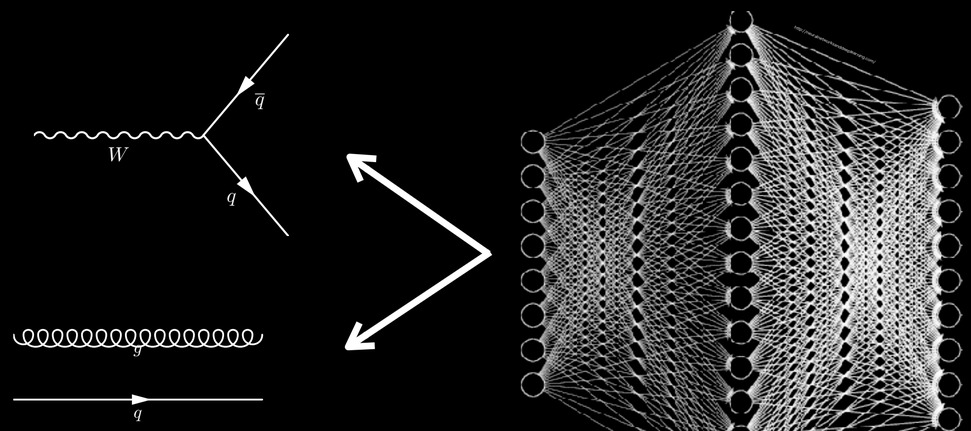
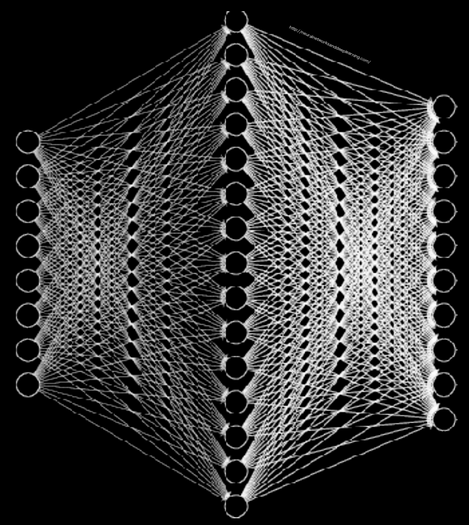
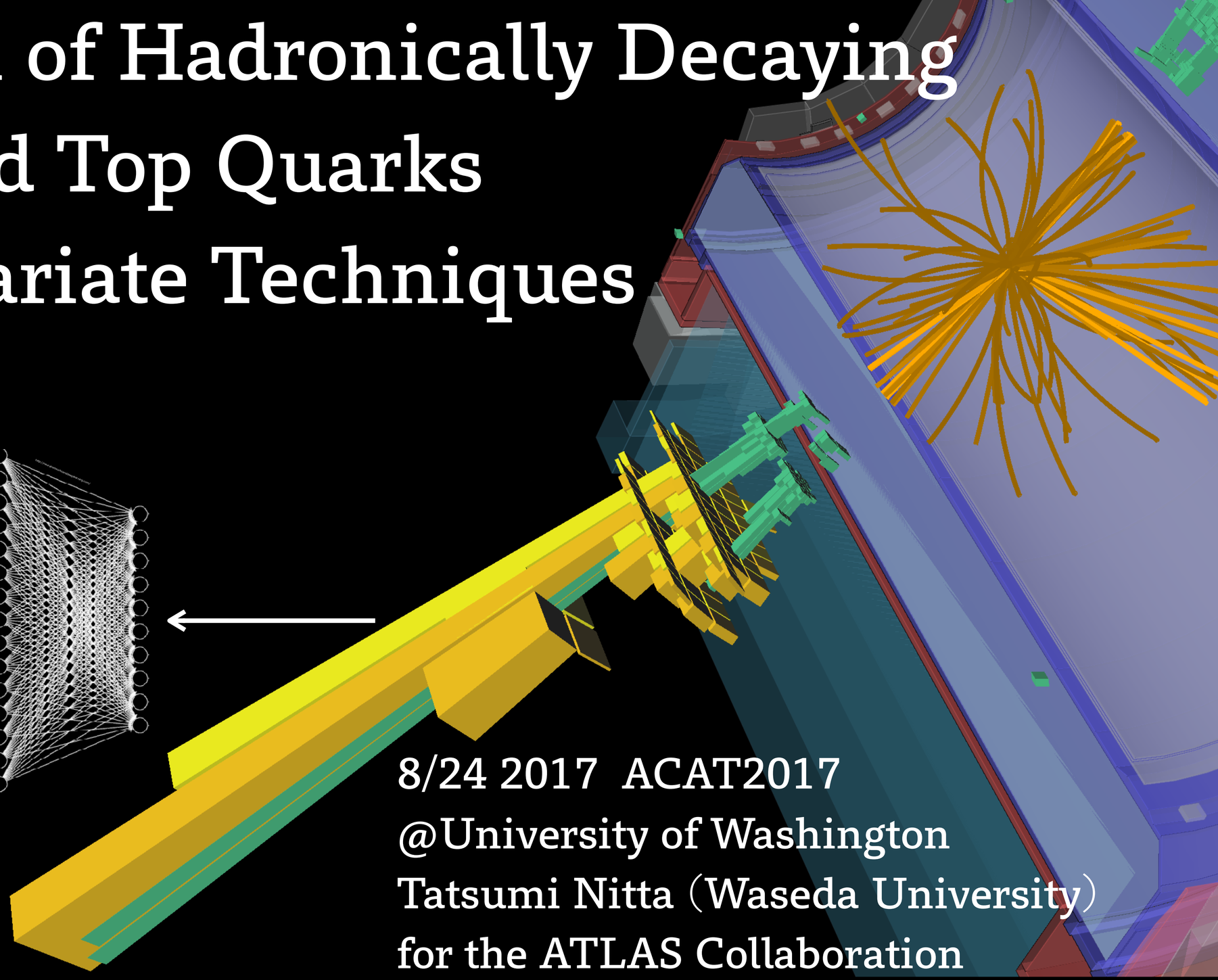
Identification of Hadronically Decaying W Bosons and Top Quarks using Multivariate Techniques at ATLAS







8/24 2017 ACAT2017
@University of Washington
Tatsumi Nitta（Waseda University）
for the ATLAS Collaboration

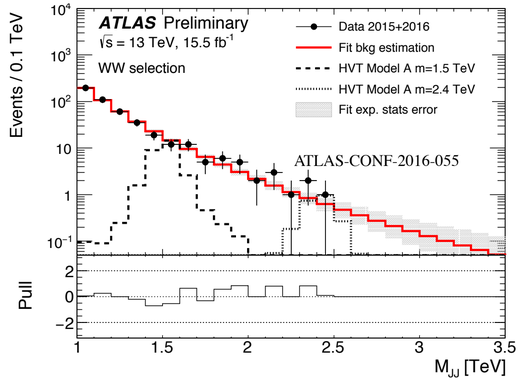
ATLAS EXPERIMENT

Introduction

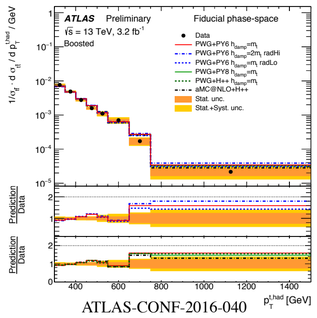
This talk mainly based on
ATLAS-CONF-2017-064
ATL-PHYS-PUB-2017-004

Rich physics in Boosted W/top

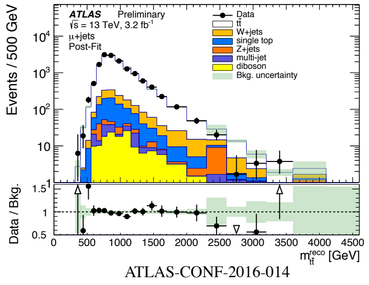
Diboson Resonance



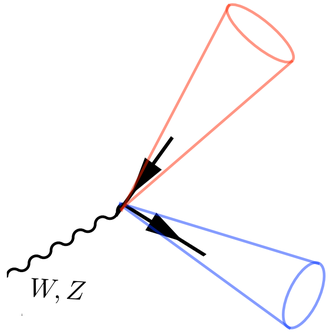
top pair production cross section measurement



