# Electron Identification Based on Boosted Decision Trees

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

ATLAS Egamma Meeting
October 2, 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

# MC Samples for e-ID studies

| **e Signal** | **Dataset** | **SW Version** |
|---|---|---|
| **W → eν** | **5104** | **V13** |
| **W → eν** | **5104** | **V12** |
| Z → ee | 5144 | V12 |
| WW → eνμν | 5922, 5925 | V12 |
| ZZ → 4l | 5931 | V12 |

| **Jet samples** | **Dataset** | **SW Version** |
|---|---|---|
| J0: di-jet (8<Pt<17 GeV) | 5009 | **V12, V13** |
| J1: di-jet (17<Pt<35 GeV) | 5010 | **V12, V13** |
| J2: di-jet (35<Pt<70 GeV) | 5011 | **V12, V13** |
| J3: di-jet (70<Pt<140 GeV) | 5012 | **V12** |
| J4: di-jet (140<Pt<280 GeV) | 5013 | **V12, V13** |
| J5: di-jet (280<Pt<560 GeV) | 5014 | **V12, V13** |
| J6: di-jet (560<Pt<1120 GeV) | 5015 | **V12, V13** |
| ttbar → Wb Wb → all jets | 5204 | **V12** |

# Electron Identification Studies

**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, and Likelihood for SW release v12 samples; add BDT for v13).

**New BDT e-ID development at U. Michigan**

- Based on version 12 datasets (talk by H. Yang)
  http://indico.cern.ch/conferenceDisplay.py?confId=38991

-- Further study based on version 13 datasets

**Performance comparisons**

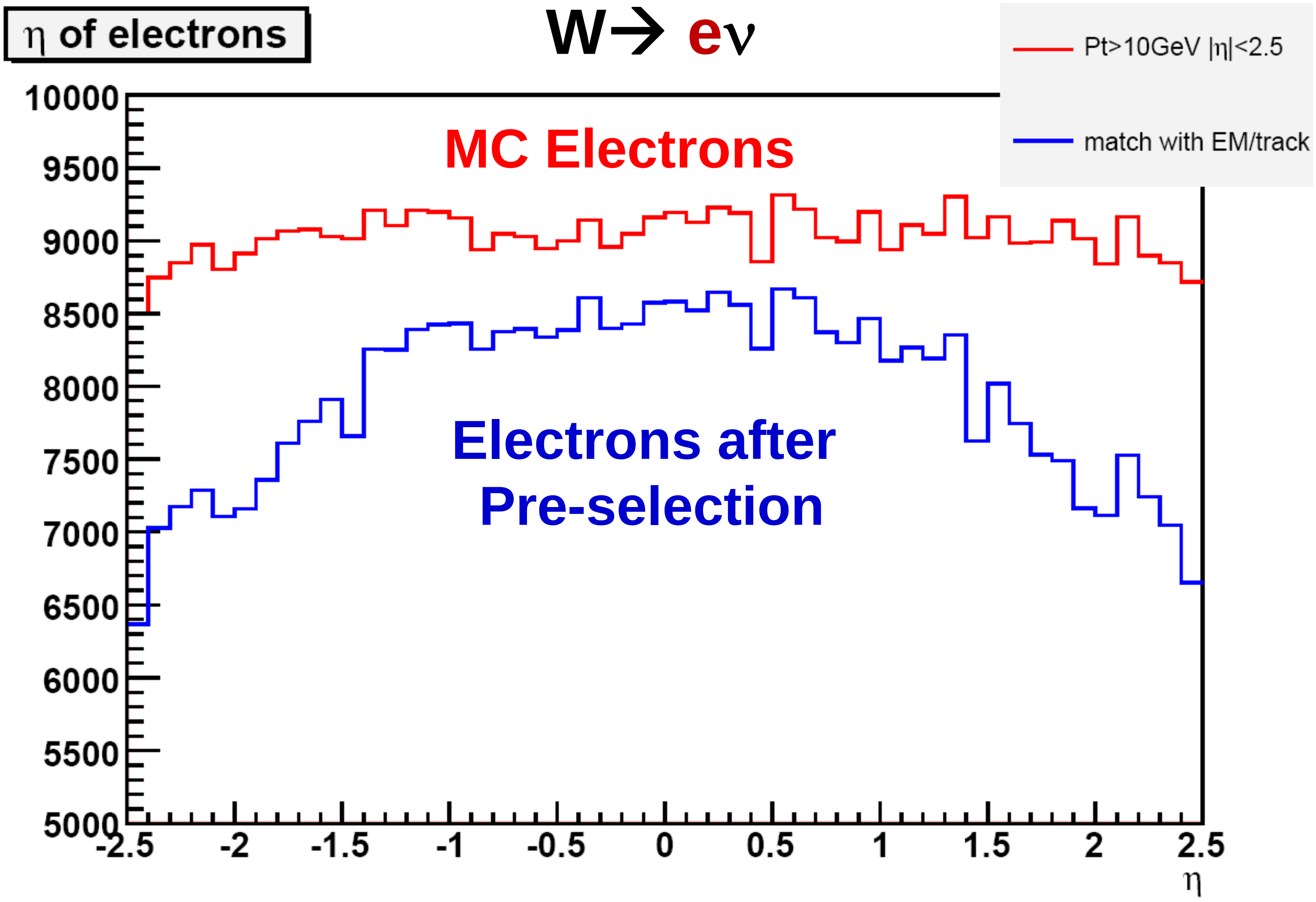
-- electron ID efficiency

-- jet fake rate

# Signal Pre-selection: MC electrons

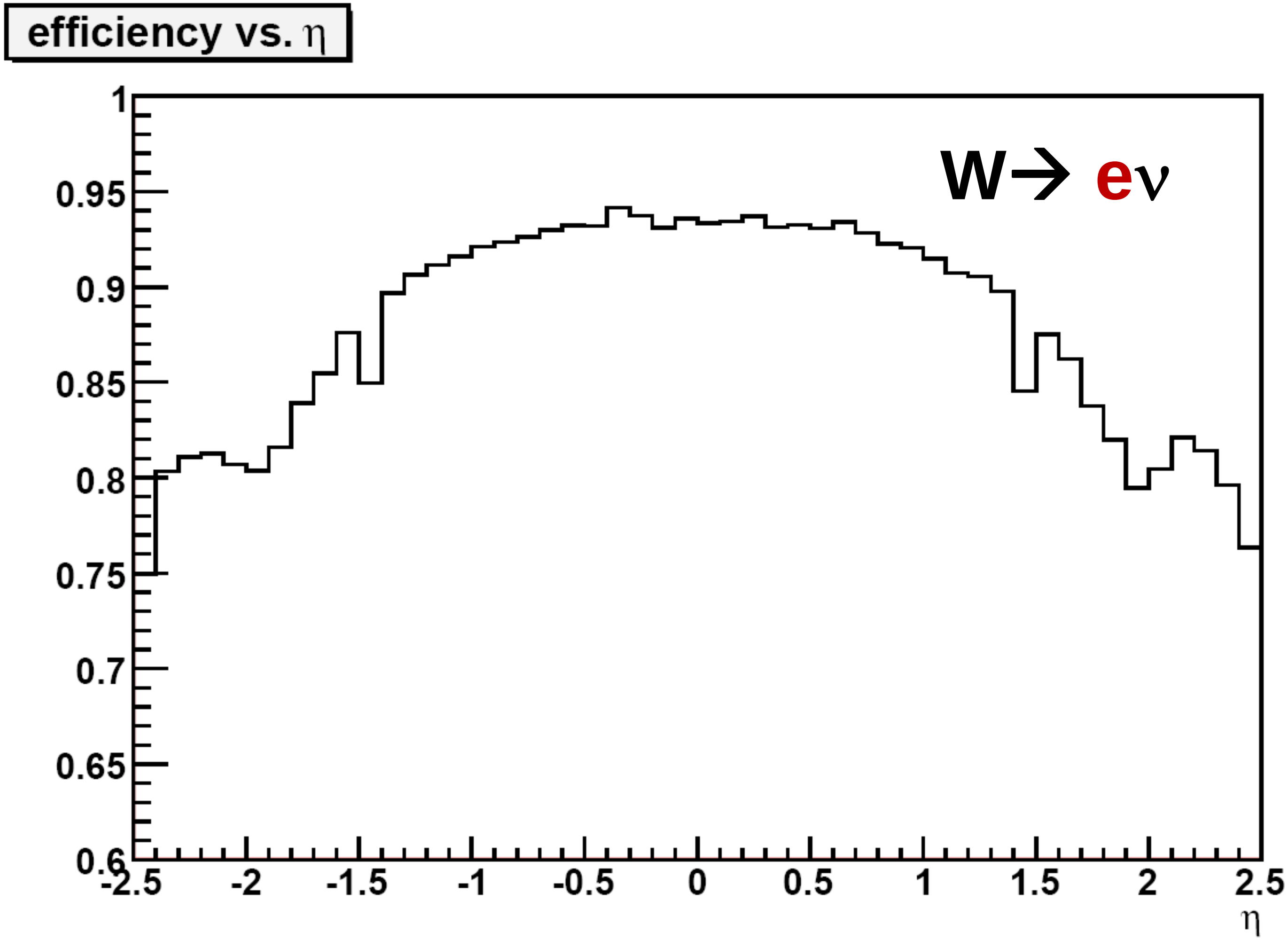
- MC True electron from W→e$\nu$ by requiring
  - $|\eta_e| < 2.5$ and $E_T^{true} > 10$ GeV ($N_e$)
- Match MC e/$\gamma$ to EM cluster:
  - $\Delta R < 0.2$ and $0.5 < E_T^{rec} / E_T^{true} < 1.5$ ($N_{EM}$)
- Match EM cluster with an inner track:
  - eg_trkmatchnt > -1 ($N_{EM/track}$)
- Pre-selection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

# 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**



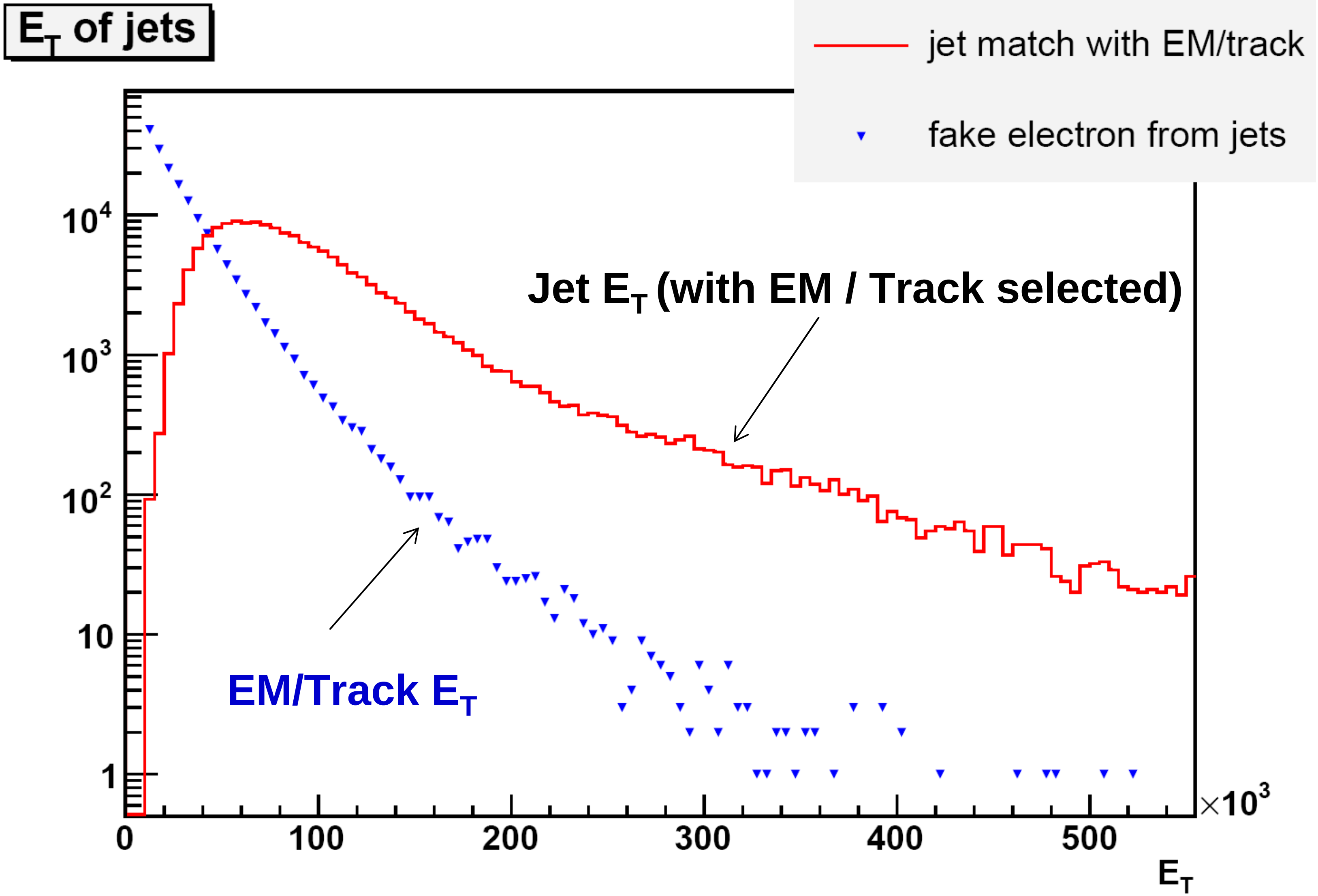
# Electron Pre-selection Efficiency

| e from process | Dataset | Software Version | EM / Track Match |
|---|---|---|---|
| **W → $e\nu$ ( $N_e$ = 135000)** | **5104** | **V13** | **89.1%** |
| **W → $e\nu$ ( $N_e$ = 485489)** | **5104** | **V12** | **88.2%** |
| Z → ee ($N_e$ = 29383) | 5144 | V12 | 87.3% |
| WW → $e\nu\mu\nu$ ( $N_e$ = 39822) | 5922<br>5925 | V12 | 87.8% |
| ZZ → 4l ( $N_e$ = 97928) | 5931 | V12 | 87.4% |

