Treasure pipeline discussion

Merve Nazlim Agaras

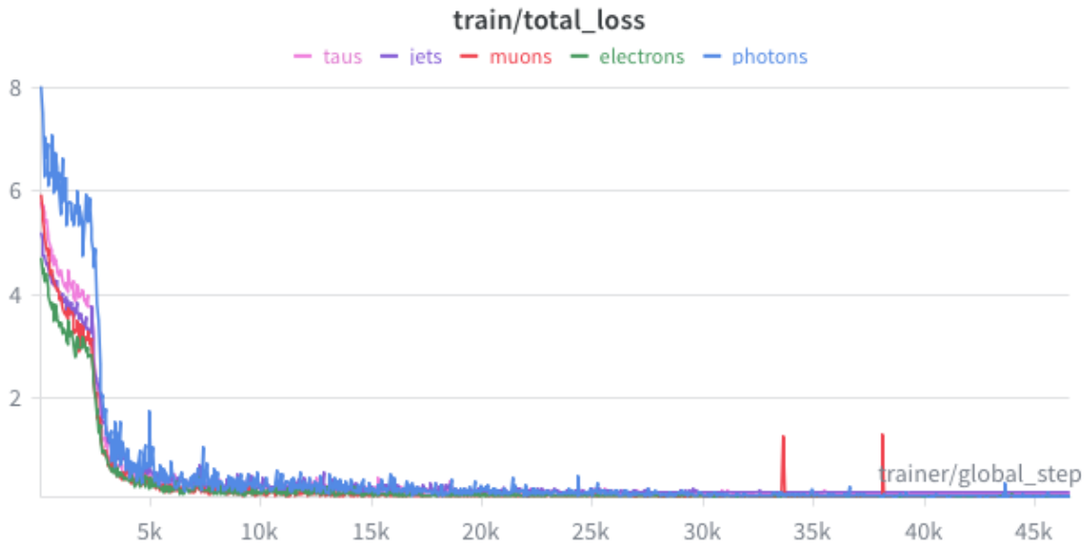
27.5.26

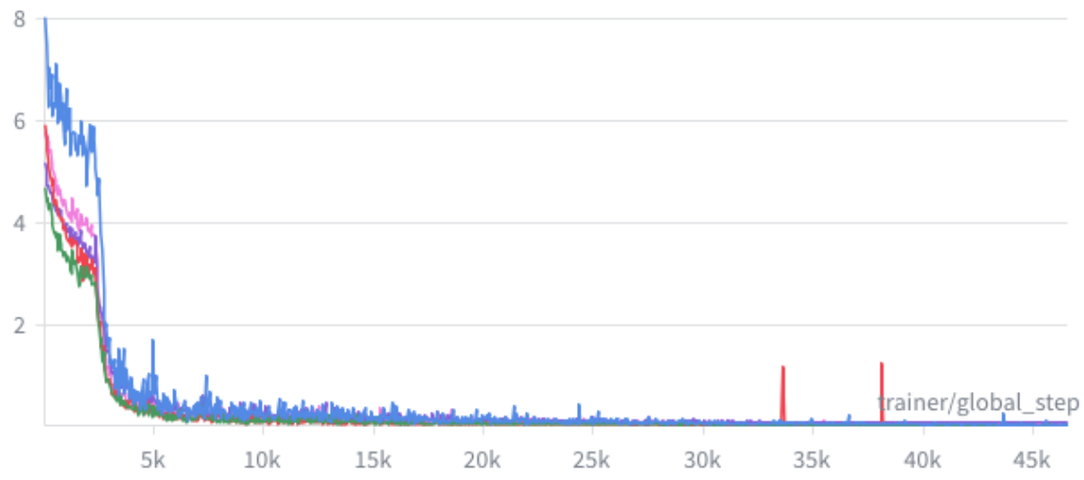
Intro

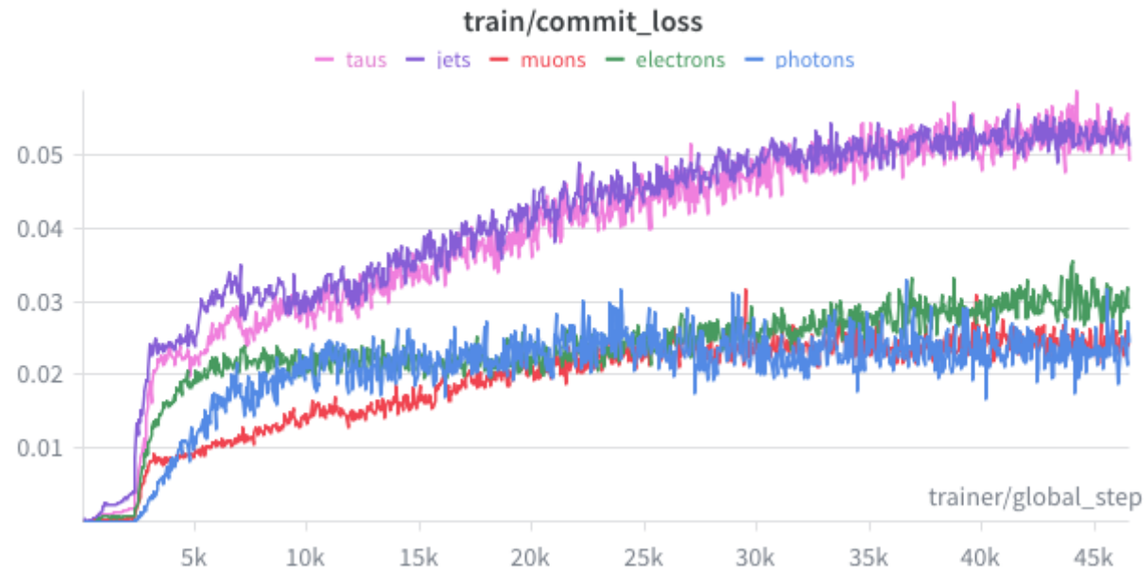
* Last week, we agreed to create a PR for event-level tokenisation to ‘heptokens’ repo.
* Link to PR: https://github.com/Treasure-AmSC/heptokens/pull/5
* PR adds ATLAS event-object dataloading for per-object VQ-VAE tokenizer training (jets, electrons, muons, photons, taus), with configurable H5 feature paths and tokenizer diagnostics.

* Ran over the recent files from Viviana, next slides for the results

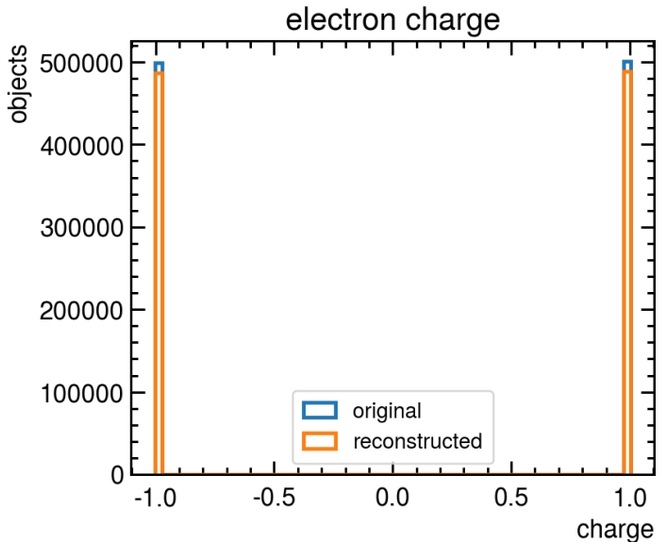
```
jets       events=3,558,000 objects=14,344,784
electrons  events=3,558,000 objects=3,150,807
muons      events=3,558,000 objects=3,945,218
photons    events=3,558,000 objects=688,912
taus       events=3,558,000 objects=4,260,772
```

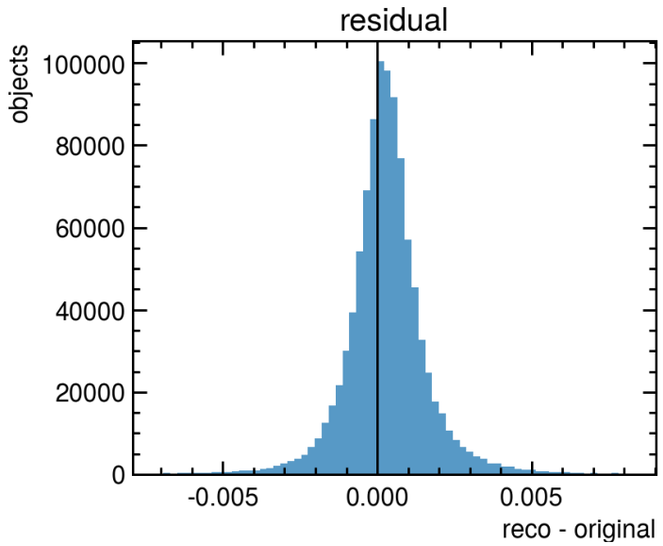

train/total_loss
taus
jets
muons
electrons
photons
8
6
4
2
5k
10k
15k
20k
25k
30k
35k
40k
45k
trainer/global_step

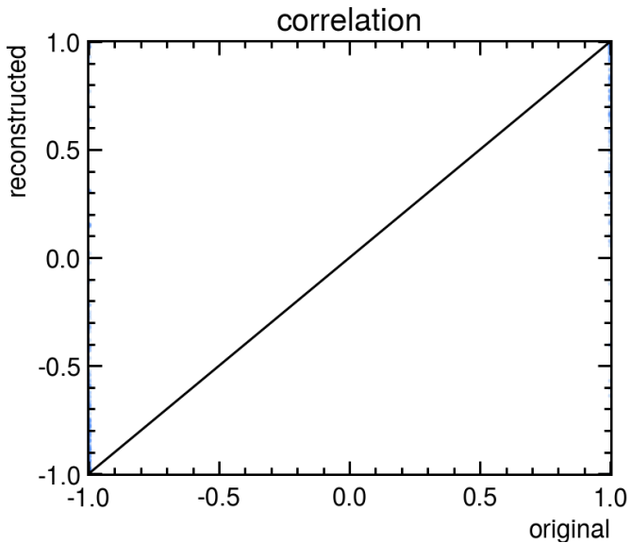

train/recon_loss
taus
jets
muons
electrons
photons
8
6
4
2
5k
10k
15k
20k
25k
30k
35k
40k
45k
trainer/global_step

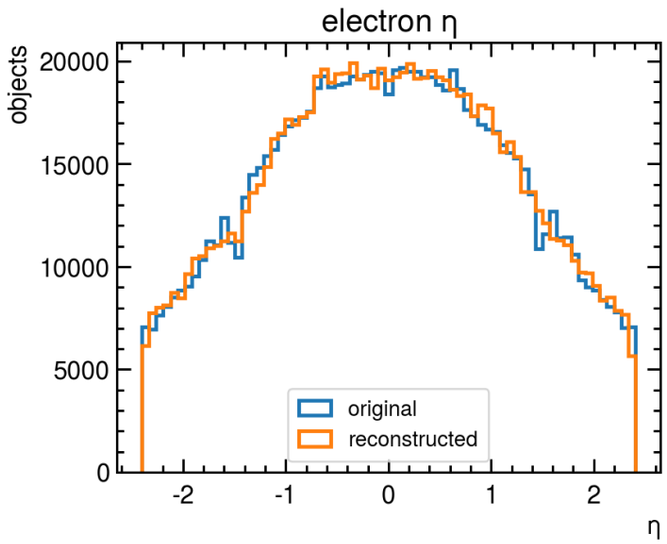

train/commit_loss
taus
jets
muons
electrons
photons
0.05
0.04
0.03
0.02
0.01
5k
10k
15k
20k
25k
30k
35k
40k
45k
trainer/global_step

