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

Albert Guo, 12/8/2016

Supervisor: Prof. Bin Zhou[1]

Dr. Yusheng Wu[1,2]

[1] University of Michigan
[2] Institute of Physics, Academia Sinica

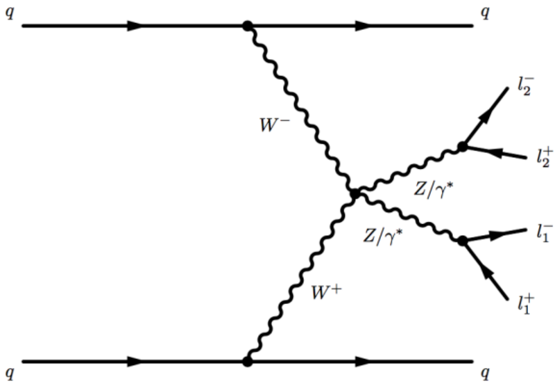
Physics background and purpose of measuring the VBS ZZ + 2 jets Production

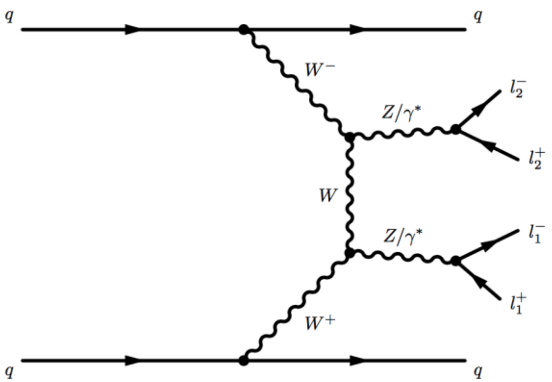
- Vector-boson-scattering (VBS) two-boson production gives a unique opportunity to examine the nature of electroweak symmetry breaking (EWSB) in the Standard Model (SM).
  - The scattering of gauges bosons violates unitarity at the TeV scale if there is no Higgs boson.
  - The unitarity can be restored 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
  - whether the EWSB mechanism is as predicted in the SM

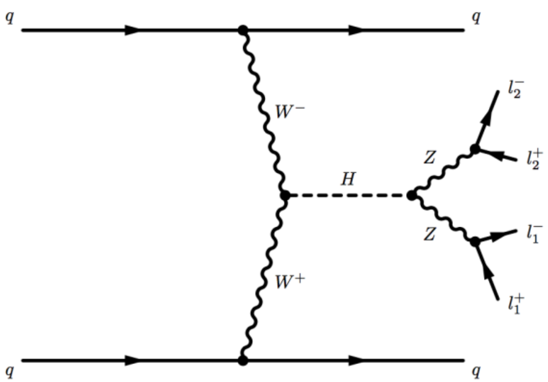
Signal: VBS

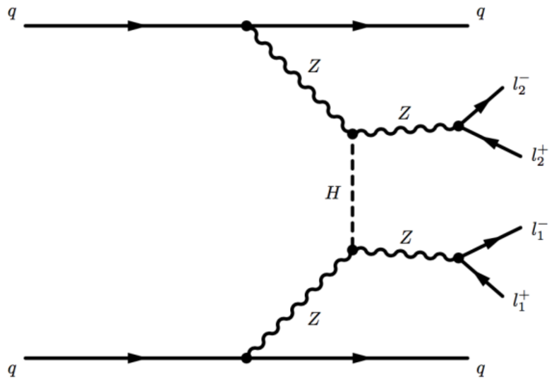
pp -> Z Z jj -> l+ l- l+ l- jj

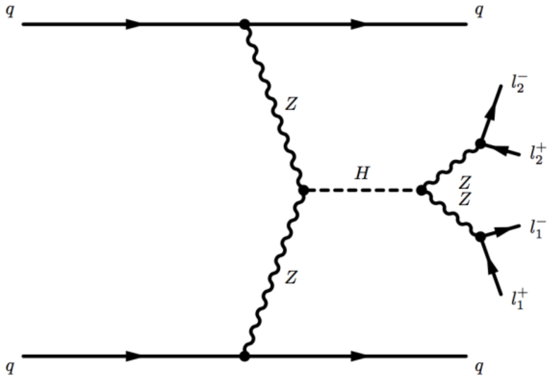
Z pair + 2 jets production in vector boson fusion with decay into charged leptons







