Classifying hadronic objects in ATLAS with ML/AI algorithms

Robert Les, IJCLab
on behalf of the ATLAS Collaboration

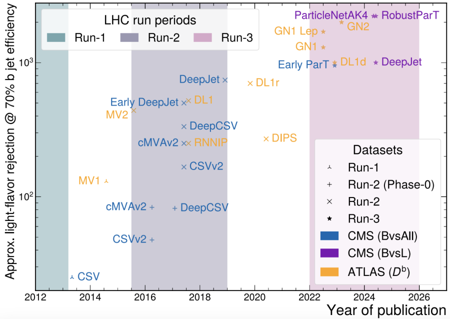
ML4Jets, Vienna
September 14, 2026

Funded by the European Union

Introduction

The historical trend for hadronic identification (jet-tagging):

1) High-level taggers on jet substructure observables
   - Robust and easy to interpret at analysis/theory level
2) Machine learning taggers on jet-substructure
   - Non-trivially combine several observables for better discrimination
3) Machine learning taggers on low-level inputs
   - Lose interpretativeness for performance

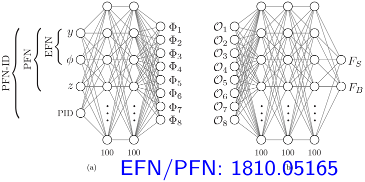


Will cover some recent results in ATLAS using ML for

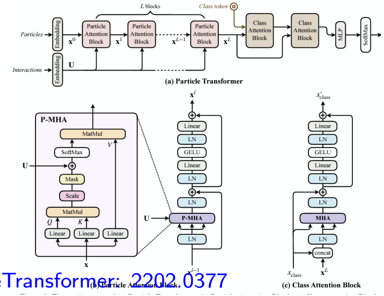
- Hadronic object classification
- Calibration of these results

ML/pheno community is developing faster then we can test on data!

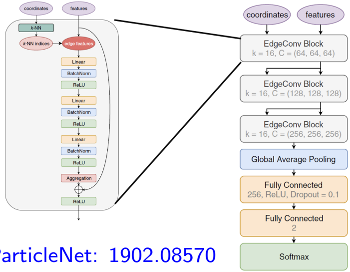
- Normal dense neural networks
- ResNet: CNN Architecture representing jet as image
- Energy/Particle Flow networks (EFN/PFN): General decomposition of IRC-safe observables
- ParticleNet: Graph network on point cloud
- ParticleTransformer: Transformer
- GN2: Transformer with auxiliary tasks
- LundNet: Graph on declustering history
- PELICAN: Lorentz invariant network



EFN/PFN: 1810.05165

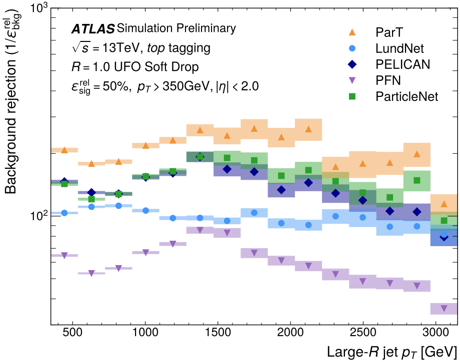


ParticleTransformer: 2202.0377



ParticleNet: 1902.08570

ATLAS conducted a semi-inclusive literature comparison for t/W-tagging



ParticleTransformer performs best for this problem *so far*

- Not all models sizes were equal
- No hyper-paramater turning
- Physics inspired networks on QCD or symmetry didn't improve
  - *"The bitter lesson"*
- Some different results w.r.t. external papers
  - Full detector simulation can be important

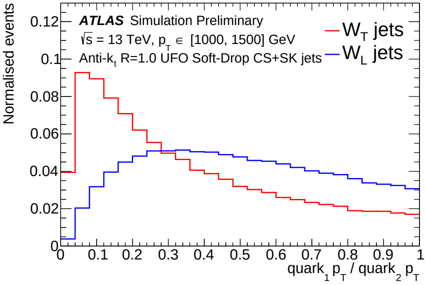
Polarized W/Z Tagger

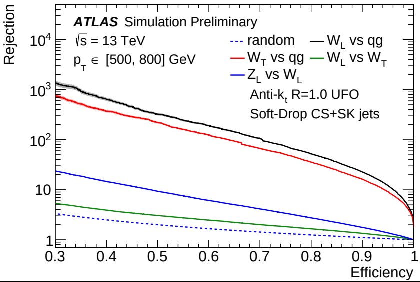
Polarized boson-tagging very interesting for physics, but tricky to tag

