Progress on CMS Open Data

Oz Amram (Fermilab)
On behalf of CMS / Fermilab TREASURE Collaborators

April 28th, 2026
TREASURE Workshop

Aspen Open Jets Dataset

2412.10504

- Jets selected from open CMS data
- **JetHT** dataset from 2016 G & H
- Used publicly released PFNano tool to process MiniAOD
- Selected large radius (AK8) jets with **minimal requirements**
  - OR over all available triggers
  - Golden JSON & Standard MET filters
  - Jet $p_T > 300$ GeV, $|\eta| < 2.5$
  - Pass ‘tight’ & ‘tightLepVeto’ JetID
- Save all jets in event passing these requirements (separate entries)

178 M jets total

Dataset

- For each jet, save p_T, η, φ, softdrop mass
- Save up to 150 PF Candidates in the jet
  - p_x,p_y,p_z,E, d0, d0err, dz, dzerr, charge, pdgID, puppiWeight
- Also some extra substructure information
  - $\tau_{1\text{-}4}$, ParticleNet scores
- Saved information in h5 file format → easy usage for broader community
- Hosted on Hamburg's Zenodo-like platform (link)

Data Format

- Simple h5 data format
- Chosen to be similar to previous community datasets (JetClass)

Events are stored in h5 format with 4 keys:

- 'event_info', shape (N_jets, 3): [Run Number, LumiBlock, Event Number]
- 'jet_kinematics', shape (N_jets, 4): [pt, eta, phi, softdrop mass]
- 'PFCands', shape (N_jets, 150, 11): Zero padded list of up to 150 PFcandidates inside the jet.
  Info for each candidate is [px, py, pz, E, d0, d0Err, dz, dzErr, charge, PDG ID, PUPPI weight]
- 'jet_tagging', shape (N_jets, 13): Tagging info/scores for the AK8 jet.
  Info for each jet: [nConstituents, tau1, tau2, tau3, tau4, ParticleNet H4q vs QCD, ParticleNet Hbb vs QCD, ParticleNet Hcc vs QCD, ParticleNet QCD score, ParticleNet T vs QCD, ParticleNet W vs QCD, ParticleNet Z vs QCD, ParticleNet regressed mass]

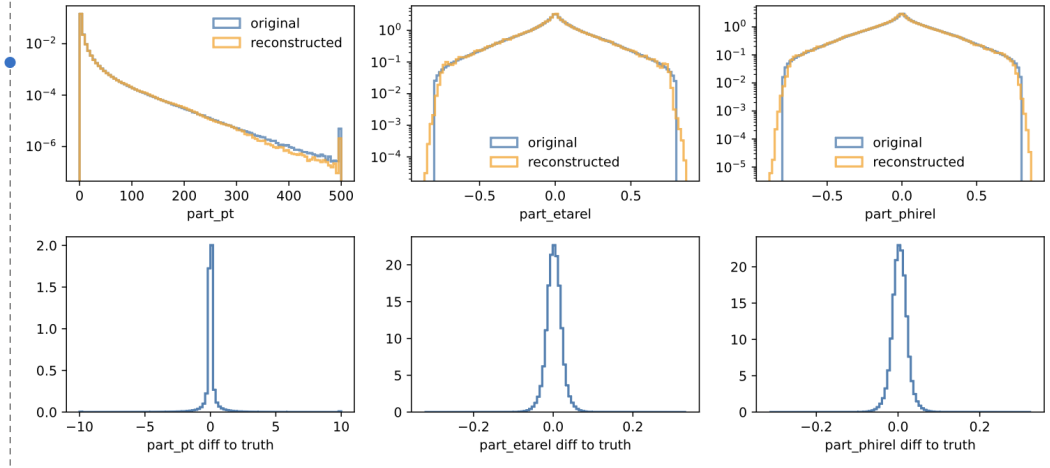
Scaling

- Have infrastructure in place to scale up to run over all CMS Open Data Datasets
  - 2.1 Billion events in 2016
  - Spreadsheet
- Can also think about what MC we want to add

