Study of Quark/Gluon Jet Tagging Performance and Optimisation Using GN3

02 September 2025

Melik Kaan Şelale, Neelam Kumari, Krisztian Peters

Overview

Part I: Background

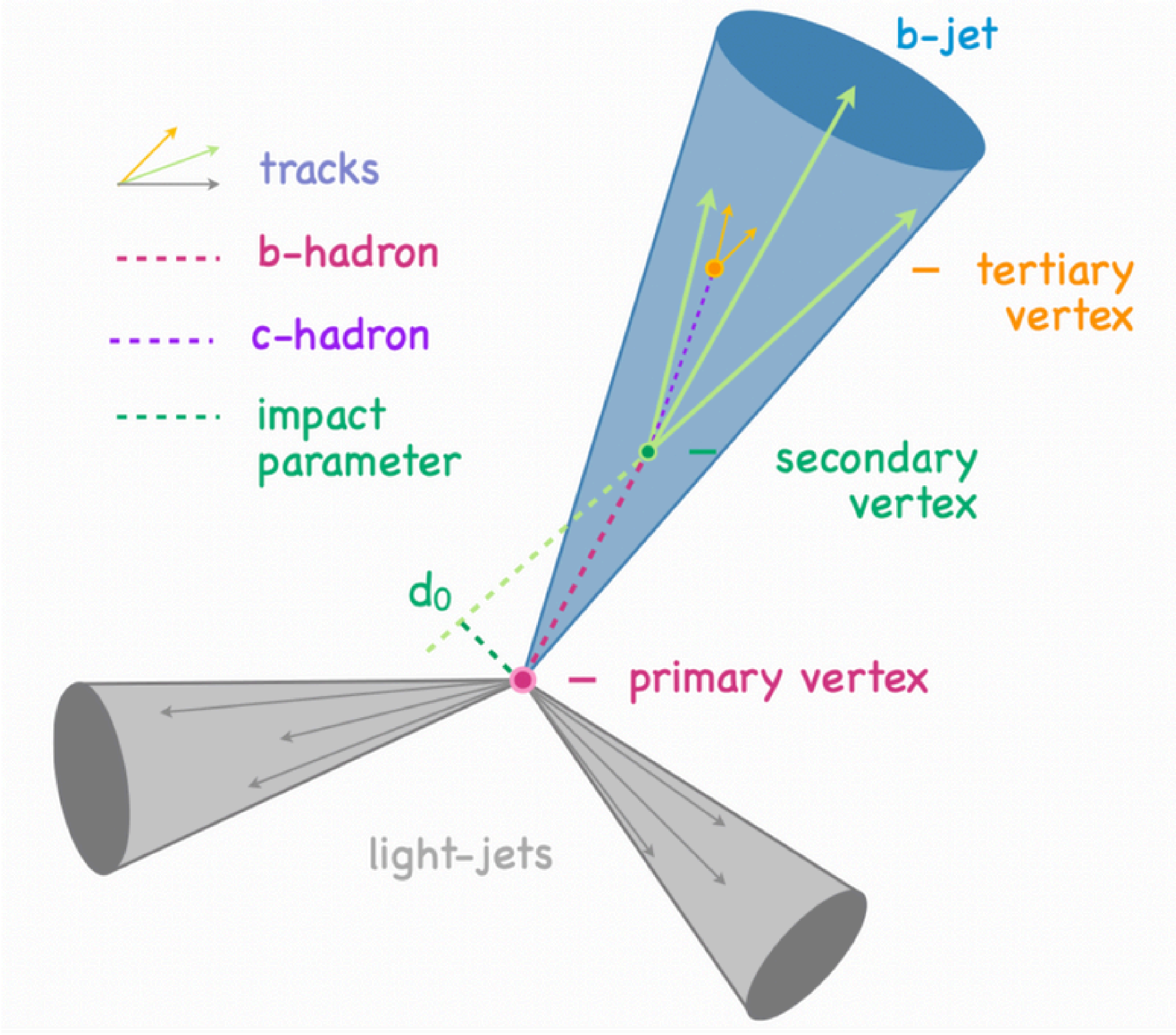
Introduction to Flavour Tagging

Objective: The primary goal of **Flavour Tagging** is to identify the flavour of a jet — specifically, whether it contains a b-quark, c-quark, tau-lepton, or from neither, in which case it is classified as a light jet.

Why: Flavour tagging is essential for many physics analyses, including Standard Model measurements and searches for new physics.

Primary Vertex is a reference point where this first jets are formed

TRACK	JET
• The reconstructed trajectory of a single charged particle • Properties: pT, charge, impact parameters etc.	• A collimated spray of many particles • Properties: Total energy, momentum, flavour etc.



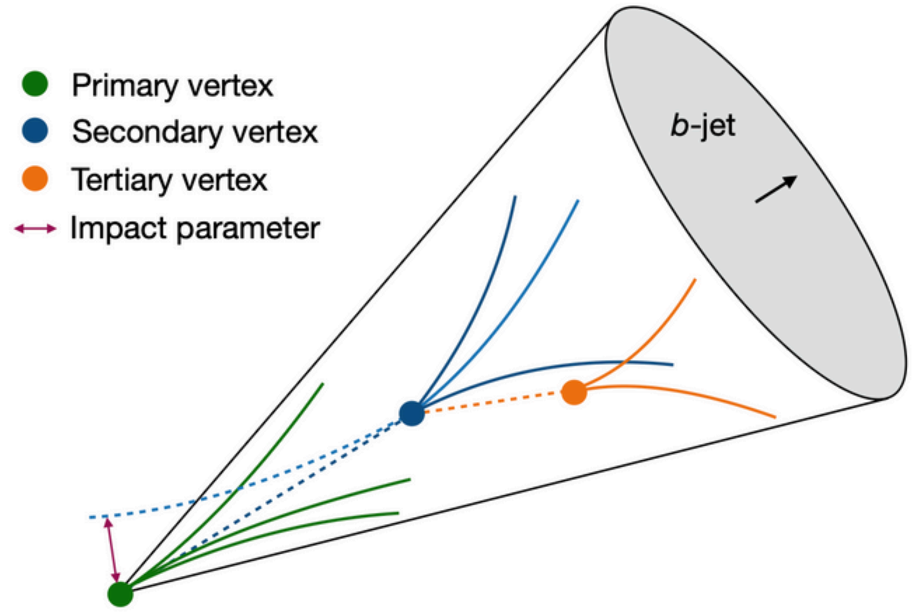
Track-Jet Association: Tracks are associated to a jet via a **Ghost Association**

Physical Signatures

Physical Property	Signature in the Detector	Distinguishing Power
• Long Lifetime: ◦ b-hadrons: ~1.5 ps ◦ c-hadrons: Shorter lifetime ◦ light-hadrons: Very short lifetime	• Large Impact Parameters (d_0, z_0) and a reconstructable, displaced Secondary Vertex.	• Heavy-flavour (b/c) jets from light-flavour jets.
• The high mass (~5 GeV) of the b-hadron.	• SV --> high invariant mass and a high number of associated tracks.	• Distinguish b-jets from c-jets and from low-mass background vertices in light-jets.
• The unique physical decay chain b → c.	• c-hadron decay --> Tertiary Vertex.	• Powerful signature to separate b-jets from c-jets.

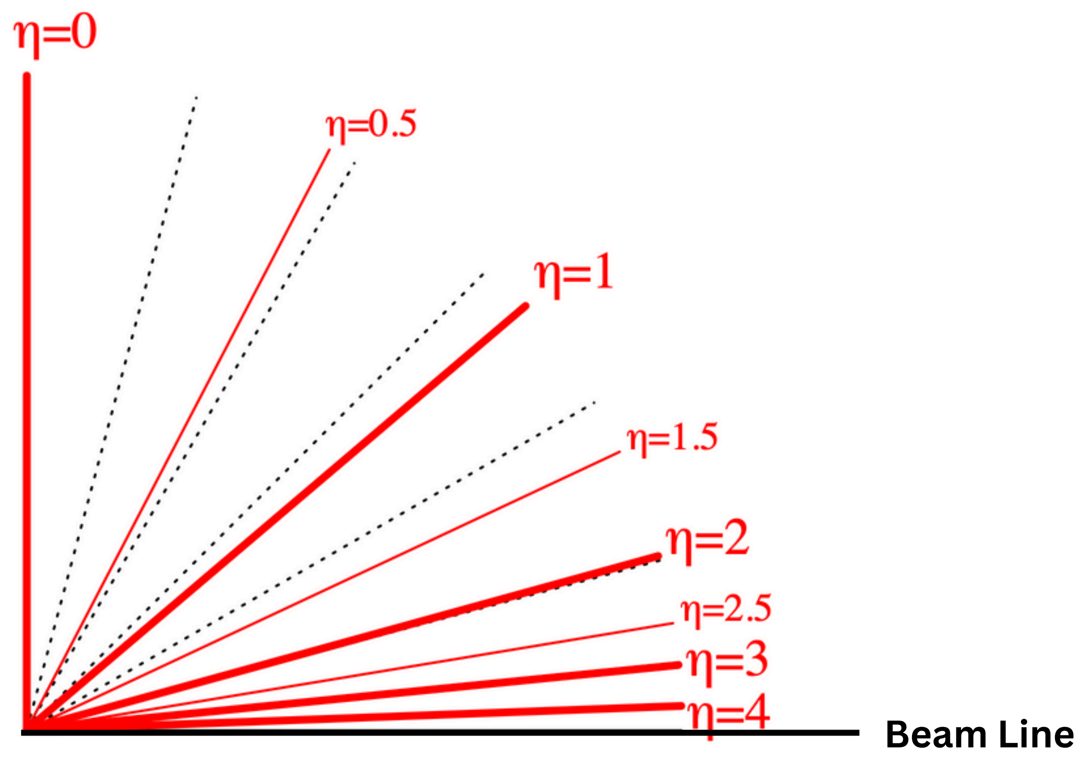
b vs. light-jet separation is strong: Based on the presence vs. absence of a displaced, high-mass secondary vertex.

b vs. c-jet separation is more challenging: Based on statistical differences (mass, lifetime) and the unique topological signature of the tertiary vertex.



Motivation of This Study

GN3 enables separate identification of u/d-, s-hadrons, and gluons — providing direct gluon vs. light-quark discrimination



Extending Eta Coverage to |η|<4.5 : Allows forward jet tagging!

- **QCD suppression:** forward jets are more quark-like; backgrounds more gluon-rich
- **Sensitivity:** access to potential BSM effects at high |η|
- **Robustness:** to make model generalised over eta coverage

Adding Dijet Samples

- **Gluon-rich input:** strengthens light-quark vs. gluon discrimination.
- **Forward coverage:** more jets in Forward Eta Region
- **Better generalisation:** across jet flavours and varying kinematics.
- **More statistics:** improved precision at phase-space edges.

Group's first attempt at studying the forward η region and quark–gluon discrimination.

There are still many aspects to optimize in order to evaluate performance in this region.

Part II: Technical Setup

Selections

	Baseline Samples	Extended Eta Samples	Extended Eta + Dijet Samples
Eta Coverage	Central Region	Central + Forward Regions	Central + Forward Regions
Low pT	ttbar	ttbar	ttbar+dijet
High pT	Zprime	Zprime	Zprime + Dijet

Transverse Momentum Definitions

Low pT: 20 GeV to 250 GeV
High pT: 250 GeV to 6000 GeV

Pseudorapidity Definitions

Central Region: $|\eta|<2.5$
Forward Region: $2.5<|\eta|<4.5$

Why are specific samples chosen for specific pT regions? --> Good Fraction

Train Setup and Input Variables

Input Variables

Tracks: Charged-particle trajectories reconstructed in the detector, carrying kinematic, displacement, and quality information useful for heavy-flavour identification with **ghost-association.**

Jets: Calibrated sprays of tracks, described by overall momentum and direction.

Electrons: Identified electron tracks within jets.

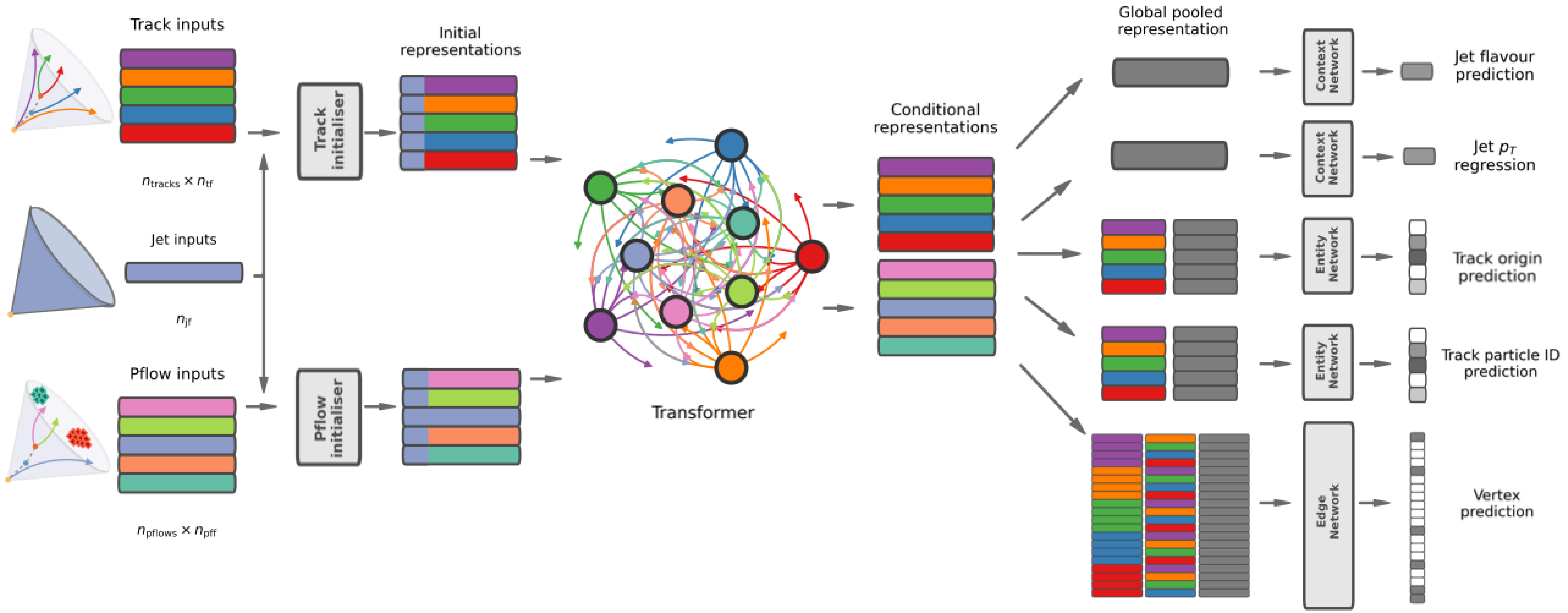
Flows: Jet constituents (charged or neutral) with basic kinematic and angular variables relative to the jet axis.

Training and Test Setups:

- **Baseline:** Standard samples used in GN3
- **Extended Eta:** Pseudorapidity coverage increased up to 4.5
- **Extended Eta + Dijet:** In the low pT region dijet samples are added alongside ttbar, and in the high pT region dijet samples are added alongside Zprime.

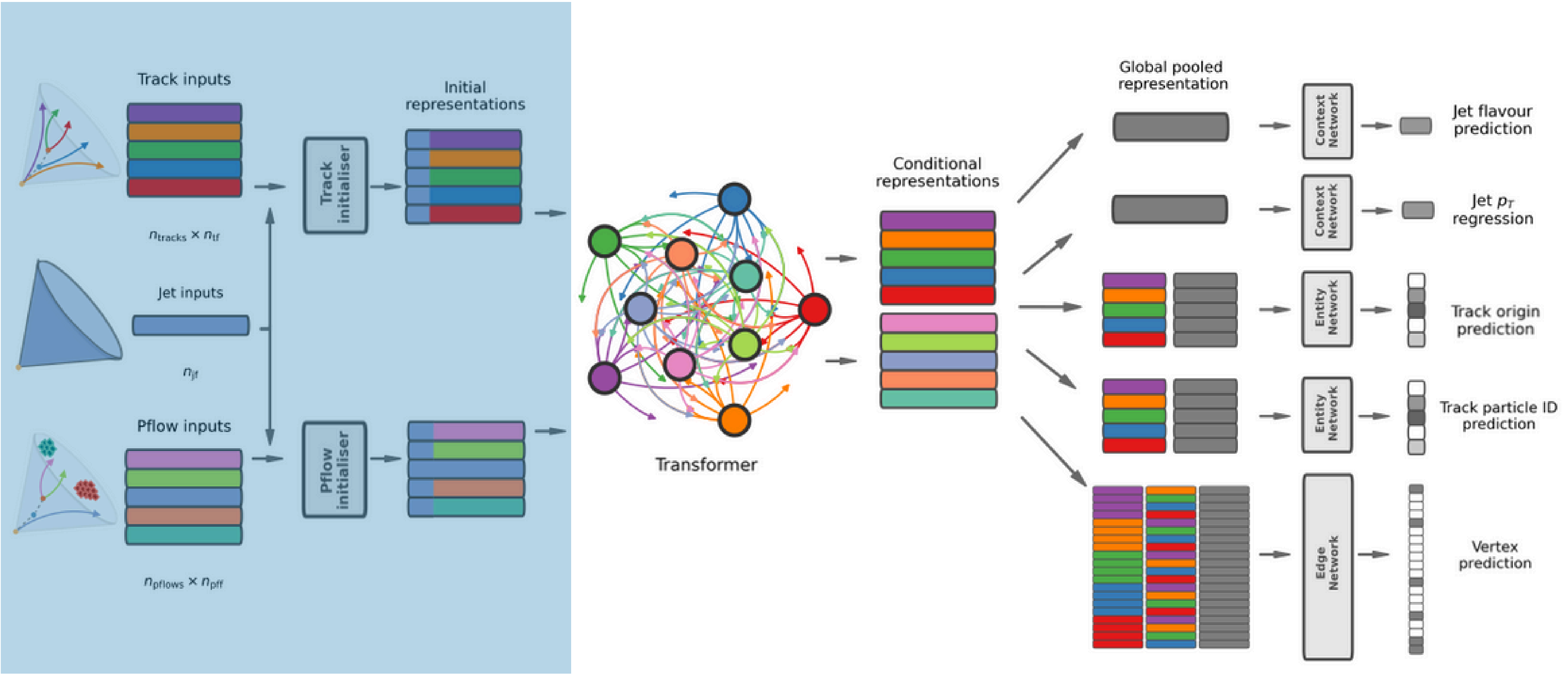
Pre-Process	Training	Evaluation
• Baseline - 10 Millions Sample	• Baseline - 10 Millions Sample	• Extended Eta - ttbar - ~2.5 Millions Sample
• Extended Eta - 30 Millions Sample	• Extended Eta - 30 Millions Sample	• Extended Eta - Zprime - ~2.5 Millions Sample
• Extended Eta + Dijet - 30 Millions Sample	• Extended Eta + Dijet - 30 Millions Sample	• Extended Eta - dijet - ~2.5 Millions Sample

