Real Time Analysis of Unstructured Data with Machine Learning on Heterogeneous Architectures

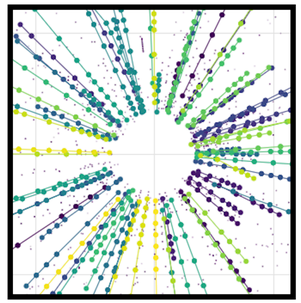
arXiv.2406.12869
arXiv.2407.12119
arXiv.2410.xxxxx

SMARTHEP Yearly Meeting 2024
Milano-Bicocca University, Oct 01, 2024

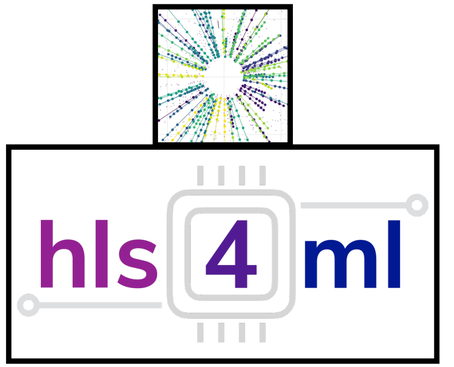
Fotis I. Giasemis

Overview

- Track Finding with ETX4VELO on GPUs



- FPGA/GPU ML Inference Throughput Comparison



- Traffic Anomaly Detection

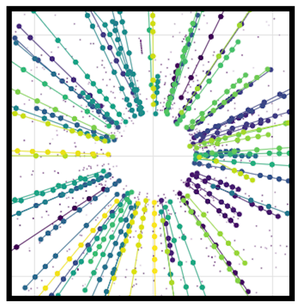
arXiv.2406.12869

ETX4VELO

Tracking in the VELO with ETX4VELO

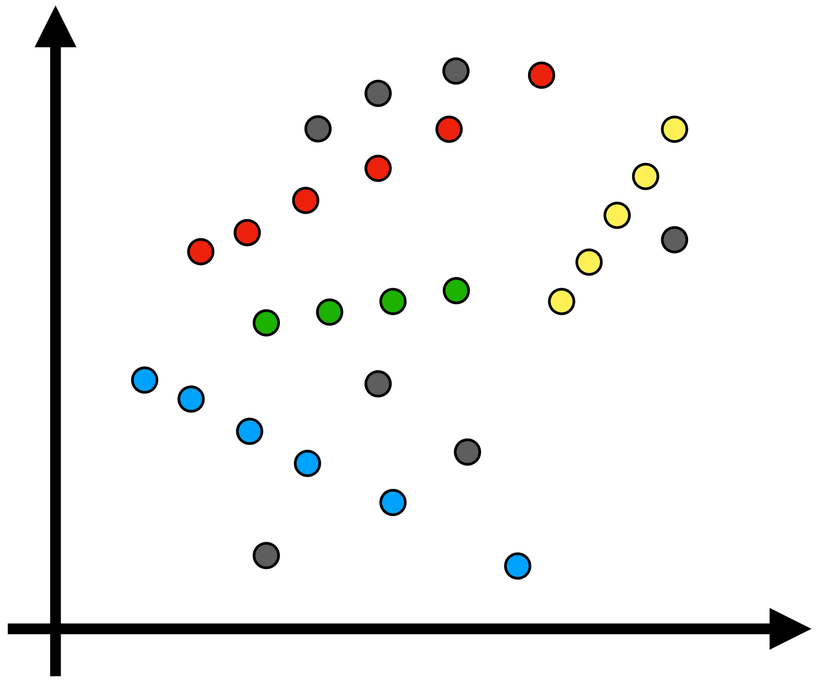
- Graph Neural Network-based pipeline for **track finding** in the VELO
- ETX4VELO

- Comparable or superior physics performance to Allen
- Excellent **electron reconstruction** achieved using **triplets**
- Significantly reduced **fake rate**



ETX4VELO

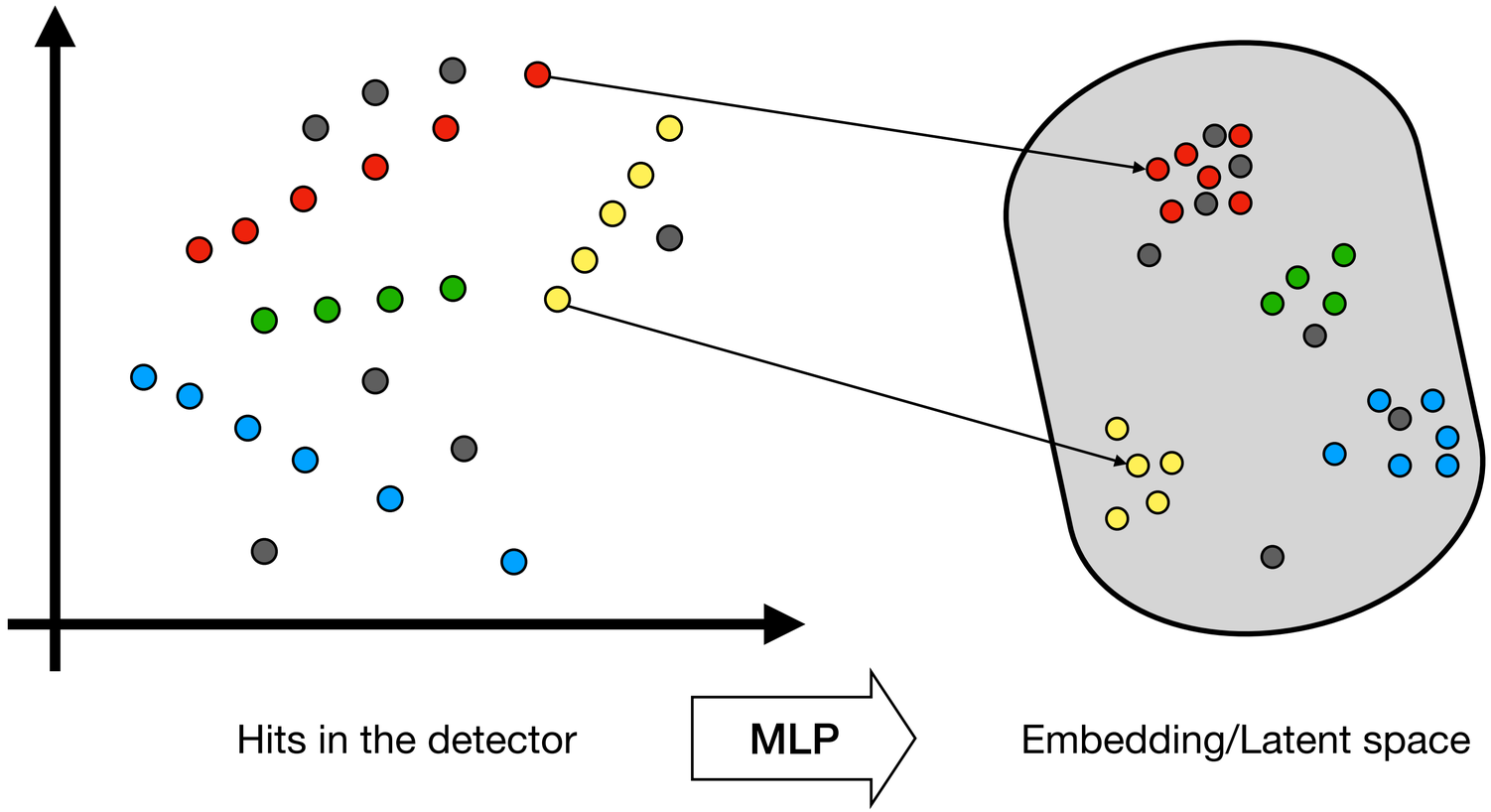
How do we get a graph from the hits?



