BACKGROUND CHARACTERIZATION IN THE 4TOP SEARCH AT THE ATLAS EXPERIMENT

By

ALEX HARRISON STOKEN

A Thesis Submitted to The Honors College

In Partial Fulfillment of the Bachelors degree
With Honors in

Physics

THE UNIVERSITY OF ARIZONA

MAY 2019

Approved by:

Dr. Erich Varnes
Department of Physics

Abstract

A data-driven approach is taken to estimate the background and find optimal cuts for the four-top-quark search at the ATLAS Experiment for $\sqrt{s} =$ 13 TeV. Using 140fb^{-1} of data generated from the Large Hadron Collider from 2015 through 2018, the Monte Carlo simulated background is reweighted by a function of H_{all}^{T} to match the distribution of data. This correction function is then used to build MC signal and background distributions. To find the best region to search for the four-top signal, the significance of the signal to background is measured for various lower bounds, and the optimal selection criteria of $N_{bjets} > 2$ jets, $N_{jets} > 9$ jets, $H_{all}^{T} > 660000$ MeV, and $jet_{pt} > 60000$ MeV are chosen. These cuts produce a significance of 1.0077 and yield 60.22 expected signal events and 3511 expected background events in the data sample.

Contents

1 Introduction 3

1.1 The ATLAS Experiment . . . 3

1.2 Top Quarks . . . 3

2 Methods 6

2.1 MC Correction Function . . . 6

2.2 Key Variables . . . 7

2.3 Cut Selection . . . 8

3 Results 9

3.1 Initial Ratio Plots . . . 9

3.2 Fitting H_{all}^{T} Correction Function . . . 12

3.3 Cut Significance . . . 14

3.4 Signal/Background Plots . . . 15

4 Discussion 15

5 Conclusion 17

1 Introduction

This paper is focused on background characterization and estimation in the four-top-quark ($t\bar{t}t\bar{t}$) search at the ATLAS Experiment at $\sqrt{s} = 13$ TeV with an integrated luminosity of 140 fb^{-1}. Four-top-quark production is a rare Standard Model (SM) process that has not yet been seen experimentally, and determining the rate of four-top-quark production via the proton-proton (pp) collisions at the Large Hadron Collider (LHC) is a vital step in both validating current SM predictions and searching for new physics processes.

1.1 The ATLAS Experiment

The ATLAS (A Toroidal LHC ApparatuS) Experiment is the largest of the European Organization for Nuclear Research's experiments (CERN) at Large Hadron Collider (LHC) in Geneva, Switzerland. The LHC, a circular particle accelerator 27-kilometers in circumference, began operations in 2008 with the de facto goal of finding evidence for the Higgs Boson, supersymmetry, and new physics processes[1].

The ATLAS Experiment features a group of 3000 physicists around the world working with data produced by the ATLAS Detector. The 7000 ton detector was designed to be general-purpose and consists of the Inner Detector, Calorimeter, Muon Spectrometer, and Magnet System. These components combine to measure energies, momenta, change, and other properties of the particles released in the pp collisions.

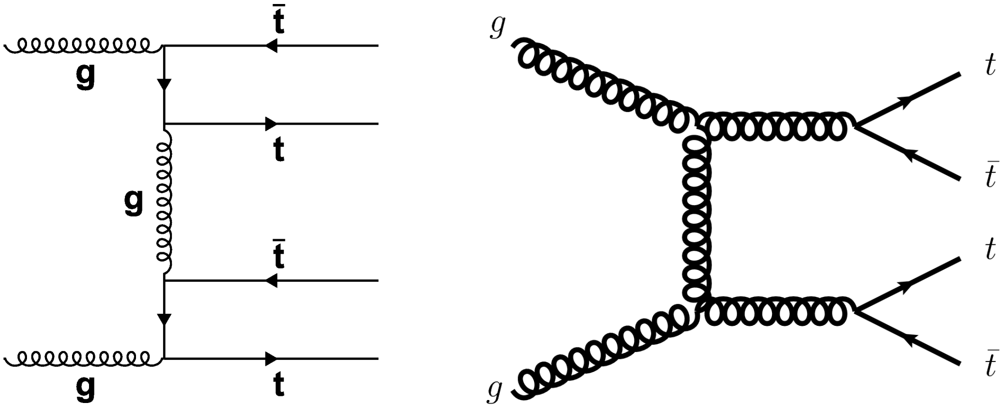
1.2 Top Quarks

The top quark is the most massive elementary particle, with a mass of 173 ± 0.4 GeV/c^2. Though the top quark experiences all four fundamental forces (gravity, eletromagnetism, weak nuclear, and strong nuclear), the top quark mainly interacts through the strong force and decays through the weak force. Due to its high mass, it is also highly coupled with the Higgs Boson and thus has proved useful in determining various Higgs properties. Because of its relative rarity, the top quark is also useful in testing the Standard Model and adjacent theories. If our analyses show, at a statistically significant level, more top/two-top/four-top quark events than predicted to be produced by SM processes, then there is a possibility of new physics processes at play. Based on the cross section attained by experimentalists, various theories are either supported or ruled out.

