Machine Learning Methods for Detecting Semi-Visible Jets

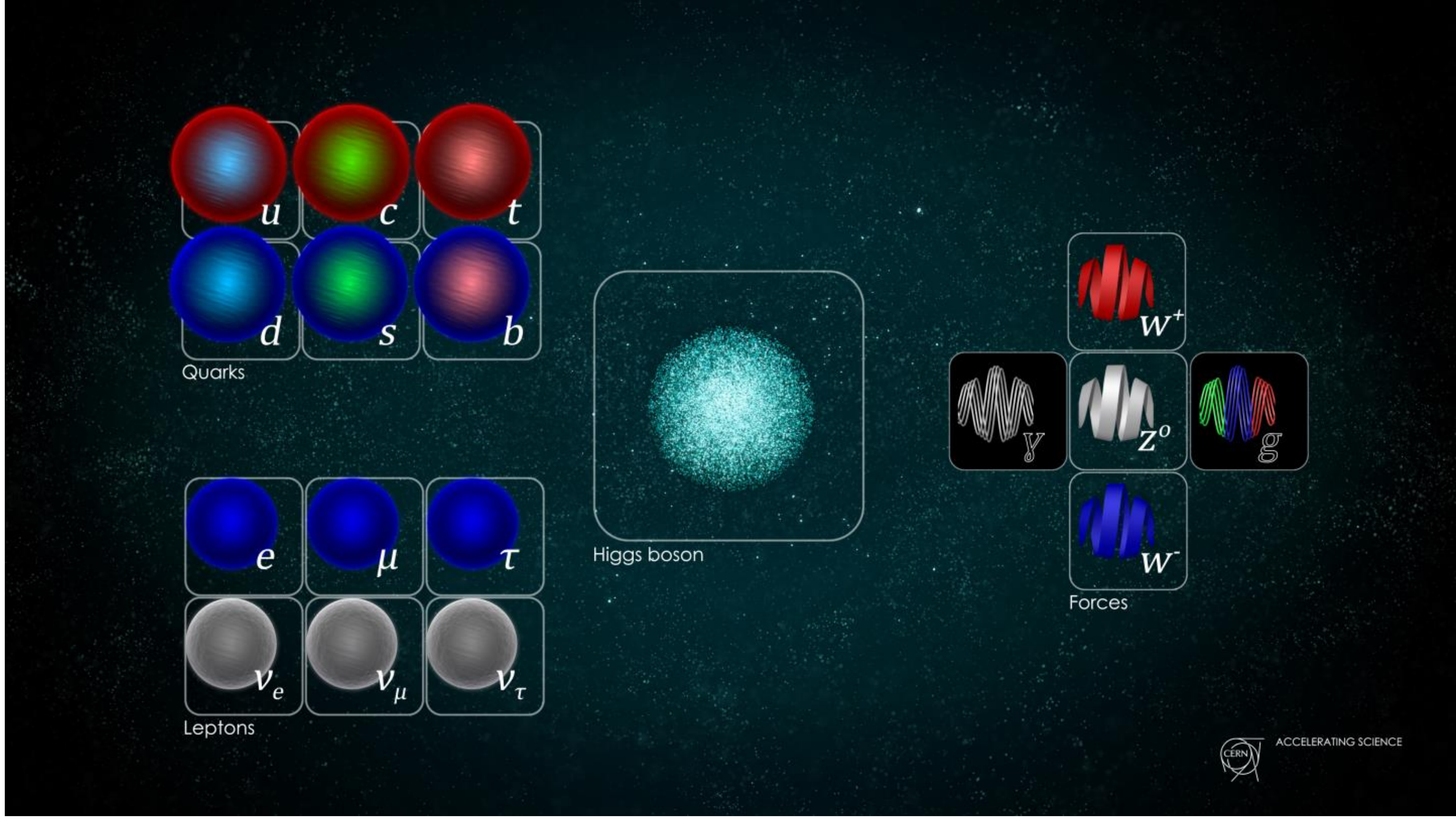
Jonah Mougoue

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.

SM shows that there are three types of fundamental particles Fermions, gauge bosons, and Higgs Bosons.

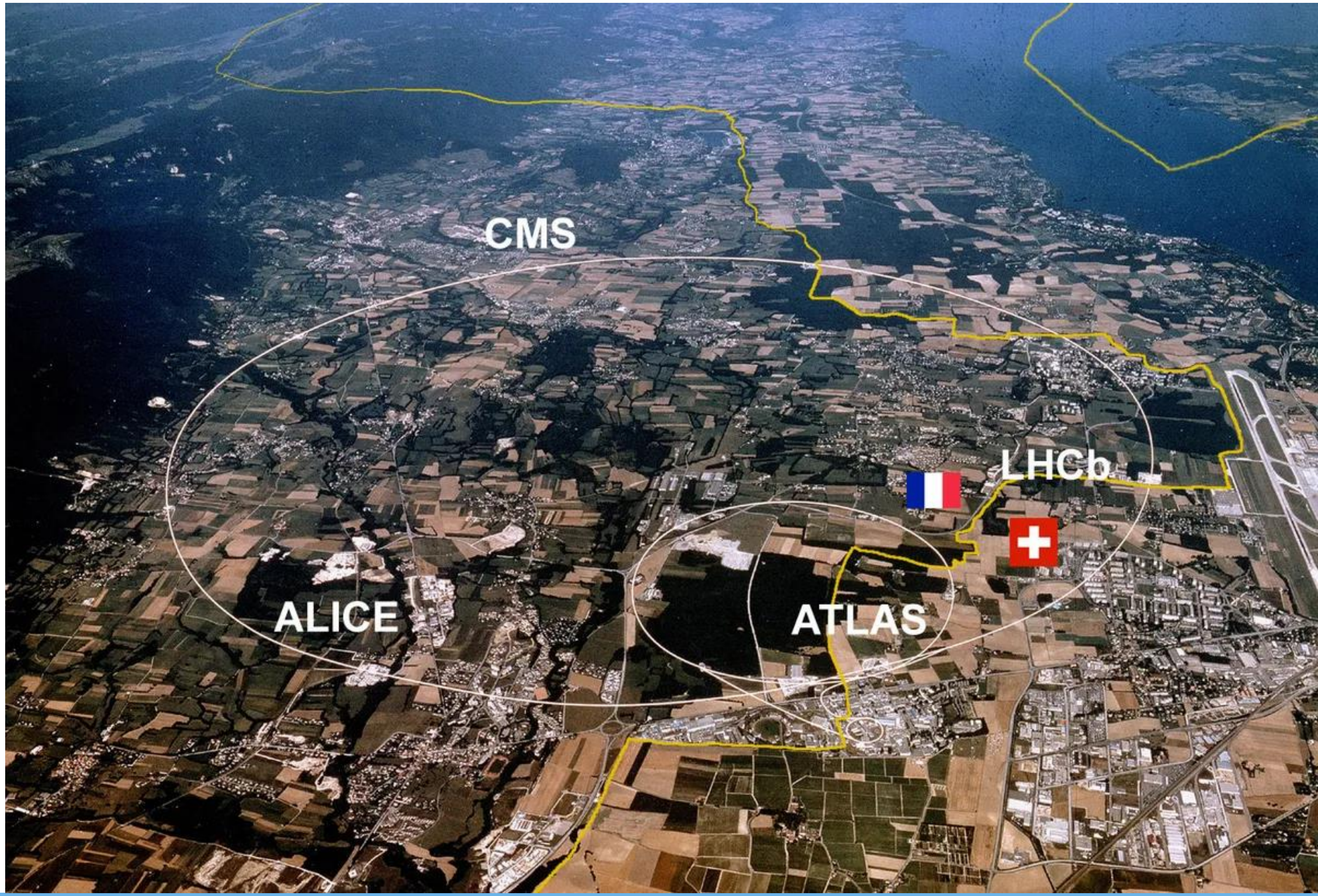
- Fermions – Particles that form matter
  - Quarks – Form hadrons
  - Leptons – Neutrinos, electrons, muons, and taus
- Guage Bosons – force carrying particles
  - Photons (γ) - Mediate the Electromagnetic (EM) force
  - Gluons(g) – Mediates the strong force
  - Z and $W^{\pm}$ bosons - Mediates the weak force
- Higgs Boson – Gives particle mass



While SM has been largely successful, it fails to account for gravity. There is no particle in SM that mediates gravity. Additionally, it doesn't account for the abundance of mass we can't observe. This unobservable mass, called dark matter, has been shown to be more abundant than SM matter.
Can dark matter interact with SM particles?

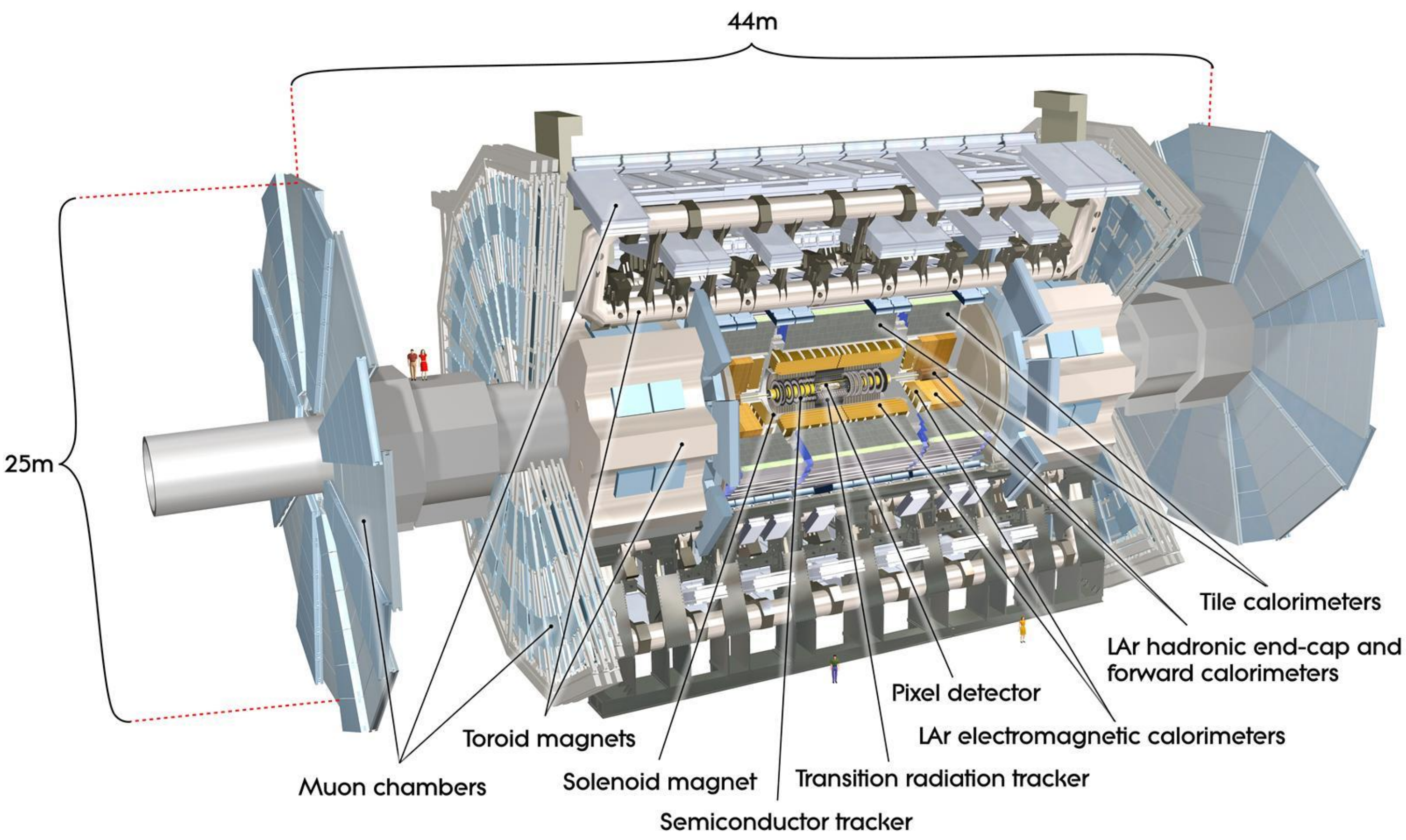
Large Hadron Collider

- The Large Hadron Collider (LHC) is the largest and most powerful particle collider in the world
- Currently on it's 3rd run, the LHC can produce collisions with an energy of 13.6 TeV
- Sends particles to one of its for detectors to record collisions



A Toroidal LHC ApparatuS (ATLAS)

- General purpose detector within the LHC
- Contains 4 detectors
  - Inner detector – Detects origin, momentum, tracks, and particle type
  - Liquid Argon Calorimeter – Detects energy of electrons and photons
  - Tile Hadronic Calorimeter – Detects energy of hadrons
  - Muon Spectrometer – Detects Muons



ATLAS Data Collection

- Low-level data contains track-level information taken directly from the detector
- High-level data is low-level data that has been reconstructed to get physical data about the jets (ex: mass, momentum, etc.)
- ATLAS is able to reconstruct high-level data using the low-level data collected from the detector
- ATLAS collaboration simulates Monte Carlo data, which we use for the analysis
- Many collisions contain missing energy (MET)

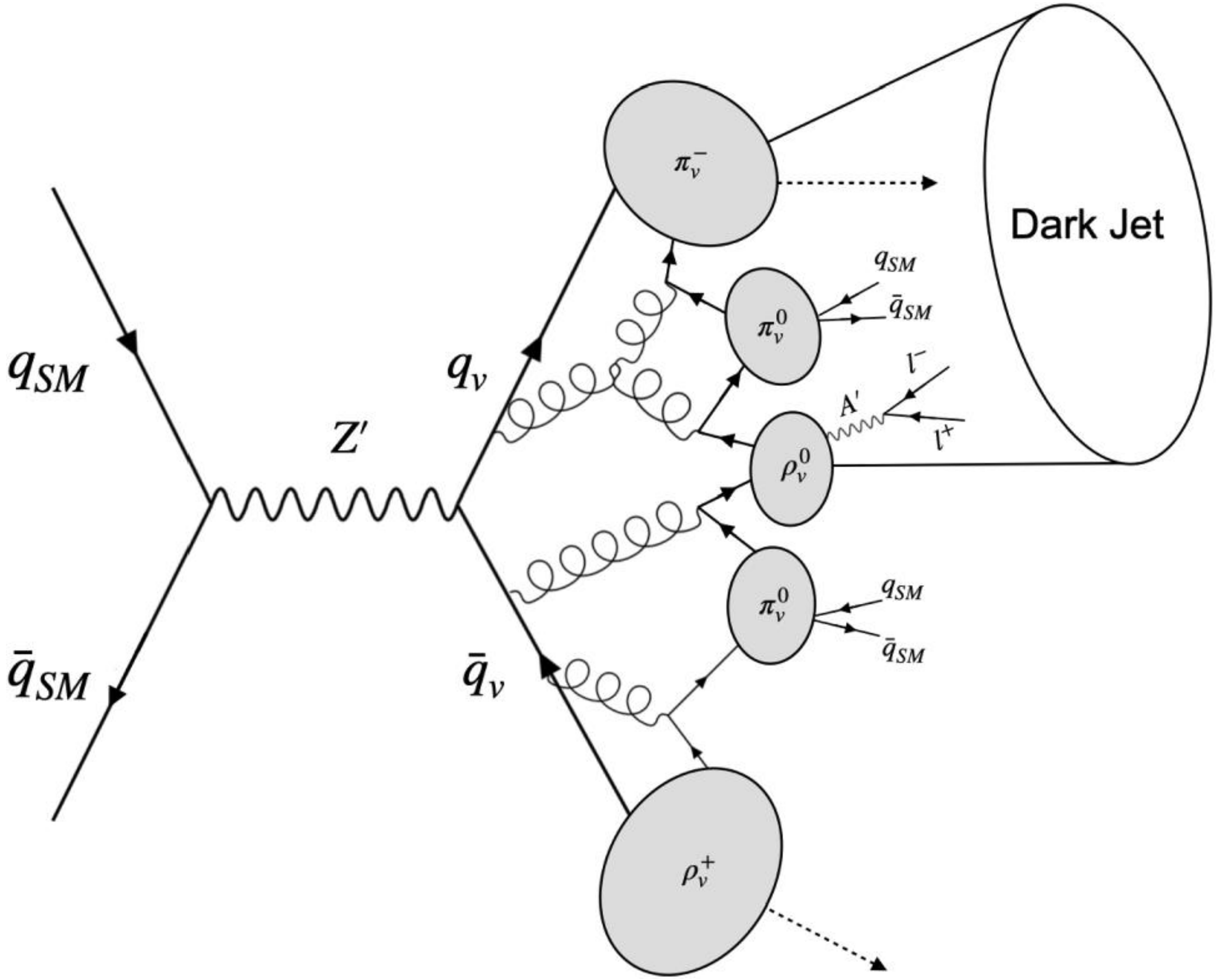
Semi-Visible Jets

Jets

- Stream of hadrons produced from quarks or gluons

Semi-Visible Jets (SVJs)

- Theorized result of Z' boson decay
- Z' bosons possibly created in proton-proton collisions
- Contain both SM particles and dark matter particles
- Since dark matter can't be detected, SVJs must contain energy that is invisible to the detector



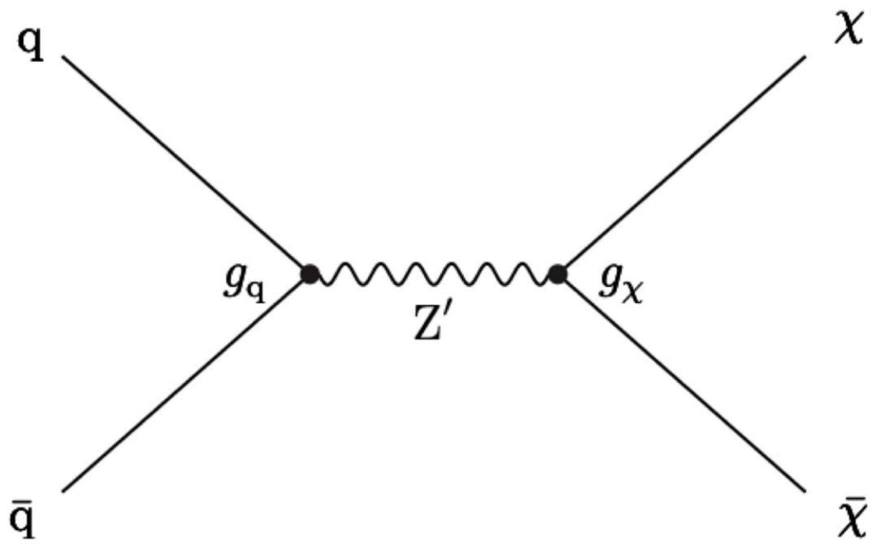
SVJ Properties

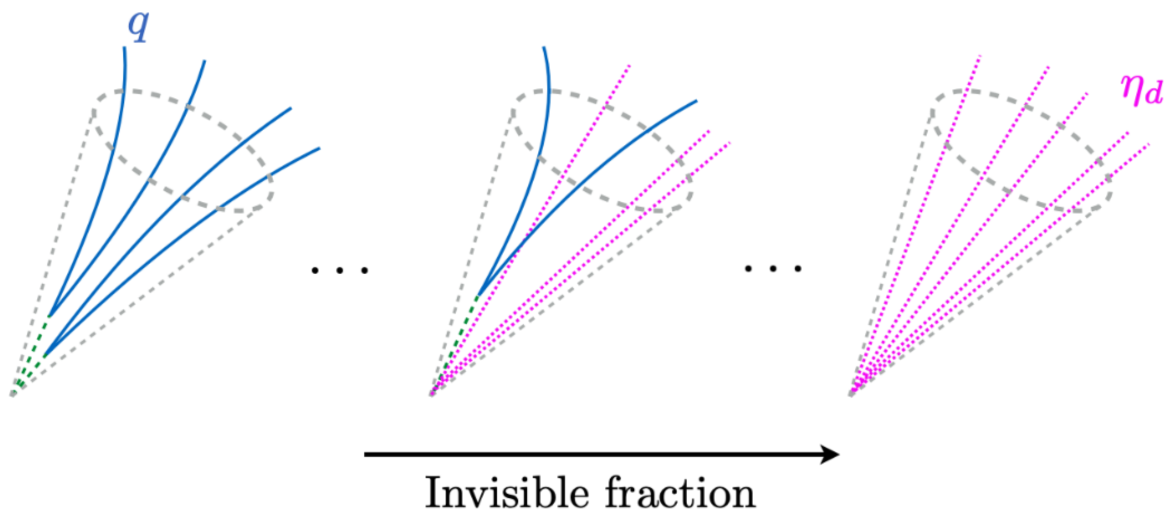
r_{inv}

- Fraction of energy that is carried by dark matter particles
- High r_{inv} means less SM hadrons and higher MET

Z’ mass

- Mass of the intermediate boson
- High Z’ mass means more energy, but also higher MET if r_{inv} is high





Why Study SVJs?

SVJs can give us insight into the nature of dark matter and how SM matter interacts with dark matter

Many events recorded in ATLAS have recorded MET, but this is often due to mismeasured SM jets

Since we’ve never detected an SVJ, how do we know how to find one?

