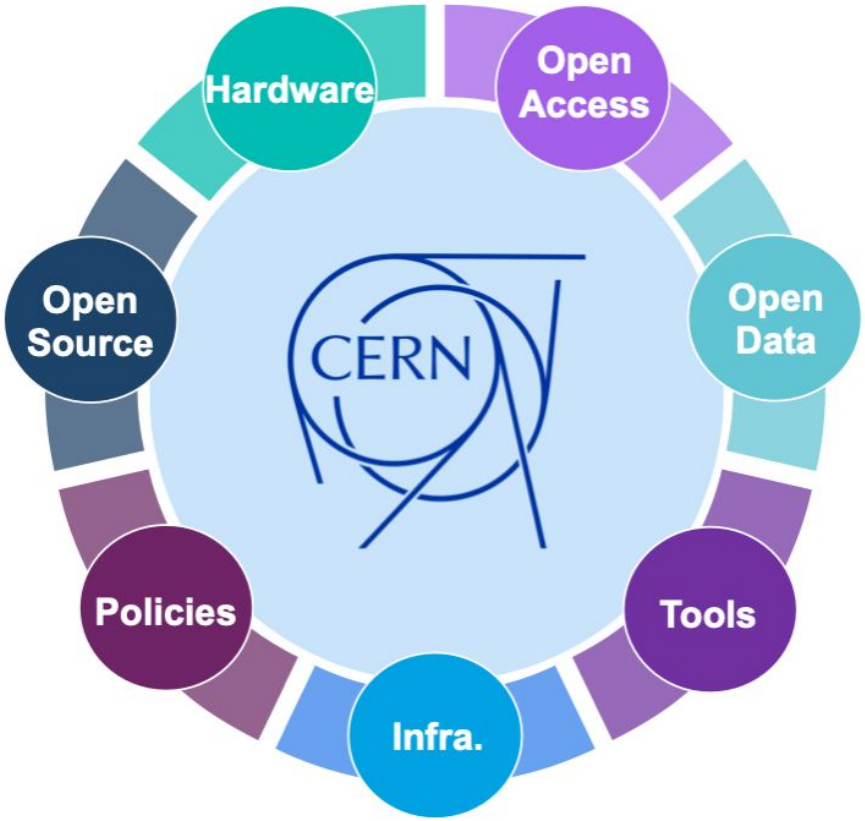
The First Release of ATLAS Open Data for Research

Mariana Vivas Albornoz (UMass)
On behalf of the ATLAS Collaboration

ICHEP
Outreach & Education
18 July 2024

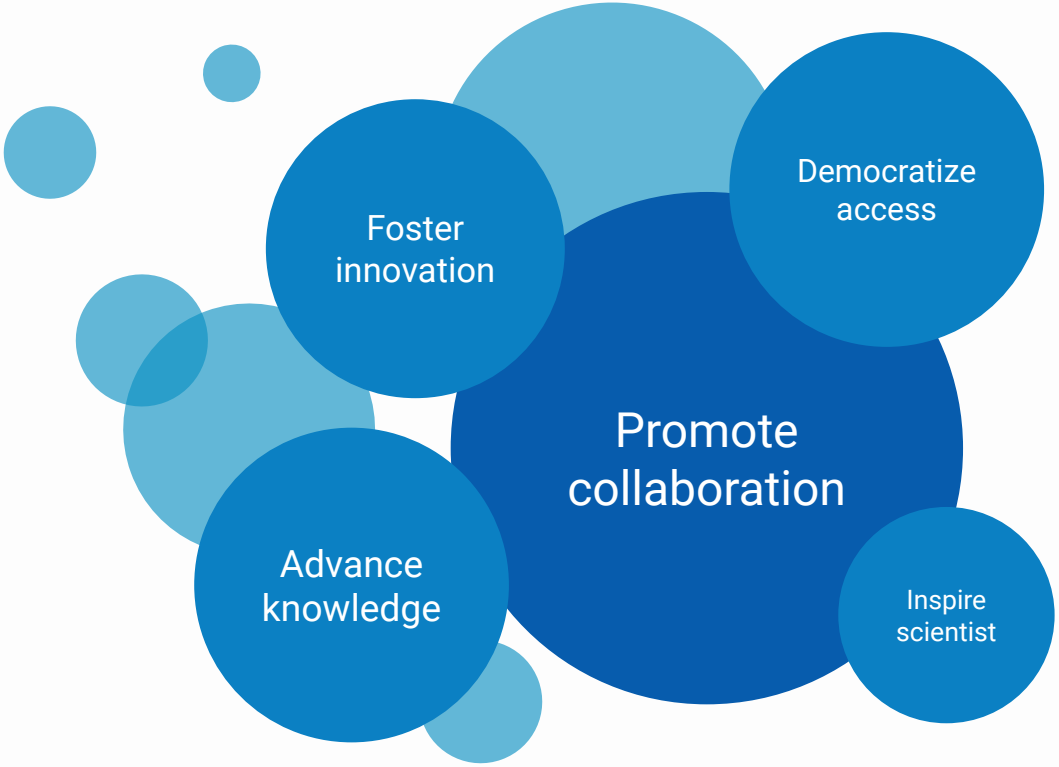
mariana.isabel.vivas.albornoz@cern.ch

Open data policy



In accordance with CERN's open data policy, 25% of the data should be made publicly available 5 years after the end of each Run.

Why open data for research?

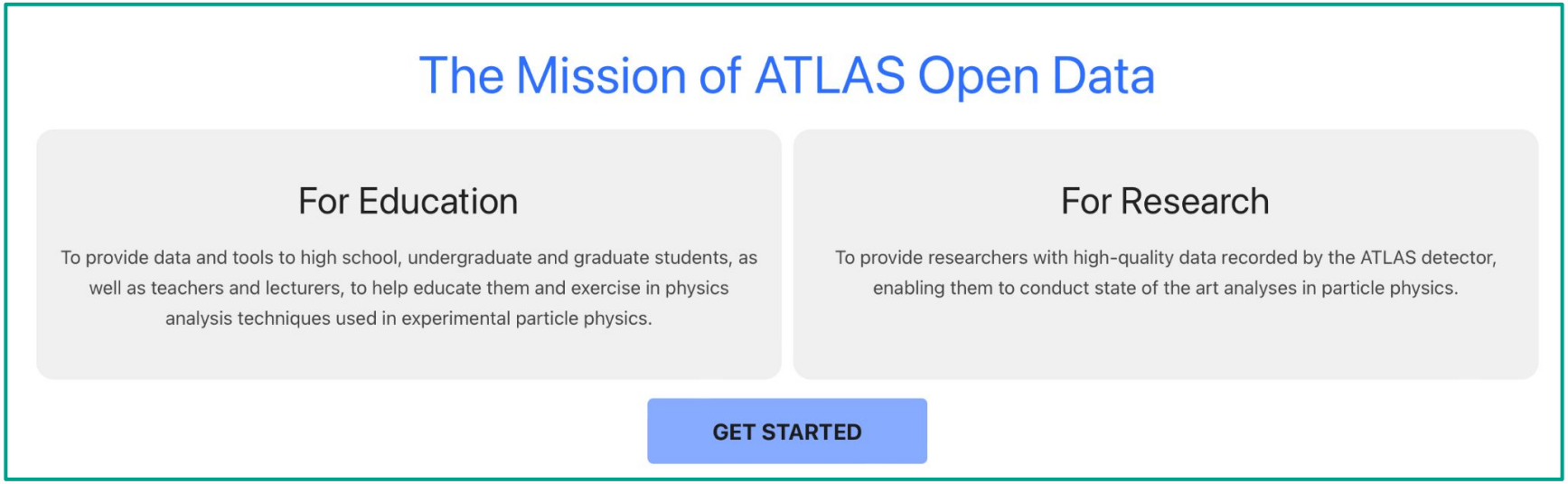


In accordance with CERN's open data policy, 25% of the data should be made publicly available 5 years after the end of each Run.

The release of our data is not the end goal but the beginning of a global collaborative quest aimed at enhancing discovery, and fostering innovation worldwide.