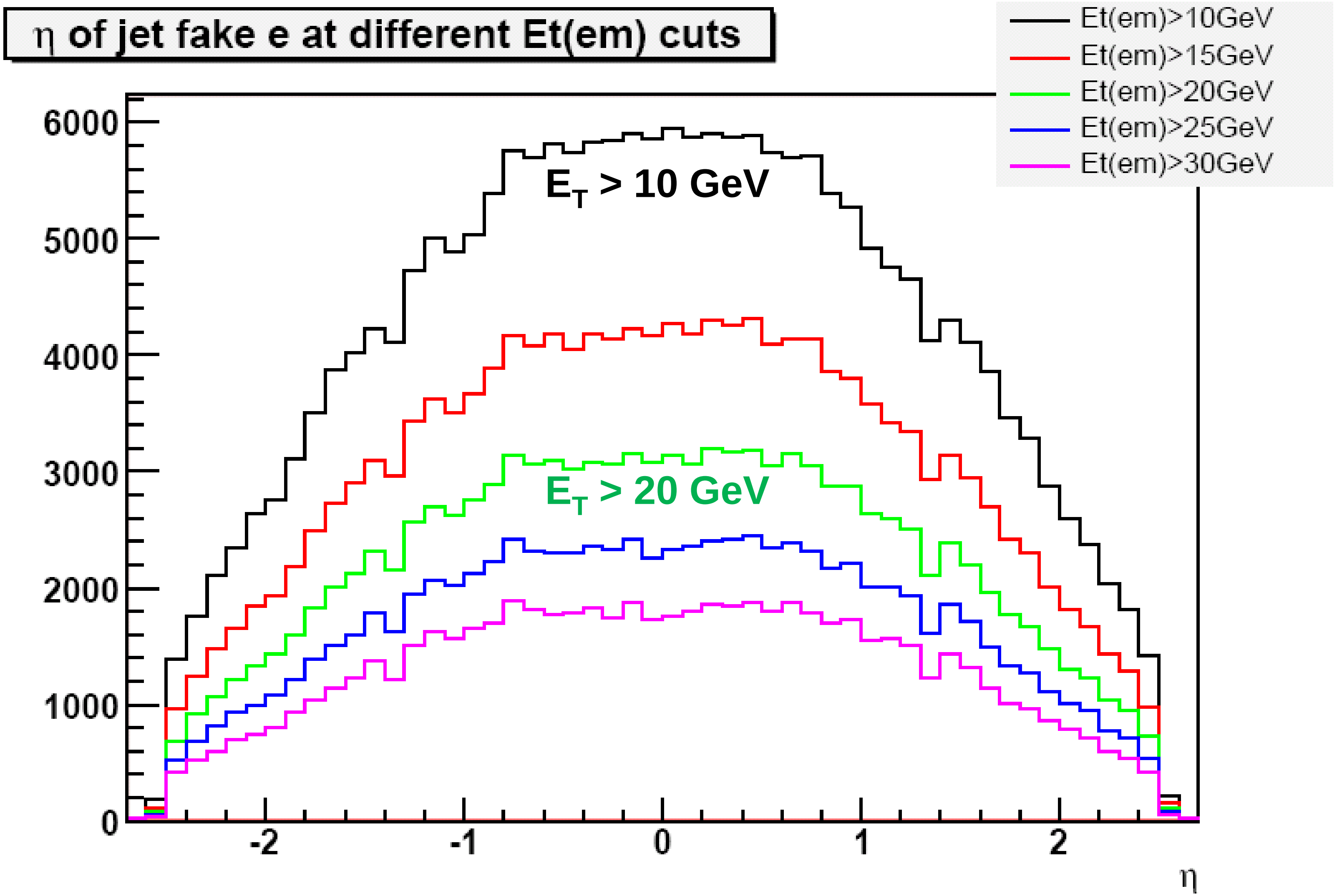
# Pre-selection of Jet Faked Electrons

- 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 = $N_{EM/Track} / N_{jet}$

# Jets (from $t\bar{t}$) and Faked Electrons



# Faked Electron from Top Jets vs Different EM $E_T$



# Jet Fake Rate from Pre-selection

**$E_T^{jet}$ > 10 GeV, $|\eta^{jet}|$ < 2.5, Match the EM/Track object to the closest jet**

| From process | Dataset | $N_{jets}$ | V13 | V12 |
|---|---|---|---|---|
| J0: di-jet (8<Pt<17 GeV) | 5009 | **404363** | **4.8E-3** | 6.0E-3 |
| J1: di-jet (17<Pt<35 GeV) | 5010 | **724033** | **1.5E-2** | 1.5E-2 |
| J2: di-jet (35<Pt<70 GeV) | 5011 | **713308** | **9.1E-2** | 1.1E-1 |
| J3: di-jet (70<Pt<140 GeV) | 5012 | **42330** | **N/A** | 3.2E-1 |
| J4: di-jet (140<Pt<280 GeV) | 5013 | **1185538** | **3.3E-1** | 4.3E-1 |
| J5: di-jet (280<Pt<560 GeV) | 5014 | **1606039** | **3.6E-1** | 5.1E-1 |
| J6: di-jet (560<Pt<1120 GeV) | 5015 | **1828401** | **3.3E-1** | 5.0E-1 |
| ttbar → Wb Wb → all jets | 5204 | **675046** | **N/A** | 3.2E-1 |

# Existing ATLAS e-ID Algorithms

**1) IsEM & 0x7FFFFF == 0 (v13)**

**2) Likelihood:**

$D_{LH}$ = **log (EMWeight / PionWeight) > 6.5 (V13)**

**3) Ele_BDTScore (Rel. v13) > 7 (v13)**

**e-ID in V12 (talk by H. Yang on Sept. 10, 2008):**

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

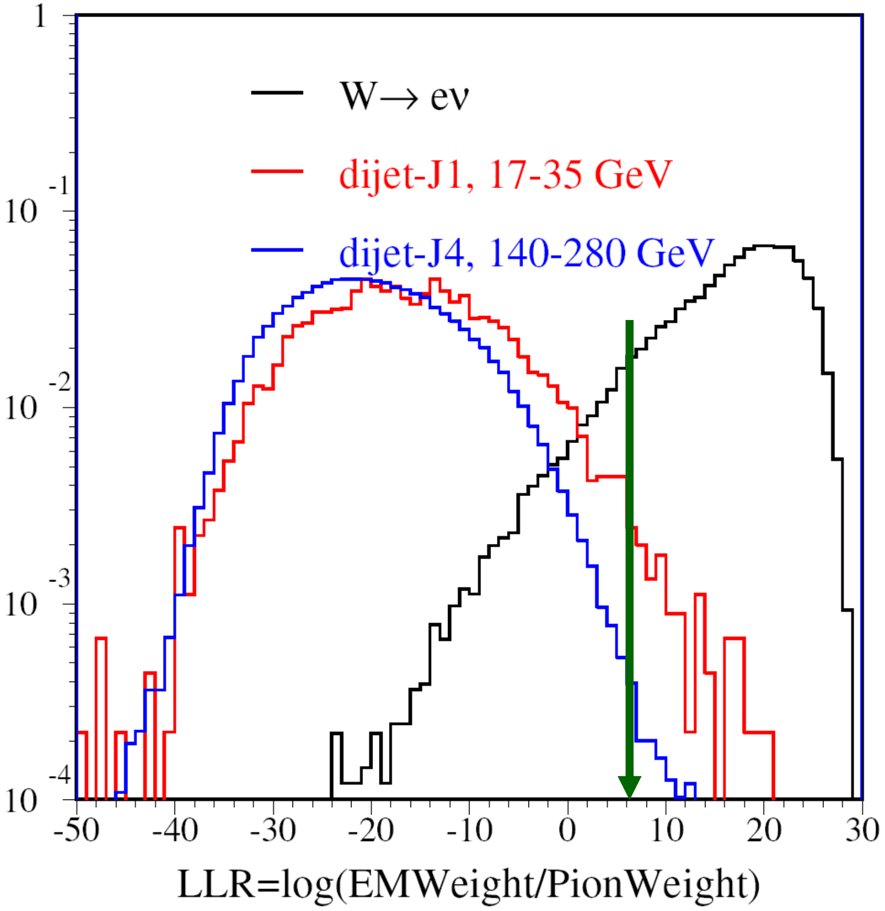
**1) IsEM & 0x7FF == 0**

**2) Likelihood:**

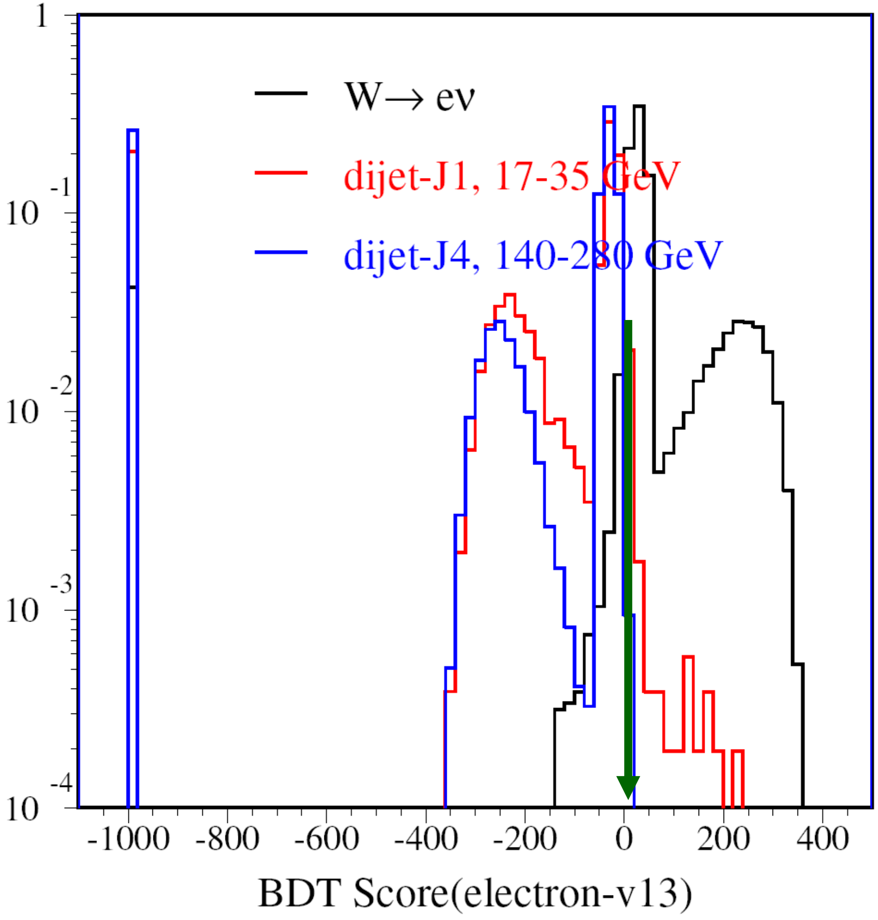
$D_{LH}$ = **EMWeight /(EMWeight+PionWeight) > 0.6**

# e-ID multivariate discriminators (v13)

## Likelihood discriminator



## Discriminator of Ele_BDTScore



# Variables Used for BDT e-ID (UM)

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

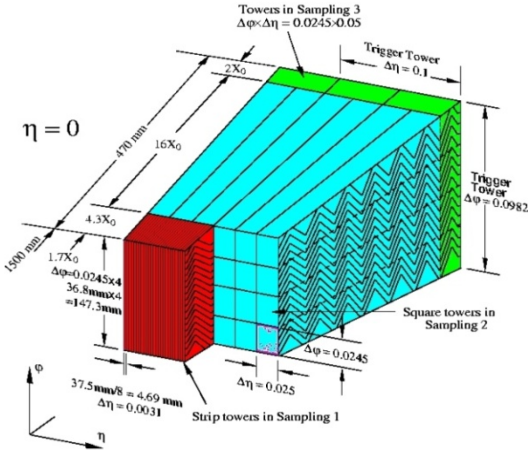
- Track Eta and EM Eta
- Electron isolation variables:

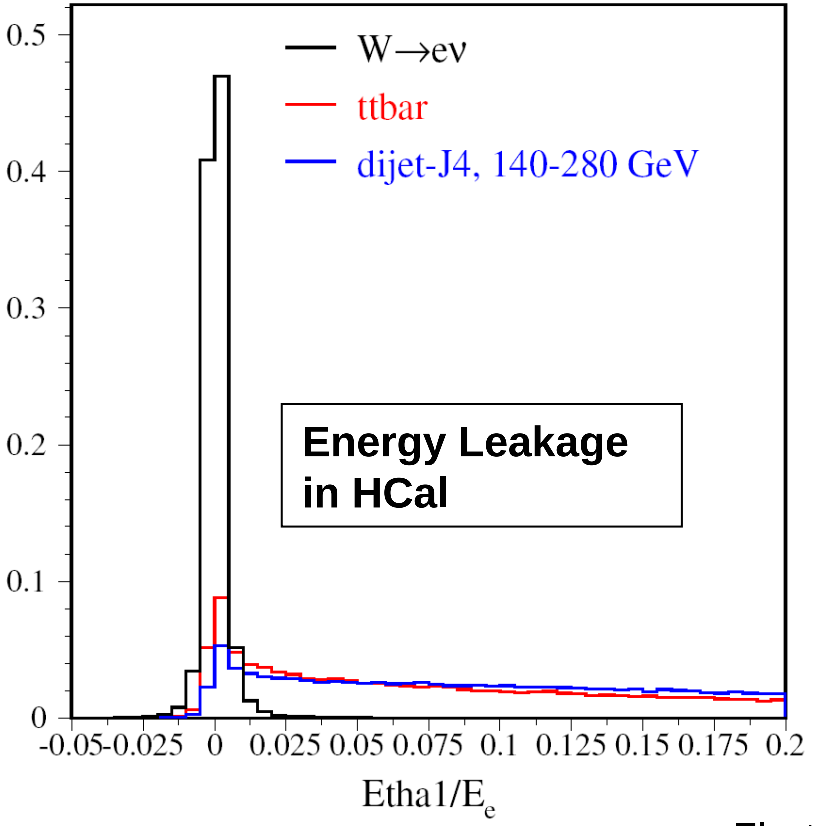
*Number of tracks ($\Delta R=0.3$)*

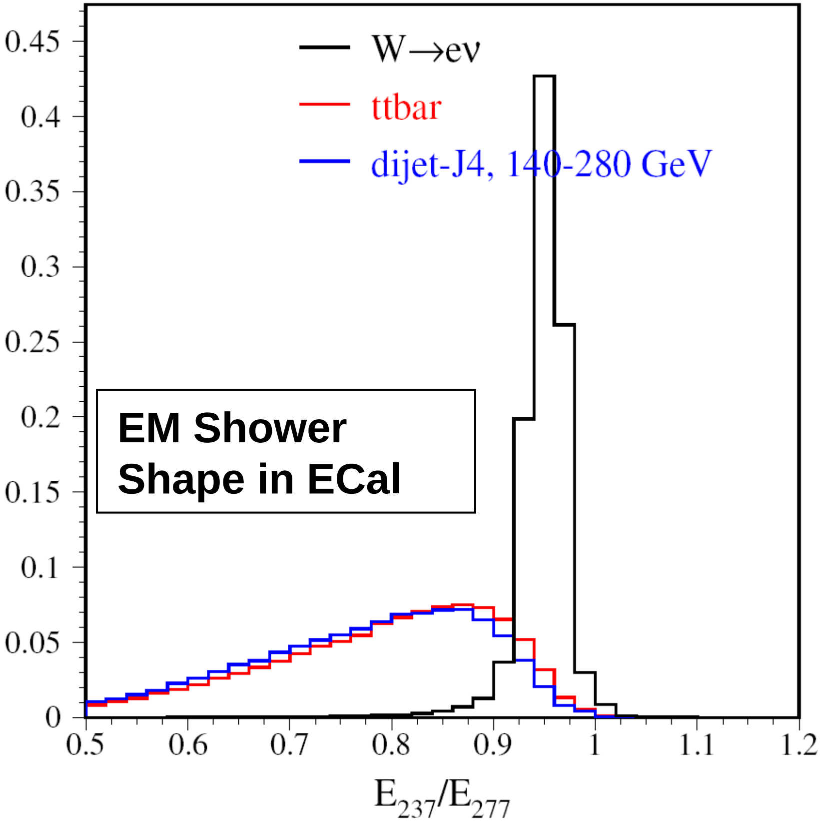
*Sum of track momentum ($\Delta R=0.3$)*

