Enhancing $ZH \rightarrow \ell\ell b\bar{b}$ Searches with Multiple Interpretations in the ATLAS Detector

A THESIS PRESENTED
BY
ROGER HUANG
TO
THE DEPARTMENT OF PHYSICS
AND
THE DEPARTMENT OF COMPUTER SCIENCE

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE JOINT DEGREE OF
BACHELOR OF ARTS
IN THE SUBJECTS OF
PHYSICS AND COMPUTER SCIENCE

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
APRIL 2016

Physics Thesis advisor: Professor John Huth
CS Thesis advisor: Professor David Parkes

Roger Huang

Enhancing $ZH \to \ell\ell b\bar{b}$ Searches with Multiple Interpretations in the ATLAS Detector

Abstract

This thesis presents an attempt at using multiple event interpretations to enhance the boosted decision tree used in the search for the Standard Model Higgs boson decay to the $b\bar{b}$ final state at a Higgs mass of $m_H = 125$ GeV with the ATLAS detector, looking in particular at the $ZH \to \ell\ell b\bar{b}$ process. Recent studies have suggested that using multiple event interpretations offered by constructing jets with different radii can offer a large improvement over the traditional deterministic approach to event interpretation. Monte Carlo datasets generated using a center of mass energy $\sqrt{s} = 13$ TeV are used for this analysis, and the data's truth labels are used to evaluate the performance of the boosted decision tree. Using a $S/\sqrt{S+B}$ measure of significance for the boosted decision tree, improvements of 1.9% in the $p_T^V < 120$ GeV region and 6.5% in the $p_T^V > 120$ GeV region are found by using additional inputs derived from the multiple event interpretations, both of which are smaller than the expected improvements from previous work.

Contents

List of figures

List of tables

Acknowledgments

I would like to express my thanks to:

John Huth, for taking me into his group and giving me the opportunity to do research in high-energy physics, and for his advice and guidance in this thesis and in the rest of my academic career.

David Lopez Mateos, for introducing me to this particular topic and making this dissertation possible, and for his mentorship from my first summer at CERN all the way until now. I am deeply indebted to him for the time he has taken to answer my questions, help me understand the problems I have come across, and guide me in my work.

Masahiro Morii, for his role as my academic advisor, where he has helped guide me through my time as an undergraduate physics concentrator at Harvard.

The rest of the Harvard ATLAS group, from whom I have learned a lot in our day-to-day conversations, and who helped make my time at CERN enjoyable.

David Parkes, who also advised this thesis, for his help and advice with the concepts and methodologies used on the machine learning side.

My friends at Harvard, who have made my time here enjoyable.

And my parents, for their advice and unconditional support in all my endeavors.

1

Introduction

This thesis describes and implements the technique of multiple interpretations to enhance the significance of the results of the search with the ATLAS experiment at the Large Hadron Collider (LHC) for the Standard Model Higgs Boson through the associated production $pp \rightarrow VH$, with the Higgs decaying via $H \rightarrow b\bar{b}$. It specifically attempts to improve the results of the boosted decision tree used in the last analysis searching for this process [1] with the addition of new physical observables constructed through the idea of multiple interpretations. The goal is to investigate how the use of these new multiple interpretations inspired variables as additional input to the boosted decision tree affects its ability to differentiate between signal and background events at reconstructed-level, using data from Monte Carlo simulations of physical events to do so. Chapter 2 describes the details of the datasets used, as well as what event reconstruction means and how it is done. Chapter 3 describes how candidate events are selected out of the entire dataset and describes all of the physical observables that are used as input variables to the boosted decision tree. Chapter 4 describes the boosted decision tree itself and how it is trained and evaluated. Finally, the results are presented in Chapter 5, with concluding thoughts in Chapter 6.

This introductory chapter will first outline the background behind the work in this thesis. Section 1.1 gives a high-level theoretical overview of the Standard Model and how the Higgs particle is relevant. Section 1.2 describes the discovery of the Higgs boson at the LHC. Section 1.3 describes the idea of multiple interpretations, as well as past work using this method.

1.1 The Standard Model

The Standard Model is our current theory describing all observed matter and fundamental forces besides gravity. According to the Standard Model, all matter is made of 3 elementary particles: leptons, quarks, and mediators. Elementary particles in the Standard Model all carry their own intrinsic angular momentum, quantized in units of the reduced Planck's constant $\hbar$. They can be categorized as bosons, which carry integer spin, and fermions, which carry 1/2-spin. Leptons and quarks are fermions, distinguished from each other foremost by the fact that leptons do not interact with the strong force but quarks do. All charged fermions interact through the electromagnetic force, and all fermions regardless of charge interact with the weak force. There are 3 generations of leptons and quarks in total, and each one also has an associated antiparticle of the same mass and opposite charge.

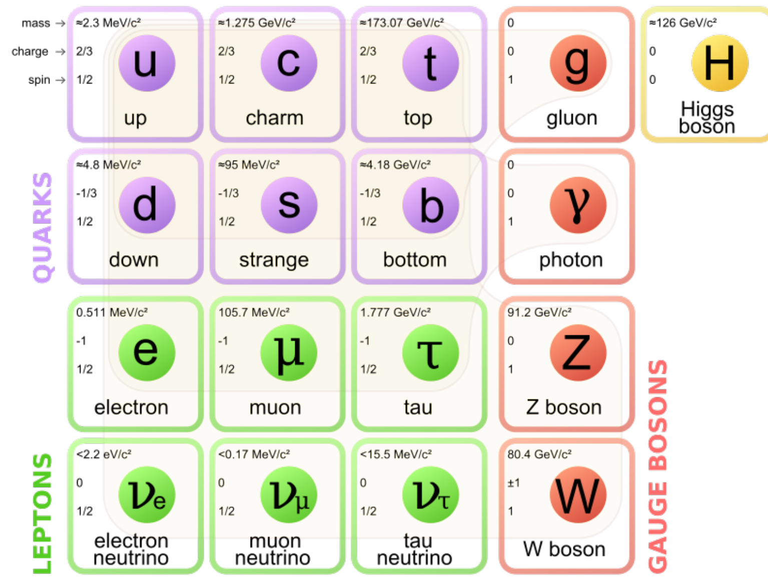

Figure 1.1: The particles of the Standard model [2].

The force mediators are bosons and carry the fundamental forces; they are also called gauge bosons. The electromagnetic force is carried by the massless photons, the weak force is carried by the massive W and Z bosons, and the strong force is carried by the massless gluons. However, the Standard Model says that the 4 bosons in the electroweak sector are all massless before the electroweak symmetry is broken. The theory then calls for at least one spin-0 Higgs particle, which is responsible for breaking the electroweak symmetry and giving the W and Z bosons their mass, as well as the mass of all other massive particles. These particles with their basic properties are pictured in Figure 1.1.

The Standard Model has survived every experimental result so far, successfully predicting and explaining the observed behavior of particles and their interactions in our experiments. Until recently, the one significant remaining prediction of the Standard Model that had not yet been experimentally confirmed was the Higgs particle. In the Standard Model, the mass of the Higgs boson (along with many other constants, including the masses of leptons and quarks) is a free parameter, and though there were some theoretical limitations on its possible values, the precise value had to be found by experiment. This last piece of the Standard Model was confirmed to be found at the LHC at CERN in 2012 [3, 4].

1.2 Higgs Boson Discovery at the LHC

One of the primary purposes of the LHC was to discover the Higgs boson and measure its properties. It aimed to do this by colliding protons at unprecedentedly high energies and looking at the resulting collision products. During Run 1 from 2009 to 2013, the LHC was run at center of mass energies of $\sqrt{s} = 7$ TeV and $\sqrt{s} = 8$ TeV. In 2012, the discovery of a new massive boson was reported at the LHC, later confirmed to be the Higgs boson. The ATLAS experiment reported the result at a significance of 5.9 standard deviations [3], and the CMS experiment reported a significance of 5.1 standard deviations [4]. Using Run 1 data in the $H \to \gamma\gamma$ (Higgs decay to 2 photons) and $H \to ZZ \to 4\ell$ (Higgs decay to 2 Z bosons decaying to 4 leptons) channels, the ATLAS and CMS collaborations later more precisely placed the mass of the Higgs boson at $125.09 \pm 0.21(\text{stat.}) \pm 0.11(\text{sys.})$ GeV [5].

Before the mass of the Higgs boson was discovered, the branching ratios for each of the possible decay channels could only be calculated as a function of its mass, shown in Figure 1.2. Now that its mass has been measured, data from the LHC has been used to confirm these branching ratios in many channels. At the measured mass, the decay to a bottom and antibottom quark ($H \to b\bar{b}$) is the dominant decay channel, but the ATLAS and CMS experiments have so far not detected this at rates high enough to match the Standard Model prediction, due largely to the high backgrounds in searching for this interaction [1].

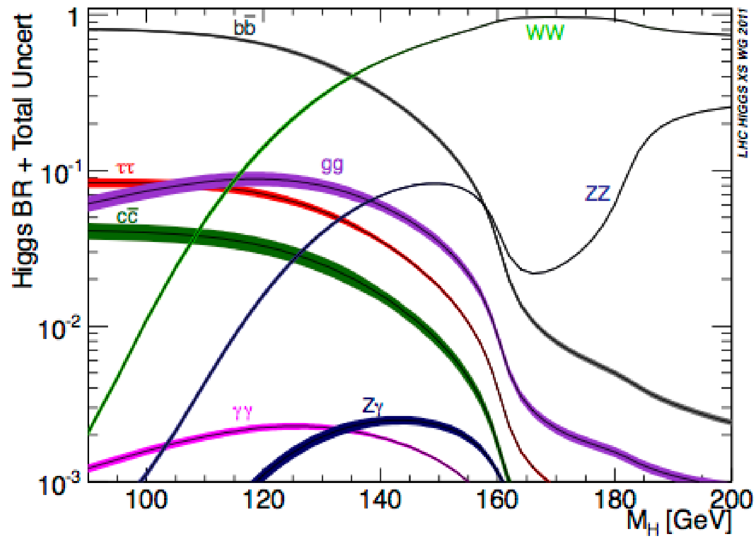

Figure 1.2: The branching ratios for each of the possible Higgs decay channels as a function of the Higgs mass. For the discovered mass of roughly 125 GeV, the $H \rightarrow b\bar{b}$ decay mode is dominant [6].

While it is expected that this decay will be detected with the Run 2 data at $\sqrt{s} = 13$ TeV that is currently being collected, new methods are needed to improve our ability to separate background from the events we are looking for. This thesis will build on the search for the Higgs $b\bar{b}$ final state after associated production of the Higgs with a vector boson W or Z, focusing in particular on the $ZH \rightarrow \ell\ell b\bar{b}$ process.

1.3 Multiple Interpretations

When the proton beams collide in the LHC, a single collision is considered to be a single event for use in analysis. To determine whether the event is of the desired type, we look at various physical observables of the event, which must be reconstructed by various methods from the raw data collected by the detector. Traditionally, a single way to construct these physical observables is selected, leaving only one way to interpret the event. However, there are new methods proposed that would exploit multiple interpretations of a single event instead of using the traditional deterministic approach. The Qjets method proposed in [7] looks at how an observable varies depending on certain parameters used in the event reconstruction, offering multiple interpretations of the event with these variations.

In particular, the properties of jets can depend quite heavily on the jet algorithms used and the parameters used in the algorithms. Jets are reconstruction-level objects that group together showers of particles that have deposited their energy in the detector. For instance, in the search for the $b\bar{b}$ final state we look for two jets that we determine to have both originated from b quarks, and use these jets to reconstruct variables like the dijet mass (details of how jets are defined, constructed, and tagged can be found in Sections 2.2 and 2.3). The reconstructed dijet mass will vary with the jet radius used in reconstruction, since larger radii will include more particles, as pictured in Figure 1.3. By constructing jets in multiple different ways for the same group of detector hits, we get multiple possible values for a single observable, and so multiple ways to look at and interpret an event. Sizable improvements in significance were found in [8] and [9] by varying how jets were created in event reconstruction to yield multiple interpretations.

