Scale Factors for Boosted Object Identification in Diboson Events at ATLAS

Vinicius Oliveira
Supervised by: Chris Malena Delitzsch, Simone Ruscelli

September 29, 2025

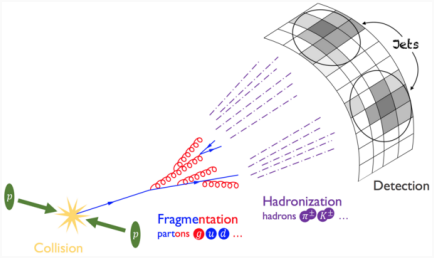
Thesis Presentation

technische universität dortmund
UNIVERSITÉ Clermont Auvergne
ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

Jets in ATLAS

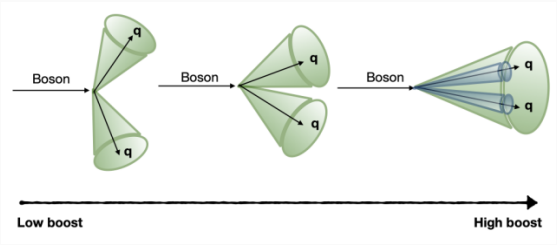
Quarks and Gluons in ATLAS:

Unlike e, μ, and γ, quarks and gluons cannot be observed as free particles in the detector. They are seen as a collimated flux of hadrons, called **jets**.



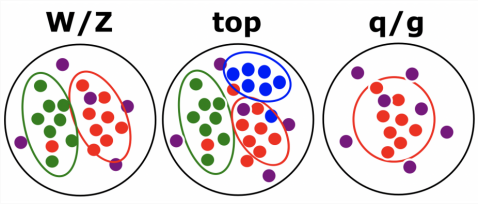
Jet Reconstruction:

- Algorithms are used to reconstruct jets from particle showers.
- Jet definitions:
  - Small-R jets: $R = 0.4$ (ATLAS/CMS)
  - Large-R jets: $R = 1.0$ (ATLAS)



Jet Tagging

It is important to identify whether jets originate from the hadronic decays of $W/Z/t$, or from q/g jets produced in pp collisions.



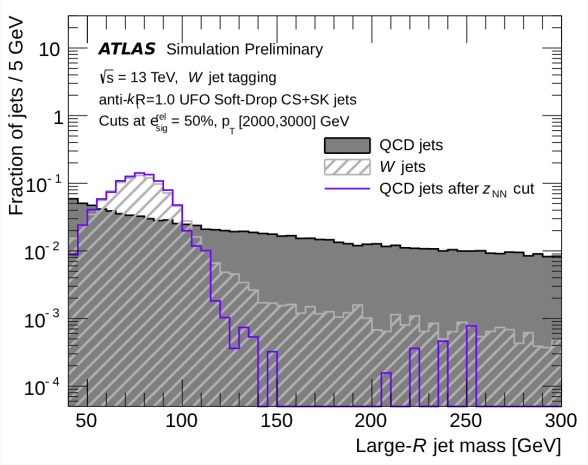
Tagging strategy:

- Jet internal structure variables can be exploited.
- Low-level taggers: based on simple kinematic cuts.
- High-level taggers: use multivariate or machine learning techniques.

Machine Learning Tagger

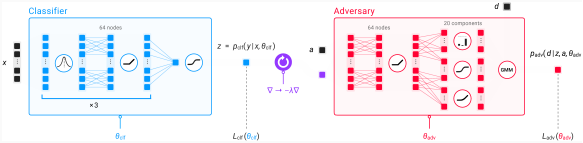
Deep Neural Network (DNN) Tagger

- Uses jet internal structure to distinguish W/Z jets from q/g jets.
- Trained on simulated data using jet substructure variables as inputs.
- When applied to background, it can sculpt the jet mass distribution into a signal-like shape.

