Performance of BDTs for Electron Identification

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

ATLAS Egamma Meeting
December 17, 2008

Motivation

- Lepton (e, μ, τ) Identification with high efficiency is crucial for new physics discoveries at the LHC
- Great efforts in ATLAS to develop the algorithms for electron identification:
  - Cut-based algorithm: IsEM
  - Multivariate algorithms: Likelihood and BDT
- Further improvement could be achieved with better treatment of the multivariate training using the Boosted Decision Trees technique

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

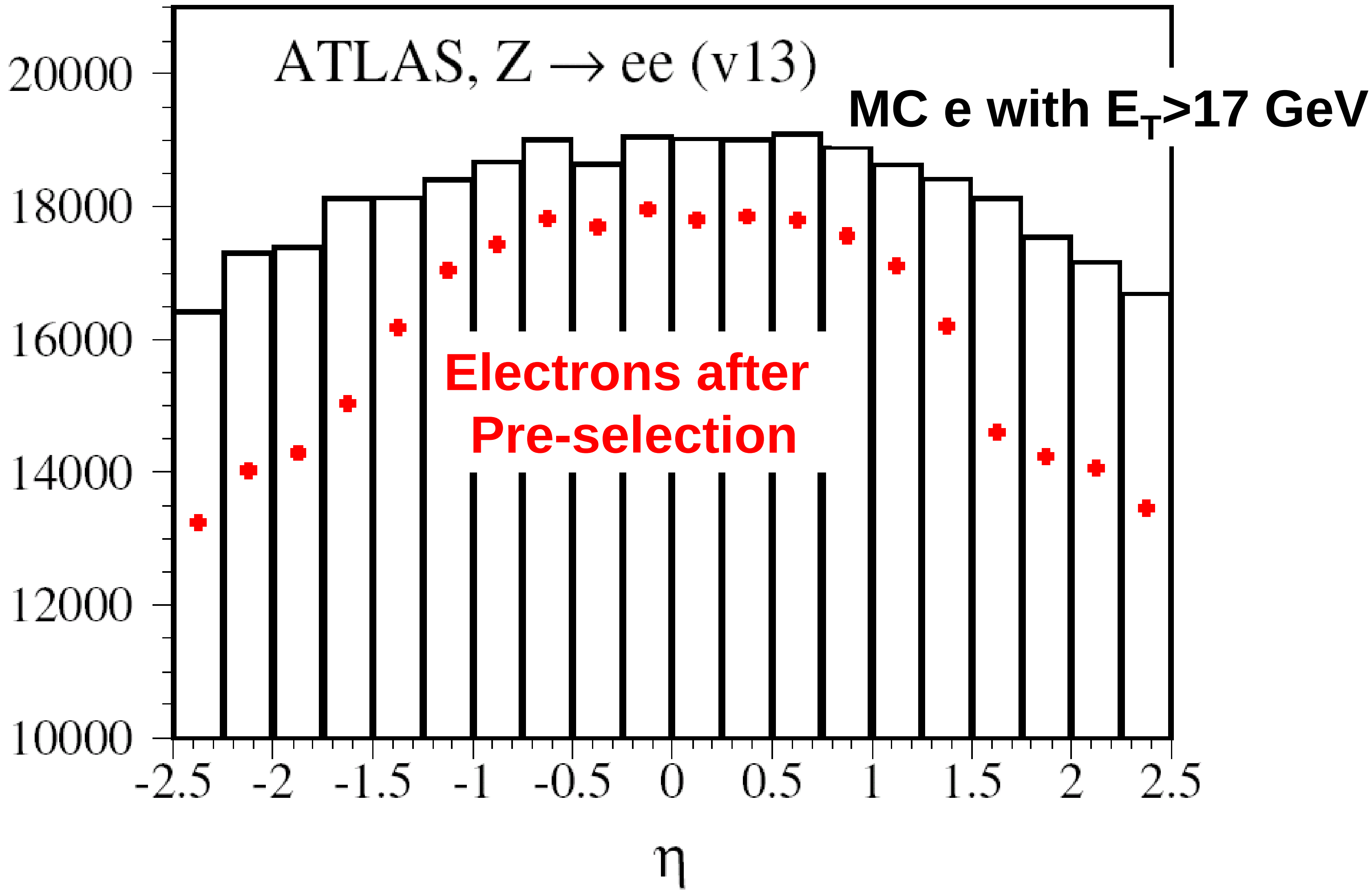
New BDT e-ID (EBoost) based on Rel. v13

- 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)

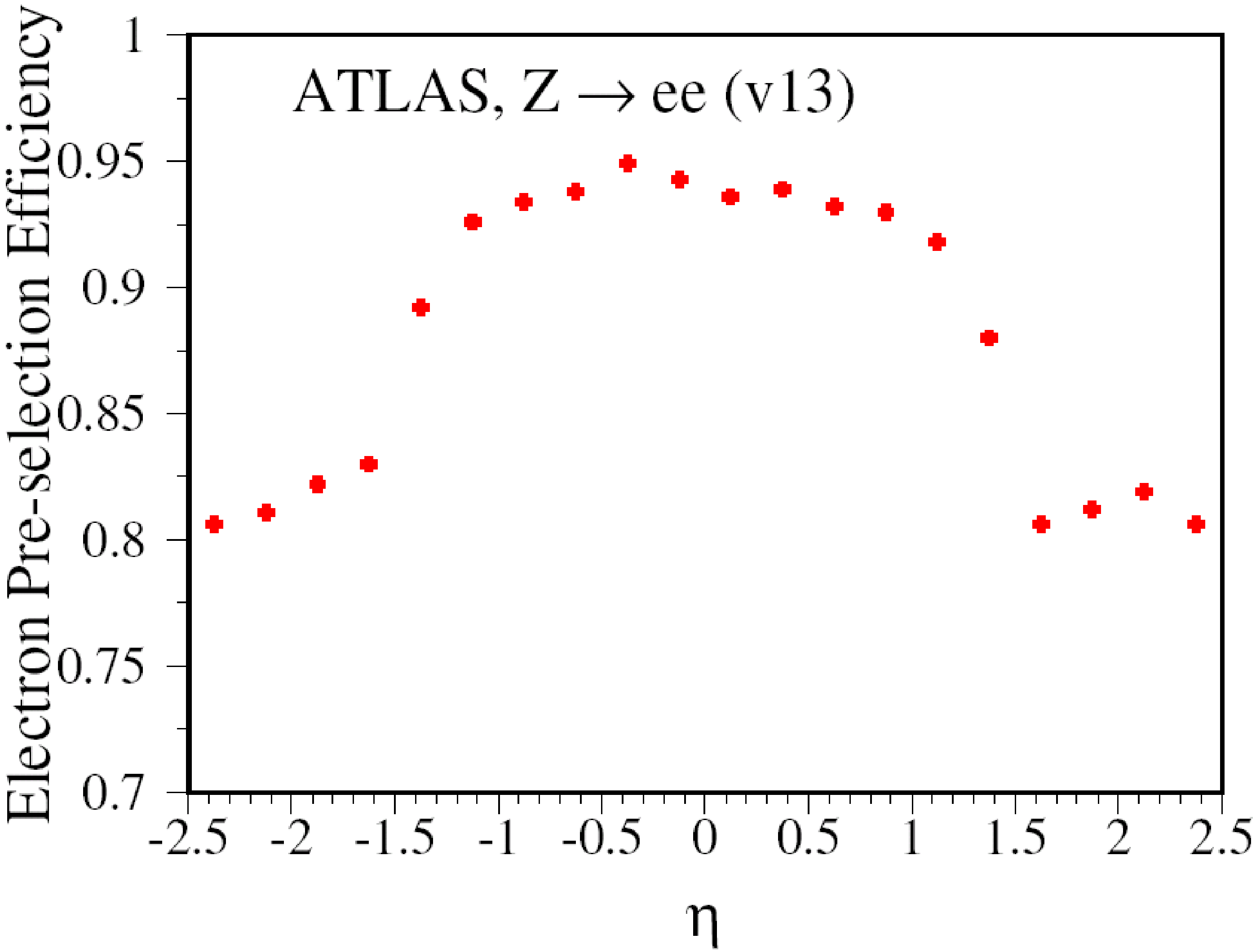
Electrons



ATLAS, Z → ee (v13)
MC e with E_T>17 GeV
Electrons after Pre-selection
20000
18000
16000
14000
12000
10000
-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



BDT e-ID (EBoost) Training

- BDT multivariate pattern recognition technique:
  - [H. Yang et. al., NIM A555 (2005) 370-385]
- BDT e-ID training signal and backgrounds (jet faked e)
  - $W \rightarrow e\nu$ as electron signal (DS 5104, v13)
  - JF17 (DS 5802, v13)
- Using the same e-ID variables as IsEM for training (see variable list in next page)

- BDT e-ID training procedure
  - 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 event weight to high P_T backgrounds to effective reduce the jet fake rate at high P_T region. Event weight training technique reference, [H. Yang et. al., JINST 3 P04004 (2008)]

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:

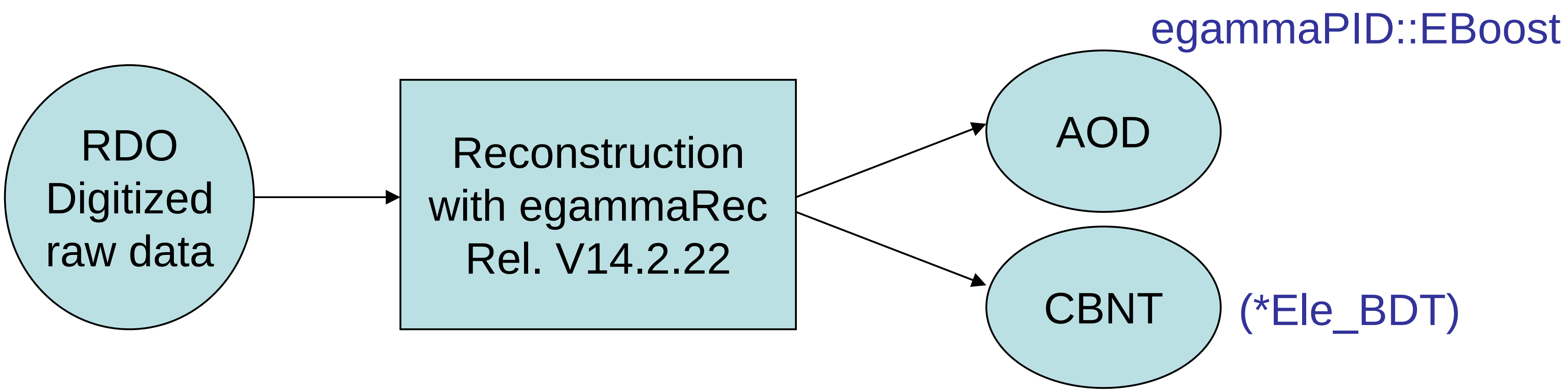
Number of tracks ($\Delta R=0.3$)

