

Brookhaven National Laboratory

U.S. DEPARTMENT of ENERGY

CMS Tokenization and ATLAS-to-CMS Transfer

Merve Nazlim Agaras

November 4, 2026

@BrookhavenLab

CMS simulation samples

2016 UL NanoAODv9 simulation from CERN Open Data

Process	CERN record	Events after cleaning
DYJetsToLL (mℓℓ > 50 GeV)	35669	2,519,857
Dileptonic top pair	67801	1,369,067
Semileptonic top pair	67993	1,399,821
GluGluHToTauTau	37444	51,057
EM-enriched QCD, 80–120 GeV	63236	422,535
Diphoton + jets (mγγ ≥ 80 GeV)	34121	140,495

5.90 million entries after event cleaning.

Event cleaning and object selection

Object	pT > (GeV)	|η| <	Identification and extra requirements
Electron	10	2.47	Loose Fall17V2 MVA without isolation
Muon	10	2.50	Loose muon ID
Photon	15	2.37	Loose cutBased ID, electron veto, gap veto
Tau	20	2.50	Loose DeepTau VSjet, |charge| = 1, gap veto
Jet	20	4.50	Tight jet ID, nonnegative mass

Photon and tau gap veto: |η| between 1.37 and 1.52. Removed 283 jets with negative mass.

No tracks or PF candidates

CMS-native tokenizer inputs

All objects use pT, η and ϕ. Electrons, muons and taus also use charge.

Object	Additional CMS feature names	Raw inputs
Electron	mvaFall17V2Iso, cutBased, miniPFRelIso_chg, pfRelIso03_all	8
Muon	mediumId, tightId, miniPFRelIso_chg, pfRelIso03_all	8
Photon	cutBased, mvaID, pfRelIso03_chg, pfRelIso03_all	7
Tau	decayMode, rawDeepTau2017v2p1VSjet, idDeepTau2017v2p1VSe, idDeepTau2017v2p1VSmu	8
Jet	mass, nConstituents, btagDeepFlavB, btagDeepFlavCvL, btagDeepFlavCvB, btagDeepFlavQG	9

Training five CMS VQ-VAEs

Object	Train	Validation	Test
Electron	2,411,908	134,575	133,280
Muon	2,673,325	148,568	148,269
Photon	420,528	23,495	23,528
Tau	3,273,825	182,254	181,296
Jet	19,808,746	1,102,639	1,099,235

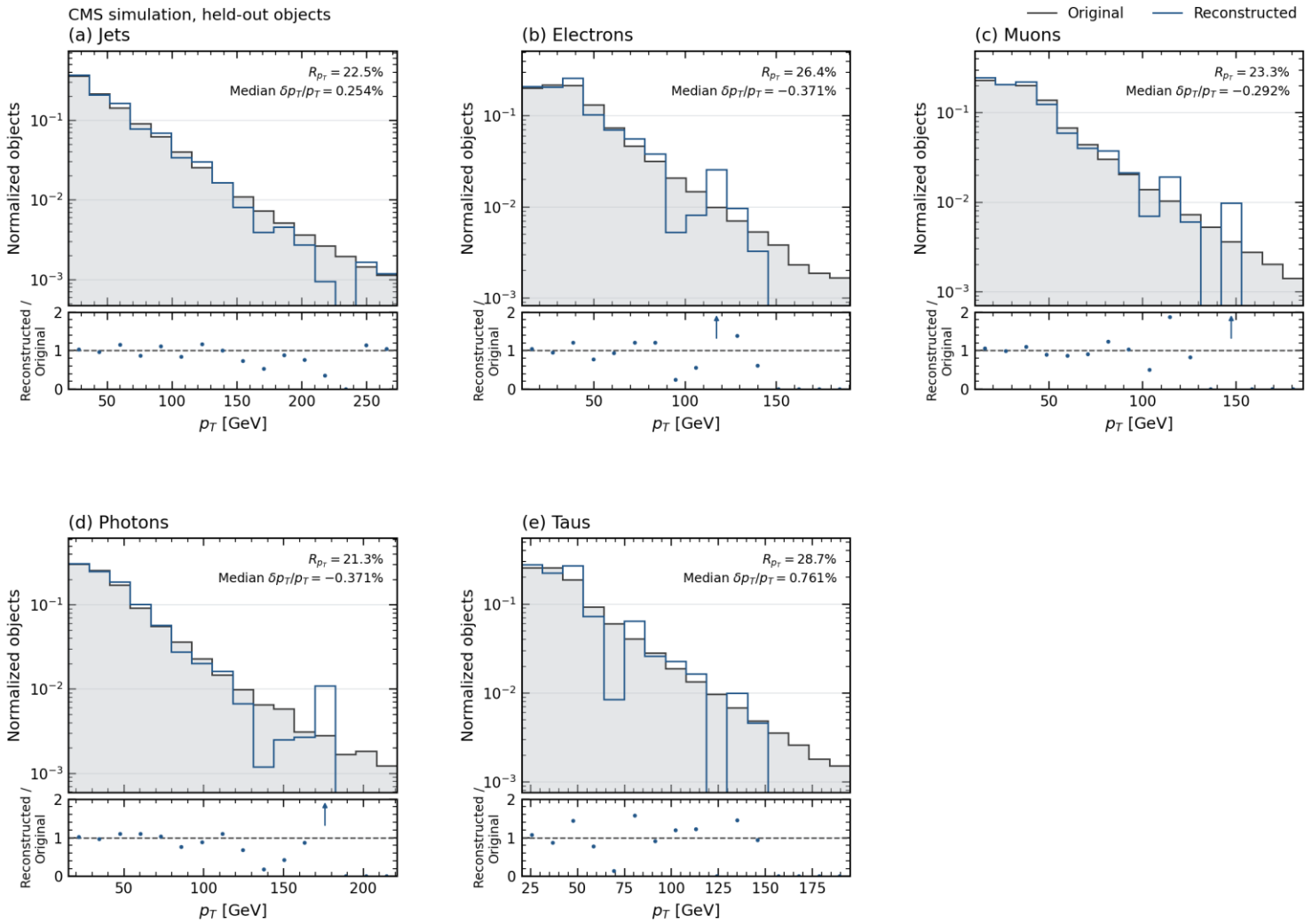
One code per object

16,384 codes, latent dimension 8. MLP widths 128/256/512 with LayerNorm and ReLU. MAE + commitment loss.

Training and selection

20 epochs, batch 1,024, seed 42. Adam, peak LR 10^{-3}, warm-up and cosine decay. Train-only preprocessing. Select by validation loss.

