ATLAS Open Data – Development of a simple-but-real HEP data analysis framework

Farid Ould-Saada^{1,*}

¹University of Oslo, Norway

Abstract. The ATLAS Collaboration at the Large Hadron Collider is releasing a new set of recorded and simulated data samples at a centre-of-mass energy of 13 TeV collected in pp collisions at the LHC. This new dataset was designed after an in-depth review of the usage of the previous release of samples at 8 TeV. That review showed that capacity-building is one of the most important and abundant uses of public ATLAS samples. To fulfil the requirements of the community and at the same time attract new users and use cases, we developed real analysis software based on ROOT in two of the most popular programming languages: C++ and Python. These so-called analysis frameworks are complex enough to reproduce with reasonable accuracy the results -figures and final yields- of published ATLAS Collaboration physics papers, but still light enough to be run on commodity hardware. With the computers that university students and regular classrooms typically have, students can explore LHC data with similar techniques to those used by current ATLAS analysers. We present the development path and the final result of these analysis frameworks, their products and how they are distributed to final users inside and outside the ATLAS community.

1. Introduction

The main purpose of the ATLAS Open Data is to provide open access to proton-proton (pp) collision data and software and analysis tools from the ATLAS experiment at the Large Hadron Collider (LHC), in accordance with the ATLAS Open Data Access policy [1], which sets out the guidelines regarding open access to ATLAS data by non-ATLAS members with a focus on education, training and outreach. The datasets and the tools are made available through the CERN and ATLAS Open Data portals [2,3]. The target audiences are high-school, undergraduate and graduate students, as well as teachers and lecturers.

The ATLAS data have been successfully deployed since 2010 in the IPPOG [4] International Masterclasses (IMC) [5] where high-school students perform various measurements based on proton-proton collisions. The ambition to bring to the “classrooms” important LHC

* Farid Ould-Saada: farido@uio.no, on behalf of the ATLAS collaboration.

Copyright 2020 CERN for the benefit of the ATLAS Collaboration.

Reproduction of this article or parts of it is allowed as specified in the CC-BY-4.0 license.

discoveries is realised using the discovery of the Higgs boson. In 2012 an ATLAS proton-proton collision data sample of 2 fb^{-1} at an energy of 8 TeV was released in XML format and used to “search”, among other particles, for the Higgs boson. This is followed by an 8-TeV data sample, corresponding to 1 fb^{-1} , in ROOT [6] format, accompanied by the corresponding analysis tools [7] and aimed for undergraduate physics students. The 8-TeV data and tools led to a wide range of activities [8]: hands-on particle-physics exercises; production of teaching materials, lectures and public talks; as well as the introduction of new “research-based particle physics” courses, such as in [9]. Following feedback from the users, a new release of ATLAS Open Data has been made public, this time using 13 TeV data from Run II of the LHC.

2. 13 TeV ATLAS Open Data release

A new set of pp collision data at $\sqrt{s} = 13 \text{ TeV}$ has been released to the public for educational purposes [10]. The dataset corresponds to an integrated luminosity of 10 fb^{-1} recorded by the ATLAS detector at the LHC in 2016. Monte Carlo (MC) simulation samples describing several Standard Model (SM) and beyond the SM (BSM) processes, which are used to model the expected distributions of different signal and background processes, are included in the release as well.

Figure 1 compares the two ATLAS Open Data releases made using the 8 TeV data in 2016 and the 13 TeV data in 2019/2020. The new data release features the highest pp collision energy available at the LHC, 10 times more luminosity, nearly 3 times more MC samples, and 7 times more data collections.

Figure 1. Evolution of ATLAS Open Data from the 8 TeV release (2016) to the 13 TeV release (2019-2020).

As demonstrated in Figure 2 the 13 TeV ATLAS Open Data SM Higgs MC signals include more production mechanisms (VH and ttH in addition to ggF and VBF) and contain more physics objects (photons and tau leptons in addition to electrons and muons). There are more BSM MC signals in the new 13 TeV release (more than 50 samples including graviton excitations, supersymmetry and dark matter, in addition to new gauge bosons Z') than in the 8 TeV release (14 Z' samples).

