Top LHC France

Large-R Jet Tagging in ATLAS

with the Particle Transformer for Run 3

Andrés Duque Bran | LPCA, Clermont-Ferrand | May 27, 2026

Focus on top quark tagging in the multiclass regime

Boosted Jet Tagging

Jet Tagging

Identify the originating particle of a jet from pp collisions:
q/g, top quark, W, Z or Higgs bosons

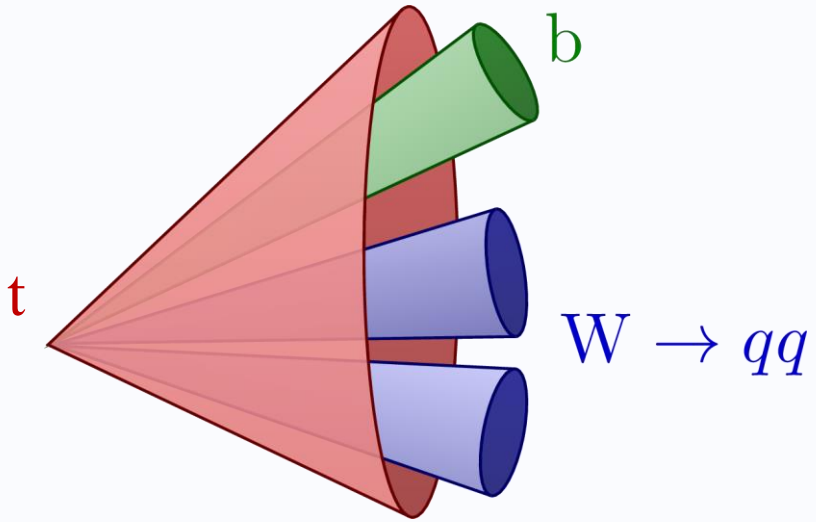
Boosted Jet Tagging

Jet Tagging

Identify the originating particle of a jet from pp collisions:
q/g, top quark, W, Z or Higgs bosons

Boosted Regime

At high p_T, massive particle decays are collimated into a single **large-R jet (R = 1.0)**



Boosted top decay

Boosted Jet Tagging

Jet Tagging

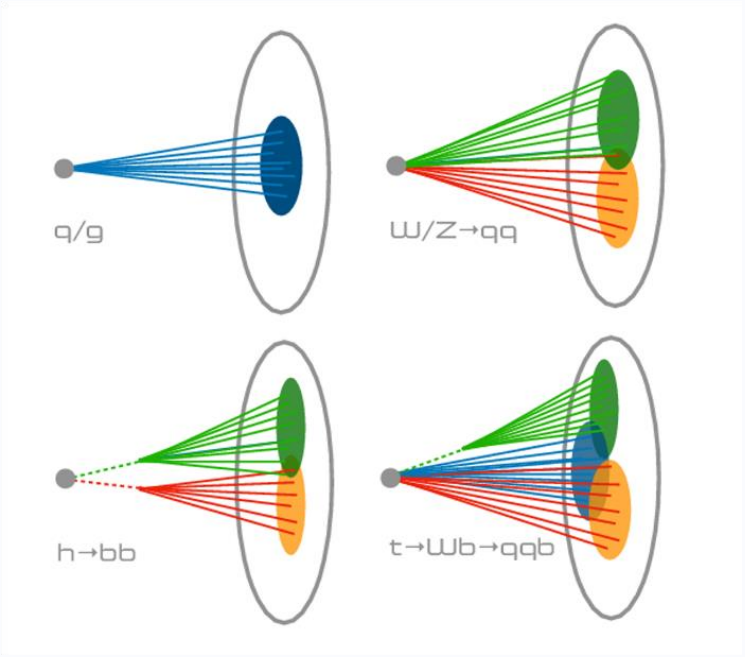
Identify the originating particle of a jet from pp collisions:
q/g, top quark, W, Z or Higgs bosons

Boosted Regime

At high p_T, massive particle decays are collimated into a single **large-R jet (R = 1.0)**

Complex Topology

Multiple hard sub-jets, overlapping energy deposits
→ **class separation non-trivial**



Boosted decay products collimated within R=1.0

Boosted Jet Tagging

Jet Tagging

Identify the originating particle of a jet from pp collisions:
q/g, top quark, W, Z or Higgs bosons

Boosted Regime

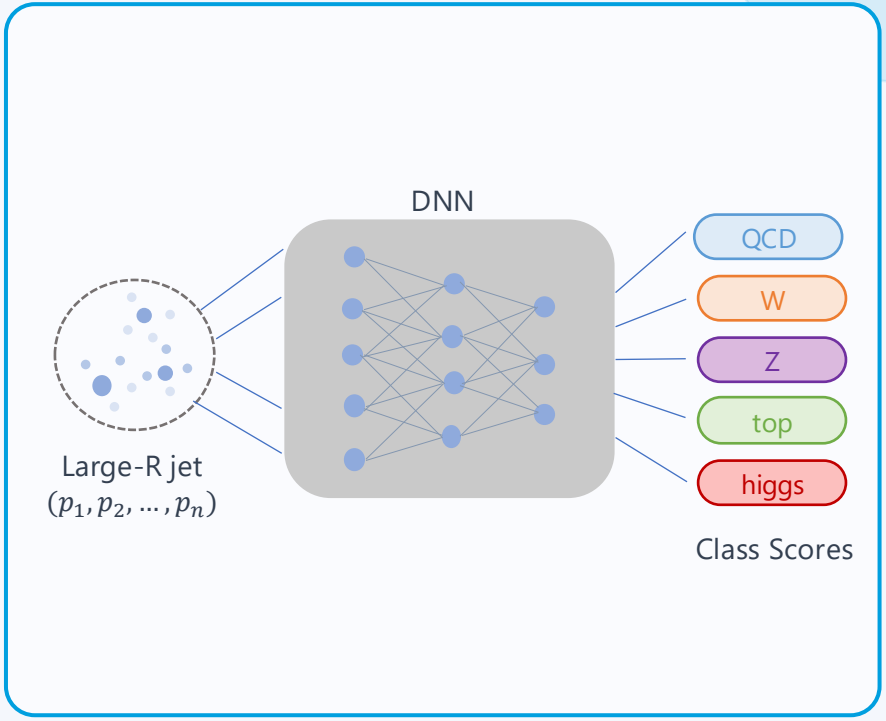
At high p_T, massive particle decays are collimated into a single **large-R jet (R = 1.0)**

Complex Topology

Multiple hard sub-jets, overlapping energy deposits
→ **class separation non-trivial**

How to tag them?

Constituent-based **deep learning** taggers take **detector-level particles** as inputs, exploiting the full substructure directly

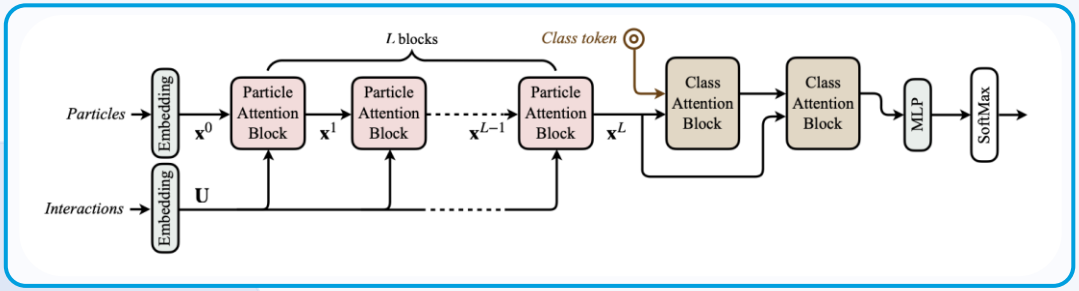


Constituent-based multiclass jet tagger

The Particle Transformer

Qu et al. 2022 • arXiv:2202.03772

- Built on **Transformer attention blocks**
- Takes **per-particle** and **pairwise interaction features**
- Uses **Particle Multi-Head Attention** — interaction features embedded directly into the attention mechanism
- Unlike NLP Transformers, jets are **unordered sets**

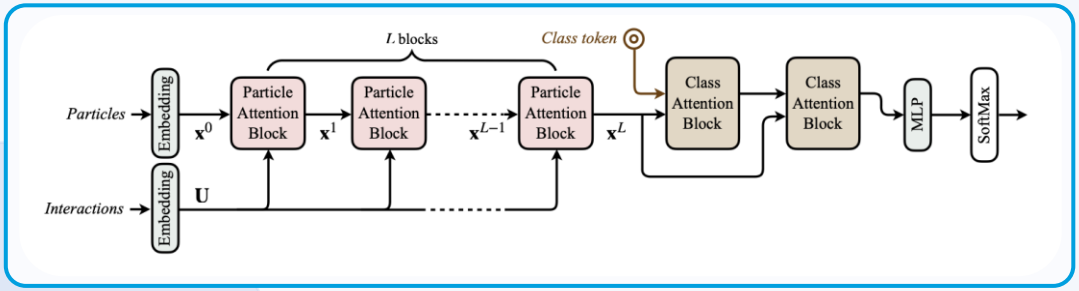


Particle Transformer architecture

The Particle Transformer

Qu et al. 2022 • arXiv:2202.03772

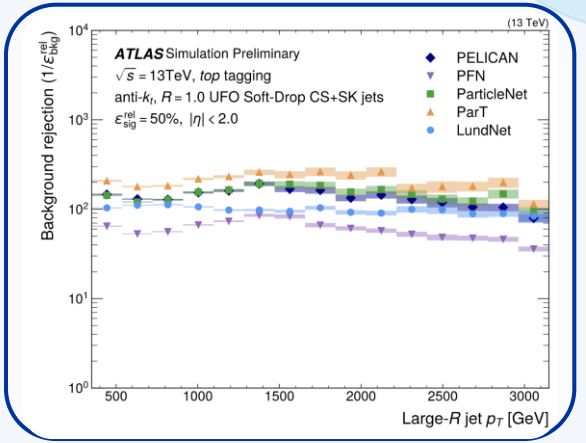
- Built on **Transformer attention blocks**
- Takes **per-particle** and **pairwise interaction features**
- Uses **Particle Multi-Head Attention** — interaction features embedded directly into the attention mechanism
- Unlike NLP Transformers, jets are **unordered sets**