CMS transverse momentum



Small median bias

Median fractional pT residual stays below 1% in for all five collections.

Object-level spread

Widths range from 21% to 29%.

More plots in backup

Brookhaven National Laboratory

Common pT metric, different samples

Object	ATLAS pT width	CMS pT width
Electron	32.63%	26.35%
Muon	33.57%	23.33%
Photon	31.16%	21.33%
Tau	28.43%	28.69%
Jet	27.52%	22.51%

Width = 100 × IQR(pT,reco / pT,orig) / median(pT,reco / pT,orig).

CMS four-lepton transfer dataset

CMS UL2016 simulation: ggF H → ZZ* → 4ℓ versus continuum ZZ → 4ℓ.

Reconstructed event selection

At least four loose e/μ with pT > 10 GeV (leading pT > 20 GeV).
Electron $|\eta| < 2.47$ and muon $|\eta| < 2.5$.

Two disjoint same-flavour, opposite-sign lepton pairs.
$40 < m(Z_1) < 120$ GeV and $12 < m(Z_2) < 120$ GeV.
Four-lepton mass > 70 GeV. No narrow Higgs mass window.

File-disjoint split	ggF H → ZZ* → 4ℓ	Continuum ZZ → 4ℓ	Total
Train	70,414	40,944	111,358
Validation	15,803	8,382	24,185
Test	16,483	8,441	24,924

What transfers from ATLAS

The two transformers differ only in how the outer encoder and its final normalization are initialized.

Model	CMS inputs	Initialization
ATLAS initialization	Frozen CMS tokenizers	ATLAS encoder and final normalization
Scratch transformer	Same frozen CMS tokenizers	Random encoder and final normalization
Raw-feature MLP	Reconstructed event observables	Random MLP (no tokens or pretraining)

Transformer setup

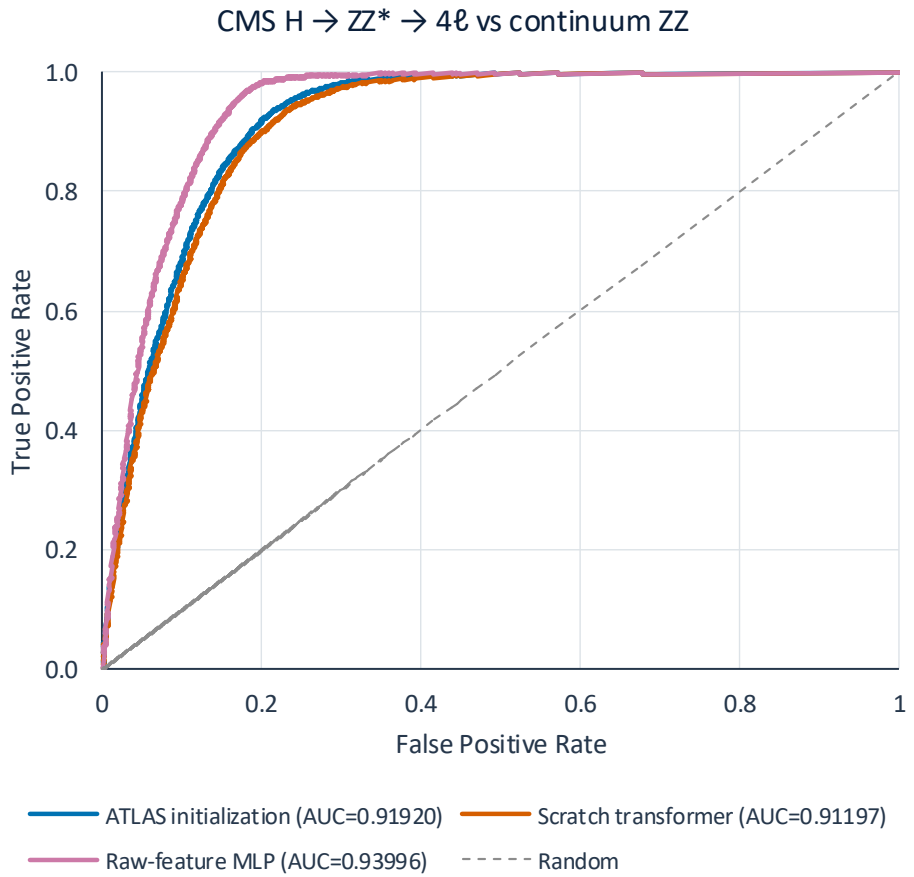
Four layers, eight heads, width 256. One code per object from five frozen CMS tokenizers. Both arms learn CMS embeddings and update the encoder.

Shared training

Same batches and event splits. 10 epochs, batch 256, AdamW at 10^{-4}

Results

All three models use the same 24,924 test events



The new MLP used all five collections' per-object features, with **no tokenizer or pretraining**.

Model	AUC
ATLAS initialization	0.9192
Scratch transformer	0.9120
Raw-feature MLP	0.9400

ATLAS initialization improves AUC by 0.0072 over scratch in this run. The raw-feature MLP achieves the highest AUC of the three.

