Brookhaven National Laboratory

Treasure event level-tokenisation

Merve Nazlim Agaras

15.7.26

* Data+MC training showed inconsistent results
  - Jets has better reconstruction with Data+MC, the rest of the objects have the opposite behaviour.
  - Tracks are particularly worse —> Reason found: # of batches were different wrt MC-only
  - Electrons, muons, taus, photons -> used epoch 0 as the best checkpoint in data+MC because there was a dip in validation loss, moved to last epoch check point. Reconstruction improved slightly but the issue is still there.

* Today: Investigation of why MC+Data training is worse in electrons.
  - Preprocessor check -> similar preprocessing for both MC and MC+Data
  - Using the last ckpt
  - Codebook usage
  - Assignment of codebook check
  - Continuous vs quantized Reconstruction
  - Train comparison
  - Sampling
  - Data-only training
  - Feature aware loss training

“

Jets

Jets

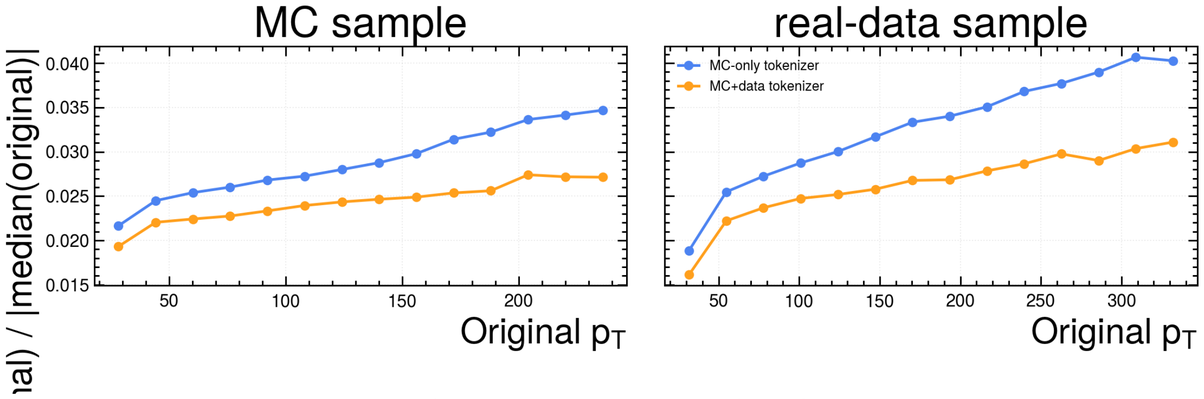
On MC samples

metric	MC only	MC+data	better
mean codebook used	78.70%	96.69%	MC+data (22.8% higher usage)
pT RMSE	2.209	1.940	MC+data (12.2% lower)
eta RMSE	0.0909	0.0774	MC+data (14.8% lower)
mass RMSE	0.2571	0.4284	MC only (40.0% lower)
n_trk RMSE	0.2018	0.1410	MC+data (30.1% lower)
GN2_pc RMSE	0.00838	0.00514	MC+data (38.7% lower)

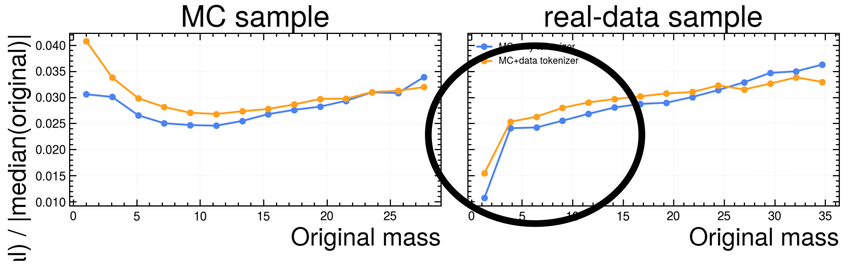
On data samples

metric	MC only	MC+data	better
mean codebook used	78.59%	96.59%	MC+data (22.9% higher usage)
pT RMSE	2.870	2.294	MC+data (20.1% lower)
eta RMSE	0.1000	0.0787	MC+data (21.3% lower)
mass RMSE	0.2978	0.4683	MC only (36.4% lower)
n_trk RMSE	0.2157	0.1350	MC+data (37.4% lower)
GN2_pc RMSE	0.00823	0.00509	MC+data (38.2% lower)

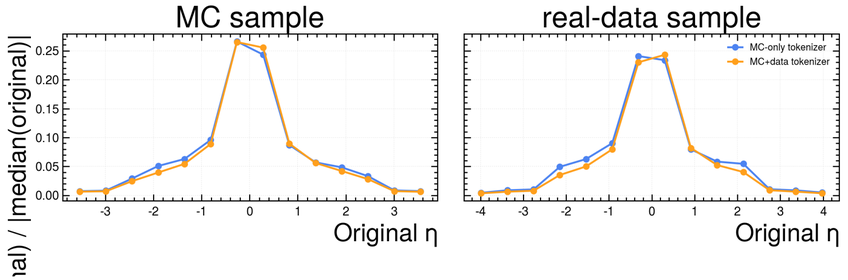
jets: p_T resolution



jets: mass resolution

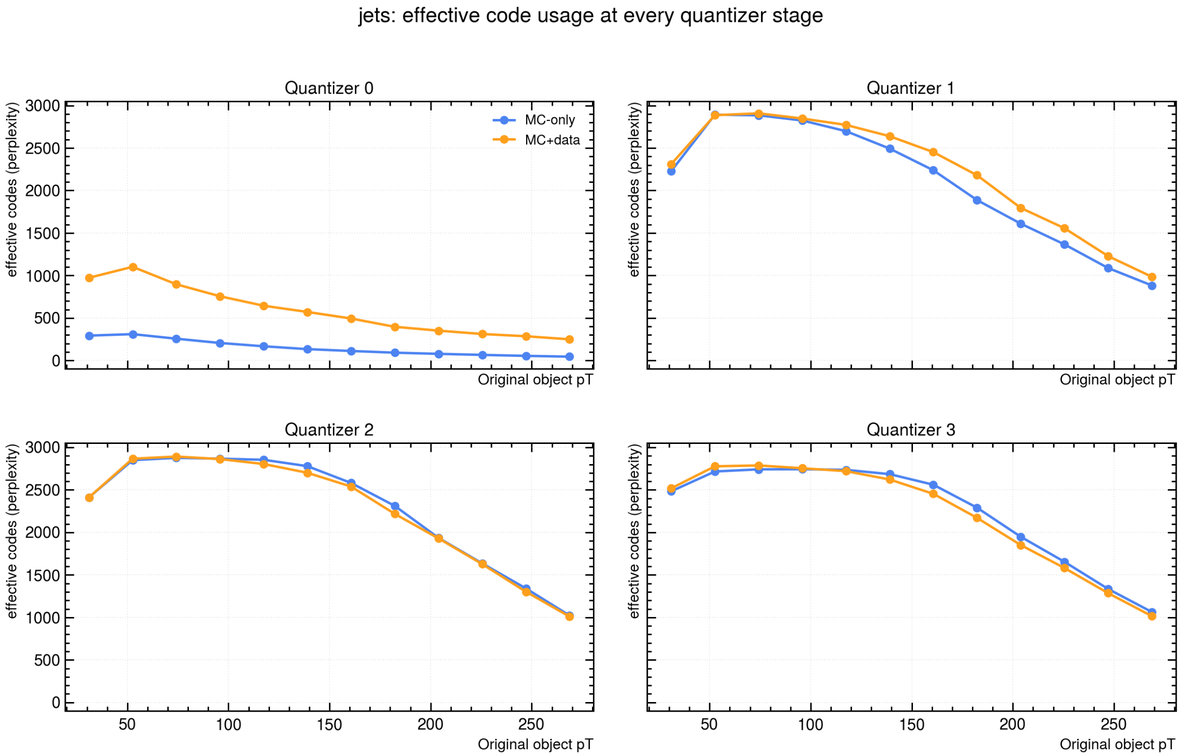


jets: η resolution



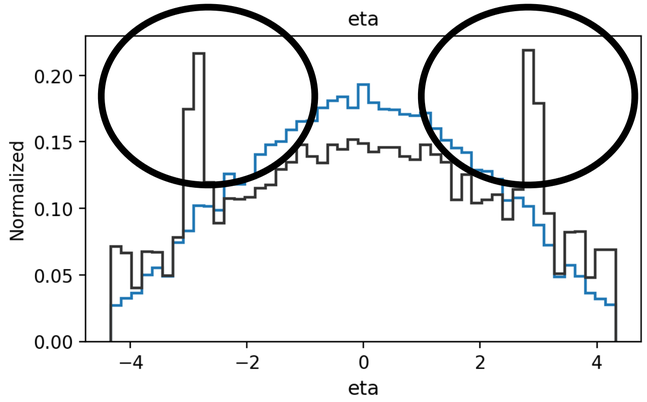
MC+data improves the features

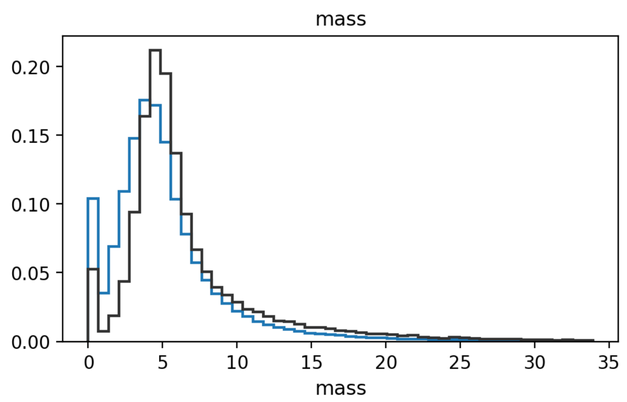
Used codes

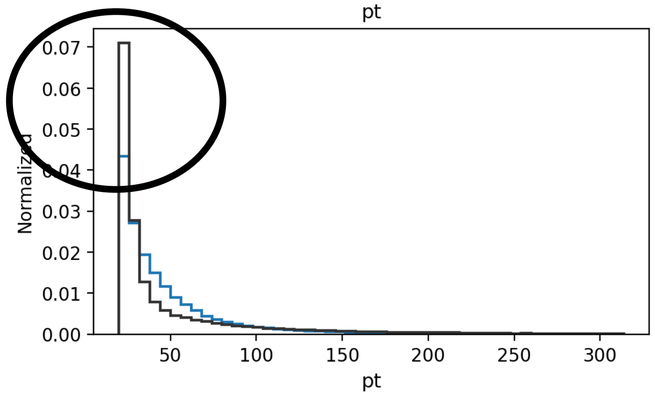


Mixed sample

common/jets/ –







- MC-only q0: 640 codes used—> usage increase driven by q0 improvement
- MC+data q0: 3583 codes used

“

Tracks

Tracks

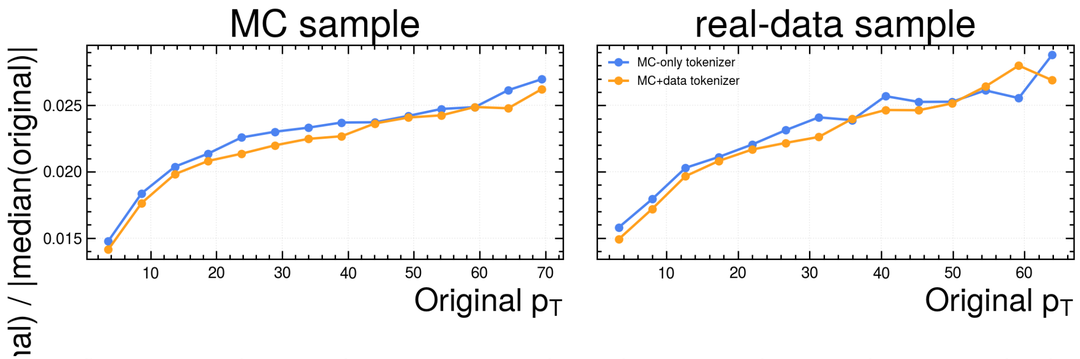
On MC samples

metric	MC only	MC+data	better
mean codebook used	79.01%	81.75%	MC+data (3.5% higher usage)
pT RMSE	5.720	5.244	MC+data (8.3% lower)
eta RMSE	0.0160	0.0162	MC only (1.1% lower)
phi RMSE	0.0313	0.0319	MC only (1.7% lower)
d0 RMSE	0.00574	0.00593	MC only (3.2% lower)
z0 RMSE	0.0858	0.0867	MC only (1.0% lower)
nDoF RMSE	0.2172	0.2198	MC only (1.2% lower)

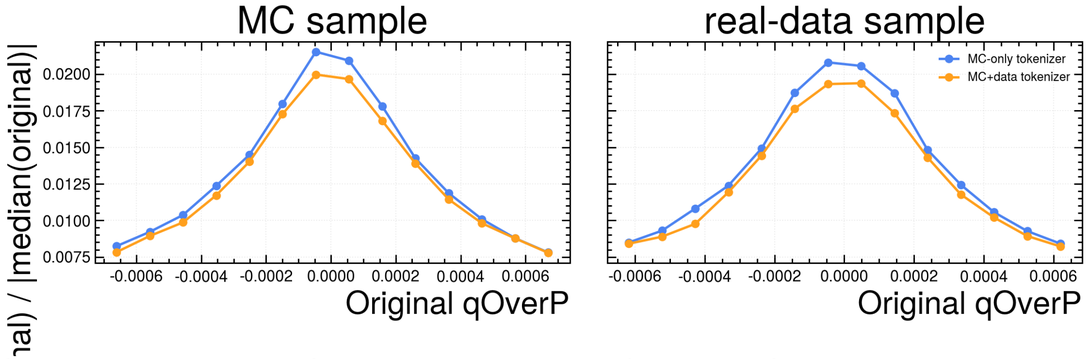
On data samples

metric	MC only	MC+data	better
mean codebook used	78.68%	81.40%	MC+data (3.5% higher usage)
pT RMSE	8.357	6.330	MC+data (24.3% lower)
eta RMSE	0.0150	0.0148	MC+data (1.3% lower)
phi RMSE	0.0307	0.0300	MC+data (2.4% lower)
d0 RMSE	0.00529	0.00523	MC+data (1.1% lower)
z0 RMSE	0.0853	0.0843	MC+data (1.1% lower)
nDoF RMSE	0.2036	0.2017	MC+data (0.9% lower)

tracks: p_T resolution



tracks: qOverP resolution



After matching the training batch size, the earlier track degradation disappears. MC+data improves track pT and most of the variables are similar.

