

# Constituent-Based W-boson Tagging with the ATLAS Detector

Quantum Computing & Machine Learning Workshop, August 11-14, 2023,  
Shandong University

Shudong Wang

Institute of High Energy Physics, CAS

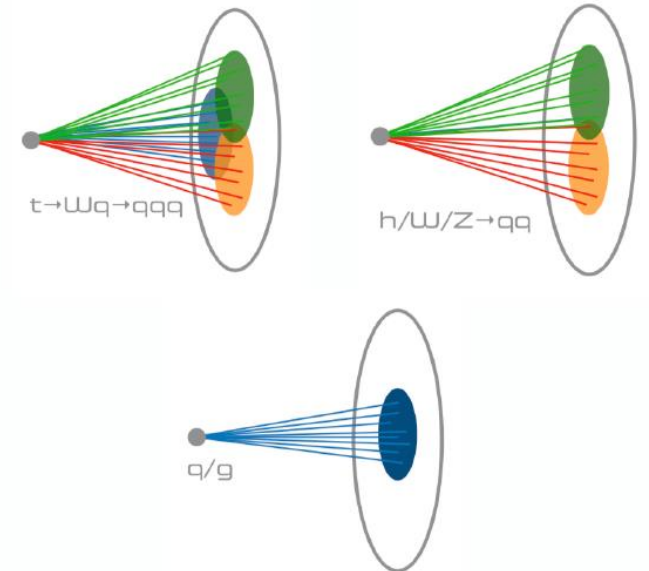
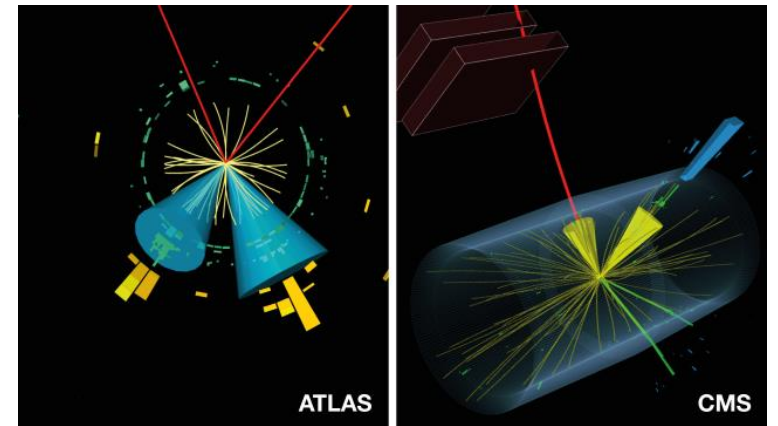
[ATL-PHYS-PUB-2023-020](#)

# Outline

- Introduction
- Monte Carlo Simulation
- Jet Reconstruction and Selection
- $W$  Jet Taggers
- Tagger Performance
- Conclusion

# Introduction

- Jets are ubiquitous at colliders, especially for hadron colliders.
- Jets are collimated sprays of particles initiated by quarks or gluons.
- Jet tagging: identifying the hard scattering particle that initiates the jet.
- High-energy particle collisions such as those produced in the Large Hadron Collider (LHC) can lead to the production of massive particles (*e.g.*  $W/Z/H$  bosons and top quarks) with much larger transverse momentum ( $p_T$ ) than rest mass.
- The decay products of such particles tend to be collimated, or ‘*boosted*’, along the direction of the progenitor particle.
- If the massive particles are sufficiently boosted, their overlapping hadronic decay products cannot be well-reconstructed with small-radius jets, and require large-radius (large-R) jet reconstruction.



*image credit*

# Introduction

- The identification of hadronically-boosted  $W$  boson decays with large- $R$  jets is vital in many physics analyses at the LHC.
- Constituent-based taggers, e.g. ParticleNet showed impressive performance and is used as official tagger by CMS. In recent ATLAS result, it is also shown as the best performance top tagger [[ATL-PHYS-PUB-2022-039](#)]. It is then natural to study  $W$  tagging with constituent-based taggers, and compare with other  $W$  jet tagger candidates.
- Conventional methods for  $W$  jet tagging:
  - Derive a set of high-level variables which describe jets, and use these variables to perform cut based tagging or ML (e.g. BDT) based tagging.
  - The construction of these variables is almost always accompanied by information loss.
- Constituent based  $W$  jet tagging:
  - Try to maximize the use of the jet constituents' information using state-of-the-art ML/DL algorithms.
- This work: test the performance of constituent-based  $W$  taggers.

# Monte Carlo Simulation

- MC samples: we follow previous note about W/Z tagging using UFO jets [[ATL-PHYS-PUB-2021-029](#)].
- For tagger training and evaluating
  - Signal: W bosons from simulated  $W' \rightarrow WZ (\rightarrow q\bar{q}q\bar{q})$  events with  $m_{W'} = 2$  TeV, Pythia8 + NNPDF2.3LO + A14 tune.
  - Background: QCD di-jet events @ LO, Pythia8 + NNPDF2.3LO + A14 tune.
- For model dependence study:
  - Background: QCD di-jet events generated using
    - Sherpa:
      - default  $p_T$ -ordered showering algorithm
      - cluster-based hadronization model & Lund string hadronization model
    - Herwig:
      - angle-ordered parton shower & dipole parton shower
      - cluster hadronization

# Jet Reconstruction and Selection

- Unified Flow Objects (UFOs) are jet input objects optimized for reconstructing large-R jets by making use of different ATLAS sub-systems in different kinematic ranges.
- Large-R jets are reconstructed from UFOs using anti- $k_t$  algorithm with the radius parameter  $R = 1.0$ .
- Both leading and sub-leading jets in an event are used.
- Jets reconstruction, grooming, truth labeling are identical to the previous work [[ATL-PHYS-PUB-2021-029](#)]

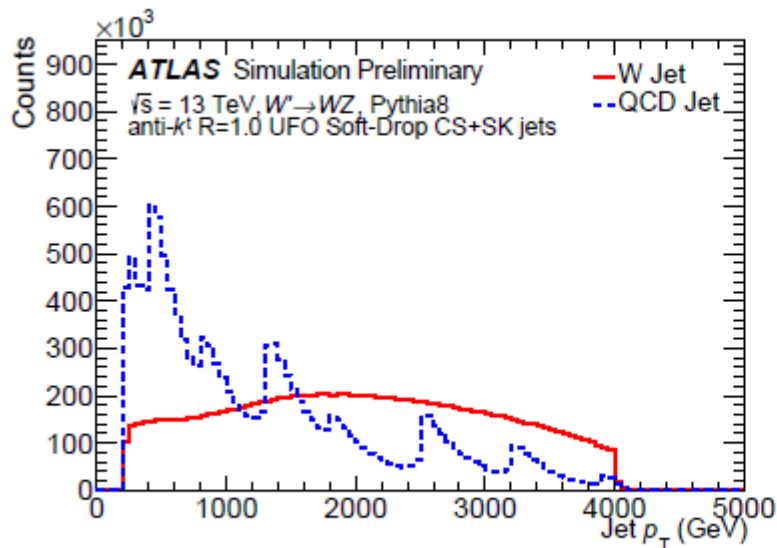
Jet Requirements	$W$ -jet requirements
Detector-level jet $ \eta  < 2.0$	$dR(\text{particle-level jet, particle-level } W) < 0.75$
Detector-level jet $p_T > 200 \text{ GeV}$	Ungroomed particle-level jet mass $> 50 \text{ GeV}$
Detector-level jet mass $> 40 \text{ GeV}$	Number of ghost associated $b$ -hadrons = 0
Number of constituents $\geq 2$	$\sqrt{d_{12}} > 55.25 \times \exp(-2.34 \times 10^{-3} \times \text{particle-level jet } p_T)$
$dR(\text{detector-level jet, particle-level jet}) < 0.75$	

Table 1: A summary of the requirements applied on the detector-level and particle-level jets in the simulation samples to produce the training and testing sets. The additional  $W$ -jet requirements constitute the truth labeling strategy, and are only applied to the signal sample of simulated  $W$ .

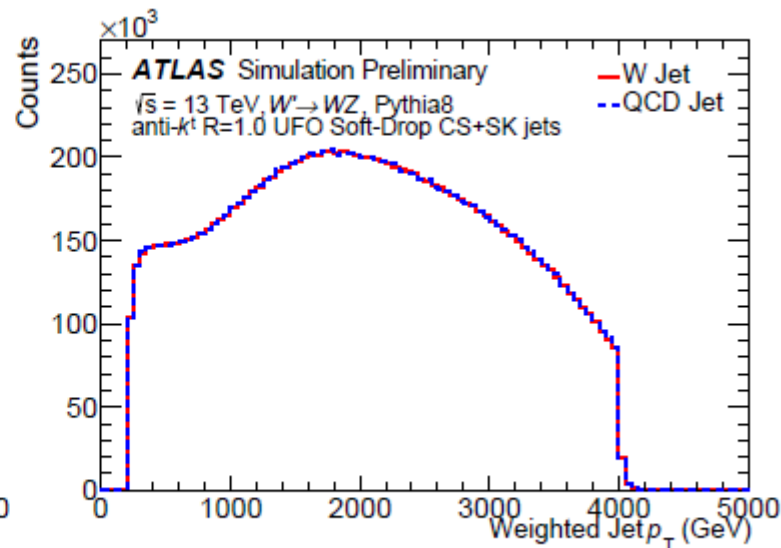
# Jet Reconstruction and Selection

## Jet $p_T$ and Training Weights

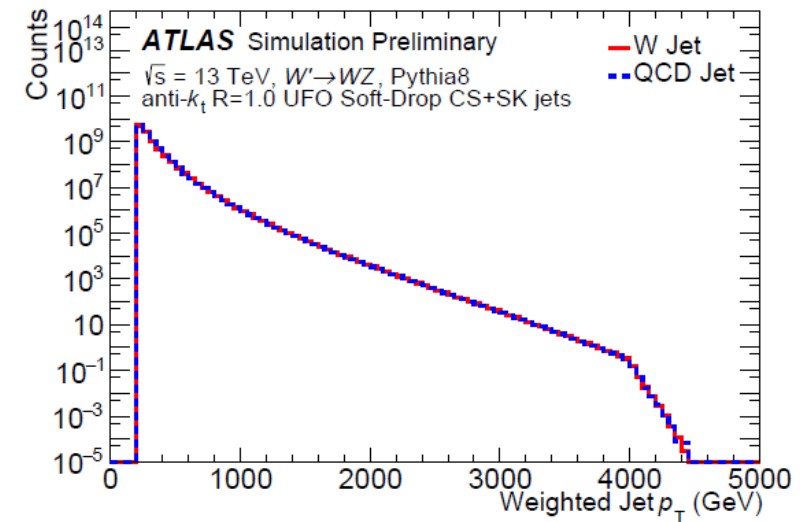
- Raw QCD jet sample contains unphysical  $p_T$  spectrum. A falling jet  $p_T$  spectrum can be obtained by applying physical weights.
- Reweight the background events to match background  $p_T$  spectrum to signal :
  - Prevents the tagger from associating signal jets with a particular  $p_T$ .
  - Helps the tagger learn to correctly classify jets across the whole  $p_T$  range.