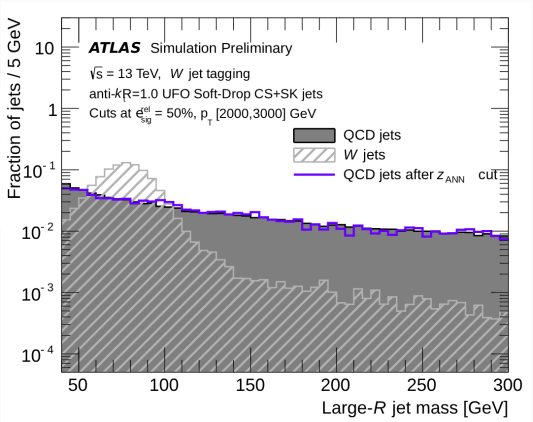


Mass Decorrelation in Jet Taggers

DNN with mass decorrelation $\Rightarrow$
Adversarial Neural Network (ANN)



$$\mathcal{L}_{\text{tagger}} = \mathcal{L}_{\text{classification}} - \lambda \mathcal{L}_{\text{adversary}}$$



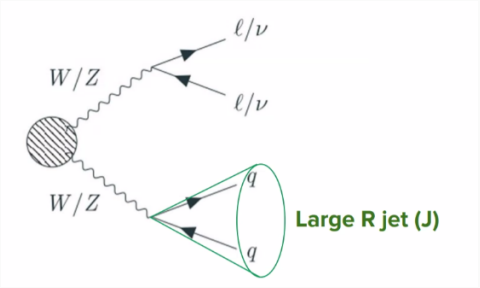
Jet Tagging Calibration

- **Scale Factors (SF)** are defined as the ratio between tagging efficiencies in data and Monte Carlo:

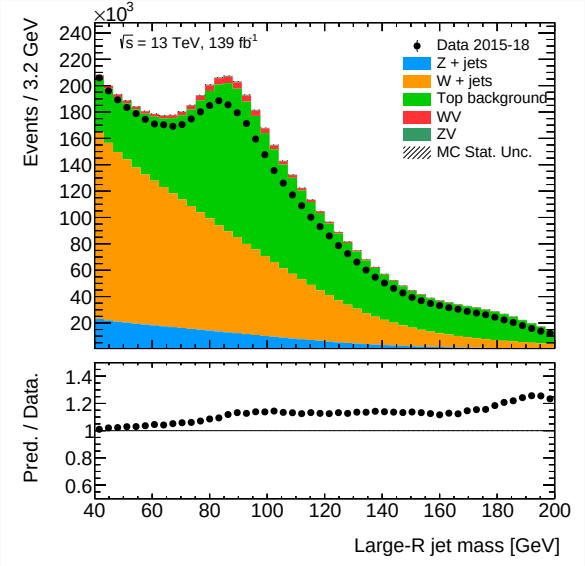
$$\mathrm{SF} = \frac{\epsilon_{\mathrm{data}}}{\epsilon_{\mathrm{MC}}}$$

- SFs have been studied in:
  - in $t\bar{t}$ events for $p_T < 350$ GeV
  - in V+jets events for $p_T > 600$ GeV

- This analysis aims to:
  - **Explore** semi-leptonic diboson events to calibrate the ANN tagger.
  - **Investigate** the intermediate-p_T region.



Full Jet Mass Distribution Data



Summary:

- Full ATLAS Run 2 dataset used ($\sqrt{s} = 13$ TeV).
- Data from 2015 to 2018.
- Leading large-R jet with $p_T > 200$ GeV and $m > 40$ GeV.

MC samples generated by:

- V+jets: Sherpa v2.2.11
- Top background: Powheg+Pythia8
- Signal: Sherpa v2.2.14

W/Z Selections

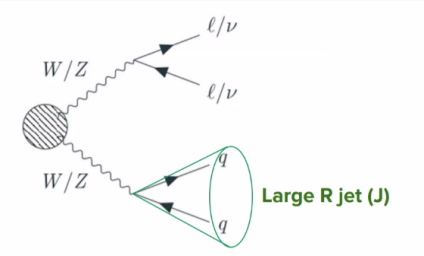
Selection Criteria for ZV and WV Channels

