PSSOH 2022, DOI: https://doi.org/10.5281/zenodo.7189233

Open Data from CMS at CERN: Status and Plans

Milos Dordevic[1] on behalf of the CMS Collaboration

1: Vinca Institute of Nuclear Sciences, National Institute of the Republic of Serbia, University of Belgrade, 11351, Vinca, Belgrade, Serbia
e-mail: milos.dordevic@cern.ch

Abstract

The CMS Collaboration at the CERN LHC has released to the public more than two petabytes of open data. The large parts of these datasets that were used in the data analyses have led to the discovery of the Higgs boson in 2012. This open data, apart from its originality and scientific value, has already facilitated public results produced by non-collaboration members, within the high energy physics or even from different fields, thus improving the knowledge exchange and supplementing the original research with new ideas and insights. The CMS open data is already used for educational and outreach purposes, through providing the hand-on exercises for CERN Masterclass and other events. A brief introduction to the LHC and the CMS Experiment is outlined, followed by the description of real-time data selection by the Trigger system and an overview of raw data to physics results pathways. Information is given on how to access the CMS open data using virtual machines and containers, followed by presenting in more detail an example of a simplified CMS analysis at the Trigger level. The possibilities to address an application of Machine Learning in high energy physics using the CMS open data are shown as well, concluding with the user feedback, and an opinion from Nature Physics.

Key words: [cern, cms, open, data, status]

1 Introduction

At the CERN Open Data portal [1], an interested subject can find more than two petabytes of open data released by the CMS Collaboration at the CERN Large Hadron Collider (LHC) [2]. These are the original datasets from high energy proton-proton collisions recorded by the CMS Experiment in 2010, 2011 and 2012. A large portion of these datasets, released now to the public view and scrutiny, was used to discover the Higgs boson in 2012 by the CMS Collaboration. According to the CMS Open Data policy [3], the CMS Collaboration has committed to release 100% of its analysable data within ten years of collecting it. The other LHC experiments, ATLAS, LHCb and ALICE, have also released their data to the public, with the corresponding usage policies [4, 5, 6]. The CMS Open Data is nowadays being used in scientific research, both within the high energy physics community [7, 8] and by non-collaboration researchers. It is also used for educational and outreach purposes all around the world, enabling them to promote the field and attracting young people to cutting-edge research being done at CERN.

Section 2 gives a brief overview of the LHC and CMS Experiment, presenting the basics of real-time data selection using the Trigger system, as well as the analysis pathways needed to convert the raw data to physics results. In Section 3, the motivation to release the CMS open data, along with the details of how to access it and examples of its usage, are outlined. Section 4 reports on the results published

using CMS open data and Section 5 gives a summary with plans for further usage and improvements.

2 Overview of the LHC and the CMS Experiment at CERN

The Large Hadron Collider (LHC) at CERN is in operation for more than a decade already, steadily delivering high intensity particle beam collisions at record breaking energies. The products of these high-energy collisions are being recorded by the most powerful ”cameras”, being the particle detectors, wrapped around the beam interaction points. The CMS detector is a multilayered, general-purpose detector designed primarily for discovery and characterisation of the Higgs boson, precision Standard Model measurements and Beyond Standard Model searches (e.g. Supersymmetry or Extra Dimensions). The particle beams colliding at the LHC are organised in bunches, with around 100 billion protons in each bunch at the design luminosity value. The bunch collisions happen at a rate of 40 MHz, while the production of the Higgs boson and potential new particles are very rare events, occurring at the rate of the order of 1 Hz or lower, respectively. Recording all the collision data would generate more than 50 terabytes each second and storing such an amount of data would be impossible and, in fact, not needed for the physics goals of the CMS experiment. Selective read-out of the particle collisions is performed in real time by the Trigger system [9]. The CMS Trigger system is implemented in two tiers, the Level-1 trigger (L1) based on custom-made fast electronics that reduces the rate to about 100 kHz, and the High Level Trigger (HLT) using a software running on computing farm, further reducing the rate to about 1.3 kHz, saving full event content and performing prompt reconstruction. Figure 1 presents a graphic image of the LHC along with the map of the area. Figure 2 shows a 3D model of the CMS experiment, revealing its very complex, multilayered internal structure.

