

ETX4VELO: Graph Neural Network-Based Pipeline for Track Finding at LHCb

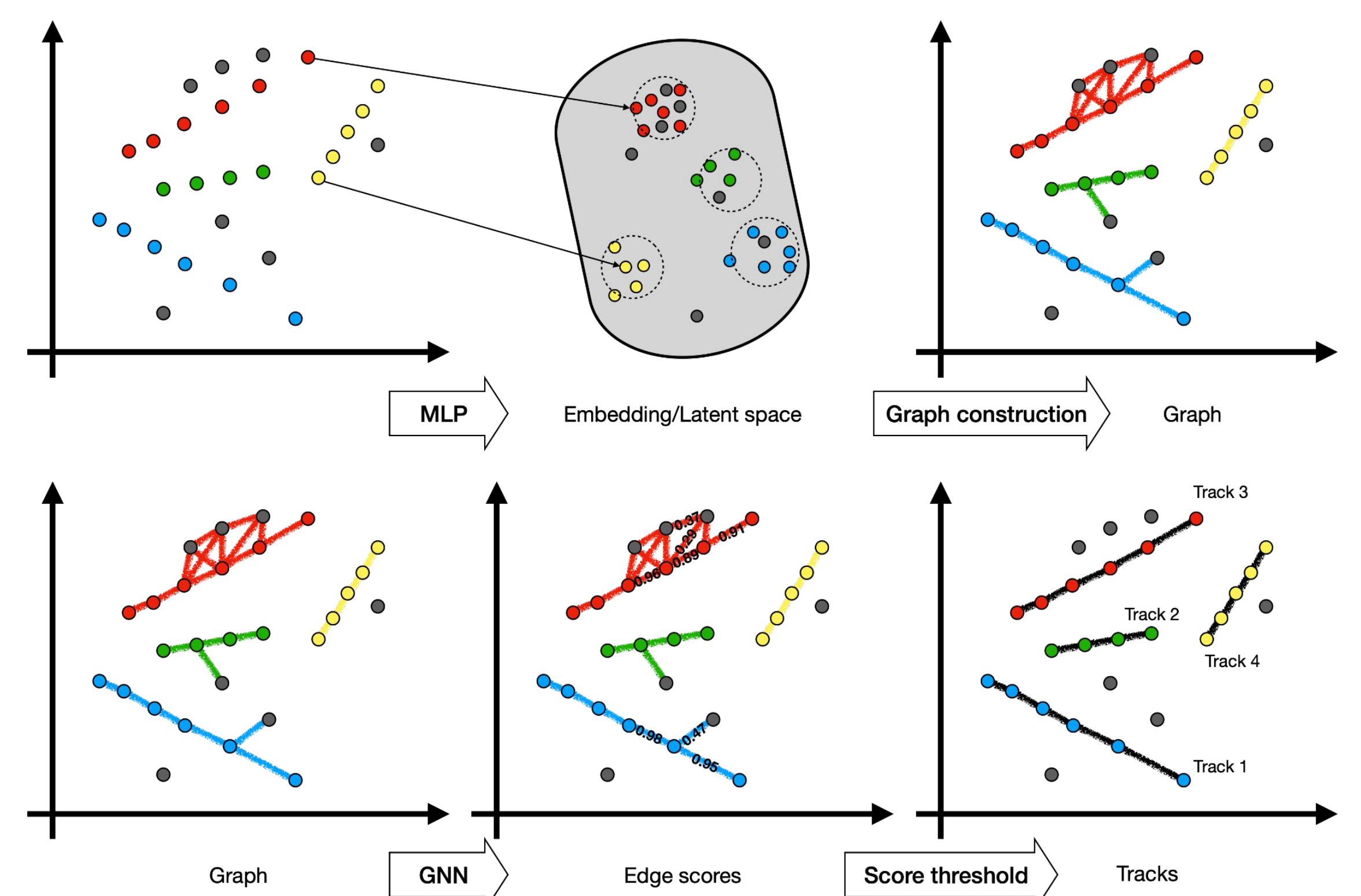
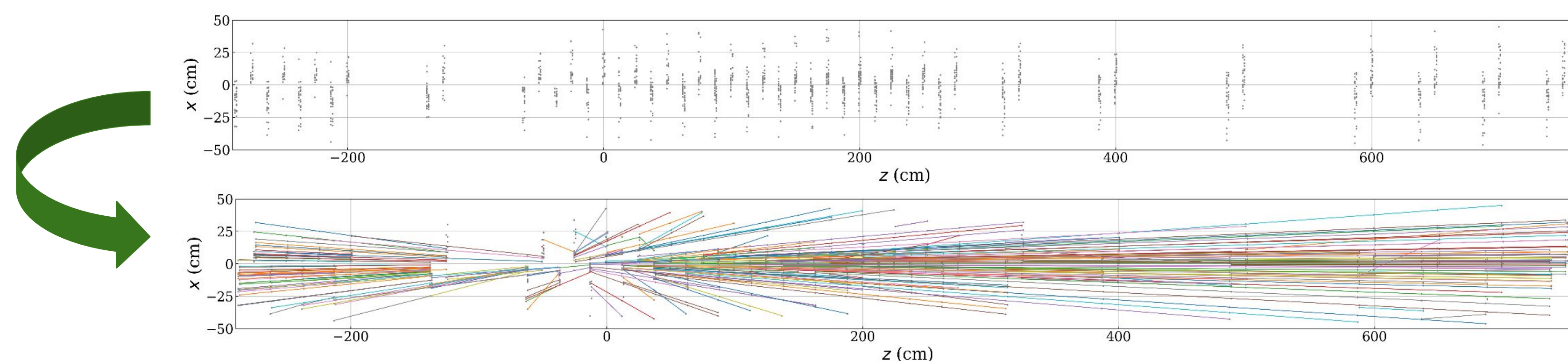
The focus is to evaluate deep-learning algorithms performance for **EFFICIENCY** and **THROUGHPUT**, and estimate how these models scale up with the increase of data rate.