Hits in the detector

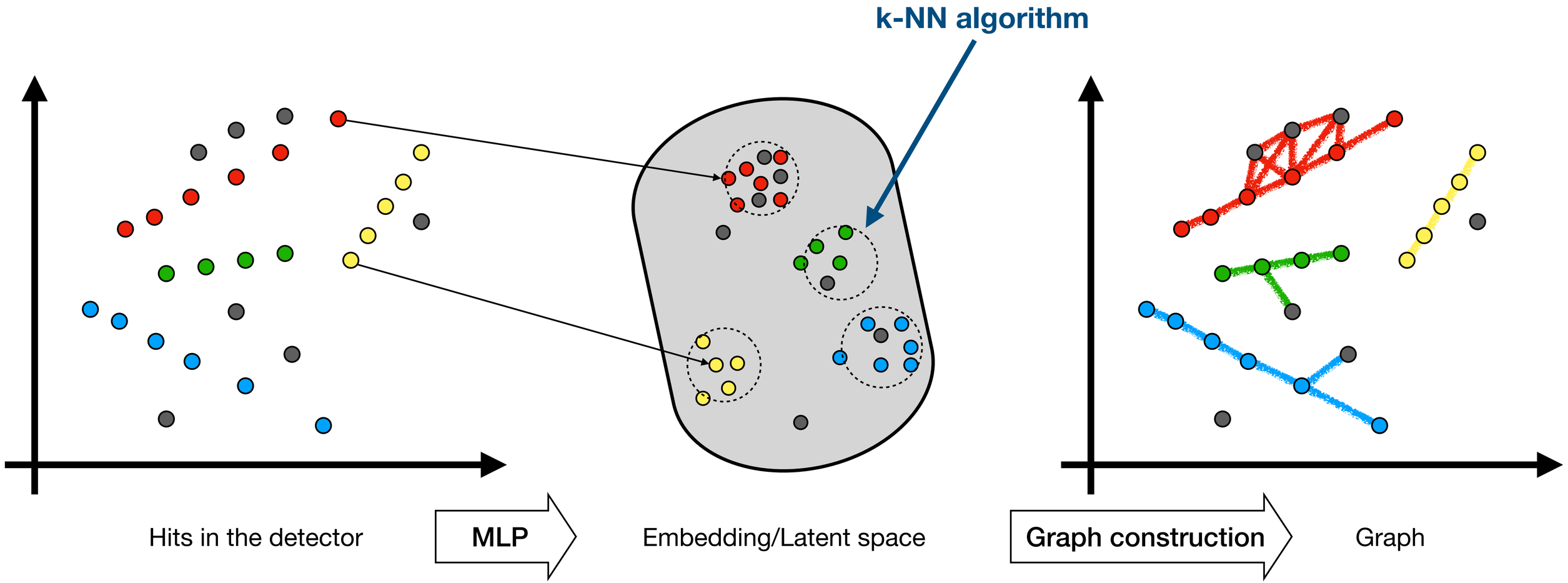
ETX4VELO

How do we get a graph from the hits?



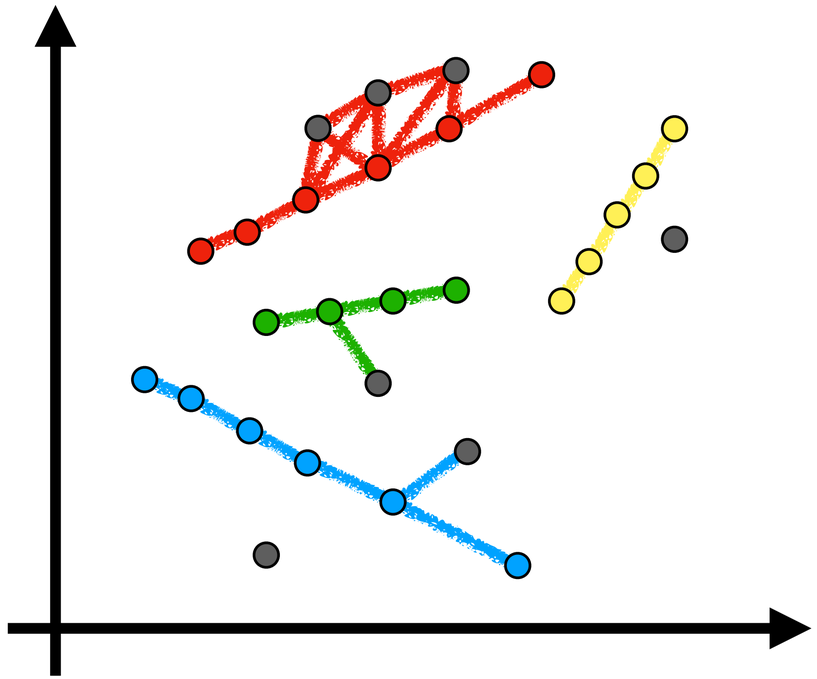
ETX4VELO

How do we get a graph from the hits?