The ATLAS Open Data project

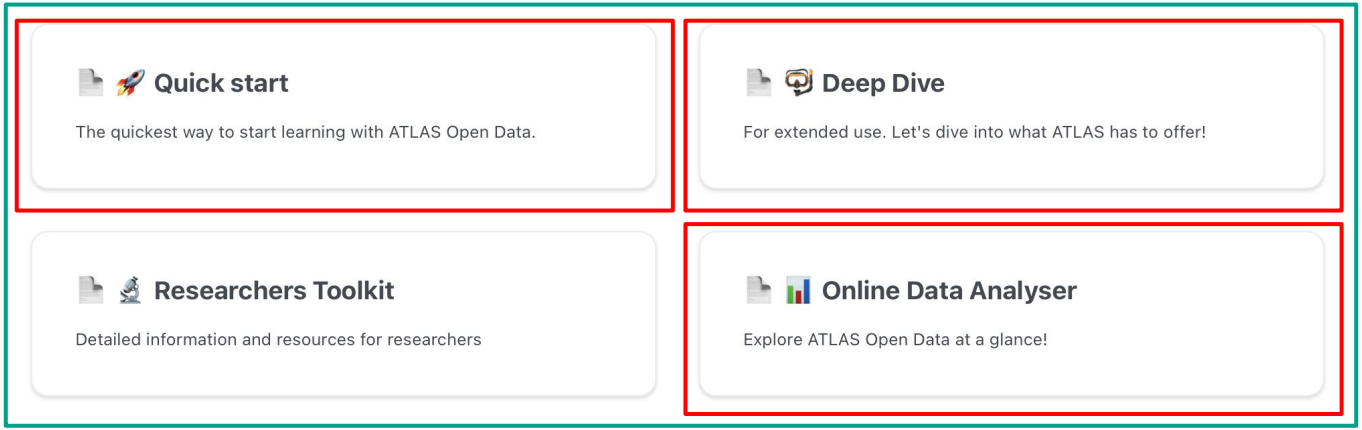
The ATLAS Open Data project started as an educational endeavor in 2016.
Since then, we released two datasets for education: 8TeV and 13TeV.



opendata.atlas.cern

Open data for education

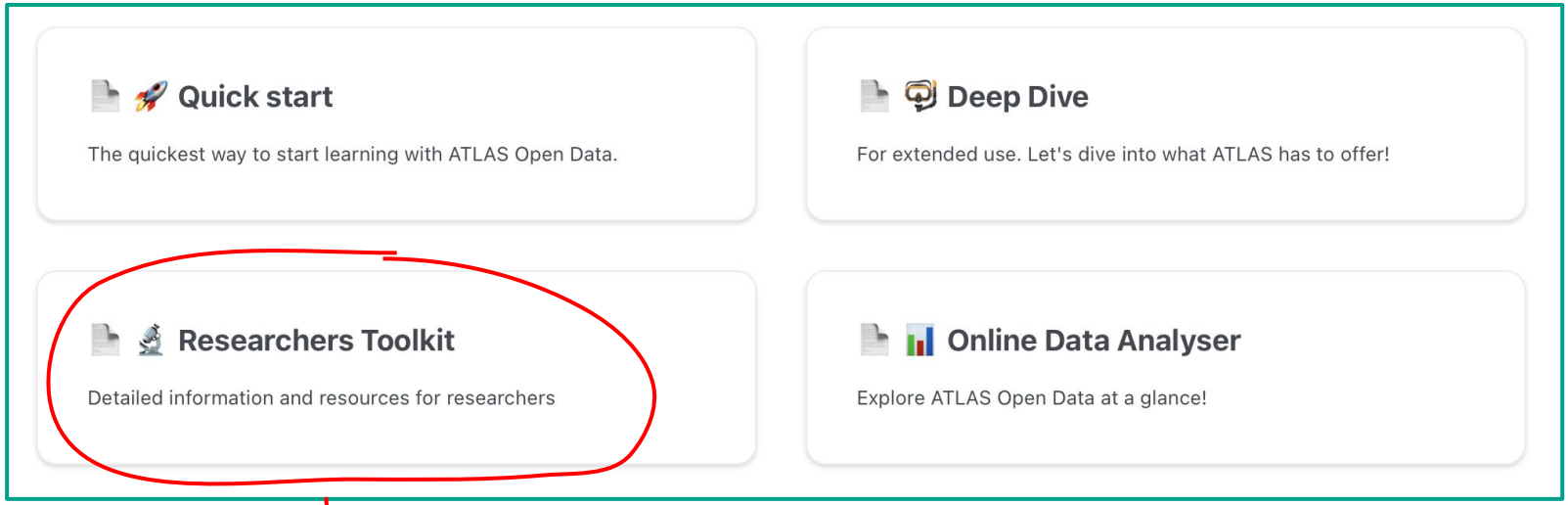
For education, we have three different paths for people to follow, depending on how they want to approach the information.



Used widely in schools, universities, events, and by individual learners worldwide!

Target audience

With a new path that we call “researchers toolkit”, we are now expanding to a wider audience, from university onwards.



People interested in making complex analyses or trying new technologies.

Considerations about the data

- **Data:** Is there enough data for enough use cases? What format should the data be in?
- **License:** What constraints will be put on data usage?
- **Accessibility:** How will users access the data?

Released detector data and simulations

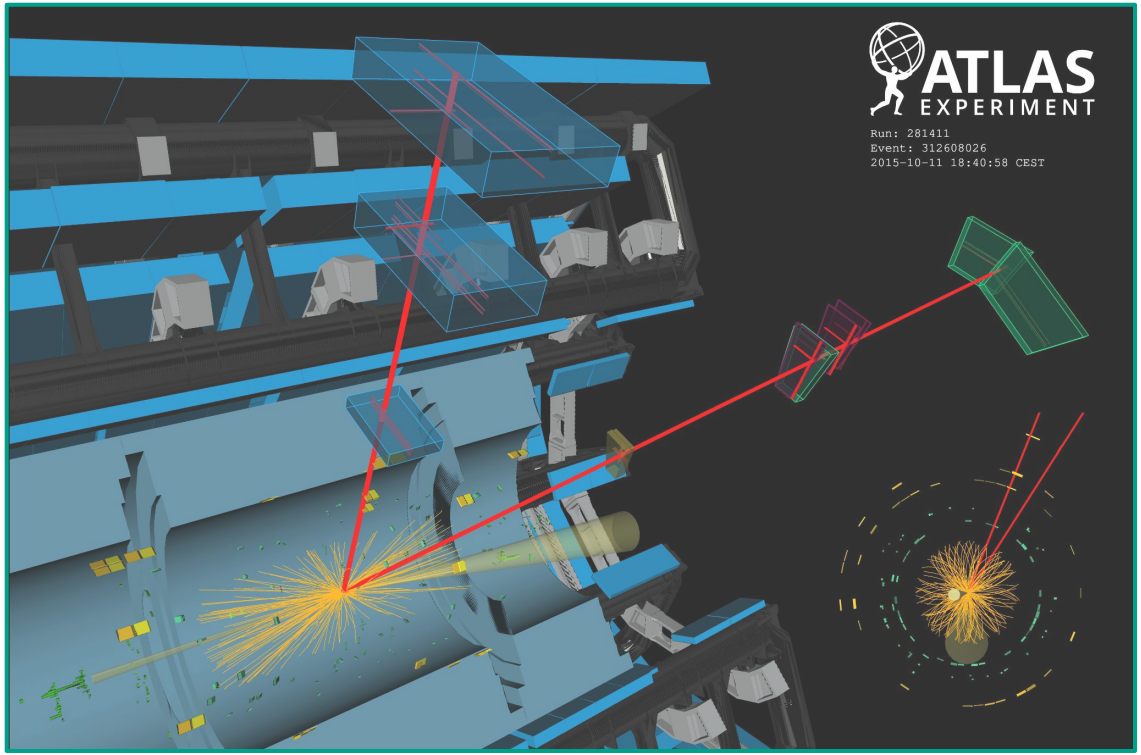
Detector data

All the data from proton-proton collisions of the **2015 and 2016 runs** with a combined luminosity of ~36 fb^{-1}.

Monte Carlo simulations

Over 300 datasets:

- Standard Model nominal samples and alternatives for systematic variations.
- Beyond the Standard Model signal samples.