Particle Transformer architecture

State of the Art — Binary Tagging

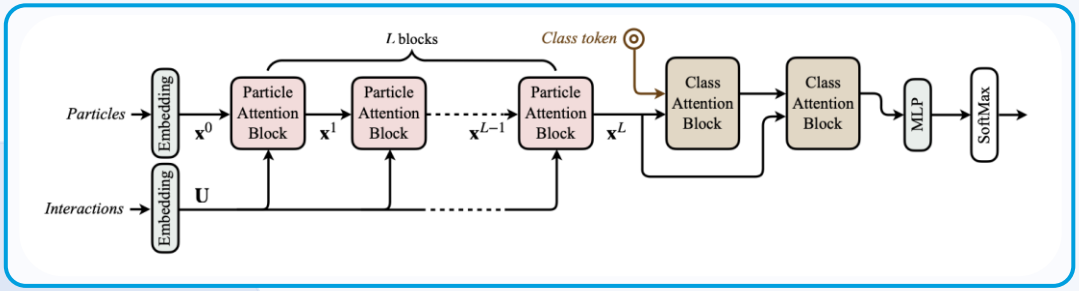
[ATLAS, JETM-2024-02]



The Particle Transformer

Qu et al. 2022 • arXiv:2202.03772

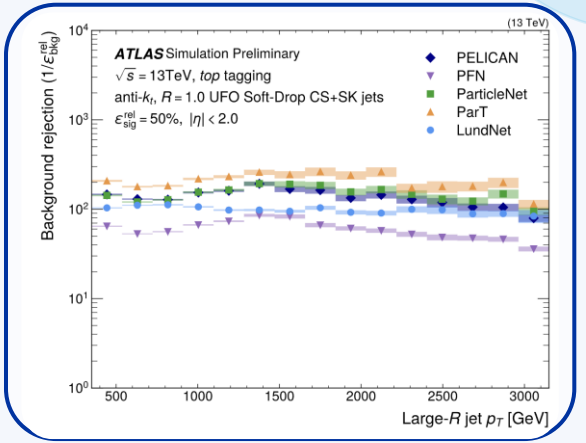
- Built on **Transformer attention blocks**
- Takes **per-particle** and **pairwise interaction features**
- Uses **Particle Multi-Head Attention** — interaction features embedded directly into the attention mechanism
- Unlike NLP Transformers, jets are **unordered sets**



Particle Transformer architecture

State of the Art — Binary Tagging

[ATLAS, JETM-2024-02]



This work: multiclass ParT — W, Z, top, H, QCD

Single model replaces multiple binary classifiers

Dataset & Model Setup

Features

Per-particle:

$log p_T$	$log E$
$log \frac{p_T}{p_T^{jet}}$	$log \frac{E}{E^{jet}}$
$\Delta\eta^*$	$\Delta\phi^*$
ΔR	
d_0	$z_0 sin\theta$

** Transformation in η-φ plane for three-prongs, following Kasieczka et al. (2019).*

Dataset & Model Setup

Features

Per-particle:

$log p_T$	$log E$
$log \frac{p_T}{p_T^{jet}}$	$log \frac{E}{E^{jet}}$
$\Delta\eta^*$	$\Delta\phi^*$
ΔR	
d_0	$z_0 sin\theta$

** Transformation in η-φ plane for three-prongs, following Kasieczka et al. (2019).*

Dataset & Model Setup

Features

Per-particle:

$log p_T$	$log E$
$log \frac{p_T}{p_T^{jet}}$	$log \frac{E}{E^{jet}}$
$\Delta\eta^*$	$\Delta\phi^*$
ΔR	
d_0	$z_0 sin\theta$

** Transformation in η-φ plane for three-prongs, following Kasieczka et al. (2019).*

Pairwise interaction:

$log\Delta^{ab} = log\sqrt{(\eta^a - \eta^b)^2 + (\phi^a - \phi^b)^2}$
$log k_T^{ab} = log(min(p_T^a, p_T^b) \cdot \Delta^{ab})$
$log z^{ab} = log\ min(p_T^a, p_T^b)/(p_T^a + p_T^b)$
$log(m^{ab})^2 = log(p^{\mu,a} + p^{\mu,b})^2$

Dataset & Model Setup

Features

Per-particle:

$log p_T$	$log E$
$log \frac{p_T}{p_T^{jet}}$	$log \frac{E}{E^{jet}}$
$\Delta\eta^*$	$\Delta\phi^*$
ΔR	
d_0	$z_0 sin\theta$

** Transformation in η-φ plane for three-prongs, following Kasieczka et al. (2019).*

Pairwise interaction:

$log\Delta^{ab} = log\sqrt{(\eta^a - \eta^b)^2 + (\phi^a - \phi^b)^2}$
$log k_T^{ab} = log(min(p_T^a, p_T^b) \cdot \Delta^{ab})$
$log z^{ab} = log\ min(p_T^a, p_T^b)/(p_T^a + p_T^b)$
$log(m^{ab})^2 = log(p^{\mu,a} + p^{\mu,b})^2$

Dataset

- **FullSim jets, Pythia 8**
  - 145M jets (90%/10% train/validation)
  - 16M jets for prediction
- Events reweighted to flatten p_T per class

Dataset & Model Setup

Features

Per-particle:

$log p_T$	$log E$
$log \frac{p_T}{p_T^{jet}}$	$log \frac{E}{E^{jet}}$
$\Delta\eta^*$	$\Delta\phi^*$
ΔR	
d_0	$z_0 sin\theta$

** Transformation in η-φ plane for three-prongs, following Kasieczka et al. (2019).*

Pairwise interaction:

$log\Delta^{ab} = log\sqrt{(\eta^a - \eta^b)^2 + (\phi^a - \phi^b)^2}$
$log k_T^{ab} = log(min(p_T^a, p_T^b) \cdot \Delta^{ab})$
$log z^{ab} = log\ min(p_T^a, p_T^b)/(p_T^a + p_T^b)$
$log(m^{ab})^2 = log(p^{\mu,a} + p^{\mu,b})^2$

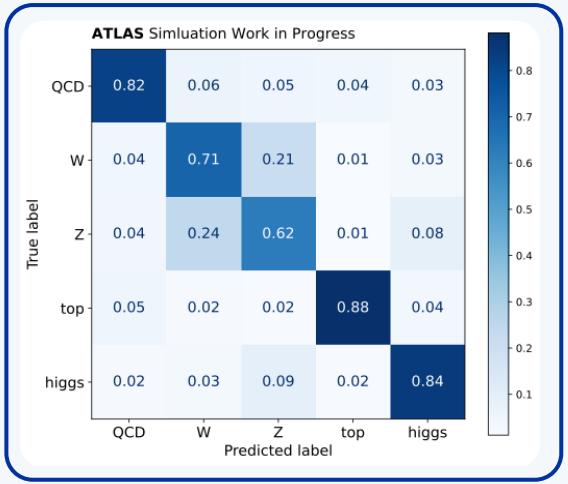
Dataset

- **FullSim jets, Pythia 8**
  - 145M jets (90%/10% train/validation)
  - 16M jets for prediction
- Events reweighted to flatten p_T per class

Class	Train+Val	Prediction
q/g	$44.9M$	$5.1M$
$W \rightarrow qq$	$25.1M$	$2.8M$
$Z \rightarrow qq, bb, cc$	$25.1M$	$2.8M$
$\boldsymbol{t \rightarrow bqq}$	$\boldsymbol{44.9M}$	$\boldsymbol{5.1M}$
$H \rightarrow bb$	$5.1M$	$0.6M$

Results: ParT Baseline Performance

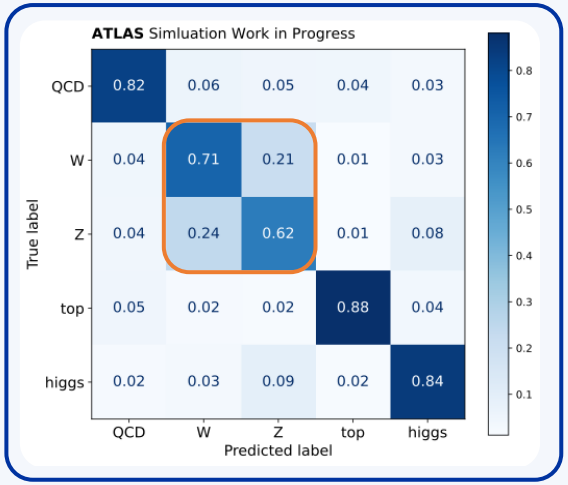
- **Good separation across all classes**
- W and Z mutually confused due to similar two-prong topology
- Some H/Z confusion: Z has mixed decay (qq, bb, cc)
- W and Z show lower QCD rejection than top and Higgs



