FERMILAB-SLIDES-24-0158-CMS-V

PV-Finder: ML based algorithm for primary vertex identification

Tetiana Mazurets, University of Puerto Rico

New Perspectives 2024

8 July, 2024

Outline

- Motivation
- Main concepts
  - Probability density of tracks
  - Target histograms
  - Structure of the model
  - Input to the model
- Performance
  - Comparison with a current reconstruction algorithm
  - Efficiency
- Data preparation in CMS
- Future work
- References

Motivation

- The CMS detector at the HL-LHC will face challenging conditions with expected pile-up of up to 200 collisions per bunch crossing, necessitating the development of a resilient primary vertex (PV) reconstruction method to maintain data integrity and triggering system efficiency.

- PV-Finder, a novel technique for CMS, is a ML based method for PV identification. The method is based on a DNN model trained using Kernel Density Estimations (KDEs) using target histograms based on MC ground truth of PVs.

- The approach, originally designed by the LHCb group, is currently being integrated into their trigger system. ATLAS is also considering this approach, having completed the initial model and now focusing on optimization.

Main Concepts: Probability density of tracks

- The position of each track at the point of closest approach (POCA) to the beam line can be characterized by a covariance matrix where z_0 and d_0 are the radial and longitudinal impact parameters of reconstructed tracks at the POCA:

$$\Sigma = \begin{pmatrix} \sigma^2(d_0) & \sigma(d_0, z_0) \\ \sigma(d_0, z_0) & \sigma^2(z_0) \end{pmatrix}$$

- **Probability density** of tracks is calculated:

$$P(r) = P(d, z) = \frac{1}{\sqrt{2\pi\Sigma}} \exp\left(-\frac{1}{2} \begin{pmatrix} d - d_0 \\ z - z_0 \end{pmatrix}^T \Sigma^{-1} \begin{pmatrix} d - d_0 \\ z - z_0 \end{pmatrix} \right)$$

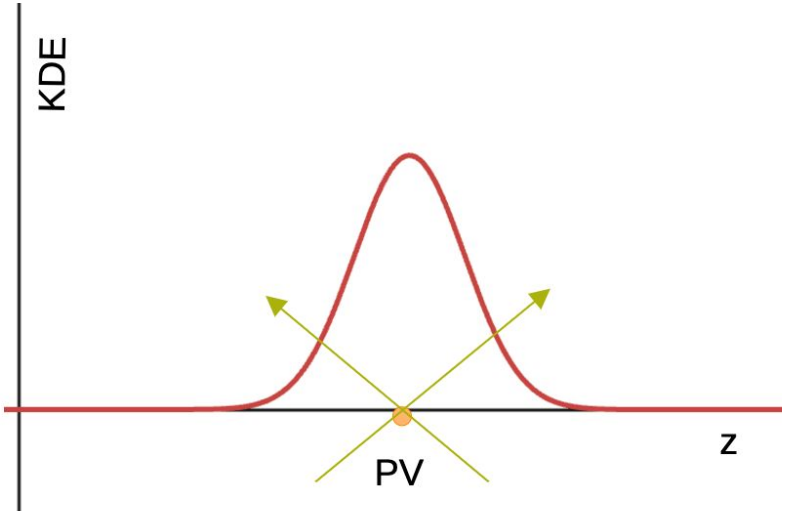


Figure 1: Simplified interpretation of KDE based on probability density

Main Concepts: Target histograms

- Target histograms: The center of Gaussian distribution is the truth PV information.
- The σ of the Gaussian is calculated using a parameterisation of the reconstructed vertex resolution (current reconstruction algorithm) as a function of the number of reconstructed tracks associated to a truth vertex. $\sigma \sim$ vertex resolution for a particular number of tracks associated with true PVs

