Treasure event level-tokenisation

Merve Nazlim Agaras

10.6.26

Last week

Summary

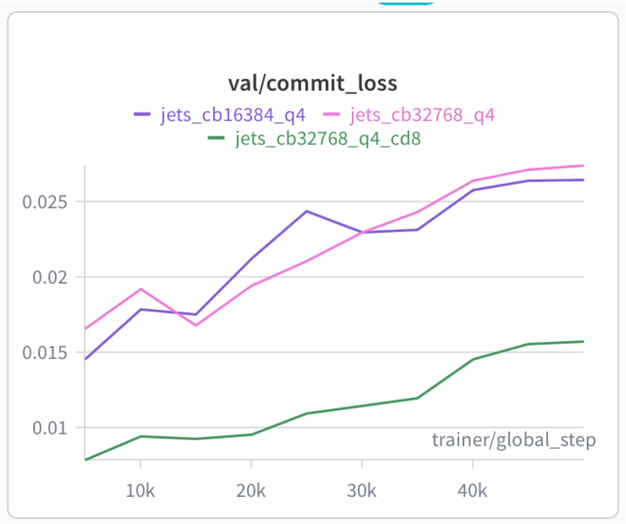
- 16384 x4 gives the best reconstruction for electrons, jets, and taus.
- **Two issues: q0 collapse, and many unused codebook entries.**
  - **Encoder outputs occupy a small number of coarse clusters.**
  - **q0 only needs a few codes to describe the broad structure.**
- The fourth quantizer is active 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.

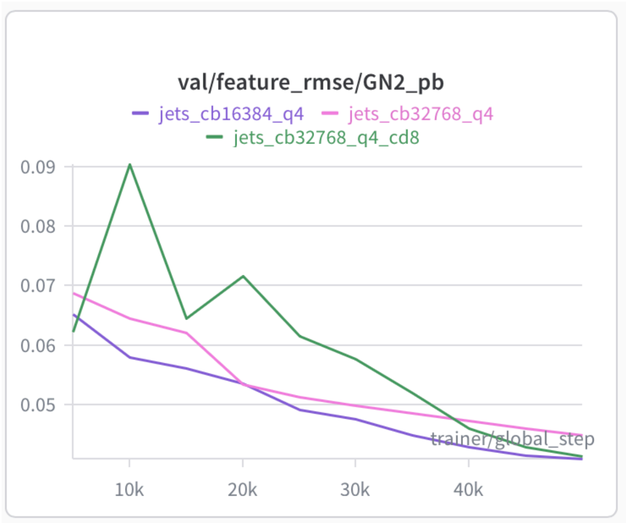
Things to try

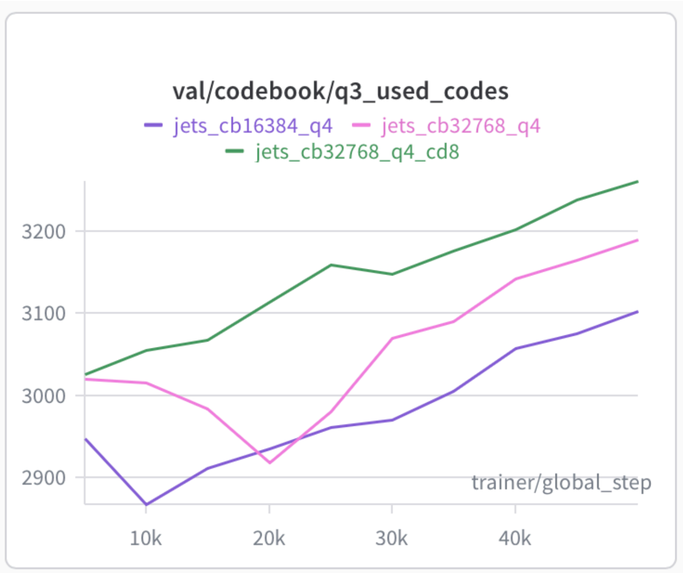
- Try feature weighting or separate losses for continuous, tagger, binary, and categorical variables.
- Increase the validation sample size
- Scan smaller codebook_dim values, such as 8.
- Add k-means codebook initialization instead of random (q0 check)
- Reset dead codes during training (Codes used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

