# Boosted regression trees in the Higgs $\rightarrow \tau\tau$ channel

Natascha S. Hedrich

*Thompson Rivers University*
*900 McGill Road, Kamloops, BC, V2C 1M0*

Supervisor: Dr. Dugan O'Neil
*Department of Physics,Simon Fraser University,*
*8888 University Dr., Burnaby, BC, V5A 1S6*

August 29, 2013

**Abstract**

This report examines the application of a multivariate analysis technique, known as Boosted Regression Trees (BRT's) to the reconstruction of the Higgs mass. BRT's are being evaluated as a competing method to the Missing Mass Calculator, which is currently being used in the H $\rightarrow \tau\tau$ channel. The effects of the regression target distribution, input variables and training parameters on the regression performance are also investigated. BRT's are a promising technique and further studies will aim to better understand potential biases.

## 1 Introduction

Particle physics, the study of the most basic constituents of matter, has evolved over many years in the process of trying to explain how the universe works on a fundamental level. These years of study have lead to the development of the Standard Model (SM), which has been tested extensively and has successfully predicted many new particles. Of these particles predicted by the SM, the Higgs boson has been the most elusive. On July 4, 2012 a new Higgs-like boson was announced and represents an impressive experimental accomplishment. After over fifty years, last summer's discovery - a 5.9 sigma excess at 126.0 $\pm$ 0.4(stat) $\pm$0.4 (syst) GeV reported by the ATLAS collaboration [1] and a 5.0 sigma excess at 125.3 $\pm$0.4 (stat)$\pm$0.5(syst) GeV reported by the CMS collaboration [2]- has shed light on the question, where does mass come from? To date, the Higgs boson has only been observed decaying to bosons and thus, the tau decay channel has the potential to be enlightening when it comes to examining the properties of this new particle.

One of the difficulties in this particular decay channel however, is that the tau is itself not stable and will further decay into either leptons or hadrons with the addition of neutrinos, which leads to difficulties in reconstructing the Higgs mass. For this reason, it will be useful to examine ways to optimize the analysis for this particular channel. One possibility, which is being explored at Simon Fraser University (SFU) in Burnaby, British Columbia is the use of boosted decision and regression trees. Of interest in the following report will be the application of boosted regression trees to the reconstruction of the Higgs mass as a competing method to the Missing Mass Calculator, which is currently used.

### 1.1 ATLAS

CERN, the European Center for Nuclear Research, represents an amazing international collaborative project and is home to the Large Hadron Collider (LHC). At 27 kilometers in circumference,

the LHC is the world's largest particle accelerator. Located 100 meters below Switzerland and France, the LHC accelerates bunches of $1.1\times10^{11}$ protons 50 nanoseconds apart with energies of 4 TeV per beam, which are then collided at one of four detectors: CMS, ATLAS, ALICE and LHCb [3]. The focus for the rest of the report will be on the ATLAS detector and collaboration, which SFU is a partner in.

The ATLAS detector is a multi-purpose detector, 46 m long and 25 m in diameter, weighing more than 7000 tonnes [4]. The detector is composed of four basic measurement systems which record energy deposits caused by particle emissions following a collision.[1] The first component is the inner detector (ID), which directly surrounds the beam line and consists of a pixel detector, transition radiation tracker (TRT) and semiconductor tracker (SCT) all of which are surrounded by a series of solenoid magnets. The ID has very high granularity and records the particle tracks immediately following the collision. Due to the magnetic field, charged particles are forced in a curved path, distinguishing them from neutral particles. The ID is then followed by the electromagnetic and hadronic calorimeters, which measure energy deposited when particles pass through these regions. The last layer is the muon detector, which measures energy deposits from muons, which pass undetected through the other levels of the detector.

Once the data has been captured, it is sent out to ATLAS collaboration centers across the globe. There are over 3000 members in the ATLAS collaboration, spread over 38 countries and 174 institutions, including SFU [4]. It is at centers such as SFU, where analyses take place to examine the data, and identify and reconstruct interesting physics events. One of the major efforts at SFU is the Higgs $\rightarrow \tau\tau$ decay channel.

## 1.2 Higgs $\rightarrow \tau\tau$

The Higgs $\rightarrow \tau\tau$ decay channel is an interesting decay channel as it has the potential to be very enlightening regarding the true nature of the new boson discovery. For masses below 140 GeV, the SM Higgs has a significant branching ratio to tau pairs. If this channel produces no evidence of the Higgs, then this new boson is potentially a new Higgs boson, but not the SM Higgs.

