Optimizing $\mathrm{t\bar{t}}$ Purity for Shower Deconstruction

Cameron Racz

Supervisor: Dr. Joseph Haley

Oklahoma State University

December 15, 2017

Abstract

This paper describes my research as part of the Oklahoma State University experimental High Energy Physics Group. My work was based on particle collision events that resulted in hadronic top quarks and an algorithm called Shower Deconstruction that can be used to find those quarks from data gained by using the ATLAS detector. Shower Deconstruction uses a likelihood ratio χ to distinguish signal jets (top jets) from background jets, but systematic uncertainties for the algorithm have not been calculated. The goal of this project was to find the best ways in which to reduce the background so that future studies can more accurately calculate these uncertainties. Signal and background events were generated with Monte Carlo simulation, and a selection was used to produce a sample of semileptonic $t\bar{t}$ events so that the hadronic top jets could be studied. The background that passes the selection consisted of Z, W, single top, and diboson events. For signal and background, I looked at the kinematics of the jets that were the best candidates for the hadronic top and explored what cuts could be made to reduce the background. I found that it can be reduced best by applying cuts on the spacing between the subjets of the candidate top jet. Cutting out events with a minimum spacing (ΔR) between the three highest transverse momentum subjets less than or equal to 0.1125 produced a background rejection rate of 9% without affecting the ratio between signal and Poisson uncertainty in the total number of events. Also, applying the same type of cut at 0.3375 gave a background rejection rate of 65% with the best signal significance, but this came at the cost of a signal acceptance rate of only 72%.

1 Introduction & Background

The LHC is used to collide protons together and study the different particles that are produced. These collisions are known as "events," and the resulting particles are detected by large detectors; one of which is the ATLAS detector used by the ATLAS Collaboration. Different particles are seen as energy deposits in different layers of the detector. Quarks and gluons created in an event, however, will hadronize before reaching the detector. These particles have color charge, so by the principle of color confinement they must cluster together to form colorless hadrons (hadronization). This process results in a shower of hadrons moving in the same direction as the quark/gluon released in the event. In the detector, this shows up as a collimated deposit of energy in one region of the detector called a "jet" [1]. Since there are many different particles that can produce jets, and jets are not as well defined as any single particles, certain algorithms are necessary to distinguish exactly what particles they originated from. One such algorithm is called Shower Deconstruction and can be used to find jets that were produced by top quarks.

Hadronically decaying top quarks are a particularly interesting example of jet-producing particles. They have a mass of about 172.44 GeV/c^2, so it is the heaviest known fundamental particle. Top quarks have a lifetime much shorter than the timescale of the strong force, so they decay before undergoing hadronization like other quarks [2]. This means that it is actually possible to find individual top quarks from the jets that they produce, rather than some hadron that contains a top quark.

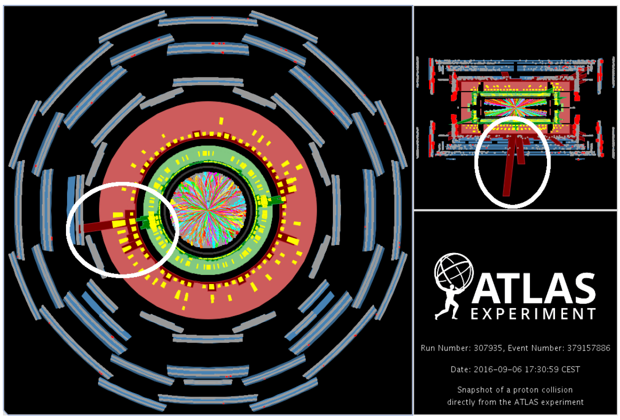


Figure 1: Example of an event that produces a jet (circled). This image shows the ATLAS detector as viewed parallel to the beam (left) and a perpendicular view (right)

The most common way in which top quarks are produced in the LHC is as a top/anti-top pair rather than just one top quark. Since an antiparticle is noted with a bar over the particle's symbol, we call this a $t\bar{t}$ (t t-bar) event. These events are often used for studies of top quarks because of their abundance. A semileptonic $t\bar{t}$ event involves one top that decays into a lepton, a neutrino, and a quark, and another top that decays hadronically. A hadronically decaying top quark will produce a b quark and a W boson, where the W then decays into two other quarks. One of the tops in a semileptonic $t\bar{t}$ event will be an antiparticle, but whether it is the leptonic or hadronic decaying top is not an important detail. A visual example of a semileptonic $t\bar{t}$ event is shown in Figure 2.

Semileptonic $t\bar{t}$ events are often preferred because they have a hadronic top to study and the presence of a lepton makes it easy to distinguish from multijet events. Multijet events are composed of many jets and could be identified as ttbar events. Leptons are not produced in such events, so requiring to have a lepton in the final state will ensure that multijet events are not identified as ttbar events. This greatly increases the purity of events with ttbar candidates.

In addition, the purity of the $t\bar{t}$ sample can be increased by identifying jets coming from a hadronic top quark decay. This can be archieved by using algorithms that check certain criteria and characteristics of the jets. As mentioned before, Shower deconstruction is one of such algorimths. A hadronically decaying top results in three jets (Fig. 2), except in the case of a top with a very high momentum where these three jets are produced

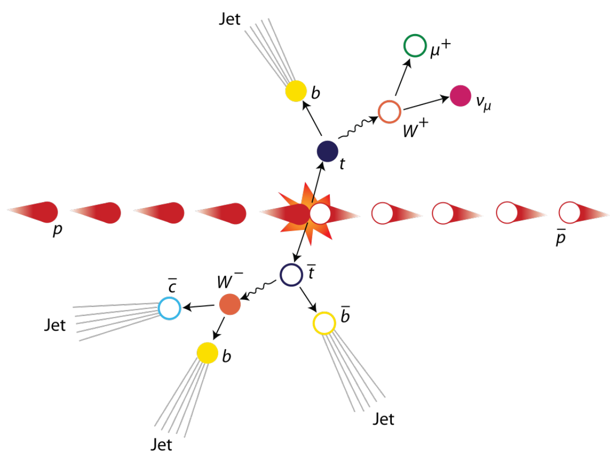


Figure 2: Example of a semileptonic $t\bar{t}$ event. The top quark is shown as the leptonic decay and the antitop as the hadronic decay.

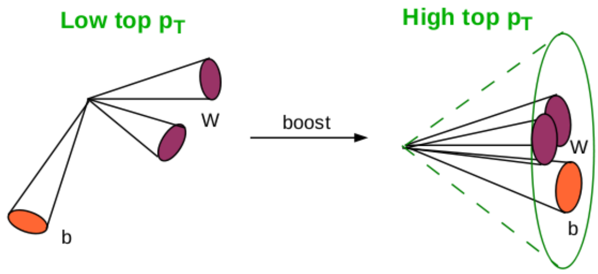


Figure 3: Comparison of a top quark decay to a boosted top decay. The green cone shows the large jet, the orange is the subjet from the b, and the purple shows the subjets from the quarks that come from the W decay.

very close together. This topology is known as a "boosted" top. For such configurations, a large cluster of energy deposited in one region of the detector is seen. This cluster is called a "large jet" and the jets coming form the quarks are called "subjets" since they are reconstructed in the same "large jet". An example of this is shown in Figure 3. The figure shows how we would normally see the products of a top quark as being very spread apart, but as particle collisions at the LHC reach higher energies, they will get closer together and harder to differentiate. This is why Shower Deconstruction is necessary for identifying top quarks.

Subjets are passed to shower deconstruction to determine if the large jet is likely coming from top quark decays. The four-momenta of the subjets $\{p\}_N$ are then compared to possible shower histories for many different particles to calculate the probability that the observed configuration $\{p\}_N$ matches those shower histories. Specifically, Shower Deconstruction creates a function

$$\chi(\{p\}_N) = \frac{P(\{p\}_N|S)}{P(\{p\}_N|B)}$$

where $P(\{p\}_N|S)$ is the probability density that the observed jet configuration matches a shower history of a top jet, and $P(\{p\}_N|B)$ is the probability density that the observed jet matches a shower history of a non-top jet. So the χ function is the likelihood ratio that a jet is a top or non-top jet. Top-like jets tends to have positive values of χ, whereas non-top jets lead to negative values of χ, therefore by applying a cut on this variable, the relative fraction of top jets can be increased[3]. It should be noted that, like any other algorithm, Shower Deconstruction will never be 100% efficient at making this separation due to the randomness of particle decays and mismeasurement in the detector. So some non-top jets could get similar χ values as signal and vice versa.

Shower Deconstruction has systematic uncertainties associated to it, but these have yet to be calculated. In order to do this, however, it is important to get a sample of jets that consists of as much signal and as little background as possible; so the sample needs to have a high signal purity. This was the goal of my research; I aimed to find the best ways in which to reduce the background of a sample to maximize the signal purity so that future studies can make better calculations of the systematic uncertainties of Shower Deconstruction.

