Treasure event level-tokenisation

Merve Nazlim Agaras

8.7.26

Intro

* Trained per-object tokenizers for event-level inputs.
* Objects: jets, electrons, muons, photons, taus, tracks.
* Each object is tokenized separately with a residual VQ-VAE.
* Final aim: combine object tokens into event sequences for downstream foundation model training.
* Objects: jets, electrons, muons, photons, taus, tracks.
* Input features per objects:

- common/jets/pt
- common/jets/eta
- common/jets/phi
- common/jets/mass
- common/jets/n_trk
- atlas/jets/QG_nTracks
- atlas/jets/QG_tracksWidth
- atlas/jets/QG_tracksC1
- atlas/jets/DL1d_pb
- atlas/jets/DL1d_pc
- atlas/jets/DL1d_pu
- atlas/jets/GN2_pb
- atlas/jets/GN2_pc
- atlas/jets/GN2_pu

- common/electrons/pt
- common/electrons/eta
- common/electrons/phi
- common/electrons/charge
- atlas/electrons/LHMedium
- atlas/electrons/LHTight
- atlas/electrons/ptvarcone30
- atlas/electrons/topoetcone20

- common/taus/pt
- common/taus/eta
- common/taus/phi
- common/taus/charge
- atlas/taus/NNDecayMode
- atlas/taus/RNNJetScore
- atlas/taus/RNNEleScore

- common/muons/pt
- common/muons/eta
- common/muons/phi
- common/muons/charge
- atlas/muons/ptvarcone30
- atlas/muons/topoetcone20

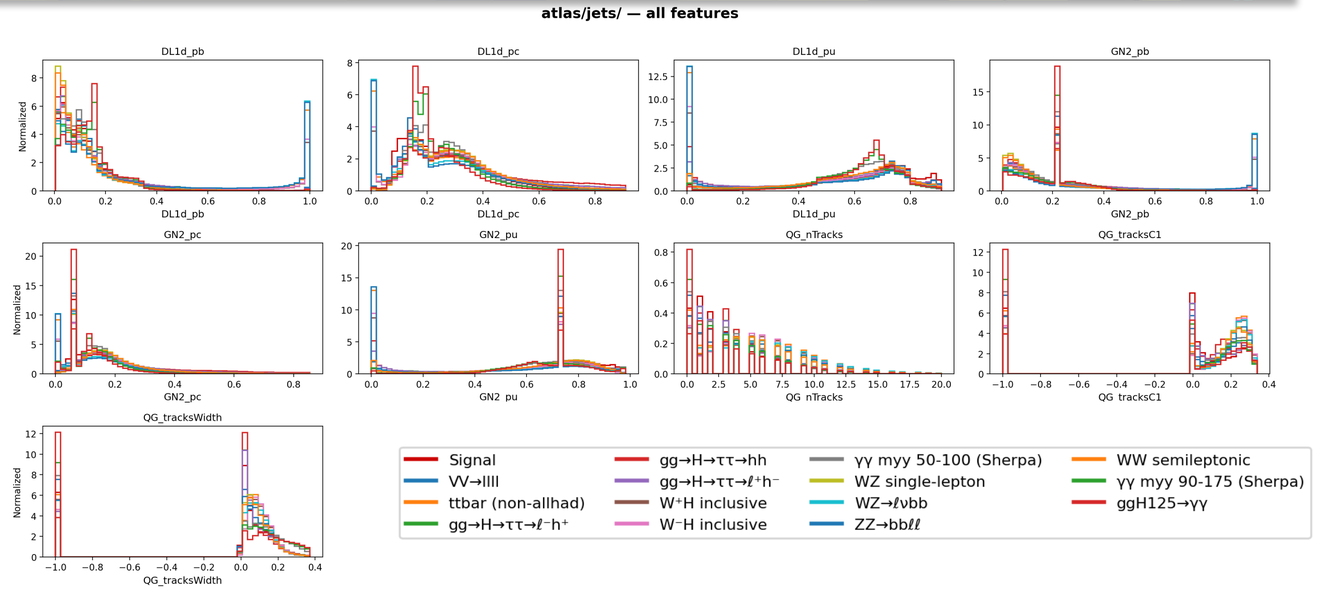
- common/tracks/pt
- common/tracks/eta
- common/tracks/phi
- common/tracks/d0
- common/tracks/z0
- atlas/tracks/qOverP
- atlas/tracks/chiSquared
- atlas/tracks/nDoF

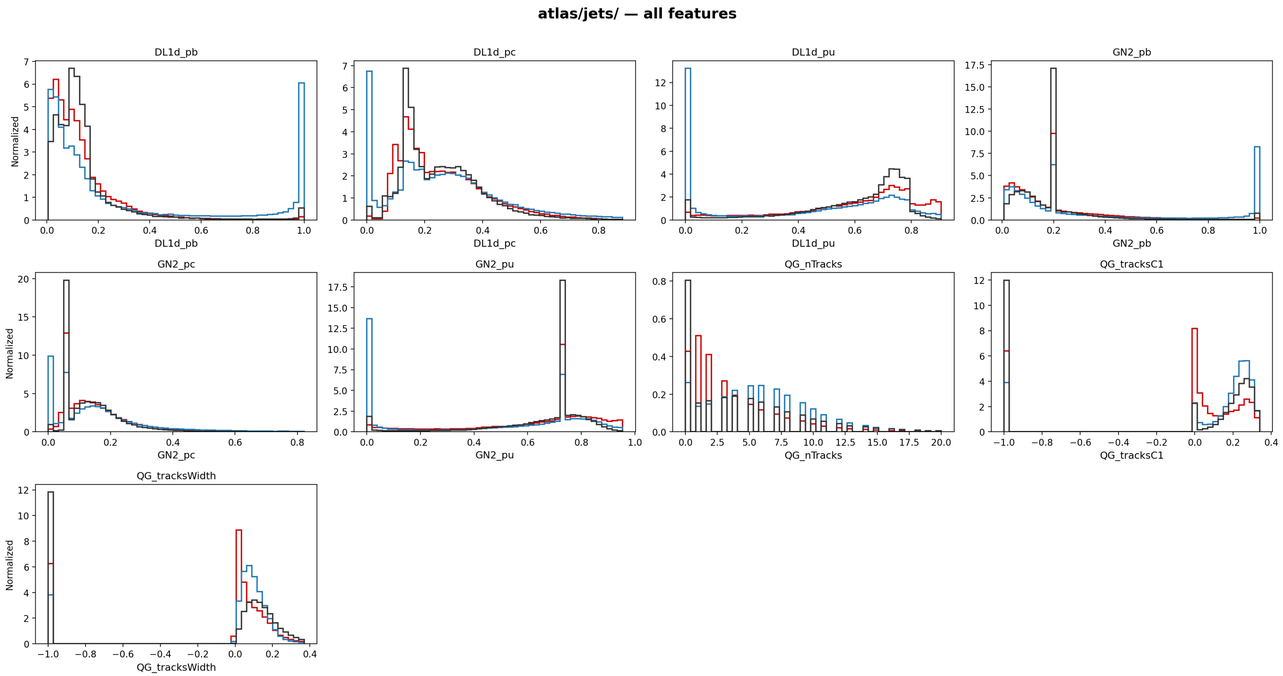
- common/photons/pt
- common/photons/eta
- common/photons/phi
- atlas/photons/isTight
- atlas/photons/ptcone20
- atlas/photons/topoetcone20
- atlas/photons/topoetcone40