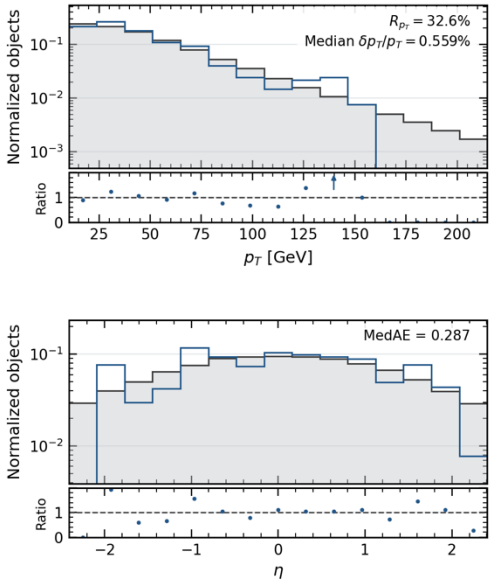
Backup

CMS dilepton transfer dataset

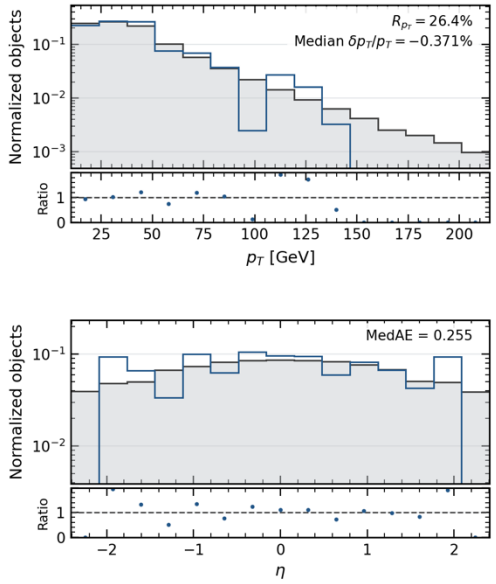
Component	What I did
Five CMS VQ-VAEs	Trained separately on CMS electron, muon, photon, tau, and jet features. Froze them before classifier training.
CMS input layers	Started the token, type, and position embeddings and the 32→256 projection from fresh weights. CMS code IDs belong to different codebooks, so we did not reuse the ATLAS embedding table.
Event transformer	Copied the **four encoder blocks** and f**inal layer normalization** from ATLAS checkpoint: 54 tensors, 3.29 million parameters. These weights were updated during CMS fine-tuning.
Classification head	Replaced the masked-code prediction task with a linear head on [CLS] for dileptonic top pairs vs Drell–Yan. The head trained together with the input layers and transformer.