2 Methods

In this study, the more specific goal was to look at the kinematics of signal and background jets and find an aspect in which both types differ in such a way that it is possible to remove most of the background by making a cut. Standard Model events created by a Monte Carlo (MC) simulation were used to produce signal and background. Also, real data recorded by the ATLAS detector in 2015 and 2016 was used, and this combination of data and simulation made it possible to make comparisons to ensure that the simulation agrees well with the data for each kinematic variable. Also, if the simulation agrees well with the data, then this shows that the data is most likely composed of the same amounts of signal and background events as the simulation.

In order to produce the right events for data and MC, all events had to pass the selection described below [4]. In the criteria, p_t is transverse momentum (relative to the beam axis), ϕ is the azimuthal angle around the beam, η is an angular coordinate measured relative to the beam axis.[1] It is defined as

$$\eta \equiv -\ln(\tan \tfrac{\theta}{2}).$$

R is a distance parameter used to measure the distance between signatures in the detector, as well as the radius of a jet. The ΔR between two signatures a and b is given by

$$\Delta R(a, b) = \sqrt{(\eta_b - \eta_a)^2 + (\phi_b - \phi_a)^2}.$$

An example of a "large-R" jet would be the full large jet produced by a top quark, and a "small-R" jet would be a subjet produced by one of the final state quarks of the top quark decay. $E_{\mathrm{T}}^{\mathrm{miss}}$ is missing transverse energy that

[1] A polar angle of $\theta = \pi/2$ (along the direction of the beam) corresponds to $\eta = 0$, and the angle $\theta = 0$ (perpendicular to the beam) corresponds to $\eta = \infty$. η is used rather than θ because particle production does not depend on η, and, for highly relativistic particles, differences in η are approximately Lorentz invariant under boosts along the beam axis [5].

would be necessary to conserve momentum for an event. This is the energy of neutrinos since they pass through the detector without leaving any signature. Also, M_{T}^{W} is the transverse mass of the leptonic W boson candidate.

- Exactly one electron or muon with $p_t > 30$ GeV, $|\eta| < 2.47$, and $1.37 \nless |\eta| \nless 1.52$ (Since this is the calorimeter transition region)
- Electron fulfills the tight ID requirement or the muon fulfills medium ID quality
- $E_{\mathrm{T}}^{\mathrm{miss}} > 20$ GeV
- $E_{\mathrm{T}}^{\mathrm{miss}} + M_{\mathrm{T}}^{W} > 60$ GeV
- At least one small-R jet (j) with $p_t > 25$ GeV and $\Delta R(\mathrm{lepton}, j) < 1.5$
- At least one large-R jet (J) with $p_t > 200$ GeV and $|\eta| < 2.0$
- $\Delta\phi(\mathrm{lepton}, J) > 2.3$
- At least one b-tagged small-R jet

After passing this selection, the data and MC samples are mostly composed of semileptonic $t\bar{t}$ events. Some background is able to pass this selection though, and this is the background for this study. It consists of W boson, Z boson, diboson (WW, ZZ, WZ), and single top quark events. Also at this stage, the large-R jet in each event that is the best candidate for the hadronically decaying top is saved in the files. There is only one of these jets in every event of every type (data, signal, background). These are the only jets that were used in this research, so all plots were made using only the best candidate jets for the top quark. A new selection that I added for all of the plots created was that the large-R jet for the candidate top must have at least 3 subjets inside of it. This is because any jets with less than three subjets are likely to be background, and if they are signal events, it would be rather difficult to know that they were since signal jets always have three decay products.

What I first did was look at the χ distribution and the different kinematic variables of the large and small jets. Then, I made different combinations of 2-dimensional plots for these variables to try and find any trends or correlations that could show a distinct difference when compared between signal and background. If anything like this was found, I made cuts according to the difference, and calculated the new signal efficiency and background rejection for the remaining events. Since the χ distributions are plotted as $\log(\chi)$, any events with $\chi = 0$ will be removed when the plots are produced. For this reason I used the number of events in the χ distributions (omitting $\chi = 0$ events) to calculate the signal efficiency and background rejection. These efficiencies were calculated using

$$\varepsilon_n = \frac{I(\chi_n^{cut})}{I(\chi_n^{uncut})}$$

where $I(\chi_n)$ is the integral of the χ distribution for the n type of sample (signal or background). The background rejection was then calculated as $1 - \varepsilon_{bkg}$.

3 Results

I first plotted different kinematic variables for large and small jets to get a sense of what the jets were like physically. The mass and momentum of the large jets are shown in Figure 4. In these and all other plots that show data compared to MC samples, the MC samples are stacked on top of each other in order to see if the data matches the sum of all MC samples like it should.

From the p_t plot in Figure 4, it is apparent that the large jet p_t is not well described by the MC. This is known by those that produced these data and MC files, and it is just because there has not been a p_t correction made for the data produced recently in 2015 and 2016 [4]. The plots also do not show any systematic uncertainties since these have yet to be calculated. When looking at the kinematics of the subjets, only the three with the highest p_t were considered since these are the jets that carry the most (if not all) of the important information about the corresponding large jet. These subjets are known as the "leading" (highest p_t), "second leading" (second highest), and "third leading" (third highest) subjets. In the following plots they are noted by a subscript of 1, 2, and 3, respectively. So, as an example, ΔR_{23} is the ΔR between the second and third leading subjets. Figure 6 shows the χ values for the large jets from Shower Deconstruction.

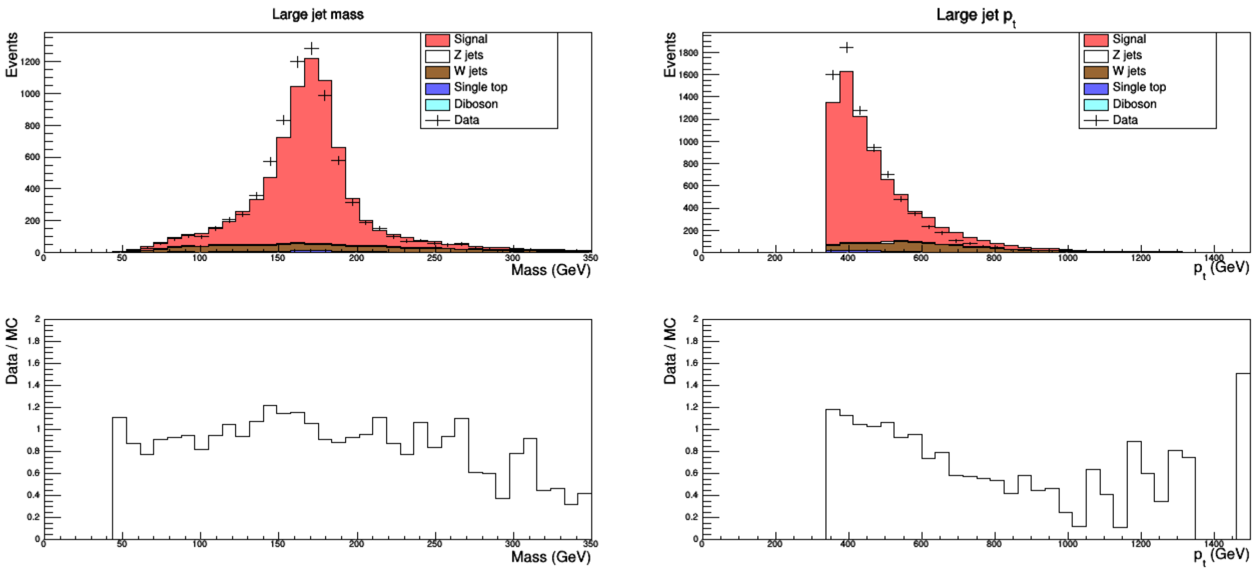
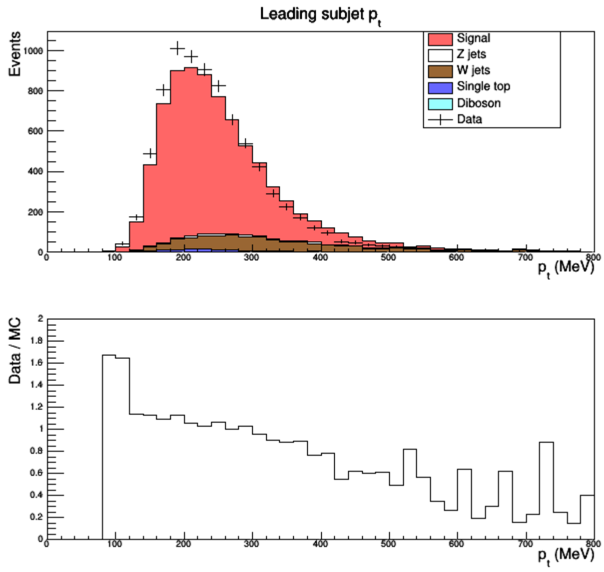
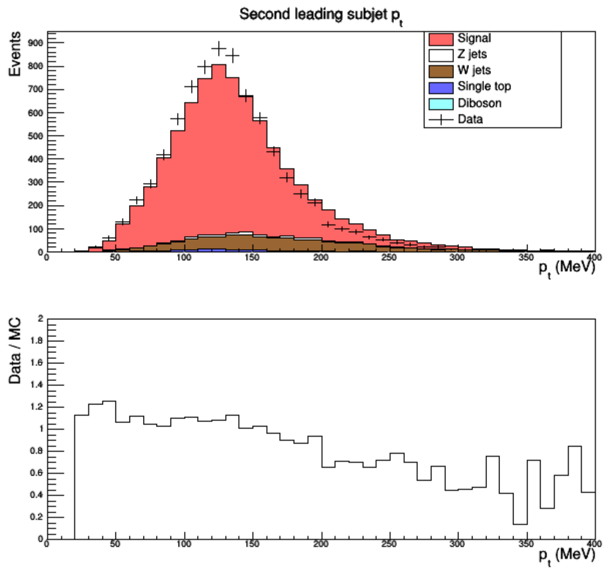


