Treasure event level-tokenisation

Merve Nazlim Agaras

3.6.26

Stats of the events

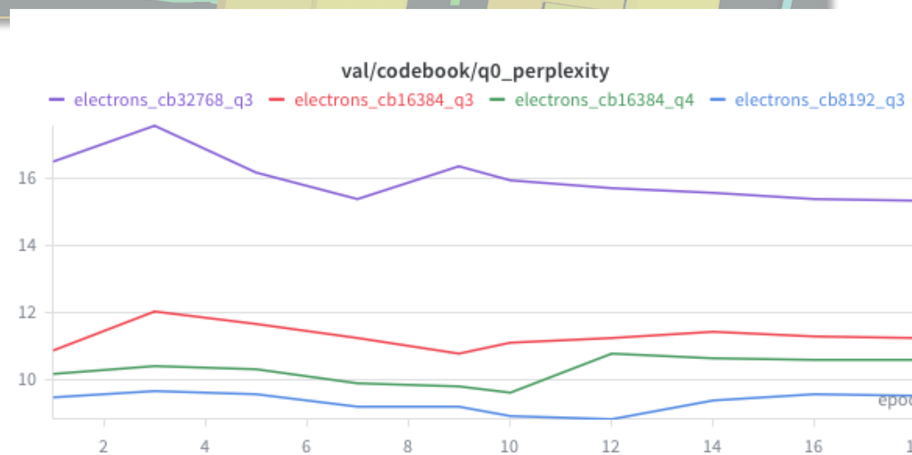
* HH->4l, ttbar and DY background
* Data15
* Event and object info

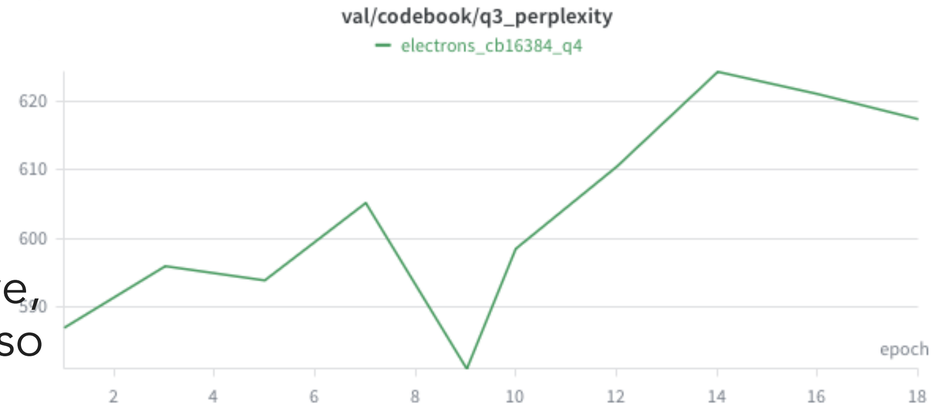
```
jets        events=8,778,484 objects=16,543,249
electrons   events=8,778,484 objects=66,818
muons       events=8,778,484 objects=124,045
photons     events=8,778,484 objects=1,016,327
taus        events=8,778,484 objects=1,678,534
```

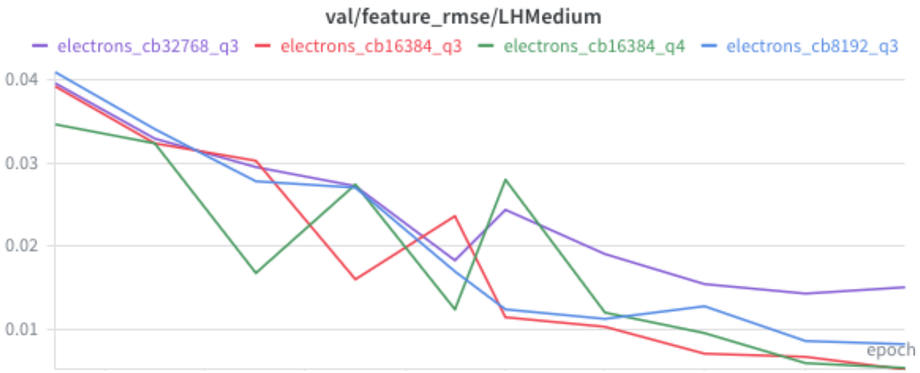
* Code: https://github.com/Treasure-AmSC/heptokens/tree/event-object-tokenization

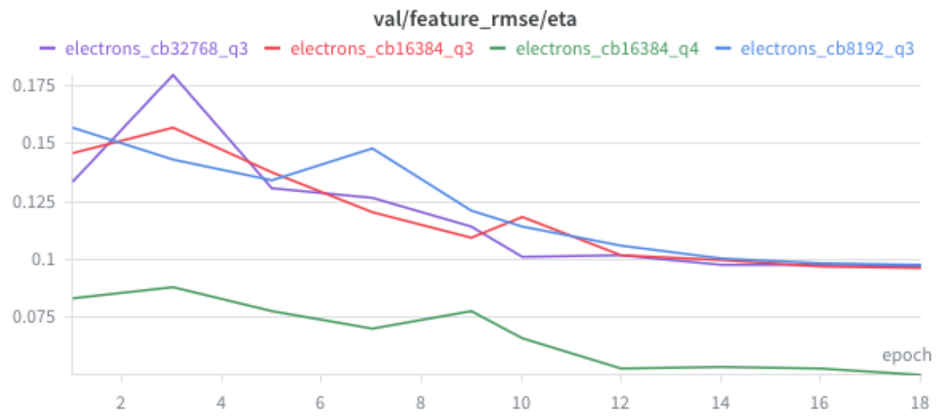
Electrons

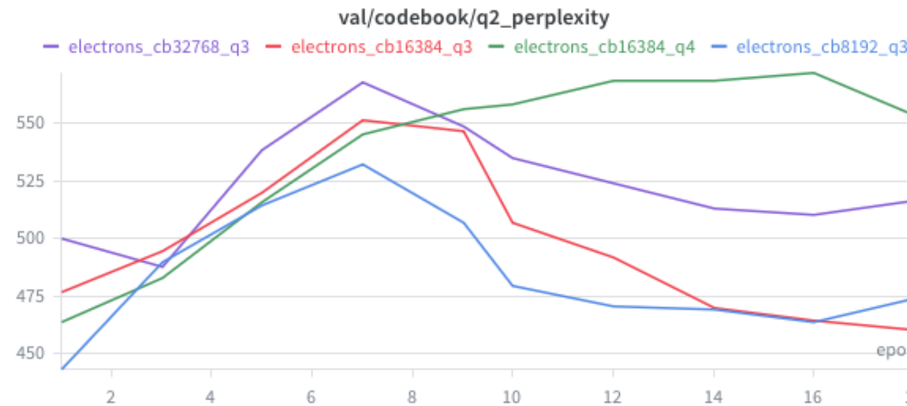
- For electrons, the result is pretty clear: **electrons_cb16384_q4** is best overall.
- It has the lowest or near-lowest val/recon_loss.
- It clearly improves pt, eta, phi, and charge.
- It improves the ID variables, especially LHMedium and LHTight.
- The 4th quantizer q3 is active, with ~650-680 used codes, so it is actually contributing.