The raw data that comes out of the online system is followed by the event reconstruction which refines this data, also applying calibrations and performing the creation of higher-level physics objects. This procedure is usually referred to as skimming (also slimming, thinning, etc.), which reduces the data size, but also increases its usability, both by reducing the number of events and compressing the event format. Event information from each step in the simulation and reconstruction chain is logically grouped into what is called a data tier. Examples of data tiers include RAW and RECO [12]. RAW represents the detector data after online formatting, the L1 trigger result, the result of the HLT selections and potentially some of the higher level quantities which are calculated during HLT processing. It would be impossible to perform physics analysis using raw primary datasets, hence the RECO data tier is made, representing first level of the event reconstruction. Detector-level information is passed through the various reconstruction algorithms. Tracking, vertexing, and Particle Flow are performed in this step and then some basic physics object collections are created. This step is very CPU intensive. The further reduction to physics objects is performed when creating the AOD (Analysis Object Data) data tier, where only some hits and other detector-level info is kept, making the physics object collections as priority and retaining some supporting information from RECO. The AOD data tier can be further skimmed into more compact data tiers, which will be described in Section 3.3. An example of tracking, ECAL and HCAL-based objects corresponding to basic data tiers is shown in Figure 3.

3 The CMS open data: motivation, access and examples

3.1 The motivation for releasing the CMS open data to the public

There is a strong motivation and multiple reasons to release the CMS data for public use. First and foremost, science results should be inclusive and open to everyone. By making the CMS data open, more people become engaged with research. There is an important educational aspect to it, since it provides an ideal platform for teaching and exercises, thus attracting students to particle physics. It is also another way to directly return something to society. The CMS open data improves communication and facilitates the exchange of knowledge with the researchers first within the same, particle physics

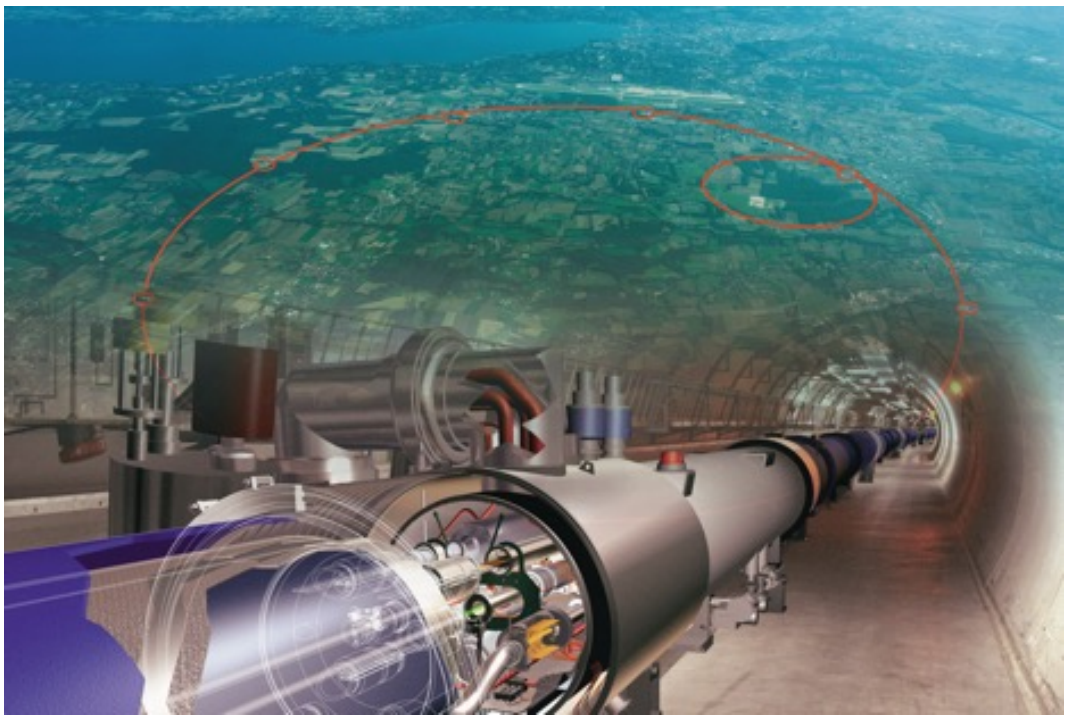

Figure 1: The Large Hadron Collider at CERN.

research field, but also from other scientific areas, such as machine learning and data science topics.

