Machine Learning Methods for Detecting Semi-Visible Jets REU Program at Columbia University - Nevis Labs

Jonah Mougoue[1]

[1]Columbia University

August 5, 2023

Abstract

A semi-visible jet (SVJ) is the decay product of a proposed Z' boson. These jets contain both matter from the standard model as well as dark matter. The Z' boson is theorized to be a product of proton-proton collisions, such as those produced at the Large Hadron Collider (LHC) at CERN. To find a SVJ, many different machine learning models have been proposed such as Boosted Decision Trees (BDTs) and a Particle Flow Networks (PFNs). In this study, we first compared a BDT with a PFN to see which machine learning model would best find SVJs. In a one-to-one comparison, the BDT correctly predicted SVJs with 88% accuracy, while the PFN correcly predicted SVJs with 91% accuracy. We next attempted to create a signal and control regions to compare with the PFN post-analysis. The background events need to be split so each region has a similar background mT_{jj} distribution. We found that jet2_{Width} is currently the best variable at cutting the signal and control regions, but more studies need to be performed on finding the best variable cut.

NEVIS LABORATORIES
COLUMBIA UNIVERSITY

Contents

1 Introduction

1.1 Standard Model

The Standard Model (SM) is a theory of subatomic particles that has been successfully used to predict new physics for decades, most recently with the discovery of the Higgs boson in 2012. The particles of in SM are split into three categories: fermions, gauge bosons, and the Higgs boson.

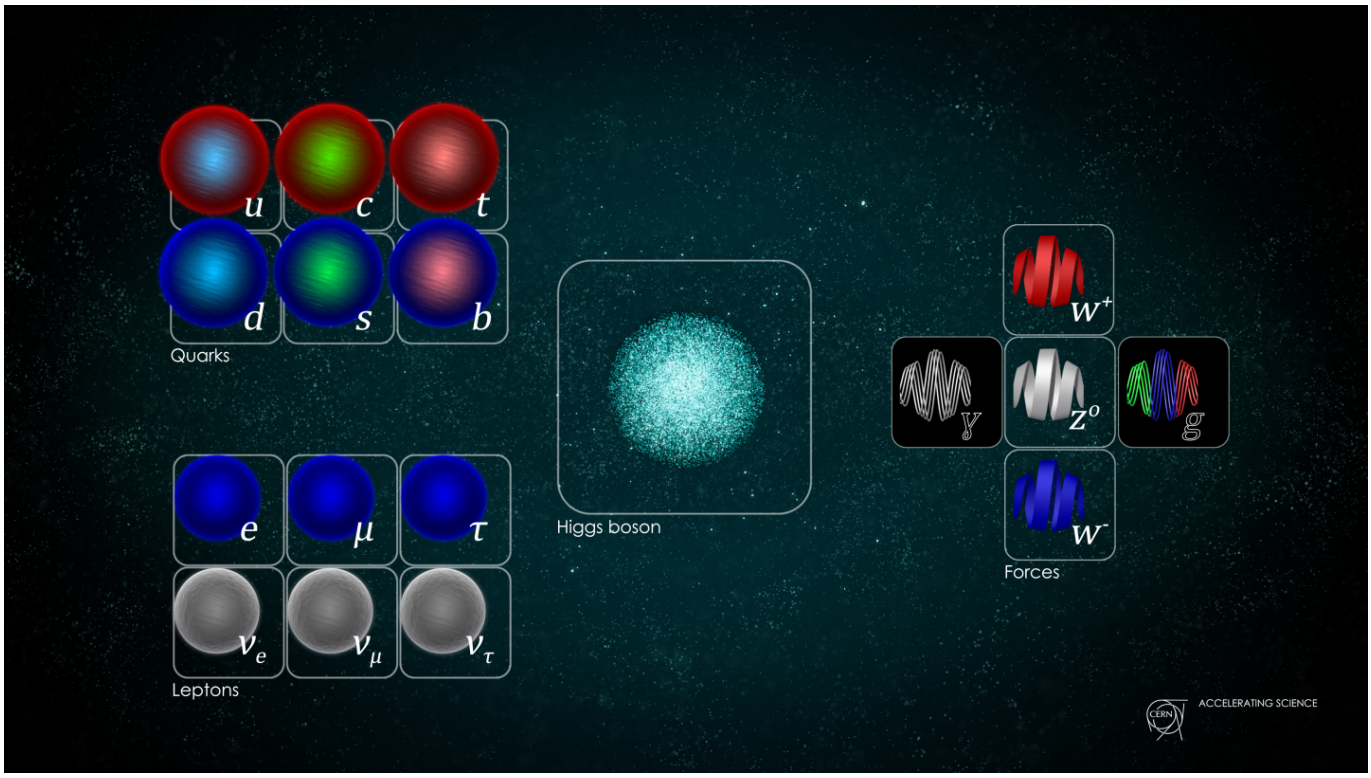


Figure 2: *The current Standard Model. Fermions are on the left, gauge bosons are on the right, and the Higgs boson is in the center. [5]*

Fermions are massive particles which all matter is made of and have $\frac{1}{2}$ integer spins. All fermions have a corresponding antifermion which have the same properties as the corresponding fermion but with opposite charge. Fermions can further be split into two categories, quarks and leptons. Quarks have fractional charges $(\pm\frac{1}{3}, \pm\frac{2}{3})$. Quarks interact with the strong force and a quark can never be separated from at least one other quark. Quarks then either form mesons or baryons (such as neutrons or protons), depending on the sum of the quark charges. These protons and neutrons form the basis for an atomic nucleus. leptons, unlike quarks, have integer charges $(0, \pm 1)$. Leptons cannot interact with the strong force, and so, unlike quarks, cannot be a part of the atomic nucleus. Leptons are created through the weal force and the products of the decay will at least include a charged lepton (an electron, muon, or tau), as well as their respective neutrino. The fermions are also split into three generations in which each generation contains particles that are more massive than the previous generation. Since muons and tauons are more massive than the electron, they're less stable and are thus significantly less likely to be found within an atom. The fermions make up the majority of the mass within an atom, but the fermions are held together due to the gauge bosons.

Gauge bosons are force carrying particles that allow particles to interact with each other, and have integer spin. Photons carry the electromagnetic force, gluons carry the strong force, while Z

and $W^{\pm}$ bosons carry the weak force. The electromagnetic force is the force that allows an electron to revolve around an atomic nucleus, and the strong force is the force that holds the protons and neutrons in an atomic nucleus together. The weak force allows quarks to change their charge, causing a nucleon to change into another, thus causing beta decay and the emission of a charged lepton and its corresponding neutrino. The gauge bosons explain the mechanics three of the four fundamental forces. As gravity is negligible at the sub-atomic scale, the gauge bosons explain how the fermions interact each other and how the particles within an atom can form a stable structure with each other. The Higgs boson doesn't carry a force like the other bosons, but instead gives particles mass. Since everything from the smallest atoms are made from fermions and bosons, the interactions between fermions and bosons are what define the fundamental physics of the universe. [5]

1.2 Beyond Standard Model Physics

The Standard Model has been used to successfully explain nearly all experimental physics results and to predict new physics. While the Standard Model is the most accurate physics model to have been created, it is still incomplete, and many scientists have thought of new theories deemed to be Beyond the Standard Model(BSM). SM doesn't explain many of the things we experimentally notice about the universe. For example, while SM explains 3 of the 4 fundamental forces well, it offers no explanation for gravity. SM also doesn't account for why there is an abundance of matter compared to antimatter. Lastly, SM doesn't explain the presence of dark matter in the universe. [5]

1.2.1 Dark Matter

In order for a galaxy to exist, there must be enough mass within the galaxy to hold it together. According to experimental data, however, it appears that there isn't enough mass within galaxies to hold them together according to our current understanding of gravity. The discrepancy between the theoretical mass needed to hold a galaxy together and the experimental mass found within galaxies has led scientists to believe that there is some matter in the universe that we are currently unable to directly detect. This matter isn't capable of interacting with SM particles in any way other than through the gravitational force. Since the matter is currently believed to be impossible to directly detect, the matter is deemed to be 'dark'. This dark matter is believed to be around six times as abundant as SM matter. While dark matter mass can't be detected, dark matter can be found by measuring the energy and mass of a system before and after a collision. If two SM particles collide and the total momentum after the collision is less than the total momentum before the collision, it is possible that the collision created a dark matter particle that can't be detected. [1]

1.3 Semi-Visible Jets

Jets are a stream of particles that arise when two particles collide with each other. The collision then forms multiple new particles move in many different directions. The particle tracks are then categorized into groups based on their trajectory called jets.

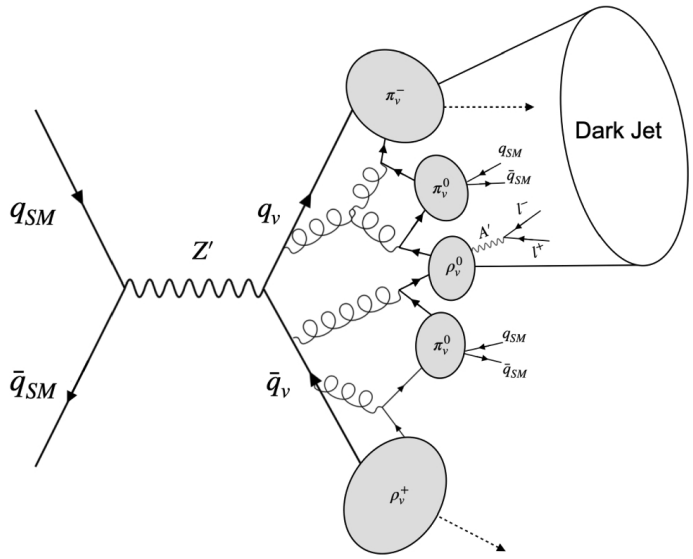


Figure 3: *Diagram showing a two SM quarks showering and forming a semi-visible jet. [8]*

Some BSM theories predict the existence of an additional boson to the Standard Model, called the Z' boson. Z' boson, which are the result of proton-proton collisions, then decay into multiple jets. These jets contain matter from SM, but also dark matter as well. This type of jet is called a Semi-Visible Jets (SVJ) since particles within the jet can't be detected directly. The production of dark matter particles leads to missing energy in the transverse plane (MET).The MET is often aligned in the same direction as a SVJ since the jet contains the dark matter particles that account for the missing energy. [7]

1.3.1 Difficulties in Finding SVJs

Currently, there hasn't been a single jet that has been correctly identified as a SVJ. Since no SVJs have been detected, there is no definite data to describe the properties of SVJs. If the chance of finding a SVJ is non-zero, then it's likely that the chance of finding a SVJ is very low. One major difficulty in finding SVJs is that the signature for a SVJ can be similar to the signature of a mismeasured SM jet. [7] These jets are produced through what's known as the QCD multijet process and these jets compose the majority of SM background. [9] SVJs contain two unknown properties about themselves that determine their MET, the ratio of stable dark hadrons to total dark hadrons (r_{inv}) and the mass of the Z' boson (Z' mass).

$$r_{inv} = \frac{N_{Stable}}{N_{Stable} + N_{Unstable}} \text{ [9]}$$

SVJ searches are particularly sensitive to events with $r_{inv} \simeq 1$, since these events have high missing energy. [9] High Z' mass also means the SVJs contain more energy, which means higher MET. The expected range of values for r_{inv} and Z' mass in a SVJ, however, are unknown. We thus don't know the expected range for MET in a SVJ.

1.4 Large Hadron Collider

Built in 2008, the Large Hadron Collider (LHC) at CERN was created with the purpose of exploring both SM and BSM physics. It is currently the largest and the most powerful particle accelerator in the world.

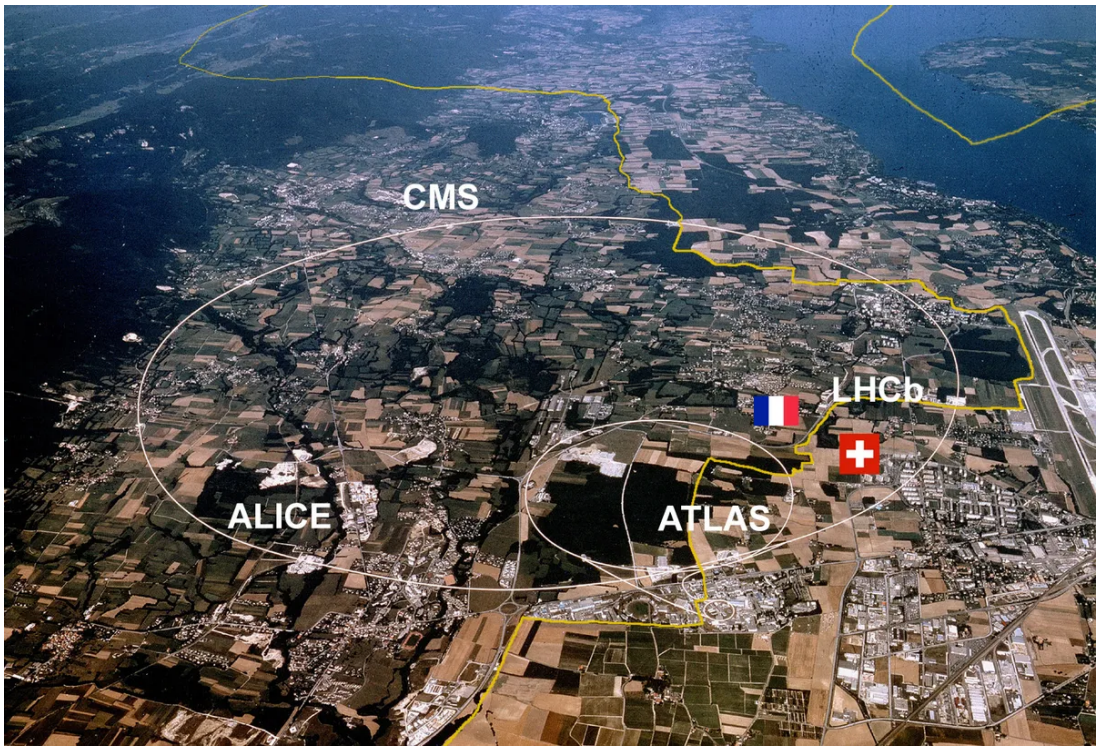


Figure 4: *A map of the LHC*

The LHC is a 27km circumference ring composed of thousands of superconducting magnets. The interior of the ring is kept in a vacuum and the magnets are cooled to a temperature of $-271.3°C$ in order to stay in a superconductive state. First, hydrogen atoms are ionized to form free protons. Using the magnets, the free protons are accelerated both clockwise and counterclockwise around the LHC. This forms two beams of particles that travel in opposite directions. The particles are then sent to collide at one of the LHC's four detectors, one of which is the Atlas detector. [2, 3] Since July 2022, the LHC has been on its third run and is able to create proton-proton collisions with a center of mass energy of 13.6 TeV. [4]