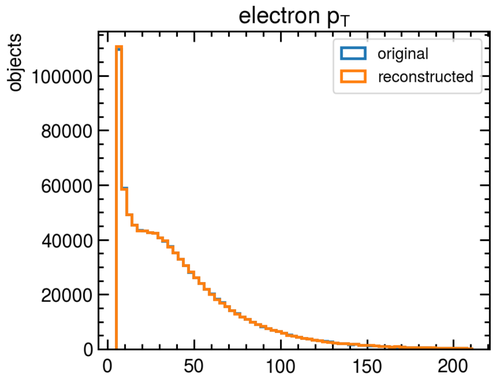


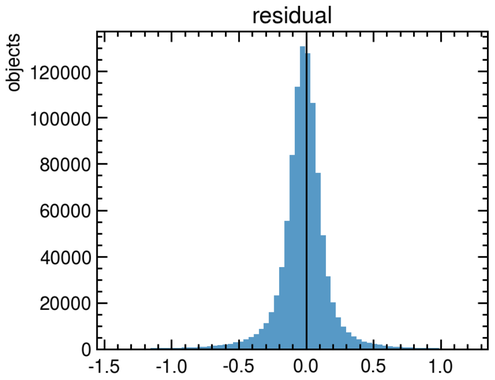


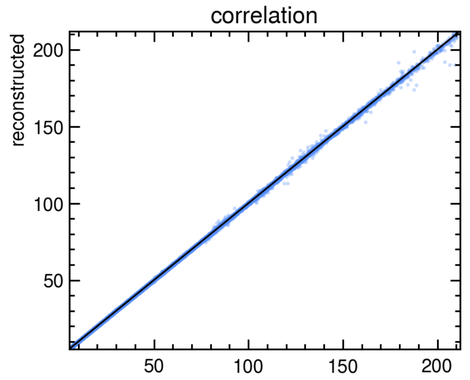


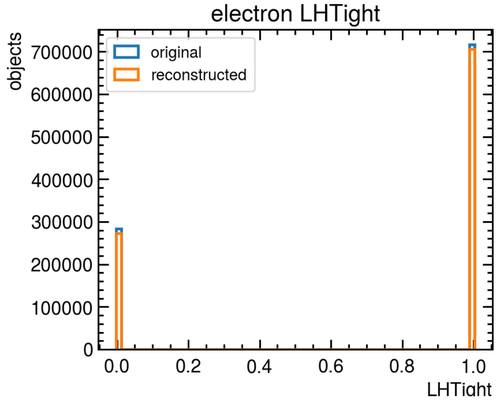


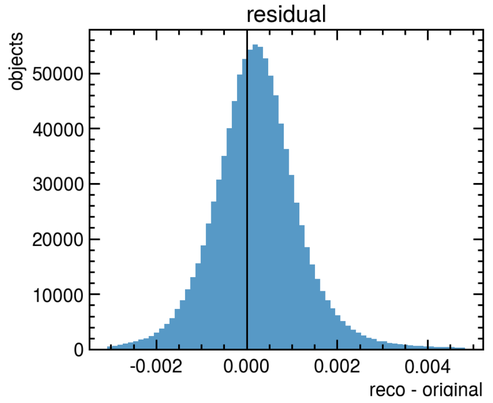
Reco - electrons

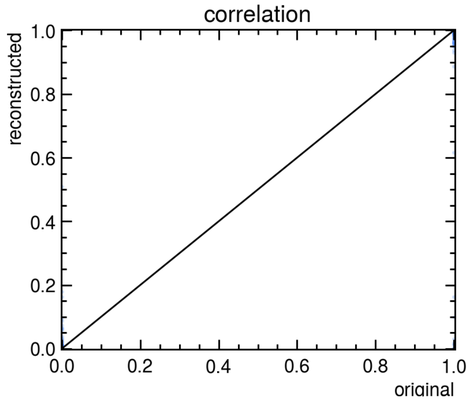


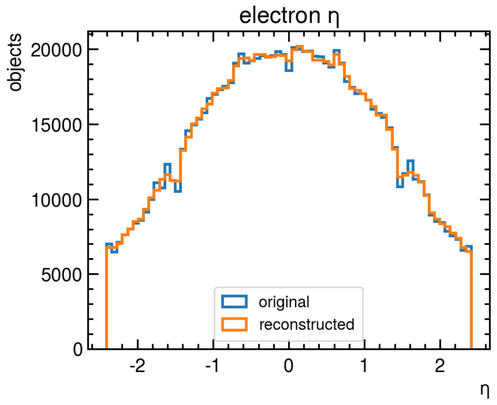


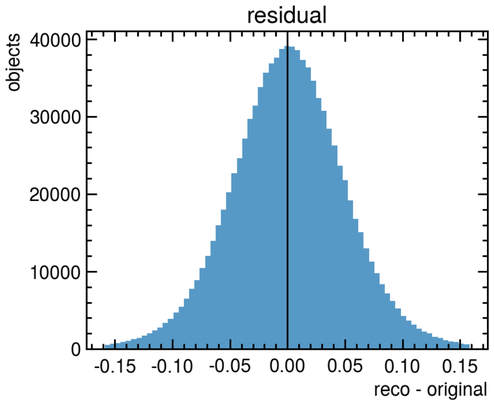


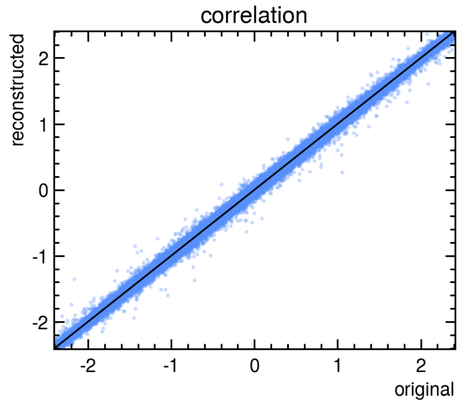


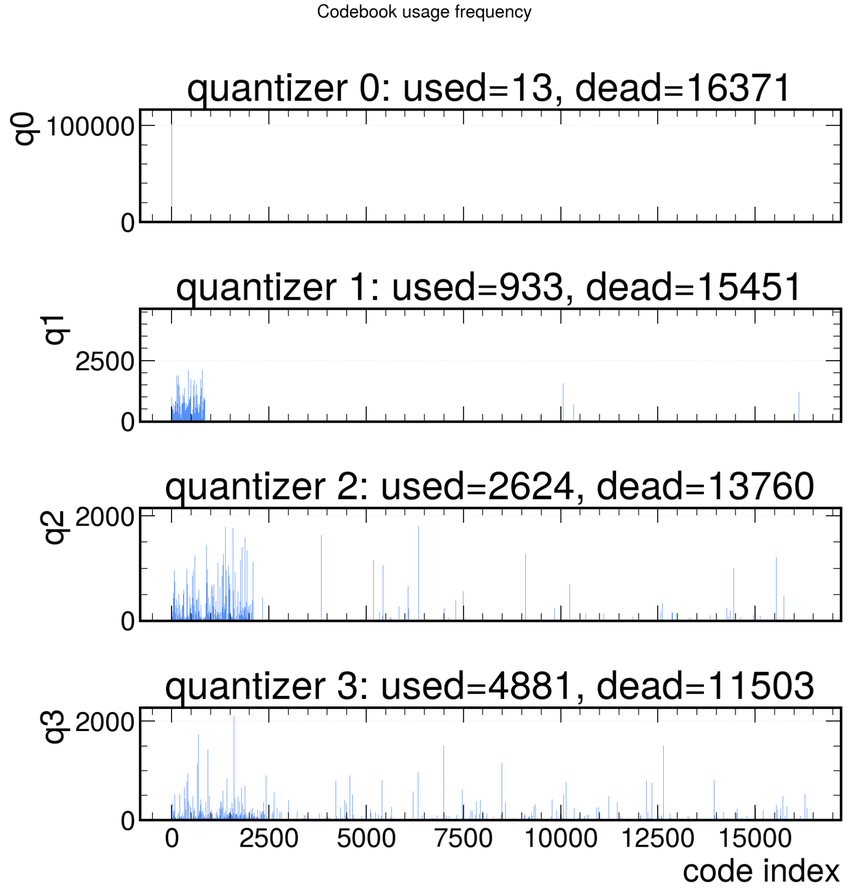


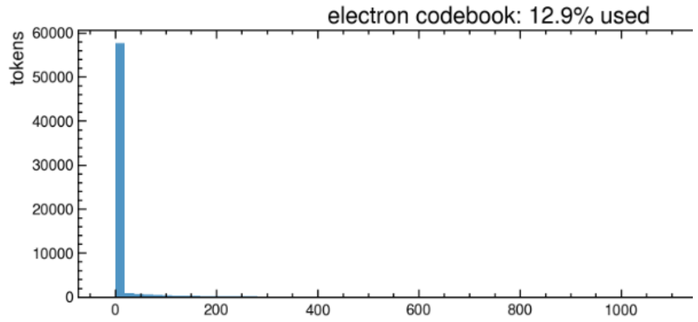






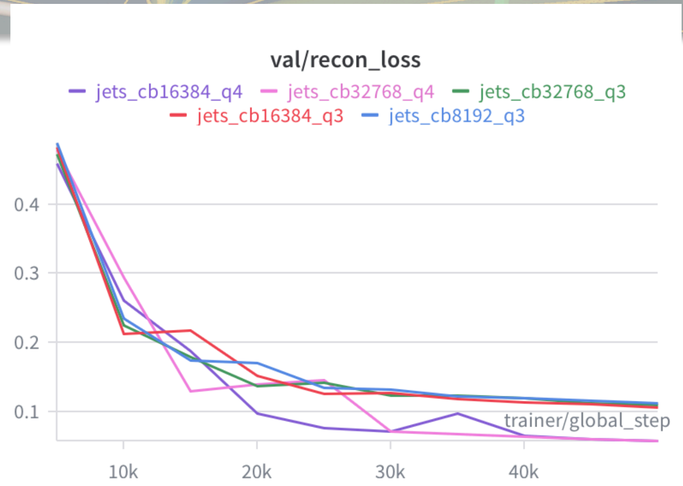