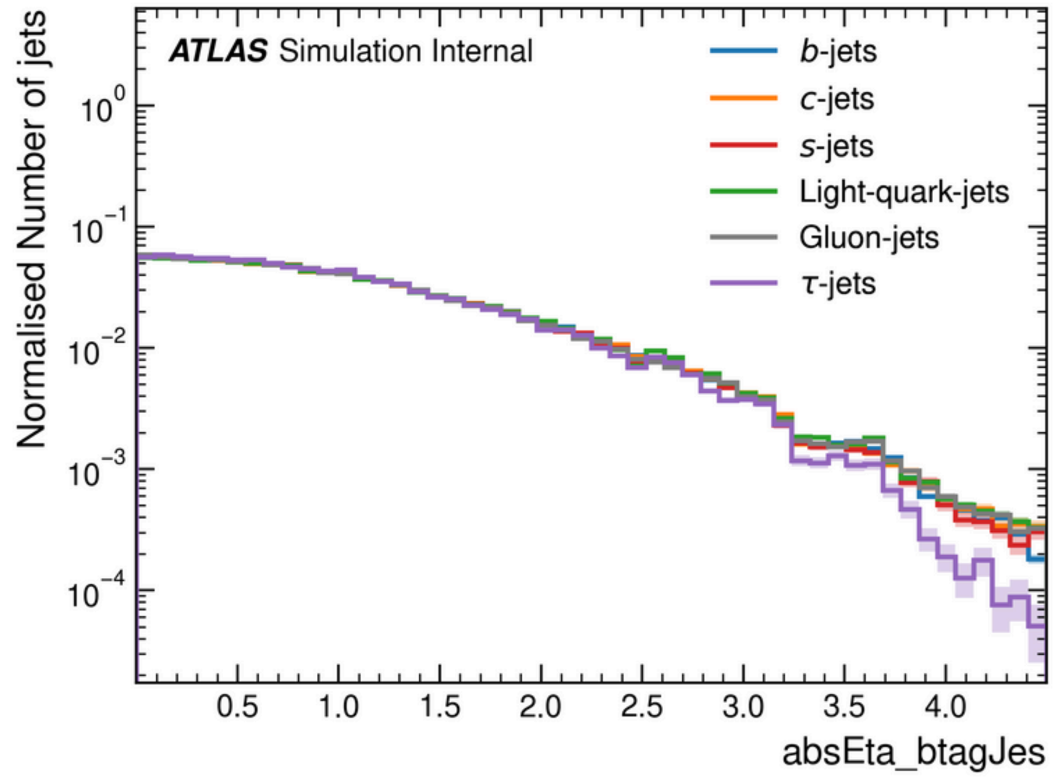
GN3 Architecture

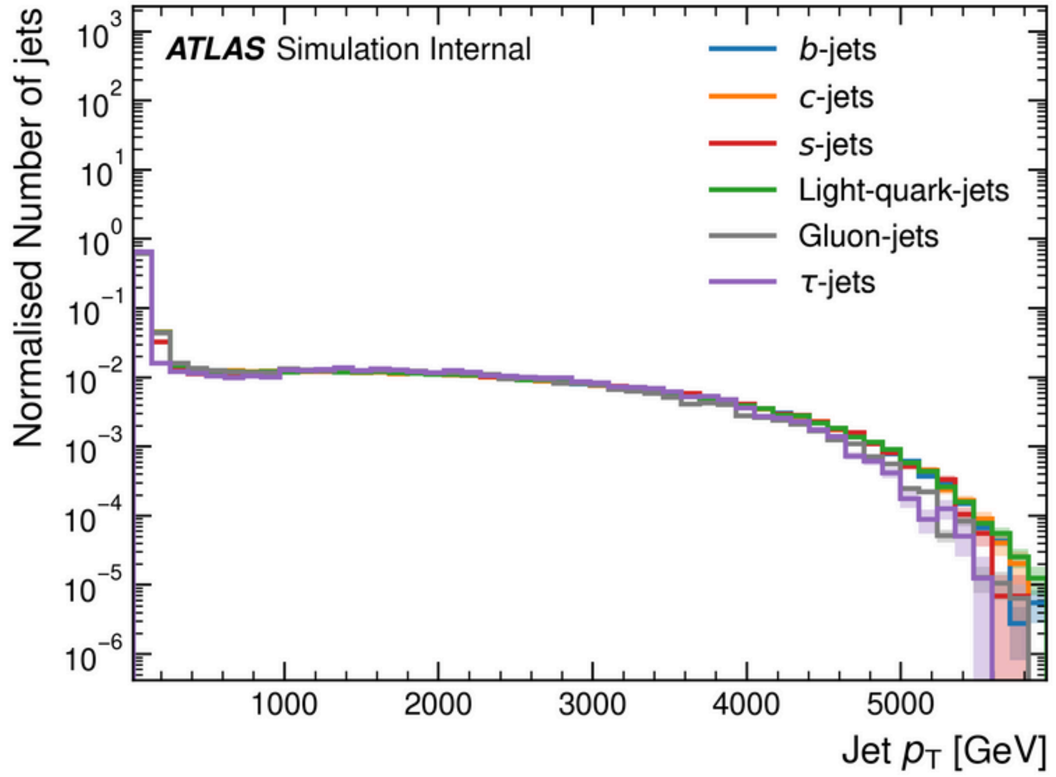


GN3 Architecture

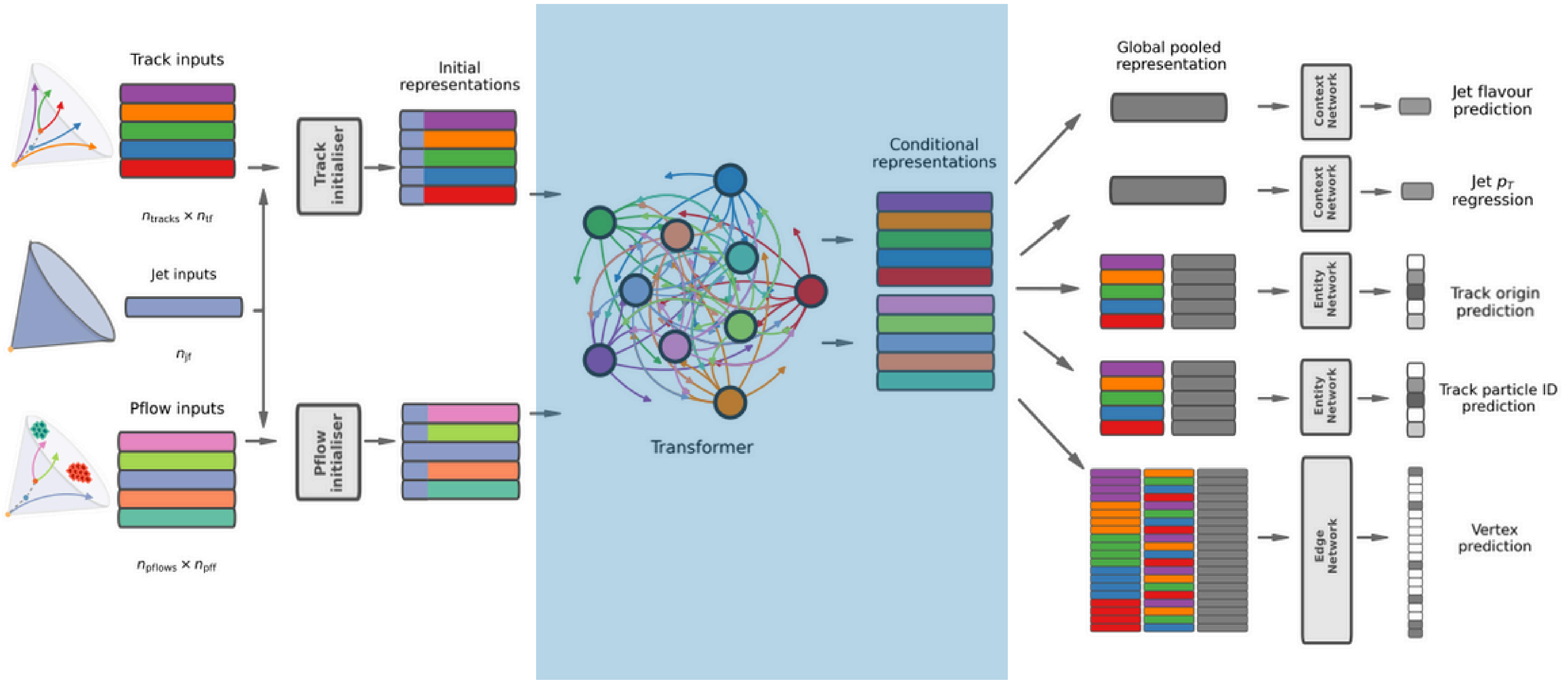
Preprocessing - Kinematic Variables





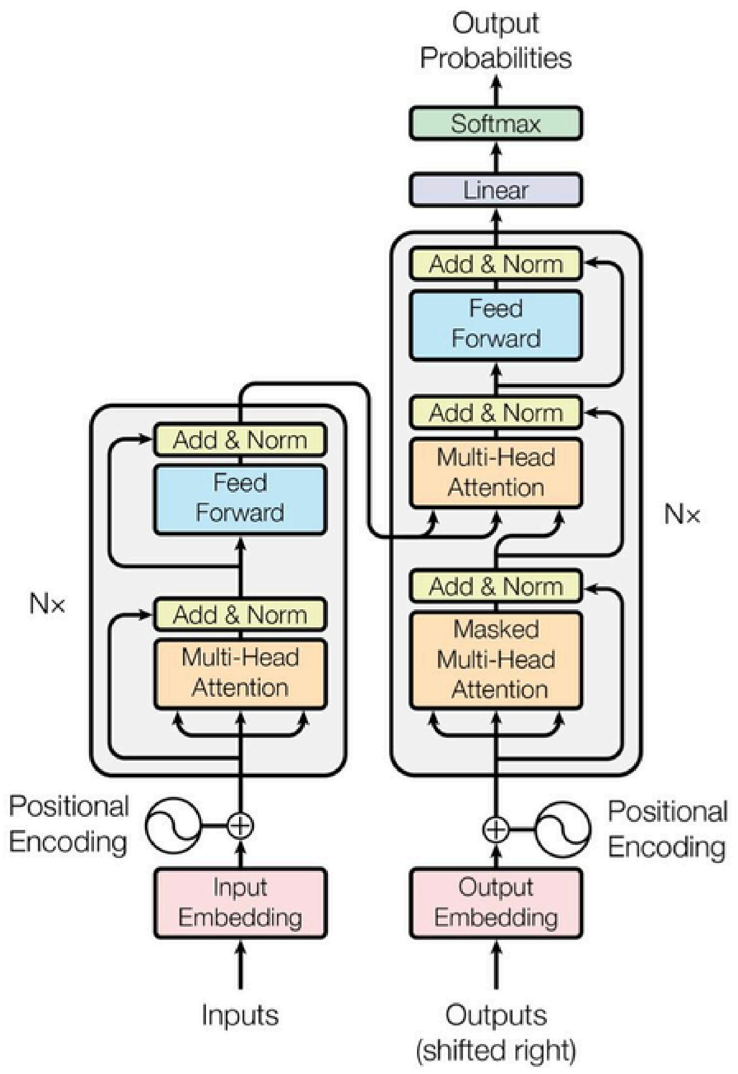


GN3 Architecture

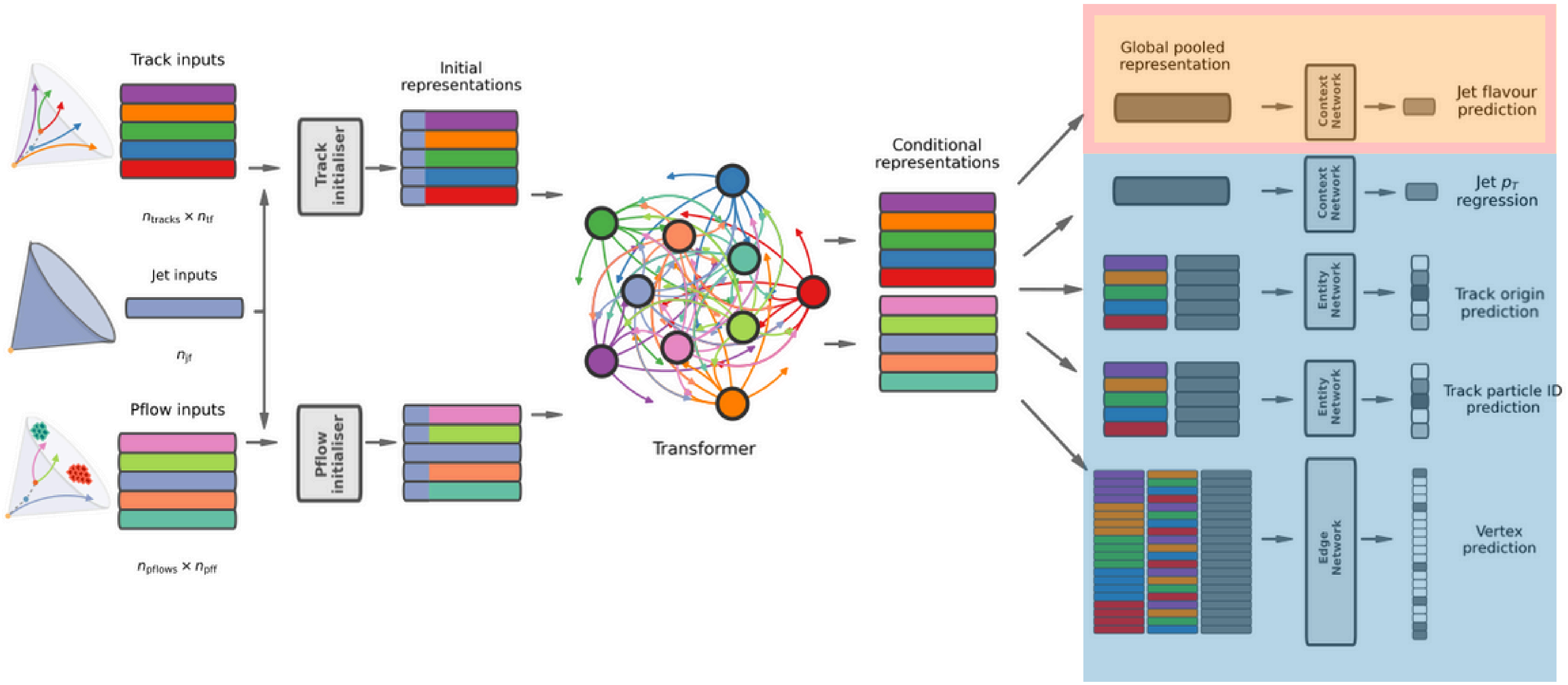


Transformer Architecture:

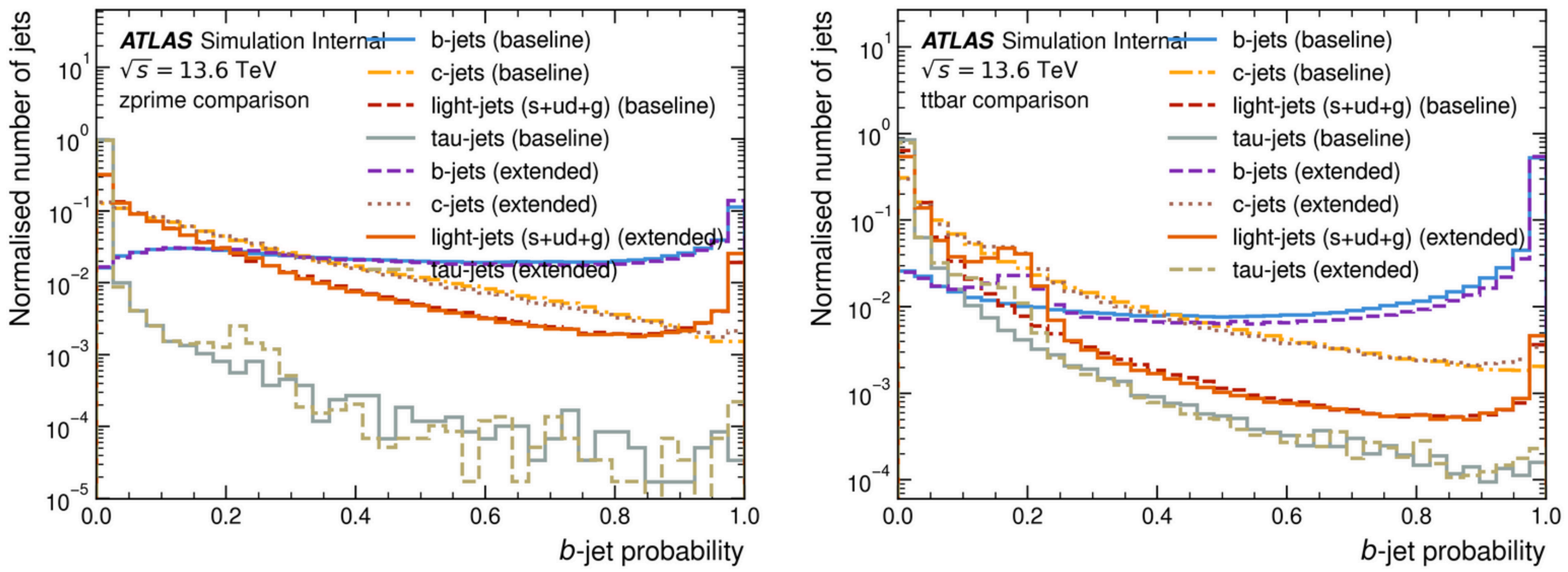
- The Transformer is an architecture that replaces recurrence with self-attention, allowing each token to attend to all others in parallel for global context modeling.
- Its core mechanism computes relevance between token pairs using query, key, and value projections, enabling dynamic weighting of contextual information.



GN3 Architecture



Main Task: Jet flavour prediction –> Probability



Four Physics-Informed Auxiliary Tasks:

- **Track-type Identification:** classify each track into particle categories.
- **Jet pT Regression:** correct reconstructed jet pT to truth level.
- **Track Origin:** estimate source of tracks
- **Vertex Prediction:** identify tracks from a common vertex.

Geometric-Loss Strategy –> $\mathcal{L}_{\text{total}} = \left(\prod_{i=1}^{5} \mathcal{L}_i\right)^{1/5}$

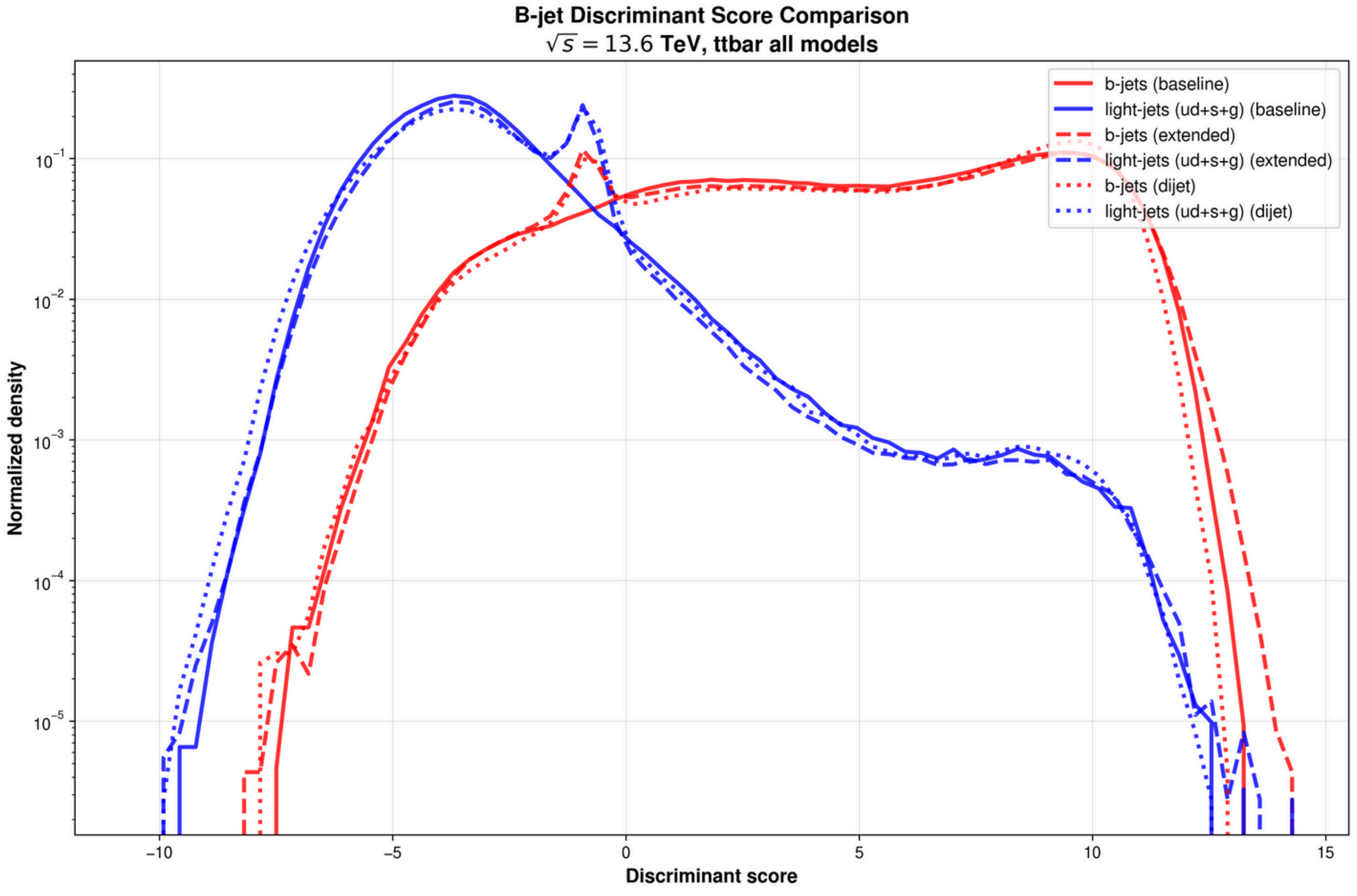
Part III: Results

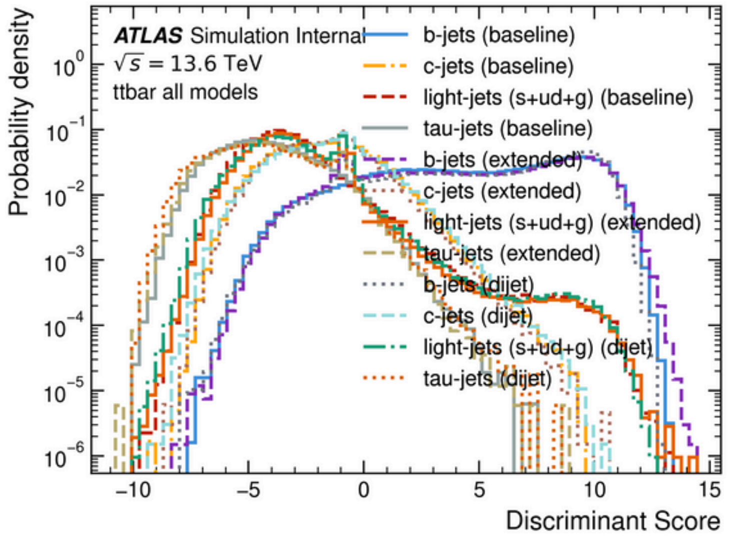
b-tagging Performance Plots

b-tagging Discriminator

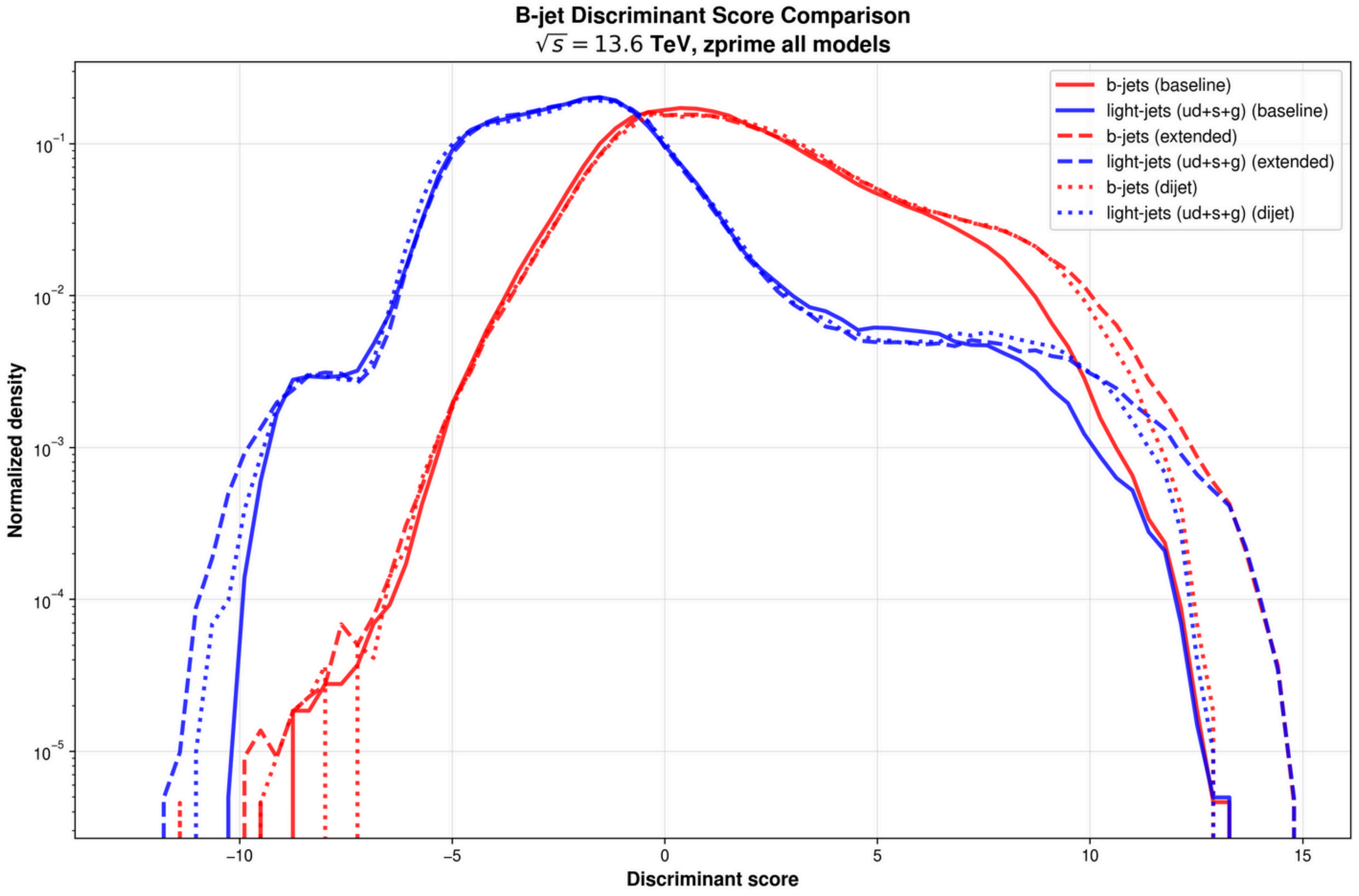
Baseline: |η|<2.5
Extended: |η|<4.5
Dijet: |η|<4.5

