Enabling Open Science with the ATLAS Open Data project at CERN

Meirin Oan Evans
9214122

School of Physics and Astronomy
The University of Manchester

MPhys Report

May 2018

This project was performed under the supervision of Dr. Darren Price

Abstract

The ATLAS Open Data project aims to release real data collected from proton–proton collisions at the LHC to the public. These data can then be used for teaching, outreach, science communication and public engagement, as well as scientific research outside the ATLAS Collaboration. The intended target audience is from high-school students upwards. There has been a previous release of 8 TeV ATLAS Open Data, which succeeded in making particle physics more accessible. The purpose of this work is to improve and build upon the first ATLAS Open Data release, by allowing users to study higher-energy 13 TeV collisions. After developing a framework for producing 13 TeV ATLAS Open Datasets in the first half of this project, the second half has gone on to use this framework to produce simplified datasets for all 2015 data and Standard Model Monte Carlo processes with leptonic final states. This has allowed analyses of Z-boson, W-boson and ZZ diboson candidates. There remains further work to be done on this project, namely extending to more analyses and including Beyond Standard Model processes. One such extension analysis will be $H \to \gamma\gamma$. SUSY samples could be included as Beyond Standard Model processes.

1 Introduction

This project stems from the work already undertaken in starting to produce data from 13 TeV proton–proton (pp) collisions from the ATLAS (A Toroidal LHC ApparatuS) [1] experiment at the Large Hadron Collider (LHC) [2] for public use in education and scientific research. Having prepared a framework to produce 13 TeV Open Data in the first semester, this part of the project involved using the framework to produce simplified datasets for 13 TeV ATLAS Open Data analyses.

1.1 Proton–proton collisions at the LHC

The LHC is the largest particle accelerator in the world, based at CERN [3], the European Organization for Nuclear Research, on the French-Swiss border near Geneva. Since its start in 2008, it has been running intermittently up to the present. It is based in an underground tunnel 27 km in circumference. The LHC accelerates protons to about 3.1 ms^{-1} less than the speed of light, before colliding them together at a primary vertex [4]. At such high energies, it is the proton constituents that collide. Each constituent collision is called an event [5]. The particles produced in events are reconstructed in dedicated detectors and used to infer the quantum-mechanical process that occurred in the collision, to study these interactions in detail and observe deviations from model predictions. First, the LHC collided protons at a centre of mass energy of 7 TeV [6], then upgraded in 2011 to 8 TeV [7], and it is currently running at 13 TeV [8].

1.2 The ATLAS Experiment and detector

The ATLAS experiment is a detector placed at one of the pp collision points in the LHC ring. Collision products travel out from the collision point into the detector. ATLAS can directly detect muons [9], electrons [10], photons [11] and hadronic jets [12]. Hadronic jets are collimated clusters of charged particles that result from the hadronisation of quarks and gluons produced in collisions. ATLAS can also infer the presence of tau leptons [13] and missing transverse momentum [14]. The ATLAS detector is shown in Figure 1. Towards the centre of the detector are trackers [15], which measure particle positions so that their momenta can be calculated. Magnets curve the tracks [16] of charged particles to enable momentum and charge measurements. Outside the trackers are Liquid Argon electromagnetic calorimeters [17], which measure energy deposits from electrons and photons. Further out are hadronic calorimeters [18], to measure jets of hadrons. At the outside of the detector are muon chambers [19].

1.3 Particle production & Data analysis

Quarks and gluons in protons can interact in a variety of ways described by quantum mechanics to produce an array of particles. The number and properties of different produced particles are being tested. The higher the energy of the colliding particles, the greater the variety of products that can be made. If all detected collision events were kept, the data outflow of a few tens of TByte/s [20] would be too large, so the detector applies a trigger [21] to select the events of highest interest and reduce the data outflow rate to about 100 MByte/s [20].

In order to suppress backgrounds and enhance data purity in processes or decays of interest, selection requirements are applied to particles. Requirements may specify the

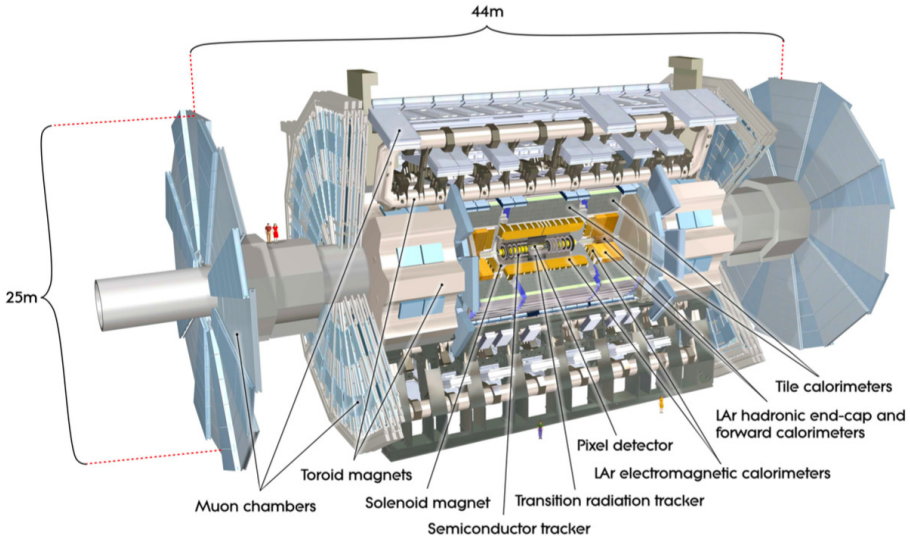


Figure 1: A schematic diagram of a cut-away section of the ATLAS detector [1]. Detector dimensions are shown and people can be seen in front for a sense of scale.

number of each particle candidate, as well as properties of those candidates being in a certain range, such as energy being over a specific value. To make data simpler to analyse, the number of events kept is reduced and objects that are of no importance to a particular analysis are skimmed out. Variables that are not of interest are not written to the output of the next stage to reduce file size.

1.4 The start of ATLAS Open Data

Once data analysis is finished and results have been obtained, physicists release papers to communicate their findings to the rest of the scientific community. Papers include a report of the findings and figures showing selected data, but the initial data used to make measurements are not made available due to resource limitations. Releasing data before most of the analysis stage can be useful for education, outreach, science communication and public engagement. Large collaborations like ATLAS have teams dedicated to these endeavours [22]. Many "citizen science" projects use public contributions to help scientific research [23]. In these projects, users can feel they have made their own discoveries. Sometimes this may be true; since data quantities are so large, scientists may not have had time to analyse them fully.

Following the first use of LHC data by the public in 2011, the ATLAS Open Data project was started. Releasing Open Data forms part of the ATLAS Data Access Policy. ATLAS Open Data come under the umbrella of CERN Open Data, which include results from the other LHC detectors CMS [24], Alice [25] and LHCb [26]. In 2016, ATLAS published a review of their 8 TeV Open Datasets, then gathered feedback over a year before reviewing their use.

1.5 First semester report summary

To summarise the first semester of this project [27], a framework was developed to convert raw data and Monte Carlo (MC) [28] to a simplified form. The framework was modified to produce datasets as similar as possible to the 8 TeV Open Datasets. This

meant adding and modifying variables to match 8 TeV Open Data. Produced datasets were ran through the 8 TeV Open Data analysis code to make plots such as a transverse mass comparison for a subset of 2015 data and an inclusive W-boson [29] MC sample.

2 Physics background for analysing heavy bosons

Fundamentally, ATLAS measures the interactions of final state particles as they travel through detector layers. These interactions are called hits. Final state particles are those with lifetimes long enough to reach the calorimeters. Connecting hits together shows the particle tracks, which can be used to reconstruct vertices. For charged particles, track curvature can be used to calculate 4-momenta. From the x and y momenta that form part of 4-momenta, transverse momenta, p_{T}, are obtained. p_{T} is an example of a final state particle measured property. From measured properties, related properties can be calculated. If an event contained two final state leptons of the same flavour and opposite charge, an interesting related property would be dilepton pair invariant mass, $M_{\ell\ell}$. Related properties allow one to understand more complex physics objects that are too short-lived to reach the calorimeters. Events with related properties that fit certain criteria are said to contain candidate particles. The candidate particle in the example of two same flavour and opposite charge leptons might be a Z-boson [30] if $M_{\ell\ell}$ is within a certain range of the Z mass.

2.1 Mass reconstruction

The mass of a parent particle that decays to stable particles visible in the ATLAS detector can be determined in processes, such as

$$Z \to \ell^+\ell^-, \tag{1}$$

where Z is a Z-boson and $\ell^\pm$ are opposite charge leptons that reach the ATLAS detector (electron or muon types). Dilepton invariant mass, $M_{\ell\ell}$, is given by

$$M_{\ell\ell} = \sqrt{(E_+ + E_-)^2 - (\mathbf{p}_+ + \mathbf{p}_-)^2}, \tag{2}$$

where $E_\pm$ and $\mathbf{p}_\pm$ are the energies and momenta of the positively and negatively charged leptons respectively. For Z candidates, the $M_{\ell\ell}$ distribution will be roughly symmetrical around the Z mass. The shape is mainly due to detector errors. Another decay where invariant mass is useful is

$$H \to \gamma\gamma, \tag{3}$$

where H is a Higgs boson [31] decaying to two photons, $\gamma\gamma$ [32]. In this case, diphoton invariant mass, $M_{\gamma\gamma}$, would be centred on the Higgs mass. A different resonant process such as neutral pion decay,

$$\pi^0 \to \gamma\gamma, \tag{4}$$

would not have an $M_{\gamma\gamma}$ distribution centred on the Higgs mass. The main background to Equation 3 is non-resonant diphoton production [33].