top pair resonance

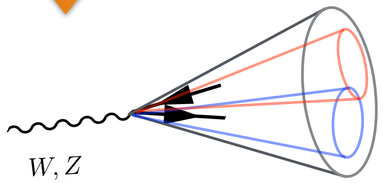


Boosted Jet



Boost

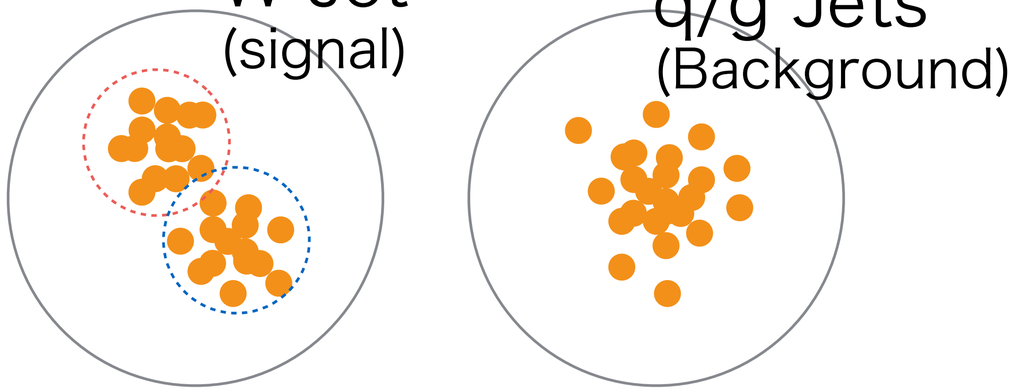
$$\Delta R \sim \frac{2M}{p_T}$$



Using Large-R Jet
to capture all constituents

Distinguish to q/g Jets

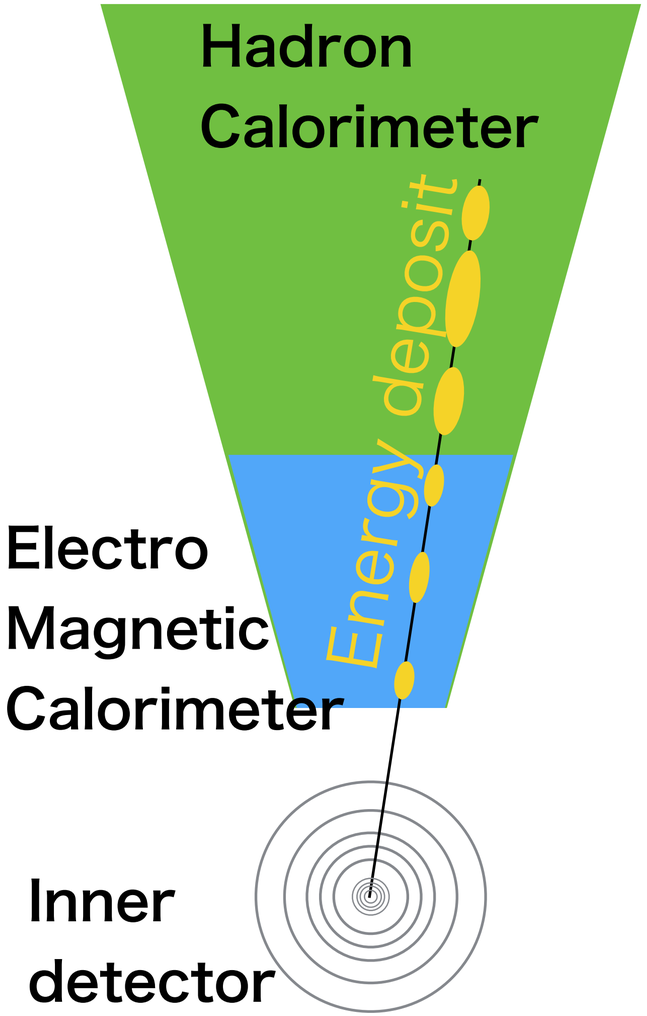
W/top jet become similar to
Quark/Gluon Jets via QCD.



Examine internal structure after jet clustering.

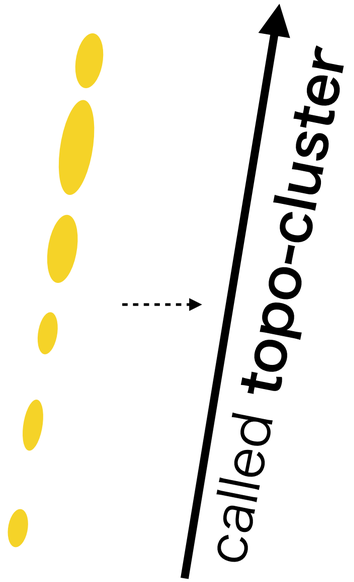
Jet Reconstruction

0. Energy deposit



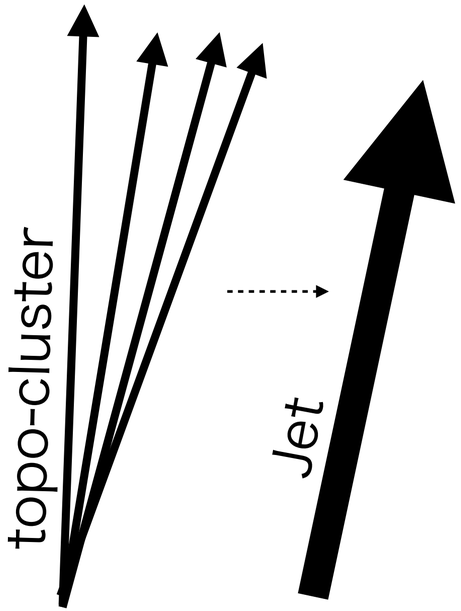
1. hadron building

= hadron
4-vector



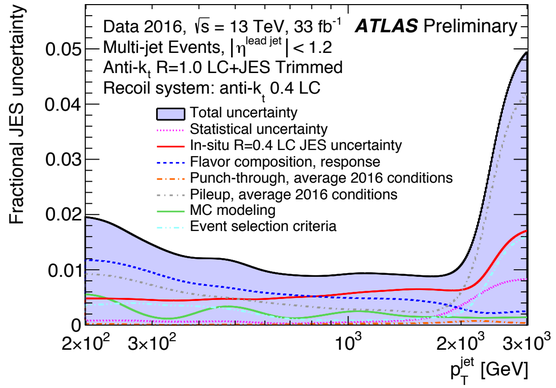
2. jet building

= Jet
4-vector



3. calibration

MC and data based calibration are applied.

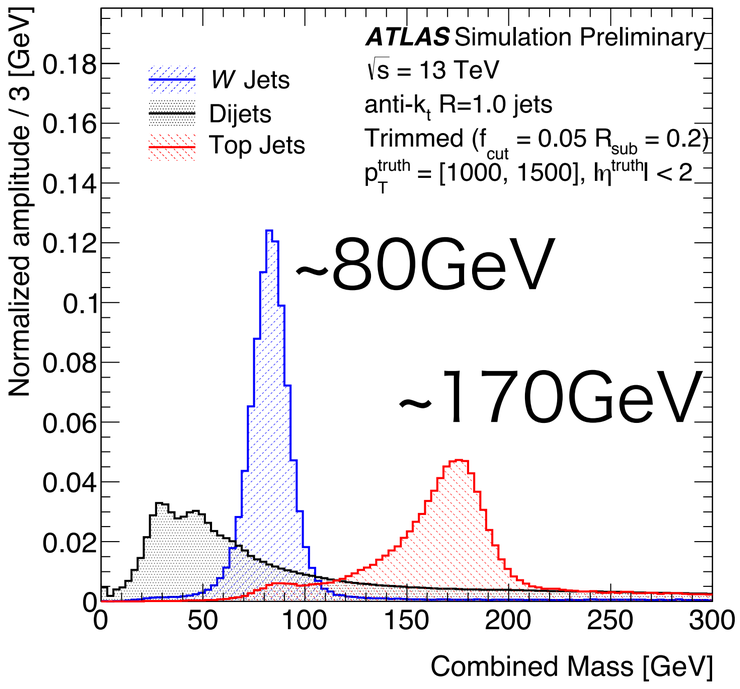


ATLAS-CONF-2017-063

Most of high level features are calculated by **topo-cluster** .

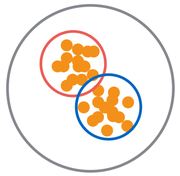
Current W/top tagger @ ATLAS

1. Jet Mass



ATLAS-CONF-2017-064

2. Number of Prong



←W/top has 2/3 prong

q/g has not clear structure →

Using “D2” for W

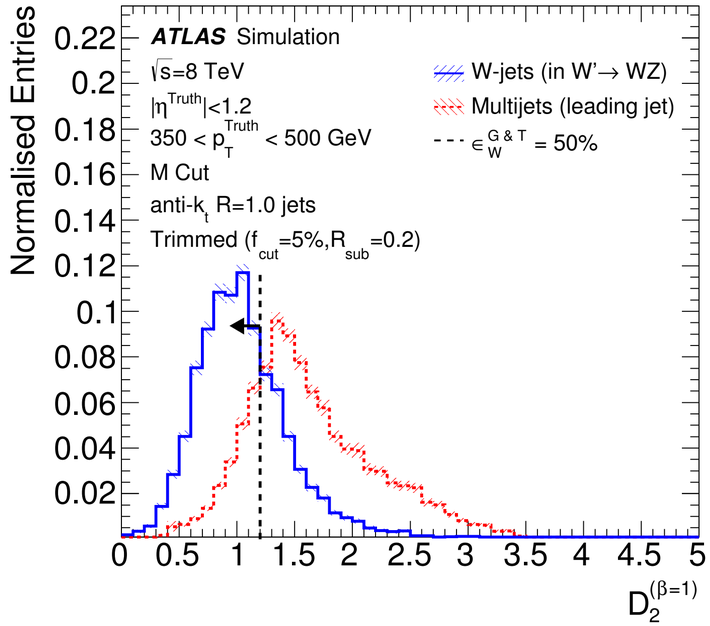
$$e_2^\beta = \frac{1}{p_{Tjet}^2} \sum_{i<j}^{N_{cl}} p_{Ti} p_{Tj} \Delta R_{ij}^\beta$$

$$e_3^\beta = \frac{1}{p_{Tjet}^3} \sum_{i<j<k}^{Ncl} p_{Ti} p_{Tj} p_{Tk} \Delta R_{ij}^\beta \Delta R_{jk}^\beta \Delta R_{ki}^\beta$$

$$D_2^\beta = \frac{e_3^\beta}{\left(e_2^\beta\right)^3}$$

similar variable

τ32 is used for top



CERN-PH-EP-2015-204

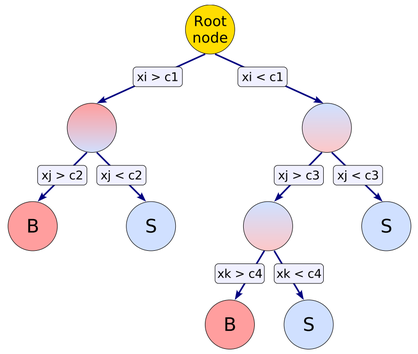
Current tagger is optimized with above 2 variables.

