



Using b-tagging to Improve $t\bar{t}H$ Searches at ATLAS

Sean Cooper, Queen Mary University of London, United Kingdom

September 8, 2016

Abstract

The possibility of including b-tagging scores as additional variables in the classification BDT for the $t\bar{t}H$, with $H \rightarrow b\bar{b}$, single-lepton channel was investigated. It was found that in the ≥ 6 jets with 3 b-jets region, the background in the highest BDT score bin could be reduced by almost half. A further study revealed that these b-tagging jets discriminate against non- $t\bar{t} + b\bar{b}$ sourced backgrounds. At present only binned b-tagging variables are feasible, however the introduction of these still gives promising performance improvements.

Contents

1. Introduction	3
1.1. Top Associated Higgs Production	3
1.2. The $t\bar{t} + b\bar{b}$ Background	3
1.3. The Single-Lepton Channel	3
2. Categorising the Data	4
2.1. Jet b-tagging	4
2.2. Defining Regions	5
3. The Multivariate Analysis	6
3.1. Why Use Multivariate Analysis?	6
3.2. Boosted Decision Trees	7
3.3. Receiver Operating Characteristic	7
3.4. BDT Overtraining	8
4. The MVA Setup	9
4.1. The MVA Variables	9
4.2. BDT Settings	10
4.3. Testing the MVA	10
5. Adding b-tagging Variables	13
5.1. ROC Improvements	13
5.2. $t\bar{t} + b\bar{b}$ Background Study	15
5.3. Binned b-tagging Variables	17
6. Conclusion	19
A. The Current Classification BDT Variables [3]	21
B. mv2c10 b-tagging Cuts	22
C. Adding Variables in Order of Least Importance	23

1. Introduction

1.1. Top Associated Higgs Production

The Higgs production mechanism that this investigation focuses on is top associated Higgs production. Precise measurement of the $t\bar{t}H$ vertex seen in figure 1 would allow us to calculate the Yukawa coupling of the top to the Higgs. Theoretically, the $H \rightarrow b\bar{b}$ decay mode pictured would produce the largest number of $t\bar{t}H$ associated events out of any decay mode. Unfortunately, the presence of very large backgrounds makes this signal one of the hardest channels to detect.

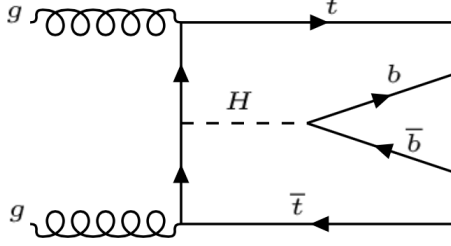


Figure 1: The $t\bar{t}H$ production process, with $H \rightarrow b\bar{b}$ decay mode.

1.2. The $t\bar{t} + b\bar{b}$ Background

The $t\bar{t} + b\bar{b}$ background is the main source of issues when it comes to measuring the $t\bar{t}H$ with $H \rightarrow b\bar{b}$ signal. As seen in figure 2 the final state particles of the $t\bar{t} + b\bar{b}$ are the same as those in the signal: t , \bar{t} , b , and \bar{b} . This is what makes the signal so hard to distinguish.

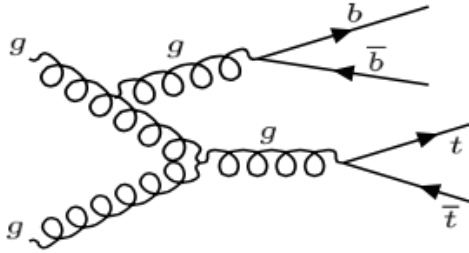


Figure 2: The $t\bar{t} + b\bar{b}$ background process.

1.3. The Single-Lepton Channel

Top quarks decay almost exclusively to a b quark and a W boson, as such we expect to see 4 b -jets in our signal. These 4 b quarks can be seen in figure 3, which displays the

single-lepton channel. This single-lepton channel, so named because only one of the W bosons decays to give a charged lepton, is the signal that this project specifically focuses on. Other possible channels include ones in which the W bosons decay only to quarks, or the dilepton state where both of the W bosons decay to leptons. It can be seen from this single-lepton channel we expect 6 jets, 4 of which are b -jets. That is providing there are no additional gluon jets, or jets lost in the detector.

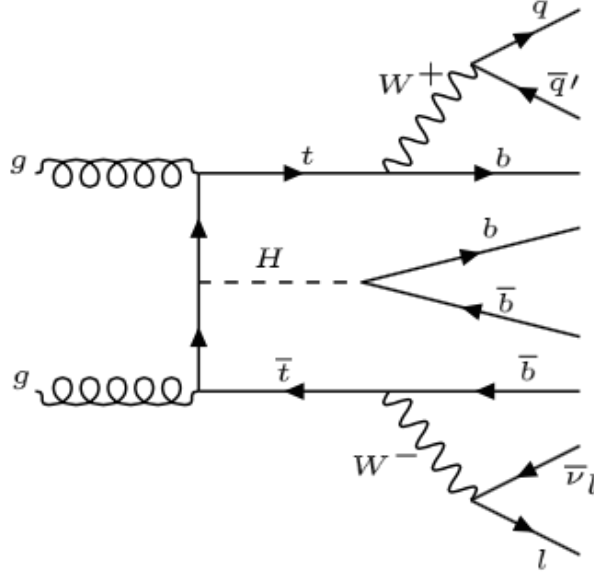


Figure 3: The expected process in the single-lepton channel.

2. Categorising the Data

2.1. Jet b-tagging

b -tagging is the flavour tagging process used to decide whether a given jet was sourced from a b quark. The bottom quark does not decay immediately after it is produced; and so it travels a short distance before it hadronises. If a jet vertex is found to be displaced from the interaction vertex, then the parent particle could be a b quark. Figure 4 shows a diagram of such a displaced vertex. Aside from this, information regarding a jet's width, and multiplicity, can also be used to classify the jet's source. To obtain optimal performance, many of these b -tagging methods are used together by implementing an artificial neural network.[1] The issue is that the b -tagging algorithms are not perfect, often a b -jet will not be tagged, and sometimes other particles such as charm quarks are incorrectly tagged as a b .