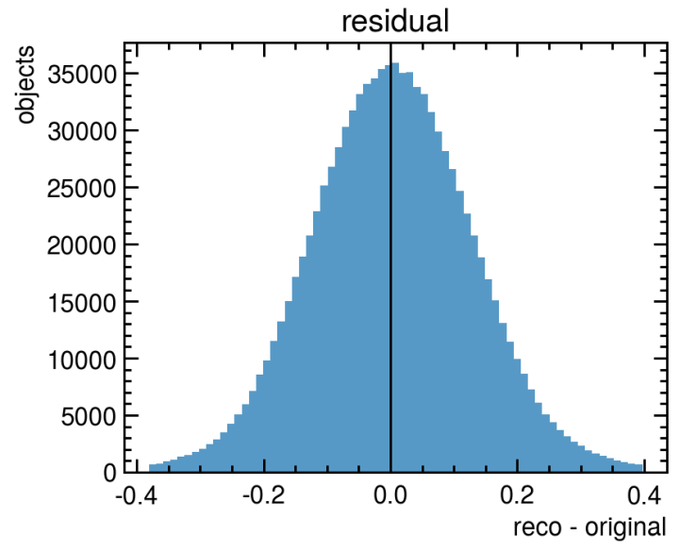
Diagnostics - electrons

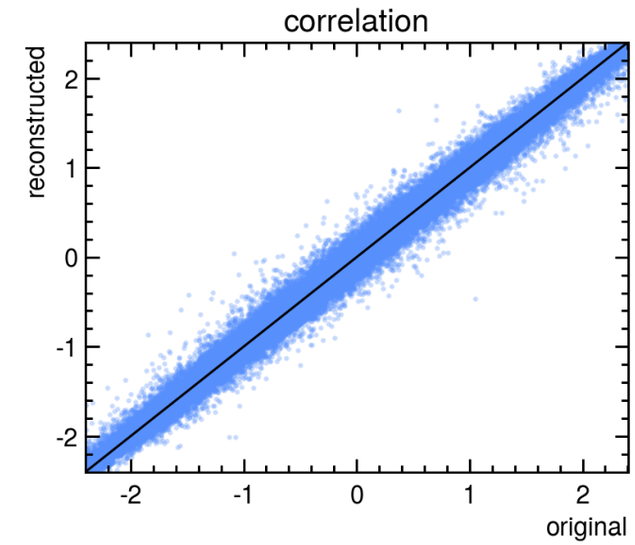


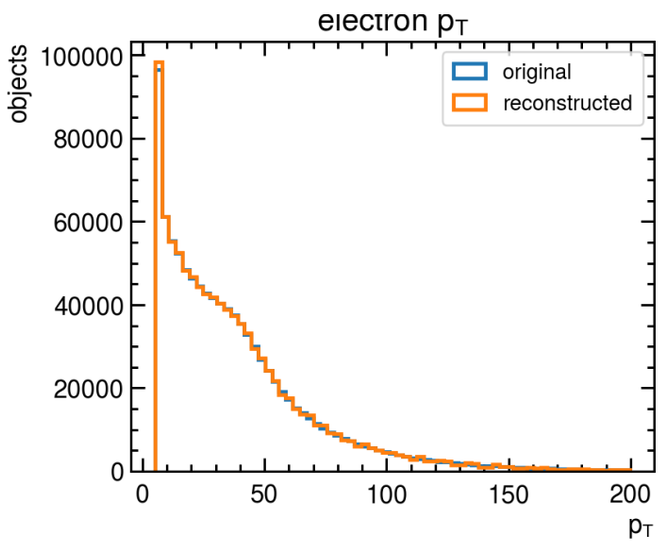


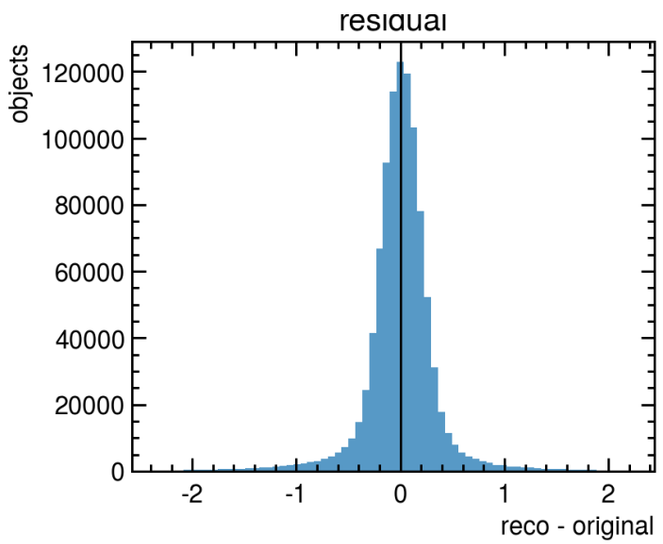


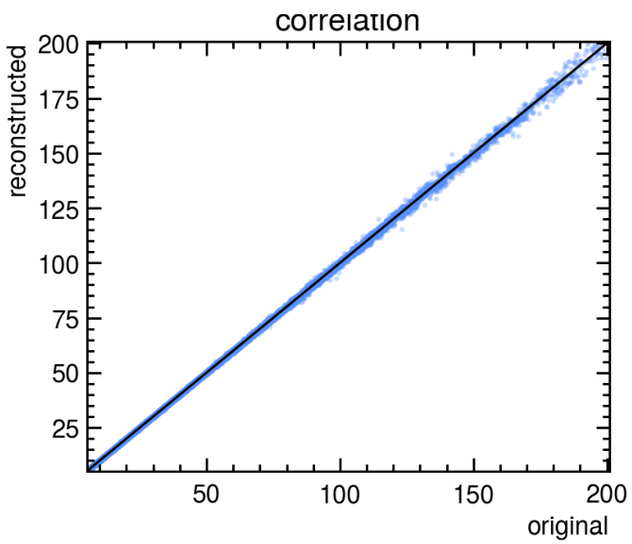




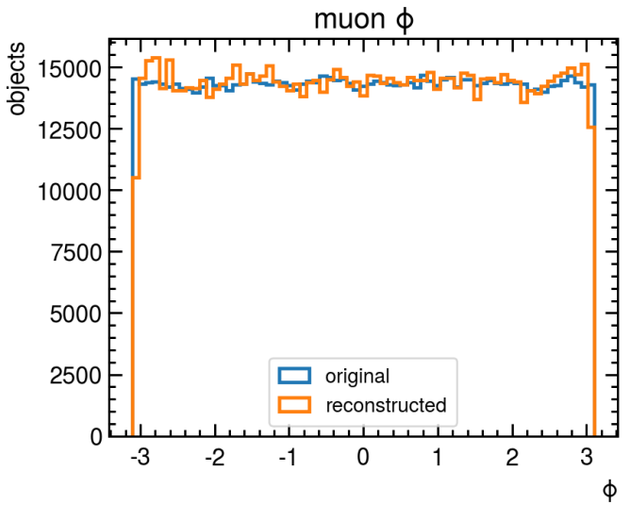


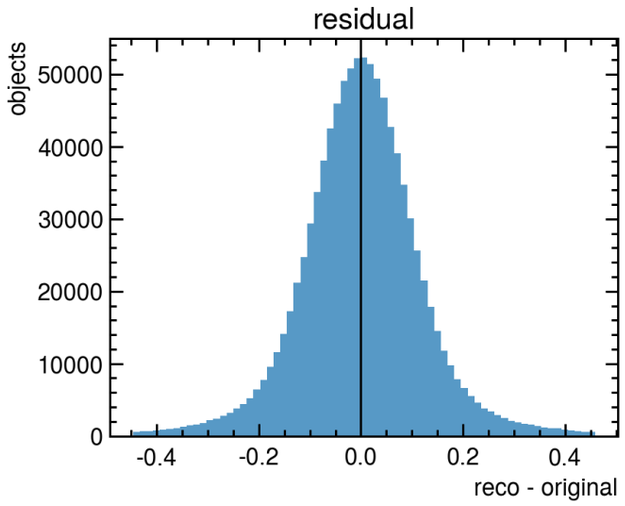


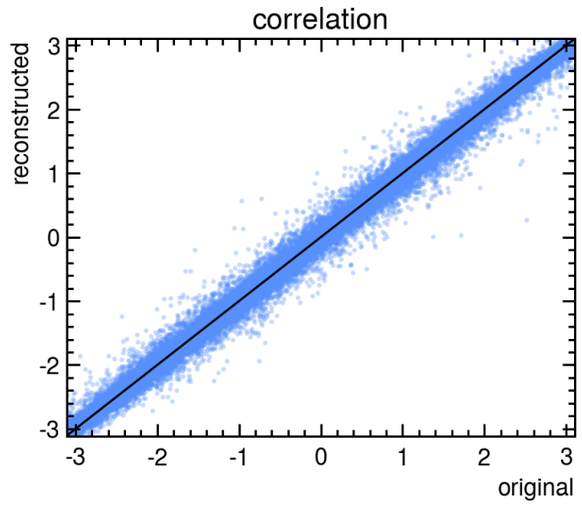


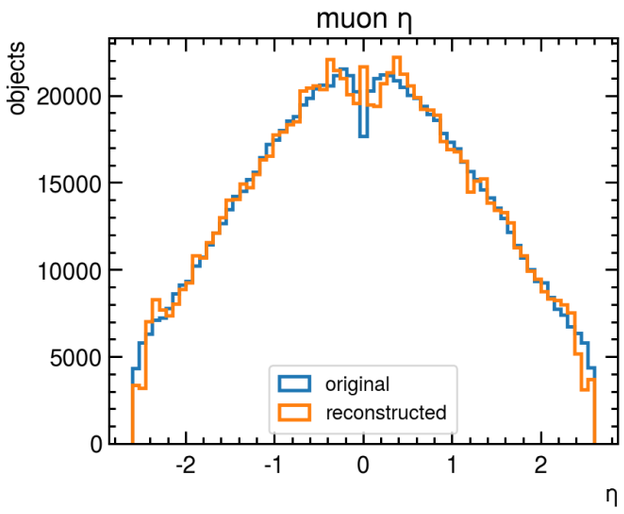


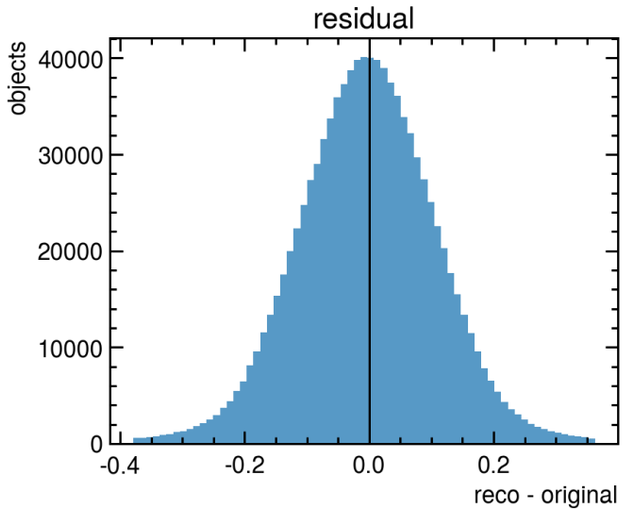
Diagnostics - muons

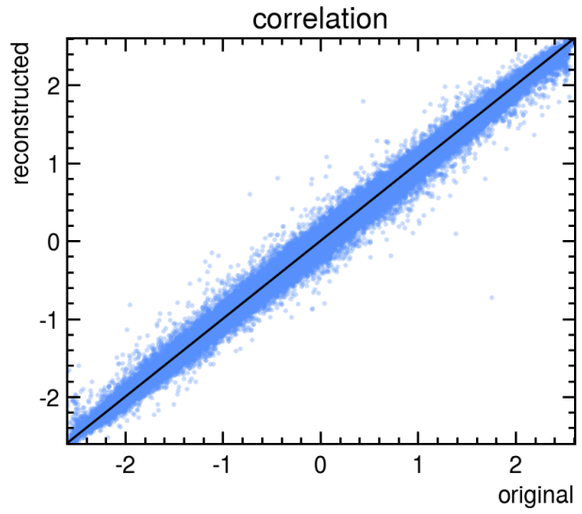


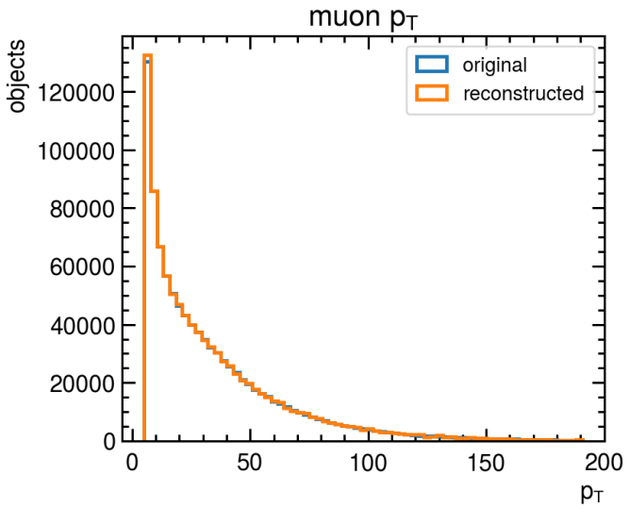


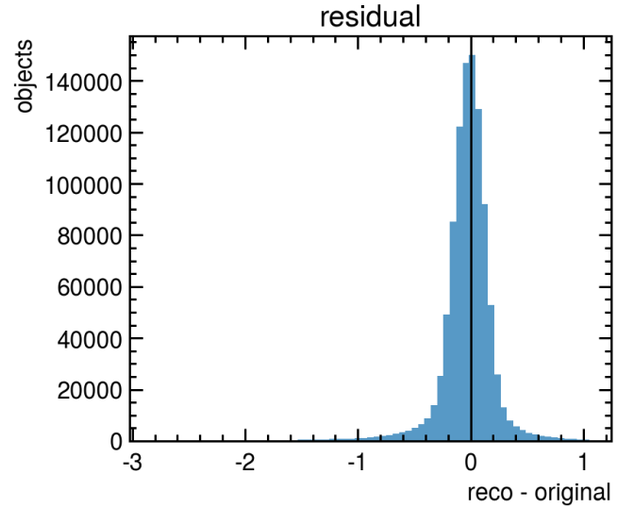


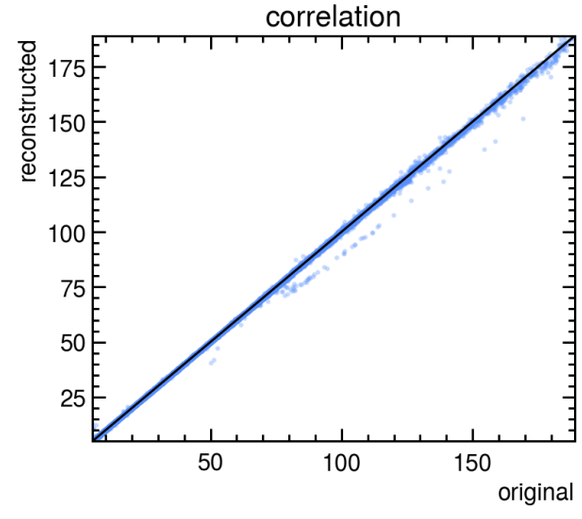




