Vector Boson Scattering (VBS) Analysis in ZZ + 2 jets Production with Machine Learning

Zunran Guo
University of Rochester

Abstract

With the recent upgrade, the luminosity at the Large Hardon Collider (LHC) is now able to detect the scattering of gauge bosons, which offers a unique opportunity to directly examine the nature of electroweak symmetry breaking (EWSB) in the Standard Model (SM).

In this research, classifiers using boosted decision trees (BDT) with various boosted algorithms are trained and tested to classify background and signal events in the vector boson scattering (VBS) process with ZZ + 2 jets production. This machine learning technique, compared with the previous manual direct variable cut, improves about 7% better at signal efficiency of around 0.80 and about 12.5% better at background rejection of around 0.89.

Background and Method

Vector-boson-scattering (VBS) ZZ + 2 jets production gives a unique opportunity to examine the nature of electroweak symmetry breaking (EWSB) in the Standard Model (SM).

The scattering of gauge bosons violates unitarity at the TeV scale if there is no Higgs boson. The unitarity can be restored by including the Higgs bosons which leads to a delicate cancellation of divergence at high energy. This mechanism can be tested via measuring the VBS production cross sections. Any anomalies will bring up questions into whether the Higgs boson is as predicted in the SM; and whether the EWSB mechanism is as predicted in the SM.

This research focuses on identifying the signal and the background events in the scattering process through a trained classifier based on selected discriminant variables using machine learning. 1 million Monte Carlo simulated signal events and background events are first generated respectively as the training data.

