

Developing a BDT for finding Dark Photon Decays to Two  
Electrons within the ATLAS Detector  
REU Program at Columbia University - Nevis Labs

Amelia Stevens<sup>1</sup>

<sup>1</sup>Haverford College, Haverford, PA, 19041

August 2, 2024



NEVIS LABORATORIES  
COLUMBIA UNIVERSITY



## Abstract

This paper searches for electron pairs originating from the decay of long-lived dark photons. This analysis utilizes data from proton-proton collisions at 13.6 TeV with  $10\text{ fb}^{-1}$  of integrated luminosity collected by the ATLAS detector in 2022. In our case, we use a Hidden Abelian Higgs Model (HAHM), which could explain dark matter and we focus on the case where the dark photon is long lived and decays to two electrons. This study proposes leveraging a boosted decision tree (BDT). This machine-learning algorithm can distinguish the signal of displaced electron pairs from prompt electrons from the Standard Model background and offer a potential avenue for testing theories predicting electron pairs, making this a promising approach to exploring the nature of dark matter at the LHC.

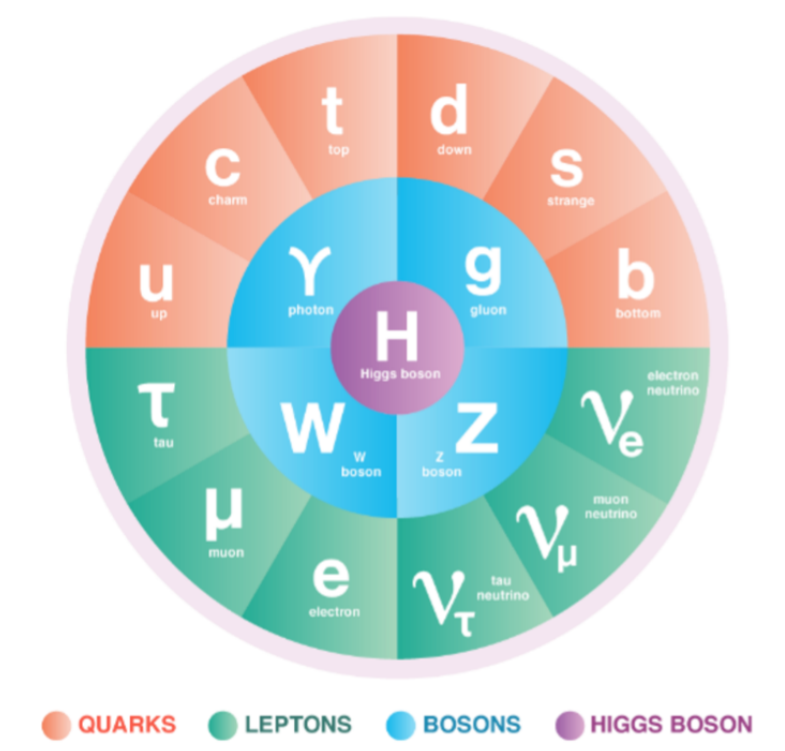
# Contents

|          |                                                                              |           |
|----------|------------------------------------------------------------------------------|-----------|
| <b>1</b> | <b>Introduction</b>                                                          | <b>2</b>  |
| 1.1      | The Standard Model . . . . .                                                 | 2         |
| 1.2      | Beyond The Standard Model . . . . .                                          | 3         |
| 1.3      | Instrumentation . . . . .                                                    | 4         |
| 1.3.1    | The Large Hadron Collider . . . . .                                          | 4         |
| 1.3.2    | The ATLAS Detector . . . . .                                                 | 4         |
| <b>2</b> | <b>The Search for Displaced Electrons from Long-Lived Dark Photon Decays</b> | <b>6</b>  |
| 2.1      | Motivation . . . . .                                                         | 6         |
| 2.2      | Data and Simulated Samples . . . . .                                         | 6         |
| 2.3      | Signal Variable Distributions . . . . .                                      | 7         |
| 2.4      | Boosted Decision Trees . . . . .                                             | 8         |
| 2.4.1    | BDT Introduction . . . . .                                                   | 8         |
| 2.4.2    | Preselection . . . . .                                                       | 9         |
| 2.4.3    | Input Variable Selection . . . . .                                           | 10        |
| 2.4.4    | Validation Procedure . . . . .                                               | 11        |
| <b>3</b> | <b>Results</b>                                                               | <b>12</b> |
| 3.1      | BDT above 300 GeV vs. BDT below 300 GeV Comparison . . . . .                 | 12        |
| 3.2      | BDT with high delta M vs. BDT with low delta M Comparison . . . . .          | 14        |
| <b>4</b> | <b>Summary and Conclusions</b>                                               | <b>16</b> |
| <b>5</b> | <b>Acknowledgements</b>                                                      | <b>17</b> |
| <b>6</b> | <b>Appendix</b>                                                              | <b>19</b> |
| 6.1      | BDT Input Variables . . . . .                                                | 19        |
| 6.2      | Unused BDT Input Variables . . . . .                                         | 20        |

# 1 Introduction

## 1.1 The Standard Model

Despite its complexity, the universe is composed of simple ingredients. The Standard Model of particle physics is our current best theory explaining these fundamental building blocks and the forces that govern their interactions. Developed in the 1970s, it has become a cornerstone of modern physics, successfully explaining experimental results and predicting new phenomena. [1].



*Figure 1: A Visual Guide to the Standard Model of Particle Physics [2]*

The Standard Model categorizes everything into three main types of particles: fermions gauge bosons and the Higgs boson. As shown in Figure 1, Fermions are the matter-makers, the ingredients that form atoms and everything we see around us. Fermions can fall into two main categories: quarks and leptons. There are six types of quarks – up, down, charm, strange, top, and bottom. Interestingly, quarks possess electric charge (except up and down) and a property called "color" (unrelated to visual color) that exists in three varieties. These color combinations are crucial – quarks can only exist in groups of three (with specific color combinations) to form protons and neutrons, the building blocks of atomic nuclei. Leptons include the familiar electron, the building block of atoms, alongside its heavier cousins, the muon and the tau. Leptons also include neutrinos, ghostly particles that barely interact with matter and have almost no mass. Unlike quarks, leptons can exist on their own. [1]

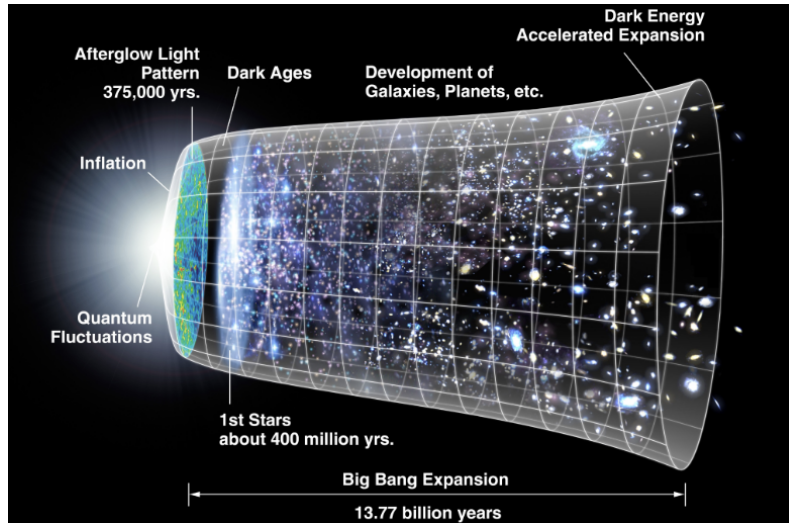
The universe is not just a collection of static particles. They interact with each other through three fundamental forces: the strong force, the electromagnetic force, and the weak force. Bosons are the force carriers. They are the messengers that mediate the interactions between fermions. The Standard model describes the vital force that binds quarks within protons and neutrons. The carrier particles responsible for this interaction are called gluons. The electromagnetic force governs the behavior of electrically charged particles, including electrons. It's responsible for the attraction between oppositely charged particles and repulsion between like charges. The force carrier is the photon, the particle of light. The weak force is responsible for certain types of radioactive decay and is essential for the nuclear fusion process that powers stars. The W and Z

bosons act as the messengers for this force. [1]

