PROCEEDINGS OF SCIENCE

The First Release of ATLAS Open Data for Research

Mariana Vivas Albornoz* **on behalf of the ATLAS Collaboration**

University of Massachusetts, Amherst,
Amherst, MA 01003, USA

E-mail: mariana.vivas.albornoz@cern.ch

The ATLAS Collaboration released for the first time open data for research. This release comprises 65 TB of proton–proton collision data from the Large Hadron Collider in 2015 and 2016, and over 300 samples of Standard Model and Beyond the Standard Model Monte Carlo simulations, in a light data format. With this release, the collaboration invites the worldwide community of scientists to participate in High Energy Physics research. To make this possible and accessible, the release includes documentation, open tools and tutorials, as well as support channels. The future plans of this project include the release of all lead–lead collisions from 2015 run in a light format, along with new documentation and tutorials.

ATL-OREACH-PROC-2024-005
24 September 2024

42nd International Conference on High Energy Physics (ICHEP2024)
18-24 July 2024
Prague, Czech Republic

*Speaker

1. Why Open Data for Research?

The CERN Open Data Policy for LHC Experiments [1], endorsed by the four main LHC collaborations, commits to publicly releasing calibrated reconstructed data with the level of detail necessary for research. Every LHC experiment is to release 25% of their data five years after the end of each run. Following the conclusion of Run 2, the ATLAS collaboration released these data.

The goal of this release is to facilitate broader access to scientific data, enabling researchers to test hypotheses, develop new tools and techniques, and contribute to current advancements in particle physics. By sharing these data, the hope is to encourage further innovation and promote collaboration in the global scientific community.

2. The ATLAS Open Data Project

The ATLAS Open Data project started as an educational endeavor in 2016, with the first release of open data for education [2]. Two distinct datasets were released, both collected by the ATLAS detector from 8 TeV proton-proton collisions in 2012: one with 2 fb^{-1} provided in XML format, and another with 1 fb^{-1} available in ROOT ntuple format.

The project was expanded in 2020 with a new data release [3]. This release includes data collected by the ATLAS detector at 13 TeV during the year 2016 and corresponds to an integrated luminosity of 10 fb^{-1}. The proton–proton collision data is accompanied by a set of Monte Carlo (MC) simulated samples describing several processes which are used to model the expected distributions of different signal and background events. The resulting format is a ROOT ntuple with more than 80 branches, almost doubling the 8 TeV ROOT ntuples, with 46 branches.

The central hub for the open data is the ATLAS Open Data website [4], where documentation, example analyses, and tools can be found. The open data and tools were designed with advanced high school students, university undergraduates, and master's students in mind. Two user paths were defined to aid the user experience: "Quick start", for those using the data for the first time or those who do not want to spend time configuring tools, and "Deep Dive", for students that want to use the tools locally and more extensively or teachers that want to use the data in class.

The open data for education has been widely used in schools, universities, events, and by individual learners worldwide [5].

3. ATLAS Open Data for Research

The ATLAS Open Data initiative aims to broaden access to the data generated by the ATLAS experiment, extending its use beyond traditional educational settings to support research activities. By releasing these data, the initiative expands beyond university-level education, inviting researchers to explore and innovate using the new resources.

To facilitate this, a new user path was introduced, the "Researchers' Toolkit", tailored for individuals interested in conducting complex analyses or experimenting with new technologies. This new pathway complements the existing educational resources, ensuring that users at various levels of expertise can effectively engage with the data.

3.1 The data

The data selection process included considerations related to:

- **Data Availability**: Ensuring there is sufficient data for a variety of use cases.
- **Data Format**: Determining the appropriate format for the data. A format that is easy to use but complex enough for authentic analysis.
- **Licensing**: Establishing the regulations that will govern data usage.
- **Accessibility**: Addressing how users can access the data easily.