The transverse momenta, p_{T}, of all visible particles are measured. When applying momentum conservation in the transverse plane before and after the pp collision, the missing transverse momentum, $p_{\mathrm{T,Miss}}$, can be calculated. $p_{\mathrm{T,Miss}}$ could be attributed to particles that pass through the ATLAS detector, such as neutrinos [34]. Transverse

mass is a useful quantity in an analysis of W-bosons decaying to a lepton and a neutrino, as in

$$W \to \ell \nu_\ell, \tag{5}$$

where ℓ is a charged lepton that reaches the ATLAS detector (electron or muon), and ν_ℓ the corresponding neutrino. Transverse mass, $M_{\mathrm{T,W}}$, is defined by

$$M_{\mathrm{T,W}} = \sqrt{(E_{\mathrm{T},\ell} + E_{\mathrm{T},\nu})^2 - (\mathbf{p}_{\mathrm{T},\ell} + \mathbf{p}_{\mathrm{T},\nu})^2}, \tag{6}$$

where $E_{\mathrm{T},\ell}$ and $E_{\mathrm{T},\nu}$ are the transverse energies of the lepton and neutrino respectively, whilst $\mathbf{p}_{\mathrm{T},\ell}$ and $\mathbf{p}_{\mathrm{T},\nu}$ are their momenta. The expected $M_{\mathrm{T,W}}$ distribution shape is similar to invariant mass, but with an asymmetric shift towards lower mass. The shift occurs because low mass events have lower p_{T}, therefore only using transverse components in the $M_{\mathrm{T,W}}$ calculation has a higher effect on these events. Equation 1 could be a background to a W analysis if one lepton were not detected, which would lead to a $p_{\mathrm{T,Miss}}$ mismeasurement, which could be misinterpreted as a neutrino.

2.2 General event requirements

In order to study heavy bosons, event selection criteria are applied according to Table 1. Some of these requirements occur at an earlier stage than the final analysis code. "Good objects" are defined following Table 2. A b-tagging algorithm [35] computes whether an event contained a b-jet [36] originating from a bottom hadron. After generic selection requirements, objects undergo further requirements specific to each analysis.

Table 1: Details of the event requirements for all analyses in the 13 TeV ATLAS Open Data project. These are applied before any object selection requirements, such as those in Table 2.

Requirement	Details
Primary vertex track cut	$N_{\mathrm{tracks}} > 2$
Trigger applied	Single lepton trigger has to be satisfied
Good run list [37]	Corrupted events [38] are not used
Veto events with bad jets	Bad jets do not originate from the primary pp collision
Preselected objects	≥ 1 lepton with $p_{\mathrm{T}} > 25$ GeV

2.3 Specific object selection requirements

Since Z, W and Higgs bosons are each of order 100 times more massive than protons, they are typically produced nearly at rest when protons collide to form bosons without associated jets. Their decay products travel at 180° to each other in the boson rest frame, meaning they often have significant p_{T} in the boson rest frame. If the bosons are produced nearly at rest, their rest frames are similar to the lab frame. Therefore, high p_{T} final state particles should be looked for when selecting heavy boson candidates.

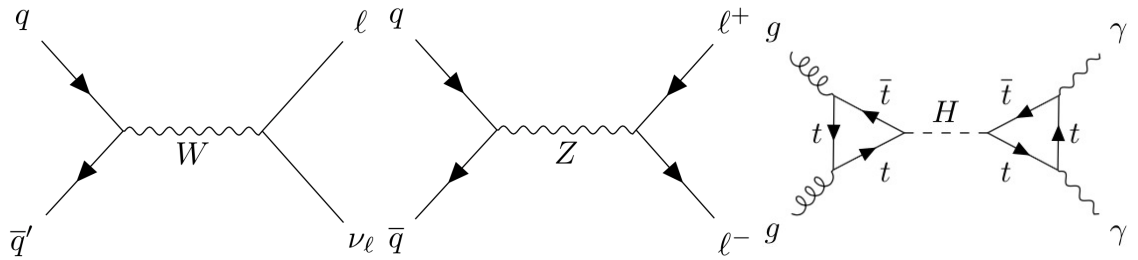
Decay products from real heavy bosons should generally be isolated. Non-isolated final state particles may originate from hadronic jets, which indicate that they have not come directly from real heavy bosons. Isolation is measured by drawing cones with vertices at the pp collision point and bases at the detector plane. The sum of p_{T} or transverse energy, E_{T}, within these cones is measured separately. Low p_{T} and E_{T} within the cones indicates isolated particles.

Table 2: Details of the requirements to select objects for all 13 TeV ATLAS Open Data analyses. Pseudorapidity [39] η is defined by $\eta = -\ln\tan\theta/2$, where θ is the polar angle defined with respect to the z-axis. Tight quality refers to objects that have likely been correctly reconstructed, whilst loose means a lower certainty in reconstruction. z_0 is the z-coordinate of a lepton track with respect to the primary vertex, whilst d_0 is the corresponding distance in the transverse plane. High Jet Vertex Tagging corresponds to jets with constituents all likely to have come from the same primary vertex. These criteria have been defined by the Standard Model [40] analysis group.

Requirement	electrons	muons	jets
$p_{\mathrm{T}} >$	20 GeV	15 GeV	25 GeV
$\lvert\eta\rvert <$	2.6	2.6	2.5
Quality	Loose or better	passes basic Muon Combined Performance [41] track requirements	Veto BadLoose
Primary Vertices	$\lvert z_0\rvert < 2.0$ mm $\lvert d_0\rvert < 2.0$ mm	$\lvert z_0\rvert < 2.0$ mm $\lvert d_0\rvert < 2.0$ mm	"Jet Vertex Tagging" > 0.59
Reconstruction Algorithm	Calorimeter & track based	Muid combined [42]	antiKt4EMTopo [43]

2.4 Selecting decays of interest

Imposing selection requirements will choose some real heavy bosons, but also some fake background. The reason for imposing selection requirements is to improve the signal to background ratio. Figure 2 shows example Feynman diagrams [44] for the production and subsequent decay of real W, Z and Higgs bosons at the LHC and Figure 3 shows the same for ZZ diboson production [45]. These are the signal processes of interest for this project. Figure 4 shows some Feynman diagrams for typical backgrounds to Z, W and Higgs searches. The $t\bar{t}$ [46] process in Figure 4(a) could fake a W signal if one lepton were not detected. The same process could also fake a Z signal if both leptons are of the same flavour. The gluon–gluon [47] process of Figure 4(b) could be a background to a W analysis if one of the hadronic jets formed by the gluons were misreconstructed as a lepton. This misreconstruction would mean a mismeasurement of p_{T} and thus $p_{\mathrm{T,Miss}}$, which could be misinterpreted as a neutrino. For gluon–gluon scattering to be misidentified as a Z signal, two gluon jets would have to be misreconstructed as leptons. The requirements imposed on events to select W and Z candidates are shown in Table 3.



(a) $W \to \ell\nu_\ell$ (b) $Z \to \ell^+\ell^-$ (c) $H \to \gamma\gamma$

Figure 2: Examples of Feynman diagrams for the processes $W \to \ell\nu_\ell$, $Z \to \ell^+\ell^-$ and $H \to \gamma\gamma$. $\bar{q}'$ denotes an antiquark of different flavour to quark q. g denotes a gluon. Other symbols are defined in Section 2.1. No arrows are drawn when a particle could be a fermion or an antifermion.

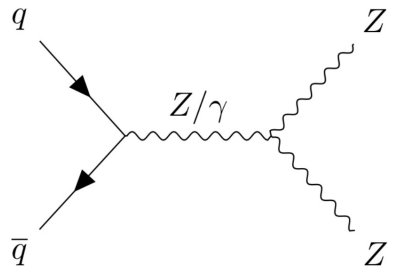
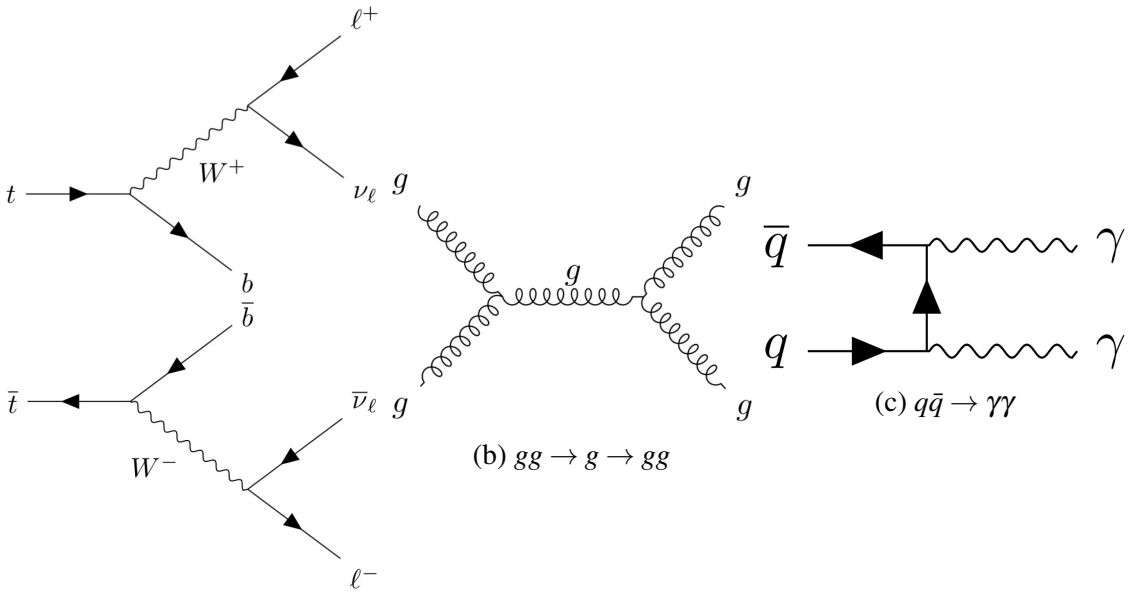


Figure 3: Example Feynman diagram for ZZ diboson production. Z/γ indicates either Z or γ.



