Electron Identification Based on Boosted Decision Trees

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

US-ATLAS e/γ Jamboree
September 10, 2008

Motivation

- Lepton (e, μ, τ) Identification is crucial for new physics discoveries at the LHC, such as $H \rightarrow ZZ \rightarrow 4$ leptons, $H \rightarrow WW \rightarrow$ 2 leptons + MET etc.
- ATLAS default electron-ID (IsEM) has relatively low efficiency (~67%), which has significant impact on ATLAS early discovery potential in $H \rightarrow WW \rightarrow l\nu l\nu$ detection (see example next page)
- 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.

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

Electron Identification Studies

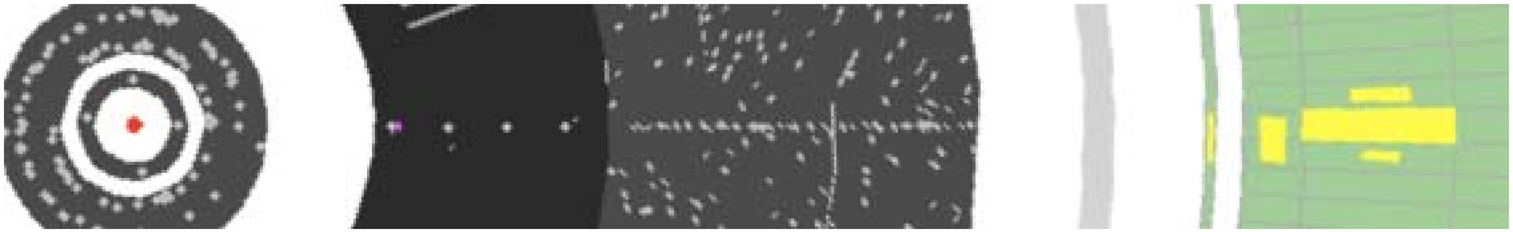
- Pre-selection: an EM cluster matching a track
- Performance based on existing ATLAS e-ID algorithms: IsEM and Likelihood(LH)
- BDT development for e-ID and compare to IsEM and LH

MC samples:

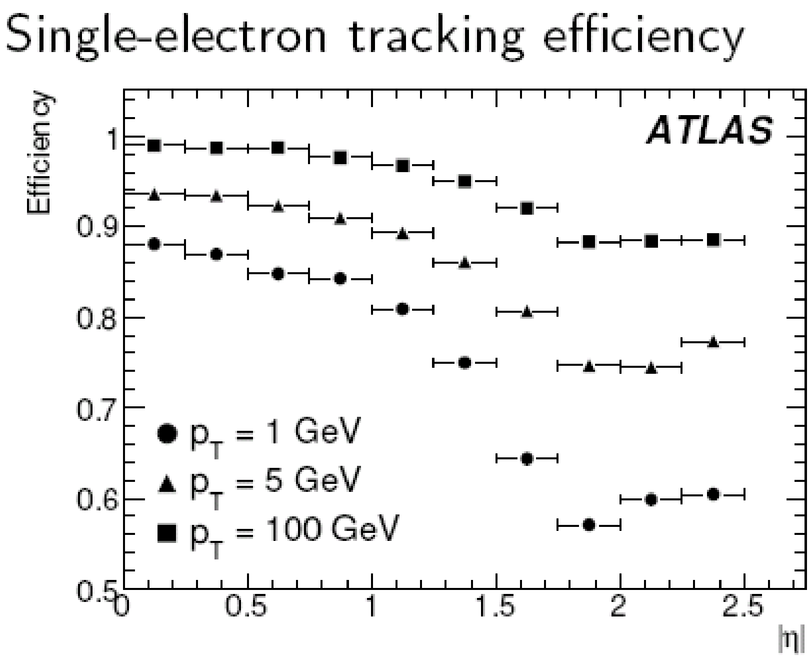
- Signal: electrons from W, Z, WW, ZZ and H→WW→$l\nu l\nu$
  - Using MC truth electron compare to the reconstructed electron to determine the efficiency, and compare the e-ID efficiency based on IsEM and LH to BDT
- Background: di-jets (Et: 8 – 1120 GeV); and ttbar → all jets, W(→$\mu\nu$)+Jets, Z(→$\mu\mu$)+Jets
  - First find EM/track objects in jet events
  - Applying e-ID (IsEM, LH, and BDT) algorithm to determine the fake electron rates from jets

e/γ Identification in Reconstruction

electron reconstructed in tracker and ECAL



pixel SCT TRT Sol LArEM

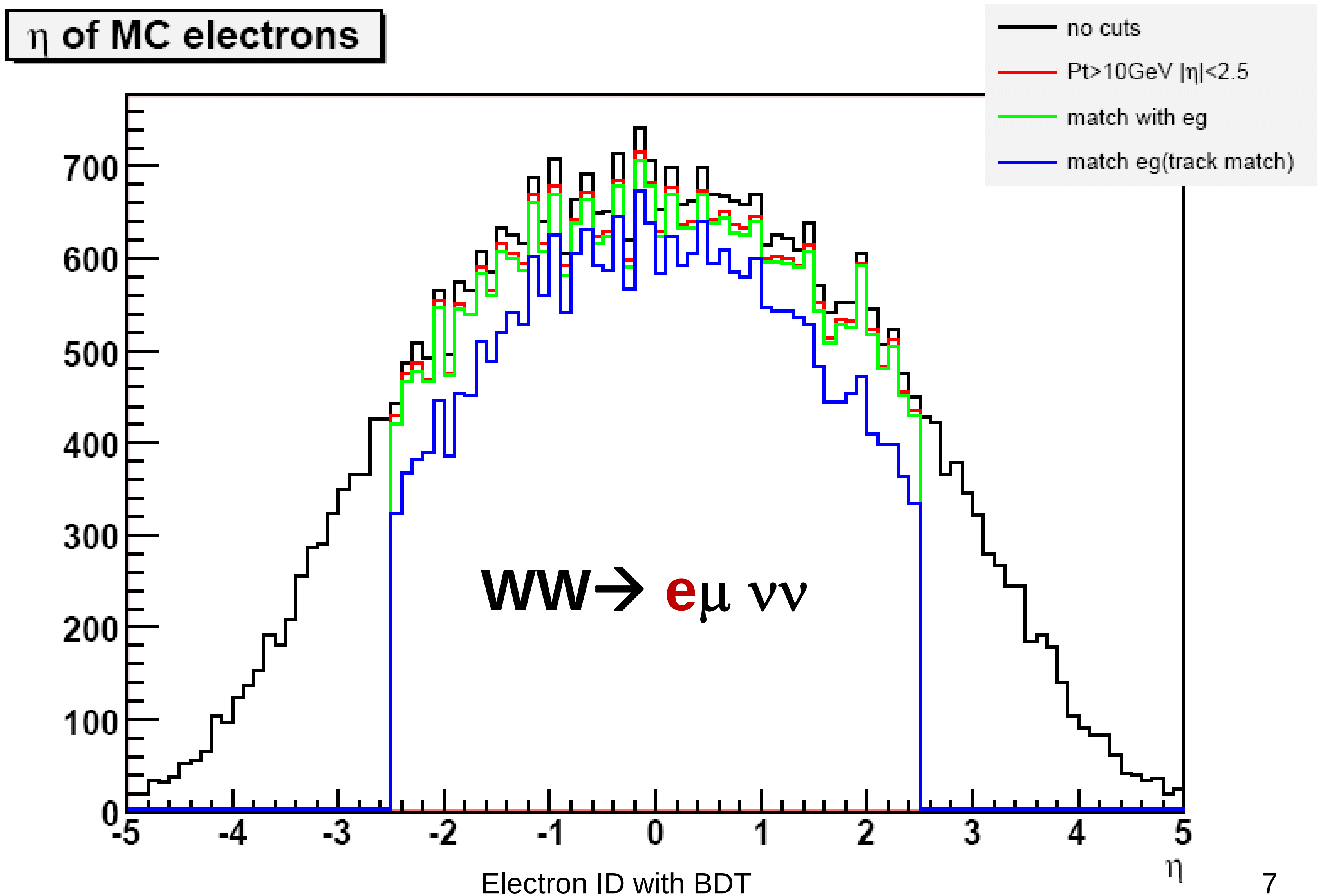


- An electron is reconstructed by matching an EM cluster with an inner detector track. Shower shape analysis is done in the calorimeter.
- The electron is identified by different algorithms using a set of variables:
  - ➢ Simple cuts on those variables: IsEM
  - ➢ Multivariate: likelihood ratio
  - ➢ ***Boosted Decision Trees (this talk)***

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 = $N_{EM/Track} / N_e$

Electrons



Electron ID with BDT

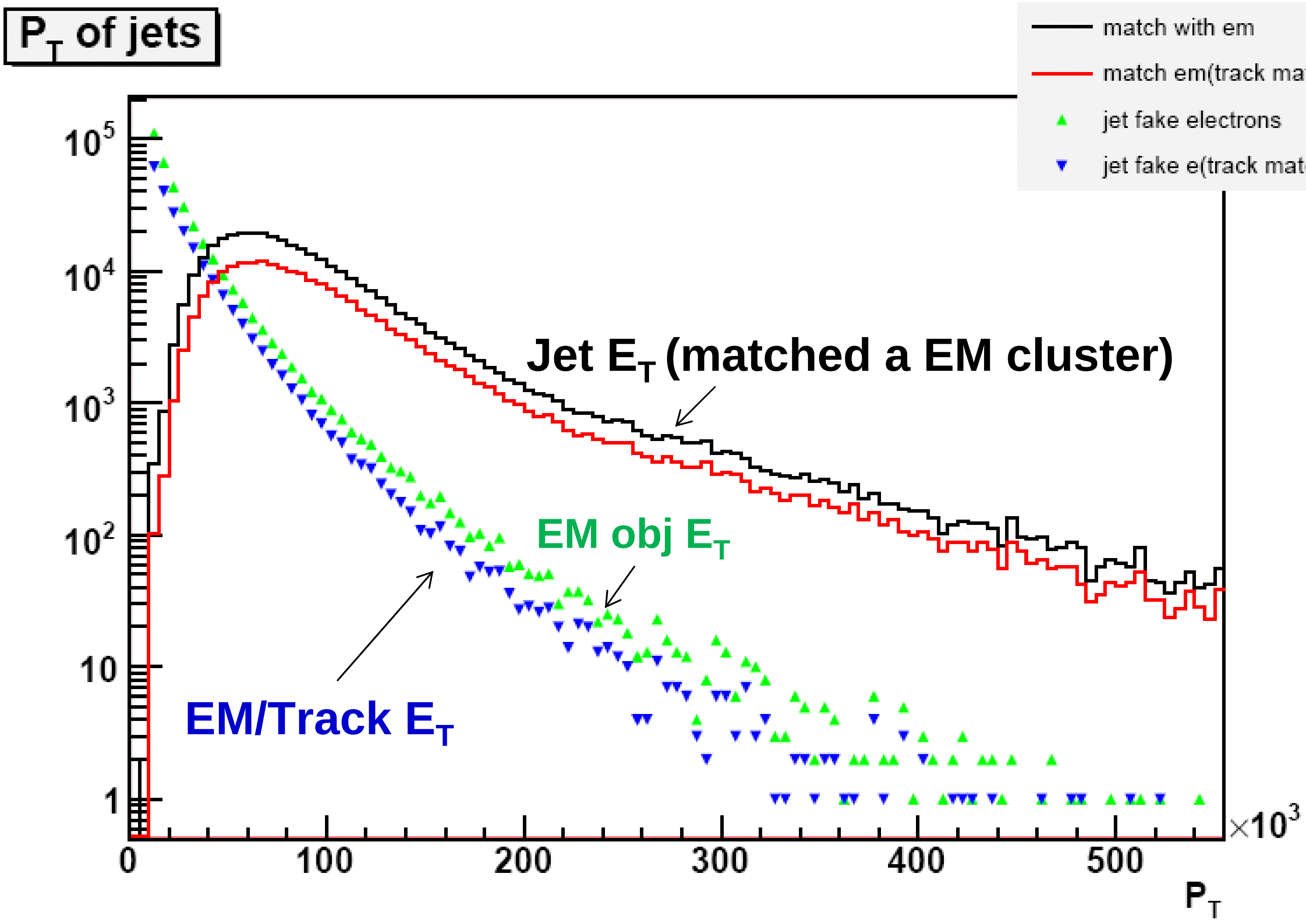
Electron Pre-selection Efficiency

From process	EM Cluster Match	Inner Track Match
W → eν (N_e = 485489)	**99.2%**	**88.2%**
Z → ee (N_e = 29383)	98.5%	87.3%
WW → eνμν (N_e = 39822)	98.9%	87.8%
ZZ → 4l (N_e = 97928)	98.1%	87.4%
H → WW → eνμν (140 GeV)	98.6%	87.5%
H → WW → eνμν (150 GeV)	98.5%	87.3%
H → WW → eνμν (160 GeV)	98.3%	87.3%
H → WW → eνμν (165 GeV)	98.4%	87.4%
H → WW → eνμν (170 GeV)	98.4%	87.5%
H → WW → eνμν (180 GeV)	98.5%	87.4%

