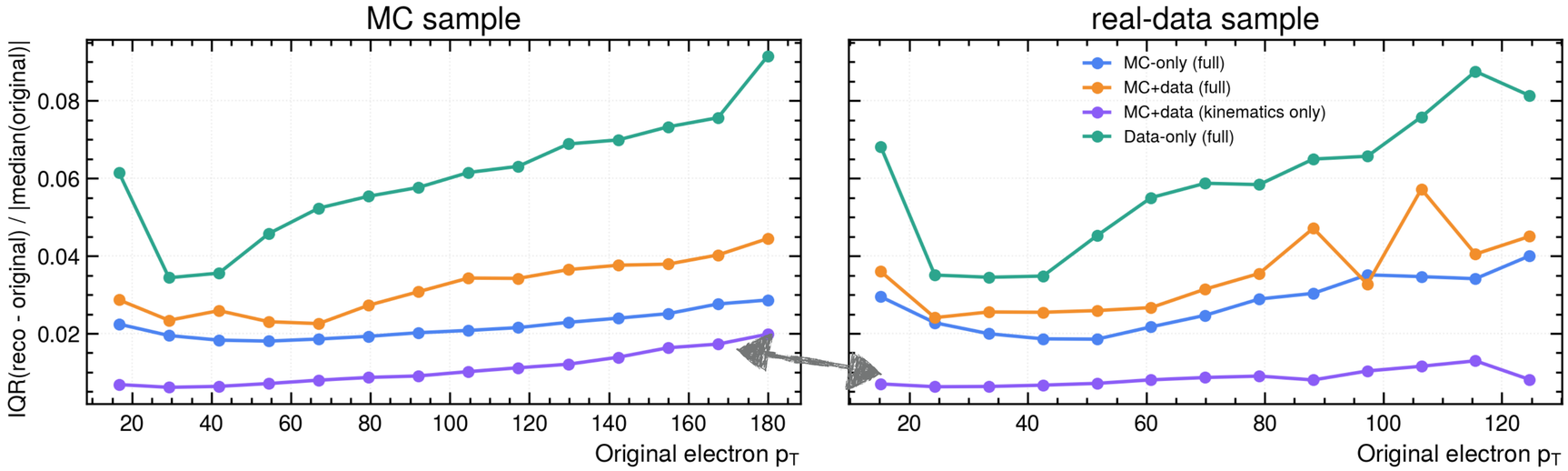
Brookhaven National Laboratory

Treasure event level-tokenisation

Merve Nazlim Agaras

29.7.26

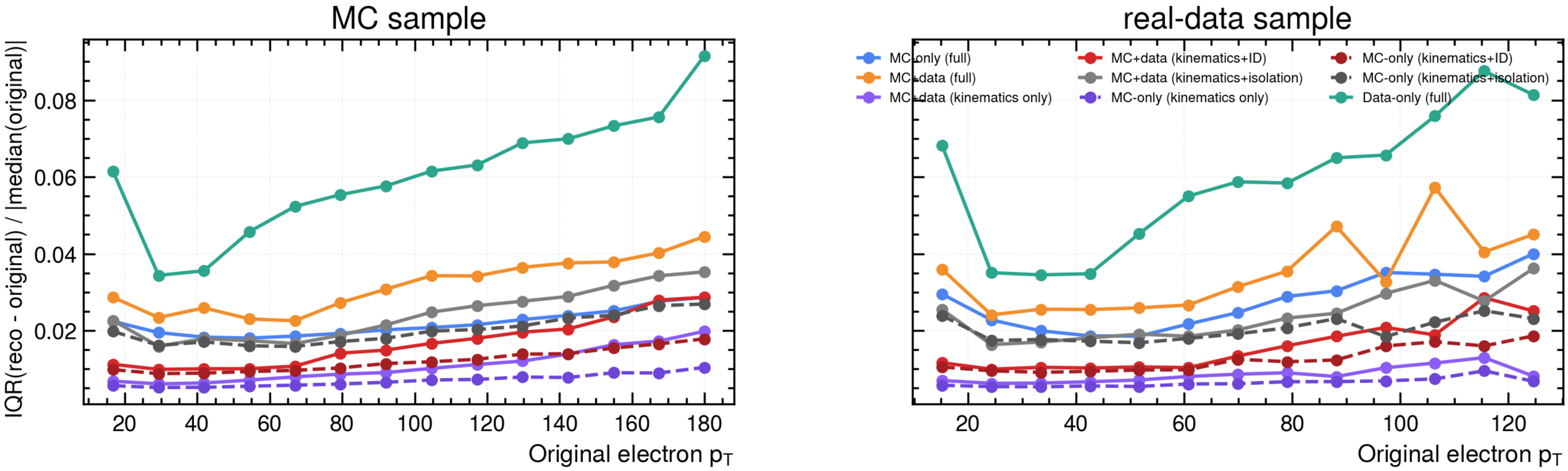
Feature controls



LAST WEEKs

* 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

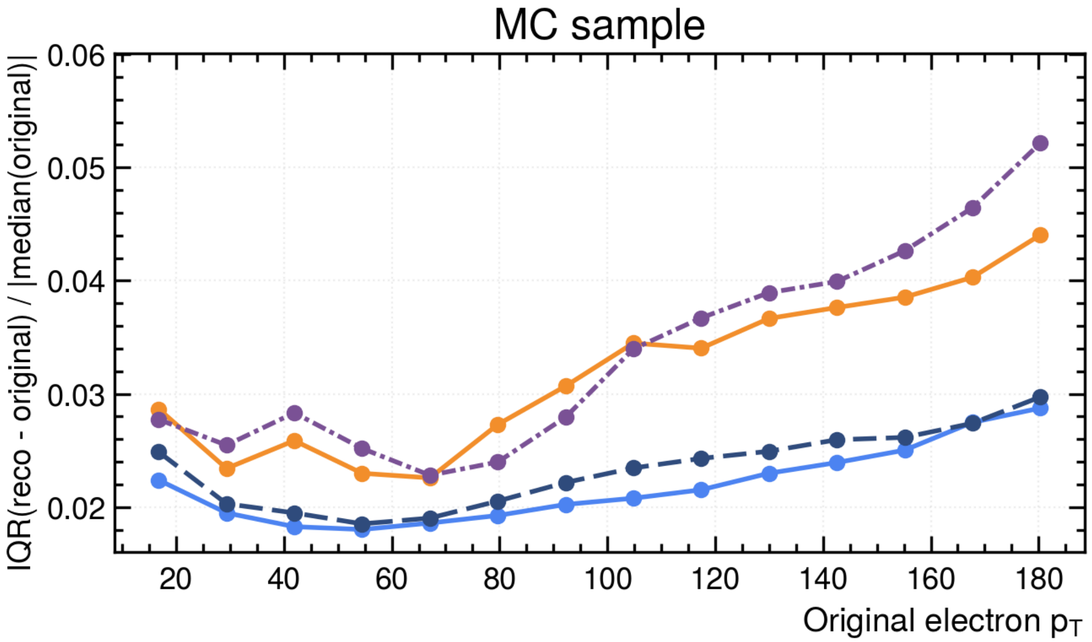
Electron tokenizers w/ different set of variables