Machine Learning Tagger

There are many variables to distinguish 2/3 prong structure proposed from theoretical community.
All of them has good discriminate power.
-> Using machine learning techniques to wrap up them.

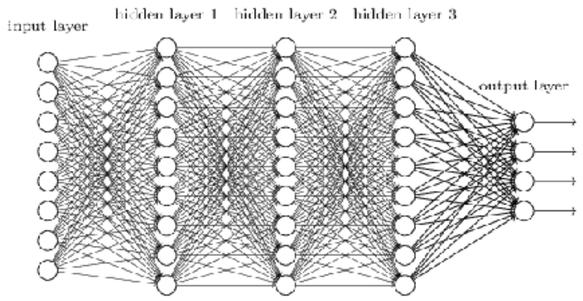
Try 2 machine learning techniques.

Boosted Decision Tree (BDT)



Binary structure classifier with Boosting techniques.
Implemented by TMVA

Deep Neural Network(DNN)



Picture

Multilayer Neural Network.
Implemented by keras

Training Samples

W Jet
W' → WZ (Pythia8)
pT=[200,2000]GeV

top Jet
Z' → tt (Pythia8)
pT=[350,2000]GeV

q/g Jet
dijet (Pythia8)
pT=[200,2000]GeV

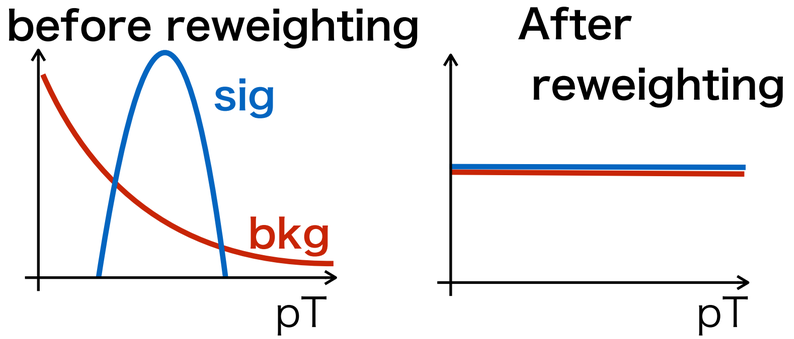
Number of events

Purpose	Training (Top quark)		Training (W boson)		Testing		
Sample	Signal	Background	Signal	Background	Top Signal	W Signal	Background
Number of jets	10^6	10^6	7×10^5	7×10^5	4×10^5	3×10^5	1×10^6

Remarks :

Some variables depends on p_T

-> Signal & Background is reweighted to flat.



Input variable selection

Tagging Type	Observable Set
15 variables W-boson tagging	m^{comb}, p_{T} e_3, C_2, D_2 τ_1, τ_2, τ_{21} R_2^{FW}, $\mathcal{P}$, a_3, A Z_{CUT}, $\sqrt{d_{12}}$ KtDR
13 variables Top-Quark Tagging	m^{comb}, p_{T} e_3, C_2, D_2 τ_1, τ_2, τ_3,τ_{32}, τ_{21} $\sqrt{d_{12}}$, $\sqrt{d_{23}}$, Q_W

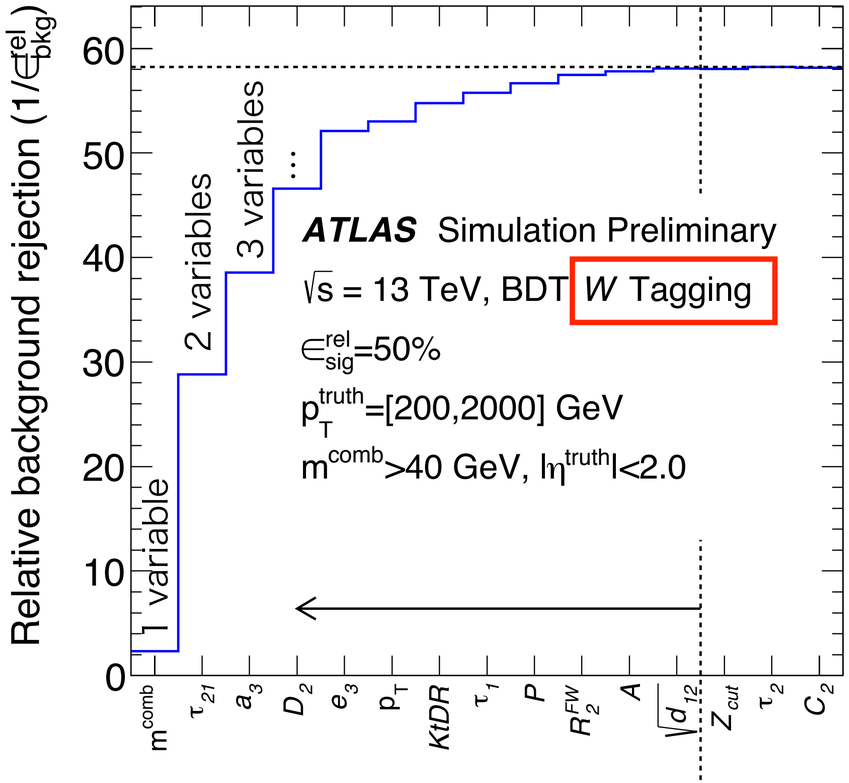
Inputs : High level features

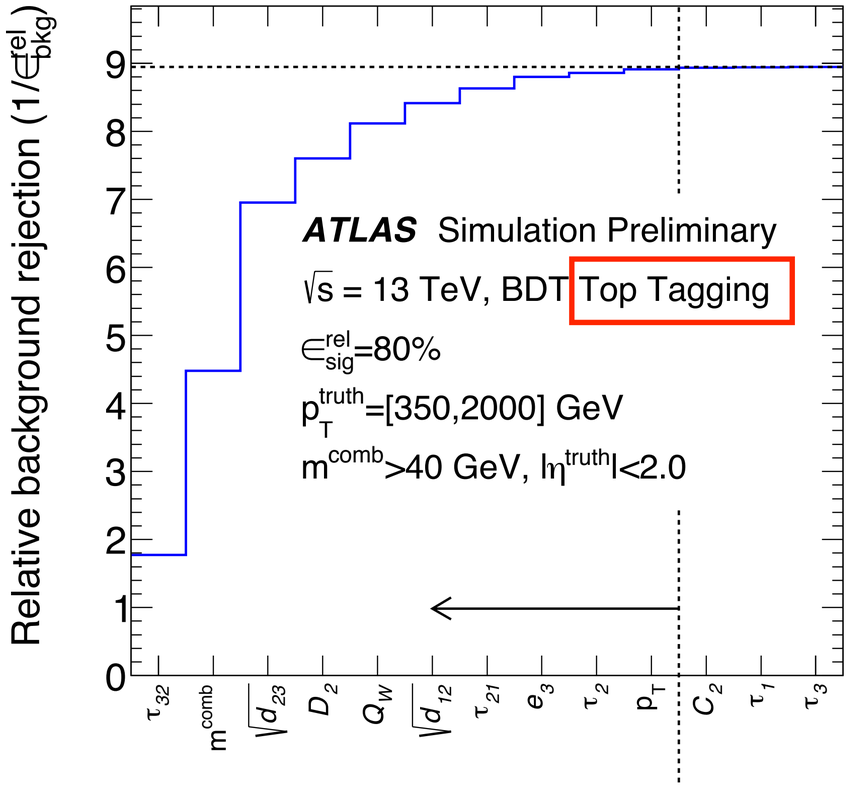
Inputting all variables should have best discriminant power.
But can lead to larger systematics uncertainties.
→ Select strongly contributing variables.

Input variables for DNN and BDT are chosen independently to get highest performance for each.

BDT input variables

Iteratively adding variable that have best background rejection in each number of variables.





Both W and top tagger can be rejected 3 variables without loss of performance.

DNN input variables

Similar variables are wrapped as group, test each of them.