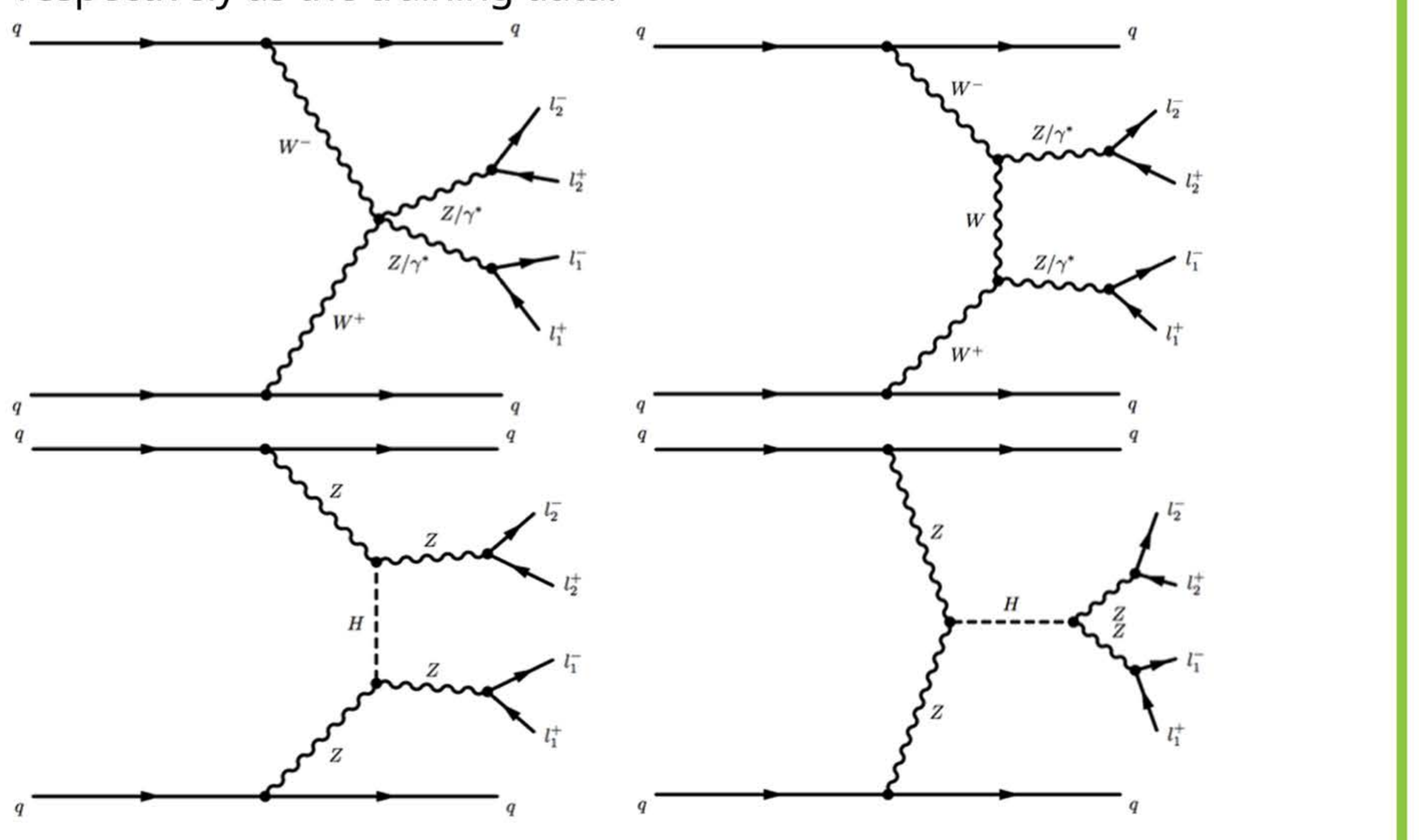


Figure 1. Signal: VBS
pp -> Z Z jj -> $l^+ l^- l^+ l^-$ jj
Z pair + 2 jets production in