*Ratio of energy in $\Delta R=0.2$-$0.45$ and $\Delta R=0.2$*

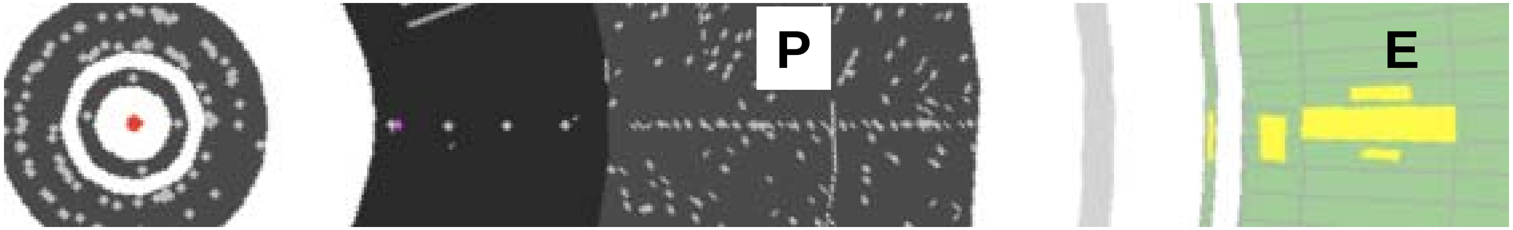
# EM Shower shape distributions of discriminating Variables (signal vs. background)

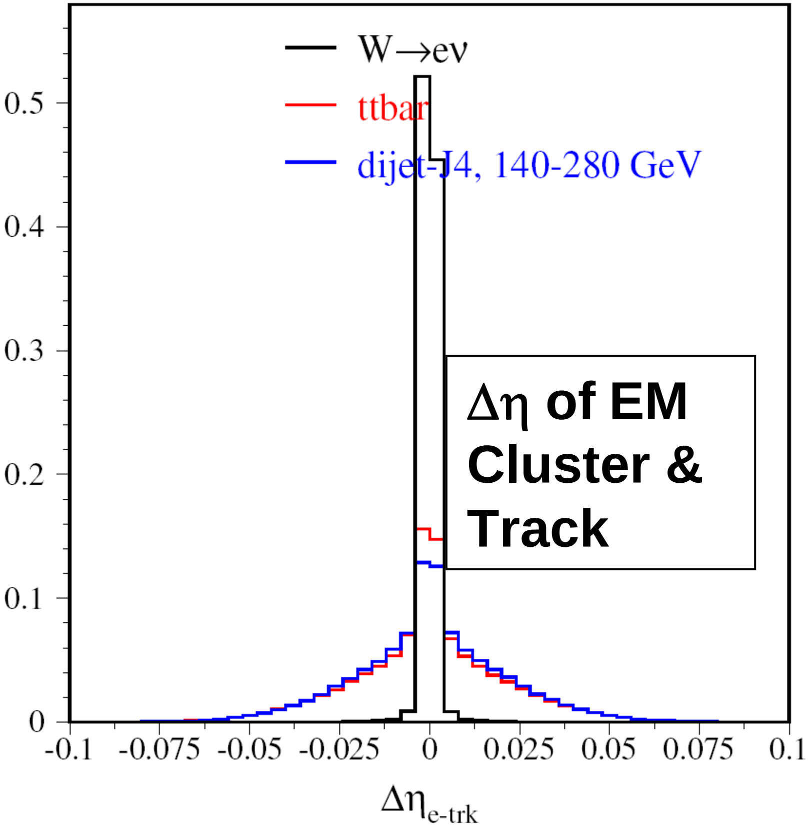




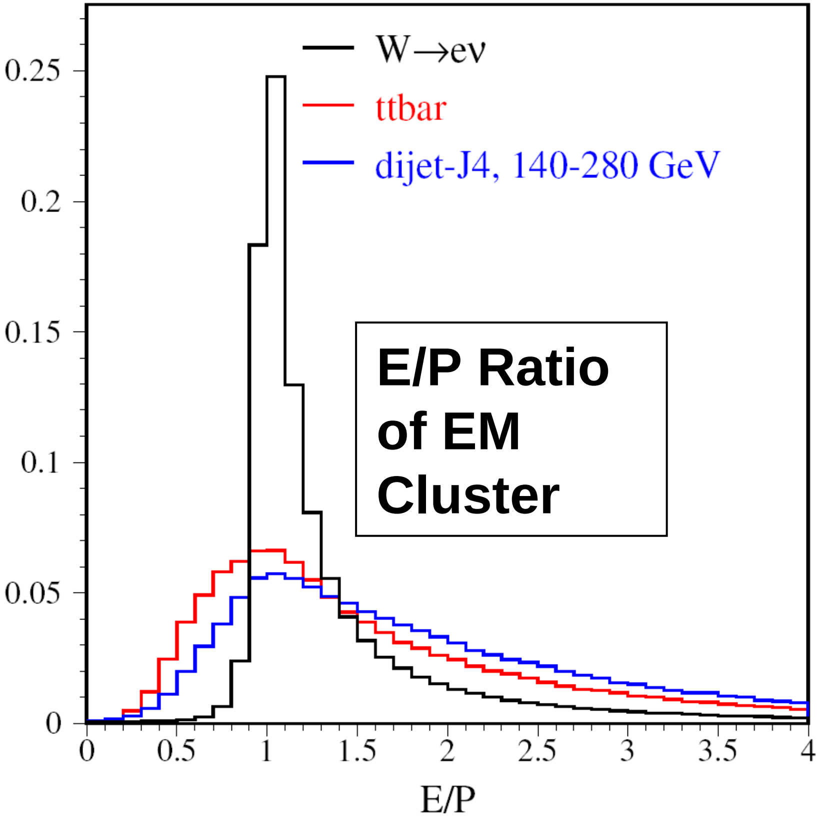


# ECal and Inner Track Match





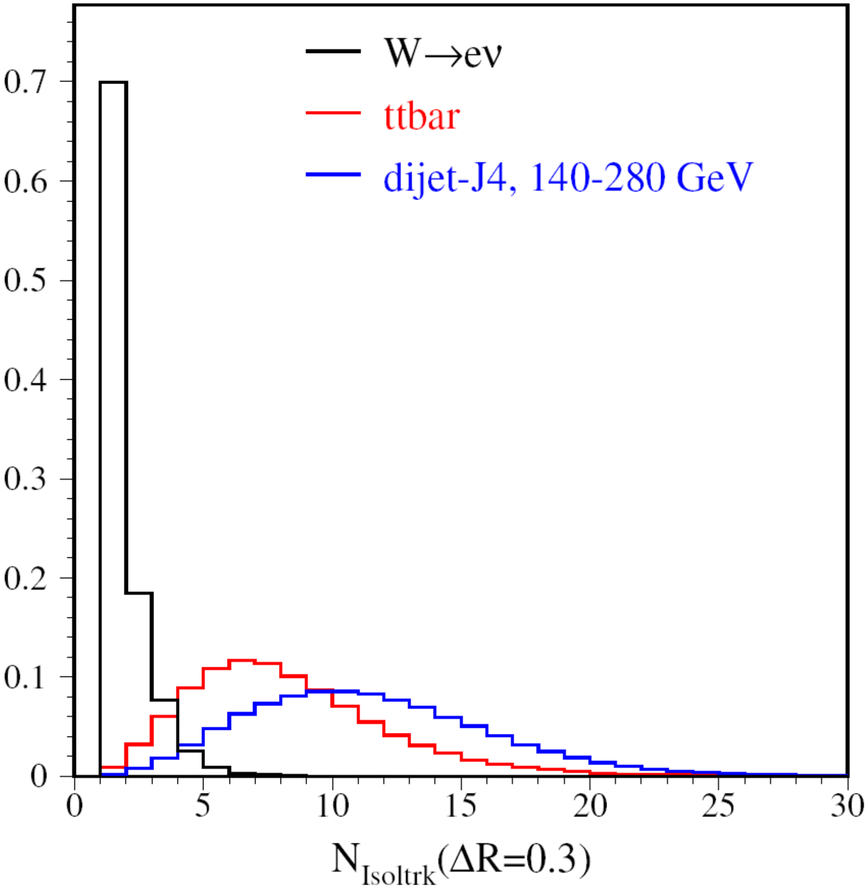
**Δη of EM Cluster & Track**



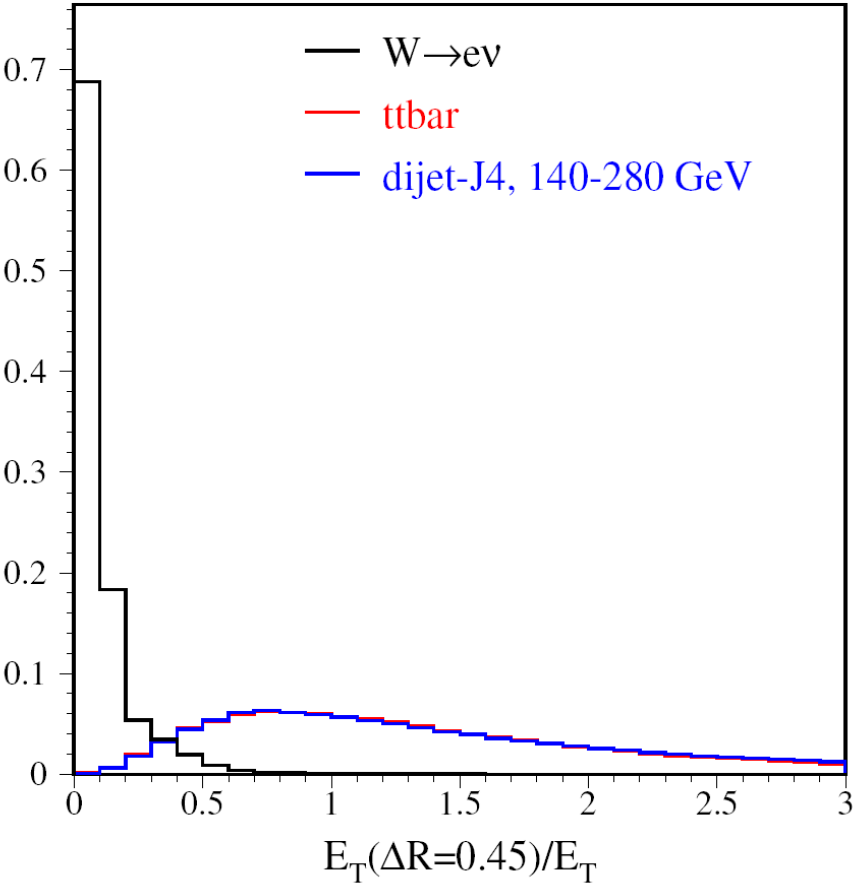
**E/P Ratio of EM Cluster**

# Electron Isolation Variables

**$N_{trk}$ around Electron Track**



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



# BDT e-ID Training (UM)

- 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→e$\nu$ as electron signal
  - Di-jet samples (J0-J6), Pt=[8-1120] GeV
  - ttbar hadronic decays samples (Rel. v12 only)

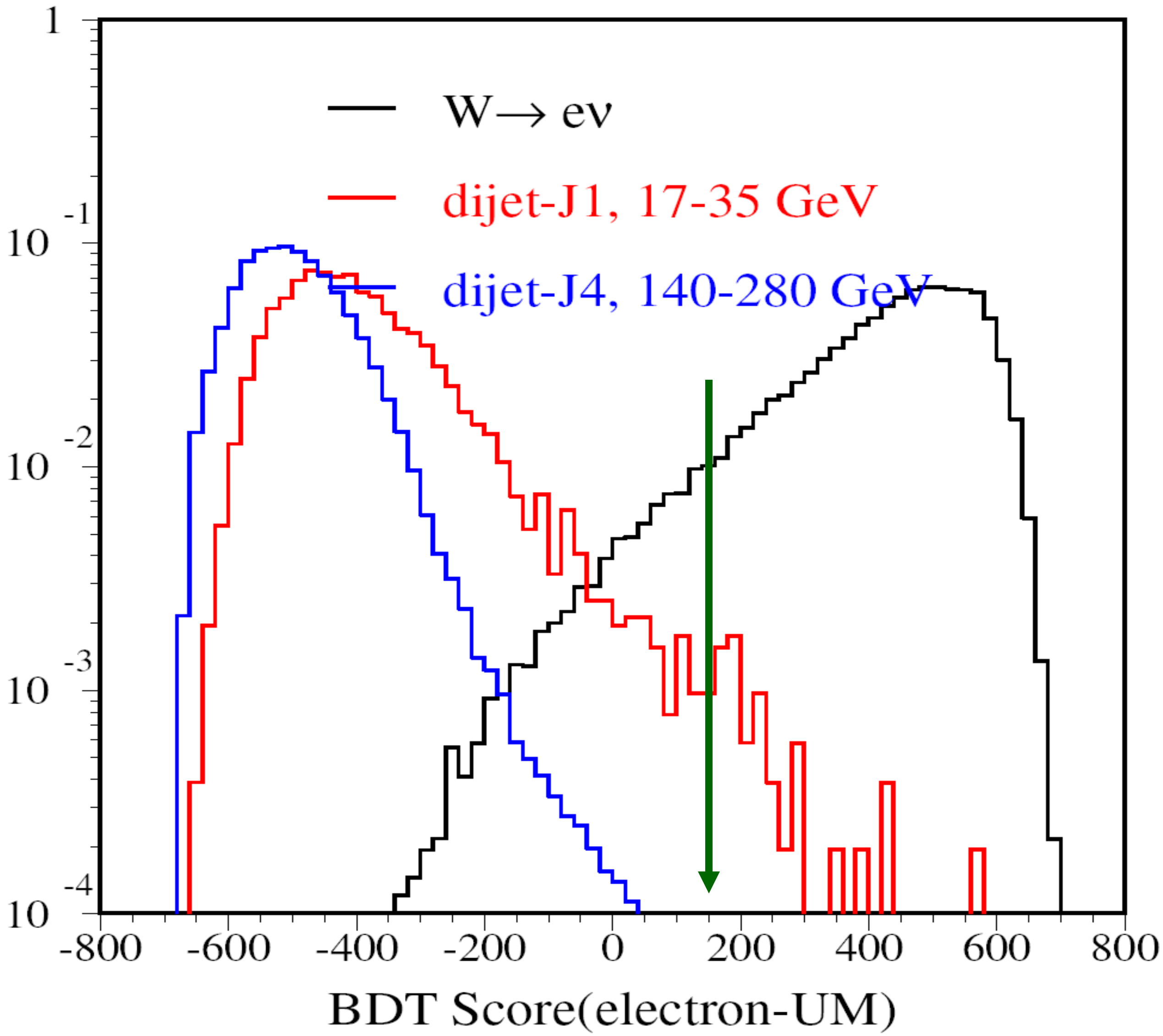
- BDT e-ID training procedure
  - Event weight training based on background cross sections [ H. Yang et. al., JINST 3 P04004 (2008) ]
  - Apply additional cuts on the training samples to select hardly identified jet faked electron as background for BDT training to make the BDT training more effective.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