The final output of these algorithms is a b -tagging score, in this case the $mv2c10$ variable, which classifies how b -like the jet is. Cuts can then be made on this value to decide which b -tagging efficiency is appropriate for the analysis. Higher efficiencies

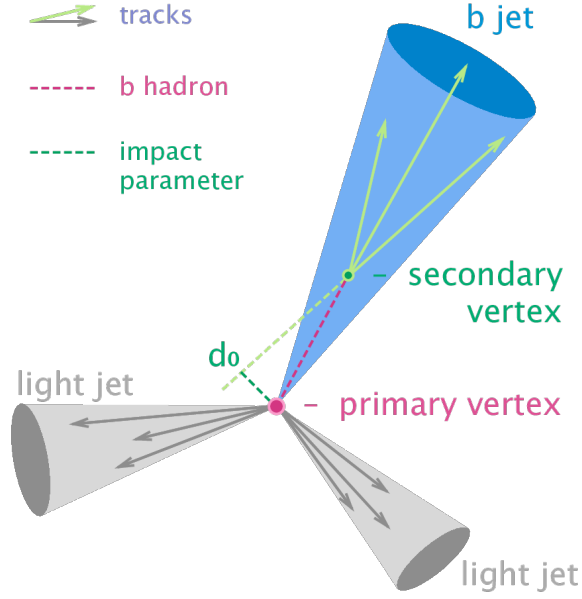


Figure 4: An example of a displaced b-jet vertex.[2]

yield more b-tagged jets, but have a lower purity. The b-tagging efficiency used in this investigation was 70%, yielding a purity of 97.46%. A list of the weight cuts at different efficiencies can be found in appendix B.

2.2. Defining Regions

The data is split into 9 different categories, or regions, based on how many jets and b-jets there are in each event. Table 1 shows how these regions are defined. The red shaded regions, 5, 7 and 8, are of most interest. These regions provide the most optimal signal, as the number of jets, and number of b-jets, is very similar to those we expect from our signal.

	4 jets	5 jets	≥ 6 jets
2 b-tags	0	3	6
3 b-tags	1	4	7
≥ 4 b-tags	2	5	8

Table 1: The definitions of the regions used to categorise the Monte Carlo data.

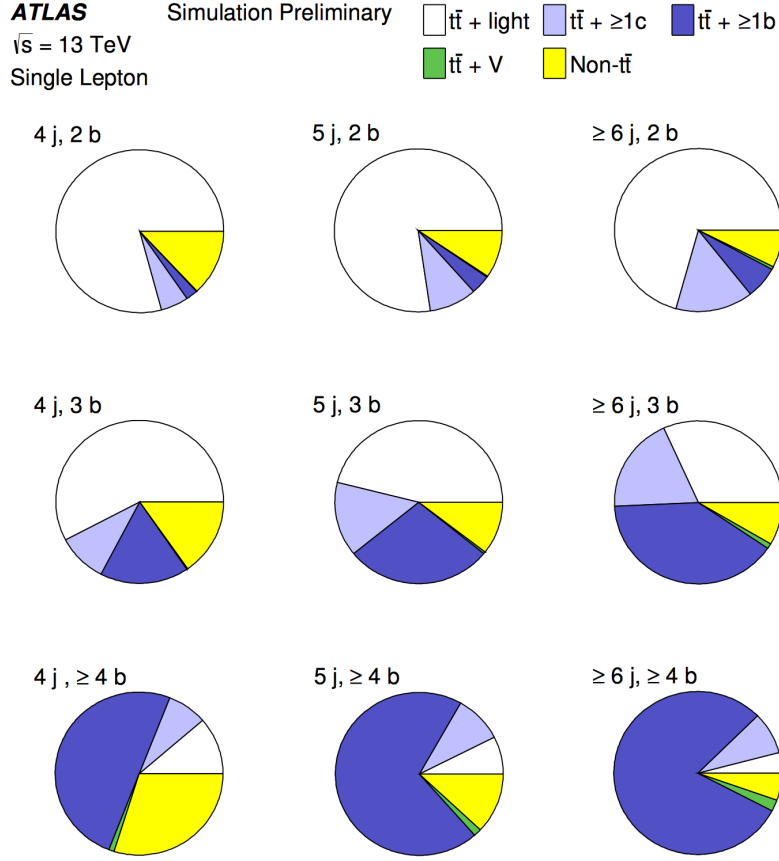


Figure 5: The components of each categorisation region's background.[3]

The pie charts shown in figure 5 outline the different sources of background in each region. Unfortunately, Regions 5 and 8 are dominated by $t\bar{t} + b\bar{b}$ sourced backgrounds. In Region 7, however, the $t\bar{t} + b\bar{b}$ is still less than half the total background.

3. The Multivariate Analysis

3.1. Why Use Multivariate Analysis?

Figure 6 shows a plot of the number of jets with $p_T \geq 40 \text{ GeV}$ in Region 7. Both the signal and background histograms peak at almost the same value. This is disappointing since this variable supposedly provides the best signal to background separation in this region. In this case, a variable's separation is defined by the ratio of overlapping and non-overlapping areas of the two histograms. Clearly then, single variable cuts will not be useful in this analysis. Instead the cuts need to be made more dynamically, varying multiple variables at the same time. In fact, the analysis done in this region uses 17 variables! This truly is a Multivariate Analysis (MVA). Beyond a 2 or 3 variables it becomes increasingly difficult for a human to set these cuts by hand. Instead machine learning techniques must be used to properly analyse the 17-dimensional hyperspace.

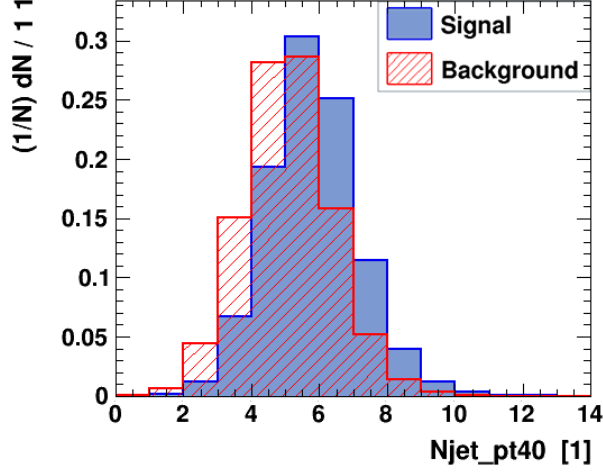


Figure 6: The number of jets with $p_T \geq 40$ GeV in Region 7, for both the signal and background samples.

3.2. Boosted Decision Trees

Boosted Decision Trees (BDTs) are the MVA method of choice in this investigation. A decision tree, is a flow-chart-like network of yes/no nodes. At each node a cut on some variable is applied, with nodes deeper often cutting in a way more specific to the events that reach it. These cuts allow an event to be classified as signal-like or background-like (or somewhere in between). Due to the large number of variable choices at each node, a single tree is far from powerful. The variables choices are largely random, albeit weighted by the separating power each one would give on the dataset. This means that any single decision tree actually has a very poor performance, and do not classify much better than a random guesser would. However, by *boosting* the decisions trees, we can combine many of these weak classifiers into one strong classifier. The boosting algorithm used in this project was the AdaBoost, or “Adaptive Boosting”, algorithm. Simply put this algorithm uses a weighted sum of the outcomes of many different trees as it’s final classification. The BDT is trained by acting it on known *training* data, and comparing the outcome to the true classification. The weighting of each tree is then adjusted, based on some error function, potentially improving its future performance. Many iterations of this process will improve the BDT’s classifying power.

