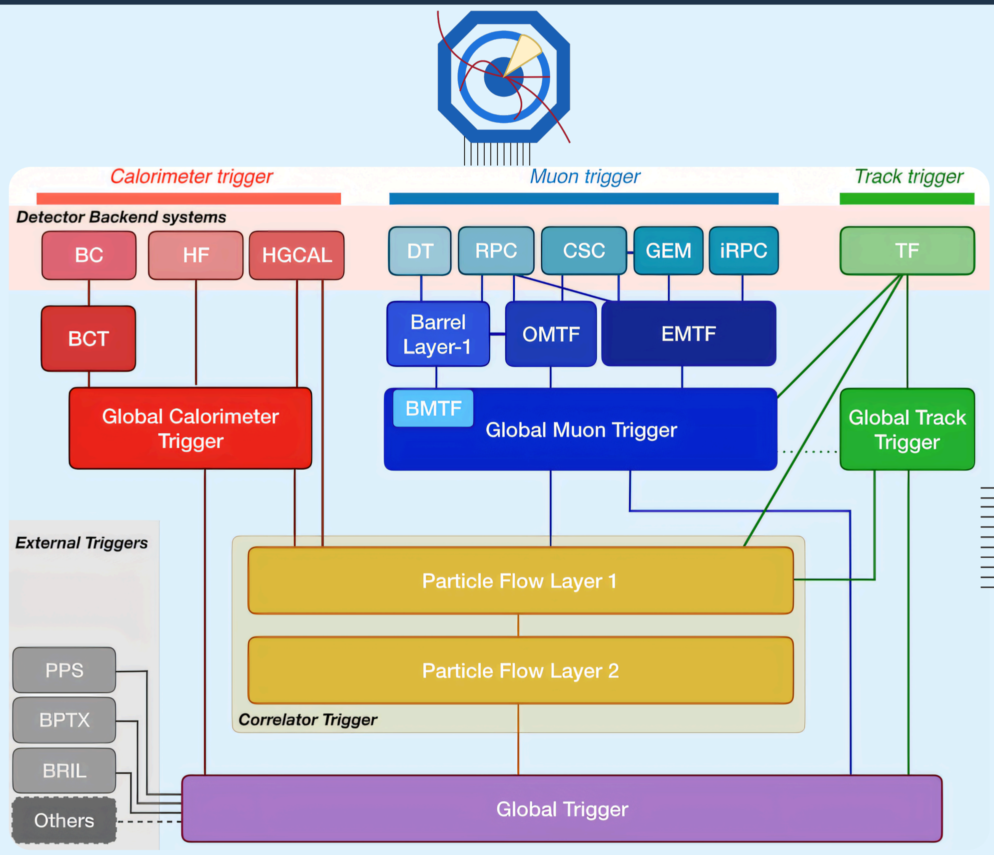
FPGA-Optimized ML for low-latency electron identification and p_T regression with the CMS Phase-2 L1 trigger

Piero Viscone[1,2] on behalf of the CMS Collaboration
[1] CERN, [2] University of Zürich
piero.viscone@cern.ch

1 CMS Phase-2 L1 trigger

The CMS **Level-1 Trigger (L1T)** is a **real-time** system that selects the most relevant collision events to reduce the data for further processing. For Phase-2 of the LHC, the trigger is being upgraded to handle higher luminosity.

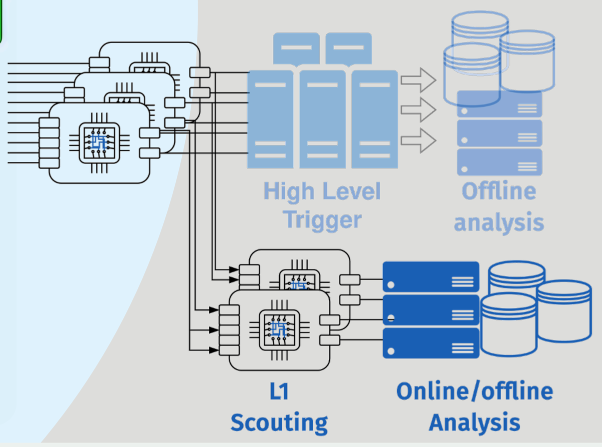
- Access to both **calorimeter and tracking information**.
- Extended **latency** frame of **12.5 µs**
- The **correlator trigger** (CT) creates **physics objects** candidates combining information from multiple subdetectors.