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

Ratio of energy in EtCone45 / E_T

Implementation of BDT Trees in EgammaRec Package and Test

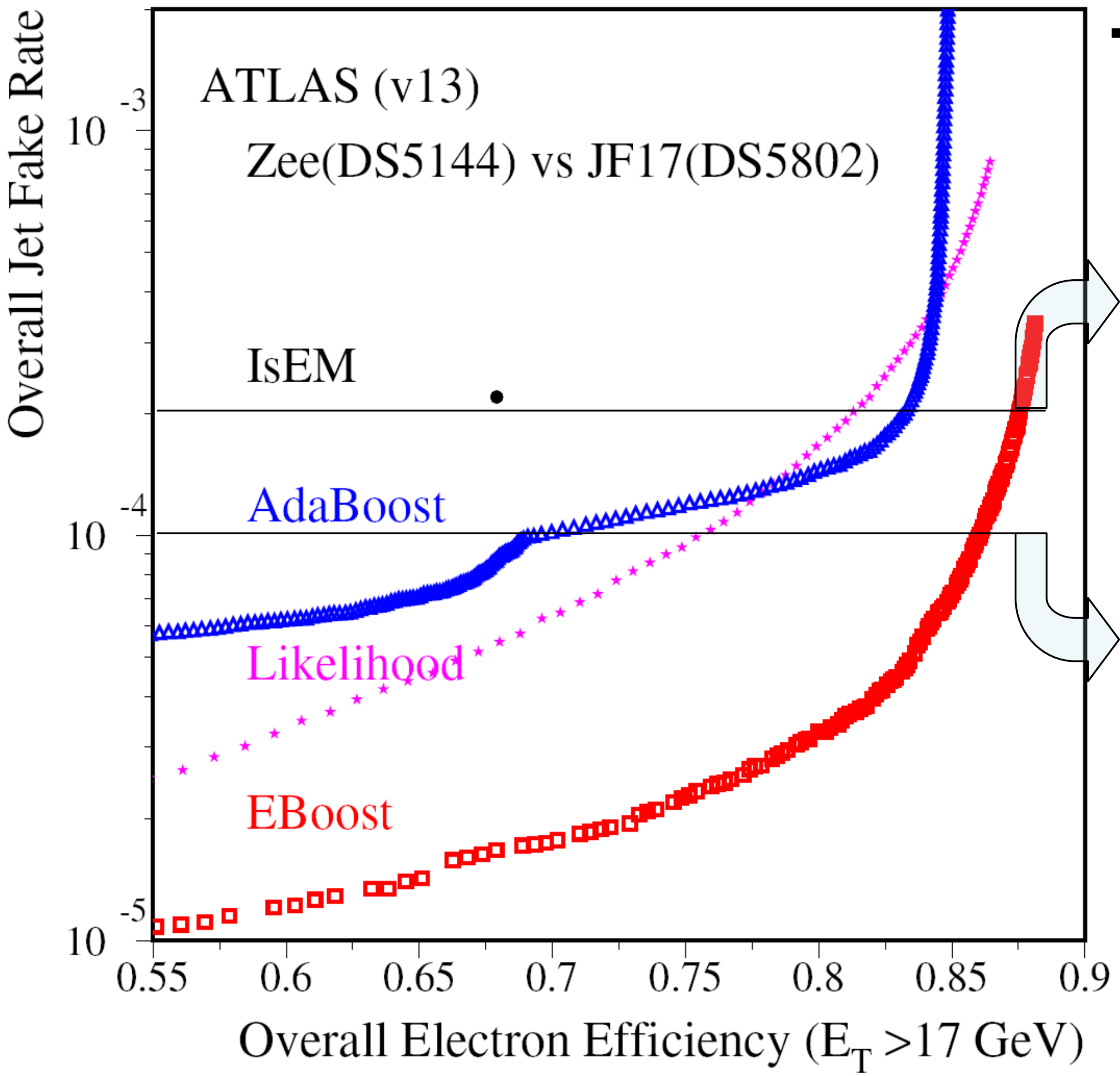
- E-ID based on BDT has been implemented into egammaRec (04-02-98) package (private).
- We run through the whole reconstruction package based on v14.2.22 to test the BDT e-ID.



E-ID Testing Samples Produced at $\sqrt{s}$ = 14 TeV (v13)

- Wenu: DS5104 (Eff_precuts = 88.6%)
  - 42020 electrons with Et>17 GeV, $|\eta|<2.5$
  - 37230 electrons after pre-selection cuts
- Zee: DS5144 (Eff_precuts = 88.6%)
  - 181281 electron with Et>17 GeV, $|\eta|<2.5$
  - 160615 electrons after pre-selection cuts
- JF17: DS5802 (Eff_precuts = 2.4%)
  - 1946968 events, 7280046 reconstructed jets
  - 176727 jets after pre-selection

Comparison of e-ID Algorithms (v13)