vector boson fusion with decay into charged leptons

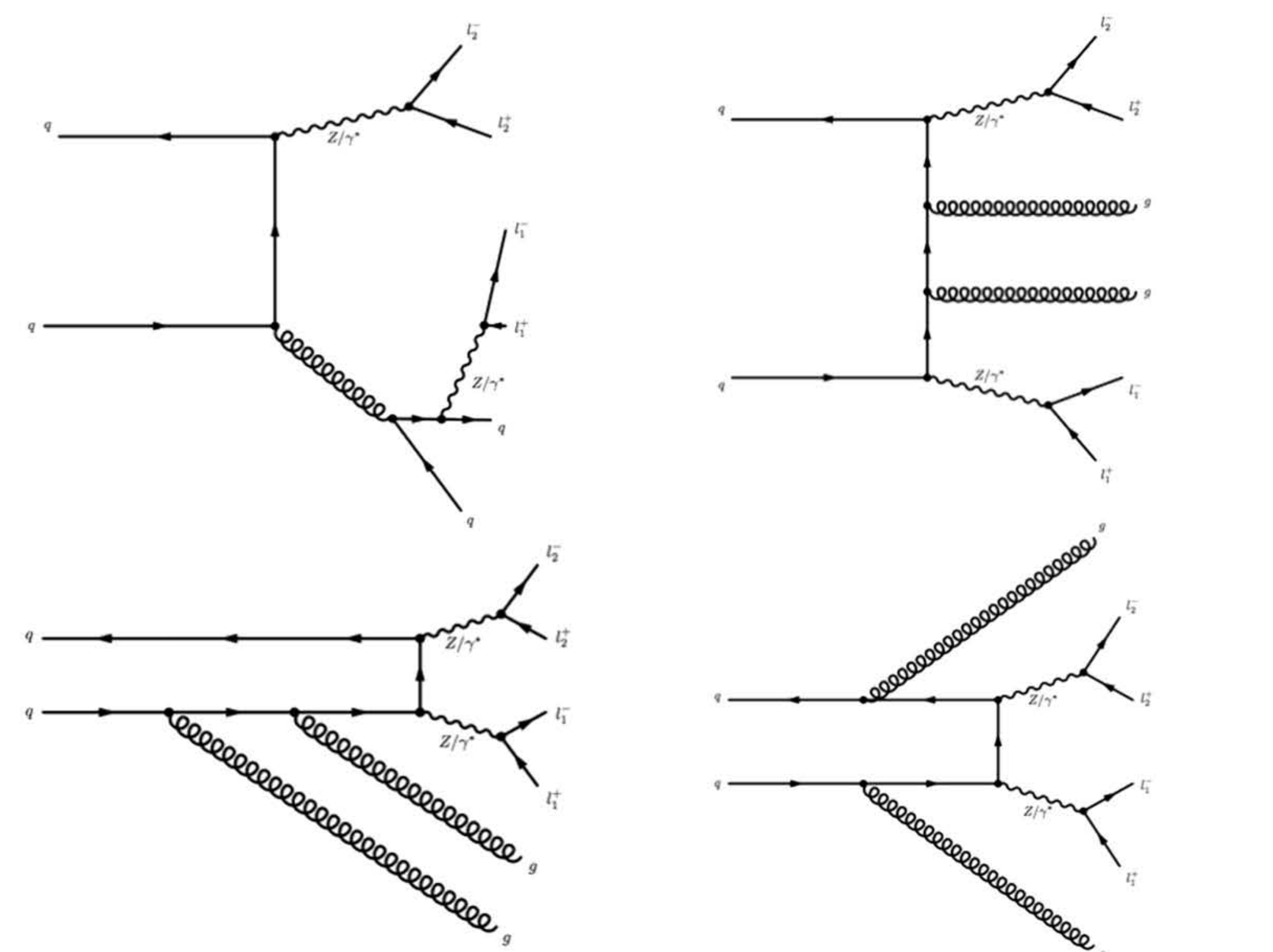
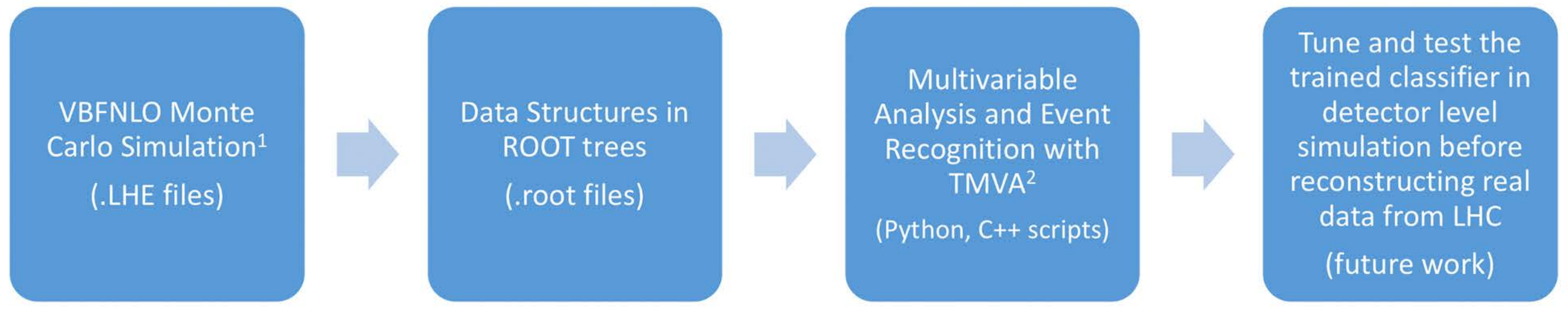