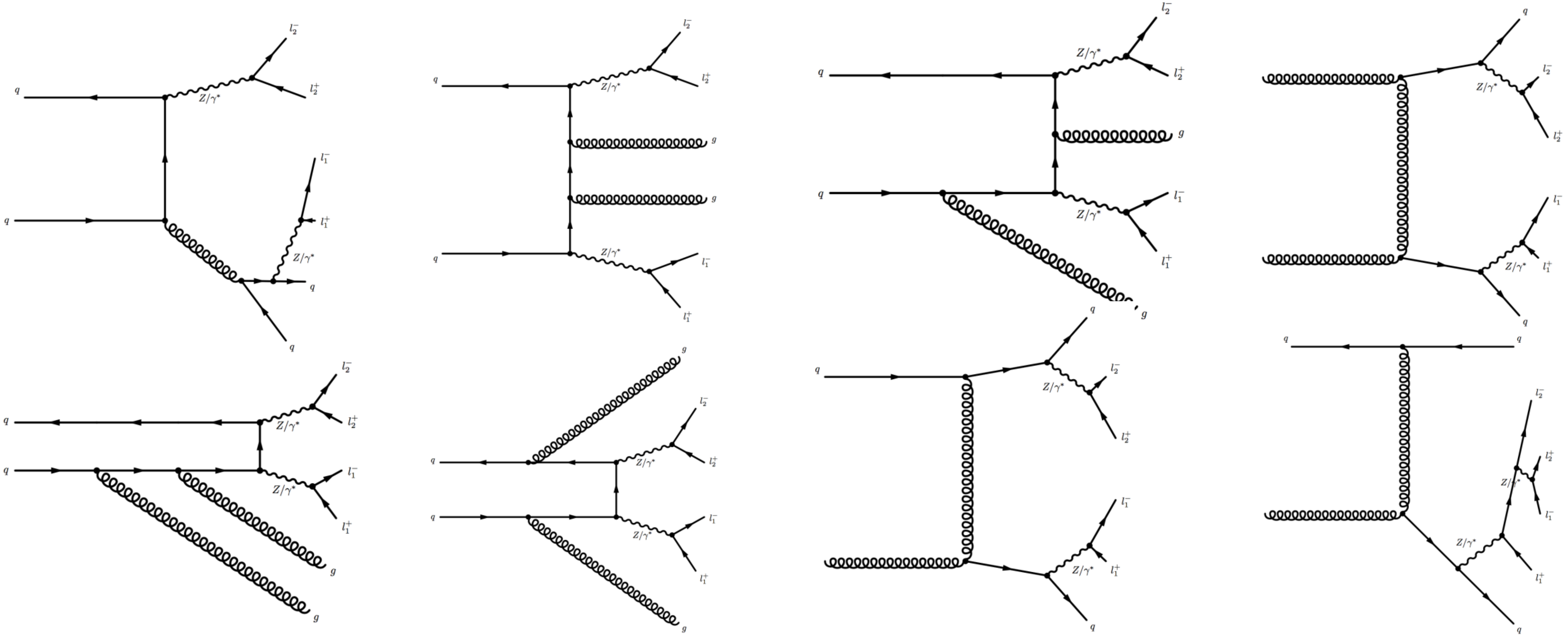




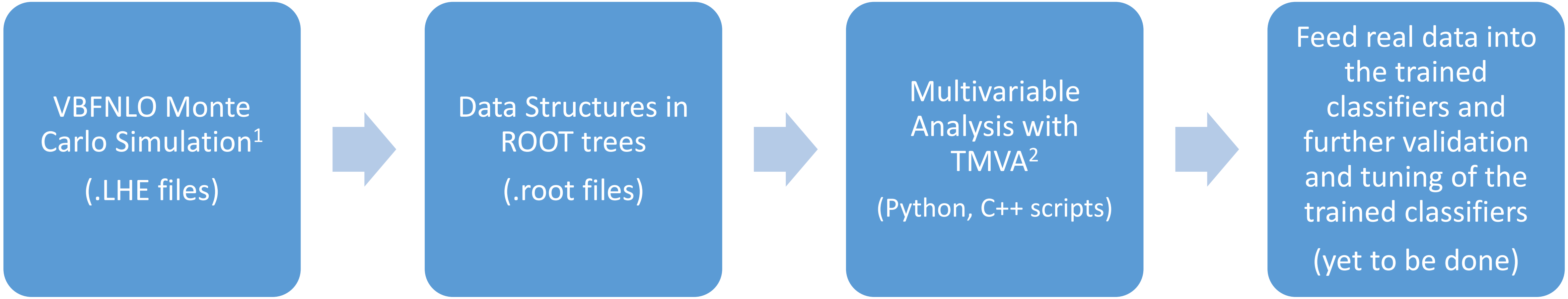
Major Background: QCD

pp -> Z Z jj -> l+ l- l+ l- jj

QCD induced Z Z + 2 jets production with fully leptonic decay.



Research Role and Workflow



[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.

VBFNLO of 1M events generation for VBS and QCD with **e+ e- e+ e-** **jj** final states

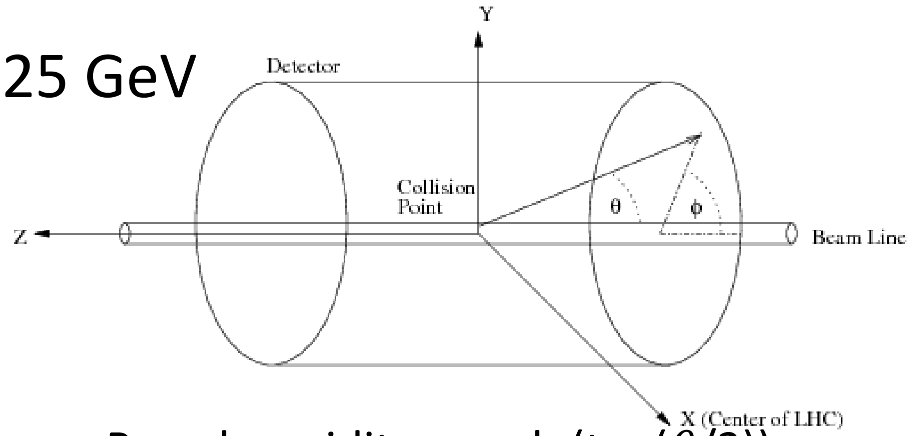
	VBS jet cuts	QCD jet cuts
min jet-jet R separation	0	0
max parton pseudorapidity (η)	5.0	5.0
exponent of generalised k_T algorithm	-1.0	-1.0
min jet transerverse momentum (pT)	15.0	15.0
max jet rapidity	9	9
	VBS lepton cuts	**QCD lepton cuts**
max lepton rapidity	2.8	2.8
min lepton transerverse mome (pT)	4.0	4.0
min. m_l+l- for any comb. of opposite charged leptons	4	4
max. m_l+l- for any comb. of opposite charged leptons	13000	13000
min lepton-lepton R separation	0.01	0.01
max lepton-lepton R separation	50.0	50.0

Rapidity: $Y = \frac{1}{2} \cdot \log((E + P_z) / (E - P_z))$; R: solid angular distance between two tracks

Pre-selection Criteria

Where θ is the angle between the particle three-momentum **p** and the positive direction of the beam axis.

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



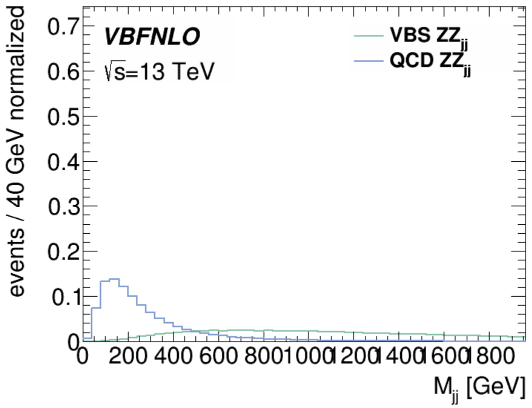
Pseudorapidity: $\eta = -\ln(\tan(\theta/2))$

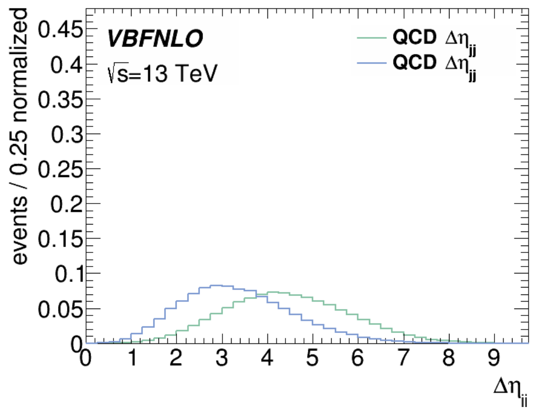
Selection Criteria	VBS Number of Events	VBS Selection efficiency	QCD Number of Events	QCD Selection efficiency
None	890100	1	1014200	1
Pre-selection	209752	0.236	115629	0.114
Pre-selection and m(jj) > 500 GeV and $\lvert\Delta\eta(jj)\rvert > 3$	125171	0.141	9509	0.009

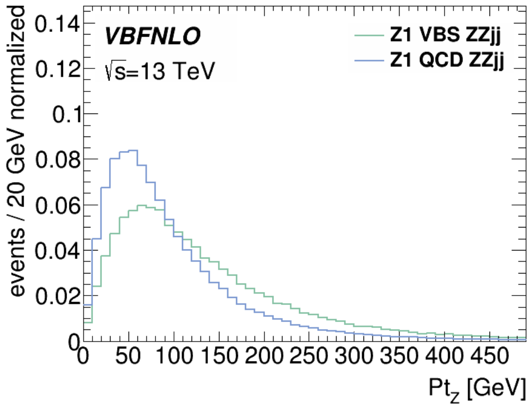
Training a Boosted Decision Tree (BDT) as the signal-background classifier

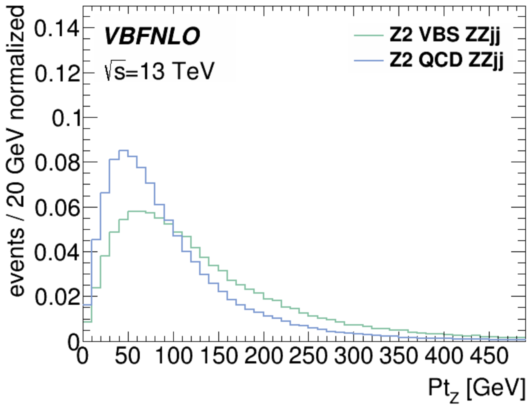
Select discriminant variables for classification

