

Technical  
University  
of Munich



# Introduction to PHYSLITE

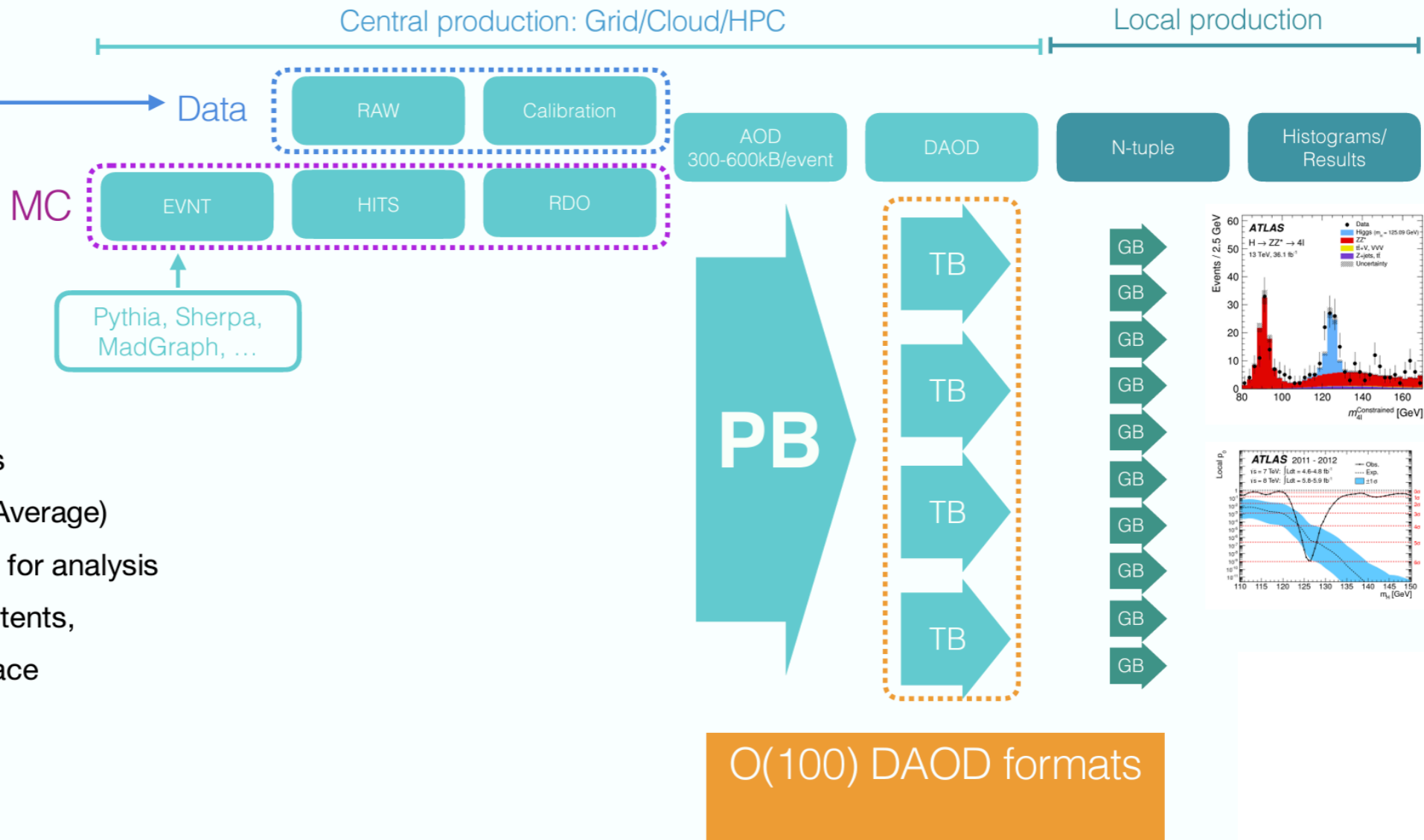
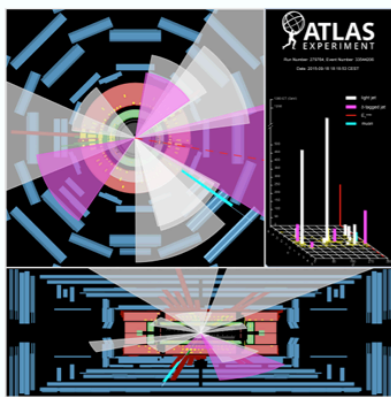
US ATLAS / IRIS-HEP Analysis Software Training  
Event 2024

