Bachelor Thesis Proposal : The Study of Physics Object Measurements using Tag-and-Probe Method With CMS Open Data

Siew-Yan Hoh

May 12, 2022

1 Executive Summary of the Research Proposal

Precision measurements of standard model (SM) processes at the Large Hadron Collider (LHC) made tremendous progress in recent years [1, 2, 3, 4, 5]. Remarkably, the per-mil precision archieved is attributed to the technological advancement of Monte Carlo (MC) generators and the detector calibration technique used to improve the modelling of high-energy hadron collisions and the instrumental systematic effects. It is worth mentioning that during RUN-I, the archieved energy resolution of photon reconstructed from the CMS (Compact Muon Selenoid) detector is between 1% and 3% [6], making the $H \to \gamma\gamma$ decay one of the golden decay channels for the Higgs boson discovery [7].

Under the CMS data policy [8], the CMS experiment has periodically released research-grade datasets (MC and data) to the public domain; allowing scientists outside of the collaboration to study and understand the detector performance, and possibly to exploit the scientific potential of these data. The open data initiative also provide a fertile ground to nurture and train students to perform data analysis on measuring physics processes and physics object performance, thus forging and installing High Energy Physics (HEP) data analysis skill capability among young students in Malaysia.

One of the instrumental systematic effects are originating from the Physics Objects used in the measurement; the Physics Objects are reconstructed and identified via several sub-detection system in the CMS detector. Depending on the the type of physics objects, thier reconstruction, identification as well as selection efficiencies are affected by the background processes associated with it.

This study will demonstrate the application of common HEP data analysis method, the Tag-and-Probe technique [9], to study the Physics Objects performance using CMS open data collected at the centre of mass energy $\sqrt{s} = 8$ TeV, corresponding to the integrated luminosity of 1.8 fb^{-1}.

The objectives of the study are to demonstrate the possibility to performing HEP data analysis using CMS open data on the current computational setup; to study the selection efficiency of the Physics Object; to derive the scale factors and the associated uncertainties to account for the selection inefficiency on MC modelling; and to validate and provide those Physics Object's scale factors for full-fledge physics process measurement in the near future.

The expected outcomes of this study are the successful demonstration of the current setup is capable to perform offline HEP data analysis using CMS open data, and to provide outlook on the improvement in term of computational resources for long term HEP study; establishment of

HEP workflow to carry out physics object performance study; provide meaningful scale factor for data and MC correction; and to promote young Malaysian students to participate in frontier research such as experimental particle physics.

2 Detailed Research Proposal

2.1 Research Background Information

The open-source model perpetuated by CERN ever since the release of World Wide Web software in 1994 is a testimony of CERN's dedication on promoting open-science policy [10]. The CMS Open Data project [11] offers new opportunities to perform multitude of HEP analysis studies, ranging from cross section measurment of the SM processes, Physics Objects performance studies to computational benchmarking and scalability studies. The project breath a new life into the HEP research domain, focusing on the reproducibility of the HEP analysis, thus providing a mean for physics result validation.

Malaysia was officially accepted into the CMS Collaboration on 20 October 2013 [12], the collaboration has successfully produced several Malaysian young scientists in frontier research such as experimental particle physics in the past years. However, the local HEP research effort is not optimally organized due to the poor structural coordination, inadequate man(woman) powers and the limited availability of pedagogical materials (lecture notes, coding examples, etc) in the local university due to social economical factors. Therefore, it is high time to establish a nurture ground in Malaysia, such as protoyping a minimalistic HEP analytical model using CMS Open Data, to be used to train, promote and encourage young Malaysian students to participate in frontier research.

The Tag-and-Probe is a data-driven method used to estimate the efficiency of an event selection on the data or simulation, based on a known resonance such as the J/ψ, or the Z boson. The final state decayed from those resonances is used as a handle to perform non-bais measurement on how many lepton (assuming $Z \to ll$, where l is lepton) passes the event selection pertaining the invariant mass of the Z boson. The Tag-and-Probe study is ideal to be used as an application on Open Data datasets as the method is self-contained and pedagogical.

The proposed research is to rehash the Tag-and-Probe workflow on the open data datasets, using the current computational resource in Nuclear Science program, to demonstrate the feasibility of the current computational and analytical models, and the reproducibility of the HEP analysis results.

The proposed research can be used as a case study, providing useful comments and feedbacks on develoment of HEP capablity in Malaysia.