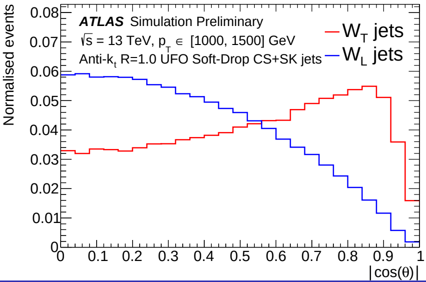
Design a ParT tagger boosted with physics info:

- Lab-frame $V \to qq$ is core discriminant, but poor reco-analogy
- **Adding regression task to improve classifier results!**

Reasonable separation power between polarization classes and QCD!

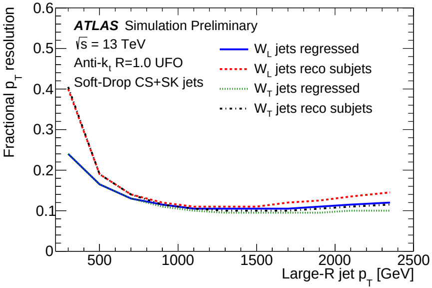




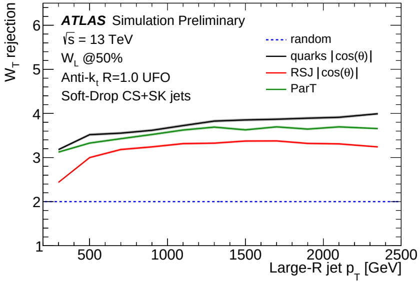


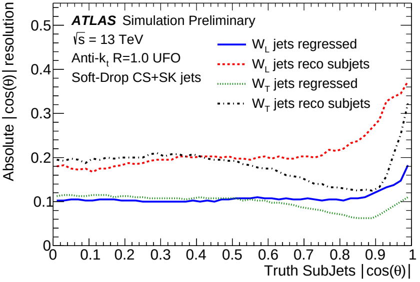
Tagger has two inputs/tasks

- Jet constituent and truth labels to classify
- Two exclusive C/A subjets regressed to truth quarks



Bottom left: **Tagger (green) close to theory bounds from quark info (black)**

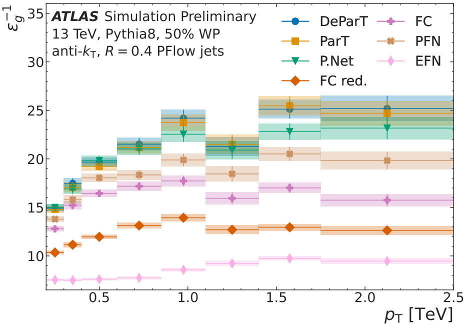


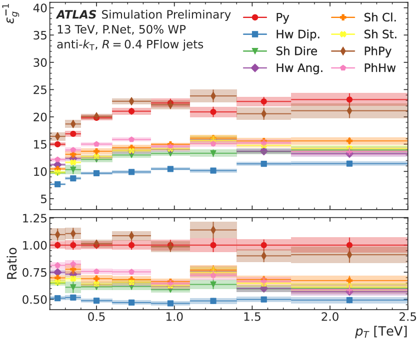


Constituent q/g-tagging

New result in ATLAS on q/g-tagging!

- Harder problem than $W/t/h/b/\tau$-tagging
  - Main differences: colour-factors $\rightarrow$ more radiation in gluon-jets
- Transformer models with additional interaction terms performs best
  - **Order magnitude better then cut-based and full pt/eta coverage**
- Differences in simulations can be large!

