Tracklet Classification with Distilled **Graph Neural Network on FPGA**

Amilqar Karam,[1,2] Philip Chang[2] and Tae Min Hong[3]

1 Johns Hopkins University, 2 University of Florida, 3 University of Pittsburgh

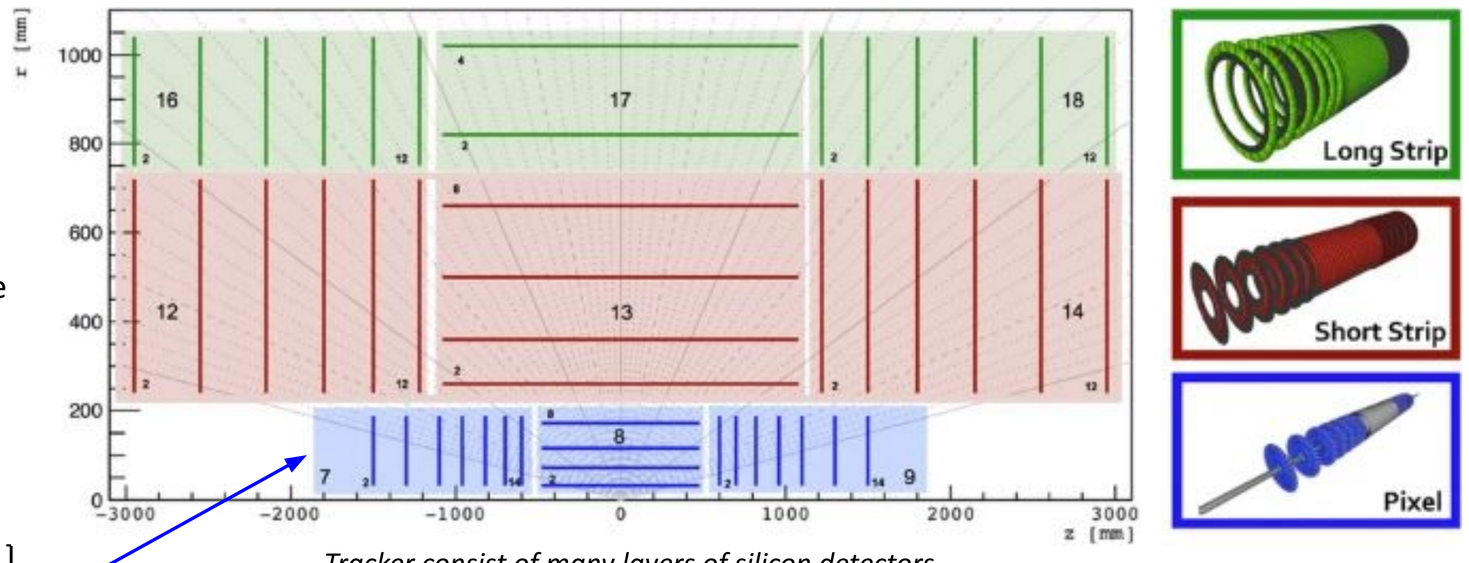
Setup

High-Luminosity LHC at CERN

- In 2029, HL-LHC will increase instantaneous luminosity by 4-5x and # of pileup pp interaction will increase proportionally
- Each pp interaction produces ~10 charged particles, which leave hits on silicon detectors in the Tracker (see right)
- Track reconstruction algorithm then must try to connect hits to create charged-particle track candidates

TrackML Dataset

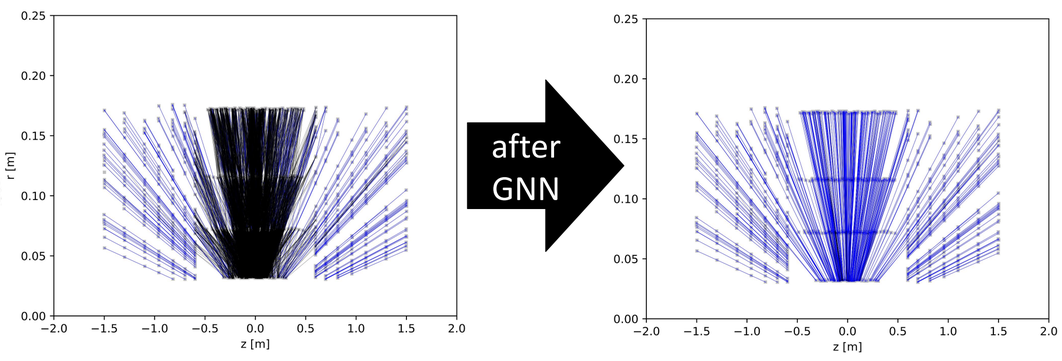
- We use a dataset from GitHub assuming a generic detector [1]
- We focus on the Si Pixel subdetector (volume #7, 8, 9 in blue)