(a) $t\bar{t} \to bW^+\bar{b}W^- \to b\ell^+\nu_\ell\bar{b}\ell^-\bar{\nu}_\ell$ (b) $gg \to g \to gg$ (c) $q\bar{q} \to \gamma\gamma$

Figure 4: Examples of Feynman diagrams for $t\bar{t}$, gluon–gluon scattering and non-resonant diphoton production, which may appear as backgrounds to W, Z and Higgs analyses.

Table 3: Details of the requirements to select W and Z-boson candidate events for the 13 TeV ATLAS Open Data project. The ptcone and etcone ratios define isolation variables.

Requirement	W-boson	Z-boson
Number of leptons with $p_{\mathrm{T}} > 25$ GeV	Exactly 1	2 opposite charge & same flavour
Lepton ptcone30/$p_{\mathrm{T}} < 0.15$	Required	Required
Lepton etcone20/$p_{\mathrm{T}} < 0.15$	Required	Required
$p_{\mathrm{T,Miss}}$	> 30 GeV	No requirement
Mass reconstruction	Transverse mass > 30 GeV	Invariant mass within 20 GeV of Z PDG mass [48]

2.5 Data and Monte Carlo

Events recorded at the LHC are referred to as "data". Data consist of candidate events that include signal and background for a given analysis. Important counterparts to data are MC simulations. MC programmes simulate one particular process, which may be a signal or background to a given analysis. A principle aim of particle physics is to understand

the MC contributions to data events that have undergone selection requirements. The integrated luminosity over the year 2015 was 3.2 inverse femtobarns (fb^{-1}), which is essentially the amount of data collected. Table 4 gives cross sections, event generators and calculation results for the MC samples used in the 13 TeV Open Data Z and W analyses.

Table 4: Cross sections, event generators and results for the percentage contributions of different Monte Carlo processes to 13 TeV ATLAS Open Data Z and W-boson analyses.

Process	$\sigma_{\text{effective}}$ (pb)	Generator	ZAnalysis(%)	WAnalysis(%)
$Z \to \ell^+\ell^-$	5805	PowhegPythia [49]	99.7	3
$W \to \ell\nu_\ell$	58800	Sherpa [50]	0.001	95
Diboson [51]	79.93	PowhegPythia	0.2	0.2
$t\bar{t}$	451.6	PowhegPythia	0.2	0.9
single top [52]	145.4	PowhegPythia	0.02	0.3
8 GeV$<M_{\ell\ell}<$40 GeV Drell-Yan [53]	6791	Sherpa	0.000008	0.06

2.6 Luminosity scaling in Monte Carlo simulated samples

MC is produced independently of data luminosity, therefore MC has to be scaled to match data. Each MC process has a cross section, σ, amounting to the quantum-mechanical amplitude presented by this process. Rather than generating MC events for every single data event, MC samples are given a weight: common processes are weighted up and rare processes weighted down. The initial weights at MC generation are summed to give an initial sum of weights. Applying selection requirements reduces the events to have a selected sum of weights. The ratio of generated to selected weights defines a selection efficiency.

Luminosity scaling between MC and data is done by calculating the MC sample luminosity, L_{sample}, and taking the ratio with the data target luminosity, L_{data}. The luminosity scale factor, SF_{lumi}, for a MC process with effective cross section $\sigma_{\text{effective}}$ and initial sum of weights $\Sigma w_{\text{initial}}$ is given by

$$SF_{\text{lumi}} = \frac{L_{\text{data}}\sigma_{\text{effective}}}{\varepsilon_{\text{red}}\Sigma w_{\text{initial}}}, \tag{7}$$

where ε_{red} is the efficiency from generation to selection, which is the number of selected events divided by the number of initial events, giving a reduction in the number of events. The MC sample luminosity is thus

$$L_{\text{sample}} = \frac{\varepsilon_{\text{red}}\Sigma w_{\text{initial}}}{\sigma_{\text{effective}}}. \tag{8}$$

The denominator of Equation 7 gives an estimate for the selected sum of weights, $\Sigma w_{\text{selected}}$. $\sigma_{\text{effective}}$ itself is given by

$$\sigma_{effective} = kFact\,\sigma_{MC}\varepsilon_{filter}, \tag{9}$$

where kFact is a correction factor applied from theory, σ_{MC} is the cross section calculated by the MC event generator and ε_{filter} is an efficiency of selection cuts within the MC generator, relative to selecting all events.

2.7 Jets

Even when analyses do not impose selection requirements on hadronic jets, there remain interesting variables related to jets. Heavy bosons can be produced in association with jets, but the probability decreases by about 10% with each extra jet. The number of b-tagged jets is also interesting. b-jets are useful since top quarks decay to bottom quarks 95.7% of the time [48], and the $H \to b\bar{b}$ decay [54] is the most probable Higgs boson decay, with a probability of 65.36% [48]. b-jets are discriminated from jets of other types by a multi-variate technique [55] that uses boosted decision trees [56].

2.8 Number of vertices

Once variables defining selection requirements have been compared between data and MC, other variables are checked. One such variable is the number of primary vertices. To increase instantaneous luminosity, one method is to increase the number of protons per proton bunch. Each proton constituent collision defines a primary vertex. The number of proton constituents that collide per bunch crossing is referred to as pile-up [57]. The number of primary vertices depends on LHC accelerator conditions, therefore changes from one year to the next as instantaneous luminosity is increased.

3 Results

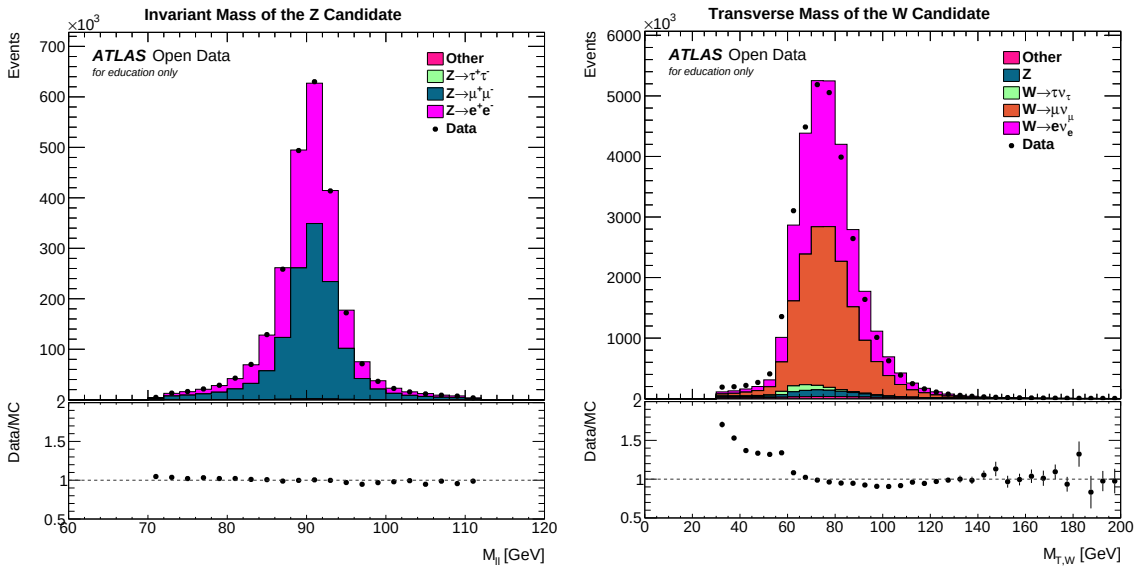
The primary results for this section of the 13 TeV ATLAS Open Data project were obtained after running over all 2015 data and the main Standard Model processes given in Table 4. W production includes MC samples for $W\to e\nu_e$, $W\to\mu\nu_\mu$ [58] and $W\to\tau\nu_\tau$ [59], with jets of all energies and flavours. Diboson includes the MC samples $W^+W^-\to\ell^+\nu_\ell\ell^-\overline{\nu}_\ell$, $W^+W^-\to\ell\nu_\ell qq$ [60], $WZ\to\ell\nu_\ell\nu_\ell\nu_\ell$, $WZ\to qq\ell^+\ell^-$, $WZ\to\ell\nu_\ell qq$, $WZ\to\ell\nu_\ell\ell^+\ell^-$ [61], $ZZ\to\ell^+\ell^-\ell^+\ell^-$, $ZZ\to\nu_\ell\nu_\ell\ell^+\ell^-$ and $ZZ\to qq\ell^+\ell^-$. Single top includes both t and $\bar{t}$ decays via t, s [62] or wt [63] channels.

3.1 W and Z candidate selections

To compare 13 TeV datasets in MC and data, a plot of Z candidate invariant mass is shown in Figure 5(a) for a 13 TeV ATLAS Open Data Z analysis. Similarly, Figure 5(b) shows W candidate transverse mass. The distribution shapes in Figure 5 are as expected. There is some discrepancy at low $M_{\mathrm{T,W}}$ in Figure 5(b) due to not including background QCD multijet processes such as Figure 4(b). These processes were not available in the same datatype as the other MC used. Plots of variables that undergo selection requirements in the analysis code to identify either W or Z candidate events according to Table 3 are given in Figures 6 and 9. Only one plot for each selection requirement variable is shown, though such plots exist for both W and Z analyses. Z lepton plots are split into leading lepton and trailing lepton. Plots such as Figures 5 were produced using **ROOT** [64] interfaced with **Python** [65]. "Data" in all Figures correspond to all of 2015 and error bars represent statistical uncertainties on data. "Data/MC" subplots are ratios of data to MC.