(a)



(b)



(c)

Figure 1: The jet  $p_T$  spectrum for signal and background, without weights (a), after applying the training weights (b) and after applying the testing weights (c).

# Jet Reconstruction and Selection

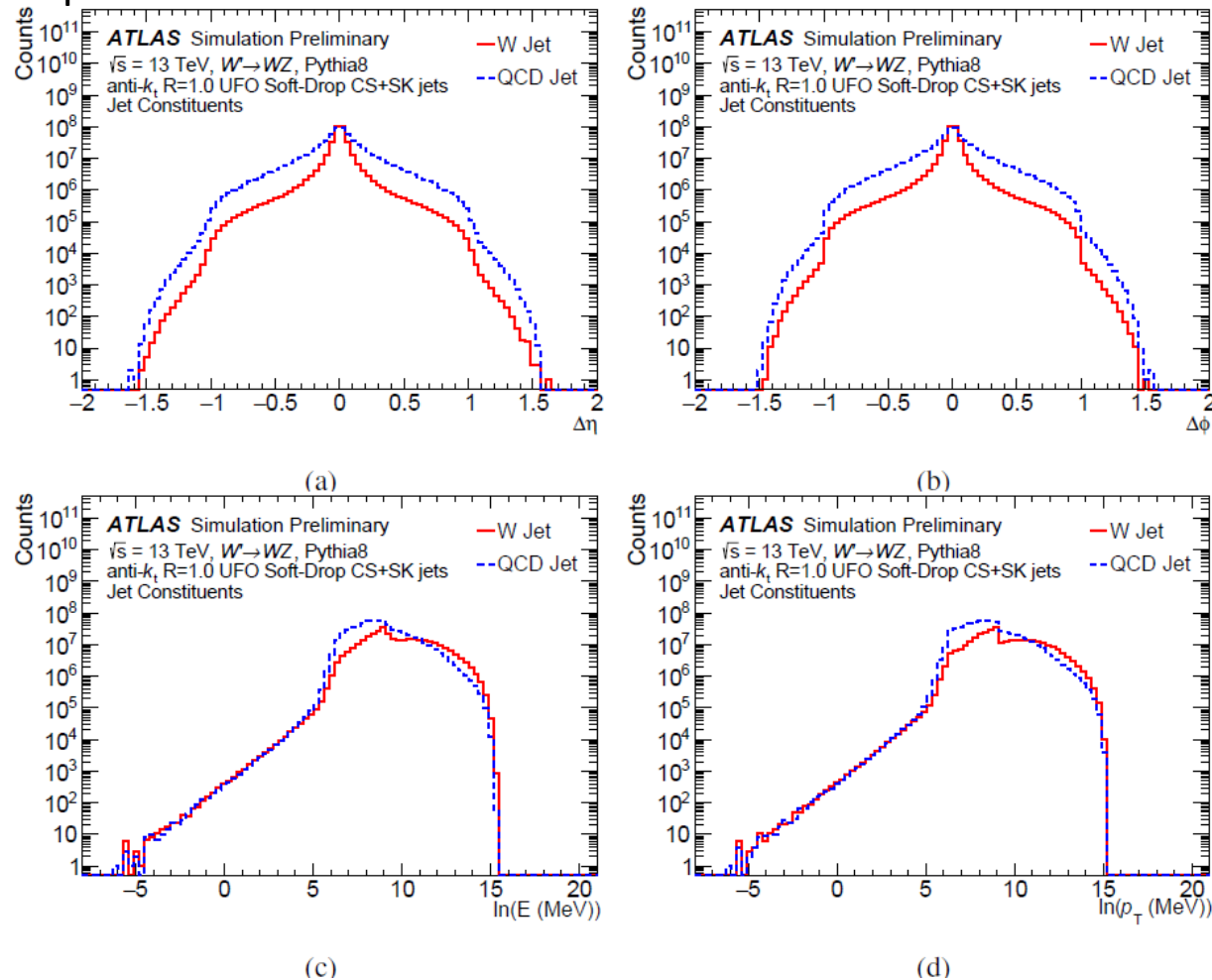
## Data Pre-processing

- The dataset used for tagger optimization consists of 25 million jets, half of the jets are signal ( $W$  jets) and half are background (QCD jets). It is split into orthogonal training, validating and testing datasets by a ratio of 6:2:2.
- To facilitate tagger training, input quantities are pre-processed to fit into a relatively reasonable numerical range, eliminate irrelevant features and capitalize on well-known symmetries.
- Constituent level inputs:
  - $\Delta\eta$  Difference in pseudorapidity between the jet constituents and the jet axis
  - $\Delta\phi$  Difference in azimuthal angle between the jet constituents and the jet axis
  - $\ln p_T$  Logarithm of the jet constituents'  $p_T$
  - $\ln E$  Logarithm of the jet constituents' energy
  - $\ln \frac{p_T}{\sum_{\text{jet}} p_T}$  Logarithm of the jet constituents'  $p_T$  relative to the total  $p_T$  in jet
  - $\ln \frac{E}{\sum_{\text{jet}} E}$  Logarithm of the jet constituents' energy relative to the total energy in jet
  - $\Delta R$  Angular separation between the jet constituent and the jet axis  $\sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$
  - $(E, p_x, p_y, p_z)$  4-momentum of jet constituent in this certain form.

# Jet Reconstruction and Selection

## Data Pre-processing

- Distributions of input features

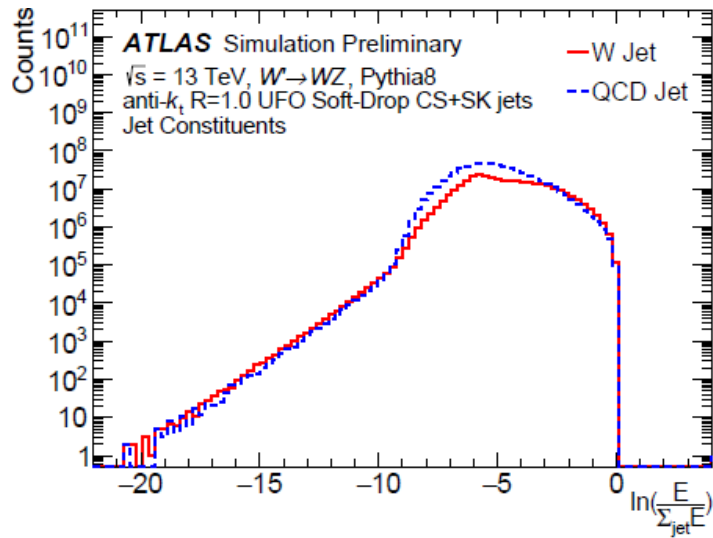


Distributions of the seven constituent-level quantities used as inputs to the  $W$  tagger training.

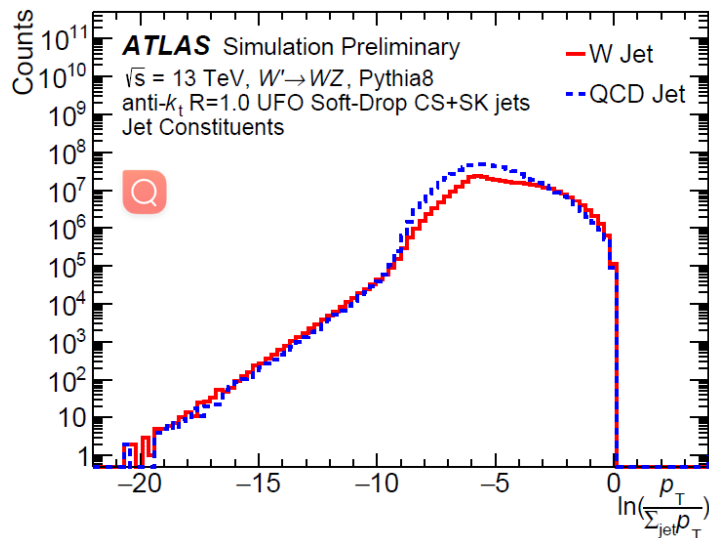
# Jet Reconstruction and Selection

## Data Pre-processing

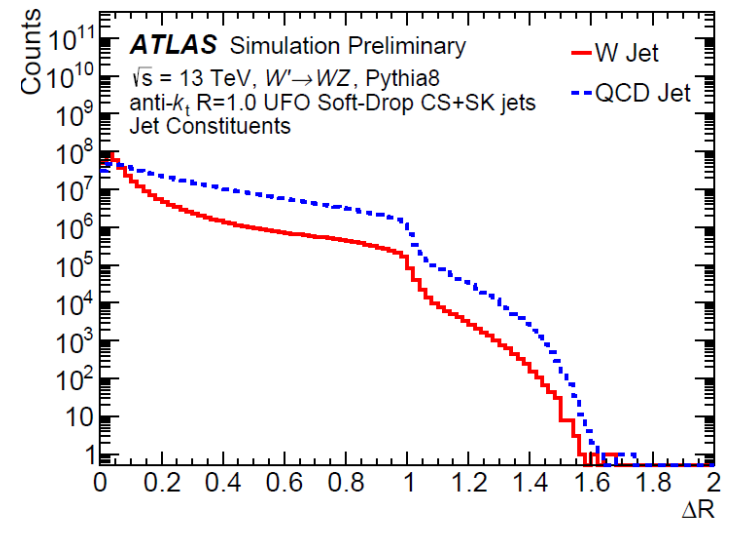
- Distributions of input features



(e)