While the Standard Model excels in explaining these three forces, it falls short in explaining the fourth fundamental force – gravity. We are not looking for theories beyond the Standard Model to resolve the various shortcomings. Gravity, the force that holds galaxies together, remains an outsider. Though weak at the subatomic level, gravity plays a dominant role in our everyday lives. Unifying gravity with the other forces within the framework of the Standard Model remains a significant challenge for physicists. Its influence at the subatomic level is negligible compared to the different forces, making it challenging to incorporate it within the Standard Model’s framework. One of the Standard Model’s triumphs was the prediction and subsequent discovery of the Higgs boson in 2012. This elusive particle is linked to the Higgs field, which permeates all space. As particles interact with this field, they acquire mass. The Higgs boson’s discovery provided a crucial missing piece in understanding how particles get their mass. [1]

## 1.2 Beyond The Standard Model

Despite its explanatory power, the Standard Model still needs to be completed. It does not explain gravity, the dominant force at the macroscopic level. It remains silent on most of the universe’s energy and mass, which exist as dark matter and energy. The Big Bang should have created equal amounts of matter and antimatter, particles with opposite properties. However, we see very little antimatter in the universe, as shown in Figure 2. The Standard Model needs help to explain this asymmetry. Lastly, the Standard Model describes three generations of quarks and leptons with vastly different masses. The reason for this remains unclear. The discovery of the Higgs boson in 2012, a key prediction of the Standard Model, provided a missing piece in our understanding of how particles acquire mass. [1]



*Figure 2: 13.77 Billion Years of Cosmic History [2]*

These unanswered questions point toward new physics beyond the Standard Model. Particle accelerators like the Large Hadron Collider (LHC) constantly search for new particles and forces that might complete our understanding of the universe’s building blocks. Despite these shortcomings, the Standard Model has revolutionized our understanding of the subatomic world and laid the foundation for future discoveries. As we delve deeper into the universe’s mysteries with powerful particle accelerators like the Large Hadron Collider, the Standard Model will likely serve as a stepping stone to a more comprehensive theory that unifies all the forces and explains the full spectrum of matter and energy in our cosmos. [1]

## 1.3 Instrumentation

### 1.3.1 The Large Hadron Collider

The Large Hadron Collider (LHC) is the largest particle accelerator in the world. It is on the border between France and Switzerland, between Lake Geneva and the Jura Mountains. It is known to be the largest and most complex machine that humanity has been able to construct. The LHC is designed to smash protons together at high energies, recreating the conditions in the first microseconds after the Big Bang. The LHC is a circular tunnel with a circumference of 27 kilometers (17 miles) as seen in Figure 3. It contains two beams of protons that circulate in opposite directions, each traveling at 99.9 percent of the speed of light. The protons are accelerated by superconducting magnets to an energy of 14 teraelectronvolts (TeV) and collide at 40 MHz. The LHC has four main experiments: ALICE, ATLAS, CMS, and LHCb. These experiments are designed to detect different particles produced when the protons collide. The LHC has been used to discover the Higgs boson, a fundamental particle responsible for other particles' mass. The LHC is also used to search for new particles and study the Higgs boson's properties. The LHC is a primary scientific instrument that helps us understand nature's fundamental laws. It is one of the most complex and ambitious scientific projects ever undertaken, and it is expected to continue producing groundbreaking discoveries for many years. [3]

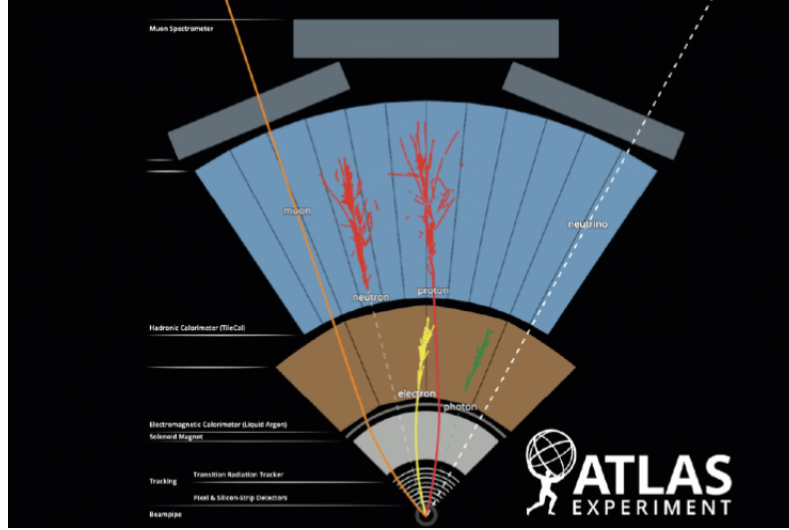


*Figure 3: Large Hadron Collider at CERN [3]*

### 1.3.2 The ATLAS Detector

ATLAS is one of the four main experiments at the Large Hadron Collider (LHC) at CERN. The LHC accelerates protons in two different directions and collides in four points. ATLAS is one of the four colliding points. It is an extensive collaboration of around 3000 people and people from more than 42 countries. ATLAS has played a significant role in understanding the origins of matter and the universe. The detector itself is similar to a fancy camera. It takes a snapshot when two protons collide. All the energy released can be transformed into mass  $E = mc^2$ . Scientists can create new particles through ATLAS and study how they behave. It is a general-purpose detector, meaning that it is designed to detect a wide range of particles and events produced in high-energy proton collisions. ATLAS is located in a cavern 100 meters (330 feet) underground near Meyrin, Switzerland. It is a cylindrical detector with a diameter of 44 meters (144 feet) and a length of 25 meters (82 feet). ATLAS weighs approximately 7,000 tonnes (15 million pounds)

and comprises multiple layers of detectors to track particles, measure their energy, and identify their type. ATLAS is designed to detect a wide range of particles, including photons, electrons, muons, hadrons, and jets. It is also intended to measure these particles' energy and momentum and reconstruct the events that produced them. [4]



**Figure 4:** *ATLAS Detector Event Display* [4]

The detector itself, displayed in Figure 4 consists of six main systems. [4]

1. Inner Detector: Located closest to the beam pipe, this cylindrical detector is crucial for tracking the paths of charged particles. Composed of high-resolution silicon pixel and strip detectors, it provides precise position information for each particle interaction. The pixel detector, closest to the beam pipe, uses millions of tiny pixels to record detailed spatial information about particle tracks. The strip detector, located further outwards, consists of long, thin strips of silicon that measure the particle's trajectory in the other dimension. By combining data from these layers, scientists can reconstruct the 3D tracks of charged particles, revealing their momentum and origin.
2. Calorimeter: The Electromagnetic Calorimeter (EMCal) detector measures the energy of electrons and photons (particles that interact with the electromagnetic force). When a high-energy electron or photon enters the EMCal, it interacts with the material, creating a cascade or "shower" of lower-energy particles. The EMCal is made of layers of lead or tungsten and liquid argon, which efficiently absorbs the photons produced. By measuring the shower's intensity, scientists can determine the initial energy of the electron or photon.
3. Hadron Calorimeter: Designed to measure the energy of hadrons (particles that interact strongly with the strong nuclear force, like protons and neutrons). Unlike electrons and photons, hadrons typically shower and interact within the HCal, depositing all their energy. The HCal is constructed with alternating layers of steel or iron (absorbers) and plastic scintillators (active material). As hadrons pass through the steel or iron, they produce secondary particles. These secondary particles then ionize the scintillator material, emitting light, which is detected by photodetectors. The amount of light detected is proportional to the energy of the hadron.
4. Muon Spectrometer: The outermost layer of the ATLAS detector, this component is dedicated to identifying muons. Muons are much heavier than electrons and can penetrate the calorimeters relatively easily. The MS combines three large air-core toroid magnets and chambers filled with gaseous detectors. The magnets bend the trajectories of muons,

allowing scientists to measure their momentum based on the curvature. Gaseous detectors like drift tubes or resistive plate chambers track the paths of the muons, enabling the reconstruction of their tracks.