$Z(\ell\ell)V(qq)$ Selection Cuts

- Exactly 2 leptons (e^+e^- or $\mu^+\mu^-$)
- ANN 50% WP tagging

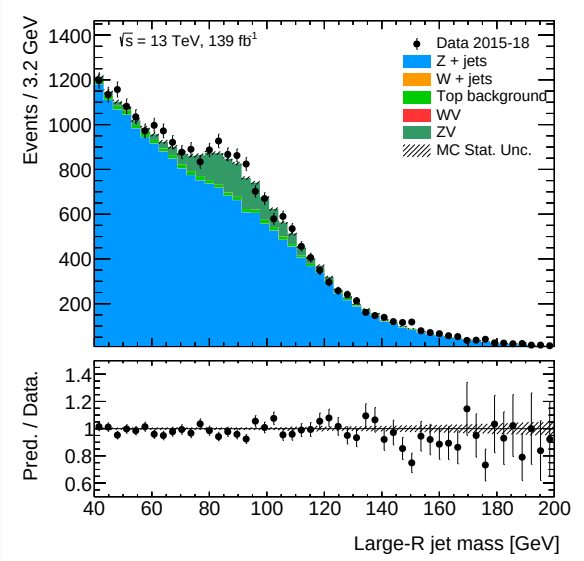
$W(\ell\nu)V(qq)$ Selection Cuts

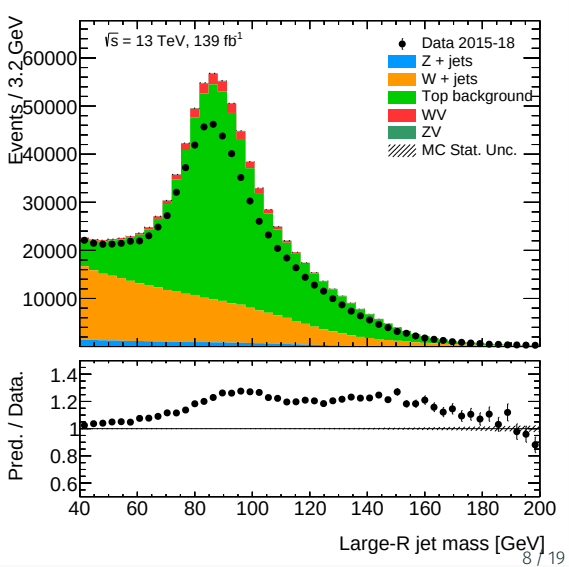
- Exactly 1 lepton (e or μ)
- ANN 50% WP tagging



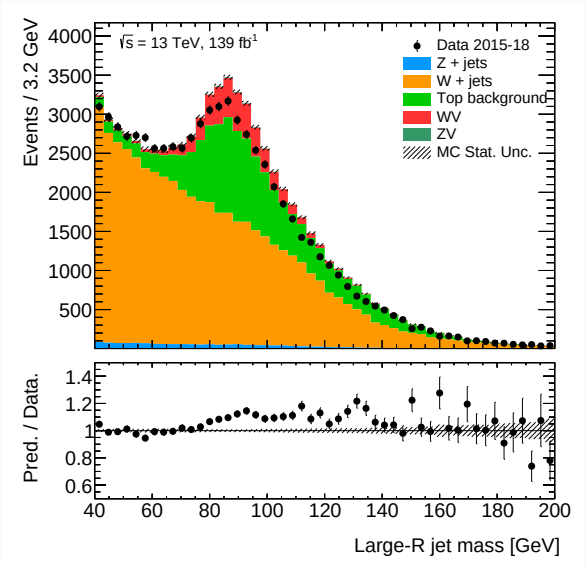
Note: Simone Rusceli's analysis focused on the $Z(\mu\mu)V(qq)$ channel with exactly 2 muons and an ANN 80% WP tagger.

