PROCEEDINGS OF SCIENCE

Latest releases of ATLAS Open Data for Education and Research

Leonardo Toffolin[a,b,c] **on behalf of the ATLAS Collaboration**

[a] *Department of Physics, University of Trieste,*
Via Valerio 2, Trieste, Italy
[b] *INFN Trieste - Gruppo Collegato di Udine,*
Dipartimento Politecnico di Ingegneria ed Architettura, University of Udine, Udine, Italy
[c] *European Organisation for Nuclear Research (CERN)*
Esplanade des Particules, 1211 Geneva 23, Geneva, Switzerland

E-mail: leonardo.toffolin@cern.ch

The ATLAS Collaboration has recently released a large volume of data for use in research publications, and has now extended this effort with a new release focused on education. The proton–proton collision open datasets, together with matching simulated samples, are available in two formats: the light research format PHYSLITE, and a simplified version for educational use. In addition, the Collaboration has published for research purposes, for the first time at the LHC, a heavy-ion (Pb–Pb) dataset collected in 2015. To support these releases, the corresponding software and detailed documentation have been made public, addressing the needs of both newcomers to particle physics and experienced researchers. These proceedings present the data, metadata, software, and documentation, as well as the first experiences with non-ATLAS users and highlights of the new educational release.

13th Large Hadron Collider Physics Conference 2025,
5-9 May 2025
GIS National Taiwan University Convention Center, Taipei, Republic of China (Taiwan)

PoS(LHCP2025)178

1. The ATLAS Open Data mission

Open Data is an effective way for scientific collaborations to share knowledge and create a lasting legacy. It enables research to continue beyond the lifetime of experiments and supports education by providing students and educators with authentic datasets to analyse and learn from.

Following CERN's approved policy, which encourages LHC collaborations to publish experimental data for open public access [1], the ATLAS Collaboration [2] at the LHC [3] established its Data Access Policy, to provide the guidelines for a project which aims to grant open access to real ATLAS data by non-ATLAS members. The aim of the ATLAS Open Data project [4] is to provide data and a basic computing infrastructure and tools to the general public, in order to help users understand how a high energy physics analysis with real ATLAS data can be performed.

This project implements the FAIR principles (Findable, Accessible, Interoperable, Reusable) [5], and honours four main features:

- Accessibility: make the data and tools package accessible to everyone, keeping in mind the range of internet bandwidths, computer operating systems, mobile access, memory and RAM, and access to experts. It should still be analysable on older commodity hardware which might have a low technical entry barrier.
- Transferable skills: along with learning objectives related to particle physics, the Open Data releases can be used for capacity building in terms of programming, software and machine learning which are skills in demand inside and outside academia by providing examples of the downstream aspects of data analysis in C++ and Python.
- Usability and versatility: ensure that many different target audiences, with different backgrounds and skills are able to use the data and tools for a wide range of learning objectives and project goals.
- Impactful and wide reach: push communication campaigns and prepare and deliver online and in person events.

The project embeds two philosophies. The education and outreach philosophy targets high school, undergraduate and graduate students, as well as teachers, lecturers and the general public. Goals are to provide simplified data from the ATLAS Experiment, tools and techniques inherited from the collider physics expertise. Alongside with it, the ATLAS Open Data programme presents also a research philosophy, with the idea of providing researchers with high-quality data recorded by the ATLAS detector, enabling them to conduct state-of-art analyses in particle physics.

2. ATLAS Open datasets

The ATLAS Open Data releases feature a collection of datasets collected by the ATLAS detector during the various data acquisition runs at the LHC. These datasets include 8 TeV and 13 TeV proton–proton collision data, with both real collision events and Monte Carlo simulations, along with dataset variations to estimate systematic uncertainties. The datasets are curated with calibrated and simplified information about reconstructed physics objects, making them accessible

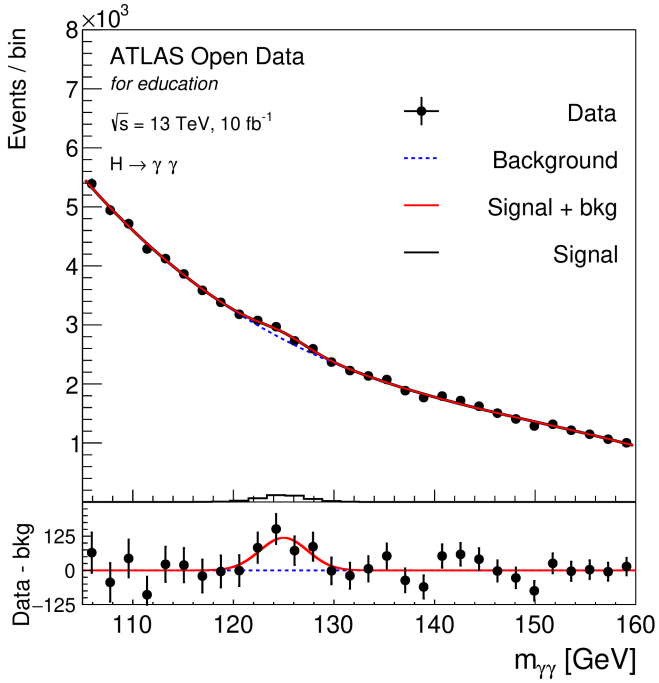

Figure 1: Diphoton invariant-mass spectrum for a $H \to \gamma\gamma$ example analysis performed with the second ATLAS Open Data Release for education, published in 2020. The label on the top-left of the plot depicts the education usage of those data. The solid red curve shows the fitted signal-plus-background model when the Higgs boson mass is constrained to be 125 GeV. The background component of the fit is shown with the dotted blue curve. The signal component of the fit is shown with a solid black curve. The bottom plot shows the residuals between the data and the background component of the fitted model. Figure taken from Ref. [8].