W tagging

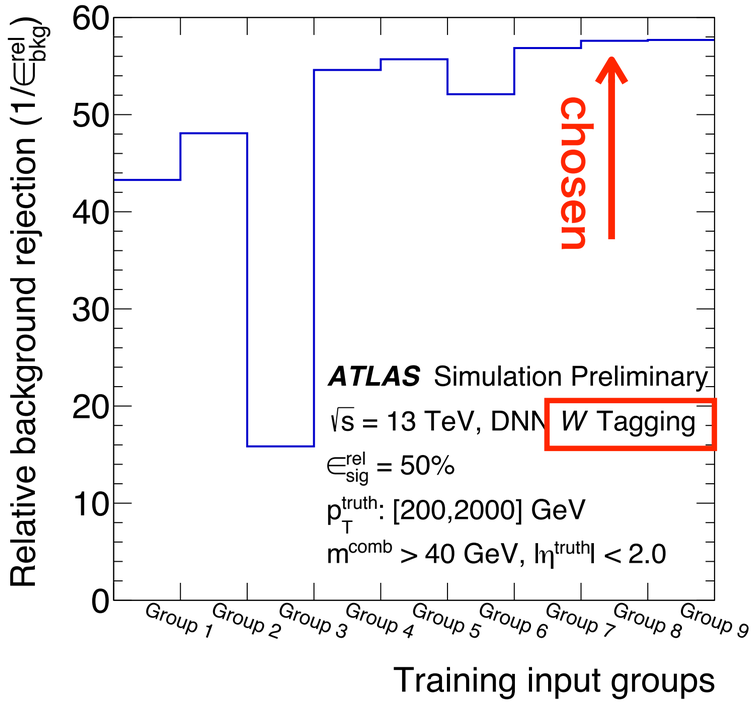
Group 1	τ_1, τ_2, e_3, m^{comb}, p_{T}
Group 2	τ_1, τ_2, e_3, m^{comb}, p_{T}, $\sqrt{d_{12}}$, KtDR
Group 3	τ_{21}, C_2, D_2, R_2^{FW}, $\mathcal{P}$, a_3, A, Z_{CUT}
Group 4	τ_{21}, C_2, D_2, R_2^{FW}, $\mathcal{P}$, a_3, A, Z_{CUT}, m^{comb}
Group 5	τ_{21}, C_2, D_2, R_2^{FW}, $\mathcal{P}$, a_3, A, Z_{CUT}, m^{comb}, p_{T}
Group 6	τ_1, τ_2, e_3, m^{comb}, p_{T}, R_2^{FW}, $\sqrt{d_{12}}$, KtDR, a_3, A
Group 7	τ_{21}, C_2, D_2, R_2^{FW}, $\mathcal{P}$, a_3, A, Z_{CUT}, m^{comb}, $\sqrt{d_{12}}$, KtDR
Group 8	τ_{21}, C_2, D_2, R_2^{FW}, $\mathcal{P}$, a_3, A, Z_{CUT}, m^{comb}, p_{T}, $\sqrt{d_{12}}$, KtDR
Group 9	τ_1, τ_2, τ_{21}, $\sqrt{d_{12}}$, C_2, D_2, e_3, m^{comb}, p_{T}, R_2^{FW}, $\mathcal{P}$, a_3, A, Z_{CUT}, KtDR

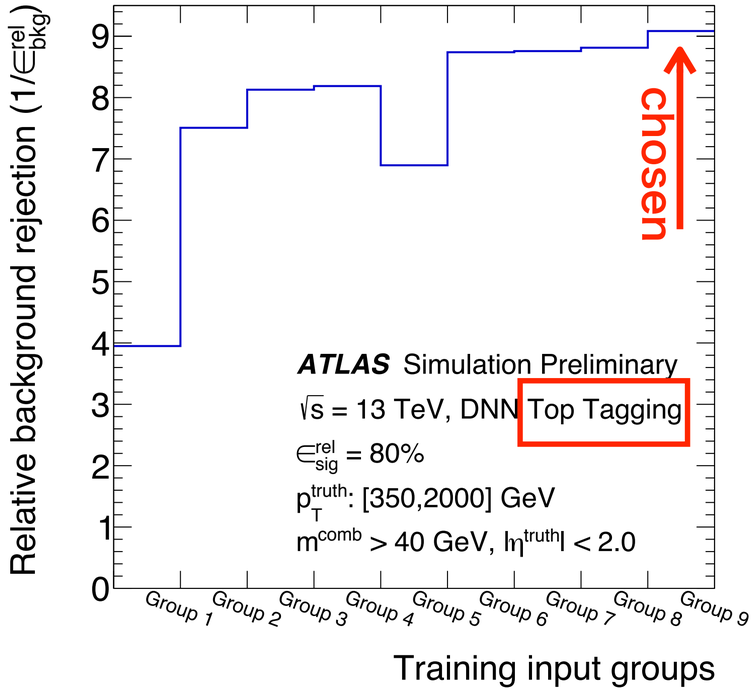
3 variables are discarded.

top tagging

Group 1	C_2, D_2, τ_{21}, τ_{32},
Group 2	C_2, D_2, τ_{21}, τ_{32}, m^{comb}
Group 3	C_2, D_2, τ_{21}, τ_{32}, m^{comb}, p_{T}
Group 4	τ_1, τ_2, τ_3, e_3, m^{comb}, p_{T}
Group 5	C_2, D_2, τ_{21}, τ_{32}, $\sqrt{d_{12}}$, $\sqrt{d_{23}}$, Q_W
Group 6	C_2, D_2, τ_{21}, τ_{32}, $\sqrt{d_{12}}$, $\sqrt{d_{23}}$, Q_W, m^{comb}
Group 7	τ_1, τ_2, τ_3, e_3, m^{comb}, p_{T}, $\sqrt{d_{12}}$, $\sqrt{d_{23}}$, Q_W
Group 8	C_2, D_2, τ_{21}, τ_{32}, $\sqrt{d_{12}}$, $\sqrt{d_{23}}$, Q_W, m^{comb}, p_{T}
Group 9	τ_1, τ_2, τ_3, τ_{21}, τ_{32}, $\sqrt{d_{12}}$, $\sqrt{d_{23}}$, Q_W, C_2, D_2, e_3, m^{comb}, p_{T}

DNN exploits all variables.





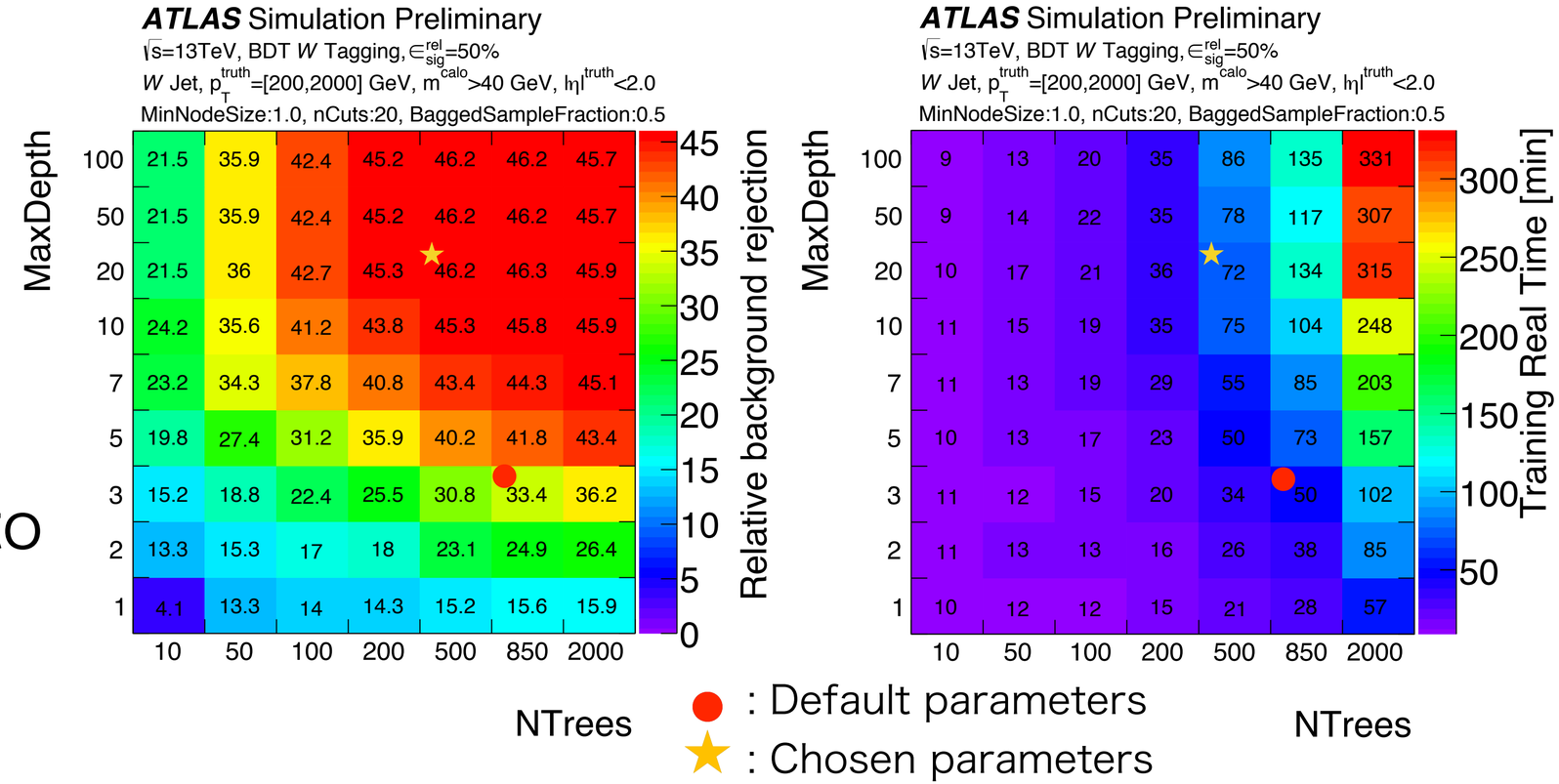
BDT Optimization

Hyper-parameters of machine learning drastically change result.

→ Optimize these with parameter scan.

Setting Name	Value Tested
NTrees	[10, 50, 100, 200, 500, 850, 2000]
MaxDepth	[1, 2, 3, 5, 7, 10, 20, 50, 100]
MiniNodeSize	[0.5, 1.0, 2.5, 5.0, 10, 20]
nCuts	[5, 10, 20, 50, 100, 500]
Bagged Fraction	[0.1, 0.3, 0.5, 0.7, 0.9]
Shrinkage	[0.05, 0.1, 0.3, 0.5, 0.7, 0.9]

Most of parameter related to how to make or trim trees.



Big & **many** trees cause of unrealistic training time.

Chose parameter set that have <u>most performant</u> and <u>minimum training time</u>.