3.3. Receiver Operating Characteristic

A good way of determining the classification power of a BDT is it’s Receiver Operating Characteristic (ROC) curve. An example ROC curve is shown in figure 7. These plots are a representation of the BDT’s profile, at various signal efficiency thresholds. Any one point on the curve gives the signal efficiency (true positive rate) and the background rejection (true negative rate) of the BDT. The black line on this example plot represents a random guesser. A perfect classifier would be one that extends all the way to the

top-right of the plot. The red line is a much more realistic curve for a BDT classifier. The further it extends towards the top-right corner, and the larger the area underneath the ROC curve, the better the MVA has performed.

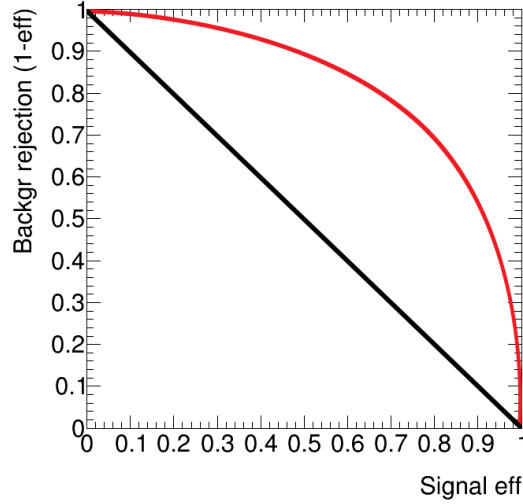


Figure 7: An example ROC curve, the black line shows the behaviour of a random guesser. The red line is a much better classifier.

3.4. BDT Overtraining

An issue with using BDT classifiers is that they have a tendency to overtrain. Overtraining is where the BDT has begun to focus on random fluctuations and noise in the training dataset. An overtrained BDT will perform better with the training data than with a second, *testing*, sample. Figure 8 shows an overtraining plot of the BDT acting in Region 7. If the training histogram (dots) are at higher values than the testing histograms (shaded) then the BDT has been overtrained. Whilst this plot does show some of this behaviour, this is actually a good example of a BDT in which the overtraining level is acceptable. The amount of separation in the signal and background histograms on this plot corresponds to the classification power of the BDT. A larger separation is generally better here.

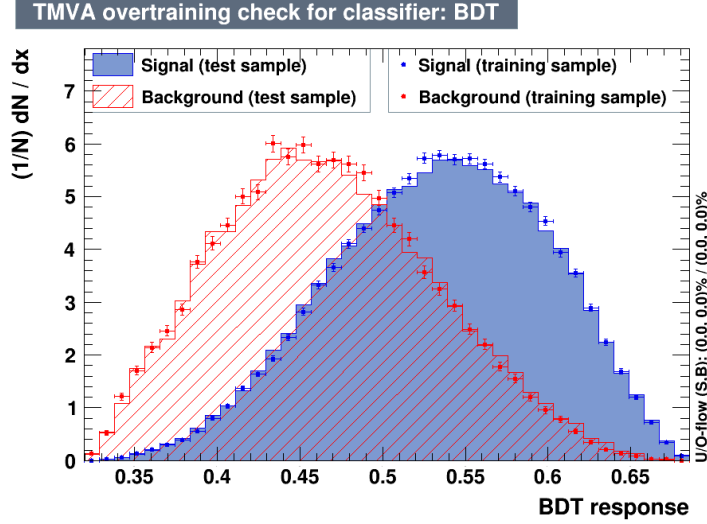


Figure 8: The effect of b-tagging variables in Region 7 for $t\bar{t} + b\bar{b}$ only backgrounds, as well as non- $t\bar{t} + b\bar{b}$ sourced backgrounds.

4. The MVA Setup

4.1. The MVA Variables

At present, the classification BDT uses just two types of variables: general kinematic variables, and variables from the Higgs reconstruction BDT.[3] In accordance with current analysis, these same variables were used in this investigation (all full list can be found in appendix A). In addition to these variables, the $mv2c10$ values from the jet b-tagging stage were also introduced. Whilst the b-tagging results had already been used to define the categorisation regions, it was postulated that these scores could still provide further signal vs background separation power.

For each given event, there are multiple jets. As such, the jets are sorted by $mv2c10$, with each jet providing an separate b-tagging variable for the event. The highest $mv2c10$ value would become that event's $mv2c10_1$, the second highest would fill $mv2c10_2$, and so on. These new variables are what are of particular interest for the classification BDT.

Consider now Region 7, with ≥ 6 jets and 3 b-tagged jets. In the optimal case, each signal event would have 4 b-tagged-jets. Hence, it is expected that the signal in this region will have a b-jet that was missed by the b-tagging. As such, one would imagine that a signal event's 4th highest $mv2c10$ value would still be high in this region, albeit just below the b-tagging cutoff. This is indeed what is seen in figure 9d. A peak is clearly seen around 0.8 in the signal data, that is not present in the background data. The preselection was done here at 70% b-tagging efficiency, so as expected this peak represents the jets that fell just below the tagging cutoff.

Testing the effectiveness of these b-tagging variables in the BDT revealed that both `mv2c10_3` and `mv2c10_4` were effective at improving the classifier’s performance. This is somewhat intuitive, since these two variables describe the jets that are nearest to the b-tagging cutoff. The remaining plots in figure 9 show the histograms of these variables. One can see that in all regions there is some separation between the signal and background plots of `mv2c10_4`. The separation in the `mv2c10_3` is much more modest, with the exception of in Region 7 where it looks to have some discriminating power. It was decided that both `mv2c10_3` and `mv2c10_4` would be investigated further as additional MVA variables.

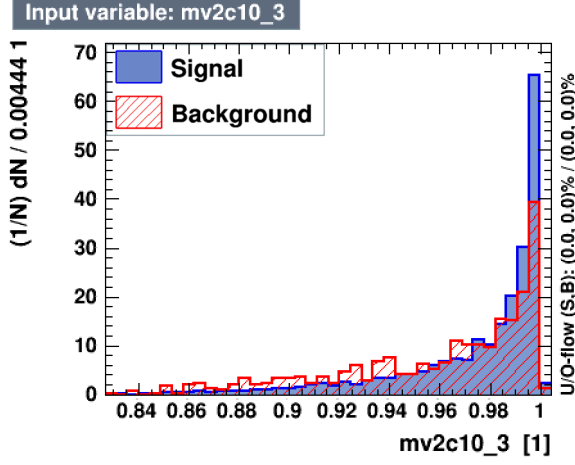
4.2. BDT Settings

Toolkit for Multivariate Analysis (TMVA) was used to generate the BDTs that were investigated. As with the basic set of variables used in this project, the BDT settings are the same as those used in the current analysis. These settings are as follows:

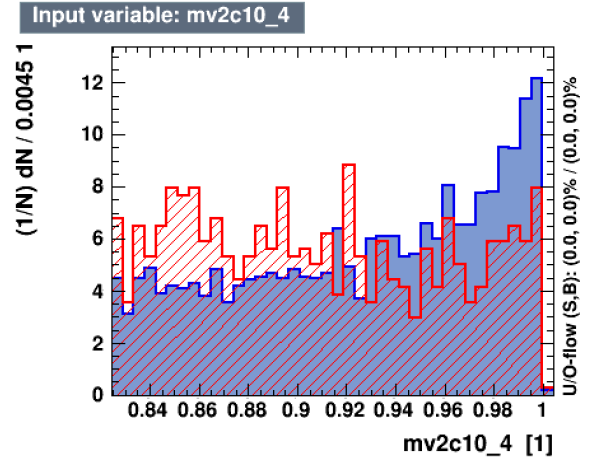
- Boost type: AdaBoost
- AdaBoostBeta: 0.15
- NTrees: 400 in Regions 7 & 8, and 250 in Region 5.
- MaxDepth: 5 in Regions 7 & 8, and 4 in Region 5.
- nCuts: 80
- MinNodeSize: 4% in Regions 7 & 8, and 5% in Region 5.

