B-tagging based on Boosted Decision Trees

Haijun Yang
University of Michigan
(with Xuefei Li and Bing Zhou)

ATLAS B-tagging Meeting
CERN, July 7, 2009

Outline

- Introduction
- Boosted Decision Trees
- B-tagging discriminating variables
- BDT B-taggers (for light-jets, C-jets, τ-jets)
- Performance comparison and cross checks of ATLAS B-taggers
- Further improvement by combining B-taggers
- Summary

Introduction

- Physics analyses at LHC need to tag b-jet as signal or veto b-jet as background, it is crucial to develop high performance b-tagging methods.

- ATLAS already developed many B-taggers, our work is to further improve performance of B-taggers based on Boosted Decision Trees.

- **Major efforts in our development include:**
  - Evaluate many b-tagging discriminating variables, including those used in previous work (ATL-PHYS-PUB-2007-019), and a new set of variables built by us using the ATLAS reconstructed track, vertex and jet information.
  - Built Boosted Decision Trees (BDT: e-boost algorithm (epsilon=0.01),1000 trees, 20 leaves) by using different number of input discriminating variables
  - Compare the performance with the existing B-taggers, and exam both b-jet signal and light-jet background overlapping rate to check the reliability and to find possible improvement
  - Make further improvement by combining different b-taggers

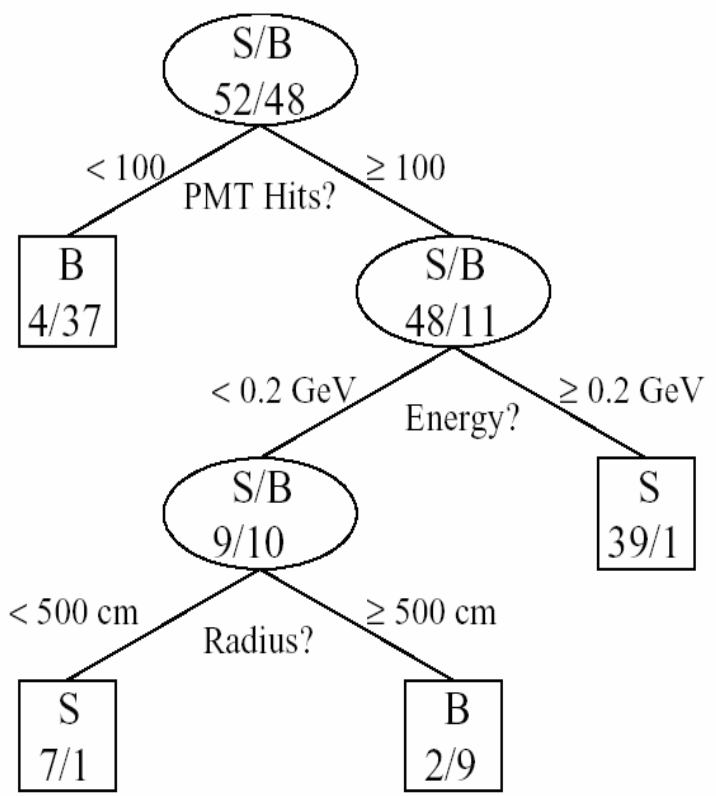
- We started BDT b-tagging with MC V12, and continued with V13 and V14 MC samples. Only present the V14 results today. Details can be found in our note: ATL-COM-PHYS-2009-274.

Boosted Decision Trees

A multivariate technique

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

➔Advantages: Transparent, naturally take care of variable correlations, robust,…



How to build a decision tree

•Split data recursively based on input variables until a stopping criterion is reached (e.g. purity, too few events)

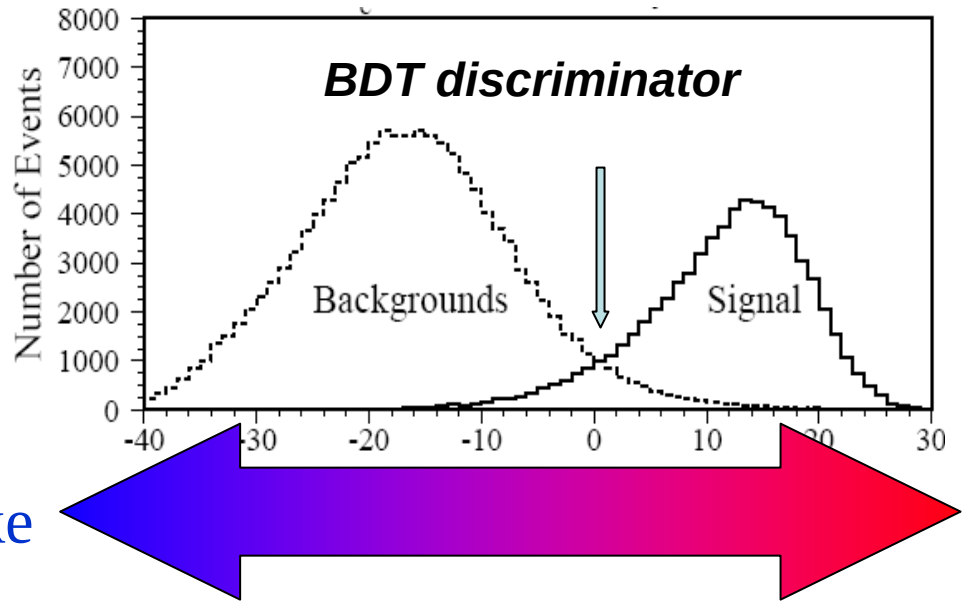
• Every event ends up in a "signal" or a "background" leaf

• 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

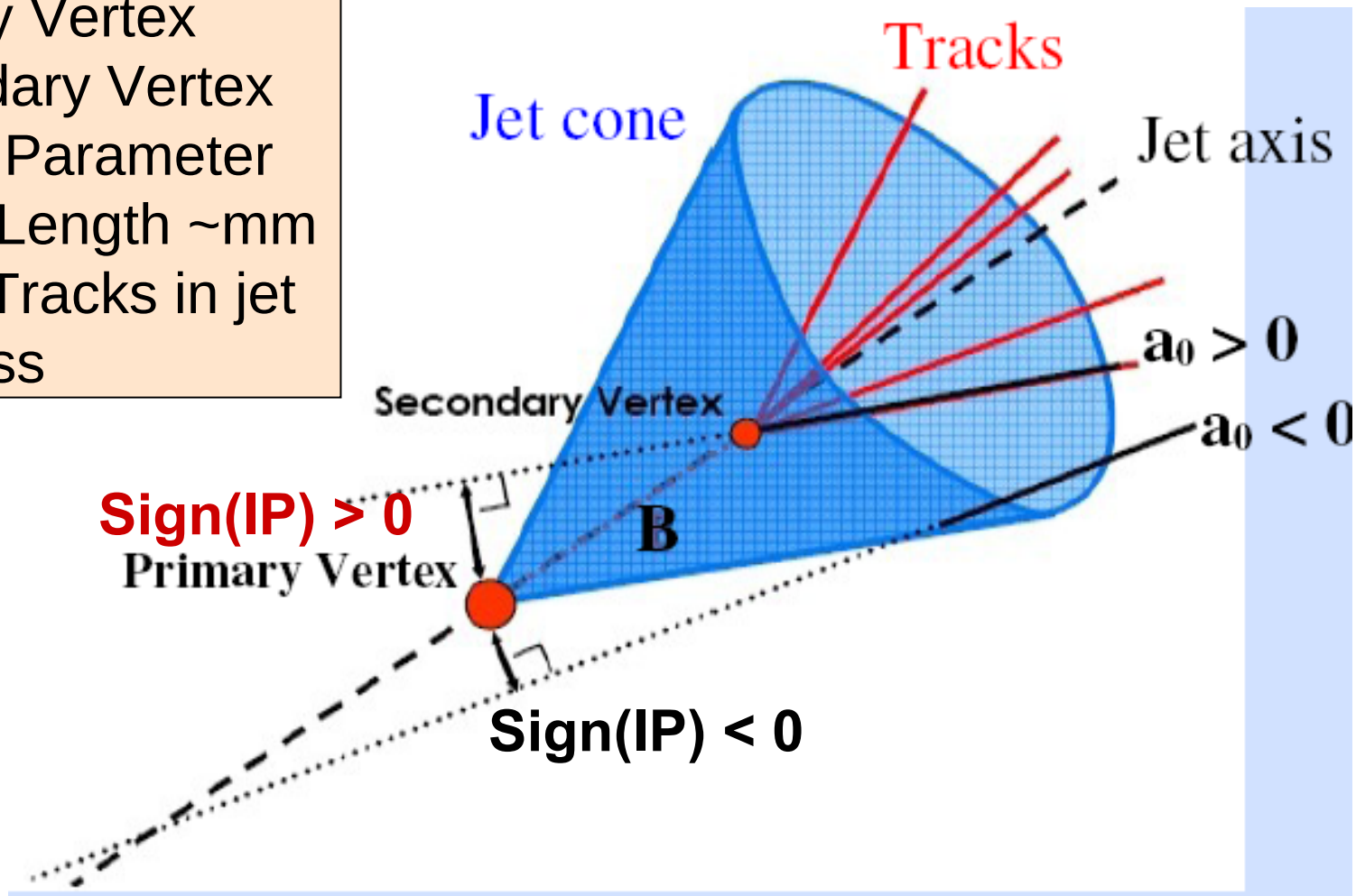
Boosting - A procedure that combines many classifiers (decision trees) to form a powerful discriminator

- **A set of decision trees:** each re-weighting the events to enhance identification of misidentified by earlier trees ("boosting")
- **For each tree,** the test 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 "discriminator"**



B-tagging based on ‘long’ life-time information

- Primary Vertex
- Secondary Vertex
- Impact Parameter
- Decay Length ~mm
- No. of Tracks in jet
- Jet mass



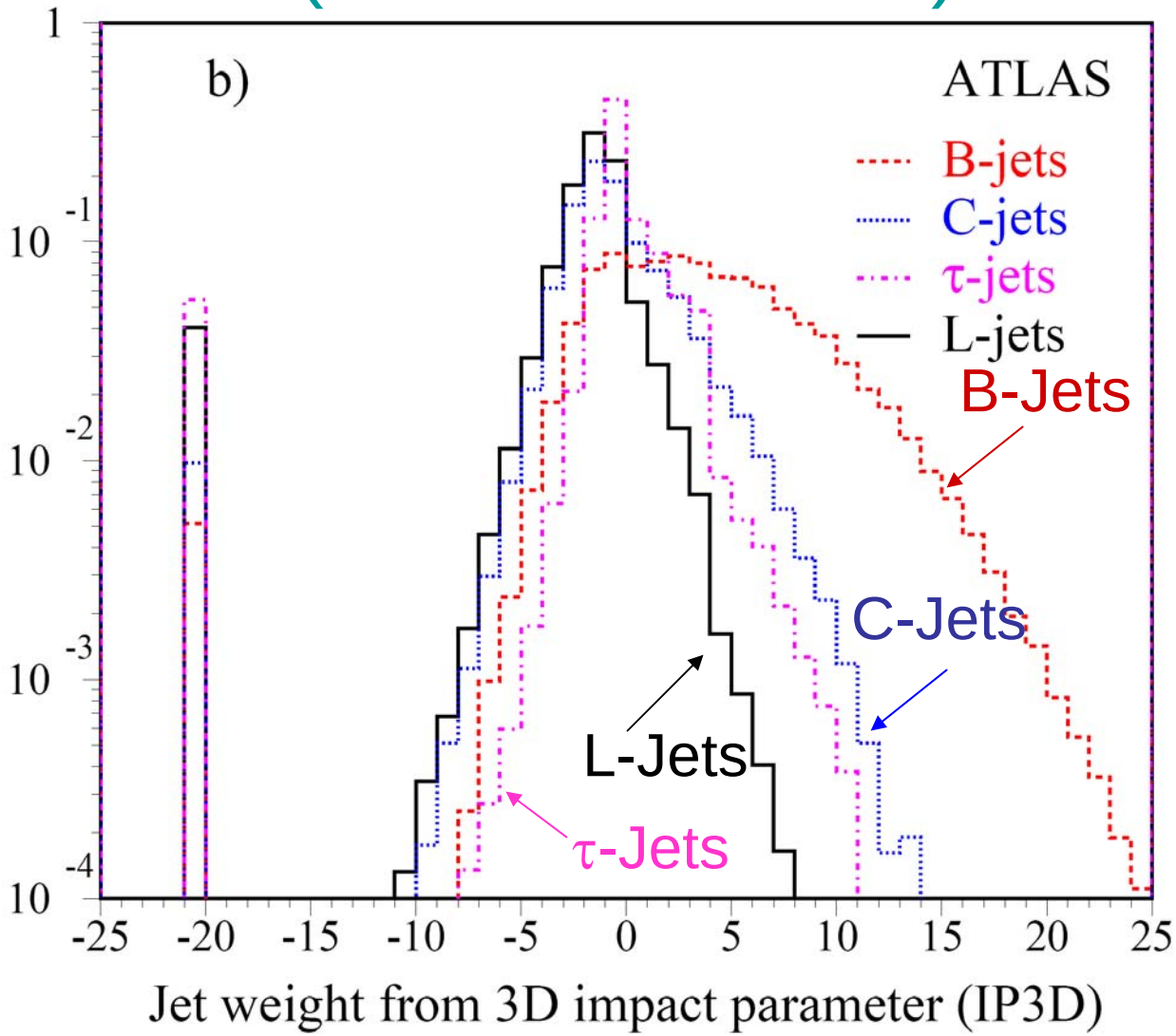
Discriminating variables for BDT b-tagging

- **Performance of BDT algorithm depends on**
  - **selecting a set of discriminating variables;**
  - **using advanced training process.**

