

IDENTIFICATION OF HADRONICALLY DECAYING W  
BOSONS AND TOP QUARKS USING HIGH-LEVEL FEATURES  
AS INPUT TO BOOSTED DECISION TREES AND DEEP NEURAL  
NETWORKS IN ATLAS AT  $\sqrt{s} = 13\text{ TeV}$   
ATL-PHYS-PUB-2017-004

Ece Akilli  
For the ATLAS Collaboration

Université de Genève

IML Workshop  
22.03.2017

# OUTLINE

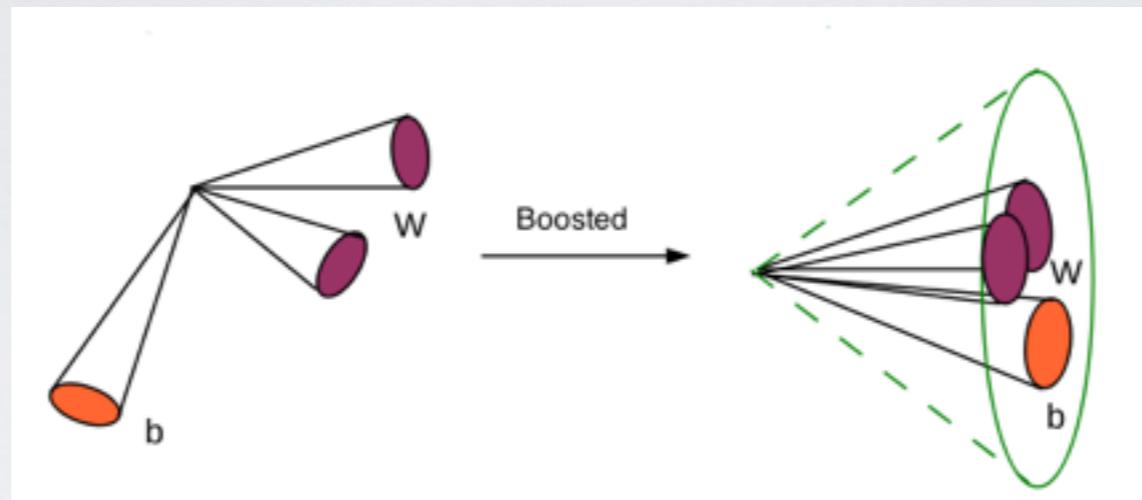
1. W Boson and Top Quark Tagging in ATLAS

2. Application of Boosted Decision Trees and Deep Neural Networks to W Boson and Top Quark Tagging Using High-Level Features

- Optimization & Analysis
- Results

3. Conclusions

## Top Quark Decay

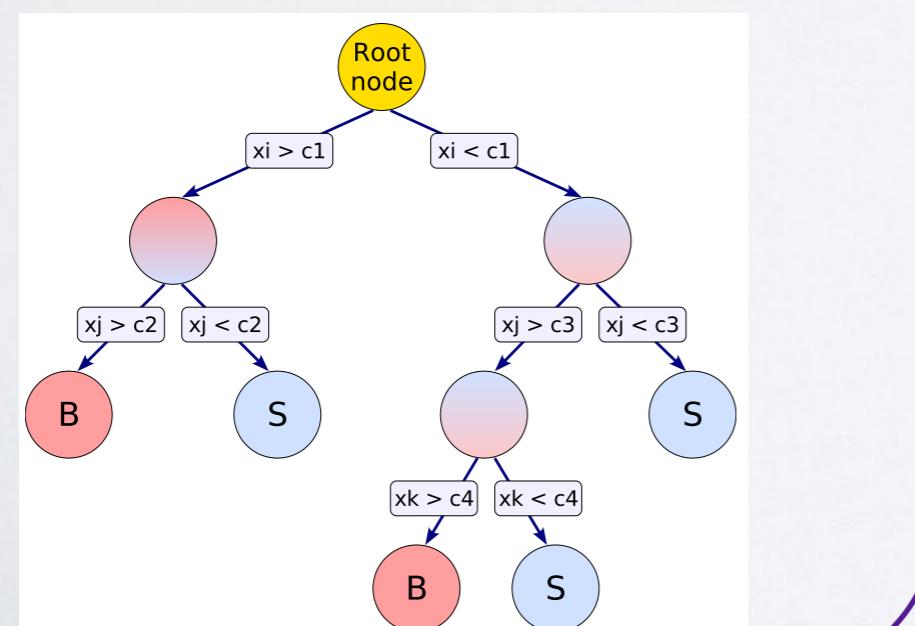


- W bosons and top quarks have short lifetime
- Decay products of high-momentum (boosted) hadronically decaying W bosons and top quarks are collimated
- Conventional object identification techniques are not successful
- Construct large-R jets
- Use substructure information to identify the W boson and top quark within an enormous QCD background

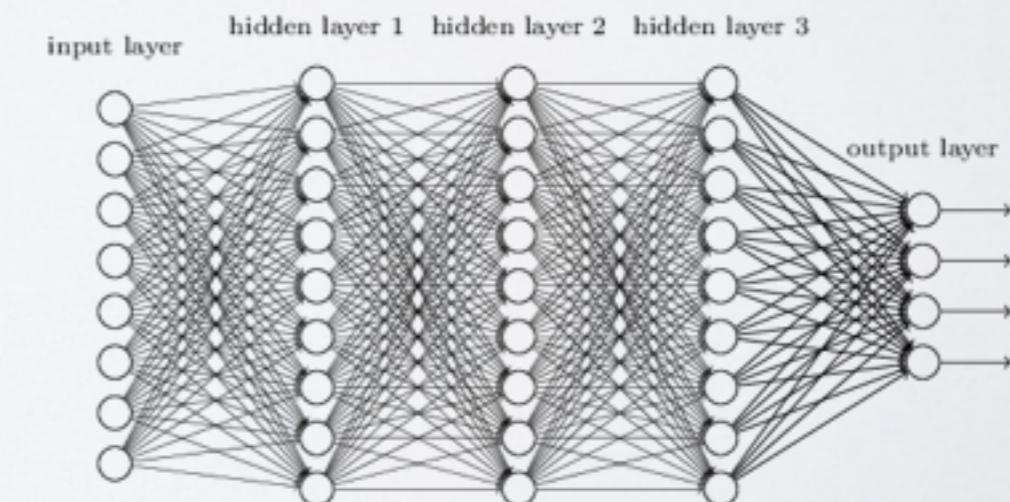
# APPLICATION OF BDTS AND DNNs TO W AND TOP TAGGING USING HIGH-LEVEL FEATURES

- Numerous substructure variables are available and are used by ATLAS
- Construct a classifier by using available substructure variables
- Feed the ML algorithms with jet substructure variables (high-level features). This was studied by CMS in Run I
- Study the performance of W and top tagging with two Machine Learning (ML) techniques in parallel

## I. Boosted Decision Trees (BDT) using TMVA

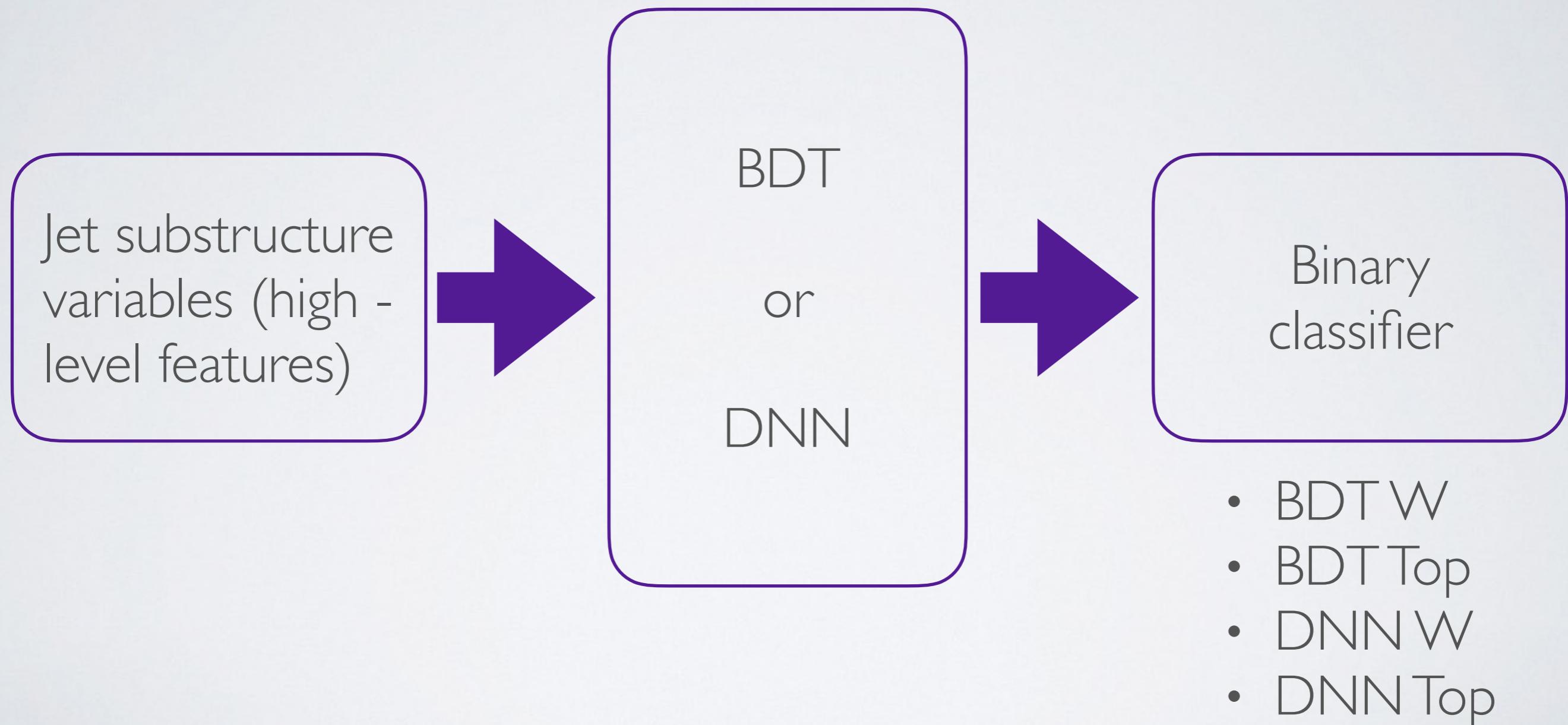


## 2 . Deep Neural Networks (DNN) using Keras with Theano backend



# APPLICATION OF BDTS AND DNNs TO W AND TOP TAGGING USING HIGH-LEVEL FEATURES

Training classifiers to discriminate between signal ( $W$  or top) and background (QCD), results in 4 binary classifiers



# APPLICATION OF BDTS AND DNNs TO W AND TOP TAGGING USING HIGH-LEVEL FEATURES

- Optimize set of inputs for BDT, DNN
- Optimize BDT, DNN architectures and training hyper-parameters
- Compare 3 tagging techniques for W and top tagging separately
  - Fixed mass cut & BDT
  - Fixed mass cut & DNN
  - Fixed mass cut & single substructure variable (W:  $D_2$ , top:  $\tau_{32}$ )

# OPTIMIZATION & ANALYSIS

# SAMPLES

## Training & Testing Samples

- Split signal and background (QCD) samples as: 70% training, 30% testing
- Use equal number of signal and background jets for training
- Train in 1  $p_T$  bin due to limited statistics

**Training Event Weights:** Signal and background samples are weighted to flat truth  $p_T$  distribution

**Testing Event Weights:** Signal samples (separately for Ws and tops) are weighted to match background (QCD) truth  $p_T$  distribution

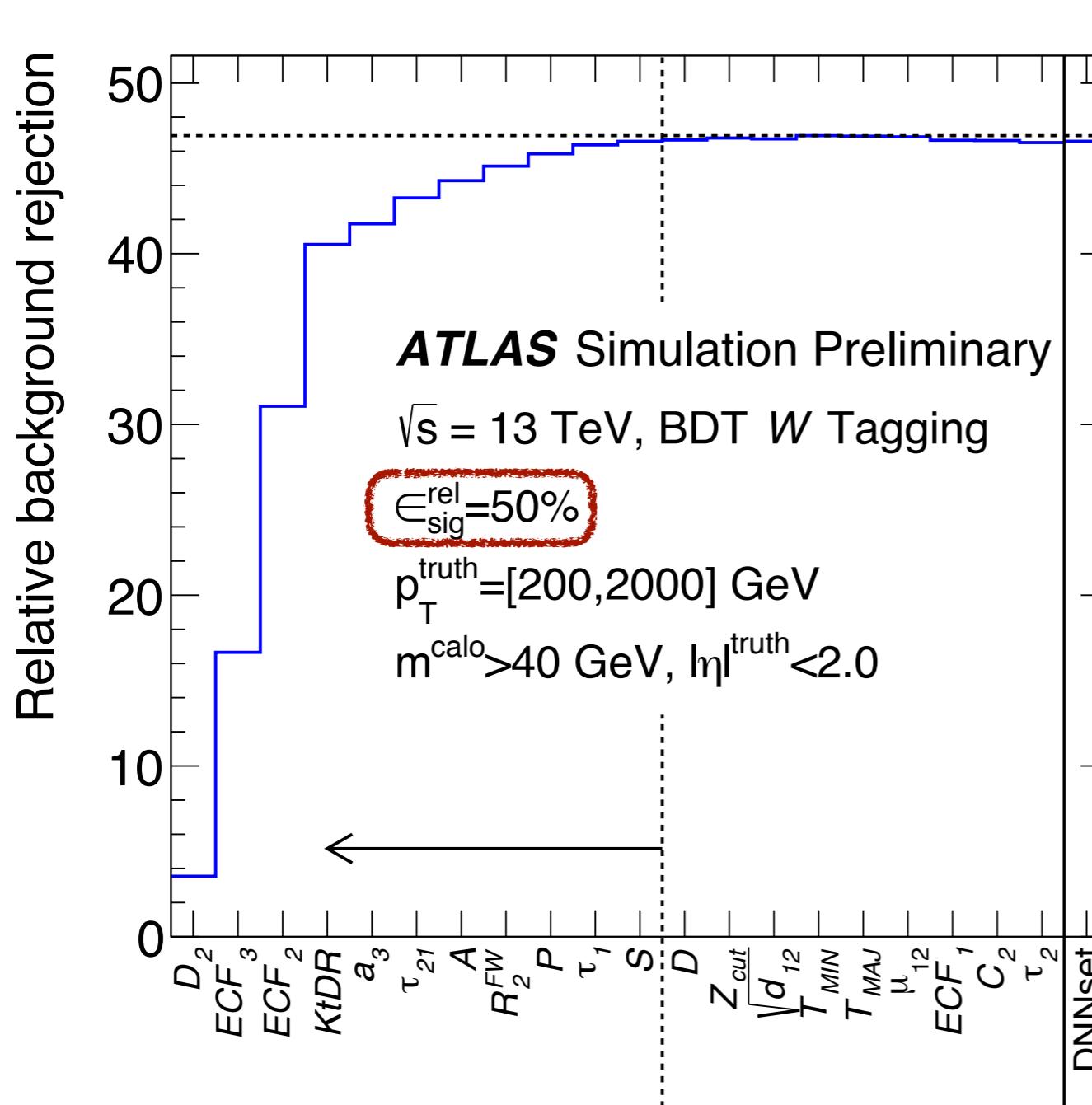
### W Tagging

- $p_T = [200, 2000]$  GeV,  $\eta = [-2, 2]$
- # Training signal jets =  $7 \times 10^5$
- # Training QCD jets =  $7 \times 10^5$

### Top Tagging

- $p_T = [350, 2000]$  GeV,  $\eta = [-2, 2]$
- # Training signal jets =  $10^6$
- # Training QCD jets =  $10^6$

