

# Optimizing Higgs Boson Analysis at DØ

John Sandy (Texas Tech University)

Michael P. Cooke (Fermi National Accelerator Laboratory)

Ryuji Yamada (Fermi National Accelerator Laboratory)

August 3, 2012

The Higgs boson has been evading physicists for over 40 years, but that has done little to deter the search. As a key component of the Standard Model (SM), it has been the focus of many experiments and collaborations around the world. With exciting results beginning to stream in from the forefronts of particle physics, this summer is an extremely exciting time to be searching for the Higgs boson. At DØ, improvement continues on the analyses that use the data that was collected during Run II of the Tevatron. As final data analysis is being completed, visualization of events becomes important, and so the use of event display software is examined. In addition, to improve sensitivity of the search in the lepton, neutrino and 2- or 3-

jet final states of the Higgs boson, several multivariate discriminators are trained to select the Higgs boson signal over specific backgrounds.

## I. INTRODUCTION

### A. The Higgs boson

Ever since Peter Higgs and his colleagues published their respective papers concerning the Higgs mechanism in 1964, physicists have been trying to find the elusive Higgs boson<sup>1</sup>. It has been a focus of massive collaborative experiments of hundreds of scientists at places like Fermilab and CERN. Theorists around the world have been scrutinizing the theory for decades, looking for insights and alternatives. But just what is a Higgs boson and why should we, as scientists, care so deeply to find it?

When J.J. Thompson discovered the electron in 1897, he unwittingly discovered what is now considered the first fundamental particle in nature<sup>1</sup>. Since then, physicists have been hard at work discovering a veritable zoo of other particles and have categorized them in what is now called the Standard Model (SM) of particle physics. The fundamental particles of the SM are shown in Fig. 1. The SM contains six quarks, each with an antiparticle, and six leptons, with antiparticles of their own<sup>1</sup>. All matter that we know of is made up of these quarks and leptons. The SM also contains four force carriers that account for all the interactions that affect matter, except for gravity, which the SM leaves out entirely. This theory of matter and its interactions brought with it a disturbing prediction: the SM particles should be massless. This, of course, we know to be untrue; many of the SM particles have mass, some even nearly the mass of a gold atom! This meant one of two things: the SM was either completely wrong or it was missing something important. Then Higgs, Guralnick, Hagen, Kibble, Englert and Brout made a critical addition to the SM. This addition adds a new scalar field that is non-zero in its ground state, i.e.,

“the vacuum,” which became known as the Higgs field<sup>1</sup>. This field permeates all space and gives mass to the weak force carrying bosons, quarks and charged leptons. Just as the photon is the excitement of the electromagnetic field, the Higgs boson is the excitement of the Higgs field.

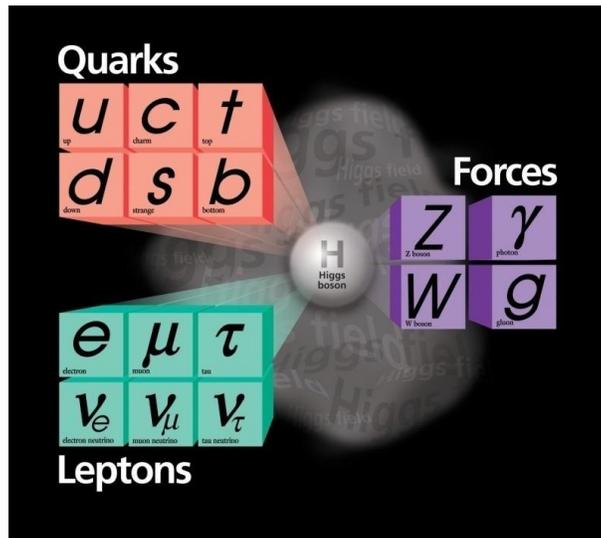


FIG 1. The fundamental particles of the Standard Model of particle physics.<sup>3</sup>

This patched the SM to its current state and this is why physicists have been so diligently searching for the SM Higgs boson. This elusive particle carries with it the life or death of the SM we know today. Without it, the SM is doomed as a broken theory. Yet, even if a Higgs is discovered, it will leave many questions to be answered. The SM predicts only one Higgs boson, but it is not the only theory that uses a Higgs-like particle to add mass.<sup>2</sup> This means that, only experimentally, will we be able to pin down exactly what kind of Higgs boson might exist. There exist many extensions to the SM, such as supersymmetry (SUSY), which predicts super symmetric partners to all the SM particles and includes five (or more) Higgs bosons.<sup>2</sup> Clearly, exploration into the Higgs is essential to further our fundamental understanding of the universe.