Confusion matrix

Results: ParT Baseline Performance

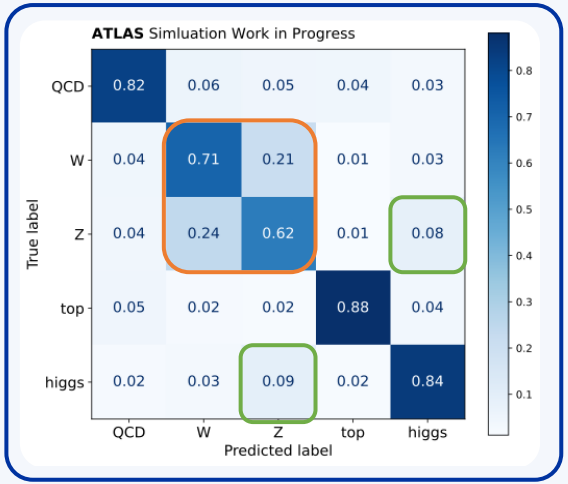
- Good separation across all classes
- **W and Z mutually confused due to similar two-prong topology**
- Some H/Z confusion: Z has mixed decay (qq, bb, cc)
- W and Z show lower QCD rejection than top and Higgs



Confusion matrix

Results: ParT Baseline Performance

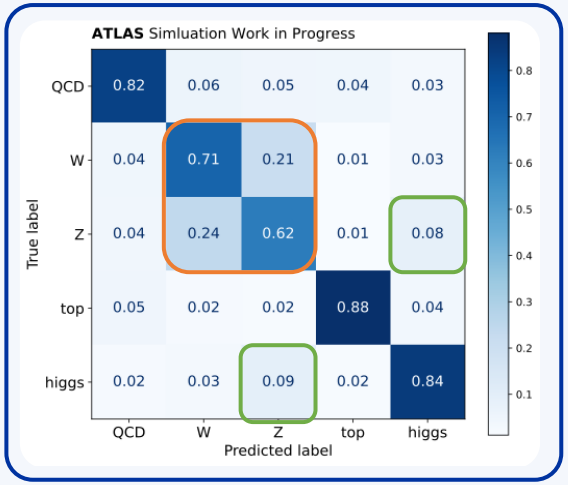
- Good separation across all classes
- W and Z mutually confused due to similar two-prong topology
- **Some H/Z confusion: Z has mixed decay (qq, bb, cc)**
- W and Z show lower QCD rejection than top and Higgs

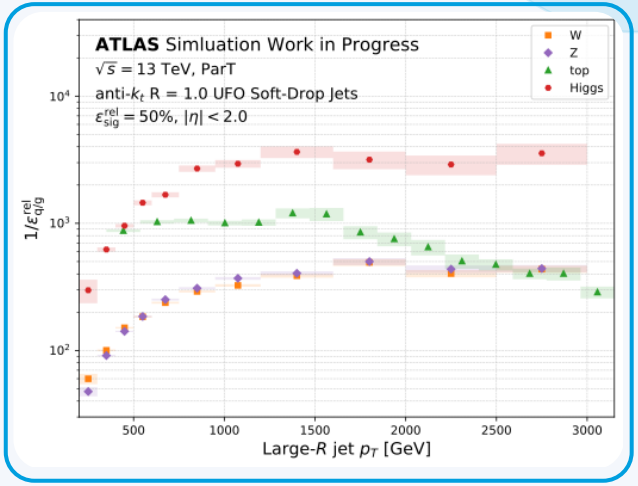


Confusion matrix

Results: ParT Baseline Performance

- Good separation across all classes
- W and Z mutually confused due to similar two-prong topology
- Some H/Z confusion: Z has mixed decay (qq, bb, cc)
- W and Z show lower QCD rejection than top and Higgs



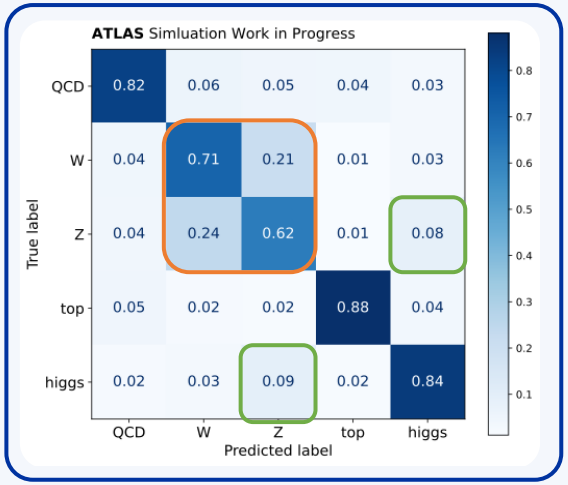


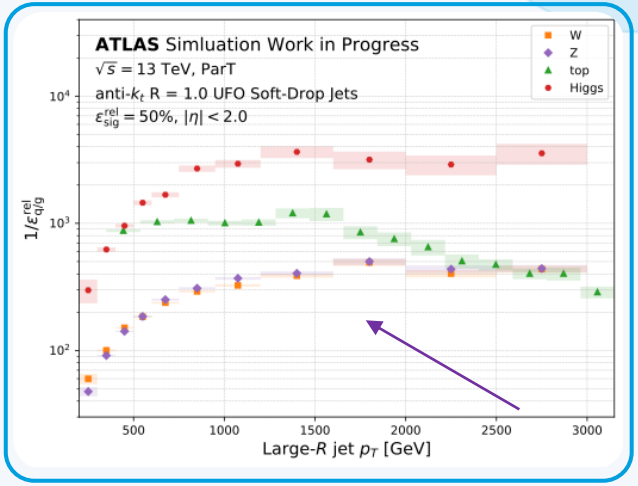
Confusion matrix and QCD rejection against p_T

Discriminant: log-likelihood ratio $\log(p_{sig}/p_{QCD})$ used as binary discriminant per class.

Results: ParT Baseline Performance

- Good separation across all classes
- W and Z mutually confused due to similar two-prong topology
- Some H/Z confusion: Z has mixed decay (qq, bb, cc)
- **W and Z show lower QCD rejection than top and Higgs**



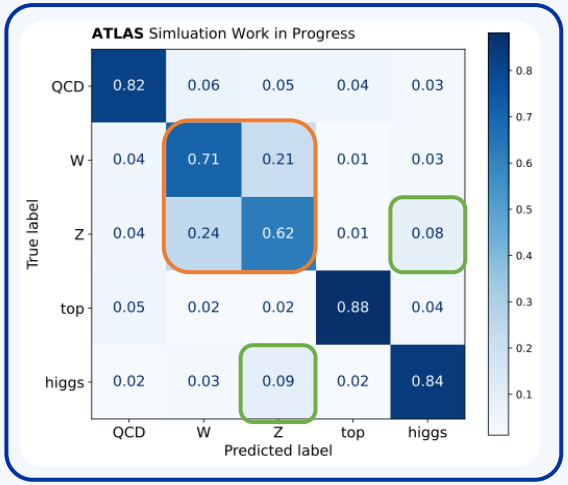


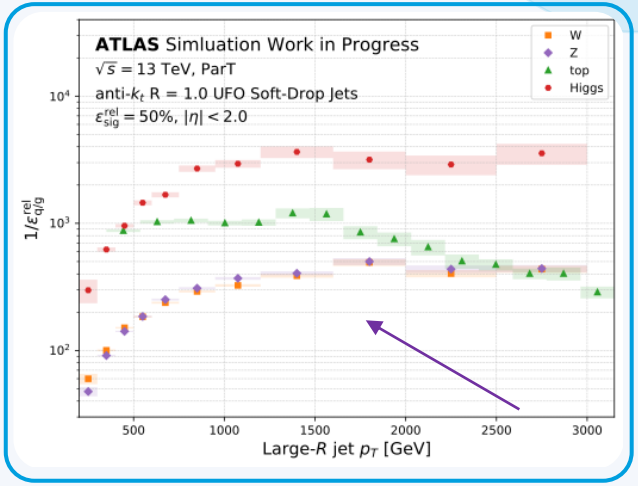
Confusion matrix and QCD rejection against p_T

Discriminant: log-likelihood ratio $\log(p_{sig}/p_{QCD})$ used as binary discriminant per class.

Results: ParT Baseline Performance

- Good separation across all classes
- W and Z mutually confused due to similar two-prong topology
- Some H/Z confusion: Z has mixed decay (qq, bb, cc)
- W and Z show lower QCD rejection than top and Higgs





Confusion matrix and QCD rejection against p_T

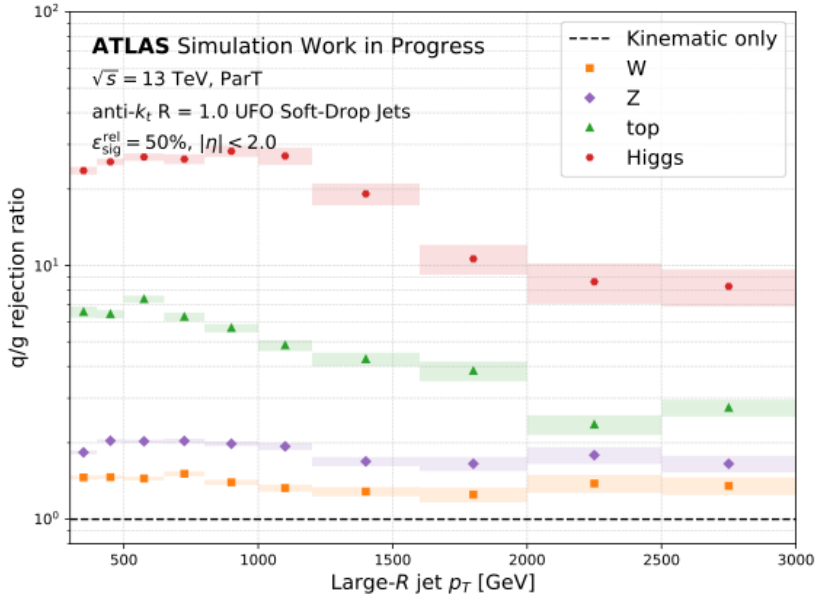
Discriminant: log-likelihood ratio $\log(p_{sig}/p_{QCD})$ used as binary discriminant per class.

