Implementation of e-ID based on BDT in Athena EgammaRec

Hai-Jun Yang
University of Michigan, Ann Arbor
(with T. Dai, X. Li, A. Wilson, B. Zhou)

ATLAS HSG3 Meeting
November 19, 2008

Motivation

- Lepton (e, μ, τ) Identification is crucial for new physics discoveries at the LHC, such as H→ ZZ→4 leptons, H→WW→ 2 leptons + MET etc.

- ATLAS default electron-ID (IsEM) has relatively low efficiency (~67%), which has significant impact on ATLAS early discovery potential in H→WW, ZZ detection with electron final states.

- It is important and also feasible to improve e-ID efficiency and to reduce jet fake rate by making full use of available variables using BDT.

Electron ID Studies with BDT

Select electrons in two steps

1) Pre-selection: an EM cluster matching a track

2) Apply electron ID based on pre-selected samples with different e-ID algorithms (IsEM, Likelihood ratio, AdaBoost and **EBoost**).

New BDT e-ID development at U. Michigan (Rel. v12)

- H. Yang's talk at US-ATLAS Jamboree on Sept. 10, 2008

  http://indico.cern.ch/conferenceDisplay.py?confId=38991

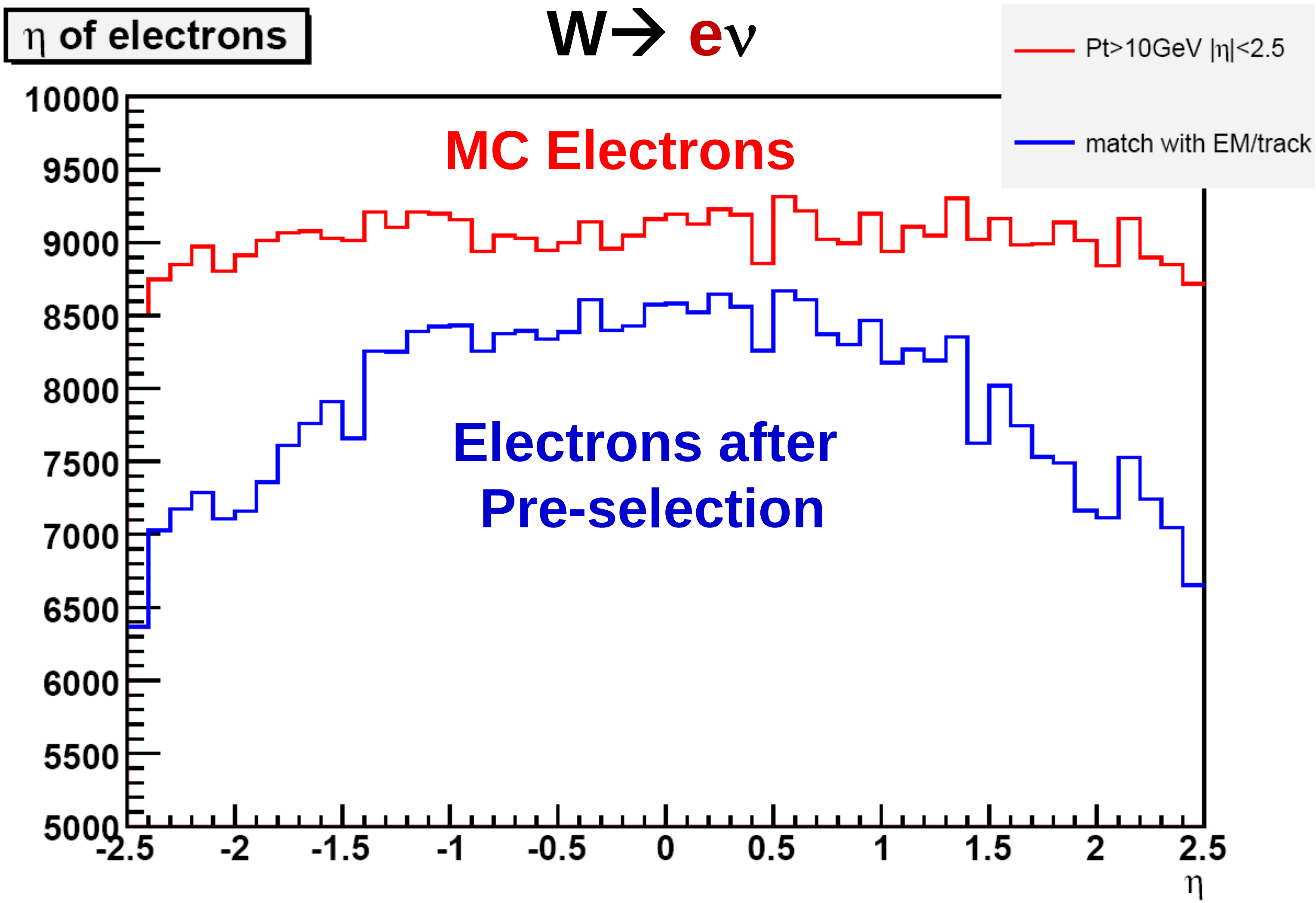
New BDT e-ID (EBoost) based on Rel. v13

- H. Yang's talk at ATLAS performance and physics workshop at CERN on Oct. 2, 2008

  http://indico.cern.ch/conferenceDisplay.py?confId=39296

Implementation of EBoost in EgammaRec (Rel. v14)

