IDENTIFICATION OF BOOSTED W BOSONS AND TOP QUARKS USING MACHINE LEARNING IN ATLAS

Ece Akıllı
Université de Genève

Machine learning for phenomenology workshop
April 4, 2018

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

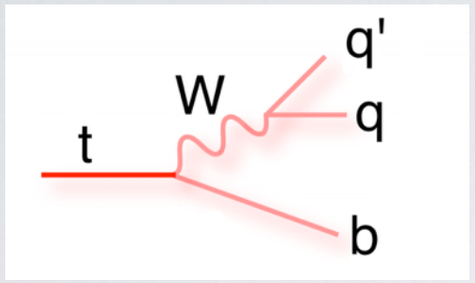
3. Conclusions

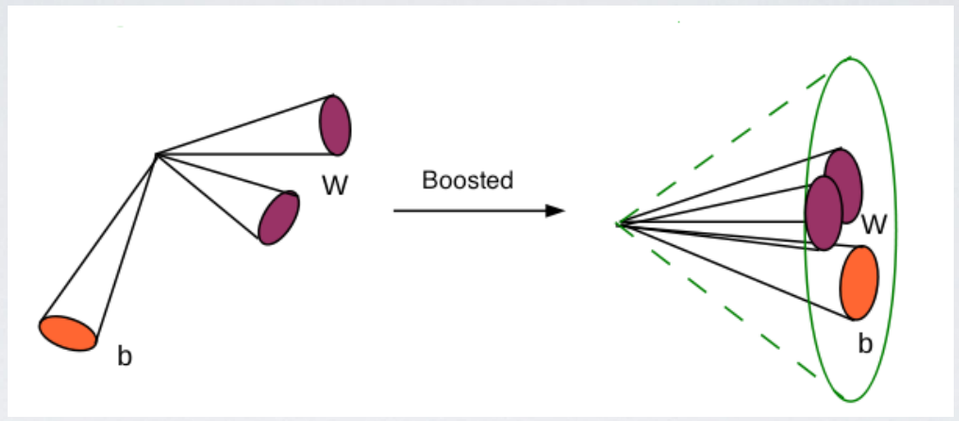
Two sets of results

MC-based early studies: ATL-PHYS-PUB-2017-004

Further studies and performance in data: ATLAS-CONF-2017-064

W-BOSON AND TOP-QUARK TAGGING IN ATLAS

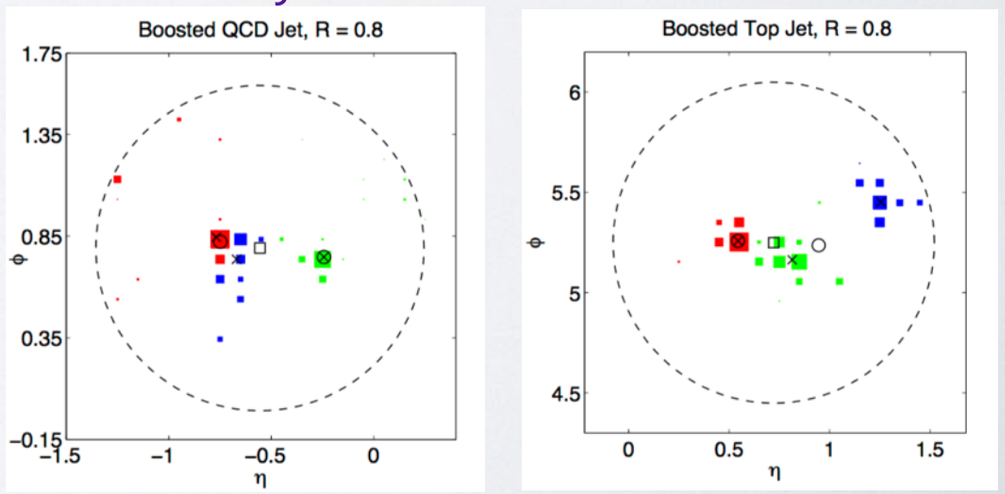




Top-Quark Decay

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

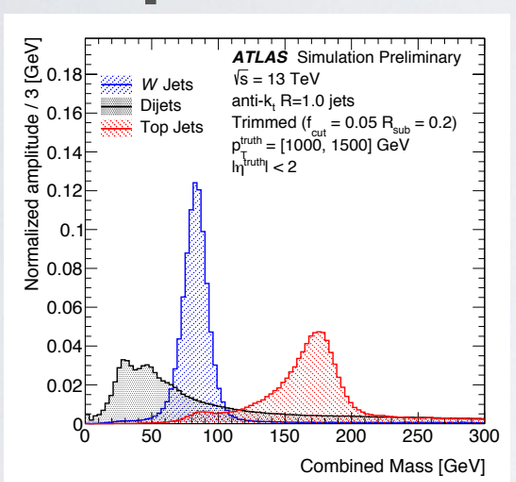
Jet Substructure

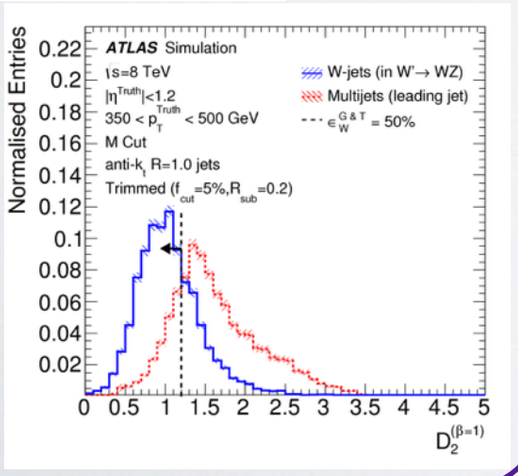


APPLICATION OF BDTS AND DNNS TO W AND TOP TAGGING USING HIGH-LEVEL FEATURES

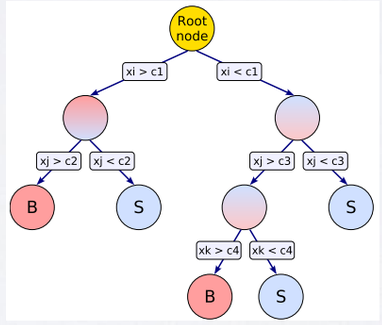
- Construct a classifier by using **substructure variables**
  - Combine available information to obtain good discrimination
- Two Machine Learning (ML) techniques in parallel
  - Boosted Decision Tree (BDT)
  - Deep Neural Network (DNN)
- Train binary classifiers: W/top vs Dijet (4 classifiers)

Input variables



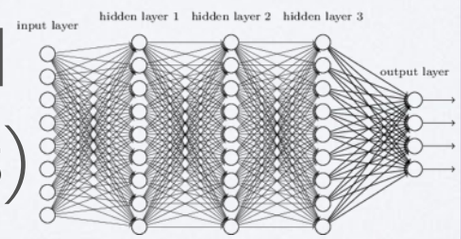


BDT (TMVA)

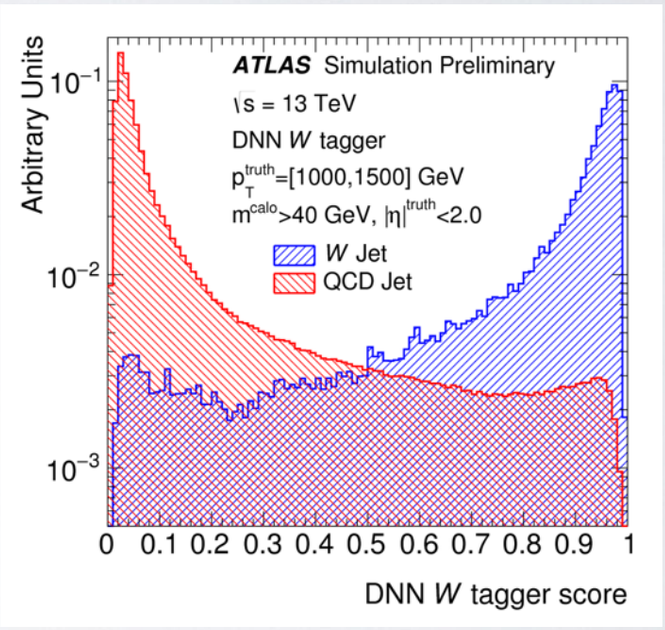


or

DNN (Keras)



Binary classifier



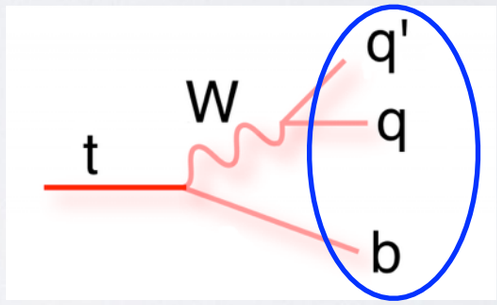
- BDT W
- BDT Top
- DNN W
- DNN Top

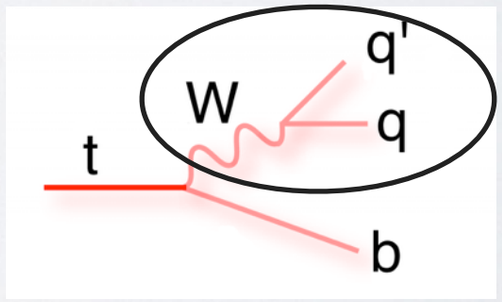
STRATEGY

- Split Monte Carlo (MC) simulation samples in training and testing sets

- Optimize BDT, DNN using MC
  - Set of inputs
  - Architecture and training hyper-parameters

- Performance comparison of ML taggers with reference taggers in MC
  - Current taggers in ATLAS
    - 2-variable taggers
    - HEPTopTagger (Only for tops)
    - Shower Deconstruction (Only for tops)
  - BDT tagger
  - DNN tagger





- Study the performance in data

OPTIMIZATION AND PERFORMANCE STUDIES IN MC

ATL-PHYS-PUB-2017-004

SAMPLES

Training & Testing Samples

- Split signal and background in training and testing samples

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

Testing Event Weights: Signal samples are weighted to match background (dijet) truth p_T distribution

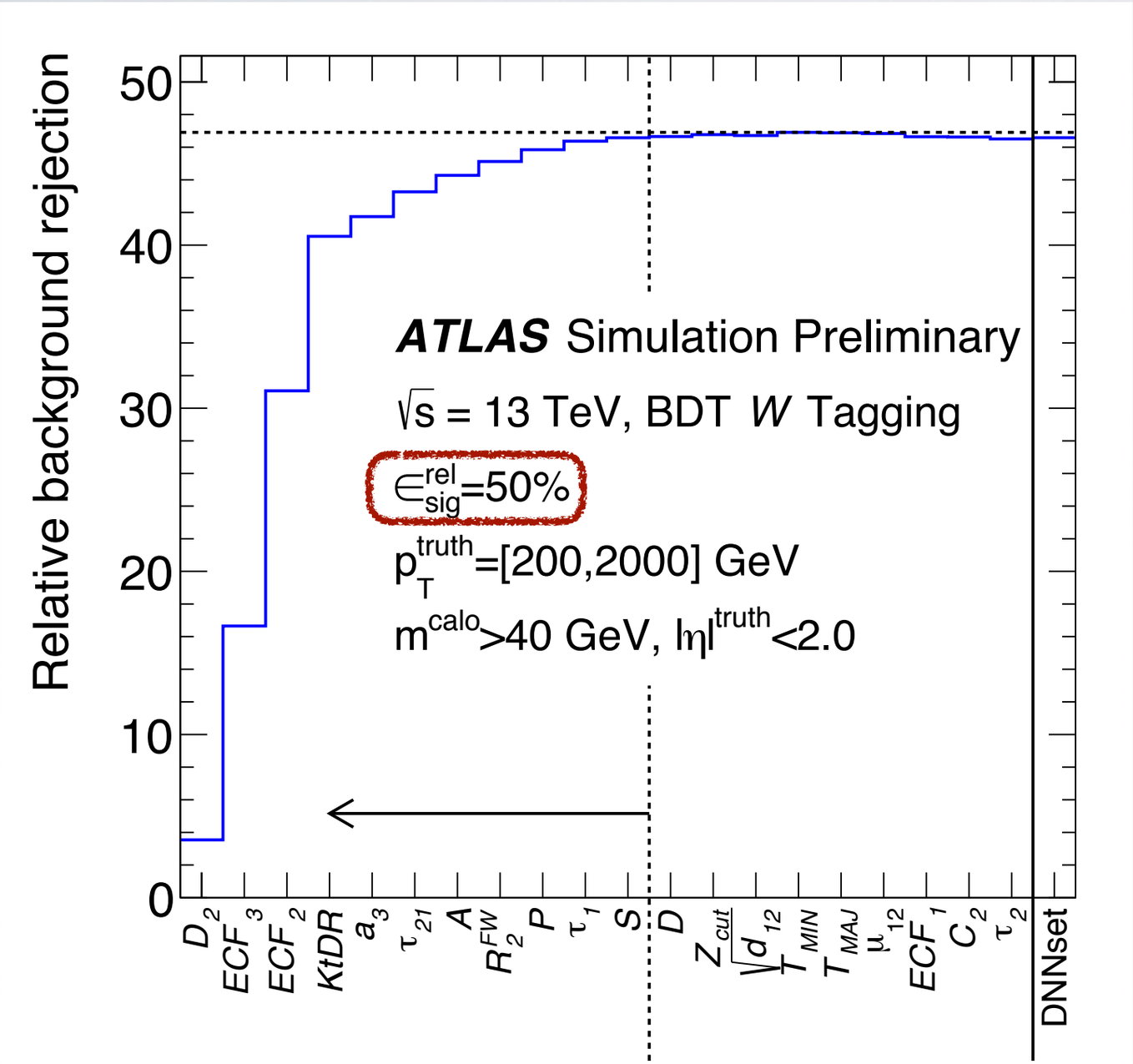
W Tagging

- p_T = [200, 2000] GeV, η = [-2,2]
- # Training signal jets = 7×10^5
- # Training light jets = 7×10^5

Top Tagging

- p_T = [350, 2000] GeV, η = [-2,2]
- # Training signal jets = 10^6
- # Training light jets = 10^6

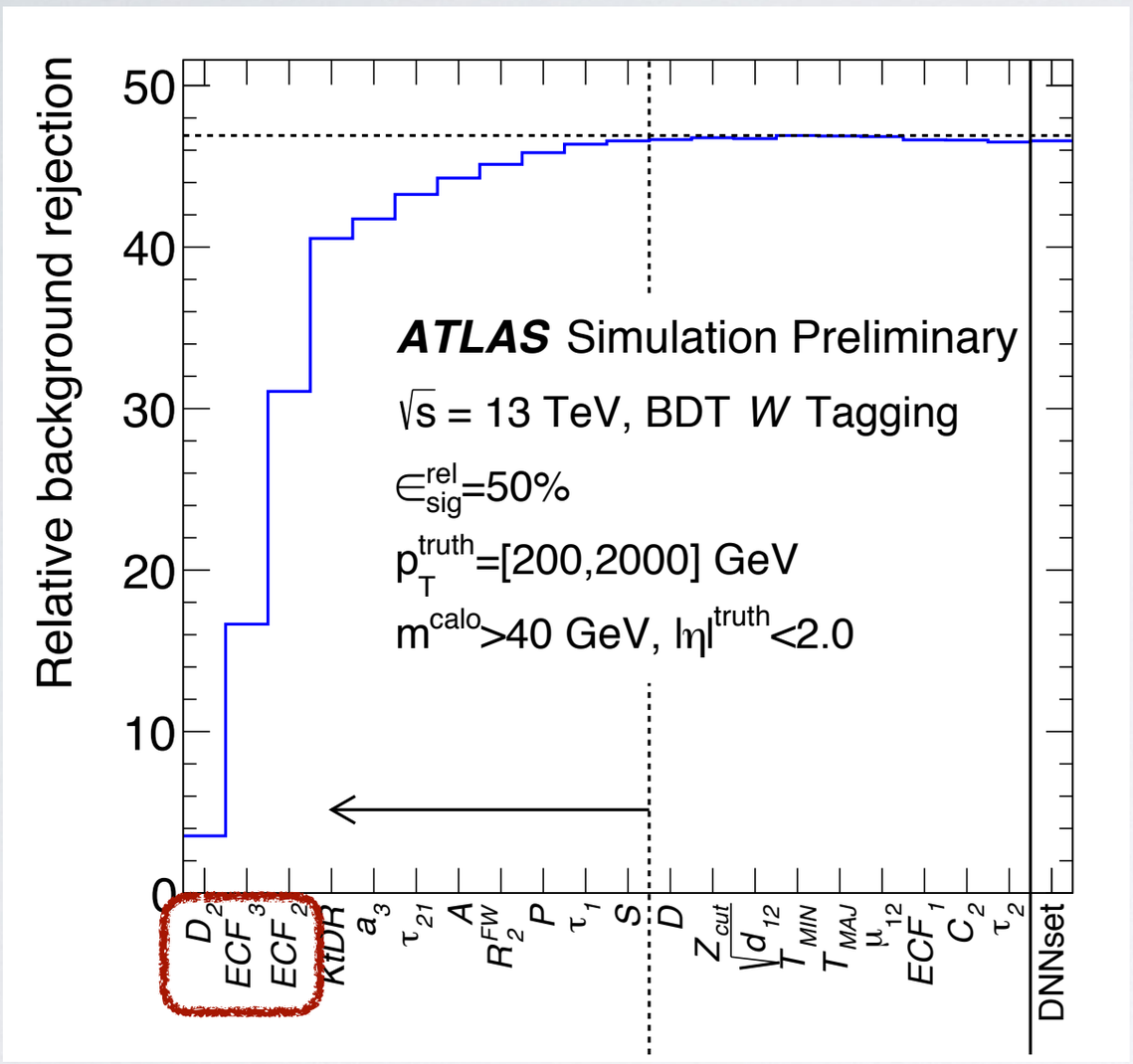
BDT TRAINING - INPUTS OPTIMIZATION



- Add variables in order of importance (improvement in rejection)
- Use a flat p_T spectrum (evaluation)
- Saturation of rejection