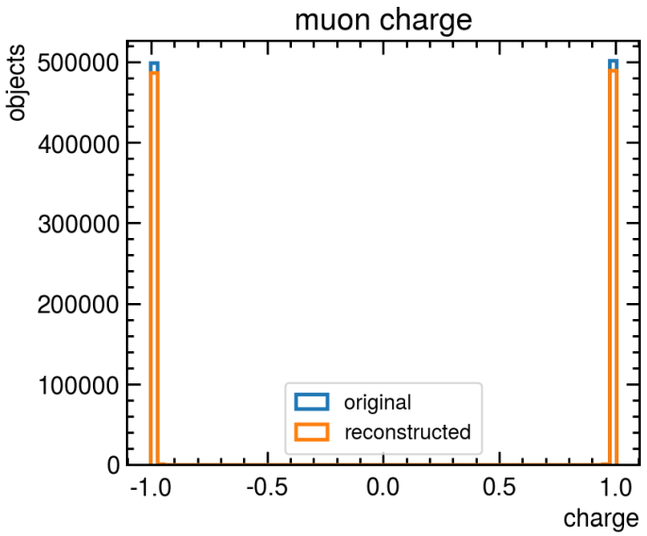


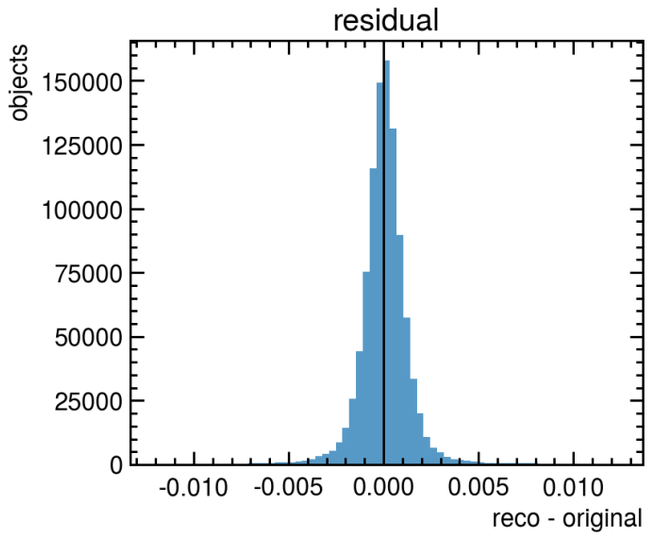


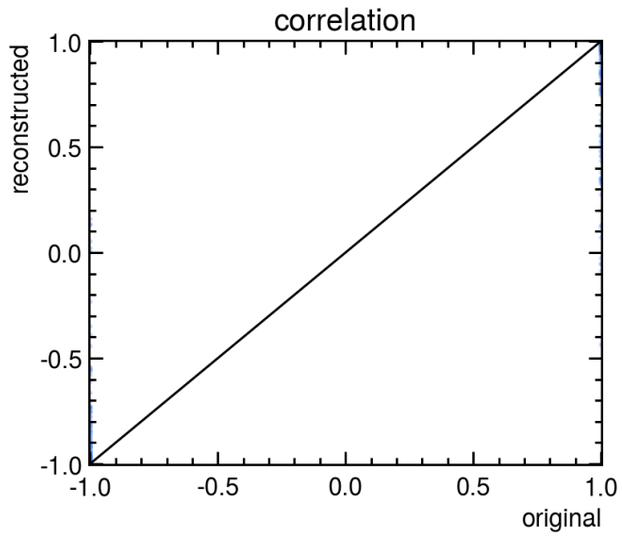


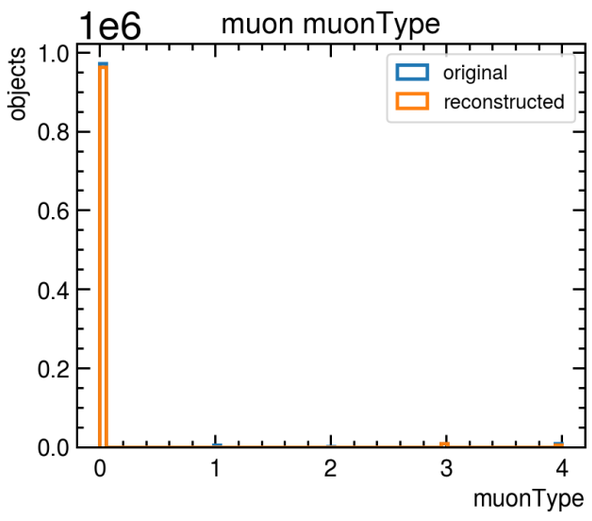


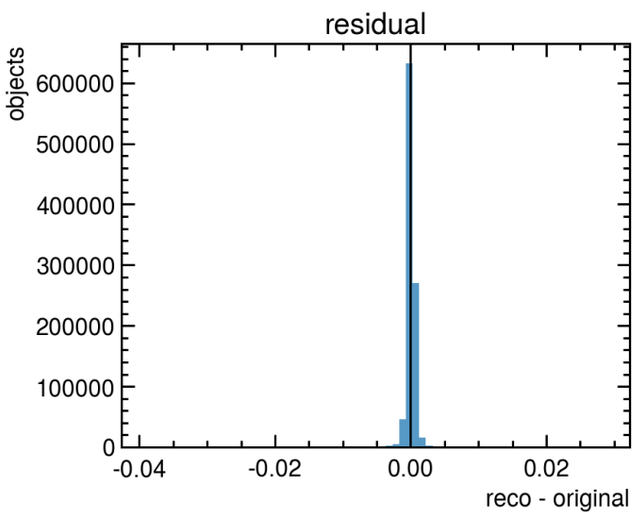
Diagnostics - muons

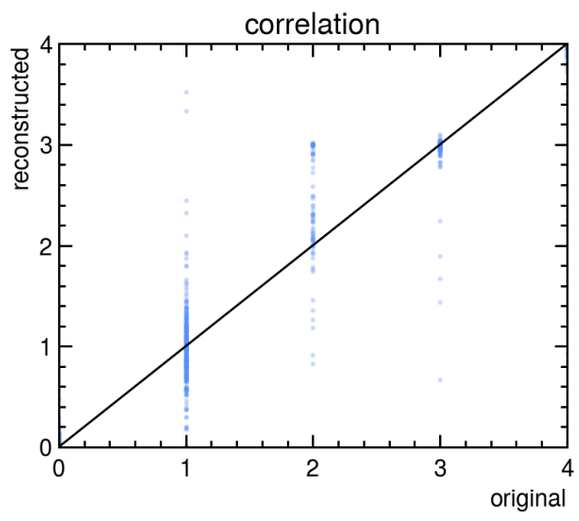


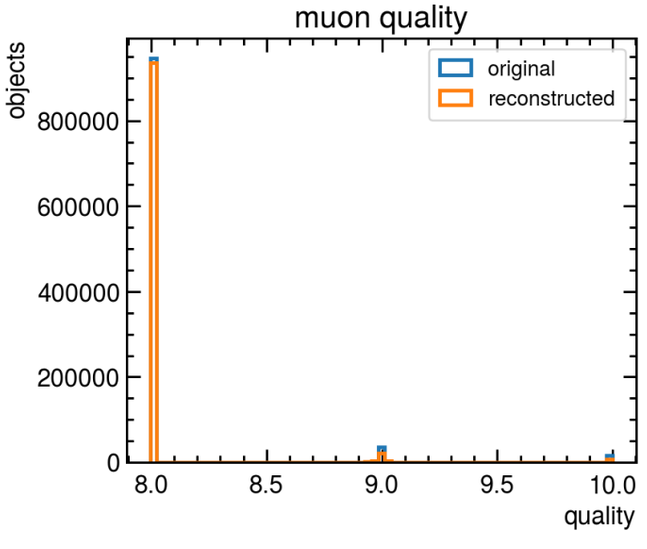


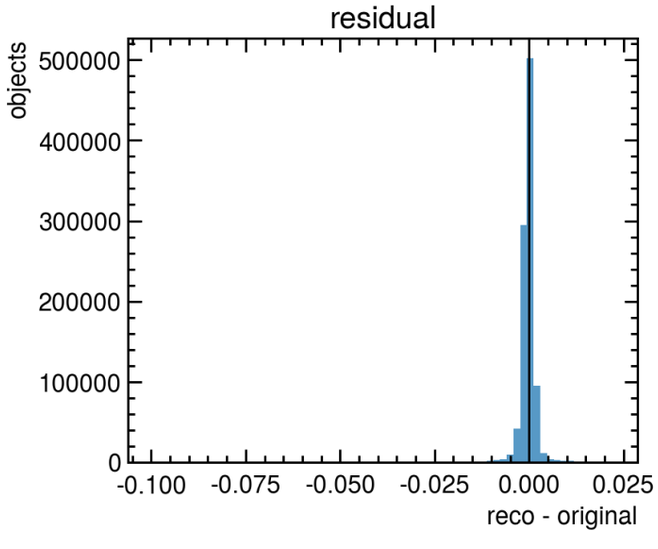


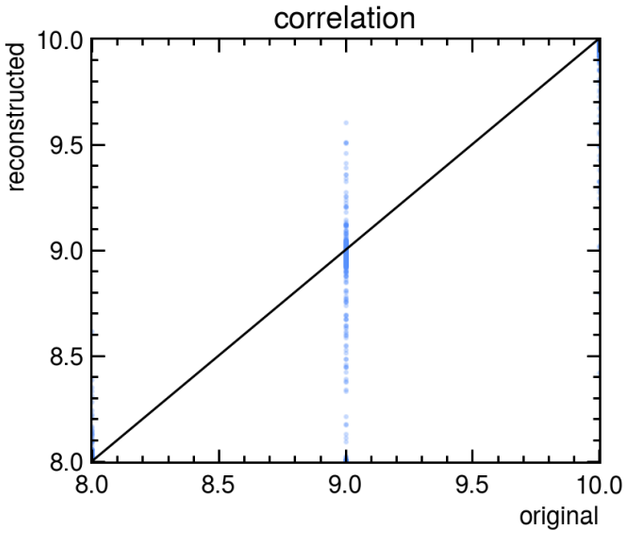




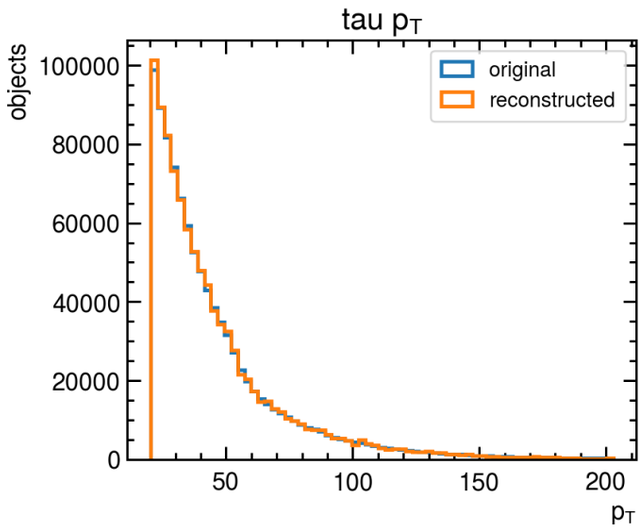


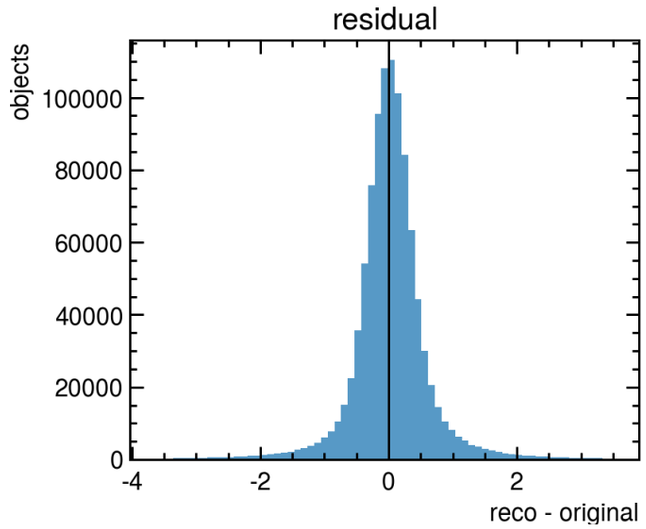


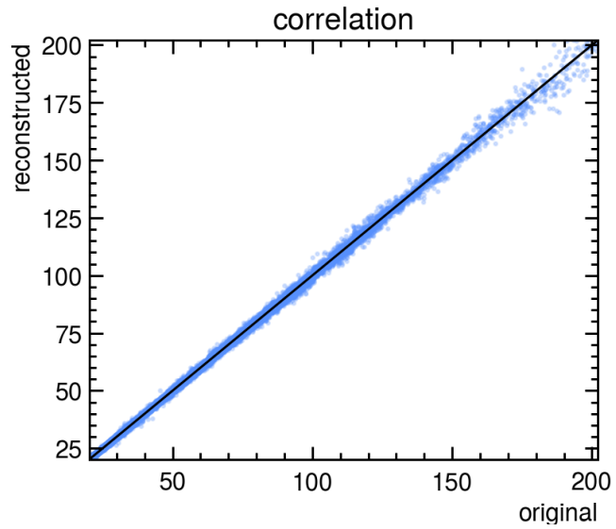


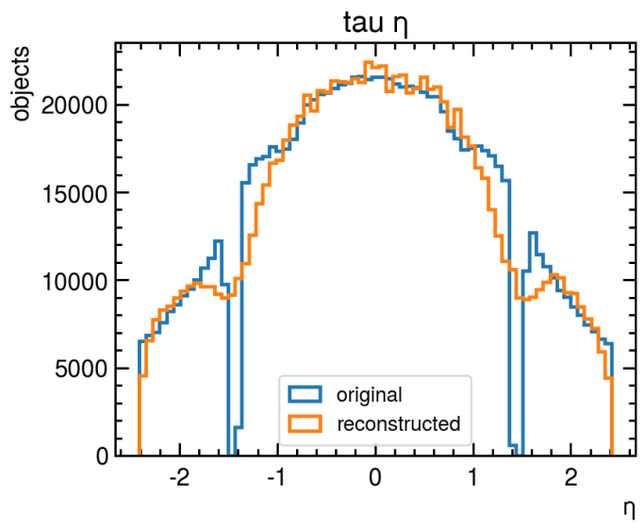


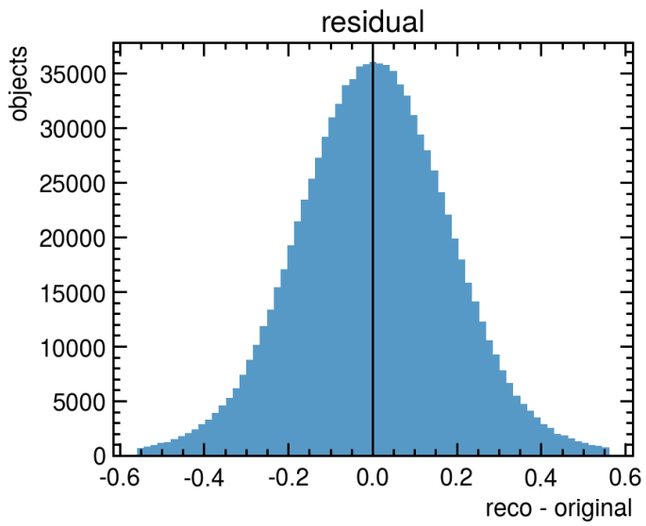
Diagnostics - taus

