



U.S. DEPARTMENT
of **ENERGY**



BERKELEY LAB



NATIONAL
ACCELERATOR
LABORATORY



TREASURE

Gabriele D'Amen (BNL)

on behalf of the TREASURE team

September 14, 2026

CLARIPHY - AI Collaboration Meeting - Pyle Center, Madison (WI)



@BrookhavenLab

The TREASURE project



Tokenized Representations for Energy-frontier AI Searches via Understanding and REasoning

DOE HEP American Science Cloud (AmSC) Intelligent Data Activities Pilot

Contacts

Gabriele D'Amen (BNL, co-PI)

Haider Abidi (BNL, co-PI)

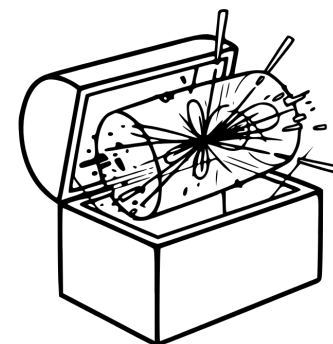
Paolo Calafiura (LBNL)

Walter Hopkins (ANL)

Michael Kagan (SLAC)

Kevin Pedro (FNAL)

Abhijith Gandrakota (FNAL)

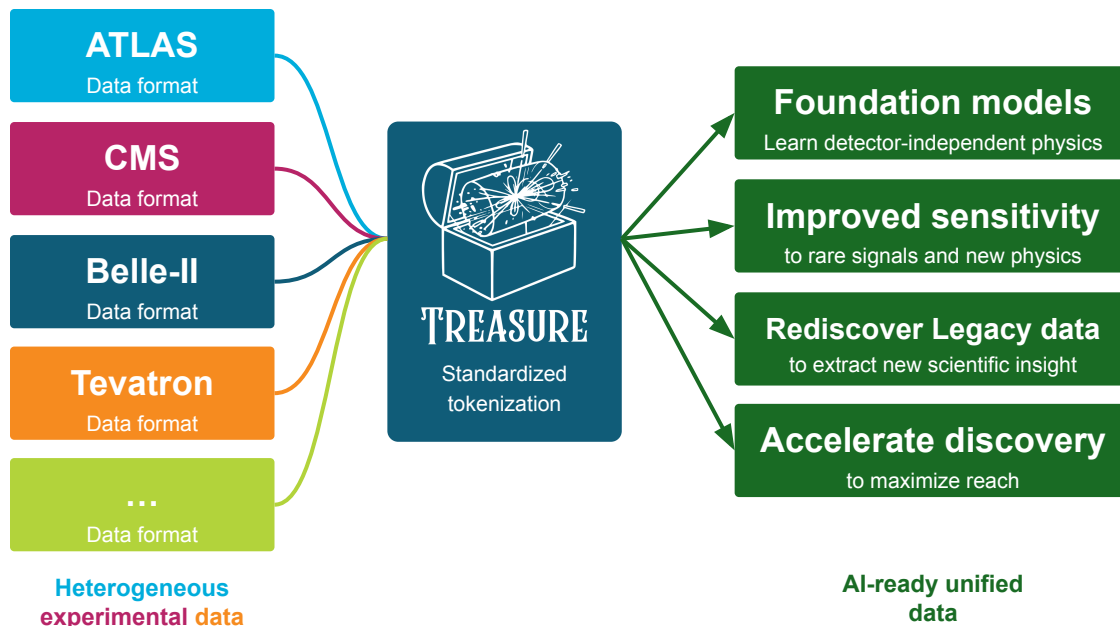


TREASURE

...but > 25 people work on TREASURE!



What is TREASURE?



- *Treasure prepares **collider data** so it can be effectively used by modern AI methods*
- *It transforms heterogeneous, experiment-specific data into standardized, **tokenized**, AI-ready representations*
- **Builds Foundation Models** to learn across experiments, detectors, and eras of data

Goal: unlock new discovery potential from both current and legacy high-energy physics datasets

Recent updates_ object-level tokenization

Tokenization: Vector-quantized variational autoencoder (VQ-VAE) compresses high-dimensional object features into discrete codes chosen from a learned codebook.

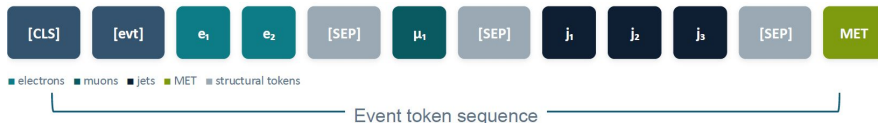
Finalizing and expanding Object-Level Tokenization!