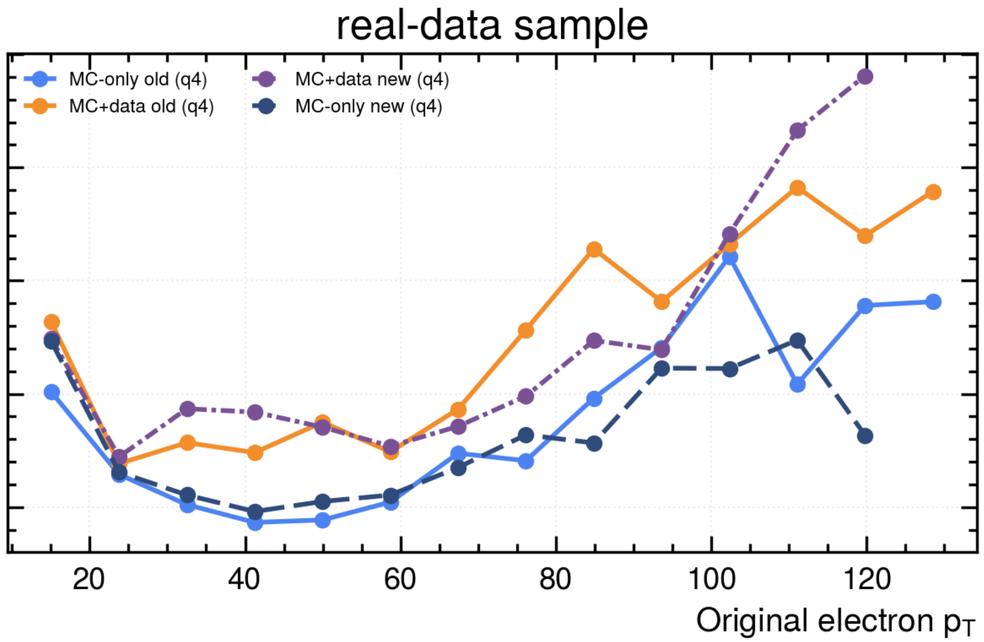


* Focused electron tokenizers beat full-feature tokenizers. The best reconstruction is not obtained by giving the VQ-VAE every electron variable at once; it improves when the feature set is split by purpose. —> **Not enough capacity to reconstruct different kind of features??**
* MC-only and MC+Data kinematics only are quite close ~X%
* Also figured out that data does not use GRL —> **re-ran the samples with the correct selections (next slide)**

Old vs new data sample tests

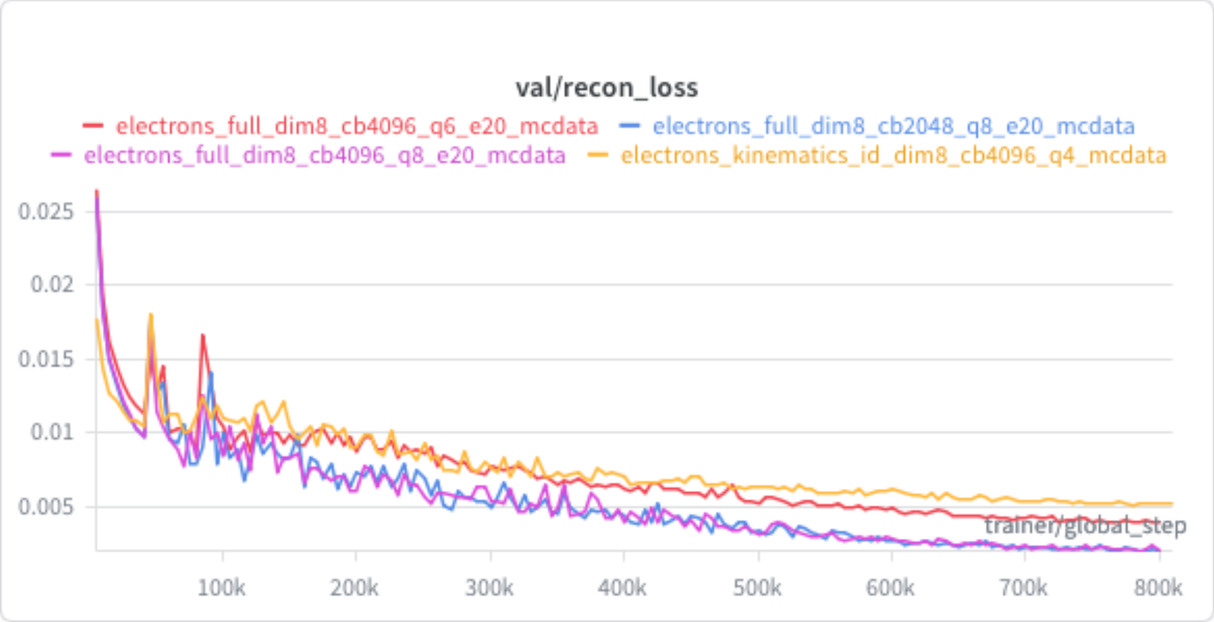
- Re-ran the data samples with GRL and similar selections as given in the link below (framework to create ntuples for opendata)
  - https://gitlab.cern.ch/atlas-outreach-data-tools/physlitetoopendata/
- Test whether the degradation is coming from the selection bias
- The similar degradation level is remained
- The final possible explanation of the degradation
  - The model is not enough to reconstruct all variables and data+MC together