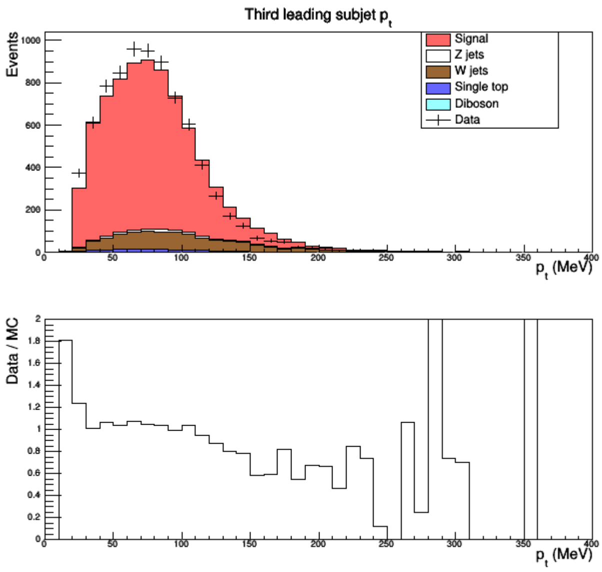
Figure 4: Mass and transverse momentum of the large-R jet. The bottom plots show the ratio of the data to the sum of all MC samples.



(a) Leading



(b) Second leading



(c) Third leading

Figure 5: p_t distributions of the first three subjets.

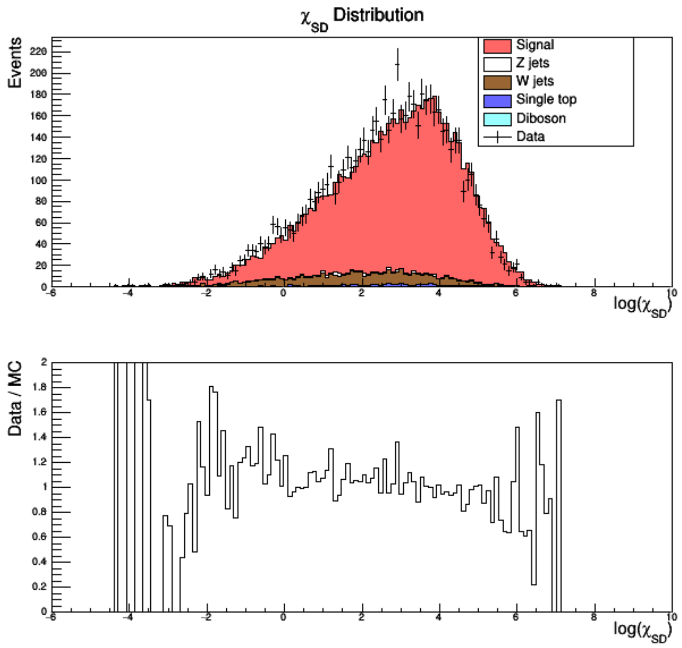


Figure 6: Distribution of χ for the candidate top quark large-R jets.

Figures 4 and 5 show that these variables are not useful for distinguishing background from signal since their distributions are spread over the same regions. The next possibility was to look at the separation of the three subjets, and these plots can be seen in Figure 7. ΔR_{min} is defined as

$$\Delta R_{min} \equiv \min\{\Delta R_{12}, \Delta R_{13}, \Delta R_{23}\}.$$

These plots show that the background for ΔR_{12} and ΔR_{min} is mostly concentrated below 0.2, so a cut on ΔR of about 0.2 could remove a majority of the background. This cut would also remove the relatively small concentration of signal events shown at that same position in the plot. The presence of this peak was a bit of a surprise since it is expected that the three constituents from the top decay are expected to be well separated. An idea for the reason behind this peak was that the b and W separated, but when the W decayed, those two quarks separated and one of the quarks moved close to the b. Unfortunately, the information necessary for determining which subjet originated from which quark was not present in the files, so it is not known exactly what this peak is from. This is a point that should be investigated further in future studies.

To get a better idea for how to apply this ΔR cut, or to find any other cuts, 2-dimensional histograms were produced to find any useful correlations in the kinematics of the jets. An example of these plots are shown in Figures

8 and 9. Figure 9 showed a possible way to cut out the background, and the results of that cut are shown in Figure 10. This proved to be a very ineffective cut due to the fact that, while the background is mostly in one region, it is still too spread out to cut out a large majority of the background without removing a large portion of the signal.

Some important 2D plots made for this study were those in Figure 11. This shows the same peaks as Figure 7 and how the peaks relate to the p_t of the large jet. I made a single cut at $\Delta R_{12} = 0.2$ which is seen in Figure 12. The signal plot in Figure 11 also shows a diagonal trend moving up and to the left. Because of this, I also tried two different diagonal cuts just below this trend in the signal. These cuts can be seen in Figures 13 and 14. The new χ plots corresponding to these cuts can be seen in Figures 15, 16, and 17. These were the most significant cuts that were worth mentioning, and while the diagonal cuts performed better than others, it seemed that the solid cut on ΔR was performing best. Because of this, I made a more formal search for the optimal place to cut on any of the mentioned ΔR values between the subjets. This is shown in Figures 18, 19, and 20. In these plots, S and B stand for the integrals of the signal and background in the corresponding plots from Figure 7, and each bin shows the values evaluated from that bin to the last bin. So for some bin k with maximum bin K, the signal purity is calculated as

$$\text{Purity}(k) = \frac{S\big|_k^K}{S\big|_k^K + B\big|_k^K},$$

the signal-to-noise ratio is

$$\text{Signal-to-Noise}(k) = \frac{S\big|_k^K}{\sqrt{S\big|_k^K + B\big|_k^K}},$$