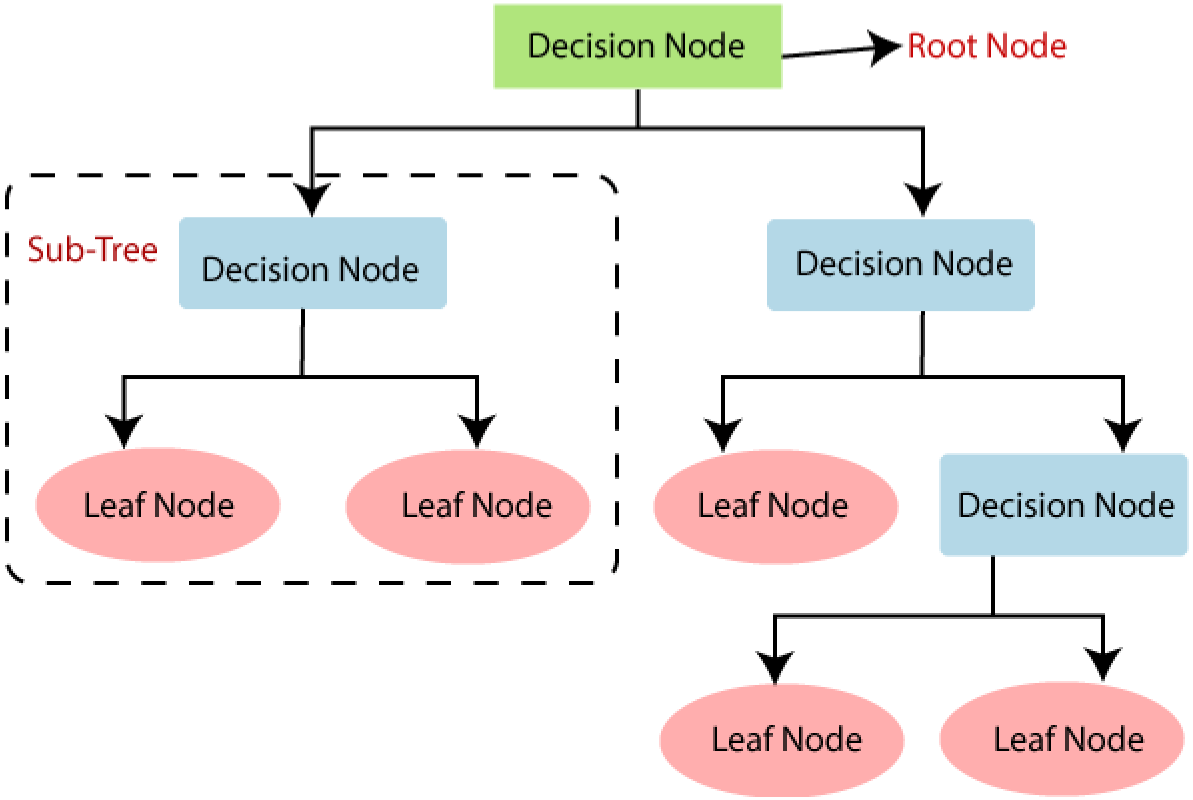
Boosted Decision Tree

Decision trees

- A powerful machine learning model
- Splits data into signal and background based on cuts on different variables and chooses the cut that most accurately splits the data

Boosting

- Improved way of finding good cuts
- Looks at multiple weak classifiers (decision trees that perform slightly better than random) and iterates over them to attempt to make a strong classifier



BDT vs PFN

In 2022, Compact Muon Solenoid (CMS) attempted a search for SVJs using a BDT, but found none

Are BDTs the best machine learning model for finding SVJs?

Particle Flow Network (PFN) is a neural network that uses low-level track data from the leading and subleading jets to find correlations and separate signal from background

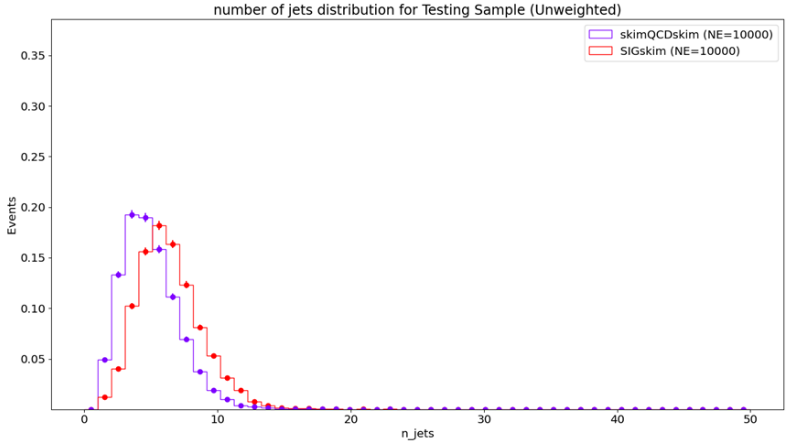
BDT uses high-level data from multiple jets and is tested with multiple sets of variables to see if the PFN using low-level data can make correlations between signal and background that the BDT using high-level data can't

BDT is trained and tested over files that contain SVJs of multiple different r_{inv} and Z' mass values

Variables

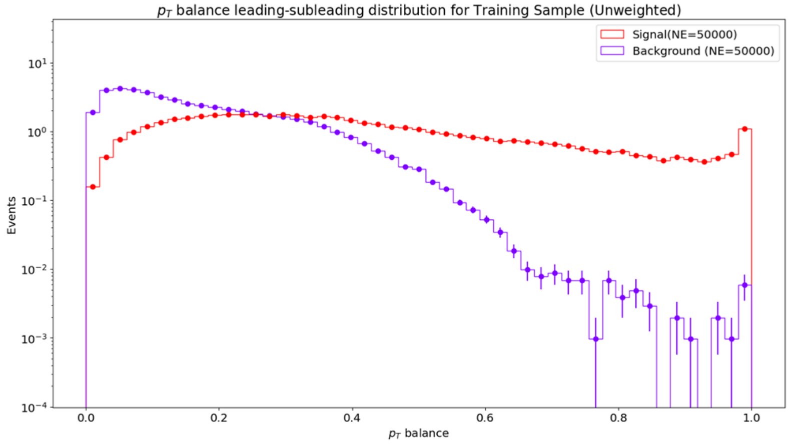
n_{jets}

- Number of jets detected



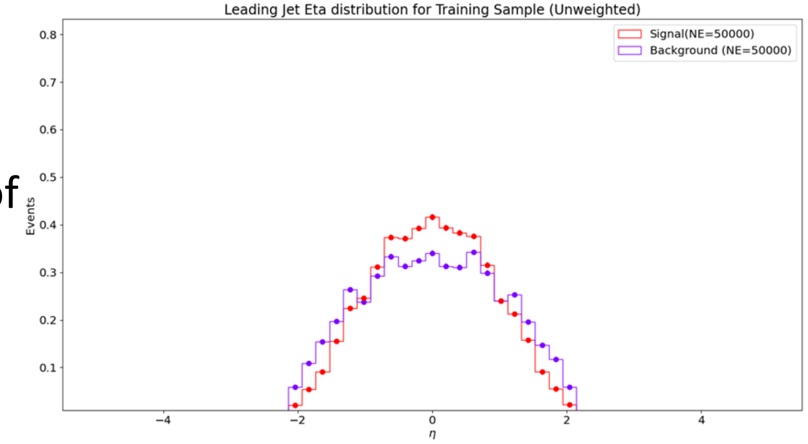
$Pt\_balance_{12}$

- Difference in momentum between the jets divided by the momentum of the leading jet



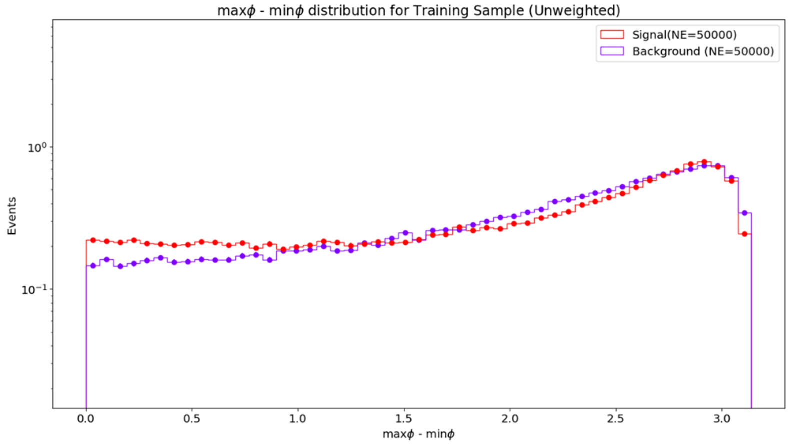
$Jet1_{\eta}$

- The pseudorapidity of the leading jet



Maxϕ_minϕ

- Difference in transverse angles between the jets



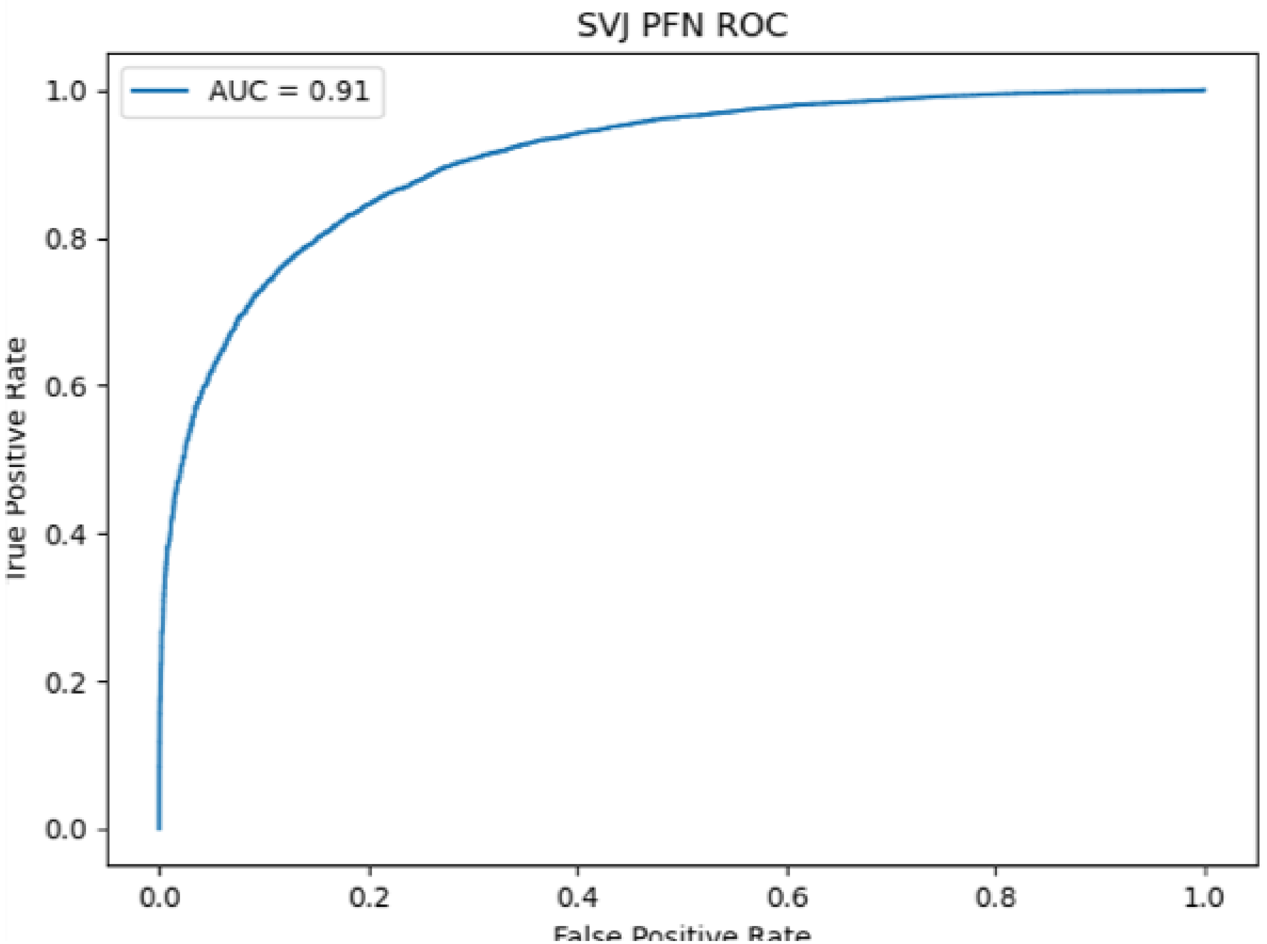
Variable Tables

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	The total reconstructed mass
rT	MET/mT
$\Delta\phi_{min}$	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

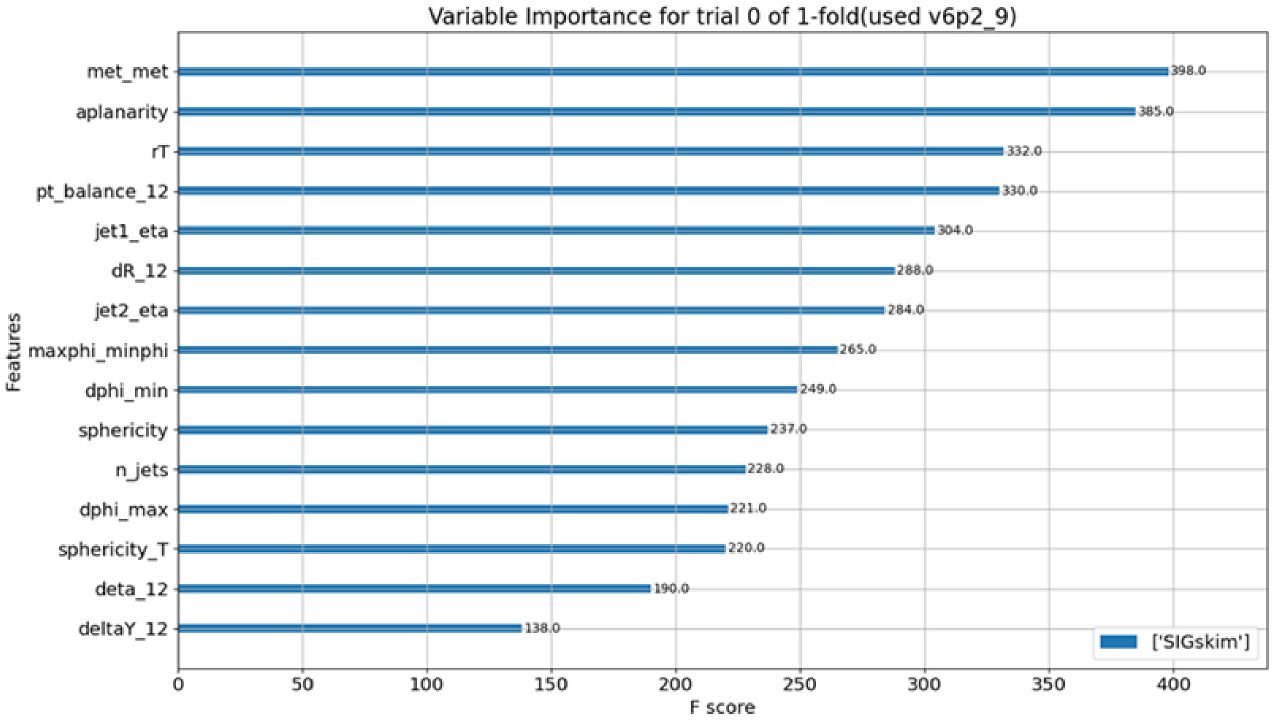
Jet1/2_width	Jet width calculated using calorimeter data
Jet1/2_TrackWidthPt1000PV	The width between the two furthest tracks with pt over 1000 MeV within the jet from the primary vertex
Jet1/2_SumPtTrkPt500PV	The pt sum of each track of at least 500 MeV within the jet from the primary vertex
Jet1/2_NumTrkPt1000PV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

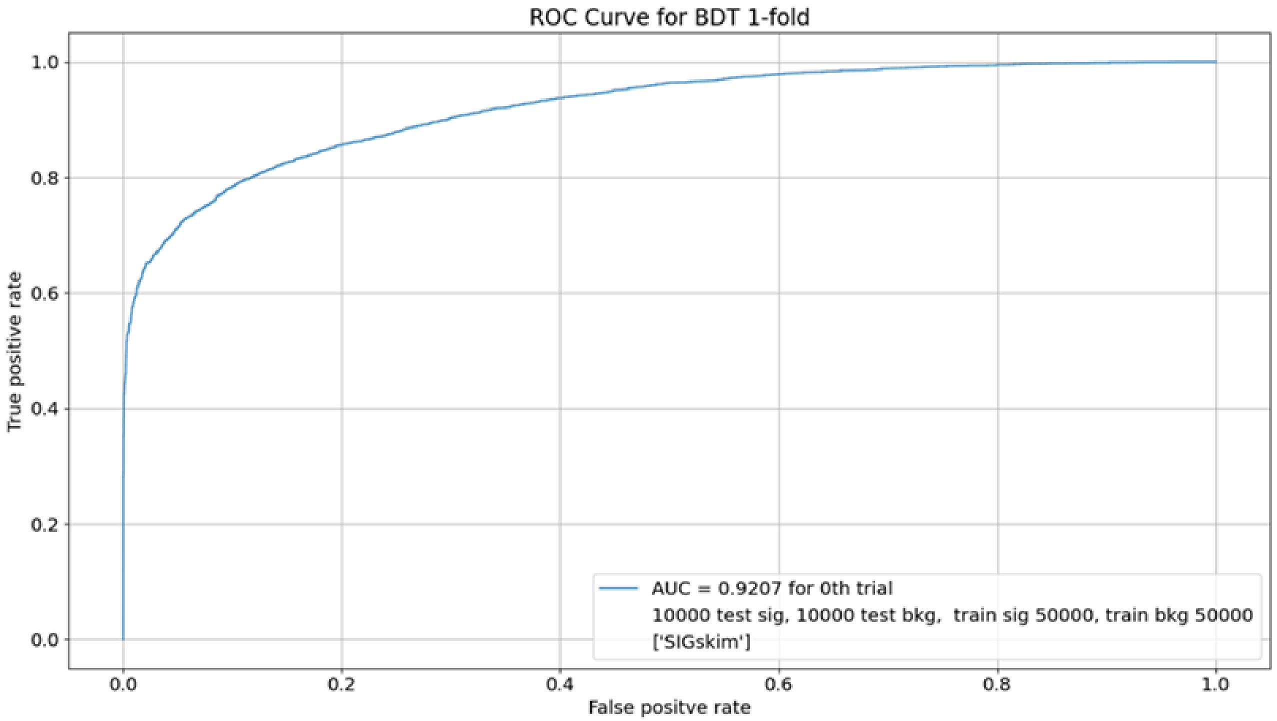
PFN ROC

Receiving operating characteristic (ROC) curves show the false positive vs true positive rate for a machine learning (ML) model. The Area Under the Curve (AUC) shows the percentage chance that the ML model successfully identifies signal from background. The PFN AUC score of 0.91 serves as a benchmark to measure the BDT by.



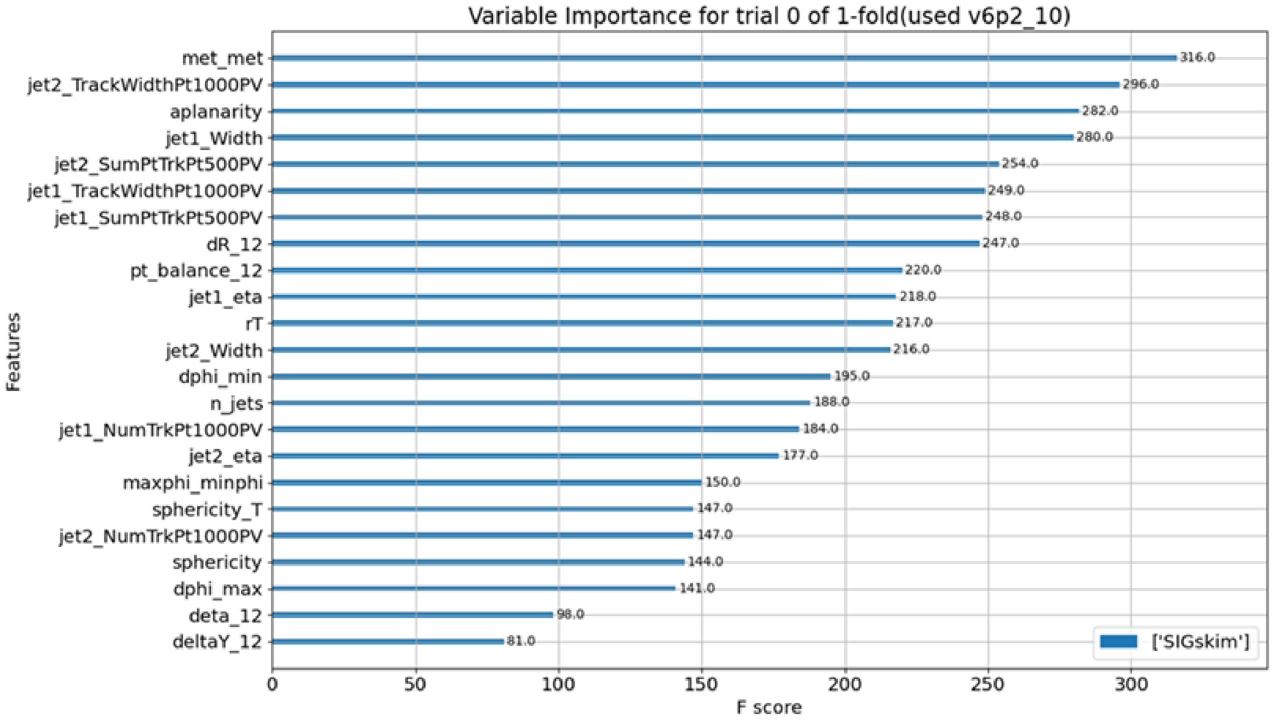
BDT Without Track Variables

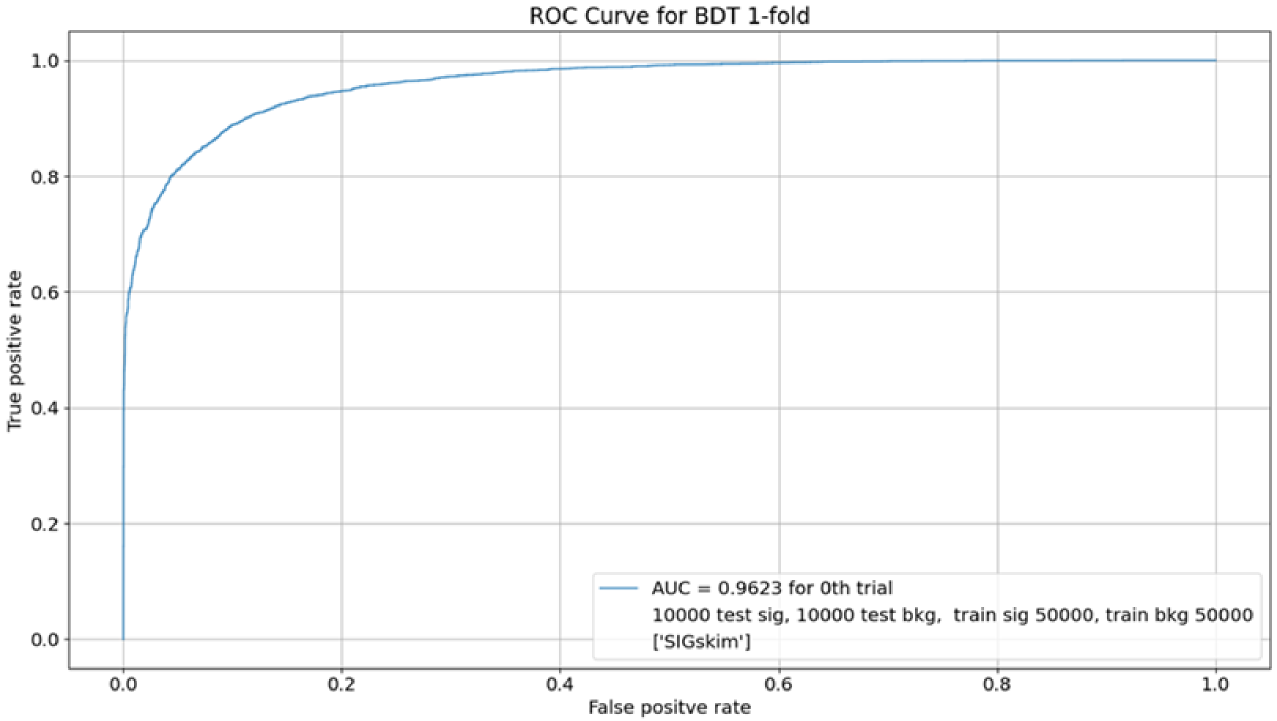




Variable importance charts rank variables by how well they discriminate signal from background.