5. Solenoid Magnet: This giant superconducting magnet creates a strong, uniform magnetic field within the central region of the ATLAS detector. Charged particles like electrons, muons, and pions are forced to curve as they travel through the magnetic field. The amount of curvature is directly related to the particle's momentum. By measuring the curvature, scientists can precisely determine the momentum of these charged particles.
6. Toroid Magnets: Large, air-core toroid magnets surround the muon spectrometer. Toroids have a donut-like shape with a coil that produces a magnetic field within the hole in the center. These are large, air-core toroid magnets surrounding the muon spectrometer. Toroids have a donut-like shape with a coil that produces a magnetic field within the hole in the center. Similar to the solenoid magnet, the curvature of the muon tracks allows scientists to measure their momentums. Identical to the solenoid magnet, the curvature of the muon tracks will enable scientists to measure their momentum.

## 2 The Search for Displaced Electrons from Long-Lived Dark Photon Decays

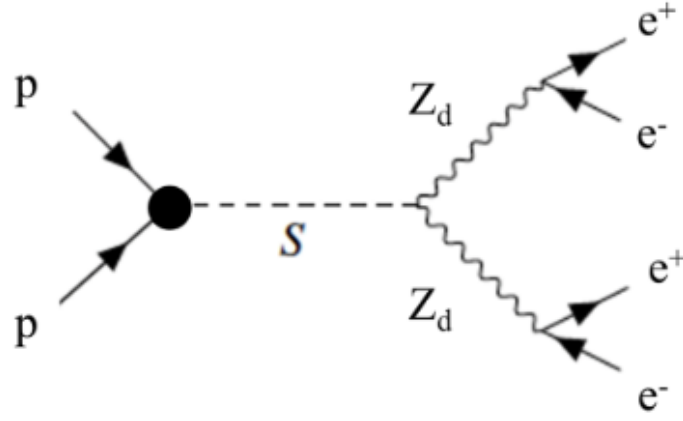
### 2.1 Motivation

Most previous searches for dark photons have focused on the case where the dark photon decays promptly. The focus is on instances where the dark photon has a long lifetime, allowing it to travel a measurable distance before decaying. The research aims to identify a potential new particle, the "dark photon," by examining its decay into electrons. We are investigating a theoretical model that introduces additional particles to the Standard Model to achieve this. The focus is on instances where the dark photon has a long lifetime, allowing it to travel a measurable distance before decaying. Currently, the focus is on understanding the electron variables before integrating them into the model. [7] [8] [10]

We are trying to look at a dark matter photon in the Feynman diagram. We have a Hidden Abelian Higgs Model (HAHM) that adds two new particles to the standard model, namely the "dark" Higgs (S) located in the diagram and the "dark photons"  $Z_d$ . We are looking for signals where the dark photons are long-lived and where they decay into two electrons (an electron and a positron). Regarding the electron's final state, we are looking for a long-lived particle, "dark photon," where instead of being produced and decaying immediately in the center of the detector, we make an electron that travels through the tracker. If it is long-lived, it will travel a distance and then decay. If it decays into two electrons, it will have a shorter track on the outer part, called a displaced electron. Electrons and photons look similar, but the only difference is that an electron leaves a track while a photon does not. With a displaced electron, we can mistakenly identify an electron as a photon. My research relates to the electron final state and looks into electron variables (pt, d0, time, etc). These will be the inputs for the BDT model. [7] [8] [10]

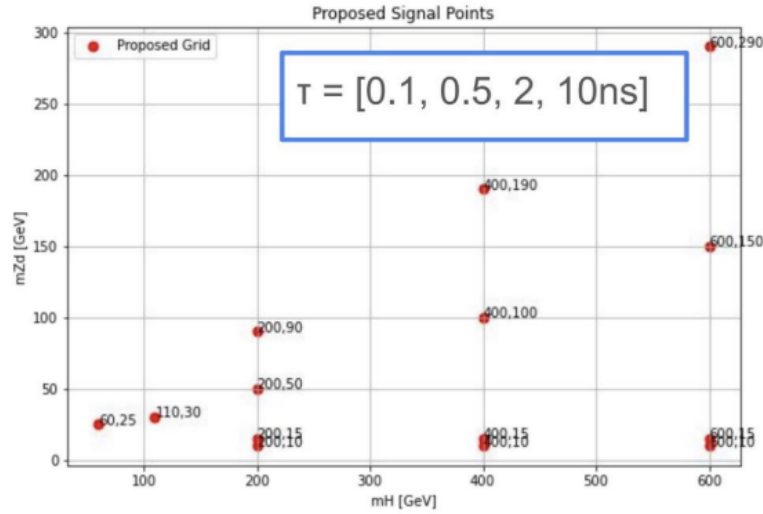
### 2.2 Data and Simulated Samples

Training is done on simulation of HAHM dark photons which decay to electrons (signal) and data taken with the ATLAS detector in 2022. The code incorporates cuts to isolate specific final states based on electron count (e.g., "ee" for two electrons). I select only events with at least 2 electrons. These simulated samples serve as the training data for the algorithm, while the performance of the algorithm will be evaluated using withheld, or blinded, data at a later



**Figure 5:** Feynman Diagram of electron pairs induced by Gauge Boson  $Z_d$  in the Dark Higgs Boson ( $S$ ) model [10]

stage. The signal sample, denoted as "400, 100, 2ns", represents a Monte Carlo simulation of a Dark Higgs boson with a mass of 400 GeV and a Dark Photon with a mass of 100 GeV and a 2 nanosecond lifetime. This signal is contrasted against background data, which include 3 million events of data 22. The core parameters for analysis within this dataset are mass and lifetime.



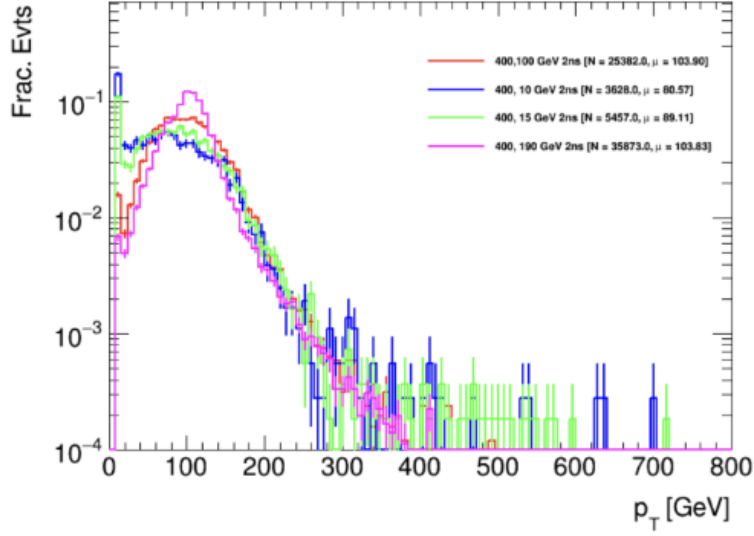
**Figure 6:** Proposed Search Grid for a  $Z_d$  Boson with Varying Decay Times [14]