The work done in [10] with the $ZH \to \ell\ell b\bar{b}$ process found an improvement of 21% in significance by using multiple interpretations instead of a single-dimensional cut on the dijet mass from 2 b jets using Monte Carlo data. It did this by constructing multiple jets of different radii around each jet axis to create multiple interpretations of each event, giving a different dijet mass for each radius used, and then applying a multi-dimensional cut based on the multiple possibilities for the dijet mass. The work in [11] found that at truth level

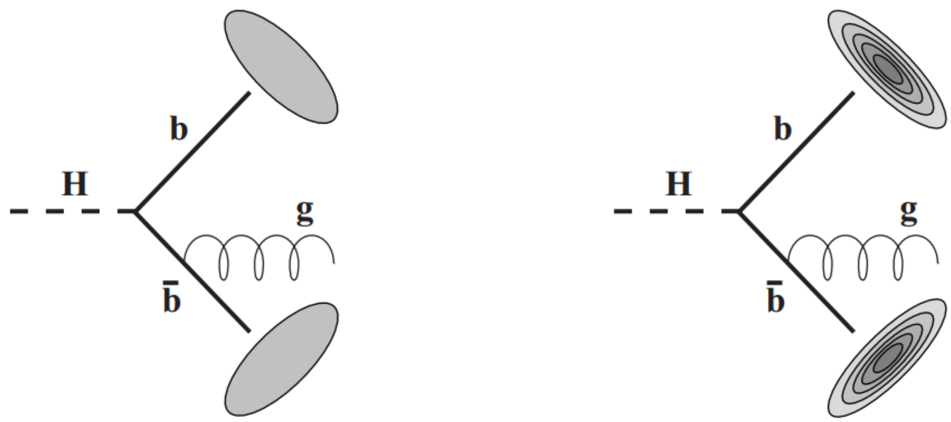

Figure 1.3: Using different jet radii to interpret a single event. Different radii include different information within each jet, giving different results when physical observables are reconstructed from the jets.

(directly accessing the properties of the simulated particles from the Monte Carlo data), using the multiple interpretations offered by different jet radii in addition to other kinematic variables in a boosted decision tree can give improvements of up to 12% in the low p_T region and 20% in the high p_T region over using these other kinematic variables with just a single jet radius. When the actual analysis is done with real data from the detector, it will have to use variables at reconstructed-level, where it must do the full process of converting data from the detector into physical objects. Additionally, it will use a boosted decision tree (BDT) in the multi-variate analysis, whose details can be found in Section 4.1. Because of this, the results of these two studies are promising, but they are not direct evidence that the multiple interpretations will be helpful in the real analysis. This thesis aims to build upon [10] and [11] by applying the method of multiple interpretations using different jet radii at reconstructed level to a boosted decision tree, using Monte Carlo data passed through the ATLAS detector simulation and closely emulating the methods of the actual analysis so that the results may be directly applicable. The rest of this thesis will describe the datasets used, the event selection and reconstruction process, the BDT training process, and the results of the study.

2

Datasets and Event Reconstruction

Event reconstruction is the process by which signals from the ATLAS detector are converted into physics objects. This includes reconstructing muons, electrons, jets, and missing energy through a variety of algorithms, all of which are necessary for this analysis. Although this thesis uses Monte Carlo data, it looks at the data at the reconstructed level for reasons described in the previous chapter, and so this chapter provides a summary of the datasets used in the analysis and the methods used to reconstruct all these physics objects. Section 2.1 describes the datasets used and the Monte Carlo methods used to generate them. Section 2.2 describes how jets are reconstructed and calibrated. Section 2.3 describes the algorithm used for *b*-tagging the jets. Section 2.4 summarizes how the remaining objects (electrons, muons, and missing energy) are constructed.

2.1 Monte Carlo Simulation and Datasets

The analysis in this thesis is done with data generated by Monte Carlo methods, which simulate physics events and objects. These objects are then run through a simulation of the ATLAS detector [12] based on GEANT-4 [13], which is a toolkit for simulating the passage of particles through matter. The output of this simulation is of the same form as that which is used for real data coming in from the detector, allowing us to see the simulated events as the ATLAS detector would. Signal and background processes for proton-proton collisions are generated at center of mass energy $\sqrt{s} = 13$ TeV, to correspond to the Run 2 data currently being taken. The signals and backgrounds are normalized to a luminosity of 30 fb^{-1}, which is a rough estimate of the amount of data that is anticipated to be taken in 2016. The signal sample is the $ZH \rightarrow \ell\ell b\bar{b}$ process discussed previously, generated in Pythia8 [14] configured with the AU2 tune [15] using the CTEQ6L1 parton distribution functions [16], and using a Higgs mass of $m_H = 125$ GeV. Since this signal is characterized by the detection of two b-tagged jets in association with two leptons from the Z decay, the primary background process is Z+jets, with the jets possibly originating from any of b, c, or light (u, d, or s) quarks. The background samples for this process are generated in Sherpa [17] for $Z \rightarrow e^+e^-$ and $Z \rightarrow \mu^+\mu^-$ for all of these possible jets, and for the transverse momentum p_T of the Z boson ranging from 0 to 1000 GeV. In addition, $t\bar{t}$ samples generated in the Powheg generator [18] were also included in the background processes, as were the diboson processes $WZ \rightarrow qq\ell\ell$ and $ZZ \rightarrow qq\ell\ell$ generated using Sherpa.

2.2 Jet Reconstruction and Calibration

Jets are detector-level objects that are seen when high-energy partons formed from the initial proton-proton collision shower into many lower energy particles, which are then picked up in the detector's calorimeter. This process can be seen in Figure 2.1, which depicts a sample event in the detector with the reconstructed jets. The energy deposits in the individual calorimeter cells are then grouped into topological clusters [19], from which jets are constructed by some choice of algorithm. Clusters are created by first picking a high-threshold energy cell as the seed, and then adding neighboring cells whose energy exceed some threshold repeatedly, forming a 3-dimensional cluster of cells. This is repeated for all of the seed cells until everything has been clustered appropriately. Due to the fact that particles do not always deposit all of their energy in the calorimeters, there are additional calibrations that need to be performed on the energies of the cells comprising these clusters. The two primary calibration methods are ElectroMagnetic scale (EM) and Local Calibration scale (LC), both of which are used in this analysis for practical reasons. The LC calibration brings the cluster energy closer to the actual amount of energy deposited, while the EM calibration gives a more accurate reading of the energy of the electrons and photons.

These topological clusters are then used as input to the anti-k_t jet clustering algorithm [21]. The anti-k_t algorithm takes a radius R as an input parameter, and then iterates over each cluster, attempting to merge it with other remaining nearby clusters based on a distance parameter dependent on the input radius R. Using a larger R results in each individual jet including more of the topological clusters in the calorimeter, which can capture additional information but can also result in the inclusion of clusters that did not originate from the same parton in truth. Previous searches for the $H \rightarrow b\bar{b}$ decay [22] have

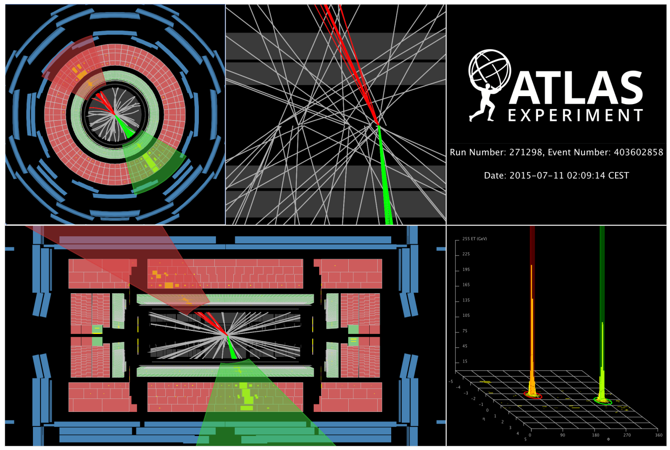

Figure 2.1: A sample event from the ATLAS detector. The top left is a transverse view of the detector, and the bottom left is a longitudinal view. The red and green shaded regions show the reconstructed jets, and the yellow markings are hits in the calorimeter. The bottom right shows the energy readings in the calorimeter for different η and φ. The gray marks are the reconstructed tracks, and the red and green highlighted ones are tracks belonging to particles included in the two shown jets. [20]

used R=0.4 for jet reconstruction as the optimal radius. In this analysis, jets are constructed for radius R=0.2 through R=0.8 at intervals of 0.1, selected because these are the radii for which LC calibrations currently exist. The R=0.4 jets are constructed with both the EM and LC calibrations, and the other radii are constructed with the LC calibration only, as EM calibrations do not currently exist for the other radii.

2.3 b-Tagging

Because the final state in this analysis includes 2 b-jets from which the Higgs is reconstructed, it is highly reliant on being able to accurately b-tag jets. This entails having high efficiency in identifying jets that originated from B-hadrons (the hadrons formed after hadronization of a b quark) as well as high efficiency in rejecting jets that did not. B-hadrons have two particularly distinguishing characteristics. They have relatively long lifetimes, which causes the secondary vertex reconstructed from the resulting jets to be significantly displaced from the primary vertex where they are formed, and they have relatively high masses, which results in a greater angular spread in their resulting jets. These properties are exploited by a variety of algorithms, such as IP3D, SV1, and JetFitter [23].

This analysis uses the MV2 algorithm, which combines the outputs of the 3 previously mentioned algorithms using a boosted decision tree, and which offers an improvement in performance over the MV1 algorithm used for *b*-tagging in the Run 1 analysis [22]. Specifically, *b*-tagging is done using MV2c20, which is the result of the MV2 algorithm when trained on a sample with *b*-jets as signal and a mixture of 80% light-flavored jets and 20% *c*-jets as background, offering better rejection of *c*-jets than versions trained with a higher percentage of light-flavored jets. The MV2c20 algorithm's output is a value on a continuum from -1 to 1, which can be interpreted as the likelihood that a jet is a *b*-jet. A cut is applied at a point to optimize the balance between efficiency of *b*-jet identification and efficiency of rejecting other jets; jets with a MV2c20 score above this cutoff are then classified as *b* jets. Further details on the tagging algorithm can be found in the *b*-tagging section of [24]. The datasets being used in this analysis only contain the MV2c20 score for the anti-k_t EM-calibrated jets for R=0.4; the other jets are dealt with by the method described in Section 3.2.

2.4 Electrons, Muons, and Missing Energy

Electrons in the ATLAS detector are reconstructed using data from the inner detector and the electromagnetic calorimeter, the details of which can be found in [25]. Muons are reconstructed using data from both of the prior detectors in addition to the muon spectrometer. Both types of leptons are subject to 3 tiers of quality controls, called loose, medium, and tight, whose details can be found in [22]. Each tier imposes additional constraints on the reconstructed lepton, cutting out more potential candidates and suppressing additional backgrounds. In this analysis looking for decay products of the Z boson, out of the two selected leptons, one is required to pass the medium requirement while the other is required to pass the loose requirement.

Though the ATLAS detector can absorb all electrons, photons, and jets, and can accurately measure muons through the muon spectrometer, it cannot directly detect neutrinos at all due to their extremely weakly interacting nature. The energy and momentum of neutrinos (in addition to other non-interacting particles) can instead be inferred by looking at the reconstructed transverse energy and momentum from all other objects and other soft components that were not classified as part of another object; conservation of momentum then gives us a quantity called missing transverse energy (MET or E_T^{miss}). This quantity is thus dependent on the quality of the calibrations of everything else in the detector.