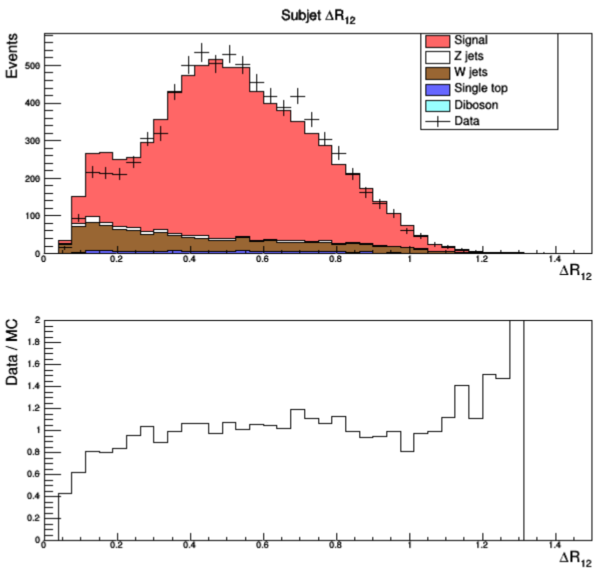
and the signal significance is

$$\text{Significance}(k) = \frac{S\big|_k^K}{\sqrt{B\big|_k^K}}.$$

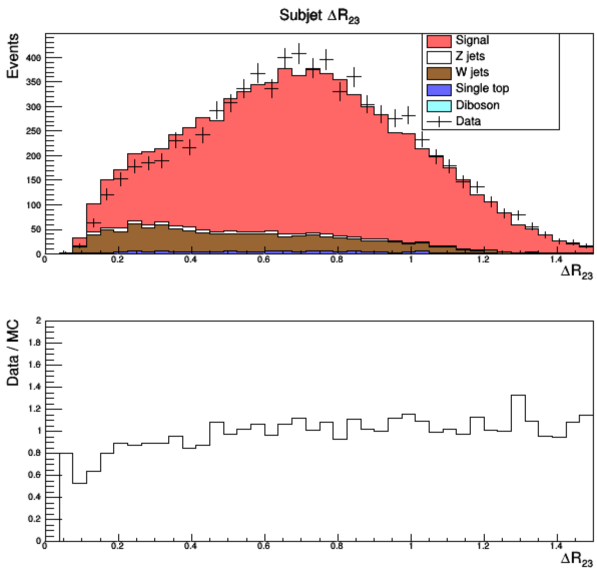
The statistical uncertainty in the signal or background is just the square root of the integral of that data type, so the signal-to-noise ratio shows the signal as a fraction of the statistical uncertainty, or “noise”, in the total MC. This tells you how large your signal sample is compared to the uncertainty in that sample. Finding the maximum of this plot shows us the ΔR cut that results in the portion of signal that is largest when compared to the uncertainty. The signal significance is the signal as a fraction of the uncertainty in the background, so this can show you how certain you can be about whether or not your signal exists; how significant your signal is. If it is within the uncertainty of the background, then it is impossible to differentiate a signal from a fluctuation of the background. This isn't the focus of this project since we know $t\bar{t}$ events exist, so this quantity isn't the most useful for this study. The signal purity is the signal as a fraction of the total MC, so it just tells you how much of your MC is signal. This also isn't exactly the best way to find our cut since only maximizing this quantity will greatly reduce our sample size and thus increase statistical uncertainties. What we want to do is find the bins that give us the highest signal-to-noise ratio, and then find which of those bins has the highest signal purity. This way we can increase the signal purity without increasing any statistical uncertainties.

The signal-to-noise ratio plots never increase, but they do start as a plateau at their highest point before decreasing. So at this plateau, we can make a cut at any of those points without affecting the signal-to-noise ratio. Looking at these same bins in the signal purity plots, we see that the highest signal purity happens at the last bin of the plateau in the signal-to-noise ratio plots. So the best cuts will happen at the last bin where the signal-to-noise ratio is at maximum. This gives us a best place to cut ΔR_{12}, ΔR_{23}, or ΔR_{min}. However, the overall best cut is whichever produces the highest signal purity. From the figures, it is apparent that cutting ΔR_{min} at the last bin that has the highest signal-to-noise ratio is the best cut. When making the cut, I applied it to the leftmost side of that bin since that bin corresponds to taking the integrals from the leftmost point on. This way I still keep all of the events in that bin. So I found that the optimal place to make a cut to increase the signal purity without changing the signal-to-noise ratio is at $\Delta R_{min} = 0.1125$. As shown in Figure 21, this cut creates a 9% background rejection rate with 99% signal acceptance.

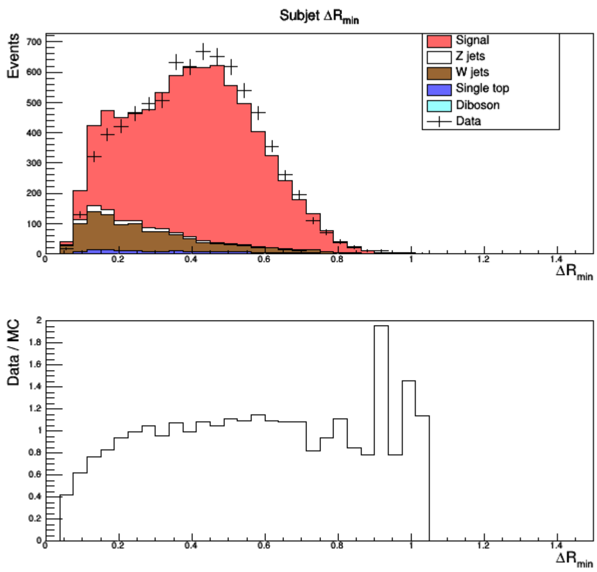
I also investigated what making a cut at the highest point in the signal significance plots would produce for signal acceptance and background rejection. I went through a similar process as before and found that cutting ΔR_{min} was still the best variable to cut on. The peak in signal significance that produced the highest signal purity was in the ΔR_{min} plots in Figure 20. The optimal cutting point was at $\Delta R_{min} = 0.3375$, and as shown in Figure 22, it produced a background rejection of 65% but this came at the cost of a signal acceptance of 72%, and this in turn significantly reduces the signal-to-noise ratio.



(a) Subjets 1 and 2

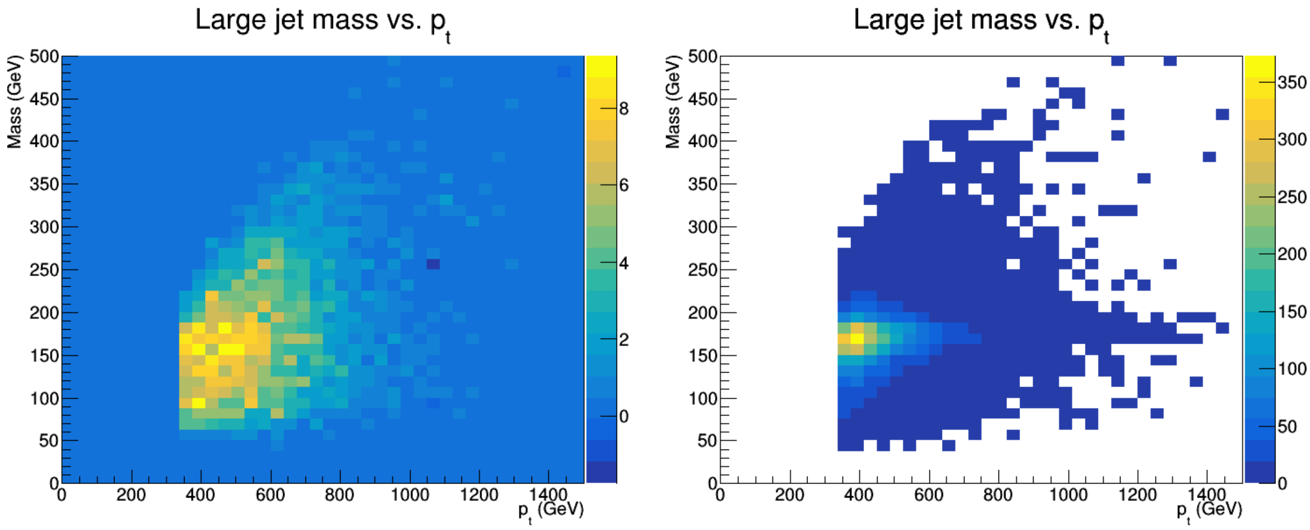


(b) Subjets 2 and 3



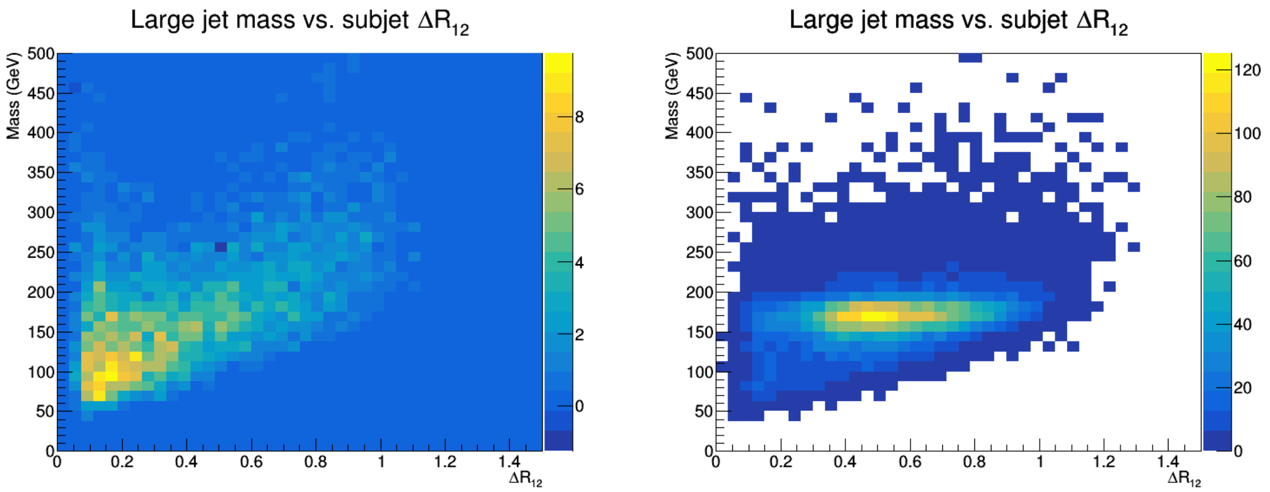
(c) Minimum between all three subjets

Figure 7: ΔR distributions of three different combinations between the first three subjets.