BDT TRAINING - INPUTS OPTIMIZATION

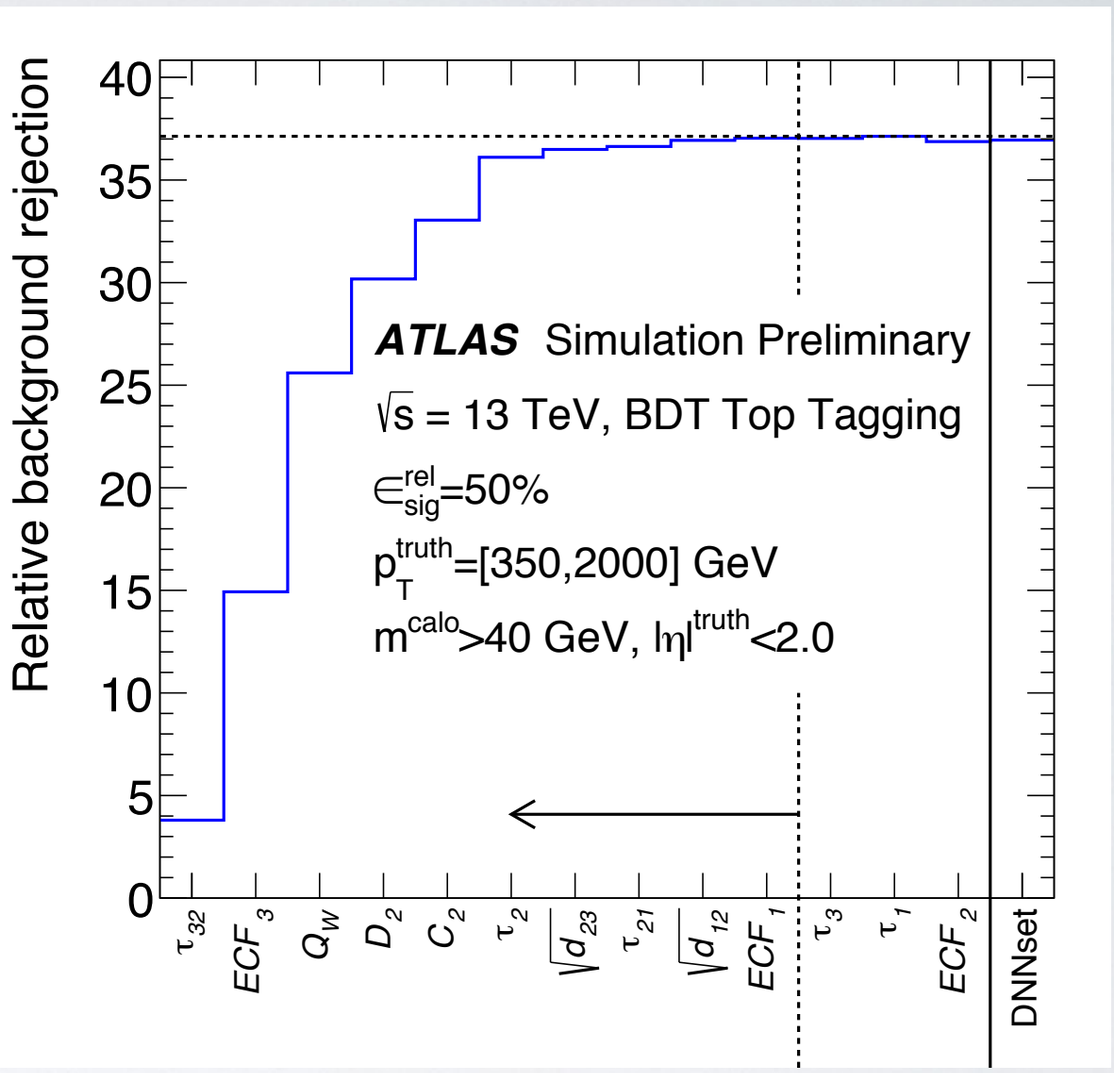
W Tagging



11 variables

Learning features from ECF beyond D2

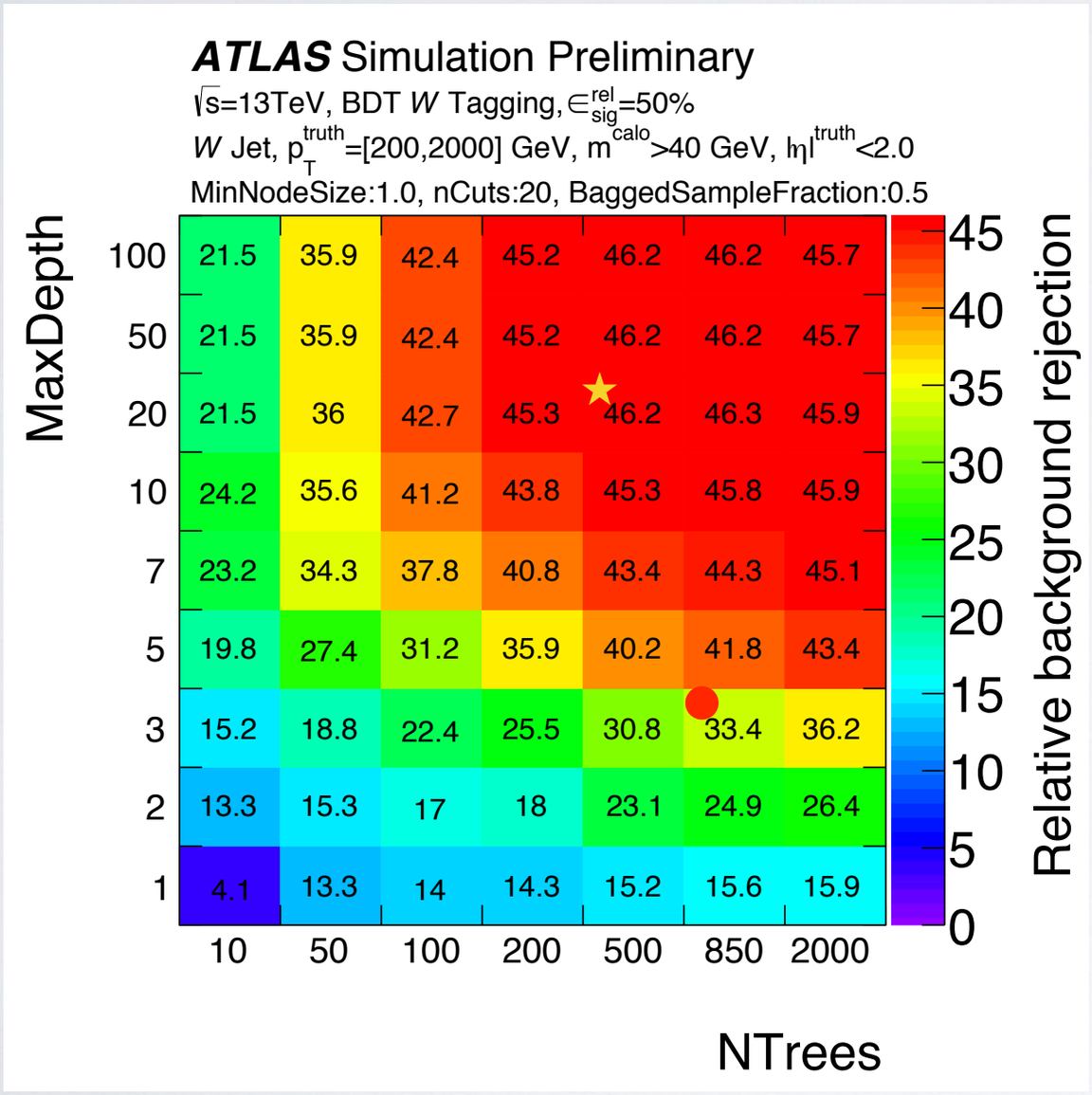
Top Tagging

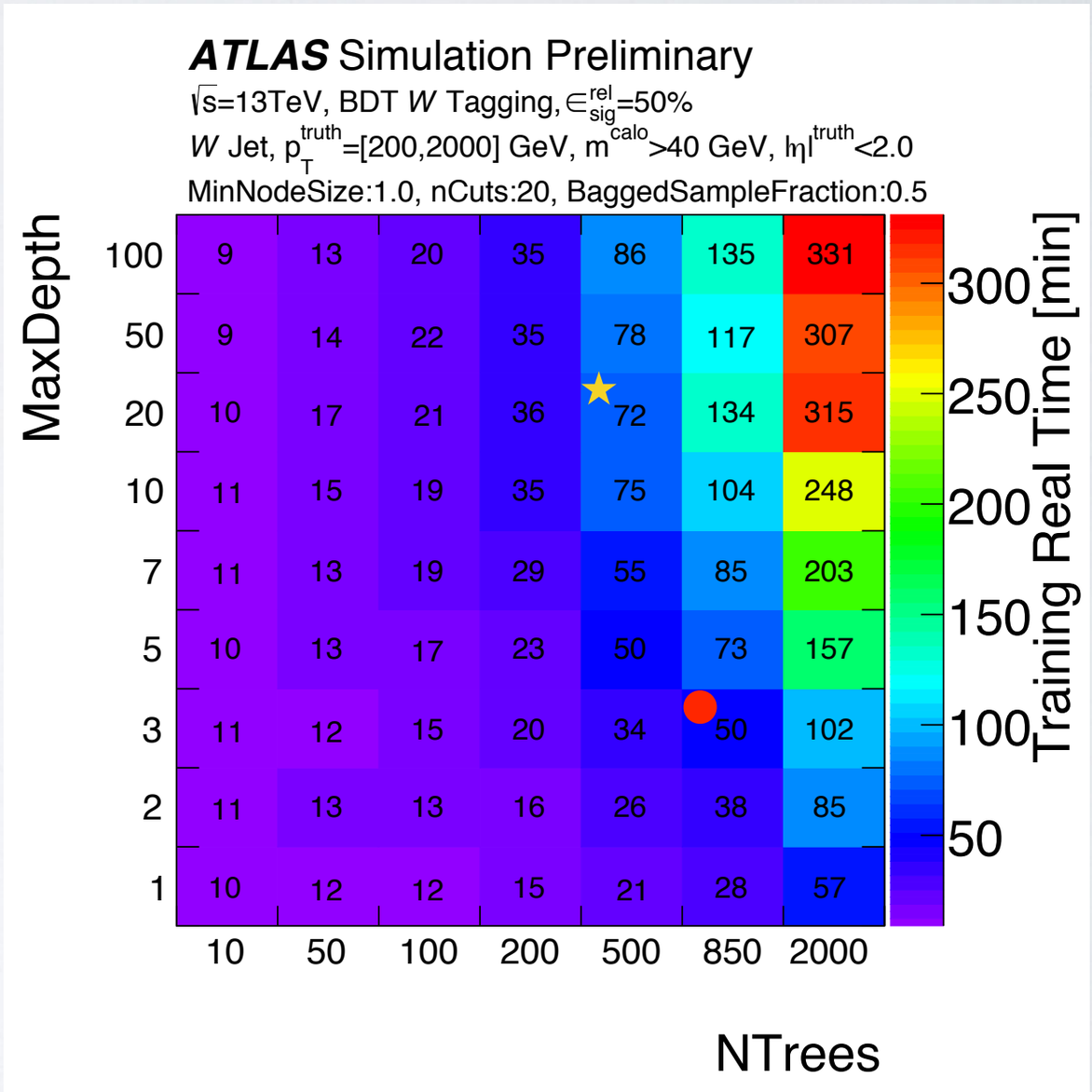


10 variables

BDT TRAINING - HYPER-PARAMETER OPTIMIZATION

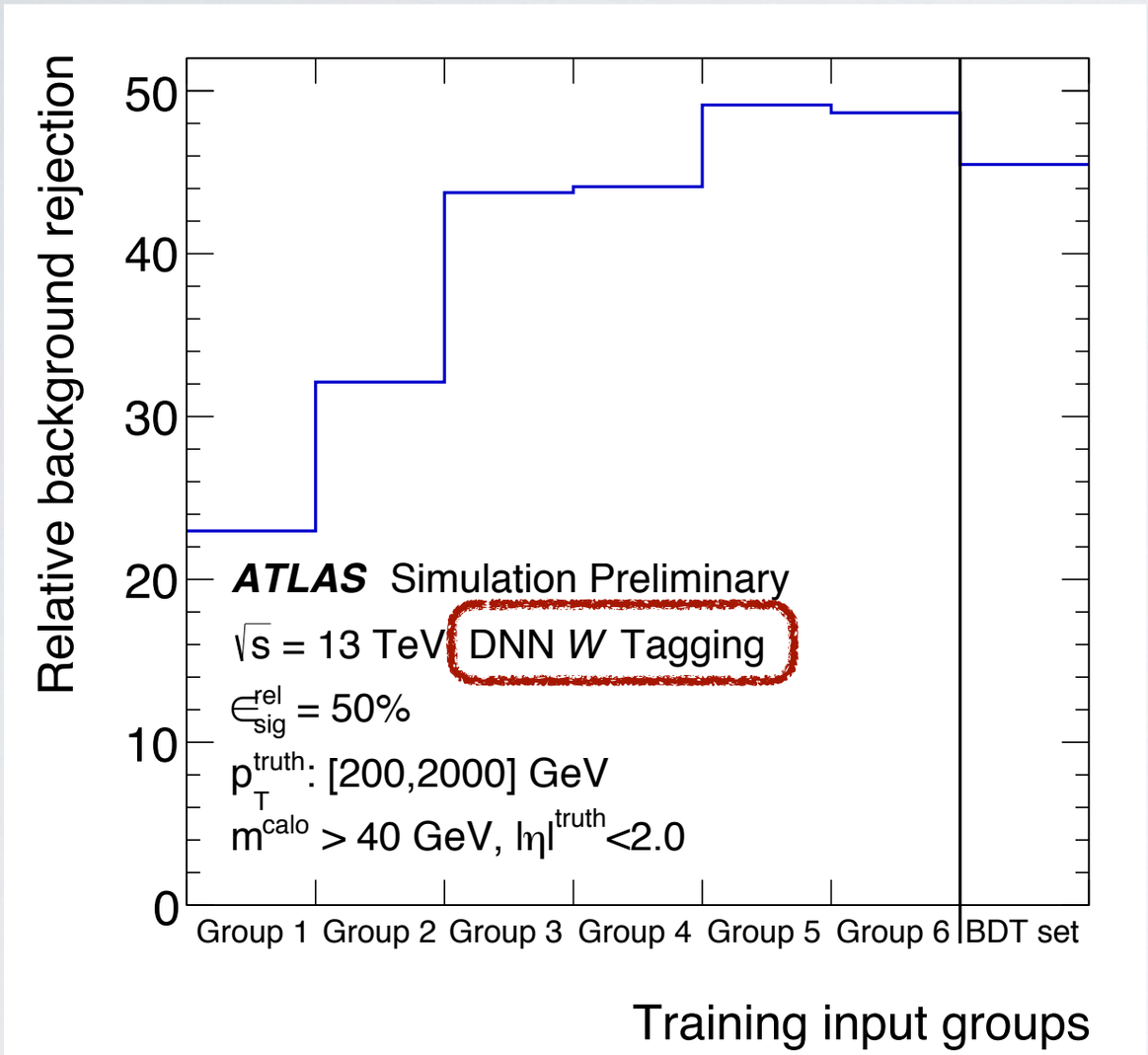
- Performed a parameter scan over many variables, most significant differences observed for NTrees and MaxDepth
- Shown for W, similar for top
- Star = Optimum settings found, Circle = TMVA default





DNN TRAINING - INPUTS OPTIMIZATION

W Tagging

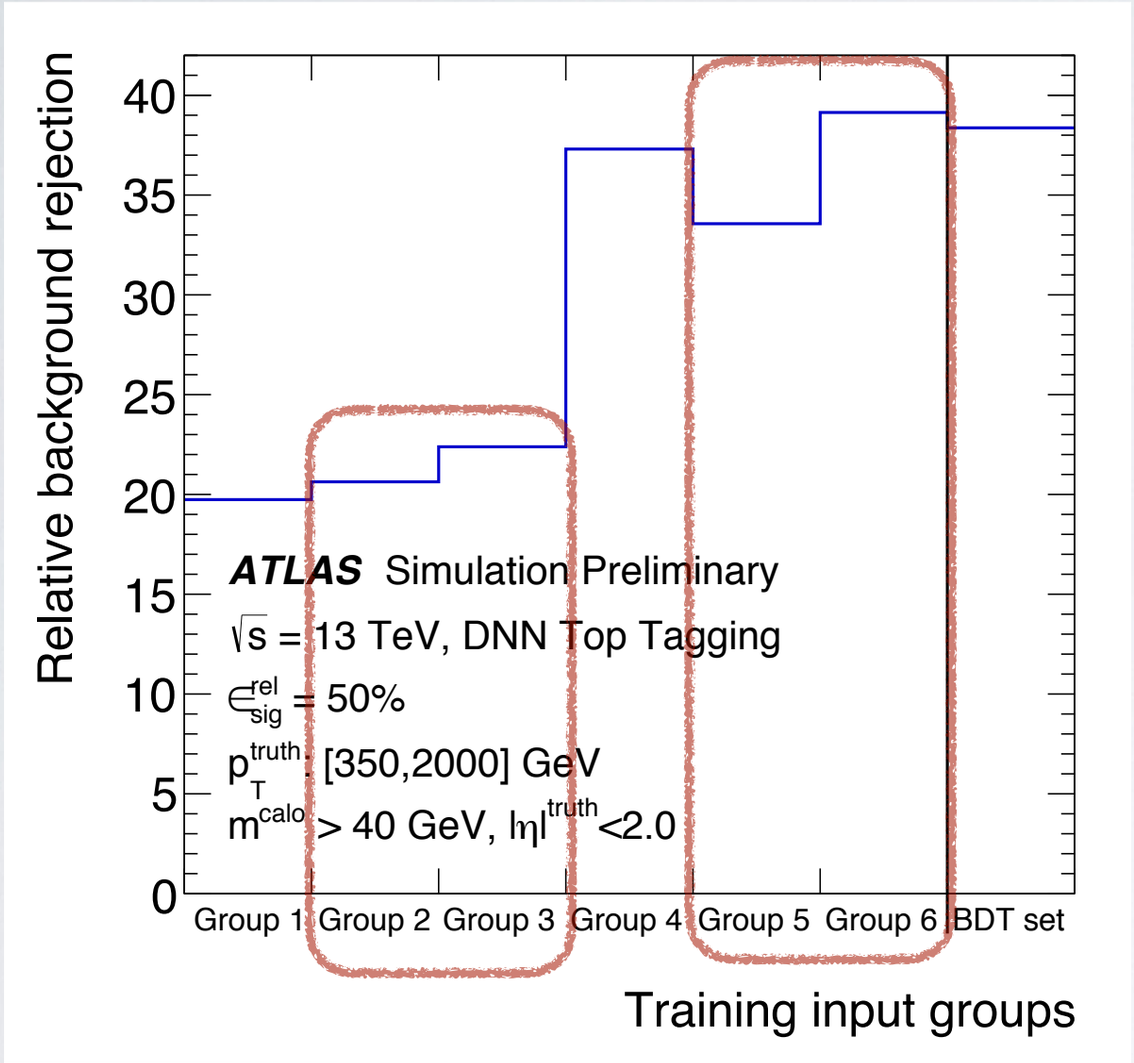


Group 5 with 18 variables

- Study different groups of input variables
- Groups are defined by varying features
  - the physical information they provide (pronginess, scale…)
  - if the observable is defined as a function of the other observables.
    Example: D2 vs ECF
- Use a flat p_T spectrum (evaluation)
- Choose the set with the highest background rejection

DNN TRAINING - INPUTS OPTIMIZATION

Top Tagging

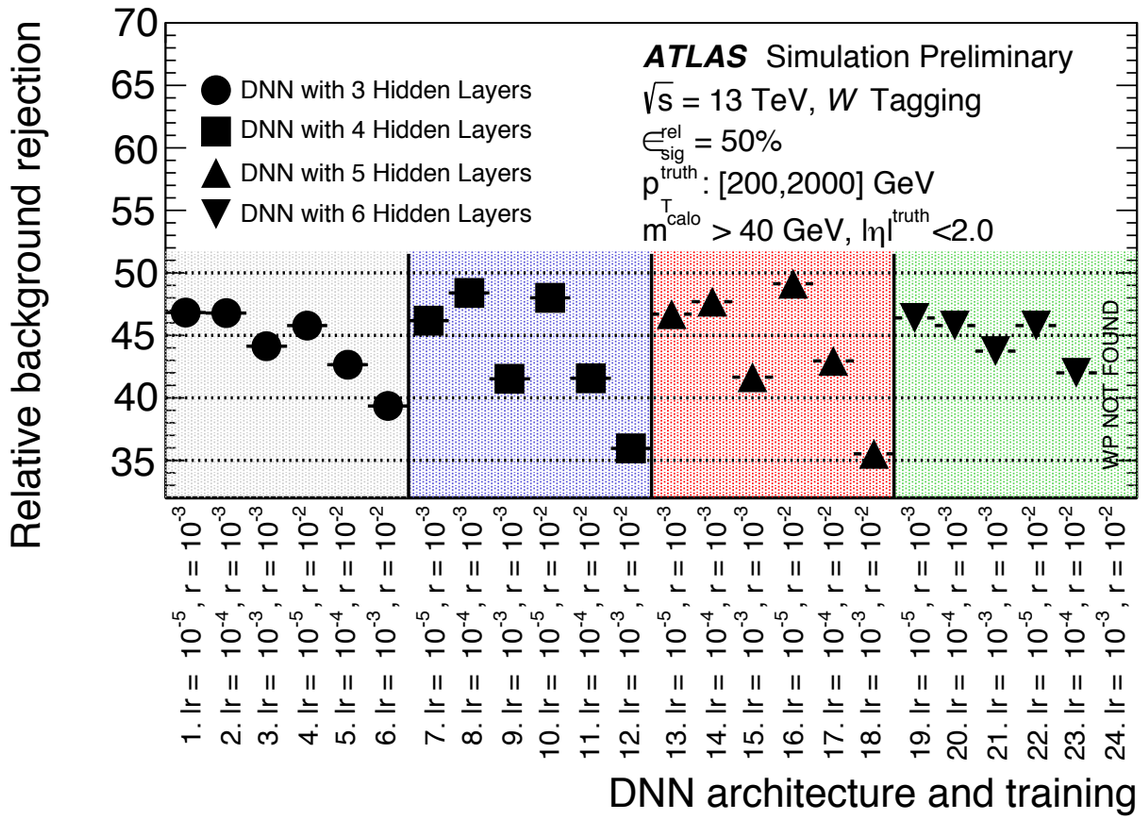


Observable	Top Tagging Observable Groups 1	2	3	4	5	6	7 (BDT)
ECF_1				o		o	o
ECF_2				o		o	
ECF_3				o		o	o
C_2					o	o	o
D_2					o	o	o
τ_1		o	o	o	o	o	
τ_2		o	o	o	o	o	o
τ_3		o	o	o	o	o	
τ_{21}	o		o		o	o	o
τ_{32}	o		o		o	o	o
$\sqrt{d_{12}}$	o	o	o	o	o	o	o
$\sqrt{d_{23}}$	o	o	o	o	o	o	o
Q_w	o	o	o	o	o	o	o

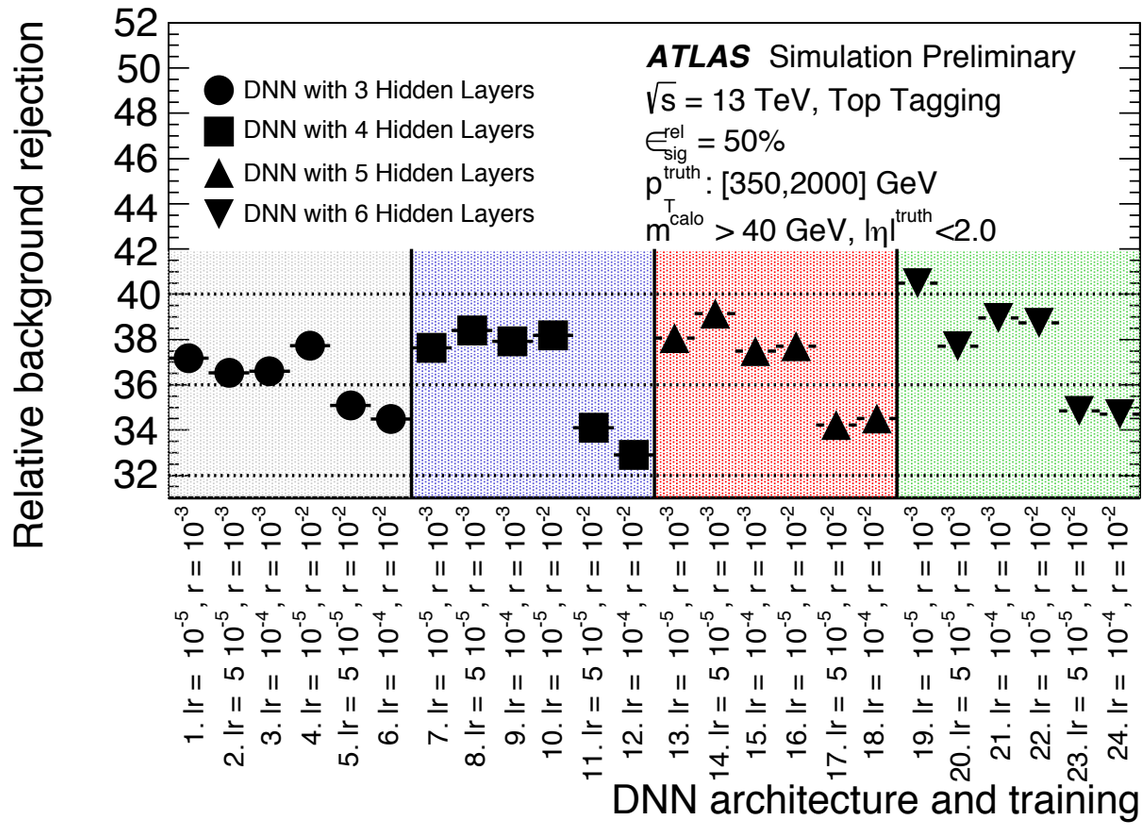
Group 6 with 13 variables

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 = 5×10^{-5} , L1 regularizer = 10^{-3} , Hidden layers = 5

BDT & DNN CHOSEN VARIABLES

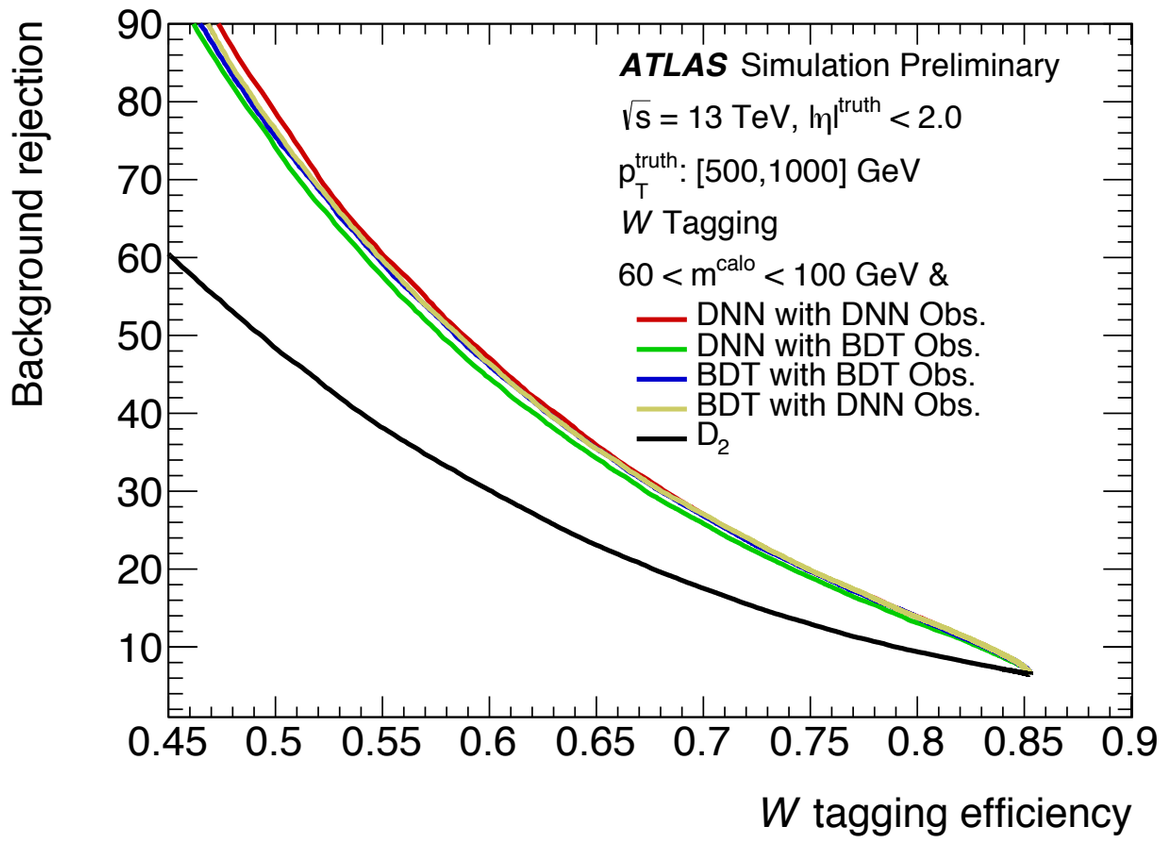
- BDT and DNN find different sets of inputs to be optimal

- 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

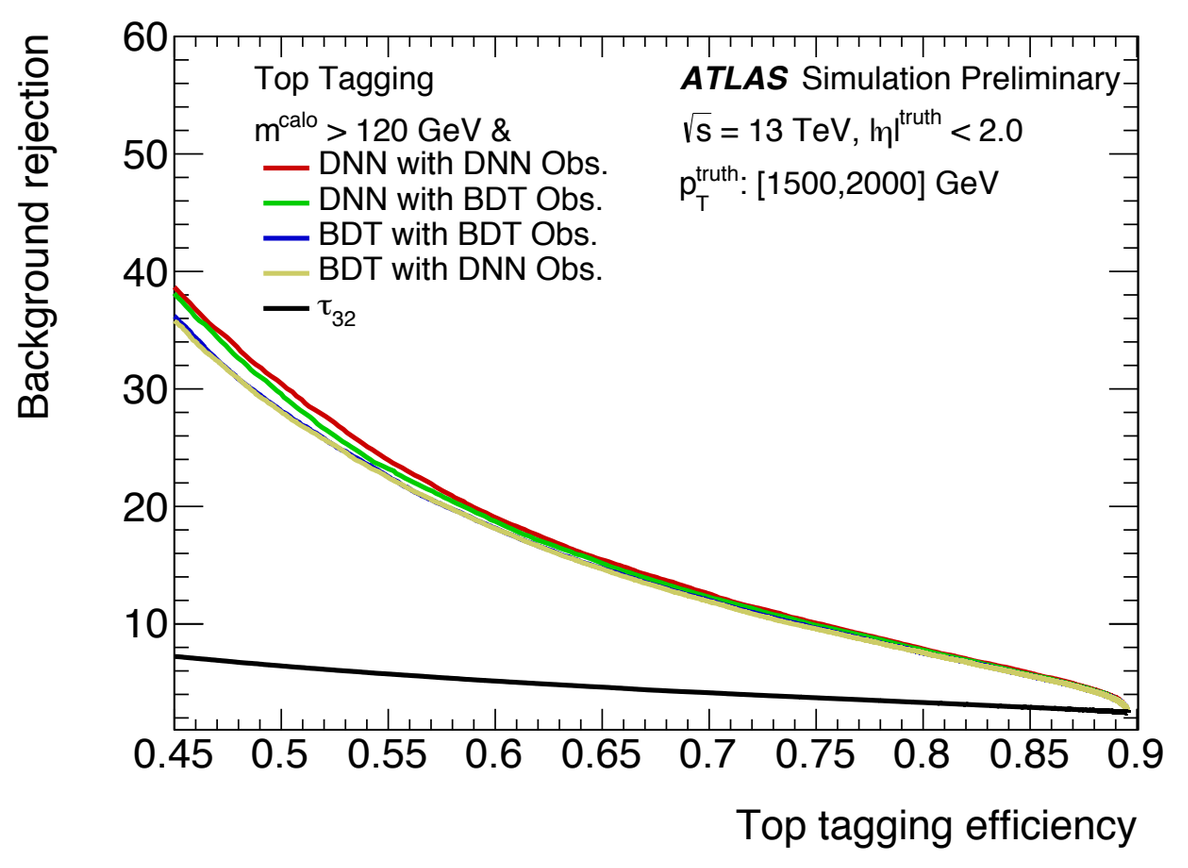
- Full list of variables for each case is in the backup