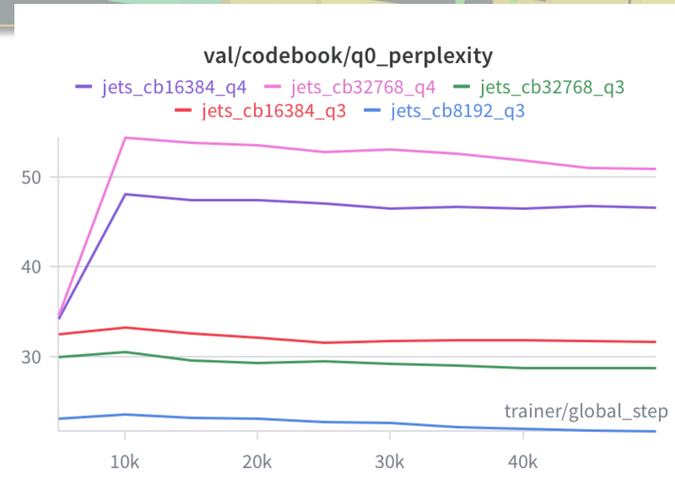


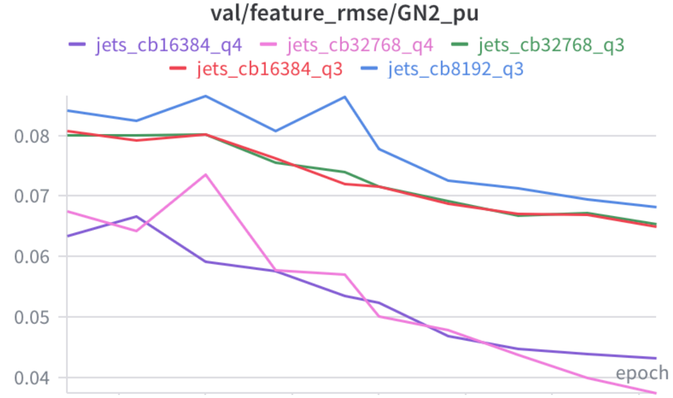


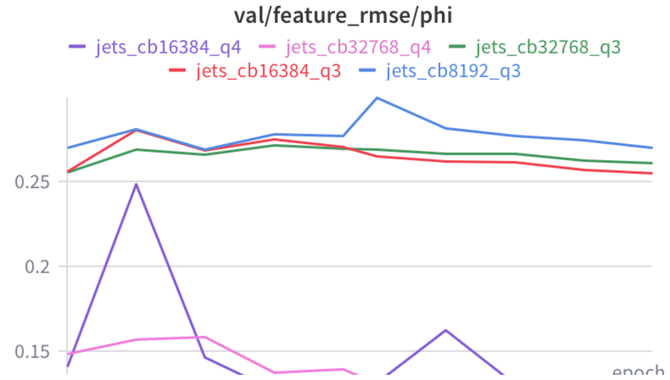
Jets

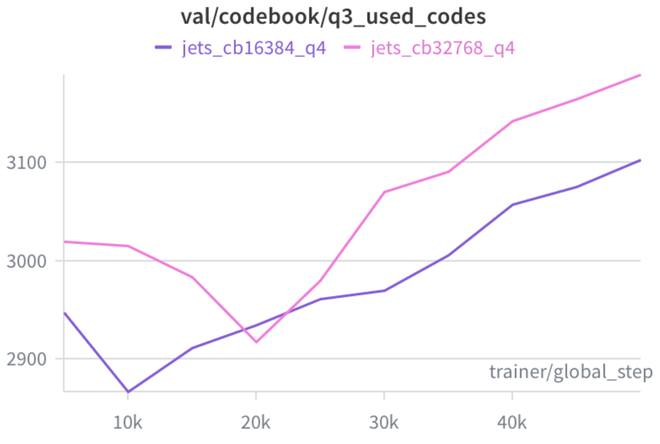
- 32k q4 is not clearly better
- It Improves over the q3 runs, but it usually does not beat 16384 q4
  - For q4, q3 is active and its perplexity increases, so the 4th quantizer is actually helping
- It also uses a smaller fraction of the codebook
- 8192 q3 is worse
- q0 is still collapsed for everything
- Best reco: **16384_q4**

