Advances in developing deep neural networks for finding primary vertices in proton-proton collisions at the LHC

S. Akar [1], M. Elashri [1], R.B. Garg [2], E. Kauffman [3], M. Peters [1], H.F. Schreiner [3], S. Shinde [1], M.D. Sokoloff [1], W. Tepe [1], L. Tompkins [2]

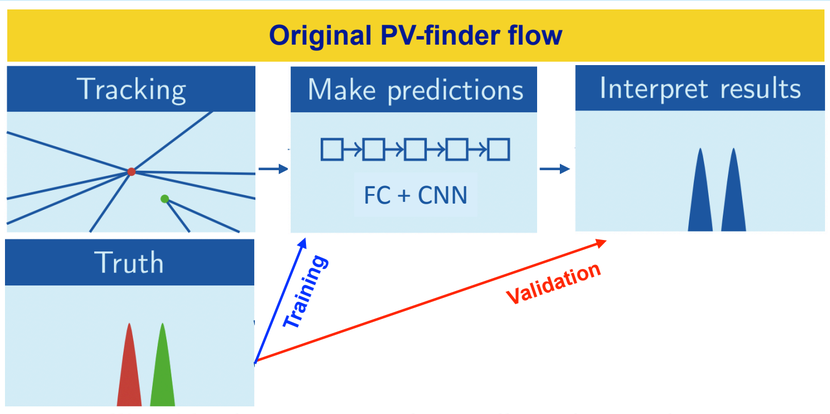
University of Cincinnati [1], Stanford University [2], Princeton University [3]

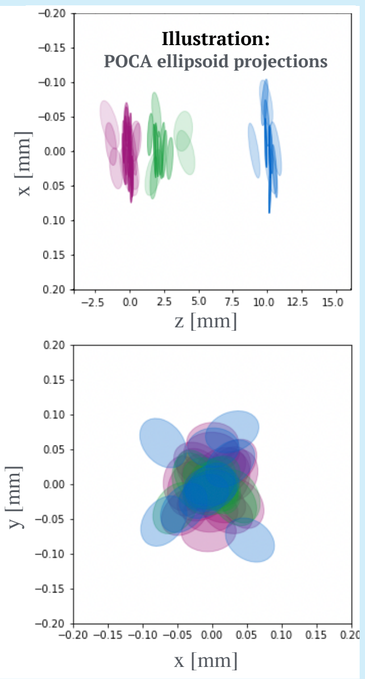
Supported by NSF awards OAC-1740102 & OAC-1739772 and via sub-awards under Cooperative Agreement OAC-1836650

Motivations and context

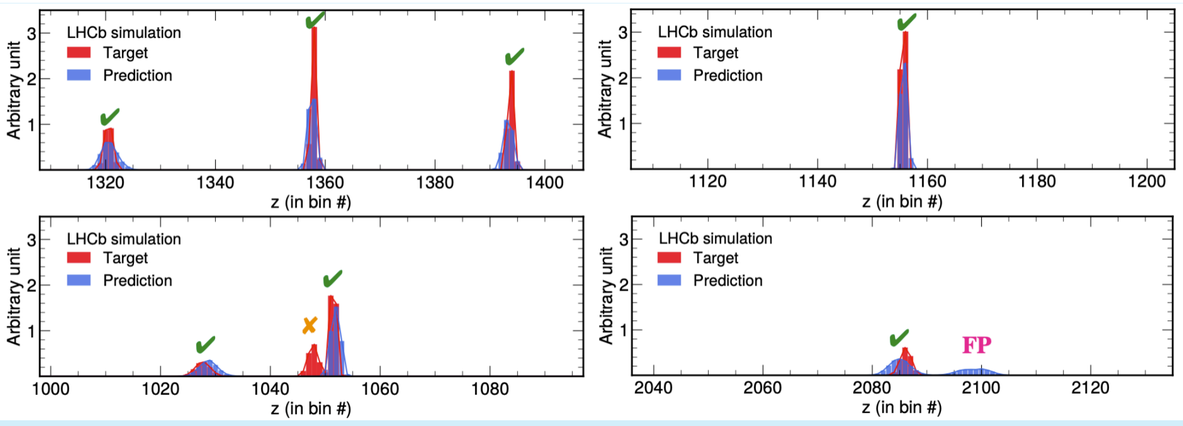
- In the next years, LHC detectors will face significantly increased luminosities
- We have developed **deep neural network (DNN)** algorithms to identify primary and secondary vertices in *pp* collisions in this high pile-up environment
- Previous models (**hybrid FC+CNN**) architecture and performances
  - **ACAT 19** J.Phys.Conf.Ser. 1525 (2020) 1, 012079 ; **CDT 20** arXiv:2007.01023 ; **CHEP 21** EPJ Web Conf. 251 (2021) 04012 ; **ACAT 22** arXiv:2304.02423 ; **CHEP 23** arXiv:2309.12417
- Here we report **new results** from a novel approach based on a **Graph Neural Network (GNN)** model