3.2 Rescaling the number of vertices

Since the number of primary vertices has no dependence on particle-level processes, rescaling MC to match data in this variable is justified. If the number of primary vertices



(a) Z candidate invariant mass against number of events. The signal MC samples are $Z \to e^+e^-$ and $Z \to \mu^+\mu^-$. "Other" includes backgrounds for dibosons, low-mass Drell-Yan, W, single top and $t\bar{t}$. $Z \to \tau\tau$ is also background.

(b) W candidate transverse mass against number of events. The signal MC samples are W + jets processes. "Other" includes backgrounds for dibosons, low-mass Drell-Yan, single top and $t\bar{t}$. $W \to \tau\nu_\tau$ and Z are also backgrounds.

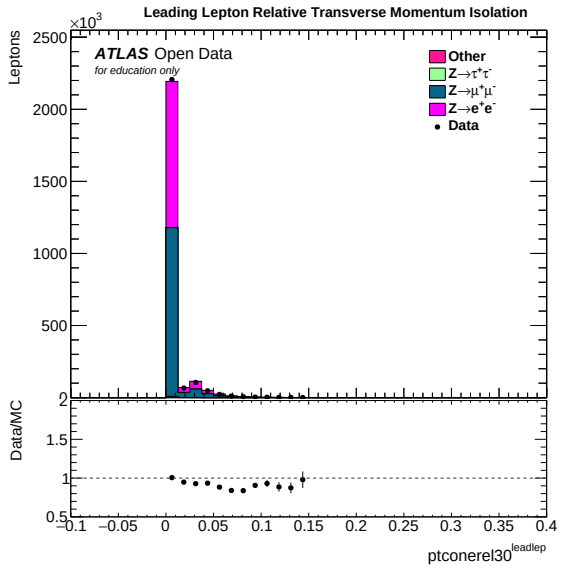
Figure 5: Stacked histograms of mass variables for 13 TeV ATLAS Open Data Z and W analyses. Both distribution shapes are as expected from Section 2.1. and demonstrate the mass reconstruction selection requirements from Table 3.

is decreased, final state particles should be more isolated. It was realised that there was a large discrepancy in the number of vertices between data and MC. Figure 7 shows an example for a W analysis. This was due to the fact that the MC samples used were produced to match with more recent data than 2015. The simplest solution is to rescale MC to better reproduce the distributions in data.

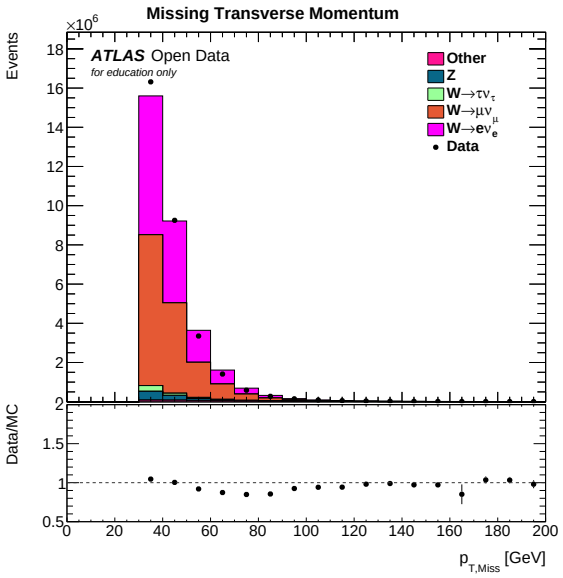
A number of polynomial fits were attempted to best describe the distribution seen in Figure 7(b). The coefficients and reduced-χ^2 of these fits are given in Table 5. All fits were made between 1 and 19 primary vertices, which is the range of histogram bins where data are shown in Figure 7(b). MC events with more than 19 vertices were rescaled to 0 since no data are seen in this region. These fits were made for the combination of all analyses. Using the combination produces suitable rescales for each analysis separately.

The most thorough solution to correcting a discrepancy between the scales of data and MC, is to generate new MC events to match data. This is expensive in CPU time, therefore the simpler solution of rescaling MC is sufficient for this project. After finding a suitable fit to the number of vertices against the ratio of Data/MC, this fit was applied as a rescaling to MC. In Figure 8 are plots of the number of vertices, before and after rescaling.

To see the effect of rescaling number of vertices on other variables, lepton transverse momentum, p_{T}, and lepton etconerel20 isolation from Table 3 are plotted before and after rescaling in Figure 9. Lepton p_{T} was chosen as a variable that should be independent of number of vertices. Lepton etconerel20 isolation was chosen as a variable that should

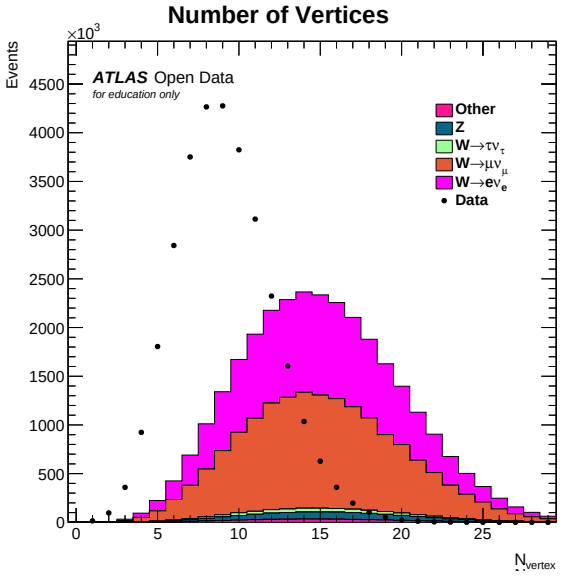


(a) Leading lepton p_{T} isolation against number of leptons, for a Z analysis. ptconerel30 is defined as the ptcone selection requirement in Table 3. The signal MC samples are $Z \to e^+e^-$ and $Z \to \mu^+\mu^-$. "Other" includes backgrounds for dibosons, low-mass Drell-Yan, W, single top and $t\bar{t}$. $Z \to \tau\tau$ is also background.

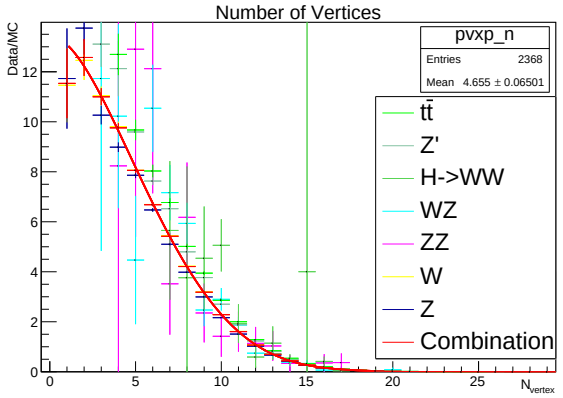


(b) Missing transverse momentum against number of events, for a W analysis. The signal MC samples are 13 TeV W + jets processes. "Other" includes backgrounds for dibosons, low-mass Drell-Yan, single top and $t\bar{t}$. $W \to \tau\nu_\tau$ and Z are also backgrounds.

Figure 6: Stacked histograms of certain variables that undergo selection requirements in 13 TeV ATLAS Open Data Z or W analyses, according to Table 3. Both Data/MC ratios are close to 1.



(a) Number of vertices against number of events, before rescaling.



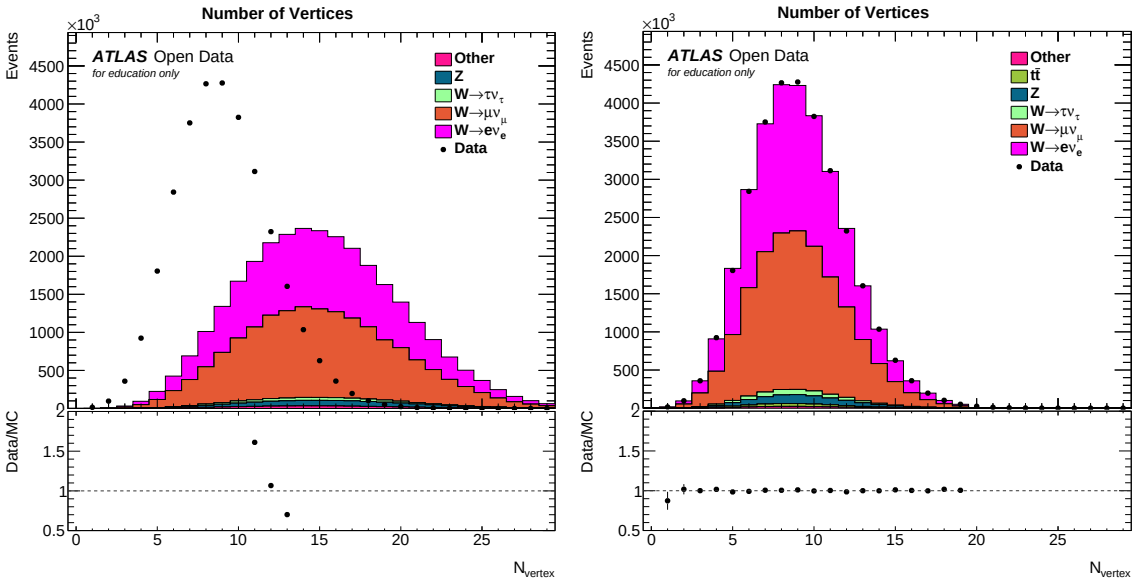
(b) Ratio of number of vertices in data to Monte Carlo before rescaling. Legend entries correspond to analysis names. Z' is a hypothetical dark matter [66] particle. "Combination" is the ratio of data events from all analyses to Monte Carlo events from all analyses. pvxp_n is another label for N_{vertex}.