Figure 2. Evolution of ATLAS Open Data SM Higgs and BSM MC signals from the 8 TeV release (2016) to the 13 TeV release (2019/2020).

Table 1 gives details about the Monte Carlo samples released in the 13 TeV ATLAS Open Data. The other BSM samples are on their way to be released.

Table 1. Description of the MC samples released in the 13 TeV ATLAS Open Data taken from [10], where the references to the various generators in the table can also be found.

Process	Unique “channelNumber”	Generator, hadronisation	Additional information
<i>Top-quark production</i>			
$t\bar{t}$ +jets	410000	POWHEG-Box v2 [68] + PYTHIA 8 [69]	only 1 ℓ and 2 ℓ decays of $t\bar{t}$ -system
single (anti)top t -channel	(410012) 410011	POWHEG-Box v1 + PYTHIA 6 [70]	
single (anti)top Wt -channel	(410014) 410013	POWHEG-Box v2 + PYTHIA 6	
single (anti)top s -channel	(410026) 410025	POWHEG-Box v2 + PYTHIA 6	
<i>W/Z (+ jets) production</i>			
$Z \rightarrow ee, \mu\mu, \tau\tau$	361106 – 361108	POWHEG-Box v2 + PYTHIA 8	LO accuracy up to $N_{\text{jets}} = 1$
$W \rightarrow e\nu, \mu\nu, \tau\nu$	361100 – 361105	POWHEG-Box v2 + PYTHIA 8	LO accuracy up to $N_{\text{jets}} = 1$
$W \rightarrow e\nu, \mu\nu, \tau\nu$ + jets	364156 – 364197	SHERPA 2.2 [71]	LO accuracy up to 3-jets final states
$Z \rightarrow ee, \mu\mu, \tau\tau$ + jets	364100 – 364141	SHERPA 2.2	LO accuracy up to 3-jets final states
<i>Diboson production</i>			
WW	363359, 363360	SHERPA 2.2	$qq'\ell\nu$ final states
WW	363492	SHERPA 2.2	$\ell\nu\ell'\nu'$ final states
ZZ	363356	SHERPA 2.2	$qq'\ell^+\ell^-$ final states
ZZ	363490	SHERPA 2.2	$\ell^+\ell^-\ell'^+\ell'^-$ final states
WZ	363358	SHERPA 2.2	$qq'\ell^+\ell^-$ final states
WZ	363489	SHERPA 2.2	$\ell\nu qq'$ final states
WZ	363491	SHERPA 2.2	$\ell\nu\ell^+\ell^-$ final states
WZ	363493	SHERPA 2.2	$\ell\nu\nu\nu'$ final states
<i>SM Higgs production ($m_{\text{H}} = 125$ GeV)</i>			
$ggF, H \rightarrow WW$	345324	POWHEG-Box v2 + PYTHIA 8	$\ell\nu\ell'\nu'$ final states
$VBF, H \rightarrow WW$	345323	POWHEG-Box v2 + PYTHIA 8	$\ell\nu\ell'\nu'$ final states
$ggF, H \rightarrow ZZ$	345060	POWHEG-Box v2 + PYTHIA 8	$\ell^+\ell^-\ell'^+\ell'^-$ final states
$VBF, H \rightarrow ZZ$	344235	POWHEG-Box v2 + PYTHIA 8	$\ell^+\ell^-\ell'^+\ell'^-$ final states
$ZH, H \rightarrow ZZ$	341947	PYTHIA 8	$\ell^+\ell^-\ell'^+\ell'^-$ final states
$WH, H \rightarrow ZZ$	341964	PYTHIA 8	$\ell^+\ell^-\ell'^+\ell'^-$ final states
$ggF, H \rightarrow \gamma\gamma$	343981	POWHEG-Box v2 + PYTHIA 8	
$VBF, H \rightarrow \gamma\gamma$	345041	POWHEG-Box v2 + PYTHIA 8	
$WH(ZH), H \rightarrow \gamma\gamma$	345318, 345319	POWHEG-Box v2 + PYTHIA 8	
$t\bar{t}H, H \rightarrow \gamma\gamma$	341081	aMC@NLO [72] + PYTHIA 8	
<i>BSM production</i>			
$Z' \rightarrow t\bar{t}$	301325	PYTHIA 8	$m_{Z'} = 1$ TeV
$\tilde{\ell}\tilde{\ell}' \rightarrow \ell\bar{\chi}_1^0\ell'\bar{\chi}_1^0$	392985	aMC@NLO + PYTHIA 8	$m_{\tilde{\ell}} = 600$ GeV, $m_{\tilde{\chi}_1^0} = 300$ GeV