Following these considerations, the collaboration released all the detector data of proton–proton collisions during the 2015 and 2016 runs, from the main physics trigger stream, with a combined luminosity of 36 fb^{-1}, ensuring enough data for a comprehensive analysis. Additionally, over 300 MC simulation samples are provided, including Standard Model nominal samples and alternatives for systematic variations. These samples are categorized as follows:

- *Higgs boson*: Any process that includes the Higgs boson. Samples are placed in this category even if they contain weak bosons, top quarks, or jets. The Higgs boson takes precedence over other classifications.
- *Top quark*: Processes containing top quarks, but excluding any that already fall under the Higgs boson category.
- *Weak boson*: Processes involving Z or W bosons, excluding those categorized as either Higgs boson or top quark processes.
- *Jets*: Processes that primarily involve the creation of jets, and which don't fit into the previous categories.

Additionally, Beyond the Standard Model signal samples are provided, categorized in:

- *SUSY*: Signals from super-symmetric models.
- *Exotics*: Other non-Standard Model signals.

The data were released in the smallest format used internally for analyses [6], known as PHYSLITE [7]. This format is compact, containing already calibrated and pre-selected objects and high-level information, corresponding to 10 kB per event for data and 12 kB for MC simulation. An important advantage: it can be analyzed directly, without the need for further processing, decreasing users' storage needs. All these characteristics guarantee a format that is easy to use and suitable for complex analyses.

The detector data and MC simulation samples were released under a Creative Commons CC0 waiver [8], which states that the work can be copied, modified, distributed, and performed, even for commercial purposes, without requiring permission. However, ATLAS requests that users provide proper citation when using the data, as outlined in the accompanying documentation.

In total, the collaboration released 65 TB of detector data and MC simulation, which is over 9 billion collisions. The datasets can be found in the CERN Open Data portal [9], where they can be easily downloaded or accessed using the cernopendata client, ensuring accessibility for the users.

3.2 The documentation

To maximize the usability of the released data, the need to provide comprehensive documentation was recognized. The focus was on providing users with clear guidance to explore and utilize the open data. Three main points were emphasized:

- **General Documentation**: Ensuring there is introductory information available about the data and the experiment.
- **Open Tools**: Guaranteeing the accessibility of tools for accessing and analysing the data.
- **Tutorials**: Addressing how users can learn to use the available tools.

As mentioned earlier, all documentation is hosted on the ATLAS Open Data website. By centralizing these resources, this approach ensures a seamless transition between educational resources and research tools, maintaining user engagement as they advance and understand more about the data. The ultimate goal is for non-ATLAS members to comprehend the procedure by which an analysis is performed within the ATLAS collaboration. Internally, this approach not only avoids duplicated effort but also fosters continuous improvement and collaboration within both sections of the project.

The documentation that accompanied the release of data for research can be organized in three categories: documentation about *the data*, about *the format*, and about *open tools*.

The information *about the data* includes general information about how data are taken, how the MC simulations are created, which tools are used, and what considerations go into the definition of physics objects. For the use of the MC simulation, metadata details about the simulation are included: the numerical ID of the datasets, the short version of the name of the sample, cross section, filter efficiency, K-factor, number of events, sum of weights, sum of weights squared, generators, keywords and a link to the code that was used to generate the sample. This ensures that the user have the whole picture of the datasets. All this information is summarized in a searchable table, for accessibility.

In addition, information about the naming conventions of MC simulations is provided, aiding users in locating the data they need. This information is also valuable for new ATLAS members, exemplifying the benefit for the collaboration internally that good documentation for external users can also provide.

Furthermore, detailed information is provided about all the variables included in a PHYSLITE file, including the variable type and a brief description. This resource is similarly beneficial for new ATLAS members, reinforcing the advantages of well-structured documentation.

To aid with the use of the format, a PHYSLITE tutorial is provided. It contains the basic information about how to use the PHYSLITE file and perform a simple analysis, through a basic $t\bar{t}$ analysis. It shows how to read a PHYSLITE directly using Python and how to reconstruct the top quark by doing object selection and overlap removal.

In the *open tools* category there are two main tools.

The main tools for analysis are contained in the Athena [10] repository. This repository holds the main offline software used by ATLAS and has been public for many years. To understand how

to use this software for analysis, the user can follow the already-public Analysis Software Tutorial, with the provided standalone containers. This tutorial is widely used inside the collaboration.