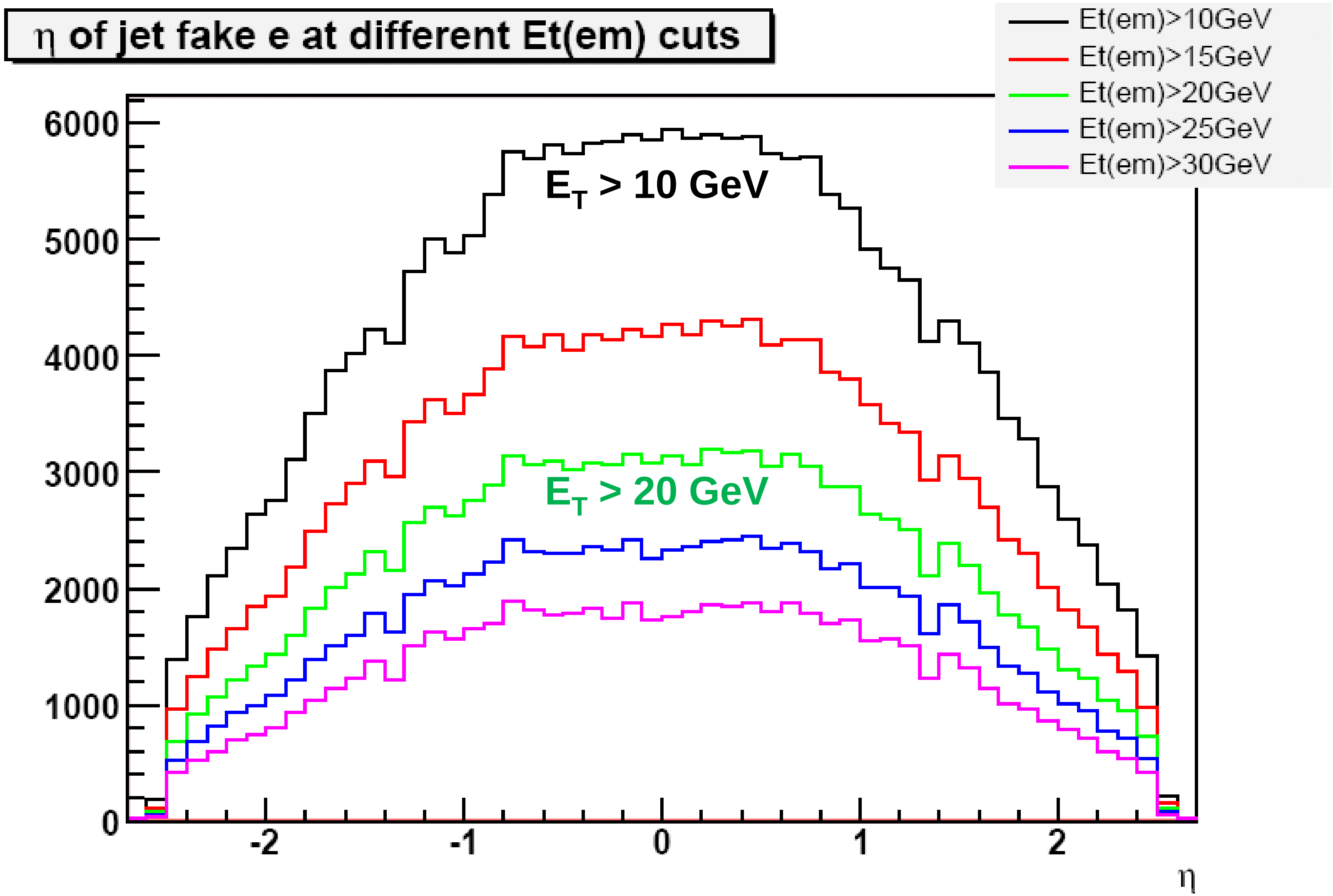
Pre-selection of Jet Faked Electrons

- 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	EM Cluster Match	Inner Track Match
J0: di-jet (8<Pt<17 GeV)	1.4E-2	6.0E-3
J1: di-jet (17<Pt<35 GeV)	3.7E-2	1.5E-2
J2: di-jet (35<Pt<70 GeV)	2.1E-1	1.1E-1
J3: di-jet (70<Pt<140 GeV)	5.3E-1	3.2E-1
J4: di-jet (140<Pt<280 GeV)	6.6E-1	4.3E-1
J5: di-jet (280<Pt<560 GeV)	7.6E-1	5.1E-1
J6: di-jet (560<Pt<1120 GeV)	8.0E-1	5.0E-1
ttbar → Wb Wb → all jets	5.1E-1	3.2E-1

Electron Identification

Based on Pre-selection

- Use the existing ATLAS e-ID algorithms, IsEM and Likelihood to check the e-ID efficiencies and the jet fake rate

- Develop and apply the Boosted Decision Trees Technique for e-ID and test the performance

- Comparison of the performance for three different e-ID methods

Existing ATLAS e-ID Algorithms

IsEM

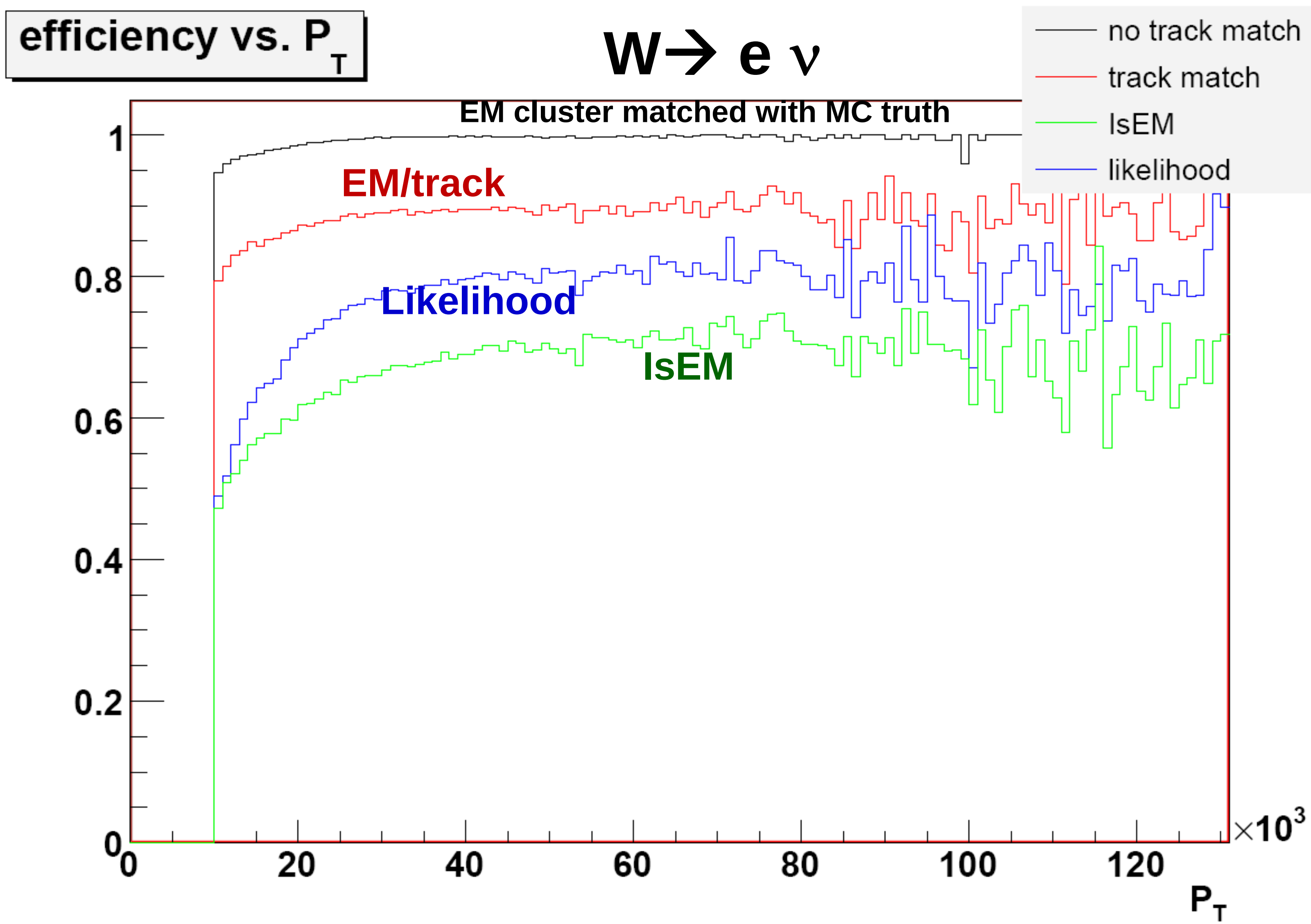
0x2	Only had. Leak	0xF00	Only track cuts
0x4	Only 2nd sampling	0xFFD	All but had. Leak
0x8	Only 1st sampling	0xFFB	All but 2^{nd} sampling
0xFF	Only Ecal	0xFF7	All but 1^{st} sampling
0x200	Only track quality	0xDFF	All but track quality
0x400	E/P	0xBFF	All but E/P
0x800	Only TRT	**0x7FF**	**All but TRT**

Likelihood

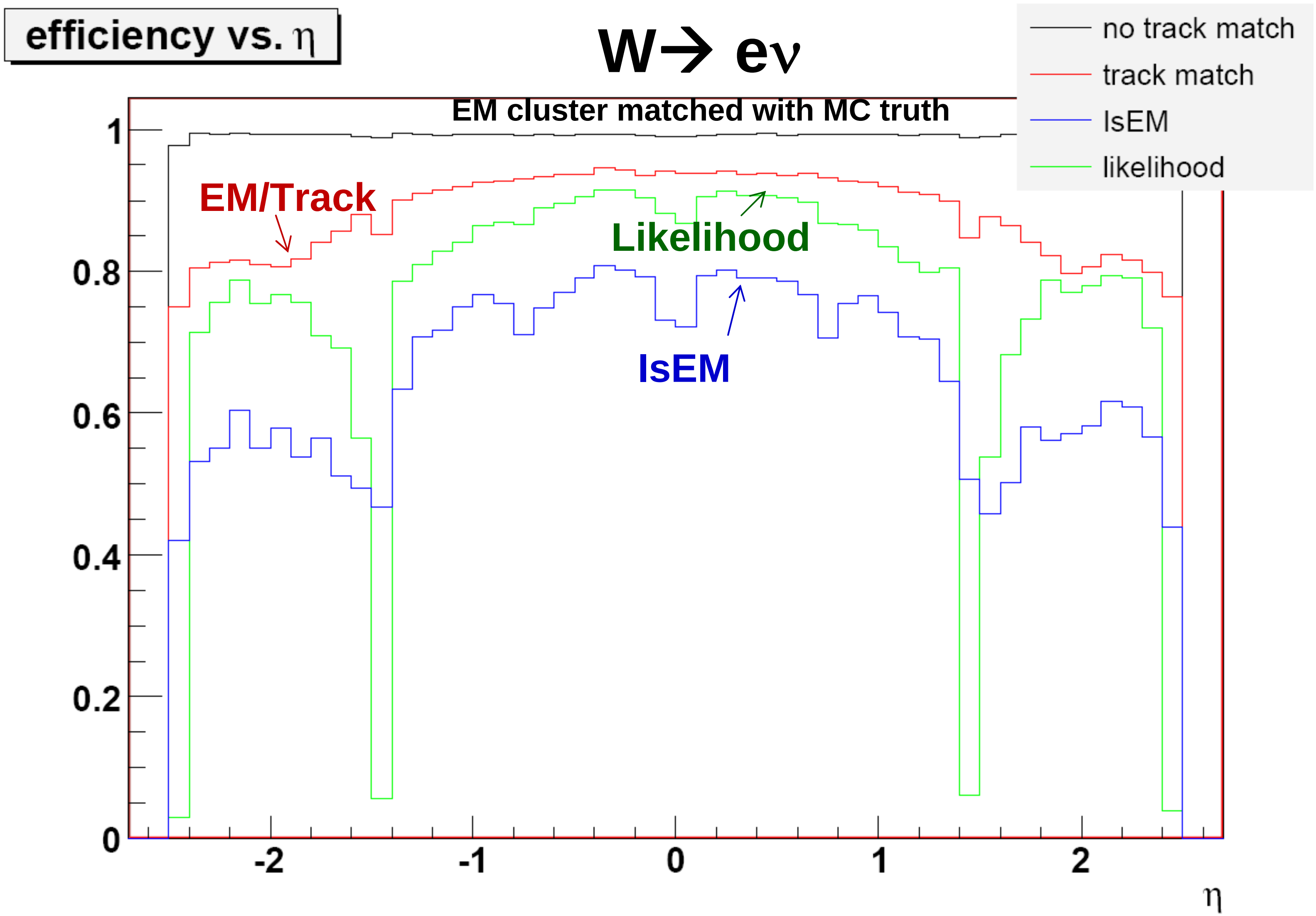
In software release V12 we used Likelihood ratio as the discriminator for e-ID:

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

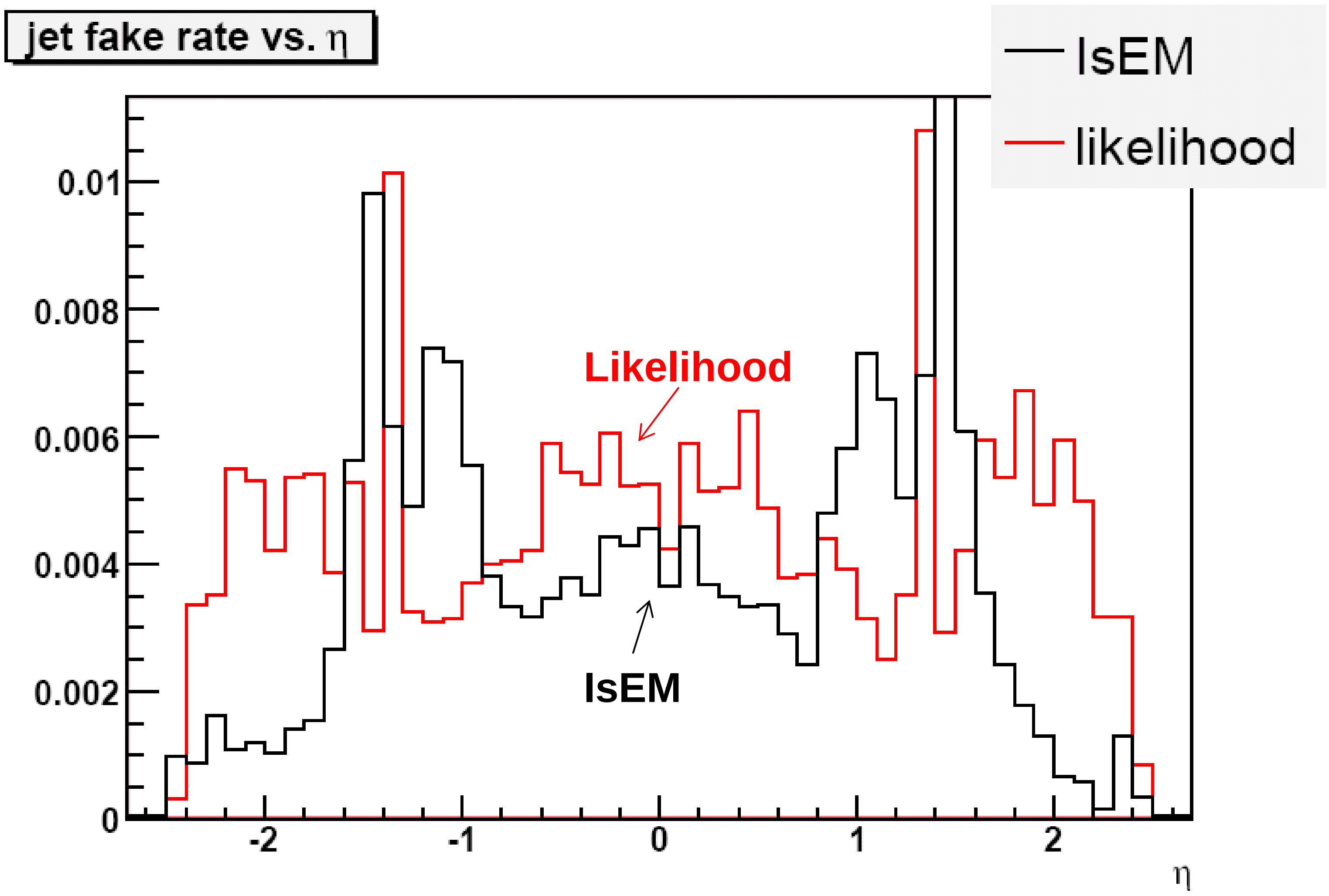
e-ID Efficiencies vs. P_T



e-ID Efficiencies vs. η



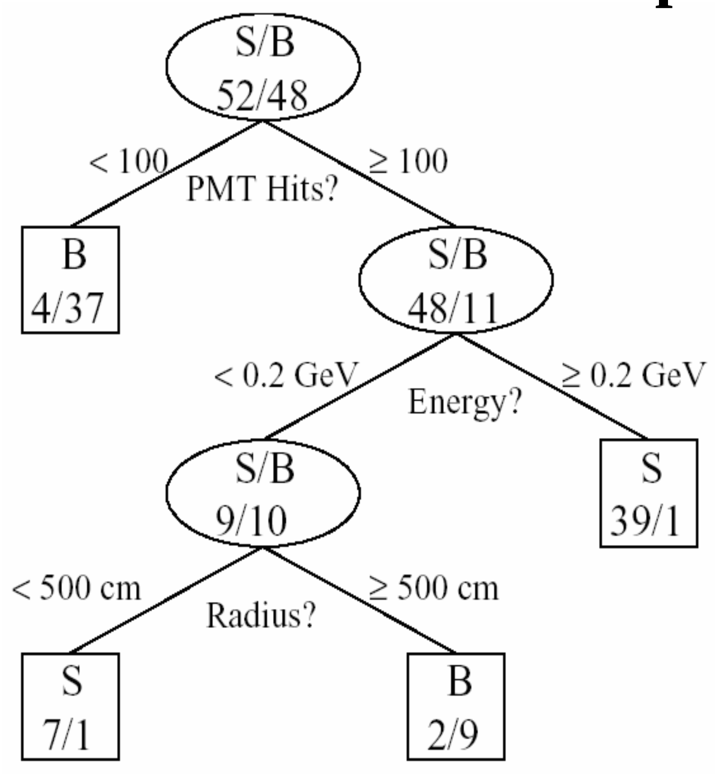
Jet Fake Rate from ttbar Events



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,

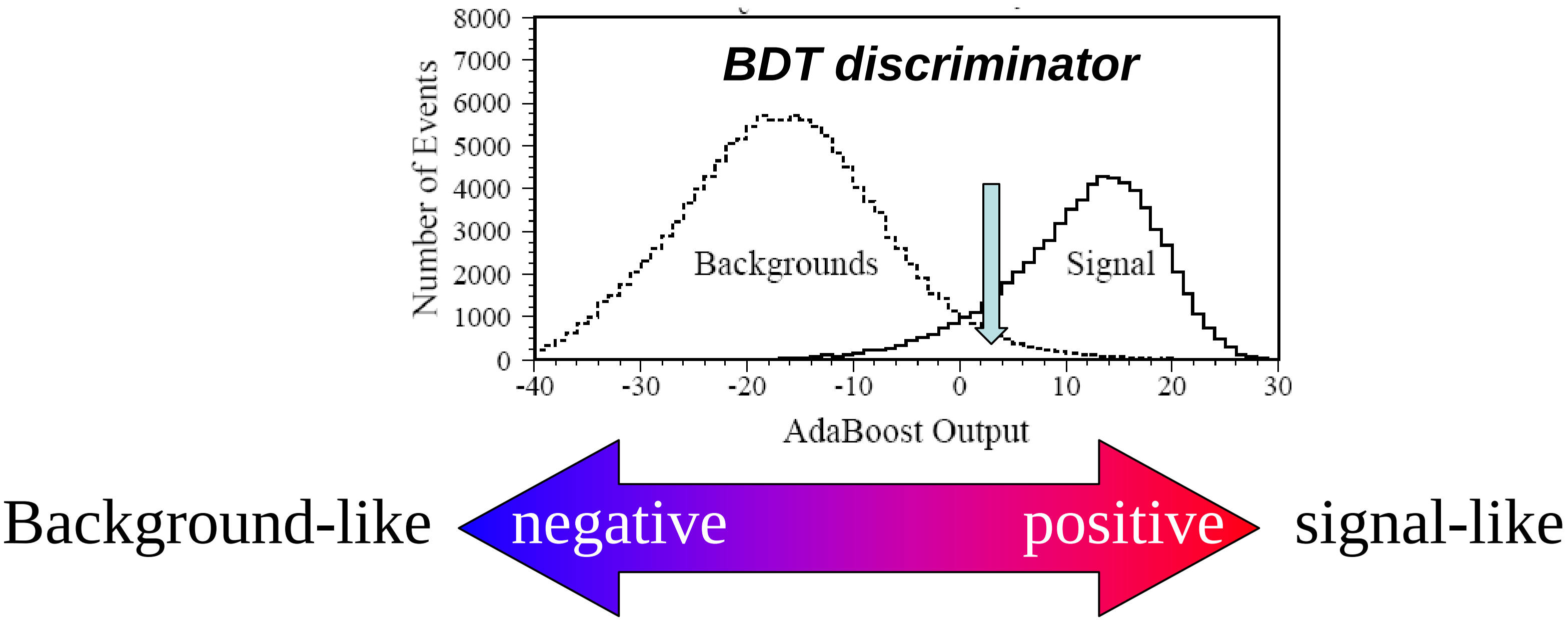
each re-weighting the events to enhance identification of backgrounds misidentified by earlier trees ("boosting")

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**"



Variables Used for BDT e-ID Analysis

IsEM consists of a set of cuts on discriminating variables. These variables are also used for BDT.

- egammaPID::ClusterHadronicLeakage

fraction of transverse energy in TileCal 1st sampling

- egammaPID::ClusterMiddleSampling

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

Shower width 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

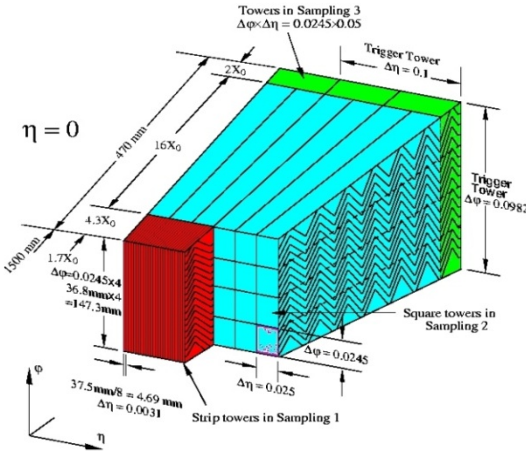
Delta eta between Track and egamma

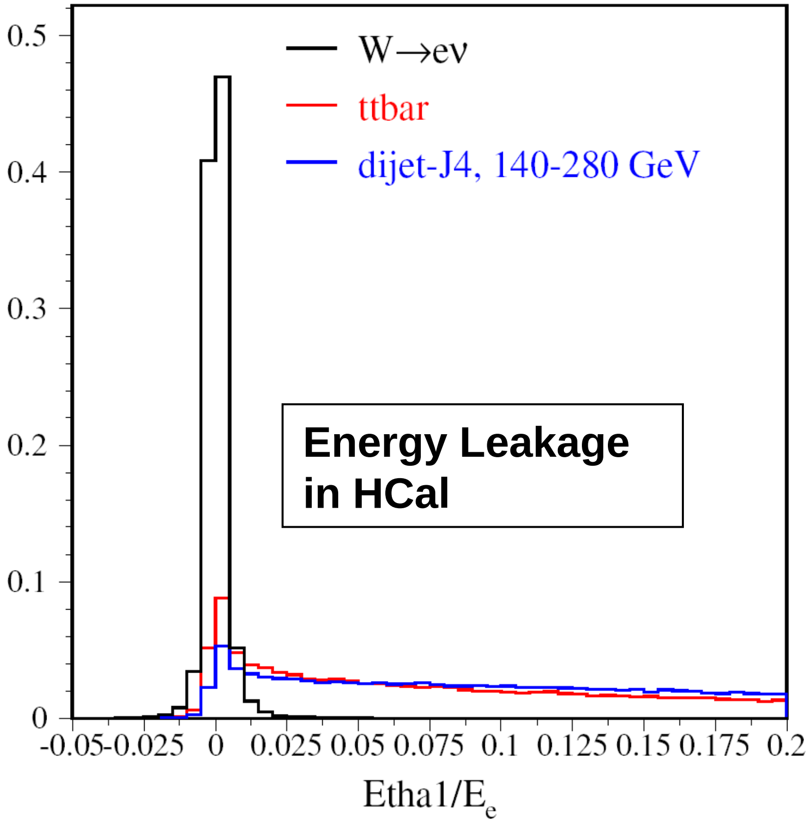
Delta phi between Track and egamma

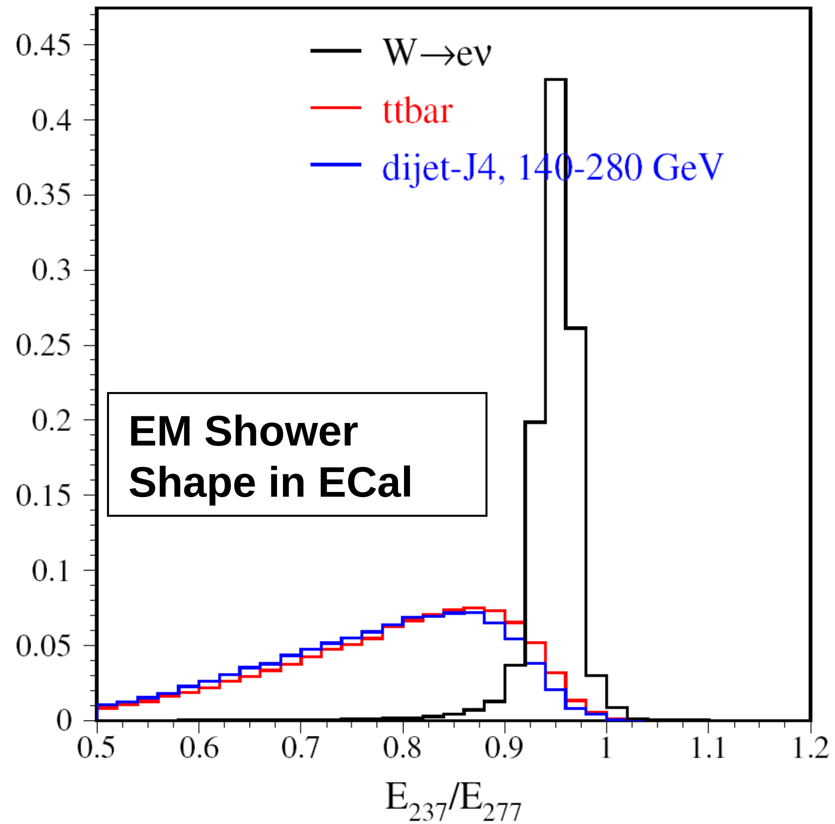
E/P – egamma energy and Track momentum ratio