“

Electrons

Studies:

- *Using the last ckpt*
- *Codebook usage*
- *Assignment of codebook check*
- *Continuous vs quantized Reconstruction*
- *Train comparison*
- *Sampling*
- *Data-only training*
- *Feature aware loss training*

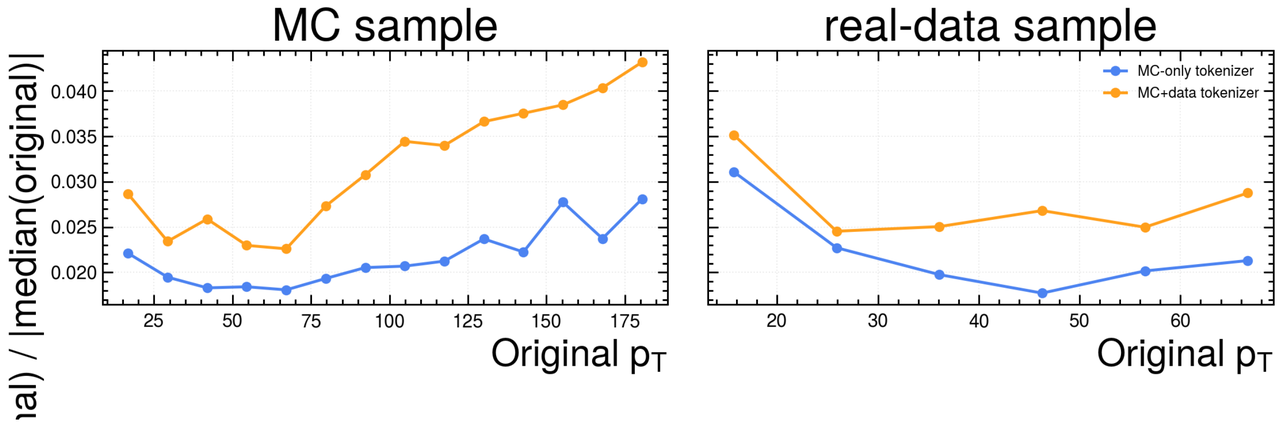
Electrons

On MC samples

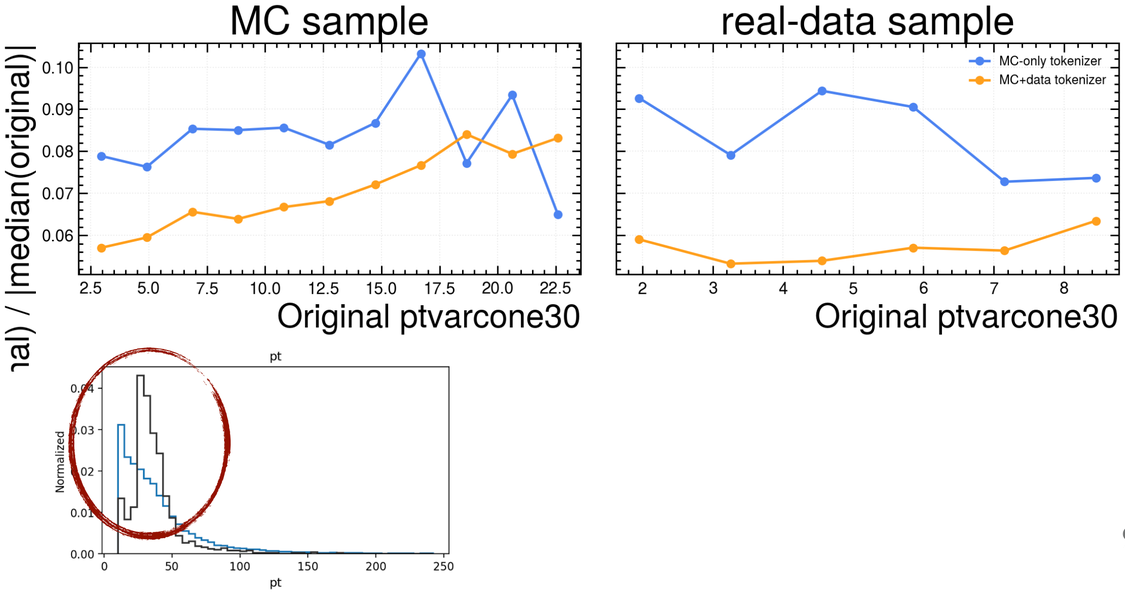
metric	MC only	MC+data	better
mean codebook used	38.16%	24.07%	MC only (58.5% higher usage)
pT RMSE	3.030	3.278	MC only (7.6% lower)
eta RMSE	0.0414	0.0461	MC only (10.2% lower)
charge RMSE	0.00674	0.0179	MC only (62.3% lower)
ptvarcone30 RMSE	1.999	2.222	MC only (10.1% lower)

On data samples

metric	MC only	MC+data	better
mean codebook used	36.73%	23.39%	MC only (57.0% higher usage)
pT RMSE	2.711	2.621	MC+data (3.3% lower)
eta RMSE	0.0442	0.0465	MC only (5.1% lower)
charge RMSE	0.00091	0.00017	MC+data (81.3% lower)
ptvarcone30 RMSE	0.6175	0.8448	MC only (26.9% lower)

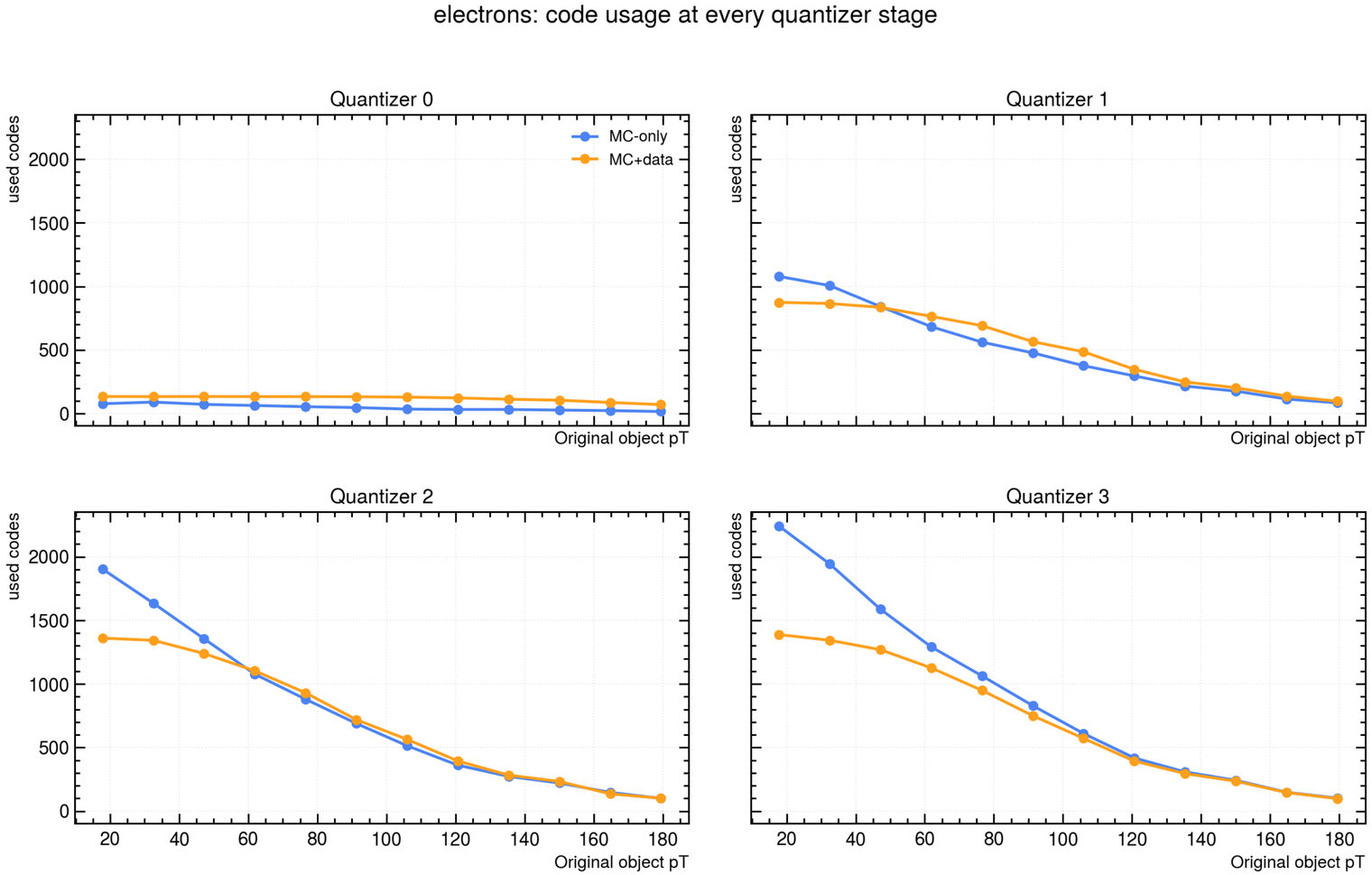


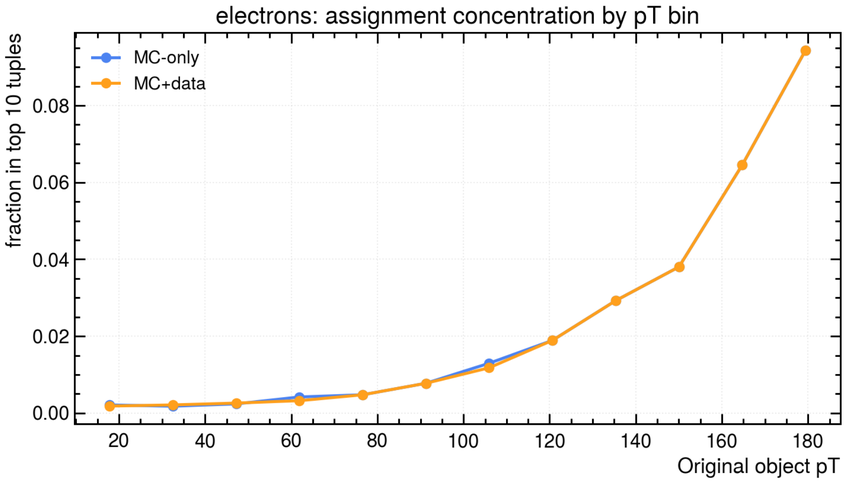
electrons: ptvarcone30 resolution



Unlike jets and corrected tracks, the full-feature MC+data electron tokenizer uses fewer codes and reconstructs most variables less precisely.