BDT With All Variables

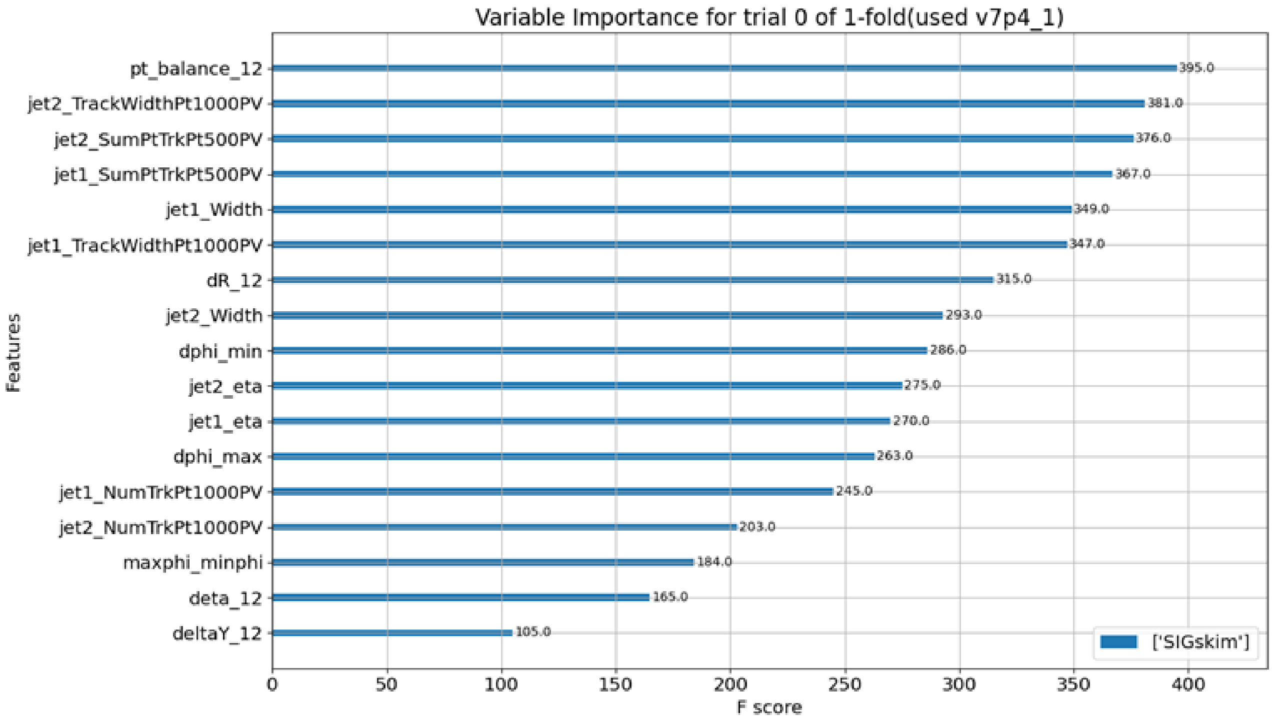


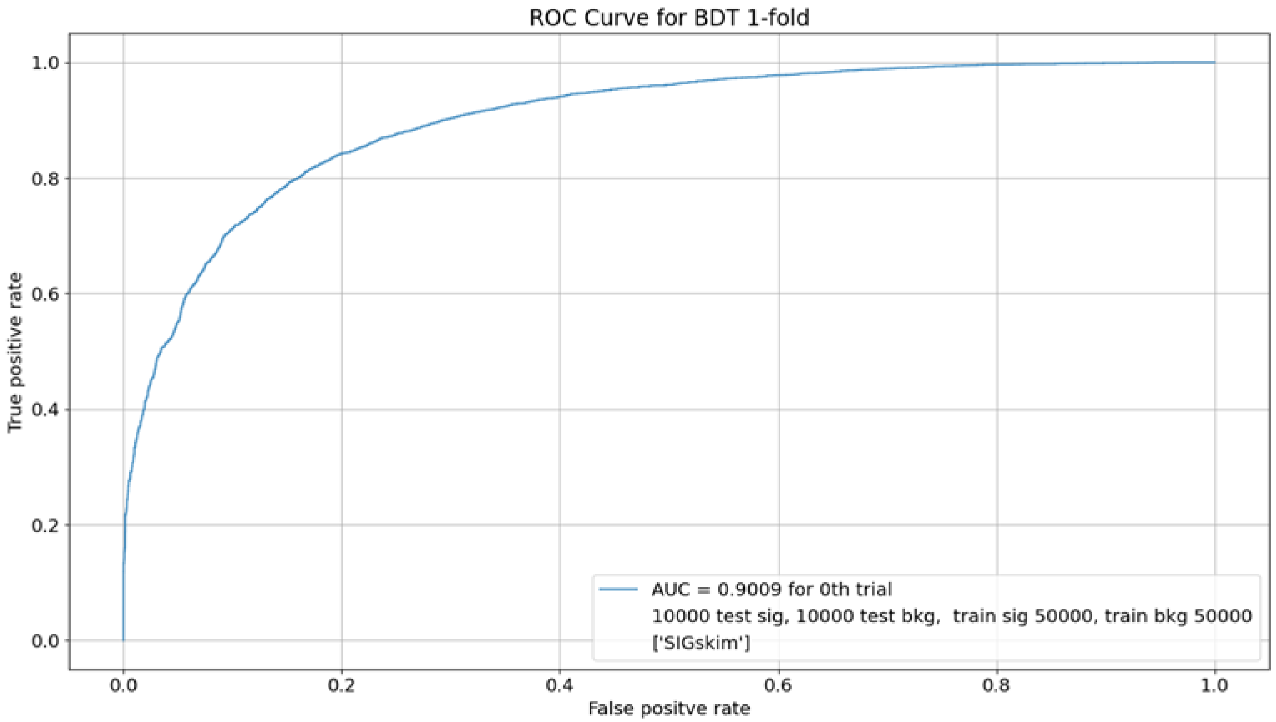


BDT performs great when tested over signal files with SVJs of many r_{inv} and Z’ mass

But, since r_{inv} and Z’ mass are unknown for SVJs, MET could be biased against SVJs

BDT Without MET

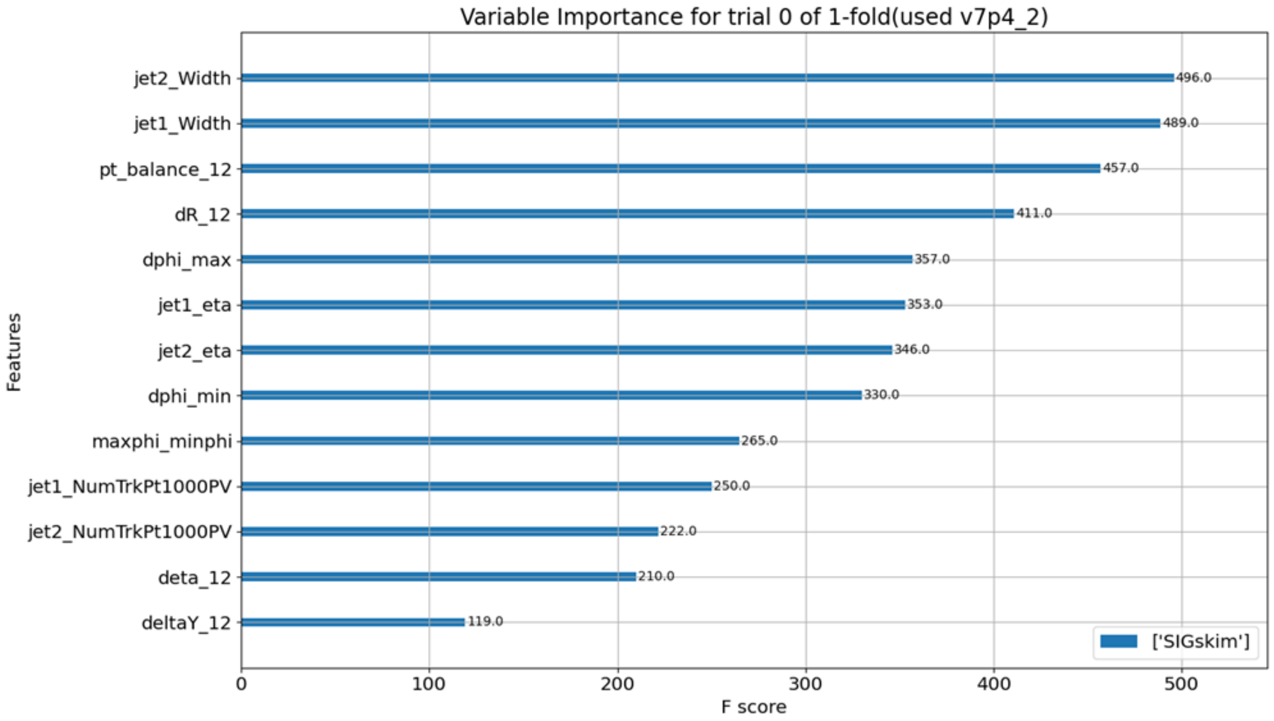


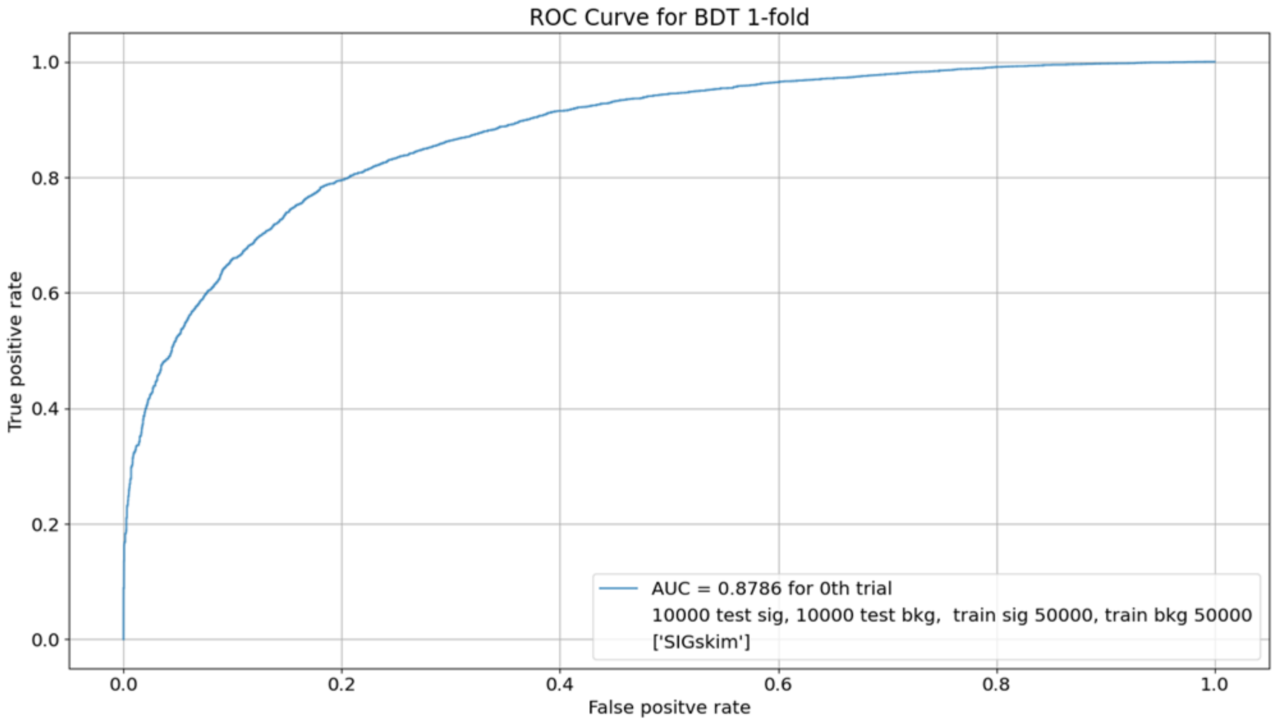


Removed MET and rT

Also removed variables involving more than 2 jets (n_{jets}, aplanarity, sphericity, $sphericity_T$) since PFN can only analyze 2 jets

BDT With the Least Amount of Variables





No n_{jets}, rT, MET aplanarity, sphericity, sphericity$_T$, jet1/2_SumPtTrkPt500PV, or jet1/2_TrackWidthPt1000PV

Performs significantly worst compared to the PFN

Results

BDT performs better than the PFN when given more high-level variables but performs worse than the PFN when missing multiple discriminatory high-level variables

MET was the strongest variable out of all

Since we don't know what r_{inv} or Z' mass is, using MET could cause our BDT to reject SVJs if they have a MET that is similar to background

Since the BDT only produces an AUC of 0.88 when compared to the PFN's 0.91 in a one-to-one test, the PFN has been shown to be superior to the BDT.

Since the PFN has access to low-level data, the PFN can reconstruct variables that don't make physical sense yet make good discriminators.

Signal and Control Region Study

Since the PFN is superior to the BDT, we move forwards with the PFN

How do we further test the efficiency of the PFN?

Find a variable which splits the data into a control region with little signal and a signal region that contains most signal, while not discriminating between background between the two groups

Need a variable highly discriminatory at detecting signal but not discriminatory at detecting background

After the PFN creates a signal region using experimental data, we unblind the experimental signal region and see how it compares to the test signal region

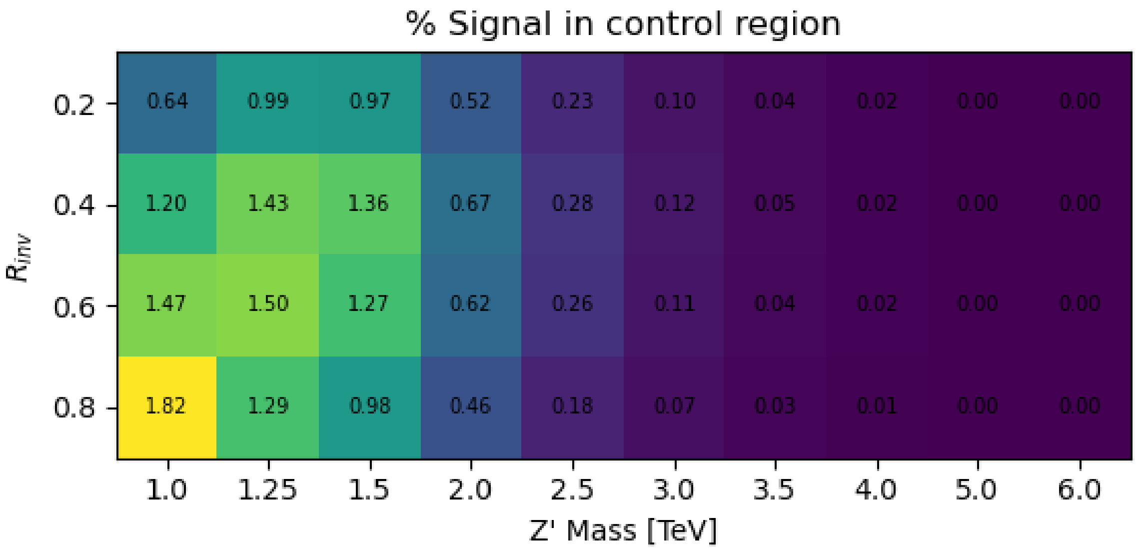
$\Delta\eta_{12}$ cut

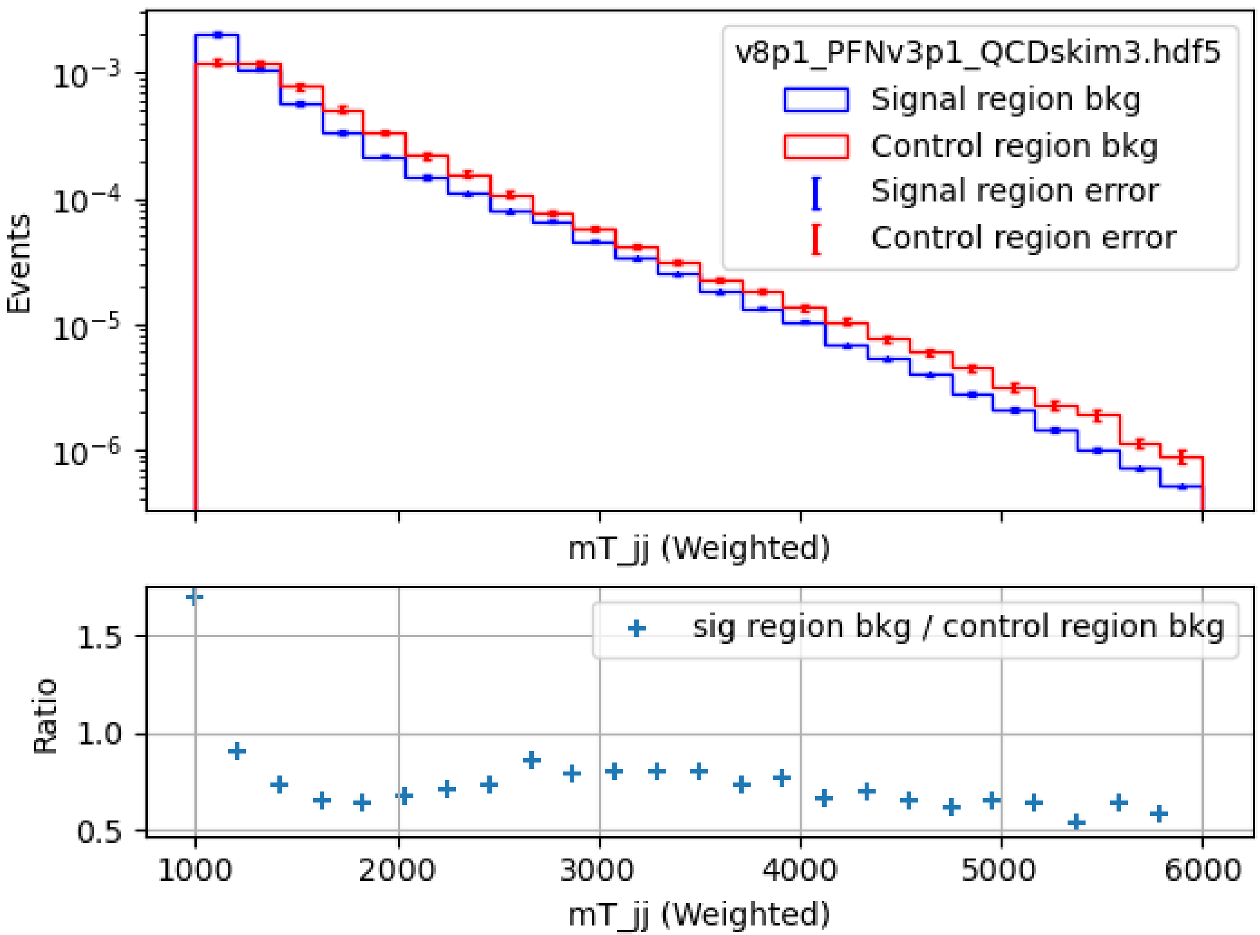
Control region: $\Delta\eta_{12} > 1$

Signal region $\Delta\eta_{12} < 1$

PFN score cut > 0.92

Pseudorapidity (η) is an angular coordinate describing a particles angle relative to the beam.