(f)

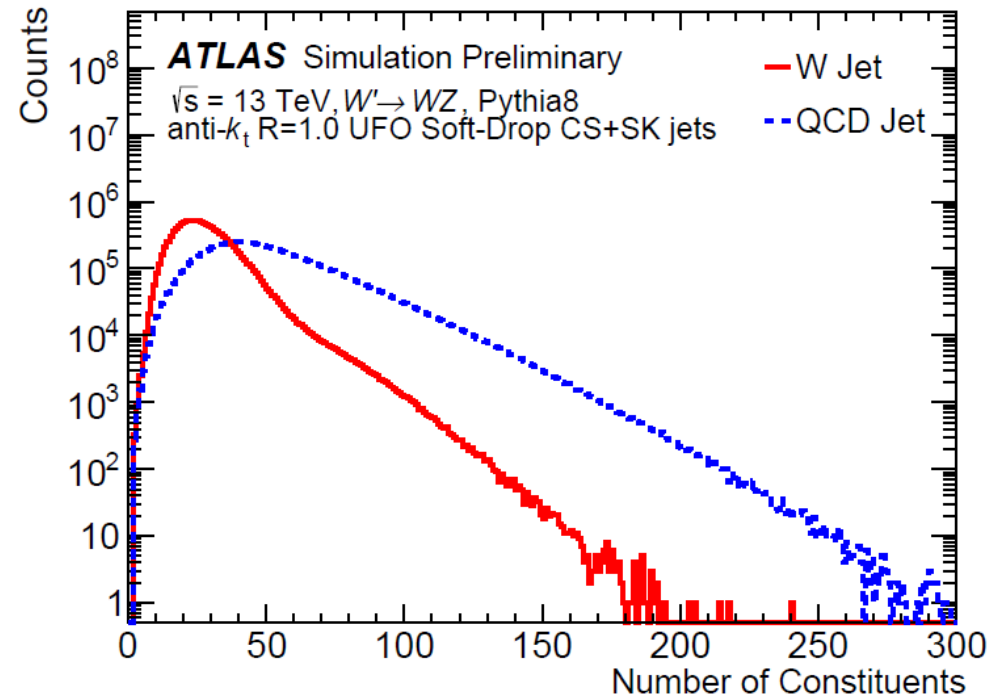


(g)

Distributions of the seven constituent-level quantities used as inputs to the  $W$  tagger training.

# W Jet Taggers

- In this study, a maximum of 200 constituents are considered by all constituent-based taggers. Only a small portion of jets in the dataset have more than 200 constituents (less than 0.04%). As jet constituents are sorted by decreasing  $p_T$ , truncation eliminates the softest constituents of the jet.



Distributions of the number of constituents in a large- $R$  jet.

# W Jet Taggers

- Particle Flow Network(PFN)/Energy Flow Network(EFN)
  - Based on Deep Sets Theorem
  - [JHEP01\(2019\)121](#)
- ParticleNet
  - Customized graph neural network architecture for jet tagging with the point cloud approach
  - [Phys.Rev.D 101 \(2020\) 5, 056019](#)
- ParticleTransformer
  - Transformer designed for particle physics
  - [arxiv: 2202.03772](#)
- All models trained to minimize cross entropy loss with Ranger optimizer.

Models	Input variables
EFN	$\Delta\eta, \Delta\phi, \ln p_T$
PFN	$\Delta\eta, \Delta\phi, \ln p_T, \ln E, \ln \frac{p_T}{\sum_{jet} p_T}, \ln \frac{E}{\sum_{jet} E}, \Delta R$
ParticleNet	$\Delta\eta, \Delta\phi, \ln p_T, \ln E, \ln \frac{p_T}{\sum_{jet} p_T}, \ln \frac{E}{\sum_{jet} E}, \Delta R$
ParticleTransformer	$\Delta\eta, \Delta\phi, \ln p_T, \ln E, \ln \frac{p_T}{\sum_{jet} p_T}, \ln \frac{E}{\sum_{jet} E}, \Delta R$ $(E, p_x, p_y, p_z)$

Input features used in each tagger.

# W Jet Taggers

- **Energy Flow Network (EFN) / Particle Flow Network (PFN)**

Jet: An *unordered, variable length* set of particles

- **Deep Sets [1703.06114]**

- Namespace for symmetric function parametrization
- A general permutation-symmetric function is additive in a latent space

Permutation invariance

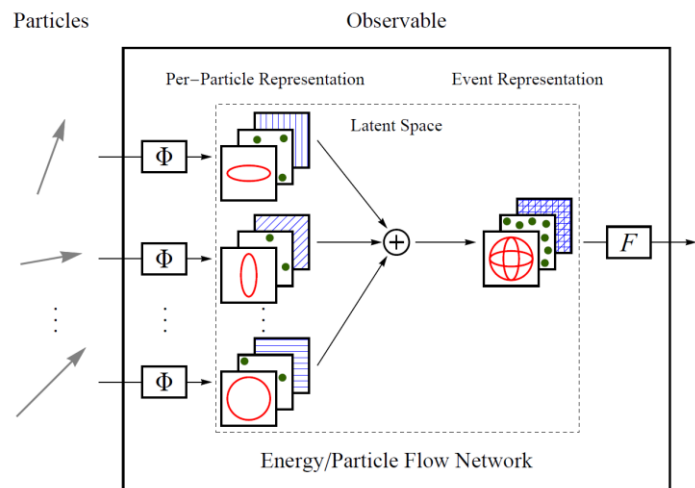
- **EFN (Energy Flow Network) / PFN (Particle Flow Network)**

- EFN: IRC-safe latent space

$$\text{EFN: } F\left(\sum_{i=1}^M z_i \Phi(\hat{p}_i)\right)$$

- PFN: Fully general latent space

$$\text{PFN: } F\left(\sum_{i=1}^M \Phi(p_i)\right)$$



P. T. Komiske, E. M. Metodiev and J. Thaler [[JHEP01\(2019\)121](#)]

Feature space

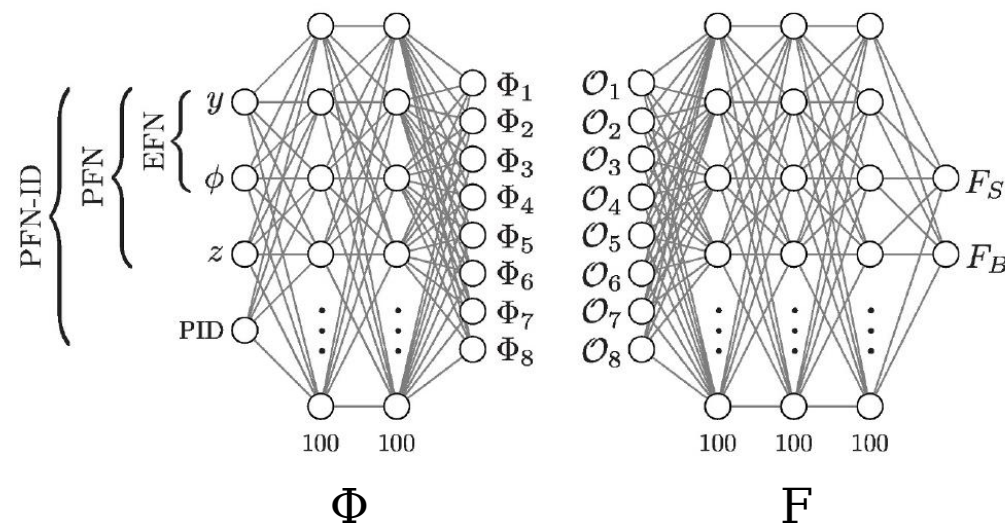
Variable length

Deep Sets Theorem [63]. Let  $\mathfrak{X} \subset \mathbb{R}^d$  be compact,  $X \subset 2^{\mathfrak{X}}$  be the space of sets with bounded cardinality of elements in  $\mathfrak{X}$ , and  $Y \subset \mathbb{R}$  be a bounded interval. Consider a continuous function  $f : X \rightarrow Y$  that is invariant under permutations of its inputs, i.e.  $f(x_1, \dots, x_M) = f(x_{\pi(1)}, \dots, x_{\pi(M)})$  for all  $x_i \in \mathfrak{X}$  and  $\pi \in S_M$ . Then there exists a sufficiently large integer  $\ell$  and continuous functions  $\Phi : \mathfrak{X} \rightarrow \mathbb{R}^\ell$ ,  $F : \mathbb{R}^\ell \rightarrow Y$  such that the following holds to an arbitrarily good approximation:<sup>1</sup>

$$f(\{x_1, \dots, x_M\}) = F\left(\sum_{i=1}^M \Phi(x_i)\right)$$

Latent space

General parametrization for a function of sets



This page is excerpted from P. T. Komiske's talk

# W Jet Taggers

- **The architecture of ParticleNet**

- **ParticleNet**

- customized graph neural network architecture for jet tagging with the point cloud approach, based on Dynamic Graph CNN (DGCNN) [Y. Wang et al., [arXiv:1801.07829](https://arxiv.org/abs/1801.07829)]
- explicitly respects the permutation symmetry of the point cloud