PERFORMANCE EVALUATION - ROC CURVES

W Tagging
p_T = [500, 1000] GeV



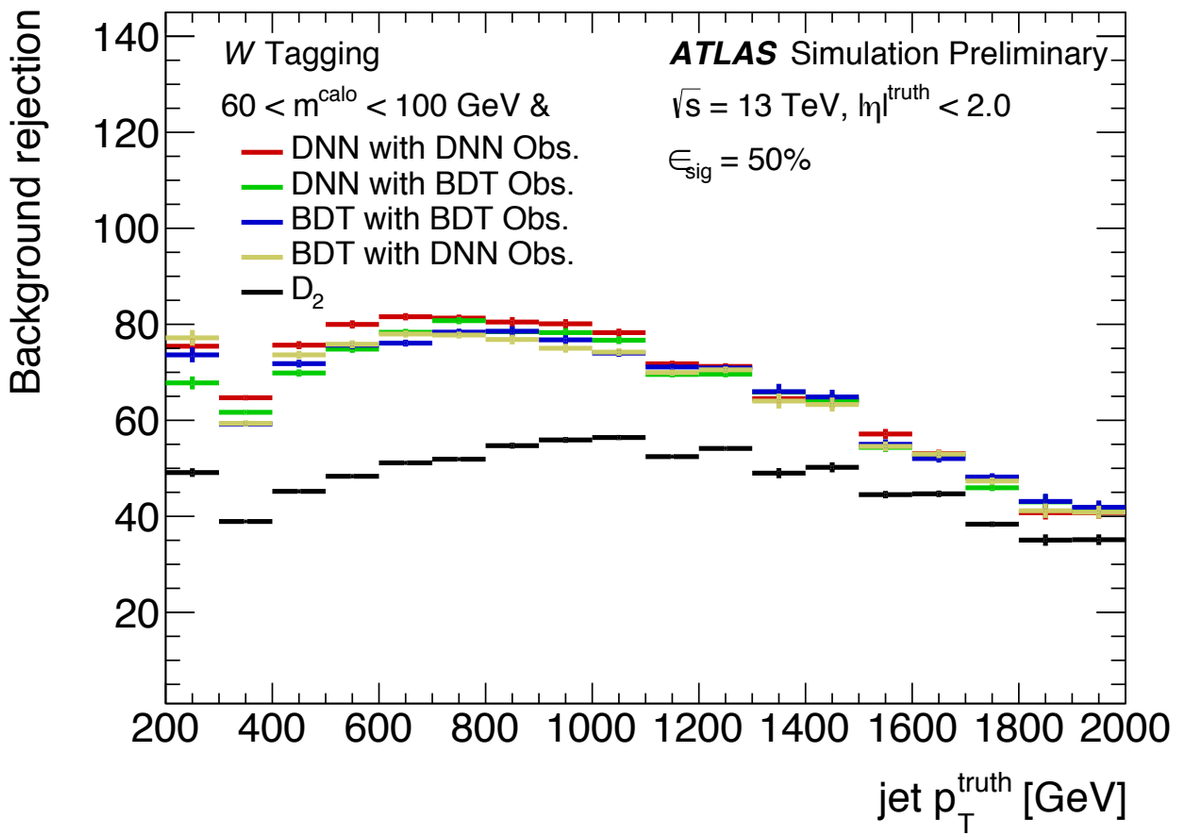
Top Tagging
p_T = [1500, 2000] GeV



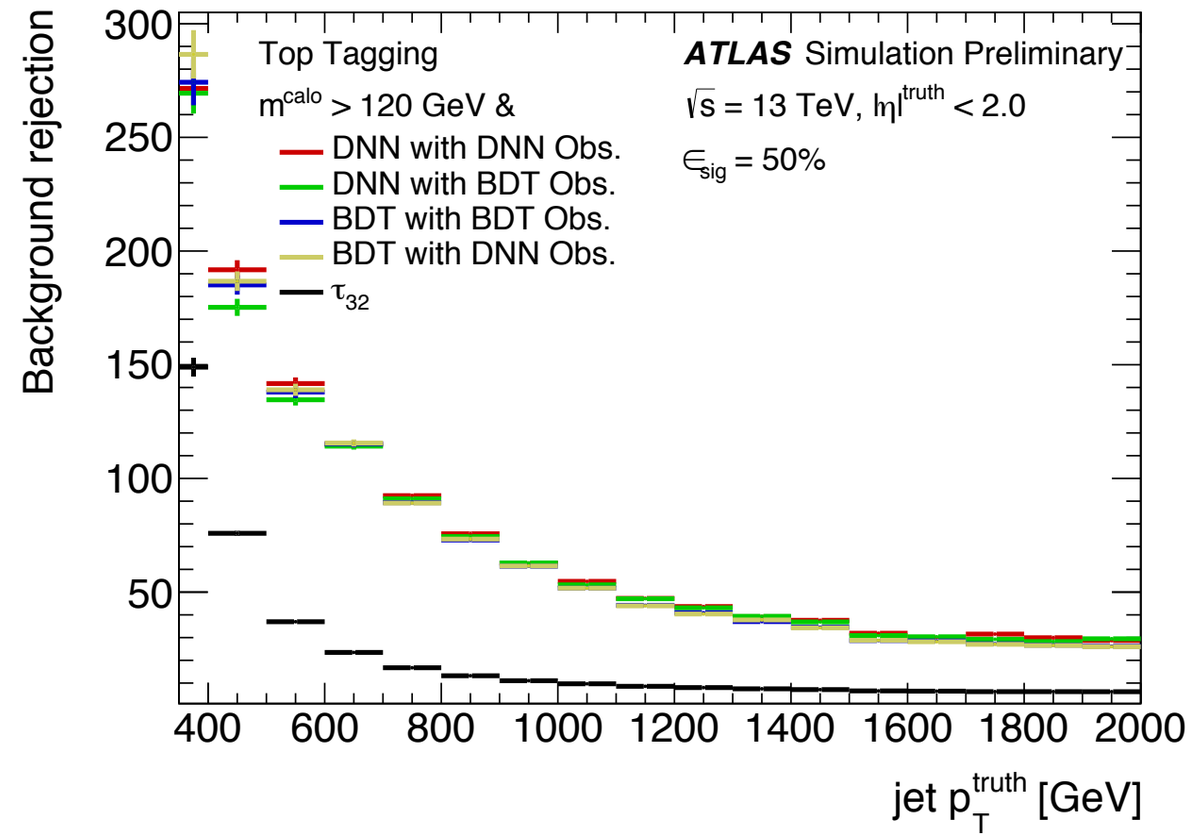
PERFORMANCE EVALUATION - WORKING POINTS

Background Rejection at 50% Fixed Efficiency WP

W Tagging



Top Tagging



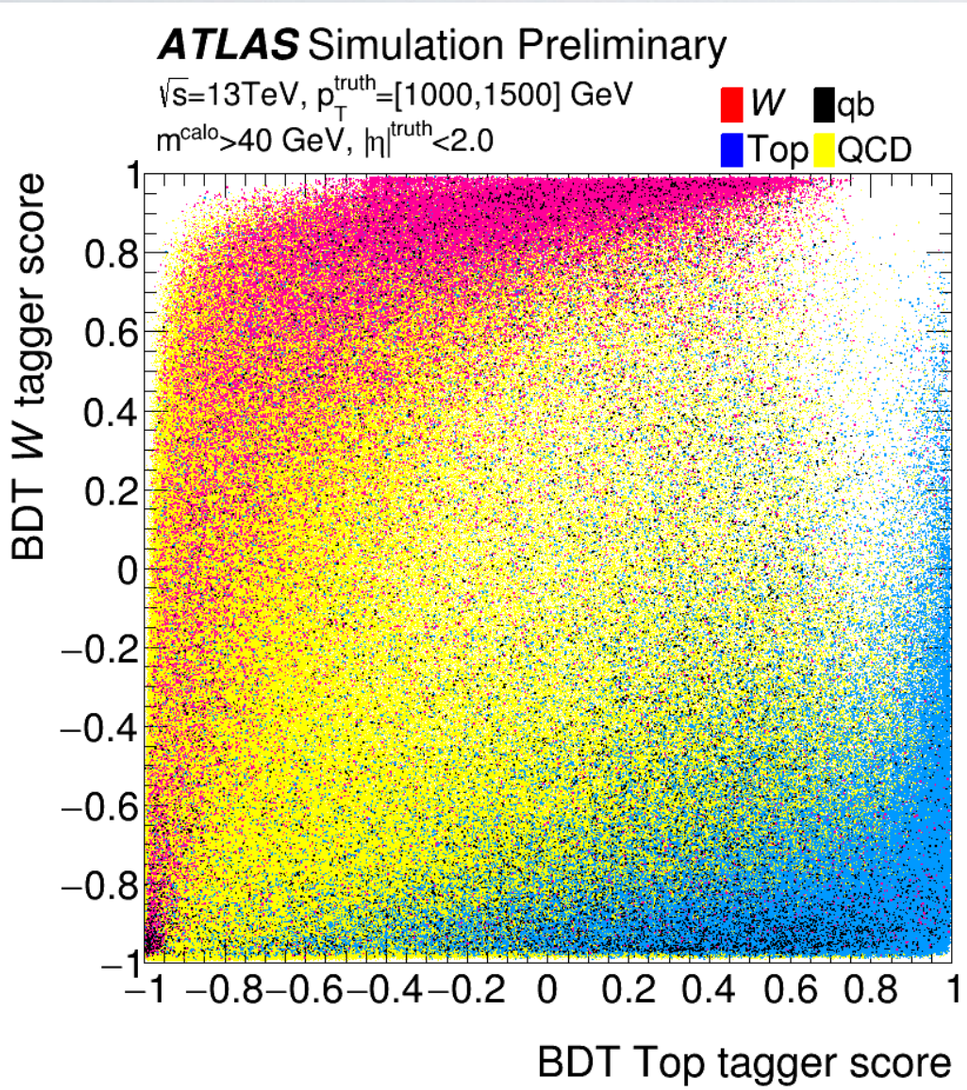
- Improvements observed for both W and top tagging
- Magnitude of impact differs for W and top tagging, but not the overall benefit of using a BDT or DNN

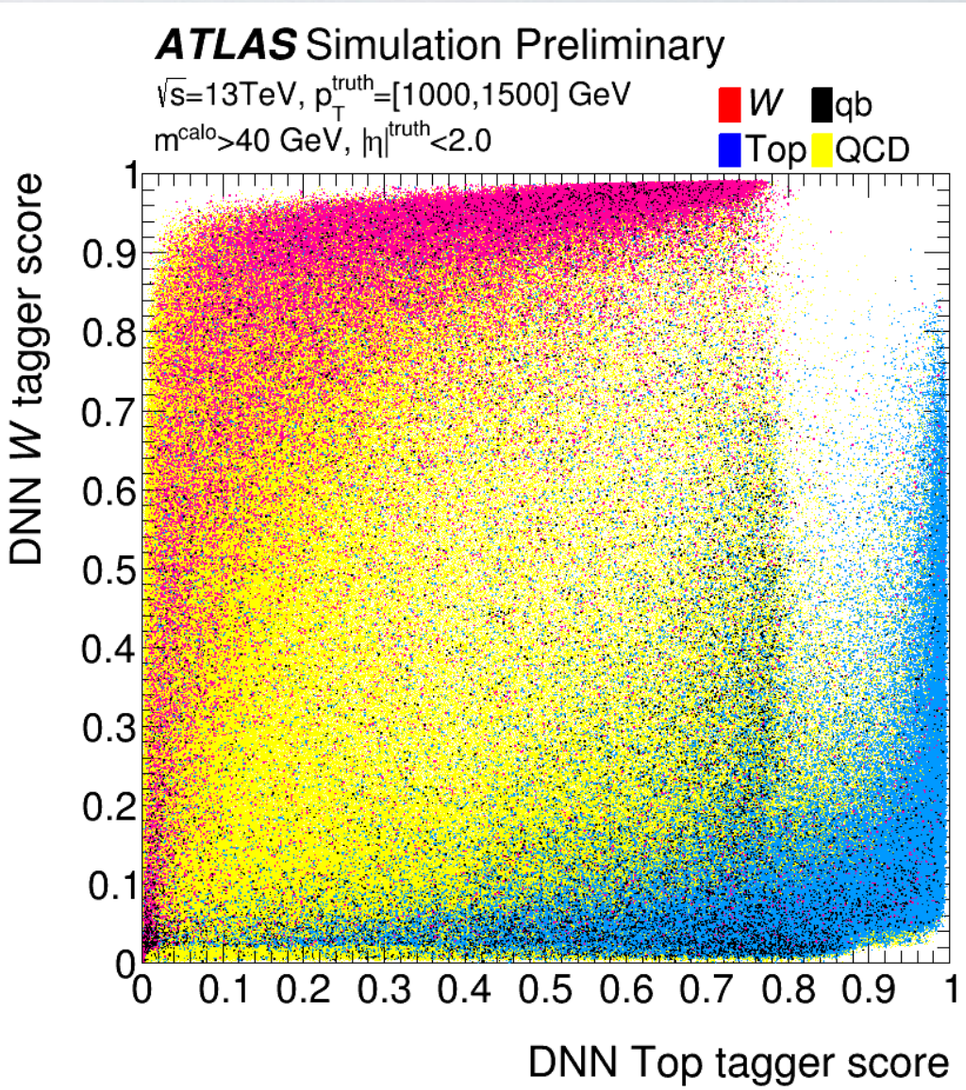
DISCRIMINANT CORRELATIONS

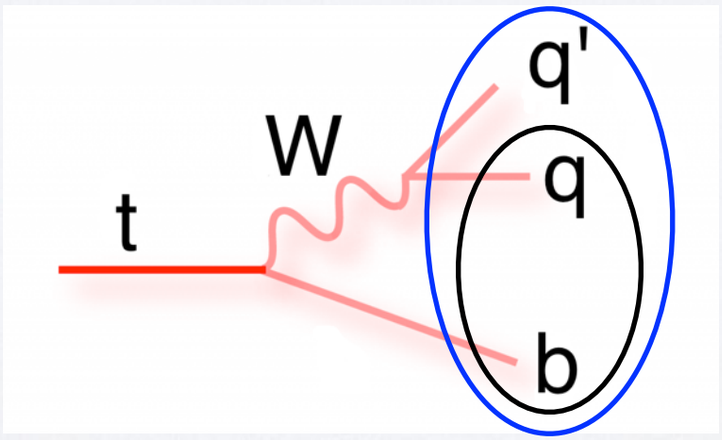
- ML techniques are expected to learn
  - linear correlations
  - non-linear correlations

- Study
  - correlations between W and top tagging outputs
  - impact on mass
  - linear correlations