ETX4VELO

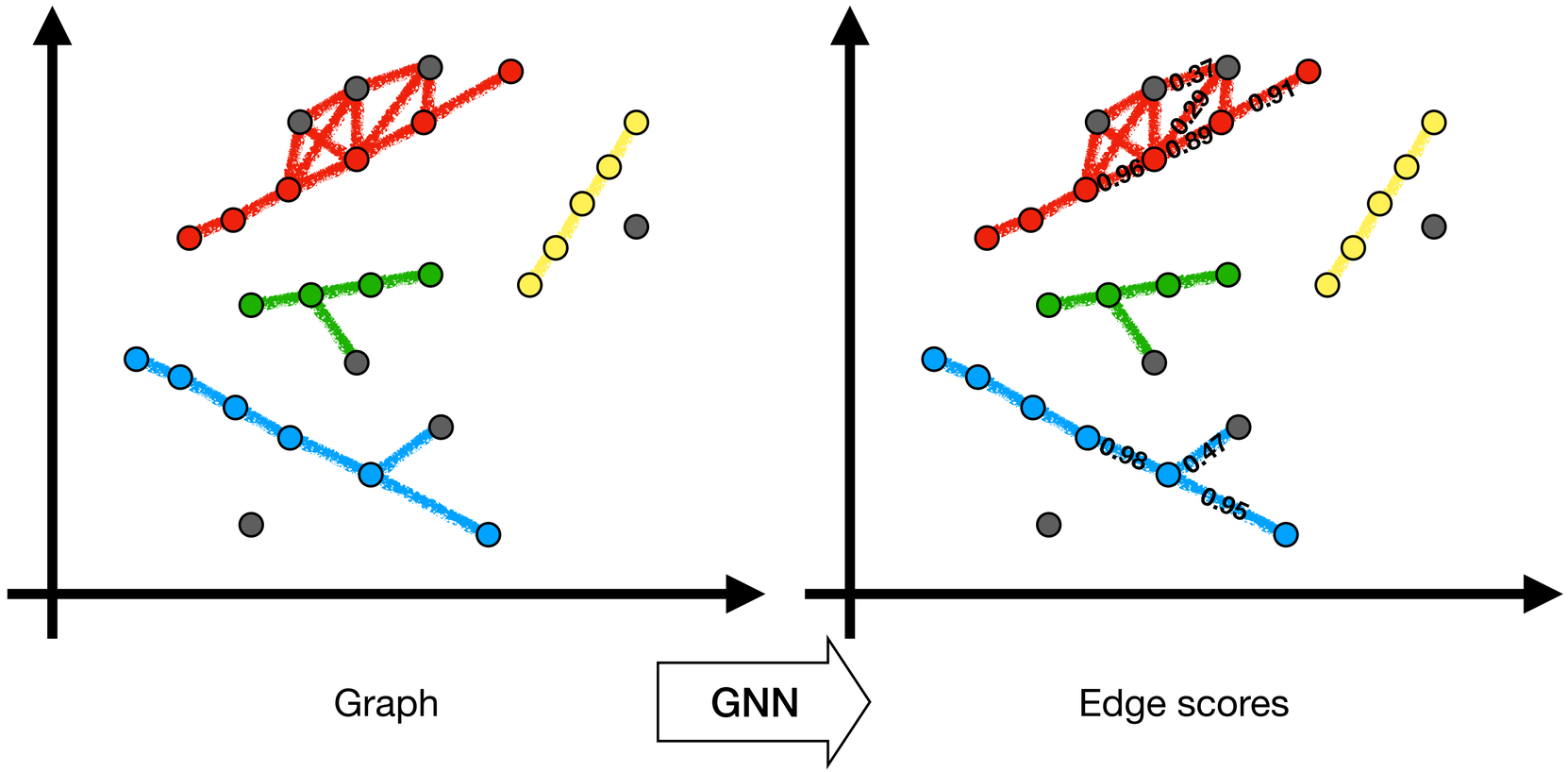
How do we get tracks?



Graph

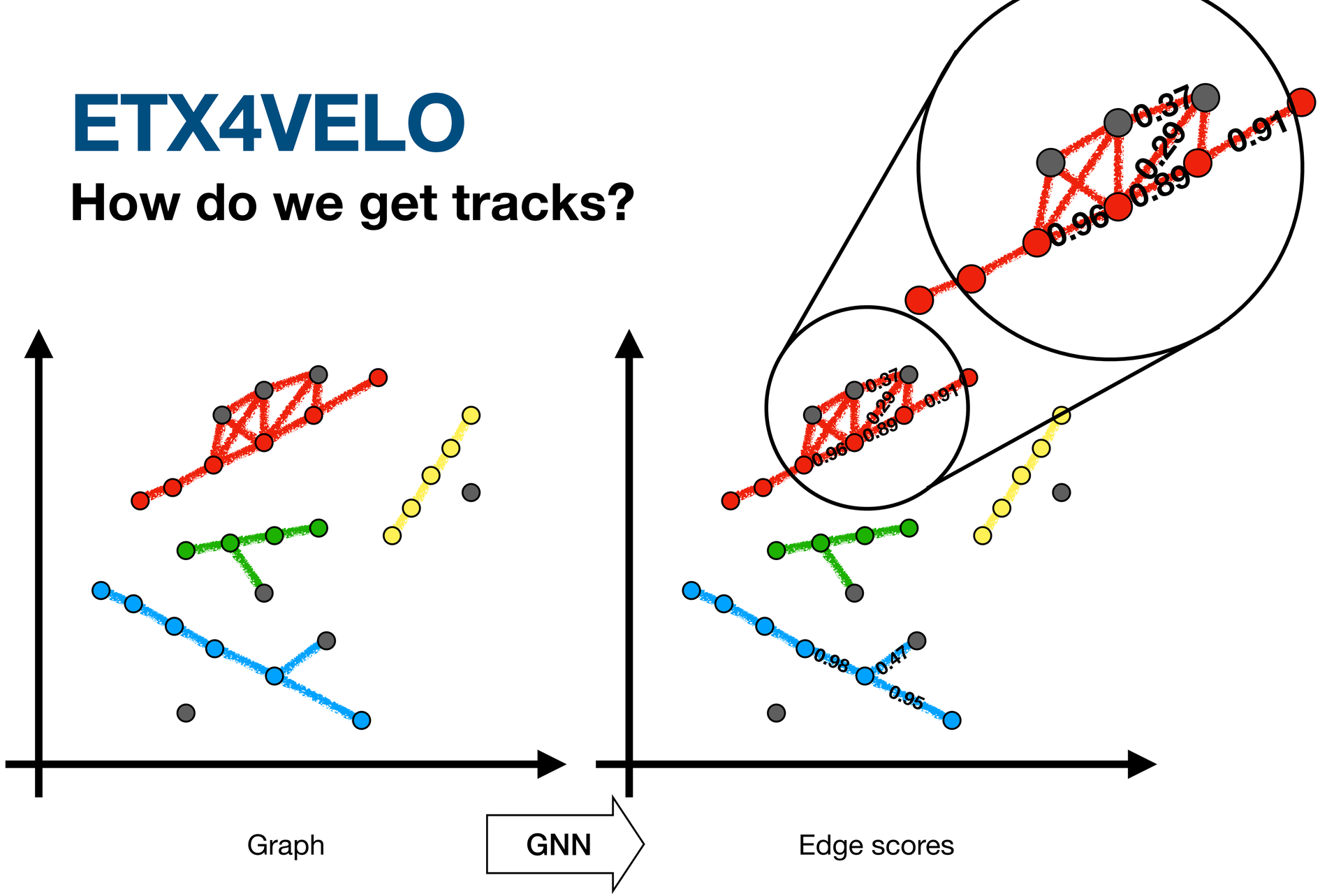
ETX4VELO

How do we get tracks?