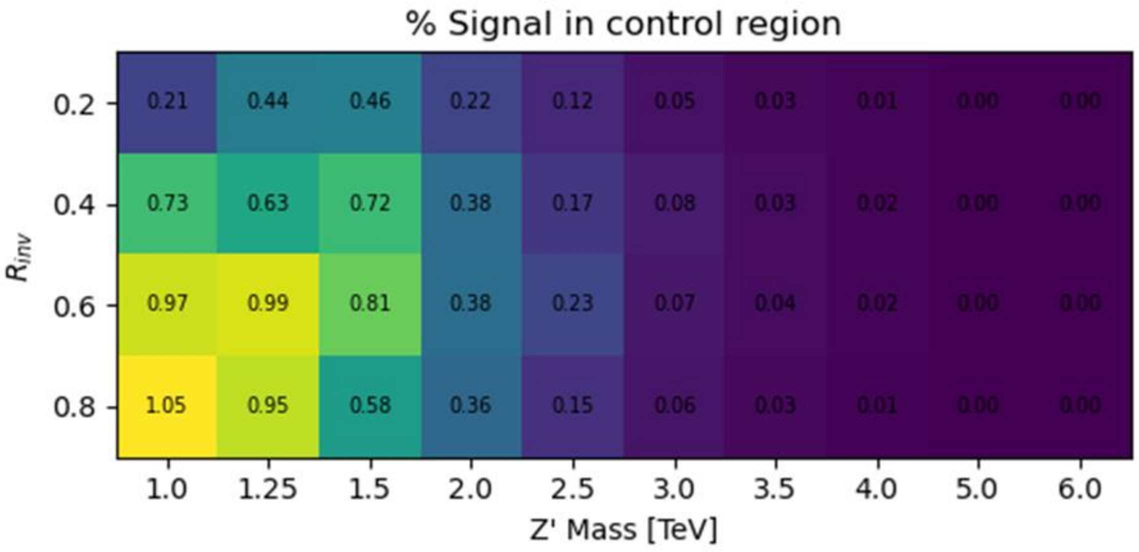


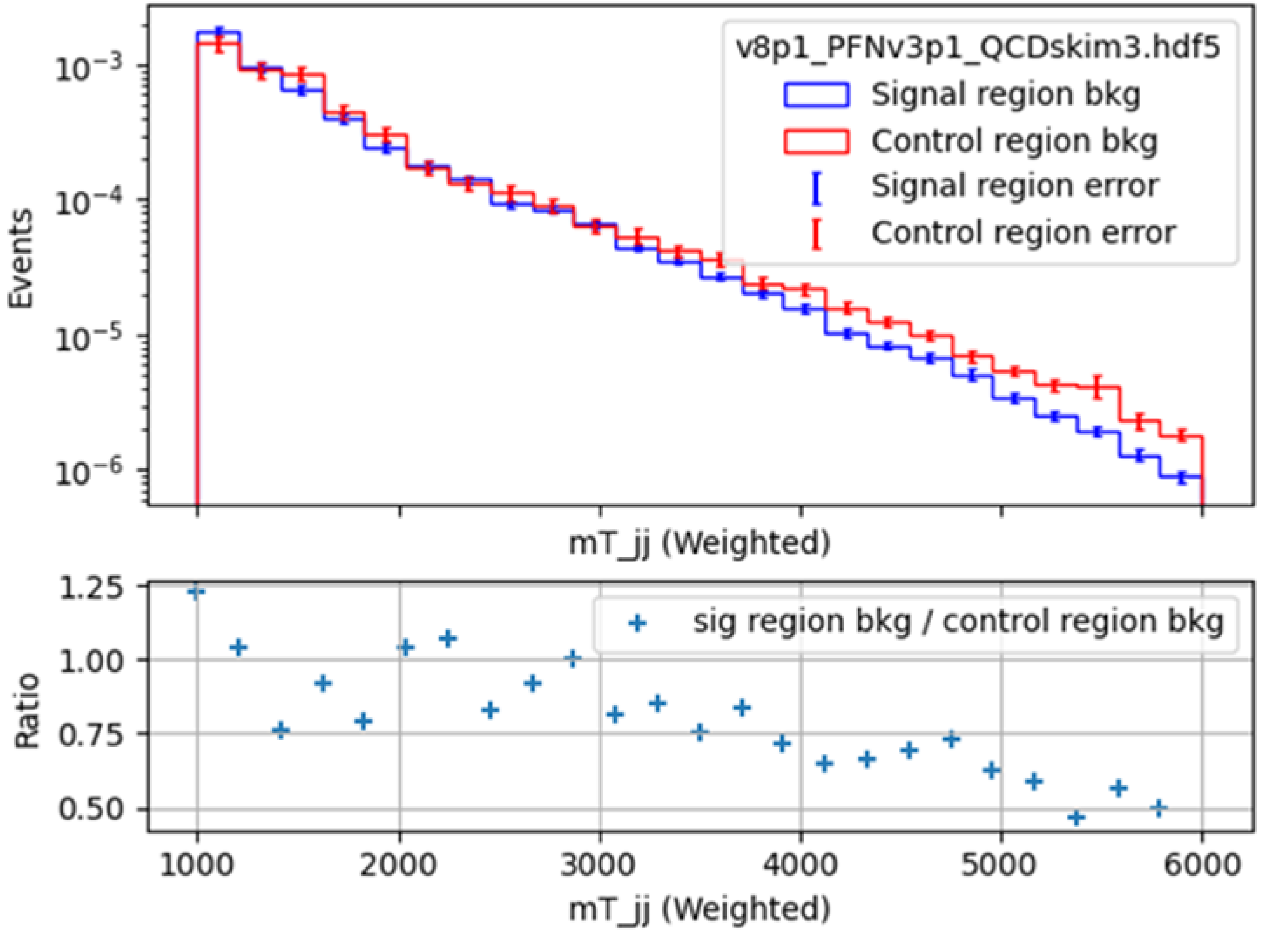
$\text{Jet2}_{\text{width}}$ cut 0.05

Signal region: $\text{jet2}_{\text{width}} > 0.05$

Control region: $\text{jet2}_{\text{width}} < 0.05$

PFN cut > 0.97



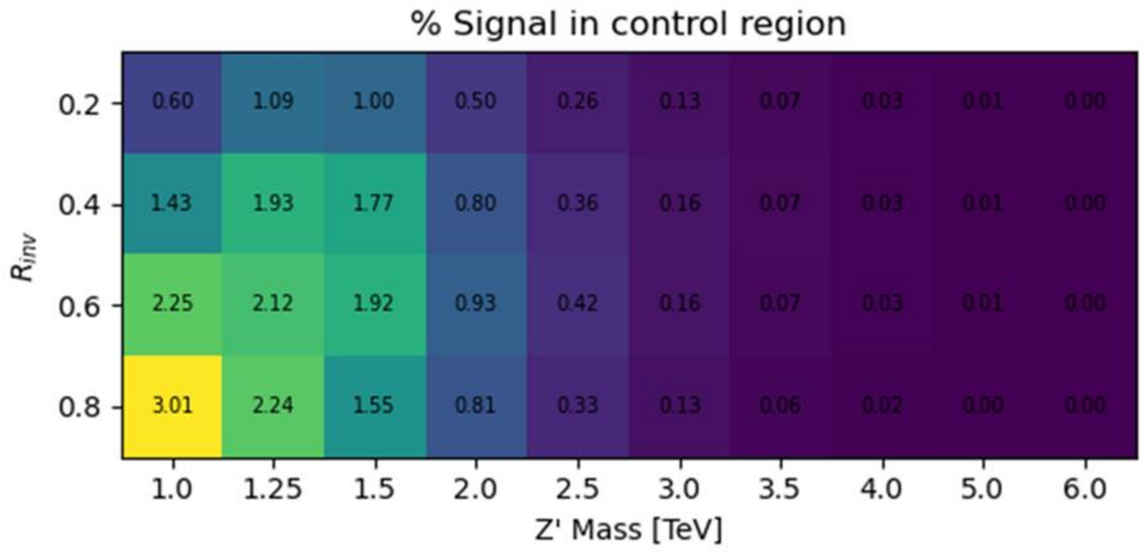


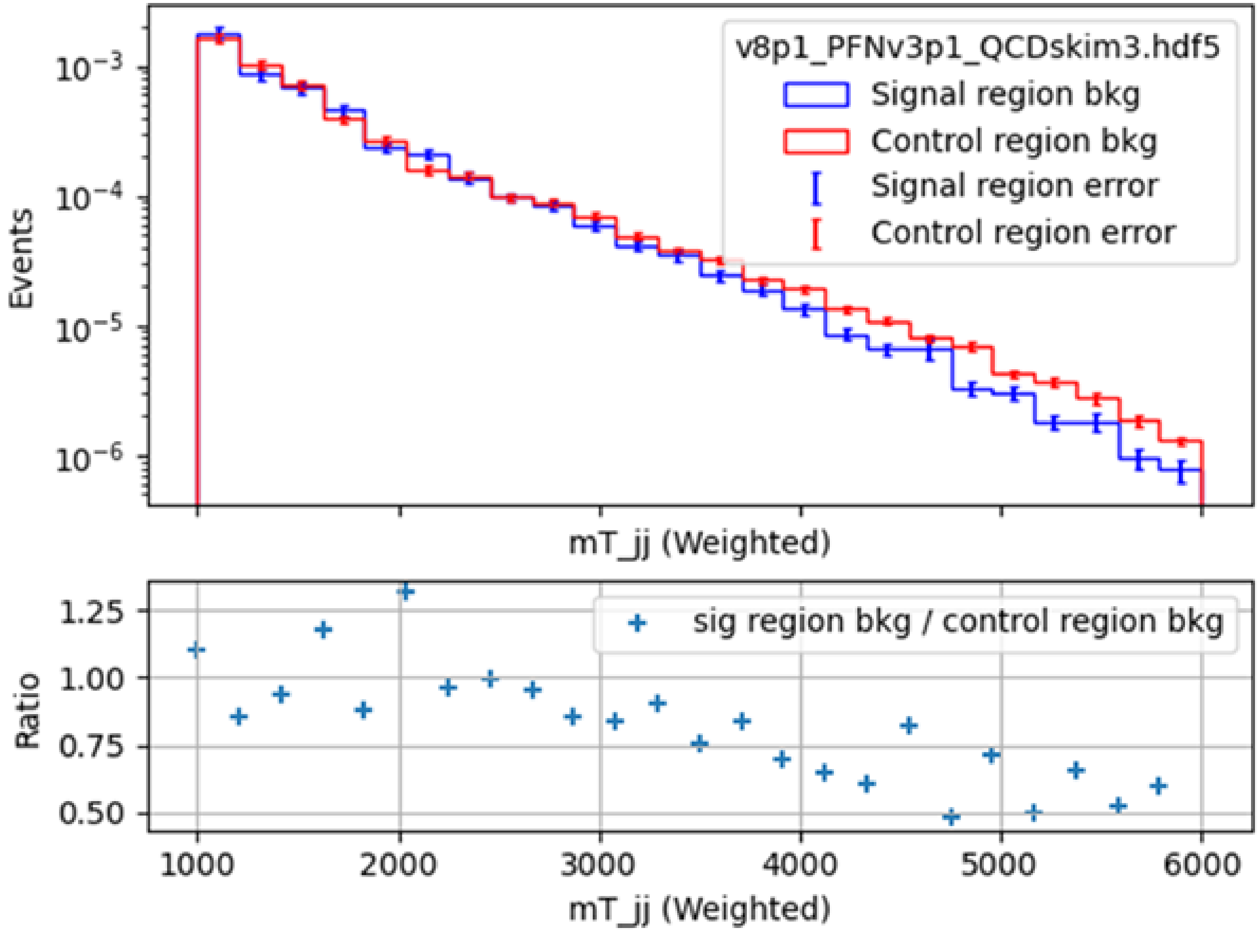
$Jet2_{width}$ cut 0.1

Signal region: $jet2_{width} > 0.1$

Control region: $jet2_{width} < 0.1$

PFN cut > 0.97



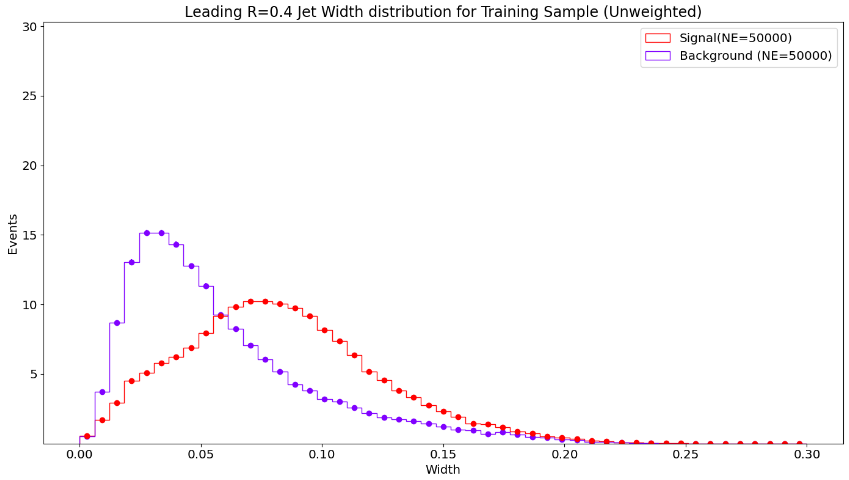


Results

The Jet2$_{width}$ cuts provides better separation than the $\Delta\eta_{12}$ cut

Both Jet2$_{width}$ cuts separate about the same despite using different values

More studies on signal/control regions need to be made



Conclusion/Future

The PFN performs better than the BDT since it's able to find correlations in the track-level data that the BDT can't

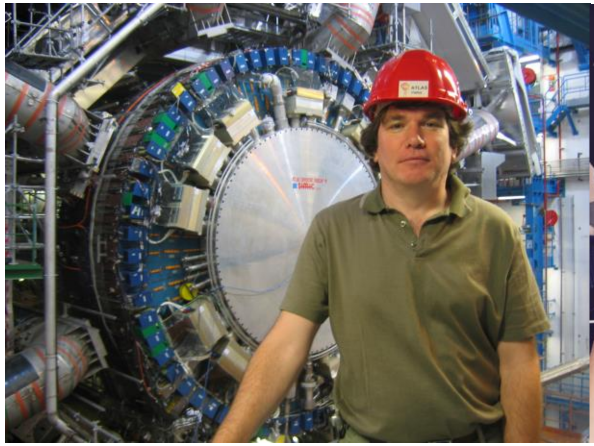
$Jet2_{width}$ is currently the best signal region split found, but more research needs to be done

We hope to unblind the signal region in fall

Next spring, we hope to publicly report results of the experiment

Acknowledgements

SVJ group members:

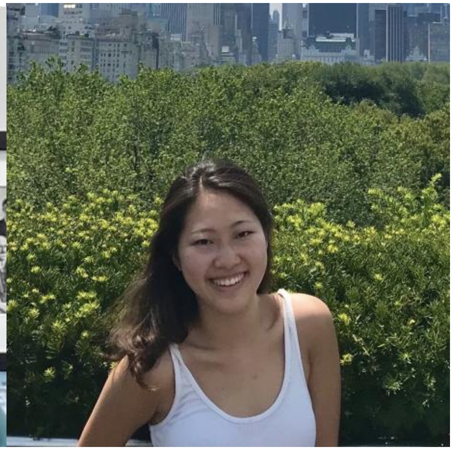


Prof. John Parsons

Dr. Julia Gonski

Elena Busch

Gabriel Matos



Kiryeong Park

I would like to thank:

- My SVJ group members who helped mentor, teach, and support me throughout my research
- Prof. John Parsons, Prof. Georgia Karagiorgi, and Amy Garwood for organizing the REU
- 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

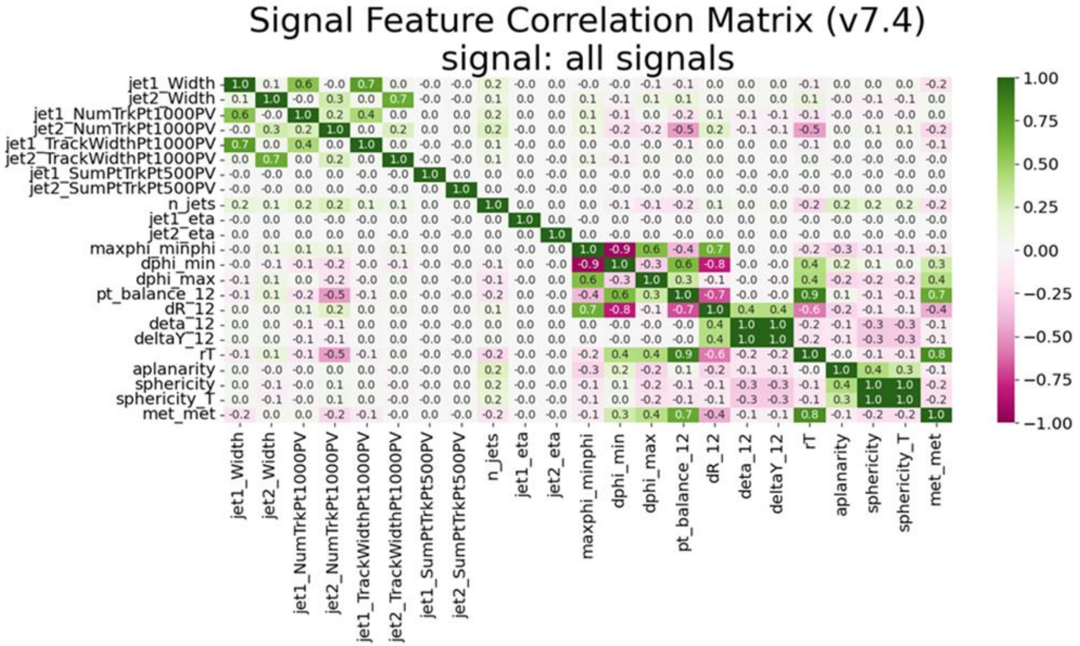
References

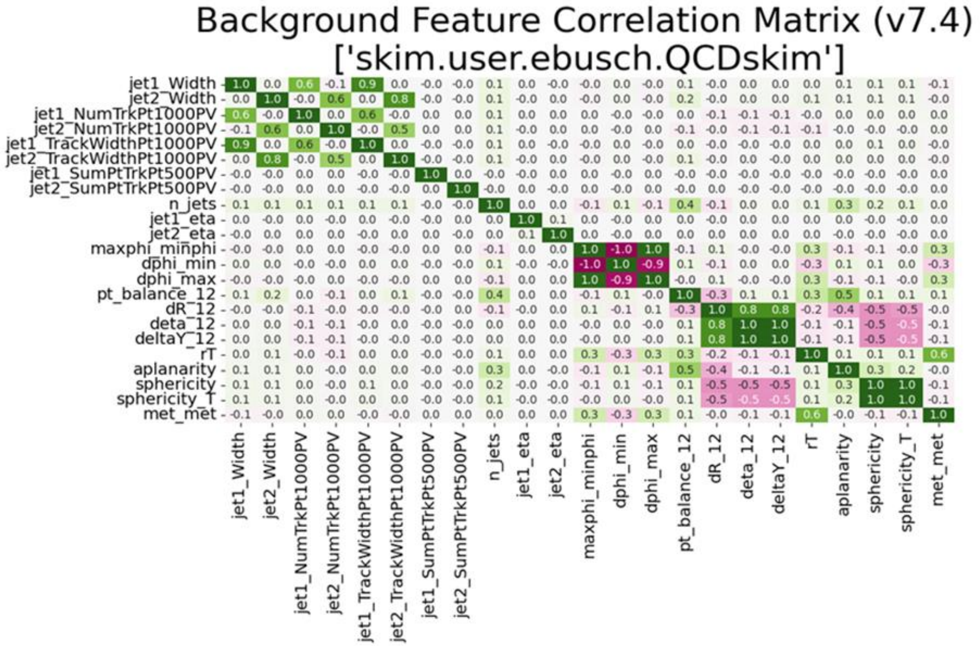
- CERN. Dark matter. https://home.cern/science/physics/dark-matter.
- CERN. The large hadron collider. https://home.cern/science/accelerators/large-hadron-collider.
- CERN. Lhc the guide faq. https://home.cern/resources/brochure/knowledge-sharing/lhc-facts-and-figures.
- CERN. Run 3 of the large hadron collider. https://home.cern/press/2022/run-3.
- CERN. The standard model. https://home.cern/science/physics/standard-model.
- Yann Coadou. Boosted decision trees. In Artificial Intelligence for High Energy Physics, pages 9–58. WORLD SCIENTIFIC, feb 2022.
- 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.
- Cesare Cazzaniga & Annapaola de Cosa. Leptons lurking in semi-visible jets at the lhc. Eur. Phys. J., 2022.
- 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.
- ATLAS Experiment. Calorimeter. https://atlas.cern/Discover/Detector/Calorimeter.
- ATLAS Experiment. Glossary. http://opendata.atlas.cern/books/current/get-started/_book/GLOSSARY.html.
- ATLAS Experiment. The inner detector. https://atlas.cern/Discover/Detector/Inner-Detector.
- ATLAS Experiment. Muon spectrometer. https://atlas.cern/Discover/Detector/Muon-Spectrometer.
- JavaTPoint. Decision tree classification algorithm. https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm.
- 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.
- 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.