## B. Higgs searches

The Higgs boson is a notoriously difficult particle to search for. The SM predicts how the Higgs will decay once produced in collisions of sufficiently high energy to create it. This

prediction includes its production rate and total set of branching ratios, which describe how often a Higgs will decay into all other particles. Simply put, the branching ratio for a particular decay process is the relative probability of how often the Higgs will decay into that specific set of particles. These branching ratios are dependent on what mass the Higgs boson actually is, so searching over a wide mass range becomes complicated quickly. The branching ratios for the most relevant mass range are shown in Fig. 2. For example, at a mass of  $125 \text{ GeV}/c^2$ , the Higgs is roughly 1 000 times more likely to decay into a pair of bottom quarks ( $b\bar{b}$ ) than two photons ( $\gamma\gamma$ ).<sup>3</sup> To complicate things more, there are many SM processes that are very similar to the signal that a Higgs leaves behind, meaning the SM background is very hard to separate from the Higgs signal.

In order to overcome these obstacles, Higgs search teams are split into many different groups. Each group is responsible for searching for the Higgs that is produced by a certain decay mode within the SM. In order to do this, data from collisions is categorized based on very basic properties, such as “does the event contain a pair of leptons?” These different categories of data are then used by the groups who would expect to see those objects in their decay channel. This allows experimenters to focus on eliminating only the backgrounds that are present in their category and mimicking the decay process they are hunting for. Each group’s exact search method is different due to the variety of problems that arise in this complicated search. Yet, this difficulty allows for the development of many new analysis techniques.

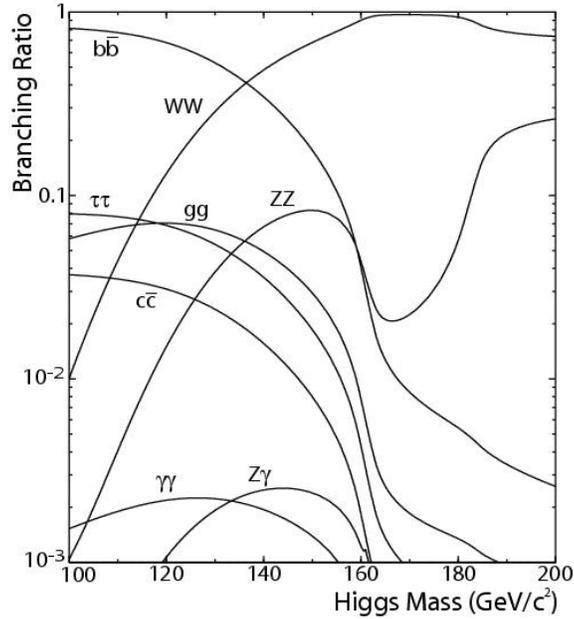


FIG. 2. The branching ratios of the SM Higgs boson at masses ranging from 100–200  $\text{GeV}/c^2$ .<sup>3</sup>

### C. Recent results

The summer of 2012 will always be remembered as the summer of the Higgs boson by the particle physics community. The Tevatron’s two experiments, DØ and CDF, have finished their analysis on the Higgs search using the full Run II dataset and have announced their results. As the preliminary results publication states, the combined results of the entire Higgs search at the Tevatron shows evidence for a new particle in the mass range around 130 GeV with a global significance of 3.1 standard deviations from the SM background<sup>4</sup>. Higgs analyses are ongoing and further improvement may yet come to their studies of Higgs boson decaying to  $b\bar{b}$ .

On July 4, 2012 at 9 AM local time at CERN in Geneva, Switzerland, the two main LHC experiments, CMS and ATLAS, presented their latest results on the Higgs search. Each experiment observed signal for a new boson of mass around 126 GeV, with a local statistical significance of 5.0 standard deviations away from the SM background.<sup>5</sup> They had made the discovery of a new particle that matched the profile of a SM Higgs. This announcement has been the source of much excitement in the particle physics community. However, there is still a great

deal of analysis left to determine just what this new particle has in store for the future of the SM and particle physics in general.