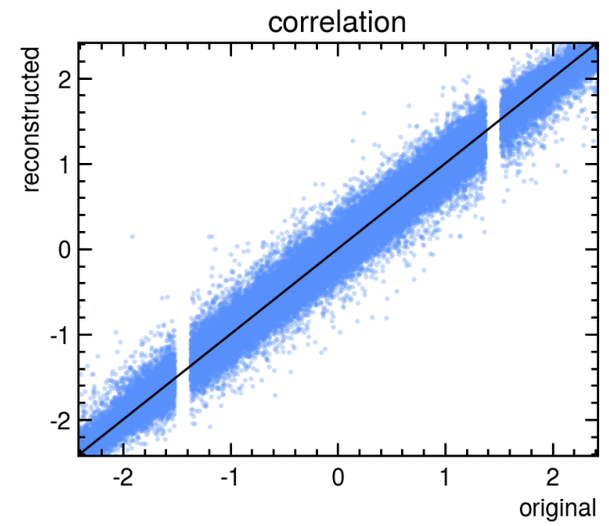


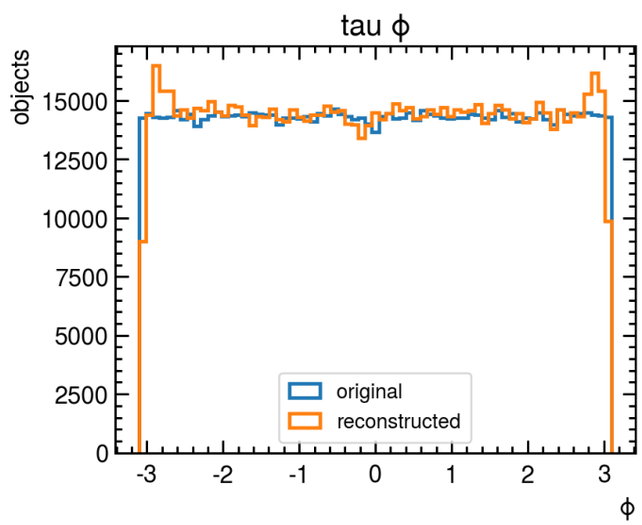


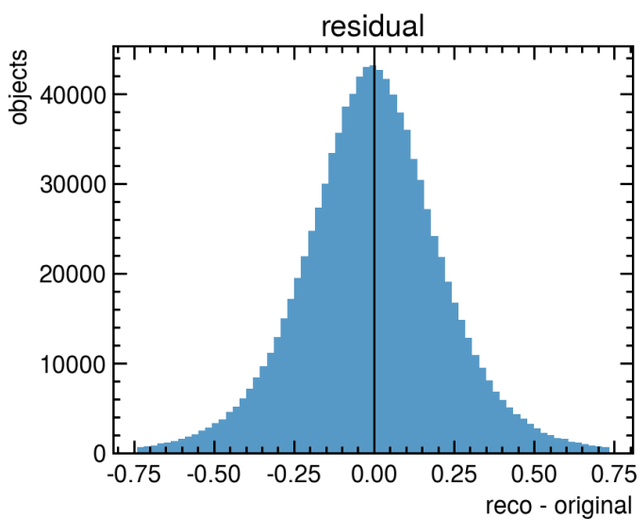


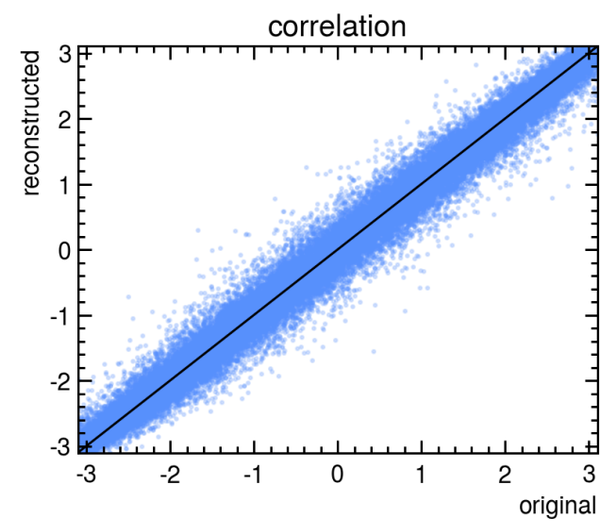




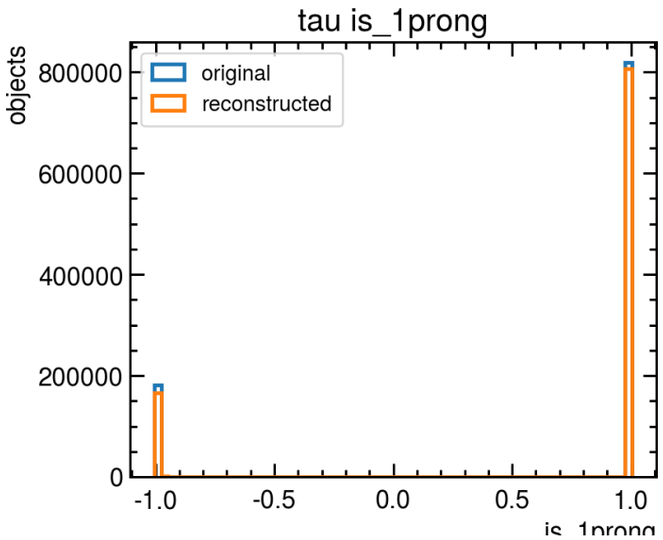


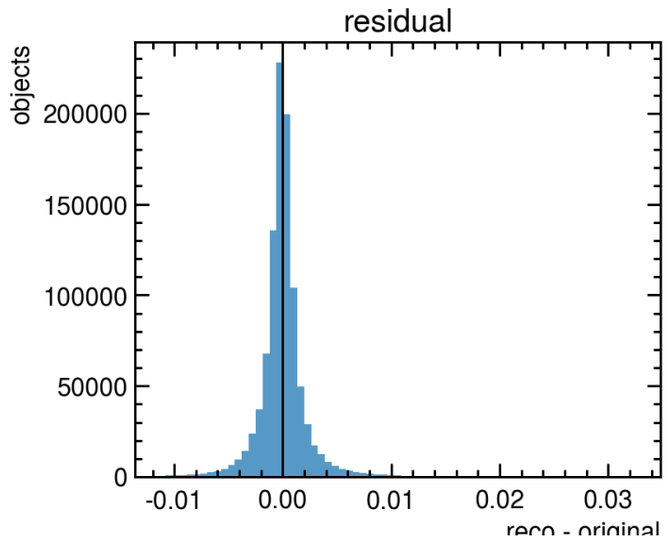


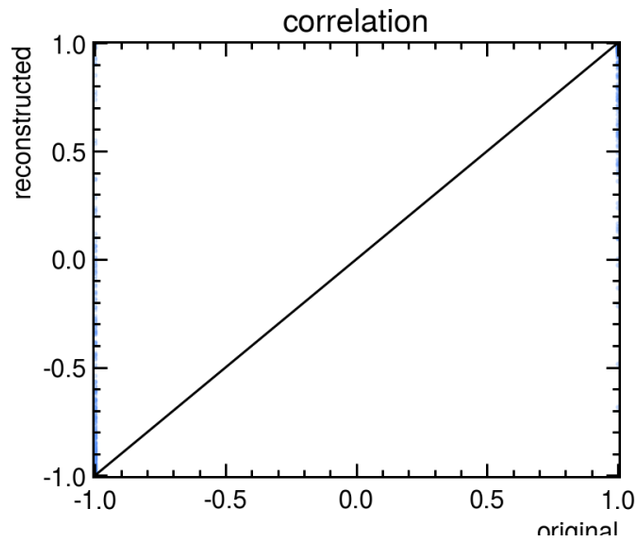


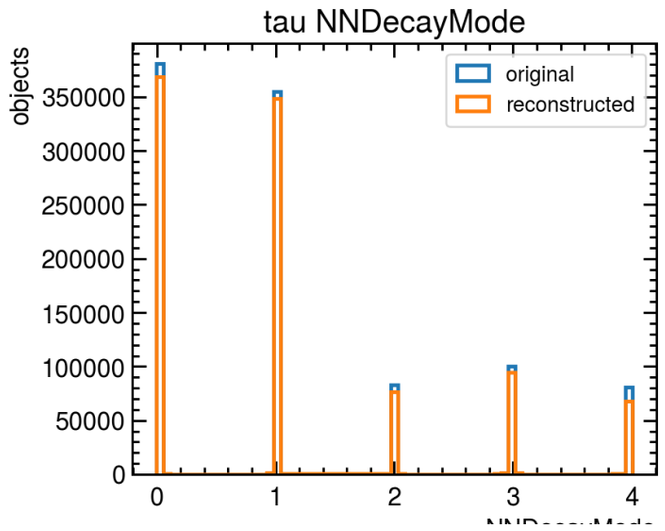


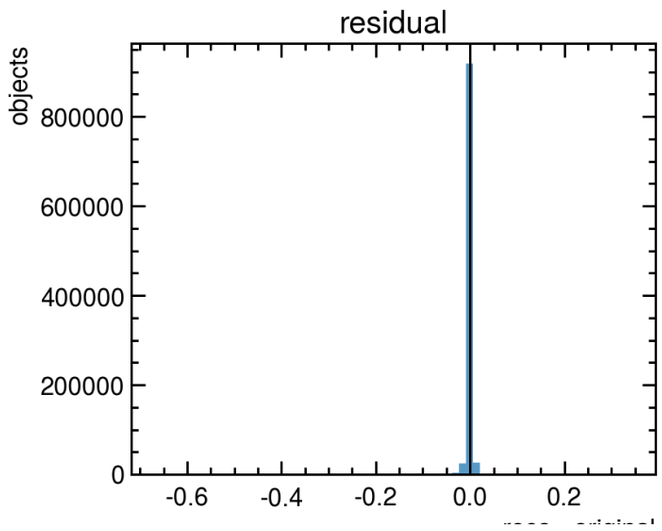
Diagnostics - taus

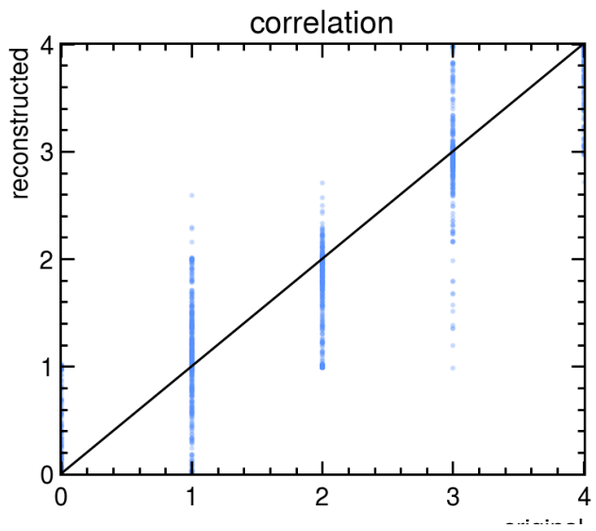


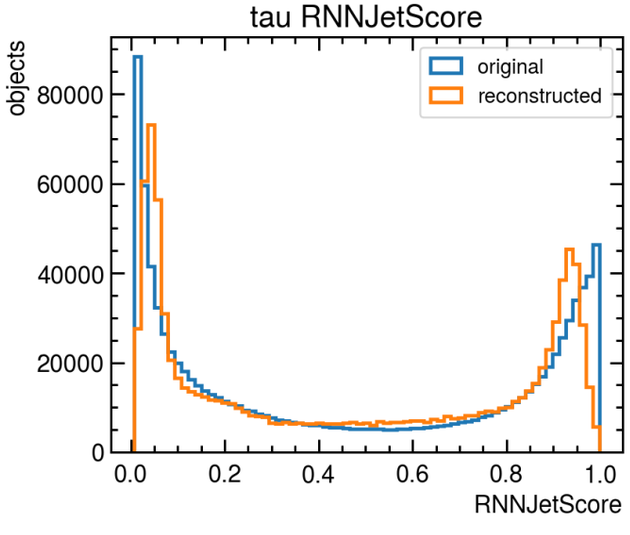


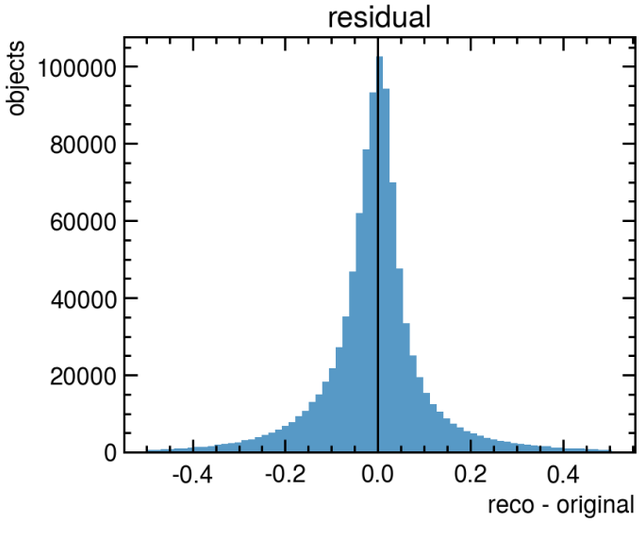


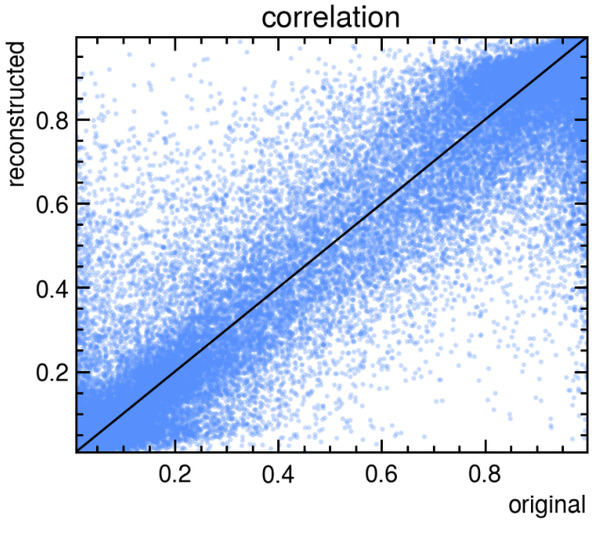




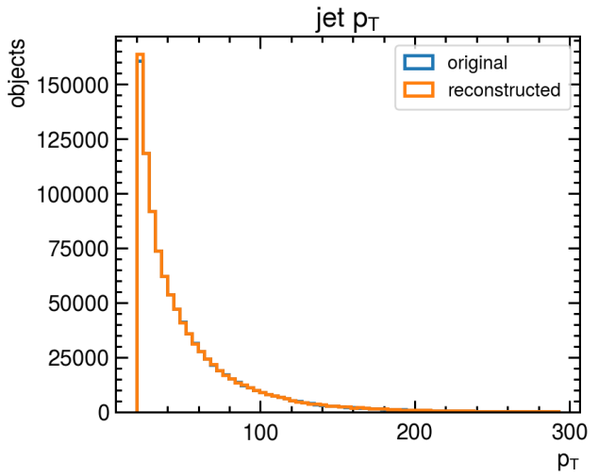


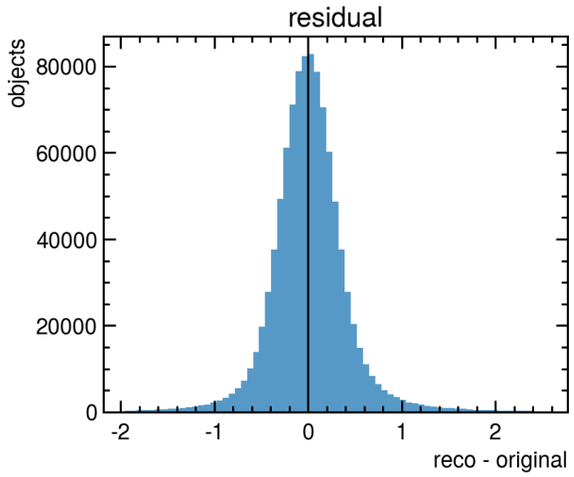