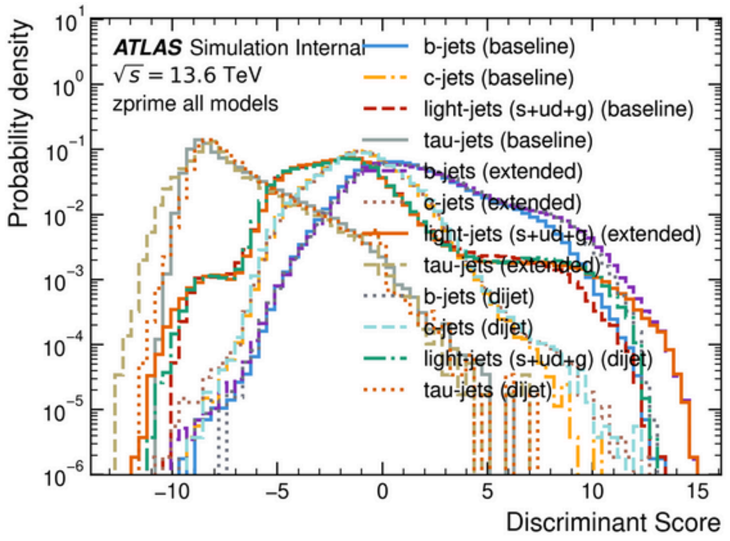
ttbar Samples





Zprime Samples





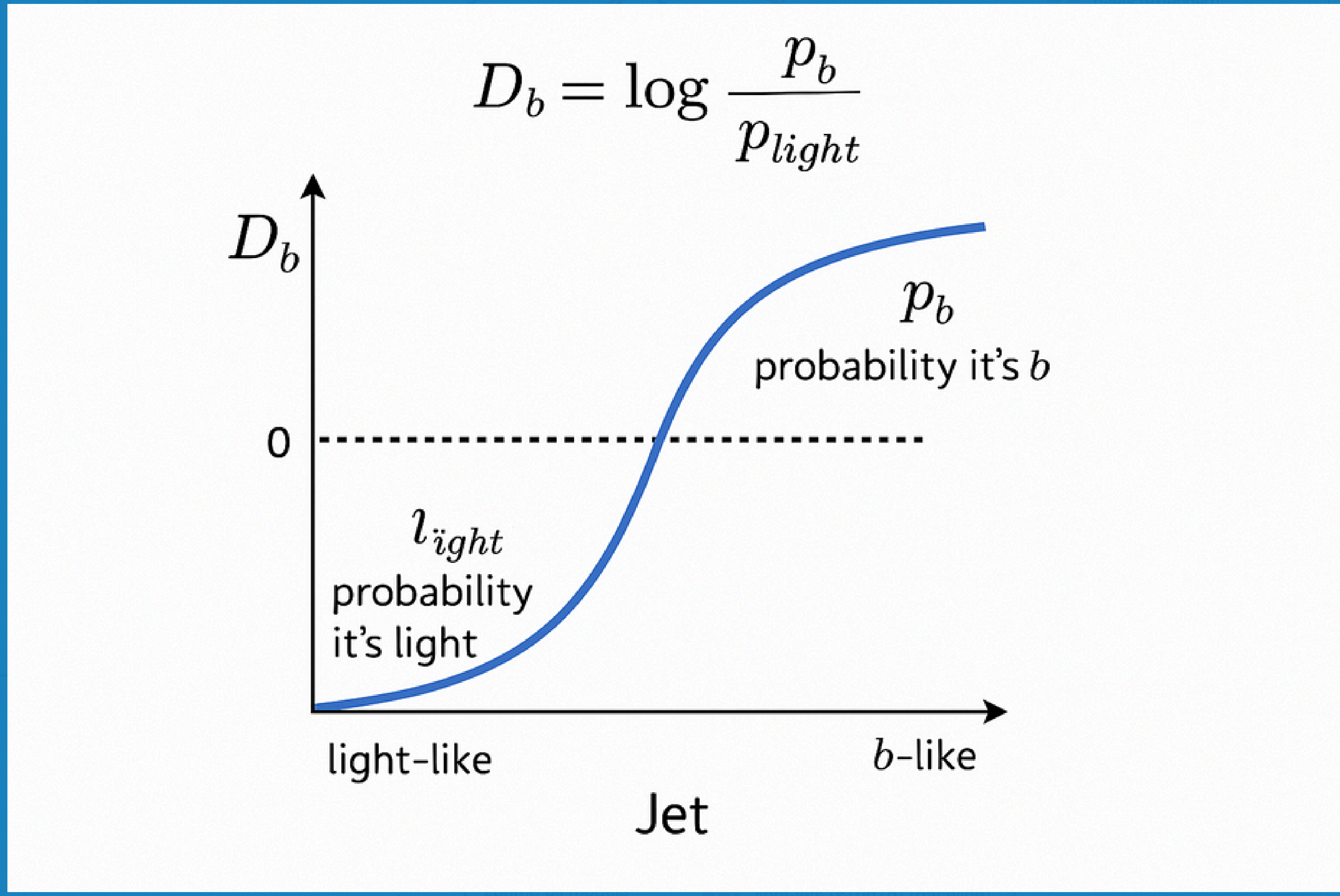
b-tagging Discriminator

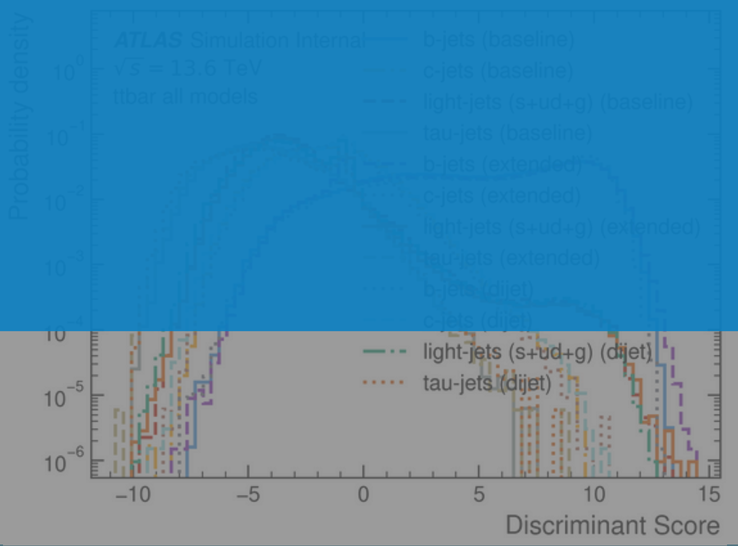
Baseline: |η|<2.5
Extended: |η|<4.5
Dijet: |η|<4.5

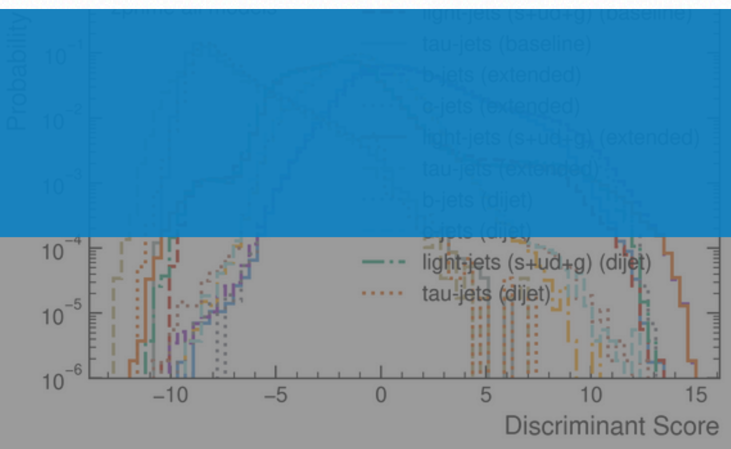
b-jet probability –> Discrimination

b-jet Discriminant

$$D_b = \log\left(\frac{p_b}{f_c\,p_c + f_\tau\,p_\tau + (1 - f_c - f_\tau)\,p_{\text{light}}}\right)$$



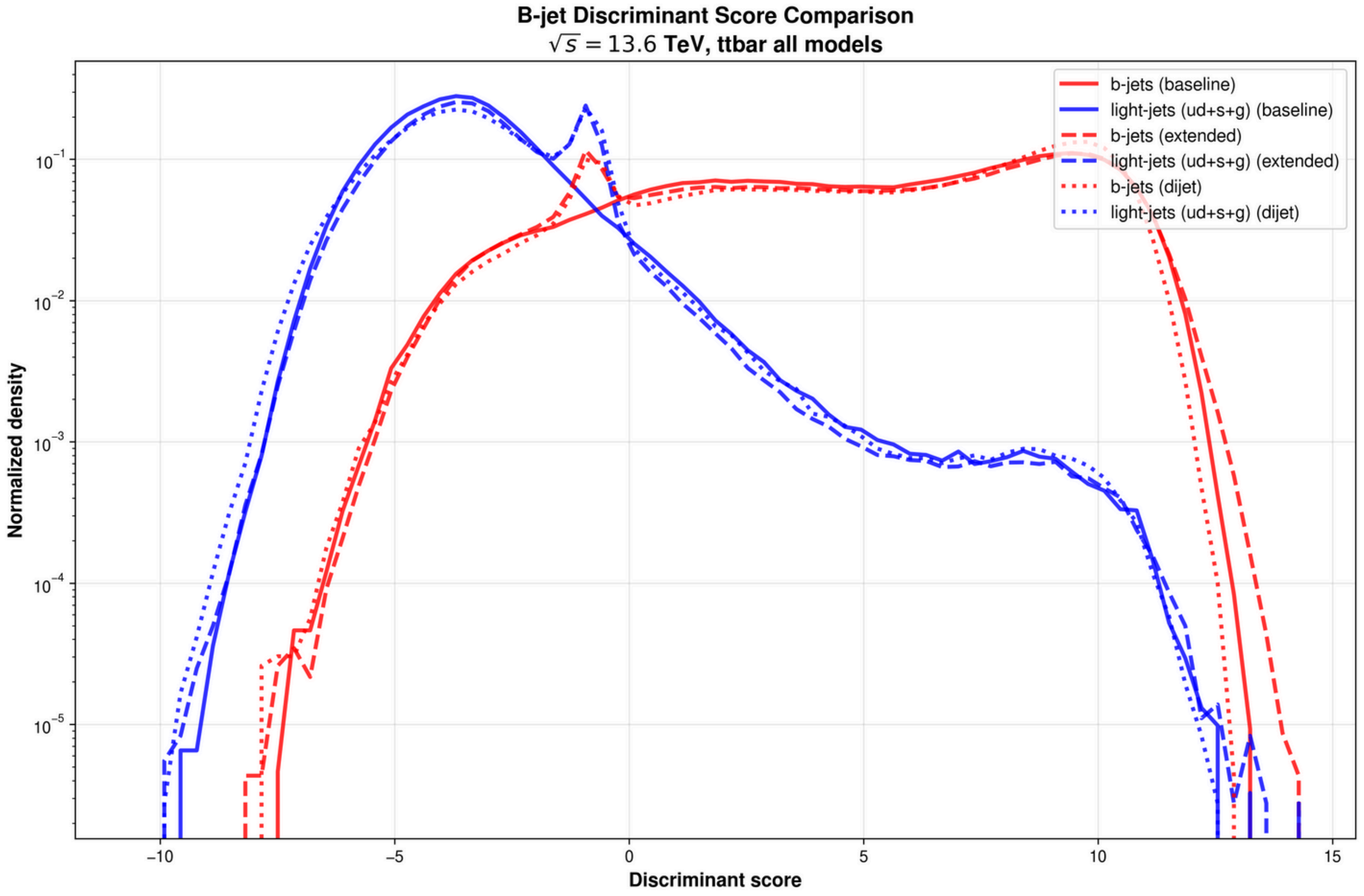


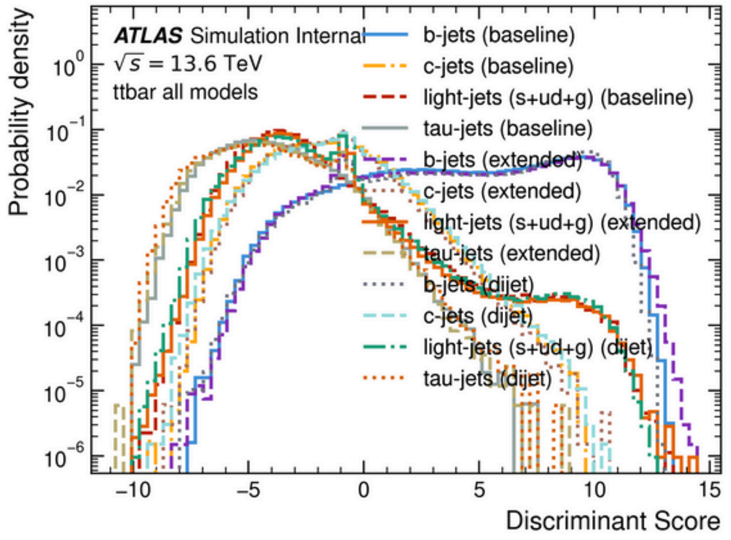


b-tagging Discriminator

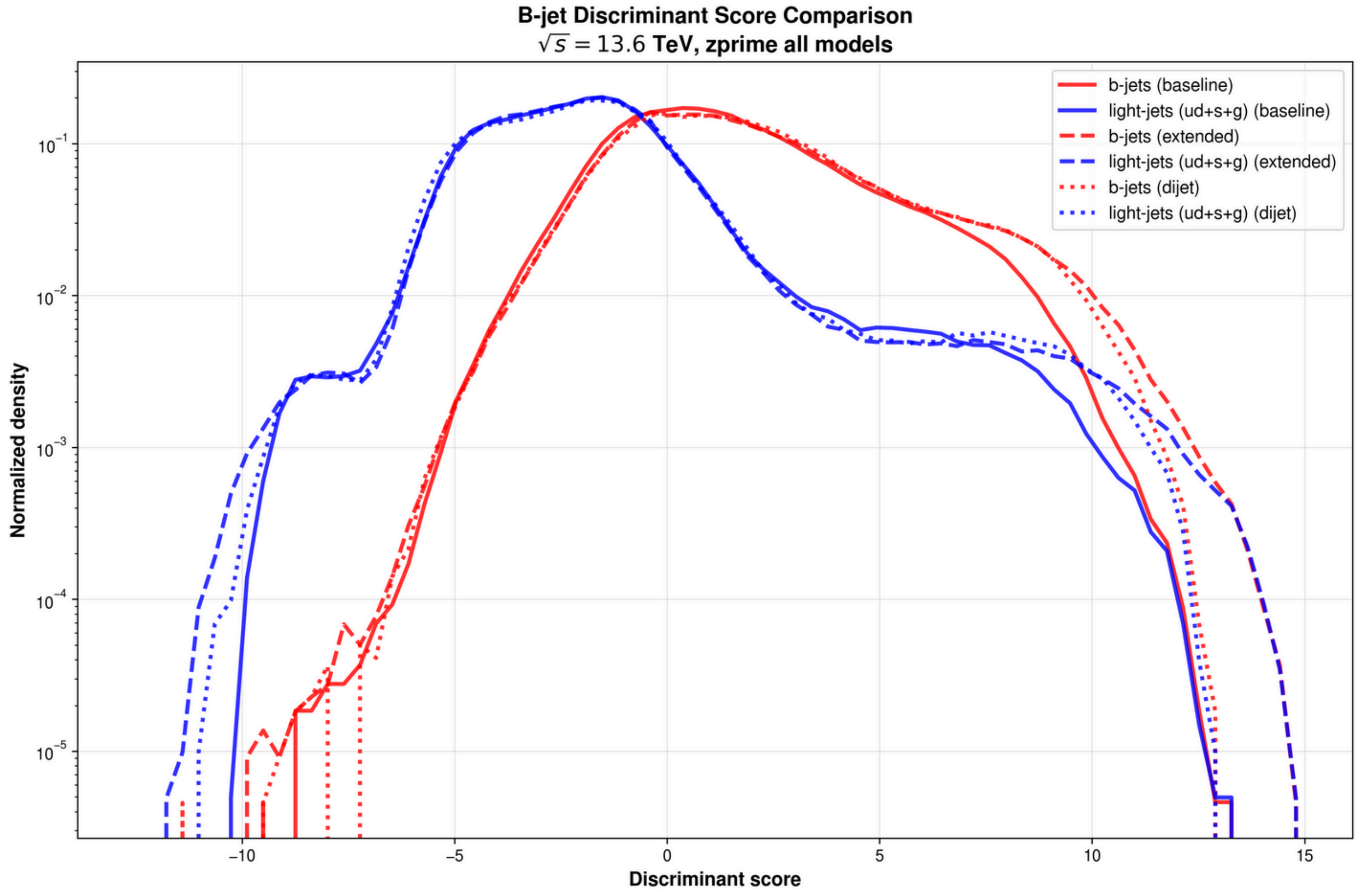
Baseline: |η|<2.5
Extended: |η|<4.5
Dijet: |η|<4.5

