Treasure event level-tokenisation

Merve Nazlim Agaras

1.7.26

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

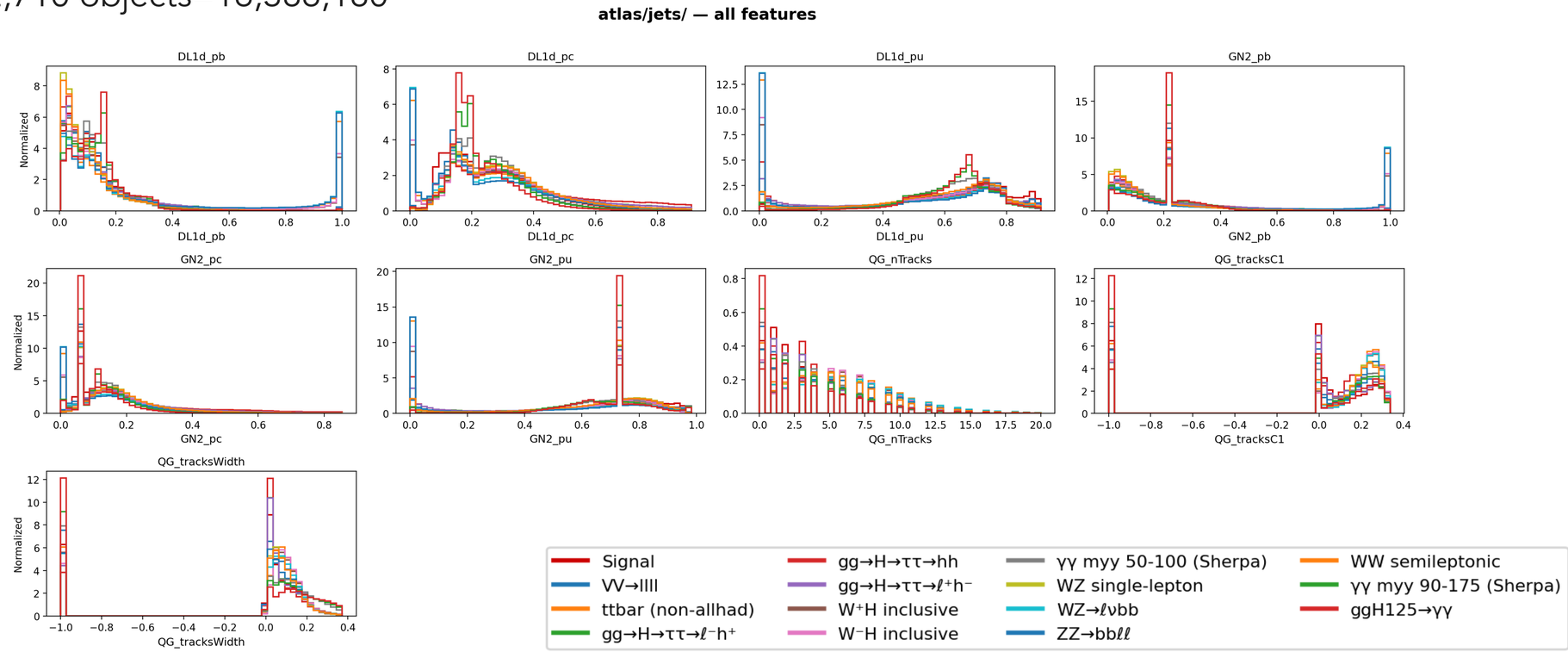
- common/muons/pt
- common/muons/eta
- common/muons/phi
- common/muons/charge
- atlas/muons/ptvarcone30
- atlas/muons/topoetcone20

- 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

- 27,892,710 total events
  - 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



Previous results

- Initial scans used raw object features.
- Increasing capacity improved some hard variables, especially jet tagger scores.
- But we saw two issues:
  - q0/codebook collapse, especially first quantizer using very few codes.
  - Some variables had poor reconstruction or smoothed discrete structure.
- Standardisation Fix:
  - Apply preprocessing before tokenizer training.
  - For skewed positive variables, use log transform first.
  - Then apply StandardScaler to all features.
  - Example for jets:
    - log-standard: pt, mass, n_trk, QG_nTracks
    - standard-only: eta, phi, taggers, QG shape variables.
- This improved reconstruction stability and codebook behavior.