Format

Detector data
Monte Carlo simulations

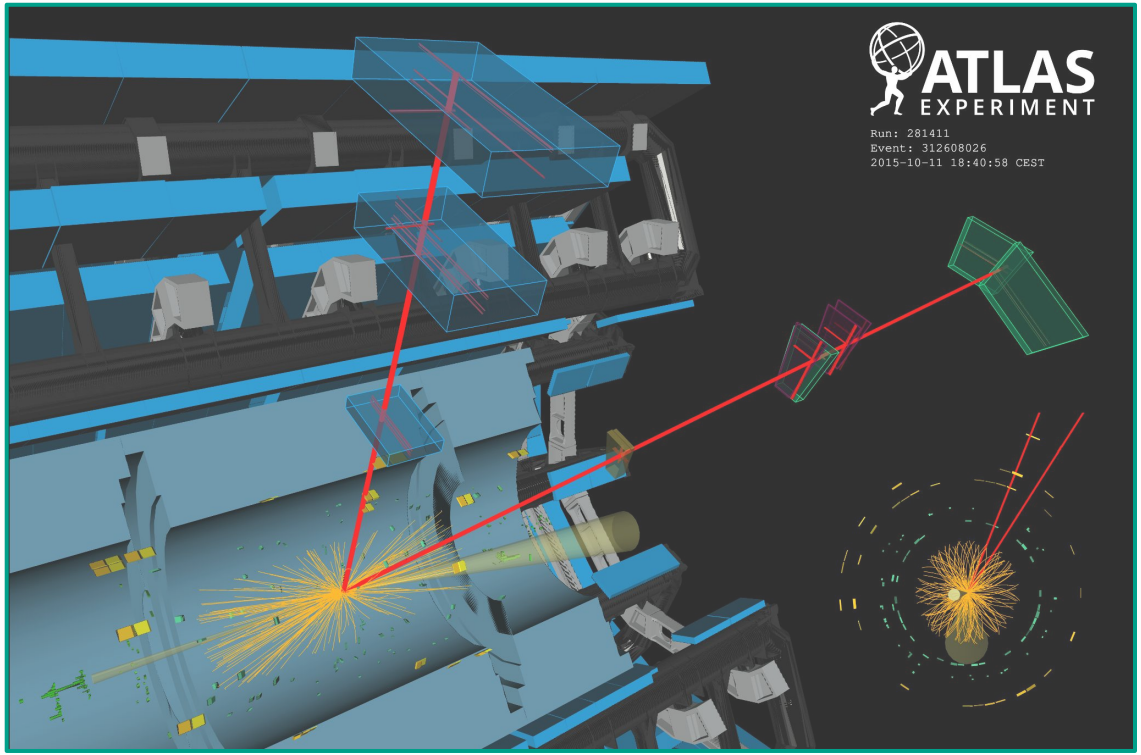
DAOD_PHYSLITE

Used for ATLAS analyses.

Size: 10kb per event for data and 12kb for MC.

It **contains already calibrated and pre-selected objects** and high level information.

Can be analyzed directly, which decreases the storage needs.



License

Detector data

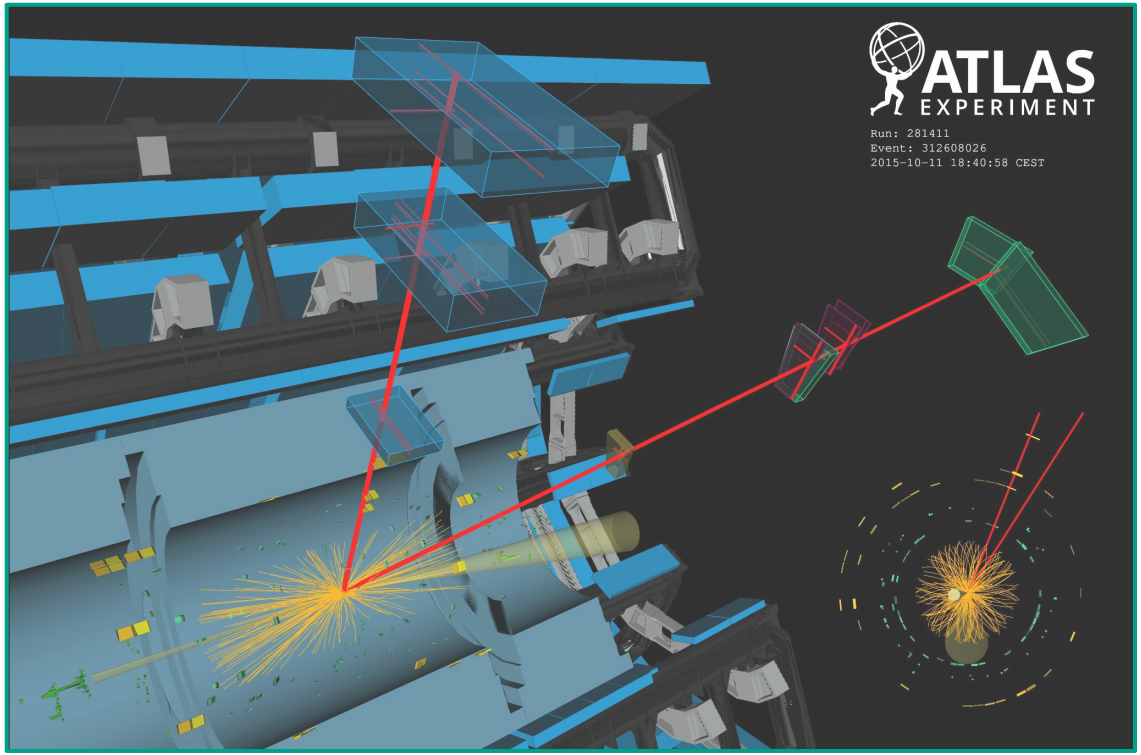
Monte Carlo simulations

DAOD_PHYSLITE

Under the CC0 waiver

"You can copy, modify, distribute and perform the work, even for commercial purposes, all without asking permission."

ATLAS **asks** for citation.



