EPJ Web of Conferences **245**, 06020 (2020) https://doi.org/10.1051/epjconf/202024506020
CHEP 2019

The CMS approach to Analysis Preservation

Lara Lloret Iglesias[1,*] on behalf of the CMS Collaboration.

[1]Institute of Physics of Cantabria – CSIC, Avda. de los Castros S/N 39005, Santander (Spain)

Abstract. The CERN analysis preservation portal (CAP) comprises a set of tools and services aiming to assist researchers in describing and preserving all the components of a physics analysis such as data, software and computing environment. Together with the associated documentation, all these assets are kept in one place so that the analysis can be fully or partially reused even several years after the publication of the original scientific results. An experiment-specific submission and retrieval interface has been developed for the CMS Collaboration. It integrates with the CMS internal analysis registry (CADI) to capture all analyses with basic information, complemented with a detailed submission form for full information. The CMS data aggregation system (DAS) is interfaced to the deposit form to assist in filling in exact dataset names used in the analysis to ensure searchability. Efforts are ongoing to describe physics content for an intelligent retrieval, and to interface with container solutions for full reproducibility for selected test cases.

1 Introduction

High Energy Physics experiments are often unique machines of their kind. This peculiarity, makes specially valuable the preservation of the data and all the different assets and pieces of information needed to get the final physics results. The CMS experiment [1] is a notable example of this: a unique and extremely complex machine that is constantly updated to achieve certain requirements in terms of energy, precision and accuracy. This means that, not only it is important to preserve the datasets and the analysis workflow, but also to somehow encapsulate all the knowledge around it, including the machine conditions, calibration parameters, and anything that can be useful for the full understanding and reproducibility of the results. The high complexity of the analyses create major challenges in terms of capturing and preserving them. Much of the information is readily available, and even obvious, at the time of the immediate data analysis but, due to the amount of details to take into consideration, it is easily forgotten and potentially lost without the dedicated tools to preserve it. This motivates the need, not only to keep the information safe, but also to design preservation tools to easily find it and access to it in a centralized manner. In parallel to this internal demand, there is also an increasing demand from outside of the collaboration asking for a comprehensive set of tools allowing knowledge preservation and aiming either reuse or even fully reproducibility of the research results. With these goals in mind, these proceedings summarize the current efforts in CMS regarding the preservation of the analysis assets. The document is structured as follows: the first section covers the reinterpretation efforts done by the CMS Collaboration. The second section goes over the Analysis Preservation Portal (CAP) and its different

*e-mail: lara@cern.ch

functionalities nowadays. Finally, the last section deals with REANA, a platform for reproducible research data analysis based on docker containers and its future integration with the CAP system.

2 Reinterpretation effort at CMS

In the CMS publications, the results of new physics searches are usually interpreted in terms of a subset of models used as benchmarks for the sensitivity of the search. These searches can be later re-interpreted to provide constraints on other models, that were not included in the original publication. Also, the results of a certain analysis can be combined with other measurements or searches from the same or a different experiment to interpret the result in terms of a more complete model. The CMS Collaboration is making an enormous effort by providing a meaningful set of tools and information on the published analyses with the goal of easing their re-interpretability. The central place for accessing the material is the CMS public results webpage [2]. Each analysis has its own webpage on which one can find:

1. All figures and tables that are part of the publication
2. Additional figures and tables
3. Links to the HEPData [3] entry including Rivet [4] analysis if available

The information provided in HEPData are mostly figures and tables, covering thus the easiest part in an analysis preservation workflow. Often in particle physics, the results to be preserved can be formulated generically (e.g. as a cross section limit) by parametrizing certain free parameters like, for instance, the resonance mass. Challenges in term of reproducibility arise when considering, for instance, different resonant widths or production mechanisms.

2.1 Object efficiencies

For standard physics objects such as leptons, the selection efficiencies are provided as material allowing for re-interpretation. These efficiencies present the advantage that they can be used directly in Delphes [5], a detector response simulation framework widely used by the Collaboration. These efficiencies can be used to some extent also for more complex objects than leptons, such as object with substructure like jets. However, uncommon objects that rely on specialised reconstruction, such as displaced vertices, are still a challenge. An example of muon efficiencies from a SUSY analysis can be seen in figure 1.