One of the difficulties in this particular decay channel is that the tau is itself not stable. At 1.78 GeV, the tau is the heaviest of the leptons, and will quickly decay into lighter particles, with an average life time of $290.6 \pm 1.0 \times 10^{-15}$ seconds [5]. The tau will either decay into a lepton with two neutrinos (to conserve lepton flavour) or a hadron (usually a mixture of charged and neutral pions) with one neutrino. Thus, the Higgs $\rightarrow \tau\tau$ decay is split into lepton-lepton, lepton-hadron and hadron-hadron categories based on how the two taus decay.

Unfortunately, there are some difficulties in reconstructing the Higgs from the decay products due to neutrinos, which can only be identified by a missing momentum or energy in the transverse plane. Since the momentum prior to the collision should be in the z direction, conservation of momentum states that all the momenta in the transverse plane must add up to zero. Of course, due to neutrinos, this is not true, and so the tau decay channel is complicated by the introduction of missing transverse energy, $\not{E}_T$. Currently, the neutrino information is determined from the $\not{E}_T$ using a tool called the Missing Mass Calculator (MMC) [6].

## 1.3 Missing Mass Calculator

The Missing Mass Calculator (MMC) is a method developed around 2011 for better reconstructing the Higgs mass, specifically in the $\tau\tau$ decay channel. Depending on the decay method of the two taus, the MMC solves for 6-8 unknowns using the following four equations [6].

[1] The ATLAS detector uses a right-handed set of coordinates, whose origin lies in the center of the detector at the nominal interaction point (IP). The beam line defines the z-axis, the x-axis points towards the center of the LHC ring from the IP and y-axis points upwards. The transverse plane is defined using cylindrical coordinates (r, $\phi$), where $\phi$ is the azimuthal angle around the beam line. All "transverse" variables are projected in the x-y plane. Finally, we define the pseudorapidity as $\eta = -ln(tan(\theta)/2)$

$$\not{E}_{T_x} = p_{mis1} sin\theta_{mis1} cos\phi_{mis1} + p_{mis2} sin\theta_{mis2} cos\phi_{mis2}$$
$$\not{E}_{T_y} = p_{mis1} sin\theta_{mis1} sin\phi_{mis1} + p_{mis2} sin\theta_{mis2} sin\phi_{mis2}$$
$$M^2_{\tau_1} = m^2_{mis1} + m^2_{vis1} + 2\sqrt{p^2_{vis1} + m^2_{vis1}}\sqrt{p^2_{mis1} + m^2_{mis1}} - 2p_{vis1}p_{mis1}cos\Delta\theta_{vm1}$$
$$M^2_{\tau_2} = m^2_{mis2} + m^2_{vis2} + 2\sqrt{p^2_{vis2} + m^2_{vis2}}\sqrt{p^2_{mis2} + m^2_{mis2}} - 2p_{vis2}p_{mis2}cos\Delta\theta_{vm2}$$

Here, the subscripts 1, 2 refer to the two taus coming from the Higgs decay. The $p_{vis}, m_{vis}, \theta_{vis}$ and $\phi_{vis}$ refer to the momentum, mass, polar and azimuthal angles of the visible tau decay products respectively. The unknown variables are therefore all those with the *mis* subscript, and represent the missing mass and momentum carried away by the neutrino(s). The $\theta_{vm}$ represents the polar angle between the missing and visible momentum vectors. Therefore, the system of equations is underdetermined. However, due to the underlying physics, some situations are more likely than others. Other information such as the $\Delta R$ between the neutrino and visible mass is used to determine the most likely solutions, and thereby reconstruct the mass of the Higgs, $M_{\tau\tau}$.

This technique is a great improvement on previous methods in terms of resolution and accuracy. It has also been particularly successful in reconstructing an unbiased Z mass peak. Unfortunately, the optimization process involves retuning for each new data set, which requires a lot of work on the part of a small group of experts, familiar with the tau energy and MET resolution. On an event-to-event basis, the computing time required to calculate the MMC can also dominate the total analysis time depending on the particular situation. Therefore, it would be interesting to explore other methods, which might be more straightforward and faster in both training and calculation. This is where the Boosted Regression Trees have the potential to help.

