PROCEEDINGS OF SCIENCE

Overview of the HL-LHC Upgrade for the CMS Level-1 Trigger

Jona Motta[*] on behalf of the CMS Collaboration

Laboratoire Leprince-Ringuet, CNRS/IN2P3, Ecole Polytechnique, Institut Polytechnique de Paris, Route de Saclay, Palaiseau, France

E-mail: jona.motta@cern.ch

The High-Luminosity LHC (HL-LHC) will open an unprecedented window on the weak-scale nature of the universe, providing high-precision measurements of the standard model as well as searches for new physics beyond the standard model. Such precision measurements and searches require information-rich datasets with statistical power that matches the high luminosity provided by the Phase-2 upgrade of the LHC. Efficiently collecting those datasets will be a challenging task, given the harsh environment of 200 simultaneous proton-proton interactions per HL-LHC bunch crossing. For this purpose, CMS is designing an efficient data-processing hardware trigger (Level-1) that will include tracking information and high-granularity calorimeter information. Trigger data analysis will be performed through sophisticated algorithms such as particle flow reconstruction, including the widespread use of Machine Learning. The current conceptual system design is expected to take full benefit of advances in FPGA and link technologies over the coming years, providing a high-performance, low-latency computing platform for large throughput and sophisticated data correlation across diverse sources. The expected impact on the physics reach of the experiment will be summarized in these proceedings and illustrated with selected benchmark channels.

PoS(EPS-HEP2023)534

The European Physical Society Conference on High Energy Physics
21-25 August 2023
Hamburg, Germany

[*]Speaker

1. Introduction

The High-Luminosity LHC (HL-LHC) is scheduled to start in 2029, and it will constitute the Phase-2 of the LHC operations. It is designed to operate at a centre-of-mass energy of 14 TeV while delivering an instantaneous luminosity of $5 - 7.5 \cdot 10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$. These conditions correspond to a number of simultaneous collisions (pileup, PU) per bunch crossing (BX) of $O(200)$.

Efficiently collecting datasets to be used in the HL-LHC physics program will be a challenging task. Therefore, as part of the Phase-2 upgrade, the CMS Collaboration [1] is redesigning its hardware-implemented Level-1 Trigger (L1T) [2]. The Phase-2 L1T builds on the other subdetectors' upgrades, exploiting tracking and highly-granular calorimetric information. Algorithms such as Particle Flow (PF) [3] reconstruction and the use of Machine Learning (ML) techniques will be employed. The Phase-2 L1T will exploit state-of-the-art Field Programmable Grid Arrays (FPGAs) and link technologies, providing a high-performance, low-latency, and high-throughput system.

These proceedings are structured as follows. Section 2 presents the Phase-2 upgrade of the L1T and the associated hardware choices; Sections 3 and 4 present how the upgraded L1T system can be exploited to preserve the physics reach and extending the discovery potential of the CMS experiment, respectively; Section 5 closes the discussion with conclusions and outlook.

2. The Phase-2 Level-1 Trigger

The architecture of the Phase-2 L1T is reported in Figure 1. It aims at optimizing processing board numbers, interconnections, and latency while ensuring flexibility and robustness of the system. Optimal division of labour is achieved by implementing intermediate global triggers. The system capitalizes on new hardware technologies to deliver computing power and high-speed data transfer for a global detector view. It uses generic stream-processing engines as data processing units to leave ample room for further algorithm optimization. This design enables both regional and time-multiplexed architecture options to be exploited to reach close to real-time analysis.

The architecture of the Phase-2 L1T is composed of four independent data processing paths converging in a single global trigger, complemented by the new scouting system. This design reflects the need for producing complementary trigger objects to achieve the best physics selectivity.

- **Calorimeter Trigger path**: takes advantage of the upgraded barrel and endcap calorimeters, taking as input the highly granular calorimeter Trigger Primitives (TPs) over the entire η coverage. The crystal-based TPs from the ECAL barrel are received by the Barrel Calorimeter Trigger (BCT), and they are processed to build e/γ candidates. The crystal TPs are then merged into TTs and sent to the Global Calorimeter Trigger (GCT), where are implemented the reconstruction algorithms for τ, jet, and sum objects.
- **Track Trigger path**: this represents the first great innovation of the Phase-2 system; track TPs are produced in the tracker back-end from the information of the outer tracker only and sent to the Global Track Trigger (GTT). In the GTT, the vertex and track-only object reconstruction is performed with the goal of subsequent muon and calorimeter deposit matching. This path is invaluable in identifying PU contributions and reducing the L1T rate.

