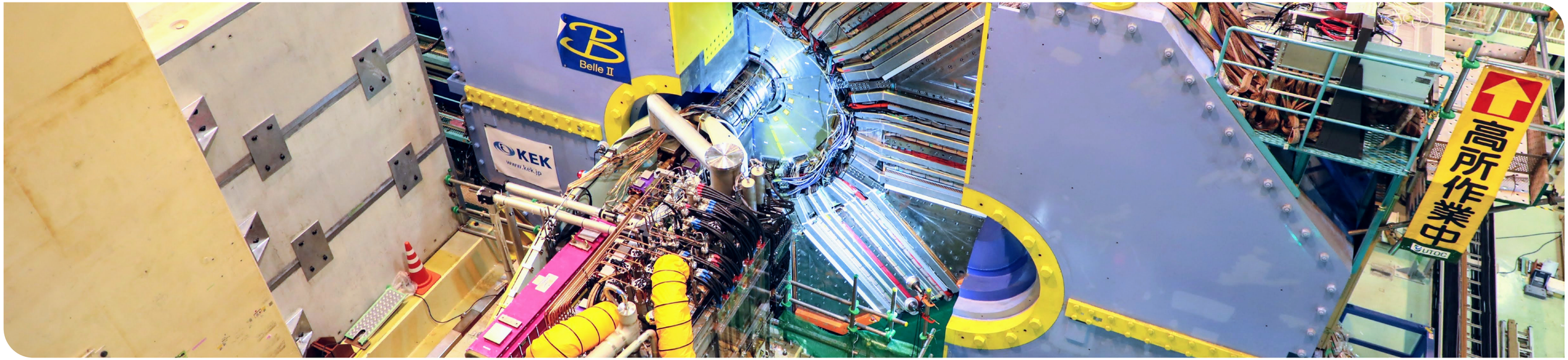
Karlsruhe Institute of Technology

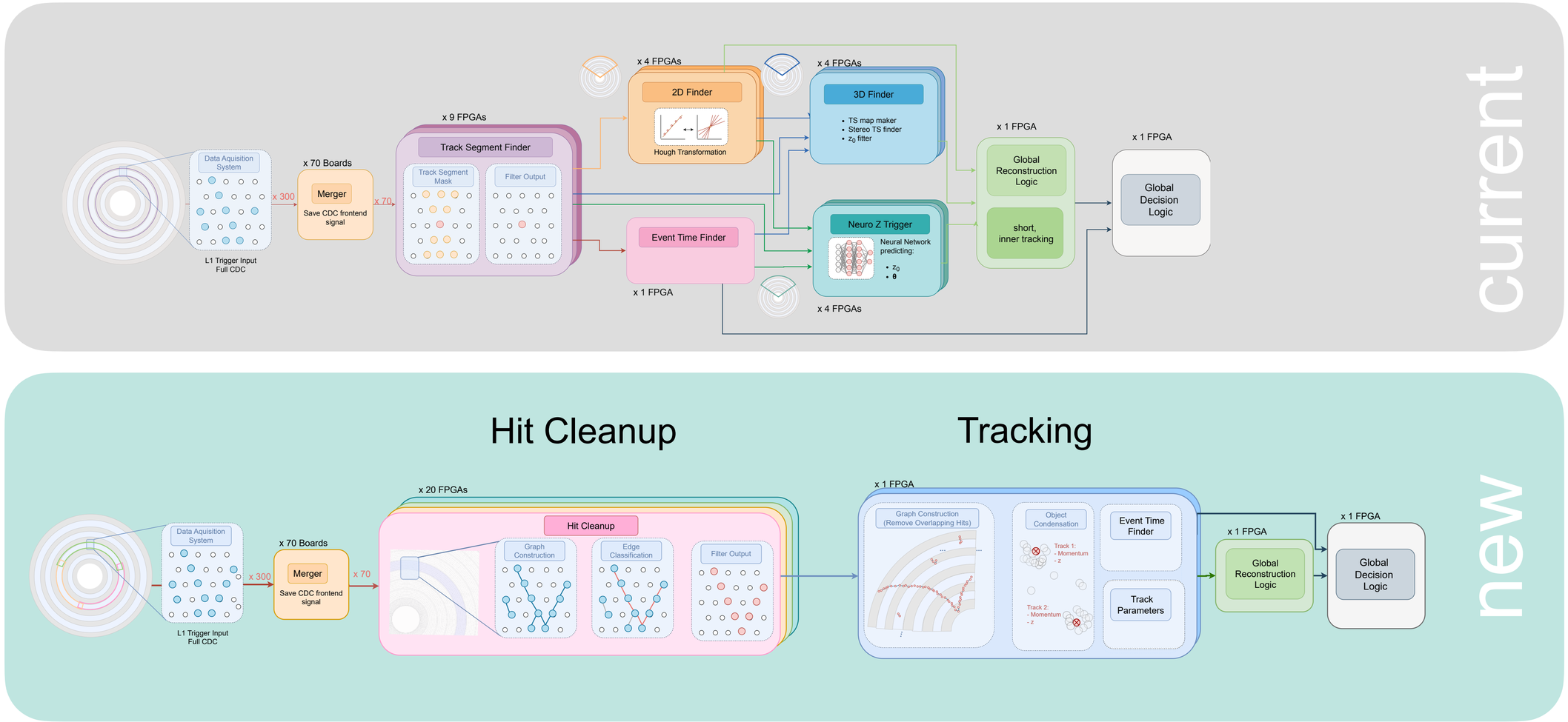
Institute of Experimental Particle Physics