DISCRIMINANT CORRELATIONS

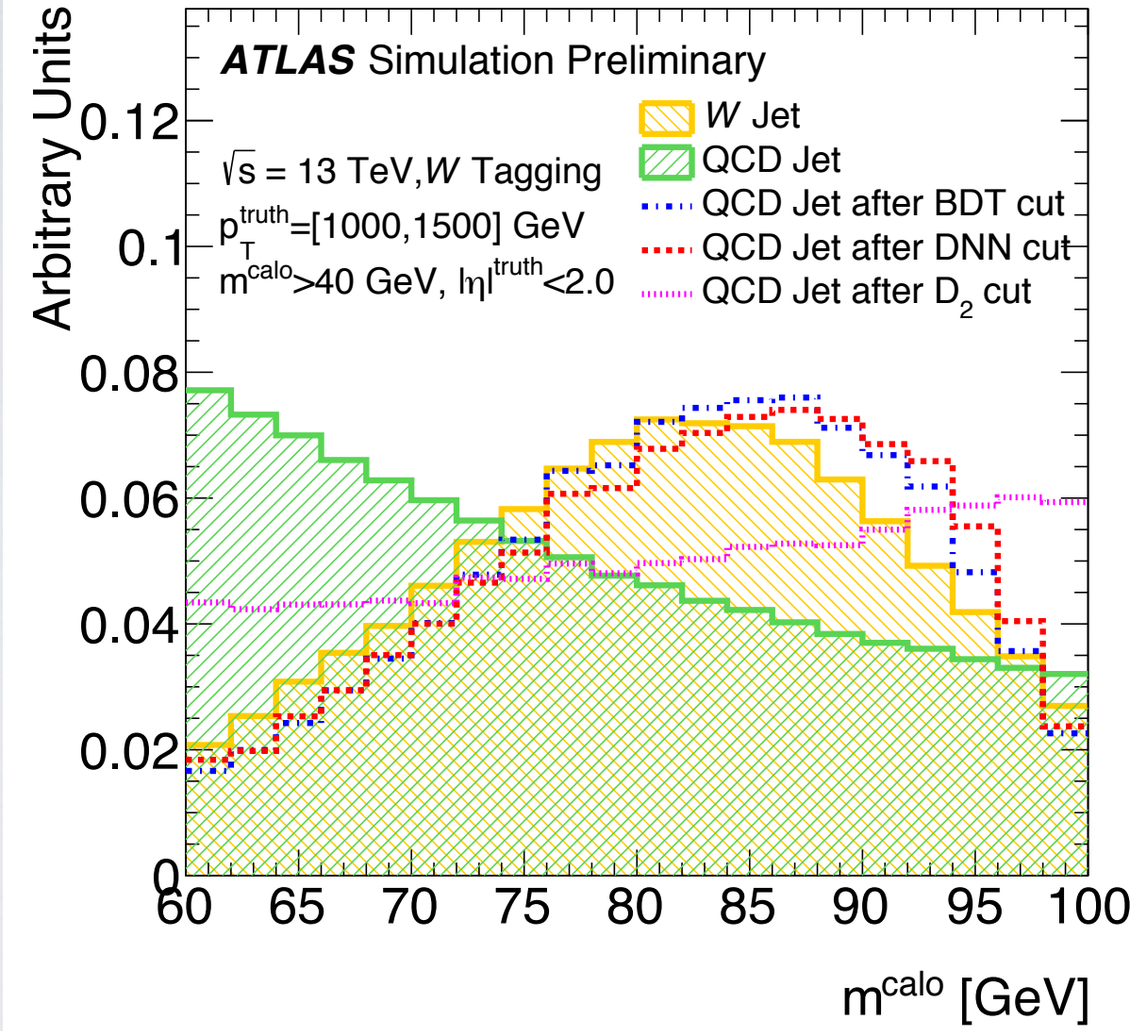




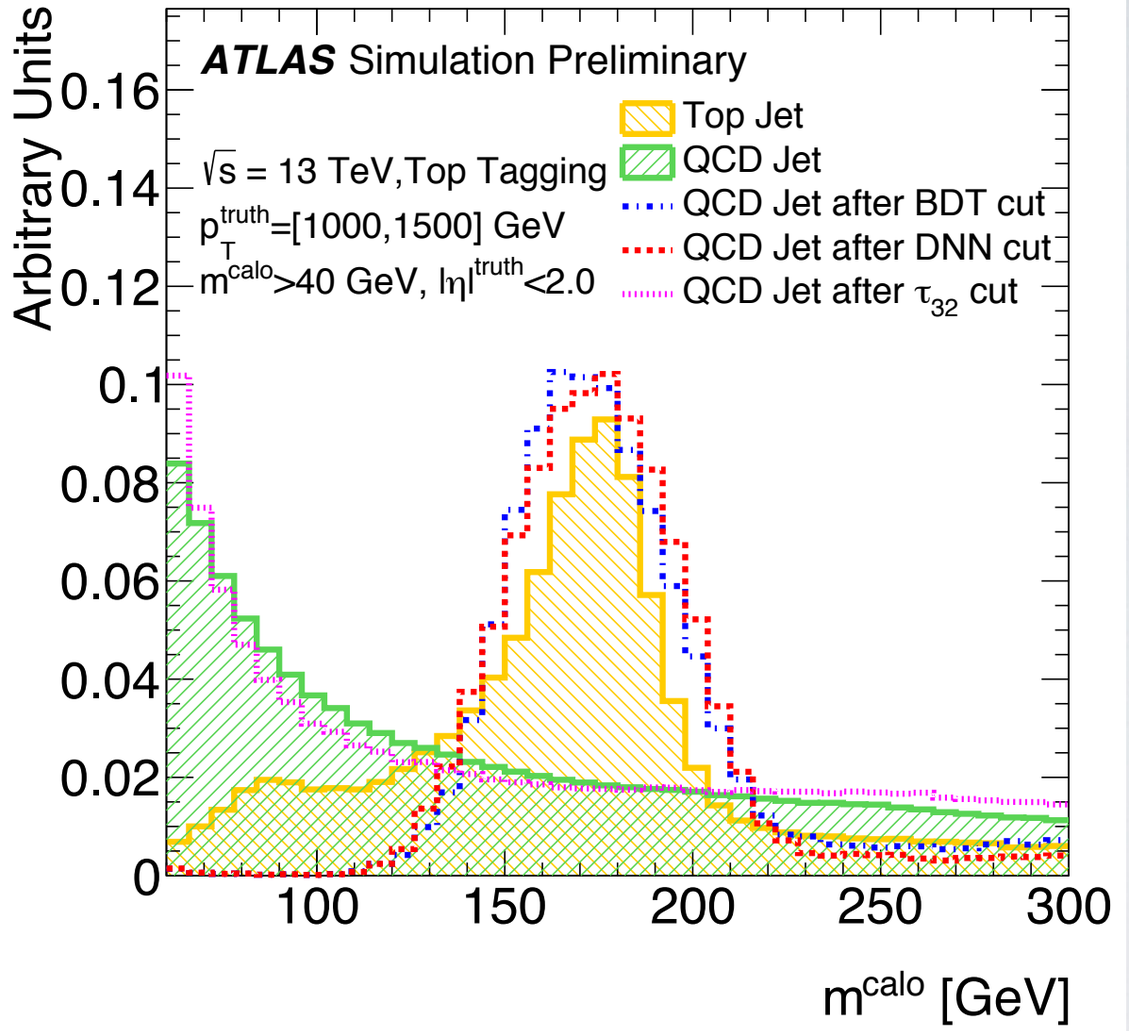


BACKGROUND MASS DISTRIBUTION BEFORE & AFTER TAGGING

W Tagging

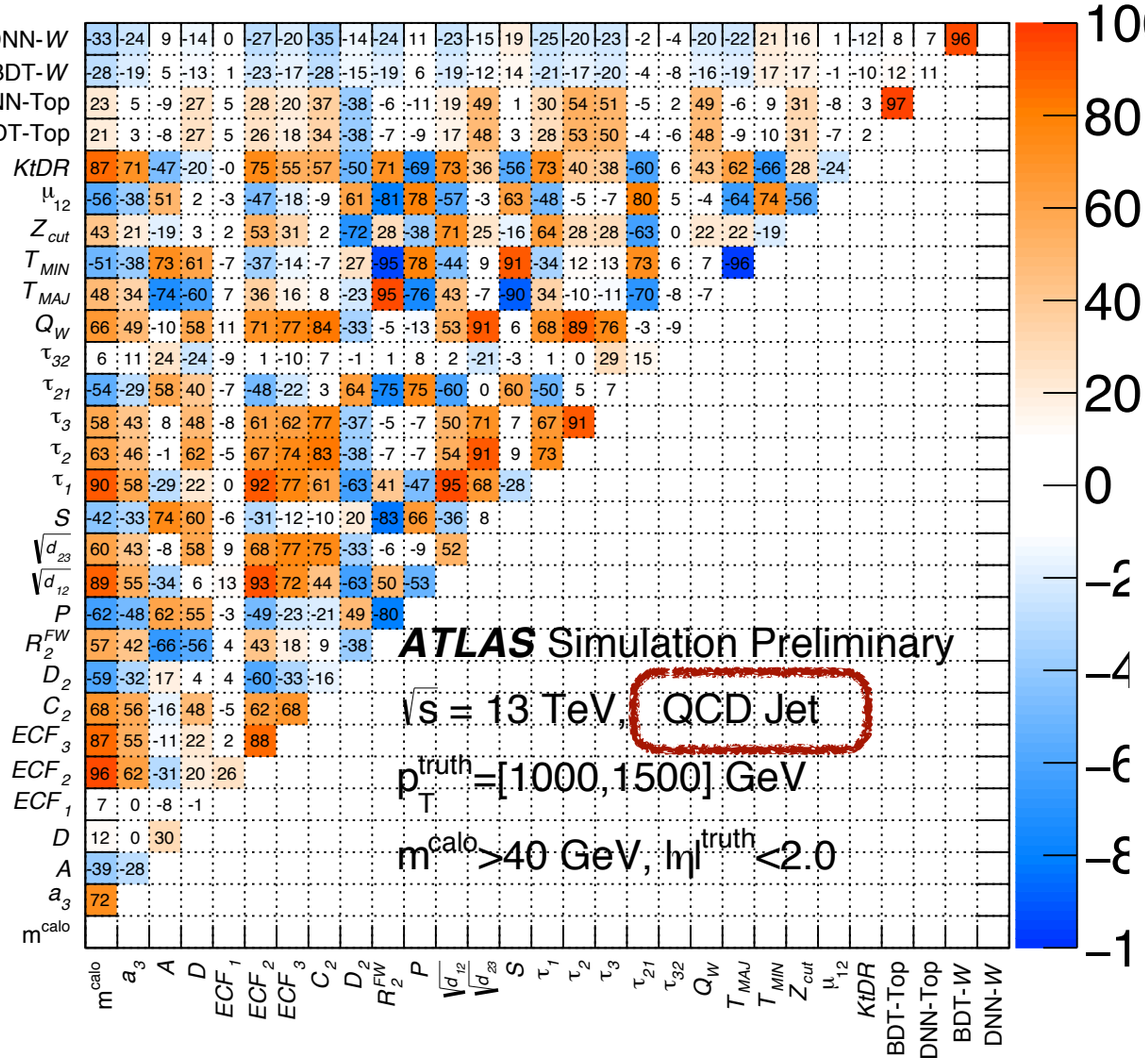


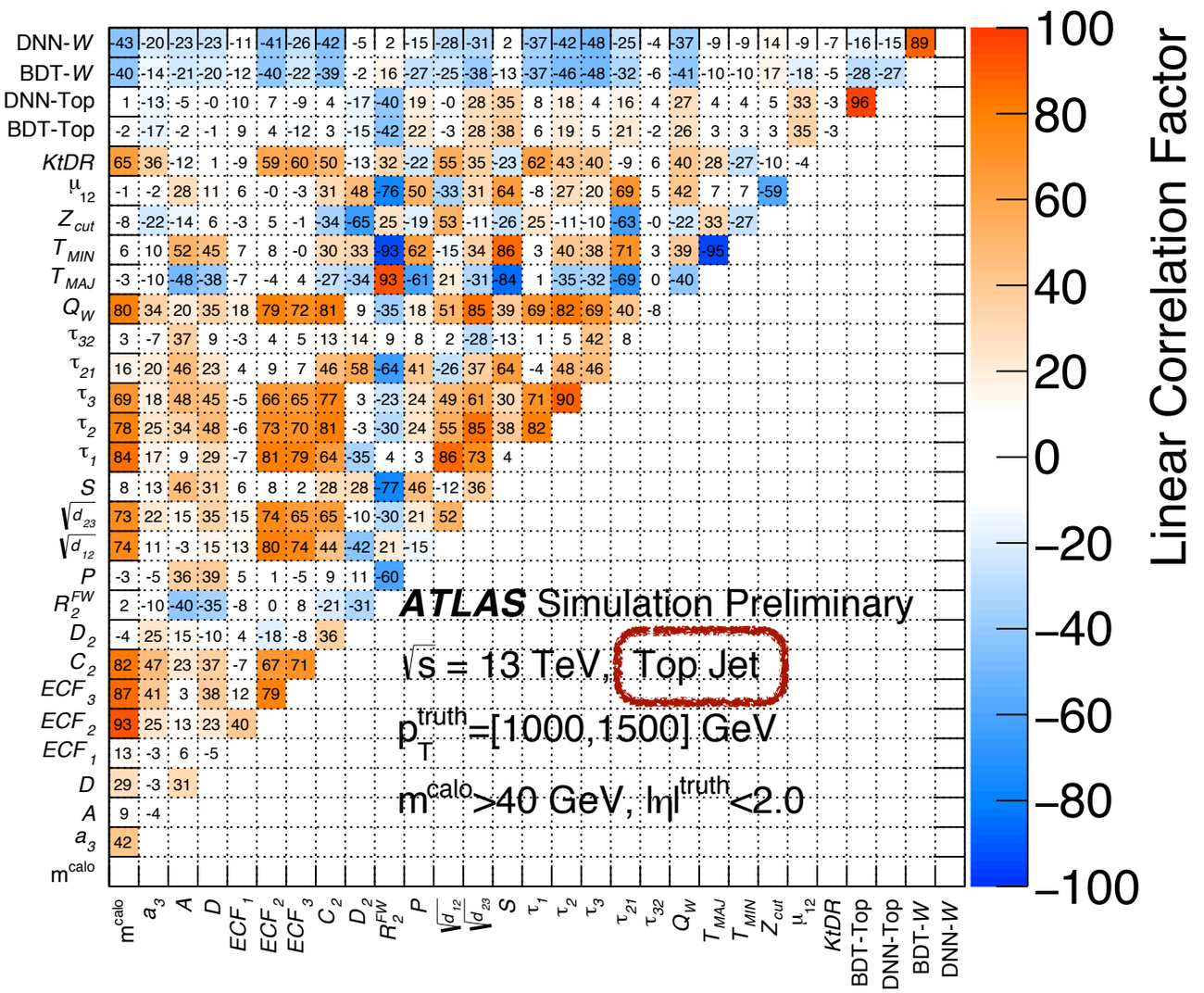
Top Tagging



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

DISCRIMINANT CORRELATIONS



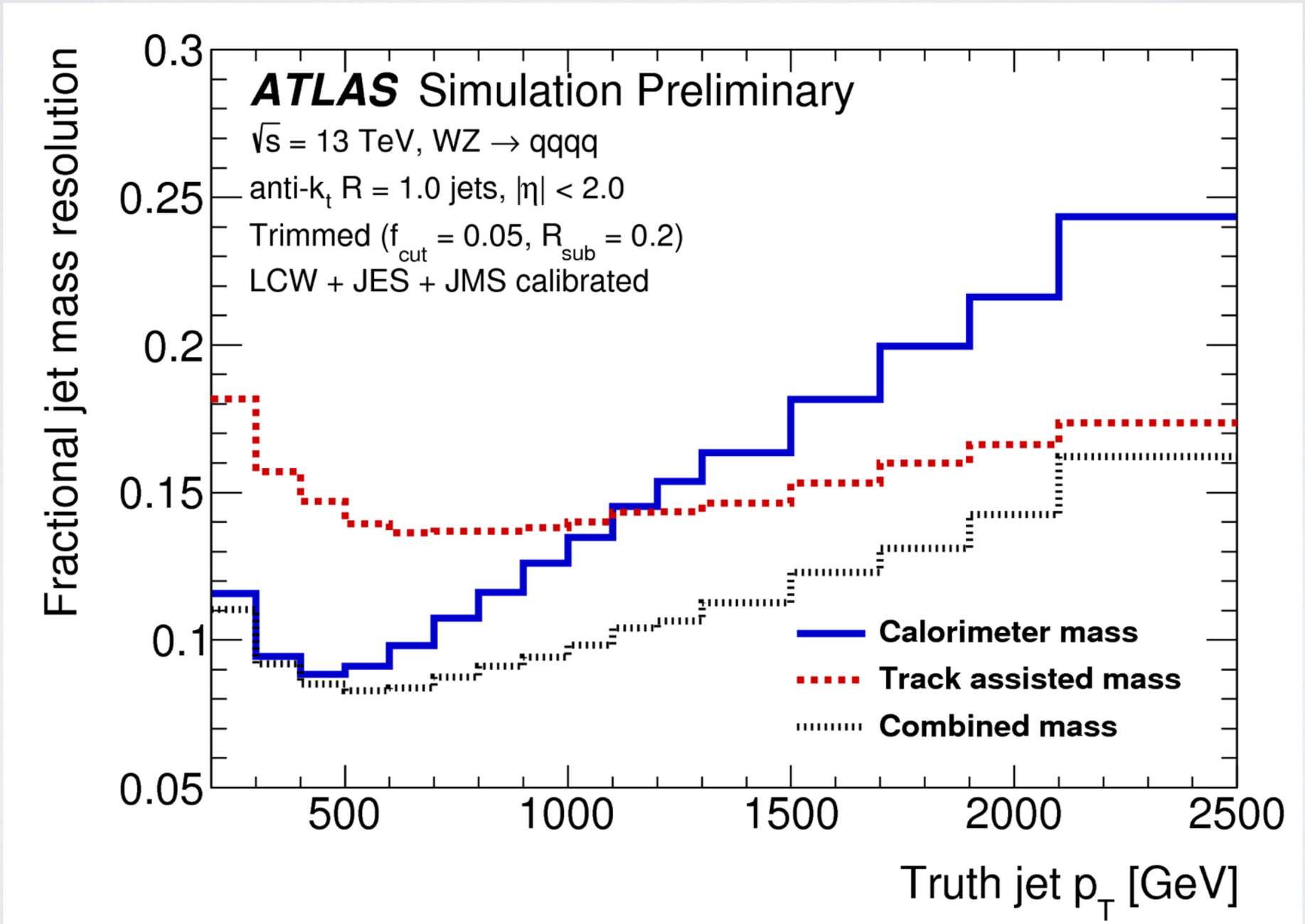


OPTIMIZATION AND PERFORMANCE STUDIES IN MC

ATLAS-CONF-2017-064

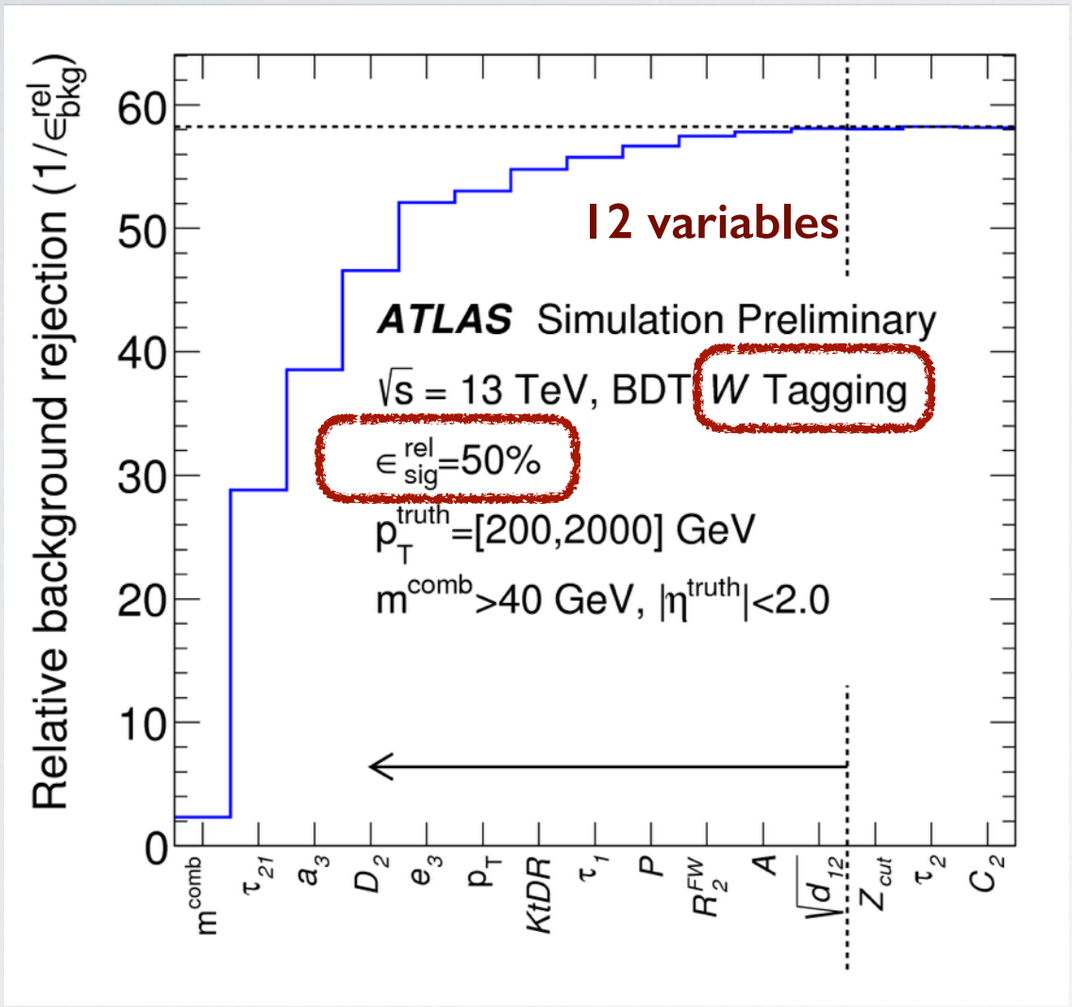
INPUTS OPTIMIZATION

- Updated the training inputs
- One important change: Include combined (calo+track-assisted) mass

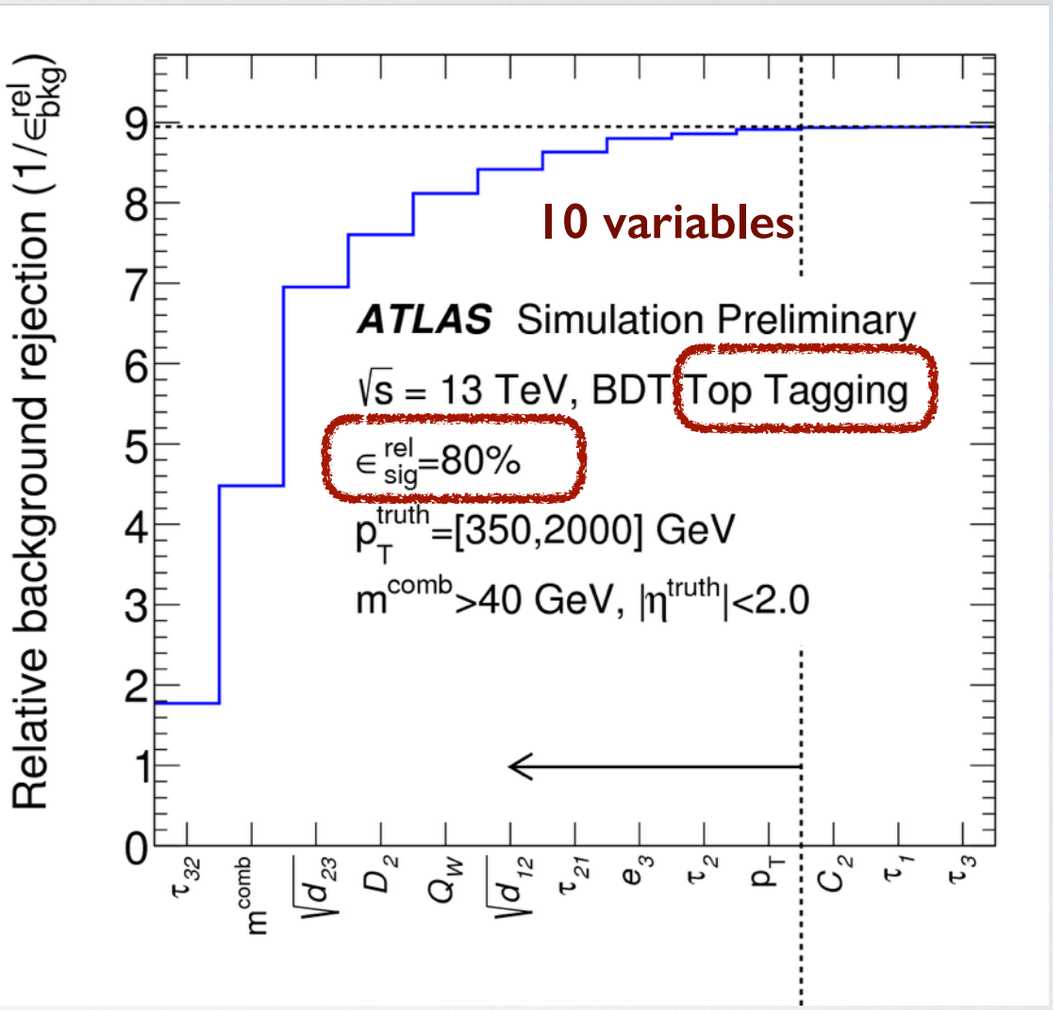


BDT TRAINING - INPUTS OPTIMIZATION

W-Boson Tagging



Top-Quark Tagging

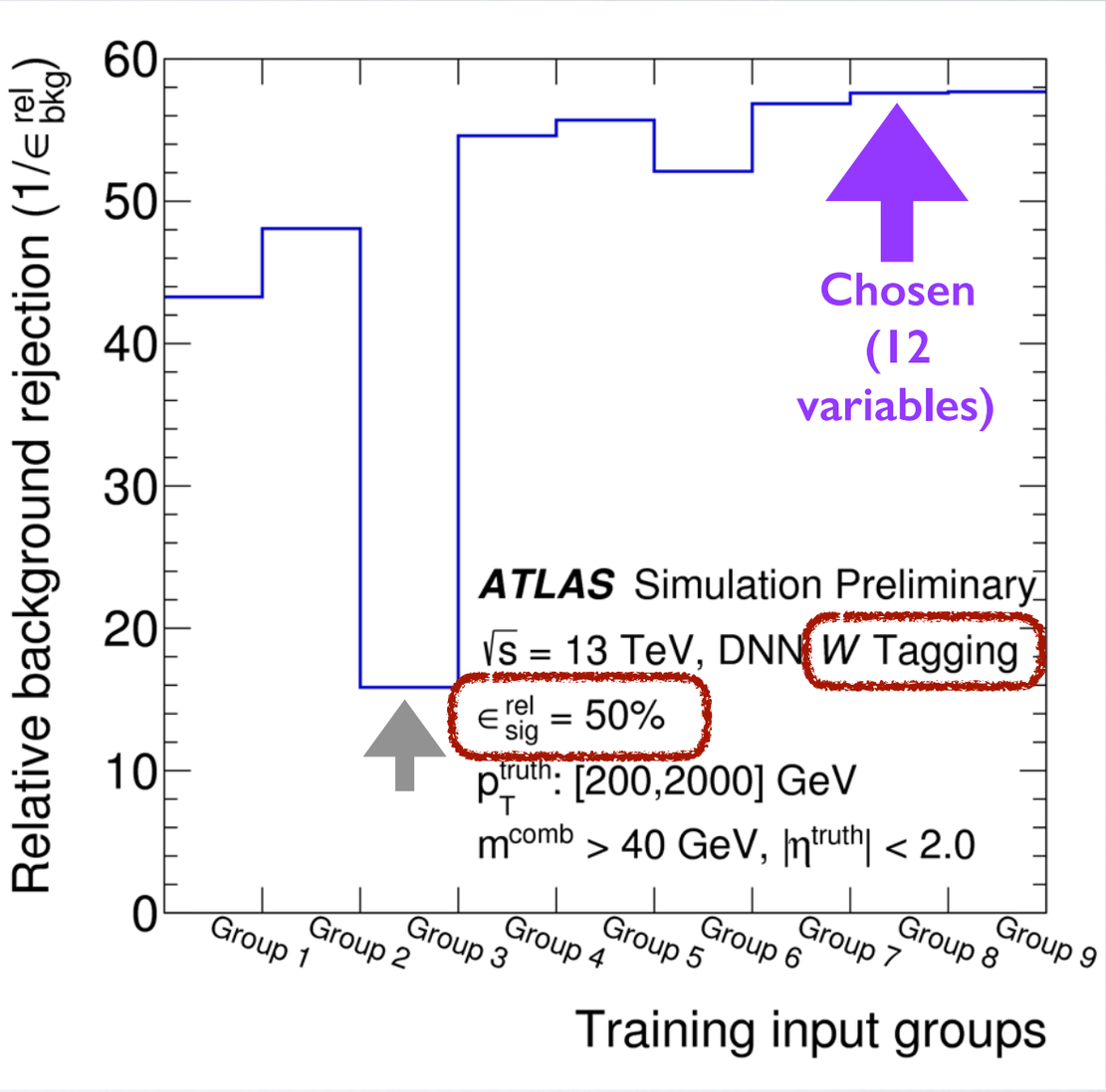


- Add variables in order of importance (improvement in rejection)
- Use a flat p_T spectrum (evaluation)
- Saturation of rejection

DNN TRAINING - INPUTS OPTIMIZATION

W-Boson Tagging

- Study different groups of input variables
- Groups are defined by varying features (scale-dependence, …)
- Use a flat p_T spectrum (evaluation)
- Choose the set with the highest background rejection
- Observed the significance of the scale and jet mass
  - Example: Group 4 = Group 3 + mass

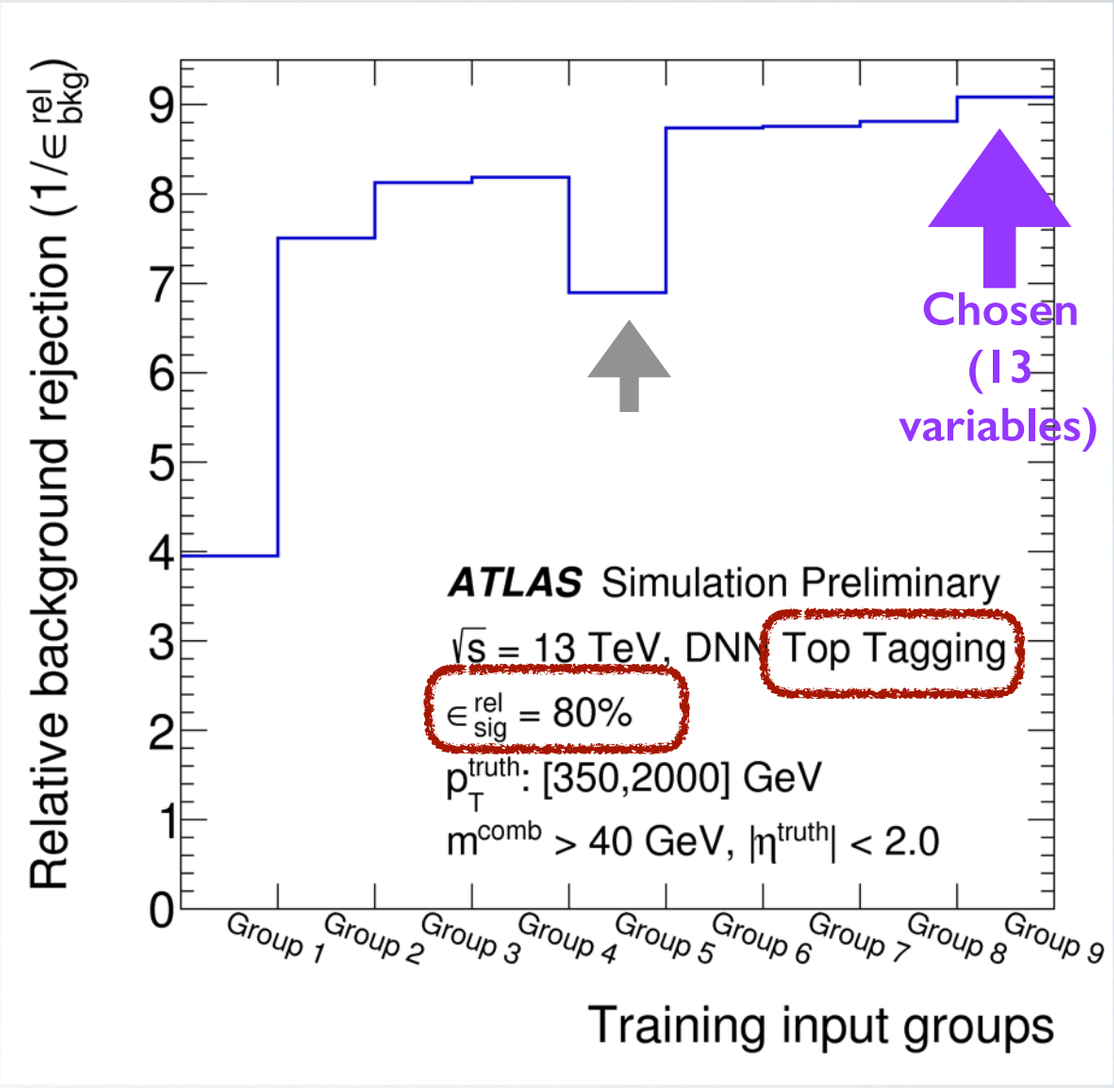


Group 1	τ_1, τ_2, e_3, m^{comb}, p_{T}
Group 2	τ_1, τ_2, e_3, m^{comb}, p_{T}, $\sqrt{d_{12}}$, KtDR
Group 3	τ_{21}, C_2, D_2, R_2^{FW}, $\mathcal{P}$, a_3, A, Z_{CUT}
Group 4	τ_{21}, C_2, D_2, R_2^{FW}, $\mathcal{P}$, a_3, A, Z_{CUT}, m^{comb}
Group 5	τ_{21}, C_2, D_2, R_2^{FW}, $\mathcal{P}$, a_3,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

DNN TRAINING - INPUTS OPTIMIZATION

Top-Quark Tagging

- Study different groups of input variables
- Groups are defined by varying features (scale-dependence, …)
- Use a flat p_T spectrum (evaluation)
- Choose the set with the highest background rejection
- Observed the significance of the scale and jet mass
  - Example: Group 6 = Group 5 + mass