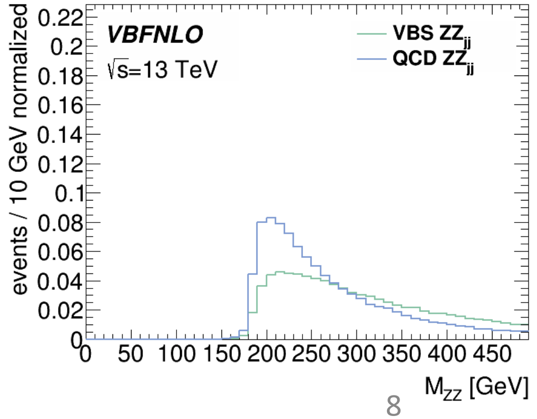
1. mass of 2 jets (m_{jj})
2. change of Pseudorapidity ($\Delta\eta$)
3. mass of 2 bosons (m_{ZZ})
4. transverse momentum of 2 bosons (pt_{Z1}, pt_{Z2})
5. transverse momentum of 4 leptons (pt_{l1}, pt_{l2}, pt_{l3}, pt_{l4})





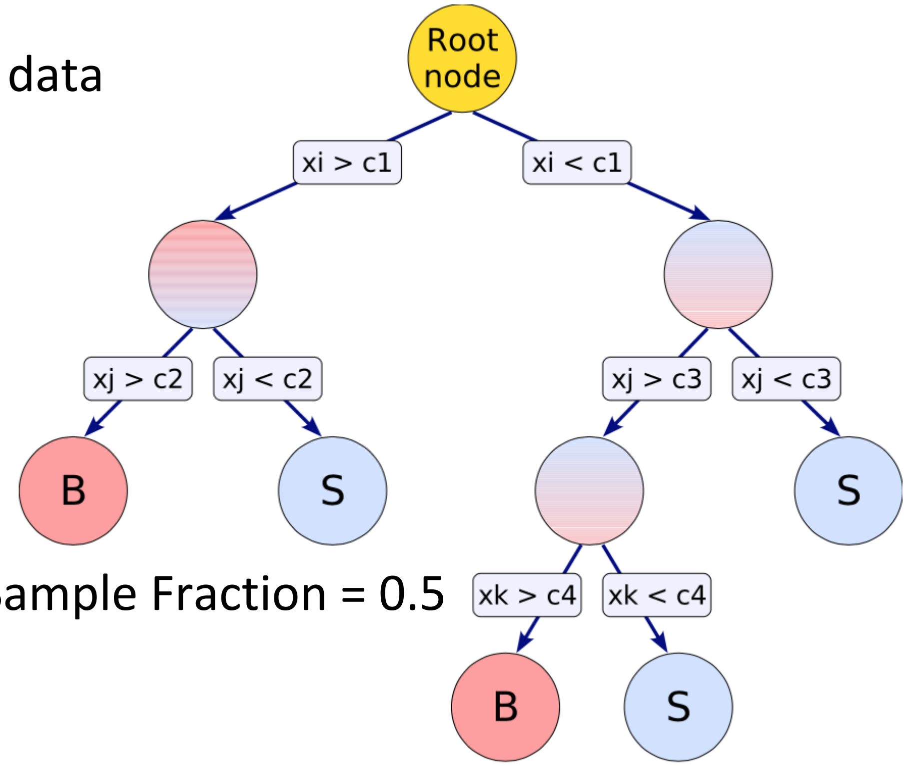






Boosting Decision Trees (BDT) Configuration

- Transformations: **D**ecorrelation; **P**rinciple Component Analysis (PCA); **G**aussian, **D**ecorrelation transformation:
- Training set: 60% of signal data and the same amount of background data
- Testing set: the rest of signal and background data
- Number of trees: 500/1000
- Boost type:
    - Adaptive
    - Adaptive + Decorrelation
    - Adaptive + Fisher discriminant
    - Gradient
    - Bagging
- For Adaptive Boost: AdaBoostBeta = 0.5, use Bagged Boost, Bagged Sample Fraction = 0.5
- Separation type: Gini Index: $p \cdot (1 - p)$ where p is purity
- Number of cuts = 20
- Maximum depth = 5
- MinNodeSize = 2.5



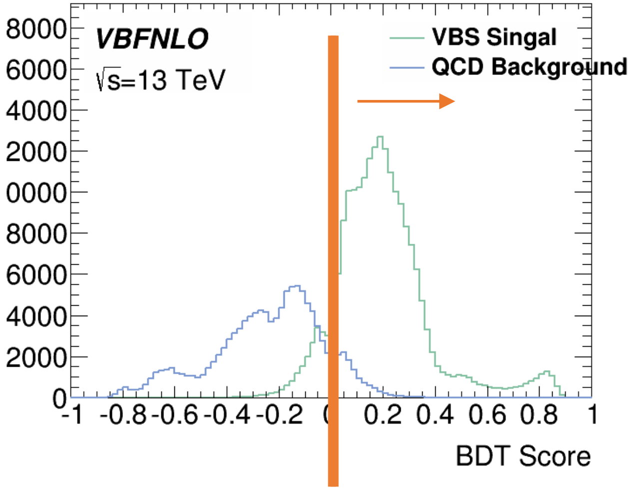
Boosting Decision Trees (BDT) VS. Variable Cuts at m(jj) > 500 GeV and |Δη(jj)| > 3

CLASSIFICATION	Signal region cut	N(S)	N(B)
by cuts on variables	m(jj) > 500 GeV and |Δη(jj)| > 3 cut	1.352+/- 0.0033	0.901+/- 0.01217
by BDT 500 trees	0	1.539 +/- 0.0035	0.938 +/- 0.00827
by BDT 500 trees	0.040	1.460 +/- 0.0034	0.648 +/- 0.00688
by BDT 500 trees	0.080	1.309 +/- 0.0033	0.354 +/- 0.00508
by BDT 500 trees	0.120	1.145 +/- 0.0030	0.190 +/- 0.00372
by BDT 500 trees	0.160	0.964 +/- 0.0028	0.096 +/- 0.00264
by BDT 500 trees	0.200	0.763 +/- 0.0025	0.042 +/- 0.00176
by BDT 500 trees	0.240	0.576 +/- 0.0022	0.018 +/- 0.00114
by BDT 500 trees	0.280	0.419 +/- 0.0018	0.0073 +/- 0.00073
by BDT 500 trees	0.320	0.292 +/- 0.0015	0.0034 +/- 0.00050
by BDT 500 trees	0.360	0.204 +/- 0.0013	0.0015 +/- 0.00033
by BDT 500 trees	0.400	0.159 +/- 0.0011	0.00088 +/- 0.00025
by BDT 500 trees	0.440	0.103 +/- 0.0010	0.00073 +/- 0.00023
by BDT 500 trees	0.480	0.120 +/- 0.0009	0.00051 +/- 0.00019
by BDT 500 trees	0.520	0.103 +/- 0.0008	0.00029 +/- 0.00015