➔IsEM (tight)
Eff = 67.9%
jet fake rate = 2.2E-4

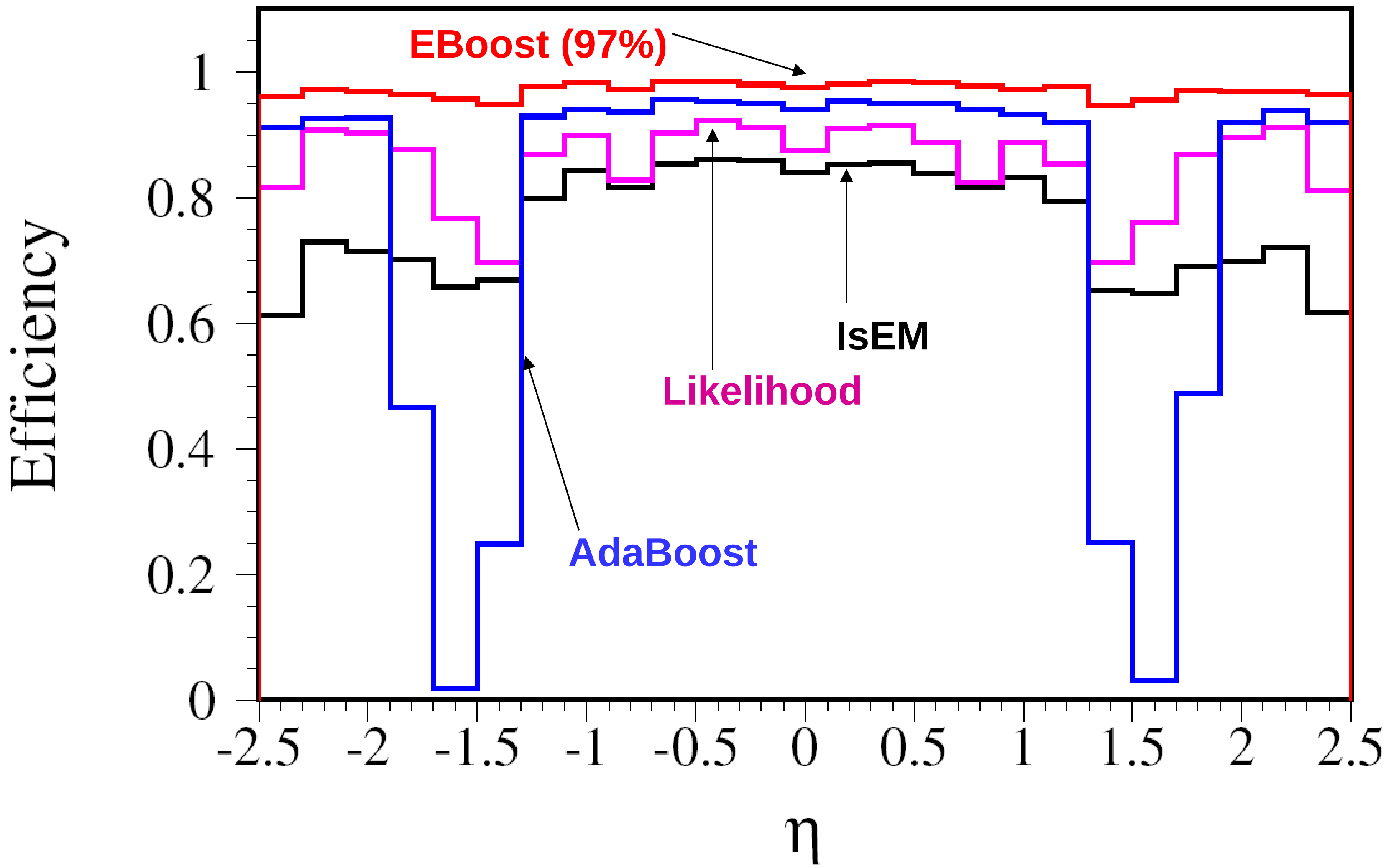
➔jet fake rate = 2.0E-4

Likelihood: Eff = 81.3%
AdaBoost: Eff = 83.2%
EBoost: Eff = 87.5%

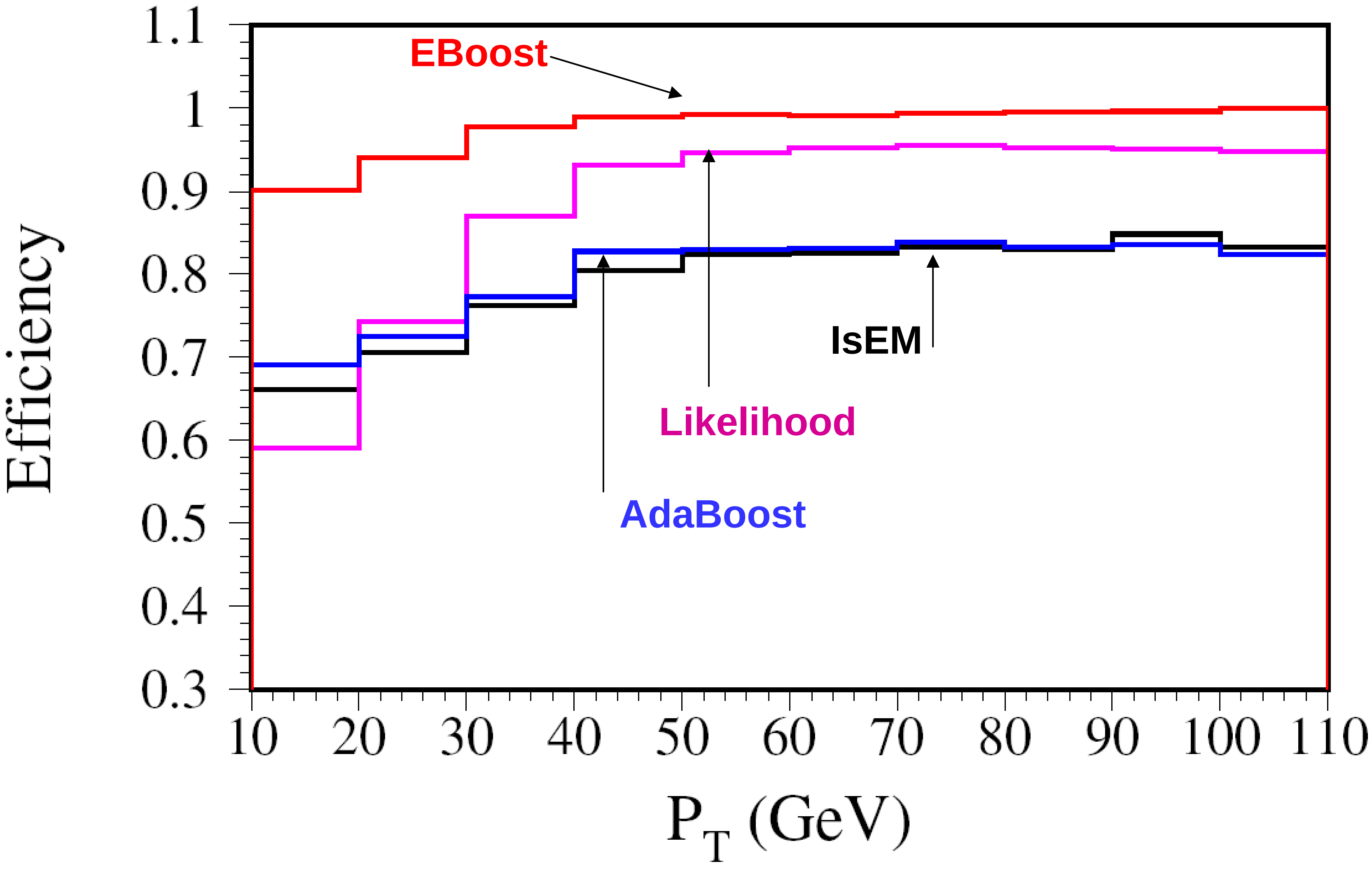
➔jet fake rate = 1.0E-4

Likelihood: Eff = 75.9%
AdaBoost: Eff = 69.3%
EBoost: Eff = 86.0%

E-ID Efficiency after pre-selection vs η (v13, jet fake rate=1.0E-4)



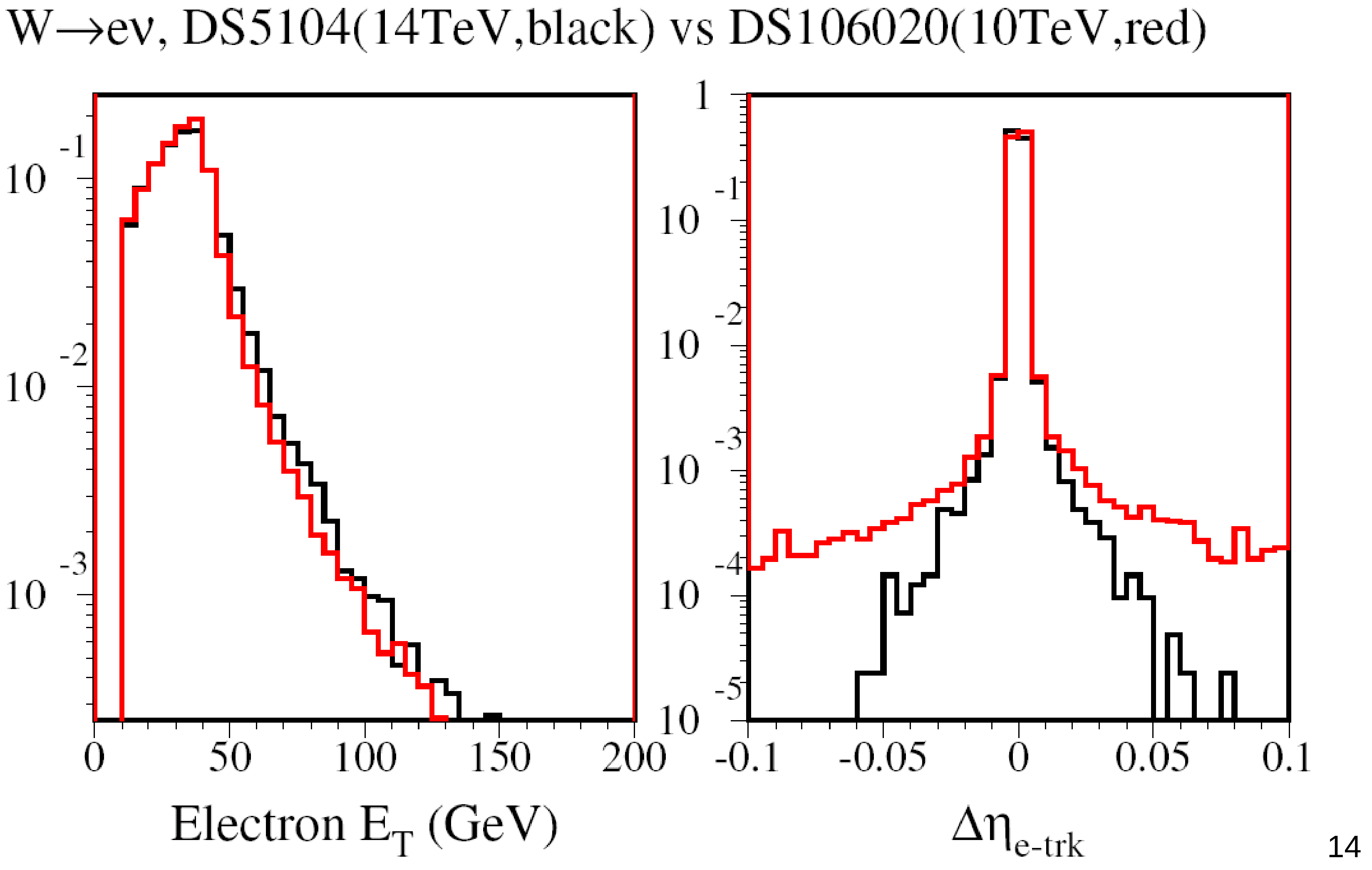
E-ID Efficiency after pre-selection vs Pt (v13, jet fake rate=1.0E-4)



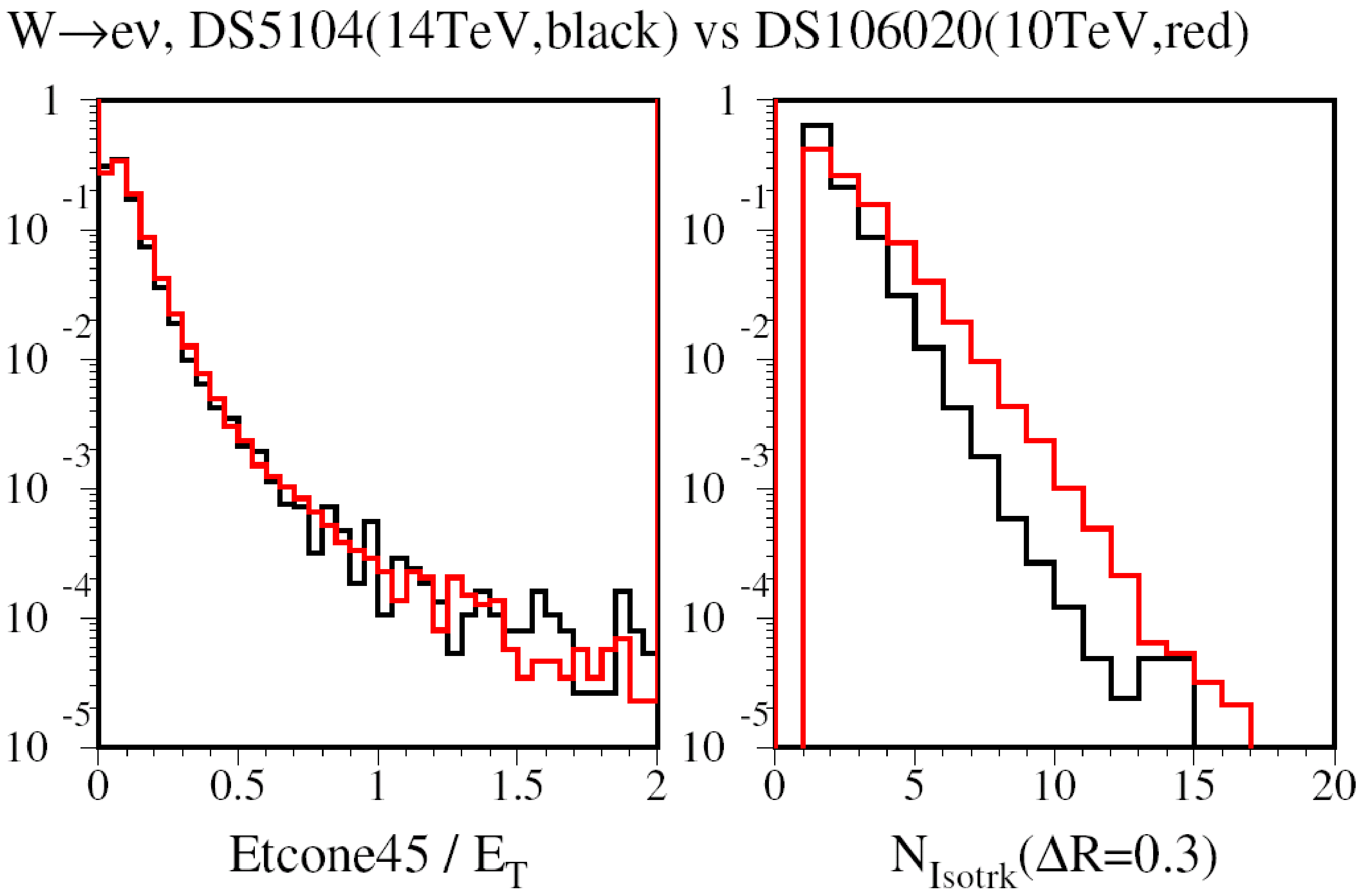
E-ID Testing Samples Produced at $\sqrt{s}$ = 10 TeV (v14)