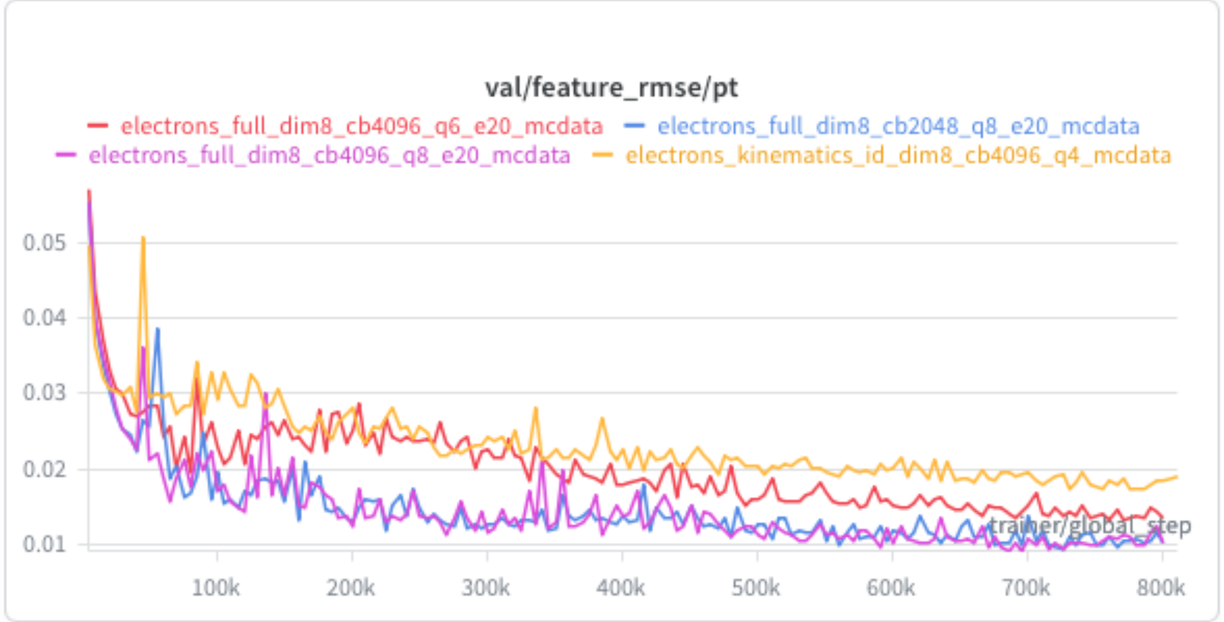




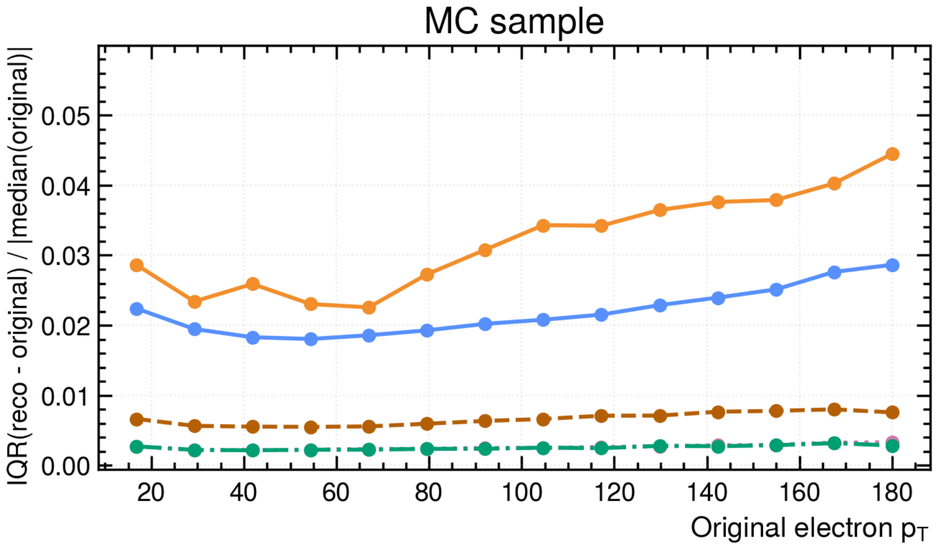
Increasing the capacity

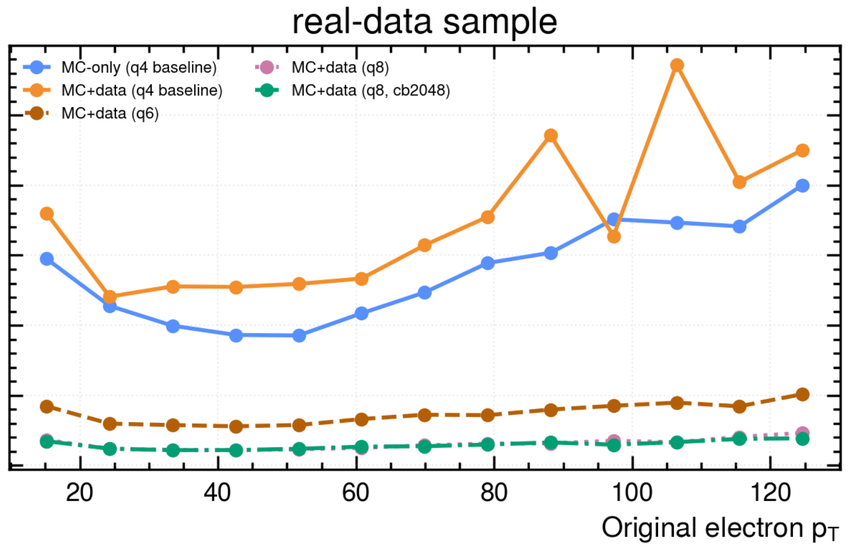
* More number of quantizers can describe the object in more refinement steps than q4
* Lower loss and RMSE for more quantisers
* Cons: More tokens per object -> higher downstream cost, less compression
  - q4 = more compact, cheaper, worse reconstruction
  - q8 = better reconstruction, more expensive, less compressed