Increment by 0.04

	N(S)	N(B)
NO SELECTION	7.2	74.0
AFTER PRE-SELECTION	1.70	8.43

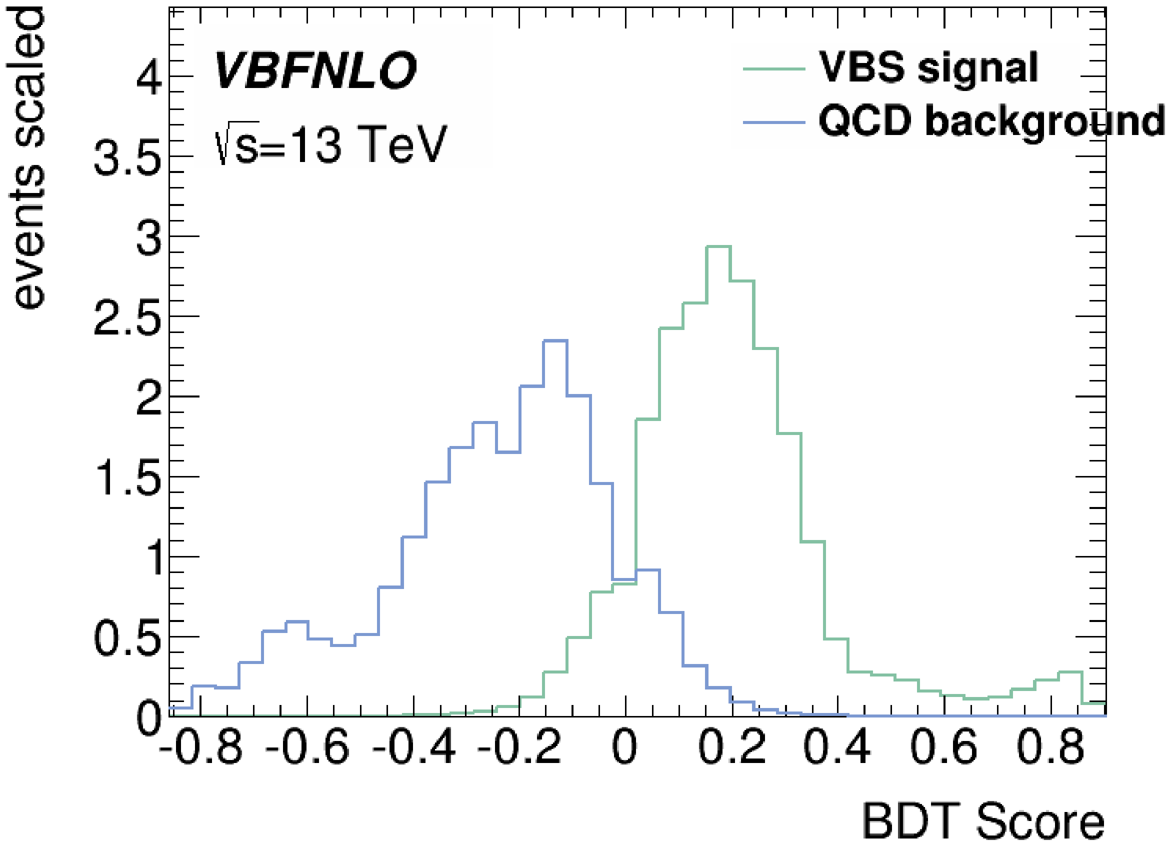
All yields are normalized to 40fb-1*, which is close to data luminosity at 13 TeV



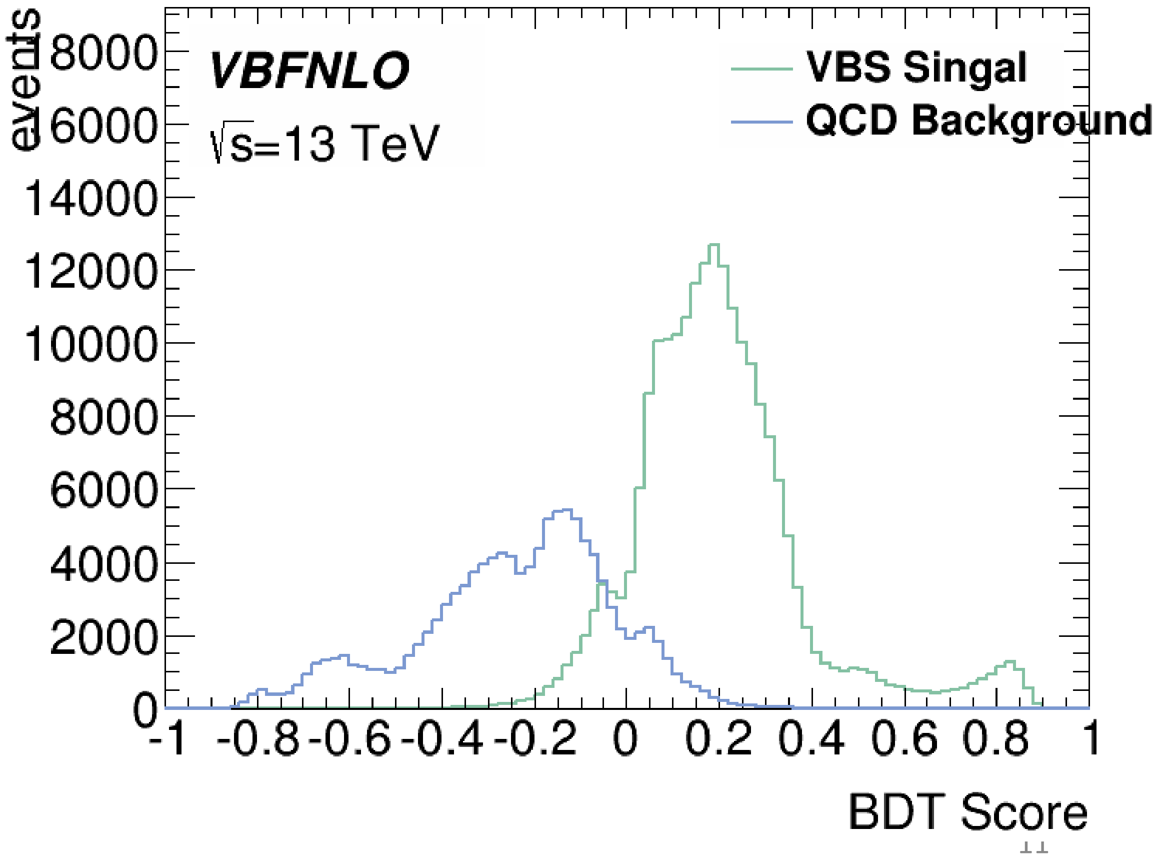
*10^{-28} m^2 (100 fm^2)