# Use Independent Samples to Test the BDT e-ID Performance

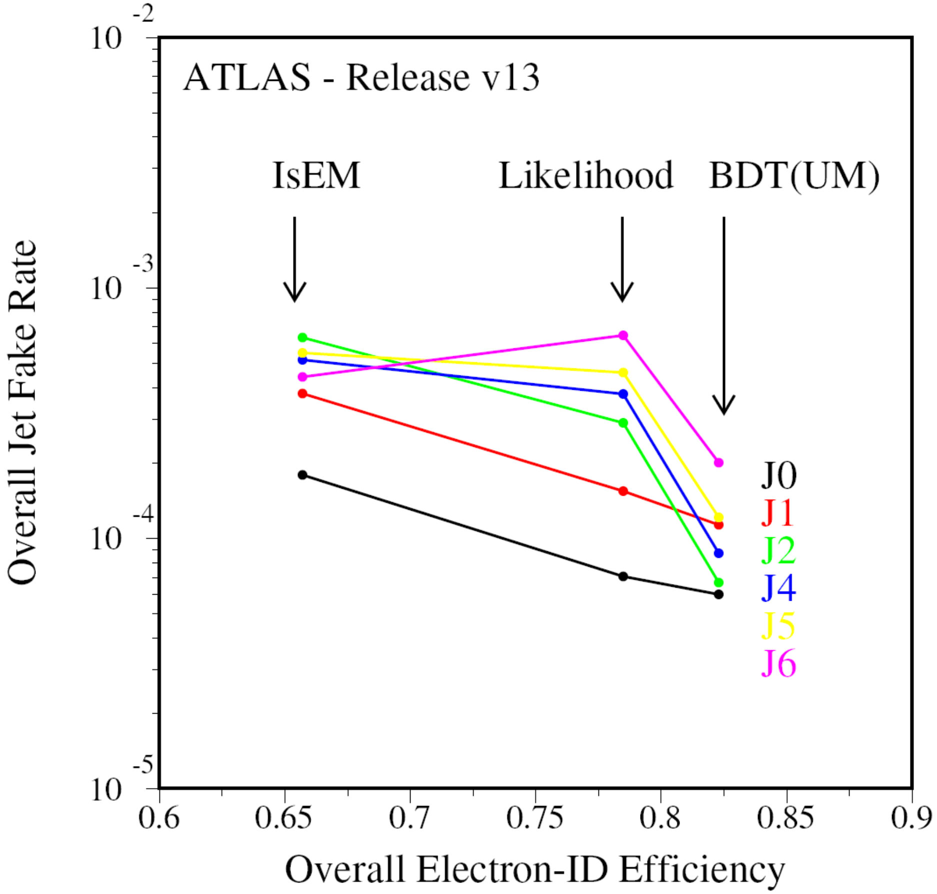
- BDT Test Signal (e) Samples:
  - **W → eν (Rel. v12, v13)**
  - WW → eνμν (Rel. v12)
  - Z → ee (Rel. v12)
  - ZZ → 4l (Rel. v12)

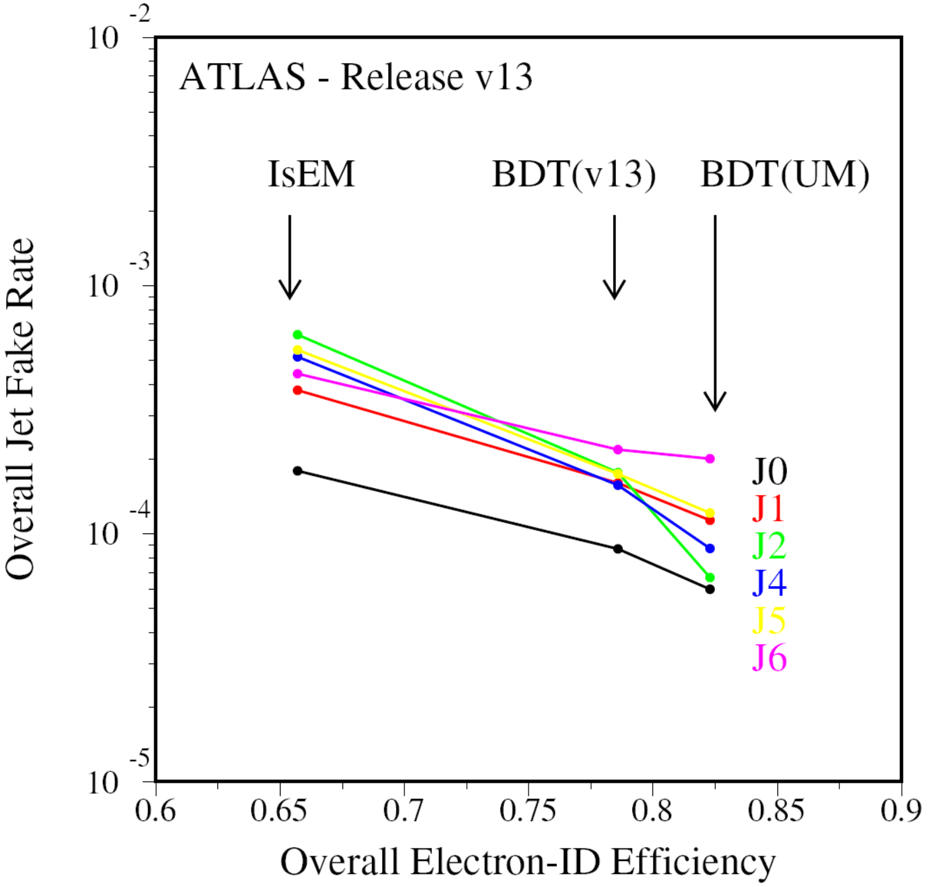
- BDT Test Background (jet faked e) Samples:
  - **Di-jet samples, Pt=[8-1120] GeV (Rel. v12, v13)**
  - ttbar hadronic decays samples (Rel. v12)
  - W→μν + Jets (Rel. v12)
  - Z→μμ + Jets (Rel. v12)

# BDT e-ID discriminator (UM)



# Comparison of e-ID Algorithms

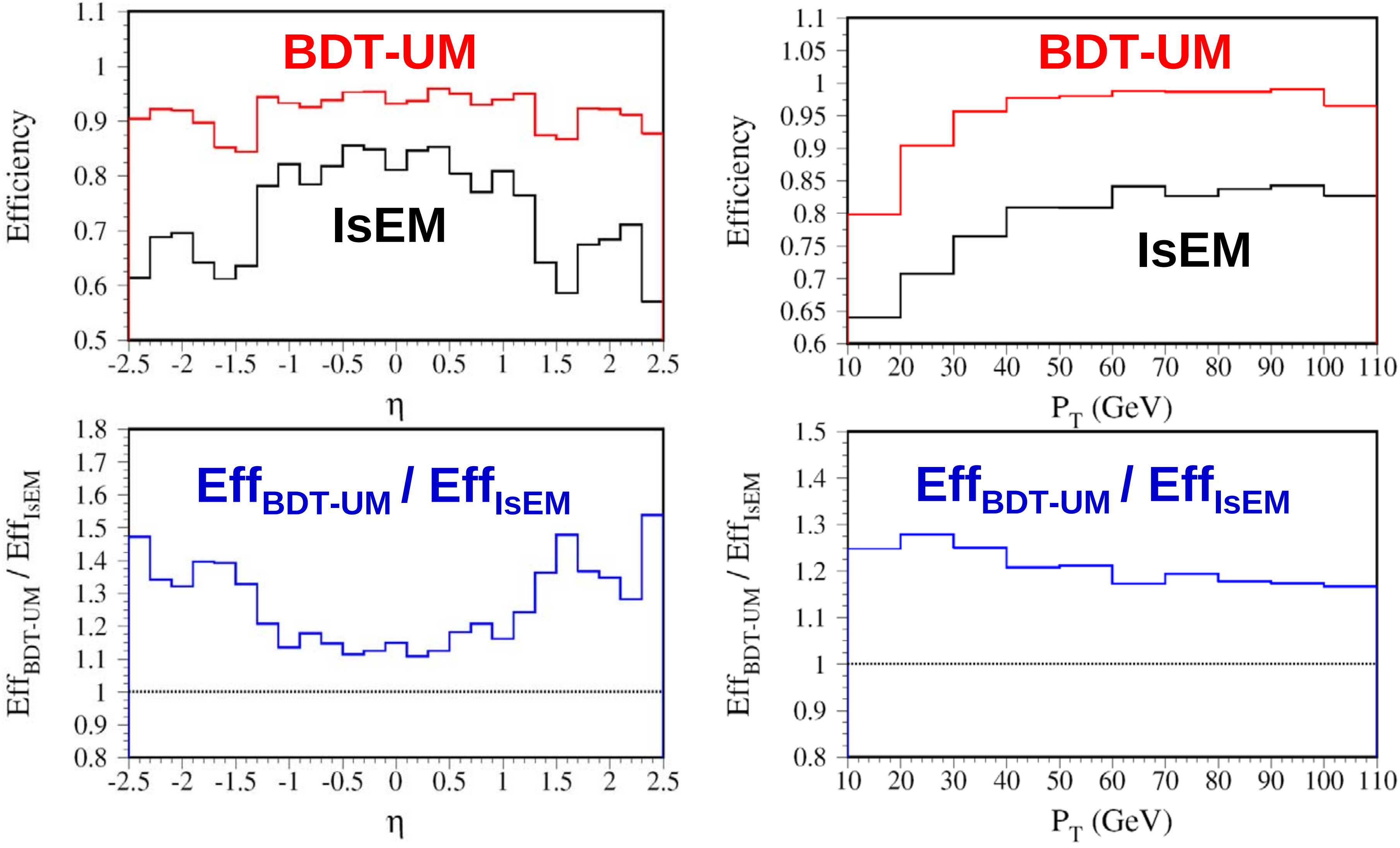




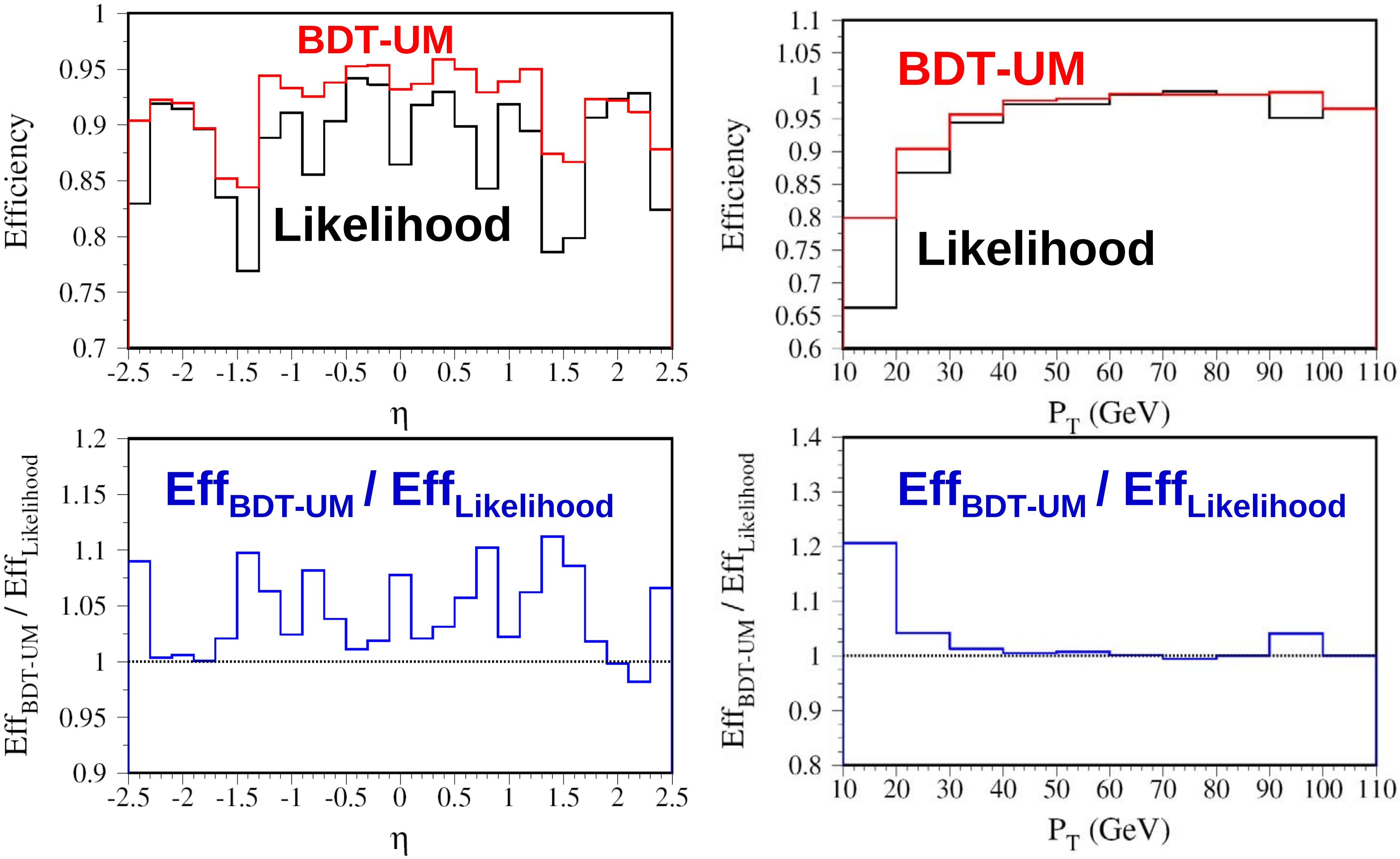
➔ BDTs have high e-ID efficiency and low jet fake rate

