EPJ Web of Conferences **245**, 08026 (2020) https://doi.org/10.1051/epjconf/202024508026
CHEP 2019

The release of the 13 TeV ATLAS Open Data: using open education resources effectively

Leonid SERKIN[1,*] on behalf of the ATLAS Collaboration

[1]INFN Gruppo Collegato di Udine, Sezione di Trieste, Udine and ICTP, Trieste
Strada Costiera 11, Trieste 34151, Italy

Abstract. The ATLAS Collaboration is releasing a new set of proton–proton collision data to the public for educational purposes. The data was collected by the ATLAS detector at the Large Hadron Collider at a centre-of-mass energy $\sqrt{s}$ = 13 TeV during the year 2016 and corresponds to an integrated luminosity of 10 fb^{-1}. This dataset is accompanied by simulated events describing several Standard Model processes, as well as hypothetical Beyond Standard Model signal processes. Associated computing tools are provided to make the analysis of the dataset easily accessible. In the following, we summarise the properties of the 13 TeV ATLAS Open Data set and the available analysis tools. Several examples intended as a starting point for further analysis work by users are shown. The general aim of the dataset and tools released is to provide user-friendly and straightforward interactive interfaces to replicate the procedures used by high-energy-physics researchers and enable users to experience the analysis of particle-physics data in educational environments.

1 Introduction

The aim of ATLAS Open Data is to provide data and tools to high school, undergraduate and graduate students, as well as teachers and lecturers, to help educate and give them experience them in physics analysis techniques used in experimental particle physics. By sharing data collected by the ATLAS experiment [1], this project aims to generate interest and enthusiasm for fundamental research, inspiring physicists of the future.

Since the endorsement of the ATLAS Data Access Policy [2] in June 2014, a number of groups within the ATLAS Collaboration as well as external collaborators and users reported [3] a large range of activities based on the 1 fb^{-1} of proton–proton (pp) collisions at $\sqrt{s}$ = 8 TeV, referred to as "8 TeV ATLAS Open Data", and released in 2016 [4].

Given this wide interest, a new set of pp collision data at $\sqrt{s}$ = 13 TeV has been released by the ATLAS Collaboration to the public for educational purposes [5]. The dataset corresponds to an integrated luminosity of 10 fb^{-1} recorded by the ATLAS

*e-mail: Leonid.Serkin@cern.ch

detector at the LHC in 2016. The dataset is accompanied by Monte Carlo (MC) simulation samples describing several Standard Model (SM) and beyond the Standard Model (BSM) processes.

These datasets, referred to as "13 TeV ATLAS Open Data", are intended to provide the means for doing hands-on particle-physics exercises in the context of higher education, for example laboratory courses or introductory exercises for undergraduate and graduate students. The released data may also prove beneficial for the production of teaching materials, lectures and public talks. Furthermore, it may be used for further development of analysis methods and techniques, such as Higgs and particle-tracking machine learning (ML) challenges.

2 Overview of 13 TeV ATLAS Open Data

The 13 TeV ATLAS Open Data events belong to the first four periods of the 2016 pp data-taking and contain approximately 270 million collision events. Only events for which all relevant subsystems were operational are considered. After applying quality criteria for the beam, detector and data, the publicly released dataset corresponds to an integrated luminosity of 10 fb^{-1}. The events from pp collisions are accompanied by MC simulation samples describing several SM processes.

In the following, a summary of the main features of the dataset is presented:

- The released samples are provided in a simplified data format, reducing the information content of the original data-analysis format used within the ATLAS Collaboration. The resulting format is a ROOT tuple with more than 80 branches.
- Several reconstructed physics objects (electrons, muons, photons, hadronically decaying tau-leptons, and two collections of jets with different values of jet radius parameters) are contained within the 13 TeV ATLAS Open Data, and their preselection requirements are detailed in Table 1.

Table 1: Preselection requirements for electron, muon, photon, hadronically decaying τ-lepton, small-R-jet and large-R-jet candidates within the 13 TeV ATLAS Open Data. Reconstruction (rec.), identification and isolation criteria of the objects, as well as other additional requirements are given in Ref. [5].

Electron (e)	Muon (μ)	Photon (γ)
InDet & EMCAL rec. loose identification loose isolation $p_{\text{T}} > 7$ GeV $\lvert\eta\rvert < 2.47$	InDet & MS rec. loose identification loose isolation $p_{\text{T}} > 7$ GeV $\lvert\eta\rvert < 2.5$	InDet & EMCAL rec. tight identification loose isolation $E_{\text{T}} > 25$ GeV $\lvert\eta\rvert < 2.37$

Hadronically decaying τ-leptons (τ_h)	Small-R jets	Large-R jets
InDet & EMCAL rec. medium identification $p_{\text{T}} > 20$ GeV $\lvert\eta\rvert < 2.5$ 1 or 3 associated tracks	EMCAL & HCAL rec. anti-k_t, R = 0.4 $p_{\text{T}} > 20$ GeV $\lvert\eta\rvert < 2.5$	EMCAL & HCAL rec. anti-k_t, R = 1.0 $p_{\text{T}} > 250$ GeV $\lvert\eta\rvert < 2.0$ trimming: $R_{\text{sub}} = 0.2$, $f_{\text{cut}} = 0.05$

2.1 Evolution of the ATLAS Open Data

The evolution of the ATLAS Open Data structure from the 8 TeV release (2016) to the 13 TeV release (2020) is presented in Figure 1. The 13 TeV ATLAS Open Data contains a much larger set of simulated SM and BSM processes, and the preselected events are classified into seven different final-state collections that contain several new reconstructed objects with respect to the 8 TeV release, such as photons, hadronically decaying τ-leptons and large-R-jet candidates. A simplified single-component systematic-uncertainty estimate related to object transverse momentum reconstruction is included in the datasets.