2 L1 Scouting

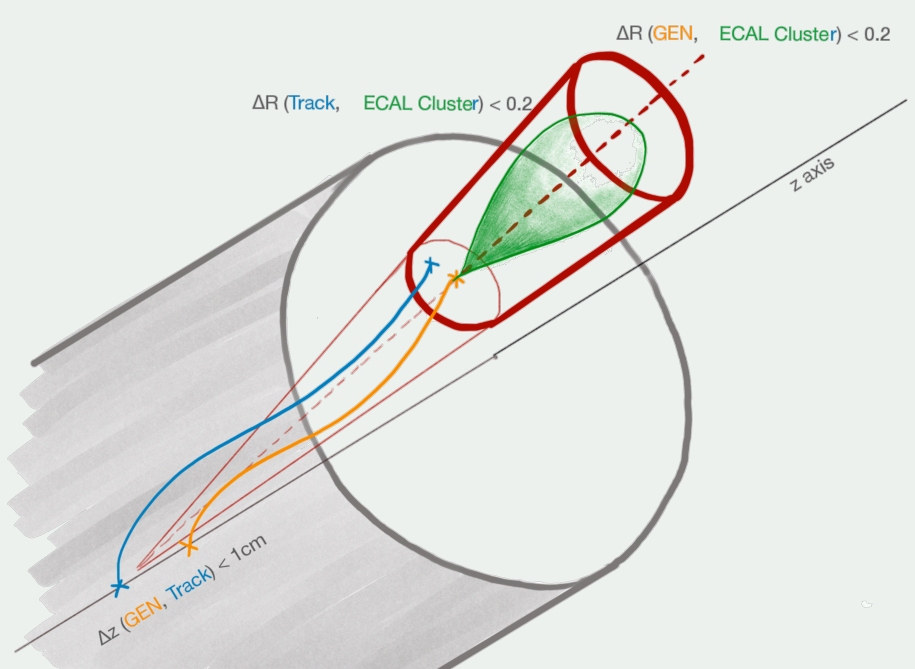
40MHz **online analysis system** that complements the L1 Trigger capturing the events that would be lost otherwise.

It exploit improved efficiency of low-pt trigger objects not fitting standard menu (too much for rate budget)



3 Track-matched electron candidates

Track-matched Electron candidates (TkEle) are built by the **correlator** trigger, linking tracks to calorimiter clusters with an **elliptical matching** in η-ϕ



Baseline approach:

- $p_T^{Tk} > 10$ GeV
- (|Δη|<0.025, |Δϕ|<0.07) for |η| < 0.9
  (|Δη|<0.015, |Δϕ|<0.07) for 0.9 < |η| < 1.479

  Tight geometrical match (Elliptic)

Novel approach:

- Loose elliptic march (|Δη|<0.03, |Δϕ|<0.3)
- Candidates classified by a BDT **(Composite ID)**