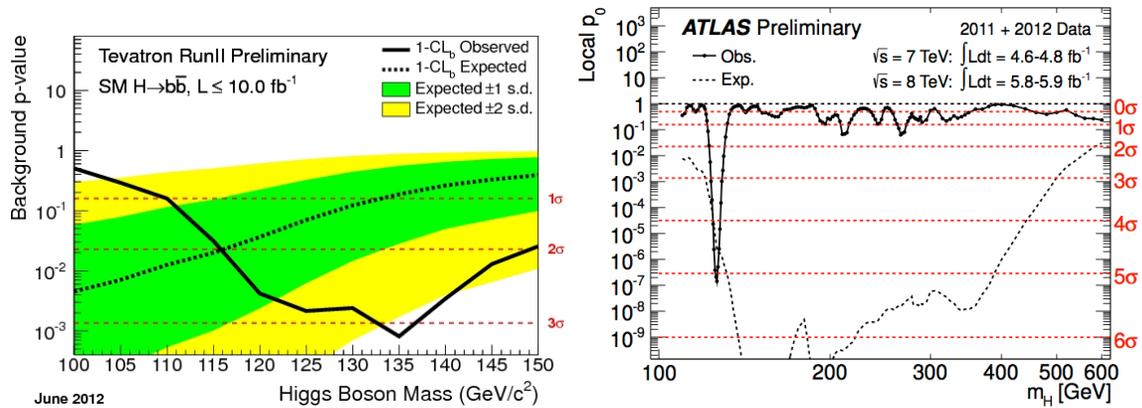


FIG. 3. The most recent Tevatron Higgs search results<sup>4</sup> (right) and the results ATLAS showed on July 4th<sup>5</sup> (left).

## II. DØ HIGGS ANALYSIS

At DØ, the Higgs search team is still actively seeking improvement over the summer 2012 results. The Tevatron provided proton-antiproton ( $p\bar{p}$ ) collisions with a center of mass (CM) energy of 1.96 TeV, which were recorded by the DØ detector. The detector consists of an inner silicon detector, scintillating fiber tracking system, a uranium liquid argon calorimeter and a large muon system as shown in Fig. 4. All the detector elements amount to over 900 000 individual channels that record data for use in analysis.<sup>6</sup>

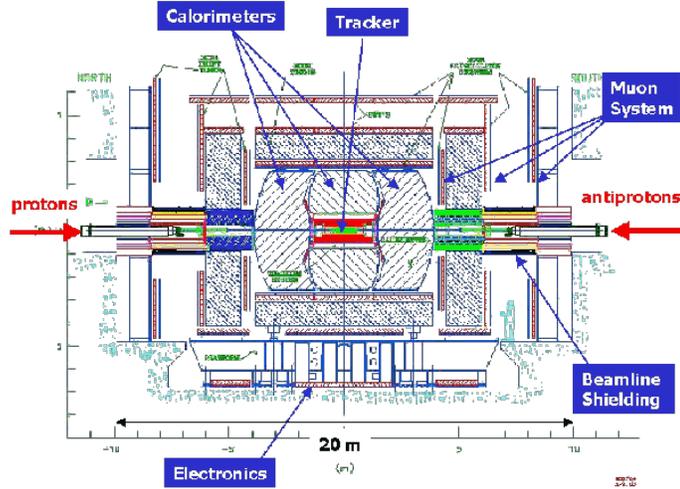


FIG. 4. A cross section of the DØ detector.<sup>6</sup>

Over the 10-year span of the detector’s “Run II,” over  $10 \text{ fb}^{-1}$  of data was recorded, which amounted to over 10 billion events to use in analysis.<sup>7</sup> As Fig. 2 indicates, the  $H \rightarrow b\bar{b}$  and  $H \rightarrow WW$  branching ratios dominate the low mass region, where experiments at CERN have discovered a new particle. In the DØ experiment, sensitivity of these two final states is dominated by one of four decay methods:  $H \rightarrow WW \rightarrow l\nu l\nu$ ,  $WH \rightarrow l\nu b\bar{b}$ ,  $ZH \rightarrow \nu\nu b\bar{b}$  or  $ZH \rightarrow ll b\bar{b}$ <sup>3</sup>, where  $l$ ’s are some charged lepton and  $\nu$ ’s are a neutrino. Whenever a quark is produced in high-energy collisions, it creates a spray of particles known as “jets.” Because the  $b$  quark is present in several of the final states of the Higgs boson, experimenters at DØ developed a process called “b-tagging.” This tool examines events from the detector and categorizes them based on their apparent number of  $b$  quarks. Events are tagged with 0, 1, or 2  $b$  quarks and categorized as 0-tag, 1-tag or 2-tag respectively. The 1 and 2 tag categories are then further subdivided into how well the jets match the theoretical profile of a jet produced by a  $b$  quark. This results in 1 loose and tight tag and 2 loose, medium and tight tag categories. An event display showing a 2 tight tag category Higgs candidate event in the DØ detector is shown in Fig. 5.