Figure 2. Major Background: QCD
pp -> Z Z jj -> $l^+ l^- l^+ l^-$ jj
QCD induced ZZ + 2 jets production with fully leptonic decay

Analysis Workflow



VBFNLO Monte Carlo Simulation[1] (.LHE files)

Data Structures in ROOT trees (.root files)

Multivariable Analysis and Event Recognition with TMVA[2] (Python, C++ scripts)

Tune and test the trained classifier in detector level simulation before reconstructing real data from LHC (future work)

[1] **VBFNLO** is a fully flexible parton level Monte Carlo program for the simulation of vector boson fusion, double and triple vector boson production in hadronic collisions at next to leading order in the strong coupling constant.

[2] The Toolkit for Multivariate Data Analysis with ROOT (**TMVA**) is a standalone project that provides a ROOT-integrated machine learning environment for the processing and parallel evaluation of sophisticated multivariate classification techniques.

Event Pre-selection Criteria

1. Transverse momentum of lepton 1 (leading lepton) > 25 GeV
2. Transverse momentum of leptons 2, 3, 4 > 7 GeV
3. Pseudorapidity of leptons 1, 2, 3, 4 < 2.5
4. Mass of Z_1 and Z_2 in the range of (66, 116) GeV

Selection Criteria	VBS Number of Events	VBS Selection efficiency	QCD Number of Events	QCD Selection efficiency
None	890100	1	1014200	1
Pre-selection	209752	0.236	115629	0.114
Pre-selection and m(jj) > 500 GeV and |Δη(jj)| > 3	125171	0.141	9509	0.009

Training a Boosted Decision Tree (BDT) Signal-background Classifier

Select discriminant variables
1. mass of dijets (m_{jj})
2. change of Pseudorapidity (Δη)
3. mass of dibosons (m_{ZZ})
4. transverse momentum of 2 bosons (pt_{Z1}, pt_{Z2})
5. transverse momentum of 4 leptons (pt_{l1}, pt_{l2}, pt_{l3}, pt_{l4})