Composite ID input features

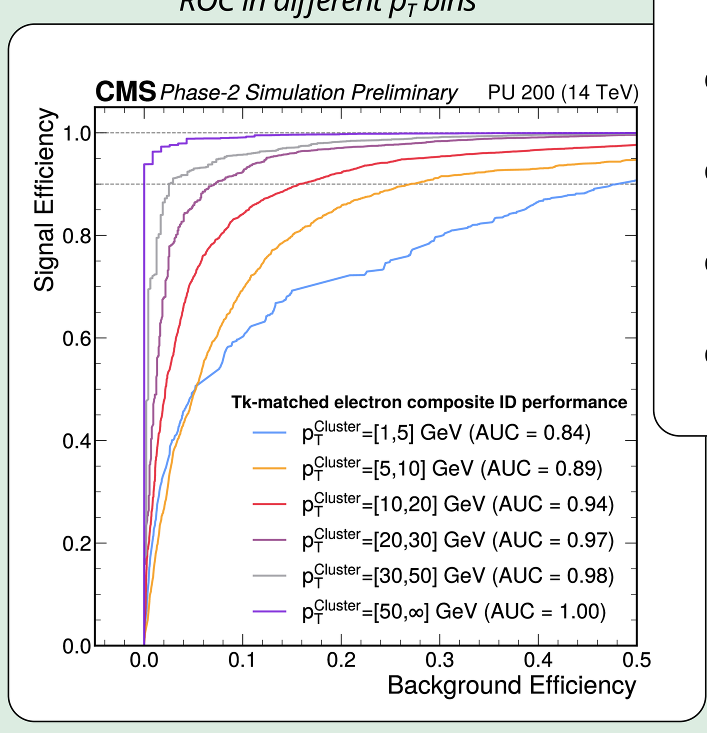
4 Identification: Composite ID

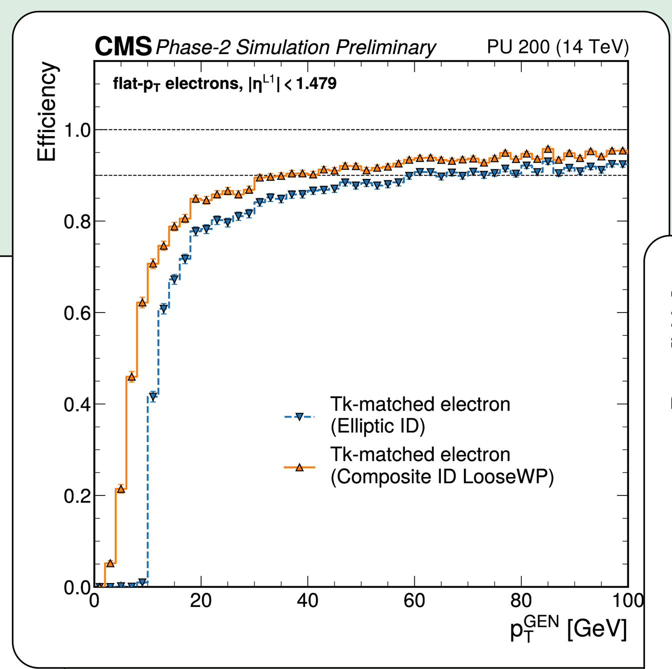
BDT classifier

- **Signal: TkEle matched to generated electrons** in a flat-p_T, flat-η sample at PU200
- **Background:** TkEle candidates in a PU200 **MinBias** sample

Background candidates are **reweighted** in order to **flat the p_T distribution** to avoid the model from imposing a tight cut on the p_T