## 1.4 Boosted Regression Trees

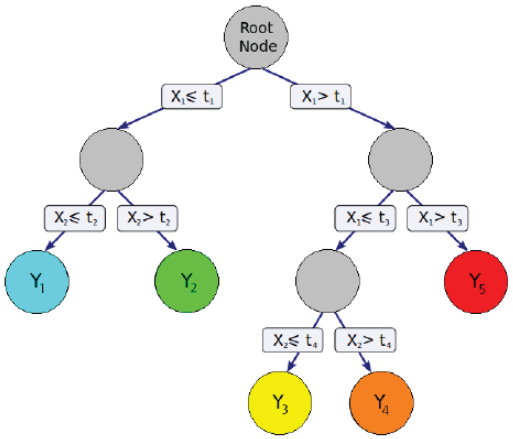


Figure 1: A graphical interpretation of a regression tree trained on two input variables $X_1$ and $X_2$ [7].

Boosted Regression Trees (BRT's) are a multivariate technique based on binary decisions. They aim to estimate a target variable given a series of input variables $X_1, X_2, .., X_n$ by applying a set of cuts $t_1, t_2, ..., t_n$. An algorithm is applied to determine the best cut that can be made on one of those variables by minimizing the average squared error [8].

$$\frac{1}{N}\sum^{N}(y - \hat{y})^2 \tag{1}$$

In Equation 1, y is the regression target of each event and $\hat{y}$ is the mean over all of the events in the node. This process of selecting cuts continues until some pre-defined end criterion is reached,

such as a maximum depth of the tree. In this way, at the end there will be a series of nodes representing one possible estimate of the target. Simply put, what is being done is that the BRT's take a phase space dictated by the input variables and chop it up into regions each of which point to a particular value of the target. Graphically, this process is shown in Figure 1. After all cuts are completed, there will still be events whose average squared error is quite large. In order to minimize the effect of these events, they are reweighted more strongly, and then a new regression tree is trained. This process is what is referred to as **boosting**. Therefore, the regression trees are also quite robust against statistical fluctuations. Furthermore, when the boosting is then done, the final result is an average of the "forest" of individual tree outputs. These methods are implemented practically using the Toolkit for MultiVariate Analysis (TMVA) in ROOT.

### 1.5 TMVA

TMVA provides a simple interface for specifying training and testing parameters in BRT's. Regression analyses using TMVA are split into two sections of training and application [8].

The first component, training, is implemented using the TMVA Factory. This tool trains, tests and evaluates the BRT's. Before the training can begin however, the input variables, target and input data set must be specified and given to the TMVA Factory. The factory then prepares the data set by splitting it evenly between a training and a testing section. Then the optimized cuts and weights are determined and stored in a binary file, called a weight file, which can later be accessed for further analysis. The testing and evaluation steps, also implemented using the factory, are used as checks of the training.

After training, the BRT's can be applied to other data samples using the TMVA Reader. It is necessary to give the reader the input variables in the same order as specified in the training in an array form. Once these variable arrays are filled and reader takes the weight file produced during training and evaluates the regression output, which can be read out for each event and compared to expected values.

## 2 Analysis and Results

Though the primary goal of this analysis was to determine whether BRT's could be used as an effective replacement for the MMC, in the process, it has progressed further into an examination of the method itself. The initial analysis is somewhat naive, using default training parameters and existing data samples. Unfortunately, these data samples have no information about the undecayed taus without the detector simulation effects, known as "truth" information, and so the input variables used initially were fully reconstructed and based on a signal/background type analysis.

| **Variable** | **Description** |
|---|---|
| $MET$ | Missing transverse energy |
| $dphi\_met\_lep$ | Delta phi between the missing transverse energy and lepton |
| $dr\_tau\_lep$ | Delta R between the tau and the lepton |
| $leadJetPt$ | Momentum of the leading jet |
| $mass\_transverse\_met\_lep$ | Transverse mass from the lepton and missing transverse energy |
| $mass\_transverse\_met\_tau$ | Transverse mass from the missing transverse energy and tau |
| $mass\_vis\_tau\_lep$ | The visible mass from the tau and lepton |
| $pt\_ratio\_tau\_lep$ | Ratio of the tau momentum to the lepton momentum |
| $pt\_vector\_sum\_all$ | Vector sum of all momentum in the decay |
| $sumPt$ | Sum of the momenta |
| $tau\_fourvect.fE$ | Energy of the tau |
| $lep\_fourvect.fE$ | Energy of the lepton |

