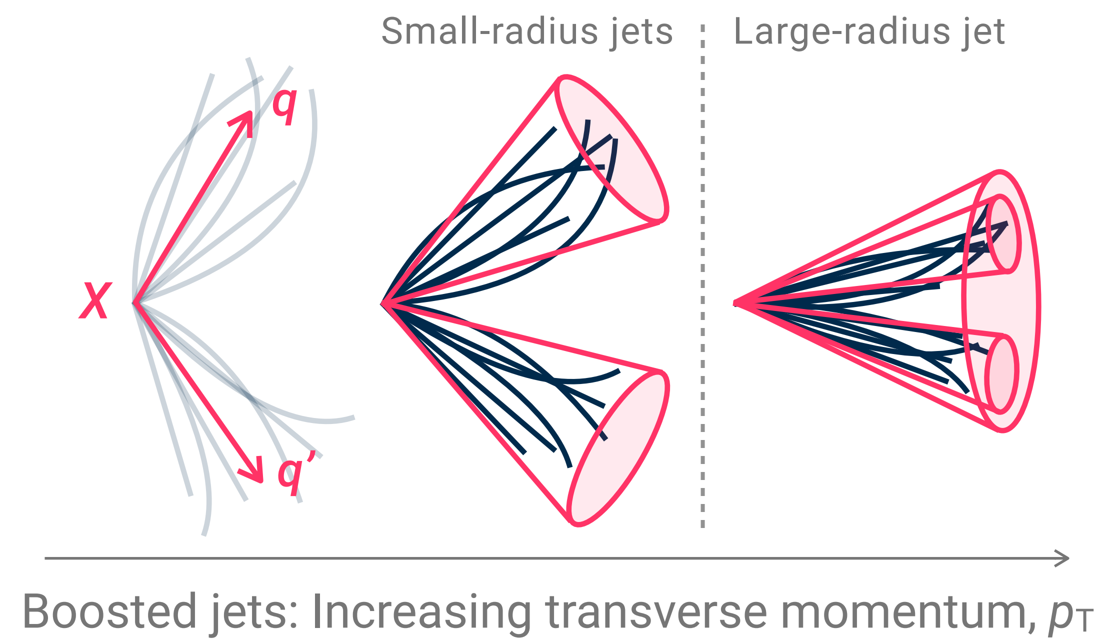
Boosted *W*/top tagging in ATLAS

BOOST 2018 / UPMC, Paris / 17 July 2018

Andreas Søgaard
on behalf of the ATLAS Collaboration

Introduction

- Identifying boosted hadronic resonance decays is guided by
  - precision **jet substructure calculations** — Talk by J. Roloff
- and is important for
  - probing **top physics** — Talk by W. Hopkins
  - searches for **physics beyond the SM** — Talk by F. Guescini
- Using trimmed anti-k_t R = 1.0 jets*: — Talk by A. Arce — Talk by S. Ganguly
  - Study **performance, modelling** of *W*/top taggers in data
  - Present **new ideas** in jet tagging



*Trimming params.: R_{sub} = 0.2, f_{cut} = 5% [0912.1342]. HEPTopTagger uses trimmed C/A jets w. R = 1.5

Performance in simulation

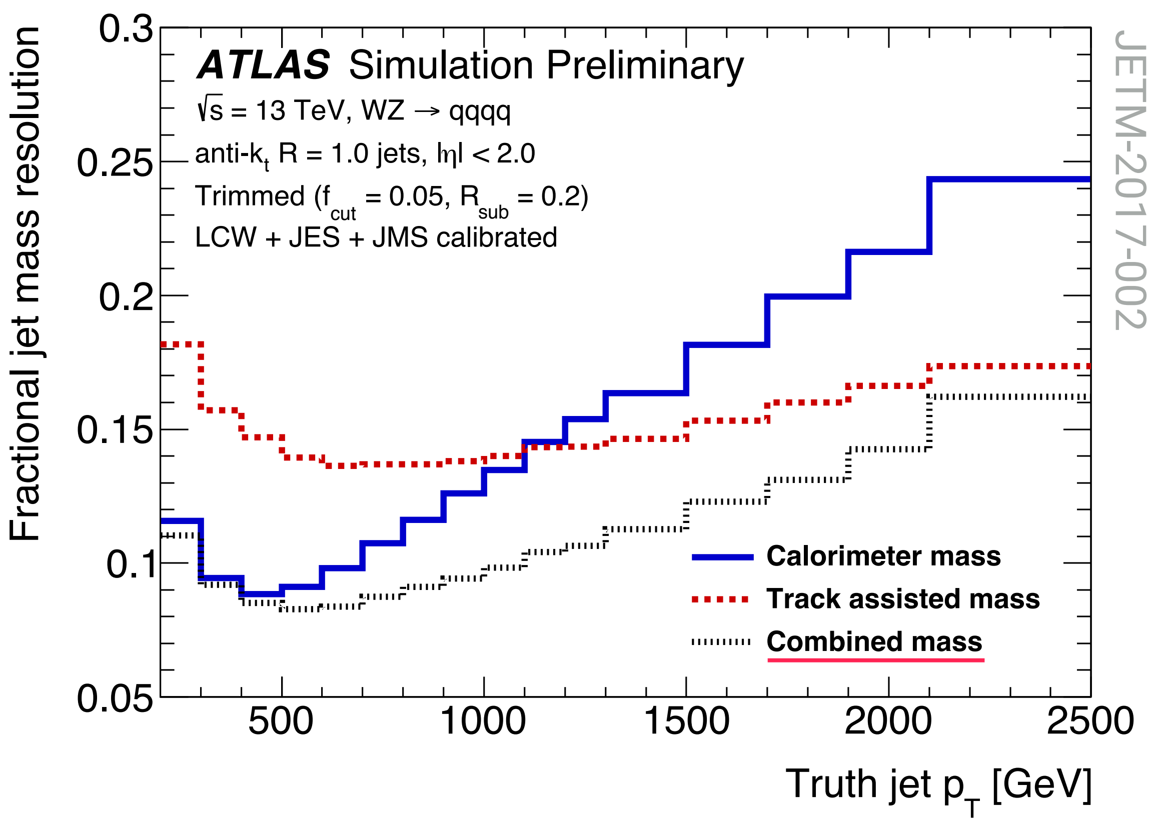
ATLAS Run 2 *W*/top tagging paper (*in prep.*)

ATL-PHYS-PUB-2017-004

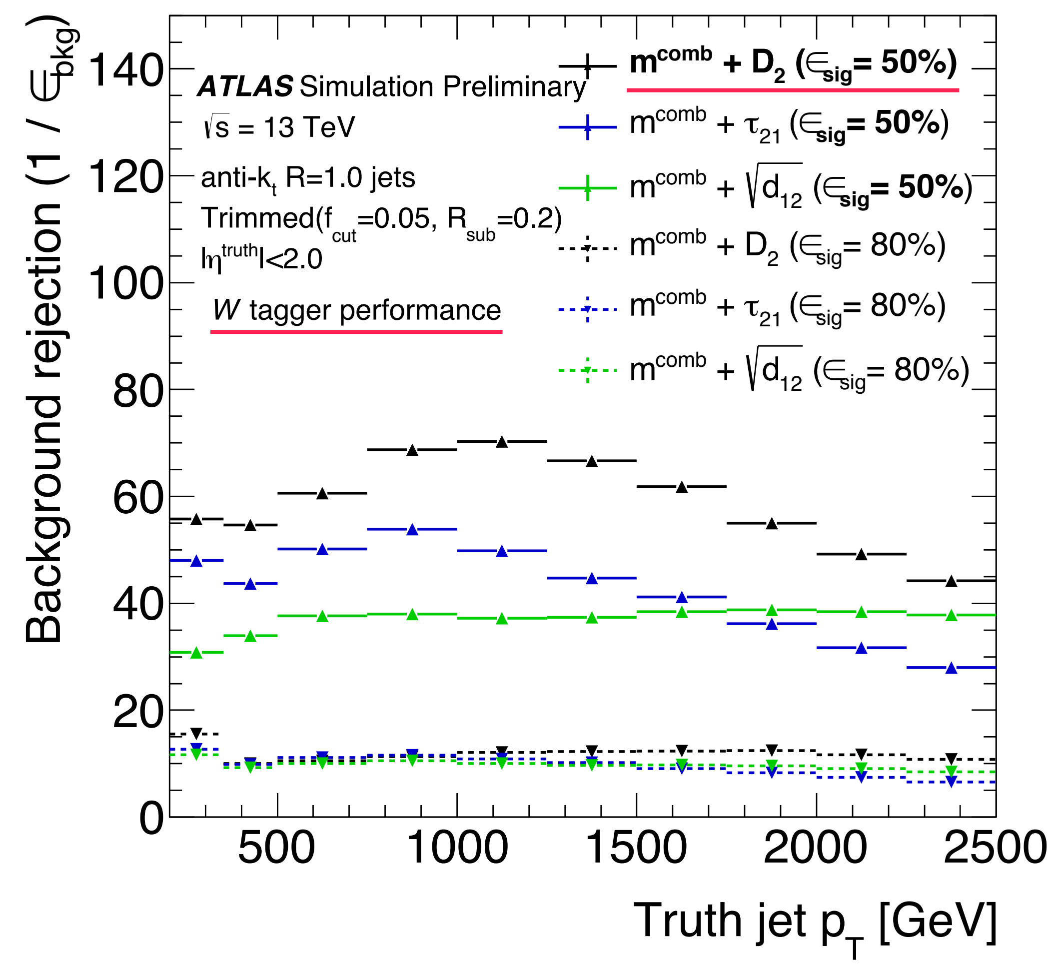
Two-variable cut-based taggers

- Powerful classification by tagging on **combined jet mass** (m^{comb}; track ⊕ calo) and additional **substructure moment**
- Optimise pair-wise combinations at fixed $\varepsilon_{\mathrm{sig}}$ WP vs. p_T
  - ***W*:** 50% WP $m^{\mathrm{comb}} + D_2$
  - **top:** 80% WP $m^{\mathrm{comb}} + \tau_{32}$

Ca. x60 rejection across p_T

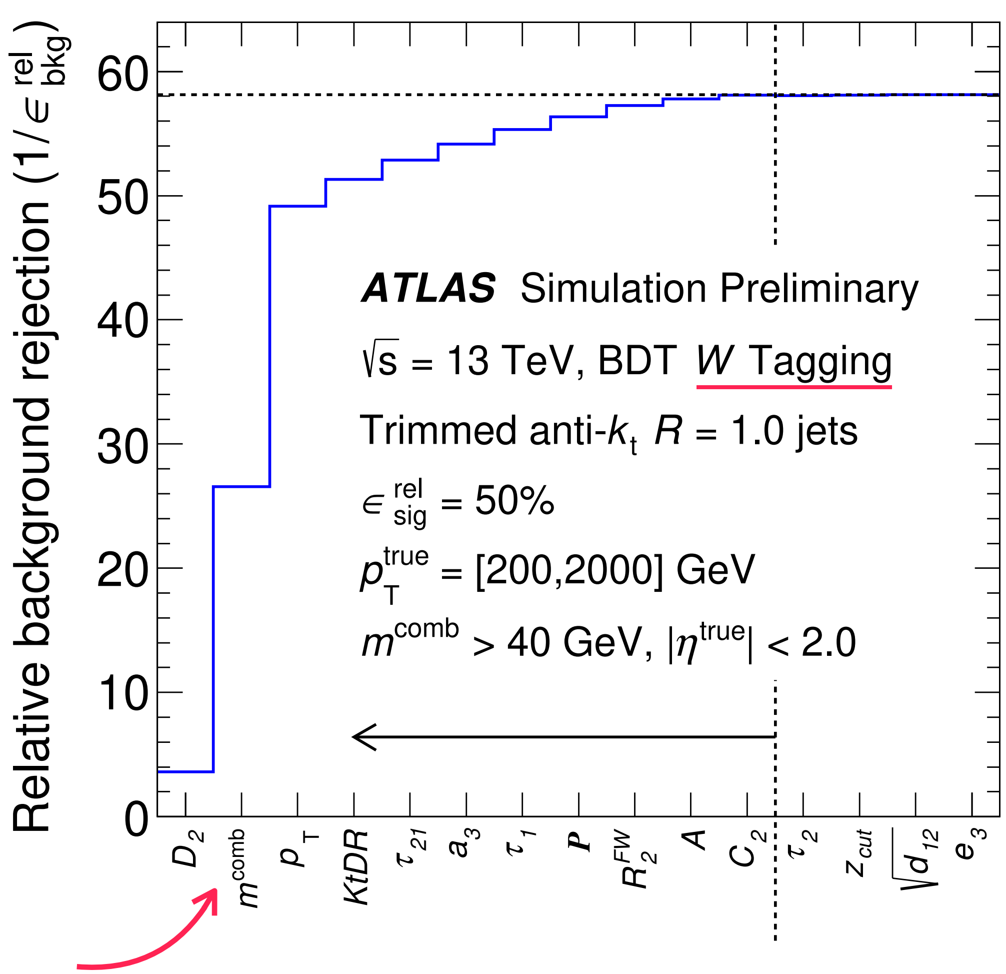


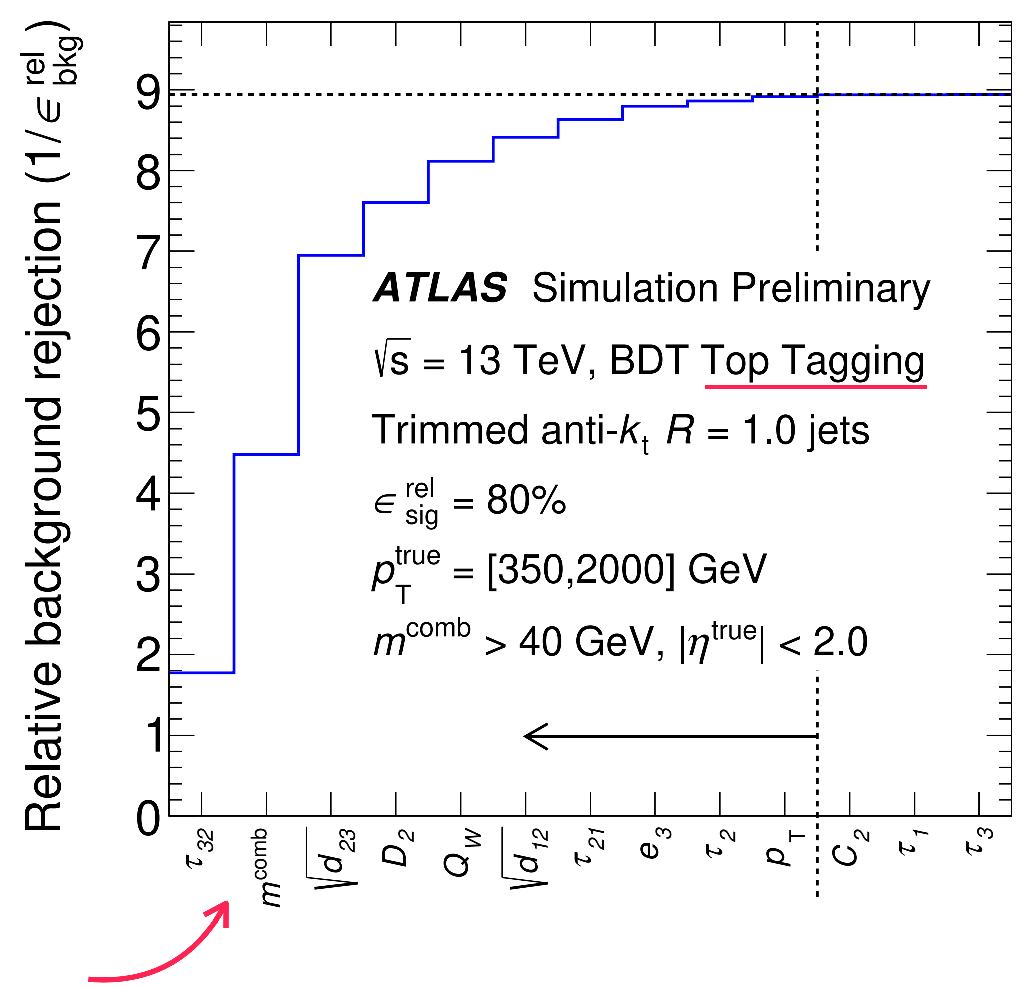
ATLAS-CONF-2017-064



Machine learning taggers

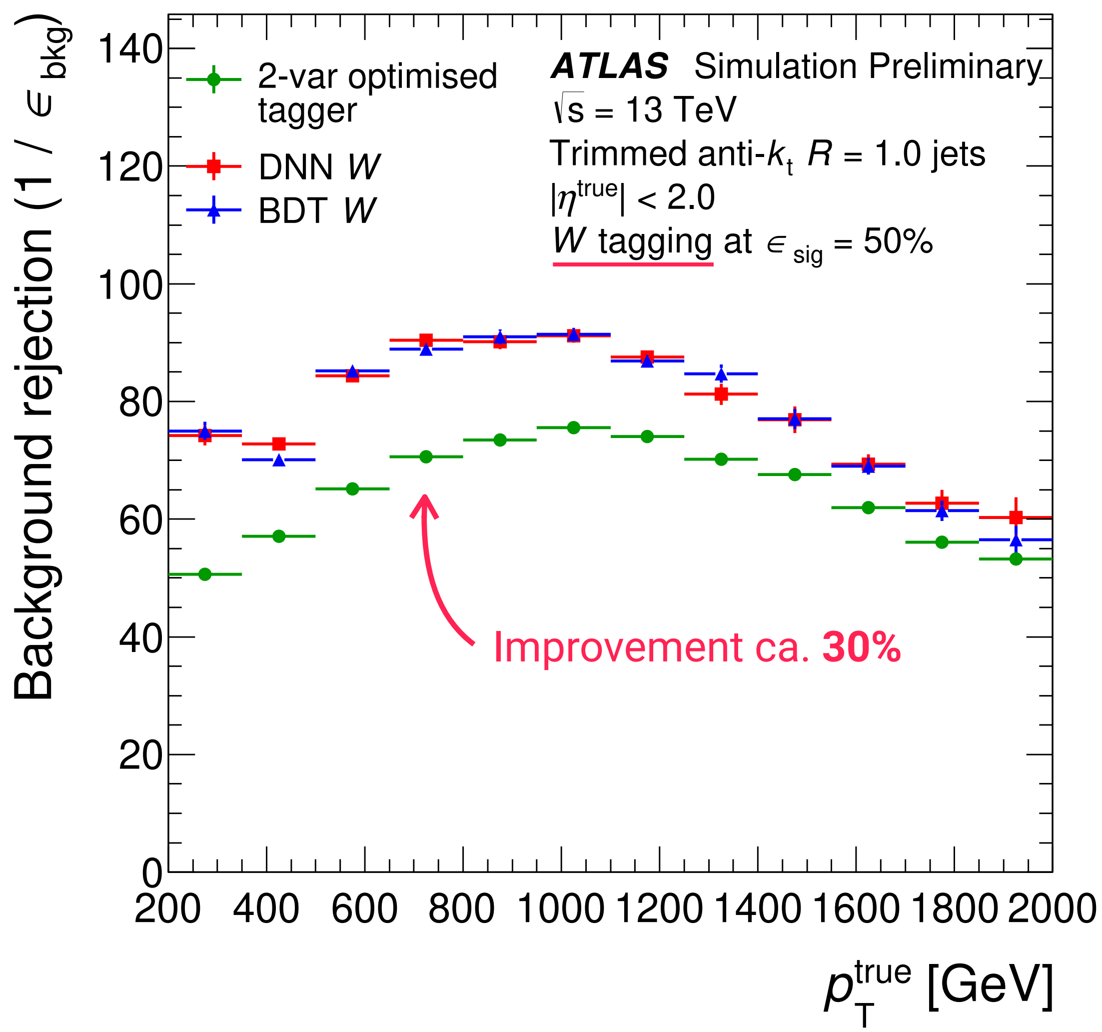
- **Multivariate** (MVA) combination of inputs possible with deep neural networks (**DNN**) and boosted decision trees (**BDT**)
- For BDT, sequentially adding features improves classification beyond two-variable combinations. Similar inputs for DNN.