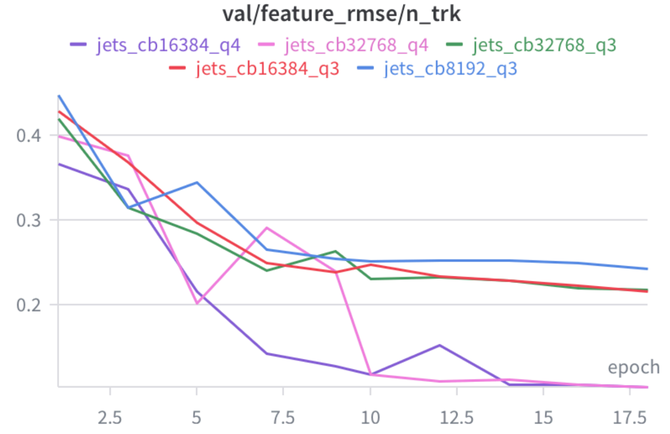




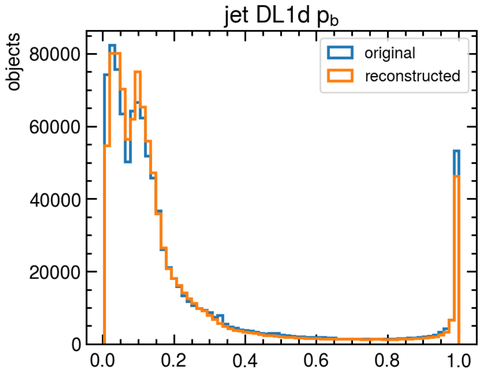


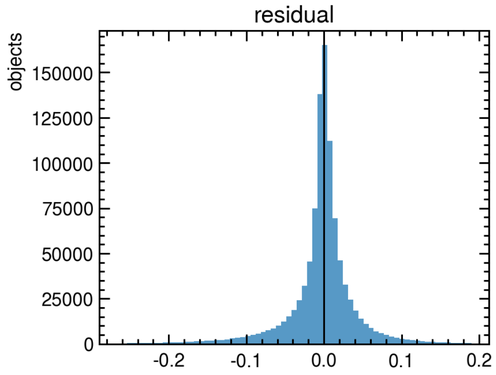


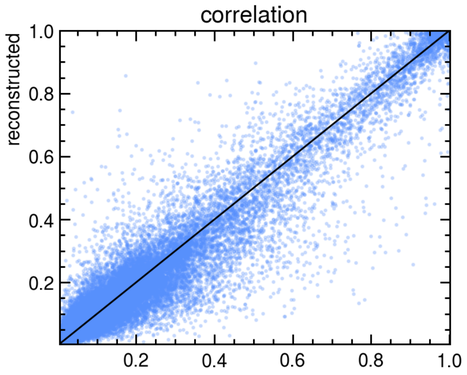


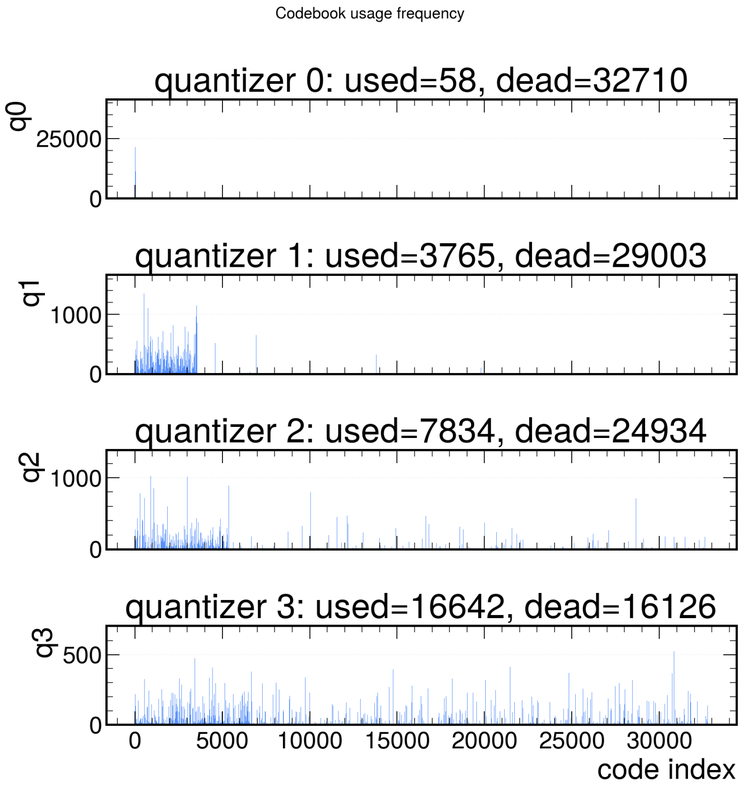


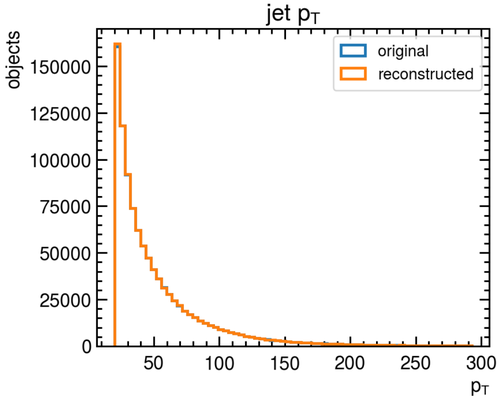
Reco - Jets

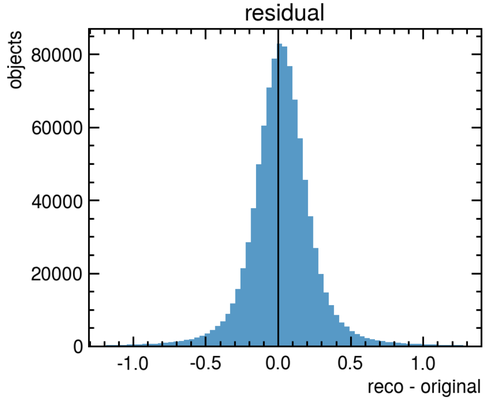

jet DL1d p_b
original
reconstructed
objects

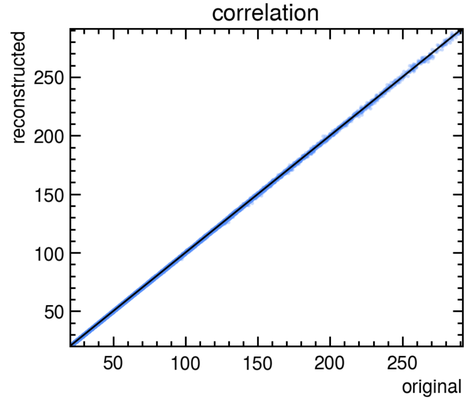

residual
objects

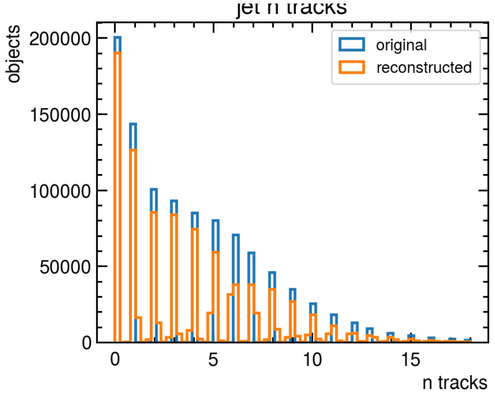

correlation
reconstructed

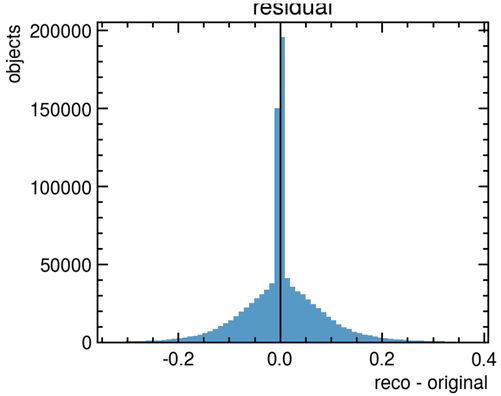

Codebook usage frequency
quantizer 0: used=58, dead=32710
quantizer 1: used=3765, dead=29003
quantizer 2: used=7834, dead=24934
quantizer 3: used=16642, dead=16126
code index