Figure 7: Plots showing the initial discrepancy in number of vertices between data and Monte Carlo, for a 13 TeV ATLAS Open Data W analysis.

have some dependence on number of vertices, but also depends on the physics of the boson decay.

Table 5: Results for the fitting of Figure 7(b), to be able to rescale Monte Carlo samples following the original discrepancy in number of vertices between data and Monte Carlo.

Polynomial order	Coefficients	N_{dof}	χ^2/N_{dof}
1	a = 1.625 ± 0.003	17	9085.94
$a+bx$	b = -0.0877 ± 0.0002		
2	a = 9.51 ± 0.02	16	1442.36
quadratic	b = -1.051 ± 0.003		
$a+bx+cx^2$	c = 0.02918 ± 0.00008		
3	a = 21.5061 ± 0.08	15	23.0982
cubic	b = -3.45 ± 0.02		
$a+bx+cx^2+dx^3$	c = 0.186 ± 0.001		
	d = -0.00338 ± 0.00002		
4	a = 22.0 ± 0.2	14	24.3699
quartic	b = -3.60 ± 0.07		
$a+bx+cx^2+dx^3+ex^4$	c = 0.203 ± 0.007		
	d = -0.0041 ± 0.0003		
	e = 0.000013 ± 0.000006		
5	a = 13.4 ± 0.5	13	1.66206
quintic	b = 0.0 ± 0.2		
$a+bx+cx^2+dx^3+ex^4+fx^5$	c = -0.36 ± 0.03		
	d = 0.039 ± 0.002		
	e = -0.00158 ± 0.00009		
	f = 0.000023 ± 0.000001		



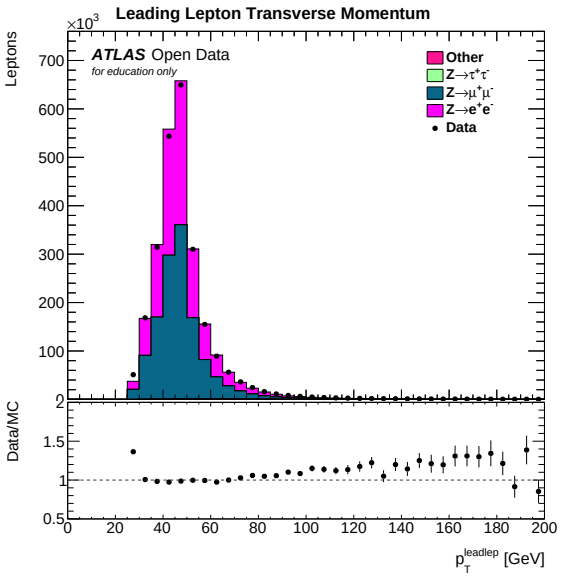
(a) Number of vertices before rescaling.

(b) Number of vertices after rescaling.

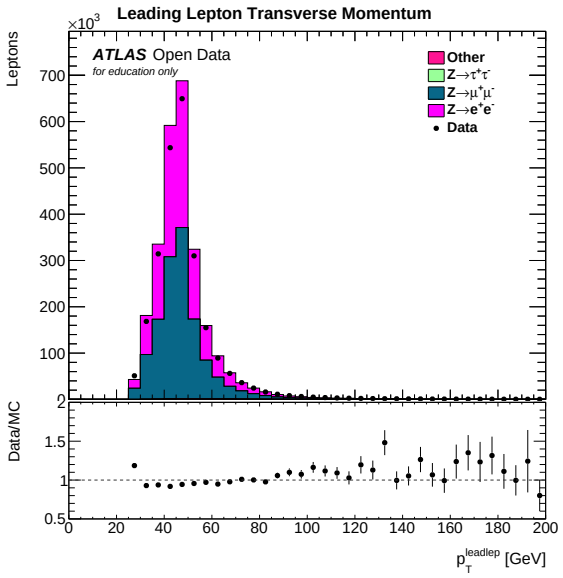
Figure 8: Stacked histograms showing the difference in the number of vertices between data and Monte Carlo, before and after rescaling, for a 13 TeV ATLAS Open Data W analysis. Rescaling makes Data/MC closer to 1. "Other" includes backgrounds for low-mass Drell-Yan, single top and dibosons in both (a) and (b), as well as $t\bar{t}$ in (a). $W\to\tau\nu_\tau$ and Z are also backgrounds.

3.3 Jets

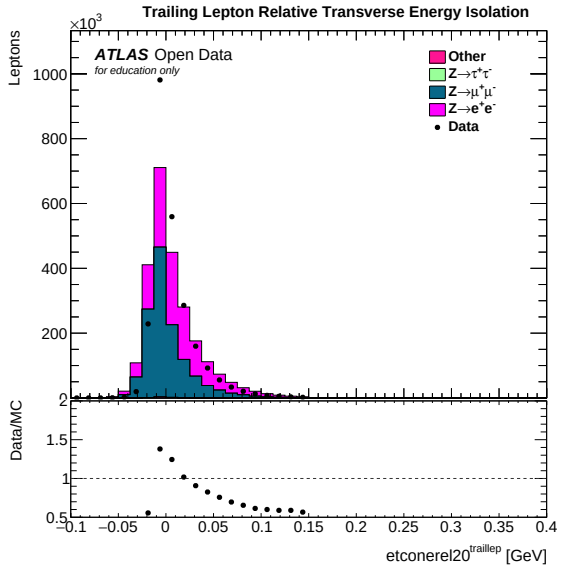
Shown in Figures 10 to 12 are several jet variables. Figure 10(a) shows the number of events decreasing by approximately a factor of 10 with each extra jet. Figure 11(a)



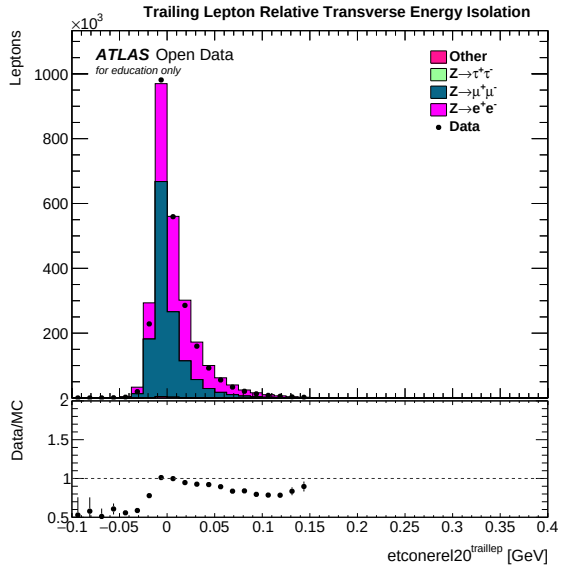
(a) Leading lepton p_{T} before rescaling.



(b) Leading lepton p_{T} after rescaling.



(c) Trailing lepton etconerel20 before rescaling.



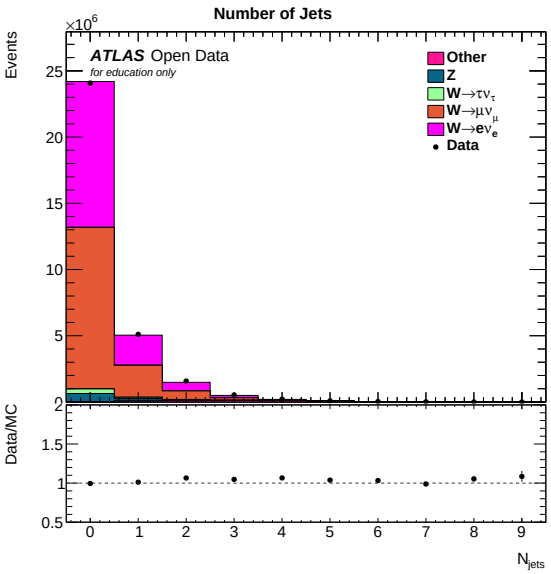
(d) Trailing lepton etconerel20 after rescaling.

Figure 9: Stacked histograms showing the difference in selected lepton variables between data and Monte Carlo, before and after rescaling the number of vertices, for an ATLAS Open Data Z analysis. etconerel20 is defined as the etcone selection requirement in Table 3. "Other" includes backgrounds for dibosons, low-mass Drell-Yan, W, single top and $t\bar{t}$. $Z \to \tau\tau$ is also background.

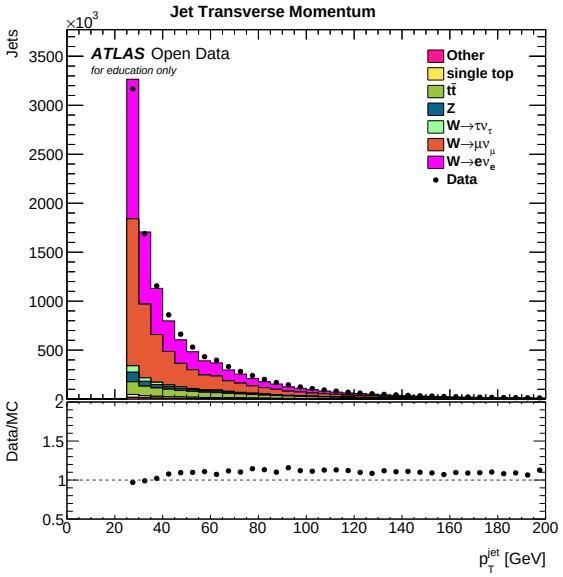
demonstrates that few events passing the $Z \to e^+e^-$ selection requirements contain b-jets. The contribution of $t\bar{t}$ can be seen in Figure 12(a), since b-jets have a high MV2c10 weight. One might wish to tag b-jets when searching for t quarks or $H \to b\bar{b}$.