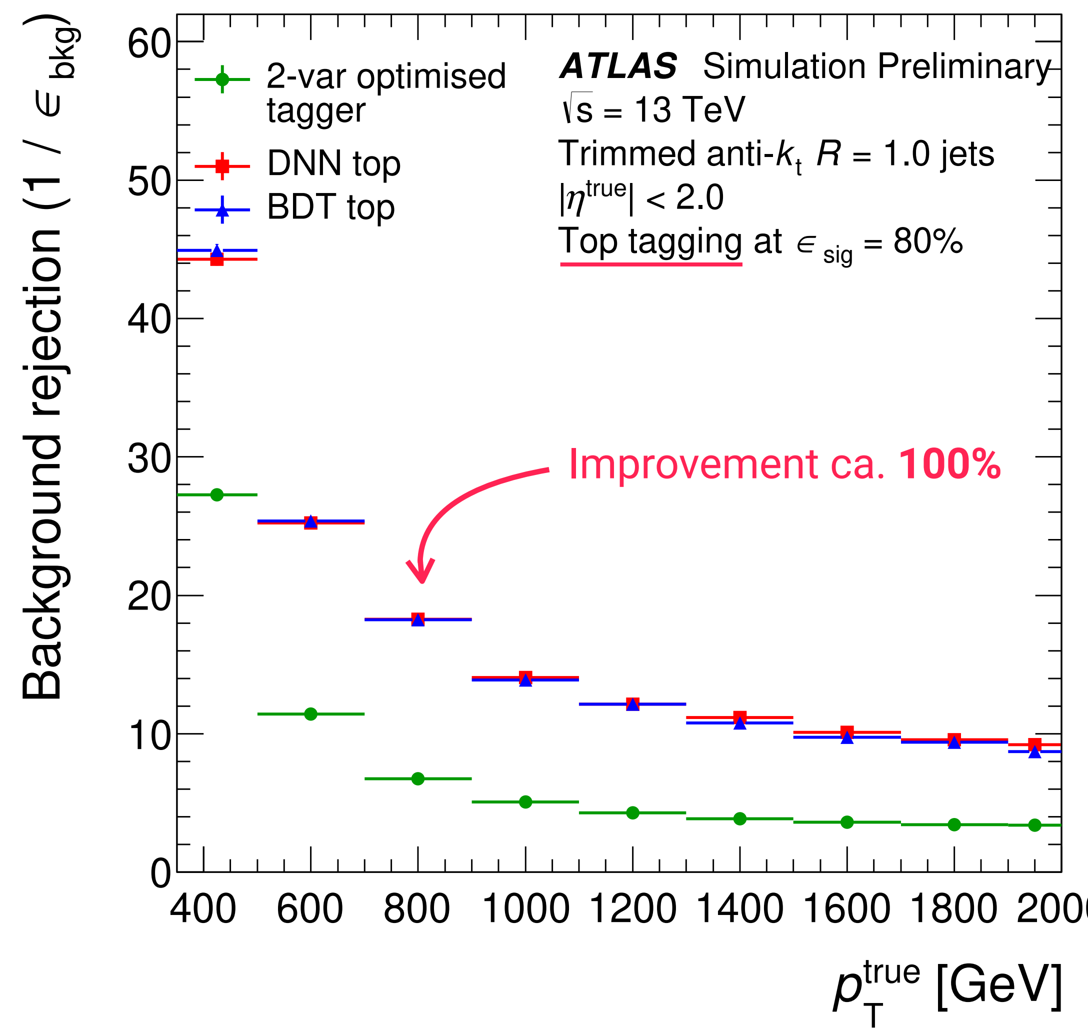




Comparison: Cut-based vs. MVA

- For fixed-ε_{sig} WPs vs. jet p_T, MVA taggers ( ) perform similarly, and better than optimised two-variable tagger ( )
  - Ability to use extra information results in better taggers





Alternative top tagging techniques

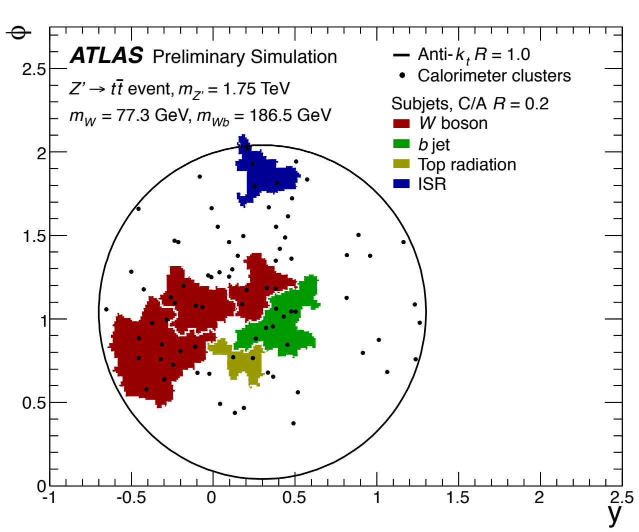
- **HepTopTagger v1 (HTT)**
  [0910.5472, 1006.2833]
- Use trimmed C/A w. R = 1.5 to capture entire top decay for $p_T > 200$ GeV
- Test compatibility with 3-prong top decay hypothesis
- Tag: $m_{HTT} \in [140, 200]$ GeV

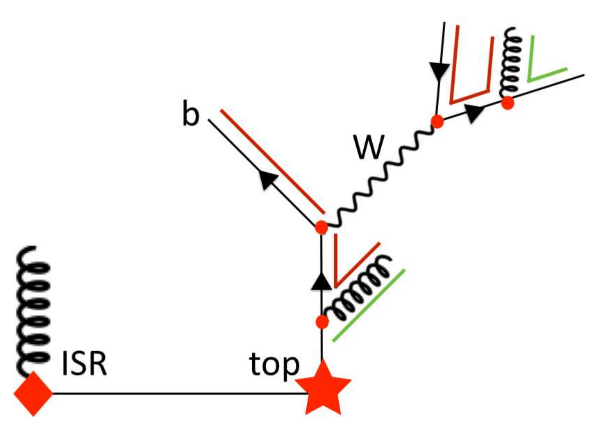
- **Shower Deconstruction (SD)**
  [1102.3480, 1211.3140]
- Re-cluster 3−6 excl. k_t top decay-compatible subjets
- Compare subjets *(≈ partons)* to parton shower histories:

$$\chi(\{p\}_N) = \frac{\sum_{\text{histories}} P(\{p\}_N|S)}{\sum_{\text{histories}} P(\{p\}_N|B)}$$

- **Encoding physics knowledge** (~ME calc.) in observable → between cut-based and ML taggers

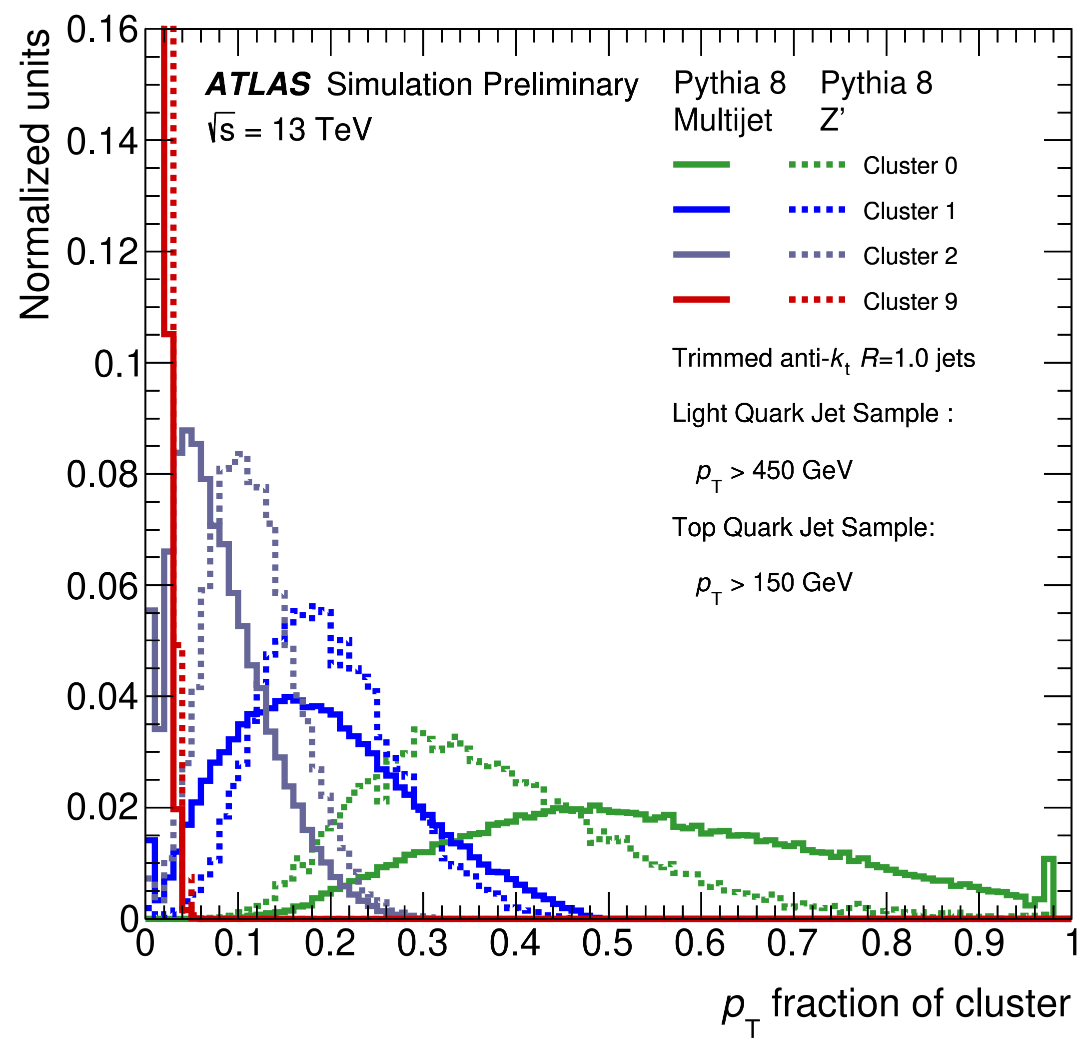
ATLAS-CONF-2014-003





TopoDNN tagger

- **Lessons:**
  - Lower-level inputs can yield more powerful classification [1511.05190, 1603.09349, 1701.08784]
  - Greater impact of ML methods for top tagging
- **TopoDNN:** [1704.02124]
  - Top tagging using jet LC topo cluster constituents directly, implemented and **evaluated in ATLAS**
  - **Targeting high-p_T $t\bar{t}$** analysis, so focus on this kinematic regime

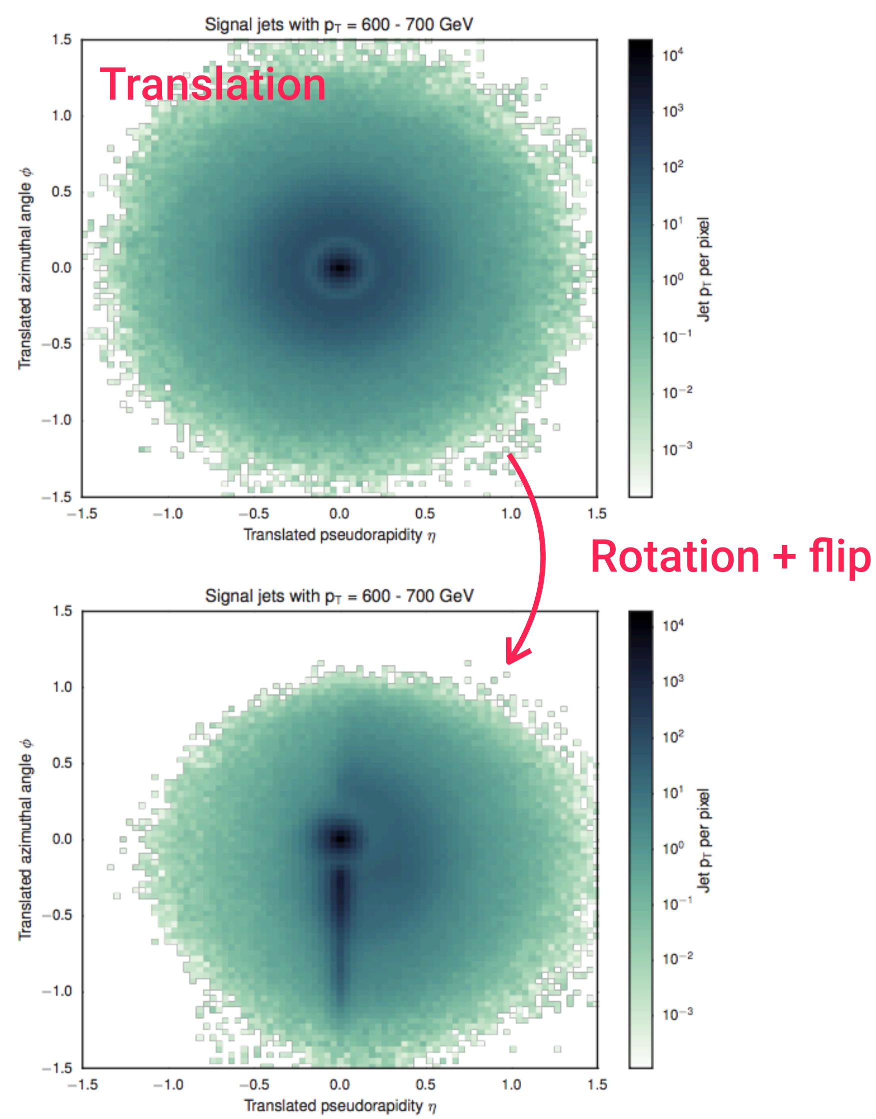


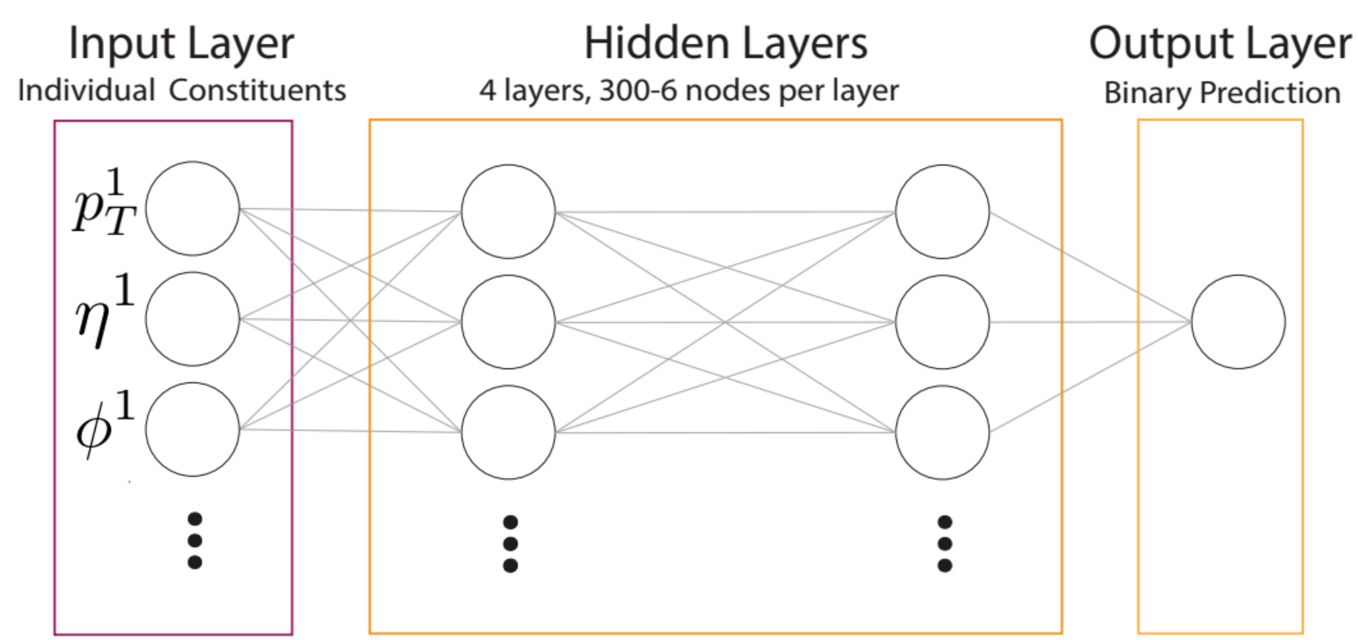
TopoDNN tagger

- Using (p_T, η, φ) of **10 leading LC topo cluster constituents** in trimmed large-R jet ($m_{\text{cluster}} \to 0$)
- Preprocessing: **translation**, **rotation**, and **flip** of assumed three-subjet topology

[1704.02124]

[1704.02124]



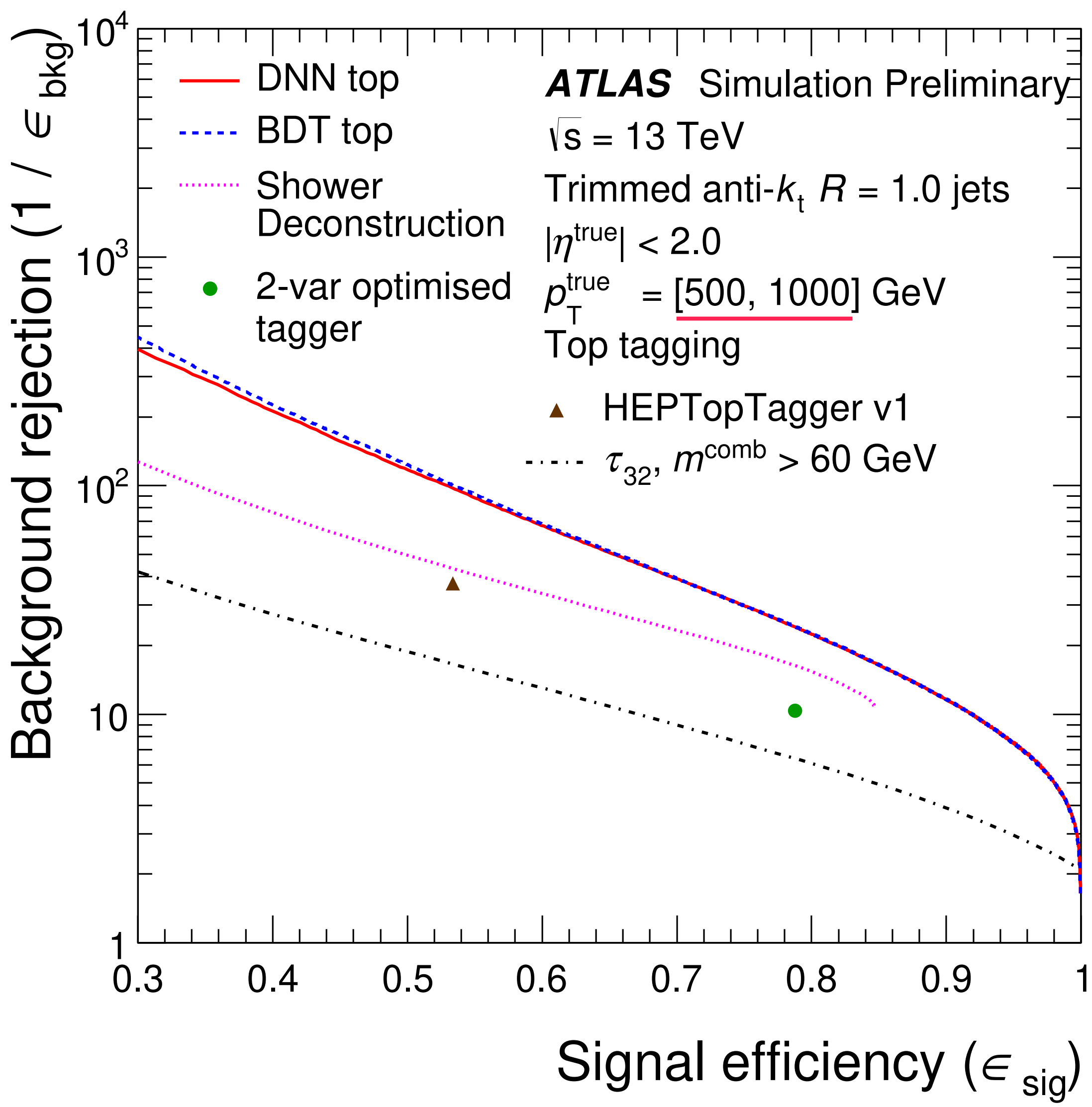


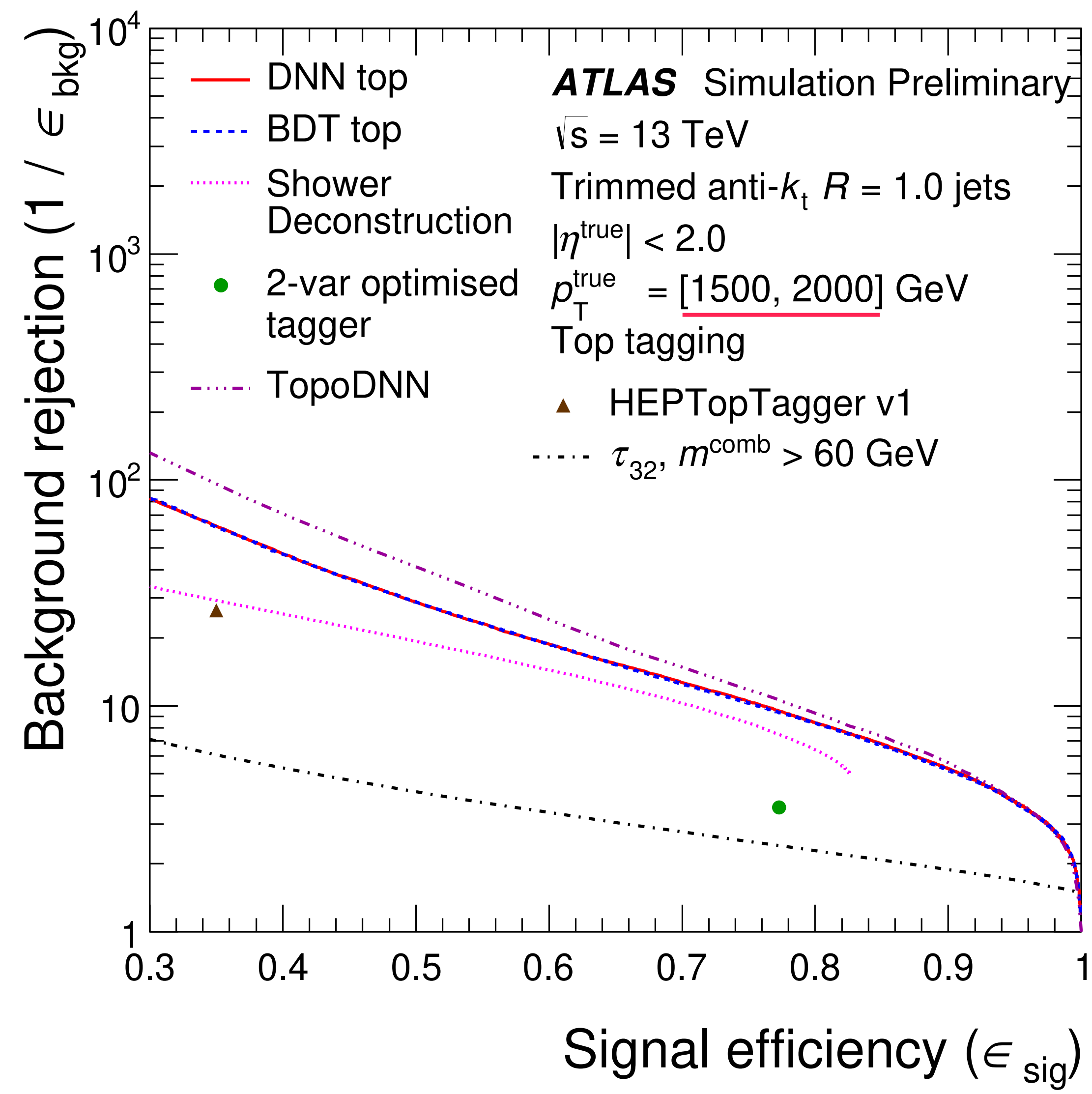
[1704.02124]