Data vs MC Comparison in ZV and WV Selections



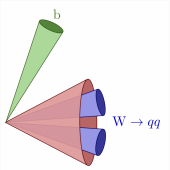


$W(\ell\nu)V(qq)$ Selection after ΔR and E_T^{miss} Cuts



Selection:

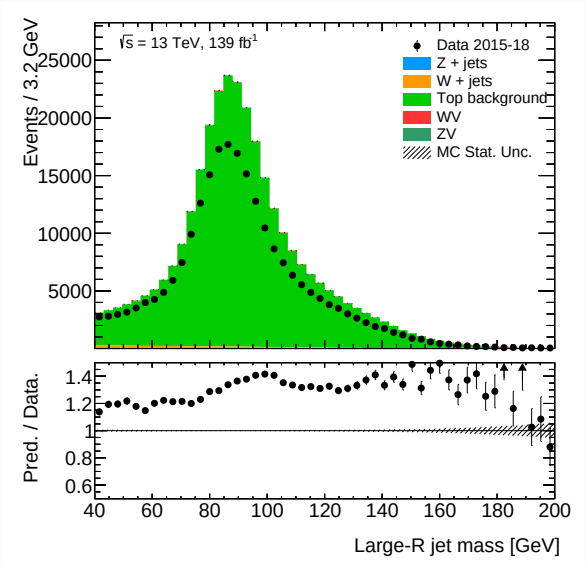
- Exactly 1 lepton (e or μ).
- ANN 50% WP tagging.
- No Pair $\Delta R(j, J) > 1.0$
- $E_T^{\text{miss}} > 50$ GeV



$$\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$$

Control Regions

Control Region (Top Background)

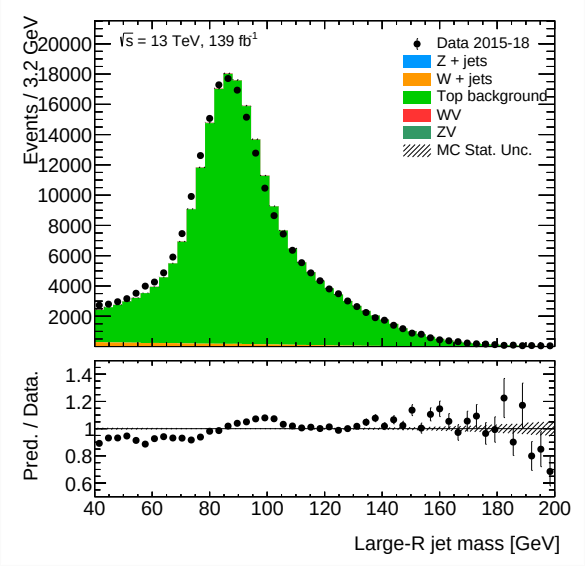


Selection:

- Exactly 1 lepton (e or μ).
- ANN 50% WP tagging.
- At least one pair with $\Delta R(j, J) < 1.0$.
- At least one b-tagged small-R jet.
- At least one small-R jet with $p_T > 25$ GeV.

Note: 98% purity in this selection.

Scale Factor (Top CR)

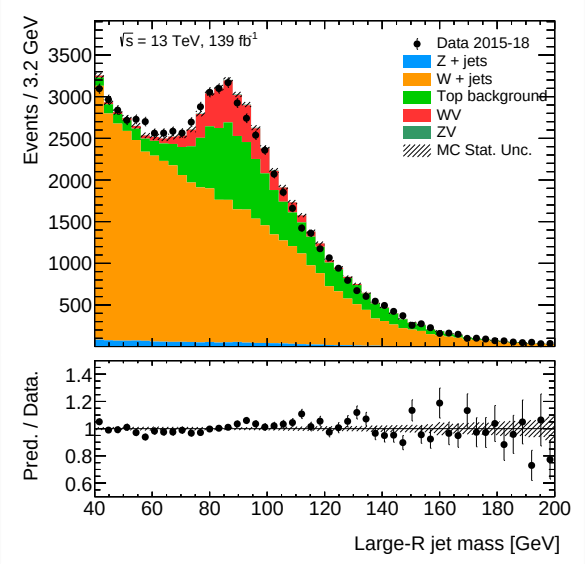


$$\mathrm{SF}_{\mathrm{top}} = \frac{\mathrm{Data} - \mathrm{MC}_{\mathrm{no\ top}}}{\mathrm{MC}_{\mathrm{top}}}$$