- trackEtaRange

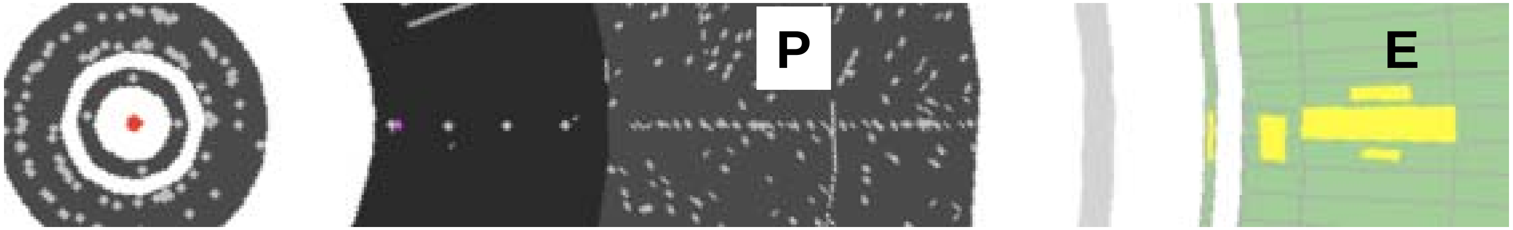
EM Shower shape distributions of discriminating Variables (signal vs. background)

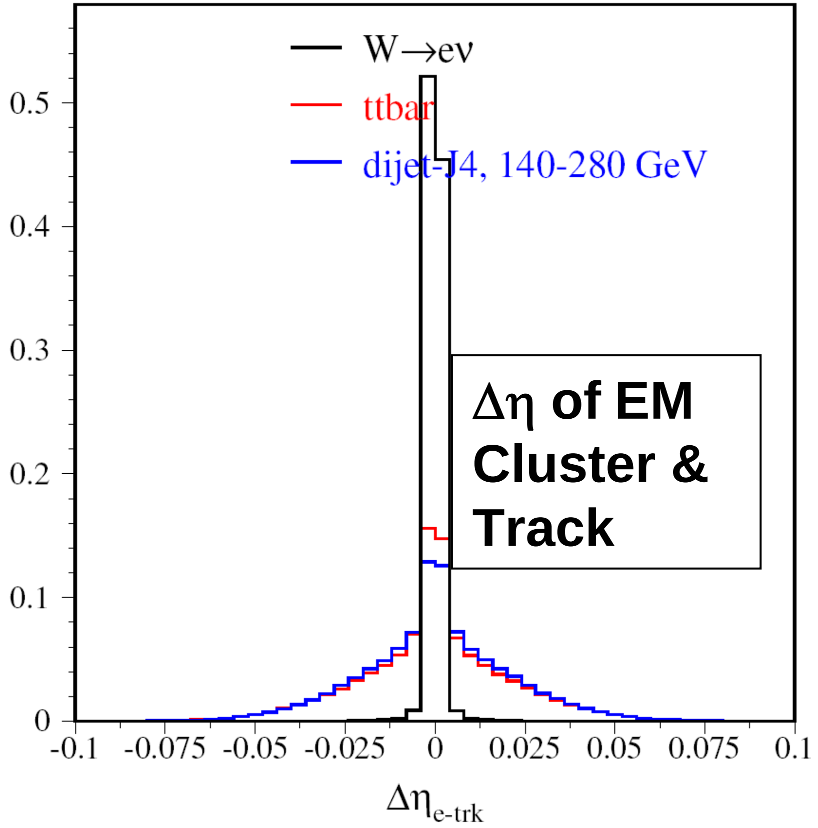


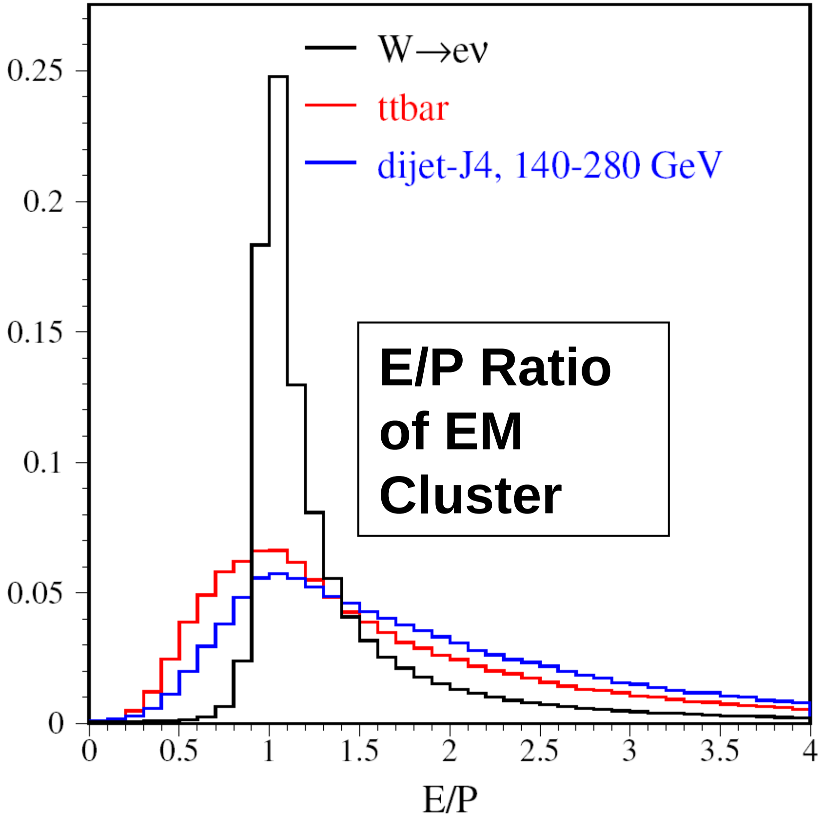




ECal and Inner Track Match

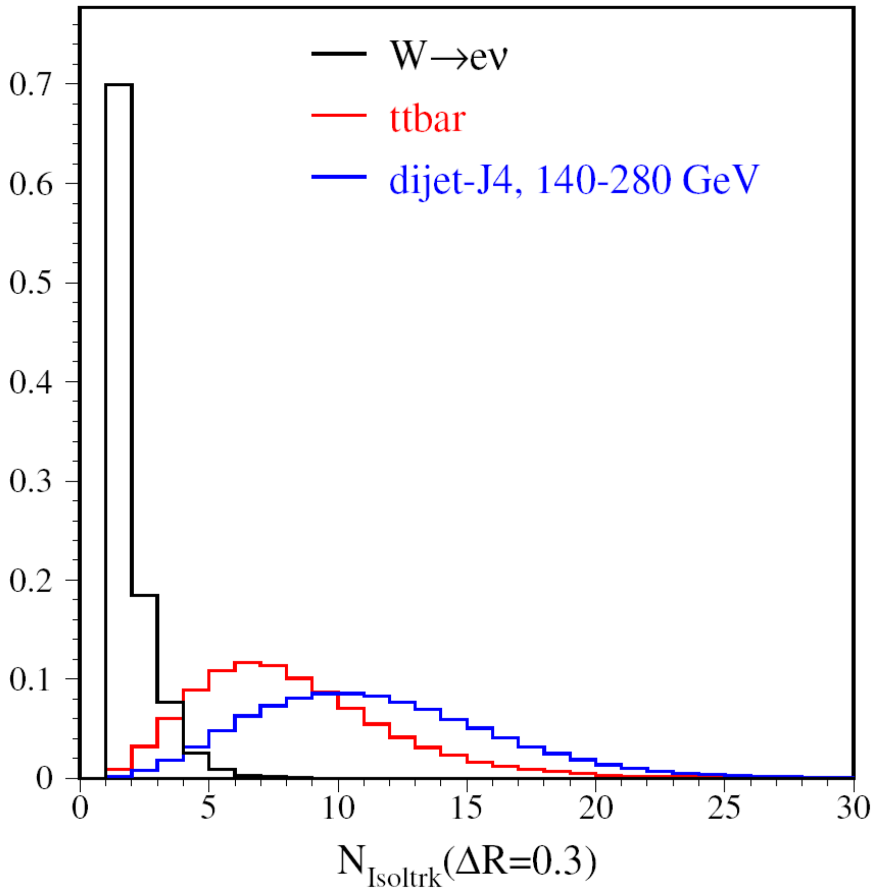




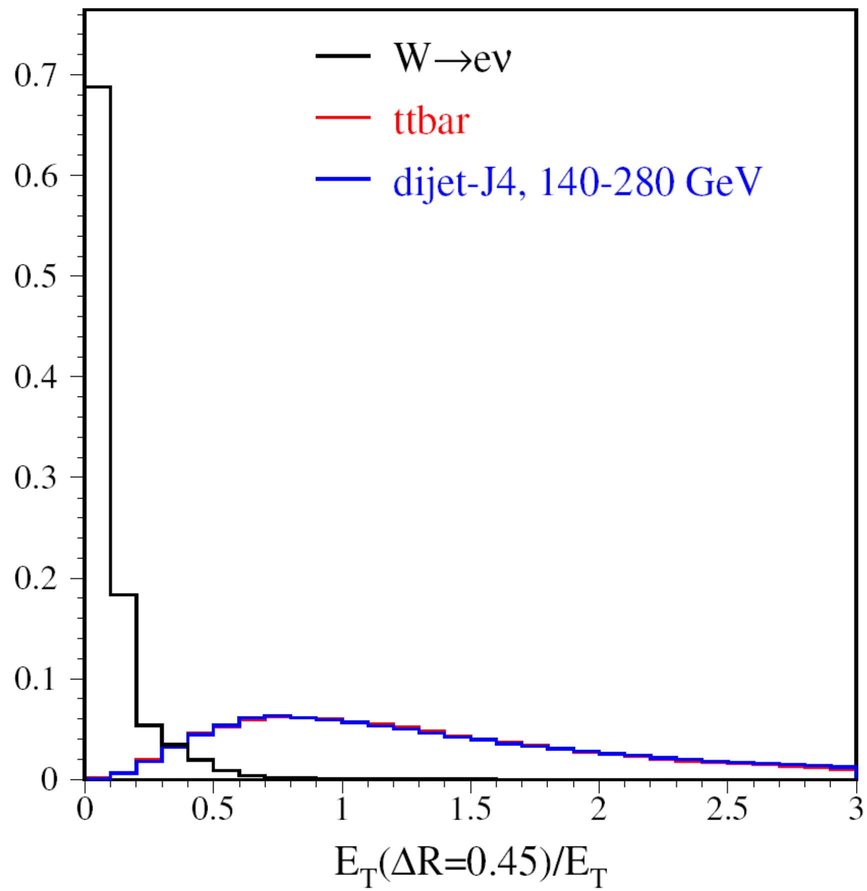


Electron Isolation Variables

N_{trk} around Electron Track



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



BDT e-ID 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
  – Di-jet samples (J0-J6), Pt=[8-1120] GeV
  – ttbar hadronic decays samples

- 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.