Because of its high mass, top quark production requires large amounts of energy and is extremely rare. While top quarks are naturally produced in the upper atmosphere via cosmic ray interactions, they can only be produced and studied extensively on Earth in particle accelerators. Four-top-quarks are primarily produced in SM gluon fusion (see Figure 1) in quantum chromodynamics. The gluons required to produce four-top-quarks come directly from the protons themselves - gluons are the exchange particles (gauge bosons) for the strong nuclear force, which hold the quarks in the protons together.

Once the four-top-quarks are produced, they decay into a few characteristic final states. Because the four top quarks decay very quickly, it is these decay products that are used to determine if an event contained four-top-quarks. The easiest to detect decay products of four-top-quarks are single leptons, opposite sign dileptons, same sign dileptons, and trileptons, along with lighter quarks. The difficulty in identifying four-top-events is that these decay products are also shared by events that do not have four-top-quarks, like $t\bar{t} + W, Z$ events and other many jet events. Analyzing the properties of the decay products to distinguish four-top (signal) events from all other events is the main goal of the ATLAS study. In particular, this paper will characterize the primary components of the background events, emphasizing E_{miss}^T, jet_{pt}, N_{jets},

N_{Bjets}, and H^T_{miss}.



(a) Process 1

(b) Process 2

Figure 1: Feynman Diagrams for two SM $t\bar{t}t\bar{t}$ production via quantum chromodynamics [2][3]

The cross section of a particle (σ), the likelihood that that particle will be produced in a given situation, is small for even single top quark production under Standard Model assumptions. For four top quarks, the cross section is five orders of magnitude smaller. The cross sections for various particles related to $t\bar{t}t\bar{t}$ is summarized in Table 1. Given this cross section, we only expect 1675.8 four-top events in our data sample. This highlights the difficulties in the four-top search, and is only further complicated because the signature for four-top-quark production is very similar to that of two-top-quarks (other processes also have similar signatures, but the two-top production is the most common of these similar processes). Processes with similar signatures - processes which decay into similar numbers and types of particles as four-top-quarks - are together labeled background processes, and must be well understood and characterized in order to differentiate them from true four-top events.

Particle	Cross Section σ [pb]
t	216.99
$t\bar{t}$	831.76
$t\bar{t}t\bar{t}$	0.01197
Higgs Boson	55.24

Table 1: Cross section for top quarks and other particles [4][5]

2 Methods

In order to find the optimal cuts - the cuts which enhance the signal to background ratio and make finding four top events more likely - we first had to ensure that our background estimation techniques produced good results. While the data for this experiment comes from the *pp* collisions at the LHC, we first simulate expected outcomes given knowledge of other physics processes and the theory being tested. This simulated data comes from the Monte Carlo method (MC)[6]. This method takes the The background events were generated by PYTHIA8, and the signal events were generated MADGRAPH5 and PYTHIA8. These generators use the Lagrangian of the model (SM or theory) combined with the Feynman rules to simulate event distributions, which are then sampled via the MC method. Finally, the specific geometry and sensitivity of the detector is modeled by simulators such as GEANT4, taking into account pileup and other detector level events to reconstruct objects. This is the final step the simulated data goes through before it is used for analysis.

2.1 MC Correction Function

This MC generated background is then scaled via event weights. These weights account for the luminosity of the data, and other small weight adjustments to account for fluctuations in the simulation process. However, further corrections are needed, and we will use data to apply these corrections. Since the four-top signal is so small compared to the background, the ratio of the background MC to data, before selection criteria are applied, should be ≈ 1. A ratio plot shows that this is not the case. To correct for this, we can fit the ratio of the two distributions with a polynomial function, and use the resulting function (hereafter the correction function) as an additional weight on the MC generated distributions.

We choose H^T_{all} to find the correction function. H^T_{all} is the ideal choice because it is the most in need of correction (there is the largest signal background ratio (see Figures 3-6). The effectiveness of the H^T_{all} correction function is then verified by comparing the ratios for the other kinematic variables E^T_{miss} and jet_{pt}.

2.2 Key Variables

The variables chosen to analyze the background all have theoretical motivation. Starting with N_{jets} and N_{bjets}, we expect high numbers of both due to gluon radiation. For four-top production, we expect 4 bjets. Some considerations when looking at b-tagged jets, as opposed to bjets, is the tag efficiency - for various reasons, bjets can be mistagged, so we are more liberal in the number of b-tagged jets that we pre-select for. We also expect a high overall number of jets for all of the final states from four-top-quarks.