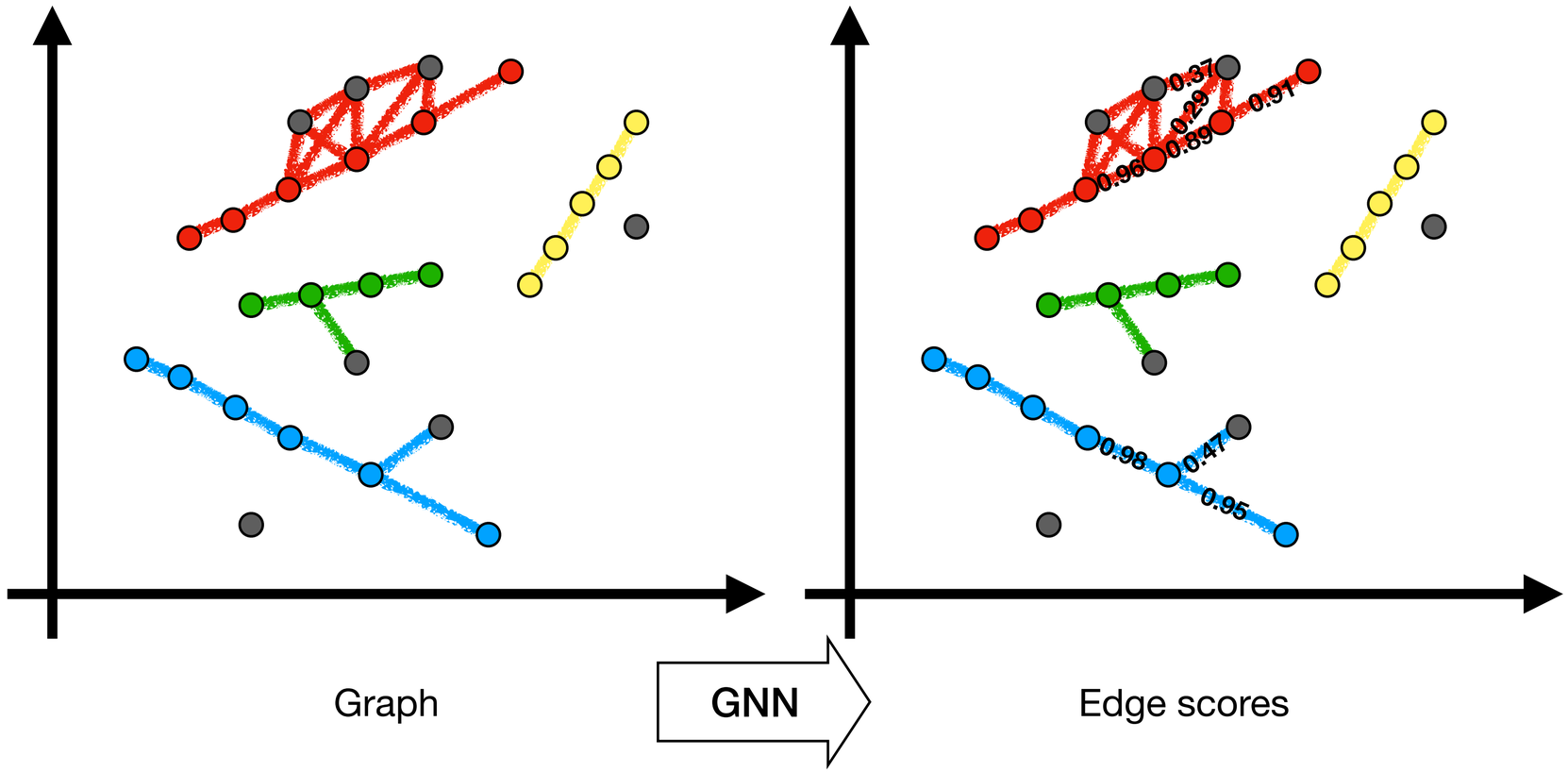
ETX4VELO

How do we get tracks?



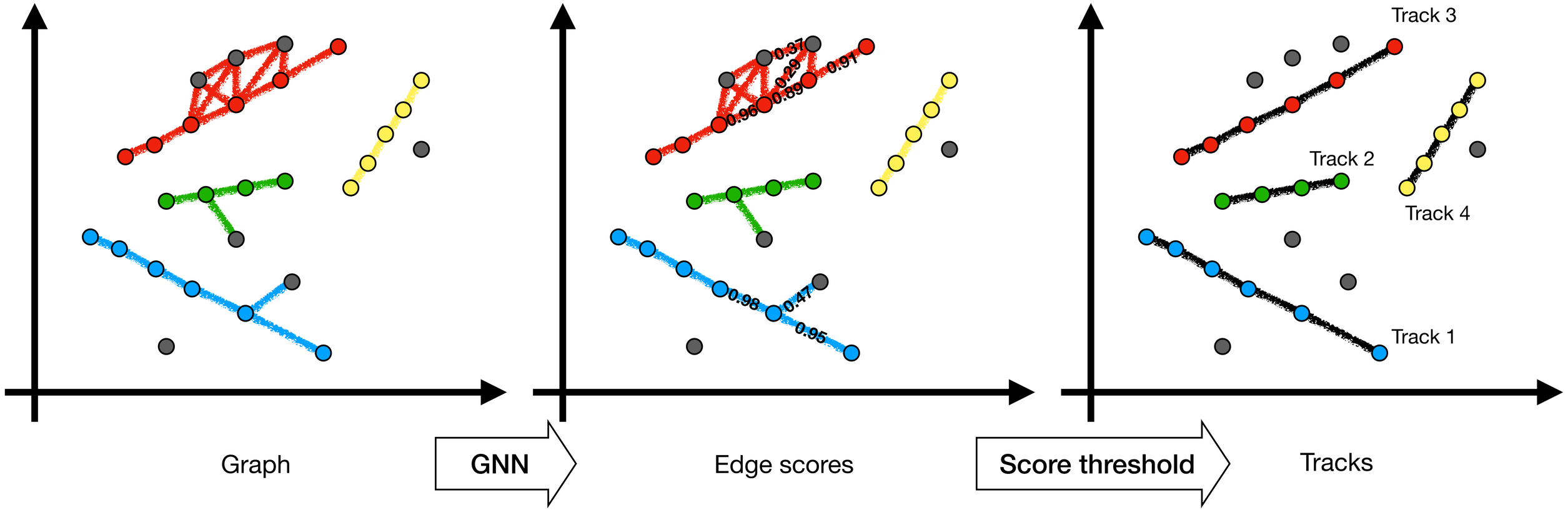
ETX4VELO

How do we get tracks?