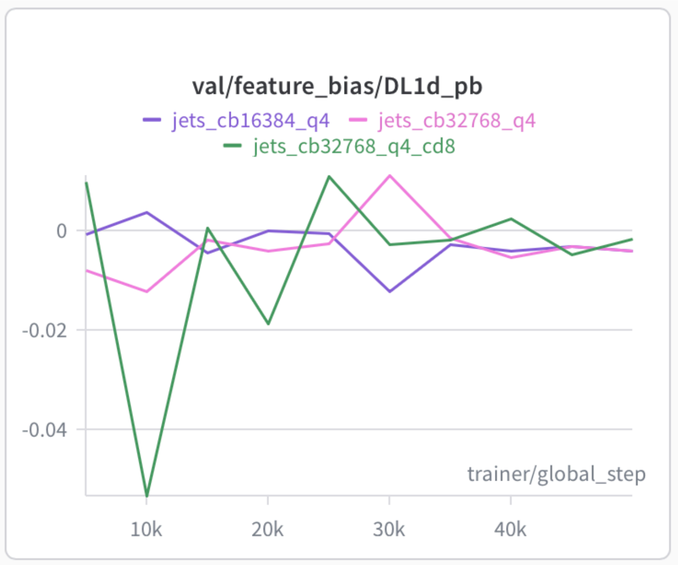
Codebook dimension studies

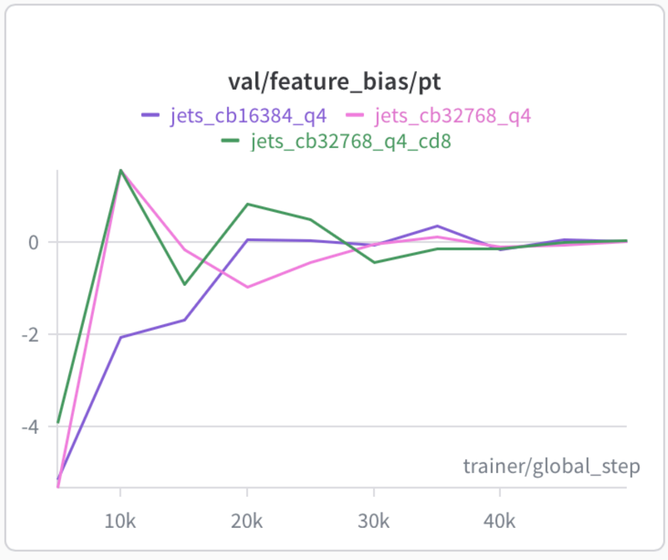
* Reducing the codebook dimension from 16 to 8 gives similar overall reconstruction performance and slightly increases q3 code usage.