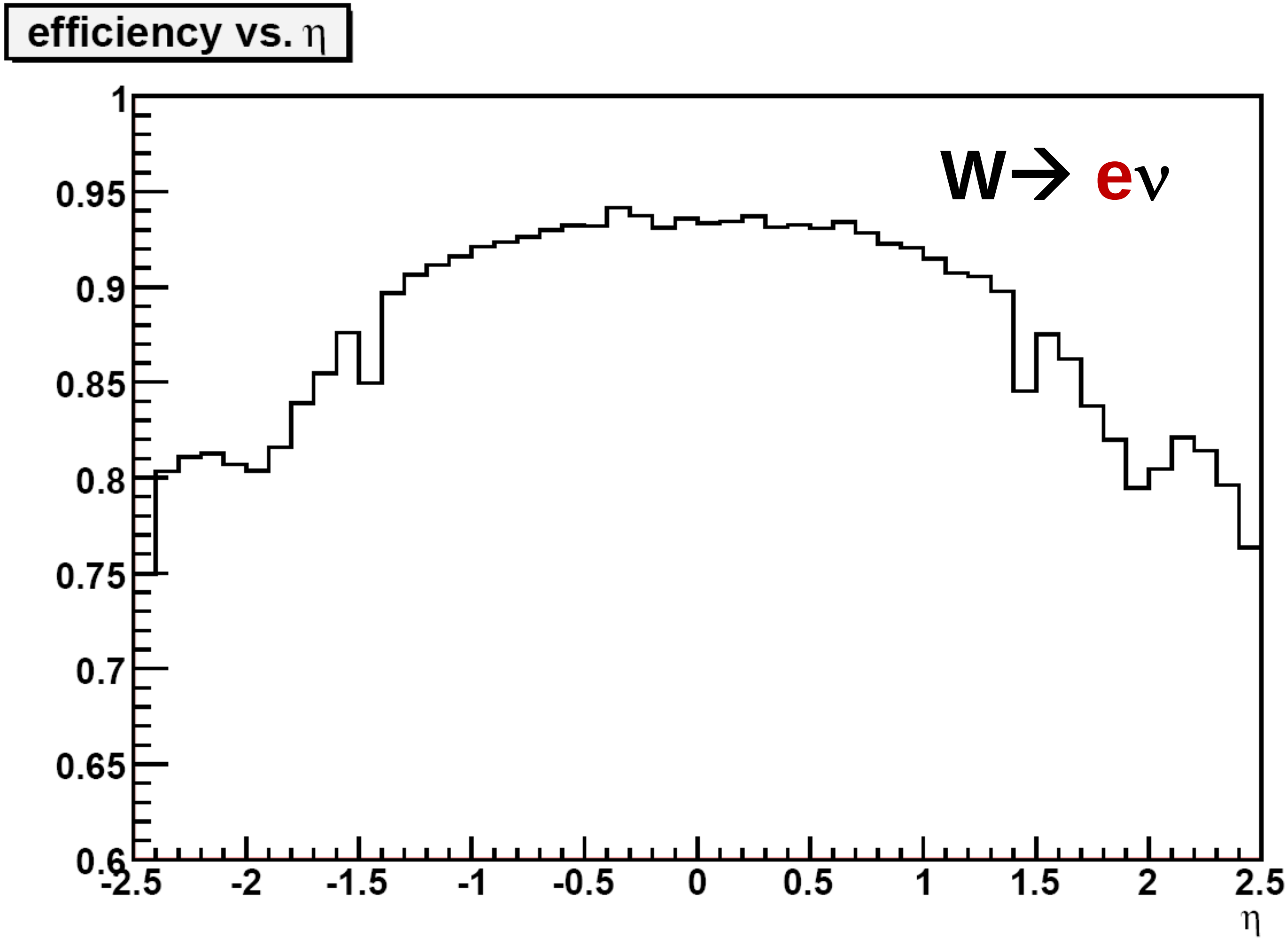
Electrons



η of electrons
W→ eν
Pt>10GeV |η|<2.5
match with EM/track
MC Electrons
Electrons after
Pre-selection
10000
9500
9000
8500
8000
7500
7000
6500
6000
5500
5000
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5
η

Electron Pre-selection Efficiency

The inefficiency mainly due to track matching



Variables Used for BDT e-ID (EBoost)

The same variables for IsEM are used

- egammaPID::ClusterHadronicLeakage

fraction of transverse energy in TileCal 1st sampling

- egammaPID::ClusterMiddleSampling

Ratio of energies in 3*7 & 7*7 window

Ratio of energies in 3*3 & 7*7 window

Shower width in LAr 2nd sampling

Energy in LAr 2nd sampling

- egammaPID::ClusterFirstSampling

Fraction of energy deposited in 1st sampling

Delta Emax2 in LAr 1st sampling

Emax2-Emin in LAr 1st sampling

Total shower width in LAr 1st sampling

Shower width in LAr 1st sampling

Fside in LAr 1st sampling

- egammaPID::TrackHitsA0

B-layer hits, Pixel-layer hits, Precision hits

Transverse impact parameter

- egammaPID::TrackTRT

Ratio of high threshold and all TRT hits

- egammaPID::TrackMatchAndEoP

Delta eta between Track and egamma

Delta phi between Track and egamma

E/P – egamma energy and Track momentum ratio

- Track Eta and EM Eta
- Electron isolation variables:

Number of tracks ($\Delta R=0.3$)

Sum of track momentum ($\Delta R=0.3$)

Ratio of energy in $\Delta R=0.2\text{-}0.45$ and $\Delta R=0.45$

BDT e-ID (EBoost) Training (v13)

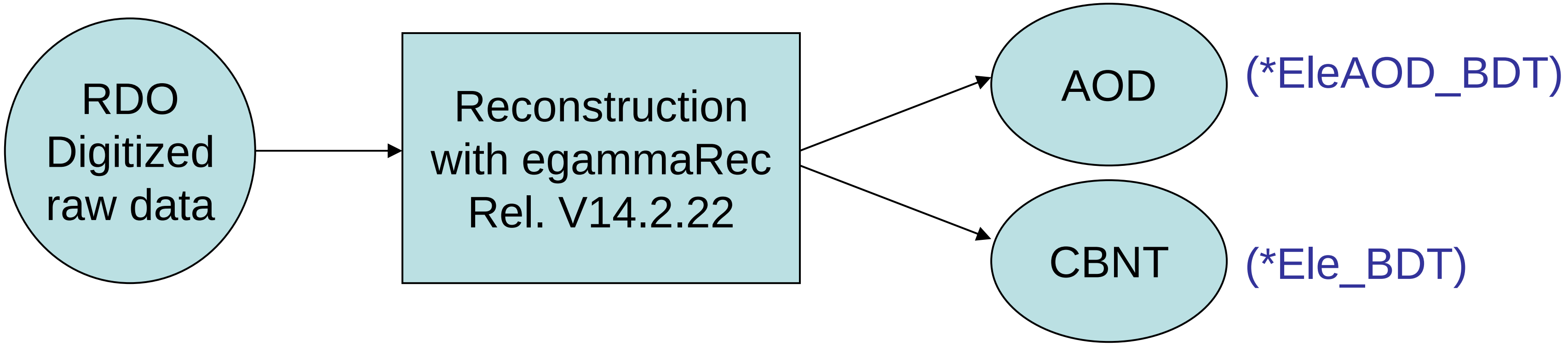
- BDT multivariate pattern recognition technique:
  - [H. Yang et. al., NIM A555 (2005) 370-385]

- BDT e-ID training signal and backgrounds (jet faked e)
  - W→eν as electron signal (DS 5104, v13)
  - Di-jet samples (J0-J6), Pt=[8-1120] GeV (DS 5009-5015, v13)

- BDT e-ID training procedure
  - Event weight training based on background cross sections [H. Yang et. al., JINST 3 P04004 (2008)]
  - Apply additional cuts on the training samples to select hardly identified jet faked electron as background for BDT training to make the BDT training more effective.
  - Apply additional event weight to high P_T backgrounds to effective reduce the jet fake rate at high P_T region.

Implementation of BDT Trees in EgammaRec Package and Test

- E-ID based on BDT has been implemented into egammaRec (04-02-98) package (private).
- We run through the whole reconstruction package based on v14.2.22 to test the BDT e-ID.