3.4 Electron vs muon selections

Since electrons are detected in a more messy detector region than muons, there may be a difference in isolation between electrons and muons. Figure 13 compares electron and muon isolation for a Z analysis, and Figure 14 for a W analysis. Though the data to MC ratios show a slight difference between electrons and muons, the overall distribution shapes are similar. The conclusion from this is that the differences between electron and

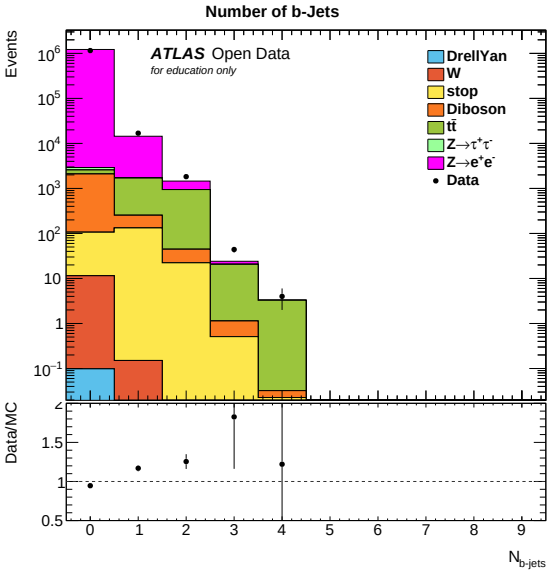


(a) Number of jets against number of events. "Other" includes backgrounds for $t\bar{t}$, single top, dibosons and low-mass Drell-Yan.

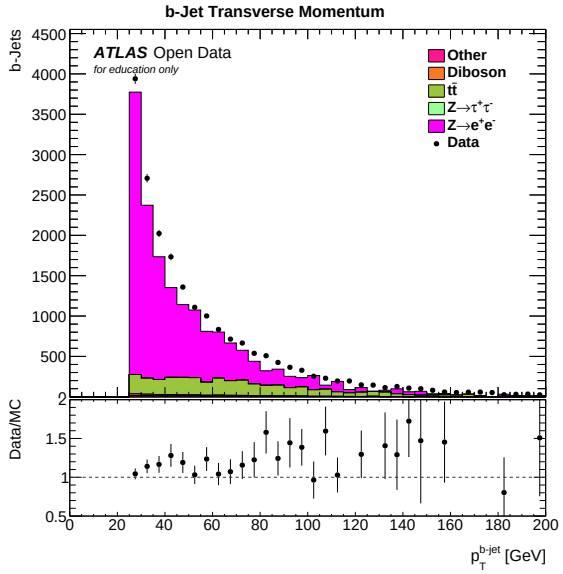


(b) Jet transverse momentum against number of jets. "Other" includes backgrounds for dibosons and low-mass Drell-Yan.

Figure 10: Stacked histograms of selected jet variables for a 13 TeV ATLAS Open Data W analysis. The signal Monte Carlo samples are W + jets processes. $W\to\tau\nu_\tau$ and Z are backgrounds.



(a) Number of b-jets against number of events. Legend entries are detailed in Table 4.



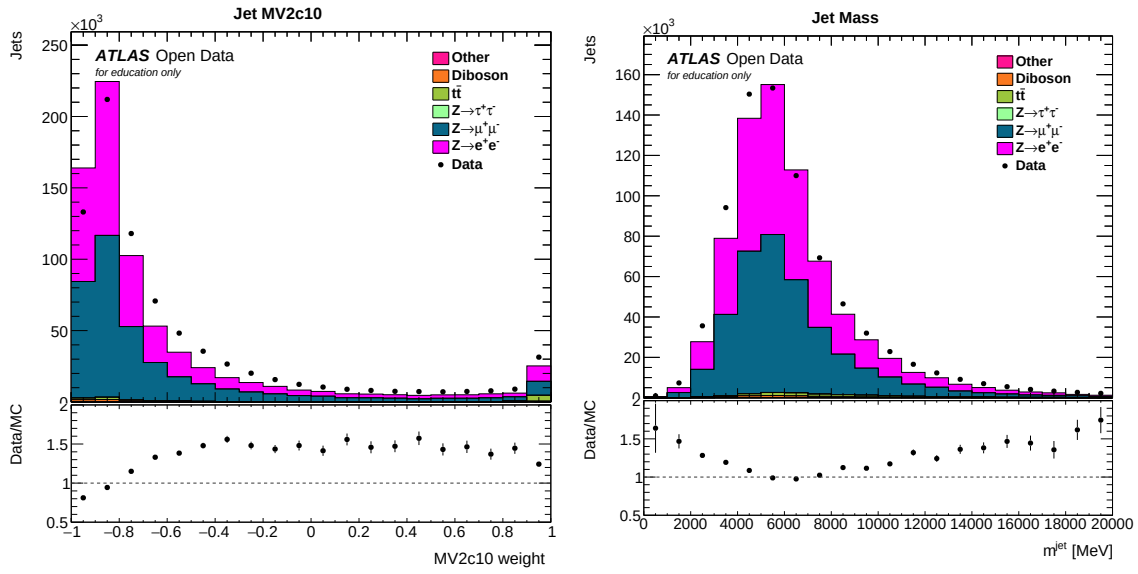
(b) b-jet transverse momentum against number of b-jets. "Other" includes backgrounds for W, single top and low-mass Drell-Yan.

Figure 11: Stacked histograms of selected b-jet variables for a 13 TeV ATLAS Open Data $Z\to e^+e^-$ analysis. $Z\to\tau\tau$, $t\bar{t}$ and diboson are also backgrounds.

muon isolation have been well modeled in MC.

3.5 ZZ diboson analysis

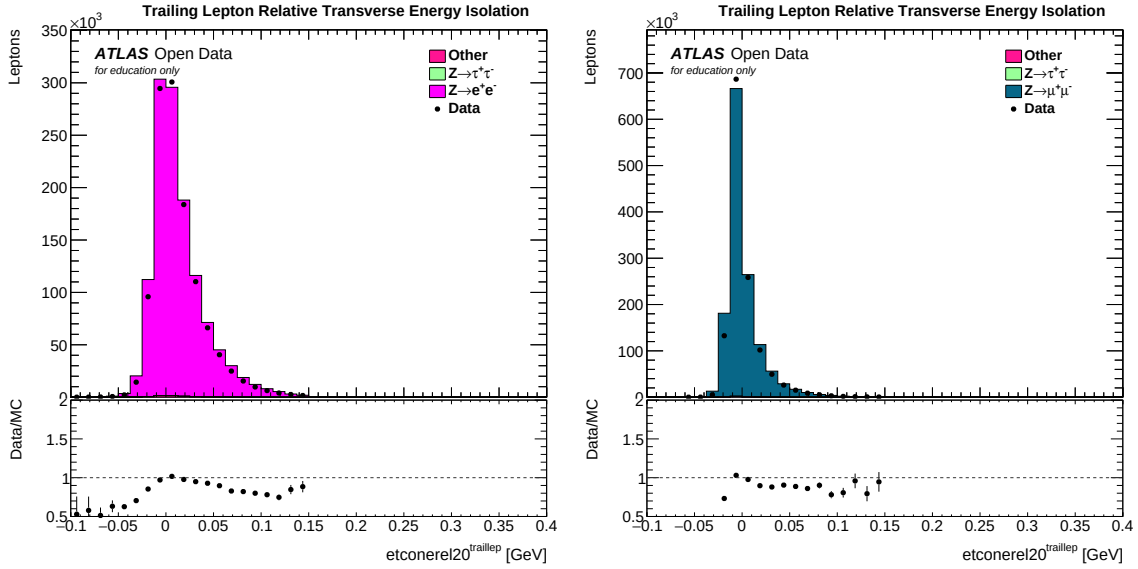
A ZZ analysis is interesting since it allows the study of triple boson couplings. Figure 15 for the invariant masses of the two Z-bosons in a ZZ analysis is to be contrasted



(a) Jet MV2c10 b-tagging weight.

(b) Jet mass.

Figure 12: Stacked histograms of further jet variables for a 13 TeV ATLAS Open Data Z analysis. The signal Monte Carlo samples are $Z \to e^+e^-$ and $Z \to \mu^+\mu^-$. "Other" includes backgrounds for W, low-mass Drell-Yan and single top. $Z \to \tau\tau$, $t\bar{t}$ and diboson are also backgrounds.



(a) $Z \to e^+e^-$

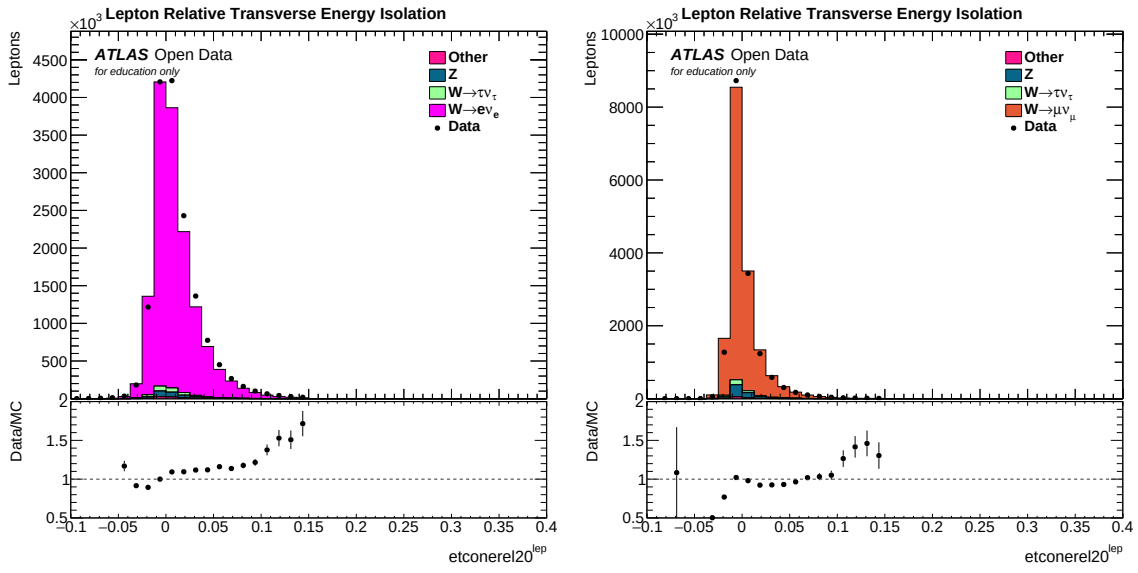
(b) $Z \to \mu^+\mu^-$

Figure 13: Trailing lepton etconerel20 against number of leptons. "Other" includes background processes for dibosons, low-mass Drell-Yan, W, single top and $t\bar{t}$. $Z \to \tau\tau$ is also background.

with Figure 5(a). Diboson production is rarer than single boson production, therefore the number of events passing ZZ selection requirements is much smaller than for a Z analysis.