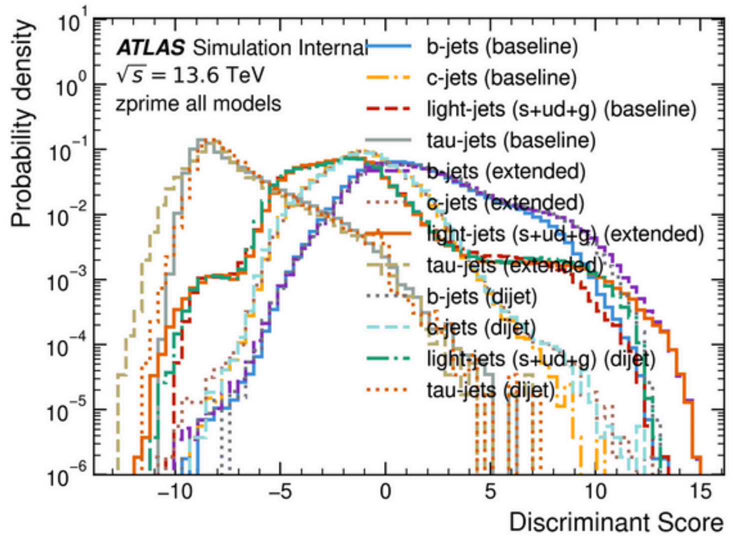
ttbar Samples





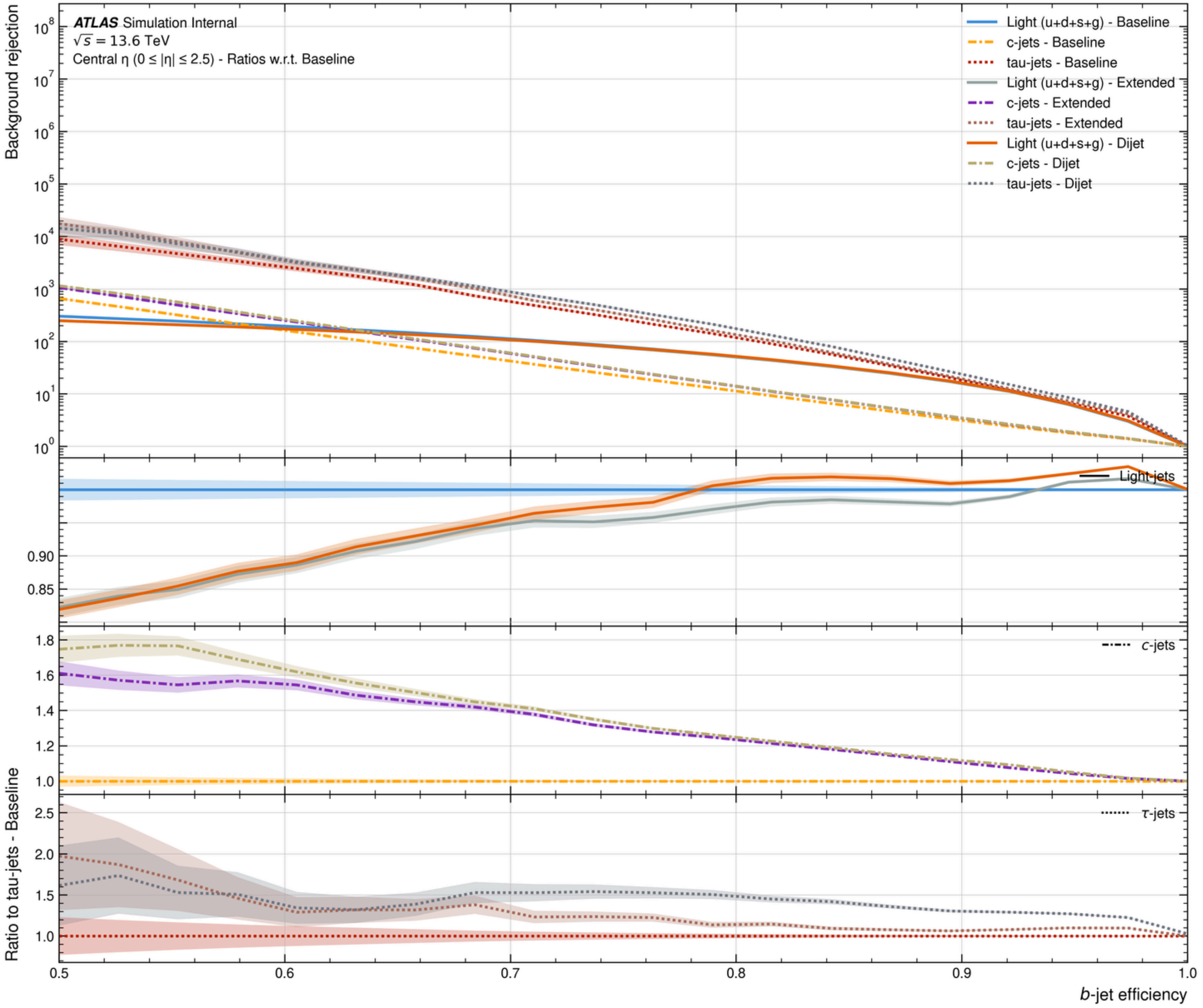
Zprime Samples



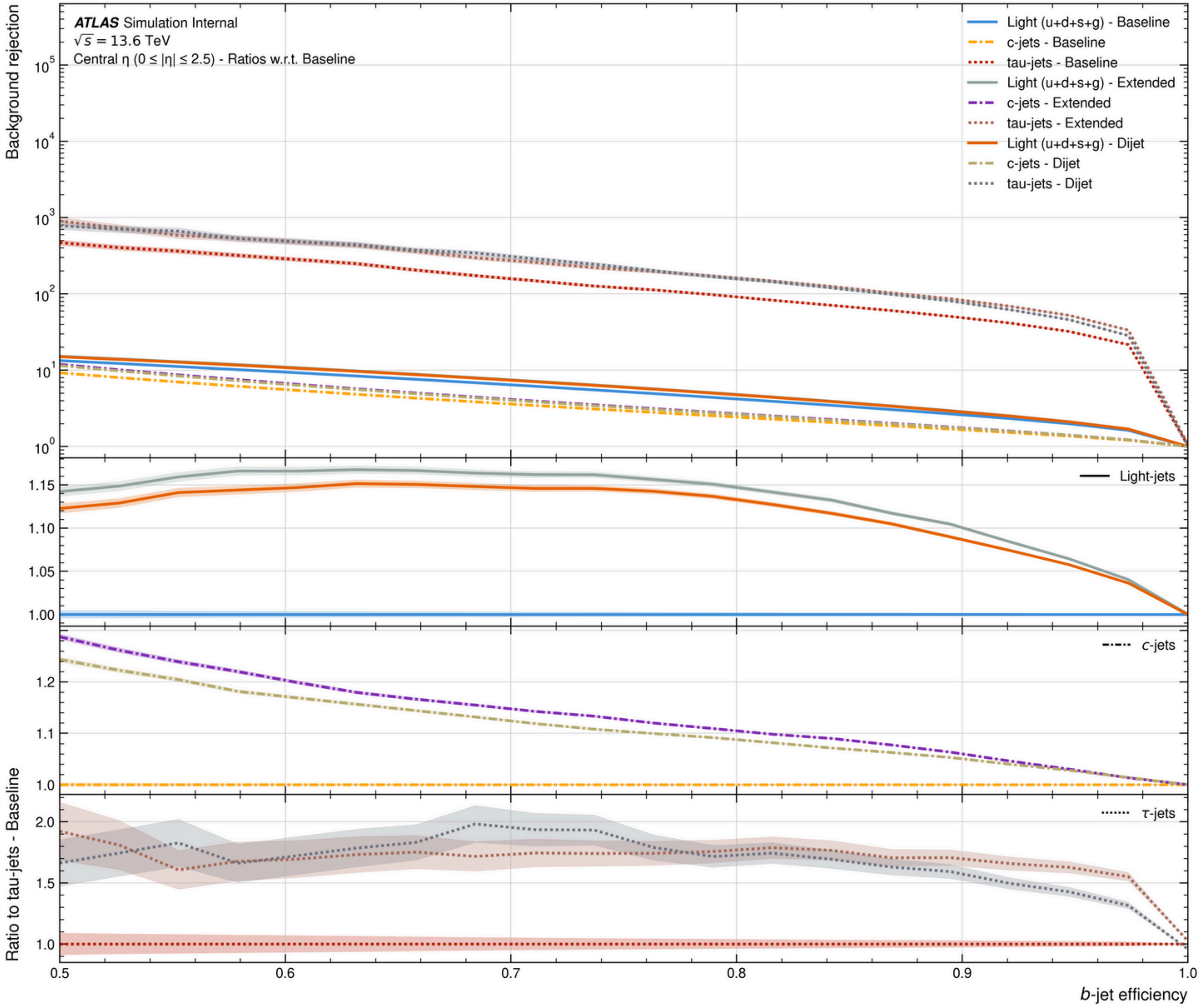


b-tagging Performance - Central

ttbar Samples - ROC Curve



Zprime Samples - ROC Curve



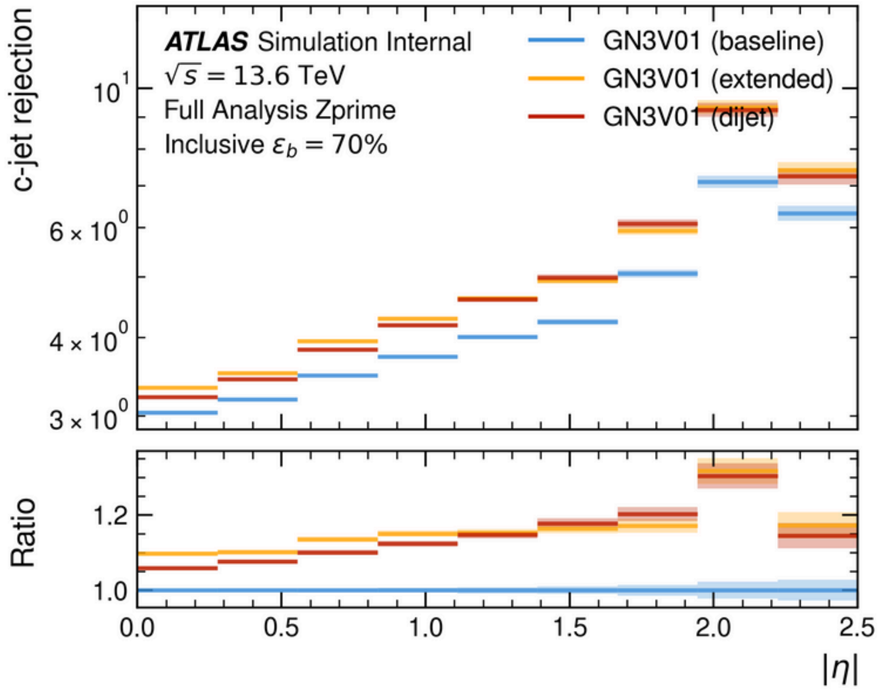
b-tagging Performance - Central

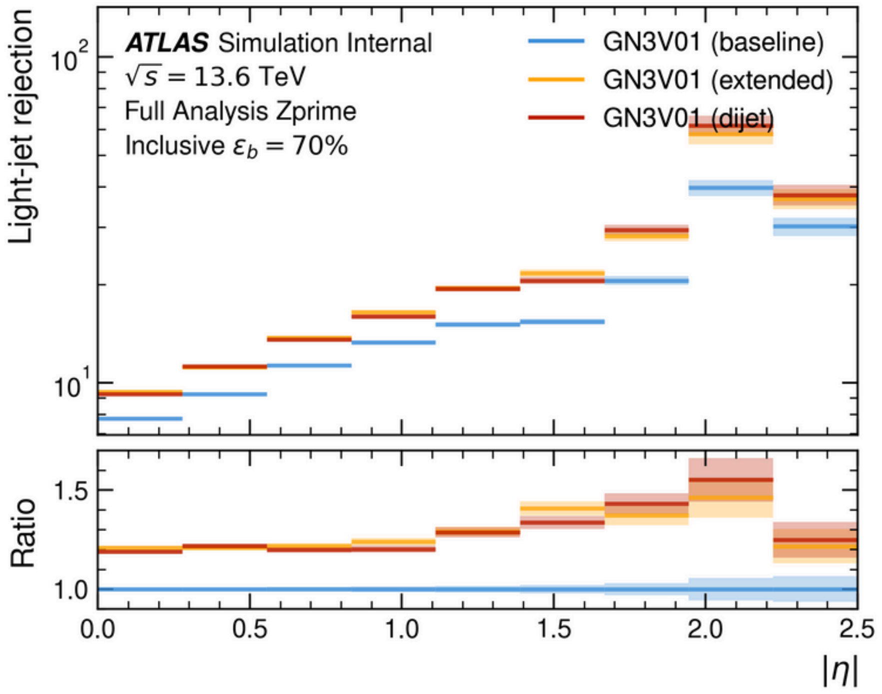
c-jet and light-jet Rejection at %70 Light-Jet Efficiency - Central Region - Zprime Samples

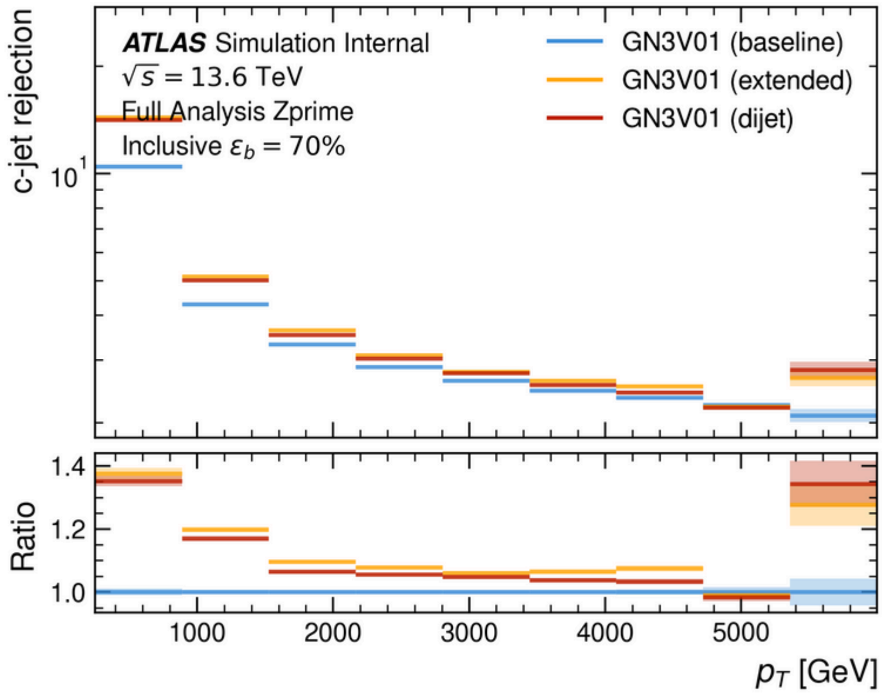
Central: |η|<2.5

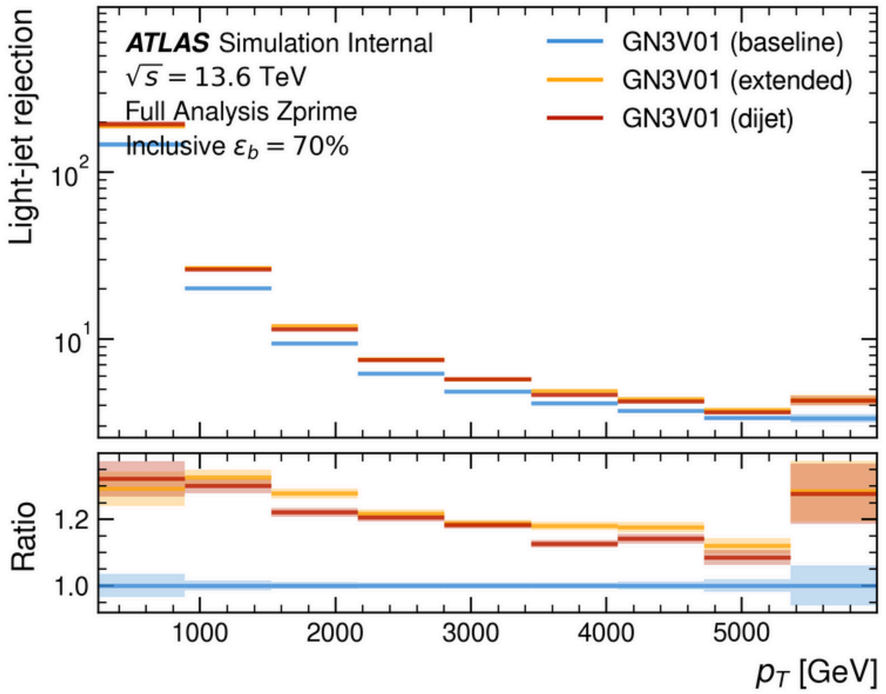
vs.Eta

vs. pT









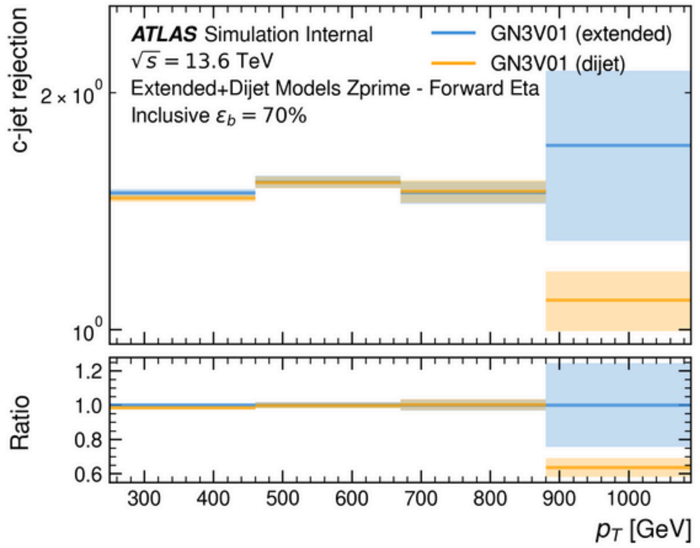
b-tagging Performance - Forward

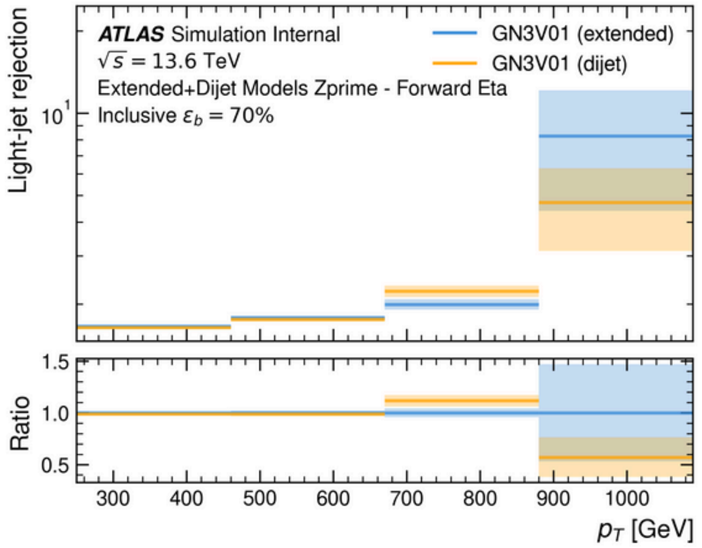
c-jet and light-jet Rejection at %70 Light-Jet Efficiency - Forward Region - Zprime

ttbar performance results are similar

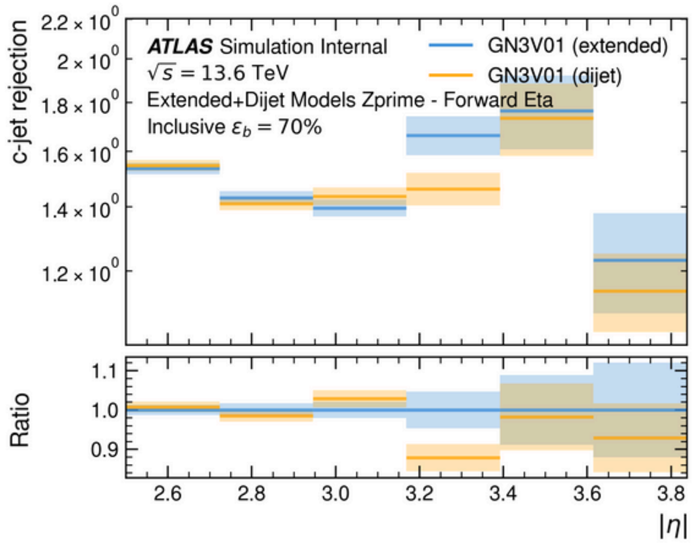
Forward: 2.5 <|η| <4.5

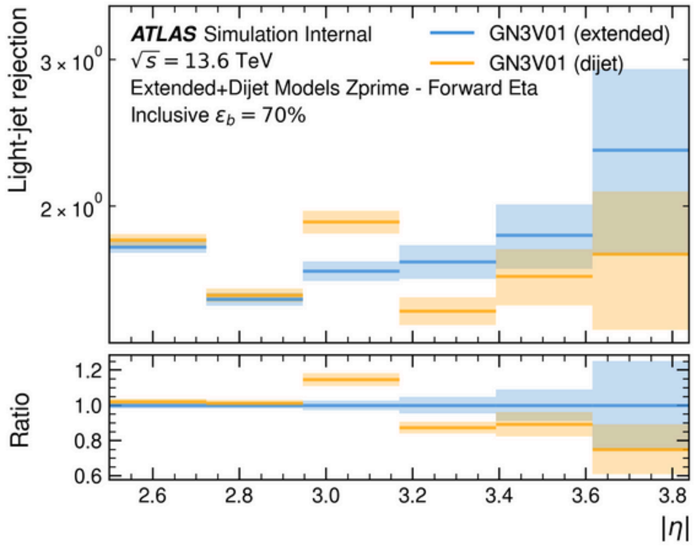
vs. pT



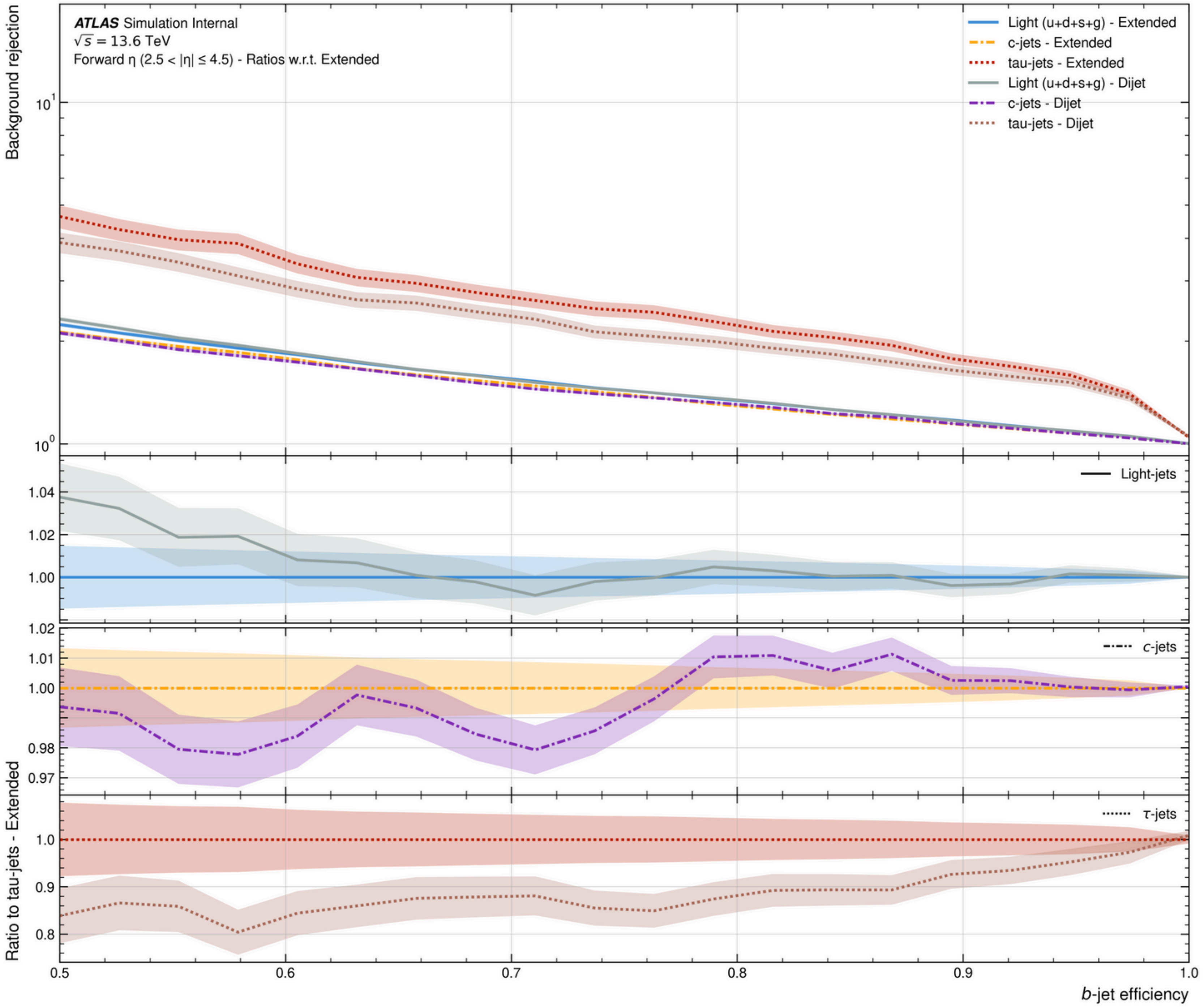


vs. eta





Zprime Samples - ROC Curve



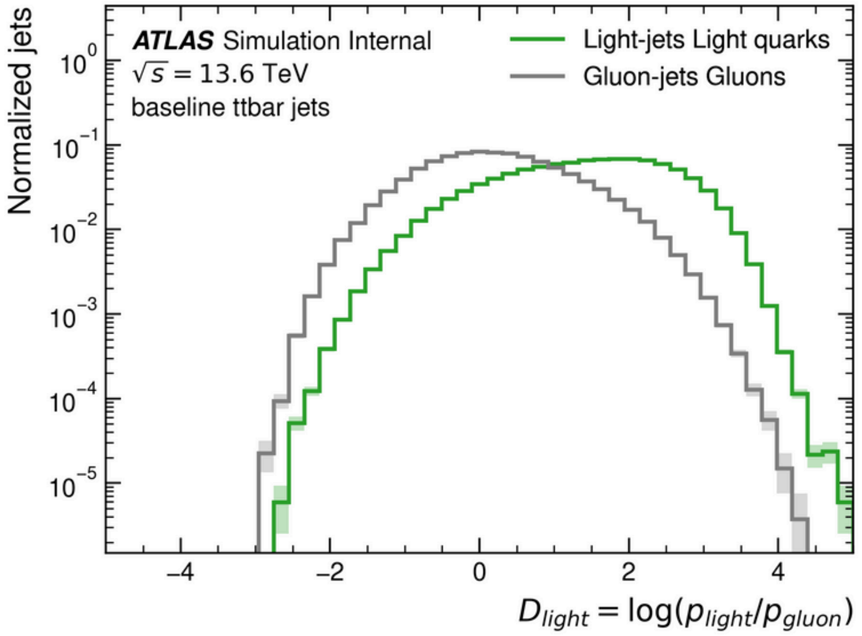
High pT!

