Identifying Boosted Hadronic W Bosons

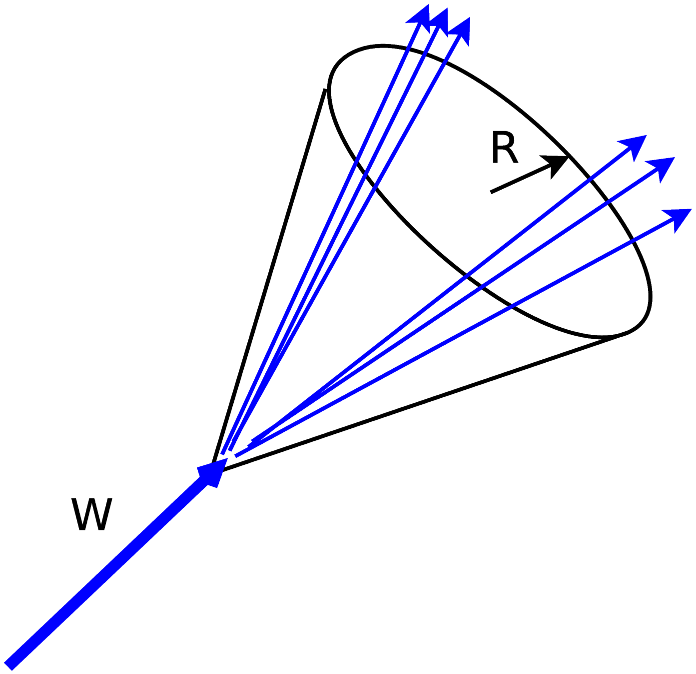
Zhenyu Han
Harvard University

With Yanou Cui and Matt Schwartz
(arXiv:1012.2077)

1/14/2011, Boston Jet Physics Workshop

Motivation: Tagging W-jets

- Boost W's: WW scattering, Z'->WW, t'->W+b, b'->W+t....
- Hadronically decaying W looks like a single fat jet in a collider detector

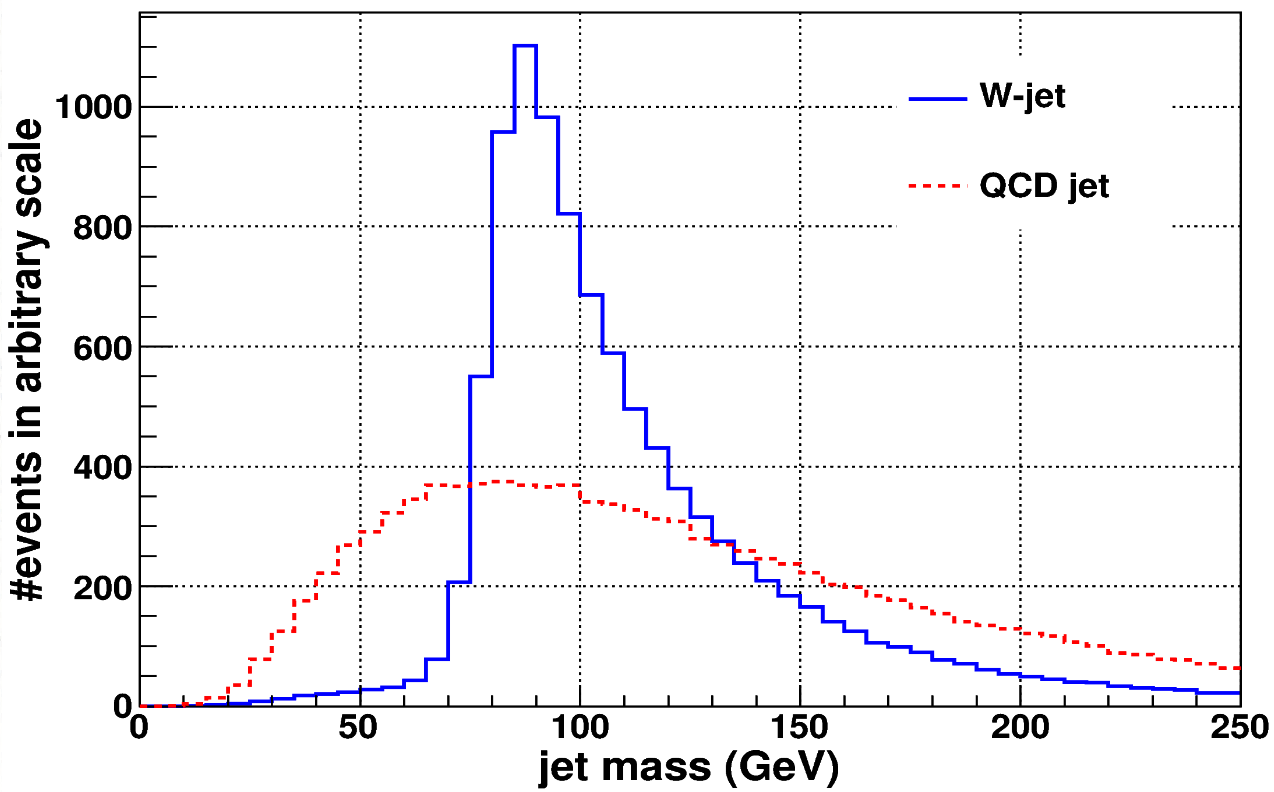


$$R = \sqrt{\Delta\phi^2 + \Delta y^2}$$

Experimentally R: 0.4 ~ 0.7

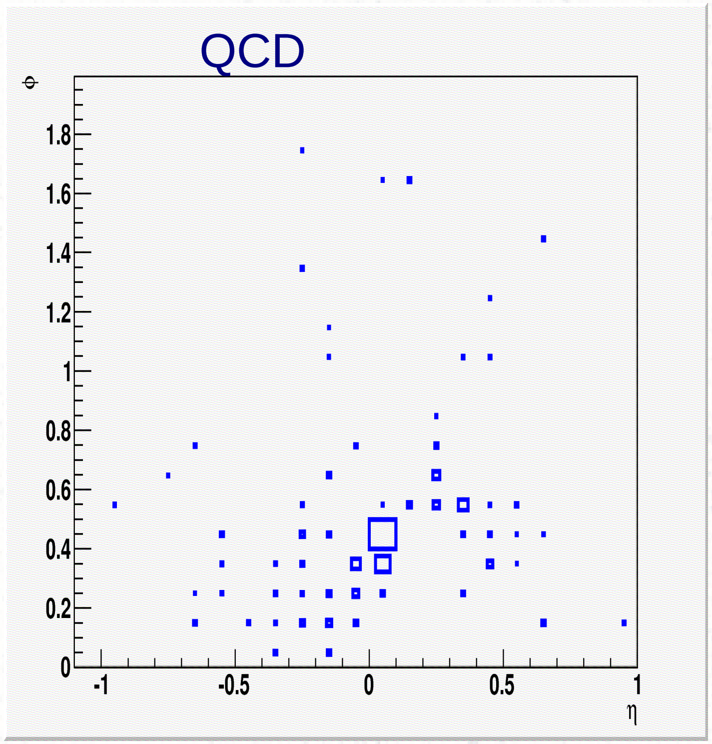
$$R_{ud} \sim 2m_W/p_T$$

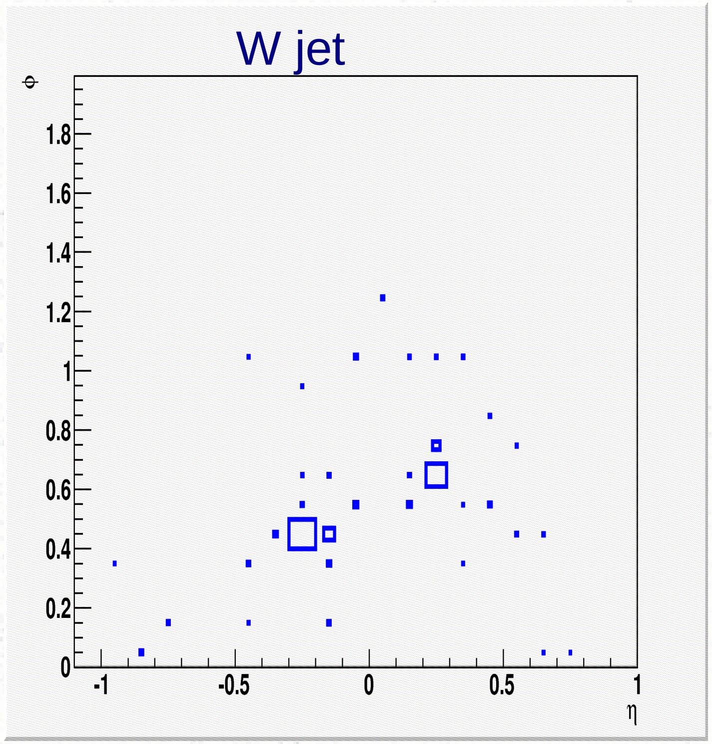
Jet Mass



Pt ~(500, 550)GeV, R=1.2

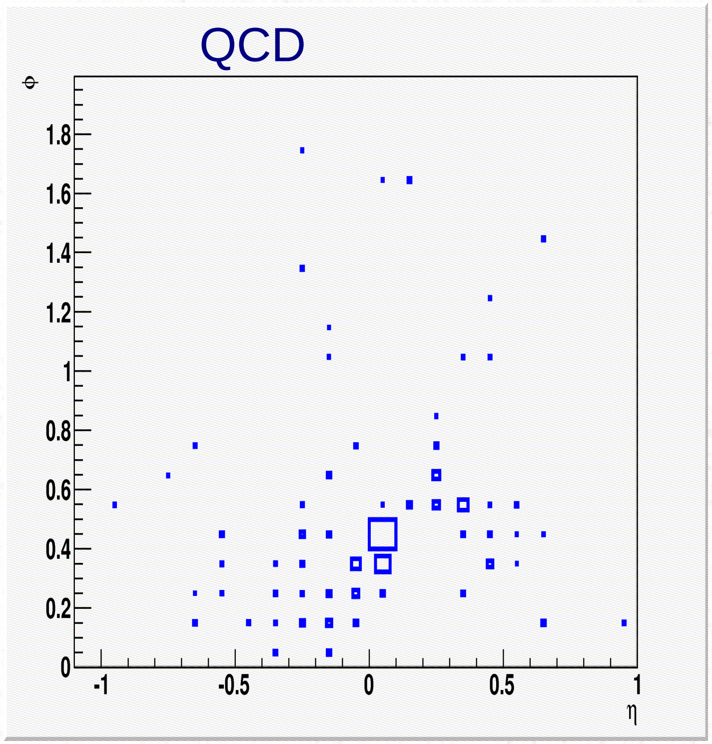
QCD jet vs W-jet

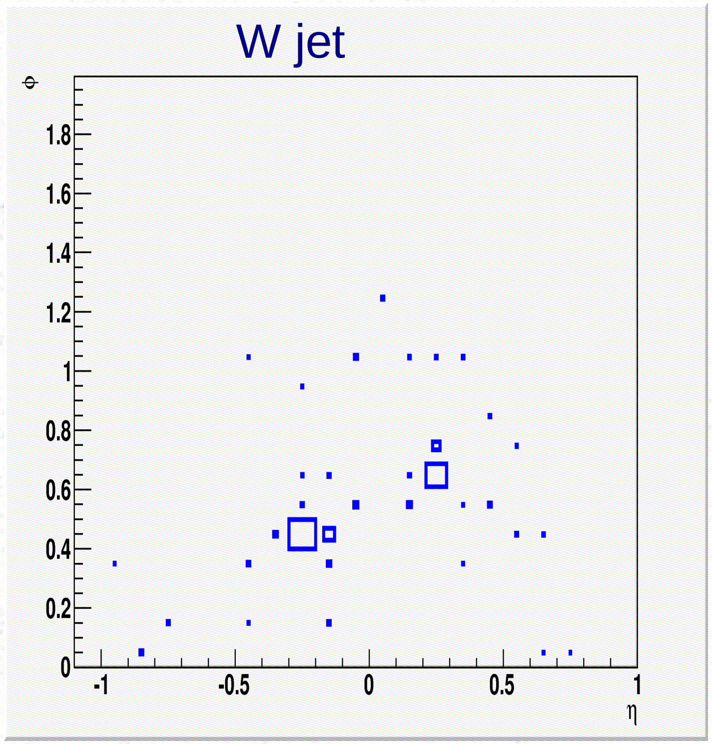




Group the energy in 0.1x0.1 bins on (eta, phi) plane.
Jets found using R=1.2, C/A.
QCD jet from W+j->lvj, W-jet from WW->lvjj, Madgraph+Pythia 8