1.5 Atlas Detector

A Toroidal LHC ApparatuS (ATLAS) is a general-purpose detector in the LHC. Weighing 7000 tons, the 46m x 25m x 25m detector is designed to detect and record data from the billions of collisions that occur within the LHC. As particles in ATLAS collide and form multiple other particles, ATLAS is able to collect data on the particles through its multiple layers of detectors and can use the data to reconstruct the tracks that the particles traveled in. ATLAS contains four layers of sub-detectors. They are, from the innermost to the outermost, the inner detector, the Liquid Argon Calorimiter (LAr), the Tile Hadronic Calorimeter (THC), and the Muon Spectrometer respectively. [15]

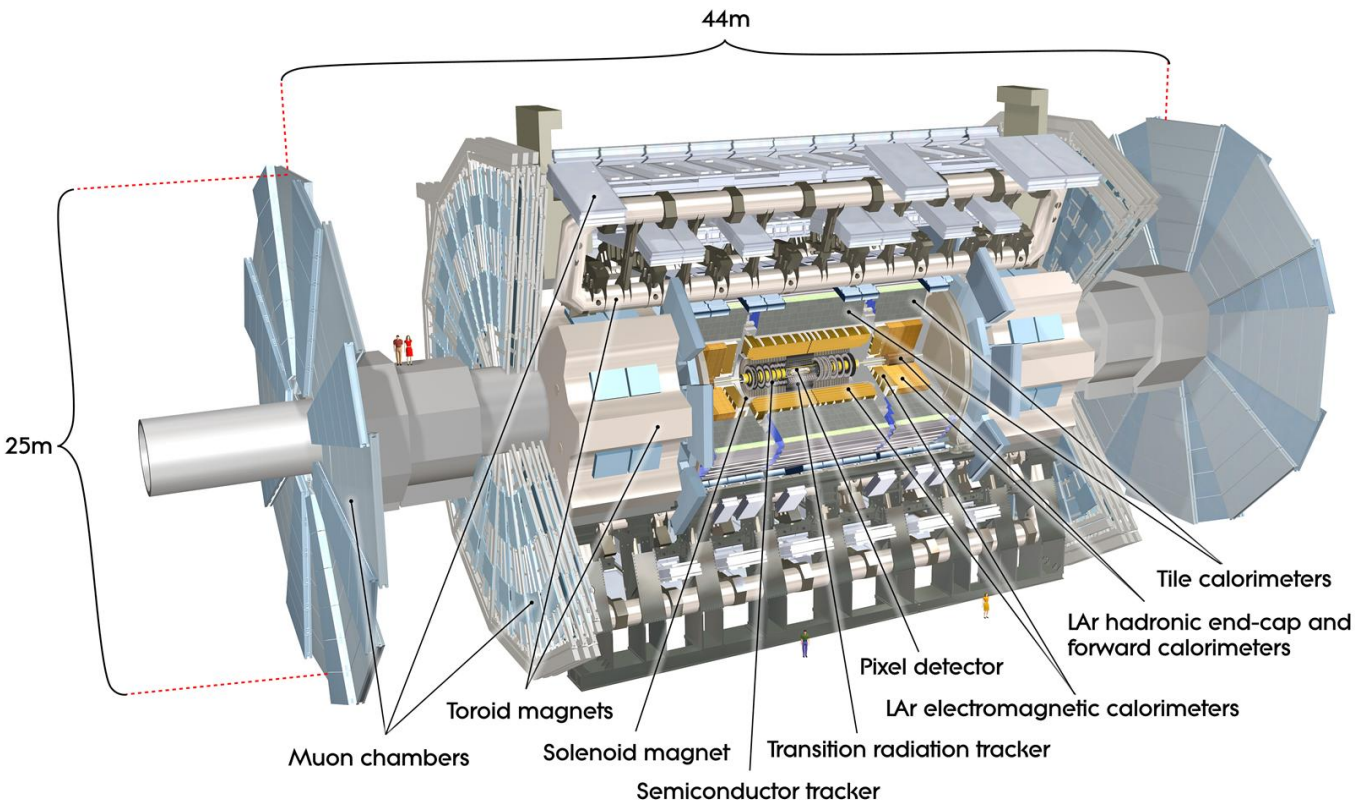


Figure 5: *A labeled diagram of ATLAS [11]*

The inner detector itself is made of three different mechanisms, the Pixel Detector, the Semiconductor Tracker(SCT), and the Transition Radiation Detector (TRT). These mechanisms are designed to measure the origin, momentum, track, and type of a particle produced during the collision. The inner detector is surrounded by a Central Solenoid Magnet, a superconducting magnet which bends the tracks of particles after collision and allows for the momentum and charge of particles to be measured. [15]

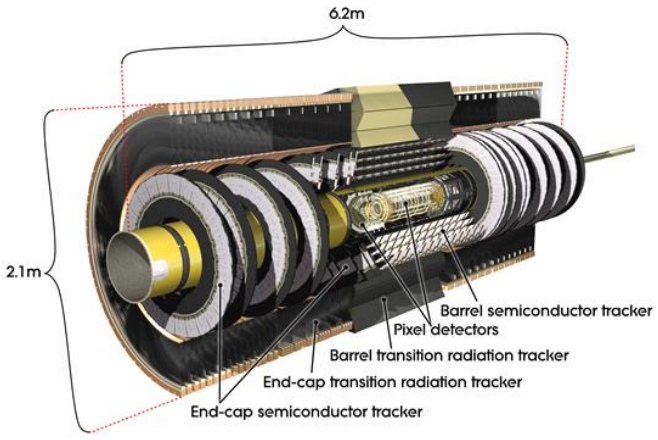


Figure 6: *The inner detector of ATLAS [12]*

The inner detector is surrounded by the LAr, which is an electromagnetic calorimeter built to detect the energy of photons, electrons, and positrons produced in the collision. [3] When these particles interact with the LAr, they form into a shower of lower energy particles and the LAr is able to measure the energies of the low energy particles to reconstruct the energy of the particle before showering. [15]

The LAr is further surrounded by the THC, which is used to measure the energies of hadrons that aren't stopped by the LAr. [3] When particles collide with the THC, they form showers of lower energy particles. The THC detects these particles and produces electric currents proportional to the masses of the particles. The currents are then used to reconstruct the energy of the hadron. [15]

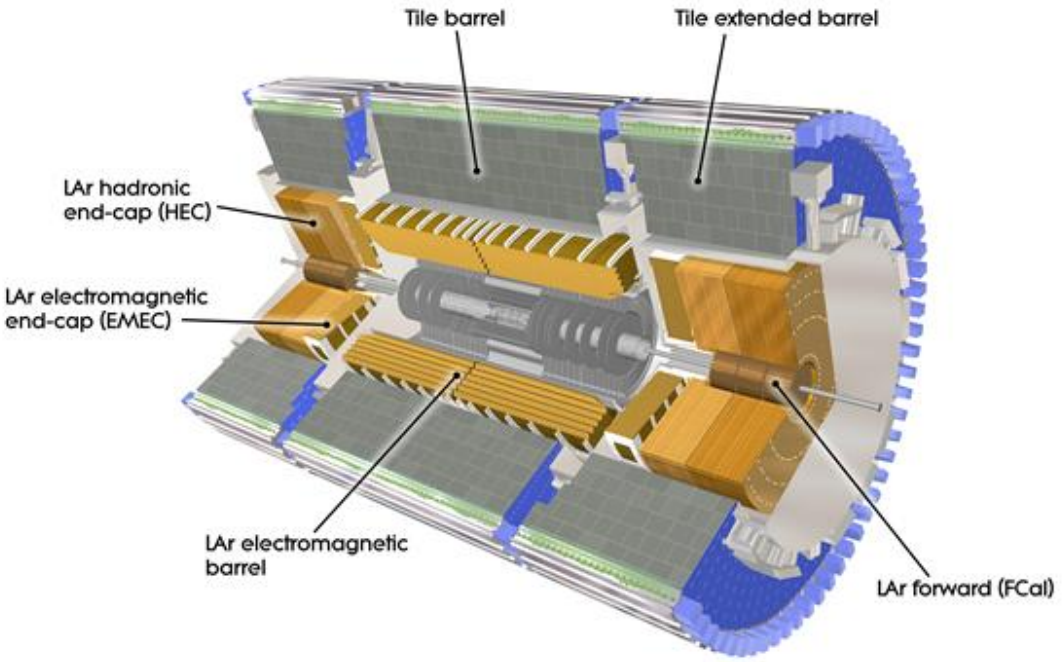


Figure 7: *The Calorimeters within ATLAS [10]*

The outermost layer of ATLAS is the muon spectrometer. The precision detectors in the muon spectrometer determine the position of muons and the fast-response detectors measure the muons' momentum and decide whether to keep the collision event. ATLAS is further surrounded by three Toroid Magnets, which are used to further measure the momentum of muons. [15]

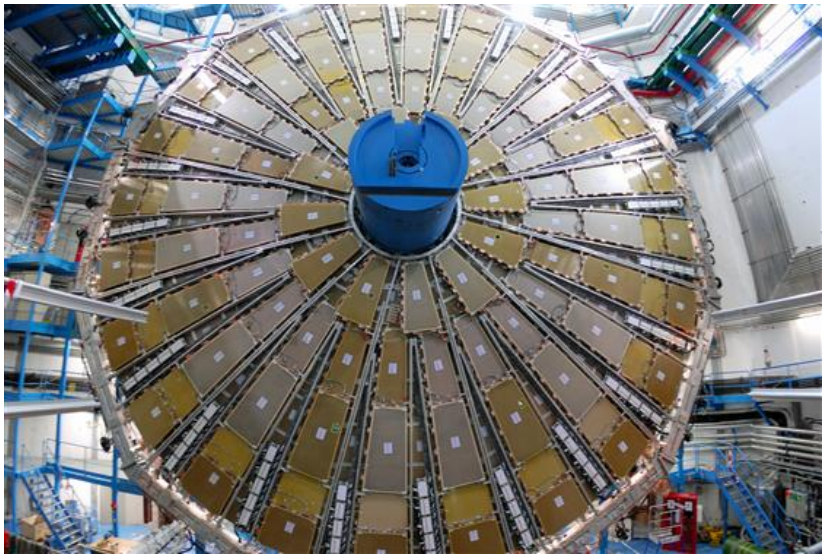


Figure 8: *The outside of the muon spectrometer. [13]*

ATLAS can detect up to 1.7 billion collisions a second, so to significantly reduce the volume of data, Atlas has two built-in trigger systems which determines which collisions to record. The first trigger uses information from the calorimeters and the muon spectrometer to select up to 100,000 events per second. The second trigger analyzes each collision and selects about 1000 events per second to record. Over 10,000 TB of data per year is recorded by ATLAS. The data ATLAS directly records is known as low-level data. ATLAS is then able to use low-level data to reconstruct properties of the particles, known as high-level data, such as tracks, momentum, and type of particle. [15]

While ATLAS records data from billions of collisions, the majority of the ATLAS collaborations' computing resources are used to run simulations in what is called the Monte Carlo method. These simulations are used to test theories and to collect simulated data, which is used to run test data analysis before real data is collected. [15]

1.6 Boosted Decision Tree

Boosted Decision Trees (BDTs) are a machine learning technique that has become popular among high energy particle physicists in the past few decades. Boosted decision trees separate rare signals from large amounts of background by using high-level variables reconstructed using low-level ATLAS data.

1.6.1 Decision Tree Algorithm

Decision trees are algorithms designed to split data into signal and background regions when the majority of events neither share all the properties of the signal or the background by finding strong classifiers that divide the events into signal and background.

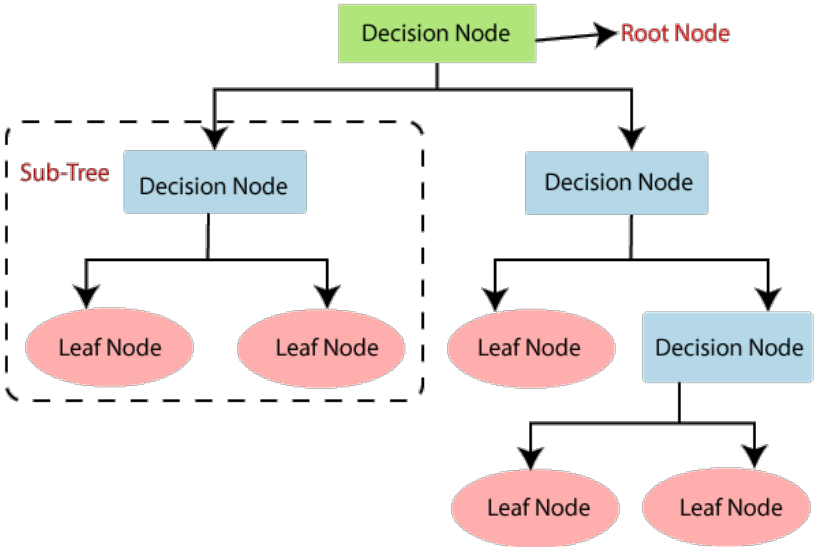


Figure 9: *A diagram of a decision tree [14]*

The basic algorithm for a decision tree, starting with the root node, is as stated:

1. If a node satisfies a stopping condition, the node becomes a leaf
2. Sort all events according to each variable
3. Find a value for each variable that best divides the set of events into signal and back round regions. If the separation can't be improved further, exit the algorithm.
4. Select the variable that leads to the best separation and create two children which contain each group respectively.
5. Repeat recursively. [6]

1.6.2 Boosting

Boosting algorithms are algorithms designed to make strong classifiers out of weak classifiers. It's difficult to find a variable cut that can accurately predict the vast majority of events, but there are many variable cuts that can be made to successfully predict at least 50% of event. These cuts are known as weak classifiers. Multiple weak classifiers, however, can be combined to form strong classifiers. A boosting algorithm attempts to find and combine multiple weak classifiers into a strong classifier and this classifier is used as variables for splitting in the decision tree. [6]

2 Boosted Decision Tree vs Particle Flow Network

In 2022, Compact Muon Solenoid (CMS), the sister experiment to ATLAS at the LHC, attempted the first search for Semi-Visible jets and used BDTs to attempt to discriminate SVJs from background jets. [9] As CMS did not detect any SVJs, the effectiveness of BDTs in searching for SVJs compared to other machine learning models has been questioned. One proposed machine learning method for detecting semi-visible jets is a Particle Flow Network (PFN). While BDTs are decision trees, PFNs are neural networks, and while BDTs use high-level data that describes physical properties of the jets and the collision as a whole, PFNs use low-level data used to reconstruct the particle tracks for

the leading and subleading jets. [16] This means that the PFN is able to make correlations between variables based on data from jet substructures.

In this study, we use ATLAS Monte Carlo data to compare BDT performance with PFN performance. This study focuses on changes made to the BDT and how it impacts the effectiveness of the BDT. The BDT is tested using multiple different combinations of variables in an attempt to maximise the efficiency of the BDT. The Monte Carlo data is split into multiple signal files with with different values for r_{inv} and Z' mass and a background file. The BDT is trained and tested over files that contain SVJs of multiple different rinv and Z' mass values.