2.1.1 Problem Statement

The CMS experiment releases research quality datasets (Open Data) periodically, allowing scientists outside of the collaboration to explore the scientific potential of the Open Data. In particular, the Tag-and-Probe method is a Physics Object specific study used to evaluate the correction factor incurred by the MC selection efficiency. The study is a self-contained and pedagogical, it is deemed suitable to be used to benchmark on the current computational setup. The working example will be assessed if it is suitable to use for develop local experimental HEP capability.

2.1.2 Research Significance

The successful demonstration of HEP workflow using the designed analytical model implies the current computational model is feasible to be used for pedagogical purposes. Students will be exposed to the Tag-and-Probe technique used in experimental particle physics research such as CMS experiment, which is crucial in a full fledge physics analysis. On the physics's aspect, the scale factor measured using CMS Open Data can be used as a correction factor for physics analysis.

To explore this possibility, the proposed research is to design an analytical model to compute the MC event efficiency and the scale factors, on the current computational setup, using the CMS Open Data collected at $\sqrt{s} = 8$ TeV corresponding to the integrated luminosity of 1.8 fb^{-1}.

2.1.3 Research Hypotheses

The problem statements beg to validate the following hypotheses:

1. the computational and analytical models proposed for the offline HEP data analysis performed optimally,
2. the Tag-and-Probe study is suitable to be used for pedagogical material,
3. computation of Physics Objects reconstruction and identification efficiencies is feasible using CMS Open Data,
4. the derived scale factors and thier associated systematics uncertainties are compatible to published results.

2.1.4 Research Questions

The research questions are:

1. does the current computational setup favourable for offline HEP data analysis, using CMS Open Data,
2. how to use HEP data analysis such as Tag-and-Probe study as pedagogical materials for a class, to develop local experimental HEP capability,
3. is the Physics Objects reconstruction and identification efficiencies computation feasible using CMS Open Data,
4. is the derived scale factors and thier associated systematics uncertainties compatible to published results.

2.1.5 Literature Review

The first data release of the CMS experiment was announced in 2014, bringing research-quality particle collision data into the public domain for the first time. Regular releases of CMS data have taken place ever since with modalities as defined in the data preservation, re-use and open access policy [8]. As of 2021, more than 2 PB of data from the 2010-2012 run period are available to users external to the CMS collaboration, served through the CERN Open data portal (CODP) [13].

The first research papers using CMS open data were published in 2017 on jet substructure studies [14, 15], and authors included valuable feedback and advice to the community. Further studies have been performed including searches for new particles [16, 17], Standard Model analyses [18], and several studies on machine learning and methodology.

CMS releases a full reprocessing of data from each data-taking period in the Analysis Object Data (AOD) format, based on the ROOT framework [19] and processed through CMS software CMSSW [20]. The data are made available in the format and with the same data quality requirements that analyses of the CMS collaboration start from. AOD is the main format used in CMS for Run-1 (2010–2012) data analysis. Starting from Run-2 (2015–2018), new reduced data formats called MiniAOD [21] and NanoAOD [22] have been developed, and Run-2 data will be released in these slimmer formats.

The collision data are stored in "primary datasets" based on the event content identified at the time of data taking. The dataset name is an indication of its physics content, and each dataset record lists of the selection algorithms, the High-Level Trigger (HLT) streams, that were used to direct the data to that specific dataset. On the other hand, the simulated datasets are generated by MC generator programs, undergo detector simulation using CMSSW, and are subsequently processed into the same format as the collision data. During this processing chain, additional events are added on top of the simulated process to take into account the pile-up in the same beam crossing. The dataset names are identical to those used internally in CMS, and give an indication of the simulated process.

The CMS software, CMSSW, is open source and available on GitHub [20]. It is also accessible to the CMS open data environment through the CernVM file system (CVMFS) [23]. This software is used for data taking, event reprocessing, and analysis, as well as for the generation of simulated events.

The Tag and Probe method [9] is a data-driven technique for measuring particle detection efficiencies. It is based on the decays of known resonances to pairs of the particles being studied. The pair, with one designated as Tag particle, defined as well identified, triggered Physics Object (tight selection criteria); and the other one designated as Probe particle, defining a unbiased set of Physics Object candidates (very loose selection criteria), either passing or failing the criteria for which the efficiency is to be measured. The efficiency is given by the fraction of Probe particle that pass a given criteria:

$$\epsilon = \frac{\text{Passing probe physics object criteria}}{\text{All probe particle}} \tag{1}$$

The denominator corresponds to the number of resonance candidates (tag+probe pairs) reconstructed in the dataset. The numerator corresponds to the subset for which the probe passes the criteria.