4.3. Testing the MVA

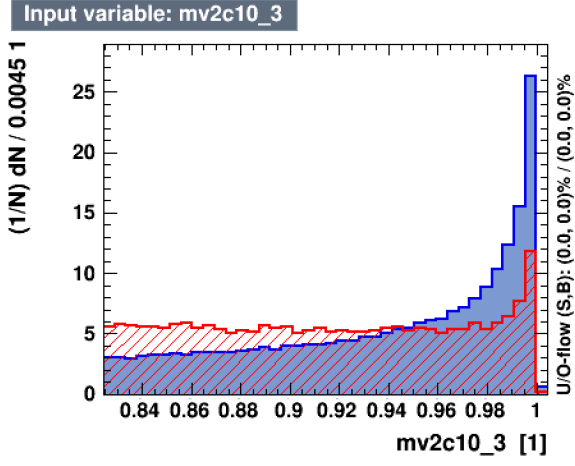
In order to test the MVA, the BDT was trained with a gradually increasing number of variables, starting with just 1 variable, all the way up the full set of 17 variables. Figure 10 shows the ROC curves for 4 of such BDTs in Region 7. Clearly, the BDT using just 1 variable is a very weak classifier, barely performing better than random guessing. However, as more variables are added, the ability of the BDT to distinguish signal and background events increases.



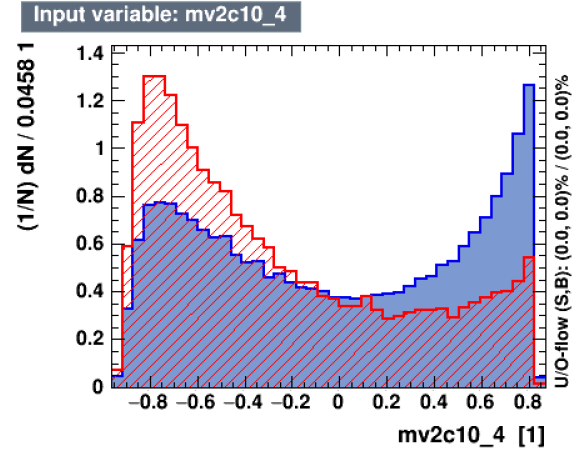
(a) Region 5: mv2c10_3



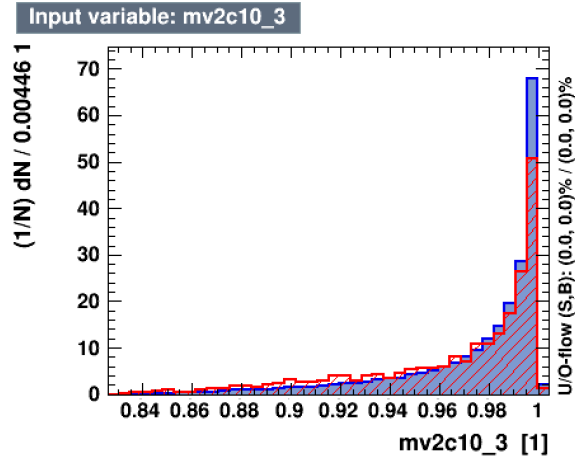
(b) Region 5: mv2c10_4



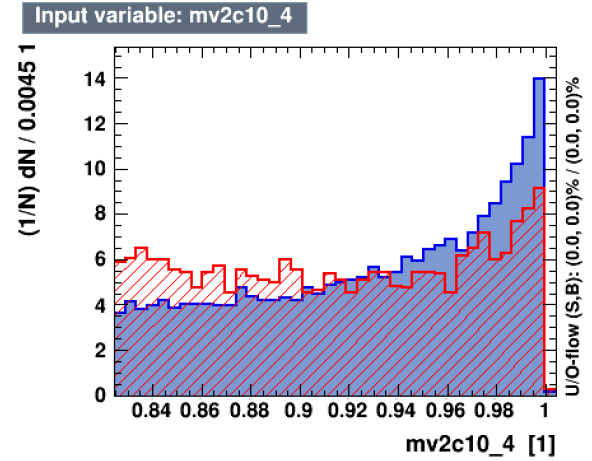
(c) Region 7: mv2c10_3



(d) Region 7: mv2c10_4



(e) Region 8: mv2c10_3



(f) Region 8: mv2c10_4

Figure 9: The 3rd and 4th highest b-tagging values, in each region of interest.

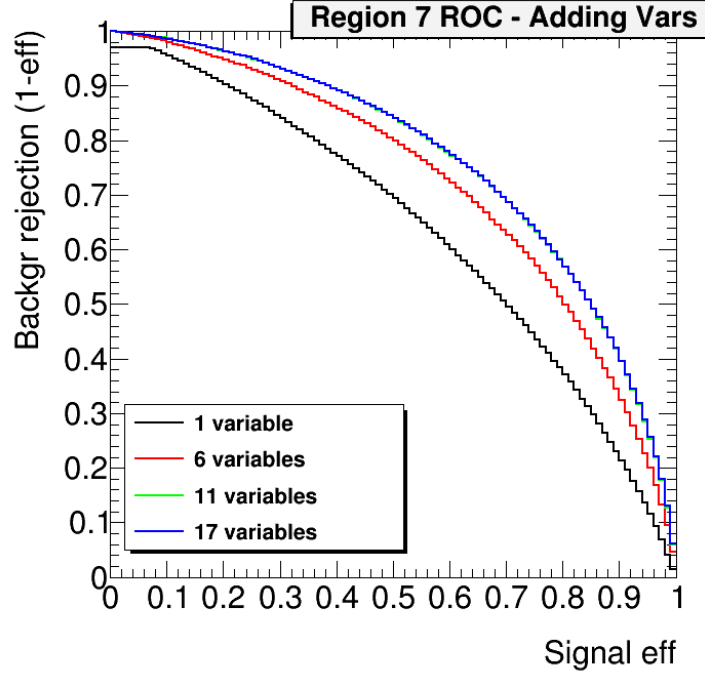


Figure 10: The effect of adding more variables to the analysis in Region 7.