Result

$$\mathrm{SF}_{\mathrm{top}} = 0.76$$

Scaled Jet Mass Distribution

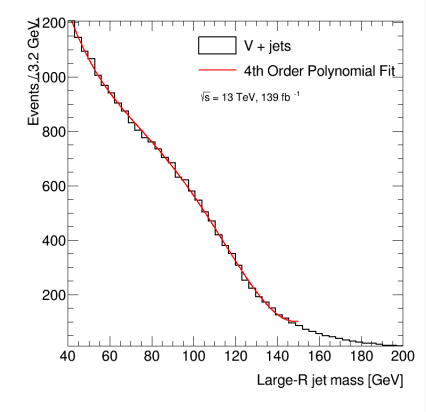


- The $\mathrm{SF}_{\mathrm{Top}}$ and $\mathrm{SF}_{\mathrm{W+jets}}$ are applied to the background.
- Better agreement in the signal region.

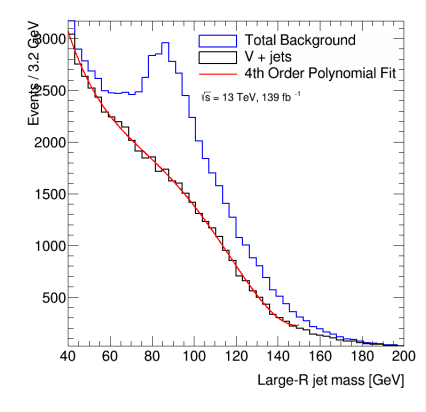
Background Estimation

V + jets Background Fit

$Z(\ell\ell)V(qq)$ Selection



$W(\ell\nu)V(qq)$ Selection

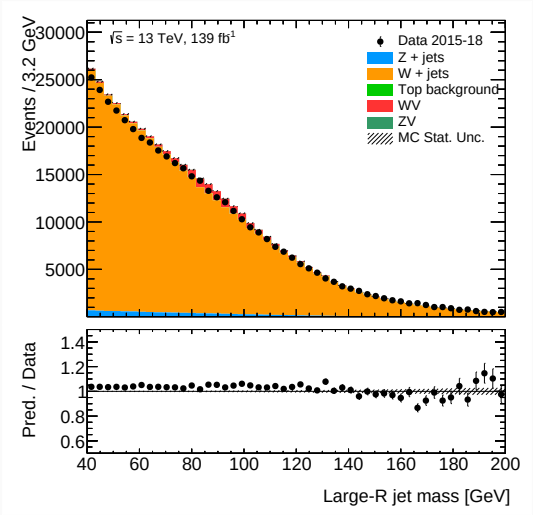


Background: modeled with a 4th-order polynomial

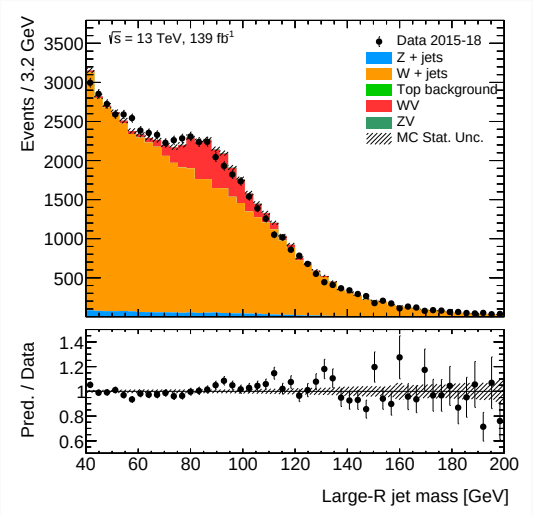
$$P(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + a_4 x^4$$