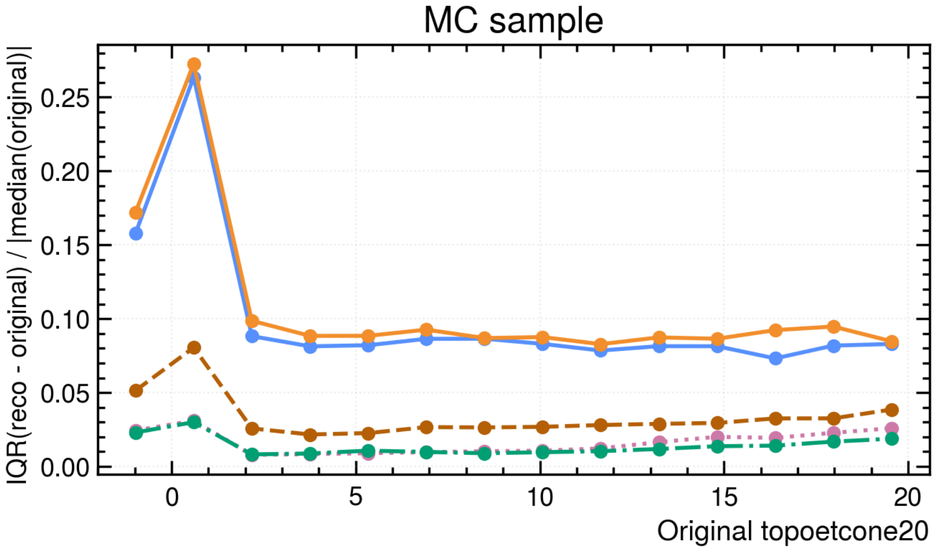


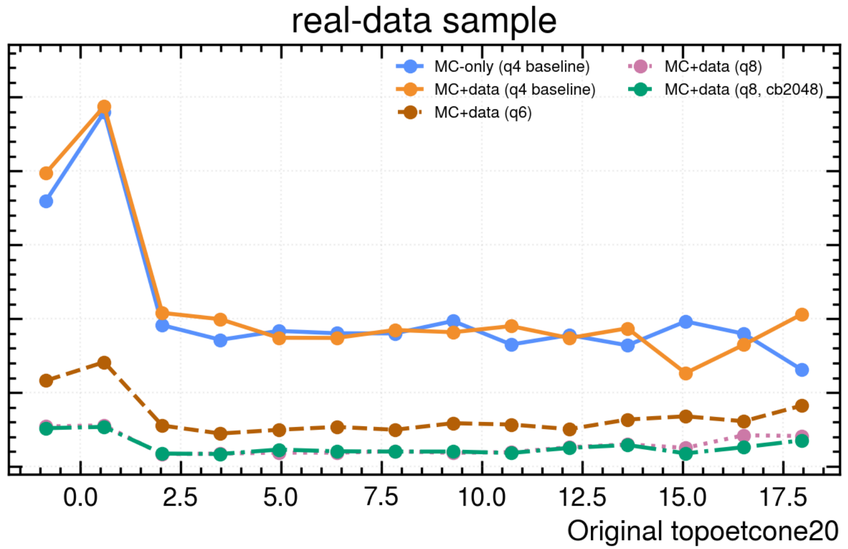


Increasing the capacity



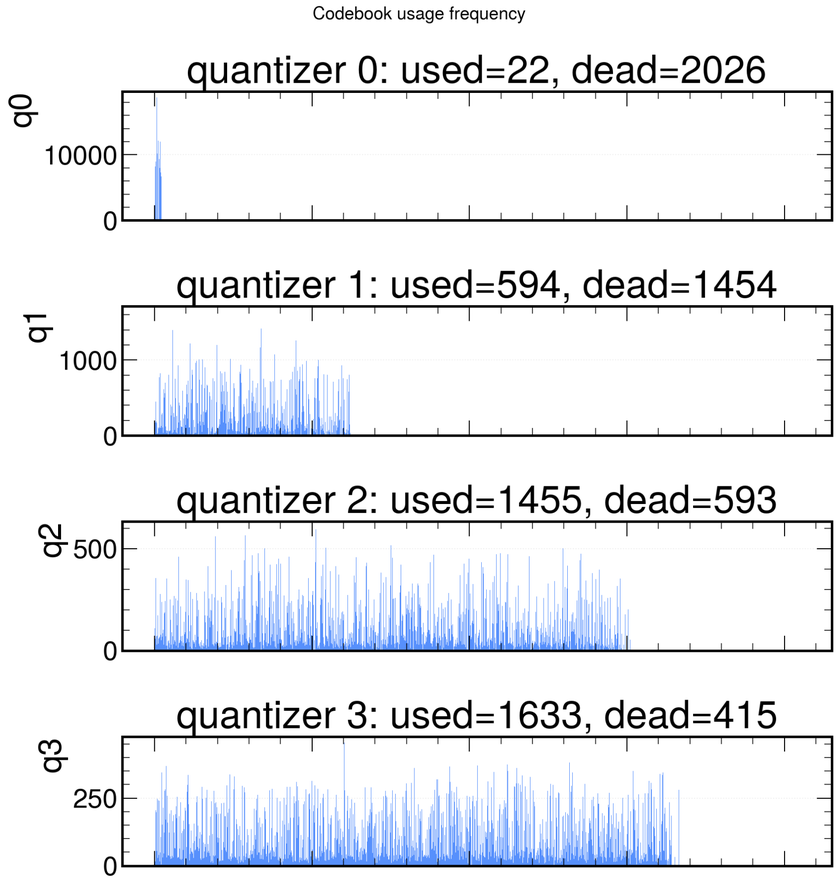


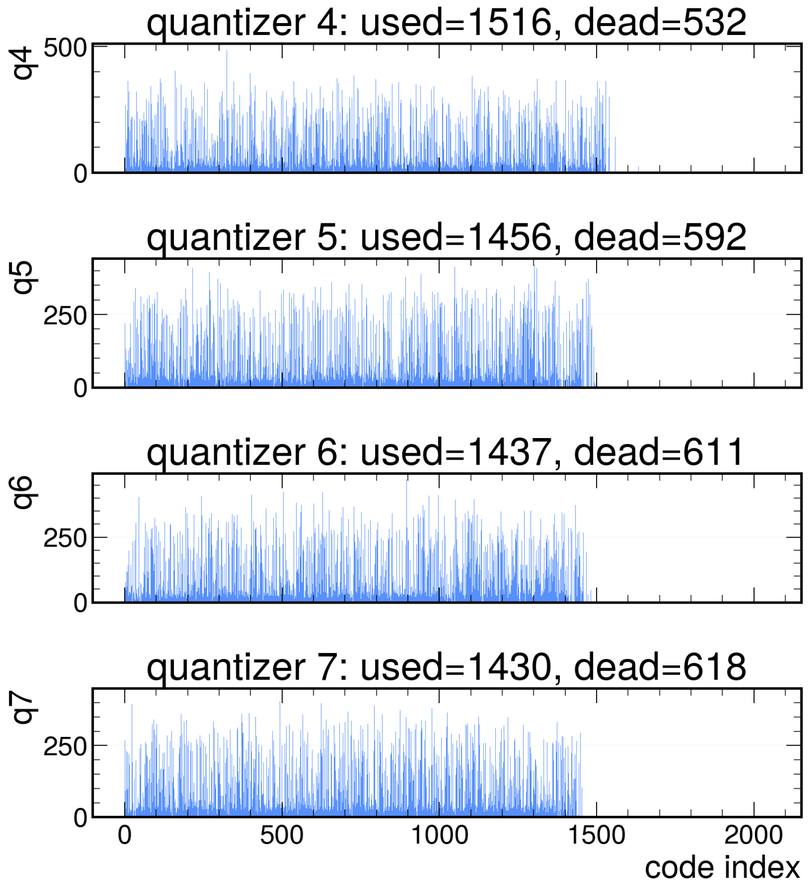


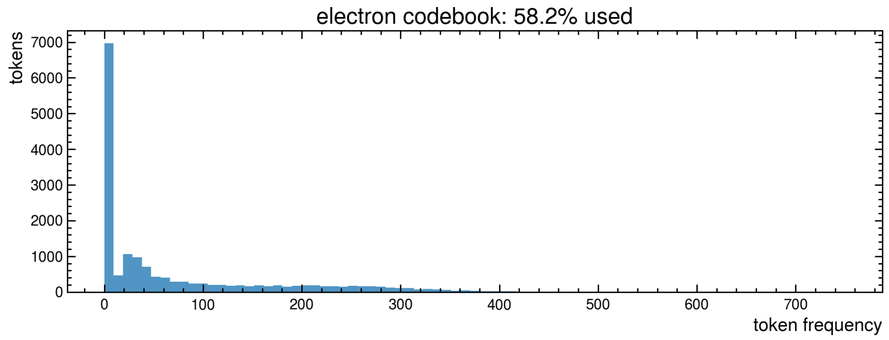


* Clearly increasing capacity fixes the problem
* Is this something we want to consider?

Usage

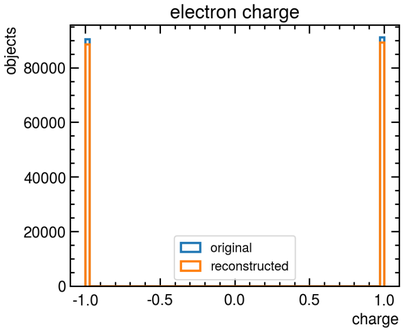


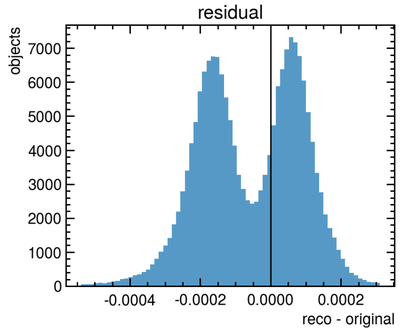


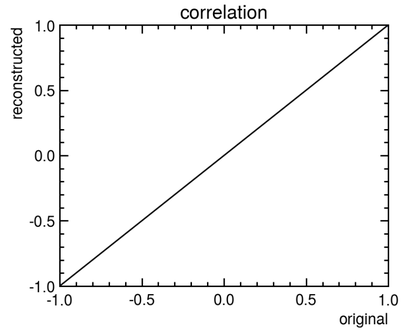


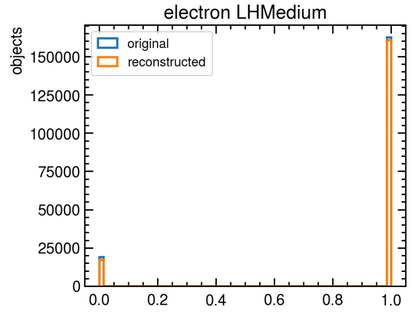
- MC+data q8 2k codebook size
- Will try new q6 2k codebook size
- The added quantizers are not empty: q4-q7 each use ~1400-1500 of 2048 codes, so the extra quantization stages are actively carrying information rather than just sitting unused.