## 2.3 Signal Variable Distributions

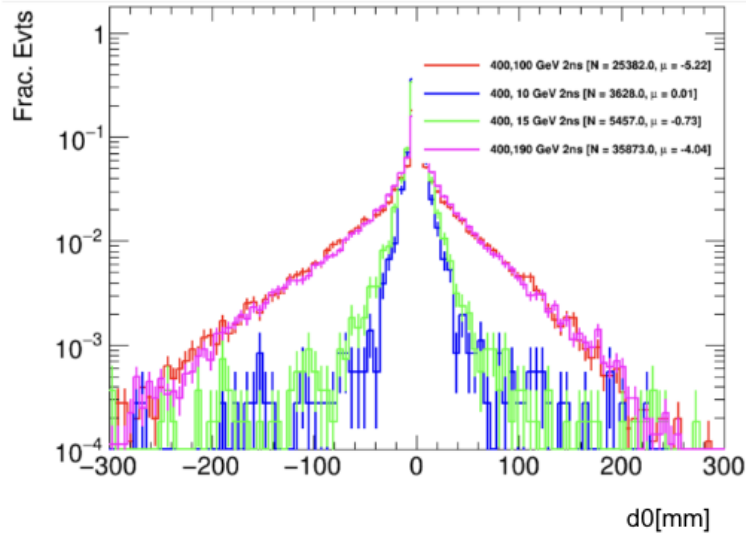
Three crucial properties of the leading electron in a particle collision.

1.  $p_t$  (transverse momentum): This measures the momentum of the electron perpendicular to the beam axis. High  $p_t$  electrons are more likely to come from a hard scattering process (like a Dark Photon decay) compared to background electrons from random interactions.
2.  $d_0$  (transverse impact parameter): This measures the distance of closest approach in the transverse plane of the track from the primary vertex (assumed collision point). Long-lived particles like the theorized Dark Photon might decay further away from the collision point, leading to a larger  $d_0$  compared to background electrons from prompt decays.

- time - calibrated time of detection of the electron as measured by the LAr calorimeter.



**Figure 7:** Transverse Momentum ( $p_T$ ) for Different 400 GeV Signals



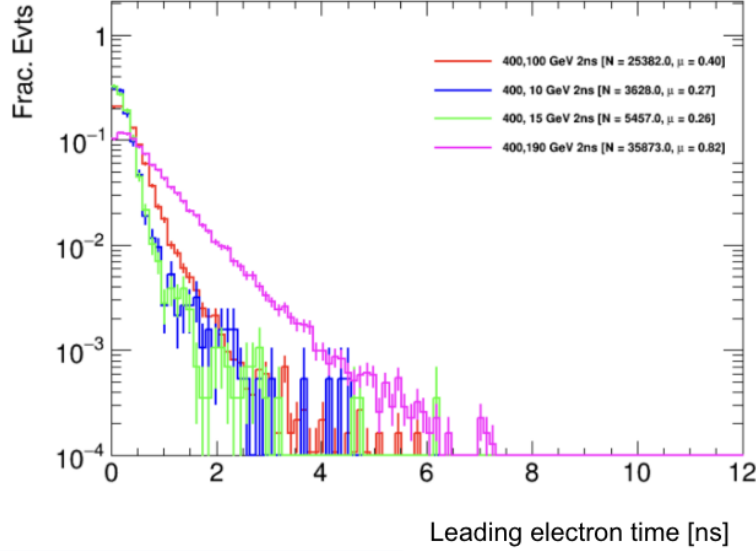
**Figure 8:** Impact Parameter ( $d_0$ ) Distribution for Different 400 GeV Signals

These variables ( $p_t$ ,  $\phi$ ,  $d_0$ ) along with potentially others, would be extracted from both the simulated signal and background events. The machine learning algorithm (likely a Boosted Decision Tree - BDT) would then learn to identify patterns in these distributions that differentiate signal from background. For example, the BDT might learn that a high  $p_t$  electron with a large  $d_0$  is more likely to be a signal candidate from a Dark Photon decay compared to a low  $p_t$  electron with a small  $d_0$  (typical of background). By studying the distributions of these variables for a large number of electrons, we can identify patterns that might indicate the presence of new particles, such as the dark photon.

## 2.4 Boosted Decision Trees

### 2.4.1 BDT Introduction

BDT sets up a training process for a machine learning algorithm (TMVA) to distinguish between signal (presence of Dark Photon) and background (events not containing the Dark Photon)



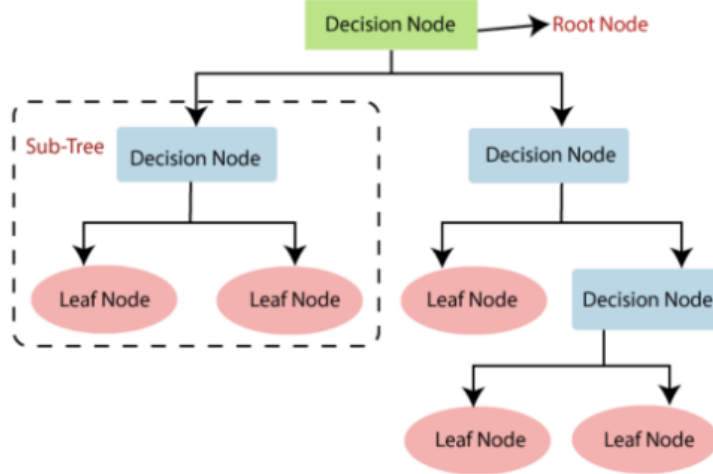
**Figure 9:** Leading Photon Time Distribution for Different 400 GeV Signals

events. The BDT is a type of supervised learning algorithm. This means it needs to be trained on a dataset with labeled examples (signal and background events in this case) before it can make predictions on unseen data. The goal here is to classify events into two categories: signal or background. BDT excels at this type of problem. BDTs are ensemble methods, meaning they combine the predictions of multiple weaker decision trees to create a stronger final model. Each decision tree in the ensemble makes a simple yes/no decision based on a specific feature of the data (e.g., energy of an electron). The "boosting" part refers to how the individual decision trees are trained sequentially. Each tree focuses on learning from the mistakes of the previous ones, leading to a more robust final model.

We can set up the training process for a TMVA (Toolkit for Multivariate Data Analysis) which can utilize BDTs. The code defines cuts to select features (like electron energy, time of flight) from the simulated signal and background events. These features are fed into the BDT training process. The BDT learns by iteratively splitting the data based on the features, aiming to create a series of decision trees that effectively separate signal from background. Once trained, the BDT can then be used to analyze real detector data (the blinded data) by evaluating the features of each event and predicting whether it belongs to the signal or background category. By employing a BDT, the researchers hope to achieve a more accurate separation of the rare Dark Photon signal events from the abundant background noise in the real detector data.

