MACHINE LEARNING FOR 40 MHZ SCOUTING AT CMS

Summer 2021

AUTHOR:
Gianvito Losapio
University of Genoa, Italy

SUPERVISORS:
Thomas Owen James
Emilio Meschi

CERN openlab

PROJECT SPECIFICATION

The LHC delivers collisions to CMS at a 40 MHz bunch crossing rate, generating hundreds of Tb/s of data in the detector but just a tiny fraction of them can be effectively read and stored. The Level-1 (L1) trigger, implemented in custom hardware using Field Programmable Gate Array (FPGA) devices, uses coarse-grained information from the calorimeter and muon subdetectors to search for signatures of interesting physics, and selects events at a maximum rate of 100 kHz.

Future upgrades will enable an L1 scouting system to capture intermediate data from the tracking, calorimeter and muon systems. This information would help with further improvements and analysis of the entire trigger system, as well as provide comprehensive and detailed detector diagnostics in real-time.

Machine Learning (ML) models are ideal candidates for the L1 scouting system as they can be implemented on FPGAs for close-to real-time analysis. By using the offline reconstructed parameters as targets, they can be used for the re-calibration of muon track parameters provided by the Global Muon Trigger (GMT). The same strategy can be potentially applied for the analysis of calorimeter data.

ABSTRACT

The Level 1 (L1) trigger at CMS uses coarse-grained information to search for signatures of interesting physics. L1 scouting is a new paradigm for data collection at CMS which could help in the early identification of promising potential signals, independently of any trigger selection bias. It will for the first time enable the reading out of trigger objects at the full collision rate (40 MHz), in order to perform studies and take measurements not possible within the constraints of the 100 KHz Level 1 accept rate.

The objective of this project is to investigate efficient machine learning algorithms with fast inference time for the L1 scouting system. Deep learning models have been compared to another class of machine learning models – namely kernel methods – as viable solutions to be implemented on FPGA devices. Several tests have been performed to compare accuracy on the re-calibration of muon track parameters. An analysis of the floating-point operations required by both models has been carried out. Preliminary tests have also been conducted for the re-calibration of jet transverse momentum.

TABLE OF CONTENTS

1. INTRODUCTION

The Large Hadron Collider (LHC) located at CERN, Geneva, is the world's largest particle accelerator, consisting of a 26.7km ring. The LHC accelerates protons to nearly the speed of light and then collides them at four points around its ring, each hosting a particle detector.

One of these detectors is CMS (Compact Muon Solenoid), which is a general-purpose particle detector, i.e., designed to enable searches for a wide variety of new physics. The CMS detector consists of several concentric layers of components (as shown in Figure 1) that exploit the different properties of particles to measure their energy and momenta. A highly detailed description of the CMS detector can be found in [1].

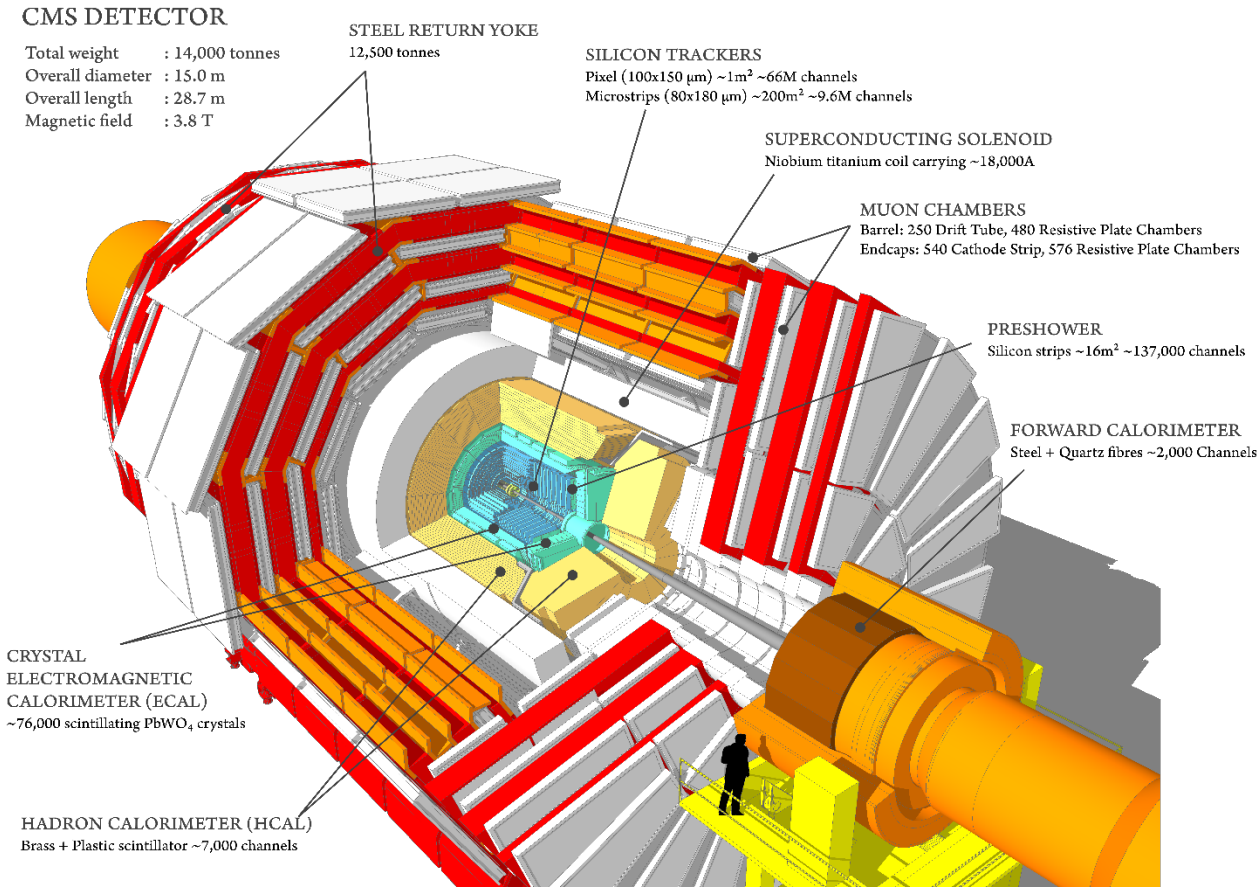

Figure 1 – Overview of the CMS detector. Credits: https://cms.cern/detector

A solenoid magnet is used to bend charged particles as they fly outwards from the collision point. Bending the trajectories of the particles helps to identify the charge and the momentum of each particle. The steel "yoke" that forms the bulk of the detector's mass is used to confine the 3.8 Tesla magnetic field generated by the solenoid magnet to the volume of the detector. A silicon tracker made of around 75 million individual electronic sensors generates electromagnetic interactions with the traversing particles and produces hits that can then be joined together to identify the track of the traversing particle.