DNN Optimization

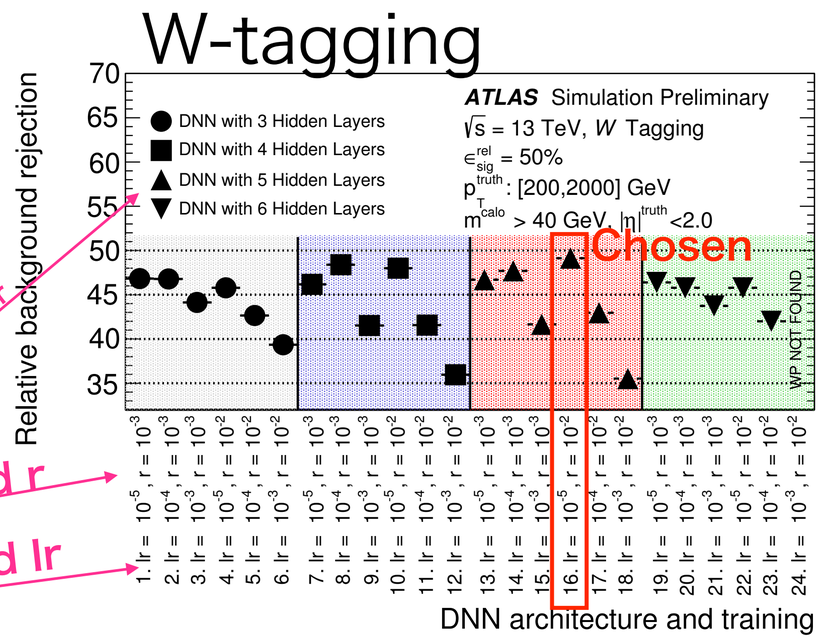
DNN is done grid search for several parameters.

Layer type
Number of hidden layers
Activation function
Learning rate
L1 Regularizer
NN weight initialization
Architecture

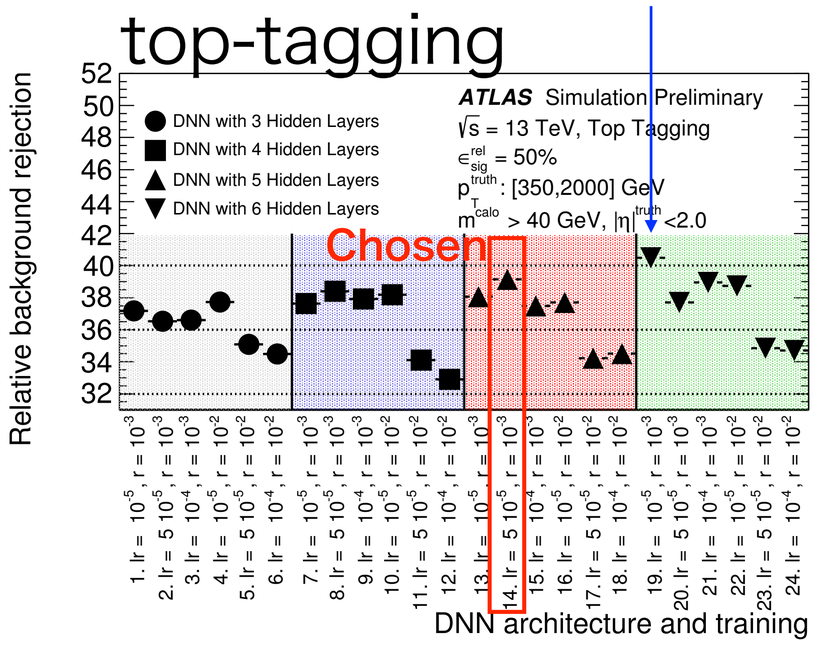
color&marker

named r

named lr



Very spiky output



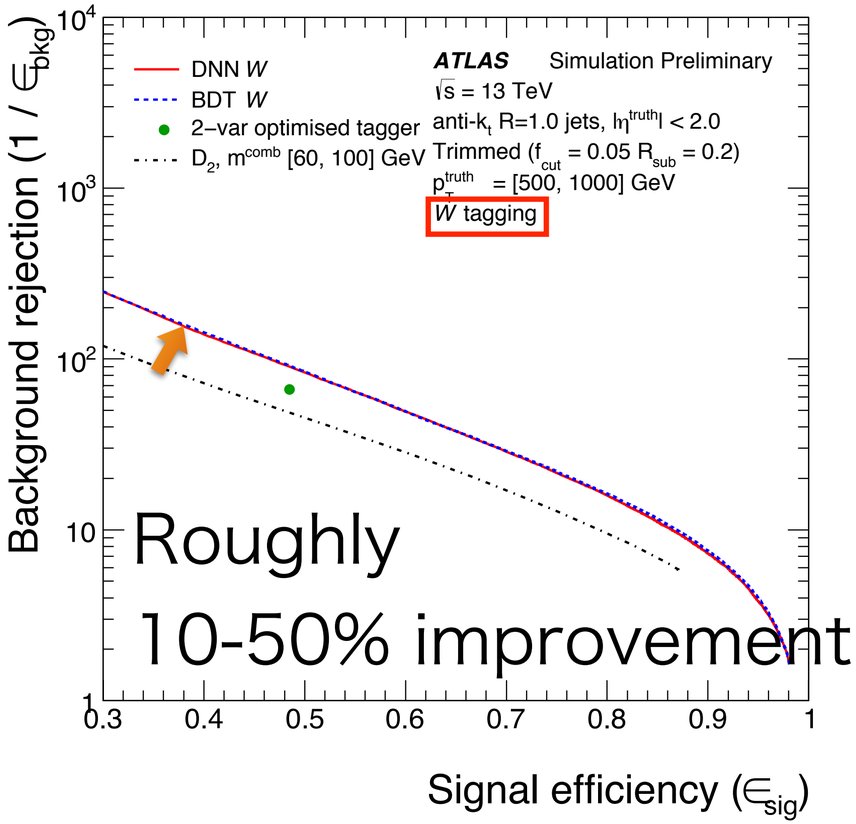
- Not depending on #layer

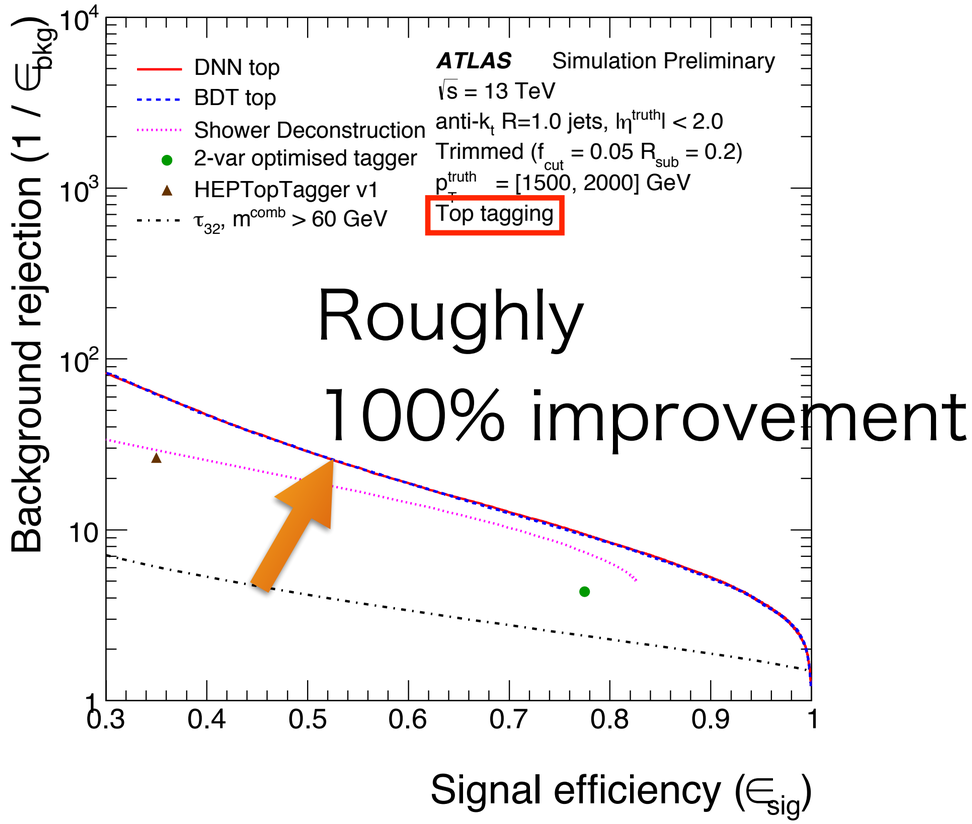
  → Not warranted to use **D**NN in this problem.

Results with MC simulation

DNN/BDT results compared to current optimization tagger.

(Statistically independent to training sample)

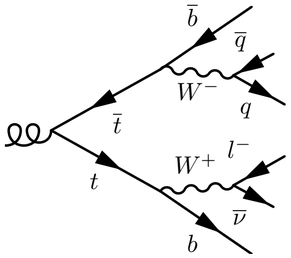




DNN and BDT achieve huge improvement, and exactly same performance.