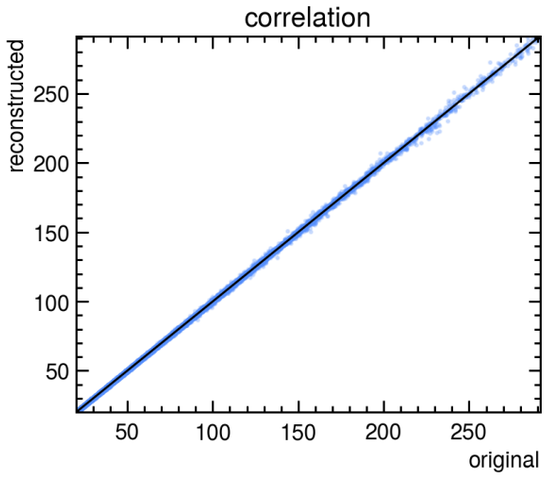


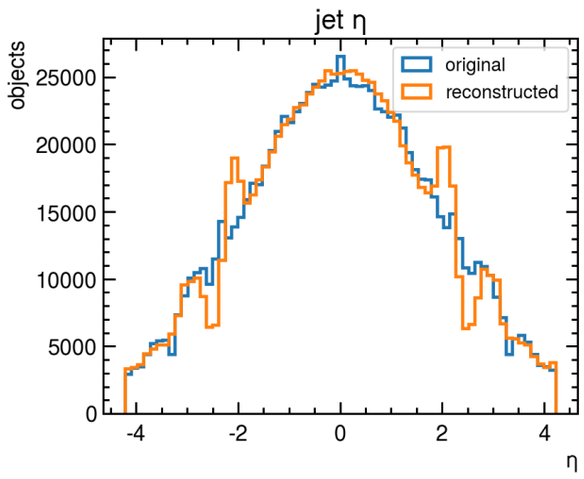


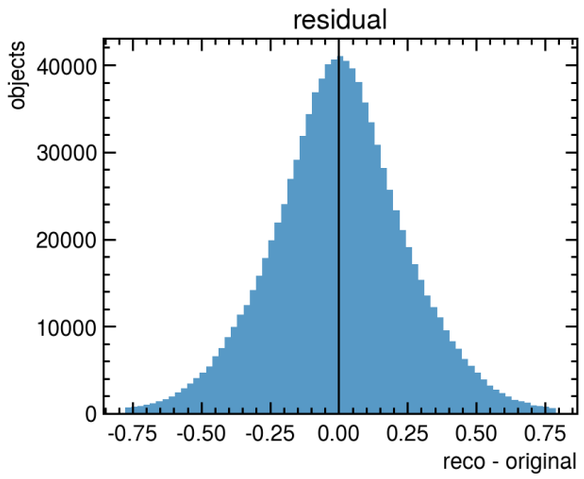
Diagnostics - jets

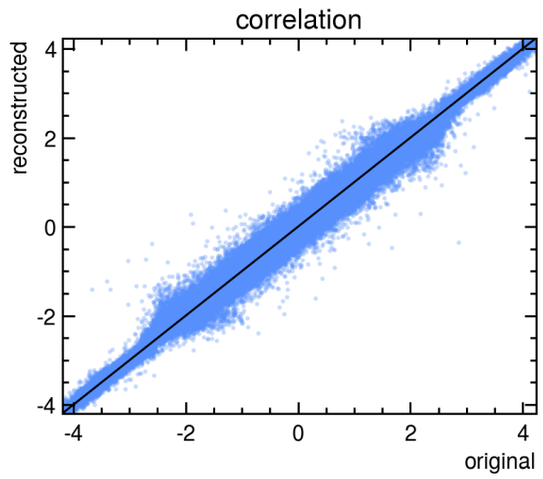


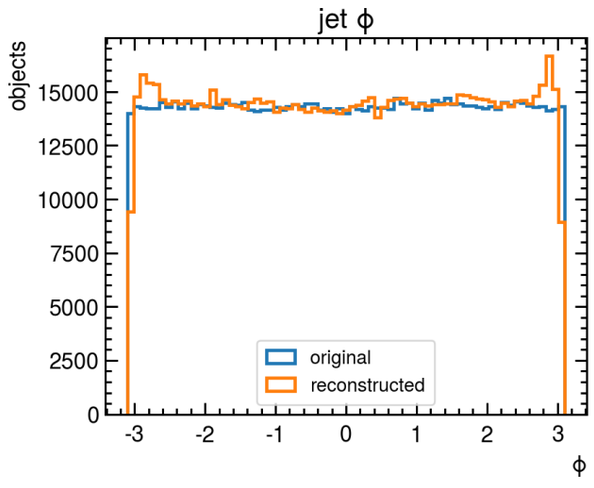


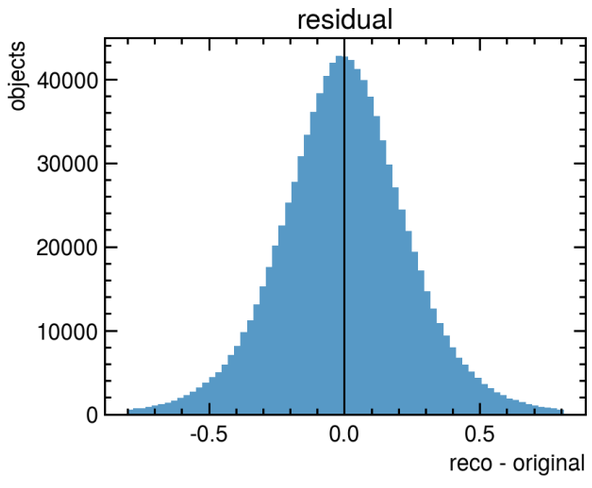


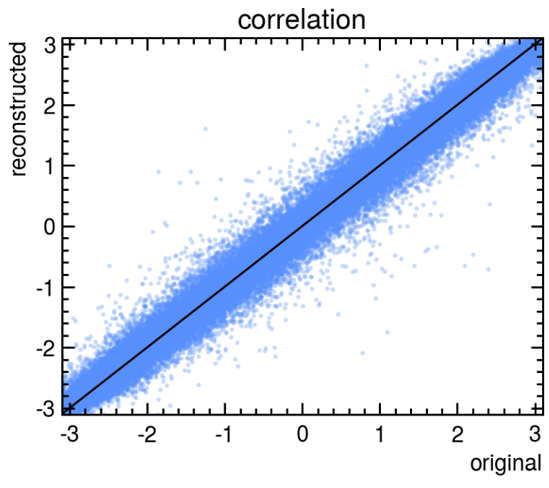




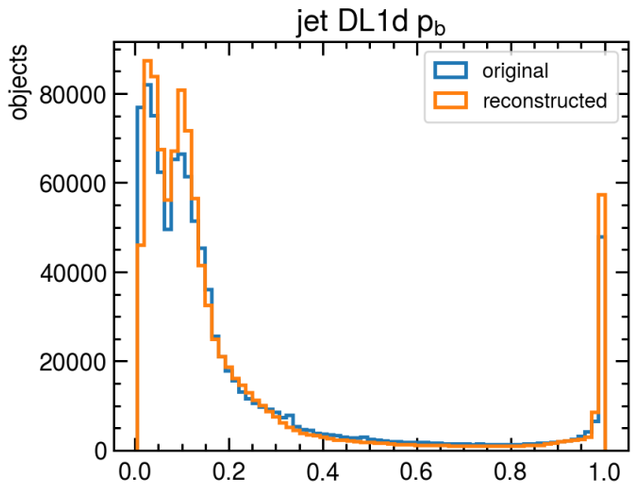


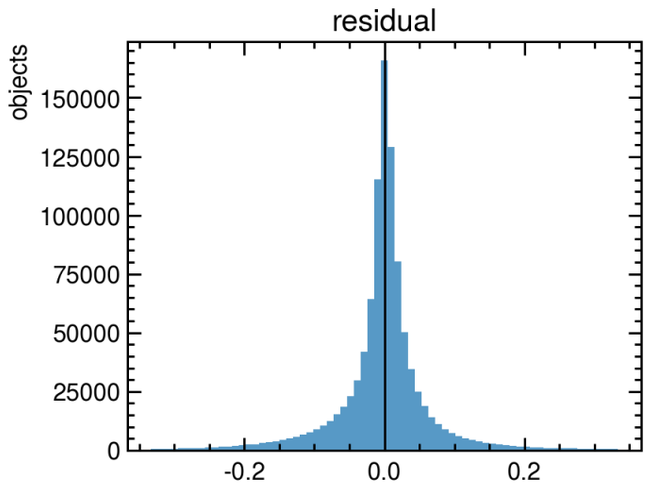


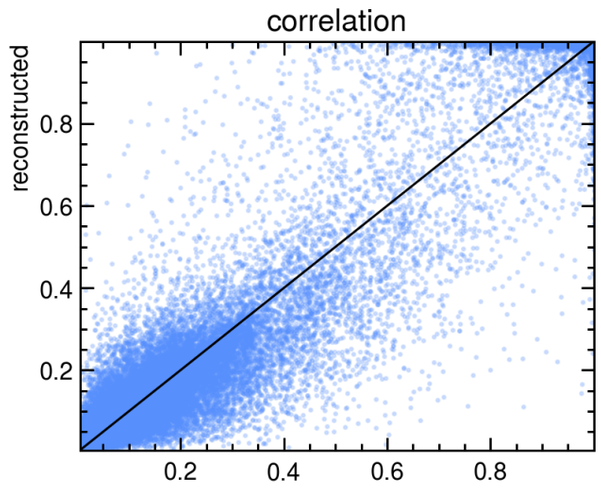


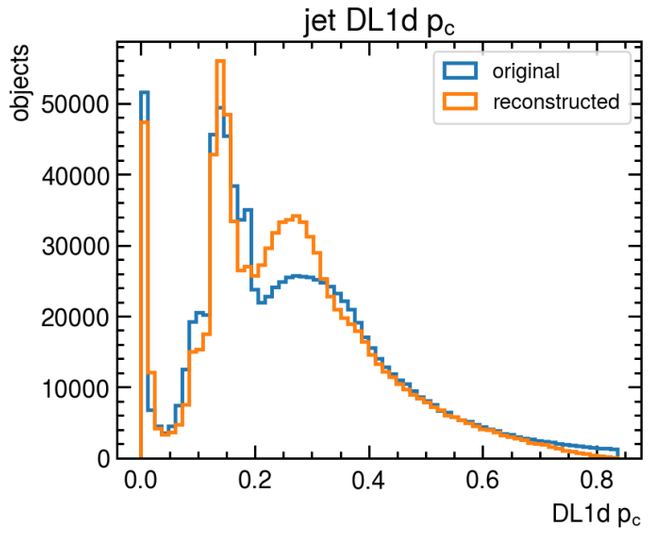


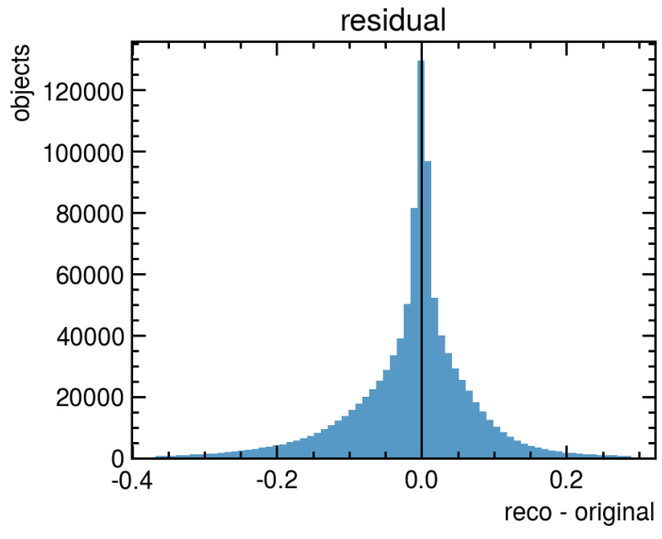
Diagnostics - jets

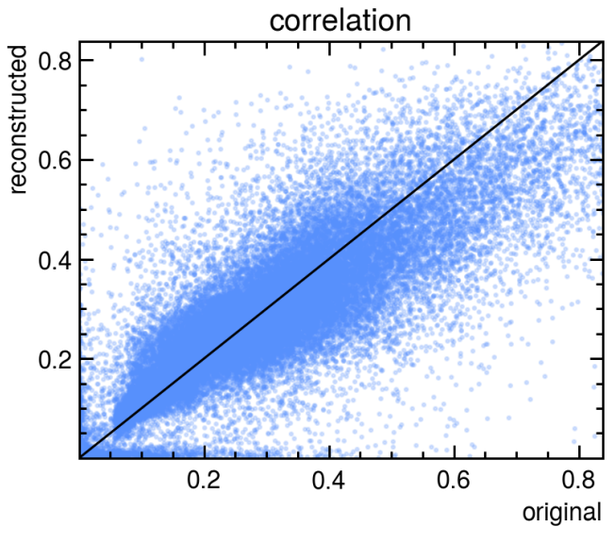


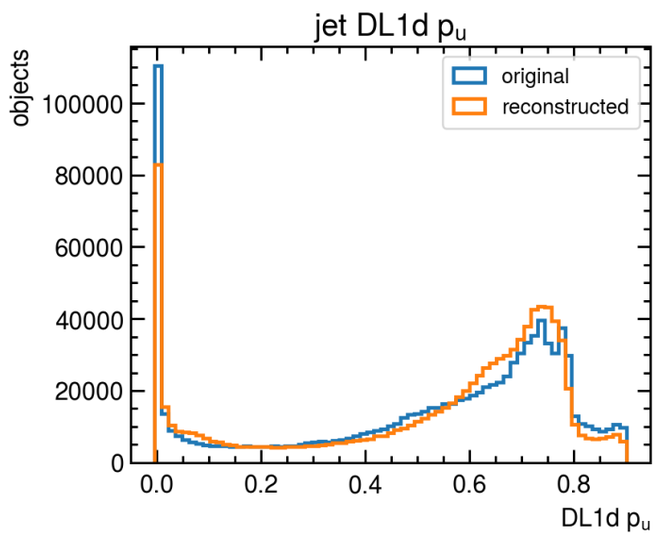