Backup

Questions?

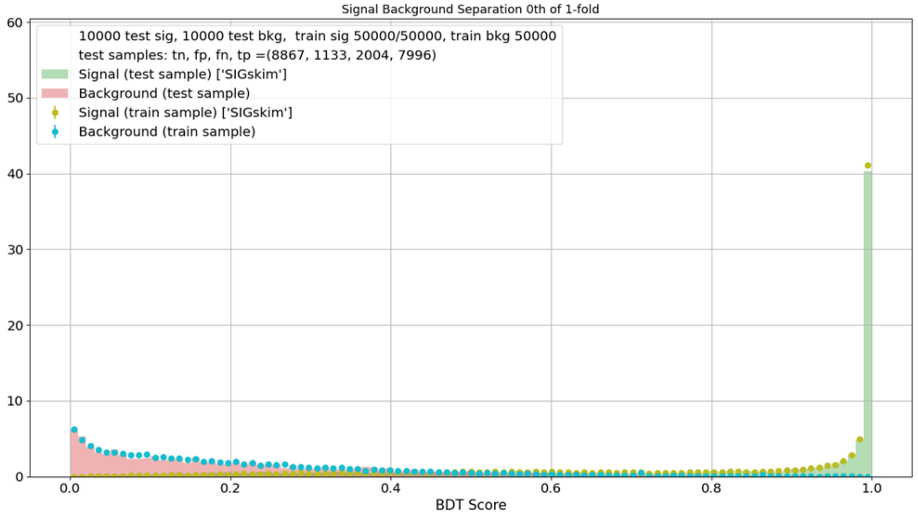
Correlation Matrices



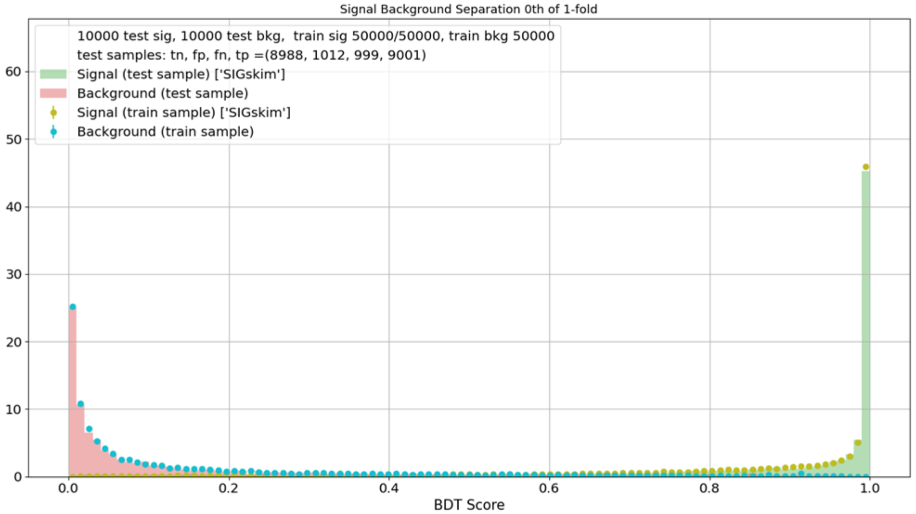


Correlation Matrices show the relation between each variable to each other. A score of 1 means perfect positive correlation, -1 means perfect negative correlation, and 0 means no correlation

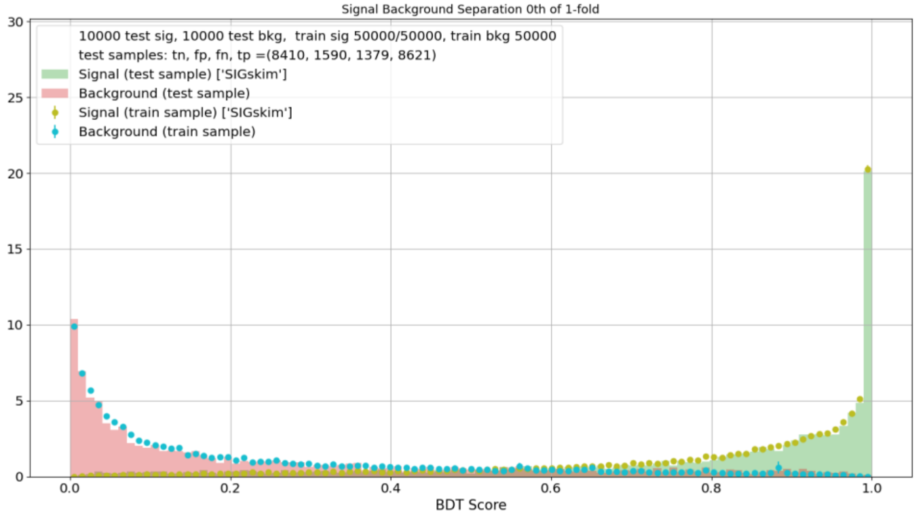
BDT without track variables



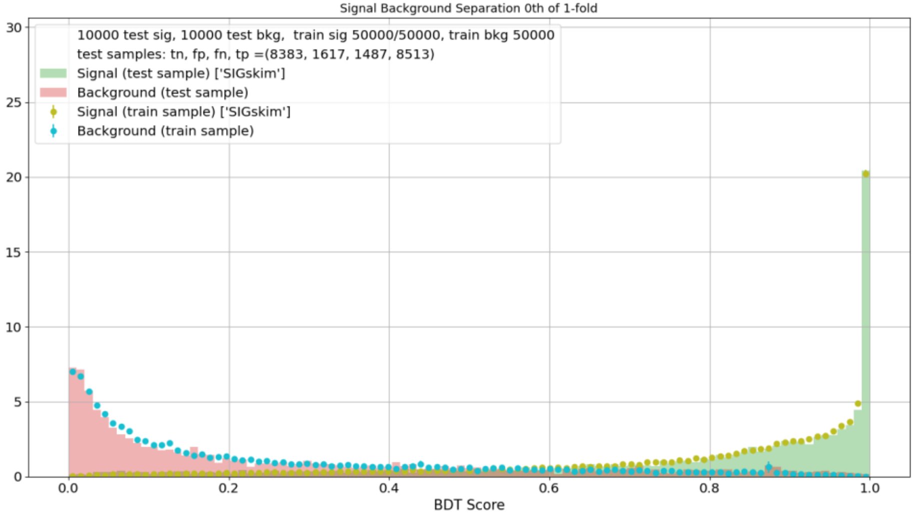
BDT with track variables



BDT with energy distribution variables

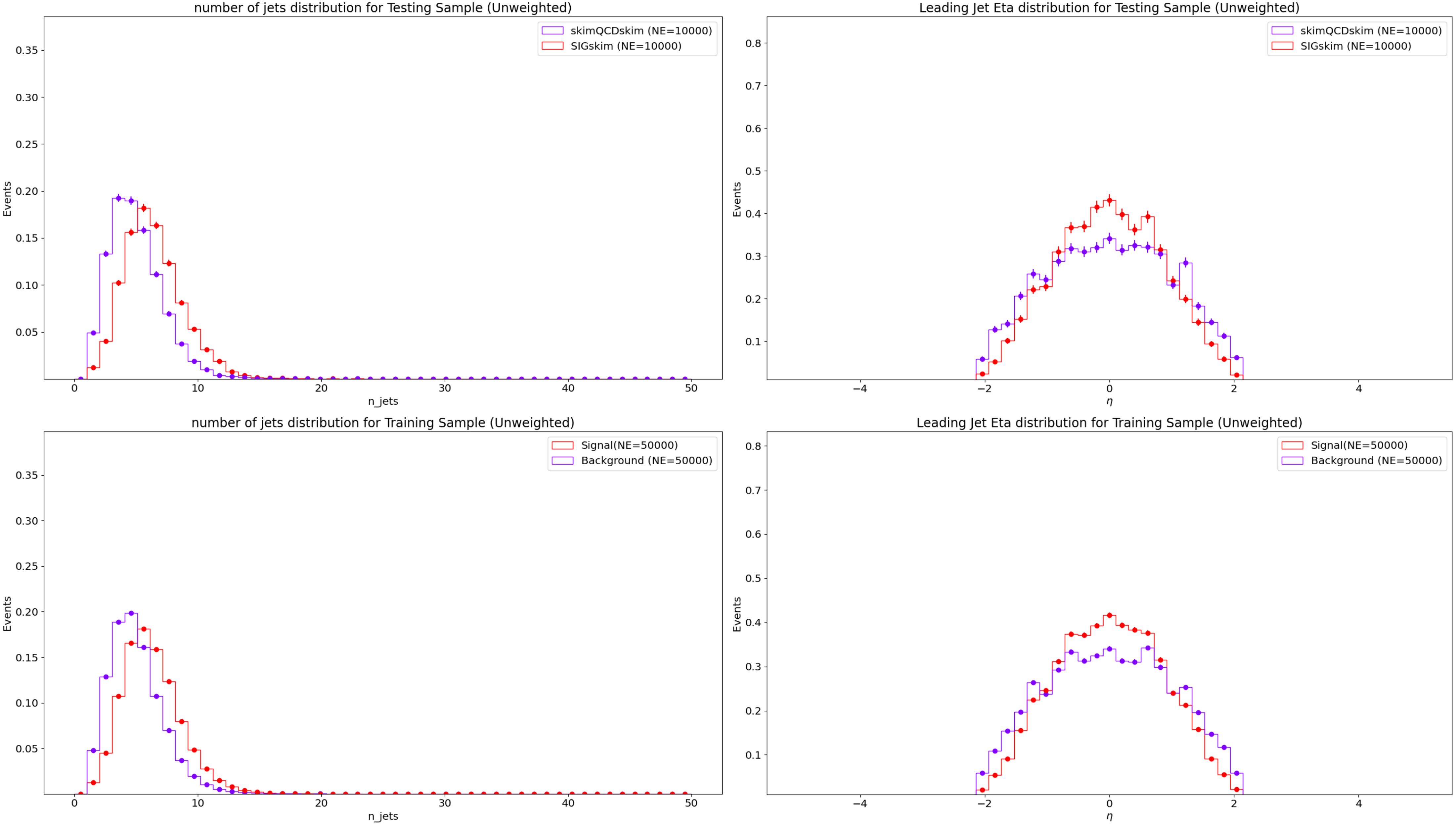


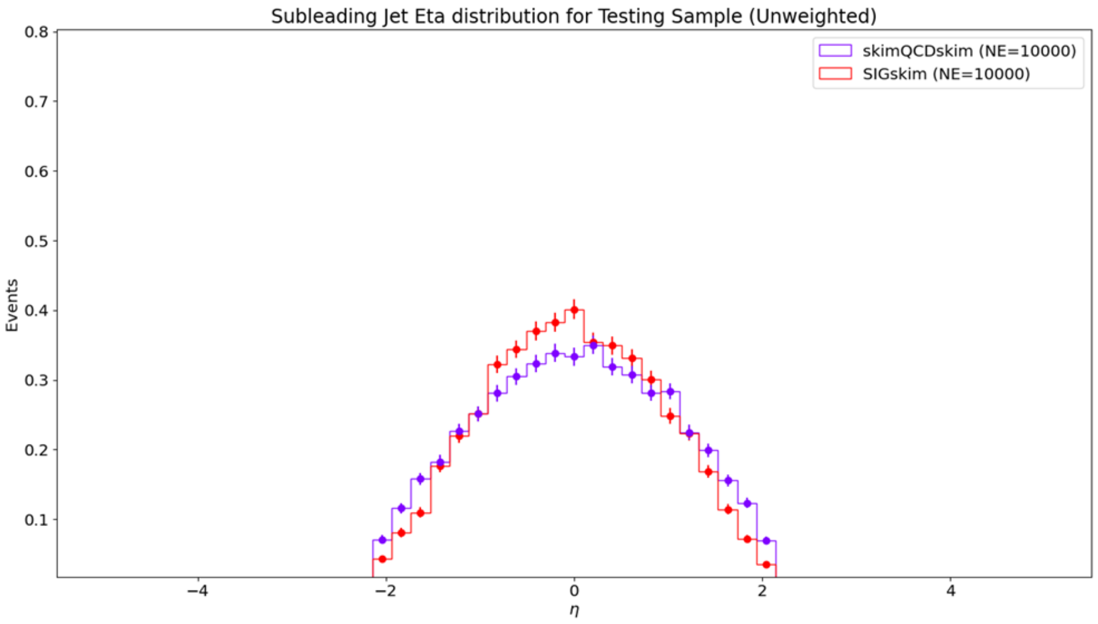
BDT without energy distribution variables

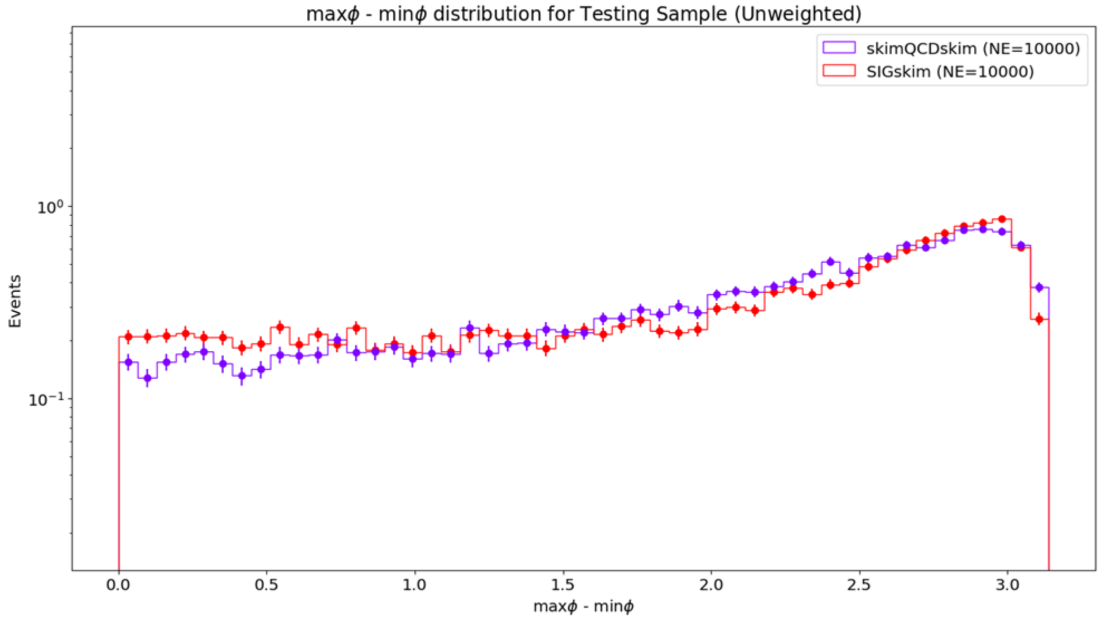


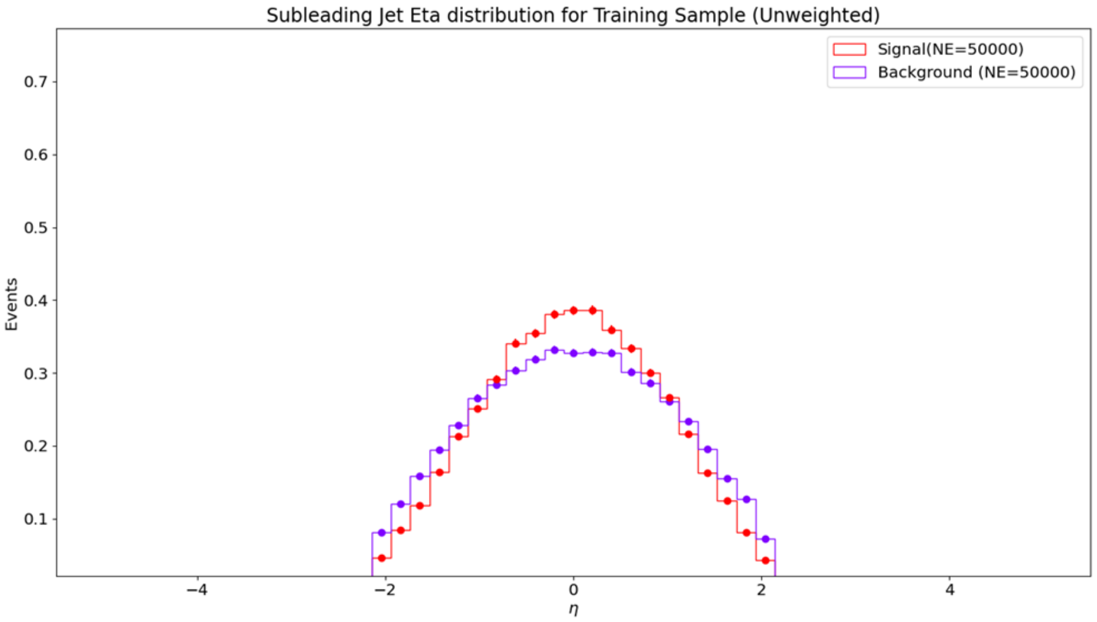
Testing

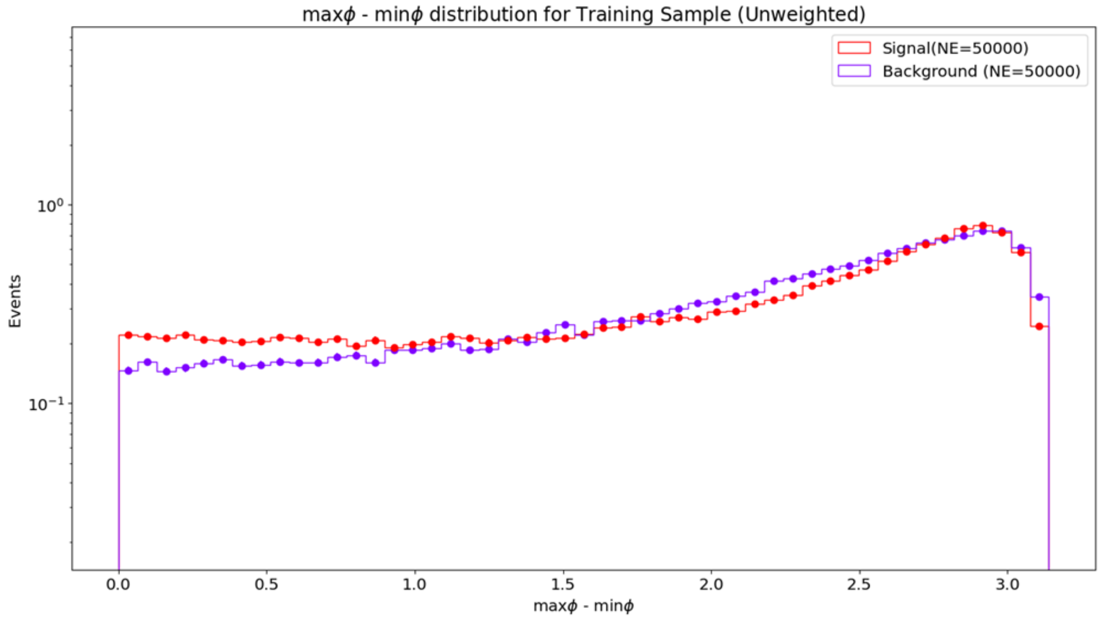
Training

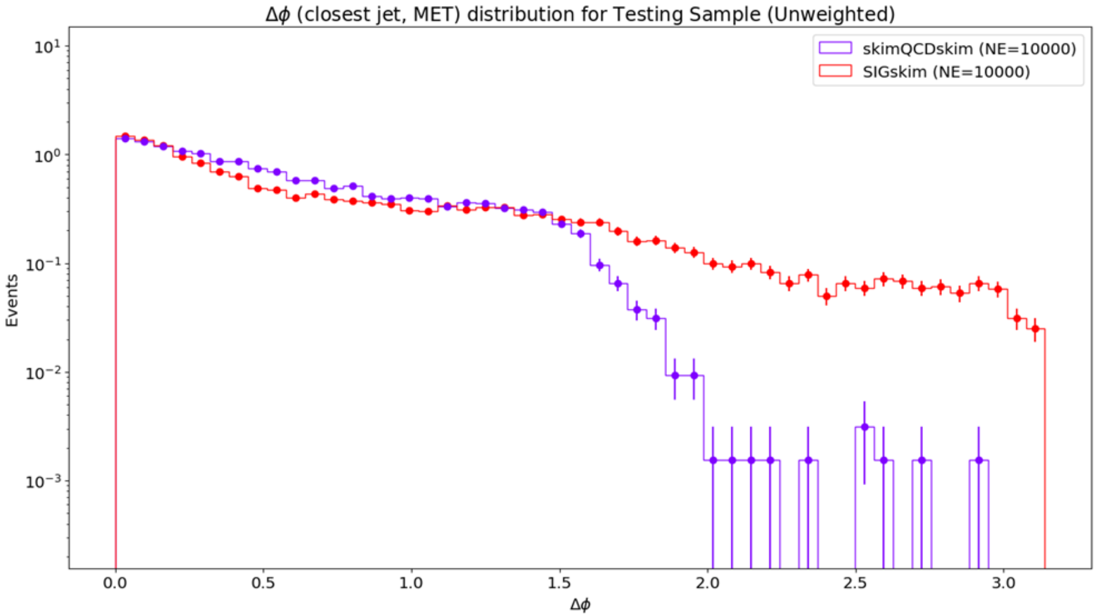


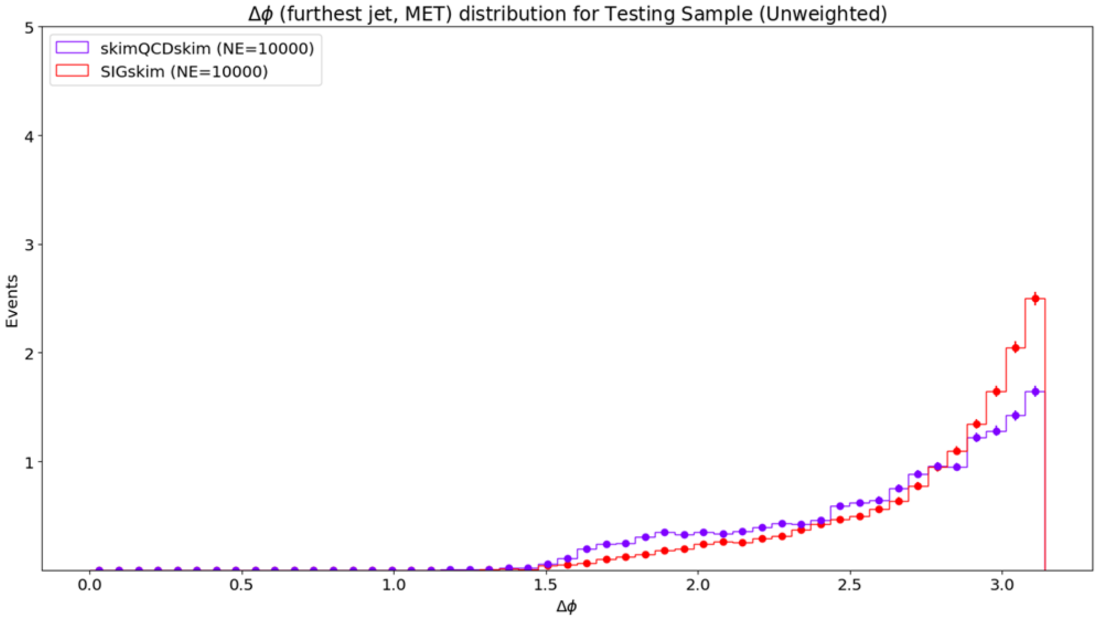

Subleading Jet Eta distribution for Testing Sample (Unweighted)
skimQCDskim (NE=10000)
SIGskim (NE=10000)
Events
η