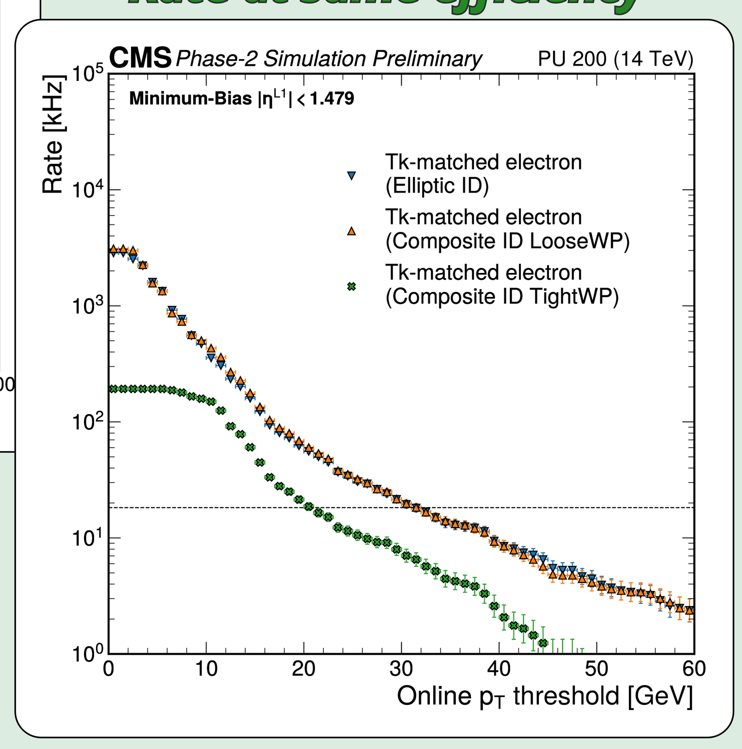
ROC in different p_T bins





Efficiency at same rate

Rate at same efficiency



Quantization & Implementation

The models have been synthetized in firmware using Conifer and Vivado™ HLS

All the input features are rescaled to the [-1,1] interval to use less bits

```
cfg = {
  "Input Precision" : ap_fixed<8,1>,
  "Score Precision" : ap_fixed<11,4>,
  "Board" : "Xilinx Virtex UltraScale+ VU13P",
  "Clock" : 4.166 #ns (240 MHz)
}
```

	BRAM	DSP	FF	LUT
SLR	0.0%	0.0%	0.9%	7.5%
Total	0.0%	0.0%	0.02%	1.9%

	Clock cycles	Latency
Clock (4.166 ns)	7	29.17 ns

```
cfg = {
  "Input Precision" : ap_fixed<11,1>,
  "Score Precision" : ap_fixed<12,3>,
  "Board" : "Xilinx Virtex UltraScale+ VU13P",
  "Clock" : 4.166 #ns (240 MHz)
}
```

	BRAM	DSP	FF	LUT
SLR	0.0%	0.0%	1.13%	6.2%
Total	0.0%	0.0%	0.03%	1.5%

	Clock cycles	Latency
Clock (4.166 ns)	6	25.0 ns

5 p_T regression

Target y_i: p_T^{Gen}/p_T^{TkEle} (output of the model = pT corrective factor)

$$\mathscr{L} = \sum_{i=1}^{N} \underbrace{w_i^{(\mathbf{cls}_i)}}_{\text{flat-}p_T\text{ reweight}} \left[\overbrace{\mathbf{cls}_i \frac{\left| y_i - p_{T_i}^{Gen}/p_{T_i}^{Calo} \right|}{\underbrace{\text{Med}\left[\left|1 - p_{T_j}^{Gen}/p_{T_j}^{Calo}\right|\right]}_{\text{Heteroscedastic-adjustment term}}}}^{\text{L1 signal regression term}} + \overbrace{(1-\mathbf{cls}_i) \cdot \underbrace{\left(\alpha + \beta \, \text{Min}\left[p_{T_i}^{Calo}, p_T^{thr}\right]\right)}_{\text{Penalization linear in } p_T \text{ (up to } p_T^{thr})} \cdot \underbrace{\text{Max}\left[0, y_i - t\right]}_{\text{Penalize output} \geq t}}^{\text{Background penalization term}} \right] \quad \forall j \in \text{Bin}\left[p_{T_i}^{Gen}\right]$$

$\alpha = 8$, $\beta = 1.05$ GeV^{-1}, $t = 0.975$, $p_T^{thr} = 15$ GeV

Signal term:

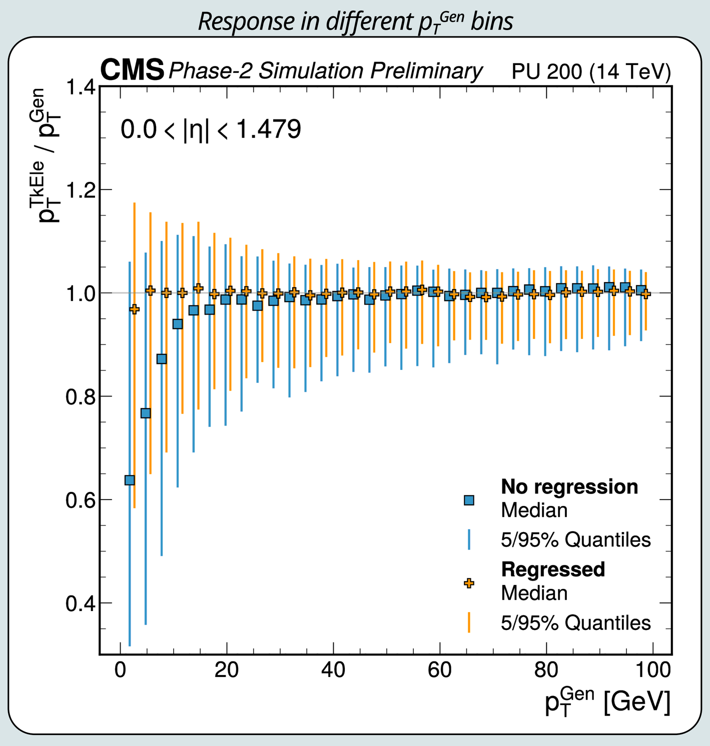
- L1 regression loss (insensitive to outliers, optimize median response)
- Heteroscedaticity-adjusted (Non constant variance)

Background term:

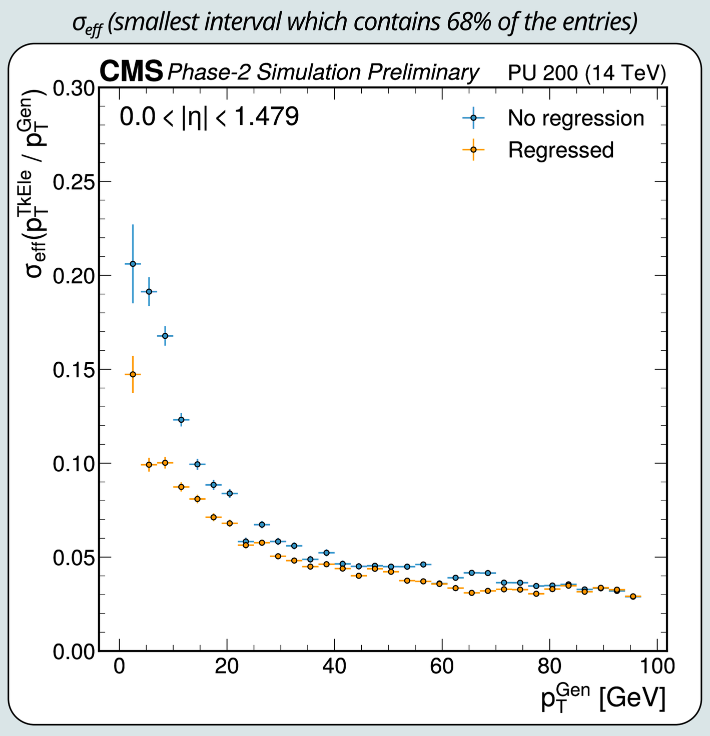
- p_T increment is penalized
- Penalization linear in p_T (up to p_T^{thr}) (Higher p_T background TkEle are signal-like+ typical trigger threshold at 12/25/36 GeV)