The energy of the particles is measured by two kinds of "calorimeters". The Electromagnetic Calorimeter (ECAL) is the innermost of the two and measures the energy of electrons and photons by stopping them completely. Hadrons, which are composite particles made up of quarks and gluons, fly through the ECAL and are stopped by the Hadronic Calorimeter (HCAL).

Detecting muons is one of CMS's most important tasks. Muons are charged particles that are just like electrons and positrons but are 200 times heavier. Unlike most particles, muons are not stopped by either of the two calorimeters. Therefore, chambers to detect muons are placed at the very edge of the experiment where they are the only particles likely to register a signal.

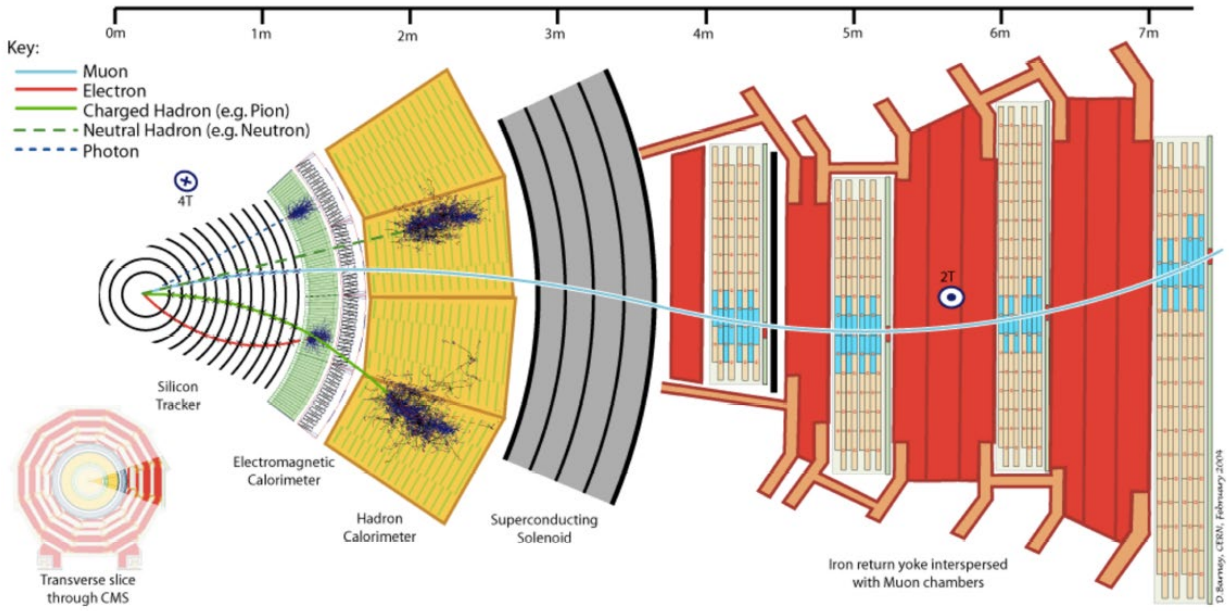

Figure 2 – A transverse slice through the CMS detector. Credits: [13]

The LHC delivers collisions to CMS at a 40 MHz bunch crossing rate. Each bunch crossing generates hundreds of Tb/s of data in the detector but just a tiny fraction of them can be effectively read and stored. A two-level trigger system selects the potentially interesting events to be read out for permanent storage and subsequent analysis [2]:

- The Level-1 (L1) trigger [2], implemented in custom hardware using field programmable gate array (FPGA) devices, uses coarse-grained information from the calorimeter and muon subdetectors to search for signatures of interesting physics, and selects events at a maximum rate of 100 kHz.
- The High-Level Trigger (HLT) is a farm of processors analysing full events read out at the L1-accept rate, using complex software algorithms to further reduce the event rate to about 1 kHz to be stored for offline analysis.

A schematic representation of the data flow is depicted in Figure 3.

The Global Muon Trigger (GMT, part of the L1 trigger) accumulates muons candidates from different regions: the barrel region (equipped with drift tube technology), the endcap regions (equipped with cathode strip chamber technology) and overlap regions (equipped with resistive plate chamber technology). Based on their quality and transverse momentum, the best eight candidates are sent to the Global Trigger (also part of the L1 trigger) to make the final decision.

After the planned Phase-2 upgrade of CMS [3], around 2027, the L1 trigger will include information from the tracking detectors. The L1 and HLT accept rates will be increased to 750 and 7.5 kHz respectively. Most importantly in regard to this work, the CMS L1 trigger will also include a new paradigm for data collection, called 'L1 Scouting' or '40 MHz Scouting'. It will for the first time enable the reading out of trigger objects at the full collision rate (40 MHz), in order to perform studies and take measurements not possible within the

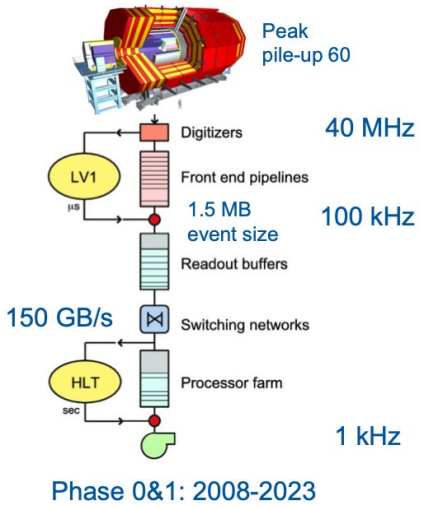

Figure 3 – A schematic representation of the data flow at the CMS detector. Credits: [5]

constraints of the Level 1 accept rate. In this way, L1 scouting could help in the early identification of promising potential signals, independently of any trigger selection bias [4].

Recently, deep learning models have been proposed to facilitate the early identification of potential physics signals, as well as improve the current data reconstruction pipeline in the L1 scouting system. A scouting demonstrator system has been built to show the feasibility of the studies, with deep learning models implemented on FPGAs for close-to real-time analysis. [5].