H^T_{all} gives even more information. H^T_{all} is the total energy from the jets, so we also look for high H^T_{all} values to indicate that many jets are in the final state. H^T_{all} can give more resolution that just a number of jets/bjets. We also study E^T_{miss}, which is indicative of neutrino production. A single top quark can decay into a W boson which then decays into, among other things, a neutrino. Thus, we analyze E^T_{miss} because many neutrinos would cause high E^T_{miss} values, which could indicate a four-top event.

2.3 Cut Selection

Once the MC distributions have been reprocessed with the correction function, we can then determine the proper event selection criteria (cuts) that will maximize the signal to background ratio. To first order, we can look at distribution of signal and background for key variables, normalized to unity, and qualitatively find cuts that have the most separating power. This separating power is a measure known as significance, and can be closely approximated with Equation (1).

$$\text{Significance} = \frac{N^{sig}_{events}}{\sqrt{N^{sig}_{events} + N^{bkg}_{events}}} \tag{1}$$

The number of events in each histogram can be represented as a function of the cuts on each variable, as

$$N_{events} = N_{events}(cuts) \tag{2}$$

As such, Significance is also a function of the specific cuts we choose. Thus, we can choose cuts that maximize Significance, and solve for

$$\operatorname*{argmax}_{\text{cuts}} \text{Significance(cuts)} \tag{3}$$

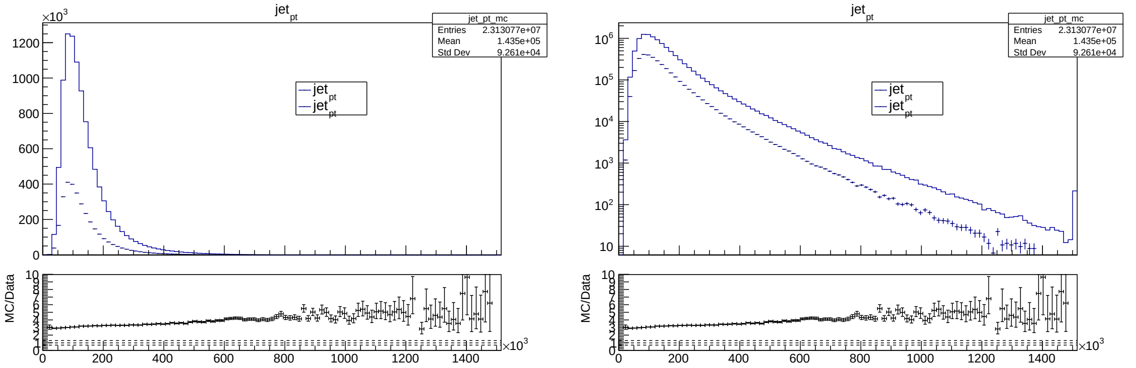
To find the maximal cuts, we first sweep through all cuts for all variables and find the cut with the overall most significance. After making that cut and refilling the histograms with that cut in place, we recalculate significance and find the next maximal cut on the remaining, uncut variables. We repeat this process until we've established a cut on all variables. A final check of significance is then done to check the overall improvement due to cuts.

3 Results

Below are the key results for this background charaterization. For all results, only events with N_{bjets} >2 are considered.

3.1 Initial Ratio Plots

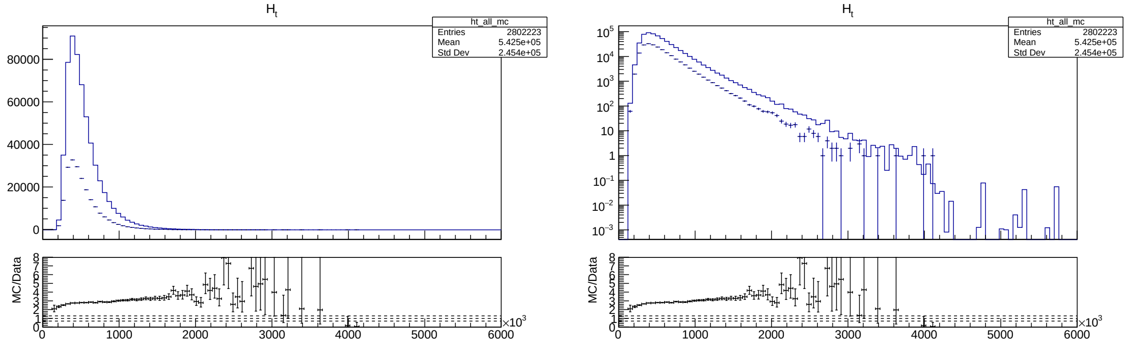
First are the plots of the MC and data distributions and ratios for the relevant kinematics variable and numbers of jets before the H_{all}^{T} correction function has been applied.