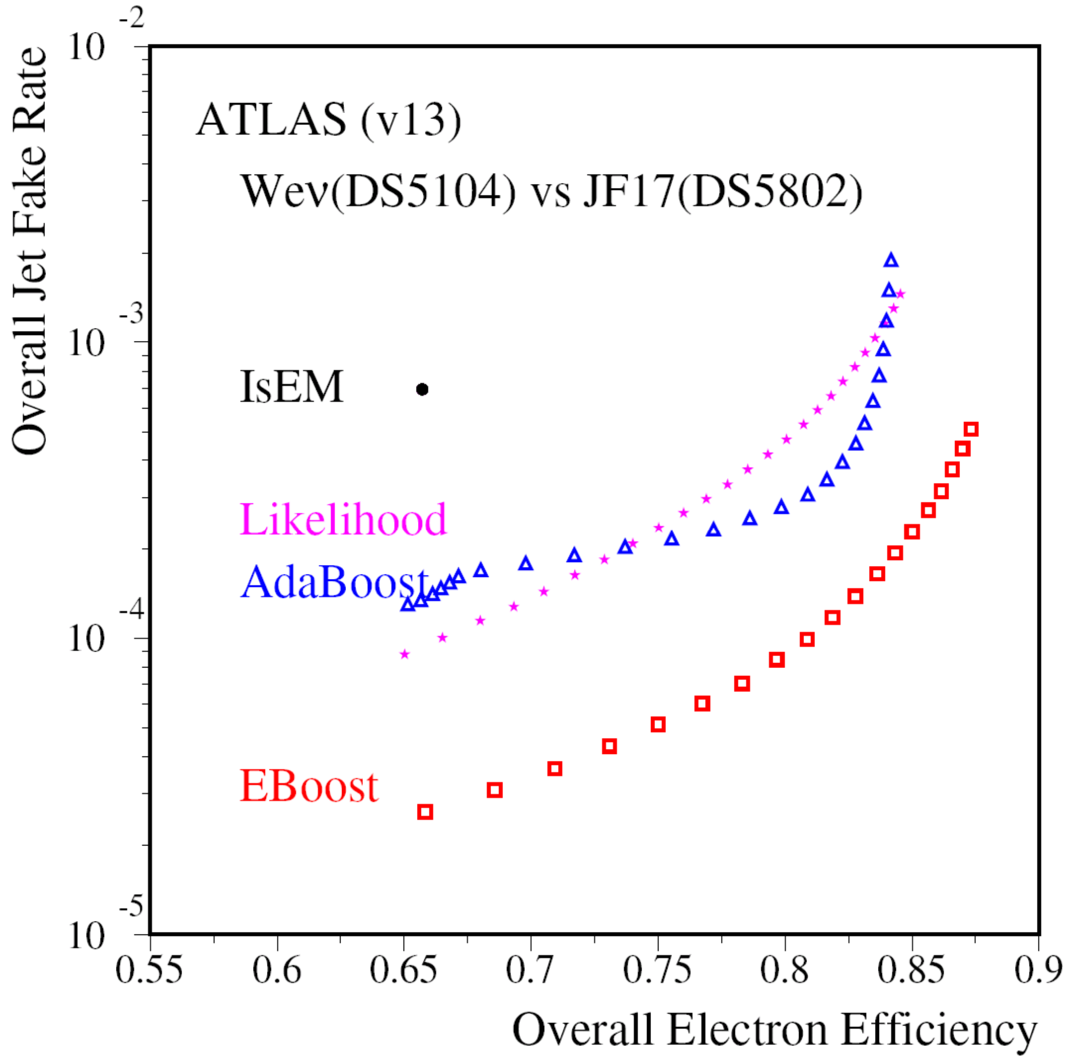


E-ID Testing Samples (v13)

- Wenu: DS5104 (Eff_precuts = 89.1%)
  - 46554 electrons with Et>10 GeV, |η|<2.5
  - 41457 electrons after pre-selection cuts

- JF17: DS5802 (Eff_precuts = 7.7%)
  - 14560093 jets with Et>10 GeV, |η|<2.5
  - 1123231 jets after pre-selection

Comparison of e-ID Algorithms (v13)



➔ IsEM (tight)
Eff = 65.7%
jet fake rate = 6.9E-4

➔ Likelihood Ratio (>6.5)
Eff = 78.5%
jet fake rate = 3.7E-4

➔ AdaBoost (>6)
Eff = 79.8%
jet fake rate = 2.8E-4

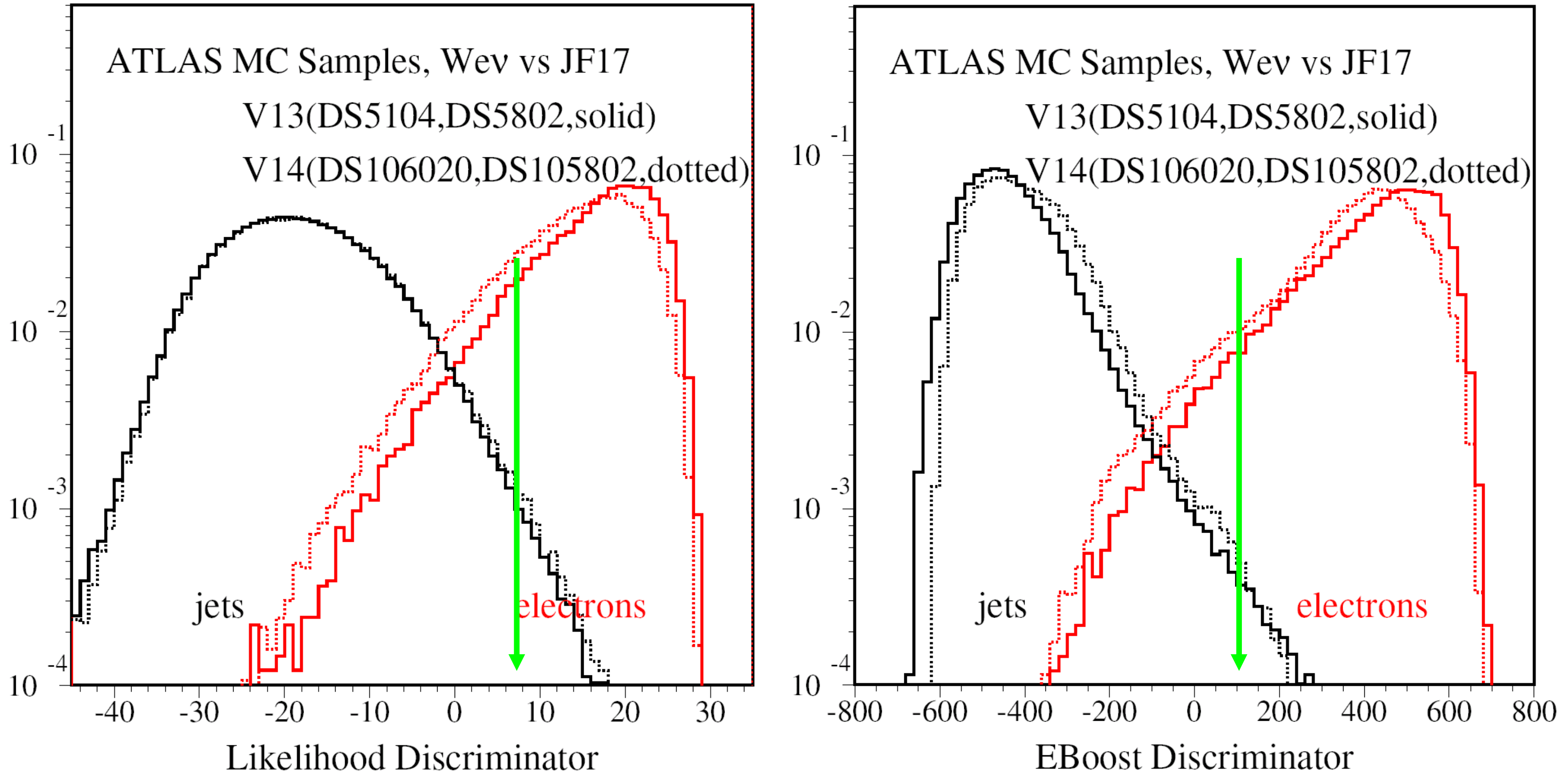
➔ EBoost (>100)
Eff = 84.3%
jet fake rate = 1.9E-4

E-ID Testing Samples (v14)