In the figure, the green curve is mostly obscured by the blue curve. This suggests that the BDT did not improve with the addition of the final 6 variables. This was to be somewhat expected, since variables were added in decreasing order of importance. This importance rating is based on the structure of the final, 17 variable BDT. As such, the conclusion should be that less important variables improve the BDT less, not that over 11 variables is somehow disadvantageous. This can be further seen in the plot in appendix C, in which the variables are added in the opposite order. In said plot the converging behaviour of the plot no longer manifests. A further study could be one which investigates removing these less important variables.

Variable importance rankings are determined by comparing how often each of the variables appear in the decision trees. This calculation is a sum that counts the number of nodes a variable occurs in, weighted by the number of training events that reached that node and by the square of the separation gained by including said node. Hence, the importance ranking often differs from the separation ranking. Separation tests the variables independent of the MVA, whereas the importance ranks how strongly a variable features in the BDT.

5. Adding b-tagging Variables

5.1. ROC Improvements

Figure 11 shows how the analysis in Region 7 improved after both `mv2c10_3` and `mv2c10_4` were introduced to the MVA. Introducing these variables clearly improves the BDT's performance; much more so than many of the preexisting variables. Both the `mv2c10_3` and `mv2c10_4` variables appear to improve the classification by approximately the same amount when used separately, however the biggest improvement comes from adding them both.

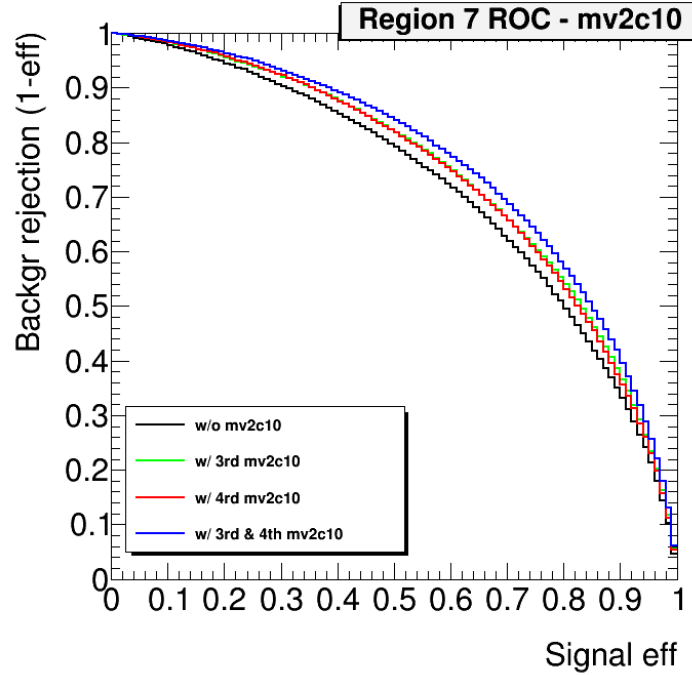
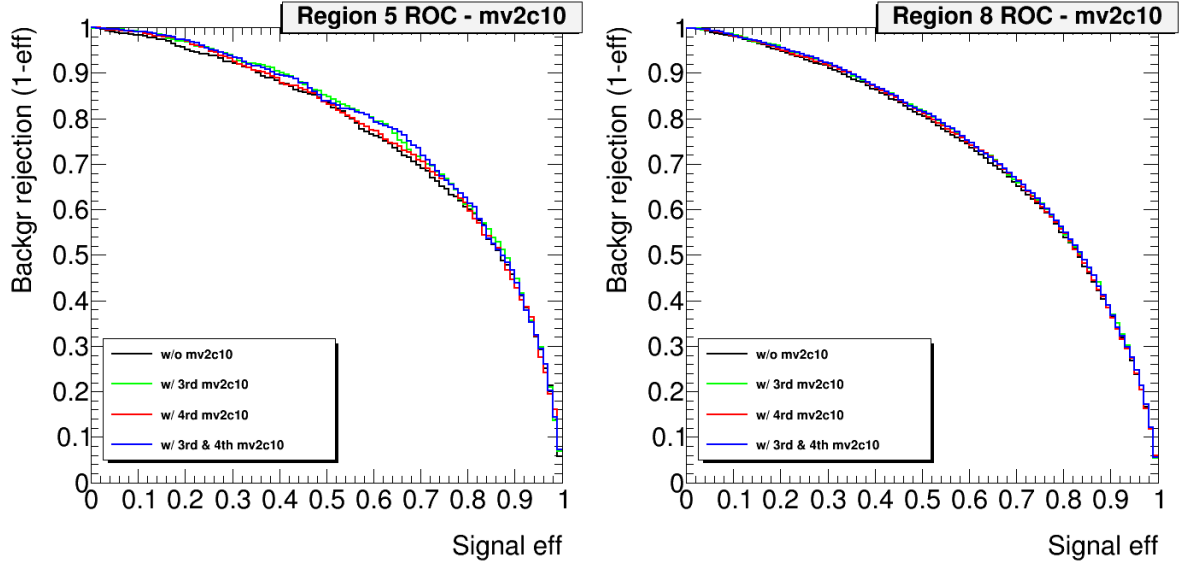


Figure 11: The effect of introducing b-tagging variables to the analysis in Region 7.

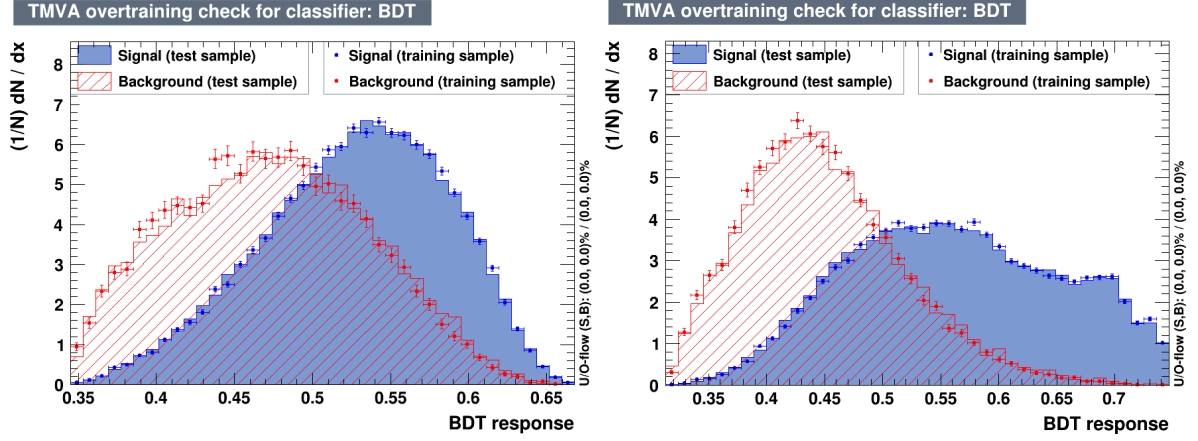
The highest BDT scoring bin in the currently ATLAS analysis, in Region 7, is at a signal efficiency of 18%.^[3] This corresponds to the bin with the highest statistical significance. Reading off from this plot we see that introducing the b-tagging variables increases the background rejection from $\sim 95\%$ to $\sim 97\%$. This means that adding these two b-tagging variables has almost halved our background in this bin, taking it from $\sim 5\%$ to $\sim 3\%$. Figures 12a and 12b show the effects of adding b-tagging variables to the analysis in Regions 5 and 8 respectively. The improvements here are much more modest than those seen in Region 7, however in both cases introducing `mv2c10_3` and `mv2c10_4` does improve the performance of the BDTs.