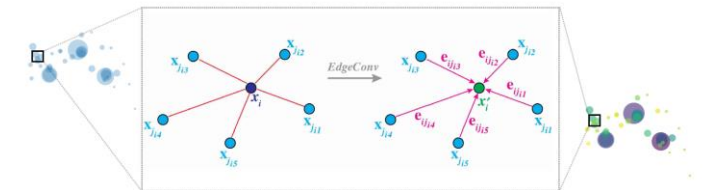
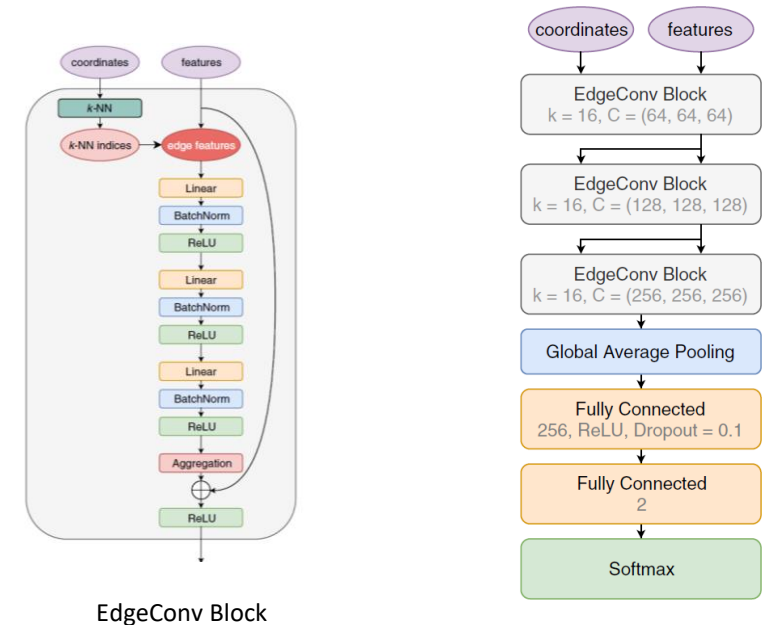
- **Key building block: EdgeConv**

- treating a point cloud as a graph: each point is a vertex
  - for each point, a local patch is defined by finding its k-nearest neighbors
  - designing a permutation-invariant “convolution” function
    - define “edge feature” for each center-neighbor pair:  $e_{ij} = h_{\Theta}(x_i, x_{ij}) = \bar{h}_{\Theta}(x_i, x_{ij} - x_i)$ 
      - same  $h_{\Theta}$  for all neighbor points, and all center points, for symmetry
    - aggregate the edge features in a symmetric way:  $x'_i = \square_{j=1}^k h_{\Theta}(x_i, x_{ij}) = \frac{1}{k} \sum h_{\Theta}(x_i, x_{ij})$

- **EdgeConv can be stacked to form a deep network**

- learning both local and global structures, in a hierarchical way

H. Qu and L. Gouskos [[Phys.Rev.D 101 \(2020\) 5, 056019](https://arxiv.org/abs/2005.05601)]



# W Jet Taggers

- The architecture of ParticleTransformer

H. Qu , C. Li, S. Qian [[2202.03772](#)]

- ParticleTransformer

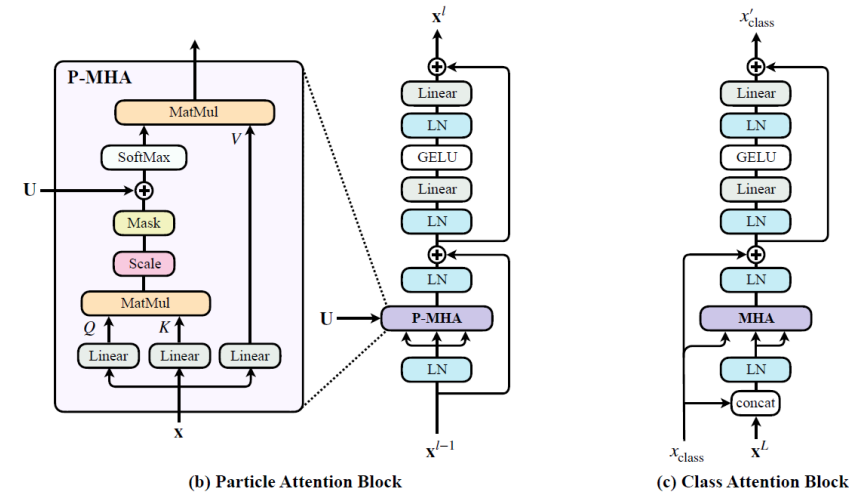
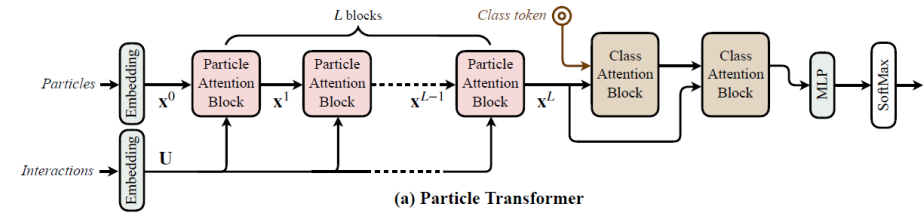
- Transformer designed for particle physics
- TWO sets of inputs
  - *Particle*: Features of every single particle
  - *Interaction*: Pair-wise features

- Particle Attention Block

- Multi-Head Attention (MHA) Module
- Pair-wise feature are introduced as the attention mask (P-MHA)

- Class Attention Block

- Multi-Head Attention (MHA) Module
- Class token is used for the MHA calculation



$$P\text{-MHA}(Q, K, V) = \text{SoftMax}(QK^T / \sqrt{d_k} + \mathbf{U})V$$

Choice of the pair-wise features: from LundNet

$$\Delta = \sqrt{(y_a - y_b)^2 + (\phi_a - \phi_b)^2}$$

$$k_T = \min(p_{T,a}, p_{T,b}) \cdot \Delta$$

$$z = \min(p_{T,a}, p_{T,b}) / (p_{T,a} + p_{T,b})$$

$$m^2 = (E_a + E_b)^2 - |\mathbf{p}_a + \mathbf{p}_b|^2$$

$d_k$ : dimension of  $K$

$$MHA_C(Q_C, K_C, V_C) = \text{SoftMax}(Q_C K_C^T / \sqrt{d_{kC}}) V_C$$

$$Q_C = W_{qC} x_{\text{class}} + b_{qC} \quad K_C = W_{kC} \mathbf{z} + b_{kC} \quad V_C = W_{vC} \mathbf{z} + b_{vC} \quad d_{kC}: \text{dimension of } K_C$$

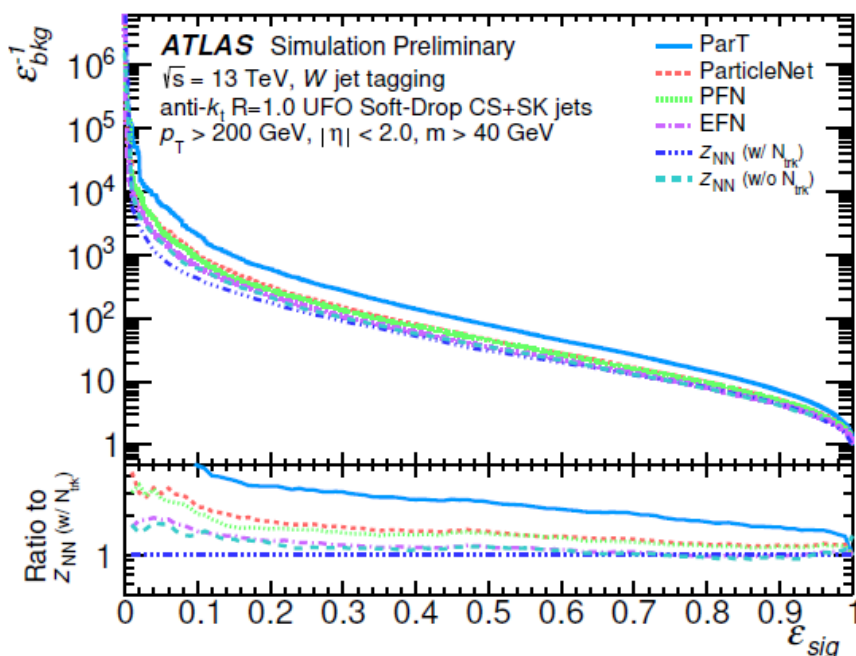
$\mathbf{z} = [x_{\text{class}}, \mathbf{x}^L]$

Concatenate class information and particle embedding

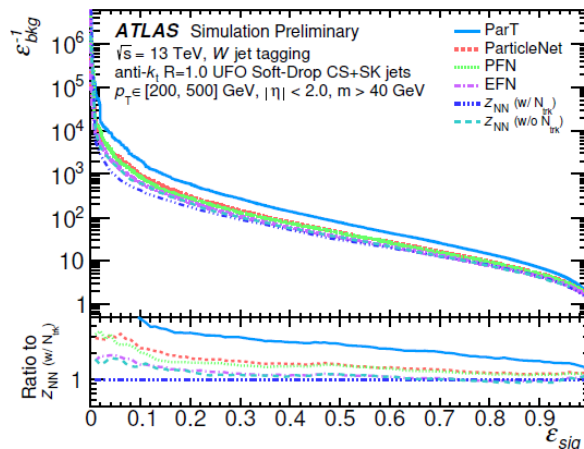
# Tagger Performance

## Comparison of different taggers

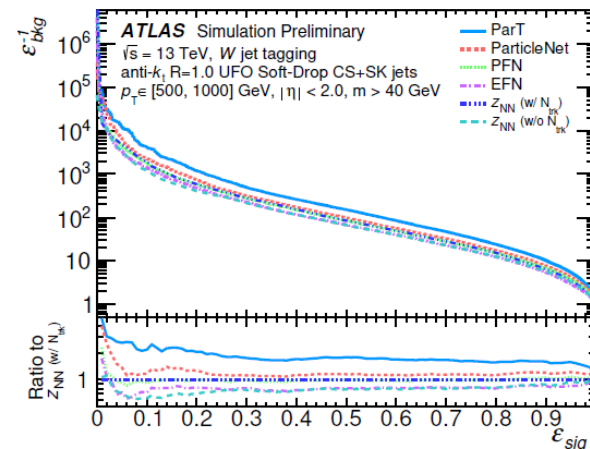
Calculated using samples with steeply falling pT spectra, i.e. both sig & bkg are weighted to have falling pT spectra.