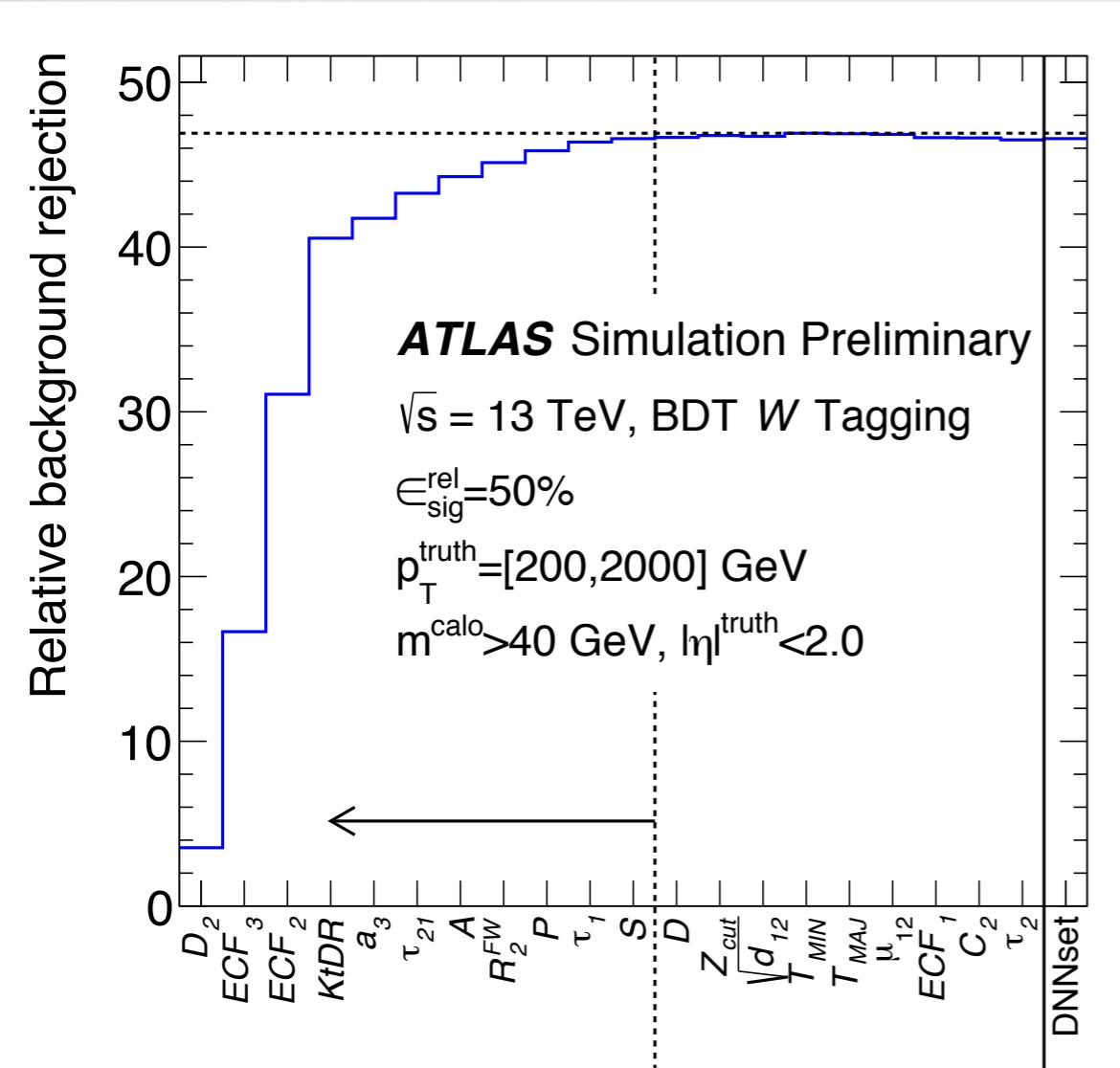
# BDT TRAINING - INPUTS OPTIMIZATION



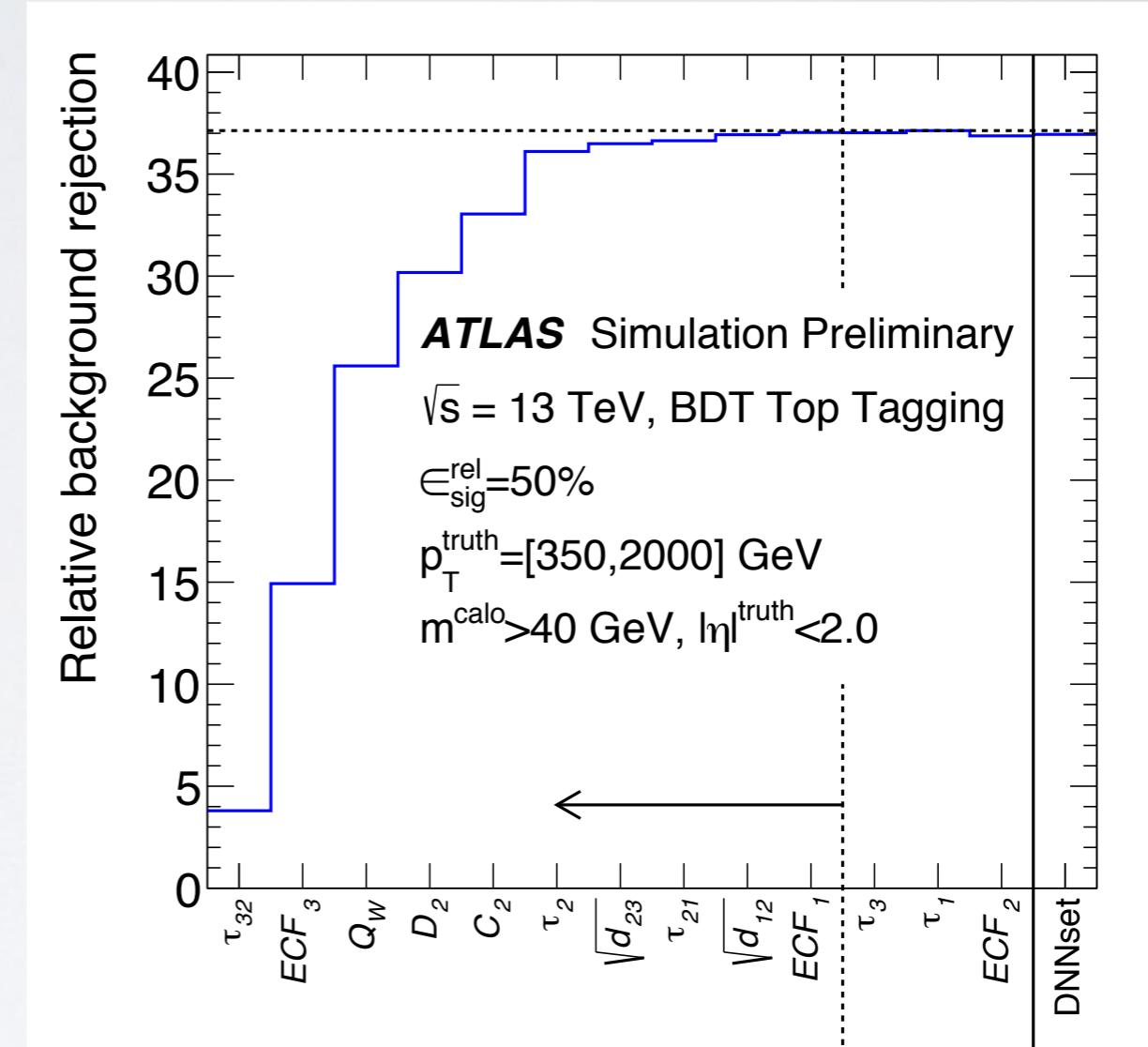
- Study the relative background rejection improvement by addition of new input variables
- The order of variables to be added are defined by the improvement in relative background rejection
- Use a flat  $p_T$  spectrum
- Stop adding variables once they stop adding significant information

# BDT TRAINING - INPUTS OPTIMIZATION

## W Tagging



## Top Tagging

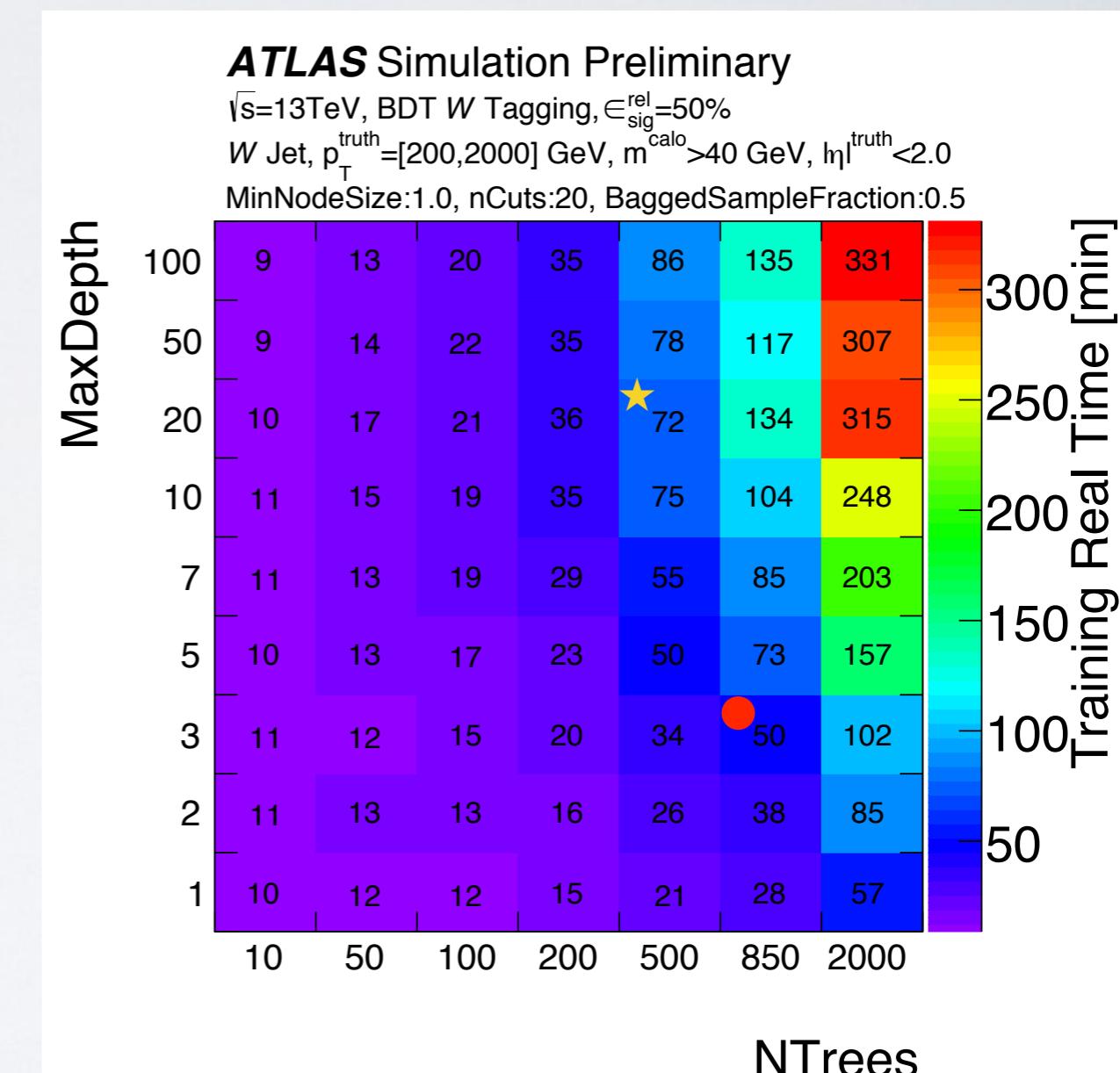
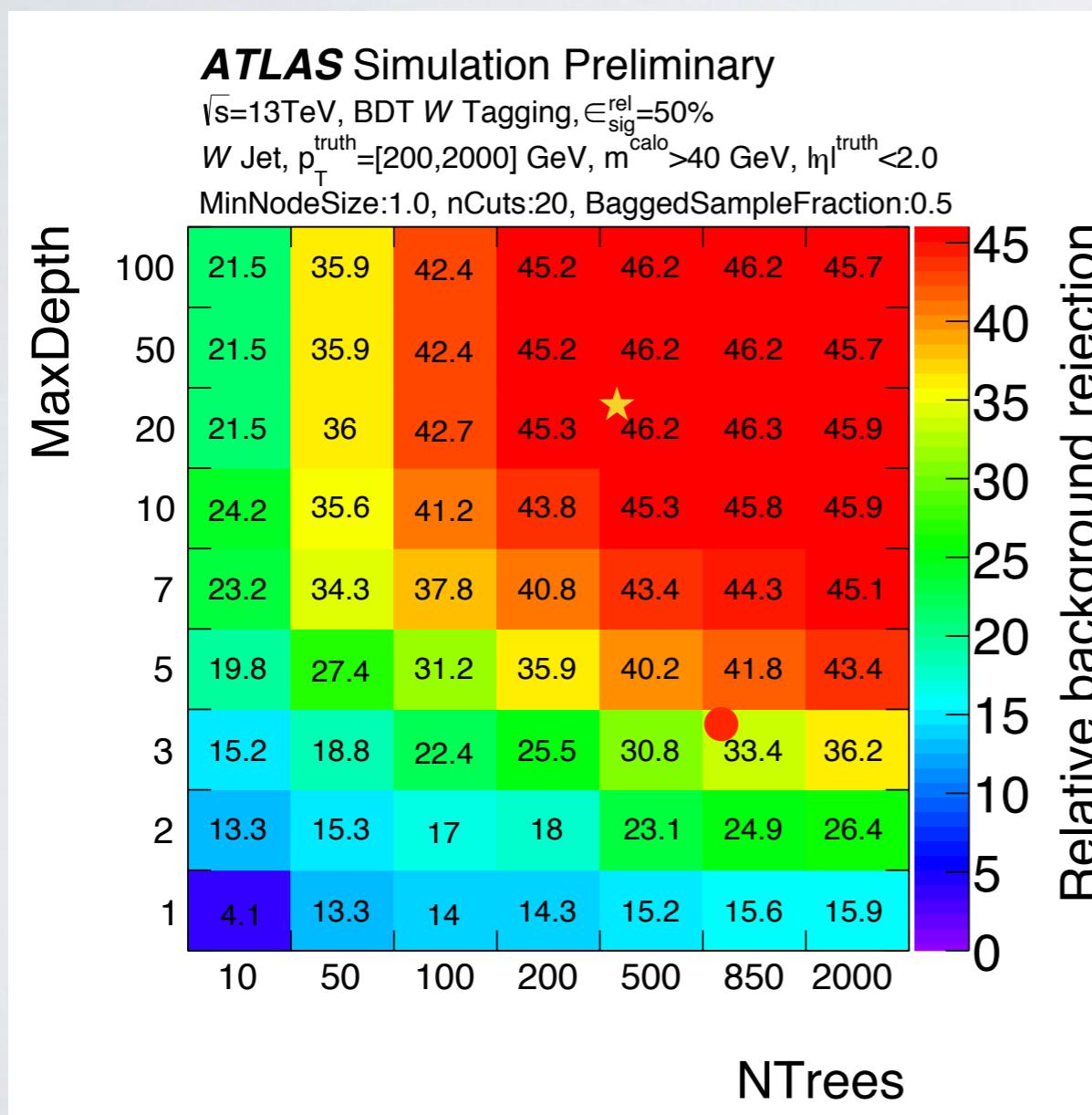


11 variables

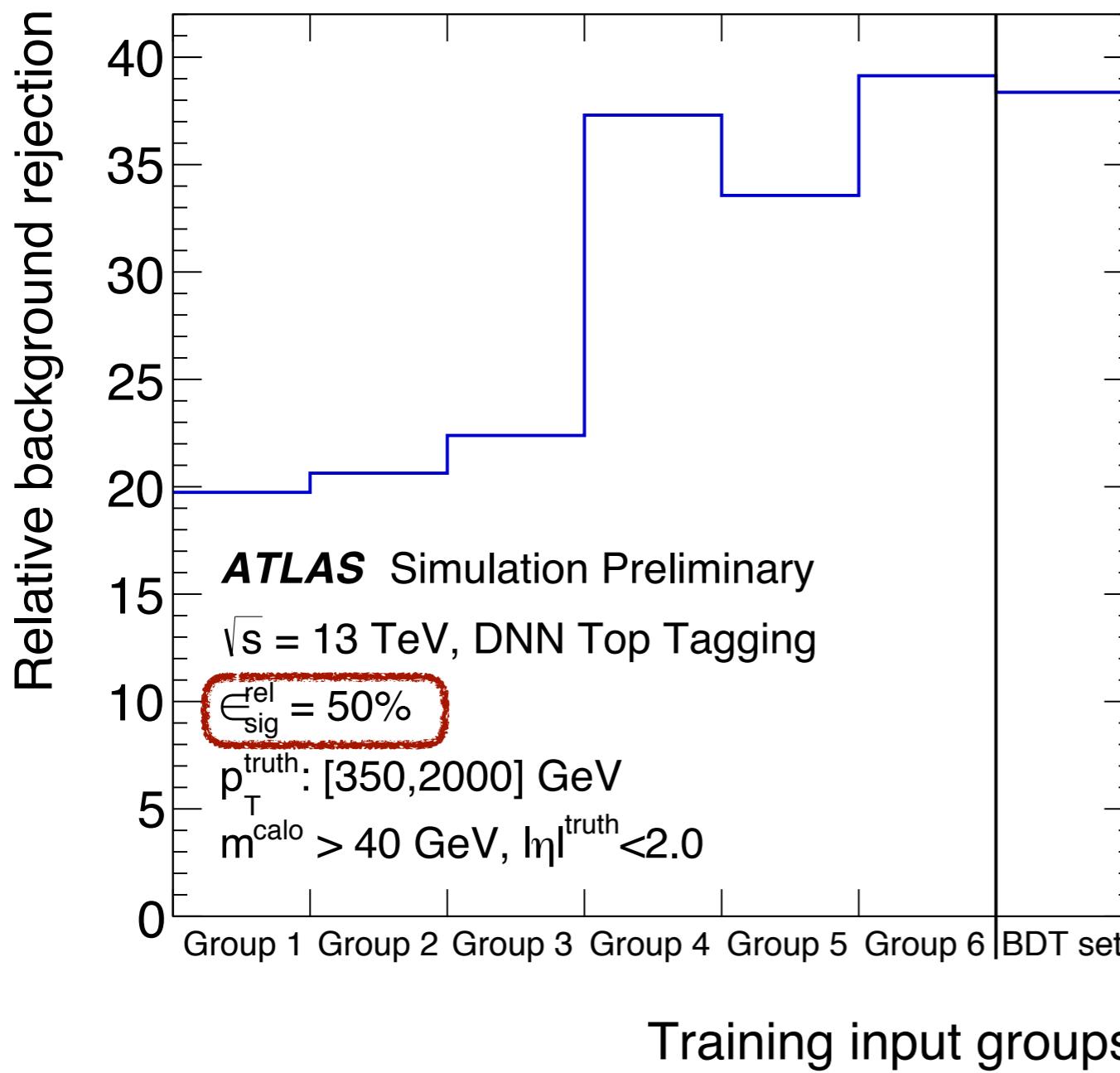
10 variables

# BDT TRAINING - HYPER-PARAMETER OPTIMIZATION

- Performed a parameter scan over many variables, most significant differences observed for NTrees and MaxDepth
- Shown for W, similar for top
- Star = Optimum settings found, Circle = TMVA default