- **Muon Trigger path**: profits from the redundant muon detectors and improved η coverage. The TPs are processed with algorithms which analyse different areas of the detector, namely the Barrel Muon Track Finder (BMTF), Overlap Muon Track Finder (OMTF), and Endcap Muon Track Finder (EMTF). The Global Muon Trigger (GMT) receives both the muon track finders' and tracker track finders' output to build matched candidates.
- **Particle Flow Trigger path**: also referred to as Correlator Trigger (CT), is the second major innovation of the upgraded system. It receives as input the information from the previous three trigger paths and exploits it to implement a processing *a la* PF. It is subdivided into two layers, the first building the high-level candidates and the second implementing identification and isolation algorithms.
- **Global Trigger**: it gathers the output from the GCT, GMT, GTT, and CT to produce the event accept/reject decision based on a menu of algorithms. It exploits correlation variables among different objects, often with algorithms designed for analysis-specific purposes. Owing to the large computing power, sophisticated ML-based *topological* algorithms are being developed to target specific signatures of rare processes in order to enhance their selection efficiency.
- **Scouting system**: it is the third central innovation of the new L1T; it guarantees the ability to perform trigger-less analysis of L1T data at the 40 MHz BX rate by profiting from the spare optical links of the various processing boards. Moreover, its access to each L1T subsystem makes it a great tool for real-time diagnostics of the whole system.

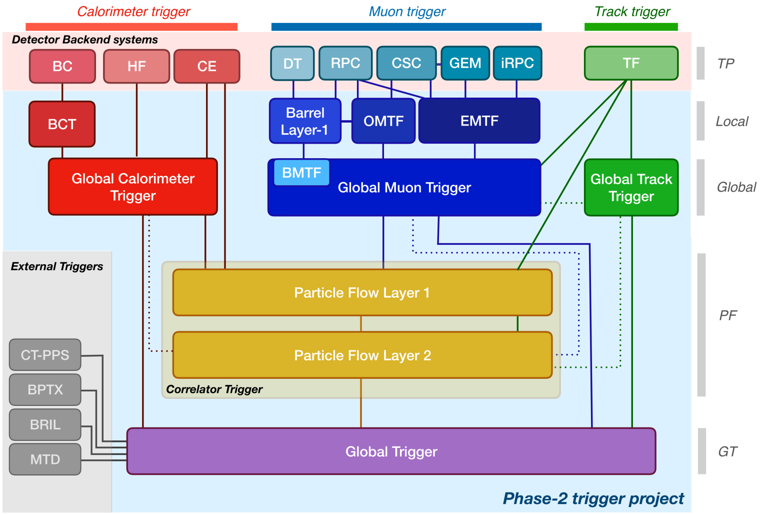

Figure 1: Schematic representation of the foreseen L1T upgraded system architecture in Phase-2. The information from the Barrel Calorimeters (BC), Hadron Forward Calorimeter (HF) and High Granularity Calorimeter (CE) detectors are used to reconstruct the e/γ, τ, and jet candidates, as well as global HT and $p_{\mathrm{T}}^{\mathrm{miss}}$ quantities in the BCT/GCT. The Drift Tube (DT), Resistive Plate Chamber (RPC), Cathode Strip Chamber (CSC), and Gaseous Electron Multiplier (GEM) detectors send their information to the BMTF, OMTF, and EMTF, which build tracks to be identified as μ candidates in the GMT. The GTT uses the Track Finder (TF) TPs to increase the L1T performance. The CT exploits the high-level objects built by each global sub-trigger and employs a PF approach to event reconstruction. The GT collects the outputs from all previous steps and takes the event accept/reject decision [2].

This architecture will have a latency of 12.5 μs, corresponding to roughly three times that of Phase-1. This latency is dictated by the track hardware reconstruction and matching time. The output bandwidth is of 750 kHz, being 7.5 times larger than the Phase-1 output rate, thus allowing for energy thresholds comparable to Run-2 and Run-3 values to be retained in the busier PU conditions. Finally, exploiting state-of-the-art FPGAs and optical links will enable the selection of events based on input data as high as $\sim$ 60 TB/s, corresponding to a 30-fold increase compared to Phase-1. Wherever possible, the Phase-2 L1T will use industry standards concerning boards, FPGAs, and optics, thus greatly facilitating long-term operations while reducing maintenance costs.