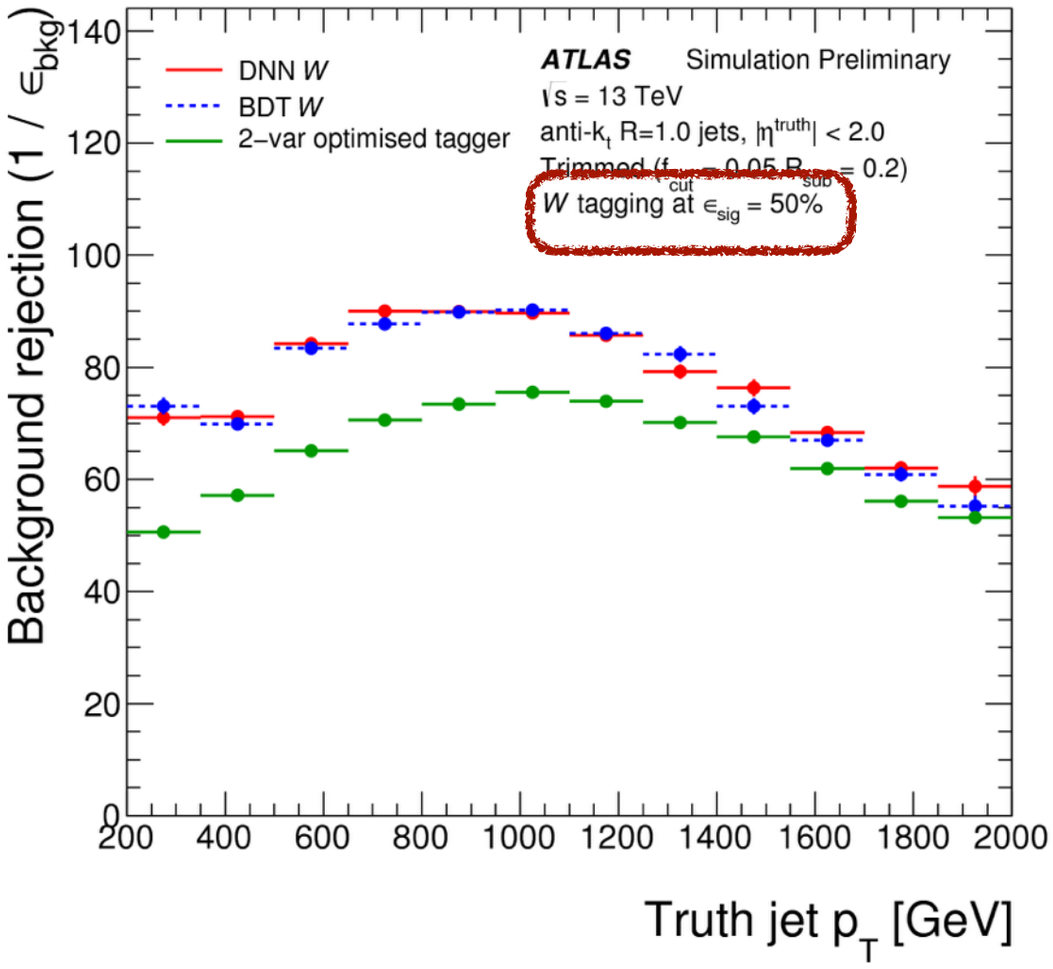


Group 1	C_2, D_2, τ_{21}, τ_{32},
Group 2	C_2, D_2, τ_{21}, τ_{32}, m^{comb}
Group 3	C_2, D_2, τ_{21}, τ_{32}, m^{comb}, p_{T}
Group 4	τ_1, τ_2, τ_3, e_3, m^{comb}, p_{T}
Group 5	C_2, D_2, τ_{21}, τ_{32}, $\sqrt{d_{12}}$, $\sqrt{d_{23}}$, Q_W
Group 6	C_2, D_2, τ_{21}, τ_{32}, $\sqrt{d_{12}}$, $\sqrt{d_{23}}$,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

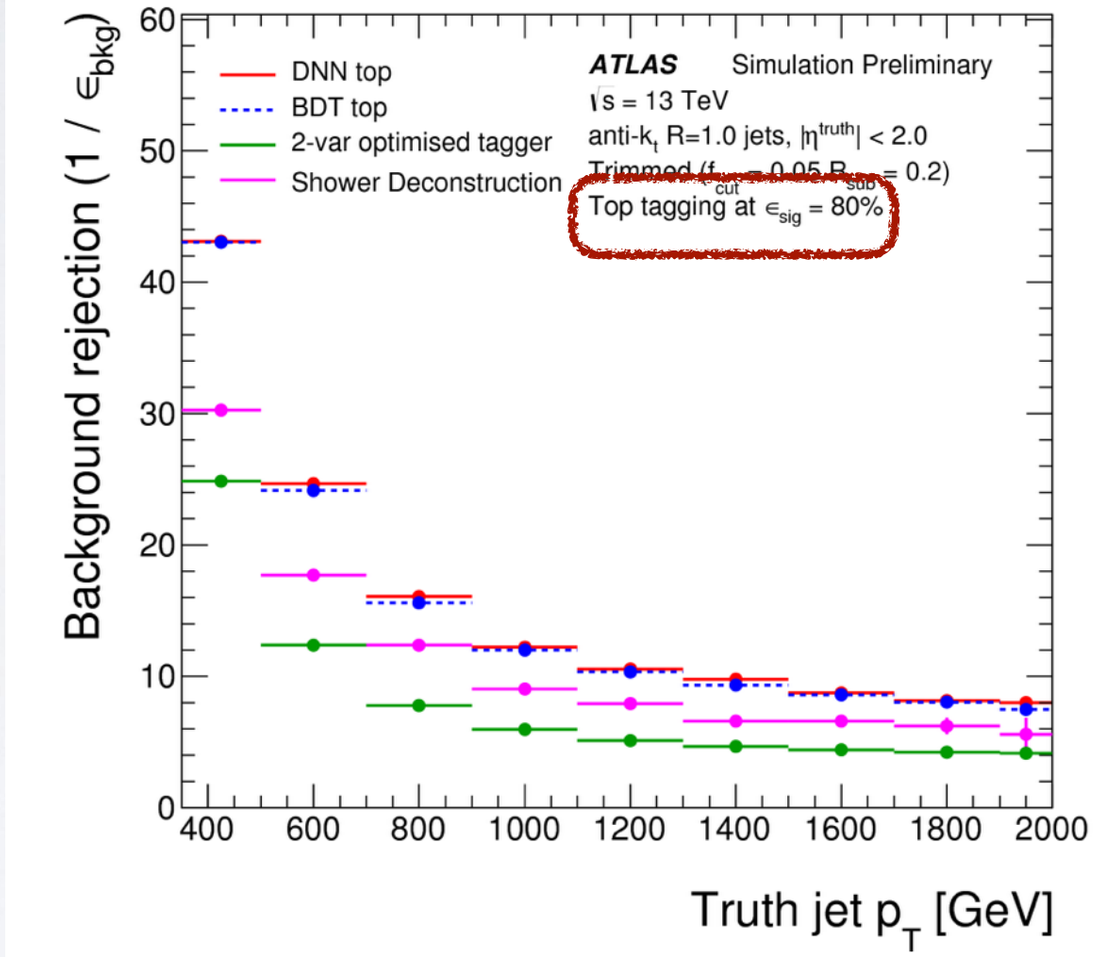
PERFORMANCE COMPARISON

Background Rejection at Fixed-Efficiency Working Point

W-Boson Tagging



Top-Quark Tagging



- BDT & DNN: Improvements observed for both W and top tagging
- Improvement is more significant for top tagging

PERFORMANCE STUDIES IN DATA

- Measure signal efficiency and background rejection in data
- Full ATLAS 2015+2016 dataset: $(36.1 - 36.7)\text{fb}^{-1}$

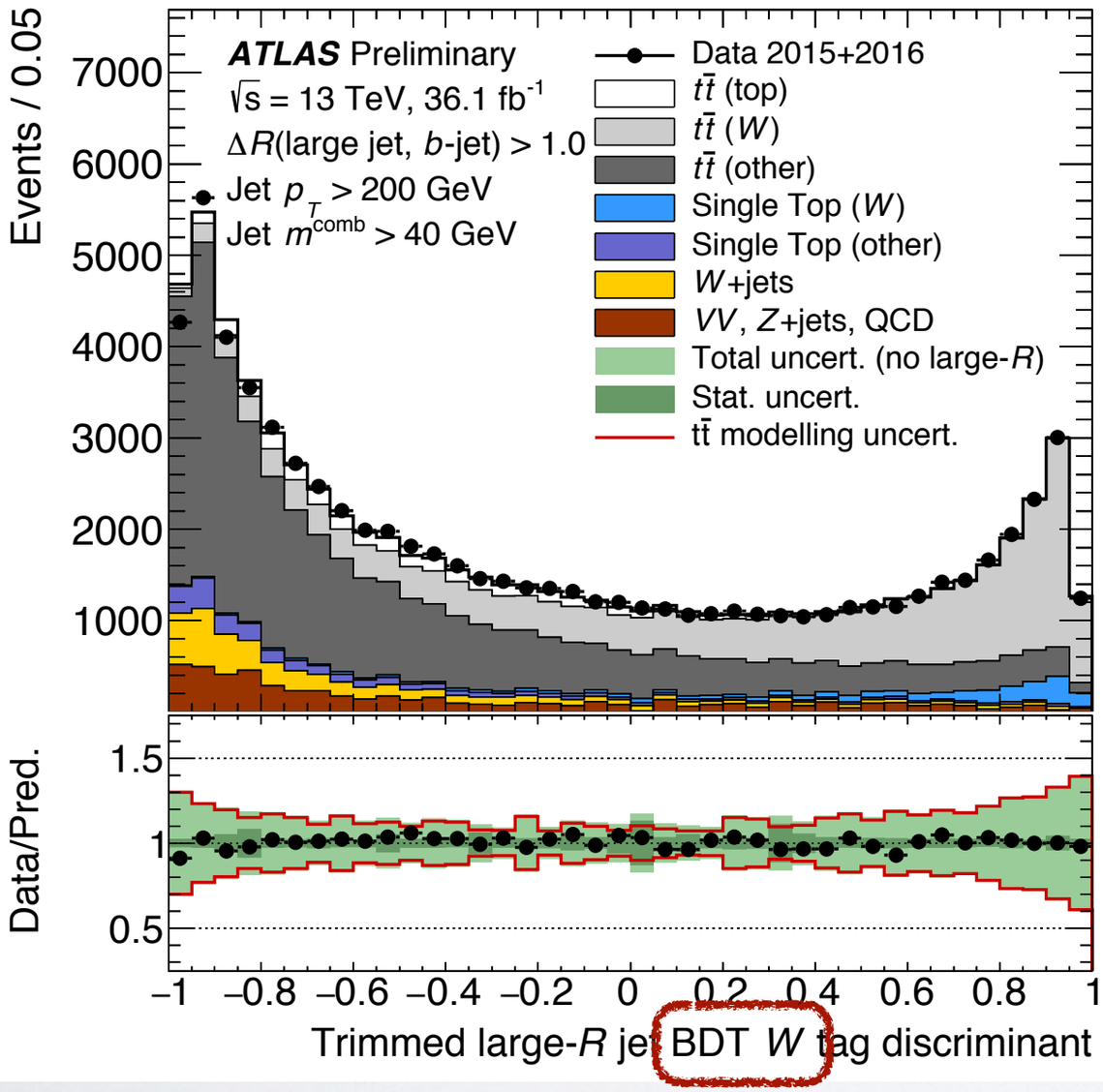
Signal: ttbar with single leptonic top decay