Vangelis Kourlitis - 18 July 2024

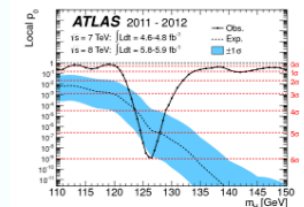
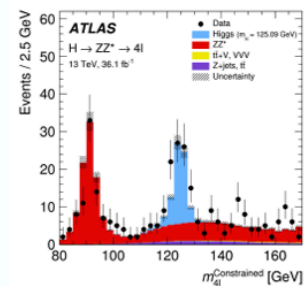
# Data formats

## Run 2

Figure: J. Catmore



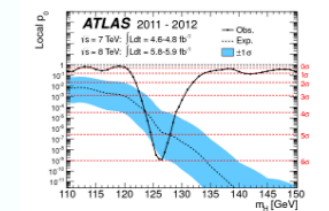
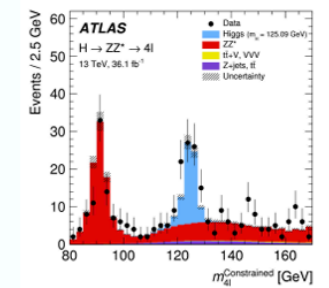
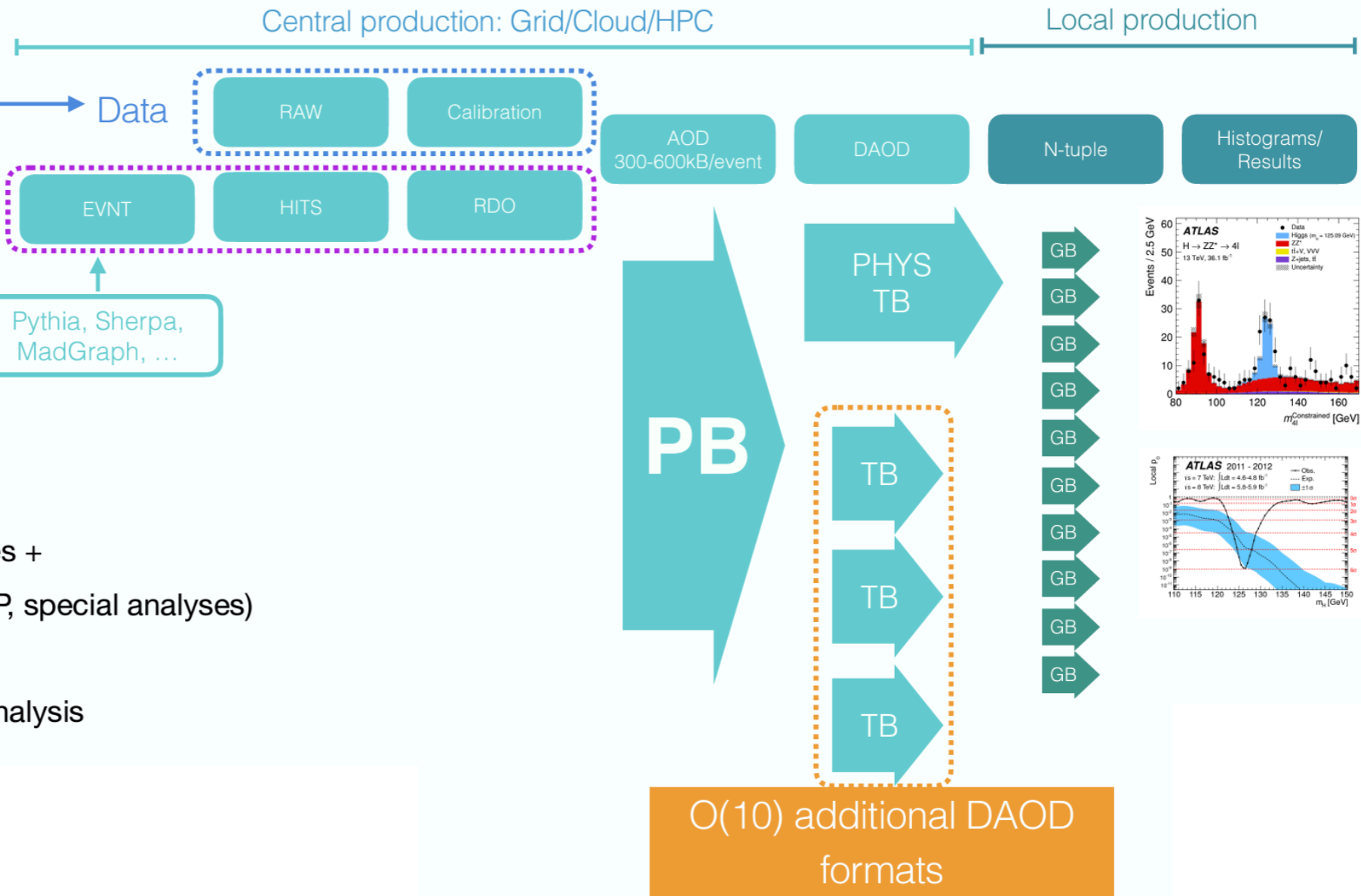
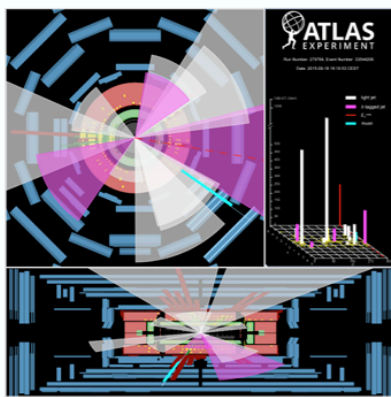
- O(100) DAOD formats
  - **30-50 kB/event** (Average)
- final flat ROOT tuples for analysis
- Large overlaps in contents,
- using a lot of disk space