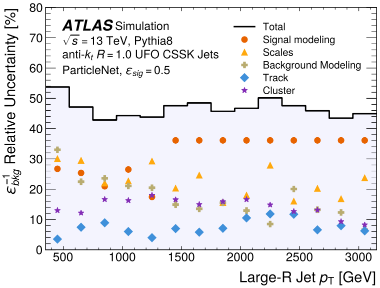


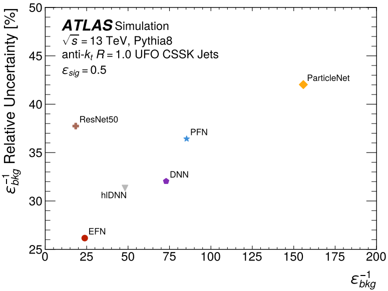


As we provide lower-level information taggers start to learn generator specific features $\rightarrow$ **higher uncertainties at analysis level**

Paper evaluating *approximate* bottom-up experimental uncertainties and theory uncertainities on top-taggers

- Better taggers have higher uncertainties by almost factor 2
- Want to develop ways to break this trend
  - EFN less model dependency since focus on IRC-safe observables?





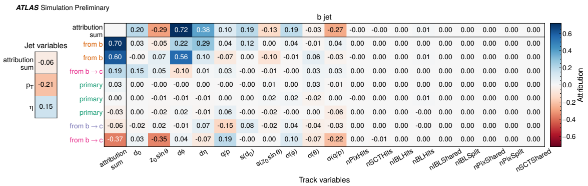
Open-data with documentation available!

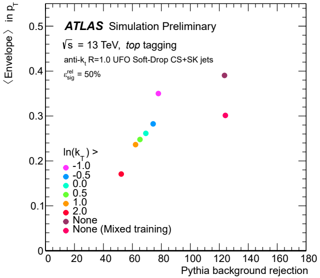
Train a network inspired by ParticleNet on the Lund-plane points:

- Graph neural-network on C/A re-clustering history of jet formation
- Can clip out regions of Lund jet plane to remove IRC unsafe regions
  - **Reduces modelling uncertainty, but also the performance**

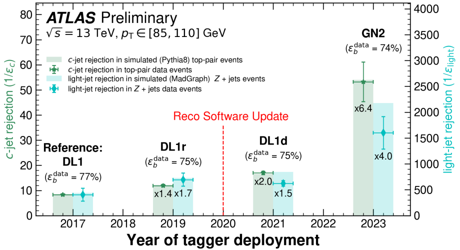
Also used integrated gradients as a feature importance metric for b-tagging

- Explains what physical features the network is learning
- **Helps build trust in network performance**

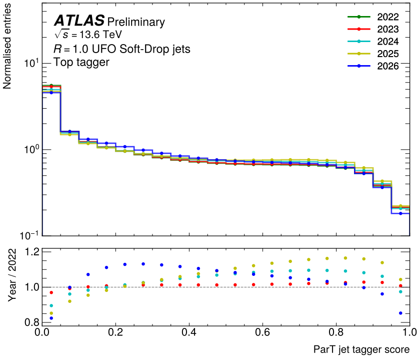




Our real results perform worse in data!



Can vary across detector conditions!



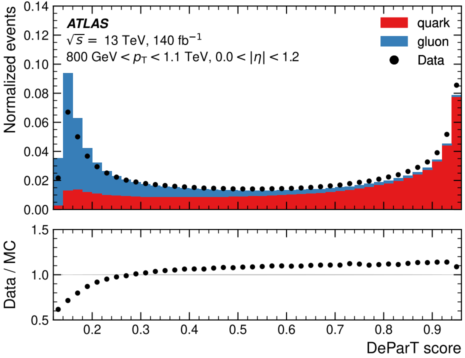
$\Longrightarrow$ **Needs calibrations!**

Can use data-science ideas to reduce the calibration uncertainty from demixing forward (gluon-like) vs central (quark-like) jets

Matrix-method:

$$p_C(x) = f_C^q p_q(x) + f_C^g p_g(x)$$
$$p_F(x) = f_F^q p_q(x) + f_F^g p_g(x)$$