3

Event Selection and Variables

Event selection is the process by which the entire list of events is narrowed down to a smaller list of candidate events in a way that suppresses certain backgrounds and ensures the candidate events were not the result of uncertain or poor-quality reconstructions. This chapter describes how the various physics objects used in the analysis are defined and used in the selection requirements. Section 3.1 defines the physics objects used in the selection and the criteria imposed upon them for use as candidate events. Section 3.2 describes how the reconstructed jets of different radii are matched to each other as different reconstructions of the "same" jet. Lastly, section 3.3 lists the variables that are used in the multivariate analysis to separate signal from background.

3.1 Object Definition and Event Selection

Only electrons and muons are considered as the lepton candidates, and they only pass event selection if one passes at least the medium requirement and the other passes at least the loose requirement, as described in Section 2.4. In addition, there is a requirement that the dilepton mass m_{ll} be between 71 and 121 GeV. This is slightly wider than the regular cut of 83 to 99 GeV, because the BDT is expected to be able to use the m_{ll} information to predict the magnitude of the $t\bar{t}$ background.

The reconstructed R=0.4 jets are split into two exclusive categories of signal jets and forward jets, whose criteria are shown in Table 3.1. Only signal jets are considered as candidates for the results for the Higgs decay. $\text{jet}_{i=1}$ denotes the leading jet, with $i > 1$ denoting the other jets. The leading jet has greater transverse momentum p_T than the jets with $i > 1$, with these other jets ranked by their transverse momentums as well.

Variable	Signal Jet	Forward Jet
p_T (GeV)	$\text{jet}_{i=1} > 45$ $\text{jet}_{i>1} > 20$	>30
$\lvert\eta\rvert$	< 2.5	2.5 < 4.5

Table 3.1: Transverse momentum p_T and pseudorapidity η requirements for signal and forward jets

In b-jet identification, there are again 3 tiers of quality called loose, medium, and tight, depending on what value of the MV2c20 algorithm is chosen as the cutoff. These correspond respectively to 80%, 70%, and 50% identification efficiency of b-jets, where the trade-off of higher efficiency identification is an increase in false-positives from misidentified light-flavor or c jets. b-tagging is only done for the EM-calibrated R=0.4 jets, and an event must have at least 2 R=0.4 signal jets passing the medium (70% efficiency) b-tagging requirement to be considered.

3.2 Matching Jets of Different Radii

Because the datasets used in this analysis only have b-tagging information for the R=0.4 EM-calibrated jets, a method is required to match all of the LC-calibrated reconstructed jets with R=0.2 through R=0.8 to these b-tagged jets. Intuitively, this is an attempt to match up the jets of different radii as different ways of reconstructing the "same" jet in the detector, and treating them as either all b-tagged or not b-tagged depending on the R=0.4 EM-calibrated jet in the set of matched jets.

To do this, first the two leading b-tagged R=0.4 EM-calibrated jets are identified. Then for the LC-calibrated jets of each other radius, the jets that are closest to the previously mentioned b-jets (with distance defined by $(\Delta R)^2 = (\Delta\eta)^2 + (\Delta\varphi)^2$ between the jets) are found, and these are then treated as the matched jets. If the two b-tagged jets are matched to the same LC-calibrated jet for any particular R by this method, then the event is rejected. When this occurs, it generally means that when the jets are reconstructed at a larger radius, the jet algorithm "merges" two jets of smaller radius into one, causing them to get matched to the same jet by this method. The acceptance rates using these criteria are shown in Figure 3.1. The acceptance rates peak around R=0.4 as expected, and fall as the used jet radius R deviates more from the R=0.4 comparison point. The smallest R jets likely have higher rejection rates because the R=0.4 jets can be split into two or more smaller jets in a way that causes the matching to fail, and the largest R jets likely have high rejection rates because the

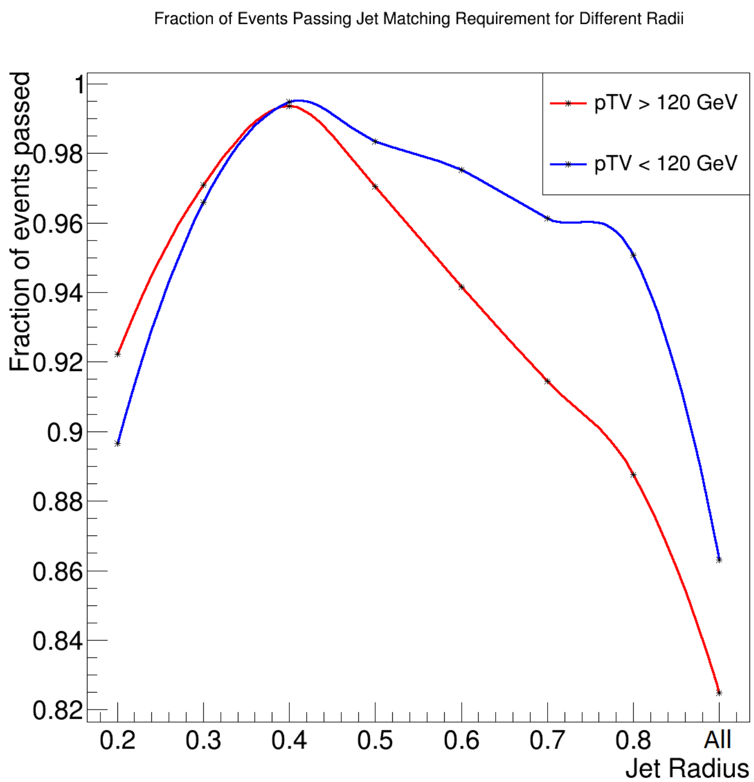

Figure 3.1: Fraction of events that pass the described jet matching requirement, for both low and high p_T^V, for each radius R of LC-calibrated jets. The right-most point denotes the fraction of events passed when the jet matching requirement is imposed for all radii from R=0.2 to R=0.8.

two R=0.4 jets might get merged into one larger one. These rates are considered to be acceptable under the assumption that the rejections occur somewhat randomly. However, this is not necessarily a safe assumption, and for future analyses it would be better to have *b*-tagging information for every jet radius so that they can be selected independently, and events don't have to be rejected by this method.

3.3 Variables Used in Multivariate Analysis

Variable	Name	Description
m_{jj}	mBB	Higgs Boson mass reconstructed from the two leading b-jets
$\Delta R(jet_1, jet_2)$	dRBB	ΔR between the two leading b-jets
$p_T^{jet_1}$	pTB1	Transverse momentum p_T of the leading b jet
$p_T^{jet_2}$	pTB2	Transverse momentum p_T of the subleading b jet
$\|\Delta\eta(jet_1, jet_2)\|$	dEtaBB	Difference in pseudorapidity $\|\Delta\eta\|$ between the two leading b jets
$MV2c20(jet_1)$	MV2cB1	Output of MV2c20 algorithm on leading b jet
$MV2c20(jet_2)$	MV2cB2	Output of MV2c20 algorithm on subleading b jet
p_T^V	pTV	Transverse momentum p_T of the Z boson
$m_{\ell\ell}$	mLL	Mass of the Z boson reconstructed from the two leptons
$\|\Delta\Phi(V, H)\|$	dPhiVBB	Difference in azimuth $\|\Delta\Phi\|$ between the Z boson and the Higgs boson
$\|\Delta\eta(V, H)\|$	dEtaVBB	Difference in pseudorapidity $\|\Delta\eta\|$ between the Z boson and the Higgs boson
E_T^{miss}	MET	Missing transverse energy

Table 3.2: Variables used in BDT from Run 1 analysis [22]. These are used again in this analysis to provide a base level of performance.

Multivariate analysis methods take in a number of user-defined input variables and use these to make predictions about data. The variables chosen for use in the BDT in Section 4 are typical physical observables for an event, that are expected to have high discriminating power due to the physical predictions of theory. The variables used in the Run 1 analysis [22], along with a description of what each of them are, are listed in Table 3.2. Furthermore, events are split into low and high p_T^V regions, with the cutoff at $p_T^V = 120$ GeV, as past studies have shown this improves the results. Separate BDTs are trained and evaluated for each of these regions.

In addition to these variables, this analysis also adds in 7 new ones: mBBR2, mBBR3, mBBR4, mBBR5, mBBR6, mBBR7, and mBBR8. Each of these is the Higgs mass reconstructed from the two LC-calibrated jets for the corresponding radius R that were matched to the two leading R=0.4 *b*-jets. Figures 3.2 and 3.3 show the 2D distributions for mBB against mBBR2 and mBBR8 in the high and low p_T^V regions, to show how the Higgs mass differs when reconstructed from jets of different radii. These plots show that as expected, the value of mBBR2 is generally smaller than than the value of mBB, which is constructed from larger jets, and that mBB is generally smaller than mBBR8, which is constructed from even larger jets. The distributions of all the input variables for both background and signal samples are shown in Figures 3.4 and 3.5, where it can be seen that under naive expectations, mBB, dRBB, and dEtaBB have some of the best discriminating power. More detailed plots of these variable distributions can be found in Appendix A. Correlation matrices for all of the input variables are shown in Figure 3.6. The values of mBBR2 through mBBR8 are less strongly correlated with mBB in the signal region than they are in the background region, which provides weak evidence in favor of the idea that they may help discriminate signal from background, as found in [11].