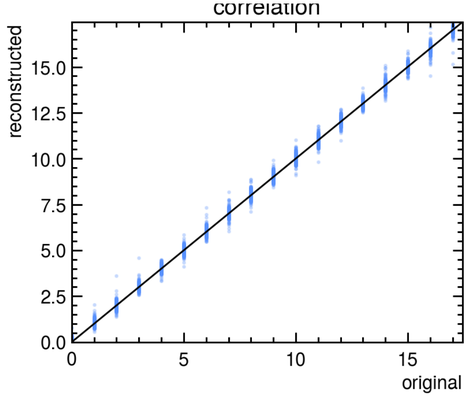

jet p_T
original
reconstructed
objects
p_T

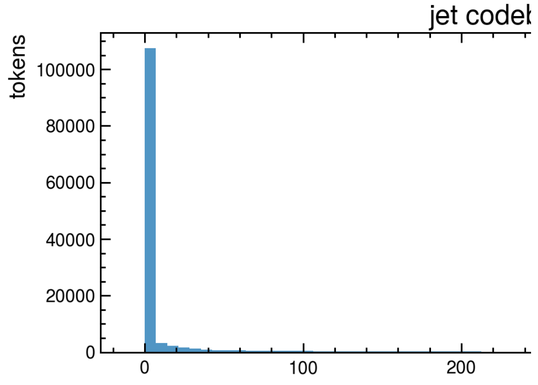

residual
objects
reco - original


correlation
reconstructed
original


jet n tracks
original
reconstructed
objects
n tracks

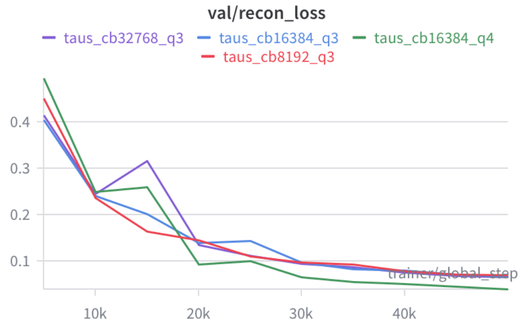

residual
objects
reco - original

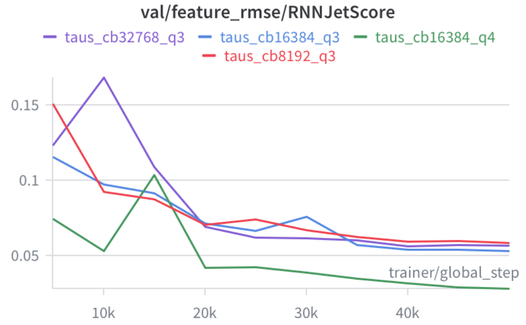

correlation
reconstructed
original

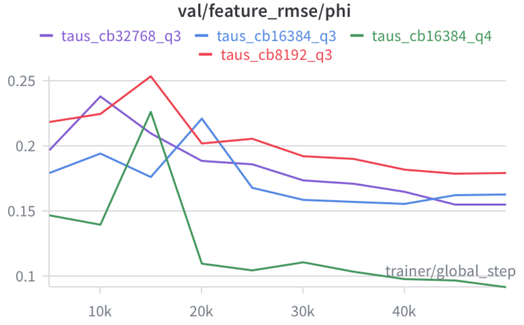

jet codes
tokens

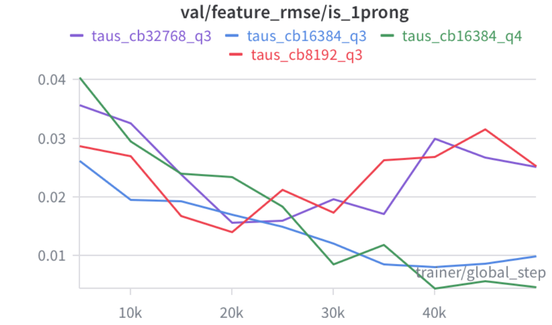
Taus

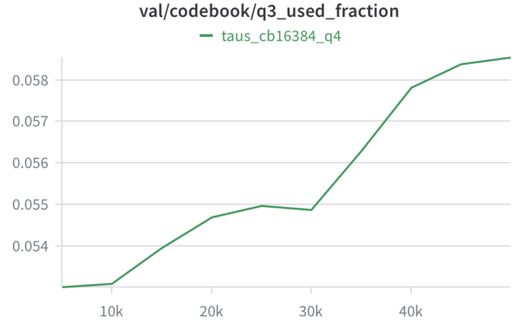
- The best reco: **taus_cb16384_q4**
- Lowest val/total_loss, val/recon_loss
- Best reconstruction for pt, eta, phi
- Clear improvement for RNNjetScore
- An active fourth quantizer, with about 950 used q3 codes and increasing perplexity