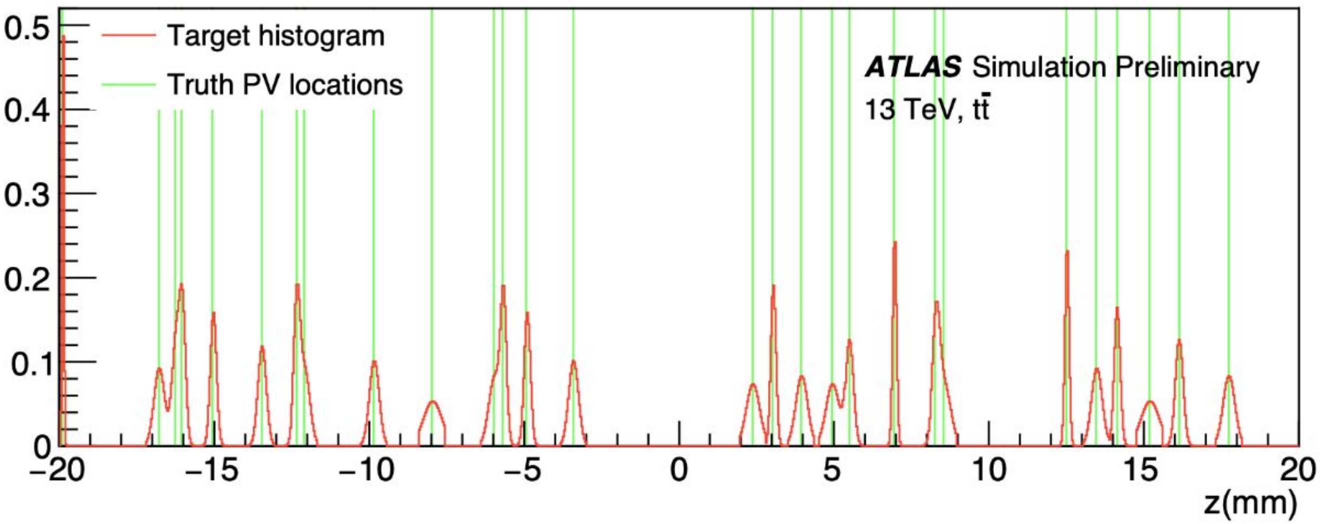


Figure 2: Target histograms with PVs truth information [1]

Main Concepts: Structure of the model

Inputs

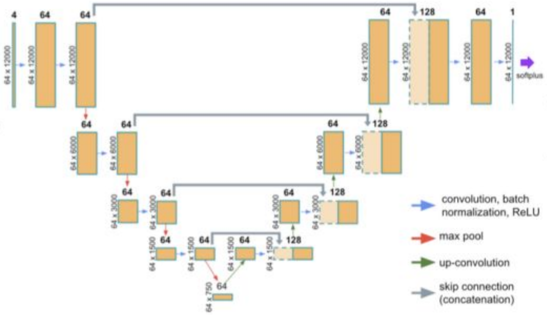
Track parameters and their uncertainties (max. (NTracks) ~ 1000) + Truth vertices

tracks-to-KDE

Analytically calculated KDE_A, KDE_B, XMax, YMax (4 1-D histgrams, 12000 bins) + Target truth histograms (1-D histogram, 12000 bins

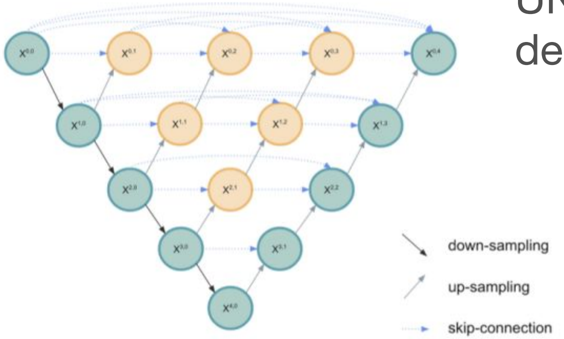
6 Fully Connected layers

UNet: KDE-to-hists


convolution, batch normalization, ReLU
max pool
up-convolution
skip connection (concatenation)
softplus

OR

UNet++: KDE-to-hists



Output

Predicted PVs (1-D histogram, 12000 bins)

Two DNNs: UNet and UNet++ (UNet with denser skip connections)

Figure 3: The full model used by LHCb (FC+UNet): tracks-to-hist [2]