and manageable for a wide range of users. The inclusion of the "for education" label in plot titles ensures clarity of purpose, distinguishing these datasets as tools tailored for learning and exploration, such as the one shown in Fig. 1.

The first ATLAS Open Data release corresponds to approximately 1 fb^{-1} of LHC proton–proton data recorded in 2012 at a centre-of-mass energy of 8 TeV, called the ATLAS Open Data 2016 dataset [6]. It was launched in 2016 together with an analysis framework and seven different examples, written in Python language. The dataset corresponds to approximately 15 million events. This is part of the data that allowed ATLAS Collaboration to discover the Higgs boson [7]. For this reason, this fraction of the 2012 data has an important scientific, educational and historical value. Simulated data are also made available for various Standard Model and new physics signal models, for comparison with LHC data.

The second release comprises approximately 10 fb^{-1} of LHC pp data recorded in 2016 at a centre-of-mass energy of 13 TeV, called the ATLAS Open Data 2020 dataset [8]. It corresponds to approximately 61 runs from the first four periods of the 2016 pp data-taking, with a total amount of about 270 million proton–proton collisions. Similar to the previous release, it was launched in 2020 together with an analysis framework.

The third and upcoming release, foreseen in Fall 2025 and currently being documented, is

going to offer 36 fb^{-1} of pp data, further enhancing the scope for educational analysis. The data are provided as a flat ROOT NTuple format [9].

The education and outreach releases allow for a superficial comparison with the ATLAS published results, however do not qualify for complete analysis replications due to the missing complete event information, which is present instead on the Open Data for Research release, including for instance full systematic uncertainties. Moreover, it cannot be used to publish journal-level analysis results. The limited information provided in the datasets therefore sets the desired bounds of usage.

Alongside with these releases, an additional release for research purposes has been published by the ATLAS Collaboration between 2024 and 2025. It comprises the full set of pp data from 2015 and 2016, around 45 TB, and a heavy-ion (Pb-Pb) 221-million collisions dataset from 2015 [10]. It represents the first LHC Run-2 heavy-ion open dataset ever published.

All releases are accompanied by datasets hosted on the CERN Open Data Portal [11] and documentation, tutorials, and supporting resources accessible on the ATLAS Open Data website [4]. These materials guide learners in analysing particle physics data in educational environments, offering a hands-on experience that emulates real-world research processes.

3. Software and tools

In the ATLAS Open Data effort, a suite of software and tools [12] is provided to support online and offline clients with different levels of analysis complexity, both in online Jupyter Notebooks [13] and in Docker containers and downloadable virtual machines [14] preloaded with all the software needed for the analysis.

Among these tools is the Histogram Analyser [15], a web-based platform designed to quickly and intuitively interact with cut-based data analysis. This tool allows users to visualise datasets through interactive histograms, focusing on how physicists distinguish between different physics processes. By adjusting the range of data for different variables, users can isolate specific signals, such as Higgs boson events, from background processes. This hands-on approach is particularly useful for building expertise in interpreting data distributions and optimising selection criteria for advanced analyses.

The Jupyter notebooks [13] accompanying the Histograms Analyser are designed for interactive data analysis, cover a wide range of topics, from Standard Model (SM) analyses to Beyond the Standard Model (BSM) physics. As an example, with the help of the `ROOTbook` technology the diphoton invariant mass spectrum can be easily extracted from data, in order to give users the possibility to look at one of the Higgs boson golden channels, as reported in Fig. 2.

The notebooks are accessible directly through a web browser, and target users with with varying levels of expertise. They also support multiple programming frameworks, including Python, C++, RDataFrame, Scikit HEP, and Coffea, ensuring that users can work in a flexible environment and have examples in the environment with which they are most familiar. Beyond physics applications, these resources include tools for training and statistical analyses that extend to other fields, such as data science and related curricula. This versatility makes the ATLAS Open Data resources valuable for broader educational purposes. A series of YouTube tutorials [16] is also provided, offering step-by-step guidance on using the tools, conducting analyses, and understanding the datasets.