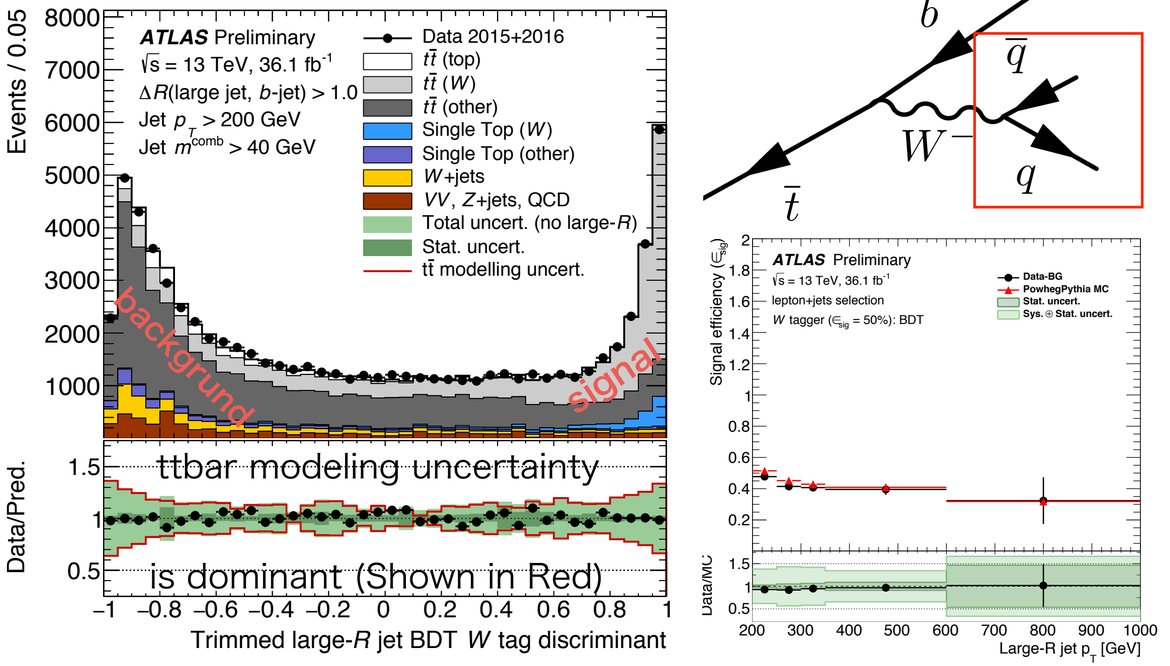
Signal Modeling check with data

W/top is collected top pair production with 1 lepton.



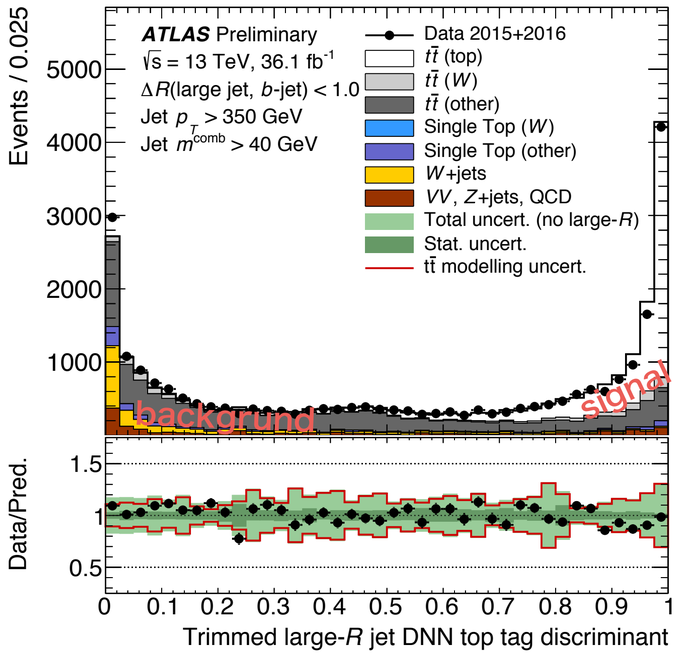
W-tagging

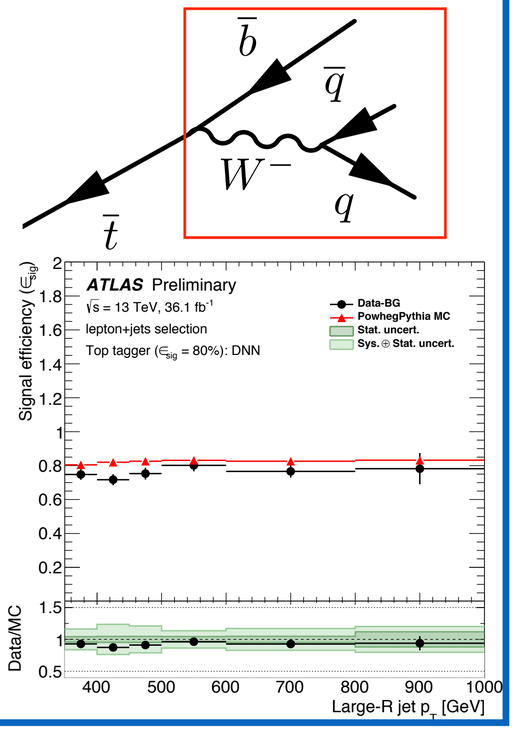
Require B-jet is outside of candidate jet



top-tagging

Require B-jet is inside of candidate jet

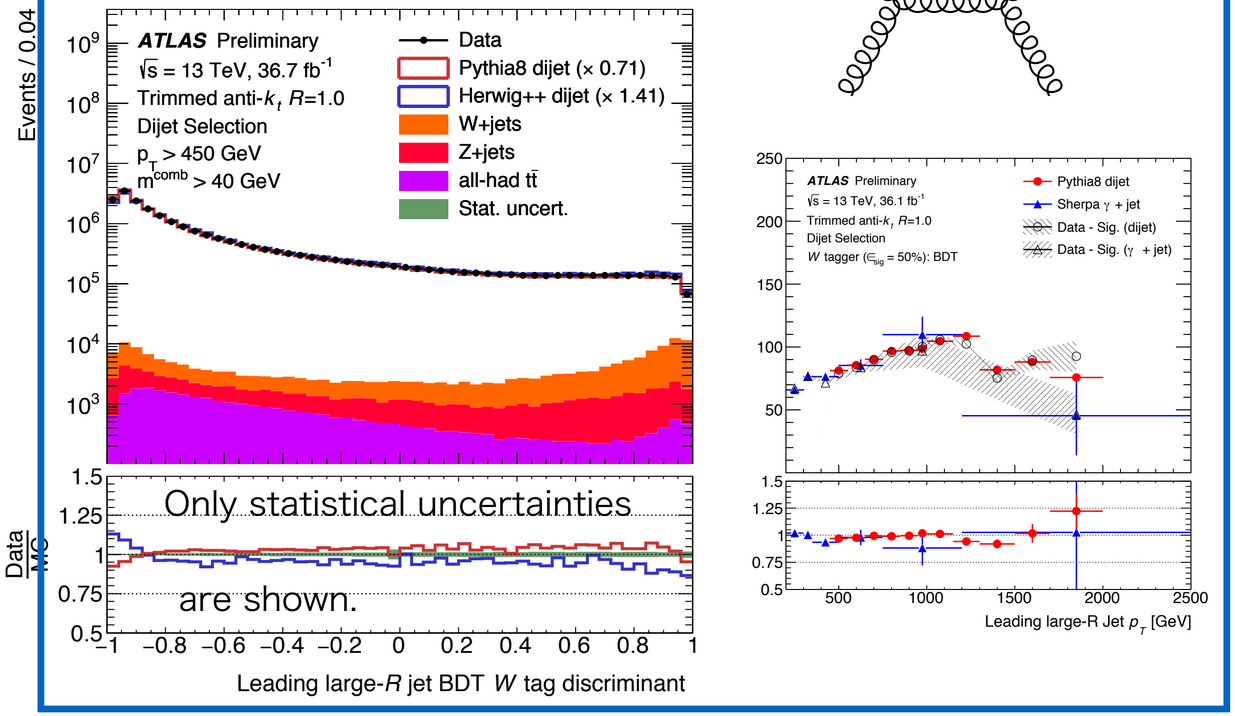




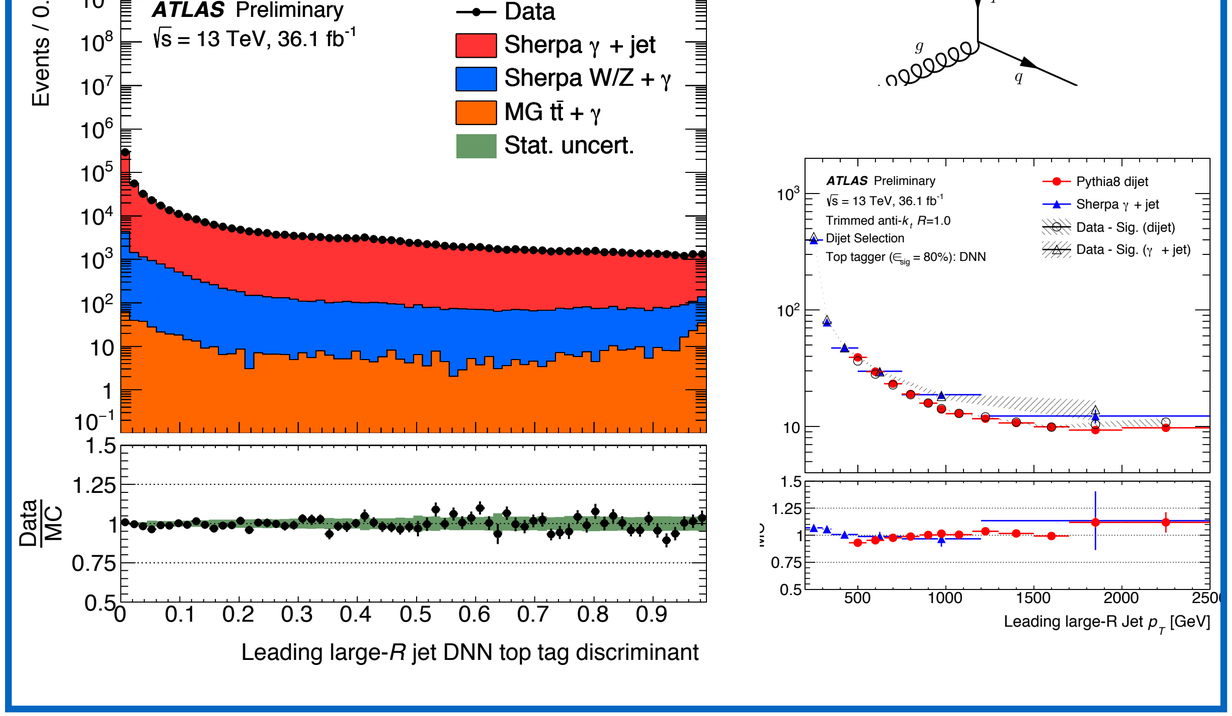
Both W/top tagging cases, modeling are very good.