QCD jet vs W-jet





Two major differences

- 2 balanced “subjets” in W-jets
- W-jet cleaner: color singlet

W-tagging

- 2 balanced subjets:
  - Filtering/mass drop: *Butterworth, Davison, Rubin & Salam*

    (see talks by Christopher/Adam/Minho/Jing...)
  - trimming/pruning *(Krohn, etal/Ellis, etal)*

    * *Extensively studied*
- Color singlet
  - Jet shape variables: planar flow/angularity/nsubjettiness

    * *Not sufficiently explored*
- Combining variables to optimize W-tagging
- Same method for Higgs/Z

Outline

- Optimizing procedure
  - The goal: maximize the statistical significance
  - Variables distinguishing W-jets from QCD-jets
  - Multivariate Analysis
- Application
  - Z'->WW
  - W+jet in dijet events
- Pythia 8 vs Herwig++
- Conclusion

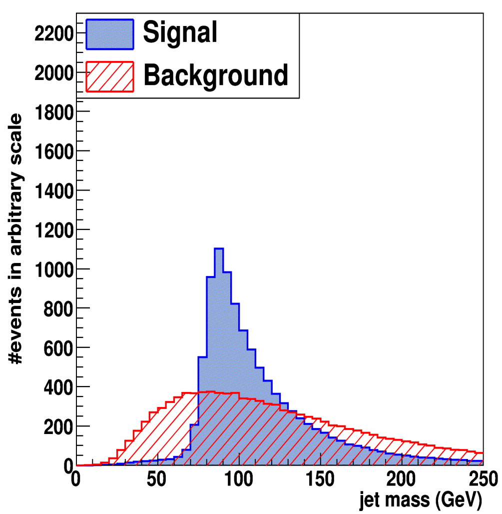
Maximize the Significance

- Data samples: SM WW->lvqq (signal), Wj->lvj (background), Madgraph+Pythia8
  - Binned in 0.1x 0.1 calorimeter cells
  - FastJet, R=1.2 C/A
  - Jet PT 200~1000 GeV, divided in 50 GeV bins
- Initial number of high pt jets: $n_S^0, \quad n_B^0$
- Final number after cuts: $n_S, \quad n_B$
- Efficiency: $\varepsilon_S = n_S/n_S^0, \varepsilon_B = n_B/n_B^0$
- Significance Improvement Characteristic:

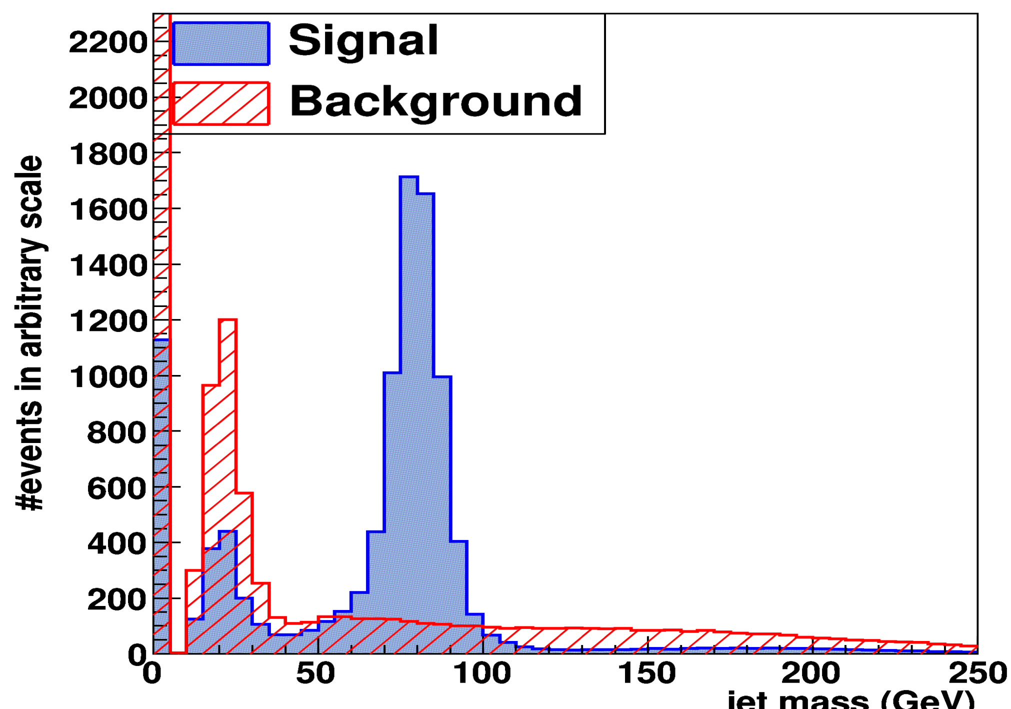
$$\mathrm{SIC} = \frac{\varepsilon_S}{\sqrt{\varepsilon_B}}$$