Background: Different background topologies, different features

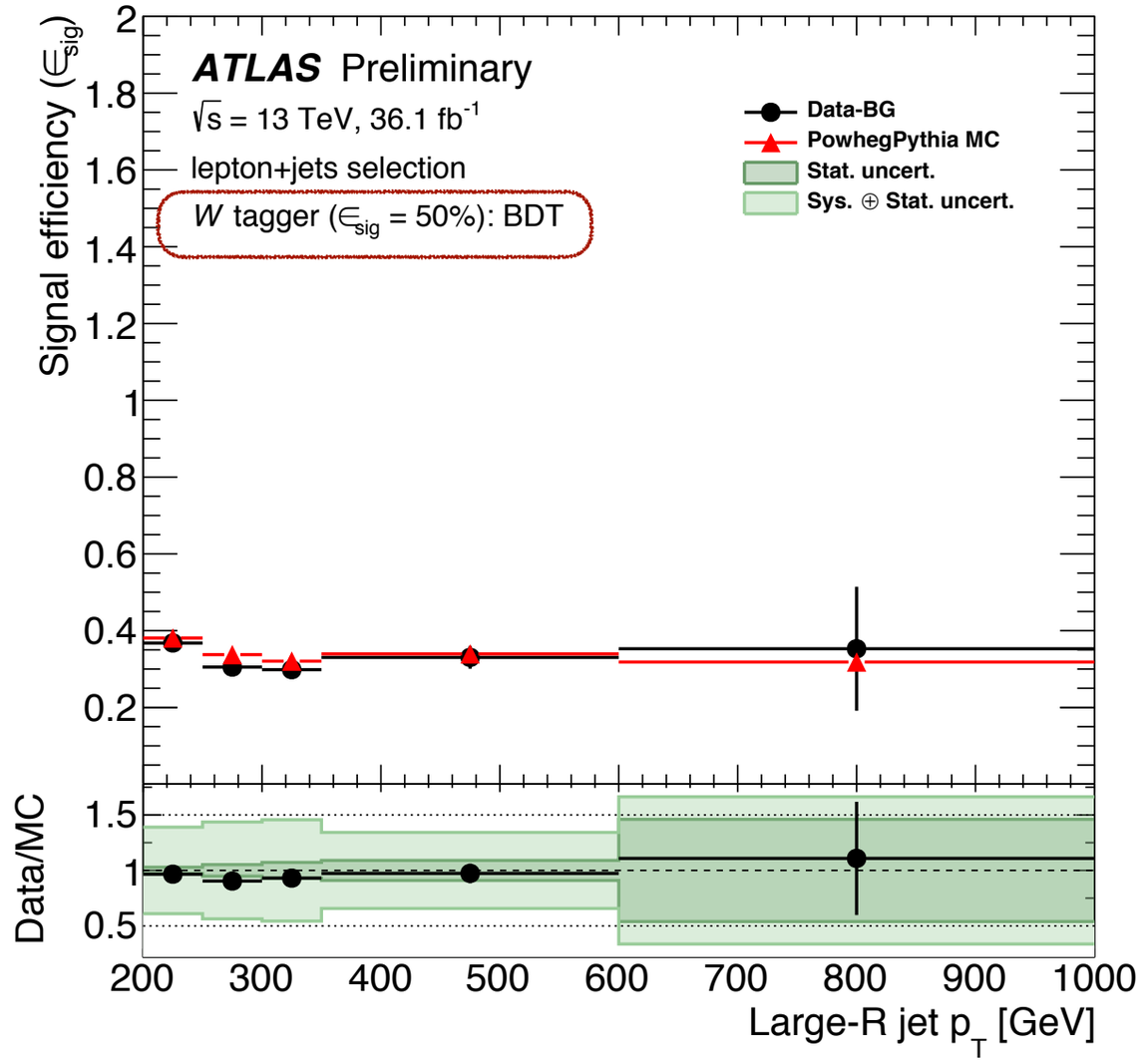
- Dijet events
- Photon + jet events

W TAGGING PERFORMANCE IN DATA - SIGNAL

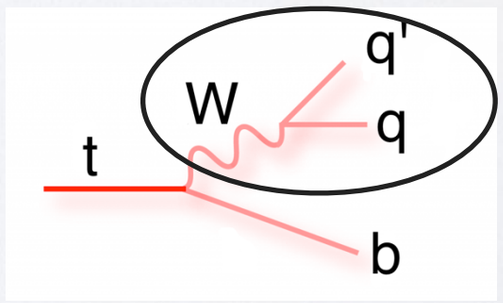
BDT Discriminant



Signal Efficiency

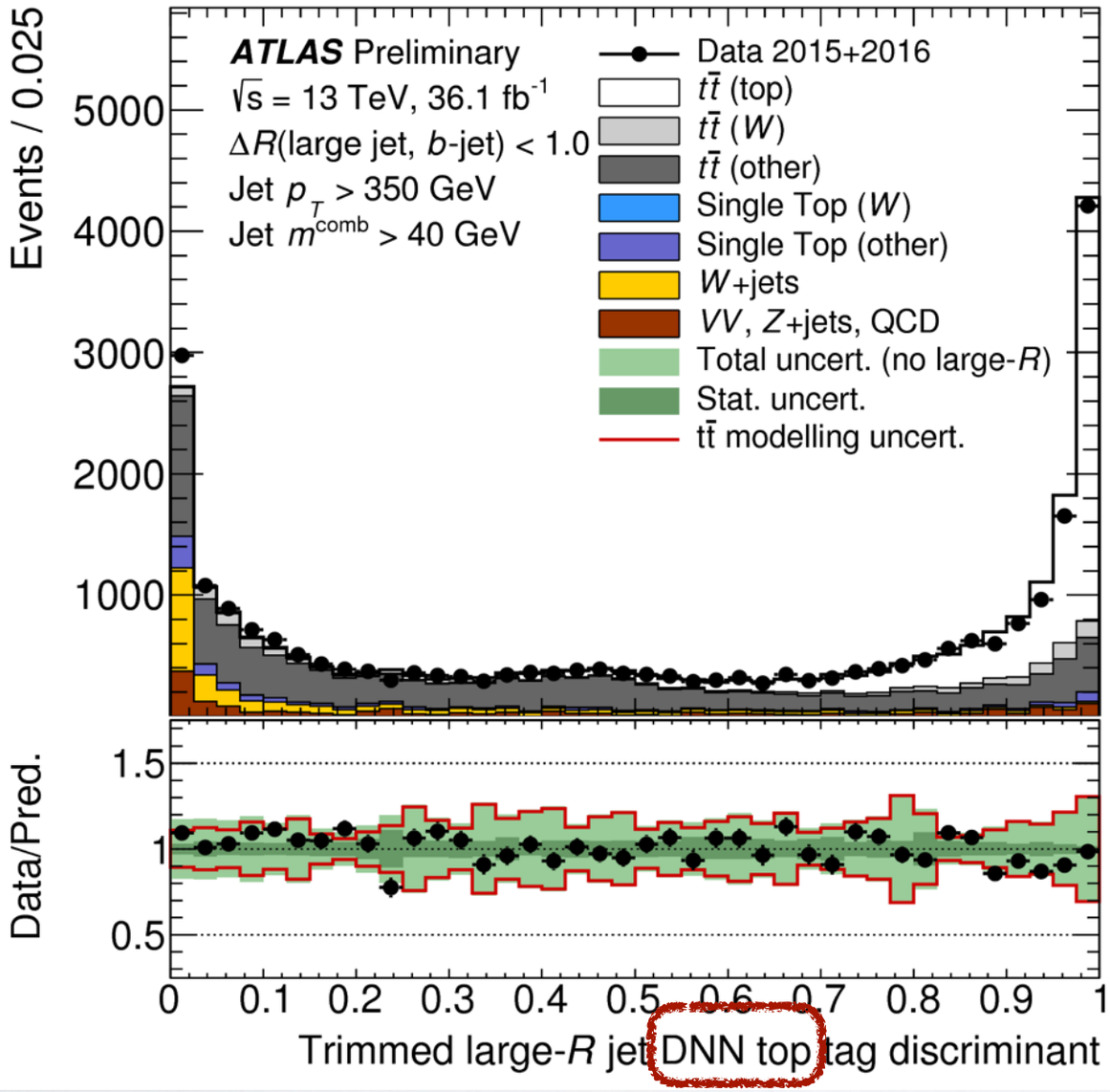


- Well modelled

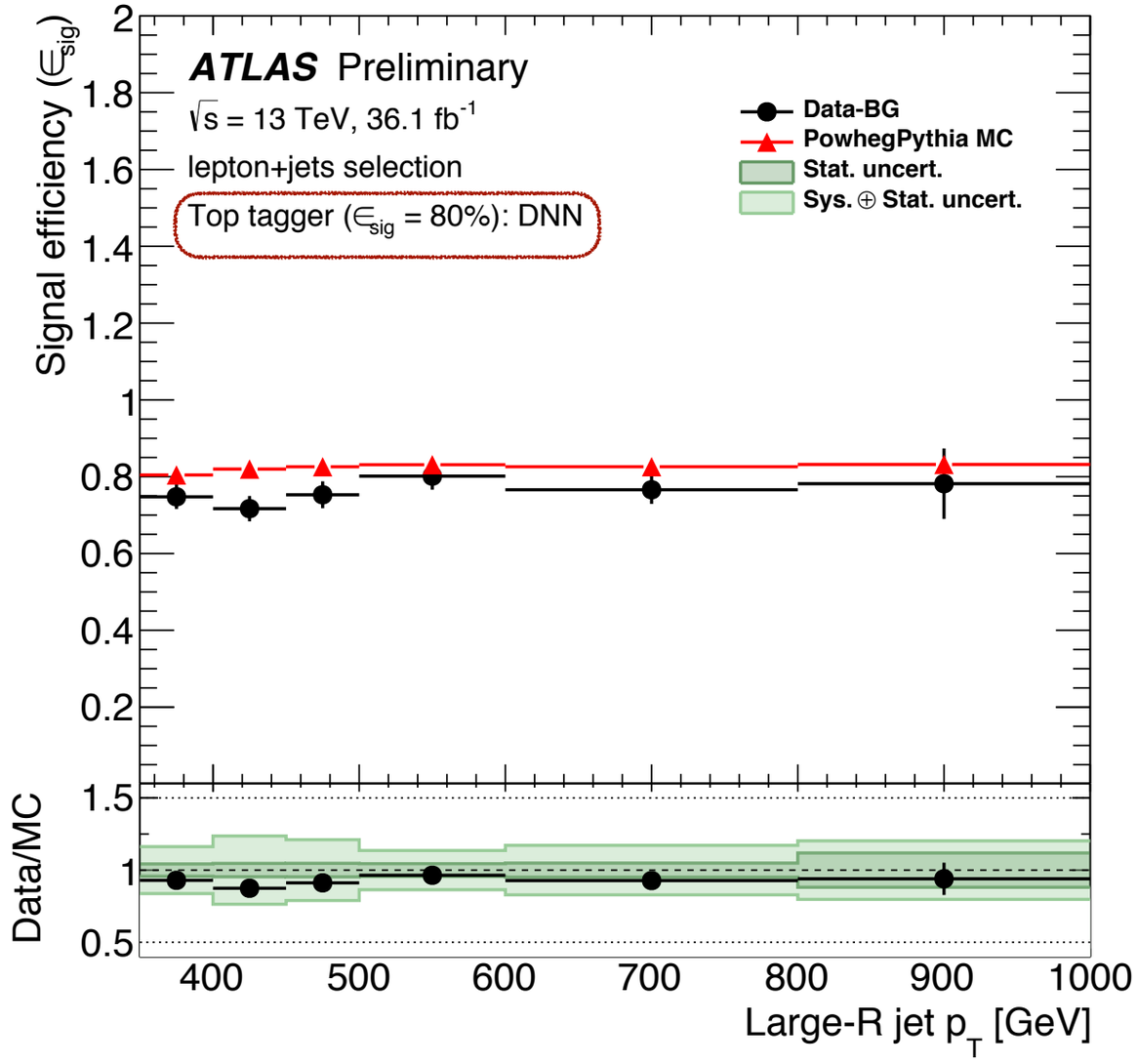


TOP TAGGING PERFORMANCE IN DATA - SIGNAL

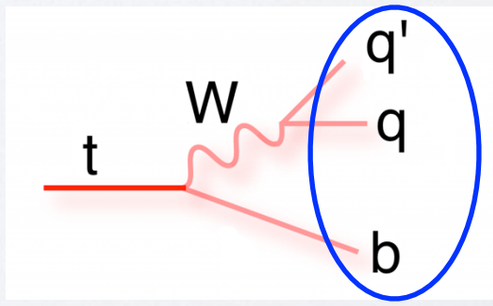
DNN Discriminant



Signal Efficiency

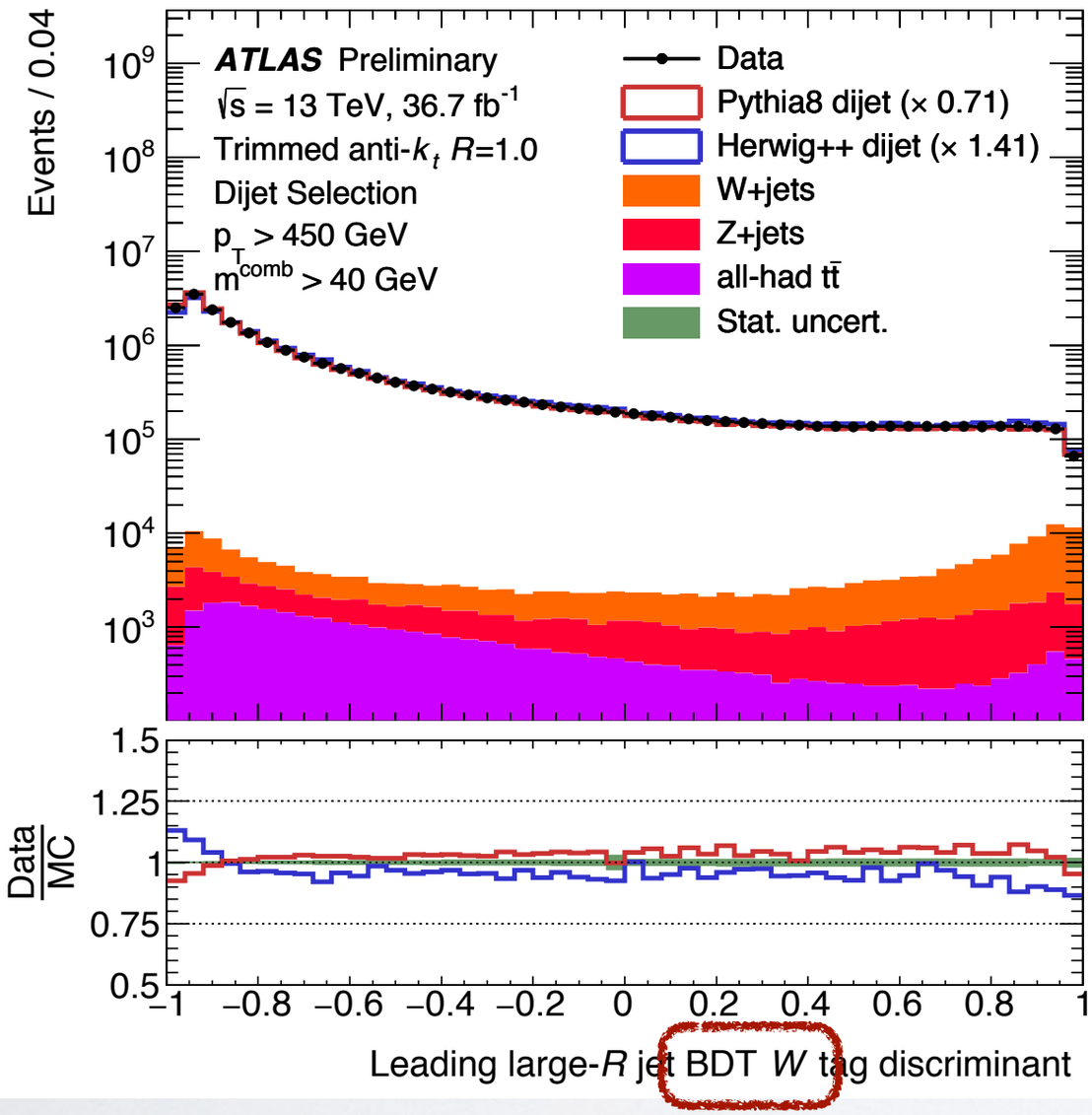


- Well modelled

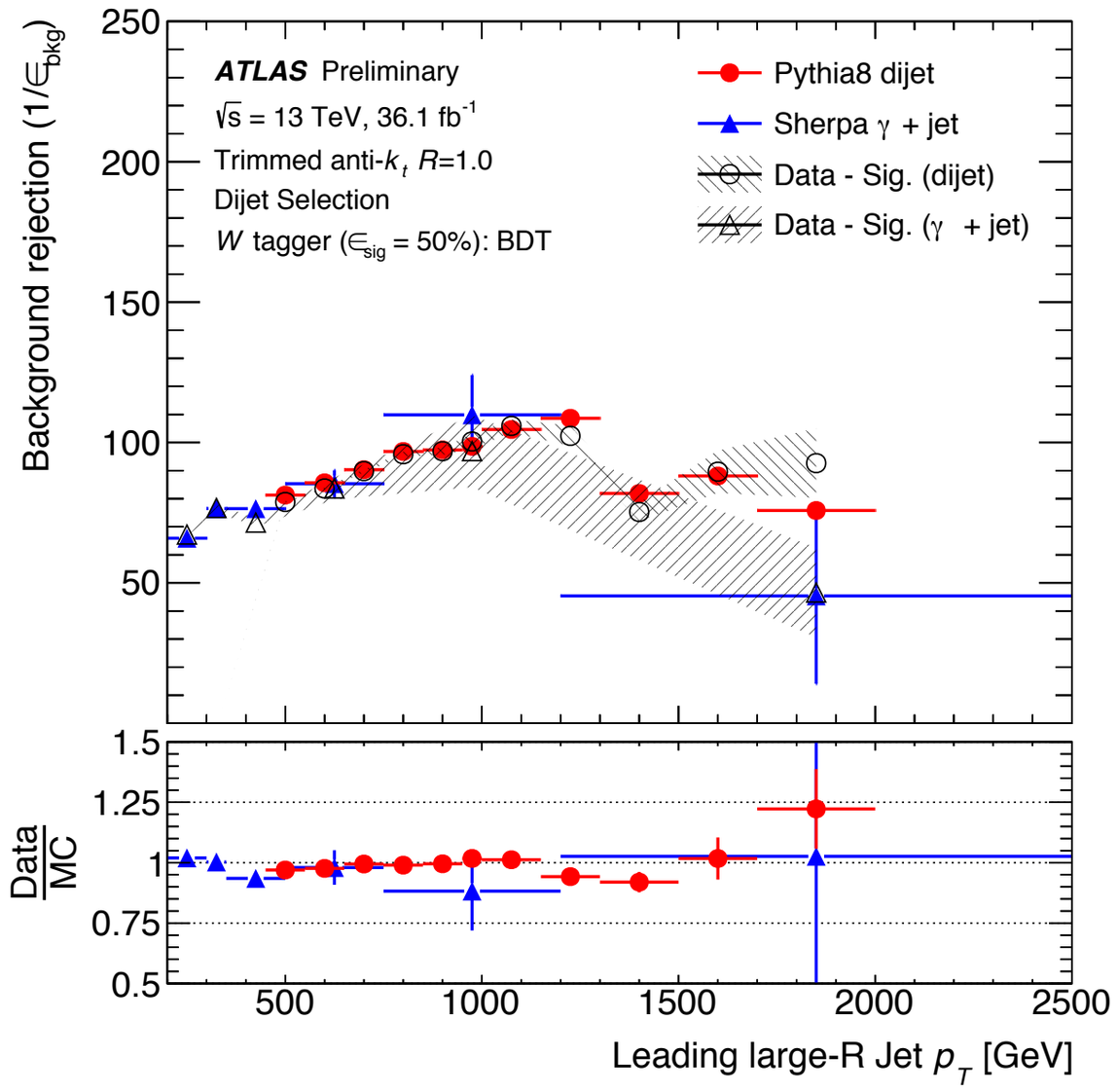


W TAGGING PERFORMANCE IN DATA - BACKGROUND

BDT Discriminant



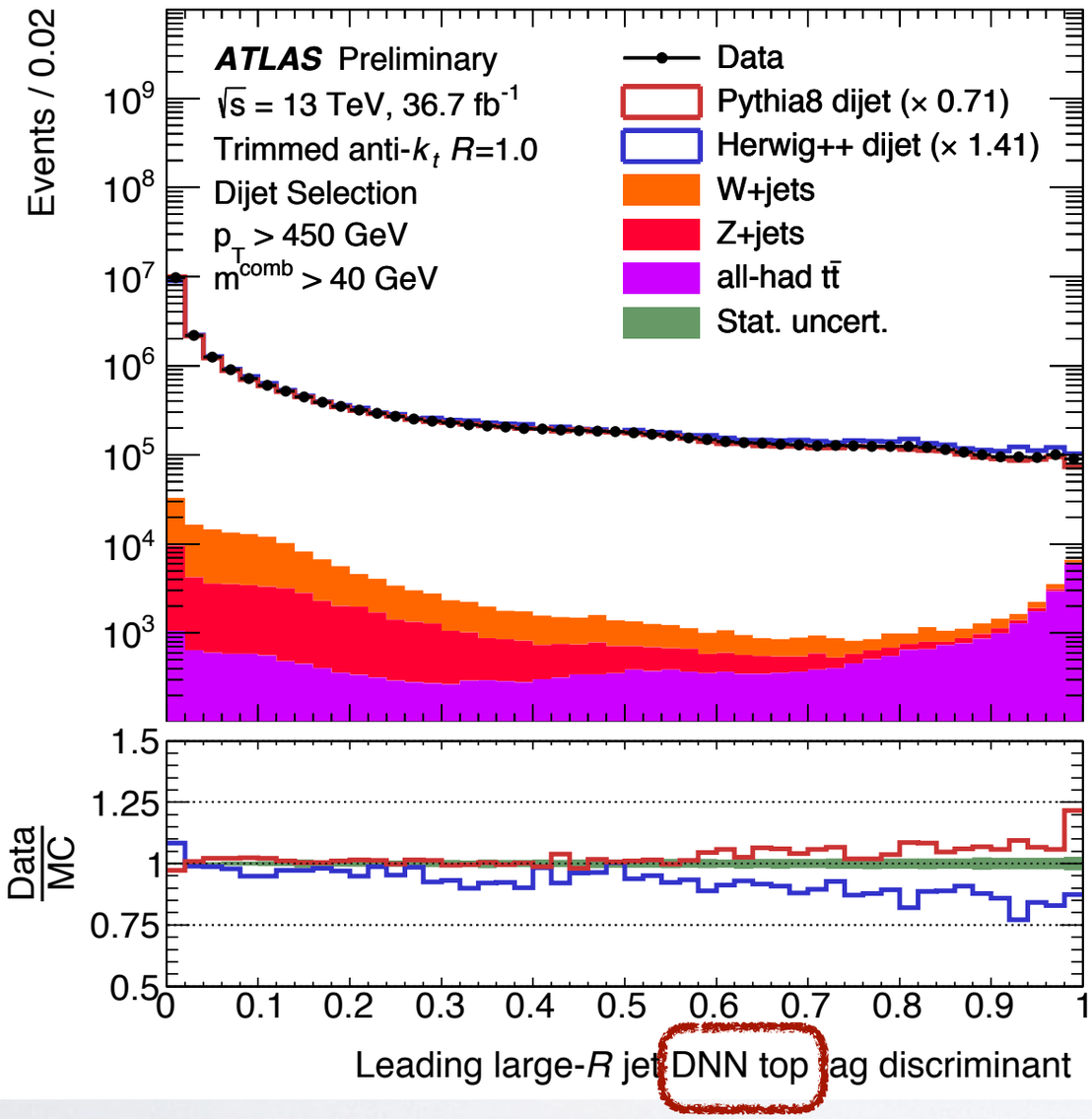
Background Rejection



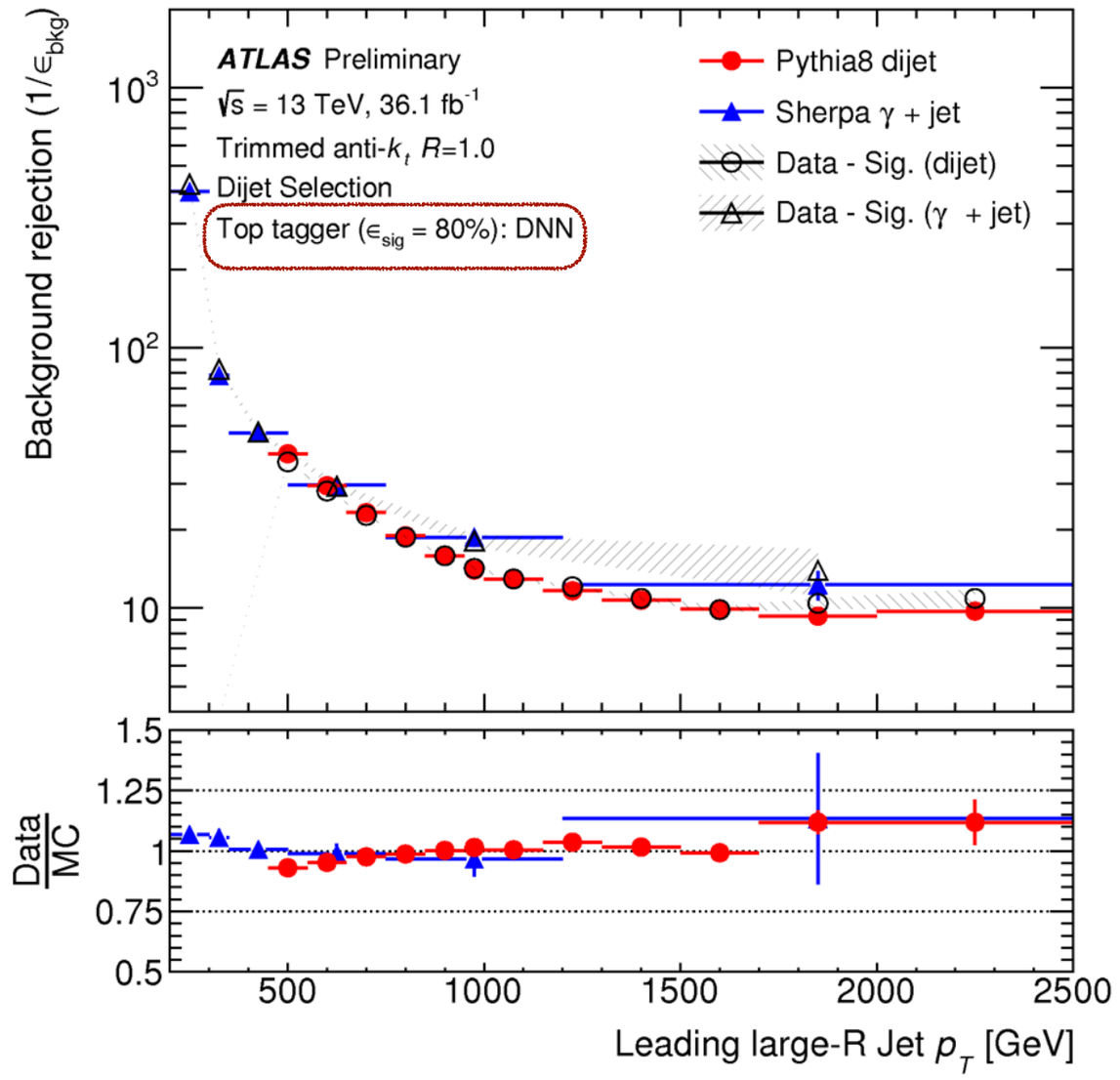
- Well modelled

TOP TAGGING PERFORMANCE IN DATA - BACKGROUND

DNN Discriminant



Background Rejection



- Well modelled

CONCLUSION

Combining high-level inputs in BDT and DNN improves background rejection

- Observed similar performance for BDT and DNN

Signal efficiency measurement in data & MC

- Modelling in agreement with data within uncertainties

Background rejection measurement in data & MC

- Modelling in agreement with data for baseline MC generators
- Similar background rejection in the common region

THANK YOU!

BACKUP

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

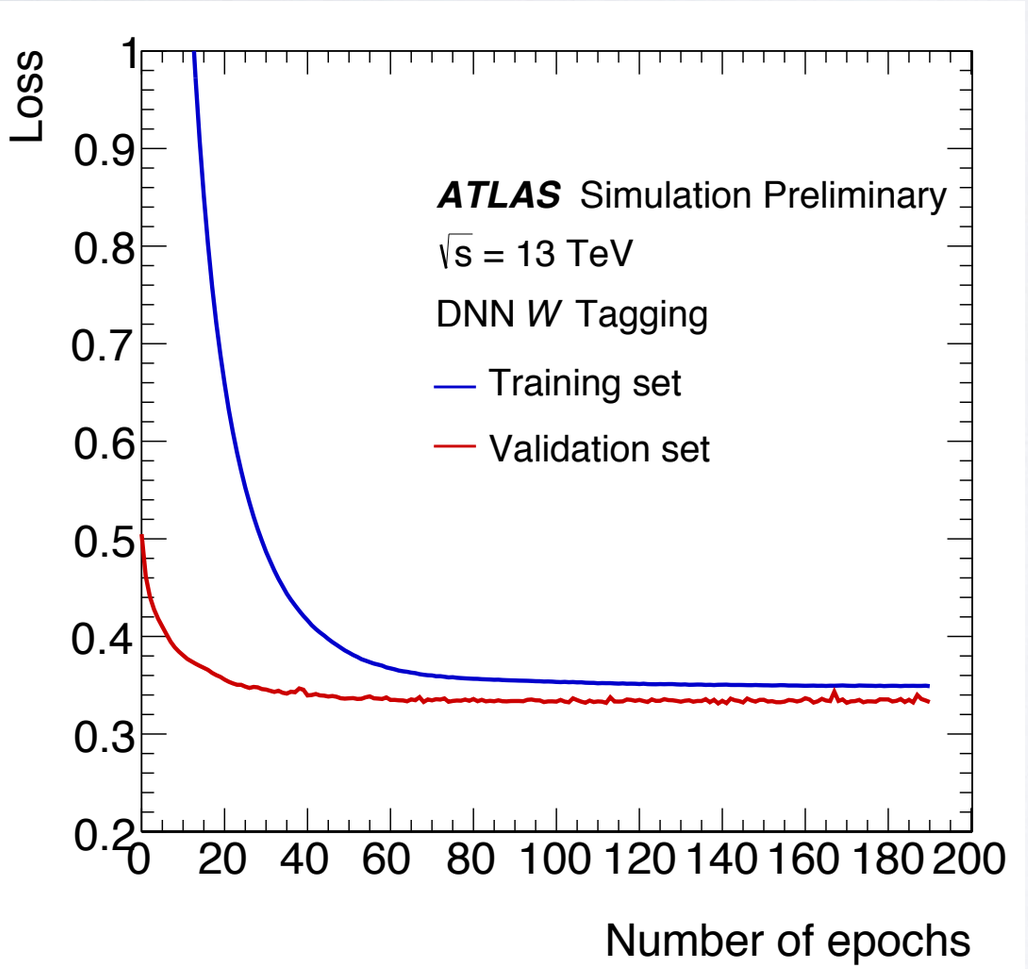
- Layer type = Dense with Batch Normalization, Maxout with Batch Normalization
  - Number of Maxout layers = 5, 10, 15, 20, 25
- Number of hidden layers = 3, 4, 5, 6
- Activation function = Rectified linear units (relu), tanh
- Weight initialization = Glorot uniform, He normal
- Learning rate
  - W =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

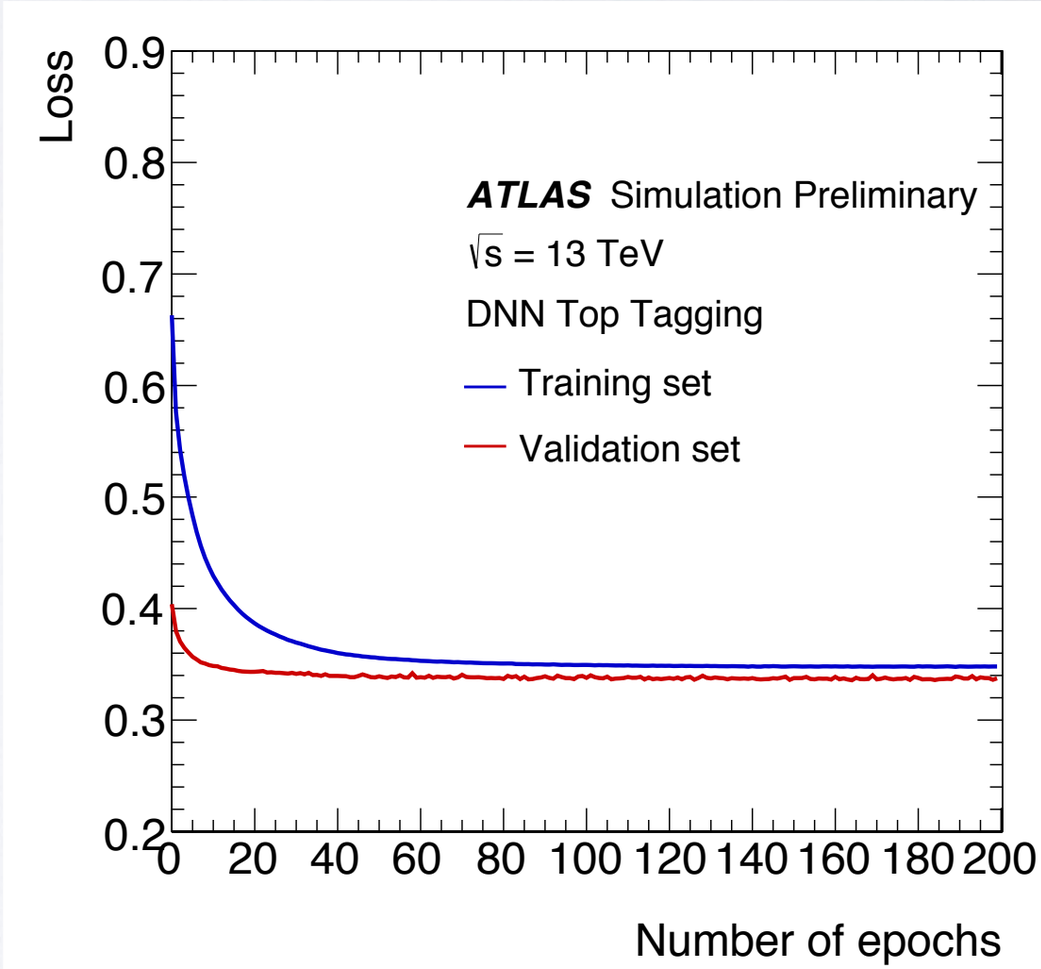
DNN TRAINING - OVERTRAINING

- 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

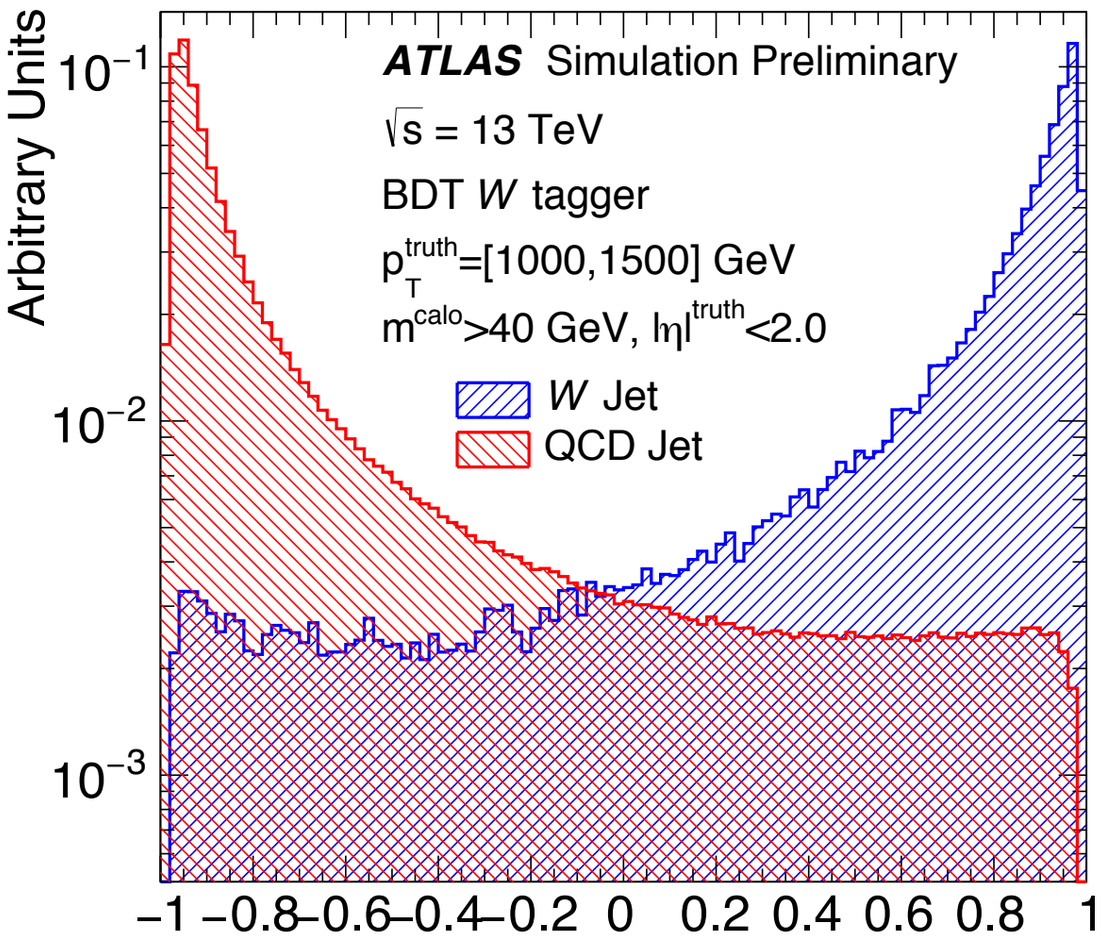
1. Training set: DNN uses this set to optimize the classifier
2. Validation set: Independent of the training set