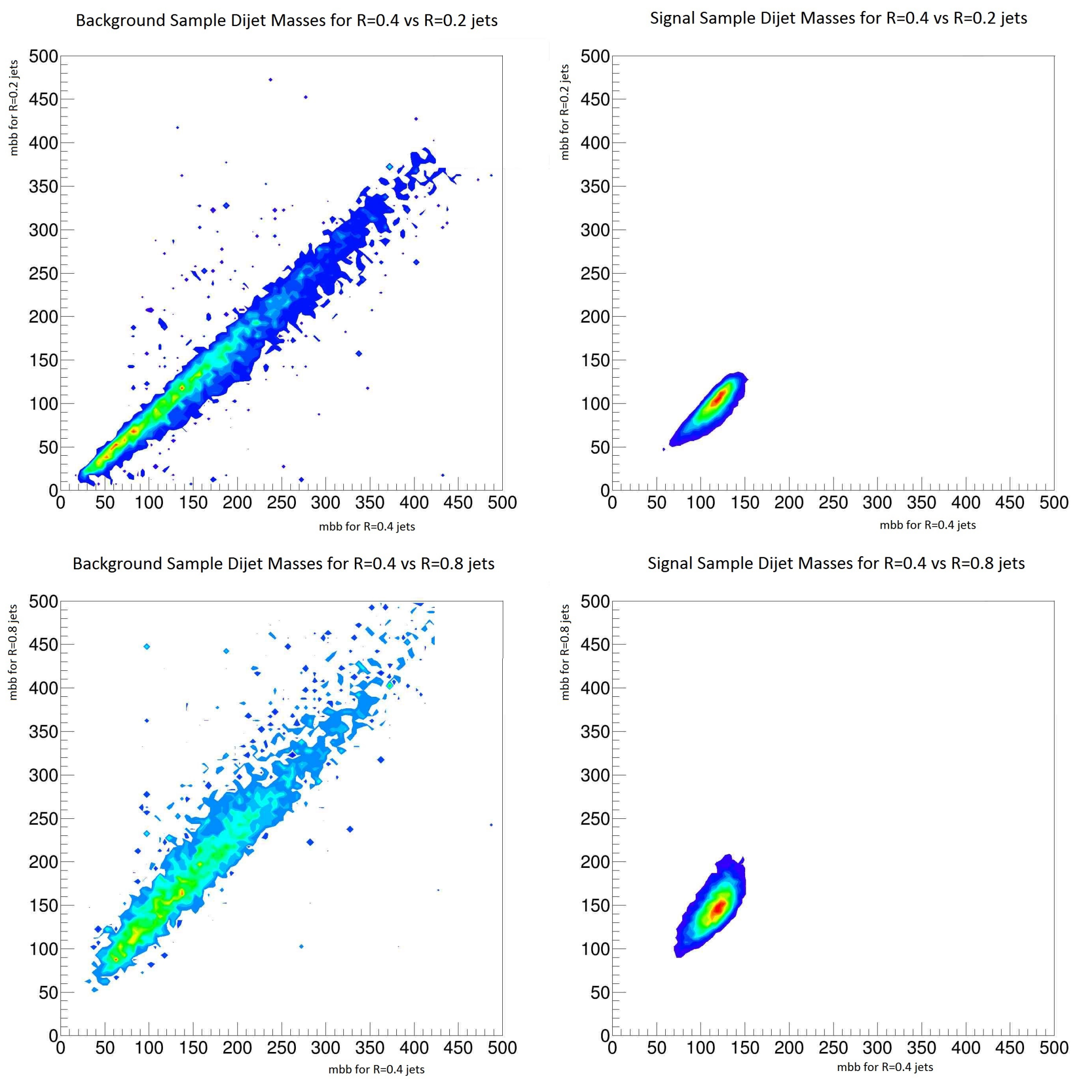

Figure 3.2: 2D distribution of the dijet mass reconstructed from different jet radii in the high p_T^V region. The top row has mBB on the x-axis and mBBR2 on the y-axis, and the bottom row has mBB on the x-axis and mBBR8 on the y-axis. The left column is for the background samples, and the right column is for the signal sample.

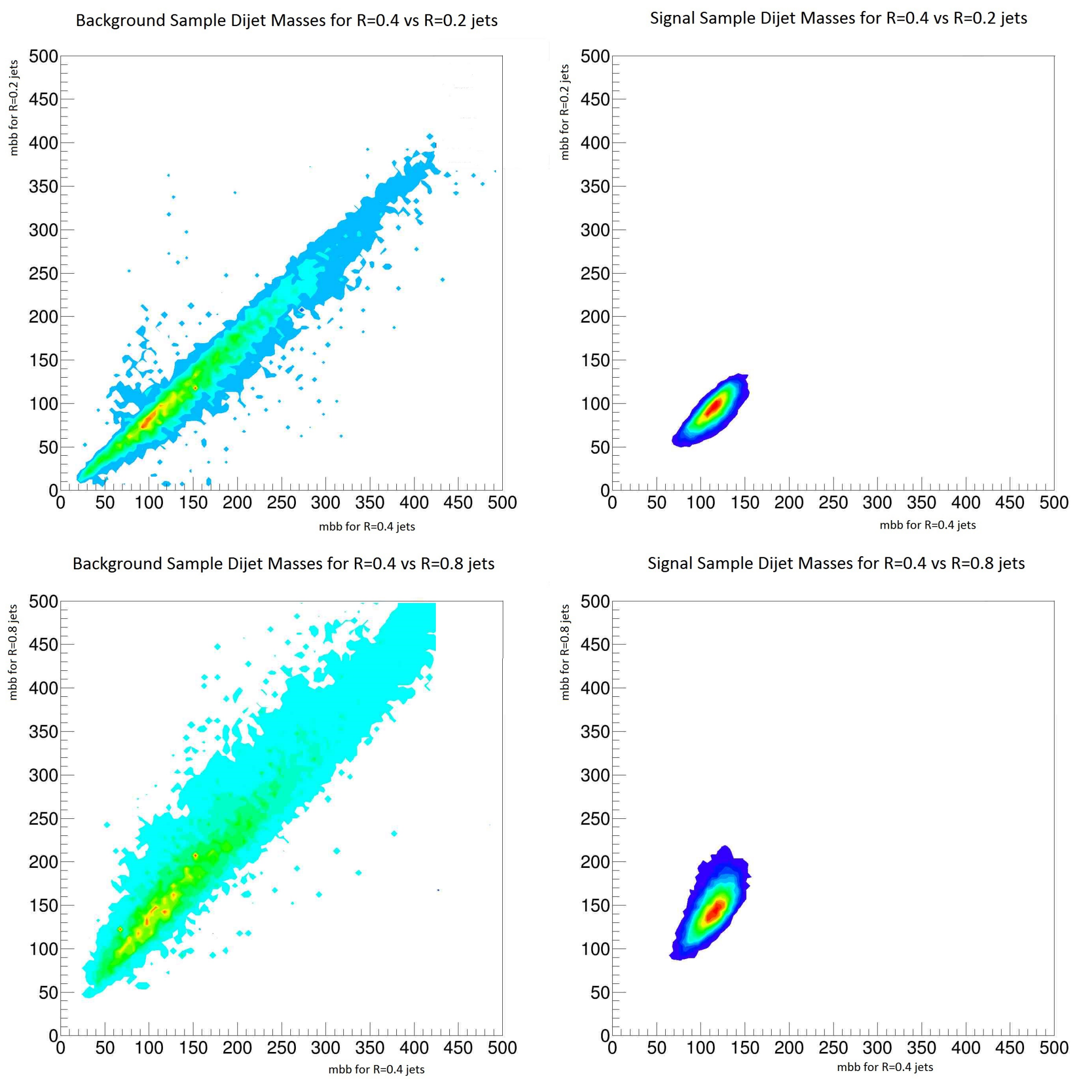

Figure 3.3: 2D distribution of the dijet mass reconstructed from different jet radii in the low p_T^V region. The top row has mBB on the x-axis and mBBR2 on the y-axis, and the bottom row has mBB on the x-axis and mBBR8 on the y-axis. The left column is for the background samples, and the right column is for the signal sample.

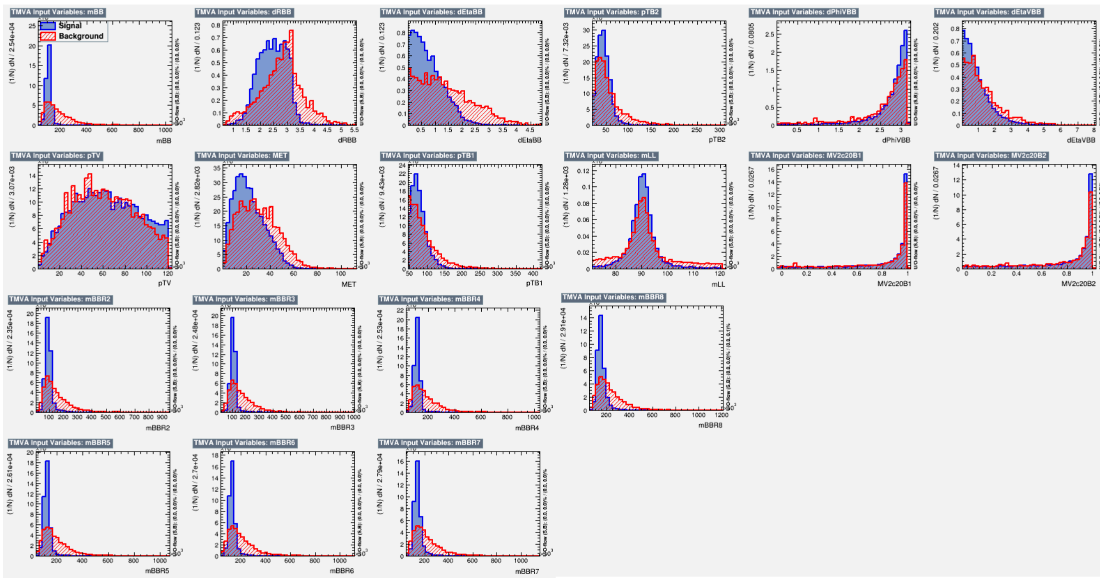

Figure 3.4: Distribution of all input variables to the BDT for background and signal samples overlaid, in the low p_T^V region.

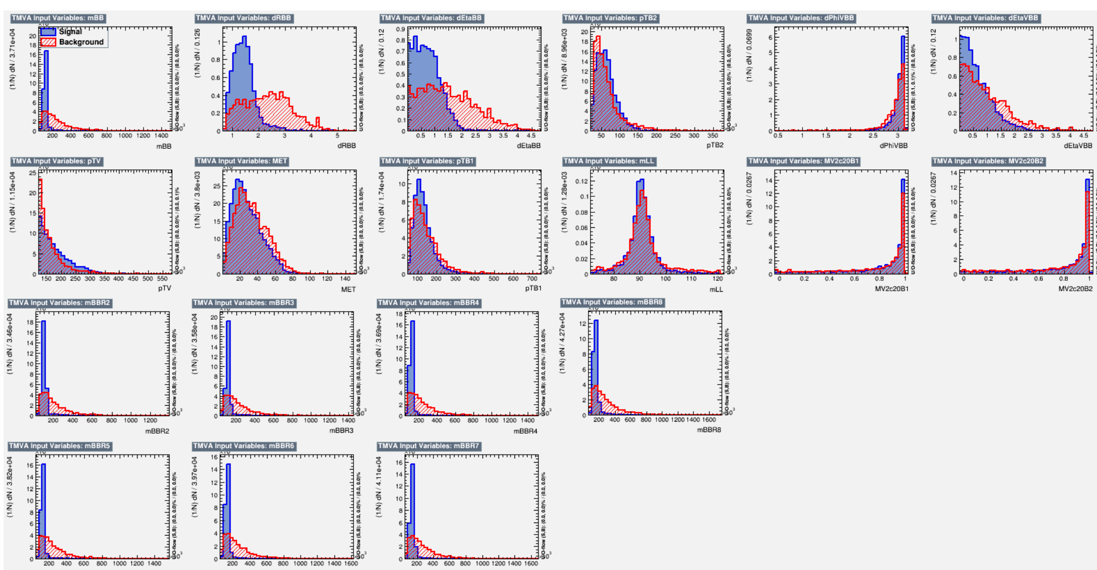

Figure 3.5: Distribution of all input variables to the BDT for background and signal samples overlaid, in the high p_T^V region.