Input statistics

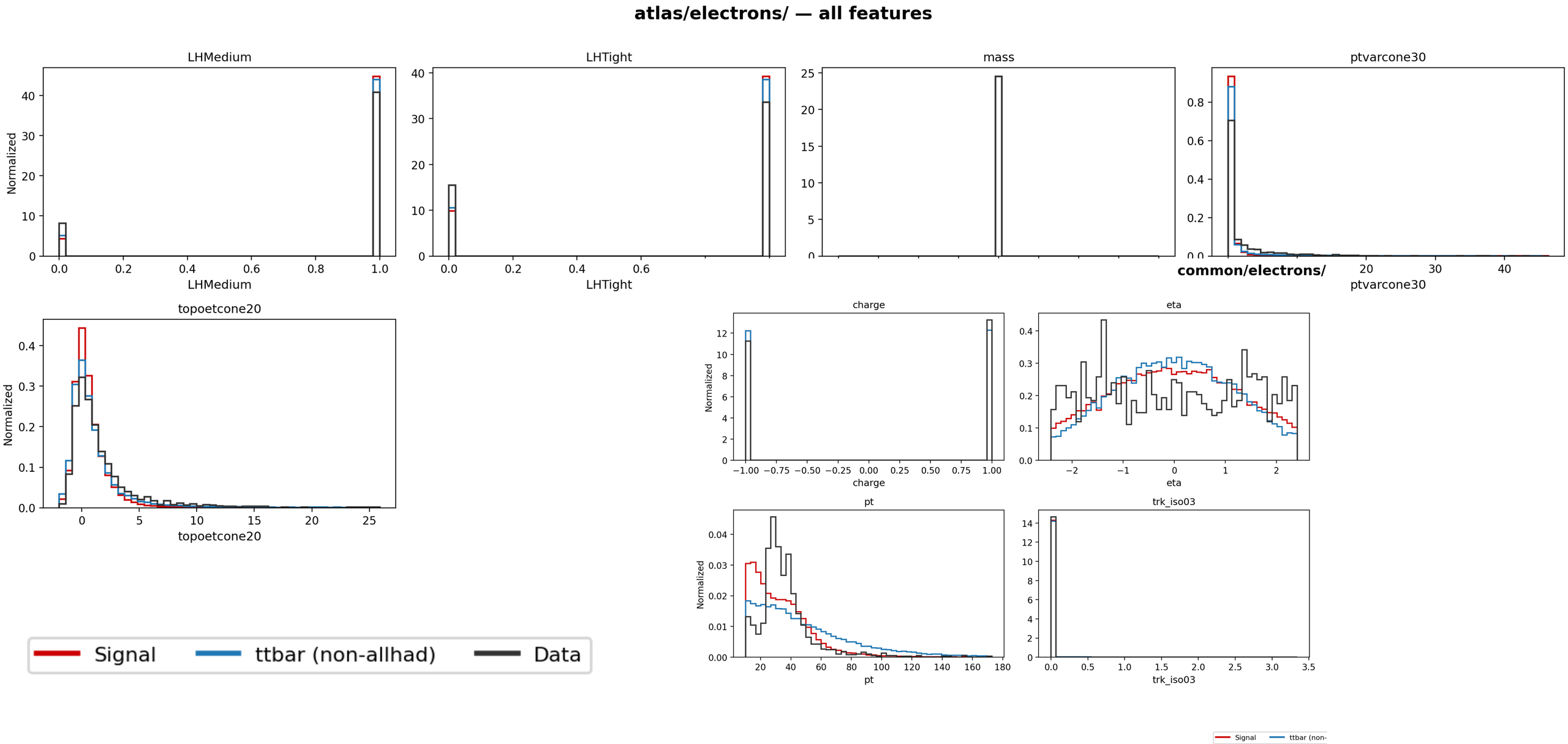
* MC
  - jets events=27,892,710 objects=103,392,186
  - electrons events=27,892,710 objects=9,094,894
  - muons events=27,892,710 objects=10,952,188
  - photons events=27,892,710 objects=14,698,617
  - taus events=27,892,710 objects=16,386,180

* Data
  - jets events=30,870,322 objects=99,158,868
  - electrons events=30,870,322 objects=3,834,447
  - muons events=30,870,322 objects=1,424,257
  - photons events=30,870,322 objects=4,096,533
  - taus events=30,870,322 objects=5,753,068





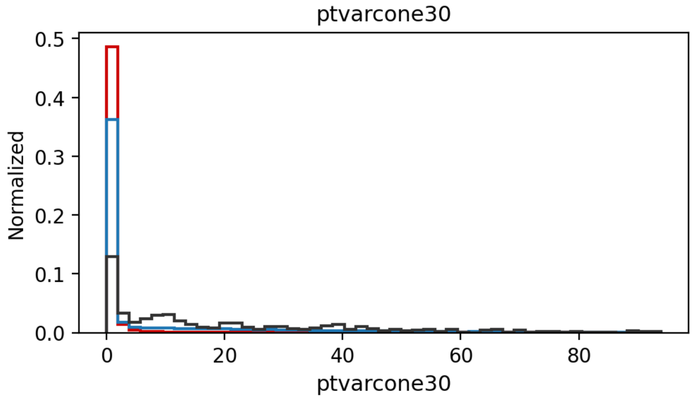
Electrons

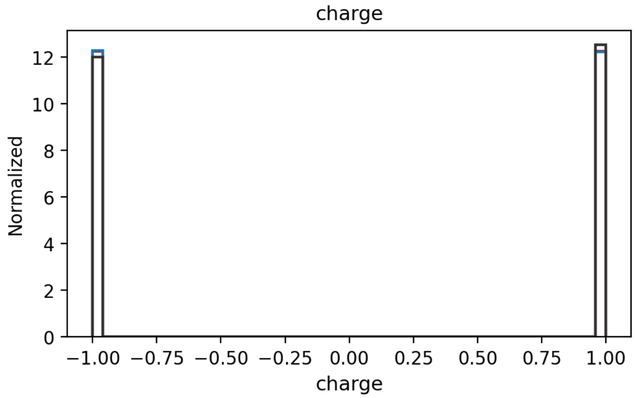

atlas/electrons/ — all features
LHMedium
LHTight
mass
ptvarcone30
Normalized
LHMedium
LHTight
common/electrons/
ptvarcone30
topoetcone20
topoetcone20
charge
eta
pt
trk_iso03
Signal
ttbar (non-allhad)
Data