Filtering with mass drop

- “Clean” the jets, reduce background
- Define subjets

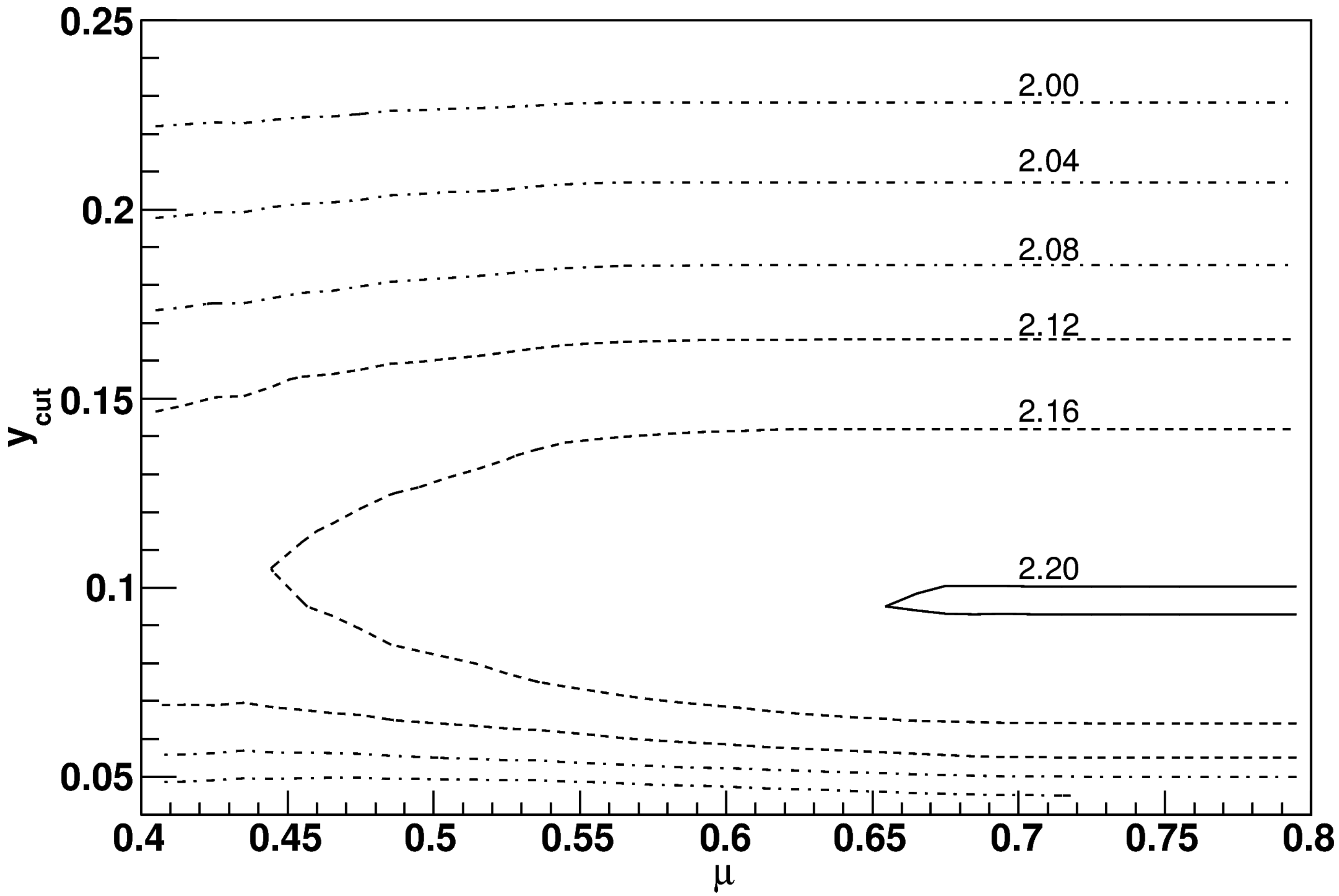


Jet mass (R=1.2), pt~(500, 550)GeV



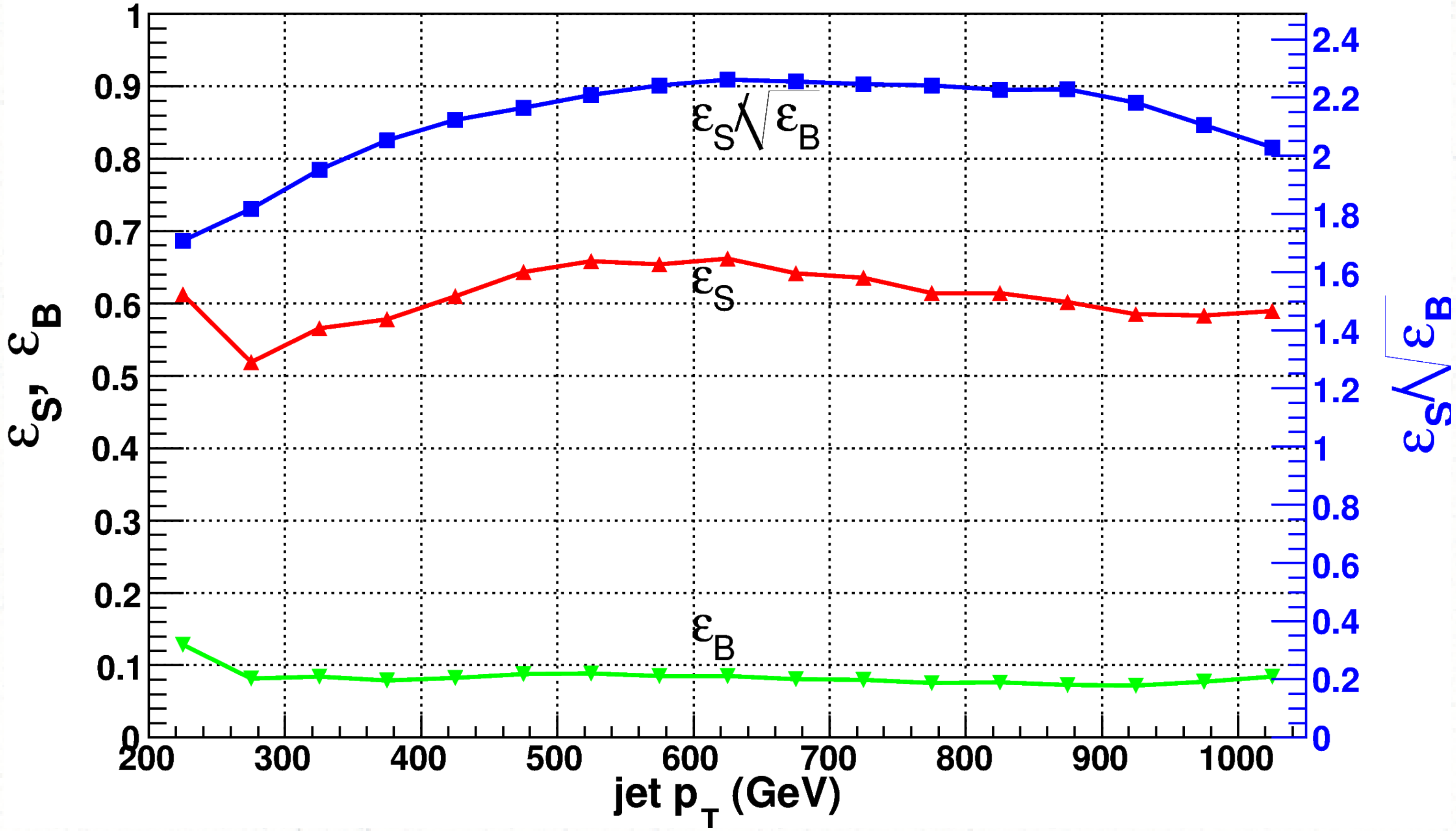
jet mass after filtering w/ mass drop

Filtering/mass drop parameters



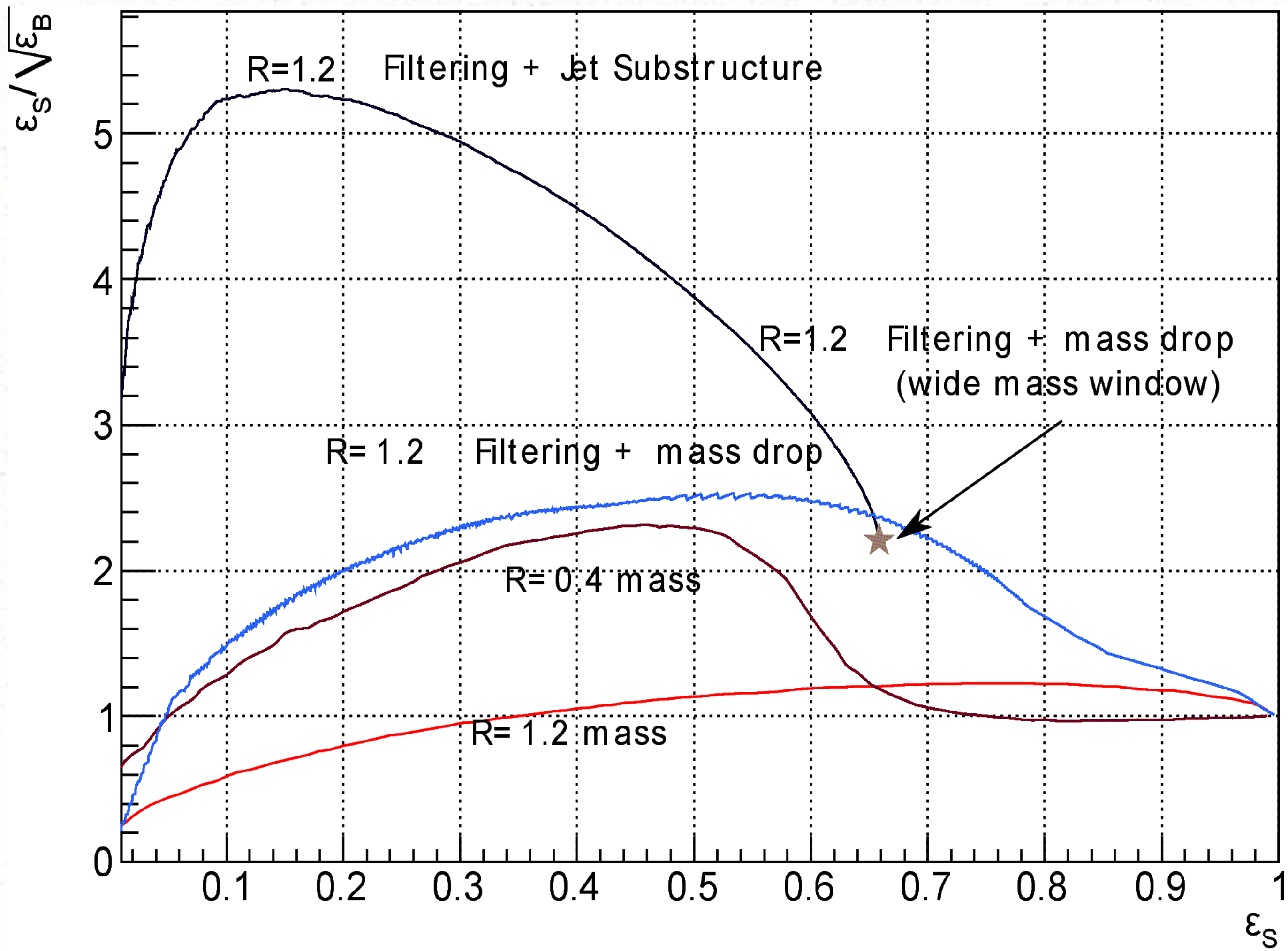
For jet pt (500, 550) GeV, filtered mass cut (60, 100) GeV

Significance (SIC)



Gain a factor of ~2 using filtering.
trimming/pruning works similarly

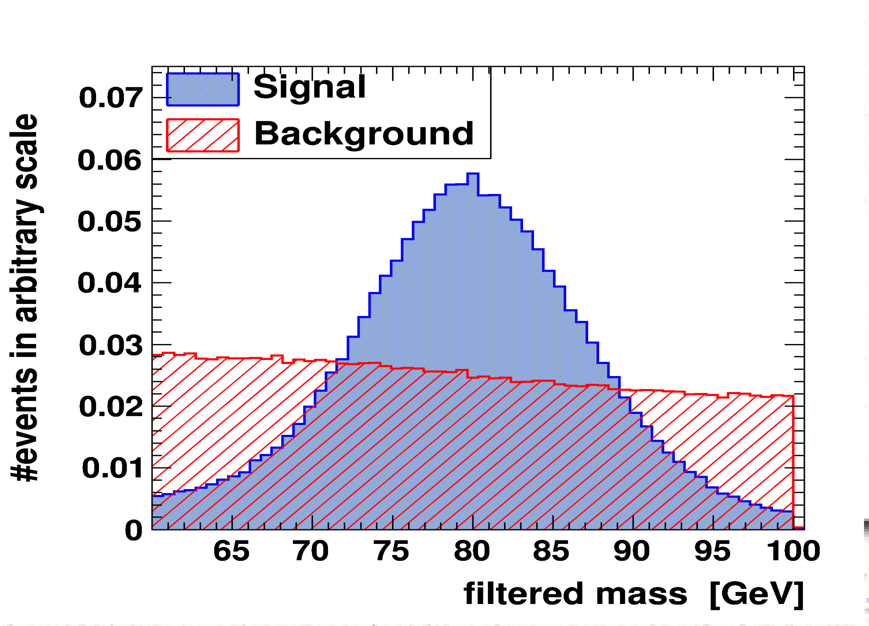
Comparing SIC



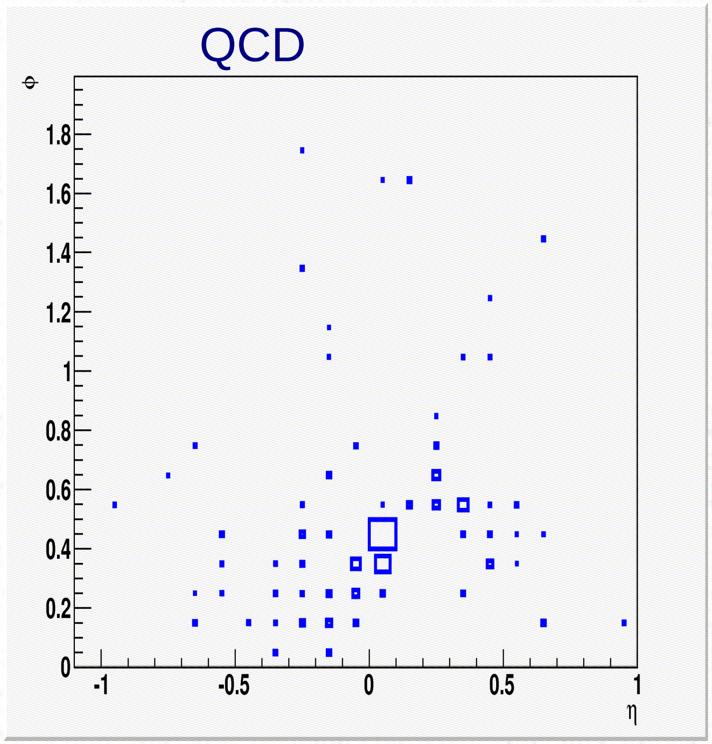
PT (500, 550)GeV

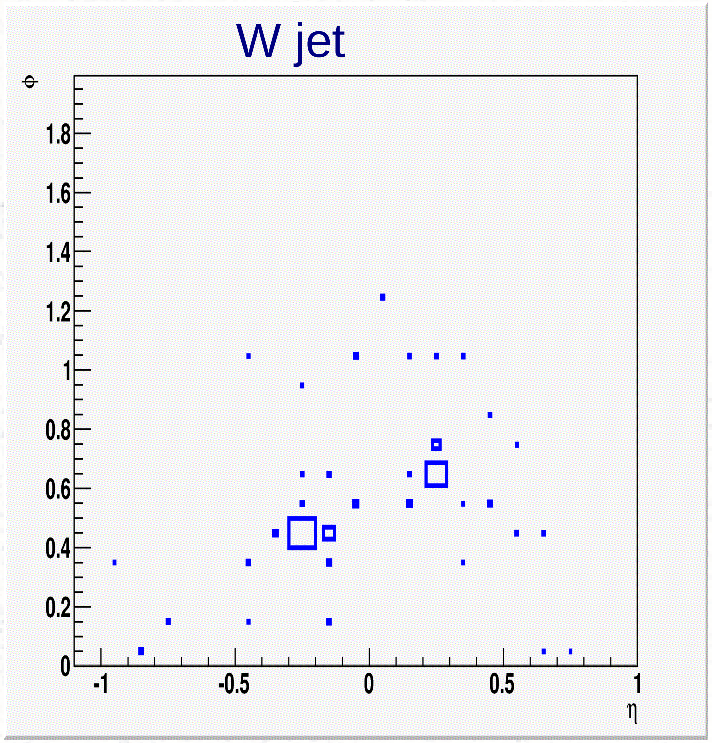
The Variables

- Keep events passing the filtered mass window cut (60, 100) GeV
    - Filtered mass
    - subjet pt ratio
    - Number of subjets; subjet pt's, masses
    - jet pt/mass for different R's (R-cores)
    - Planar flow, pull...