➔ BDT (UM) has achieved better performance

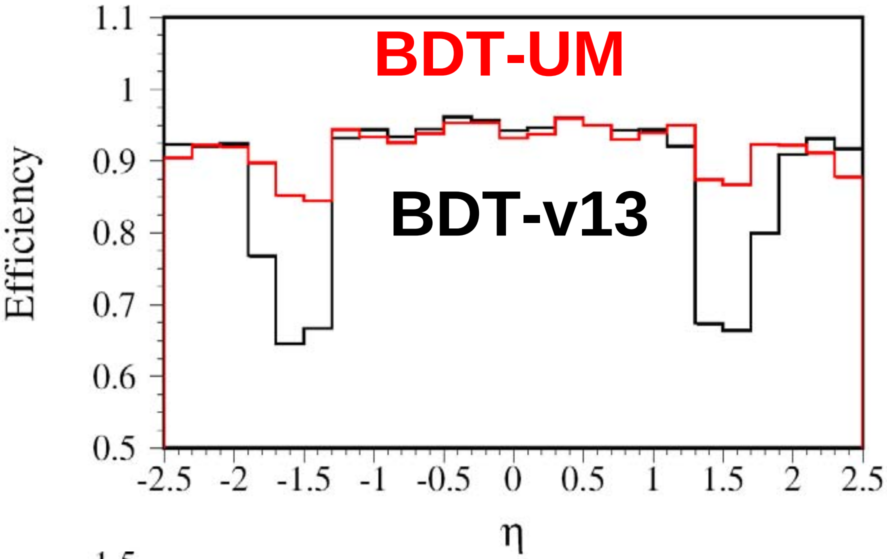
# Comparison of IsEM vs BDT-UM

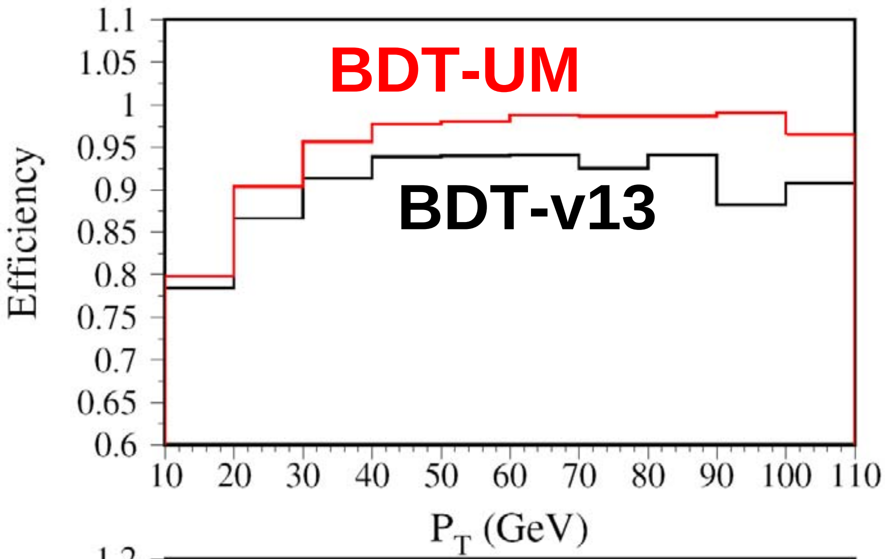


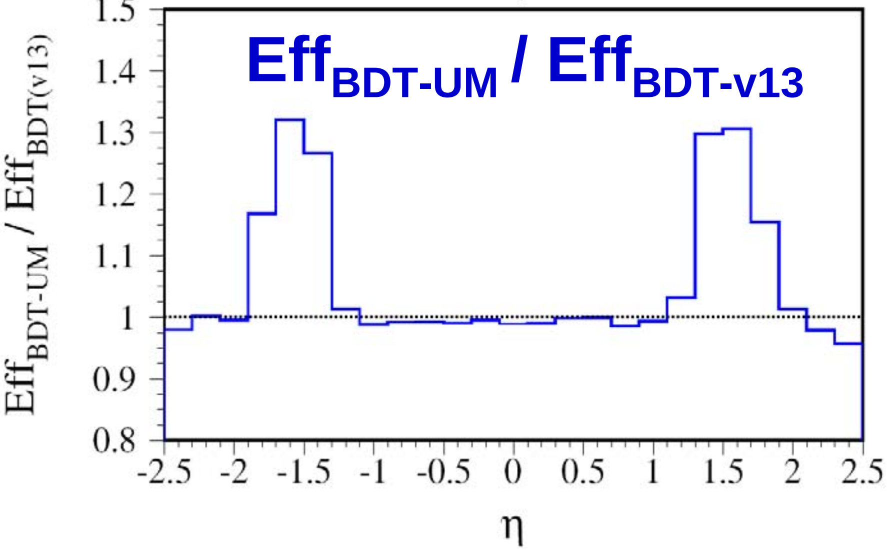
# Comparison of Likelihood vs BDT-UM

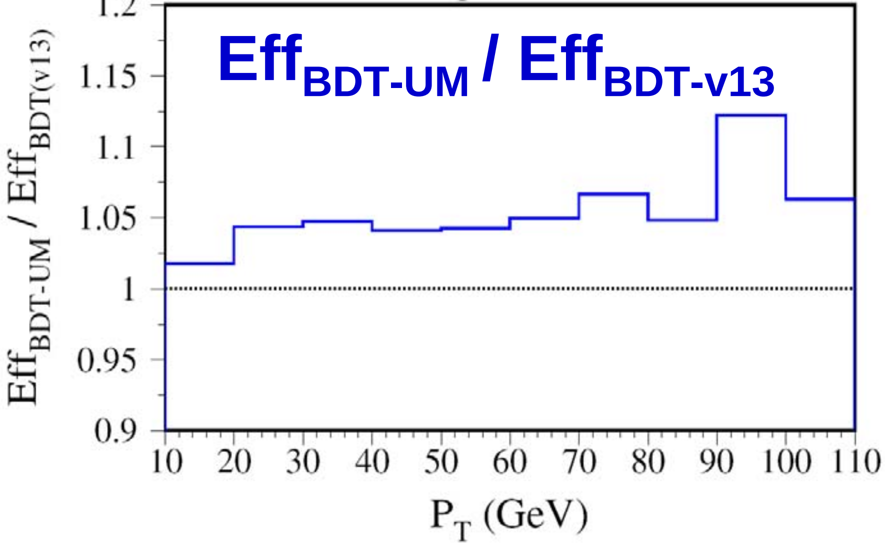


# Comparison of BDT-v13 vs BDT-UM

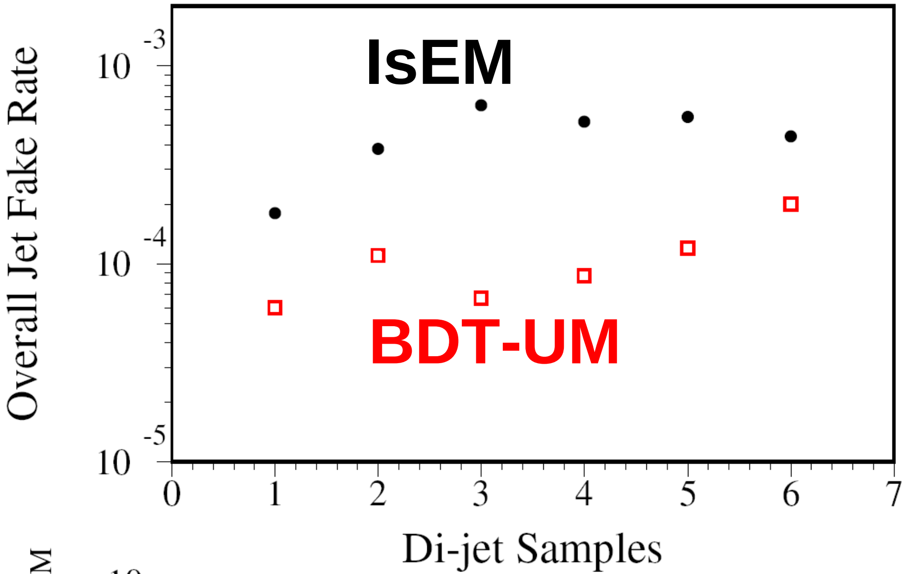


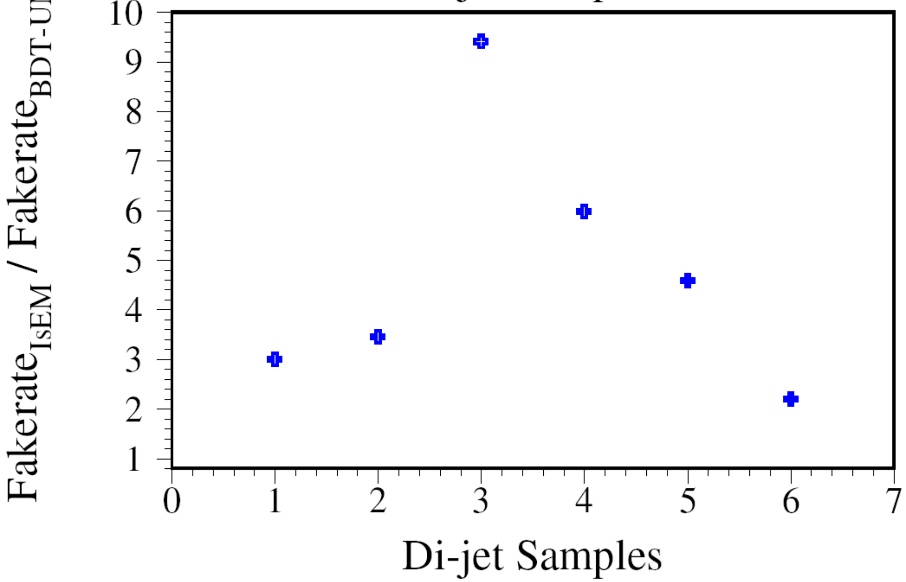






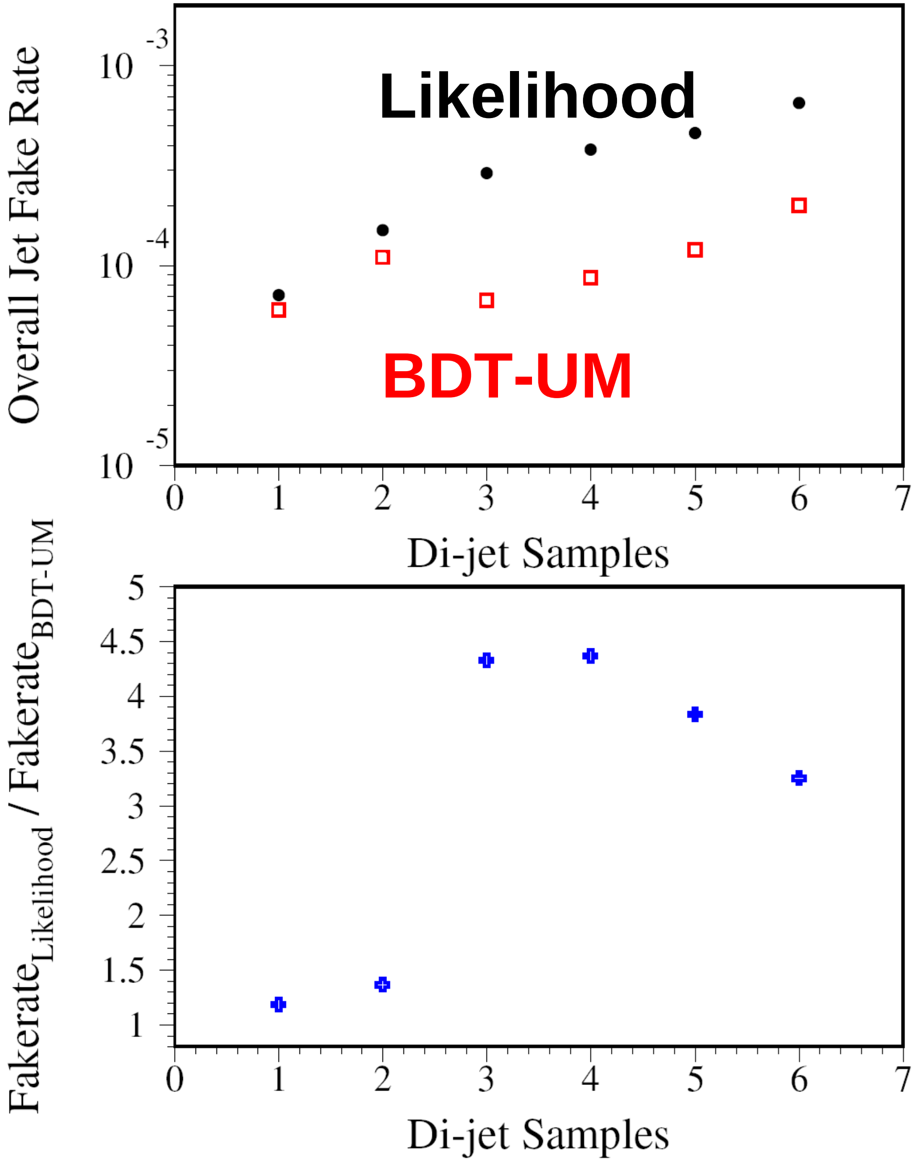
# Jet Fake Rate (IsEM vs BDT-UM)





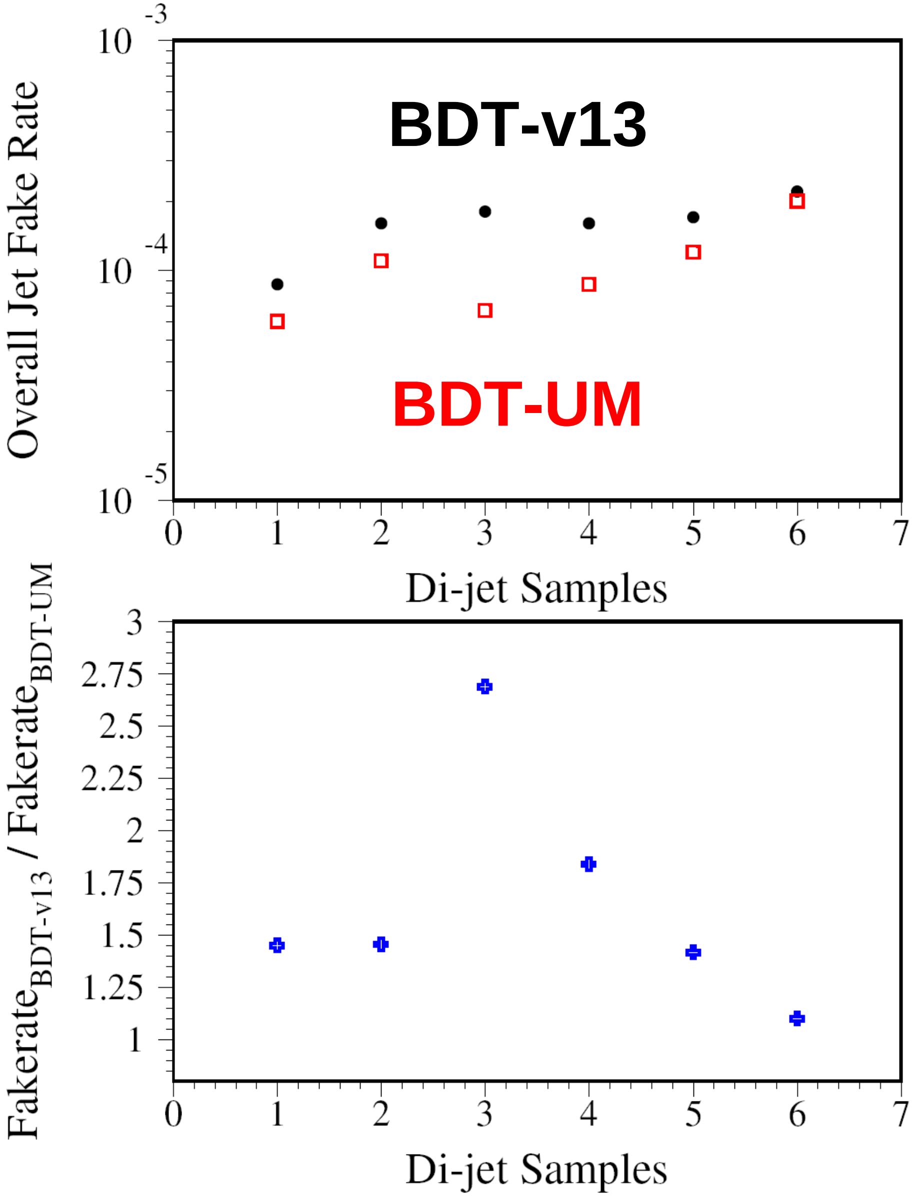
$$\frac{\textbf{Fakerate}_{\textbf{IsEM}}}{\textbf{Fakerate}_{\textbf{BDT-UM}}}$$

# Jet Fake Rate (Likelihood vs BDT-UM)



$$\frac{\textbf{Fakerate}_{\textbf{Likelihood}}}{\textbf{Fakerate}_{\textbf{BDT-UM}}}$$

# Jet Fake Rate (BDT-v13 vs BDT-UM)



$$\frac{\textbf{Fakerate}_{\textbf{BDT-v13}}}{\textbf{Fakerate}_{\textbf{BDT-UM}}}$$

# Overall Electron Efficiency and Fake Rate from Jets ($E_T$ (EM) > 10 GeV)

| From process | IsEM | Likelihood | BDT (Rel. v13) | BDT (U. Michigan) |
|---|---|---|---|---|
| **W → eν (Signal)** | **65.7%** | **78.5%** | **78.6%** | **82.3%** |
| J0: di-jet (8<Pt<17 GeV) | 1.8E-4 | 7.1E-5 | 8.7E-5 | **6.0E-5** |
| J1: di-jet (17<Pt<35 GeV) | 3.8E-4 | 1.5E-4 | 1.6E-4 | **1.1E-4** |
| J2: di-jet (35<Pt<70 GeV) | 6.3E-4 | 2.9E-4 | 1.8E-4 | **6.7E-5** |
| J3: di-jet (70<Pt<140 GeV) | N/A | N/A | N/A | **N/A** |
| J4: di-jet (140<Pt<280 GeV) | 5.2E-4 | 3.8E-4 | 1.6E-4 | **8.7E-5** |
| J5: di-jet (280<Pt<560 GeV) | 5.5E-4 | 4.6E-4 | 1.7E-4 | **1.2E-4** |
| J6: di-jet (560<Pt<1120 GeV) | 4.4E-4 | 6.5E-4 | 2.2E-4 | **2.0E-4** |

# Summary and Future Plan

- Electron ID efficiency can be improved by using BDT multivariate particle identification technique
  - e Eff = 65.7% (IsEM) → 78.5% (LH) →82.3% (BDT).