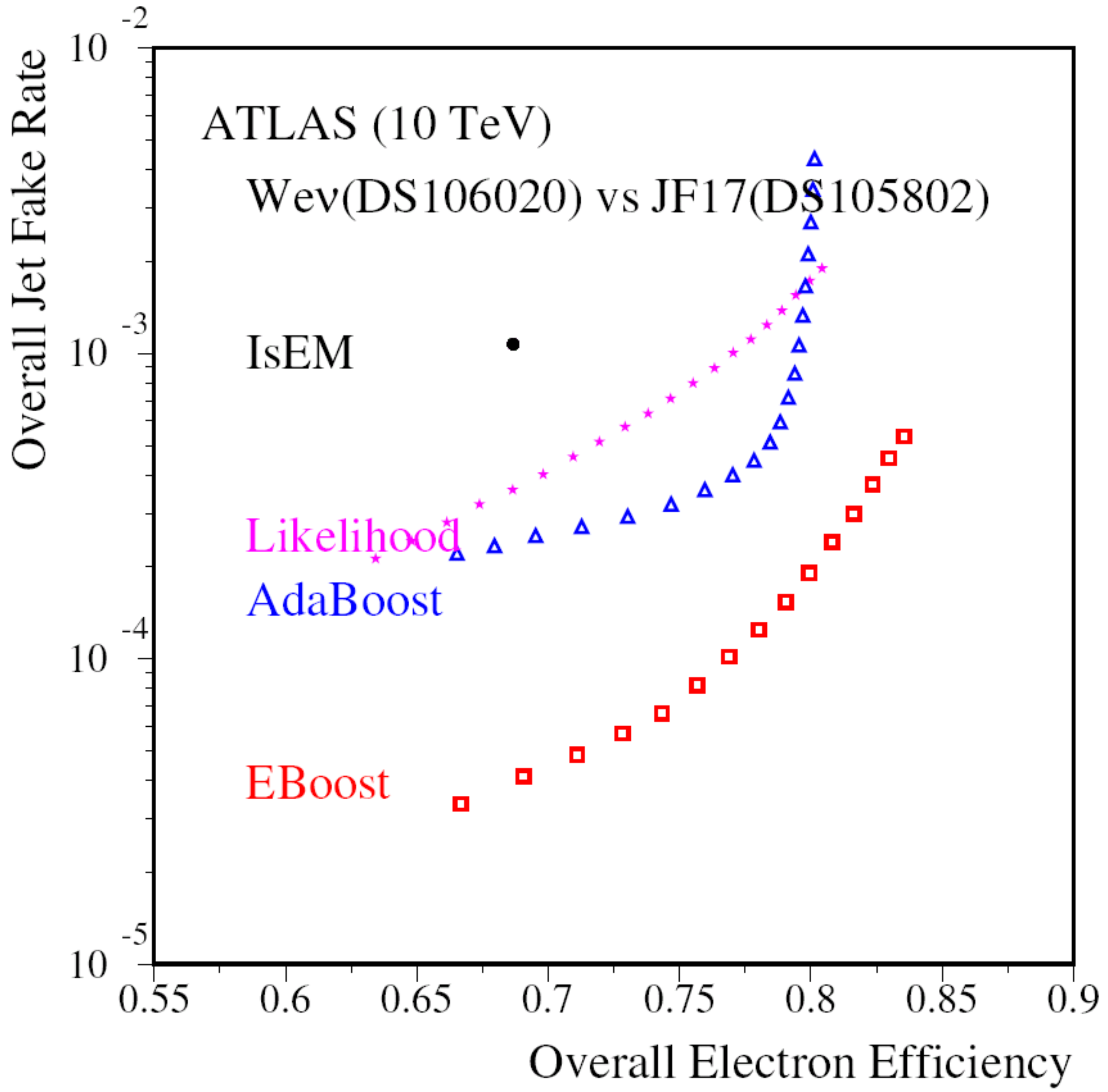
- Wenu: DS106020 (Eff_precuts = 86.9%)
  - 173930 events, 173822 electrons
  - 130589 electrons with Et>10GeV, $|\eta|<2.5$
  - 113500 electrons with pre-selection cuts

- JF17: DS105802 (Eff_precuts = 8%)
  - 475900 events, 1793636 jets
  - With pre-selection, 143167 jets

E-ID Discriminators (v13 vs v14)



Comparison of e-ID Algorithms (v14)

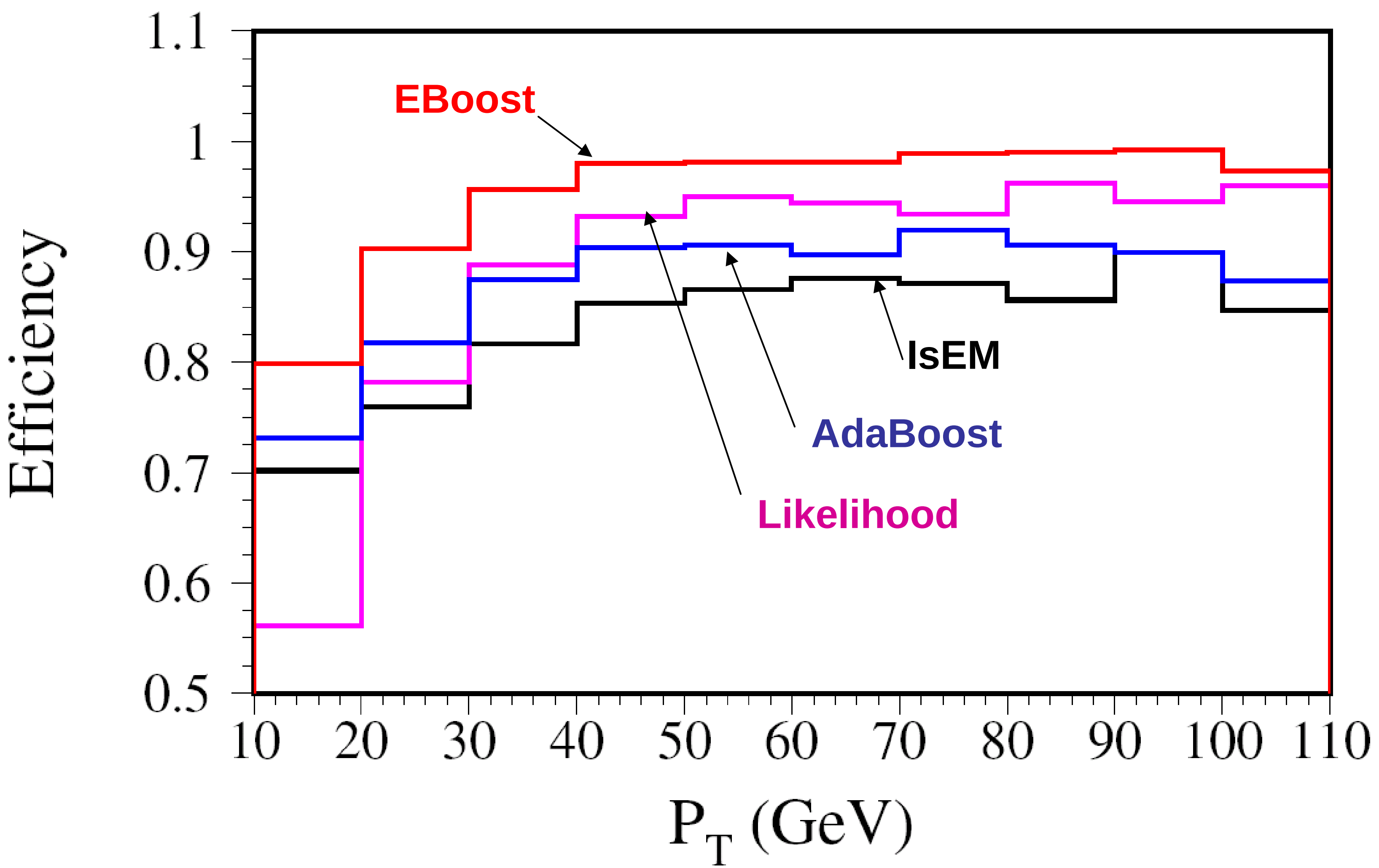


- ➔IsEM (tight)
  Eff = 68.7%
  jet fake rate = 1.1E-3
- ➔Likelihood Ratio (>6.5)
  Eff = 70.9%
  jet fake rate = 4.6E-4
- ➔AdaBoost (>6)
  Eff = 73%
  jet fake rate = 2.9E-4
- ➔EBoost (>100)
  Eff = 80%
  jet fake rate = 1.9E-4