Main Concepts: Input to the model

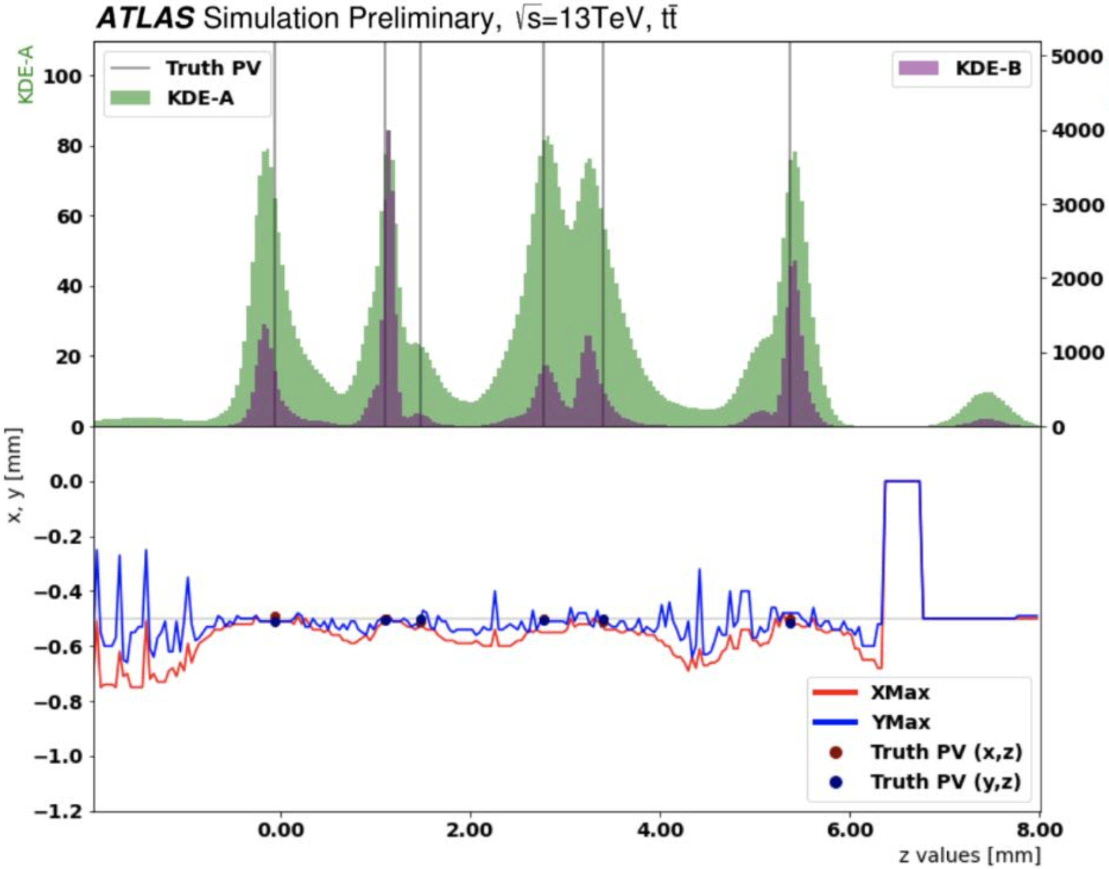


Figure 4: Input features to KDE-to-hists model [3]

KDE-A ≈ Σ probability densities

KDE-B ≈ Σ (probability densities)^2

XMax, YMax - location of the maximum summed track probability in x and y (mm)

Performance: Comparison with a current reconstruction algorithm

In [1] the $\sigma_{\text{vtx-xtx}}$ was calculated by measuring the difference in z between pairs of closely reconstructed primary vertices. Approximated with the function:

$$y = \frac{a}{1 + \exp(b \cdot (R_{cc} - |x|))} + c$$

where a, b, c are free fitting parameters, x is equivalent to $\Delta z_{\text{vtx-xtx}}$ and R_{cc} is the cluster–cluster resolution, defined as the half-width at the half-depth of the dip.