Classification / Top

(See backup for *W*)

- Shower deconstruction ( ) most powerful non-ML method
- TopoDNN ( ) improves on “standard” ML taggers ( ) at high p_T



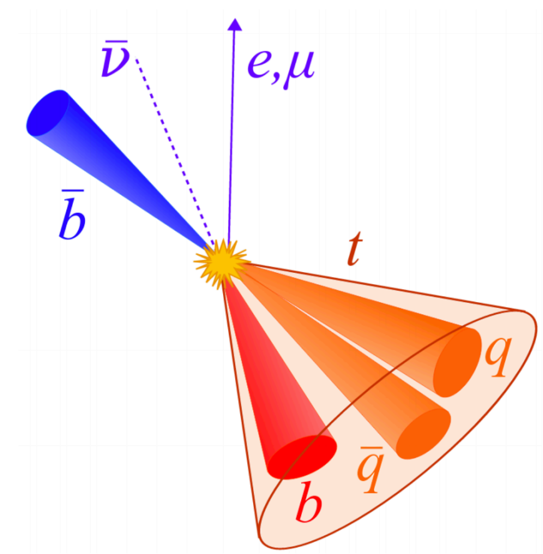


Measurements in data

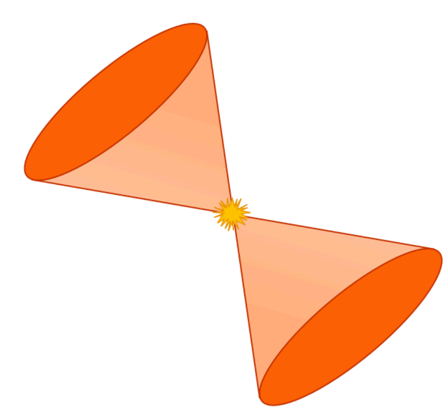
ATLAS Run 2 W/top tagging paper (*in prep.*)

Overview

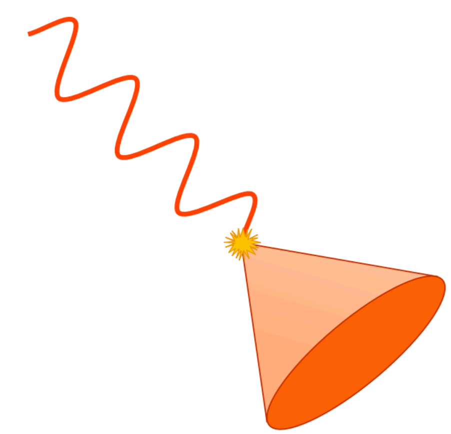
- To use taggers for physics, **must evaluate uncertainty in data**
- Measure *W*/top tagging efficiency and multijet rejection
- Full ATLAS 2015 – 2016 dataset, L = 36.1 – 36.7 fb^{-1}
- ***W*/top tagging efficiency**
  - $t\bar{t}$ decay to single lepton + jets topology
  - ***W*:** ΔR(*b*-jet, large-*R* jet) > 1.0
    $p_T(J)$ > 200 GeV
  - **Top:** ΔR(*b*-jet, large-*R*/HTT jet) < 1.0/1.5
    $p_T(J)$ > 350 GeV
- **Multijet rejection**
  - Dijets: $p_T(J)$ > 450 GeV
  - γ + jets: $p_T(J)$ > 200 GeV



Single-lepton $t\bar{t}$ decay



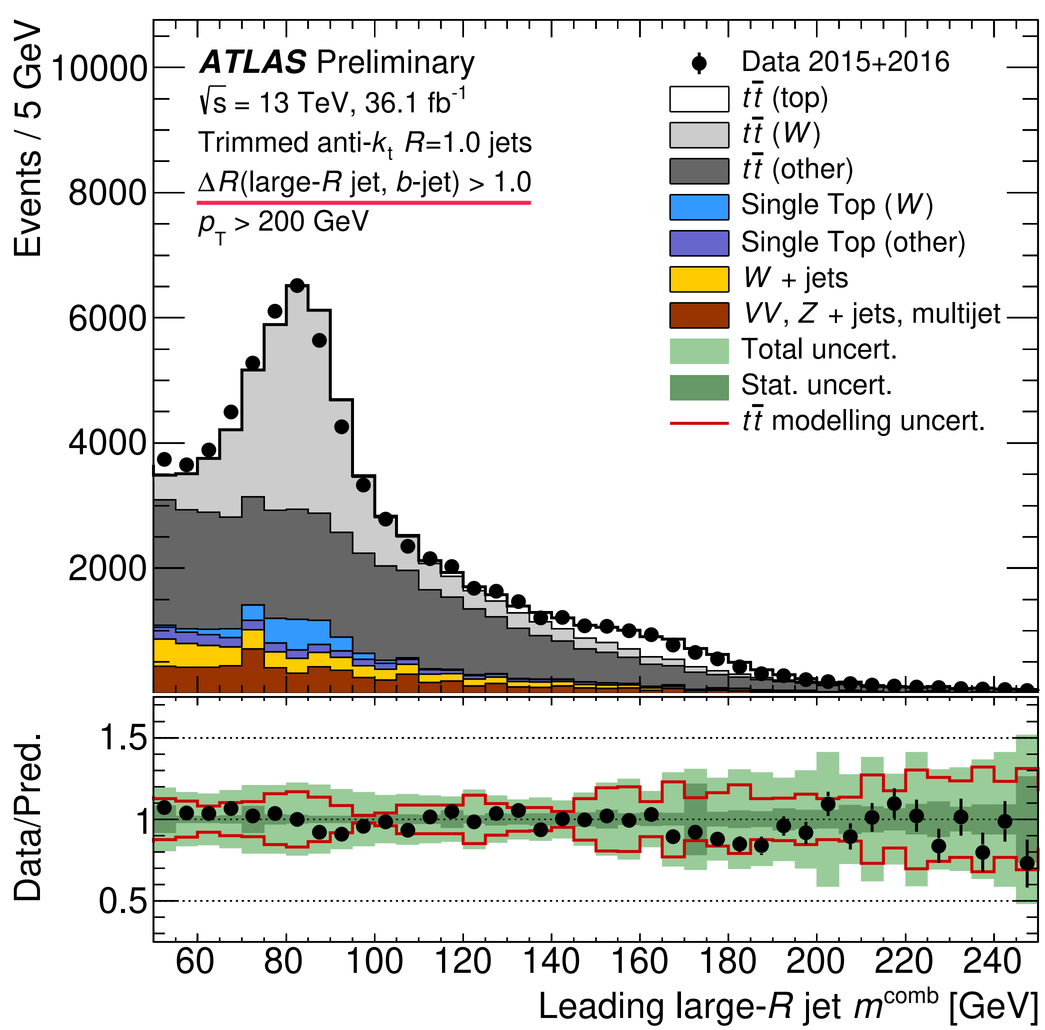
Dijets

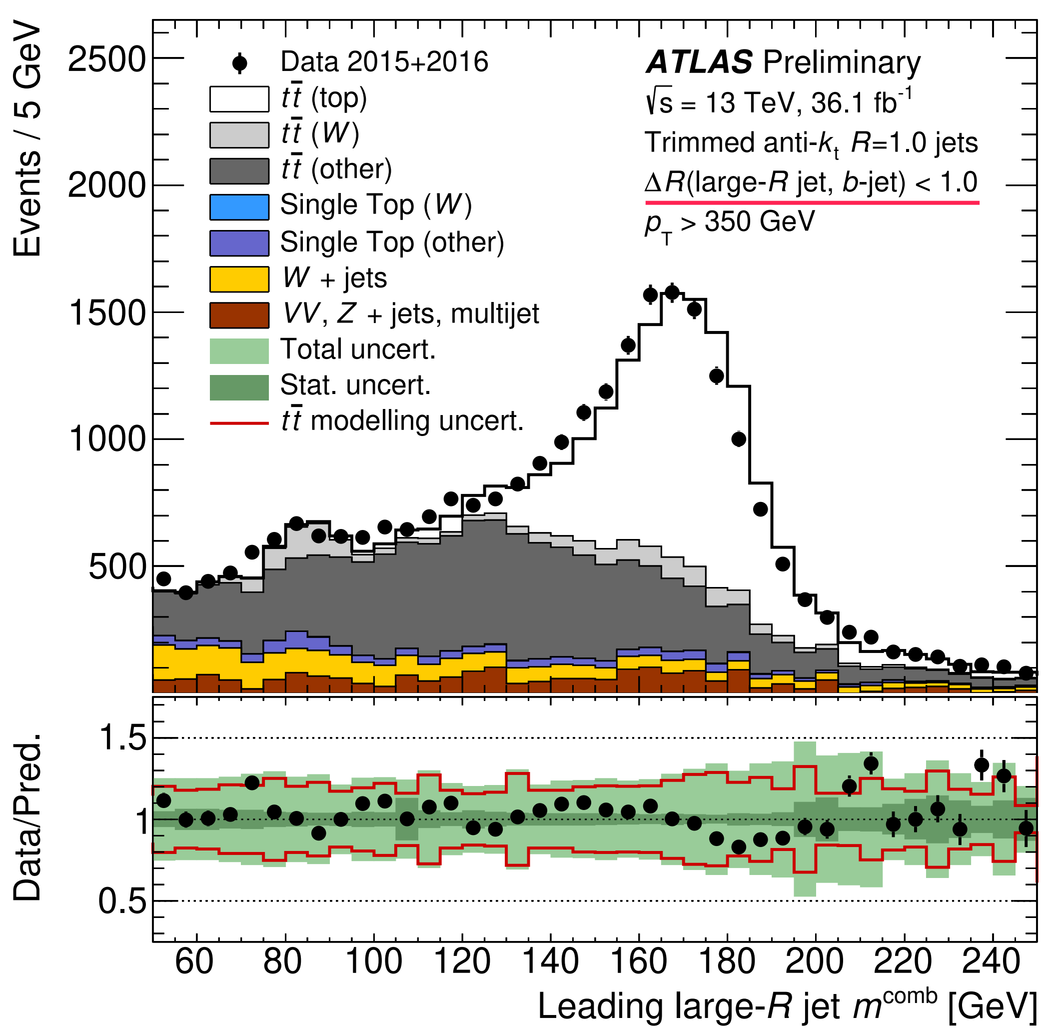


γ+jet

Distributions / Large-*R* jet mass

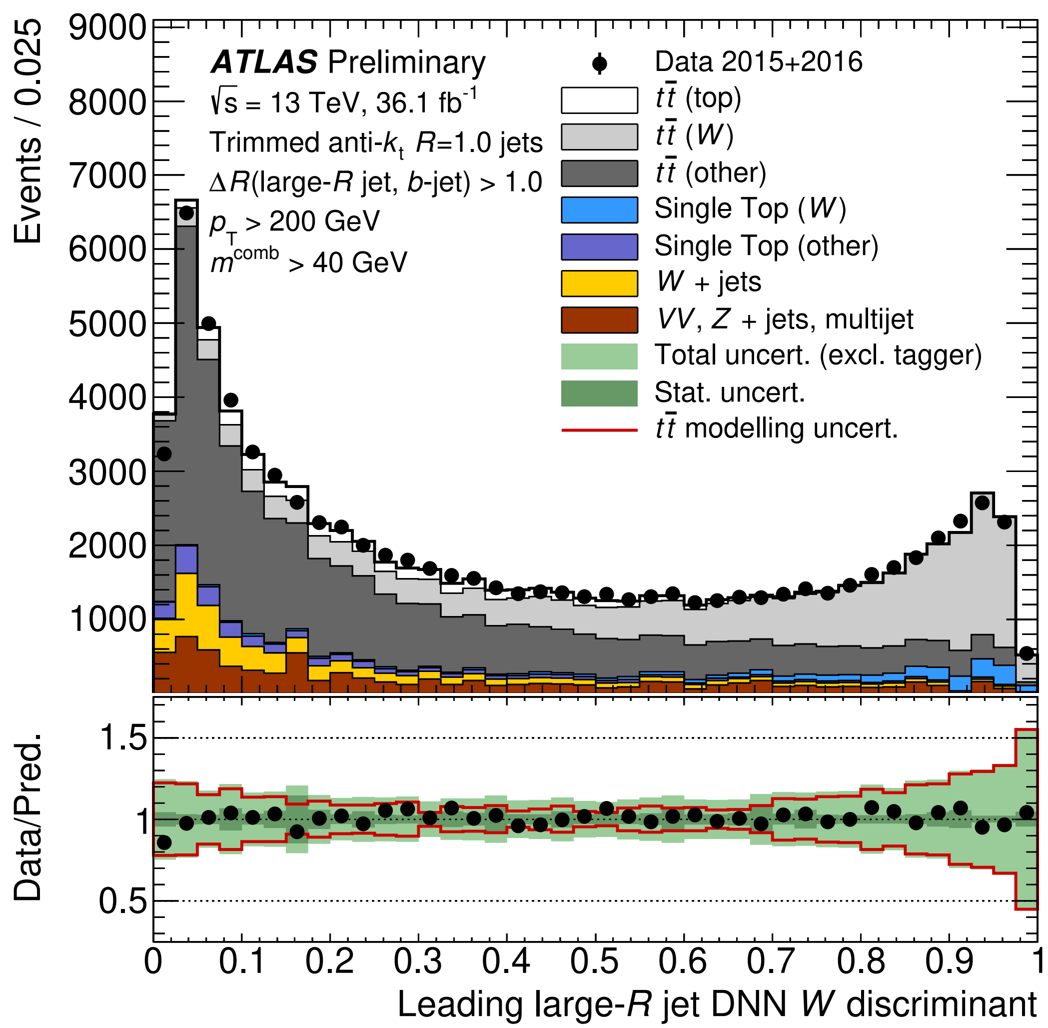
- Enriched samples of *W* and top jets in data, before tagging
- Good MC/data agreement; large $t\bar{t}$ modelling uncertainty

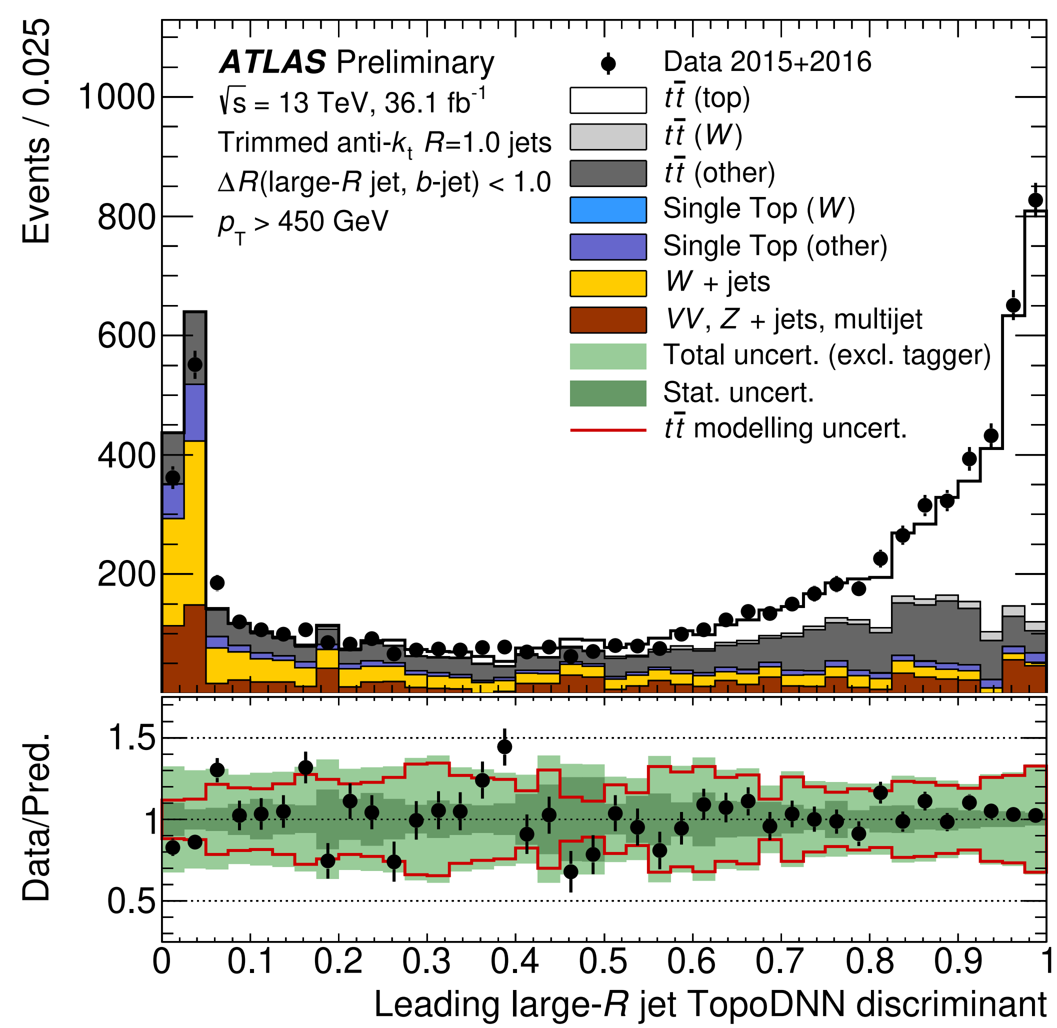




Distributions / ML taggers

- Good modelling of ML tagging variables in simulation:
  - *(left)* High-level ***W* DNN** in *W*-enriched selection
  - *(right)* Low-level **TopoDNN** in top-enriched selection



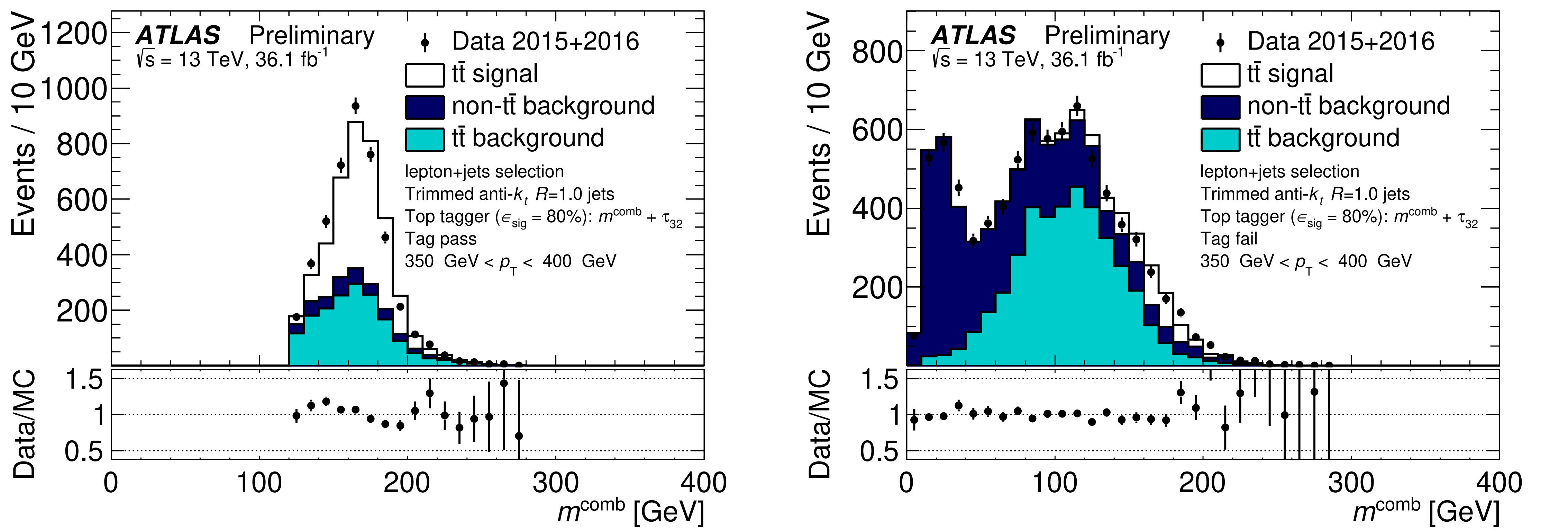


W/top tagging efficiency

- Measure signal-like events in data using signal/background templates to fit large-R jet mass distributions for tag pass/fail

