

Exploring Boosted Decision Trees for an ATLAS Search for Dark Mesons

Master thesis of 45 credits

Eva Mayer

Supervisors: Olga Sunneborn Gudnadottir, Rebeca Gonzalez Suarez

July 2023



Acknowledgements

I want to thank everyone who has contributed to the success of this thesis.

First and foremost I want to thank my supervisor Olga Sunneborn Gudnadottir for her always immediate, hands-on help with any problem I faced, and for her calm style of supervision that both gave me the freedom to try and go my own path and the necessary guidance to still end up in the right place.

I also want to thank my supervisor Rebeca Gonzalez Suarez for contributing heavily with her expertise and experience that gave a broader perspective to the project, and for motivating and inspiring me with her never-ending enthusiasm and encouragement.

Another big thank you goes out to the whole dark meson analysis team that gave useful input to my work in our weekly meetings.

I want to thank everyone in the ATLAS group in Uppsala for providing a welcoming working environment and helpful discussions on my results. Special thanks go to Arnaud Ferrari, my subject reader, whose thoughtful and detailed comments improved this thesis.

Last but not least I want to thank my friends for all the fun, the lunch breaks, and the emotional support that kept me going even in the darkest winter, and my family for their continuous support over the years that made all of this possible.

Abstract

This project explores the usage of the machine learning method of boosted decision trees (BDTs) for the ATLAS search for dark mesons, which are a dark matter candidate.

As this search tries to extract a tiny signal from a huge and very similar background using a set of sensitive kinematic variables, the question whether a multivariate method could improve the sensitivity has been studied. For this purpose a BDT model has been trained and tested on simulated signal and background samples. For evaluation receiver operating characteristic (ROC) curves and precision recall curves (PRC) have been studied. As an estimation of whether or not the sensitivity could be improved, selections on the BDT discriminant have been applied and evaluated using the resulting significance.

This method has successfully been applied and the significances using one common selection for the whole parameter space mostly exceed the ones that are reached without the usage of BDTs. However, one common selection for the whole parameter space is not the ideal choice, as the signal distributions depend heavily on the free parameters of the theory.

Sammfattning

Standardmodellen för partikelfysik är en teori som beskriver de minsta byggstenarna som materia består av: elementarpartiklarna. Den beskriver också de grundläggande krafter som verkar mellan dessa partiklar. Det finns bara en känd grundläggande kraft i naturen som inte beskrivs av standardmodellen: gravitation, en attraktionskraft mellan massiva föremål.

Även om denna teori kan förutsäga många processer med mycket hög precision finns det vissa saker som den inte kan förklara, och en av dem är mörk materia.

Mörk materia är extra materia som är nödvändig för att förklara astronomiska observationer av gravitationseffekter, till exempel i galaxrotationer. Dessa observationer visar att det finns en mycket större mängd materia i universum än vad som är synligt och som kan beskrivas av standardmodellen.

En teoretisk modell som förklarar vad mörk materia består av utökar standardmodellen med en ny så kallad mörk sektor, som sällan interagerar med resten av de kända elementarpartiklarna. Denna teori förutsäger existensen av nya fundamentala objekt som kallas mörka mesoner.

För att testa denna teori utförs sökningar vid ATLAS-experimentet vid the Large Hadron Collider. The Large Hadron Collider är en ring med 27 km omkrets som accelererar protoner till mycket höga energier och låter dem kollidera vid någon av kollisionspunkterna som omges av detektorer som tillhör fyra olika experiment. Det största detektorsystemet är ATLAS-experimentet. Kollisionerna mellan protoner leder till att protonerna delas upp och deras beståndsdelar skapar nya partiklar, som lämnar spår och energiavlagringar i de olika detektorerna och kan rekonstrueras.

Om mörka mesoner existerar kan de produceras i de processer som följer på protonkollisionerna i ATLAS och sedan sönderfalla till standardmodellpartiklar som skulle detekteras i detektorsystemet.

Antalet sådana händelser skulle vara mycket lågt jämfört med andra processer som sker i kollisionerna, och de upptäckta signaturerna skulle vara mycket lika helt normala processer i standardmodellen. Detta gör sökandet väldigt svårt. För fall som dessa är det vanligt att använda maskininlärningstekniker, som bygger på datoralgoritmer som är tränade för att klassificera datamängder som är svåra att klassificera med endast kinematiska urval.

Detta projekt har tillämpat en maskininlärningsmetod som kallas Boosted Decision Trees (BDTs) på en sökning efter mörka mesoner vid ATLAS. Resultaten visar att BDTs är framgångsrika när det gäller att förbättra känsligheten i denna sökning.

Contents

1	Introduction	6
1.1	Standard Model of Particle Physics	6
1.2	Dark Matter	7
1.3	The ATLAS Experiment at the LHC	7
1.4	The ATLAS Search for Dark Mesons	8
1.5	Boosted Decision Trees	16
1.6	Methods of Evaluation	17
1.7	Previous Results	18
2	Method	19
2.1	Simulation	19
2.2	Setup	20
2.3	Inclusion of the Full Set of Backgrounds	21
2.4	Variable Choice	24
3	Results	25
3.1	Comparison of ROC and PRC Curves for Evaluation	25
3.2	Testing on Individual Signal Points	27
3.3	Applicability of the Model to a Larger Parameter Space	34
3.4	Maximising Signal Significance	37
3.4.1	Raw Event Numbers	40
3.5	Distributions	42
4	Conclusion	47
A	Code	50

1 Introduction

Large parts of this section are taken from the final report of my 15 credits project "Application of a Boosted Decision Tree to signals and backgrounds of the search for dark mesons at the ATLAS experiment" [1], as this thesis is a direct continuation of said project.

1.1 Standard Model of Particle Physics

The Standard Model (SM) of Particle Physics describes all currently known elementary particles and forces with the exception of gravity. There are two types of elementary particles: The fermions, which have spin-1/2, and the bosons, which have integer spin values of either 0 or 1.

Fermions, which are the matter particles, are divided into quarks and leptons. The interactions between the fermions are carried by the gauge bosons: The gluon is the carrier of the strong force, which acts only on quarks. The photon is the carrier of the electromagnetic force, which acts on quarks and charged leptons (electrons, muons, and taus). The W- and Z-bosons are the carriers of the weak force, which acts on all fermions.

Additionally there is a scalar boson, the Higgs boson. It appears as an excitation of the Higgs field, which generates the masses of the elementary particles through the Brout-Englert-Higgs mechanism.

A summary of the elementary particles and their properties is shown in Figure 1.

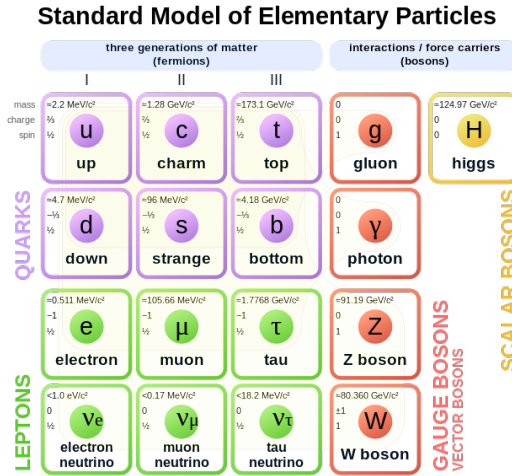


Figure 1: Overview of the elementary particles of the SM. Figure taken from [2].

1.2 Dark Matter

Even though the SM predicts a large range of physical processes with a high precision, there are some experimental features that are not explained by or that even contradict the SM. Among these are e.g. neutrino oscillations and the baryon-antibaryon asymmetry.

A major shortcoming is the dark matter (DM) problem. Astronomical observations differ from the expectations by our current knowledge of physics. Gravitational effects, like the velocity distributions in the rotation of galaxies and gravitational lense effects, suggest a significantly larger amount of matter in the universe than we observe through other channels. This could be explained if our understanding of gravitation is flawed or if there is matter in the universe that is not described by the SM.

Modified Newtonian Dynamics (MOND) theories take the first approach and modify Newton's law of gravitation for low acceleration. While being successful in explaining the behaviour of galaxies, these theories struggle to be compatible with general relativity and fail to describe the behaviour of galaxy clusters without postulating at least some amount of Dark Matter [3].

Another possible approach is to postulate particles that could make up DM and explain the gravitational effects as well as why they do not interact in any other visible way with the SM. The first suggestion of SM neutrinos, as they are only weakly interacting, was ruled out by cosmological models. Collider searches have been focusing on weakly interacting massive particles (WIMPs), especially as a part of Supersymmetry (SUSY). WIMPs as SUSY particles have been experimentally excluded since 2019 and the constraints on other variants of WIMPs grow stricter and stricter with results from various experimental searches piling up.

At collider experiments there is a growing interest in alternative theories that extend the SM by a dark sector. Stealth DM predicts fermions that are strongly coupled and transform under the electroweak group. Stable confined states of these particles provide a promising DM candidate that is not ruled out by current constraints, and their coupling with the Higgs boson together with weak interaction provides a window for detection. [4, 5]

1.3 The ATLAS Experiment at the LHC

The Large Hadron Collider (LHC) is the largest particle collider in the world, located at CERN in Geneva. The LHC is a synchrotron that accelerates protons and makes them collide at four interaction points with a center-of-mass energy of up to 13.6 TeV. Each interaction point is surrounded by a detector: ALICE, CMS, LHCb, and ATLAS.

The ATLAS experiment is the largest of these detectors. It is a multipurpose experiment consisting of an inner tracker surrounded by a solenoid, electromagnetic and hadronic calorimeters, and a muon spectrometer. An illustration of the detector is shown in Figure 2 and a full description of it can be found in [6].

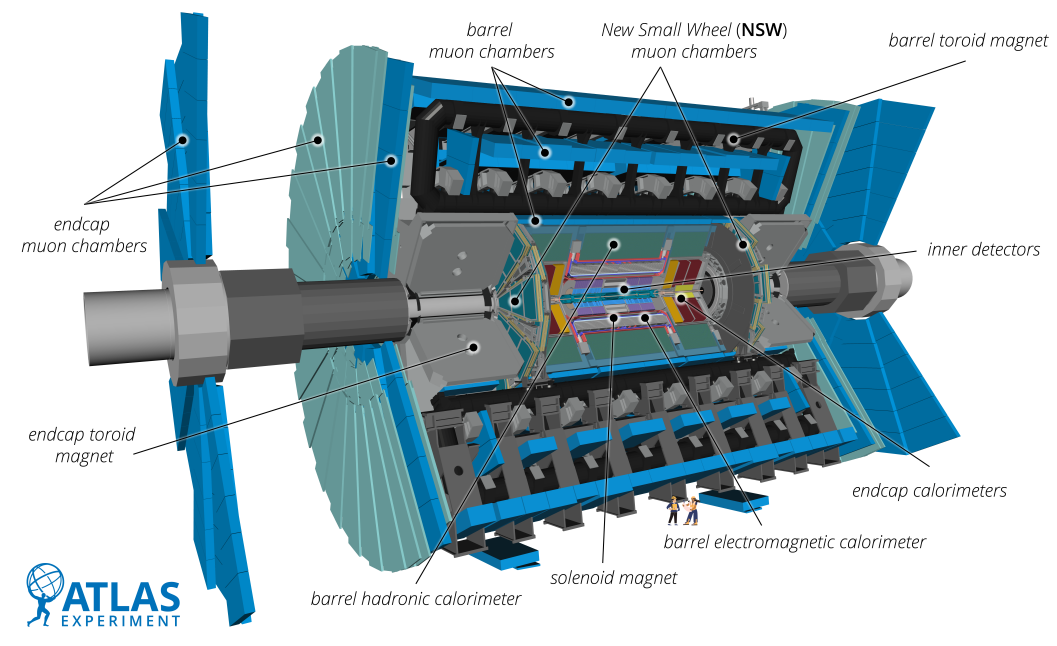


Figure 2: Schematic illustration of the ATLAS experiment. From [7].

1.4 The ATLAS Search for Dark Mesons

A Stealth DM theory (described in [5] and [8]) that extends the SM by a new, strongly-confined dark sector is currently tested at the ATLAS experiment by an ongoing data-analysis [9].

The theory predicts meson-like states of particles that are candidates for DM. They can decay into SM particles through interactions with the Higgs boson. The ATLAS data-analysis searches for pair production of dark pions through a resonant state of a dark rho, or alternatively through Drell-Yan production.

Dark pions and dark rhos are the meson states of this dark sector with the lowest (or second to lowest, respectively) mass. Therefore, the dark pion cannot decay into any other DM states and is the most likely to be observed to decay into SM particles.

The free parameters of this model are the mass of the dark pion m_{π_D} and the ratio of the

dark pion and the dark rho mass $\eta = \frac{m_{\pi_D}}{m_{\rho_D}}$. A third parameter, the number of generations, does not impact the phenomenology significantly and is for simplicity fixed at 4 for this search.

Two different ways of gauging the global flavor symmetry yield two distinct models: In the $SU(2)_L$ model, the whole $SU(2)$ global flavor symmetry is gauged, while in the $SU(2)_R$ only the $U(1)$ subgroup is gauged. In the ongoing analysis both models are considered.

In the $SU(2)_L$ model, the decay into SM particles can be gaugephilic, which means the decay products are predominantly gauge bosons, or gaugephobic, which means the decay products are exclusively fermions. The $SU(2)_R$ model on the other hand only allows for gaugephobic decays. The analysis at ATLAS is only studying the gaugephobic model.

For simplicity, antiparticles are denoted in the same way as particles in the following.

The highest branching ratio over the largest range of masses for the decay of charged dark pions into SM particles is the one into a top- and bottom-quark, as can be seen in Figure 3. For the neutral dark pions the highest branching ratio into SM particles is into two top-quarks over the largest range of masses, as can be seen in Figure 3.

Therefore the decays of the two dark pions considered in this work are the ones into two top- and two bottom-quarks ($t\bar{t}b\bar{b}$) and into one bottom-quark and three top-quarks ($t\bar{t}t\bar{t}b$), see Figure 4.