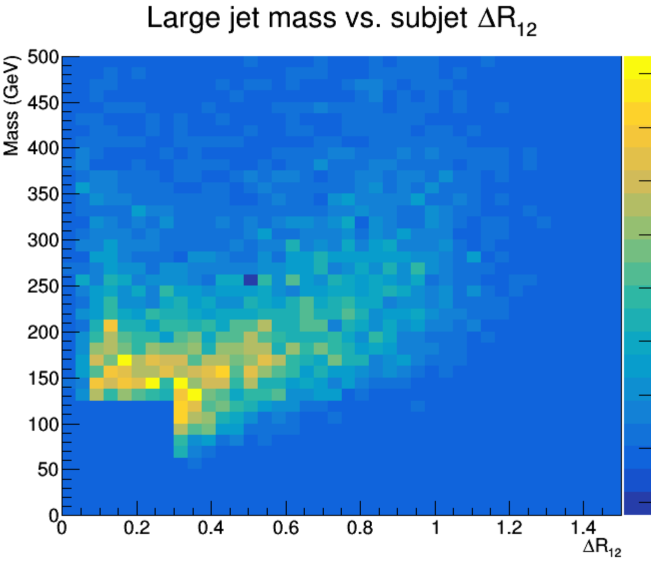
(a) Background (b) Signal

Figure 8: Mass vs p_t for the large jets in signal and background.

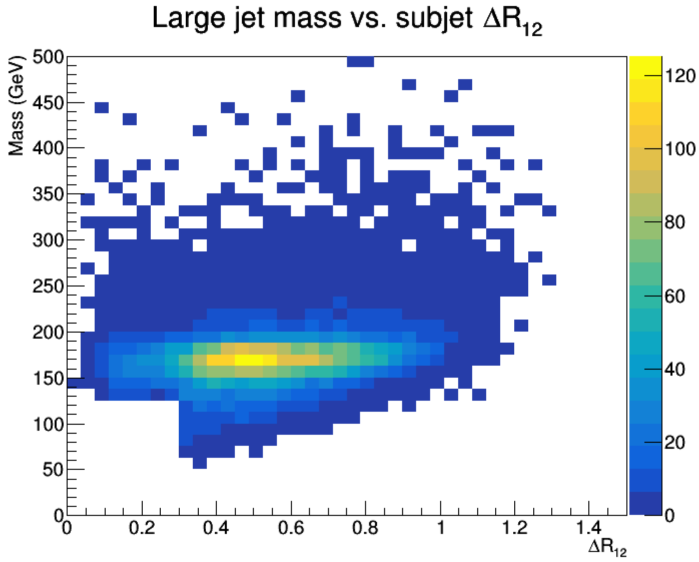


(a) Background (b) Signal

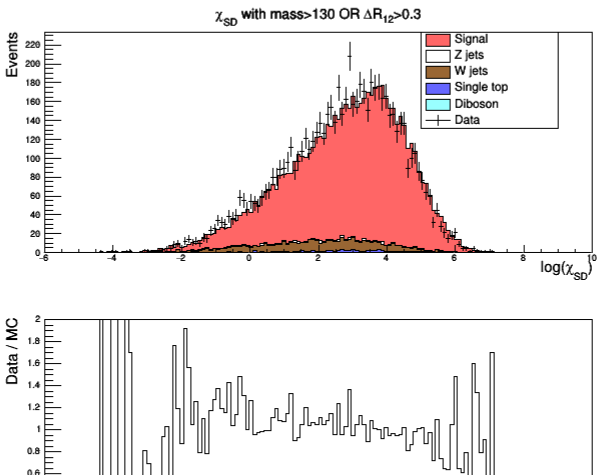
Figure 9: Mass vs ΔR_{12} for the large jets in signal and background.



(a) Background

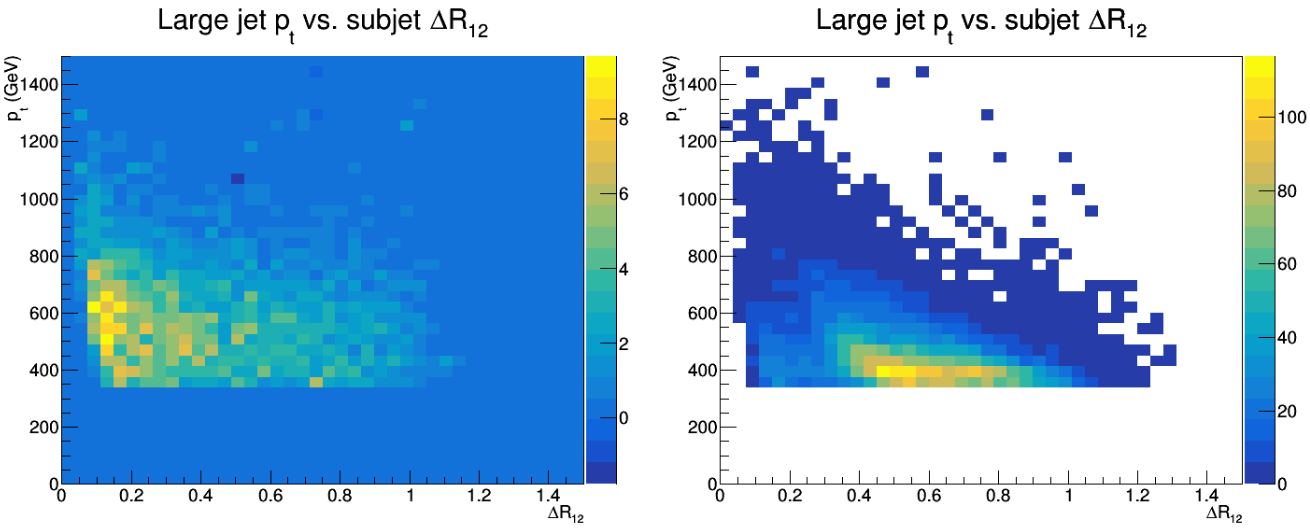


(b) Signal



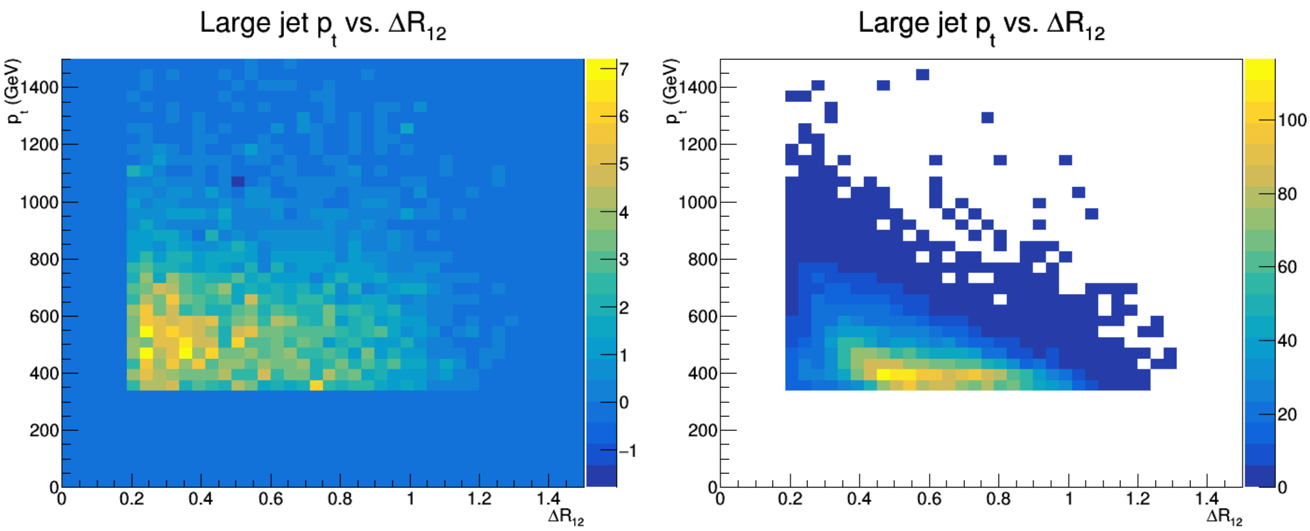
(c) χ after mass cut of 130 GeV and ΔR_{12} cut of 0.3.

Figure 10: Results of cutting the mass and ΔR_{12}. This cut gave $\varepsilon_{sig} = 0.99$ and $1 - \varepsilon_{bkg} = 0.04$.