Hit Cleanup with GNNs for real-time tracking at Belle II

Greta Heine, Torben Ferber, Lea Reuter, Slavomira Stefkova | 10.04.2024

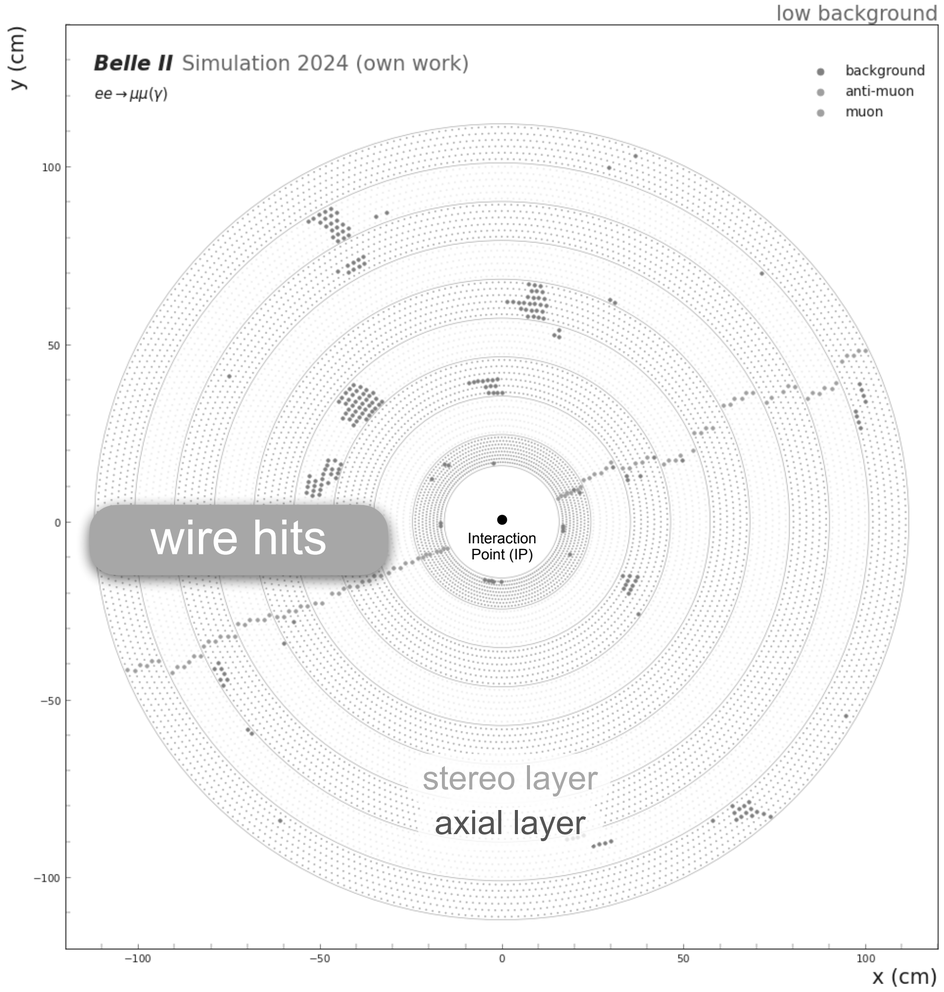


New CDC Trigger

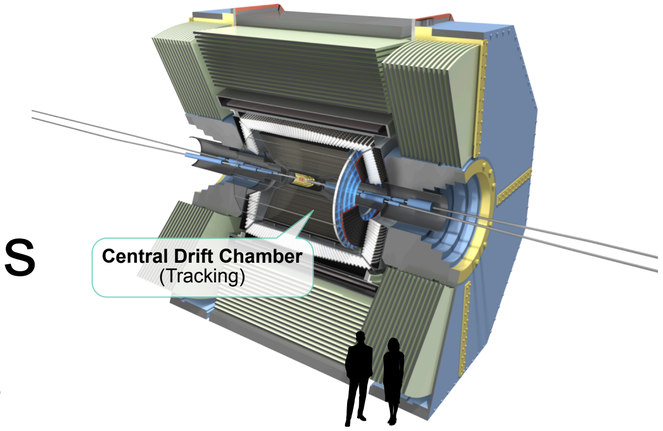


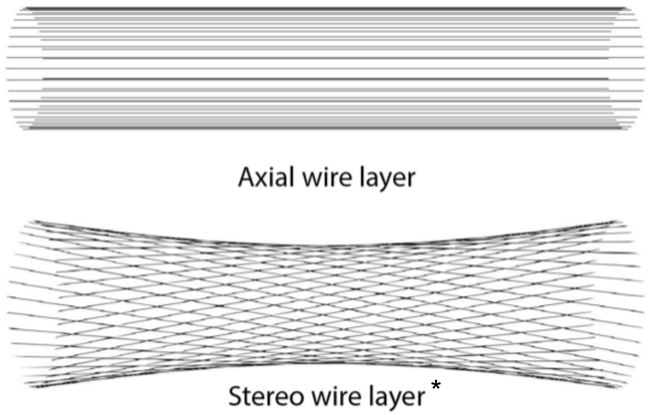
The Problem Challenge

Central Drift Chamber Event Display



- xy projection
- ≈ 15 000 wires
- 9 super layers
- alternating axial and stereo layers