Original hybrid ML approach to finding primary vertices

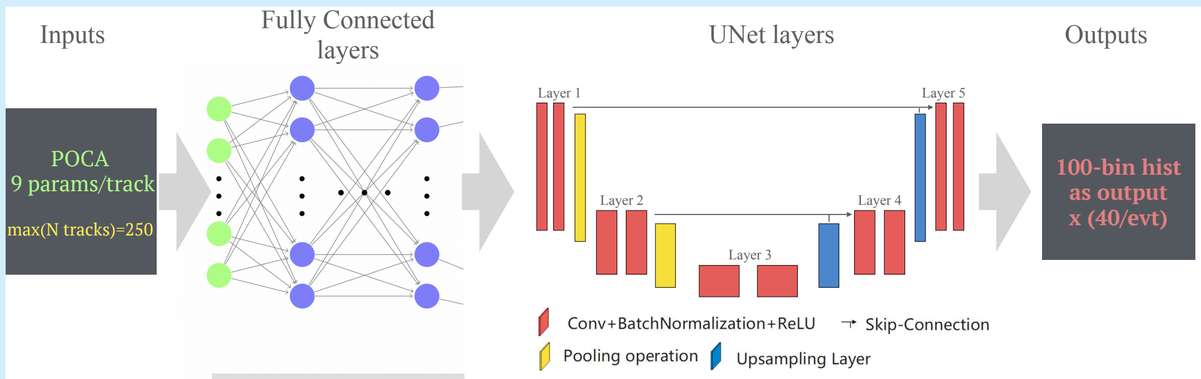




- **poca-ellipsoids:** the positions and error ellipsoids at tracks' positions of closest approach to the beamline.
- **target histograms:** proxies that are Gaussian distributions whose heights and widths reflect the expected PV resolutions
- Hybrid model is trained to predict distributions similar to the target histograms
- Heuristic algorithms extract PV positions from the predicted histograms