Events Distributions

distribution obtained directly from TMVA

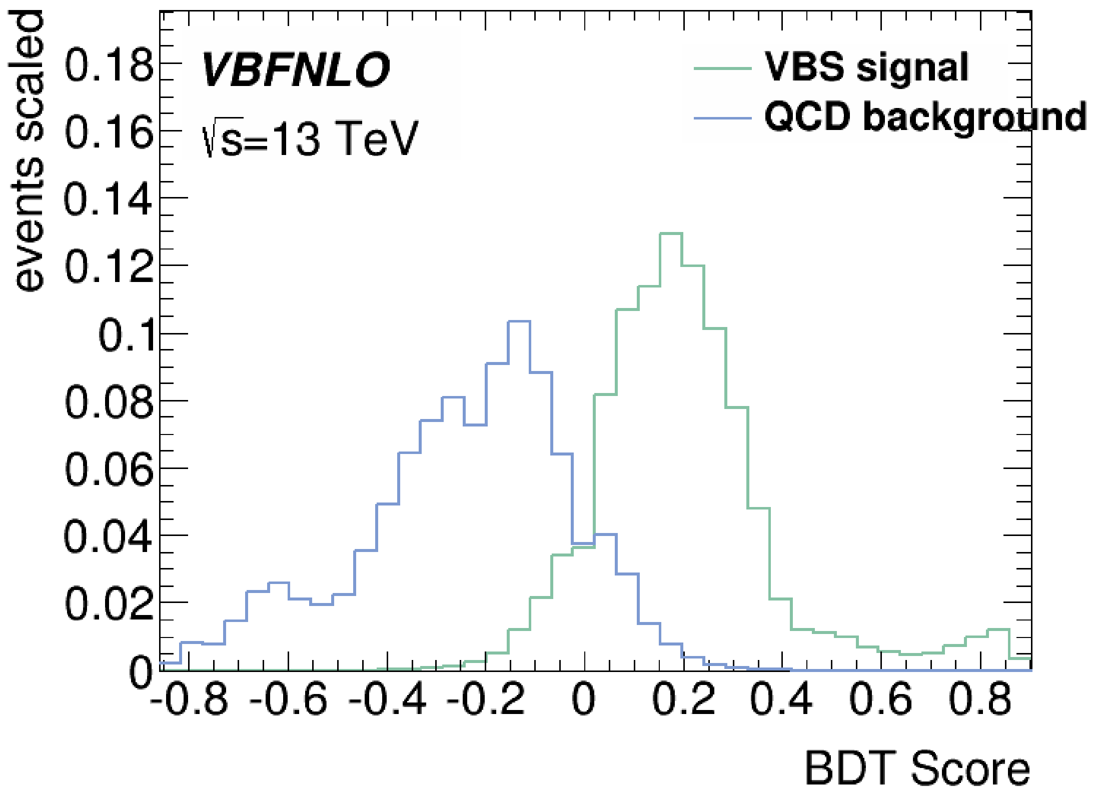


distribution by directly counting events from ROOT trees

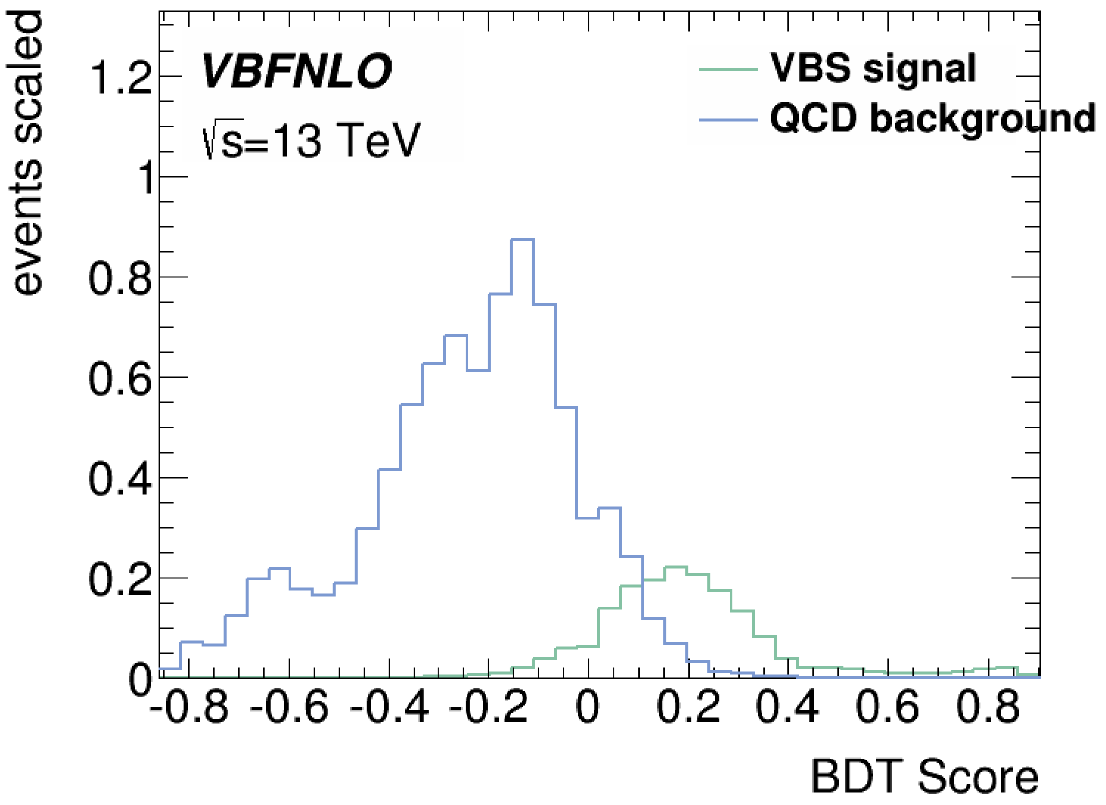


BDT Distributions from TMVA

normalized to unity area

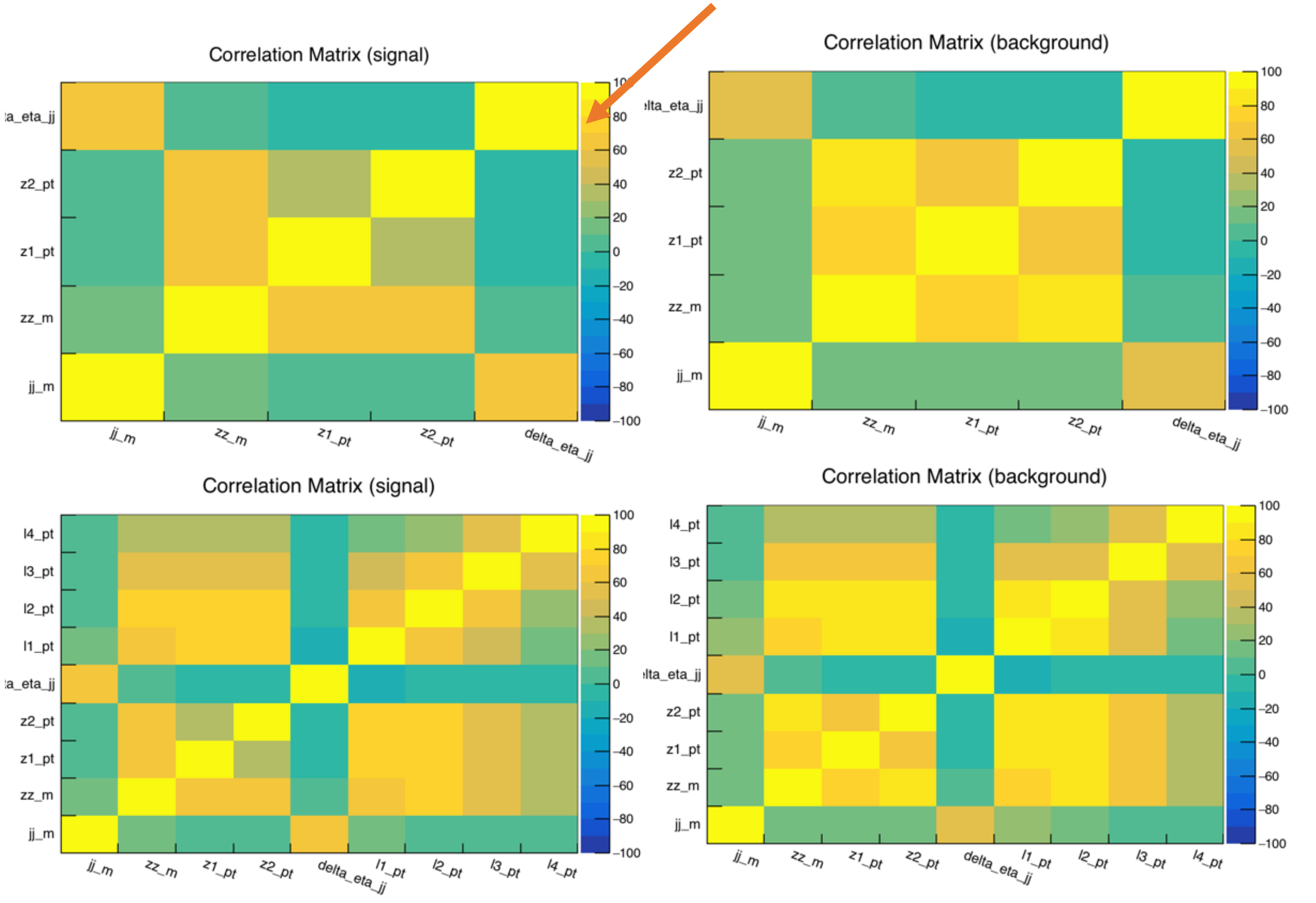


normalized to 40fb-1

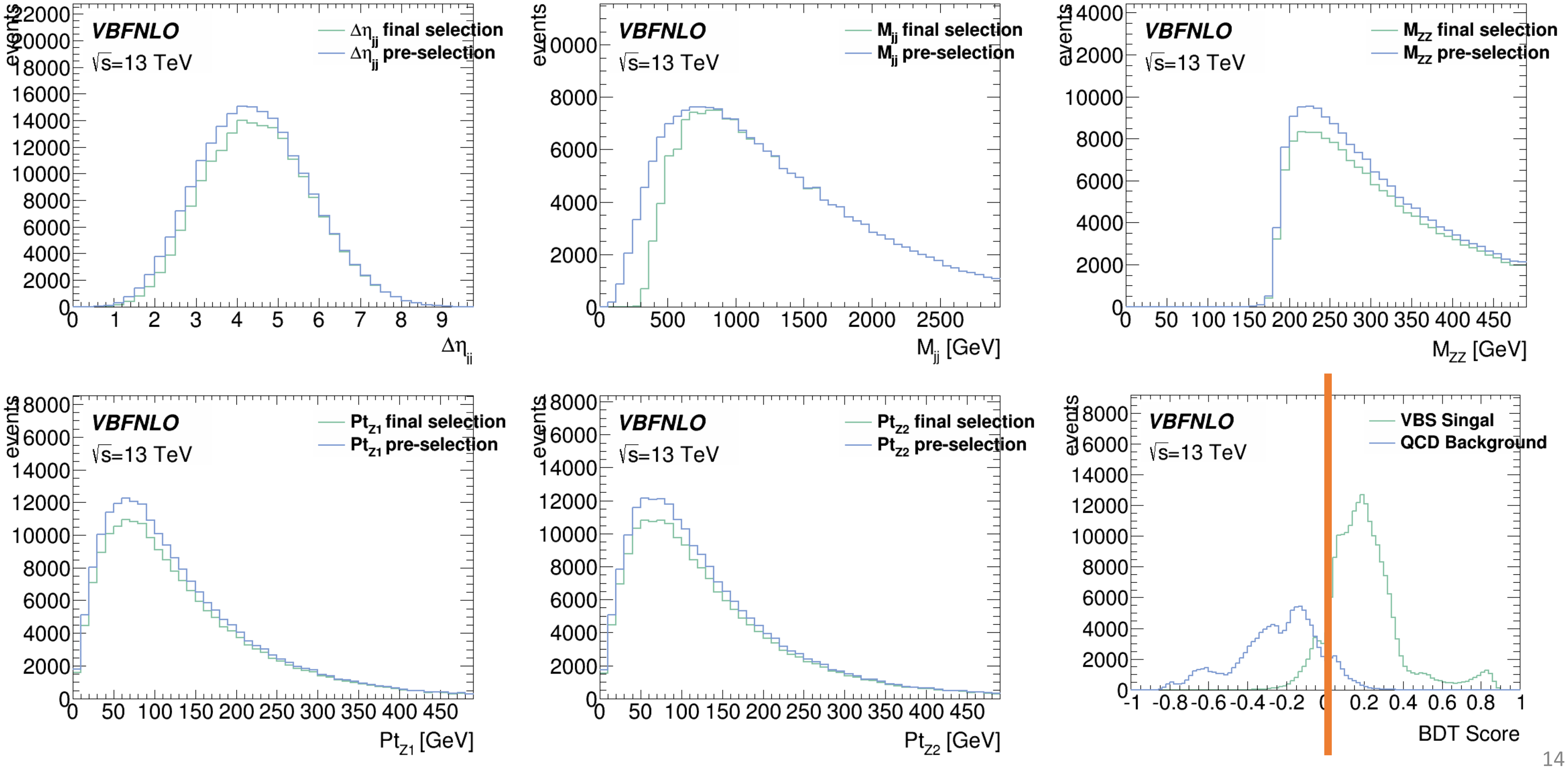


BDT Ranking of Variables and Correlation Matrices

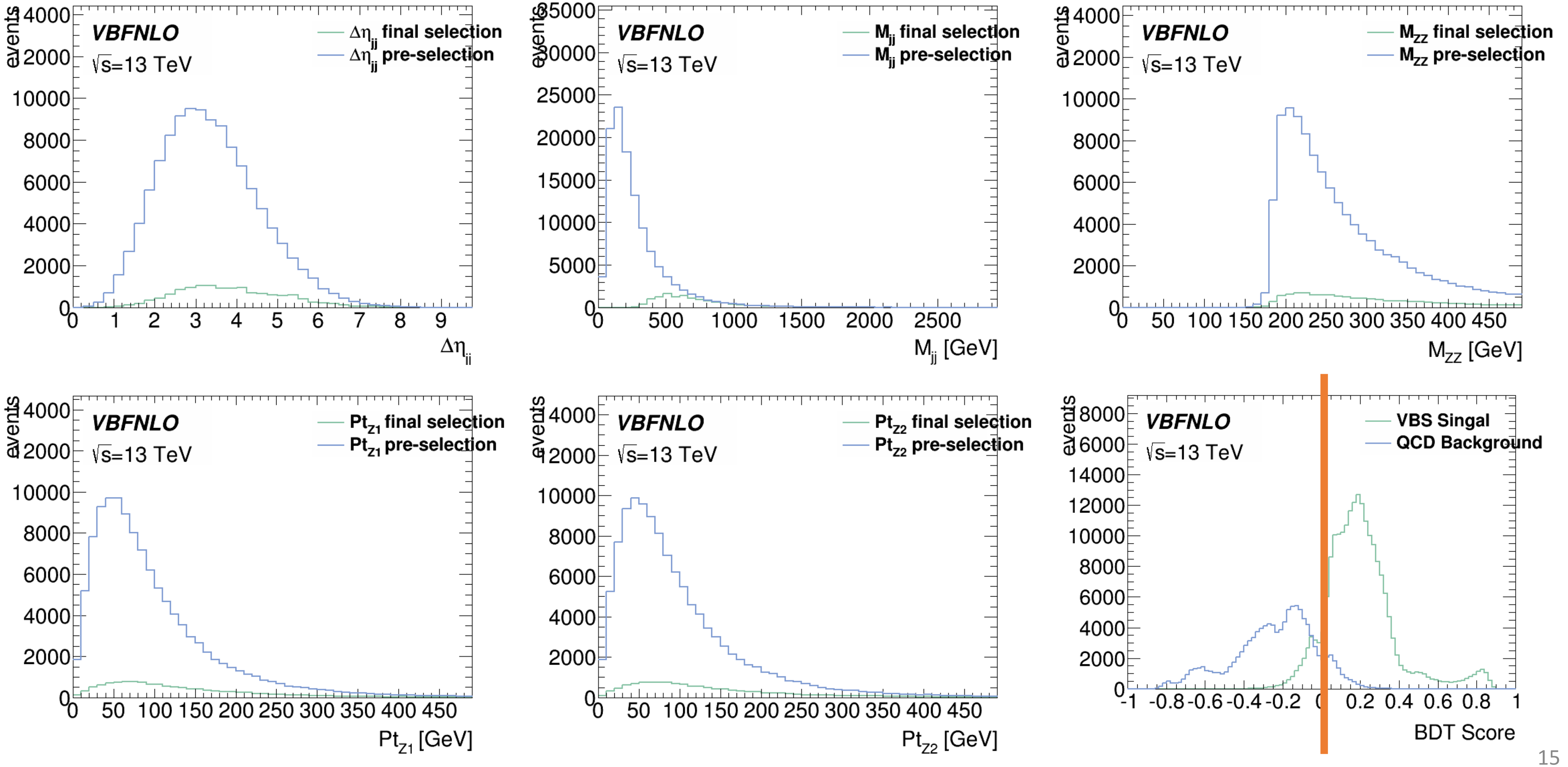
```
Ranking result (top variable is best ranked)
--------------------------------------------------
Rank : Variable      : Variable Importance