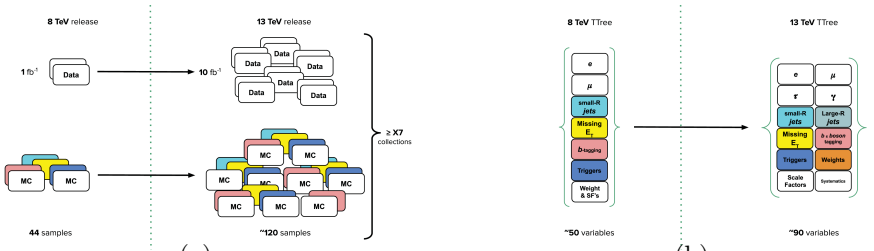


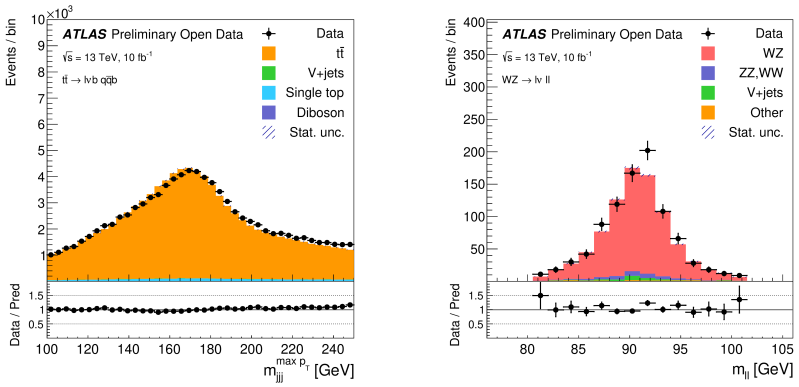
Figure 1. Evolution of ATLAS Open Data (a) dataset and (b) tuple structure from the 8 TeV release (2016) to the 13 TeV release (2020). From Ref. [5].

3 Physics-analysis examples using the 13 TeV ATLAS Open Data

The general aim of the 13 TeV ATLAS Open Data and tools is to provide a straightforward interface to replicate the procedures used by high-energy-physics (HEP) researchers and enable users to experience the analysis of particle-physics data in educational environments. Hence, several examples of physics analyses using the 13 TeV ATLAS Open Data were created, inspired by and following as closely as possible the procedures and selections taken in already published ATLAS physics results. In total, *twelve analyses*, grouped into different final states, have been prepared:

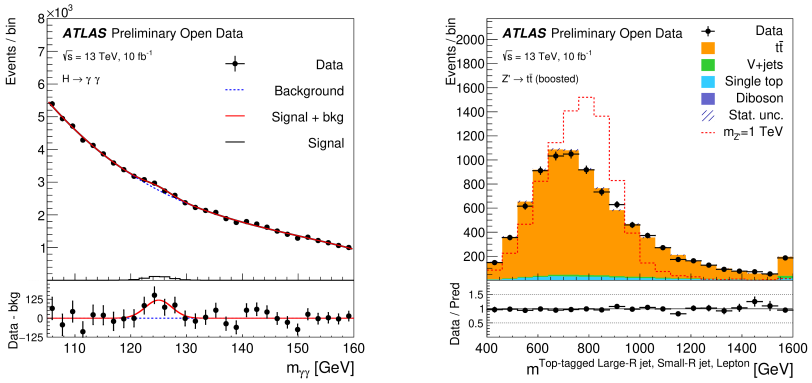
- *Four high-statistics* analyses with a selection of $W^{\pm} \to \ell\nu$ leptonic-decay events; single-Z-boson events, where the Z boson decays into an electron–positron or muon–antimuon pair, or into a τ-lepton pair with a hadronically decaying τ-lepton accompanied by a τ-lepton that decays leptonically; and top-quark pairs in the $t\bar{t} \to W^+W^-b\bar{b} \to \ell\nu q\bar{q}^{'}b\bar{b}$ final state. All of these analyses have sufficiently high event yields to study the SM processes in detail, and are intended to show the general good agreement between the released 13 TeV data and the MC prediction. They also enable the study of SM observables, such as the mass of the W and Z bosons, and that of the top quark, as shown in Figure 2(a).
- *Three low-statistics* analyses with a selection of single top quark produced in the t-channel in the $t+q \to Wb+q \to \ell\nu b+q$ final state; diboson events with $W^{\pm}Z \to \ell\nu\ell\ell^{'}$, shown in Figure 2(b); and $ZZ \to \ell^+\ell^-\ell^+\ell^-$ fully-leptonic final states. These analyses illustrate the statistical limitations of the released dataset given the low production cross section of the rare processes, where the variations between data and MC prediction are attributed to sizeable statistical fluctuations.