Evaluate fraction from simulation and invert matrix

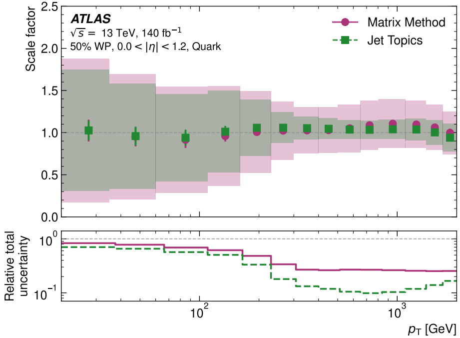


Jet-Topics:

$$p_q(x) = (p_F(x) - \kappa(F|C)p_C(x))/(1-\kappa(F|C))$$
$$p_g(x) = (p_C(x) - \kappa(C|F)p_F(x))/(1-\kappa(C|F))$$

$$\kappa(F|C) = \min_x \frac{p_F(x)}{p_C(x)} \quad \kappa(C|F) = \min_x \frac{p_C(x)}{p_F(x)}$$

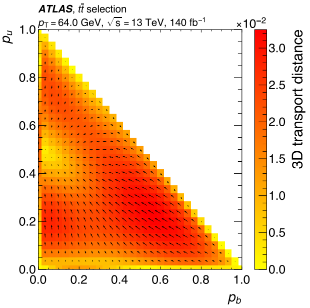
Find optimal demixing points in data

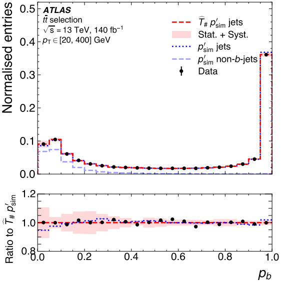


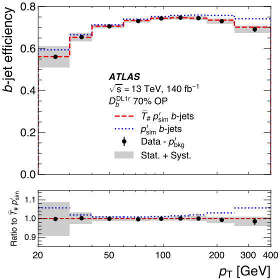
Optimal Transport for Calibrations

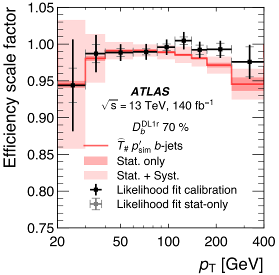
Can phrase calibration as an optimal transport problem

- Uses partially input convex neural network (PICNN) and normalizing flows
- Provide **calibrated continuous output nodes** as opposed to fixed working points
- Extensive tests on new methods in real sig/bkg mixed data
- Good closure to data, and conventional calibration

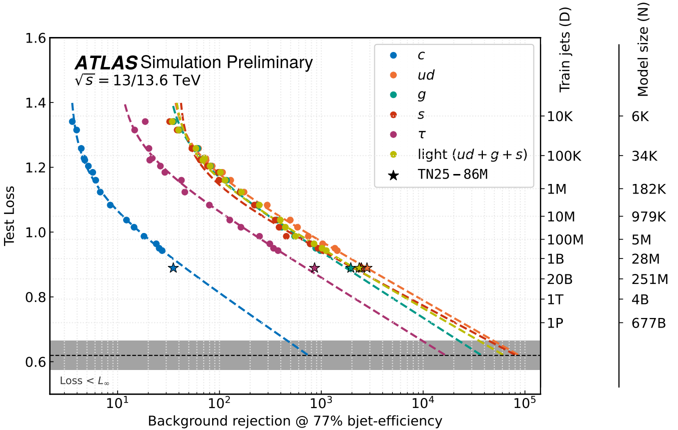








Network performance follows industry results of neural scaling laws



Summary

Presented several results on ATLAS ML techniques for hadronic object identification:

- ML approaches in W/top and q/g-tagging
- And their calibration with ML techniques

Usage of ML techniques have provided improvements almost everywhere

- Multi-task networks seem to be next step

But it is not the end of the line:

- New ML algorithms always incoming
  - Full detector performance is important for these
- Robustness may soon become a bottleneck
  - Need to consider the uncertainty and not just ROC-like metrics
- Interpretability can help with this
  - ATLAS making first studies in this area

BACKUP

Calibrations

The tagger approach to substructure limited by capability to calibrate

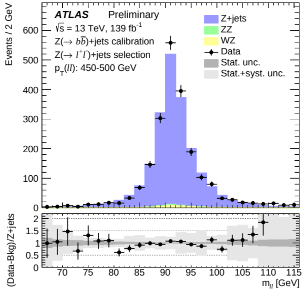
- Train on simulation, but also apply to data