Passing $m^{\text{comb}} + \tau_{32}$ (80% WP) tag

Failing $m^{\text{comb}} + \tau_{32}$ (80% WP) tag



Pre-fit

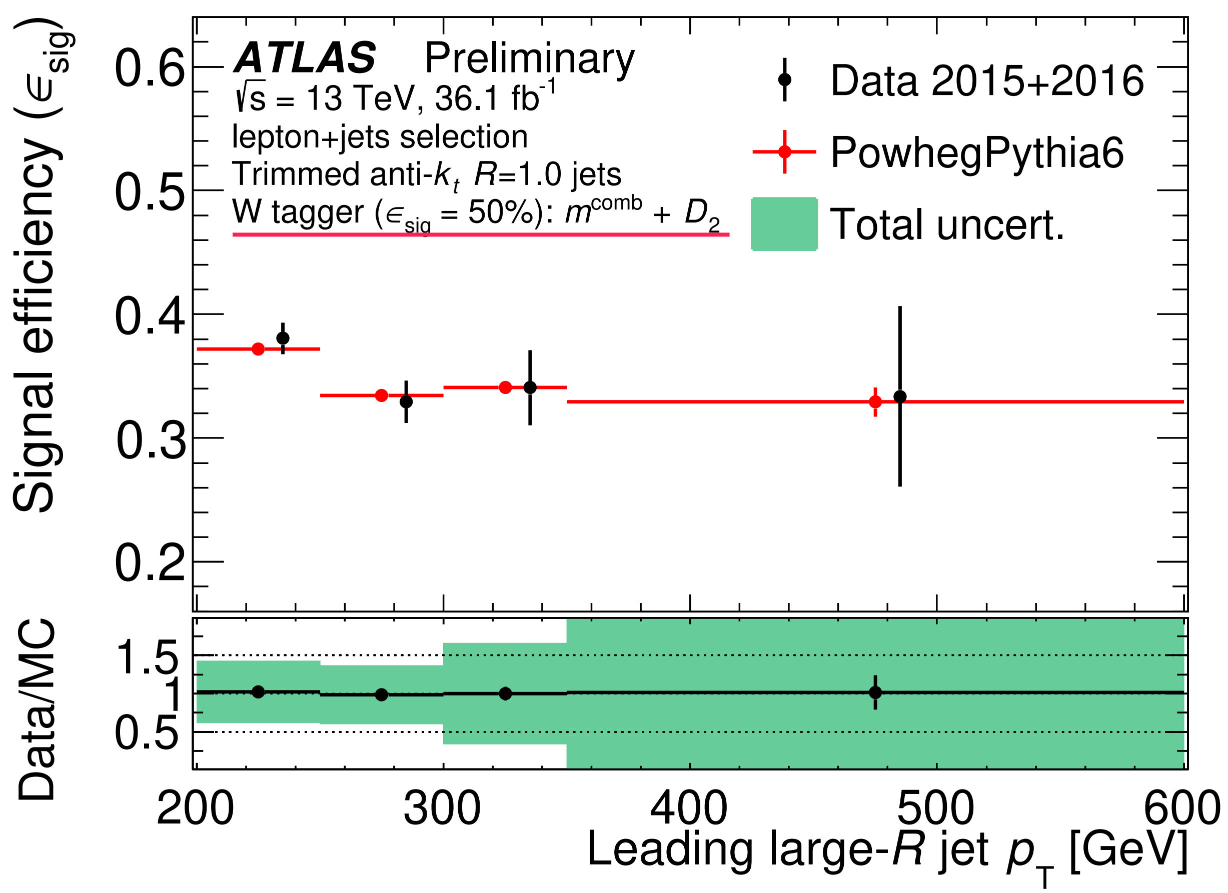
$$\epsilon_{\text{MC}} = \frac{N^{\text{tagged}}_{\text{signal}}}{N^{\text{tagged}}_{\text{signal}} + N^{\text{not tagged}}_{\text{signal}}}$$

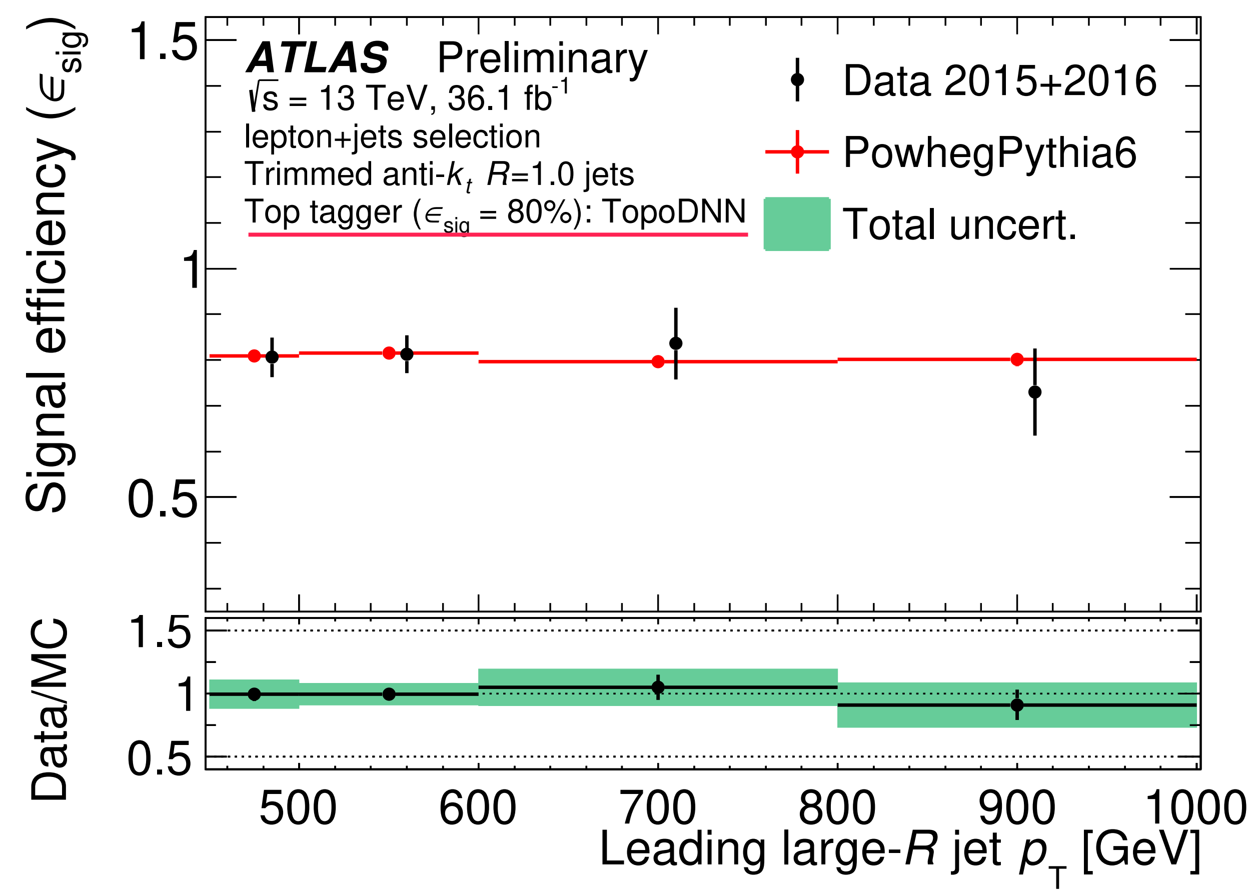
$$\epsilon_{\text{data}} = \frac{N^{\text{tagged}}_{\text{fitted signal}}}{N^{\text{tagged}}_{\text{fitted signal}} + N^{\text{not tagged}}_{\text{fitted signal}}}$$

Post-fit

W/top tagging efficiency / p_T

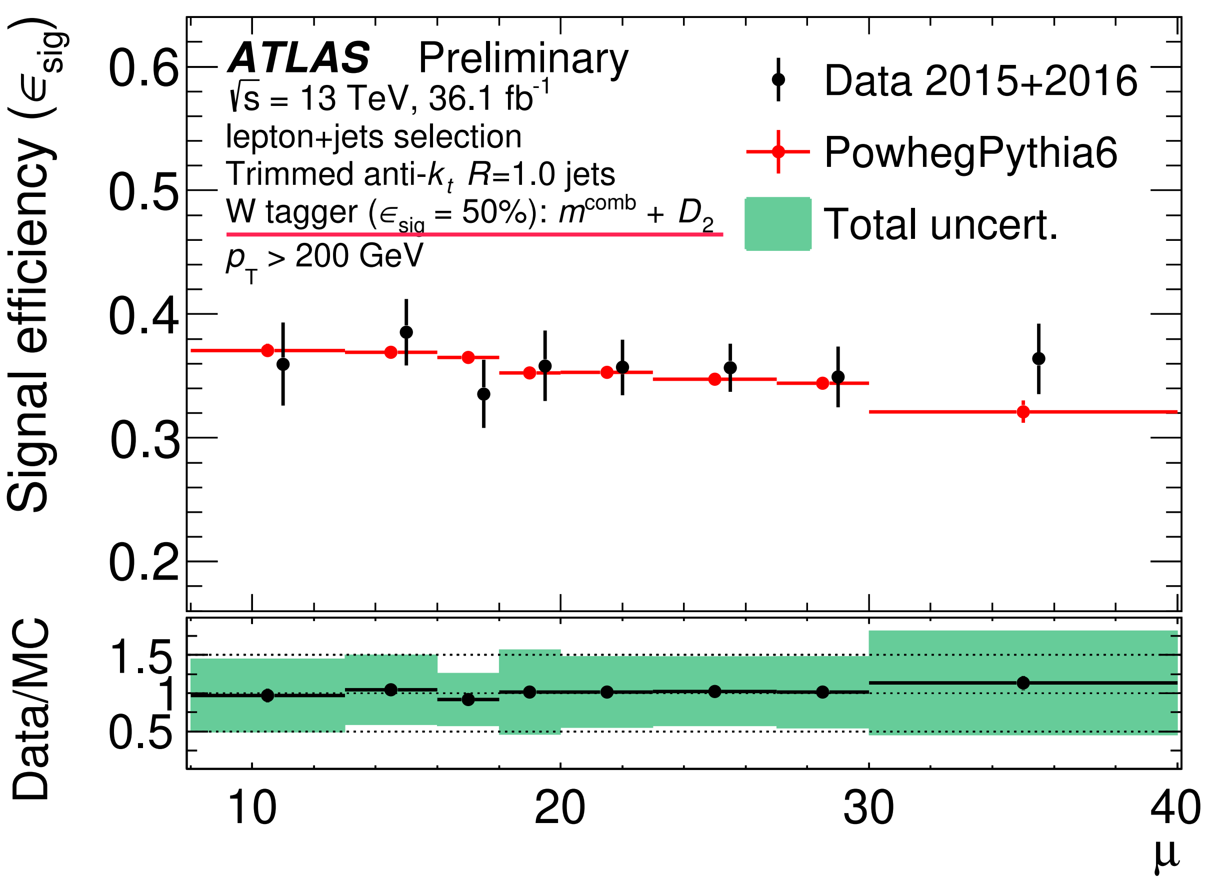
- **Signal efficiency vs. p_T** in MC and data for example taggers
- Tagging **scale-factor: $\varepsilon_{data}/\varepsilon_{MC}$,** with stat. and syst. uncertainties

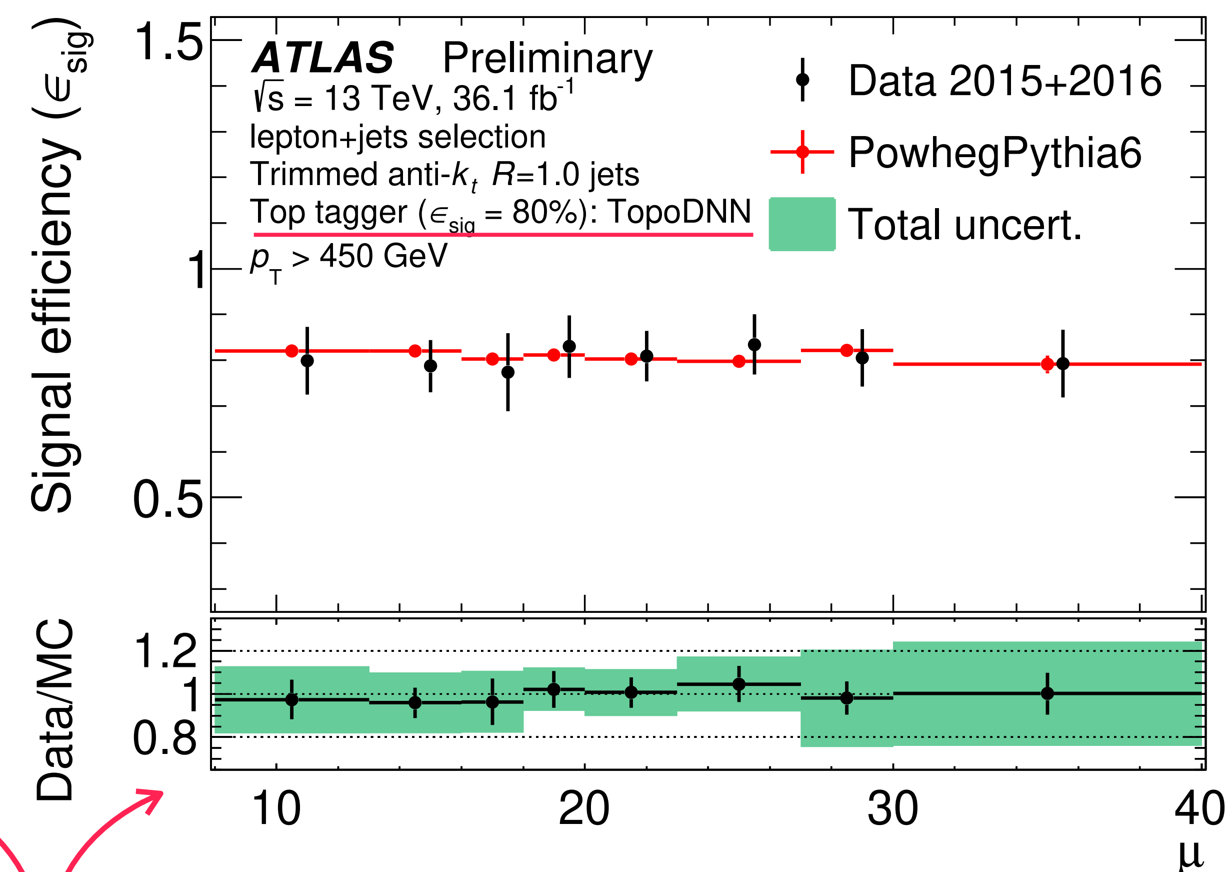




W/top tagging efficiency / Pile-up

- **Signal efficiency vs. mean number of pile-up interactions (*μ*)** in MC and data for example taggers
- Tagging efficiency **robust wrt. pile-up** within even stat. uncertainties