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.

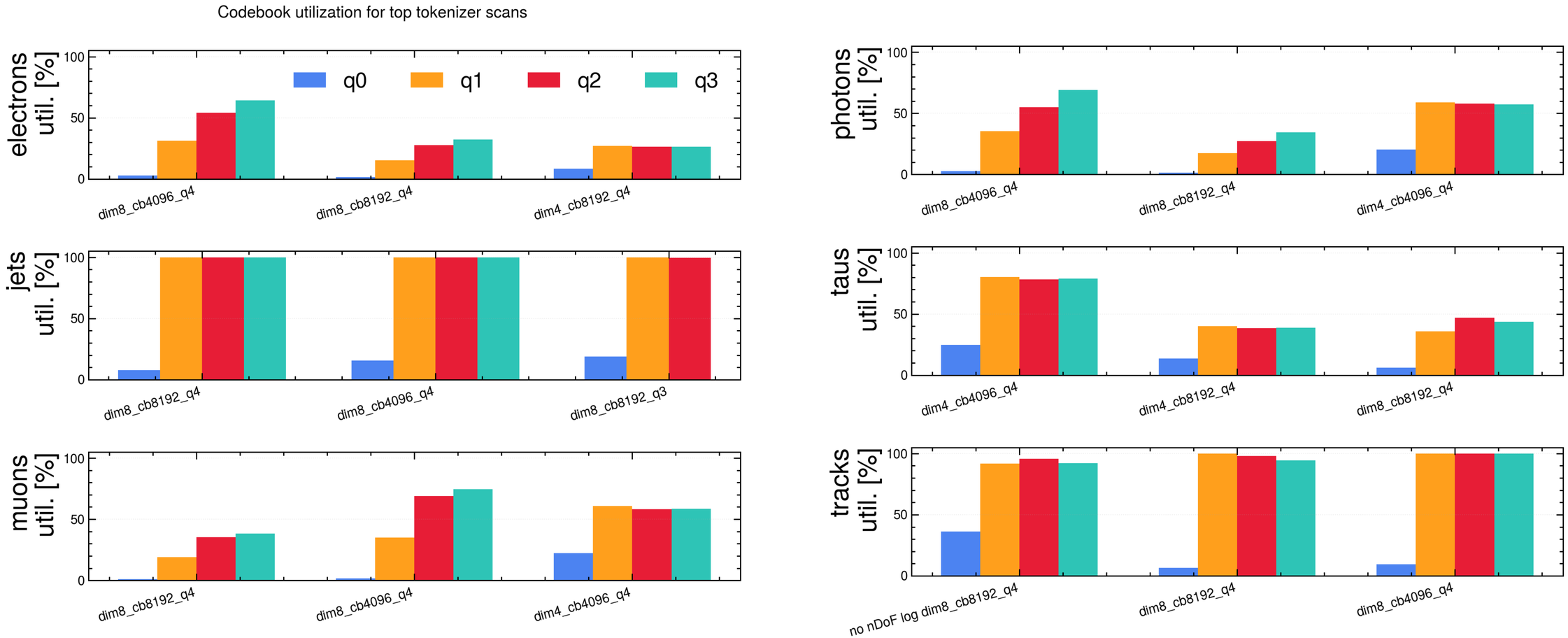
object	current preferred setup
jets	dim8_cb4096_q4
electrons	dim8_cb4096_q4
muons	dim8_cb4096_q4
photons	dim8_cb4096_q4
taus	dim8_cb4096_q4 or dim8_cb8192_q4
tracks	dim8_cb8192_q4, no log on nDoF

Feature-Level Findings

* Jets: tagger variables and n_trk need enough capacity; dim8 works better than dim4.
* Electrons/muons/photons: isolation/cone variables are the hard ones.
* Taus: pt dominates; RNN scores are okay with dim8.
* Tracks: nDoF should not be log-transformed; standard-only works better.