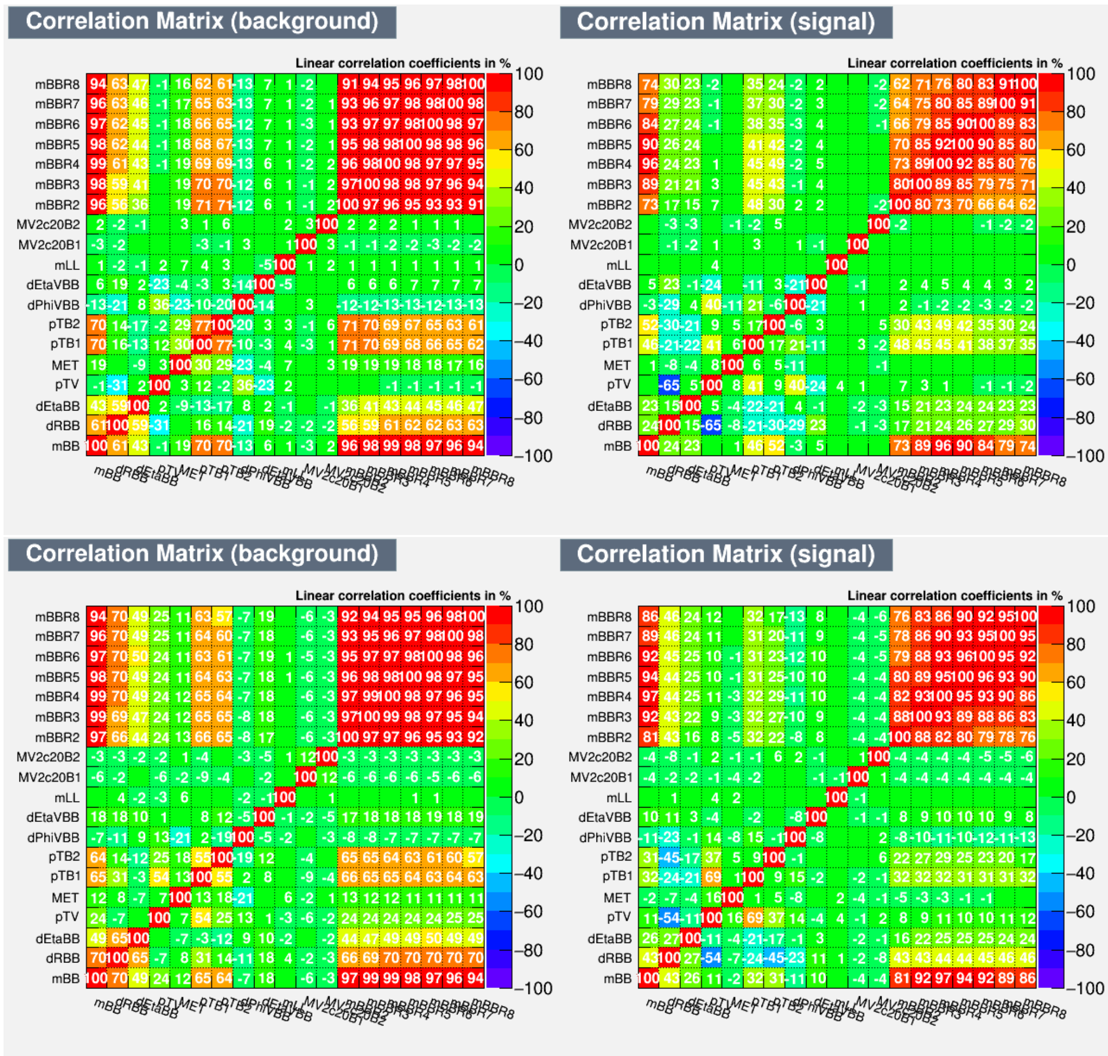

Figure 3.6: Correlation matrices for the input variables for background (left) and signal (right) samples. The top row is for the low p_T^V region, and the bottom row is for the high p_T^V region. The correlations of the mBB values reconstructed from jets of different radii are roughly the expected values found in [11].

4

BDT Training and Optimization

Multivariate methods are now fairly widely used in high-energy physics analyses, because there is rarely just one or two physical observables that give all the information needed to classify the events. In machine learning terms, the analysis has a binary classification problem, where events are classified as either signal or background. Since we are using data from Monte Carlo simulations, each event has an accurate truth label, which can be used for training and testing. This analysis uses the boosted decision tree (BDT) implemented in TMVA [26], the same method used in the Run 1 analysis [22]. However, with machine learning methods, care must be taken to not overfit the training data. This chapter offers an overview of the strategies used to train the BDT. Section 4.1 describes the BDT algorithm and how its performance is measured. Section 4.2 describes the method used to split up the data for training and testing. Section 4.3 describes the parameters used to configure the BDT and how the optimal settings were found.

4.1 Boosted Decision Tree Algorithm

Boosted decision trees use the idea of *boosting*, by which an ensemble of weaker methods (decision trees in this case) can be combined to create a more robust and more powerful method. A basic binary decision tree selects the variable with the highest separation power at each node, and then makes a cut at some value, separating the remaining events into two pools, which are passed to its children in the tree. The separation gain is defined by some separation index chosen by the user; here, the Gini Index is used, defined as $p(1-p)$, where p is the purity of the node. The leaves at the end of the tree then each contain a number of events, and the leaves are each labelled Signal or Background according to the majority of events that have ended up in that particular node. This process is shown in Figure 4.1.

In a boosted decision tree, the events that were misclassified by the first decision tree are given an additional weight depending on the boosting procedure being used before the entire event set is fed into the next decision tree. This process is repeated sequentially for however many trees have been specified. For the Adaptive Boost algorithm (AdaBoost) being used in this BDT, the weights of each misclassified event are multiplied by α^{β}, where α is given by Eq. 4.1, β is a user-specified learning rate, and *err* is the misclassification rate of the previous tree. The weights are renormalized for each tree so that their sum remains constant.

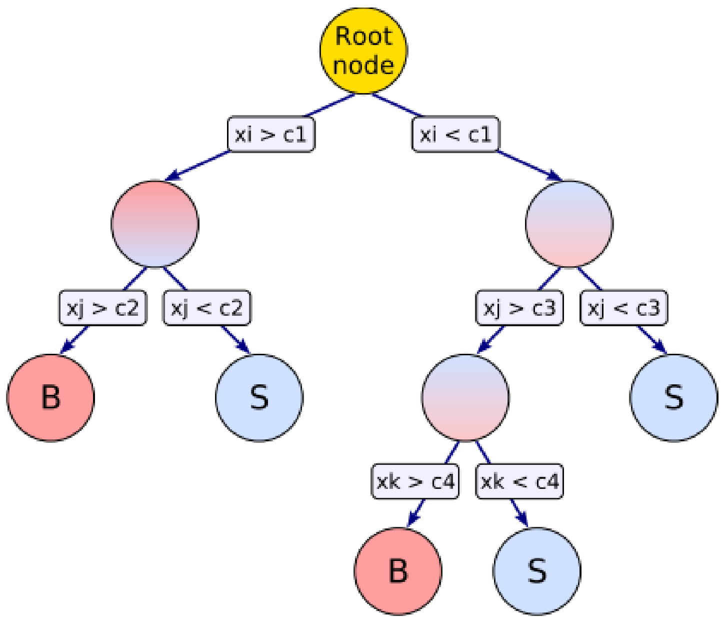

Figure 4.1: Model of a simple binary decision tree. The root node contains all of the input events, and at each level it picks the one input variable with the best discriminating power, and makes a cut on it to split the events into two pools passed to its children. This is repeated until some stopping criteria is reached, and the final leaves are labelled Signal or Background. The final leaf that an event ends up in is the classification that the decision tree assigns to it. [26]

$$\alpha = \frac{1 - err}{err} \tag{4.1}$$

This ensemble of trees, also called a forest, then gives the boosted decision tree's final output for each event by taking a weighted average of the result of each of the consituent decision trees. For the AdaBoost, this output on an event x is given by Eq. 4.2, where $h_i(x)$ denote the binary outputs 1 or -1 of each individual tree.

$$BDT(x) = \frac{1}{N_{trees}} * \sum_{i=1}^{N_{trees}} \ln{(\alpha_i)} h_i(x) \tag{4.2}$$

This procedure works well with weak classifiers like shallow decision trees, as the likelihood of overtraining one of the component classifiers is small, and the final ensemble is more stable with respect to statistical fluctuations in the training samples. Boosted decision trees also have the benefit that they tend to be robust against unuseful variables, since they will just avoid using them.

Finally, there must be a measure of the performance of the BDT. Regular decision trees can be judged by a simple accuracy measure, since they just output yes/no for binary classification problems, but a BDT gives an output between -1 and 1 for each event as described in Eq. 4.2, with the output value indicating the likelihood that the event is background or signal. To translate this output into a classification, a cut is applied at some value, with BDT outputs above this value indicating signal and outputs below it indicating background. A standard $S/\sqrt{S+B}$ measure for significance is used, where for a chosen cut, S is the number of signal events above this cut and B is the number of background events above the cut. In machine learning terms, S is the number of true positives, and B is the number of false positives. This measure of significance is preferable over something like a plain accuracy statistic because the problem in question has far more background events than signal events. For any particular configuration, the cut that maximizes $S/\sqrt{S+B}$ is chosen, and this is treated as the performance of that configuration.

4.2 Holdout Method

In an ideal world, there would be a wealth of data available for training and testing. Unfortunately, this is not the case in the real world, and so there exist methods to make the best use of the available data while avoiding overtraining. This analysis employs a holdout method, depicted in Figure 4.2. The data is randomly split into 3 equal parts. Since each event in Monte Carlo data is labelled with an EventNumber, which is unrelated to the actual contents of the event, this splitting is done by using the value of EventNumber % 3. These three divisions are labelled training, validation, and test data. Step 1 is to tune the parameters of the BDT using the training data for training and the validation data for testing, with the optimal configuration determined by looking at the BDT's performance on the validation data. Step 2 is to train the BDT on the training and validation data together using this optimal configuration, and then testing it with the test data. The performance on the test data is measured by applying the cut that maximized significance on the validation data, and this is then used as the final evaluation of the BDT's performance. This means that the test data is never used for tuning, and so the result can be trusted as the true performance of the BDT, rather than a consequence of overtraining.

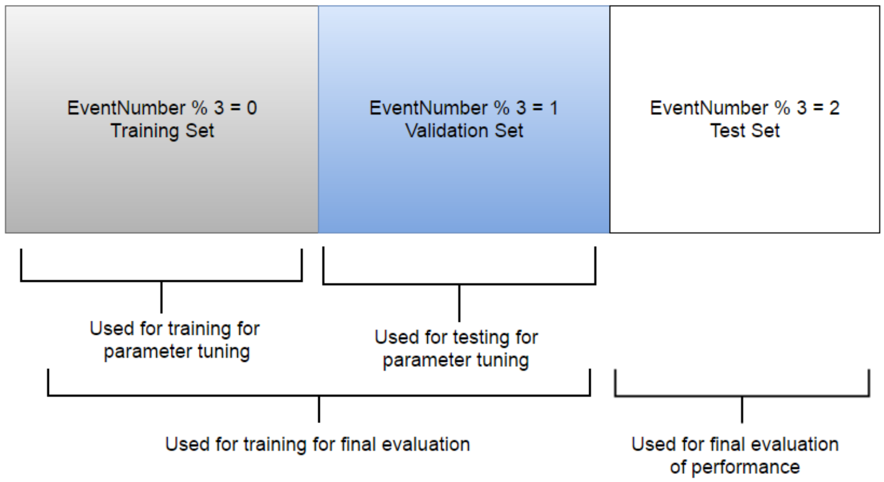

Figure 4.2: Diagram of how the available data is split 3-ways for the holdout method. The training data is only ever used for training; the validation data is used for testing when tuning the BDT and also included in training for final evaluation. The test data is never used for training and never used for tuning; it is only used for the final evaluation of the BDT's performance after a configuration has been chosen.

4.3 BDT Parameter Tuning

Machine learning methods generally have a number of free parameters with which they can be configured, and the configuration which yields the best results varies with the problem and the inputs. The relevant parameters, their meanings, and their settings in the Run 1 analysis BDT are shown in Table 4.1. nCuts is set to -1 for every BDT used in this analysis, which tells the tree to algorithmically find the optimal cut. The settings that are considered for tuning are the AdaBoostBeta, MinNodeSize, NTrees, and MaxDepth.

TMVA Setting	Value	Meaning
BoostType	AdaBoost	boost procedure (see Section 4.1)
SeparationType	GiniIndex	measure of separation gain in nodes (see Section 4.1)
PruneMethod	NoPruning	pruning method (normally not used in boosted ensembles)
nCuts	100	number of equally spaced cuts tested per variable per node
AdaBoostBeta	0.15	learning rate for boosting
NTrees	200	number of trees in the ensemble
MaxDepth	4	maximum depth of a tree
nEventsMin	100	minimum number of events per node (now replaced by MinNodeSize)
MinNodesize	N/A	minimum node size as a percentage of total number of events

Table 4.1: BDT parameters used by the TMVA implementation along with their meanings and values for the BDT used in the Run 1 analysis [22]. The parameter nEventsMin, whose value was an absolute number of events, has been replaced by MinNodeSize, which is a percentage, in the TMVA implementation since the previous analysis.