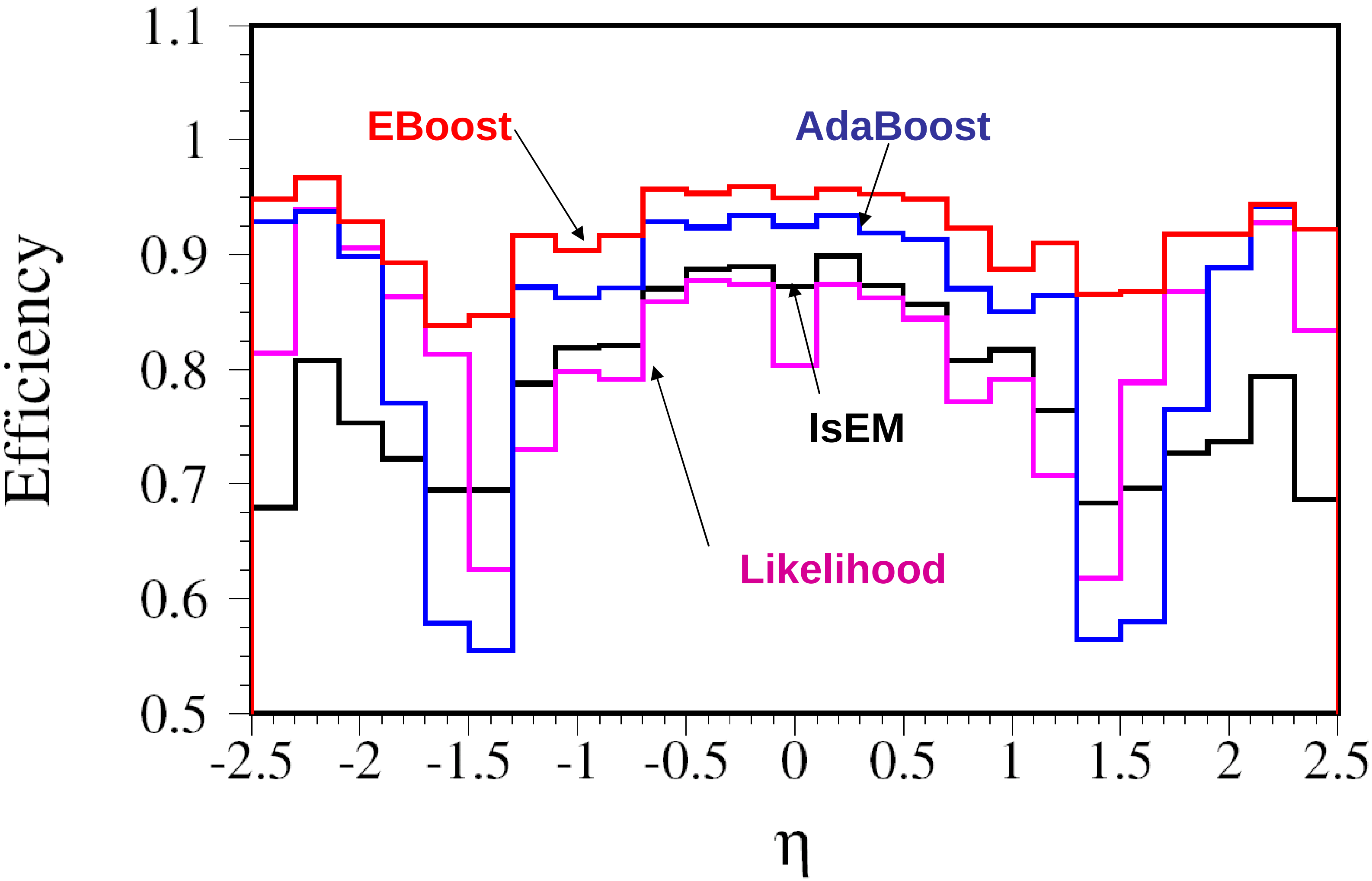
Overall E-ID Efficiency and Jet Fake Rates (v13 vs. v14)

Test MC	Precuts	IsEM(tight)	LH>6.5	AdaBoost > 6	EBoost > 100
W→eν (v13)	89.1%	65.7%	78.5%	79.8%	84.3%
W→eν (v14)	86.9%	68.7%	70.9%	73.0%	80.0%
JF17 (v13)	7.7E-2	6.9E-4	3.7E-4	2.8E-4	1.9E-4
JF17 (v14)	8.0E-2	11E-4	4.6E-4	2.9E-4	1.9E-4

E-ID Efficiency vs Pt (v14)



E-ID Efficiency vs η (v14)



Future Plan

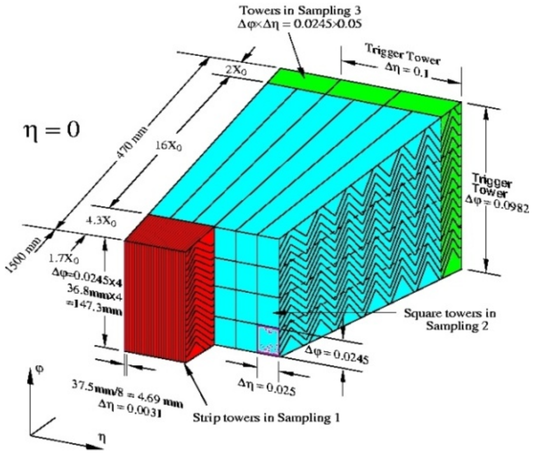
- We have requested to add EBoost in ATLAS official egammaRec package and make EBoost discriminator variable available for physics analysis.
- We will provide EBoost trees to ATLAS egammaRec for each major software release
- Explore new variables and BDT training techniques to further improve the e-ID performance