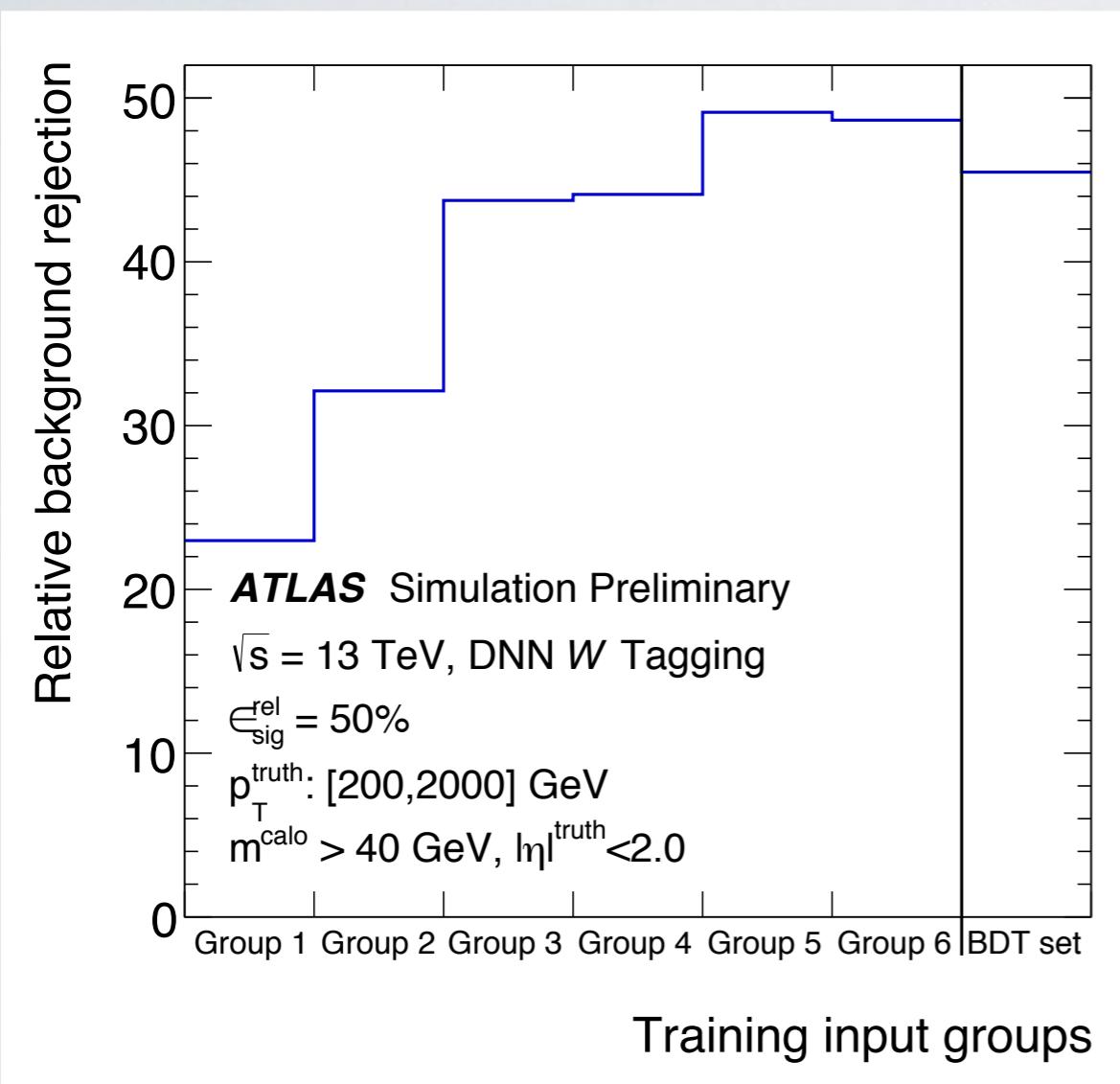
# DNN TRAINING - INPUTS OPTIMIZATION



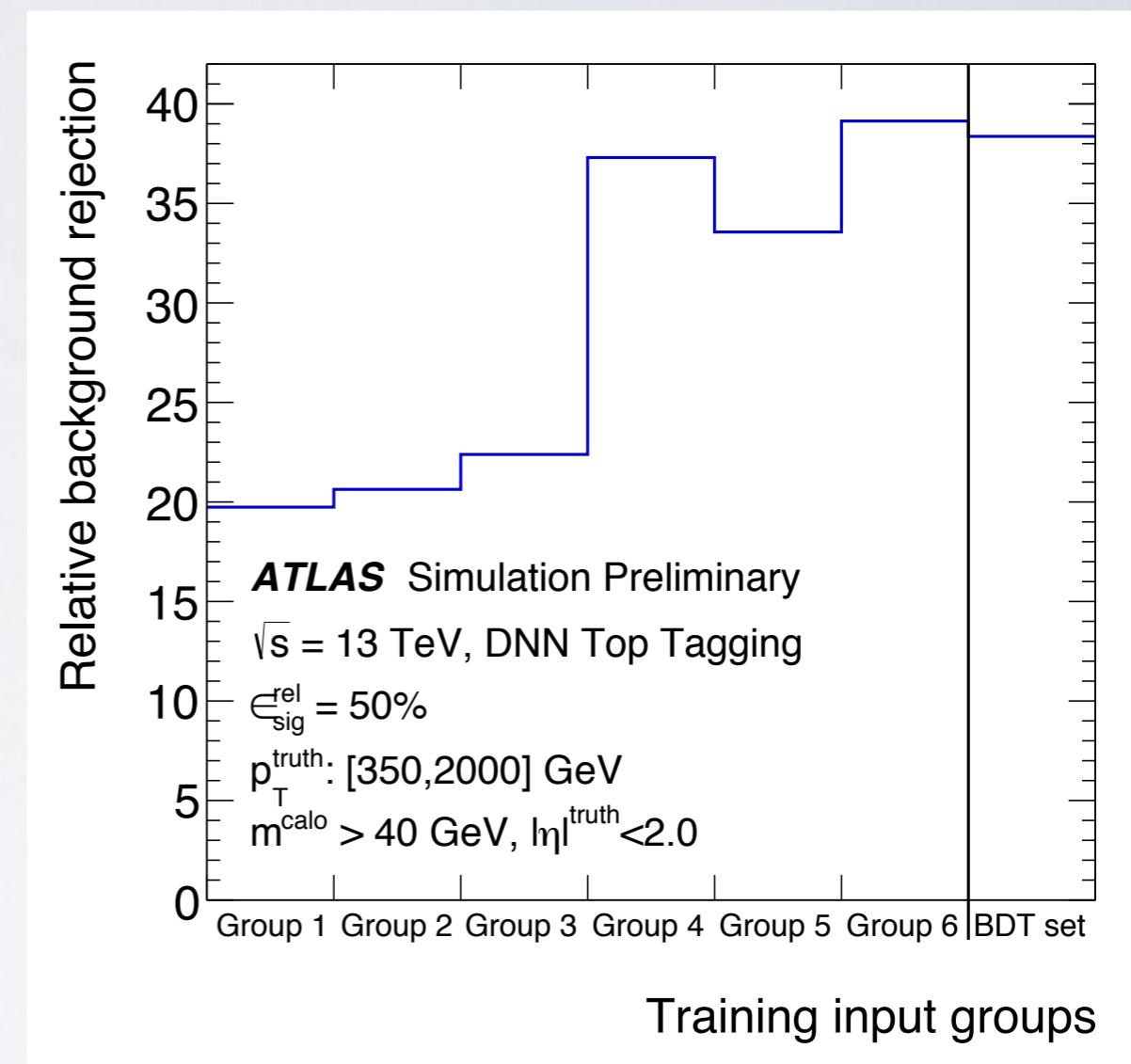
- Study the relative background rejection improvement by using different sets of input groups
- Groups are defined by
  - the physical information they provide (pronginess, scale...)
  - if the observable is defined as a function of the other observables
- Use a flat  $p_T$  spectrum
- Choose the set with the highest background rejection

# DNN TRAINING - INPUTS OPTIMIZATION

## W Tagging



## Top Tagging



Group 5 with 18 variables

Group 6 with 13 variables

# DNN TRAINING - HYPER-PARAMETER OPTIMIZATION

## Fixed hyper-parameters

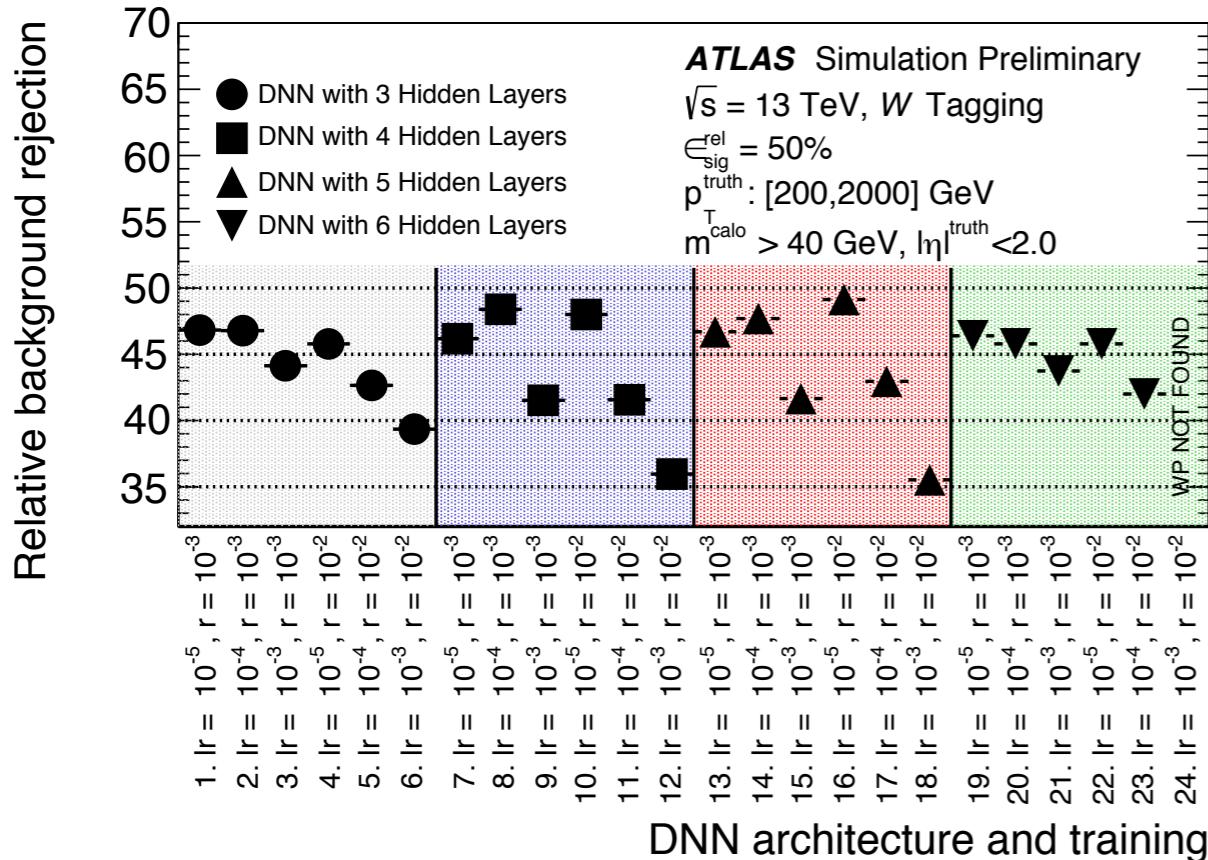
- Batch size = 200
- Number of epochs = 200, Early stopping = 50
- Optimizer = Adam

## Performed a grid search over many variables

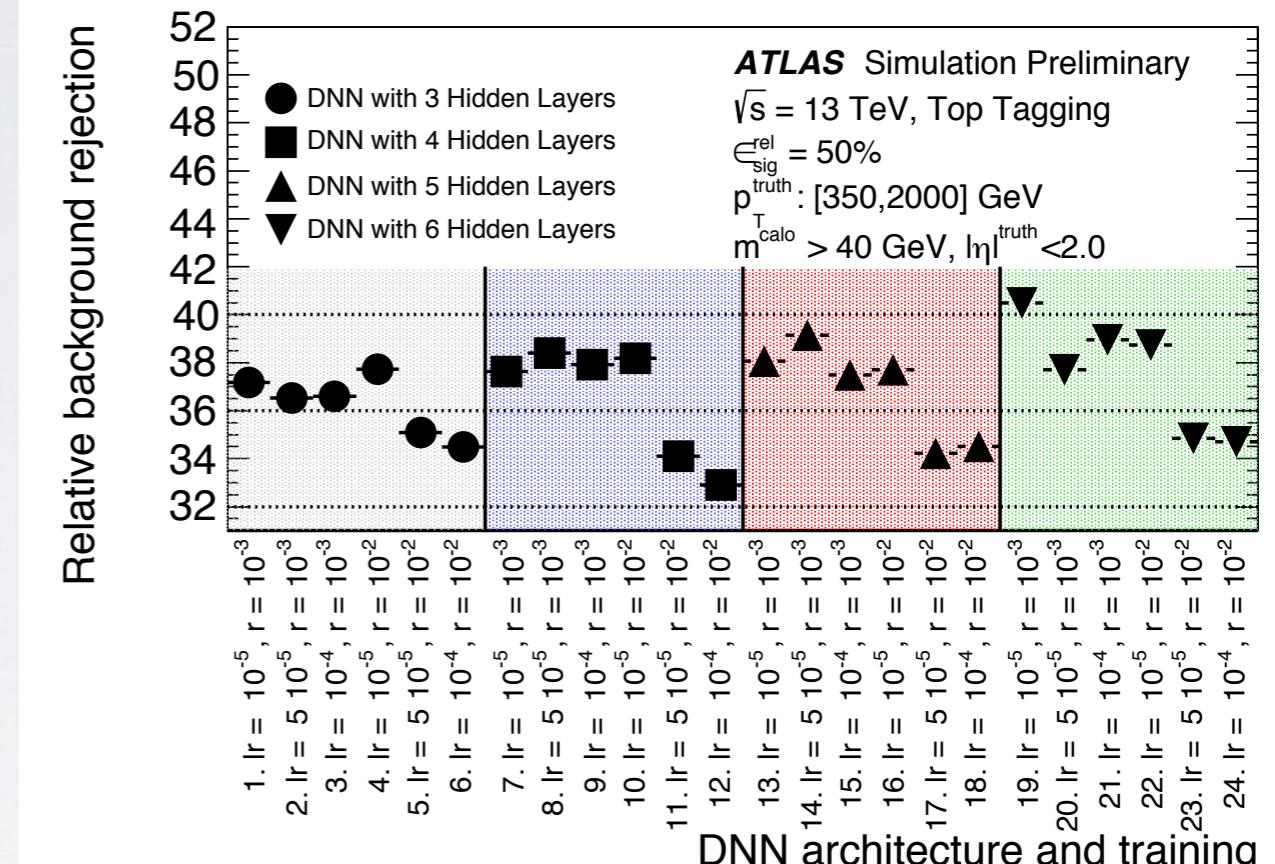
- Layer type = Dense with Batch Normalization, Maxout with Batch Normalization
  - Number of Maxout layers = 5, 10, 15, 20, 25
- Number of hidden layers = 3, 4, 5, 6
- Activation function = Rectified linear units (relu), tanh
- Weight initialization = Glorot uniform, He normal
- Learning rate
  - $W = 10^{-5}, 10^{-4}, 10^{-3}$
  - $Top = 10^{-5}, 5 \times 10^{-5}, 10^{-4}$
- L1 regularizer =  $10^{-3}, 10^{-2}$