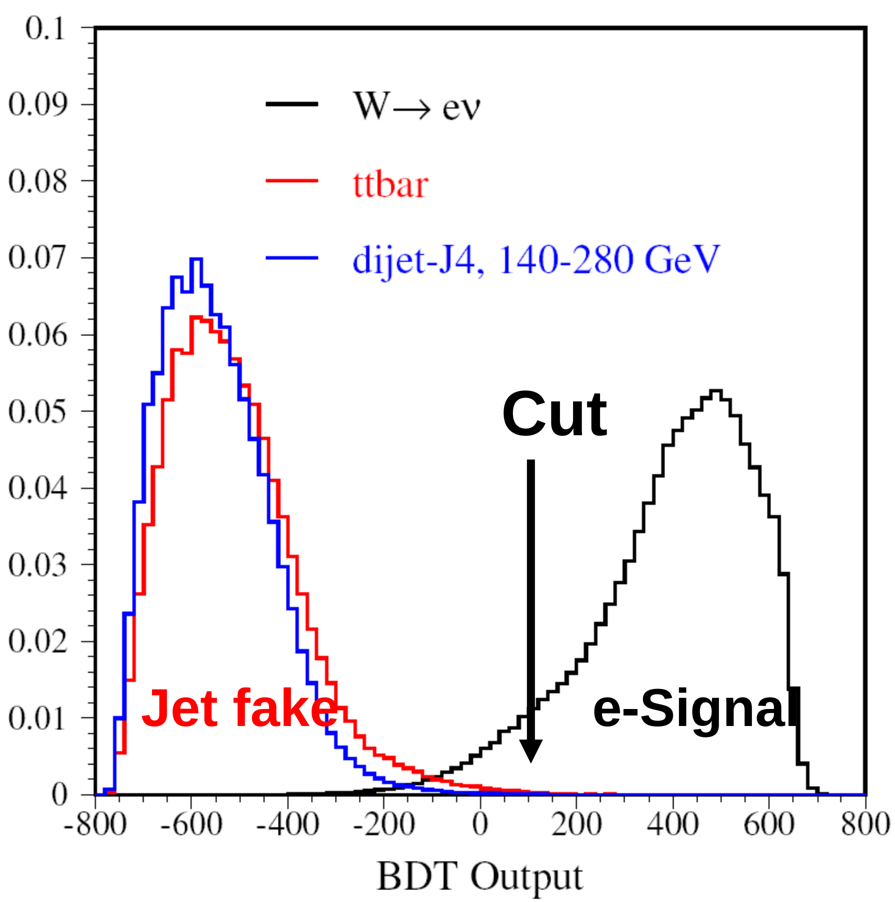
Use Independent Samples to Test the BDT e-ID Performance

- BDT Test Signal (e) Samples:
  - W → eν
  - WW → eνμν
  - Z → ee
  - ZZ → 4l
  - H → WW → lνlν, M_H=140,150,160,165,170,180

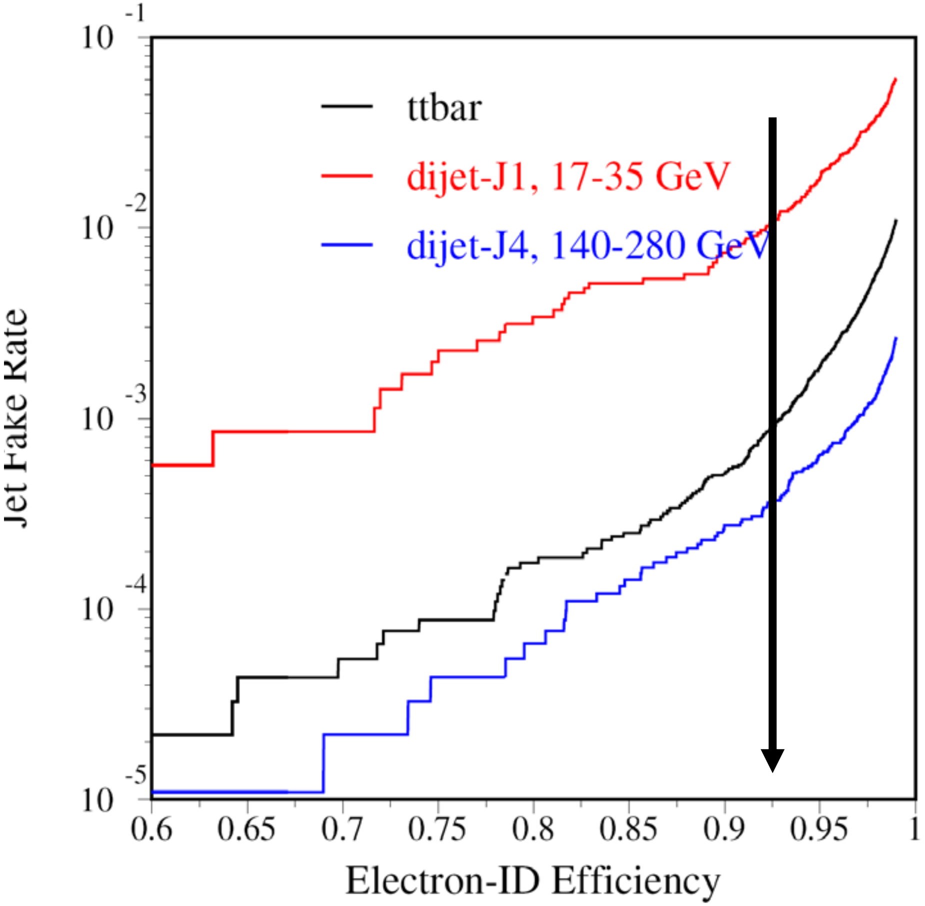
- BDT Test Background (jet faked e) Samples:
  - Di-jet samples (J0-J6), Pt=[8-1120] GeV
  - ttbar hadronic decays samples
  - W→μν + Jets
  - Z→μμ + Jets

Performance of The BDT e-Identification

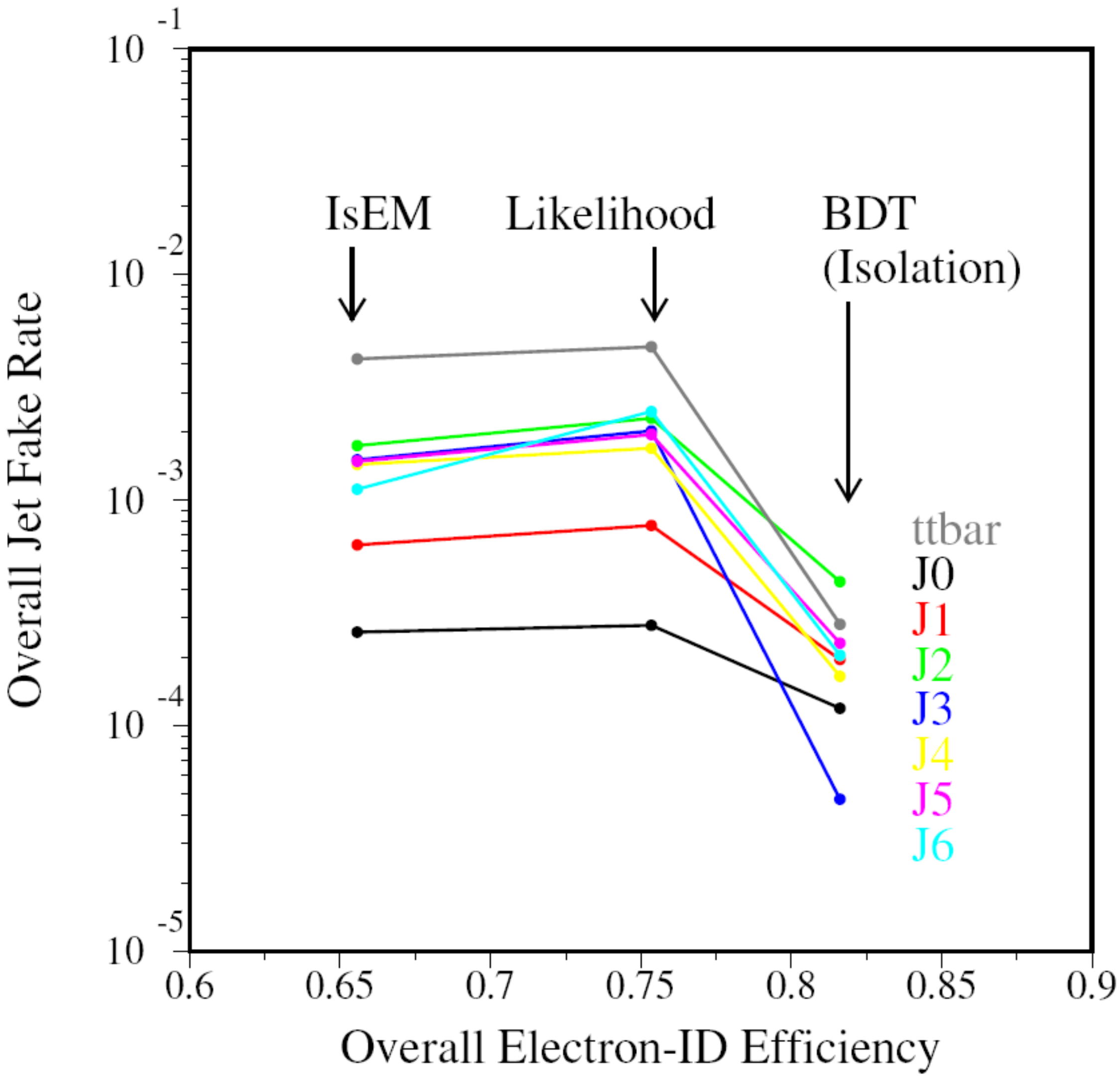
BDT Output Distribution



Jet Fake Rate vs e-ID Eff.