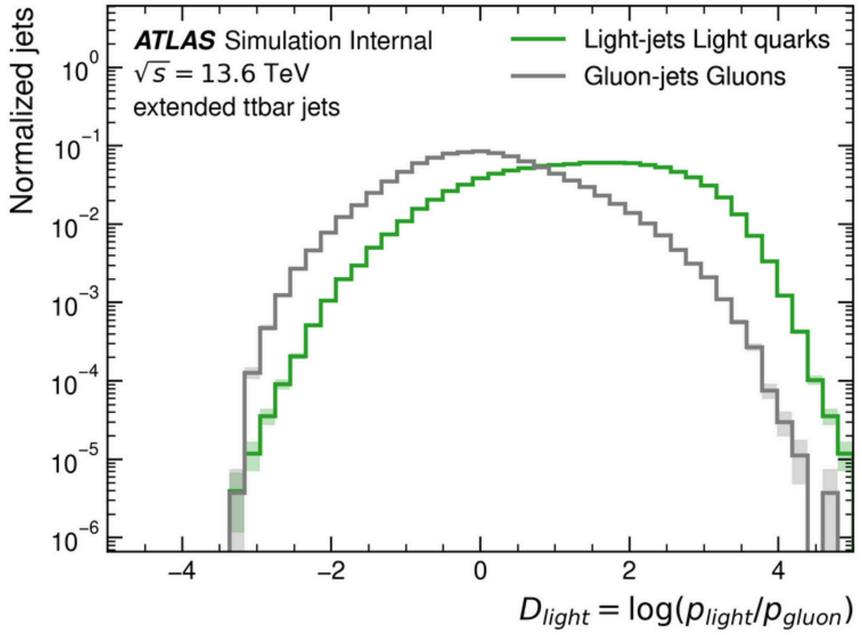
quark/gluon-tagging Performance Plots

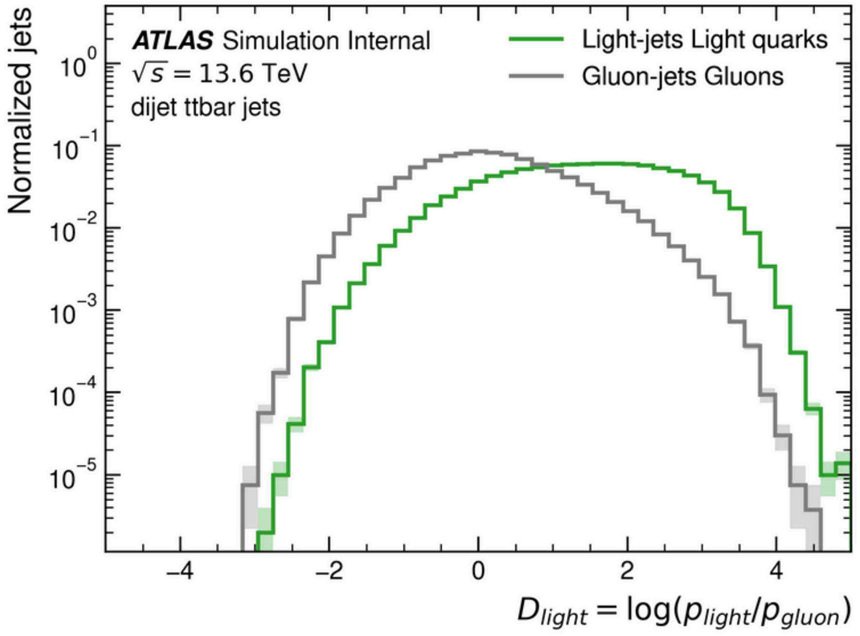
quark/gluon-tagging Discriminator

Baseline: |η|<2.5
Extended: |η|<4.5
Dijet: |η|<4.5

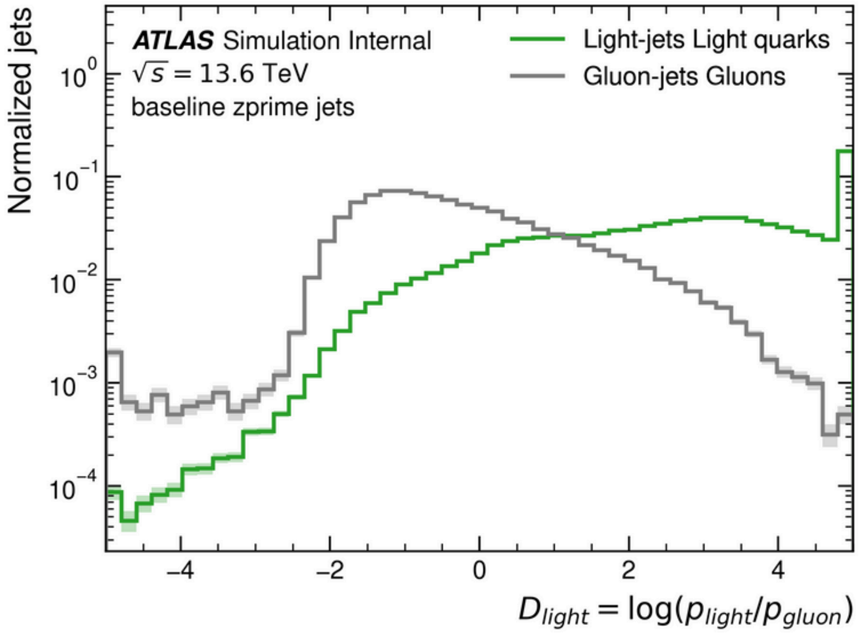
ttbar Samples

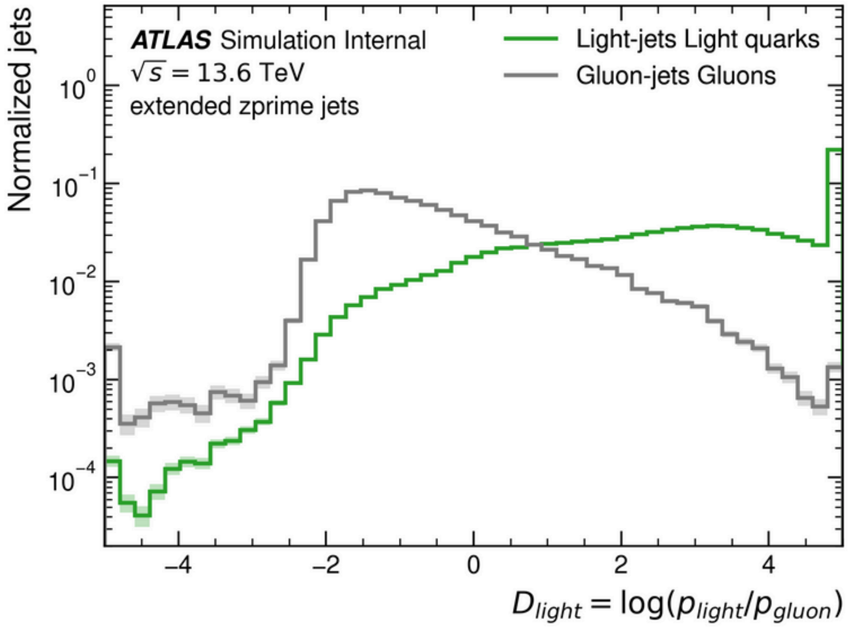


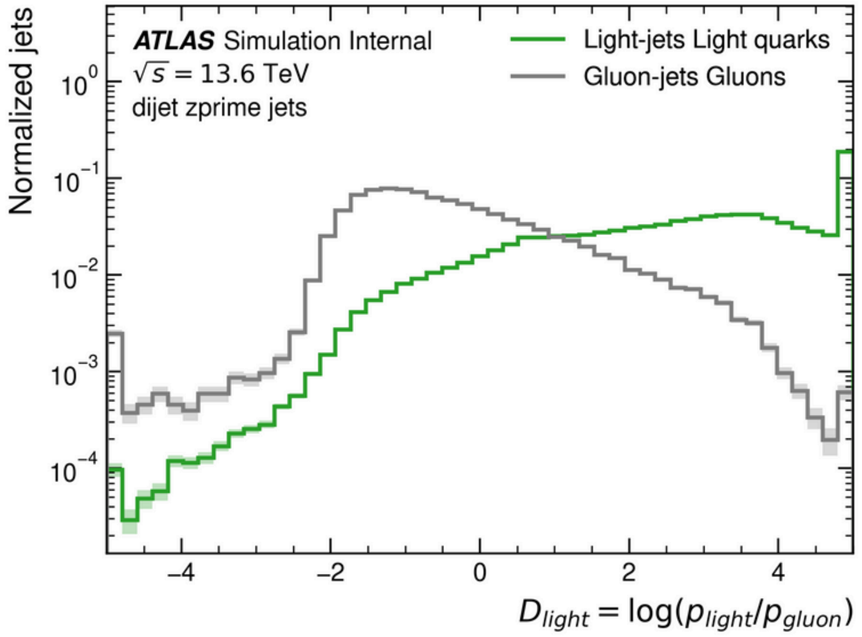




Zprime Samples







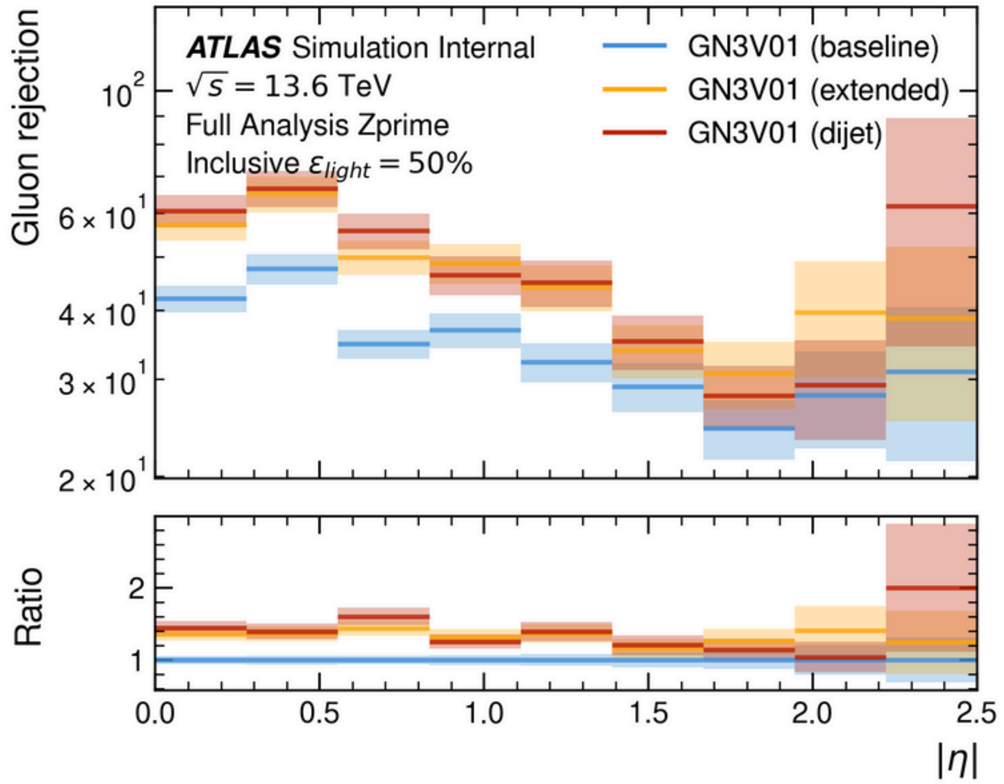
$$D_{\ell/g} = \log\frac{P_\ell}{P_g} = \log\frac{p_{ud}+p_s}{p_{gluon}}$$

quark/gluon-tagging Performance - Central

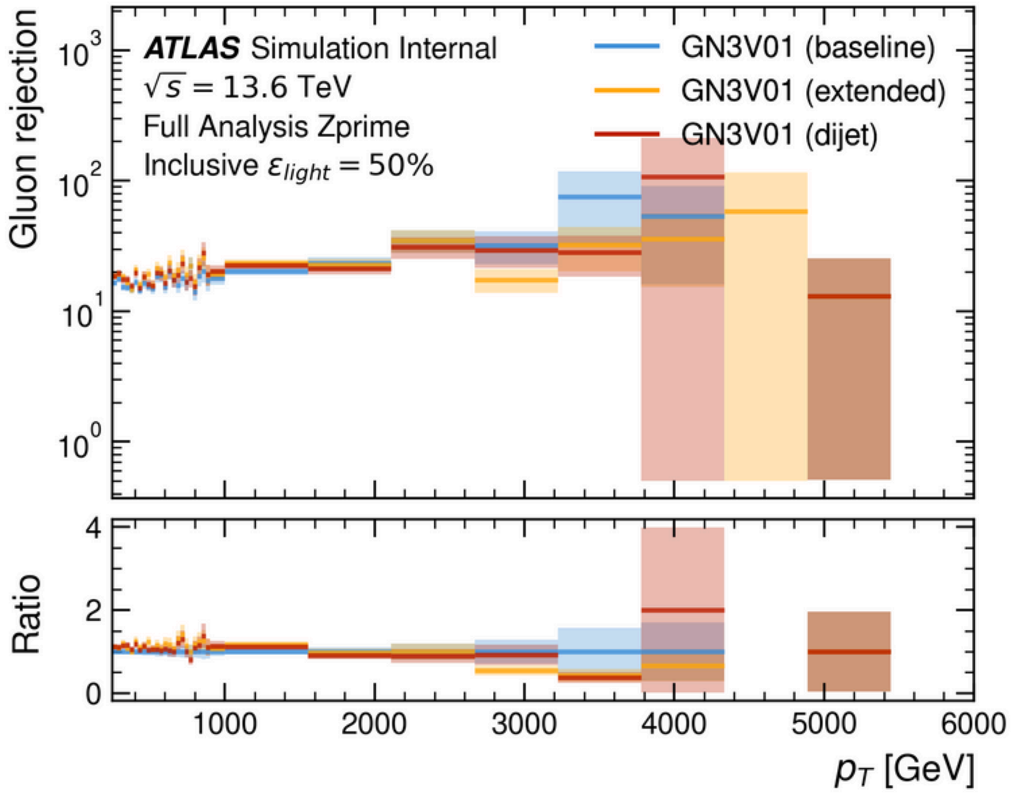
Zprime Samples

Central: |η|<2.5

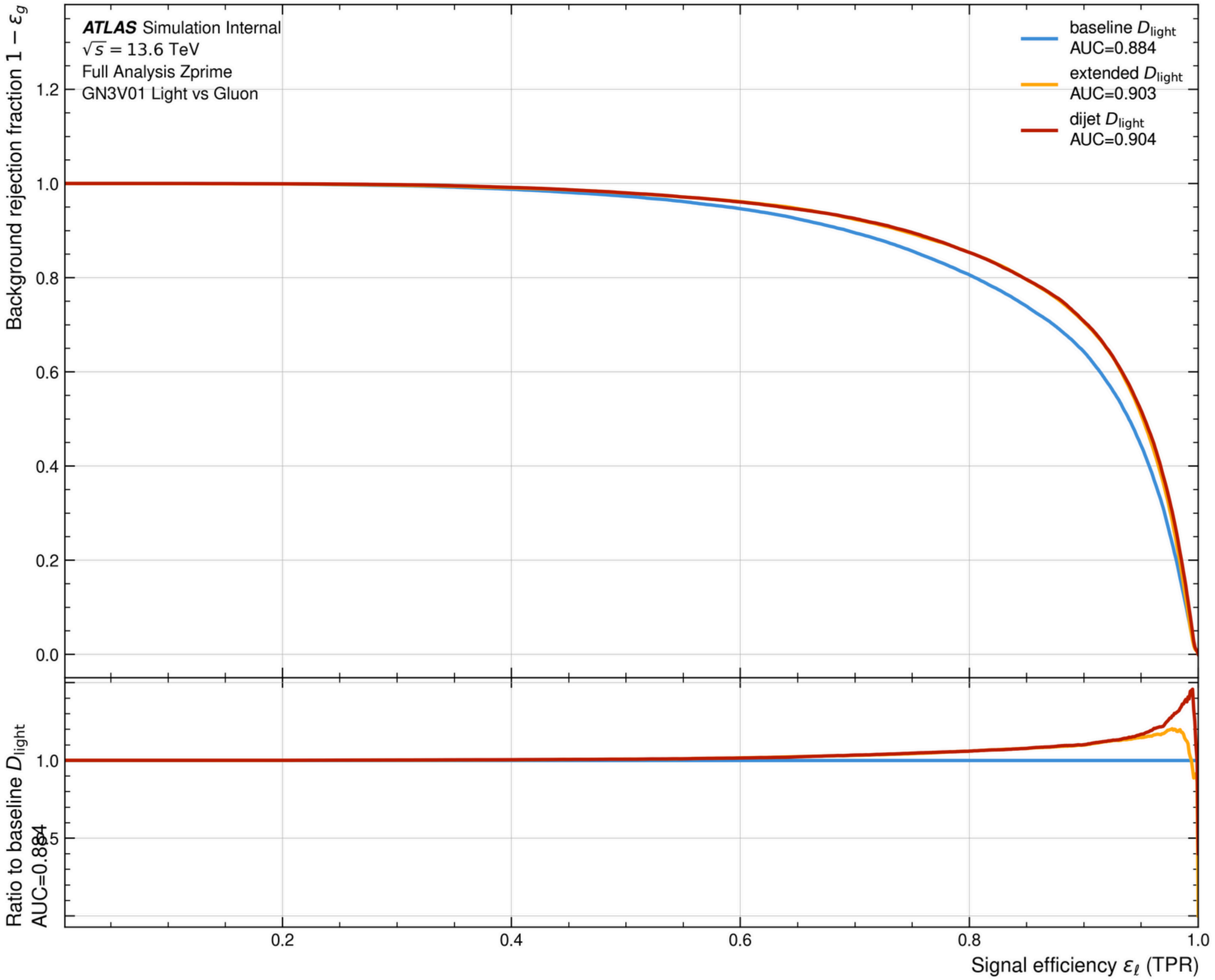
Gluon Rejection vs. Eta



Gluon Rejection vs. pT



ROC Curve



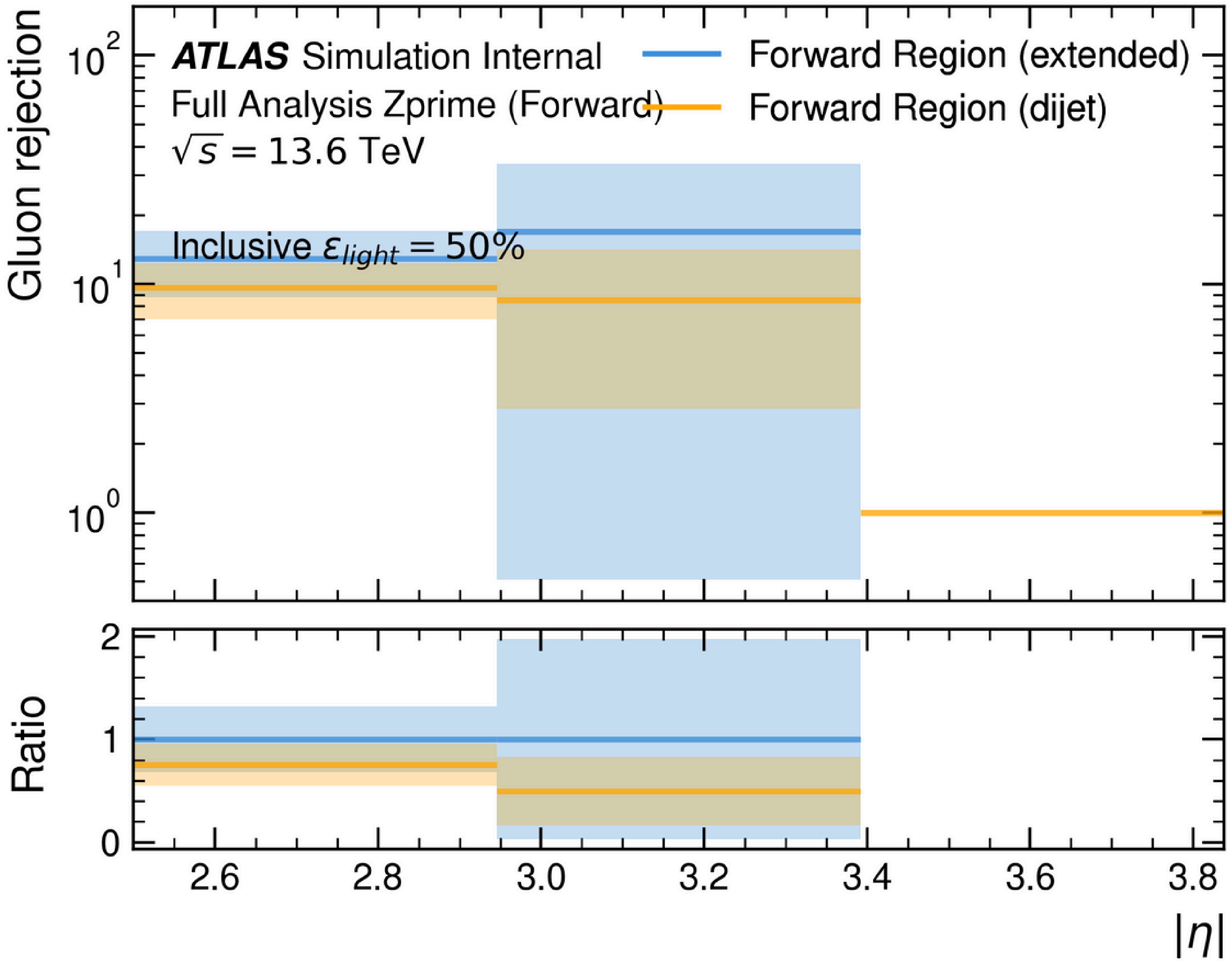
quark/gluon-tagging Performance - Forward

Forward: 2.5 <|η| <4.5

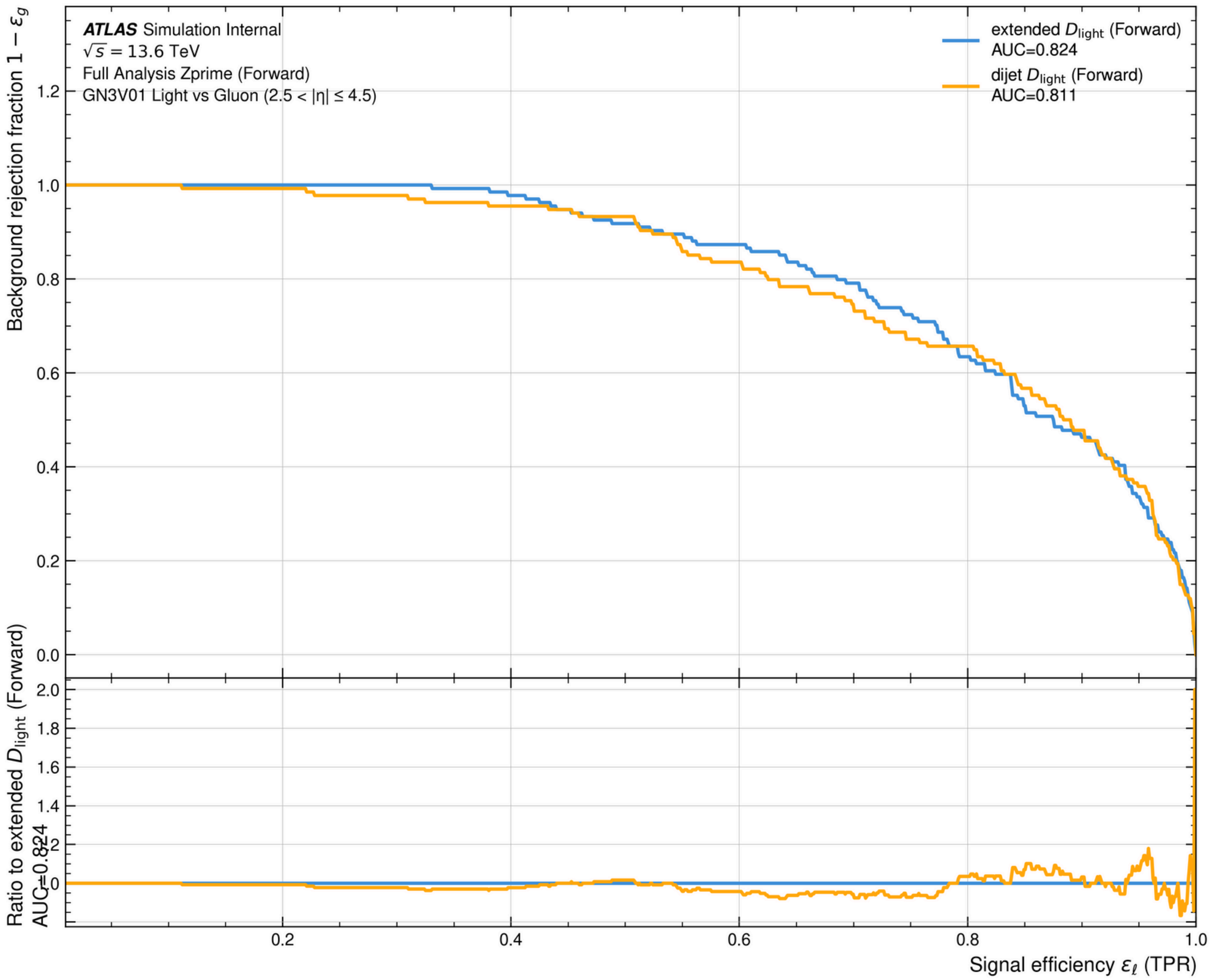
Zprime Samples

ttbar performance results are similar

Gluon Rejection vs. Eta



ROC Curve



High pT!