The objective of this project is to investigate efficient machine learning algorithms with fast inference time for the L1 scouting system. Deep learning models have been compared to another class of machine learning models – namely kernel methods - for the re-calibration of L1 trigger parameters. Several tests have been performed to compare accuracy and inference operations for the re-calibration of muon track parameters. Preliminary tests have also been conducted for the re-calibration of jet transverse momentum.

2. DATA ANALYSIS

L1 trigger measurements follow the CMS coordinate system (shown in Figure 4). The beam direction is parallel to the z axis, and collisions occur at approximately x = y = 0. The main parameters used in this study are:

- The azimuthal angle ϕ, measured in the x-y plane
- The pseudorapidity $\eta = -\ln(\tanh(\theta/2))$, with θ being the polar angle in the z-y plane
- The transverse momentum p_T, which is the component of the momentum p in the x-y plane

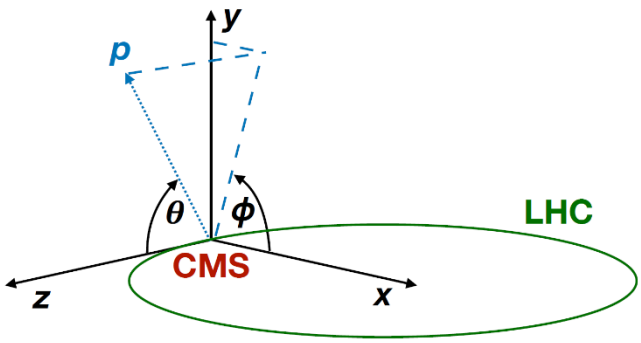

Figure 4 – CMS coordinate system. Credits: [13]

a. Muons

Muon track parameters have been collected from a preprocessed output of a special L1 trigger called ZeroBias during LHC Run-2 (2017-2018). ZeroBias (ZB) is a beam bunch crossing-time trigger, without physics signal requirements, used to understand the underlying event structure of collisions occurring at CMS [6]. Each row contains L1 trigger measurements of a single CMS event. Each event contains measurements related to a specific number of particles: 8 muons, 12 jets, 12 EGammas and 12 taus. An EGamma is an object that could be either an electron or a photon (gamma particle), but the resolution of the L1 trigger is not good enough to differentiate between those two possibilities. Each particle is characterized by a different number of features. The detailed number of features for each particle is shown in Table 1.

Table 1 – Schematic structure of the dataset

Muons		Jets		Egammas		Taus	
nMatchedMuons	8 x 15 features	nJets	12 x 3 features	nEGammas	12 x 4 features	nTaus	12 x 5 features

Table 2 shows the details of the four L1 measurements contained in the dataset which are used as inputs for the muon re-calibration problem. Each variable is stored as an integer value which represents a bin index over a specific measurement range [7]. In order to obtain the real values of the measurements, each integer number has to be mutliplied by the corresponding bin width.

Table 2 – Variable specifications for the L1 muon track parameters

Name	Description	Dtype	Range	Bin width	Min-Max
hwPtL1	transverse momentum	int64	$[0;\ 2^9 - 1]$	0.5 GeV	[0; 255] GeV
hwPhiL1	azimuthal angle	int64	$[0;\ 2^{10} - 1]$	$2\pi/576 \sim 0.011$ rad	$[0;\ 2\pi]$ rad
hwEtaL1	pseudorapidity	int64	$[-2^8;\ 2^8 - 1]$	$0.0870/8 = 0.010875$	$[-2.45;\ 2.45]$
hwSignL1	particle charge	int64	$\{0, 1\}$	-	-

The reconstructed values have been generated with the so-called Kalman filter Barrel Muon Track Finder (K-BMTF), which uses a Kalman filter based reconstruction algorithm that will be deployed for LHC Run 3 . Table 2 shows the details of all the reconstructed variables contained in the dataset. Only three of them, namely *ptReco, etaVtxReco* and *phiVtxReco* are used as targets for the muon re-calibration problem.

Table 3 – Variable specifications for the reconstructed muon track parameters

Name	Description	Dtype	Range
ptReco	reconstructed transverse momentum	float	$[0, 255]$ GeV
etaExtRecoSt1	reconstructed pseudorapidity (Extrapolated to Station 1)	float	$[-2.45;\ 2.45]$
phiExtRecoSt1	reconstructed azimuthal angle (Extrapolated to Station 1)	float	$[-\pi;\ \pi]$ rad
etaVtxReco	reconstructed pseudorapidity (Vertex)	float	$[-2.45;\ 2.45]$
phiVtxReco	reconstructed azimuthal angle (Vertex)	float	$[-\pi;\ \pi]$ rad
dXYReco	reconstructed collision point XY distance	float	-
chargeReco	reconstructed particle charge	int64	$-1,1$

Each muon has previously been matched to reconstructed values. L1 measurements corresponding to unmatched muons are reported in specific columns, whose labels contain the original feature name followed by the 'Unmatched' keyword. As a result, only a fraction of muons per row (described by the variable *nMatchedMuons*) have a corresponding match. The reconstructed values of unmatched muons are zero-padded (values set to zero).

The muon re-calibration problem consists of correcting the L1 measurements in order to match the offline reconstructed values. Starting from the original dataset, a new dataset has been generated in order to retain only the relevant measurements for the muon re-calibration problem. For each matched muon, a single independent row is created into the new dataset containing its four L1 measurements along with the three corresponding reconstructed values (*ptReco, etaVtxReco, phiVtxReco*). The new dataset contains 1 336 160 rows.