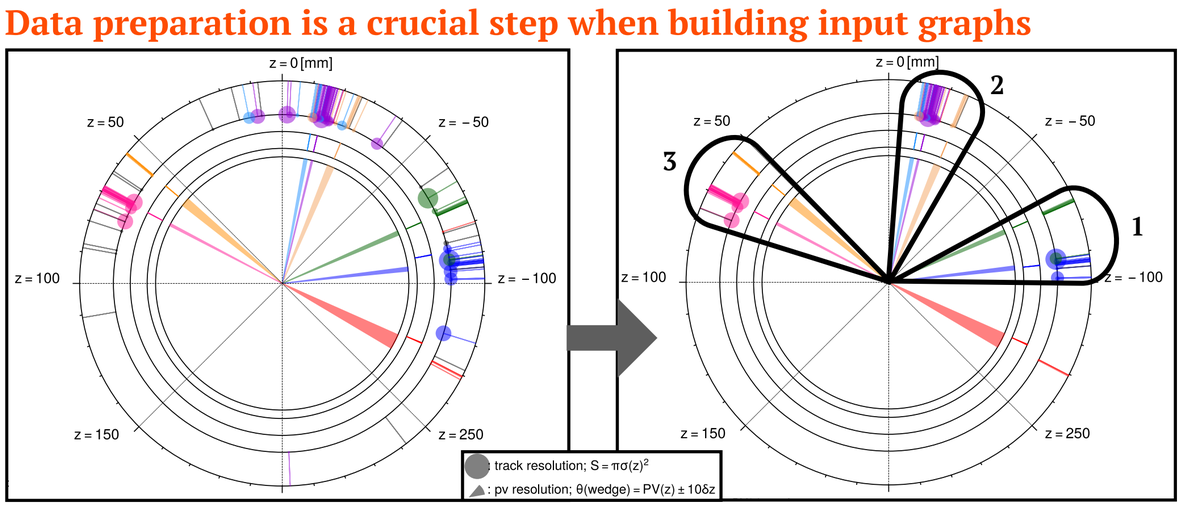


Hybrid model architecture



Graph building for GNN implementation

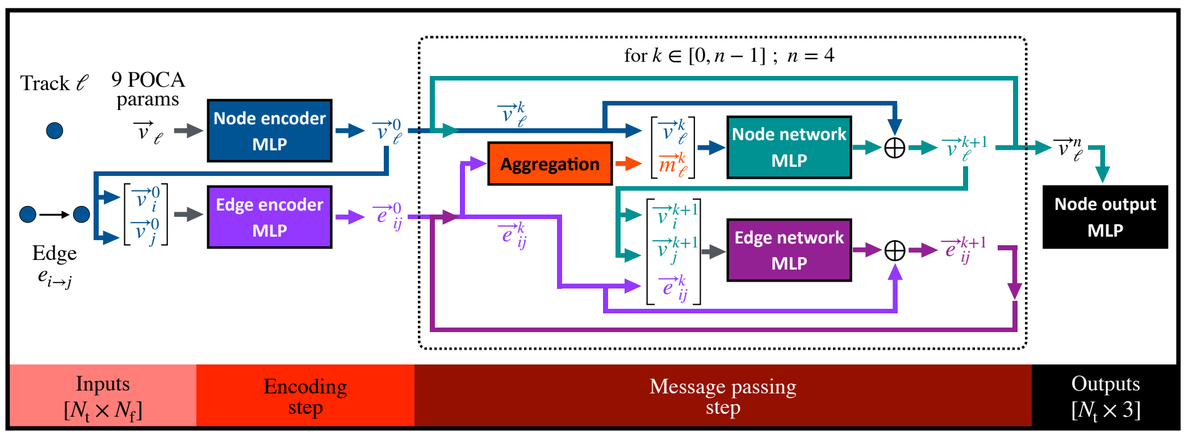
Data preparation is a crucial step when building input graphs



1. **Ordering tracks** in z values along the beamline.
2. **Filtering tracks** with simple isolation criteria ($|\Delta z_{i-1}|$ **or** $|\Delta z_{i+1}|$) < 0.5 mm
3. **Forming groups** based on gaps, $d_{ij} > 20$ mm, between consecutive tracks
4. **Build graphs** from groups with **edges connecting nodes (i.e. tracks) if** $d_{ij} < 5$ mm

GNN model architecture and loss function

Same model as used for **track finding** from hits in the Velo in LHCb **except final output layer!**

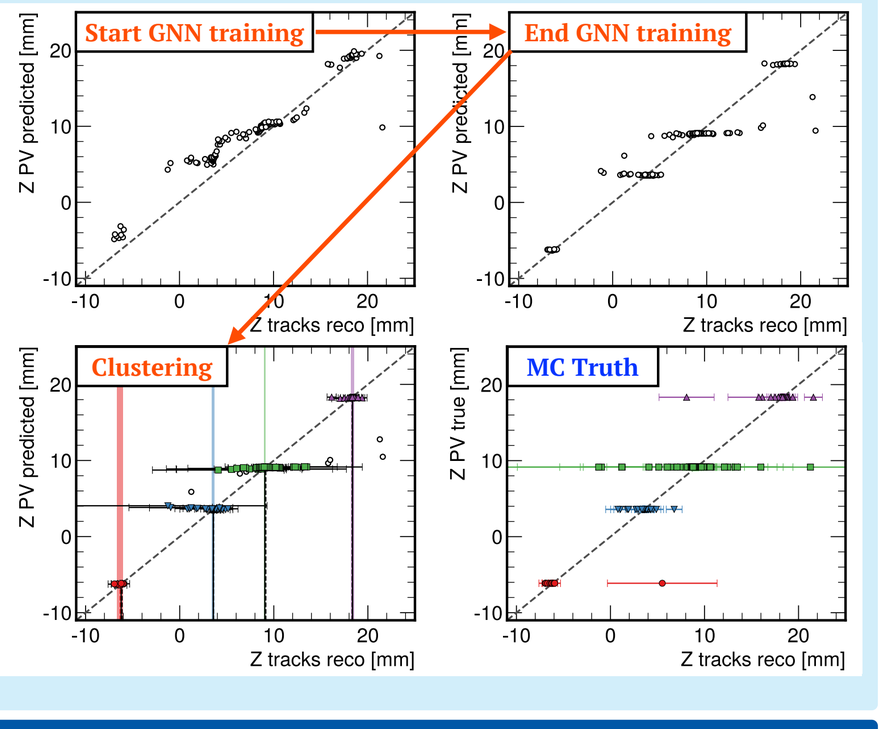


- **Same input features for GNN as for hybrid model!**
- **Target** defined as true PV coordinates (x_{pv}, y_{pv}, z_{pv}) from which a track originates
- **Custom loss function** defined as the weighted sum of distances between each track predicted PV coordinates and true PV coordinates:

$$L_D = \left(\sum_i w_i \sqrt{(\hat{x}_i - x_{pv,i})^2 + (\hat{y}_i - y_{pv,i})^2 + (\hat{z}_i - z_{pv,i})^2}\right) / \sum_i w_i$$
$$w_i = 1/(\sigma_i^x \sigma_i^y \sigma_i^z)$$