For this purpose we developed the **ETX4VELO** pipeline which focuses on developing Graph Neural Networks (GNNs) algorithms for track reconstruction inside the VELO subdetector of the LHCb experiment.

Steps of the pipeline:

- **Embedding**
- **Graph Construction**
- **GNN**
- Triplets (not shown in the diagram)
- **Score cut**

The efficiency reached at this level is on par with the current algorithms in production inside the first-level trigger of LHCb. The **ETX4VELO** pipeline is based on **edge and triplet GNN** model that can reconstruct harsh cases of shared hits between tracks.



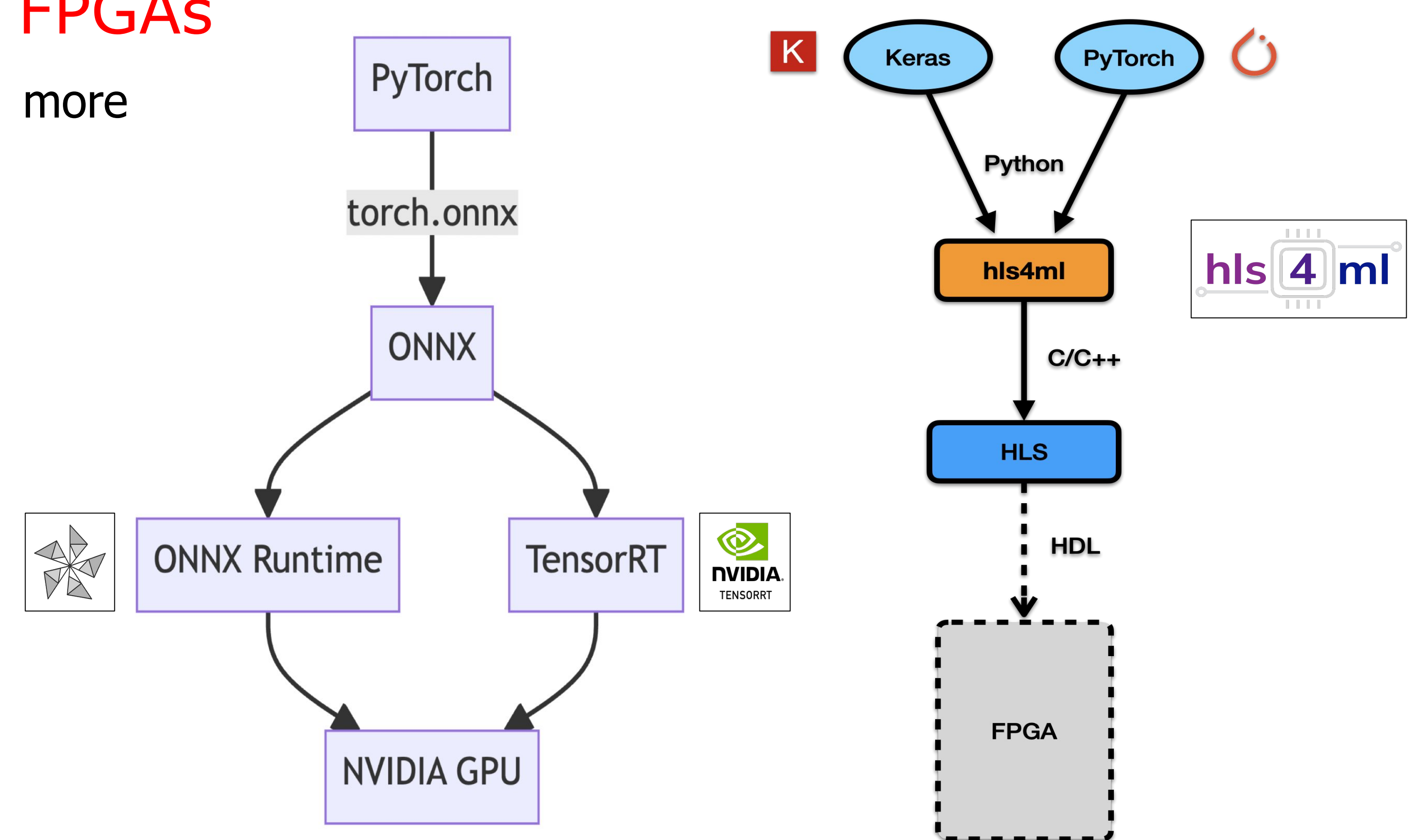
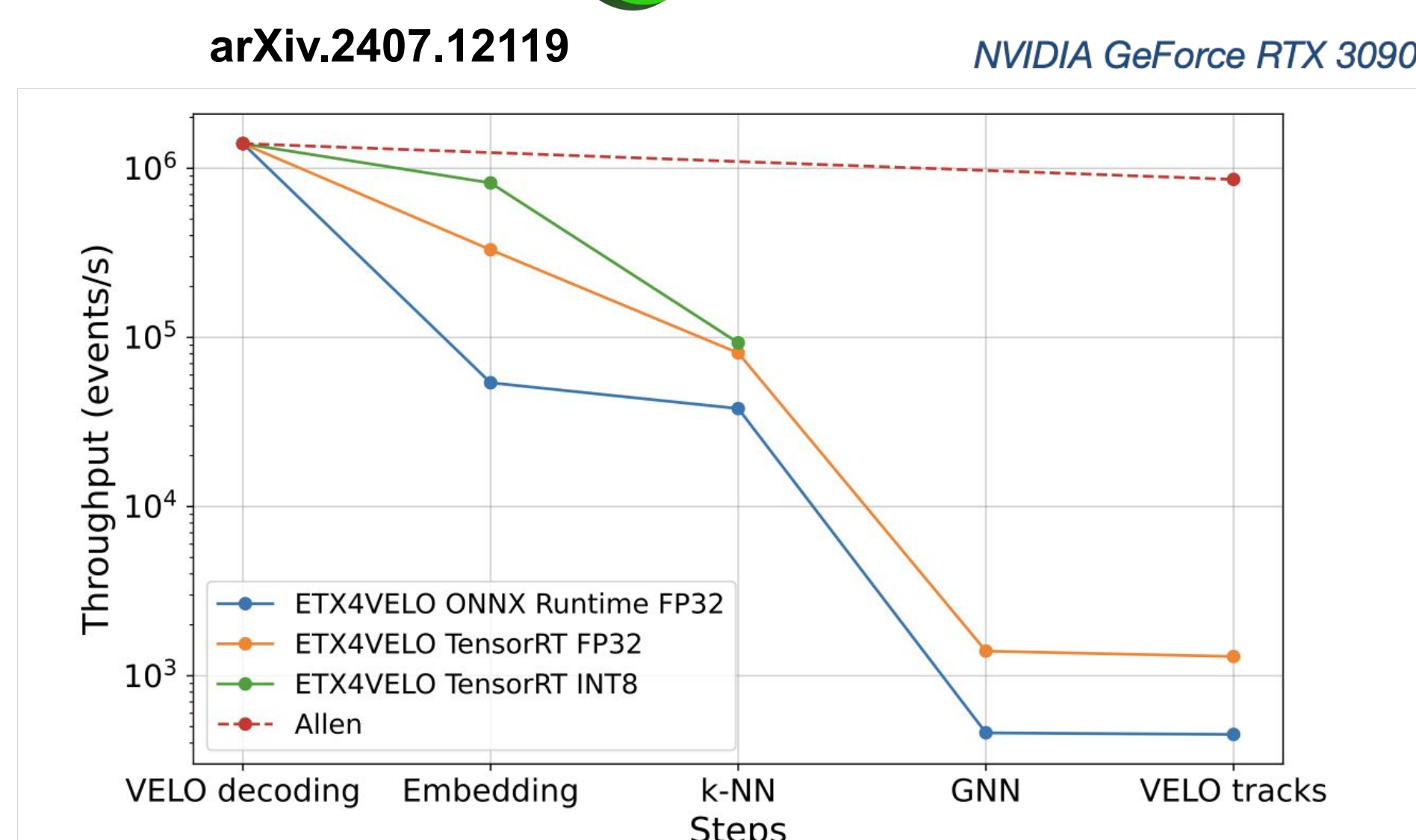
Category	Metric	Allen	ETX4VELO smaller graph	ETX4VELO Larger graph
Long, no electrons ✓ In acceptance ✓ Reconstructible in the velo ✓ Reconstructible in the SciFi ✓ Not an electron	Efficiency	99.26%	99.28%	99.51%
	Clone rate	2.54%	0.96%	0.89%
	Hit efficiency	96.46%	98.73%	98.90%
	Hit Purity	99.78%	99.94%	99.94%
Long electrons ✓ In acceptance ✓ Reconstructible in the velo ✓ Reconstructible in the SciFi ✓ Electron	Efficiency	97.11%	98.80%	99.22%
	Clone rate	4.25%	7.42%	7.31%
	Hit efficiency	95.24%	96.54%	96.79%
	Hit purity	97.11%	98.46%	98.46%
Long, from strange ✓ In acceptance ✓ Reconstructible in the velo ✓ Decays from a strange <i>Good proxy for displaced tracks</i>	Efficiency	97.69%	97.50%	98.06%
	Clone rate	2.50%	0.92%	0.81%
	Hit efficiency	97.69%	98.22%	98.77%
	Hit purity	99.34%	99.68%	99.68%
X	Ghost rate	2.18%	0.76%	0.81%