Used codebooks in MC samples



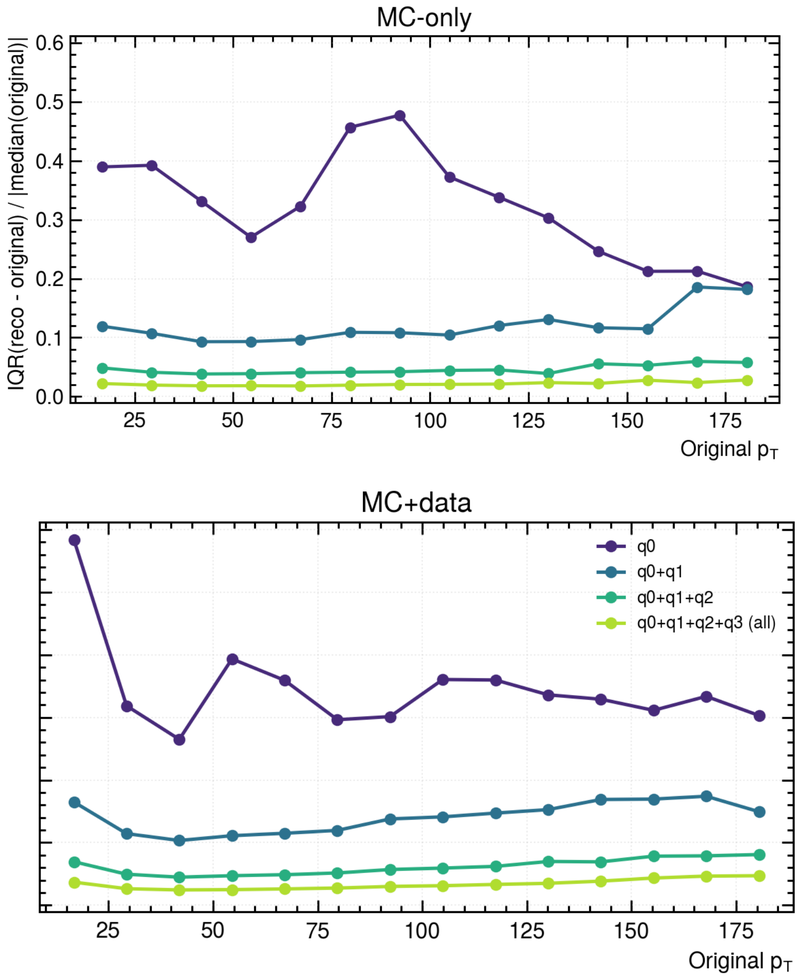


* Assignment of codebooks in each training is similar

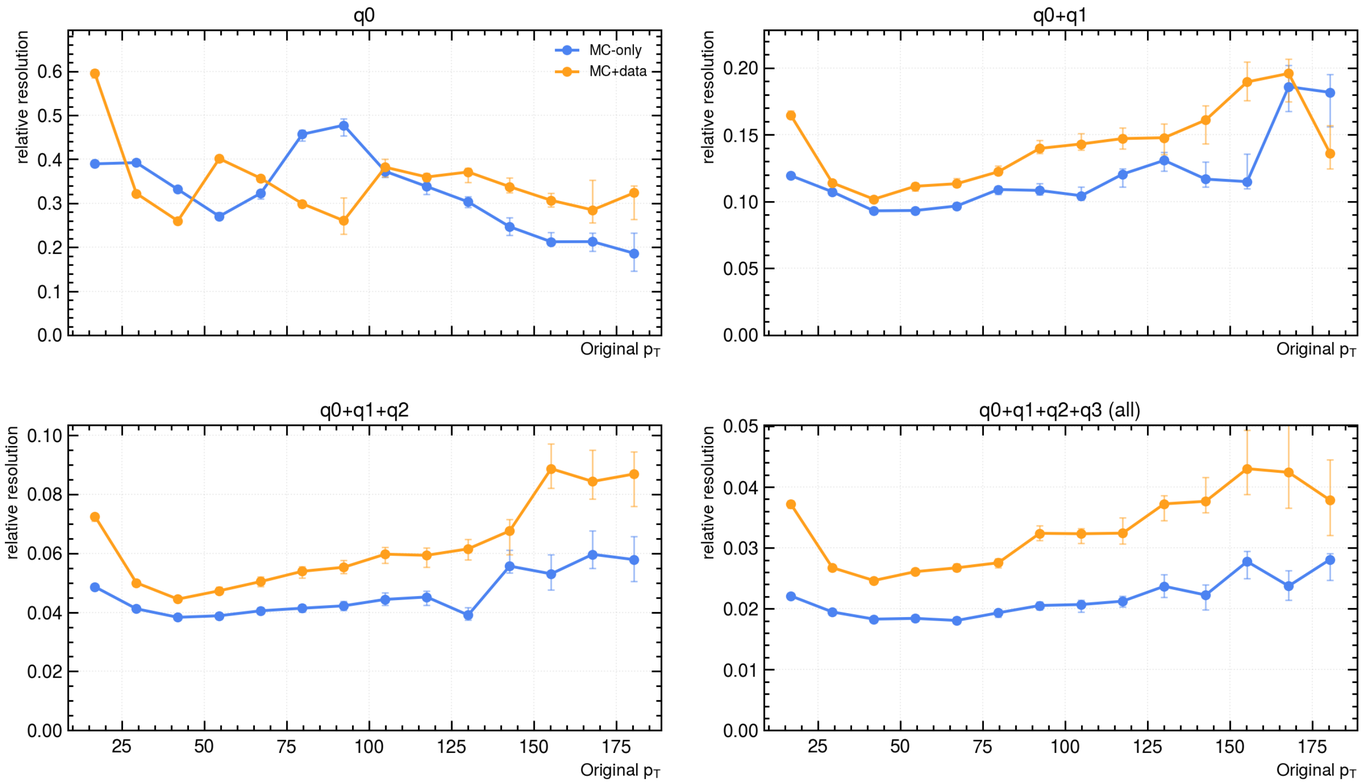
* MC+data uses more coarse q0 codes, capturing broad MC/data differences.
* MC+data uses fewer effective codes in later q2/q3 stages.
* Those later quantizers provide the fine corrections needed for precise reconstruction.
* Consequently, orange remains worse after every added stage.

Quantizer reconstructions

electrons: cumulative quantizer reconstruction of p_T

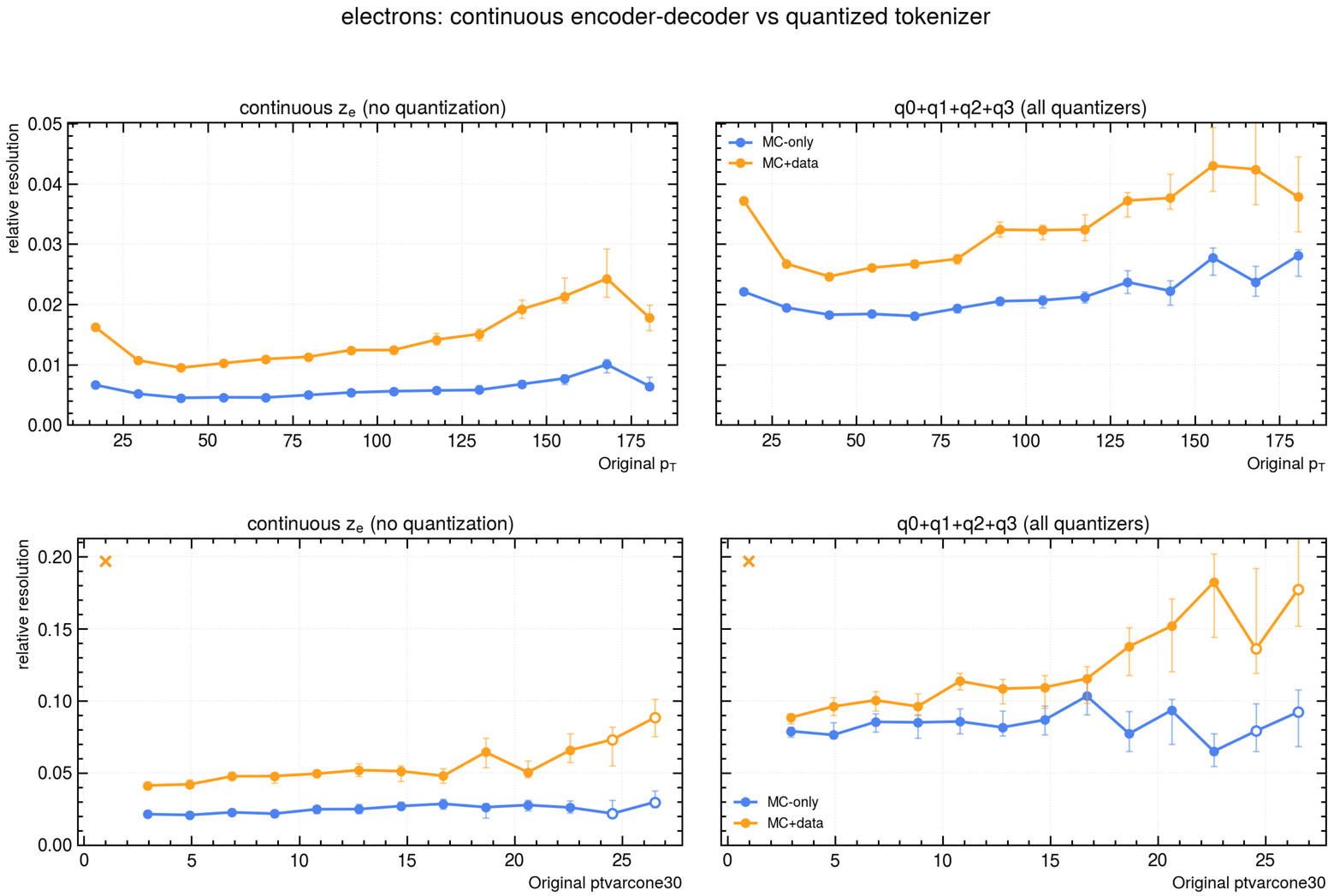


electrons: MC-only vs MC+data at each cumulative quantizer stage



* Adding quantisers improving the reconstruction.
* but the MC+data model remains worse after all four stages.
* Increasing the number of residual stages alone would not close the gap.

Continuous vs Quantized Reconstruction



- The main problem is already present before quantization.
- Continuous z_e (encoder output): MC+data is roughly 2–3× worse than MC-only.
- Full quantization degrades both models further by a similar amount.
- Therefore, reduced codebook usage is **secondary**, not the root cause.
- The mixed distribution is broader and harder, with unchanged model capacity.
- The objective optimizes the average over MC and data, potentially sacrificing pT precision.

```
z_q, indices, commit_loss, z_e = model.encode_with_encoder_output(batch)

reco_continuous = model.decode(z_e, batch)  # no quantization
reco_quantized = model.decode(z_q, batch)   # all q0–q3
```

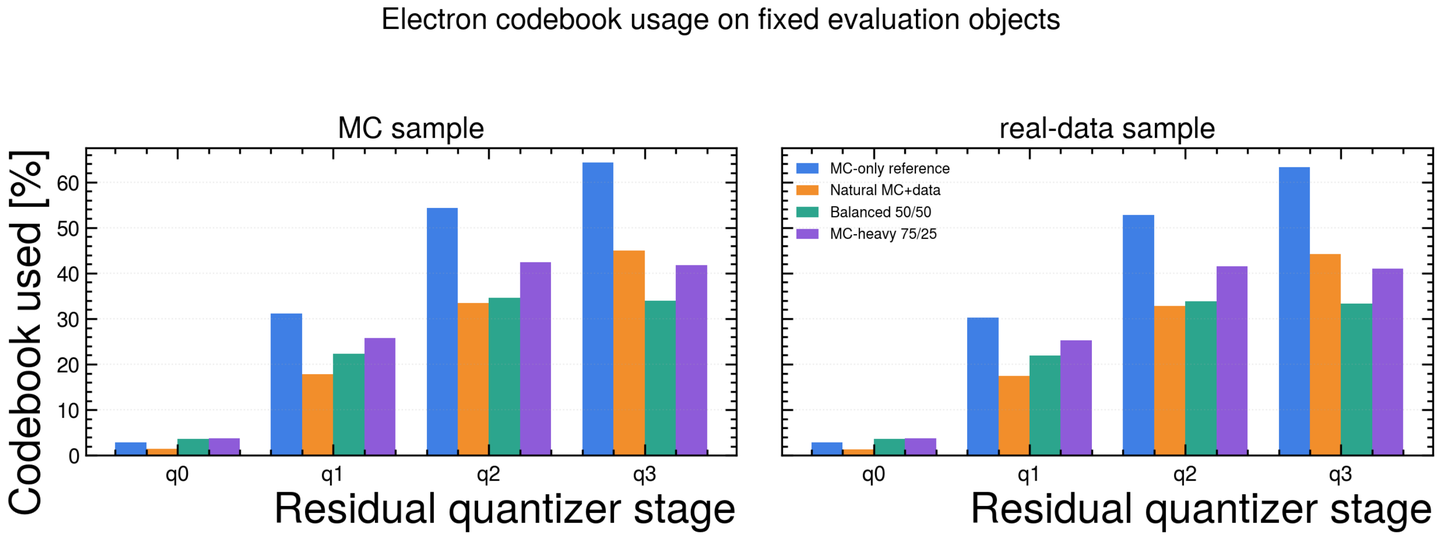
Data/MC difference inputs

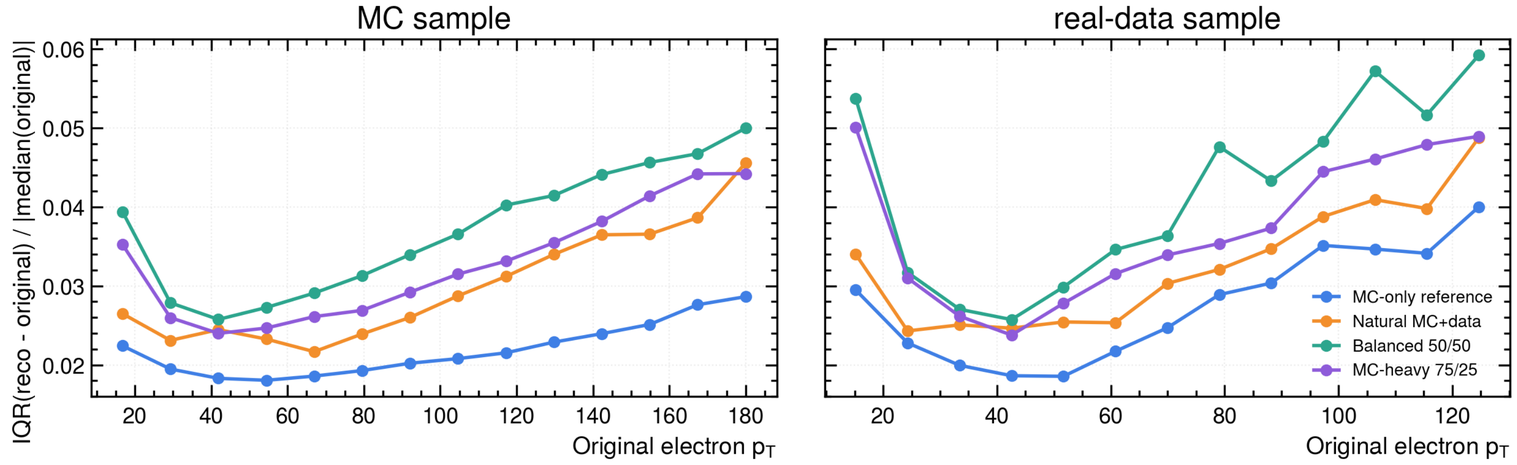
feature	MC median	data median	MC IQR	data IQR	data/MC IQR	KS distance
pt	39.962	33.909	40.087	14.851	0, 37	0, 235
eta	0.0036065	0.013849	1.7596	2.5052	1.424	0, 090
LHMedium	1	1	0	0	nan	0, 061
LHTight	1	1	0	1	nan	0, 080
ptvarcone30	0	0	0	1.4766	nan	0, 137
topoetcone20	0.39982	0.6412	1.8306	2.2303	1.218	0, 071