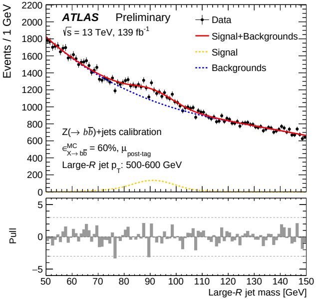
Often use top-down approach, where calibrate to a well known resonance

- $t\bar{t}$, $V + \text{jets}$, $V + \gamma$
- Can measure QCD jet rejection in multijet topologies

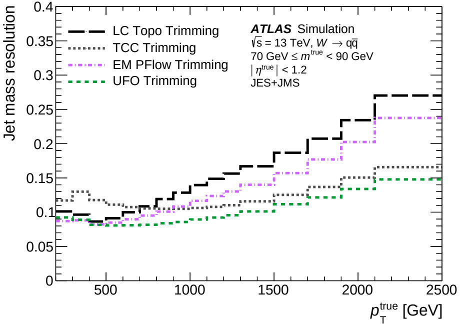
Difficulties can arise:

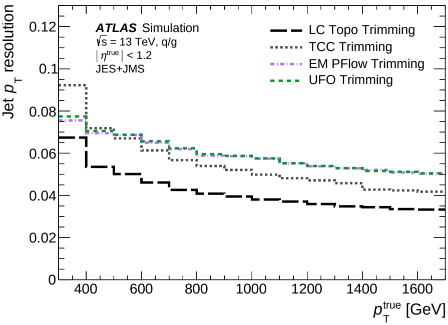
- No easy SM resonance: e.x. calibrating to h
- Limited probe statistics at high-p_T
- Limited trigger/selection capabilities at low-p_T





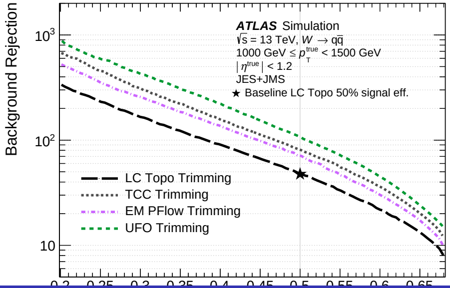
Jet collections





Tagger performance deeply connected to jet input definition!

- Over Run2 moved from Topological Clusters → Track-Calo Clusters → Unified Flow-Objects
- Better mass+substructure resolution



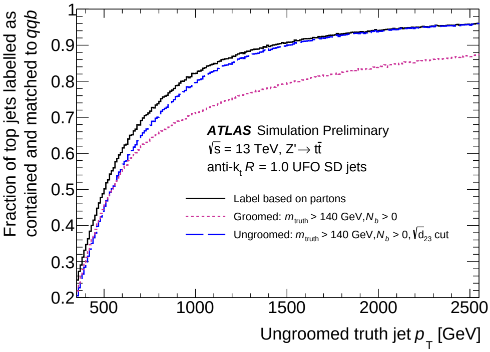
Taggers Definitions

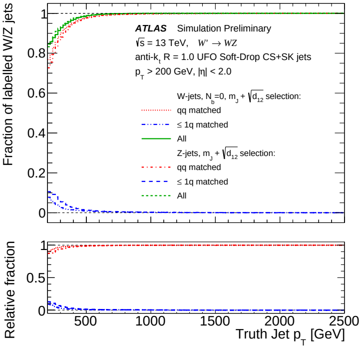
Over several publications iterated on top and W truth label defintions

- Move from parton history to ghost-associated matching between truth and reco jets