QCD jet vs W-jet

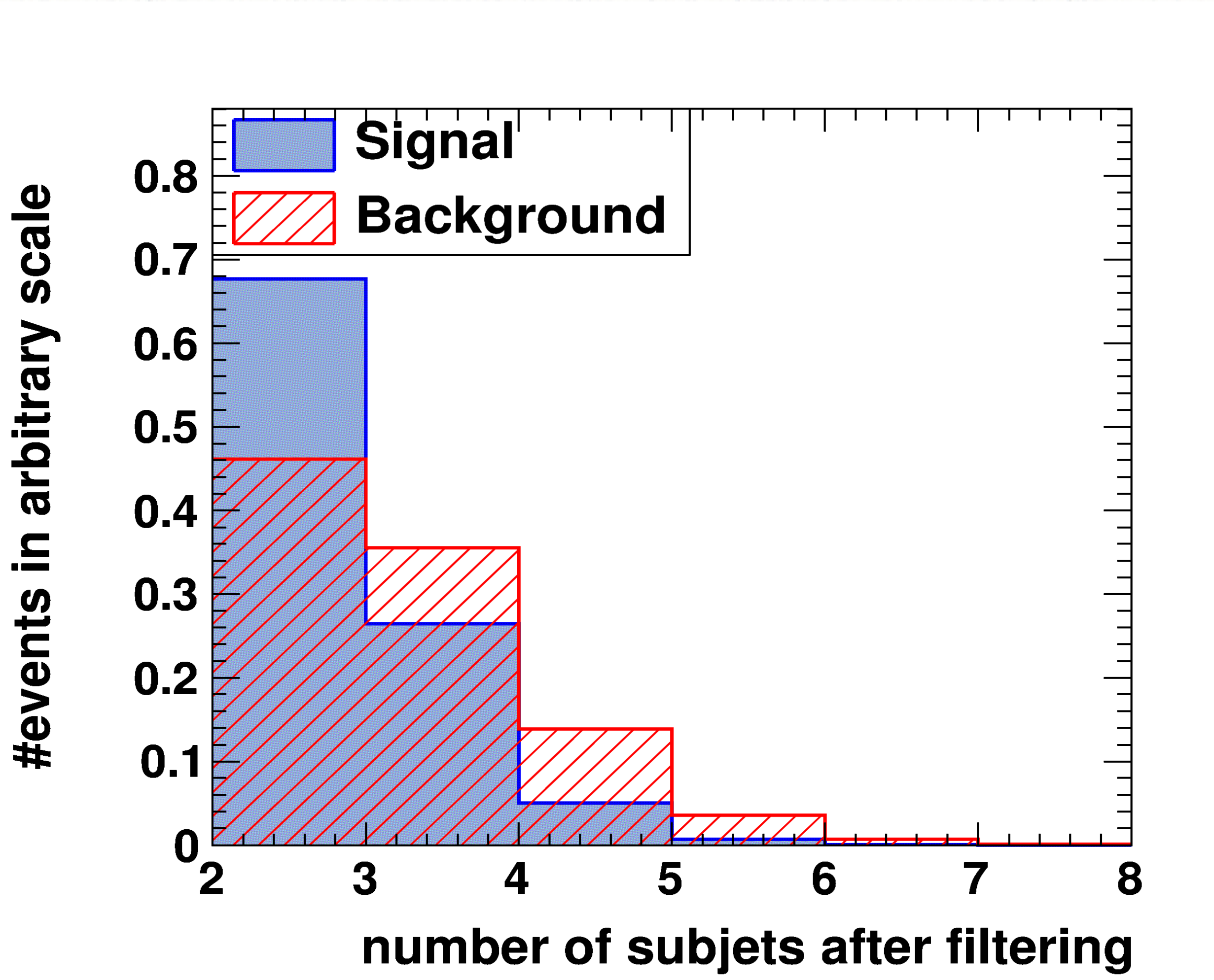




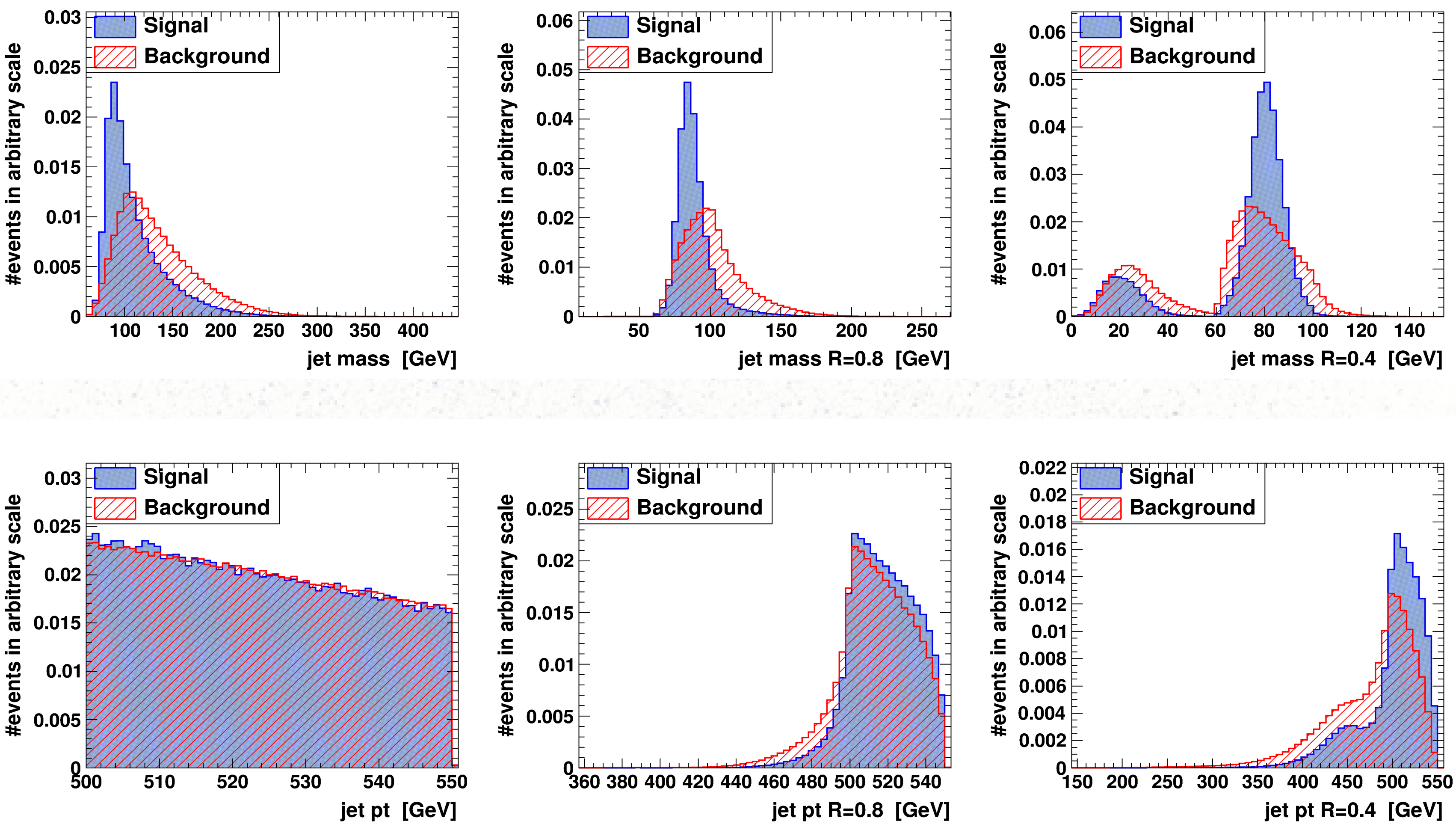
Two major differences

- 2 balanced “subjets” in W-jets
- W-jet cleaner: color singlet

Number of subjets (pt>10GeV)



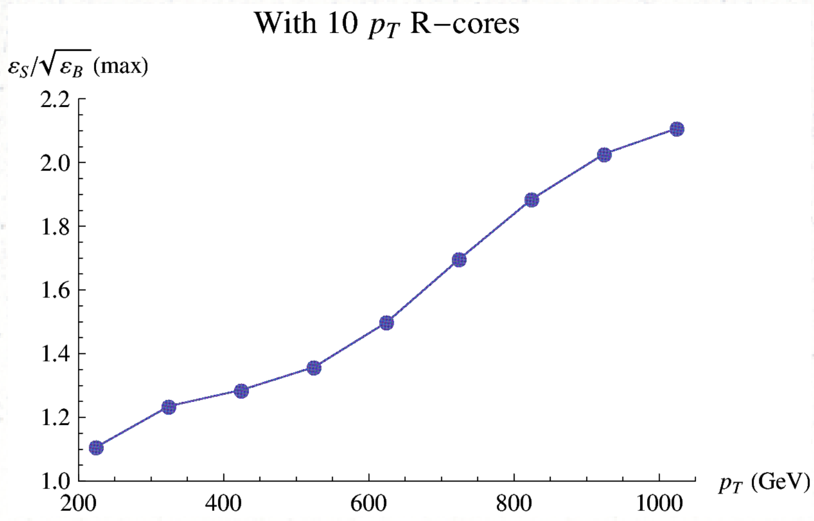
Jet mass/pt for different R



Recluster with different R, take leading jet

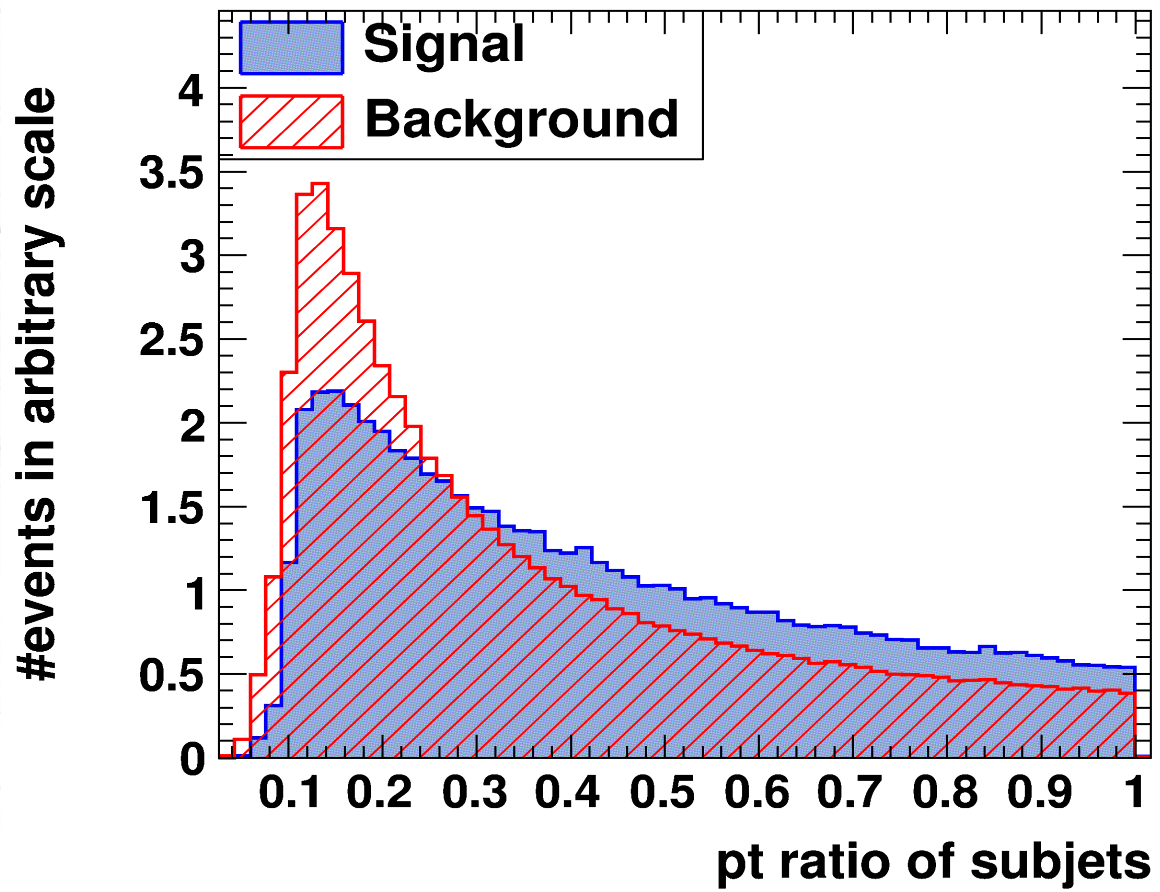
R-cores

- PT R-cores: $C_{P_T}(R) \equiv P_T(R)/P_T(R_{\text{fat}})$
- Mass R-cores: $C_m(R) \equiv m(R)/m(R_{\text{fat}})$
- R_fat = 1.2, R=0.2~1.1
- Not very useful when individually used
- Combined give good discriminating power (on top of filtering)