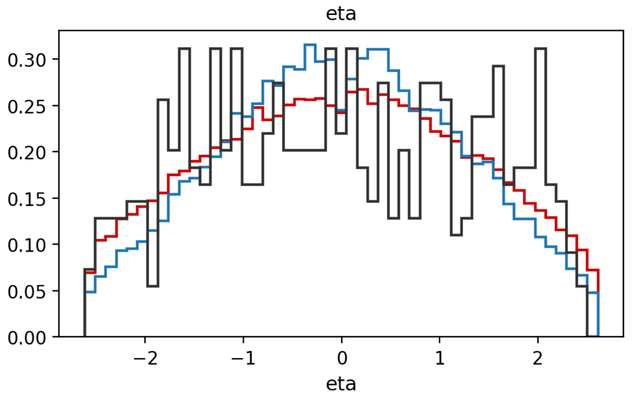
Muons

atlas/muons/ -

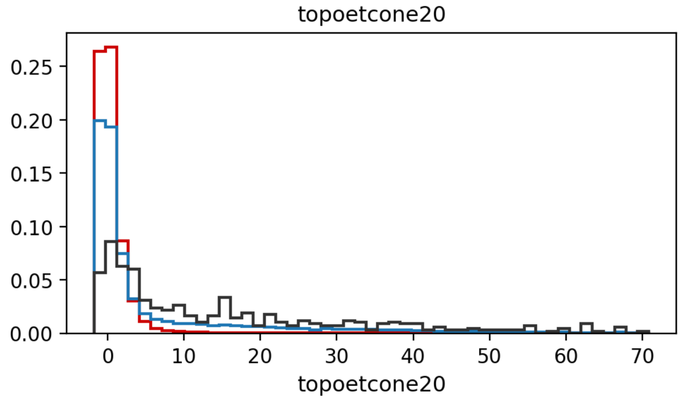
common/muons/ ·

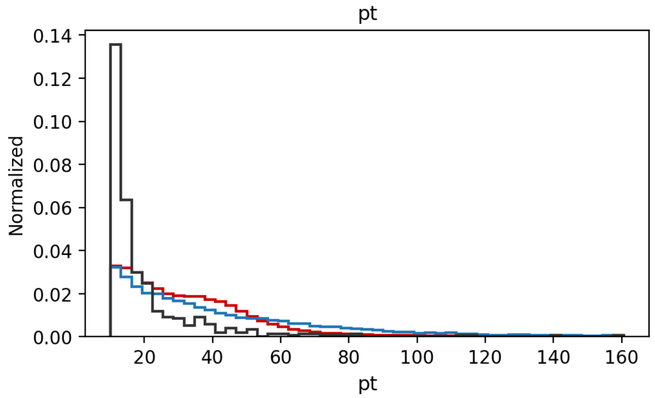


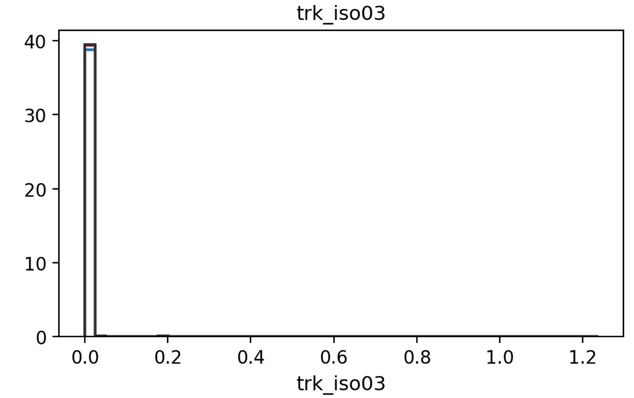




– all features



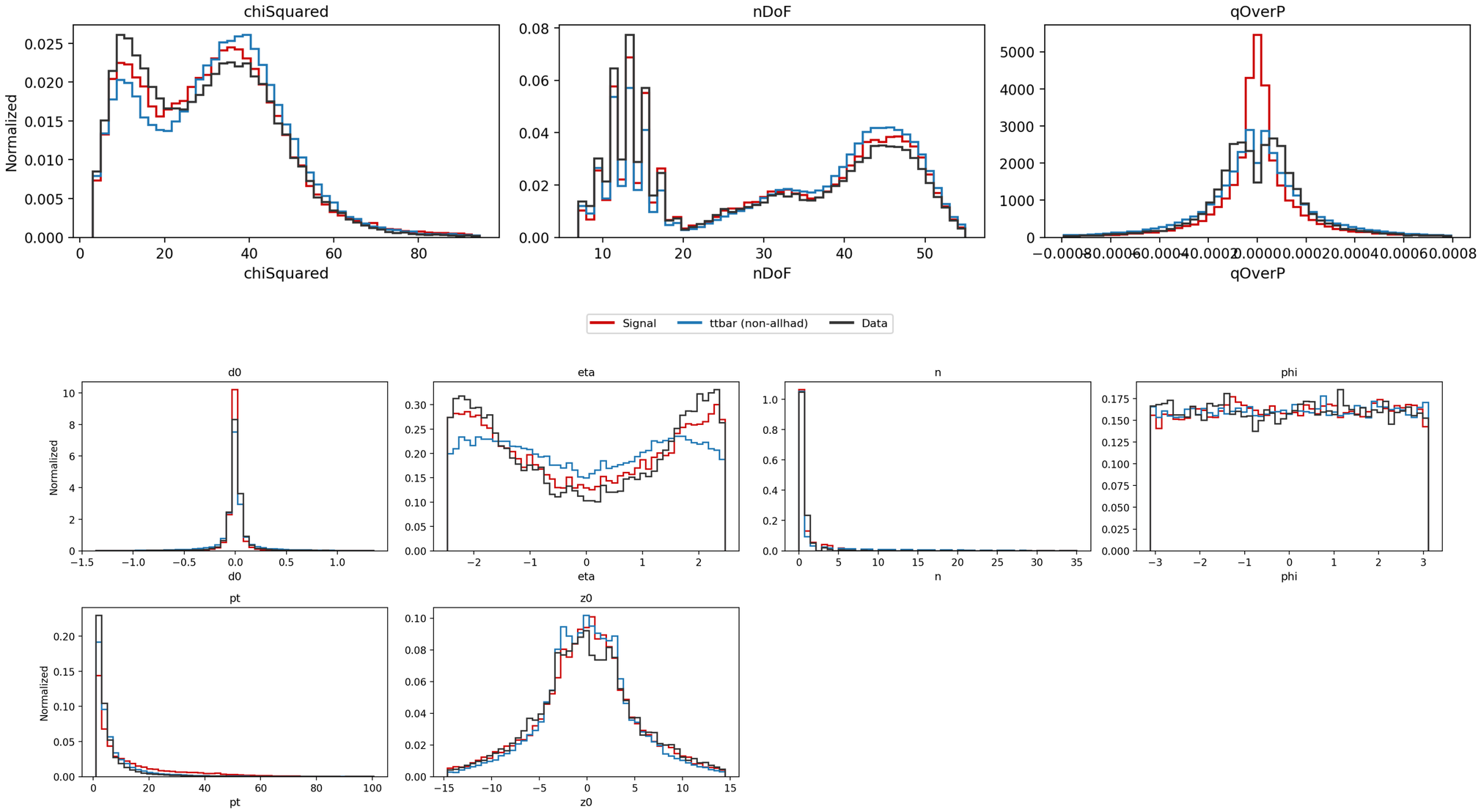




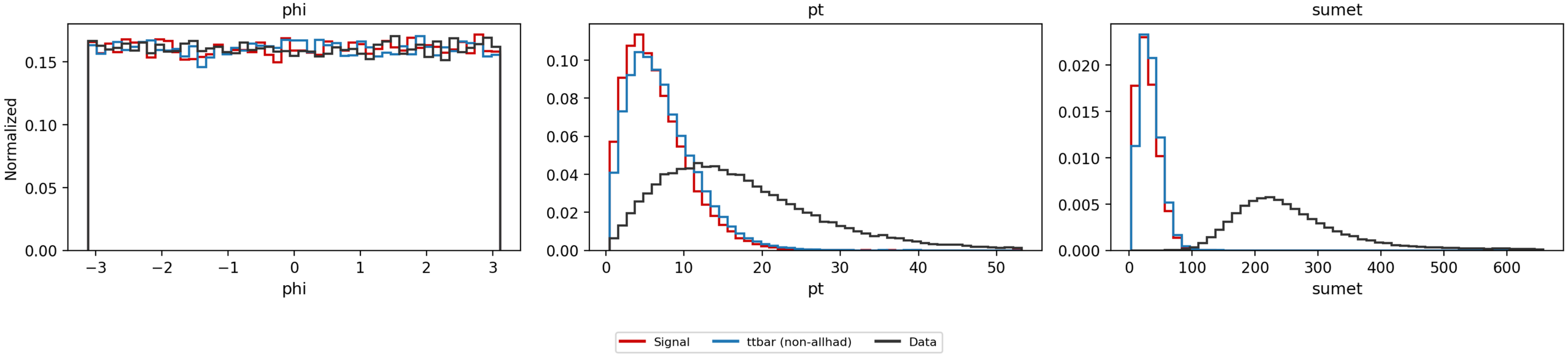
Signal ttbar (non-allhad) Data

Tracks

atlas/tracks/ — all features



MET


common/met/ — all features
phi
pt
sumet
Normalized
phi
pt
sumet
Signal
ttbar (non-allhad)
Data

Capacity Scans

- Scanned:
  - codebook dimension: 4, 8, 16 for jets; 4/8 for others.
  - codebook sizes: mostly 4096 and 8192.
  - quantizers: mainly q4 after q4 showed better reconstruction.

- Compared:
  - overall MAE/RMSE,
  - selected feature RMSE,
  - q0 and final quantizer utilization.

object	current preferred setup
jets	dim8_cb4096_q4
electrons	dim8_cb4096_q4
muons	dim8_cb4096_q4
photons	dim8_cb4096_q4
taus	dim8_cb4096_q4 or dim8_cb8192_q4
tracks	dim8_cb8192_q4, no log on nDoF

Training on MC+Data

- Trained the chosen setup per object on MC+Data
- Compared the resolutions on each tokeniser training in data and MC only setups
- For some objects, Data+MC training gives weaker reconstruction than MC only training
  - Including real data improves jet tokenization, likely because jets have broad detector-level structure and benefit from additional phase-space coverage.
  - For electrons, the MC-only tokenizer gives better pT resolution, suggesting that the current MC+data training either introduces domain-shift/noisy tails or spreads limited codebook capacity over regions that do not improve the core electron reconstruction.

Jets

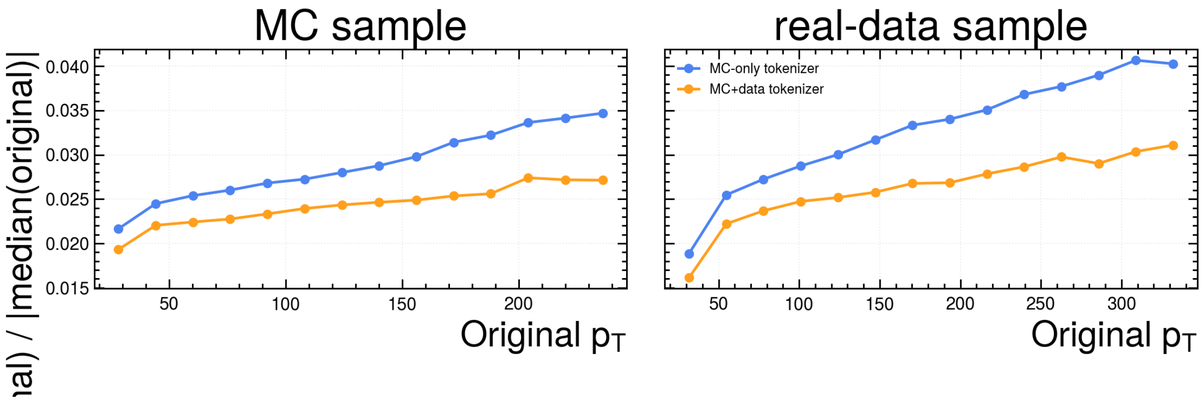
On MC samples

metric	MC only	MC+data	better
mean codebook used	78.63%	96.66%	MC+data
pT RMSE	2.26	1.84	MC+data
eta RMSE	91	77	MC+data
mass RMSE	265	407	MC only
n_trk RMSE	200	150	MC+data
GN2_pc RMSE	0.0084	0.0052	MC+data