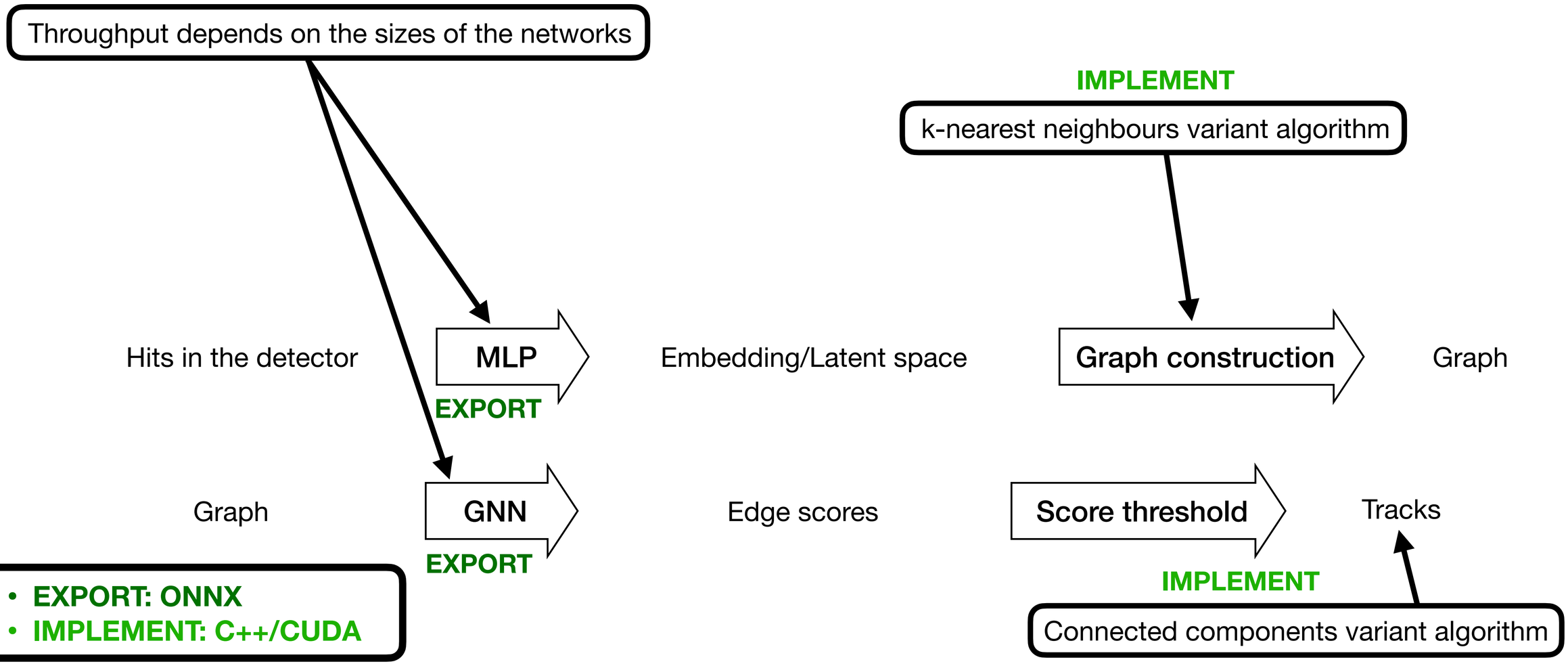
ETX4VELO

How do we get tracks?



ETX4VELO GPU Version

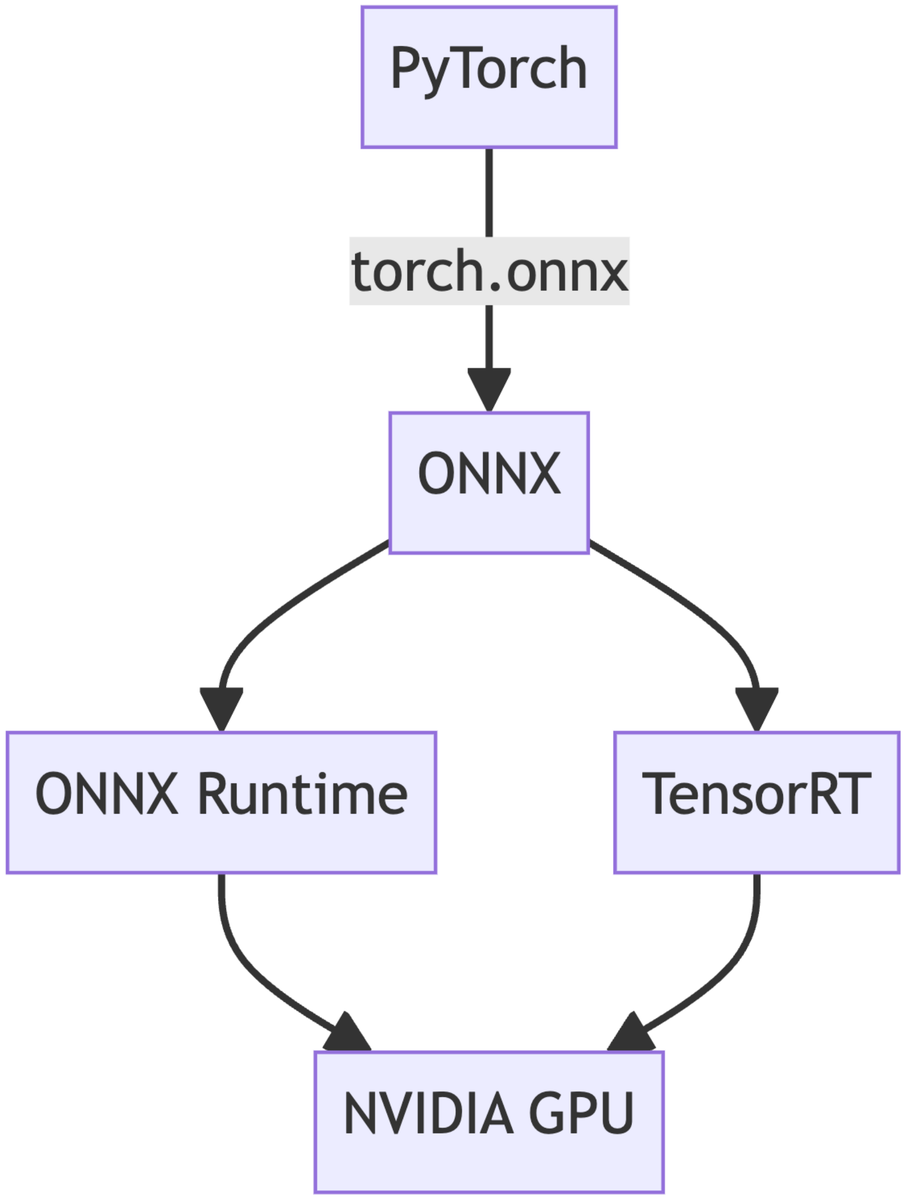
Inference steps



ML Inference on GPU (Allen)