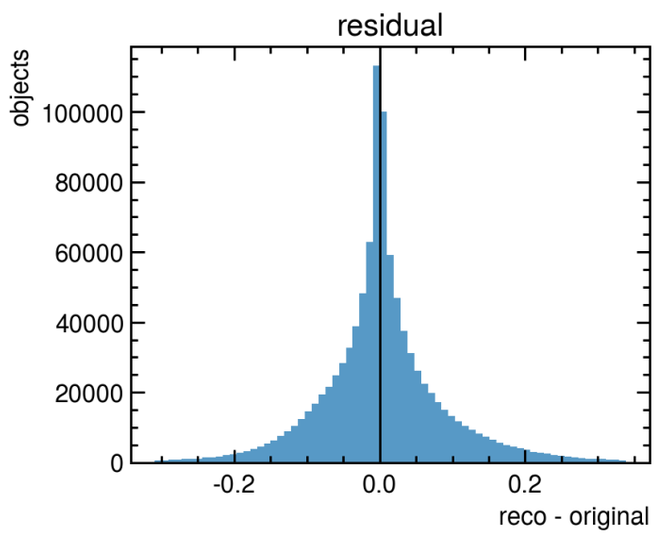


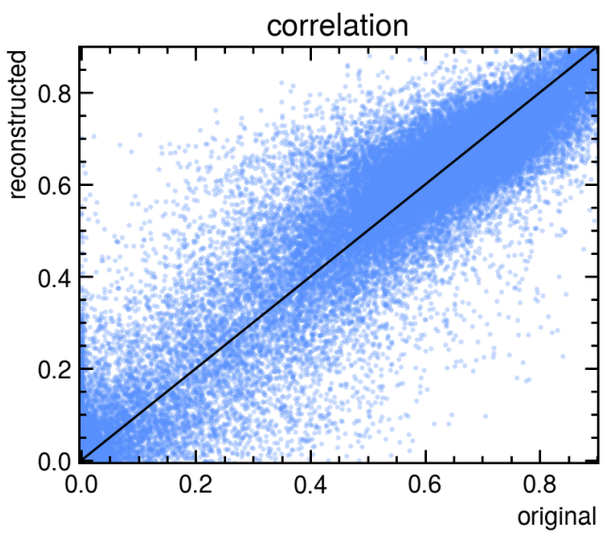




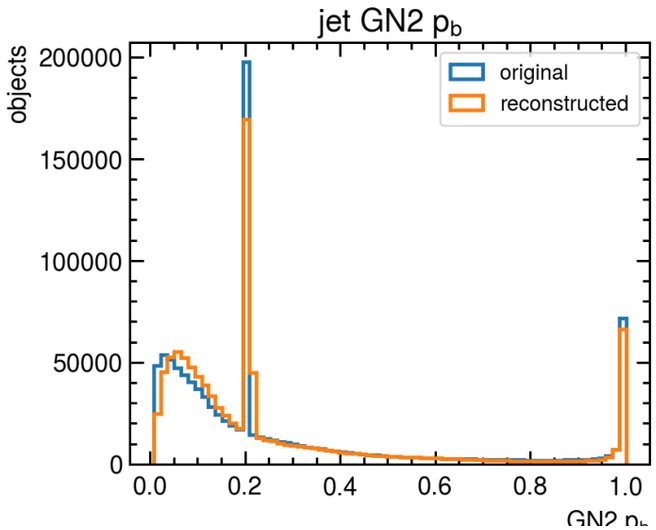


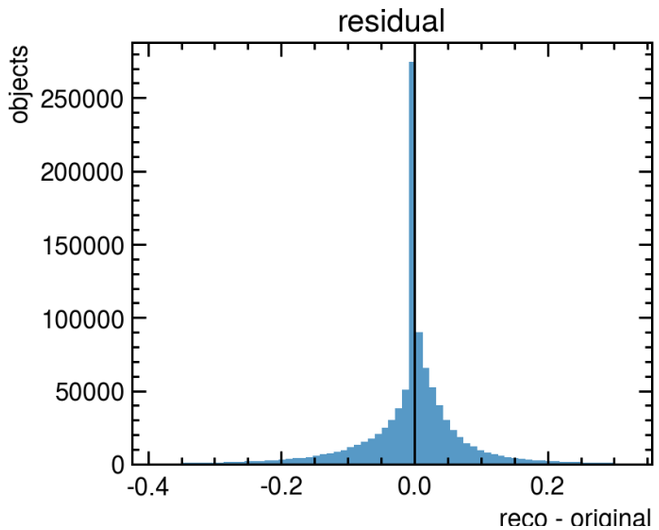


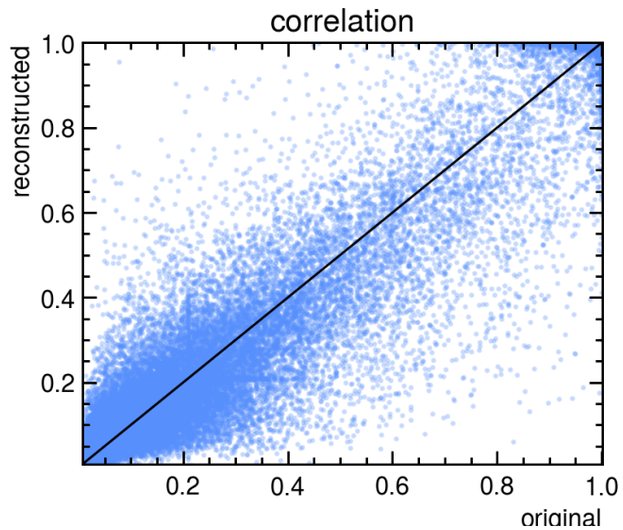


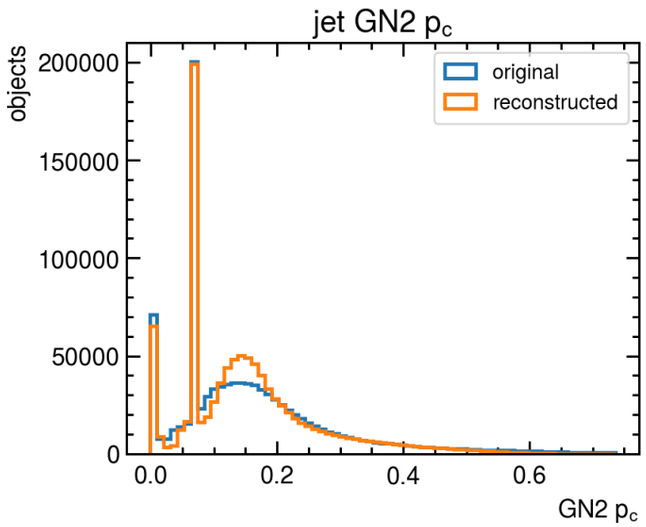


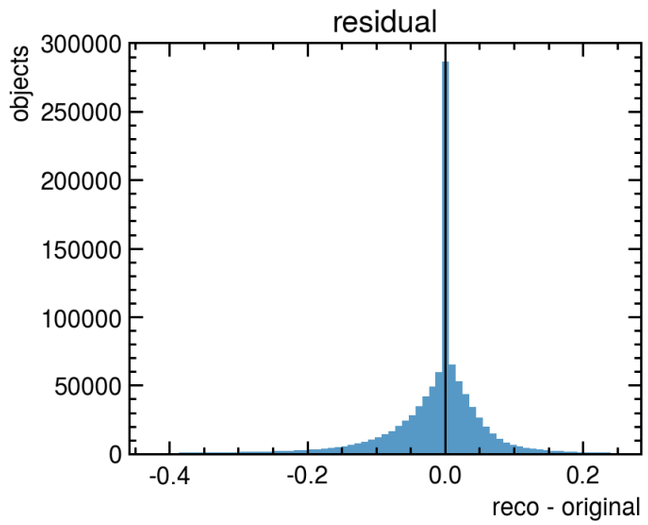
Diagnostics - jets

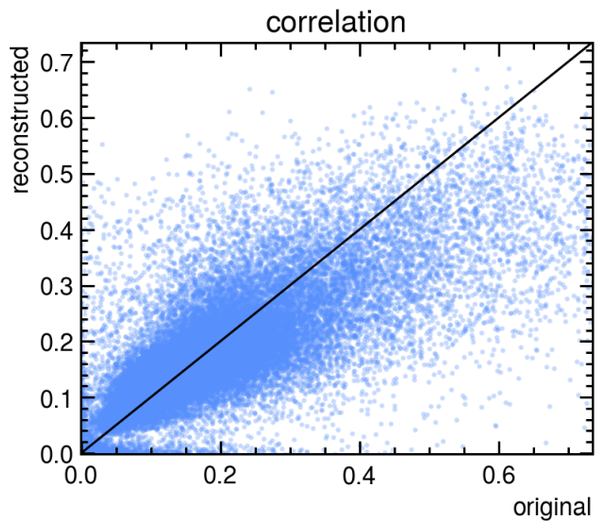


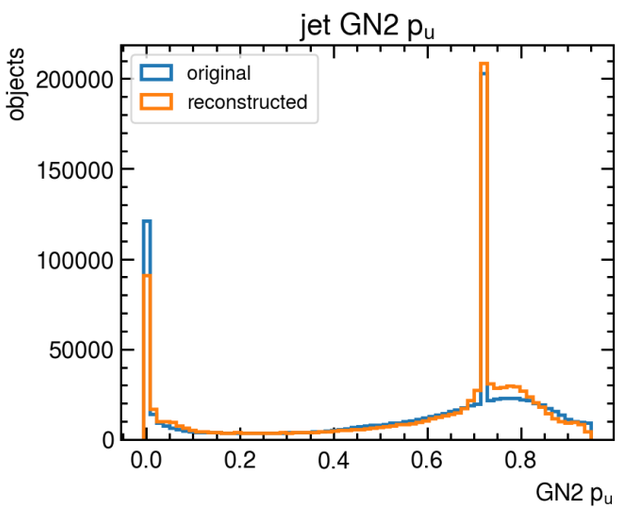


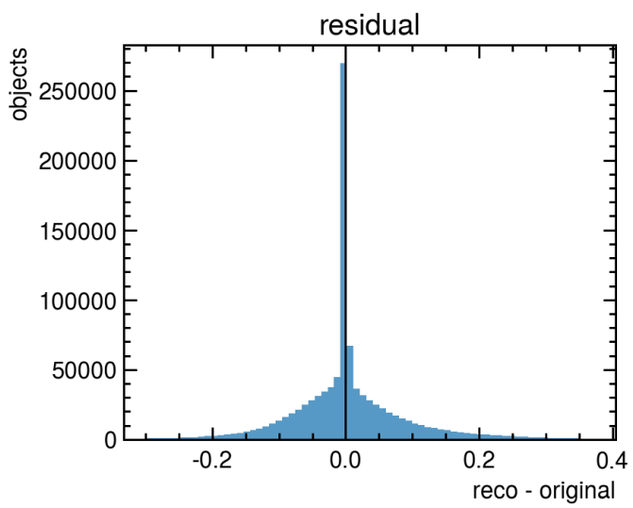


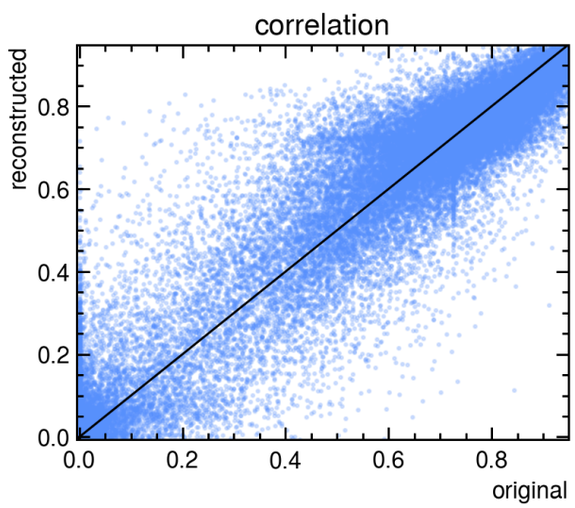




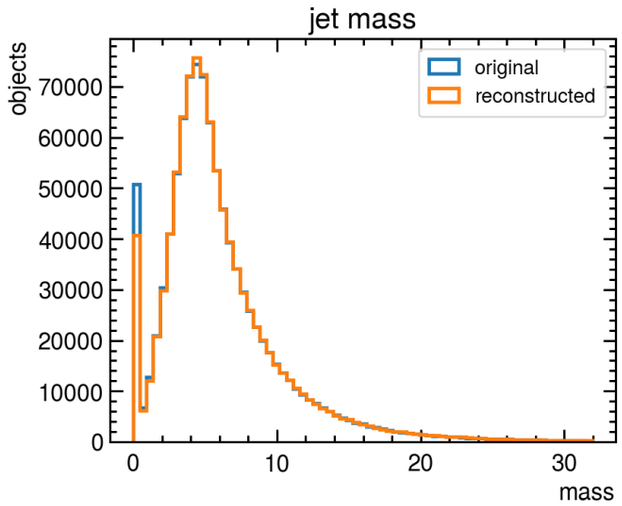


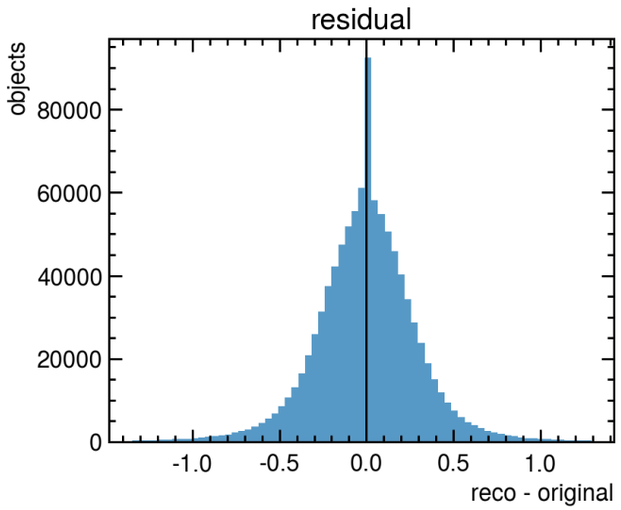


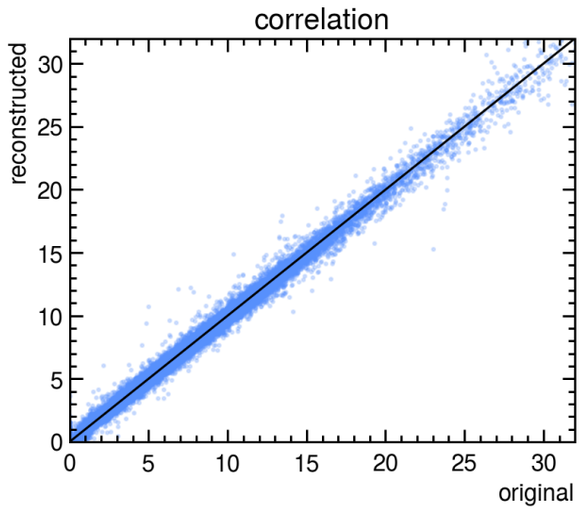