--------------------------------------------------
   1 : delta_eta_jj  : 3.689e-01
   2 : jj_m          : 1.875e-01
   3 : zz_m          : 7.758e-02
   4 : z1_pt         : 6.298e-02
   5 : l2_pt         : 6.292e-02
   6 : z2_pt         : 6.136e-02
   7 : l1_pt         : 6.100e-02
   8 : l4_pt         : 6.053e-02
   9 : l3_pt         : 5.731e-02
```



Variable Selection for VBS for BDT score cut at 0.08 in linear scale



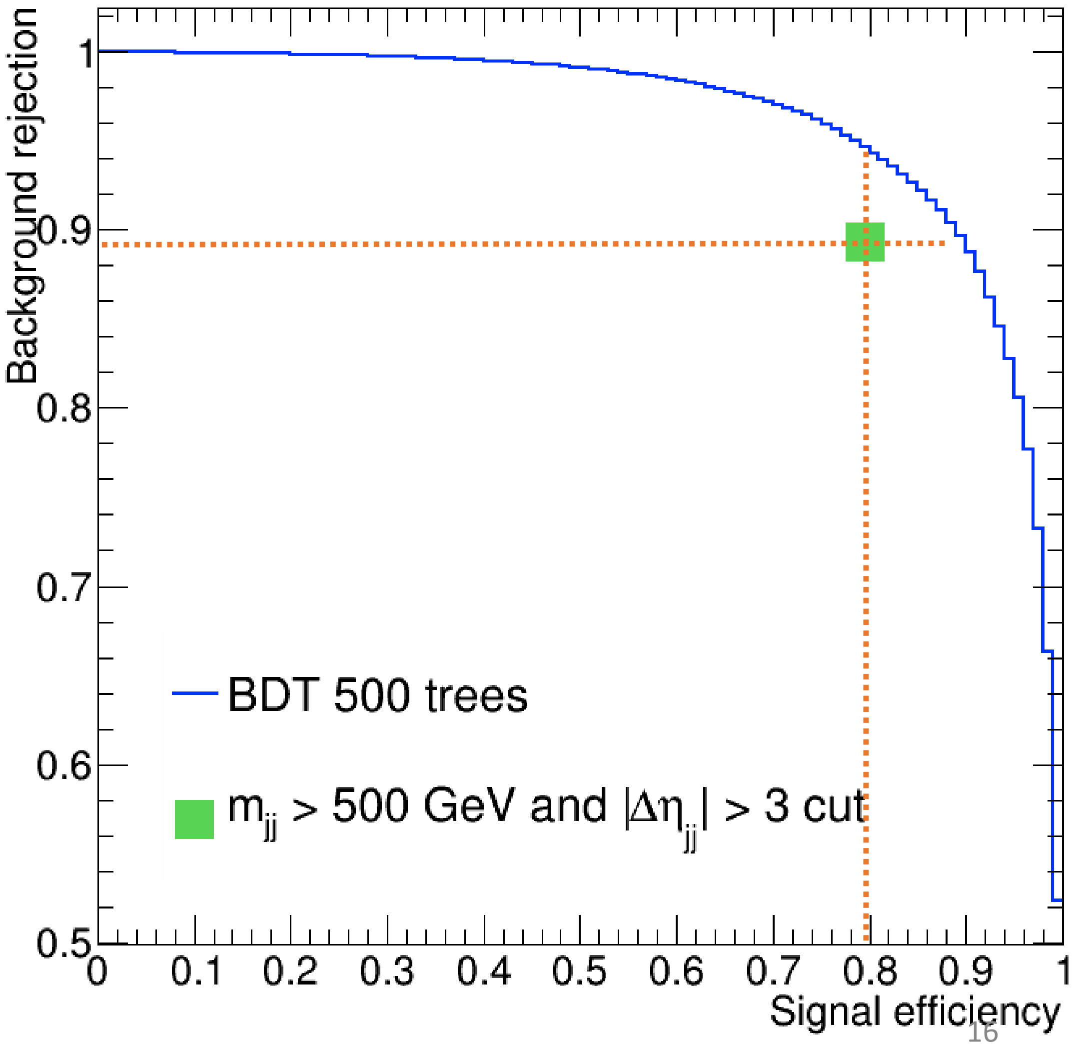
Variable Selection for QCD for BDT score cut at 0.08 in linear scale



Conclusion

- BDT is better than the direct variable cut of m_{jj} > 500 GeV and $|\Delta\eta_{jj}|$ > 3:
  - At signal efficiency of around 0.80, BDT is about 7% better than the direct cut
  - At backgorund rejection of around 0.89, BDT is about 12.5% better than the direct cut
- The final yields obtained in slide 10 need to multiply 4 to account for all decay channels: eeee, eemumu, and mumumumu
  - For 40 fb-1, 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 with other channels and using more luminosity in 2017)

ROC curve of the one of the optimal configurations







BÄRIN URSINA
SCUSI...
NOI DORMIAMO





Questions?

Thank you!

