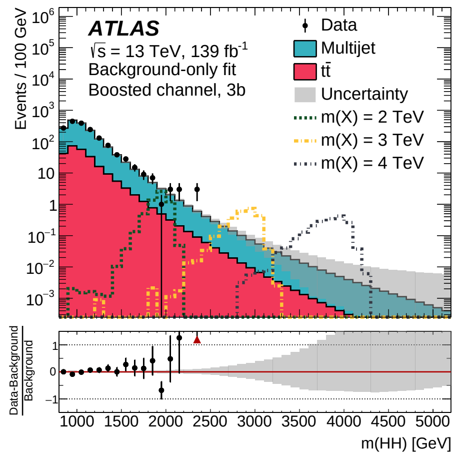


Figure 10: Some hh production modes at leading order. Source: [4].

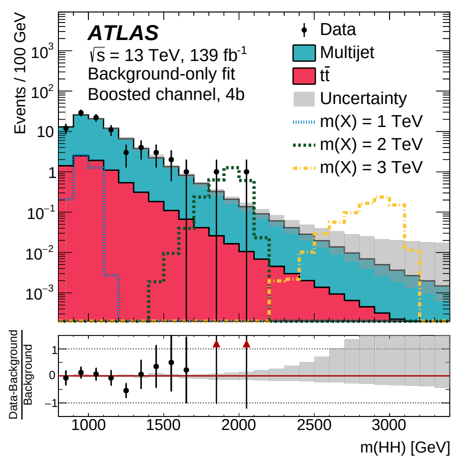
5.2 Results from Previous Study



(a)



(b)



(c)

Figure 11: Background simulations from previous study at three different tagger cuts. Source: [3]