Top well separated: distinctive three-prong structure.

Effect of Impact Parameters

$log p_T$	$log E$
$log \frac{p_T}{p_T^{jet}}$	$log \frac{E}{E^{jet}}$
$\Delta\eta^*$	$\Delta\phi^*$
ΔR	
d_0	$z_0 sin\theta$

Kinematic only



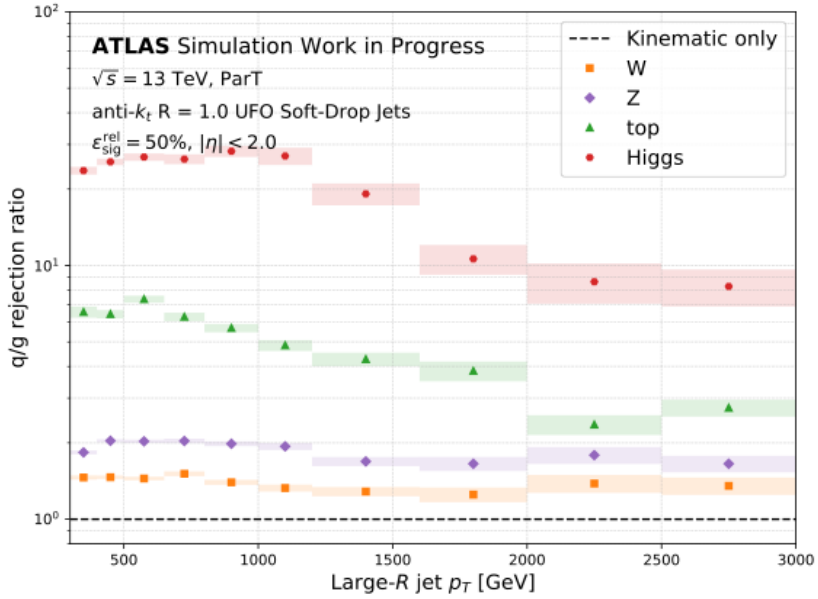
QCD rejection ratio between baseline and only kinematic features

Effect of Impact Parameters

$log p_T$	$log E$
$log \frac{p_T}{p_T^{jet}}$	$log \frac{E}{E^{jet}}$
$\Delta\eta^*$	$\Delta\phi^*$
ΔR	
d_0	$z_0 sin\theta$

Kinematic only

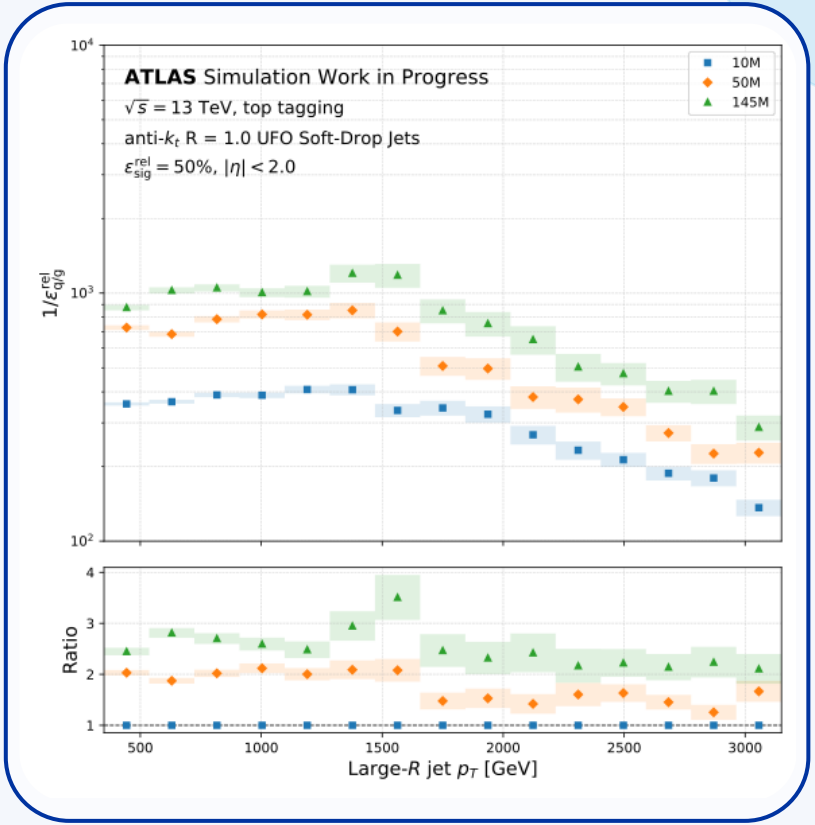
- Adding IP features consistently improves QCD rejection across all classes and pT bins
- **Largest gains for top and Higgs:** b-quark decay products carry large impact parameters
- W and Z benefit less, as b-quarks are not in their dominant decay modes



QCD rejection ratio between baseline and only kinematic features

Statistics Impact

- Performance scales consistently with training statistics across all p_T bins.
- Most improvement at low p_T.
- Performance gain stabilizes from 50M to 145M



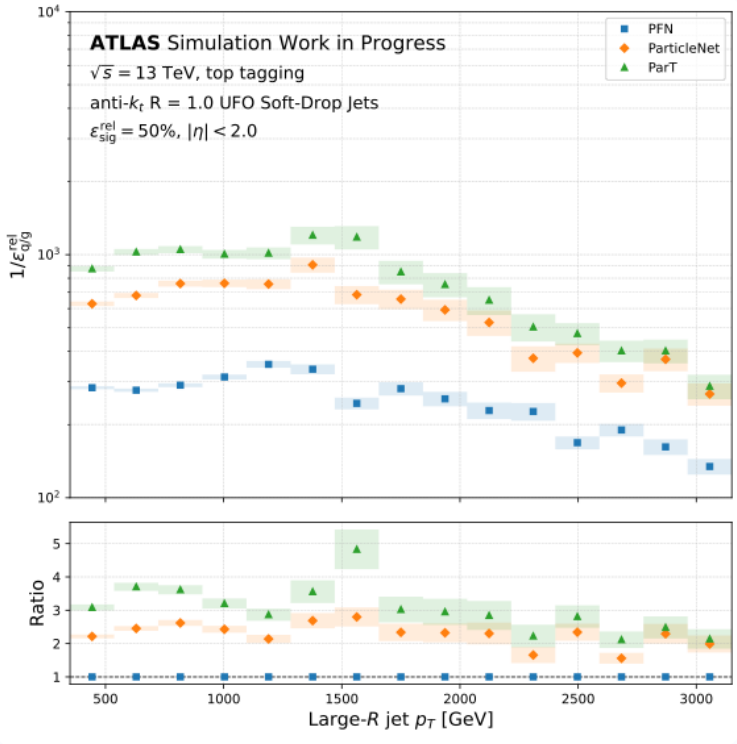
Comparison to other Architectures

- ParT compared against two baselines at identical training statistics.
- **PFN:** per-particle MLP + aggregation. Per-particle features only.
- **ParticleNet:** GNN with dynamic k-NN, angular pairwise features.

Comparison to other Architectures

- ParT compared against two baselines at identical training statistics.
- **PFN:** per-particle MLP + aggregation. Per-particle features only.
- **ParticleNet:** GNN with dynamic k-NN, angular pairwise features.

- ParT outperforms both across all p_T bins.
- Gains over PFN are substantially larger.
- Gains over ParticleNet moderate but consistent.
- **Pairwise interaction features provide clear advantage.**



MC Generator Dependence

- The QCD samples were replaced by different MC generators.
- Testing statistics were kept identical across all generators.

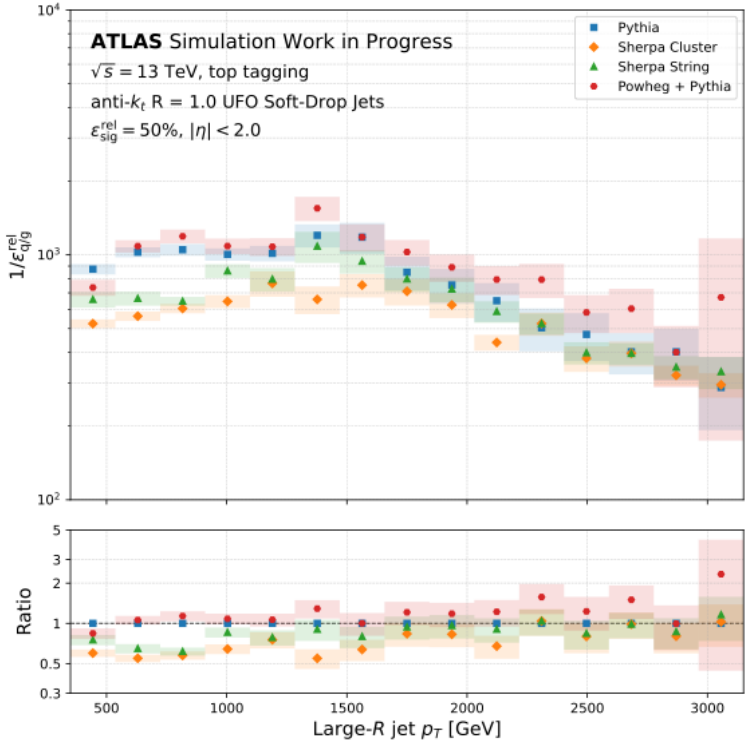
Generator	Hadronisation	Role
Pythia8	String	**Training**
Sherpa 2.2.5	Cluster	Testing
Sherpa 2.2.5	String	Testing
Powheg + Pythia8	String	Testing

MC Generator Dependence

- The QCD samples were replaced by different MC generators.
- Testing statistics were kept identical across all generators.