The optimal configuration found in [22], listed in Table 4.1, did not use the variables mBBR2 through mBBR8 described in Section 3.3, and was trained and evaluated on a different dataset. Consequentially, the BDT must be tuned again to properly take advantage of this new information on this new data. It is computationally impractical to perform a multi-dimensional parameter scan over everything to find the optimal configuration, so instead a 2-dimensional scan is first performed over the NTrees and AdaBoostBeta parameters, and a 1-dimensional scan is performed over each of MaxDepth and MinNodeSize at the optimal value found using the 2D scan. The parameter-scan results for the high p_T^V region can be found in Figure 4.3, and for the low p_T^V region in Figure 4.4. Points with abnormally high significance relative to the surrounding points were discarded as likely statistical fluctuations. The final settings chosen for each BDT using all of the variables, including mBBR2 through mBBR8, are shown in Table 4.2.

TMVA Parameter	Setting for $p_T^V > 120$ GeV	Setting for $p_T^V < 120$ GeV
NTrees	1500	1100
AdaBoostBeta	0.15	0.05
MaxDepth	4	4
MinNodeSize	1.5%	2.5%
nCuts	-1	-1

Table 4.2: Settings for the BDT parameters used for the low and high p_T^V samples for the final evaluation. Any other parameters not shown here used the values shown in Table 4.1.

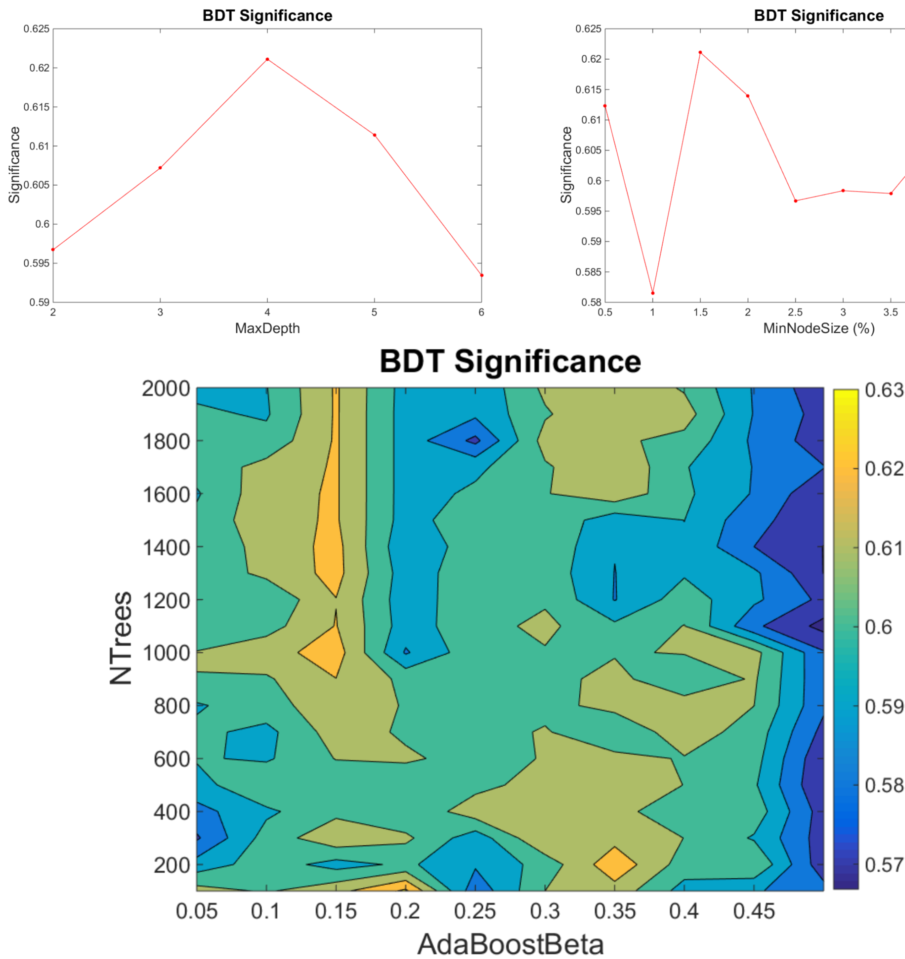

Figure 4.3: Plot of how the BDT's performance varies with various parameters using all of the input variables in the high p_T^V region, with training done on the training set and performance measured on the validation set described in Section 4.2. The settings used were NTrees=1500, AdaBoostBeta=0.15, MaxDepth=4 and MinNodeSize=1.5% as the default, and then 1 or 2 of these parameters were varied in the plots. The 2D scan over NTrees and AdaBoostBeta (bottom) was performed first to find this point, and then 1D scans (top) were performed around this point.

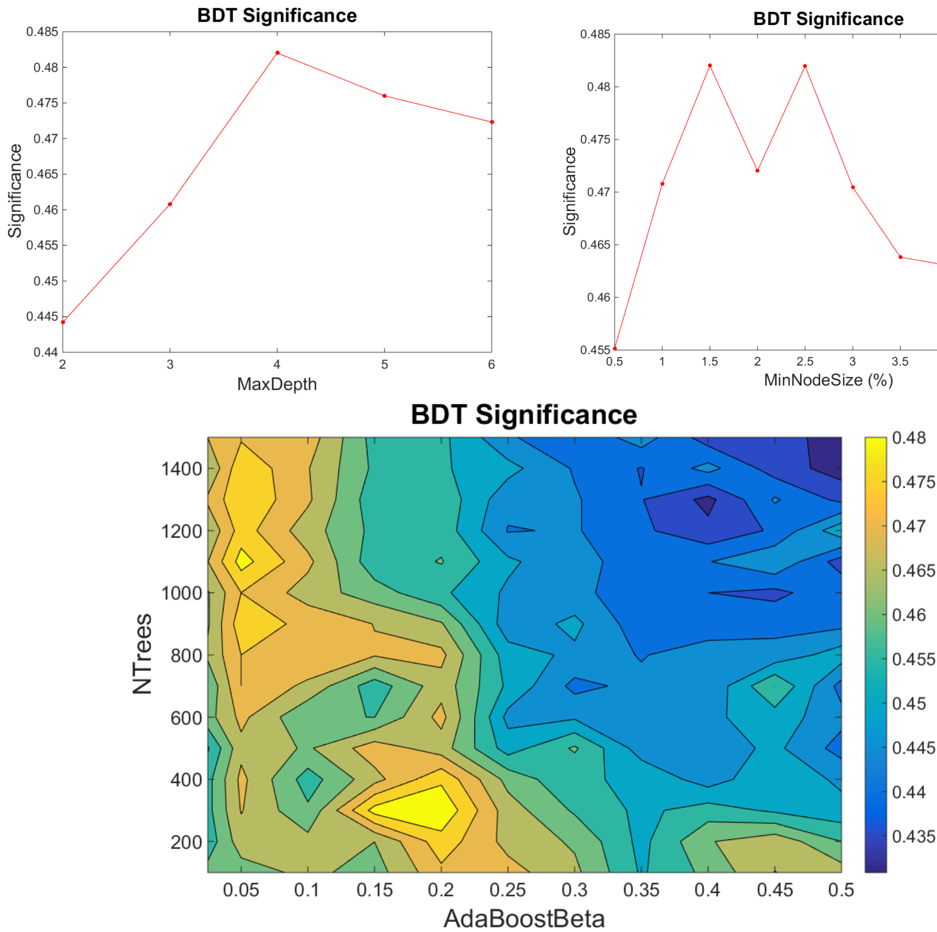

Figure 4.4: Plot of how the BDT's performance varies with various parameters using all of the input variables in the low p_T^V region, with training done on the training set and performance measured on the validation set described in Section 4.2. The settings used were NTrees=1100, AdaBoostBeta=0.05, MaxDepth=4 and MinNodeSize=1.5% as the default, and then 1 or 2 of these parameters were varied in the plots. The 2D scan over NTrees and AdaBoostBeta (bottom) was performed first to find this starting point, and then 1D scans (top) were performed around this point.

5
Results

Region	Base 12 Variables	Base + 7 R's Unoptimized	Base + 7 R's Optimized	Percent Improvement
$p_T^V < 120$ GeV	0.416	0.420	0.424	1.9%
$p_T^V > 120$ GeV	0.664	0.682	0.707	6.5%

Table 5.1: Significances of the BDT evaluated on the test dataset, and the total percent improvement found by adding in the additional mBB from multiple interpretations.

The significance of the BDT trained on the training and validation datasets and tested on the test dataset is shown in Table 5.1. The significance value for each BDT is obtained by finding the cut that optimizes $S/\sqrt{S+B}$ on the validation sample, and then applying that same cut on the test sample. The base 12 variables are the ones used in the Run 1 analysis [22], using the same BDT settings, which are shown in Table 4.1. The unoptimized result with the additional mBB reconstructed from jets of different R was evaluated using the same parameters, and the optimized result uses the settings found by parameter scan in Section 4.3. The magnitude of these significances are smaller than those found in [22], but this should not be a concern, since this analysis used a much smaller dataset and thus has poorer statistics. We are only concerned with the relative magnitude of the significance of the BDT with just the base 12 variables compared to the significance of the BDT with the additional R's. The shapes of the BDT outputs can be found in Figure 5.1, and it is seen that adding in the new inputs does not change the shape of the BDT output significantly.

The results show a small improvement in the significance of the BDT results after adding in inputs based on multiple event interpretations using jets of different R, but the improvements are significantly smaller than the ones found using a BDT in the truth-level study in [11]. As a validation check on the approach taken in this analysis, the work in [10] using a cut-based analysis with multiple interpretations is replicated in Appendix B, and similarly high levels of improvement are found. This suggests the event reconstruction and selection was done properly in this analysis, and that the problem likely lies elsewhere.