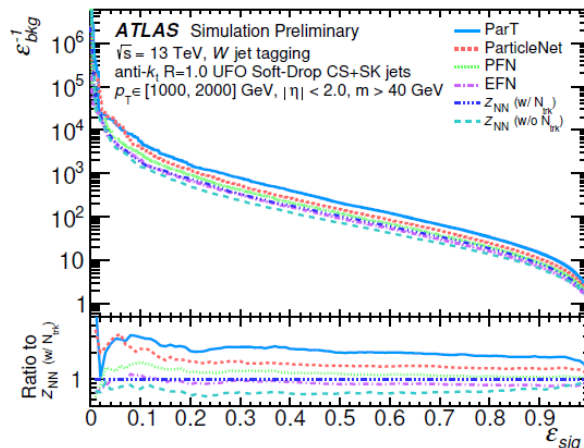
(a)



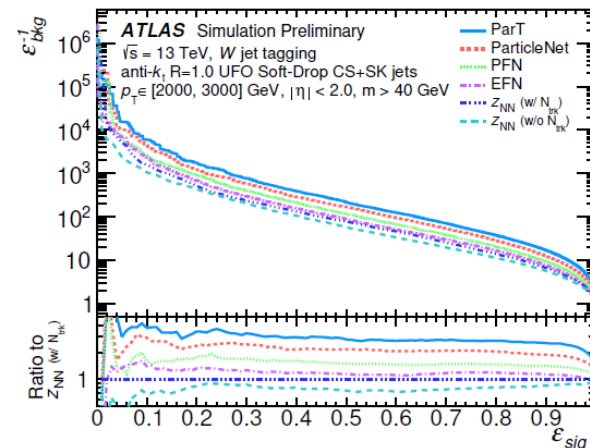
(b)



(c)



(d)



(e)

For a signal efficiency of 0.5 (0.8) case, the background rejection of ParticleTransformer is about 1.8-2.8 (1.6-2.7) times better than the baseline tagger.

Figure 3: The QCD jets background rejection ( $\epsilon_{bkg}^{-1}$ ) versus the  $W$ -jets signal efficiency ( $\epsilon_{sig}$ ) for all the taggers studied. All of the constituent-based taggers studied surpass the performance of the high-level-feature-based tagger (noted as  $Z_{NN}$  in the figure) in the previous study [52].

# Tagger Performance

## Comparison of different taggers

Calculated using samples with steeply falling pT spectra

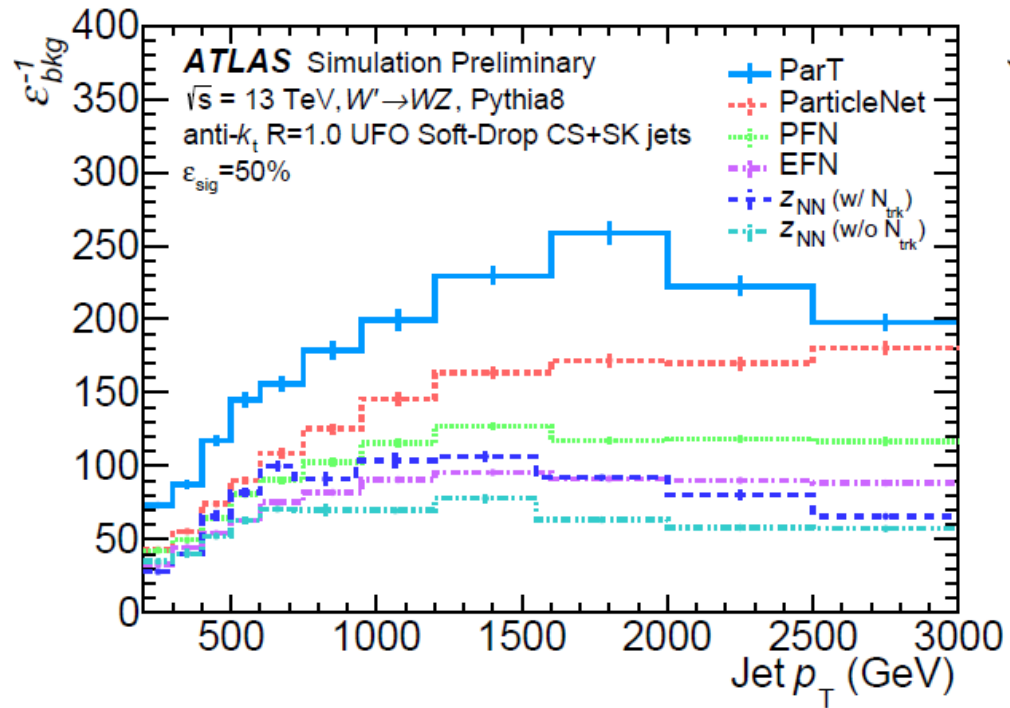
Model	AUC	ACC	$\varepsilon_{bkg}^{-1}$ @ $\varepsilon_{sig} = 0.5$	$\varepsilon_{bkg}^{-1}$ @ $\varepsilon_{sig} = 0.8$	# Params	Inference Time
EFN	0.920	0.835	35.1	7.95	56.73k	0.065 ms
PFN	0.931	0.853	44.7	9.50	57.13k	0.11 ms
ParticleNet	0.933	0.826	46.2	9.76	366.16k	0.36 ms
ParticleTransformer	0.951	0.880	77.9	14.6	2.14M	0.28 ms

Table 3: The performance of each  $W$  jet tagger is measured with several metrics evaluated on the testing set.

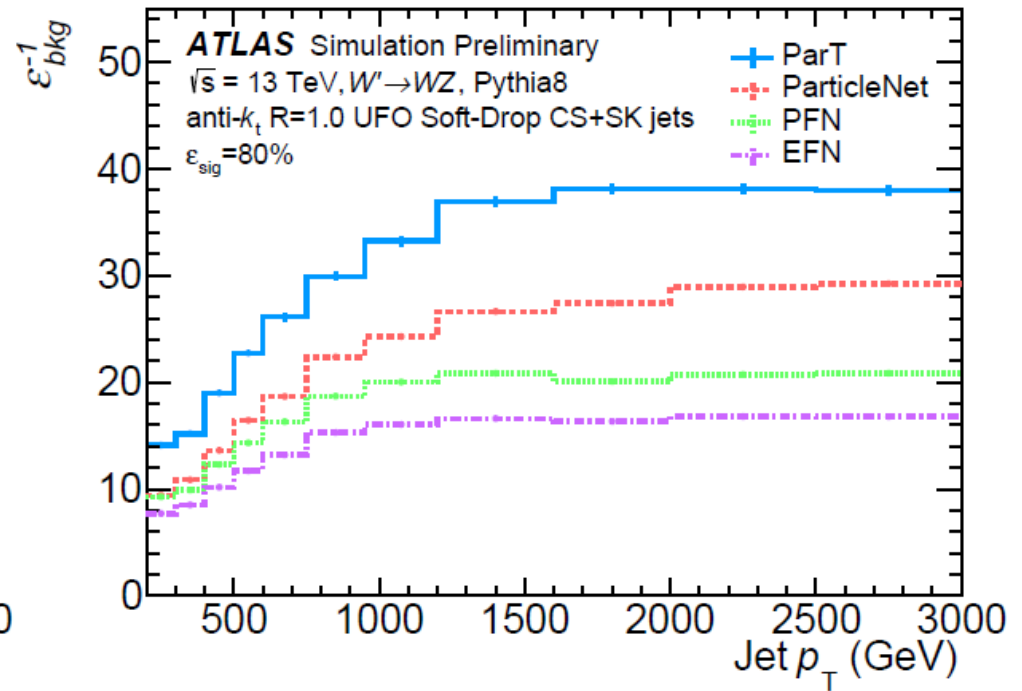
# Tagger Performance

## Comparison of different taggers

Calculated using samples with steeply falling pT spectra



(a)



(b)

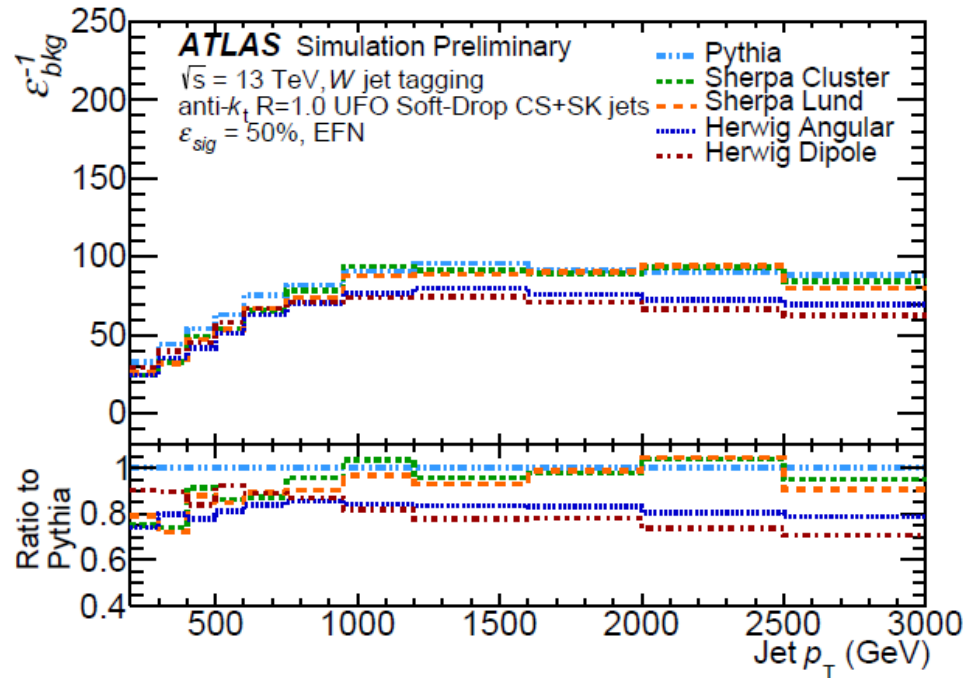
Figure 4: Background rejection ( $\epsilon_{bkg}^{-1}$ ) as a function of the jet  $p_T$  of studied  $W$  taggers for  $\epsilon_{sig} = 0.5$  (a) and  $\epsilon_{sig} = 0.8$  (b) working points.