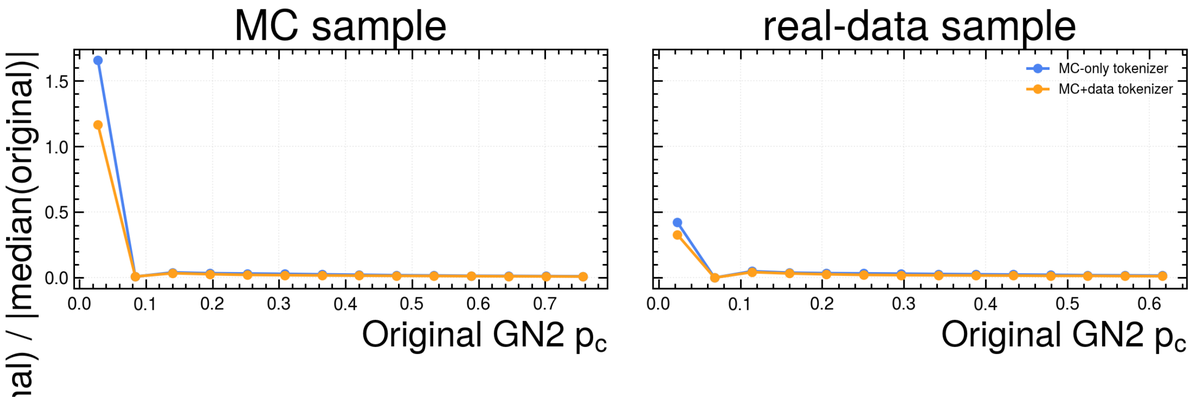
On data samples

metric	MC only	MC+data	better
mean codebook used	78.67%	96.59%	MC+data
pT RMSE	2.80	2.20	MC+data
eta RMSE	101	75	MC+data
mass RMSE	293	446	MC only
n_trk RMSE	214	141	MC+data
GN2_pc RMSE	0.0083	0.0052	MC+data

jets: p_T resolution



jets: GN2 p_c resolution



Electrons

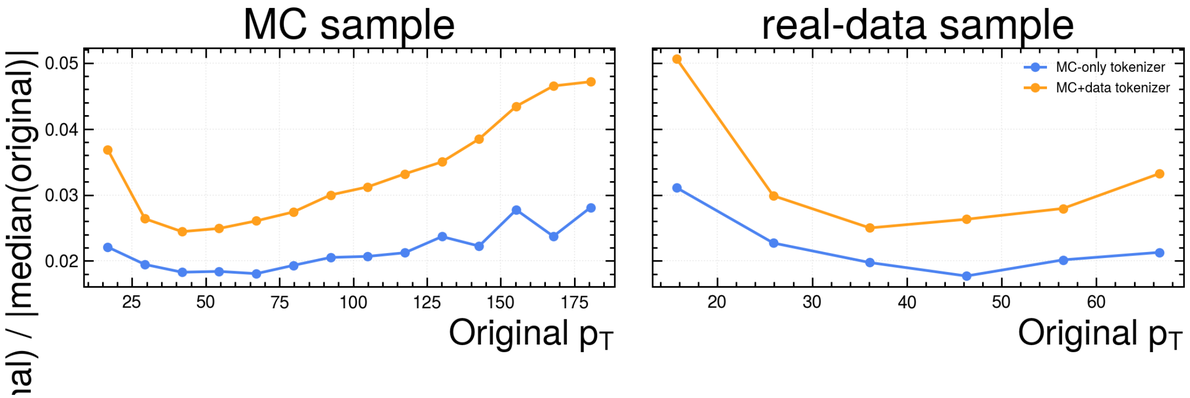
On MC samples

metric	MC only	MC+data	better
mean codebook used	38.17%	26.98%	MC only
pT RMSE	3.03	3.68	MC only
eta RMSE	41	58	MC only
ptvarcone30 RMSE	2.00	2.77	MC only
topoetcone20 RMSE	1.21	2.68	MC only

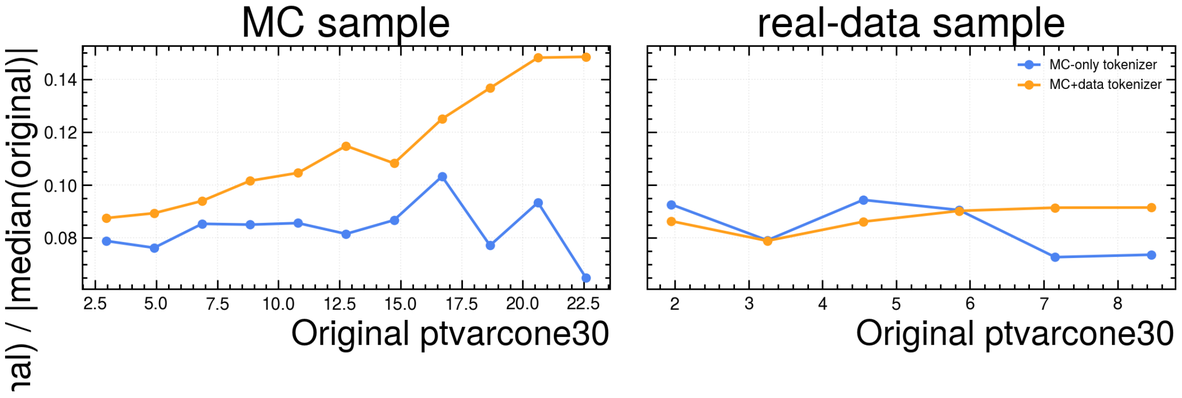
On data samples

metric	MC only	MC+data	better
mean codebook used	36.73%	26.16%	MC only
pT RMSE	2.71	3.64	MC only
eta RMSE	44	61	MC only
ptvarcone30 RMSE	617	1.720	MC only
topoetcone20 RMSE	377	487	MC only

electrons: p_T resolution



electrons: ptvarcone30 resolution



Muons

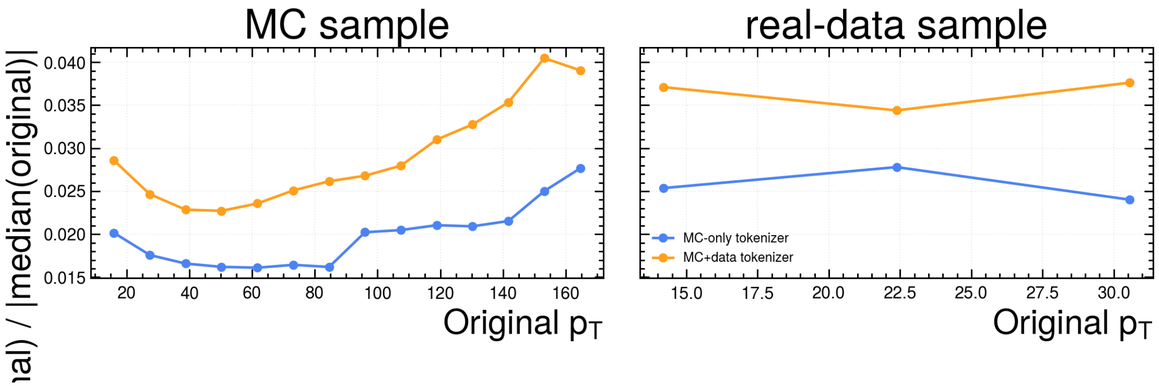
On MC samples

metric	MC only	MC+data	better
mean codebook used	45.05%	27.45%	MC only
pT RMSE	3.23	6.17	MC only
eta RMSE	34	42	MC only
ptvarcone30 RMSE	5.34	11.05	MC only
topoetcone20 RMSE	847	1.818	MC only

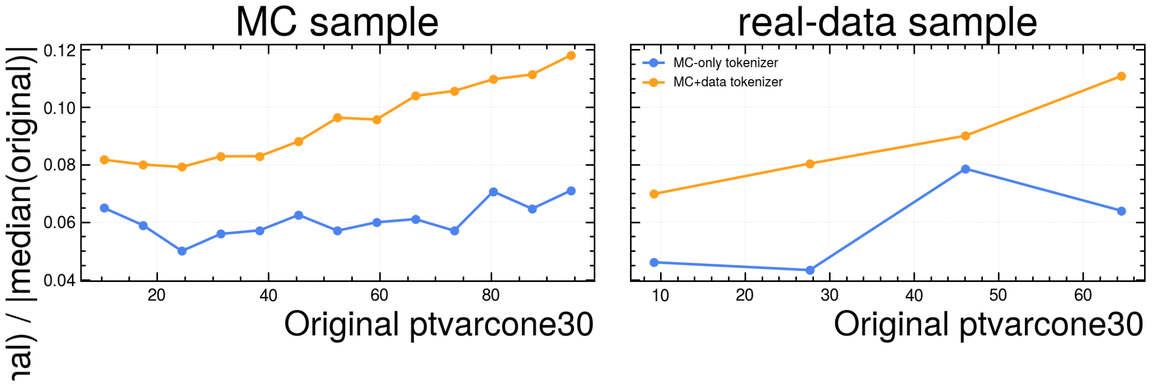
On data samples

metric	MC only	MC+data	better
mean codebook used	40.18%	26.13%	MC only
pT RMSE	3.80	8.88	MC only
eta RMSE	73	67	MC+data
ptvarcone30 RMSE	85.83	86.09	MC only, tiny
topoetcone20 RMSE	3.41	9.08	MC only