3.2 Data levels and access: virtual machines and containers

The starting point for the usage of the CMS open data is the webpage [1] where a user can browse more than two petabytes of particle physics data, allowing to inspect different datasets, environments, software and many examples, each well documented. The CMS open data has the following levels:

- Level 1 : Open access publication and additional numerical data
- Level 2 : Simplified data for Outreach and Education
- Level 3 : Reconstructed data and the software to analyse them
- Level 4 : Raw data, and the software to reconstruct and analyse it

as defined in the CMS data policy [3], starting either from raw experimental or simulated data, going through the reconstructed data and the datasets with the higher level of abstraction generated by analysis workflows, and ultimately all the way to data which are represented in scientific publications. Each of these levels enables different opportunities for long-term re-use, but also poses different challenges for data preservation. Level 1 corresponds to publications, with additional documentation provided, in order to put the results in context and understand the analysis procedures, some additional numerical data which did not or could not appear in the publications. Examples of these would be the cross sections of different physics processes, given as a function of multiple variables. The Level 2 includes simplified data formats, such as, for example, multi-dimensional distributions of analysis variables, or the four- vectors of particles or jets, energy clusters and tracks. These could be re-used immediately for theory interpretations, some limited physics analysis, but also for educational and outreach purposes. Level 3 represents the reconstructed data and simulations, released together with the corresponding

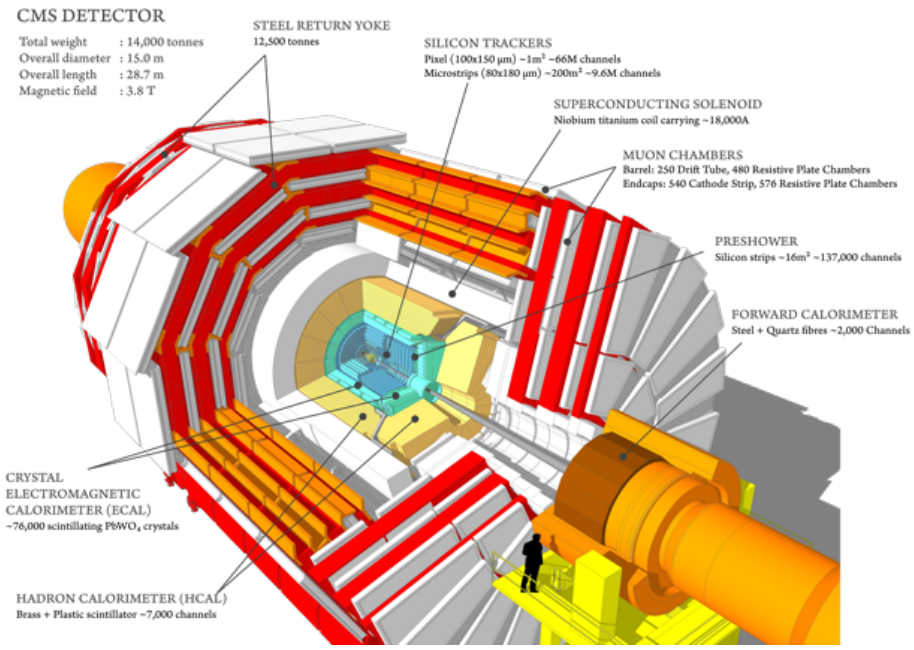

Figure 2: The CMS Experiment at CERN LHC.

software, analysis workflows and documentation that is needed to access this data, reproduce published analyses, or to perform new analyses without requiring re-reconstruction of the data or running the new simulations. Level 4 is the raw data and the software and documentation needed to access, reconstruct and analyse this data. This is the most complex level, which also requires the highest computational effort. An example of using Level 3 CMS open data will be presented in Section 3.4.

Access to the CMS open data is provided via the usage of the CERN Virtual Machine (VM) [10], or the Docker container [11], providing the usage of the CMSSW [12] environment and the ROOT [13] framework. The setup of the CERN VM is enabled through the usage of the Oracle VM VirtualBox software, a free, open source and multiplatform application to run virtual machines. It provides a base for the operating system which is compatible with the CMSSW environment needed to run the analysis on the open data. Following the download of the required software, the VM file size on the host machine is at the order of a few GBs. Docker is a free, commercial implementation of a software container, as an alternative to the VM images. There are different kinds of container images available at the Docker Hub and CERN GitLab, from the light-weight ones to the images containing full CMSSW installation that is at the order of several GB, allowing them to preserve complete CMS physics analyses. Both VM and Docker organisation and structures compared, are shown in Figure 4.