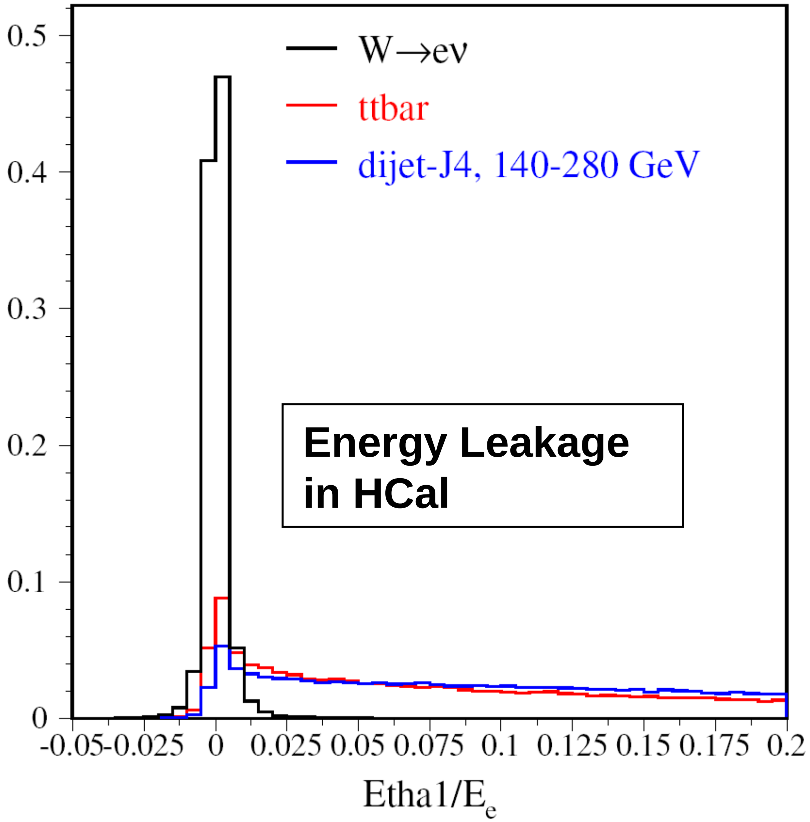
Backup Slides

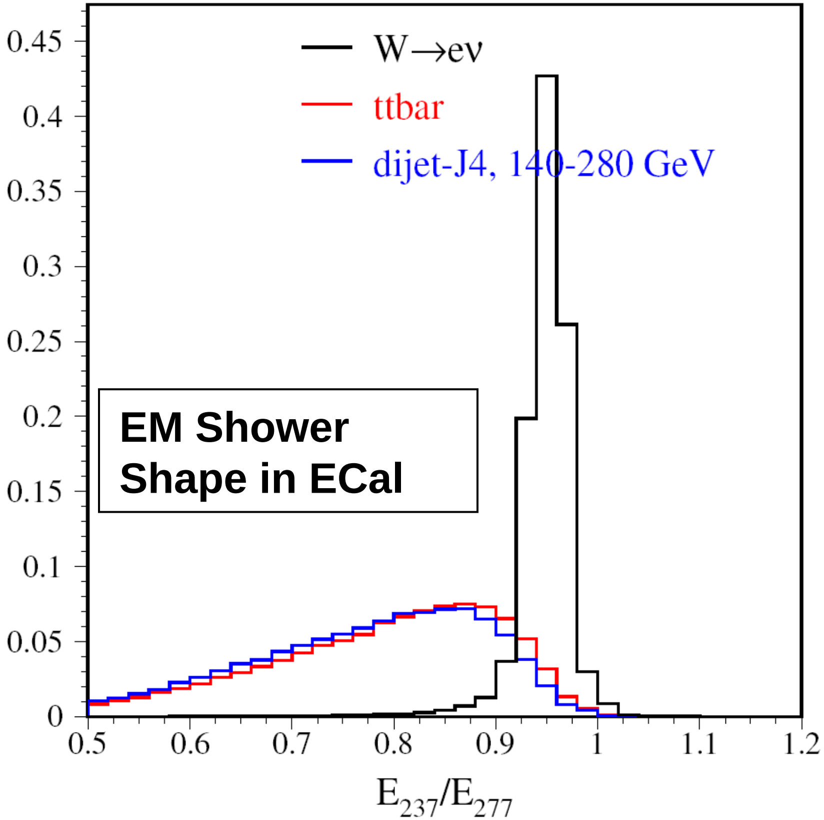
List of Variables for BDT

1. Ratio of Et(ΔR=0.2-0.45) / Et(ΔR=0.2)
2. Number of tracks in ΔR=0.3 cone
3. Energy leakage to hadronic calorimeter
4. EM shower shape E237 / E277
5. $\Delta\eta$ between inner track and EM cluster
6. Ratio of high threshold and all TRT hits
7. Number of pixel hits and SCT hits
8. $\Delta\phi$ between track and EM cluster
9. Emax2 – Emin in LAr 1st sampling
10. Number of B layer hits
11. Number of TRT hits
12. Emax2 in LAr 1st sampling
13. EoverP – ratio of EM energy and track momentum
14. Number of pixel hits
15. Fraction of energy deposited in LAr 1st sampling
16. Et in LAr 2nd sampling
17. η of EM cluster
18. D0 – transverse impact parameter
19. EM shower shape E233 / E277
20. Shower width in LAr 2nd sampling
21. Fracs1 – ratio of (E7strips-E3strips)/E7strips in LAr 1st sampling
22. Sum of track Pt in DR=0.3 cone
23. Total shower width in LAr 1st sampling
24. Shower width in LAr 1st sampling

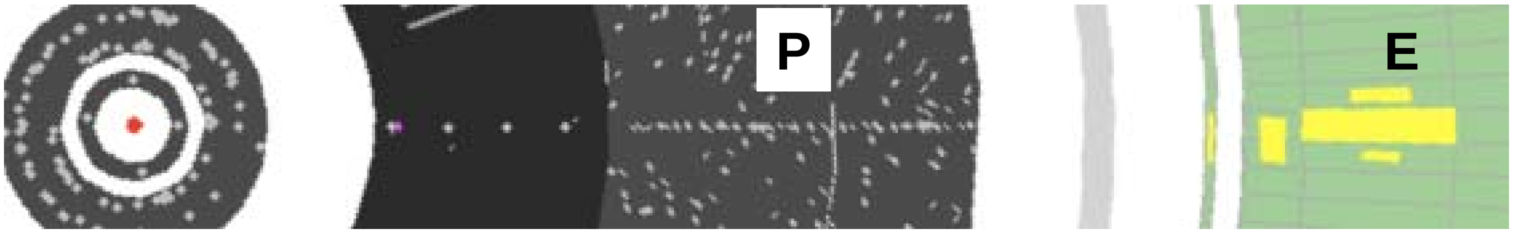
EM Shower shape distributions of discriminating Variables (signal vs. background)