Additionally, information for using Phoenix [11], an open-source event visualization tool, is provided in the form of a tutorial that guides users through transforming PHYSLITE data into the JSON format required for Phoenix. This enables users to create their own event displays using the public data.

3.3 User support channels

Providing sufficient data that is easy to use, with robust documentation, is the main way to help the user utilize the data. However, it is recognized that users may still require additional support. Two key support mechanisms were identified:

- **Self-Help Resources**: Resources for solo-troubleshooting.
- **User Support Mechanisms**: Other ways a user can get help.

To facilitate user support, the CERN open data forum serves as the primary contact channel. Here, users can ask questions, report issues, and engage in peer support, with voluntary assistance from ATLAS scientists.

3.4 Collaborating

Collaboration through the CERN open data forum is encouraged. Users can share interesting materials they have produced using our data and tools.

For people interested in deeper involvement with the experiment, they can join the collaboration as a short-term associate [12]. Non-ATLAS scientists who gain this status can be involved in the analysis work, which may include access to internal ATLAS collision data and MC simulations. This work generally culminates in a public result, such as a published paper or public note.

3.5 Future plans

The future of this project includes the release of heavy ion data, specifically lead–lead collisions from the 2015 run. These data require some new documentation in order to make heavy ion data analysis more accessible to the users. Additional documentation and tutorials for the proton–proton collision data will be provided as needed to ensure that accessing the data becomes easier.

For educational purposes, a new subset of the research data in a simplified format will be released, allowing analyses with 36 fb^{-1}[13]. This release is designed to be more accessible for students and educators. In addition, more example analyses and tools will be developed, along with detailed tutorials, facilitating for educational institutions to incorporate real-world particle physics data into their curricula.

References

[1] *CERN Open Data Policy for the LHC Experiments*, DOI, 2020.

[2] ATLAS collaboration, *Review of ATLAS Open Data 8 TeV datasets, tools and activities*, https://cds.cern.ch/record/2624572, 2018.

[3] ATLAS collaboration, *Review of the 13 TeV ATLAS Open Data release*, https://cds.cern.ch/record/2707171, 2020.

[4] ATLAS Collaboration, "ATLAS Open Data." https://opendata.atlas.cern. Last accessed 29 August 2024.

[5] ATLAS collaboration, M.O. Evans, *ATLAS Open Data – a genuinely collaborative approach for the creation of educational resources*, https://cds.cern.ch/record/2783039, 2021.

[6] ATLAS Collaboration, "ATLAS measures ZZ production using Run-3 data and a new slim data format." https://atlas.cern/Updates/Physics-Briefing/ZZ-PHYSLITE. Last accessed 29 August 2024.

[7] ATLAS collaboration, *PHYSLITE - A new reduced common data format for ATLAS*, Tech. Rep. ATL-SOFT-PROC-2023-029, CERN, Geneva (2023).

[8] Creative Commons, "CC0 1.0 UNIVERSAL." https://creativecommons.org/publicdomain/zero/1.0/. Last accessed 28 August 2024.

[9] ATLAS Collaboration, "DAOD_PHYSLITE format 2015-2016 Open Data for Research from the ATLAS experiment." DOI. Last accessed 29 August 2024.

[10] ATLAS Collaboration, *Software and computing for Run 3 of the ATLAS experiment at the LHC*, https://arxiv.org/abs/2404.06335, 2024.

[11] F. Ali, E. Moyse, M.H. Khan, E.C. Labra, A. Pappas, J. Smiesko et al., *HSF/phoenix: v2.14.1*, DOI, May, 2023.

[12] ATLAS Collaboration, "How to collaborate with ATLAS." https://atlas.cern/Discover/Collaboration/External-Collaboration. Last accessed 29 August 2024.

[13] M. Vivas Albornoz, "Open Data at ATLAS: Bringing TeV collisions to the World." Proceedings of the 42nd International Conference on High Energy Physics (ICHEP2024), 2024.