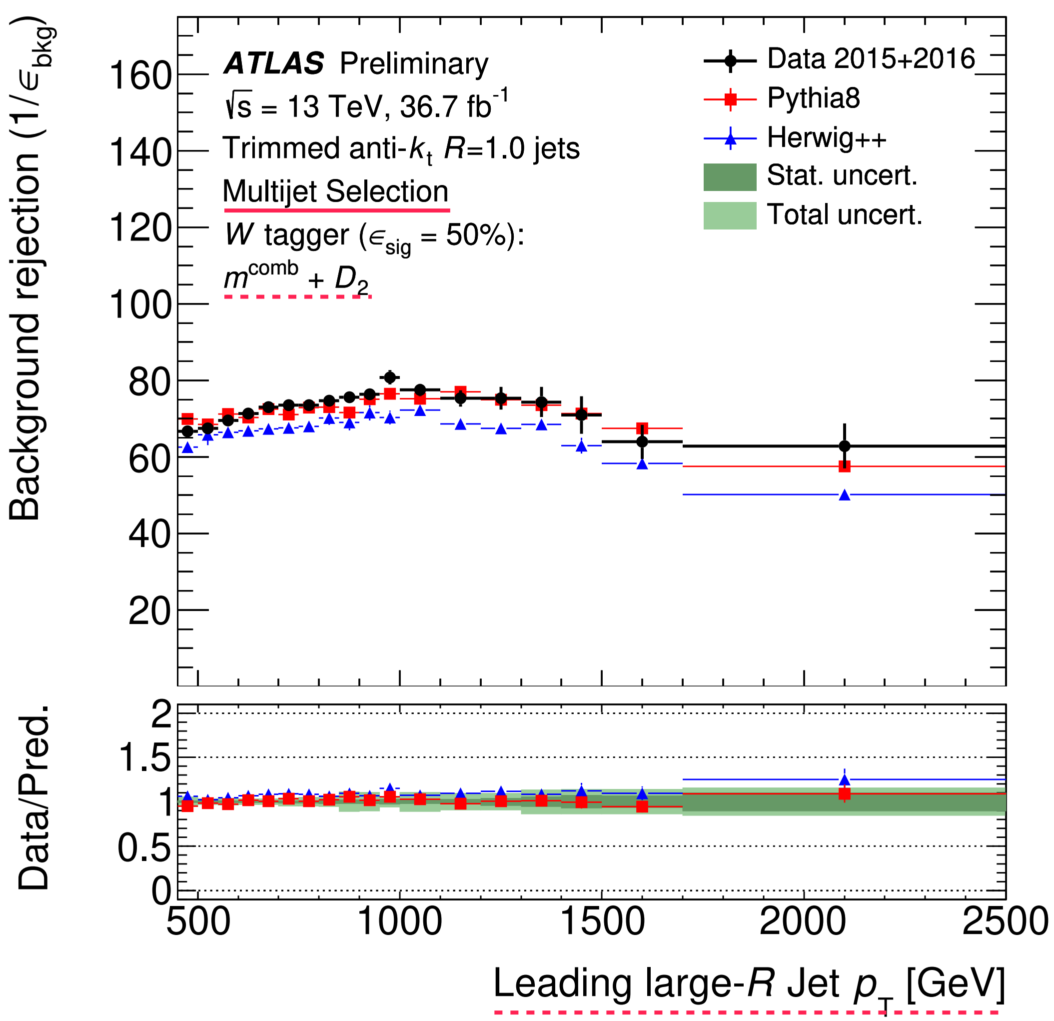


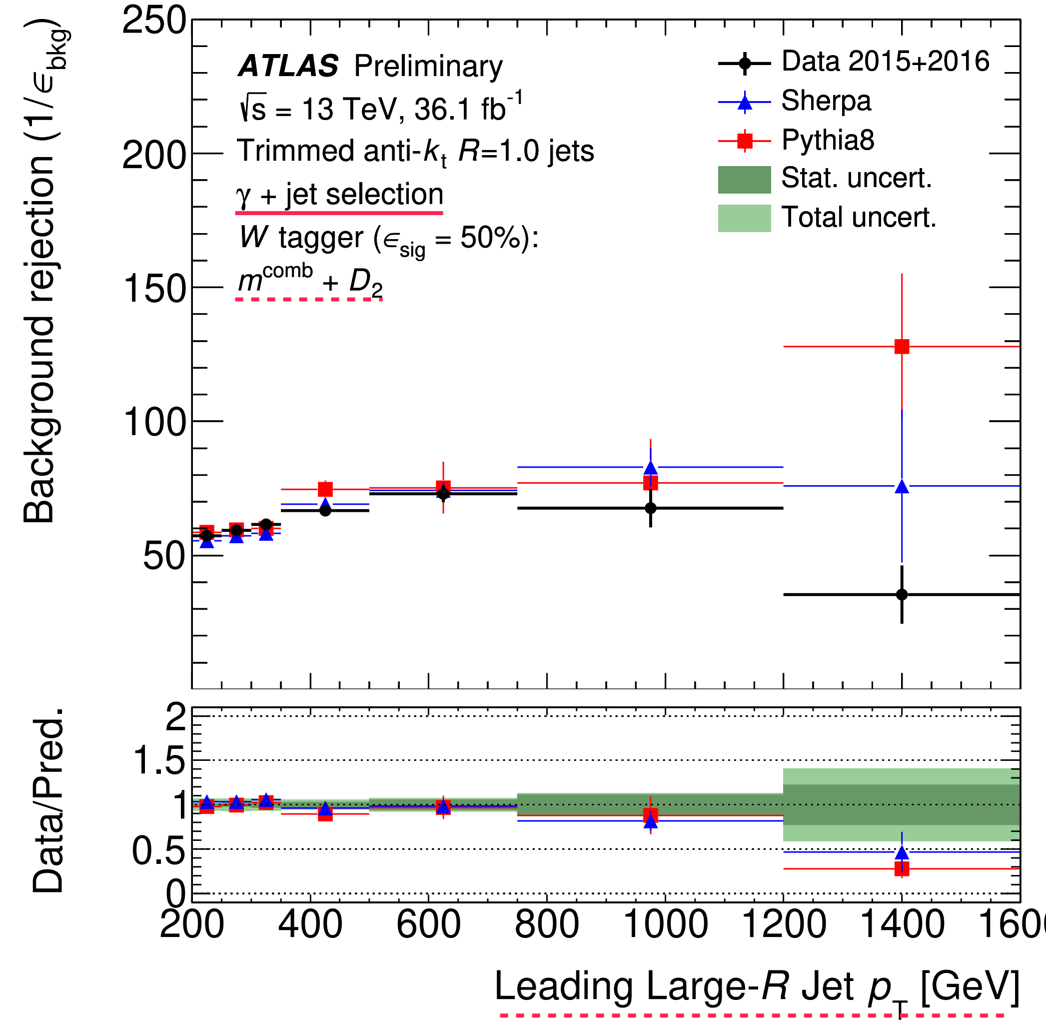


Different axis scales

Multijet rejection / *W*: $m^{\text{comb}} + D_2$

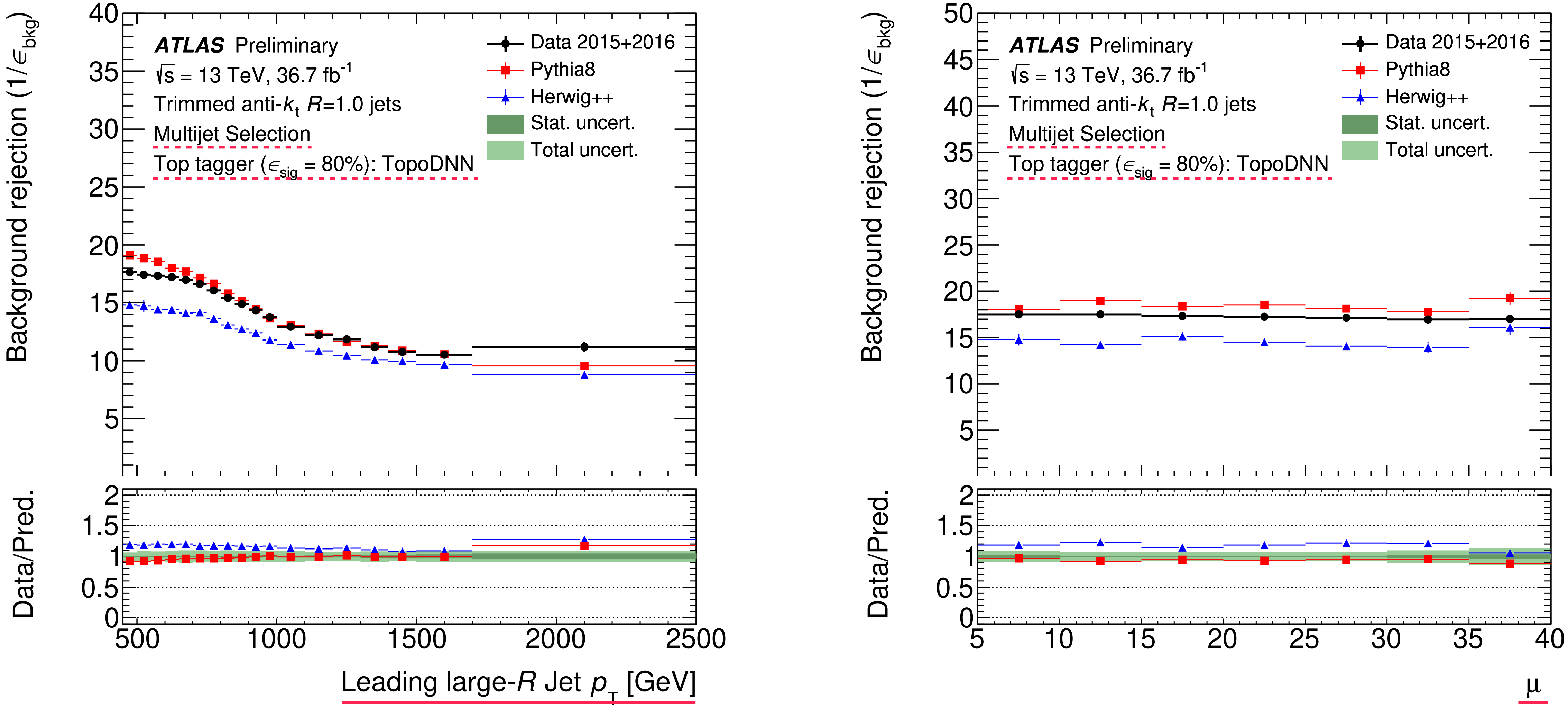
- Rejection $1/\varepsilon_{\text{bkg}}$ measured directly in signal-subtracted data:
  - **Dijet events:** Good data/Pythia8 agreement; Herwig++ lower
  - **γ + jets:** Both Sherpa, Pythia8 agree within uncertainties





Multijet rejection / Top: TopoDNN

- **Larger Herwig++ disagreement** for MVA than analytical taggers
- Small downwards trend vs. μ for TopoDNN, but no systematic trend across different tagging variables
- μ-dependence generally **well-modelled**



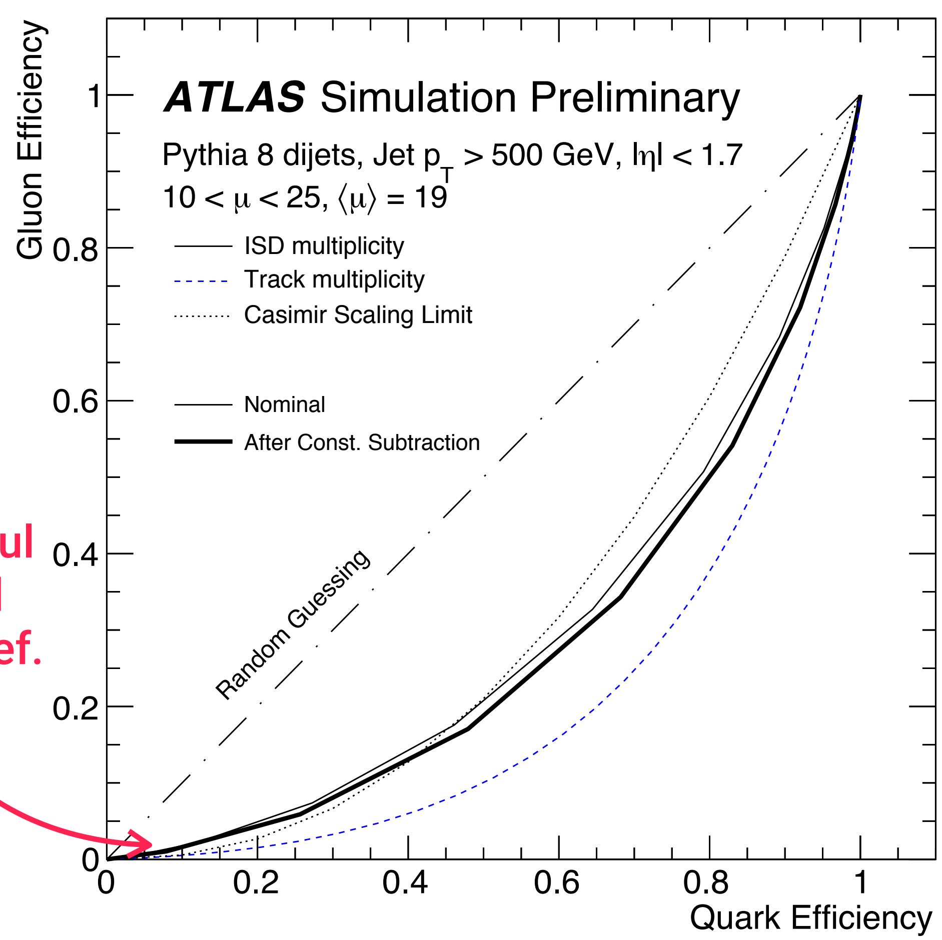
New ideas in jet tagging

ATL-PHYS-PUB-2018-011
EXOT-2017-14
ATL-PHYS-PUB-2018-014

Iterative Soft Drop (ISD) multiplicity

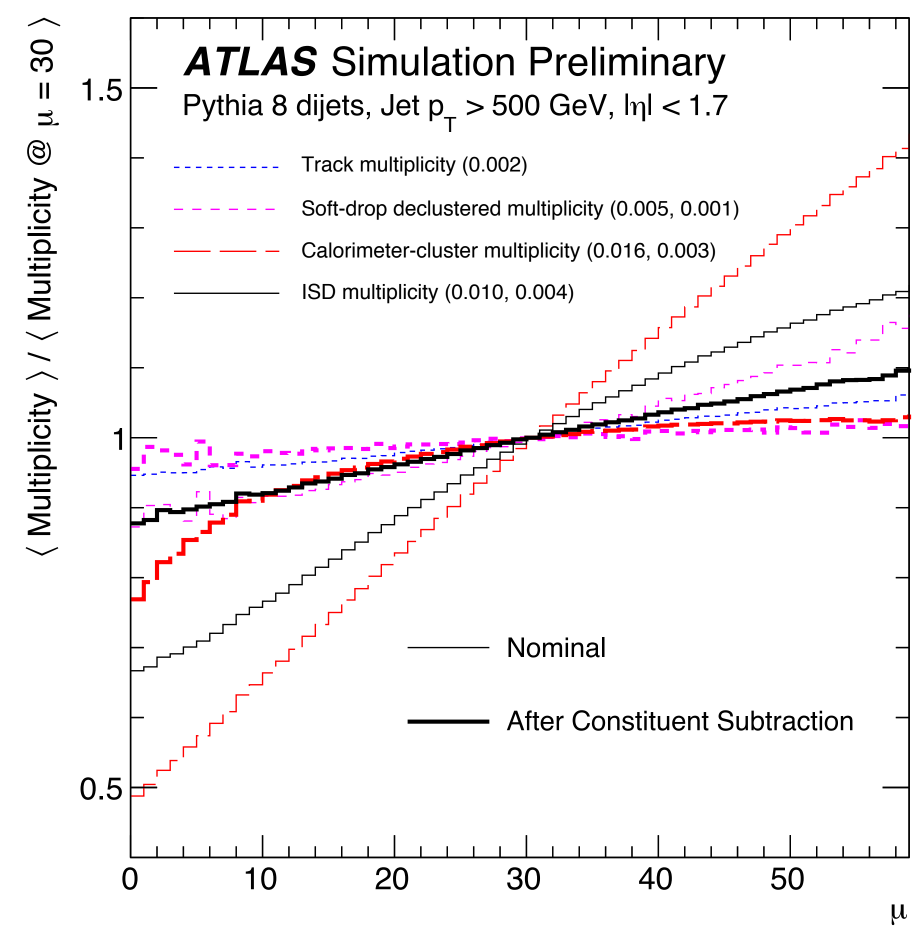
- **Constituent multiplicity**: single most important observable for *q*/*g* jet tagging, but **challenged by pile-up** contamination
- **ISD:** Undo clustering steps of C/A jet, counting hard splittings
- **Constituent subtraction:** Subtract ghost $p_T{}^g = A_g \times \rho$ from consts.

ATL-PHYS-PUB-2018-011



Most powerful calo.-based multiplicity def.

ATL-PHYS-PUB-2018-011



Stable wrt. μ with CS

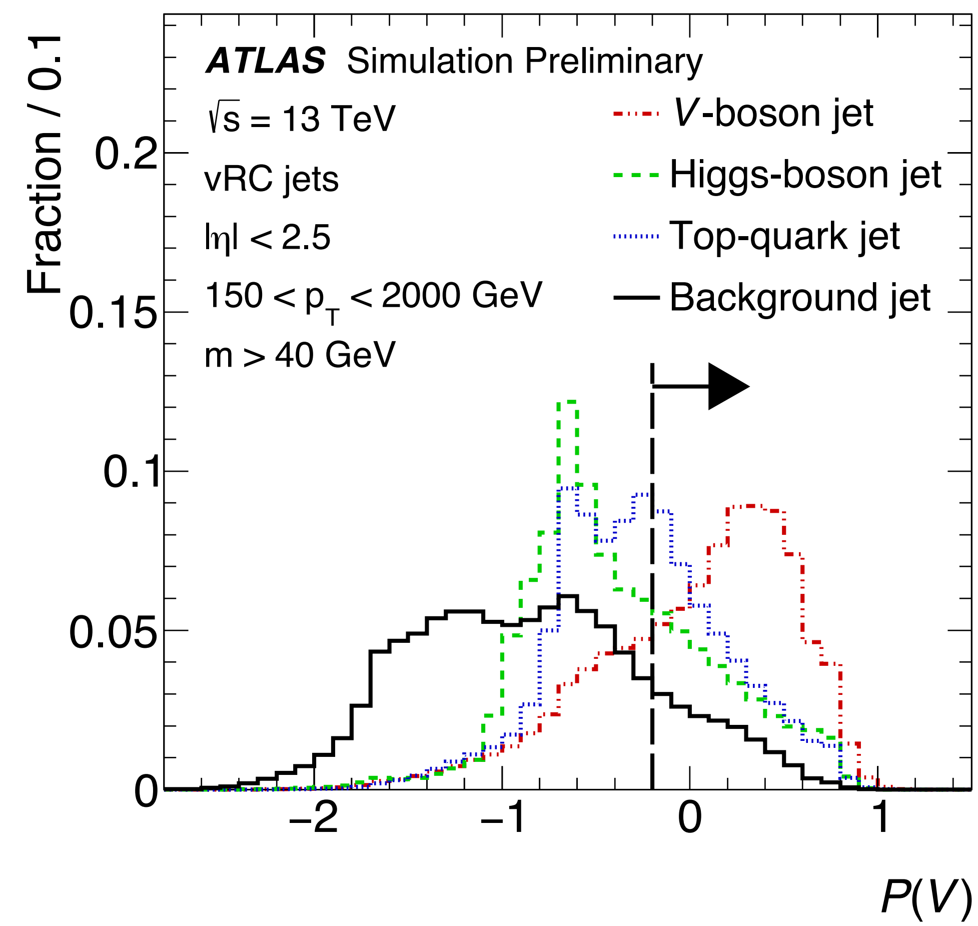
Multi-class jet classification

- Perform **multi-class classification** of *V**/*H*/top vs. multijets for **VLQ search** using variable-*R* reclustered (vRC) jets. See also: Talk by F. Guescini

- Train DNN with one classifier score (D_{DNN}) for each of the 4 jet classes, compute **per-class likelihoods**, e.g. *P*(*V*):

EXOT-2017-14

$$P(V) = \log_{10}\left(\frac{D^{V}_{\mathrm{DNN}}}{0.9\cdot D^{\mathbf{background}}_{\mathbf{DNN}} + 0.05\cdot D^{t}_{\mathrm{DNN}} + 0.05\cdot D^{H}_{\mathrm{DNN}}}\right)$$



- For multiply-tagged jets, **resolve ambiguity** by additional classifier variable ratios (*see backup*)

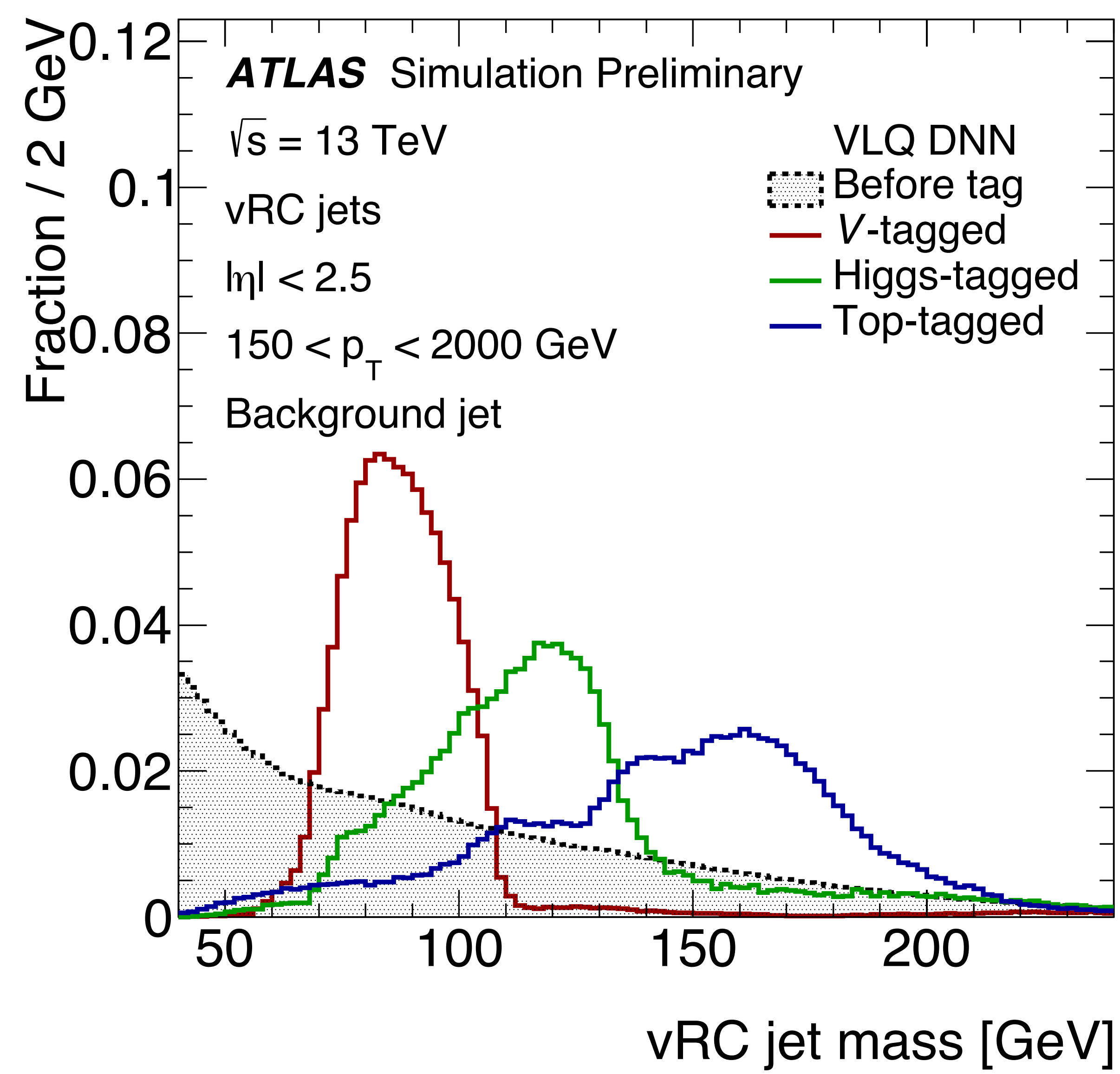
- Reclustered jets: **uncertainties on calibrated input** anti-k_t R = 0.4 jets can be used directly

* *V* = *W* or *Z*.