Table 1: A list of all input variables used in the initial reconstructed data analysis

## 2.1 Reconstructed Data Analysis

Using previously simulated data sets of $H \rightarrow \tau\tau$ for Higgs masses ranging from 100 GeV to 150 GeV in 5 GeV increments, the target for the regression was chosen to be the true Higgs mass. The input variables are as specified in Table 1. Furthermore, since this analysis was begun based on an example regression analysis, the training parameters, as specified in Table 2, are somewhat arbitrary.

| Parameter | Description | Value |
|---|---|---|
| NTrees | Number of trees used in boosting | 100 |
| nEventsMin | Minimum number of events required in an end node | 5 |
| BoostType | Boosting method | AdaBoostR2 |
| AdaBoostBeta | Boosting parameter | 0.2 |
| SeparationType | The quantity being minimized at each step | RegressionVariance |
| nCuts | Number of cuts being applied | 20 |

Table 2: A list of default training parameters obtained from an example regression analysis in TMVA [8] and used in the mass reconstruction analysis.

The analysis is based on a series of data samples with Higgs masses varying from 100 GeV to 150 GeV in 5 GeV increments. To ensure independent training and testing events, the input samples are divided in two parts prior to training. Following this, the BRT's are trained on the true Higgs mass target using the distribution shown in Figure 2a. The BRT's are then applied to the remaining events, the regression output is obtained and filled into a histogram. The mean and RMS of the regression histogram is obtained and plotted as a function of the true mass of the Higgs for each mass sample.

Ideally, in a plot of the regression output as a function of the true Higgs mass, the slope of the best fit line should be close to 1. Unfortunately, this is not what is seen in Figure 2b. The slope of the plot is in fact $0.226 \pm 0.002$ and biased very strongly towards 125 GeV- the center of the target distribution. Since the target distribution was suspiciously discrete, in can be useful to examine the effect changing the target distribution has on the slope of the plot.

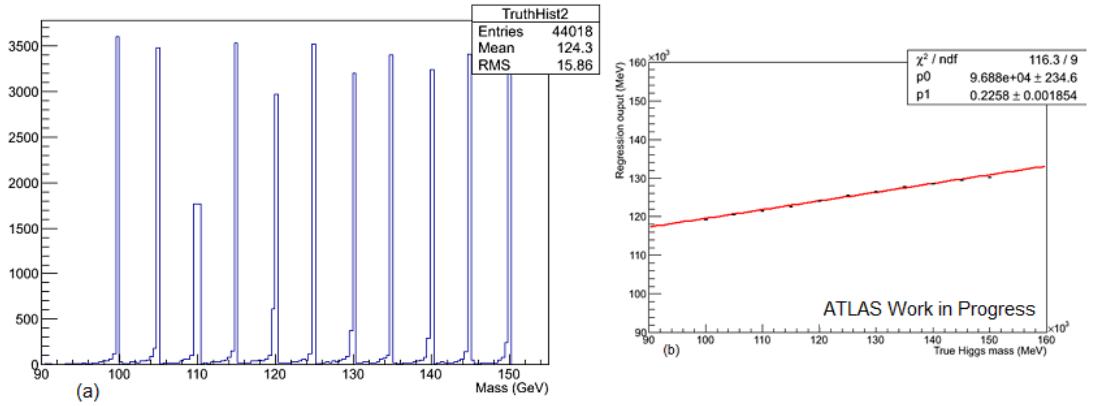


Figure 2: **(a)** The initial distribution of the Higgs masses used for training. **(b)** The regression output of the Boosted Regression Tree as a function of the true Higgs mass. Note that the error bars, which represent the error on the mean of the regression output, are very small.