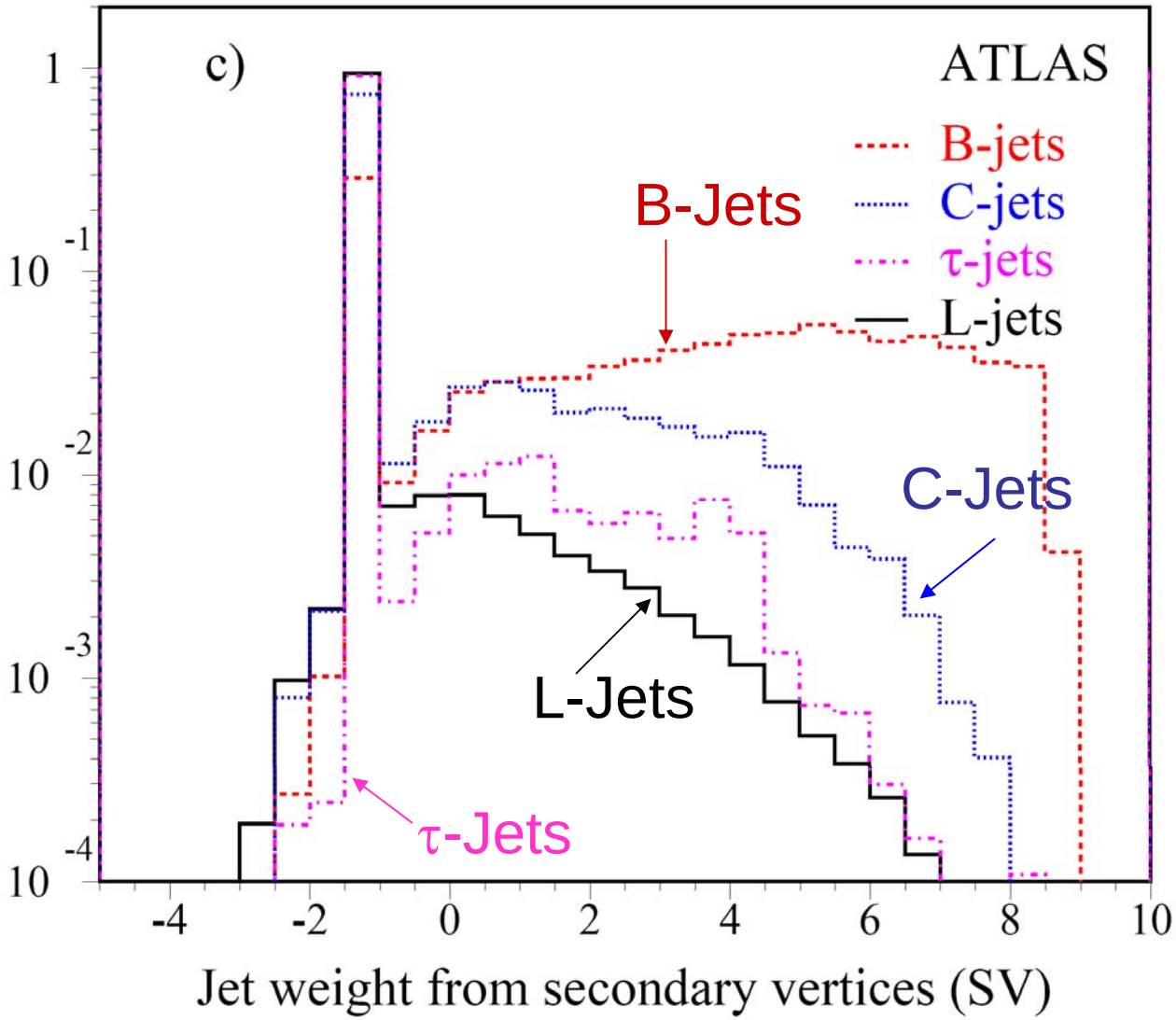
- **11 variables are selected from previous b-tagging work (17 variables):**
  - IP2D, IP3D, SV1: jet weights from IP and secondary vertices
  - Softe: jet weight from soft electron based tagger
  - jet-mass: mass of particles which participate in the vertex fit
  - Efrac: ratio between the total energy of charged particles in the vertex and the total energy of all particles in the jet
  - d0sig_max, z0sig_max: the largest transverse and longitudinal impact parameter significance of tracks in the jet
  - ptTrk_max: the largest transverse momentum of tracks in the jet
  - Nvertex_2track: Number of two-track vertices
  - Ntrack: Number of tracks in the jet

Jet Weight from 3D Impact Parameters

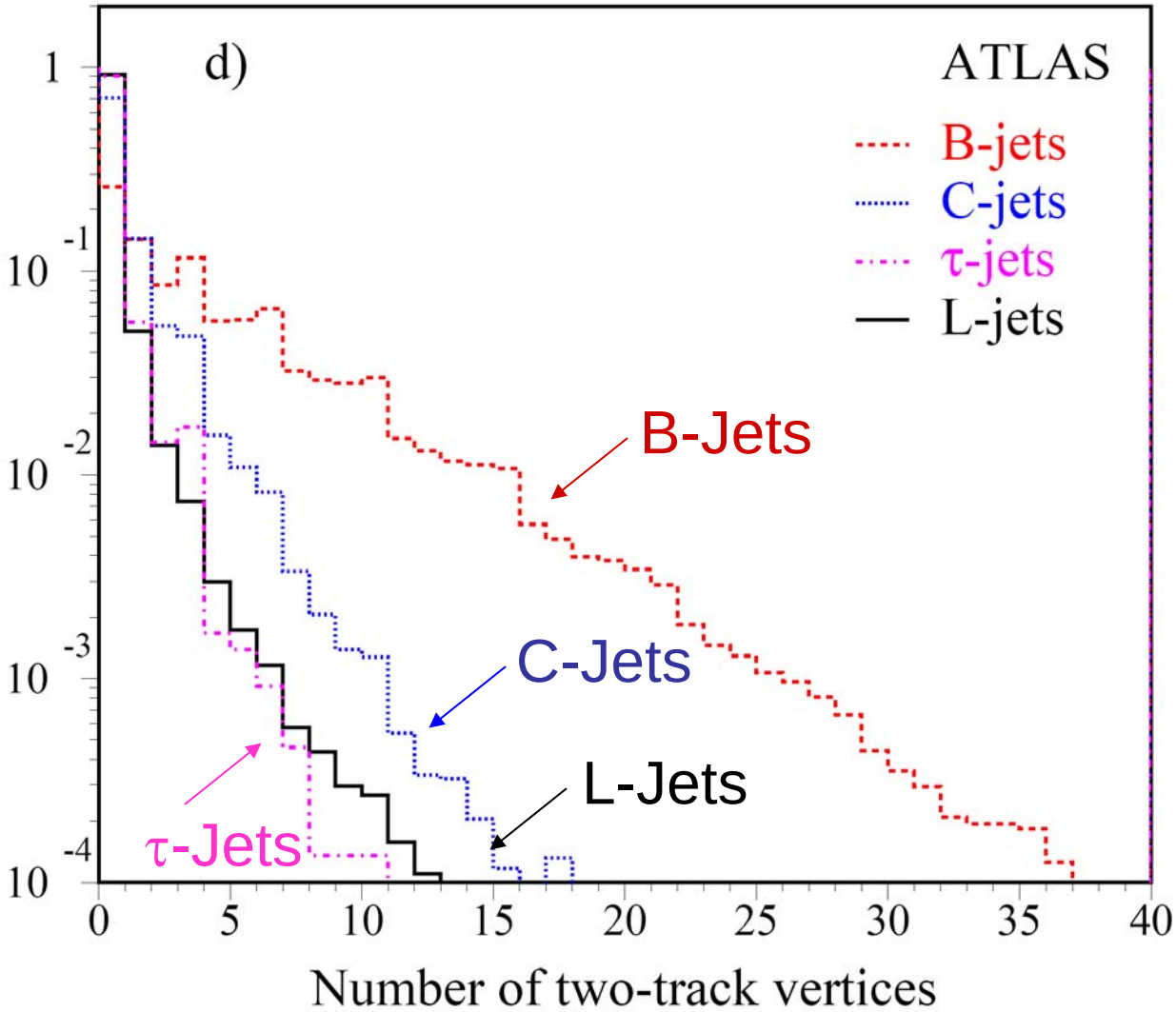
(Likelihood based)



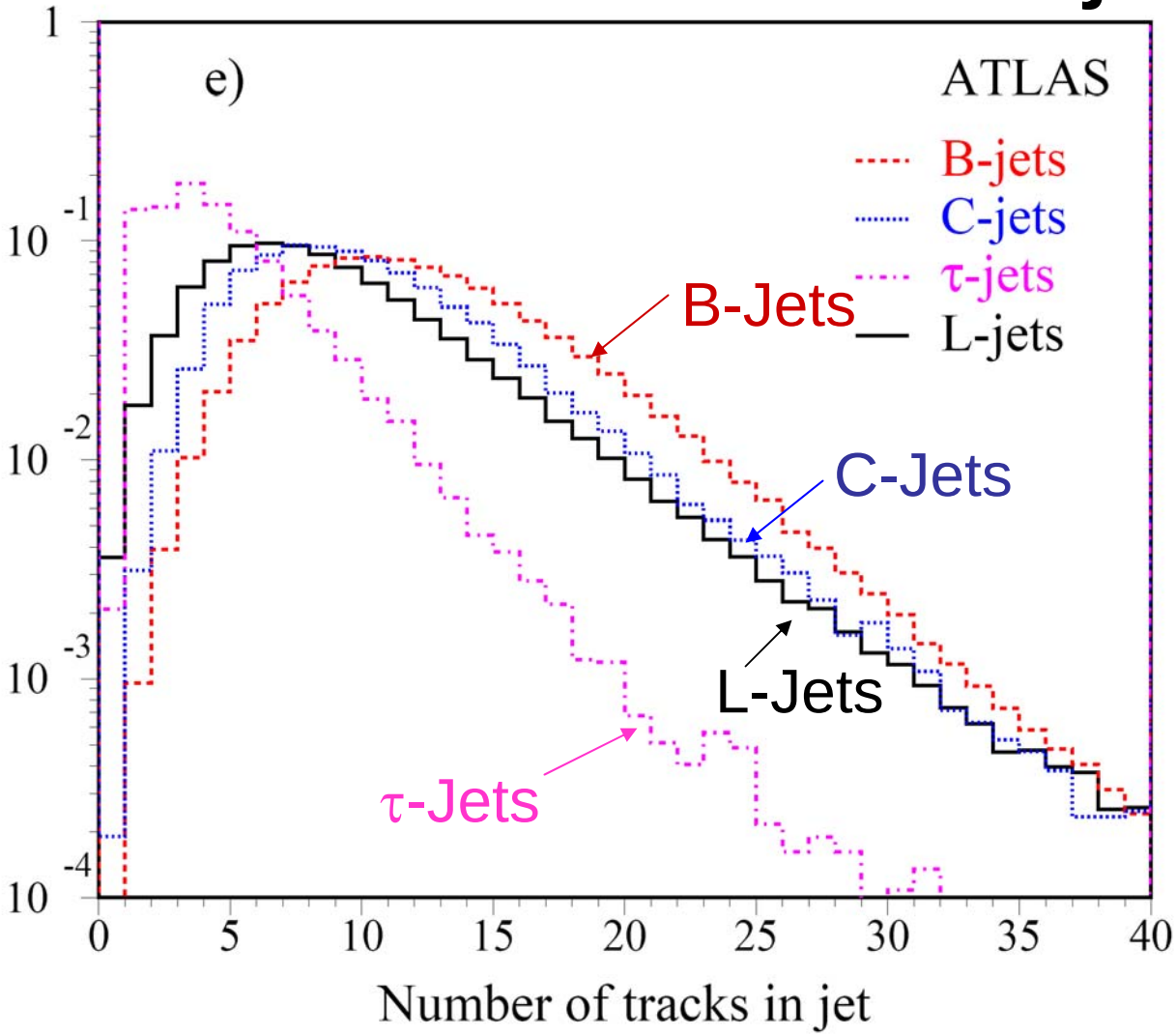
Secondary Vertex (SV1, Likelihood based)



