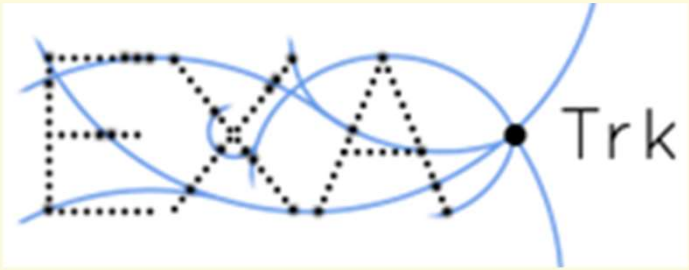
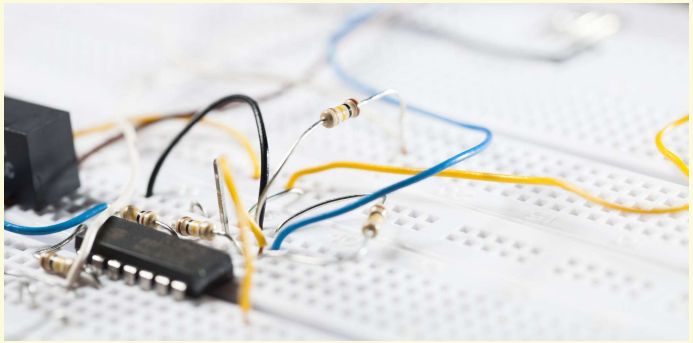
Improving the Inference of the Graph Neural Networks for Track Reconstruction



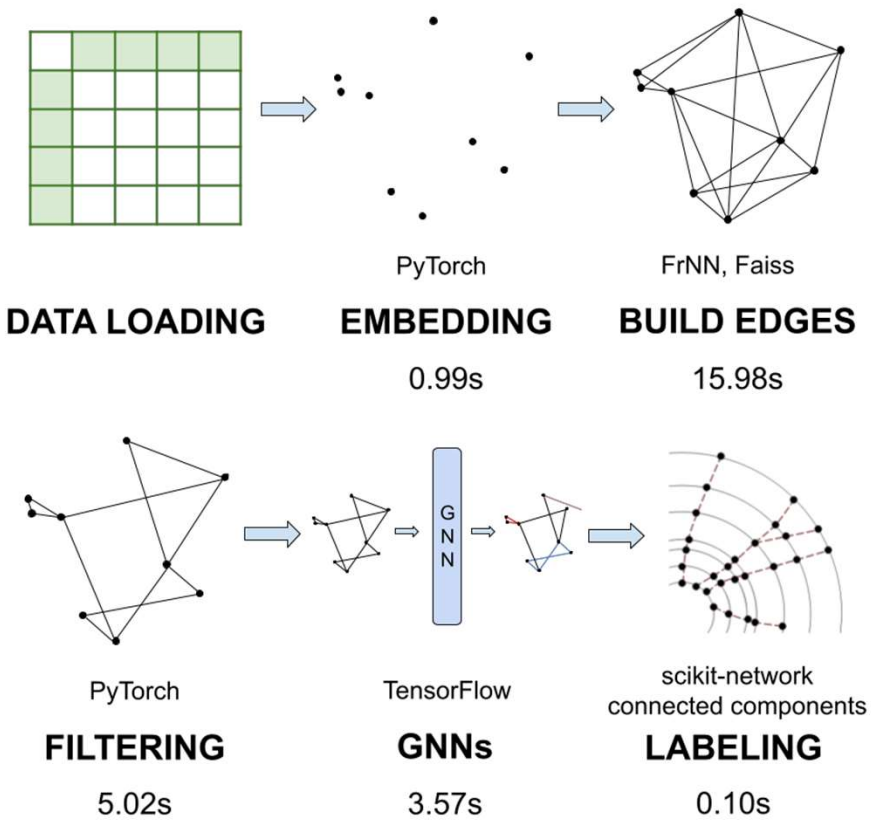
James S Gaboriault-Whitcomb, Henry H Paschke and Alina Lazar

on behalf of the **Exa.TrkX** collaboration

Youngstown State University,



The Exa.TrkX GNN Inference Pipeline



	GPU (ms)	CPU (s)
Data Loading	2.2	
Metric Learning	6.7	0.99
Graph building	40 ± 10	15.98
Filtering	370 ± 80	5.02
GNN	170 ± 30	3.57
Track Building (CC)	90 ± 8	0.1
Total	700 ± 100	25.66

MPI was used to run events in parallel, using multiple cores.

The most time-consuming steps of the pipeline are Build Edges and Filtering. To speed-up Build Edges we used Faiss with 2 threads and multiprocessing for the Filtering for-loop.

The results indicate that it is best to use between 10 and 15 cores per event, however running it on the GPU is still 27 times faster.