The distributions of the L1 measurements (Global Muon Trigger) versus the offline reconstructed values in the resulting dataset are reported in Figure 5. As can be seen from the upper row, the distributions of ϕ and η present a few different peaks, whereas the distribution of p_T presents a difference over the tail. In the

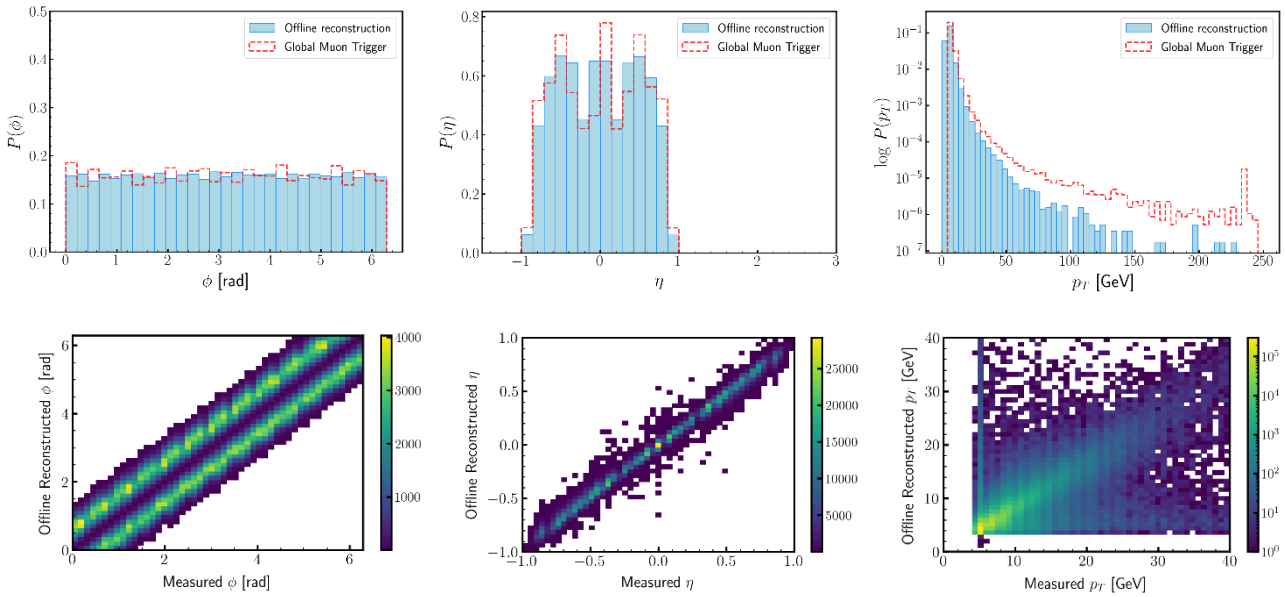

Figure 5 – Two different visualisation of the distributions of L1 measurements vs offline reconstructed values

second row, 2D histograms give a more detailed view of how the divergence of single values is distributed across the features range. The histogram of η is the only one which more closely resembles a straight line (i.e., the ideal outcome). The double band in the histogram of ϕ is explained by the fact that differently charged muons are bent in opposite directions by the magnetic field. In the histogram of p_T, the majority of the points are concentrated around zero, with the remaining ones mainly scattered along the two axes.

b. Jets

Jet track parameters have been collected from a preprocessed version of the ZeroBias (ZB) dataset with Charged Hadron Subtraction applied [8]. The structure of the dataset is similar to the one described in the previous section. Each row contains up to twelve L1 jet measurements, with each jet represented by three features. The specifications of jet features are reported in Table 4. Each row also contains 140 reconstructed values (denoted by *PF* in the column names, which stands for Particle Flow, which is the name of the algorithm used for reconstruction [13]).

Table 4 – L1 jet measurement specifications

Name	Description	Dtype	Range	Bin width	Min-Max
hwPt	transverse momentum	int64	$[0;\ 2^{11}-1]$	0.5 GeV	$[0;\ 1023]$ GeV
hwEta	pseudorapidity	int64	$[-2^7;\ 2^7-1]$	$0.0870/2 = 0.0435$	$[-5;\ 5]$
hwPhi	azimuthal angle	int64	$[0;\ 2^8-1]$	$2\pi/144 \sim 0.044$ rad	$[0;\ 2\pi]$ rad

The following procedure was applied to match L1 measurements to reconstructed values:

1. For each reconstructed - L1 measurement pair compute $\Delta R = \sqrt{\Delta\phi^2 + \Delta\eta^2}$
2. Each reconstructed jet is matched to the L1 jet corresponding to the minimum ΔR value as long as that minimum is below a threshold $\Delta R < 0.4$
3. Each L1 jet is allowed to be matched to multiple reconstructed jets

Figure 6 shows the distributions of the offline reconstructed values ϕ, η, p_T of matched versus unmatched jets. It can be seen that a lot of the unmatched reconstructed jets are at high $|\eta|$, whereas the reconstructed jets that have high p_T are mostly matched.

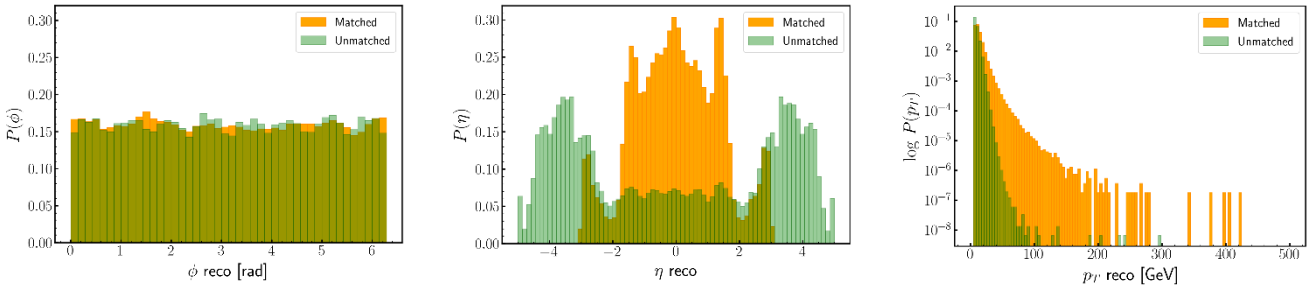

Figure 6 – Offline reconstructed ('reco') values: matched vs unmatched jets

The same matching procedure has also been applied to a subset of the dataset, determined by a preliminary cut $|\eta| < 3$. Quantitative results for both datasets are reported in Table 5. As confirmed by the previous plot, the number of unmatched jets significantly decreases when the cut $|\eta| < 3$ is applied. The last column contains the percentage of matched p_T, i.e., the ratio between the sum of p_T of all the matched jets and the sum of p_T of all the reconstructed jets.

Table 5 – Quantitative results of the jet matching procedure

	Number of matched jets	Number of unmatched jets	Percentage of matched jets	Percentage of matched p_T
No cut	1 383 135	35 546 552	3.75 %	5.47 %
$\|\eta\| < 3$	1 278 076	865 802	59.62 %	61.52 %

The jet re-calibration problem consists of correcting the L1 p_T measurement in order to approximate the sum of the matched reconstructed p_T. Starting from the original dataset, a new one has been generated in order to retain only relevant information for the jet re-calibration problem. For each L1 jet and its corresponding matched reconstructed jets, a single independent row is created into the new dataset containing the three L1 measurements along with the sum of the reconstructed p_T. The new dataset contains 1 234 480 rows.