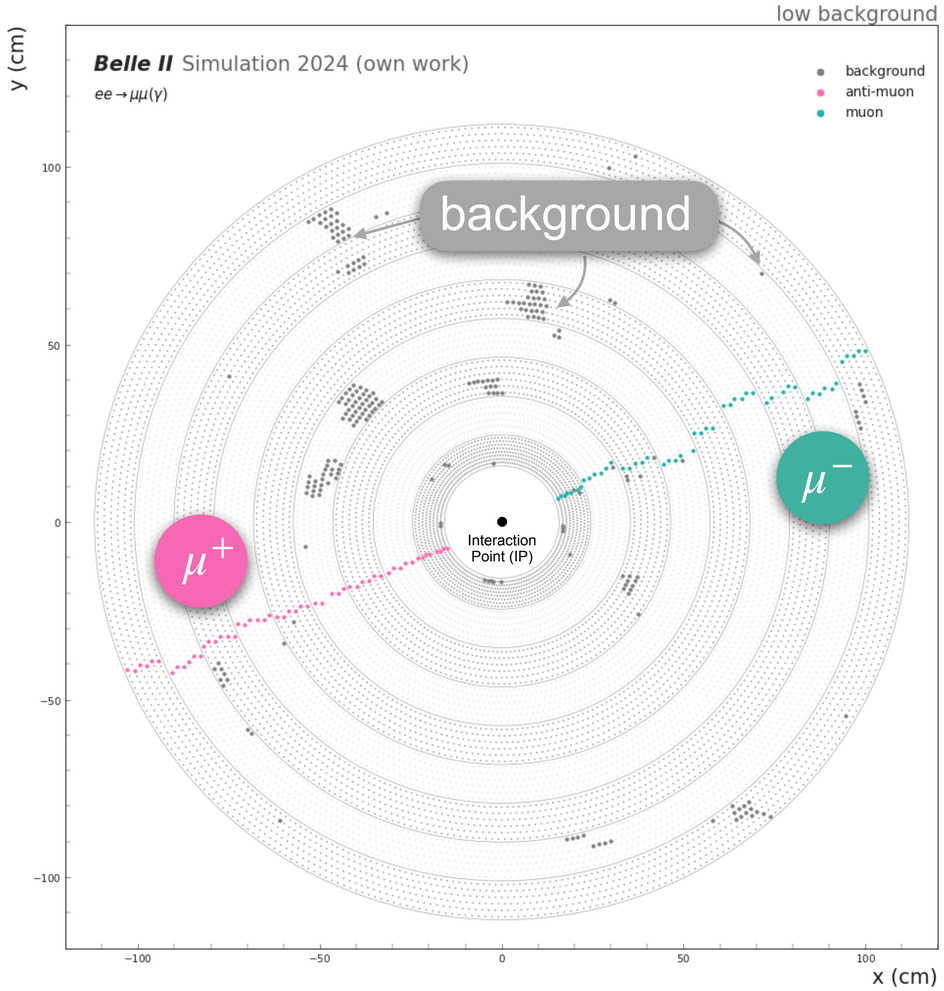




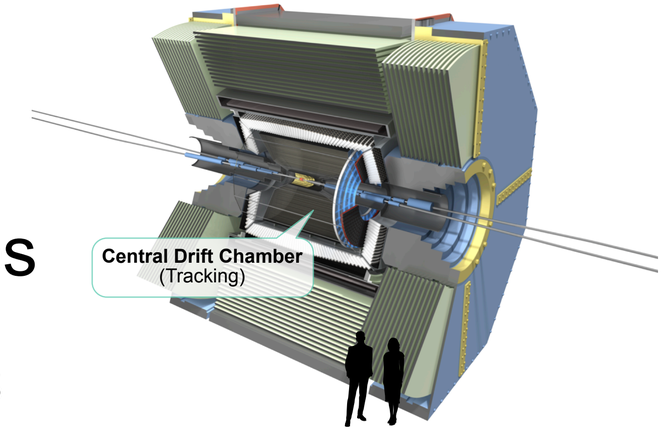
3D information

*rotation is hugely exaggerated for illustration

Central Drift Chamber Event Display

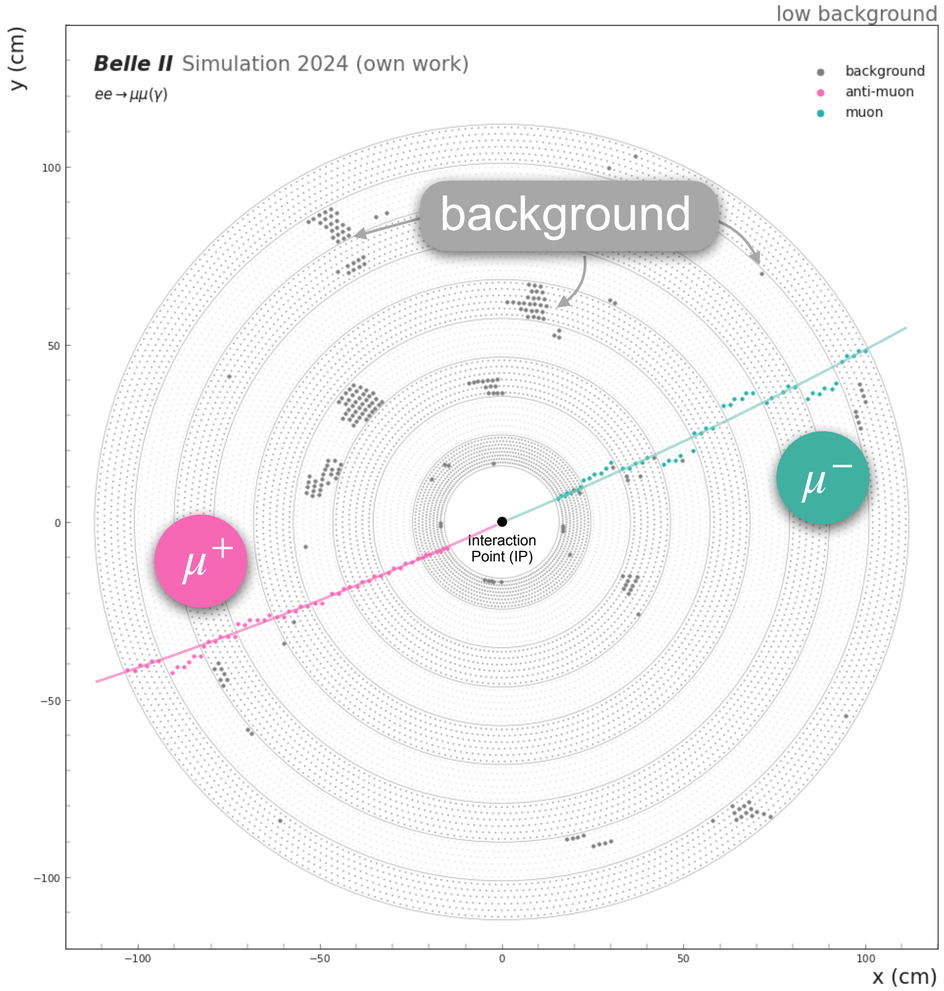


- xy projection
- ≈ 15 000 wires
- 9 super layers
- alternating axial and stereo layers



- MC simulated $\mu^+\mu^-$ pair
- overlayed with real background

Central Drift Chamber Event Display

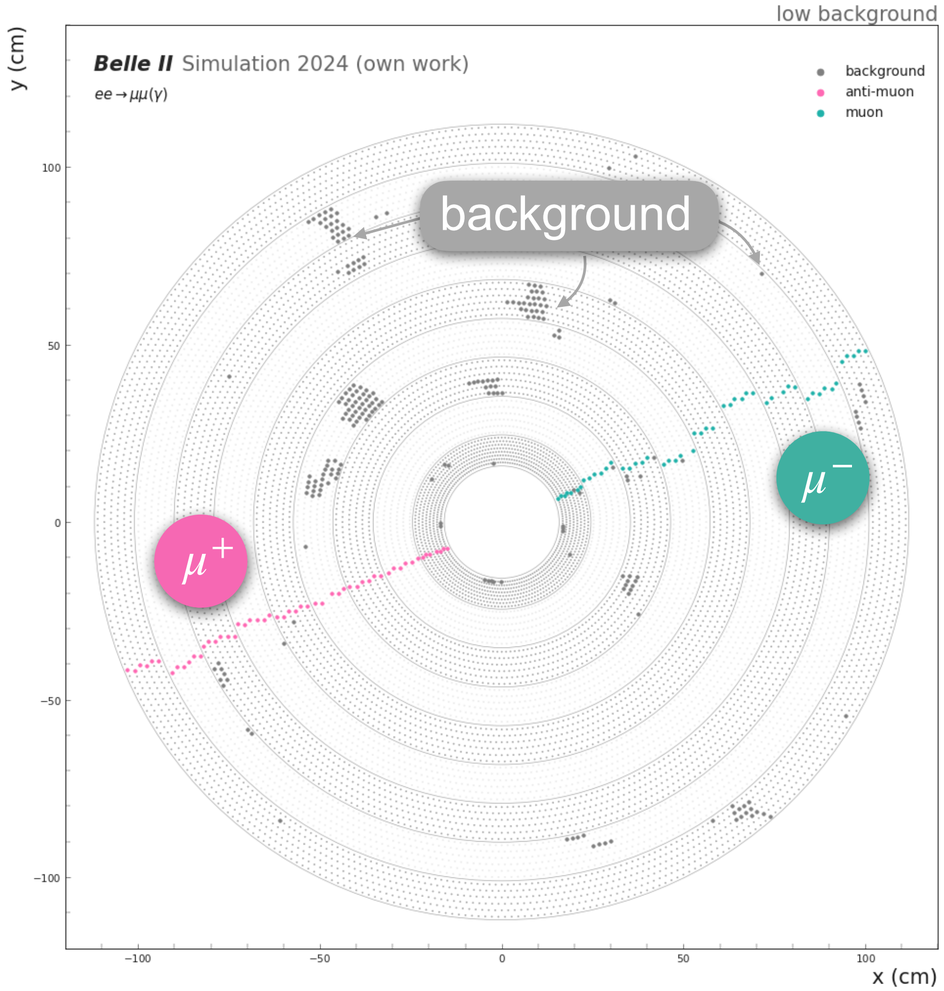


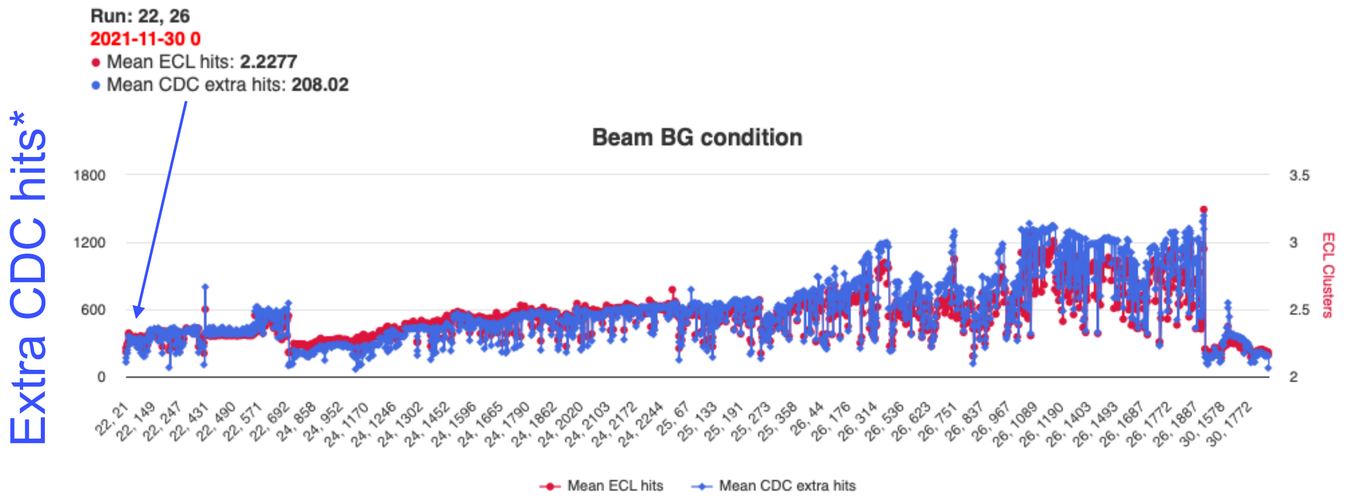
Tracking

1) reject background hits
2) identify signal hits (track finding)
3) estimate curvature $\propto p_T$ for trigger decision

