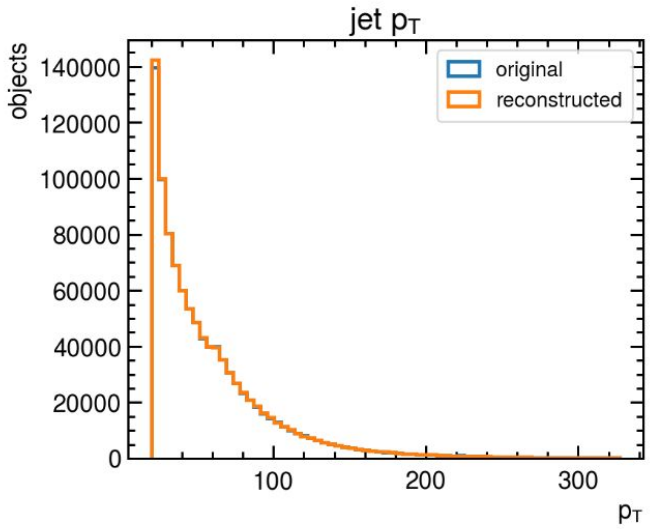
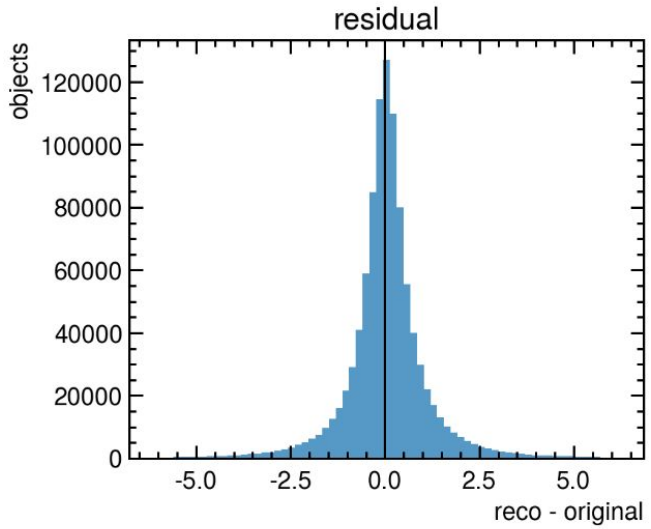
From ATLAS Events to Tokens

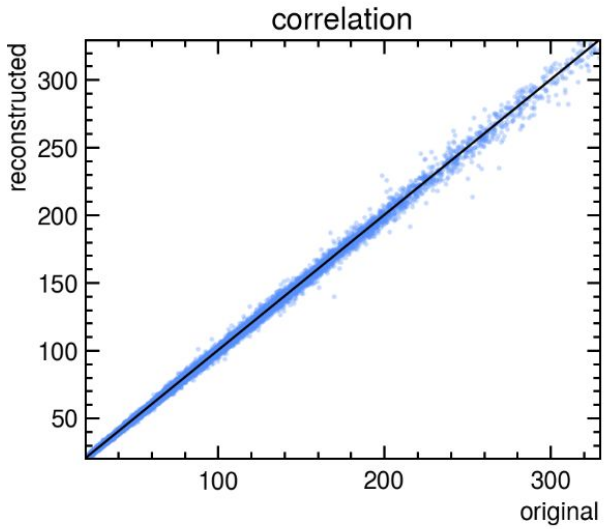
- Each reconstructed physics object is compressed into a few discrete tokens.
- Objects: Jets, electrons, muons, photons, taus, and tracks.
- Event-level information is also included: pileup, MET.
- This creates a common representation of MC and real data.
- Data: ATLAS Open Data Run-2 MC and Data samples.
- Features: Object kinematics ID/isolation variables, b-tagging, track-quality variables.
- Final aim: Combine object tokens into event sequences for downstream foundation model training.

Physics Information is Preserved

- The tokenizer compresses each object into a few discrete token, then decodes back to continuous physics variables.
- Object tokenizers are able to reconstruct key physics variables with small resolution loss.







Distribution shape recovered

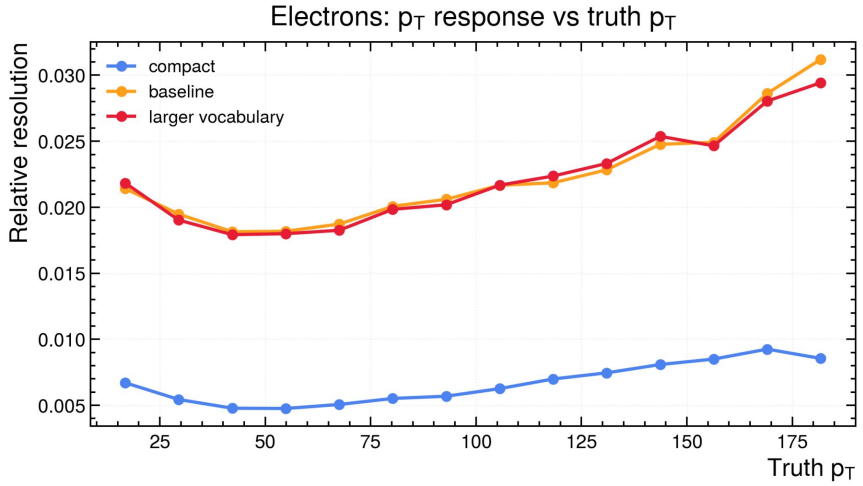
Residuals centered near zero
Jet pT bias ≈ 0