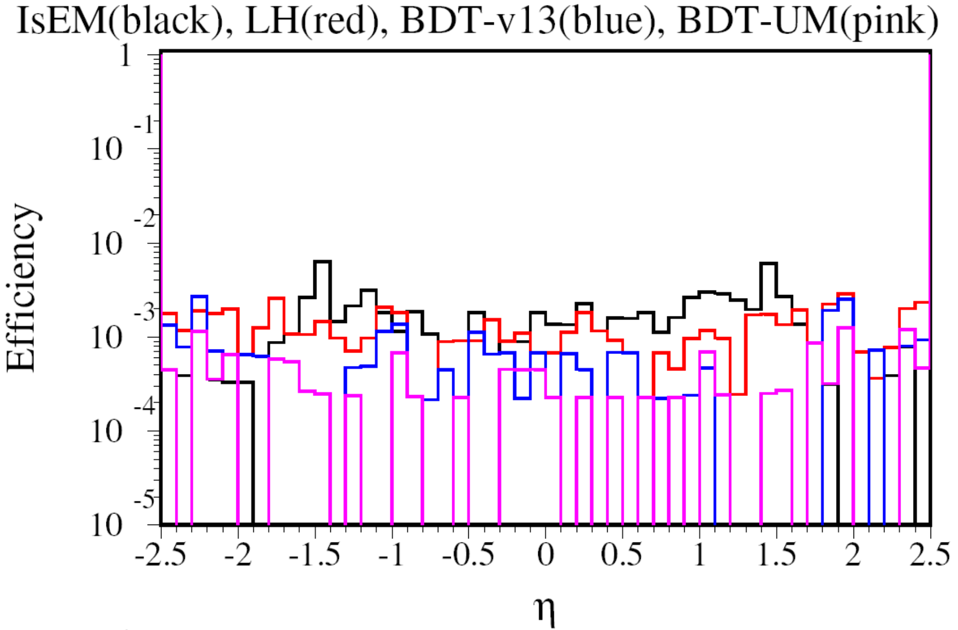
- BDT technique also reduce the jet fake rate

- Incorporate the Electron ID based on BDT into ATLAS official reconstruction package

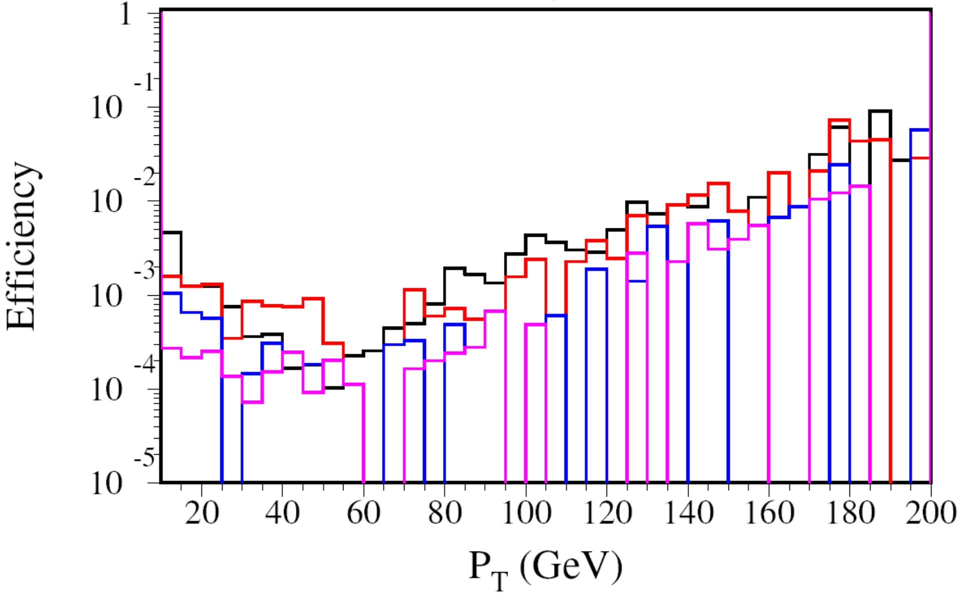
- BDT training with real data:
  - Select electron signals Z→ ee (Tag-Prob)
  - Select fake electron from di-jet samples

# Backup Slides

# Comparison of e-ID Algorithms (v13)



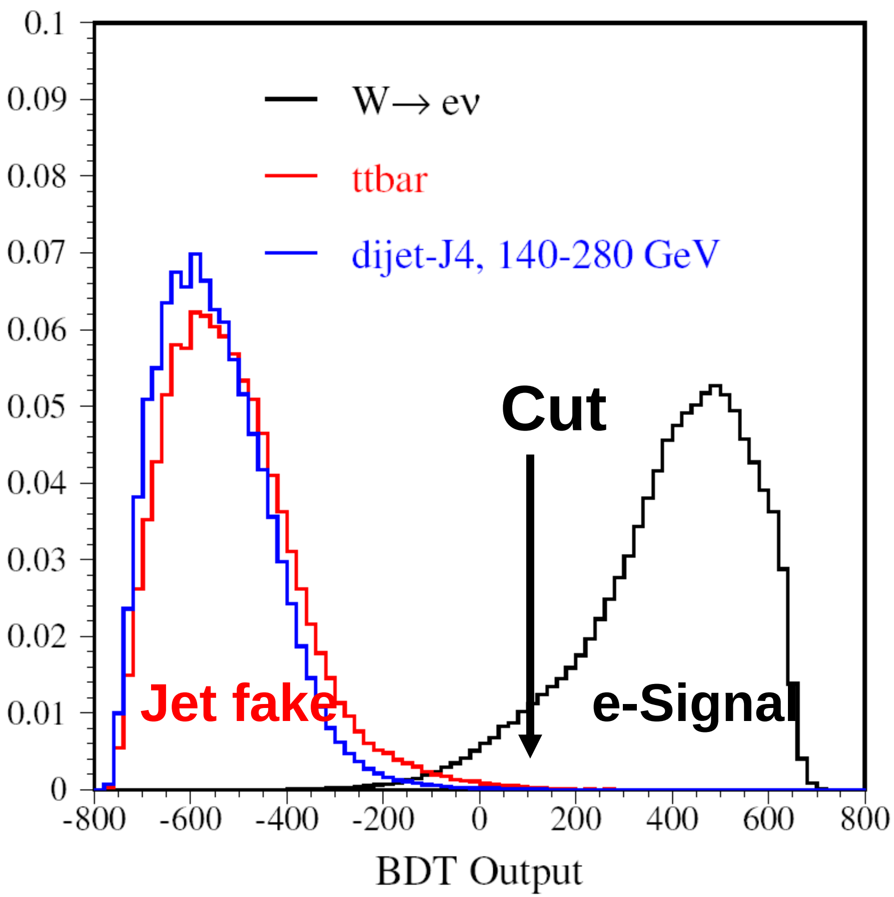
**Dijet: $P_T$=140-280 GeV**



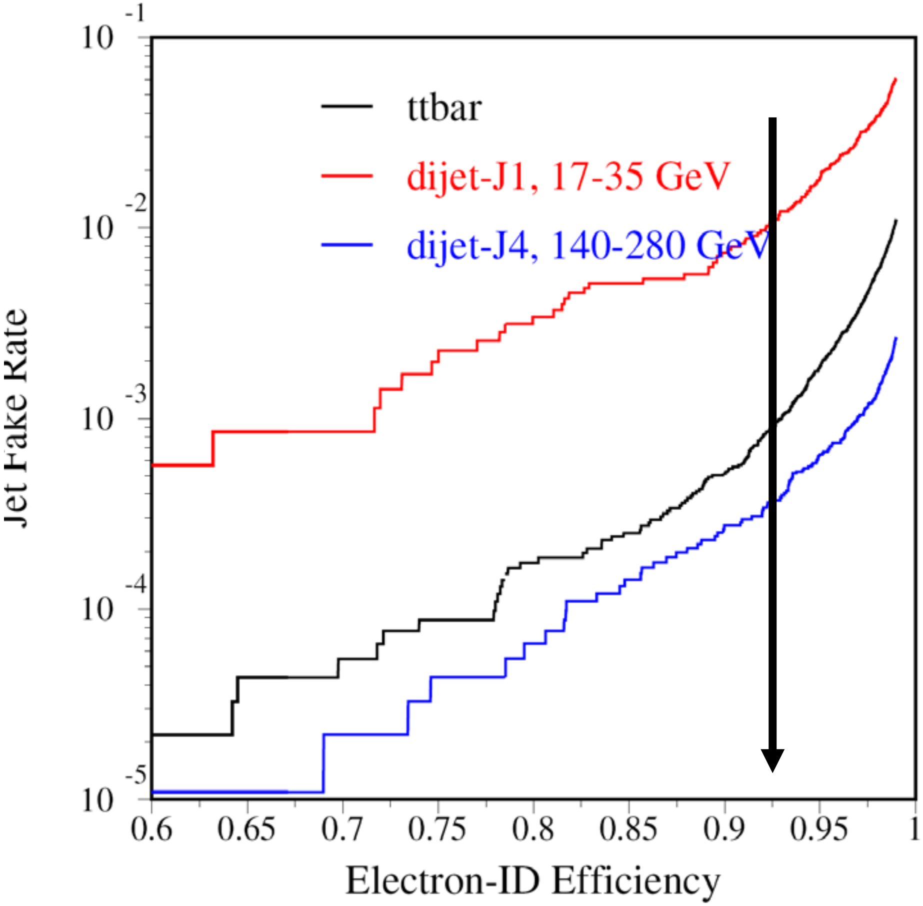
**Dijet: $P_T$=140-280 GeV**

# Performance of The BDT e-ID (v12)

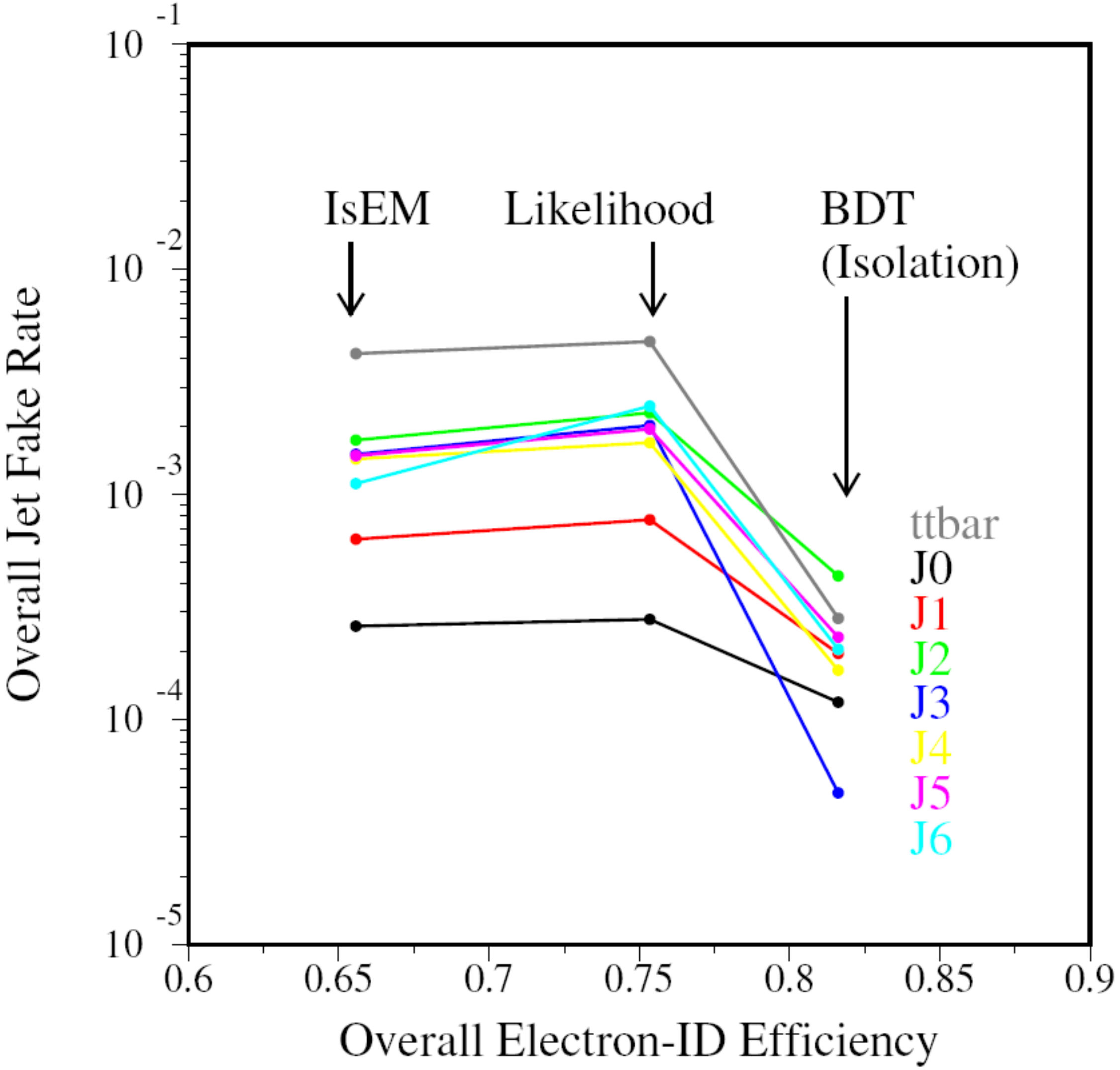
**BDT Output Distribution**



**Jet Fake Rate vs e-ID Eff.**



# Comparison of e-ID Algorithms (v12)



**Di-jet Samples**

J0: Pt = [8-17] GeV
J1: Pt = [17-35] GeV
J2: Pt = [35-70] GeV
J3: Pt = [70-140] GeV
J4: Pt = [140-280] GeV
J5: Pt = [280-560] GeV
J6: Pt = [560-1120] GeV