Electron: shared reconstruction features

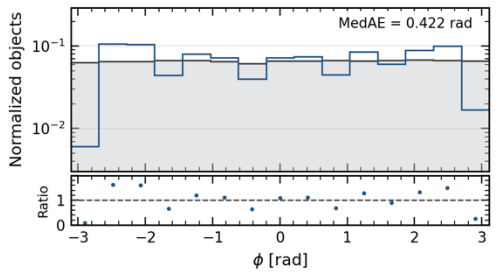
ATLAS

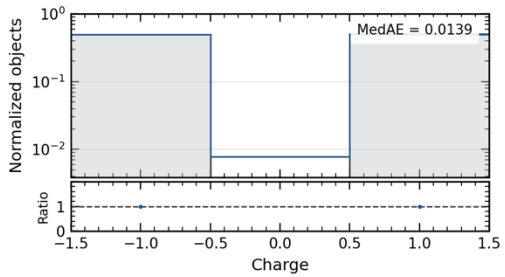


CMS

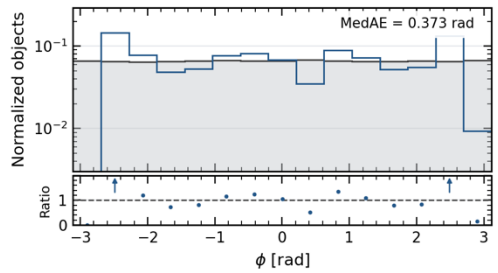


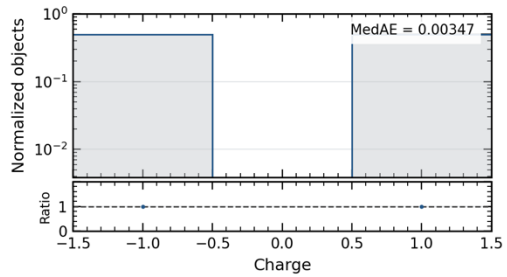
ATLAS





CMS

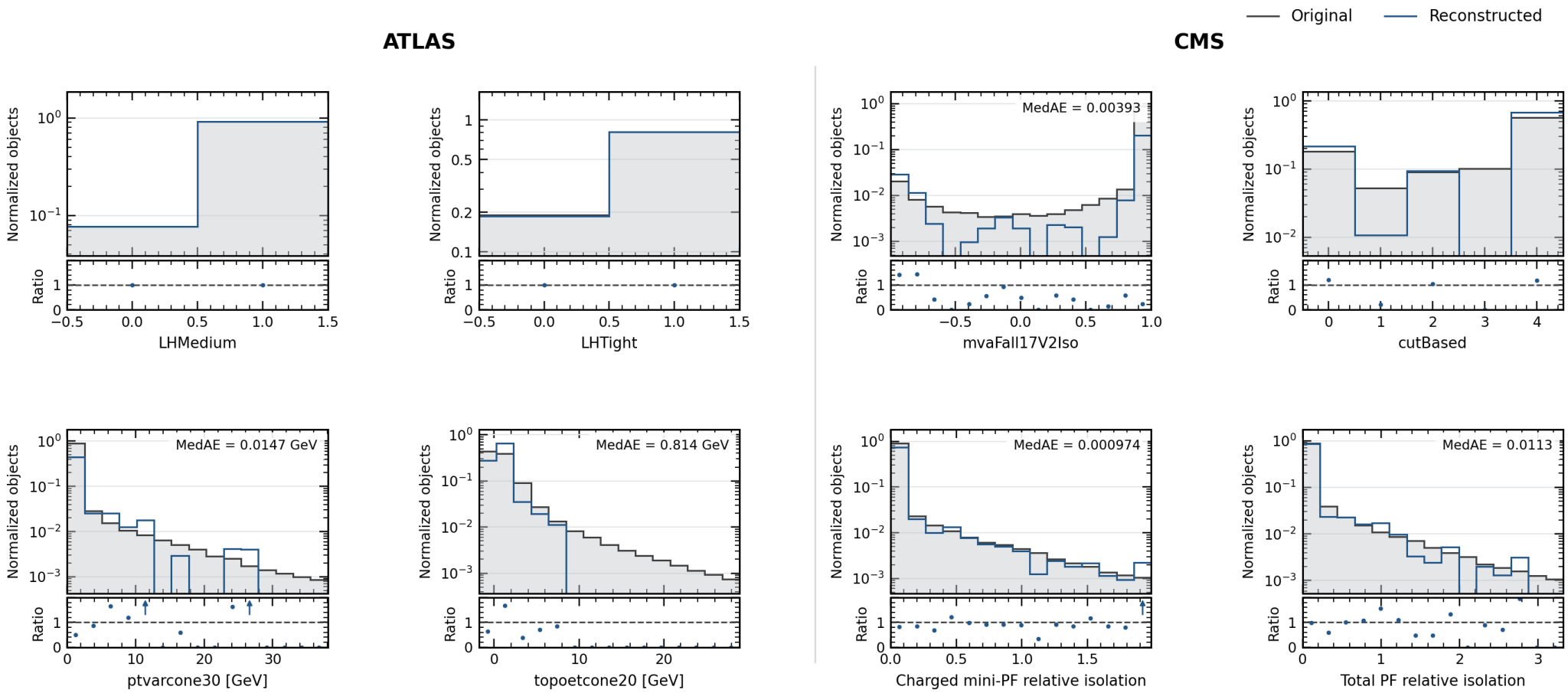




Gray: original Blue: reconstructed Ratio: reconstructed / original

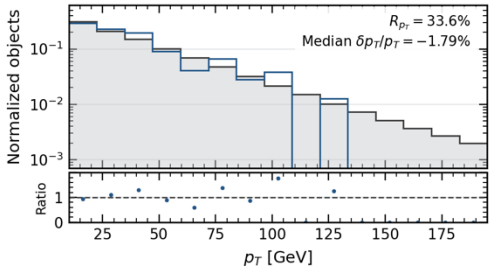
Electrons

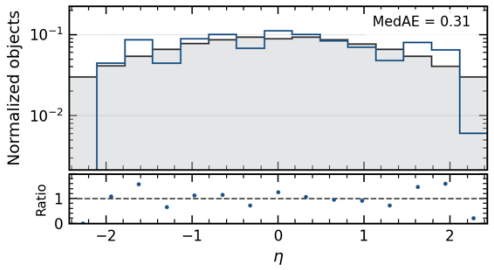
Identification and isolation. CMS uses relative isolation; ATLAS uses isolation in GeV.