Table 1 contains a description for each high-level variables that was given to the BDT. Table 2 shows each of the track variables given to the BDT. These variables were reconstructed manually using low-level data so the BDT could have information on the jet substructures.

jet1	The leading jet / the jet with the highest pt
jet2	The subleading jet / the jet with the second highest pt
n_{jets}	Number of jets detected
jet1/2$_{pt}$	Transverse momentum of jets1 and jet2
pt_balance$_{12}$	(jet1$_{pt}$ - jet2$_{pt}$)/jet1$_{pt}$
jet1/2$_{\eta}$	Pseudorapidity of jet1 and jet2
$\Delta\eta_{12}$	The difference between jet1$_{\eta}$ and jet2$_{\eta}$
MET	Missing energy in the transverse direction
mT$_{jj}$	The total reconstructed mass
rT	MET/mT$_{jj}$
$\Delta\phi_{min}$	The minimum transverse angle from either jet to the direction of MET
$\Delta\phi_{max}$	The maximum transverse angle from either jet to the direction of MET
$max\phi\_min\phi$	The difference between $\Delta\phi_{max}$ and $\Delta\phi_{min}$
ΔR	The solid angle between the two leading jets
deltaγ_{12}	the difference in rapidity between jet1 and jet2
Aplanarity	How well the jets are distributed in the transverse plane
Sphericity	A measure of the spherical symmetry of the distribution of jets
Sphericity$_T$	Sphericity in the transverse plane

Table 1: *All high level variables given to the BDT*

Jet1/2_width	Jet width calculated using calorimeter data
Jet1/2_TrackWidthPt1000PV	The width between the two furthest tracks with pt over 1000 MeV within the jet from the primary vertex
Jet1/2_SumPtTrkPt500PV	The pt sum of each track of at least 500 MeV within the jet from the primary vertex
Jet1/2_NumTrkPt1000PV	The amount of tracks of at least 1000 MeV within the jet from the primary vertex

Table 2: *Table of track variables - These are variables that we manually constructed using track data to test the BDT*

2.1 Preselection

Before the BDT performance analysis was ran, we preselected events that had jet1$_\eta$ and jet2$_\eta$ between -2.1 and 2.1 and $\Delta\gamma$ below 2.8.

2.2 Results

2.2.1 PFN Performance

Figure 10 shows that the AUC of the PFN is 0.91. This ROC curve serves as a benchmark to compare with the BDT.

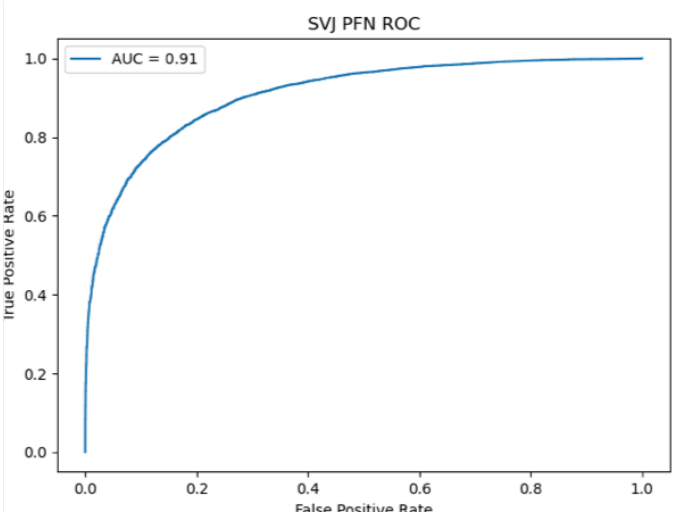


Figure 10: *A receiver operating curve (ROC) plots the false positive rate vs the true positive rate of a machine learning algorithm. The Area Under the Curve (AUC) shows how likely a machine learning algorithm is to correctly detect signal from background.*

2.2.2 Correlation Matrices

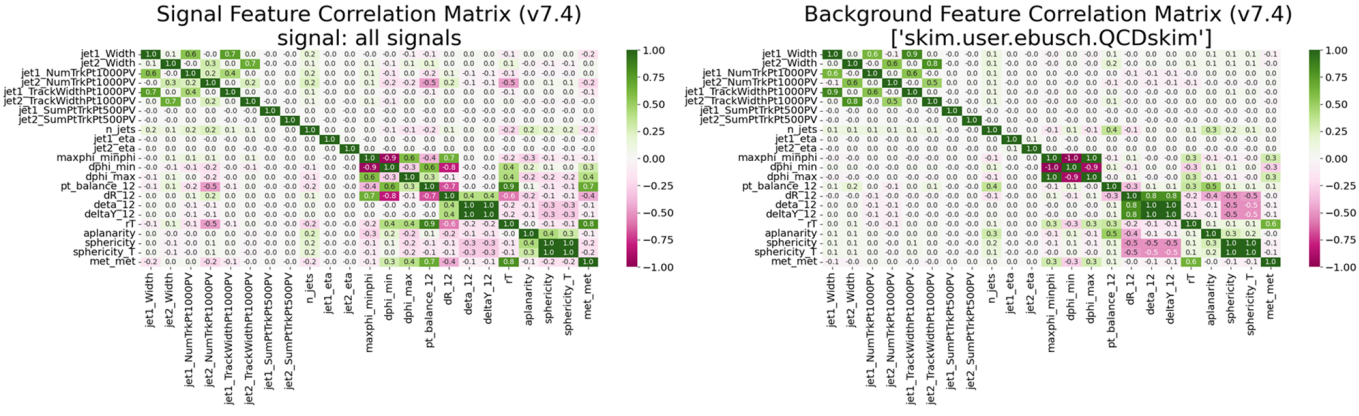
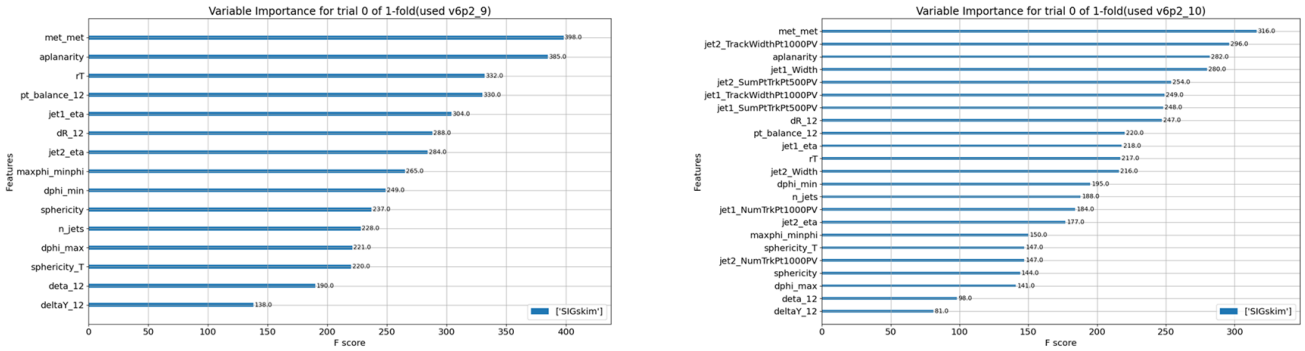


Figure 11: *Correlation matrices for signal and background respectively*

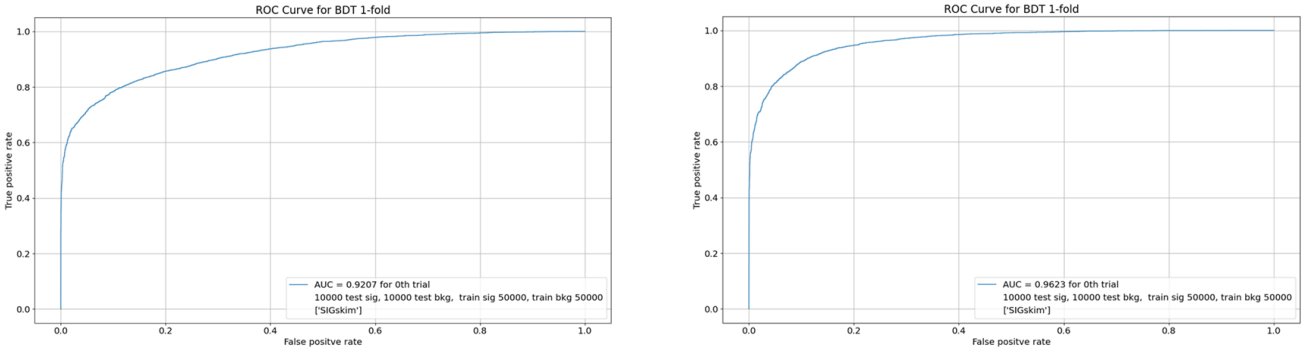
2.2.3 BDT With Track Variables vs Without Track Variables

These results show the difference in BDT performance when adding 8 new variables describing the properties of the jet tracks (jet1/2_Width, jet1/2_NumTrkPt1000PV, jet1/2_TrackWidthPt1000PV, and jet1/2_SumPtTrkPt500PV). These variables were manually reconstructed using low-level track data from the PFN.



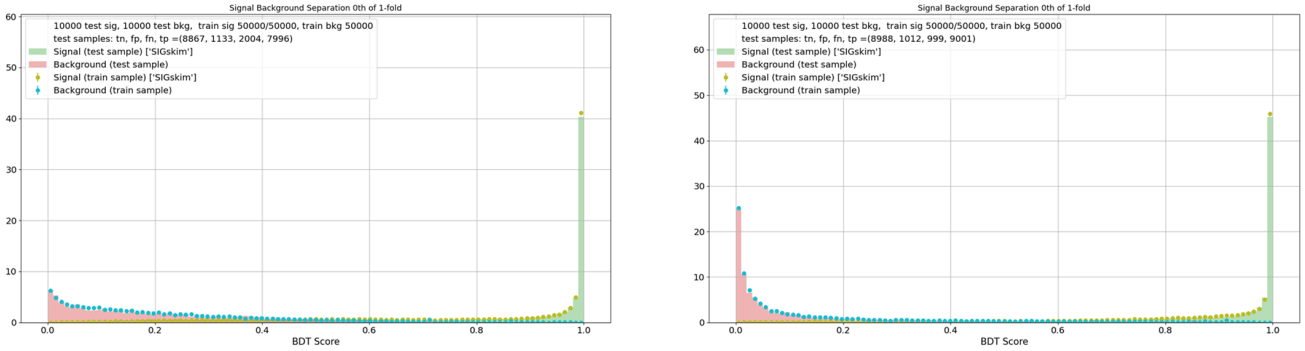
(a) *Variable importance with no track variables* ***(b)*** *Variable importance with track variables*

Figure 12: *Variable importance rankings rate each variable by how discriminating it is*



(a) *ROC Curve Without Tracks* ***(b)*** *ROC Curve with Tracks*

Figure 13: *A receiver operating curve (ROC) plots the false positive rate vs the true positive rate of a machine learning algorithm. The Area Under the Curve (AUC) shows how likely a machine learning algorithm is to correctly identify signal from background.*

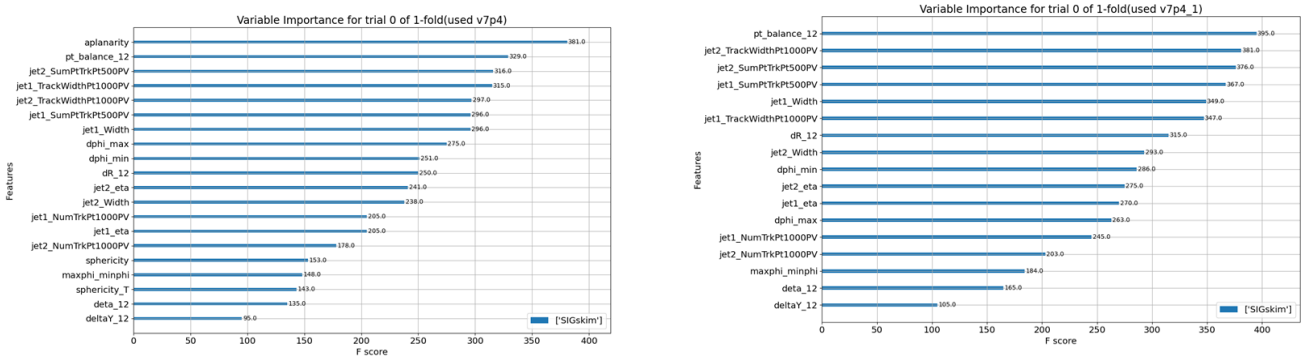


(a) SBS with no track variables *(b) SBS with track variables*

Figure 14: *Signal Background Separation (SBS) shows a histogram of each event and its respective BDT score. High separation between signal and background BDT scores indicates better performance.*

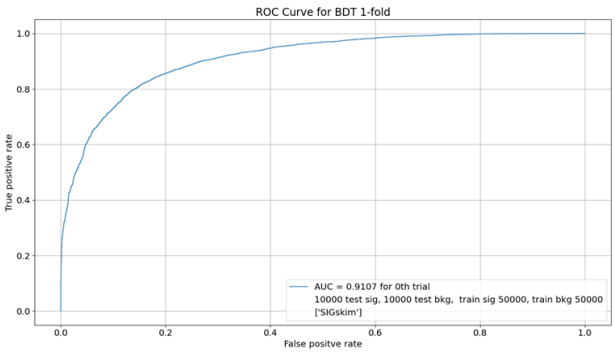
2.2.4 BDT With vs Without Energy Distribution Variables

These results show the difference in BDT performance when removing certain high-level variables (n_{jets},rT,MET) and when further removing energy distribution variables (aplanarity, sphericity, sphericity_T).

