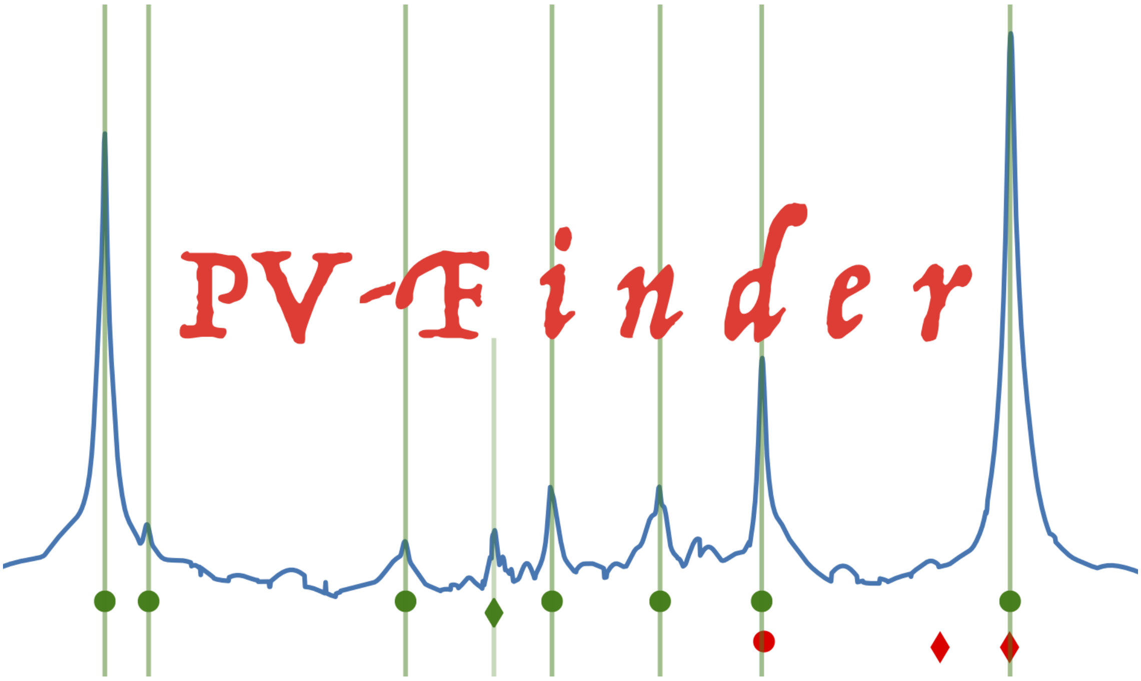
IRIS-HEP Proposal:
Adapting PV-Finder to the CMS and ATLAS Experiments

Elliott Kauffman
elliott.kauffman@duke.edu

Mentors: Henry Schreiner, Mike Sokoloff

April 7, 2022



1 Introduction

PV-Finder is a hybrid deep learning algorithm which identifies primary vertices (proton-proton collision points in LHC detectors). This algorithm was developed for use in conjunction with the LHCb detector in Run 3 of the LHC, which will experience a luminosity that is 5.5 times that of Run 2. This change motivates a more efficient method to locate primary vertices in LHCb data. Reconstructed particle tracks are used to calculate one-dimensional Kernel Density Estimators (KDE), representing the density of tracks along the direction of the beamline. The peaks of this KDE are correlated with the z-positions of primary vertices. One method of generating the KDE is to calculate the maximum track density for the transverse plane of each bin along the z-direction. This KDE is referred to as ”KDE-A” and is used alongside ”KDE-B”, which is the square of the original bin values. Some models have included the x and y locations of the maximum as perturbative features. Another method of finding the KDE is to compute a 3-dimensional error ellipsoid representing each track's contribution for each bin along the z-direction. The latter method is different in that we can encode x and y minimization without the computational effort of scanning the $x-y$ planes to locate the maximum density point. This KDE is then used as an input to a convolutional neural network (CNN) to predict primary vertex locations. To train the network, Monte Carlo data is used so that the truth information of the primary vertices is available. The CNN predicts Gaussian peaks around primary vertex locations. In LHCb data, the efficiency of the CNN has inreased from to 90% to past 98% over the course of the past few years (Fang *et al* (2020), Akar *et al* (2021)). The success of this project motivates its extension to both the ATLAS and CMS experiments.

The ATLAS and CMS detectors provide different challenges with respect to finding primary vertices. These detectors have about 50 times more vertices per event, so they have higher track multiplicity and therefore more peaks in the KDE. This will make it more difficult to distinguish individual peaks. On the other hand, The ATLAS and CMS detectors have finer resolution in the z-direction, which should help counteract this problem as the peaks in the KDE will be more precise. During a previous DIANA-HEP fellowship, the generation of the KDE for both the ATLAS and CMS experiments has been accomplished. This upcoming stage of the project will be focused on using these KDEs to train a neural network.

2 Related work

Multiple machine learning models have been developed thus far in conjunction with LHCb data, which will provide the basis for the extension of the project to the ATLAS and CMS experiments. The first model is referred to as the All-CNN. This CNN takes the 4000-bin LHCb KDE as an input, which then goes through 8 convolutional layers with a ”Leaky ReLu” activation function in between, until the output, which uses the ”softplus” function to convert into probabilities. The output is a 4000-bin histogram with peaks corresponding to predicted primary vertices. If these peaks overlap with the truth vertices, the primary vertex is said to be located (Fang *et al* (2020), Akar *et al* (2021)).