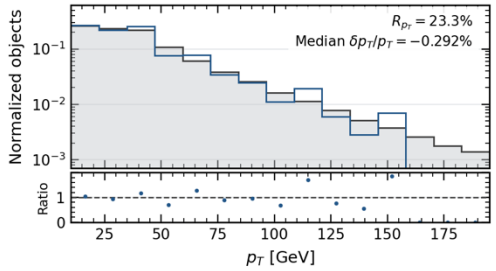
Muons

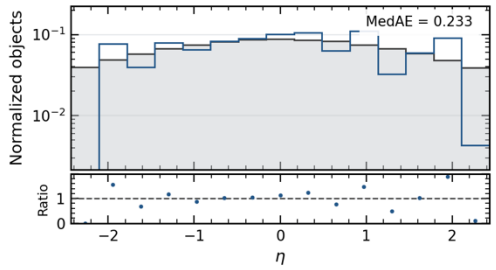
ATLAS



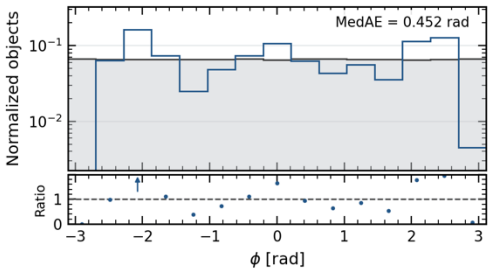


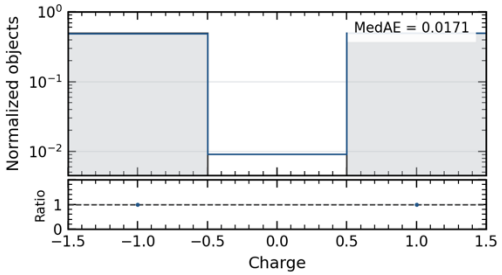
CMS



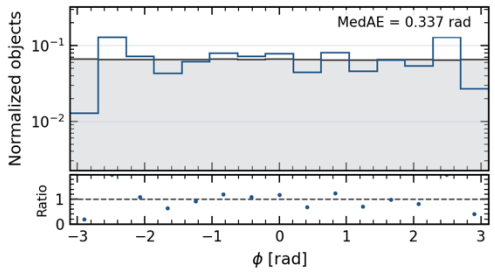


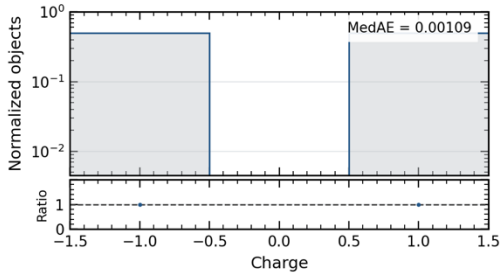
ATLAS





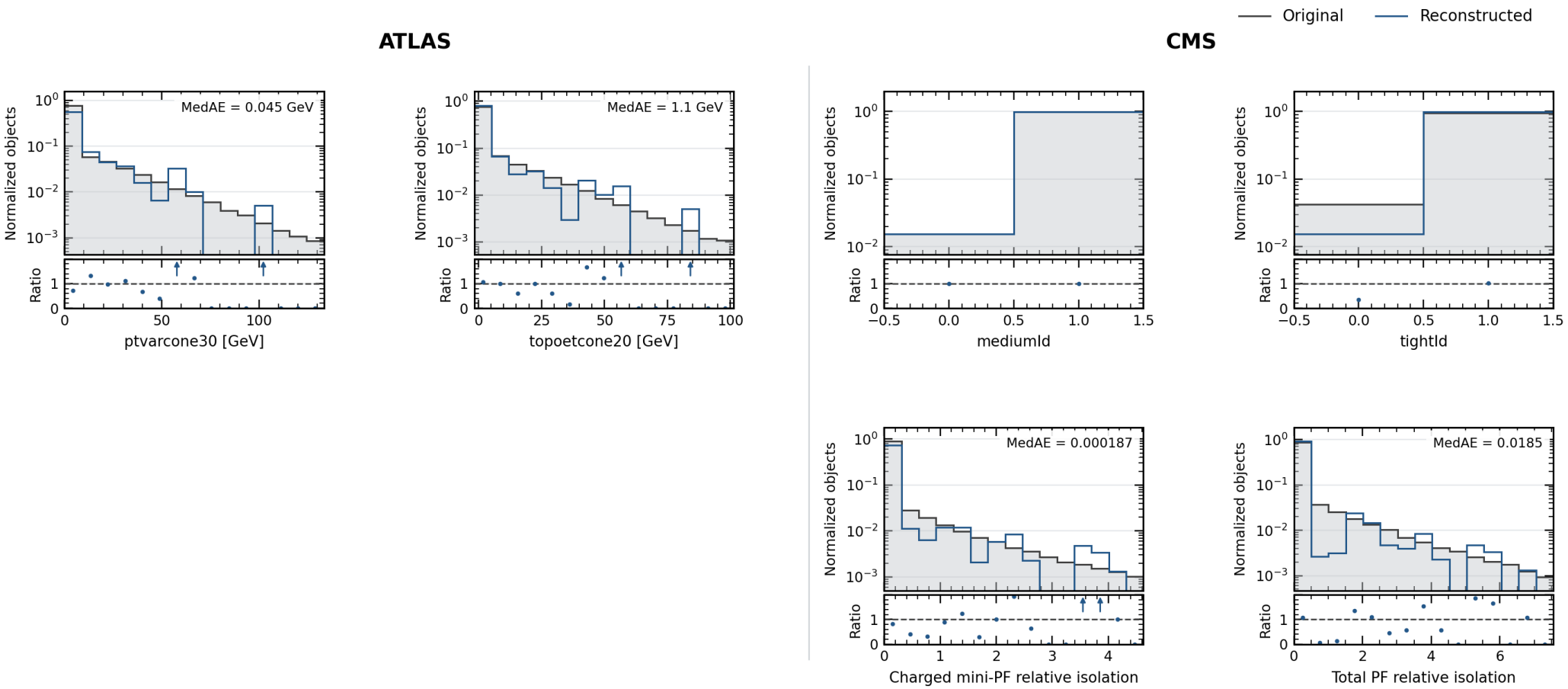
CMS



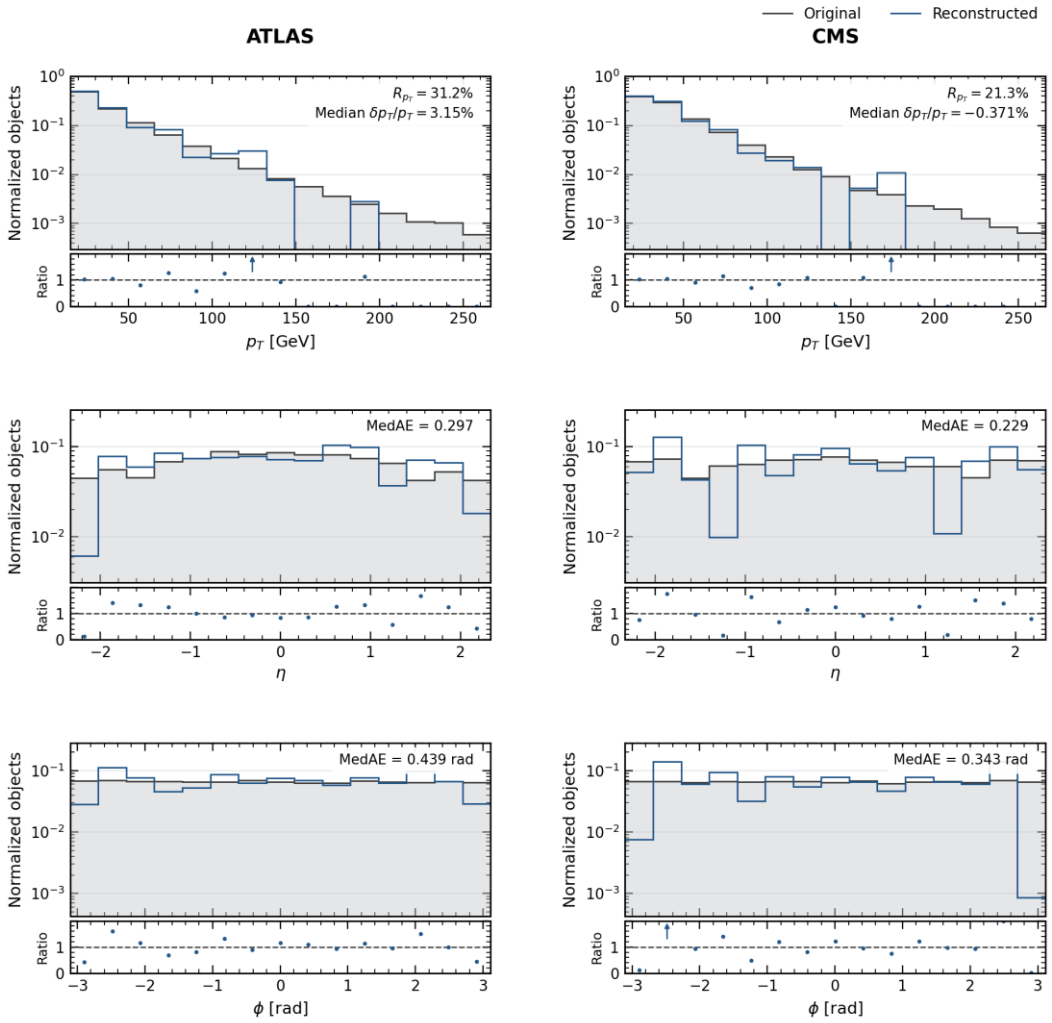


Muons

Isolation in both experiments. CMS also includes medium and tight identification flags.