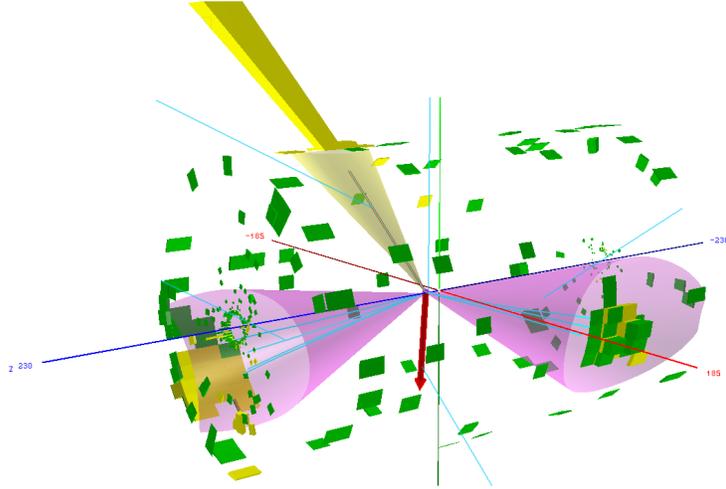


FIG. 5. An event display showing a Higgs candidate event with a final state containing an electron (yellow cone), 2 jets (purple cone), and neutrino (red arrow).

In order to obtain event displays like the one in Fig. 4, an adaptation of D0CAFVis is used. D0CAFVis is a ROOT-based [8] software package that was developed by DØ's Dr. Michael Wang to generate 3D recreations of collisions in the DØ detector, but last updated in August 2010. As such, its ability to read and display the most recent data was originally hindered. With help from DØ support staff, D0CAFVis was recompiled and is now able to read and display all events from Run II. This newly compiled version, as well as new and detailed instructions on how to use D0CAFVis, will be submitted to the DØWiki and/or as a collaboration note. This will allow DØ collaborators to continue using the D0CAFVis software to generate event displays for use in future talks and publications. Several such event displays have already been incorporated into several DØ presentations.

Before visualizing these events, however, they must be picked from the sea of background. The number of Higgs events expected at DØ for dominant decay methods per  $\text{fb}^{-1}$  is shown in Fig. 6 for a large mass range.

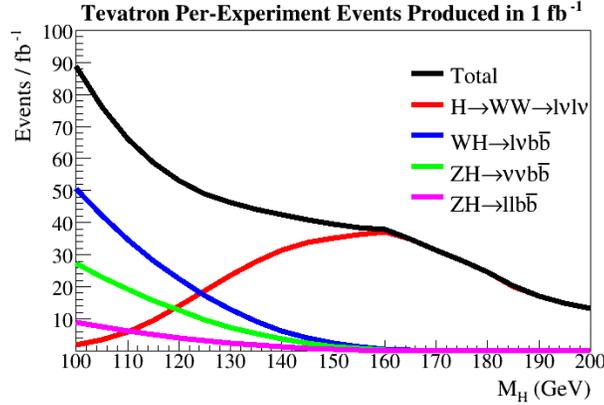


FIG. 6. The expected number of SM Higgs boson events produced in an experiment at the Tevatron per  $\text{fb}^{-1}$  of data.<sup>3</sup>

As Fig.4 demonstrates, with roughly  $10 \text{ fb}^{-1}$  of data, the amount of events that are actual Higgs production, is absolutely miniscule in comparison with the amount of data that is being analyzed, leaving the rest of the events as background. In order to separate the background from the data, a tool called a multivariate analysis (MVA) is used. As the name implies, this tool uses an analysis technique that is dependent on multiple input variables to separate data and background. The variables selected for use in the MVA are very important, as they decide how well the MVA will be able to separate signal from the background. Variable selection is carried out concurrently while “training” the MVA. Training is an automated learning process where the framework creates a set of variable specific parameters for the MVA to use in data analysis. At the beginning of the summer, concerns were raised about how well trained the 2- and 3-jet MVAs at  $D\emptyset$  were trained for separating Higgs data from the backgrounds of top quark pair production ( $t\bar{t}$ ), two vector bosons (VV), and a vector boson plus jets (V+jets). Each of these different backgrounds mimics the Higgs signal in the channels we are concerned with. In order to improve analysis sensitivity, they have been retrained.

The process of retraining the MVAs began with an entirely new variable list for each background to examine. The utilized  $D\emptyset$  analysis framework contains hundreds of variables that

could be used. However, it would be impractical to sort through every single one. Simulated data obtained via modeling how the DØ detector should record data and how the actual data should behave is used to obtain a variable list of manageable size to begin training. This set of simulated data is analyzed to obtain how well each variable contributes to separating background from data in each background category. While this method does not perfectly predict how important every variable is, it allows a short list of variables to be generated.