Four types of processing boards are being developed depending on the area of implementation, all abiding by the ATCA (Advanced Telecommunications Computing Architecture) industry standard: the Serenity, the AP-x, the BMT-L1, and the X2O boards. All of them underwent substantial evolution since the TDR, achieving increased input/output bandwidth and computing power, as well as newer and denser on-board optics. The hardware testing is underway with both single-flavour and multi-flavour board tests. Finally, the L1T will be equipped with a total of more than 200 Xilinx Virtex UltraScale+ VU13P with up to 28 Gb/s transceivers, allowing the extensive development of advanced ML algorithms that can be implemented in FPGA firmware with the `hls4ml` [4] tool.

3. Preserving the physics reach

The central feature of the Phase-2 L1T to preserve the CMS physics reach is the inclusion of track information. Using the full outer tracker ($|\eta| < 2.4$) volume and a time-multiplexed architecture, the GTT reconstructs charged particle tracks at the BX rate, allowing precise reconstruction of primary and secondary vertexes. The tracking and vertex matching efficiency are reported in Figure 2, showcasing a tracking efficiency > 95% over the entire η range, and a very high vertex reconstruction performance. The tracking and vertexing algorithms have already been demonstrated in hardware tests [2].

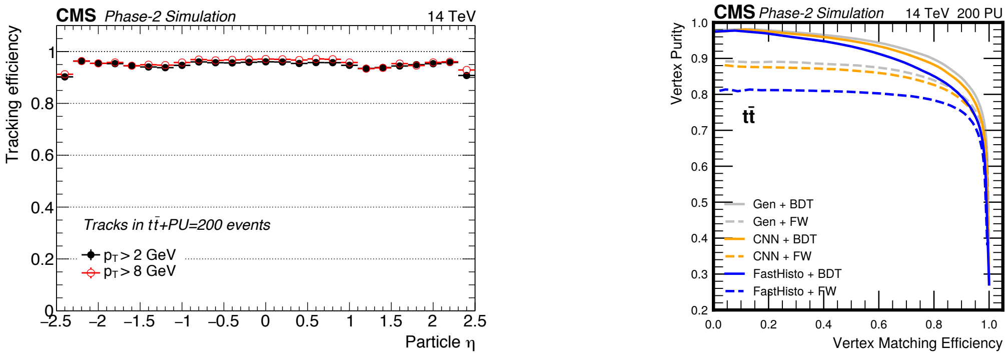

Figure 2: L1T tracking efficiency versus η for tracks above 2 GeV (black) or 8 GeV (red) defined with respect to truth-level particles (left). Track-to-vertex association performance curves for the FastHisto algorithm (blue) when the association is performed with fixed window (FW) or BDT; for reference, the performance when the association uses the true ("Gen") vertex position is shown; the orange curves illustrate the performance of an algorithm using convolutional neural networks (CNNs) (right). Obtained in $t\bar{t}$ events at 200 PU [2].

The tracks and vertex information, in association with the calorimeter clusters and muon candidates, is exploited vy the CT to perform a full event reconstruction with a PF approach to identify individual particles within each event. Moreover, the PF candidates are used as input to the Pile-Up Per Particle (PUPPI) algorithm, which subtracts PU contribution on a particle-by-particle basis. Both PF and PUPPI approaches have been extensively demonstrated in all hardware subsystems involved, showcasing a 98% hardware-emulator agreement [2]. Figure 3 reports the large performance improvement granted by the PUPPI algorithm compared to standard triggering approaches for jets, total energy sum (HT), and τ leptons. The great gain achieved on all objects guarantees the preservation of sensitivity for all CMS analyses involving such objects.