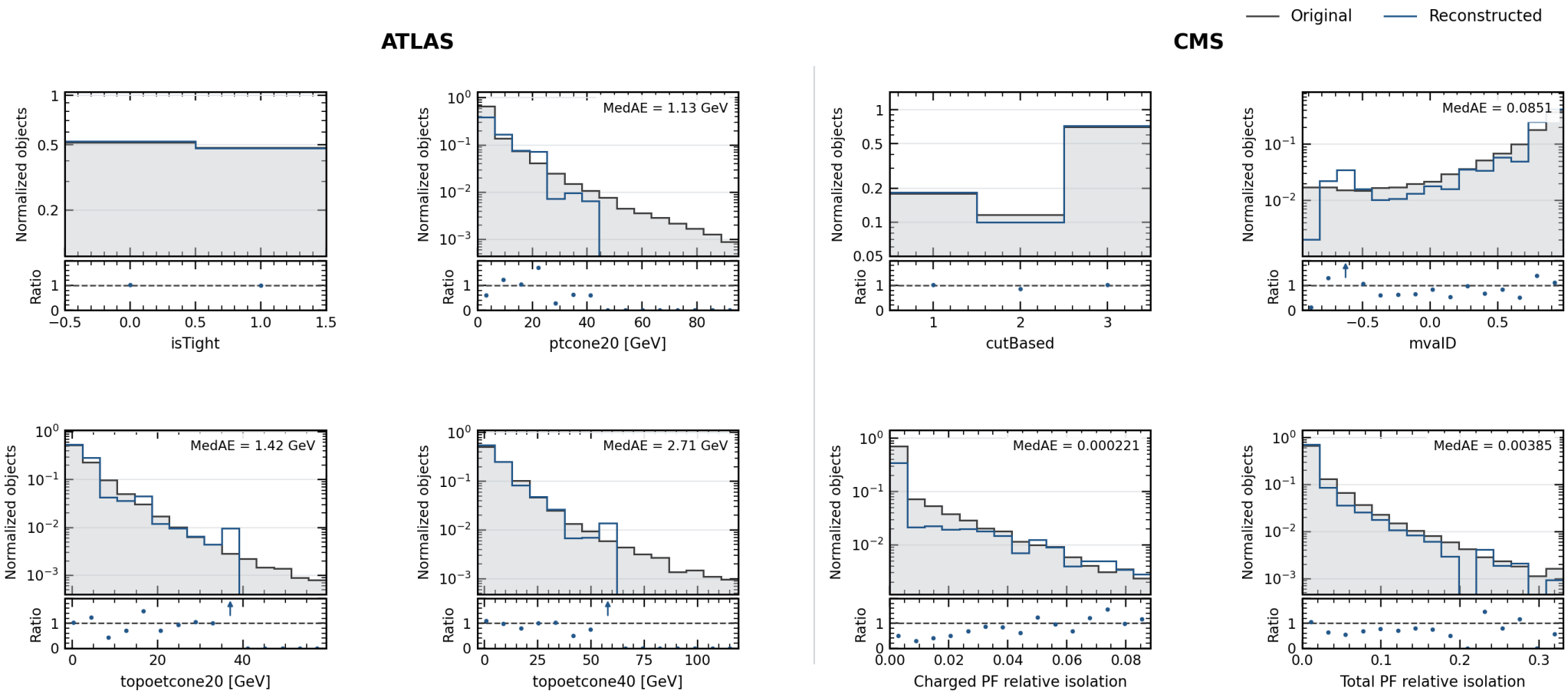
Photons



Shared bins and y scale within each row; separate samples. Ratio = reconstructed / original.

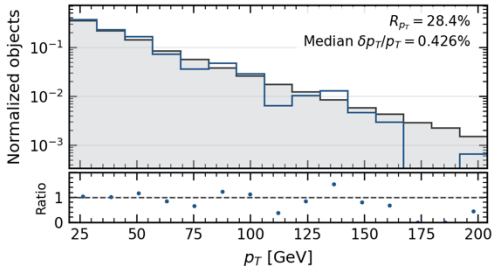
Photons

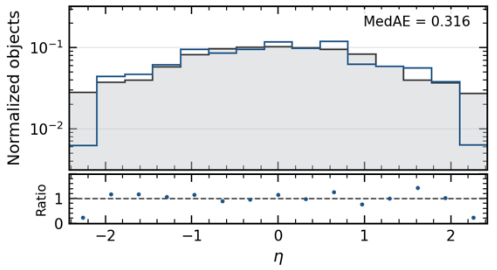
Identification and isolation. CMS and ATLAS use different isolation definitions.