# Tagger Performance

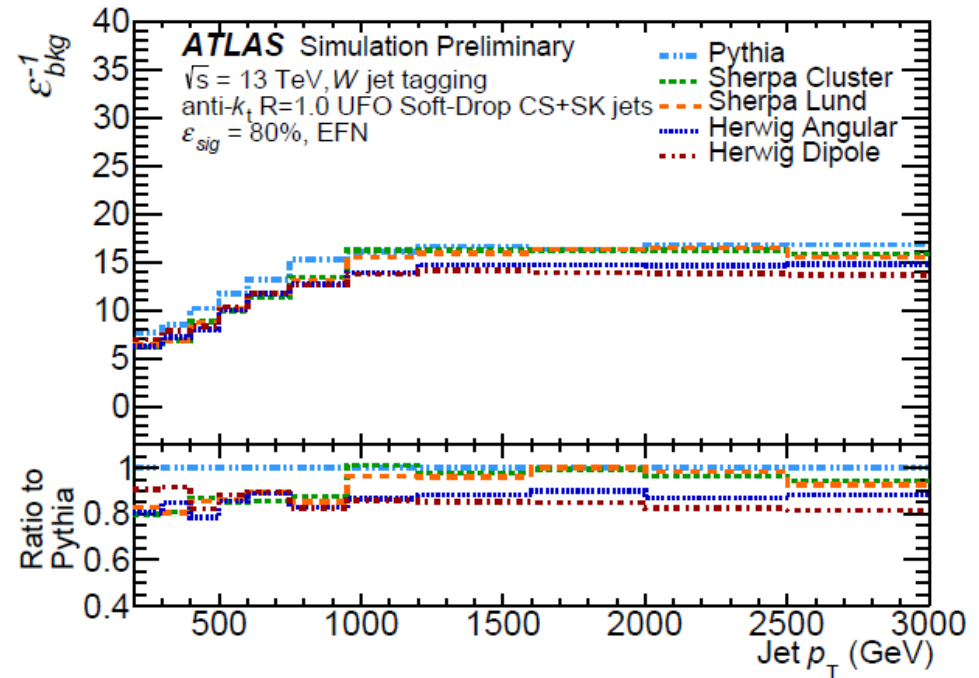
## Model dependence

- To estimate the dependence of tagger performance on physics modeling of the parton shower and hadronization, taggers are evaluated on alternative background samples. Sherpa and Herwig models are used to evaluate the dependence on the modeling of hadronization and parton showering, respectively.

EFN:



(a)



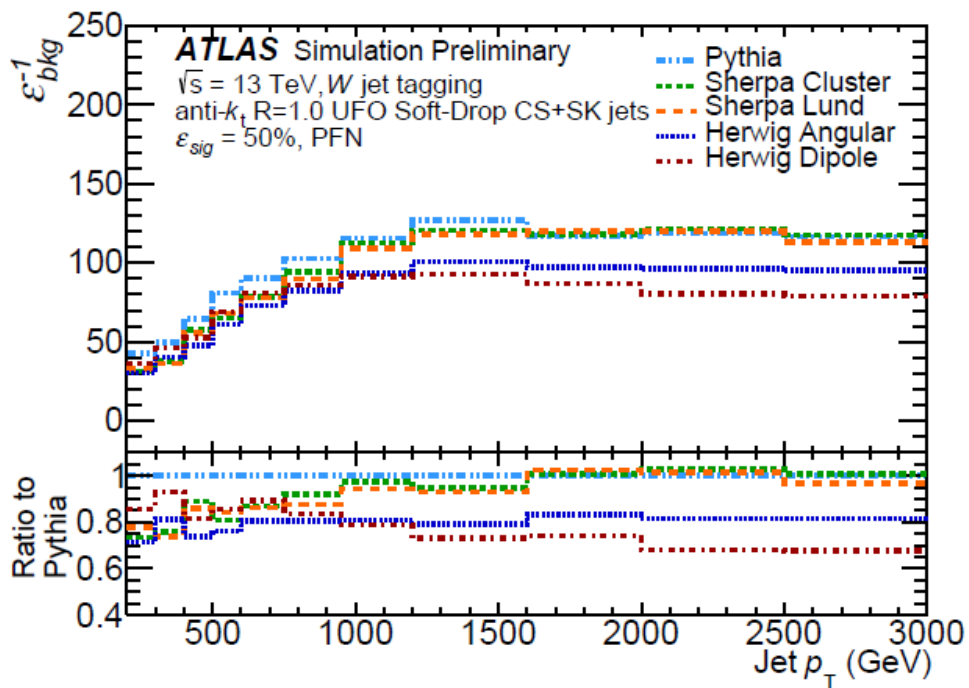
(b)

Figure 5: Comparison of the background rejection ( $\epsilon_{bkg}^{-1}$ ) of  $W$  taggers in different samples of simulated QCD jet, as a measure of model dependence. Shown is the background rejection using the threshold which results in an signal efficiency of 50% (a,c) or 80% (b,d) in each  $p_T$  bin for  $W' \rightarrow WZ$  testing sample. The top (bottom) row stands for the EFN (PFN) tagger.

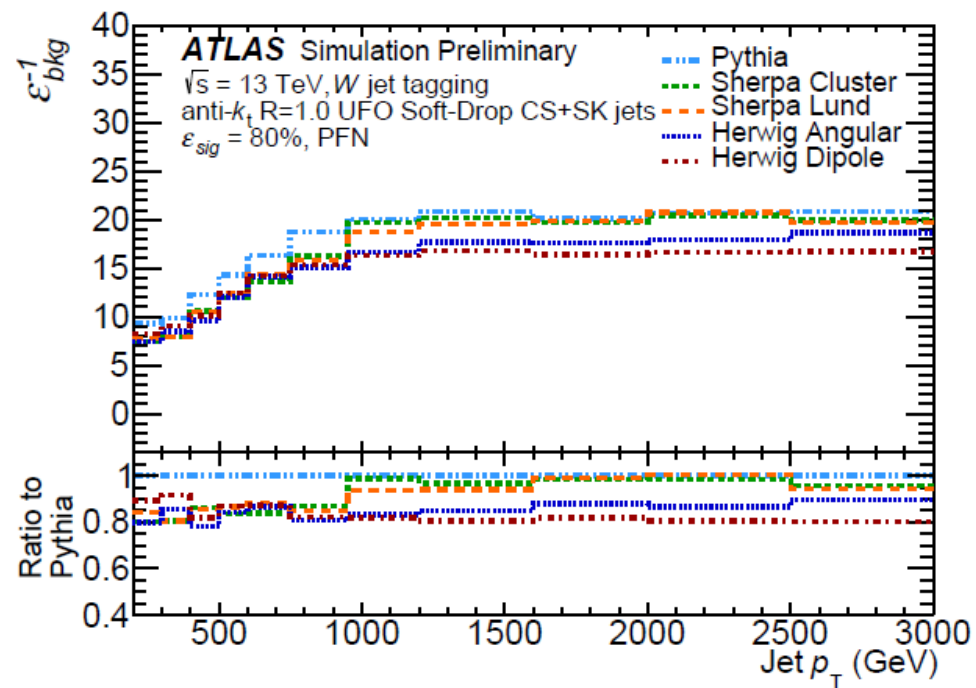
# Tagger Performance

## Model dependence

PFN:



(c)



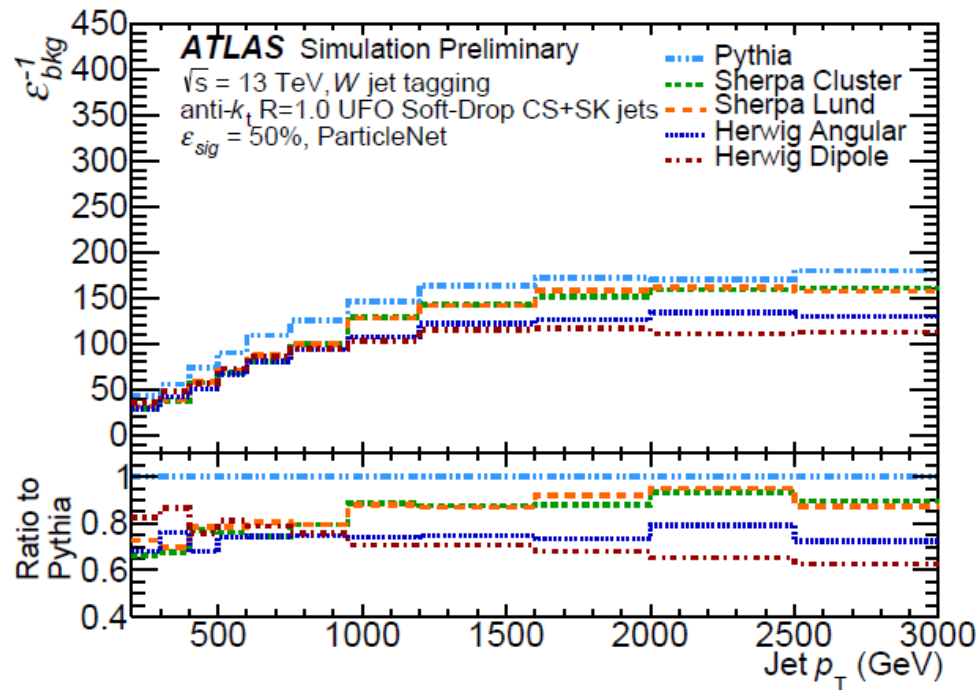
(d)

Figure 5: Comparison of the background rejection ( $\epsilon_{bkg}^{-1}$ ) of  $W$  taggers in different samples of simulated QCD jet, as a measure of model dependence. Shown is the background rejection using the threshold which results in an signal efficiency of 50% (a,c) or 80% (b,d) in each  $p_T$  bin for  $W' \rightarrow WZ$  testing sample. The top (bottom) row stands for the EFN (PFN) tagger.