To this end, an artificial mass variable is used so that the distribution could be controlled very accurately without changing any other inputs. The first step is to smear the target distribution, and so for each mass sample, the Higgs mass was filled with a randomly generated Gaussian distribution centered at the mass in question. Between 100 GeV and 150 GeV (non inclusive), the Gaussians are given a standard deviation of 50 GeV whereas the endpoints are given a standard

deviation of 100 GeV to extend the tails. When testing on this new distribution, the regression performance improved slightly, giving a slope of 0.288 ± 0.006 as seen in Figure 3. However, this

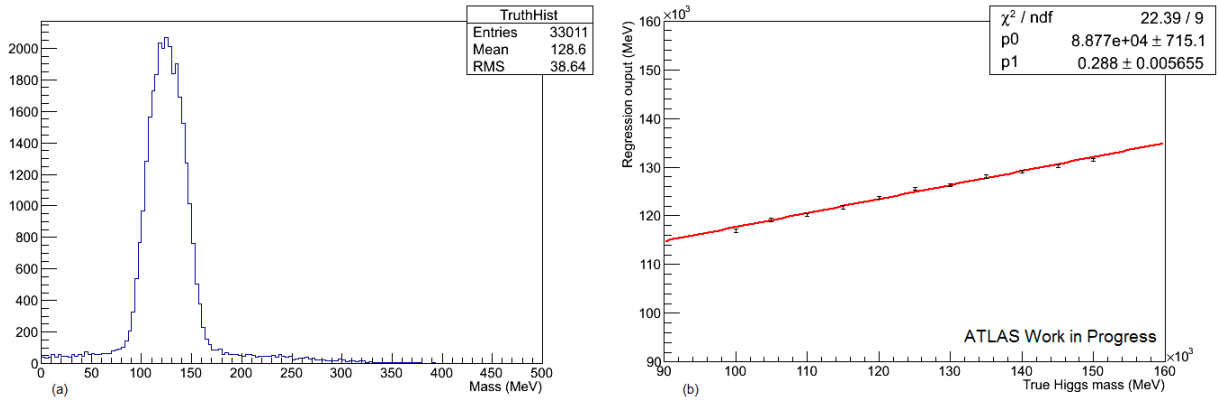


Figure 3: **(a)** The artificial Higgs mass distribution using randomly generated Gaussians for each mass. **(b)** The regression output of the Boosted Regression Tree as a function of the true Higgs mass.

is not a very large improvement. The next test is therefore to increase the tails relative to the center of the distribution since the mass points still seem to be biased quite strongly towards the 125 GeV point.

In order to do this, the Higgs mass is filled with a randomly generated uniform distribution, once again centered at the mass of interest. For masses between 100 GeV and 150 GeV the distribution has a width of 5 GeV but the end points, the tails are extended down to 0 GeV and up to 250 GeV. This leads to a large improvement in the slope, up to 0.628 ± 0.006 as shown in Figure 4. These results show that the tails of the distribution do indeed play an important role in

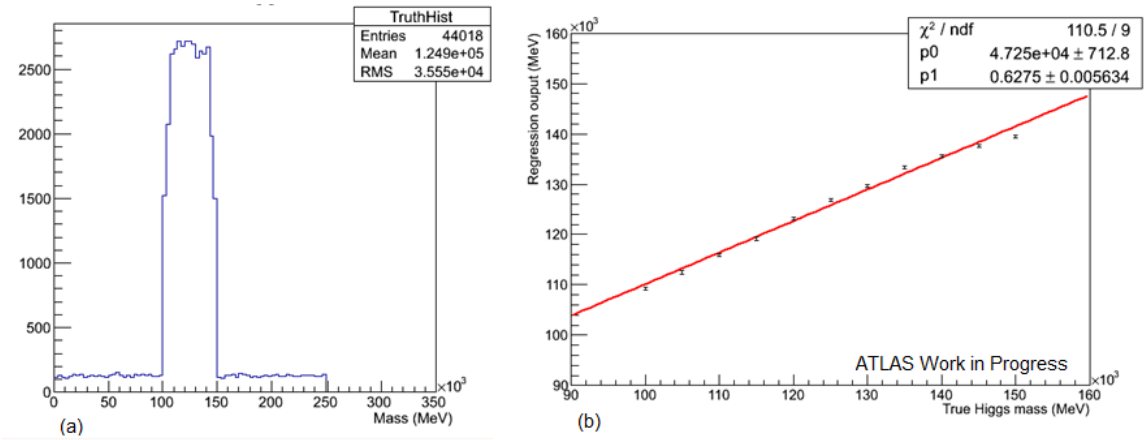


Figure 4: **(a)** The distribution of the artificial Higgs mass variable using randomly generated uniform distributions. **(b)** The regression output of the Boosted Regression Tree as a function of the true Higgs mass.

the regression performance. Intuition would suggest that when events filter through the regression tree, they are pulled towards the side of the distribution with more events. Therefore, at the tail ends of the training distribution, the regression output would be high for the low masses and low for high masses. By smearing the distribution, this effect is mitigated.