More robust truth-labeling helps with cleaner classification labels

- Less generator dependence

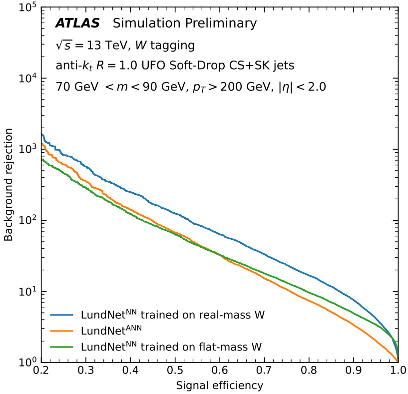


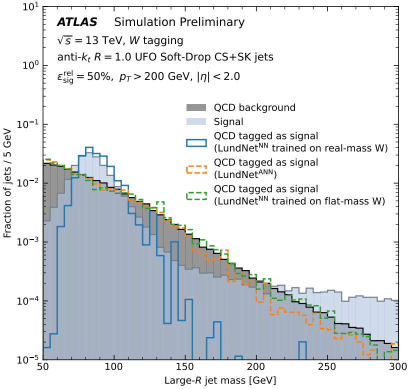


Mass Decorrelation

Often analyses want a mass deccorelated W-tagger

- But the network learns that the mass is very good discriminant
  - Sculpts background to look like signal!
  - Not an input to the network!
- Tried adversarial techniques, but found that custom simulation with flat-mass coverage and reweighing the background performed best
  - Similar rejections, but better performance in tails



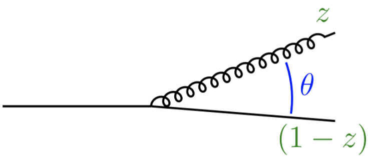


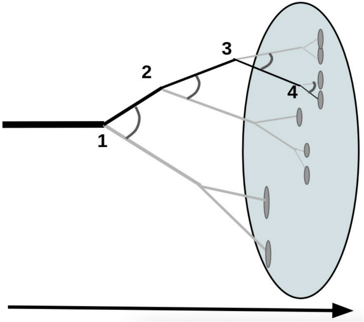
Can recluster jet with C/A algorithm and each split can be represented on lund-plane based on:

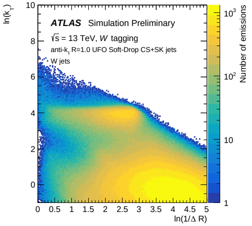
transverse momentum k_T, angle Δ, and momentum fraction z

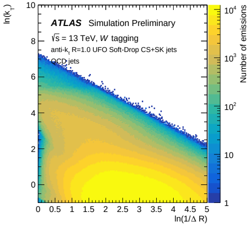
Variables can distinguish splitting type:

- ISR: low $\ln(1/\Delta)$
- Non-perturbative: low $\ln(k_T)$
- hard colinear: high $z \rightarrow$ top graph edge







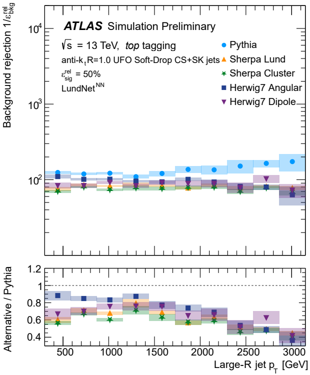


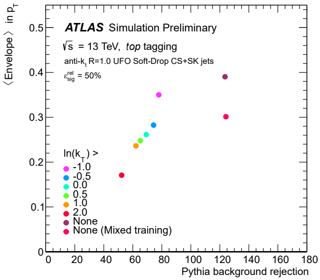
Train a network inspired by ParticleNet on the Lund-plane points:

- Graph neural-network on C/A re-clustering history of jet formation

But also see comparably modeling dependency as other networks

- Lund-plane point position can be suggestive if IR/collinear splitting
- Can clip out regions of Lund jet plane to remove IRC unsafe regions?

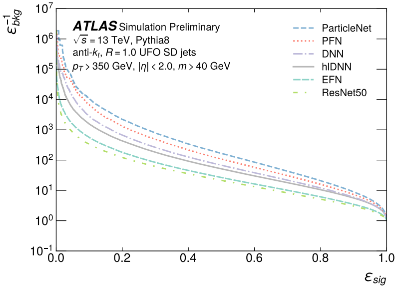


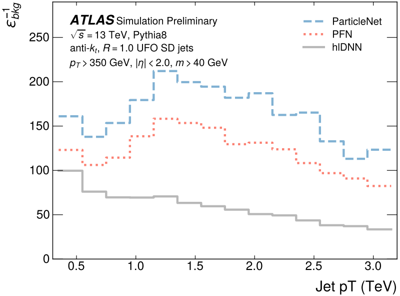


Constituent based top taggers outperform high-level features ones:

- Provide network the lowest level information available: the jet constituents themselves
- Another factor 2-3 improvement!
- ResNet/EFN under-perform w.r.t theoretical performance
  - Real simulation studies important!