Number of 2-track vertices



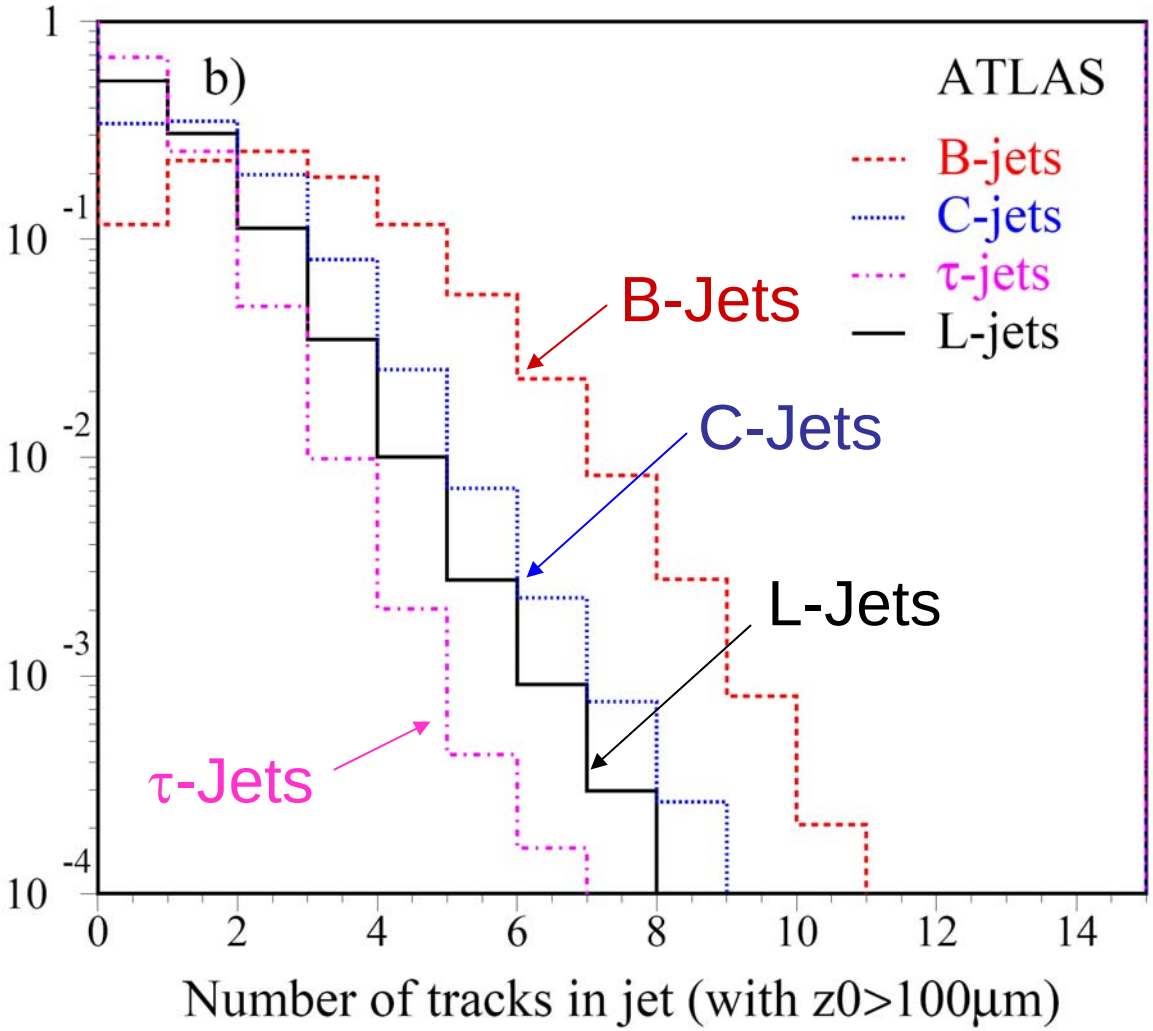
Number of tracks in jet



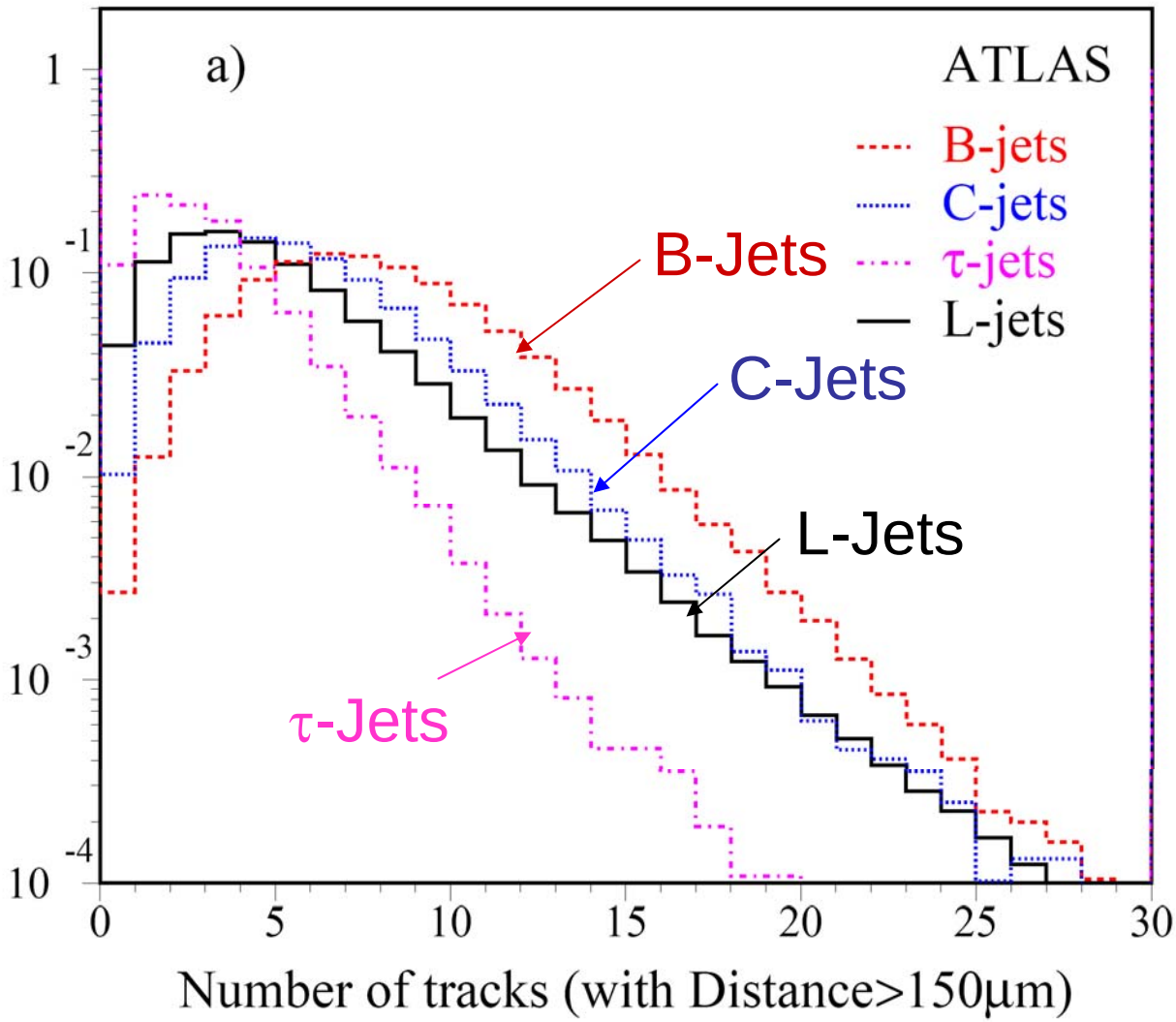
NEW variables for BDT b-tagging

- **Building additional 27 new variables, 9 are selected:**
  - Ntrack_distance_150: number of tracks in the jet with distance between PV and track-jet cross point greater than 150 microns
  - Ntrack_z0_100: number of tracks in the jet with longitudinal IP greater than 100 microns
  - Ntrack_z0_05: number of tracks in the jet with longitudinal IP significance greater than 0.5
  - 2d_dl: 2D decay length of the jet (mm)
  - d0sig_avg: average of transverse IP significance from two leading tracks which have the largest d0sig
  - d0_avg: average of transverse IP from two leading tracks
  - z0_avg: average of longitudinal IP from two leading tracks
  - Sumtrkpt_jetE: ratio between sum of track Pt w.r.t jet axis direction and jet energy
  - SumEpt_jetE: ratio between sum of Electron Pt w.r.t jet axis direction and jet energy

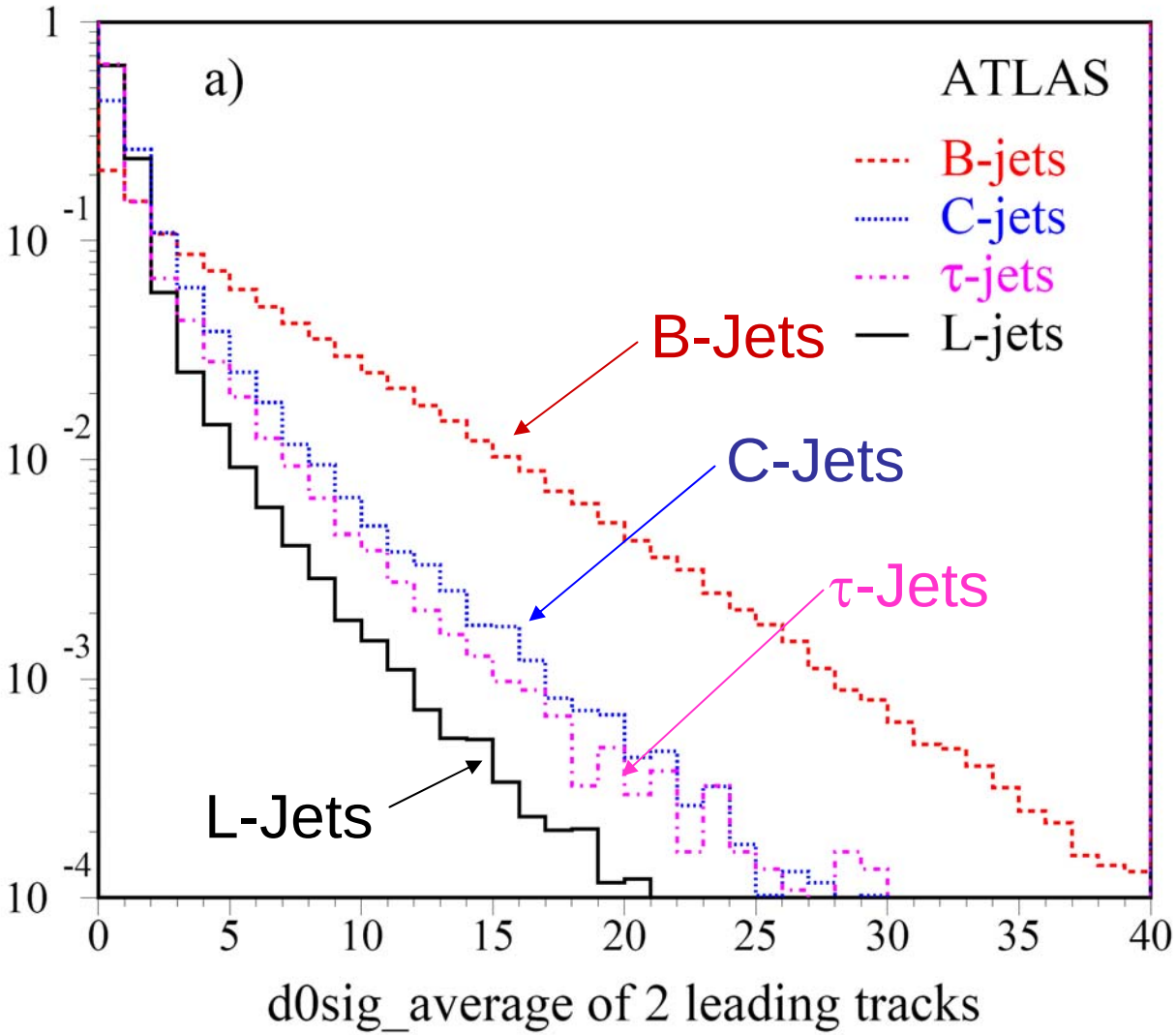
Number of tracks in jet with longitudinal IP > 100 microns (new)



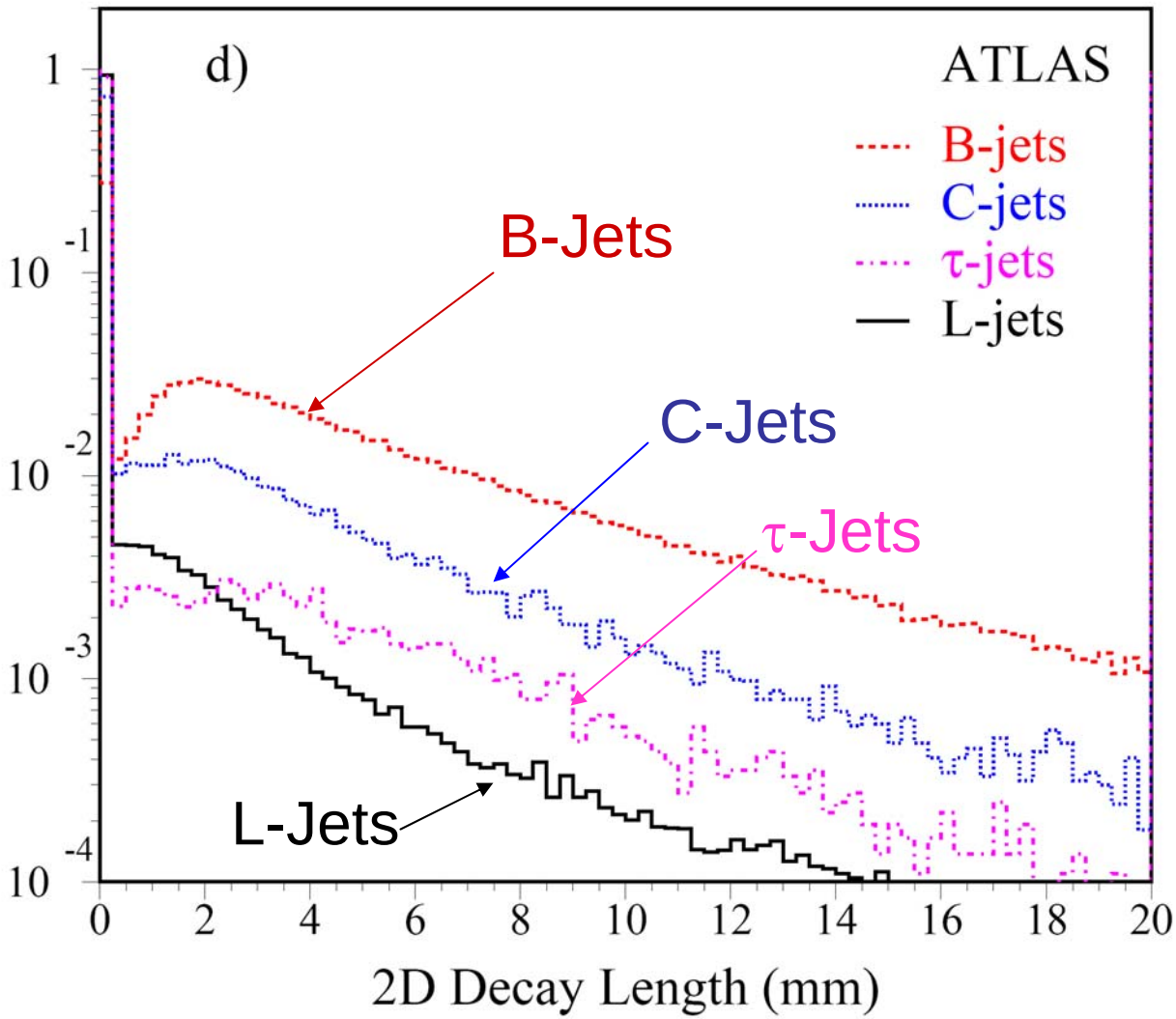
Number of tracks with distance from PV to track-jet cross point > 150microns (new)