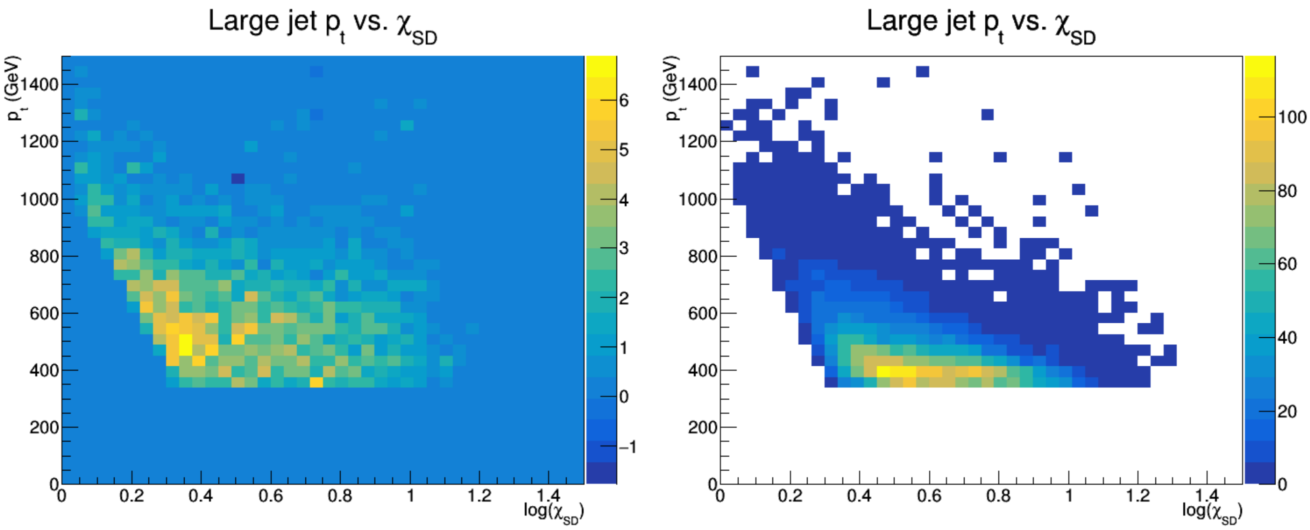
(a) Background (b) Signal

Figure 11: Large jet p_t vs ΔR_{12}



(a) Background (b) Signal

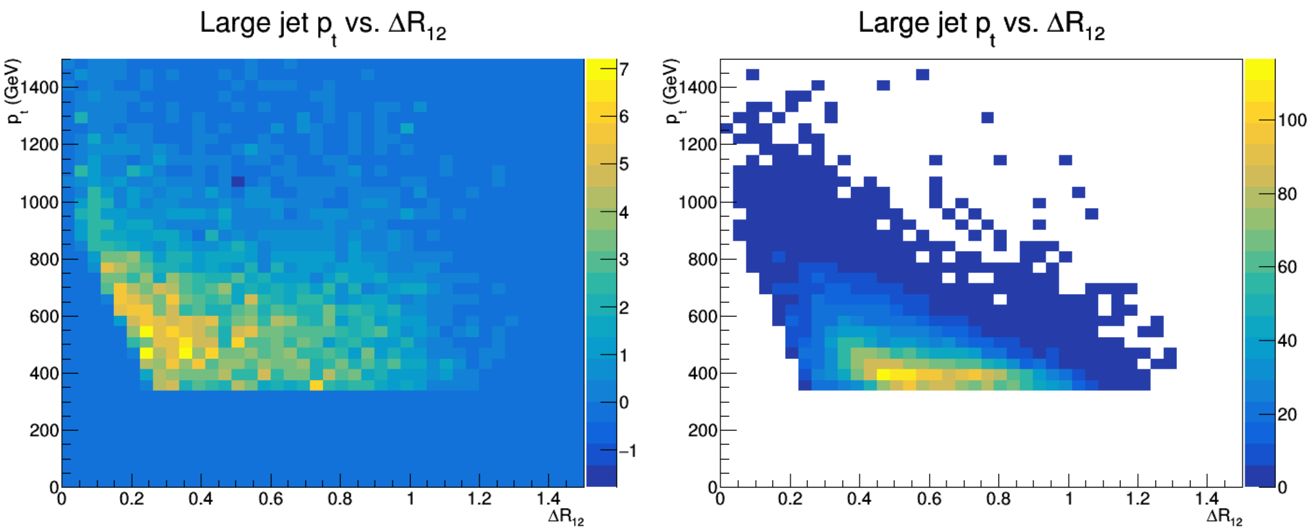
Figure 12: Large jet p_t vs ΔR_{12} after cutting all events with $\Delta R_{12} < 0.2$



(a) Background

(b) Signal

Figure 13: Large jet p_t for all jets with $p_t > -2200(\Delta R_{12}) + 1100$.



(a) Background

(b) Signal

Figure 14: Large jet p_t for all jets with $p_t > -2625(\Delta R_{12}) + 1050$.

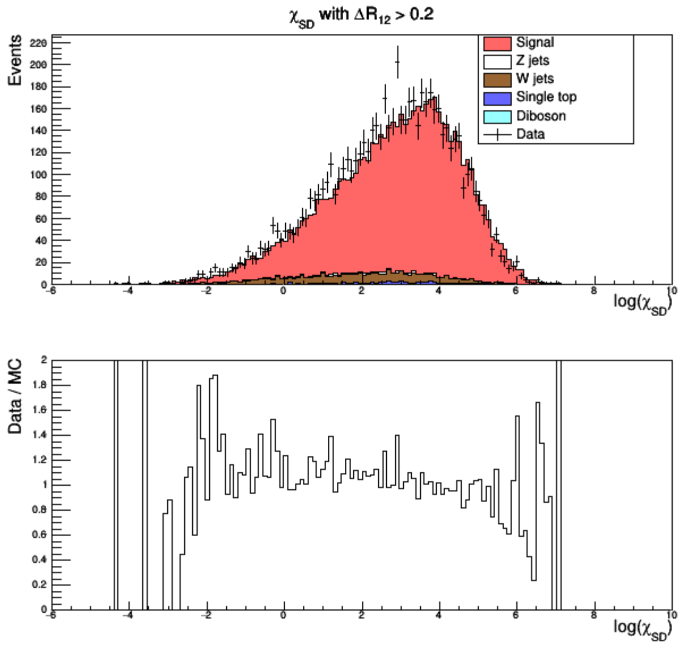


Figure 15: χ for only $\Delta R_{12} > 0.2$. This cut gave $\varepsilon_{sig} = 0.95$ and $1 - \varepsilon_{bkg} = 0.22$.

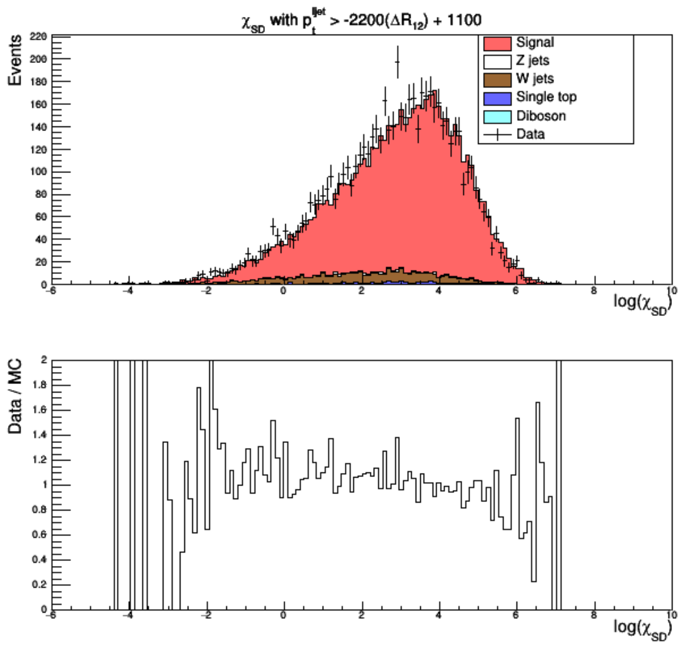


Figure 16: χ for Large jet $p_t > -2200(\Delta R_{12})+1100$. This cut gave $\varepsilon_{sig} = 0.92$ and $1 - \varepsilon_{bkg} = 0.20$.