To begin retraining the Higgs vs.  $t\bar{t}$ , VV, and V+jets MVAs, a short list of 12 variables for each tag category of each background in each jet multiplicity was obtained and run through the analysis framework. The framework reads a configuration file which tells it how to run the MVA and with what variables. Training an MVA is a different process than actually analyzing data, so their configuration files are different. Upon completion of a training run, several useful outputs are created. The first is a log file of the training run, which lists, among other things, how much each variable contributed to the separating of data and background, or its “importance.” The second is a series of plots. The first plot conveys how closely related each variable is in the information it is contributing, or a “correlation plot.” The remaining plots show how well the variable is predicted to separate the background from signal.

Using these training outputs, the variable list can be shortened, or pruned, to obtain the shortest list of most important and least noisy variables possible for actual data analysis. The first step in pruning the variable lists is to check the importance rankings of each one. In this retraining, any variable that was an entire order of magnitude less than the most important variable or lower was cut from its respective list. This ensures that the MVA is not using any variable that is not contributing in a meaningful way to separating the data from the background. Second, correlation plots were examined. In this retraining, the less important of any two

variables that were higher than 90% correlated were cut from their respective lists. An example of a correlation plot is shown in Fig. 7.

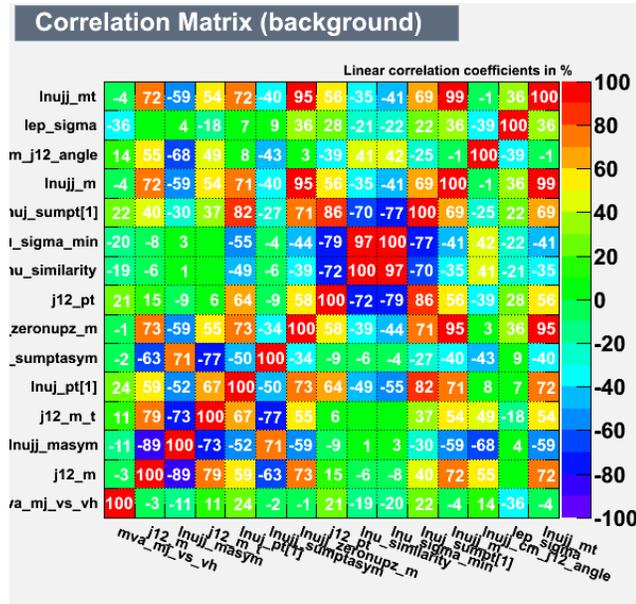


FIG. 7. An example of a correlation matrix that shows several highly correlated variables.

Pruning highly correlated variables removes redundant information and can help reduce noise in the final training. Finally, the modeling of the remaining variables was examined. In an earlier stage of analysis, plots were created which show how well each variable’s simulated results agree with what the detector actually records. When these plots are examined, certain variables do not show good agreement between simulation and data. If a variable is too poorly modeled, it must be thrown out of the training or else it will lead to a poorly modeled MVA output. An example of a poorly modeled variable is shown in Fig. 8.