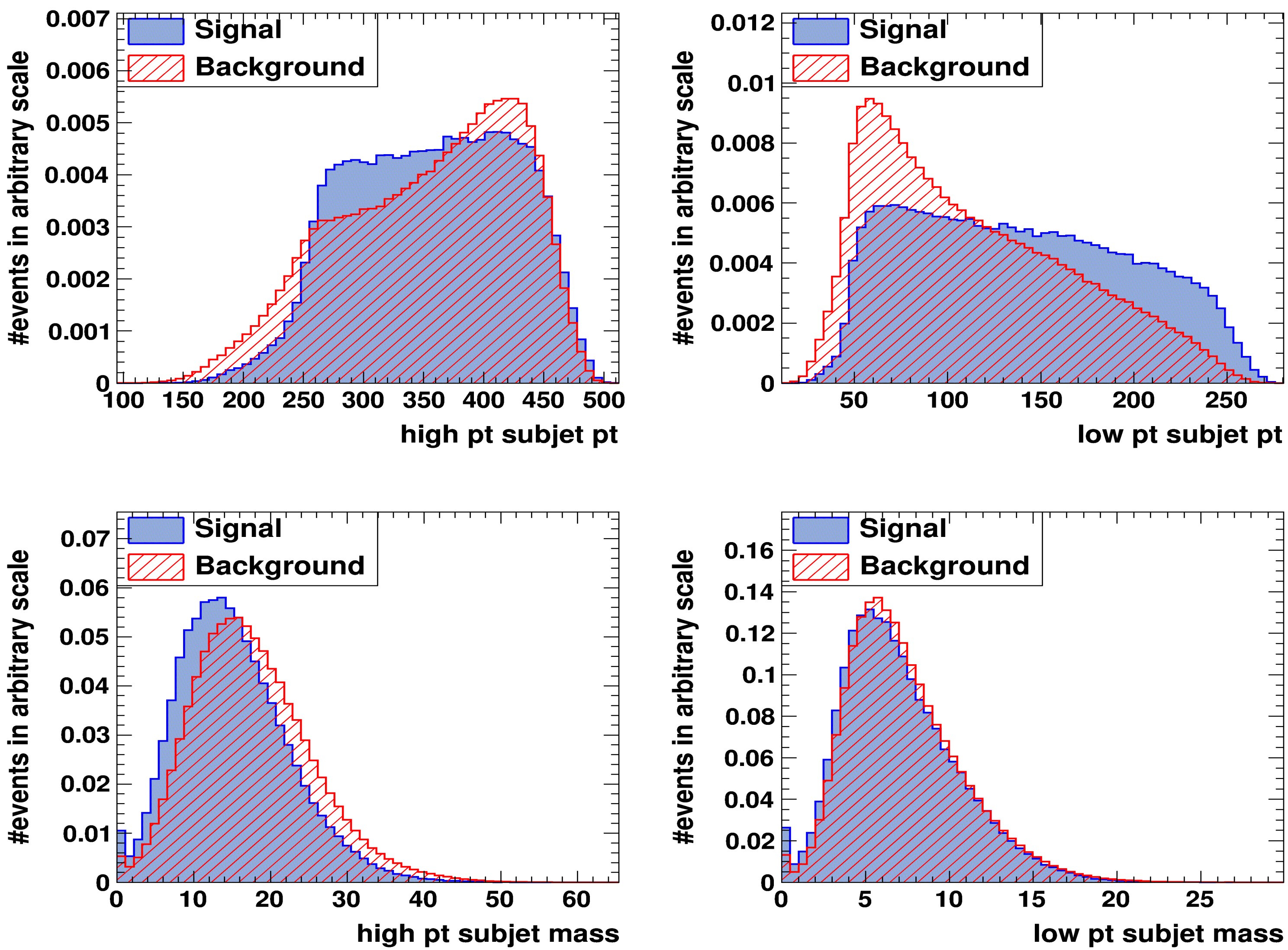


Subjets Pt ratio

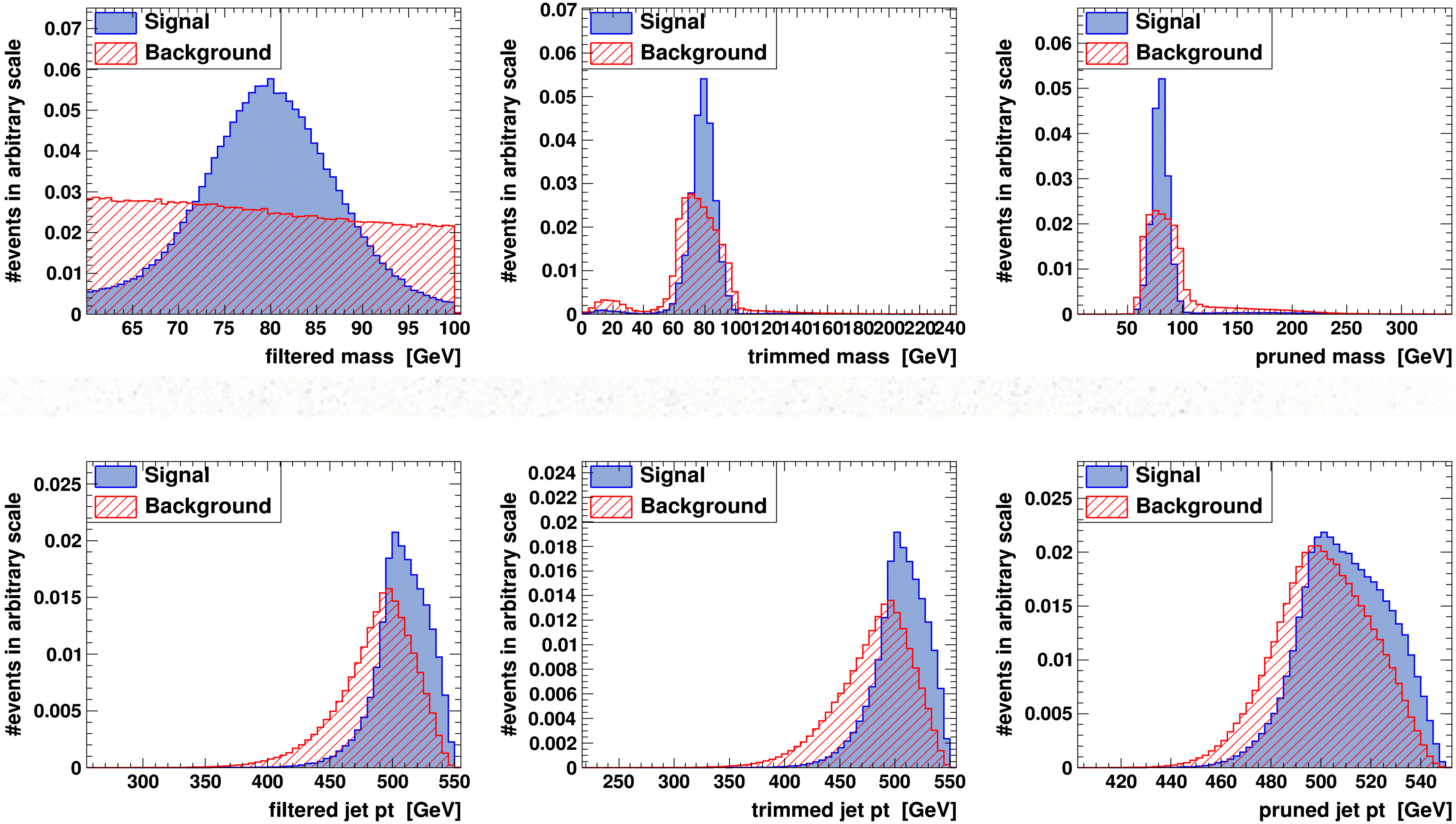
- 2 highest pt subjets, signal more balanced (leftover from filtering)