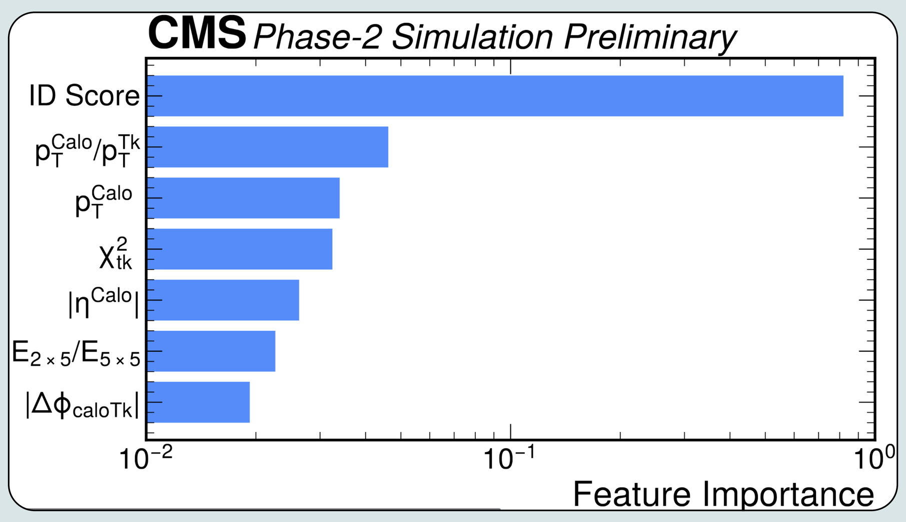
Avoid the model from increasing the p_T of background electrons more than signal electron (i.e. **increase** the **rate**)

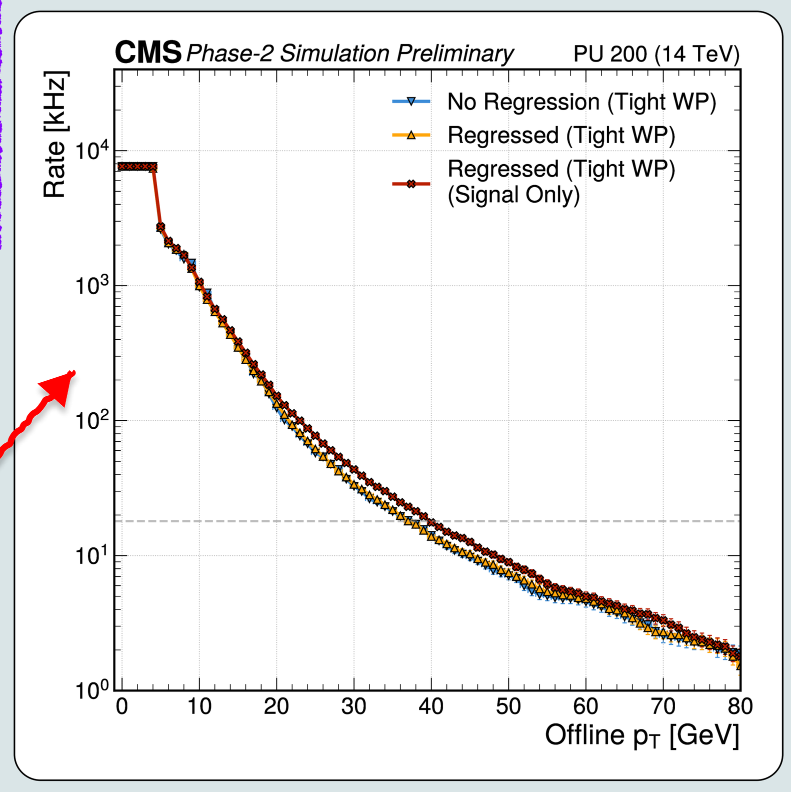
Response in different p_T^{Gen} bins



σ_{eff} (smallest interval which contains 68% of the entries)







The loss has **always null second derivative**

XGBoost employs the **Newton-Raphson** method to perform the minimization, which **require a positive definite Hessian**.