2.2 Simplified likelihood

The observed data are interpreted using a likelihood formalism. The calculation of this likelihood implies that the independent physical uncertainty sources are well known, accounting for the free parameters of the likelihood. This presents a problem in terms of what to preserve for the re-interpretation of the results since, in order to get the inputs needed for the likelihood calculation, it is necessary to run the full analysis machinery on the new signal samples to be considered. The simplified likelihood approach allows to easily recalculate final limits changing the signal hypotheses without the need of having access to the whole analysis workflow. In this simplified approach, the free parameters θ are taken as the deviation from the central value in each bin under the assumption of a Gaussian distribution for the bin content. The equiation for the simplified likelihood is the following:

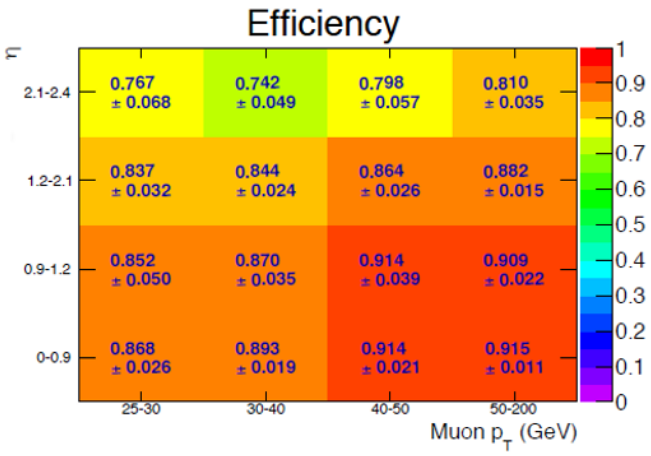

Figure 1. Example of muon efficiencies from a SUSY analysis as a function of the momentum (p_T) and the pseudorapidity (η).

$$\mathcal{L}(\mu,\theta) = \prod_{i=1}^{N(bins)} \frac{(\mu \cdot s_i + b_i + \theta_i)^{n_i} e^{-(\mu \cdot s_i + b_i + \theta_i)}}{n_i!}$$

This simplified approach implies that, for obtaining the likelihood, one just needs to calculate the signal yields in each bin s_i. The available re-interpretation material provides the background yields (b_i) and its uncertainties (θ_i) together with the data yields and the covariance matrix (V). An example of V for an exotica search analysis [9] in final states with an energetic jet or a hadronically decaying W or Z boson and transverse momentum imbalance is provided in figure 2.

A full reinterpretation workflow has already been successfully tested in a search for physics beyond the standard model in events with high-momentum Higgs bosons and missing transverse momentum [8].

2.3 Additional material

The datacards used for the signal extraction using the CMS internal statistical combination tool (Higgs Combine [11]) are also preserved for a large number of analyses, including event distributions and systematic variations. The preservation of these datacards allows reusing the results to perform statistical combination and summary plots in an easy manner. Furthermore, the CMS internal analysis management tools store links to the analysis twiki (mainly for review), links to pre-approval and approval presentations, analysis notes and, once the paper has been published, the publication metadata (DOI, arXiv, HEPData, Rivet, CDS, data set used). Finally, the development of an analysis description language capable of describing the contents of an analysis in a standard and unambiguous way, independent of any computing framework, is already being studied. The full analysis preservation would imply having all this information accessible just from one single place, indexed and searchable by final state, triggers, datasets, etc. This is where the CERN Analysis Preservation portal comes into play.