Tracker consist of many layers of silicon detectors

Graph Neural Network (GNN)

- Si hit: graph node with global cylindrical position coordinates $x_k = (r_k, \phi_k, z_k)$
- Tracklets: edges that connect two hits + their differences [2] $a_{ij} = (\Delta r_{ij}, \Delta\phi_{ij}, \Delta z_{ij}, \Delta R_{ij})$
  where $\Delta R_{ij} = \sqrt{\Delta\eta_{ij}^2 + \Delta\phi_{ij}^2}$
- Graph: all graph nodes and edges that represent all tracklets in per pp bunch crossing
- GNN: receives graph, outputs tracklet scores (black: fake tracklet, blue: true tracklet)

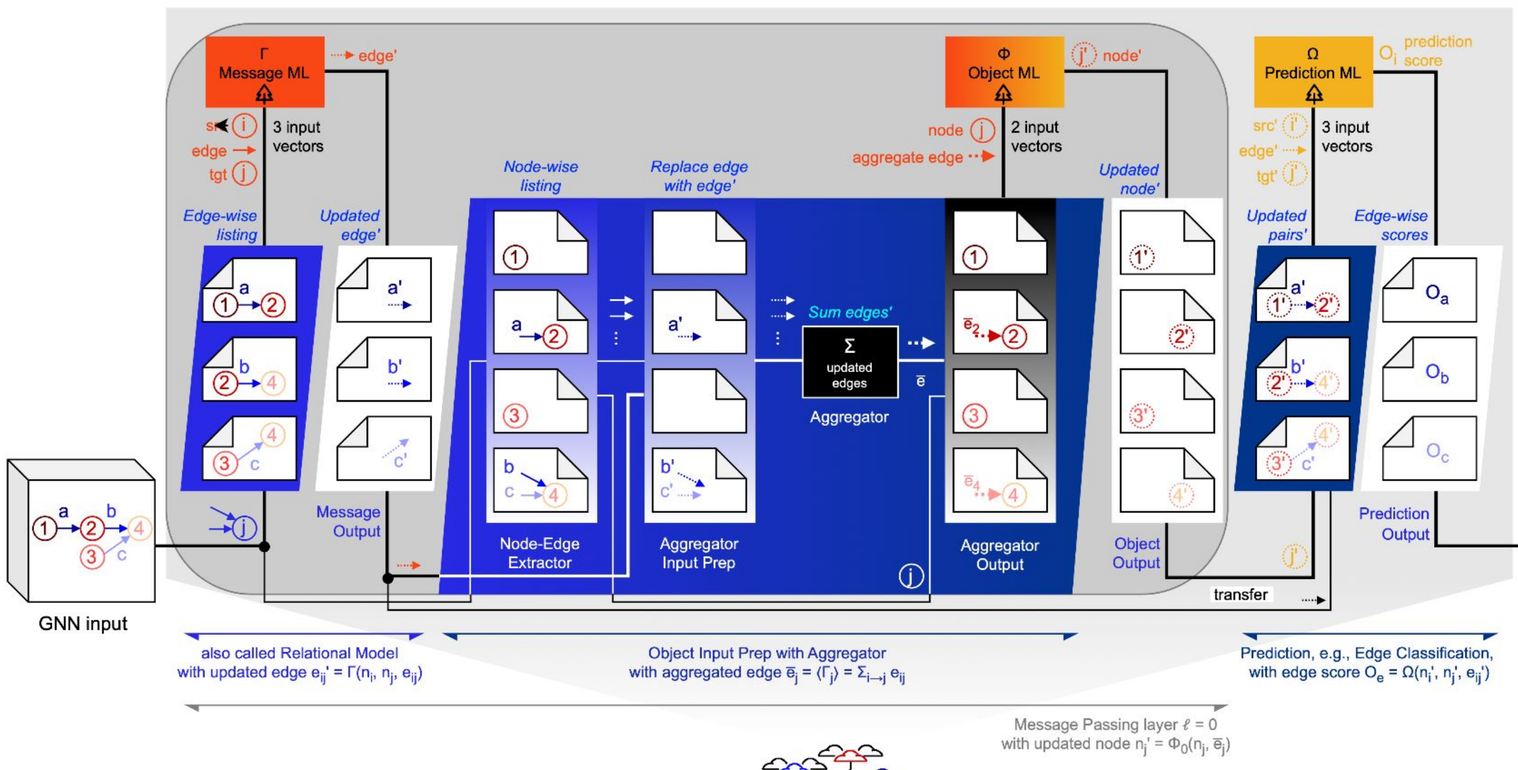


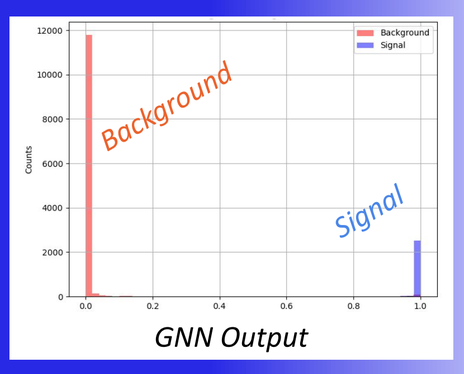
Input graph *After threshold*

Method

GNN Model

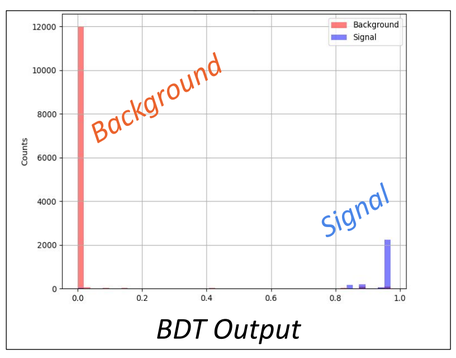
- Multilayer perceptrons (MLP)
  - Relational Model, "Message"
  - Object Model, "Object"
  - Edge Classifier, "Prediction"
- MLP
  - 2 hidden layers
  - 10 NN nodes for each layer
- Message passing: one time
- GNN shows good S/B separation





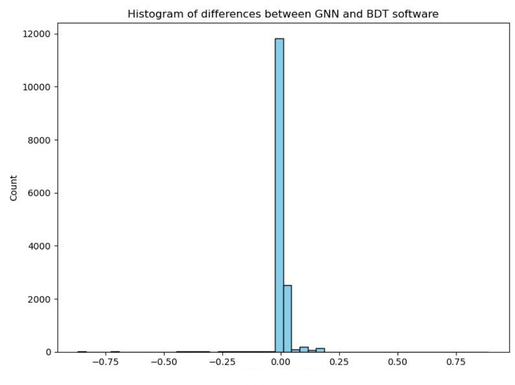
GNN Output

Distill

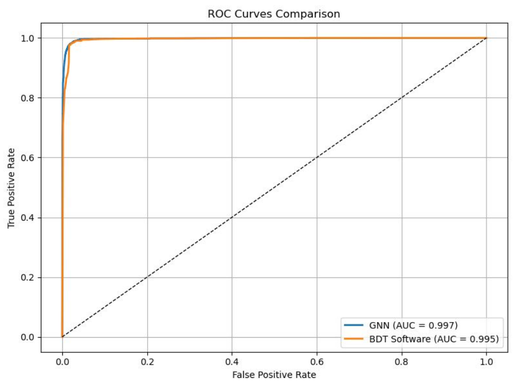


BDT Output

Bons



Difference



GNN Distillation

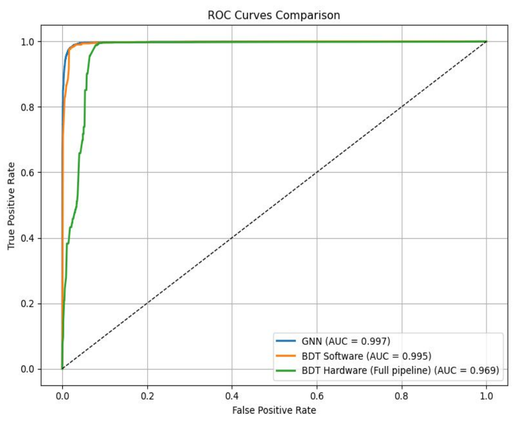