Conclusion

- The baseline performance of GN3 was first evaluated, confirming strong flavour-tagging and reliable quark/gluon separation.
- The η coverage was extended to include forward jets, with central-region performance remaining stable and showing a slight increase at high pT, highlighting GN3's overall robustness.
- Dijet samples were introduced to further optimize the forward region, but no additional gain was observed beyond the extended setup.
- Overall, the tagging performance remained stable, with quark–gluon discrimination at high pT continuing to be a clear strength.
- This study points to clear opportunities to optimize forward-jet tagging.
- **Next Step:** Evaluating models performance with dijet test samples.

BACKUPS

Datasets

For the training and evaluation part three different sample types are used: ttbar, Zprime and dijet

ttbar (low-pT) Samples:

- Tops decay t → W b → gives two high-purity b-jets
- Extra c- and light-jets from hadronic W decays (W → u d, W → c s)
- Realistic environment: pile-up, underlying event, central low-pT kinematics

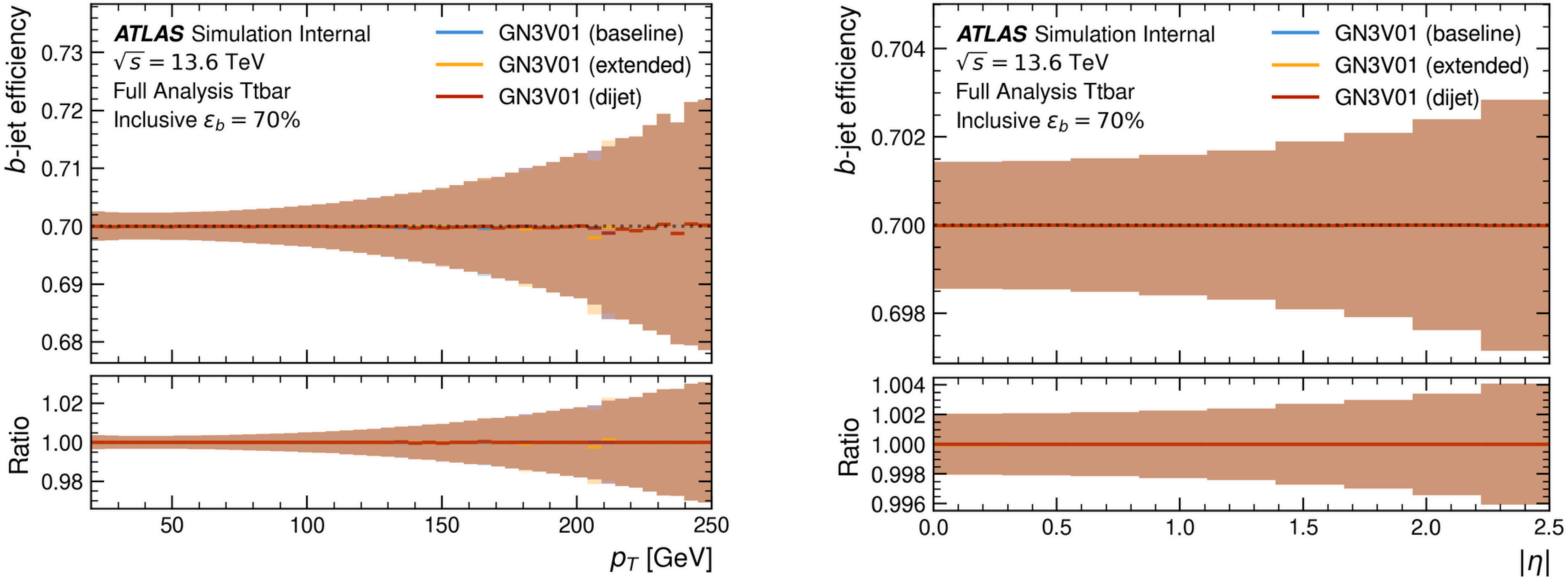
Z′ (high-pT) Samples:

- Z′ → q qbar (includes b bbar, c cbar) → broad flavour coverage
- Probes high-pT where tracks are more collimated and substructure changes
- Complements ttbar to cover the full pT range

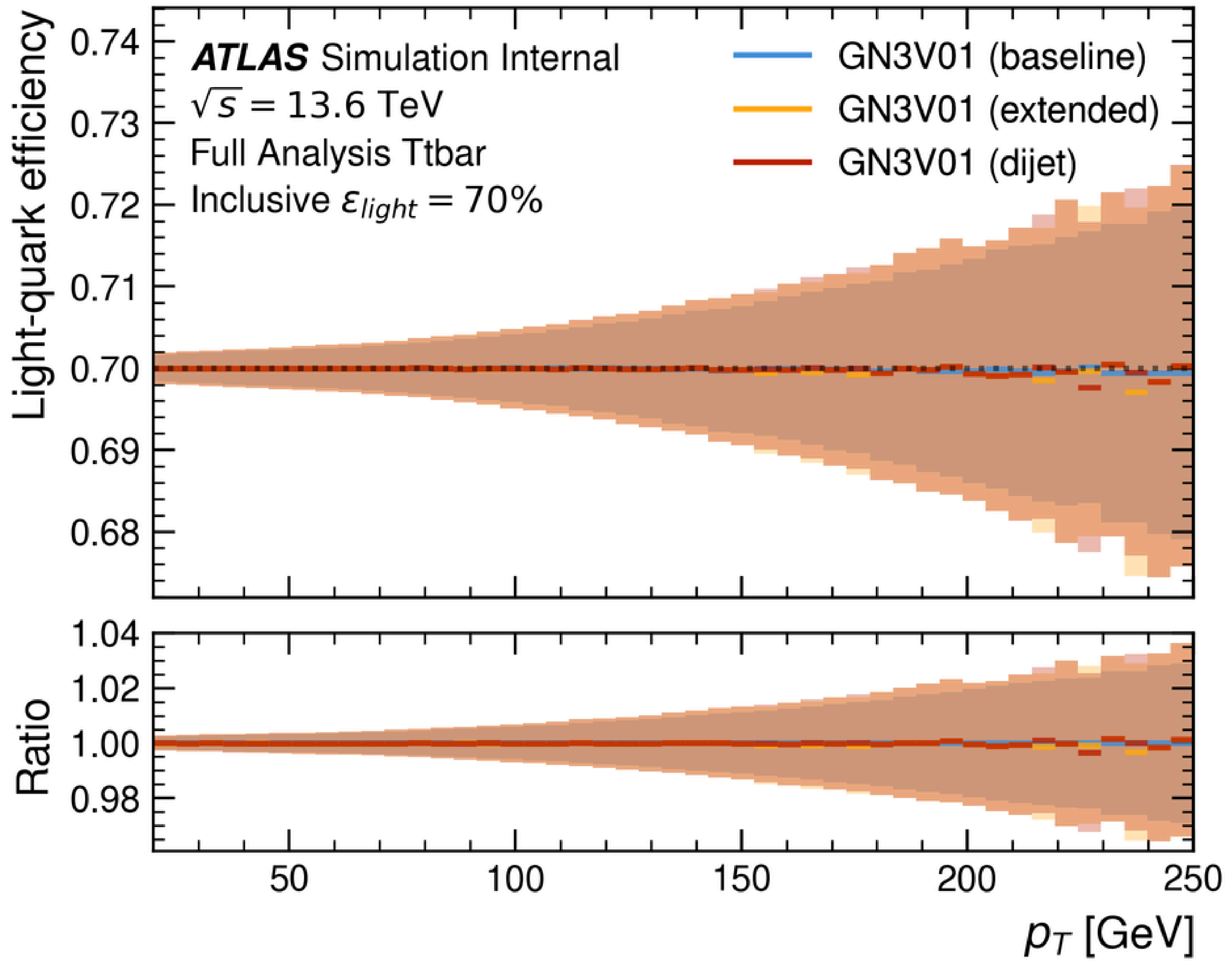
Dijet Samples:

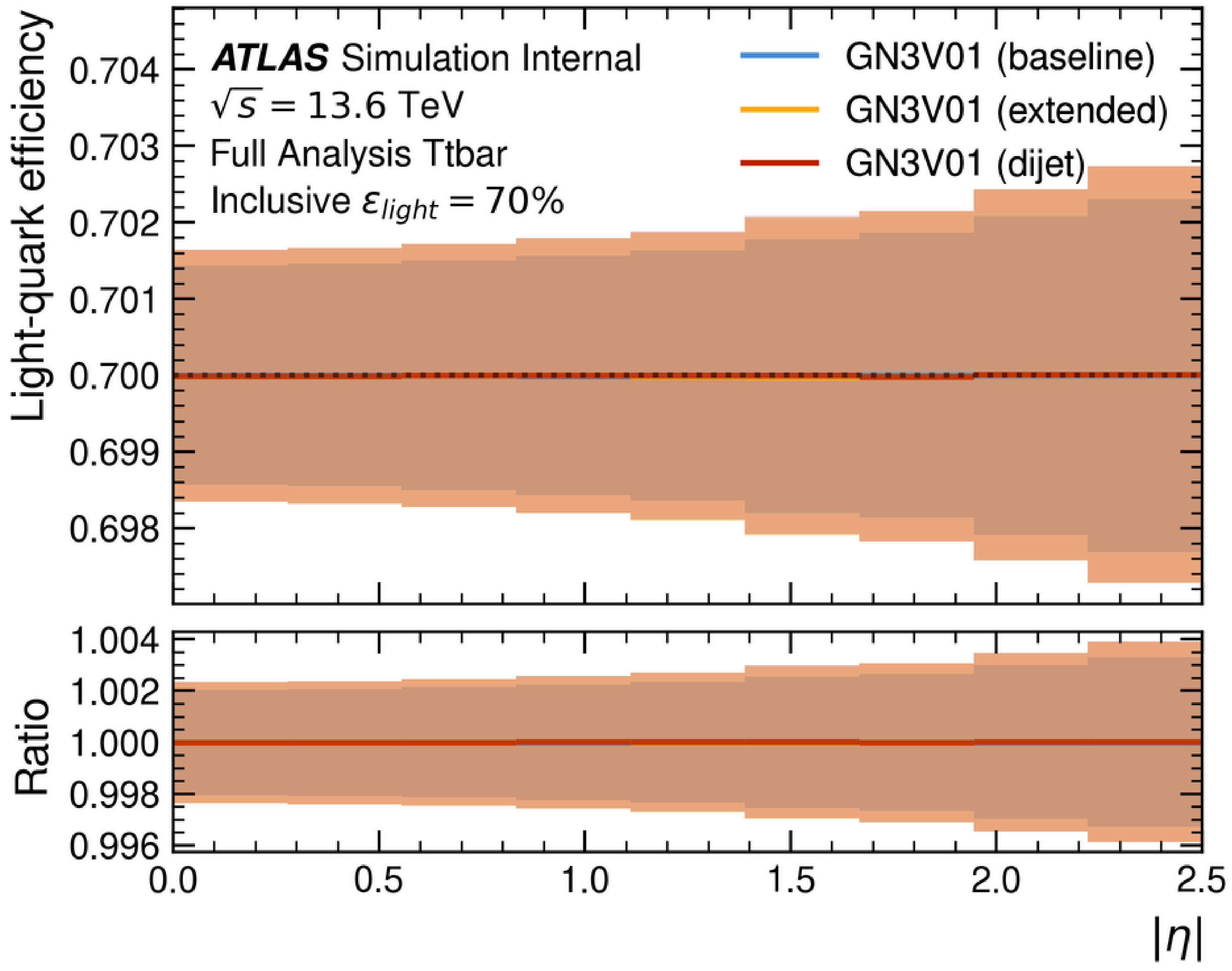
- Typically gluon-dominated → strengthens light-quark vs gluon discrimination
- Forward-heavy (many jets with 2.5 < |eta| < 4.5) → extends coverage beyond central
- Adds statistics at phase-space edges → better robustness and generalisation

Fixed 70% b-jet Efficiency at each bin!



Fixed 70% Light-quark Efficiency at each bin!

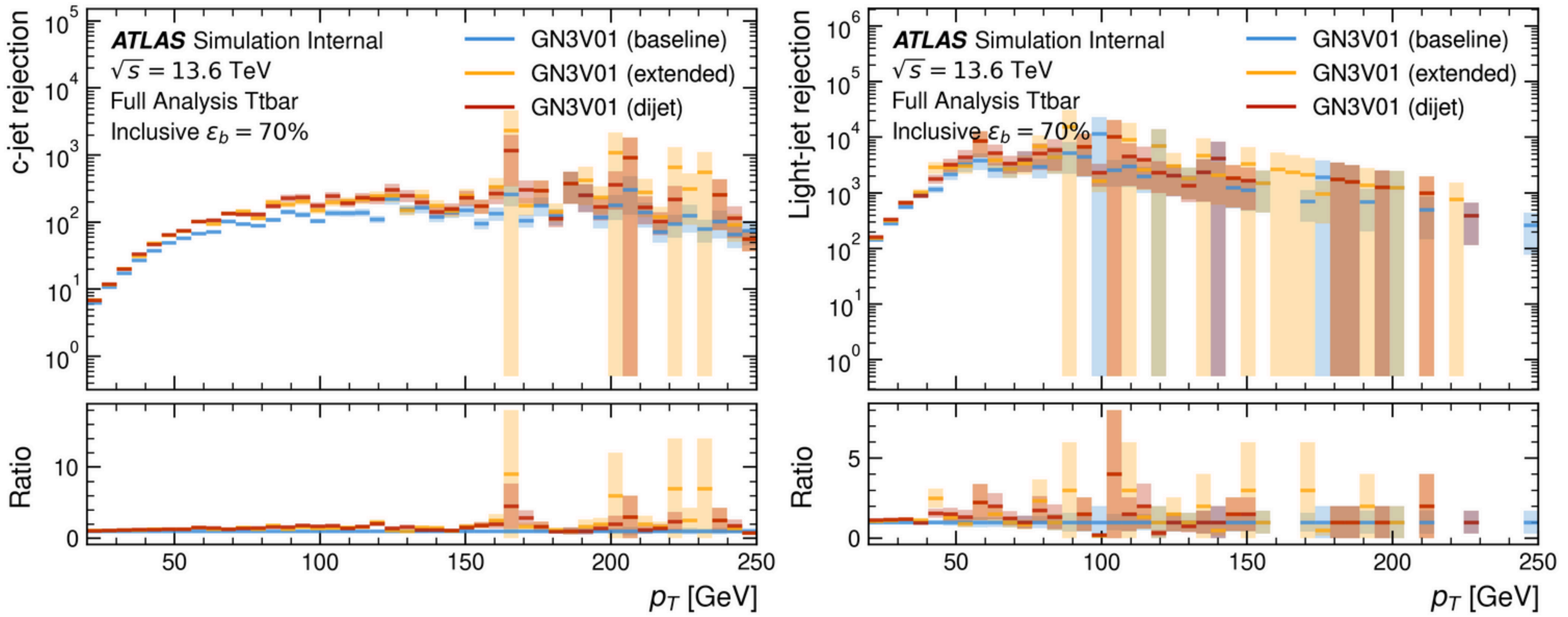




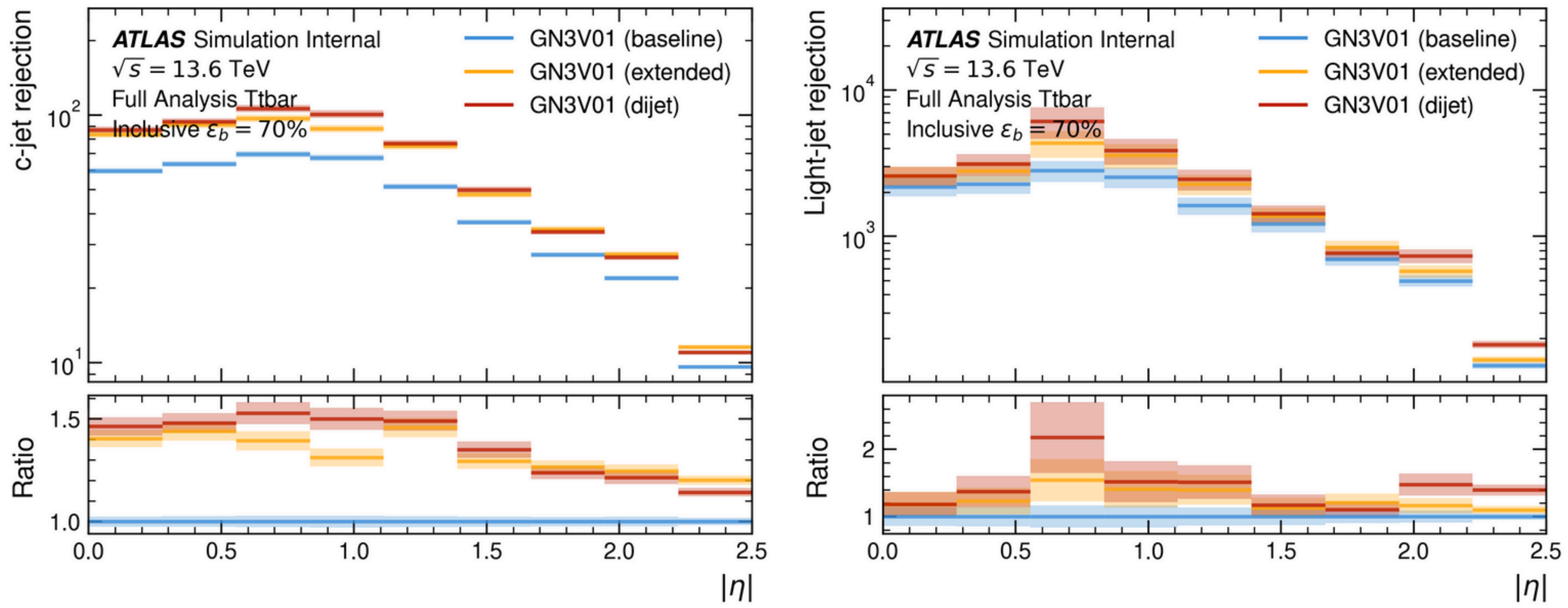
b-tagging Efficiency/Rejection Plots

c-jet and light-jet Rejection vs. pT at %70 Light-Jet Efficiency - Central Region