Released detector data and simulations

Detector data

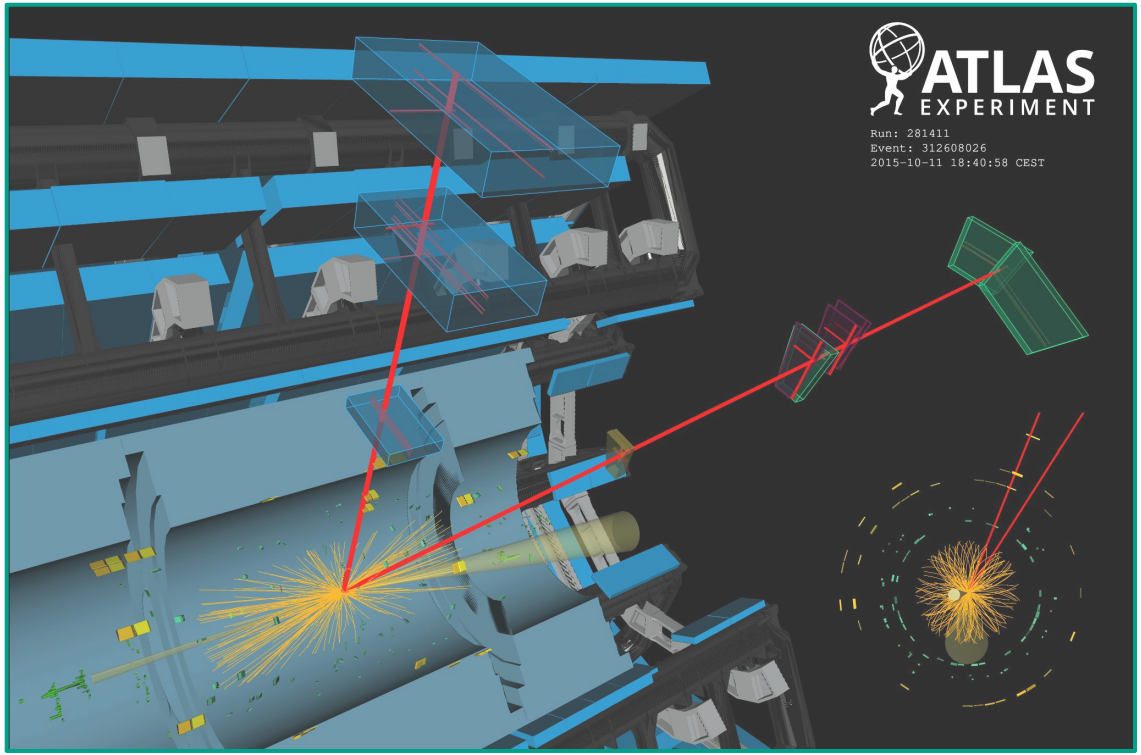
Monte Carlo simulations

DAOD_PHYSLITE

Under the CC0 waiver

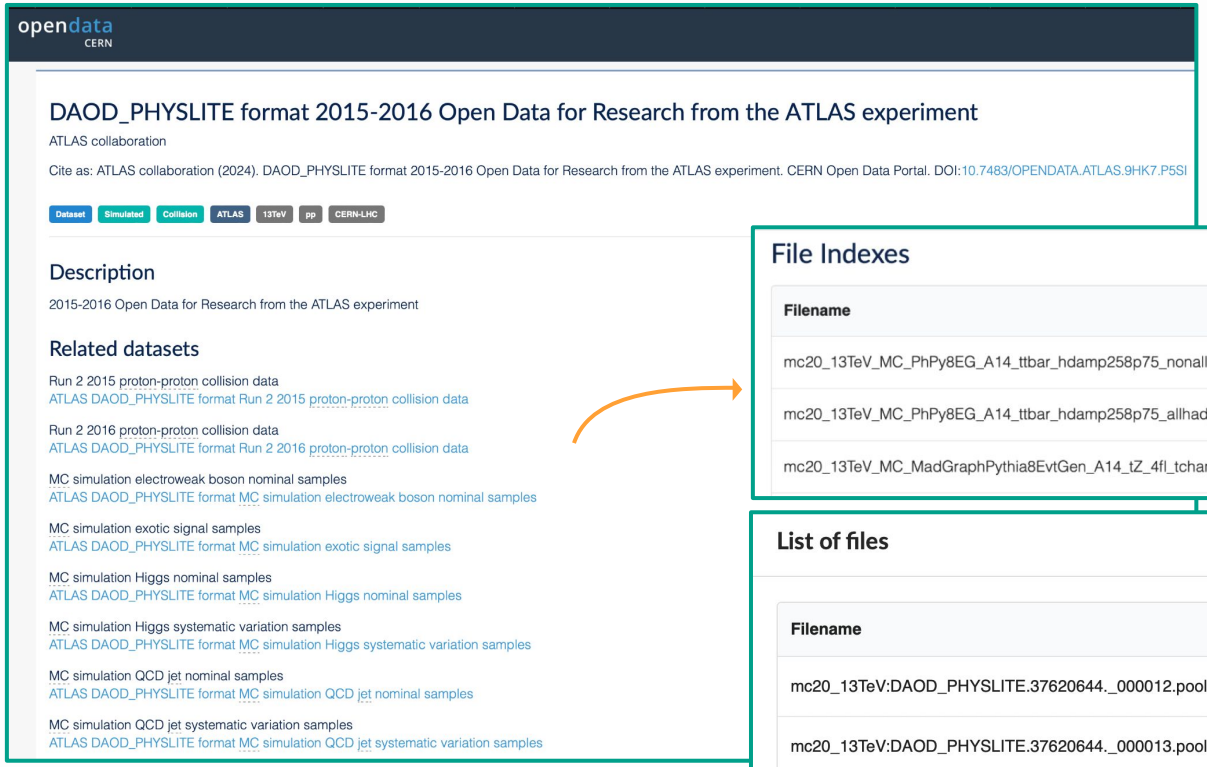
65 TB of data+MC.
Over 9 billion collisions

!



Where are the datasets?

The datasets are available in the CERN open data portal, where **they can be easily downloaded** on the website or using the cernopendata client.



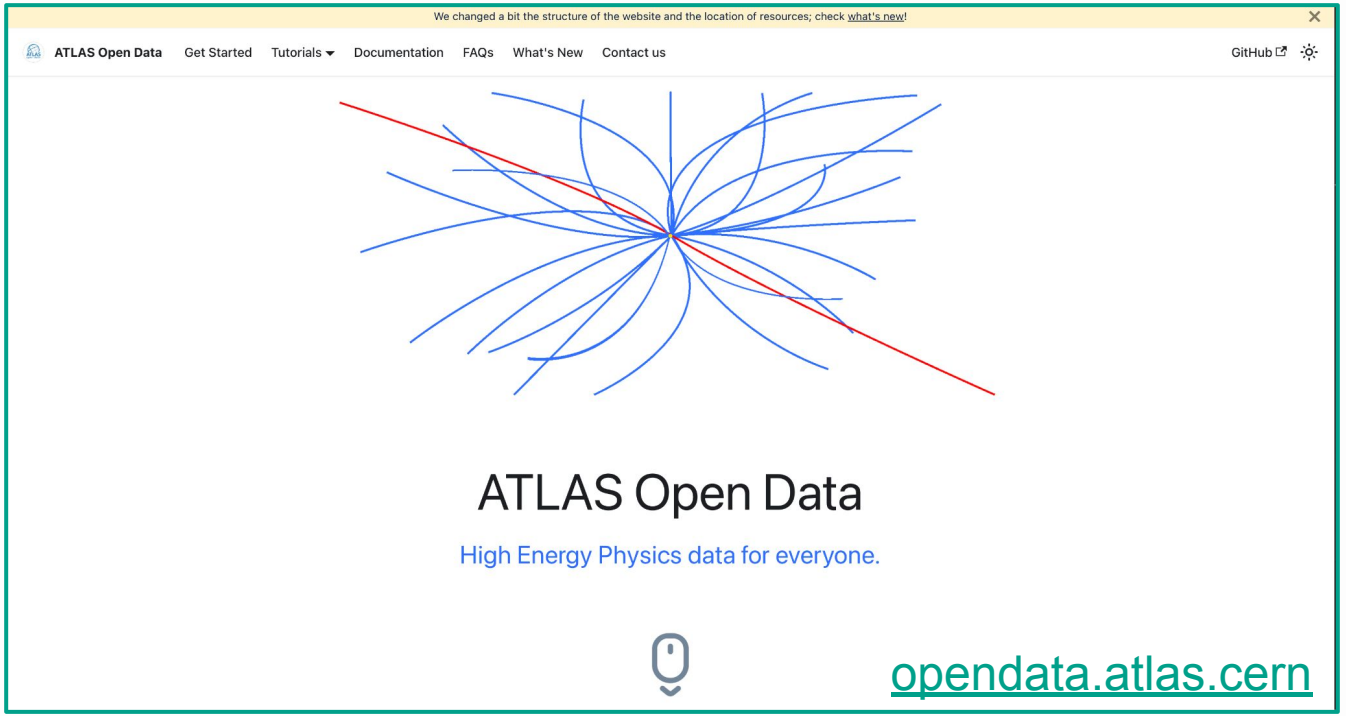
How can we help people use the data?

We have to make it easy for people to use the data!

- **General Documentation:** How will users get more information about the data?
- **Open Tools:** What tools are available for data access and analysis?
- **Tutorials:** How can users learn to use the data and tools?

Documentation

Ultimately, we want non-ATLAS members to be able to understand the procedure by which an analysis can be performed



Documentation about the datasets

The metadata for the Monte Carlo datasets, so that people have the complete information that is necessary for an analysis.