Top Background Subtraction in $W(\ell\nu)V(qq)$ selection

Without 50% WP ANN Tagging

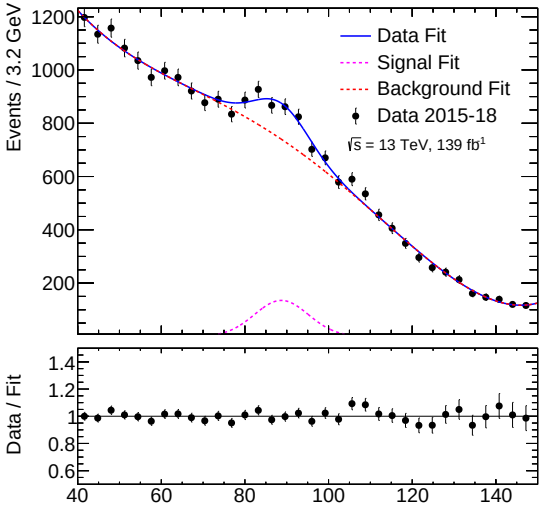


With 50% WP ANN Tagging



Signal Extraction

Fit on $Z(\ell\ell)V(qq)$ Selection Data



Fit Description

- **Background:** modeled with a 4th-order polynomial:

$$P(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + a_4 x^4$$

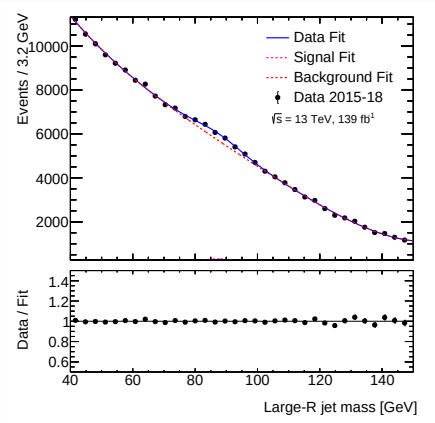
- **Signal:** modeled as a single Gaussian.
- Background parameters are fitted on MC simulation before the full data fit.
- Signal yield obtained as:

$$N_{\text{signal}} = A \cdot \sigma \cdot \sqrt{2\pi}$$

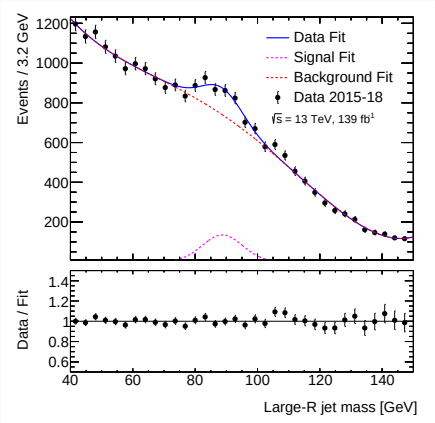
Tagging Efficiency in Data

Tagging Efficiency Extraction in $Z(\ell\ell)V(qq)$ Selection

Without 50% WP ANN Tagging



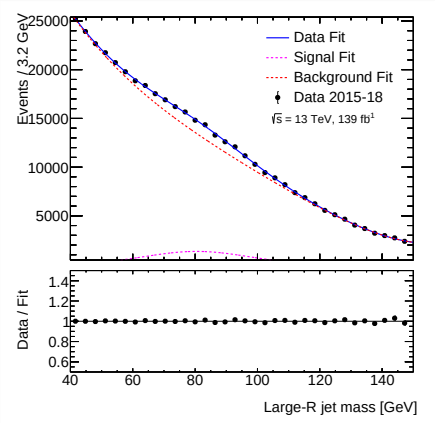
With 50% WP ANN Tagging



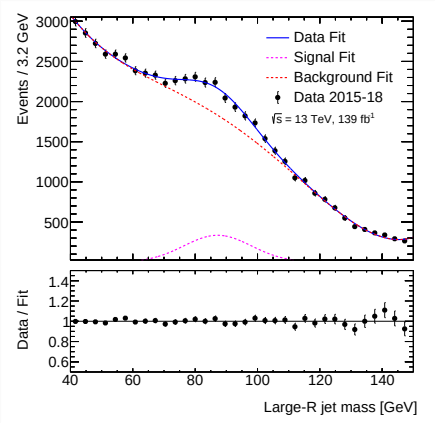
$$\epsilon_{\text{data}} = \frac{N_{\text{signal}}^{\text{tagged}}}{N_{\text{signal}}^{\text{pre-tagged}}} \approx 36\% \qquad \epsilon_{\text{MC}} \approx 29\%$$