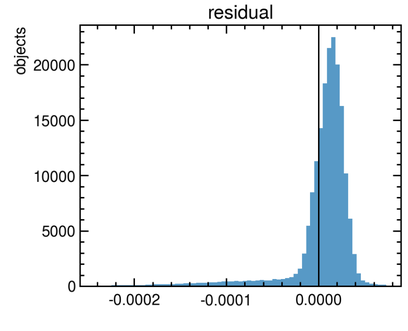
Reconstruction

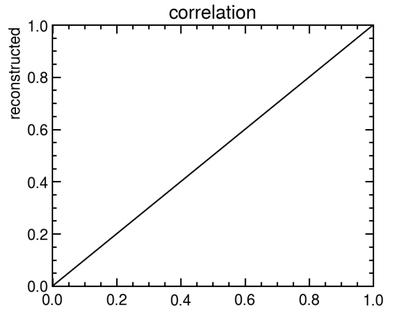


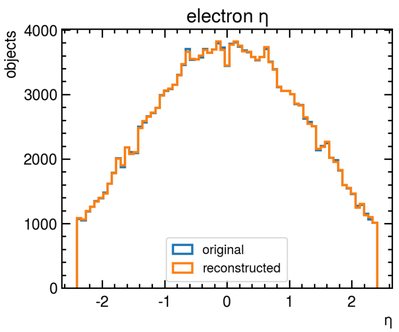


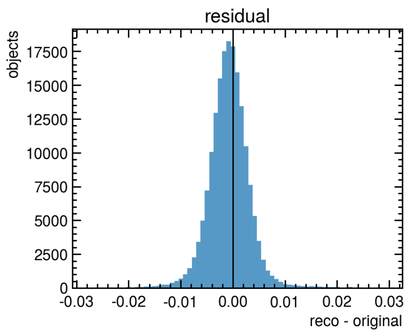


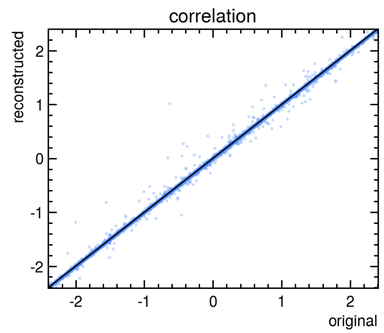


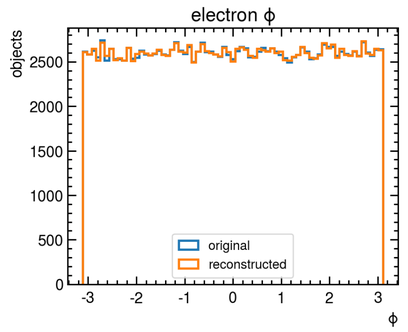


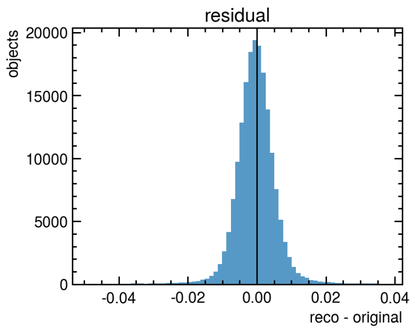


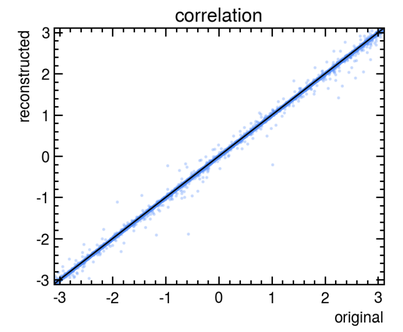


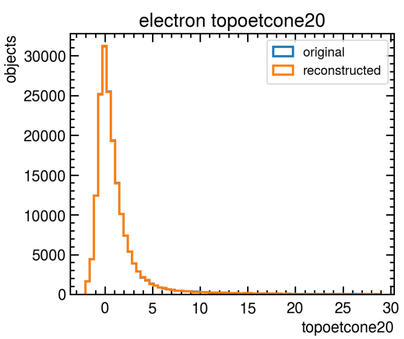


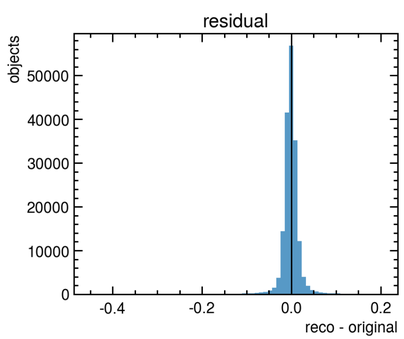


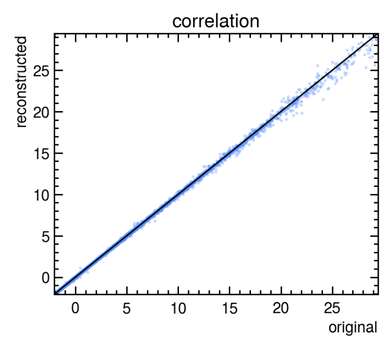


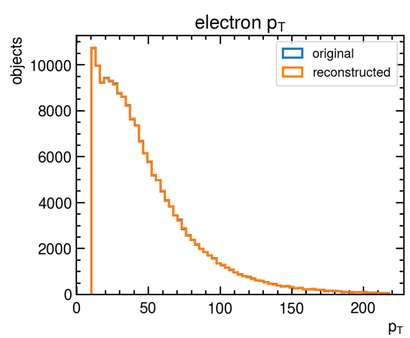


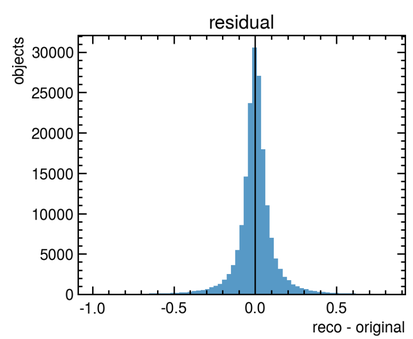


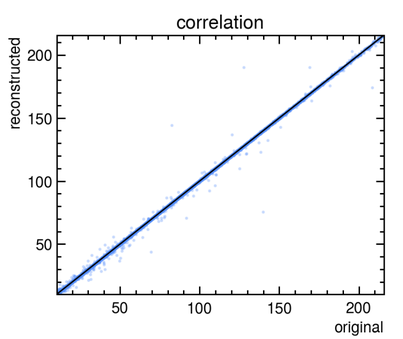






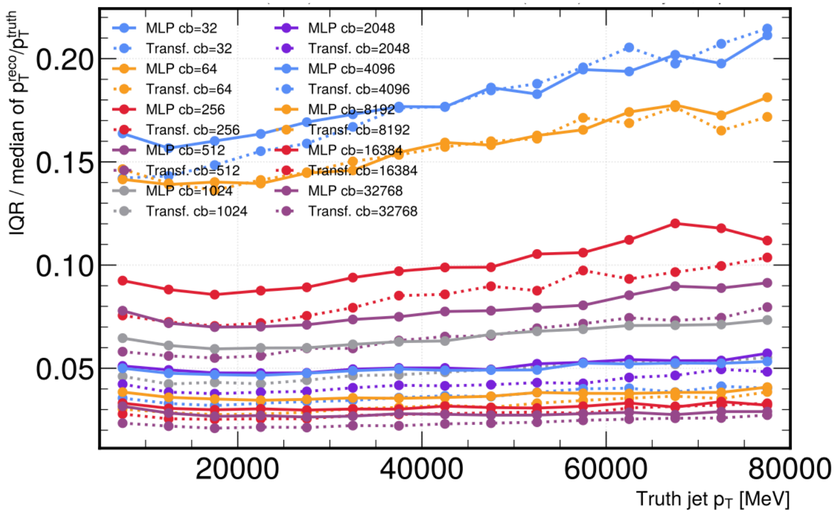






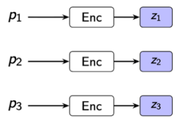
Event context aware VQ-VAE

In-context Encoding Improves Reconstruction



- Transformer outperforms MLP
- Both benefit from larger codebook size → push to 8192, 32768 codebook sizes

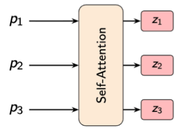
No-context (MLP)



$z_i = f_\theta(p_i)$

Each token is independent

In-context (transformer)



$z_i = f_\theta(p_1, \ldots, p_N)_i$

Tokens carry context of entire jet

* Jet studies showed improvement when using tranformer instead of MLP

1. Easy first implementation
   - **Per-object in-context tokenizers**: For per VQ-VAE use transformer instead of MLP
2. But the application of this in event-level is not simple:
   - A. if we use only kinematic variables (pT, eta,phi), it should be easy, but we should add **type embeddings**; otherwise an electron and a jet with similar kinematics could look identical to the tokenizer and different preprocessing per object.
   - B. If we use different feature sets per object, we need **type-specific input projections** and r**econstruction heads**.
   - C. We also need to think about loss balancing, since object multiplicities are very different, e.g. many tracks/jets but fewer electrons or muons.

Option 1 is duable in a short time, option 2 needs more time and iterations (long term)

Summary/next steps

Summary

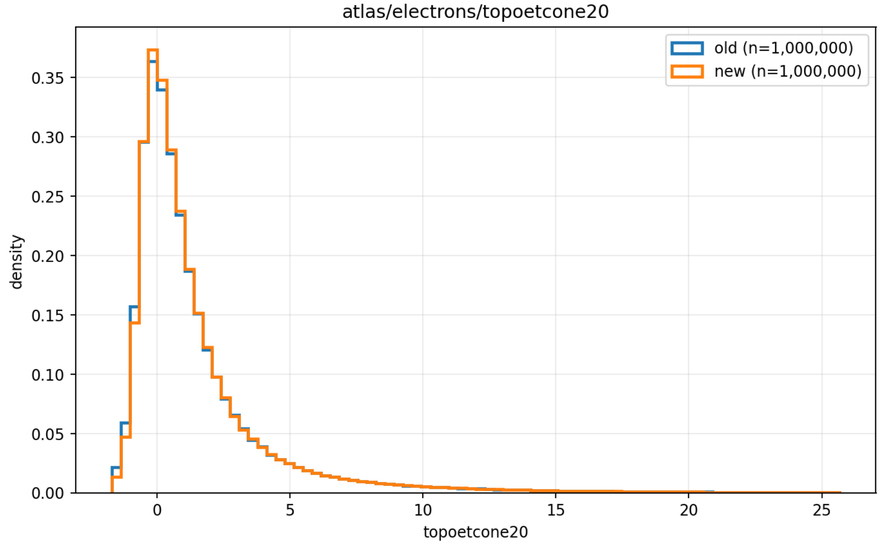
* Adding data selection (mainly GRL and some additional cuts) did not change the degradation on the reconstruction for electrons
* Studies showed splitting different groups of features increases affect largely the reconstruction efficiency
  - Leads to not enough capacity of reconstructing all variables + data/MC differences
* Increasing number of quantisers and reducing codebook sizes gave the optimal reconstruction and codebook usage statistics