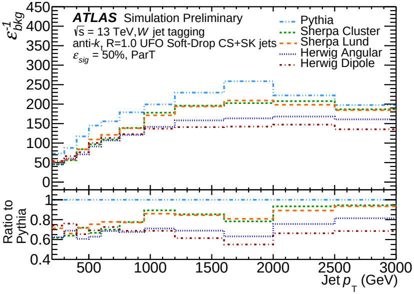
Model	AUC	ACC	ε_{bkg}^{-1} @ $\varepsilon_{sig} = 0.5$	ε_{bkg}^{-1} @ $\varepsilon_{sig} = 0.8$	# Params	Inference Time
ResNet 50	0.885	0.803	21.4	5.13	1,486,209	9 ms
EFN	0.901	0.819	26.6	6.12	1,670,451	4 ms
hlDNN	0.938	0.863	51.5	10.5	93,151	3 ms
DNN	0.942	0.868	67.7	12.0	876,641	3 ms
PFN	0.954	0.882	108.0	15.9	689,801	4 ms
ParticleNet	0.961	0.894	153.7	20.4	764,887	38 ms





Similar results for constituent based W-tagger

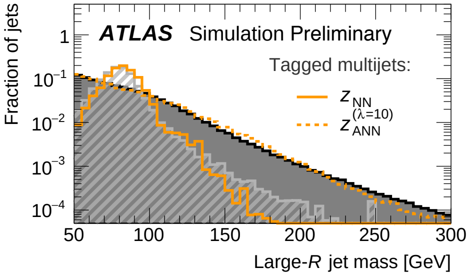
- Transformers performing best on this dataset
- But still large modeling uncertainty

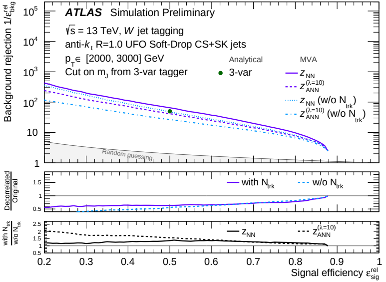


Model	AUC	ACC	ε_{bkg}^{-1} @ $\varepsilon_{sig} = 0.5$	ε_{bkg}^{-1} @ $\varepsilon_{sig} = 0.8$	# Params	Inference Time
EFN	0.920	0.835	35.1	7.95	56.73k	0.065 ms
PFN	0.931	0.853	44.7	9.50	57.13k	0.11 ms
ParticleNet	0.933	0.826	46.2	9.76	366.16k	0.36 ms
ParticleTransformer	0.951	0.880	77.9	14.6	2.14M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

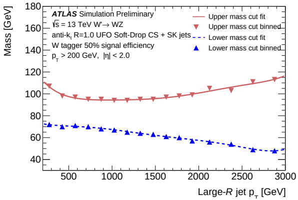
Similar to top-tagger, developed a DNN W-tagger over 10 substructure variables

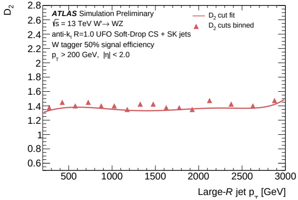
- But obviously the DNN learns that the mass is very good discriminant
  - Sculpts background to look like signal!
  - Not an input to the network
  - Difficult to use sideband regions in analysis
- Train against a second "adversarial" network to force network to decorrelate the feature: Loss function $L = L_{classifier} - \lambda L_{adv}$
- Decorrelated tagger similar to cut-based.
- Correlated tagger has 10x better background rejection

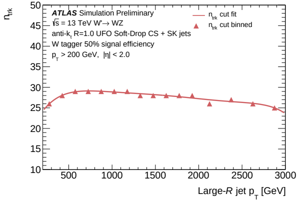




Historic taggers: W/Z-tagger







Baseline W/Z-taggers provided at fixed $50/80\%$ signal efficiency

- Fixed-cut 3-variable tagger: Jet mass, $D_2^{\beta=1}$, $N(tracks)$
- Flat efficiency as a function of jet p_{T}
- Different performance for transverse/longitudinal bosons