Method	$\sigma_{\text{vtx-xtx}}$ (mm)
PV-Finder UNet	0.23 ± 0.01
PV-Finder UNet++	0.37 ± 0.01
AMVF (current reco algorithm in ATLAS)	0.76 ± 0.02

Table 1: Vertex-vertex reconstruction results [1]

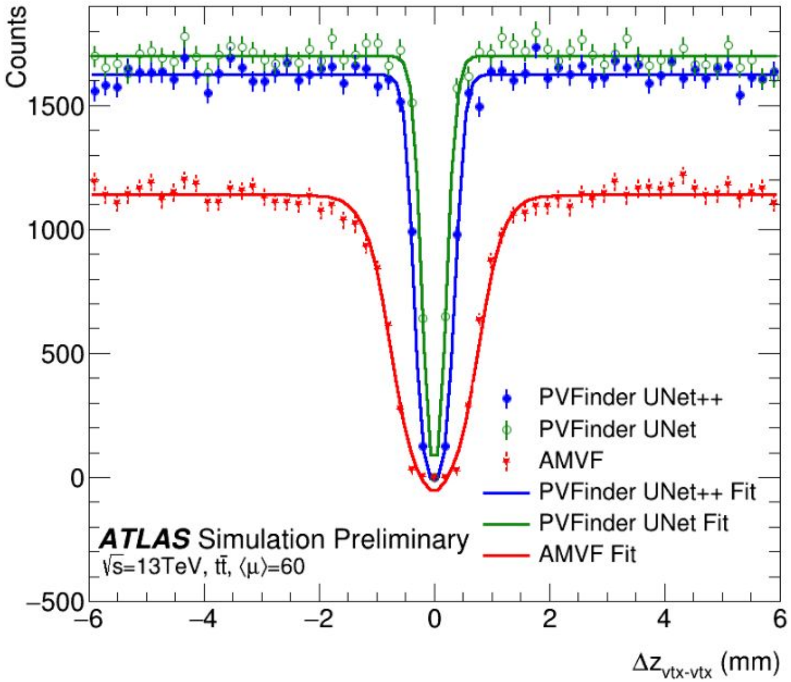


Figure 5: Distribution of the pairs separation of all nearby reconstructed primary vertices [1]

Performance: Comparison with a current reconstruction algorithm

PV-Finder reconstructs more total, more clean and less merged vertices with a slight increase in the number of fake vertices.

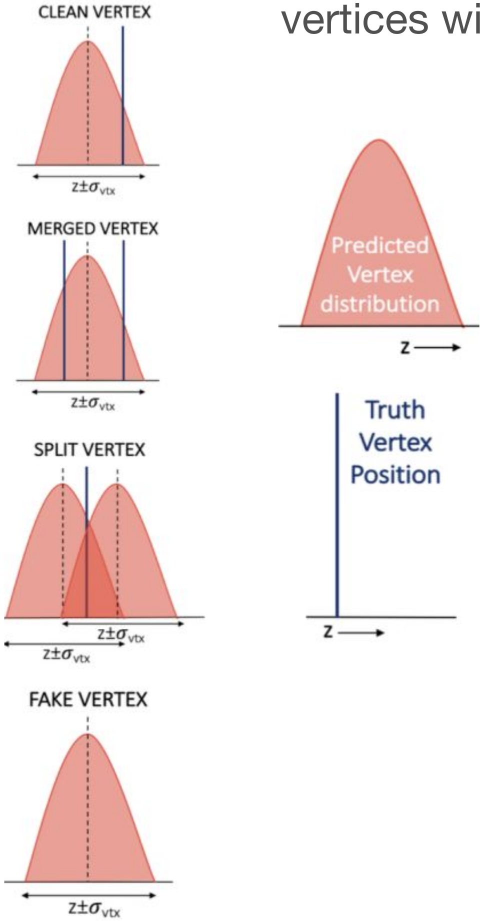


Figure 6: Vertex classification [3]