# DNN TRAINING - HYPER-PARAMETER OPTIMIZATION

## W Tagging



## Top Tagging



Grid search for DNN chosen variables

- Layer type = Dense with Batch Normalization
- Activation function = Rectified linear units
- Weight initialization = Glorot uniform

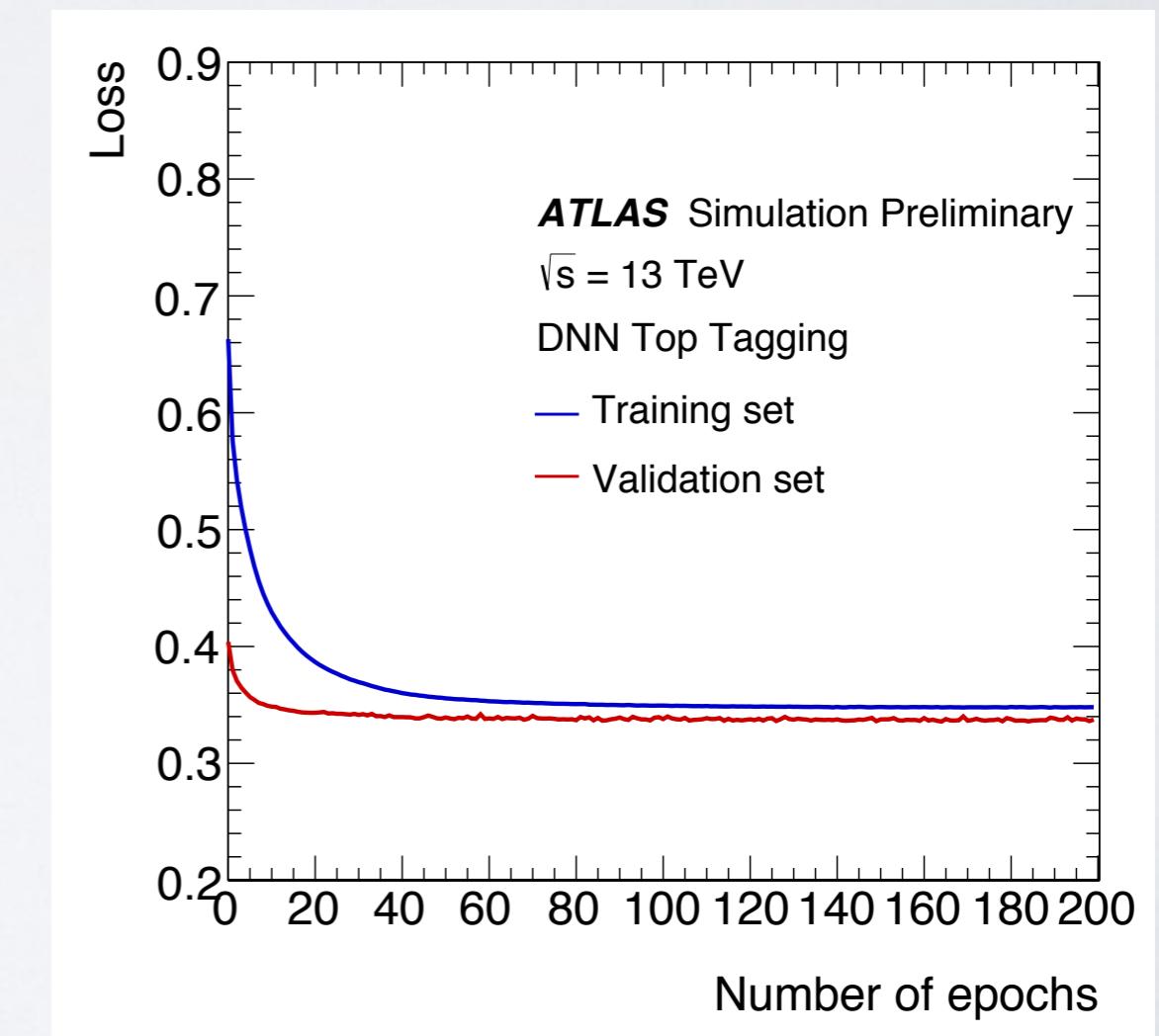
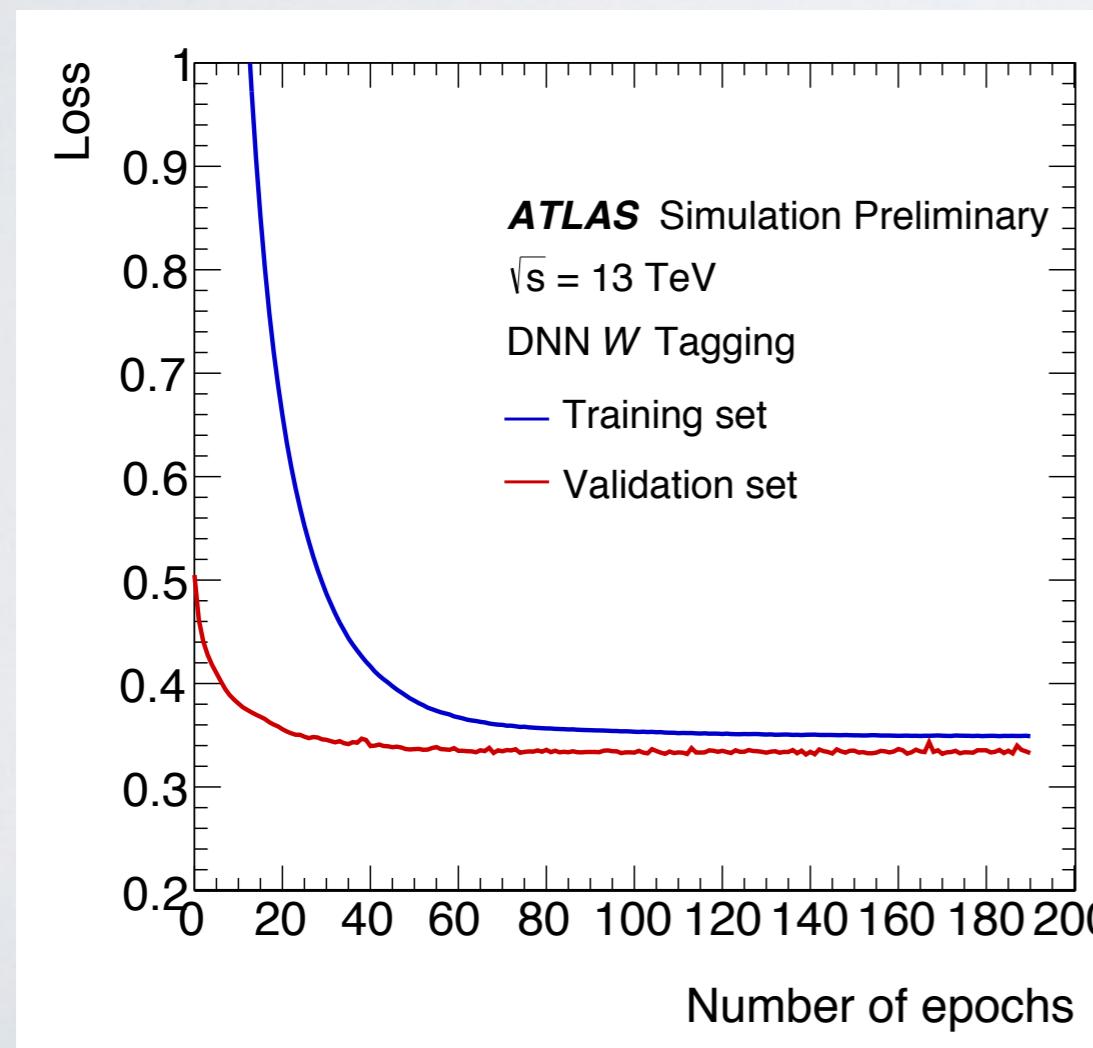
**W Chosen** Learning rate =  $10^{-5}$ , L1 regularizer =  $10^{-2}$ , Hidden layers = 5

**Top Chosen** Learning rate =  $5 \times 10^{-5}$ , L1 regularizer =  $10^{-3}$ , Hidden layers = 5

# DNN TRAINING - OVERTRAINING

- DNN minimizes the training loss
- In order to have a handle on the over-training while training, the loss is calculated in 2 different sets
  1. **Training set**: DNN uses this set to optimize the classifier
  2. **Validation set**: Independent of the training set

**Loss of the validation set  $\leq$  Loss of the training set**

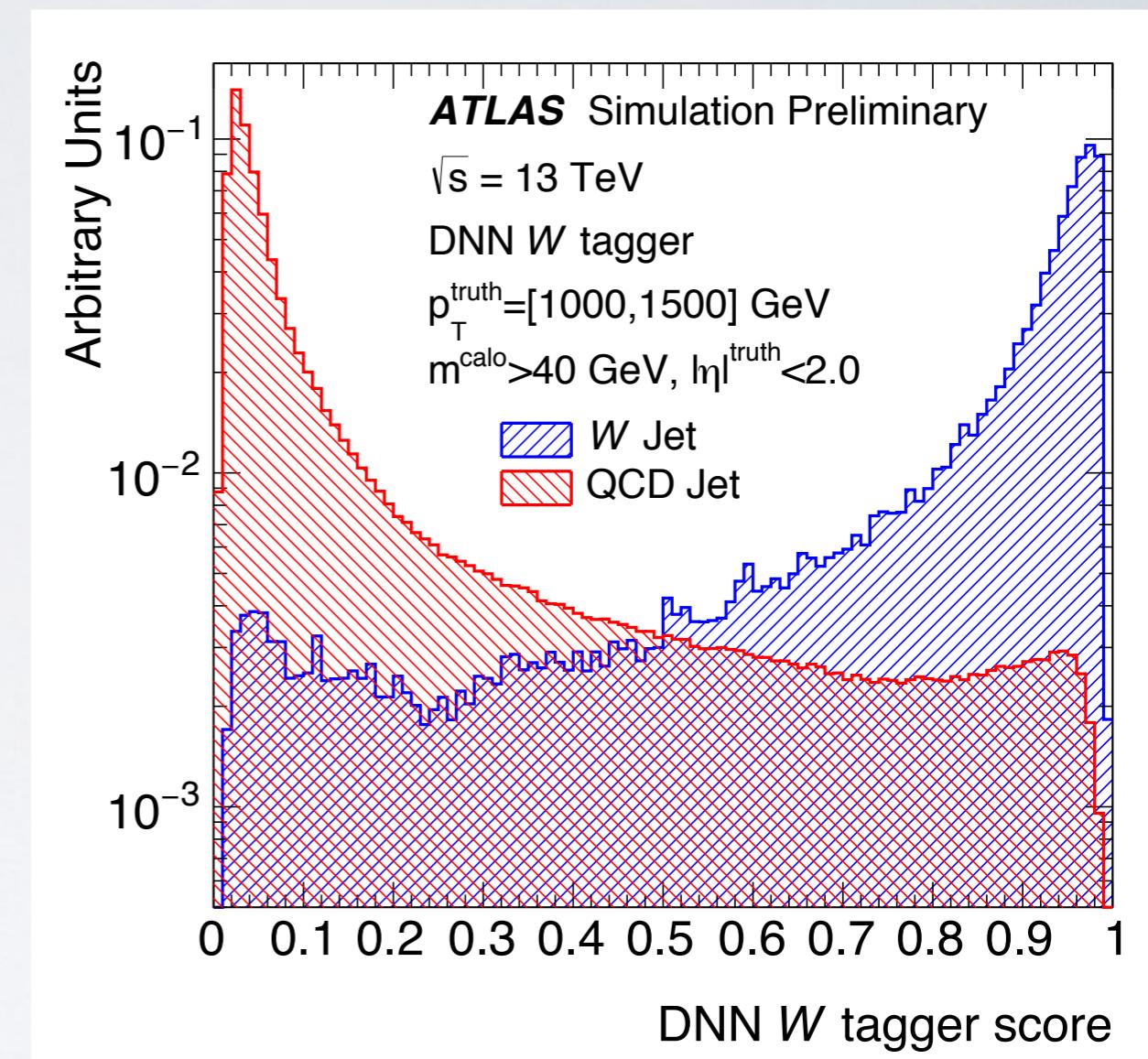
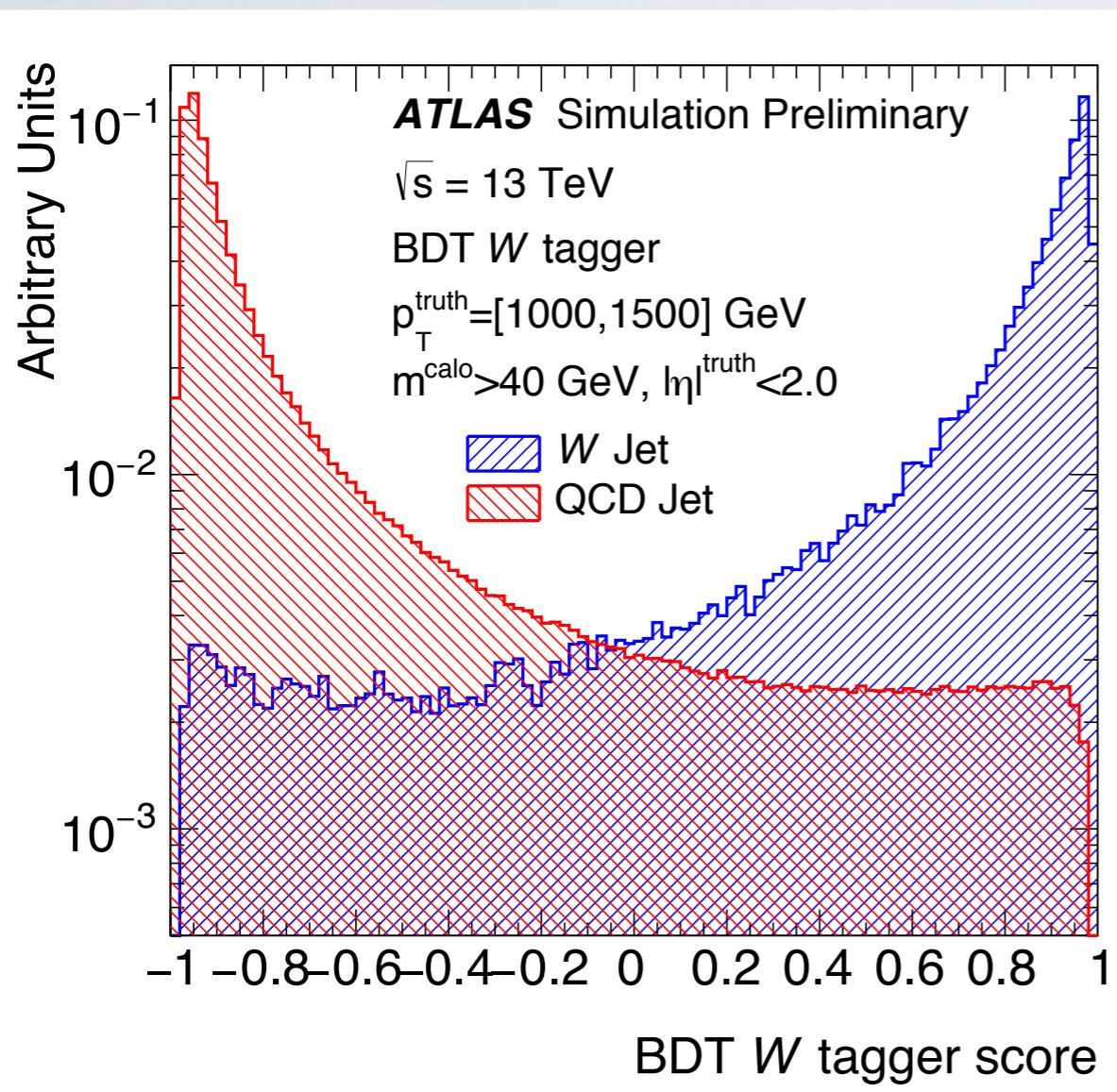