(a) Introducing b-tagging to Region 5.

(b) Introducing b-tagging to Region 8.

Figure 12: The introduction of b-tagging variables to the analysis in Regions 5 & 8.



(a) $t\bar{t} + b\bar{b}$ sourced background only.

(b) Non- $t\bar{t} + b\bar{b}$ sourced background.

Figure 13: How the background sources effect the overtraining plots in Region 7.

5.2. $t\bar{t} + b\bar{b}$ Background Study

In order to discover why the b-tagging variables were so effective, a study was conducted on the different components comprising Region 7's background. The pie chart for Region 7, shown earlier in figure 5, shows the different sources of Region 7's background. The proportion of $t\bar{t} + b\bar{b}$ sourced background is much less than half. This is in contrast to the backgrounds of Regions 5 & 8, in which the vast majority of the background comes from $t\bar{t} + b\bar{b}$ sources. It was therefore hypothesised that mv2c10_3 and mv2c10_4 best discriminate against the non- $t\bar{t} + b\bar{b}$ components of the background.

The two overtraining plots in figure 13 show the performance of the BDT with $t\bar{t} + b\bar{b}$ sourced background only, and without any $t\bar{t} + b\bar{b}$ sourced background at all. Note that the $t\bar{t} + b\bar{b}$ sourced background plot looks very similar to the overtraining plot seen earlier in figure 8. The plot excluding $t\bar{t} + b\bar{b}$ sourced background however looks very different. The BDT performs much better, and the signal proves much more separable than previously. The plots in this background study includes the b-tagging variables, which could be the source of improvement seen in the plot excluding the $t\bar{t} + b\bar{b}$ backgrounds.

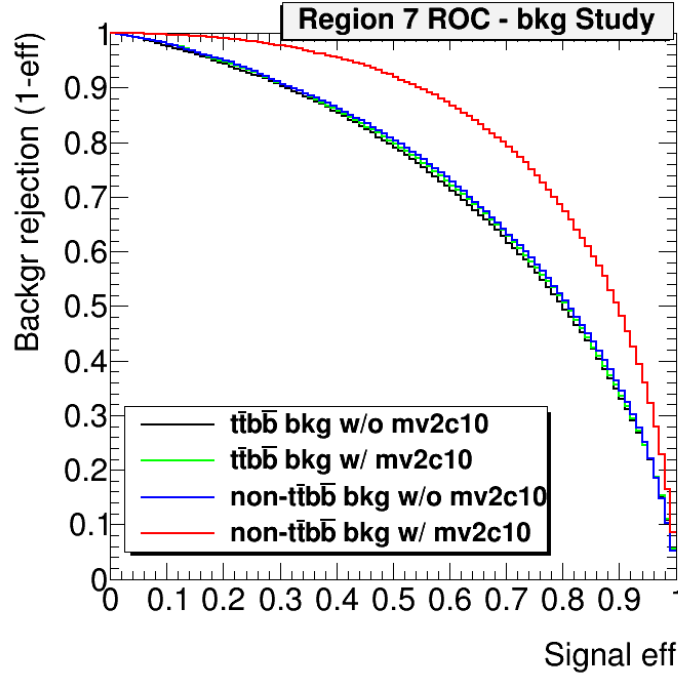
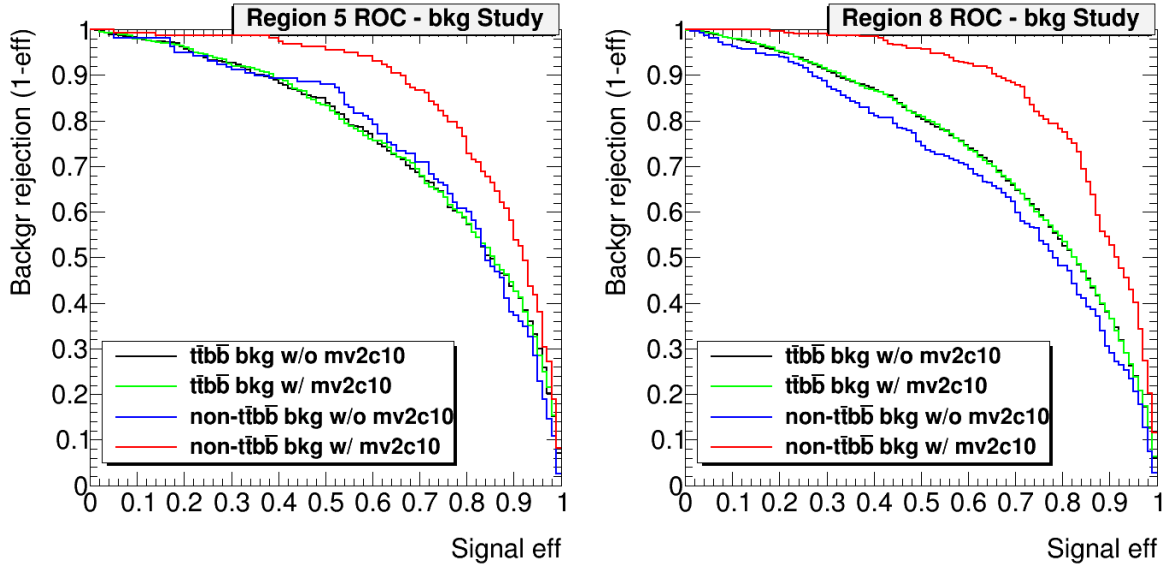


Figure 14: The effect of b-tagging variables in Region 7 for $t\bar{t} + b\bar{b}$ only backgrounds, as well as non- $t\bar{t} + b\bar{b}$ sourced backgrounds.

