Combined Tokenization Schemes

Nicholas Luongo

TREASURE Weekly Meeting

July 1, 2026

Argonne NATIONAL LABORATORY

ATLAS EXPERIMENT

Tokenization Schemes

Investigate per-object type and combined schemes for event-level tokenization

Per-object

- Each type of object (jets, electrons, etc) gets its own tokenizer
- Pros – handles heterogenous object features, allows for specialization

Combined

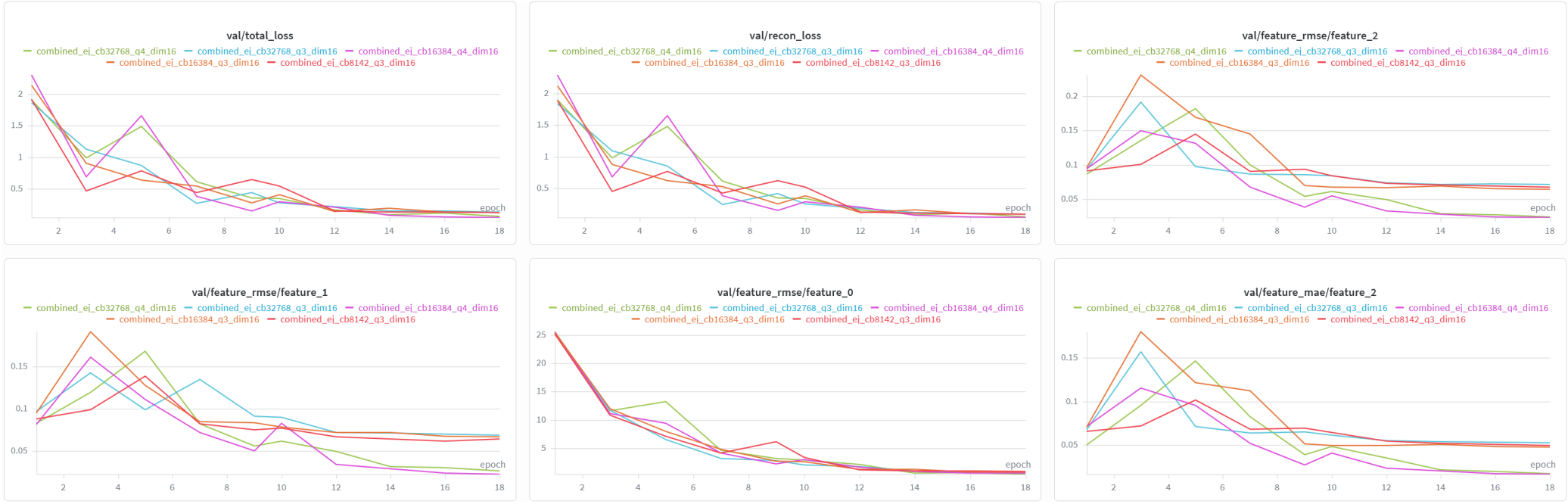
- Single tokenizer used for all objects in the event
- Pros – simpler training procedure, more data in training set, no duplicated effort