Performance Comparison of e-ID Algorithms



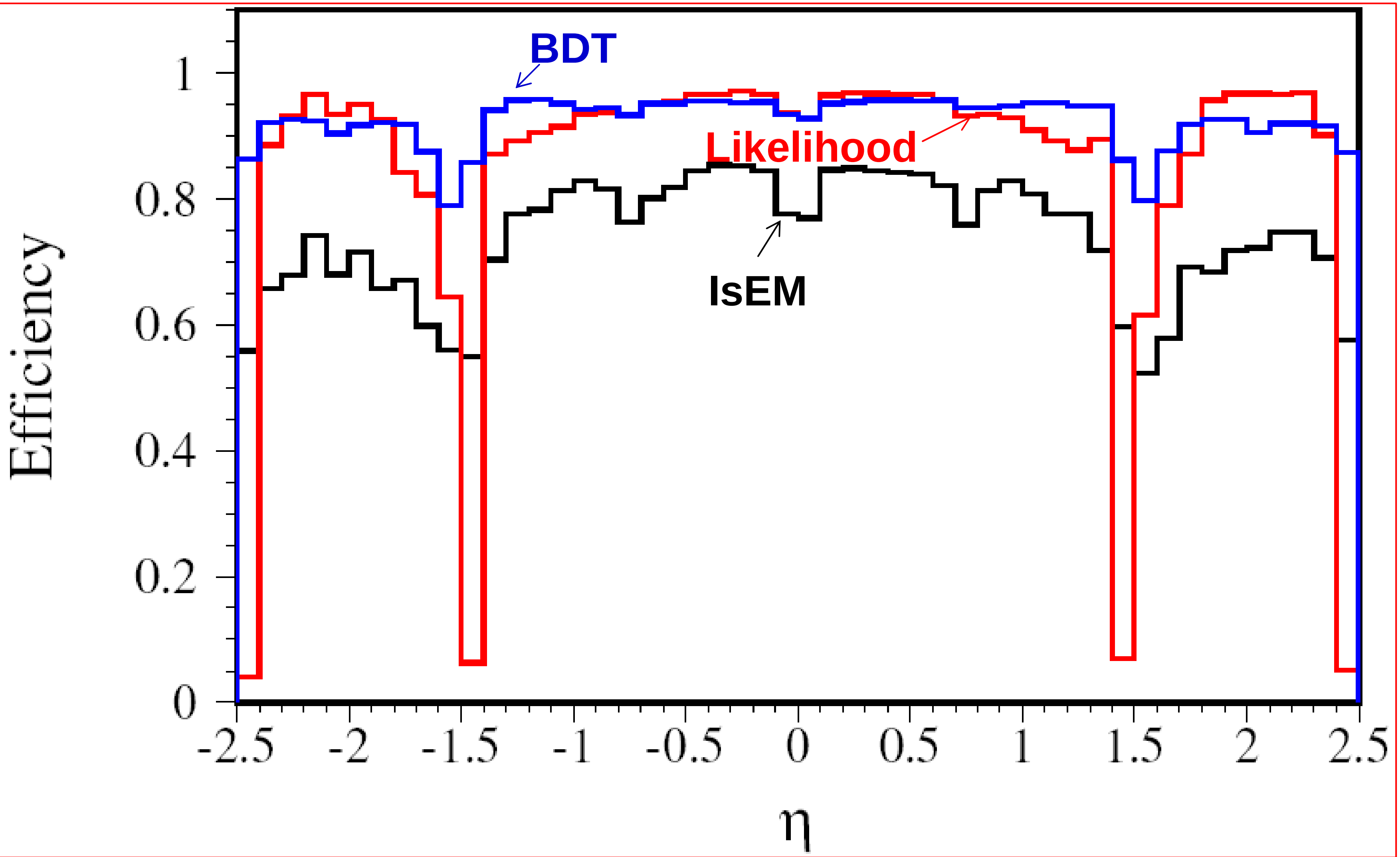
Di-jet Samples

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

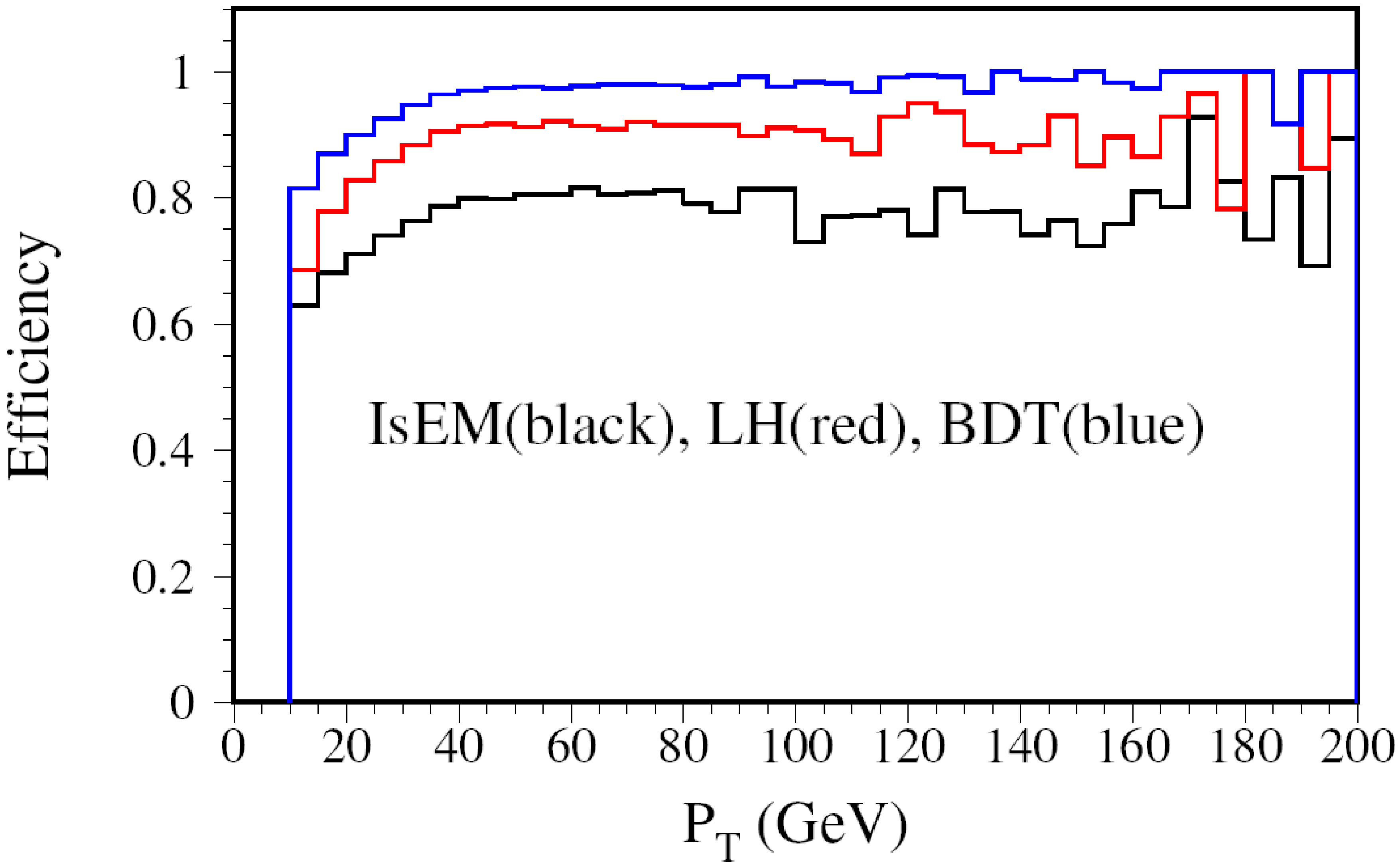
ttbar:
All hadronic decays

BDT e-ID:
– High efficiency
– Low fake rate

Electron ID Eff vs. η (W → eν)

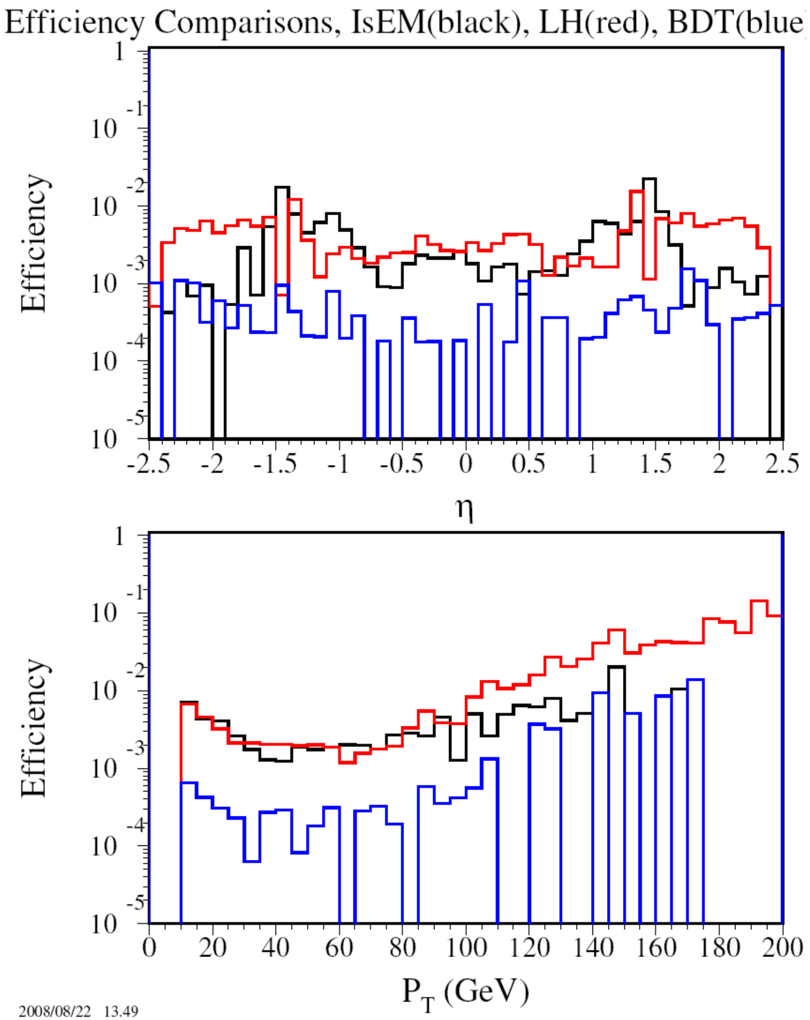


Electron ID Eff vs P_T ($W \rightarrow e\nu$)

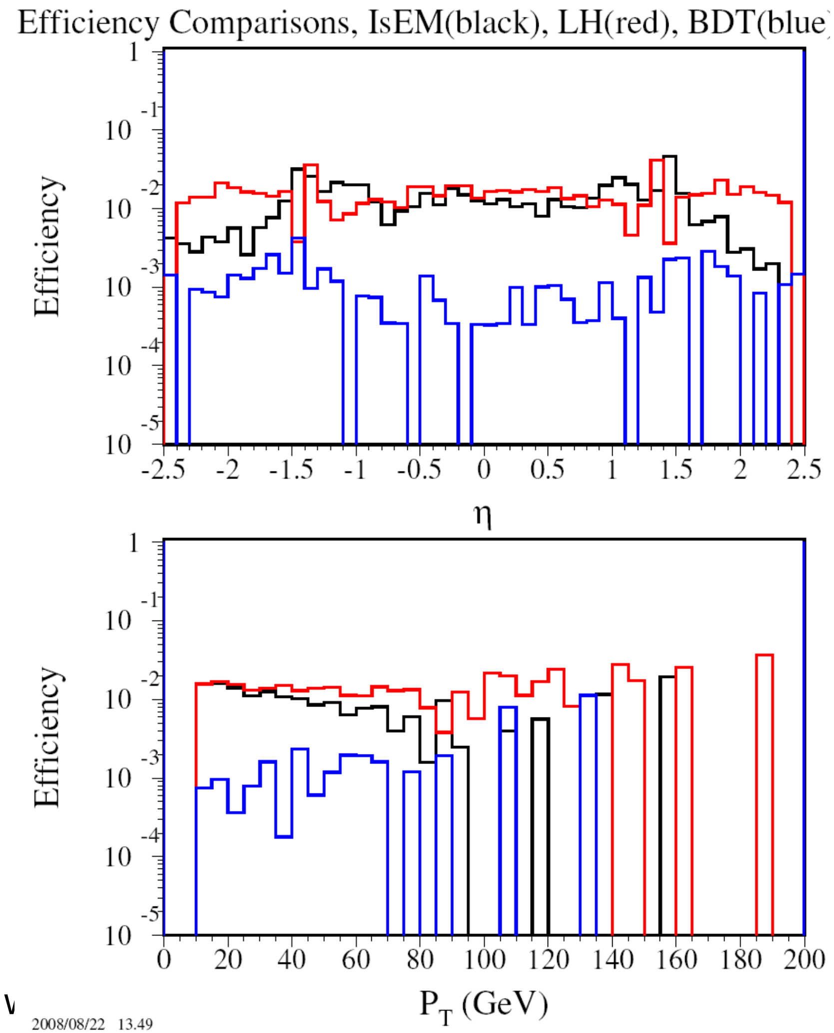


Jet Fake Rate (after EM/Track matching)

J4: di-jet (P_T = 140-280 GeV)



ttbar: all hadronic decays



ID v

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 (no isolation)	BDT (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

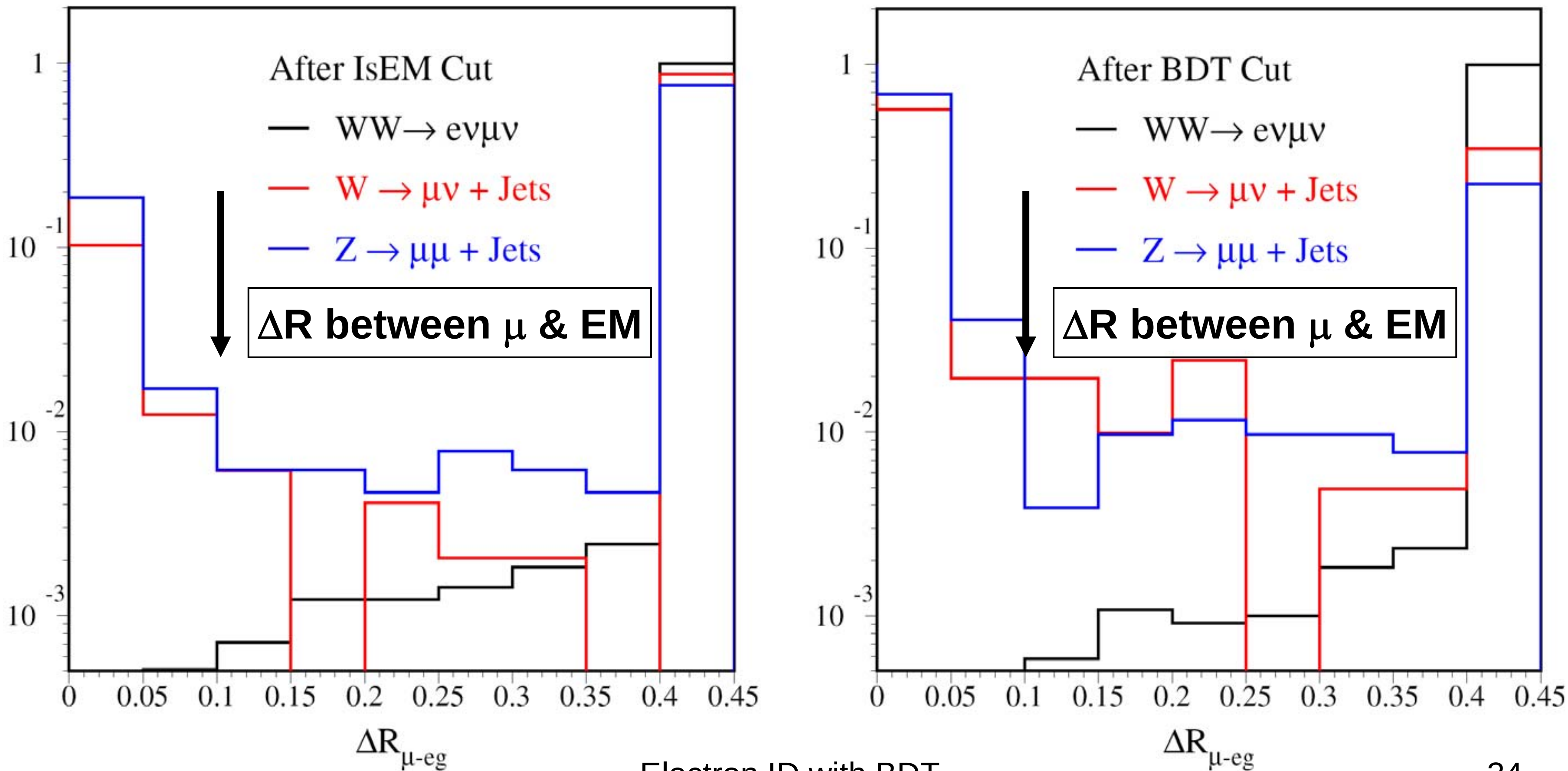
Overall Electron Fake Rate from μ +Jets Events

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	N_e	$Eff_{EM/Track}$	Eff_{IsEM}	Eff_{LH}	Eff_{BDT1}	Eff_{BDT2}
Test Samples	Candidates	Matching	no Isloation	no Isloation	no Isloation	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