feature	MC median	data median	MC IQR	data IQR	data/MC IQR	KS distance
pt	48.001	30.565	47.033	36.155	769	0, 251
eta	0.0059123	0.023255	2.2365	3.731	1.668	0, 125
mass	6.3725	5.5612	4.9493	4.233	855	0, 105
n_trk	5	3	6	6	1.000	0, 237
GN2_pc	0.1309	0.11477	0.15042	0.1182	786	0, 180
GN2_pu	0.61848	0.72523	0.73579	0.12353	168	0, 268

* Data adds structured, lower-dimensional population to jets.
* While electron data introduce competing zero-inflated, categorical variables.

Different sampling

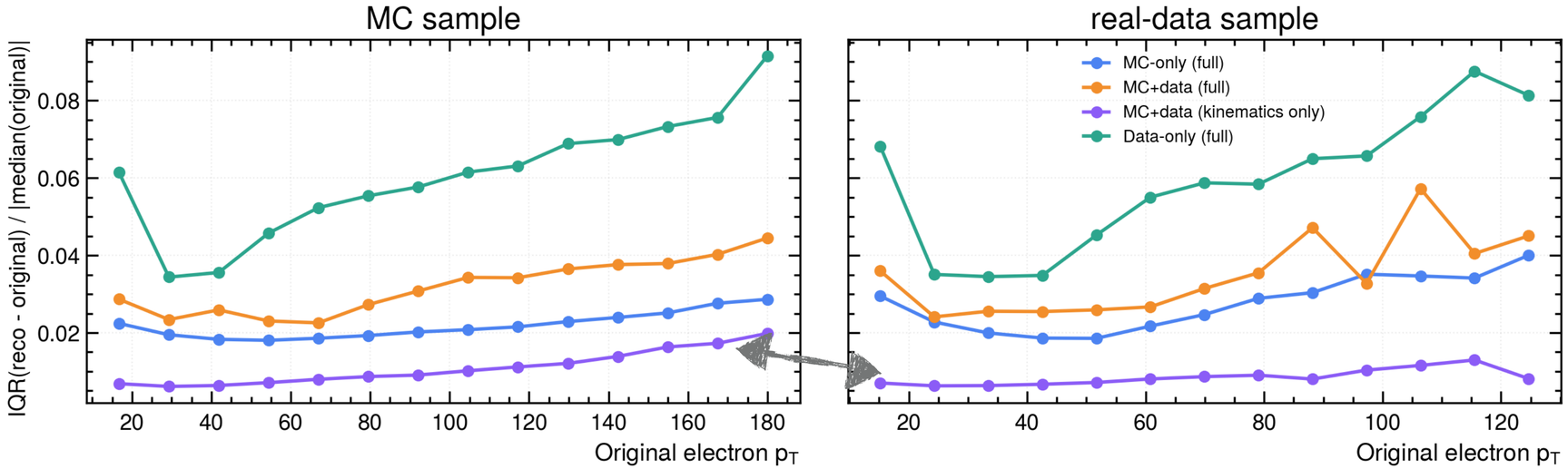




* Tried 4 different sampling setups in each batch
  - 50/50 MC and Data
  - 25/75 data/MC
  - MC only
  - Natural MC+Data (random)

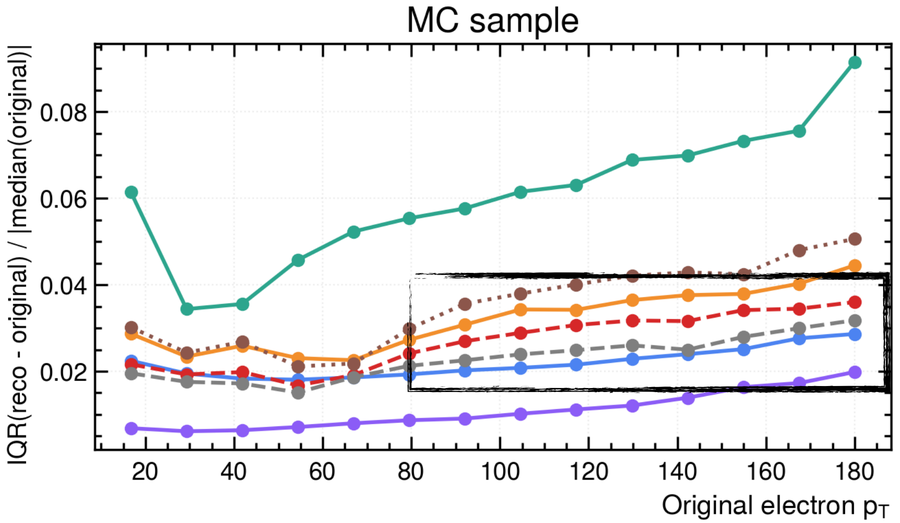
* Changing the MC/data mixture does not recover the precision.
* 50/50 performs worst, while MC-heavy training is below MC-only, so simple domain reweighting is not sufficient.

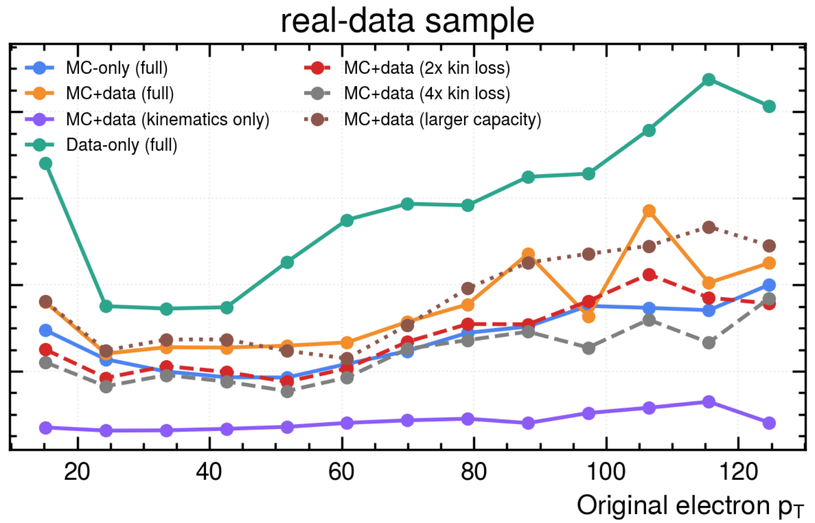
Feature controls

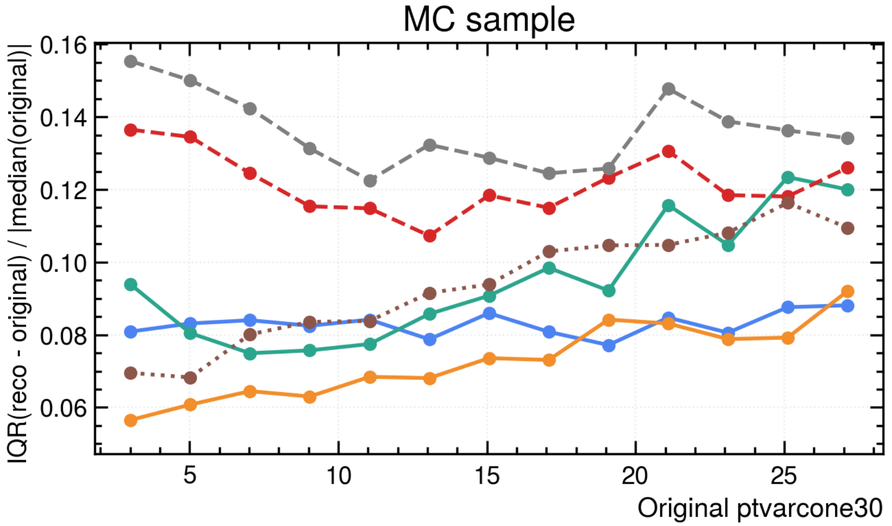


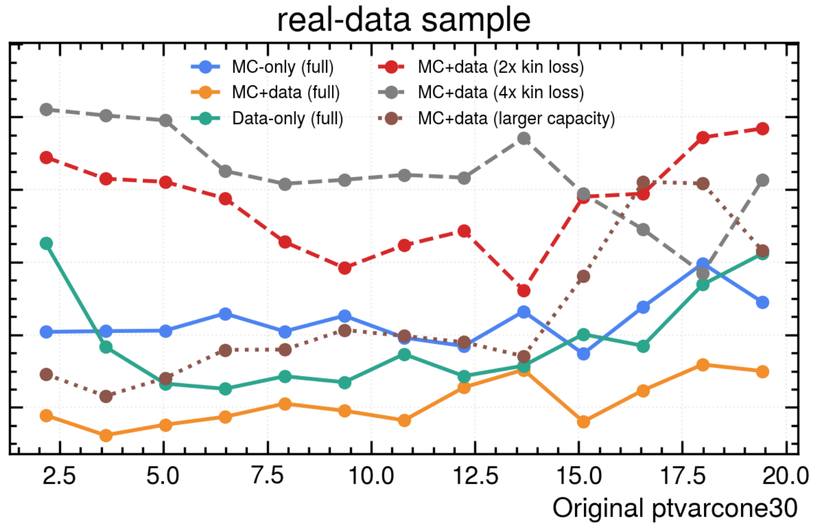
* MC+data kinematics-only training improves pT reconstruction by **~50–81% across both MC and data**, indicating competition between kinematic and isolation/ID features within the tokenizer capacity. Therefore, adding data itself is not the problem. The degradation appears when the model jointly reconstructs the all the feature groups.
* Things to try:
  - Keep the MC+data kinematics-only tokenizer and train a separate auxiliary tokenizer for isolation and ID variables and combine both token groups in the event sequence eg. Electron → 4 kinematic tokens + 4 auxiliary tokens → distinct group/type ID → event transformer
    - Cons: longer event sequence
  - Hierarchical Feature-Group Tokenization: The inner transformer learns how to combine the feature-group tokens belonging to one object into a single object representation Feature-group tokens → inner transformer → one object vector → event transformer
    - Cons: adds a new trainable aggregation stage
  - Feature-aware loss function

Feature-aware loss weighting









* The model normalizes weights to mean one:
  - **2×:** kinematic weights become 1.33; other weights become 0.667.
  - **4×:** kinematic weights become 1.6 ; other weights become 0.4.

* Neither increased model capacity nor 2x–4x kinematic weighting reproduces the kinematics-only performance.
* But weighting is improving the pT reconstruction as expected.

Data/MC difference inputs

feature	MC median	data median	MC IQR	data IQR	data/MC IQR	KS distance
pt	39.962	33.909	40.087	14.851	0, 37	0, 235
eta	0.0036065	0.013849	1.7596	2.5052	1.424	0, 090
LHMedium	1	1	0	0	nan	0, 061
LHTight	1	1	0	1	nan	0, 080
ptvarcone30	0	0	0	1.4766	nan	0, 137
topoetcone20	0.39982	0.6412	1.8306	2.2303	1.218	0, 071

feature	MC median	data median	MC IQR	data IQR	data/MC IQR	KS distance
pt	48.001	30.565	47.033	36.155	769	0, 251
eta	0.0059123	0.023255	2.2365	3.731	1.668	0, 125
mass	6.3725	5.5612	4.9493	4.233	855	0, 105
n_trk	5	3	6	6	1.000	0, 237
GN2_pc	0.1309	0.11477	0.15042	0.1182	786	0, 180
GN2_pu	0.61848	0.72523	0.73579	0.12353	168	0, 268

* Both jets and electrons show sizeable MC/data distribution shifts, yet jets improve while electrons degrade. Distribution shift alone therefore cannot explain the result.
* The important thing is that electron inputs mix continuous kinematics, binary ID flags, and zero-inflated isolation variables within one flat regression objective.