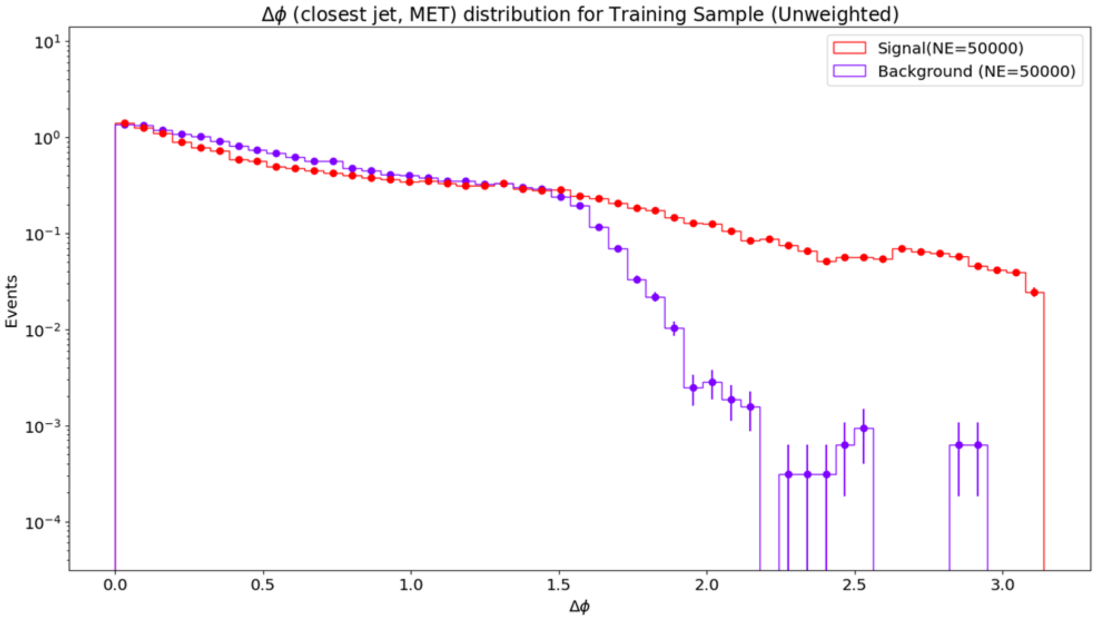

maxϕ - minϕ distribution for Testing Sample (Unweighted)
skimQCDskim (NE=10000)
SIGskim (NE=10000)
Events
maxϕ - minϕ

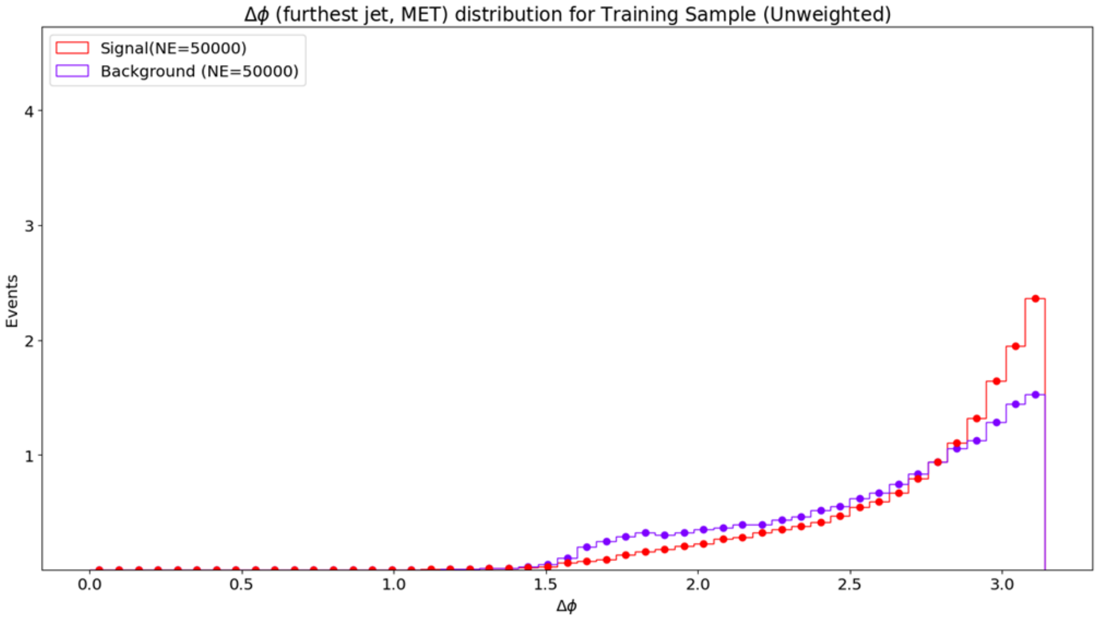

Subleading Jet Eta distribution for Training Sample (Unweighted)
Signal(NE=50000)
Background (NE=50000)
Events
η

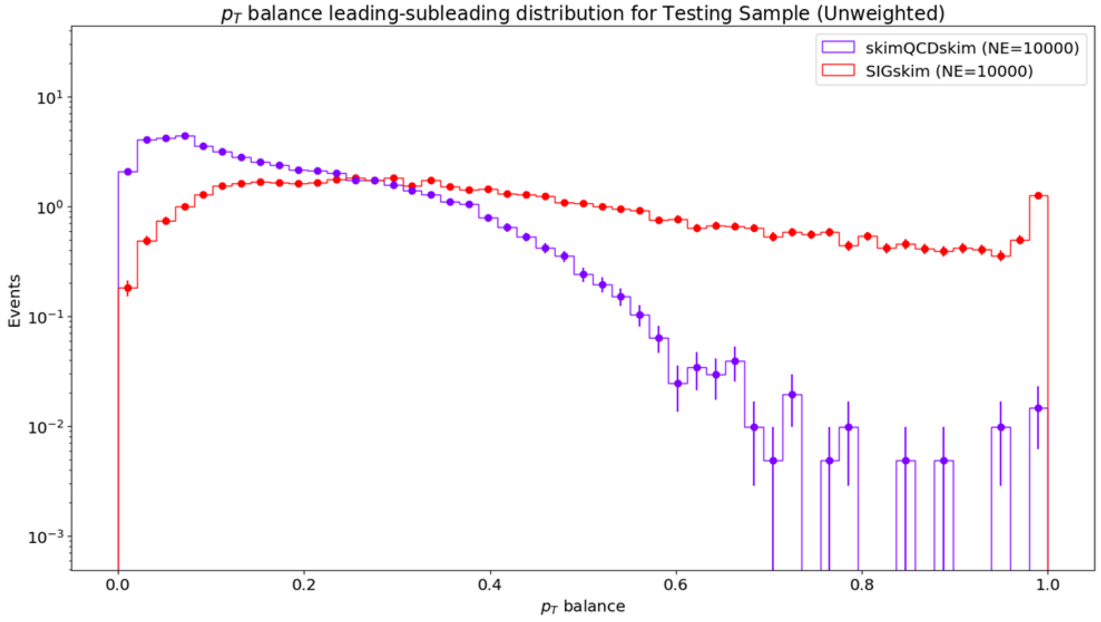

maxϕ - minϕ distribution for Training Sample (Unweighted)
Signal(NE=50000)
Background (NE=50000)
Events
maxϕ - minϕ


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

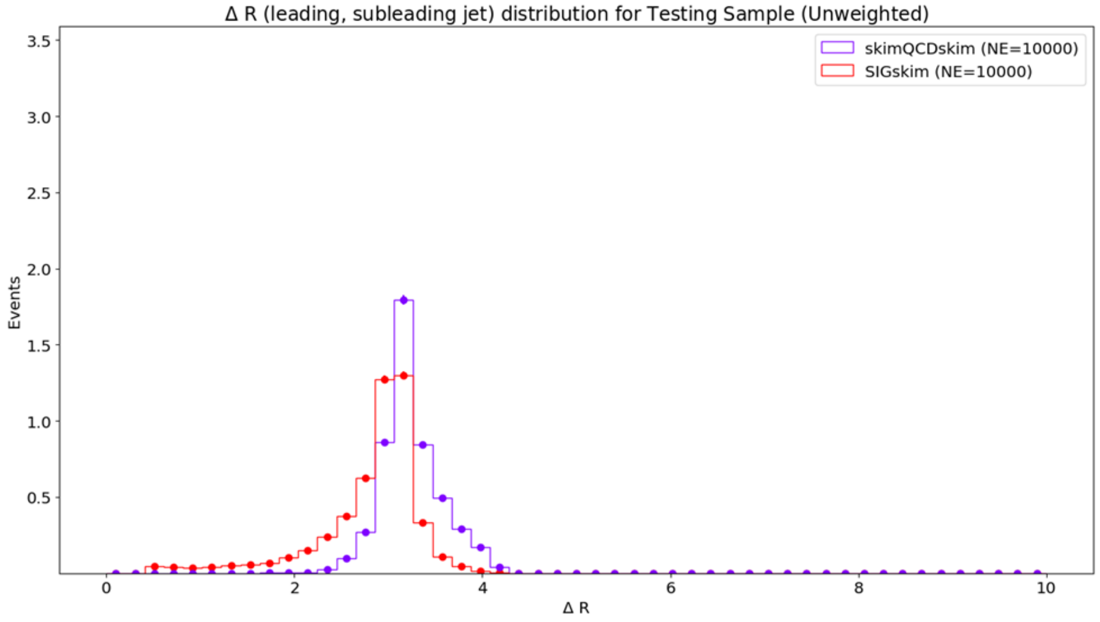

Δϕ (furthest jet, MET) distribution for Testing Sample (Unweighted)
skimQCDskim (NE=10000)
SIGskim (NE=10000)
Events
Δϕ


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

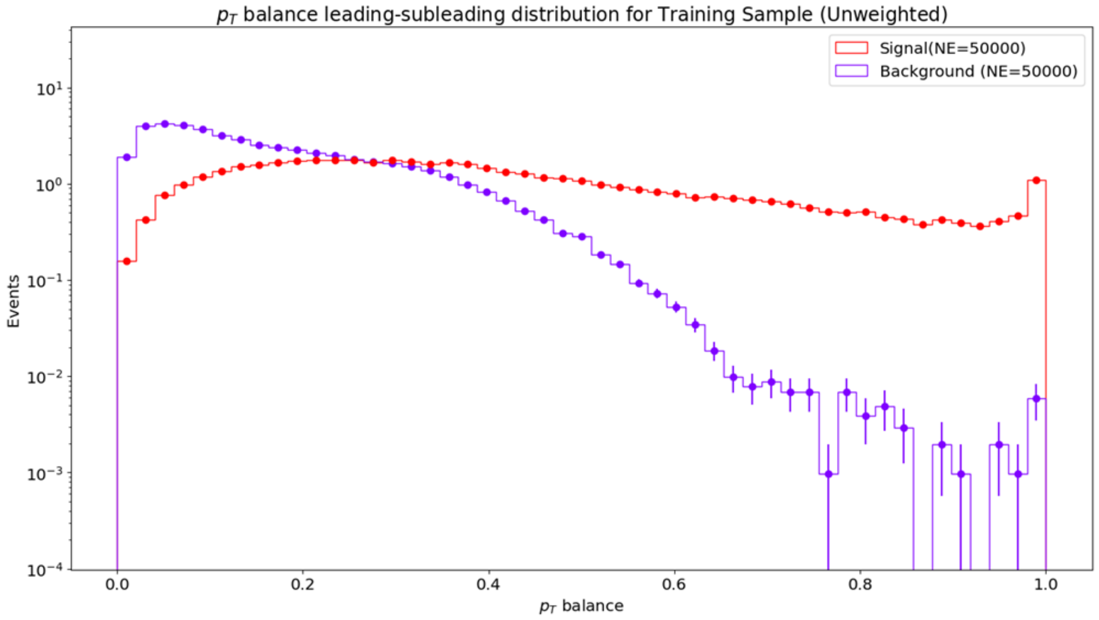

Δϕ (furthest jet, MET) distribution for Training Sample (Unweighted)
Signal(NE=50000)
Background (NE=50000)
Events
Δϕ



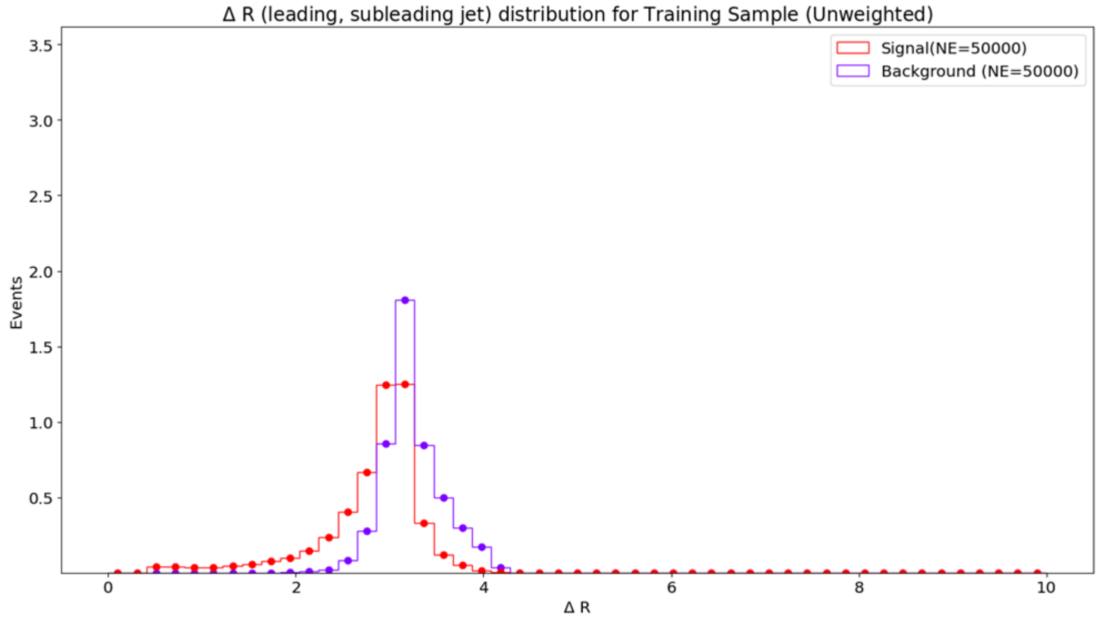
p_T balance leading-subleading distribution for Testing Sample (Unweighted)
skimQCDskim (NE=10000)
SIGskim (NE=10000)
Events
p_T balance



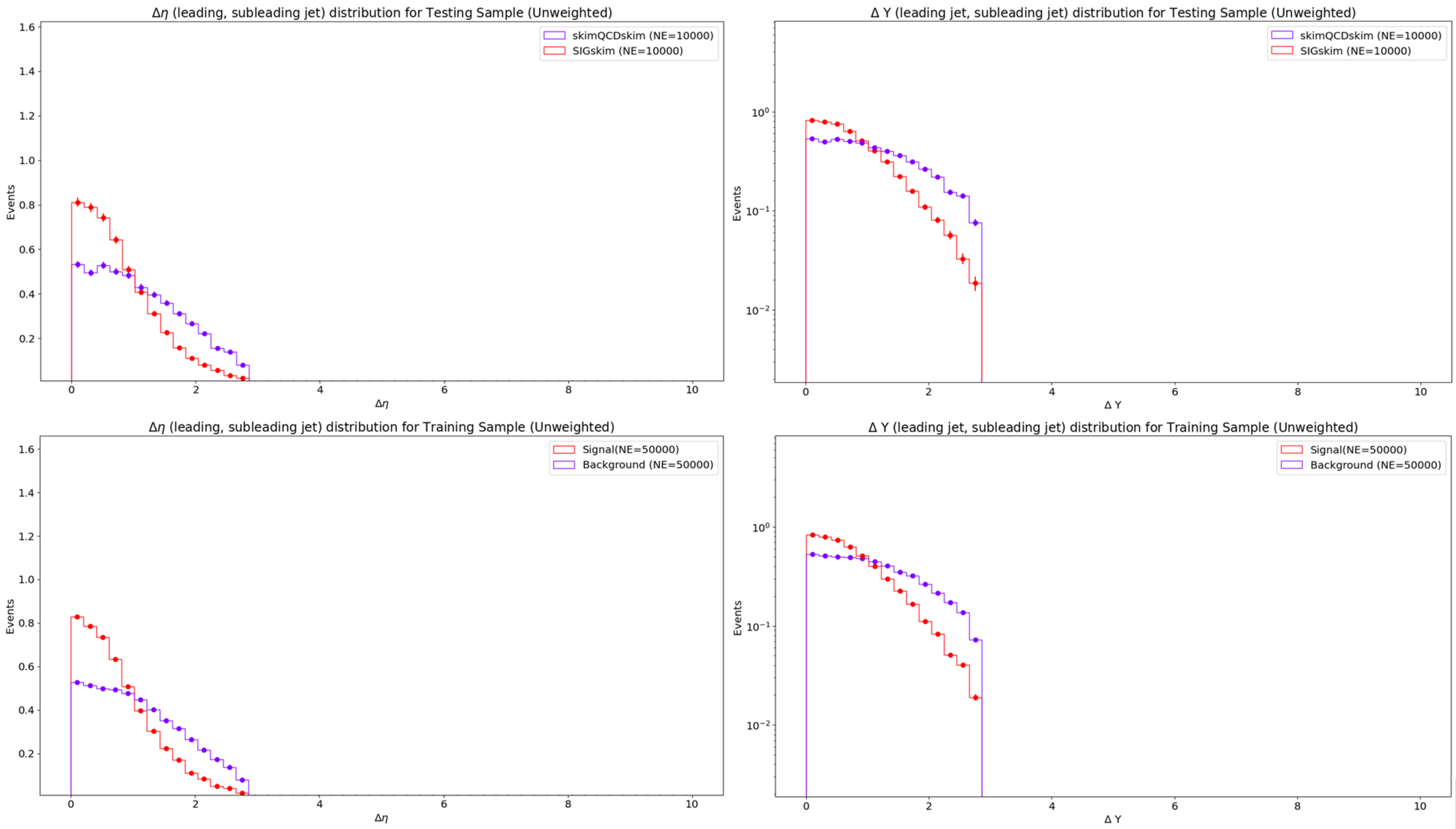
Δ R (leading, subleading jet) distribution for Testing Sample (Unweighted)
skimQCDskim (NE=10000)
SIGskim (NE=10000)
Events
Δ R



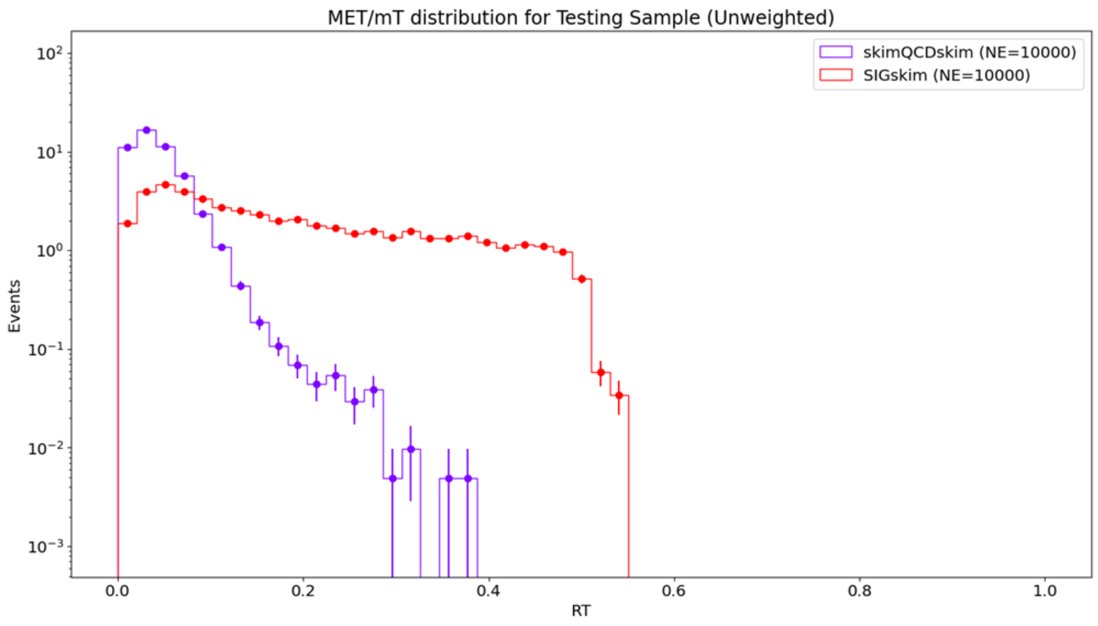
p_T balance leading-subleading distribution for Training Sample (Unweighted)
Signal(NE=50000)
Background (NE=50000)
Events
p_T balance