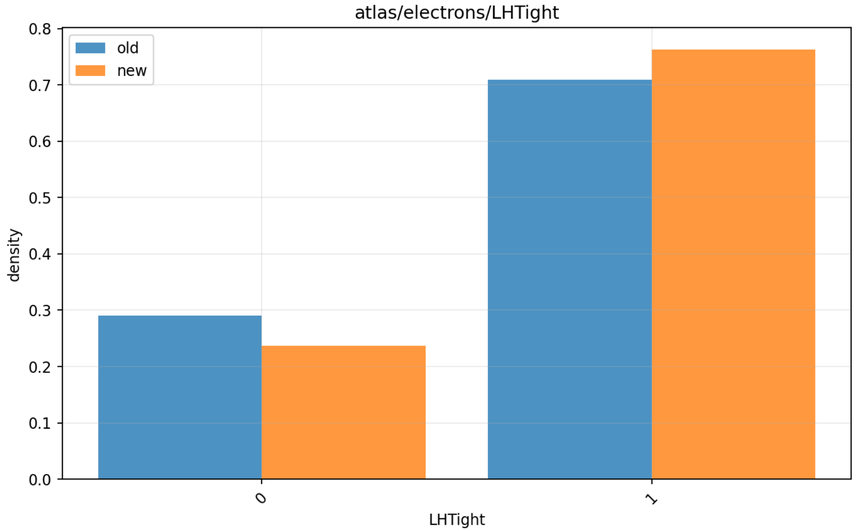
Next steps

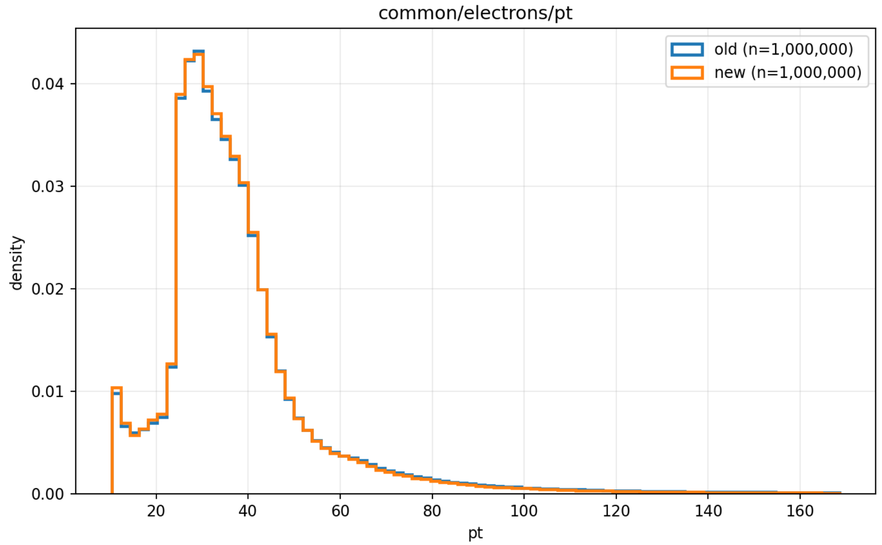
* Freeze the individual VQ-VAE trainings per object for two setups
  - q4 and q8
* Start training the downstream task for these two setups
* Downstream task will be H->ZZ->4l vs ZZ* vs ttbar+Zjets multi-class classfication
* If time allows, adding production modes to transformer heads
* In parallel, try per-object in-context tokenizers and/or hierarchical feature group tokenizers