Subjet pt/mass

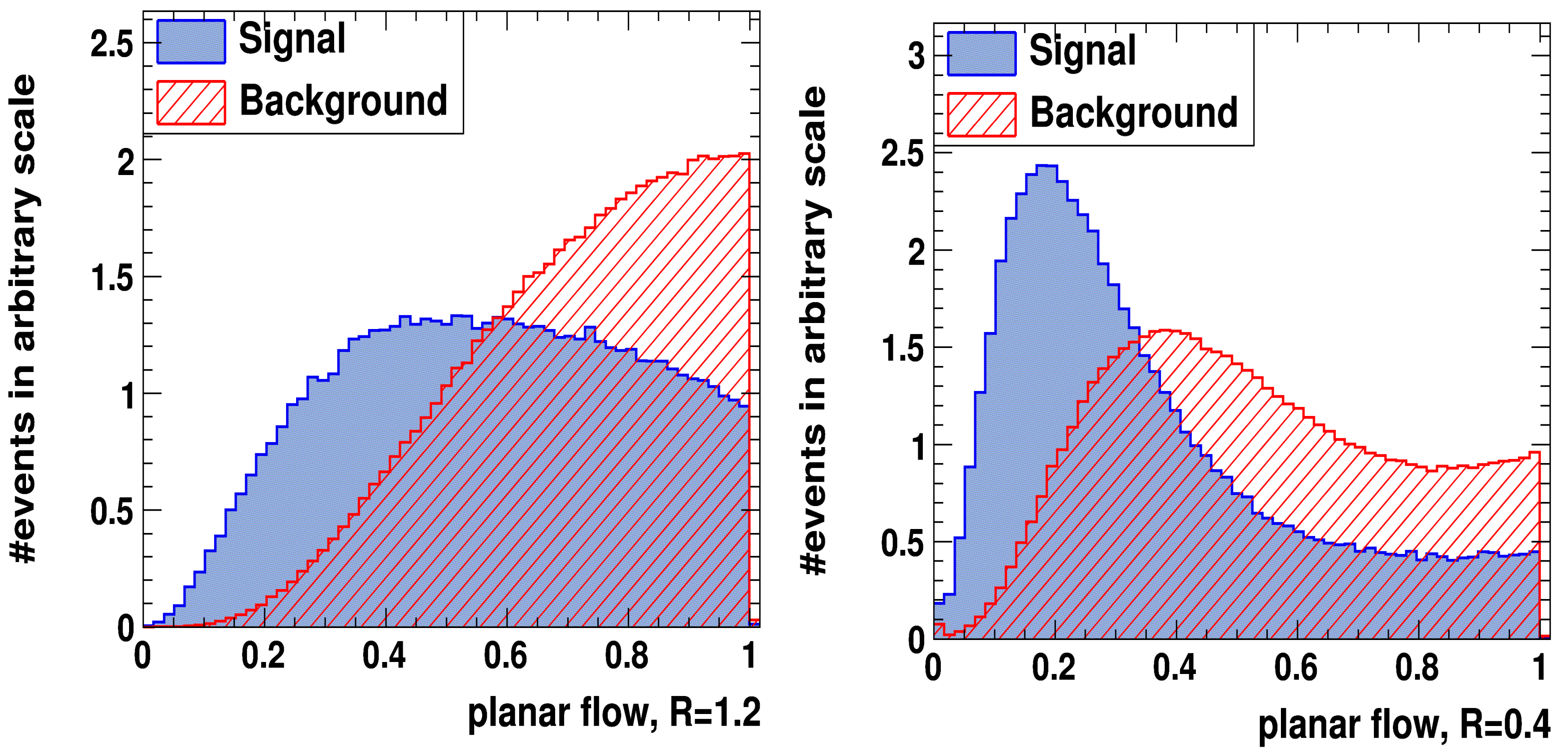


filtered/trimmed/pruned mass/pt



Soper & Spannowsky: combining different algorithms enhance ZH detection

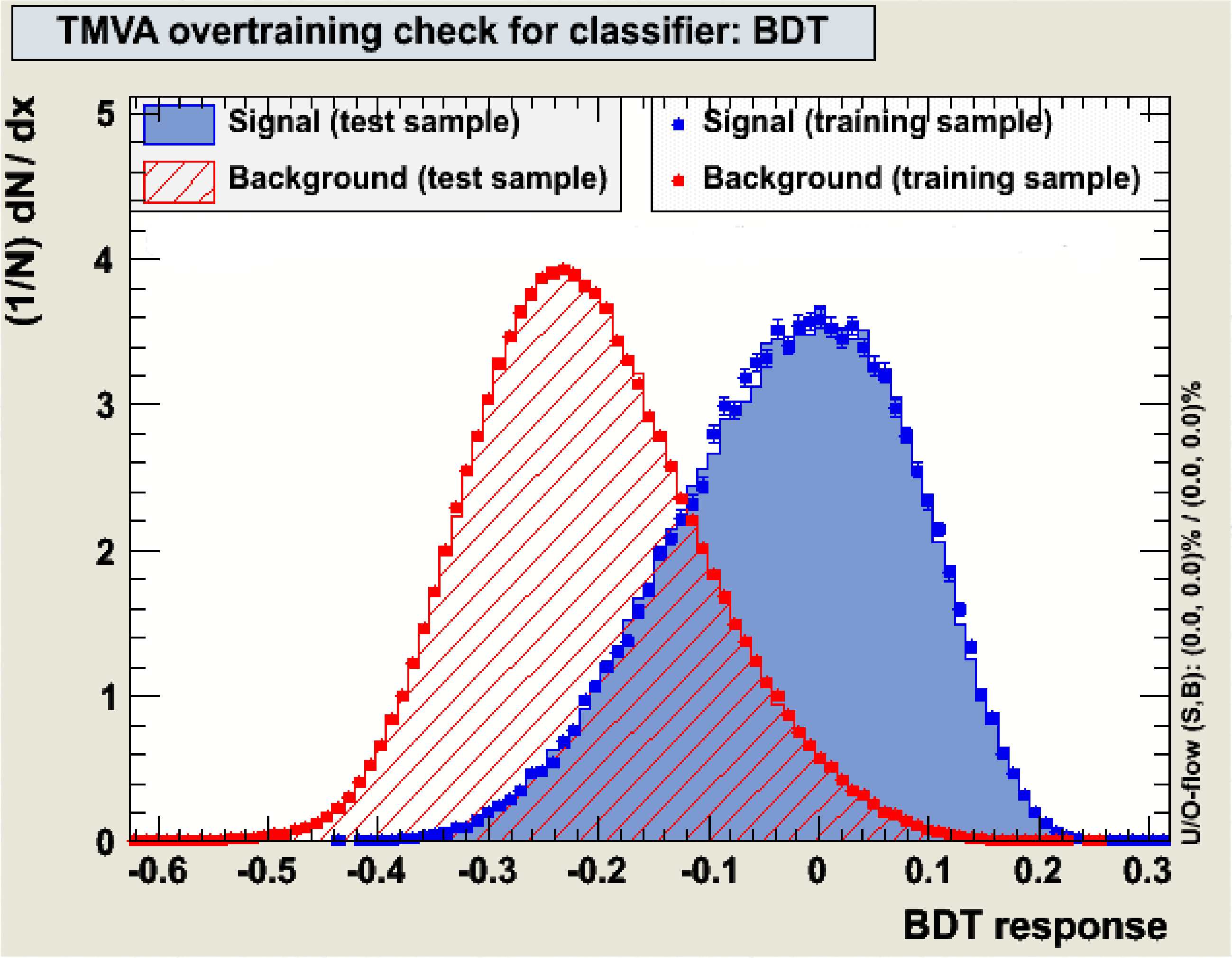
Planar flow


Signal
Background
#events in arbitrary scale
0
0.5
1
1.5
2
2.5
0
0.2
0.4
0.6
0.8
1
planar flow, R=1.2
Signal
Background
#events in arbitrary scale
0
0.5
1
1.5
2
2.5
3
0
0.2
0.4
0.6
0.8
1
planar flow, R=0.4

Multi-variable analysis

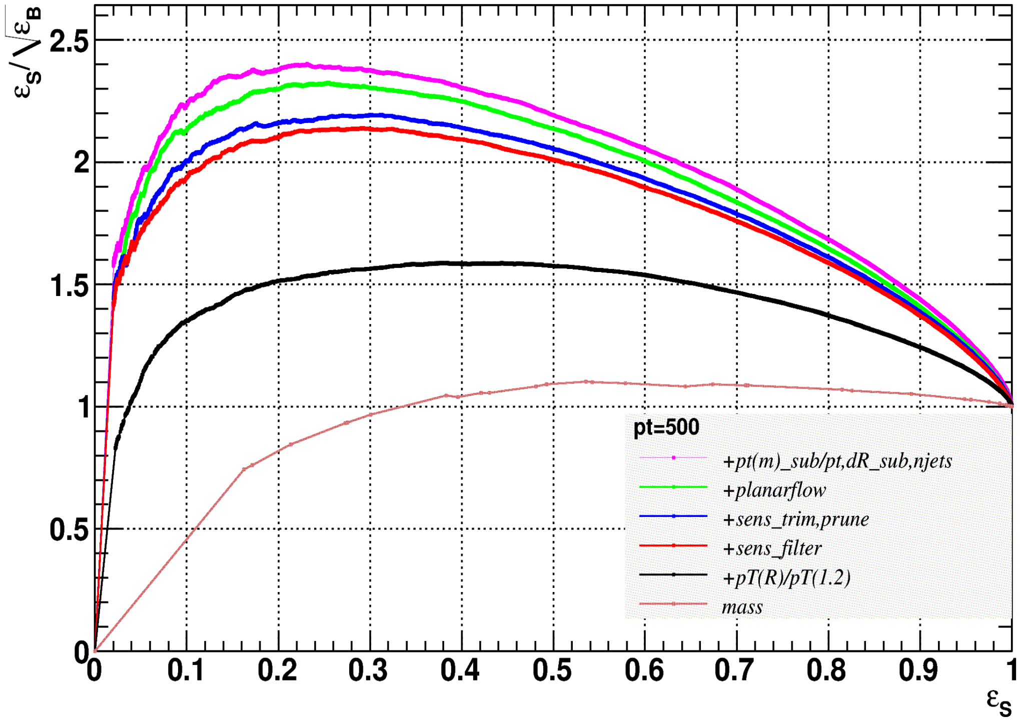
- Simple cuts do not improve significantly, <20% improvement
- Variables correlated
- Use TMVA (Toolkit for Multivariate Data Analysis with ROOT)
- Boosted decision tree (BDT)
  - train and test with signal and background data

BDT response



PT (500, 550)GeV

Efficiency and significance from MVA



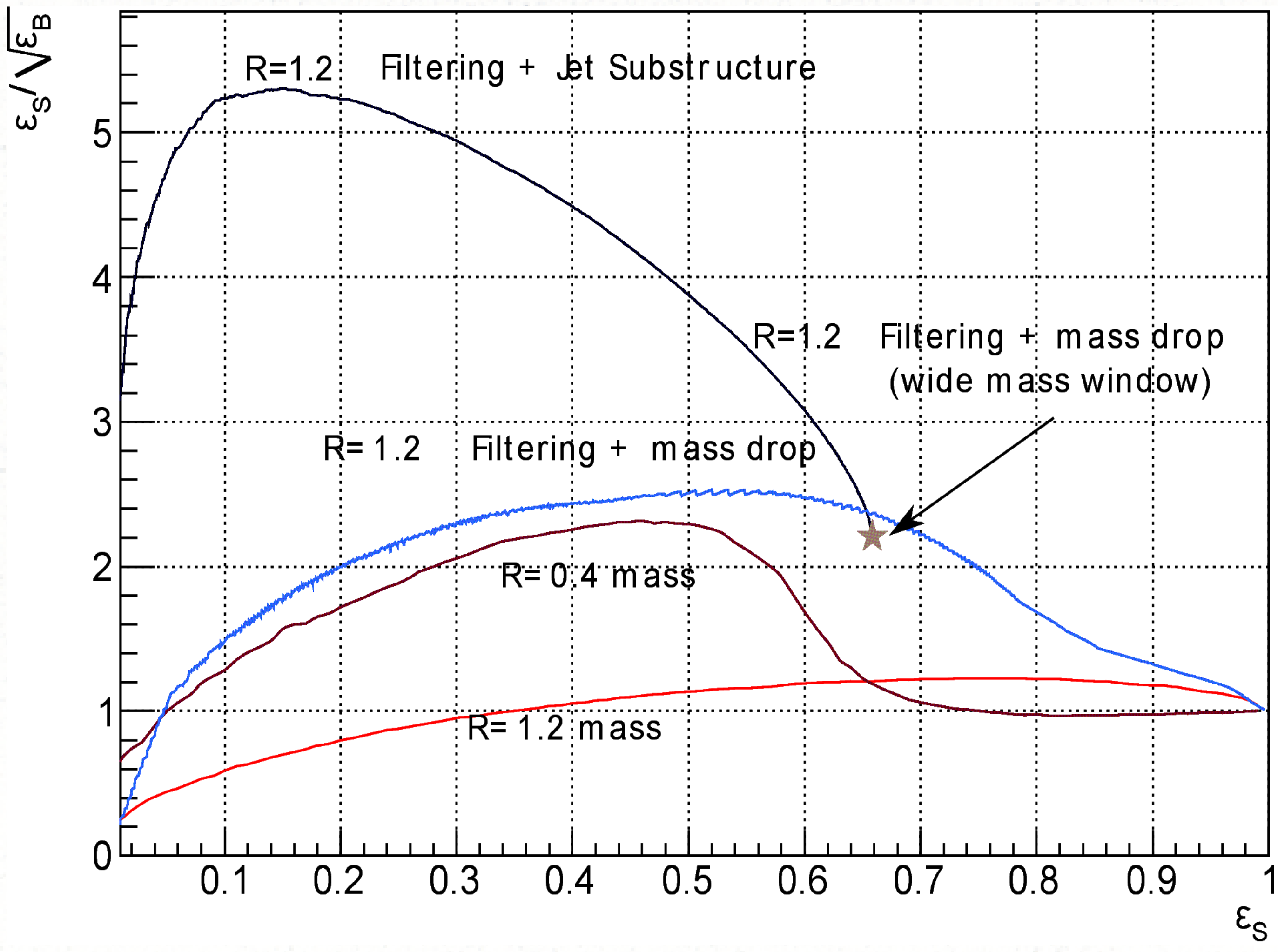
PT (500,550) GeV
On top of filtering

* 25 variables are used

$$m_{\rm jet},\ c_{p_T}(0.2-0.11),\ {\rm sens}^{m,p_T}_{\rm filt,trim,prun},\ P_f,\ P_f(0.4),\ \frac{p_T^{\rm sub1,sub2}}{p_T},\ \frac{m^{\rm sub1,sub2}}{m},\ \Delta R_{\rm sub},\ n_{\rm sub}$$

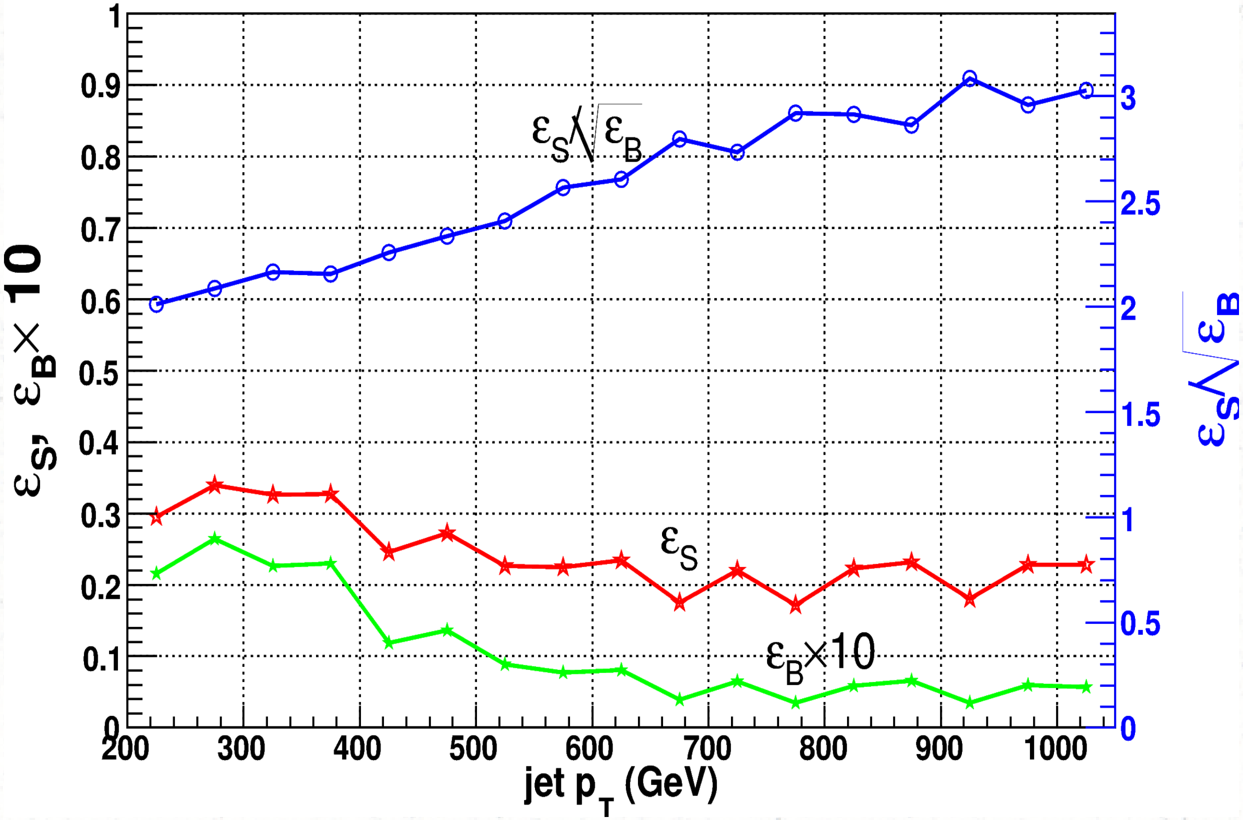
* A smaller set of 7 variables give ~1.9 for maximum SIC