Background modeling check

W-tagging



top-tagging

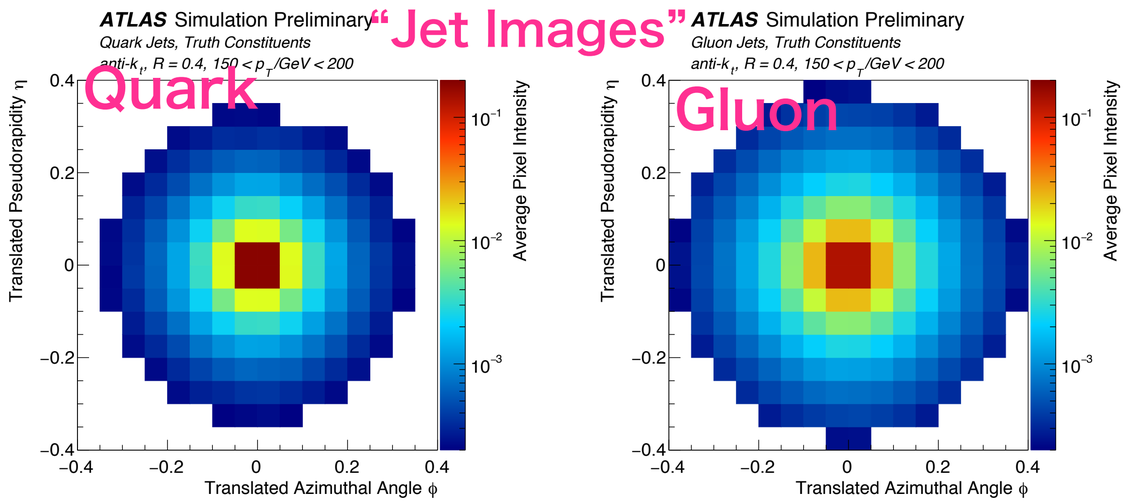


Both W/top tagging cases, modeling are very good.

More advanced techniques for q/g separation in ATLAS

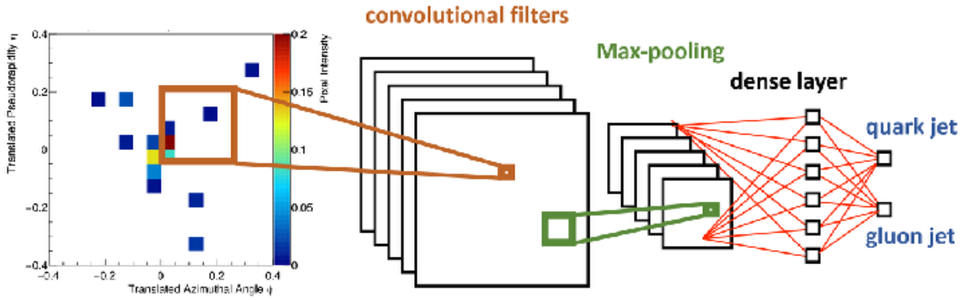
ATL-PHYS-PUB-2017-017

Make 2D histogram of energy deposit

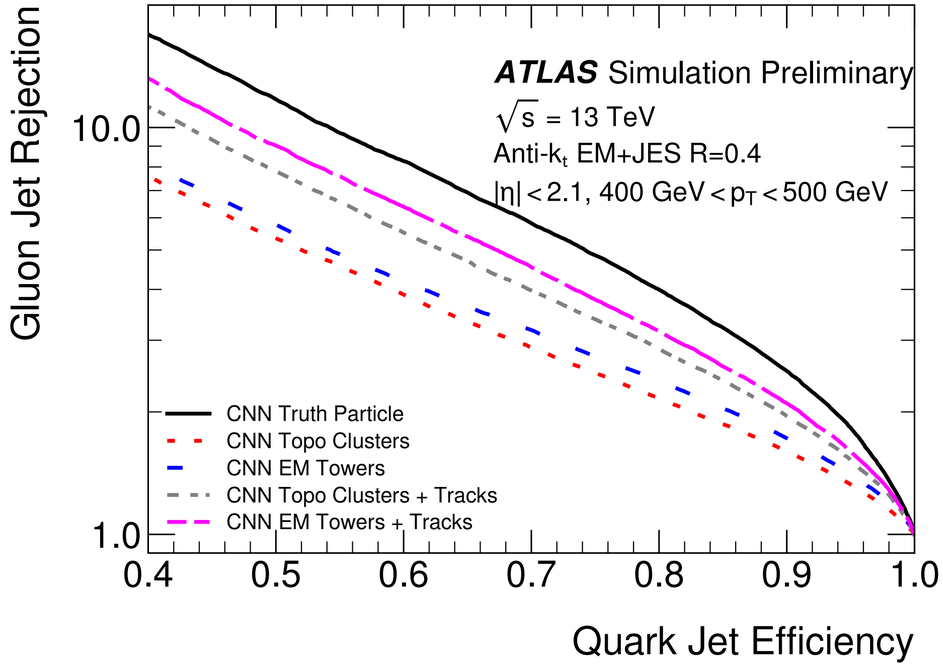


"Jet Images"

Input jet images to Convolutional NN.



These have good performance!



Adding to study simple performances, MC modeling & pile-up dependence are studied. That's promising!

Summary

Constructing **W/top jet tagger** by **Machine Learning** techniques.

[Results]

- **Performance fairly improve** than non-Machine Learning tagger.
- DNN and BDT has exactly same performance.
  → DNN does not overwhelm BDT in this problem.
- **Modeling is confirmed by Data.**
  → Data is under control of our hands!
- **Jet Image** based q/g separation has already studied with MC.
  → It is interesting study that W/top tagging with Jet Image.

Thanks for listening!

Backup

Summary of JSS variables

Variables calculating Center of mass System

ThrustAxis
1112.2567

$$T = \max_{\vec{n}_T}\left(\frac{\sum_i |\vec{p}_i \cdot \vec{n}_T|}{\sum_i |\vec{p}_i|}\right)$$

$$T_{\mathrm{Ma}} = \max_{\vec{n}_{\mathrm{Ma}} \perp \vec{n}_T}\left(\frac{\sum_i |\vec{p}_i \cdot \vec{n}_{\mathrm{Ma}}|}{\sum_i |\vec{p}_i|}\right)$$

$$T_{\mathrm{Mi}} = \frac{\sum_i |\vec{p}_i \cdot \vec{n}_{\mathrm{Mi}}|}{\sum_i |\vec{p}_i|} \quad \vec{n}_{\mathrm{Mi}} = \vec{n}_T \times \vec{n}_{\mathrm{Ma}}$$

SphericityTensor
1112.2567

$$S^{\alpha,\beta} = \frac{\sum_i p_i^\alpha p_i^\beta}{\sum_i |\vec{p}_i|^2}$$

$$S = \frac{3}{2}(\lambda_2 + \lambda_3) \qquad A = \frac{3\lambda_3}{2}$$

FoxWolfram Moments
1112.2567

$$H_l = \sum_{i,j} \frac{|\vec{p}_i||\vec{p}_j|}{E^2} P_l(\cos\theta_{ij})$$

$$\mathrm{FoxWolfram2} = \mathrm{H}_2/\mathrm{H}_0$$

Variables calculating Laboratory System

Energy Correlation Function
1510.05821

$$E_{CF0}(\beta) = 1,$$

$$E_{CF1}(\beta) = \sum_{i\in J} p_{\mathrm{T}_i},$$

$$E_{CF2}(\beta) = \sum_{i<j\in J} p_{\mathrm{T}_i} p_{\mathrm{T}_j} \left(\Delta R_{ij}\right)^\beta,$$

$$E_{CF3}(\beta) = \sum_{i<j<k\in J} p_{\mathrm{T}_i} p_{\mathrm{T}_j} p_{\mathrm{T}_k} \left(\Delta R_{ij}\Delta R_{ik}\Delta R_{jk}\right)^\beta$$

$$e_2^{(\beta)} = \frac{E_{CF2}(\beta)}{E_{CF1}(\beta)^2}, \qquad C_2^{(\beta)} = \frac{e_3^{(\beta)}}{(e_2^{(\beta)})^2},$$

$$e_3^{(\beta)} = \frac{E_{CF3}(\beta)}{E_{CF1}(\beta)^3}. \qquad D_2^{(\beta)} = \frac{e_3^{(\beta)}}{(e_2^{(\beta)})^3}.$$