Conclusions and next steps

Conclusion

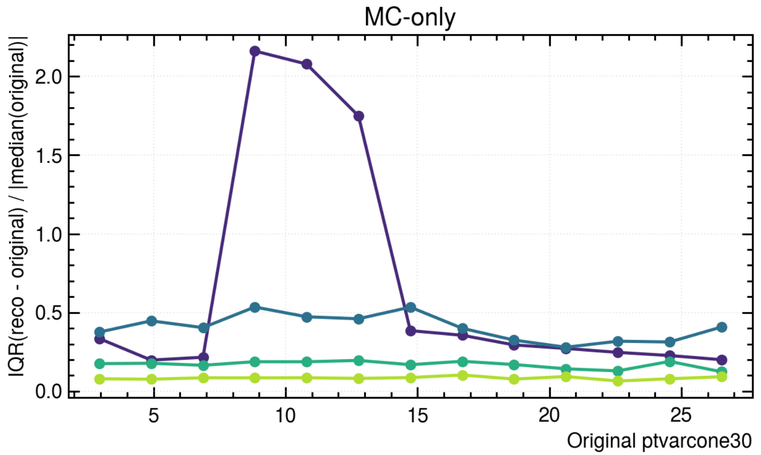
* Many checks have been performed to understand the MC-MC/Data training difference.
* For jets, MC+data improves overall reconstruction.
* For electrons, the degradation begins before quantization and is not resolved by checkpoint choice, sampling balance, increased capacity, or simple loss weighting.
* The kinematics-only MC+data tokenizer is much better than every full-feature model on both MC and real data.
* The studies point to competition between electron feature groups.
  - Feature competition exists in both trainings, but it becomes **more severe when MC and data are combined** because the feature relationships differ between the two.

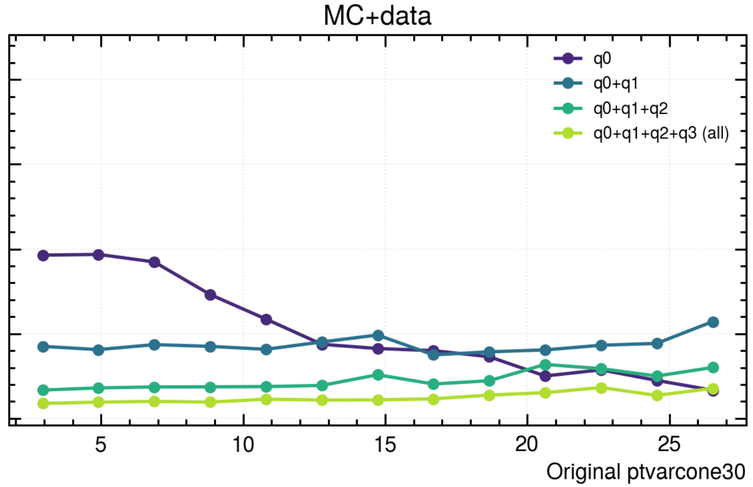
Next steps:

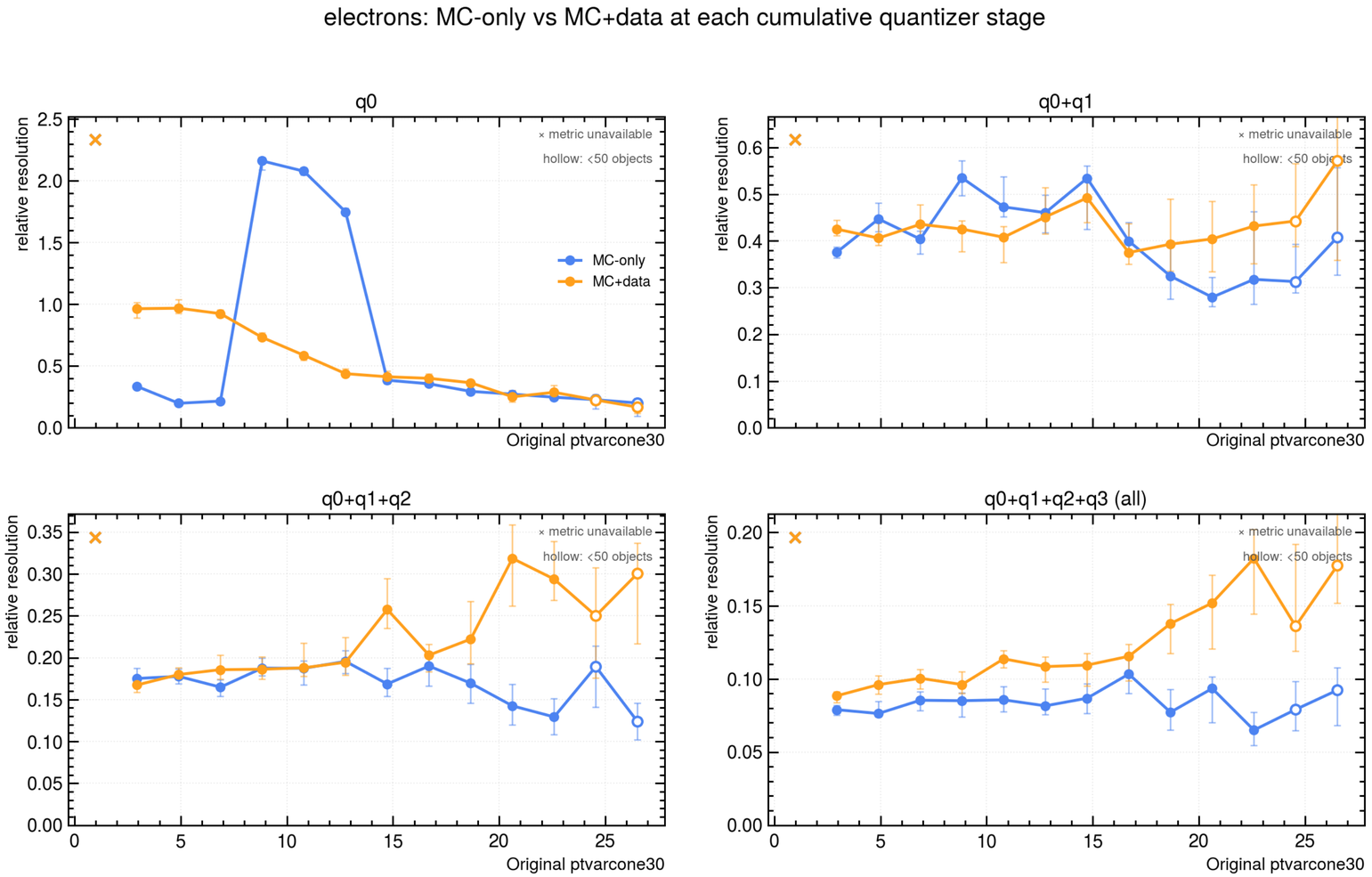
* Use regression losses for continuous kinematics and binary classification losses for ID flags.
* Test separate kinematic and ID/isolation token groups with distinct type IDs.
* Repeat the feature-group control for muons, photons, and taus.
* Validate different representations in downstream classification.

Backup

Quantizer reconstructions







* q0 is more stable in MC+Data
* MC+data must represent a broader, more heterogeneous distribution with the same model capacity.

Muons

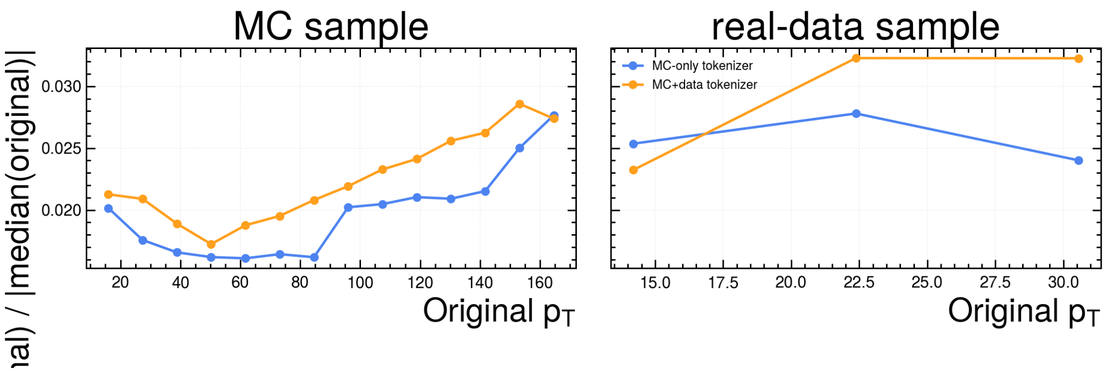
On MC samples

metric	MC only	MC+data	better
mean codebook used	45.05%	23.86%	MC only (88.8% higher usage)
pT RMSE	3.232	5.542	MC only (41.7% lower)
eta RMSE	0.0341	0.0358	MC only (4.8% lower)
charge RMSE	0.00562	0.0203	MC only (72.3% lower)
ptvarcone30 RMSE	5.337	8.708	MC only (38.7% lower)

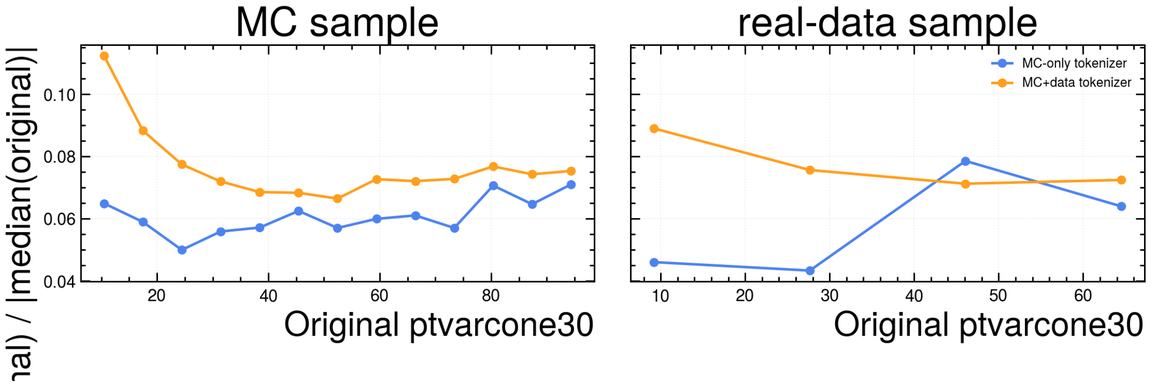
On data samples

metric	MC only	MC+data	better
mean codebook used	40.18%	23.23%	MC only (73.0% higher usage)
pT RMSE	3.801	4.057	MC only (6.3% lower)
eta RMSE	0.0733	0.0737	MC only (0.6% lower)
charge RMSE	0.0352	0.0436	MC only (19.3% lower)
ptvarcone30 RMSE	85.8	92.5	MC only (7.2% lower)

muons: p_T resolution



muons: ptvarcone30 resolution



Taus

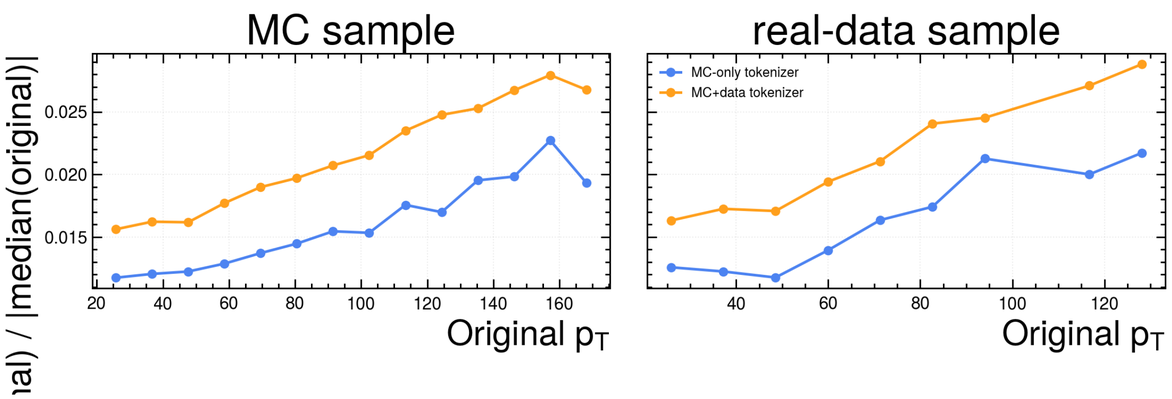
On MC samples

metric	MC only	MC+data	better
mean codebook used	33.01%	19.92%	MC only (65.8% higher usage)
pT RMSE	1.155	1.846	MC only (37.5% lower)
eta RMSE	0.0219	0.0281	MC only (21.9% lower)
charge RMSE	0.00067	0.00016	MC+data (76.1% lower)
NNDecayMode RMSE	0.0118	0.0115	MC+data (2.8% lower)