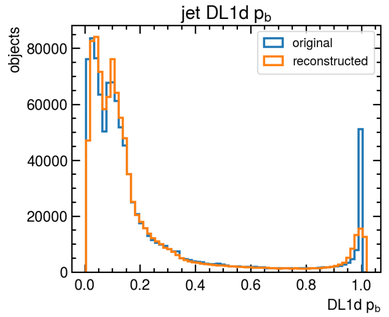


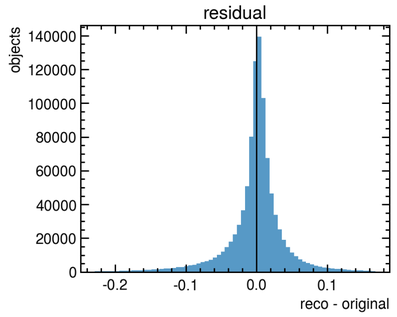


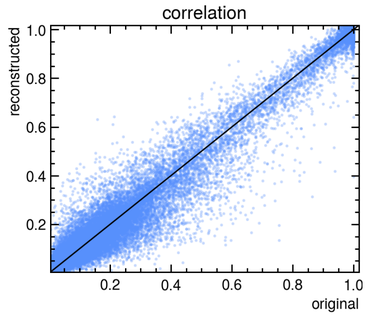


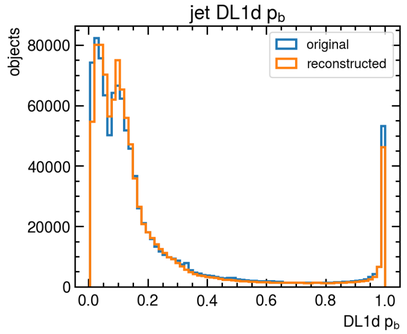


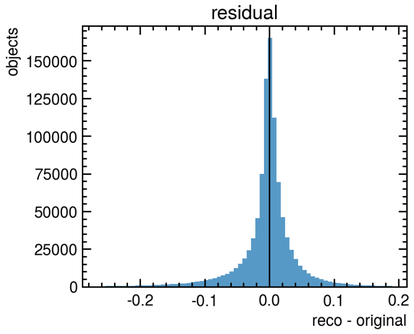
Codebook dimension studies

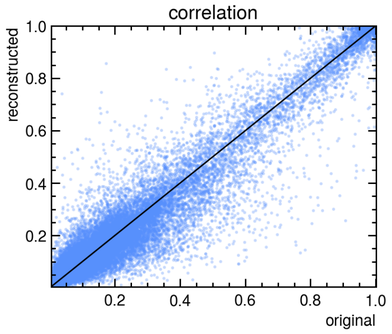


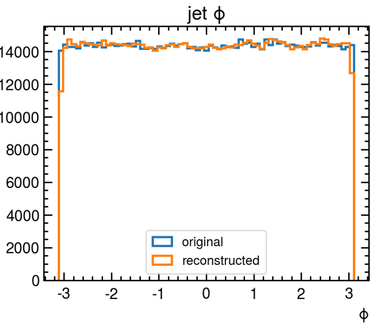


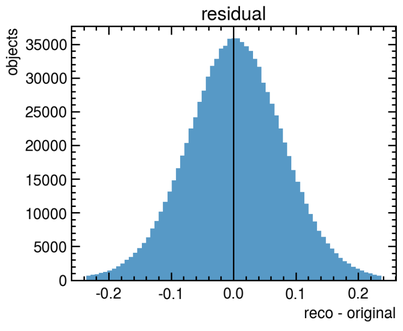


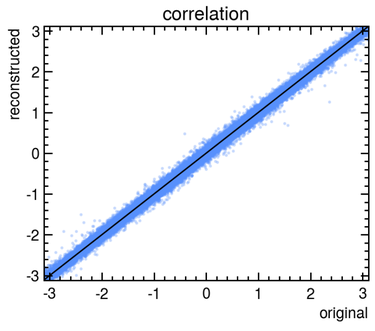


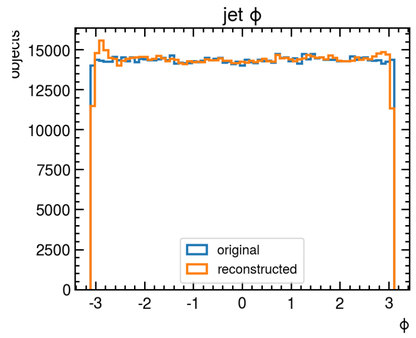


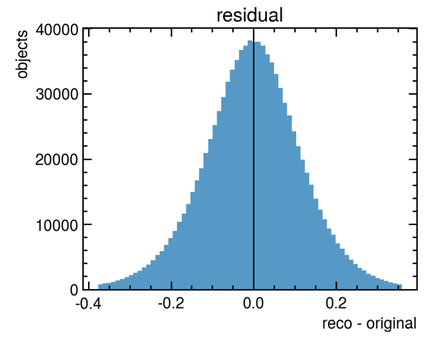


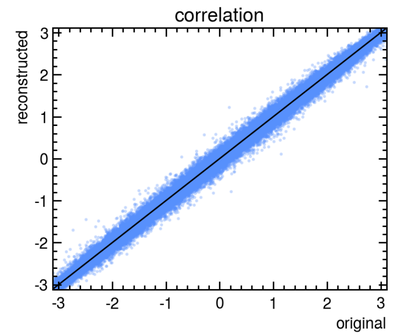


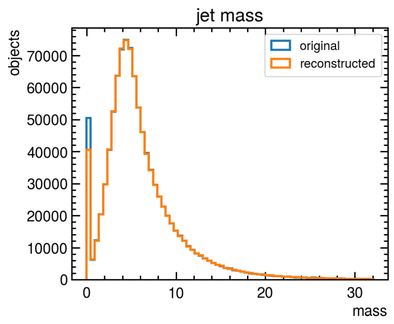


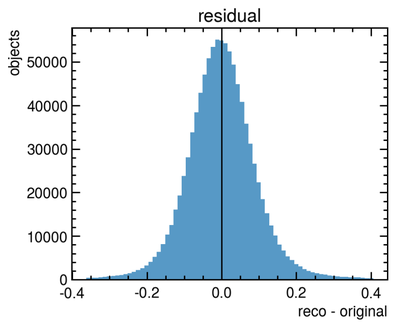