arXiv.2406.12869

Inference of the ETX4VELO Models on GPUs and FPGAs

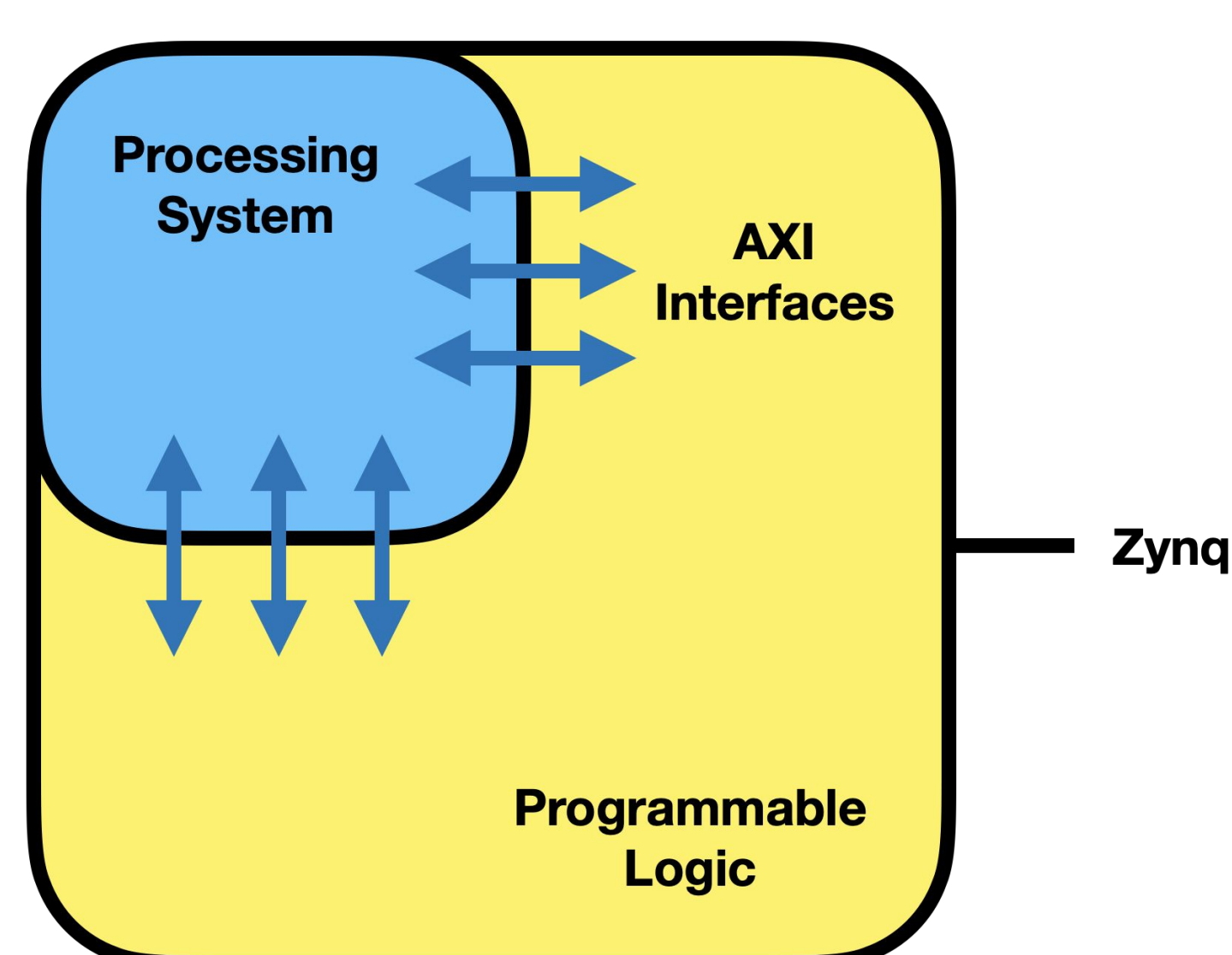
Then we focused on running these algorithms on architectures with more potential for performance.

- Integration inside the **LHCb's first-level GPU trigger (HLT1)**
- Inference of the MLP and GNN on **GPU**
- Inference of the MLP on **FPGA**



FPGA GPU Comparison

- MLP with INT8 precision implemented on PYNQ-Z2 **FPGA** using **HLS4ML**
- Synthesized with **Vivado HLS**
- Deployed using **PYNQ**
- Extrapolated to **Alveo** boards
- Using resource utilization and latency estimates
- Compared to **GPU** implementation on NVIDIA GeForce RTX 3090



arXiv.2502.02304

Accelerator Implementation	Alveo U50	Alveo U250	GeForce RTX 3090
Throughput (Events/s × 10 ⁶)	0.55	1.10	0.82
Active Power Draw (W)	75	230	350
Energy per Event (micro J)	140	210	430
Energy Gain	3.1x	2.0x	1.0x
Price (USD)	3,000	~ 10,000	1,500

Device	Memory (GB)	Price (USD)	TDP (W)
AMD Alveo U50	8	3,000	75
AMD Alveo U250	64	~10,000	230
NVIDIA GeForce RTX 3090	24	1,500	350