Figure 7 shows a qualitative analysis of the outcome of the p_T matching procedure. On the top left panel it is possible to observe that the distribution of the matched p_T closely follows the distribution of the L1 measurements. The histogram on the top right panel shows that most of the matched values are concentrated at around $p_T < 20$ GeV. On the bottom left plot, the distribution of the residuals between the L1 measurements and the reconstructed value is shown having a bell-shaped curve around zero and a reported root mean square error of 1.102. The last plot on the bottom right shows that most of the L1 jets are matched to just one reconstructed jet, whereas a smaller quantity is matched to a couple of reconstructed jets. Only a few L1 jets are matched to three reconstructed jets.

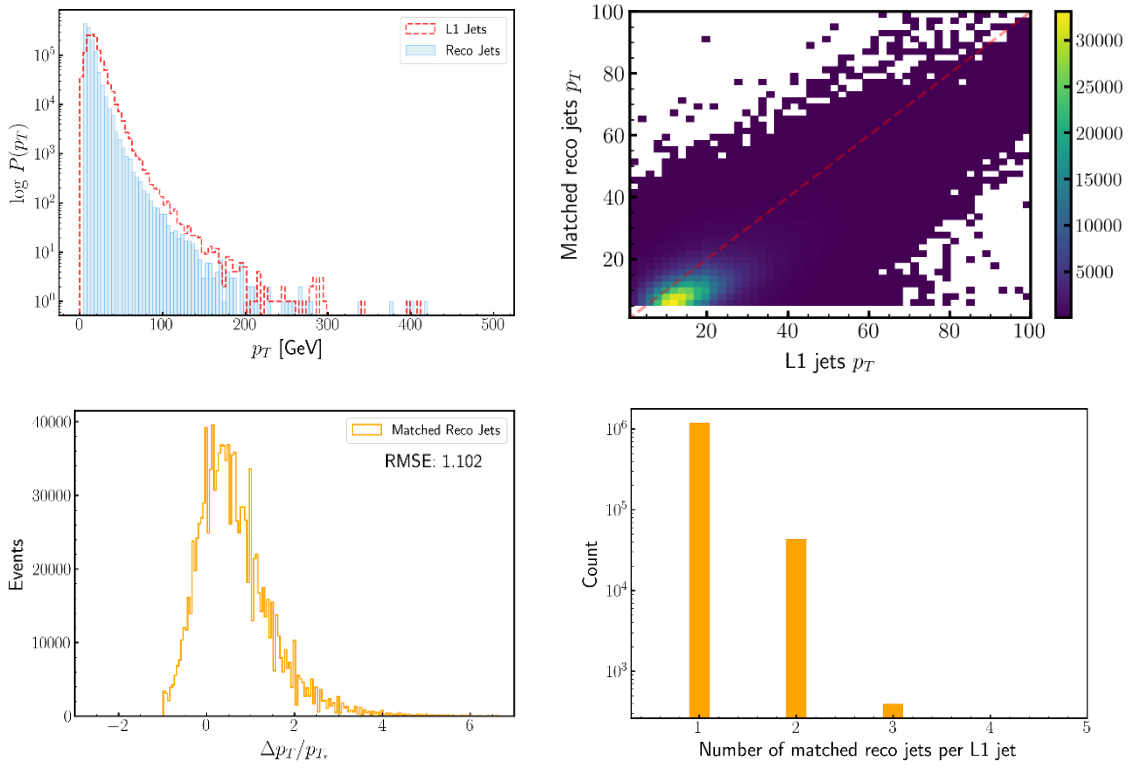

Figure 7 - Analysis of the outcome of the p_T matching procedure

3. MACHINE LEARNING MODELS

Machine learning is the study of computer algorithms that allow computer programs to automatically improve through experience [9]. Supervised learning is a class of machine learning problems whose goal is to find an input-output relation

$$f: X \rightarrow Y$$

given examples from a training set $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ such that it can be generalized to any new data point.

The problem of particle track re-calibration presented in this project can be formally described as a multivariate regression problem, i.e.

- the input space is $\mathbb{R}^D$, with D being the number of input track parameters
- the output space is $\mathbb{R}^T$, with T being the number of parameters to be re-calibrated

The solution f is computed by finding an approximate solution to

$$\arg\min_f \sum_{i=1}^{n} \ell(f(x_i), y_i)$$

by means of iterative methods based on the gradient of a loss function ℓ which measures the approximation error.

In the following sections, two different machine learning models are briefly presented with a preliminary analysis on the approximate number of operations required at inference time.

c. NEURAL NETWORKS

Neural networks are machine learning models which are loosely inspired by information processing in the human brain. A neural network is composed by a set of layers, called hidden layers, composed of a certain number of computational units, called neurons. Each layer takes as input the output of the previous layer and transforms it by a function

$$h = g(Wx + b)$$

where x is the input, g is a non-linear activation function, W is a matrix of weights and b is an array of biases.

Starting from raw data, neural networks can extract efficient intermediate representations and approximate any function of "arbitrary" complexity [10].

Figure 8 shows a schematic representation of the composition of hidden layers as matrix multiplications. The approximated number of operations required at inference time is:

$$O(SH_n(D + N_l H_n + T))$$

where S is the number of input points, N_l is the number of hidden layers, H_n is the number of neurons per layer, D and T are the input and output dimensionality, respectively.

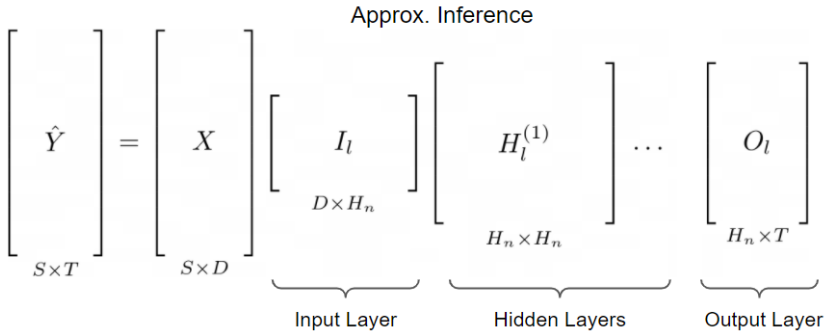

Figure 8 – Schematic representation of the inference operations required by a neural network

d. FALKON

Falkon is a fast, efficient large-scale kernel method developed at MaLGa Center, Genoa [11]. The approximated function is

$$f(x) = \sum_{i=1}^{M} \alpha_i \, k(x, x_i)$$

where $k(x, x_i)$ is a kernel function which acts as a similarity measure between input points, α_i are a set of learned coefficients, M is the number of Nystrom centers which constitute an arbitrary subset of the training data.