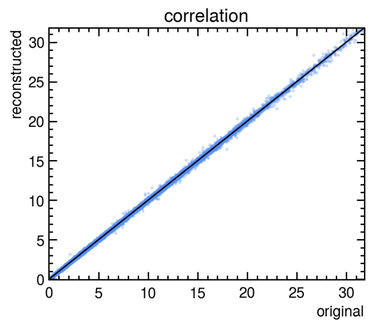


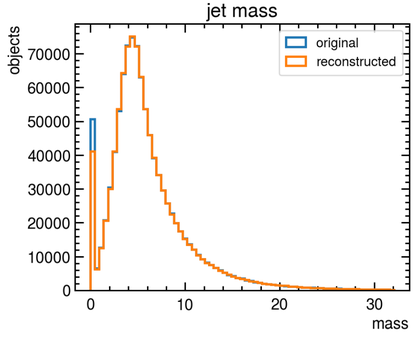


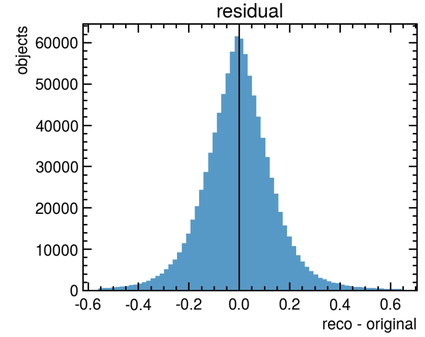


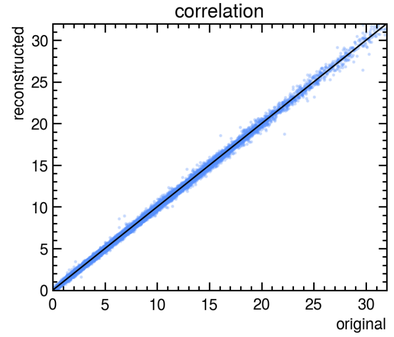










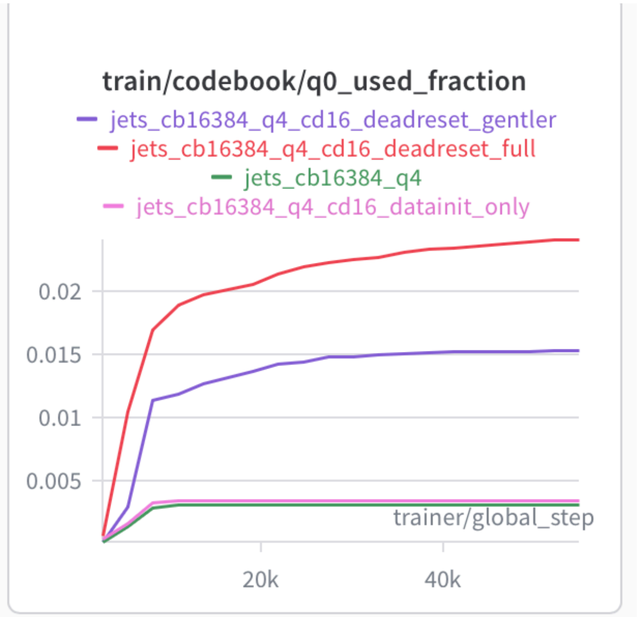


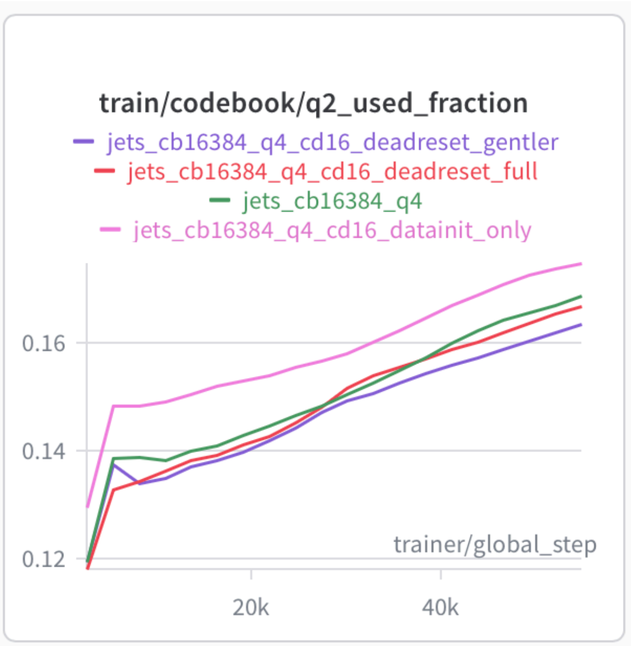
Codebook_dim = 8

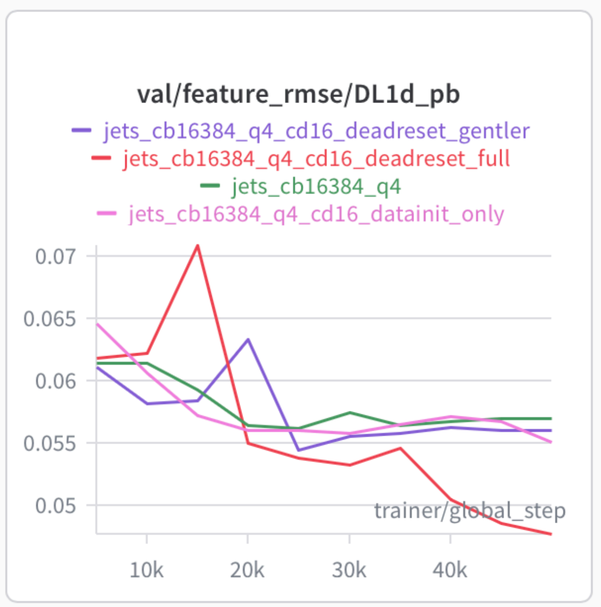
Codebook_dim = 16

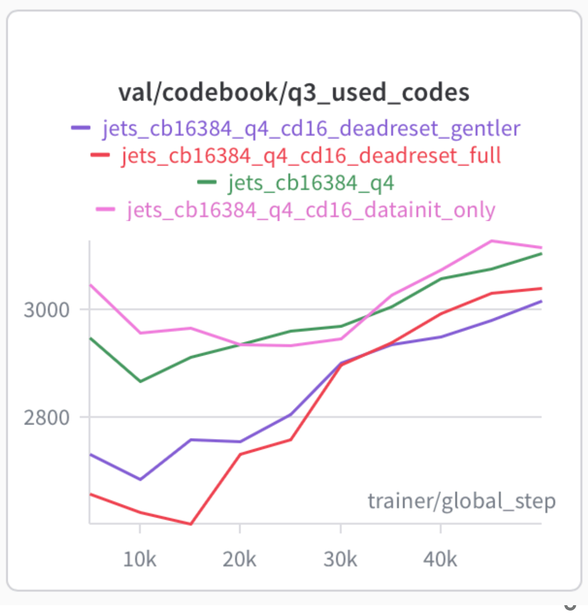
Reset deadcodes and initialisation

- Added dead-code resetting for q0:
  - q0 usage increased from roughly 40 codes to around 3k.
  - Reconstruction of some jet features stays the same/ slightly better.
  - However, approximately 85% of q0 remained unused between resets, and every reset reached its configured limit.
- Tested gentler resets every 5k steps with a 10% reset limit. q0 still collapsed, still used only around 1.4% of the codebook
- Tested data-based q0 initialization using 65k encoder outputs. q0 still collapsed to approximately 58 used codes, showing that random initialization was not the main cause.
  - the initial codebook vectors from real encoder outputs instead of starting with random vectors
- Concluded that q0 naturally uses a small coarse vocabulary. Whether this harms performance still needs to be tested downstream.