In order to truly discover the effect of the b-tagging variables on these backgrounds, BDTs were trained with and without the inclusion of mv2c10_3 and mv2c10_4. Figure 14 shows the ROC curves of in Region 7, for different background sources, with and without the b-tagging variables. Whilst there is a small improvement made to the $t\bar{t} + b\bar{b}$ only performance, it is only slight. The real improvement however comes with

the non- $t\bar{t} + b\bar{b}$ background. Here the BDT's performance increases dramatically once the b-tagging variables are added. This is logical, since we expect fewer b-jets in the non- $t\bar{t} + b\bar{b}$ backgrounds, which means that their mv2c10 are likely to be much lower. This is especially true for mv2c10_3 and mv2c10_4, since we would expect the b-jets sourced from the top quarks to occupy the mv2c10_1 and mv2c10_2 slots. The plots in figure 15 show similar results for Regions 5 and 8. The curves for non- $t\bar{t} + b\bar{b}$ are very noisy in these regions. It is assumed that this is due to the lower number of events available for these backgrounds, in these regions, however that has not been confirmed. The slight increase in performance seen by introducing the b-tagging variables to the Region 7 $t\bar{t} + b\bar{b}$ only study is not seen here. This could suggest that the b-tagging variables can only consistently improve the classification of signal against non- $t\bar{t} + b\bar{b}$ backgrounds.



(a) Introducing b-tagging variables to Region 5, for (b) Introducing b-tagging variables to Region 8, for different background sources.

Figure 15: The effect of b-tagging variables in Regions 5 & 8 for $t\bar{t} + b\bar{b}$ only backgrounds, as well as non- $t\bar{t} + b\bar{b}$ sourced backgrounds.

5.3. Binned b-tagging Variables

In order to have continuous b-tagging scores, the Monte Carlo data must be calibrated with the real detector data. This means that, in reality, it is somewhat infeasible to attain continuous b-tagging values. Instead, binned b-tagging scores are used. As the calibration is improved, the binning will become higher and higher resolution, however at present the binning is quite a low resolution.

An investigation was conducted to ascertain how these binned b-tagging variables would perform compared to the previously tested continuous ones. The binning used was as follows:

- Bin 0: $mv2c10 < \text{the 85\% b-tagging efficiency cut.}$
- Bin 1: $mv2c10 < \text{the 77\% b-tagging efficiency cut.}$
- Bin 2: $mv2c10 < \text{the 70\% b-tagging efficiency cut.}$
- Bin 3: $mv2c10 < \text{the 60\% b-tagging efficiency cut.}$
- Bin 4: $mv2c10 < \text{the 50\% b-tagging efficiency cut.}$
- Bin 5: $mv2c10 < \text{the 30\% b-tagging efficiency cut.}$
- Bin 6: $mv2c10 \geq \text{the 30\% b-tagging efficiency cut.}$

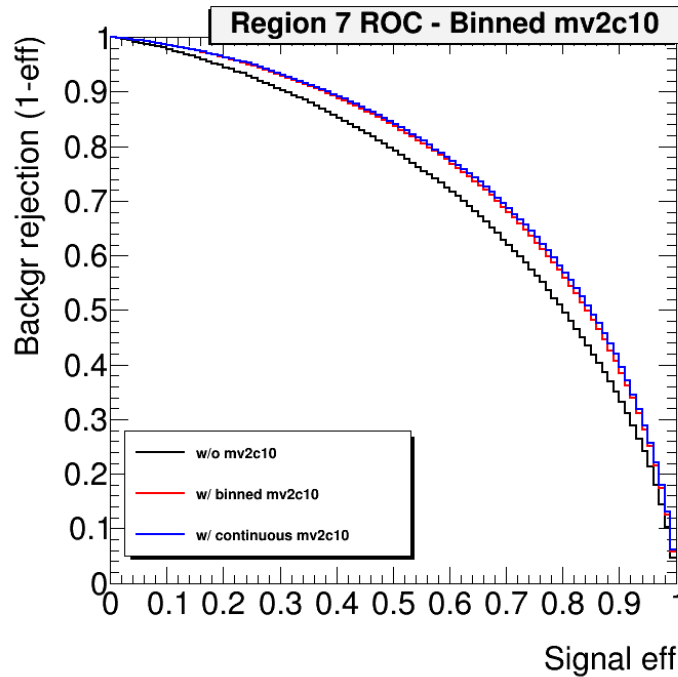


Figure 16: The effect of binned vs continuous b-tagging variables in Region 7.

The plot in figure 16 shows there is only a very small depreciation in BDT performance when the b-tagging variables are binned instead of continuous. This result is important as it means that b-tagging variables could potentially be added to the current BDT classifiers. The plot seen in figure 17 show that Regions 5 and 8 also perform similarly when the b-tagging variables are binned. The greatest difference is seen in the Region 5 plot. This indicates that a higher resolution binning than the one considered here is required for the performance to better match that of the continuous variables. Despite this, there is still a distinct improvement over not using the b-tagging variables at all.

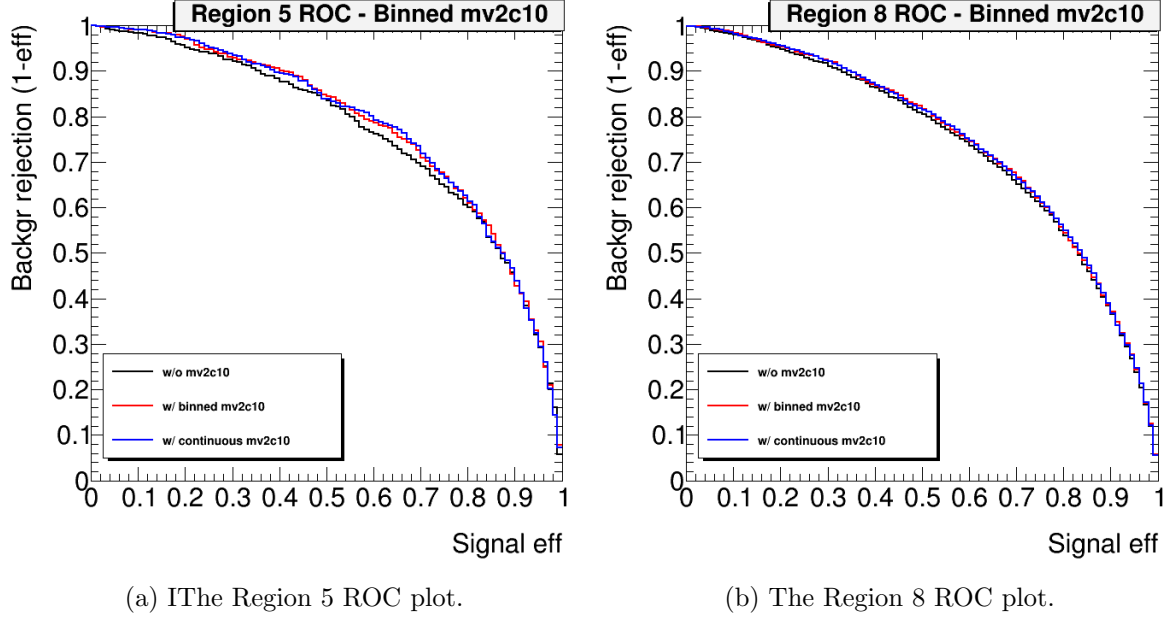


Figure 17: The effect of binned vs continuous b-tagging variables in Regions 5 & 8.