- Throughput: infer events in **batches**
- Maximum number allowed by GPU memory

- **ONNX Runtime** + CUDA Execution Provider
- **TensorRT**

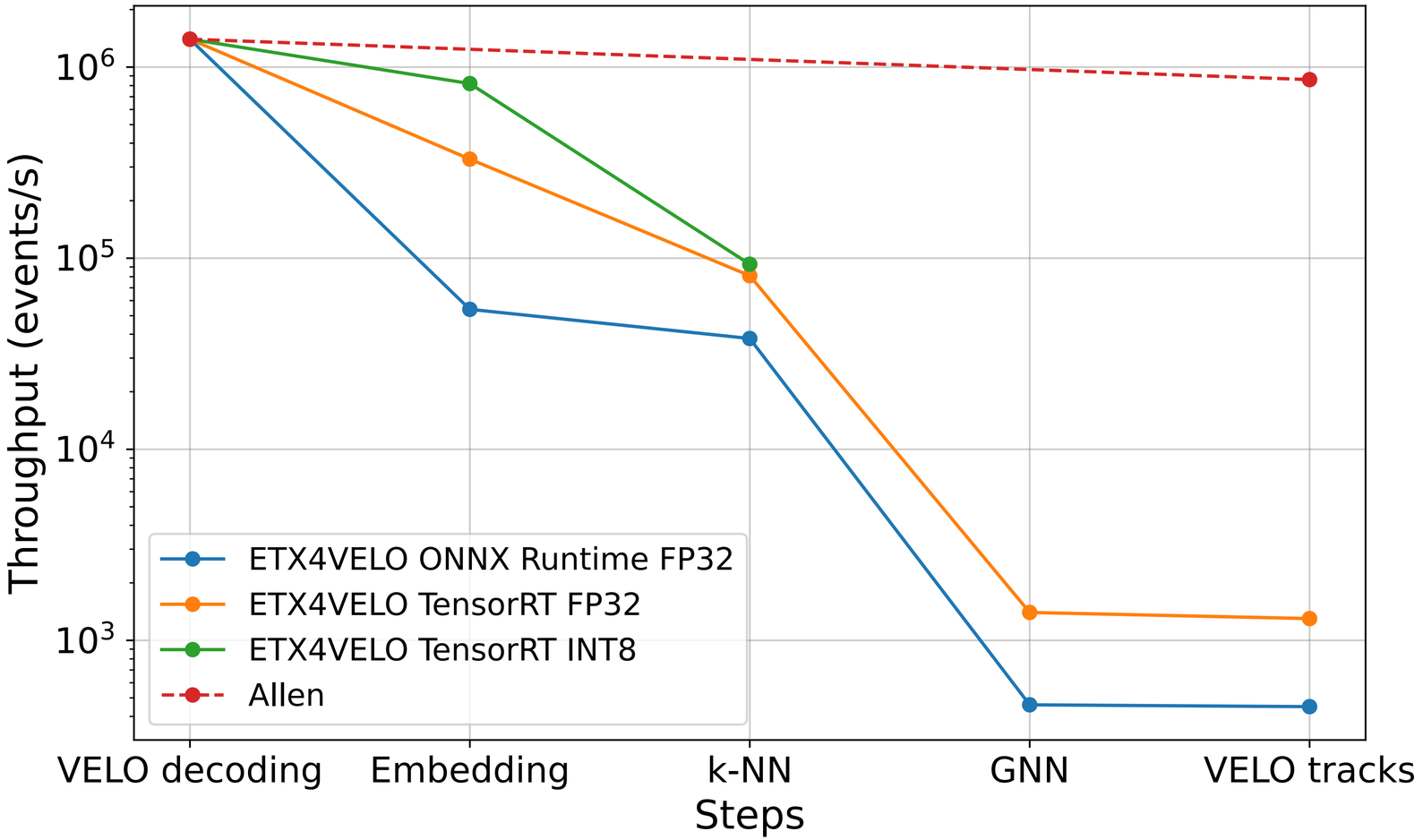


arXiv.2407.12119

ETX4VELO GPU Version

Computational performance

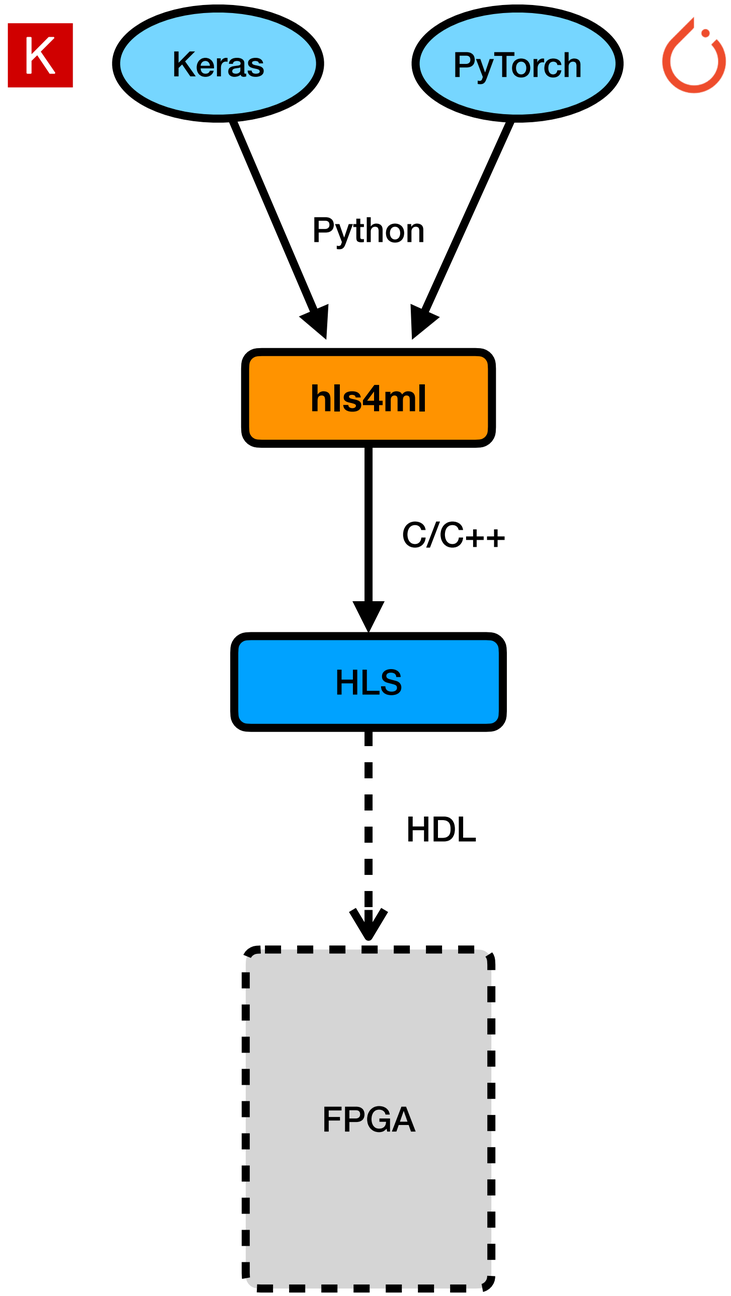
NVIDIA GeForce RTX 3090



CERN Secondment

hls4ml

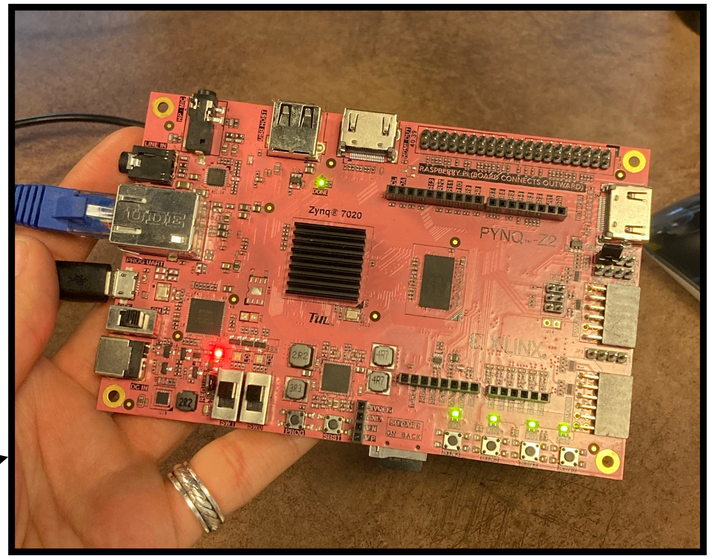
- hls4ml
- **Machine learning** on FPGAs
- Using High-Level Synthesis (HLS)



CERN Secondment

hls4ml-etx4velo

- Started January 2024
- Currently finishing up paper
- Repository
- **ETX4VELO** MLP model on **FPGA board**
- **Comparative studies** GPU/FPGA
- FPGA **outperforms** GPU with **60% of the power** consumption