3.3 Data format of CMS open data

Most of the CMS open data is published in the Analysis Object Data (AOD) format which includes serialised C++ objects requiring the CMSSW environment and ROOT framework to be read. Each of the AOD events holds about 500 kB of information, resulting generally with large files. The CMS Collaboration has thus developed derived data formats called MiniAOD [14], and its successor NanoAOD [15]. While the MiniAOD is similar to AOD and also storing serialised C++ objects, NanoAOD is based on storing the basic types such as floats, integers and arrays thereof. Table 1 presents an example of the variable content of the muons collection, embedded in the NanoAOD data format.

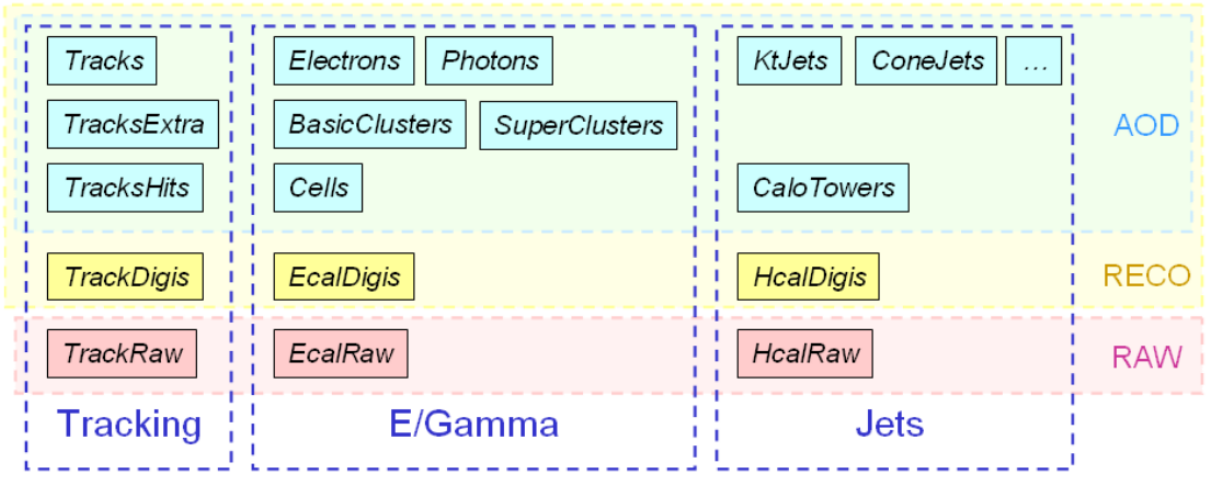

Figure 3: Tracking, ECAL, and HCAL-based objects corresponding to different CMS data tiers.

Table 1: Data format of a muon collection in the NanoAOD.

Variable	Type	Description
nMuon	unsigned int	Number of muons in this event
Muon_pt	float[nMuon]	Transverse momentum of the muons
Muon_eta	float[nMuon]	Pseudorapidity of the muons
Muon_phi	float[nMuon]	Azimuth of the muons
Muon_mass	float[nMuon]	Mass of the muons
Muon_charge	int[nMuon]	Charge of the muons (either 1 or -1)

3.4 Examples of usage: Higgs boson and CMS Trigger System

The CMS open data has provided an example [16] of a strongly simplified reimplementation of parts of the original CMS Higgs to four lepton analysis, published in [17]. This example enables different levels of complexity, from the one useful for educational purposes, to the more complex one requiring at least some minimal understanding of the content of the paper [17]. The example uses legacy versions of the original CMS datasets in the AOD data format, but slightly different than in the original publication. Many of the data selection cuts are the same, however this CMS open data analysis is still a much simplified reimplementation of the original CMS Higgs to four leptons analysis. The four lepton invariant mass spectrum in $H \rightarrow 4l$, as obtained using CMS open data, is outlined in Figure 5.