There are a number of factors that could explain the disparity between the significance improvement found here and the improvements found at truth-level in [11]. First, the jet matching method described in Section 3.2 is not perfect, and the loss of events from failed matching in addition to the possible information loss from this indirect method of b-tagging the jets could be a problem. In addition, this reconstructed-level analysis has to

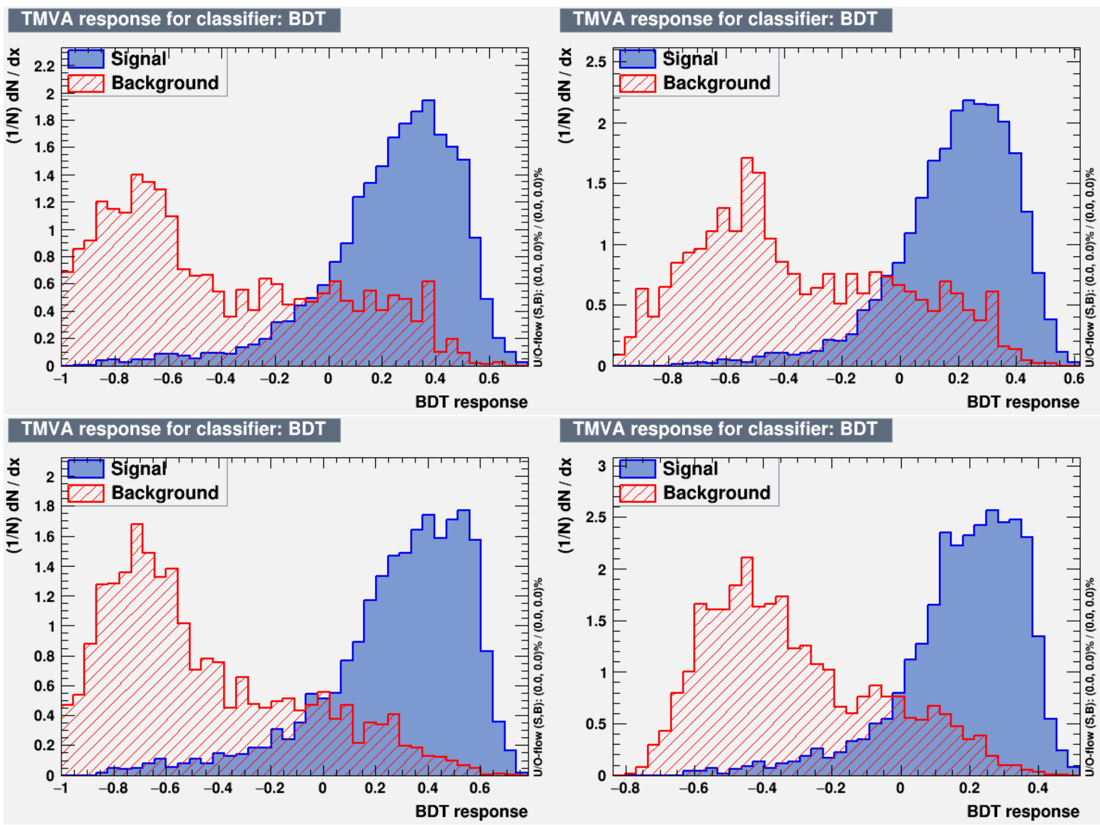

Figure 5.1: BDT Output on the test dataset. The top row is for the low p_T^V region, and the bottom row is for the high p_T^V region. The left column has the outputs using only the base variables from the Run 1 analysis, and the right column has the outputs including the additional inputs from the different R jets.

use the MV2 algorithm to b-tag the R=0.4 jets in the first place, which is also imprecise compared to using the truth-level flavor of the jets. The truth-level analysis also used 12 different radii jets, from R=0.4 up through R=1.5 at intervals of 0.1, as opposed to the 7 radii used in this analysis from R=0.2 to R=0.8. It is possible that much of the information gain responsible for the improvements in the truth-level BDT output came from the higher radii jets, which were not present in this analysis.

Lastly, in theory the holdout method employed in this analysis and described in Section 4.2 means that these results should be valid, since it is blinded to the test dataset until the final evaluation. However, the relatively small amount of data available in total for this analysis means the statistics are rather poor. Consequentially, the BDT results can be unstable and prone to statistical fluctuations, as evidenced by the non-smooth distribution of the BDT output in Figure 5.1. Significance gains as small as the ones seen here can plausibly be attributed to such fluctuations, as a result of a "lucky" split of the data. It is also possible that the significance gains are small because of an "unlucky" split of the data, and that the real gains would be higher. This issue could potentially be addressed with the use of a different measure of the significance; a preliminary study where the significance is found by adding the $S/\sqrt{S+B}$ values for each of the BDT output bins in quadrature has found improvements of 8-9% with the addition of multiple interpretations variables. The combination of the improvements found in this analysis with the results in Appendix B suggests that it is likely that the poor statistics are working against us, and is promising for the use of multiple interpretations in the BDT once more data is available.

6

Conclusion

Previous work [10, 11] has suggested that the use of multiple interpretations by constructing jets at multiple different radii and using the dijet masses from each of them could offer sizable improvements in significance in the $H \to b\bar{b}$ search in association with a W or Z boson. This thesis has attempted to use multiple dijet masses $m_{bb,R=0.2}$ through $m_{bb,R=0.8}$ at intervals of 0.1 as additional input to the boosted decision tree used in the $ZH \to \ell\ell b\bar{b}$ search while under conditions close to those of the real analysis. However, the improvements of 1.9% in the $p_T^V < 120$ GeV region and 6.5% in the $p_T^V > 120$ GeV region are significantly smaller than the improvements found in previous work at truth-level. Possible causes of this disparity may be the jet matching strategy used in this analysis, the fact that this analysis uses fewer different radii for practical reasons than previous work did, and the relatively small datasets available for training and testing, with the latter 2 issues probably being the biggest culprits. The fact that this analysis was able to find small improvements despite all of these barriers is promising for the use of multiple interpretations under realistic analysis conditions. It is likely that further investigation with proper *b*-tagging information for the jets of every radius used, with additional radii jets being constructed, and with better statistics available will yield higher levels of improvement that are more consistent with the expectations of previous work.

A

BDT Input Variable Distributions by Sample

This appendix contains stacked plots of the base input variables to the BDT described in Section 3.3. These plots show how the distributions vary for the different types of background, and it is seen that the largest background is that of a Z boson with two b-jets, as expected. The distributions in the $p_T^V < 120$ GeV region are shown in Figure A.1, and the $p_T^V > 120$ GeV region is shown in Figure A.2.

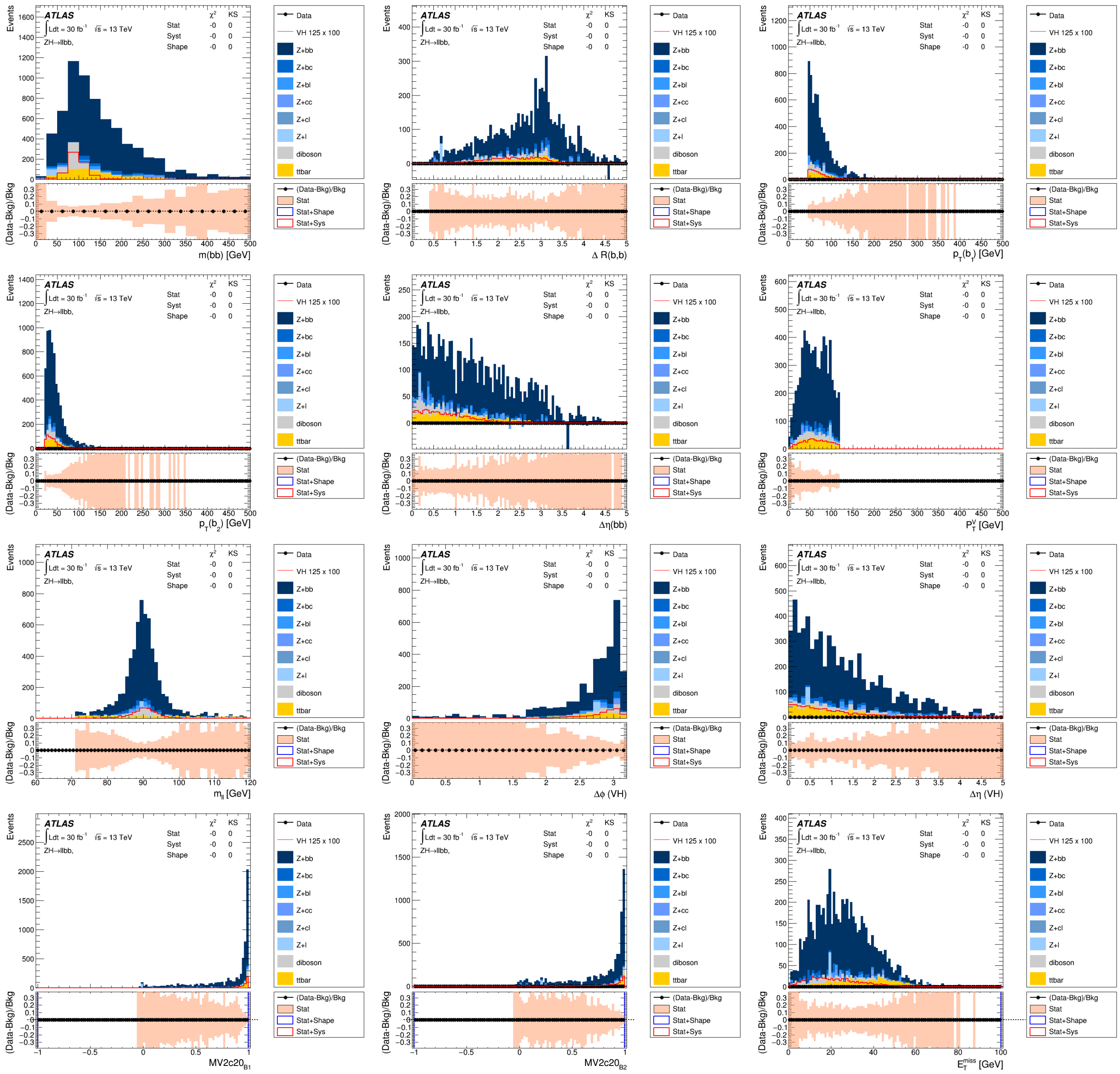

Figure A.1: Distributions for all of the BDT variables in the $p_T^V < 120$ GeV region, separated by the different types of background.

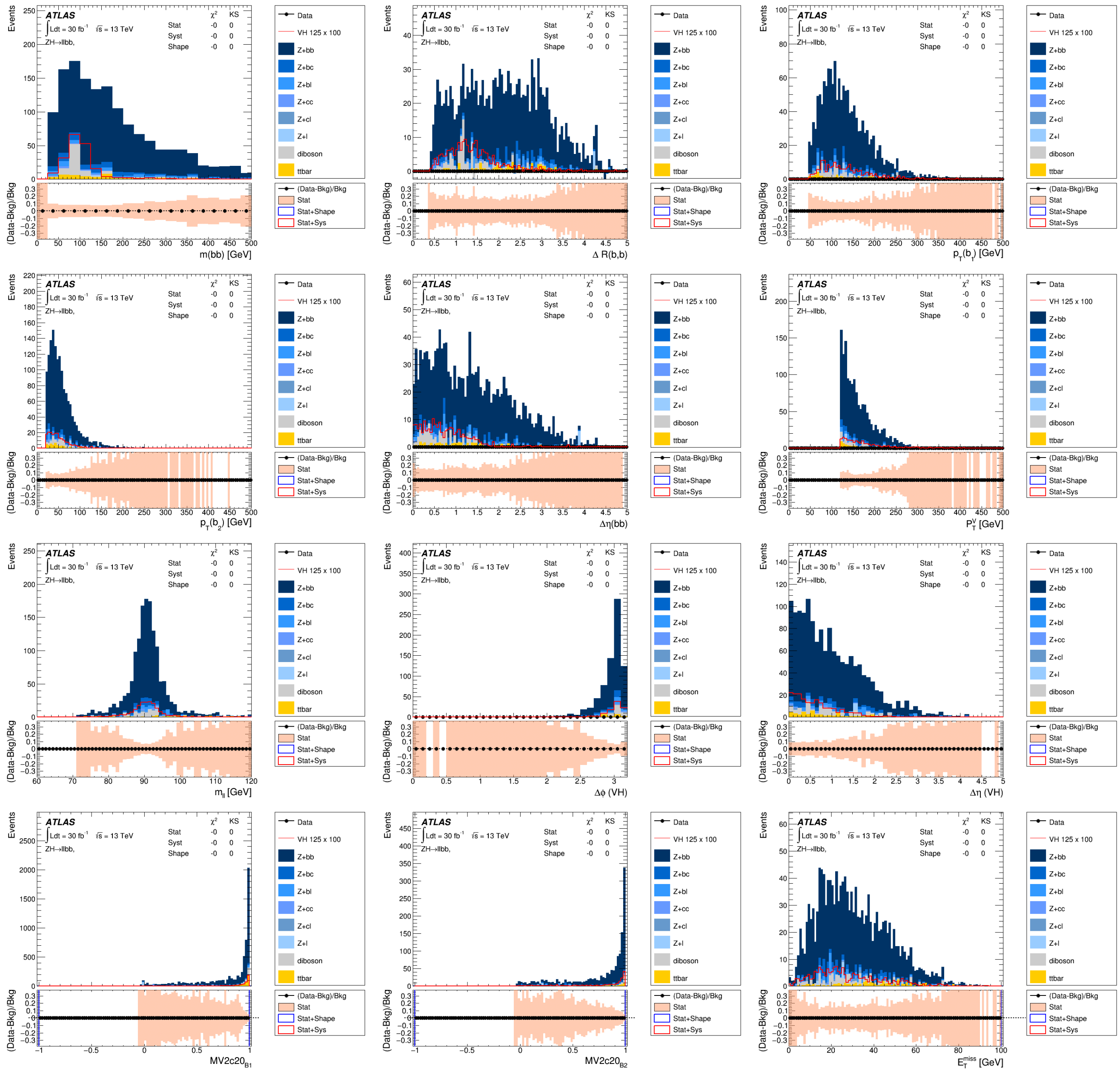

Figure A.2: Distributions for all of the BDT variables in the $p_T^V > 120$ GeV region, separated by the different types of background.