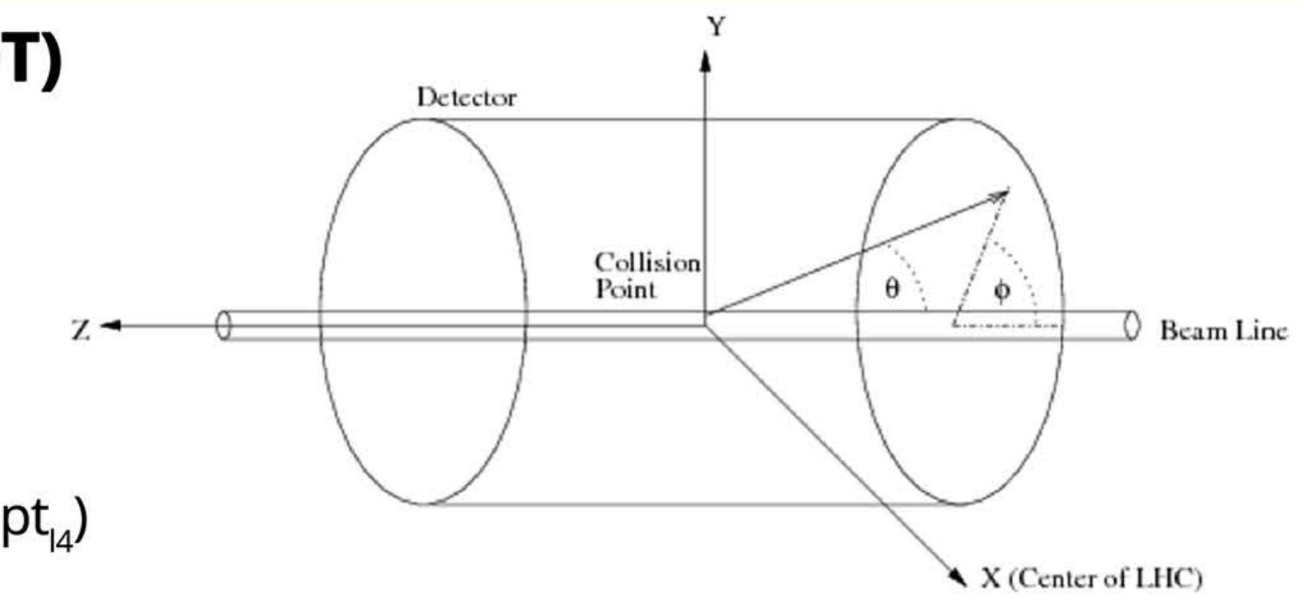


Figure 3. Collision Schematics
Where θ is the angle between the particle three-momentum p and the positive direction of the beam axis.

Pseudorapidity: η = -ln(tan(θ/2))