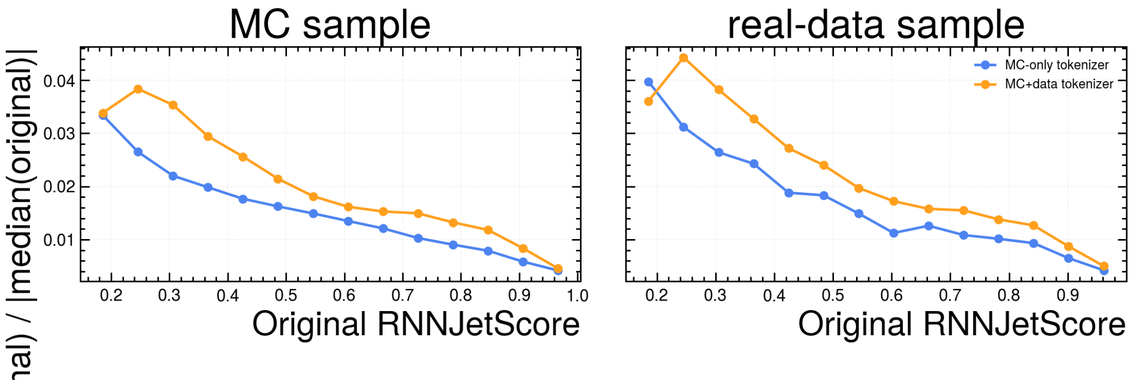
On data samples

metric	MC only	MC+data	better
mean codebook used	32.33%	19.59%	MC only (65.1% higher usage)
pT RMSE	1.568	1.930	MC only (18.8% lower)
eta RMSE	0.0262	0.0305	MC only (14.3% lower)
charge RMSE	0.00092	0.00013	MC+data (85.9% lower)
NNDecayMode RMSE	0.0212	0.0161	MC+data (24.3% lower)

taus: p_T resolution



taus: RNNJetScore resolution



Photons

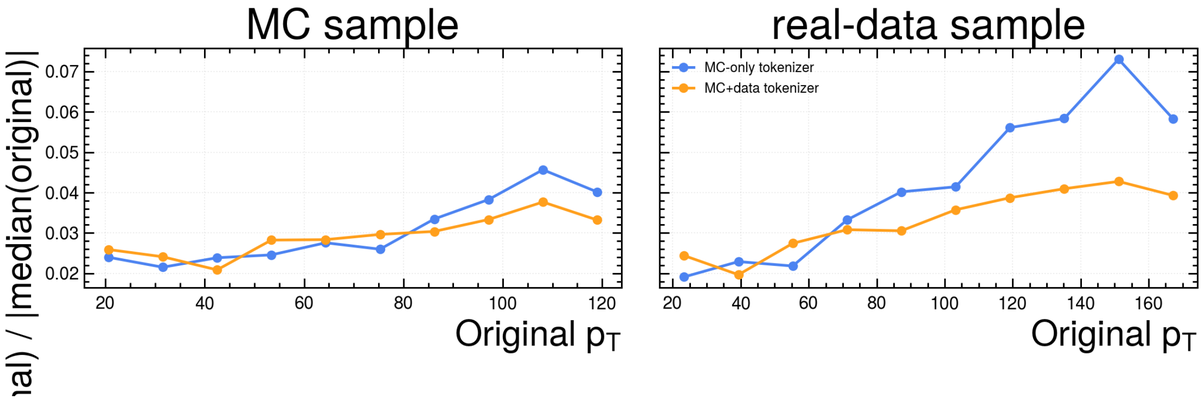
On MC samples

metric	MC only	MC+data	better
mean codebook used	40.42%	33.05%	MC only (22.3% higher usage)
pT RMSE	5.369	3.603	MC+data (32.9% lower)
eta RMSE	0.0632	0.0557	MC+data (11.7% lower)
isTight RMSE	0.00843	0.0156	MC only (46.1% lower)
ptcone20 RMSE	37.964	31.324	MC+data (17.5% lower)
topoetcone20 RMSE	1.485	1.127	MC+data (24.1% lower)
topoetcone40 RMSE	2.212	1.988	MC+data

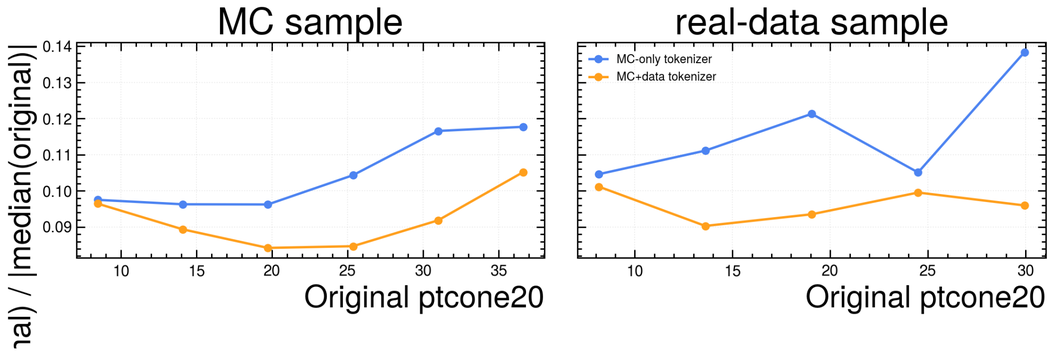
On data samples

metric	MC only	MC+data	better
mean codebook used	40.12%	32.81%	MC only (22.3% higher usage)
pT RMSE	7.028	4.962	MC+data (29.4% lower)
eta RMSE	0.0733	0.0565	MC+data (23.0% lower)
isTight RMSE	0.0122	0.0137	MC only (11.1% lower)
ptcone20 RMSE	7.019	4.476	MC+data (36.2% lower)
topoetcone20 RMSE	0.8364	0.6888	MC+data (17.7% lower)
topoetcone40 RMSE	3.775	2.209	MC+data (41.5% lower)

photons: p_T resolution



photons: ptcone20 resolution



Tracks

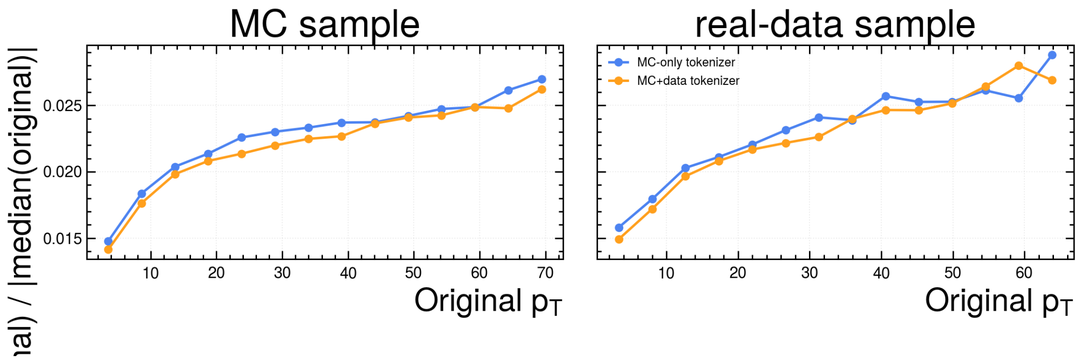
On MC samples

metric	MC only	MC+data	better
mean codebook used	79.01%	81.75%	MC+data (3.5% higher usage)
pT RMSE	5.720	5.244	MC+data (8.3% lower)
eta RMSE	0.0160	0.0162	MC only (1.1% lower)
phi RMSE	0.0313	0.0319	MC only (1.7% lower)
d0 RMSE	0.00574	0.00593	MC only (3.2% lower)
z0 RMSE	0.0858	0.0867	MC only (1.0% lower)
nDoF RMSE	0.2172	0.2198	MC only (1.2% lower)

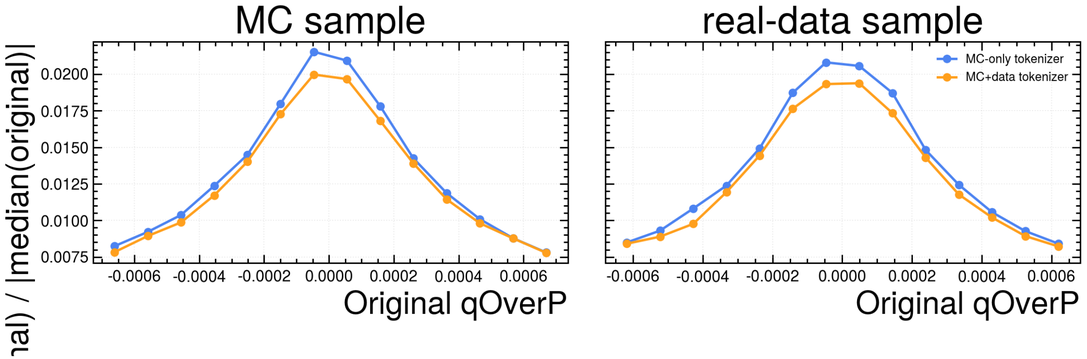
On data samples

metric	MC only	MC+data	better
mean codebook used	78.68%	81.40%	MC+data (3.5% higher usage)
pT RMSE	8.357	6.330	MC+data (24.3% lower)
eta RMSE	0.0150	0.0148	MC+data (1.3% lower)
phi RMSE	0.0307	0.0300	MC+data (2.4% lower)
d0 RMSE	0.00529	0.00523	MC+data (1.1% lower)
z0 RMSE	0.0853	0.0843	MC+data (1.1% lower)
nDoF RMSE	0.2036	0.2017	MC+data (0.9% lower)

tracks: p_T resolution



tracks: qOverP resolution



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/RNNJetScore
- atlas/taus/RNNEleScore

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

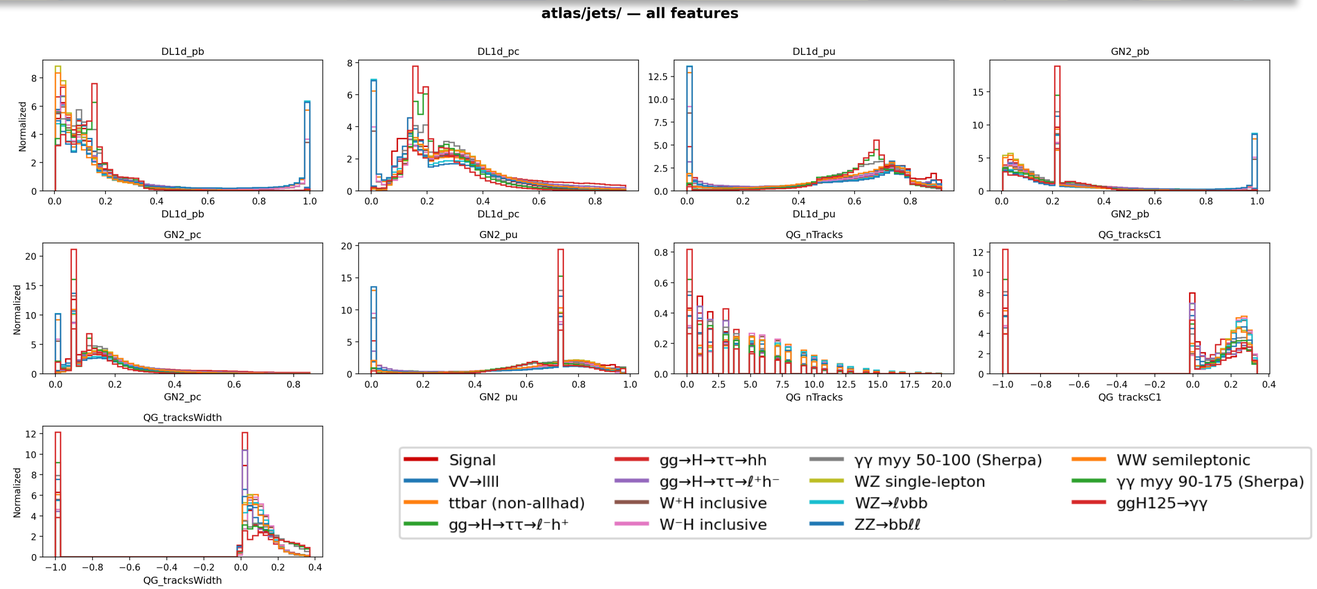
- common/tracks/pt
- common/tracks/eta
- common/tracks/phi
- common/tracks/d0
- common/tracks/z0
- atlas/tracks/qOverP
- atlas/tracks/chiSquared
- atlas/tracks/nDoF