Average of impact parameter significance from two tracks in jet with max. d0sig (new)



2D Decay Length (new)



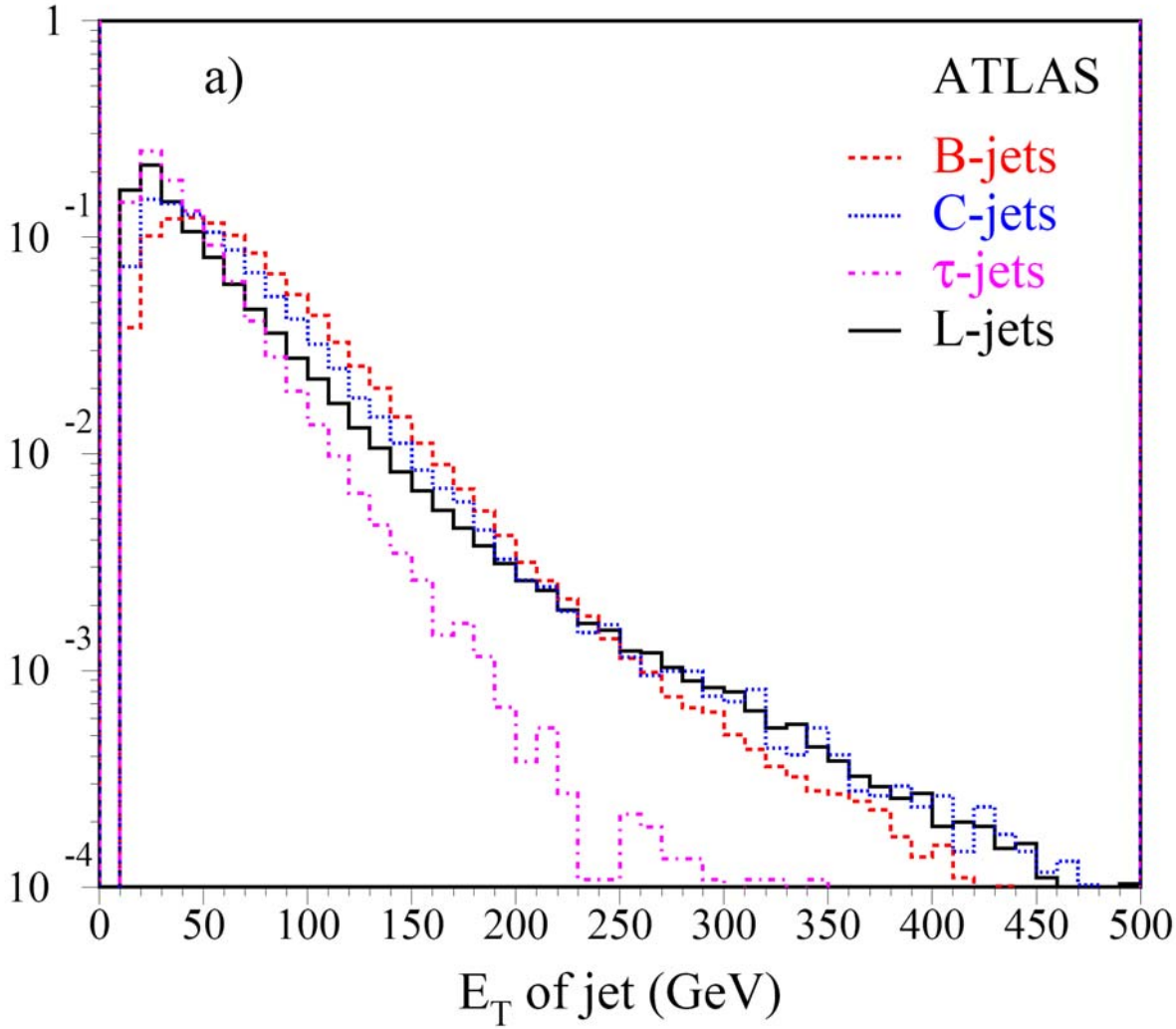
BDT B-taggers

- ➔ MC Training sample – jets with different flavors from tt events
- ➔ Test Samples (tt, WH120, WH400)

- ➔ Three – b-taggers are built :

- BDT_bl (20 vars): B Jets vs Light Jets
- BDT_bc(20 vars): B Jets vs C Jets
- BDT_bt(20 vars): B Jets vs τ Jets

Pt of jets from ttbar events



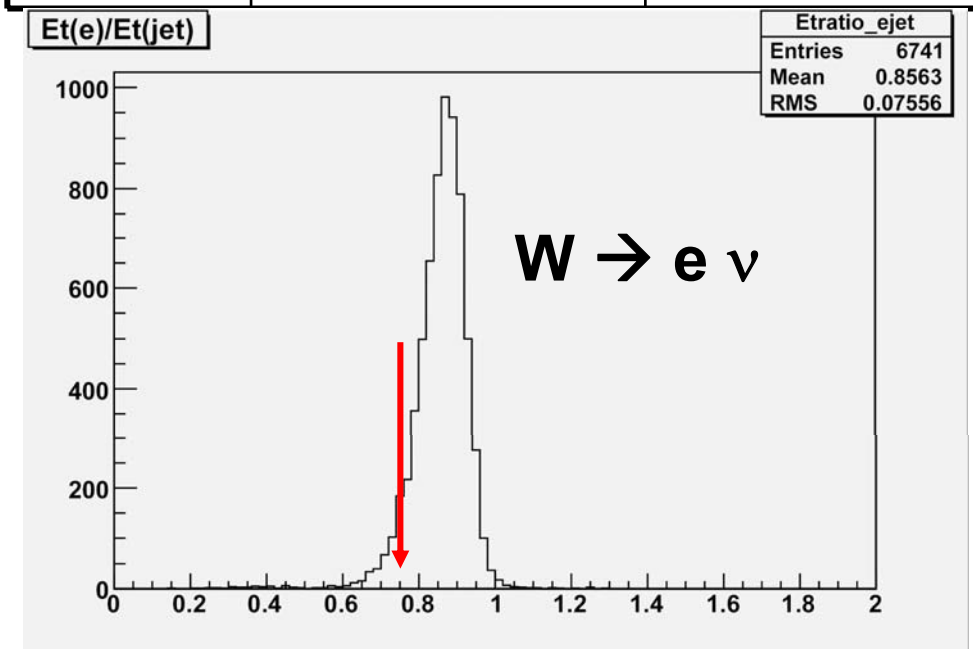
Jets for BDT Training and Test

Number of Jets $E_T > 15$ GeV, $\|\eta\| < 2.5$	For BDT Training	For Performance Test		
	$t\bar{t}$	$t\bar{t}$	WH120	WH400
b jets	150001	327222	74235	96020
c jets	31549	68480	159964	203201
τ jets	19120	37000	-	-
Light jets	223332	437369	526532	745084

Note: electrons are removed from AOD jet container
thanks Laurent to point this to us !

Effect of electron-jet removal

MC (ttbar)	No. jets before e-jet removal	No. jets after e-jet removal	Change
B-jet	327222	326953	0.08%
Light-jet	486742	437369	10.1%
C-jet	68610	68480	0.2%
τ-jet	42507	37000	13.0%



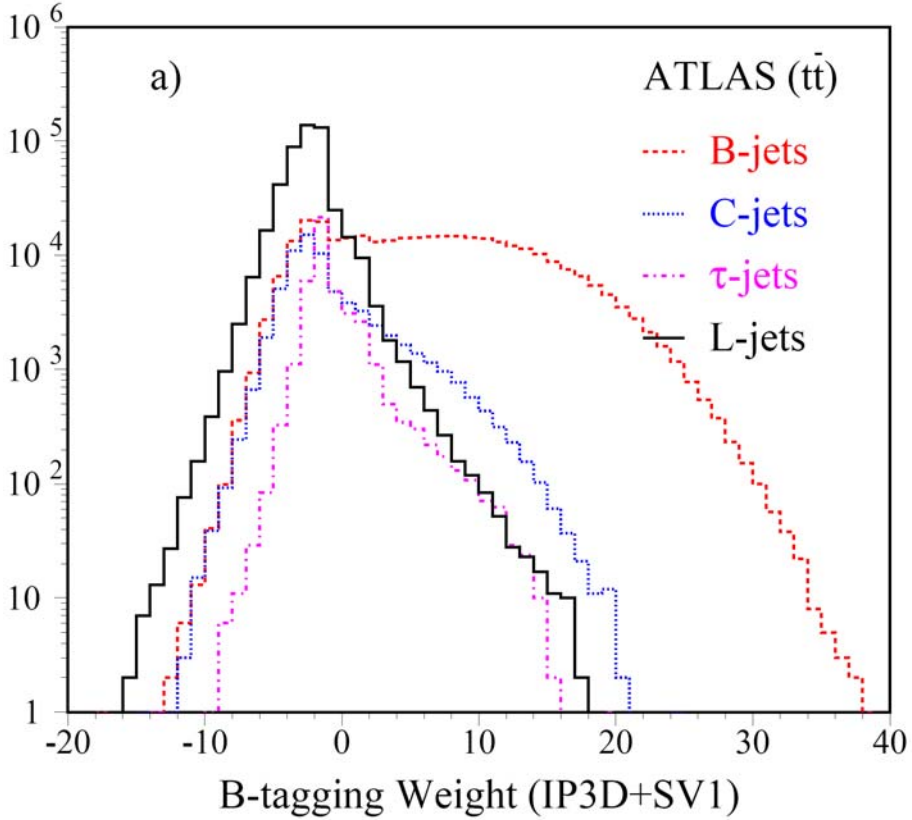
➔**The "change" reflects the non-b-jet rejection rate drops compared to those without electron-jet removal**