# BDT & DNN CHOSEN VARIABLES

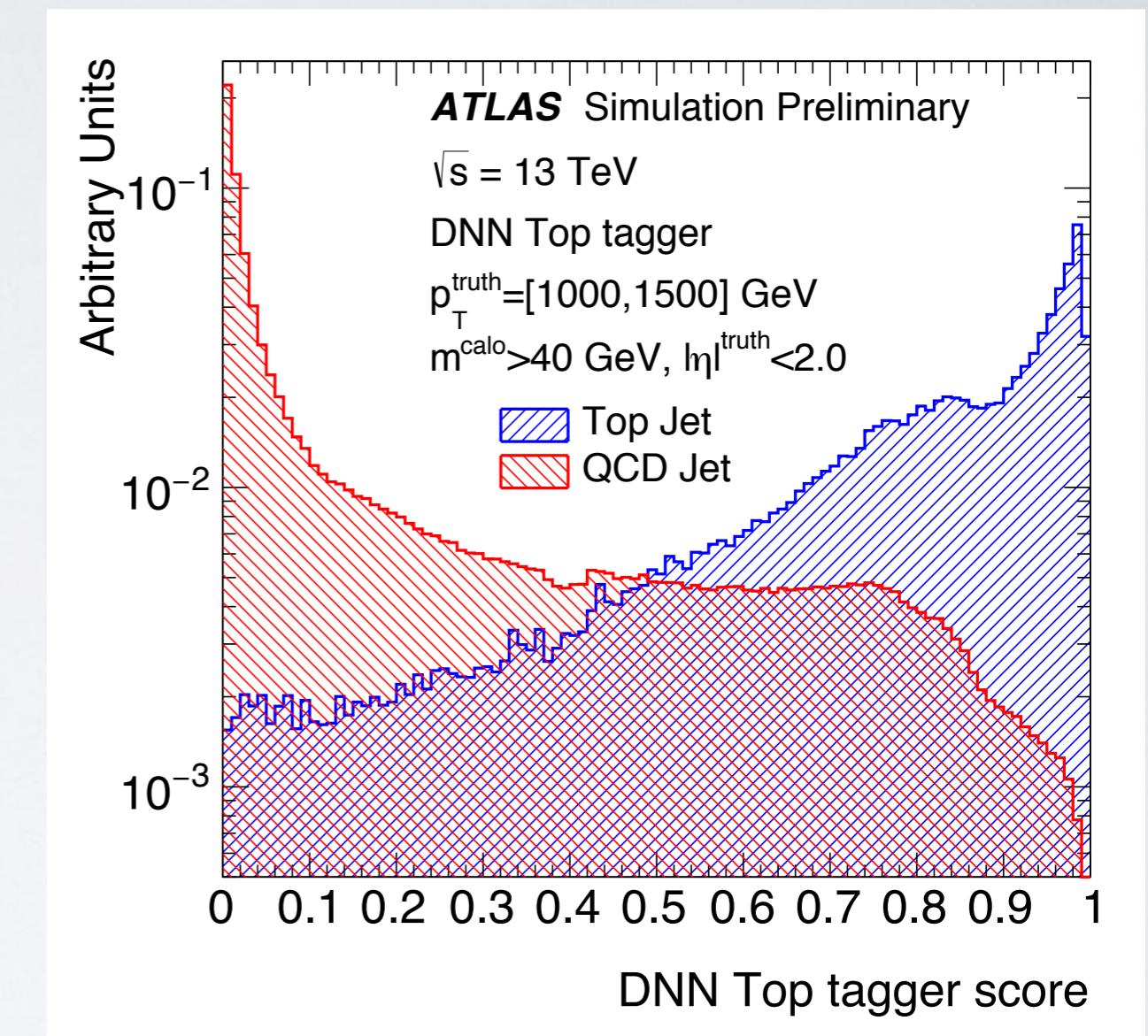
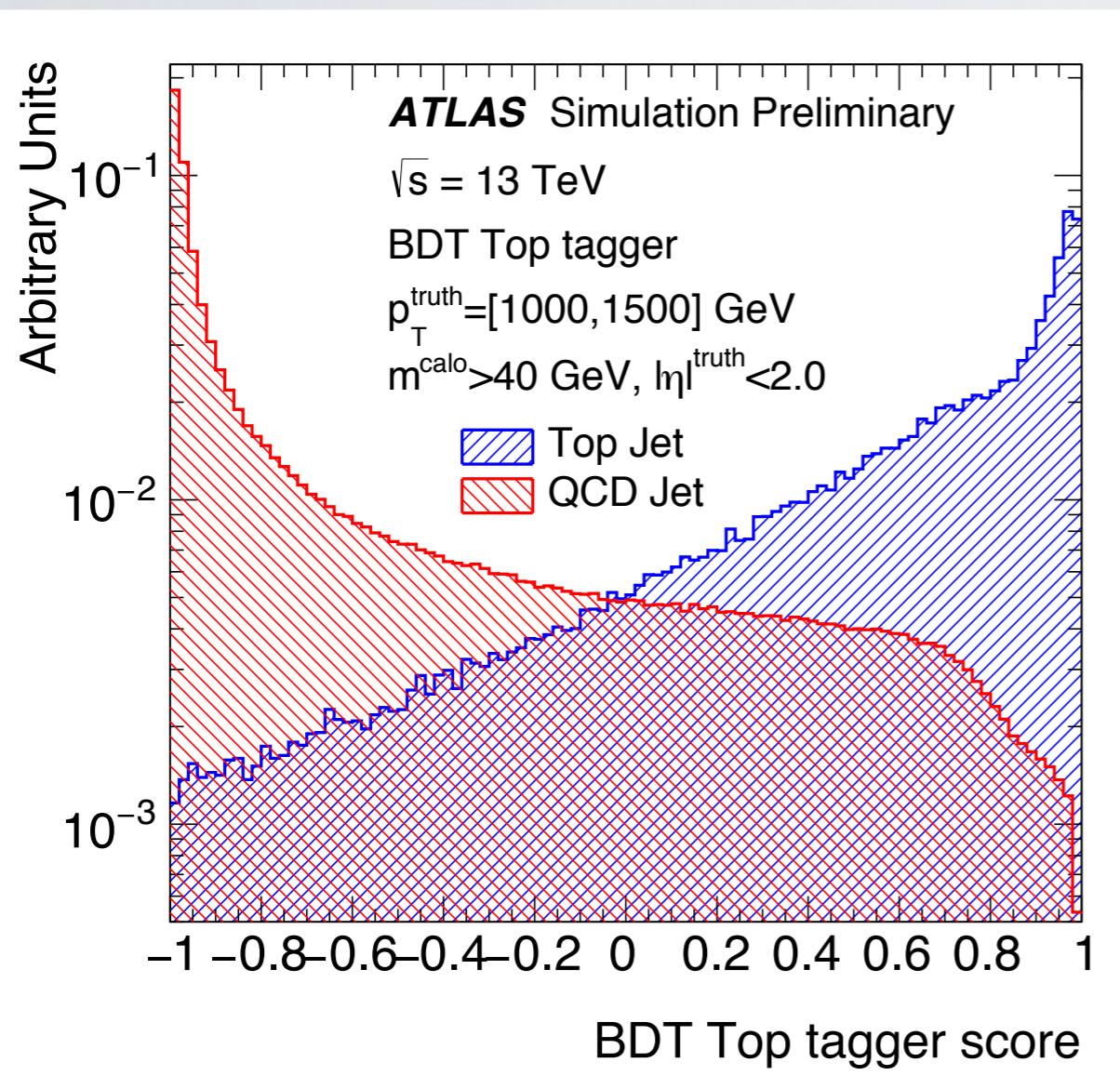
- BDT and DNN find different sets of inputs to be optimal
- Fair comparison on same set of inputs → Train DNN and BDT on 2 different set of observables for each tagger
  - DNN with DNN Obs., DNN with BDT Obs.
  - BDT with BDT Obs., BDT with DNN Obs.
  - If not stated explicitly, each method is trained with its own optimized observables
- Full list of variables for each case is in the backup