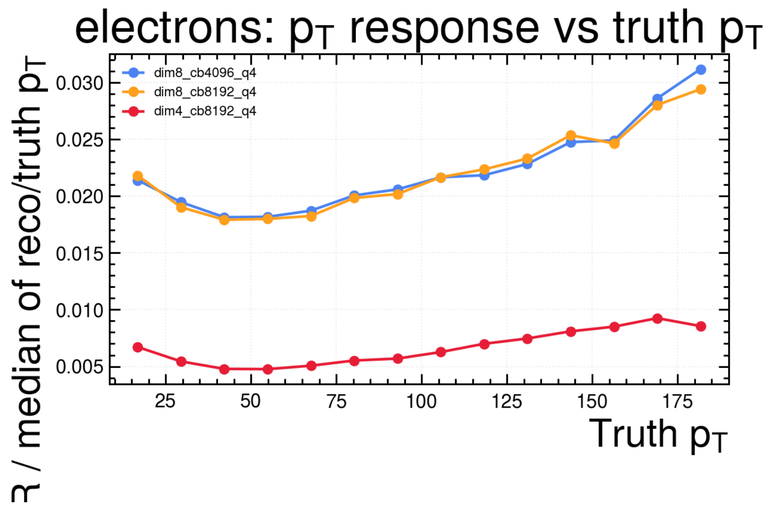
Summary Plots

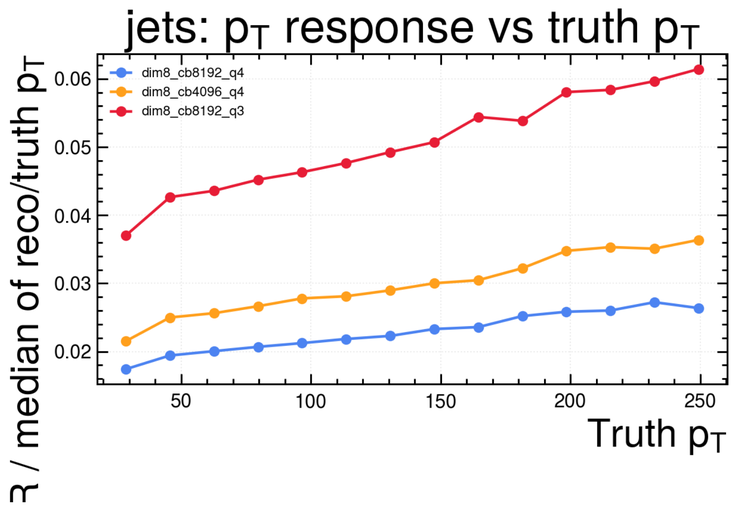
* Codebook utilisation per object/top scans.