The tag+probe invariant mass distribution is used to select only signal, that is, only true J/ψ or Z candidates decaying to dimuons. This is achieved in this exercise by the usage of two methods: fitting and side-band-subtraction.

The determination of the detector efficiency is a critical ingredient in any physics measurement. It accounts for the particles that were produced in the collision but escaped detection (did not reach the detector elements, were missed by the reconstructions algorithms, etc). It can be in general estimated using simulations, but simulations need to be calibrated with data. The Tag-and-Probe method here described provides a useful and elegant mechanism for extracting efficiencies directly from data!

[1] Morad Aaboud et al. "Precision measurement and interpretation of inclusive W^+ , W^- and Z/γ^* production cross sections with the ATLAS detector". In: *Eur. Phys. J. C* 77.6 (2017), p. 367. DOI: 10.1140/epjc/s10052-017-4911-9. arXiv: 1612.03016 [hep-ex].

[2] Serguei Chatrchyan et al. "Measurement of inclusive W and Z boson production cross sections in pp collisions at $\sqrt{s} = 8$ TeV". In: *Phys. Rev. Lett.* 112 (2014), p. 191802. DOI: 10.1103/PhysRevLett.112.191802. arXiv: 1402.0923 [hep-ex].

[3] Serguei Chatrchyan et al. "Measurement of the $t\bar{t}$ production cross section in the dilepton channel in pp collisions at $\sqrt{s} = 8$ TeV". In: *JHEP* 02 (2014), p. 024. DOI: 10.1007/JHEP02(2014)024. arXiv: 1312.7582 [hep-ex].

[4] Georges Aad et al. "Measurement of the angular coefficients in Z-boson events using electron and muon pairs from data taken at $\sqrt{s}$=8 TeV with the ATLAS detector". In: *JHEP* 08 (2016), p. 159. DOI: 10.1007/JHEP08(2016)159. arXiv: 1606.00689 [hep-ex].

[5] Georges Aad et al. "Measurement of the transverse momentum and ϕ^*_η distributions of Drell–Yan lepton pairs in proton–proton collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector". In: *Eur. Phys. J. C* 76.5 (2016), p. 291. DOI: 10.1140/epjc/s10052-016-4070-4. arXiv: 1512.02192 [hep-ex].

[6] CMS Experiment. "Performance of photon reconstruction and identification with the CMS detector in proton-proton collisions at $\sqrt{s} = 8$ TeV". In: *Journal of Instrumentation* 10.08 (2015), P08010–P08010. DOI: 10.1088/1748-0221/10/08/p08010. URL: https://doi.org/10.1088/1748-0221/10/08/p08010.

[7] Serguei Chatrchyan et al. "Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC". In: *Phys. Lett. B* 716 (2012), pp. 30–61. DOI: 10.1016/j.physletb.2012.08.021. arXiv: 1207.7235 [hep-ex].

[8] CMS Experiment. *CMS Data Preservation, re-use and Open access policy.* 2020. URL: https://cms-docdb.cern.ch/cgi-bin/PublicDocDB/ShowDocument?docid=6032.

[9] Olaf Behnke et al. *Data analysis in high energy physics: a practical guide to statistical methods.* Weinheim: Wiley-VCH, 2013. DOI: 10.1002/9783527653416. URL: https://cds.cern.ch/record/1517556.

[10] CERN. *CERN Open Data Policy for the LHC Experiments.* 2020. DOI: 10.7483/OPENDATA.0XO6.HYY1.

[11] Kati Lassila-Perini et al. "Using CMS Open Data in research – challenges and directions". In: *EPJ Web Conf.* 251 (2021), p. 01004. DOI: 10.1051/epjconf/202125101004. arXiv: 2106.05726 [hep-ex].

[12] et al. Faridah Mohamad Idris. *The CMS Collaboration in Malaysia.* 2013. URL: https://www.inustec.my/uploads/1/3/8/2/138265193/cms_collaboration.pdf.

[13] CERN. 2022. URL: https://opendata.cern.ch/.

[14] Aashish Tripathee et al. "Jet Substructure Studies with CMS Open Data". In: *Phys. Rev. D* 96.7 (2017), p. 074003. DOI: 10.1103/PhysRevD.96.074003. arXiv: 1704.05842 [hep-ph].

[15] Andrew Larkoski et al. "Exposing the QCD Splitting Function with CMS Open Data". In: *Phys. Rev. Lett.* 119.13 (2017), p. 132003. DOI: 10.1103/PhysRevLett.119.132003. arXiv: 1704.05066 [hep-ph].