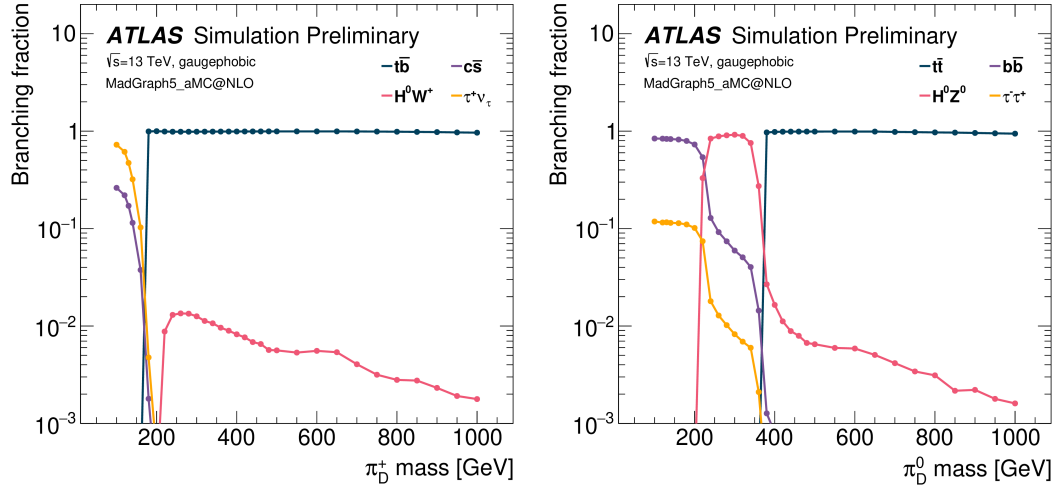


Figure 3: Branching ratios of the charged (left) and neutral (right) dark pion decays into SM particles. Taken from [9].



Figure 4: Production channels. Figure by Jochen Jens Heinrich.

The two main decay channels are the all-hadronic channel, which is when the final state consists solely of hadrons, and the 1-lepton channel, which is when the final state is composed of one charged lepton, one neutrino, and additionally only hadrons. This work is studying the 1-lepton channel, which contains the following two processes:

$$qq \rightarrow \pi_D^- \pi_D^+ \rightarrow t\bar{t}b\bar{b} \rightarrow WbWb\bar{b}\bar{b} \rightarrow \nu q q b \bar{b} \bar{b} \bar{b}$$

and:

$$qq \rightarrow \pi_D^0 \pi_D^+ \rightarrow t\bar{t}t\bar{b} \rightarrow WbWbWb\bar{b} \rightarrow \nu q q q b \bar{b} \bar{b} \bar{b}$$

The experimental signature then consists of either six or eight jets, at least four of which originating from b-quarks, exactly one charged lepton (either an electron or a muon, as the tau decays to hadrons are not considered), and missing transverse energy due to the neutrino(s).

The physics objects that are used to reconstruct these signatures are the charged lepton, either an electron or a muon, missing transverse energy, and jets. The jets are reconstructed using the anti- k_T algorithm [10] with a radius parameter $R=0.4$. Jets originating from b-quarks are b-tagged with a 77% efficiency working point. To indirectly extract properties of the dark mesons the jets with $R=0.4$ are reclustered as large- R jets with $R=0.8$ or $R=1.2$. A large- R jet containing the lepton is referred to as $\mathbb{J}^{lep}$ and a large- R jet consisting only of hadronic jets is referred to as $\mathbb{J}^{had}$.

The production of a top-antitop and a bottom-antibottom pair, $t\bar{t}b\bar{b}$, is the SM process most similar to the signal, and makes up the irreducible background in the analysis. Its experimental signature consists of exactly the same particles as the signal. However, the kinematics show differences that can be used for reducing the background.

To grasp these differences a set of kinematic variables has been explored:

- N_{jets} : The number of jets in the event, reconstructed and identified as described in section 3.3 of [11].
- N_{b-jets} : The number of jets identified as originating from a B-hadron using a 77% efficiency working point, following section 3.5 of [11].
- p_{Tl} : The transverse momentum of the lepton.
- H_T : The scalar sum of transverse momentum for jets in an event.
- p_{Tj_0} : The transverse momentum of the leading jet.
- $\Delta R(l, b1)$: ΔR^1 between the lepton and the leading b-tagged jet.
- $\Delta R(l, b2)$: ΔR between lepton and second closest b-tagged jet. Note that the second closest and the leading b-tagged jet could be the same.
- $\Delta R^{min}(b, b)$: The minimal ΔR between any two b-tagged jets.
- $m_{j^{lep}+b1, R=0.8}$: The invariant mass of the large-R jet containing the lepton and the leading b-tagged jet for $R=0.8$.
- $m_{j^{lep}+b1, R=1.2}$: The invariant mass of the large-R jet containing the lepton and the leading b-tagged jet for $R=1.2$.
- $m_{j^{lep}, R=0.8} + m_{j^{had}, R=0.8}$: The sum of the masses of the large-R jet containing the lepton and the fully hadronic large-R jet for $R=0.8$.
- $m_{j^{lep}, R=1.2} + m_{j^{had}, R=1.2}$: The sum of the masses of the large-R jet containing the lepton and the fully hadronic large-R jet for $R=1.2$.
- m_T : transverse mass $\sqrt{2 \cdot p_{Tl} \cdot E_T^{miss} (1 - \cos(\Delta\phi(\vec{l}, \vec{p}_T^{miss}))}$, where E_T^{miss} is the missing transverse energy, and $\Delta\phi(\vec{l}, \vec{p}_T^{miss})$ is the angle between the transverse momentum of the charged lepton and the missing transverse momentum.
- $m_{j_0^{had}, R=0.8}$: Mass of the leading large-R jet containing the lepton with $R=0.8$.
- $m_{j_0^{had}, R=1.2}$: Mass of the leading large-R jet containing the lepton with $R=1.2$.
- E_T^{miss} : The missing transverse energy.

¹ATLAS uses a right-handed coordinate system with its origin at the nominal interaction point (IP) in the centre of the detector and the z -axis along the beam pipe. The x -axis points from the IP to the centre of the LHC ring, and the y -axis points upwards. Cylindrical coordinates (r, ϕ) are used in the transverse plane, ϕ being the azimuthal angle around the z -axis. The pseudorapidity is defined in terms of the polar angle θ as $\eta = -\ln \tan(\theta/2)$. Angular distance is measured in units of $\Delta R \equiv \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$.

- $E_{T,Sig}^{miss}$: The missing transverse energy normalized by the square root of H_T .
- m_{bb}^{min} : The smallest invariant mass of any two b-tagged jets.
- $m_{bb\Delta R_{min}}$: The invariant mass of the two b-tagged jets closest to each other.

For further details about the variables, see [11].

Previous constraints on the dark pion mass for the different theoretical models and by different experimental data-analyses can be seen in Figure 5. The expected sensitive region of the 1-lepton channel from the current data-analysis without the usage of BDTs is shown in Figure 6. The current data-analysis extracts the signal through selections on some of the most sensitive variables and then performs a fit on $m_{\mathbb{J}^{lep},R=1,2} + m_{\mathbb{J}^{had},R=1,2}$.

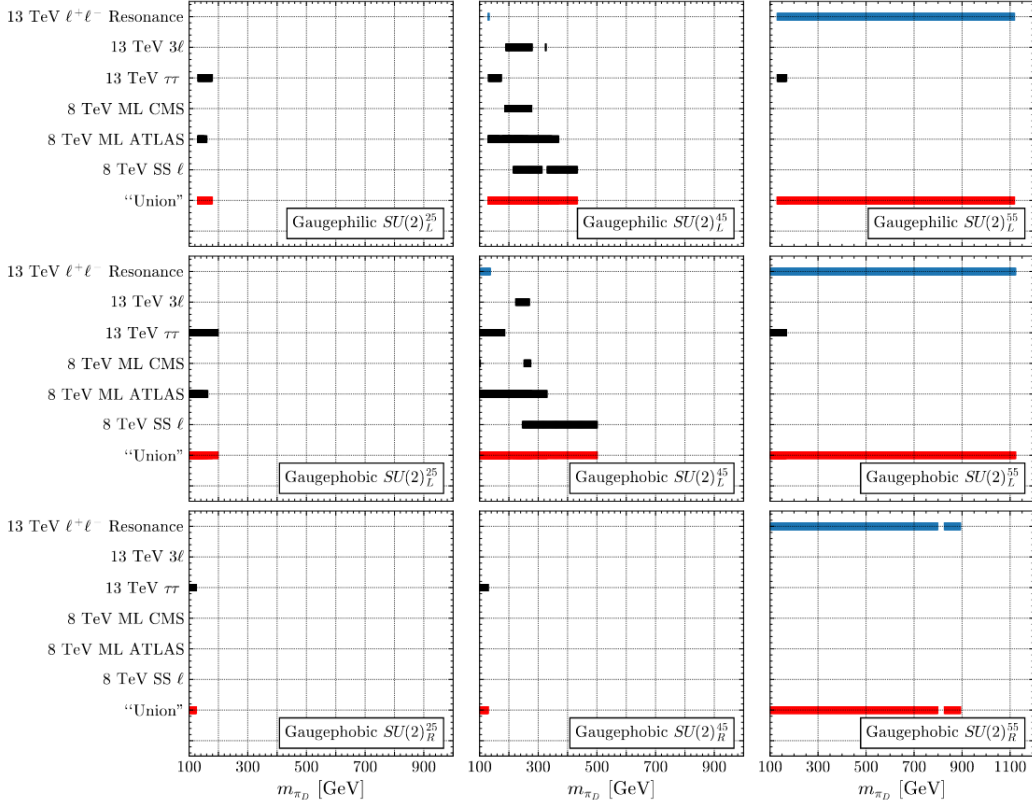


Figure 5: Constraints on the dark pion masses for the $SU(2)_L$ model in the gaugephilic and gaugephobic theory and for the $SU(2)_R$ model only for the gaugephobic theory. The parameter η takes the values 0.25, 0.45, and 0.55. From [11].

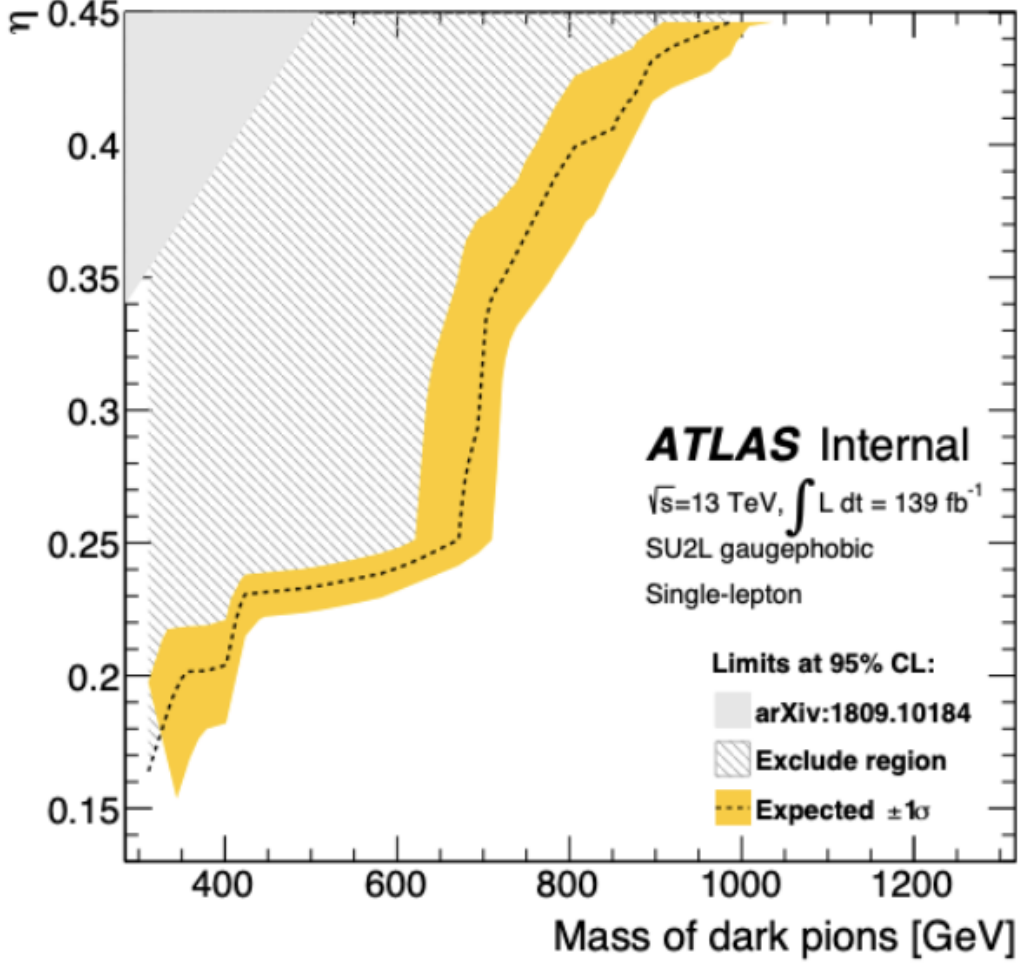


Figure 6: Sensitive region of the 1-lepton channel of the ATLAS dark meson analysis in the $SU(2)_L$ model. Figure by Olga Sunneborn Gudnadottir.

The expected signal cross sections for the Run-2 center of mass energy of 13 TeV are shown in Figure 7. The upper left regions with high η and low m_{π_D} values have the largest cross sections, and the cross section decreases with increasing m_{π_D} and decreasing η . The cross sections of the $SU(2)_L$ model are significantly higher than the ones of the $SU(2)_R$ model, which is the reason why more attention was given to the $SU(2)_L$ model in this project.

The search is using the full Run-2 data, of integrated luminosity 139 fb^{-1} .

As a first step the analysis makes a broad selection to remove as much of the background

Table 1: Pre-selection of the events used for this project.

Number of leptons	==1
Number of additional loose leptons	==0
Number of jets	>= 5
Number of b-tagged jets	>= 3
H_T >= 300 GeV	

as possible while keeping most of the signal. This selection is called preselection. The preselection used in this project is the one of the nominal dark meson analysis [11] for the 1-lepton channel, shown in Table 1. The number of leptons is set to 1 with no additional loose leptons, as vetoing additional leptons removes background. Furthermore, the number of jets is set to at least five, and the number of b-tagged jets to least three. As a first kinematic requirement the value of H_T is set to at least 300 GeV.