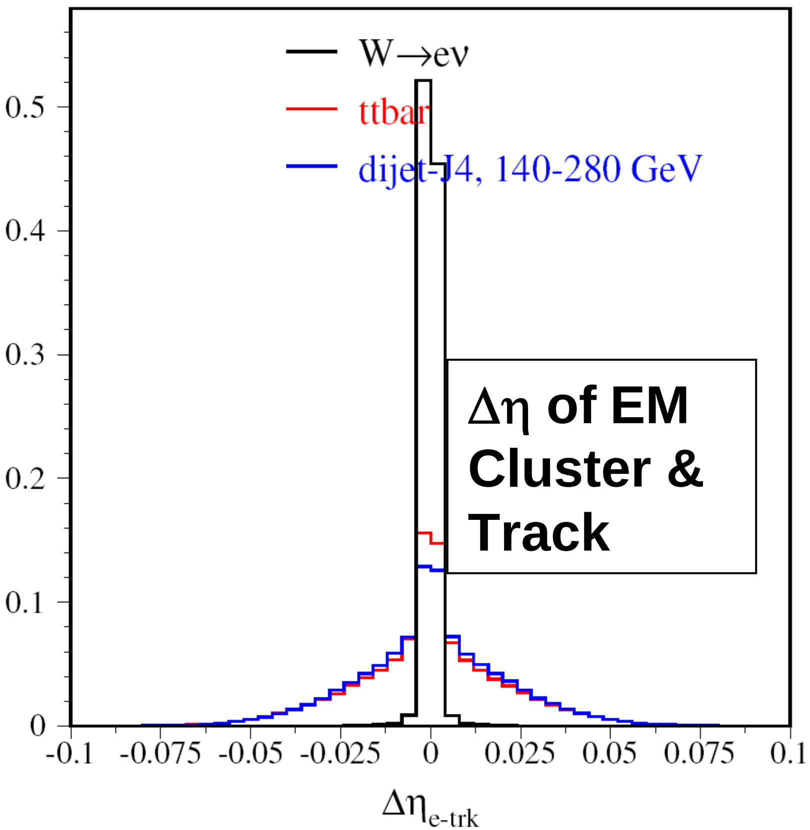


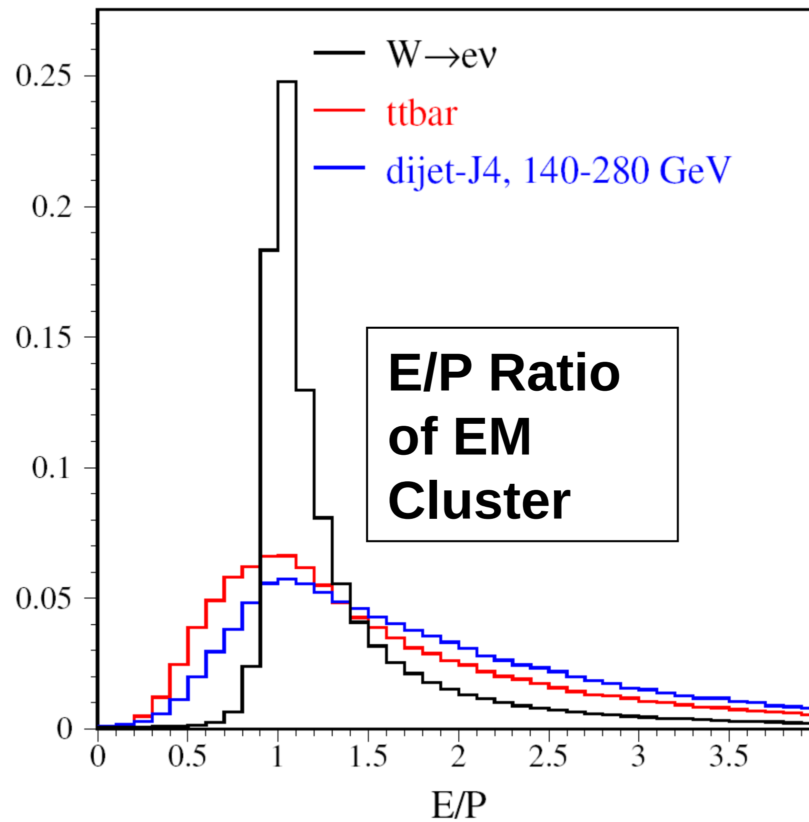




ECal and Inner Track Match



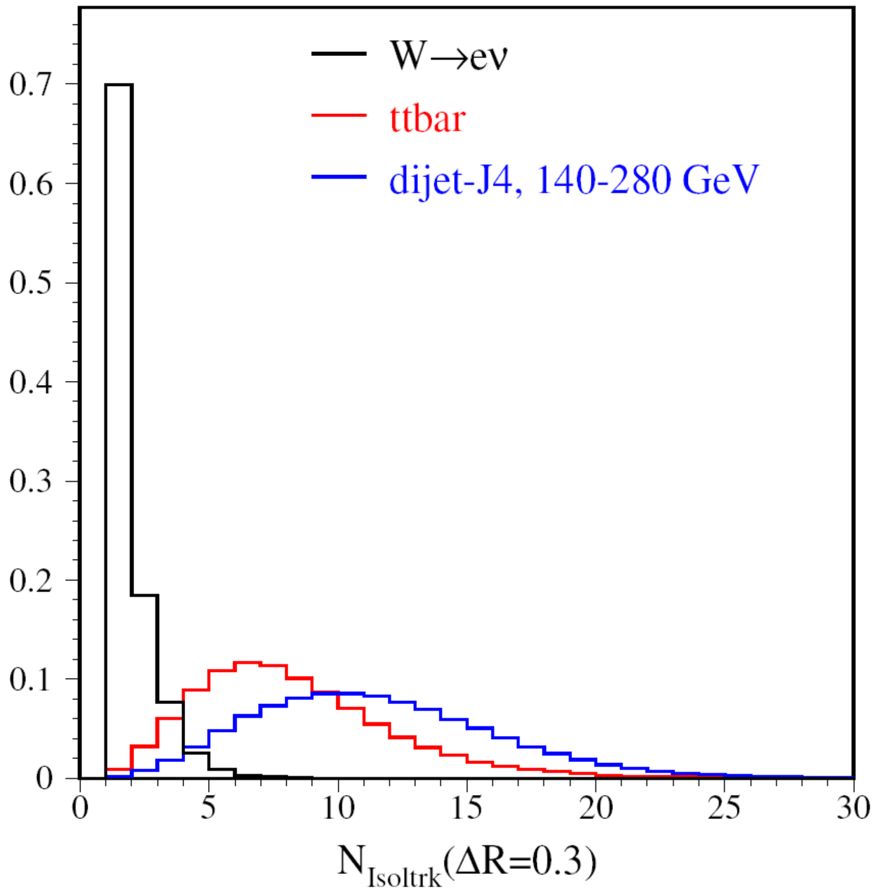


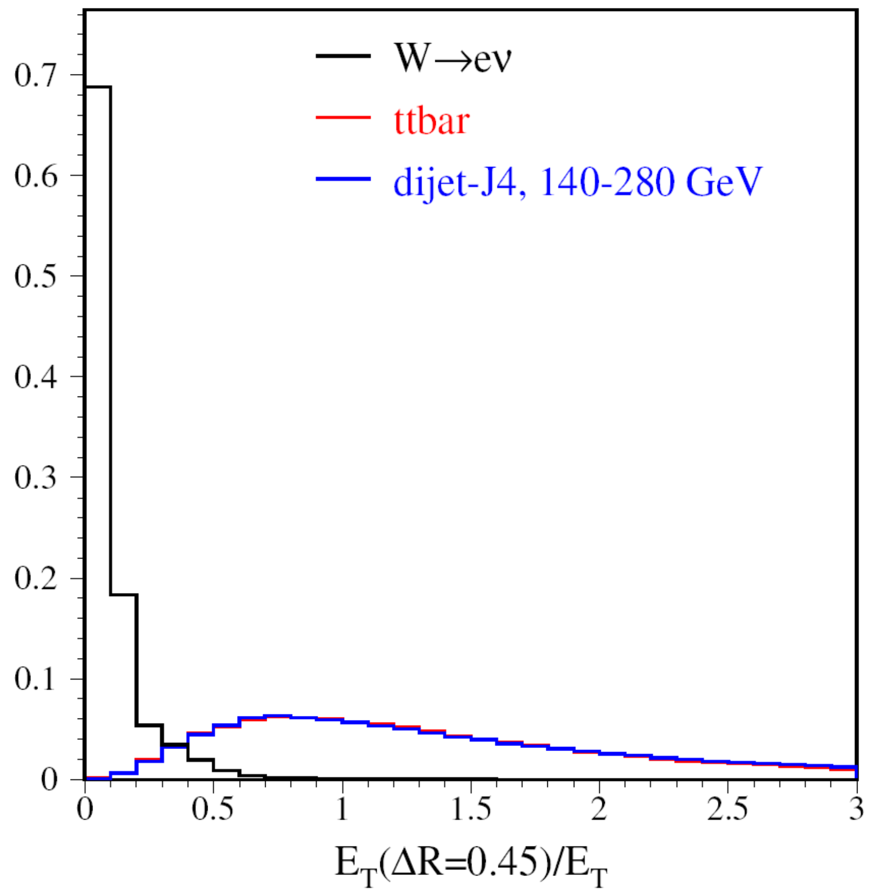


Electron Isolation Variables

N_{trk} around Electron Track

$E_T(\Delta R=0.2\text{-}0.45)/E_T$ of EM





Example: H→ WW →lνlν Studies

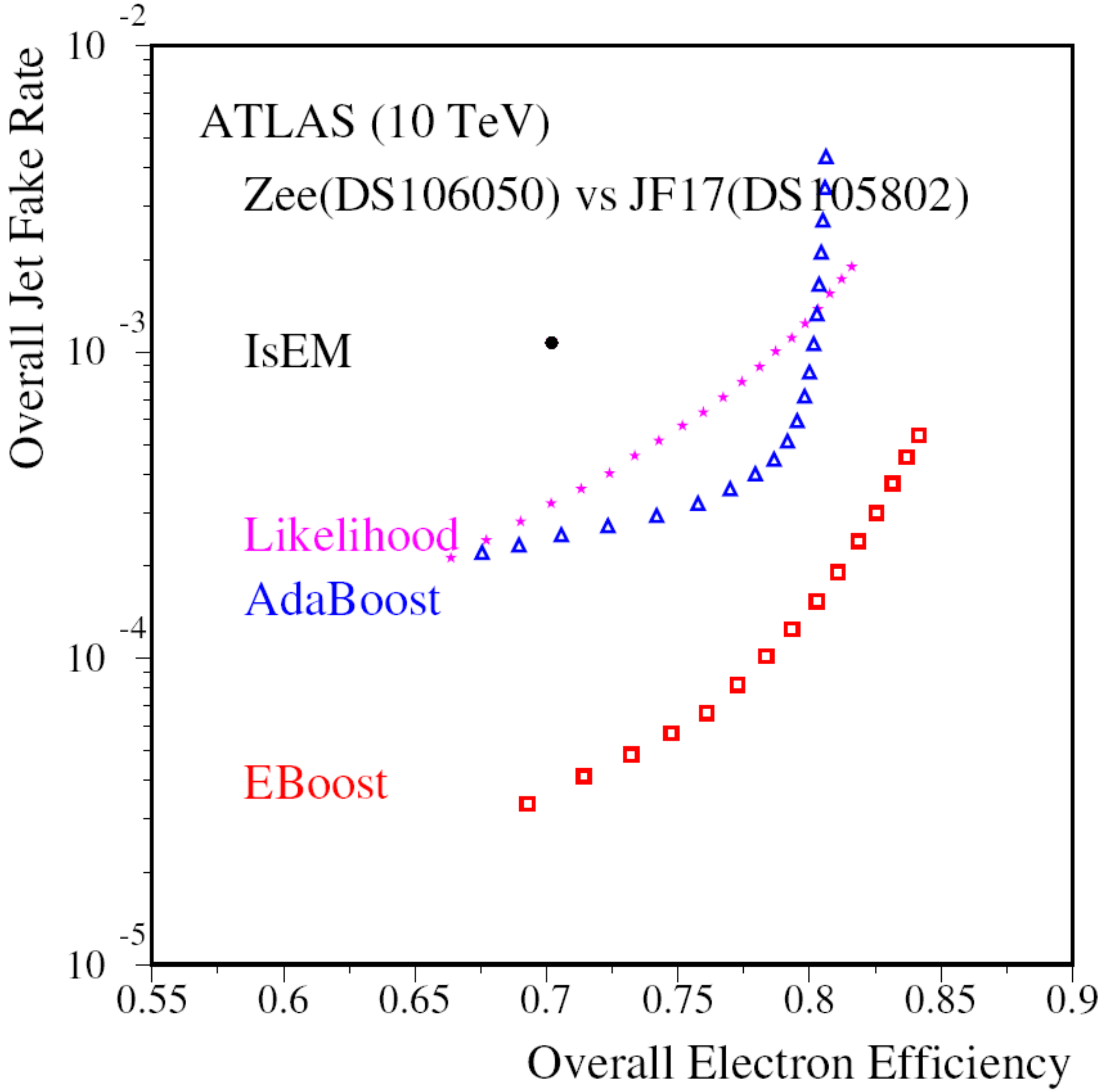
[H. Yang et.al., ATL-COM-PHYS-2008-023]