6. Conclusion

The possibility of including the mv2c10 b-tagging scores as additional MVA variables was investigated. It was shown that in Region 7, ≥ 6 jets with 3 b-jets, the background in the highest BDT score bin could be reduced by almost half. These variables are most effective in this region due to the high proportion of non- $t\bar{t} + b\bar{b}$ sourced backgrounds. Whilst the b-tagging variables do not provide much discriminating power against $t\bar{t} + b\bar{b}$ backgrounds, they were still able to improve the BDT performance in regions where the $t\bar{t} + b\bar{b}$ backgrounds feature heavily.

Currently the b-tagging variables attainable from real data are binned, as opposed to continuous. It was shown that binned b-tagging values can still improve the classification BDT. A further study would have to be conducted to investigate how the b-tagging variables perform at even lower binning resolutions, which are much closer to what is currently possible.

References

- [1] G. Aad *et al.* [ATLAS Collaboration], JINST **11** (2016) no.04, P04008 doi:10.1088/1748-0221/11/04/P04008 [arXiv:1512.01094 [hep-ex]].
- [2] Nazar Bartosik, “b-jet tagging,” http://bartosik.pp.ua/hep_sketches/btagging.
- [3] The ATLAS collaboration [ATLAS Collaboration], ATLAS-CONF-2016-080.

A. The Current Classification BDT Variables [3]

Variable	Definition	Region		
		$\geq 6j, \geq 4b$	$\geq 6j, 3b$	$5j, \geq 4b$
General kinematic variables				
$\Delta R_{\text{bb}}^{\text{avg}}$	Average ΔR for all b -tagged jet pairs	✓	✓	✓
$\Delta R_{\text{bb}}^{\text{max } p_T}$	ΔR between the two b -tagged jets with the largest vector sum p_T	✓	—	—
$\Delta \eta_{\text{jj}}^{\text{max}}$	Maximum $\Delta \eta$ between any two jets	✓	✓	✓
$m_{\text{bb}}^{\text{min } \Delta R}$	Mass of the combination of any two b -tagged jets with the smallest ΔR	✓	✓	—
$m_{\text{jj}}^{\text{min } \Delta R}$	Mass of the combination of any two jets with the smallest ΔR	—	—	✓
$m_{\text{bj}}^{\text{max } p_T}$	Mass of the combination of a b -tagged jet and any jet with the largest vector sum p_T	—	✓	—
$p_{\text{T}}^{\text{jet5}}$	p_T of fifth leading jet	✓	✓	✓
$N_{\text{bb}}^{\text{Higgs } 30}$	Number of b -jet pairs with invariant mass within 30 GeV of the Higgs boson mass	✓	—	✓
N_{40}^{jet}	Number of jets with $p_T \geq 40$ GeV	—	✓	—
$H_{\text{T}}^{\text{had}}$	Scalar sum of jet p_T	—	✓	✓
$\Delta R_{\text{lep-bb}}^{\text{min } \Delta R}$	ΔR between the lepton and the combination of the two b -tagged jets with the smallest ΔR	—	—	✓
Aplanarity	$1.5\lambda_2$, where λ_2 is the second eigenvalue of the momentum tensor built with all jets	✓	✓	✓
Centrality	Scalar sum of the p_T divided by the sum of the E for all jets and the lepton	✓	✓	✓
$H1$	Second Fox-Wolfram moment computed using all jets and the lepton	✓	✓	✓
Variables from Reconstruction BDT output				
BDT Output		✓	✓	✓
m_{H}	Higgs boson mass	✓	✓	✓
$m_{\text{H},b_{\text{lepton}}}$	Mass of Higgs boson and b -jet from leptonic top	✓	—	—
$\Delta R_{\text{Higgsbb}}$	ΔR between b -jets from the Higgs boson	✓	✓	✓
$\Delta R_{\text{H},t\bar{t}}$	ΔR between Higgs boson and $t\bar{t}$ system	✓	✓	✓
$\Delta R_{\text{H},\text{lepton}}$	ΔR between Higgs boson and leptonic top	✓	—	—
$\Delta R_{\text{H},b_{\text{hadtop}}}$	ΔR between Higgs boson and b -jet from hadronic top	—	✓	✓

B. mv2c10 b-tagging Cuts

b-jet Efficiency [%]	Purity [%]	Weight Cut
29.99	99.95	0.9977155
50.05	99.62	0.9769329
60.03	99.00	0.934906
69.97	97.46	0.8244273
76.97	95.17	0.645925
84.95	89.66	0.1758475

C. Adding Variables in Order of Least Importance

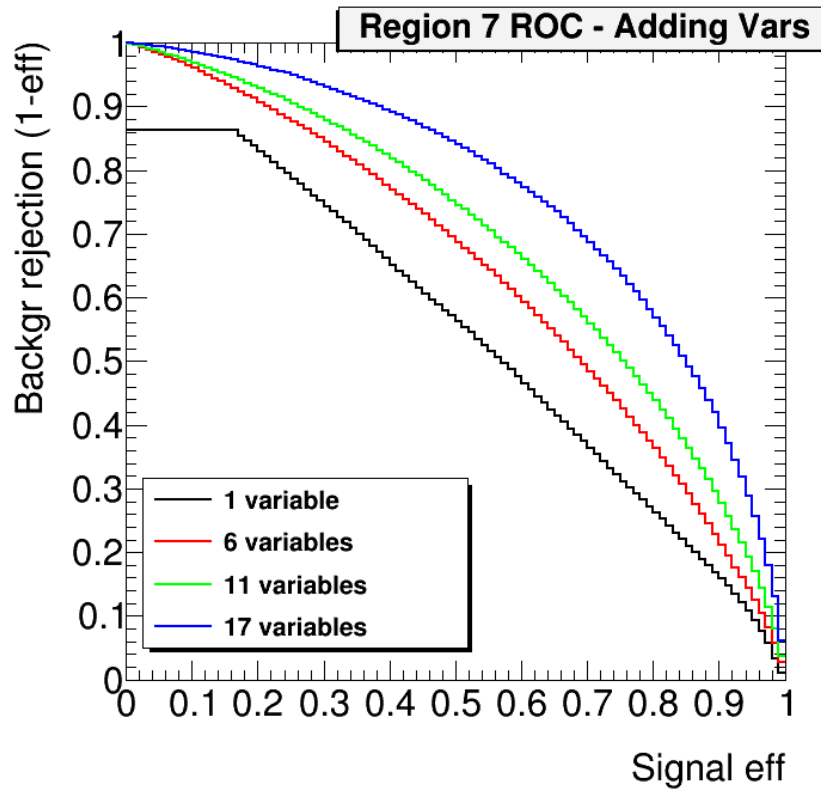


Figure 18: The effect of adding more variables to the analysis in Region 7. Unlike in figure 10, the variables were added in ascending order of importance, instead of descending.