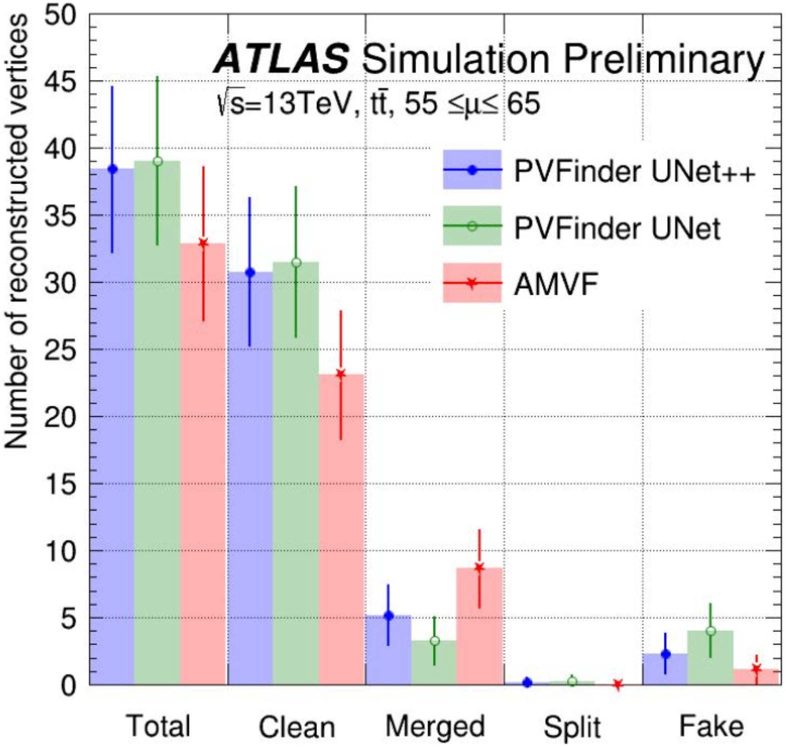


Figure 7: Reconstructed vertices with each quality classification [1]

Performance: Efficiency

- The efficiency is defined as the number of truth vertices assigned to reconstructed vertices as “clean” and “merged” divided by the total number of reconstructed truth vertices.
- The false positive rate is defined as the average number of predicted vertices not matched to any truth vertex.

Model	Efficiency	False Positive Rate (# / event)
PV-Finder UNet	94.2%	1.5
PV-Finder UNet++	88.7%	2.6
AMVF	93.9%	0.8

Table 2: Efficiency results[1]

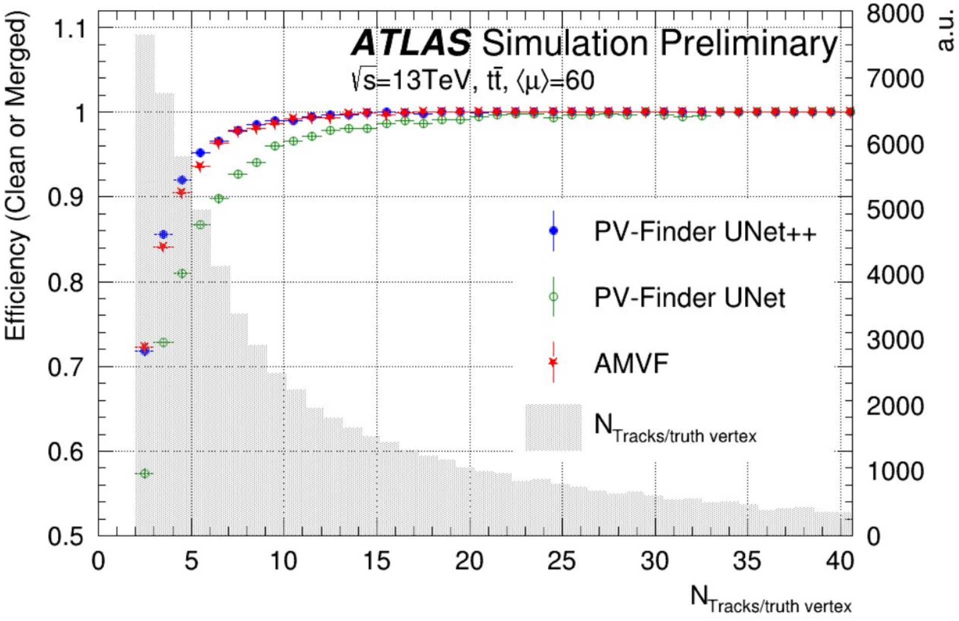


Figure 8: Vertex reconstruction efficiency [1]