Object by object agreement

Brookhaven National Laboratory

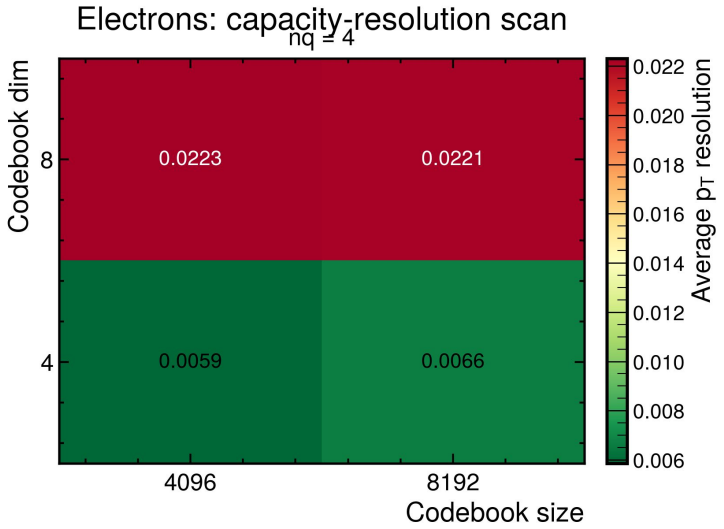
Optimisation studies

- We scan tokenizer capacity to choose a working point that preserves physics-feature resolution while keeping the event representation compact.
  - Scanned tokenizers from O(10^3) to O(10^5) discrete codes per object type, varying number of tokens, codebook size, and code dimension.
- Optimised working points are selected separately for each **object type.**
- Higher capacity tokenizers better preserve object kinematics.



Resolution vs pT: lower curves mean better reconstruction across the kinematic range

compact: dim4, 4k codebook, q4
baseline: dim8, 4k codebook, q4
larger: dim8, 8k codebook, q4

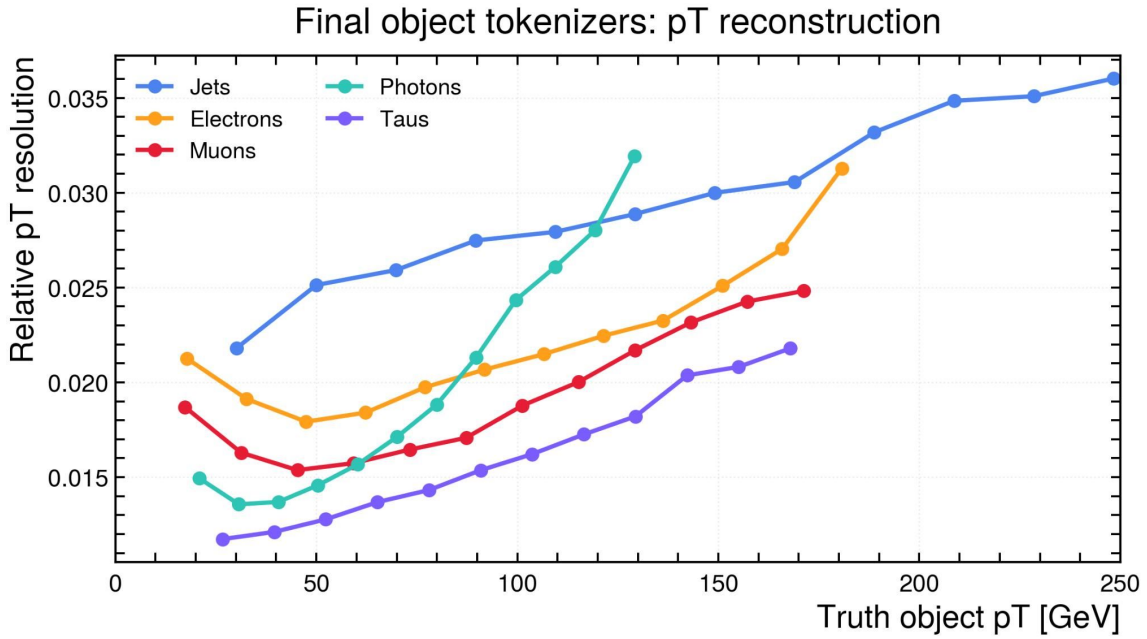


Capacity scan summary: choose the most compact setup with good reconstruction

Brookhaven National Laboratory

Object-Level Reconstruction

- Final selected tokenizers are evaluated across all reconstructed object types.
- Decoded tokens preserve object pT at the few-percent level.
- Resolution remains within the target range across the main kinematic spectrum.



Brookhaven National Laboratory