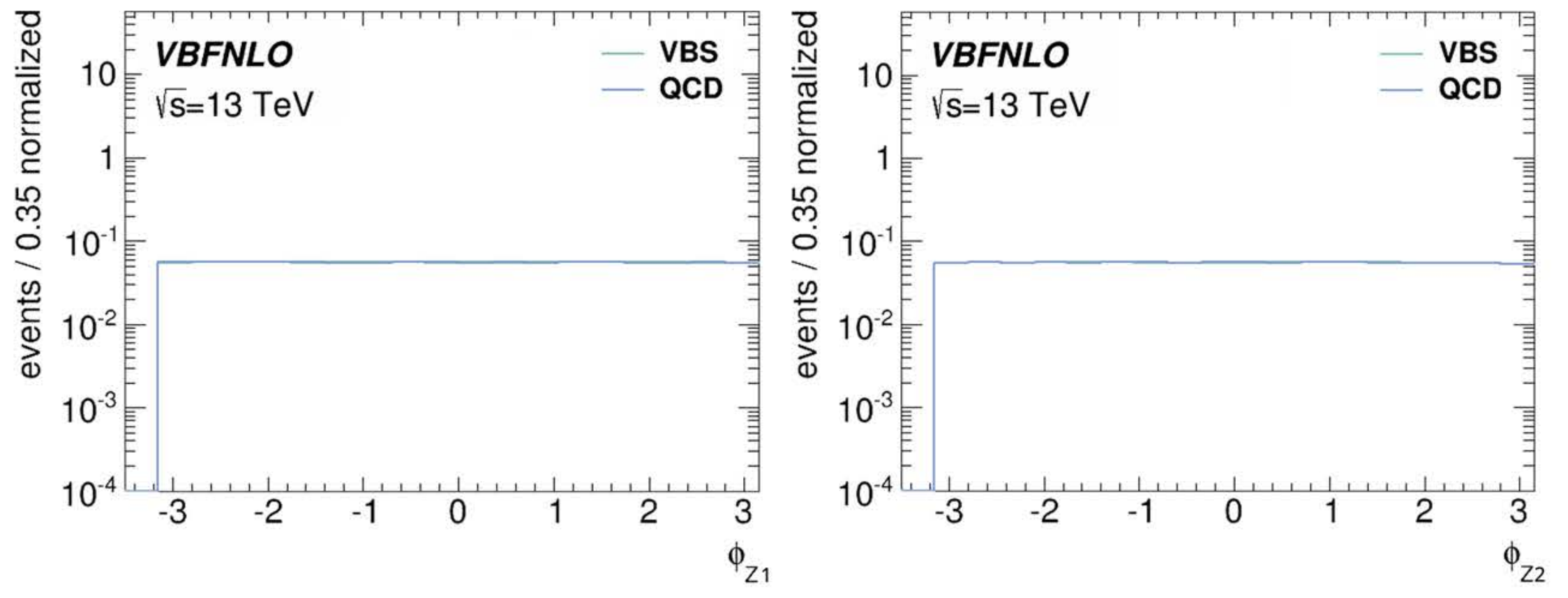


Figure 4-5. Plots of Unselected Variables whose distribution patterns are not noticeably different at all

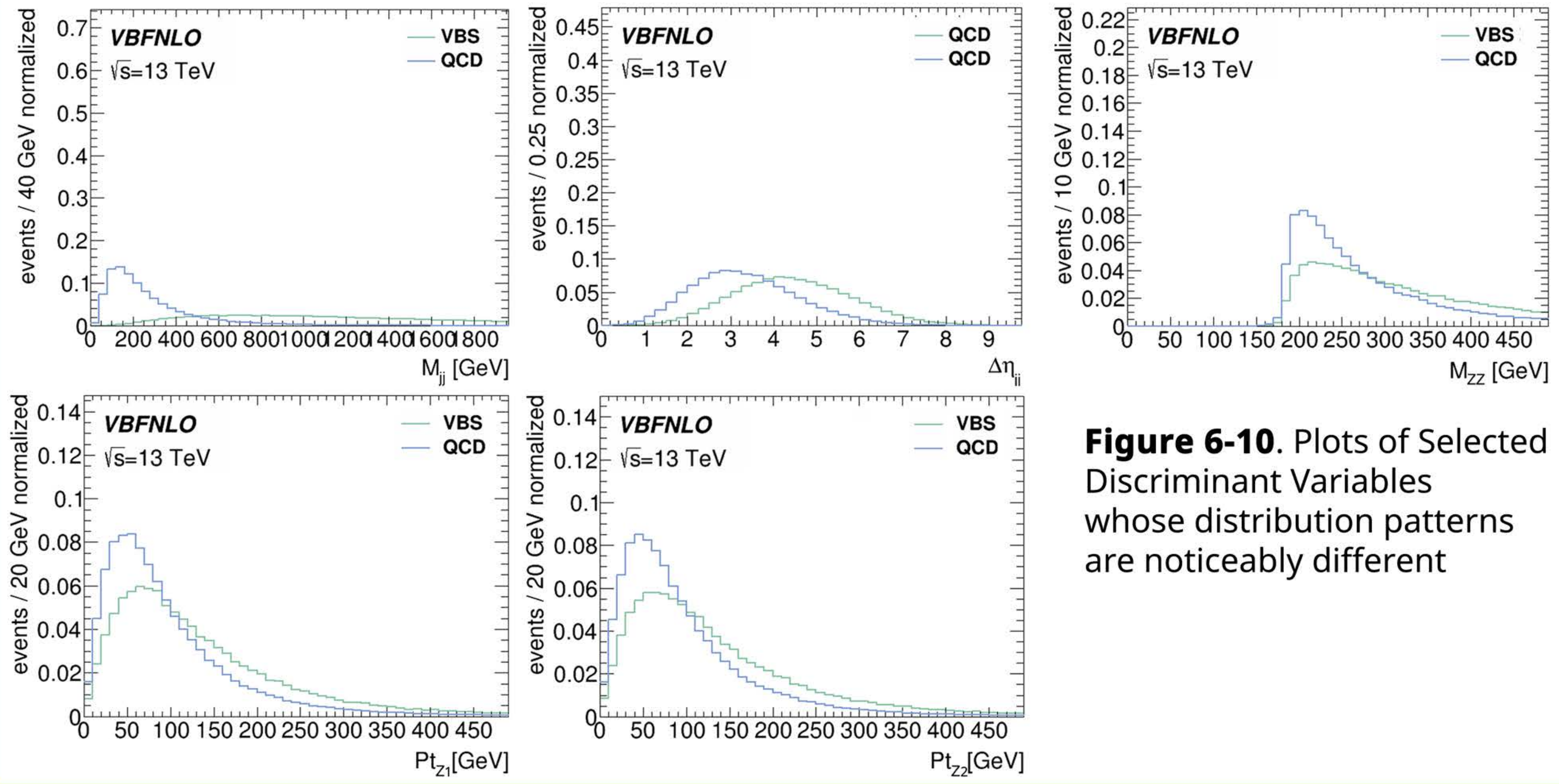


Figure 6-10. Plots of Selected Discriminant Variables whose distribution patterns are noticeably different