# Data formats

## Run 3

Figure: J. Catmore

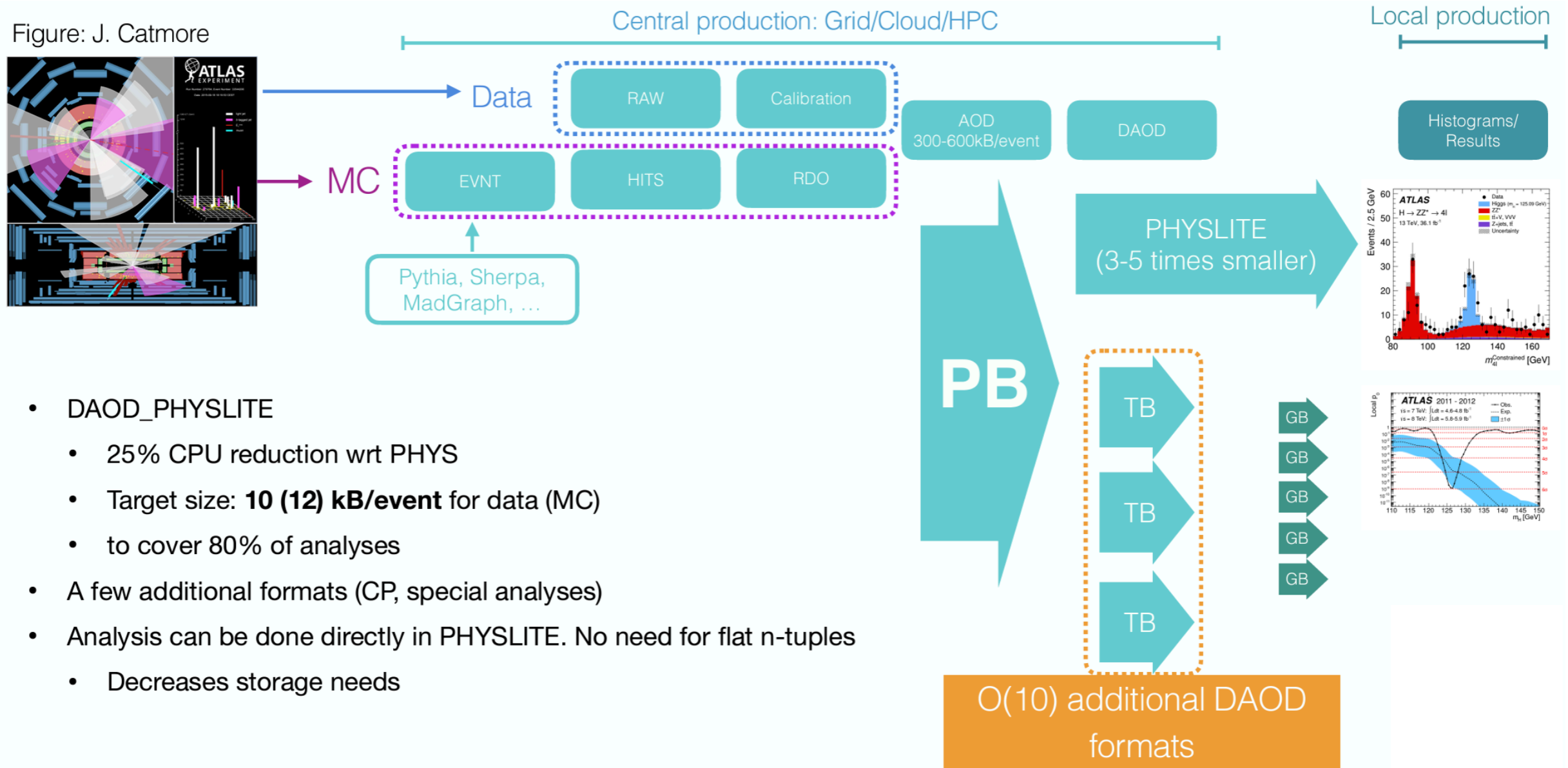


- DAOD\_PHYS
  - **30-50 kB/event**
  - to cover 80% of analyses +
- A few additional formats (CP, special analyses)
- Highly reduces overlaps!
- Final flat ROOT tuples for analysis
- Less total disk space

# Data formats

## Run 4+

Figure: J. Catmore

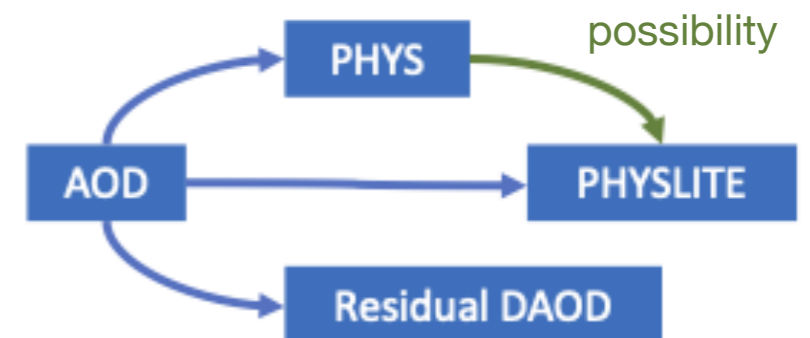


- DAOD\_PHYSLITE
  - 25% CPU reduction wrt PHYS
  - Target size: **10 (12) kB/event** for data (MC)
  - to cover 80% of analyses
- A few additional formats (CP, special analyses)