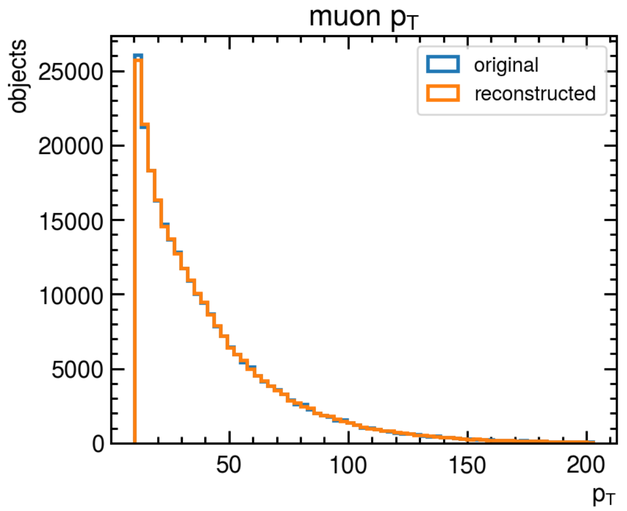
muons: p_T resolution

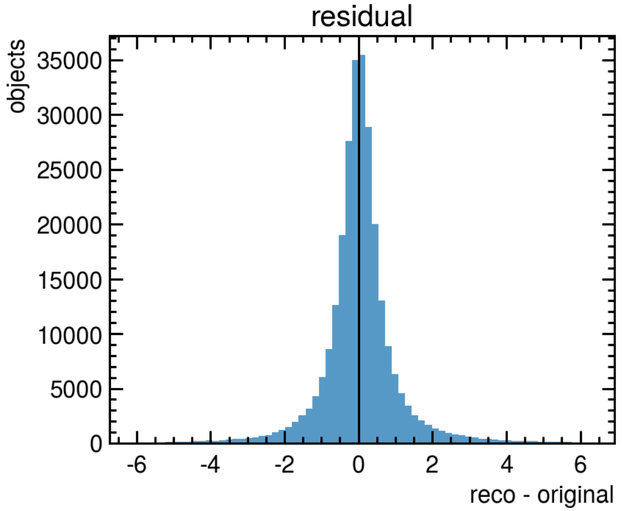


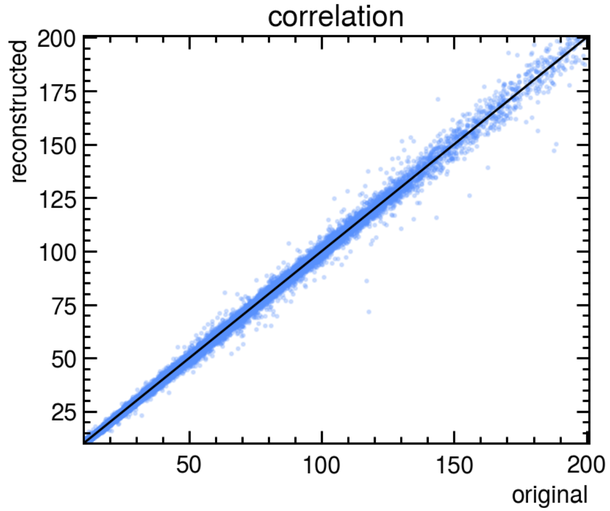
muons: ptvarcone30 resolution



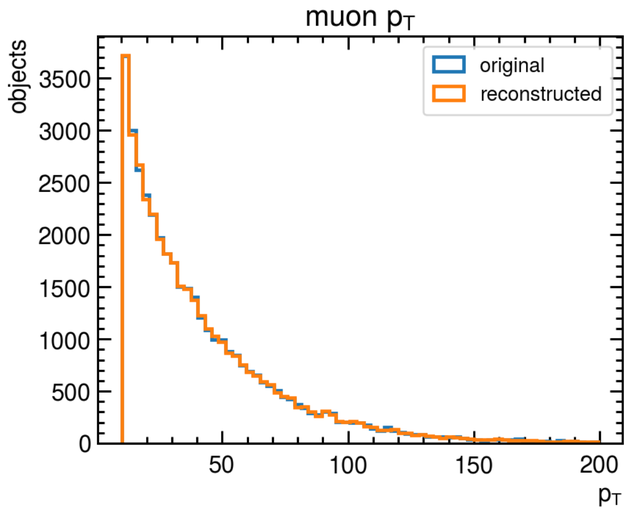
Muon pT

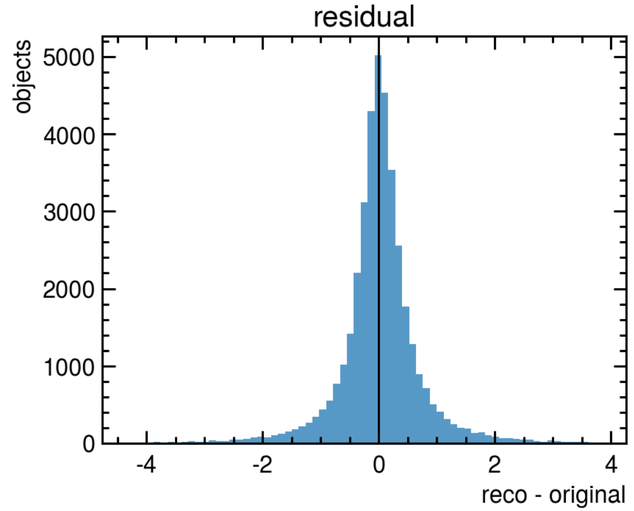


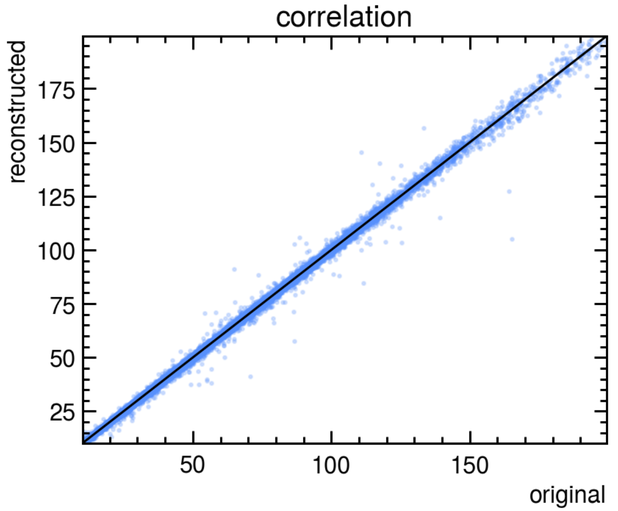




MCOnly training







Data+MC

Taus

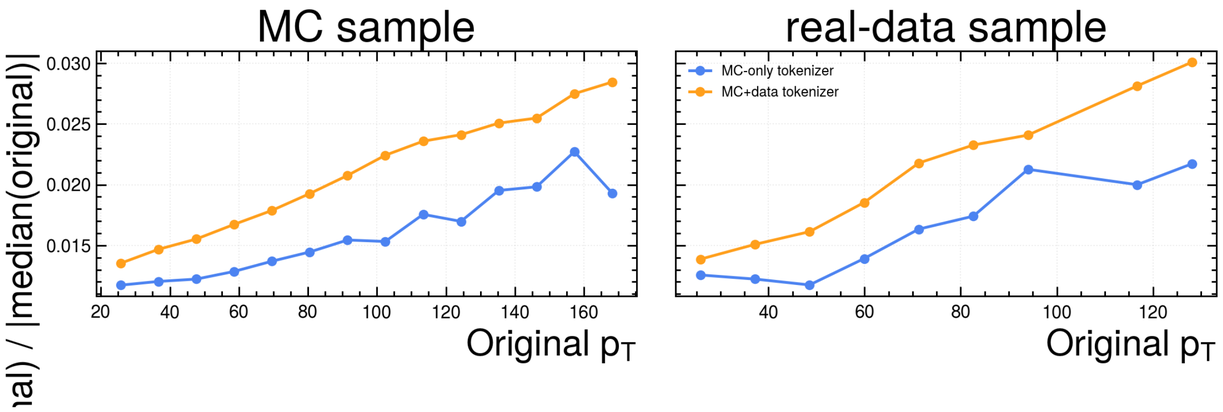
On MC samples

metric	MC only	MC+data	better
mean codebook used	33.01%	21.55%	MC only
pT RMSE	1.15	1.31	MC only
eta RMSE	22	28	MC only
charge RMSE	0.00067	0.00239	MC only
NNDecayMode RMSE	0.0118	0.0191	MC only

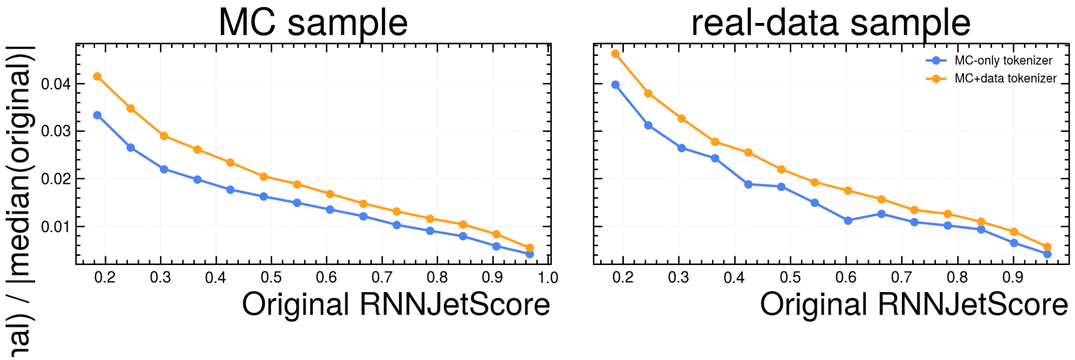
On data samples

metric	MC only	MC+data	better
mean codebook used	32.33%	21.14%	MC only
pT RMSE	1.57	1.36	MC+data
eta RMSE	26	29	MC only
charge RMSE	0.00092	0.00255	MC only
NNDecayMode RMSE	0.0212	0.0226	MC only