An example [18] needed to extract the Trigger information from the CMS Open/Legacy data is provided, with its implementation in C++ code and configuration in Python [19]. In this example one can find the following analysers, each performing some of basic Trigger analyses using CMS open data:

- GeneralInfoAnalyzer : several C++ snippets on how to access trigger information such as metadata, prescales, module information, etc.
- ModuleInTriggerAnalyzer : shows how to dump all the modules for a specific trigger and/or obtain the last active module (filter) of a trigger.
- TriggerMatchingAnalyzer : how to match reconstructed tracks to objects that fired a trigger (or possible set of triggers) that contain a specific module.
- TriggerSimplePrescalesAnalyzer : use wildcards to access different versions of the same trigger, check their L1 and HLT prescales, and whether the trigger accepted the event or not.

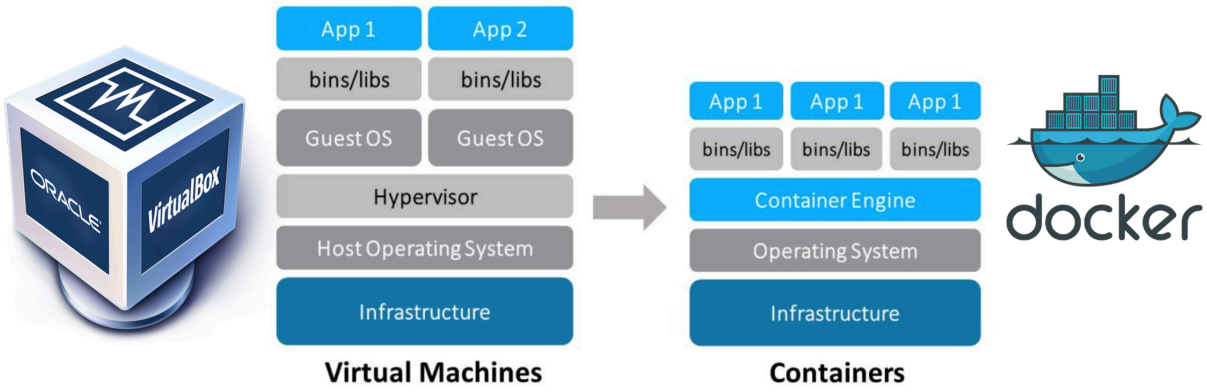

Figure 4: Oracle VM VirtualBox (left). Docker software container (right).

Here we will focus on the ModuleInTriggerAnalyzer and present an example of running it, together with the required commands and the obtained output, using the CERN VM. Example of a typical HLT path used in CMS Trigger is shown in Figure 6. After setting up the VM, in the terminal, the list of commands is required to be typed in to run the analyser. These commands include the creation of the CMSSW environment, downloading of the corresponding analyser from the github repository, navigation to the directory where the example is installed, linking required databases for replicating the analysis conditions and finally running the examples. The described workflow is the following:

```
cmsrel CMSSW_5_3_32
cd CMSSW_5_3_32/src/
cmsenv
git clone -b 2011 git://github.com/cms-legacydata-analyses/TriggerInfoTool.git
cd TriggerInfoTool
cd {packagename}
scram b
ln -s python/{configname} .
ln -sf /cvmfs/cms-opendata-conddb.cern.ch/FT_53_LV5_AN1_RUNA
    FT_53_LV5_AN1
ln -sf /cvmfs/cms-opendata-conddb.cern.ch/FT_53_LV5_AN1_RUNA.db
    FT_53_LV5_AN1_RUNA.db
ls -l
ls -l /cvmfs/
cmsRun {configname} > full.log 2>&1 &       (checking w/ "tail -f full.log")
```