➔E-jet removal criteria:
ΔR (e, jet) < 0.1
ET(e) / ET(jet) > 0.75

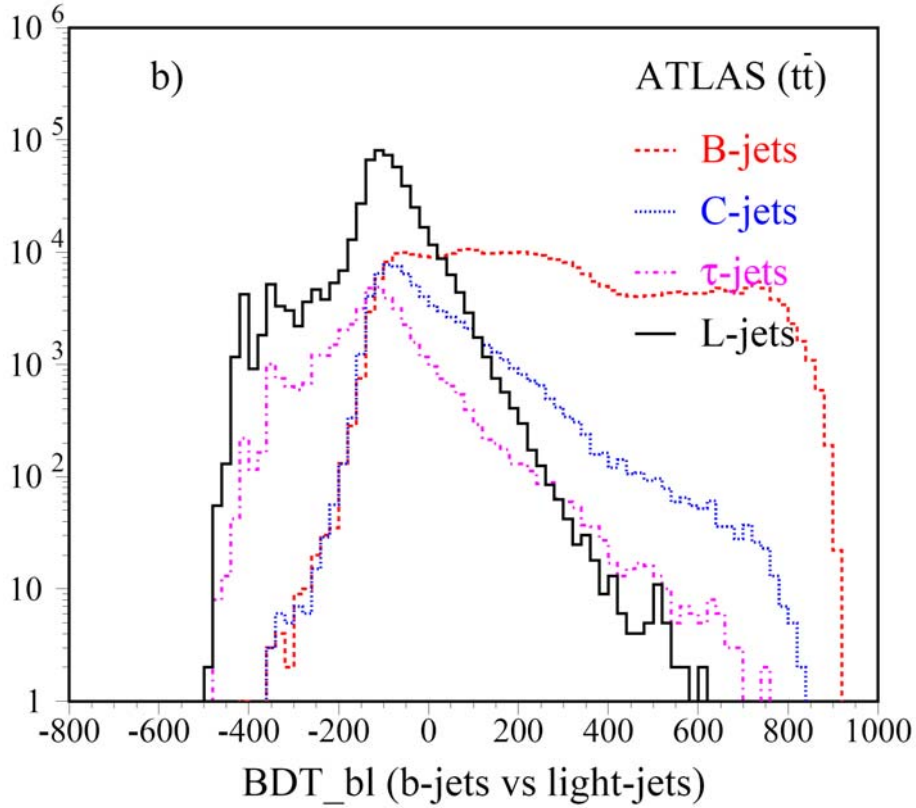
➔About 6% e-jet left

B-tagging Discriminators

IP3D+SV1



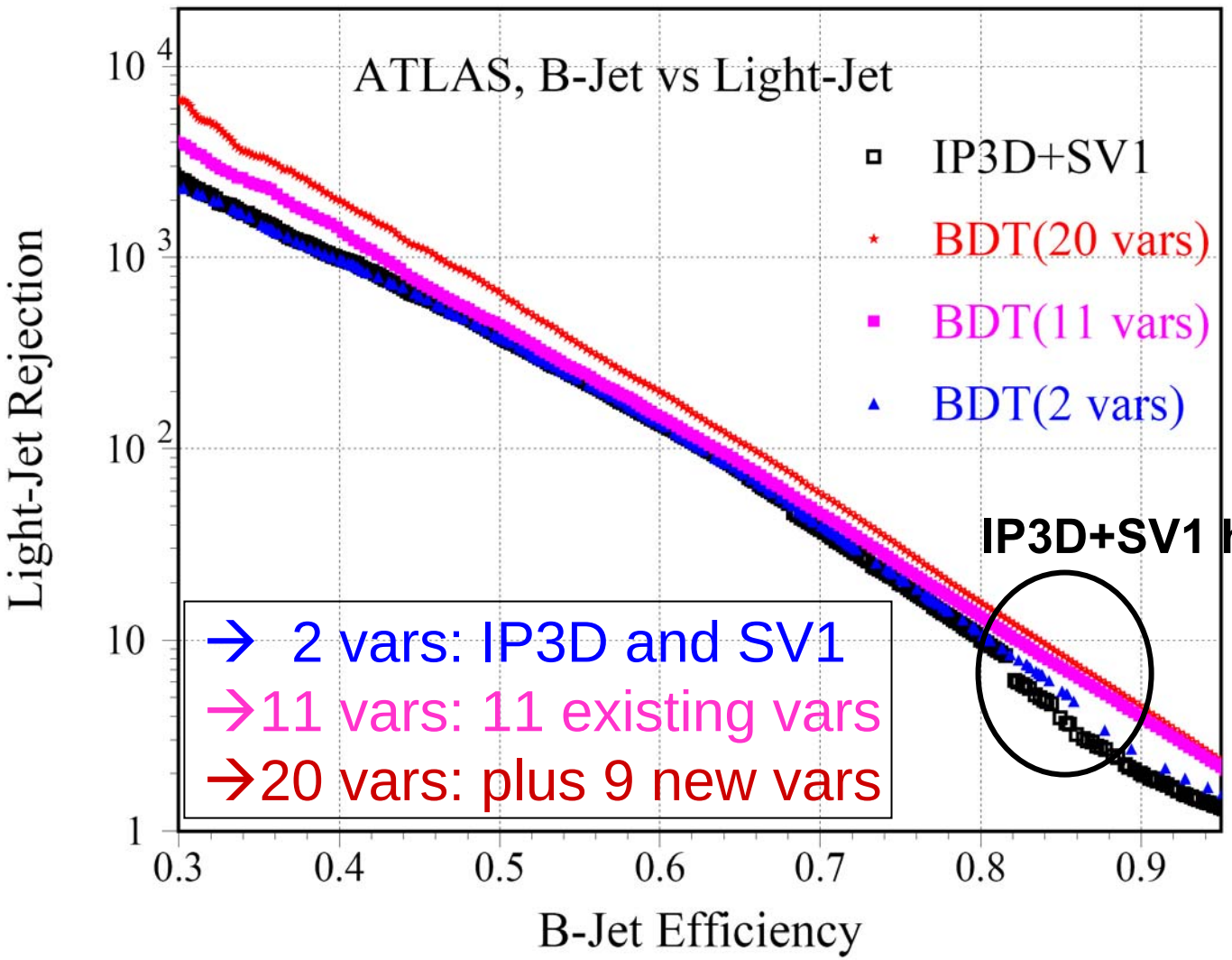
BDT_bl



Goodness of BDT Input Variables

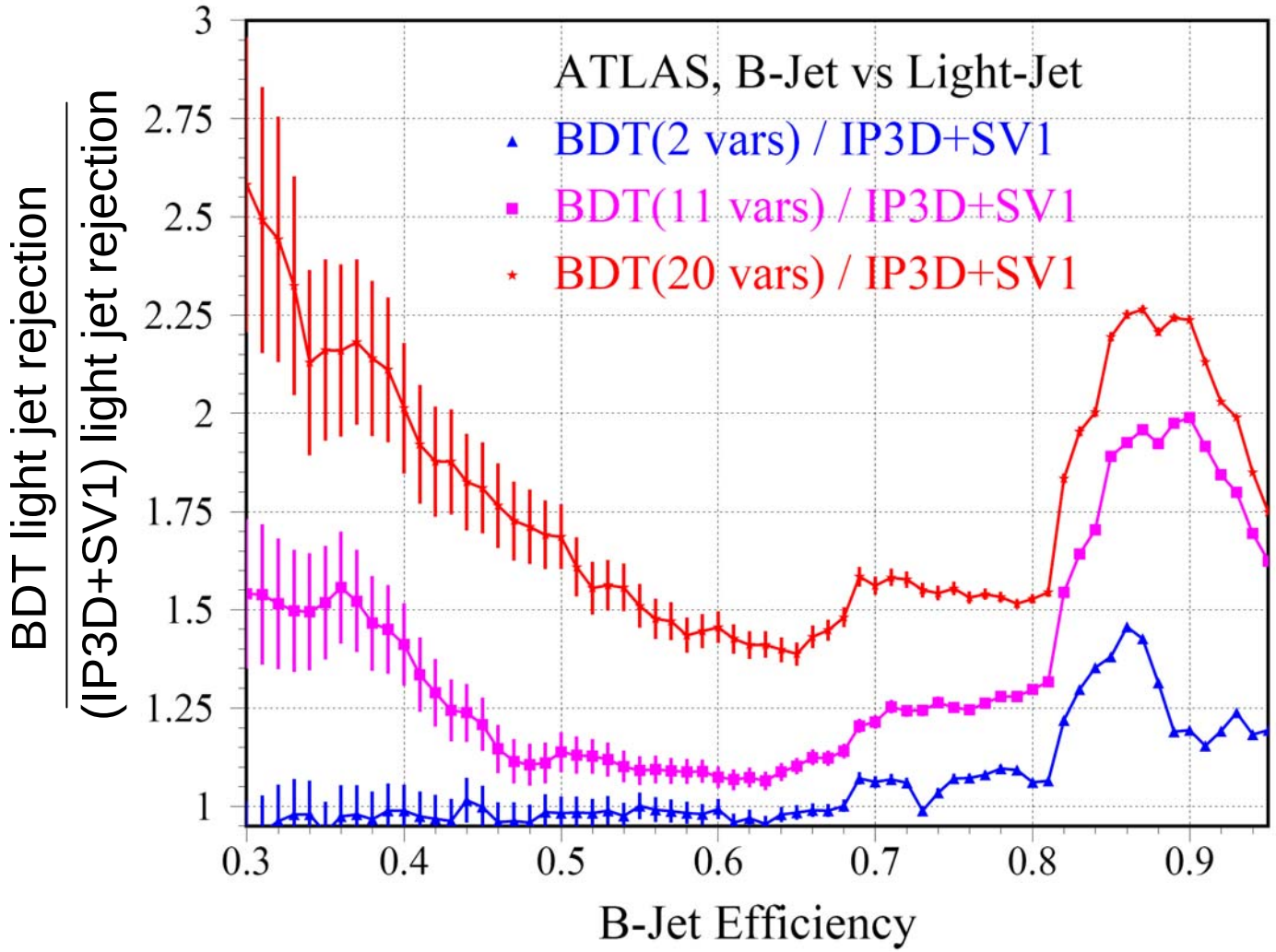
BDT Input Variables	Relative *gini index* Contribution (%)		
	BDT_bl(for light-jets)	BDT_bc(for c-jets)	BDT_bt(for τ-jets)
Eleven Existing Discriminating Variables			
IP2D	0.4	0.73	0.56
IP3D	**26.3**	**12.75**	**1.97**
SV1	**42.24**	**55.18**	**20.06**
softe	**1.11**	0.83	0.37
d0sig_max	0.4	0.65	0.55
z0sig_max	**1.63**	**1.14**	0.77
mass	0.18	**8.75**	**8.14**
efrac	**13.39**	**1.16**	1.54
pt_max	**1.31**	0.61	**4.80**
nvertex_2track	0.08	**1.31**	**3.26**
ntrack	**2.84**	0.60	**45.43**
Nine New Discriminating Variables			
ntrack_distance_150	**5.76**	**2.56**	**2.46**
ntrack_z0_100	**1.44**	0.26	**1.93**
ntrack_z0sig_05	0.12	0.38	1.62
2d_dl	**0.76**	**4.76**	**1.98**
d0sig_avg	0.36	0.63	0.36
d0_avg	0.20	1.04	1.14
z0_avg	0.41	**4.77**	**2.09**
sumtrkpt_jetE	0.39	0.29	0.30
sumept_jetE	0.67	**1.60**	0.69

Light-jet Rejection vs. B-jet Eff.



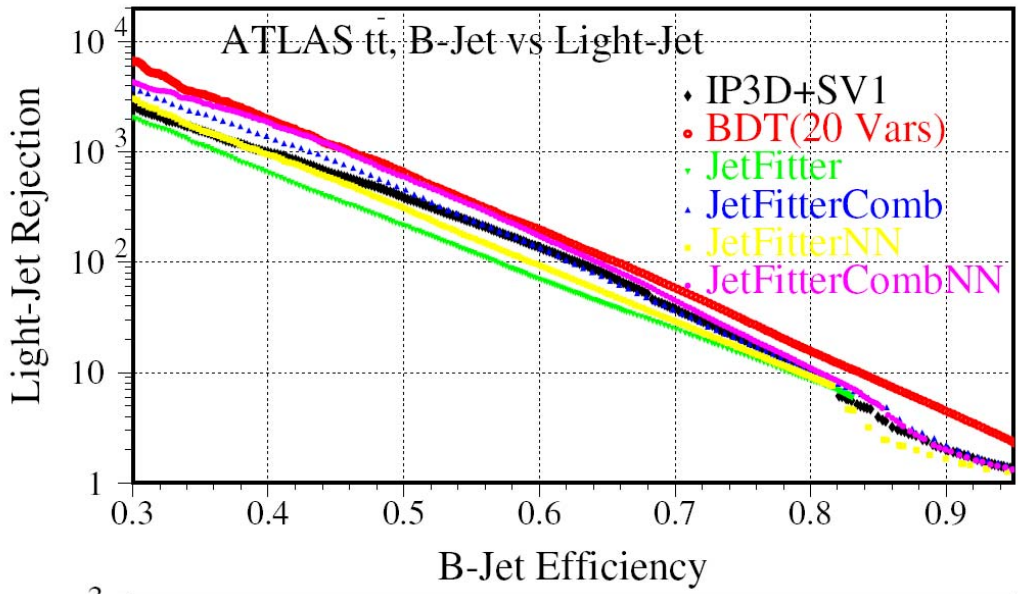
IP3D+SV1 has structure

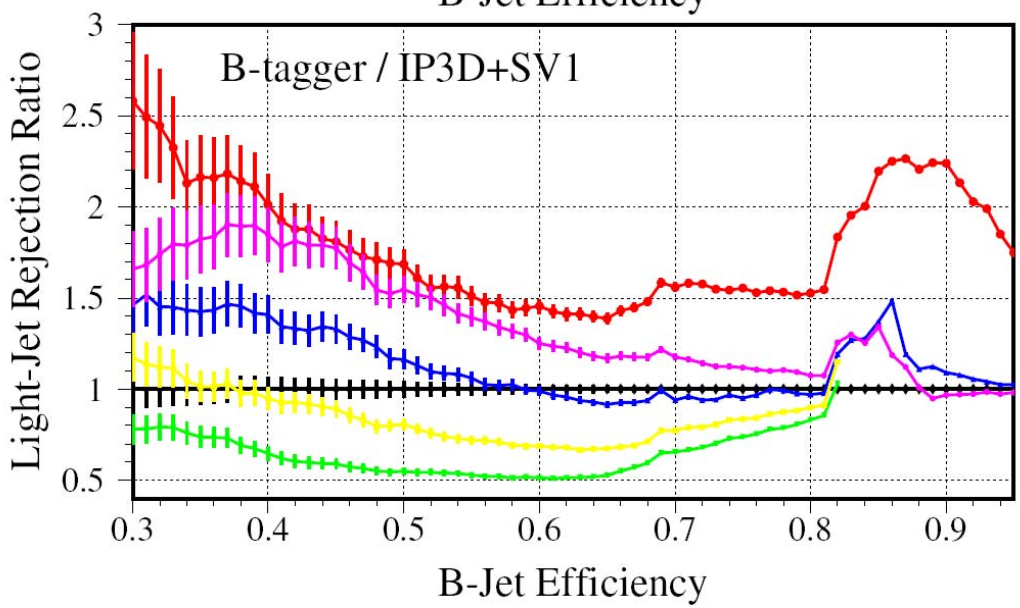
B-tagging Performance Comparison vs. Number of BDT Input Variables



More Comparisons use ttbar samples

Light jet rejection vs b-tagging efficiency





Light-Jet Rejection (B-tagger)

Light-Jet Rejection (IP3D+SV1)

Light-jet rejection for 50% -70% b-tagging efficiencies

Test Sample $\sqrt{s} = 10$ TeV	b-jet Efficiency	light-jet Rejection		
		IP3D+SV1	JetFitterCombNN	BDT_bl
$t\bar{t}$	70%	38.0 ± 0.4	44.7 ± 0.5	59.4 ± 0.7
$t\bar{t}$	60%	136.2 ± 2.4	170.5 ± 3.4	198.2 ± 4.2
$t\bar{t}$	50%	389.3 ± 11.6	601.9 ± 22.4	656.4 ± 25.5
$WH(120\ \text{GeV})$	70%	30.5 ± 0.2	33.9 ± 0.3	39.5 ± 0.4
$WH(120\ \text{GeV})$	60%	123.3 ± 1.9	151.5 ± 2.6	167.9 ± 3.0
$WH(120\ \text{GeV})$	50%	474.4 ± 14.3	666.5 ± 23.7	740.9 ± 27.8
$WH(400\ \text{GeV})$	70%	44.7 ± 0.4	50.3 ± 0.4	56.7 ± 0.5
$WH(400\ \text{GeV})$	60%	142.6 ± 2.0	173.6 ± 2.7	193.6 ± 3.1
$WH(400\ \text{GeV})$	50%	426.6 ± 10.2	555.0 ± 15.2	651.3 ± 19.3

- BDT_bl is trained using b-jets as signal and light-jets as background with 20 input variables