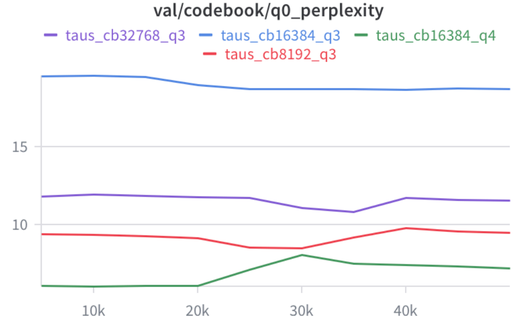




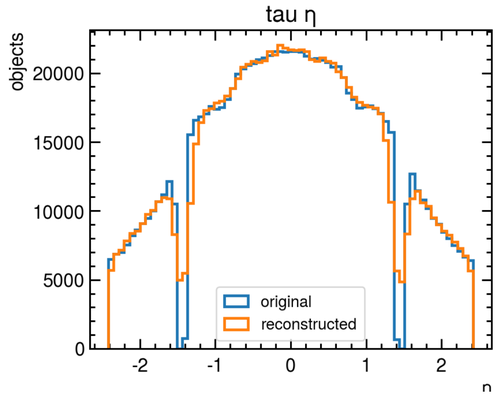


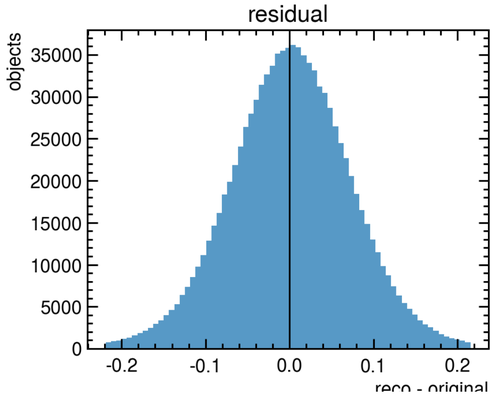


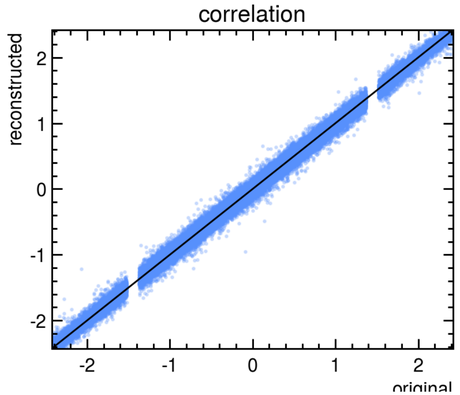


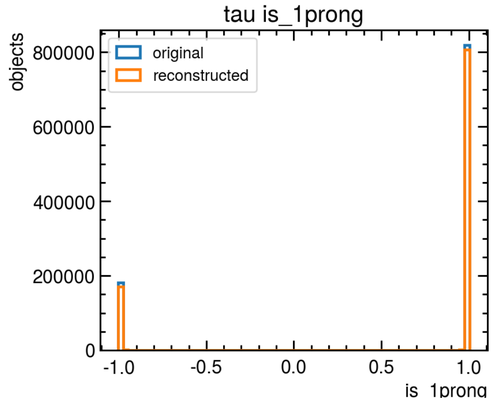


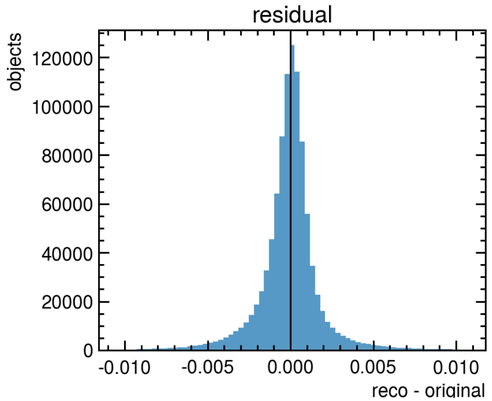
Reco

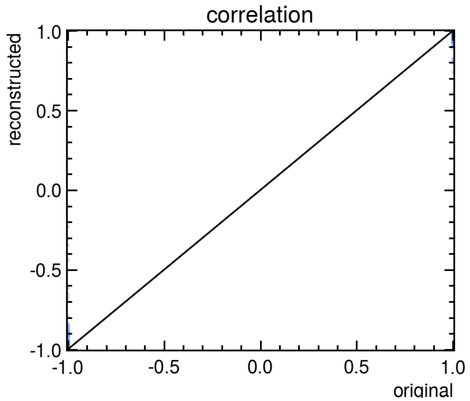

tau η
objects
original
reconstructed
η

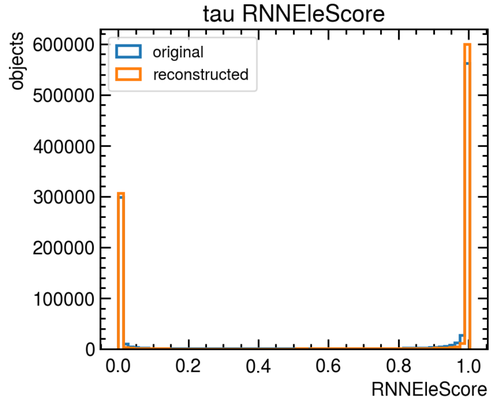

residual
objects
reco - original

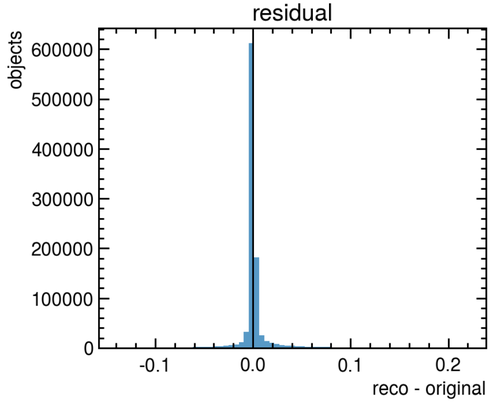

correlation
reconstructed
original

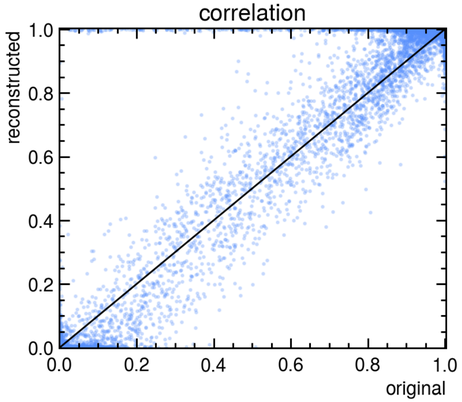

tau is_1prong
objects
original
reconstructed
is_1prong

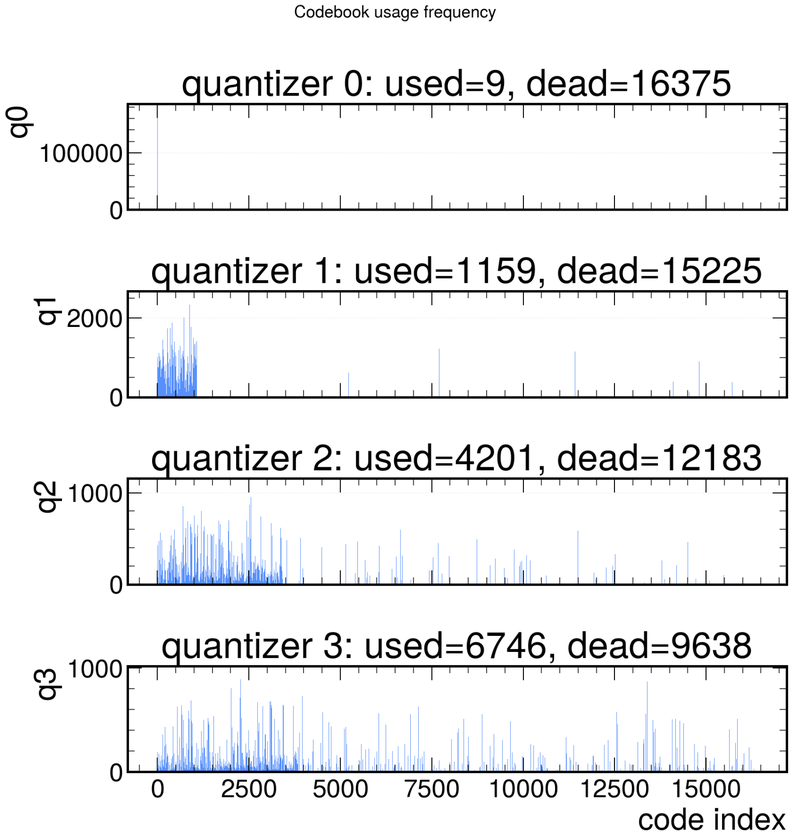

residual
objects
reco - original

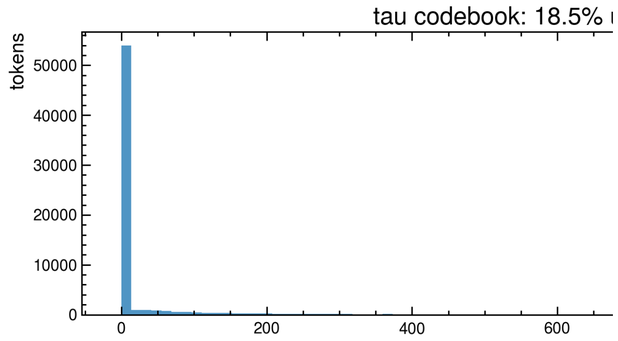

correlation
reconstructed
original


tau RNNEleScore
objects
original
reconstructed
RNNEleScore


residual
objects
reco - original


correlation
reconstructed
original


Codebook usage frequency
quantizer 0: used=9, dead=16375
q0
quantizer 1: used=1159, dead=15225
q1
quantizer 2: used=4201, dead=12183
q2
quantizer 3: used=6746, dead=9638
q3
code index


tau codebook: 18.5%
tokens

Summary

* Summary

- 16384 x4 gives the best reconstruction for electrons, jets, and taus.
- **Two issues: q0 collapse, and many unused codebook entries.**
  - **Encoder outputs occupy a small number of coarse clusters.**
  - **q0 only needs a few codes to describe the broad structure.**
- The fourth quantizer is active and improves reconstruction.
- Larger codebooks improve some difficult features, but much of the capacity remains unused.
- Dead codes show inefficient use of capacity, but do not necessarily mean poor reconstruction.

* Possible Reasons

- The feature space may only contain a small number of coarse object patterns
- Continuous, binary, and categorical variables are all trained with the same reconstruction loss.
- Some features have very different scales or contain little useful variation.
- Codebook initialization, commitment loss, or encoder output scale may contribute to q0 collapse.
- Data and simulation differences may affect reconstruction.