Include only (p_T, η, ϕ) features consider electrons and jets for initial tests

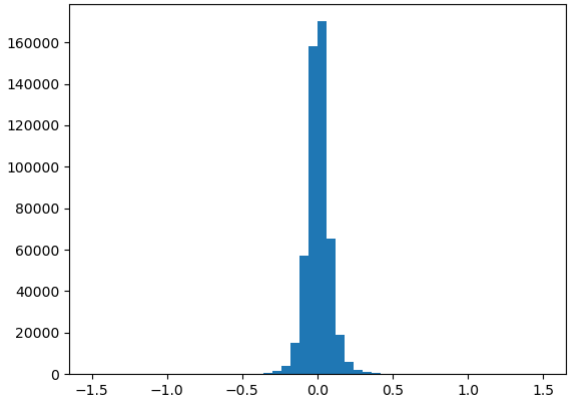
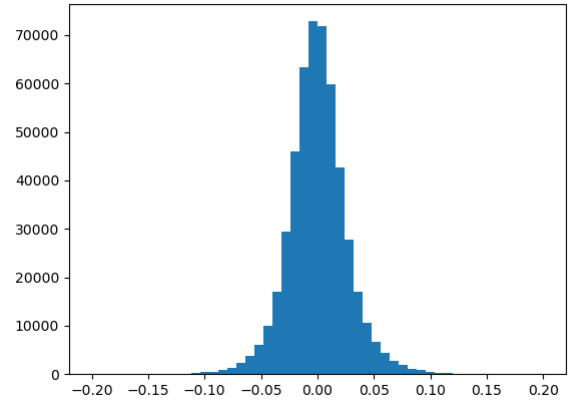
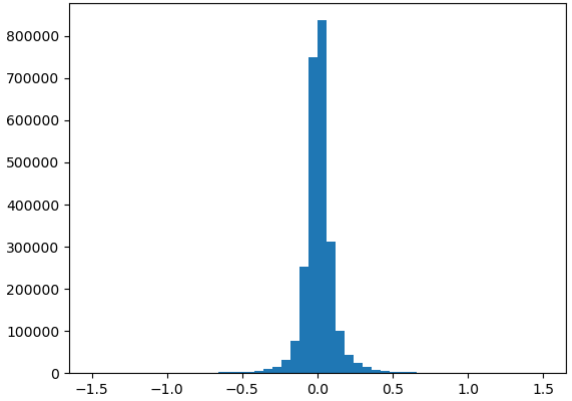
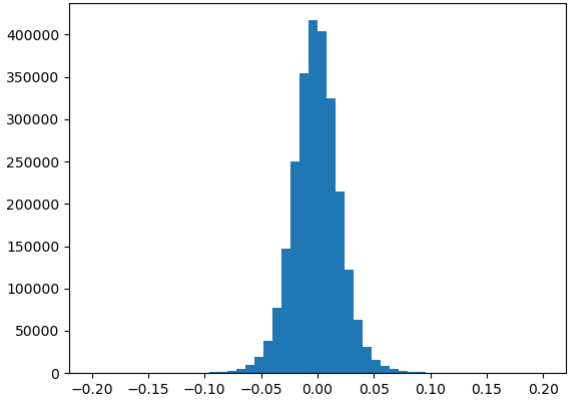
Compare reconstruction accuracy using residual VQ-VAE for both schemes

Combined Codebook Scan

Tested same tokenization hyperparameters as Nazlim for combined tokenizer, found that **cb16384_q4_dim16** also performed optimally, used for following studies



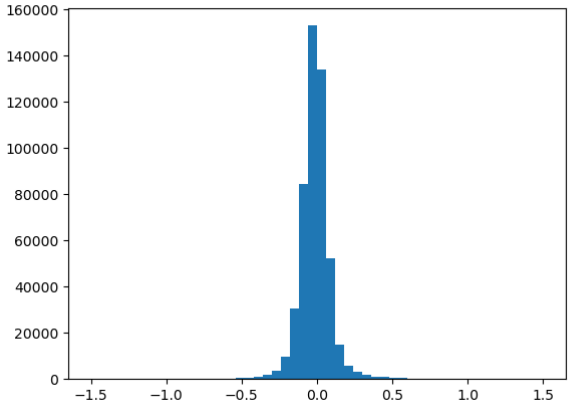
e+*j* tokenizer

	p_T Residual		η Residual	
Electrons		Mean: 0.00519 StDev: 0.0994		Mean: 0.000255 StDev: 0.0261
Jets		Mean: 0.0157 StDev: 1.513		Mean: -0.00127 StDev: 0.0216