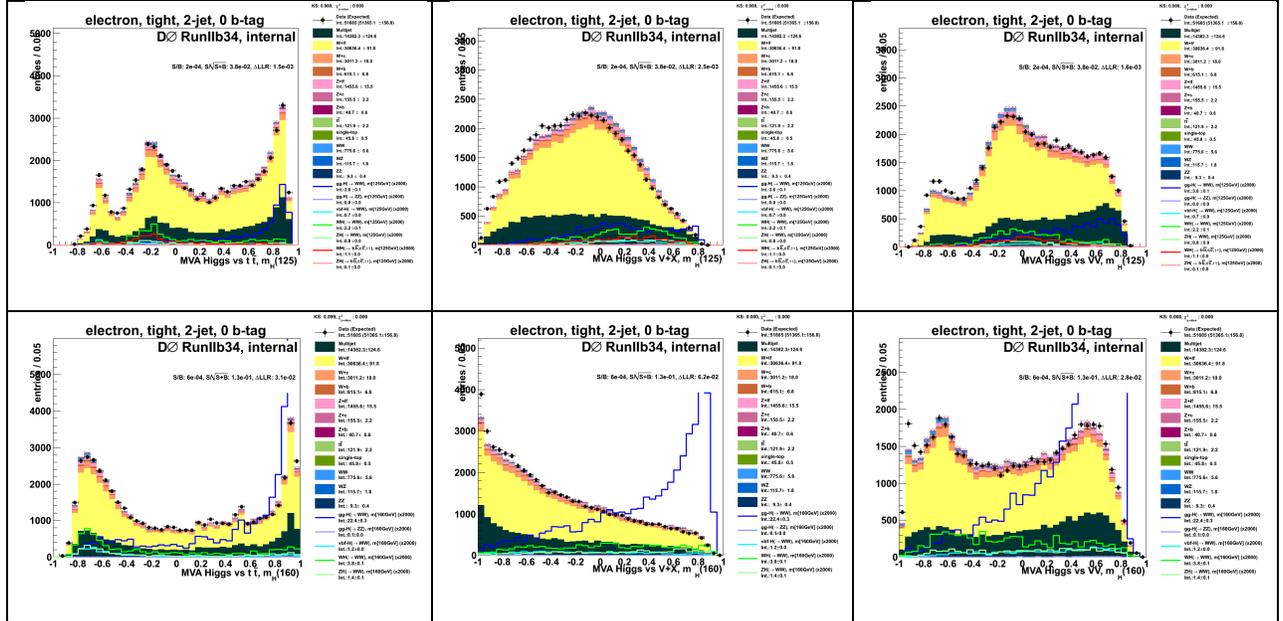


FIG. 9. MVA output distributions for 0-tag, 2-jet Higgs vs.  $t\bar{t}$  (left), V+jets (middle) and VV (right) for Higgs boson mass points of 125 GeV (top) and 160 GeV (bottom). Points on the plot are data, colored histograms represent expected background contributions, solid lines are overlaid contribution from expected signal.

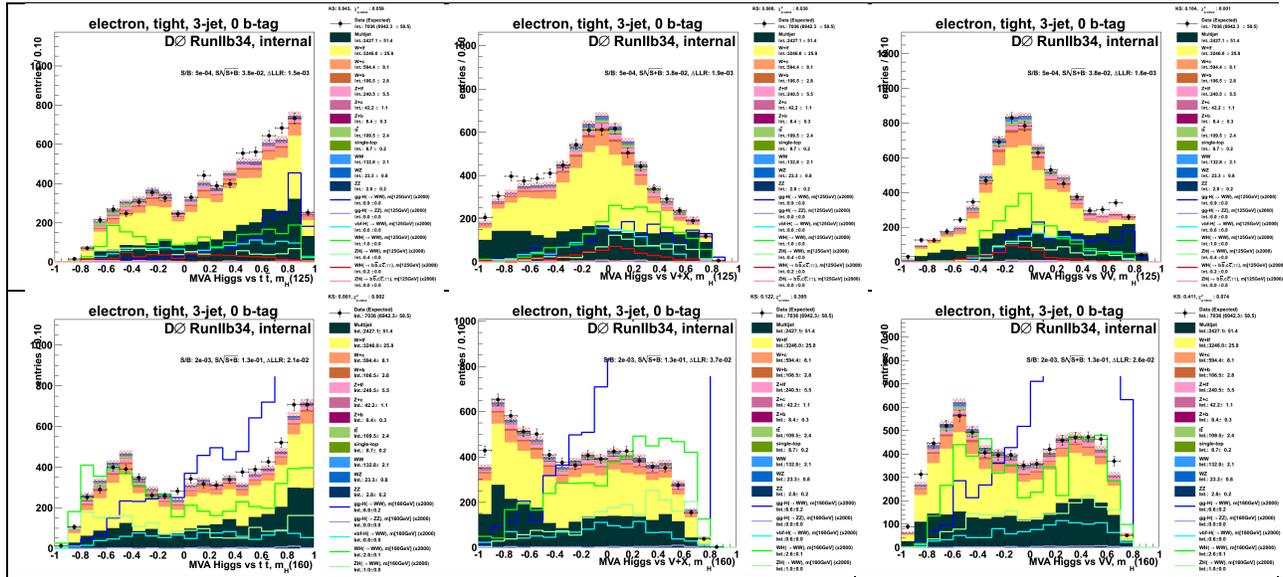


FIG. 10. MVA output distributions for 0-tag, 3-jet Higgs vs.  $t\bar{t}$  (left), V+jets (middle) and VV (right) for Higgs boson mass points of 125 GeV (top) and 160 GeV (bottom). Points on the plot are data, colored histograms represent expected background contributions, solid lines are overlaid contribution from expected signal.

These final MVA outputs, specifically because they are the high statistics 0-tag category, show how well the MVA has been modeled. If these plots show poor modeling, then more

variable pruning can be done to improve the tool until it is fully optimized. Figures 9 and 10 show the preliminary results of the MVA training. Further variable pruning was completed to improve poor modeling in the low MVA region of the low mass final MVAs.