Comparison



R=1.2 Filtering + Jet Substructure
R=1.2 Filtering + mass drop
(wide mass window)
R= 1.2 Filtering + mass drop
R= 0.4 mass
R= 1.2 mass
$\varepsilon_S/\sqrt{\varepsilon_B}$
ε_S
0
1
2
3
4
5
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Efficiency and significance from MVA



The total significance improvement after filtering + MVA: 3.4~6.7

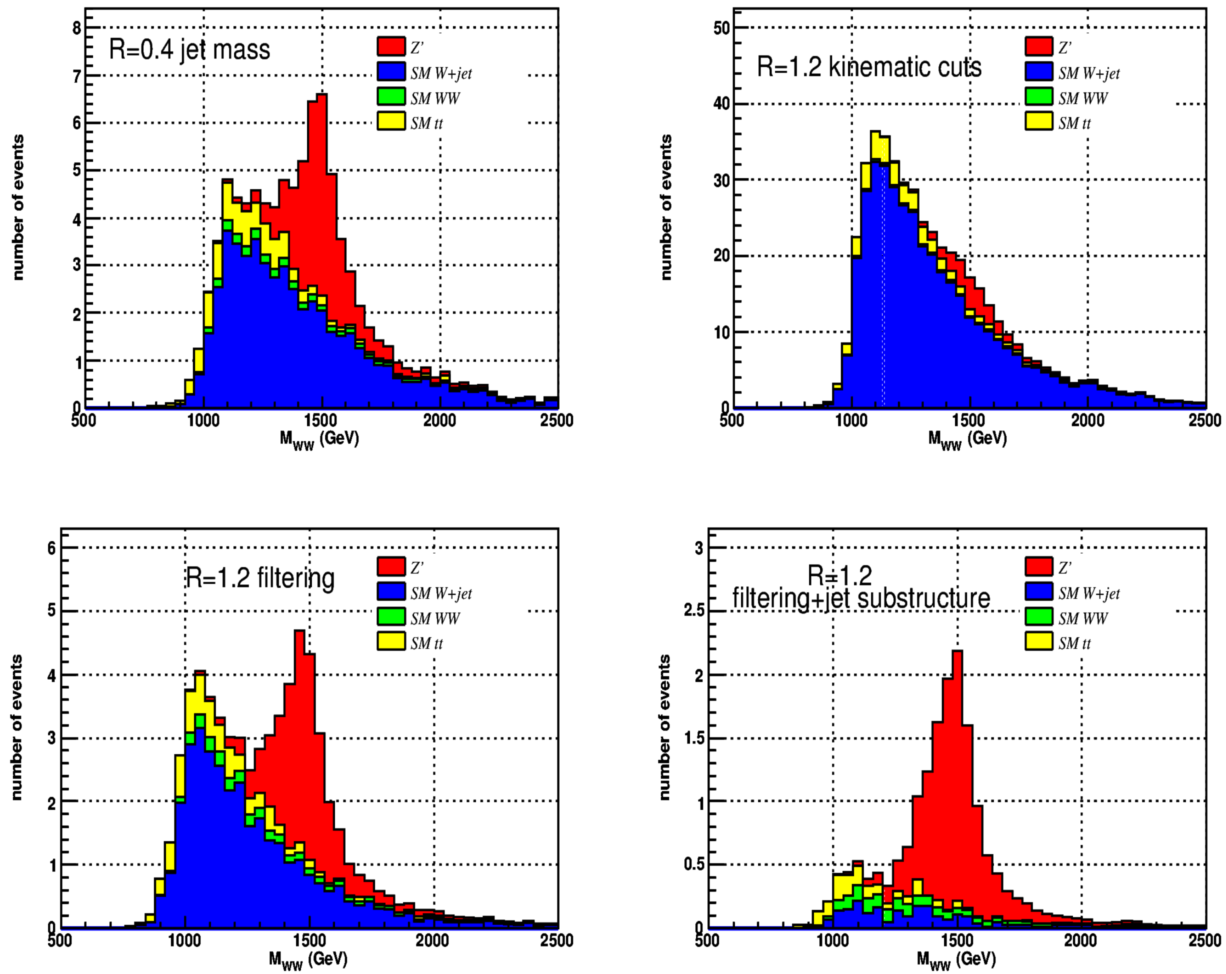
Test the robustness

- Consider other processes, use exactly the same
  - Jet grooming parameters
  - Filtered mass window cut (60, 100) GeV
  - Weight files from training WW/Wj
- Different Monte Carlo tools

Application: Z'

R=0.4 mass cut

R=1.2



Filtering

MVA

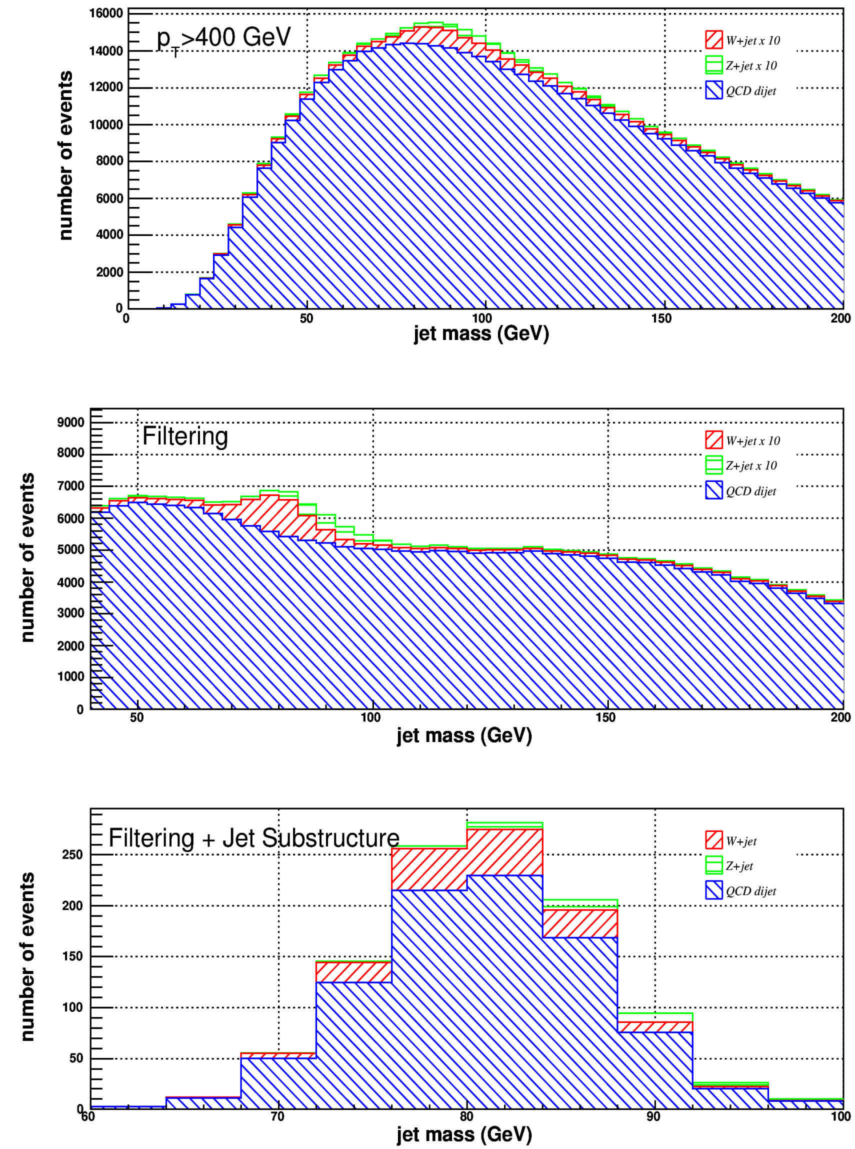
14 TeV LHC, MZ'=1.5TeV,
Z'->WW->lvj

$g_{Z'ff} = 0.2 g_{Zff}$ $\int \mathcal{L} = 2 fb^{-1}$

Events after MVA (Z':Wj:WW:tt) = 13:1.3:0.5:1.1

Dijet vs Wj at 7TeV

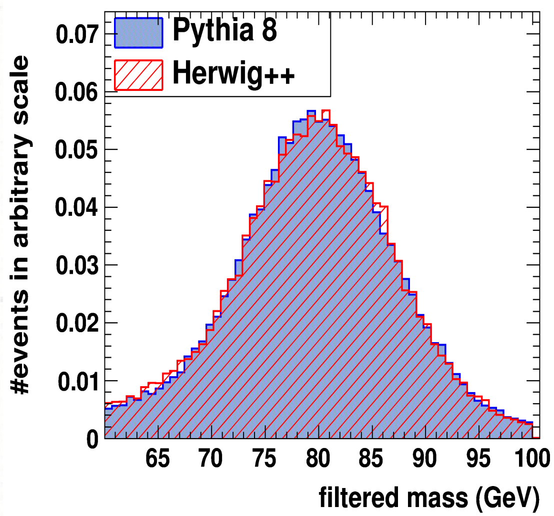
- R=1.2, pt cut > 400 GeV,
  1 inverse fb.
- Original:
  - S/B = 1.6k(/2)/0.50M
  - S/sqrt(B) = 1.1
- After filtering and MVA
  - 150 vs 940
  - S/sqrt(B) = 5.1, S/B=0.17
- S eff 26%, B fake rate: 0.34%
- hadronic W+j can be 'discovered' at 7TeV.

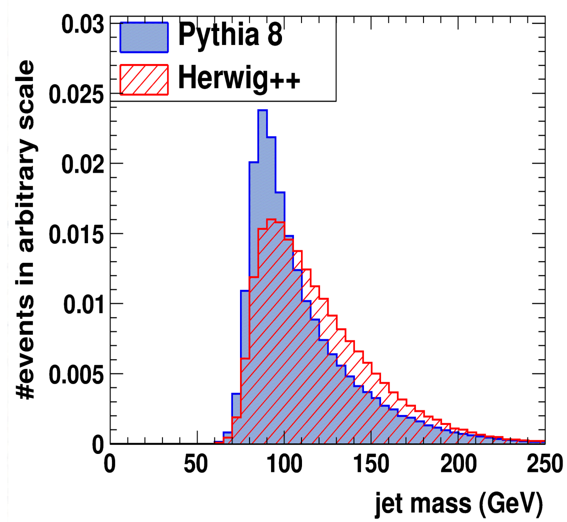


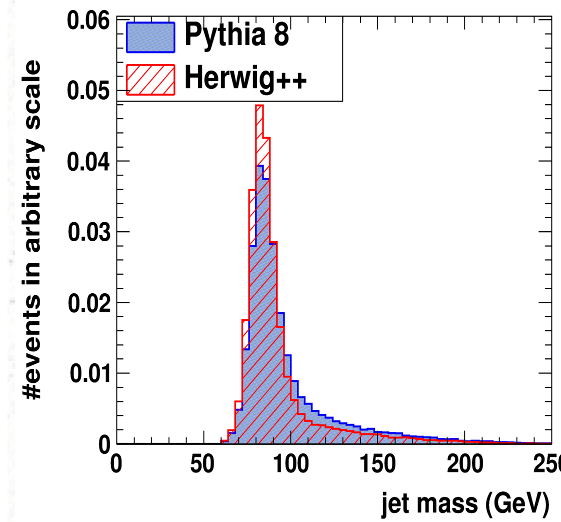
Different Monte Carlo's

- Herwig++ vs Pythia 8, same processes: WW/Wj
- Similar results from filtering:
  - efficiency (S/B): 0.64/0.087 (Herwig++), 0.66/0.089(Pythia 8)
  - significance: both 2.2
- Differences in MVA
  - Underlying events modeled differently