e-only tokenizer

Electrons

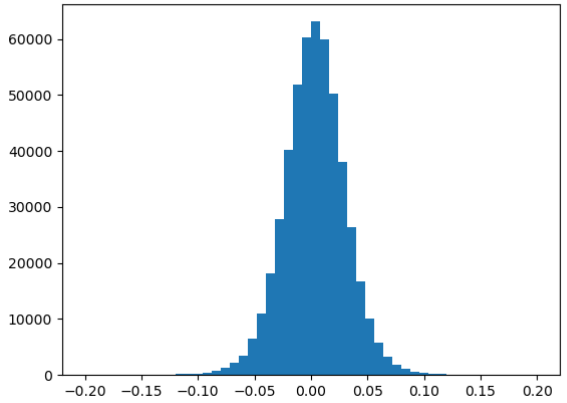
p_T Residual



Mean: -0.00691

StDev: 1.644

η Residual



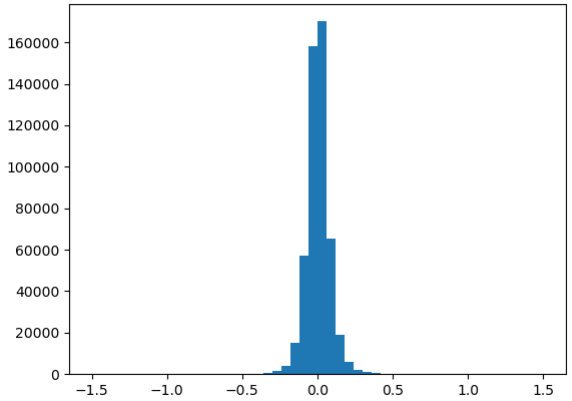
Mean: 0.00323

StDev: 0.284

e Tokenizer Comparison

e+j

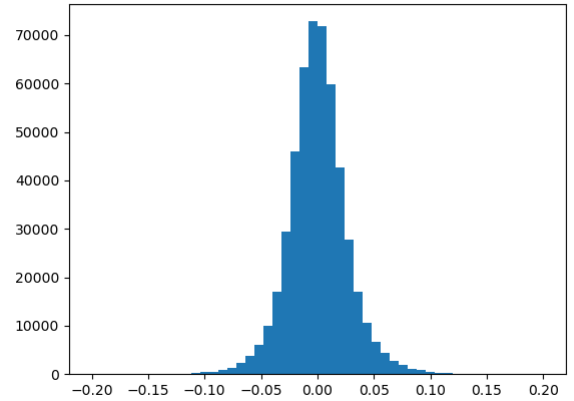
p_T Residual



Mean: 0.00519

StDev: 0.0994

η Residual

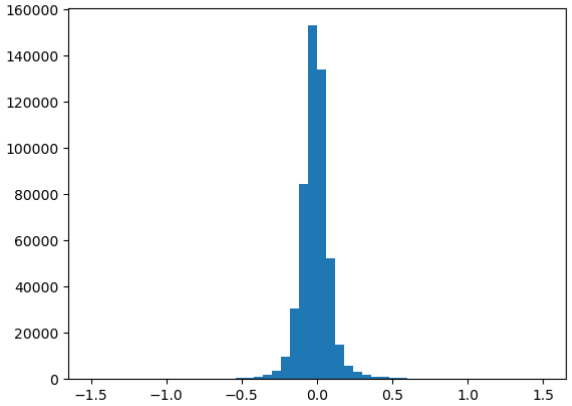


Mean: 0.000255

StDev: 0.0261

e-only

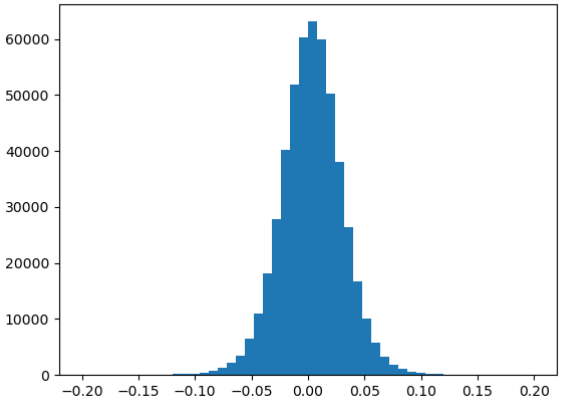
p_T Residual



Mean: -0.00691

StDev: 1.644

η Residual