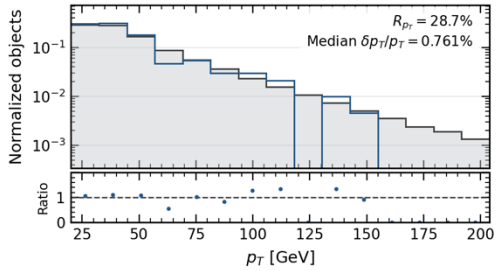
Taus

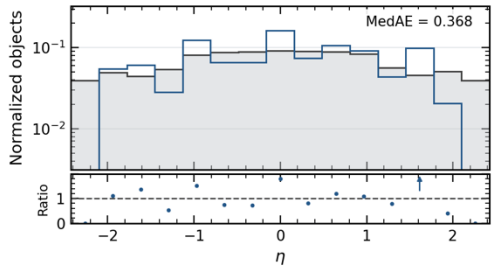
ATLAS



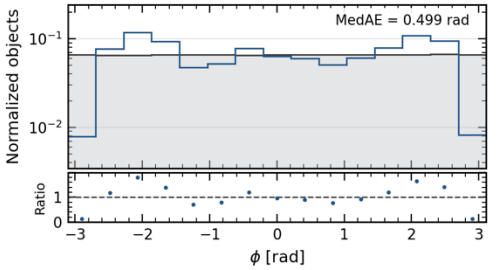


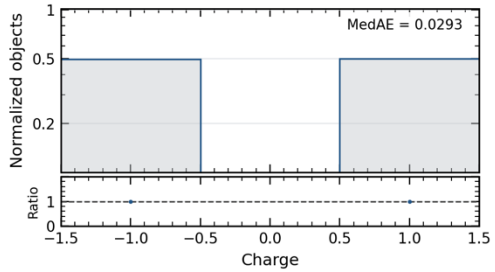
CMS



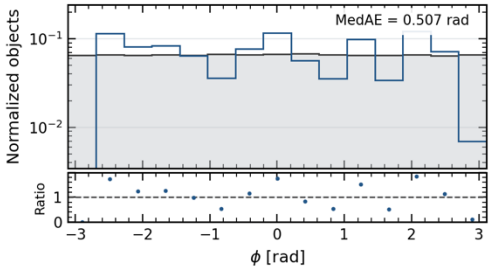


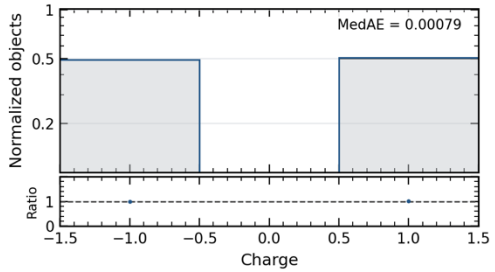
ATLAS





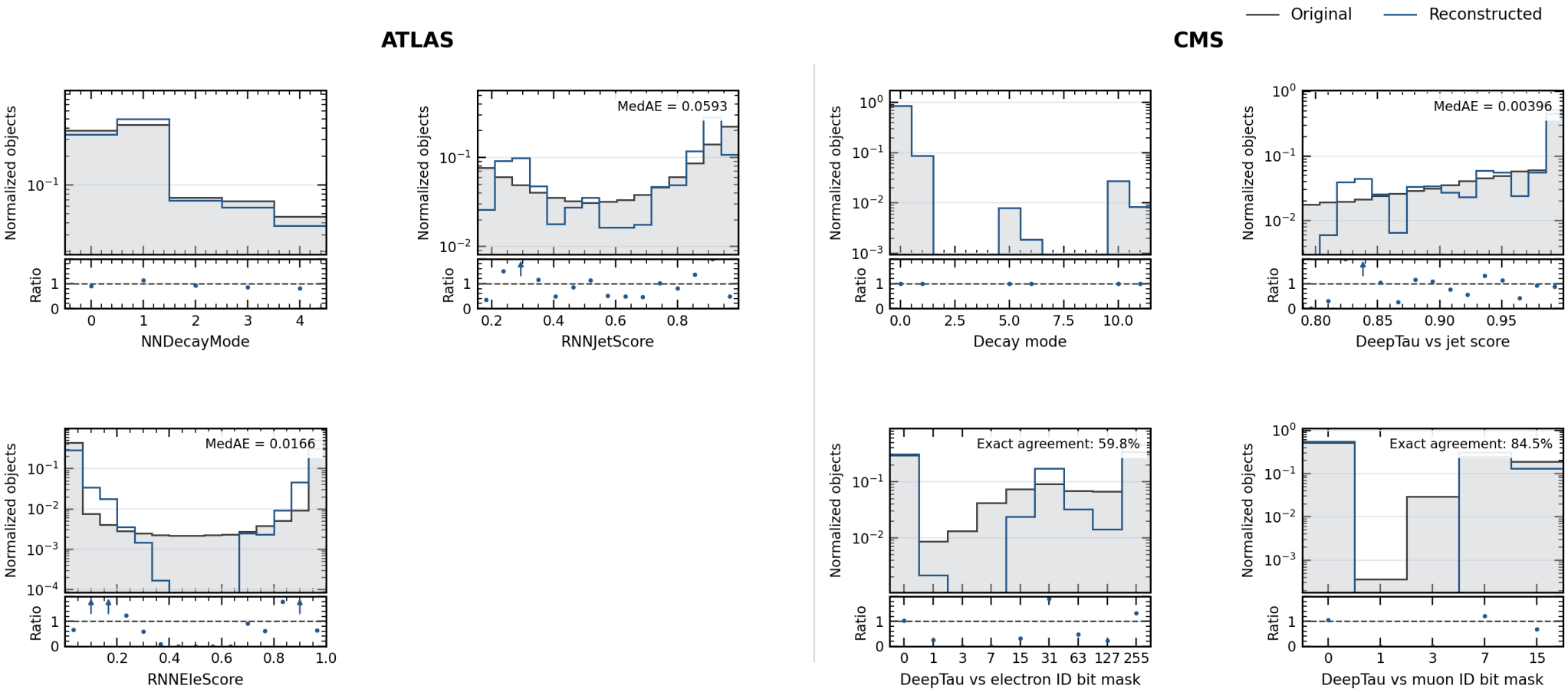
CMS





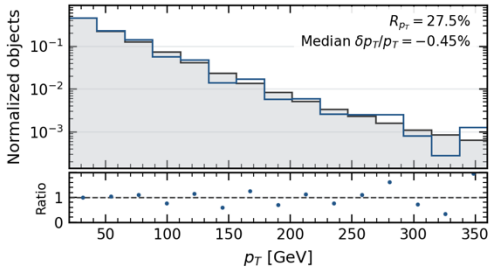
Taus

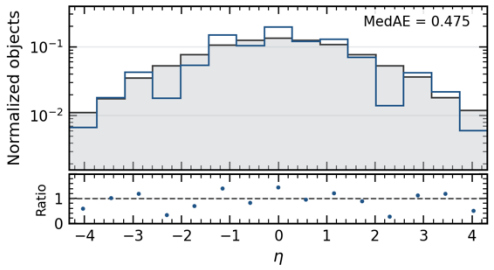
Decay mode and identification. The two experiments use different identification scores and working points.