CERN Secondment

Results

TABLE VII

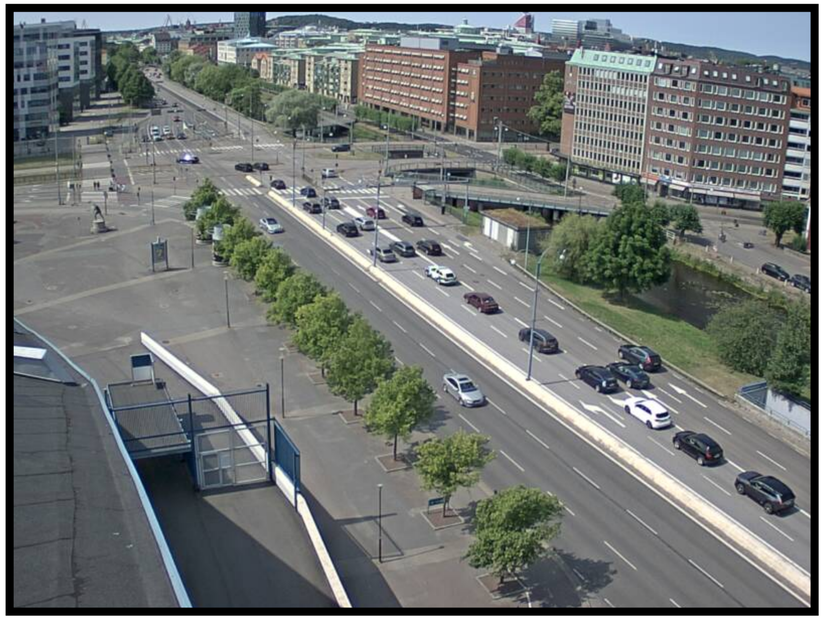
THROUGHPUT COMPARISON OF THE EMBEDDING MLP BETWEEN THE FPGA IMPLEMENTATION THEORETICAL THROUGHPUT AND THE GPU IMPLEMENTATION. FOR THE ALVEO U250 IMPLEMENTATIONS, <A,B> REFERS TO THE PRECISION BEING AP_FIXED<A,B>. THE POWER USAGE, WHILE RUNNING THE INFERENCE AND WHILE IDLE, IS ALSO COMPARED.

Accelerator	Alveo U250		GeForce RTX 2080Ti			GeForce RTX 3090		
Implementation	<8,3>	<16,6>	ORT FP32	TRT FP32	TRT INT8	ORT FP32	TRT FP32	TRT INT8
Throughput (events/s $\times 10^3$)	1 100	470	46	260	540	54	330	820
Active Power Consumption (W)	230			250			350	
Idle Power Consumption (W)	24			40			50	