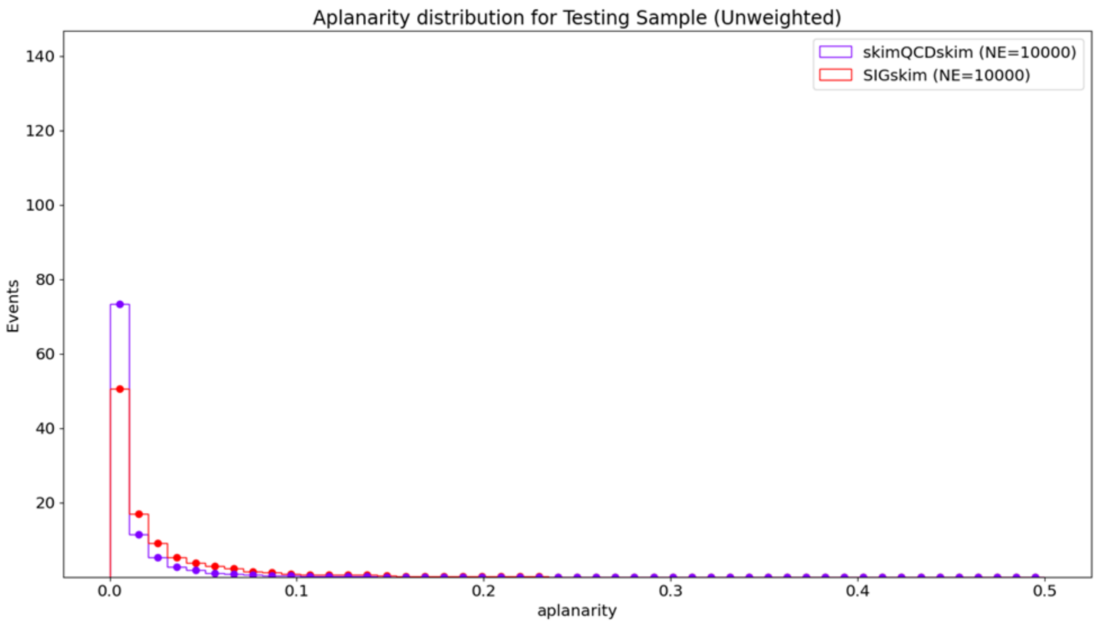


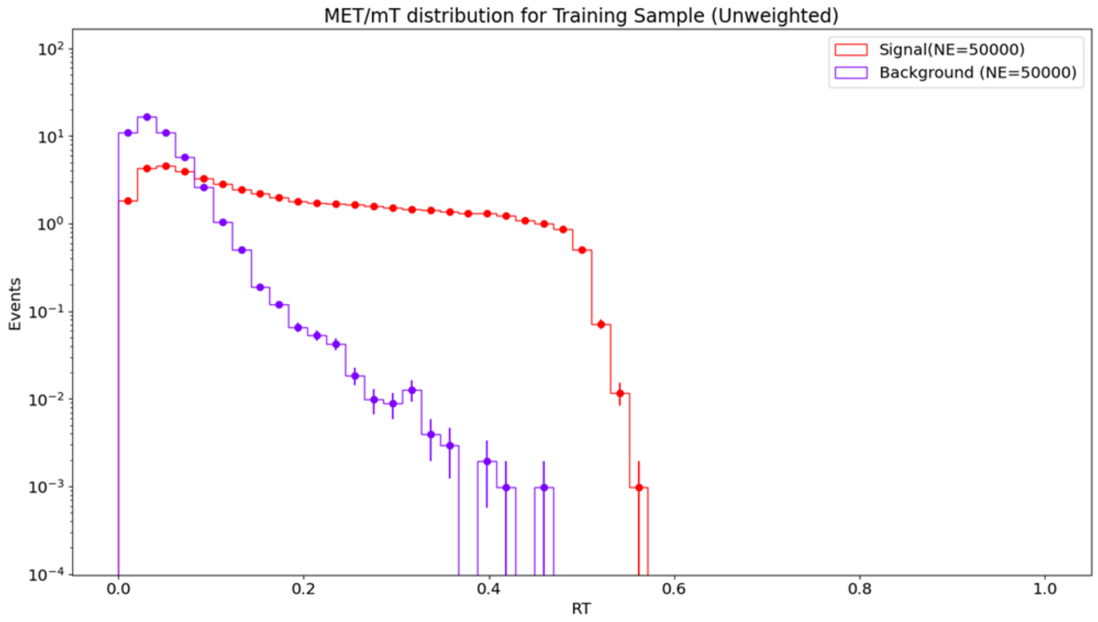
Δ R (leading, subleading jet) distribution for Training Sample (Unweighted)
Signal(NE=50000)
Background (NE=50000)
Events
Δ R

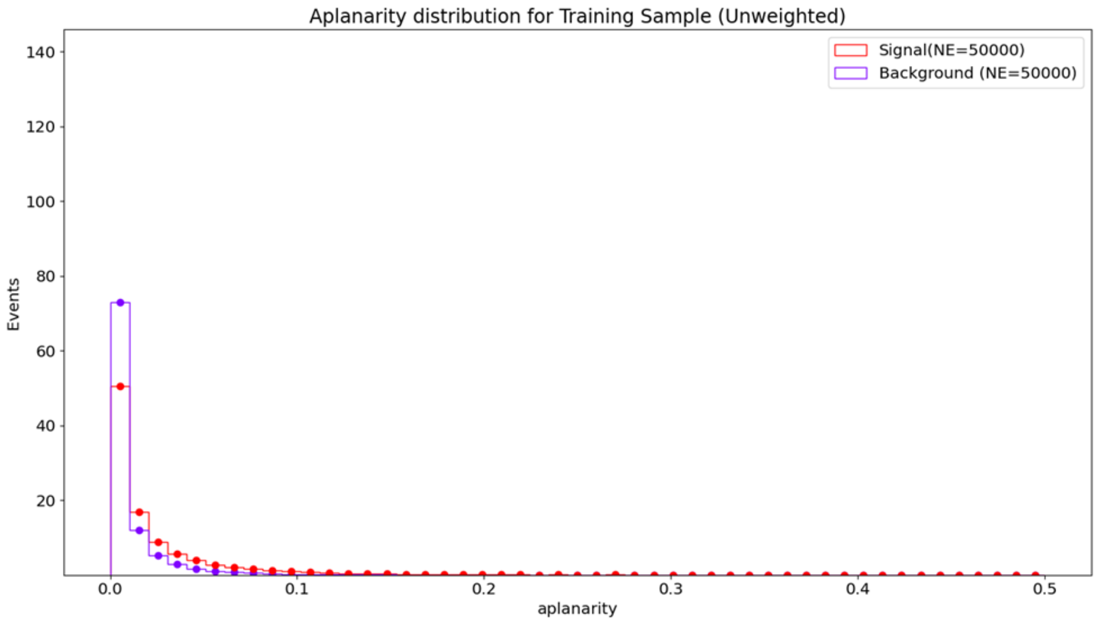


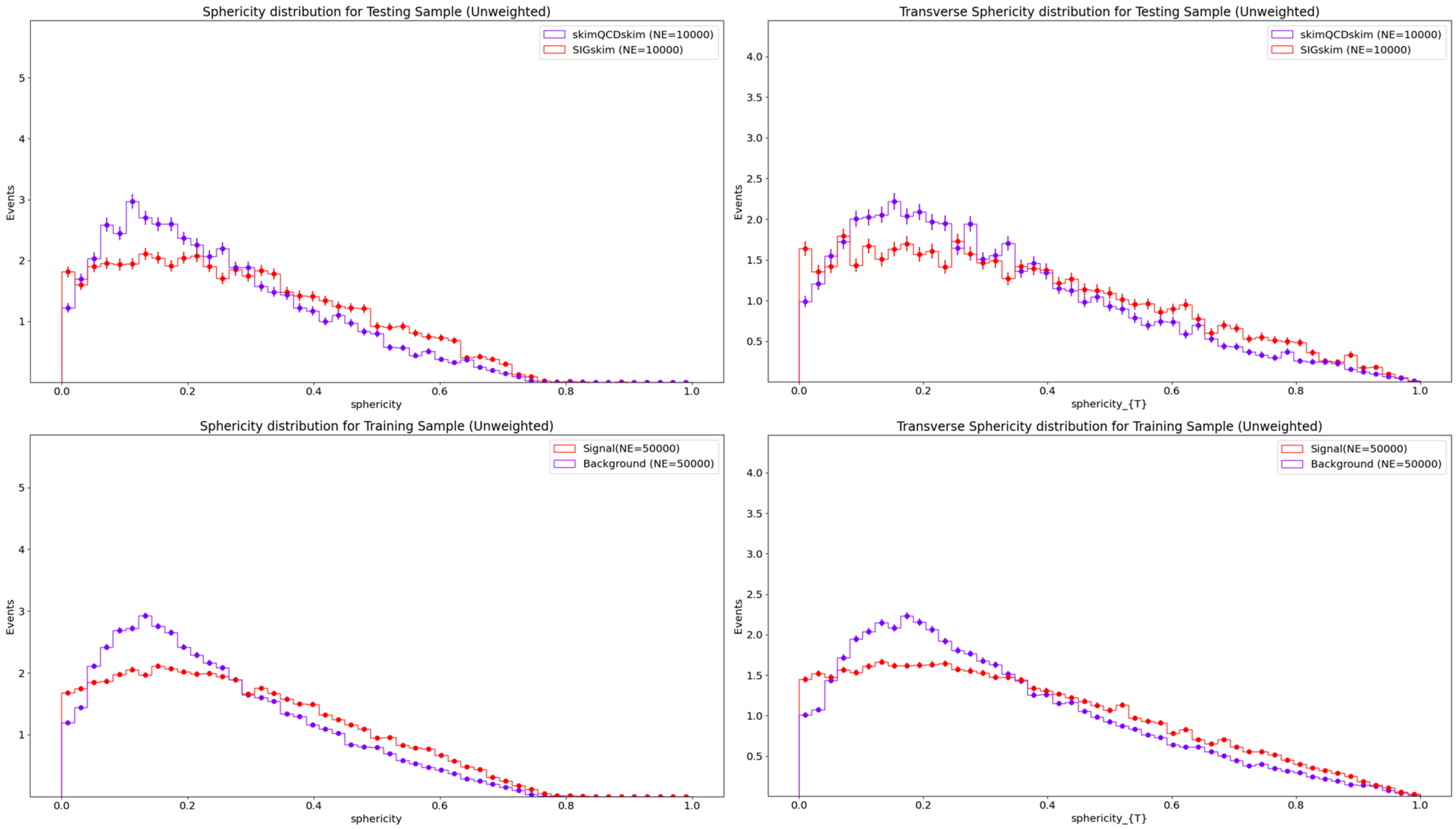
Δη (leading, subleading jet) distribution for Testing Sample (Unweighted)
skimQCDskim (NE=10000)
SIGskim (NE=10000)
Events
Δη
Δ Y (leading jet, subleading jet) distribution for Testing Sample (Unweighted)
skimQCDskim (NE=10000)
SIGskim (NE=10000)
Events
Δ Y
Δη (leading, subleading jet) distribution for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

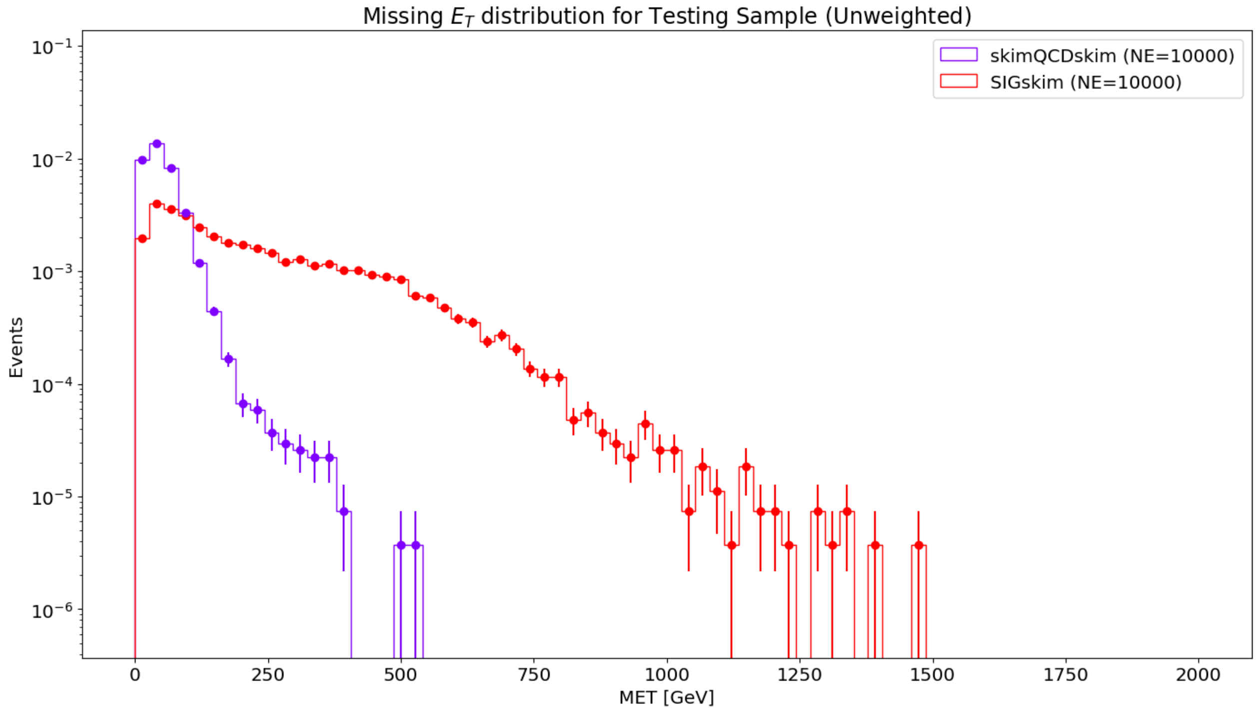

MET/mT distribution for Testing Sample (Unweighted)
skimQCDskim (NE=10000)
SIGskim (NE=10000)
Events
RT

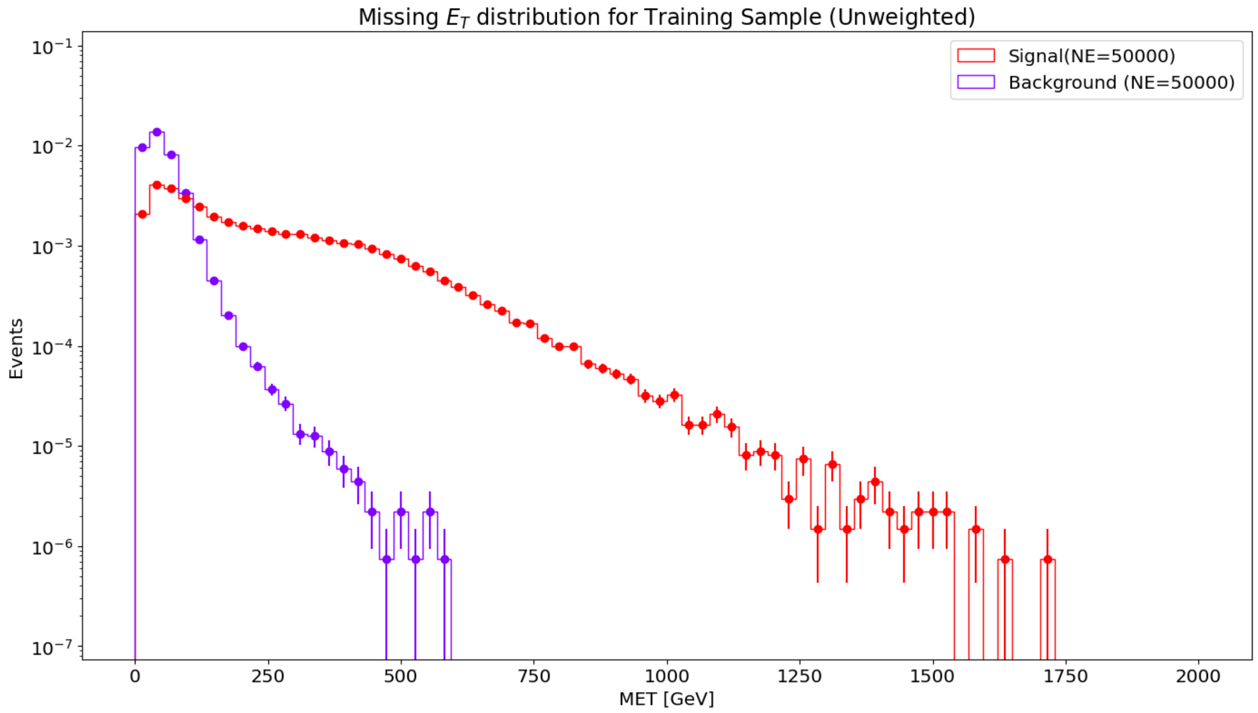

Aplanarity distribution for Testing Sample (Unweighted)
skimQCDskim (NE=10000)
SIGskim (NE=10000)
Events
aplanarity

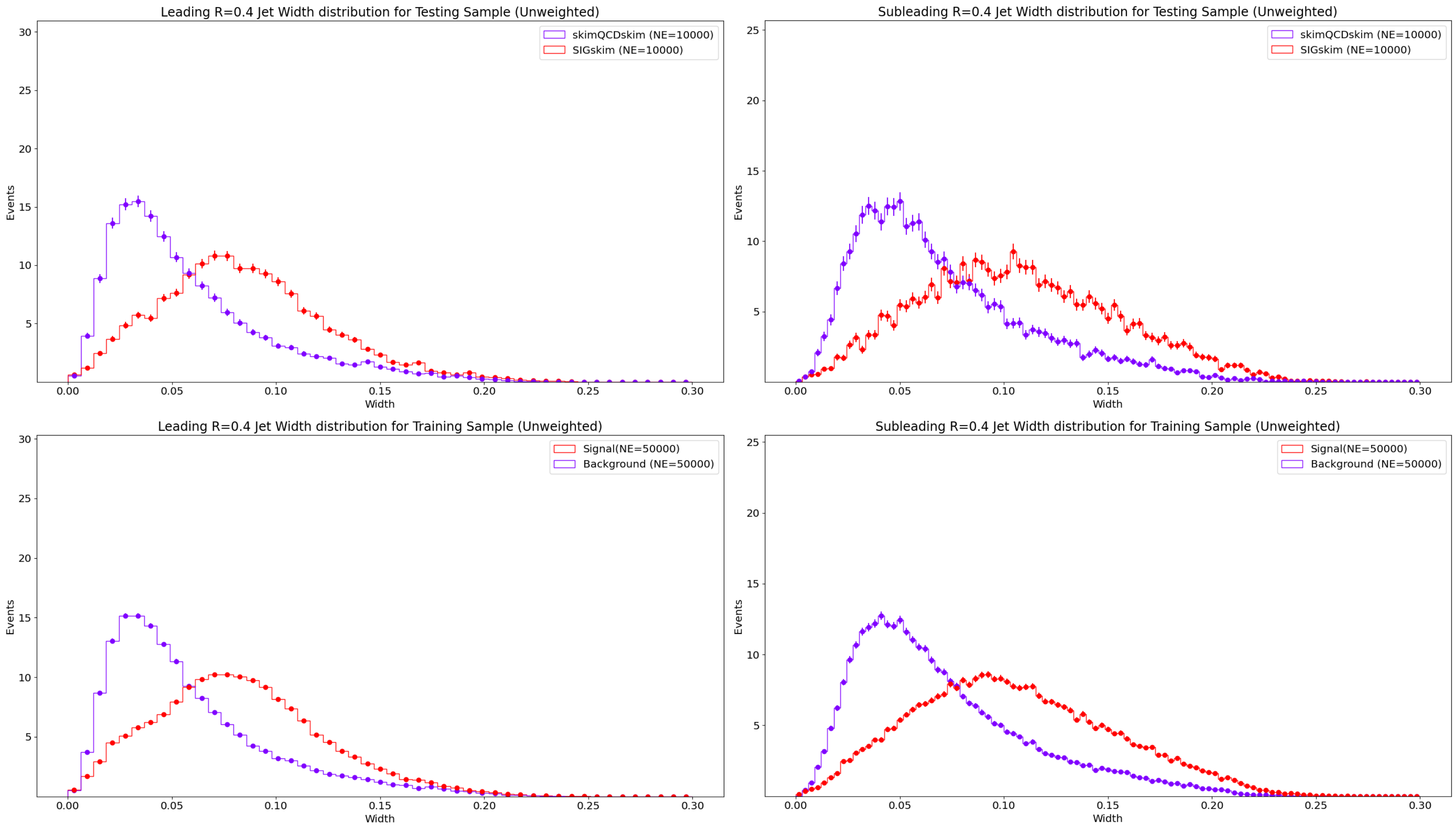

MET/mT distribution for Training Sample (Unweighted)
Signal(NE=50000)
Background (NE=50000)
Events
RT