- MC simulated $\mu^+\mu^-$ pair
- overlayed with real background

Back in 2021…

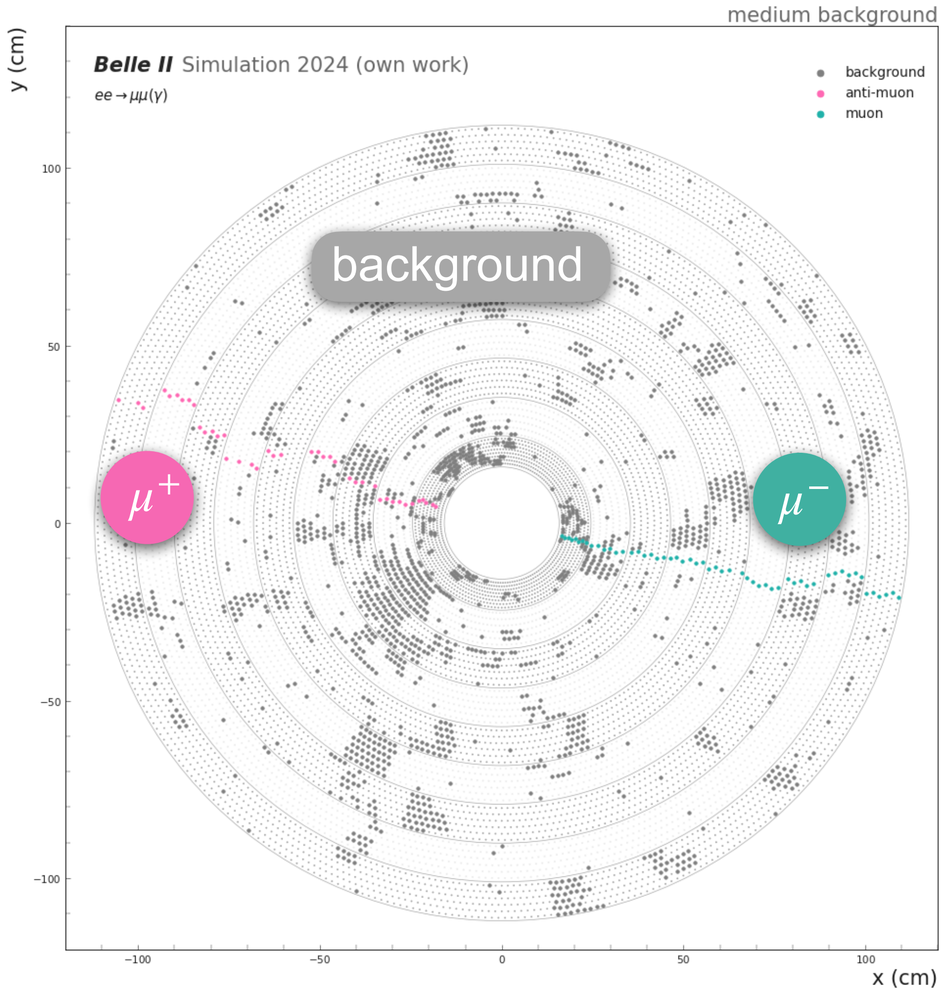


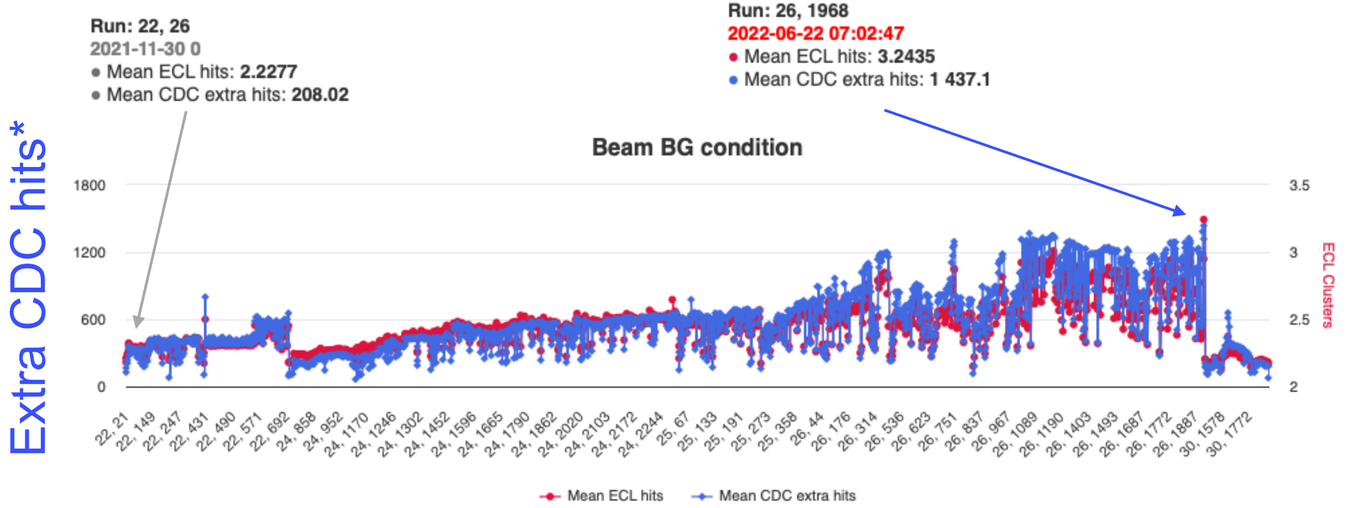


*The number of CDC hits in the event not assigned to any track.