### 2.4.2 Preselection

BDT preselection is the process of applying a set of cuts to a dataset before training a Boosted Decision Tree (BDT). The goal of preselection is to reduce the amount of data that the BDT needs to process, which can improve the BDT's training speed and performance. Depending on displacement, could be electrons or photons. The specific cuts that are applied in BDT preselection will depend on the specific analysis. However, there are some general guidelines that can be followed. First, removing events that are unlikely to be signal events can be done by applying cuts on variables that are known to be different between signal and background events. For example, in the dark photon analysis, a cut could be applied on the electron time-of-flight to remove events with poorly reconstructed electrons. Next, reducing the dimensionality of the data can be done by removing variables that are not informative for the BDT. For example, if two variables are highly correlated, then only one of them may need to be included in the BDT training. [9]



**Figure 10:** Basic Structure of a Boosted Decision Tree [11]

The code snippet Eleanor provided in Table 1 defines a set of preselection cuts for a dark photon analysis at ATLAS. The preselection cuts are applied to both the signal and background data before the data is used to train the BDT. Here are some of the specific preselection cuts that are applied in the code snippet: The final state selection code maps different final states (e.g., "ee", "eg", "gg") to specific cuts on the number of electrons and photons in the event. I am selecting only events with two electrons "ee". This ensures that only events with the desired final state are considered for further analysis. The baseline cut code defines a set of baseline cuts that are applied to all final states. These cuts may include requirements on the electron time-of-flight  $< 12.5$  ns to remove events with poorly reconstructed electrons. Transverse momentum ( $p_T$ ) + trigger cuts selects high momentum electrons. Electron identification (ID) and isolation cuts ensures good quality electrons and reduces background from jets. Truth matching (signal MC only) ensures signal events correspond to generated particles. By applying preselection cuts, the analysis reduces the amount of data to be processed in subsequent stages, making the analysis more computationally efficient and focusing on events with a higher probability of containing the signal of interest. [9]

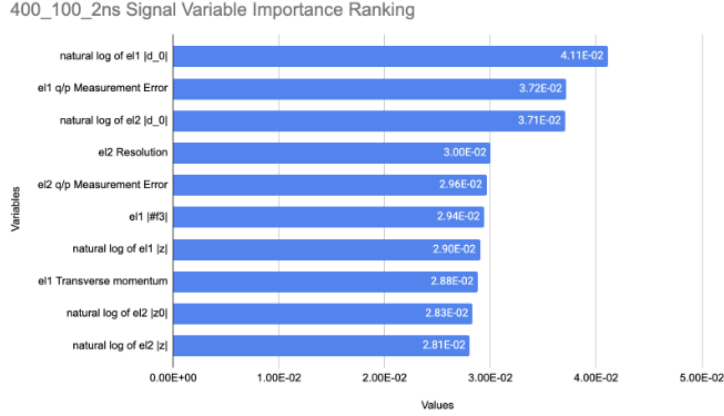
| Variable        | Requirement               |
|-----------------|---------------------------|
| $p_T$           | $> 10$ GeV                |
| eta             | $< 2.47$                  |
| ID              | $< \text{VeryLooseNoPix}$ |
| Crack Veto      | True                      |
| passOR          | True                      |
| nTrackParticles | $> 0$                     |
| track z0        | $< 500$                   |

**Table 1:** Electron Selection Criteria

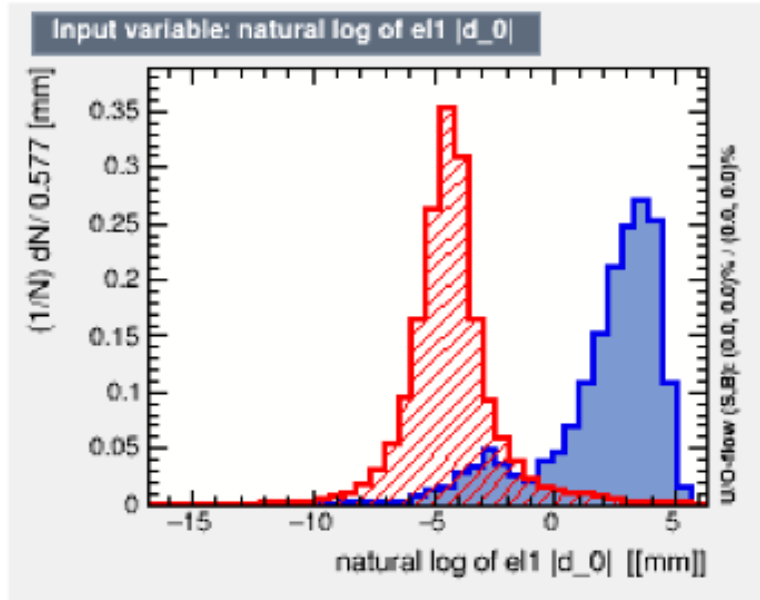
### 2.4.3 Input Variable Selection

Training the BDT with all available variables and then analyzing the feature importance scores can reveal which variables contribute most to the BDT's discrimination between signal and background events. Less important variables can then be removed. If two variables are highly correlated, then only one of them may be needed. Including both variables can increase the

computational cost of training the BDT without significantly improving its performance. We can also consider the physical separation of the signal process and how it might differ from background processes. Variables that are expected to be sensitive to these differences are more likely to be informative for the BDT. We began with a comprehensive list of approximately 80 input variables for the leading (first) and subleading (second most energetic) electrons. The "subleading" electron refers to the second most energetic electron in the event. Through this process, we reduced the initial variable pool to 46 for the final BDT models. To optimize the BDT's performance in different signal regions, we trained separate models for the 400-100-2ns signal, the 200-50-2ns signal, High mass signals, Low mass signals, High delta M signals, and Low delta M signals.



*Figure 11: Variable Importance Ranking for 400 GeV Signal*

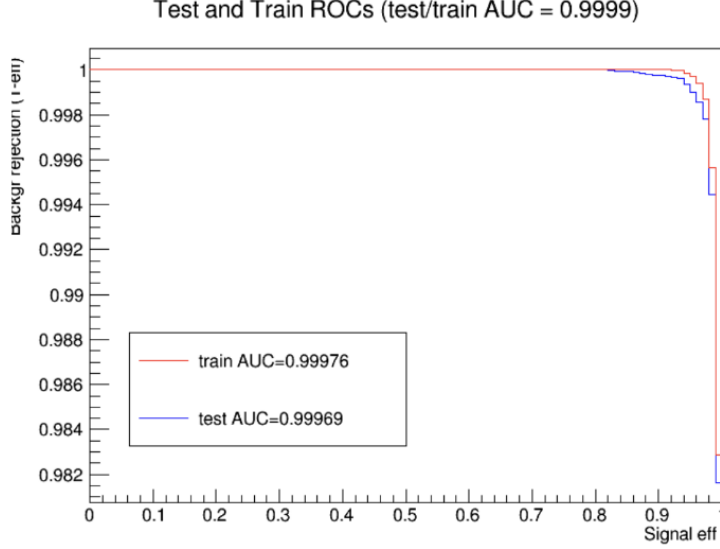


*Figure 12: Natural Log of Leading Electron Impact Parameter Distribution at 400 GeV [12]*

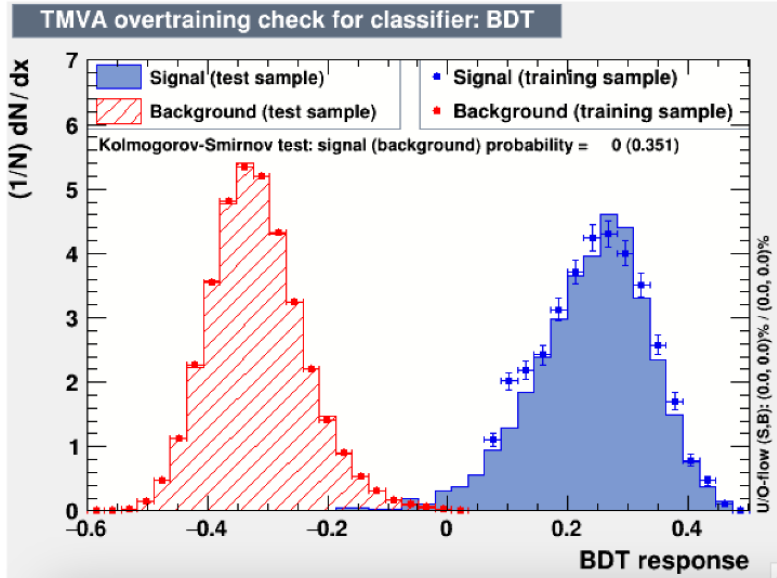
#### 2.4.4 Validation Procedure

Figure 14 provides a visual representation of the validation process for the 400, 100, 2ns signal using a Boosted Decision Tree (BDT) classifier. Receiver operating characteristic (ROC) curves are plotted for both the training and test samples. The Area Under the Curve (AUC) for the training sample is 0.99976, indicating excellent performance on the data used to train the model. The test AUC of 0.99969 is similarly high, suggesting good generalization of the model to unseen

data. The Kolmogorov-Smirnov test, with a probability of 0.351, implies that the distributions of the signal and background probabilities are not significantly different, further supporting the model's effectiveness in discriminating between the two classes. Performing a test for overtraining, a machine learning model becomes too specialized in the training data and performs poorly on new data, is important.