The training parameters of the BRT's may also affect the method's performance. In order to test these effects, all the parameters in Table 2 are used as a default while varying only one at a

time. The results are summarized in Table 3. There seems to be little improvement from changing

| Parameter | | Value | | Slope |
|---|---|---|---|---|
| NTrees (more trees) | = | 1000 | → | 0.6527 |
| nEventsMin (more events per node) | = | 50 | → | 0.71 |
| MaxDepth(deeper trees) | = | 200 | → | 0.5943 |
| AdaBoostBeta (larger boosting parameter) | = | 0.5 | → | 0.6295 |
| AdaBoostBeta (smaller boosting parameter) | = | 0.05 | → | 0.5917 |

Table 3: A table outlining the various changes to input parameters when testing regression performance.

the training parameters other than when increasing the minimum number of events per node. It should be noted as well that though increasing the number of trees used in boosting does improve the slope somewhat, the training time also increases considerably, so there is some trade off.

The variable nEventsMin, specifying the minimum number of events per node is interesting because it has a strong influence on overtraining. Overtraining can occur when BRT's are trained on statistical fluctuations particular to the training sample, which can lead to other samples being estimated incorrectly. This is usually clearest when looking at the deviation between the regression output and the truth value for the training and testing. If there is overtraining, the testing deviation should be much larger than the training. This indeed is present for a minimum value of 5, but it disappears for larger values as shown in Table 4. It is interesting to note that

| nEventsMin | Training RMS (GeV) | Testing RMS (GeV) | Slope |
|---|---|---|---|
| 150 | 30.12 | 34.3 | 0.6520 |
| 200 | 31.03 | 33.52 | 0.6382 |
| 250 | 31.61 | 33.78 | 0.5711 |
| 300 | 31.33 | 33.85 | 0.5548 |
| 350 | 31.61 | 33.67 | 0.4973 |

Table 4: The RMS values of the training and testing distributions as the minimum number of events in each end node is varied.With a large number of events required in each end node, overtraining was avoided.

the slope of the regression as a function of the true Higgs mass decreases with increasing value of the nEventsMin variable. Therefore, a balance must be struck between having a good slope and preventing overtraining. For this reason, all future tests are conducted using a minimum of 200 events.

## 2.2 Truth Level Data Analysis

At this point, new data samples containing truth-level variables were generated using PowHeg [9, 10, 11] and Herwig [12] to determine whether it was possible to better understand the results seen from the reconstructed variables. According to physical reasoning, it is simple to reconstruct the mass of the Higgs boson from the four vectors of its decay products, in this case two taus. Therefore, the energy and three momenta of each of the taus at truth level are used as input variables in the further analysis. The target is also changed to be the true Higgs mass once again rather than the artificial mass variable used previously. Due to the very narrow peaks for each Higgs mass however, the distribution of the target is in fact nearly discrete. The result is a slope of $0.909 \pm 0.0003$ as shown in Figure 5. This suggests that the BRT's effectively approximate the true Higgs mass using the truth level tau four vectors.

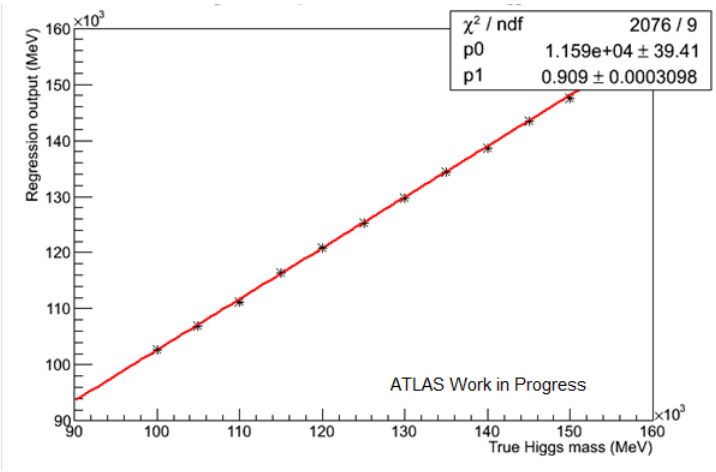


Figure 5: The regression output of the Boosted Regression Tree as a function of the true Higgs mass using the tau four vectors as the input variables.