# RESULTS

# BDT & DNN CLASSIFIERS - W TAGGING



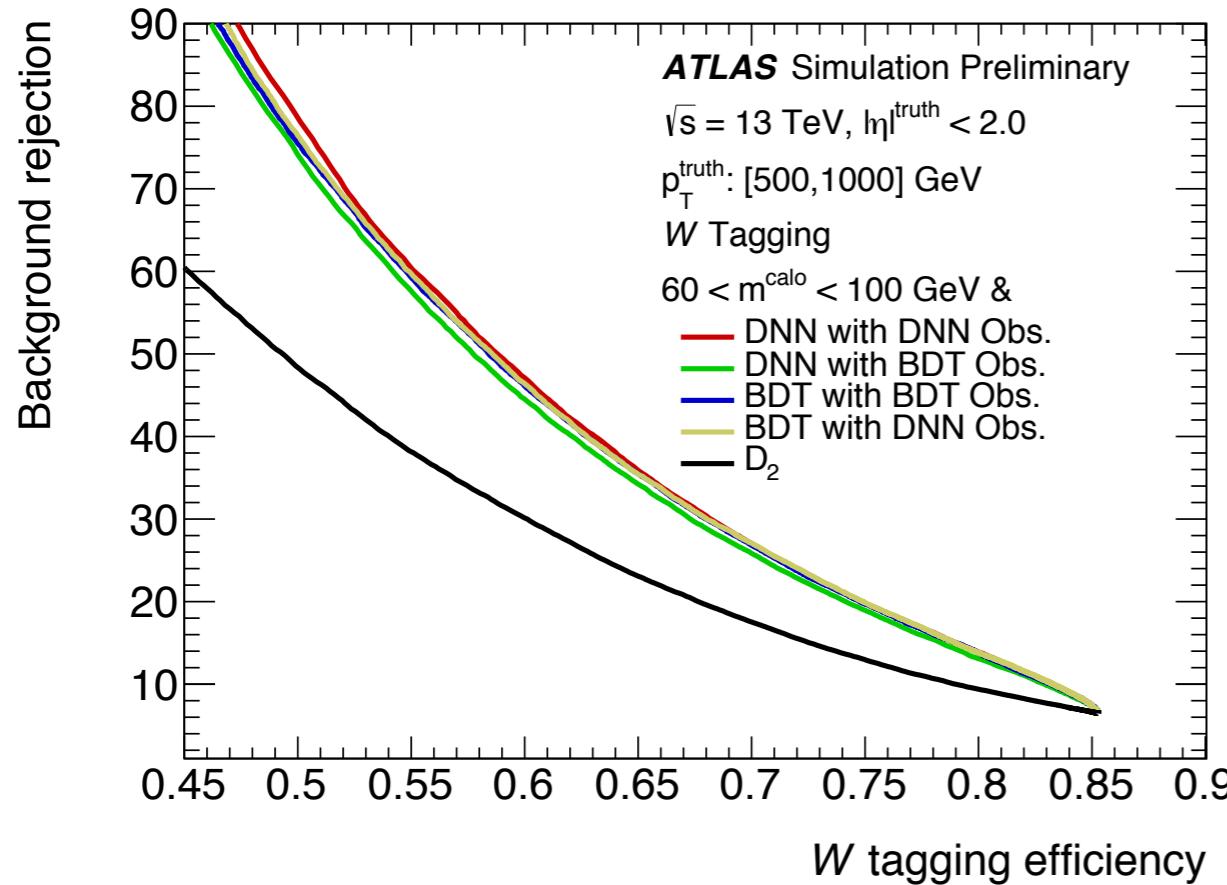
# BDT & DNN CLASSIFIERS - TOPTAGGING



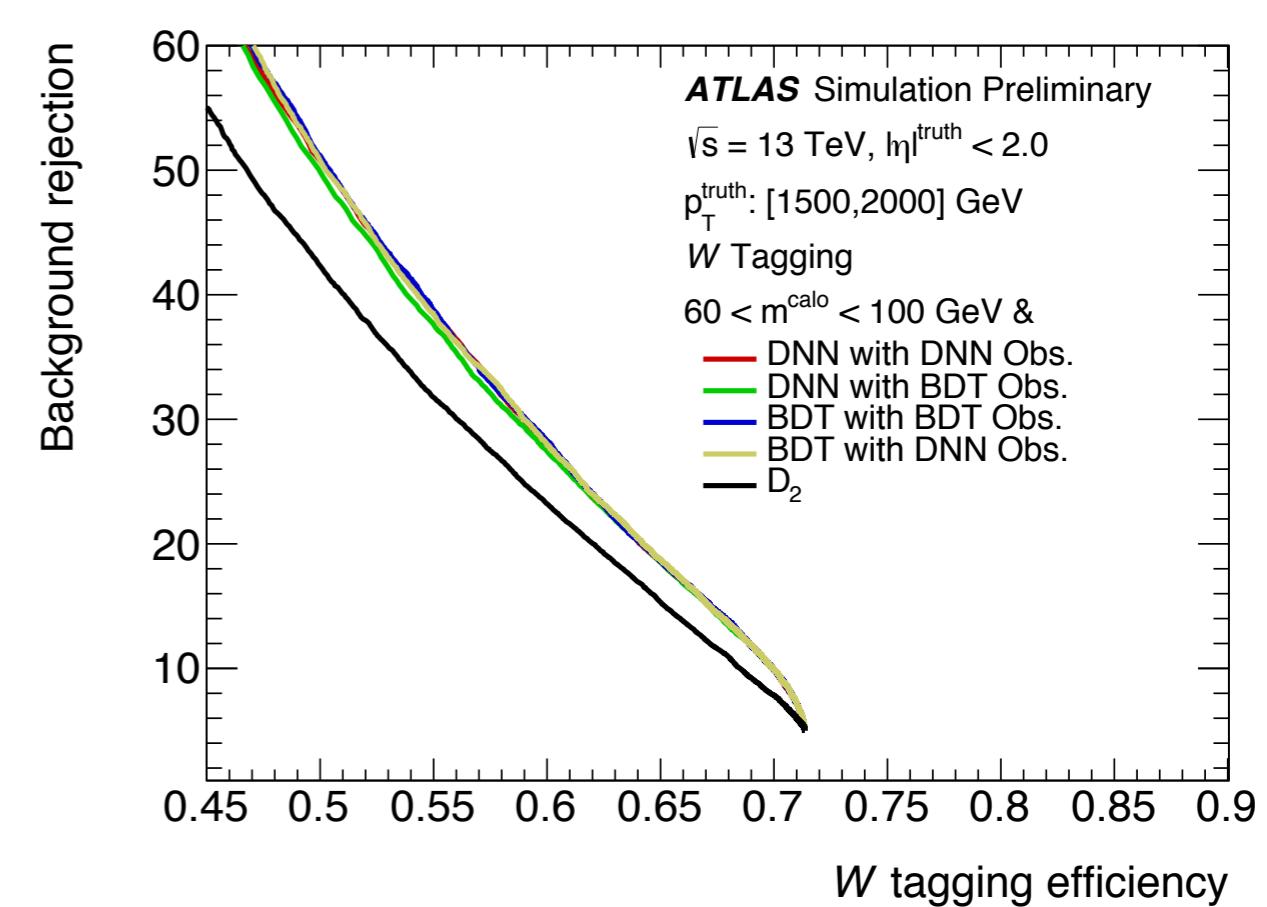
# PERFORMANCE EVALUATION - W TAGGING

## ROC Curves

$p_T = [500, 1000]$  GeV



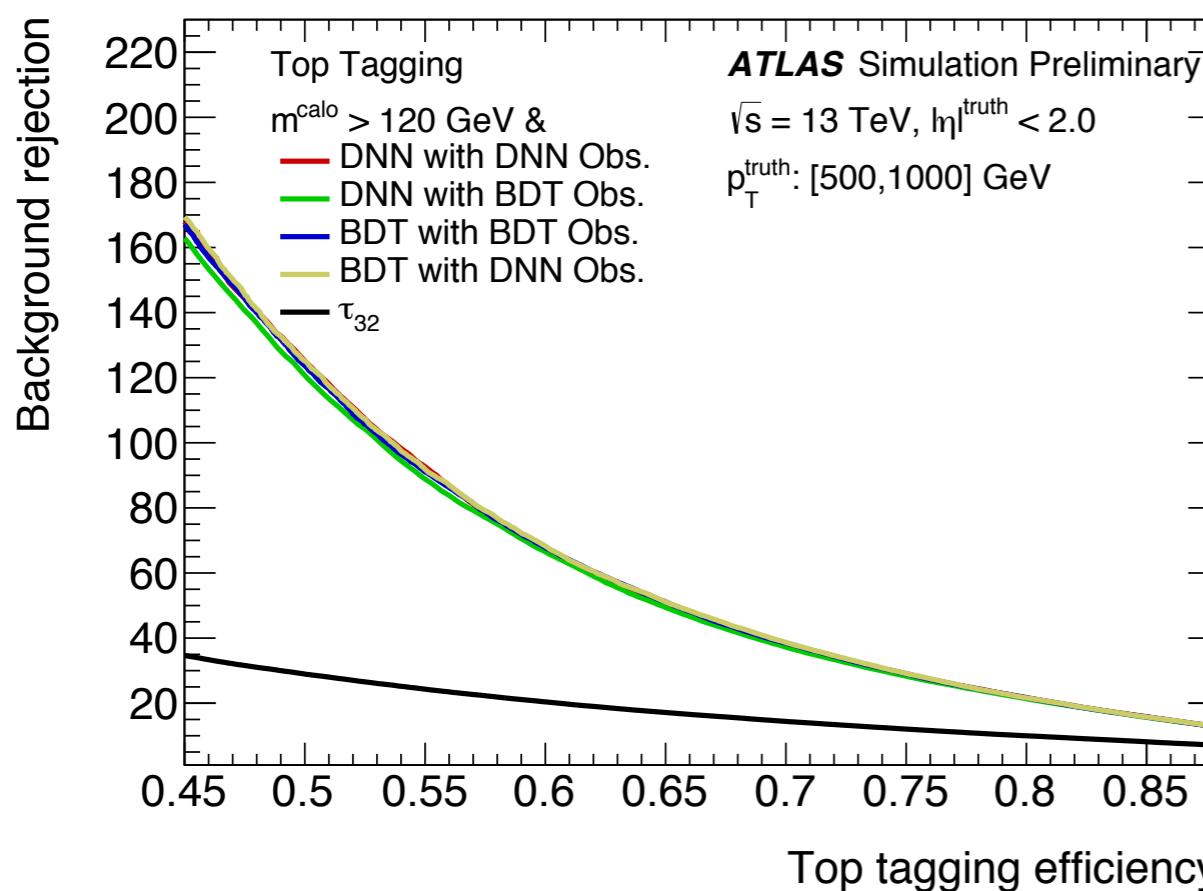
$p_T = [1500, 2000]$  GeV



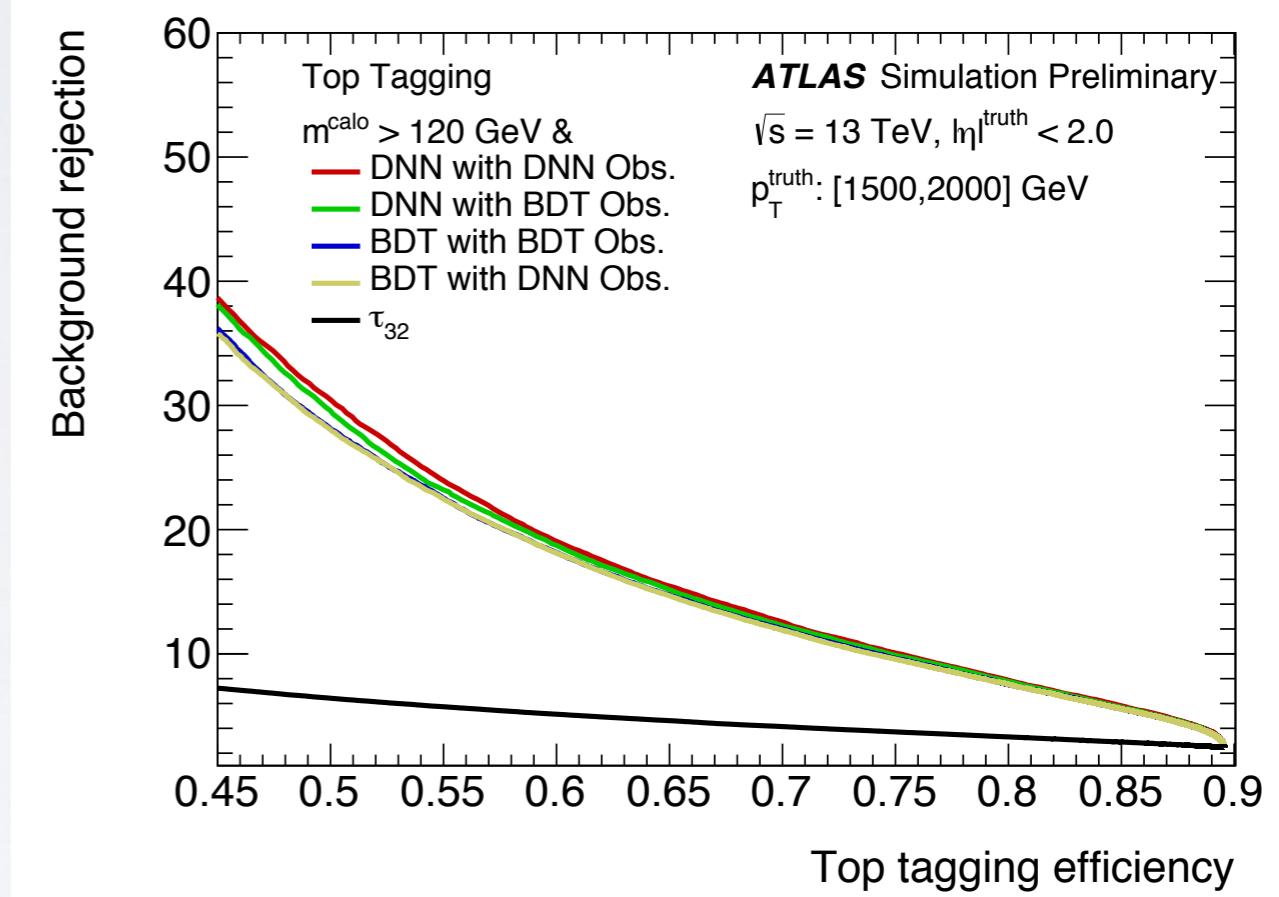
# PERFORMANCE EVALUATION - TOP TAGGING

## ROC Curves

$p_T = [500, 1000] \text{ GeV}$



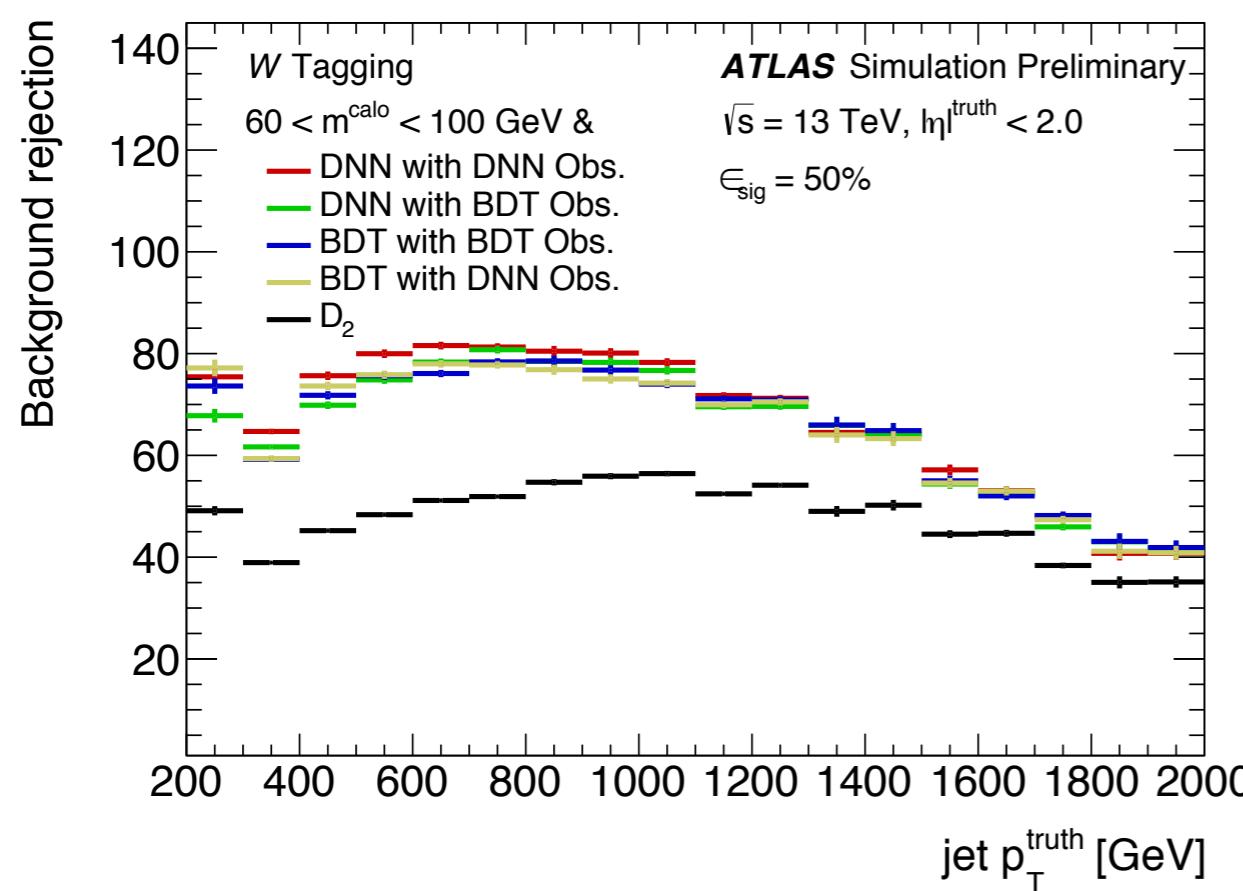
$p_T = [1500, 2000] \text{ GeV}$



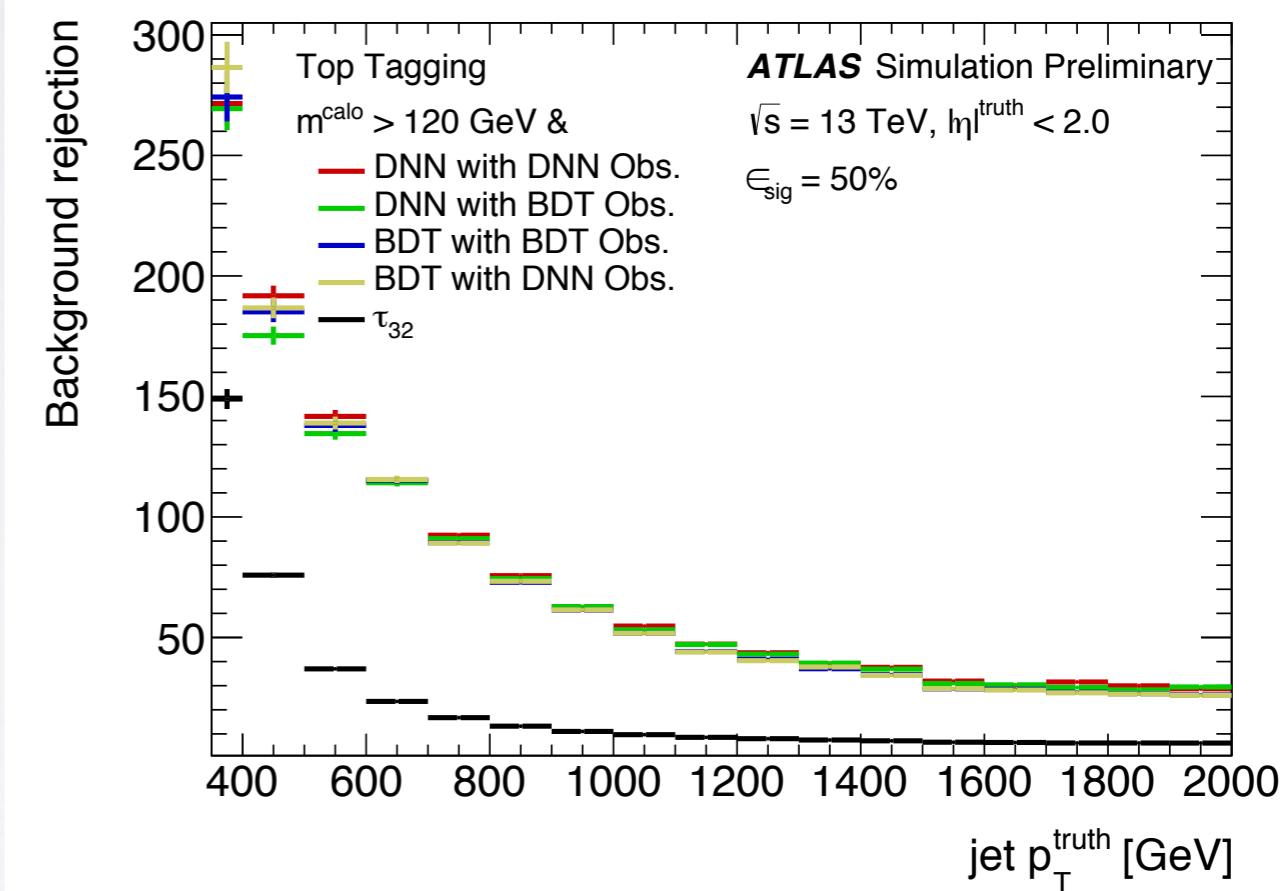