Data preparation in CMS

- The dataset was constructed with a ground truth information of the PVs based on RelVal ttbar dataset
- 989 reco PVs matched with ground truth PVs
- Currently working on mix process for adding PU to the dataset using MinBias dataset

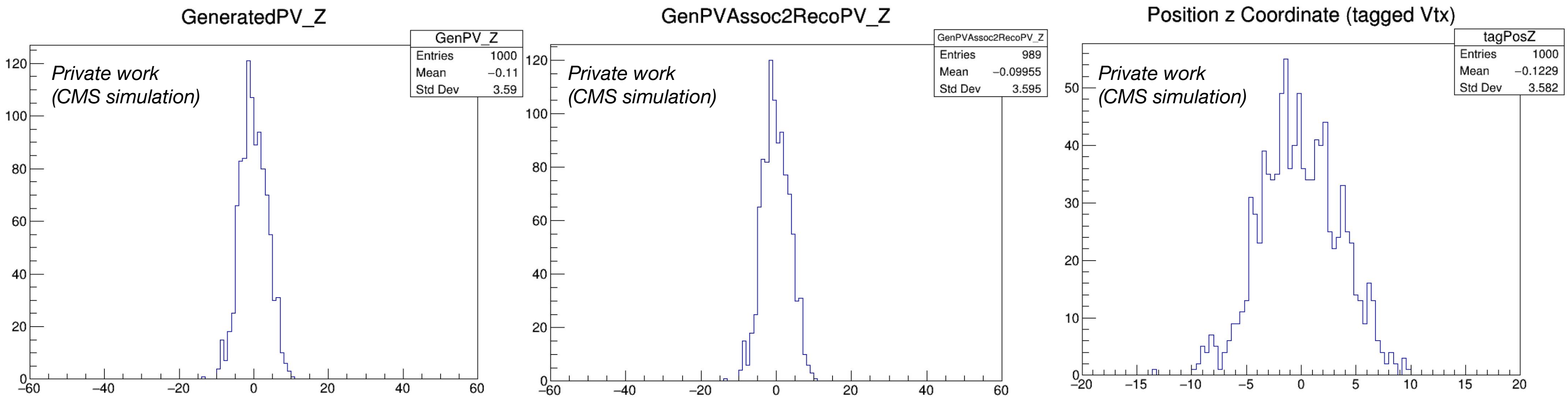


Figure 9: Gen and Reco vertices along z position (no PU)

Future work

- Add pile-up to the dataset, currently using MinBias with <μ>=200
- Construct target histograms based on the information of the Deterministic Annealing reconstruction algorithm (current in CMS)

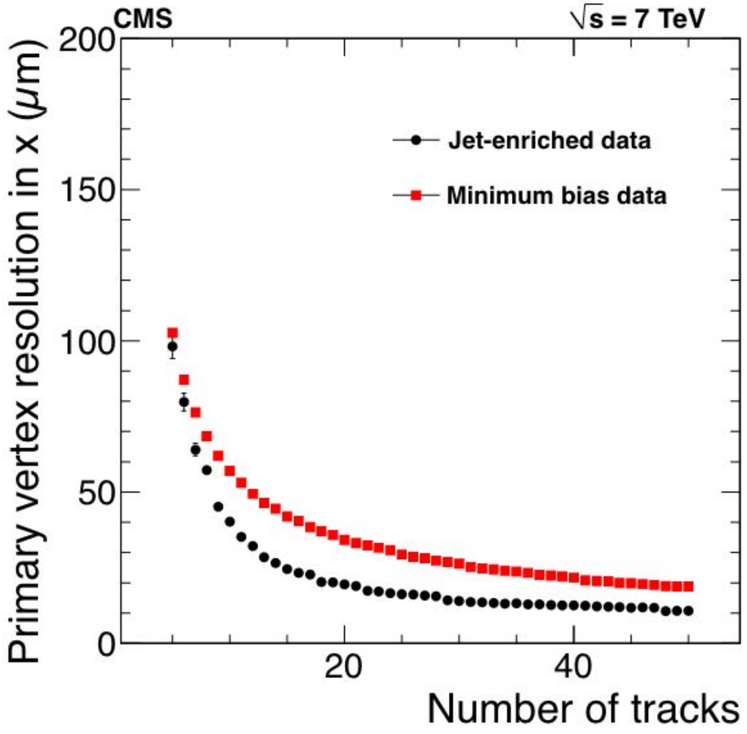


Figure 10: Primary-vertex resolution as a function of the number of tracks at the fitted vertex [4]