- Wenu: DS106020 (Eff_precuts = 86.7%)
  - 58954 electrons with Et>17 GeV, $|\eta|<2.5$
  - 51100 electrons after pre-selection cuts
- Zee: DS106050 (Eff_precuts = 86.7%)
  - 108550 electrons with Et>17 GeV, $|\eta|<2.5$
  - 94153 electrons after pre-selection cuts
- JF17: DS105802 (Eff_precuts = 2.34%)
  - 237950 events, 896818 reconstructed jets
  - 20994 jets after pre-selection cuts

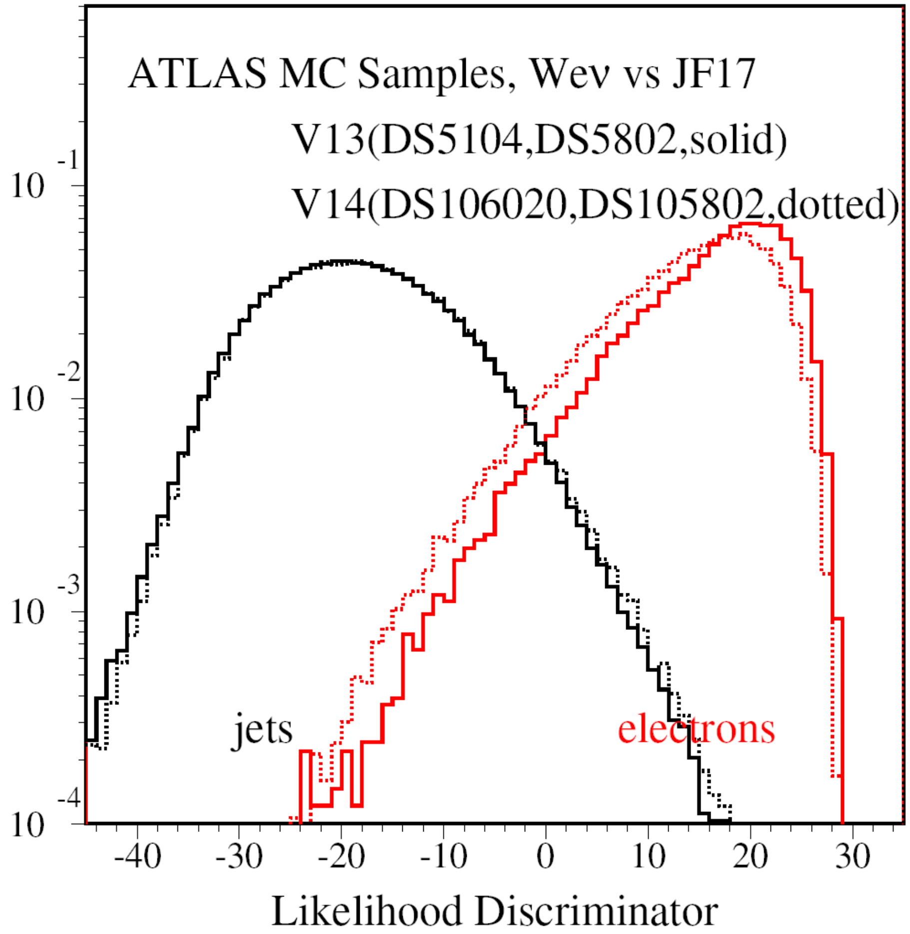
Variable distribution Comparison
14 TeV vs 10 TeV