C-jet rejection comparison with BDT_bl and BDT_bc

Test Sample $\sqrt{s} = 10$ TeV	b-jet Eff.	c-jet Rejection					
		IP3D+SV1	JetFitter	BDT_bl (for light-jets)		BDT_bc (for c-jets)	
			CombNN	20 Vars	22 Vars	20 Vars	22 Vars
$t\bar{t}$	70%	4.9 ± 0.05	5.0 ± 0.05	5.1 ± 0.05	5.3 ± 0.05	6.3 ± 0.06	6.6 ± 0.07
$t\bar{t}$	60%	8.4 ± 0.10	8.3 ± 0.10	8.4 ± 0.10	8.7 ± 0.1	12.1 ± 0.17	12.9 ± 0.2
$t\bar{t}$	50%	14.7 ± 0.22	14.1 ± 0.21	14.0 ± 0.21	14.9 ± 0.2	25.5 ± 0.50	29.8 ± 0.6
WH(120 GeV)	70%	4.5 ± 0.03	4.5 ± 0.03	4.4 ± 0.03	4.6 ± 0.03	5.2 ± 0.03	5.2 ± 0.03
WH(120 GeV)	60%	7.7 ± 0.06	7.8 ± 0.06	7.5 ± 0.05	7.9 ± 0.06	9.9 ± 0.08	10.4 ± 0.09
WH(120 GeV)	50%	14.3 ± 0.14	13.8 ± 0.13	13.6 ± 0.13	14.2 ± 0.14	20.9 ± 0.24	23.8 ± 0.3
WH(400 GeV)	70%	4.9 ± 0.03	5.3 ± 0.03	4.9 ± 0.03	5.5 ± 0.03	5.9 ± 0.03	6.4 ± 0.04
WH(400 GeV)	60%	8.8 ± 0.06	9.1 ± 0.06	8.5 ± 0.06	9.3 ± 0.07	11.3 ± 0.09	12.5 ± 0.1
WH(400 GeV)	50%	16.2 ± 0.15	16.1 ± 0.15	15.2 ± 0.14	16.4 ± 0.15	23.8 ± 0.26	28.3 ± 0.34

- BDT_bl is trained using b-jets as signal and light-jets as background
  → BDT_bl performance is comparable with other b-tagger

- BDT_bc is trained using b-jets as signal and c-jets as background
  → BDT_bc performance has significant improvement (~40%)

τ-jet rejection comparison with BDT_bl and BDT_bt

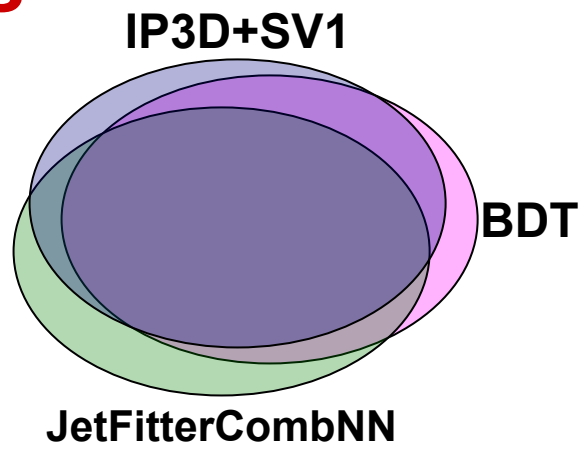
Test Sample $\sqrt{s} = 10$ TeV	b-jet Eff.	τ-jet Rejection					
		IP3D+SV1	JetFitter	BDT_bl		BDT_bt	
			CombNN	20 Vars	22 Vars	20 Vars	22 Vars
$t\bar{t}$	70%	8.6 ± 0.1	10.3 ± 0.2	17.9 ± 0.4	18.0 ± 0.4	67 ± 2.9	75.7 ± 3.5
$t\bar{t}$	60%	24.2 ± 0.6	22.2 ± 0.6	30.2 ± 0.9	31.7 ± 0.9	141.0 ± 8.7	172.9 ± 11.9
$t\bar{t}$	50%	46.9 ± 1.7	46.3 ± 1.7	51.0 ± 1.9	55.7 ± 2.2	338.2 ± 32.4	501.4 ± 58.4

- BDT_bl is trained using b-jets as signal and light-jets as background
- BDT_bt is trained using b-jets as signal and τ-jets as background

→Both BDT_bl and BDT_bt have better performance compared to other b-taggers

Cross checks of ATLAS Btaggers: Overlapped B-jets

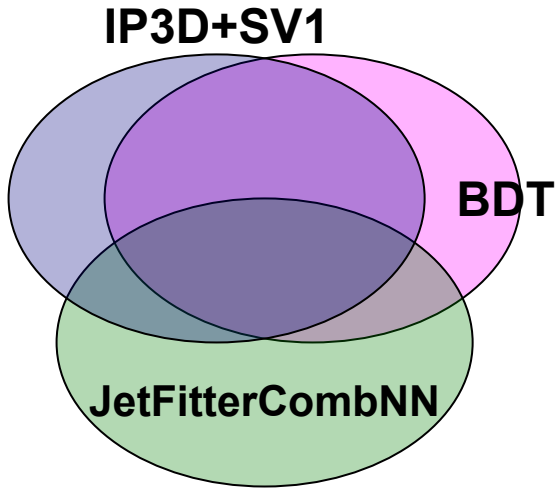
- Apply cuts for IP3D+SV1, BDT and JetFitterCombNN with 60% of B jet efficiency, respectively.
- Then calculate the overlapped B-jets passed these cuts. ***Overlapped efficiency = (A.and.B) / (A.or.B)***



No. of B-jets	IP3D+SV1	JetFitterCombNN	BDT_bl
IP3D+SV1	198196	182993/213399 = 85.8%	190827/205564 = 92.8%
JetFitterCombNN		198196	182386/214006 = 85.2%
BDT_bl			198196

Btaggers: Overlapped Light-jets

- Apply cuts for IP3D+SV1, BDT and JetFitterCombNN with 60% of B jet efficiency, respectively.
- Then calculate the overlapped light-jets passed these cuts. ***Overlapped efficiency = (A.and.B) / (A.or.B)***



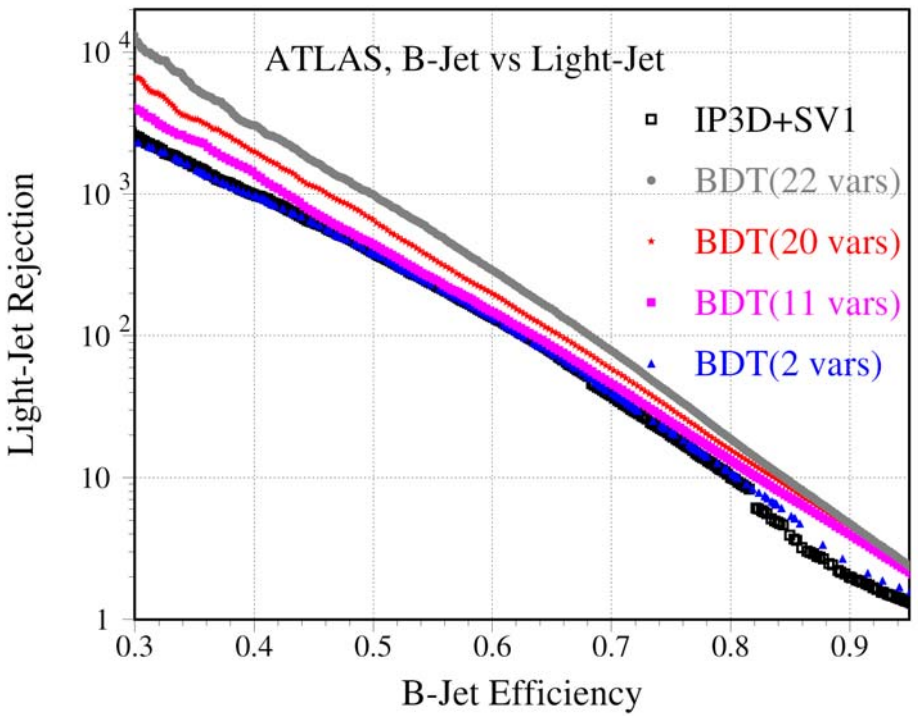
No. of L-jets	IP3D+SV1	JetFitterCombNN	BDT_bl
IP3D+SV1	3536	1446/4875=30%	2027/3886=52%
JetFitterCombNN		2785	1178/3984=30%
BDT_bl			2377

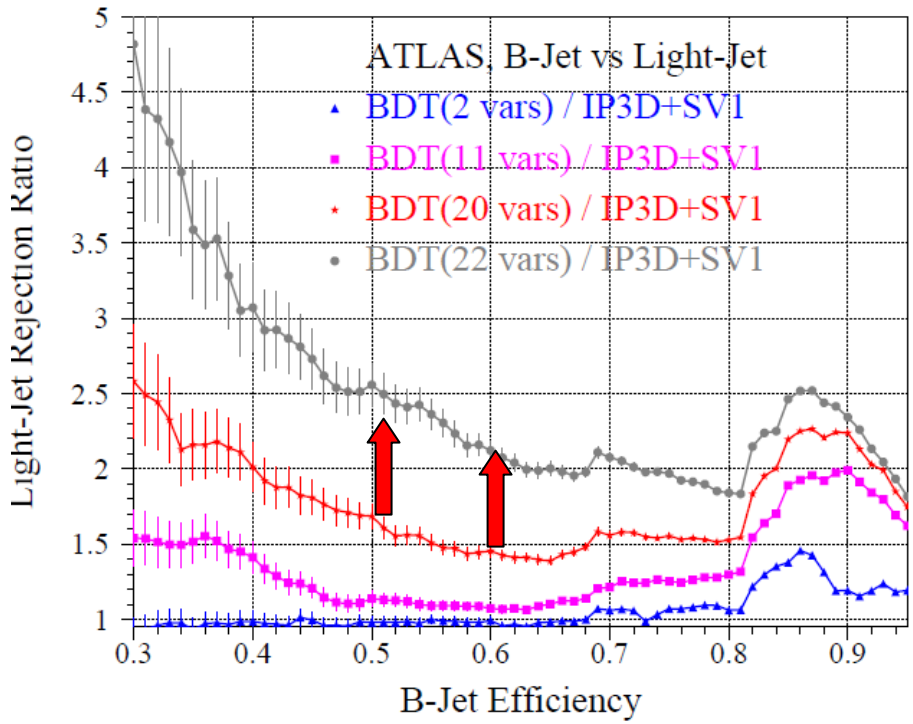
BDT B-tagger: Further Improvement

→ High b-jet overlapped efficiency (>85%) indicates that B-taggers have reliable performance to tag b-jet

→ Low light-jet overlapped rate (~30%) indicates that better light-jet rejection can be achieved by combining B-taggers

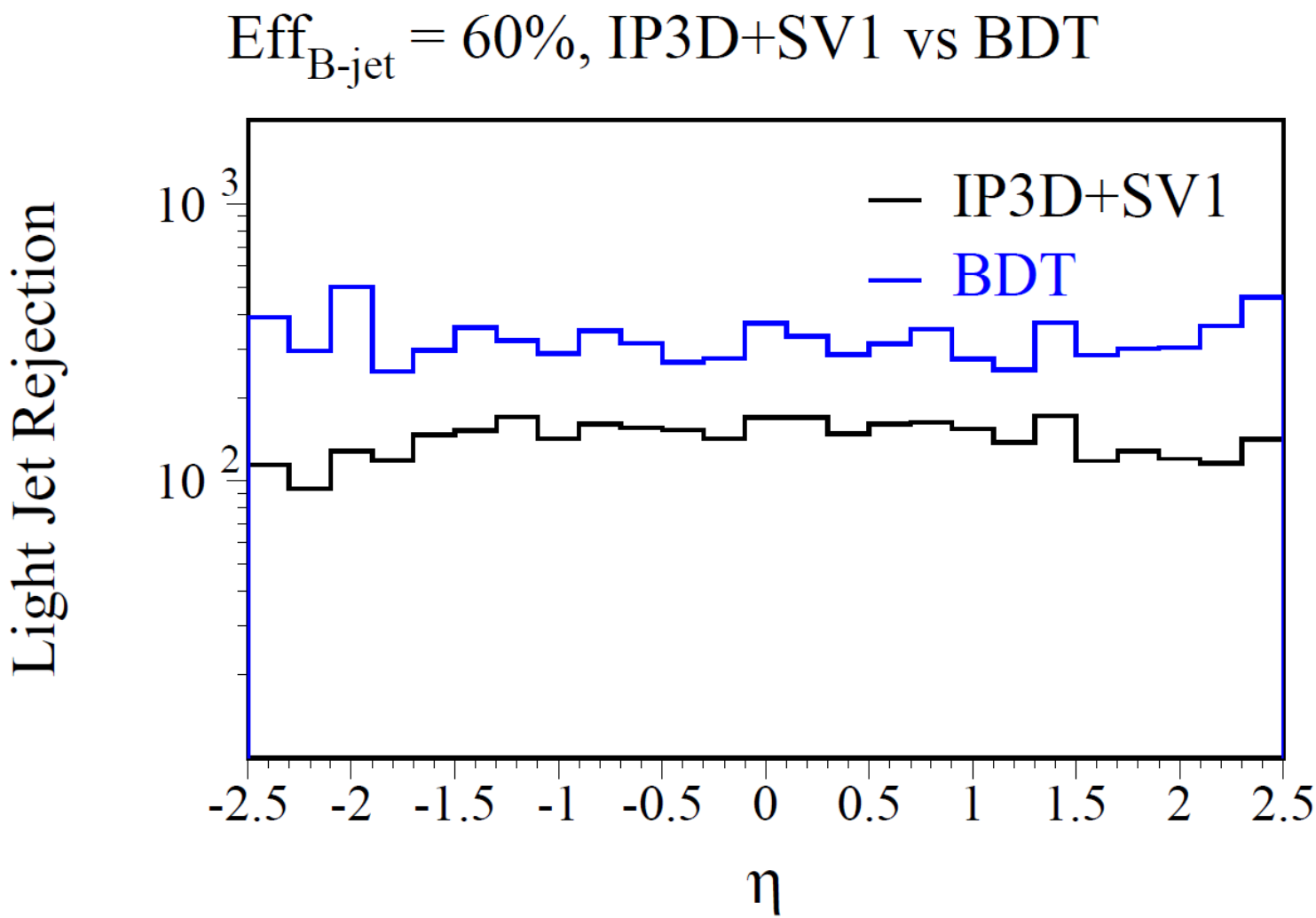
→ 22 vars: 20 vars plus IP3D+SV1 and JetFitterCombNN