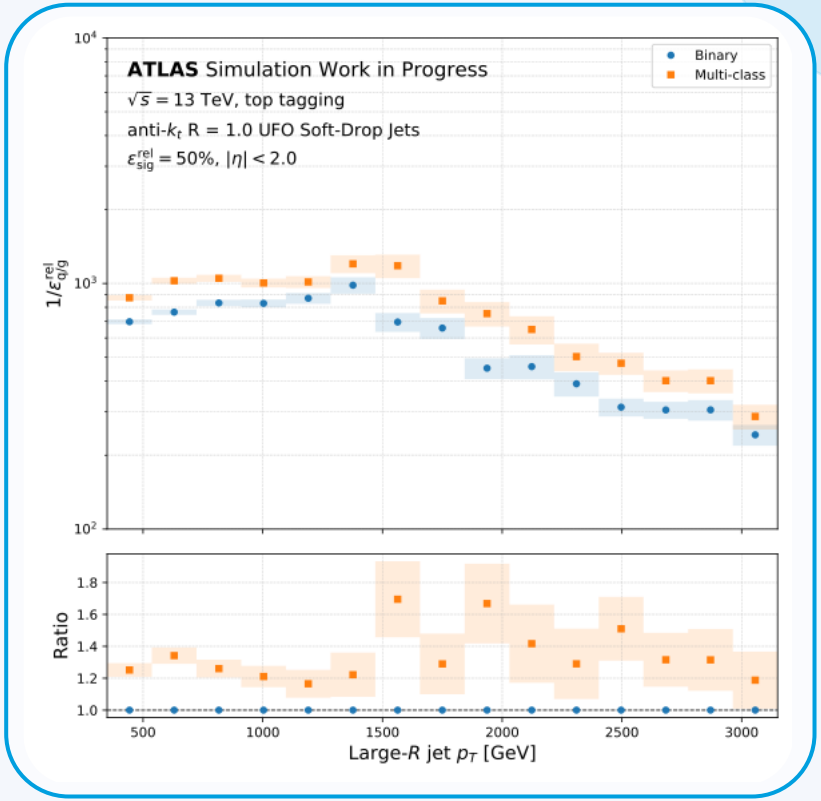
Generator	Hadronisation	Role
Pythia8	String	**Training**
Sherpa 2.2.5	Cluster	Testing
Sherpa 2.2.5	String	Testing
Powheg + Pythia8	String	Testing

- Pythia-based samples show comparable performance across full p_T.
- Sherpa comparable at high p_T where hard substructure dominates.



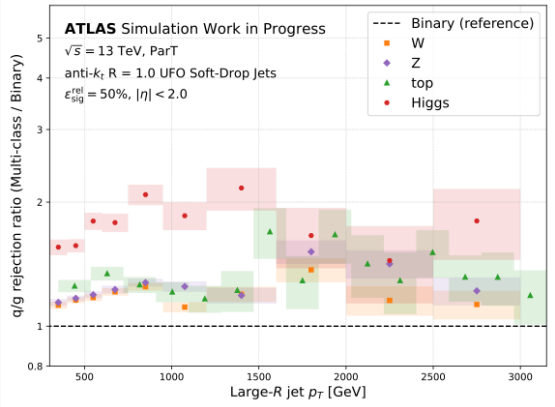
Comparison to binary ParT

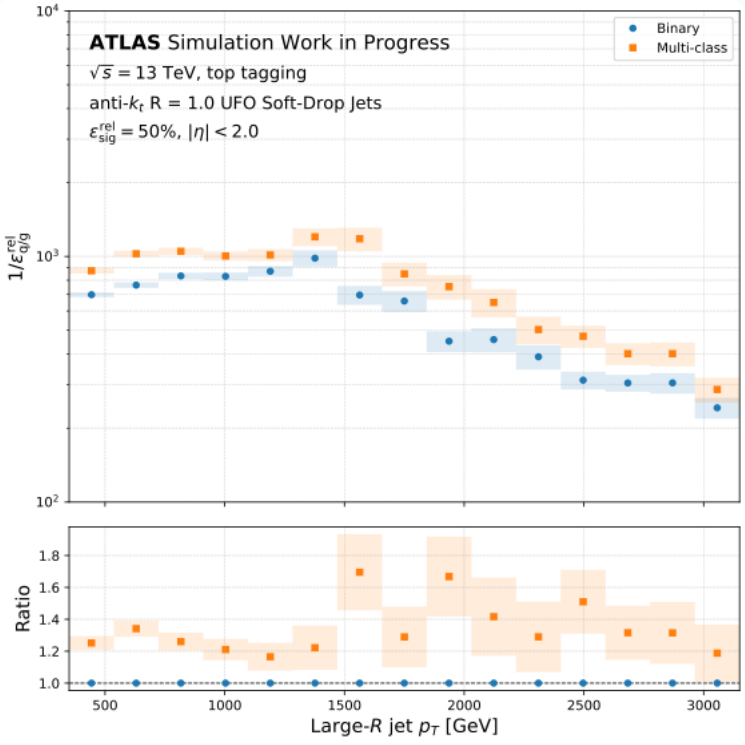
- A separate binary ParT model was trained for each signal:
  QCD vs W, QCD vs Z, QCD vs top, QCD vs Higgs
- Training and test statistics were kept identical per class.
- **Multiclass outperforms the binary model.**



Comparison to binary ParT

- A separate binary ParT model was trained for each signal: QCD vs W, QCD vs Z, QCD vs top, QCD vs Higgs
- Training and test statistics were kept identical per class.
- **Multiclass outperforms the binary model.**





Summary & Outlook

Summary

- Multiclass constituent-based jet tagging with ParT evaluated under ATLAS Run 3 conditions
- Impact parameters crucial → largest gains for top and Higgs
- Performance improves with statistics → mostly at low p_T
- ParT outperforms PFN, competitive with ParticleNet
- **Multiclass outperforms binary — one model, richer representations**

Outlook

Mass decorrelation for W tagging: ensure the tagger does not sculpt the jet mass distribution

Thank you!

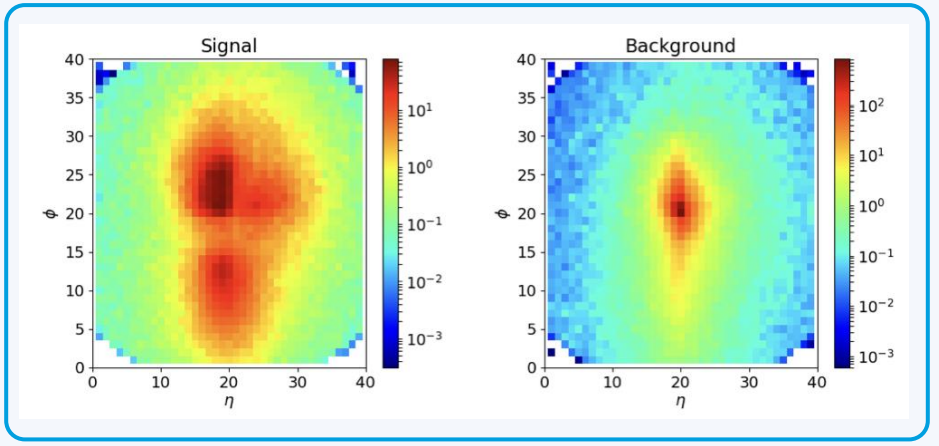
Questions?

Backup

$\Delta\eta, \Delta\phi$ Transformation

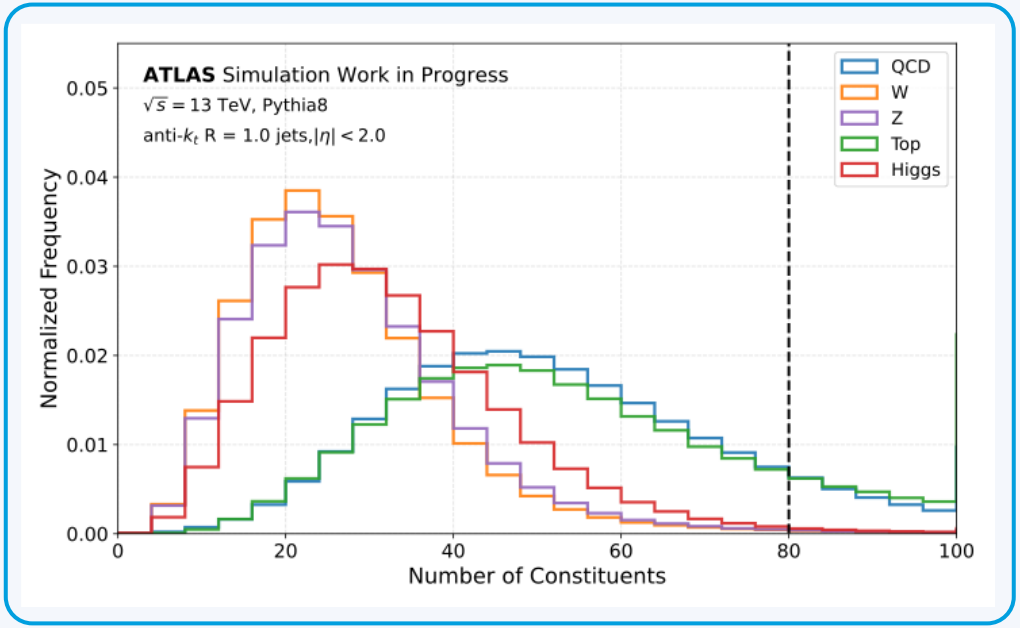
Preprocessing for booster top taggers:

- **Center –** translate η and ϕ so the hardest constituent sits at the origin.
- **Rotate –** rotate so the second hardest constituent lies on the negative ϕ axis.
- **Reflect –** flip η if needed so the third hardest constituent is in the positive η half-plane.