ttbar Samples

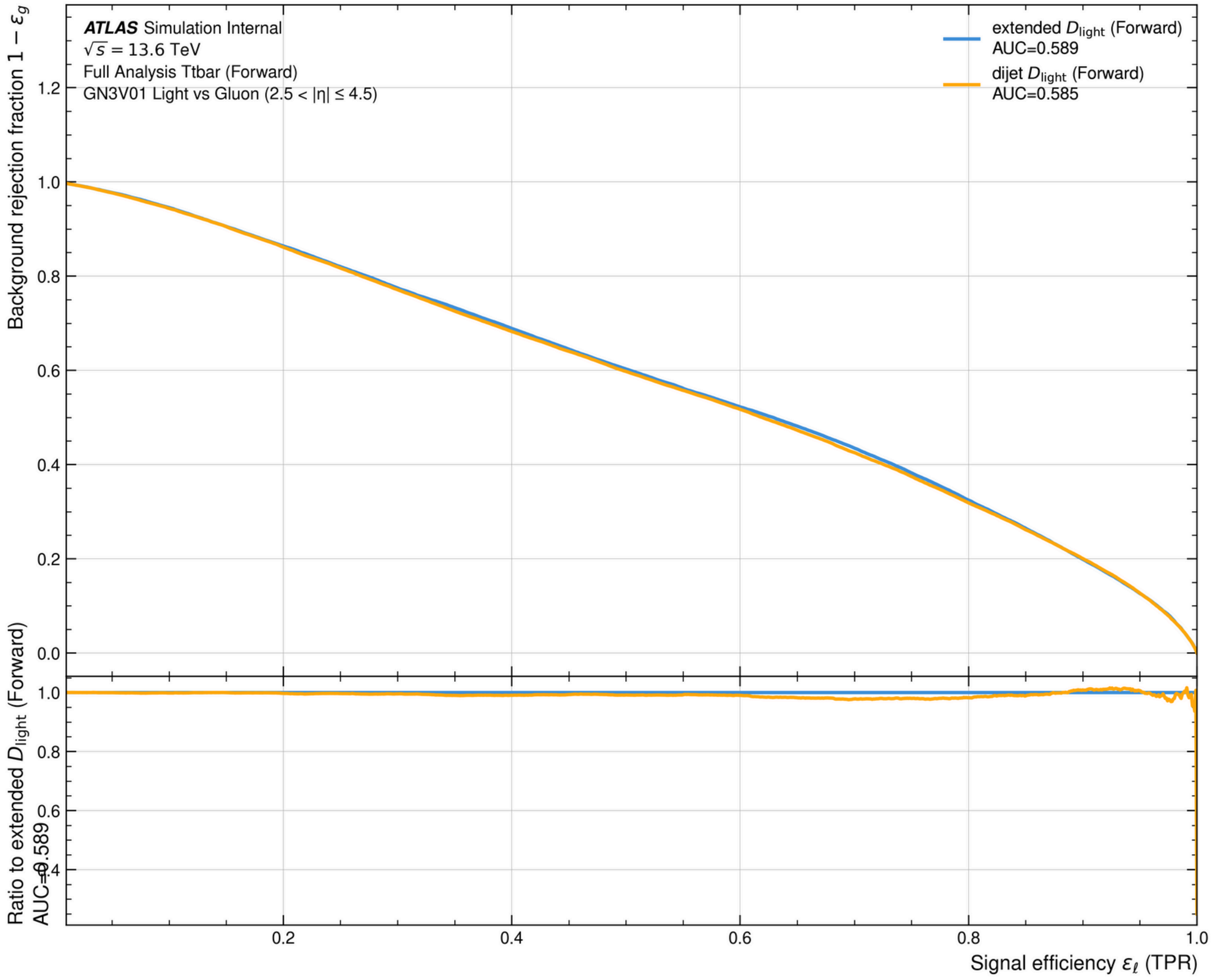


ttbar Samples



light/quark-tagging ROC Curves - Forward

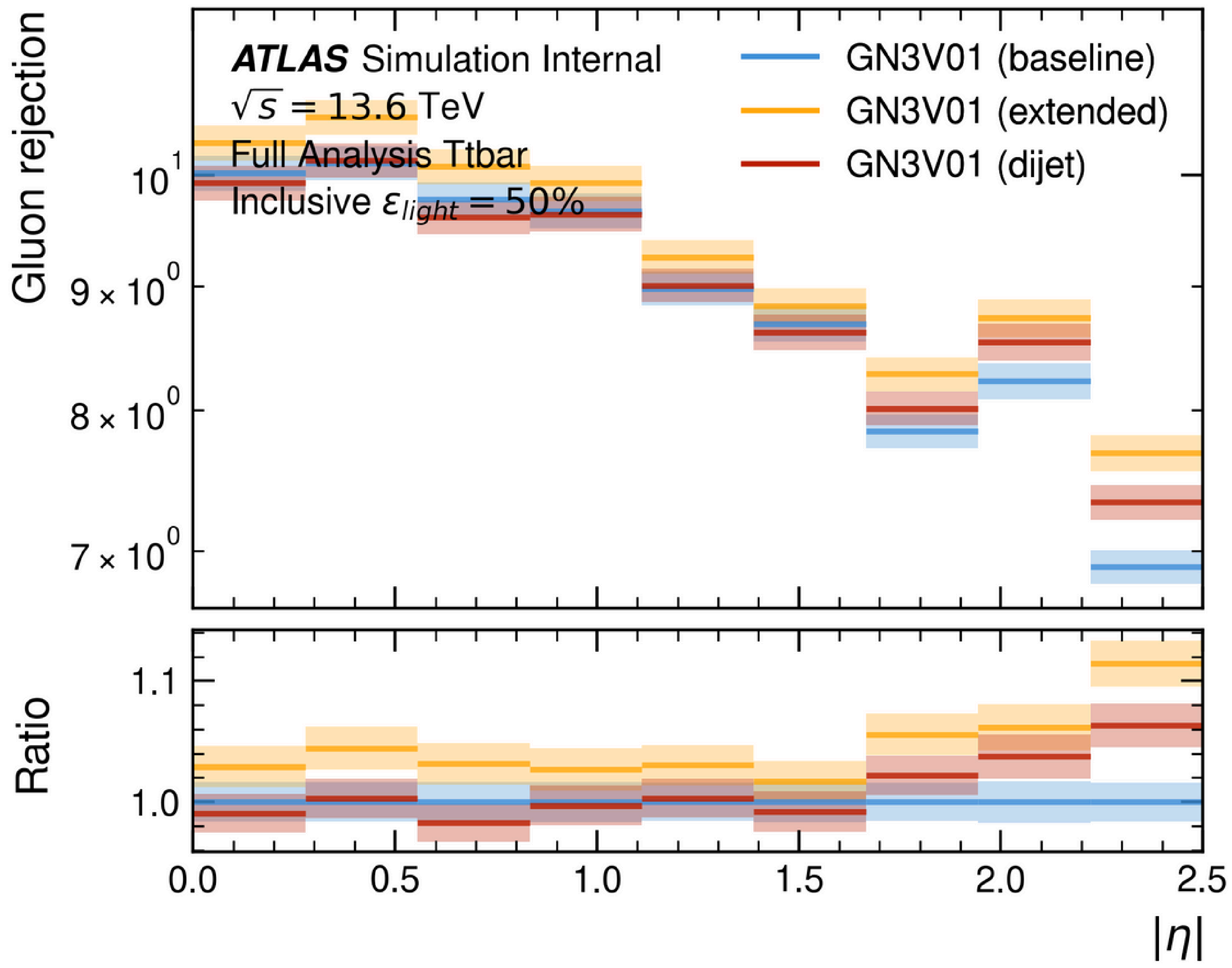
ttbar Samples



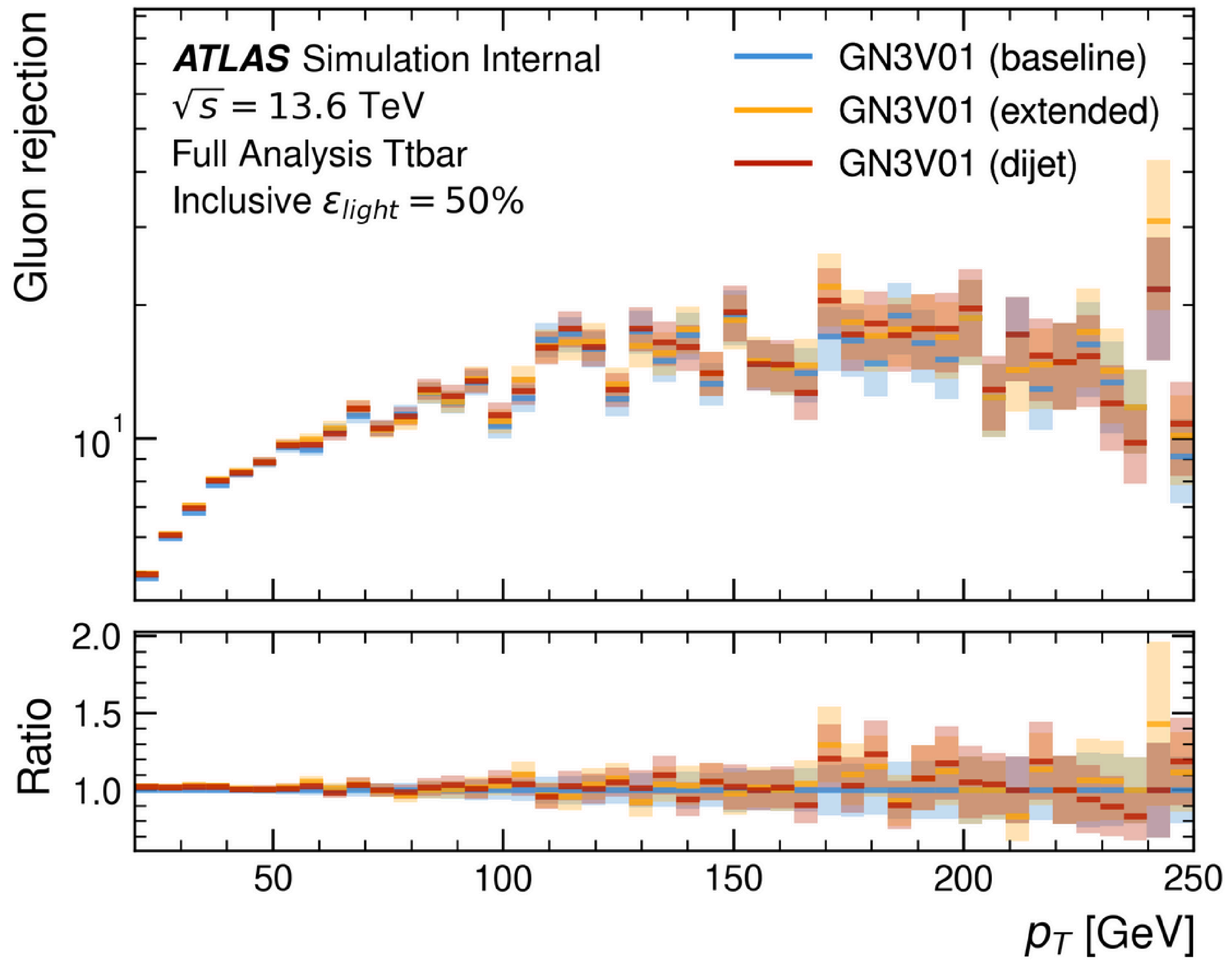
b-tagging Efficiency/Rejection Plots

Gluon Rejection vs. Eta at %50 Light-Jet Efficiency - Central Region - ttbar

vs. eta



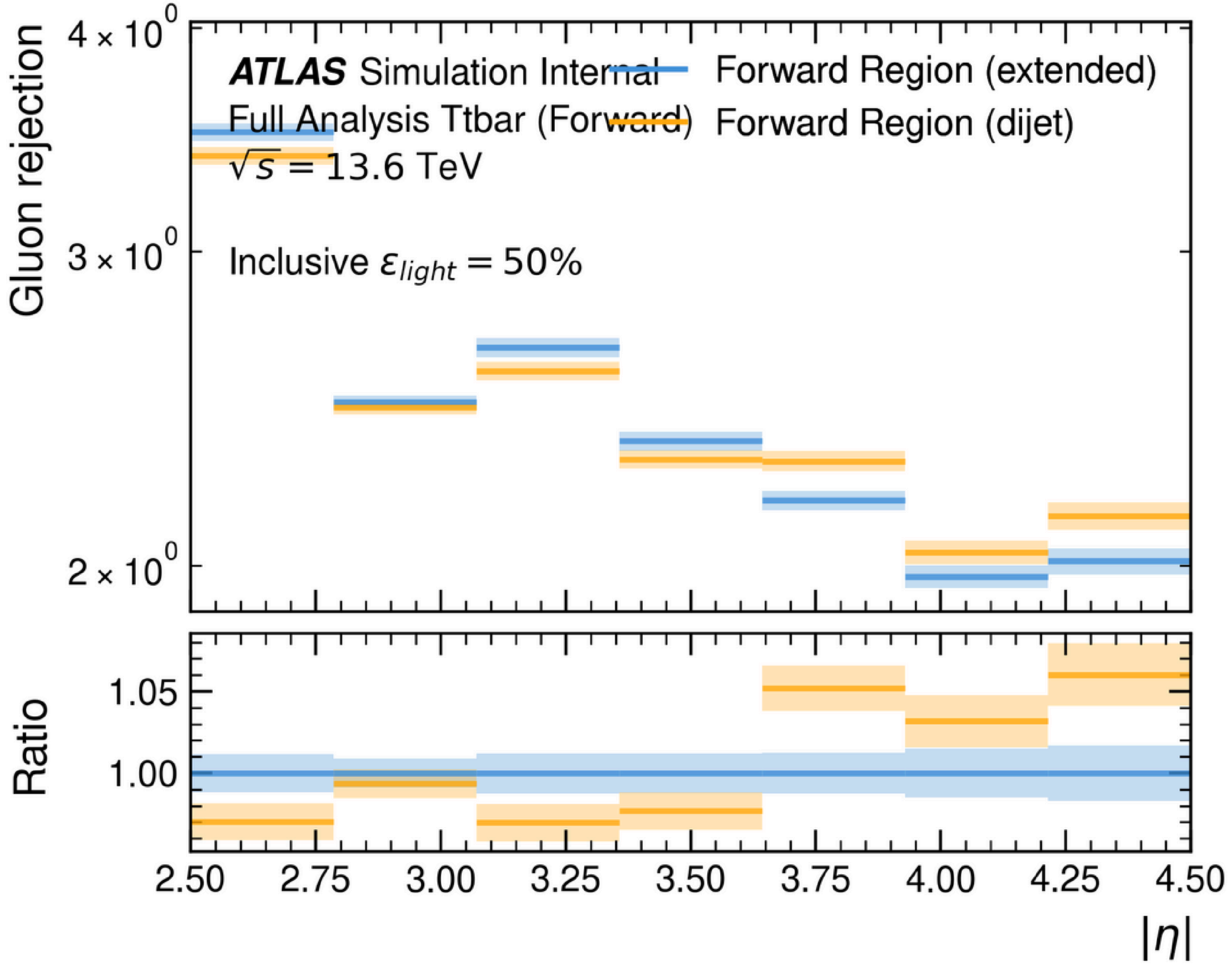
vs. pT



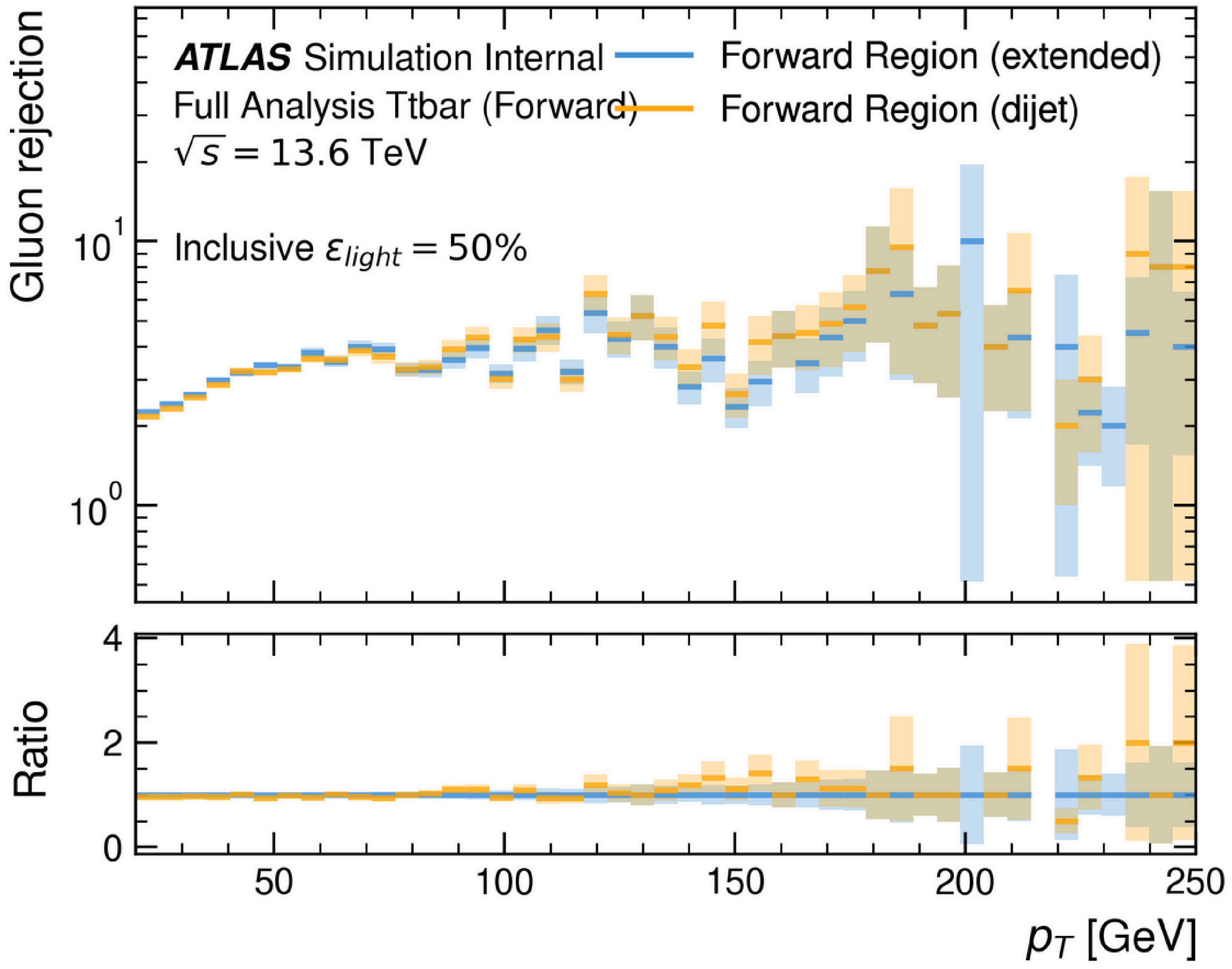
b-tagging Efficiency/Rejection Plots

Gluon Rejection at %50 Light-Jet Efficiency - Forward Region - ttbar

vs. Eta

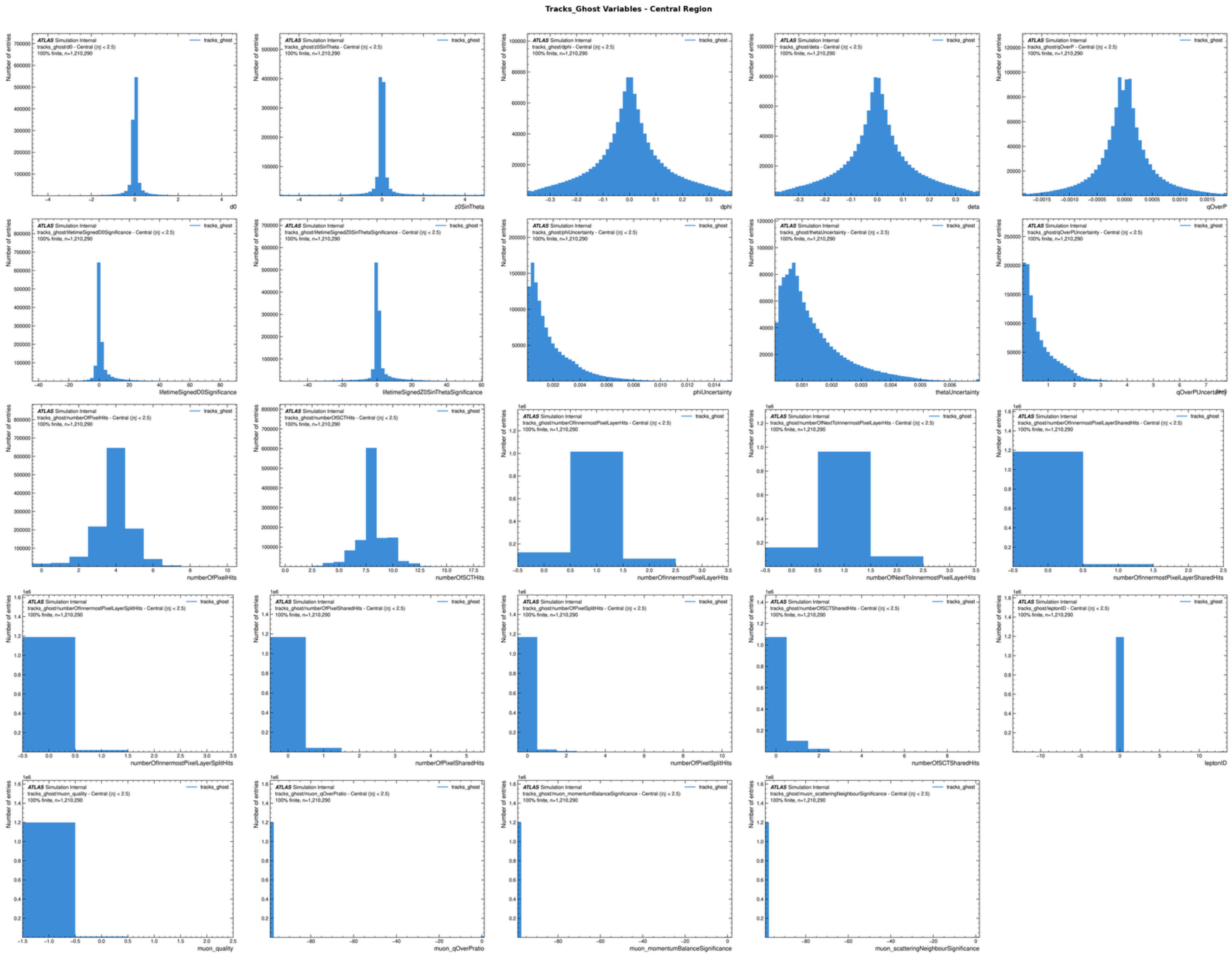


vs. pT

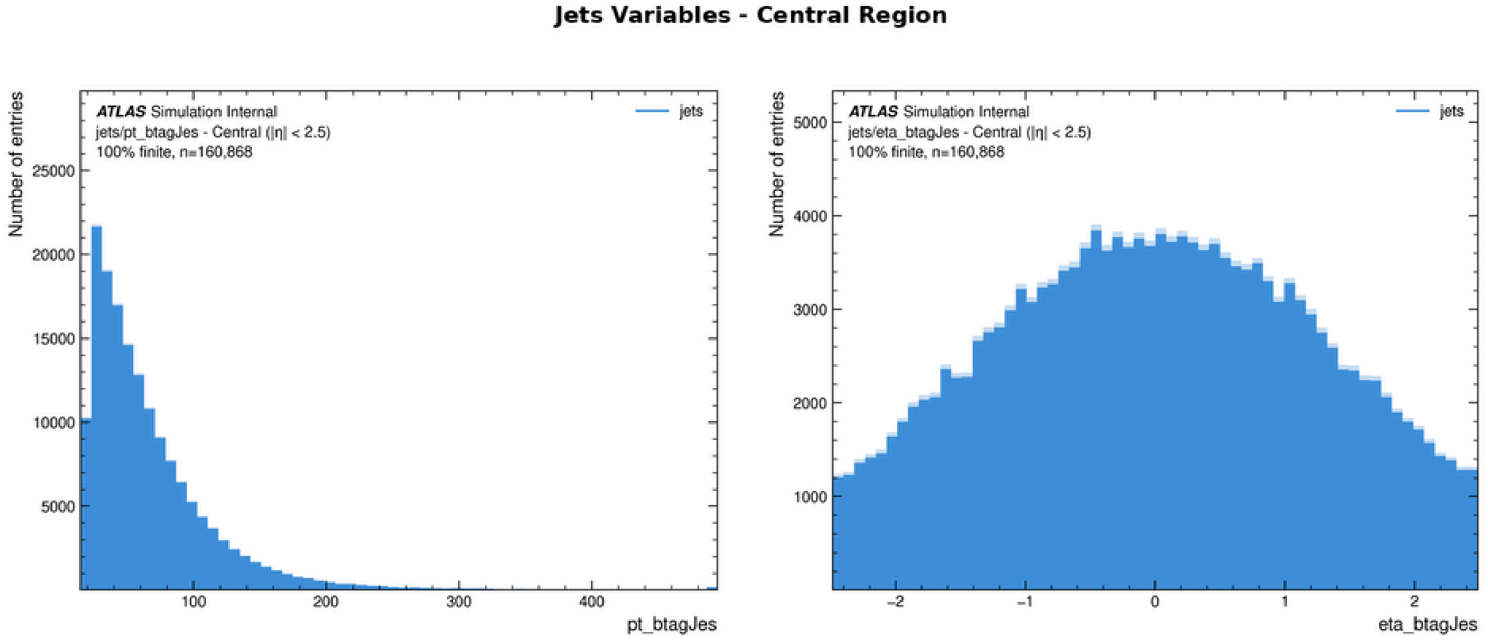