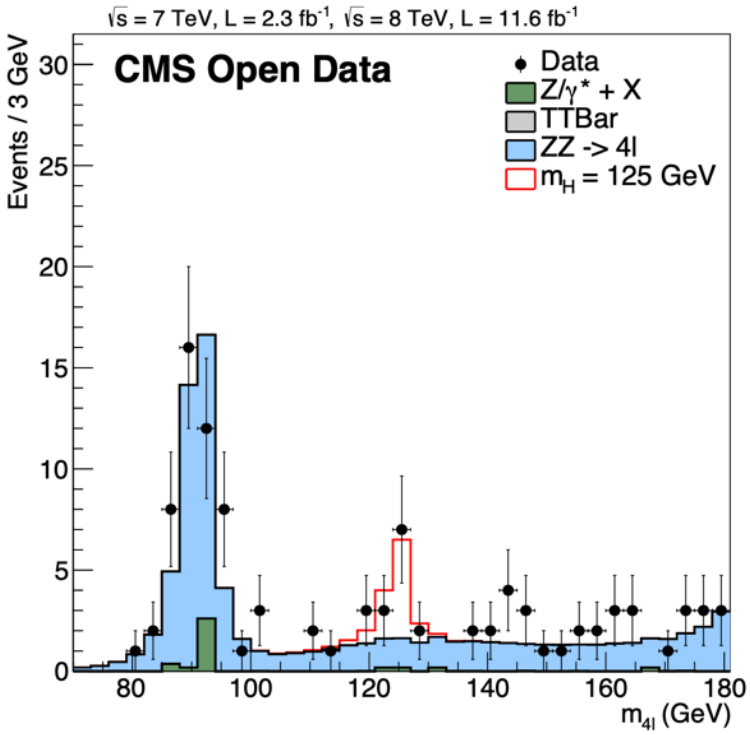

Figure 5: The four lepton invariant mass in the $H \rightarrow 4l$ analysis using CMS open data.

Following the execution of these commands, the corresponding output file is provided. The selection of this output, listing all the modules contained in the Trigger path HLT_Jet190_v6, as well as the event processing information with the selected event range are printed, as shown in Figure 7. The last module that has fired in this trigger path for each event is highlighted with the corresponding colour. This is a basic example of Trigger analysis, but such evaluations are very often made in CMS Trigger.

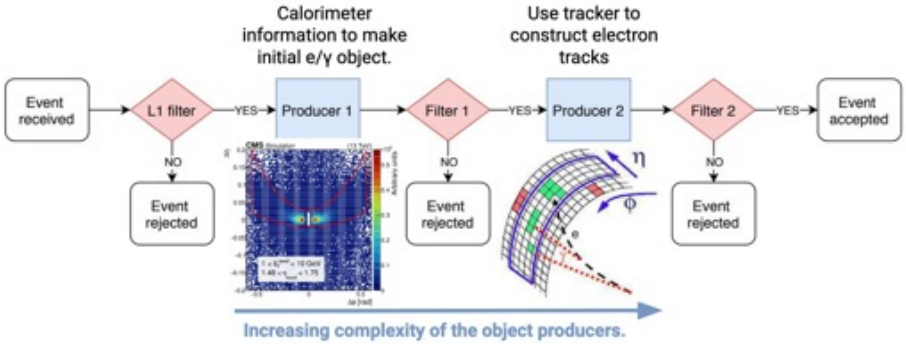

Figure 6: The example of the HLT path, showing increasing complexity of the object producers.

3.5 CMS Open data for machine learning in high energy physics

The CMS Open Data has also addressed the continuously growing application of machine learning (ML) to various challenges in high energy physics [20]. In the paper [21], it is clearly outlined that collaboration with data science and ML community is taken as a high priority in helping to advance the application of state-of-the-art algorithms to high energy physics. The ML datasets, derived from millions of CMS simulation events, focus on solving a number of problems in particle identification, tracking and distinguishing between multiple collisions in each bunch crossing (pileup). Reconstructed data and simulations released are from the CASTOR calorimeter, used by CMS in 2010,