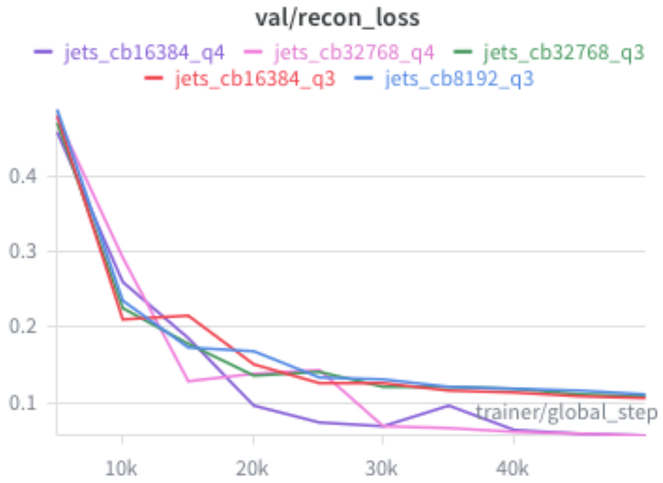
Next things to try

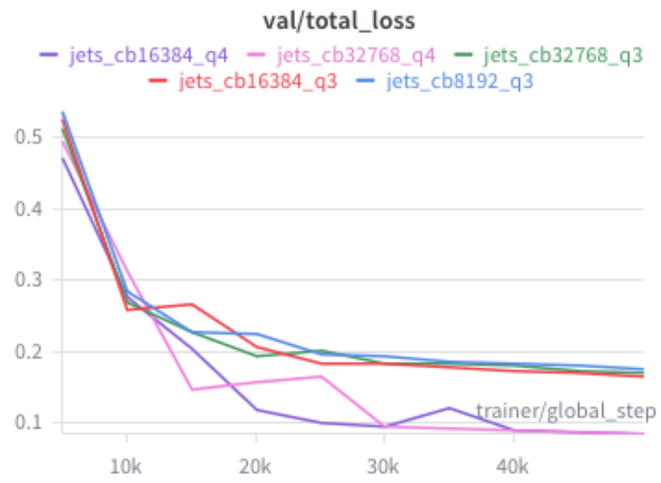
* Try feature weighting or separate losses for continuous, tagger, binary, and categorical variables.
* Scan smaller codebook_dim values, such as 8.
* Add k-means codebook initialization instead of random (q0 check)
* Reset dead codes during training (Codes used fewer than the threshold are replaced with vectors from the current batch)
* Try cosine-similarity quantization
* Log the encoder output before changing more parameters
* Compare tokenizers using downstream transformer
* Add derived physics checks such as invariant masses and angular separations after decoding

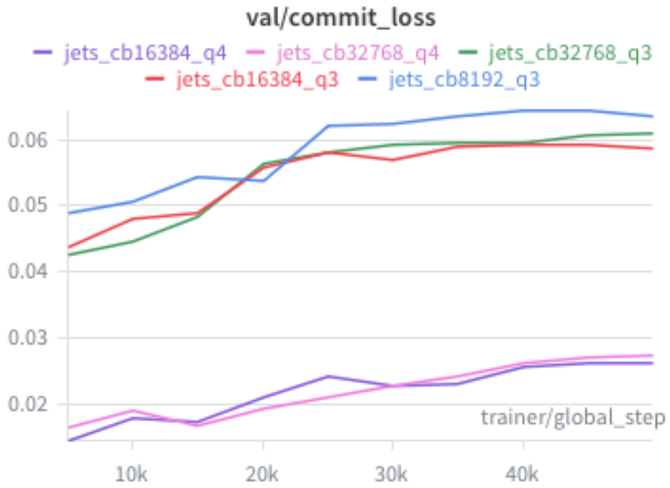
Backup

Loss

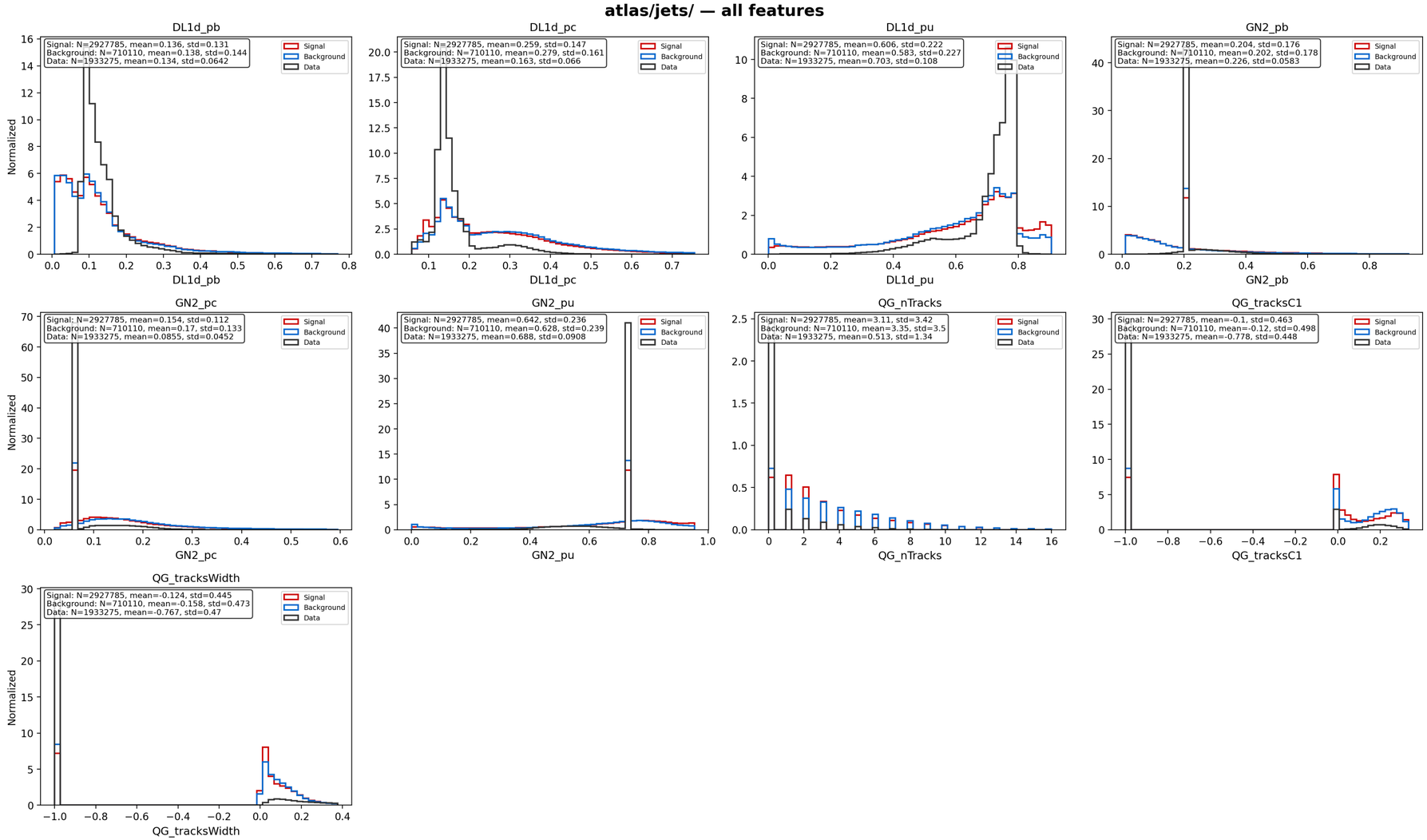
* Commitment loss encourages encoder outputs to stay close to selected code vectors
* total_loss = reconstruction_loss + commitment_loss
* commitment_weight=1.0
* Commitment weight is contributing to optimisation but still smaller
* Try different weights?
* Try different initialisation of the codebooks?