- Analysis can be done directly in PHYSLITE. No need for flat n-tuples
  - Decreases storage needs

# PHYSLITE

## Main characteristics

1. Unskimmed and monolithic
2. Contains already calibrated objects
3. Objects are loosely preselected
4. Available since Release 22
5. Centrally produced ~4 times per year  
*April '23, September '23, January '24, June '24, ...*



# PHYSLITE

## Status

In general PHYSLITE model seems to provide good coverage of the bulk of analyses

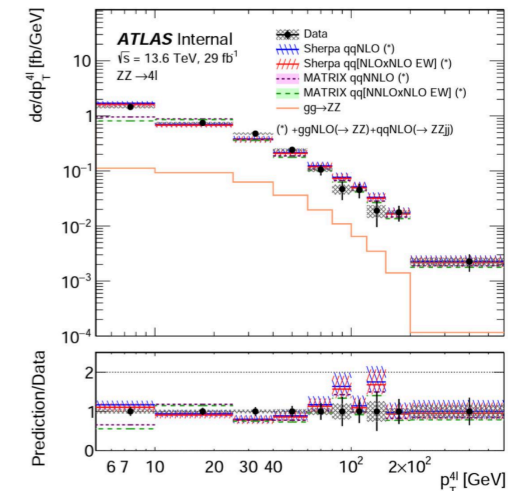
*Adoption plans from various PA groups (Top, di-Higgs, SM, etc.)*