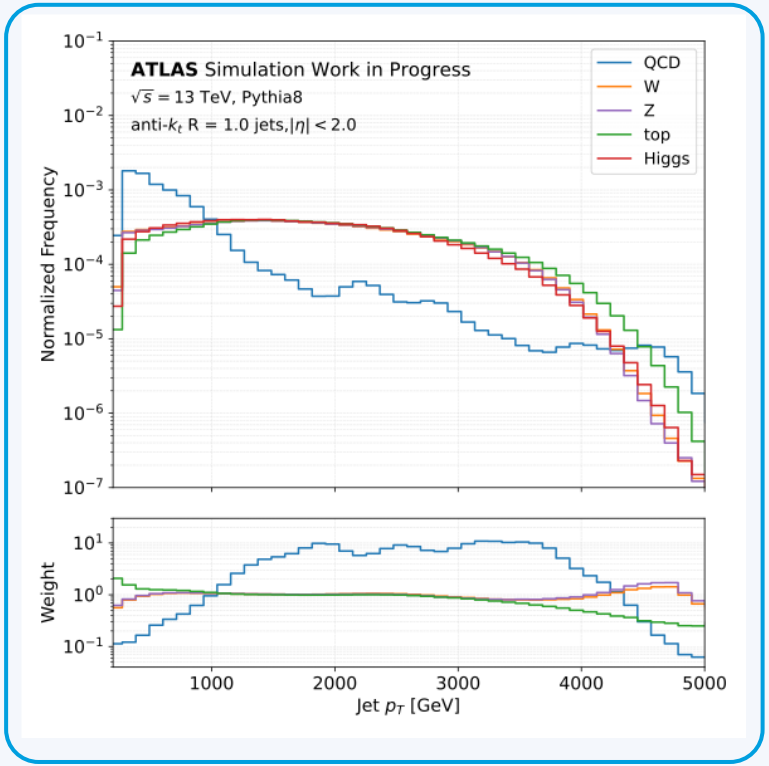
Top and q/g distributions after $\eta - \phi$ transformation

Dataset & Model Setup



Distribution of number of constituents per class

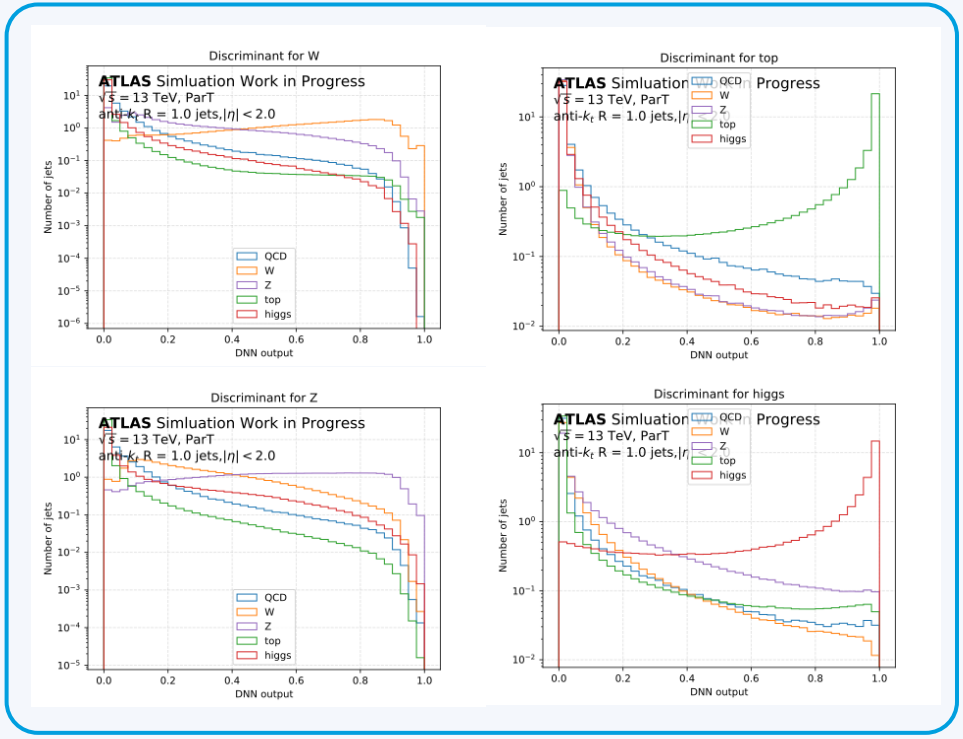
$\Delta\eta, \Delta\phi$ Transformation



pT distribution per jet and weights applied during training

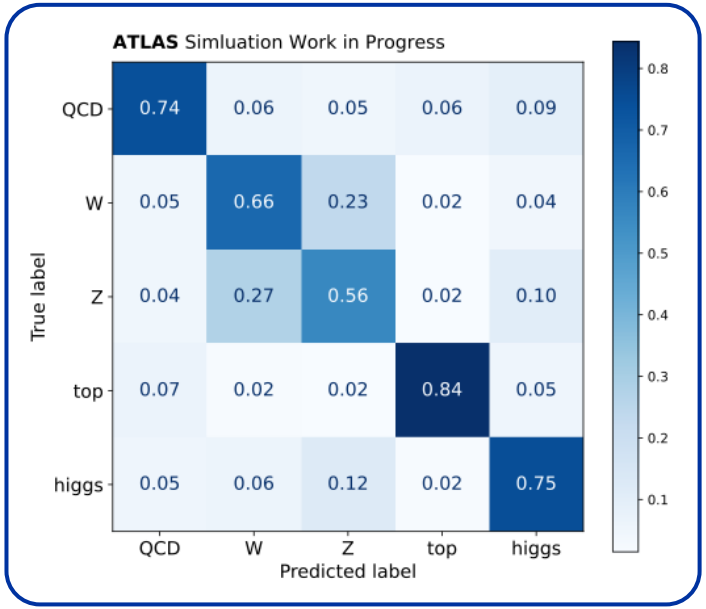
- Jet p_T is reweighted before training to ensure that all classes share a similar p_T distribution, preventing the model from learning kinematic biases instead of physics features.
- This reweighting stabilizes training and improves generalization, ensuring that performance differences across classes are not driven by a mismatched p_T spectra.

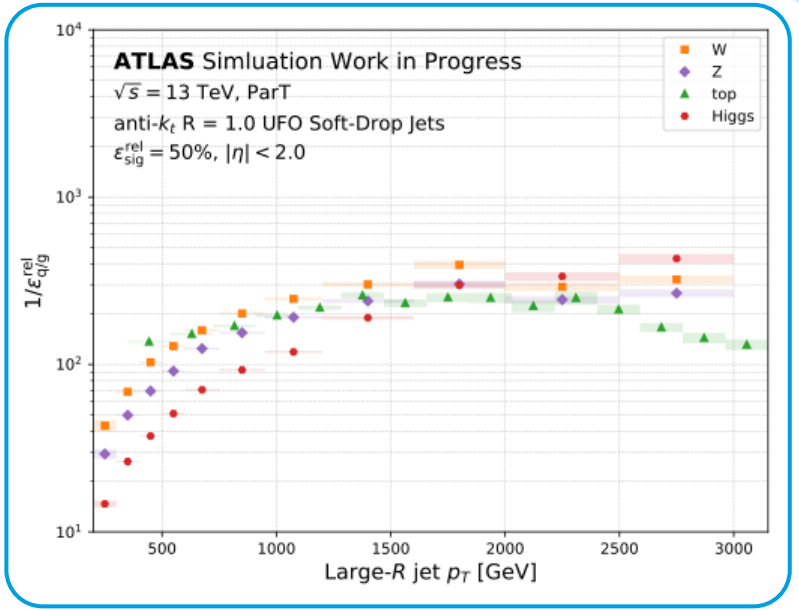
ParT: Baseline Performance



Discriminant per signal class

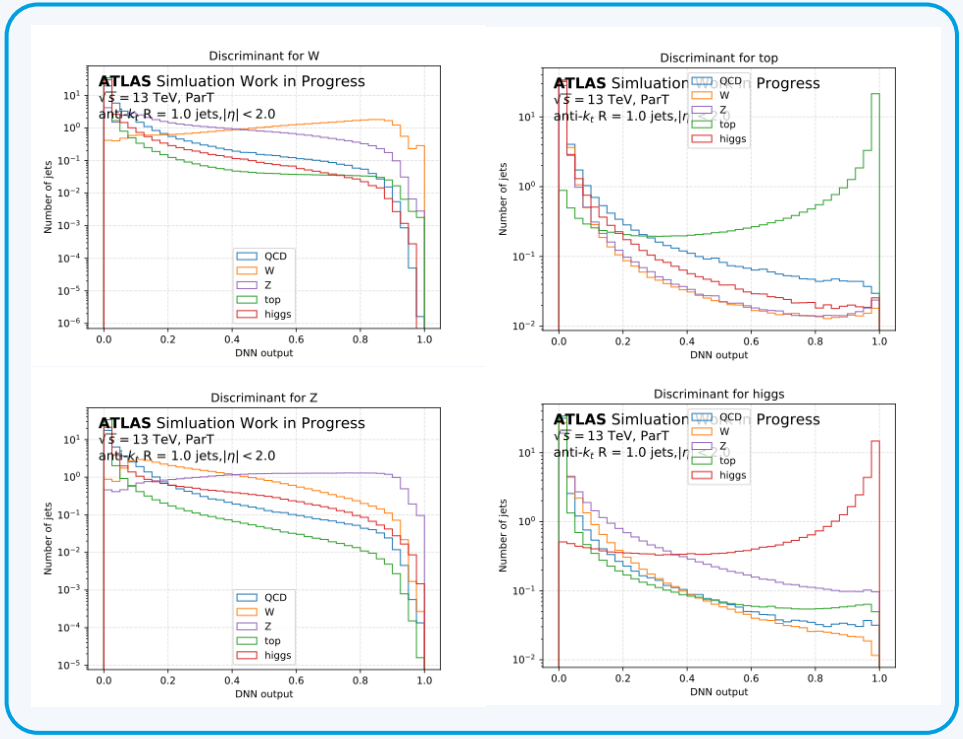
Effect of Impact Parameters: Only kinematics performance