(a) Linear scale

(b) Log scale

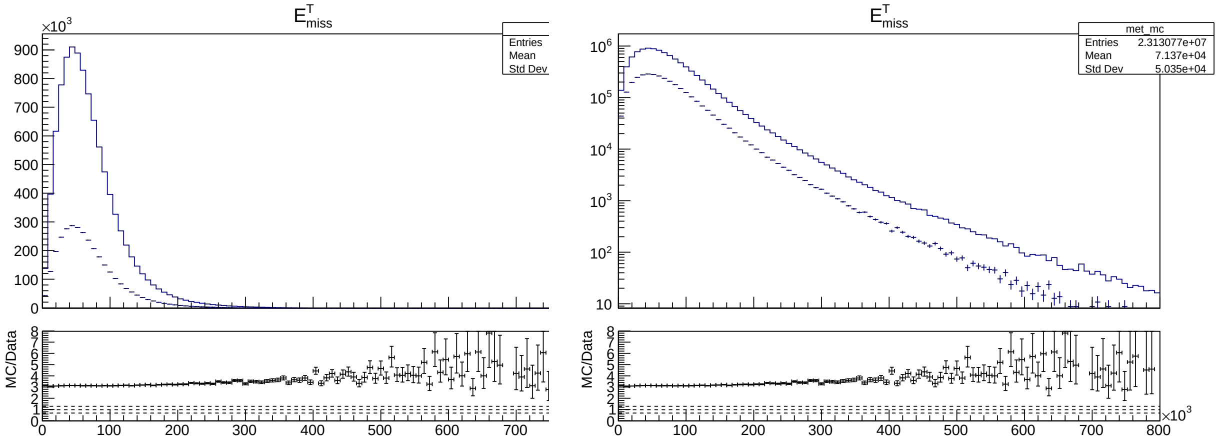
Figure 2: jet_{pt} Background and MC distributions



(a) Linear scale

(b) Log scale

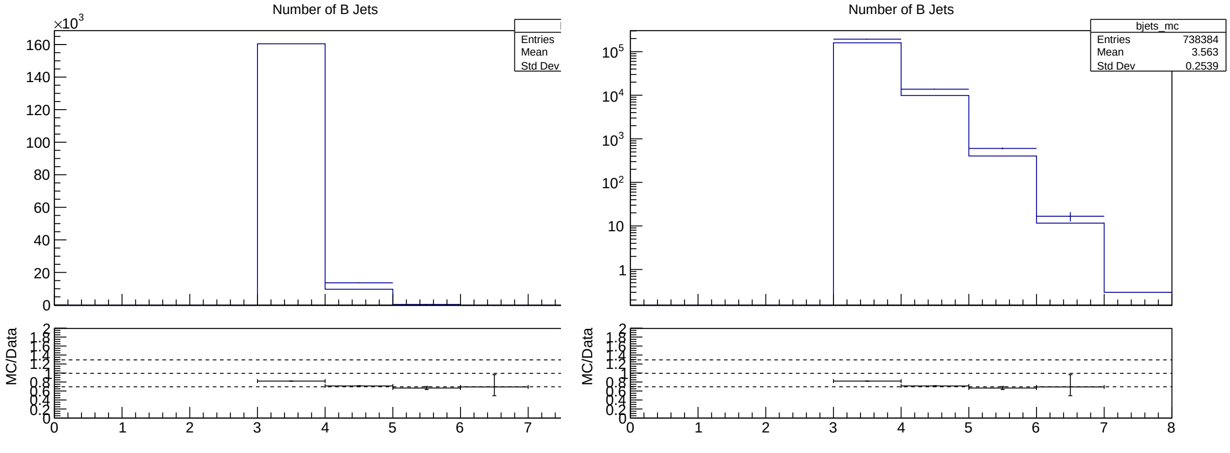
Figure 3: H_{all}^{T} Background and MC distributions