*Current PHYSLITE usage is comparable to the most used SM-specific format*

*Special format needs only from BLS, HION and LLP searches*

First ATLAS result using PHYSLITE

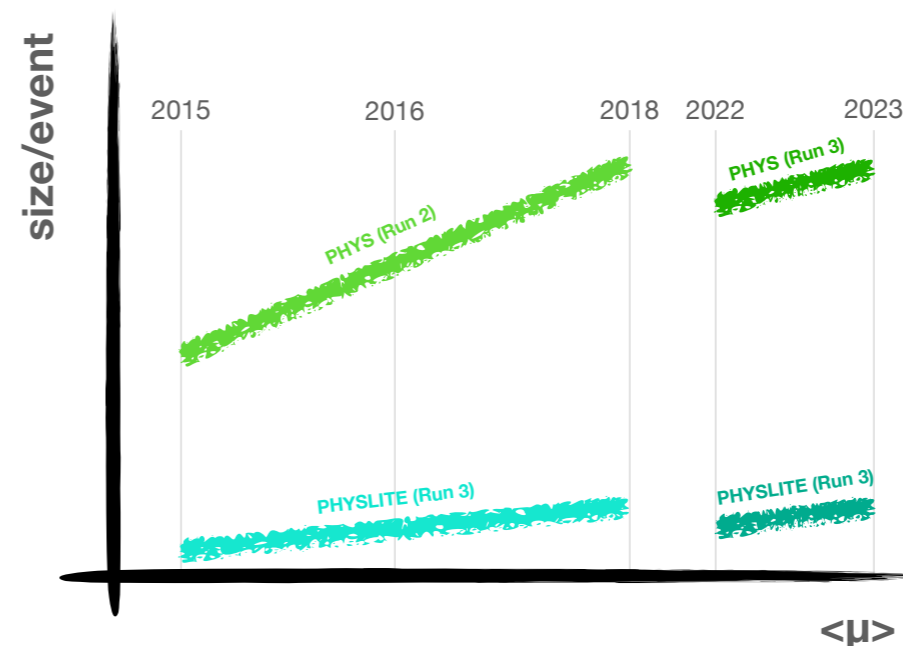
[arXiv:2311.09715](https://arxiv.org/abs/2311.09715)