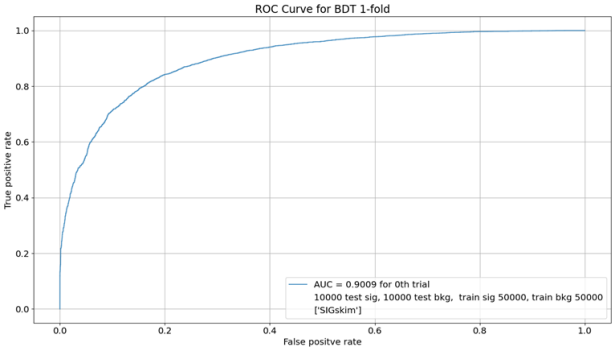


(a) Variable importance with no n_{jets}, rT, MET *(b) Variable importance with no n_{jets}, rT, MET, aplanarity, sphericity, sphericity_T*

Figure 15: *Variable importance rankings rate each variable by how discriminating it is*

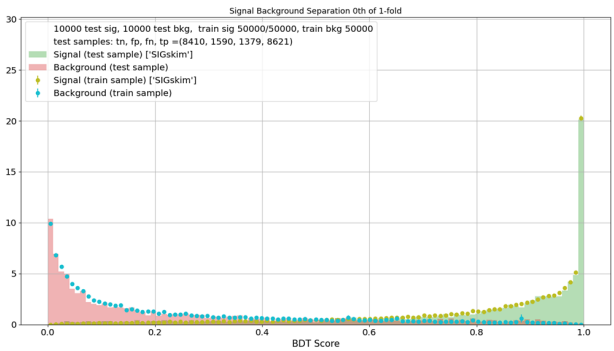


(a) ROC curve with no n_{jets}, rT, MET

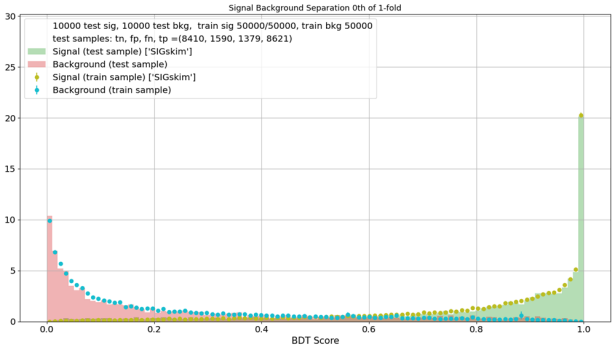


(b) roc curves with no n_{jets}, rT, MET, aplanarity, sphericity, sphericity$_T$

Figure 16: *A receiver operating curve (ROC) plots the false positive rate vs the true positive rate of a machine learning algorithm. The Area Under the Curve (AUC) shows how likely a machine learning algorithm is to correctly identify signal from background.*



(a) SBS with no n_{jets}, rT, MET

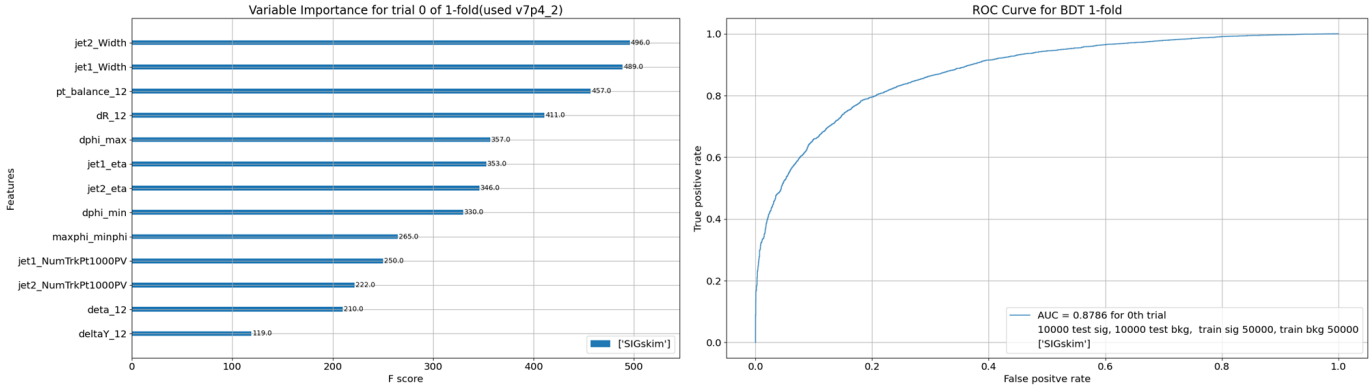


(b) SBS with no n_jets, rT, MET, aplanarity, sphericity, sphericity$_T$

Figure 17: *Signal Background Separation (SBS) shows a histogram of each event and its respective BDT score. High separation between signal and background BDT scores indicates better performance.*

2.2.5 BDT Without Multiple Variables

These results show BDT performance when multiple variables are removed (n_{jets}, rT, MET, aplanarity, sphericity, sphericity$_T$, jet1/2_SumPtTrkPt500PV, and jet1/2_TrackWidthPt1000PV).



(a) *(b)*

Figure 18: *Variable importance and ROC curve with no n_{jets}, rT, MET, aplanarity, sphericity, $sphericity_T$, jet1/2_SumPtTrkPt500PV, and jet1/2_TrackWidthPt1000PV*

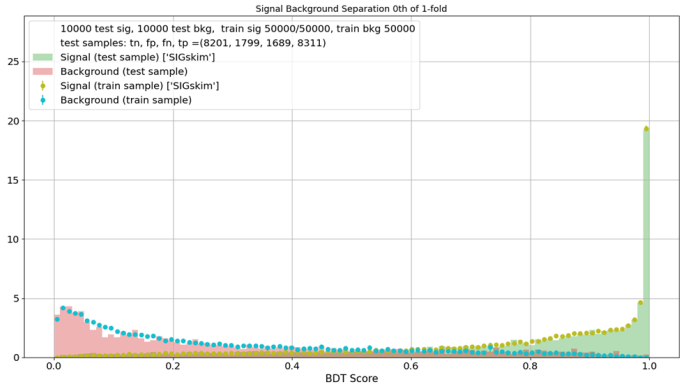


Figure 19: *SBS with no n_{jets}, rT, MET, aplanarity, sphericity, $sphericity_T$, jet1/2_SumPtTrkPt500PV, and jet1/2_TrackWidthPt1000PV*

2.3 Conclusion

Figure 13b shows that when the BDT is given all the variables, it performs better than the PFN with an AUC of 0.96 compared to an AUC of 0.91 shown in figure 10. The BDT only obtains an AUC of 0.96, however, when given jet1/2_TrackWidthPt1000PV and jet1/2_SumPtTrkPt500PV, two track-level variables which the BDT doesn't have access to. Figure 12b shows that these variables accounted for four of the eight top discriminating variables. The BDT without track variables got an AUC of 0.92 as shown in figure 13a. Even with jet1/2_TrackWidthPt1000PV and jet1/2_SumPtTrkPt500PV, the BDT performed similarly to the PFN when n_{jets}, rT, and MET were removed as shown in figure 16a. Since MET has been the best discriminating variable, removing MET significantly reduces the AUC of the BDT. Since r_{inv} and Z' mass are unknown for SVJs and the BDT was trained over signal files with multiple r_{inv} and Z' mass values, the BDT assumed events with high MET tend to be signal. This poses a problem, however, if SVJs have r_{inv} and Z' mass that leads them to have a low MET. While the BDT can discriminate well if SVJs have a high met, the BDT would perform significantly worse if SVJs have a low MET. Therefore, MET has to be removed so the BDT won't discriminate against SVJs with low MET. rT is also removed since it is calculated using MET. Lastly, n_{jets} is removed from the BDT since the PFN can only access properties of the leading and subleading jets, so the PFN doesn't have access to the number of

jets.12a, The performance of the BDT drops to 0.91 when these variables are removed removed. The performance of the BDT drops further when n_{jets}, rT, MET, aplanarity, sphericity, and sphericity$_T$ are all removed as shown in figure 16b. Aplanarity, sphericity, and sphericity$_T$ are variables that involve energy distribution after a collision due to the jets. Since the PFN can only access properties of the leading and subleading jets, the PFN can't reconstruct these variables, and so these variables are removed to make the BDT comparable to the PFN. Figure 18b shows that the BDT AUC falls to 0.88, significantly worse than the PFN, when n_{jets}, rT, MET, aplanarity, sphericity, sphericity$_T$, jet1/2_SumPtTrkPt500PV, and jet1/2_TrackWidthPt1000PV are all removed. Since this BDT only has variables the PFN can reconstruct, the AUC of 0.88 shows that the BDT performs worse than the PFN in a one-to-one comparison. These results show when the BDT isn't given certain discriminating high-level variables, the PFN performs better. When in a one-to-one comparison, the PFN was able to find correlations in the low-level track data that aren't related to the most discriminatory high level variables in the BDT. The PFN is therefore superior than the BDT as it is capable of discriminating between signal and background using correlations in the low-level data that the BDT doesn't have access to.

Since we determined that the PFN was superior to the BDT in discriminating signal from background, we assigned PFN scores to every signal and background event which is used in the next analysis.

3 Signal and Control Region Study

Now that we have decided to use a PFN instead of a BDT for our SVJ search, we need to be able to tell how accurate the PFN performs at the SVJ search. To accomplish this, we create a test signal with the majority of signal events and a test control region with less than 5% of events being signal in the control region. The background events need to be split so each region has a similar background mT_{jj} distribution. Once the PFN performs its analysis and creates its own signal region, the PFN data can be compared to the signal and control region. Since the PFN will likely discriminate some background as signal, the PFN signal region can be compared to the test signal and control regions to attempt to find a discrepancy in the PFN signal region distribution which could indicate the presence of SVJs. These regions can be made by applying cuts to variables which separate most of the signal into the signal region while not being strongly correlated to whether or not an event is signal or background. In this study, we use the PFN scores assigned to each event and attempt to find a variable cut that separates separates most signal into the signal region while having a similar background mT_{jj} distribution in each region.

3.1 Results

3.1.1 $\Delta\eta_{12}$ Cut

For the $\Delta\eta_{12}$ cut, we cut events with PFN scores below 0.92 and split events with $\Delta\eta_{12} < 1$ into the signal region.

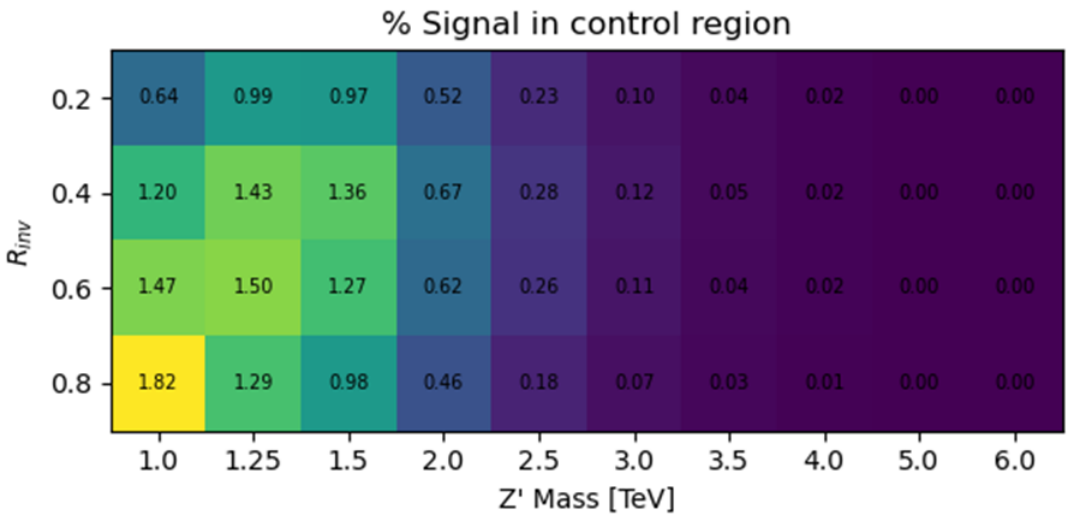


Figure 20: *% Signal in each control group for each signal file*

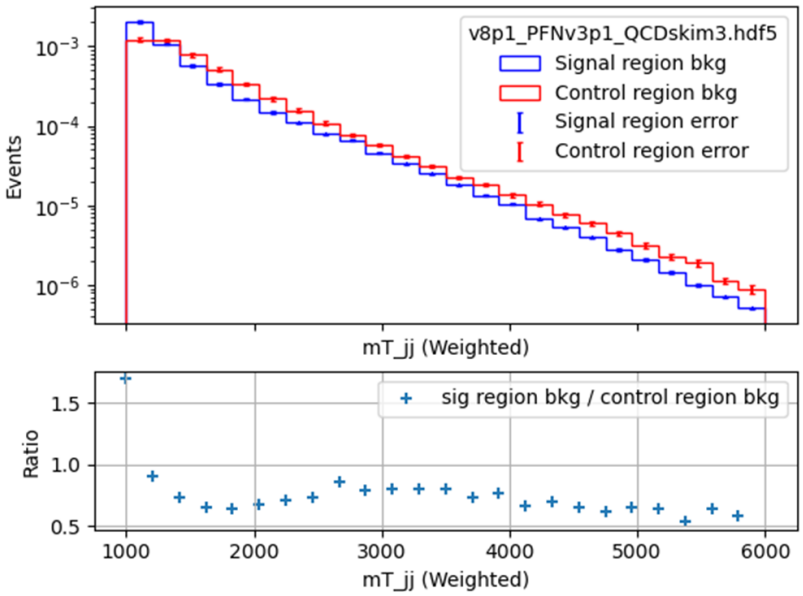


Figure 21: *Normalized background mT_{jj} distribution made with signal region $\Delta\eta_{12} < 1$*

3.1.2 jet2$_{Width}$ 0.05 Cut

For the jet2$_{Width}$ cut, we cut events with PFN scores below 0.97 and split events with jet2$_{Width}$ > 0.05 into the signal region.

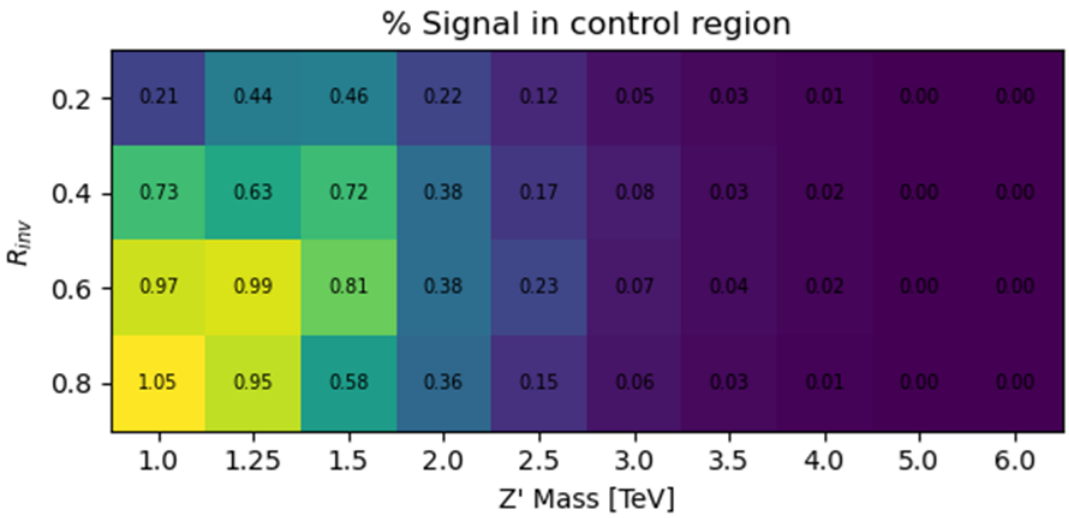


Figure 22: *% signal in each control group for each signal file*

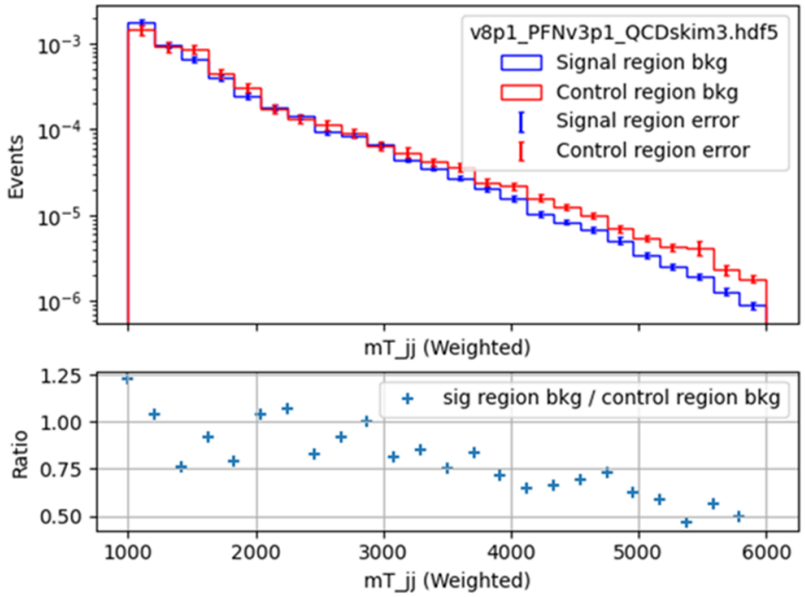


Figure 23: *Normalized background mT_{jj} distribution made with signal region $jet2_{Width} > 0.05$*

3.1.3 jet2_{Width} 0.1 Cut

Lastly we try a jet2_{Width} cut where we cut events with PFN scores below 0.97 and split events with $\text{jet2}_{Width} > 0.1$ into the signal region.