- *Three SM Higgs-boson* analyses with a selection of events in the $H \to WW^* \to e\nu\mu\nu$, $H \to ZZ^* \to 4\ell$ and $H \to \gamma\gamma$ decay channels, the latter illustrated in Figure 3(a), which serve as examples to implement simplified analyses in different final-state scenarios and "re-discover" the production of the SM Higgs boson.
- *Two BSM physics* analyses searching for new hypothetical particles: one implementing the selection criteria of a search for direct production of superpartners of SM leptons, and the second one implementing the selection criteria of a search for new heavy neutral Z' particles that decay into top-quark pairs, shown in Figure 3(b).



(a) (b)

Figure 2. Comparison between data and MC prediction for the distribution of (a) the invariant mass of the three-jets combination with the highest vector p_T in the $t\bar{t} \to W^+W^-b\bar{b} \to \ell\nu q\bar{q}'b\bar{b}$ selection, and (b) the invariant mass of the reconstructed Z-boson candidate in the $W^{\pm}Z \to \ell\nu\ell\ell'$ selection. From Ref. [5].



(a) (b)

Figure 3. Comparison between data and MC prediction for the distribution of (a) the diphoton invariant-mass spectrum in the $H \to \gamma\gamma$ selection, and (b) the approximate mass of the $t\bar{t}$ system in the new heavy $Z' \to t\bar{t}$ single-lepton boosted selection. From Ref. [5].

4 General capabilities and educational tools

The released datasets can be used for educational purposes with different levels of task difficulty and using a set of associated educational tools, as summarised below:

- At a *beginner level*, one could visualise the content of the datasets and produce simple distributions. An *intermediate-level* task would consist of making histograms with collision data after some basic selection, while *advanced-level* tasks would allow for a deeper look into the ATLAS data, with possibilities of measuring real event properties and physical quantities.
- An important aspect of the 13 TeV ATLAS Open Data is that it is prepared specifically for educational purposes. To this end, precision has been traded for simplicity of use. For example, no data-driven estimation of the multijet background is provided. Designing and implementing approximate data-driven methods to estimate multijet backgrounds is left as a *challenging exercise* to advanced students.
- The 13 TeV ATLAS Open Data is hosted on the CERN and ATLAS Open Data online portals [6, 7] and is accompanied by a set of associated educational tools:
  - *analysis framework*, written in C++ and interfaced with ROOT, which implements the protocols needed for reading the datasets, making an analysis selection, writing out histograms and plotting the results;
  - *online notebooks* allowing data analysis to be performed directly in a web browser by integrating the ROOT framework with the Jupyter notebook technology;
  - *virtual machine* that contains the analysis framework and the complete samples needed to carry out educational analysis of the released datasets;
  - *online documentation* platform [7] that provides introductory material and detailed information for a wide audience about the ATLAS experiment and the released 13 TeV ATLAS Open Data sets.

5 Summary

A new set of *pp* collision data has been released to the public for educational purposes, and constitutes the first public release of *pp* collision data samples recorded at 13 TeV by an LHC experiment. The data has been collected by the ATLAS detector at the LHC at $\sqrt{s} = 13$ TeV during the year 2016 and corresponds to an integrated luminosity of 10 fb^{-1}. The *pp* collision data is accompanied by a set of MC simulated samples describing several processes which are used to model the expected distributions of different signal and background events.

Associated educational tools are released to make the analysis of the dataset easily accessible. These educational tools provide a straightforward interface to replicate the procedures used by HEP researchers while enabling users to experience the analysis of particle-physics data in educational environments. A number of physics-analysis examples inspired by published ATLAS results are presented to demonstrate the wide range of final-state scenarios provided within the 13 TeV ATLAS Open Data.

References

[1] ATLAS Collaboration, JINST **3** S08003 (2008).
[2] ATLAS Collaboration, *ATLAS Data Access Policy*, ATL-CB-PUB-2015-001 (2015), https://cds.cern.ch/record/2002139.
[3] ATLAS Collaboration, *Review of ATLAS Open Data 8 TeV datasets, tools and activities*, ATLOREACH-PUB-2018-001 (2018), https://cds.cern.ch/record/2624572.
[4] ATLAS Collaboration, *Review of the ATLAS Open Data Dataset*, ATL-OREACH-PUB-2016-001 (2016), https://cds.cern.ch/record/2203649.
[5] ATLAS Collaboration, *Review of the 13 TeV ATLAS Open Data release*, ATL-OREACH-PUB-2020-001 (2020), https://cds.cern.ch/record/2707171.
[6] CERN Open Data Portal, http://opendata.cern.ch.
[7] ATLAS Open Data and Tools for Education, http://opendata.atlas.cern.