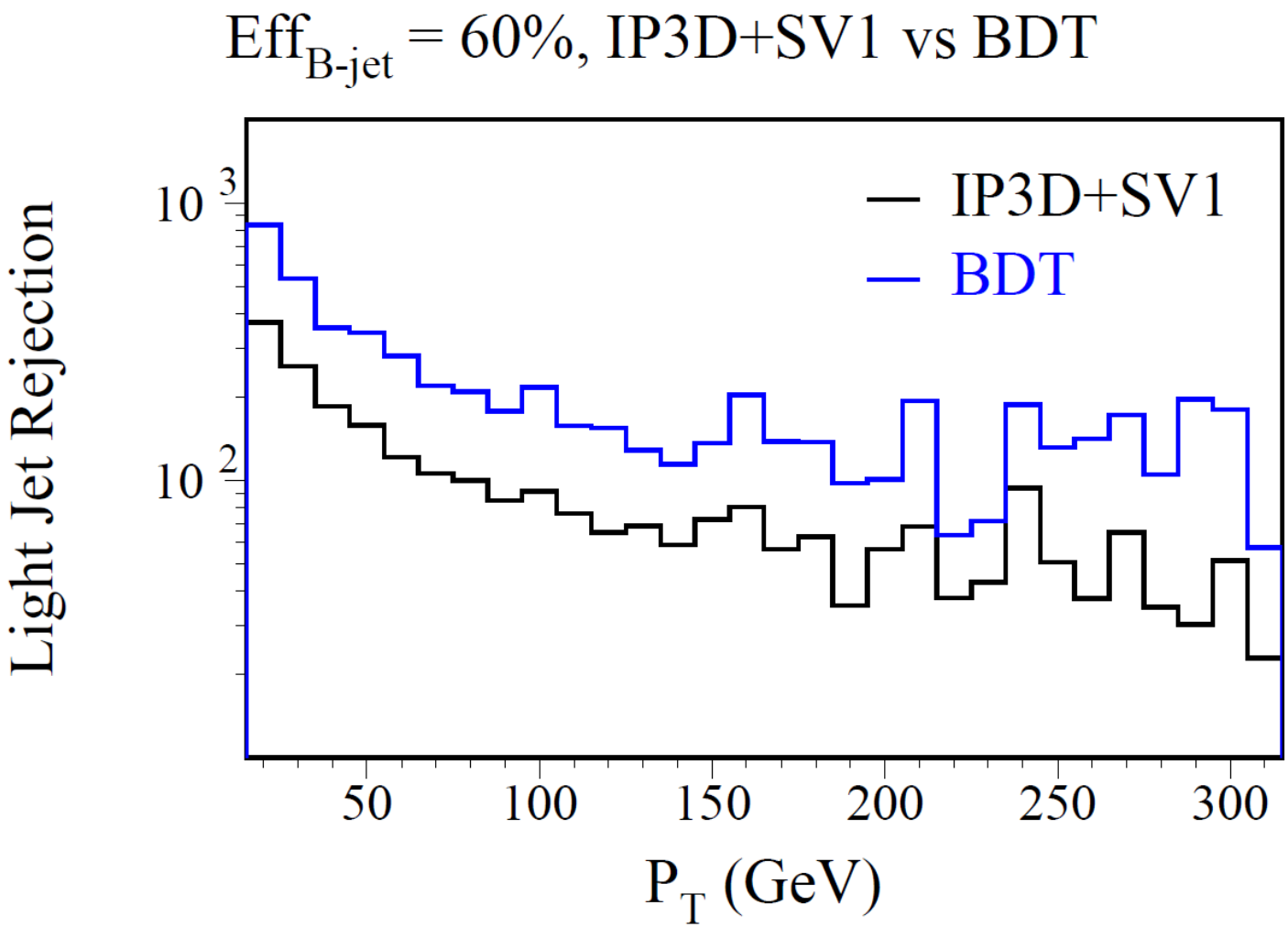


BDT_bl (22 Vars) vs IP3D+SV1

(light jet rejection vs η)



BDT_bl (22 Vars) vs IP3D+SV1
(light jet rejection vs P_T)



Summary and Plan

- BDT b-tagging development and performance comparisons are presented based on v14 MC samples produced at 10 TeV center-of-mass energy.

- We plan to continue test the BDT b-tagging and to perform the b-tagging algorithm comparison based on v15 samples.

- We implemented and fully tested BDT b-tagging in offline Physics Analysis programs (private).

- We will follow suggestion of b-tagging group conveners to implement the BDT b-tagging into the ATLAS official b-tagging package.

Backup Slides

Criterion for “Best” Tree Split

- Purity, *P*, is the fraction of the weight of a node (leaf) due to signal events.
- Gini Index: Note that Gini index is 0 for all signal or all background.

$$Gini = (\sum_{i=1}^{n} W_i) P(1-P)$$

- The criterion is to minimize

  Gini_left_node+ Gini_right_node.

Criterion for Next Node to Split

- Pick the node to maximize the change in Gini index. Criterion =

$$Gini_{parent\_node} - Gini_{right\_child\_node} - Gini_{left\_child\_node}$$

- We can use Gini index contribution of tree split variables to sort the importance of input variables.
- We can also sort the importance of input variables based on how often they are used as tree splitters.

Signal and Background Leaves

- Assume an equal weight of signal and background training events.
- If event weight of signal is larger than ½ of the total weight of a leaf, it is a signal leaf; otherwise it is a background leaf.
- Signal events on a background leaf or background events on a signal leaf are misclassified events.

How to Boost Decision Trees ?

- For each tree iteration, same set of training events are used but the weights of misclassified events in previous iteration are increased (boosted). Events with higher weights have larger impact on Gini index values and Criterion values. The use of boosted weights for misclassified events makes them possible to be correctly classified in succeeding trees.
- Typically, one generates several hundred to thousand trees until the performance is optimal.
- The score of a testing event is assigned as follows: If it lands on a signal leaf, it is given a score of 1; otherwise -1. The sum of scores (weighted) from all trees is the final score of the event.

Two Boosting Algorithms

- AdaBoost Algorithm:

1. Initialize the observation weights $w_i = 1/\mathrm{n}$, i = 1, 2,..., n
2. For m = 1 to M:
   2.a Fit a classifier $T_m(x)$ to the training data using weights w_i
   2.b Compute

$$err_m = \frac{\sum_{i=1}^{n} w_i I(y_i \neq T_m(x_i))}{\sum_{i=1}^{n} w_i}$$

I = 1, if a training event is misclassified; Otherwise, I = 0

   2.c Compute $\alpha_m = \beta \times log((1 - err_m)/err_m)$
   2.d Set $w_i \leftarrow w_i \times exp(\alpha_m I(y_i \neq T_m(x_i)))$, i=1, 2,...,n
   2.e Re-normalize $w_i = w_i / \sum_{i=1}^{n} w_i$
3. Output T(x) $= \sum_{m=1}^{M} \alpha_m T_m(x)$

- $\epsilon-$boosting Algorithm:

1. Initialize the observation weights $w_i = 1/\mathrm{n}$, i = 1, 2,..., n
2. For m = 1 to M:
   2.a Fit a classifier $T_m(x)$ to the training data using weights w_i
   2.b Set $w_i \leftarrow w_i \times exp(2\epsilon I(y_i \neq T_m(x_i)))$, i=1, 2,...,n
   2.c Re-normalize $w_i = w_i / \sum_{i=1}^{n} w_i$
3. Output T(x) $= \sum_{m=1}^{M} \epsilon T_m(x)$

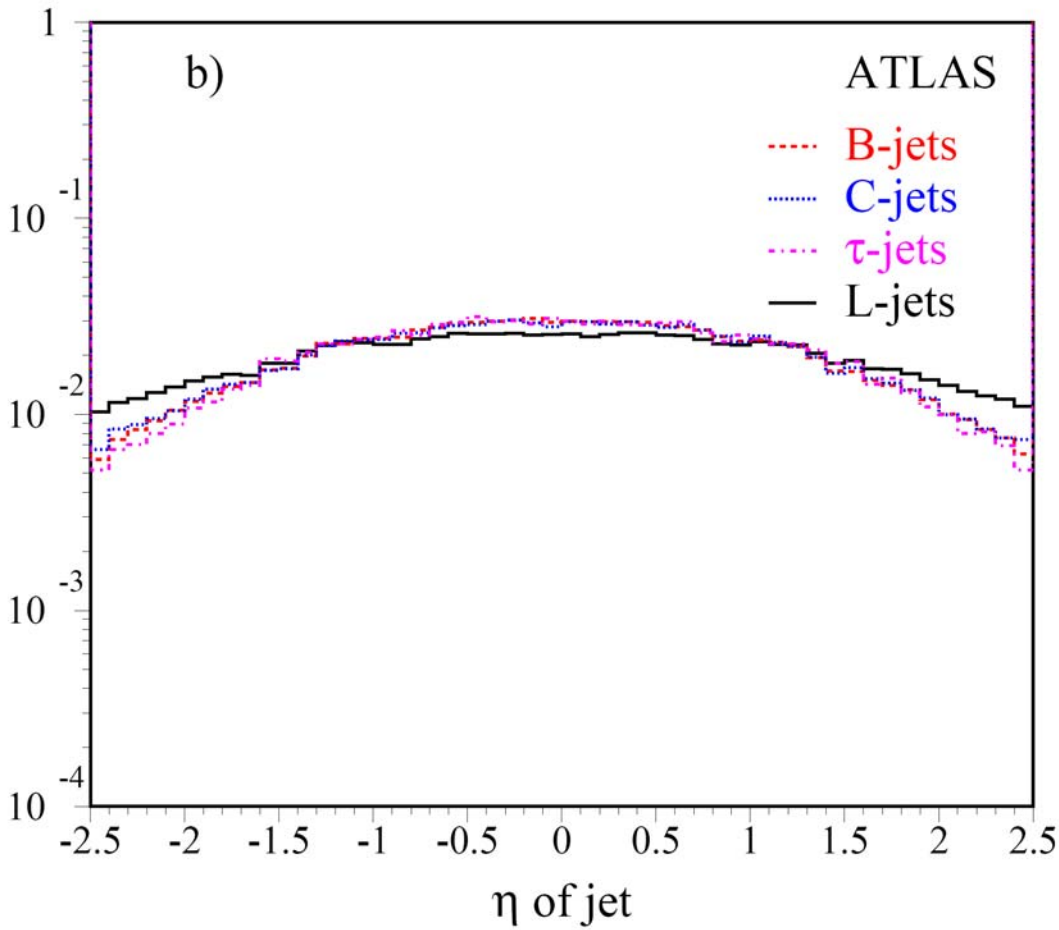
Example

- **AdaBoost: the weight of misclassified events is increased by**
  - error rate=0.1 and $\beta = 0.5$, $\alpha_m = 1.1$, exp(1.1) = 3
  - error rate=0.4 and $\beta = 0.5$, $\alpha_m = 0.203$, exp(0.203) = 1.225
  - Weight of a misclassified event is multiplied by a large factor which depends on the error rate.

- **ε–boost: the weight of misclassified events is increased by**
  - If $\varepsilon = 0.01$, exp(2*0.01) = 1.02
  - If $\varepsilon = 0.04$, exp(2*0.04) = 1.083
  - It changes event weight a little at a time.

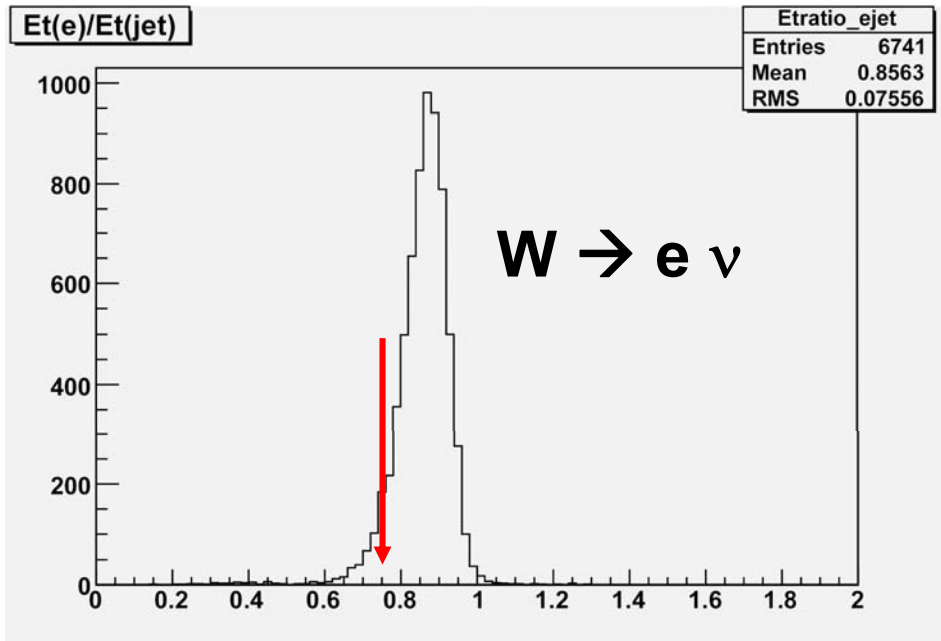
➔ AdaBoost converges faster than ε-boost. However, the performance of AdaBoost and ε–boost are very comparable with sufficient tree iterations.