References

[1] ATL-PHYS-PUB-2023-011 Primary Vertex identification using deep learning in ATLAS

[2] S. Akar , M. Elashr, R.B. Garg, E. Kauffman, M. Peters, H.F. Schreiner, S. Shinde, M.D. Sokoloff, W. Tepe, L. Tompkins "Advances in developing deep neural networks for finding primary vertices in proton-proton collisions at the LHC" https://tinyurl.com/yh7m9cyt

[3] Description and performance of track and primary-vertex reconstruction with the CMS tracker, The CMS Collaboration, Journal of Instrumentation 9, 80 (2014)

[4] Offline Primary Vertex Reconstruction with Deterministic Annealing Clustering, CMS Internal Note CMS IN-2011/014

Backup slides

Additional information

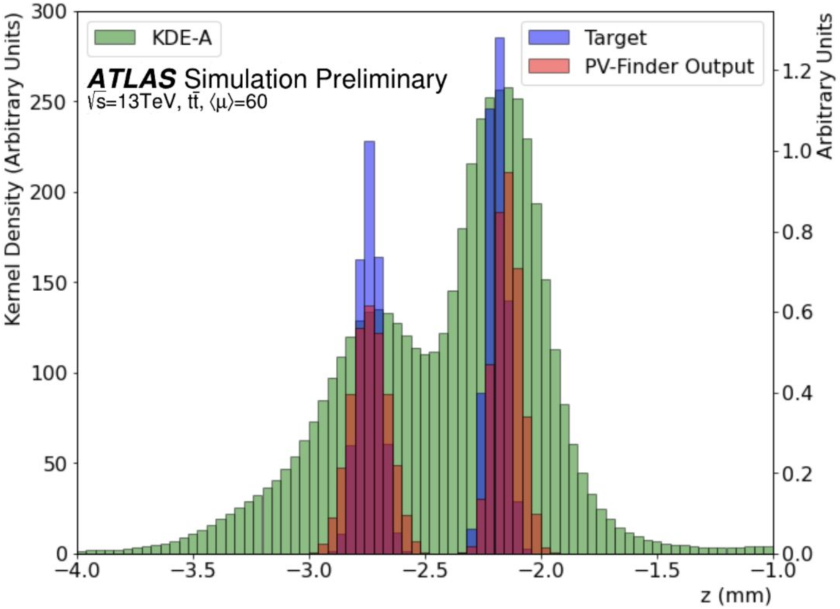


Figure 11: Example of the model output (red) compared with KDE-A and -B (green and orange respectively) with target histograms (blue) [1]

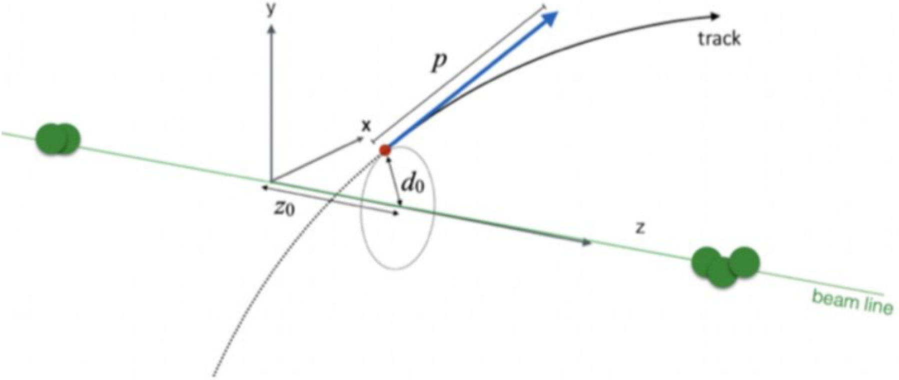


Figure 12: Coordinate system, POCA point (z_0; d_0)[2]