(a) Linear scale

(b) Log scale

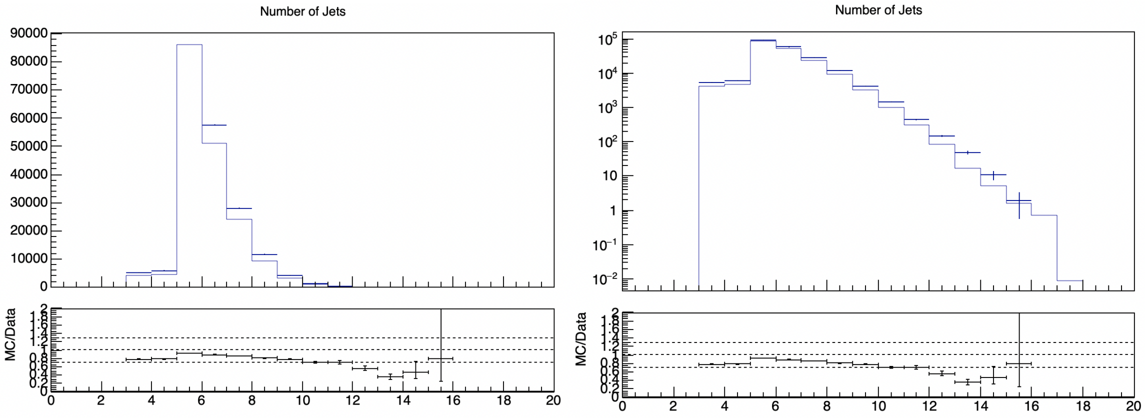
Figure 4: E^T_{miss} Background and MC distributions



(a) Linear scale

(b) Log scale

Figure 5: N_{bjets} Background and MC distributions



(a) Linear scale

(b) Log scale