ZZ production cross-sections in the four-lepton final state

Current event size very close to the ~12kB target

Event size	C DAOD_PHYS	C DAOD_PHYSLITE
	29.2 kB	11.9 kB

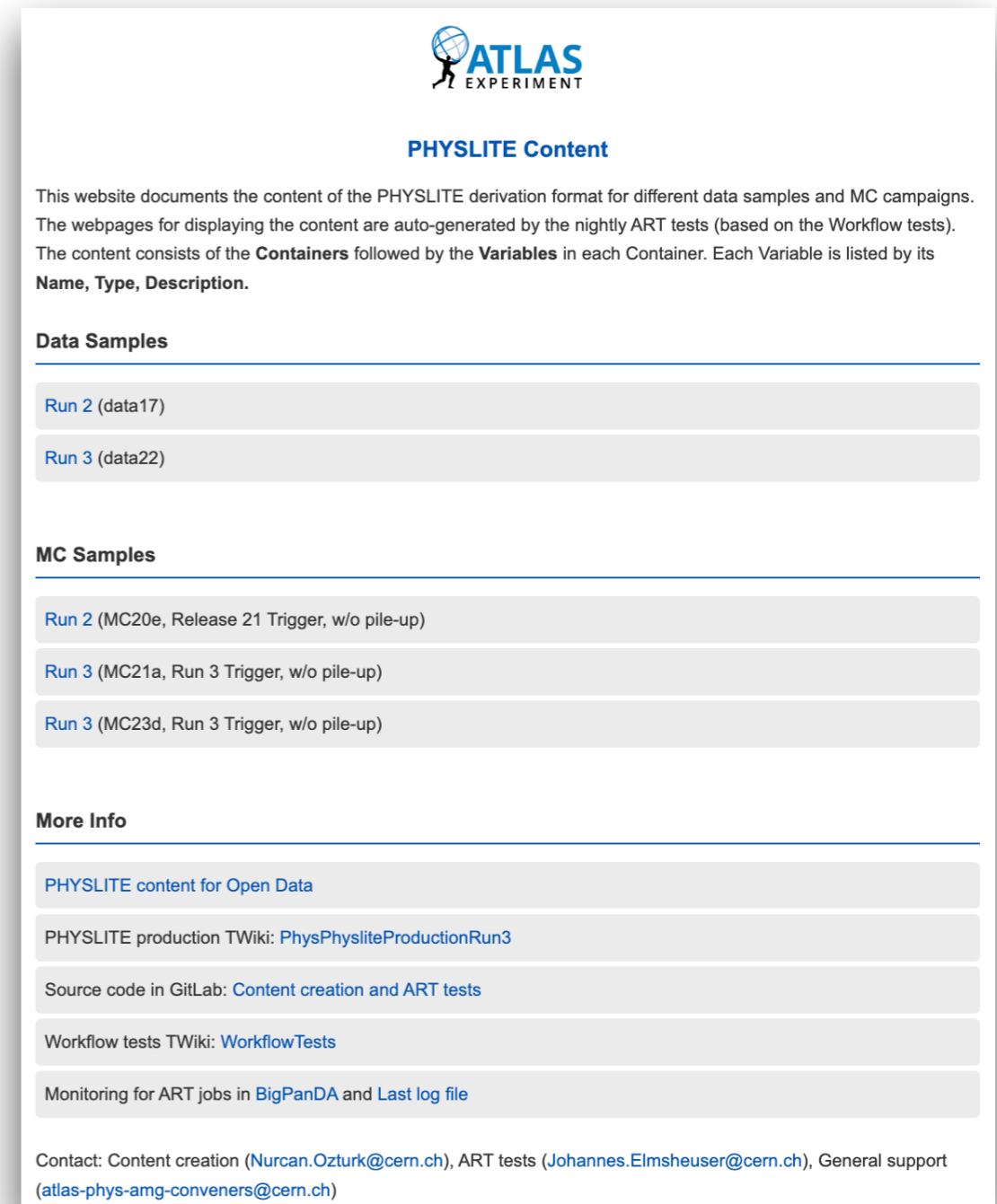


~85% more stable in pile-up increase, however projections show 15kB/event in HL-LHC  $\langle\mu\rangle$   
R&D continues!

# PHYSLITE Content

## New webpage!

- Summarising information from various CP and PA groups
  - Auto-generated by ART tests
- PHYSLITE is an **evolving data format**
  - [TWiki](#) on how to request modifications to PHYSLITE
  - The more analyses your request helps, the more likely is to be accepted!



The screenshot shows the 'PHYSLITE Content' webpage. At the top right is the ATLAS EXPERIMENT logo. Below it is the title 'PHYSLITE Content'. A paragraph explains that the website documents the content of the PHYSLITE derivation format for different data samples and MC campaigns, and that the webpages are auto-generated by nightly ART tests. It states that the content consists of Containers followed by Variables in each Container, with each Variable listed by its Name, Type, and Description.

**Data Samples**

- Run 2 (data17)
- Run 3 (data22)

**MC Samples**

- Run 2 (MC20e, Release 21 Trigger, w/o pile-up)
- Run 3 (MC21a, Run 3 Trigger, w/o pile-up)
- Run 3 (MC23d, Run 3 Trigger, w/o pile-up)

**More Info**

- [PHYSLITE content for Open Data](#)
- PHYSLITE production TWiki: [PhysPhysliteProductionRun3](#)
- Source code in GitLab: [Content creation and ART tests](#)
- Workflow tests TWiki: [WorkflowTests](#)
- Monitoring for ART jobs in [BigPanDA](#) and [Last log file](#)