- Metadata
- Naming Convention
- General information

Metadata

Below, you will find the metadata for all the samples, which includes comprehensive information such as:

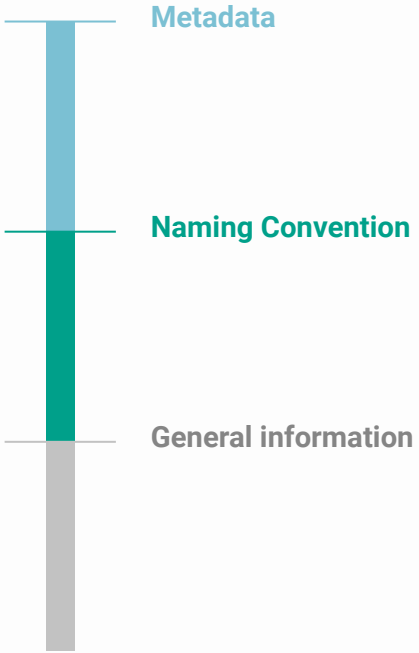
- **Dataset ID**: This is a unique identifier assigned to each dataset. It ensures that each dataset can be uniquely referenced and accessed.
- **Physics Short**: This is an abbreviated name that provides key information about the dataset. To know more about how to read them, check the subsection about naming convention.
- **Cross-Section (in pb)**: Represents the probability of a particular interaction occurring, measured in picobarns (pb). It is a fundamental parameter that helps understanding the likelihood of specific particle interactions under given conditions.
- **Filter Efficiency**: Measure of the effectiveness of the selection criteria applied to the data. It indicates the fraction of events

Search...

Dataset ID	Physics short	Cross section (pb)	Filter efficiency	K-fa
301204	Pythia8EvtGen_A14MSTW2008LO_Zprime_NoInt_ee_SSM3000	0.001762	1.0	1.0
301209	Pythia8EvtGen_A14MSTW2008LO_Zprime_NoInt_mumu_SSM3000	0.0017718	1.0	1.0

Documentation about the datasets

Documentation about the naming convention of the Monte Carlo simulations and the detector data, so that anyone can find the file they need.



Naming convention

In ATLAS, we use specific nomenclature for naming files to ensure they are easily identifiable. The naming conventions vary based on the type of file (Monte Carlo simulations or experimental data) to maintain clarity and organization.

Monte Carlo Simulations

The names for Monte Carlo simulations are composed by different substrings, separated by a dot:

```
campaign.dataset_id.short_description.production_step.data_format.processing_tags
```

Each part represents the following:

1. **campaign**: Indicates the MC simulation campaign and center of mass energy, when relevant. For example, for the released data from the MC20 campaign of proton-proton collisions at 13TeV of center of mass energy is "mc20_13TeV".
2. **dataset_id**: An 6 to 8 character numerical identifier, different for each dataset.
3. **short_description**: Indicates the simulation tools used and the physical process described by the dataset. Common simulation tools are Powheg, Pythia, Sherpa, among others. You can check the list of simulation tools or common abbreviations for more information about you can find on this substring.
4. **production_step**: The production step that generated the dataset. For the release data it is always "deriv" from derivation.
5. **data_format**: The dataset format. All the released data is in PHYSLITE format, so this substring is always "DAOD_PHYSLITE".
6. **processing_tags**: These tags indicate the configuration of the software used in each production step in the creation of the

Documentation about the datasets

Metadata

Naming Convention

General information

General information about how data is taken, how the Monte Carlo simulations are created, which tools are used, what considerations go into the physics objects.

ATLAS Data Collection

The ATLAS experiment at the Large Hadron Collider (LHC) is one of the most ambitious scientific experiments of our time, aiming to understand the fundamental particles and forces that shape our universe. The process of data collection, capturing the interactions of proton-proton (pp) and ion-ion collisions at very high energies, and analysis, understanding these collisions, is vital for this search. This section outlines how the ATLAS experiment works: how the experiment runs and how the collision events are processed for reconstruction, which is part of the process to eventually offer insight into the Standard Model and beyond.

ATLAS Runs and LHC Operations

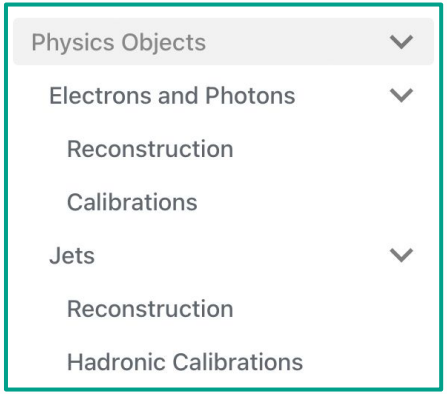
An ATLAS run is a coordinated data adquisition effort that coincides with the LHC fill cycle. It last tipically around 12 hours, capturing

Introduction to Monte Carlo Simulations

Monte Carlo (MC) simulations are computer-generated models that mimic particle collisions as measured by a detector. In high energy physics (HEP), they are used to model how theoretical particle interactions would manifest in the detector. These simulations take into account the complex physics of particle collisions, as well as the geometry and material properties of the detector. However, they also include approximations and assumptions about both the physics processes and the detector response.

The role of MC simulations in physics analysis can be broken down into the following aspects:

- **Event Selection**: Through MC simulations, researchers develop and test criteria for selecting events from the vast data generated by particle collisions experiments. Simulating various scenarios allows for the fine-tuning of event selection algorithms to isolate rare processes or signals indicative of new physics, while filtering out undesired or uninteresting events.
- **Background Estimation**: MC simulations allow analysers to model and understand background processes in detail, which helps

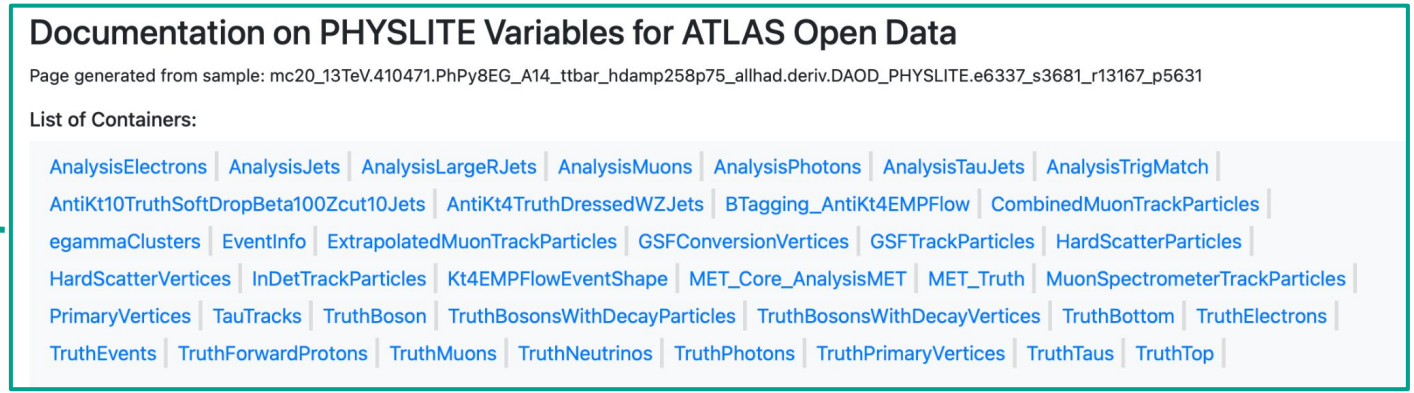


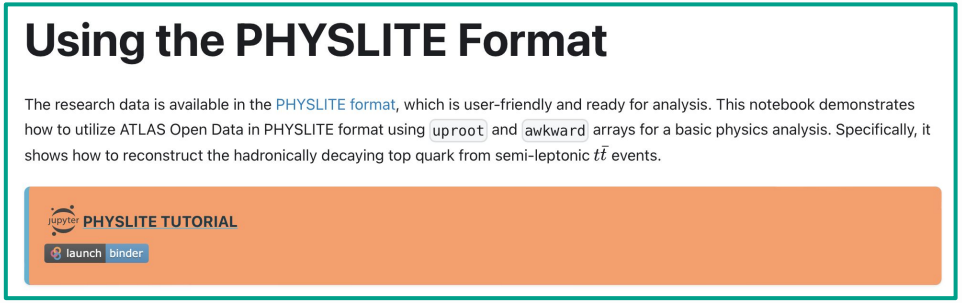
And much more!

Documentation about the format

The datasets are in **DAOD_PHYSLITE** format and have related documentation:

Variables documentation





PHYSLITE tutorial

PHYSLITE tutorial

Contains the basic information on how to use the file and perform a simple analysis.