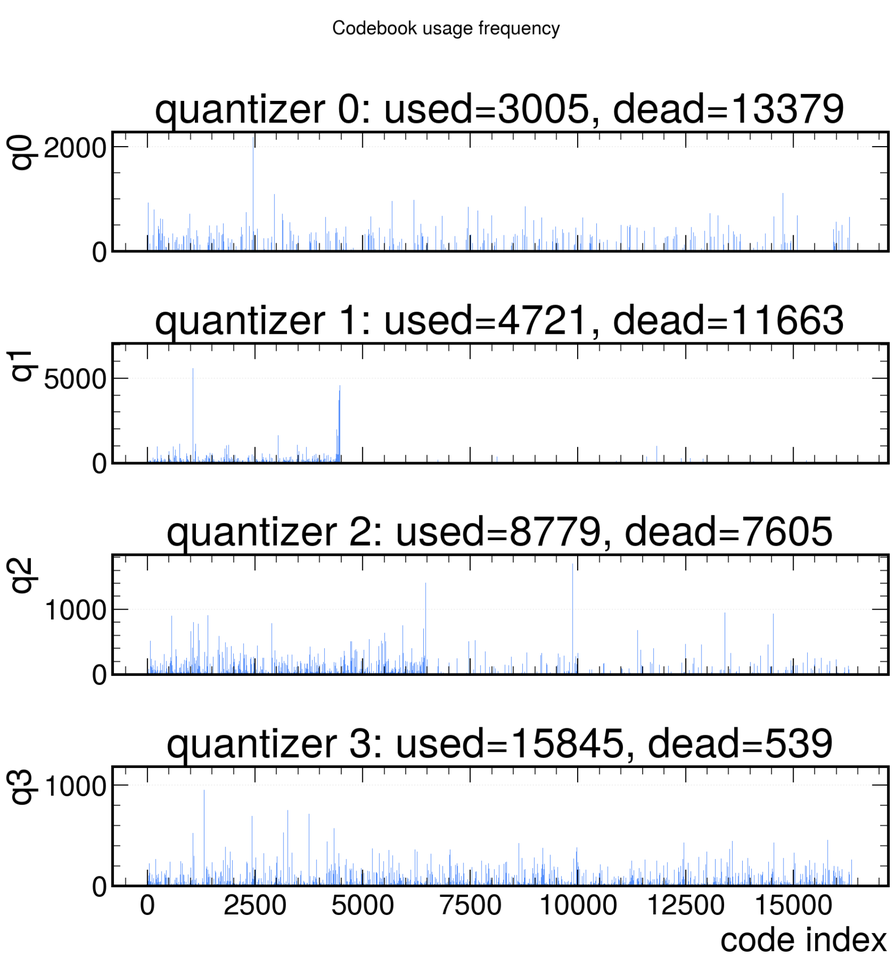


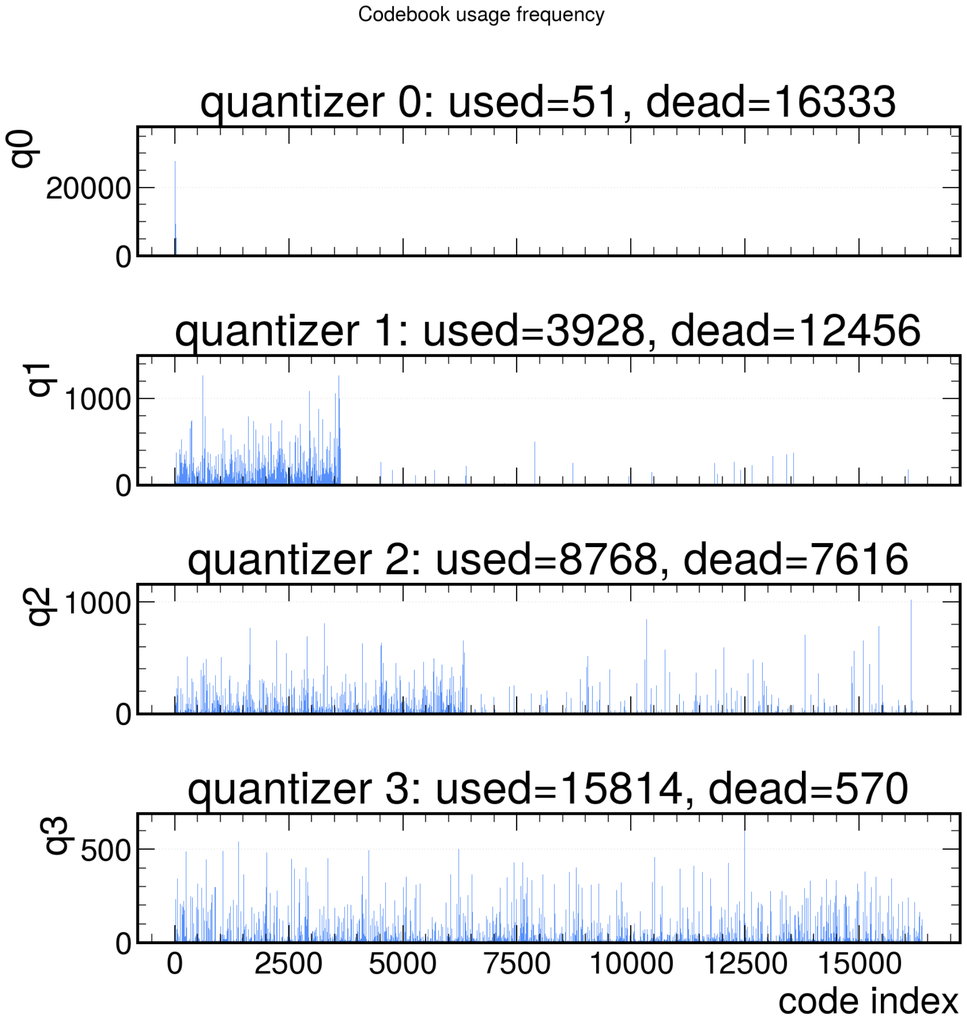






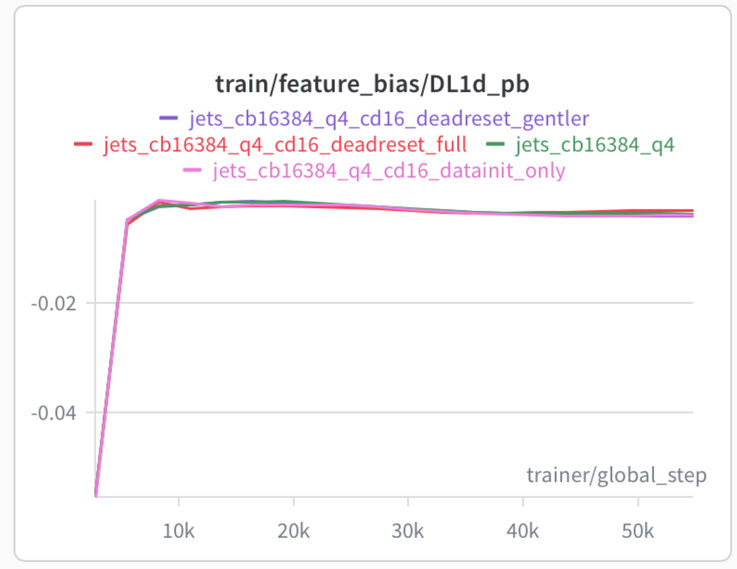
Q0 hard reset vs no reset

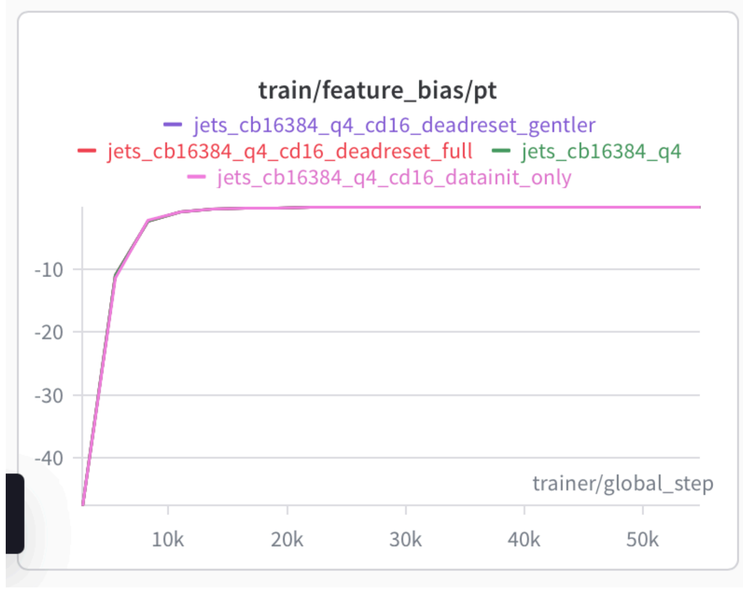


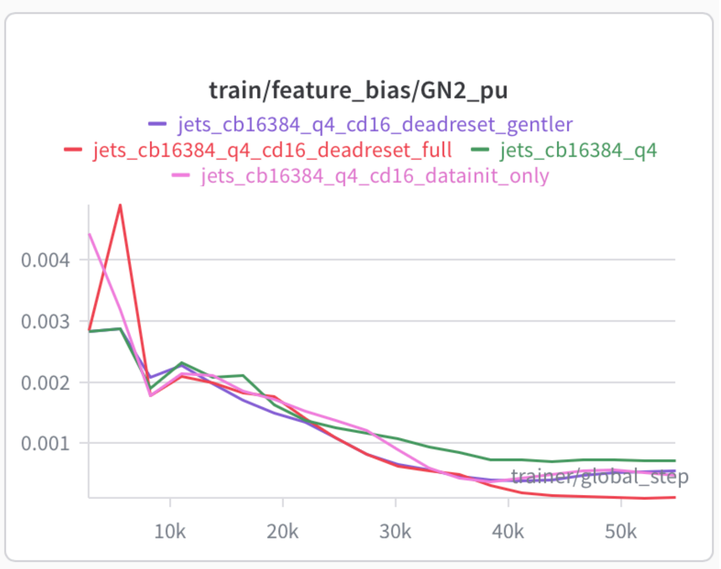


Reconstruction loss

* Current loss is one unweighted L1 average across every feature
* The decoder quickly learns the feature's overall mean/location.
* Reconstructed features remains systematically about 0.005 below the original value.
* The remaining biases plateau and vary by feature, suggesting that longer training alone is unlikely to remove them.
* Feature-weighted or feature-specific losses could be tested to reduce residual biases (bias should be interpreted relative to each feature's scale)