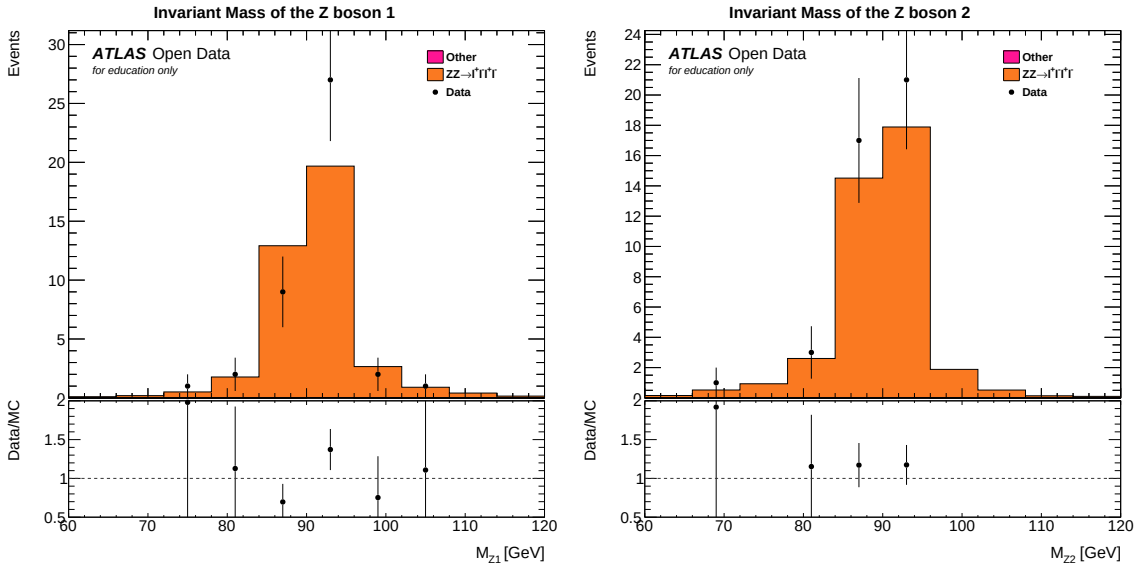
4 Discussion

Figures 5, 6 and 9 show that selection requirements have been correctly implemented in 13 TeV Open Data Z and W analyses. These analyses produce clear peaks in the mass



(a) $W \to e\nu_e$ (b) $W \to \mu\nu_\mu$

Figure 14: Stacked histograms of lepton etconerel20 against number of leptons, as defined by the etcone selection requirement in Table 3. The signal Monte Carlo samples are 13 TeV W + jets processes. "Other" includes backgrounds for dibosons, low-mass Drell-Yan, single top and $t\bar{t}$. $W \to \tau\nu_\tau$ and Z are also backgrounds. Both distributions are well matched between data and Monte Carlo in the bins where most events are present, after rescaling the number of vertices as described in Section 3.2.



(a) Z 1. (b) Z 2.

Figure 15: Histograms of Z candidate invariant mass against number of events, for a ZZ analysis. The signal Monte Carlo sample is $ZZ \to \ell^+\ell^-\ell^+\ell^-$, since the ZZ analysis searches for fully leptonic final states. "Other" represents the backgrounds $ZZ \to qq\ell^+\ell^-$ and $WZ \to \ell\nu_\ell\ell^+\ell^-$.

distributions of Figure 5. Figures 5, 6 and 9 also show that data and MC have been well matched for the most pertinent variables to 13 TeV Open Data Z and W analyses. Since there are no systematic uncertainties on MC included yet, a complete comparison between data and MC has not been possible. Reference [30] suggests a 5% theoretical uncertainty on MC and a 2.1% experimental uncertainty on data luminosity.

4.1 Next steps from 1st semester & improvements over 8 TeV Open Data

In the first semester report, a number of next steps for progression were laid out. Modifying selection requirements, plots of data subsets and running over all data on the grid [67] are done. Running over all MC is partly done, but no Beyond Standard Model processes have been included yet. GitHub documentation has been maintained throughout the second semester. Photon inclusion has been started by looking into datatypes that contain photons. Making plots of smaller phase space regions has not yet been started. $HH \rightarrow bb\tau\tau$ dependency removal and implementing the CxAOD and TupleMakers in one directory have become low priorities in this second semester.

With an increase in centre of mass energy comes an increase in cross-sections. Releasing all 2015 data with integrated luminosity 3.2 fb^{-1}, compared to 1 fb^{-1} that formed the 8 TeV Open Data release, would be a further improvement. In documenting the 13 TeV Open Data production on GitHub, ATLAS members will be able to access and modify it for future purposes. The 8 TeV Open Data production code was not as well documented.

4.2 Further work

An integrated luminosity of 3.2 fb^{-1} would allow the study of SUSY [68], such as gluino [69] production. Gluinos decay to stop squarks [70], which then decay to dark matter, such as neutralinos [71]. Another interesting SUSY process would be chargino-neutralino [72] production. Large-radius jets [73] would allow the reconstruction of boosted Z-bosons, W-bosons and top quarks. Photon information would permit Higgs analyses through $H \rightarrow \gamma\gamma$. Systematics in the datasets would allow users to better quantify errors in their analyses. By including truth information, a more complete comparison between data and MC simulation will be possible.

5 Conclusion

This project makes a further contribution to the first public release of 13 TeV ATLAS Open Data. Improved Z and W candidate mass distributions have been obtained compared to those available from 8 TeV ATLAS Open Data. The 13 TeV ATLAS Open Data Z and W analyses have been thoroughly studied and validated, so would therefore be nearly ready to be released. A ZZ diboson analysis has also been conducted. 13 TeV ATLAS Open Data will be extended to include $t\bar{t}$ and $\gamma\gamma$ analyses.

References

[1] ATLAS Collaboration, *Journal of Instrumentation*, Volume 3, Issue 3, pages 1-380, 2008.
[2] Evans, L. & Bryant, P., *Journal of Instrumentation*, Volume 3, Issue 1, pages 1-151, 2008.
[3] Coblans, H., *Scientific World*, Volume 5, Issue 1, pages 4-8, 1961.
[4] ATLAS Collaboration, Borissov, G., et al, *Journal of Physics: Conference Series*, Volume 664, pages 1-8, 2015.
[5] Fonseca-Martin, T., et al, *Journal of Physics: Conference Series*, Volume 119, Issue 2, pages 1-10, 2008.
[6] ATLAS Collaboration, Onyisi, P., *Proceedings of the 35th International Conference of High Energy Physics*, pages 1-6, 2010.
[7] Anastasiou, C., Buehler, S., Herzog, F. & Lazopoulos, A., *Journal of High Energy Physics*, Volume 72, Issue 4, pages 1-9, 2012.