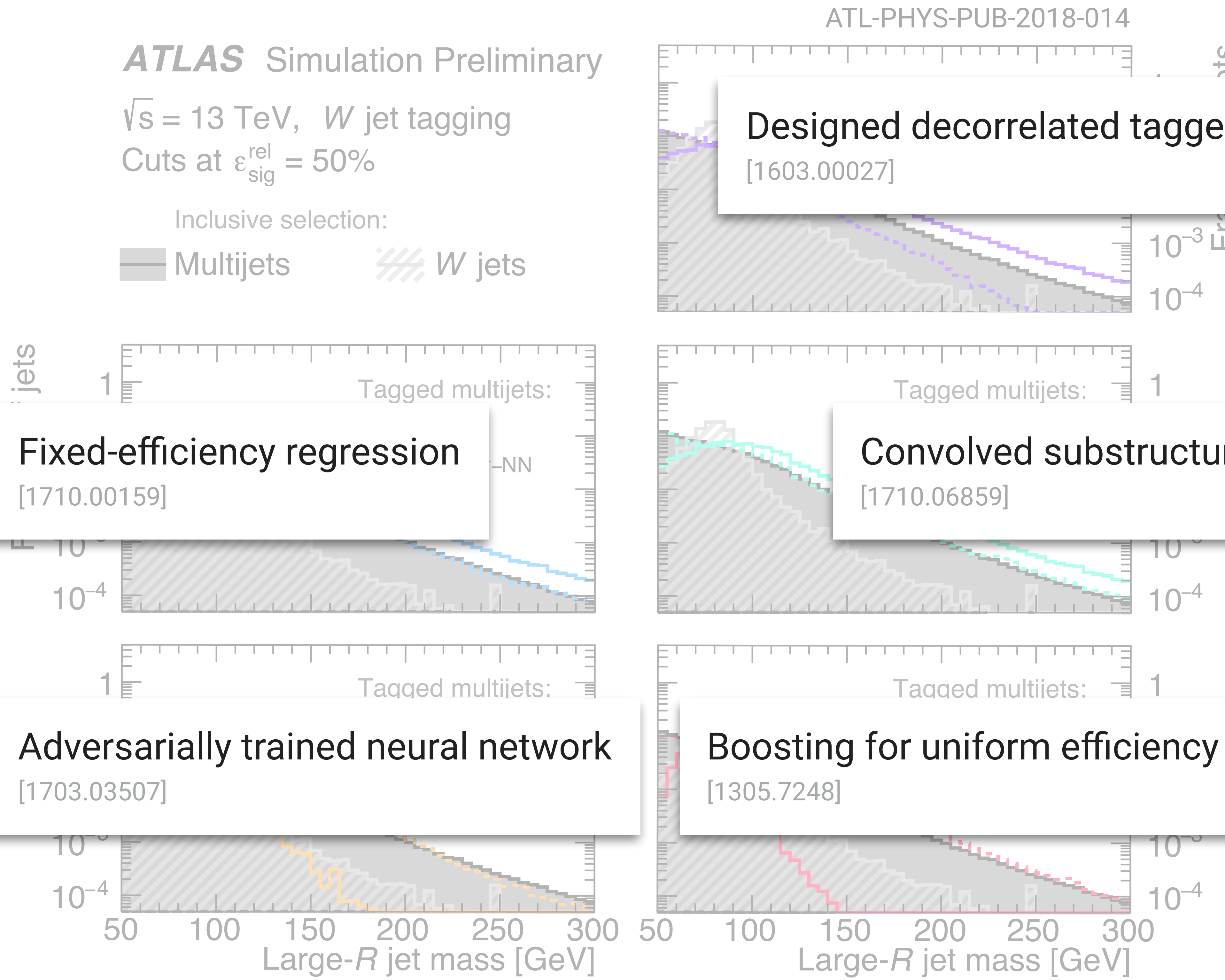
Mass-decorrelation / Overview

- Heavy resonance taggers prove powerful in data, MC
- But standard substructure taggers, particularly MVA, **sculpt the background jet mass distribution** to resemble signal peaks
- Problematic if search relies on jet mass spectrum
- Compare five methods for **mass-decorrelation**

EXOT-2017-14



Mass-decorrelation / Methods



Analytical / single-variable

DDT — Designed decorrelated taggers [1603.00027]

k-NN — Fixed-efficiency regression [1710.00159]

CSS — Convolved substructure [1710.06859]

Multivariate (MVA)

ANN — Adversarially trained neural network [1703.03507]

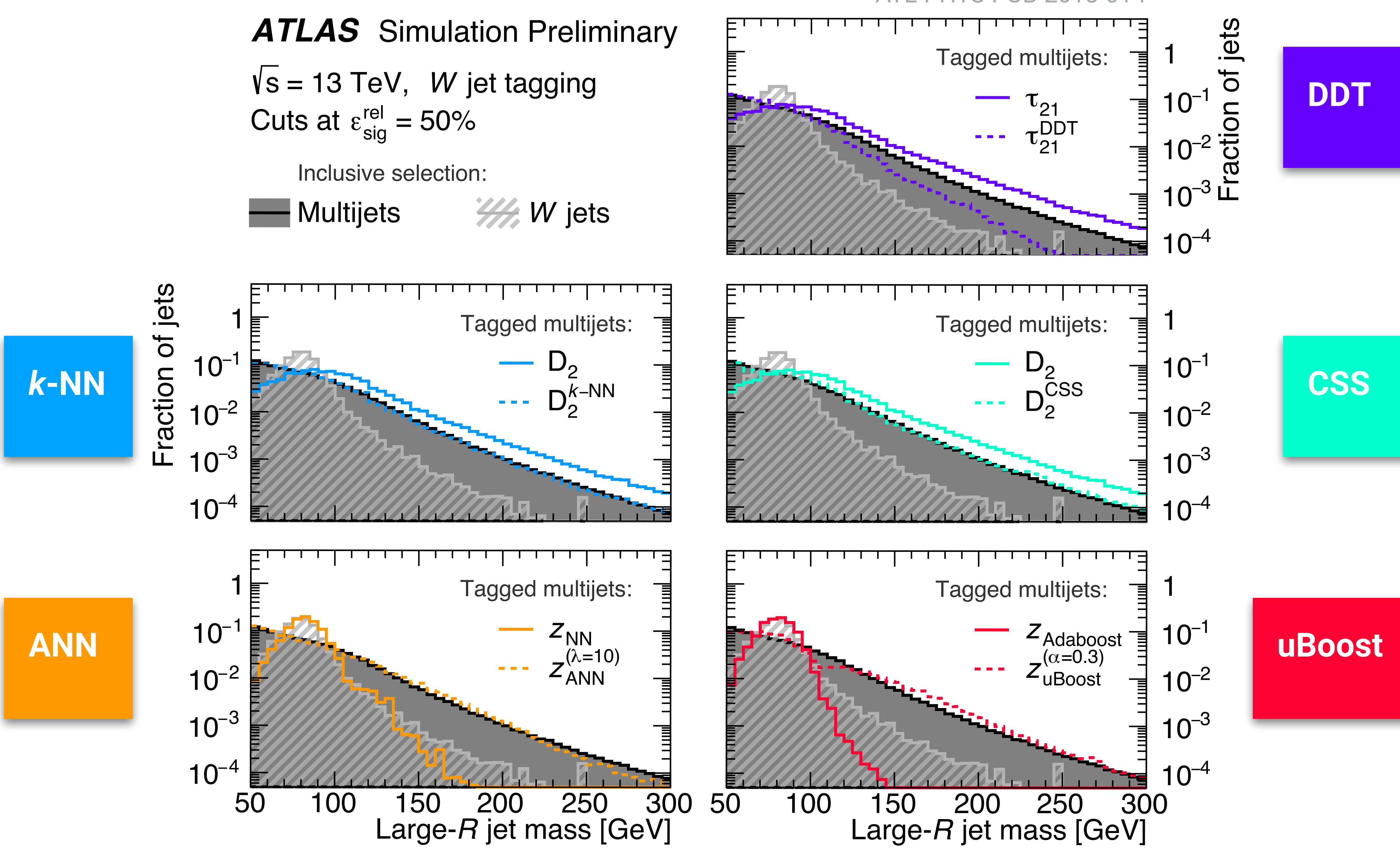
uBoost — Boosting for uniform efficiency [1305.7248]

Mass-decorrelation / Methods

Analytical / single-variable

Multivariate (MVA)

ATL-PHYS-PUB-2018-014



Mass-decorrelation / Results

- Simultaneously study

x-axis: **classification**

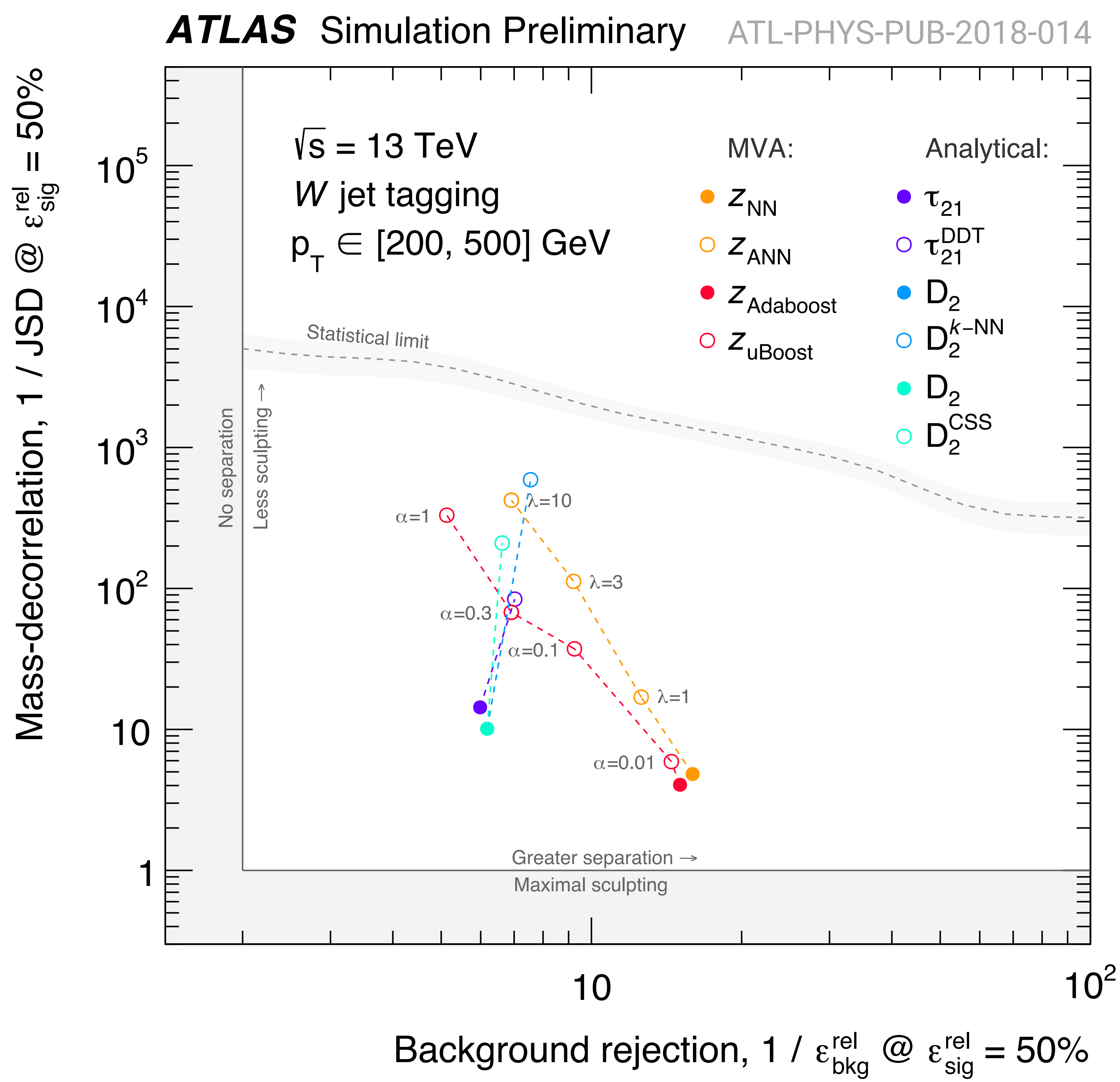
$1/\varepsilon_{bkg}$ @ ε_{sig} = 50%

y-axis: **mass-decorrelation**

1/JSD* @ ε_{sig} = 50%

(with no mass selection)

- ***k*-NN** leads to ~full mass-decorrelation
- **CSS** decorrelates more than **DDT**, due to no linearity requirement
- Mass-decorrelated **MVAs** are regularised, allowing for parametrised trade-off.



*JSD = Jensen-Shannon divergence; ~ relative entropy of pass/fail multijet mass distributions.

Mass-decorrelation / Results

- Simultaneously study

x-axis: **classification**

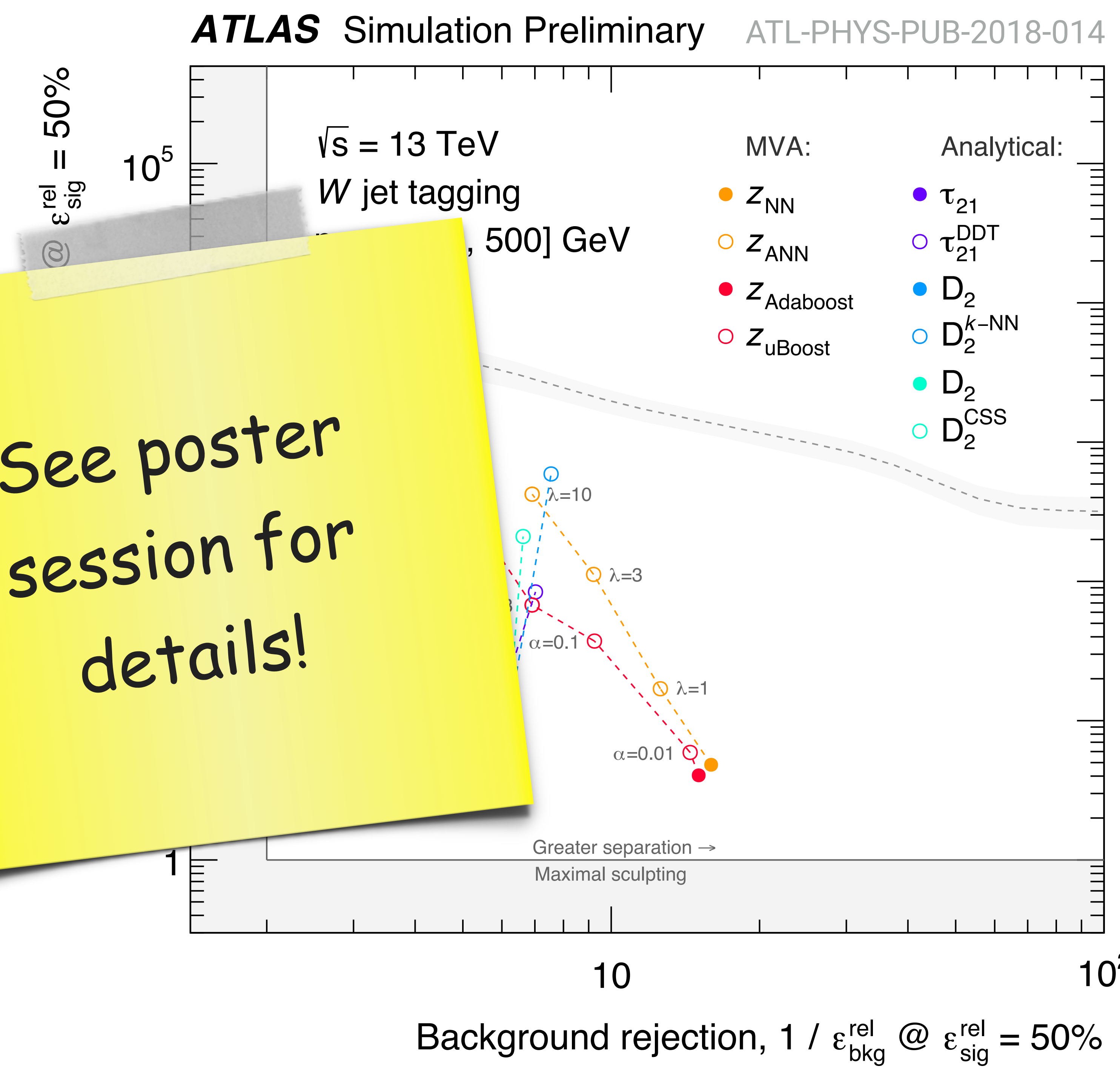
$1/\varepsilon_{bkg}$ @ $\varepsilon_{sig} = 50\%$

y-axis: **mass-decorrelation**

$1/\text{JSD}^*$ @ $\varepsilon_{sig} = 50\%$

(with no mass selection)

- ***k*-NN** leads to ~full mass-decorrelation
- **CSS** decorrelates more than **DDT**, due to no linearity requirement
- Mass-decorrelated **MVAs** are regularised, allowing for parametrised trade-off.



*JSD = Jensen-Shannon divergence; ~ relative entropy of pass/fail multijet mass distributions.

Conclusion

Conclusion

- Highlights of new results for boosted *W*/top tagging in ATLAS
- Novel high-p_T **topocluster-based top-tagging**, TopoDNN
- Evaluated taggers in data, **providing systematic uncertainties**
  - **W/top tagging** efficiencies and scale factors vs. p_T, μ
  - **Multijet rejection** vs. p_T, μ in two final states
- **New ideas**: ISD multiplicity and multi-class jet tagging
- Studied various approaches to **mass-decorrelated** jet tagging
  - Robust decorrelation with fixed-efficiency regression
  - Trade-off between classification and decorrelation with MVAs

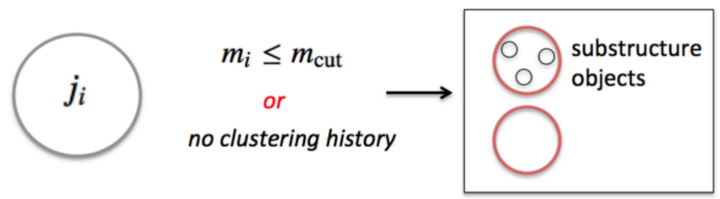
Thank you.

Backup

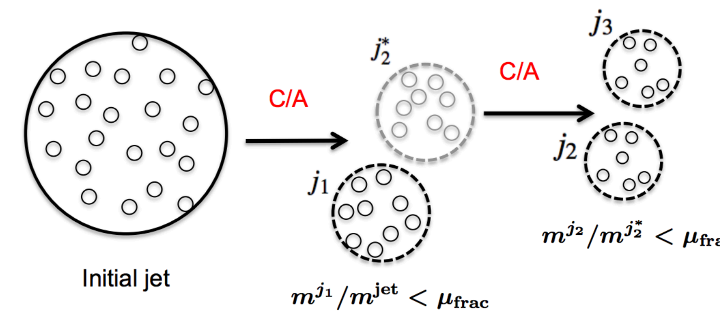
W/top tagging / HepTopTagger

Parameter	Value
m_{cut}	$50\ GeV$
$R_{\text{filt}}^{\text{max}}$	0.25
N_{filt}	5
f_W	15%

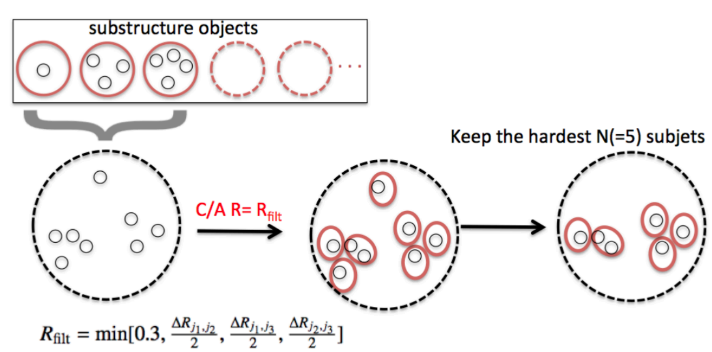
ATLAS-CONF-2017-064



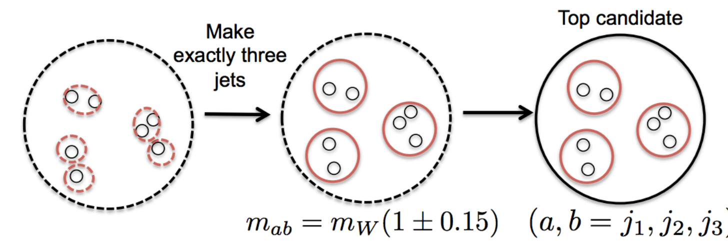
(a) Every object encountered in the declustering process is considered a 'substructure object' if it is of sufficiently low mass or has no clustering history.



(b) The mass-drop criterion is applied iteratively, following the highest subjet-mass line through the clustering history, resulting in N_i substructure objects.



(c) For every triplet-wise combination of the substructure objects found in (b), recluster the constituents into subjets and select the N_{subjet} leading-p_{T} subjets, with $3 \leq N_{\text{subjet}} \leq N_i$ (here, $N_{\text{subjet}} = 5$).