Tokenization

- Baseline: Start with simple VQ-VAE tokenizer used in first AOJ publication
- New student was able to reproduce & expand upon previous AOJ tokenization workflow
- Next steps:
  - Integrate CMS datasets into HEPTokens
  - Event level tokenization studies

Current Tokenization

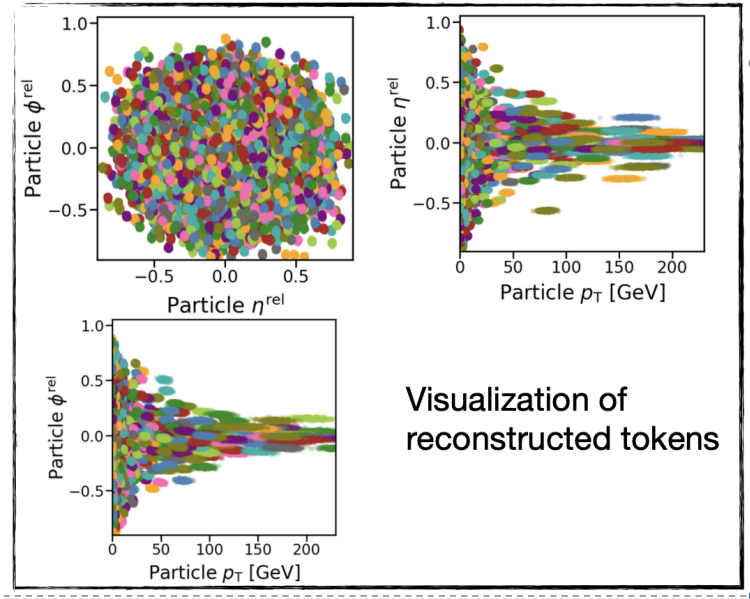


Comparison of original and reconstructed particle pt, η^{rel}, ϕ^{rel} inside jets

$\eta^{rel} = \eta^{particle} - \eta^{jet}$, $\phi^{rel} = \phi^{particle} - \phi^{jet}$

Credit to Pranati Jana!