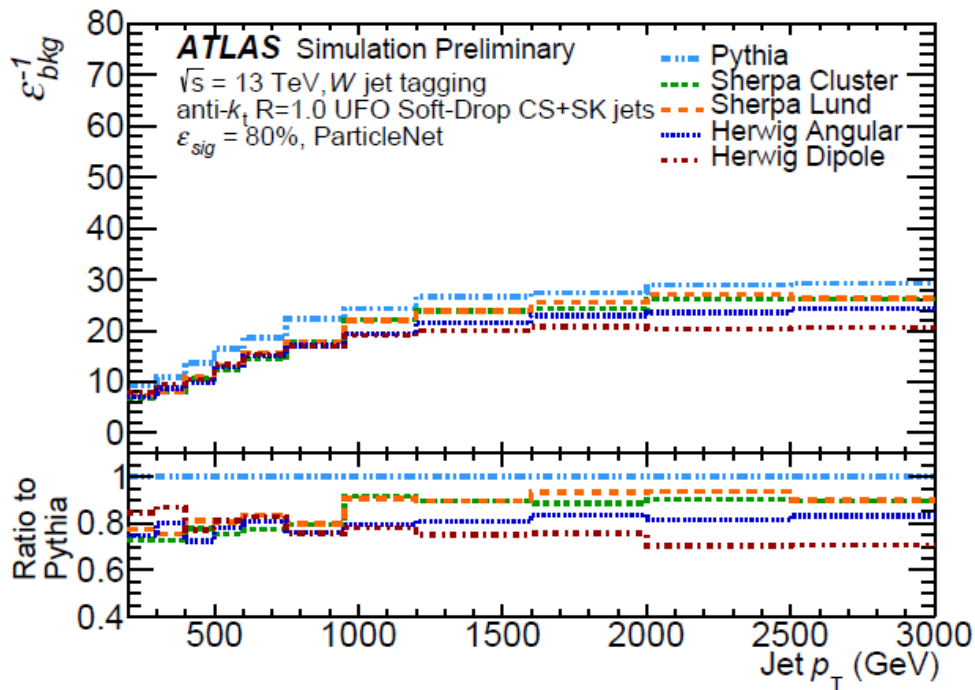
# Tagger Performance

## Model dependence

### ParticleNet:



(a)



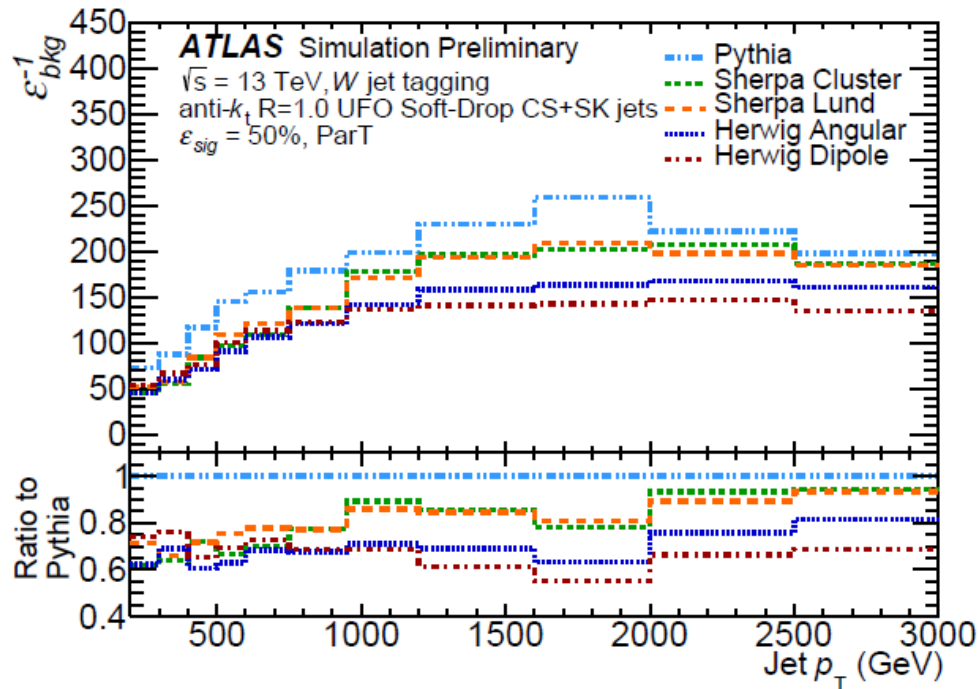
(b)

Figure 6: Comparison of the background rejection ( $\epsilon_{bkg}^{-1}$ ) of  $W$  taggers in different samples of simulated QCD jet, as a measure of model dependence. Shown is the background rejection using the threshold which results in an signal efficiency of 50% (a,c) or 80% (b,d) in each  $p_T$  bin for  $W' \rightarrow WZ$  testing sample. The top (bottom) row stands for the ParticleNet (ParticleTransformer) tagger.

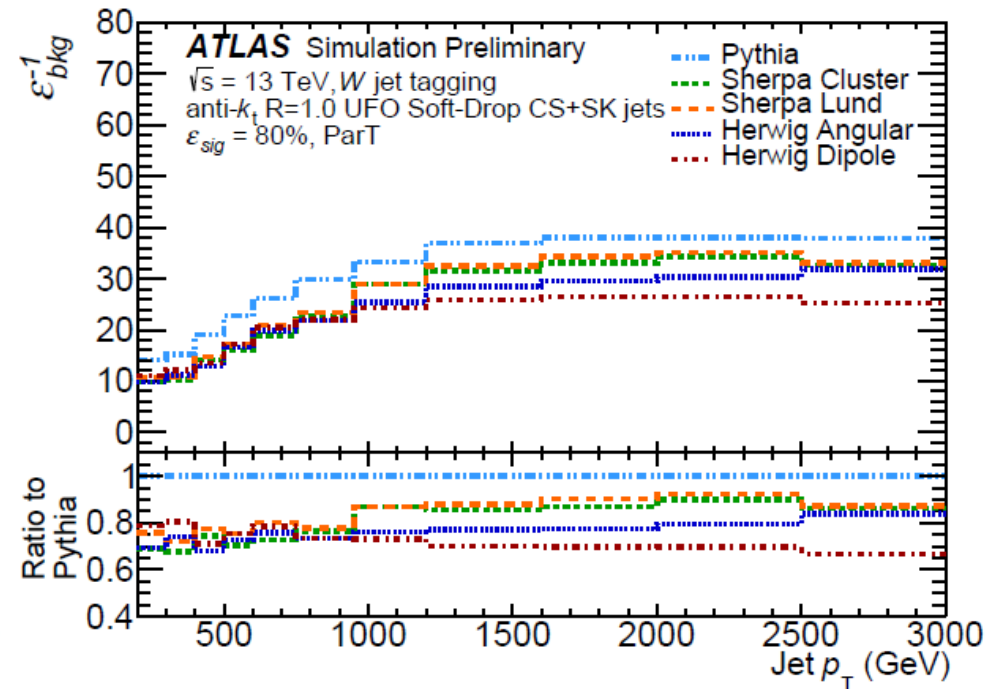
# Tagger Performance

## Model dependence

### ParticleTransformer:



(c)



(d)

- The background rejection of each tagger is different in each background sample.
- For 50% working point, the differences in background rejection are about 10%-40%. For 80% working point, the differences are smaller, around 10%-30%.

# Tagger Performance

## Model dependence

- The envelope constructed with the maximum ratio between the Pythia background rejection and the set of four alternative models is presented in bins of jet  $p_T$ , for each of the studied taggers.

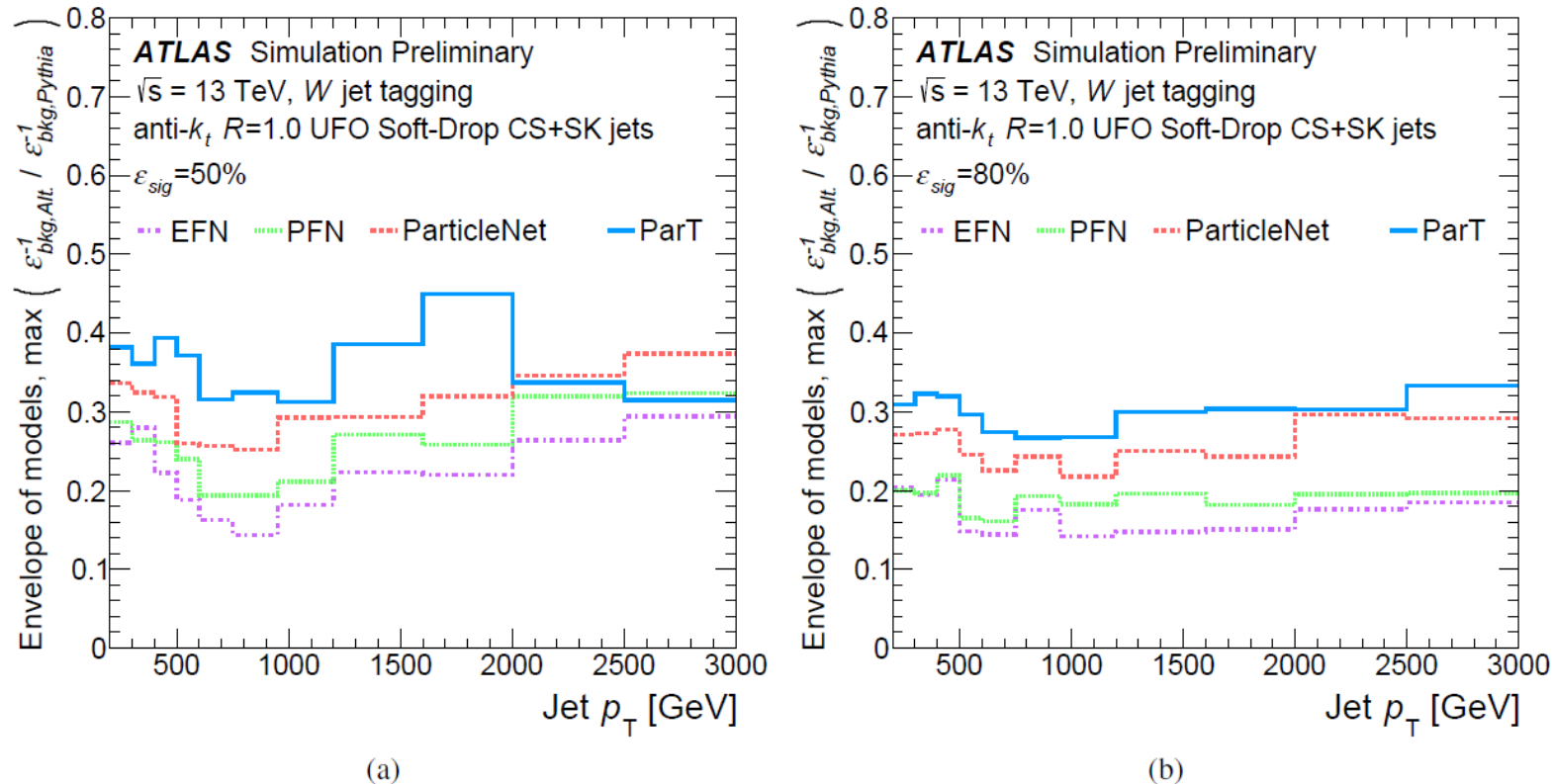


Figure 7: The envelope constructed with the maximum ratio between the PYTHIA background rejection and the set of four alternative models is presented for bins of jet  $p_T$ , for each of the studied taggers, for classifiers with a fixed 50% signal tagging efficiency (a), or 80% signal tagging efficiency (b) in the nominal sample.