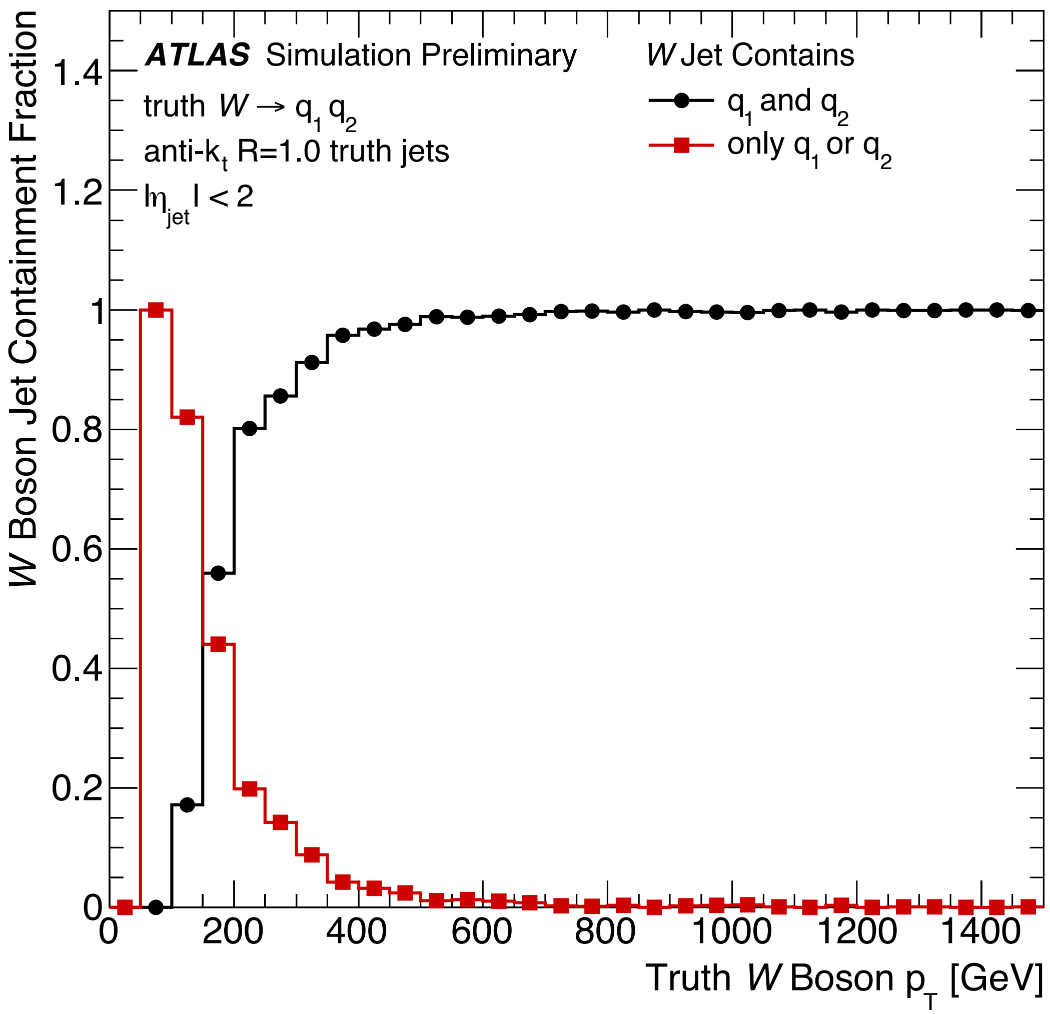


(d) Recluster the constituents of the N_{subjet} subjets into exactly three subjets to make the top candidate for this triplet-wise combination of substructure objects.

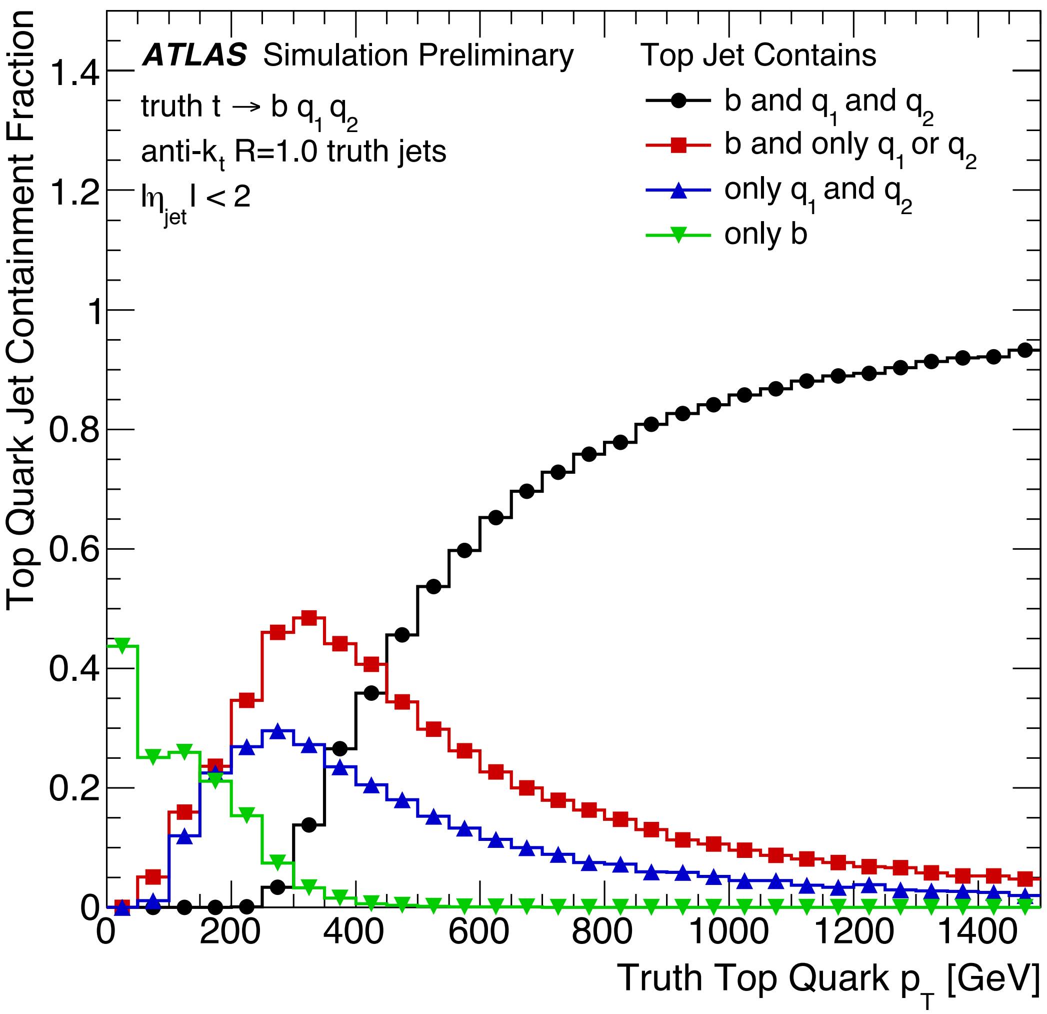
Figure from [1306.4945]

W/top tagging / Jet labelling

ATLAS-CONF-2017-064



ATLAS-CONF-2017-064



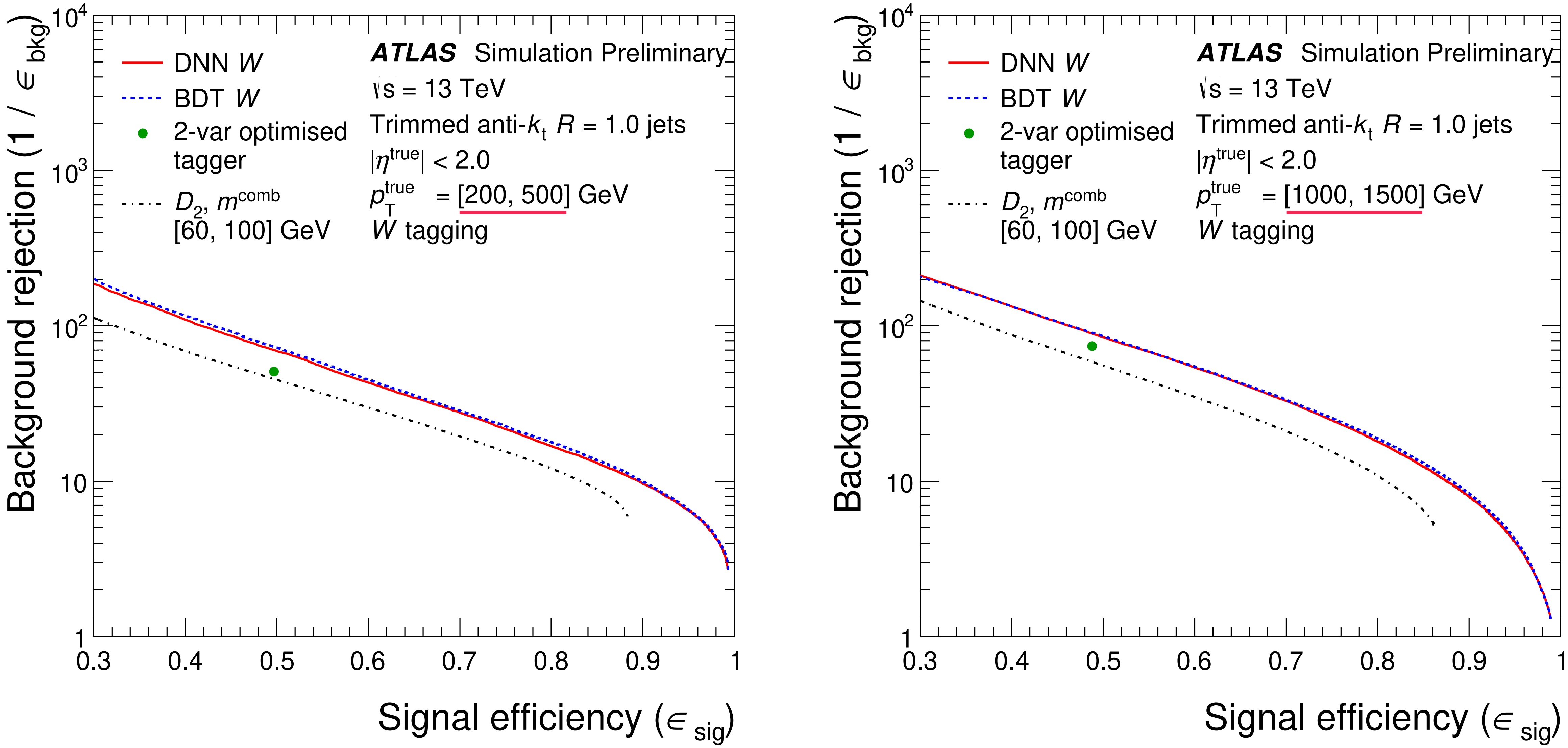
***W*/top tagging** / Chosen ML inputs

	W-Boson Tagging		Top-Quark Tagging	
Observable	BDT	DNN	BDT	DNN
m^{comb}	○	○	○	○
p_{T}	○	○	○	○
e_3			○	○
C_2		○		○
D_2	○	○		○
τ_1	○			○
τ_2			○	○
τ_3				○
τ_{21}	○	○	○	○
τ_{32}			○	○
R_2^{FW}	○	○		
$\mathcal{P}$	○	○		
a_3	○	○		
A	○	○		
Z_{CUT}		○		
$\sqrt{d_{12}}$	○	○	○	○
$\sqrt{d_{23}}$			○	○
$KtDR$	○	○		
Q_w			○	○

Table adapted from ATLAS-CONF-2017-064

Classification / *W*

- Two-variable optimised tagger (●) improves on D_2 + m-cut (⋮)
- MVA taggers (| ⋮) outperform cut-based taggers across p_T



ISD multiplicity

ATL-PHYS-PUB-2018-011

ISD Multiplicity – Number of Splittings while Grooming

1. Undo the last clustering step.
2. If $\Delta R_{ij} < \Delta R_{\mathrm{cut}}$, the algorithm terminates.
3. If $z_{ij} > z_{\mathrm{cut}}(\Delta R_{ij}/R)^{\beta}$, increment the ISD constituent multiplicity by one and iterate on the harder proto-jet. If not, iterate on the harder proto-jet without incrementing.

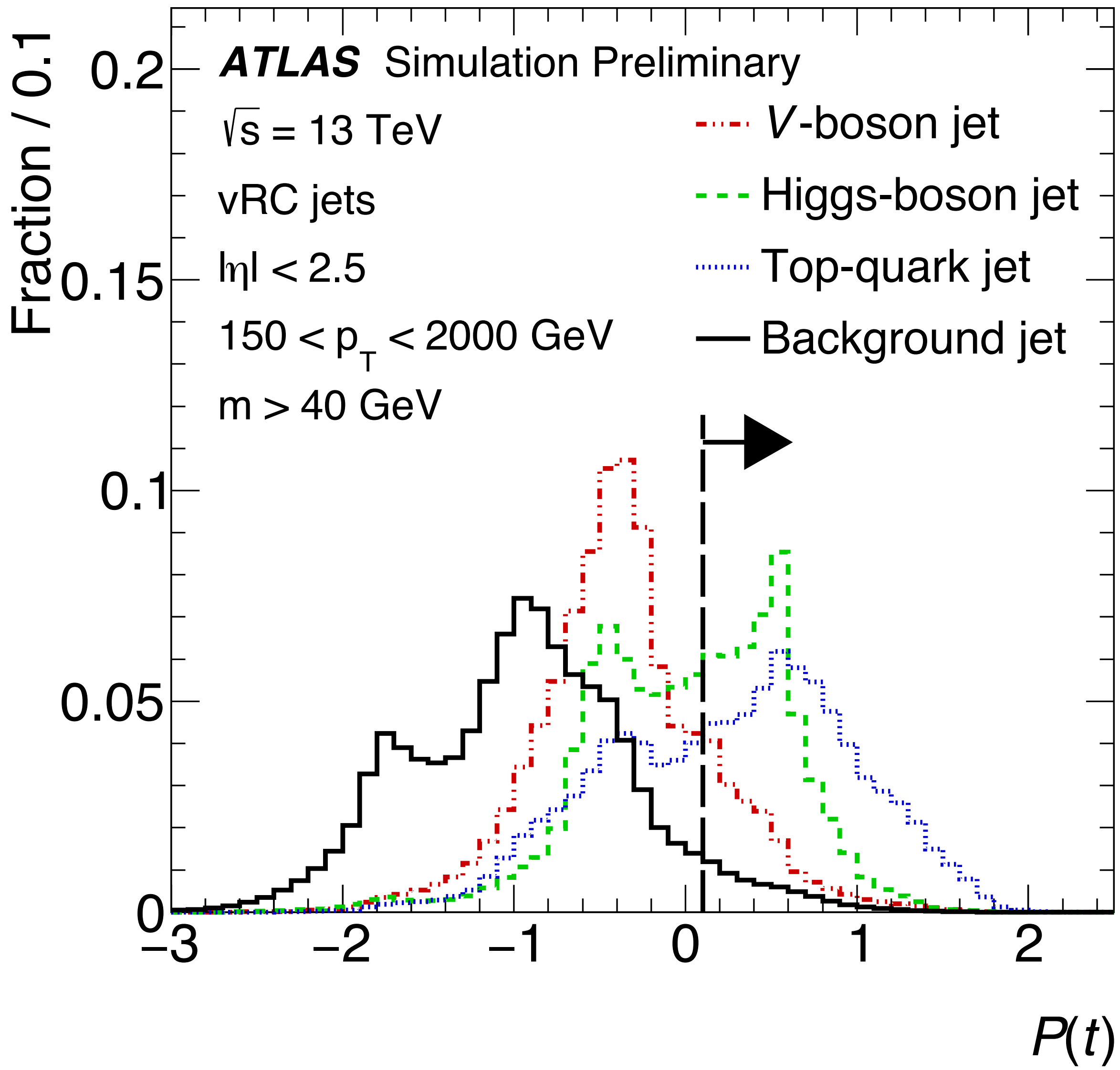
The parameters z_{cut}, β, and ΔR_{cut} are algorithm-specific. For the ISD Multiplicity, the parameters $z_{\mathrm{cut}} = 0.007$, $\beta = -1$, and $\Delta R_{\mathrm{cut}} = 0$ were chosen, which are the values found to maximize the quark-versus-gluon discrimination power while maintaining calculability from Ref. [20].

Multi-class jet tagging

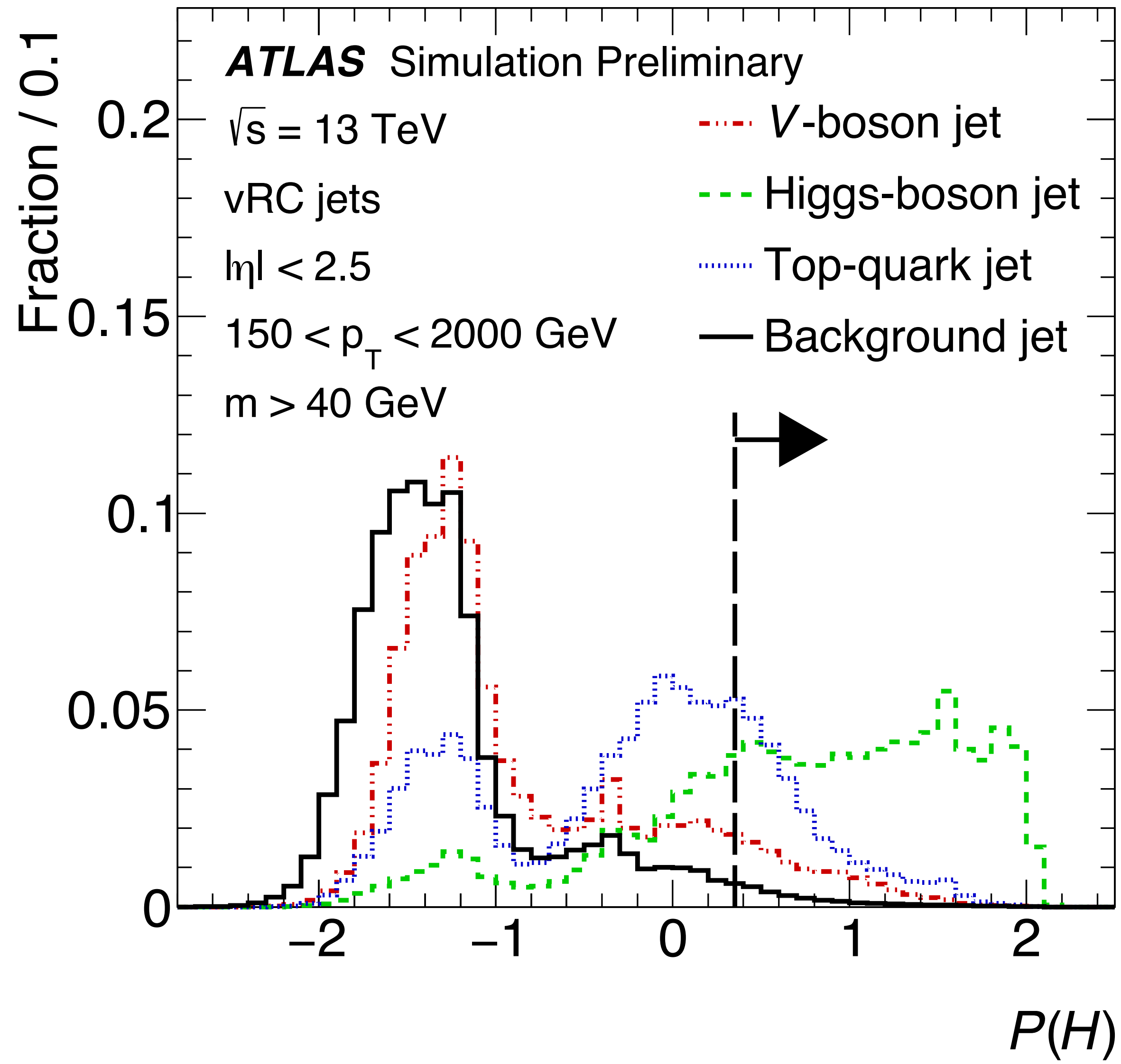
$$P(t) = \log_{10}\left(\frac{D^{t}_{\mathrm{DNN}}}{0.9\cdot \boldsymbol{D}^{\mathbf{background}}_{\mathbf{DNN}} + 0.05\cdot D^{H}_{\mathrm{DNN}} + 0.05\cdot D^{V}_{\mathrm{DNN}}}\right)$$

$$P(H) = \log_{10}\left(\frac{D^{H}_{\mathrm{DNN}}}{0.9\cdot \boldsymbol{D}^{\mathbf{background}}_{\mathbf{DNN}} + 0.05\cdot D^{V}_{\mathrm{DNN}} + 0.05\cdot D^{t}_{\mathrm{DNN}}}\right)$$