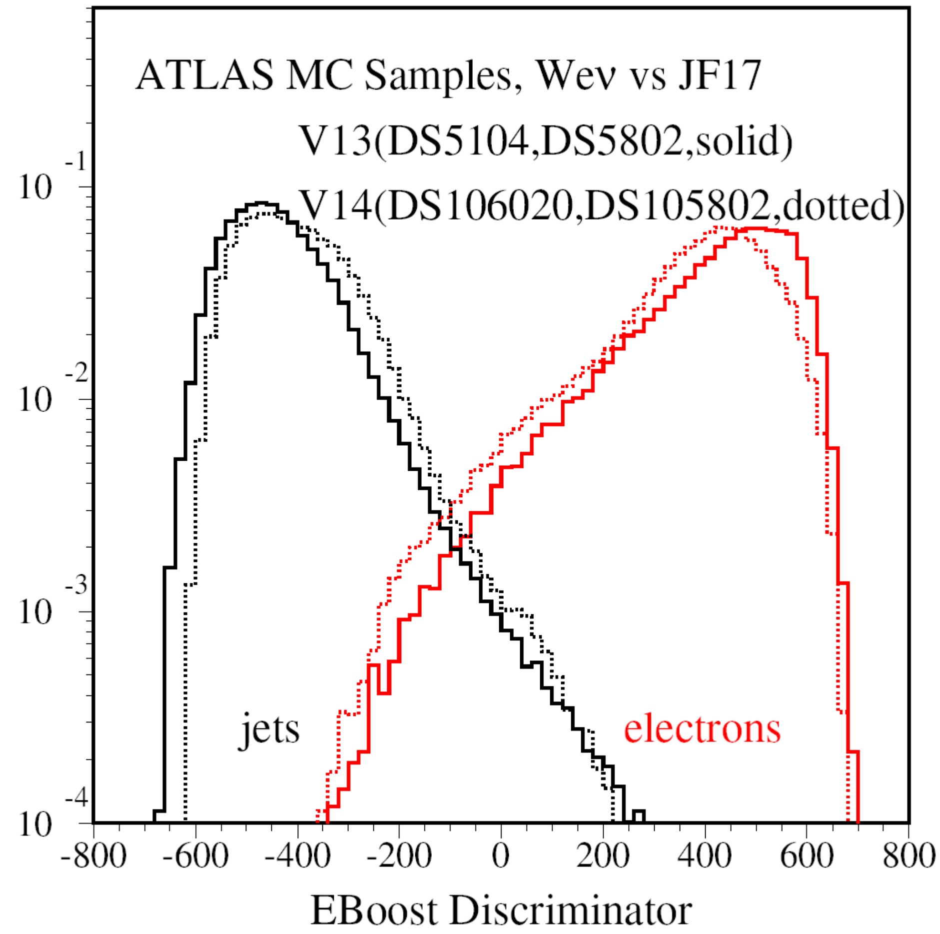


Variable distribution Comparison
14 TeV vs 10 TeV

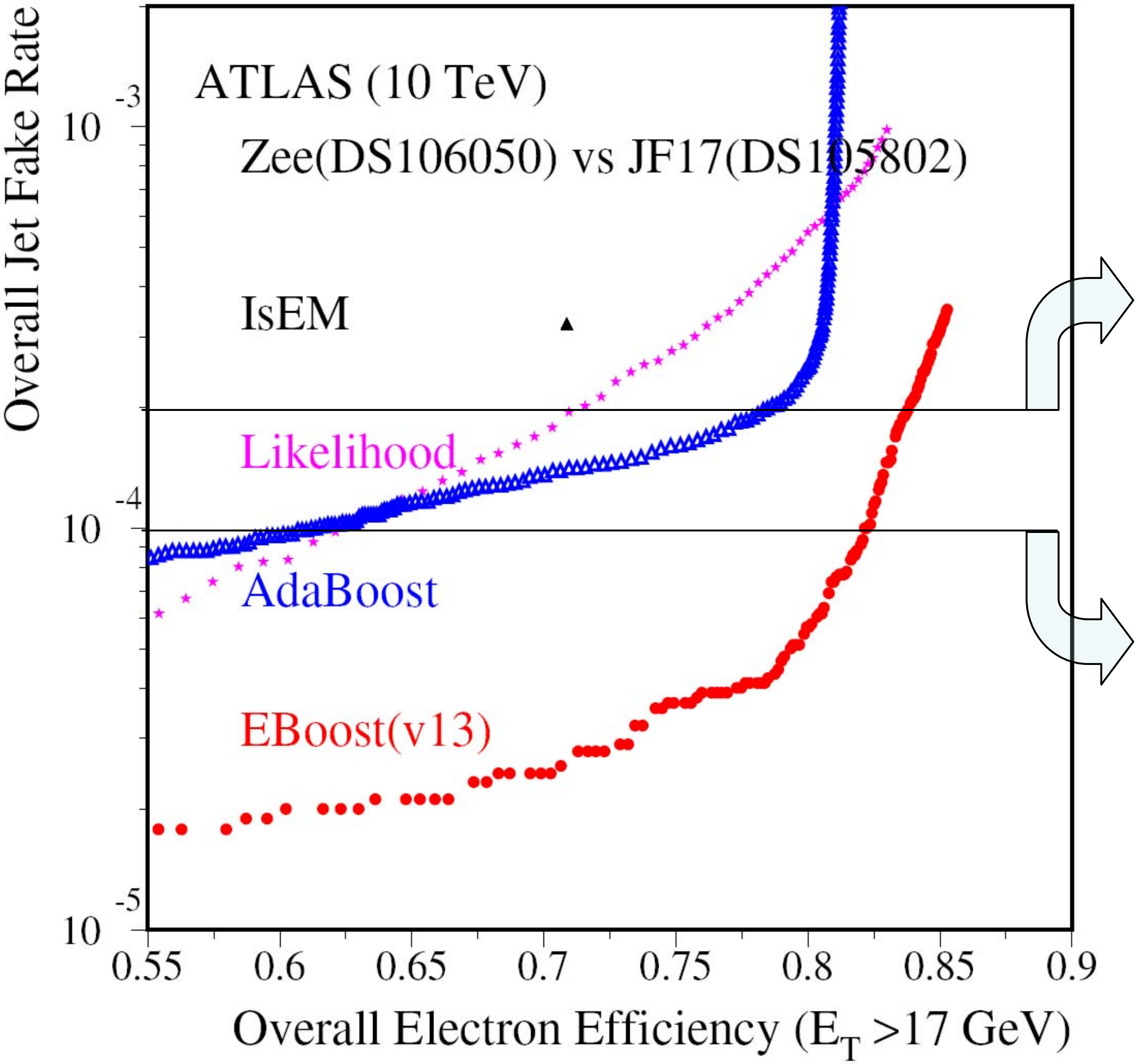


E-ID Discriminators with no retraining for 10 TeV MC Samples





Comparison of e-ID Algorithms (v14)



➔IsEM (tight)
Eff = 70.9%
jet fake rate = 3.2E-4

➔jet fake rate = 2.0E-4
Likelihood: Eff = 71.6%
AdaBoost: Eff = 78.5%
EBoost: Eff = 83.9%

➔jet fake rate = 1.0E-4
Likelihood: Eff = 62.9%
AdaBoost: Eff = 61.2%
EBoost: Eff = 82.2%

Robustness of Multivariate e-ID

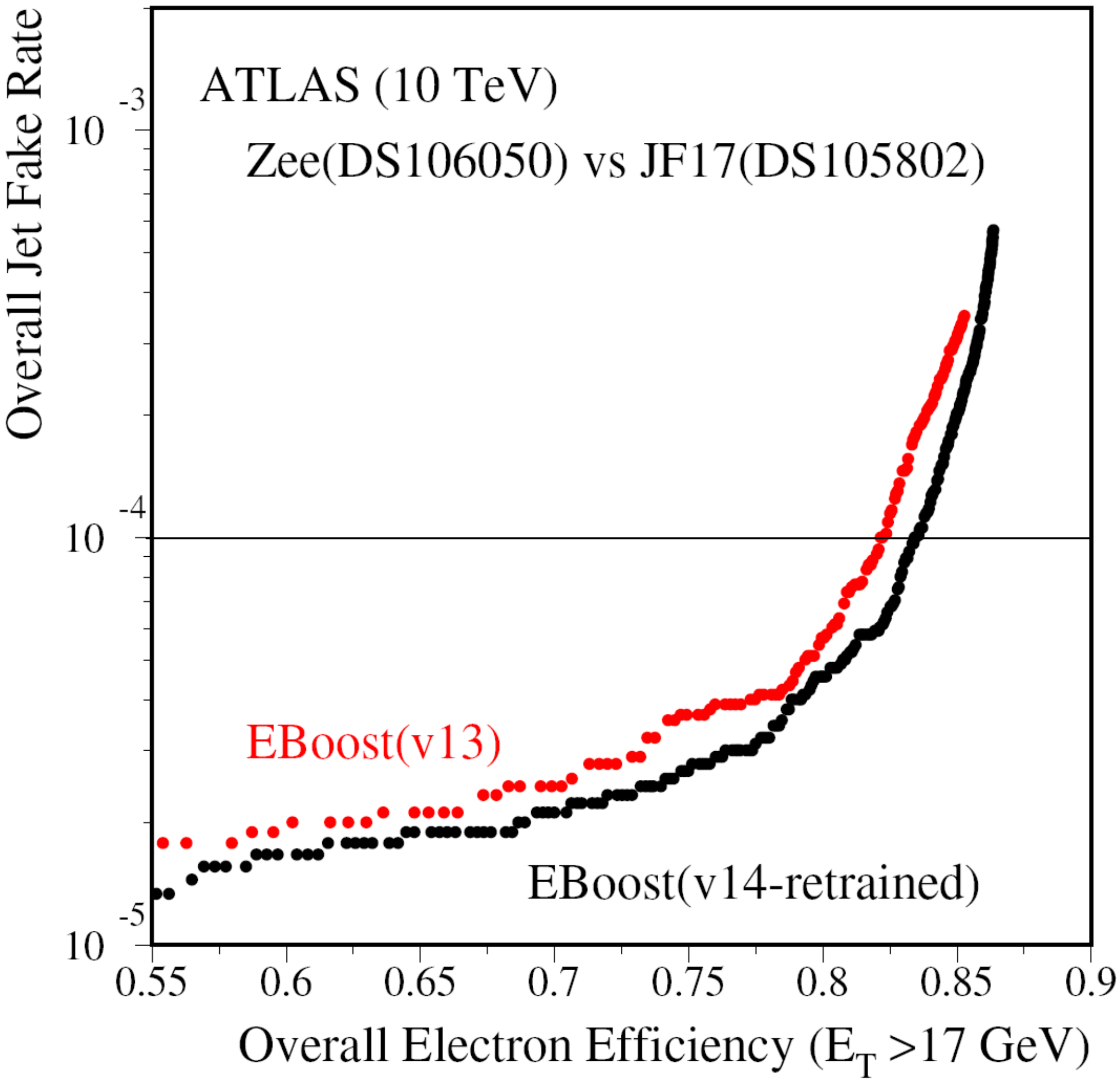
(√s =14 TeV vs. 10 TeV without retraining)

Test MC	Precuts	Likelihood	AdaBoost	EBoost
Z→ee (v13) √s =14 TeV	88.6%	75.9%	69.3%	86.0%
Z→ee (v14) √s =10 TeV	86.7%	62.9%	61.2%	82.2%
Eff. Change after pre-sel	-1.9%	-13.0% -11.1%	-8.1% -6.2%	-3.8% -1.9%
JF17 (v13) √s =14 TeV	2.4E-2	1.0E-4	1.0E-4	1.0E-4
JF17 (v14) √s =10 TeV	2.3E-2	1.0E-4	1.0E-4	1.0E-4

Robustness of Multivariate e-ID
(√s =14 vs 10 TeV without retraining)

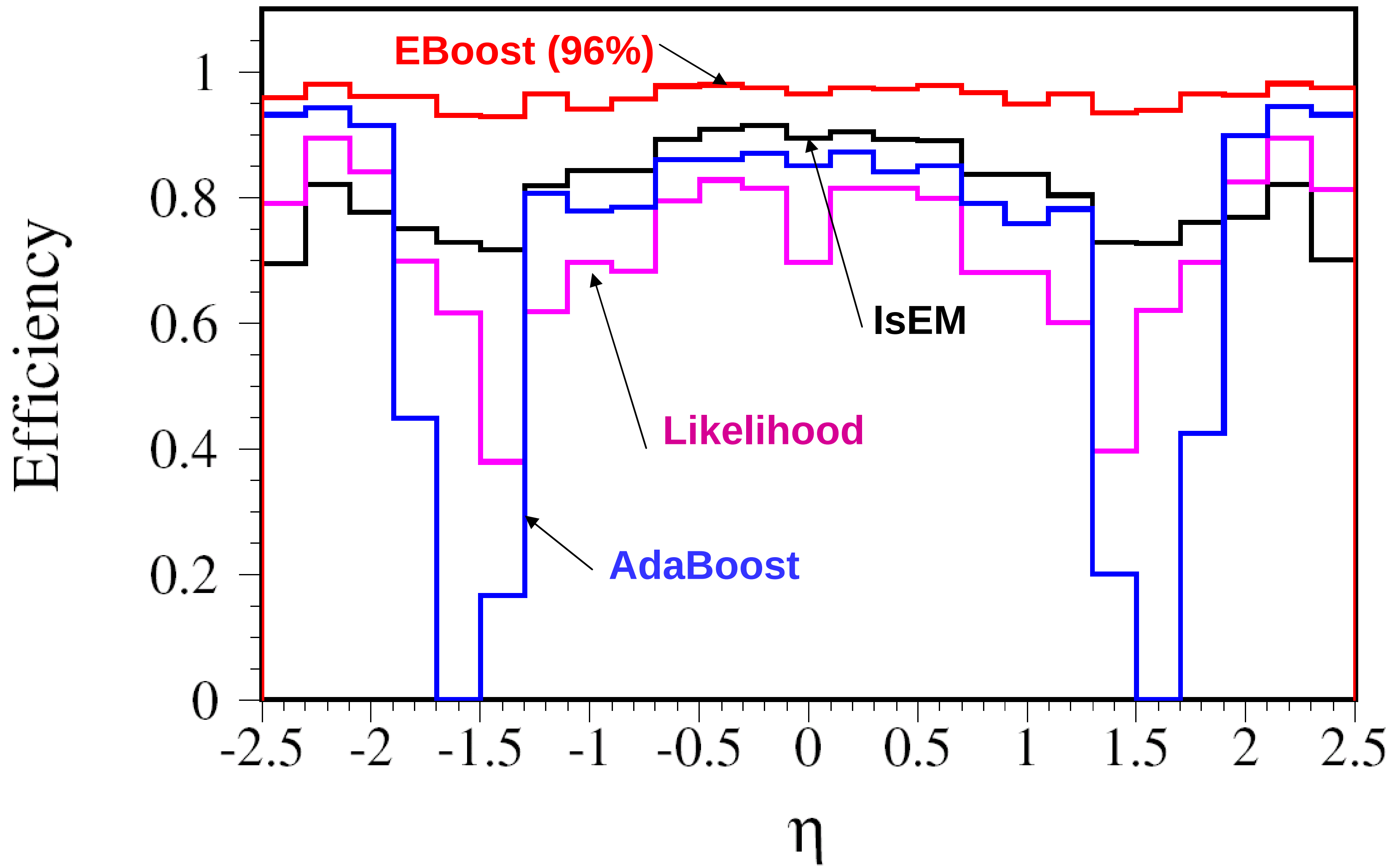
Test MC	Precuts	Likelihood	AdaBoost	EBoost
Z→ee (v13) √s =14 TeV	88.6%	81.3%	83.2%	87.5%
Z→ee (v14) √s =10 TeV	86.7%	71.6%	78.5%	83.9%
Eff. Change after pre-sel	-1.9%	-9.7% -7.8%	-4.7% -2.8%	-3.6% -1.7%
JF17 (v13) √s =14 TeV	2.4E-2	2.0E-4	2.0E-4	2.0E-4
JF17 (v14) √s =10 TeV	2.3E-2	2.0E-4	2.0E-4	2.0E-4