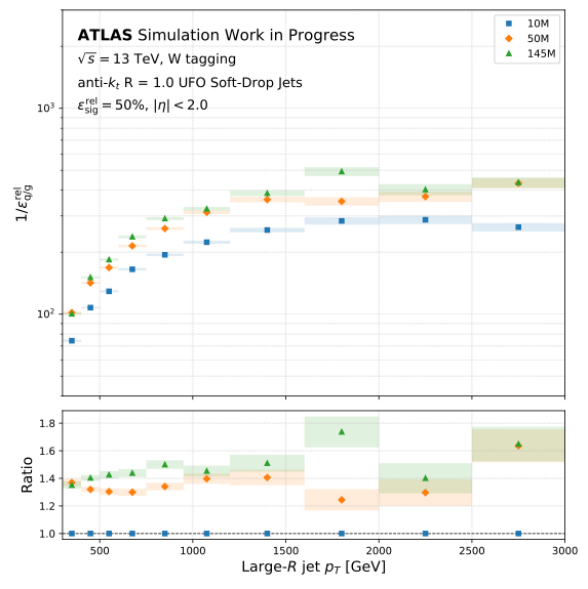
Confusion matrix and QCD rejection against p_T

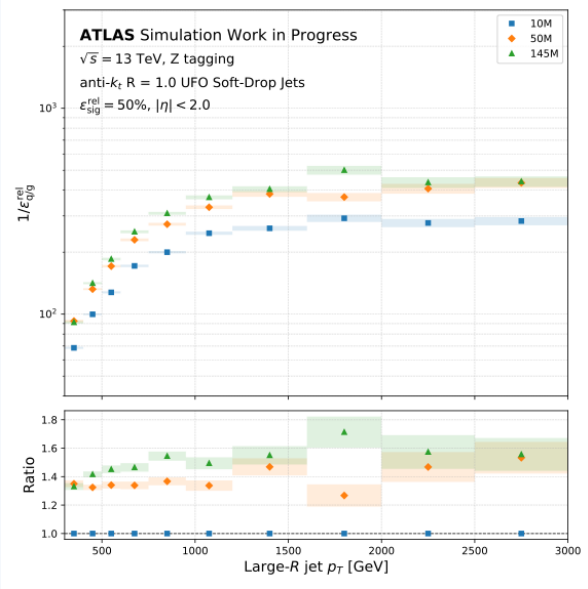
Effect of Impact Parameters

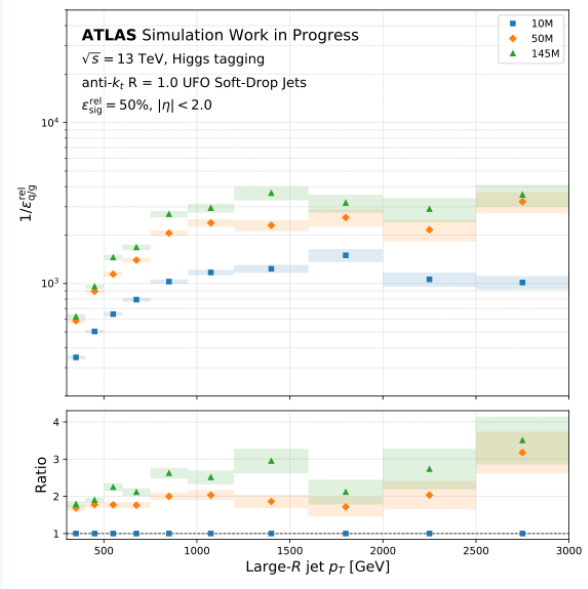


Discriminant per signal class

Statistics Impact



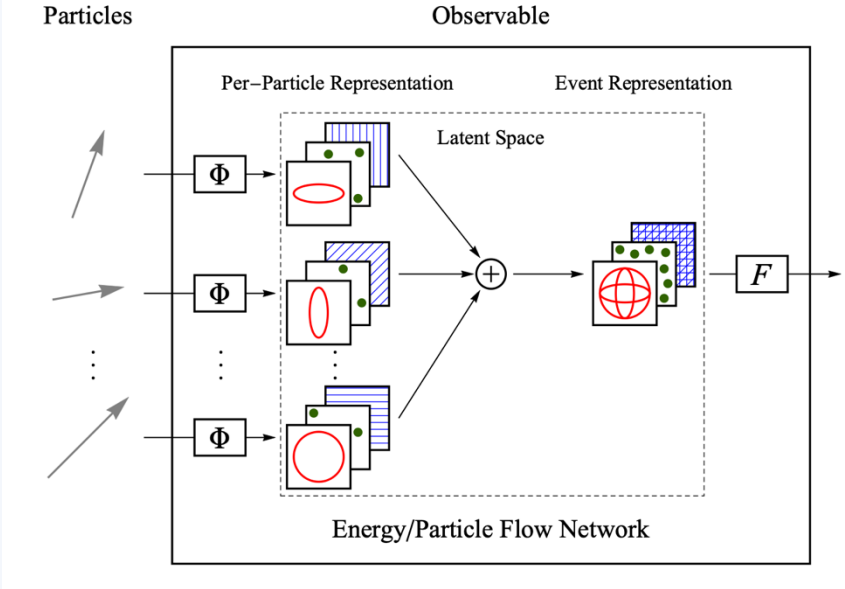




QCD rejection against p_T for other classes

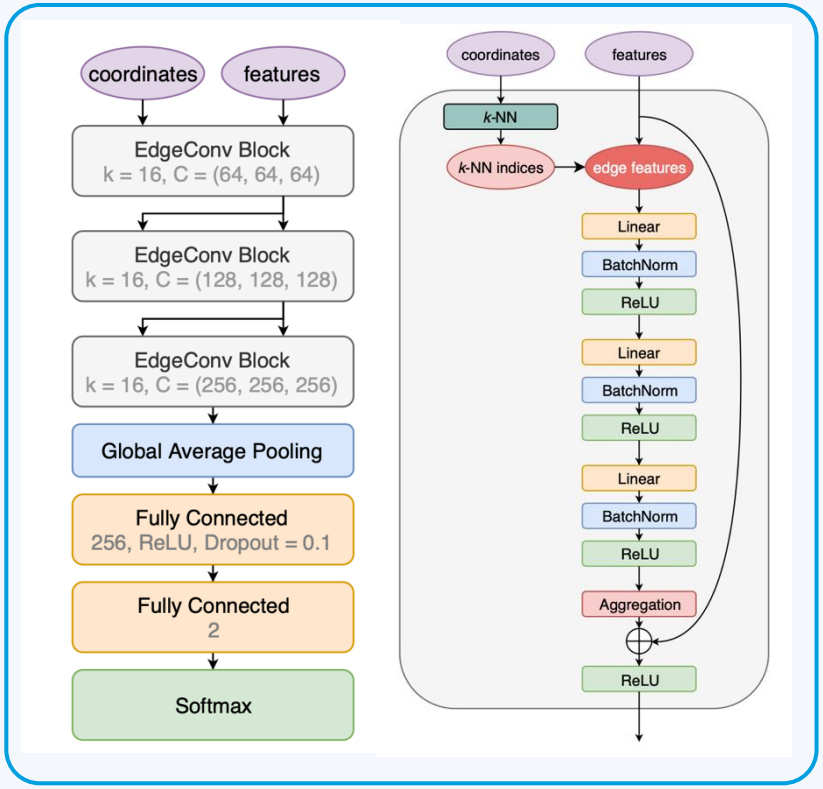
The Particle Flow Network Architecture

- Permutation-invariant architecture that processes jets as unordered sets of particles.
- Uses a per-particle embedding network followed by a sum to aggregate information.
- A final classifier network that learns global jet features from the aggregated representation for tagging tasks.