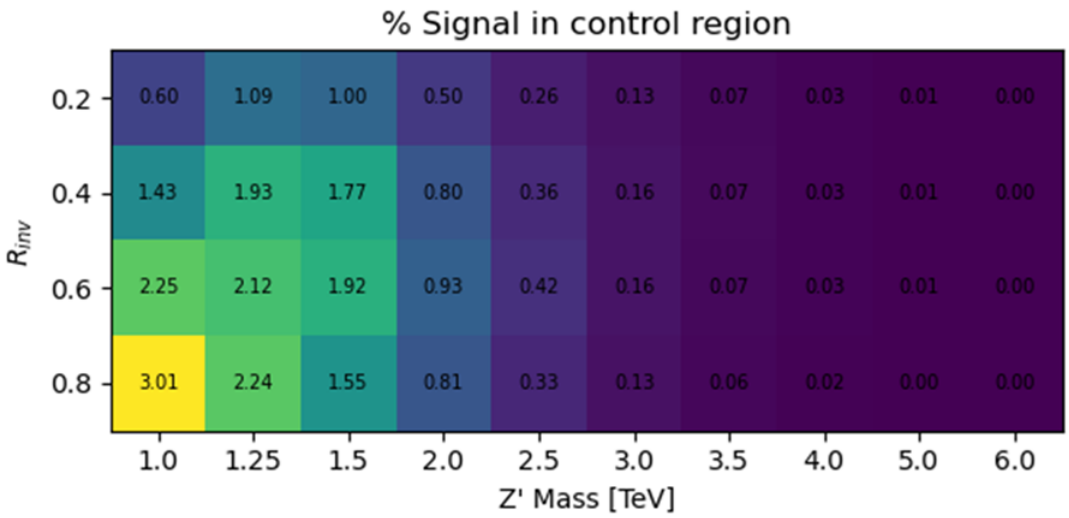


Figure 24: *% signal in each control group for each signal file*

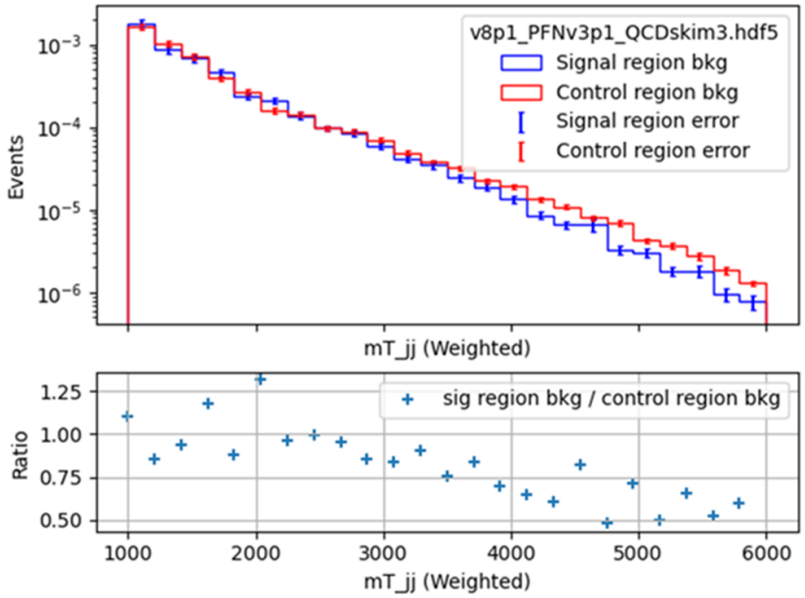


Figure 25: *Normalized background mT_{jj} distribution made with signal region $jet2_{Width} > 0.1$*

3.2 Conclusion

All cuts produced a maximum signal contamination of less than 5%, so all cuts are usable. The $\Delta\eta_{12}$ cut, however, produced an uneven ratio plot as seen in figure 20. There is a larger relative amount background in the signal region with an mT_{jj} of 1000, while the background has a relative larger amount of mT_{jj} above 1000. For the $jet2_{Width}$ signal region cut of > 0.05, the ratio plot is uneven, but is more focused around 1, as opposed to the $\Delta\eta_{12}$ cut. The $jet2_{Width}$ signal region cut of > 0.1 also provides an uneven ratio plot that is relatively focused around 1 and is similar to $jet2_{Width}$ signal region cut of > 0.05. This means that despite using different values for $jet2_{Width}$, the variable produces a similar background mT_{jj} distribution. Further studies, however, should be conducted to determine the ideal variable cut.

4 Conclusion and Next Steps

In this study, we found that PFNs perform better than BDTs when discriminating between SVJ and background events. In a one-to-one comparison, the BDT produced an ROC curve with an AUC of 0.88, while the PFN produced an ROC curve with an AUC of 0.91. We have thus decided to use a PFN to attempt to discriminate between SVJ and background events. We also attempted to find a variable cut that can produce a signal and control region where background events need are split so each region has a similar background mT_{jj}. Currently, jet2_{Width} signal region cuts of > 0.05 and > 0.1 have been found to be the most effective. More studies, however, need to be conducted before a definitive variable cut can be declared.

5 Acknowledgements

I would like to thank Prof. John Parsons, Prof. Georgia Karagiorgi, and Amy Garwood for organizing the REU program. I would like to thank Prof. John Parsons, Dr. Julia Gonski, Kiryeong Park, Gabriel Matos, Elena Busch, and the ATLAS community for their mentorship, patience and support throughout the summer. Lastly, I would like to thank the National Science Foundation for making this research possible. This material is based upon work supported by the National Science Foundation under Grant No. PHY/1950431.

References

[1] CERN. Dark matter. https://home.cern/science/physics/dark-matter.

[2] CERN. The large hadron collider. https://home.cern/science/accelerators/large-hadron-collider.

[3] CERN. Lhc the guide faq. https://home.cern/resources/brochure/knowledge-sharing/lhc-facts-and-figures.

[4] CERN. Run 3 of the large hadron collider. https://home.cern/press/2022/run-3.

[5] CERN. The standard model. https://home.cern/science/physics/standard-model.

[6] Yann Coadou. Boosted decision trees. In *Artificial Intelligence for High Energy Physics*, pages 9–58. WORLD SCIENTIFIC, feb 2022.

[7] ATLAS collaboration. Not a jet all the way: Is dark matter hiding in plain sight? https://atlas.cern/Updates/Briefing/Semi-Visible-Jets, May 2023.

[8] Cesare Cazzaniga & Annapaola de Cosa. Leptons lurking in semi-visible jets at the lhc. *Eur. Phys. J.*, 2022.

[9] A. Tumasyan et al. Search for resonant production of strongly coupled dark matter in proton-proton collisions at 13 TeV. *Journal of High Energy Physics*, 2022(6), jun 2022.

[10] ATLAS Experiment. Calorimeter. https://atlas.cern/Discover/Detector/Calorimeter.

[11] ATLAS Experiment. Glossary. http://opendata.atlas.cern/books/current/get-started/_book/GLOSSARY.html.

[12] ATLAS Experiment. The inner detector. https://atlas.cern/Discover/Detector/Inner-Detector.

[13] ATLAS Experiment. Muon spectrometer. https://atlas.cern/Discover/Detector/Muon-Spectrometer.

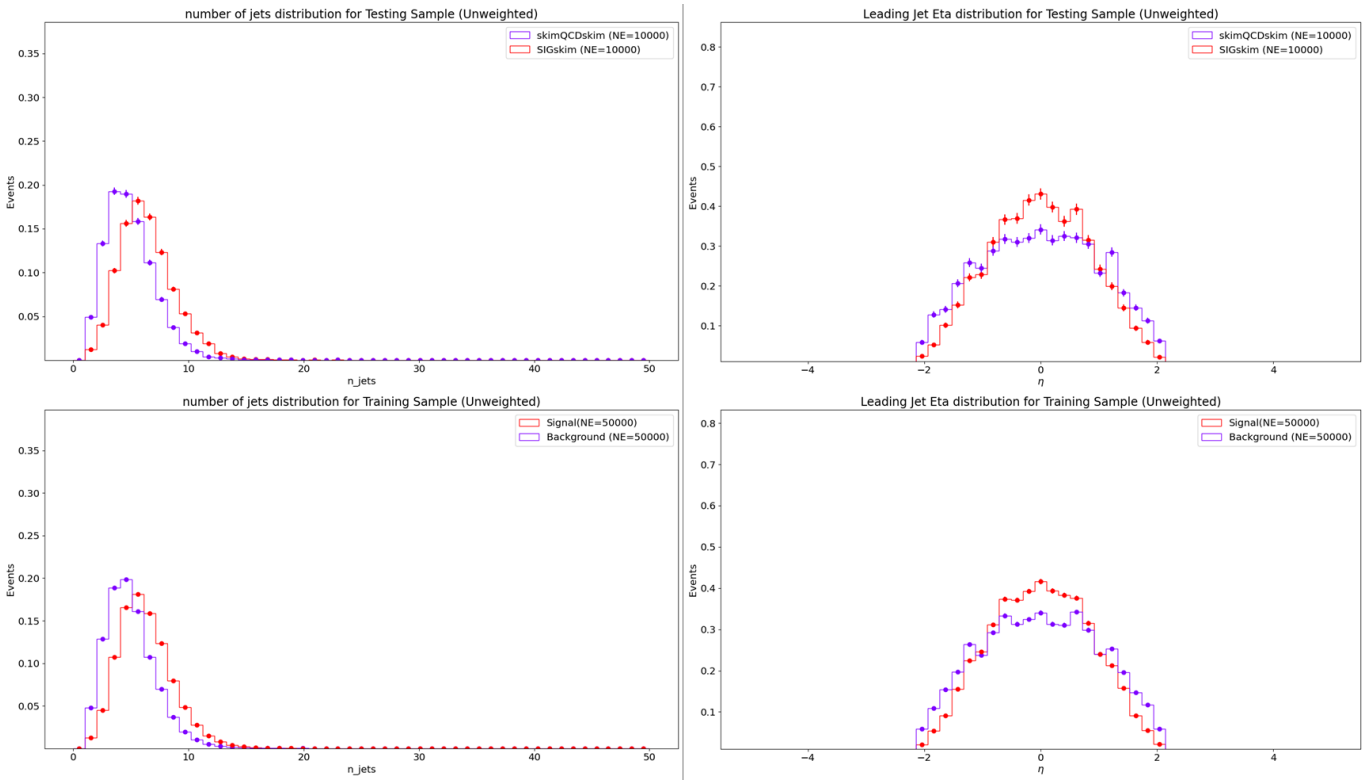
[14] JavaTPoint. Decision tree classification algorithm. https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm.

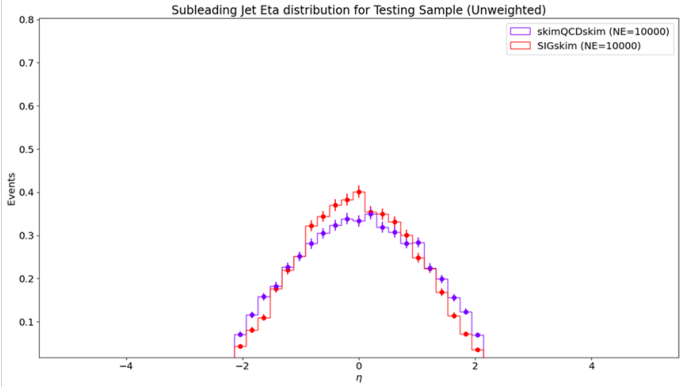
[15] Rebeca Gonzalez Suarez Katarina Anthony, Sascha Mehlhase and Ana Maria Rodriguez Vera on behalf of the ATLAS Collaboration. Atlas fact sheets. https://atlas.cern/Resources/Fact-sheets.

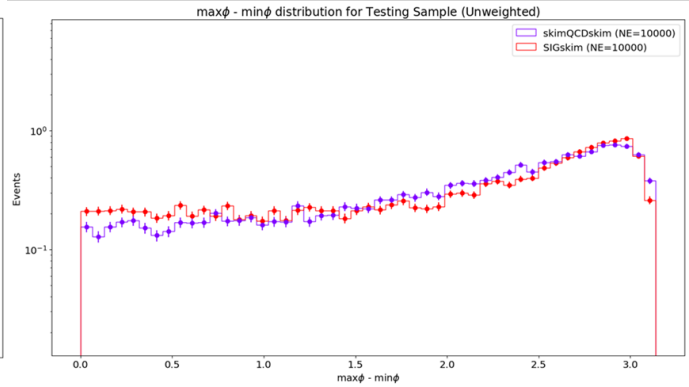
[16] Patrick T. Komiske, Eric M. Metodiev, and Jesse Thaler. Energy flow networks: deep sets for particle jets. *Journal of High Energy Physics*, 2019(1), jan 2019.