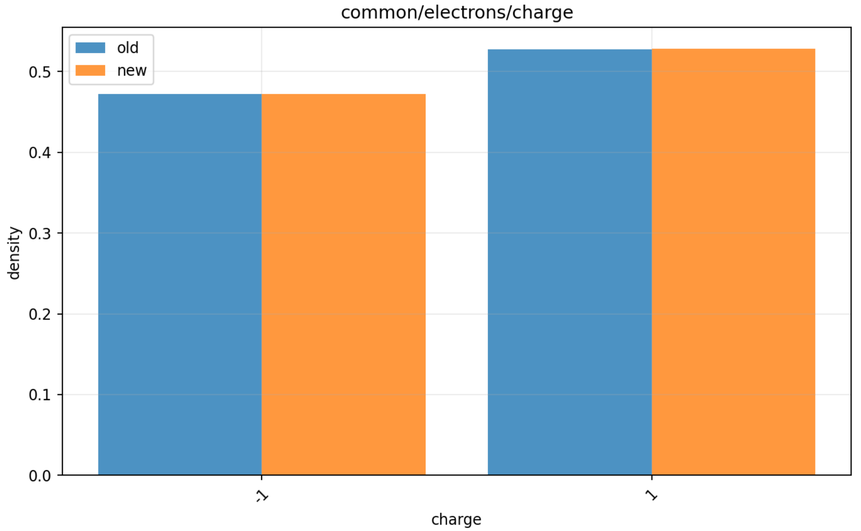
Backup

Data old vs new



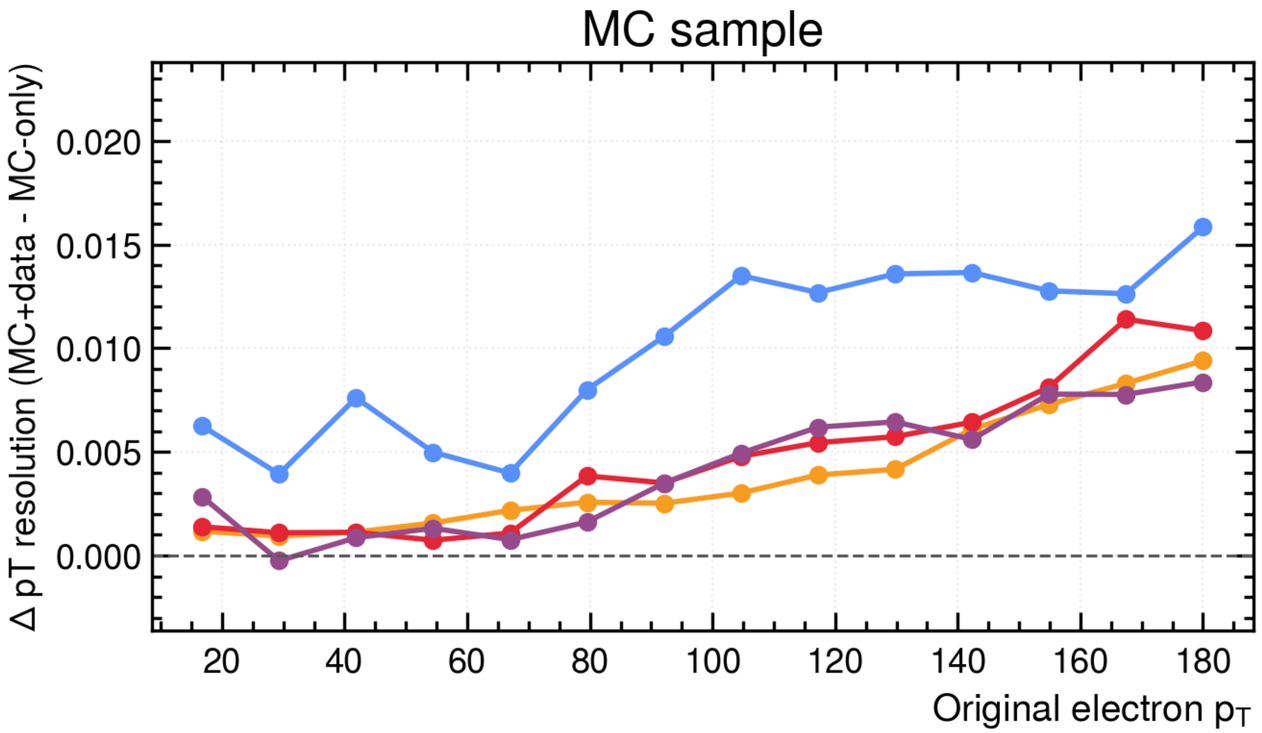


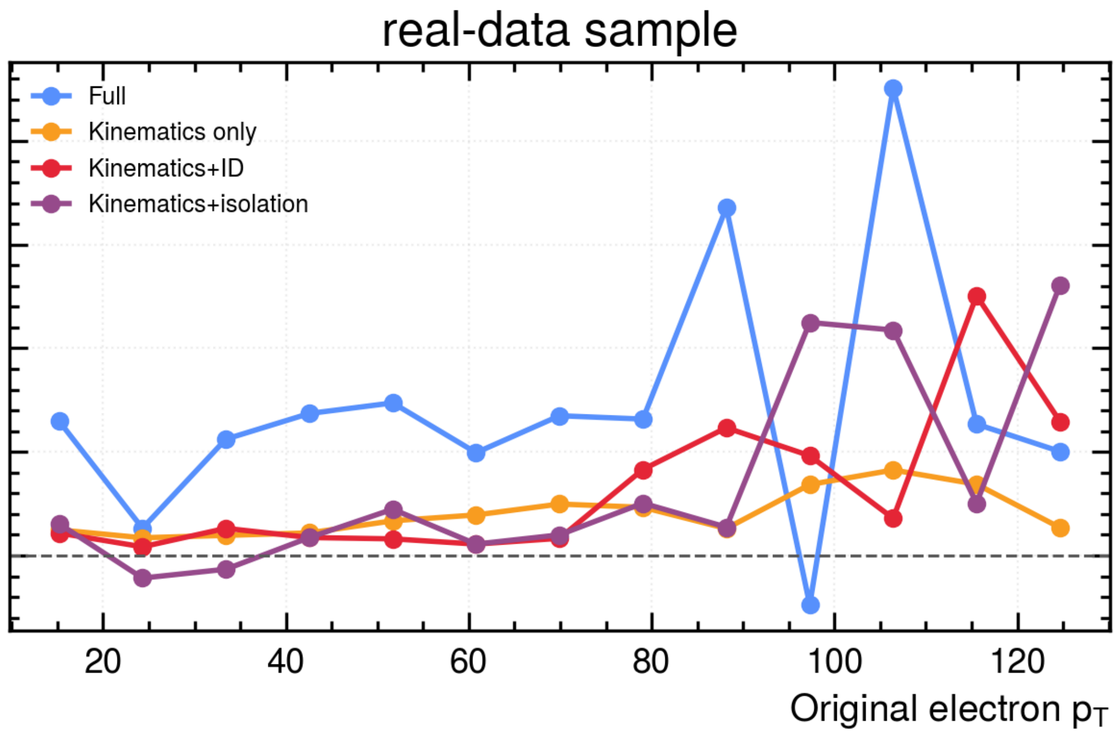




Delta resolution

MC+data minus MC-only pT binned 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/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

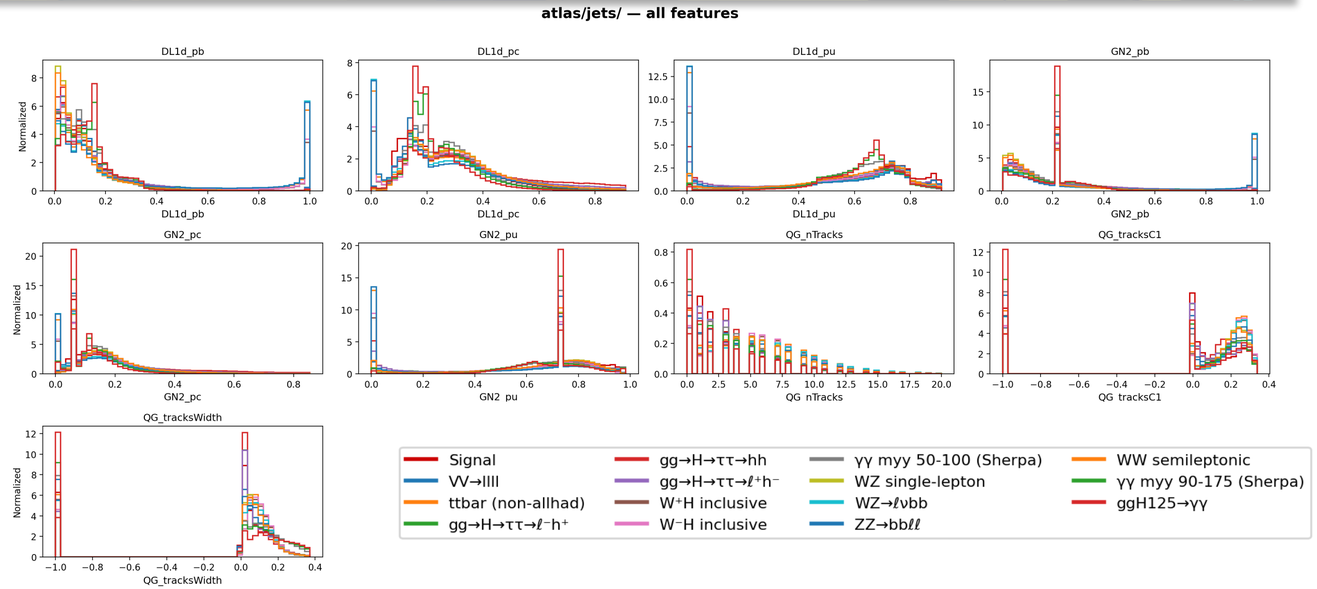
- common/tracks/pt
- common/tracks/eta
- common/tracks/phi
- common/tracks/d0
- common/tracks/z0