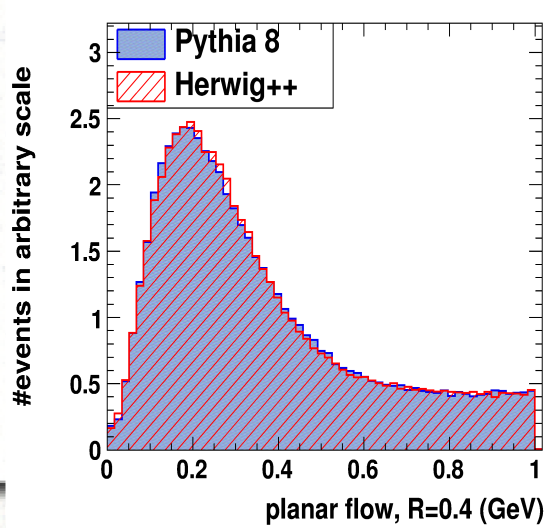
Herwig++ vs Pythia 8 (signal)

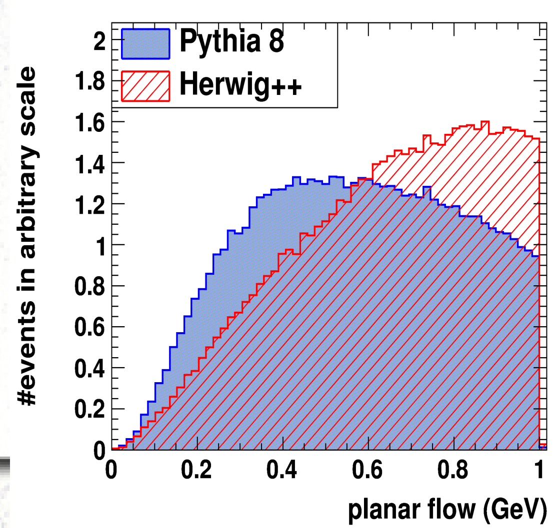






No UE

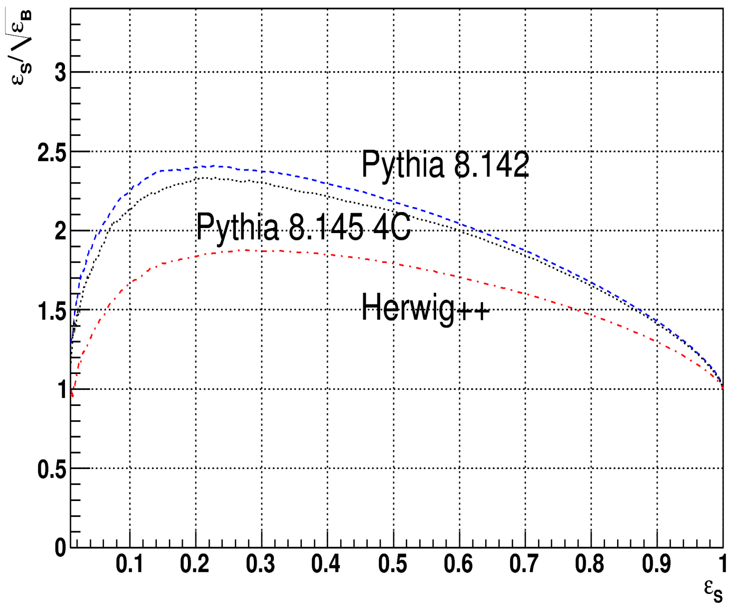


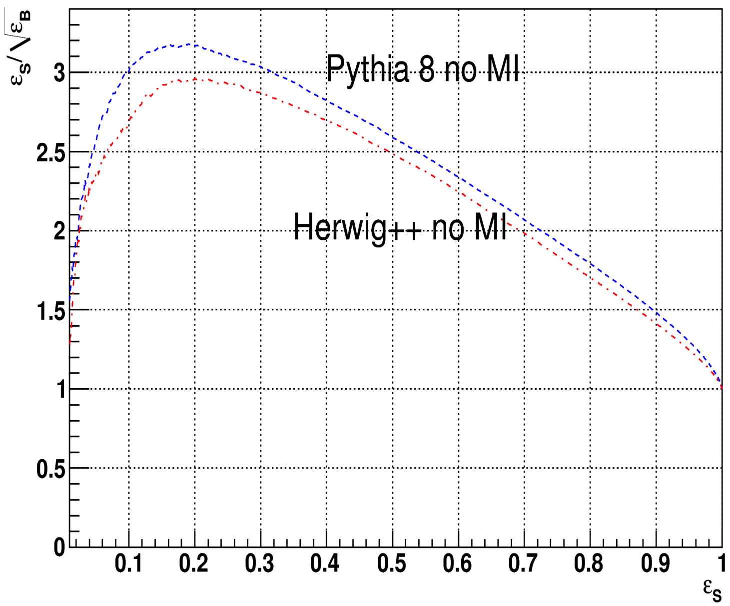


Herwig++ vs Pythia8

- Apply BDT weight files from training Pythia8 data on Herwig++ data (and a different Pythia tune)

PT (500, 550) GeV





- Need to be resolved using the LHC data

Conclusion

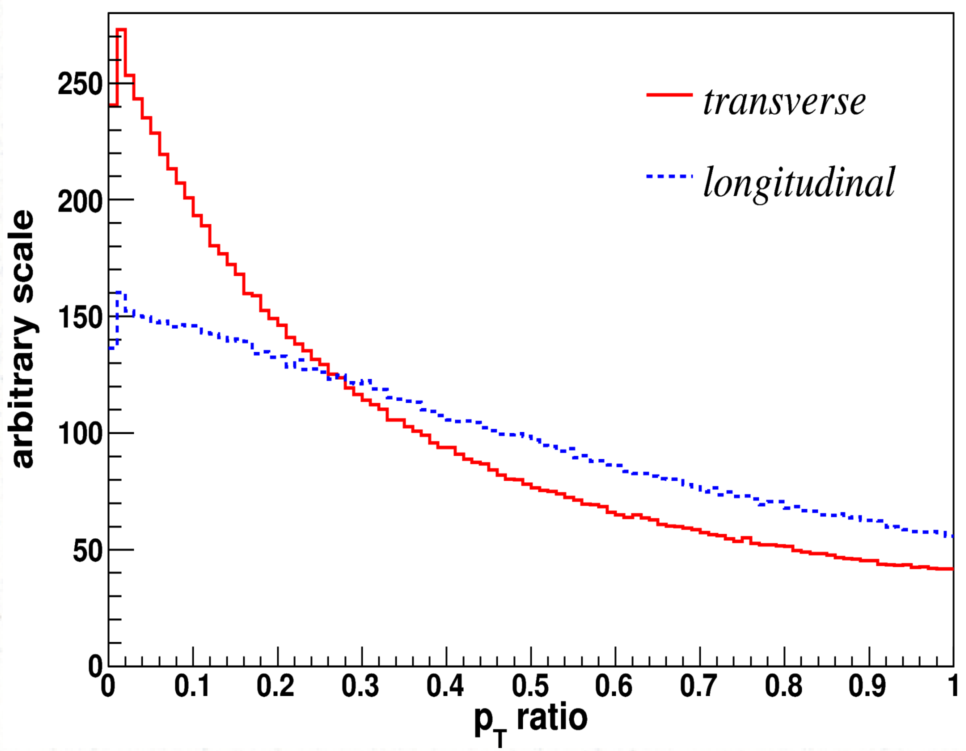
- Boosted hadronic W bosons can be efficiently distinguished from QCD jet using jet substructure.
- Starting from high PT fat jet, jet grooming algorithms can improve the significance by a factor of ~2.
- Multi-variable analysis improves further by ~2.
- Many applications, and awaiting tests at the LHC.
- Code publicly available:

  http://jets.physics.harvard.edu/wtag/

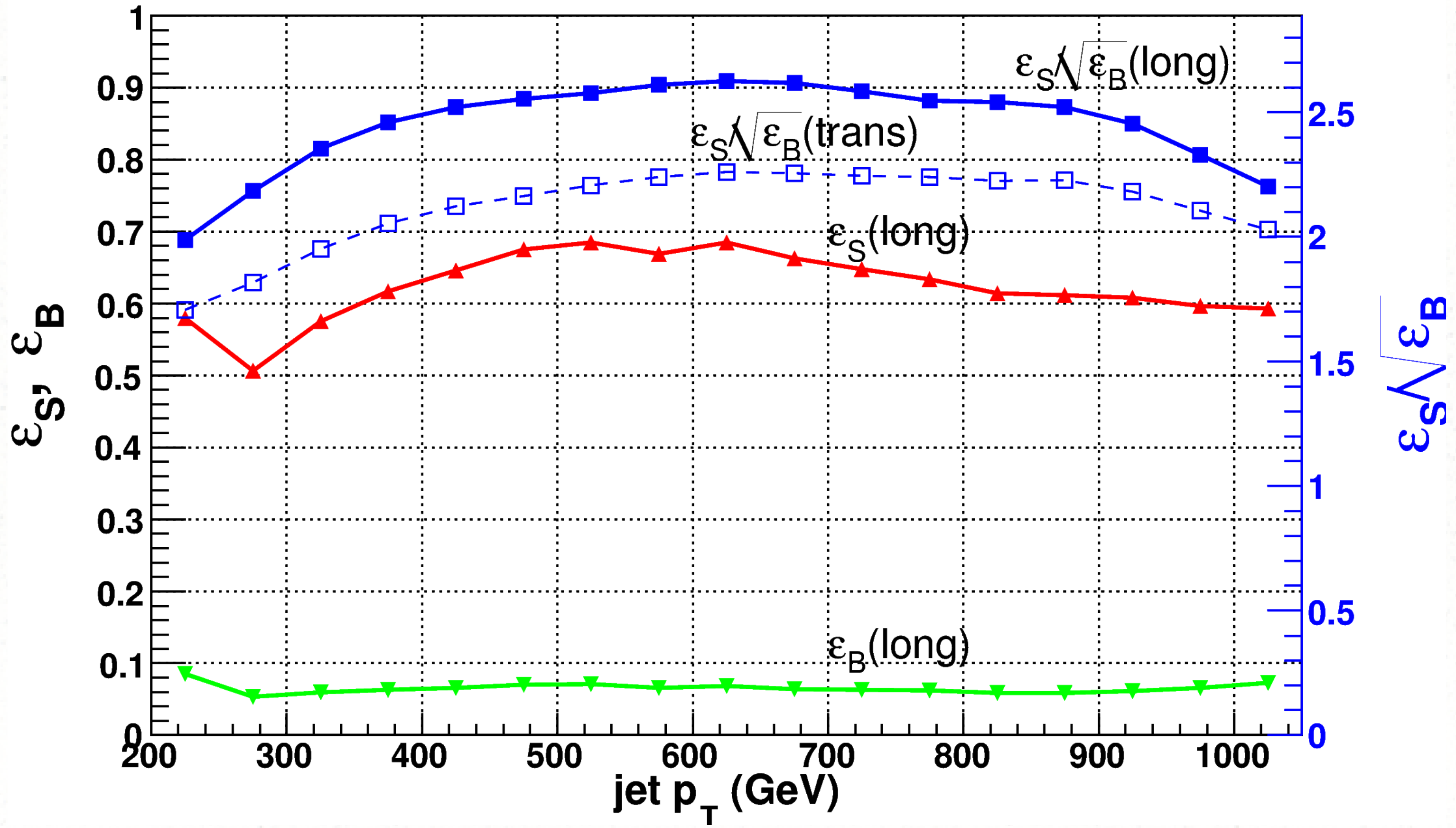
backup

Discussion-W polarization effect

- SM WW mostly transverse (92% for pt>200 GeV)
- Longitudinal W more balanced.



Better significance from filtering



MVA slightly better for longitudinal W