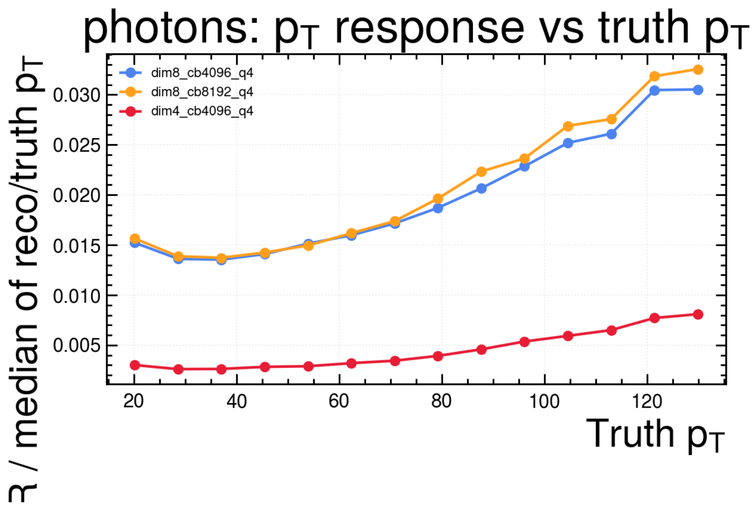


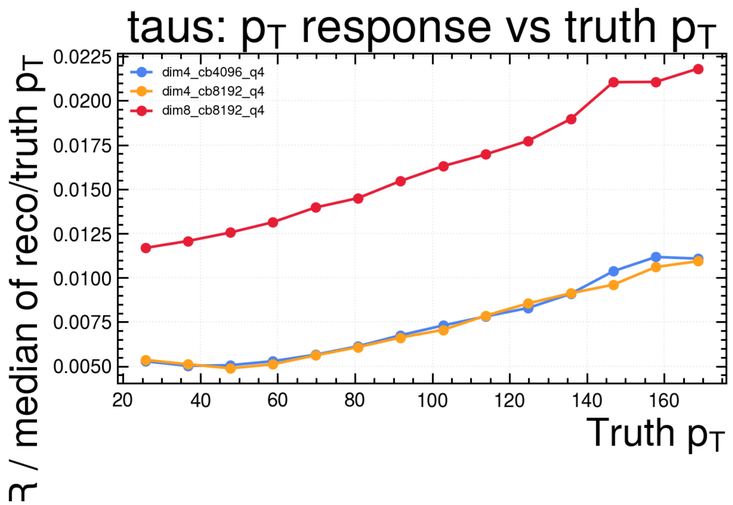
Summary Plots

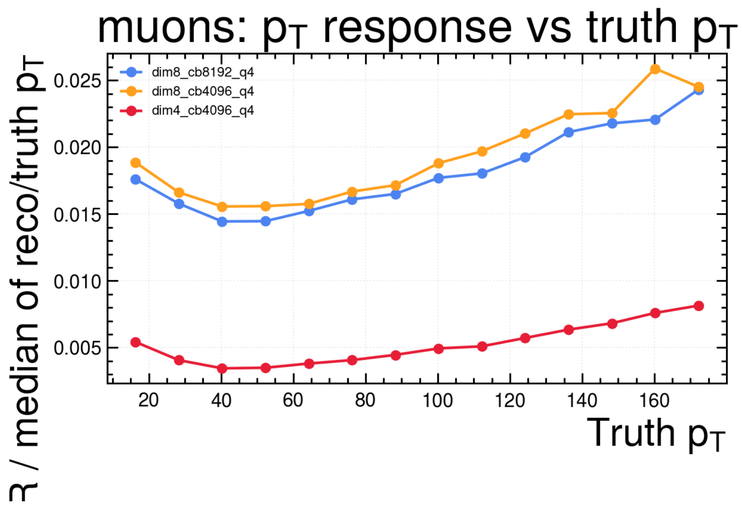
* binned pT resolution vs truth pT

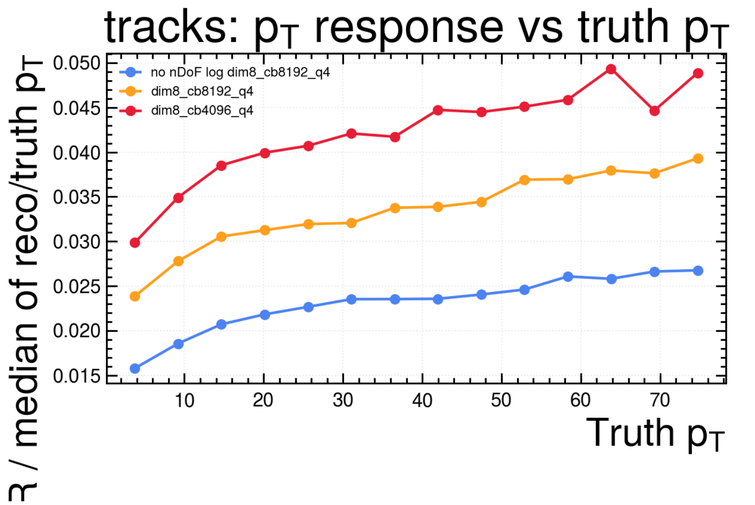






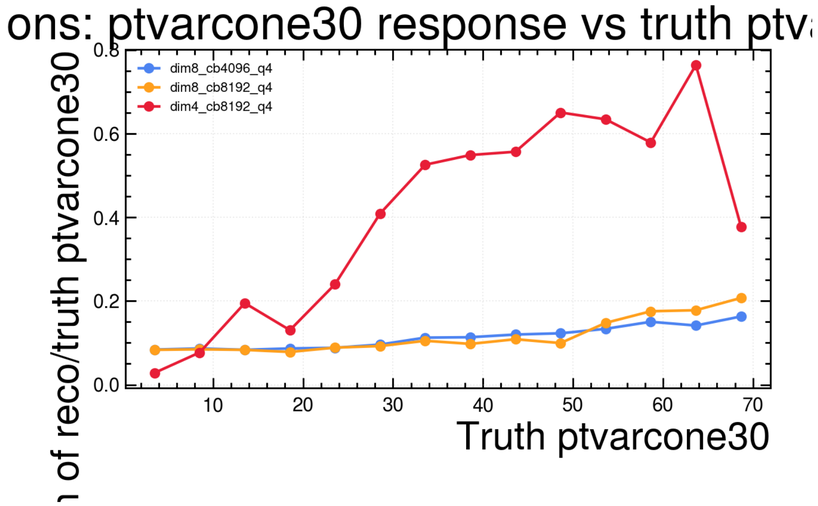


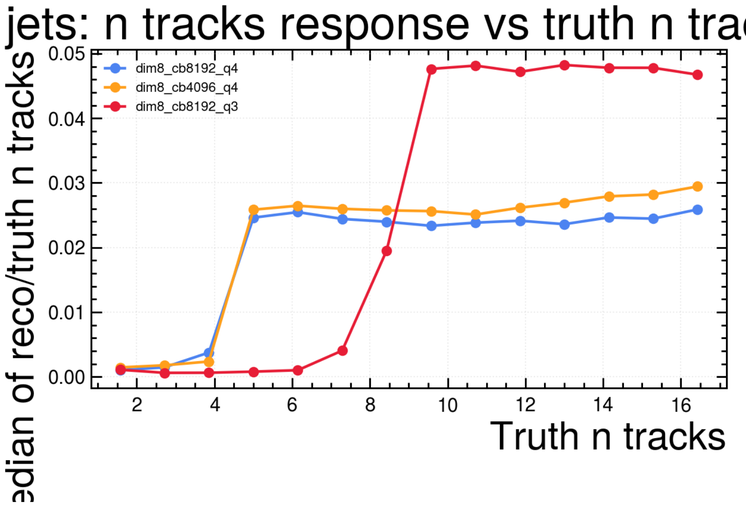


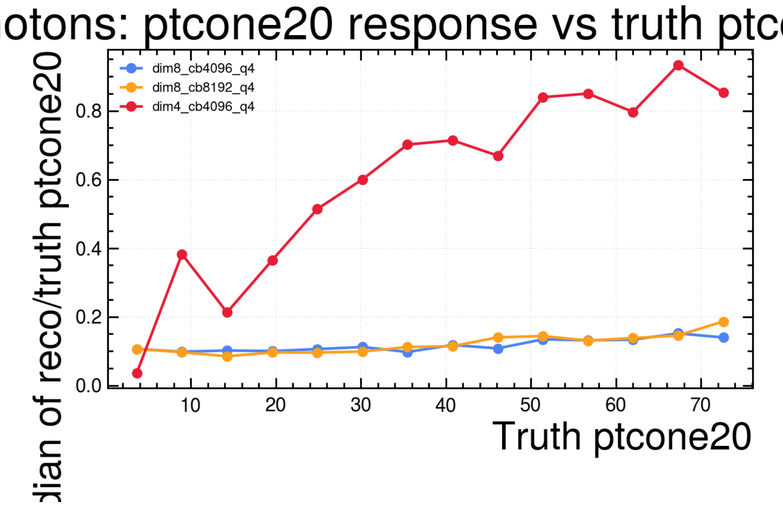