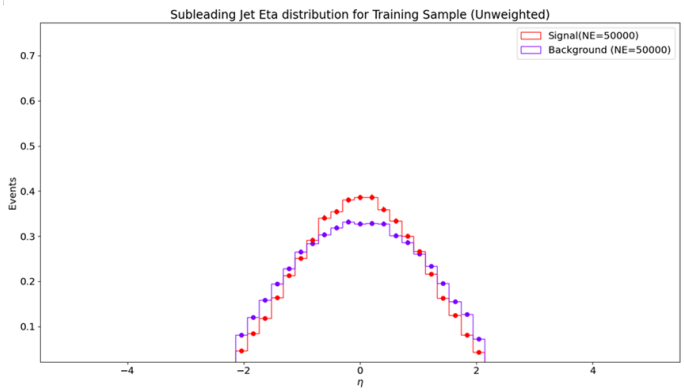
6 Appendix

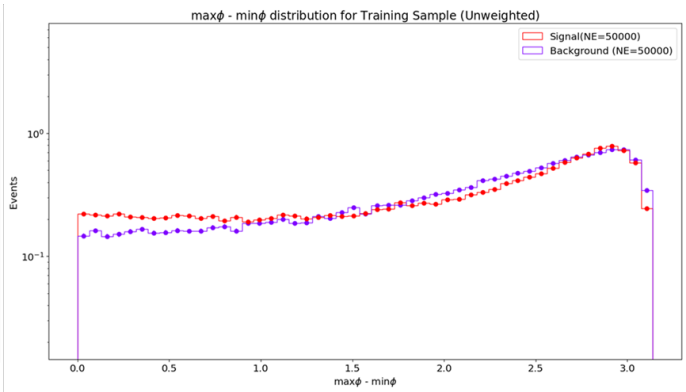
6.1 BDT Input Plots

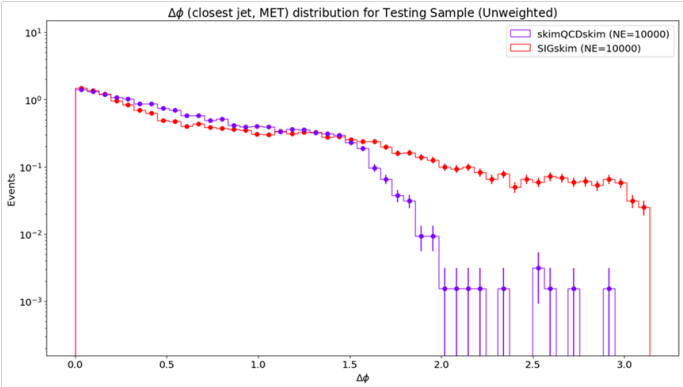


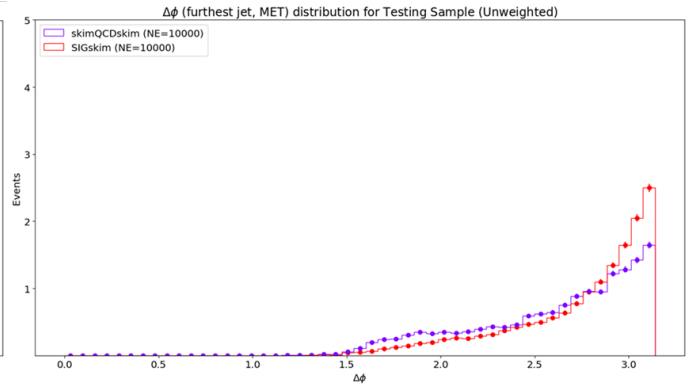

Subleading Jet Eta distribution for Testing Sample (Unweighted)
skimQCDskim (NE=10000)
SIGskim (NE=10000)
Events
η

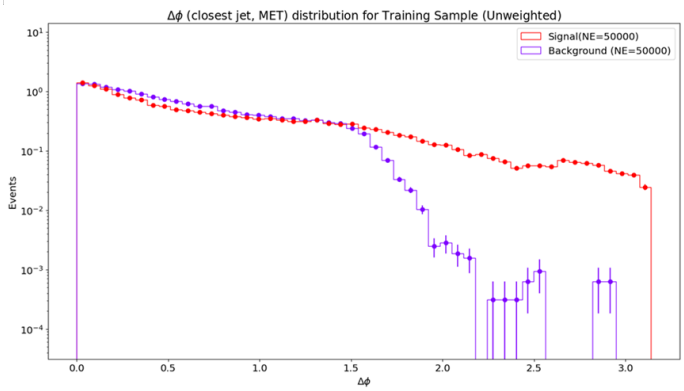

maxϕ - minϕ distribution for Testing Sample (Unweighted)
skimQCDskim (NE=10000)
SIGskim (NE=10000)
Events
maxϕ - minϕ

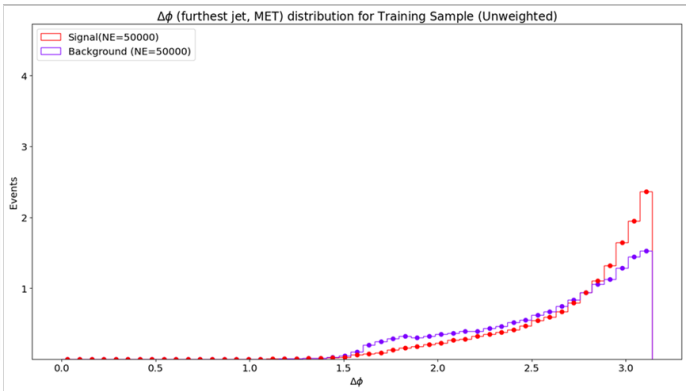

Subleading Jet Eta distribution for Training Sample (Unweighted)
Signal(NE=50000)
Background (NE=50000)
Events
η

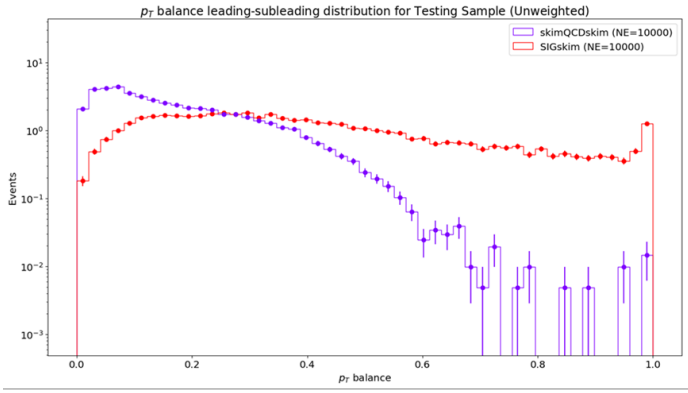

maxϕ - minϕ distribution for Training Sample (Unweighted)
Signal(NE=50000)
Background (NE=50000)
Events
maxϕ - minϕ

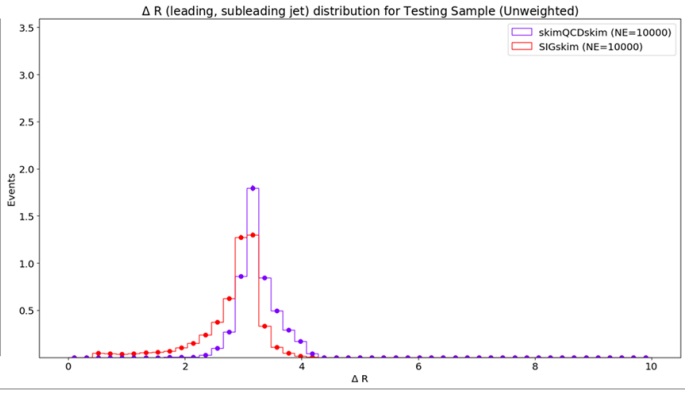

Δϕ (closest jet, MET) distribution for Testing Sample (Unweighted)
skimQCDskim (NE=10000)
SIGskim (NE=10000)
Events
Δϕ

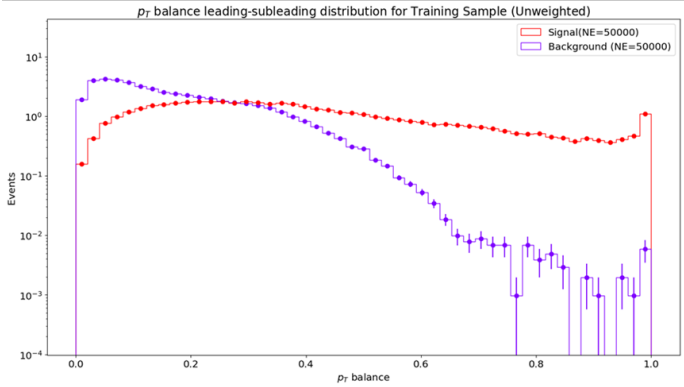

Δϕ (furthest jet, MET) distribution for Testing Sample (Unweighted)
skimQCDskim (NE=10000)
SIGskim (NE=10000)
Events
Δϕ

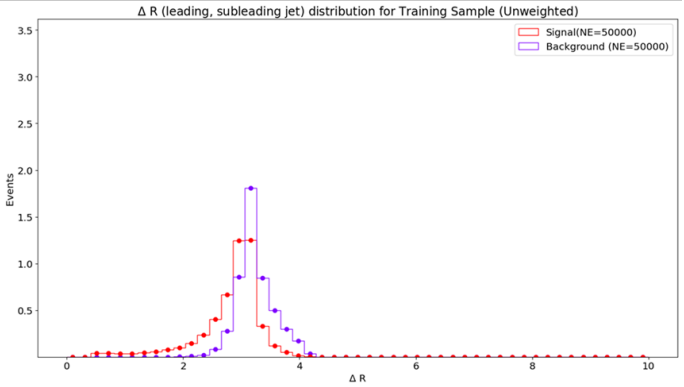

Δϕ (closest jet, MET) distribution for Training Sample (Unweighted)
Signal(NE=50000)
Background (NE=50000)
Events
Δϕ

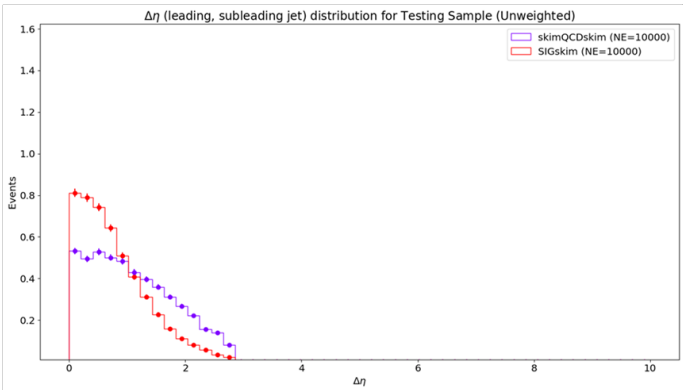

Δϕ (furthest jet, MET) distribution for Training Sample (Unweighted)
Signal(NE=50000)
Background (NE=50000)
Events
Δϕ

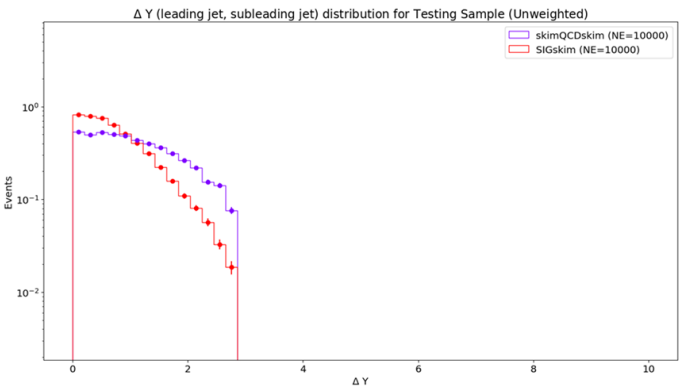

p_T balance leading-subleading distribution for Testing Sample (Unweighted)
skimQCDskim (NE=10000)
SIGskim (NE=10000)
Events
p_T balance

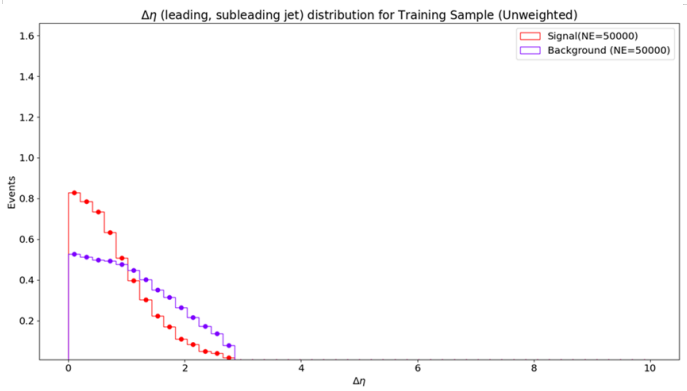

Δ R (leading, subleading jet) distribution for Testing Sample (Unweighted)
skimQCDskim (NE=10000)
SIGskim (NE=10000)
Events
Δ R

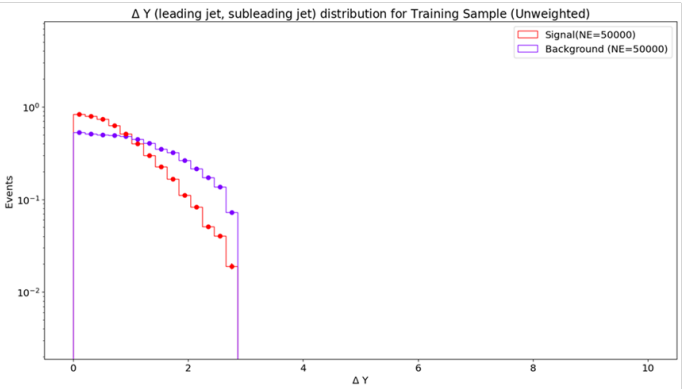

p_T balance leading-subleading distribution for Training Sample (Unweighted)
Signal(NE=50000)
Background (NE=50000)
Events
p_T balance

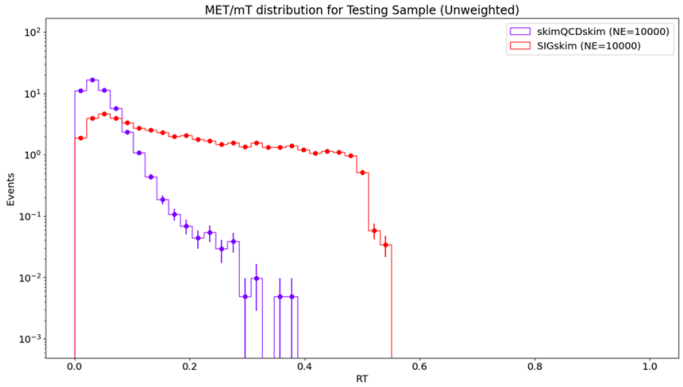

Δ R (leading, subleading jet) distribution for Training Sample (Unweighted)
Signal(NE=50000)
Background (NE=50000)
Events
Δ R

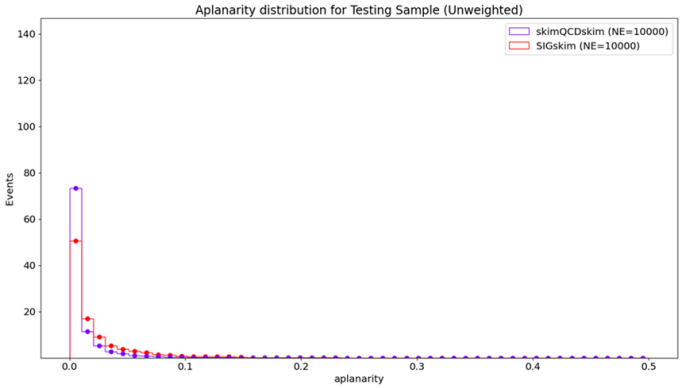

Δη (leading, subleading jet) distribution for Testing Sample (Unweighted)
skimQCDskim (NE=10000)
SIGskim (NE=10000)
Events
Δη