Boosted Decision Trees (BDT) Configuration

- Transformations: Decorrelation; Principle Component Analysis (PCA); Gaussian, Decorrelation transformation:
- Training set: 60% of simulated signal data and equivalent amount of background data
- Testing set: the rest of simulated signal and background data
- Number of trees: 500/1000
- Boosttypes:
  - § Adaptive
  - § Adaptive + Decorrelation
  - § Adaptive + Fisher discriminant
  - § Gradient
  - § Bagging
- For Adaptive Boost: AdaBoostBeta=0.5, useBaggedBoost, BaggedSampleFraction=0.5
- Separationtype: Gini Index: p·(1-p) where p is purity
- Numberofcuts = 20/Optimal cuts
- Maximumdepth = 5
- MinNodeSize = 2.5

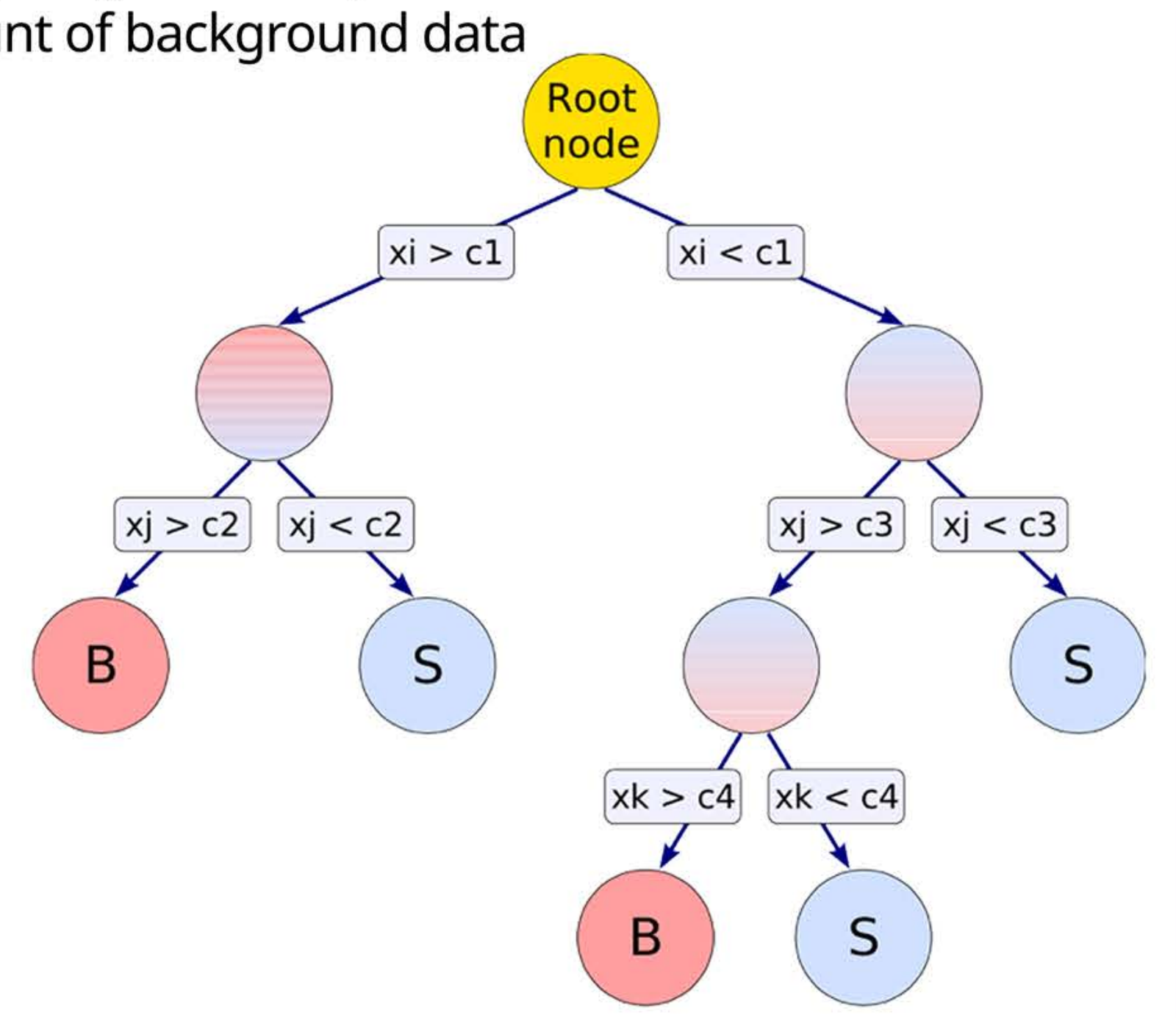


Figure 11. Boosted Decision Tree Schematics
Starting from the root node, a sequence of binary splits using the discriminating variables xi is applied to the data. Each split uses the variable that at this node gives the best separation between signal and background when being cut on. The same variable may thus be used at several nodes, while others might not be used at all. The leaf nodes at the bottom end of the tree are labeled "S" for signal and "B" for background depending on the majority of events that end up in the respective nodes.

Boosting Mechanism Example

AdaBoost (Adaptive Boost):

```
input  : an unlabelled event $X_i$
output: labelled event $X_i$ (either signal or background)
Start same weight for all data points: $\alpha_i = \frac{1}{N}$;
for $t = 1, ..., T$ do                                  // loop through all features of each data point
    $\hat{y}_i = sign(f_t(X_i))$      // predict by feature classifier $f_t(x)$ with data point weight $\alpha_i$;
    $weighted\_error = \frac{\sum_{i=1}^{N} \alpha_i \mathbb{1}(\hat{y}_i \neq y)}{\sum_{i=1}^{N} \alpha_i}$;
    $\hat{w}_t = \frac{1}{2} \ln(\frac{1-weighted\_error(f_t)}{weighted\_error(f_t)})$          // compute coefficient $\hat{w}_t$;
    if $f_t(X_i) = y_i$ then $\alpha_i = \alpha_i e^{-\hat{w}_t}$ ;         // decrease the weight if predicts correctly
    if $f_t(X_i) \neq y_i$ then $\alpha_i = \alpha_i e^{\hat{w}_t}$ ;          // increase the weight if predicts wrongly
    $\alpha_i = \frac{\alpha_i}{\sum_{j=1}^{N} \alpha_j}$                              // normalize weight $\alpha_i$;
end
Final model is predicted by an ensemble of feature classifiers $\hat{y}_i = sign(\sum_{t=1}^{T} \hat{w}_t f_t(X_i))$
```

Classification Result and Conclusion

This research project is heavily driven by machine learning algorithms for event classification. In contrast to following a carefully primed model to analytically search for target events and interactions in LHC, an algorithmic appraoch is taken to quickly identify the patterns of target events and interactions by "learning" from the truth events generated from the theory-level simulation.

In the previous studies, the classification in comparison was made based on direct threshold cuts on certain discriminant variables, such as mass of jets, change of pseudorapidity, transverse momentum of bosons, etc. as Figure 12 also illustrated.

The result of one of the optimal classifications confirms the usefulness of this machine learning approach as the the parameters of the selected discriminant variables (features) in the classifier would be automatically tuned to values close to the thresholds of direct variable cut during the training process. With more discrimant variables added in the classifier than the direct variable cut has, the classification accuracy indeed improves.