This project will explore if extra sensitivity can be gained by using multivariate methods.

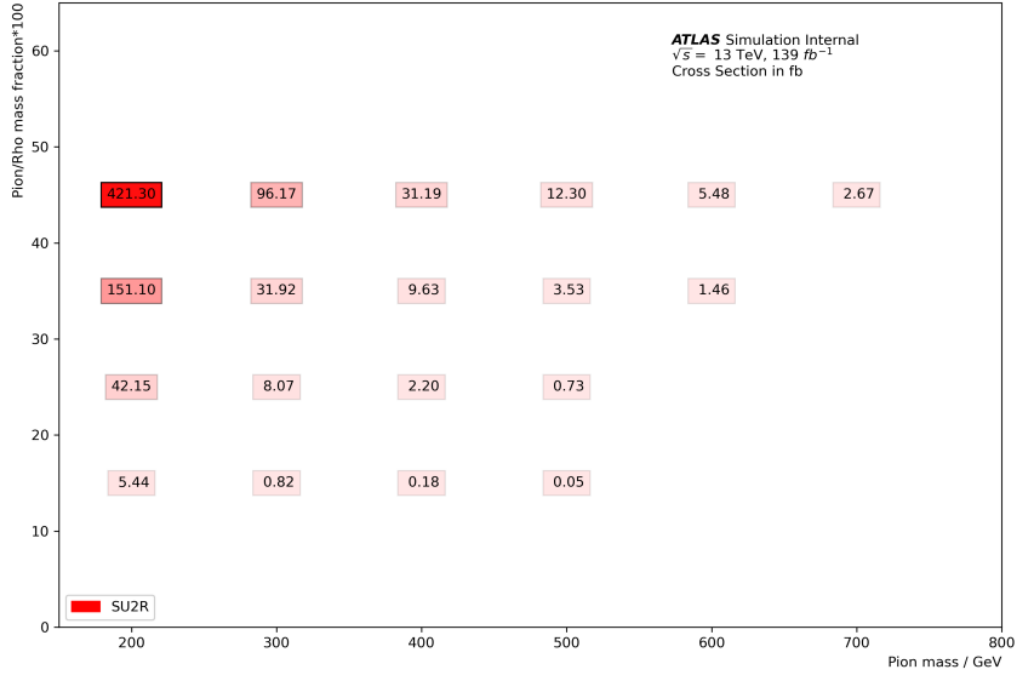
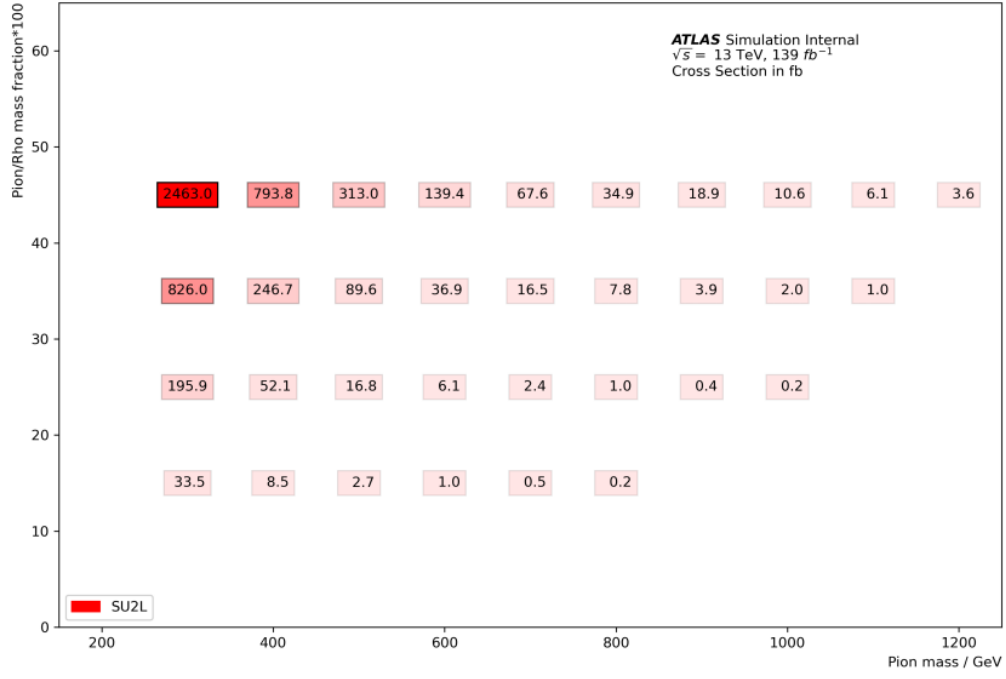


Figure 7: Signal cross sections of the $SU(2)_L$ and $SU(2)_R$ model. From [11].

1.5 Boosted Decision Trees

Algorithms that are built to learn from datasets and subsequently build a model to make predictions for similar datasets are called machine learning (ML) algorithms. They are predominantly used in cases where correlations and patterns in datasets are too complex for humans to understand and describe them with a simple model.

In high energy physics the use of ML methods is becoming increasingly popular, especially in searches with tiny signals and a large background that are very similar to each other.

Decision Tree learning is a ML method that is based on a tree-shaped system of rules. A set of training data that are labelled with different categories is analysed with respect to a set of features. In this project the categories are "signal" and "background" and the features are the variables describing the kinematics and the invariant masses in simulated events (see section 1.4).

A model is created by choosing and applying a threshold on a feature to discriminate between the categories in the best possible way. The thereby separated data are then split up further by applying more thresholds on more features, which creates a tree-shaped structure of decisions. A prediction score is assigned to the end of every branch to indicate the strength of a tendency towards either one of the two categories. The number of times a threshold is applied along the tree is called the depth of a tree. Boosted Decision Trees (BDTs) are a variation of this method where many Decision Trees of small depth are combined. After one decision tree is developed the data are reweighted according to how well they are classified by the model. A second decision tree is trained based on the reweighted data, and the procedure is repeated until a sufficiently large number of trees has been developed. The prediction scores of all trees are added and normalised to a number between 0 and 1.

This final prediction score, called discriminant, is the main output of the model. If the discriminant for an event is closer to 0 or closer to 1 indicates the confidence of the prediction towards either one of the two categories. A binary decision of classifying data into the two categories can be made by making a selection based on a threshold on the discriminant.

The structure of the BDT method is shown in Figure 8.

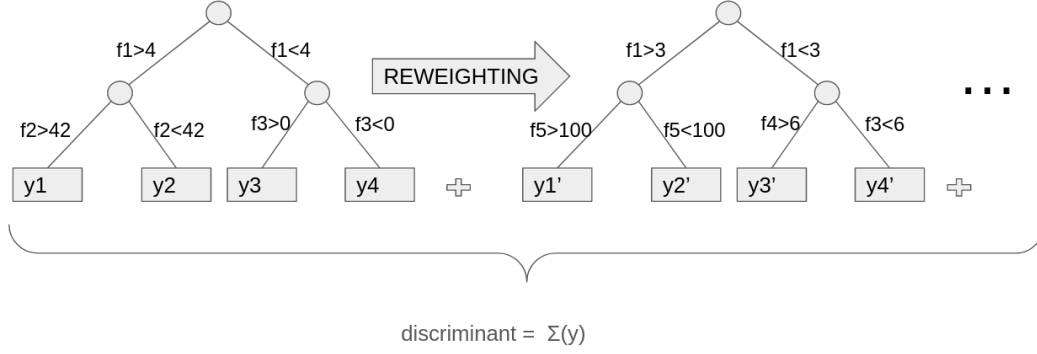


Figure 8: Schematic image of the working principle of a BDT. Source: own work.

The ensemble nature of the BDTs makes them resilient to the risk of strong bias, which is specifically useful for high energy physics analyses with extremely low signal event numbers compared to the background yields. In addition, they are by far more transparent and less complex than convolutional neural networks (CNNs) which are another machine learning technique that is commonly used in high energy physics. [12, 13, 14]

As the signal extraction of the dark meson analysis is challenged by the sparsity of the signal in a huge and very similar background, it is the ideal use case for BDTs. They have been used before in similar applications and therefore there is a lot of material and software specialized on signal extraction in high energy physics, which makes them easy to implement.

1.6 Methods of Evaluation

For evaluating the performance of the BDT in this analysis, two types of curves are used: the receiver operating characteristic (ROC) curve, and the precision recall curve (PRC).

In the ROC curve the false-positive rate is plotted against the true-positive rate (also called efficiency) for all the possible selections based on the discriminant that can be used for classifying the data. The false-positive rate is calculated by dividing the number of background events that were falsely classified as signal through the number of true background events. The true-positive rate is calculated by dividing the number of correctly classified signal events through the number of true signal events. Good performance corresponds to a high true-positive rate and a low false-positive rate.

The integral (the area under the curve, also called AUC) is a measure of the performance. It is a number between 0 and 1, where everything between 0 and 0.5 suggests that the BDT will tend to make a wrong decision more often than a correct one. An AUC value of 0.5

means that the BDT is just as powerful as a random classifier. If the AUC is between 0.5 and 1 the BDT has some discriminating power and the closer it is to 1 the better.

The weakness of this curve as a method of evaluation, however, comes from imbalanced datasets. If the background sample is by far larger than the signal sample, the significance of the relation between the false-positive rate and the true-positive rate decreases. Indeed even if a very large part of the true signals is classified as the signal and a very small part of the background is classified as signal, the ratio of signal to background in the signal class can be very low.

The second commonly used curve for evaluation is the precision recall curve (PRC). In this curve the x axis is the efficiency, which is another term for the true-positive rate. The y axis shows the purity, which is the ratio of signal to background in the signal class. This eliminates the weakness of the ROC curve, but the drawback here is that the curve and the AUC of this curve are less intuitive in the interpretation. The no-skill curve in this case would be a constant with the value of the total ratio between signal and background events as the purity. The AUC of this line would be interpreted as the AUC of 0.5 for a ROC, and everything above this value shows that the BDT classifier has some level of useful skill. The AUC scores of different curves cannot necessarily directly be compared, as the no-skill curve and therefore the no-skill AUC score might vary from case to case. They can only be compared if the ratio of signal events to background events in the datasets are the same.

1.7 Previous Results

This work builds on the results of a 15 credits project previously undertaken by the author on the same topic. The results from that project [1] show that the BDT is able to distinguish well between the signal and the ttbb background. From the three tested signal samples, SU2L-45-400, SU2L-25-800, and SU2R-45-300², the one with the highest dark pion mass is the easiest to discriminate from the background, while the one with the lowest dark pion mass is the hardest to discriminate. This was expected, as the distributions of the variables for low values of m_{π_D} resemble the ones of the background more closely.

An importance ranking and a correlation matrix were produced to study which variables were most and least relevant. Based on these findings the number of variables used was reduced by five. The performance of the BDT barely decreased, which shows that the chances are high that the number of variables can be decreased even more.

²The notation of the signal samples includes the following information: if it belongs to the $SU(2)_L$ or the $SU(2)_R$ model, the value of η multiplied by 10, and the value of m_{π_D} in GeV. E.g. SU2L-45-400 is a sample from the $SU(2)_L$ model with $\eta=0.45$ and $m_{\pi_D}=400$ GeV.

2 Method

2.1 Simulation

The signal samples used are simulated using MADGRAPH5_aMC@NLO v2.4.3 [15]. The matrix elements are calculated to next to leading order and interfaced with PYTHIA8.212 [16] using the NNPDF2.3lo set of parton distribution functions (PDF) and the A3 set of tuned parameters. They are described in further detail in [11].

Figure 9 shows the complete collection of signal samples that have been simulated with different numbers of events.

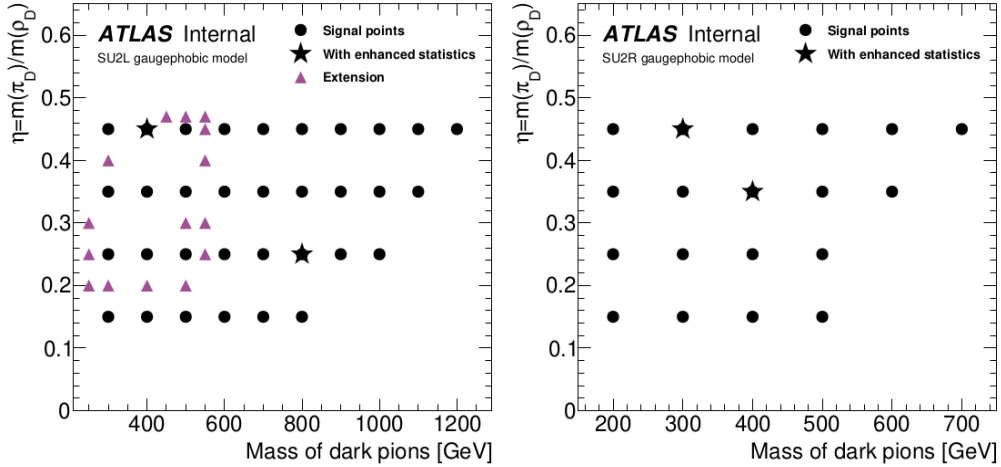


Figure 9: Distribution of simulated signal samples with different values of η and m_{π_D} . The left plot shows the samples for the $SU(2)_L$ model and the right one the samples for the $SU(2)_R$ model. Original signal samples (black dots) consist of ten thousand events, samples of the extension for the all-hadronic channel (purple triangles) of twenty thousand events, and samples with enhanced statistics of a hundred thousand events. This project does not use the purple samples. From [11].