The function f is computed by solving a kernel ridge regression problem. The algorithm is optimized for GPUs and makes use of several techniques, namely Nystrom approximation, conjugate gradient, and preconditioning to compute a fast iterative solution.

Figure 9 shows the inference operation required by Falkon. It is mainly a matrix multiplication between the kernel matrix K_{SM} – that is the matrix containing the kernel function computed between the S input points and the M Nystrom centers – and the matrix of learned coefficients A. T denotes the output dimensionality.

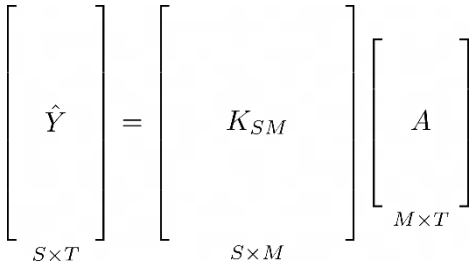

Figure 9 - Inference operation required by Falkon

The approximate number of operations required at inference time is thus:

$$O(SMT) + \text{number of kernel operations}$$

4. EXPERIMENTS AND RESULTS

In the following sections, details of the experiments on the re-calibration of the L1 trigger parameters are reported. The two machine learning models described above are trained to predict a correction to be applied to the L1 measurements, in order to have a more precise value in the future L1 scouting system.

Muon or jet track parameters from real data are used as training inputs. The differences between L1 measurements and offline reconstructed values are used as targets, defined as follows:

$$\Delta p_T = p_{T_{\text{pred}}} - p_{T_{\text{reco}}}$$

$$\Delta\eta = \eta_{\text{pred}} - \eta_{\text{reco}}$$

$$\Delta\phi = \begin{cases} \phi_{\text{pred}} - \phi_{\text{reco}} - 2\pi, & \phi_{\text{pred}} - \phi_{\text{reco}} > \pi \\ \phi_{\text{pred}} - \phi_{\text{reco}} + 2\pi, & \phi_{\text{pred}} - \phi_{\text{reco}} < -\pi \\ \phi_{\text{pred}} - \phi_{\text{reco}}, & \text{otherwise} \end{cases}$$

a. MUON RE-CALIBRATION

Data pre-processing. The dataset for muon re-calibration described in section 2.a has been processed as follows:

- Duplicates have been discarded (i.e., rows having the same values for all the variables)
- L1 integer parameters ($hwPtL1$, $hwPhiL1$, $hwEtaL1$) have been multiplied by their corresponding bin widths in order the values in physical units.
- A filter based on the measured transverse momentum has been applied: only muons with $5.5 < hwPtL1 < 45$ GeV are considered for further analysis.
- The offline reconstructed azimuthal angle ($phiVtxReco$) has been rescaled to the interval $[0;\ 2\pi]$ in order to match the range of the L1 measurements.

Figure 10 shows an input-output diagram for the muon re-calibration problem.

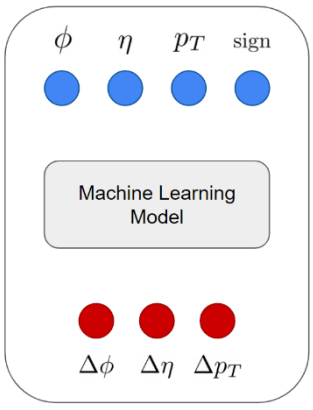

Figure 10 – Input-output diagram for the muon re-calibration problem

The dataset has been randomly split into training, validation and test set, with 65-20-15% of the data in each set, respectively. A 0-1 normalization has been applied to the features, whereas a standard scaling has been applied to the targets.

Neural networks. Different neural network models have been trained with TensorFlow [12]. Each neural network is composed of N_l hidden layers made of repeated building blocks, namely a dense layer with a constant number H_n of neurons, a batch normalization layer and a ReLU activation function. Concerning the learning procedure, the Adam optimizer has been used with default parameters and different batch sizes (256, 512, 8192). The early stopping method has been employed to avoid overfitting.

Table 6 – List of tested neural network models

Model ID	Loss function	Regularization	Number of hidden layers N_l	Number of neurons per hidden layer H_n
1	MSE	-	4	128
2	MSE	10^{-5}	4	128
3	Logcosh	-	4	128
4	Logcosh	10^{-5}	4	128
6	MSE	-	3	32
7	MSE	10^{-5}	3	32
8	Logcosh	-	3	32
9	Logcosh	10^{-5}	3	32

Figure 11 shows the results of the best neural network model, identified by ID 4 in the previous table.