Ximantis Secondment

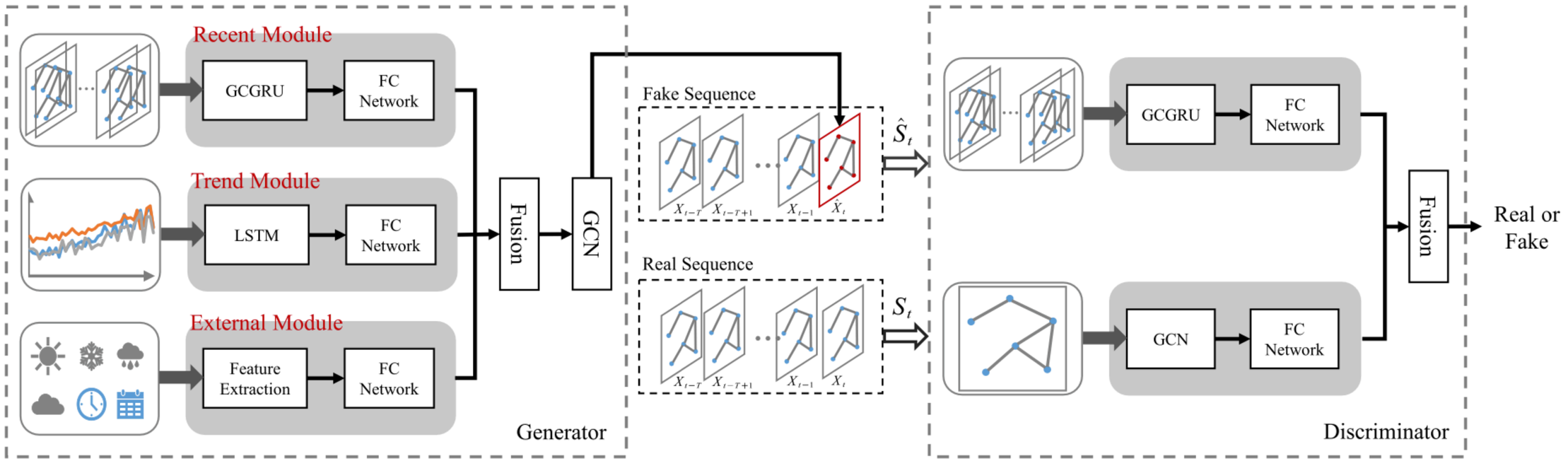
Traffic anomaly detection

- Started March 2024
- Currently finishing up paper
- Repository
- Anomaly detection on traffic data
- **125 cameras in Gothenburg**, Sweden
- Results with STGAN



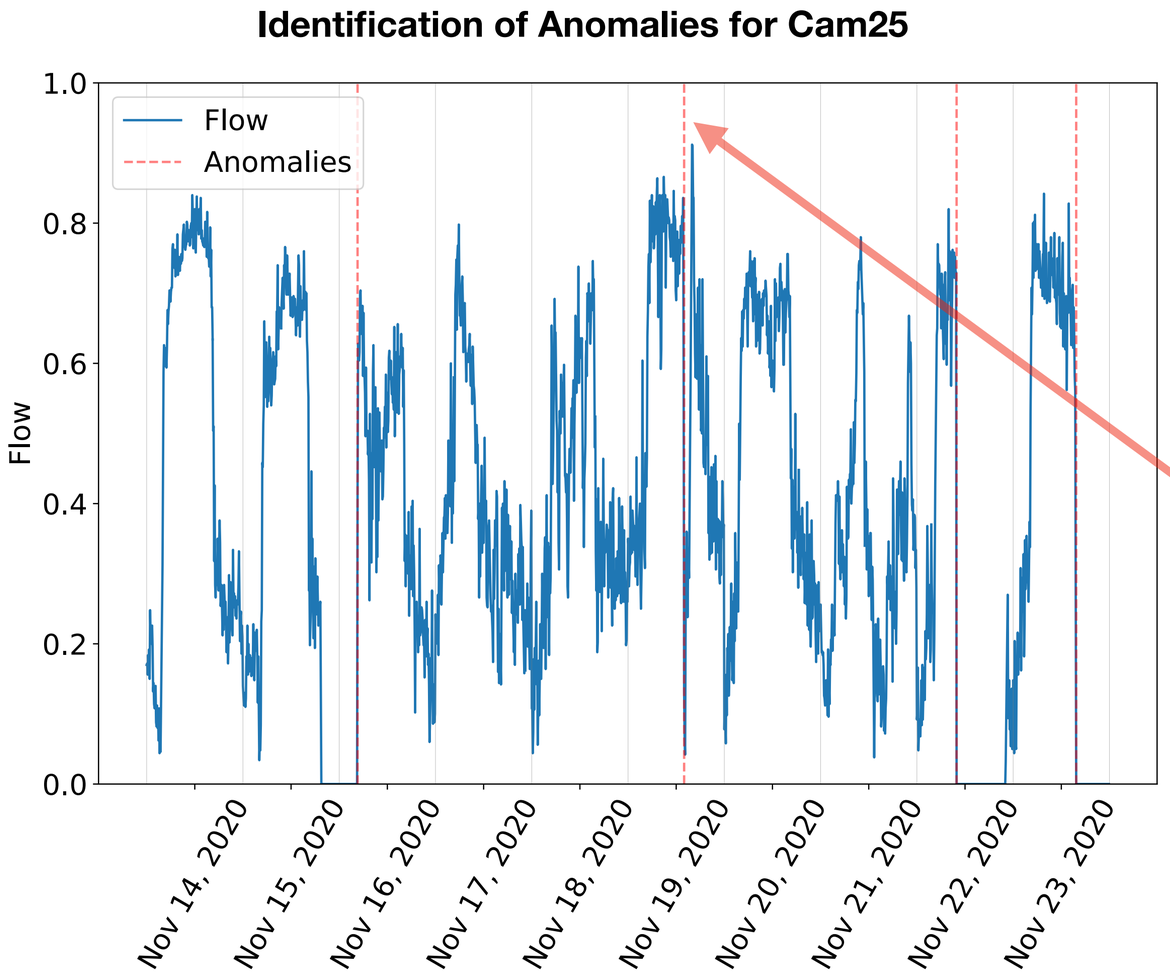
Ximantis Secondment

STGAN



Ximantis Secondment

Results



Nov 19, 2020, Cam14, **14:10**

Nov 19, 2020, Cam14, **14:20**

arXiv.2410.xxxxx

Future Work

Next 12 months

- **Submission** of FPGA and traffic AD papers
- **Quantization** of GNN
- Acceleration of GNN on **FPGA**
- Submission of **whitepaper 2**
- PhD thesis

Real Time Analysis of Unstructured Data with Machine Learning on Heterogeneous Architectures

Conclusion

arXiv.2406.12869

arXiv.2407.12119

arXiv.2410.xxxxx

Track Finding with ETX4VELO on GPUs

- **End-to-end** implementation in **LHCb** and first throughput results

FPGA/GPU ML Inference Throughput Comparison

- FPGA **outperforms** GPU with **60% of the power** consumption

Traffic Anomaly Detection

- Successfully **identified** traffic anomalies due to **extreme weather**

Future work

- Quantization of the **GNN**
- Acceleration of GNN on **FPGA**

Thank you!

This work is part of the SMARTHEP network and it is funded by the European Union's Horizon 2020 research and innovation programme, call H2020-MSCA-ITN-2020, under Grant Agreement n. 956086, and in collaboration with Ivan Kisel and FIAS under the ANN4Europe project.