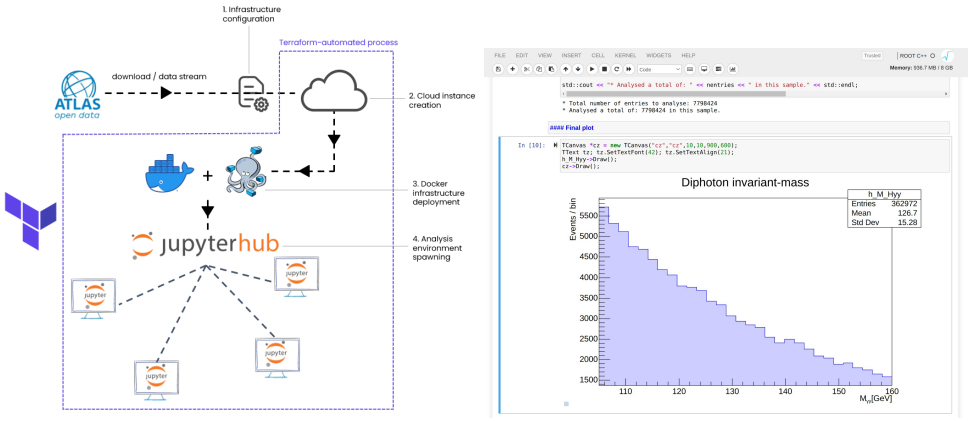

Figure 2: Schematic view of the Jupyter-based infrastructure (on the left); diphoton invariant mass distribution from the 13 TeV Open Data release dataset (on the right), obtained by running one of the Jupyter notebook tutorial released alongside the datasets. Taken from [8].

Several notebooks with analysis examples and an interface to launch the framework mentioned above, are available using SWAN (**S**ervice for **W**eb-based **AN**alysis), Binder and Colab executable platforms.

The usage of Jupyter notebooks, without downloading anything, can be a good option if a solid internet connection is available; local virtual machines with locally downloaded data and analysis code can be used in a completely offline environment as well; Docker containers [17] can combine the two approaches, presenting a notebook interface on top of a locally running web service.

4. Summary and future plans

ATLAS Open Data initiatives have been serving educational purposes since 2016, providing computing resources, tutorials, and documentation that support the datasets. These materials are widely used in training sessions, workshops, and masterclasses, helping to spread HEP-based learning across disciplines. A new 13 TeV Open Data for education release is coming soon, with updated data formats and higher integrated luminosities, offering opportunities to adopt new data formats and expand the available dataset to higher integrated luminosities. In parallel, the recent Open Data release for research opens up further opportunities. Finding synergies between the education and research efforts will be important for the future. To sustain this bottom-up, community-driven approach, monitoring and evaluation are essential. Finally, the Open Data community is constantly expanding, with new examples and contributions being collected from around the world. This collaborative approach ensures that the ATLAS Open Data initiative continues to evolve and serve a wide range of needs and use cases.

References

[1] Data Preservation and Long Term Analysis in High Energy Physics (DPHEP) Study Group, *CERN Open Data Policy for the LHC Experiments*, CERN-OPEN-2020-013, https://cds.cern.ch/record/2745133.

[2] ATLAS Collaboration, *The ATLAS Experiment at the CERN Large Hadron Collider*, *JINST* **3** (2008) S08003.

[3] O. S. Brüning *et al.*, *LHC Design Report*, CERN Yellow Reports; Monographs, CERN, Geneva, 2004, doi:10.5170/CERN-2004-003-V-1.

[4] *ATLAS Open Data website*, https://opendata.atlas.cern.

[5] M. Wilkinson, M. Dumontier, I. Aalbersberg *et al.*, *The FAIR Guiding Principles for scientific data management and stewardship*, *Sci. Data* **3** (2016) 160018.

[6] ATLAS Collaboration, *Review of ATLAS Open Data 8 TeV datasets, tools and activities*, https://cds.cern.ch/record/2624572/.

[7] ATLAS Collaboration, *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC*, *Phys. Lett. B* **716** (2012) 1-29.

[8] ATLAS Collaboration, *Review of the 13 TeV ATLAS Open Data release*, https://cds.cern.ch/record/2707171.

[9] ROOT Collaboration, *ROOT - An Object Oriented Data Analysis Framework*, *Nucl. Inst. & Meth. in Phys. Res. A* **389** (1997) 81–86.

[10] ATLAS Collaboration, *DAOD_HION14 format 2015 Pb-Pb Open Data for Research from the ATLAS experiment*, 2020, CERN Open Data Portal, doi:10.7483/OPENDATA.ATLAS.IKCT.HH2K.

[11] *CERN Open Data portal*, https://opendata.cern.ch.

[12] *ATLAS outreach data and tools repository*, https://github.com/atlas-outreach-data-tools.

[13] ATLAS Open Data 13 TeV documentation, 13 TeV Open Data Jupyter Notebooks, https://opendata.atlas.cern/docs/13TeVDoc/13tutorial.

[14] ATLAS Open Data 13 TeV documentation, Virtual machines, https://opendata.atlas.cern/docs/13TeVDoc/enviroments/hybrid.

[15] *The ATLAS Open Data Histogram Analyser webpage*, https://opendata.atlas.cern/docs/webapps/histanalyser.

[16] *The ATLAS Experiment YouTube channel*, https://www.youtube.com/@ATLASExperiment.

[17] *ATLAS Open Data, Docker containers*, https://hub.docker.com/r/atlasopendata/root_notebook.

PoS(LHCP2025)178