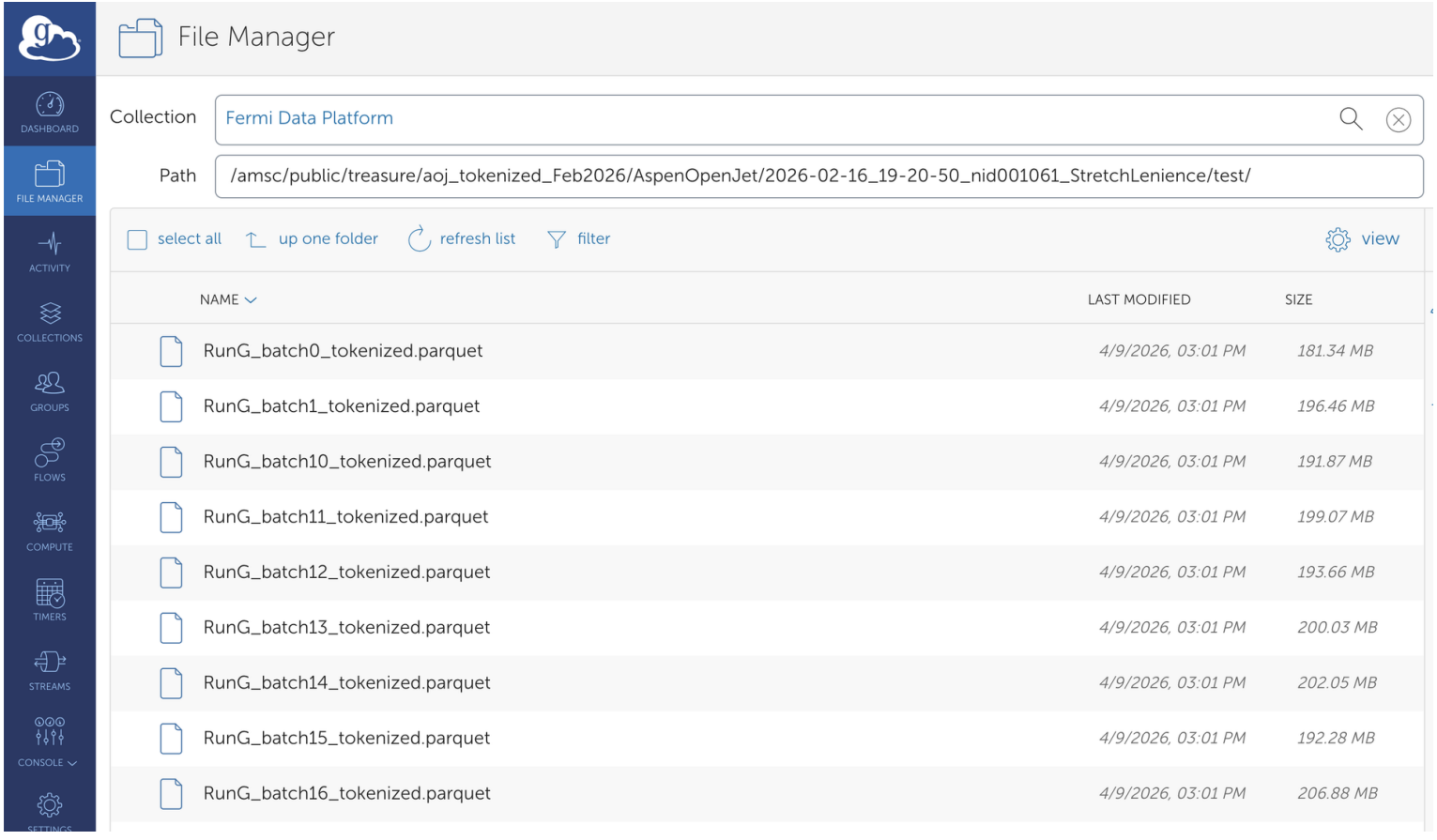
Codebook size 8192



Visualization of reconstructed tokens

Integration with FDP

Current tokenized data is now hosted on FPD!



Integration with HEPTokens

- Have additionally processed CMS QCD and boosted top MC samples
  - Using AOJ h5 format

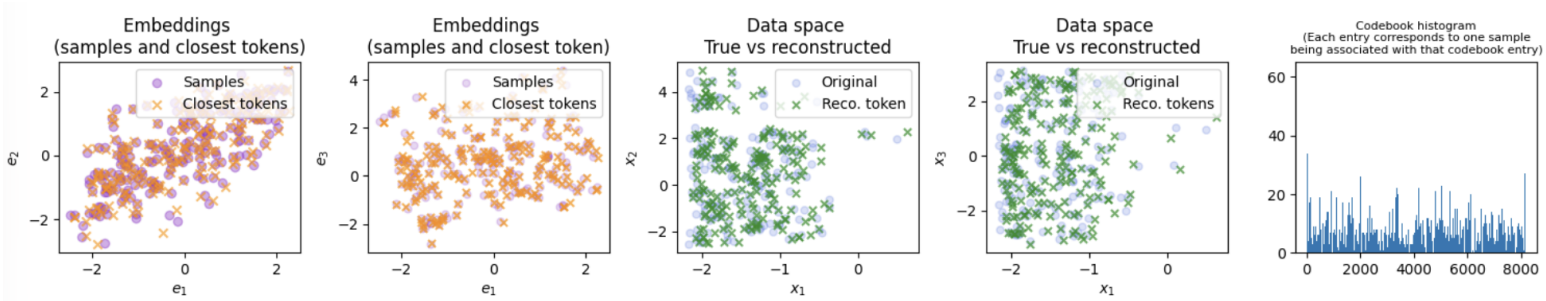
Credit to Aaron Wang!