The points in this parameter space at which samples have been simulated are referred to as signal points. For both models two signal points were produced with enhanced statistics. One point was chosen to be in the high-eta and low- m_{π_D} region, while the other one is in the low- η and high- m_{π_D} region, in order to cover different ends of the spectrum of different signatures. The signal points with enhanced statistics are:

- "SU2L-45-400": $SU(2)_L$ model, $\eta = 0.45$, $m_{\pi_D} = 400$ GeV
- "SU2L-25-800": $SU(2)_L$ model, $\eta = 0.25$, $m_{\pi_D} = 800$ GeV

- "SU2R-45-300": $SU(2)_R$ model, $\eta = 0.45$, $m_{\pi_D} = 300$ GeV
- "SU2R-35-400": $SU(2)_R$ model, $\eta = 0.35$, $m_{\pi_D} = 400$ GeV

The background samples are composed of $t\bar{t}$ production in association with various other particles, $t\bar{t}t\bar{t}$ production, single- t production, production of a vector boson in association with jets, and the production of multiple bosons. The background components were individually produced using different generators [17, 18, 15]. A summary of the background samples with their generators and configurations is shown in Table 2. They are also described in great detail in [11].

Table 2: Overview of the configuration of all nominal background samples used in the analysis. Overlaps between the samples were removed. Taken from [11].

Process	Generator	PDF	Showering	Tune	Cross section
$t\bar{t}$	POWHEGBOX v2	NNPDF3.0NLO	PYTHIA8	A14	NNLO+NNLL
$t\bar{t}$ +HF	POWHEG BOX RES	NNPDF3.0NLO	PYTHIA8	A14	NNLO
V + jets	SHERPA v2.2.11	NNPDF3.0NNLO	SHERPA	Def.	NLO
Single top	POWHEGBOX v2	NNPDF3.0NLO	PYTHIA8	A14	NLO+NNLL
$t\bar{t}t\bar{t}$	MADGRAPH5_aMC@NLO v2.4.3	NNPDF3.1NLO	PYTHIA8	A14	NLO
$t\bar{t}V$	MADGRAPH5_aMC@NLO v2.3.3	NNPDF3.0NLO	PYTHIA8	A14	NLO
$t\bar{t}H$	POWHEGBOX v2	NNPDF3.0NLO	PYTHIA8	A14	NLO
Other $t\bar{t}$ + X	MADGRAPH5_aMC@NLO	NNPDF2.3LO	PYTHIA8	A14	NLO
Multiboson	SHERPA v2.2.1/v2.2.2	NNPDF3.0NNLO	SHERPA	Def.	NLO

For the final results these samples are normalised to the target luminosity, which is the full Run-2 luminosity. Not all figures and results are normalised to the target luminosity, each figure will have a note specifying if it is the case.

In this kind of analysis, weights are typically added to the simulated samples in different ways. This is done to estimate the contributions of different processes accurately, to reproduce experimental conditions, or to account for differences in reconstruction efficiencies between data and simulation among other purposes. Weights are not considered in the training, while they are added later in the evaluation. Each result will also have a note specifying if this is the case.

2.2 Setup

All computations were done on the CERN LXPLUS computer cluster. The software was written in Python [19], using the ROOT [20] framework and the libraries Matplotlib [21],

Numpy [22], and Pandas [23]. For the training and testing of BDTs the XGBoost package [24] was used, and for the evaluation and visualisation the scikit-learn [25] package was employed.

The complete code can be found in [26] and some extracts are shown in the appendix A.

The work is embedded in the dark meson analysis and the status and results of the project were regularly presented and discussed in the framework of this group.

2.3 Inclusion of the Full Set of Backgrounds

In the previous project only the sample of the main irreducible background ttbb had been included. Inspired by [27] the training of two consecutive BDTs was considered.

If the kinematics of the main background and the signal would be more similar to each other than the kinematics of the main background and the other backgrounds, it could be fruitful to first train a BDT to classify the main background versus the other backgrounds and then define a signal region based on the results. In the signal region a second BDT could be used in order to classify the signal versus the main background.

To explore this option a BDT was trained to distinguish between ttbb and the other background samples. The results that can be seen in Figures 10 and 11, however, showed that the power of discriminating ttbb from the other backgrounds was worse than for the BDTs that had been trained to discriminate the signal from ttbb. Therefore in the end just one BDT was trained to discriminate the signal from the full set of backgrounds, including ttbb.

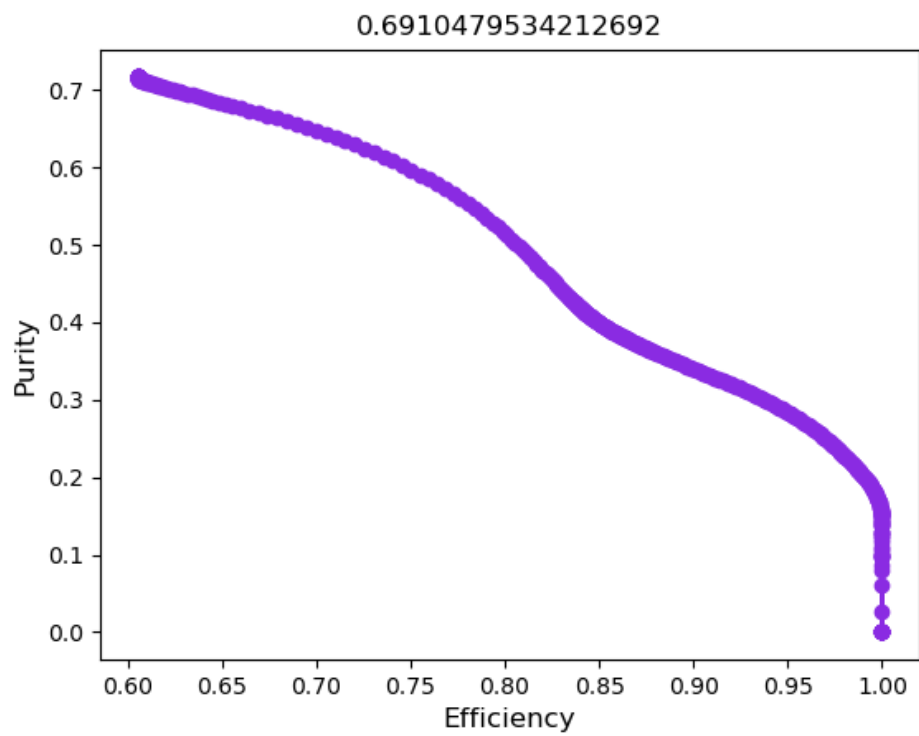


Figure 10: ROC curve with AUC score of a BDT trained on discriminating the $t\bar{t}b\bar{b}$ sample from the other background samples.

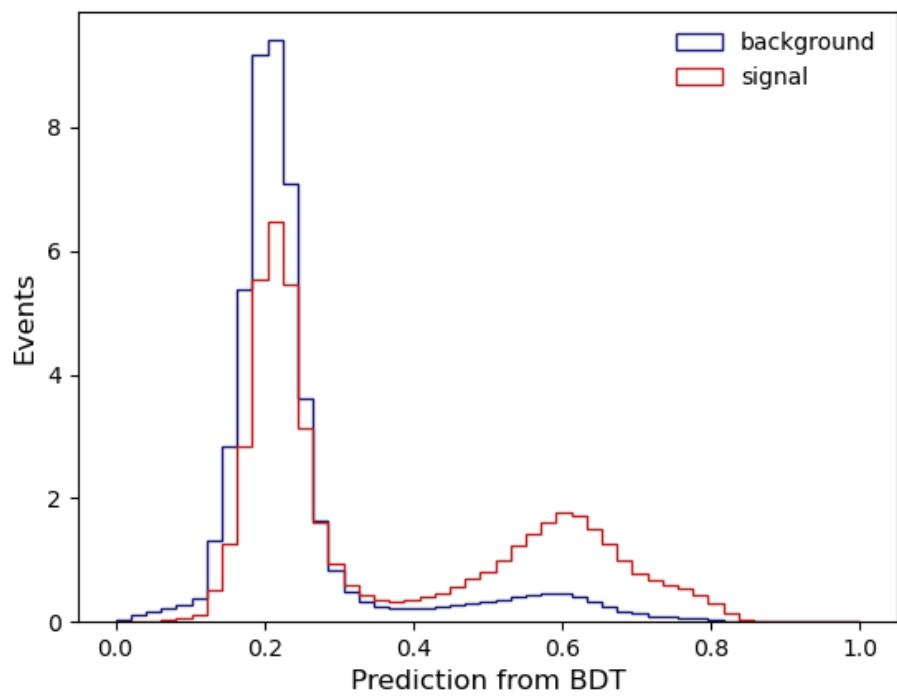


Figure 11: Histogram of the discriminant for the $t\bar{t}b\bar{b}$ sample ("signal") and the set of other backgrounds ("background"). Note: the two overlaid histograms are normalized to the same area. Neither weights nor the target luminosity are considered.

2.4 Variable Choice

With more backgrounds included the memory usage of the training and testing was too high for the available computing resources. This problem was solved by reducing the number of variables. The previous project [1] already showed that many of the variables are strongly correlated and that the number could be reduced without impacting the performance too much.

Based on the studies done in [1] the following eight variables were chosen:

- N_{Jets} : The number of jets
- N_{b-jets} : The number of b-tagged jets
- p_{Tl} : The transverse momentum of the lepton
- H_T : The scalar sum of the transverse momenta of the jets
- $\Delta R(l, b_1)$: ΔR between the lepton and the leading b-tagged jet
- $m_{j^{lep}, R=1.2} + m_{j^{had}, R=1.2}$: The sum of the masses of the large-R jet containing the lepton and the fully hadronic large-R jet for $R=1.2$
- m_T : The transverse mass $\sqrt{2 \cdot p_{Tl} \cdot E_T^{miss} (1 - \cos(\Delta\phi(\vec{l}, \vec{p}_T^{miss}))}$, where E_T^{miss} is the missing transverse energy, and $\Delta\phi(\vec{l}, \vec{p}_T^{miss})$ is the angle between the transverse momentum of the charged lepton and the missing transverse momentum.
- $m_{bb\Delta R_{min}}$: The invariant mass of the two b-tagged jets closest to each other

3 Results

3.1 Comparison of ROC and PRC Curves for Evaluation

In this project both ROC and PRC curves, as explained in section 1.6, were produced. In many cases the two curves paint a drastically different picture.

In Figures 12 and 13 both curves are depicted for the same signal sample trained on the same model. While the ROC curve indicates very good performance, the PRC is just slightly better than it would be for a random classifier.

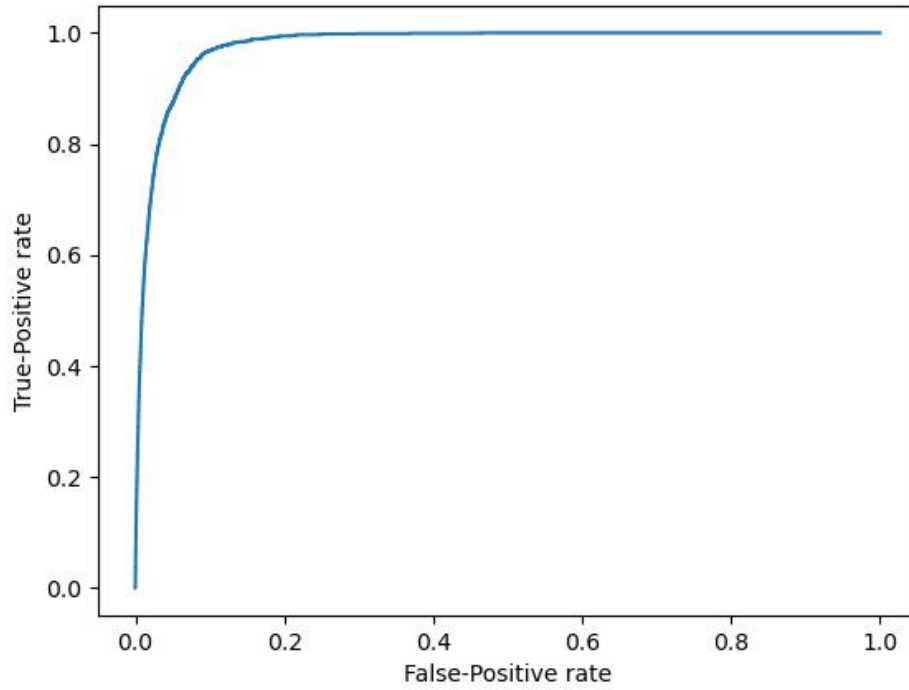


Figure 12: Receiver Operating Characteristic curve for a BDT trained on SU2L-28-800 and neighboring signal points tested on SU2L-15-800.

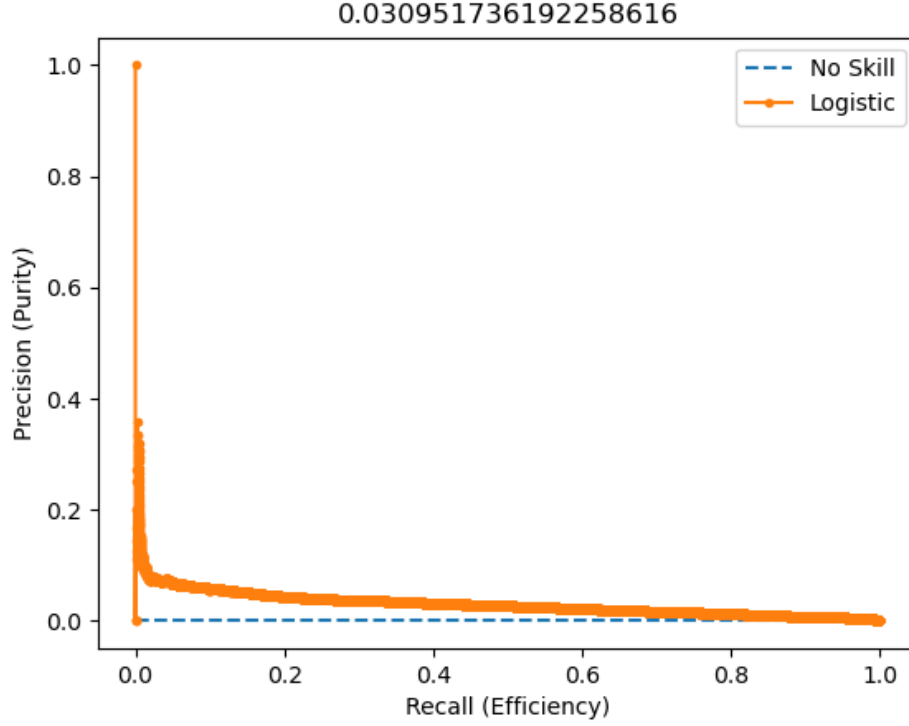


Figure 13: Precision Recall Curve for a BDT trained on SU2L-28-800 and neighboring signal points tested on SU2L-15-800.