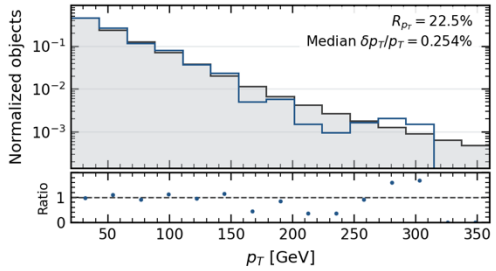
Jets

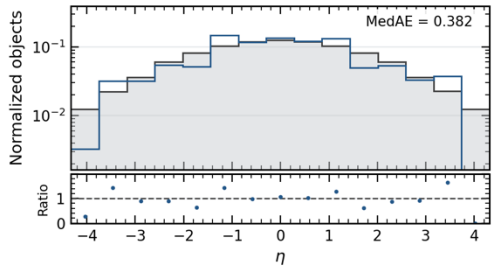
ATLAS



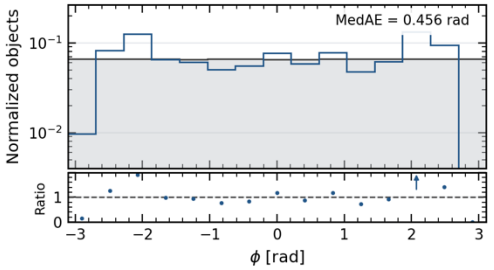


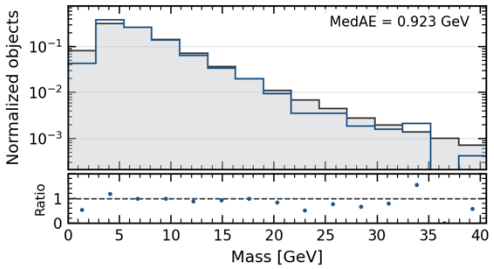
CMS



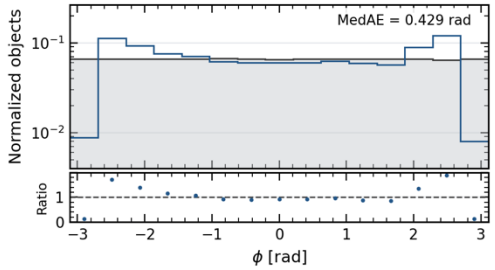


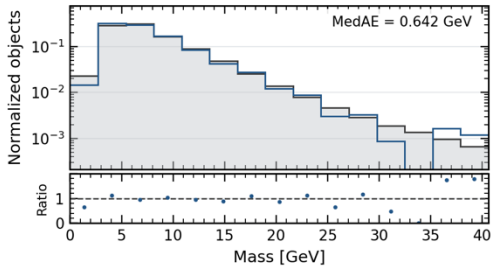
ATLAS



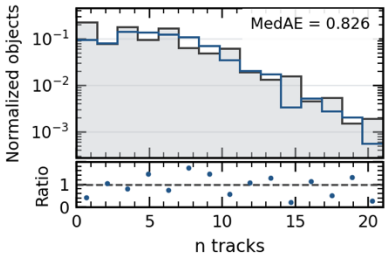


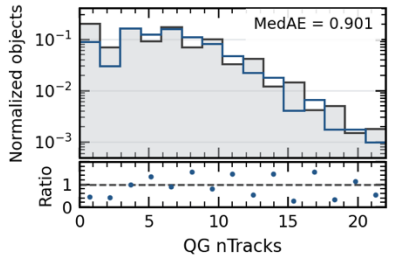
CMS

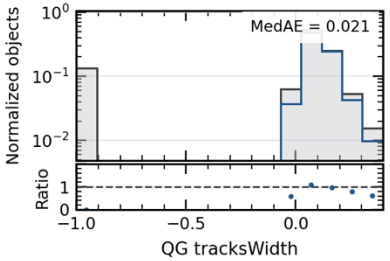


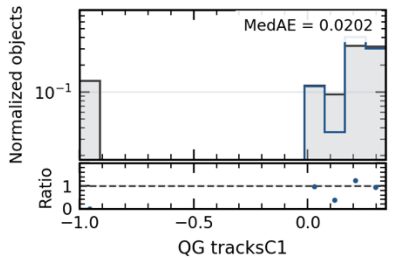


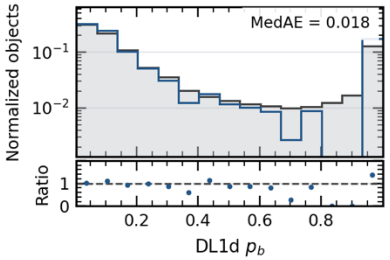
Jets: ATLAS

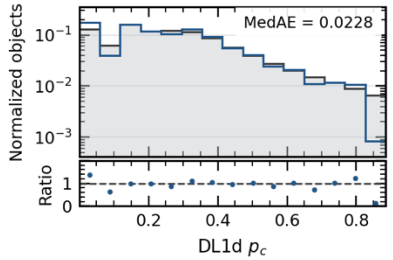


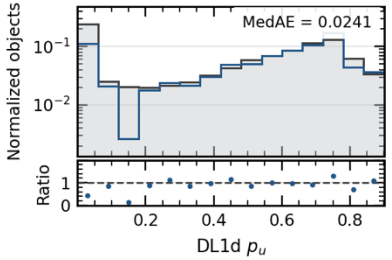


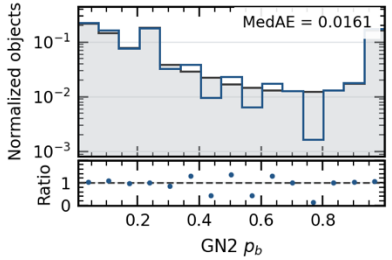


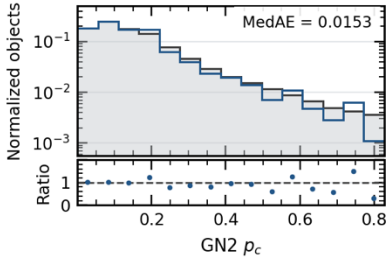


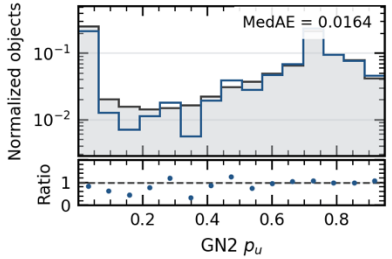






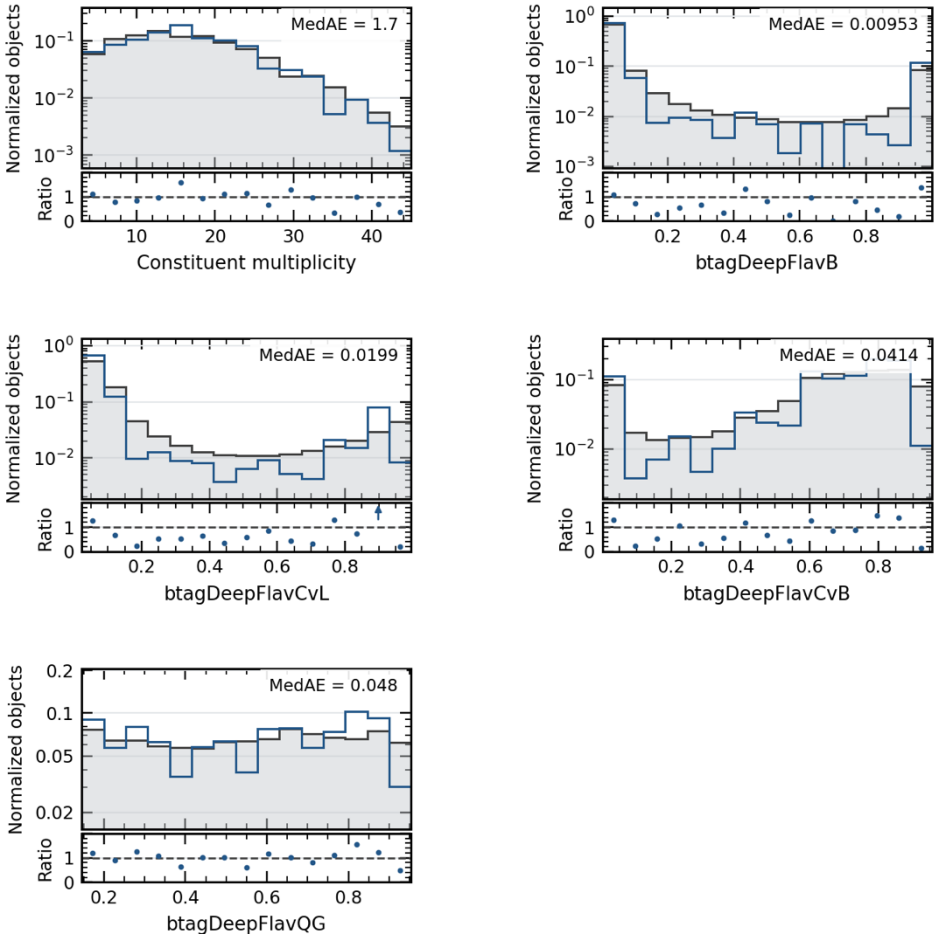






Brookhaven National Laboratory

Jets: CMS



CMS constituent multiplicity and DeepFlavour scores.