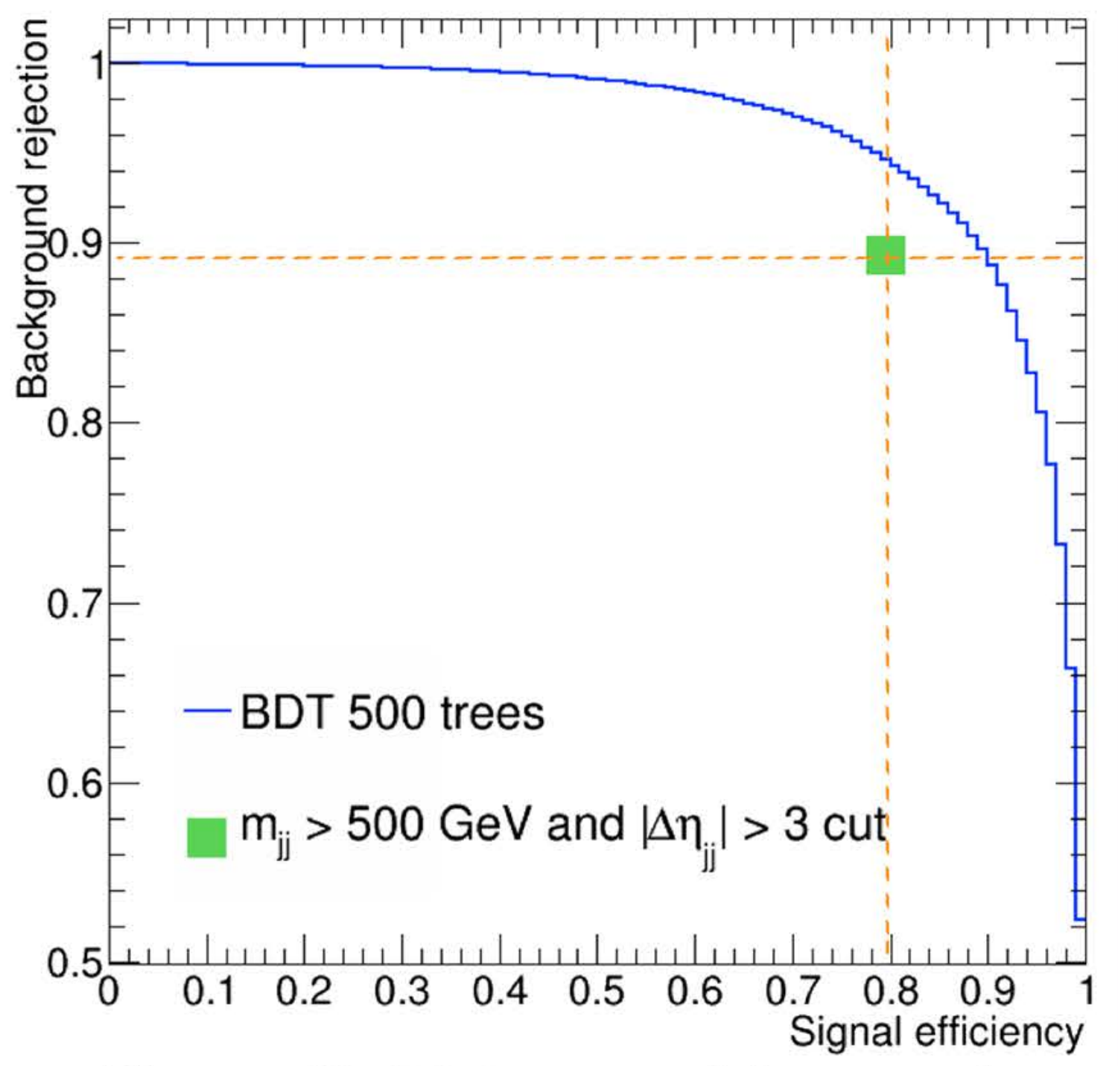


Figure 12. ROC curve of the one of the optimal classifications

- BDT performs better than the direct variable cut
- At signal efficiency of around 0.80, BDT is about 7% better than the direct variable cut
- At background rejection of around 0.89,t BDT is about 12.5% better than the direct cut
- For 40 fb^{-1} (cross section), BDT score cut > 0.08: N(S) = 5.2, N(B) = 1.1
- Considering 50% reconstruction efficiency, N(S) = 2.6, N(B) = 0.5
- Potential to observe electroweak production of ZZ + 2 jets at 13 TeV (combined witth other channels and using more luminosity in 2017)

Yield N (after pre-selection) = luminosity (40) x cross-section (S/B) x eff(pre-selection)
Yield N (after final selection) = luminosity (40) x cross-section (S/B) x eff(pre-selection) x eff(S/B selection)

Future Work

Once the classification result is proved to be reliable, this well-trained classifier can be tested on detector-level simulation or extended and adapted for other data and processes in the future. After detector-level simulation, the real data collected from the LHC could then be reconstructed to measure the final state particles in the collisions and hopefully help verify whether certain reaction processes really happened.

Acknowledgement

I would like to thank Prof. Bin Zhou* and her research group for supervising and guiding my research, as well as Prof. Jean Krisch* and Dr. Steven Goldfarb*† for their advice and support. I would also like to thank the Richard Lounsbery Foundation and the University of Michigan* for fully funding my research at the European Organization for Nuclear Research† (CERN) in Geneva, Switzerland. It was an amazing journey full of excitement and inspiration.

Major References

The ATLAS Collaboration, Studies of Vector Boson Scattering And Triboson Production (2013)
C. Bittrich, Study of Polarization Fractions in the Scattering of Massive Gauge Bosons (2015)