When looking at the histogram of the discriminant in Figure 14 the situation becomes more clear. The percentage of background events that are falsely classified is very small for some selections based on the discriminant, and the percentage of signal events that are correctly classified is quite high, hence the ROC looks good. The purity on the other hand rarely reaches a high value. This comes from the very low cross sections that are predicted for the signal of this model. Even if most of the signal is classified correctly and only a very small percentage of the background is missclassified as signal, the ratio of signal to background in the selection is still lower than 40% for most selections.

As the datasets used in this project are indeed very imbalanced, because the expected signal cross sections are way lower than the background cross sections, the PRC curves and their AUC scores are in this case the preferred choice for evaluating the performance of the BDT models. They take the relative cross sections of signal and background into account, which in the ROC curves are not considered at all.

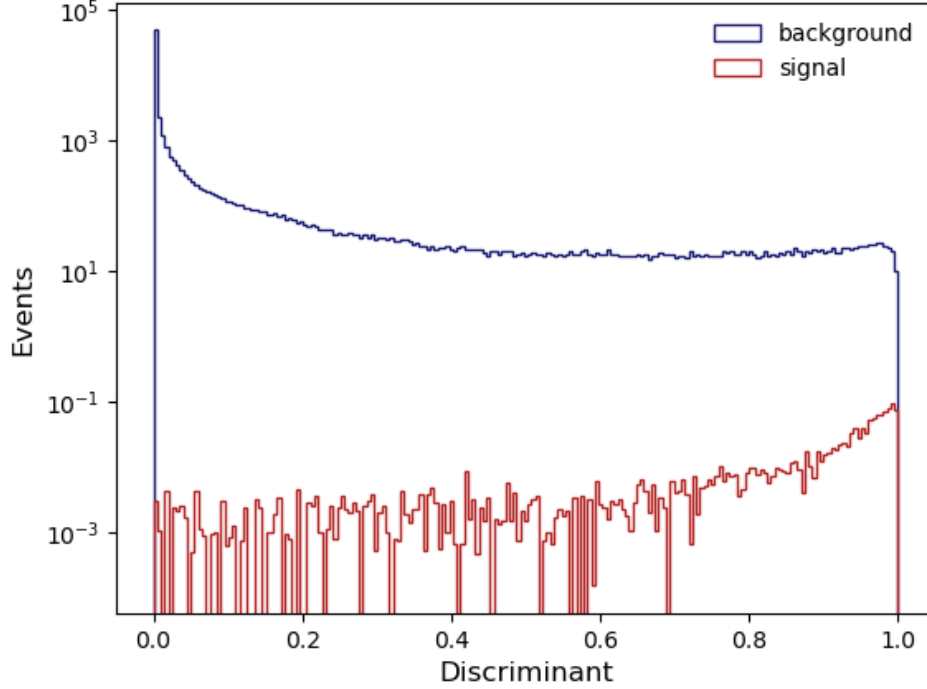


Figure 14: Discriminant distribution for a BDT trained on SU2L-28-800 and neighboring signal points tested on SU2L-15-800. Note: this histograms considers weights, but is not normalized to the target luminosity. The scale is logarithmic.

3.2 Testing on Individual Signal Points

With all backgrounds included, as a first step three different BDT models were trained and tested, one for each of the signal points with the highest event numbers.

In Figures 15 - 17 the PRC curves are shown for the three models. The PRC curve for the SU2L-25-800 signal sample looks significantly better than for the other two samples. Among the other two samples, the BDT trained and tested on SU2L-45-400 performs better than the one trained and tested on SU2R-45-300, according to the PRC curves. This behaviour coincides with the observations made in the previous project [1], where the same signal points were studied and the BDT trained and tested on SU2L-25-800 performed best and the one trained on SU2R-45-300 performed worst.

This tendency also shows up in the discriminant distributions in Figures 18 - 20. The differences between the signal points in the discriminant distributions seem to be a bit

smaller than in the PRC curves.

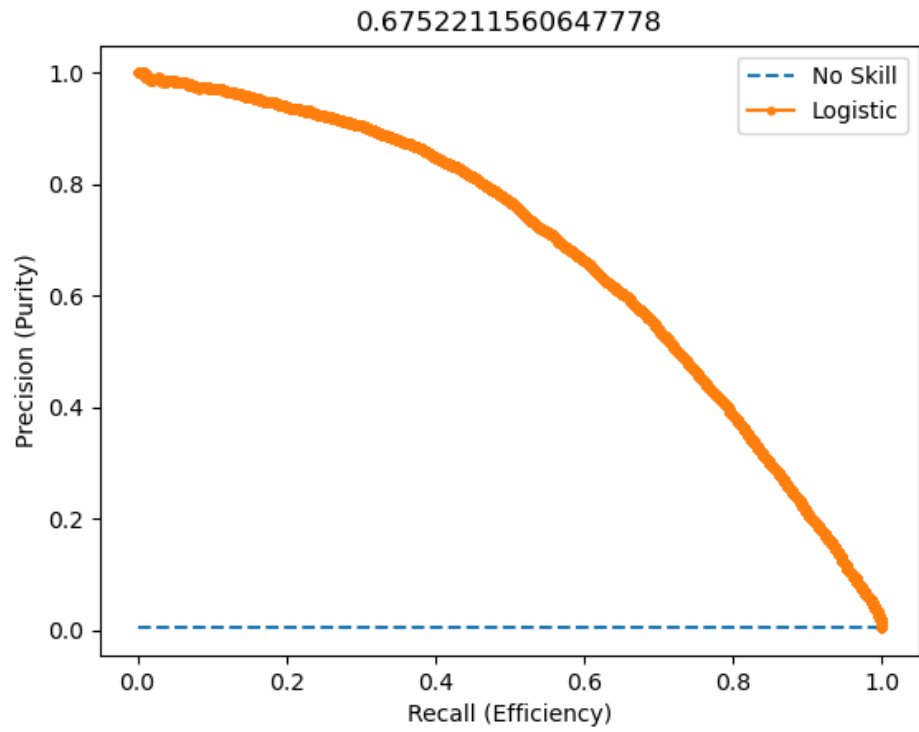


Figure 15: PRC for a BDT trained and tested on SU2L-25-800 vs. the full set of back-grounds.

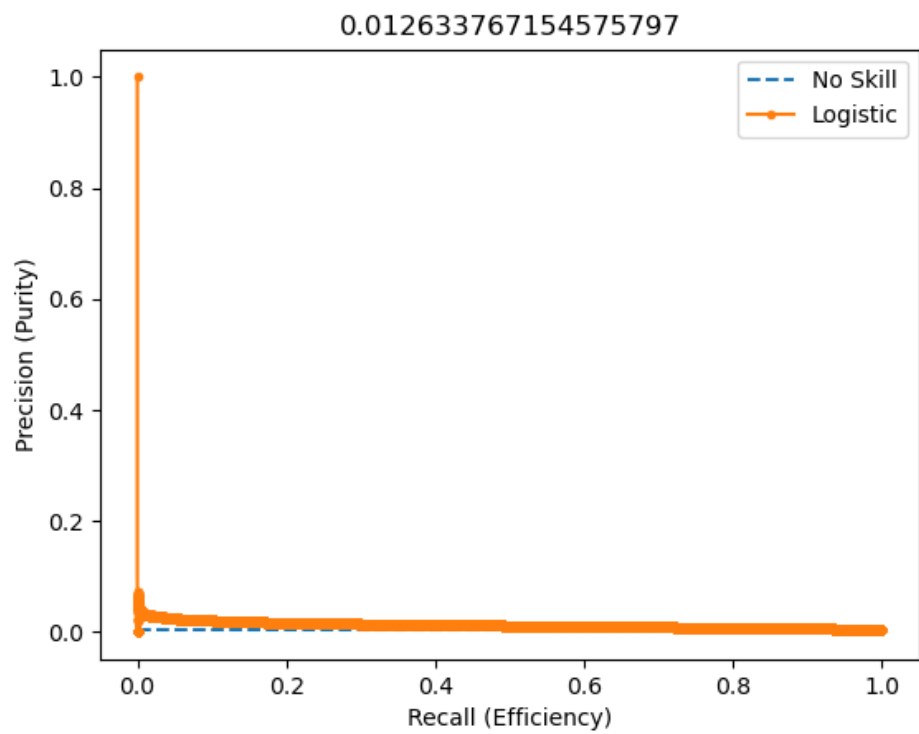


Figure 16: PRC for a BDT trained and tested on SU2L-45-400 vs. the full set of backgrounds.

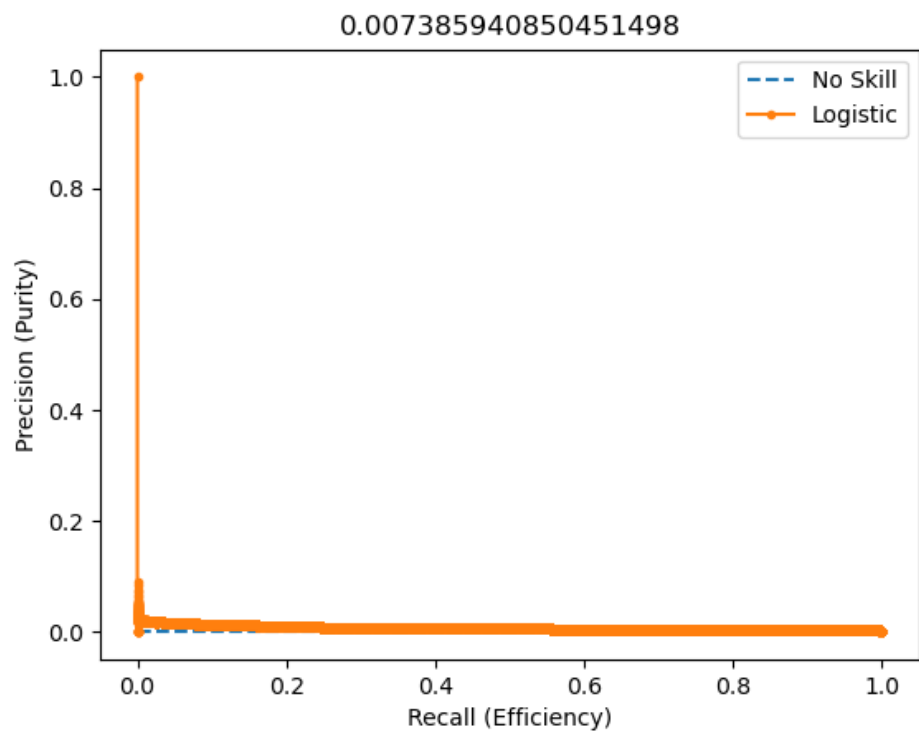


Figure 17: PRC for a BDT trained and tested on SU2R-45-300 vs. the full set of backgrounds.

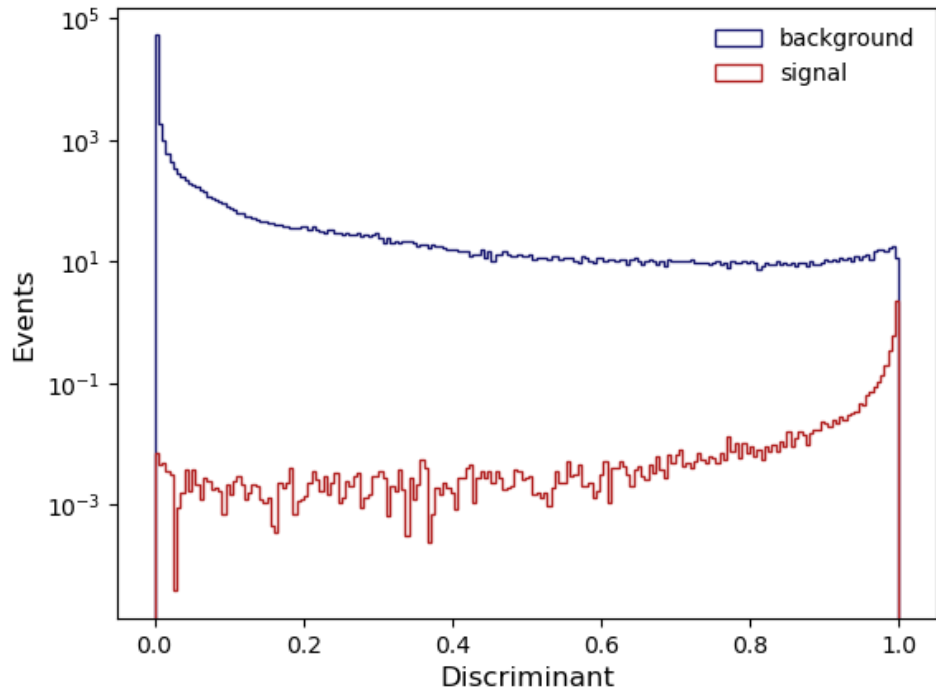


Figure 18: Histogram of the discriminant of a BDT trained and tested on SU2L-25-800 vs. the full set of backgrounds. Weights are considered, the histogram is not normalized to the target luminosity.