[8] ATLAS Collaboration, Riu, I., *8th International Workshop on Top Quark Physics*, Ischia, Italy, 14-18th September 2015.
[9] ATLAS Collaboration, *The European Physical Journal C*, Volume 76, Issue 5, pages 1-17, 2016.
[10] ATLAS Collaboration, Arnaez, O., *American Institute of Physics Conference Proceedings*, Volume 1200, pages 693-696, 2010.
[11] ATLAS Collaboration, Berglund, E., *Proceedings of Science*, Volume 84, pages 430-434, 2009.
[12] ATLAS Collaboration, *The European Physical Journal C*, Volume 77, Issue 7, pages 1-32, 2017.
[13] ATLAS Collaboration, Limbach, C., *Nuclear and Particle Physics Proceedings*, Volume 260, pages 195-198, 2015.
[14] Tomiwa, K., G., *Journal of Physics: Conference Series*, Volume 889, Conference 1, pages 1-5, 2017.
[15] ATLAS Inner Detector Alignment Community, de Renstrom, P. et al, *Nuclear Instruments and Methods in Physics Research A*, Volume 582, Issue 3, pages 800-805, 2007.
[16] Cornelissen, T. et al, *Journal of Physics: Conference Series*, Volume 119, Issue 3, pages 1-10, 2008.
[17] ATLAS Collaboration, Zolnierowski, Y. et al, *Nuclear Instruments and Methods in Physics Research A*, Volume 384, Issue 1, pages 230-236, 1996.
[18] ATLAS Collaboration, Di Girolamo, B. et al, *Nuclear Instruments and Methods in Physics Research A*, Volume 453, Issues 1-2, pages 233-236, 2000.
[19] ATLAS Collaboration, Deile, M. et al, *Nuclear Instruments and Methods in Physics Research A*, Volume 518, Issue 1-2, pages 65-68, 2004.
[20] ATLAS Collaboration, Bystrycky, J. et al, *Nuclear Science Symposium*, Anaheim, CA, USA, 1996.
[21] ATLAS Collaboration, *The European Physical Journal C*, Volume 77, Issue 5, pages 1-39, 2017.
[22] ATLAS Collaboration, Barnett, R. & Johansson, K., *Physics Education*, Volume 41, Issue 5, pages 432-436, 2006.
[23] Trumbull, D., Bonney, R., Bascom, D. & Cabral, A., *Science Education*, Volume 84, Issue 2, pages 265-275, 2000.
[24] CMS Collaboration, *Journal of Instrumentation*, Volume 3, Issue 4, pages 1-307, 2008.
[25] Alice Collaboration, *Journal of Instrumentation*, Volume 3, Issue 2, pages 1-216, 2008.
[26] LHCb Collaboration, *Journal of Instrumentation*, Volume 3, Issue 5, pages 1-187, 2008.
[27] Evans, M., O., "Enabling Open Science with the ATLAS OpenData project at CERN", Unpublished report, University of Manchester, pages 1-20, 2018.
[28] Ay, C. et al, *Journal of Physics: Conference Series*, Volume 219, Part 3, pages 1-8, 2010.
[29] ATLAS Collaboration, *The European Physical Journal C*, Volume 2018, Issue 78, pages 110-170, 2018.
[30] ATLAS Collaboration, *The European Physical Journal C*, Volume 77, Issue 361, pages 1-16, 2017.
[31] ATLAS Collaboration, *Physics Letters B*, Volume 3, Issue 1, pages 1-20, 2012.
[32] ATLAS Collaboration, *Physical Review D*, Volume 90, Issue 5, pages 1-35, 2014.
[33] Catani, S. et al, *Physical Review Letters*, Volume 108, Issue 7, pages 2001-2004, 2012.
[34] Kopp, J. & Lindner, M., *Physical Review D*, Volume 76, Issue 9, pages 1-8, 2007.
[35] Gang, C. & Lianshou, L., *Journal of Physics G: Nuclear and Particle Physics*, Volume 30, Issue 10, pages 1399-1406, 2004.
[36] ATLAS Collaboration, *Journal of Instrumentation*, Volume 11, Issue 4, pages 8-130, 2016.
[37] Golling, T. et al, *The European Physical Journal C*, Volume 72, Issue 4, pages 1-6, 2012.
[38] Dos Anjos, A. et al, *IEEE Transactions on Nuclear Science*, Volume 53, Issue 4, pages 2144-2149, 2006.
[39] ATLAS Collaboration, *Journal of High Energy Physics*, Volume 2012, Issue 11, pages 33-85, 2012.
[40] Weinberg, S., *Physical Review Letters*, Volume 19, Issue 21, pages 1264-1266, 1967.
[41] Gibson, A., *Nuclear Science Symposium Conference Record*, Dresden, Germany, 19th-25th October 2008.

[42] Akikawa, H. et al, *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, Volume 499, Issues 2-3, pages 537-548, 2003.
[43] Vukotic, I. et al, *Journal of Physics: Conference Series*, Volume 898, Issue 5, pages 1-7, 2003.
[44] Feynman, R., P., *Physical Review*, Volume 76, Issue 6, pages 749-759, 1949.
[45] ATLAS Collaboration, *Physical Review Letters*, Volume 116, Issue 10, pages 1-19, 2016.
[46] ATLAS Collaboration, *Physics Letters B*, Volume , Issue , pages 136-157, 2016.
[47] Parke, S., J. & Taylor, T., R., *Physical Review Letters*, Volume 56, Issue 23, pages 2459-2460, 1986.
[48] Particle Data Group, Patrignani, C. et al, *Chinese Physics C*, Volume 40, Issue 1, pages 11-1787, 2016.
[49] Oleari, C., *Nuclear Physics B-Proceedings Supplements*, Volumes 205-206, pages 36-41, 2010.
[50] Gleisberg, T. et al, *Journal of High Energy Physics*, Volume 2009, Issue 2, pages 1-62, 2009.
[51] ATLAS Collaboration, *Physics Letters B*, Volume 777, pages 91-113, 2018.
[52] ATLAS Collaboration, *Journal of High Energy Physics*, Volume 2017, Issue 86, pages 1-40, 2017.
[53] Drell, S. & Yan, T.-M., *Physical Review Letters*, Volume 25, Issue 5, pages 316-320, 1970.
[54] ATLAS Collaboration, *Journal of High Energy Physics*, Volume 2017, Issue 12, pages 24-92, 2017.
[55] ATLAS Collaboration, Krohn, O. et al, *American Physical Society Meeting*, Washington, DC, USA, 28th-31st January 2017.
[56] Yang, H.-J. et al, *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, Volume 574, Issue 2, pages 342-349, 2007.
[57] ATLAS Collaboration, *The European Physical Journal C*, Volume 76, Issue 11, pages 1-36, 2016.
[58] ATLAS Collaboration, *Physical Review D*, Volume 85, Issue 7, pages 1-39, 2012.
[59] ATLAS Collaboration, *The European Physical Journal C*, Volume 2012, Issue 72, pages 2062-2082, 2012.
[60] ATLAS Collaboration, *Physics Letters B*, Volume 712, Issues 4-5, pages 289-308, 2012.
[61] ATLAS Collaboration, *Physical Review D*, Volume 85, Issue 11, pages 1-21, 2012.
[62] ATLAS Collaboration, *Physics Letters B*, Volume 756, pages 228-246, 2016.
[63] Re, E., *The European Physical Journal C*, Volume 2011, Issue 71, pages 1547-1561, 2011.
[64] Brun, R. & Rademakers, F., *Nuclear Instruments and Methods in Physics Research A: Accelerators, Spectrometers, Detectors and Associated Equipment*, Volume 389, Issues 1-2, pages 81-86, 1997.
[65] `Python`, Version 3.6.2, Python Software Foundation, Beaverton, Oregon, USA.
[66] Polesello, A. & Tovey, D., *Journal of High Energy Physics*, Volume 2004, Issue 71, pages 1-12, 2004.
[67] Lamanna, M., *Nuclear Instruments and Methods in Physics Research A: Accelerators, Spectrometers, Detectors and Associated Equipment*, Volume 534, Issues 1-2, pages 1-21, 2015.
[68] Gervais, J.-L. & Sakita, B., *Nuclear Physics B*, Volume 34, Issue 2, pages 632-639, 1971.
[69] ATLAS Collaboration, *The European Physical Journal C*, Volume 72, pages 2174-2192, 2012.
[70] ATLAS Collaboration, *Journal of High Energy Physics*, Volume 2014, Issue 124, pages 1-65, 2014.
[71] ATLAS Collaboration, *Physical Review D*, Volume 88, Issue 11, pages 1-23, 2013.
[72] Baro, N. & Boudjema, F., *Physical Review D*, Volume 80, Issue 7, pages 1-24, 2009.
[73] Nachman, B., Nef, P., Schwartzman, A., Swiatlowski, M. & Wanotayaroj, C., *Journal of High Energy Physics*, Volume 2015, Issue 48, pages 75-91, 2015.

Appendix

The number of words in this document is 8725.
This document was last saved on 14/05/2018 at 22:53.

Table 6: Contents of 13 TeV ATLAS Open Datasets to use in the 13 TeV ATLAS Open Data analysis code. The content of these datasets have been modified from the 8 TeV ATLAS Open Data release. Superfluous branches not needed for the educational purposes of the data release have been dropped. Scalefactors correct for known differences between data and MC.

variable name	type	description
runNumber	int	run identifier
eventNumber	int	event identifier
mcWeight	float	simulated event weight
pvxp_n	int	number of primary vertices
vxp_z	float	primary vertex z-position
SF_Pileup	float	pileup reweighting scalefactor
SF_Ele	float	electron efficiency scalefactor
SF_Muon	float	muon efficiency scalefactor
SF_Btag	float	b-tag algorithm scalefactor
SF_Trigger	float	scalefactor to account for different operating efficiencies of triggers
SF_JVF	float	jet vertex fraction scalefactor
trigE	bool	whether trigger fired in electron stream
trigM	bool	whether trigger fired in muon stream
passGRL	bool	whether event passes Good Run List
hasGoodVertex	bool	whether event has ≥ 1 good vertex with $N_{\text{tracks}} > 2$
lep_n	int	number of preselected leptons
lep_trigMatched	vector<bool>	whether the lepton is the one triggering the event
lep_pt	vector<float>	lepton transverse momentum
lep_eta	vector<float>	lepton pseudorapidity
lep_phi	vector<float>	lepton azimuthal angle
lep_E	vector<float>	lepton energy
lep_z0	vector<float>	z-coordinate of track associated to lepton wrt primary vertex
lep_charge	vector<float>	lepton charge
lep_type	vector<int>	number signifying lepton type (e, μ, τ)
lep_ptcone30	vector<float>	scalar sum of track p_{T} in R=0.3 cone around lepton, not including lepton itself
lep_etcone20	vector<float>	scalar sum of track E_{T} in R=0.2 cone around lepton, not including lepton itself
lep_d0	vector<float>	d_0 of track associated to lepton at point of closest approach
lep_d0sig	vector<float>	d_0 significance of track associated to lepton at point of closest approach
met_et	float	transverse energy of missing momentum vector
met_phi	float	azimuthal angle of missing momentum vector
jet_n	int	number of selected jets
jet_pt	vector<float>	jet transverse momentum
jet_eta	vector<float>	jet pseudorapidity
jet_phi	vector<float>	jet azimuthal angle
jet_E	vector<float>	jet energy
jet_m	vector<float>	jet mass
jet_jvt	vector<float>	jet vertex tagging
jet_trueflav	vector<int>	simulated jet flavour
jet_truthMatched	vector<int>	whether jet matches simulated jet
jet_MV2c10	vector<float>	weight from algorithm based on Multi-Variate technique