Input Variables - Central Eta Region

Track Variables

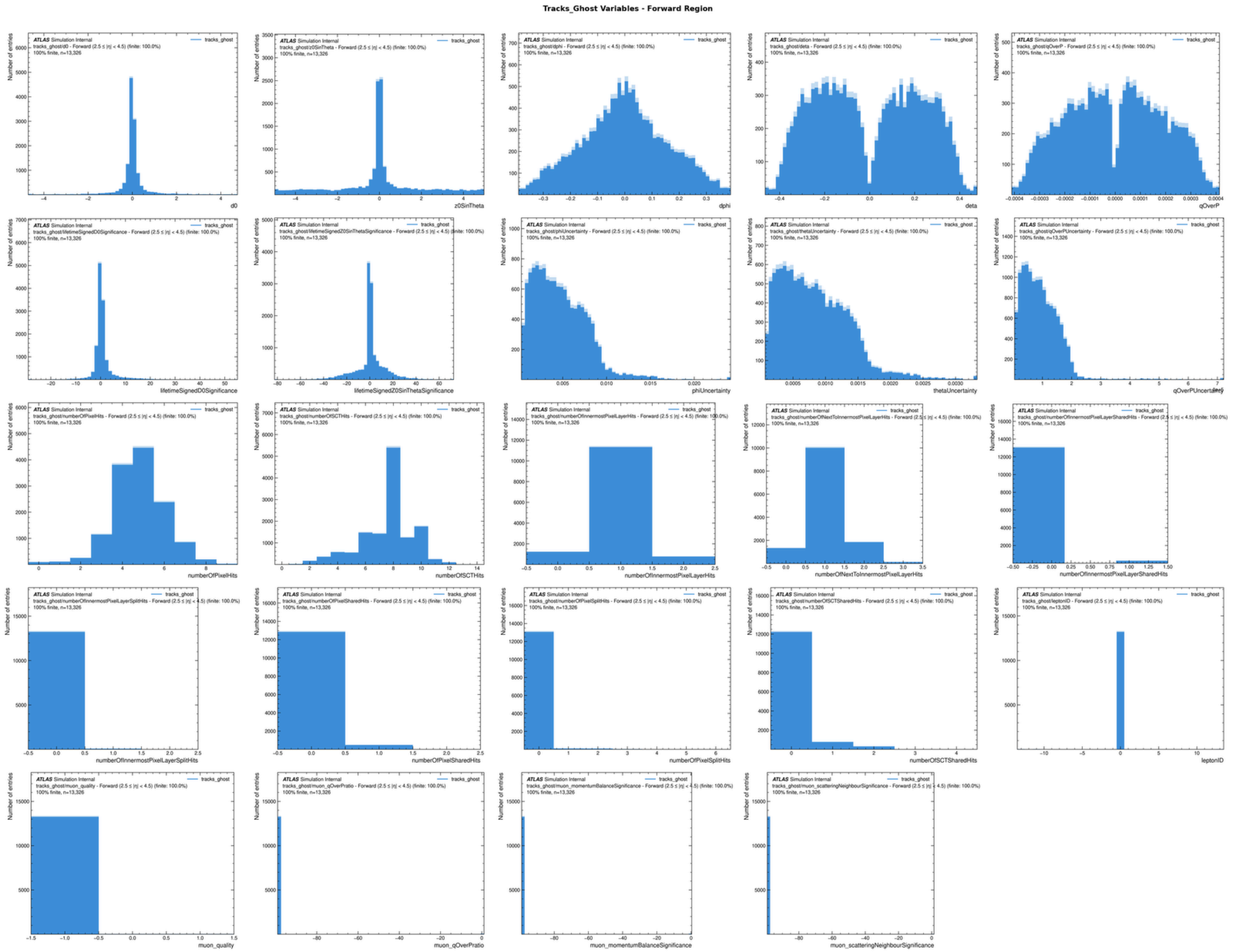


Jet Variables

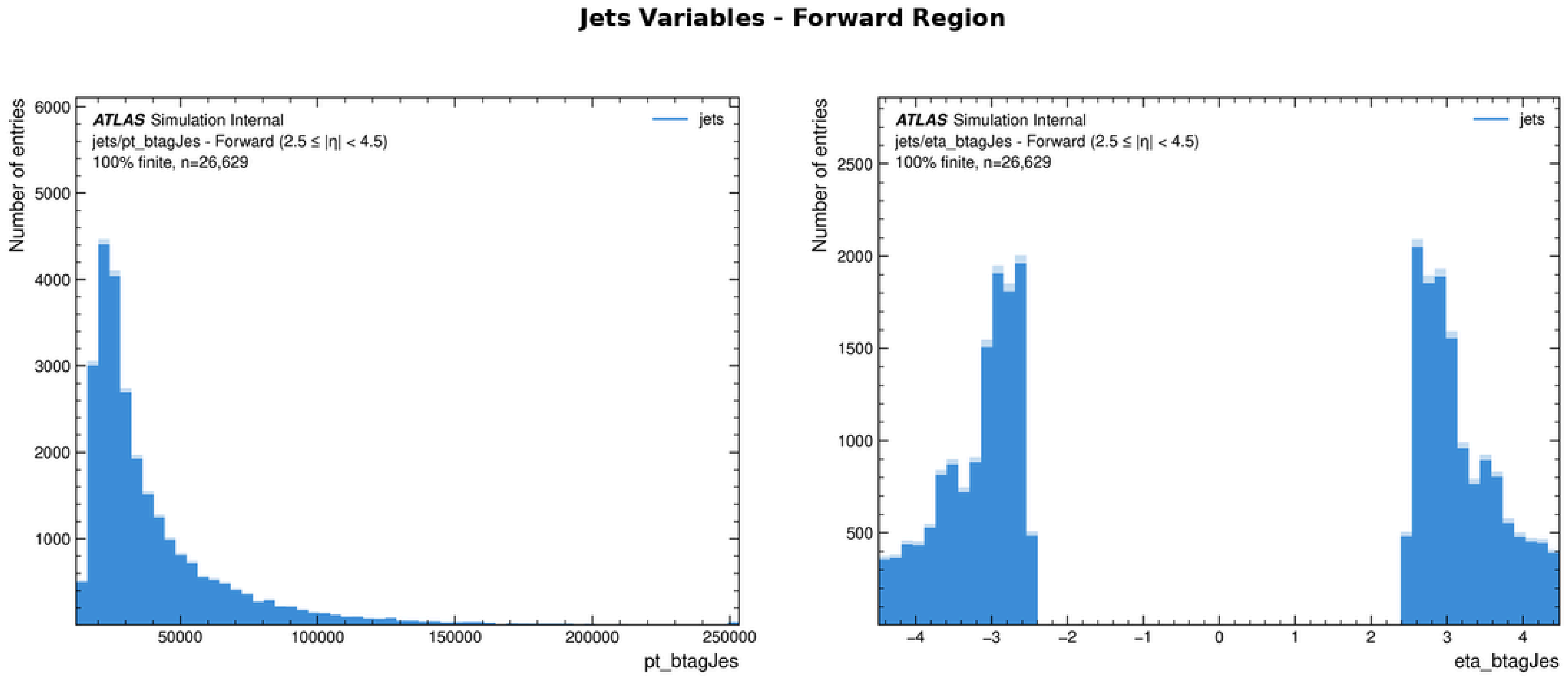


Input Variables - Forward Eta Region

Track Variables

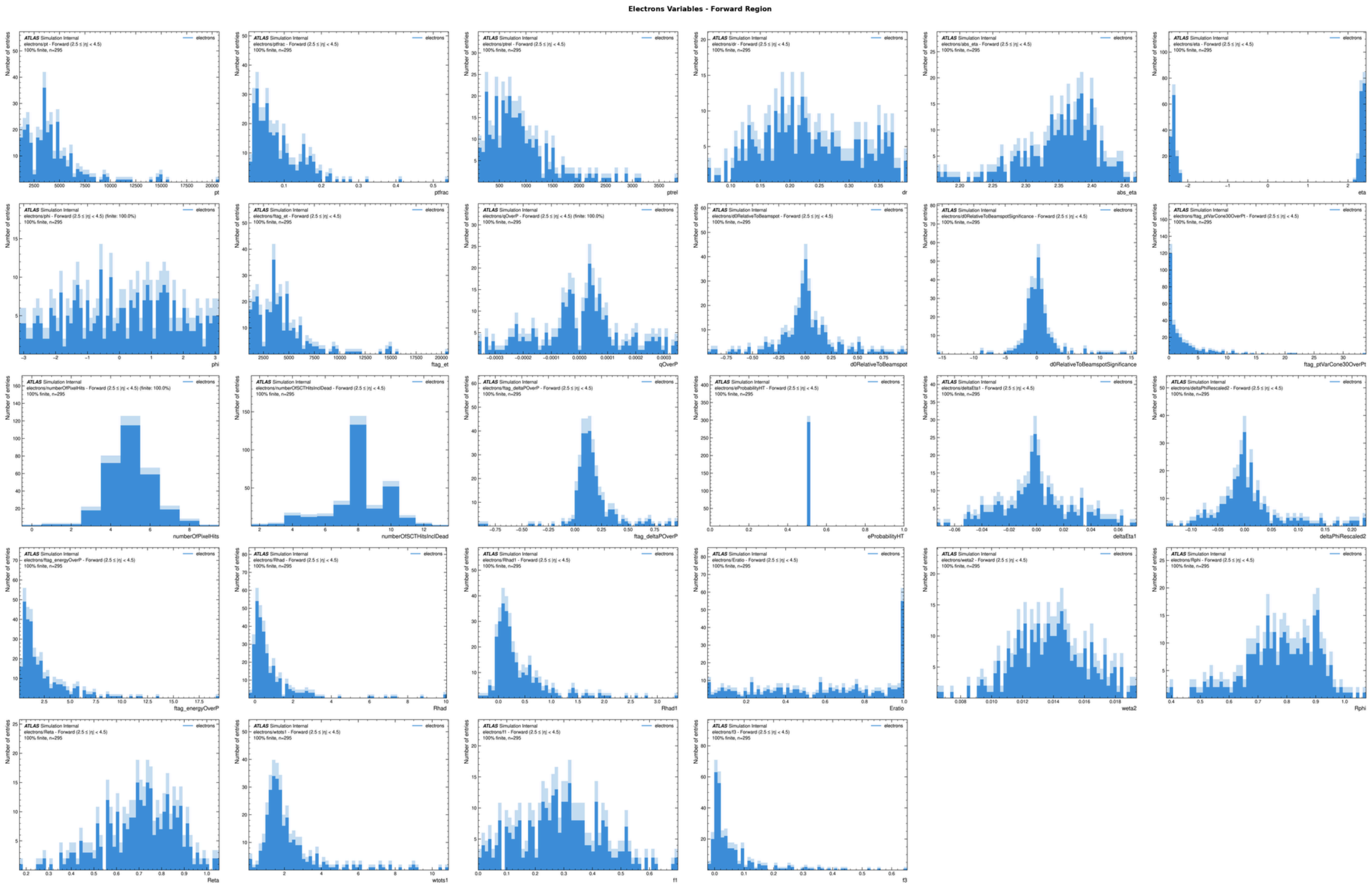


Jet Variables

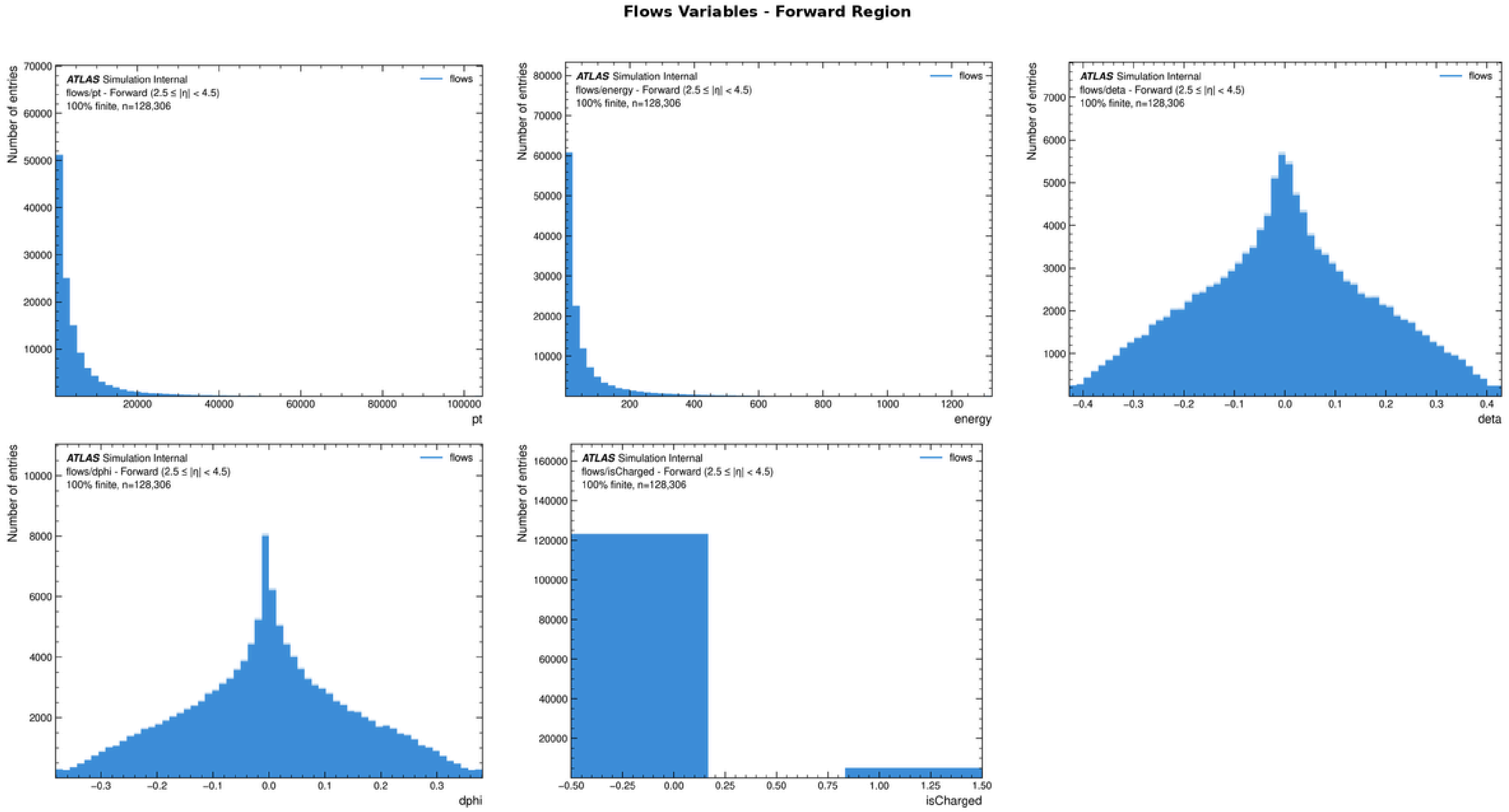


Input Variables - Forward Eta Region

Electron Variables

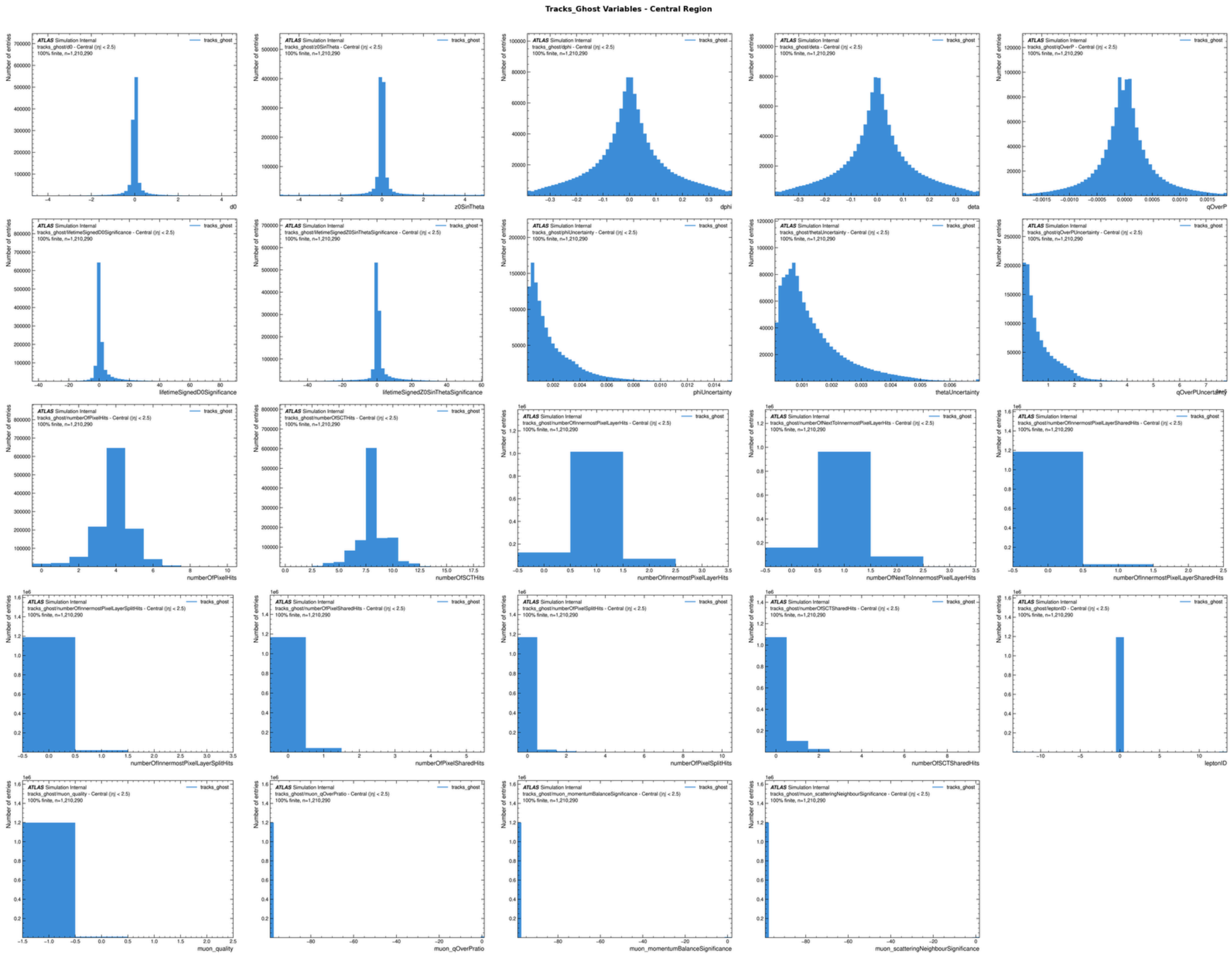


Flow Variables

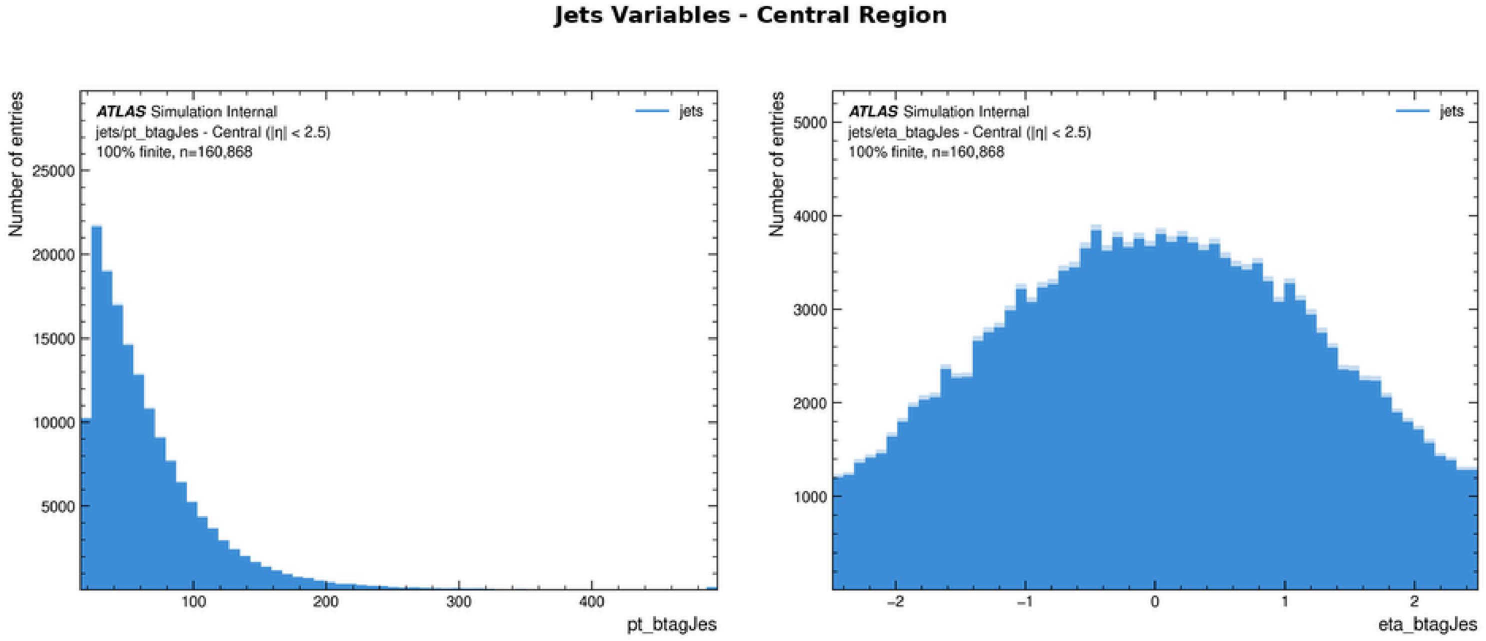


Input Variables - Central Eta Region

Track Variables

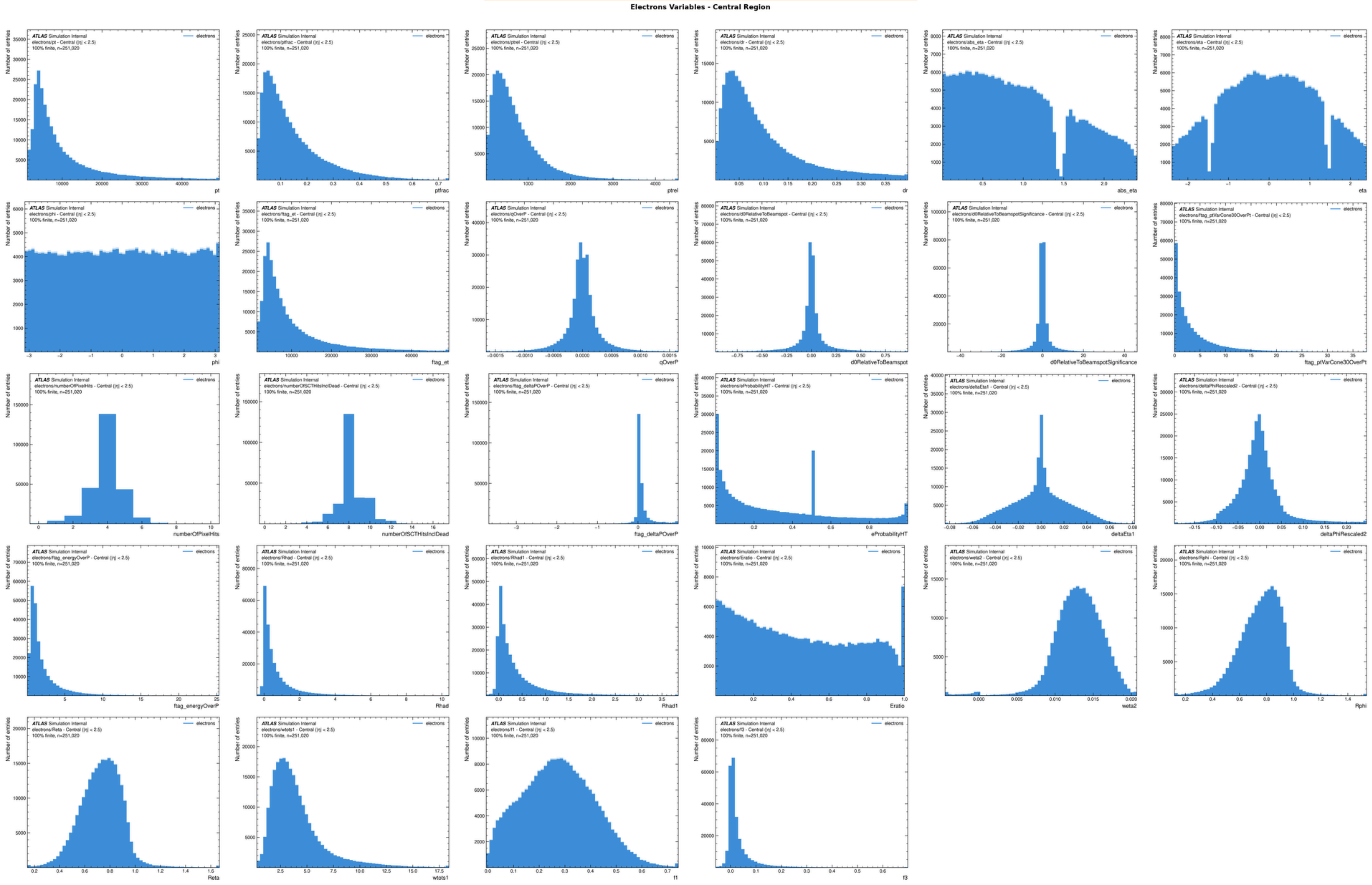


Jet Variables



Input Variables - Central Eta Region

Electron Variables



Flow Variables