Tagging Efficiency Extraction in $W(\ell\nu)V(qq)$ Selection

Without 50% WP ANN Tagging



With 50% WP ANN Tagging



$$\epsilon_{\text{data}} = \frac{N_{\text{signal}}^{\text{tagged}}}{N_{\text{signal}}^{\text{pre-tagged}}} \approx 16\% \qquad \epsilon_{\text{MC}} \approx 39\%$$

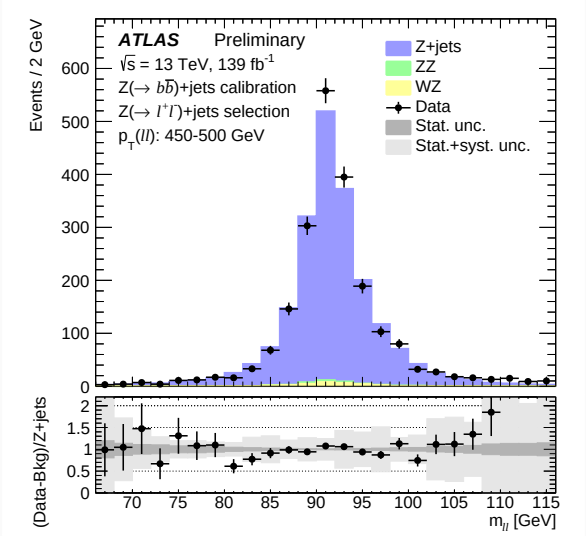
Scale Factor Results

- Extracted scale factors for the ANN tagger:
  - $SF_{ZV} = 1.28 \pm 0.47$
  - $SF_{WV} = 0.41 \pm 0.09$

- Observations on uncertainties:
  - The uncertainty in SF_{ZV} is considerably large due to error propagation from unstable fit parameters.
  - The SF_{WV} result may be affected by an apparent overestimation of the pre-tagging signal events.

$$\mathrm{SF} = \frac{\epsilon_{\mathrm{data}}}{\epsilon_{\mathrm{MC}}}$$

Pre-Tagging Event Estimation



Summary:

- Pre-tagging signal events could be estimated as using Z+jets events.
- Strategy: extract from the fully leptonic channel.
- Semi-leptonic channel then derived using the branching ratio.
- Provides an indirect cross-check for the scale factor extraction.

Conclusions and Future Work

- Scale factors for the ANN tagger were extracted in the $Z(\ell\ell)V(qq)$ and $W(\ell\nu)V(qq)$ channels.
- The $W(\ell\nu)V(qq)$ channel is especially challenging due to strong background contamination.
- Results show potential, but uncertainties remain significant.

- In further studies:
  - Systematic uncertainties should be included in the analysis.
  - The fully leptonic decay channel could be explored to estimate pre-tagging events.

Thank you for your attention!

Questions?

Backup Slides

W/Z Jets Selections

$Z(\ell\ell)$ Channel

Advantages:

- Clean signature
- Precise mass reconstruction
- Lower background contamination

Disadvantages:

- Smaller branching ratio
- Lower event statistics

$W(\ell\nu)$ Channel

Advantages:

- Higher branching ratio
- Larger event statistics

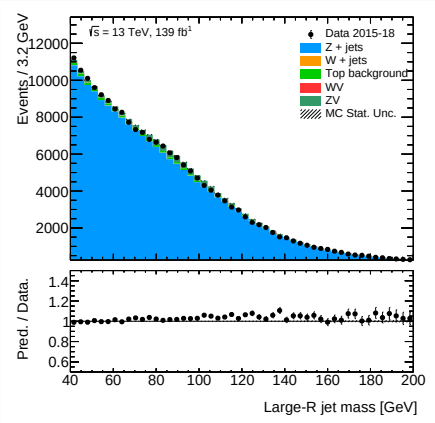
Disadvantages:

- Presence of a neutrino leads to missing transverse energy
- Larger backgrounds (e.g., $t\bar{t}$, multijet)

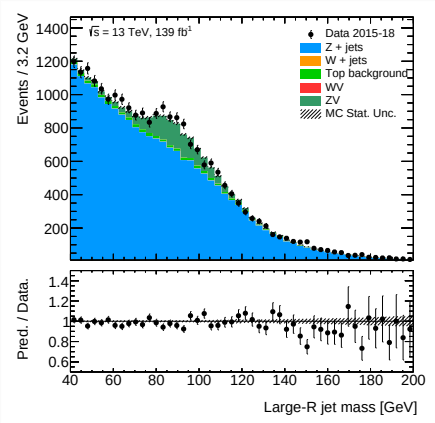
Tagging Efficiency in MC Data

Tagging Efficiency Extraction in $Z(\ell\ell)V(qq)$ Selection MC Data

Without 50% WP ANN Tagging



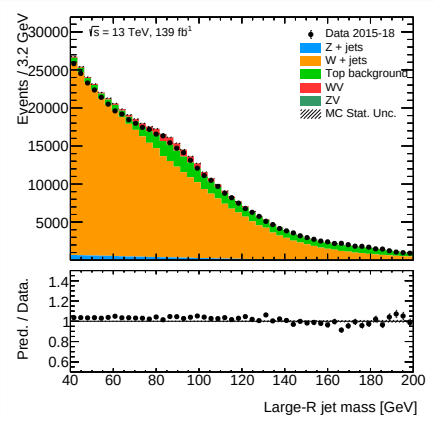
With 50% WP ANN Tagging



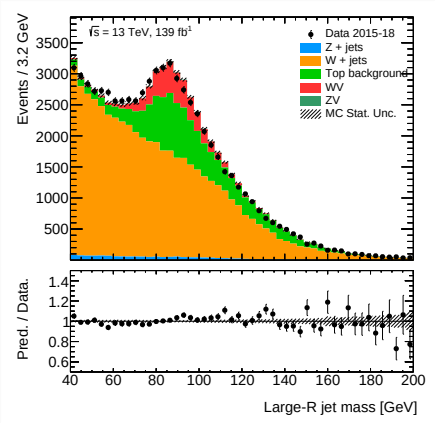
$$\epsilon_{MC} = \frac{N_{signal}^{tagged}}{N_{signal}^{untagged}} \approx 29\%$$

Tagging Efficiency Extraction in $W(\ell\nu)$ Selection MC Data

Without 50% WP ANN Tagging

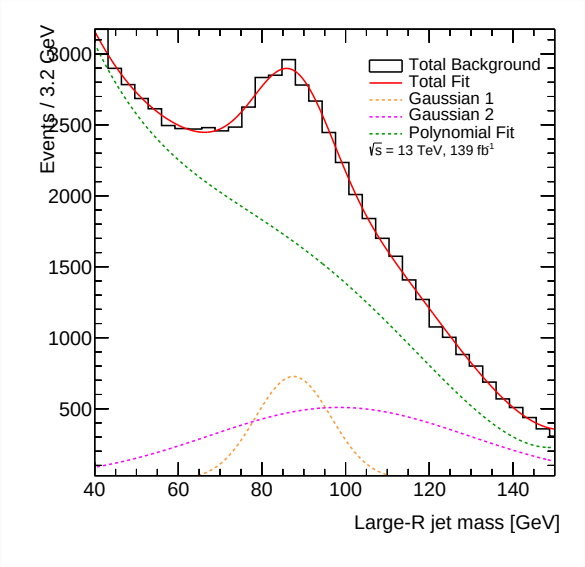


With 50% WP ANN Tagging



$$\epsilon_{MC} = \frac{N_{signal}^{tagged}}{N_{signal}^{untagged}} \approx 39\%$$

Background Fit in $W(\ell\nu)$ Channel



Fit Description

- **V + jets:** modeled with a 4th-order polynomial.
- **Top background:** modeled with two Gaussians.