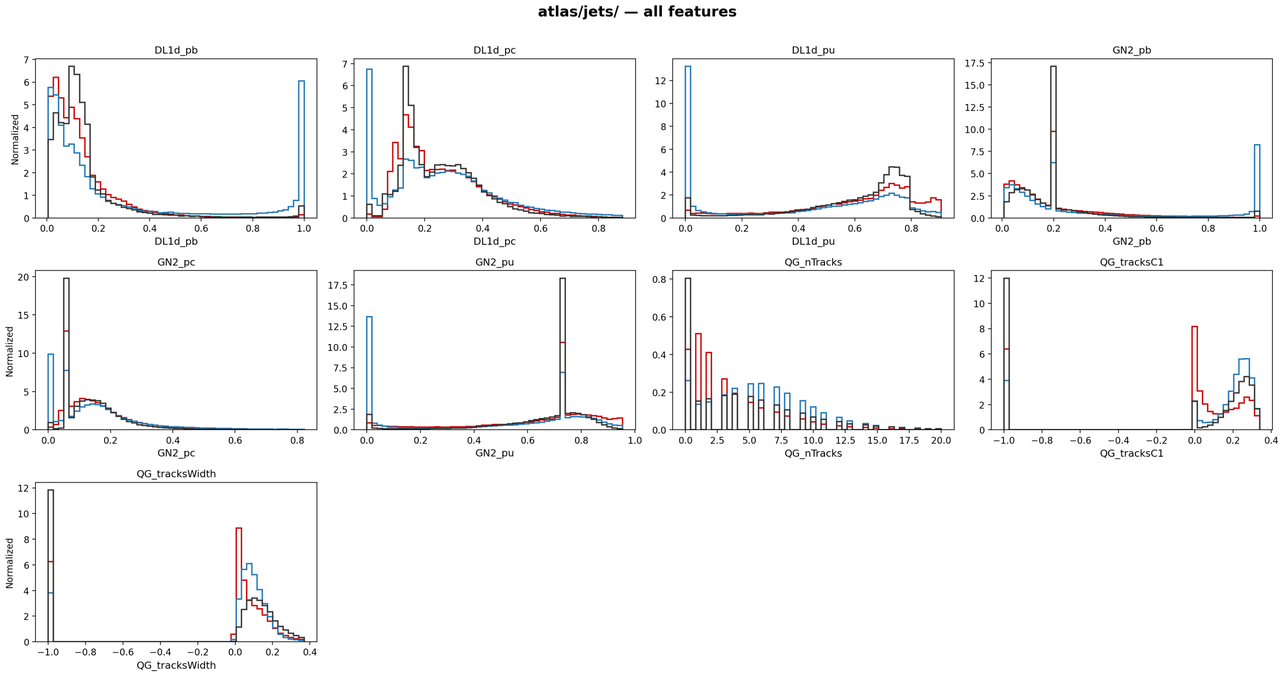
- common/photons/pt
- common/photons/eta
- common/photons/phi
- atlas/photons/isTight
- atlas/photons/ptcone20
- atlas/photons/topoetcone20
- atlas/photons/topoetcone40

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





Capacity Scans

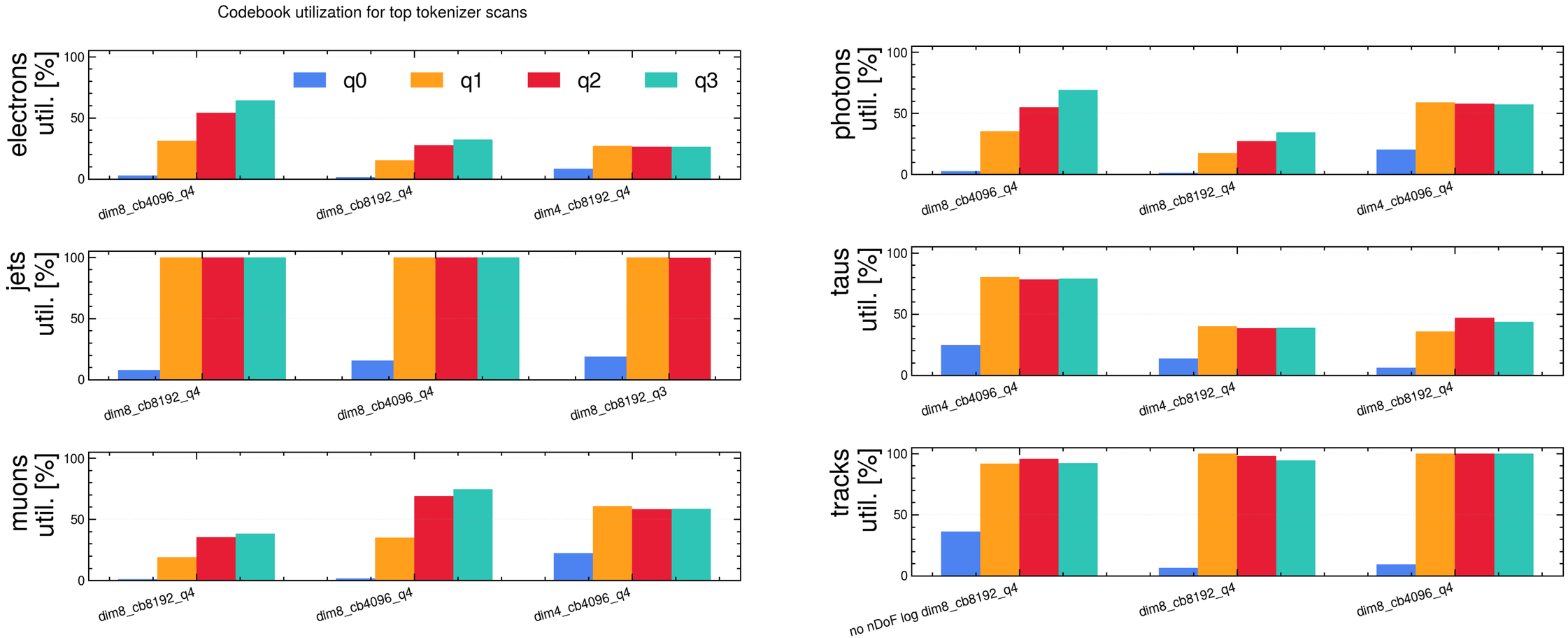
- Scanned:
  - codebook dimension: 4, 8, 16 for jets; 4/8 for others.
  - codebook sizes: mostly 4096 and 8192.
  - quantizers: mainly q4 after q4 showed better reconstruction.

- 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

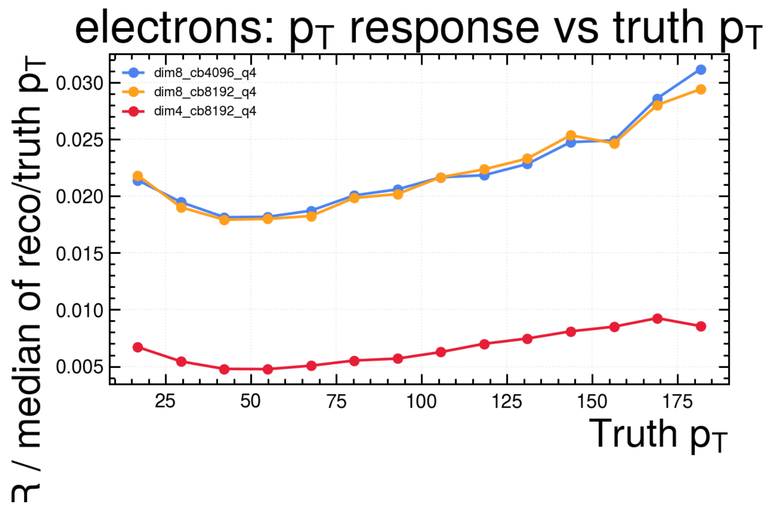
Summary Plots

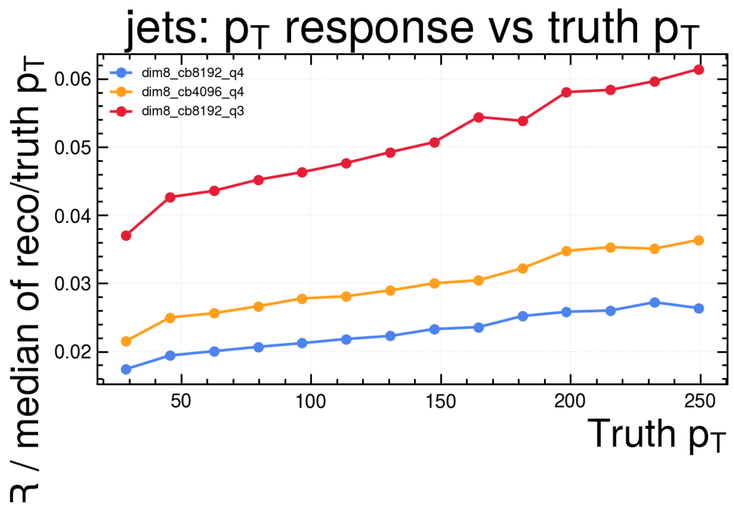
* Codebook utilisation per object/top scans.