Figure 6: N_{jets} Background and MC distributions

3.2 Fitting H^T_{all} Correction Function

Below is the ratio of MC/Data for the H^T_{all} kinematic variable.

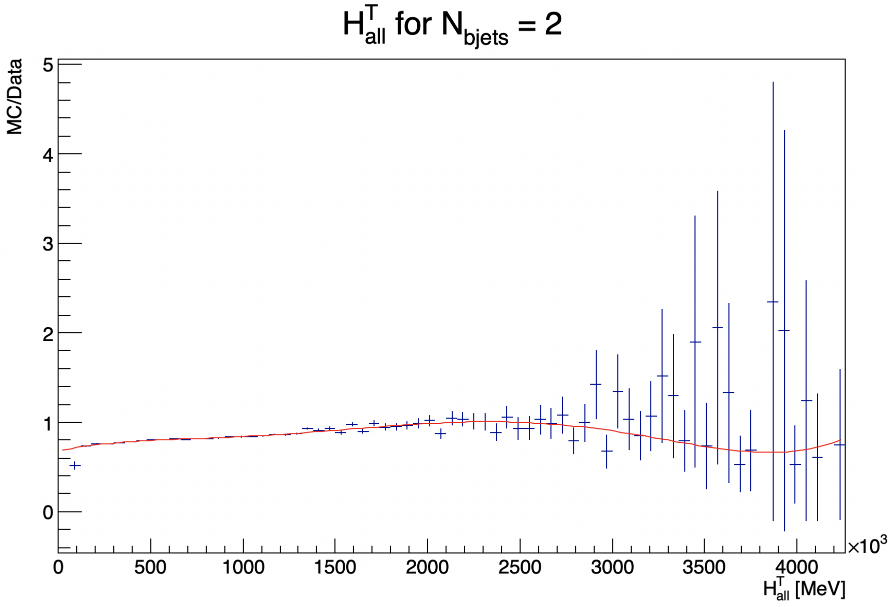
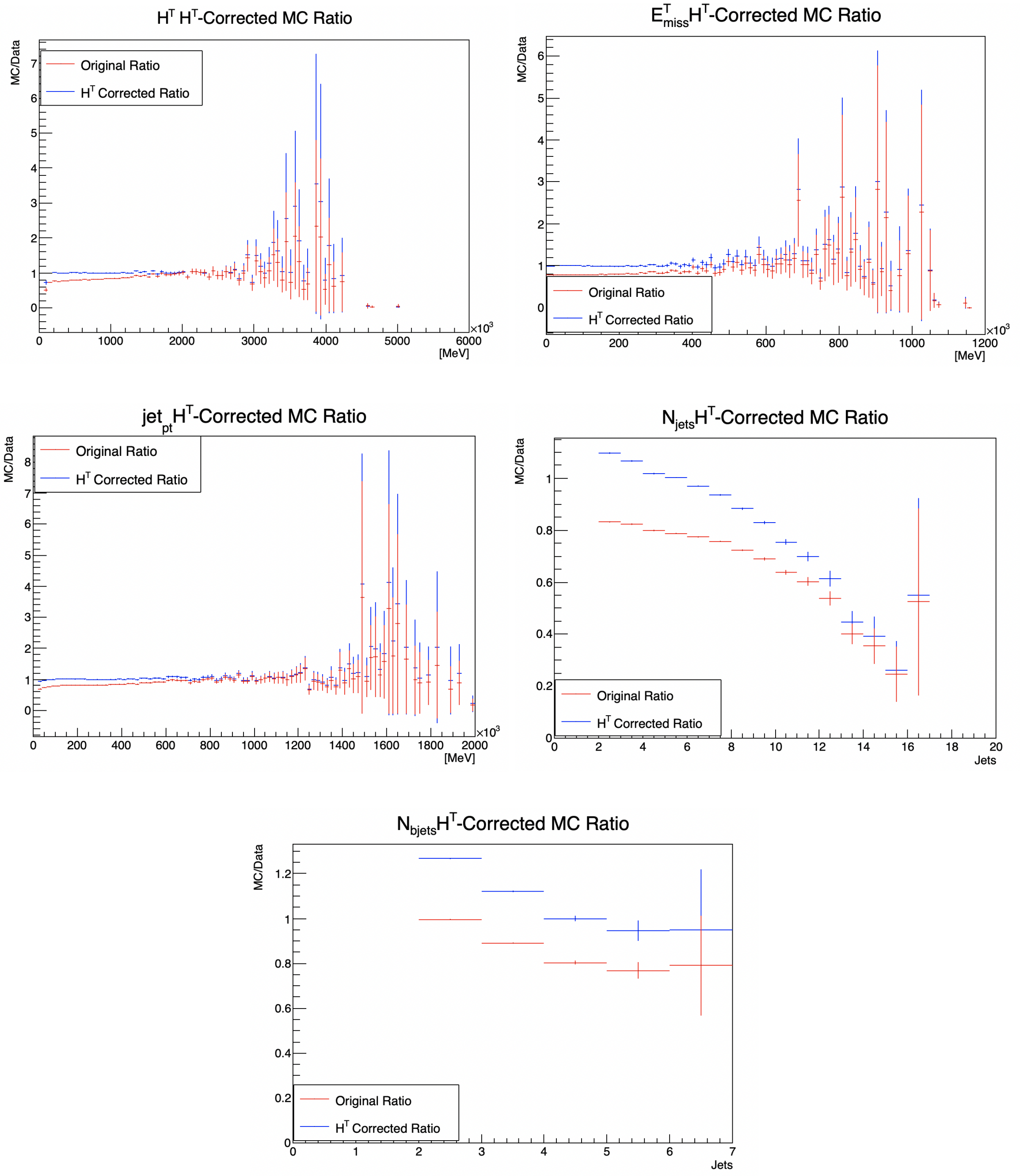