The data ntuple structure has also evolved from about 50 variables in 2016 to about 90 variables in 2019 (Figure 3). Photons, tau-leptons, large-radius jets, b-quark and boson-tagging, as well as systematics are available in the 13 TeV release, enabling more detailed studies.

Figure 3. Evolution of ATLAS Open Data tuple structure from the 8 TeV release (2016) to the 13 TeV release (2019/2020).

A total of about 150 GB of storage is needed for all collision and simulated 13 TeV ATLAS Open Data.

3. ATLAS Open Data tools

Efforts were put into the production, validation and release of the ATLAS 13 TeV data and MC simulation (in ROOT ntuples format) for training and education, including a full software and analysis framework written both in C++ and Python and interfaced to ROOT.

Figure 4. Overview of the ATLAS Open Data Virtual Machine.

The framework implements the necessary features to read the dataset, and make histograms and plots. Example analyses are provided, among which some exemplify how to do event selections, weight events appropriately, and scale MC samples to the right cross section and luminosity. One way of achieving an operating-system-independent analysis is by means of a Virtual Machine (VM), where all the necessary software is pre-installed. An overview of the ATLAS Open Data VM is shown in Figure 4. This VM runs a Linux-based operating system, and has ROOT and other relevant software pre-installed. Furthermore, the VM contains the software analysis code and the dataset, as well as relevant documentation. However, VMs containing datasets require a certain amount of disk space. Also, VMs usually require considerable graphics resources, which may slow down the running significantly. A

solution to these problems is to run the VM as a server, which is possible with the new generation of VMs. The fact that the applications of the new generation of VMs are simply opened in the browser of the host machine / laptop requires far less disk space and graphics resources than with the standard VMs. An illustration of this system is given in Figure 5. Jupyter notebooks use both Python and C++ ROOT kernels to analyse data and MC samples and produce physics-analysis results using a VM as a server. JavaScript applications allow simple cut-and-count analyses. GitHub and GitLab are used as repositories and GitBooks for documentation.

In addition to acquiring particle-physics knowledge, the students have the opportunity to work with modern software tools based on new technology and get an insight into principles of cloud and distributed computing.

Figure 5. Illustration of a VM running as a server, where the applications are opened in the browser of the host machine.

4. Analysis examples

Various physics analysis examples are extensively discussed in Reference [10]. More are planned to be released soon. All will be documented and made available through notebooks as discussed in Section 3. Here we give two examples for illustration.

4.1 Search for Higgs decaying to diphotons

The first example makes use of a final state with two well reconstructed photons with high transverse momenta p_T . The distributions of the invariant mass, $m_{\gamma\gamma}$, of selected diphoton events are shown in Figure 6. This is done for two unconverted central (pseudorapidity $|\eta| < 0.75$)¹ photons (left) and inclusively for both unconverted and converted photons (right). There is a good agreement between the 13 TeV data and the SM prediction featuring a Higgs boson of mass 125 GeV. With this example the students have for the first time the possibility not only to display Higgs candidates, as already done by high-school students within masterclasses, but to closely reproduce the analysis work that led to a discovery [11].

¹ ATLAS uses a right-handed coordinate system with its origin at the nominal interaction point in the centre of the detector, and the z axis along the beam line. Pseudorapidity is defined in terms of the polar angle θ as $\eta = -\ln \tan \theta/2$.

With the advent of the ATLAS 13 TeV data, we have the ambition to tune and expand the educational material to follow the LHC 'heartbeats' and to influence the textbooks and teaching methods already at high school level.