# 3 Discussion

These last results raise some interesting questions however concerning whether the truth level tau variables are special in some way or whether the regression would work as well for reconstructed variables or other truth level variables. Some interesting next steps to further explore this would be firstly to look at other truth-level variables. For example, using the decay products of the taus, i.e. the leptons and hadrons, as the input variables. Very likely, this would create additional complications due to the introduction of the neutrinos. The tau four vectors already take into account the missing energy carried away by the neutrinos, but when using only the visible tau decay products at truth level, the neutrino contribution would be unknown. On the other end of the spectrum, it would also be interesting to see what the regression output would look like if the input variables were simply the Higgs four vector. Past attempts with other variables seem to indicate that BRT's do not perform very well with a very small number of variables, so it would seem unlikely. During training, with very few variables, the average squared error cannot be minimized properly. However, it is not yet quite clear why this is.

Another option that could produce interesting results would be to smear the truth level variables in such a way so that they mimic the resolution of the reconstructed variables. In this way, it would be possible to determine what effect the resolution of the input variables has on the regression output. This test could potentially answer the question as to whether the truth-level tau vectors are special in some way in regards to the BRT's.

# 4 Conclusions

This analysis has explored many aspects of Boosted Regression Trees in the application of reconstructing the Higgs boson mass. BRT's are a promising technique and offer a resolution comparable to the current MMC, but further studies are still required to understand the potential biases. It is clear that the regression output is quite strongly affected by the distribution of the target variable, especially the inclusion of tails. Varying the training parameters on the other hand makes very little difference with the exception of the minimum number of events per node. This variable not only strongly affects the accuracy of the regression but also prevents overtraining when large enough and so there is a trade off between the overtraining and accuracy that needs to be dealt with. Finally, the results also indicate that the input variables have a large effect on the regression performance. However, exactly how or why certain variables work better than others is not yet quite clear, and must be explored further.

# 5 Acknowledgements

A special thanks to Dr. Dugan O'Neil and Dr. Andres Tanasijczuk for their help and guidance throughout this summer. I would also like to thank Koos van Nieuwkoop for his suggestions and insight. I am infinitely greatful to the CERN Summer Student Program and IPP for this fantastic opportunity and to all of my fellow summer students (both at CERN and SFU) for letting me bounce ideas off of them and for making this summer a memorable one.

# References

[1] ATLAS Collaboration, *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC*, Phys.Lett.B, 716 (2012) 1-29

[2] CMS Collaboration, *Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC*, Phys.Lett.B, 716 (2012)30-61.

[3] C. Lefevre,*CERN Faq: LHC, the guide* (2009).

[4] ATLAS Collaboration, *ATLAS Fact sheet* (2010).

[5] J. Beringer, et al., *Review of particle physics* Phys.Rev.D 86.1 (2012).

[6] A. Elagin, P.Murat, A.Pranko, A.Safonov, *A New Mass Reconstruction Technique for Resonances Decaying to di-tau*, NIM A 654(1),(2011)481-489.

[7] K. van Nieuwkoop, *Bringing the Higgs Boson to Rest* MSc Diss. SFU Department of Physics, (2013).

[8] A. Hoecker, et al., *TMVA-toolkit for multivariate data analysis*, (2007) arXiv:physics/0703039 [physics.data-an].

[9] P. Nason, *A New method for combining NLO QCD with shower Monte Carlo algorithms*, JHEP 1785 0411 (2004) 040, arXiv:hep-ph/0409146 [hep-ph].

[10] S. Frixione, P. Nason, and C. Oleari, *Matching NLO QCD computations with Parton Shower Simulations: the POWHEG method*, JHEP 0711 (2007) 070, arXiv:0709.2092 [hep-ph].

[11] S. Alioli, P. Nason, C. Oleari, and E. Re, *A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX*, JHEP 1006 (2010) 043, arXiv:1002.2581 [hep-ph].

[12] G. Corcella et al., *HERWIG 6.5 release note*, (2005) arXiv:hep-ph/0210213 [hep-ph].