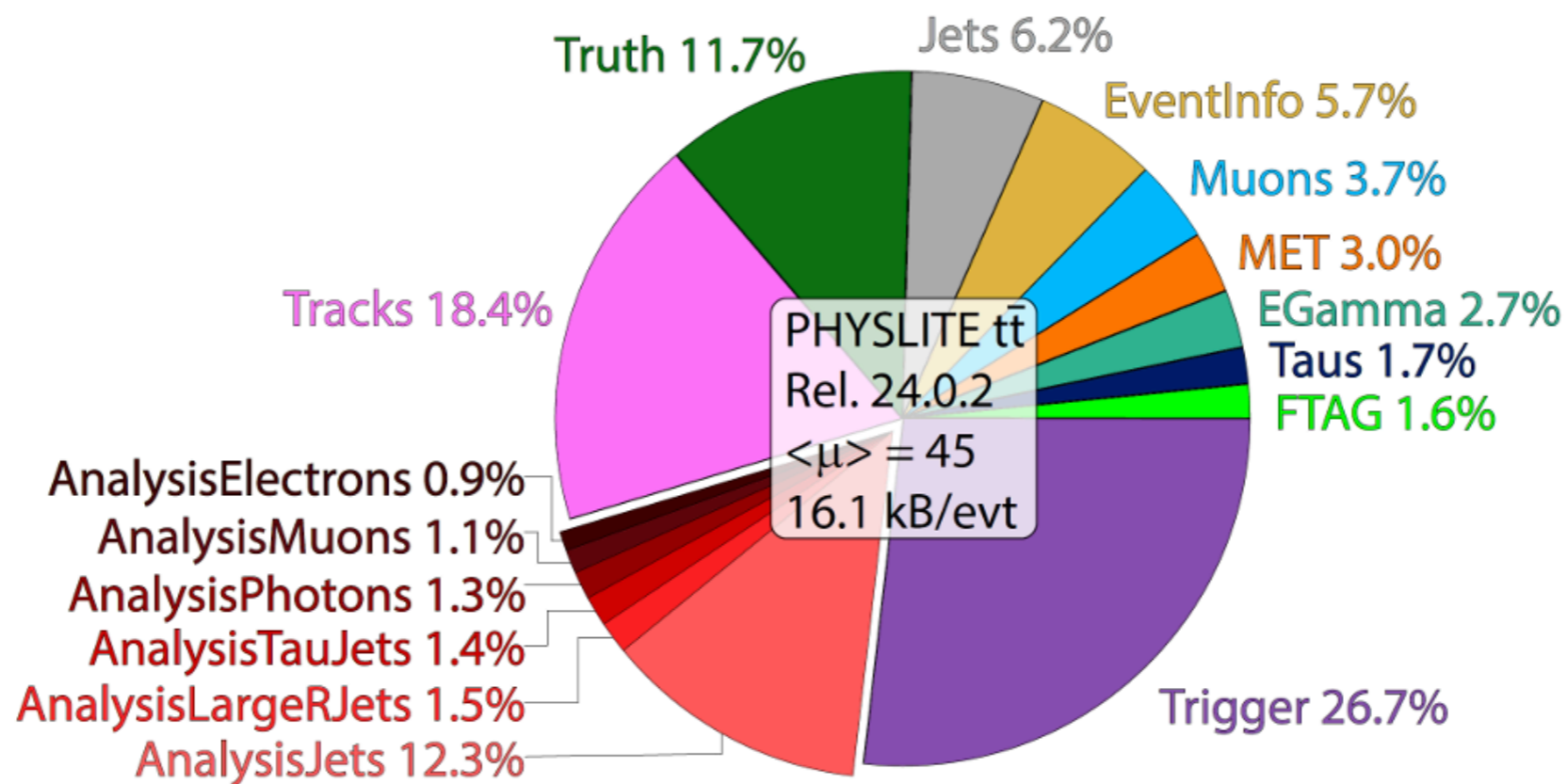
Contact: Content creation ([Nurcan.Ozturk@cern.ch](mailto:Nurcan.Ozturk@cern.ch)), ART tests ([Johannes.Elmsheuser@cern.ch](mailto:Johannes.Elmsheuser@cern.ch)), General support ([atlas-phys-amg-conveners@cern.ch](mailto:atlas-phys-amg-conveners@cern.ch))

<https://atlas-physlite-content.web.cern.ch/>

# PHYSLITE Content

## Breakdown

Mainly information from  
the [TRUTH3](#) format





# PHYSLITE

## R&D

2023

Format development and early testing

### R&D projects:

- GRL information  
*incorporate variables indicated if an events passed a GRL*
- Event augmentation  
*add additional information only to a subset of events*
- Lossy compression  
*reduce variables precision — loose information but do not compromise physics*
- RNTuple adoption  
*TTree evolution offering space savings at least 10% and commensurate speed up*
- **Columnar analysis**  
*array programming methods for on the fly systematics*