- How to analyze PHYSLITE directly.
- Reconstruct the top quark:
  - Object selection.
  - Overlap removal.

```
In [33]: plt.hist(ak.flatten(mjjj(events.Jets) / GeV, axis=None), bins=100)
         plt.xlabel("Reconstructed Top Quark Mass (GeV)")
         plt.ylabel("Number of Events")
         plt.title("Distribution of Reconstructed Top Quark Mass")
         plt.axvline(172.76, color='r', linestyle='dashed', linewidth=2, label='Expected Top Quark Mass')
         plt.legend()
         plt.show()
```

Read PHYSLITE with uproot

We can open a `TFile` using `uproot.open`. To check the Ttree objects in the file we use the `.keys()` method.

```
]: print('TTree objects inside the ROOT file:')
   for ii in uproot.open(filename).keys():
       print('-',ii)
```

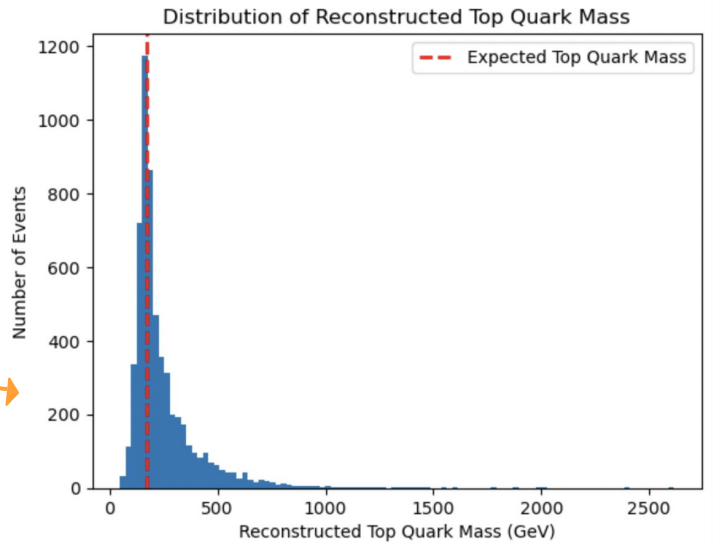
```
TTree objects inside the ROOT file:
- ##Params;3
```

Using the PHYSLITE Format

The research data is available in the PHYSLITE format, which is user-friendly and ready for analysis. This notebook demonstrates how to utilize ATLAS Open Data in PHYSLITE format using `uproot` and `awkward` arrays for a basic physics analysis. Specifically, it shows how to reconstruct the hadronically decaying top quark from semi-leptonic $t\bar{t}$ events.

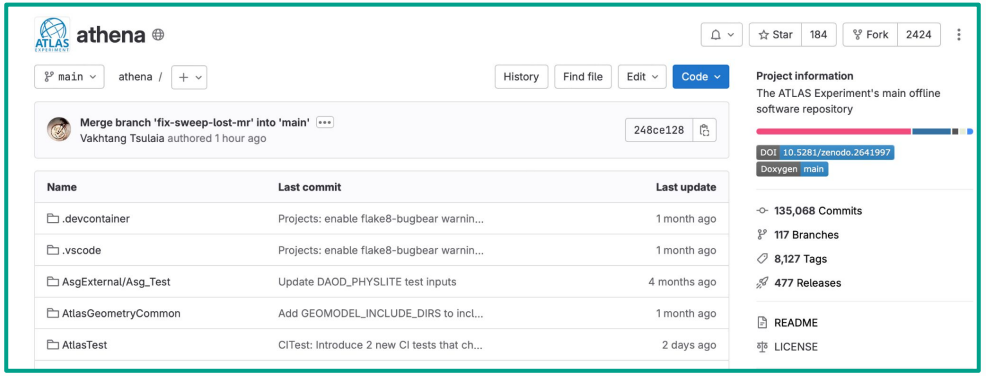
jupyter PHYSLITE TUTORIAL

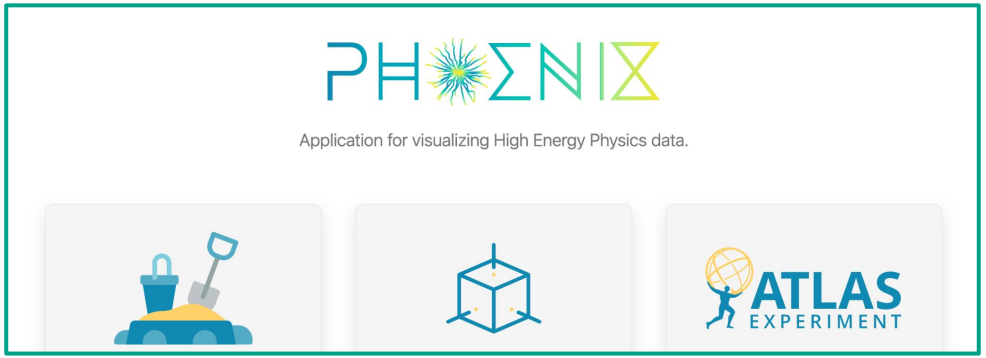
launch binder



Open tools

The software used by ATLAS has been public for many years. We also have an event visualization tool that is available for users to explore.





Projects

While the athena repository contains any and all code that could be built into an ATLAS software release, each release itself generally only consists of a consistent subset of the code base. Each particular build *flavour* is called a project and is steered from a particular subdirectory of the `Projects` directory.

The `Athena` project is a complete build of almost everything in the repository. When a particular Athena project is built the build result encodes the project name. Thus, independent of any release number, `AthSimulation` is built from different code than `AthAnalysisBase`.

The main projects are:

Project	Purpose
Athena	Reconstruction and Derivation production*
AthGeneration	For event generation
AthSimulation	For full Geant4 simulation
AthAnalysisBase	Athena based analysis
AnalysisBase	Non-athena ROOT based analysis
DetCommon	For reading trigger configuration when e.g. configuring L1 Hardware

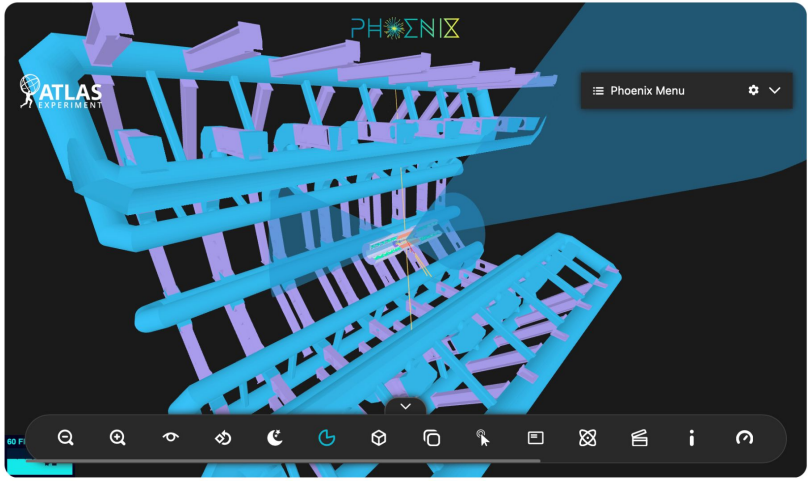
Tutorials

People can also use the already public Analysis Software Tutorial, with the standalone containers that we are providing. For event visualisation, we have a tutorial to transform PHYSLITE in the JSON needed for Phoenix.

Tutorials for Research › Phoenix for Event Visualisation

Phoenix for Event Visualisation

Visualizing an event inside the detector opens a path to a deeper understanding of the physical processes resulting from a collision. One commonly used tool for event visualisation is Phoenix, a framework that allows 3D visualisation of collision events inside the detector.



The event in the image above is an event for gluino production, which in SUSY is the supersymmetric partner of a gluon.

ATLAS Analysis Software Tutorial

ATLAS EXPERIMENT

Welcome to the **ATLAS Analysis Software Tutorial** pages.

This is the portal to the ATLAS analysis software tutorial held multiple times throughout the year.

This tutorial is aimed at (new and old) members of the ATLAS collaboration interested in learning the basics of ATLAS software and the latest physics analysis tools. There are several introductory lectures aimed at introducing the topics.