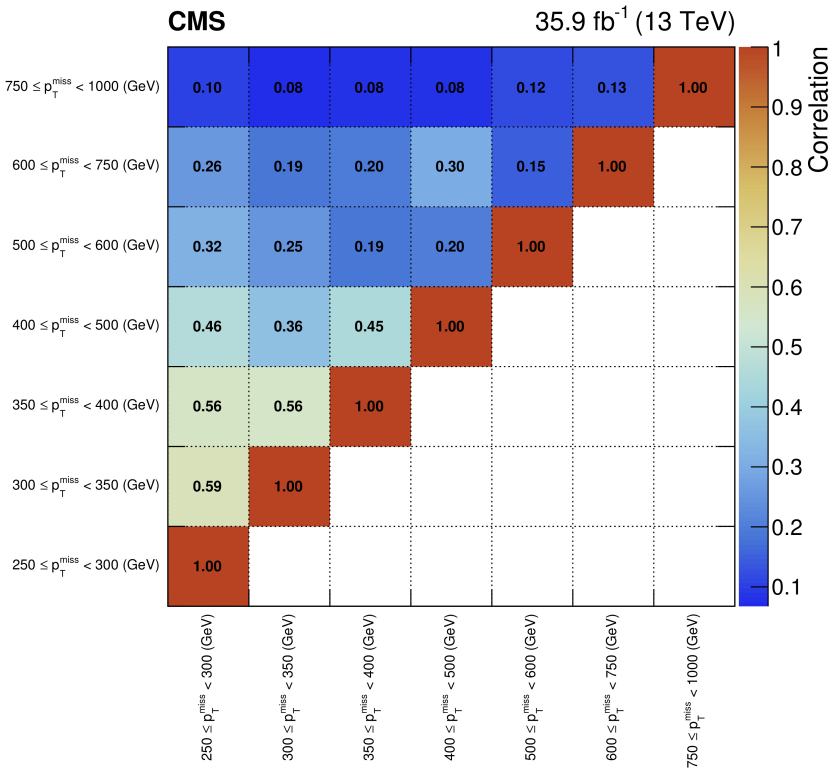

Figure 2. Example of the V matrix for an exotica search analysis in final states with an energetic jet or a hadronically decaying W or Z boson and transverse momentum imbalance. It shows the correlations between the predicted background yields in all the missing E_T bins for one of the signal regions of the analysis. The boundaries of the E_T bins, expressed in GeV, are shown at the bottom and on the left.

3 The Analysis Preservation Portal

The CERN Analysis Preservation Portal (CAP) [10] is a service for the four LHC Collaborations at CERN developed to address the need for the long-term preservation of all the digital assets and associated knowledge in the data analysis process, in order to enable future reproducibility of research results. This effort is run by CERN Scientific Information Services with the help from the different experiments. The portal is still in beta phase, but already providing many useful functionalities. Currently it is possible to automatically import information from CMS analysis management system (CADI) and to add and edit information in the portal using a command line client that allows to create/edit JSON files with the analysis information and submit it directly from the terminal, propagating it to the web portal

Apart from the information automatically updated from CADI, other information can be added such as analysis datasets, triggers, final states, statistical treatment or systematic uncertainties details, just to name a few.

It is possible to connect an external account (Github, CERN Gitlab, ORCiD, Zenodo…) with the CAP account, to automate tasks and content submission. One can just add the current repository content from the tarball or create a connection (webhook) so that everytime something is changed, the CAP is automatically updated. It also includes a harvester tool so

that the documents in CDS associated to the analysis can be directly attached to the CAP entry.

All the information stored in CAP is indexed and searchable, allowing to perform multiple queries. This implies that one can easily retrieve, for instance, all the analyses using some dataset/trigger and with a certain final state.

4 REANA

REANA [6] is a platform for reproducible research data analysis based on docker containers that encapsulate the analysis environment, software and workflow in order to run transparently for the user. The midterm goal is to integrate REANA with the CAP services into a single workflow, allowing the user to run an analysis in REANA launching it directly from the preserved assets in CAP maybe just with some small modifications. The generated output can then be automatically stored in CAP. Since the CMS software (CMSSW) has already been wrapped in a docker image, any analysis using mostly CMSSW can run on this docker and be then easily integrated into REANA. There are already several REANA examples that can be taken as reference in the REANA GitHub repository [7].

5 Conclusions

The CMS Collaboration is making an effort to preserve analyses and make them reinterpretable. Several tools exist already: Public webpages (including additional material, efficiencies, simplified likelihoods), HEPData (Rivet), CERN Analysis Preservation Portal and REANA among others. The next big step will be preserving the analysis implementation by integrating as many of these services as possible into a same workflow.

References

[1] CMS Collaboration, JINST 3 S08004 (2008)
[2] http://cms.web.cern.ch/news/cms-physics-results
[3] https://www.hepdata.net/
[4] https://rivet.hepforge.org/
[5] https://cp3.irmp.ucl.ac.be/projects/delphes
[6] http://www.reanahub.io/
[7] https://github.com/reanahub/reana
[8] CMS Collaboration, "Search for physics beyond the standard model in events with high-momentum Higgs bosons and missing transverse momentum in proton-proton collisions at 13 TeV." Physical review letters 120.24 (2018): 241801.
[9] CMS Collaboration, "Search for new physics in final states with an energetic jet or a hadronically decaying W or Z boson and transverse momentum imbalance at s= 13 TeV." Physical Review D 97.9 (2018): 092005.
[10] https://analysispreservation.cern.ch/
[11] https://github.com/cms-analysis/HiggsAnalysis-CombinedLimit