Figure 6. Diphoton invariant-mass spectrum as observed in the selected diphoton events, with (left) two unconverted photons and (right) inclusive.

4.2 Search for electroweak production of supersymmetric particles

After the discovery of the last missing particle of the SM, the next challenges at the LHC are: what kinds of new physics are to be found beyond the SM? Where is dark matter? Are there additional fundamental symmetries? The second example developed here deals with a search for supersymmetry, through the production of a pair of sleptons, the super-partners of the known leptons. Each slepton decays to a lepton and a neutralino, an invisible dark matter candidate particle. The final state studied here features two charged leptons (electrons or muons) and missing transverse energy. The analysis closely follows Reference [12].

The distributions of the transverse momentum of the leading lepton, dilepton invariant mass, missing transverse momentum E_T^{miss} and “stransverse” mass m_{T2} before and after the two signal-region selections are shown in Figure 7. Good agreement is found between data and MC prediction, even in signal regions where large statistical fluctuations are observed. The number of observed (expected) events in the “loose” and “tight” signal regions are 57 (58.5) and 4 (2.4), respectively. Interesting to note are the events observed in data at very high values of E_T^{miss} (> 900 GeV). These have been investigated further and found to contain very high- p_T calorimeter-tagged muons which pass the loose muon selection working point. By applying tight identification criteria on the muons, these events could be rejected. The display and study of such events could make an excellent technical task for the students, giving them the opportunity to learn more about detection techniques.

5. Summary and outlook

ATLAS 13 TeV Open Data samples, collected in pp collisions at the LHC and corresponding to an integrated luminosity of 10 fb^{-1} , and various Standard Model and Beyond SM physics simulation samples, have been made public for education. The samples are accompanied by data processing and analysis tools. These resources allow students to easily access and analyse data using desktops or laptops, practise programming (Python & C++) and adapt Jupyter notebooks using both Python and C++ ROOT to analyse data and MC samples and produce physics analysis results using a VM as a server. Both the data and tools have been validated and deployed in a number of university courses. The students acquire particle-physics knowledge by making SM measurements, searching for new physics and reproducing as closely as possible important results published by the ATLAS Collaboration. In addition,

the students have the opportunity to work with modern software tools based on new technology and get an insight into principles of cloud and distributed computing.

Figure 7. Comparison between data and MC prediction for several distributions in a final state with 2 leptons and missing transverse energy in a search for supersymmetric particles.

References

- [1] ATLAS Collaboration, *ATLAS Data Access Policy*, ATL-CB-PUB-2015-001, <https://cds.cern.ch/record/2002139>.
- [2] CERN Open Data Portal, <http://opendata.cern.ch>.
- [3] ATLAS Open Data and Tools for Education, <http://opendata.atlas.cern>.
- [4] International Particle Physics Outreach Group: <http://ippog.org/>.
- [5] International Masterclasses: <http://physicsmasterclasses.org>.
- [6] R. Brun and F. Rademakers, *ROOT: An object-oriented data analysis framework*, Nucl. Instrum. Meth. A **389** (1997) 81.
- [7] ATLAS Collaboration, *Review Studies for the ATLAS Open Data Dataset*, ATL-OREACH-PUB-2016-001, <https://cds.cern.ch/record/2203649>.
- [8] ATLAS Collaboration, *Review of ATLAS Open Data 8 TeV datasets, tools and activities*, ATL-OREACH-PUB-2018-001, <https://cds.cern.ch/record/2624572>.
- [9] FYS5555, *Research-based particle physics*, University of Oslo, Norway.
- [10] ATLAS Collaboration, *Review of the 13 TeV ATLAS Open Data release*, ATL-OREACH-PUB-2020-001, <http://cdsweb.cern.ch/record/2707171>.
- [11] 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.
- [12] ATLAS Collaboration, *Search for electroweak production of supersymmetric particles in final states with two or three leptons at $\sqrt{s} = 13$ TeV with the ATLAS detector*, Eur. Phys. J. C **78** (2018) 995.