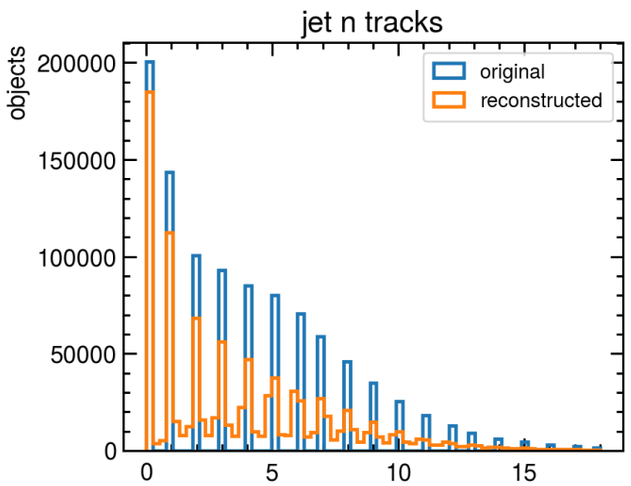


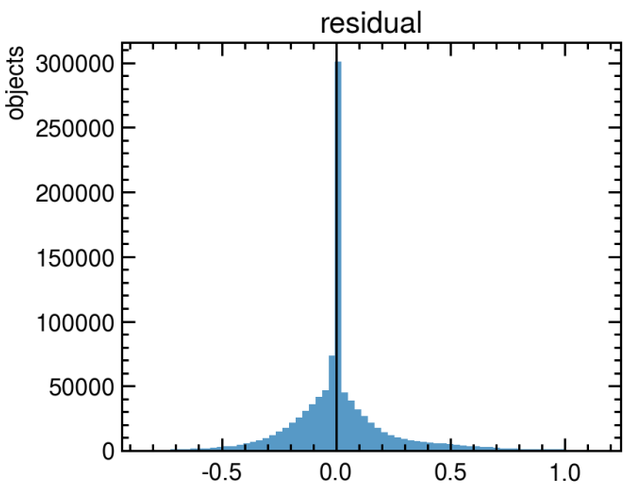
Diagnostics - jets

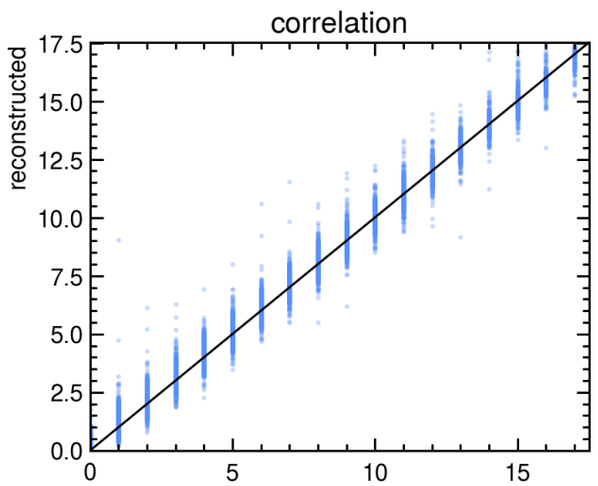


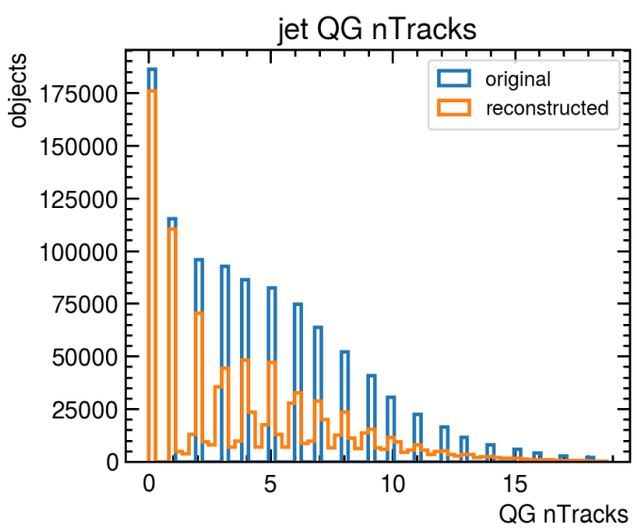


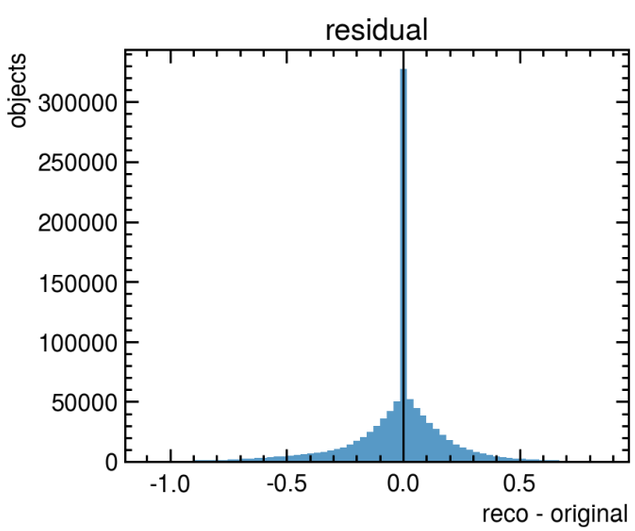


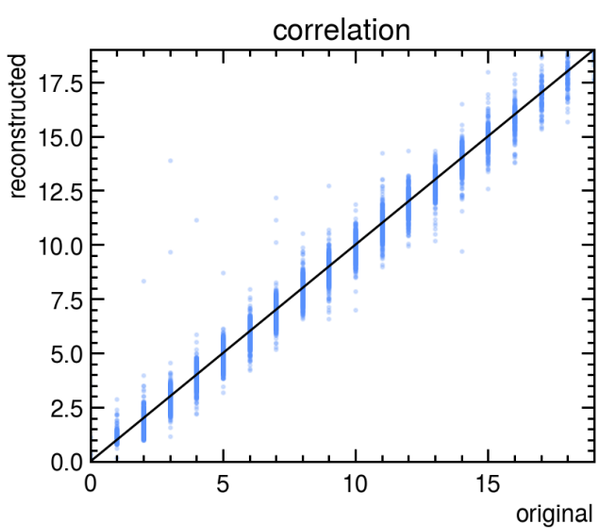




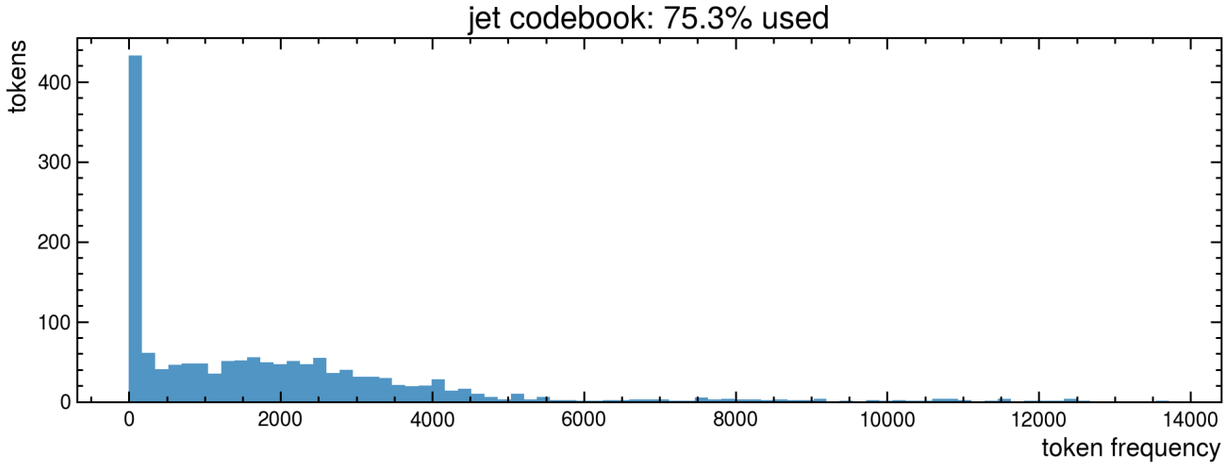




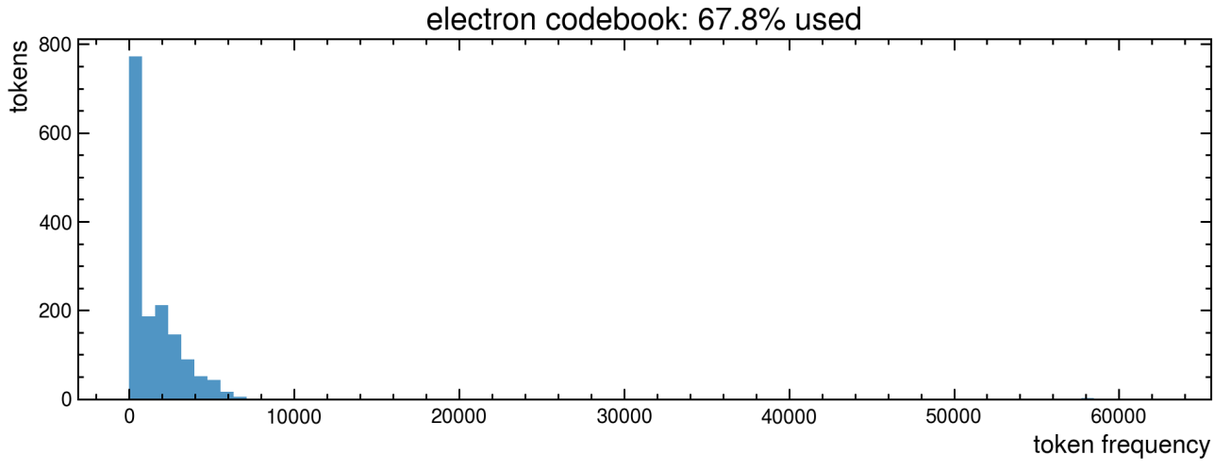




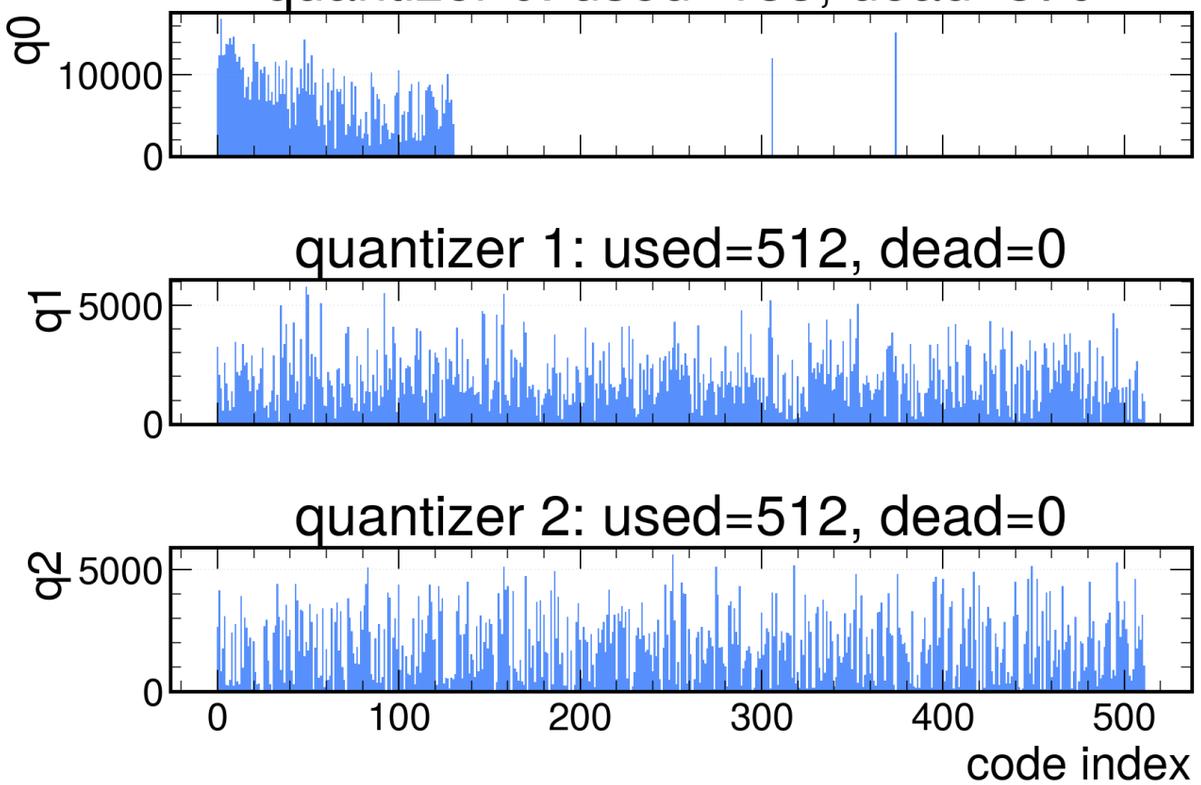
Diagnostics

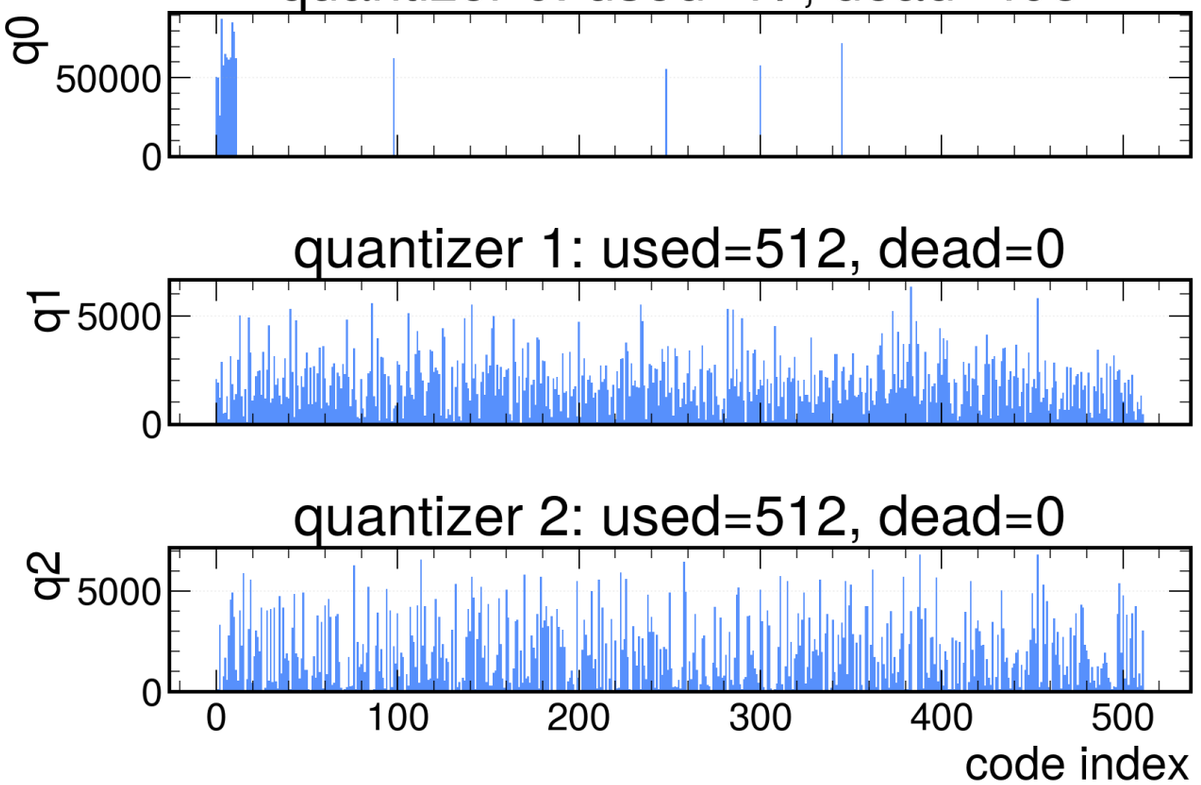


Codebook usage frequency

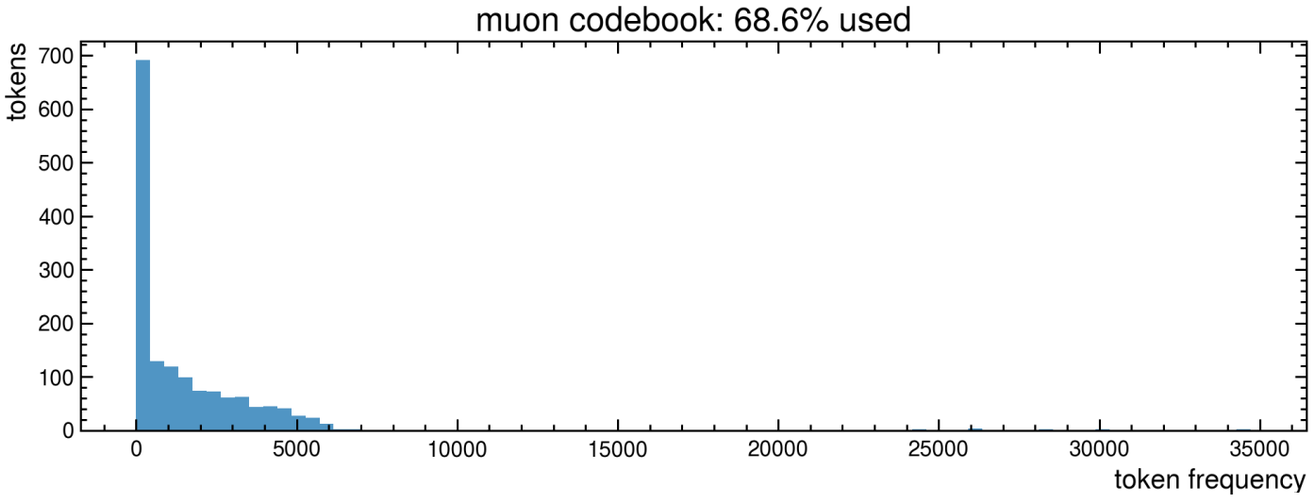


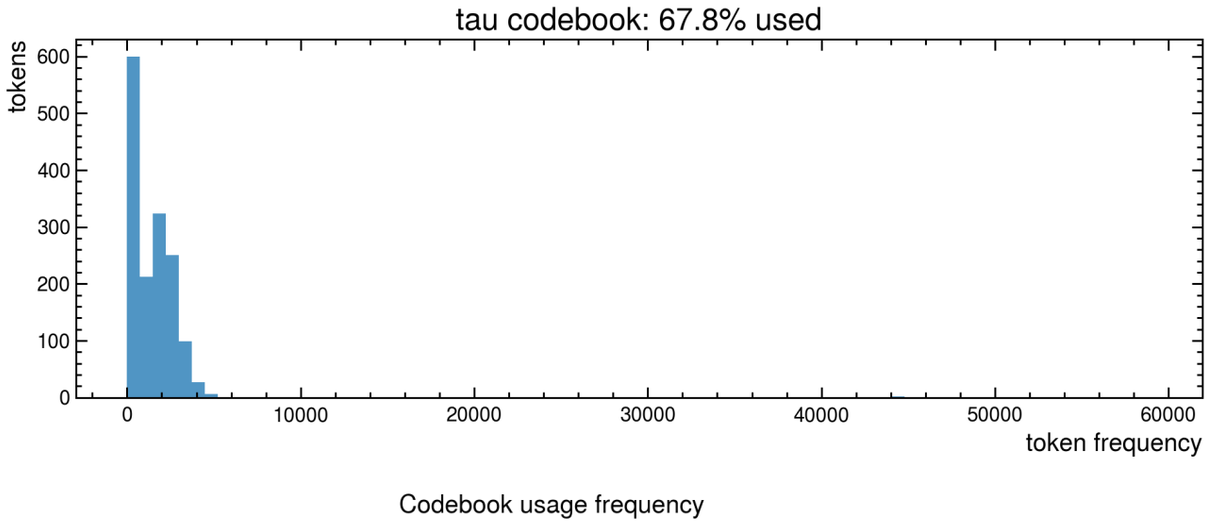
Codebook usage frequency



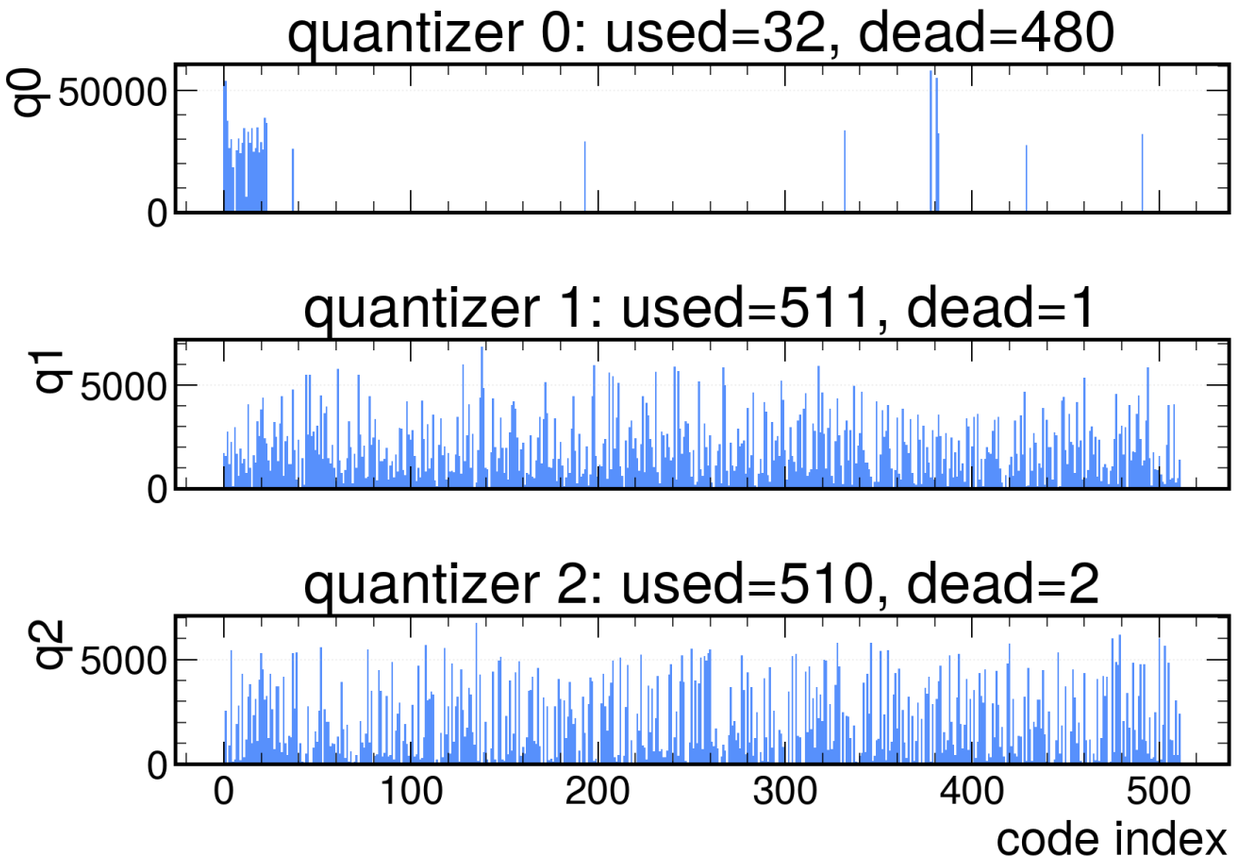


Diagnostics





Codebook usage frequency



Discussion

* First full-feature object tokenizers are now trained and diagnostics are running for all object types.
* Basic kinematics are reconstructed well, but some higher-level/discriminant features are not yet well preserved.
  - The VQ-VAE appears to smooth bounded score-like features such as DL1d/GN2 probabilities.
* Next steps: Add data, separate continuous, discrete, and bounded score-like variables more carefully; test feature-wise preprocessing/normalisation; compare **single-codebook vs residual-codebook settings**; and evaluate whether these reconstruction differences matter for downstream model.