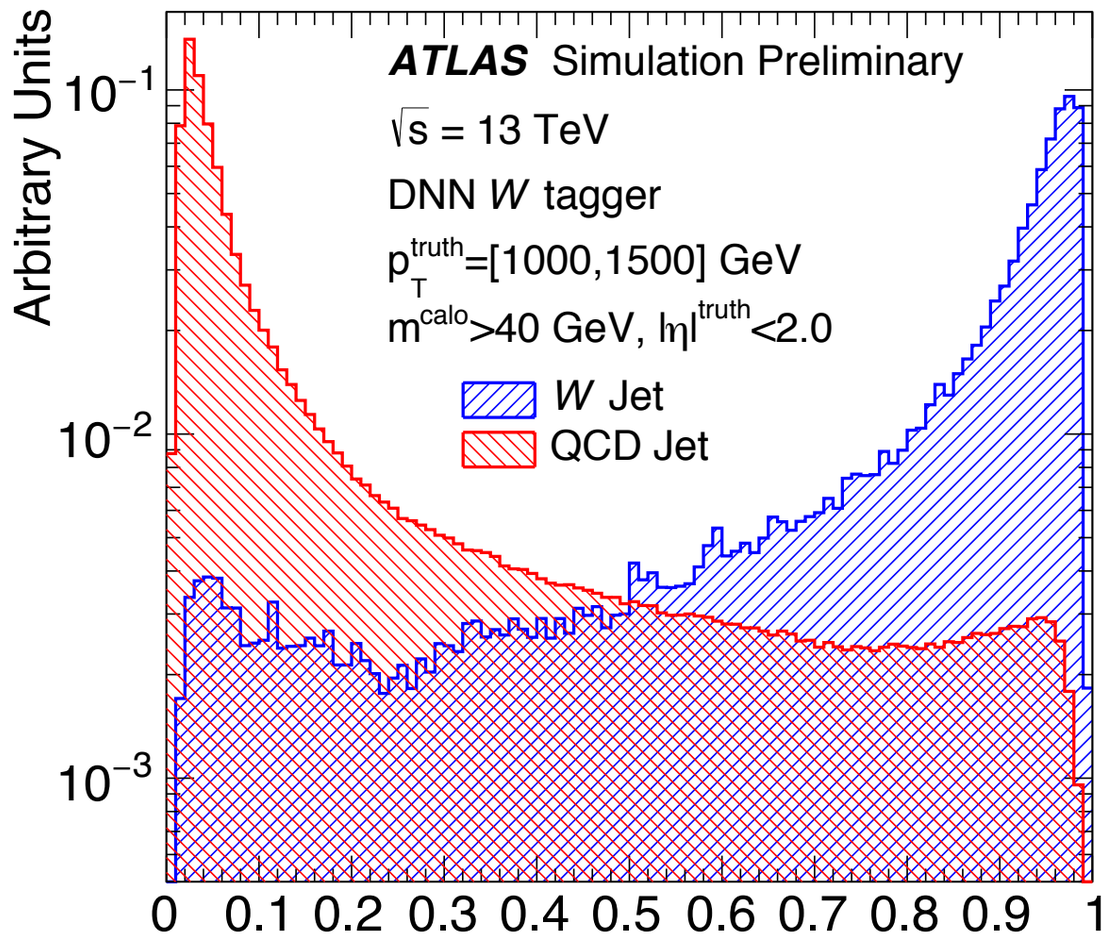
Loss of the validation set ≤ Loss of the training set



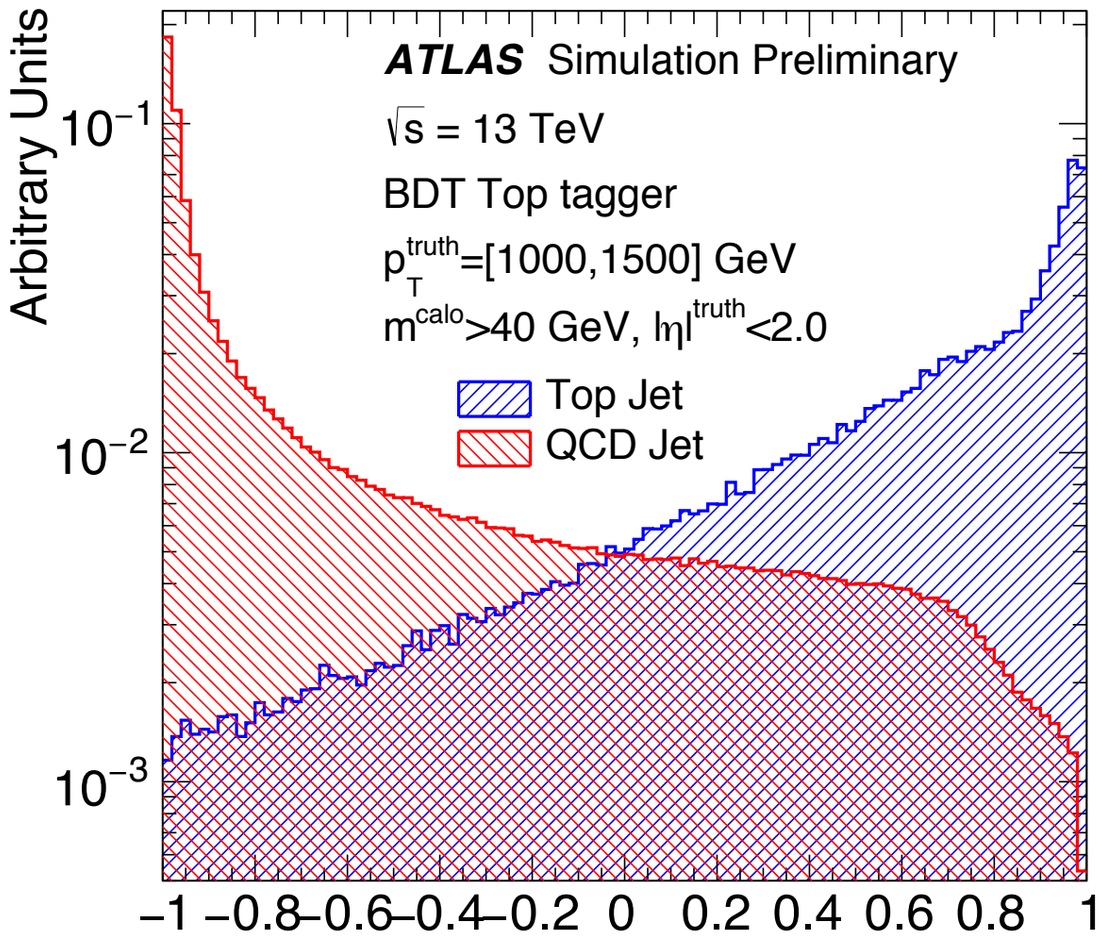


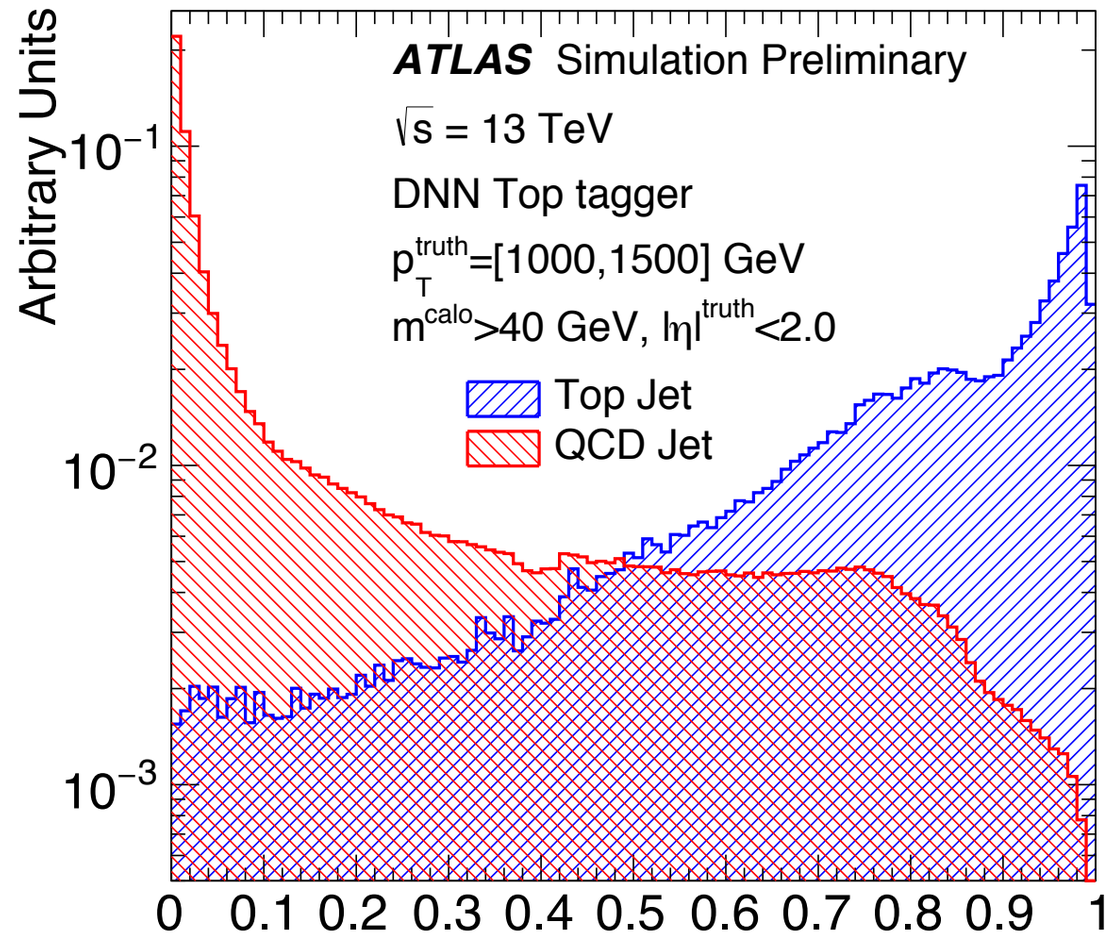
BDT & DNN CLASSIFIERS - W TAGGING



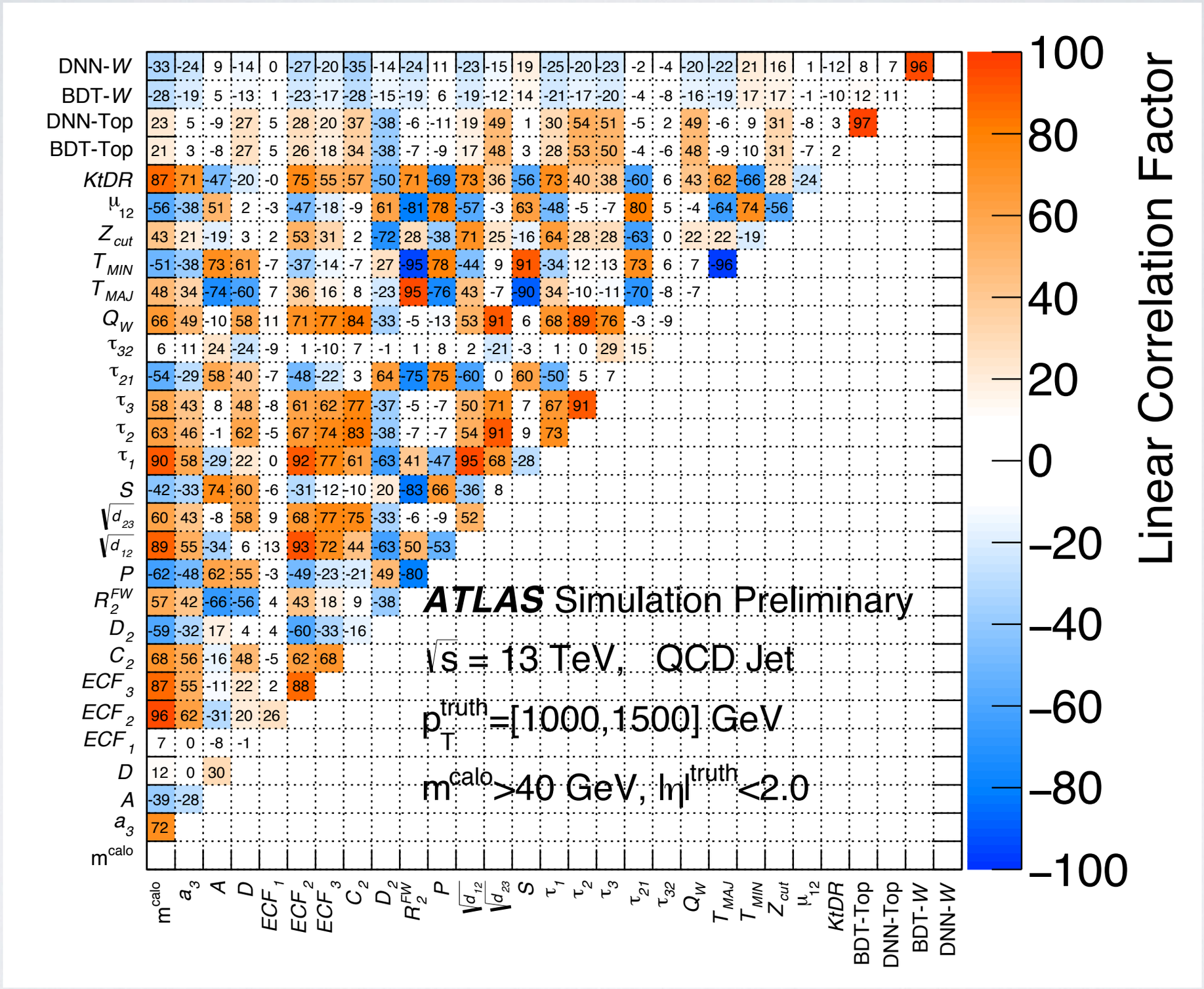


BDT & DNN CLASSIFIERS - TOP TAGGING

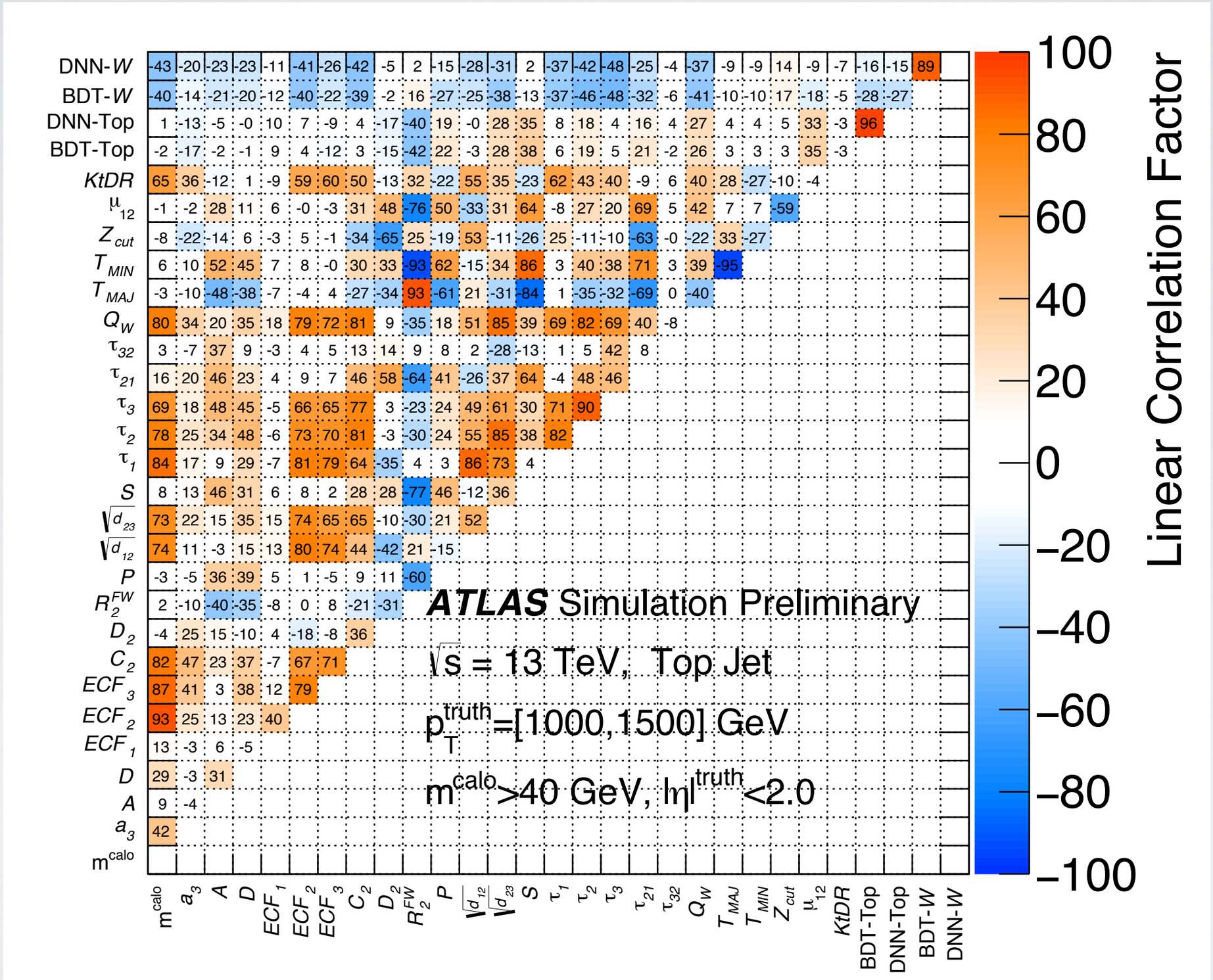




DISCRIMINANT CORRELATIONS



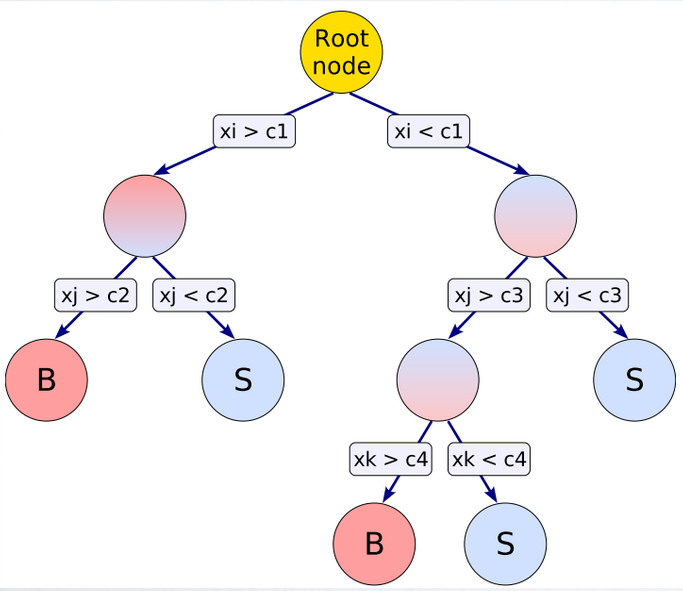
DISCRIMINANT CORRELATIONS



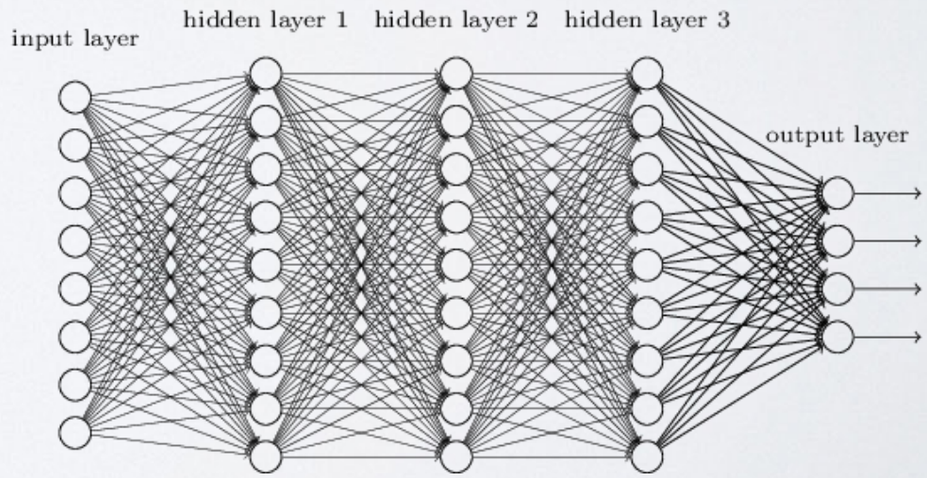
APPLICATION OF BDTS AND DNNS TO W AND TOP TAGGING USING HIGH-LEVEL FEATURES

- Numerous substructure variables are available and are used by ATLAS
- Feed the ML algorithms with jet substructure variables (high-level features)
- Study the performance of *W* and top tagging with two Machine Learning (ML) techniques in parallel

1. Boosted Decision Trees (BDT) using TMVA



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



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, η = [-2,2]
- # Training signal jets = 7×10^5
- # Training QCD jets = 7×10^5

Top Tagging

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

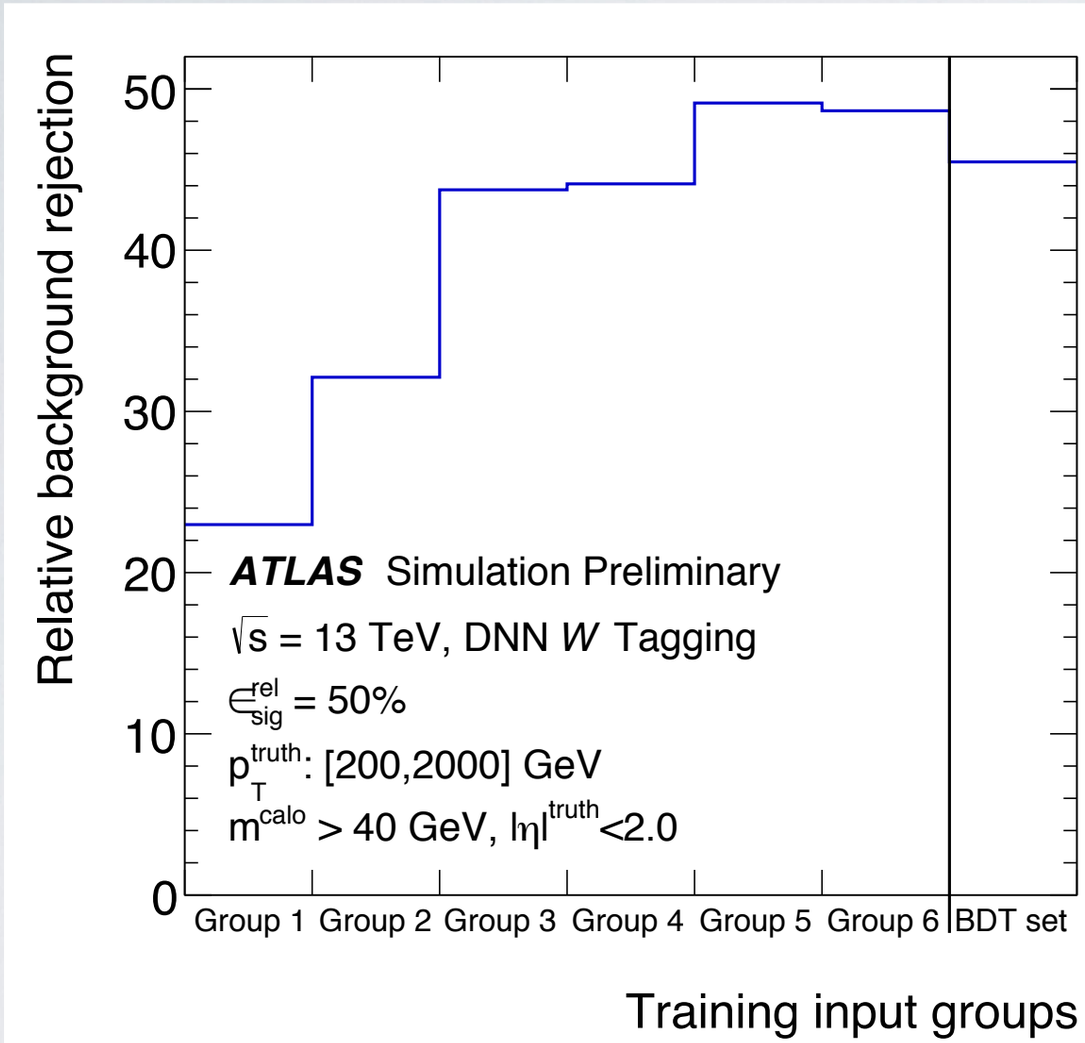
INPUTS

ATL-PHYS-PUB-2017-004

Observable	Variable	Used For	Reference
Energy Correlation Ratios	ECF_1, ECF_2, ECF_3 C_2,D_2	top,W	[26, 27]
N-subjettiness	τ_1,τ_2,τ_3 τ_{21},τ_{32}	top,W	[28, 29]
Center of Mass Observables	Fox Wolfram (R_2^{FW}) Sphericity (S) Thrust (T_{MIN},T_{MAJ})	W W W	[30] [31] [32]
Splitting Measures	Z_{CUT} μ_{12} $\sqrt{d_{12}}$,$\sqrt{d_{23}}$	W W top,W	[33] [34] [35]
Planar Flow	$\mathcal{P}$	W	[36]
Dipolarity	$\mathcal{D}$	W	[37]
Angularity	a_3	W	[38]
Aplanarity	A	W	[31]
KtDR	$KtDR$	W	[39]
Qw	Q_w	top	[33]