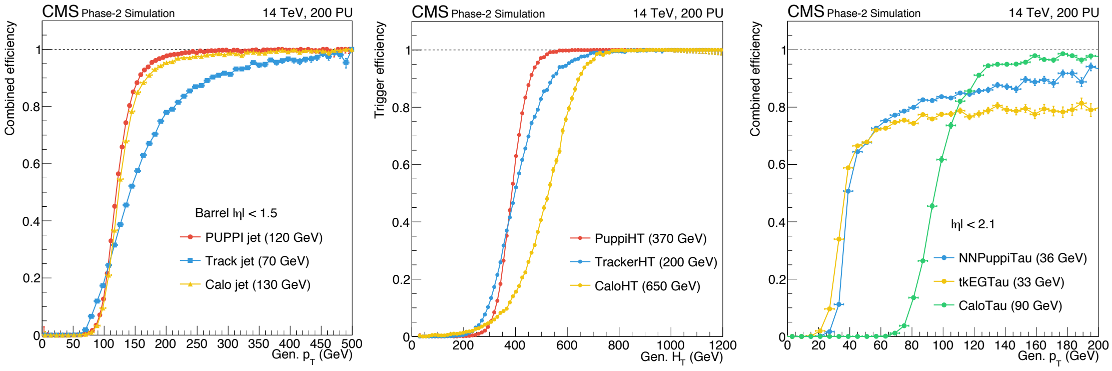

Figure 3: Comparison of combined matching and turn-on efficiency for track jets, calorimeter only jets, and histogrammed PUPPI jets, for thresholds that provide a fixed online rate of 70.1 kHz (left), and HT at 10.5 kHz (centre), in $t\bar{t}$ events at 200 PU. Single τ trigger efficiency as a function of p_T for various algorithms at 6.6 kHz, in HH → bbττ events at PU 200 (right) [2].

Preserving the Phase-1 performance of L1T electron reconstruction is paramount for several CMS analyses like the η-differential measurement of the $t\bar{t} \rightarrow bb\ell\nu_\ell qq$ production cross section, which could attain a 50% improvement in relative uncertainty [5]. The L1T includes two electron identification approaches. The baseline identification technique called *Elliptic-ID* relies on the independent selection of calorimeter clusters and tracks followed by an angular matching procedure. Electrons track quality and momentum resolution suffer because of distortions due to bremsstrahlung radiation. To overcome this inefficiency in the endcap region, a new ML approach called *Composite-ID* [6], combines information about tracks and clusters into a single Boosted Decision Tree model for matching and identification. Figure 4 reports the efficiency as a function of the electron p_T and $|\eta|$ for the *Elliptic-ID* and *Composite-ID* methods, showcasing a consistent improvement of ∼ 10% (at a fixed rate value) of the second technique over the first one.

Of utmost importance is also the preservation of the hadronically decaying τ leptons (τ_h) performance. Reaching the typical energy thresholds targeted by the L1T (50 GeV) requires using the track information. Nevertheless, algorithms exploiting only calorimetric information have a major role in increasing L1T efficiency at high thresholds (100 GeV). To this end, a new algorithm has been designed to exploit only calorimetric information and to ensure a unified treatment of all TPs from the barrel and endcap calorimeters: the TauMinator [7, 8]. At its core, the TauMinator is composed of two CNNs that cast the τ_h reconstruction into an image recognition problem. The preliminary firmware implementation of the algorithm showcases minimal FPGA resources usage

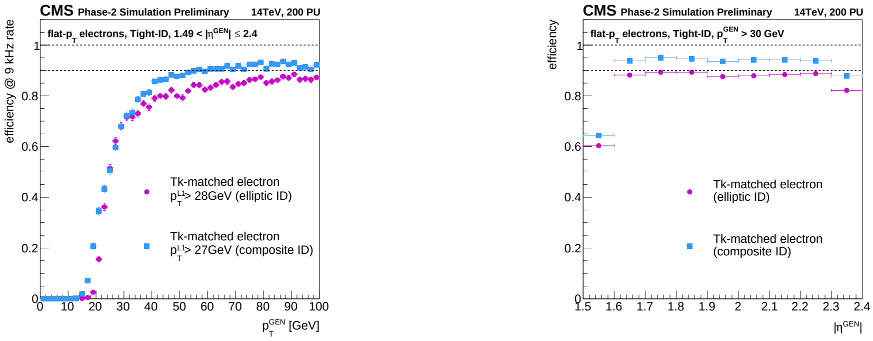

Figure 4: Performance of the Composite-ID model compared to the geometric matching (Elliptic-ID) for the “tight” WP optimized for single lepton triggers with thresholds between 20 and 30 GeV. Efficiency comparison as a function of the p_T of the generated electron, for a fixed 9 kHz rate, corresponding to the endcap bandwidth for Run-3-like thresholds (right); efficiency comparison as a function of the pseudorapidity for electrons with $p_T > 30$ GeV (left). The new Composite-ID model guarantees a substantial efficiency gain over the full pseudorapidity range [6].

and 100% hardware-emulator agreement. The substantial performance improvement granted by the TauMinator algorithm over the standard CaloTau algorithm is reported in Figure 5. This is fundamental for the HH → bbττ search, which is the second most sensitive to Higgs boson pair production at CMS [9].

