Machine Learning Methods for Detecting Semi-Visible Jets

Jonah Mougoue
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

- **Gauge Bosons** – Force carrying particles
  - Photons (\(\gamma\)) - Mediate the Electromagnetic (EM) force
  - Gluons (\(g\)) – Mediate the strong force
  - Z and W\(^{\pm}\) bosons - Mediate 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

$r_{\text{inv}}$
- Fraction of energy that is carried by dark matter particles
- High $r_{\text{inv}}$ 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 $r_{\text{inv}}$ 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
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**

- $n_{\text{jets}}$
  - Number of jets detected

- $\eta_{\text{Jet1}}$
  - The pseudorapidity of the leading jet

- $\Delta \eta$
  - Difference in transverse angles between the jets

- $\Delta p_t$
  - Difference in momentum between the jets divided by the momentum of the leading jet

- $\text{Pt\_balance}_{12}$
  - Difference in momentum between the jets divided by the momentum of the leading jet
### Variable Tables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>jet1</td>
<td>The leading jet / the jet with the highest pt</td>
</tr>
<tr>
<td>jet2</td>
<td>The subleading jet / the jet with the second highest pt</td>
</tr>
<tr>
<td>n_jets</td>
<td>Number of jets detected</td>
</tr>
<tr>
<td>jet1/2pt</td>
<td>Transverse momentum of jet1 and jet2</td>
</tr>
<tr>
<td>pt_balance(1)</td>
<td>((\text{jet}<em>{1\text{pt}} - \text{jet}</em>{2\text{pt}})/\text{jet}_{1\text{pt}})</td>
</tr>
<tr>
<td>jet1/2(\eta)</td>
<td>Pseudorapidity of jet1 and jet2</td>
</tr>
<tr>
<td>(\Delta\eta_{12})</td>
<td>The difference between jet1(\eta) and jet2(\eta)</td>
</tr>
<tr>
<td>MET</td>
<td>Missing energy in the transverse direction</td>
</tr>
<tr>
<td>mT</td>
<td>The total reconstructed mass</td>
</tr>
<tr>
<td>rT</td>
<td>MET/mT</td>
</tr>
<tr>
<td>(\Delta\phi_{\text{min}})</td>
<td>The minimum transverse angle from either jet to the direction of MET</td>
</tr>
<tr>
<td>(\Delta\phi_{\text{max}})</td>
<td>The maximum transverse angle from either jet to the direction of MET</td>
</tr>
<tr>
<td>(\max\phi_{\text{min}})</td>
<td>The difference between (\Delta\phi_{\text{max}}) and (\Delta\phi_{\text{min}})</td>
</tr>
<tr>
<td>(\Delta R)</td>
<td>The solid angle between the two leading jets</td>
</tr>
<tr>
<td>delta(\gamma_{12})</td>
<td>the difference in rapidity between jet1 and jet2</td>
</tr>
<tr>
<td>Aplanarity</td>
<td>How well the jets are distributed in the transverse plane</td>
</tr>
<tr>
<td>Sphericity</td>
<td>A measure of the spherical symmetry of the distribution of jets</td>
</tr>
<tr>
<td>Sphericity(r)</td>
<td>Sphericity in the transverse plane</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jet1/2__width</td>
<td>Jet width calculated using calorimeter data</td>
</tr>
<tr>
<td>Jet1/2__TrackWidth(\text{Pt1000PV})</td>
<td>The width between the two furthest tracks with pt over 1000 MeV within the jet from the primary vertex</td>
</tr>
<tr>
<td>Jet1/2__SumPt(\text{Trk500PV})</td>
<td>The pt sum of each track of at least 500 MeV within the jet from the primary vertex</td>
</tr>
<tr>
<td>Jet1/2__Num(\text{Trk1000PV})</td>
<td>The amount of tracks of at least 1000 MeV within the jet from the primary vertex</td>
</tr>
</tbody>
</table>
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 $n_{\text{jets}}$, $rT$, MET aplanarity, sphericity, sphericity$_T$, jet1/2$_{\text{SumPtTrkPt500PV}}$, or jet1/2$_{\text{TrackWidthPt1000PV}}$

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 $r_{\text{inv}}$ 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 ($\eta$) is an angular coordinate describing a particle's angle relative to the beam.
Jet2\textsubscript{width} cut 0.05

Signal region: jet2\textsubscript{width} > 0.05

Control region: jet2\textsubscript{width} < 0.05

PFN cut > 0.97
Jet$_{2\text{width}}$ cut 0.1

Signal region: jet$_{2\text{width}}$ > 0.1
Control region: jet$_{2\text{width}}$ < 0.1
PFN cut > 0.97
Results

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

Both Jet2\textsubscript{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\text{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

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. Run 3 of the large hadron collider. [https://home.cern/press/2022/run-3](https://home.cern/press/2022/run-3).
Backup
Questions?
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

10000 test sig, 10000 test bkg, train sig 50000, train bkg 50000

test samples: tt, ft, tt = (9410, 1590, 1379, 8621)

Signal (test sample) [9410]

Background (test sample)

Signal (train sample) [9410]

Background (train sample)

10000 test sig, 10000 test bkg, train sig 50000, train bkg 50000

test samples: tt, ft, tt = (9488, 1012, 999, 9001)

Signal (test sample) [9488]

Background (test sample)

Signal (train sample) [9488]

Background (train sample)