Eta of Jets from ttbar events



Electron-Jet Removal

- Electron was put in jet container, that should be removed in jet counting. E-jet removal based on ATLAS CSC book,
  - CERN-OPEN-2008-202 (2009), page 414



- E-jet removal criteria:
  - ΔR (e, jet) < 0.1
  - $E_T(e) / E_T(jet) > 0.75$
- About 6% e-jet left

Effect of Electron fake jet removal on Light-jet rejection

MC	Eff_b	IP3D+SV1	JetFitter CombNN	BDT_bl (22 Vars)
Ttbar	70%	35 → 38	48 →45	65 → 59
Ttbar	60%	146 → 136	188 → 171	219 → 198
Ttbar	50%	429 → 389	663 → 602	725→ 656

→The light-jet rejection rates drop ~10% after electron fake jet removal

Effect of Electron fake jet removal

MC	No jets before e-jet removal	No jets after e-jet removal	Change (Light-jets)
ttbar	486742	437369	10.1%
WH120	545190	526532	3.4%
WH400	753547	745084	1.1%

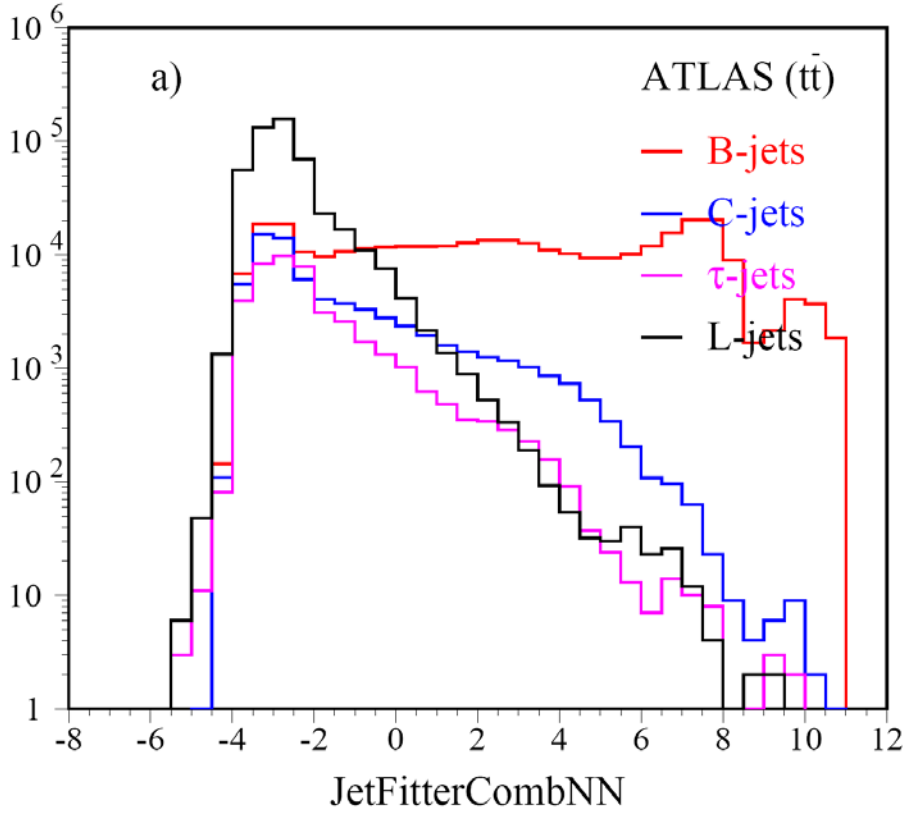
Effects of Soft Electron Tagger (softe)

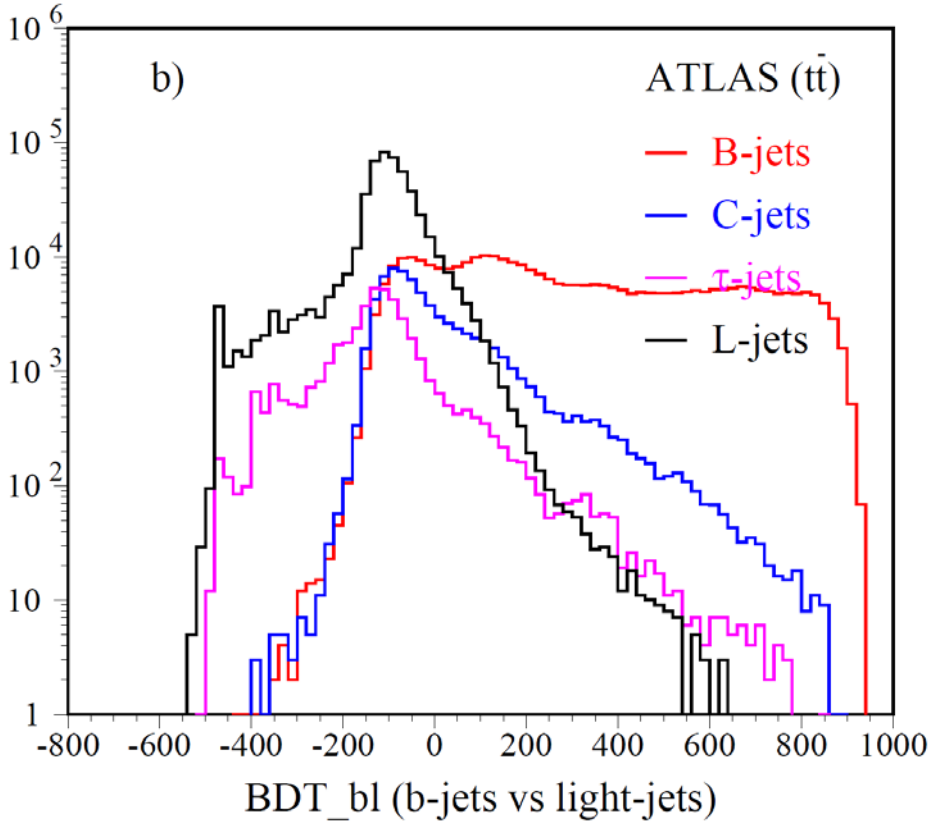
Training MC Samples	Test MC Samples	Input Vars (w/wo softe)	Light-Jet Rejection at 60% Eff_bjet	Gini Index contribution
With e-jet	with e-jet	using softe	219 ± 5	1.11 %
With e-jet	no e-jet	using softe	198 ± 4.2	1.11 %
No e-jet	no e-jet	using softe	196 ± 4.2	0.28 %
No e-jet	no e-jet	Remove softe	190 ± 4.0	0 %

➔ By removing electron fake jets (e-jet) in BDT training samples, the gini index contribution for soft electron tagger (softe) drops from **1.11% to 0.28%.**

➔ The light-jet rejection drops **3%-4%** if we removed softe from input variables.

JetFitterCombNN & BDT_bl (22Vars)

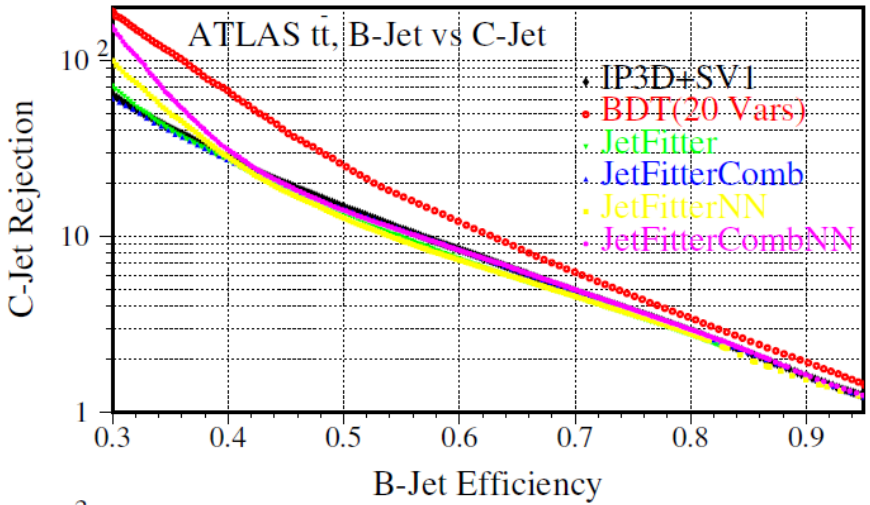


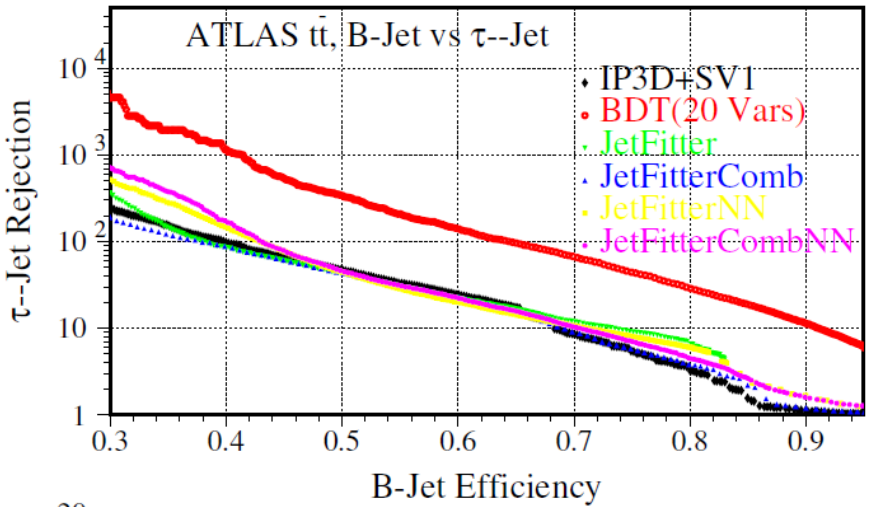


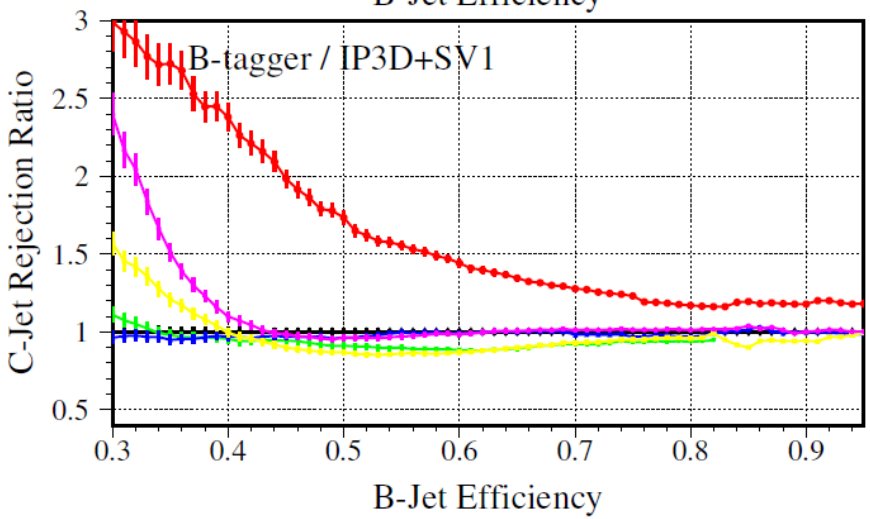
Light-jet rejection comparison
BDT combined with JetFitterCombNN

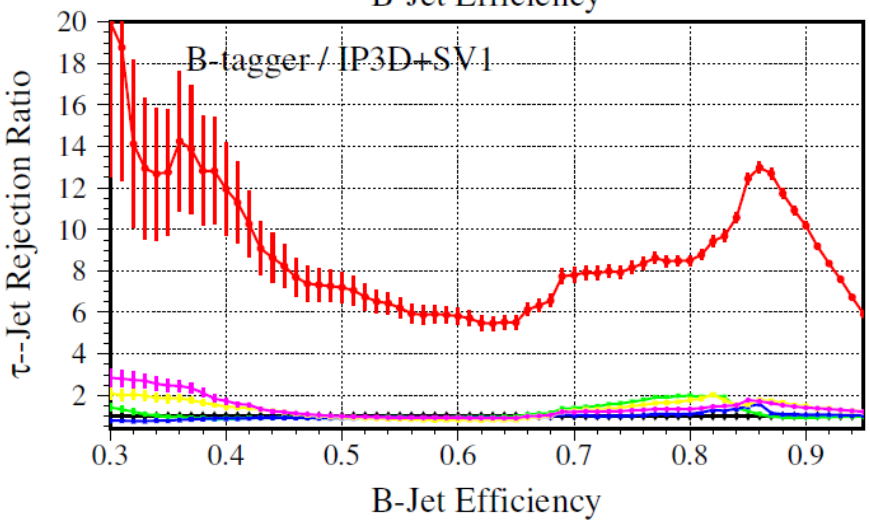
Test Sample $\sqrt{s} = 10$ TeV	b-jet Eff.	light-jet Rejection				
		IP3D+SV1	JetFitter CombNN	BDT_bl 22 Vars	BDT_bc 22 Vars	BDT_bt 22 Vars
$t\bar{t}$	70%	38.0 ± 0.4	44.7 ± 0.5	78.9 ± 1.1	38.3 ± 0.4	20.8 ± 0.2
$t\bar{t}$	60%	136.2 ± 2.4	170.5 ± 3.4	289.1 ± 7.4	104.0 ± 1.6	63.4 ± 0.8
$t\bar{t}$	50%	389.3 ± 11.6	601.9 ± 22.4	995.4 ± 47.5	286.4 ± 7.3	245.6 ± 5.8
WH(120 GeV)	70%	30.5 ± 0.2	33.9 ± 0.3	52.2 ± 0.5	30.2 ± 0.2	12.8 ± 0.1
WH(120 GeV)	60%	123.3 ± 1.9	151.5 ± 2.6	245.0 ± 5.3	98.9 ± 1.4	44.5 ± 0.4
WH(120 GeV)	50%	474.4 ± 14.3	666.5 ± 23.7	1081.2 ± 49.0	331.2 ± 8.3	194.7 ± 3.8
WH(400 GeV)	70%	44.7 ± 0.4	50.3 ± 0.4	81.6 ± 0.9	44.6 ± 0.4	22.5 ± 0.1
WH(400 GeV)	60%	142.6 ± 2.0	173.6 ± 2.7	277.8 ± 5.4	117.7 ± 1.5	64.4 ± 0.6
WH(400 GeV)	50%	426.6 ± 10.2	555.0 ± 15.2	923.6 ± 32.5	330.0 ± 7.0	263.8 ± 5.0

B-tag efficiency vs. C-Jet and τ-Jet Rejection and Comparisons









Dedicated training helps to further reject c-jet and τ-jet