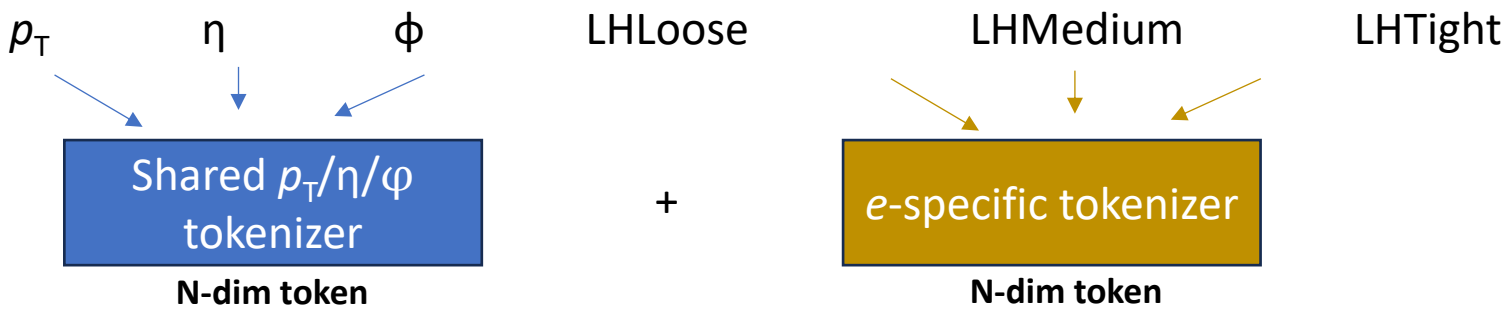


Mean: 0.00323

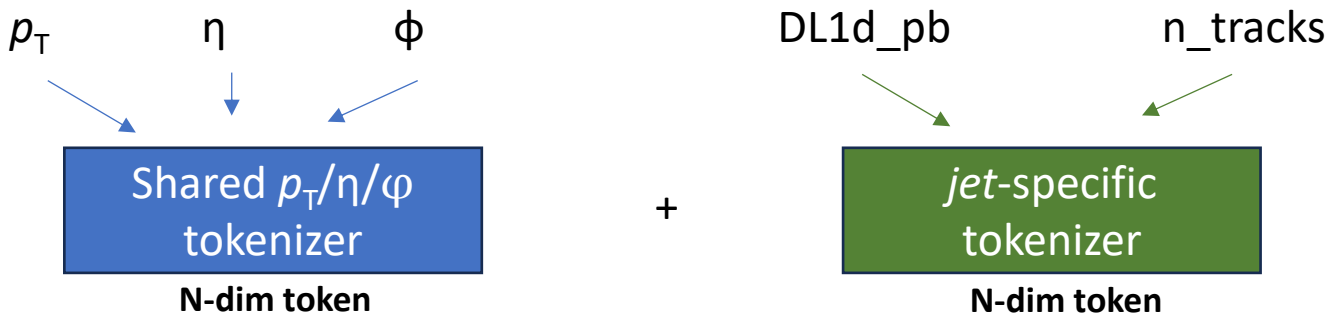
StDev: 0.284

Shared Tokenization Scheme

Electrons



Jets



Summary

- Presented comparison of combined and individual electron + jet tokenization on common features
- Combined tokenization provides better reconstruction metrics for electrons with jets to be added shortly
- Presented shared tokenization scheme for accommodating both common and unique features per object

Backup

e-only Codebook Scan