- Use boosted decision trees (BDT)
  - Regress 3 MLP scores with 3 sets of BDTs (max depth 4, # trees = 6, 7 ,10)

Implement

Results

FPGA details

- Xilinx Vertex Ultrascale+ VU9P FPGA
  - I simulated AUC: hardware and software classification performance show good agreement
  - I wrote VHDL of GNN structure that handles inputs/outputs + converted BDT into VHDL [3] + plugged the IPs in



Firmware performance

- **Successful synthesis** with 320 MHz clock on xcvu9p-flga2104-2L-e
- Resources: **sub-1% usage** (LUT 15k, FF 4k, DSP 0, BRAM: 28.5, IO: 446)
- Latency *for 739 graph nodes, 1252 graph edges*: 106 µs

Next steps

- To parallelize input preparation. Currently, the total latency & thruput scales with # tracklets.

References

[1] *TrackML Particle Tracking Challenge,* https://www.kaggle.com/competitions/trackml-particle-identification/data
[2] *Improving tracking algorithms with machine learning: a case for line-segment tracking at the High Luminosity LHC*, arXiv:2403.13166
[3] *Nanosecond machine learning regression with deep boosted decision trees in FPGA for high energy physics*, JINST 17 (2022) P09039C

CPAD 2025 at Penn

RDC 4 & 5