Backup

Rapidity: $Y = \frac{1}{2} \cdot \log((E + Pz) / (E - Pz))$

Pseudorapidity: $\eta = -\ln(\tan(\phi/2))$

R separation = solid angular distance between two tracks

ΔR: $\Delta R(l_1, l_2) = \sqrt{ (\eta(l_1) - \eta(l_2))^2 + (\phi(l_1) - \phi(l_2))^2}$

Yield (no selection) = luminosity (40) x cross-section (S/B)

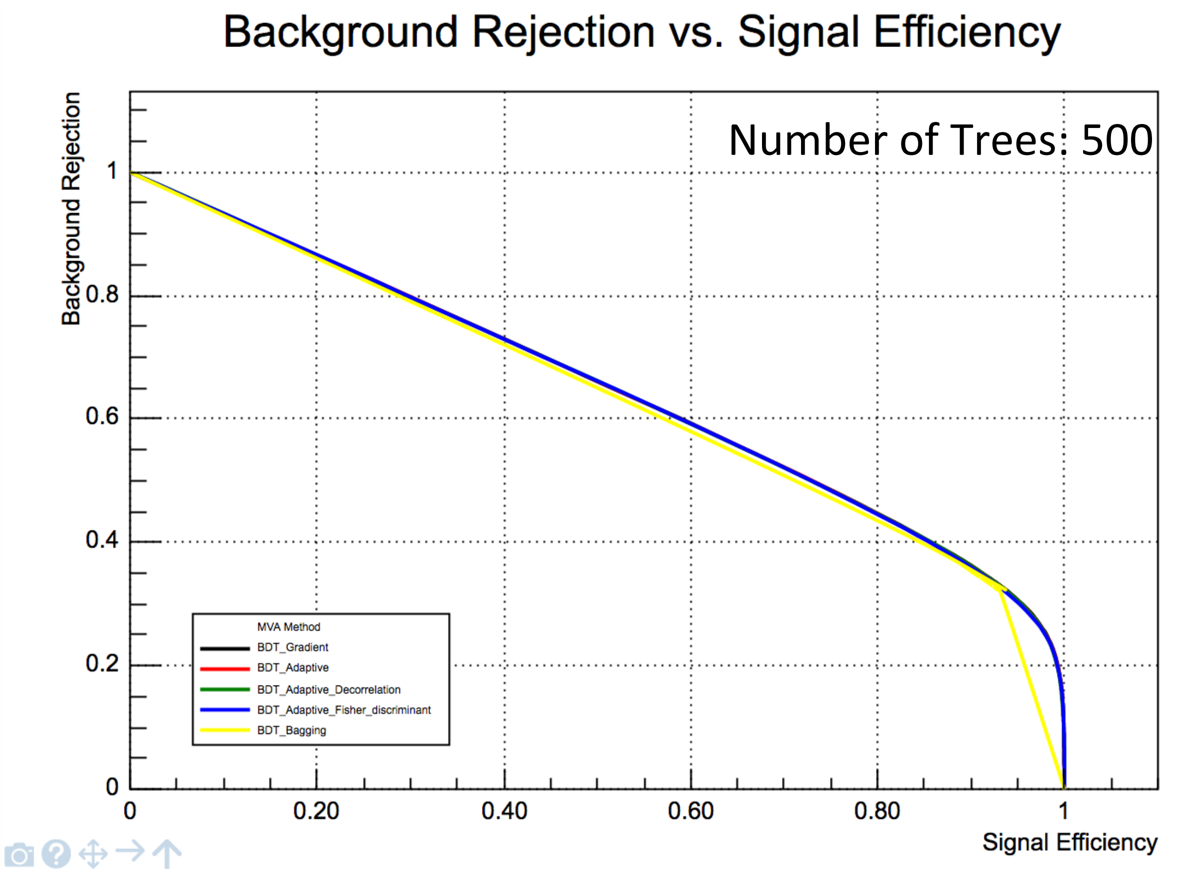
Yield (after pre-selection) = luminosity (40) x cross-section (S/B) x eff(pre-selection)

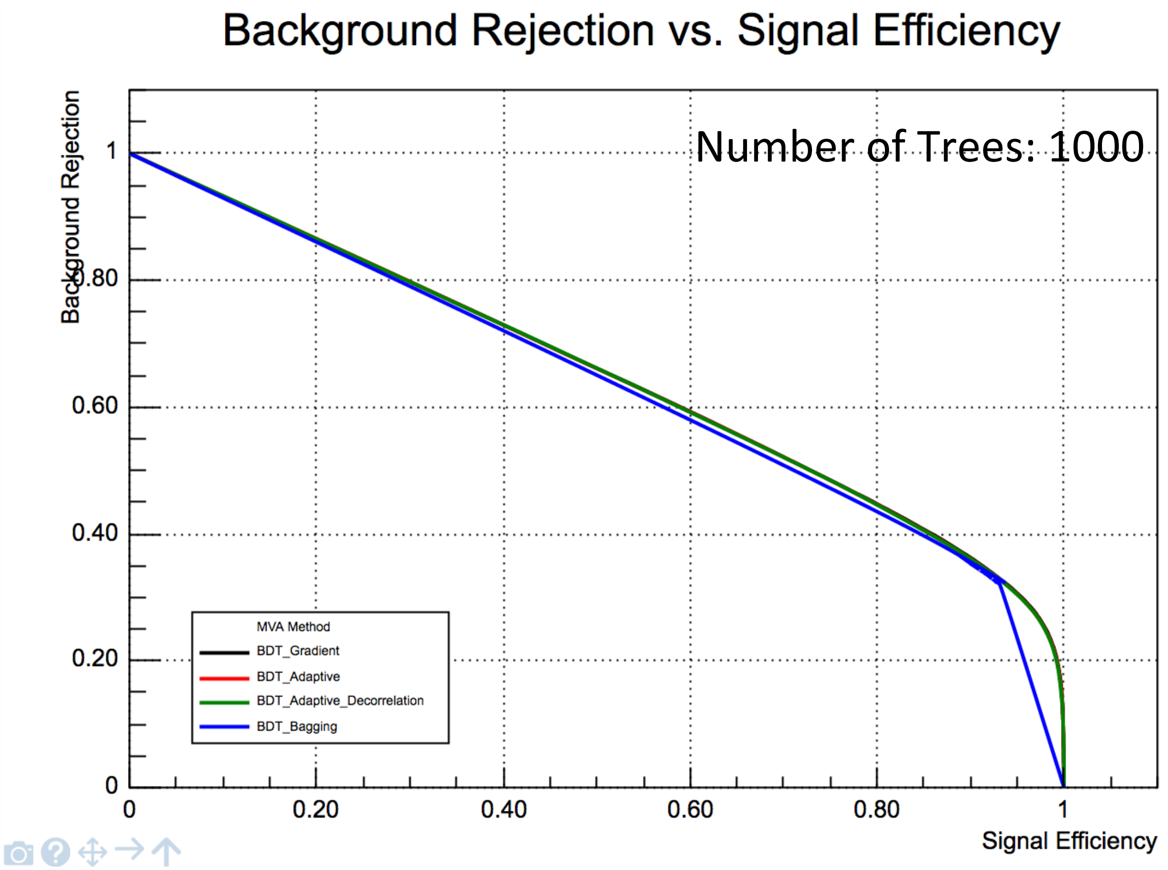
Yield (after final selection) = luminosity (40) x cross-section (S/B) x eff(pre-selection) x eff(S/B selection)

Uncertainty = √ (number of raw events) / (number of raw events)

Uncertain events = Yield x Uncertainty

Other ROC Curves obtained on Swan[1] from different training configurations





[1] SWAN (Service for Web based ANalysis) is a platform to perform interactive data analysis in the cloud.