- low background (2021)
- ≈ 1.9% of wires hit
- low computing cost

Now…



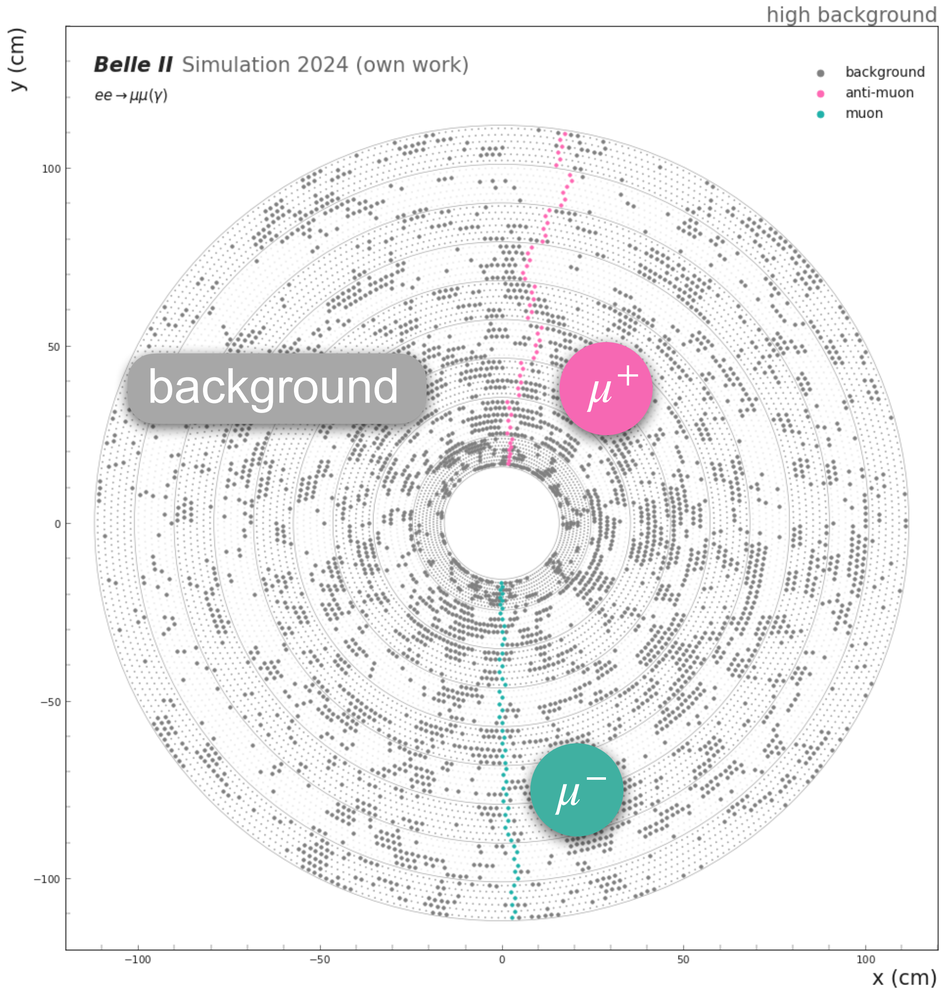


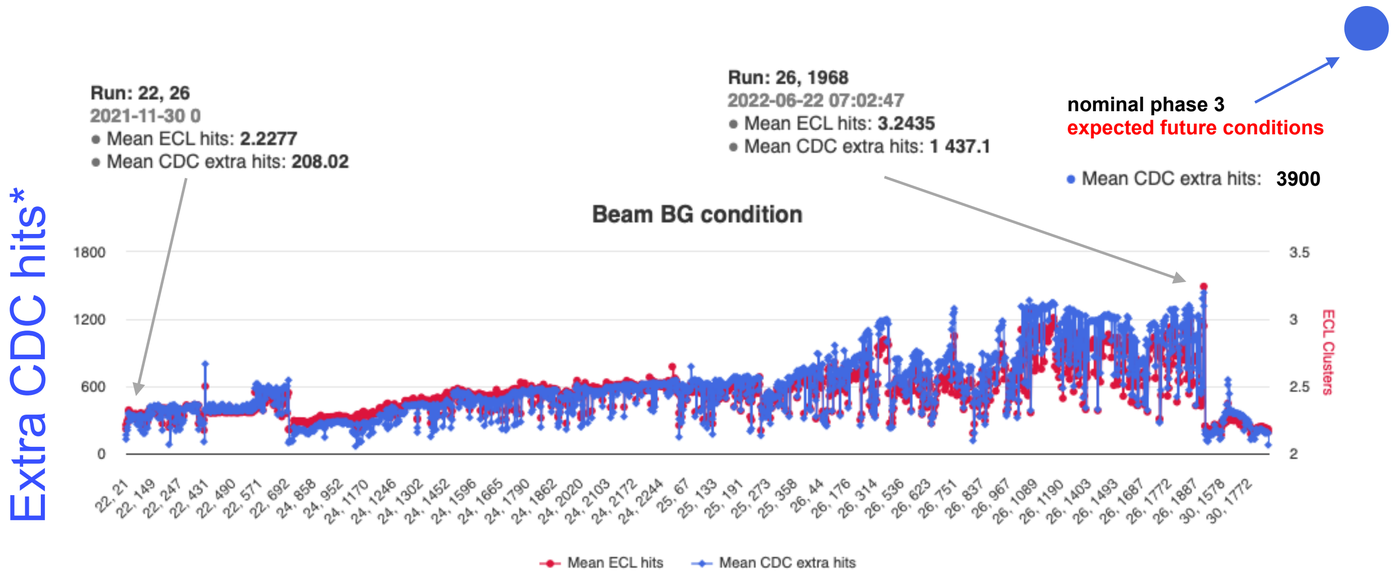
*The number of CDC hits in the event not assigned to any track.

- medium background (2022)
- ≈ 10.5% of wires hit
- moderate computing cost

a bit harder

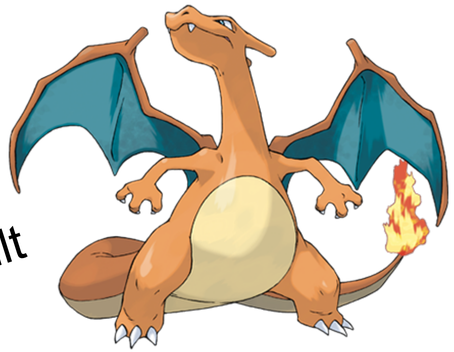
Future…





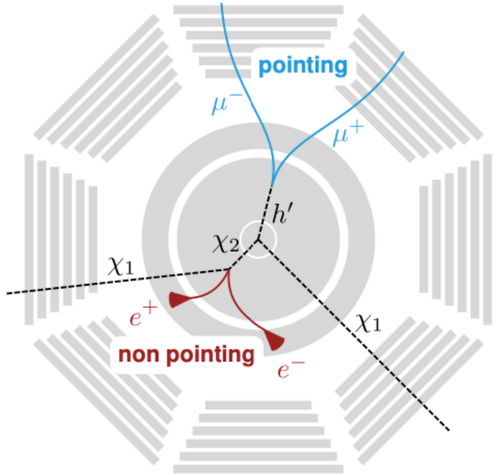
*The number of CDC hits in the event not assigned to any track.

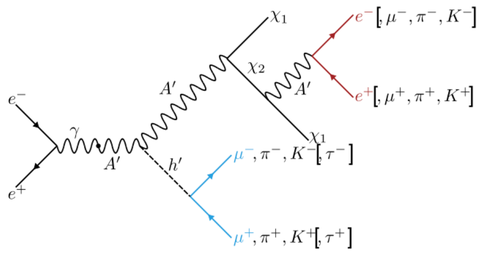
- high background (future)
- ≈ 27.1% of wires hit
- high computing cost