Sequence check

- Measure how often event sequences exceed 128 tokens.
- The sequence also contains event [CLS] and separator tokens.
- If we use 4 quantiser
  - 1 CLS + 1 event token + 5 separators+ 4 × (electrons + muons + taus + photons + jets)
  - If an event contains 35 objects: -> 35 × 4 + special tokens ≈ 147 tokens
- Current limit is 128 and order is (electrons, muons, taus, photons, jets)
- Checked 4M events
  - Median leng: 35
  - 99% below 62
  - Maximum length: 123
- No sequences were truncated

Summary

- Reducing the codebook dimension from 16 to 8 gives similar overall reconstruction and slightly higher late-quantizer usage, but does not resolve q0 collapse.
- Data-based initialization also does not prevent q0 collapse, suggesting that random initialization is not the main cause.
- Dead-code resetting increases q0 coverage and slightly improves some reconstructed features, but many codes become inactive again between resets.
- Reconstruction is generally good, although some features retain small residual biases that do not improve with longer training.
- No event sequences were truncated

Next steps

* New samples ran by Viviana, rerun the tokenisation with them.
* Test feature-weighted or feature-specific reconstruction losses.
* Evaluate derived physics quantities.
* Include track objects and move to pretraining with data15/16+MC samples (ttbar, ZZ->4l, ZZ->gamma gamma, ZZ background)