Figure 7: H^T_{all} ratio plot with 5th Degree Polynomial Fit

The value of the fitted function is

$$\frac{\text{MC}}{\text{Data}} = 1.86\text{e-}32(H^T)^5 - 1.73\text{e-}25(H^T)^4 + 5.41\text{e-}19(H^T)^3 - 6.98\text{e-}13(H^T)^2 + 4.77\text{e-}7(H^T) + 0.67$$

After applying the correction function, the ratio histograms are as follows, with the blue histograms representing the corrected histograms and the red the uncorrected histograms:


H^T H^T-Corrected MC Ratio

E^T_miss H^T-Corrected MC Ratio

jet_pt H^T-Corrected MC Ratio

N_jets H^T-Corrected MC Ratio

N_bjets H^T-Corrected MC Ratio

Figure 8: MC/Data Original and H^T_{all}-Corrected Ratios

3.3 Cut Significance

The following plots show how the significance (as defined in Equation (1)) changes with a the lower cut value on the variable in question.

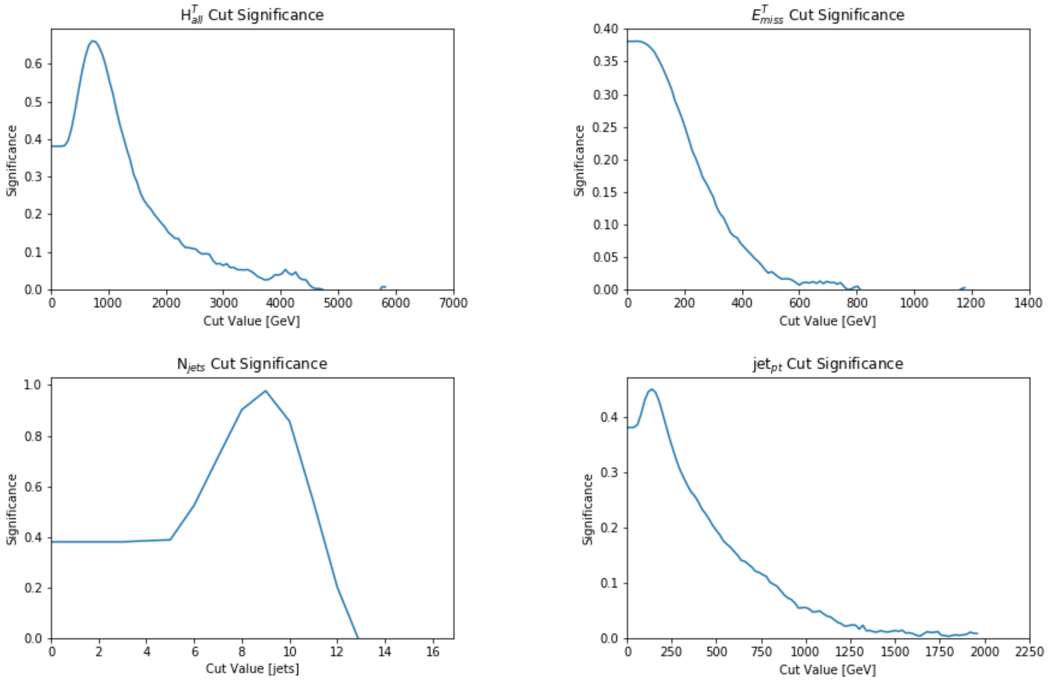


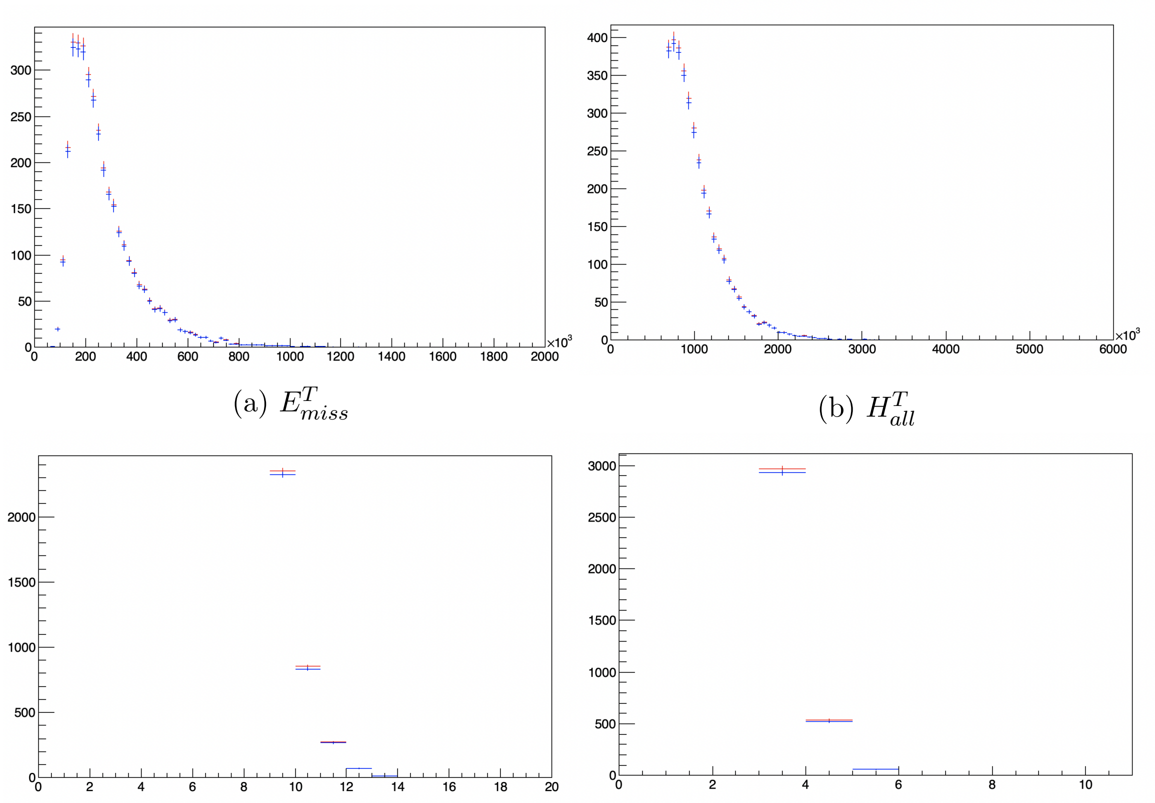
Figure 9: Significance

Cut variable	Significance			
	Prelim Cut	After 1 Cut	After 2 Cuts	After 3 Cuts
N_{jets}	**0.98**	–	–	–
H^T_{all}	0.66	**1.00**	–	–
Jet_{pt}	0.45	0.98	**1.01**	–
N_{bjets}	0.42	–	–	–
E^T_{miss}	0.38	–	–	–

Table 2: Significance of Optimal Cut at each step. (–) indicates no significance-improving cut on that variable.

3.4 Signal/Background Plots

The following plots show the signal and background MC generated distributions after the maximal significance cut on all variables. All are plotted in log scale.



(a) E^T_{miss}

(b) H^T_{all}

(c) N_{jets}

(d) N_{bjets}

Figure 10: Post-Cut Signal and Background Plots

4 Discussion

This discussion section will follow the order of the Results section and address some of the key points therein.