Summary

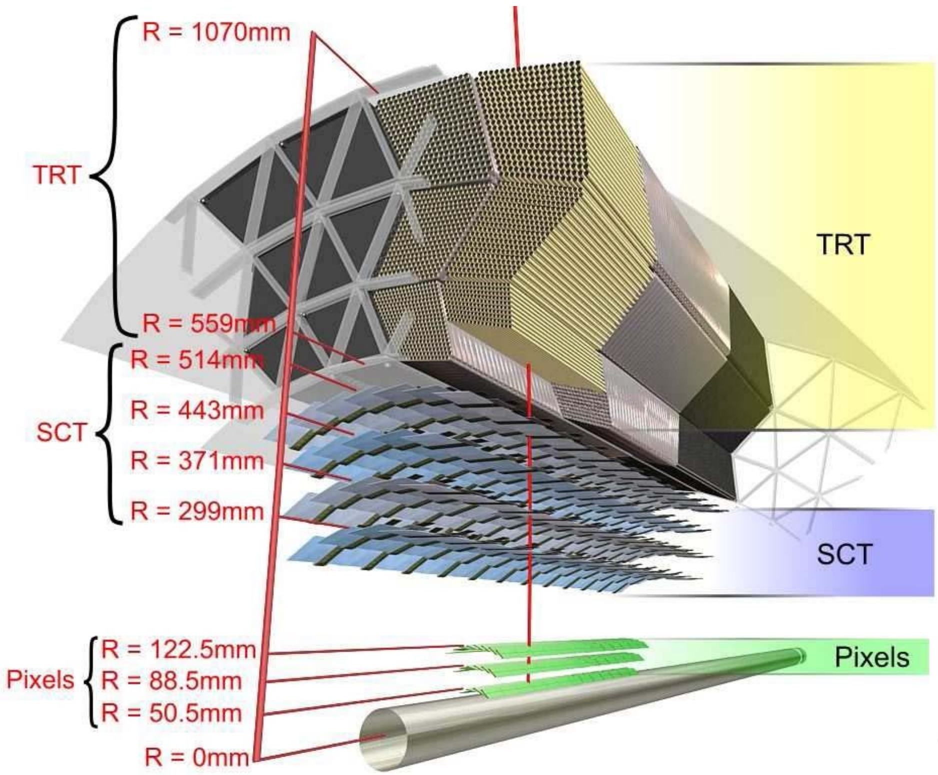
- Electron ID efficiency can be improved by using BDT multivariate particle identification technique
  - Electron Eff = 67% (IsEM) → 75% (LH) →82% (BDT).
- BDT technique also reduce the jet fake rate
  - jet fake rate = 4E-3 (IsEM) → 5E-3 (LH) →3E-3 (BDT) → 3E-4 (BDT with isolation variables) for ttbar
- Fake electron from an EM cluster associated with a muon track can be effectively suppressed

Future Plans

- Incorporate the Electron ID based on BDT into ATLAS official reconstruction package
- Test and check the performance of version 13/14
- Further improve the e-ID efficiency by training the BDTs for barrel, endcap and transition regions, separately.

Backup Slides

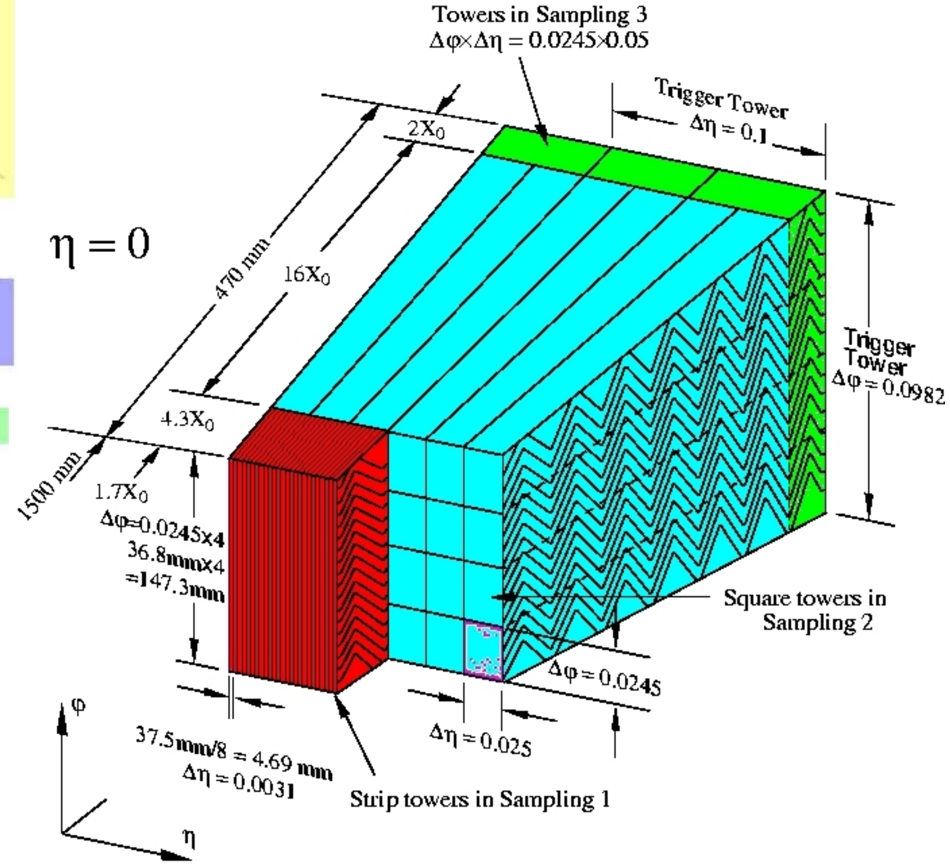
Inner Tracker & ECal for Electron-ID



Fine segmentation for Position/direction measurement

Basic cell in sampling 2:

$\Delta\eta \times \Delta\phi = 0.025 \times 0.025$



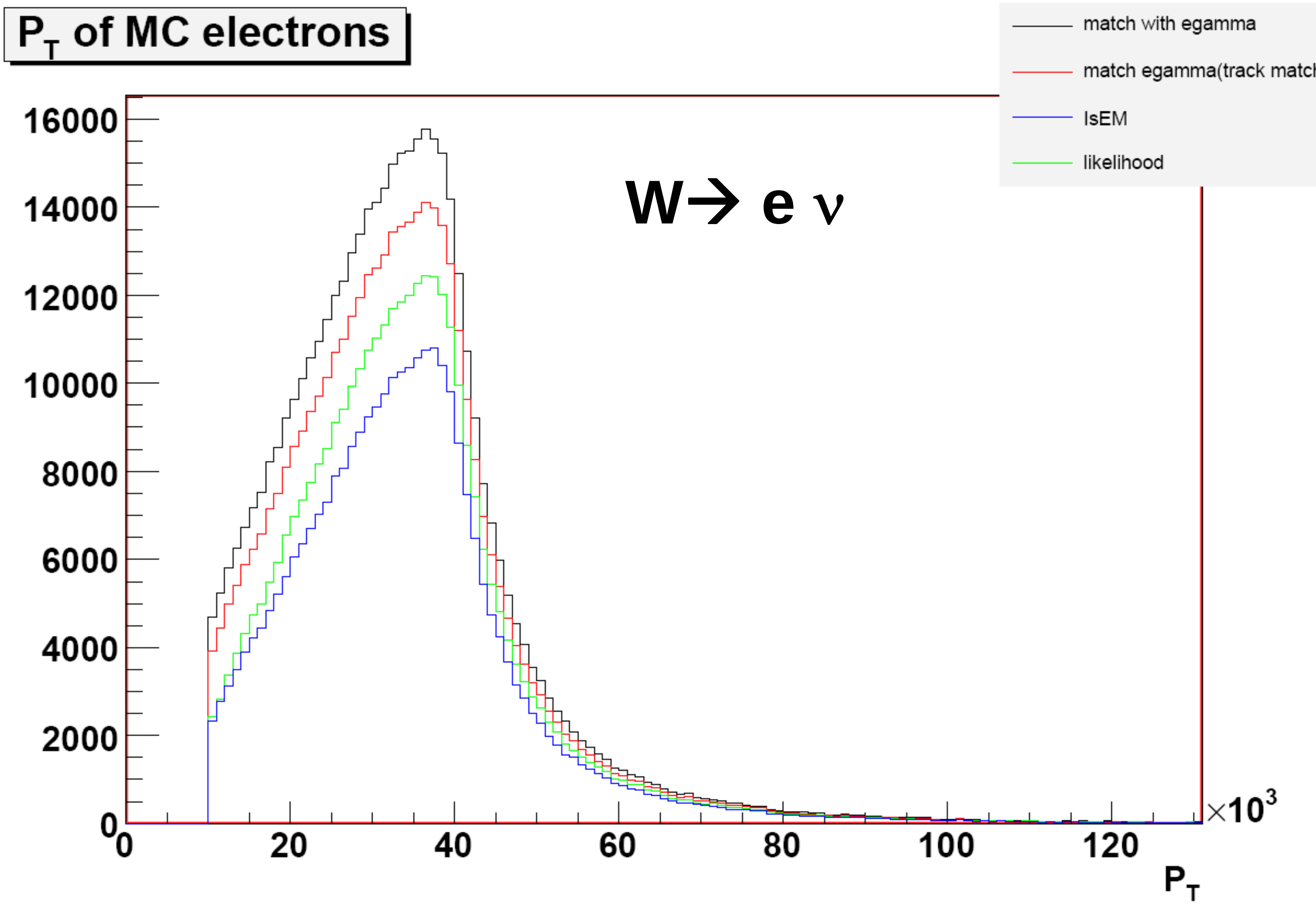
Tracking

Silicon Pixel

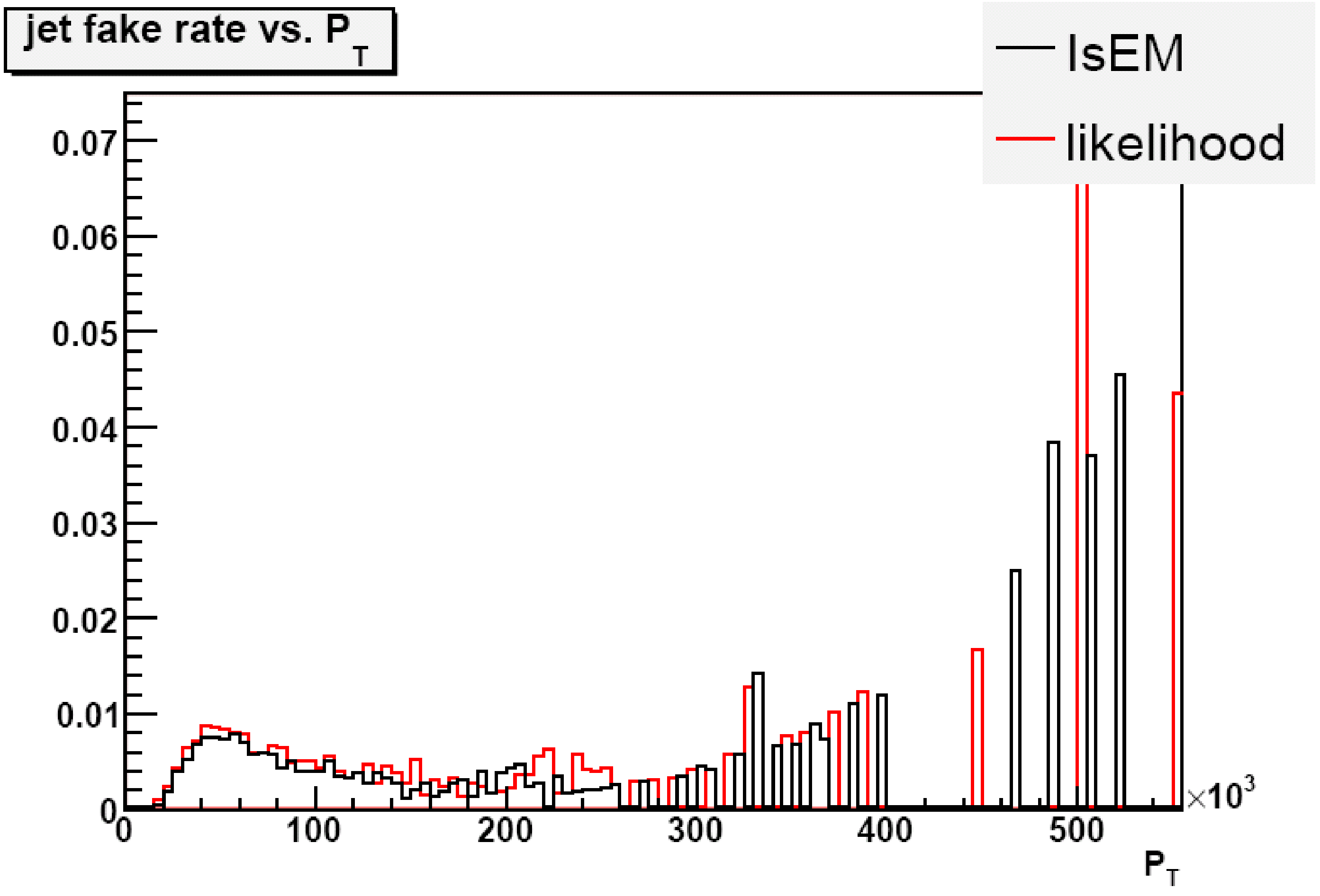
Silicon strips

Transition radiation straw tubes

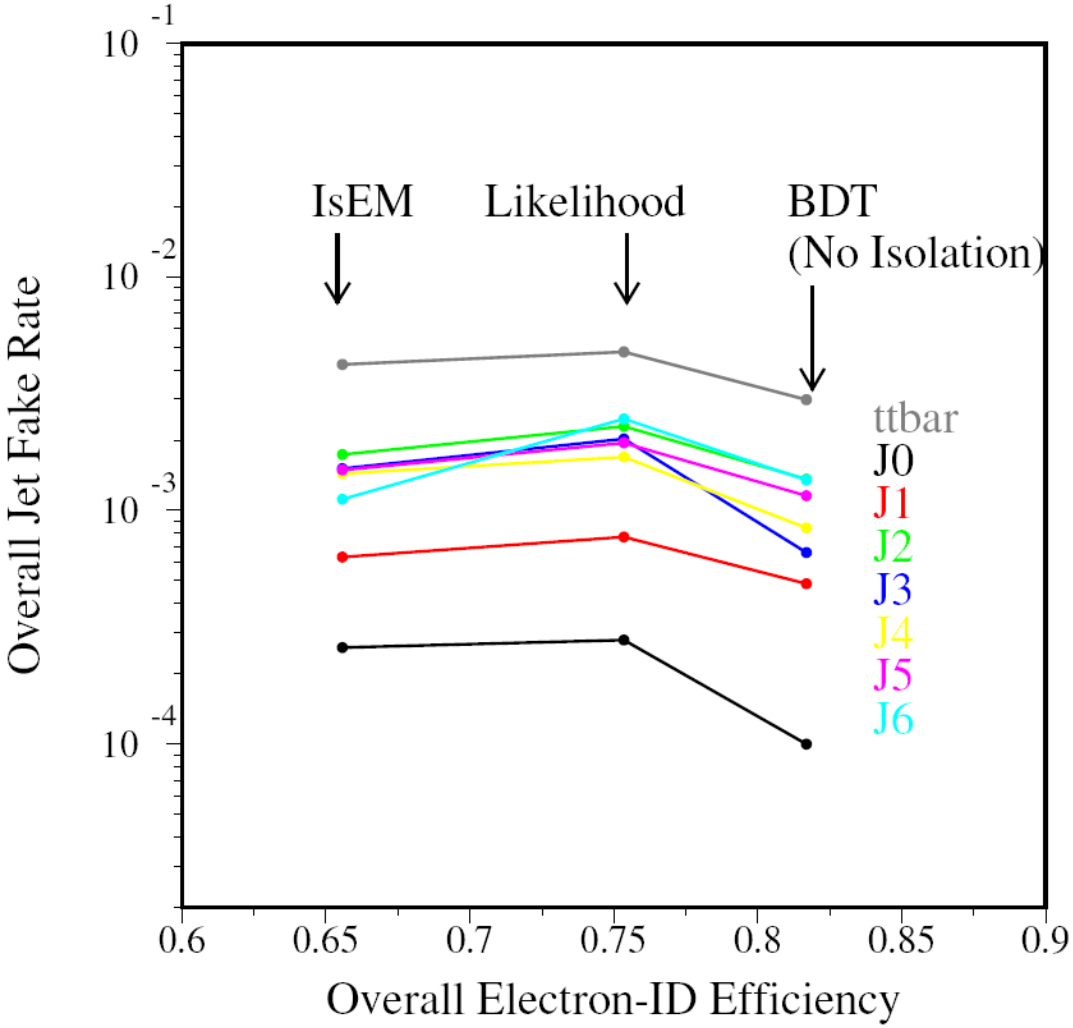
Electron P_T Distributions



Jet Fake Rate from ttbar Events



Performance Comparison of e-ID Algorithms



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

Overall E-ID Efficiency with E_T>17 GeV

MC Processes	N_e	$Eff_{EM/Track}$	Eff_{IsEM}	Eff_{LH}	Eff_{BDT1}	Eff_{BDT2}
Test Samples	Candidates	Matching	no Isloation	no Isloation	no Isloation	with Isolation
Weν	366215	0.888	0.658	0.758	0.818	0.814
HWW 140	41404	0.882	0.668	0.767	0.815	0.791
HWW 150	44554	0.880	0.670	0.771	0.820	0.790
HWW 160	66455	0.879	0.673	0.776	0.824	0.788
HWW 165	87657	0.879	0.678	0.779	0.826	0.791
HWW 170	67761	0.880	0.681	0.780	0.828	0.797
HWW 180	48541	0.879	0.682	0.781	0.829	0.804
HWW170vbf	8244	0.879	0.674	0.765	0.818	0.757
WWeμX	24659	0.883	0.675	0.774	0.831	0.822
Weν-J1	38142	0.884	0.669	0.769	0.824	0.811
Weν-J2	37769	0.884	0.660	0.762	0.823	0.794
Weν-J3	37643	0.879	0.663	0.759	0.822	0.791
Weν-J4	24168	0.882	0.663	0.759	0.822	0.778
Weν-J5	7920	0.874	0.644	0.746	0.813	0.756
Zee	10192	0.881	0.672	0.766	0.832	0.828
ZZll$\nu\nu$	17982	0.879	0.680	0.780	0.835	0.826
ZZllll	35998	0.880	0.680	0.778	0.837	0.819
Zee-J1	127340	0.881	0.679	0.780	0.833	0.817
Zee-J2	44252	0.881	0.679	0.780	0.832	0.810
Zee-J3	84422	0.878	0.673	0.771	0.830	0.797
Zee-J4	45402	0.878	0.671	0.769	0.827	0.788
Zee-J5	15232	0.877	0.674	0.766	0.826	0.776

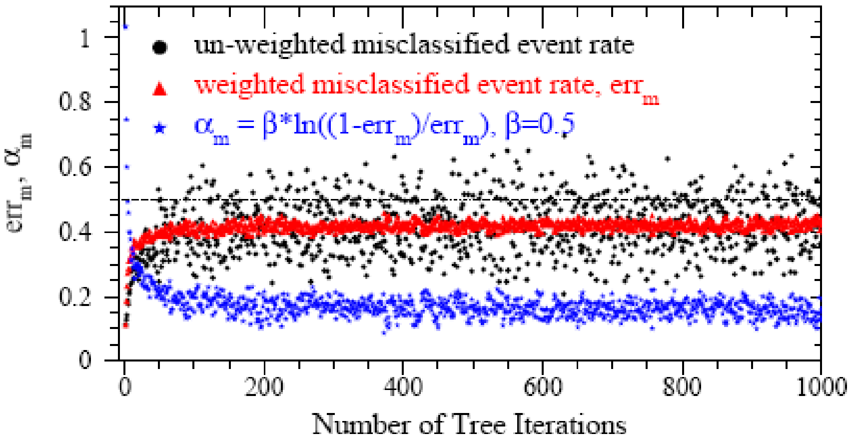
Overall e-fake rate with E_T>17 GeV

MC Processes	N_e	$Eff_{EM/Track}$	Eff_{IsEM}	Eff_{LH}	Eff_{BDT1}	Eff_{BDT2}
Test Samples	Candidates	Matching	no Isloation	no Isloation	no Isloation	with Isolation
dijet-J0	10724	0.867E-02	0.187E-03	0.280E-03	0.933E-04	0.933E-04
dijet-J1	105977	0.540E-02	0.245E-03	0.236E-03	0.236E-03	0.661E-04
dijet-J2	12149	0.435E-01	0.823E-03	0.107E-02	0.741E-03	0.412E-03
dijet-J3	17004	0.252E+00	0.123E-02	0.118E-02	0.412E-03	0.588E-04
dijet-J4	417606	0.423E+00	0.106E-02	0.143E-02	0.685E-03	0.132E-03
dijet-J5	25951	0.519E+00	0.119E-02	0.189E-02	0.886E-03	0.270E-03
dijet-J6	29620	0.506E+00	0.810E-03	0.230E-02	0.125E-02	0.135E-03
$t\bar{t}$	288653	0.237E+00	0.277E-02	0.342E-02	0.204E-02	0.187E-03
W$\mu\nu$-J1	21818	0.963E-01	0.101E-02	0.380E-02	0.101E-02	0.779E-03
W$\mu\nu$-J2	29143	0.132E+00	0.161E-02	0.377E-02	0.144E-02	0.995E-03
W$\mu\nu$-J3	63356	0.166E+00	0.134E-02	0.265E-02	0.134E-02	0.789E-03
W$\mu\nu$-J4	54328	0.202E+00	0.147E-02	0.304E-02	0.136E-02	0.773E-03
W$\mu\nu$-J5	22257	0.229E+00	0.103E-02	0.256E-02	0.144E-02	0.674E-03
Z$\mu\mu$-J2	44316	0.140E+00	0.162E-02	0.569E-02	0.257E-02	0.219E-02
Z$\mu\mu$-J3	64704	0.172E+00	0.155E-02	0.507E-02	0.193E-02	0.176E-02
Z$\mu\mu$-J4	85775	0.204E+00	0.176E-02	0.464E-02	0.219E-02	0.156E-02
Z$\mu\mu$-J5	37162	0.233E+00	0.126E-02	0.414E-02	0.188E-02	0.135E-02
Electron Fake Rate from Jets with muon veto cut $\Delta R_{\mu-eg} > 0.1$						
W$\mu\nu$-J1	21818	0.963E-01	0.825E-03	0.229E-02	0.504E-03	0.275E-03
W$\mu\nu$-J2	29143	0.132E+00	0.127E-02	0.216E-02	0.515E-03	0.412E-03
W$\mu\nu$-J3	63356	0.166E+00	0.963E-03	0.169E-02	0.474E-03	0.316E-03
W$\mu\nu$-J4	54328	0.202E+00	0.131E-02	0.190E-02	0.368E-03	0.239E-03
W$\mu\nu$-J5	22257	0.229E+00	0.988E-03	0.180E-02	0.449E-03	0.359E-03
Z$\mu\mu$-J2	44316	0.140E+00	0.948E-03	0.271E-02	0.104E-02	0.745E-03
Z$\mu\mu$-J3	64704	0.172E+00	0.958E-03	0.235E-02	0.665E-03	0.525E-03
Z$\mu\mu$-J4	85775	0.204E+00	0.129E-02	0.232E-02	0.536E-03	0.420E-03
Z$\mu\mu$-J5	37162	0.233E+00	0.102E-02	0.188E-02	0.377E-03	0.377E-03

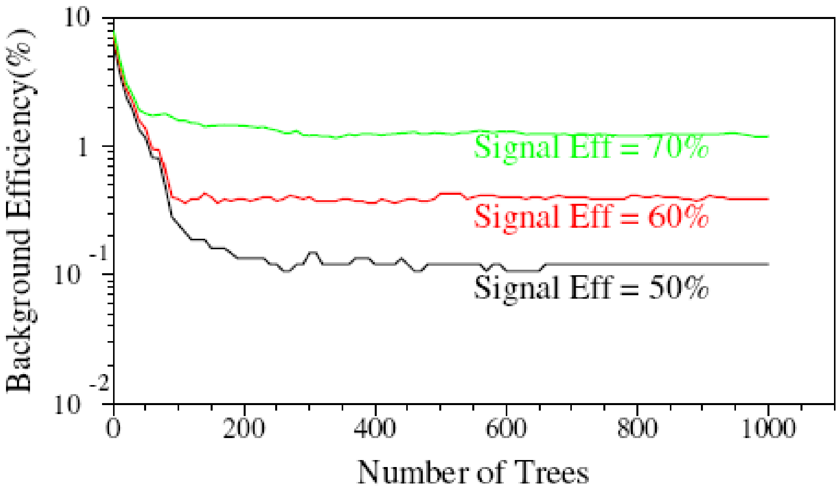
Rank of Variables (Gini Index)

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. η 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

Weak → Powerful Classifier



→ The advantage of using boosted decision trees is that it combines many decision trees, "weak" classifiers, to make a powerful classifier. The performance of boosted decision trees is stable after a few hundred tree iterations.



→ Boosted decision trees focus on the misclassified events which usually have high weights after hundreds of tree iterations. An individual tree has a very weak discriminating power; the weighted misclassified event rate err_m is about 0.4-0.45.

Ref1: H.J.Yang, B.P. Roe, J. Zhu, "*Studies of Boosted Decision Trees for MiniBooNE Particle Identification*", physics/0508045, Nucl. Instum. & Meth. A 555(2005) 370-385.

Ref2: H.J. Yang, B. P. Roe, J. Zhu, " *Studies of Stability and Robustness for Artificial Neural Networks and Boosted Decision Trees* ", physics/0610276, Nucl. Instrum. & Meth. A574 (2007) 342-349.

Major Achievements using BDT

- MiniBooNE neutrino oscillation search using BDT and Maximum Likelihood methods
  - Phys. Rev. Lett. 98 (2007) 231801
  - One of top 10 physics stories in 2007 by AIP
- D0 – discovery of single top using BDT, ANN, ME
  - Phys. Rev. Lett. 98 (2007) 181802
  - One of top 10 physics stories in 2007 by AIP
- BDT was integrated in CERN TMVA package
  - Toolkit for MultiVariate data Analysis
  - http://tmva.sourceforge.net/
- Event Weight training technique for ANN/BDT
  - H. Yang et.al., JINST 3 P04004 (2008)
  - Integrated in TMVA package within 2 weeks after my first presentation at CERN on June 7, 2007