Tutorials for Research › Getting Started with Containers for Analysis

Getting Started with Containers for Analysis

Welcome to our tutorial on setting up containers for analysis! This guide will help you get started quickly and easily.

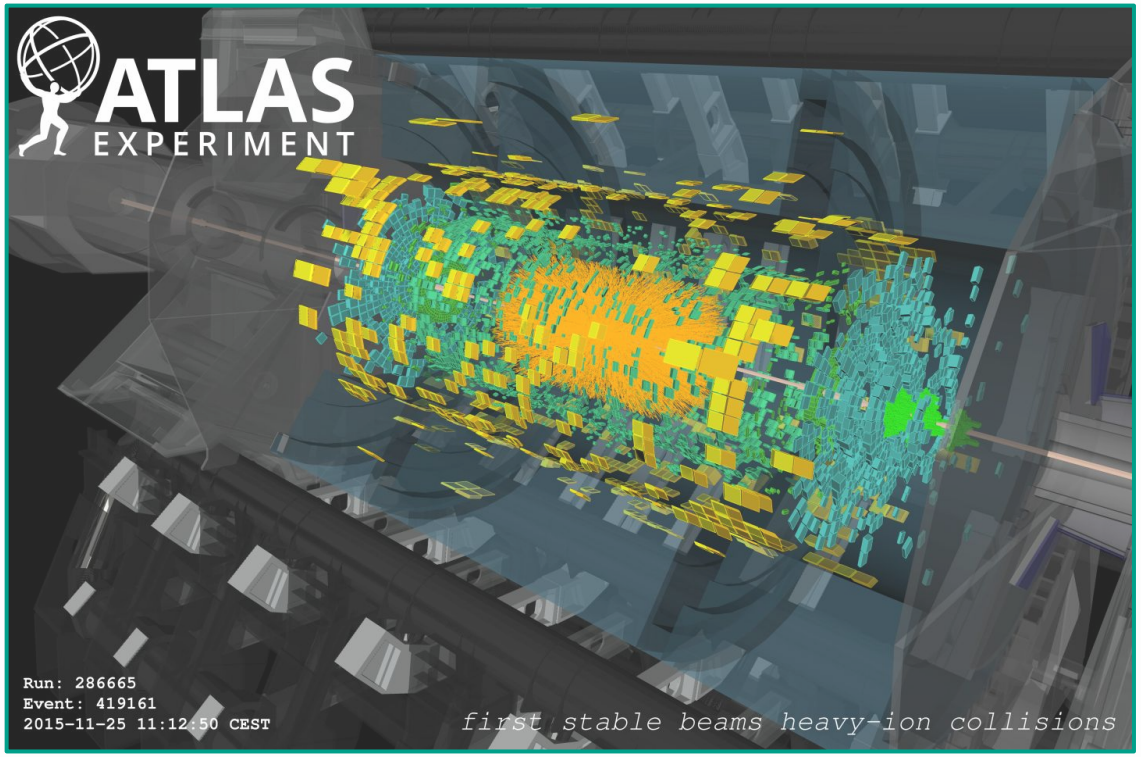
Future plans

Heavy Ions: lead-lead collisions from 2015 run.

More documentation: To provide a deeper understanding of the data.

More tutorials: So that approaching the data is easier.

New release of open data for education!

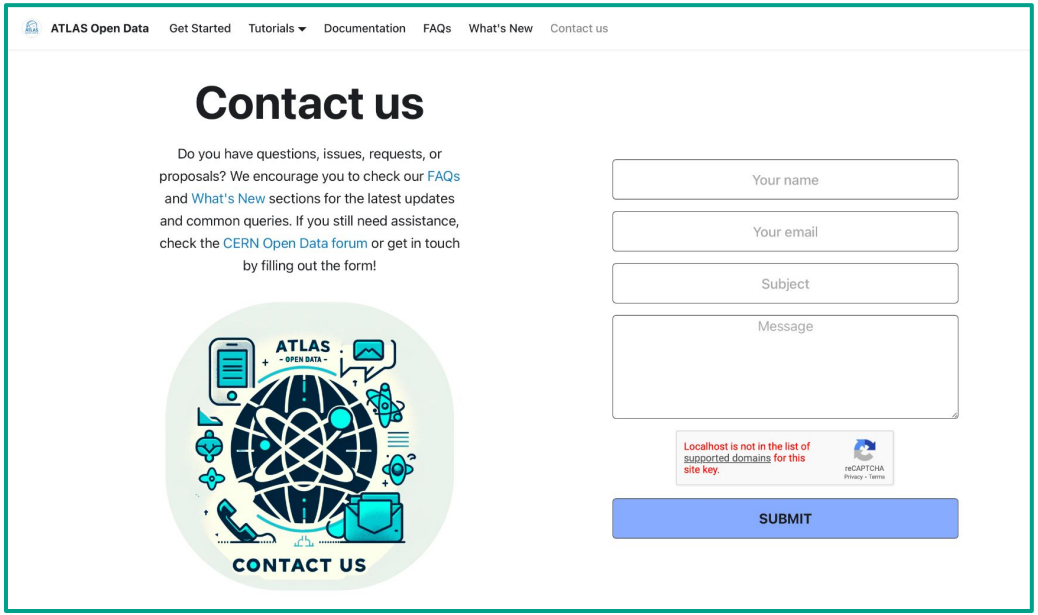


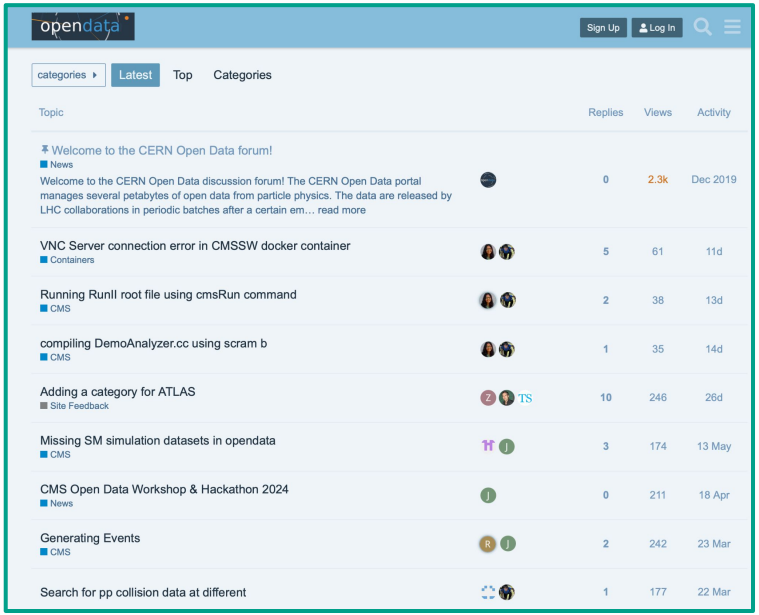
How can we help people use the data?

- **Self-Help Resources:** What resources are available for solo-troubleshooting?
- **User Support Mechanisms:** How can users get help?