*Figure 13: Receiver Operating Characteristic (ROC) Curve for 400 GeV Signal*



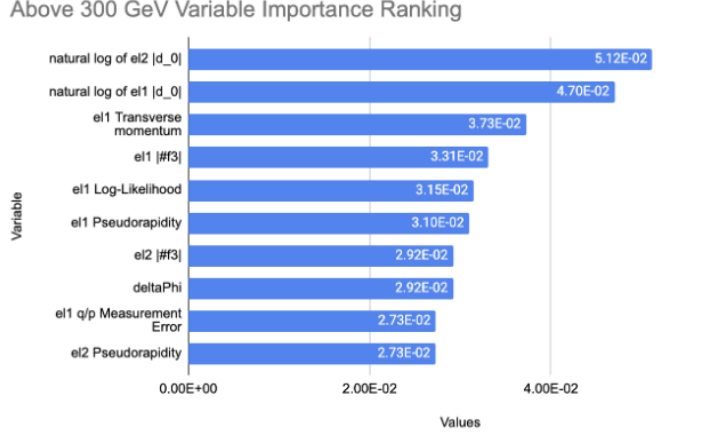
*Figure 14: BDT Response Distribution for 400 GeV Signal and Background [12]*

### 3 Results

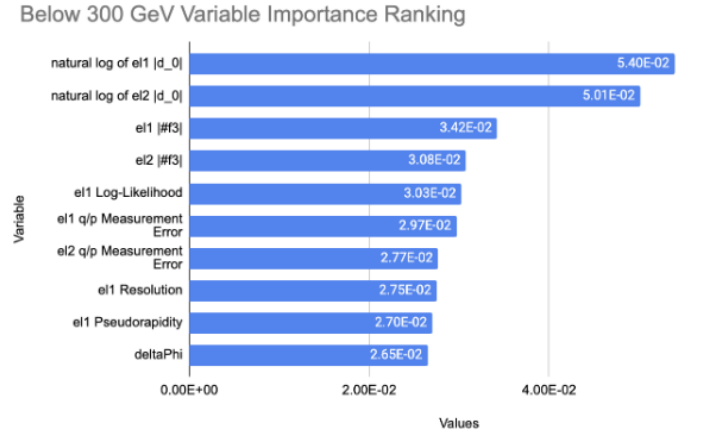
#### 3.1 BDT above 300 GeV vs. BDT below 300 GeV Comparison

The plots in Figure 15 and Figure 16 depict the variable importance ranking for signals below and above 300 GeV, respectively. Each plot shows a list of variables with their corresponding importance scores, represented by the length of the bars. The "d0" variable, which stands for

the impact parameter, is highlighted in both plots. It appears with a relatively high importance score in the "Below 300 GeV" plot, suggesting that it plays a significant role in discriminating between signal and background events in this energy range. In the "Above 300 GeV" plot, the importance of d0 seems to be slightly lower. This suggests that the impact parameter (d0) is a valuable discriminator for identifying signals below 300 GeV. However, its importance might diminish as we move to higher energy regions (above 300 GeV).



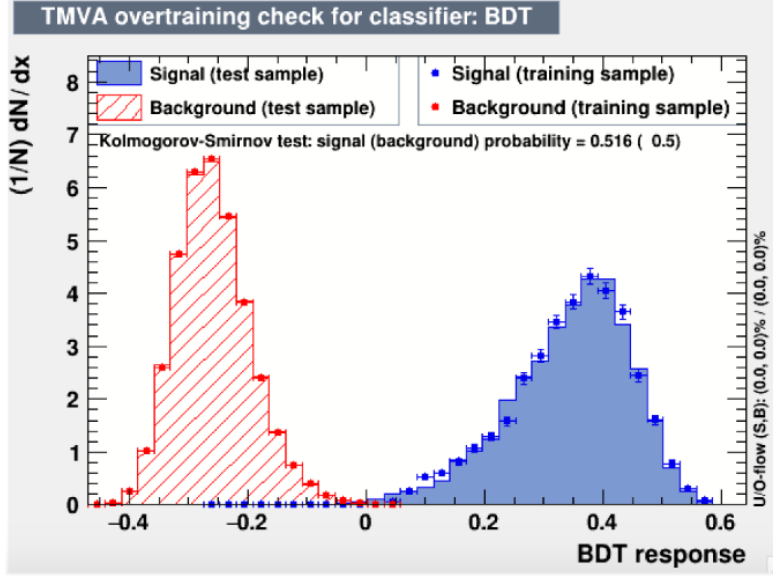
**Figure 15:** Variable Importance Ranking for Above 300 GeV Signals



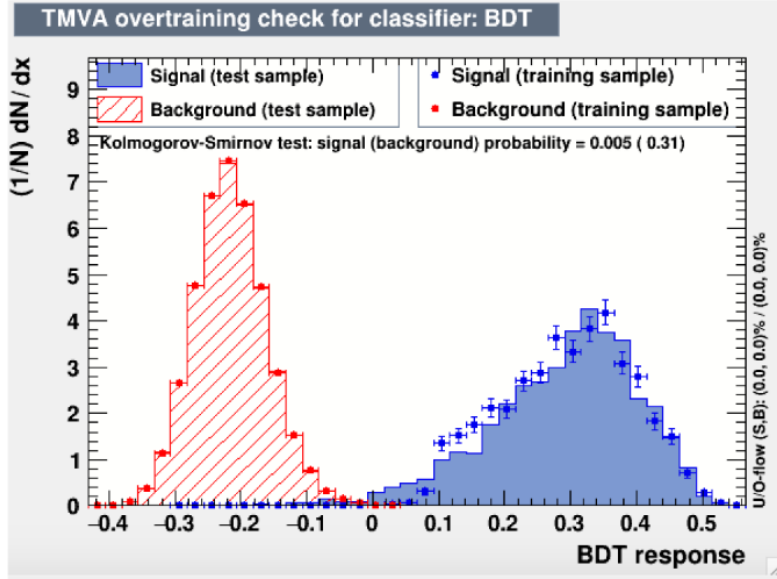
**Figure 16:** Variable Importance Ranking for Below 300 GeV Signals

Figure 17 and Figure 18 presents a comparison of signal and background distributions for a TMVA classifier, separated into "Below 300 GeV" and "Above 300 GeV" categories. The x-axis represents the BDT response, a value indicating the classifier's confidence in classifying an event as signal. The y-axis shows the normalized distribution of events. In both energy regions, the signal and background distributions exhibit some degree of separation. The signal tends to cluster towards higher BDT response values, while the background is more concentrated towards lower values. There is noticeable overlap between the signal and background distributions, especially in the "Above 300 GeV" region. This indicates that the classifier might struggle to accurately differentiate between signal and background events in this energy range. The Kolmogorov-Smirnov test probabilities suggest a better separation between signal and background for the "Below 300 GeV" category (0.005 for signal, 0.31 for background) compared to the "Above 300 GeV" category (0.516 for signal, 0.5 for background). The TMVA classifier demonstrates some effectiveness in separating signal and background events, particularly in the "Below 300 GeV" region. However, the overlap in the "Above 300 GeV" region highlights the challenges in discriminating between signal and background at higher energies. The performance of the classifier could be further

improved by optimizing the training process or incorporating additional variables. It's essential to consider the statistical significance of the observed separation when interpreting the results.