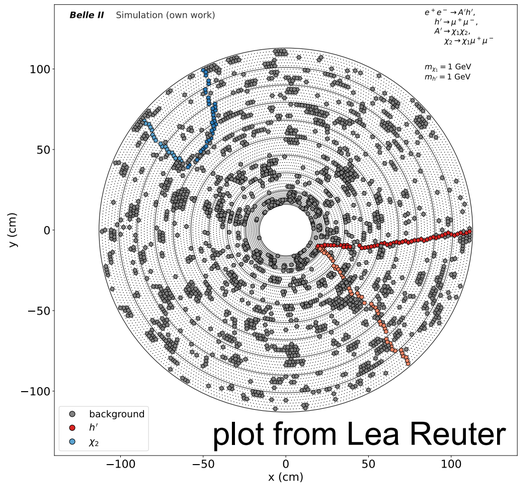
Displaced vertices

- tracks not originating from the IP



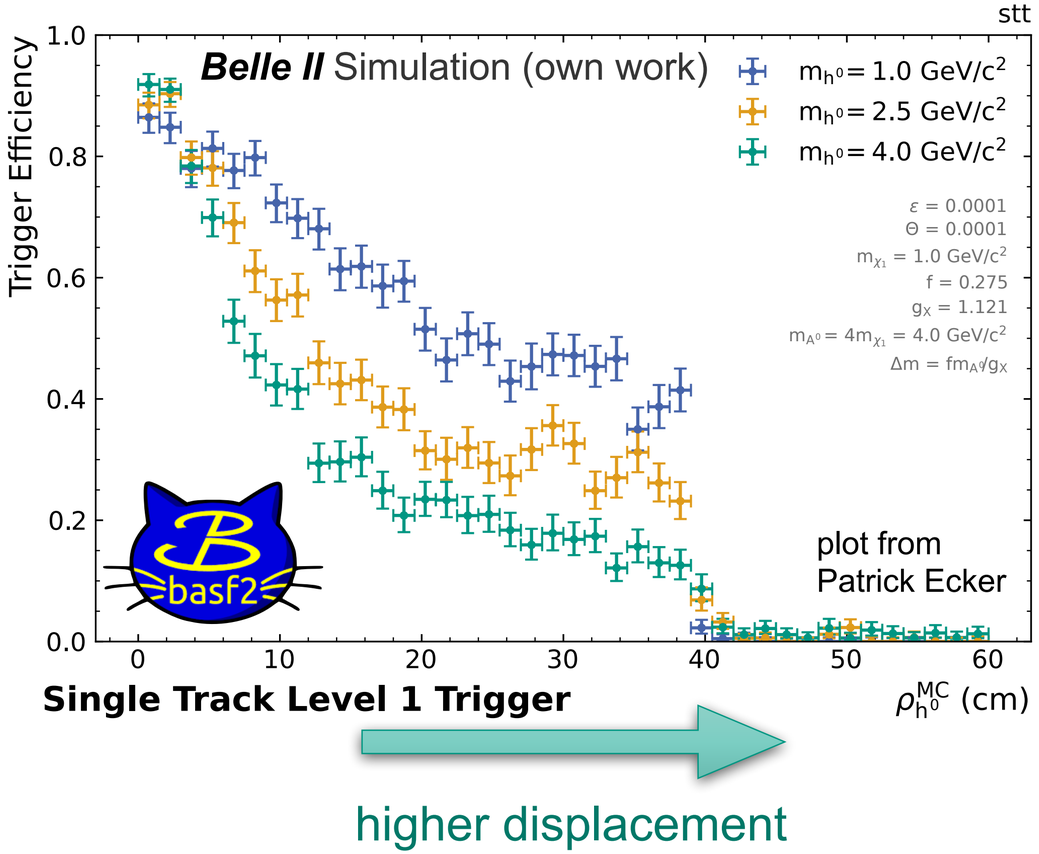


Plots from Patrick Ecker



efficiency

Current Tracking Algorithm

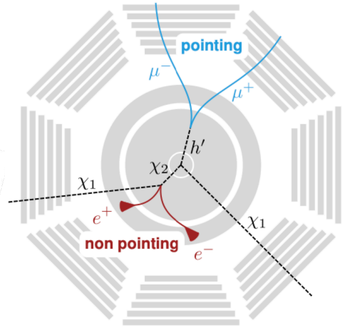


higher displacement

Real time tracking: the challenges

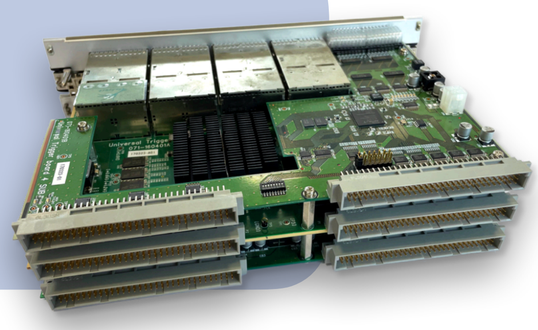
Tracking

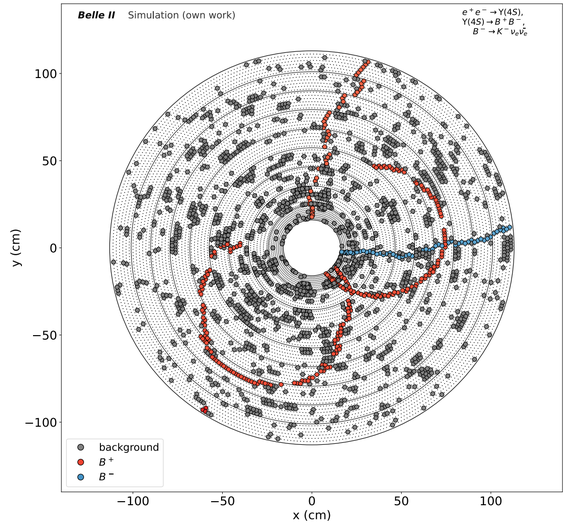
- increasing number of background hits ($t_{\text{comp}} \propto n_{\text{hits}}^2$)
- displaced vertices
- low track multiplicity
- z-resolution