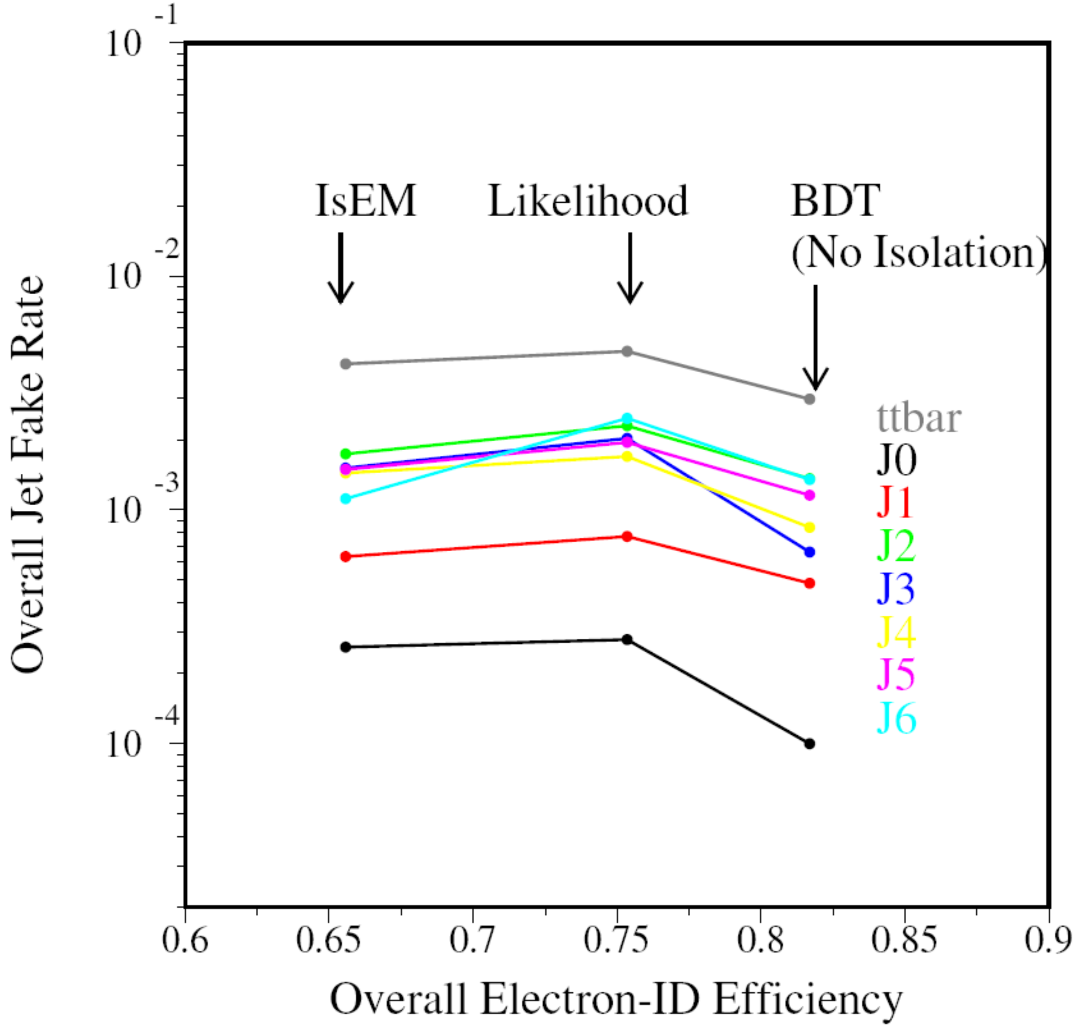
**ttbar**:

All hadronic decays

**BDT e-ID:**

- High efficiency
- Low fake rate

# Comparison of e-ID Algorithms (v12)



**Di-jet Samples**

J0: Pt = [8-17] GeV
J1: Pt = [17-35] GeV
J2: Pt = [35-70] GeV
J3: Pt = [70-140] GeV
J4: Pt = [140-280] GeV
J5: Pt = [280-560] GeV
J6: Pt = [560-1120] GeV

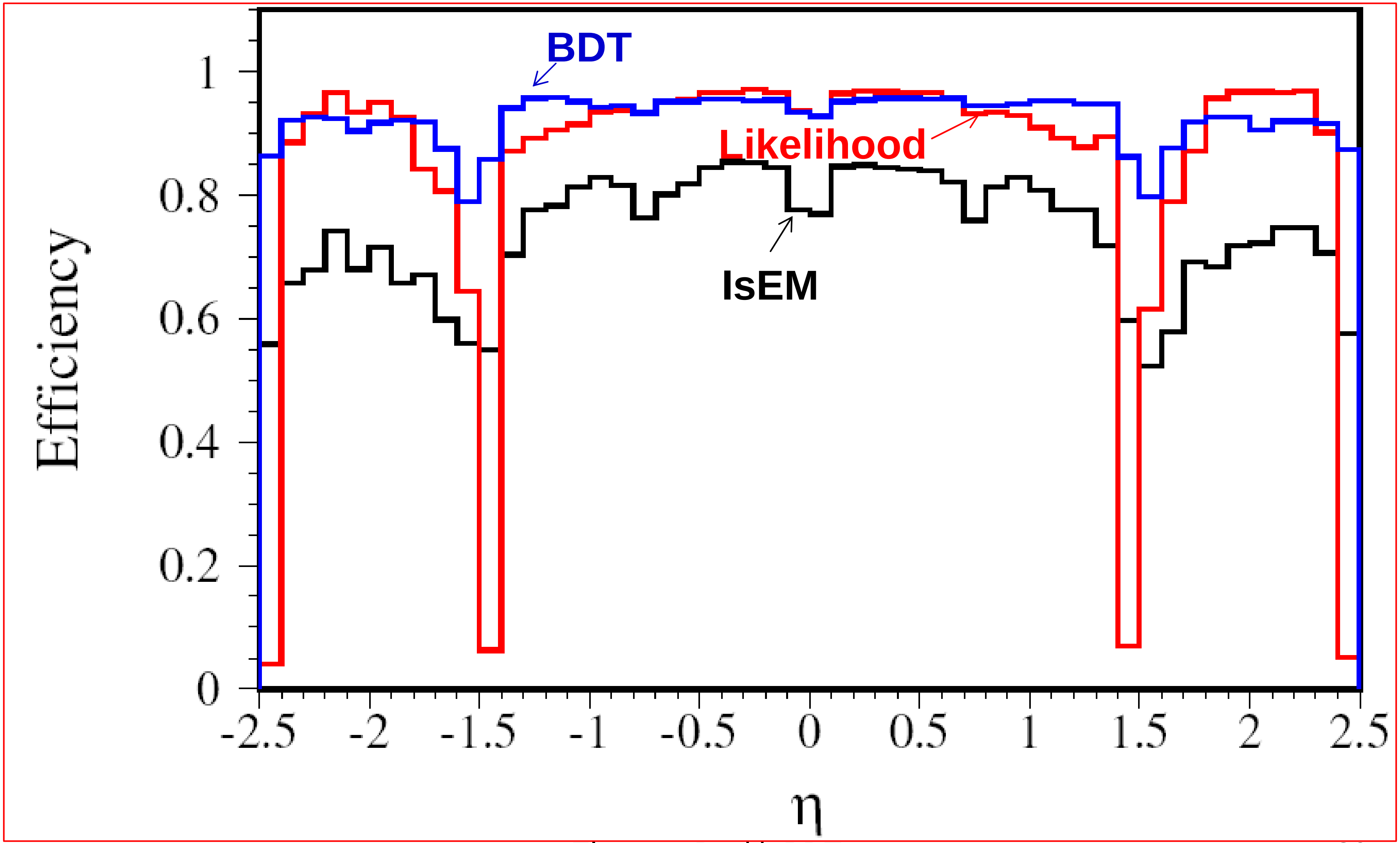
**ttbar**:

All hadronic decays

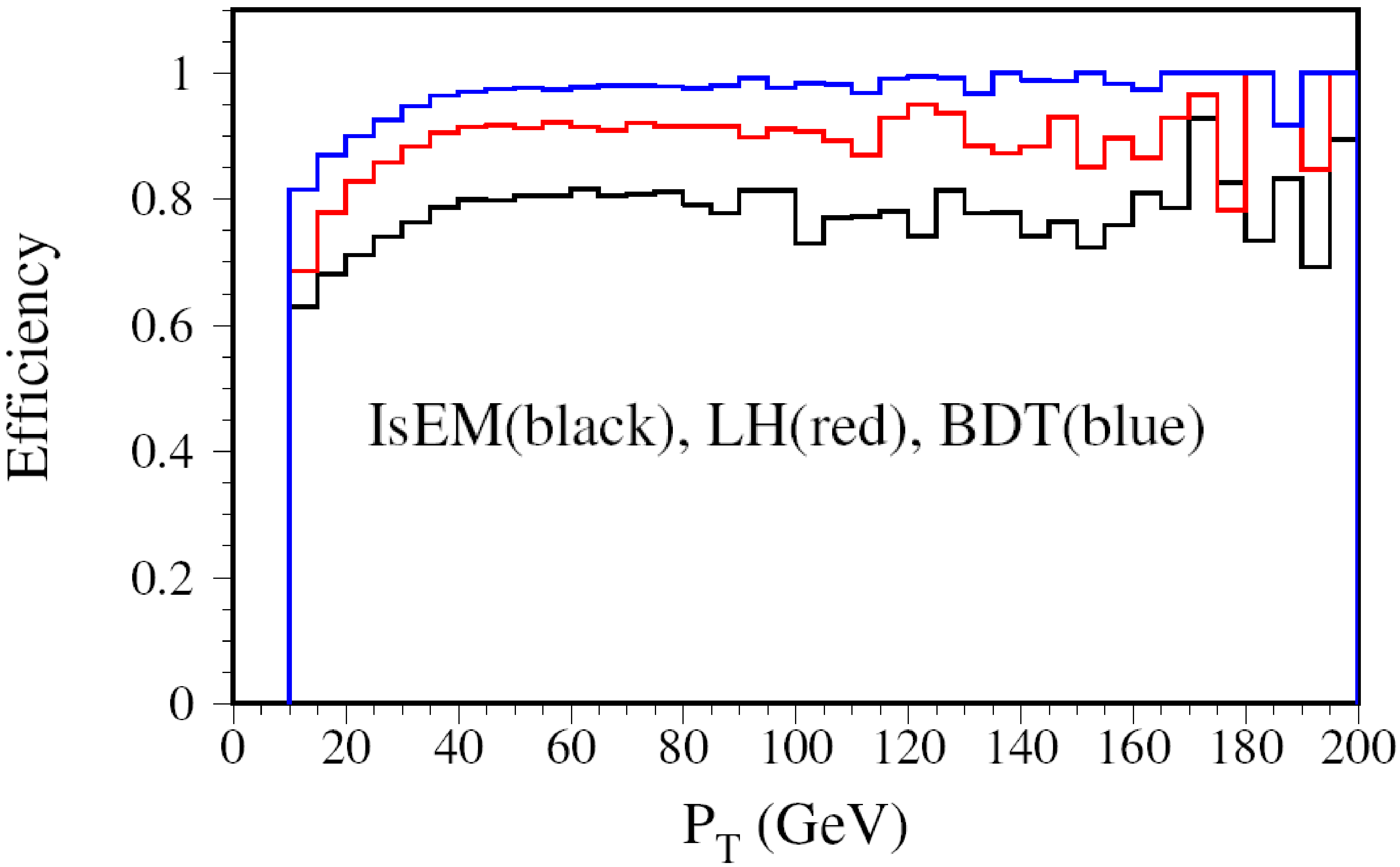
**BDT Results**

– High electron eff
– Low jet fake rate

# Electron ID Eff vs. η (W → eν)



# Electron ID Eff vs $P_T$ ($W \rightarrow e\nu$)



# Overall e-ID Efficiency ($E_T$ > 10 GeV)

| From process | IsEM | Likelihood | BDT (no Isolation) | BDT (Isolation) |
|---|---|---|---|---|
| **W → eν** | **65.6%** | **75.4%** | **81.7%** | **81.6%** |
| Z → ee | 66.7% | 75.8% | 82.6% | 82.4% |
| WW → eνμν | 66.9% | 76.4% | 82.6% | 81.7% |
| ZZ → 4l | 67.5% | 77.0% | 83.1% | 81.4% |
| H → WW → eνμν (140 GeV) | 66.1% | 75.4% | 80.7% | 78.7% |
| H → WW → eνμν (150 GeV) | 66.4% | 76.0% | 81.2% | 78.6% |
| H → WW → eνμν (160 GeV) | 66.8% | 76.7% | 81.9% | 78.6% |
| H → WW → eνμν (165 GeV) | 67.3% | 77.2% | 82.1% | 78.8% |
| H → WW → eνμν (170 GeV) | 67.7% | 77.3% | 82.3% | 79.5% |
| H → WW → eνμν (180 GeV) | 67.7% | 77.5% | 82.4% | 80.1% |

# Overall Electron Fake Rate from Jets

## $E_T$ (EM) > 10 GeV

| From process | IsEM | Likelihood | BDT<br>(no isolation) | BDT<br>(Isolation) |
|---|---|---|---|---|
| J0: di-jet (8<Pt<17 GeV) | 2.6E-4 | 2.8E-4 | 1.0E-4 | 1.0E-4 |
| J1: di-jet (17<Pt<35 GeV) | 6.3E-4 | 7.7E-4 | 4.9E-4 | 2.0E-4 |
| J2: di-jet (35<Pt<70 GeV) | 1.7E-3 | 2.3E-3 | 1.4E-3 | 4.4E-4 |
| J3: di-jet (70<Pt<140 GeV) | 1.5E-3 | 2.0E-3 | 6.6E-4 | 4.7E-5 |
| J4: di-jet (140<Pt<280 GeV) | 1.4E-3 | 1.7E-3 | 8.4E-4 | 1.7E-4 |
| J5: di-jet (280<Pt<560 GeV) | 1.5E-3 | 2.0E-3 | 1.2E-3 | 2.3E-4 |
| J6: di-jet (560<Pt<1120 GeV) | 1.1E-3 | 2.5E-3 | 1.4E-3 | 2.1E-4 |
| ttbar → Wb Wb → all jets | 4.2E-3 | 4.8E-3 | 3.0E-3 | 2.8E-4 |

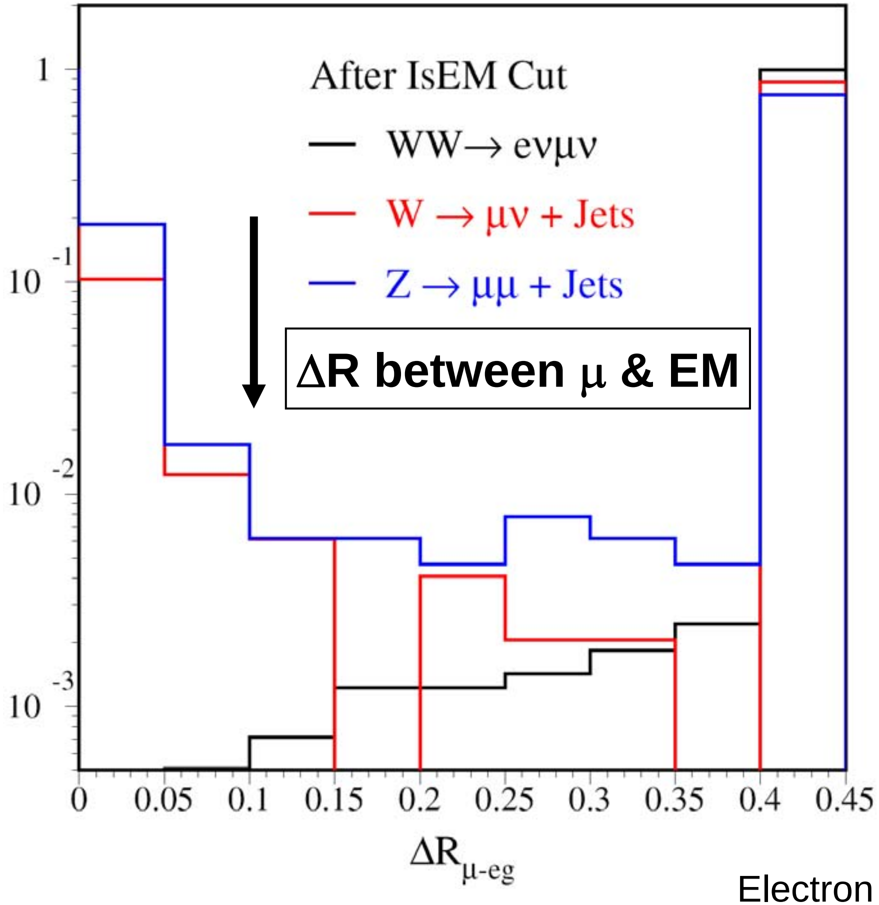
# Overall Electron Fake Rate from μ +Jets Events