B

Replication of Previous Studies

One of the studies that this thesis builds upon is [10], which found a 21% improvement in significance at the reconstructed level using a cut-based approach with multiple interpretations from jets of different radii. That approach is replicated here for comparison. For this analysis, only the Z+jets samples are used for background. The samples are further split into 5 p_T^V bins, and the analysis is done separately on each of these bins. The significance changes using the 2D cut described in the paper are shown in Figure B.1, where improvements are either very small or nonexistent, as they were in the previous study.

The study further defines a parameter ϱ, defined as the fraction of event interpretations that pass a m_{bb} cut designed to select for signal. The ϱ distributions for the signal and back-

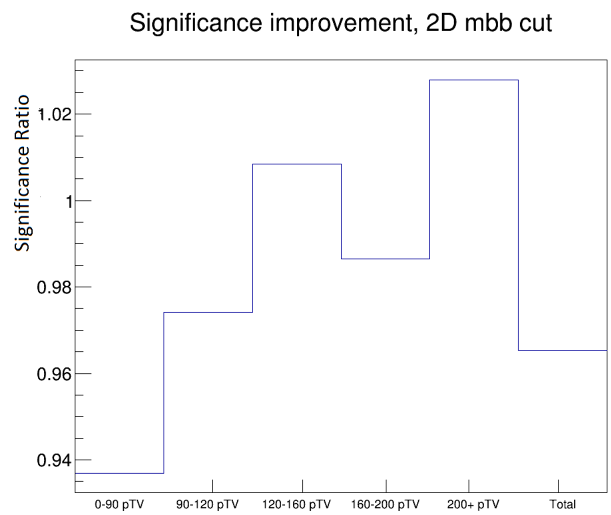

Figure B.1: Ratio of significance using the 2D cut of $95 < m_{bb,R=0.4} < 140$ GeV, $105 < m_{bb,R=0.6} < 160$ GeV compared to the 1D cut.

ground samples are shown in Figure B.2. Note that the background ϱ distribution peaks around 0, while the signal ϱ distribution peaks closer to 1, as expected.

The significance improvement using the $t^*(z)$ measure described in the study in each of the p_T^V bins is shown in Figure B.3, finding similarly high levels of improvement. On the whole, the significance gains and ϱ distributions found in this replication are similar, but not identical to those found in the original study. This disparity is likely due to the different methods in jet reconstruction, since the original study used a telescoping jets approach, whereas this analysis reconstructs the jets directly from the calorimeter topological clusters. The successful replication of the results in [10] means that the unexpectedly low improvements in significance found in Section 5 are likely the result of some other complicating factor that comes from bringing in the BDT.

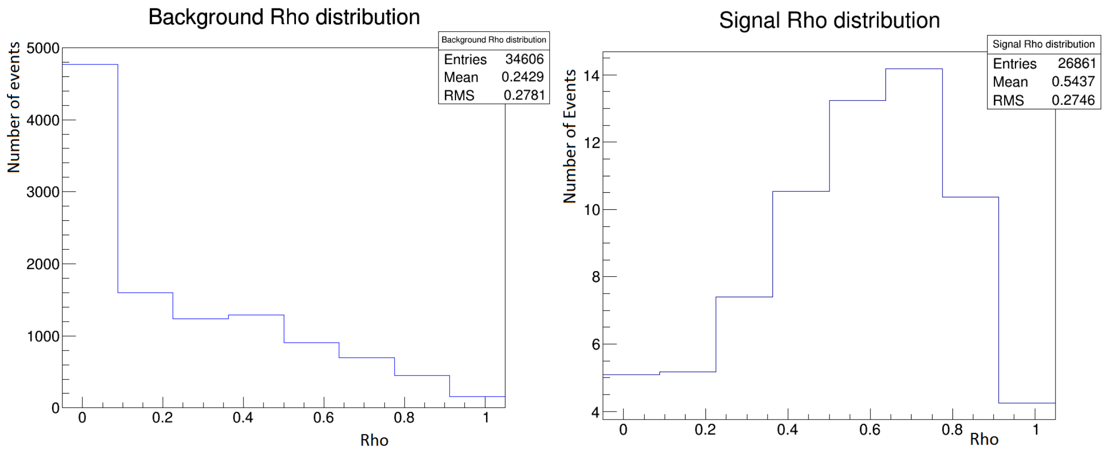

Figure B.2: ϱ distributions for the background (left) and signal (right) samples. The counts are weighted by the event weights from the Monte Carlo simulation.

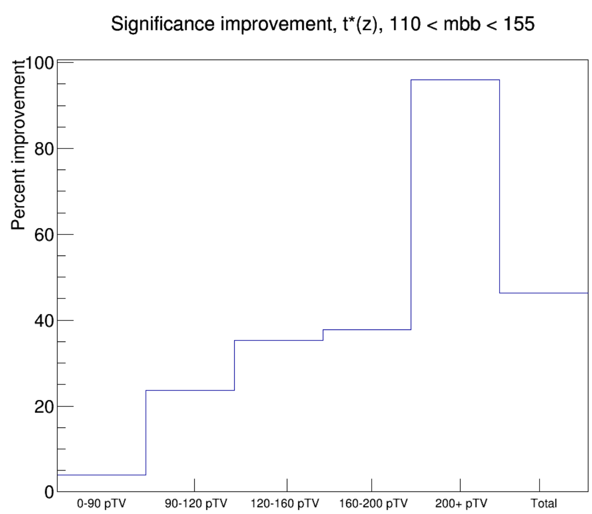

Figure B.3: Improvement in significance by using $t^*(z)$ instead of the 1D cut. The total significance gain is found by adding the significance gains from each of the other p_T^V bins in quadrature.

References

[1] ATLAS Collaboration. Search for the $b\bar{b}$ decay of the Standard Model Higgs boson in associated $(W/Z)H$ production with the ATLAS detector. 2014. http://arxiv.org/abs/1409.6212v1.

[2] Wikimedia Commons. Standard Model of Elementary Particles. 2006. https://commons.wikimedia.org/wiki/File:Standard_Model_of_Elementary_Particles.svg.

[3] ATLAS Collaboration. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. 2012. http://arxiv.org/abs/1207.7214.

[4] CMS Collaboration. Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. 2012. http://arxiv.org/abs/1207.7235.

[5] ATLAS Collaboration and CMS Collaboration. Combined Measurement of the Higgs Boson Mass in pp Collisions at $\sqrt{s} = 7$ and 8 TeV with the ATLAS and CMS Experiments. 2015. http://arxiv.org/abs/1503.07589.

[6] ATLAS Experiment Blog. A new particle is born. 2012. http://pdg3.lbl.gov/atlasblog/wp-content/uploads/2012/07/sm_xsec.png.

[7] Stephen Ellis, Andrew Hornig, David Krohn, Tuhin Roy, and Matthew Schwartz. Qjets: A Non-Deterministic Approach to Tree-Based Jet Substructure. *Physical Review Letters*, 108, 2012. http://xxx.lanl.gov/abs/1201.1914.

[8] Yang-Ting Chien. Telescoping Jets: Multiple Event Interpretations with Multiple R's. 2013. http://xxx.lanl.gov/abs/1304.5240.

[9] Dilani Kahawala, David Krohn, and Matthew Schwartz. Jet Sampling: Improving Event Reconstruction through Multiple Interpretations. *JHEP*, 1306, 2013. http://xxx.lanl.gov/abs/1304.2394.

[10] Stephen Chan, John Huth, David Lopez Mateos, and Kevin Mercurio. $ZH \to llb\bar{b}$ Analysis with Telescoping Jets. 2015. https://cds.cern.ch/record/1982439.

[11] Yang-Ting Chien, David Farhi, David Krohn, Andrew Marantan, David Lopez Mateos, and Matthew Schwartz. Quantifying the power of multiple event interpretations. 2014. http://arxiv.org/abs/1407.2892.

[12] ATLAS Collaboration. The ATLAS Simulation Infrastructure. May 2010. http://arxiv.org/abs/1005.4568.

[13] S. Agostinelli et al. GEANT4, a simulation toolkit. http://www.geant4.org/geant4/.

[14] Torbjörn Sjöstrand, Stephen Mrenna, and Peter Skands. A Brief Introduction to PYTHIA 8.1. 2007. http://arxiv.org/abs/0710.3820.

[15] ATLAS Collaboration. ATLAS tunes of PYTHIA6 and PYTHIA8 for MC11. *ATL-PHYS-PUB-2011-009*, 2011.

[16] J. Pumplin, D.R. Stump, J. Huston, H.L. Lai, P. Nadolsky, and W.K. Tung. New Generation of Parton Distributions with Uncertainties from Global QCD Analysis. 2002. http://arxiv.org/abs/hep-ph/0201195.

[17] T. Gleisberg, S. Hoeche, F. Krauss, M. Schoenherr, S. Schumann, F. Siegert, and J. Winter. Event generation with SHERPA 1.1. 2008. http://arxiv.org/abs/0811.4622.

[18] Simone Alioli, Paolo Nason, Carlo Oleari, and Emanuele Re. A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX. 2010. http://arxiv.org/abs/1002.2581.

[19] ATLAS Collaboration. Topological cell clustering in the ATLAS calorimeters and its performance in LHC Run 1. 2016. http://arxiv.org/abs/1603.02934.

[20] ATLAS Collaboration. ATLAS Experiment - Public Results. https://twiki.cern.ch/twiki/bin/view/AtlasPublic/EventDisplayRun2Collisions.

[21] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. The anti-k_t jet clustering algorithm. 2008. http://arxiv.org/abs/0802.1189.

[22] ATLAS Collaboration. Supporting Document for the Search for the bb decay of the Standard Model Higgs boson in associated (W/Z)H production with the ATLAS detector. Technical Report ATL-COM-PHYS-2014-051, CERN, Geneva, Jan 2014. https://cds.cern.ch/record/1645654.

[23] Marc Lehmacher. b-Tagging Algorithms and their Performance at ATLAS. 2008. http://arxiv.org/abs/0809.4896.

[24] ATLAS Collaboration. Expected Performance of the ATLAS Experiment - Detector, Trigger and Physics. 2009.

[25] ATLAS Collaboration. Electron reconstruction and identification efficiency measurements with the ATLAS detector using the 2011 LHC proton-proton collision data. 2014. http://arxiv.org/abs/1404.2240.

[26] A. Hoecker et al. TMVA - Toolkit for Multivariate Data Analysis. 2007. http://arxiv.org/abs/physics/0703039.