2024

2025

2026



**PHYSLITE ready for super-fast lightweight physics data analysis**

# Open Data for Research

**Imminent release of 2015+2016 physics main proton–proton data as PHYSLITE**

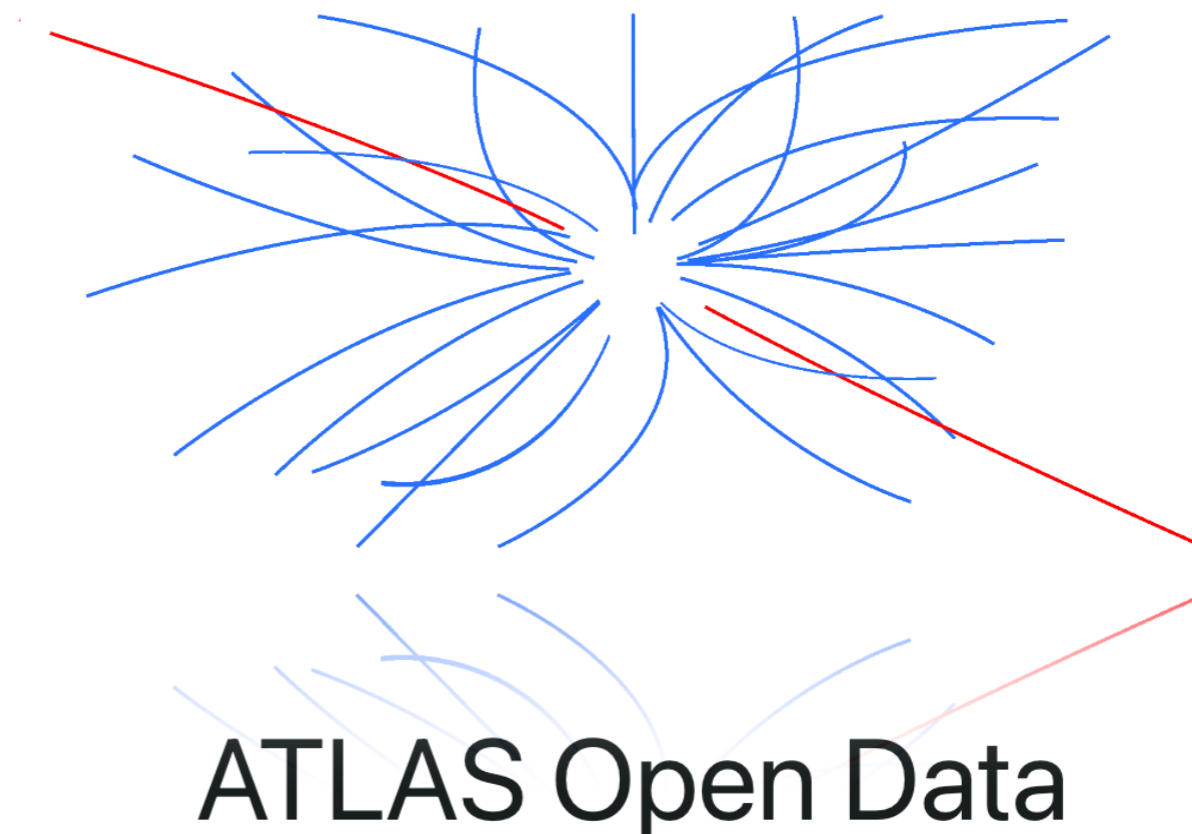
- 2015 Pb+Pb data too in PHYSLITE-like format

**Support real (or at least realistic) full-blown analyses (including systematics)**

- Accompany with “an appropriate set of simulated Monte Carlo samples”
- Distributed by [opendata.cern.ch](http://opendata.cern.ch)

**Support material**

- Analysis software (Athena) is already public
- ATLAS open data portal with extensive documentation: [opendata.atlas.cern](http://opendata.atlas.cern)



**ATLAS Open Data**

High Energy Physics data for everyone.

# Today

- Fetch a PHYSLITE file from the Open Data release
- Load it in memory using `coffea NanoEvents` and `PHYSLITESchema`
- Demonstrate basic data handling
  - More about `coffea` and actual analysis later:  
[Coffea columnar analysis framework](#)