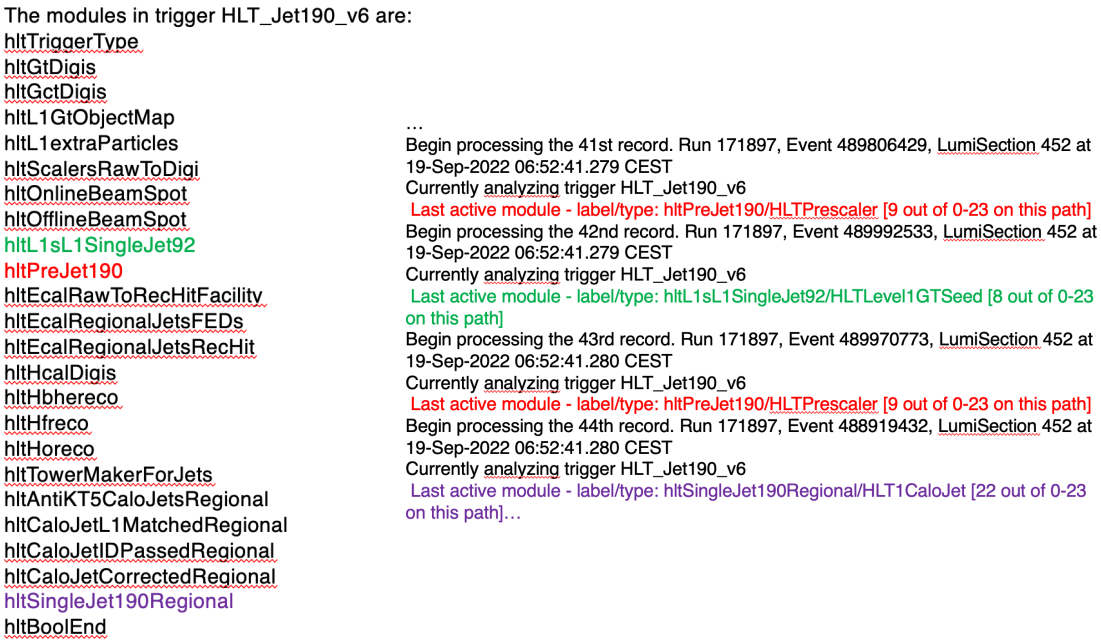

Figure 7: The output of running the ModuleInTriggerAnalyzer with CMS open data. The modules contained in trigger path HLT_Jet190_v6 (left). Output content showing event processing (right).

representing the first release of data from the very-forward region of CMS, with full instructions on access and usage.

4 Published results using CMS Open data and users feedback

There are already several scientific papers published with CMS open data. Some examples are the papers [22] and [23] which studied the QCD splitting functions and jet substructure, respectively, both being very common subjects of studies within the real LHC experiments. These are very clear examples of how the CMS open data can be used by theorists to extract (and publish) the valuable physics information. In the paper [23], the jet transverse momentum was extracted like in many original CMS analyses, and this is shown in Figure 8. We will not this time go into details of this paper. However, there are some very interesting lessons that the authors of this paper have kindly shared with the public, based on their own experience with using the CMS Open Data, quoting them precisely:

- "We then converted AOD files into a text-based MIT Open Data (MOD) format to facilitate the use of external analysis tools."

- “From the physics perspective, our experience with the CMS Open Data was fantastic”
- “From a technical perspective, though, we have encountered a number of challenges”

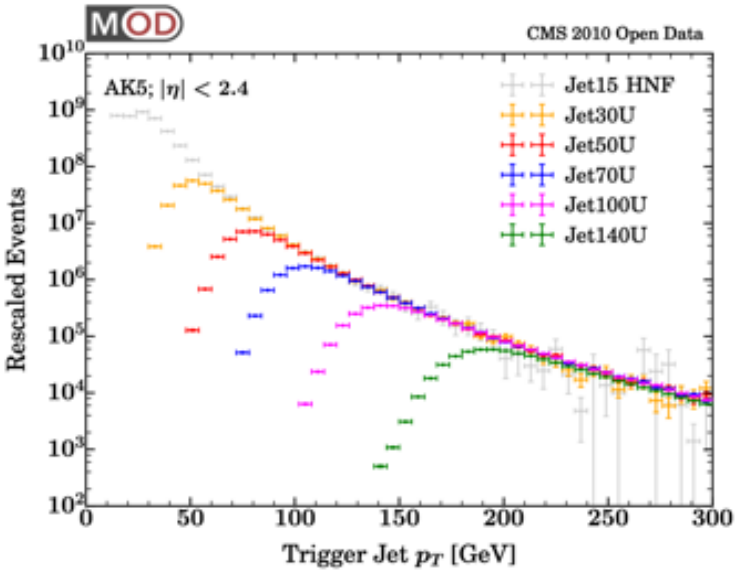

Figure 8: Leading jet transverse momentum spectrum, as obtained from CMS open data.