Implementation

- sub-microsecond latency
- limited computing resources

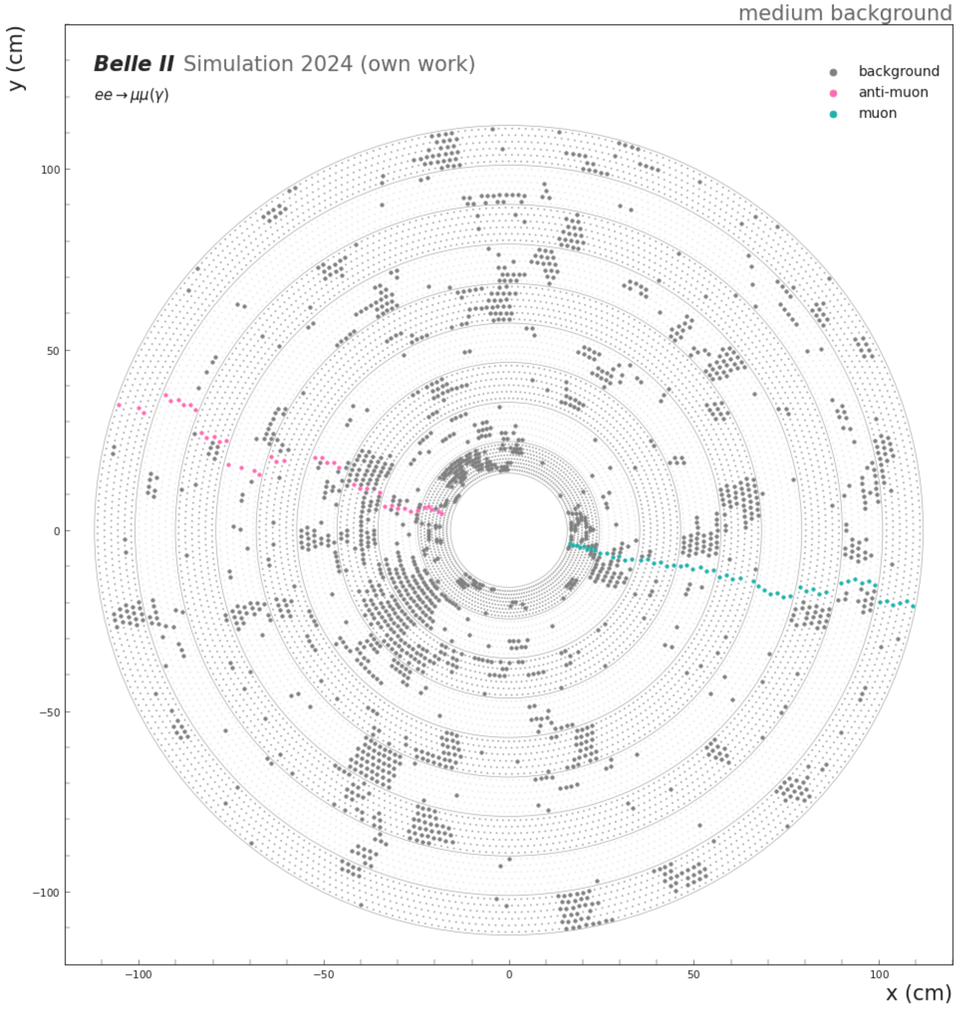


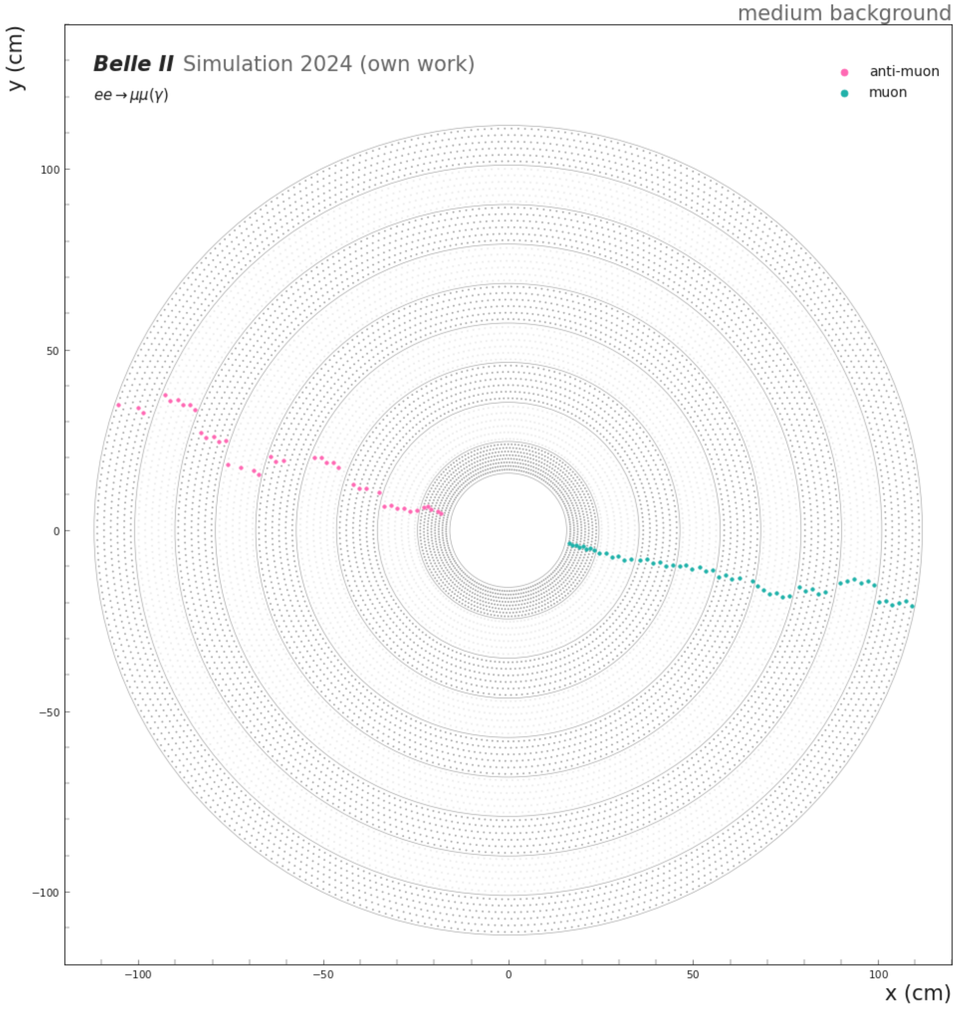


plot from Lea Reuter

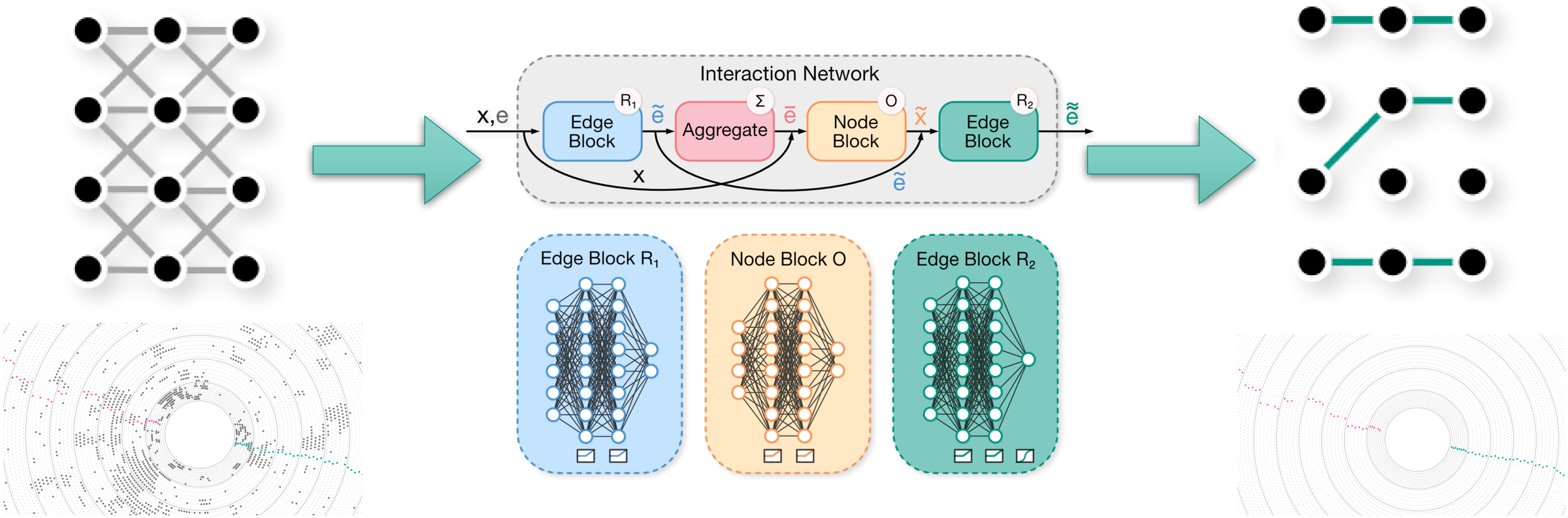
The (intermediate) Goal

Hit Cleanup


medium background
Belle II Simulation 2024 (own work)
$ee \to \mu\mu(\gamma)$
background
anti-muon
muon
y (cm)
x (cm)

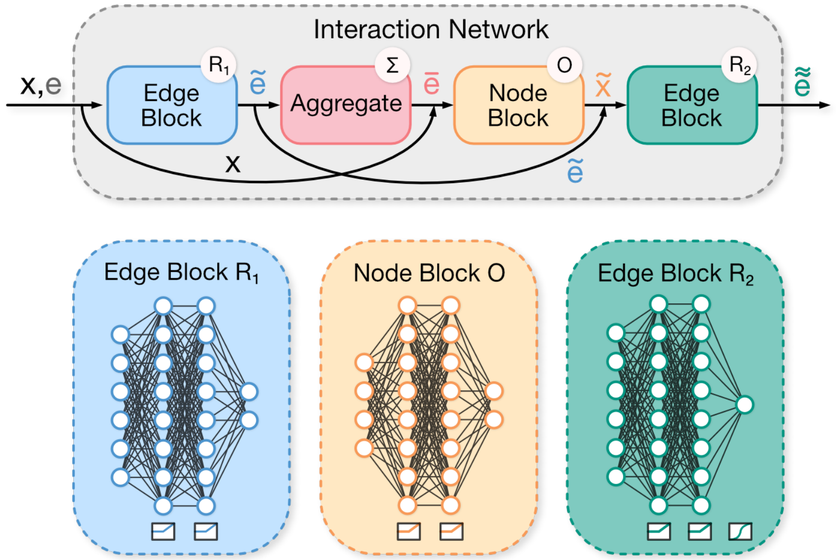

medium background
Belle II Simulation 2024 (own work)
$ee \to \mu\mu(\gamma)$
anti-muon
muon
y (cm)
x (cm)

Hit Cleanup: with a Graph Neural Network



Interaction Network

- Interaction network model by Battaglia et al. with PyTorch Geometric
- Inputs: hit information encoded in graphs
- Outputs: edge-level predictions




```
InteractionNetwork(node_dim: 4, edge_dim: 3, hidden_size: 8)
+--------------------+------------+
|      Modules       | Parameters |
+--------------------+------------+
| R1.layers.0.weight |     88     |
|  R1.layers.0.bias  |     8      |
| R1.layers.2.weight |     64     |
|  R1.layers.2.bias  |     8      |
| R1.layers.4.weight |     24     |
|  R1.layers.4.bias  |     3      |
| O.layers.0.weight  |     56     |
|  O.layers.0.bias   |     8      |
| O.layers.2.weight  |     64     |
|  O.layers.2.bias   |     8      |
| O.layers.4.weight  |     32     |
|  O.layers.4.bias   |     4      |
| R2.layers.0.weight |     88     |
|  R2.layers.0.bias  |     8      |
| R2.layers.2.weight |     64     |
|  R2.layers.2.bias  |     8      |
| R2.layers.4.weight |     8      |
|  R2.layers.4.bias  |     1      |
+--------------------+------------+
Total Trainable Params: 544
```