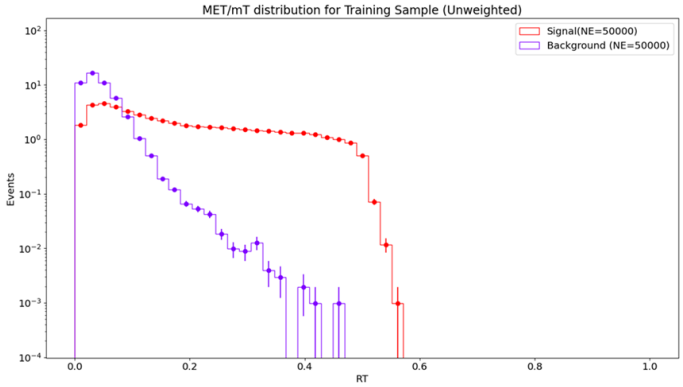

Δ Y (leading jet, subleading jet) distribution for Testing Sample (Unweighted)
skimQCDskim (NE=10000)
SIGskim (NE=10000)
Events
Δ Y

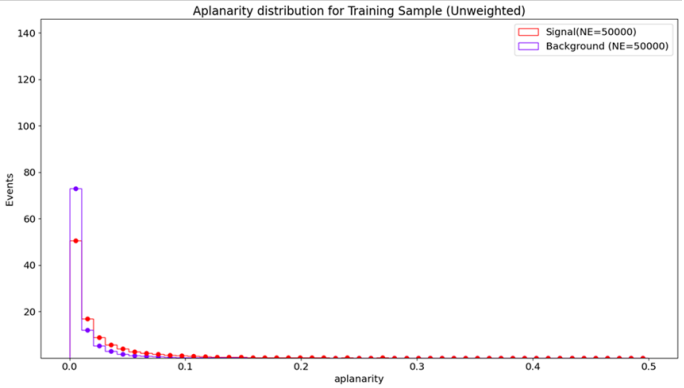

Δη (leading, subleading jet) distribution for Training Sample (Unweighted)
Signal(NE=50000)
Background (NE=50000)
Events
Δη

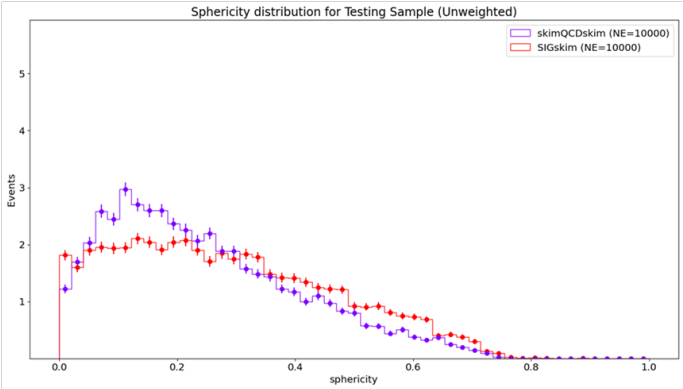

Δ Y (leading jet, subleading jet) distribution for Training Sample (Unweighted)
Signal(NE=50000)
Background (NE=50000)
Events
Δ Y

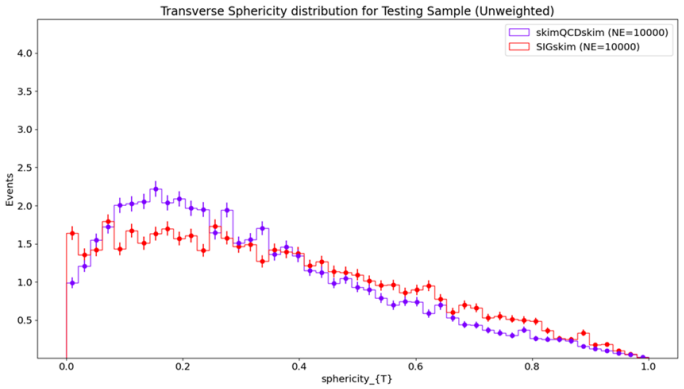

MET/mT distribution for Testing Sample (Unweighted)
skimQCDskim (NE=10000)
SIGskim (NE=10000)
Events
RT

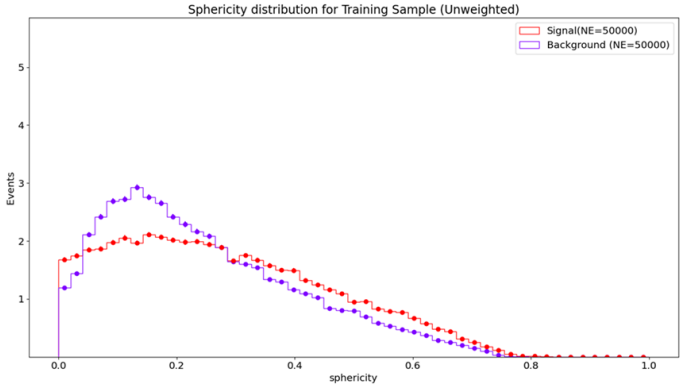

Aplanarity distribution for Testing Sample (Unweighted)
skimQCDskim (NE=10000)
SIGskim (NE=10000)
Events
aplanarity

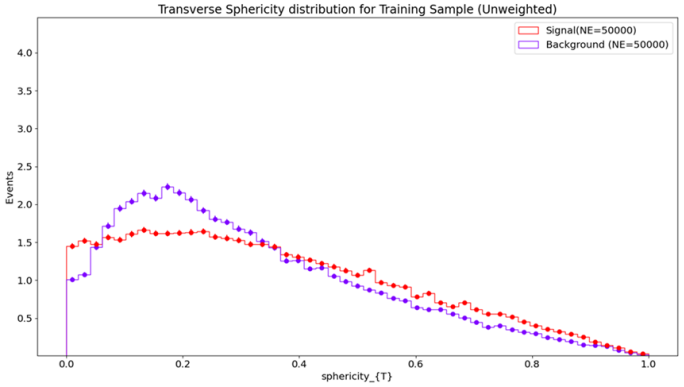

MET/mT distribution for Training Sample (Unweighted)
Signal(NE=50000)
Background (NE=50000)
Events
RT

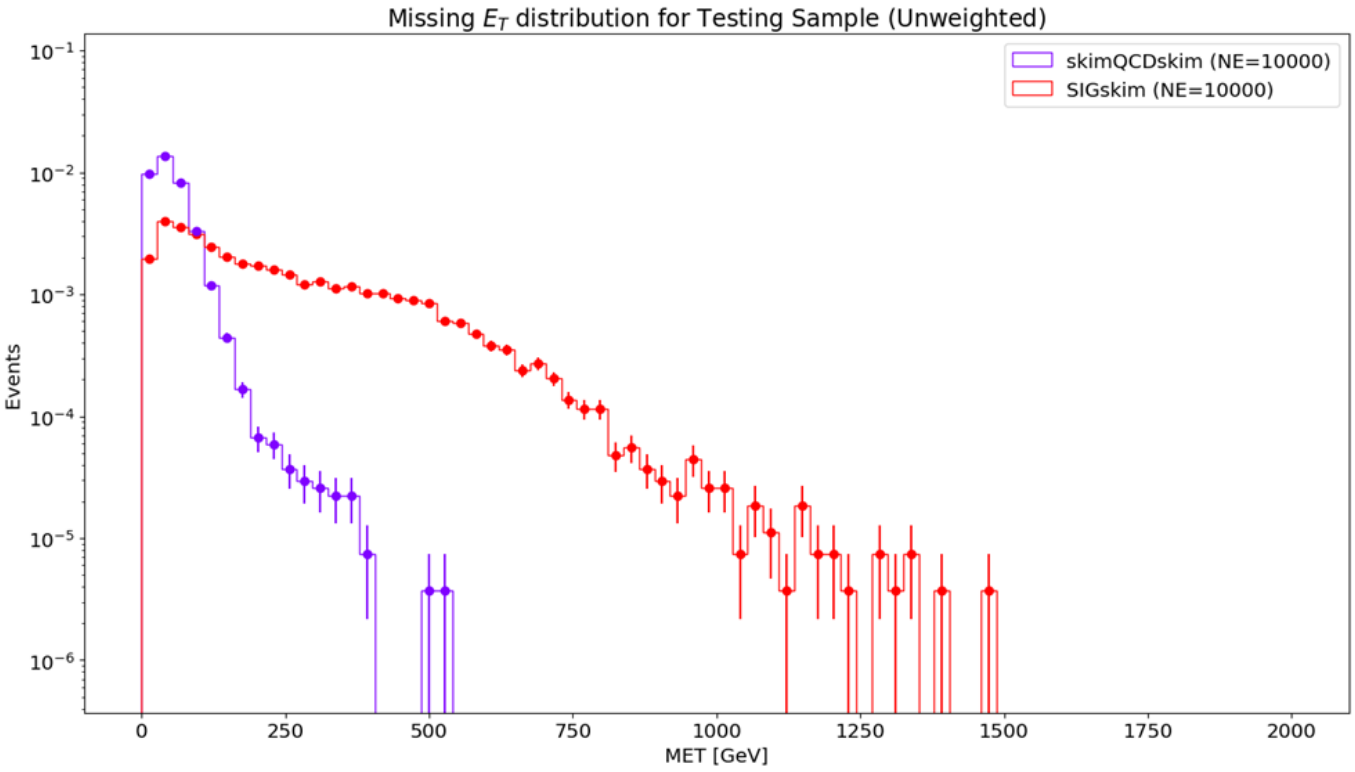

Aplanarity distribution for Training Sample (Unweighted)
Signal(NE=50000)
Background (NE=50000)
Events
aplanarity

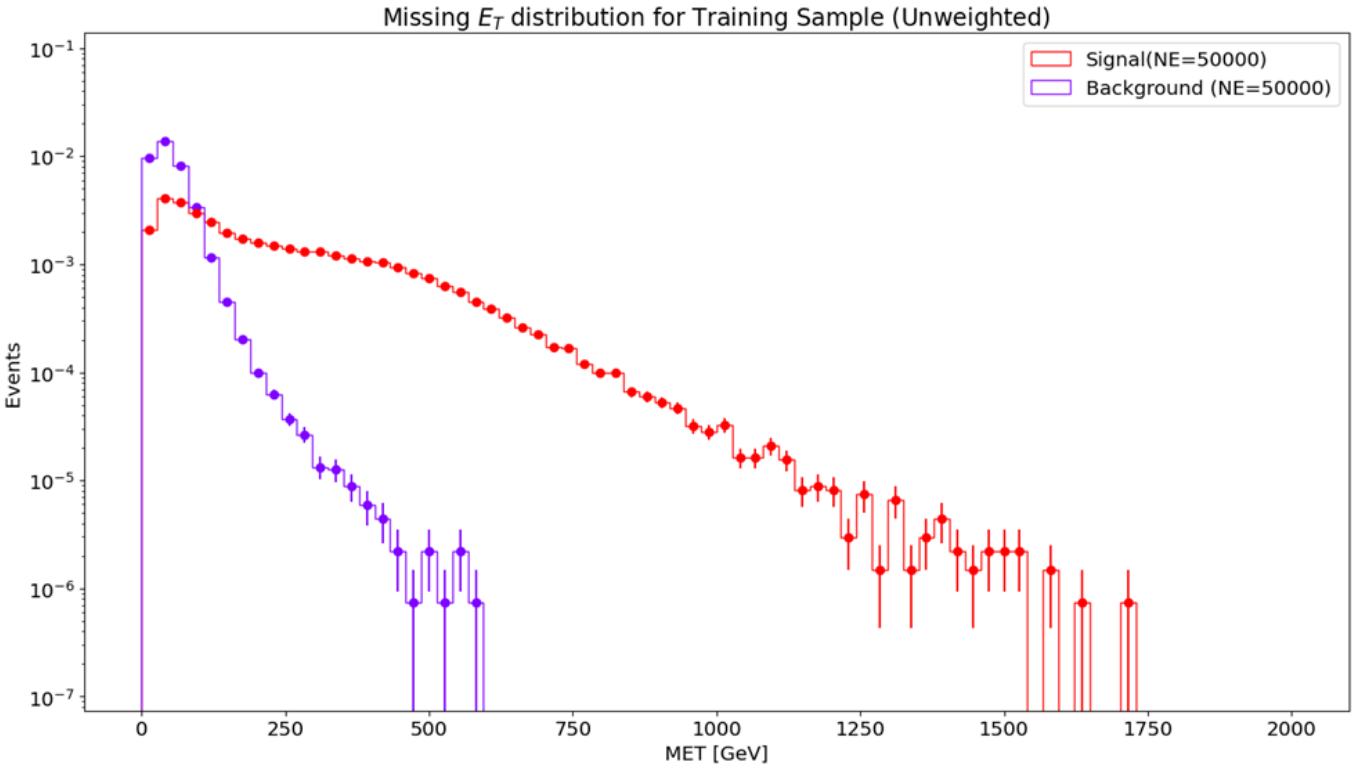

Sphericity distribution for Testing Sample (Unweighted)
skimQCDskim (NE=10000)
SIGskim (NE=10000)
Events
sphericity

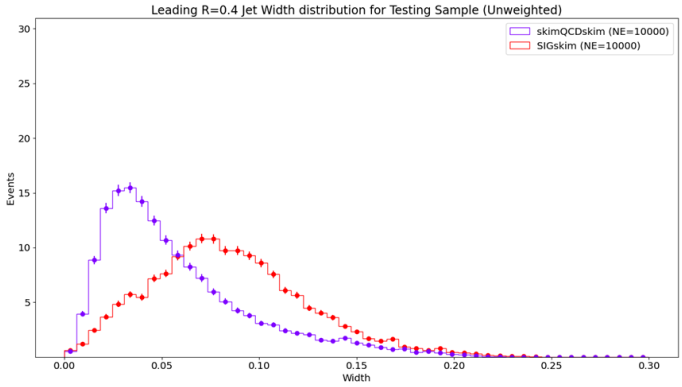

Transverse Sphericity distribution for Testing Sample (Unweighted)
skimQCDskim (NE=10000)
SIGskim (NE=10000)
Events
sphericity_{T}

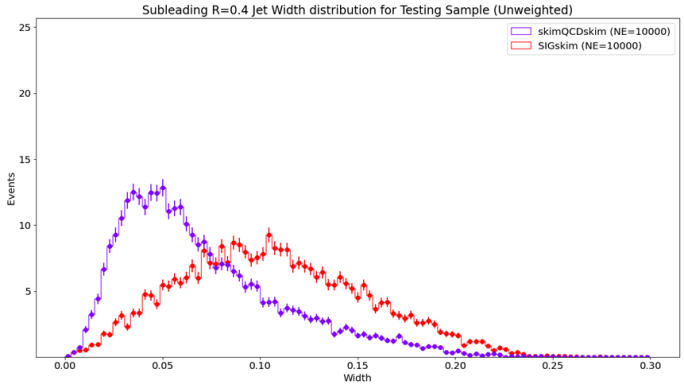

Sphericity distribution for Training Sample (Unweighted)
Signal(NE=50000)
Background (NE=50000)
Events
sphericity