There is also an opinion article on the CERN open data published in Nature Physics [24]. Here are a few quotes from this article, that the author of this manuscript found to be the most interesting to cite:

- “only those who spent years building the experiment have earned quick access”
- “other scientists can analyse the data while LHC is still running, testing unconventional strategies”
- “public data can complement the overall research effort”

These quotes stressed some of the most important points related to releasing the CMS data to the public. Even though some of the users have had difficulties in parsing the available data format, faced complexity of the CMSSW software environment or sometimes found it challenging to browse through the documentation, all the analysts who used open data have highlighted the importance of the research-like quality of the data released by the CMS Collaboration and rooted for its continuation. User experience and feedback are ranked among the most important for the future of CMS open data.

5 Summary

The CMS Collaboration is leading the Open Data effort within the LHC experiments at CERN. The services to locate, browse and detailed descriptions on how to use the CMS Open Data are all provided and broadly used for education, outreach and scientific purposes. The means of using Virtual Machines and Docker containers have allowed access to the original CMS data, with the possibility, in the latter case, even to preserve a full CMS analysis. In this paper, examples of performing a simplified Higgs to four leptons CMS analysis, as well as how to access and run the basic CMS Trigger analysis, are presented, with the latter also specifying the user commands and the output. The use of CMS Open Data is already spreading throughout the scientific community, mainly in high energy physics theory, but also in other related disciplines such as machine learning and data science. According to the CMS open data policy, the public release of new CMS data is becoming imminent and also improvements to the documentation and use are continuously being made.

References

[1] CERN, The CERN Open Data portal, https://opendata.cern.ch.
[2] The CMS Collaboration, JINST 3 (2008) S08004.
[3] The CMS Collaboration, CMS preservation policy, 10.7483/OPENDATA.CMS.7347.JDWH.
[4] The ATLAS Collaboration, ATLAS preservation policy, 10.7483/OPENDATA.ATLAS.T9YR.Y7MZ.
[5] The LHCb Collaboration, LHCb preservation policy, 10.7483/OPENDATA.LHCb.HKJW.TWSZ.
[6] The ALICE Collaboration, ALICE preservation policy, 10.7483/OPENDATA.ALICE.54NE.X2EA.
[7] A. Tripathee, W. Xue, A. Larkoski, S. Marzani, J. Thaler, Jet substructure studies with CMS open data, Phys. Rev. D 96, 074003 (2017).
[8] C. Cesarotti, Y. Soreq, M.J. Strassler, J. Thaler, W. Xue, Searching in CMS open data for dimuon resonances with substantial transverse momentum, Phys. Rev. D 100, 015021 (2019).
[9] The CMS Collaboration, The CMS trigger system, JINST 12 (2017) P01020.
[10] CERN, The CERN Open Data portal, http://opendata.cern.ch/docs/cms-virtual-machine-2011.
[11] CERN, The CERN Open Data portal, http://opendata.cern.ch/docs/cms-guide-docker.
[12] The CMS collaboration, CMS Offline Software, https://github.com/cms-sw/cmssw.
[13] R. Brun, F. Rademakers, ROOT - An object oriented data analysis framework (1997).
[14] G. Petrucciani, A. Rizzi,C. Vuosalo, Mini-AOD: A New Analysis Data Format for CMS (2015), http://dx.doi.org 10.1088/1742-6596/664/7/072052.
[15] A. Rizzi, G. Petrucciani, M. Peruzzi (CMS Collaboration), EPJ Web Conf. 214, 06021, 6 p, (2019).
[16] CERN, The CERN Open Data portal, http://opendata.cern.ch/record/5500.
[17] The CMS Collaboration, S. Chatrchyan et al., Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC, Phys.Lett.B 716 (2012) 30-61.
[18] CERN, The CERN Open Data portal, http://opendata.cern.ch/record/5004.
[19] CERN, The CERN Open Data portal, https://github.com/cms-opendata-analyses/TriggerInfoTool/tree/2011.
[20] CERN, The CERN Open Data portal, https://home.cern/news/news/knowledge-sharing/cms-releases-open-data-machine-learning.
[21] K. Albertsson et al., Machine Learning in High Energy Physics Community White Paper, arXiv:1807.02876.
[22] A. Larkoski et al., Exposing the QCD splitting functions with CMS open data, Phys. Rev. Lett. 119, 132003 (2017).
[23] A. Tripathee et al., Jet Substructure Studies with CMS Open Data, Phys. Rev. D 96, 074003 (2017).
[24] M. Strassler, J. Thaler, Slow and steady, Nature Physics 15, 725 (2019).