

TREASURE

Progress and next steps

Merve Nazlim Agaras

23 September 2026

Code

Polina, Nazlim and Jeff will have regular meetings to distribute the tasks and discuss the results

Running on 3 different machines

WIP personal code ->

https://github.com/MerveNazlim/heptokens/tree/event_level

DDP, parallel running in different GPUS included

Tokenisation + parquet + pretrain + fine tune (both from tokenized data and continuous input and able to hierarchical tokenization)

CAVEAT

Heptokens are refactored recently, this is the version that is public in <https://github.com/Treasure-AmSC>

Input h5 files

Being copied under, 20GB of MC + ~1T [15-16 ATLAS Open data](#)
<https://cernbox.cern.ch/s/4fQgnW0Eodwdj46>

What the code does today

Reads event data

Loads the common TREASURE format and prepares the features used by each object type.

Builds representations

Trains object tokenisers and writes complete events in a form that the event model can read.

Trains event models

Supports pretraining on unlabelled events, followed by training for a chosen physics task.

Evaluates the result

Checks reconstruction and classification. Code for continuous-input comparisons is also available.

All these based on the `heptokens` repo.

What the current paper will include

Complete workflow

Describe the implemented ATLAS PoC from experiment data to a trained event model.

Tokenisation study

Explain the object representation and how we selected a useful reconstruction setup. Show scans, perplexity and other tokenization studies.

Physics validation

Show reconstruction of high level variables (DR, Dll etc) from pretrained backbone.

Pretrain+finetune vs frozen comp.

Include event pretraining and the Higgs-to-four-lepton vs background discrimination with training from scratch and multi-class classification of production modes of Higgs-to-four-lepton.

Tokenisation TODOs

- **Identify the essential input features.** Test which measurements can be removed without losing reconstruction quality or physics performance.
- **Explore a shared tokeniser.** Test one model for different object types, including objects with different sets of input features.
- **Test context-aware tokenisation.** Compare encoding each object alone with encoding it using information from the rest of the event.
- **Compare with hierarchical continuous inputs.** Test whether grouping related features can match or improve on tokenisation.
- **Check robustness.** Study rare objects and differences in reconstruction and physics performance between data and simulation.

Event representation TODOs

- **Test event ordering.** Compare different ways of ordering the objects and check whether the result depends on that order.
- **Check type and position embeddings.** Measure whether object identity and sequence position provide useful information.
- **Compare event summaries.** Test different ways of combining object representations into one event representation.
- **Study model size.** Compare physics performance and training cost as the event model becomes larger.

Pretraining TODOs

- **Choose the prediction target.** Compare recovering token IDs, physical measurements or a learned representation of the hidden object.
- **Test the masking strategy.** Vary which objects are hidden and how much of each event is removed.
- **Compare learning objectives.** Test masked-object prediction against next-token prediction with a defined event ordering.
- **Test autoregressive prediction.** Compare predicting an object's codes together with predicting them one at a time.

Physics TODOs

- **Add a more demanding physics task.** Select a sample with different object content or a more complex topology than HZZ.
- **Test changes in data conditions.** Evaluate models with different simulations, detector conditions or data-taking periods.
- **Explore pretraining with more experiments.** Choose a CMS or Belle sample, define a compatible object and feature description, and test its inclusion in pretraining.