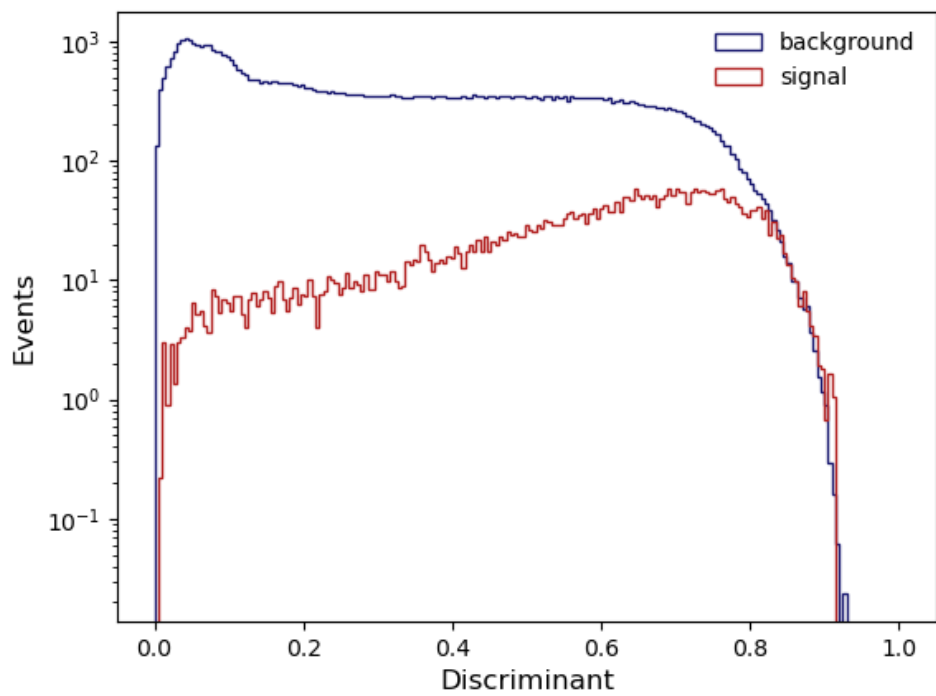


Figure 19: Histogram of the discriminant of a BDT trained and tested on SU2L-45-400 vs. the full set of backgrounds. Weights are considered, but normalization to the target luminosity is not.

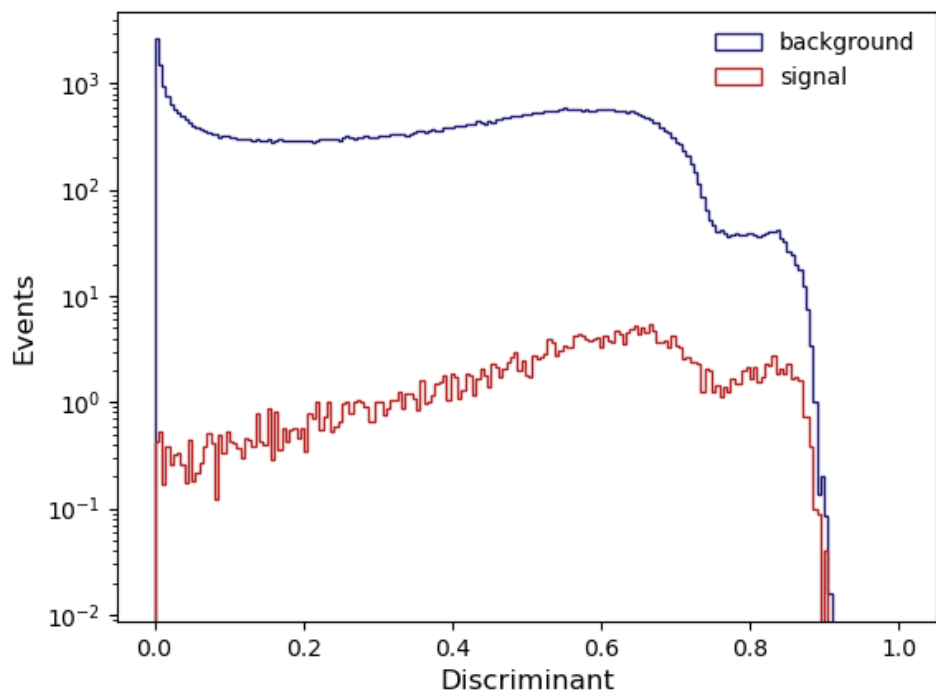


Figure 20: Histogram of the discriminant of a BDT trained and tested on SU2R-45-300 vs. the full set of backgrounds. Weights are considered, but normalization to the target luminosity is not.

3.3 Applicability of the Model to a Larger Parameter Space

As explained in section 1.4 there is a broad parameter space to cover in the analysis. For applying the BDT method in the analysis it would be desirable to train a model that can be used for all or at least many signal points at once. Therefore different sets of signal points have been used for the training and the resulting models have been applied to all signal points.

As can be seen in Figure 9 there are signal points with higher event numbers that were specifically focused on for the trainings. These points are SU2L-45-400, SU2L-25-800, and SU2R-45-300.

Models for the $SU(2)_L$ sector have been trained on the following sets of training data:

- SU2L-45-400
- SU2L-25-800
- SU2L-45-400 and SU2L-25-800
- SU2L-45-400 and its neighboring signal points (SU2L-45-300, SU2L-45-500, SU2L-35-300, SU2L-35-400, SU2L-45-500)
- SU2L-25-800 and its neighboring signal points (SU2L-15-700, SU2L-15-800, SU2L-25-700, SU2L-25-900, SU2L-35-700, SU2L-35-800, SU2L-35-900)
- SU2L-45-400 and its neighboring signal points and SU2L-25-800 and its neighboring signal points

Tables 3 to 8 show the PRC AUC scores of the different models trained on all signal points. The differences between models trained on one high-statistics point and on a high-statistics point and its neighbors are barely noticeable. While the scores are slightly different for many signal points, there is no clear trend towards either of these options. This applies as well for SU2L-45-400 as for SU2L-25-800. The reason for this is that the neighbors of the high-statistics points have 10 times lower event numbers and therefore cannot influence the training by much.

What is clearly noticeable is that the models trained on any set including SU2L-25-800 perform well for a number of surrounding signal points as well, while the models trained on SU2L-45-400 do not perform well in the metric of the PRC for any signal point, as the model does not seem to find significant differences between the signal and background signatures.

In conclusion, signal points with small event numbers do not influence the quality of the model significantly. The most important factor overall is whether or not SU2L-25-800 is

contained on the training set.

The model used for the $SU(2)_L$ sector in all following chapters is trained in SU2L-25-800 and its neighbors.

Table 3: PRC AUC scores for the whole SU2L signal grid tested with a BDT model trained on SU2L-25-800.

$\eta \backslash m_{\pi_D}(\text{GeV})$	300	400	500	600	700	800	900	1000	1100	1200
0.45	0.000	0.004	0.001	0.002	0.005	0.010	0.019	0.027	0.053	0.079
0.35	0.000	0.001	0.004	0.010	0.024	0.045	0.081	0.102	0.138	
0.25	0.001	0.007	0.026	0.116	0.274	0.694	0.372	0.346		
0.15	0.001	0.005	0.010	0.015	0.024	0.029				

Table 4: PRC AUC scores for the SU2L signal grid tested with a BDT trained on all samples with η between 0.15 and 0.35 and m_{π_D} between 700 GeV and 900 GeV.

$\eta \backslash m_{\pi_D}(\text{GeV})$	300	400	500	600	700	800	900	1000	1100	1200
0.45	0.000	0.005	0.001	0.003	0.005	0.011	0.021	0.030	0.055	0.078
0.35	0.000	0.001	0.004	0.011	0.026	0.047	0.080	0.100	0.122	
0.25	0.001	0.006	0.024	0.110	0.246	0.675	0.365	0.334		
0.15	0.001	0.004	0.010	0.017	0.027	0.031				

Table 5: PRC AUC scores for the SU2L signal grid tested with a BDT trained on SU2L-45-400.

$\eta \backslash m_{\pi_D}(\text{GeV})$	300	400	500	600	700	800	900	1000	1100	1200
0.45	0.000	0.013	0.001	0.001	0.001	0.001	0.000	0.000	0.000	0.000
0.35	0.001	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.000	
0.25	0.000	0.001	0.001	0.000	0.000	0.002	0.000	0.000		
0.15	0.000	0.001	0.001	0.000	0.000	0.000				

Table 6: PRC AUC scores for the SU2L signal grid tested with a BDT trained on all samples with η between 0.35 and 0.45 and m_{π_D} between 300 GeV and 400 GeV.

$\eta \backslash m_{\pi_D} (\text{GeV})$	300	400	500	600	700	800	900	1000	1100	1200
0.45	0.000	0.011	0.002	0.002	0.003	0.003	0.002	0.002	0.001	0.001
0.35	0.001	0.003	0.003	0.003	0.001	0.001	0.000	0.000	0.000	
0.25	0.001	0.015	0.012	0.005	0.002	0.004	0.000	0.000		
0.15	0.000	0.002	0.003	0.002	0.002	0.001				

Table 7: PRC AUC scores for the SU2L signal grid tested with a BDT trained on SU2L-45-400 and SU2L-25-800.

$\eta \backslash m_{\pi_D} (\text{GeV})$	300	400	500	600	700	800	900	1000	1100	1200
0.45	0.000	0.010	0.001	0.001	0.002	0.005	0.013	0.021	0.045	0.068
0.35	0.000	0.001	0.002	0.006	0.018	0.036	0.068	0.086	0.109	
0.25	0.000	0.005	0.024	0.107	0.243	0.666	0.358	0.322		
0.15	0.001	0.005	0.010	0.014	0.022	0.023				

Table 8: PRC AUC scores for the SU2L signal grid tested with a BDT trained on all samples with η between 0.35 and 0.45 and m_{π_D} between 300 GeV and 400 GeV as well as all samples with η between 0.15 and 0.35 and m_{π_D} between 700 GeV and 900 GeV.

$\eta \backslash m_{\pi_D} (\text{GeV})$	300	400	500	600	700	800	900	1000	1100	1200
0.45	0.000	0.007	0.001	0.003	0.005	0.011	0.020	0.029	0.054	0.078
0.35	0.000	0.002	0.004	0.011	0.025	0.047	0.077	0.098	0.120	
0.25	0.001	0.007	0.028	0.118	0.251	0.673	0.364	0.329		
0.15	0.001	0.005	0.011	0.016	0.026	0.030				

As the cross sections and the potential sensitivity of the SU2R model are way lower, not the same amount of time was spent on it. Considering that only the signal points with high event numbers were relevant in the $SU(2)_L$ case, the same was assumed for $SU(2)_R$.

3.4 Maximising Signal Significance

The most important metric to evaluate the performance of the BDT and to compare it with other methods is the significance. The significance is calculated as $sig = \frac{s}{\sqrt{s+b}}$ where s is the number of true signal events in the signal class and b is the number of true background events in the signal class. It is used as a figure of merit to approximate the sensitivity of the analysis to the signal. As the signal event numbers are so much lower than the background event numbers, this is approximately the same as the commonly used alternative $\frac{s}{\sqrt{b}}$.

To probe the potential of the BDT model to increase the significance, selections on the discriminant were tested and the resulting significances were evaluated. To find a threshold on the discriminant that leads to a selection of events with the maximal significance for a large area of the parameter space, Table 9 has been produced. It lists the threshold values that lead to the maximum reachable significance for each signal point. For some signal points in the upper left corner the best threshold values are at 0.00, while for some signal points in the lower right corner the ideal threshold is at 1.00³. This shows that the properties that the BDT uses for the classification are radically different between these two regions of the parameter space.

Table 9: Selections on the discriminant that lead to the best significances for the SU2L signal grid tested with a BDT trained on SU2L-25-800 and its neighbors.

$\eta \backslash m_{\pi_D} \text{ (GeV)}$	300	400	500	600	700	800	900	1000	1100	1200
0.45	0.00	0.00	0.00	0.04	0.34	0.64	0.87	0.93	0.96	0.98
0.35	0.00	0.01	0.23	0.67	0.90	0.96	0.98	0.98	0.99	
0.25	0.02	0.50	0.90	0.98	0.99	1.00	1.00	1.00		
0.15	0.84	0.98	1.00	1.00	1.00	0.94				

To explore if it is still possible to find a threshold on the discriminant that provides good significances over a large range of signal points the Tables 10 to 18 were produced. They show the significances for all signal points for every threshold value between 0.1 and 0.9 in steps of 0.1.

As expected the lower threshold values lead to higher significances in the upper left corner of the table, while higher threshold values improve the significance in the lower right area, but decrease it in the upper left corner.

Considering the sensitivity and exclusion limits of the current state of the analysis that can

³For these tables the BDT discriminant is evaluated in steps of 0.01 between 0.00 and 1.00. A threshold of x corresponds to a selection of all events with a discriminant of value x or higher. Accordingly, a threshold at 1.00 selects all events that have a discriminant of exactly 1.00.

be seen in Figure 6, a threshold at 0.8 would be a good choice to make a selection that could improve the sensitivity.

Table 10: Significances for a threshold on the discriminant at 0.1.

$\eta \backslash m_{\pi_D}(\text{GeV})$	300	400	500	600	700	800	900	1000	1100	1200
0.45	2.80	5.82	9.16	8.22	5.61	3.35	1.98	1.04	0.61	0.35
0.35	4.18	9.40	6.96	3.35	1.55	0.72	0.36	0.17	0.09	
0.25	4.67	3.76	1.37	0.50	0.19	0.07	0.03	0.01		
0.15	0.44	0.29	0.15	0.08	0.04	0.02				

Table 11: Significances for a threshold on the discriminant at 0.2.

$\eta \backslash m_{\pi_D}(\text{GeV})$	300	400	500	600	700	800	900	1000	1100	1200
0.45	1.98	4.58	7.23	7.73	6.00	3.83	2.33	1.24	0.73	0.42
0.35	3.56	7.75	7.12	3.82	1.82	0.86	0.43	0.21	0.10	
0.25	4.24	4.12	1.61	0.60	0.22	0.08	0.03	0.01		
0.15	0.46	0.29	0.16	0.09	0.04	0.02				

Table 12: Significances for a threshold on the discriminant at 0.3.

$\eta \backslash m_{\pi_D}(\text{GeV})$	300	400	500	600	700	800	900	1000	1100	1200
0.45	2.05	4.15	5.98	7.08	6.10	4.11	2.56	1.39	0.82	0.47
0.35	3.11	6.27	7.01	4.13	2.01	0.97	0.48	0.24	0.12	
0.25	3.72	4.25	1.77	0.67	0.25	0.09	0.04	0.01		
0.15	0.47	0.29	0.16	0.09	0.05	0.03				

Table 13: Significances for a threshold on the discriminant at 0.4.