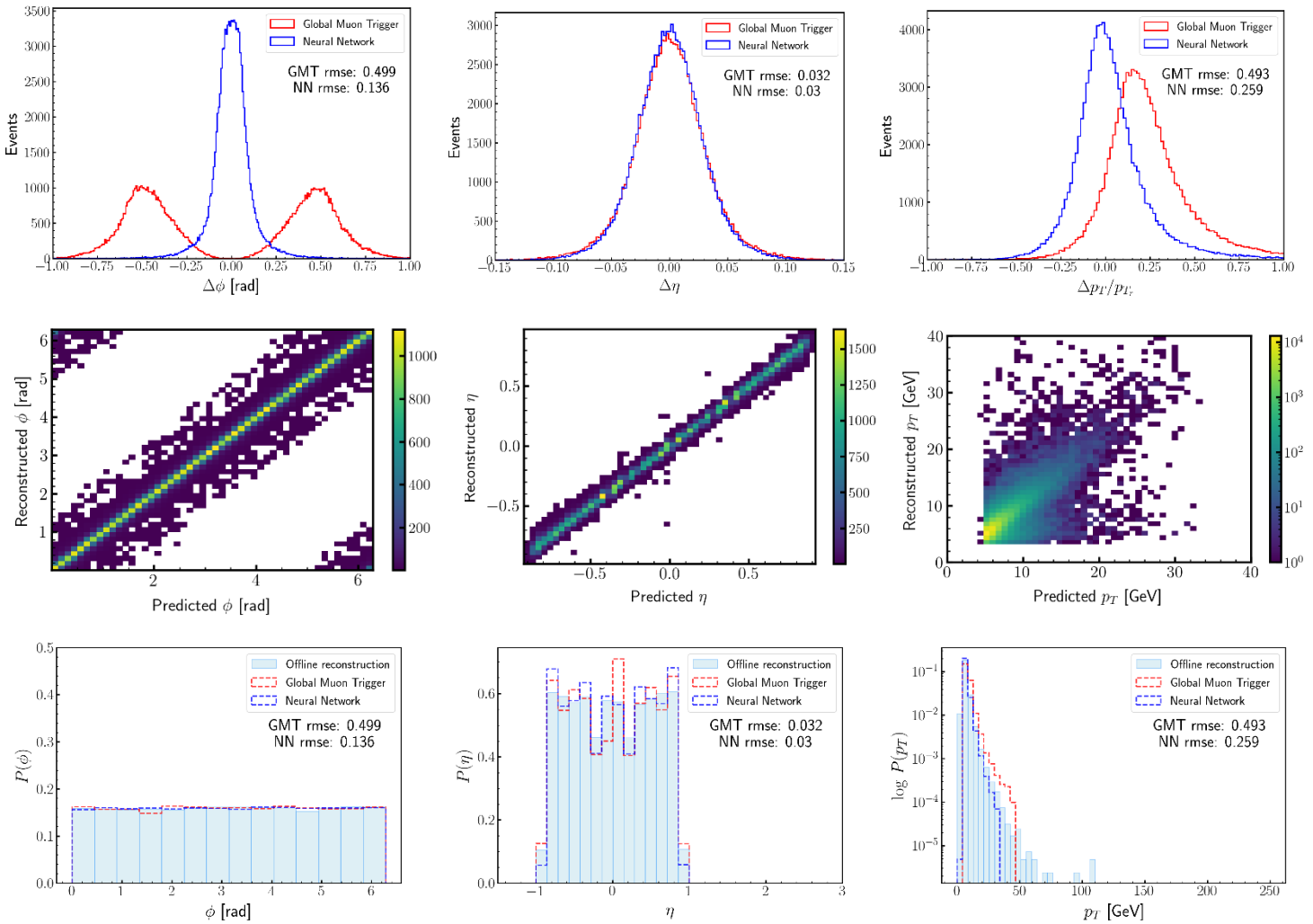

Figure 11 – Results of muon re-calibration with the best neural network model
GMT = Global Muon Trigger, NN = Neural network, rmse = root mean square error

The first row shows the residuals between L1 measurements and offline reconstructed values. A remarkable improvement can be observed in $\Delta\phi$ and Δp_T, with the curve being clearly shifted around zero. A slight improvement can also be observed in $\Delta\eta$. These results are confirmed by the 2D histograms on the second row, which appear improved compared to the ones in Figure 5. Finally, on the third row the distributions of the L1 measurements (Global Muon Trigger), the predicted variables and the offline reconstructed values are shown.

Falkon. Different kernels functions have been tested with several combination of hyperparameters and random Nystrom centers selection:

- Gaussian kernel: $k(x, x_i) = \exp\left(\frac{||(x-x_i)||^2}{2\sigma^2}\right)$
- Linear kernel: $k(x, x_i) = \beta + \frac{1}{\sigma^2} x^T x_i$
- Polynomial kernel: $k(x, x_i) = (\alpha x^T x_i + \beta)^{\text{degree}}$

Figure 12 shows the tuning curves for the Gaussian kernel. By fixing two parameters, it is possible to see, in turn, how the third one affects the accuracy of the re-calibration and consequently choose the best combination that minimizes the root mean square error.

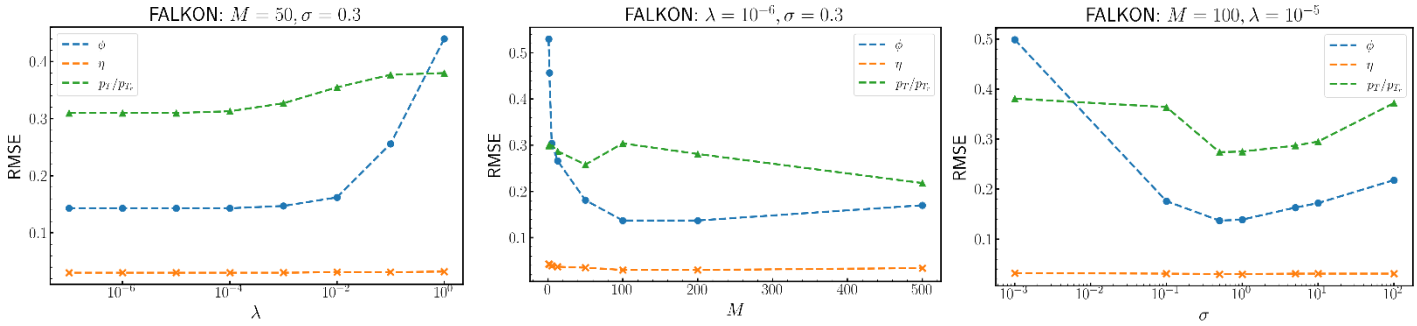

Figure 12 – Hyperparameter tuning with the Gaussian kernel

Figure 12 shows the effect of increasing the parameter M with linear and polynomial kernels. All the three curves are stable for $M > 100$.

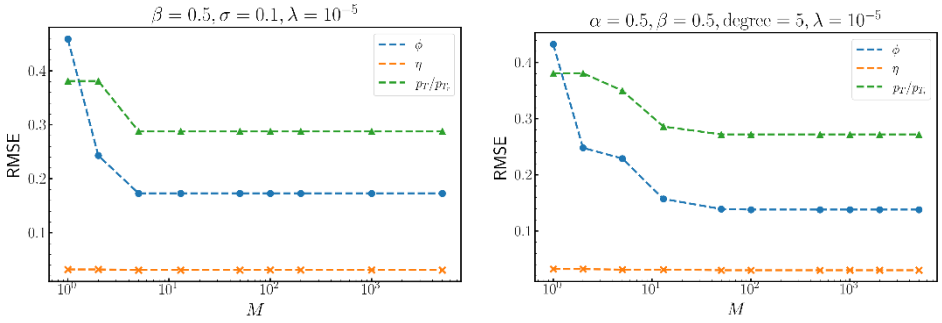

Figure 13 - Tuning the parameter M with (left) Linear kernel, (right) Polynomial kernel

As discussed in the section 2.d, the parameter M plays a crucial role in the number of operations required at inference time. The tuning curves reported before show that all the three kernels can produce good results with a relatively low M (compared to other applications [10]).