- WIP: Adding a CMS dataloader HEPTokens to read these files (branch)
- Next: Compare top tagging performance pre- vs post- tokenization

Towards Event Level

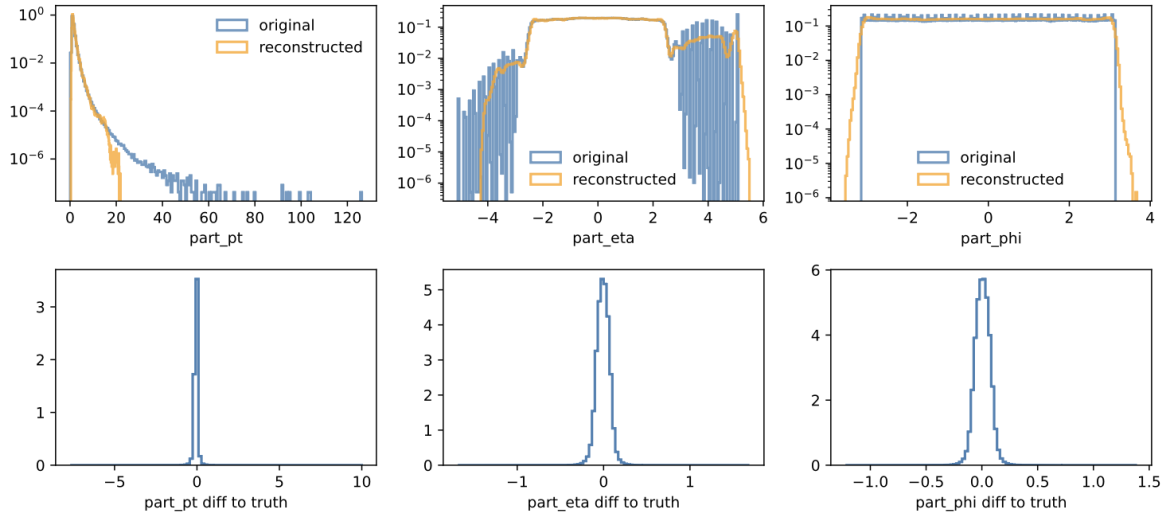
- Started building the pipeline for event-level tokenization
- First simple h5 file format, very similar to AOJ format
- Similar VQ-VAE tokenizer
- First run on 2016G ZeroBias dataset

(Very) WIP: Event-Level Tokenization

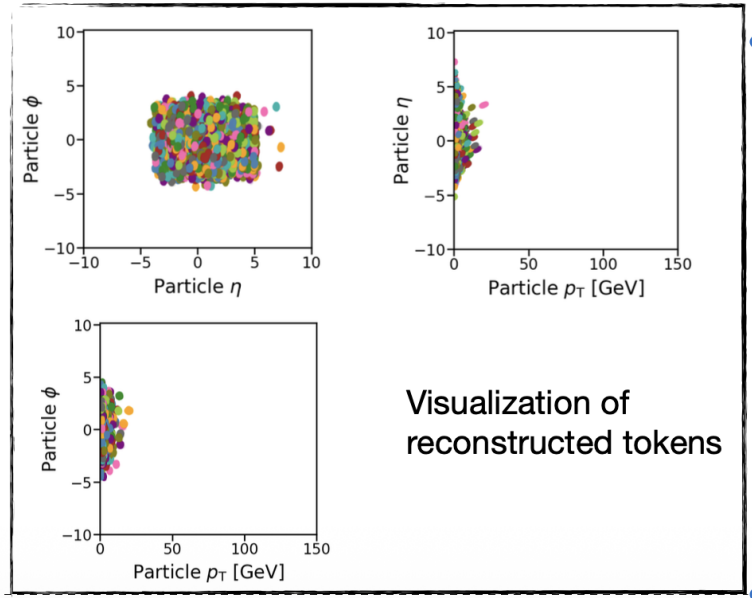


Credit to Pranati Jana!

Codebook size 8192



Comparison of original and reconstructed particle pt, η, ϕ



Visualization of reconstructed tokens

Event Level: Next Steps

- Finalize event level data format (from this workshop :)
- Run over single CMS open dataset (eg ZeroBias)
- Try out improved tokenizer from HEPTokens (?)
- Scale out to all CMS open datasets

Conclusions

- Lots of progress setting up pipeline for CMS datasets
  - Experience gained for students et al
- Next steps
  - Integration with HEPTokens
  - Move towards event level
  - … (many more ideas)

Backup