- using the L1 loss, XGBoost find the **split** of each node using a **quantile based approach** instead of Newton-Raphson
- The python API does not allow to change the split method: **the proposed loss require to be implemented in C++, recompiling XGBoost.**

L1 Scouting benchmark

Hidden Abelian Higgs model

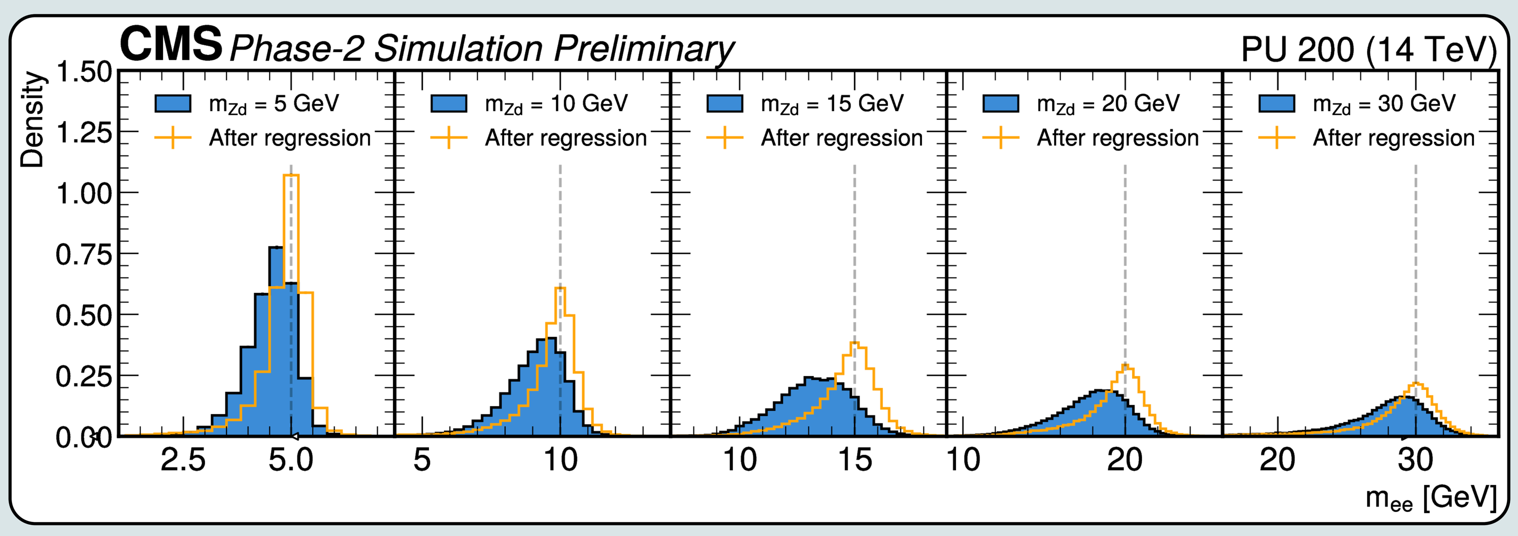
Low-mass Drell-Yan produced **Dark-Photon Z_d that decays to a pair of electrons**

5 mass points were generated.

Selections:

- TkEle matched to generated electrons
- Opposite sign pair
- $p_T > 4$ GeV
- ID score > -0.3
- $|\Delta v_z| < 0.65$

$\sigma_{eff}(m_{ee})$ 23% lower after the regression



References

[1] The Phase-2 Upgrade of the CMS Level-1 Trigger, CMS-TDR-021, https://cds.cern.ch/record/2714892

[2] Electron Reconstruction and Identification in the CMS Phase-2 Level-1 Trigger, CMS-DP-2024-098, https://cds.cern.ch/record/2916193

[3] Low-latency electron p_T regression in the CMS Phase-2 Level 1 trigger, https://twiki.cern.ch/twiki/bin/view/CMSPublic/L1TP2EleRegressionDPSNote2025