taus_8k: p_T resolution



taus_8k: RNNJetScore resolution



Photons

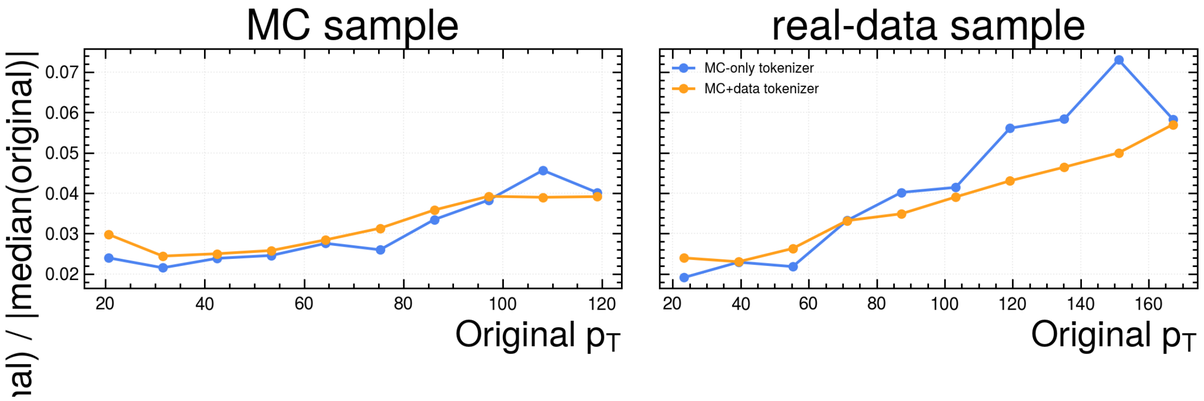
On MC samples

metric	MC only	MC+data	better
mean codebook used	40.42%	38.32%	MC only
pT RMSE	5.37	3.70	MC+data
eta RMSE	63	58	MC+data
isTight RMSE	0.0084	0.0098	MC only
ptcone20 RMSE	37.96	47.10	MC only
topoetcone20 RMSE	1.49	2.64	MC only
topoetcone40 RMSE	2.21	3.61	MC only

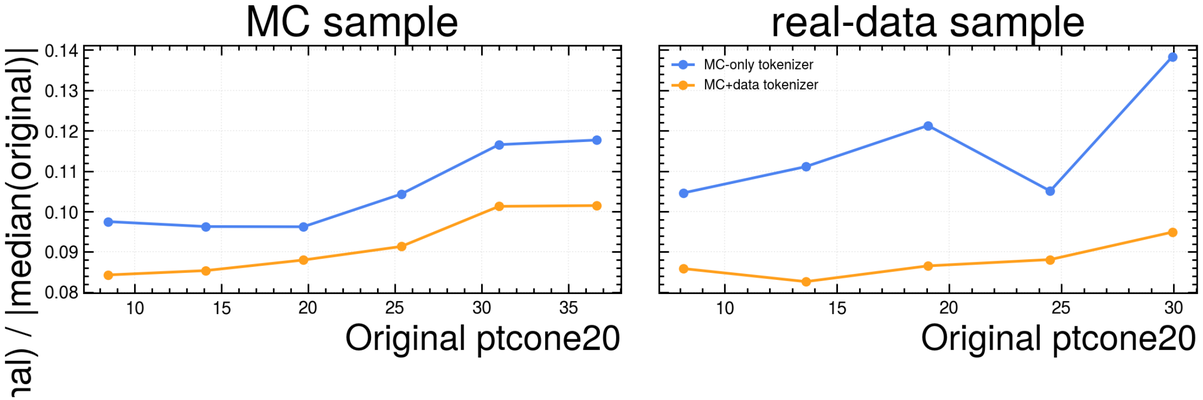
On data samples

metric	MC only	MC+data	better
mean codebook used	40.13%	38.08%	MC only
pT RMSE	7.03	6.52	MC+data
eta RMSE	73	58	MC+data
isTight RMSE	12	0.0077	MC+data
ptcone20 RMSE	7.02	7.03	MC only, tiny
topoetcone20 RMSE	836	1.685	MC only
topoetcone40 RMSE	3.77	4.41	MC only

photons: p_T resolution



photons: ptcone20 resolution



Tracks

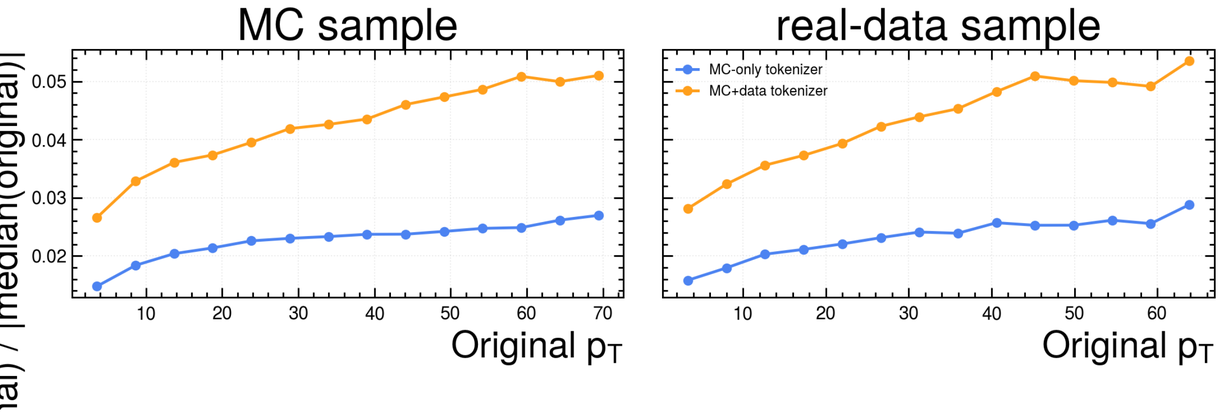
On MC samples

metric	MC only	MC+data	better
mean codebook used	79.01%	26.26%	MC only
pT RMSE	5.72	8.12	MC only
eta RMSE	16	32	MC only
phi RMSE	31	56	MC only
d0 RMSE	0.0057	0.0105	MC only
z0 RMSE	0.0858	0.1550	MC only
nDoF RMSE	217	375	MC only

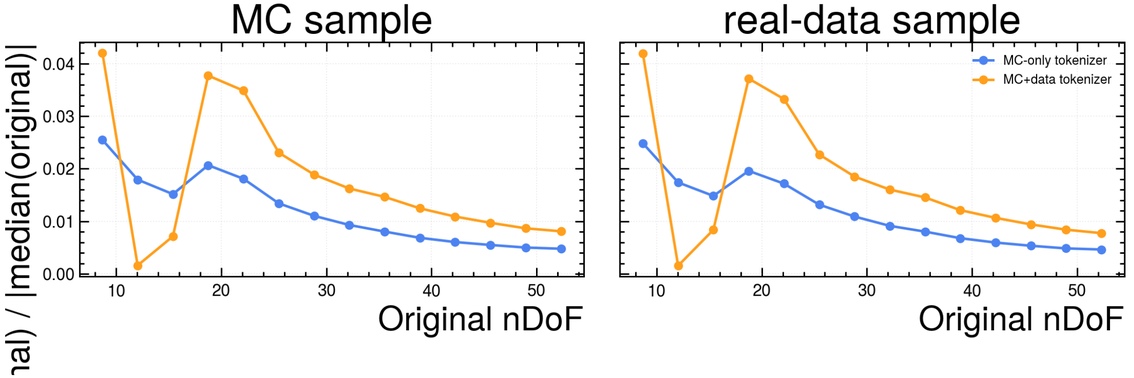
On data samples

metric	MC only	MC+data	better
mean codebook used	78.68%	26.24%	MC only
pT RMSE	8.36	12.03	MC only
eta RMSE	15	30	MC only
phi RMSE	31	54	MC only
d0 RMSE	0.0053	0.0094	MC only
z0 RMSE	0.0853	0.1525	MC only
nDoF RMSE	204	354	MC only