$\eta \backslash m_{\pi_D}(\text{GeV})$	300	400	500	600	700	800	900	1000	1100	1200
0.45	1.97	3.64	5.22	6.45	6.11	4.29	2.76	1.52	0.90	0.52
0.35	2.79	5.47	6.83	4.34	2.19	1.06	0.53	0.26	0.13	
0.25	3.32	4.36	1.92	0.74	0.27	0.10	0.04	0.02		
0.15	0.49	0.30	0.16	0.09	0.05	0.03				

Table 14: Significances for a threshold on the discriminant at 0.5.

$\eta \backslash m_{\pi_D}(\text{GeV})$	300	400	500	600	700	800	900	1000	1100	1200
0.45	1.82	3.22	4.54	5.81	5.90	4.36	2.91	1.63	0.98	0.57
0.35	2.10	4.98	6.57	4.49	2.34	1.16	0.58	0.29	0.14	
0.25	2.94	4.40	2.05	0.79	0.30	0.11	0.04	0.02		
0.15	0.50	0.30	0.16	0.10	0.05	0.03				

Table 15: Significances for a threshold on the discriminant at 0.6.

$\eta \backslash m_{\pi_D}(\text{GeV})$	300	400	500	600	700	800	900	1000	1100	1200
0.45	1.23	2.86	4.07	5.27	5.65	4.46	3.09	1.76	1.06	0.62
0.35	1.83	4.59	6.19	4.53	2.50	1.26	0.64	0.32	0.16	
0.25	2.77	4.39	2.18	0.87	0.33	0.12	0.05	0.02		
0.15	0.48	0.31	0.16	0.10	0.06	0.03				

Table 16: Significances for a threshold on the discriminant at 0.7.

$\eta \backslash m_{\pi_D}(\text{GeV})$	300	400	500	600	700	800	900	1000	1100	1200
0.45	1.39	2.61	3.83	4.75	5.15	4.50	3.25	1.89	1.17	0.69
0.35	1.59	3.83	5.54	4.61	2.67	1.38	0.71	0.36	0.18	
0.25	2.29	4.26	2.35	0.96	0.36	0.14	0.05	0.02		
0.15	0.51	0.33	0.16	0.10	0.06	0.03				

Table 17: Significances for a threshold on the discriminant at 0.8.

$\eta \backslash m_{\pi_D}(\text{GeV})$	300	400	500	600	700	800	900	1000	1100	1200
0.45	1.04	2.34	3.05	3.95	4.37	4.36	3.45	2.07	1.31	0.80
0.35	1.40	3.15	4.63	4.43	2.84	1.56	0.82	0.42	0.21	
0.25	1.79	3.93	2.54	1.10	0.43	0.16	0.06	0.03		
0.15	0.53	0.34	0.16	0.10	0.06	0.04				

Table 18: Significances for a threshold on the discriminant at 0.9.

$\eta \backslash m_{\pi_D}(\text{GeV})$	300	400	500	600	700	800	900	1000	1100	1200
0.45	1.37	1.50	2.07	2.88	3.34	3.79	3.39	2.23	1.53	0.99
0.35	1.10	1.86	3.50	3.98	2.99	1.80	1.01	0.54	0.27	
0.25	1.17	3.47	2.70	1.35	0.54	0.21	0.08	0.03		
0.15	0.53	0.38	0.16	0.10	0.06	0.04				

Compared with the significances of the analysis without the BDT that can be seen in Figure 21 there are many signal points with improved significance. Applying a selection based on the BDT discriminant with a threshold at 0.8 would very likely improve the sensitivity of the analysis.

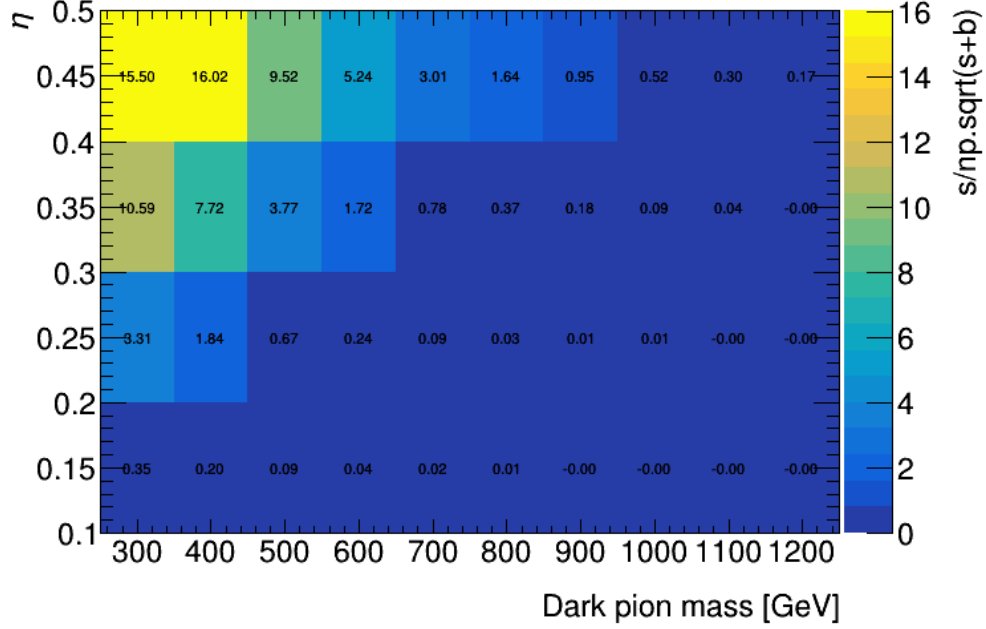


Figure 21: Expected significances of the 1-lepton channel of the ATLAS dark meson analysis in the $SU(2)_L$ model. Figure from Olga Sunneborn Gudnadottir.

3.4.1 Raw Event Numbers

If a selection on the discriminant was to be implemented in the analysis, the influence of statistical uncertainties coming from the signal yields in simulation would have to be considered. Table 19 shows how many unweighted signal events are in the signal category for a threshold on the discriminant at 0.8. As noted in Figure 9 the complete number of events for the enhanced statistics points SU2L-45-400 and SU2L-25-800 are 100 000, while for the others it is 10 000. This table however was produced using only half of each sample, because the other half had already been used for the training and could therefore not be used for the testing of the model.

Table 19: Unweighted signal event numbers in the signal category for the threshold of 0.8 on the discriminant.

$\eta \backslash m_{\pi_D}(\text{GeV})$	300	400	500	600	700	800	900	1000	1100	1200
0.45	4	198	75	205	479	905	1301	1366	1484	1541
0.35	13	96	382	886	1234	1377	1453	1450	1367	
0.25	61	552	1030	1229	1222	11066	1027	981		
0.15	113	289	386	632	928	1179				

The remaining signal event numbers are sufficiently high for the results to be significant in the relevant areas of the parameter space where the significances would likely be improved compared to the current estimates of the analysis.

In the upper left corner of the table the remaining signal event numbers would not be sufficient, but as they are already excluded this is not of high importance. If the selection was to be optimised for the whole parameter space, a different selection with a lower threshold on the discriminant would be applied to the upper left region, which would lead to much higher event numbers for the corresponding signal points.

3.5 Distributions

To visualize the way the BDT uses the variables for the classification, histograms of the variables before and after applying a threshold of 0.8 on the discriminant have been produced. The histograms in this chapter include weights and are normalized to the target luminosity. The background is split up in its constituents and several different signal points are plotted in each histogram for comparison.

Figure 22 shows the discriminant itself before and after applying the threshold.

Figure 23 to 30 show all the used variables before and after applying the threshold. It can be noticed that the BDT works in more complex ways than a selection based on the individual variables, because for most variables there is no clear threshold that separates the two histograms, but instead the shape of the distribution changes to a shape that more closely resembles the signal distribution.

The variable that has been used in the most obvious way is H_T . In this case there is a clear threshold at around 600 GeV, which fits with the distributions of the high- m_{π_D} signal points.

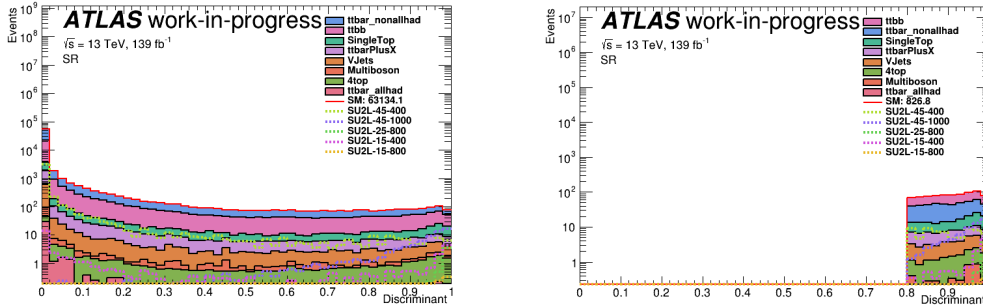


Figure 22: Discriminant of the model trained on SU2L-25-800 for all background samples and a number of signal samples. The left histogram shows the initial distribution, the right histogram contains only events with a discriminant > 0.8 .

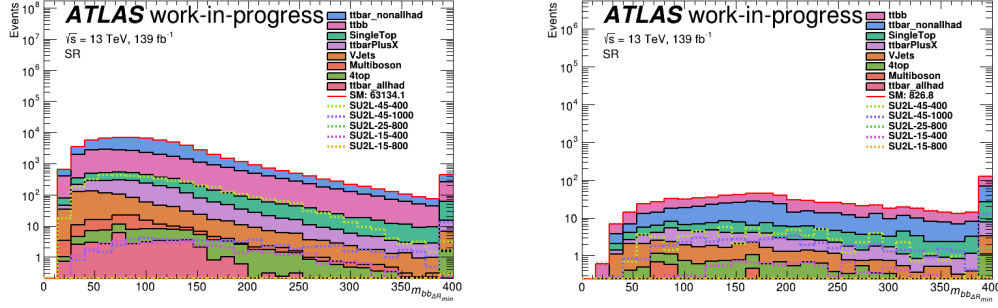


Figure 23: $m_{bb\Delta R_{min}}$ for all background samples and a number of signal samples. The left histogram shows the initial distribution, the right histogram contains only events with a discriminant > 0.8 . The last bin contains the overflow.

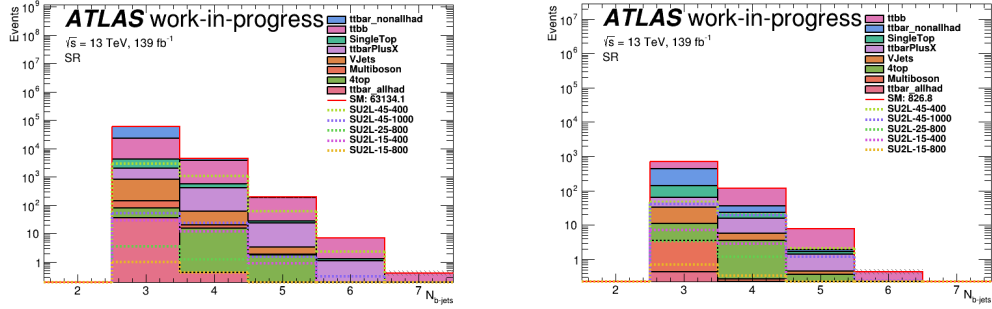


Figure 24: N_{b-jets} for all background samples and a number of signal samples. The left histogram shows the initial distribution, the right histogram contains only events with a discriminant > 0.8 .

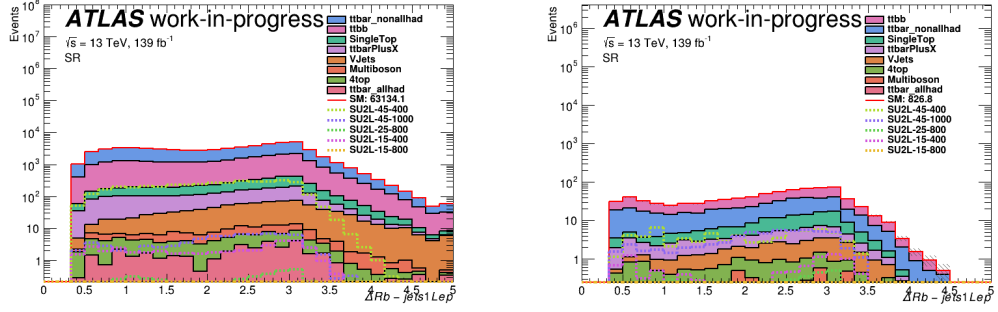


Figure 25: $\Delta R(l, b1)$ for all background samples and a number of signal samples. The left histogram shows the initial distribution, the right histogram contains only events with a discriminant > 0.8 . The last bin contains the overflow.

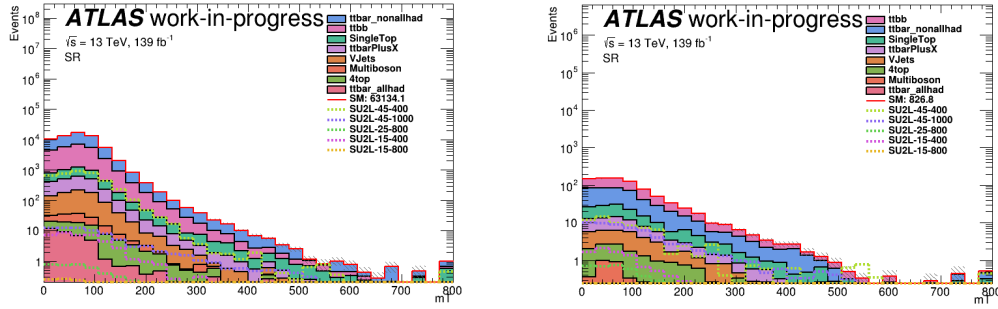


Figure 26: m_T for all background samples and a number of signal samples. The left histogram shows the initial distribution, the right histogram contains only events with a discriminant > 0.8 .