# Tagger Performance

## Model dependence

- The sensitivity of tagger performance on the **hadronization** modeling.

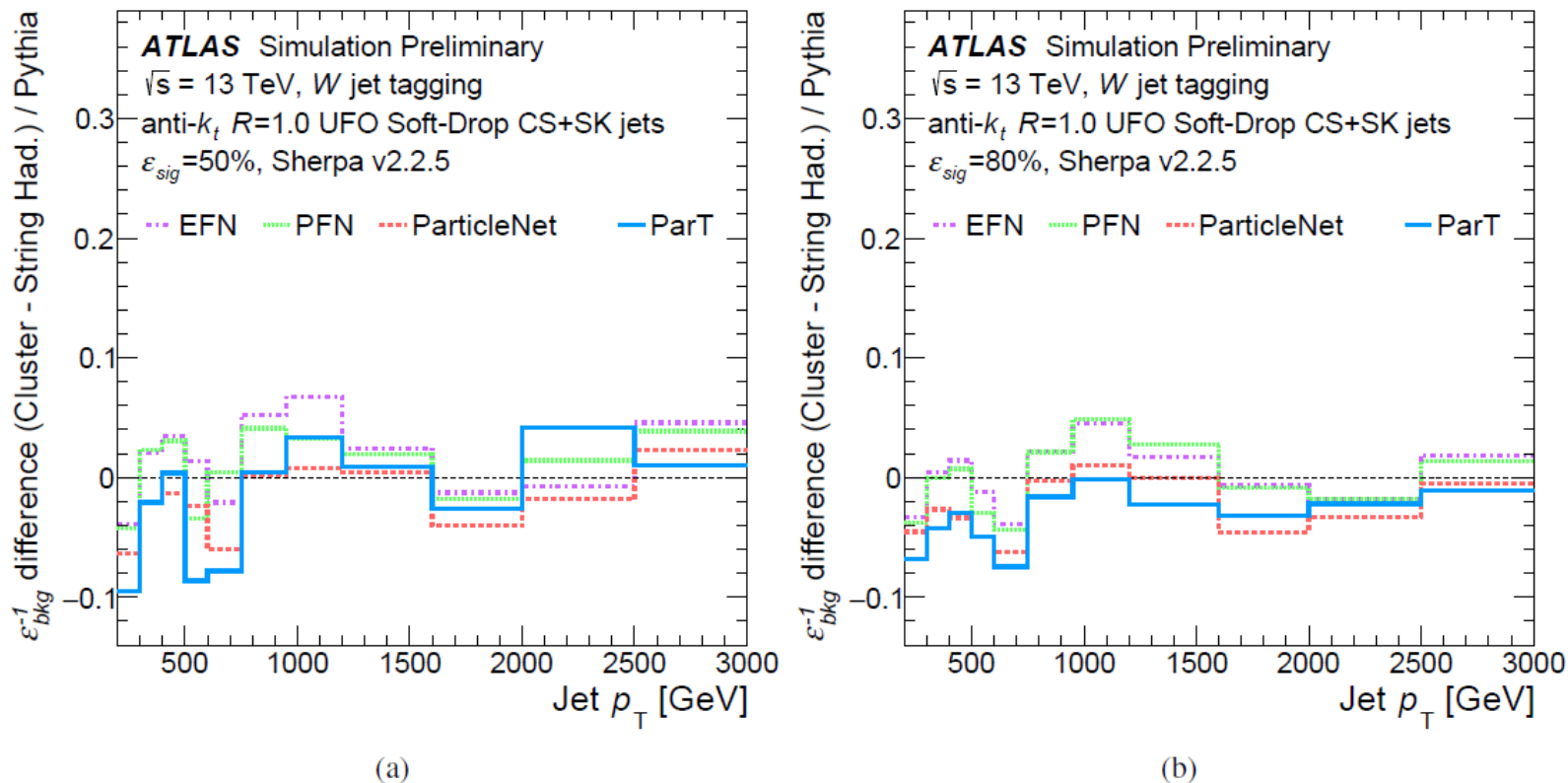


Figure 8: The sensitivity of tagger performance on the hadronisation and parton shower modeling, for each of the studied taggers, for classifiers with a fixed 50% signal tagging efficiency (a,c), or 80% signal tagging efficiency (b,d) in the nominal sample.

# Tagger Performance

## Model dependence

- The sensitivity of tagger performance on the **parton shower** modeling.

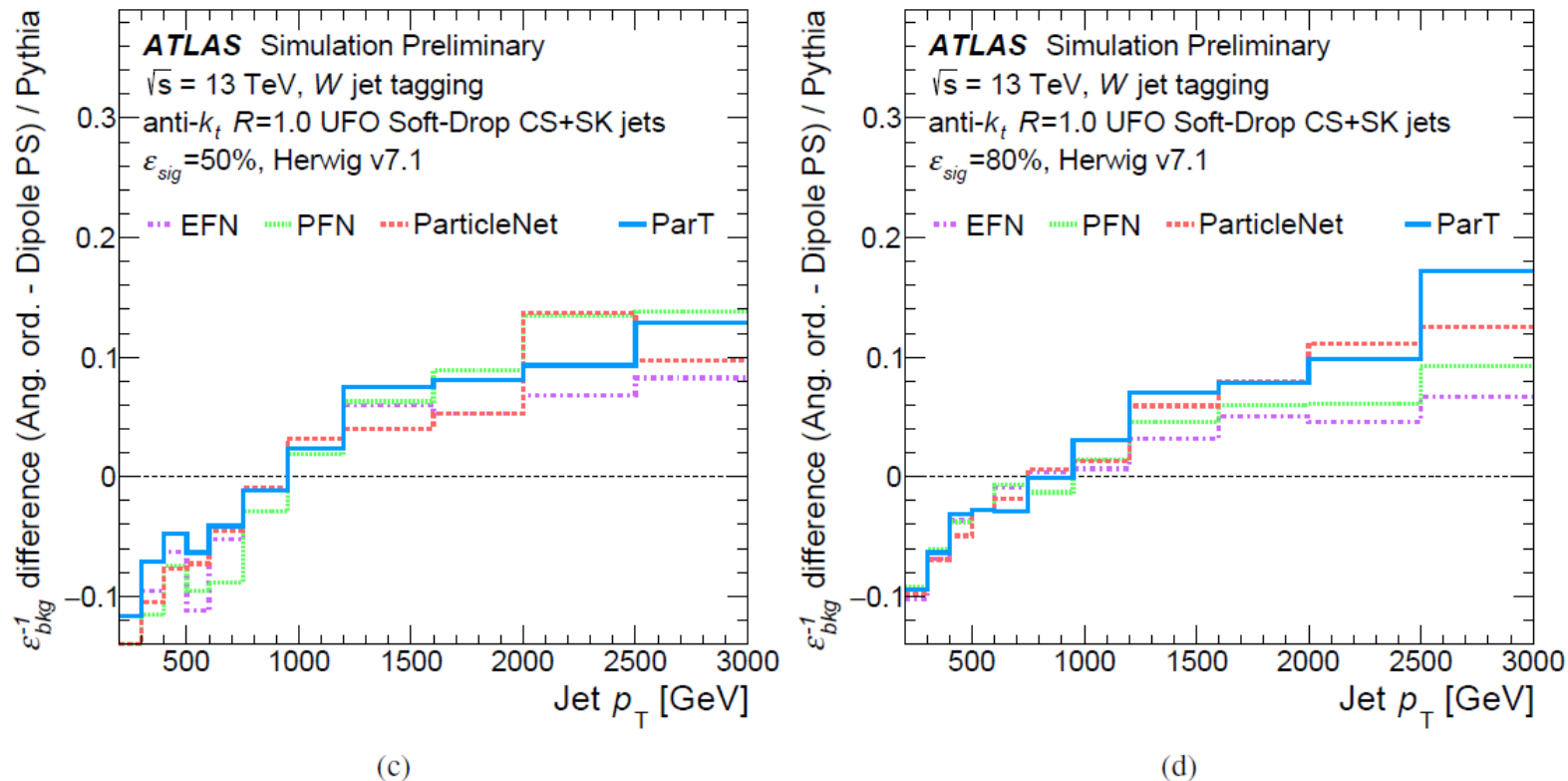


Figure 8: The sensitivity of tagger performance on the hadronisation and parton shower modeling, for each of the studied taggers, for classifiers with a fixed 50% signal tagging efficiency (a,c), or 80% signal tagging efficiency (b,d) in the nominal sample.

# Conclusion

- All of the constituent-based taggers trained in this study (EFN, PFN, ParticleNet, ParticleTransformer) show stronger performance than the tagger using high-level quantities presented in a previous study.
- In terms of performance, ParticleTransformer stands out as the top performer, with ParticleNet, PFN, and EFN following behind. Notably, ParticleTransformer achieves a improvement of about 1.8-2.8 (1.6-2.7) times in background rejection compared to the baseline tagger, for the 50% (80%) working point.
- The dependence of tagger performance on the choice of parton shower and hadronization models used in Monte Carlo simulations is also presented.
- Model dependence of tagger performance is found to increase with the complexity of the classifier.
- The performance of  $W$  tagging is found to be more susceptible to parton shower model variations than to models of non-perturbative hadronization effects.