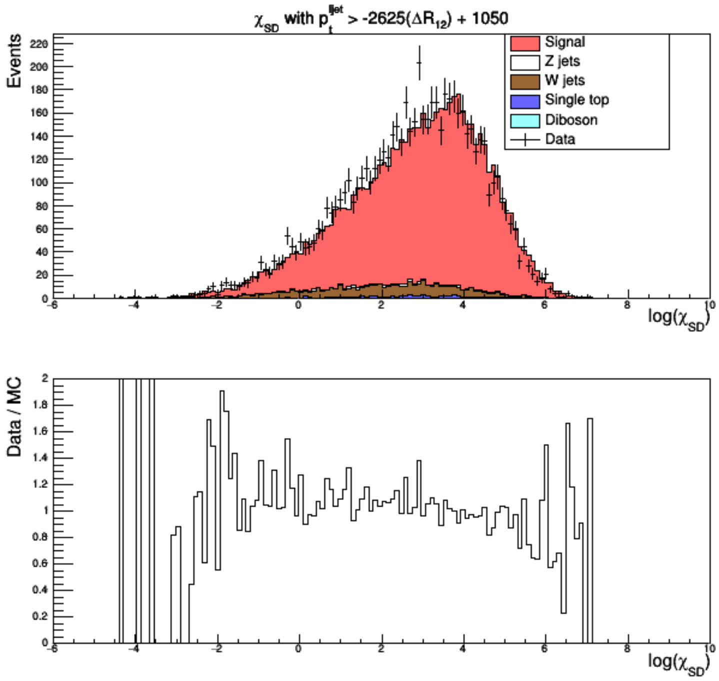
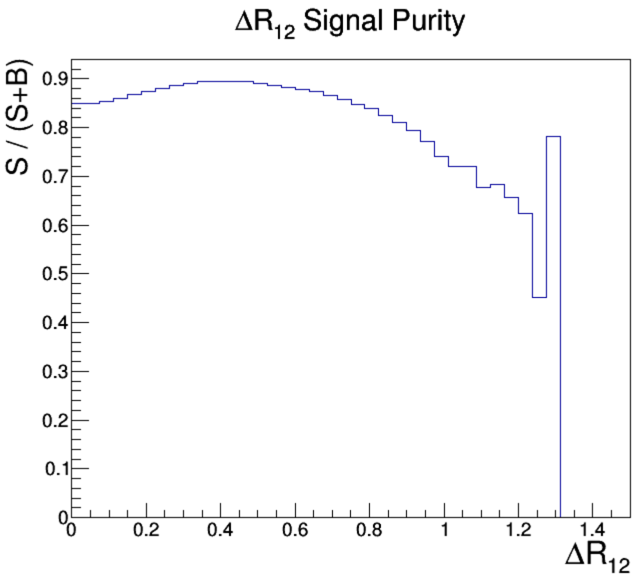
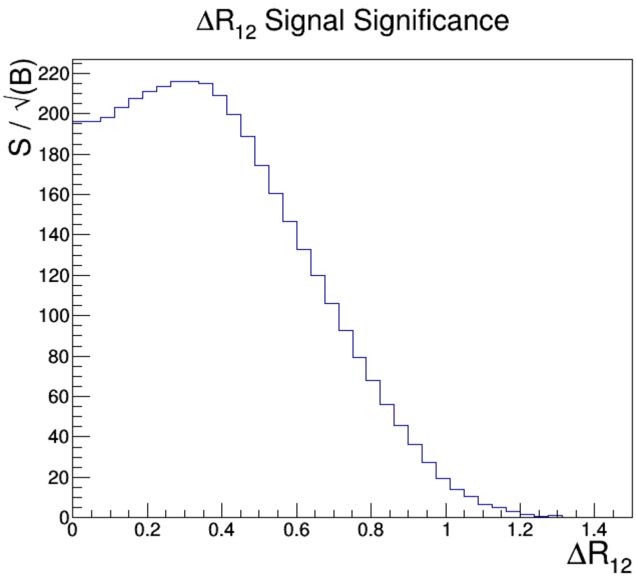


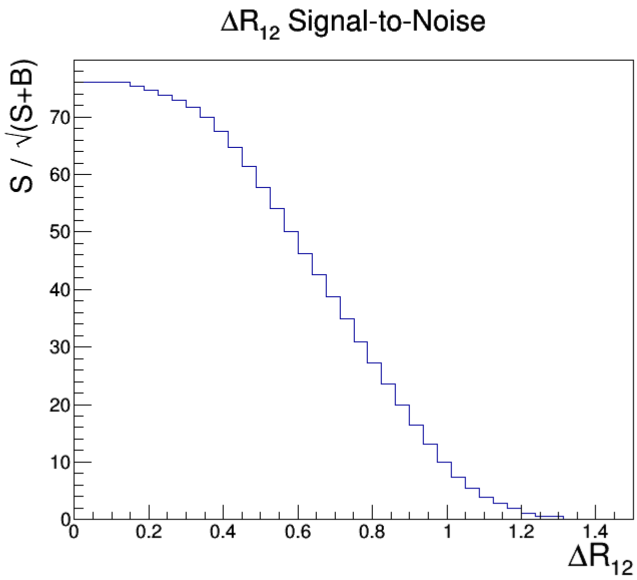
Figure 17: χ for large jets with $p_t > -2625(\Delta R_{12}) + 1050$. This cut gave $\varepsilon_{sig} = 0.95$ and $1 - \varepsilon_{bkg} = 0.12$.



(a) Signal Purity

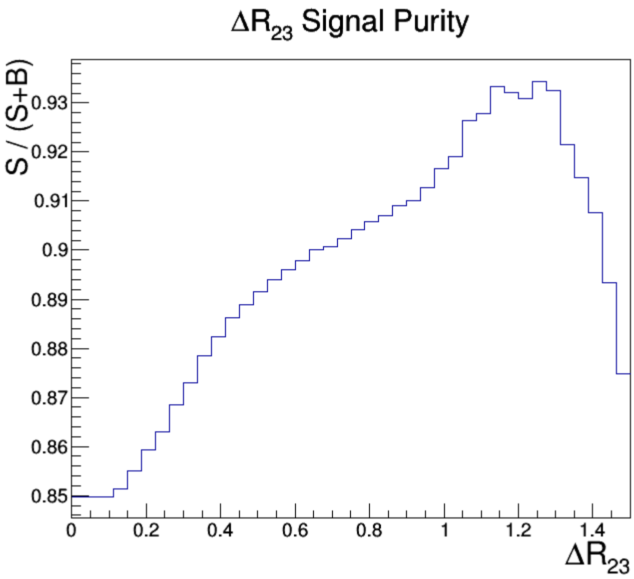


(b) Signal Significance

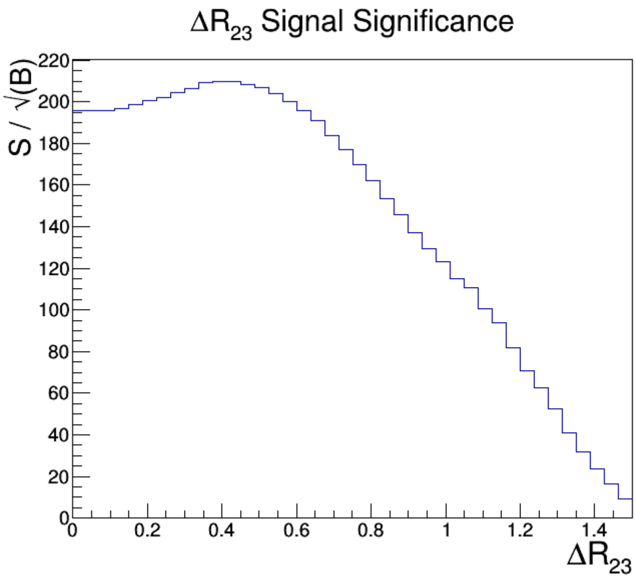


(c) Signal-to-noise ratio

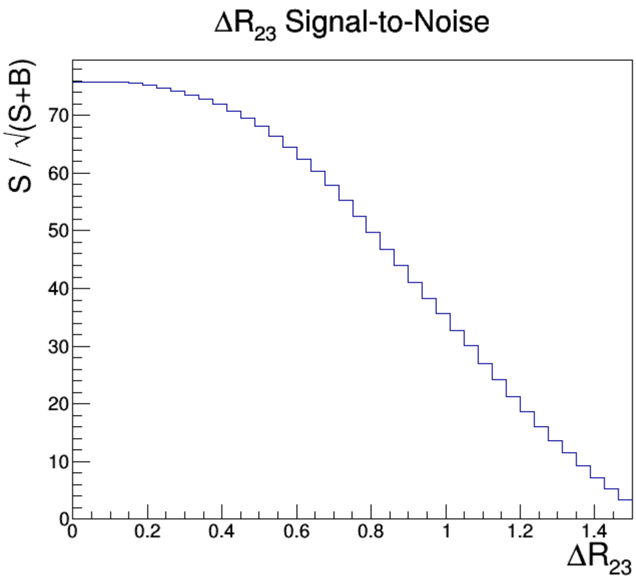
Figure 18: Ratio plots for ΔR_{12}.