*Figure 17: BDT Response Distribution for Above 300 GeV Signals [12]*

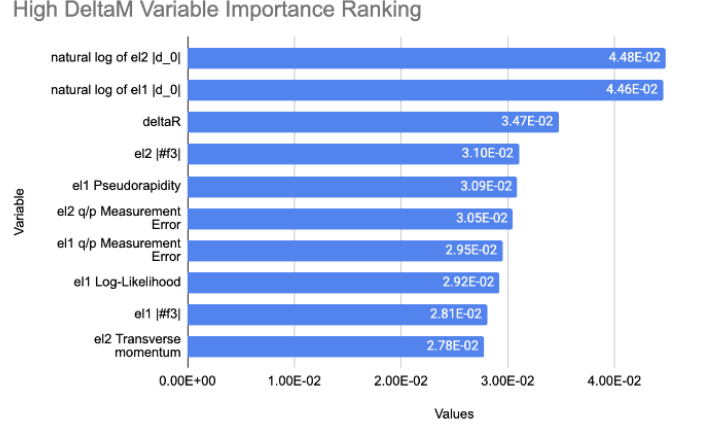


*Figure 18: BDT Response Distribution for Below 300 GeV Signals [12]*

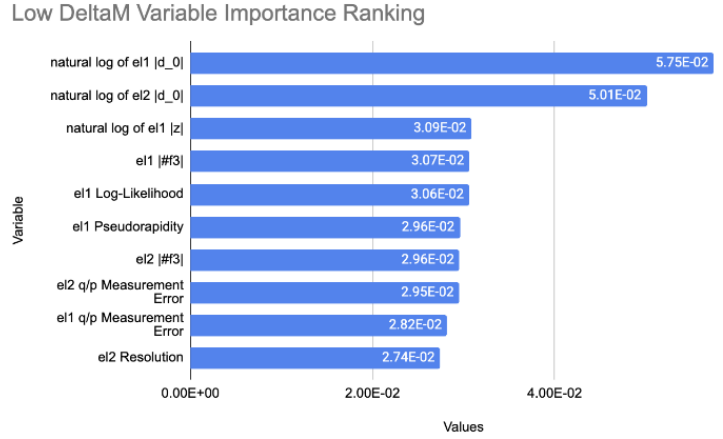
### 3.2 BDT with high delta M vs. BDT with low delta M Comparison

Figure 19 and Figure 20 presents a comparison of variable importance rankings and input variable distributions for low and high DeltaM signals. Figure 19 shows the variable importance ranking, with the most important variables at the top for both low and high DeltaM categories. Figure 20 displays the distributions of the top four variables for each category. The most important variables differ between low and high DeltaM signals. For low DeltaM, variables related to electron identification and energy measurement are prominent, while for high DeltaM, variables associated with electron pseudorapidity and isolation take precedence. The distributions of the top variables exhibit distinct shapes between the two DeltaM categories. This suggests that

these variables can effectively differentiate between low and high DeltaM signals. The analysis highlights the importance of considering DeltaM when selecting relevant input variables for signal discrimination. The observed differences in variable importance and distributions emphasize the need for tailored approaches to distinguish between low and high DeltaM signals. It would be beneficial to investigate the correlation between the top variables and DeltaM to gain further insights into their relationship. The impact of these variables on the classification performance should be evaluated using a suitable machine learning model.



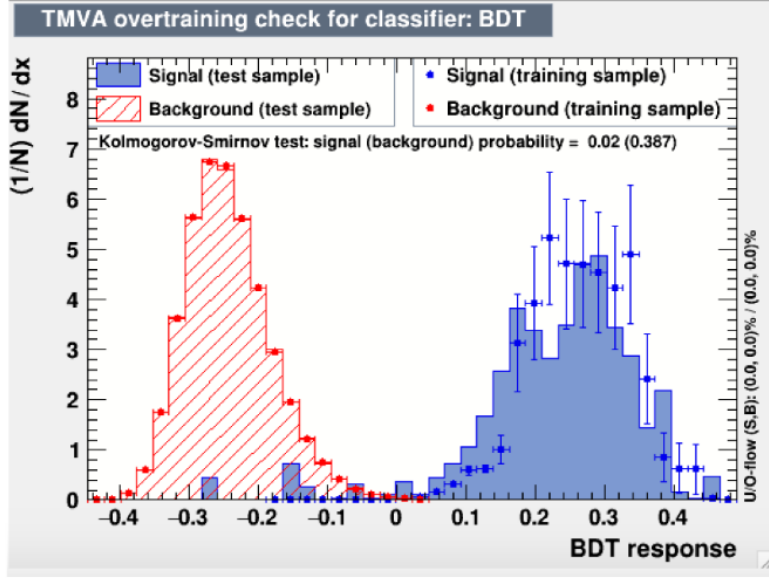
**Figure 19:** Variable Importance Ranking for High DeltaM Ranking



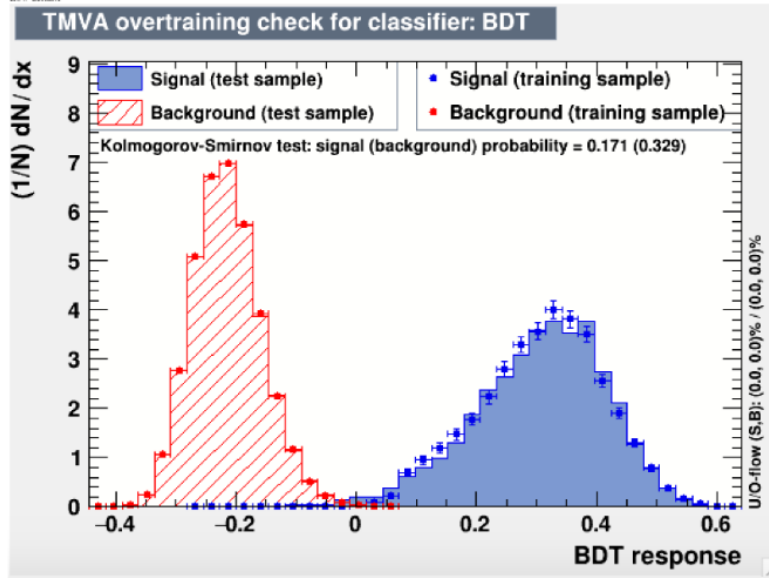
**Figure 20:** Variable Importance Ranking for Low DeltaM Ranking

Figure 21 and Figure 22 presents a comparison of two signal regions, Low DeltaM and High DeltaM, as classified by a Boosted Decision Tree (BDT) algorithm. The plots visualize the distribution of the BDT response for both signal and background events within each region.

The Low DeltaM signal exhibits a distinct separation between signal (blue) and background (red) events in the BDT response distribution. The Kolmogorov-Smirnov test for the signal yields a probability of 0.171, indicating a moderate level of agreement between the signal and background distributions. The Kolmogorov-Smirnov test for the background results in a probability of 0.329, suggesting a higher level of agreement between the background distributions. The High DeltaM signal shows a less pronounced separation between signal and background events compared to the Low DeltaM region. The Kolmogorov-Smirnov test for the signal yields a probability of 0.02, indicating a strong disagreement between the signal and background distributions. The Kolmogorov-Smirnov test for the background results in a probability of 0.387, suggesting a higher level of agreement between the background distributions. The Low DeltaM region demonstrates a clearer separation between signal and background events based on the BDT response. This



*Figure 21: BDT Response Distribution for High DeltaM Signals [12]*



*Figure 22: BDT Response Distribution for Low DeltaM Signals [12]*

suggests that the BDT is more effective in discriminating signal from background in this region. In contrast, the High DeltaM region shows a less distinct separation, indicating a potential challenge in distinguishing signal from background events using the BDT.