# PERFORMANCE EVALUATION - SUMMARY

## Background Rejection at 50% Fixed Efficiency WP

### W Tagging



### Top Tagging

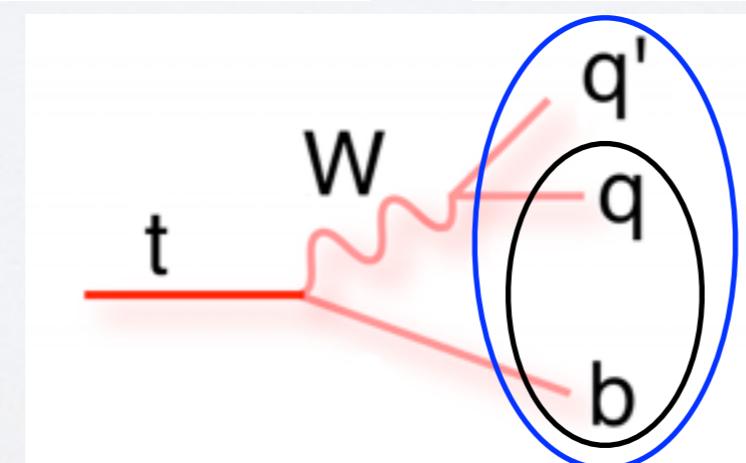
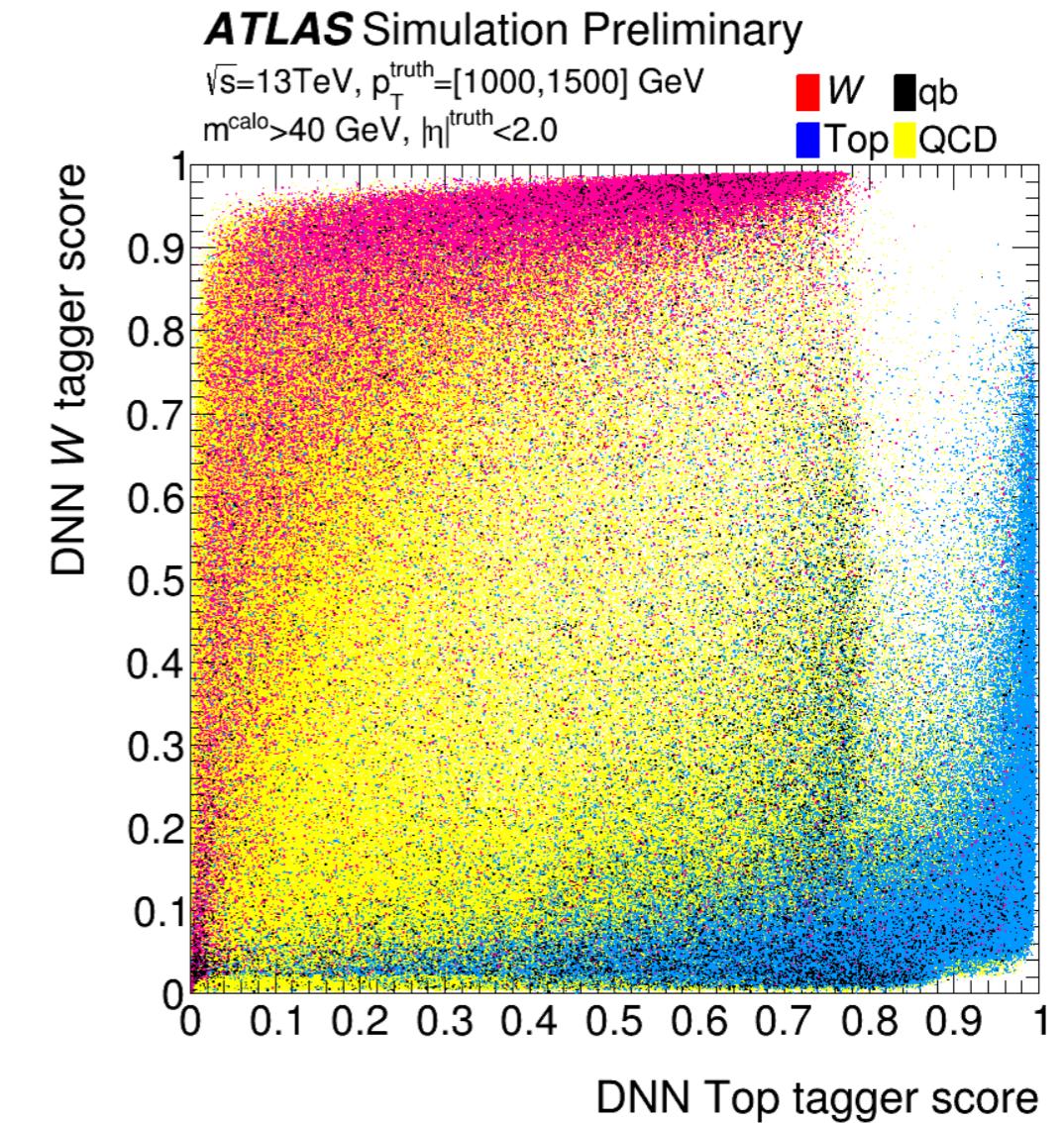
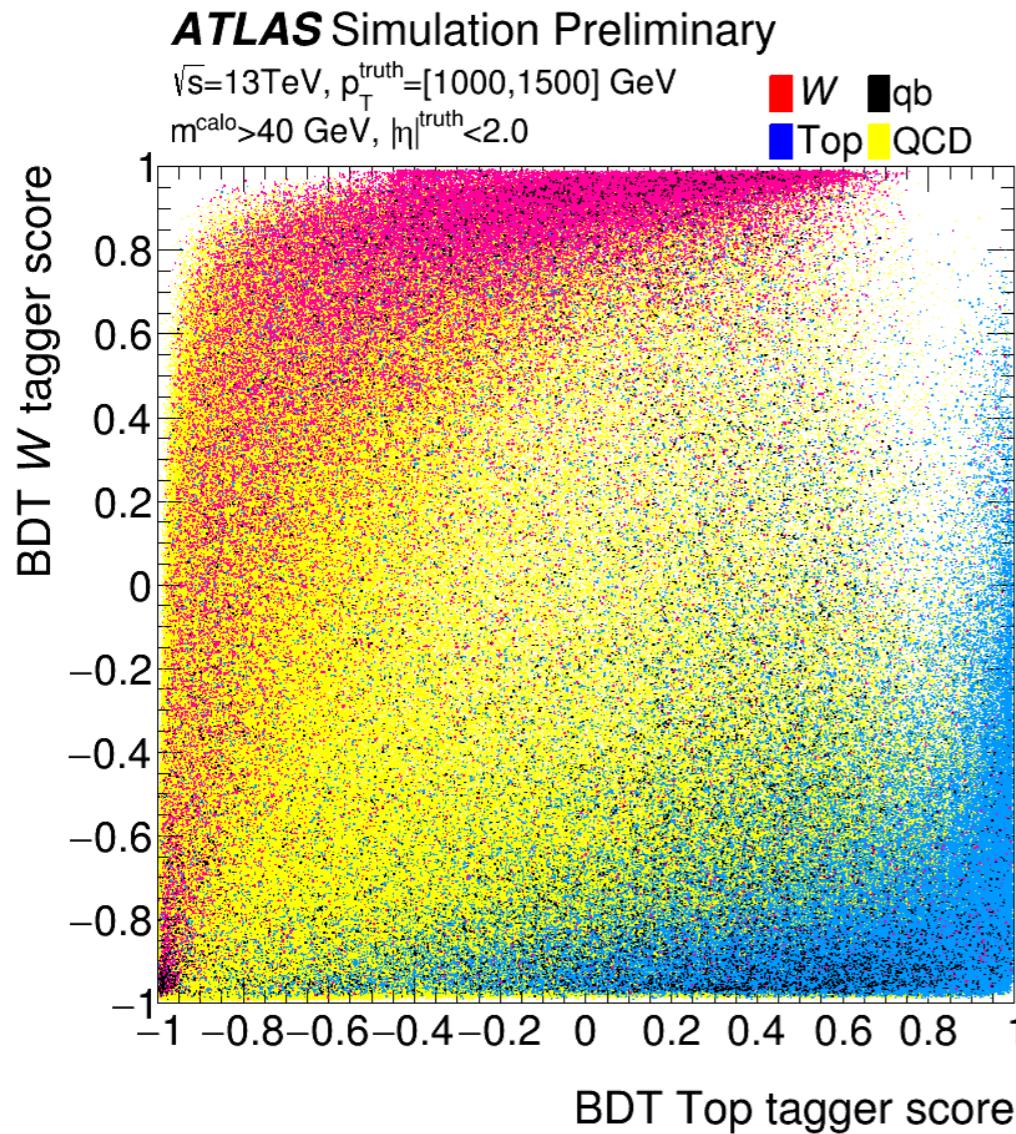


- Improvements observed for both W and top tagging
- Magnitude of impact differs for W and top tagging, but not the overall benefit of using a BDT or DNN

# DISCRIMINANT CORRELATIONS

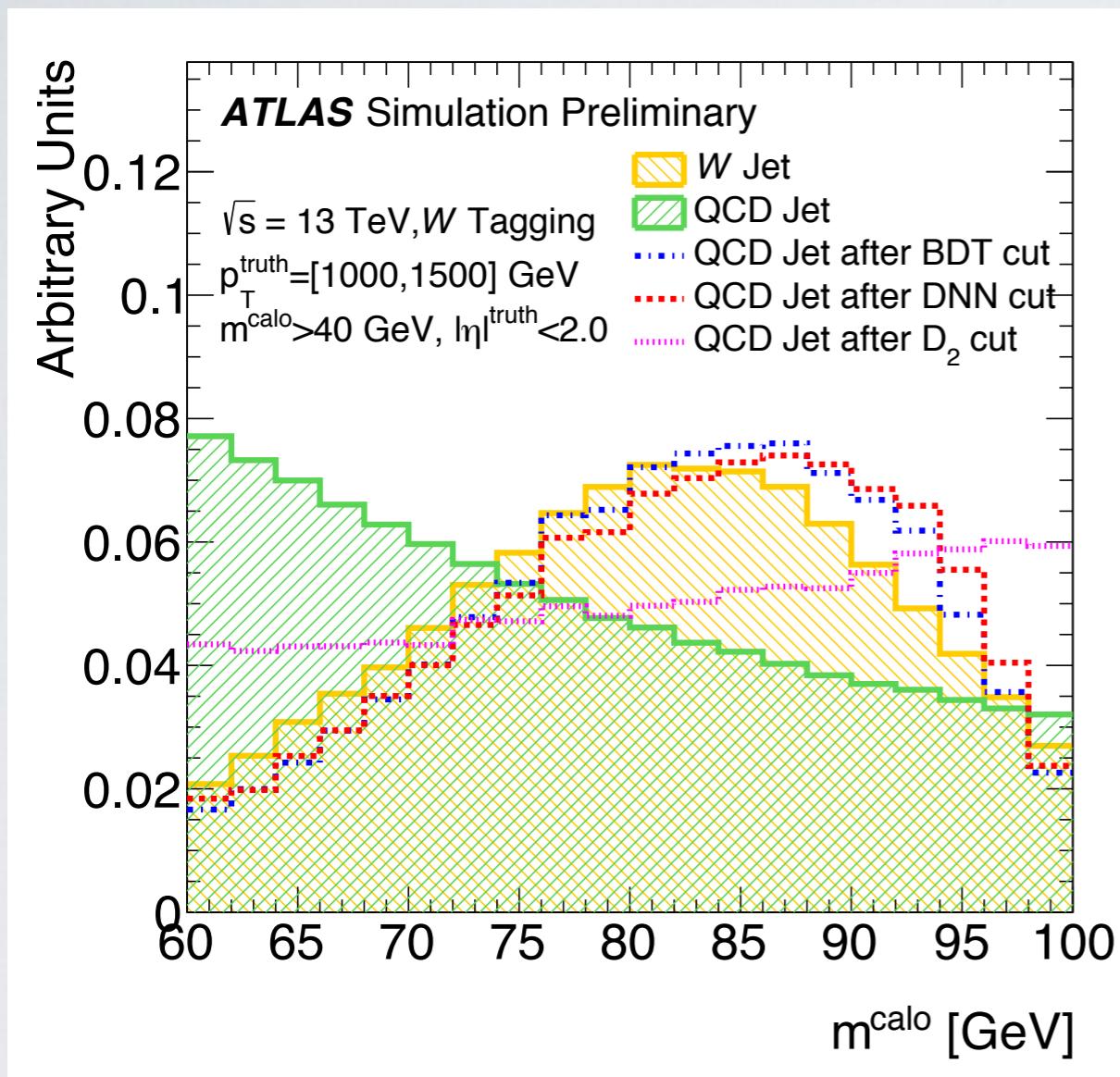
- ML techniques are expected to learn
  - linear correlations
  - non-linear correlations
- Studied linear correlations to understand what was learned