Qualitative results obtained with the best combination of hyperparameters for all three kernels are very similar to the ones shown in Figure 11.

Numerical results. Table 7 shows the Root Mean Square Error (RMSE) of the re-calibrated muon track parameters with respect to the offline reconstructed values. It is possible to notice that all the machine learning models provide a similar improvement compared to the Global Muon Trigger (GMT), whose error is reported in the first row.

Table 7 – Performance of the ML models compared to the Global Muon Trigger (GMT)

	ϕ	η	$p_T/p_{T_{\text{reco}}}$
GMT	0.499	0.032	0.493
Neural network: 128x4	**0.136**	**0.030**	**0.259**
Neural network: 32x3	**0.136**	**0.030**	0.274
Falkon with Gaussian kernel: M=500	**0.136**	**0.030**	0.270
Falkon with Gaussian kernel: M=100	0.137	**0.030**	0.304
Falkon with Linear kernel: M=100	0.173	0.031	0.288
Falkon with Polynomial kernel: M=100, deg=5	0.138	0.030	0.272

A computation of the approximate number of operations per inference is reported in Table 8. Using Falkon with the linear kernel has the benefit of reducing by orders of magnitudes the number of floating-point operations (flops) required - compared to the two neural network models.

Table 8 – Approximated number of floating-point operations (flops) per inference

	Approx. flops (per inference)
Neural Network 128x4	66 432
Neural Network 32x3	3296
Falkon M=500	2000
Falkon M=100	**300**

b. JET RE-CALIBRATION

The dataset for jet re-calibration described in section 2a has been used to conduct preliminary tests with both machine learning models described so far. Figure 15 shows on the left the input-output scheme for the jet re-calibration problem, on the right an example of results obtained with the neural network model denoted by ID 6 in Table 4. The p_T reconstruction looks improved, with the Root Mean Square Error reduced from 1.102 to 0.52. This is an encouraging result which suggests that after further investigations the method could be efficiently extended to calorimeter data.

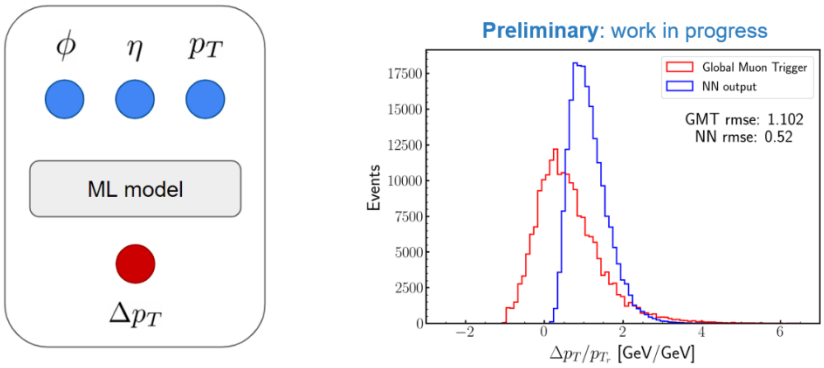

Figure 14 – (Left) Input-output scheme of the jet re-calibration problem, (Right) Preliminary results

5. CONCLUSIONS & FUTURE WORK

Neural networks and Falkon are two different Machine Learning methods which have potential for future use within the CMS L1 scouting system. Several tests have shown that they can be efficiently employed for the re-calibration of muon track measurements provided by the L1 trigger. With carefully selected parameters, Falkon can preserve accuracy and provide advantages in terms of inference time – a crucial aspect needed to meet strict latency requirements of the L1 scouting system.

The results of muon re-calibration can be easily extended to calorimeter data. Preliminary tests have shown encouraging results with both ML models for improvement in jet transverse momentum. Furthermore, for both neural networks and Falkon there is also the possibility to implement more refined strategies in the training procedure which can be used to give more importance to rare particles, crucially important areas of phase space for physics studies.

References

[1] CMS Collaboration, The CMS experiment at the CERN LHC, Aug 2008, JINST 3 S08004, doi:10.1088/1748-0221/3/08/S08004.

[2] CMS Collaboration, The CMS trigger system, JINST 12, P01020 (2017), doi:10.1088/1748-0221/12/01/P01020.

[3] CMS Collaboration, The Phase-2 upgrade of the CMS Level-1 trigger, CERN-LHCC-2020-004;CMS-TDR-021 (2020).

[4] Badaro, G., Behrens, U., Branson, J., Brummer, P., Cittolin, S., Da Silva-Gomes, D., ... & Zejdl, P. (2020). 40 MHz Level-1 Trigger Scouting for CMS. In EPJ Web of Conferences (Vol. 245, p. 01032). EDP Sciences.

[5] CMS Collaboration, D. Golubovic et al., CTD2020: 40 MHz Scouting with Deep Learning in CMS, Zenodo, Apr, 2020

[6] CMS Collaboration, Zero bias and HF-based minimum bias triggering for pp collisions at 14 TeV in CMS, tech.rep., CERN, Feb, 2009.

[7] CMS Collaboration, Scales for inputs to uGT, http://globaltrigger:hephy:at/files/upgrade/ugt/scales inputs 2 ugt 2017Aug14:pdf, [Accessed 2021-07-08].

[8] CMS Collaboration, Study of Pileup Removal Algorithms for Jets, CERN-CMS PAS JME-14-001

[9] Murphy, K. P. (2012). Machine learning: a probabilistic perspective. MIT press.

[10] Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. Mathematics of control, signals and systems, 2(4), 303-314.

[11] Meanti, G., Carratino, L., Rosasco, L., & Rudi, A. (2020). Kernel methods through the roof: handling billions of points efficiently. arXiv preprint arXiv:2006.10350.

[12] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... & Zheng, X. (2016). Tensorflow: A system for large-scale machine learning. In 12th USENIX symposium on operating systems design and implementation, 2016, pp. 265-283.

[13] James, Thomas Owen. A Hardware Track-Trigger for CMS: At the High Luminosity LHC. https://cds.cern.ch/record/2647214

[14] CMS Collaboration, Particle-flow reconstruction and global event description with the CMS detector, https://arxiv.org/abs/1706.04965