### III. CONCLUSIONS

The process of training secondary MVAs for Higgs vs.  $t\bar{t}$ ,  $v\bar{v}$  and  $v\bar{v}$  show promise for improving the overall signal sensitivity in the Higgs analysis at  $D\bar{D}$ . A similar technique has already been used to gain significant improvement in other analysis channels here, and there is reason to believe that these newly trained MVAs will bring improvement to the 2- and 3-jet channels as well. The final results of the training show how important modeling can be. Poorly modeled variables were pruned from the variable list used to produce the plots in Figs. 9-10 and the end result is shown in Figs. 11-12.

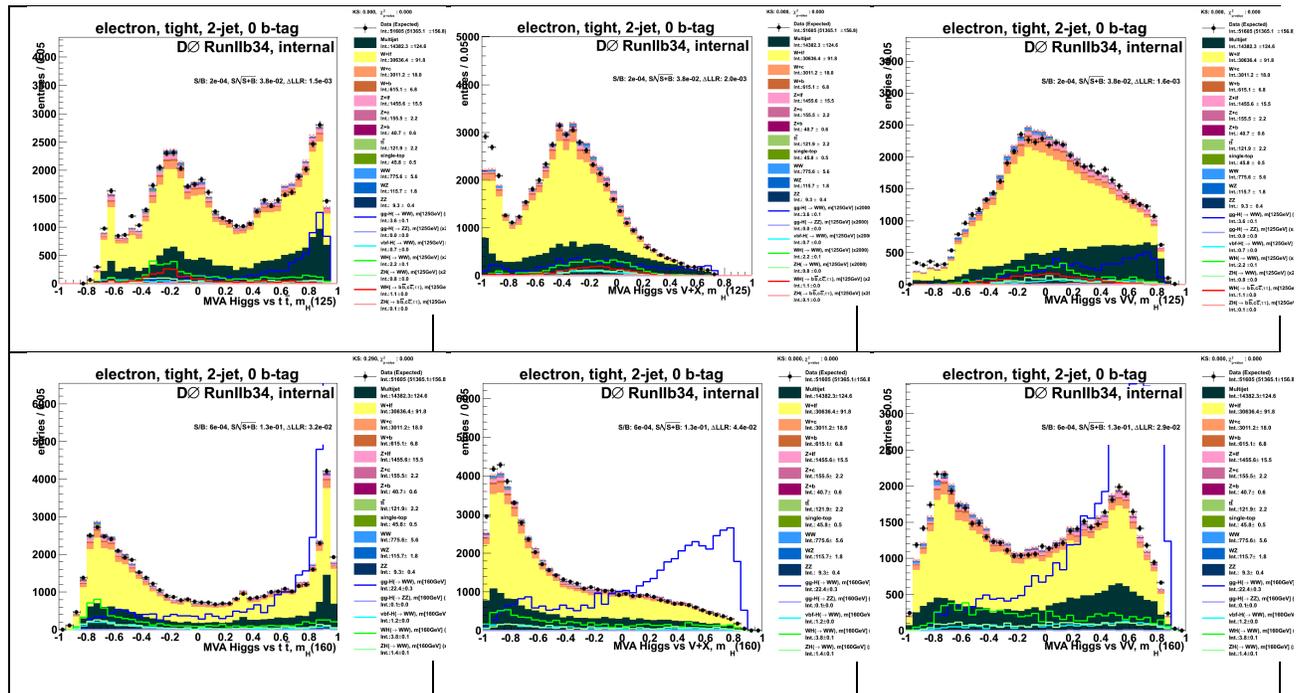


FIG 11. Further pruned MVA output distributions for 0-tag, 2-jet Higgs vs.  $t\bar{t}$  (left),  $V+\text{jets}$  (middle) and  $VV$  (right) for Higgs boson mass points of 125 GeV (top) and 160 GeV (bottom). Points on the plot are data, colored histograms represent expected background contributions, solid lines are overlaid contribution from expected signal.