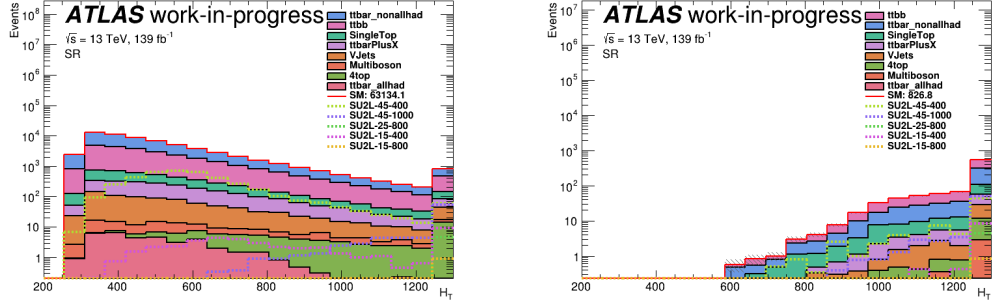


Figure 27: H_T for all background samples and a number of signal samples. The left histogram shows the initial distribution, the right histogram contains only events with a discriminant > 0.8 . The last bin contains the overflow.

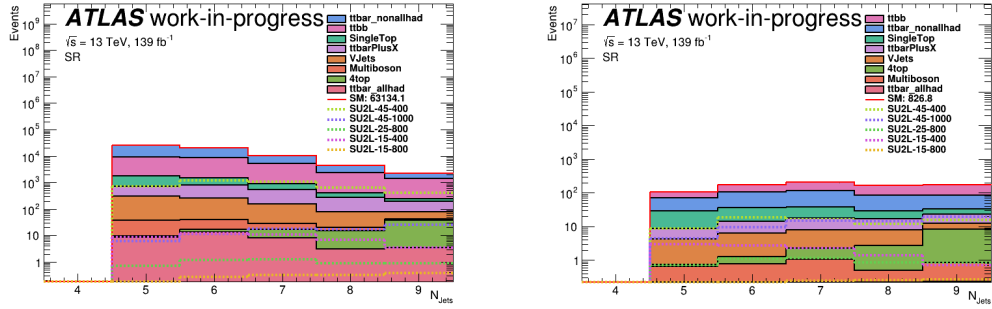


Figure 28: N_{Jets} for all background samples and a number of signal samples. The left histogram shows the initial distribution, the right histogram contains only events with a discriminant > 0.8 .

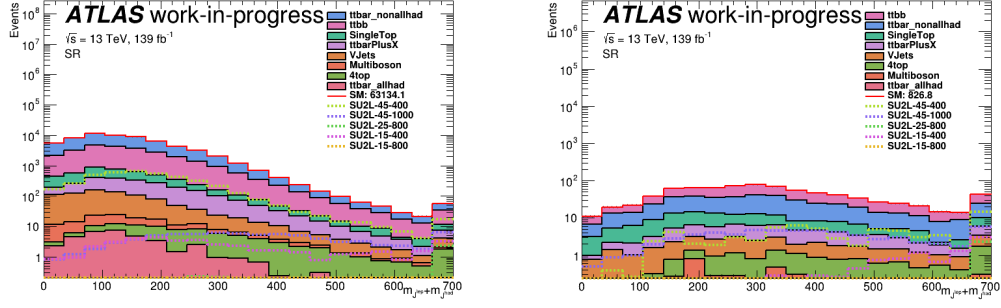


Figure 29: $m_{J^{lep}, R=1.2} + m_{J^{had}, R=1,2}$ for all background samples and a number of signal samples. The left histogram shows the initial distribution, the right histogram contains only events with a discriminant > 0.8 . The last bin contains the overflow.

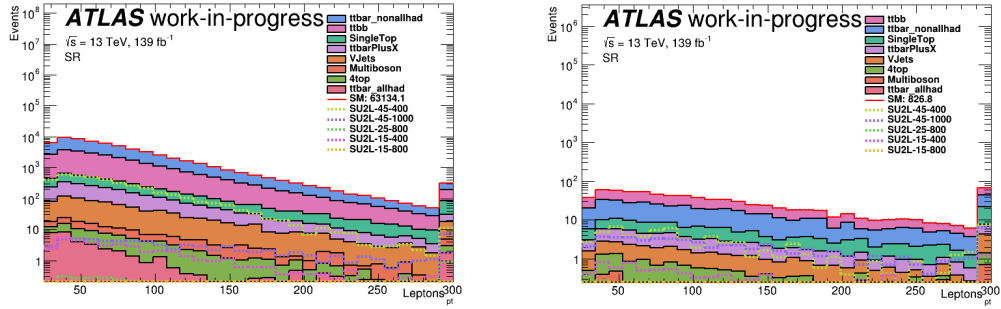


Figure 30: p_{T_l} for all background samples and a number of signal samples. The left histogram shows the initial distribution, the right histogram contains only events with a discriminant > 0.8 . The last bin contains the overflow.

4 Conclusion

This project has explored the potential of BDTs in the 1-lepton channel of the ATLAS dark meson search. For this purpose a number of different BDT models have been trained to discriminate the simulated signal samples from the most relevant simulated background samples.

As an estimation of whether or not the sensitivity can be improved the significance $\frac{s}{\sqrt{s+b}}$ has been used. Even though no statistical uncertainties on the signal were considered, this estimation shows a promising increase of the significance for the majority of the parameter space already for one selection on the discriminant of one model.

Different values for η and m_{π_D} cause huge differences in the kinematics, which lead to a significantly higher resemblance of high- η and low- m_{π_D} signals to the background. This can be explained by the fact that the discriminating variables are partly designed to reconstruct the masses of the two dark pions, which will be easier if the masses are higher. Simultaneously a high value of η minimizes the transverse momenta that the dark pions receive, because the mass of the dark rho is mostly transformed into the mass of the dark pions instead of being released as kinetic energy.

This leads to significantly worse performance of the BDT in the high- η and low- m_{π_D} region. On the other hand, the production cross sections are considerably higher for lower values of m_{π_D} , which still leads to high significances in this region. This behaviour coincides with what was observed without the usage of BDTs in the analysis.

Considering this, the ideal approach for the use of BDTs in this search would be not to apply one common selection on one discriminant for the whole parameter space, but rather to use different selections and maybe even different models for different regions of the parameter space, as is commonly done in other searches in ATLAS.

Applying selections on a discriminant is the fastest and simplest option to use a BDT in this analysis, which is why it was tested in this project. Another option that has not been tested for this case, but that could be explored if BDTs were to be implemented in the search, would be to use the discriminant as a variable to fit on.

Additionally, the method could be extended by adding m_{π_D} as an input variable for the training and thereby increasing the mass range over which the parametrized BDT is sensitive.

To conclude, the usage of BDTs in the ATLAS search for dark mesons shows great potential to increase the sensitivity of the analysis in first estimations for a simple discriminant-based selection. A more sophisticated usage can be expected to increase the sensitivity even more.

References

- [1] E. Mayer. Application of a boosted decision tree to signals and backgrounds of the search for dark mesons at the atlas experiment. <http://uu.diva-portal.org/smash/get/diva2:1714864/FULLTEXT01.pdf>, 2022.
- [2] Wikipedia. Table of the standard model. https://en.wikipedia.org/wiki/File:Standard_Model_of_Elementary_Particles.svg, 2023. [Online; accessed 29-May-2023].
- [3] G. Bertone et al. History of dark matter. *Reviews of Modern Physics*, 90(4), oct 2018.
- [4] R. C. Brower et al. Stealth dark matter confinement transition and gravitational waves. *Physical Review D*, 103(1), jan 2021.
- [5] G. D. Kribs et al. Dark mesons at the LHC. *Journal of High Energy Physics*, 2019(7), jul 2019.
- [6] ATLAS Collaboration. The ATLAS Experiment at the CERN Large Hadron Collider. *JINST*, 3:S08003, 2008.
- [7] R. M. Bianchi and ATLAS Collaboration. ATLAS experiment schematic illustration. General Photo, 2022.
- [8] G. D. Kribs et al. Effective theories of dark mesons with custodial symmetry. *Journal of High Energy Physics*, 2019(8), aug 2019.
- [9] ATLAS Collaboration. Search for dark mesons decaying to top and bottom quarks with the ATLAS detector in 140 fb^{-1} of proton-proton collisions at $\sqrt{s} = 13 \text{ TeV}$. Technical report, CERN, Geneva, 2023. <http://cds.cern.ch/record/2859005>.
- [10] M. Cacciari et al. The anti- k_t jet clustering algorithm. *Journal of High Energy Physics*, 2008(04):063–063, apr 2008.
- [11] G. R. Gledhill et al. Search for dark mesons decaying to top and bottom quarks with the ATLAS detector in 139 fb^{-1} proton–proton collisions at $\sqrt{s} = 13 \text{ TeV}$, 2023. Internal note. cds.cern.ch/record/2711578.
- [12] Y. Coadou. Boosted decision trees. In *Artificial Intelligence for High Energy Physics*, pages 9–58. WORLD SCIENTIFIC, feb 2022.
- [13] V. Lalchand. Extracting more from boosted decision trees: A high energy physics case study, 2020. arXiv:2001.06033.

- [14] A. Lindholm et al. *Machine Learning: A First Course for Engineers and Scientists*. Cambridge University Press, 2021.
- [15] J. Alwall et al. The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations. *Journal of High Energy Physics*, 2014(7), jul 2014.
- [16] C. Bierlich et al. A comprehensive guide to the physics and usage of pythia 8.3. arXiv:2203.11601, 2022.
- [17] C. Oleari. The powheg box. *Nuclear Physics B - Proceedings Supplements*, 205-206:36–41, aug 2010.
- [18] P. Freeman et al. Sherpa: a mission-independent data analysis application. In Jean-Luc Starck and Fionn D. Murtagh, editors, *SPIE Proceedings*. SPIE, nov 2001.
- [19] G. Van Rossum et al. *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam, 1995.
- [20] R. Brun et al. Root - an object oriented data analysis framework. Proceedings AI-HENP’96 Workshop, September 1996.
- [21] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.
- [22] C. R. Harris et al. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020.
- [23] The pandas development team. pandas-dev/pandas: Pandas, February 2020.
- [24] T. Chen et al. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’16, pages 785–794, New York, NY, USA, 2016. ACM.
- [25] F. Pedregosa et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [26] E. Mayer. Bdts in the atlas search for dark mesons. https://gitlab.com/eva_mayer_/bdts-in-the-atlas-search-for-dark-mesons.
- [27] ATLAS Collaboration. Search for $t\bar{t}h/a \rightarrow t\bar{t}t\bar{t}$ production in the multilepton final state in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. cds.cern.ch/record/2805212, 2022.

A Code

The complete code can be found on Gitlab [26].

Figure 31 shows an extract of the code that applies the preselection to the dataframes and divides them into test and training data.

```
192     for filename, label in Rsig:
193         print(">>> Extract the training and testing events for {} from the {} dataset.".format(
194             label, filename))
195         print(Rsig.index([filename,label]))
196         l = Rsig.index([filename,label])
197
198         # Load dataset, filter the required events and define the training variables
199         filepath = filename
200         label = label
201         df = ROOT.RDataFrame("nominal", filepath)
202         df = filter_events(df)
203         df = define_variables(df,l)
204
205         # Book cutflow report
206         report = df.Report()
207
208         # Split dataset by event number for training and testing
209         columns = ROOT.std.vector["string"](variables)
210         data_train = df.Filter("event % 2 == 0", "Select events with even event number for training")\
211             .Snapshot("Events", "train_allx2_sig_" + label + ".root")
212         data_test = df.Filter("event % 2 == 1", "Select events with odd event number for training")\
213             .Snapshot("Events", "test_allx2_sig_" + label + ".root")
214
215
216
217         # Print cutflow report
218         report.Print()
```

Figure 31: Extract from the code that applies the preselection and splits the ROOT files into test and training sets.

Figure 32 shows an extract from the code that trains the BDT and sets the configurations of the training.

```

247 param = {}
248
249 # Booster parameters
250 param['eta'] = 0.1 # learning rate
251 param['max_depth'] = d # maximum depth of a tree
252 param['subsample'] = 0.8 # fraction of events to train tree on
253 param['colsample_bytree'] = 0.8 # fraction of features to train tree on
254 param['scale_pos_weight'] = fr
255 param['tree_method'] = 'gpu_hist'
256
257 # Learning task parameters
258 param['objective'] = 'binary:logistic' # objective function
259 param['eval_metric'] = 'error' # evaluation metric for cross validation
260 param = list(param.items()) + [('eval_metric', 'logloss')] + [('eval_metric', 'rmse')]
261
262 num_trees = 100 # number of trees to make
263
264
265 #training:
266 print("training...")
267 booster = xgb.train(param, train, num_boost_round=num_trees)
268
269 booster.save_model("{}_model_v3_ttbq_weighted.txt".format(s))
270 booster.save_model("{}_model_v3_p2.json".format(s))
271 booster.save_model("{}_model_v3_ttbq.ubj".format(s))
272 booster.save_model("{}_model_v3_p2.model".format(s))
273 print(booster.best_ntree_limit)
274

```

Figure 32: Extract from the code that trains the BDT.

Figure 33 shows an extract from the code that applies the BDT model to the test data and saves it in a new ROOT file with the discriminant as a branch.

```

325 model = xgb.Booster()
326
327 model.load_model("{}_model_v3_p2.json".format(s))
328
329
330 predictions = model.predict(test)
331 #p = open('predictions_{}.txt'.format(sig), 'w')
332 #p.write(predictions)
333
334 #add discriminant to pandas dataframe
335 panda_test['discriminant']=predictions
336 panda_test_s = panda_test[panda_test.exactLabel=='sig']
337
338 #create dictionary from pandas dataframe
339 data_s = {key: panda_test_s[key].values for key in ["Jets_N_c", "BJets_70_N_c", "bb_m_for_minDeltaR_c",
"mT_c", "Leptons_pt_c", "HT_pt_c", "deltaRBJetLep_c", "LJet_m_plus_RCJet_m_12_c", "Label", "discriminant",
"weight"]}
340
341 #convert column Label from int to float
342 data_s["Label"] = data_s["Label"].astype(np.float32)
343
344 #create and save RDataFrame from dictionary
345 rdf_disc = ROOT.RDF.MakeNumpyDataFrame(data_s)
346 rdf_disc.Snapshot("Events", "test_{0}_{1}_withDisc.root".format(name, sig))
347

```

Figure 33: Extract from the code that applies the model to the test data.