tracks: p_T resolution



tracks: nDoF resolution



Summary/next steps

Summary

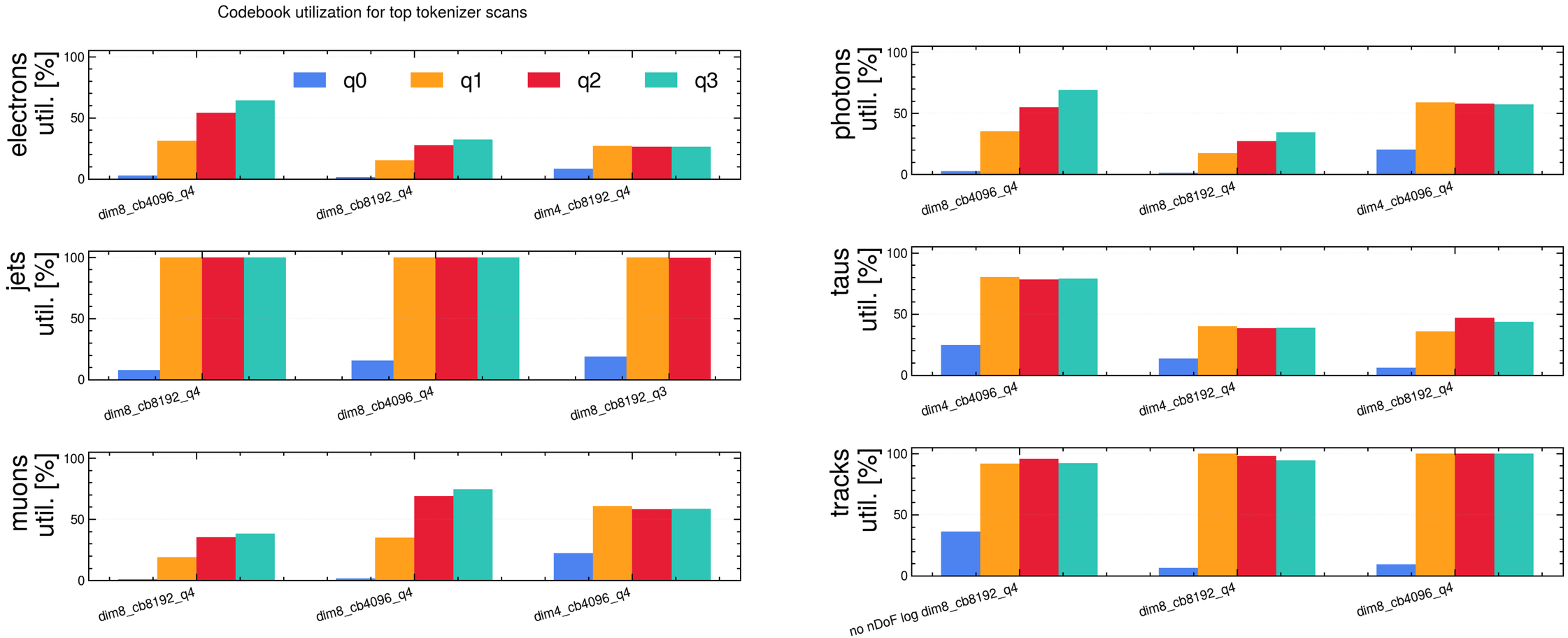
* Compared object tokenizers trained on MC only vs MC+real data using fixed MC and data samples.
* Adding real data does not automatically improve reconstruction for all objects.
* MC-only tokenizers give better reconstruction for electrons, muons, taus, and tracks.
* MC+data tokenizers improve or remain similar mainly for jets, and partly for photons.
* Differences are object-dependent and likely reflect data/MC composition, feature tails, and training balance.

* Next Steps
* Use MC-only tokenizers as the baseline for the $H \rightarrow ZZ \rightarrow 4\ell$ downstream studies.
* Check whether MC+data training needs reweighting or balanced sampling across MC and real data.
* Check preprocessing of the data samples.

Backup

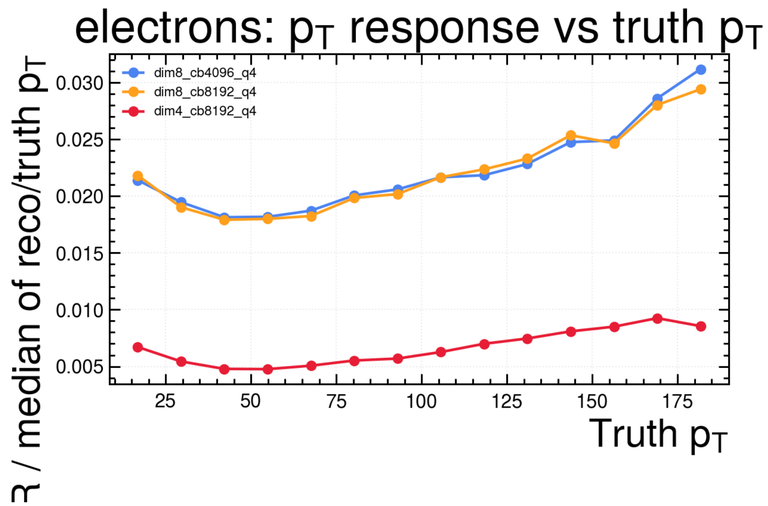
Summary Plots

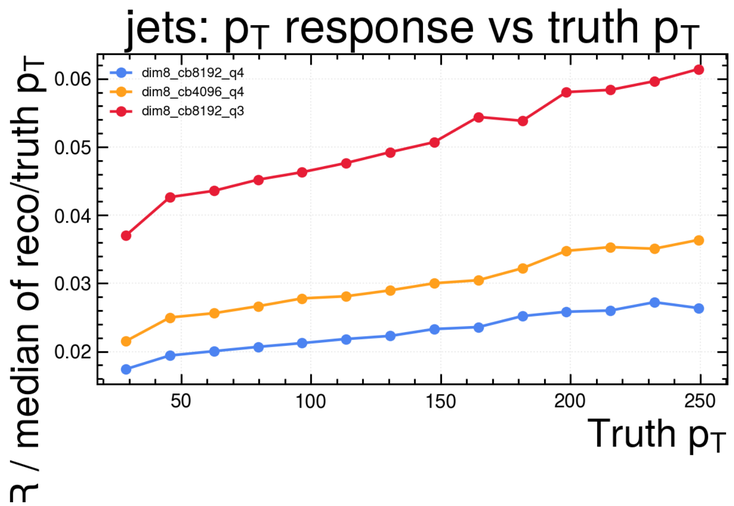
* Codebook utilisation per object/top scans.