Some progress has also been made using the U-Net architecture, which is similar in structure to the All-CNN except that it includes ”skip-connections”, in which some information bypasses certain layers and is added back in at a later point. This allows some finer details from the input to survive the down-sampling operations. Initial investigations into this architecture have achieved around 95% efficiency (Akar *et al* (2021)). This may prove a useful model for ATLAS and CMS, considering that there are often more primary vertices located closer together than in LHCb data due to the higher track multiplicity.

Some progress has also been made incorporating POCA (Point of Closest Approach) ellipsoids into the PV-Finder process. POCA ellipsoids are objects that exist for each track, and are centered around their point of closest approach in 3-dimensional space to the beamline, which runs in the z-direction. Their size is calculated using the error information from the covariance matrix associated with each track, so that smaller ellipsoids are better-known and thus more important than larger ellipsoids. These objects have multiple potential applications in PV-Finder. They can be used to speed up the KDE generation process, which is currently the slowest part of the PV-Finder process. In addition, we may be able to connect the network for KDE generation to the PV-predicting network in order to achieve better performance. The sum of the projections of these ellipsoids onto the z-axis is qualitatively similar to KDEs generated using the track density approach. In addition, their 3-dimensional nature may grant us easy access to information about the x and y components of primary vertices, without need for extra inputs into the neural network (Akar *et al* (2021)).

3 Research questions

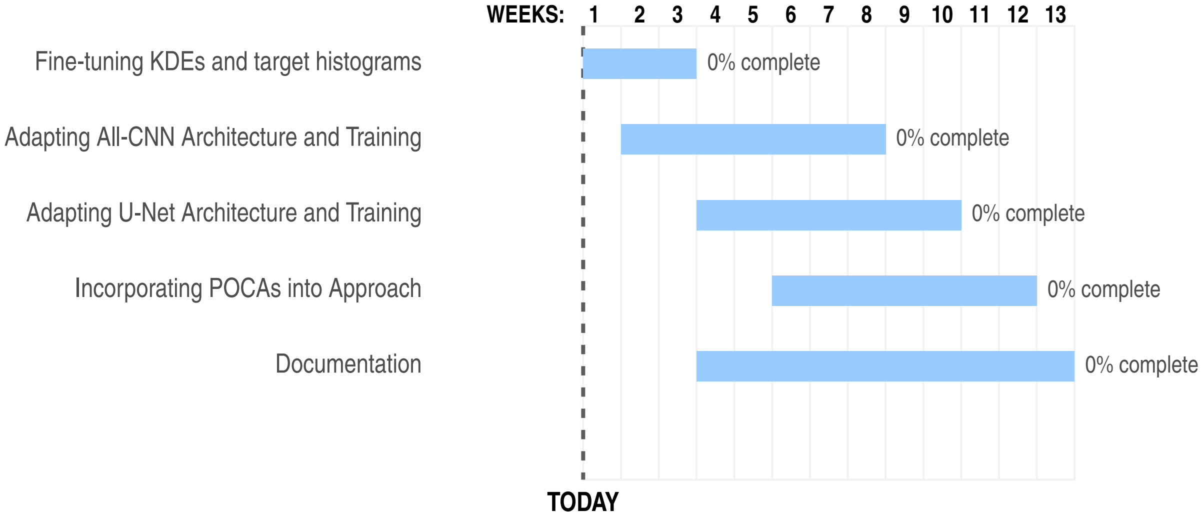
The focus of this project is to adapt the PV-Finder algorithm to the CMS and ATLAS experiments. The success of PV-Finder in LHCb data motivates the question of whether a similar process could be used in the CMS and ATLAS experiments. Initial results appear to be positive, as the KDE generation process has been developed successfully. The existing All-CNN and U-Net models will need to be adjusted to account for the different nature of ATLAS and CMS data. The performance of both models on ATLAS and CMS data will be investigated.

In addition, POCA ellipsoids have the potential to expedite the KDE generation method, so the performance of these KDEs will be compared with the performance of KDEs generated using the maximum track density method. After training, POCA ellipsoids may also be useful in a different way. Currently, PV-Finder only gives us information about primary vertices in the z-direction. Once we have this information though, we may be able to use the POCA ellipsoids from nearby tracks to predict the x and y information of the primary vertices.

4 Methodology

Similarly to the LHCb version of PV-Finder, Monte Carlo simulated data of the ATLAS and CMS experiments will be used to train neural networks designed to locate primary vertices. In order to generate KDEs, track information from the Monte Carlo data is obtained and processed. After the generation of the KDEs, the files are converted into HDF5 files for processing into the PyTorch framework. During the same step, validation histograms for the primary vertices are generated using the truth track information. PyTorch tools are then used to design and train the model.

5 Time planning



References

S. Akar *et al*. Progress in developing a hybrid deep learning algorithm for identifying and locating primary vertices. *EPJ Web Conf.*, 251:04012, 2021.

R. Fang *et al*. A hybrid deep learning approach to vertexing. *J. Phys. Conf. Ser.*, 1525:012079, 2020.