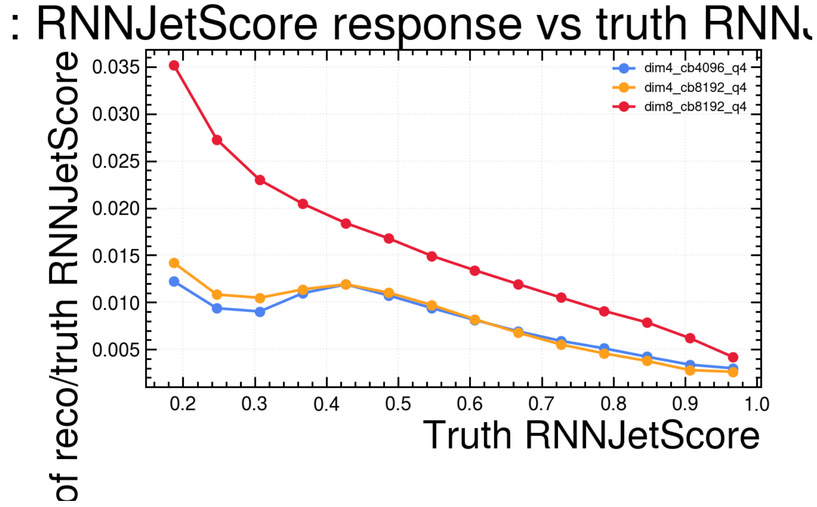
Summary Plots

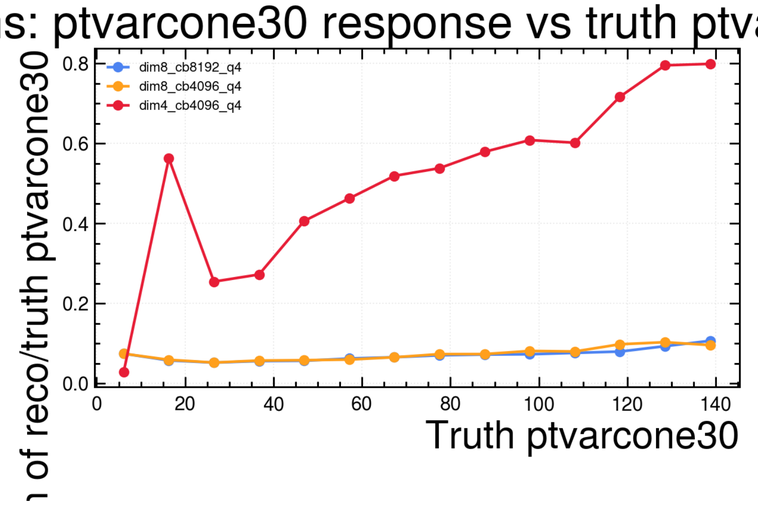
* Binned features resolution vs truth

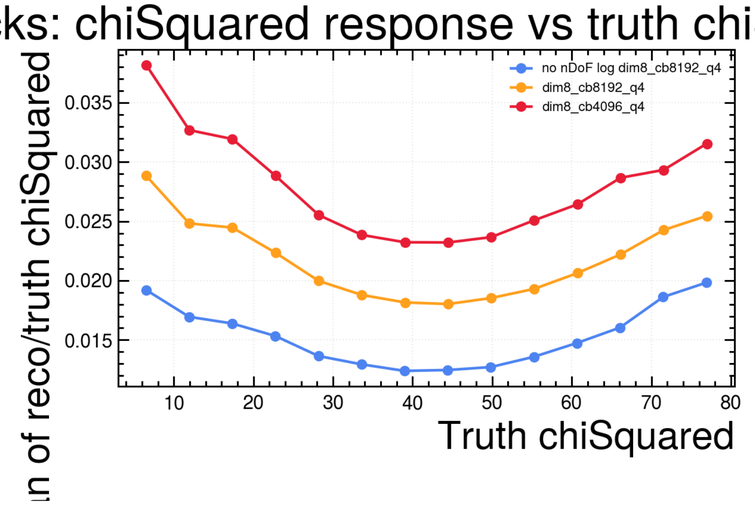




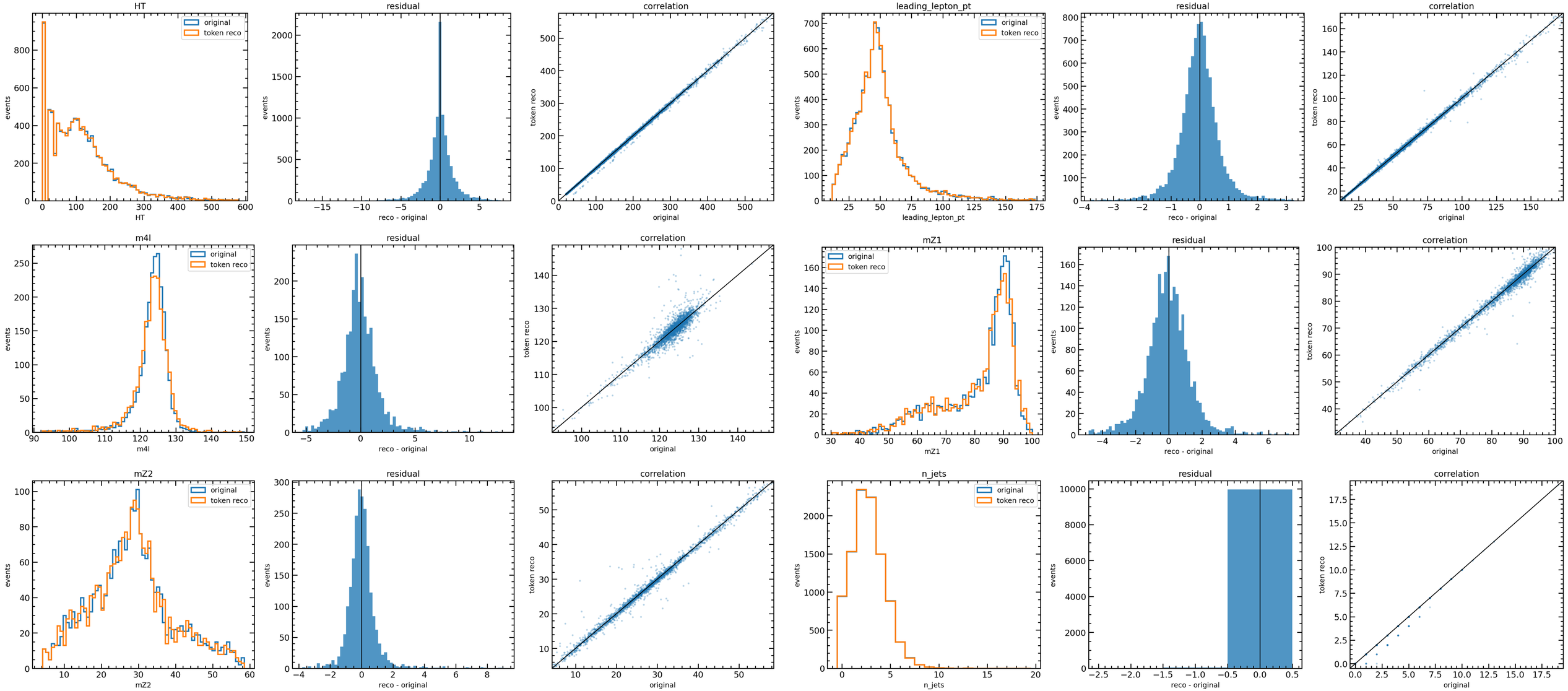








High-level variables after creating event-level tokeniser



Summary/next steps

Summary

* Built per-object ATLAS tokenizers for jets, electrons, muons, photons, taus, and tracks.
* Standardisation/log-standard preprocessing was the main improvement over the raw-input scans.
* The best current compression point is around $_{\text{dim8}}$, $_{\text{q4}}$, and moderate codebook sizes.
* Remaining reconstruction issues are feature-specific rather than global tokenizer failures.

Next steps

* Finalise one tokenizer setup per object and freeze the preprocessing choices.
* Re-train the tokenizer with data+MC (MC only so far)
* Tokenize the full dataset into event-level parquet files.
* Start downstream tests with the event-level foundation model.
* Use downstream performance to decide whether a context-aware tokenizer is worth pursuing.

Backup