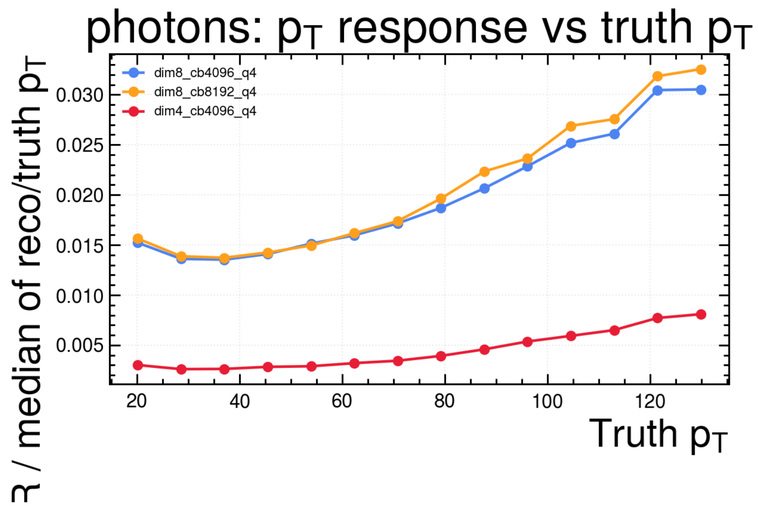


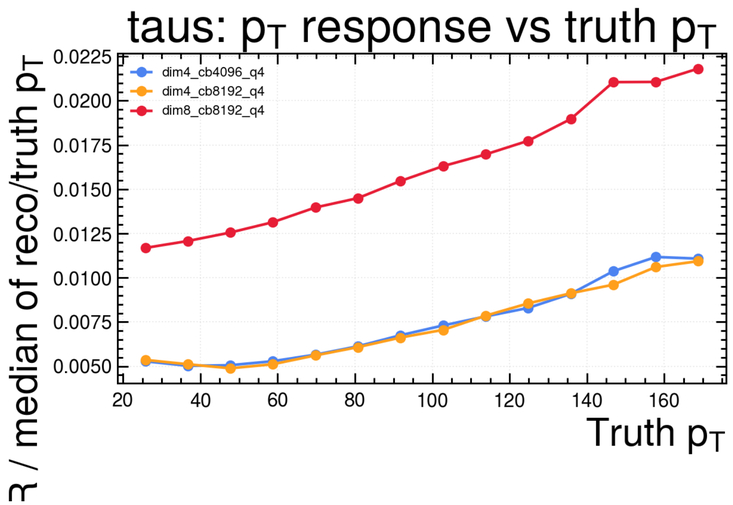
Summary Plots

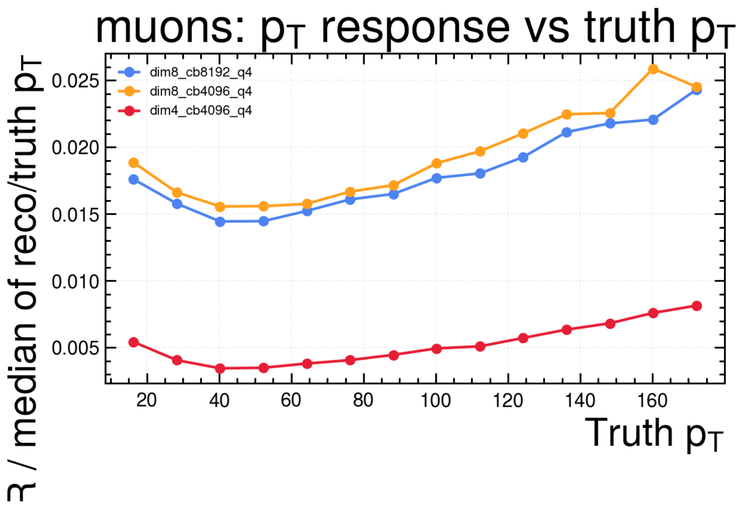
* binned pT resolution vs truth pT

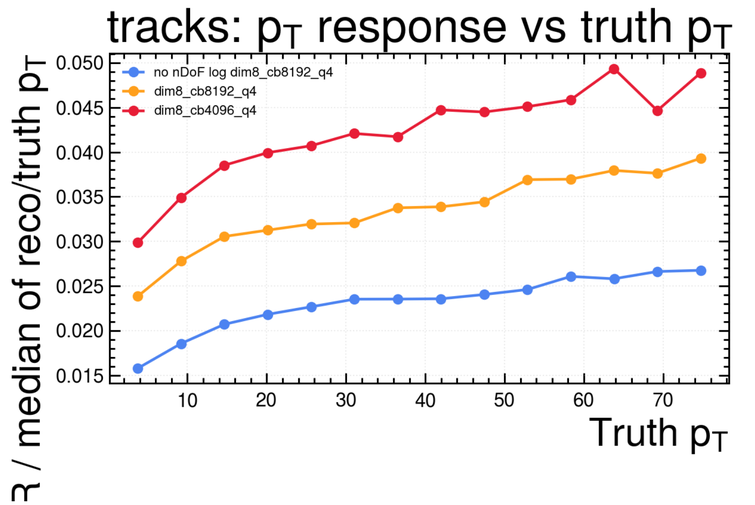






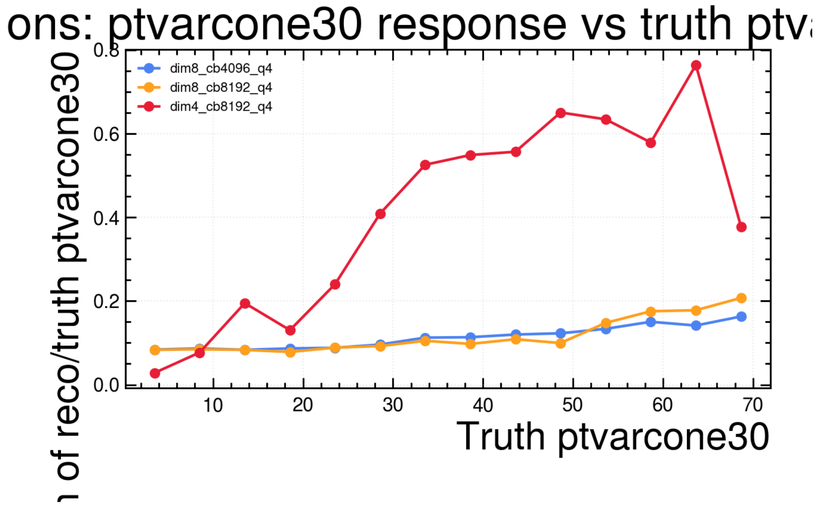


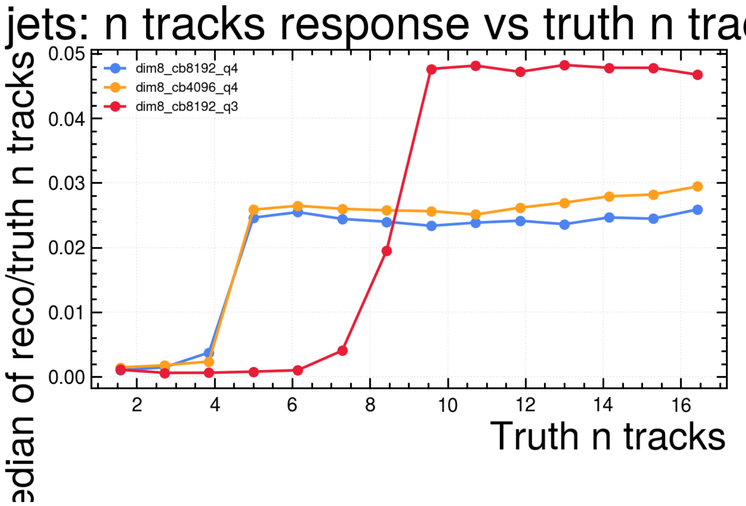


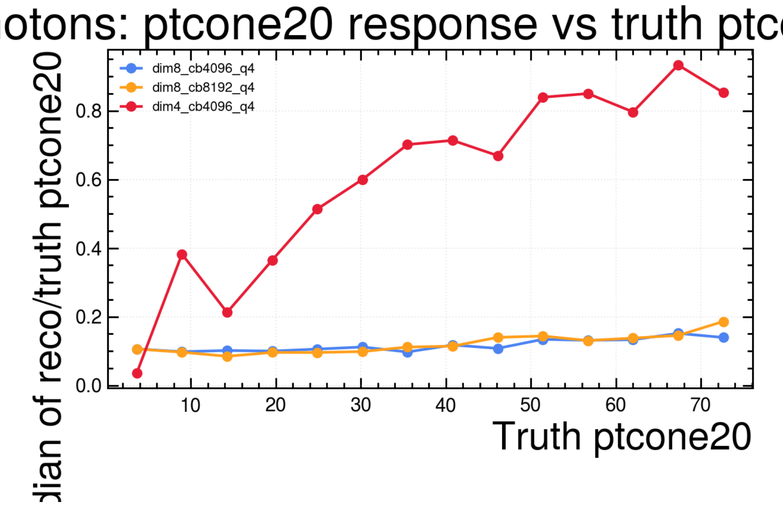


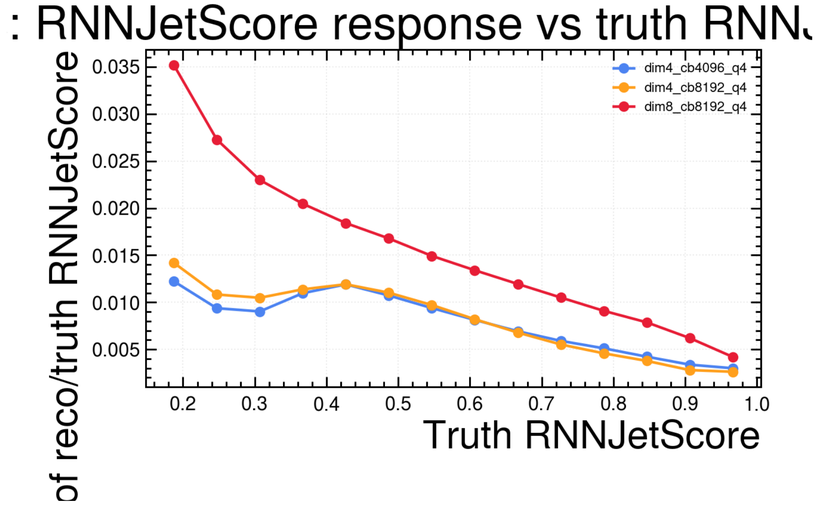
Summary Plots

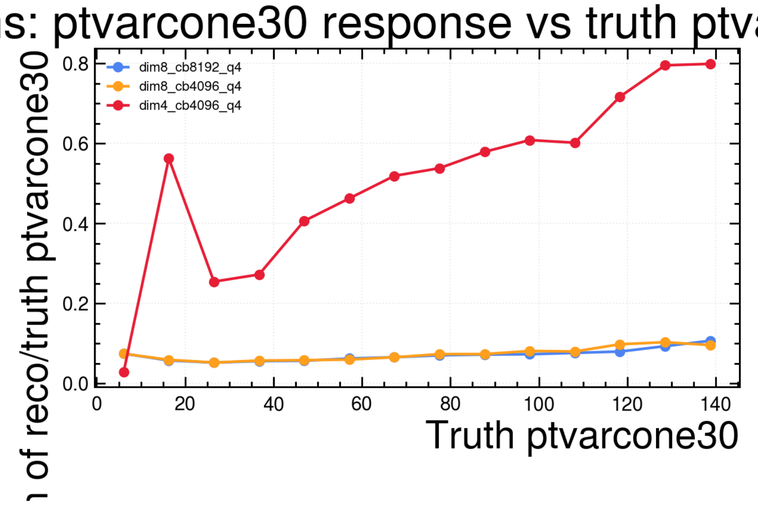
* Binned features resolution vs truth

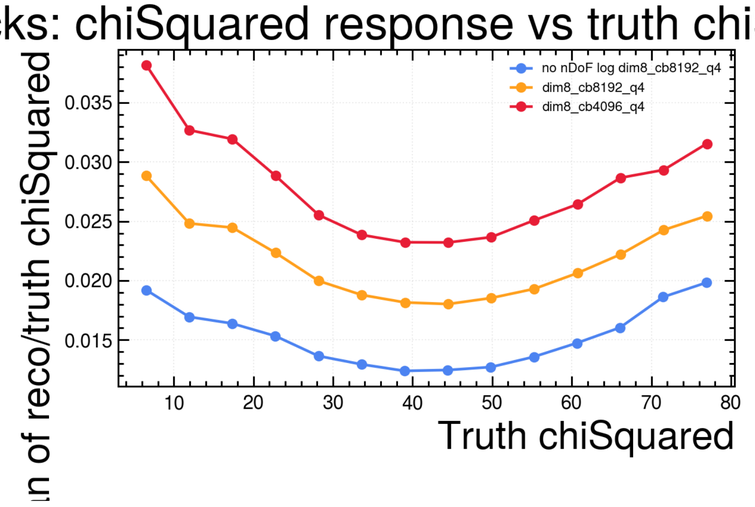




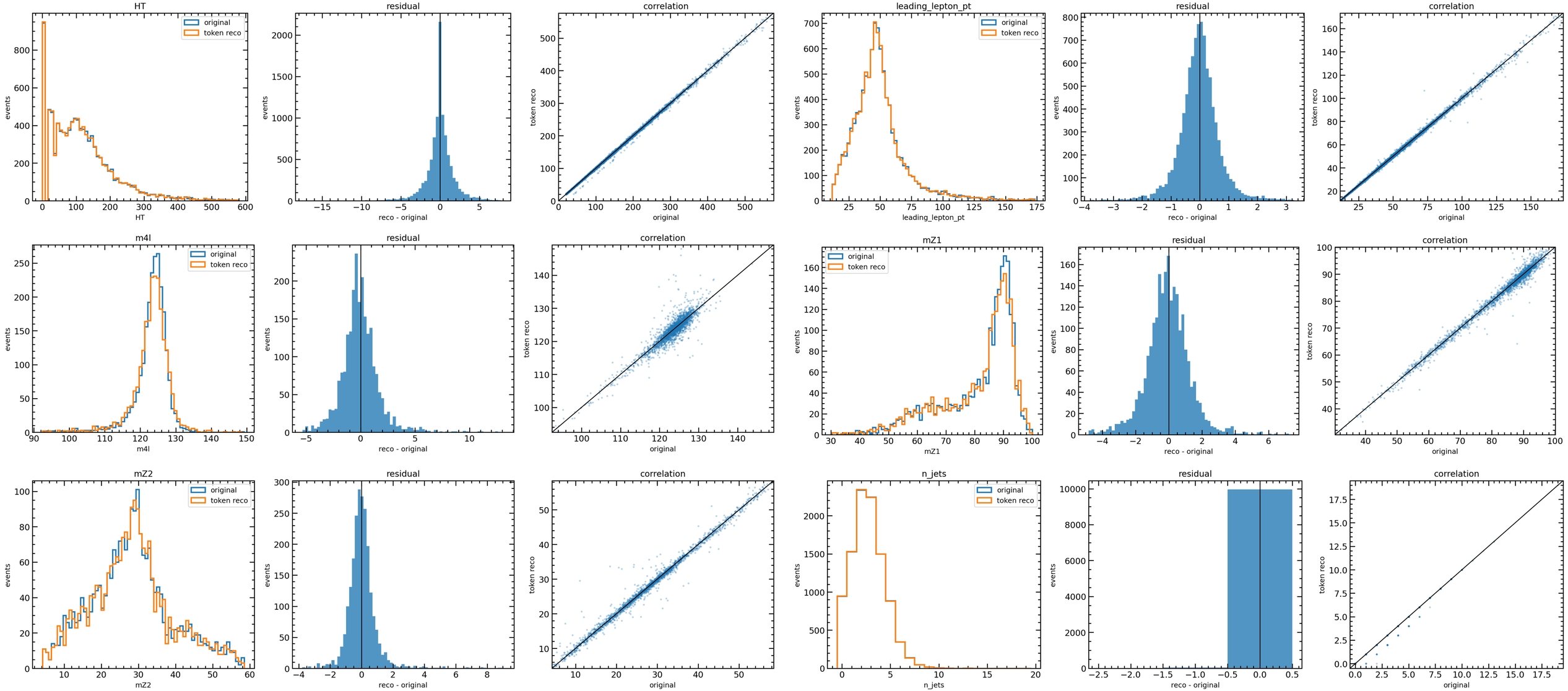








High-level variables after creating event-level tokeniser



Hierarchical Feature-Group Tokenization

Multiple sub-tokens per object, aggregated by inner transformer

Flat (current)

Each object → 1 position in the sequence.
All features are projected through a single linear layer.
The transformer must figure out which features are kinematics, which are b-tagging, etc.

Jet → Linear([pT, η, ϕ, mass, ..., GN2_pu]) → hidden dim

Hierarchical (feature groups)

Each object → 2-3 sub-tokens by semantic group.
Inner transformer aggregates sub-tokens → 1 vector.
Outer event transformer reads aggregated objects.

Jet → [kinematics: pT,η,ϕ,m] + [substructure: QG]
+ [b-tagging: DL1d, GN2]
→ inner attention → 1 jet vector → outer transformer

Feature groups per object:

Electron: [kinematics: pT,η,φ] + [ID: LHLoose/Med/Tight] Muon: [kinematics: pT,η,φ] + [ID: quality]
Tau: [kinematics: pT,η,φ] + [ID: nTracks, RNN scores] Photon: [kinematics: pT,η,φ] + [ID: isLoose/Med/Tight]
Jet: [kinematics: pT,η,φ,m] + [substructure: QG vars] + [b-tagging: DL1d+GN2] Track: [kin] + [impact: d0,z0] + [quality: χ^2,nDoF]

Brookhaven National Laboratory