Support

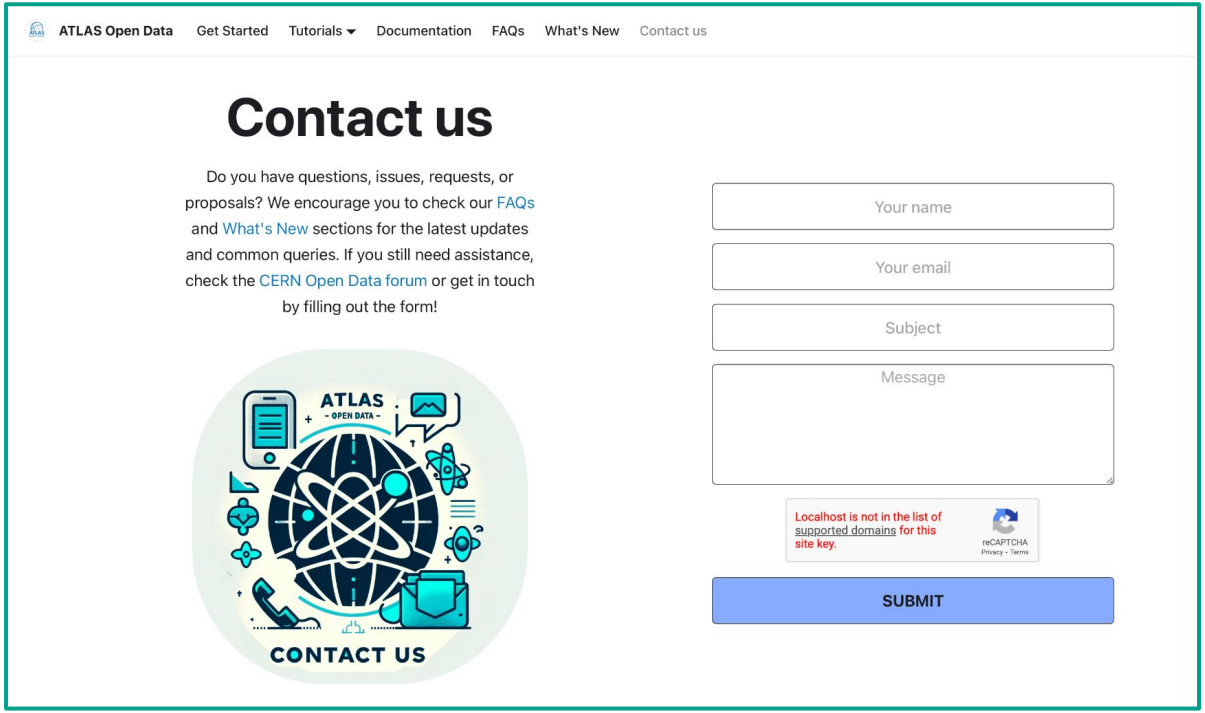
Support is based on voluntary effort inside the collaboration.





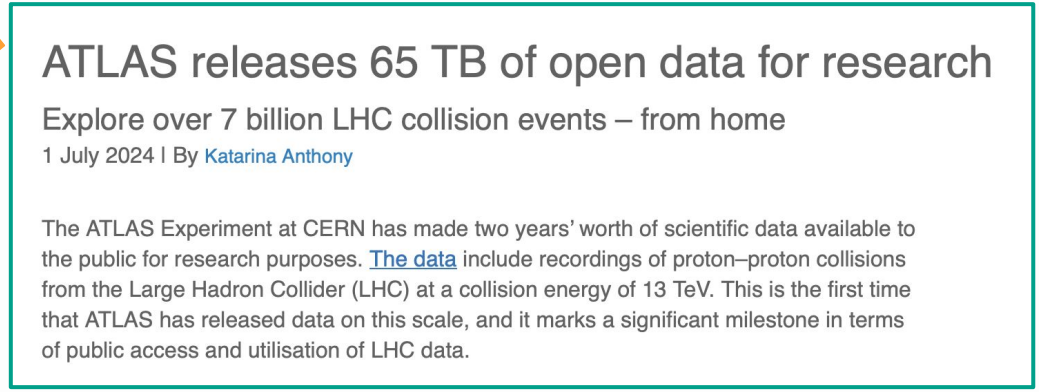
Contributing

We also encourage contributions! Did something cool with the data? Let us know.
Want even deeper analysis involvement? Come join us as a short term associate!



For more information

- Poster presentation on Friday!
- Visit the website.
- Check the briefing.
- Watch the tutorial on how to use the website.
- Public note on the last release of Open Data for Education.



Datasets

Related datasets

Run 2 2015 proton-proton collision data
ATLAS DAOD_PHYSLITE format Run 2 2015 proton-proton collision data

Run 2 2016 proton-proton collision data
ATLAS DAOD_PHYSLITE format Run 2 2016 proton-proton collision data

MC simulation electroweak boson nominal samples
ATLAS DAOD_PHYSLITE format MC simulation electroweak boson nominal samples

MC simulation exotic signal samples
ATLAS DAOD_PHYSLITE format MC simulation exotic signal samples

MC simulation Higgs nominal samples
ATLAS DAOD_PHYSLITE format MC simulation Higgs nominal samples

MC simulation Higgs systematic variation samples
ATLAS DAOD_PHYSLITE format MC simulation Higgs systematic variation samples

MC simulation QCD jet nominal samples
ATLAS DAOD_PHYSLITE format MC simulation QCD jet nominal samples

MC simulation QCD jet systematic variation samples
ATLAS DAOD_PHYSLITE format MC simulation QCD jet systematic variation samples

MC simulation SUSY signal samples
ATLAS DAOD_PHYSLITE format MC simulation SUSY signal samples

MC simulation top nominal samples
ATLAS DAOD_PHYSLITE format MC simulation top nominal samples

MC simulation top systematic variation samples
ATLAS DAOD_PHYSLITE format MC simulation top systematic variation samples

Example data

data15_13TeV_Run_00266904_file_index.txt
data15_13TeV_Run_00266919_file_index.txt
data15_13TeV_Run_00267073_file_index.txt
data15_13TeV_Run_00267148_file_index.txt
data15_13TeV_Run_00267152_file_index.txt

data16_13TeV_Run_00296939_file_index.txt
data16_13TeV_Run_00296942_file_index.txt
data16_13TeV_Run_00297041_file_index.txt
data16_13TeV_Run_00297170_file_index.txt
data16_13TeV_Run_00297447_file_index.txt

Datasets

Related datasets

Run 2 2015 proton-proton collision data
ATLAS DAOD_PHYSLITE format Run 2 2015 proton-proton collision data

Run 2 2016 proton-proton collision data
ATLAS DAOD_PHYSLITE format Run 2 2016 proton-proton collision data

MC simulation electroweak boson nominal samples
ATLAS DAOD_PHYSLITE format MC simulation electroweak boson nominal samples

MC simulation exotic signal samples
ATLAS DAOD_PHYSLITE format MC simulation exotic signal samples

MC simulation Higgs nominal samples
ATLAS DAOD_PHYSLITE format MC simulation Higgs nominal samples

MC simulation Higgs systematic variation samples
ATLAS DAOD_PHYSLITE format MC simulation Higgs systematic variation samples

MC simulation QCD jet nominal samples
ATLAS DAOD_PHYSLITE format MC simulation QCD jet nominal samples

MC simulation QCD jet systematic variation samples
ATLAS DAOD_PHYSLITE format MC simulation QCD jet systematic variation samples

MC simulation SUSY signal samples
ATLAS DAOD_PHYSLITE format MC simulation SUSY signal samples

MC simulation top nominal samples
ATLAS DAOD_PHYSLITE format MC simulation top nominal samples

MC simulation top systematic variation samples
ATLAS DAOD_PHYSLITE format MC simulation top systematic variation samples

Example MC

mc20_13TeV_MC_MadGraphPythia8EvtGen_A14NNPDF23_3top_SM_file_index.txt

mc20_13TeV_MC_MadGraphPythia8EvtGen_A14NNPDF23_ttbarWW_file_index.txt

mc20_13TeV_MC_aMcAtNloPythia8EvtGen_MEN30NLO_A14N23LO_ttW_file_index.txt

mc20_13TeV_MC_aMcAtNloPythia8EvtGen_MEN30NLO_A14N23LO_ttZnunu_file_index.txt

mc20_13TeV_MC_aMcAtNloPythia8EvtGen_MEN30NLO_A14N23LO_ttZqq_file_index.txt

mc20_13TeV_MC_PowhegPythia8EvtGen_NNLOPS_nnlo_30_ggH125_ZZ4l_file_index.txt

mc20_13TeV_MC_PowhegPythia8EvtGen_NNPDF3_AZNLO_ggZH125_HgamgamZinc_file_index.txt

mc20_13TeV_MC_PowhegPythia8EvtGen_NNPDF3_AZNLO_ggZH125_ZZ4lepZinc_file_index.txt

mc20_13TeV_MC_PowhegPythia8EvtGen_NNLOPS_nnlo_30_ggH125_mumu_file_index.txt

mc20_13TeV_MC_PowhegPythia8EvtGen_NNPDF3_AZNLO_ggZH125_Hmumu_Zinc_file_index.txt