EXOT-2017-14

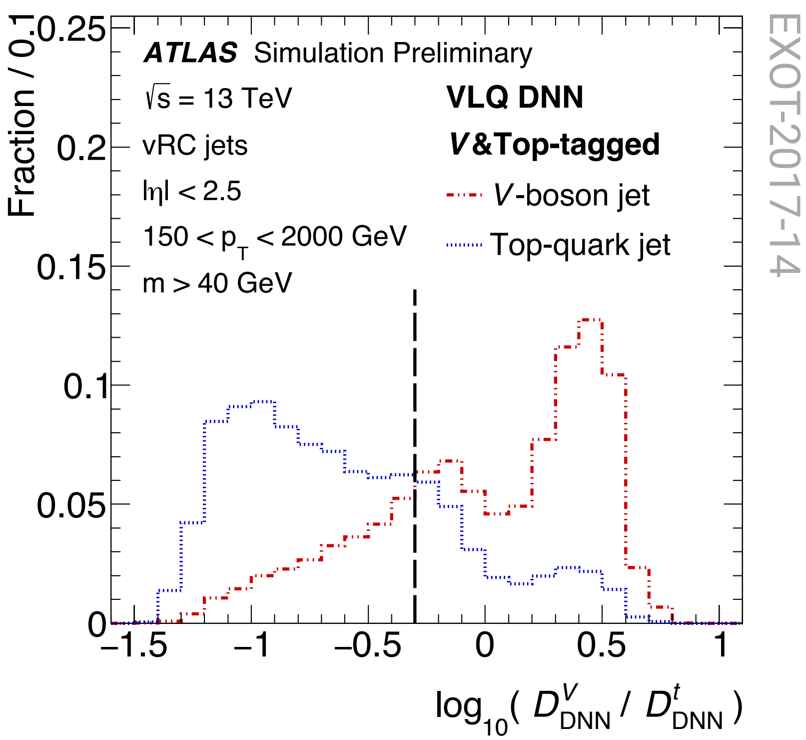


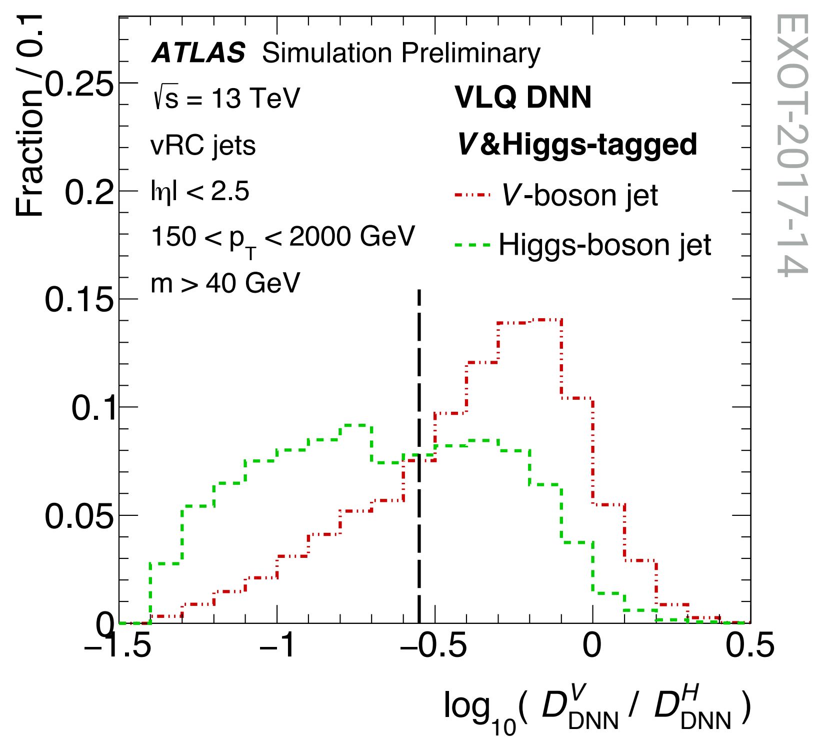
EXOT-2017-14

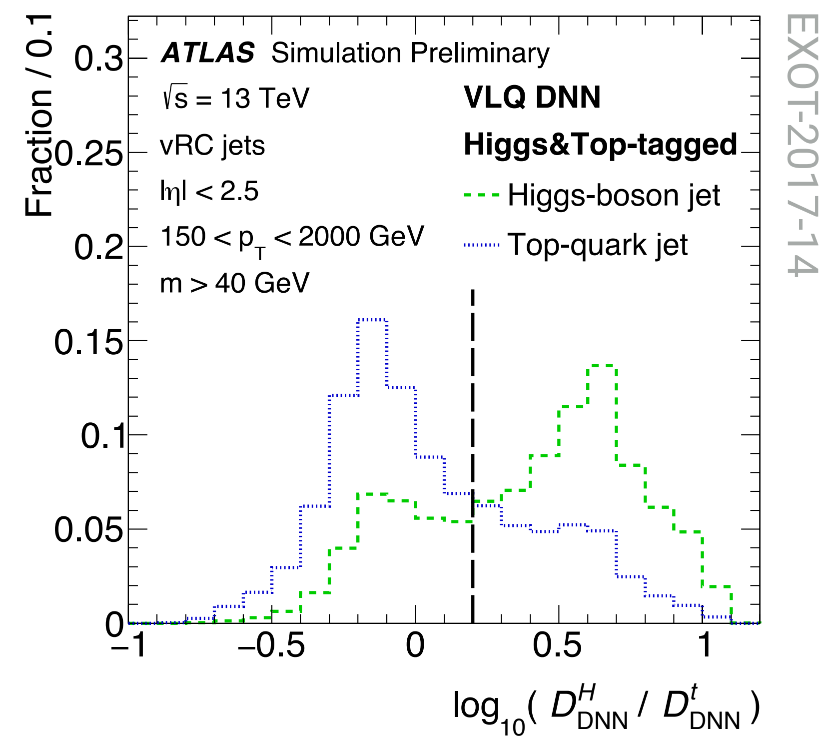


Multi-class jet tagging

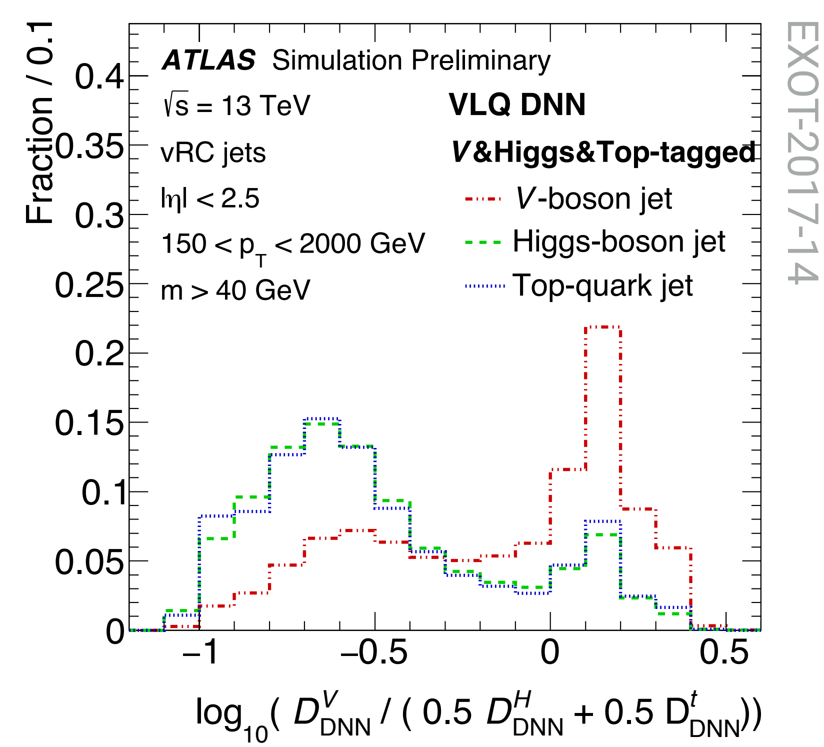
- **Double-tagged:** Likelihood ratios for each tagged class







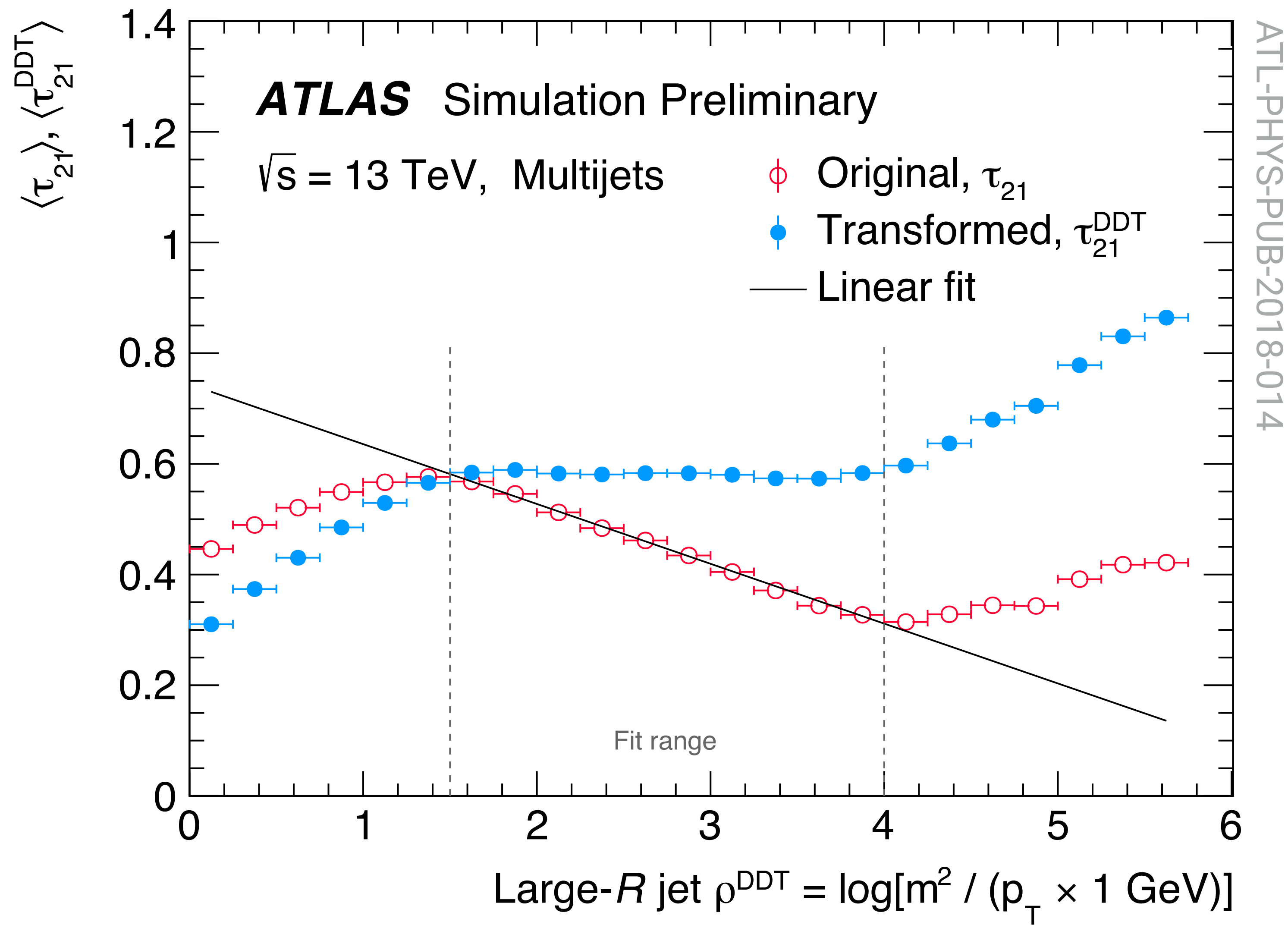
- **Triple-tagged:** Assume *H*, as this is most likely case



Mass-decorrelation / DDT

- Remove mean bias of τ_{21} by performing linear fit vs. ρ^{DDT} in restricted range for multijets. Choice of base substructure variable limited by linearity requirement.

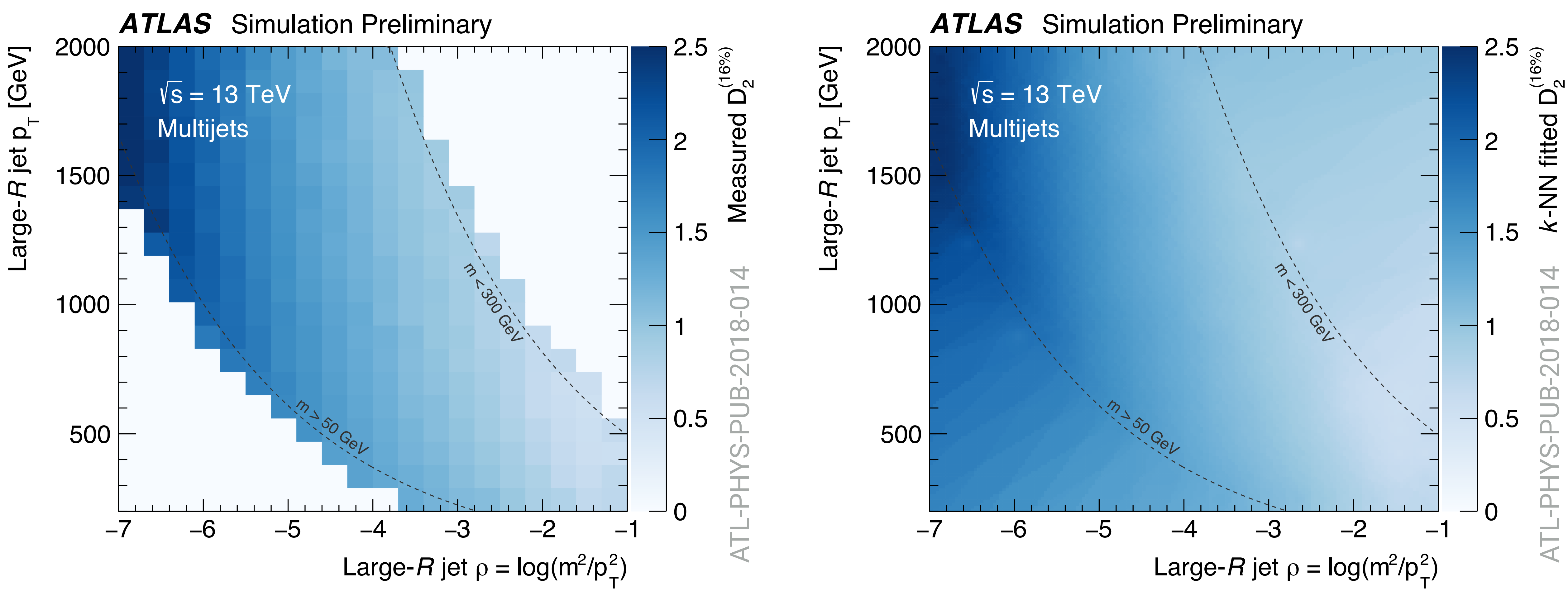
$$\tau_{21}^{\mathrm{DDT}} = \tau_{21} - a \times (\rho^{\mathrm{DDT}} - 1.5)$$



Mass-decorrelation / *k*-NN

- Use *k*-nearest neighbour regression to fit fixed-percentile value of D_2 for multijets as a function of ($\log p_T$, ρ) to remove non-linear correlations for fixed ε_{bkg}.

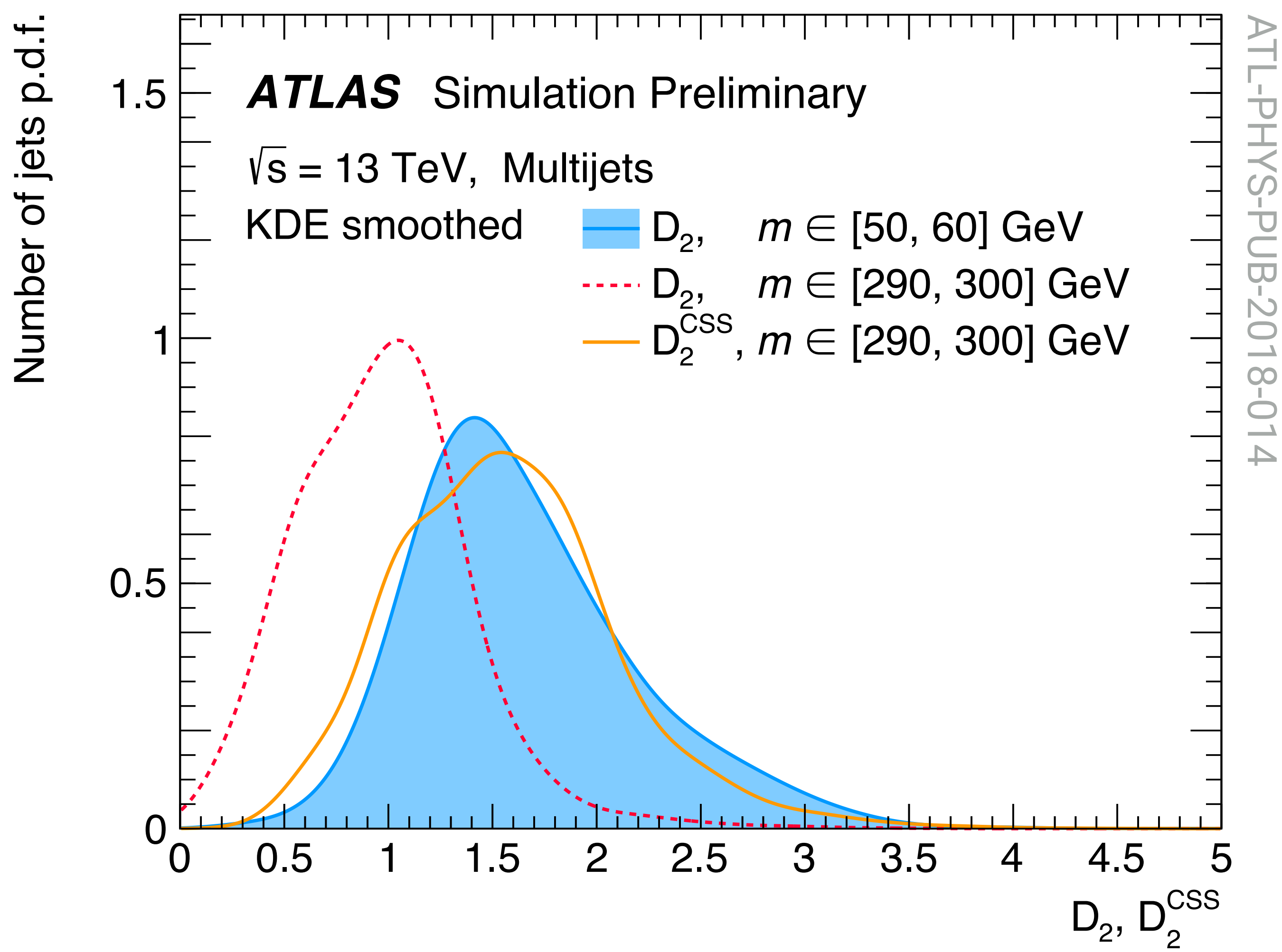
$$D_2^{k\text{-NN}} = D_2 - D_2^{(16\%)}$$



Mass-decorrelation / CSS

- D_2 distribution at jet mass m is convoluted to be coherently closer to distribution at reference mass m_0. Removes mass-dependence for higher-order moments than mean.

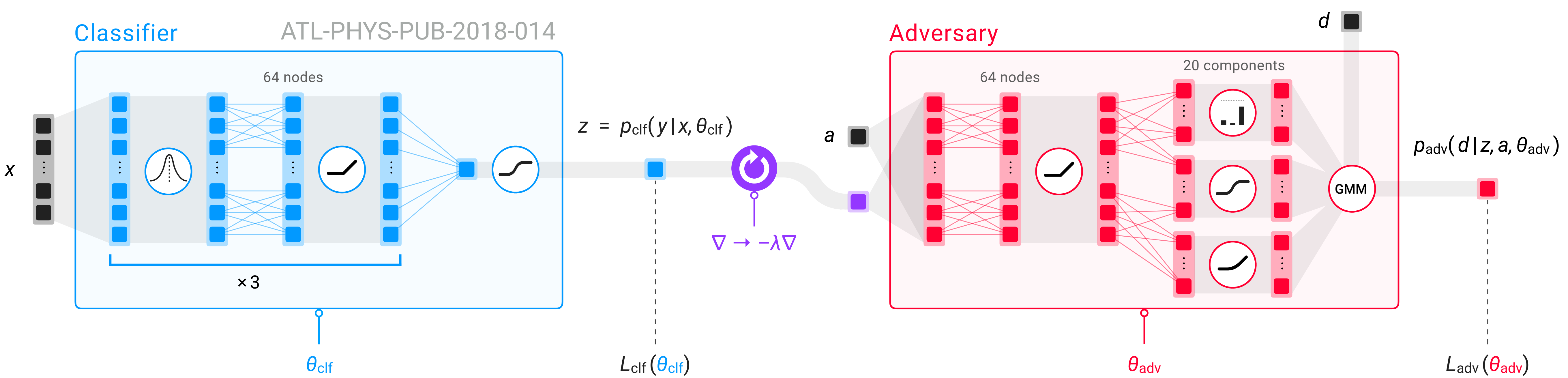
$$x_{\rm CSS} = G^{-1}(C(x)|\alpha,\Omega_D) \qquad F_{\rm CSS}(x|\alpha,\Omega_D) = \left(\frac{\alpha}{\Omega_D}\right)^{\alpha}\frac{1}{\Gamma(\alpha)}x^{\alpha-1}e^{-\frac{\alpha x}{\Omega_D}}.$$



Mass-decorrelation / ANN

- Pit neural network classifier against adversary network, tasked with inferring the jet mass from the classifier output and penalising the classifier if this is possible. Trade-off controlled by parameter λ.

$$\min_{\theta_{\mathrm{clf}}} \max_{\theta_{\mathrm{adv}}} L_{\mathrm{clf}}(\theta_{\mathrm{clf}}) - \lambda L_{\mathrm{adv}}(\theta_{\mathrm{clf}}, \theta_{\mathrm{adv}})$$



x: inputs, 10 substructure moments
y: target, jet label

a: aux. info., log p_T
d: decorrelation var(s)., jet *m*

Mass-decorrelation / uBoost

- Using adaptive boosting of decision tree to ensure uniform efficiency at fixed level $\bar{\varepsilon}$ across *m* in addition to standard classification boosting. Trade-off controlled by uniforming rate α.