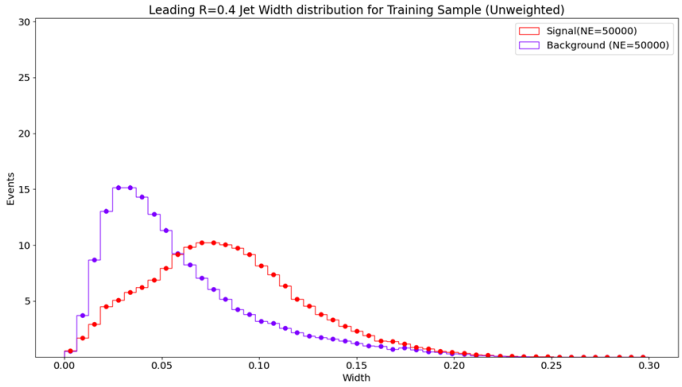

Transverse Sphericity distribution for Training Sample (Unweighted)
Signal(NE=50000)
Background (NE=50000)
Events
sphericity_{T}

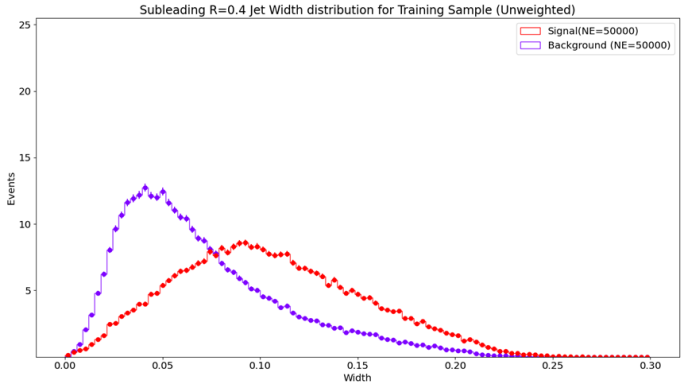

Missing E_T distribution for Testing Sample (Unweighted)
skimQCDskim (NE=10000)
SIGskim (NE=10000)
Events
MET [GeV]

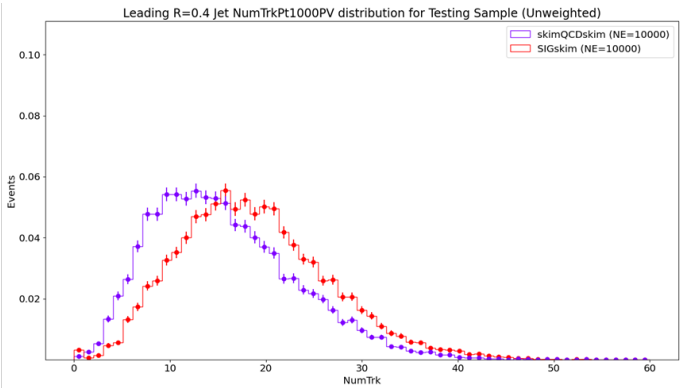

Missing E_T distribution for Training Sample (Unweighted)
Signal(NE=50000)
Background (NE=50000)
Events
MET [GeV]

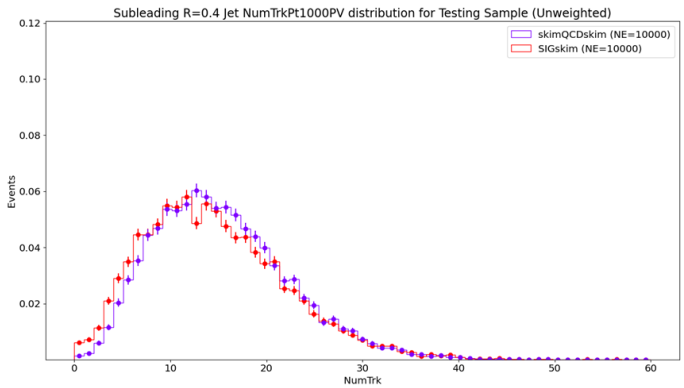

Leading R=0.4 Jet Width distribution for Testing Sample (Unweighted)
skimQCDskim (NE=10000)
SIGskim (NE=10000)
Events
Width

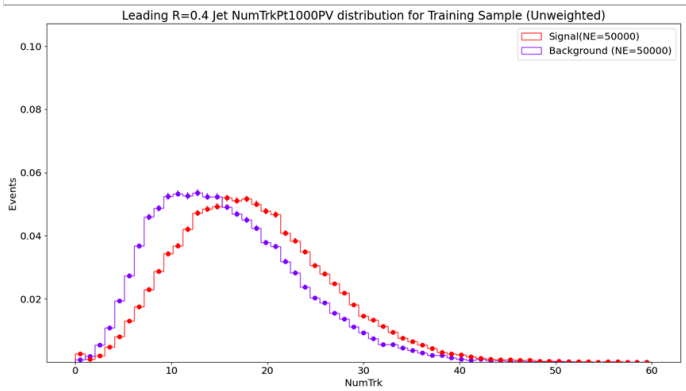

Subleading R=0.4 Jet Width distribution for Testing Sample (Unweighted)
skimQCDskim (NE=10000)
SIGskim (NE=10000)
Events
Width

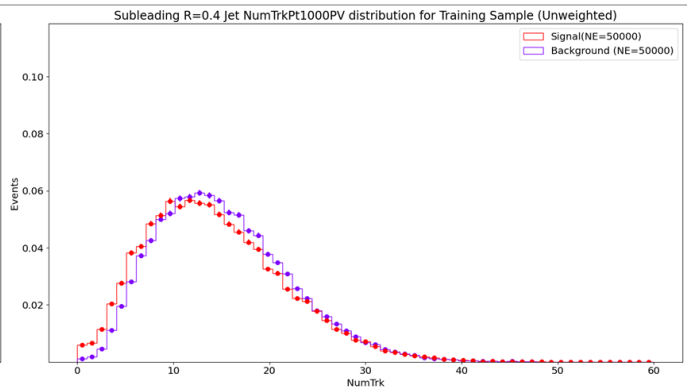

Leading R=0.4 Jet Width distribution for Training Sample (Unweighted)
Signal(NE=50000)
Background (NE=50000)
Events
Width

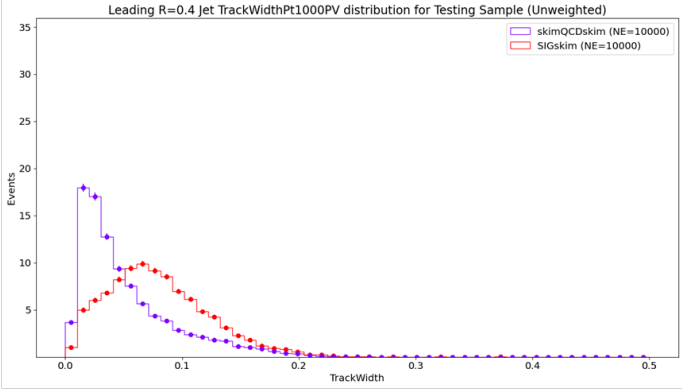

Subleading R=0.4 Jet Width distribution for Training Sample (Unweighted)
Signal(NE=50000)
Background (NE=50000)
Events
Width

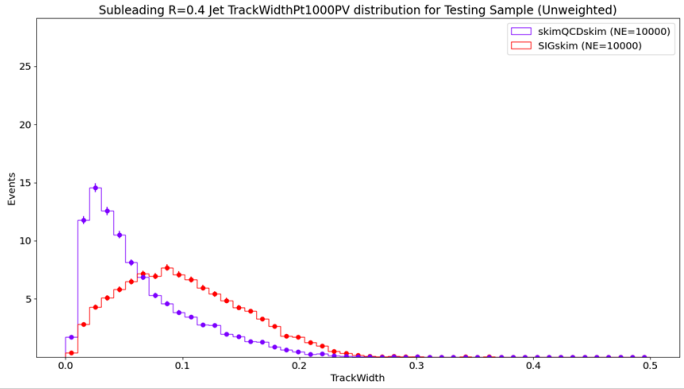

Leading R=0.4 Jet NumTrkPt1000PV distribution for Testing Sample (Unweighted)
skimQCDskim (NE=10000)
SIGskim (NE=10000)
Events
NumTrk

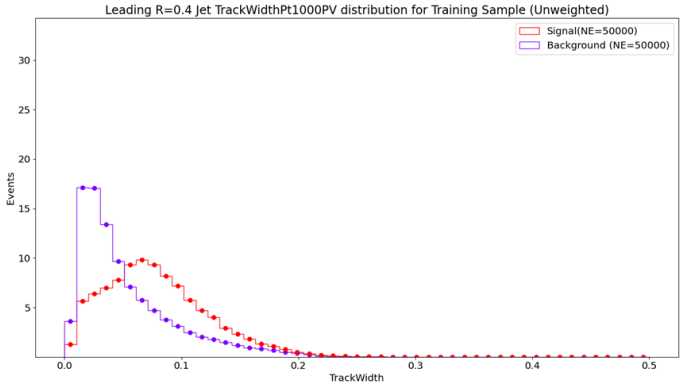

Subleading R=0.4 Jet NumTrkPt1000PV distribution for Testing Sample (Unweighted)
skimQCDskim (NE=10000)
SIGskim (NE=10000)
Events
NumTrk

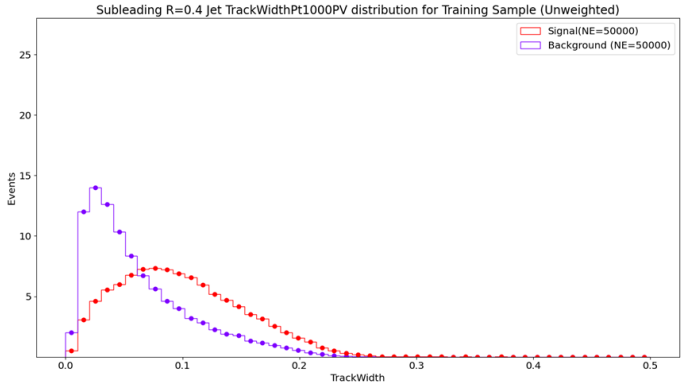

Leading R=0.4 Jet NumTrkPt1000PV distribution for Training Sample (Unweighted)
Signal(NE=50000)
Background (NE=50000)
Events
NumTrk

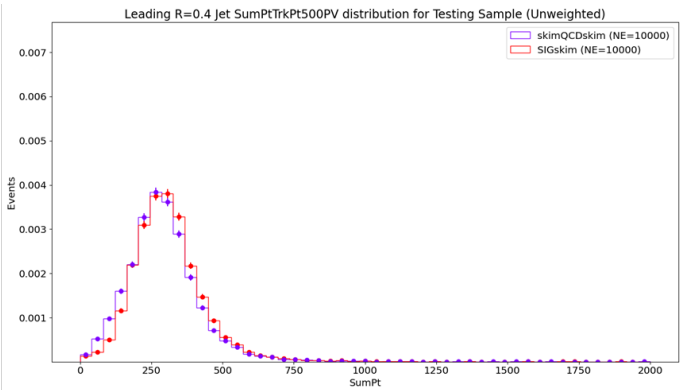

Subleading R=0.4 Jet NumTrkPt1000PV distribution for Training Sample (Unweighted)
Signal(NE=50000)
Background (NE=50000)
Events
NumTrk

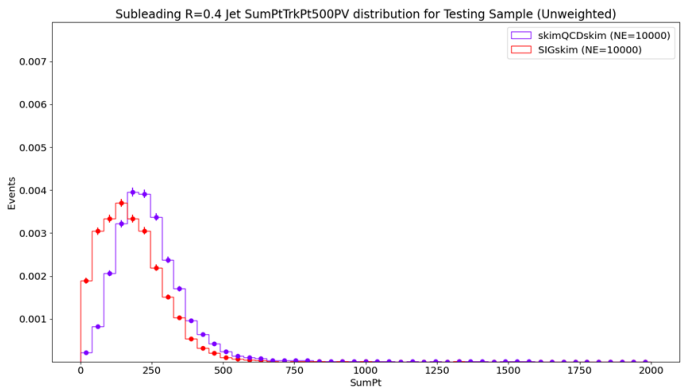

Leading R=0.4 Jet TrackWidthPt1000PV distribution for Testing Sample (Unweighted)
skimQCDskim (NE=10000)
SIGskim (NE=10000)
Events
TrackWidth

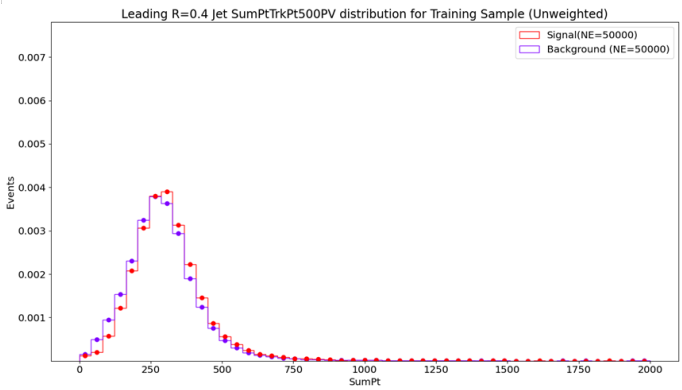

Subleading R=0.4 Jet TrackWidthPt1000PV distribution for Testing Sample (Unweighted)
skimQCDskim (NE=10000)
SIGskim (NE=10000)
Events
TrackWidth

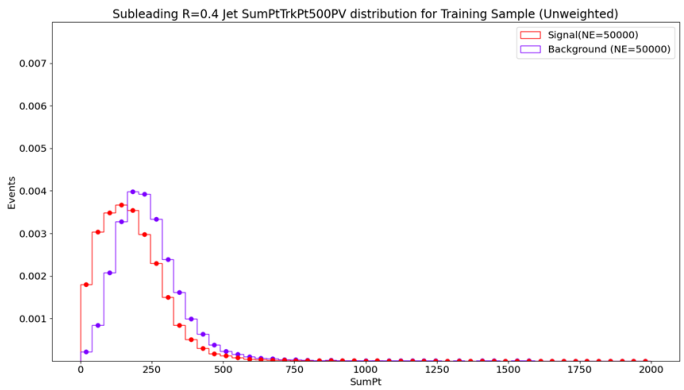

Leading R=0.4 Jet TrackWidthPt1000PV distribution for Training Sample (Unweighted)
Signal(NE=50000)
Background (NE=50000)
Events
TrackWidth


Subleading R=0.4 Jet TrackWidthPt1000PV distribution for Training Sample (Unweighted)
Signal(NE=50000)
Background (NE=50000)
Events
TrackWidth


Leading R=0.4 Jet SumPtTrkPt500PV distribution for Testing Sample (Unweighted)
skimQCDskim (NE=10000)
SIGskim (NE=10000)
Events
SumPt


Subleading R=0.4 Jet SumPtTrkPt500PV distribution for Testing Sample (Unweighted)
skimQCDskim (NE=10000)
SIGskim (NE=10000)
Events
SumPt


Leading R=0.4 Jet SumPtTrkPt500PV distribution for Training Sample (Unweighted)
Signal(NE=50000)
Background (NE=50000)
Events
SumPt


Subleading R=0.4 Jet SumPtTrkPt500PV distribution for Training Sample (Unweighted)
Signal(NE=50000)
Background (NE=50000)
Events
SumPt