Separate VQ-VAE (and codebook) for each object type to track different features:

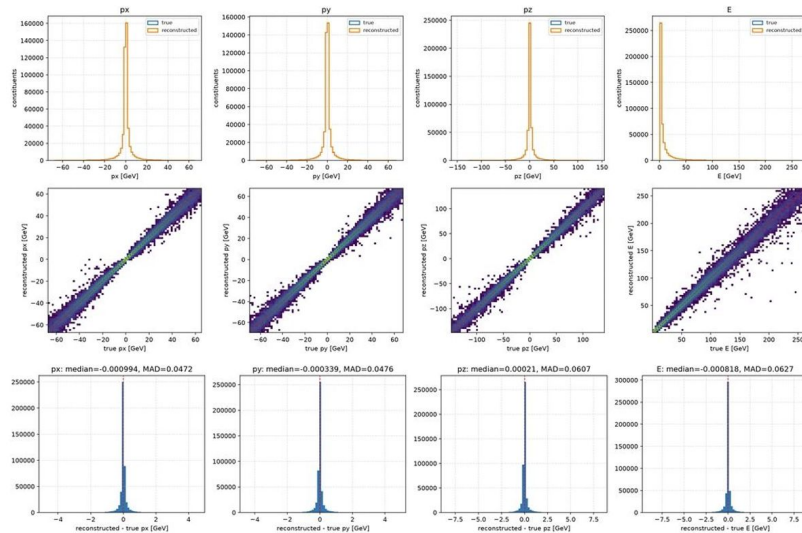
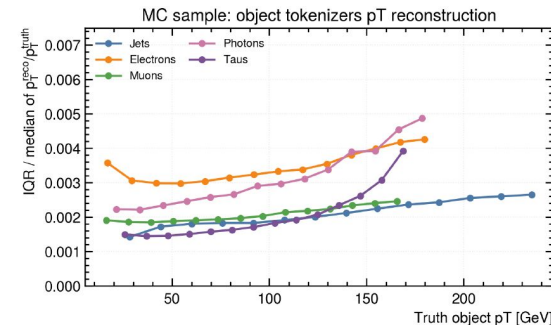
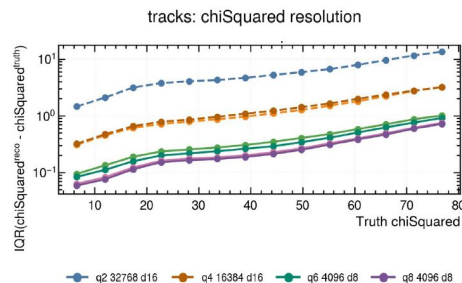
- **Electron:** p_T , η , ϕ , isolation variables
- **Jet:** p_T , η , ϕ , mass, btagging etc
- **Track:** qOverP, η , ϕ , d0, z0, p_T , χ^2 , nDoF

Handles heterogeneous object features, allows for specialization

Advanced studies ongoing for **ATLAS Open Data** and **CMS Aspen Jet Open Data (AOJ)** & **Belle-II** data (preliminary)



(left) Optimization of VQ-VAE encoder for tracks and (right) p_T reconstruction for various objects from **ATLAS OD** ([link](#))



top) Reconstruction of jet-4vectors from **CMS AOJ**

Recent updates_ advanced tokenization strategies

Ramping up Event-level tokenization!

Combined tokenization: single tokenizer for all objects

- Requires scheme to combine different object features
- Tested by combining **electron** and **jet** tokenizers
- *Figure of merit*: reconstructed-true value of tokenized feature

Preliminary result: combined tokenizer on common features shows smaller reconstruction residuals than single tokenizer

Beyond VQ-VAE!

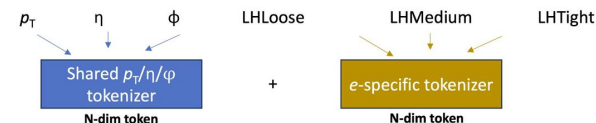
Started investigating alternative tokenization/embedding strategies:

- Particles with similar kinematics & PID should have similar tokens
- Enables **cross-detector** token alignment
- Currently deployed for object-level tokenization -> targeting event-level tokenization (soon)

Shared tokenization

Object tokenizer

Electrons



Jets



Recent updates_ **heptokens** goes public!

Quick start

```
from heptokens.data import
    SingleFileMapModule

from
heptokens.models.vq_vae
import
    LitVqVae

dm = SingleFileMapModule(
    data_path="mc-ttbar.h5",
    num_csts=40)

model = LitVqVae(
    codebook_size=512,
    num_quantizers=3)

trainer.fit(model, dm)
```

Studies presented so far share a common platform: **heptokens**

Goal of heptokens: Let physicists tokenize their detector data without building custom VQ-VAE training infrastructure from scratch

Handles HEP-specific data layouts out of the box

Handles large scale hyperparameter optimization

Available at <https://github.com/Treasure-AmSC/heptokens!>

Try it out!

Work with us!



- ...and so much more!
 - Studies of **low-level information** tokenization underway
 - Evaluation of **advanced tokenization** schemes
 - Started studying tokenization of **Tevatron** data...
 - ...and **FCce** simulations
- We are more than happy to share our **tokenization workflow**: [heptokens](#)
- ...and to provide tokenized open data from ATLAS, CMS, and beyond! (Belle-II underway, Tevatron being explored)
 - Part of ATLAS OD already tokenized. Working on a larger dataset to make public
- ...also, to share our tools to convert experimental data to HDF5 data: [HDF5 converter](#)
- Ongoing collaboration with Genesis **MOSAIC** Phase-I and continuing collaboration with **FM4NPP** (next talk)
- **Happy to collaborate!**
 - reach out to Gabriele (gabriele.d'amen@cern.ch) or Haider Abidi (sabidi@cern.ch) – or find me at the coffee break

Links



TREASURE @ Blueprint Workhop, Washington DC: [TREASURE Blueprint Workshop](#)

TREASURE @ EuCAIFCon 2026, Heidelberg (DE): [EuCAIFCon_Treasure.pdf](#)

heptokens on GitHub: <https://github.com/Treasure-AmSC/heptokens>

HDF5 converter on GitHub: [HDF5 converter](#)

Backup