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

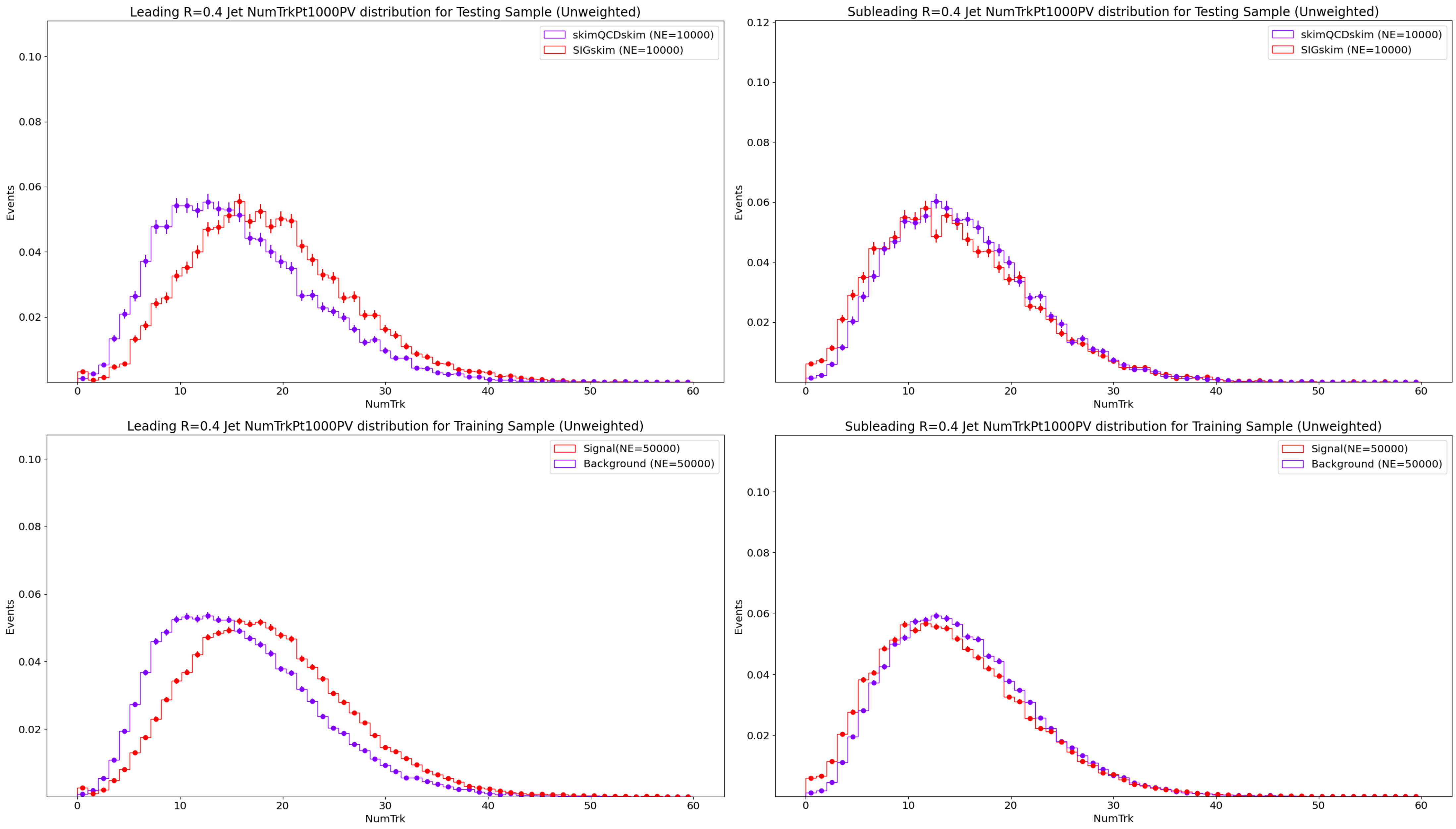

Sphericity distribution for Testing Sample (Unweighted)
skimQCDskim (NE=10000)
SIGskim (NE=10000)
Events
sphericity
Transverse Sphericity distribution for Testing Sample (Unweighted)
skimQCDskim (NE=10000)
SIGskim (NE=10000)
Events
sphericity_{T}
Sphericity distribution for Training Sample (Unweighted)
Signal(NE=50000)
Background (NE=50000)
Events
sphericity
Transverse Sphericity distribution for Training Sample (Unweighted)
Signal(NE=50000)
Background (NE=50000)
Events
sphericity_{T}

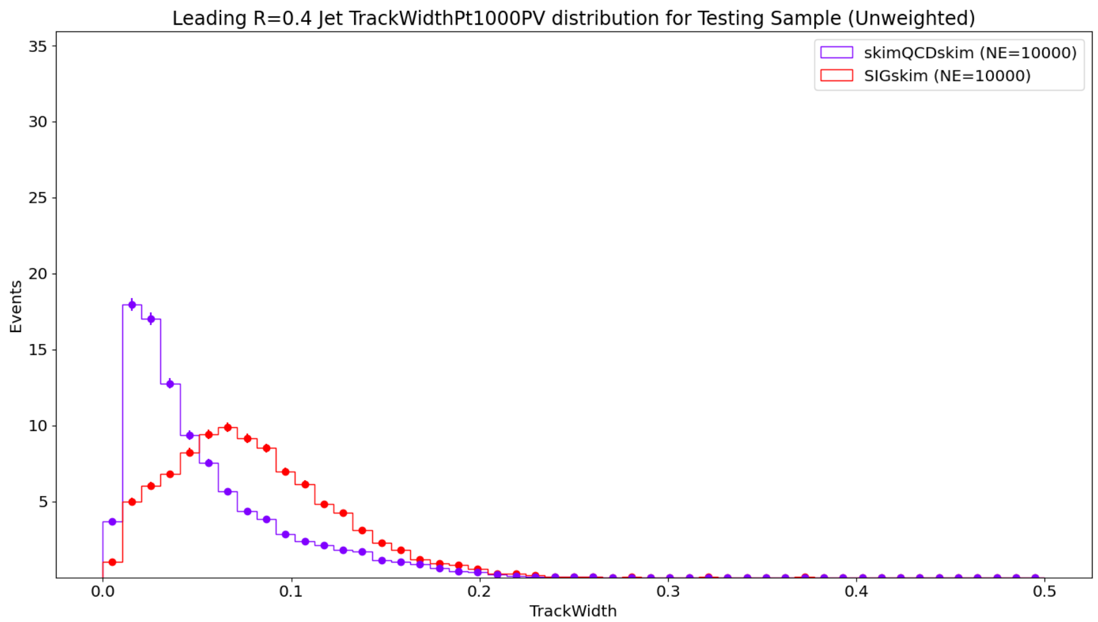

Missing E_T distribution for Testing Sample (Unweighted)
skimQCDskim (NE=10000)
SIGskim (NE=10000)
Events
MET [GeV]

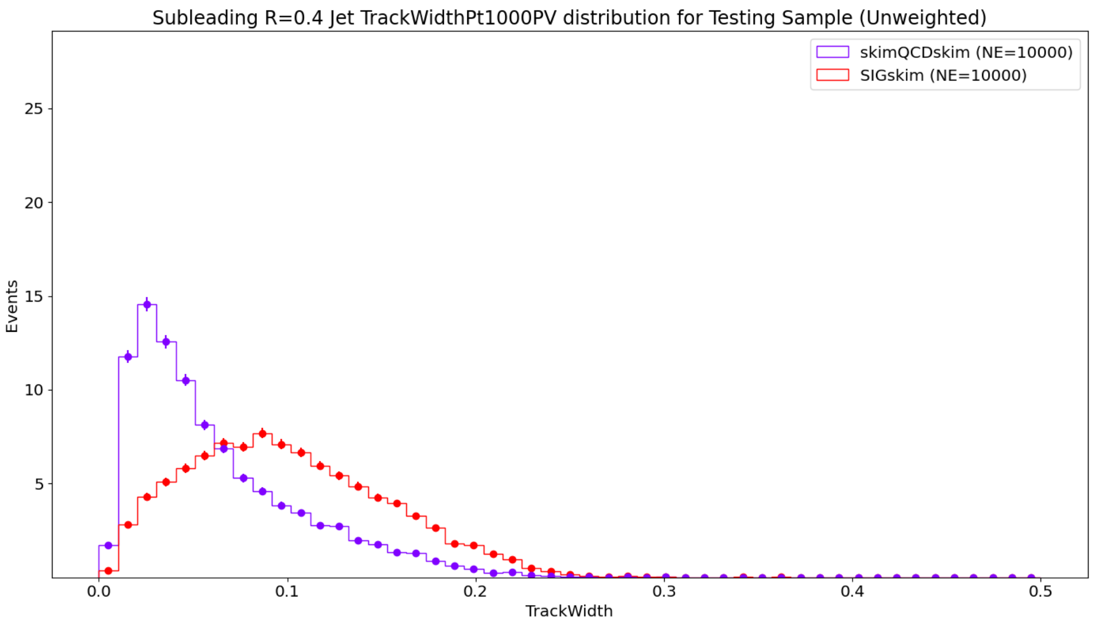

Missing E_T distribution for Training Sample (Unweighted)
Signal(NE=50000)
Background (NE=50000)
Events
MET [GeV]

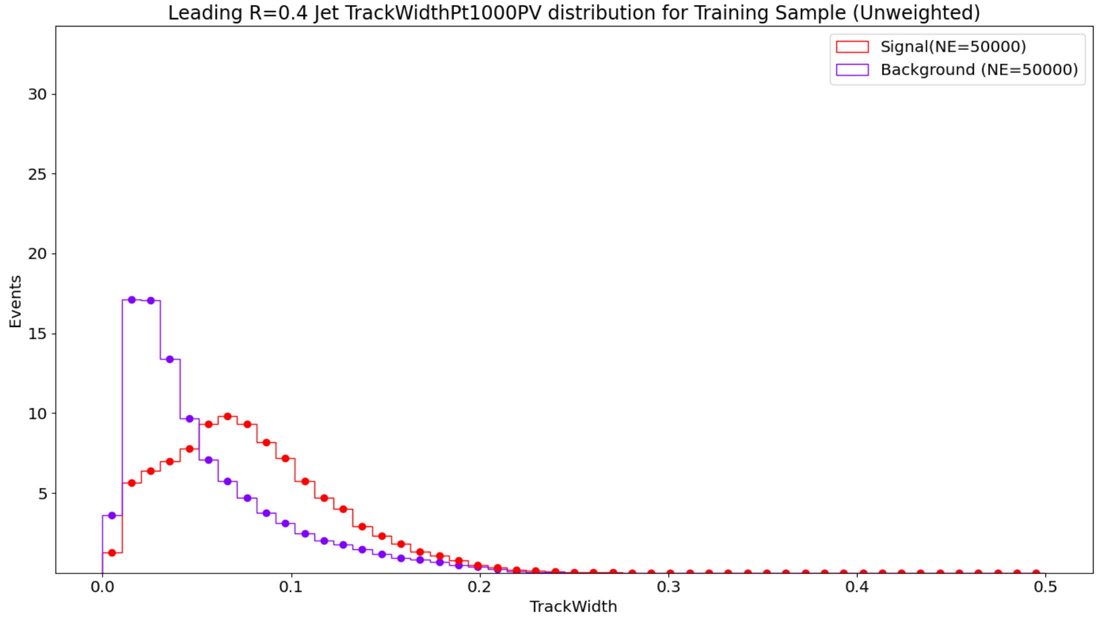


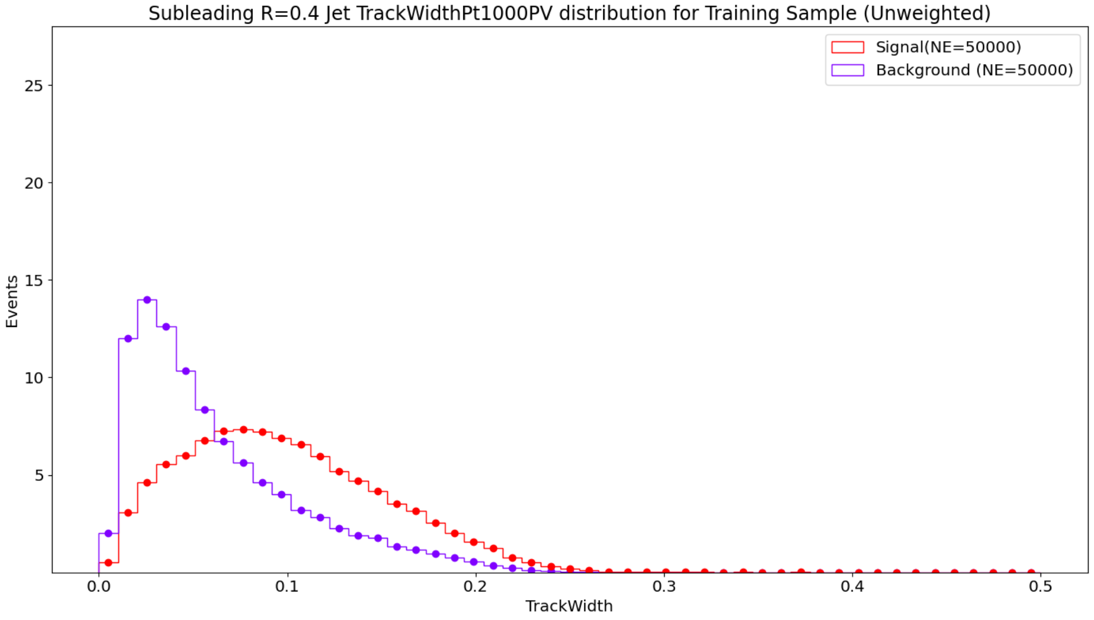
Leading R=0.4 Jet Width distribution for Testing Sample (Unweighted)
skimQCDskim (NE=10000)
SIGskim (NE=10000)
Events
Width
Subleading R=0.4 Jet Width distribution for Testing Sample (Unweighted)
skimQCDskim (NE=10000)
SIGskim (NE=10000)
Events
Width
Leading R=0.4 Jet Width distribution for Training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

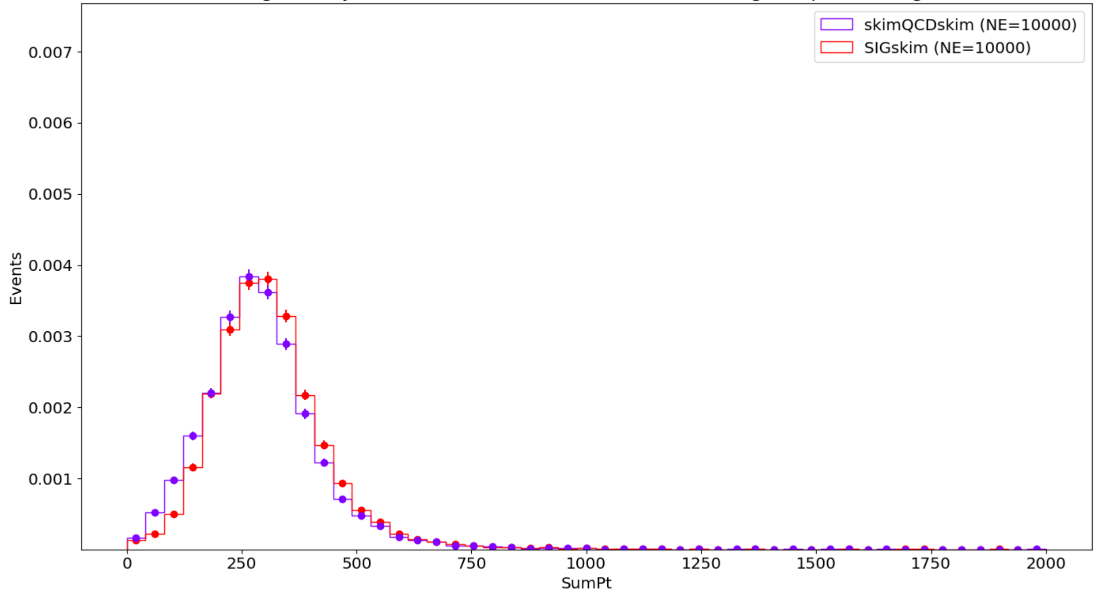

Leading R=0.4 Jet TrackWidthPt1000PV distribution for Testing Sample (Unweighted)
skimQCDskim (NE=10000)
SIGskim (NE=10000)
Events
TrackWidth


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

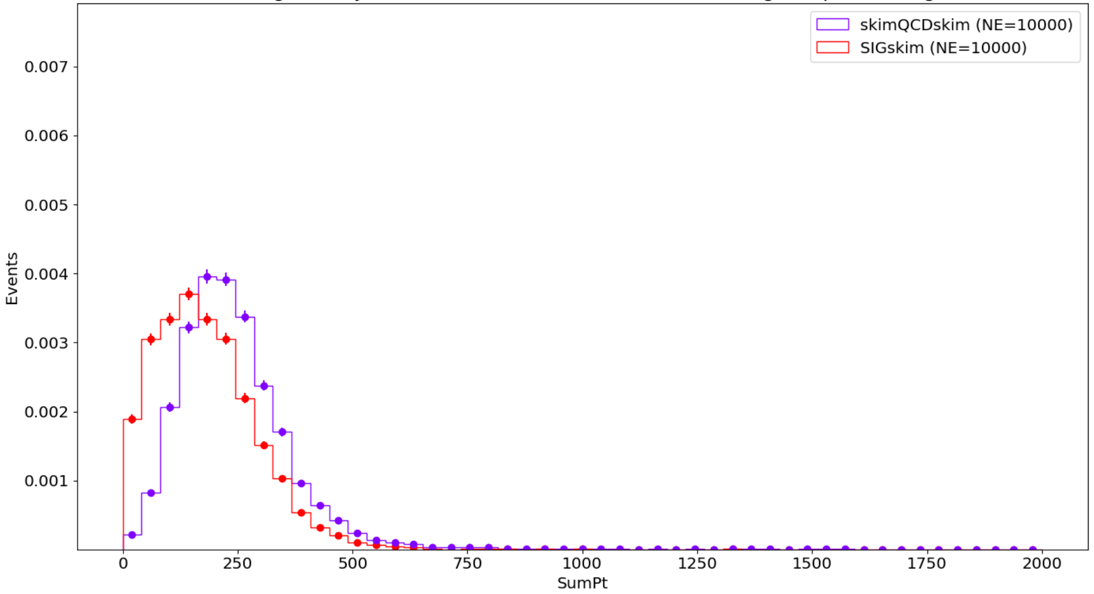

Leading R=0.4 Jet TrackWidthPt1000PV distribution for Training Sample (Unweighted)
Signal(NE=50000)
Background (NE=50000)
Events
TrackWidth


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

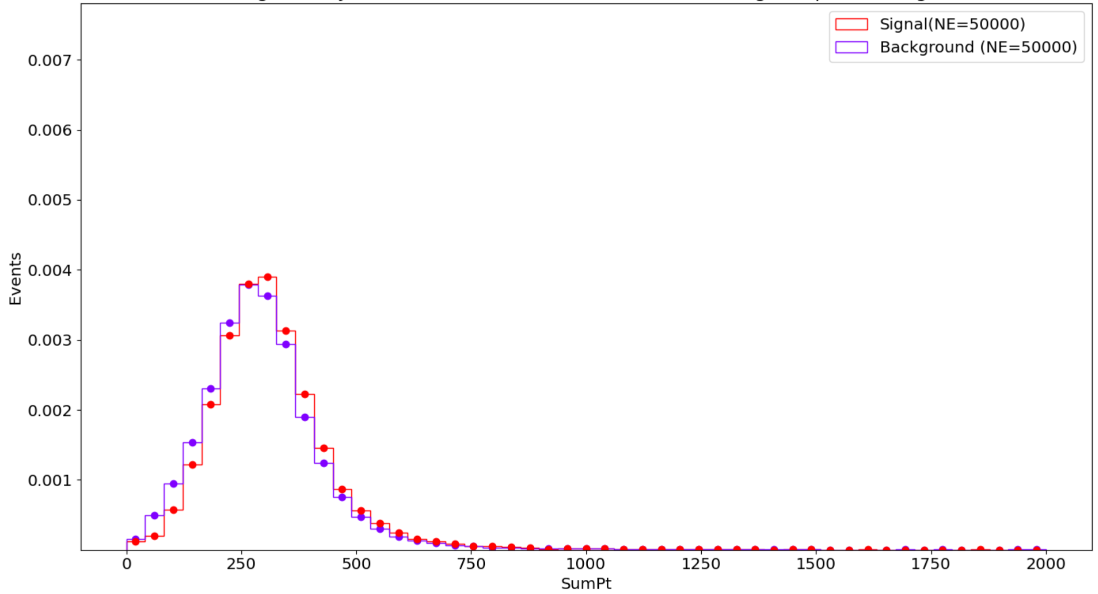
Leading R=0.4 Jet SumPtTrkPt500PV distribution for Testing Sample (Unweighted)



Subleading R=0.4 Jet SumPtTrkPt500PV distribution for Testing Sample (Unweighted)



Leading R=0.4 Jet SumPtTrkPt500PV distribution for Training Sample (Unweighted)



Subleading R=0.4 Jet SumPtTrkPt500PV distribution for Training Sample (Unweighted)