There were many choices regarding the correction function. The first was the variable upon which to do the first order correction. H^T_{all} was chosen for this first correction because it is a widely encompassing kinematic variable. There are many sub-processes that go into getting a total value for H^T_{all}, by creating an overall blanket correction to such a variable, we are essentially correcting for many of the underlying issues in the MC simultaneously.

The next step involved choosing the order of the function. This was done qualitatively, by looking at the how well the model fit in the critical regions and how sensitive the fit was to outliers in the background. A third order polynomial was sufficient to capture the general form of the ratio, but failed to account for the low event bins in the high H^T_{all} region. The fifth order polynomial showed much greater agreement in the high H^T_{all} range, and that polynomial was used in the analysis. Note that for the fit, the histogram was extended such that there was not an overflow bin.

To check the validity of the correction function, it was applied to all of the other variables and then the MC/Data ratio was taken again. We expect that the ratio has moved closer to unity, and the plots in Figure 8 (where red is uncorrected and blue is corrected) show that this is the case. The other ratios are not as close to one as H^T_{all}, but there are diminishing returns for adding an additional correction function specifically for each variable present, so we consider only the correction from H^T_{all}.

Next, we work to establish the maximally significant cuts. The Significance function in Equation 1 is a close approximation to the true significance, and is monotonic with the true significance, so maximizing this (much simpler to calculate function) will also maximize the true significance. We only consider cuts that act as a lower bound since theory (and the signal distribution) tells us that four-top-quark production is generally high in E^T_{miss}, H^T_{all}, N_{jets}, and N_{bjets}. We cut one variable at a time instead of all at once (see 2.3 for more details) because after each cut, significance could change considerably once the events that do not meet the cut criteria are removed. We can see this happening after the first cut on N_{jets}. Before that cut, each variable had cuts that would improve the significance of the resulting distributions. After the cut, however, only two of the variables still had significance-improving cuts, and those cuts improved significance only marginally.

Note that the significance for some cuts is actually below zero. This is an artifact from the signal generating process. Some signal events have negative weight, but when enough events are combined per bin, this negative weight usually doesn't show up, but appears here in bins with low event counts.

From Table 2, we can see that after each cut, significance improves. The significance without any selection criteria is 0.20, so even criteria on the weakest variable (E_{miss}^{T}) does provide some improvement over the base significance. N_{jets} shows the largest improvements, and that leads to the first criteria being a lower limit on the number of jets. The other key variables follow in turn, but no criteria affects the significance at the same magnitude as the first cut on the number of jets. After that, all other criteria only offer marginal improvements. This shows how crucial N_{jets} is in the search for four-top-quark events.

5 Conclusion

A data-driven method to correct the background estimation, and put maximally significant cuts on key variables, is presented here. The final selection criteria are presented in Table 3. From these selection criteria, the final significance is 1.0077, and we expect to see 60.22 signal events and 3511 background events in 140 fb^{-1} of data.

Further work must be done before this four-top-quark analysis is complete. The next step is to quantify systematic uncertainties. Background estimation is just one component of a full high energy physics analysis, which comes while the group is

Table 3: Maximally Significant Selection Criteria

Variable	**Selection Criteria**
N_{jets}	$>$ 9 jets
N_{bjets}	$\geq$ 2 jets
H^T_{all}	$>$ 660,000 MeV
jet_{pt}	$>$ 60,000 MeV

still blind to the data so that the estimation remains unbiased. Once the group is unblinded to the data, the group will search for signal events inside the cut regions defined in Table 3, and use the number of signal events to measure the cross section for four-top-quarks and test whether it is consistent with SM expectations or hints at new physics processes.

References

[1] *LHC Facts and Figures*. URL: https://home.cern/resources/brochure/accelerators/lhc-facts-and-figures.

[2] *Producing four top quarks at once to explore the unknown*. Nov. 2018. URL: https://atlas.cern/updates/physics-briefing/producing-four-top-quarks-once-explore-unknown.

[3] CMS Collaboration. "Search for standard model production of four top quarks in final states with same-sign and multiple leptons in proton-proton collisions at $\sqrt{s} = 13$ TeV". In: (2019).

[4] ATLAS Collaboration. "Combined measurements of Higgs boson production and decay using up to 80 fb^{-1} of proton–proton collision data at $\sqrt{s} = 13$ TeV collected with the ATLAS experiment". In: (2019).

[5] ATLAS Collaboration. "Search for four-top-quark production in the single-lepton and opposite-sign dilepton final states in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector". In: *Phys. Rev.* D99.5 (2019), p. 052009. DOI: 10.1103/PhysRevD.99.052009. arXiv: 1811.02305 [hep-ex].

[6] *Tutorials on FeynRules and MadGraph*. URL: http://susy.phsx.ku.edu/~kckong/KWS2015/KWS2015.pdf.