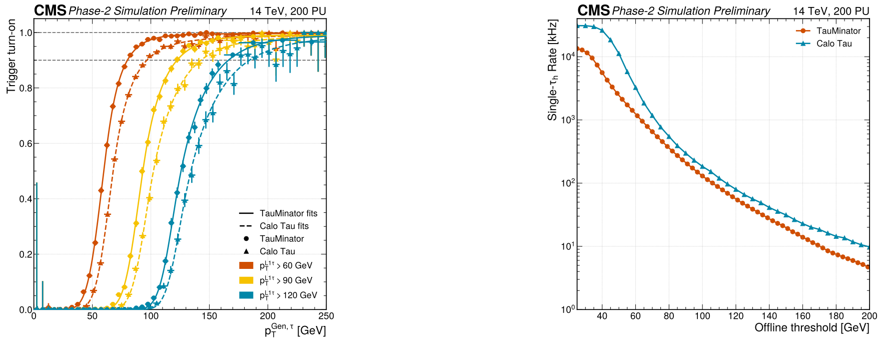

Figure 5: (left) Comparison of the trigger turn-ons of the TauMinator (circles-solid) and the CaloTau (triangles-dashed) algorithms for different values of L1T threshold, evaluated in HH → bbττ events. The functional form of the fits consists of a cumulative Crystal Ball function convolved with an arc-tangent in the high p_T region. (right) The single-τ rate as a function of the offline p_T. The TauMinator algorithm ensures the following improvements over the CaloTau algorithm: a reduction of the rate by 37% at a 150 GeV threshold ; or a reduction of the threshold by 14 GeV at a fixed rate of 31.4 kHz [7, 8].

4. Extending the discovery potential

Tack information is not only central to preserving the physics reach but also to extend the discovery potential. For this reason, a new end-to-end Neural Network (NN) approach has been developed to exploit track features to regress the vertex position and classify the associated tracks

[10, 11]. This new method guarantees a reduction of up to 50% in the tails of the vertex position residual. Such development is fundamental to reconstruct muons with $2 < p_T < 3$ GeV for the search of $\tau \rightarrow \mu\mu\mu$ decays, as well as to develop new algorithms for displaced objects crucial for long-lived particles searches [5].

To enhance the L1T performance on jets, the new Seed Cone jet algorithm [12] has been developed to use the full granularity and implemented in the FPGAs used by the CT. The algorithm begins by finding the highest p_T PUPPI candidate, then finding all particles within a ΔR cone of the seed to form the first jet. The constituents of the jet are removed from the processing before the procedure is repeated. The constituents are used to define the jet position and p_T. This approach is similar to the anti-k_T [13] algorithm and, as shown in Figure 6 (left), guarantees performances similar to it. This will highly benefit all CMS analyses employing jets. The Seed Cone jets can be used with a new NN-based L1T algorithm to identify jets originating from b quarks (b-tagging) [14], the first of its kind. The performance of the NN b-tagging algorithm is reported in Figure 6 (right), showcasing how it outperforms standard triggers in the low invariant mass regime.