τ
1510.05821

$$\tau_0(\beta) = \sum_{i\in J} p_{\mathrm{T}_i} \Delta R^\beta,$$

※aN : subjet or wta
※wta : winner takes all most energetic cluster

$$\tau_1(\beta) = \frac{1}{\tau_0(\beta)} \sum_{i\in J} p_{\mathrm{T}_i} \Delta R^\beta_{a_1,i},$$

$$\tau_2(\beta) = \frac{1}{\tau_0(\beta)} \sum_{i\in J} p_{\mathrm{T}_i} \min(\Delta R^\beta_{a_1,i}, \Delta R^\beta_{a_2,i}),$$

$$\tau_{21} = \frac{\tau_2}{\tau_1} \qquad \tau_{32} = \frac{\tau_3}{\tau_2}$$

$$\tau_{21}^{wta} = \frac{\tau_2^{wta}}{\tau_1^{wta}} \qquad \tau_{32}^{wta} = \frac{\tau_3^{wta}}{\tau_2^{wta}}$$

PlanarFlow
1206.5369

$$I_E^{kl} = \frac{1}{M}\sum_i \frac{1}{E_i} p_{i,k} p_{i,l} \qquad P = 4 \times \frac{\det(I_E)}{\mathrm{Tr}(I_E)^2}$$

Angularity
1206.5369

$$\tau_a = \frac{1}{M}\sum_i E_i \sin^a\theta_i [1-\cos\theta_i]^{1-a}$$

※a : parameter
※θ : ∠(jet axis, i th constituents)

※Variables below are calculated from subjet(kt0.2)

Splitting Scale
1302.1415

$$d_{ij} = \min(p_{\mathrm{T}i}^2, p_{\mathrm{T}j}^2)\frac{\Delta R_{ij}^2}{R^2}$$

ZCut12
0806.0023

$$d_{\mathrm{cut}} = \min(p_{TA}^2, p_{TB}^2)\Delta R_{AB}^2$$

$$z_{\mathrm{cut}} \equiv \frac{d_{\mathrm{cut}}}{d_{\mathrm{cut}} + Q_M^2}$$

Qw
0806.0023

$$Q_W = \min(m_{12}, m_{23}, m_{31})$$

Dip12
1102.1012

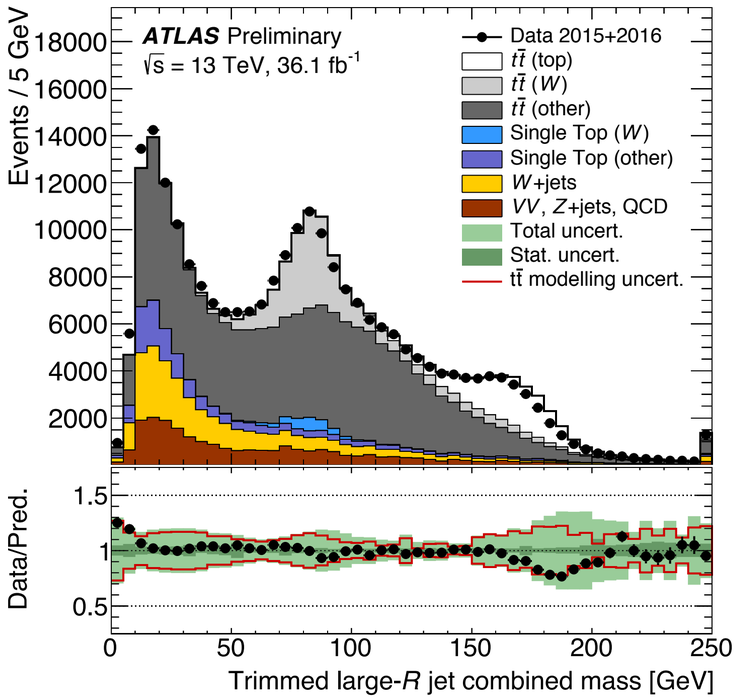
$$\mathcal{D} \equiv \frac{1}{R_{12}^2}\sum_{i\in J} \frac{p_{Ti}}{p_{T_J}} R_i^2$$

※i = 1,2 th jet
※R = ΔR

Signal Modeling check with data

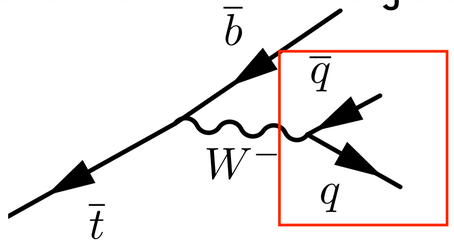
Using top pair production.

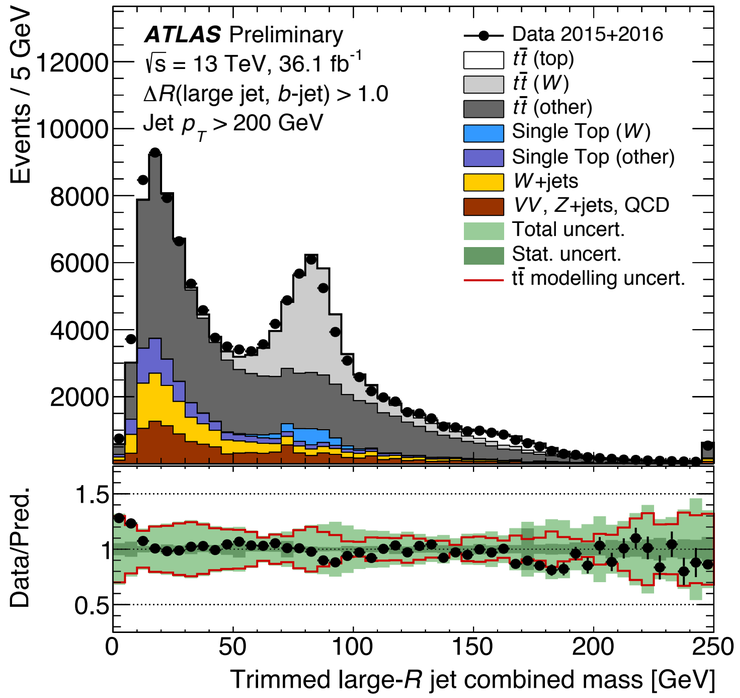
Particle-to-recojet matching
$\Delta R(x, \text{jet}) < 0.75$



W-tagging

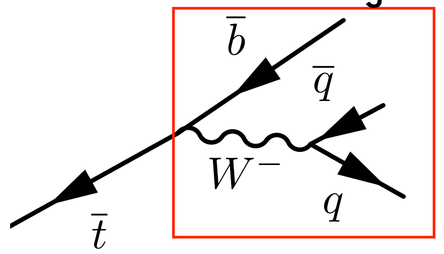
Require B-jet is
outside of candidate jet

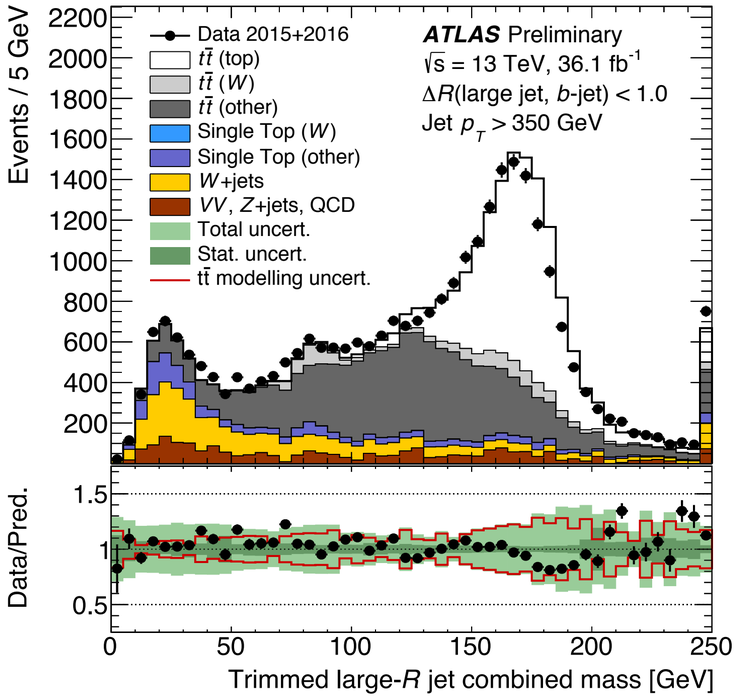




top-tagging

Require B-jet is
inside of candidate jet





Chosen parameters

	W-Boson Tagging Chosen	Top-Quark Tagging Chosen	Reference
Layer type	Dense	Dense	[24]
Number of hidden layers	5	5	[24]
Activation function	rectified linear unit (relu)	rectified linear unit (relu)	[41]
Learning rate	10^{-5}	5×10^{-5}	[43]
L1 Regularizer	10^{-2}	10^{-3}	[41]
NN weight initialization	Glorot uniform	Glorot uniform	[44]
Batch size	200	200	[41]
Batch normalization	Yes	Yes	[45]
Training groups	Group 5	Group 6	-
Architecture	18, 25, 22, 19, 14, 7, 1	13, 18, 16, 14, 10, 5, 1	-

Table 8: Chosen DNN parameters and architecture for W-boson and top-quark tagging.

Setting Name	Description	Chosen Value
BoostType	Type of boosting technique	GradientBoost
NTrees	Number of trees in the forest	500
MaxDepth	Max depth of the decision tree allowed	20
MinimumNodeSize	Minimum fraction of training events required in a leaf node	1.0%
Shrinkage	Learning rate for GradientBoost algorithm	0.5
UseBaggedBoost	Use only a random (bagged) subsample of all events for growing the trees in each iteration	True
BaggedSampleFraction	Relative size of bagged event sample to original size of the data sample	0.5
SeparationType	Separation criterion for node splitting	GiniIndex
nCuts	Number of grid points in variable range used in finding optimal cut in node splitting	500