The architecture of Particle Flow Network (PFN) [arXiv:1810.05165]

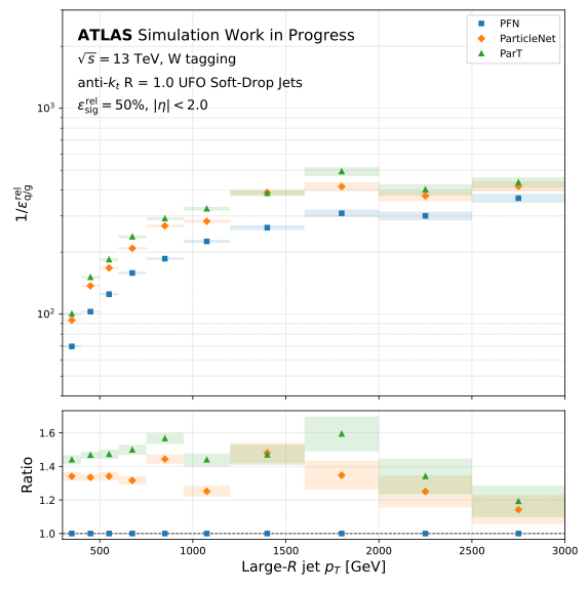
The ParticleNet Architecture

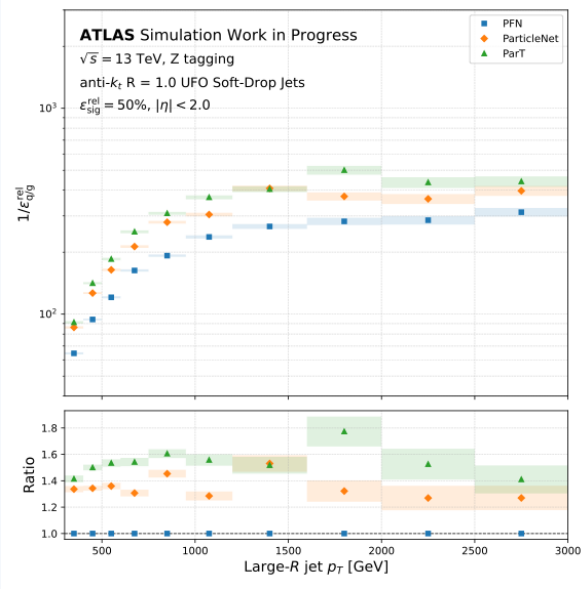


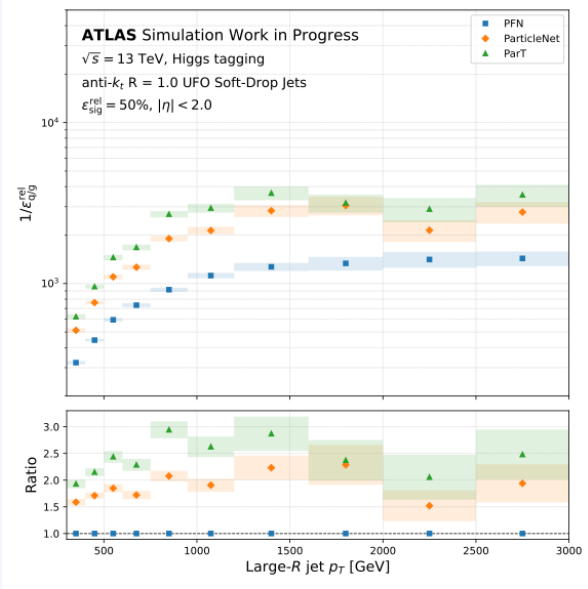
The architecture of ParticleNet [arXiv:1902.08570]

- Dynamic Graph CNN: builds and updates a particle graph using nearest-neighbor relations.
- EdgeConv blocks: learn geometric and relational features between constituents.
- Hierarchical aggregation: combines local and global information into a jet-level representation.

Model Comparison

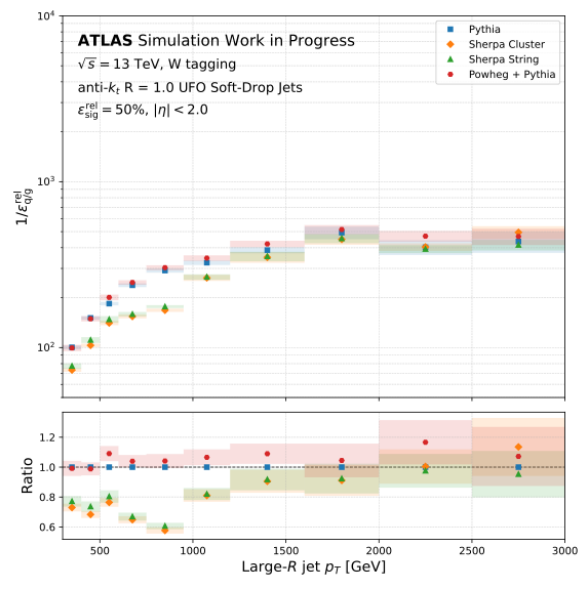


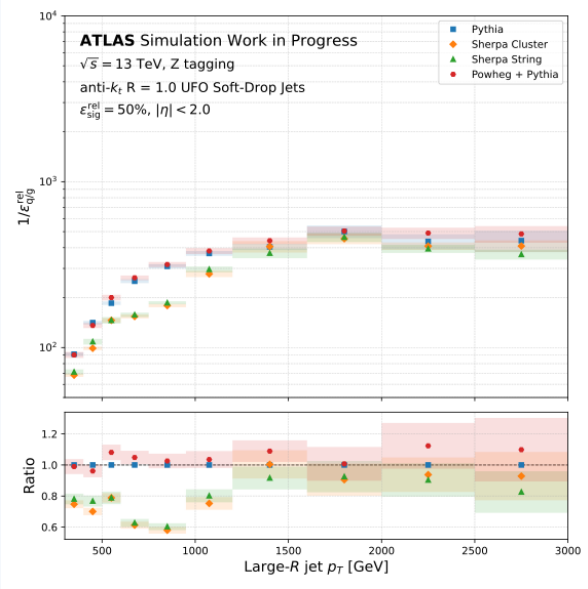


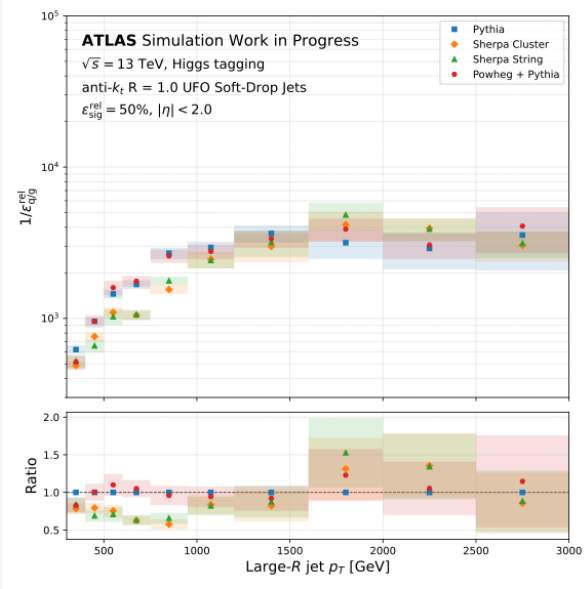


QCD rejection against p_T for other classes

MC Generator Comparison

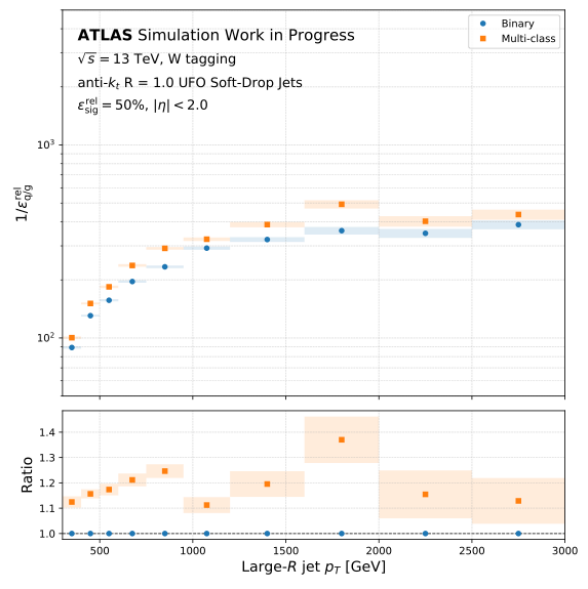


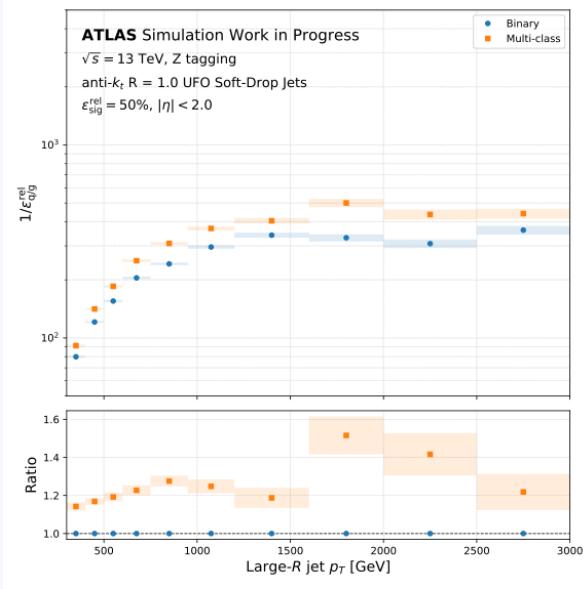


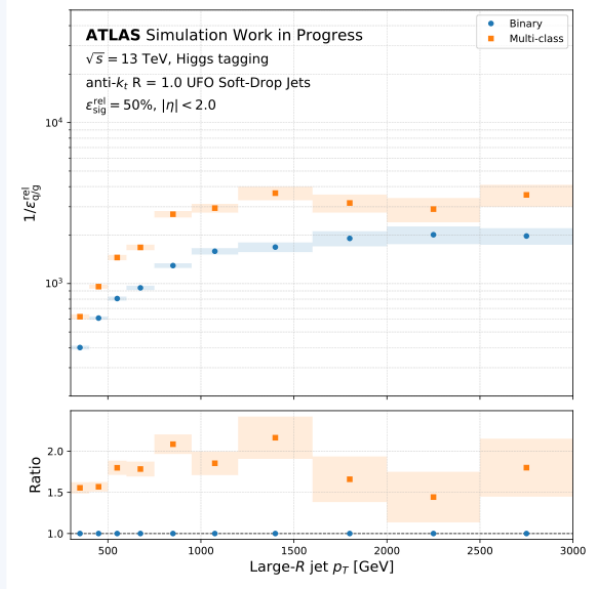


QCD rejection against p_T for other classes

Binary Comparison







QCD rejection against p_T for other classes