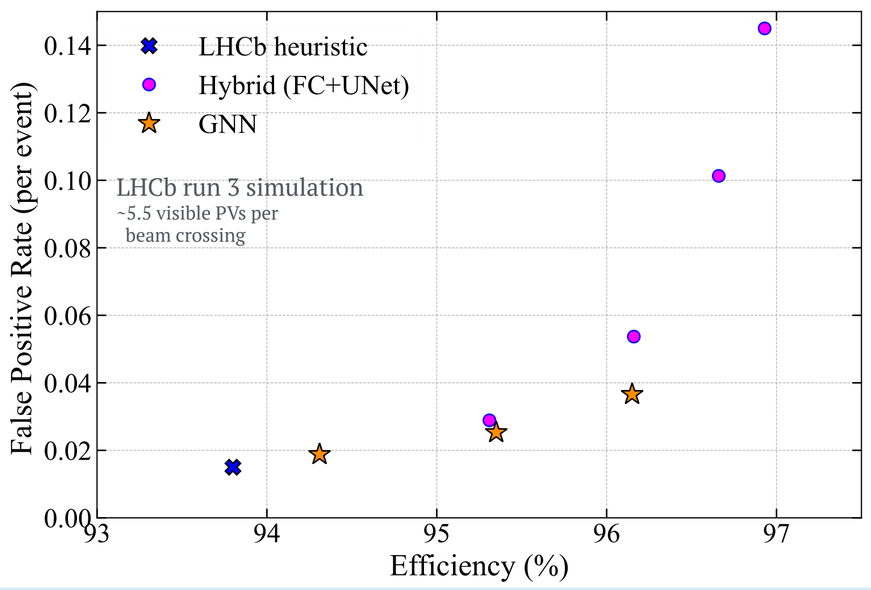
Construction of PV candidates from GNN output

- **PV candidates are formed from the GNN output using a heuristic algorithm to cluster tracks**
- **Location of candidate PV are inferred from weighted mean of the positions of tracks forming each candidate**

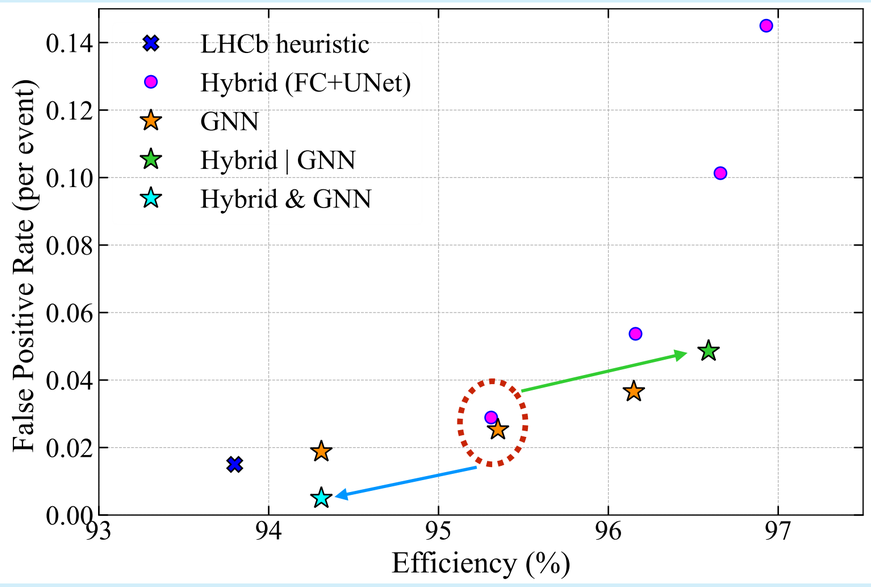


Physics performances: GNN vs Hybrid

- ☑ **Hybrid best model results** from developments over the past years
- ☐ No direct tracks to PV association
- ☑ **GNN model achieve slightly better physics performance**
- ☑ Tracks to PV association by construction



- **Compare outputs of GNN and Hybrid models at point with similar physics performances**
- **Combination of output from models allows to either significantly increase efficiency or decrease false positive rate**



Summary:

- **GNN models appear quite versatile** where similar models achieve good performances for different tasks (tracking vs PV finding)
- **GNN and hybrid models achieve similar intrinsic physics performances...**
- **...but only partial overlap meaning both models did not learn exactly the same relations from identical input data!**