Improvement with EBoost Re-training using √s =10 TeV MC Samples

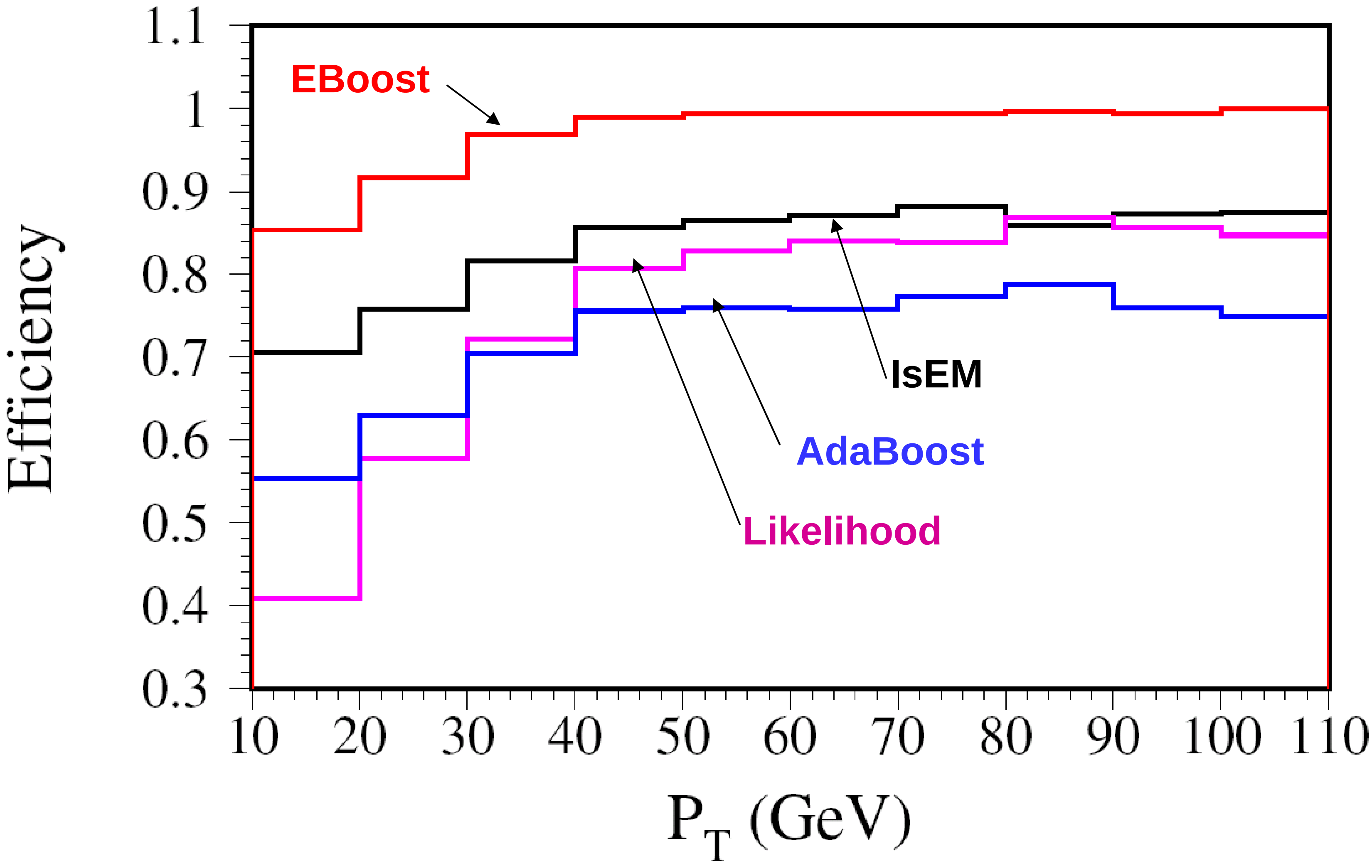


- ➔EBoost (retrained)
  Eff = 82.2 → 83.4%
  jet fake rate = 1.0E-4
- ➔For each major release multivariate e-ID should be retrained to obtain optimal performance
- ➔All multivariate e-ID should be retrained using real data

E-ID Efficiency after pre-selection vs η (v14, jet fake rate=1.0E-4)



E-ID Efficiency after pre-selection vs Pt (v14, jet fake rate=1.0E-4)



Future Plan

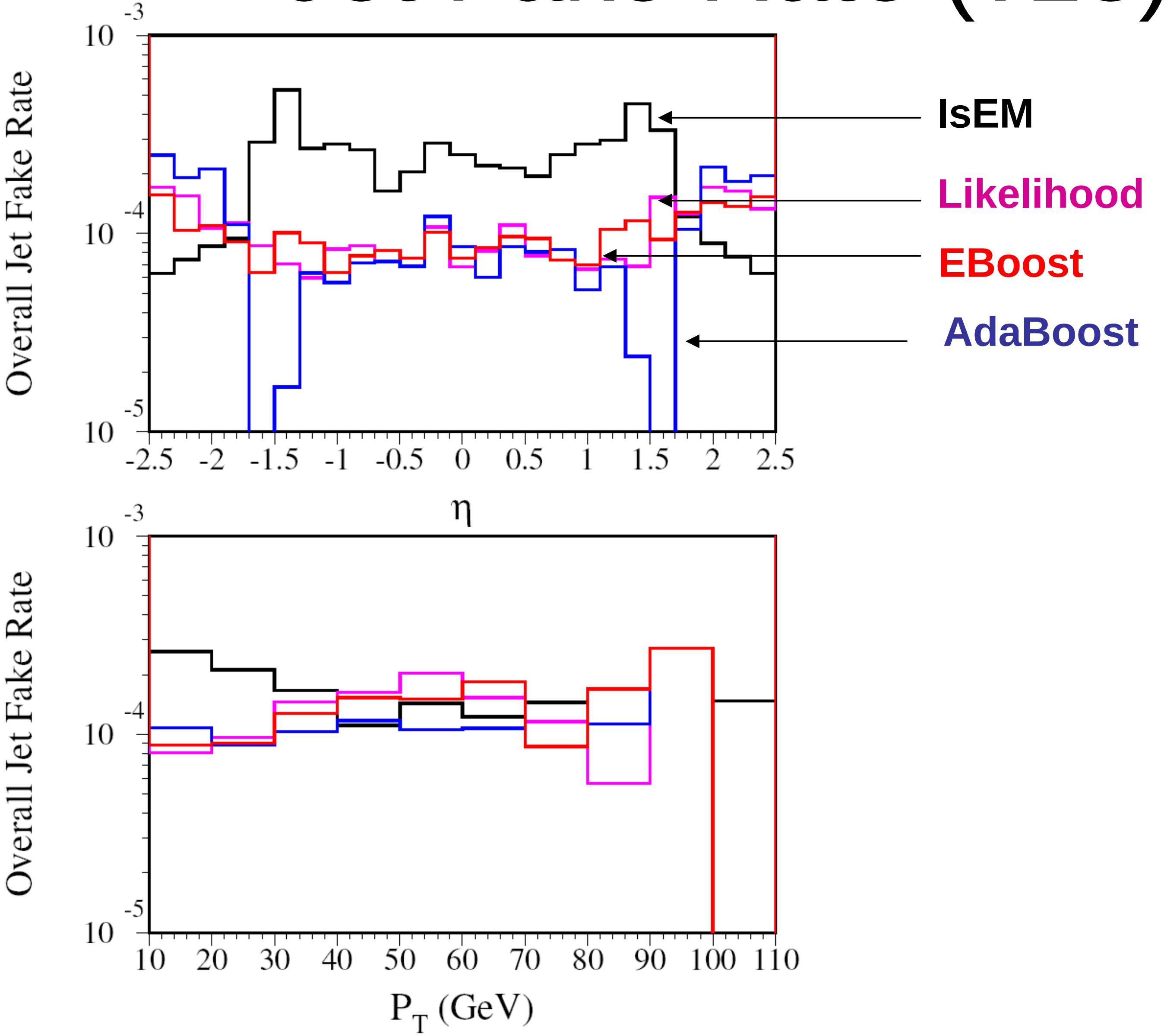
- We have requested to add the EBoost in ATLAS official egammaRec package and make EBoost discriminator variable available for more test and for physics analysis.