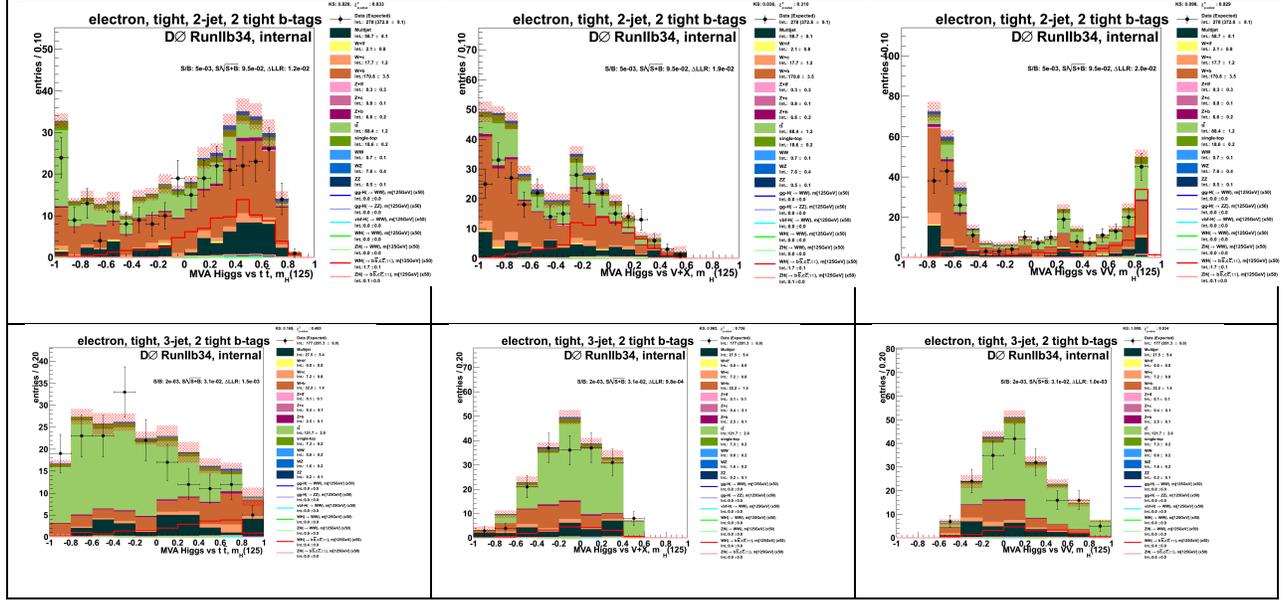


FIG. 13. MVA output distributions for 2 tight,  $b$ -tagged Higgs vs.  $t\bar{t}$  (left),  $V$ +jets (middle) and  $VV$  (right) for 2 (top) and 3 (bottom) jet channels. Points on the plot are data, colored histograms represent expected background contributions, solid lines are overlaid contribution from expected signal.

The final result of all this training will be improvement in Higgs analysis, where these new MVAs can be used to separate Higgs boson events from the backgrounds they have been trained against. These improved MVAs will allow the lepton neutrino plus jets Higgs searches at  $D\bar{O}$  to be further optimized, resulting in more definitive evidence in the SM Higgs boson searches of the Tevatron, and a better understanding of the newly observed particles branching ratios.

#### **IV. ACKNOWLEDGEMENTS**

This work was supported by the U.S. Department of Energy - Office of Science, Office of Workforce Development for Teachers and Scientists, and the Experimental Particle Physics Division at Fermi National Accelerator Laboratory. I would like to thank my mentors Ryuji Yamada and Michael P. Cooke. I would also like to thank Anthony Podkova and Herb Greenlee for their assistance.

## REFERENCES

- <sup>1</sup>D. Griffiths, *Introduction to Elementary Particles*, New York: John Wiley & Sons, Inc., 1987.
- <sup>2</sup>CDF Collaboration, "Higgs in Plain English," [http://www-cdf.fnal.gov/physics/new/hdg/Plain\\_English.html](http://www-cdf.fnal.gov/physics/new/hdg/Plain_English.html)
- <sup>3</sup>"The Tevatron's Massive Search for the Higgs," Seminar by Michael Cooke at Columbia University, Nov. 17, 2010.
- <sup>4</sup>Tevatron New Physics Higgs Working Group, CDF Collaboration, DØ Collaboration (2012), arXiv:1207.0449 [hep-ex].
- <sup>5</sup>Latest update in the search for the Higgs boson, Combined Press Conference by LHC Experiments at CERN, July 4, 2012.
- <sup>6</sup>V. M. Abazov et al. (DØ Collaboration), Nucl. Instrum. Methods Phys. Res., A 565, 463 (2006), [arXiv:physics/0507191v1], "DZero Fact Sheet."
- <sup>7</sup>M. Cooke, "DZero's Data Set" *Fermilab Today*, October 6, 2011.
- <sup>8</sup>Rene Brun and Fons Rademakers, ROOT - An Object-Oriented Data Analysis Framework, Proceedings AIHENP'96 Workshop, Lausanne, Sept. 1996, Nucl. Instrum. Methods Phys. Res., A 389 (1997), 81-86. See also <http://root.cern.ch/>.