DNN TRAINING - INPUTS OPTIMIZATION

W Tagging

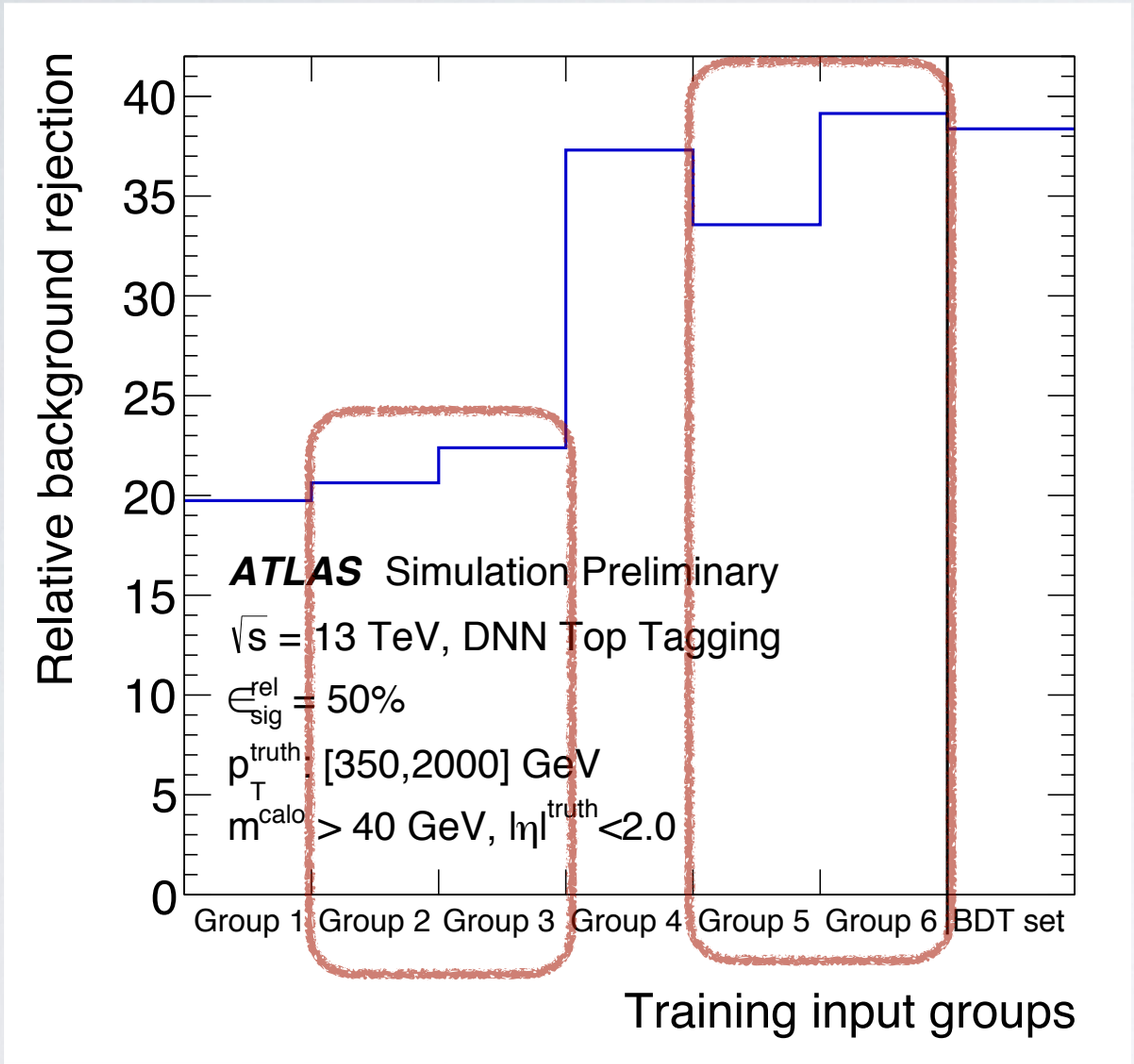


Group 5 with 18 variables

Observable	*W*-Boson Tagging Observable Groups 1	2	3	4	5	6	7 (BDT)
ECF_1			○	○	○	○	
ECF_2			○	○	○	○	○
ECF_3			○	○	○	○	○
C_2	○	○			○	○	
D_2	○	○			○	○	○
τ_1			○	○	○	○	○
τ_2			○	○	○	○	
τ_{21}	○	○			○	○	○
R_2^{FW}		○	○	○	○	○	○
S		○	○	○	○	○	○
$\mathcal{P}$					○	○	○
$\mathcal{D}$					○	○	
a_3			○	○	○	○	○
A			○	○	○	○	○
T_{MIN}		○		○		○	
T_{MAJ}		○		○		○	
Z_{CUT}					○	○	
μ_{12}					○	○	
$\sqrt{d_{12}}$	○	○	○		○	○	
$KtDR$					○	○	○

DNN TRAINING - INPUTS OPTIMIZATION

Top Tagging



Observable	Top Tagging Observable Groups 1	2	3	4	5	6	7 (BDT)
ECF_1				o		o	o
ECF_2				o		o	
ECF_3				o		o	o
C_2					o	o	o
D_2					o	o	o
τ_1		o	o	o	o	o	
τ_2		o	o	o	o	o	o
τ_3		o	o	o	o	o	
τ_{21}	o		o		o	o	o
τ_{32}	o		o		o	o	o
$\sqrt{d_{12}}$	o	o	o	o	o	o	o
$\sqrt{d_{23}}$	o	o	o	o	o	o	o
Q_w	o	o	o	o	o	o	o

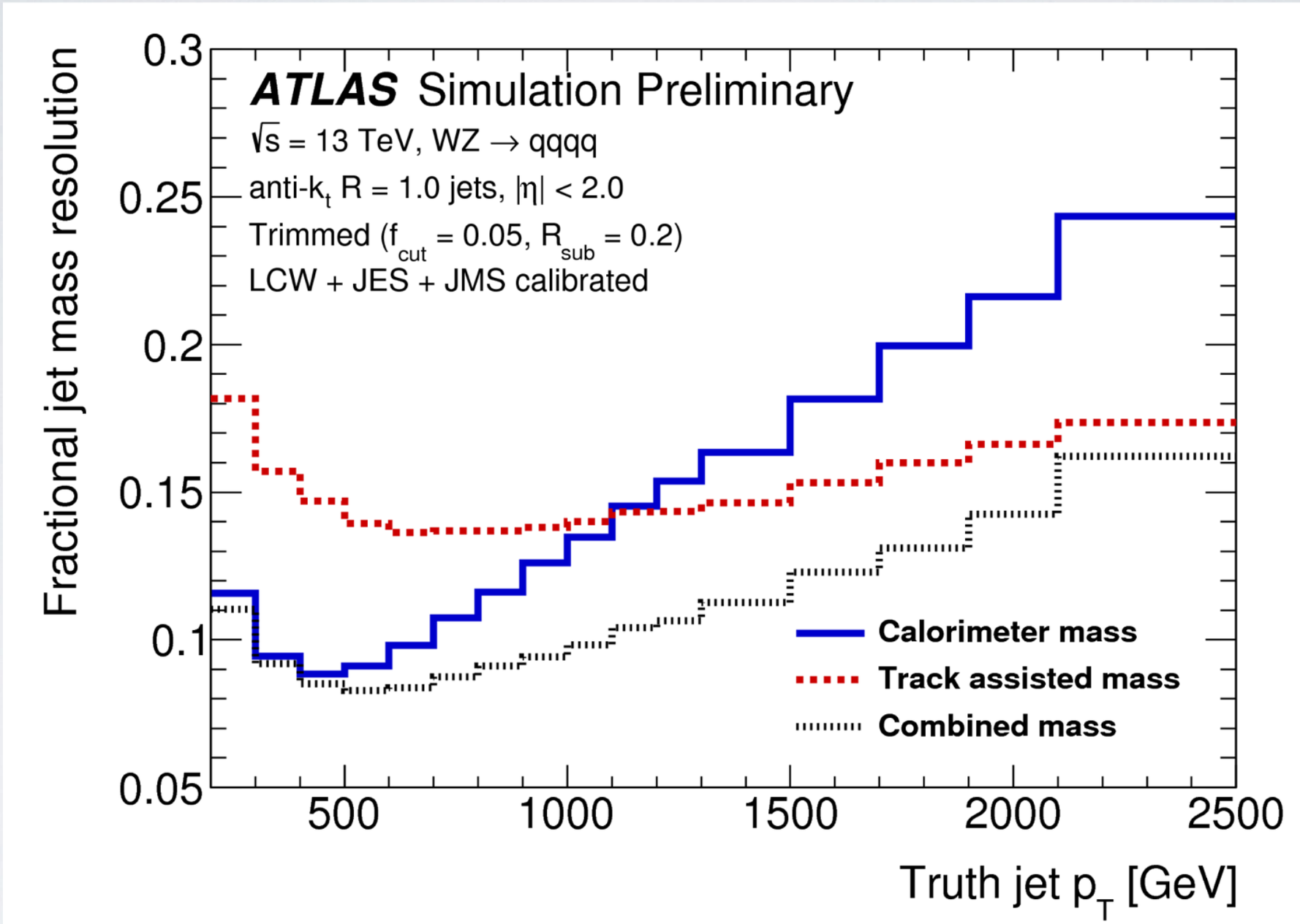
Group 6 with 13 variables

INPUTS

ATLAS-CONF-2017-064

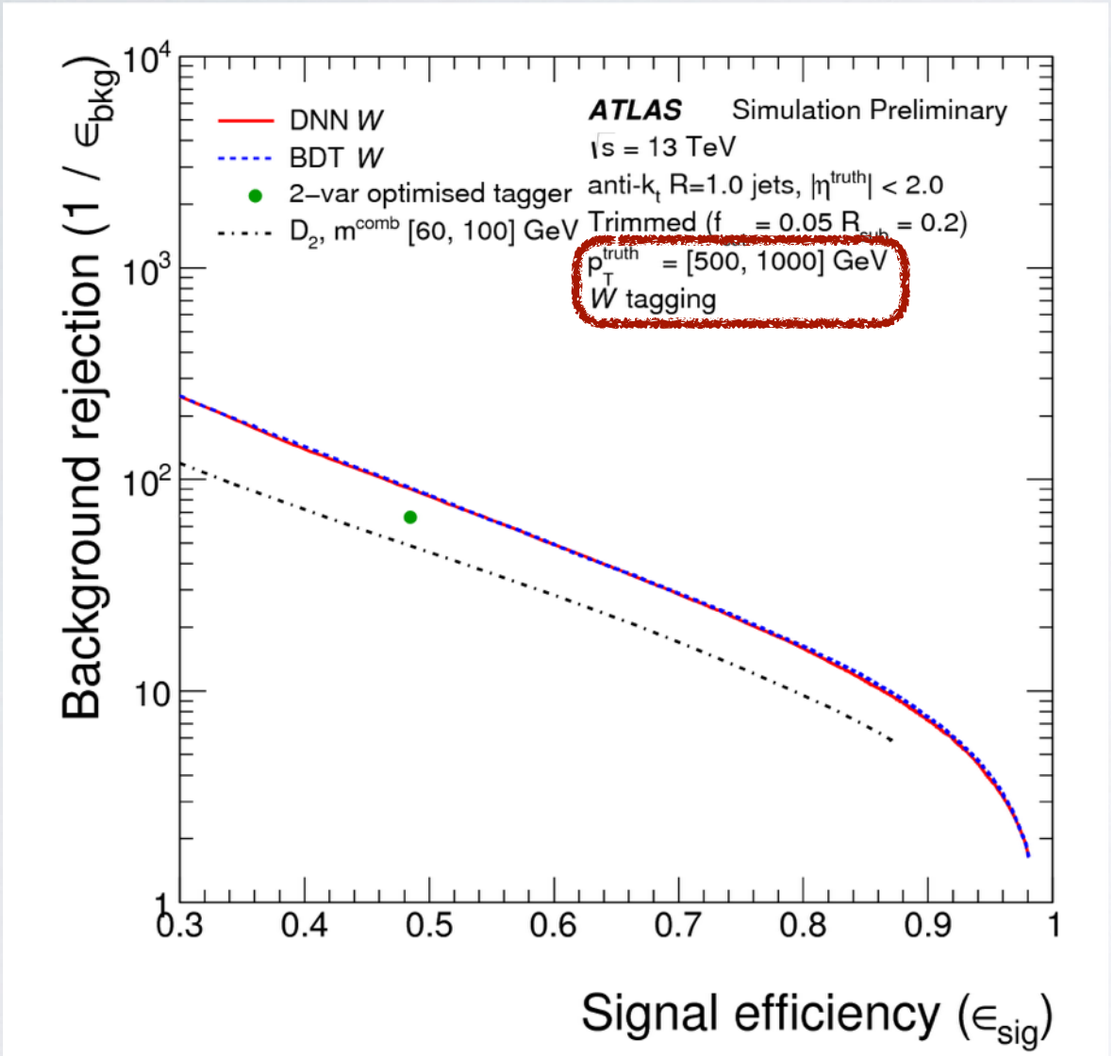
Observable	Variable	Used For	Reference
Jet mass	m^{comb}	top,W	[35]
Energy Correlation Ratios	ECF_1, ECF_2, ECF_3 C_2,D_2	top,W	[41,42]
N-subjettiness	τ_1,τ_2,τ_3 τ_{21},τ_{32}	top,W	[43,44]
Center of Mass Observables	Fox Wolfram (R_2^{FW})	W	[45]
Splitting Measures	Z_{CUT} $\sqrt{d_{12}}$,$\sqrt{d_{23}}$	W top,W	[46] [47]
Planar Flow	$\mathcal{P}$	W	[48]
Angularity	a_3	W	[49]
Aplanarity	A	W	[50]
KtDR	$KtDR$	W	[51]
Qw	Q_w	top	[46]

INPUTS - COMBINED MASS

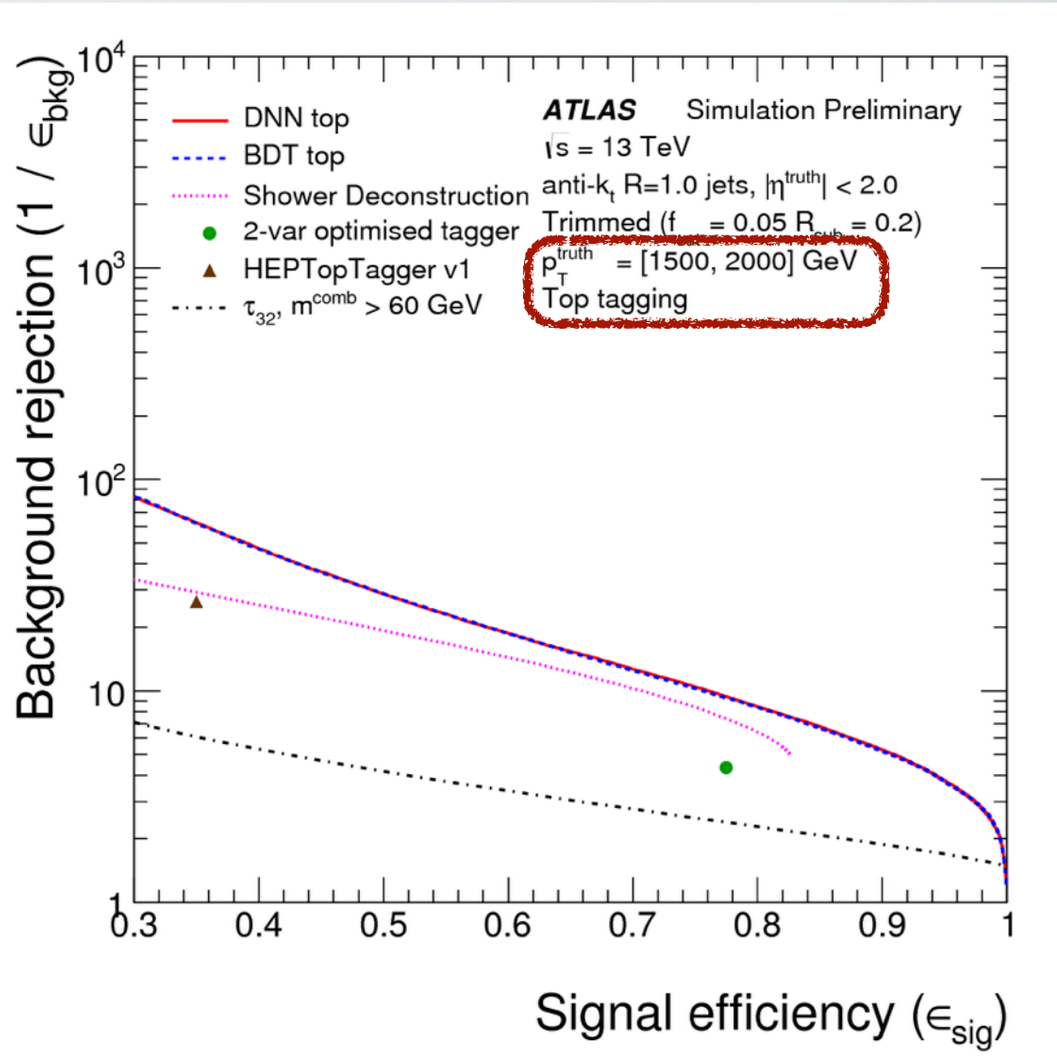


PERFORMANCE EVALUATION

W-Boson Tagging



Top-Quark Tagging

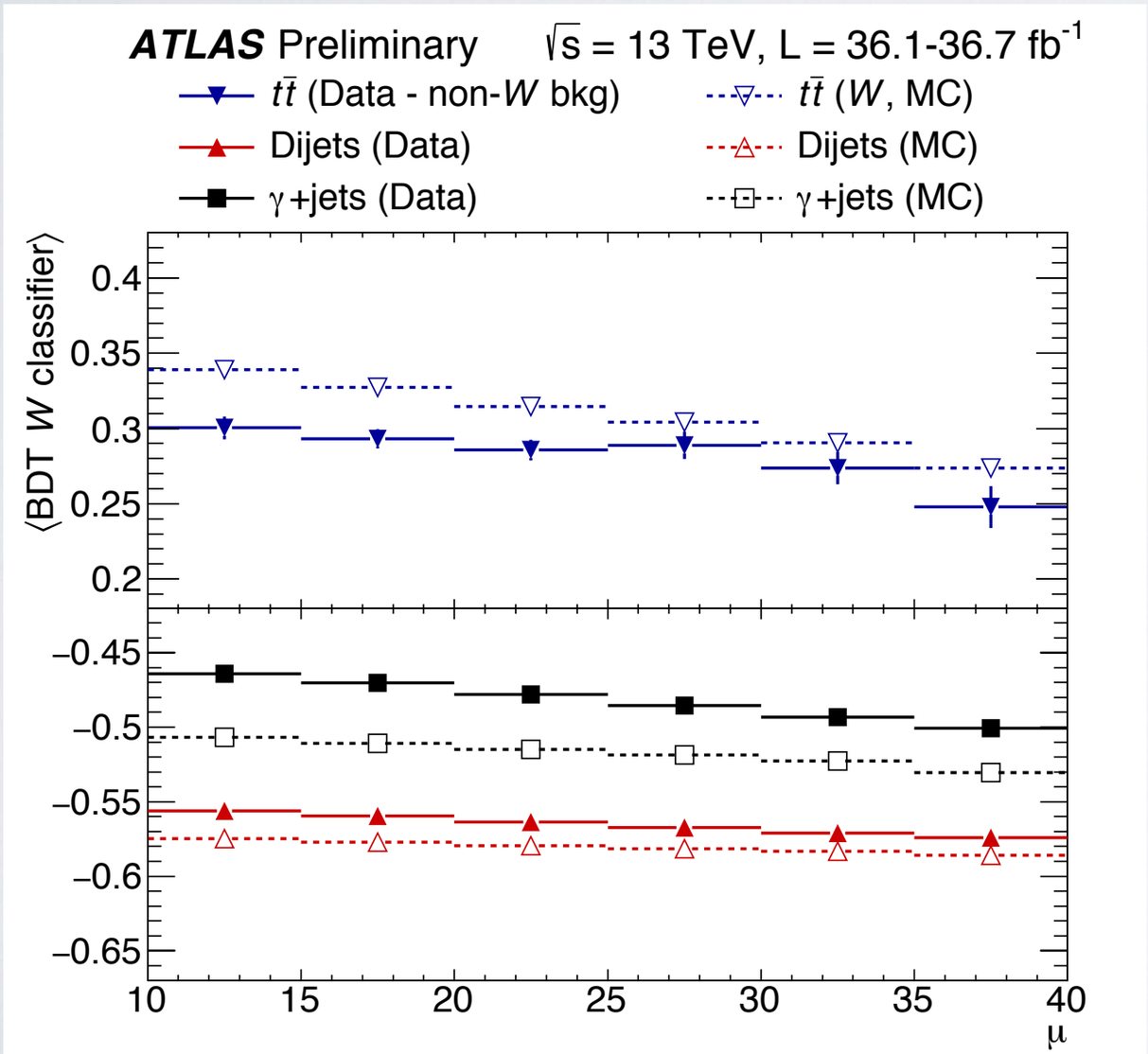


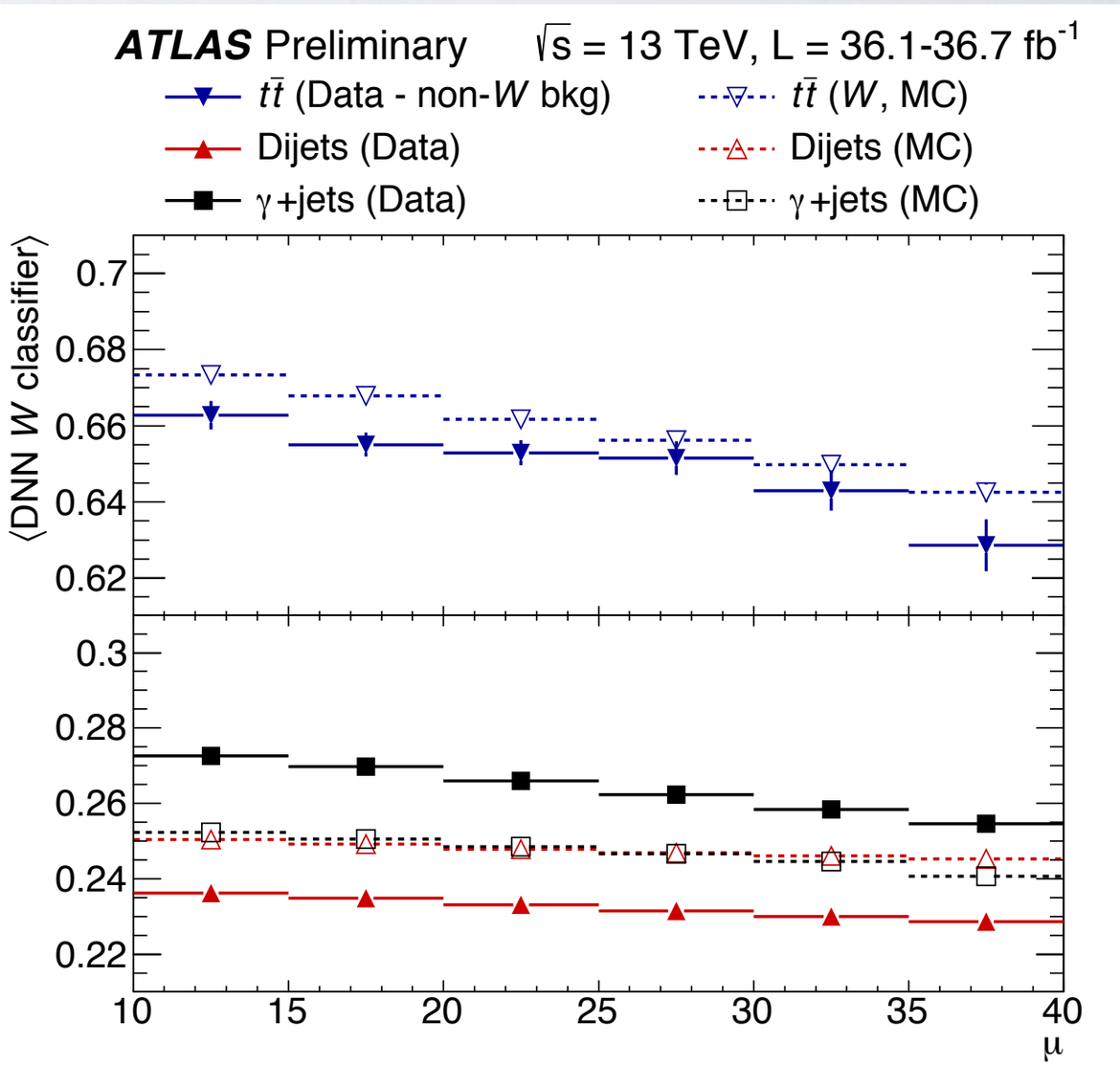
- BDT & DNN: Improvements observed for both W and top tagging
- Magnitude of improvement differs for W and top tagging, but not the overall benefit of using a BDT or DNN

PILE-UP ROBUSTNESS

Robustness against pile-up

- Further investigation and evaluation of uncertainties are pending





PILE-UP ROBUSTNESS

Robustness against pile-up

- Further investigation and evaluation of uncertainties are pending