(a) Signal Purity

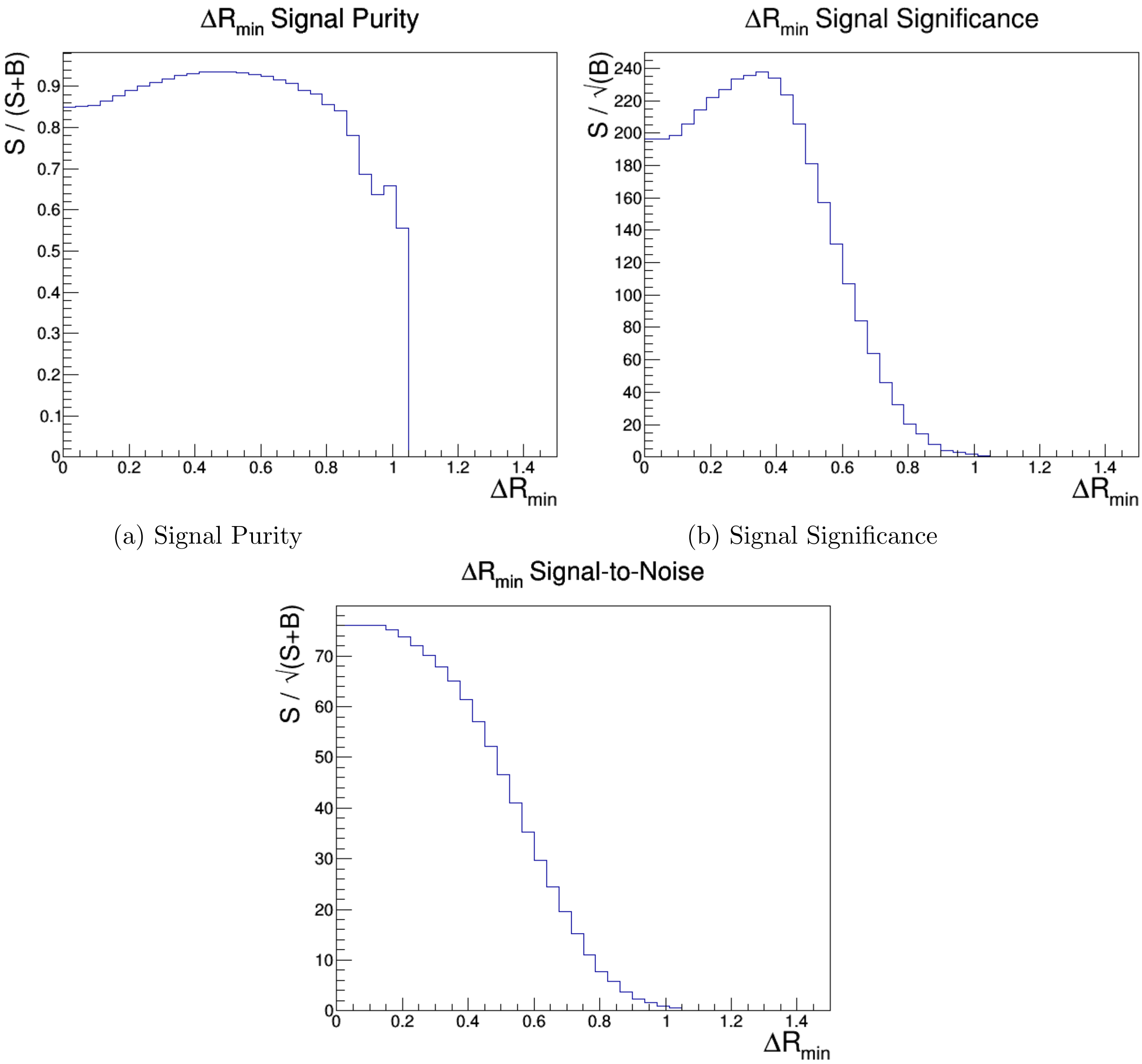


(b) Signal Significance



(c) Signal-to-noise ratio

Figure 19: Ratio plots for ΔR_{23}.



(a) Signal Purity

(b) Signal Significance

(c) Signal-to-noise ratio

Figure 20: Ratio plots for ΔR_{min}.

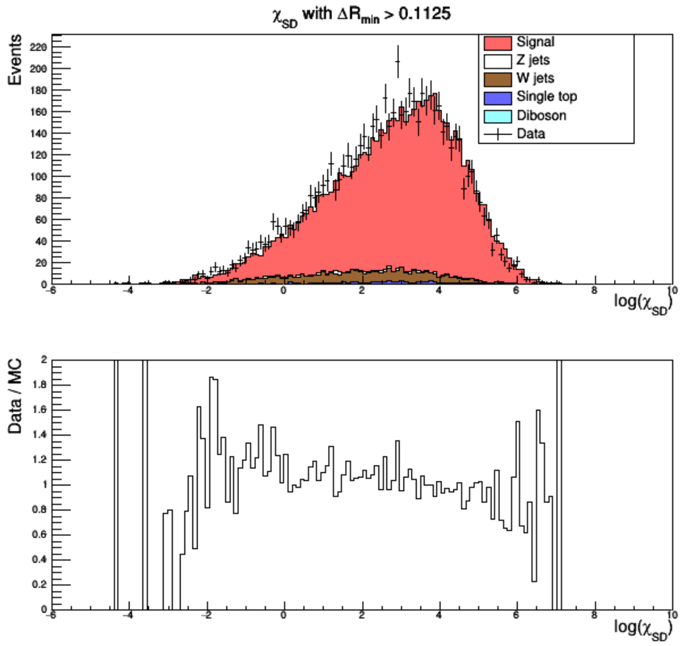


Figure 21: χ for large jets with $\Delta R_{min} > 0.1125$. This cut produces the best signal purity without changing the signal-to-noise ratio, and it results in $\varepsilon_{sig} = 0.99$ and $1 - \varepsilon_{bkg} = 0.09$.

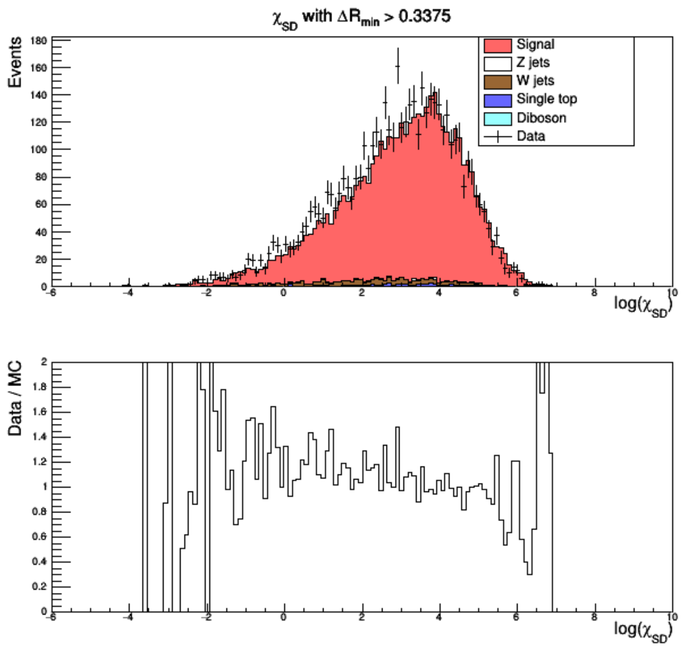


Figure 22: χ for large jets with $\Delta R_{min} > 0.3375$. This cut produces the best signal significance, and it results in $\varepsilon_{sig} = 0.72$ and $1 - \varepsilon_{bkg} = 0.65$.

4 Conclusions & Prospects

The goal of this research was to find the best way in which to remove background events from a sample that passes the $t\bar{t}$ selection cuts. This is necessary in order to maximize the purity of $t\bar{t}$ events so that systematic uncertainties in the Shower Deconstruction algorithm can be more accurately calculated in the future. I have shown that cuts on the ΔR between the subjets of the candidate hadronic top jet can be used to remove portions of this background. A cut on ΔR_{min} of 0.1125 will produce the highest possible background rejection rate of 9% without affecting the signal-to-noise ratio. This cut will reduce the background and the statistical uncertainty in the background that would contribute to the systematic uncertainty of Shower Deconstruction, leading to a small systematic uncertainty in the algorithm.

Future studies that want to use this information should try to find the exact origin of the accumulation of $t\bar{t}$ events with ΔR between the three highest p_t subjets near and below 0.2 as seen in Figure 7. This information may be useful for related studies that deal with differentiating signal from background. Also, if anyone wanted to increase the signal significance of the samples used in this project, the best cut was found to be at $\Delta R_{min} = 0.3375$. This produced a large background rejection of 65%, but it decreases the signal-to-noise ratio with a signal acceptance of 72%. This low signal acceptance is something that would definitely need to be considered when maximizing the signal significance.

5 Acknowledgments

I want to thank Dr. Josu Cantero Garcia for for help editing this paper, providing the program code and files necessary for starting this research, and for overall guidance and information throughout the process. I also want to thank Evan Van de Wall for help editing this paper, for providing the code necessary for producing plots, and for help in modifying and correcting the code.

References

[1] B.R. Webber, Int. J. Mod. Phys. **A15S1**, 577 (2000). URL: https://arxiv.org/abs/hep-ph/9912292

[2] A. Quadt, Eur. Phys. J. C **48** (3), 835 (2006). URL: https://doi.org/10.1140/epjc/s2006-02631-6

[3] D.E. Soper and M. Spannowsky, Phys. Rev. **D87** (2011). URL: https://journals.aps.org/prd/abstract/10.1103/PhysRevD.84.074002

[4] ATLAS Collaboration, *Performance of Top Quark and W Boson Tagging in Run 2 with ATLAS*, ATL-COM-PHYS-2016-531, 2016, URL: https://cds.cern.ch/record/2152725/files/ATL-COM-PHYS-2016-531.pdf

[5] X.C. Vidal and R.C. Manzano, *Momentum: Taking a Closer Look at the LHC* URL: https://www.lhc-closer.es/taking_a_closer_look_at_lhc/0.momentum