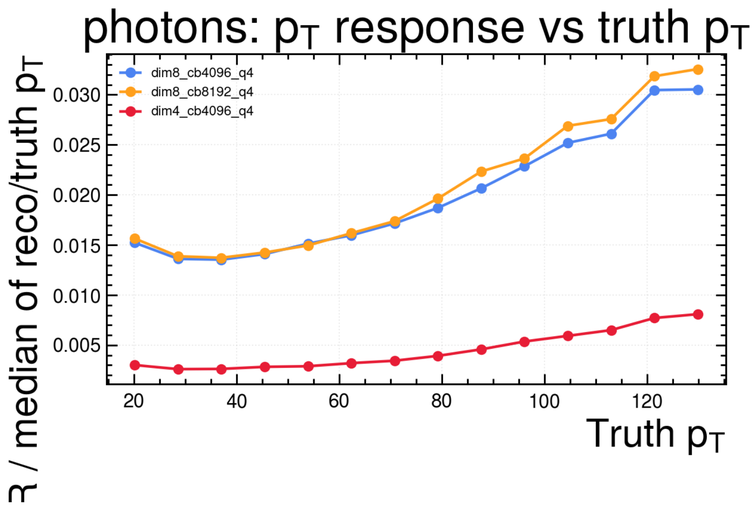


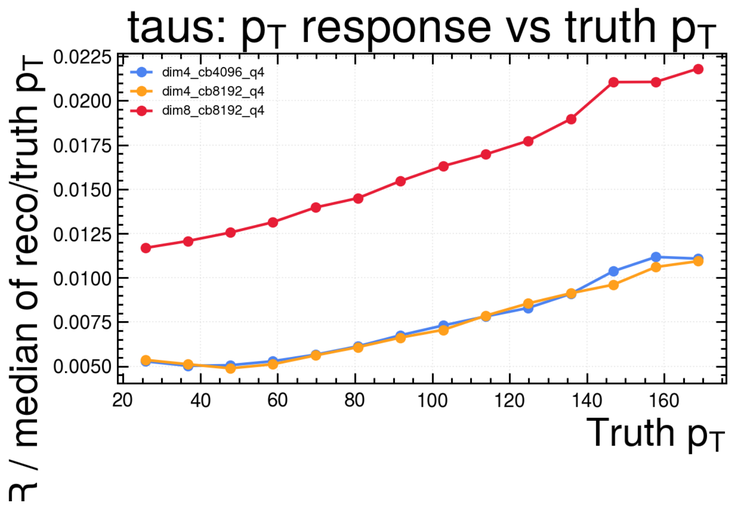
Summary Plots

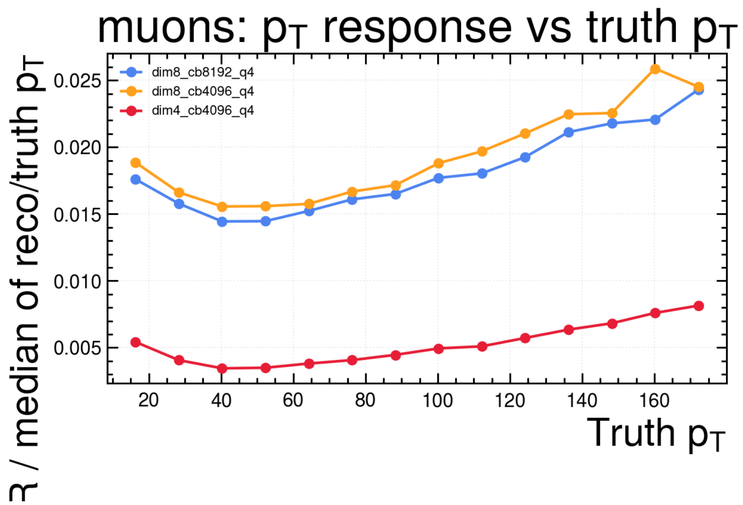
* binned pT resolution vs truth pT

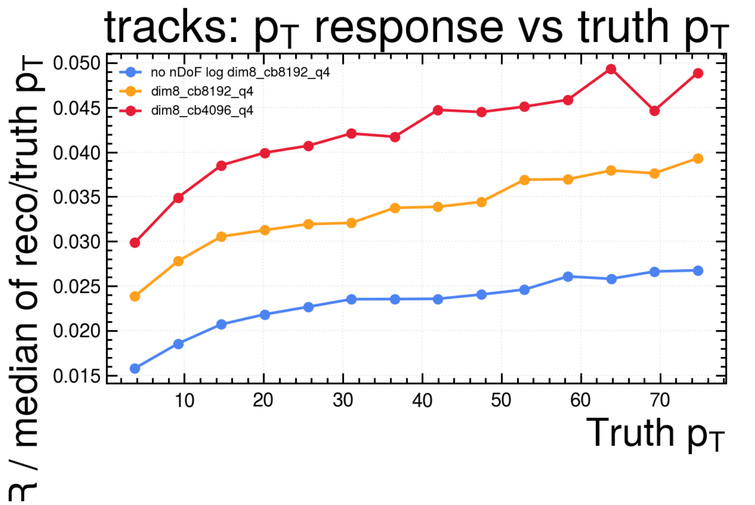






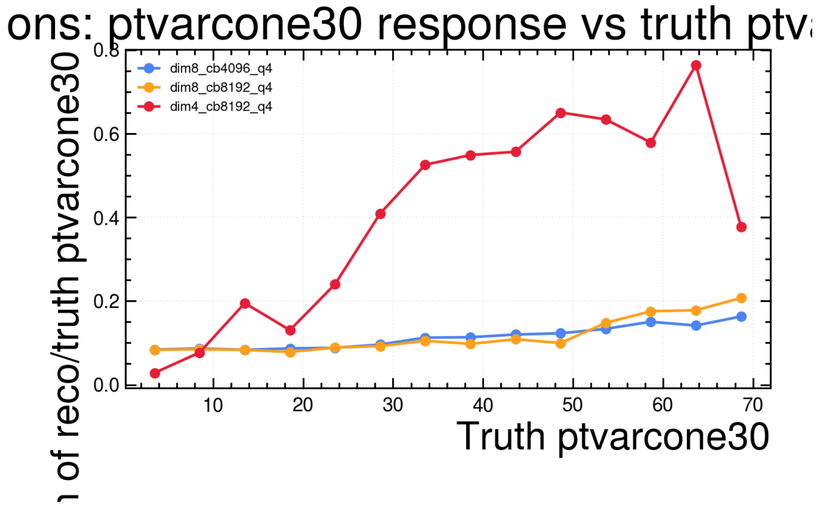


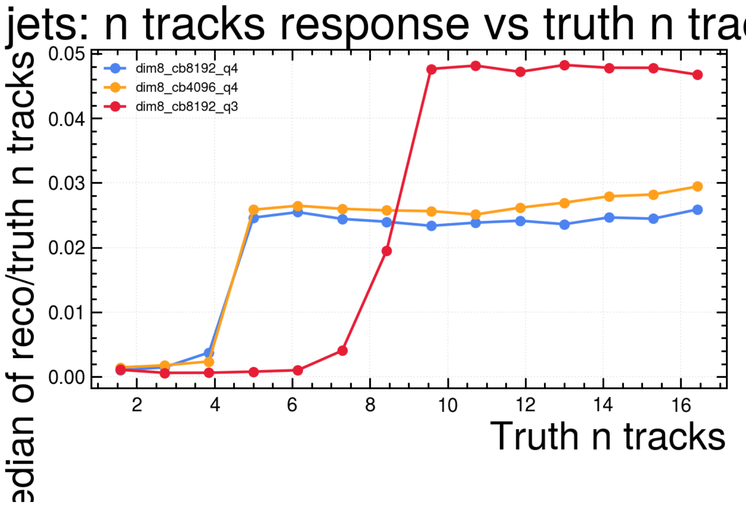


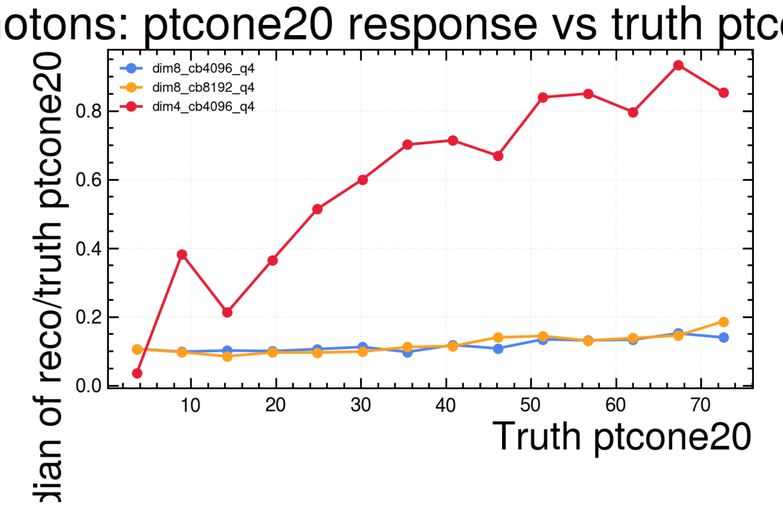


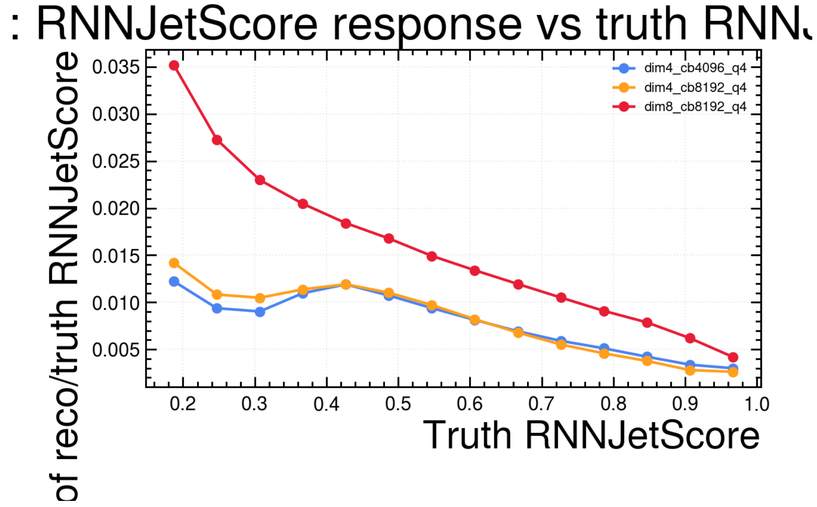
Summary Plots

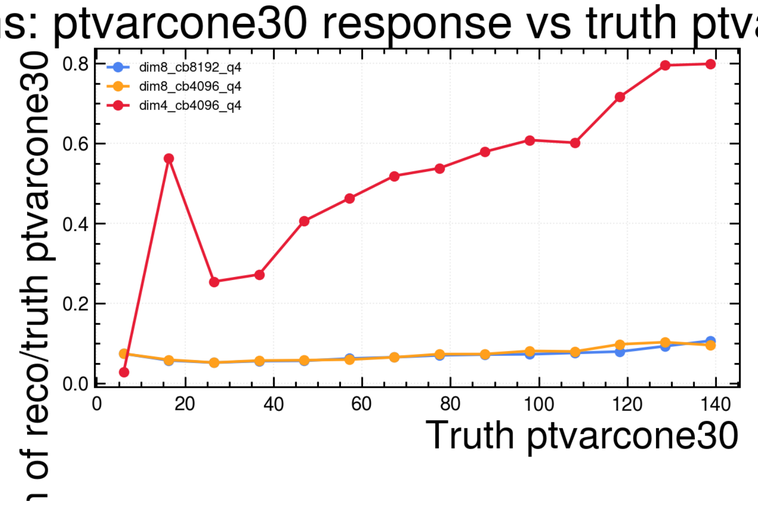
* Binned features resolution vs truth

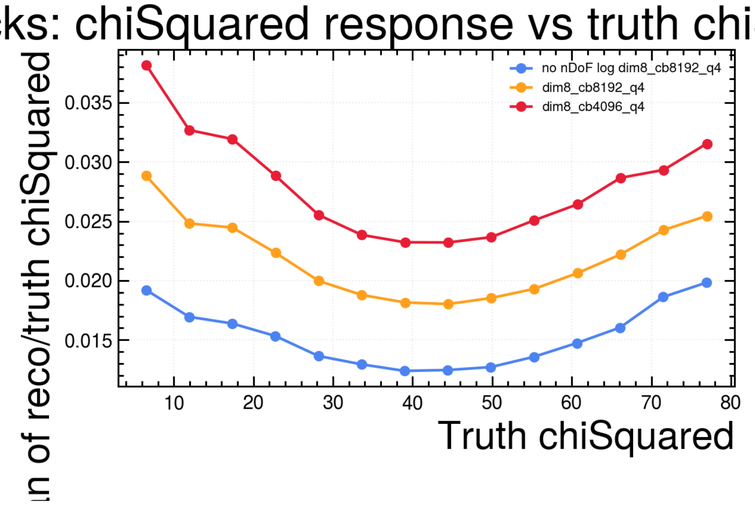












High-level variables after creating event-level tokeniser