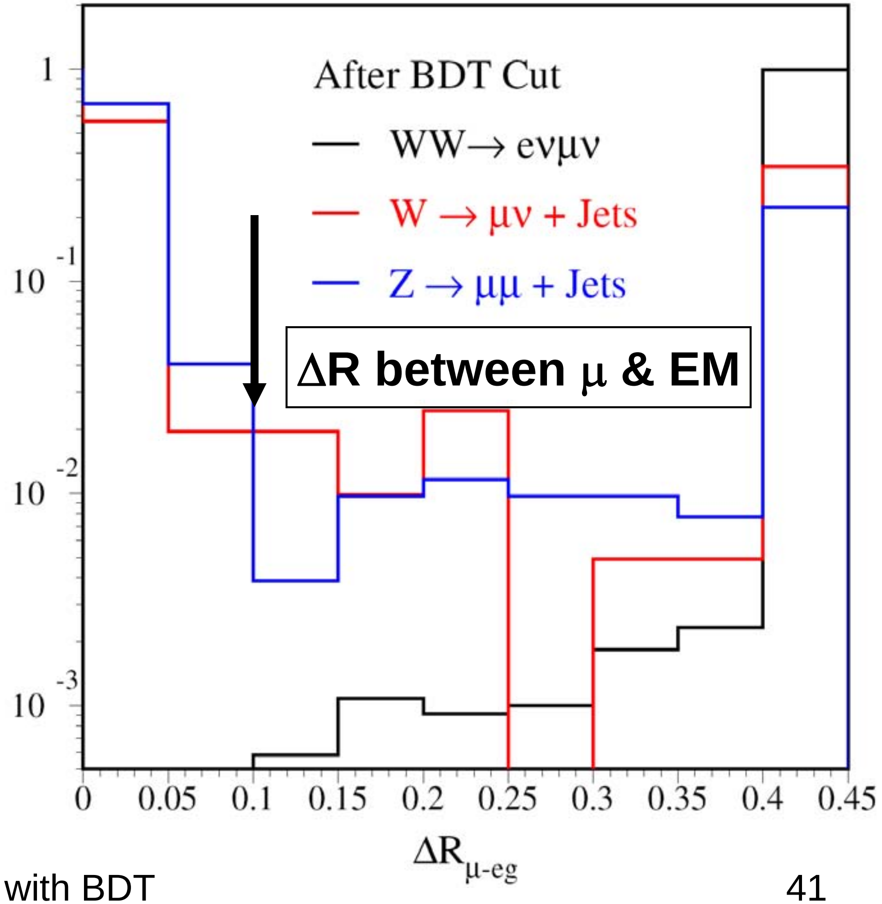
## Why the fake rate increase from single μ to di-μ events?

| From process | IsEM | Likelihood | BDT (no isolation) | BDT (Isolation) |
|---|---|---|---|---|
| W → $\mu\nu$, J1 | 1.6E-3 | 4.8E-3 | 1.7E-3 | 8.2E-4 |
| W → $\mu\nu$, J2 | 2.0E-3 | 4.6E-3 | 1.8E-3 | 9.6E-4 |
| W → $\mu\nu$, J3 | 1.8E-3 | 3.5E-3 | 1.6E-3 | 7.6E-4 |
| W → $\mu\nu$, J4 | 2.0E-3 | 4.0E-3 | 1.6E-3 | 7.8E-4 |
| W → $\mu\nu$, J5 | 2.0E-3 | 3.6E-3 | 1.8E-3 | 6.7E-4 |
| Z → $\mu\mu$, J2 | 2.3E-3 | 6.8E-3 | 2.8E-3 | 2.1E-3 |
| Z → $\mu\mu$, J3 | 2.0E-3 | 6.1E-3 | 2.1E-3 | 1.7E-3 |
| Z → $\mu\mu$, J4 | 2.2E-3 | 5.5E-3 | 2.5E-3 | 1.6E-3 |
| Z → $\mu\mu$, J5 | 2.1E-3 | 5.1E-3 | 2.3E-3 | 1.3E-3 |

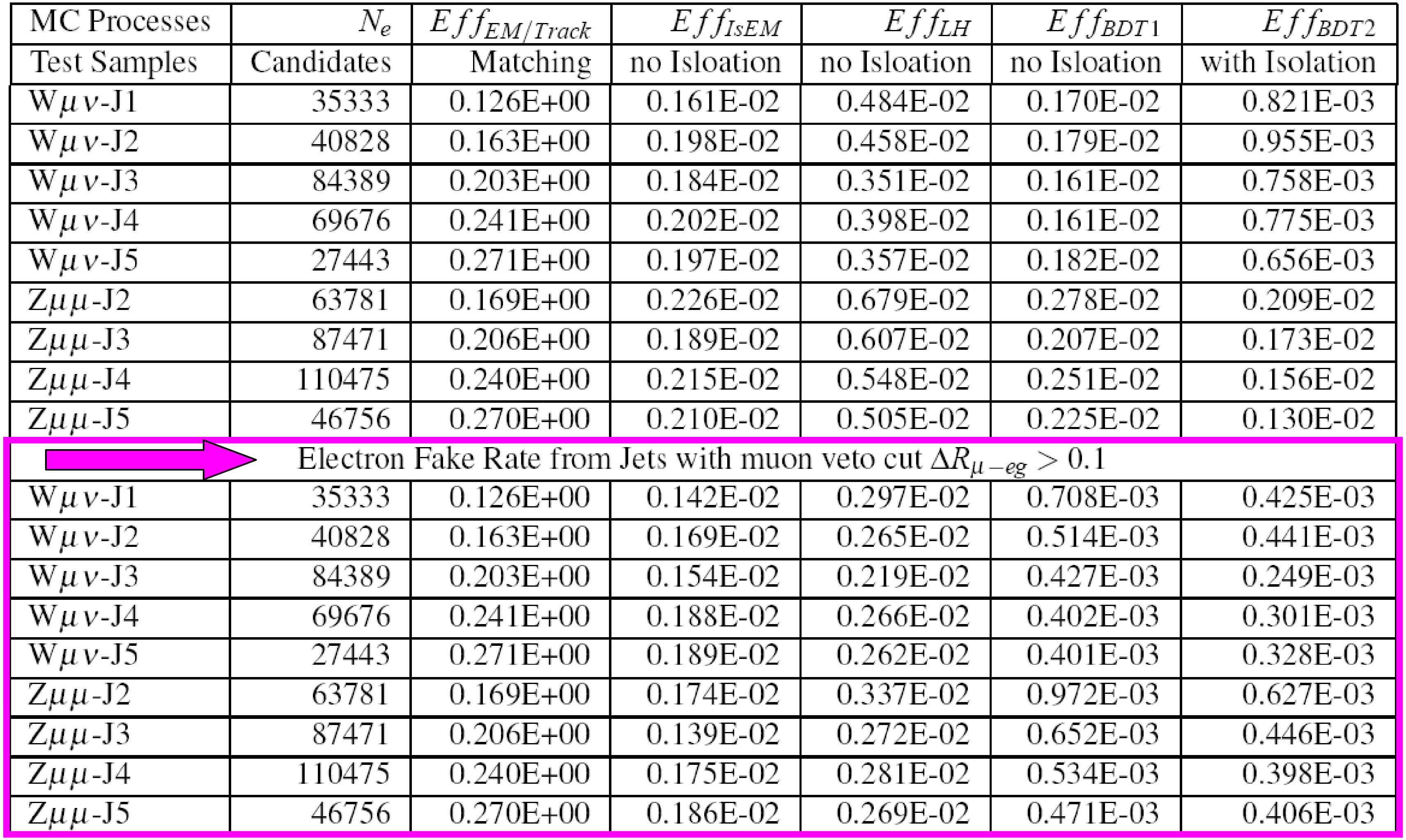
# Fake Electron from an EM Cluster associated with a muon track

**It can be suppressed by requiring ΔR between μ & EM greater than 0.1**





# Fake Electron from an EM Cluster associated with a muon track



| MC Processes<br>Test Samples | $N_e$<br>Candidates | $Eff_{EM/Track}$<br>Matching | $Eff_{IsEM}$<br>no Isloation | $Eff_{LH}$<br>no Isloation | $Eff_{BDT1}$<br>no Isloation | $Eff_{BDT2}$<br>with Isolation |
|---|---|---|---|---|---|---|
| W$\mu\nu$-J1 | 35333 | 0.126E+00 | 0.161E-02 | 0.484E-02 | 0.170E-02 | 0.821E-03 |
| W$\mu\nu$-J2 | 40828 | 0.163E+00 | 0.198E-02 | 0.458E-02 | 0.179E-02 | 0.955E-03 |
| W$\mu\nu$-J3 | 84389 | 0.203E+00 | 0.184E-02 | 0.351E-02 | 0.161E-02 | 0.758E-03 |
| W$\mu\nu$-J4 | 69676 | 0.241E+00 | 0.202E-02 | 0.398E-02 | 0.161E-02 | 0.775E-03 |
| W$\mu\nu$-J5 | 27443 | 0.271E+00 | 0.197E-02 | 0.357E-02 | 0.182E-02 | 0.656E-03 |
| Z$\mu\mu$-J2 | 63781 | 0.169E+00 | 0.226E-02 | 0.679E-02 | 0.278E-02 | 0.209E-02 |
| Z$\mu\mu$-J3 | 87471 | 0.206E+00 | 0.189E-02 | 0.607E-02 | 0.207E-02 | 0.173E-02 |
| Z$\mu\mu$-J4 | 110475 | 0.240E+00 | 0.215E-02 | 0.548E-02 | 0.251E-02 | 0.156E-02 |
| Z$\mu\mu$-J5 | 46756 | 0.270E+00 | 0.210E-02 | 0.505E-02 | 0.225E-02 | 0.130E-02 |
| Electron Fake Rate from Jets with muon veto cut $\Delta R_{\mu-eg} > 0.1$ | | | | | | |
| W$\mu\nu$-J1 | 35333 | 0.126E+00 | 0.142E-02 | 0.297E-02 | 0.708E-03 | 0.425E-03 |
| W$\mu\nu$-J2 | 40828 | 0.163E+00 | 0.169E-02 | 0.265E-02 | 0.514E-03 | 0.441E-03 |
| W$\mu\nu$-J3 | 84389 | 0.203E+00 | 0.154E-02 | 0.219E-02 | 0.427E-03 | 0.249E-03 |
| W$\mu\nu$-J4 | 69676 | 0.241E+00 | 0.188E-02 | 0.266E-02 | 0.402E-03 | 0.301E-03 |
| W$\mu\nu$-J5 | 27443 | 0.271E+00 | 0.189E-02 | 0.262E-02 | 0.401E-03 | 0.328E-03 |
| Z$\mu\mu$-J2 | 63781 | 0.169E+00 | 0.174E-02 | 0.337E-02 | 0.972E-03 | 0.627E-03 |
| Z$\mu\mu$-J3 | 87471 | 0.206E+00 | 0.139E-02 | 0.272E-02 | 0.652E-03 | 0.446E-03 |
| Z$\mu\mu$-J4 | 110475 | 0.240E+00 | 0.175E-02 | 0.281E-02 | 0.534E-03 | 0.398E-03 |
| Z$\mu\mu$-J5 | 46756 | 0.270E+00 | 0.186E-02 | 0.269E-02 | 0.471E-03 | 0.406E-03 |

# Rank of Variables (Gini Index)

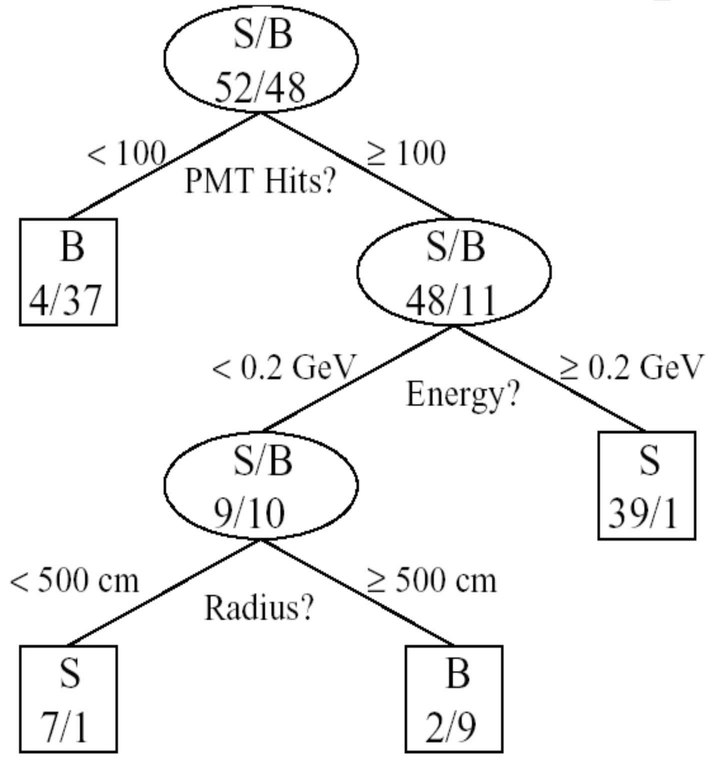
1. Ratio of Et($\Delta$R=0.2-0.45) / Et($\Delta$R=0.2)
2. Number of tracks in $\Delta$R=0.3 cone
3. Energy leakage to hadronic calorimeter
4. EM shower shape E237 / E277
5. $\Delta\eta$ between inner track and EM cluster
6. Ratio of high threshold and all TRT hits
7. $\eta$ of inner track
8. Number of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

# Boosted Decision Trees

➔Relatively new in HEP – MiniBooNE, BaBar, D0(single top discovery), ATLAS

➔Advantages: robust, understand ‘powerful’ variables, relatively transparent, …

**“A procedure that combines many weak classifiers to form a powerful committee”**



## ***BDT Training Process***

- Split data recursively based on input variables until a stopping criterion is reached (e.g. purity, too few events)
- Every event ends up in a “signal” or a “background” leaf
- Misclassified events will be given larger weight in the next decision tree (boosting)

H. Yang et.al. NIM A555 (2005)370, NIM A543 (2005)577, NIM A574(2007) 342

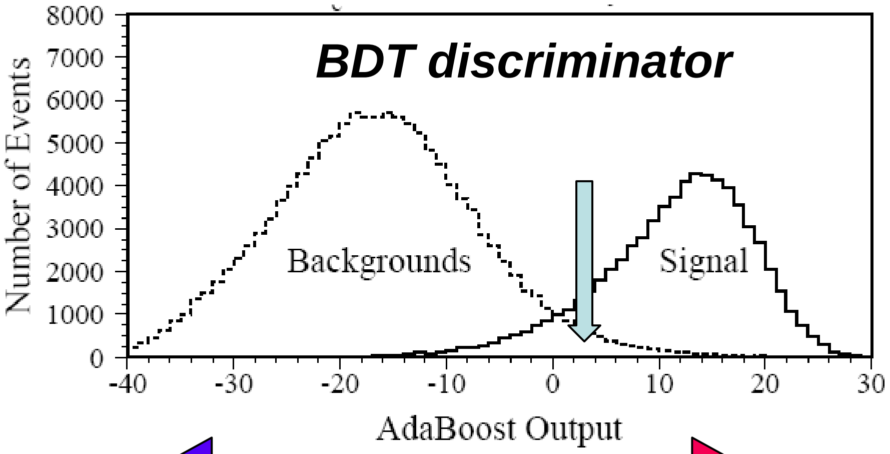
## **A set of decision trees can be developed**,

each re-weighting the events to enhance
identification of backgrounds misidentified
by earlier trees (“boosting”)

For each tree, the data event is assigned

+1 if it is identified as signal,
- 1 if it is identified as background.

The total for all trees is combined into a “**score**”



Background-like negative positive signal-like