- atlas/tracks/qOverP
- atlas/tracks/chiSquared
- atlas/tracks/nDoF

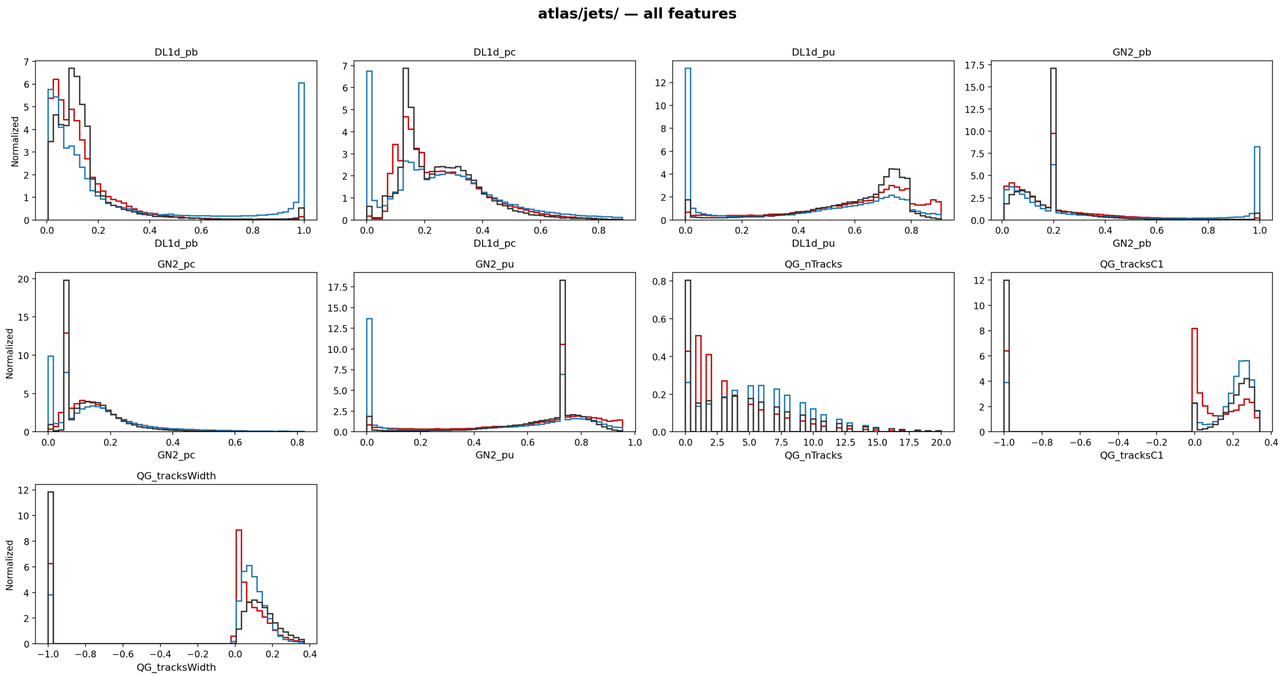
- common/photons/pt
- common/photons/eta
- common/photons/phi
- atlas/photons/isTight
- atlas/photons/ptcone20
- atlas/photons/topoetcone20
- atlas/photons/topoetcone40

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





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

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