- At least one lepton pair (ee, μμ, eμ) with P_T > 10 GeV, |η|<2.5
- Missing E_T > 20 GeV, max(P_T (l) ,P_T(l)) > 25 GeV
- $|M_{ee} - M_z|$ > 10 GeV, $|M_{\mu\mu} - M_z|$ > 15 GeV to suppress background from Z → ee, μμ

Higgs Mass (GeV)	Eff($e\nu e\nu$)	Eff($\mu\nu\mu\nu$)	Eff($e\nu\mu\nu$)
140	26.3%	49.9%	34.2%
150	28.5%	51.1%	37.0%
160	29.9%	53.3%	39.9%
165	30.5%	54.1%	40.8%
170	30.5%	52.7%	42.2%
180	29.3%	50.1%	43.2%

Used ATLAS electron ID: IsEM & 0x7FF == 0

Comparison of e-ID Algorithms (v14)



➔ IsEM (tight)
Eff = 70.2%
jet fake rate = 1.1E-3

➔ Likelihood Ratio (>6.5)
Eff = 73.4%
jet fake rate = 4.6E-4

➔ AdaBoost (>6)
Eff = 74.2%
jet fake rate = 2.9E-4

➔ EBoost (>100)
Eff = 81.1%
jet fake rate = 1.9E-4

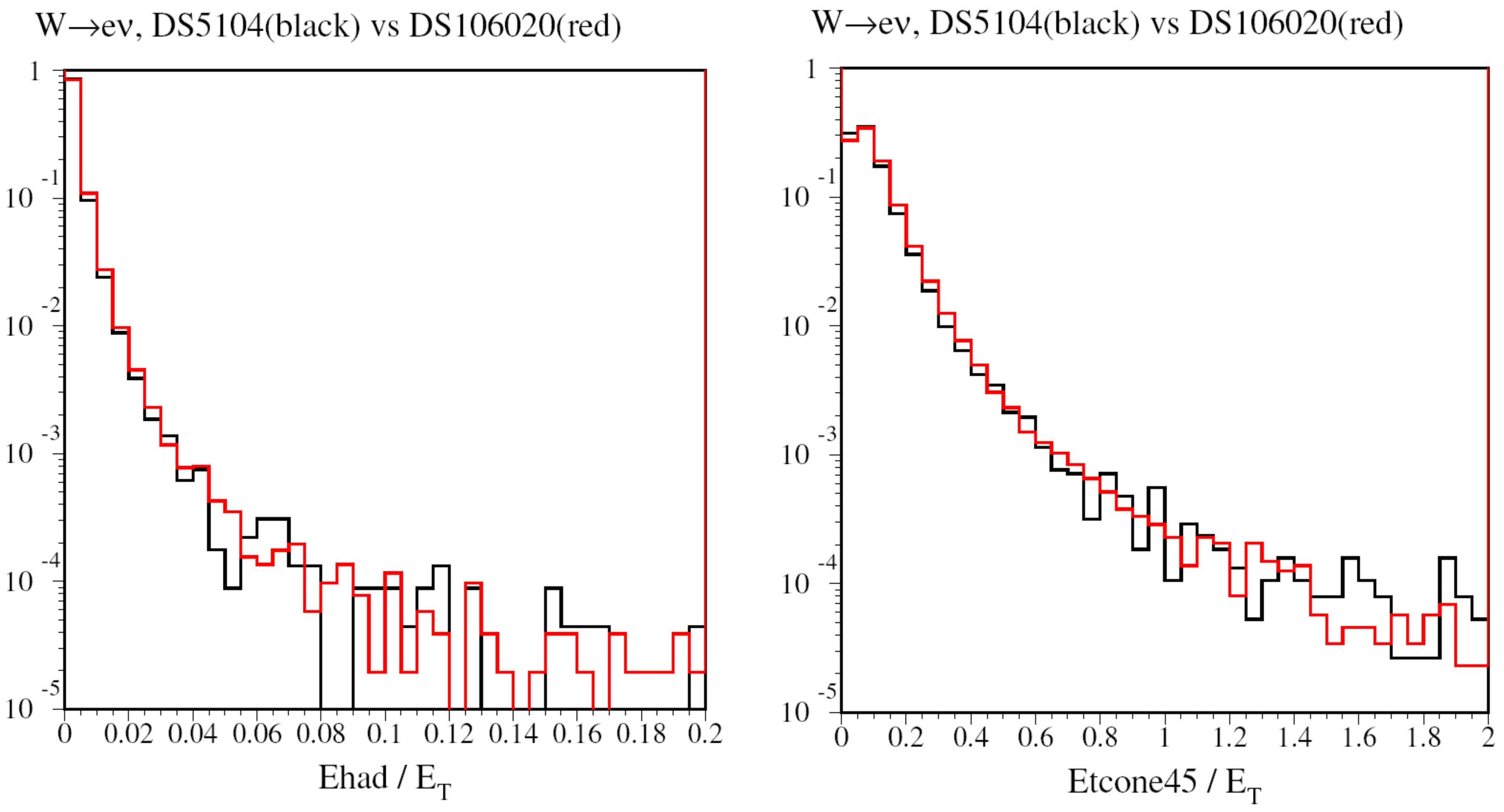
Signal Pre-selection: MC electrons

- MC True electron from $W \rightarrow e\nu$ by requiring
  - $|\eta_e| < 2.5$ and $E_T^{true} > 10$ GeV (N_e)
- Match MC e/γ to EM cluster:
  - $\Delta R < 0.2$ and $0.5 < E_T^{rec} / E_T^{true} < 1.5$ (N_{EM})
- Match EM cluster with an inner track:
  - eg_trkmatchnt > -1 ($N_{EM/track}$)
- Pre-selection Efficiency = $N_{EM/Track} / N_e$

Pre-selection of Jet Faked Electrons

- Count number of jets with
  - $|\eta_{jet}|$ < 2.5, E_T^{jet} >10 GeV (N_{jet})
- Loop over all EM clusters; each cluster matches with a jet
  - E_T^{EM} > 10 GeV (N_{EM})
- Match EM cluster with an inner track:
  - eg_trkmatchnt > -1 ($N_{EM/track}$)
- Pre-selection Acceptance = $N_{EM/Track}$ / N_{jet}

Comparisons of v13 and v14



Comparisons of v13 and v14