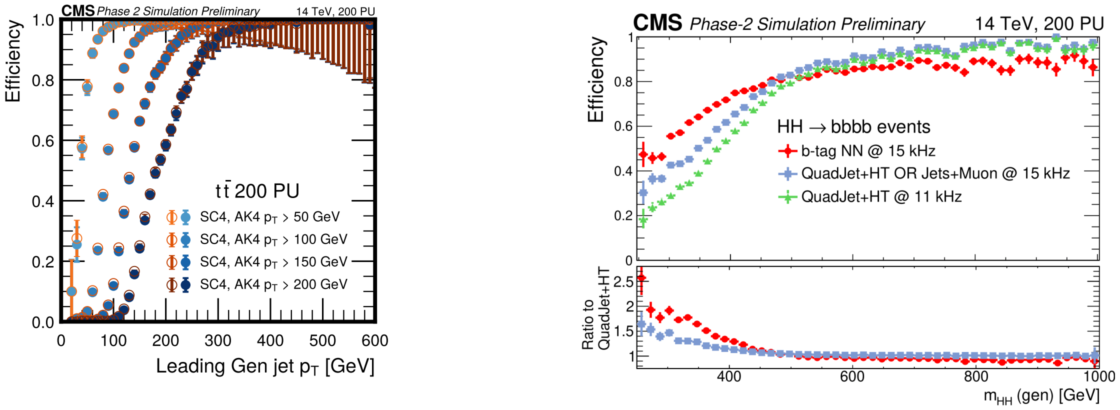

Figure 6: Performance of the Seeded Cone and anti-k_T jet algorithms for different online jet thresholds, both with a radius parameter of 0.4 and using the same L1T PUPPI candidates as inputs, in $t\bar{t}$ events at 200 PU (left) [12]. Trigger efficiency in HH → bbbb events at 200 PU as a function of the di-Higgs system mass (right); in the upper panel, the efficiency of the b-tag NN (red circles) is compared to the QuadJet+HT (green triangles), and QuadJet+HT OR Jets+Muon (blue squares). The bottom panel shows the ratio of the efficiency for each trigger option to the efficiency of the QuadJet+HT. The b-tag NN trigger increases the efficiency for events with low m_{HH} by up to a factor 1.5 over the QuadJet+HT or Jets+Muon triggers.

5. Conclusions and outlook

The HL-LHC will pose big challenges for the CMS experiment, which will undergo a substantial upgrade both in hardware and software capabilities. In this context, the Phase-2 L1T upgrade proposes solid and flexible solutions to triggering and data acquisition challenges via state-of-the-art FPGAs and very-high-speed optical links. The extensive use of ML technique is favoured by improved hardware capabilities and development tools like `hls4ml`. The hardware demonstration of the algorithms is ongoing and planned for testing with data during the LHC Run-3. The improved L1T physics selectivity will benefit all analyses envisaged for the CMS Phase-2 physics program.

References

[1] CMS Collaboration, *The CMS experiment at the CERN LHC*, *Journal of Instrumentation* **3** (2008) S08004.

[2] CMS Collaboration, *The Phase-2 Upgrade of the CMS Level-1 Trigger*, Tech. Rep. CERN, Geneva (2020), **CERN-LHCC-2020-004, CMS-TDR-021**.

[3] CMS Collaboration, *Particle-flow reconstruction and global event description with the CMS detector*, *Journal of Instrumentation* **12** (2017) P10003.

[4] J. Duarte, S. Han, P. Harris, S. Jindariani, E. Kreinar, B. Kreis et al., *Fast inference of deep neural networks in FPGAs for particle physics*, *Journal of Instrumentation* **13** (2018) P07027.

[5] ATLAS and CMS Collaborations, *Addendum to the report on the physics at the HL-LHC, and perspectives for the HE-LHC: Collection of notes from ATLAS and CMS*, Tech. Rep. CERN, Geneva (2019), **CERN-LPCC-2019-01, CMS-FTR-19-001, ATL-PHYS-PUB-2019-006, CERN-2019-007-ADD**.

[6] CMS Collaboration, *Electron Reconstruction and Identification in the CMS Phase-2 Level-1 Trigger*, **CMS-DP-2023-043** (2023) .

[7] CMS Collaboration, *Hadronic Tau Reconstruction in the CMS Phase-2 Level-1 Trigger using NNs with Calorimetric Information*, **CMS-DP-2023-062** (2023) .

[8] J. Motta, *Development and firmware implementation of a Machine Learning based hadronic τ lepton Level-1 Trigger algorithm in CMS for the HL-LHC*, *PoS* **EPS-HEP2023** (2023) 590.

[9] CMS Collaboration, *Search for nonresonant Higgs boson pair production in final state with two bottom quarks and two tau leptons in proton-proton collisions at $\sqrt{s} = 13$ TeV*, *Physics Letters B* **842** (2023) 137531.

[10] CMS Collaboration, *Performance of the End-to-End Neural Network Approach to Phase-2 Level-1 Trigger Primary Vertex Reconstruction and Track to Vertex Association*, **CMS-DP-2021-035** (2021) .

[11] CMS Collaboration, *Performance Plots Showing the Effect of Different Cuts and Weighting in the Baseline Approach and a Technical Plot Showing the Effects of Pruning and Quantisation of the End-to-End Neural Network Approach to Phase-2 Level-1 Trigger Primary Vertex Reconstruction*, **CMS-DP-2022-020** (2022) .

[12] CMS Collaboration, *Jet Reconstruction with the Seeded Cone algorithm in the CMS Phase-2 Level-1 Trigger*, **CMS-DP-2023-023** (2023) .

[13] M. Cacciari, G.P. Salam and G. Soyez, *The anti-k_T jet clustering algorithm*, *Journal of High Energy Physics* **2008** (2008) 063.

[14] CMS Collaboration, *Neural network-based algorithm for the identification of bottom quarks in the CMS Phase-2 Level-1 trigger*, **CMS-DP-2022-021** (2022) .