## 4 Summary and Conclusions

The analysis involves predicting background events to assess signal model sensitivity. We aim to compute expected signal-to-background ratios to evaluate model performance. Key electron properties considered include basic measurements like distance from the beamline ( $d_0$ ), longitudinal position ( $z$ ), pseudorapidity ( $\eta$ ),  $\phi$  coordinate, and transverse momentum ( $p_t$ ). Track quality metrics, such as associated track count and crossed detector layers, along with calorimeter isolation and other measurements like log-likelihood and momentum resolution, are incorporated. Derived quantities like the natural logarithm of  $d_0$  and  $z$ ,  $\Delta R$ , and  $\Delta \phi$  between electrons are also calculated. Importantly, the high deltaM signal exhibits distinct characteristics compared

to lower mass signals, with differing variable importance. This evidence suggests the potential need for specialized treatment of high deltaM models.

By examining the importance of electron variables, we've identified distinct characteristics in high deltaM signal models compared to others, suggesting the potential need for separate treatment. We've also streamlined the input variables for the BDT, prioritizing electron  $d_0$  for leading and subleading electrons, aligning with our search for displaced electrons. The next steps involve calculating expected signal and background ratios using the BDT, enhancing training data for the high deltaM signal, and determining the optimal approach: a universal BDT for all signals or a specialized one for high deltaM scenarios.

## 5 Acknowledgements

I would like to express my gratitude to the ATLAS Collaboration, especially the Columbia ATLAS group — John Parsons, Jonathan Long, Andrew Smith, Maria Bressan, and Eleanor Woodward — for their support, guidance, and expertise over the course of this project. To my fellow REU participants, it's been a pleasure getting to know you all these past 10 weeks. My sincere thanks to the entire team at Nevis Laboratories, particularly John Parsons, Reshmi Mukherjee, Georgia Karagiorgi, and Amy Garwood, for their invaluable support in making this REU program possible. This material is based upon work supported by the National Science Foundation under Grant No. PHY-2349438.

# References

- [1] The Standard Model. <https://home.cern/science/physics/standard-model>.
- [2] Department of Energy. DOE Explains...the Standard Model of Particle Physics. <https://www.energy.gov/science/doe-explainsthe-standard-model-particle-physics>.
- [3] The Large Hadron Collider. <https://home.cern/science/accelerators/large-hadron-collider>.
- [4] Detector and Technology. <https://atlas.cern/Discover/Detector>.
- [5] ATLAS detector slice (and particle visualisations). <https://cds.cern.ch/record/2770815>.
- [6] TMVA Users Guide. <https://root.cern.ch/download/doc/tmva/TMVAUsersGuide.pdf>.
- [7] ATLAS Collaboration. Search for Higgs-like scalar decaying into new spin-1 bosons in four-lepton final states at the ATLAS detector with  $139 \text{ fb}^{-1}$  at  $\sqrt{s} = 13 \text{ TeV}$ . *ATL-COM-PHYS-2022-260*. 2021. <https://cds.cern.ch/record/2807291/>.
- [8] ATLAS Collaboration. Search for a new scalar decays to beyond-the-Standard-Model light bosons in four-lepton final states with  $140 \text{ fb}^{-1}$  at  $\sqrt{s} = 13 \text{ TeV}$ . *ATL-COM-PHYS-2021-388*, 2021. <https://cds.cern.ch/record/2773055/export/hx?ln=en>.
- [9] Eleanor Luise Woodward. (Dark Photons BDT). [https://gitlab.cern.ch/ewoodwar/dark-photons-bdt/-/blob/master/README.md?ref\\_type=heads](https://gitlab.cern.ch/ewoodwar/dark-photons-bdt/-/blob/master/README.md?ref_type=heads).
- [10] Illuminating Dark Photons with High-Energy Colliders. *Journal of High Energy Physics: Springer Science and Business Media LLC*, [http://dx.doi.org/10.1007/JHEP02\(2015\)157](http://dx.doi.org/10.1007/JHEP02(2015)157), 10.1007/jhep02(2015)157, February 2015.
- [11] JavaTPoint. Decision tree classification algorithm. <https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm>.
- [12] Amelia Mei Stevens. (Updated Dark Photons BDT). <https://gitlab.cern.ch/amsteven/dark-photons-bdt/-/tree/master>.
- [13] Amelia Mei Stevens. (2 Electron Final State BDT Training). [https://www.nevis.columbia.edu/~amstevens/gg\\_BDT.html](https://www.nevis.columbia.edu/~amstevens/gg_BDT.html).
- [14] Eleanor Woodward. Vertexed Long-Lived Dark Photons Analysis: Signal Grid Request. April 18th, 2024. Columbia ATLAS Internal.

## 6 Appendix

### 6.1 BDT Input Variables

1. f3 - fraction of the electron energy deposited in the third layer of the calorimeter
2. pt - transverse momentum of electron (Transverse = in plane orthogonal to direction of the proton-proton beams)
3. phi - phi coordinate (azimuthal angle) of the electron trajectory
4. eta - pseudorapidity coordinate of electron
5. dz - resolution of calorimeter pointing measurement
6. nTracks - number of tracks associated to this electron
7. LHValue - continuous version of electron ID variable, corresponding to the log-likelihood of electron corresponding to a "true" electron
8. dpt - percent difference between pt of electron track and electron object
9. d0 - transverse impact parameter of the electron track
10. z - "pointing" of electron as measured using calorimeter deposits only
11. z0 - longitudinal impact parameter of electron track
12. qop err - error of electron charge over momentum measurement
13. chi2 - goodness of fit parameter for electron track from fit to the hits that form the track
14. nPIX - number of pixel layers crossed by the electron track
15. time - calibrated time of production of the electron as measured by the LAr calorimeter
16. charge - electron electric charge (+/- 1)
17. nSCT - number of layers crossed by the electron track in the Semiconductor Tracker
18. topcone20 - measurement of electron isolation in calorimeter: sum of transverse energy of topological calorimeter clusters within a cone of radius  $dR < 0.2$  of electron calorimeter cluster
19. E - calibrated electron energy measurement
20. deltaR - a measure of the Euclidean distance between two particles in a two-dimensional space defined by the pseudorapidity, and the azimuthal angle
21. deltaPhi - difference in azimuthal angle between two particles, which measures the angular separation of the particles in the plane transverse to the beam
22. deltaEta - difference in pseudorapidity between two particles, which measures the separation of the particles along the beam axis

## 6.2 Unused BDT Input Variables

1. etas1 - eta coordinate of calorimeter cluster in the first calorimeter layer
2. etas2 - eta coordinate of calorimeter cluster in second (middle) layer of the calorimeter
3. ID - electron identification discretized working point, where higher values roughly correspond to higher likelihood of corresponding to a "true" electron object
4. maxEcell E - energy deposited in the calorimeter cell that received the most energy from this electron
5. phis1 - phi coordinate of calorimeter cluster in the first calorimeter layer
6. phis2 - phi coordinate of calorimeter cluster in the second (middle) calorimeter layer
7. maxEcell t - (uncalibrated) time of electron measured by the calorimeter cell that received the most energy
8. maxEcell x - x coordinate of calorimeter cell that received the most energy
9. maxEcell y - y coordinate of calorimeter cell that received the most energy
10. maxEcell z - z coordinate of calorimeter cell that received the most energy
11. clusterE - electron energy measured in the calorimeter cluster
12. m - electron mass
13. nMissingLayers - number of Inner Detector layers not crossed by the electron track
14. qop - electron charge divided by momentum ( $q/p$ )
15. tracketa - eta of electron track
16. trackpt - transverse momentum of electron track momentum ( $q/p$ )
17. nSi - number of Silicon layers crossed by the electron track in the pixel detector