# DISCRIMINANT CORRELATIONS

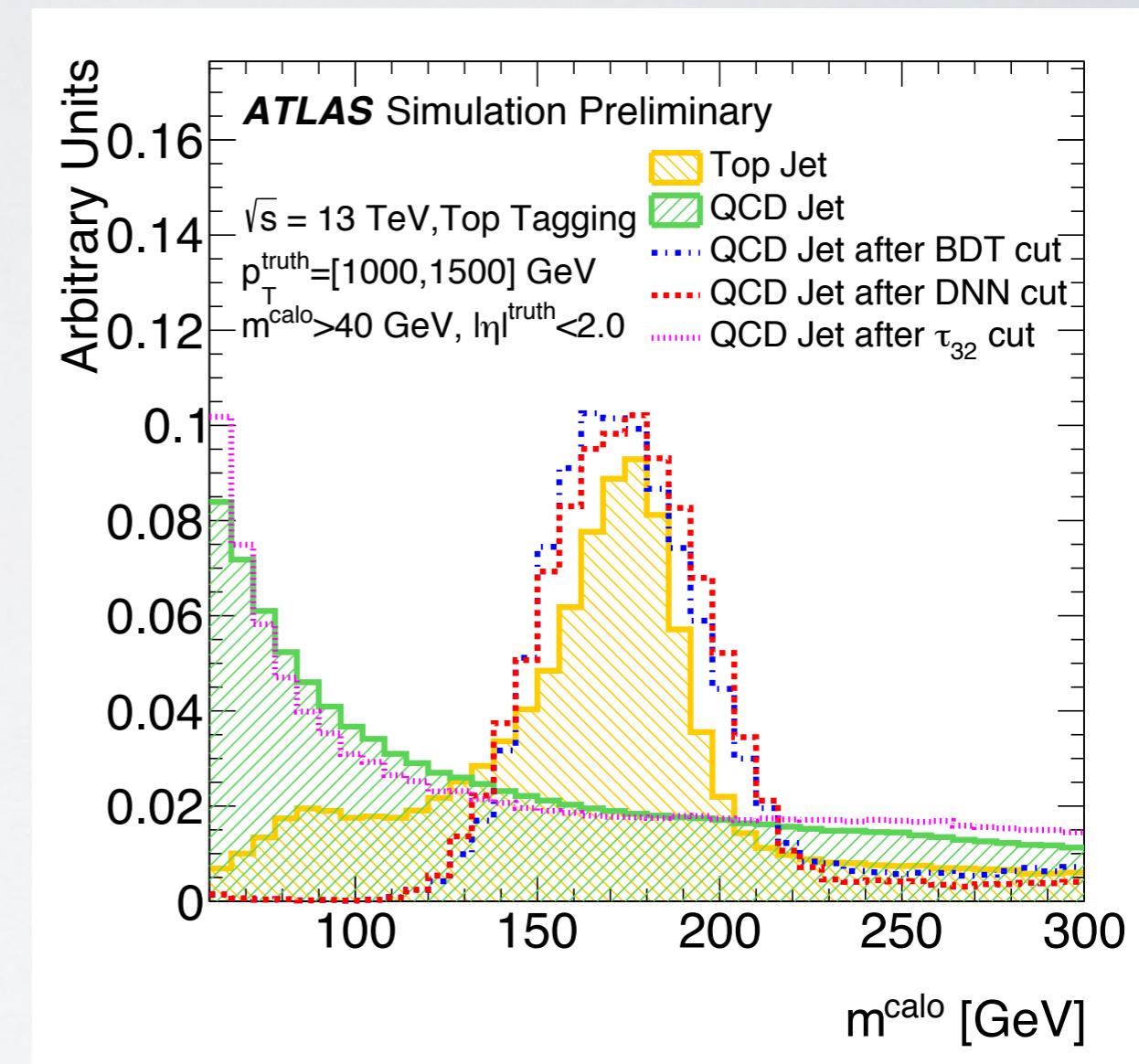


# BACKGROUND MASS DISTRIBUTION BEFORE & AFTER TAGGING

## W Tagging

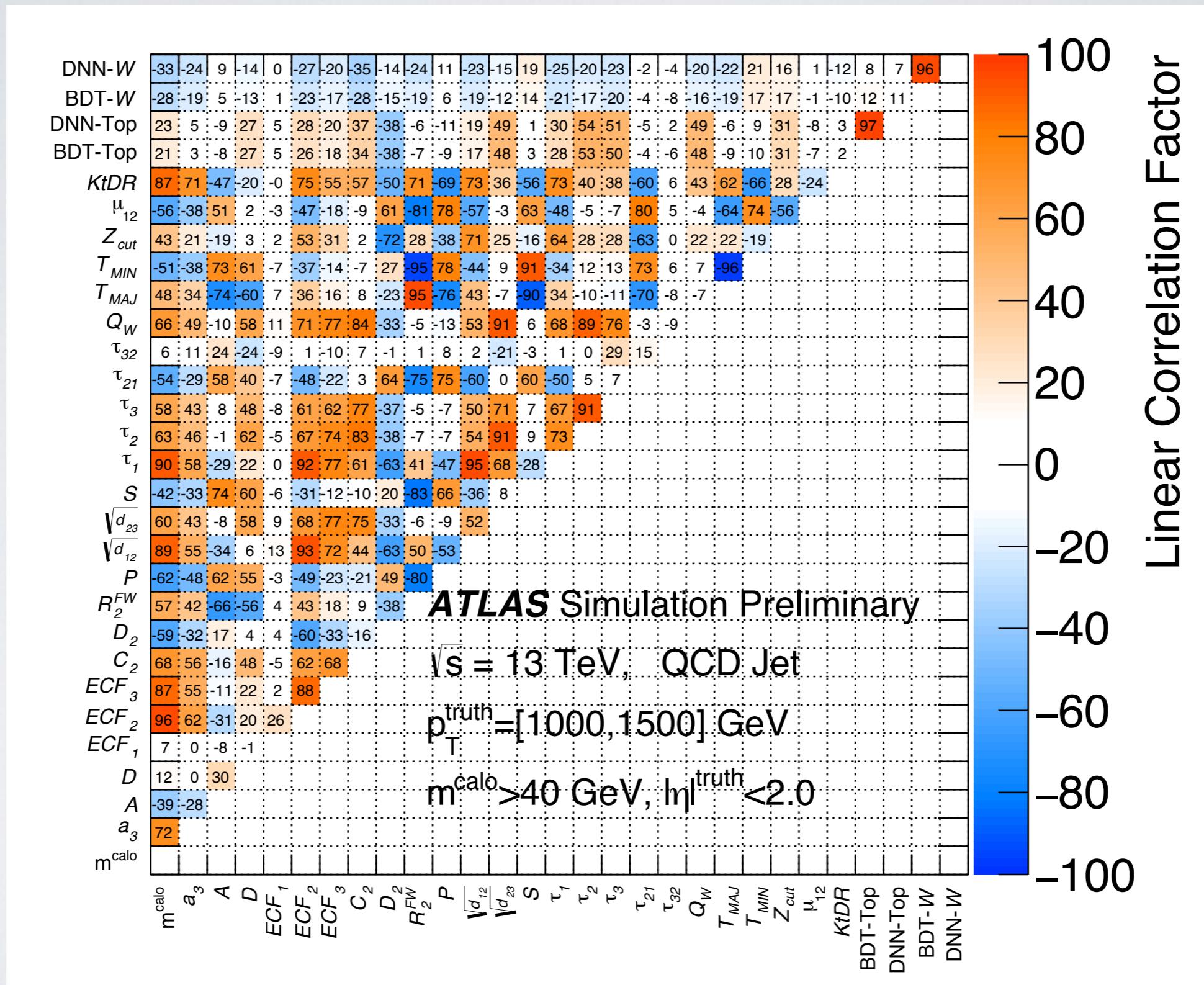


## Top Tagging

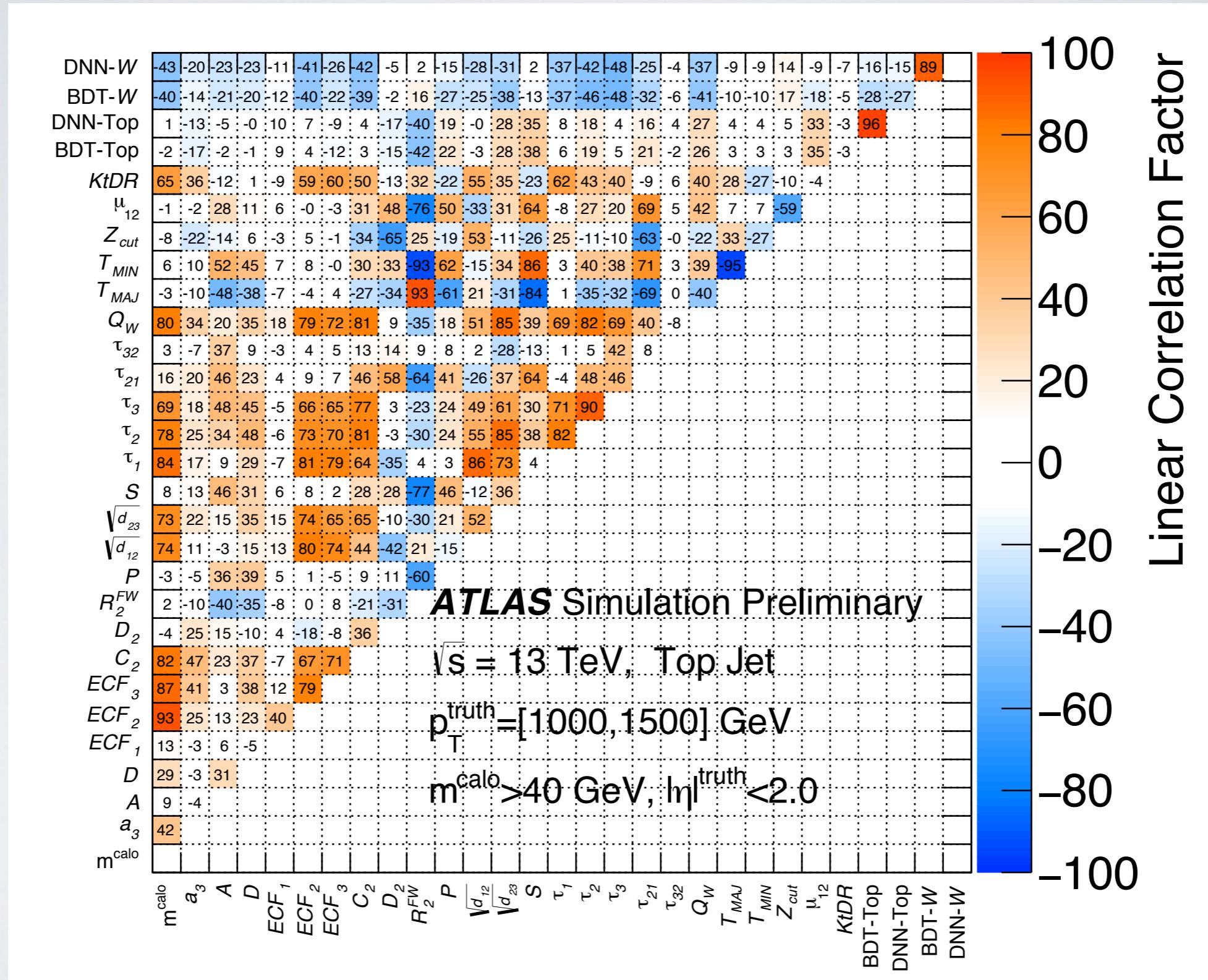


Strong mass shaping is expected as several substructure variables are highly correlated with the jet mass

# DISCRIMINANT CORRELATIONS



# DISCRIMINANT CORRELATIONS



# CONCLUSIONS

- A preliminary look at the use of BDTs and DNNs in ATLAS for W and top tagging using substructure variables
- The studied BDTs and DNNs outperform the respective reference taggers
  - Fixed mass cut & single substructure variable (W:  $D_2$ , top:  $\tau_{32}$ )
- The performance of the BDT and DNN taggers is rather similar

# THANK YOU!

# BACKUP

# SAMPLES

- **Jet Collection:** Trimmed anti- $k_t$   $R = 1.0$  jets,  $R_{\text{sub}} = 0.2$ ,  $f_{\text{cut}} = 0.05$
- Signal: Fully contained tops , Fully contained Ws

# BDT SCANNED HYPER-PARAMETERS

- 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

# DNN INPUTS GROUPS

W Tagging

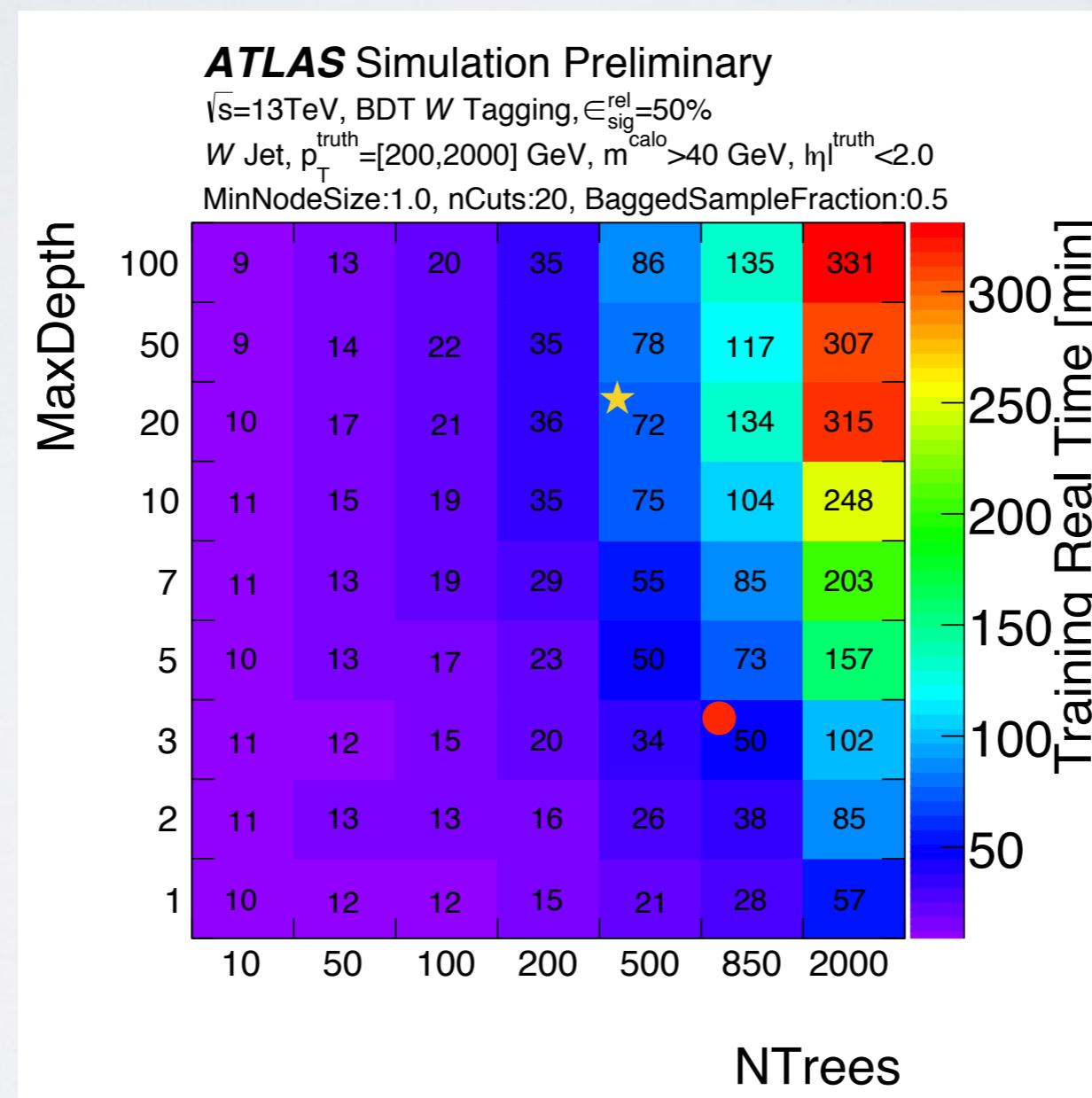
- Group1:  $C_2, D_2, \boldsymbol{\tau}_{21}, \sqrt{d_{12}}$
- Group2:  $C_2, D_2, \boldsymbol{\tau}_{21}, \sqrt{d_{12}}, R^{FW}_2, S, T_{min}, T_{maj}$
- Group3:  $ECF_1, ECF_2, ECF_3, \boldsymbol{\tau}_1, \boldsymbol{\tau}_2, R^{FW}_2, S, a_3, A, \sqrt{d_{12}}$
- Group4:  $ECF_1, ECF_2, ECF_3, \boldsymbol{\tau}_1, \boldsymbol{\tau}_2, R^{FW}_2, S, a_3, A, T_{min}, T_{maj}$
- Group5:  $C_2, D_2, \boldsymbol{\tau}_{21}, \sqrt{d_{12}}, S, ECF_1, ECF_2, ECF_3, \boldsymbol{\tau}_1, \boldsymbol{\tau}_2, R^{FW}_2, S, a_3, A, Z_{cut}, KtDR, \mu_{12}, \mathcal{D}$
- Group6:  $C_2, D_2, \boldsymbol{\tau}_{21}, \sqrt{d_{12}}, S, ECF_1, ECF_2, ECF_3, \boldsymbol{\tau}_1, \boldsymbol{\tau}_2, R^{FW}_2, S, a_3, A, Z_{cut}, KtDR, \mu_{12}, T_{min}, T_{maj}, \mathcal{D}$

Top Tagging

- Group1:  $\boldsymbol{\tau}_{21}, \boldsymbol{\tau}_{32}, \sqrt{d_{12}}, \sqrt{d_{23}}, Q_w$
- Group2:  $\boldsymbol{\tau}_1, \boldsymbol{\tau}_2, \boldsymbol{\tau}_3, \sqrt{d_{12}}, \sqrt{d_{23}}, Q_w$
- Group3:  $\boldsymbol{\tau}_1, \boldsymbol{\tau}_2, \boldsymbol{\tau}_3, \sqrt{d_{12}}, \sqrt{d_{23}}, Q_w, \boldsymbol{\tau}_{21}, \boldsymbol{\tau}_{32}$
- Group4:  $ECF_1, ECF_2, ECF_3, \boldsymbol{\tau}_1, \boldsymbol{\tau}_2, \boldsymbol{\tau}_3, \sqrt{d_{12}}, \sqrt{d_{23}}, Q_w$
- Group5:  $\boldsymbol{\tau}_1, \boldsymbol{\tau}_2, \boldsymbol{\tau}_3, \boldsymbol{\tau}_{21}, \boldsymbol{\tau}_{32}, \sqrt{d_{12}}, \sqrt{d_{23}}, Q_w, C_2, D_2$
- Group6:  $\boldsymbol{\tau}_1, \boldsymbol{\tau}_2, \boldsymbol{\tau}_3, \boldsymbol{\tau}_{21}, \boldsymbol{\tau}_{32}, ECF_1, ECF_2, ECF_3, C_2, D_2, \sqrt{d_{12}}, \sqrt{d_{23}}, Q_w$

# BDT TRAINING - HYPER-PARAMETER OPTIMIZATION

- Star = Optimum settings, Circle = TMVA default
- Machine architecture: Intel® Xeon® CPU E5-2680 v3, 2.50 GHz with 4 GB memory per processor



# BDT & DNN CHOSEN VARIABLES

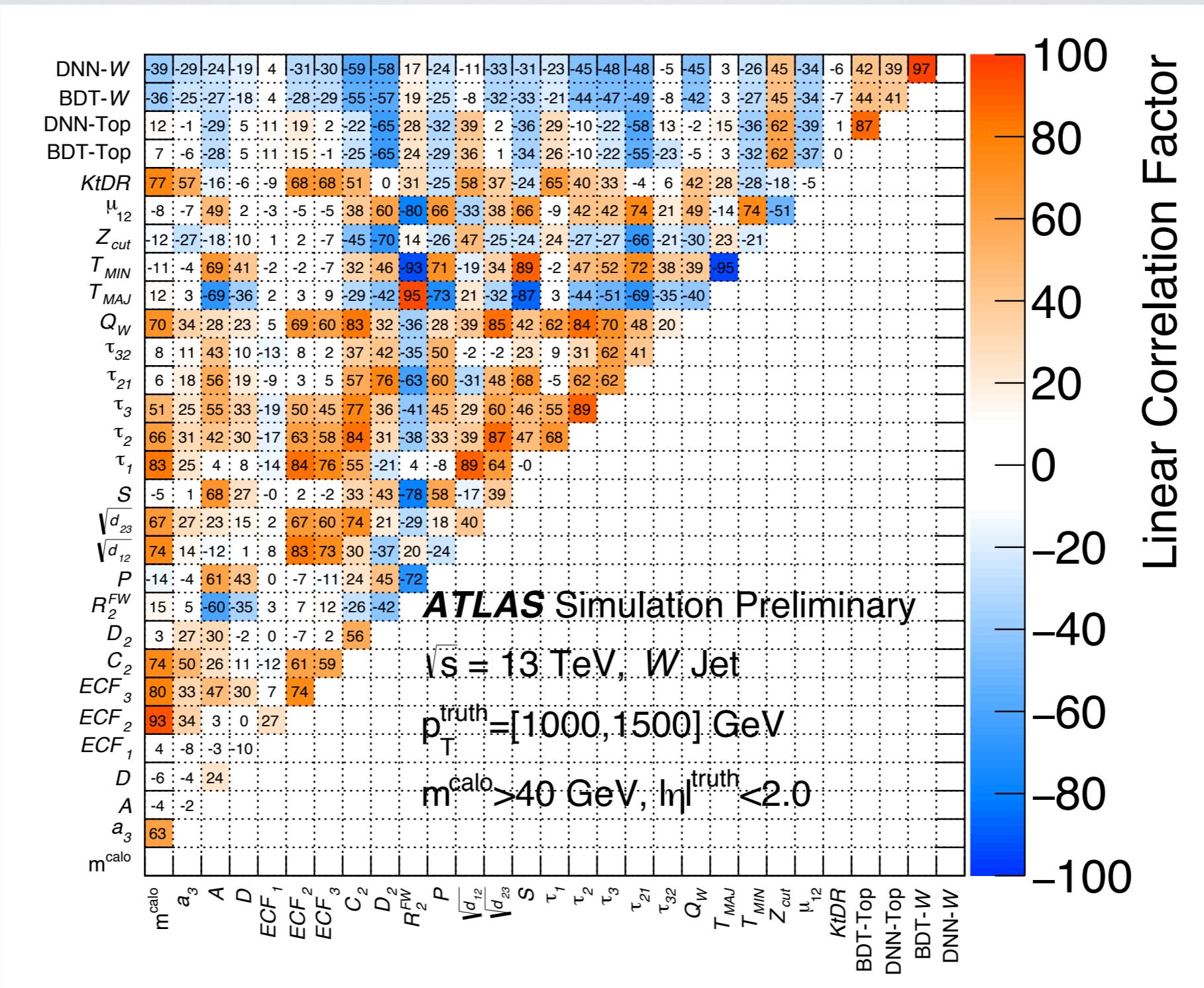
## W Tagging

- BDT set:  $D_2, \tau_{21}, S, ECF_2, ECF_3, \tau_1, R^{FW}_{2,S}, a_3, A, KtDR$
- DNN set:  $C_2, D_2, \tau_{21}, \sqrt{d}_{12}, S, ECF_1, ECF_2, ECF_3, \tau_1, \tau_2, R^{FW}_{2,S}, a_3, A, Z_{cut}, KtDR, \mu_{12}, \mathcal{D}$

## Top Tagging

- BDT set:  $\tau_2, \tau_{21}, \tau_{32}, ECF_1, ECF_3, C_2, D_2, \sqrt{d}_{12}, \sqrt{d}_{23}, Q_W$
- DNN set:  $\tau_1, \tau_2, \tau_3, \tau_{21}, \tau_{32}, ECF_1, ECF_2, ECF_3, C_2, D_2, \sqrt{d}_{12}, \sqrt{d}_{23}, Q_W$

# DISCRIMINANT CORRELATIONS



# DISCRIMINANT CORRELATIONS