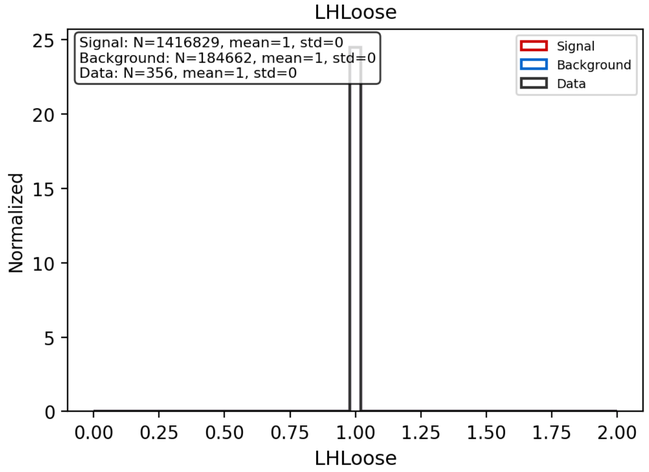


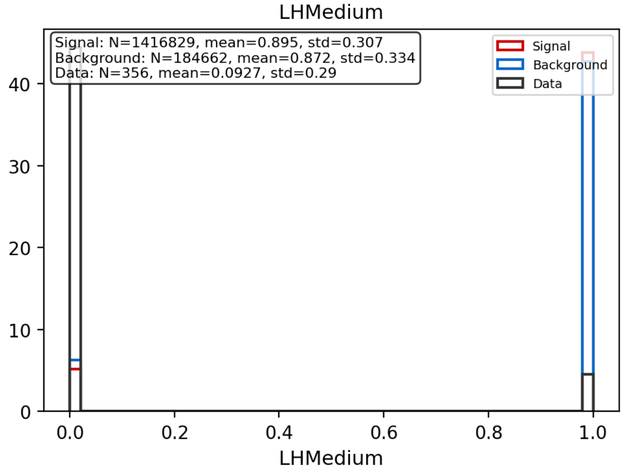
Data vs MC

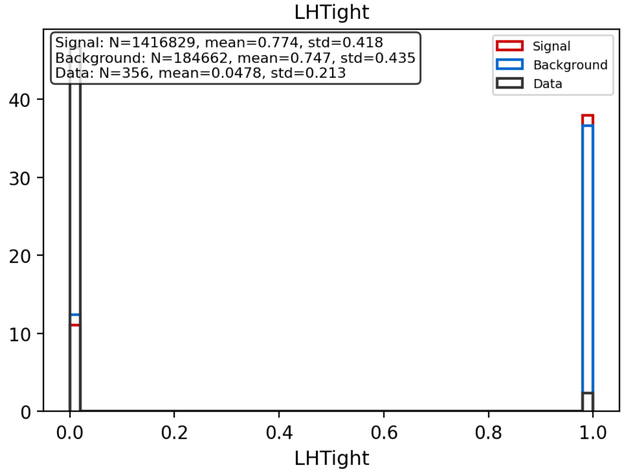

atlas/jets/ — all features
DL1d_pb
Signal: N=2927785, mean=0.136, std=0.131
Background: N=710110, mean=0.138, std=0.144
Data: N=1933275, mean=0.134, std=0.0642
DL1d_pc
Signal: N=2927785, mean=0.259, std=0.147
Background: N=710110, mean=0.279, std=0.161
Data: N=1933275, mean=0.163, std=0.066
DL1d_pu
Signal: N=2927785, mean=0.606, std=0.222
Background: N=710110, mean=0.583, std=0.227
Data: N=1933275, mean=0.703, std=0.108
GN2_pb
Signal: N=2927785, mean=0.204, std=0.176
Background: N=710110, mean=0.202, std=0.178
Data: N=1933275, mean=0.226, std=0.0583
GN2_pc
Signal: N=2927785, mean=0.154, std=0.112
Background: N=710110, mean=0.17, std=0.133
Data: N=1933275, mean=0.0855, std=0.0452
GN2_pu
Signal: N=2927785, mean=0.642, std=0.236
Background: N=710110, mean=0.628, std=0.239
Data: N=1933275, mean=0.688, std=0.0908
QG_nTracks
Signal: N=2927785, mean=3.11, std=3.42
Background: N=710110, mean=3.35, std=3.5
Data: N=1933275, mean=0.513, std=1.34
QG_tracksC1
Signal: N=2927785, mean=-0.1, std=0.463
Background: N=710110, mean=-0.12, std=0.498
Data: N=1933275, mean=-0.778, std=0.448
QG_tracksWidth
Signal: N=2927785, mean=-0.124, std=0.445
Background: N=710110, mean=-0.158, std=0.473
Data: N=1933275, mean=-0.767, std=0.47
Signal
Background
Data
Normalized

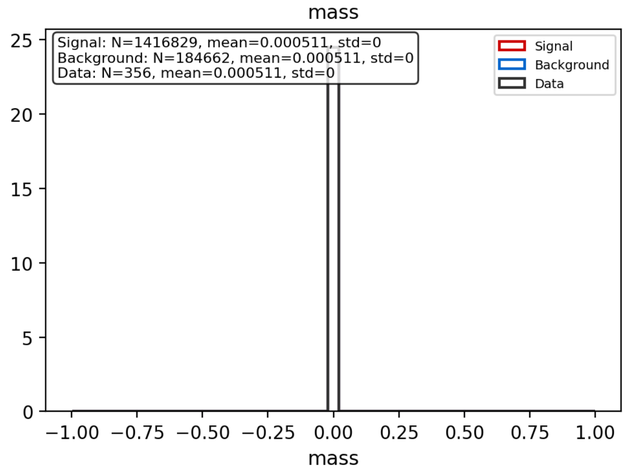
Data vs MC

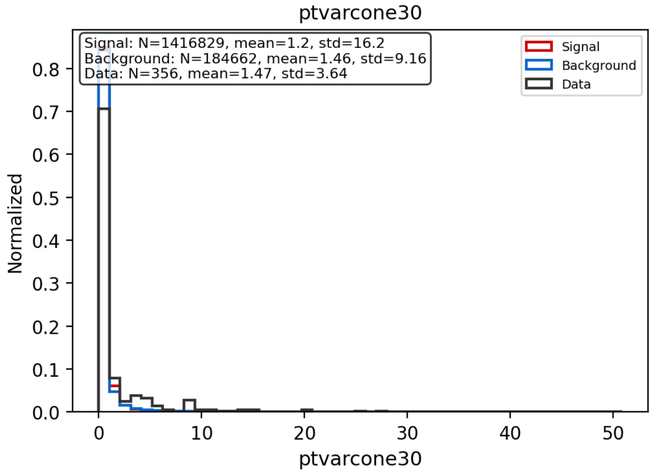
atlas/electrons/ — all features

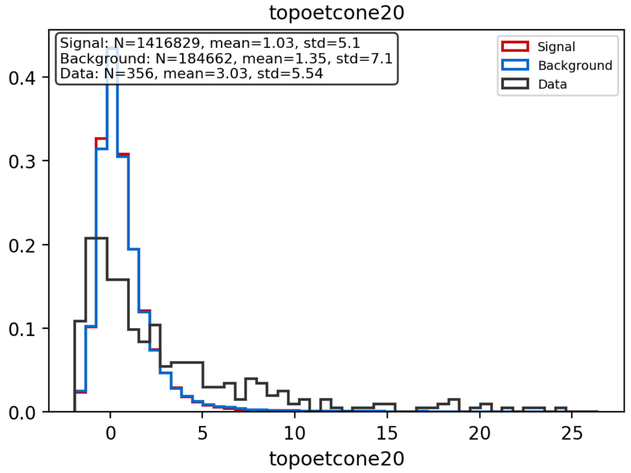












Data vs MC

common/event/ — all features