- We have proposed to provide EBoost trees to ATLAS egammaRec for each major software release

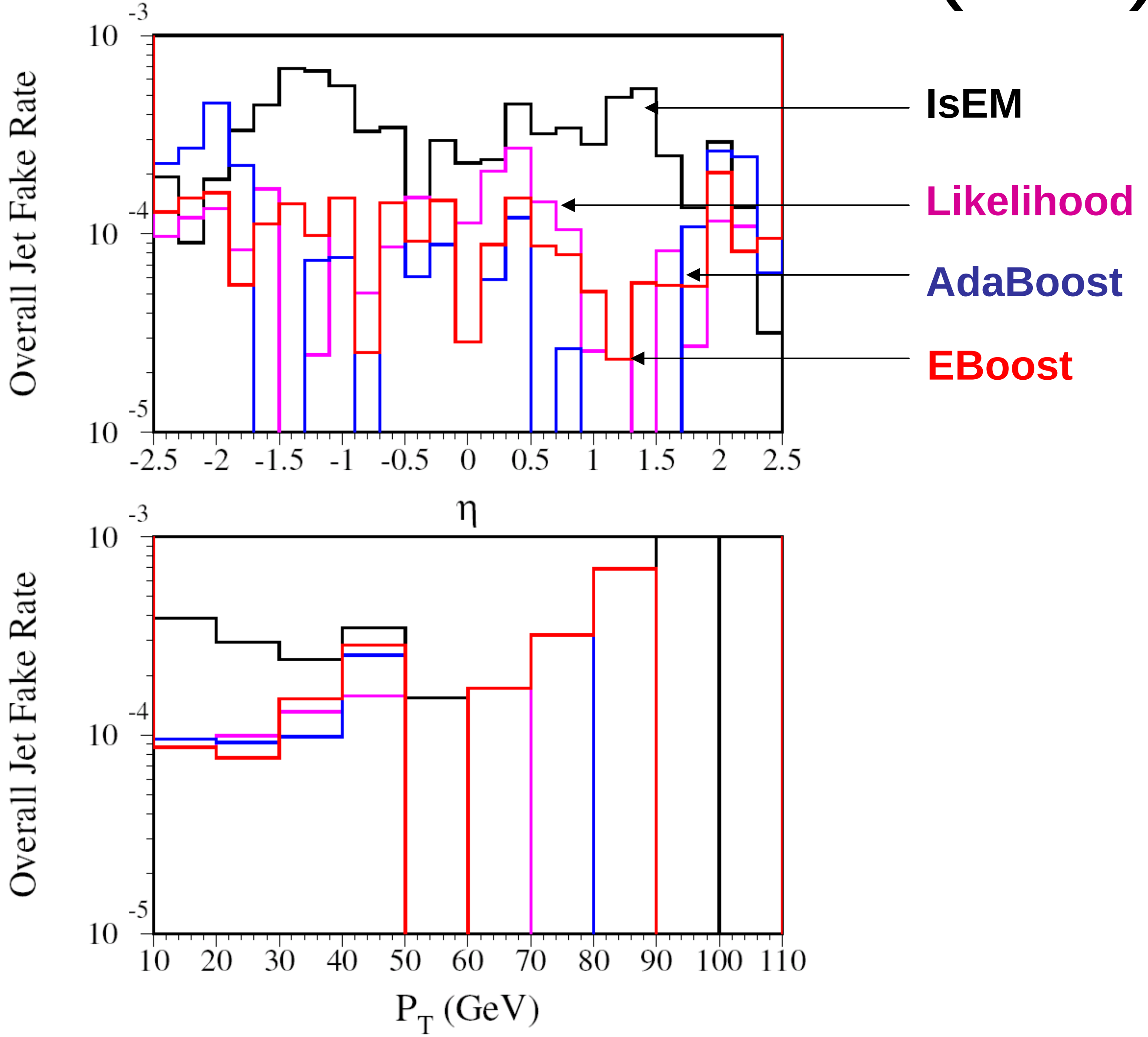
- We will explore new variables to further improve e-ID by suppressing γ converted electron etc.

Backup Slides

Jet Fake Rate (v13)



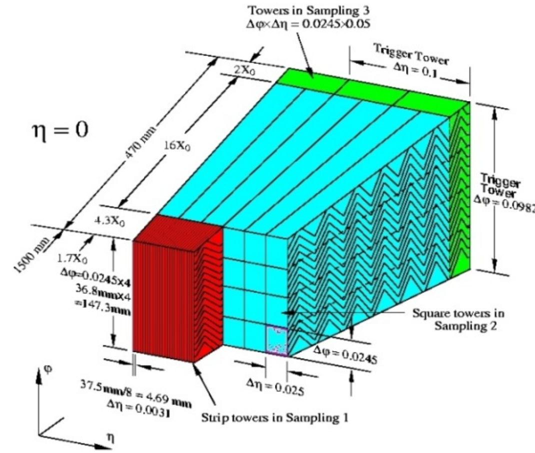
Jet Fake Rate (v14)

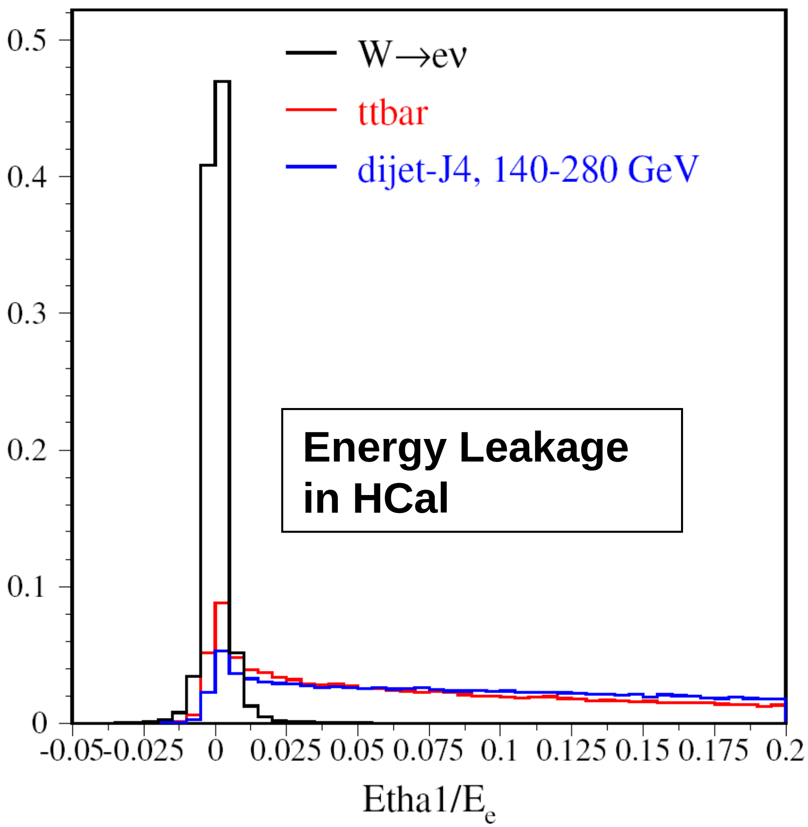


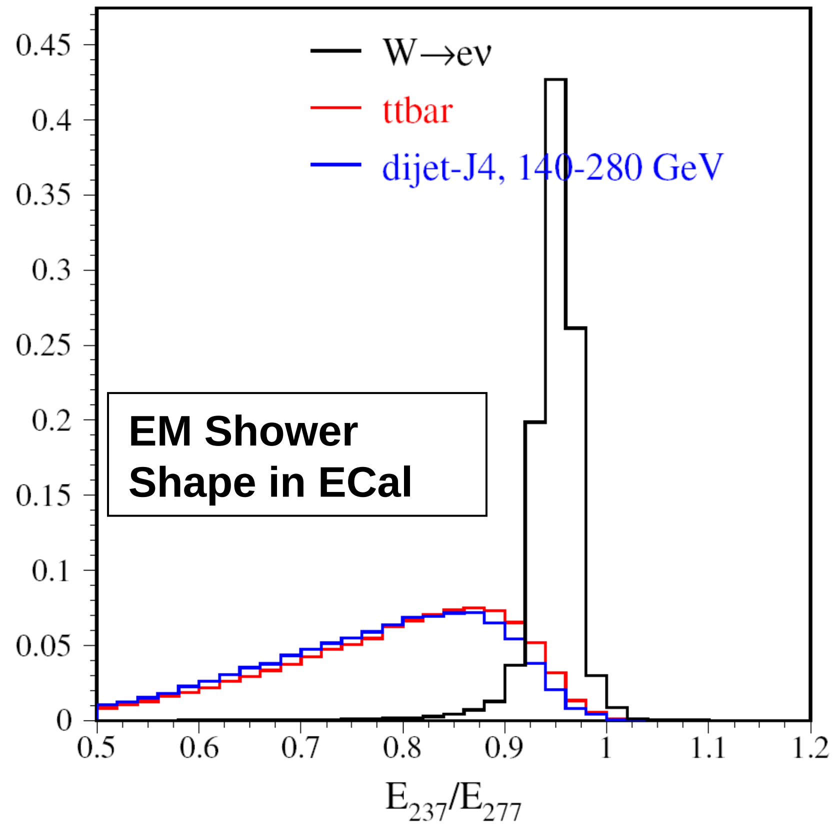
List of Variables for BDT

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

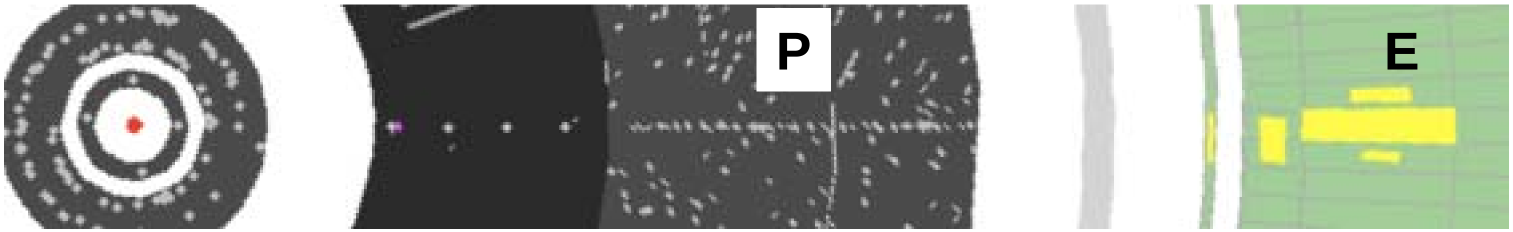
EM Shower shape distributions of discriminating Variables (signal vs. background)