[16] Cari Cesarotti et al. "Searching in CMS Open Data for Dimuon Resonances with Substantial Transverse Momentum". In: *Phys. Rev. D* 100.1 (2019), p. 015021. DOI: 10.1103/PhysRevD.100.015021. arXiv: 1902.04222 [hep-ph].

[17] Christopher G. Lester and Matthias Schott. "Testing non-standard sources of parity violation in jets at the LHC, trialled with CMS Open Data". In: *JHEP* 12 (2019), p. 120. DOI: 10.1007/JHEP12(2019)120. arXiv: 1904.11195 [hep-ex].

[18] Aram Apyan et al. "Opportunities and challenges of Standard Model production cross section measurements in proton-proton collisions at $\sqrt{s}$ =8 TeV using CMS Open Data". In: *JINST* 15.01 (2020), P01009. DOI: 10.1088/1748-0221/15/01/P01009. arXiv: 1907.08197 [hep-ex].

[19] R. Brun and F. Rademakers. "ROOT: An object oriented data analysis framework". In: *Nucl. Instrum. Meth. A* 389 (1997). Ed. by M. Werlen and D. Perret-Gallix, pp. 81–86. DOI: 10.1016/S0168-9002(97)00048-X.

[20] CMS Experiment. 2021. URL: http://cms-sw.github.io/.

[21] Giovanni Petrucciani, Andrea Rizzi, and Carl Vuosalo. "Mini-AOD: A New Analysis Data Format for CMS". In: *J. Phys. Conf. Ser.* 664.7 (2015), p. 7. DOI: 10.1088/1742-6596/664/7/072052. arXiv: 1702.04685 [physics.ins-det].

[22] Karl Ehatäht. "NANOAOD: a new compact event data format in CMS". In: *EPJ Web Conf.* 245 (2020). Ed. by C. Doglioni et al., p. 06002. DOI: 10.1051/epjconf/202024506002.

[23] Jakob Blomer et al. *The CernVM File System: v2.7.5*. Version 2.7.5. Oct. 2020. DOI: 10.5281/zenodo.4114078. URL: https://doi.org/10.5281/zenodo.4114078.

2.2 Research Objectives

The research objectives are:

1. to design a minimalistic HEP analytical model, used to assess the current computational setup,
2. to design a user friendly analytical method to promote HEP data analysis among young student,
3. to calculate the Physics Objects reconstruction and identification efficiencies,
4. to derive Physics Objects scale factors and thier associated systematics uncertainties.

2.3 Research Methodology

Analysis of the CMS data is most commonly done in two steps: first, selecting events of interest and writing them to a new, smaller format, and second, analysing the selected events.

Due to the experiment-specific data format, the first step will almost inevitably be done using the CMS software CMSSW in a computing environment compatible with the open data. For a realistic physics analysis, this step usually consists of hundreds of jobs, each taking several CPU hours. The analysts then have the option of either remaining in the open data environment, or moving their data out of the open data portal to their own computers for subsequent processing and optimization.

In the second step, an offline data analysis will be performed on the processed datasets; the institutional computational capability will be put into test. Conventionally, processed datasets are transferred locally into the workstation or equivalent, and user will write a macro or compiled code using ROOT to perform event selection and analyse the selected event.

For the estimation of the Physics Object identification selection efficiency, the tag is chosen to be a well identified and isolated Physics Object, while the probe is chosen as a Physics Object identified with loose selections. The invariant mass of the tag-probe pair is required to be within a window around the Z boson mass (the effect of changing the Z mass window is included as

a systematic uncertainty). After that, the probe is required to pass the analysis identification selections and the efficiency is computed both in data and simulation.

Both the lepton identification and the lepton isolation selection efficiencies are measured by the fitting method [9], to take into account the combinatorial background below the resonance peak. The signal plus background fit to the invariant mass distribution is performed simultaneously in two categories, corresponding to events in which the probe lepton passes or fails the identification requirements, and separately in bins of transverse momentum, p_T and peusedorapidity η.

The identification selection on Physics Object resulted in different efficiency in data and simulation. These differences are corrected by a scale factor defined as a function of lepton p_T and η respectively,

Finally, the feasibility on the computational and analyical setup is assessed in term of the PC walltime.

2.4 Expected Research Outcome

1. New theoretical finding: None
2. Specific application or potential research: Demonstration of Physics Object measurement using Open Data
3. Social economic impact: None

3 Equipment and Materials Access

Equipment	Location
Hyper-Performance Computer (HPC)	CERN
1 node Server	Nuclear Science Building
Workstation	Nuclear Science Building
laptop	Personal

4 Gantt Chart