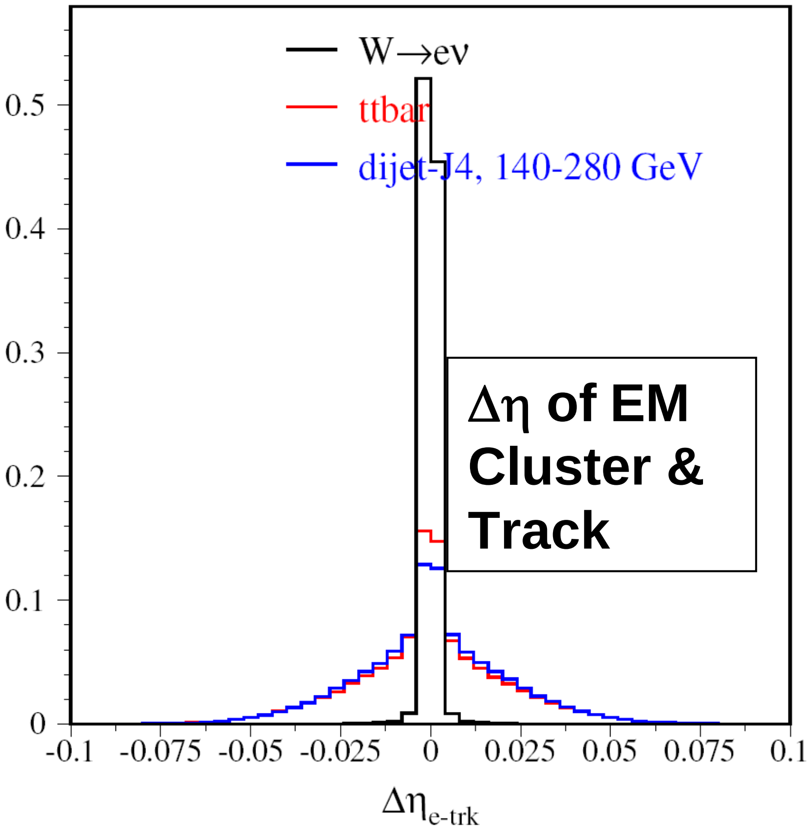


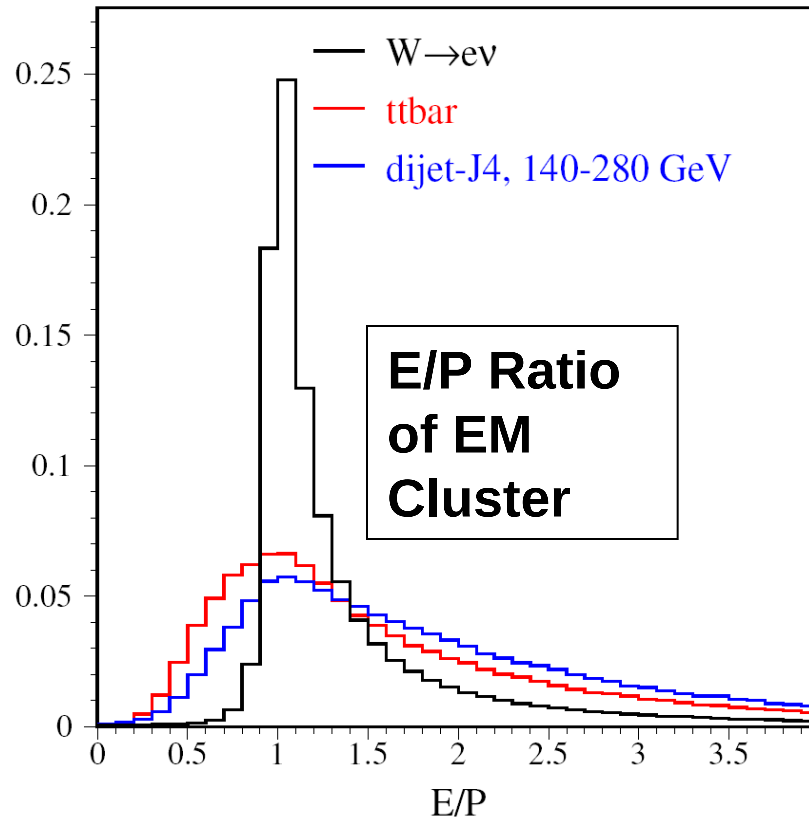




ECal and Inner Track Match

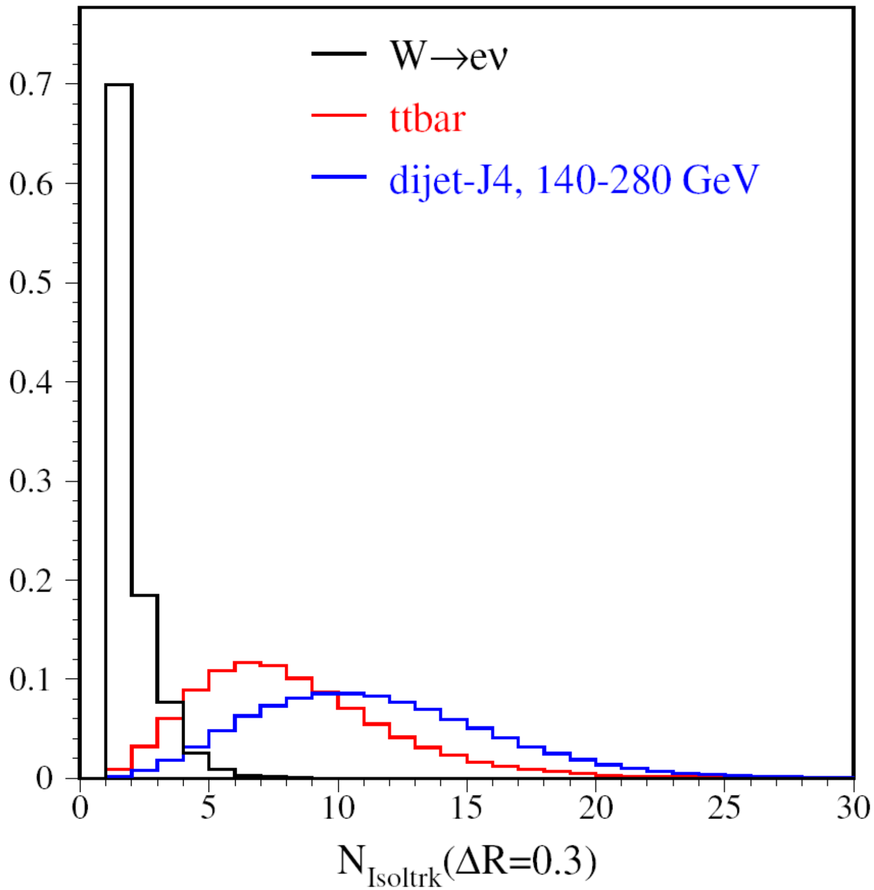




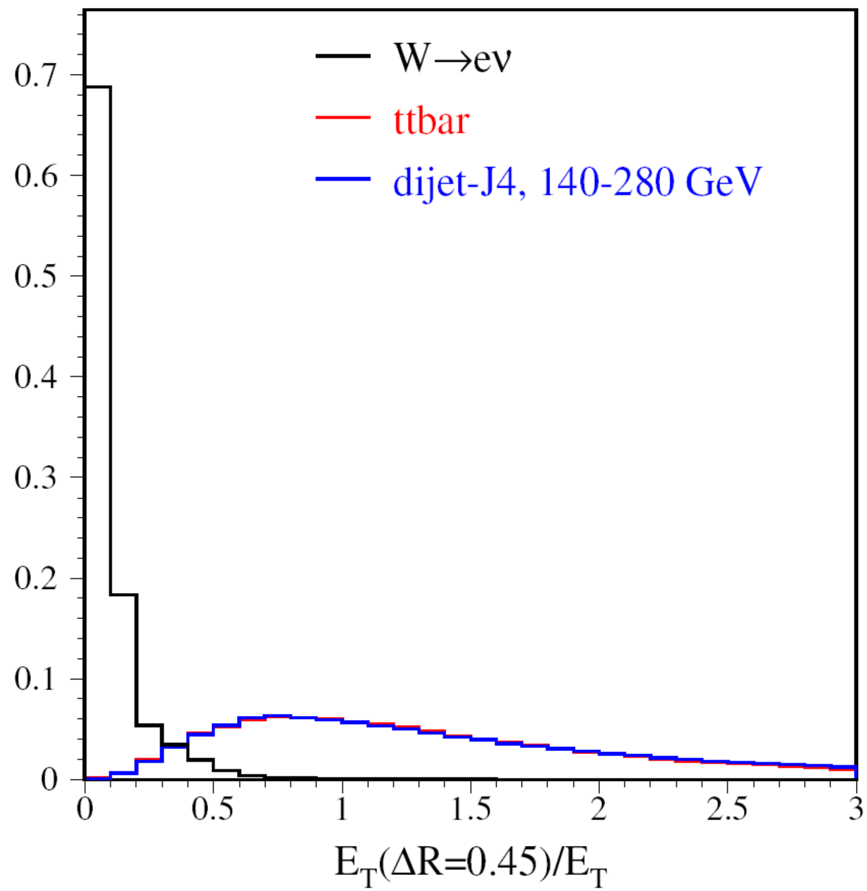


Electron Isolation Variables

N_{trk} around Electron Track



E_T(ΔR=0.2-0.45)/E_T of EM



Signal Pre-selection: MC electrons

- MC True electron from W→eν by requiring
  - $|\eta_e| < 2.5$ and $E_T^{true} > 17$ 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 reconstructed jets with
  - $|\eta_{jet}| < 2.5$ (N_{jet})
- Loop over all EM clusters; each cluster matches with a jet
  - $E_T^{EM} > 17$ GeV (N_{EM})
- Match EM cluster with an inner track:
  - eg_trkmatchnt > -1 ($N_{EM/track}$)
- Pre-selection Acceptance = $N_{EM/Track} / N_{jet}$