Edge block R1

for edge feature updates

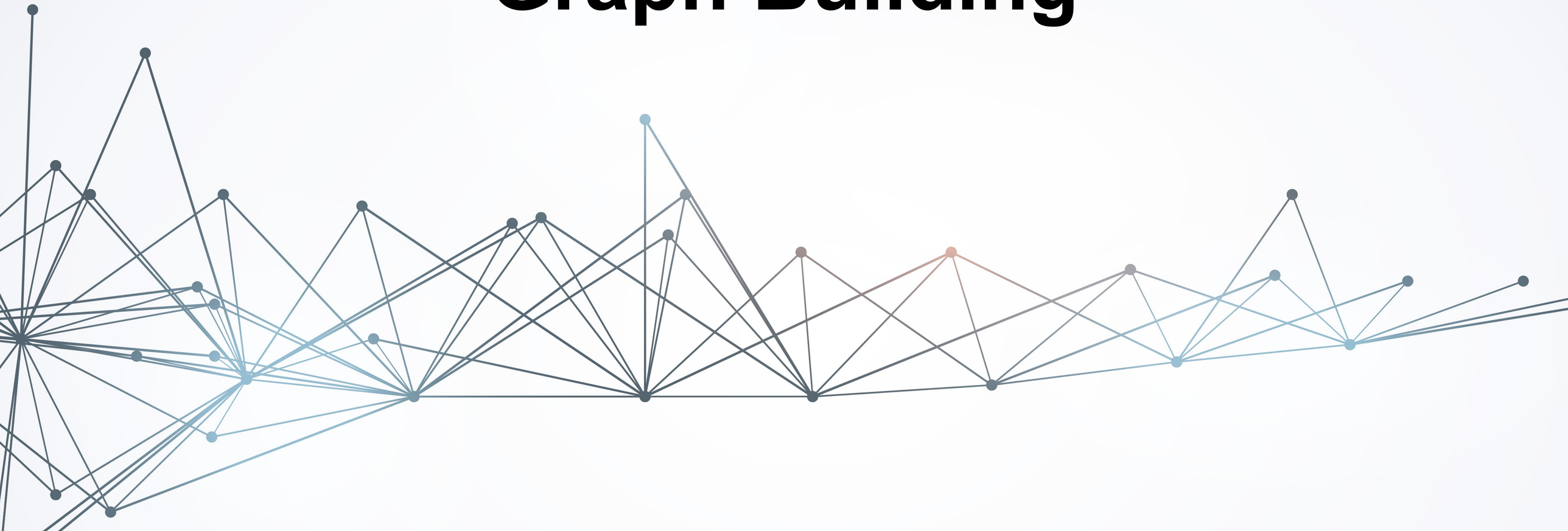
Node block O

for hit feature updates

Edge block R2

1-dim edge-level output

Graph Building



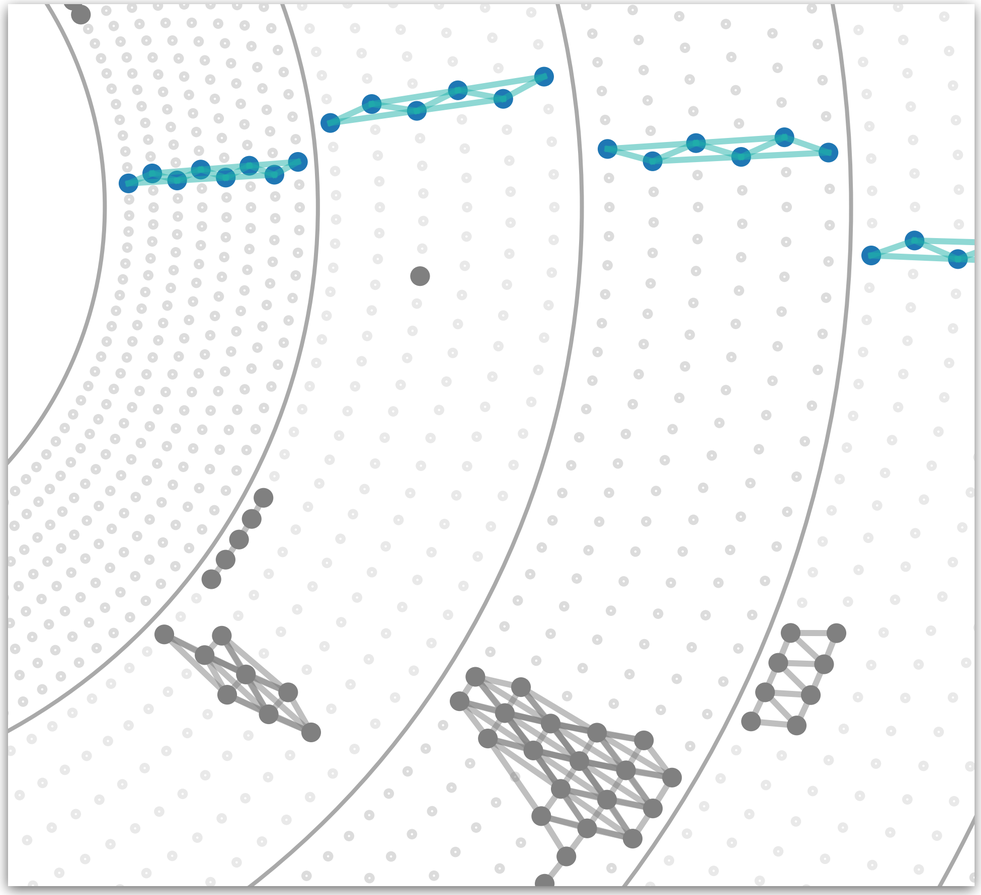
Graph Building

- connect hits to graphs
- connect *just the right amount* of hits with each other
- fully connected graph

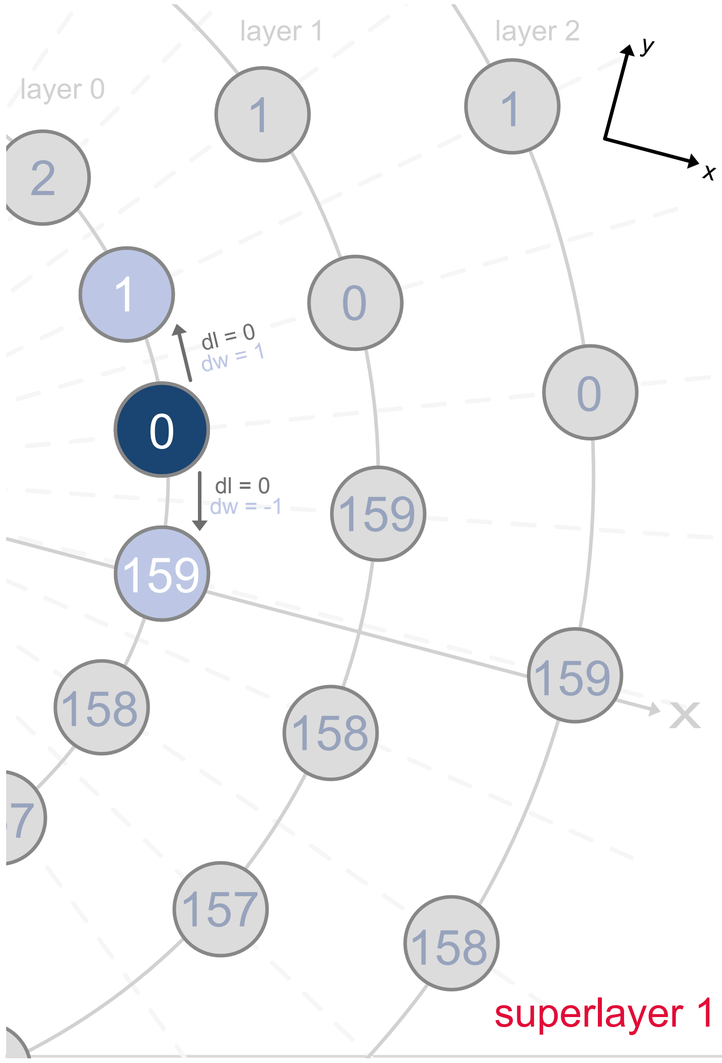
$$n_{edges} = \frac{1}{2} n_{hits}(n_{hits} - 1) \approx n_{hits}^2$$

- design goal

$$n_{edges} \approx 2 \cdot n_{hits}$$

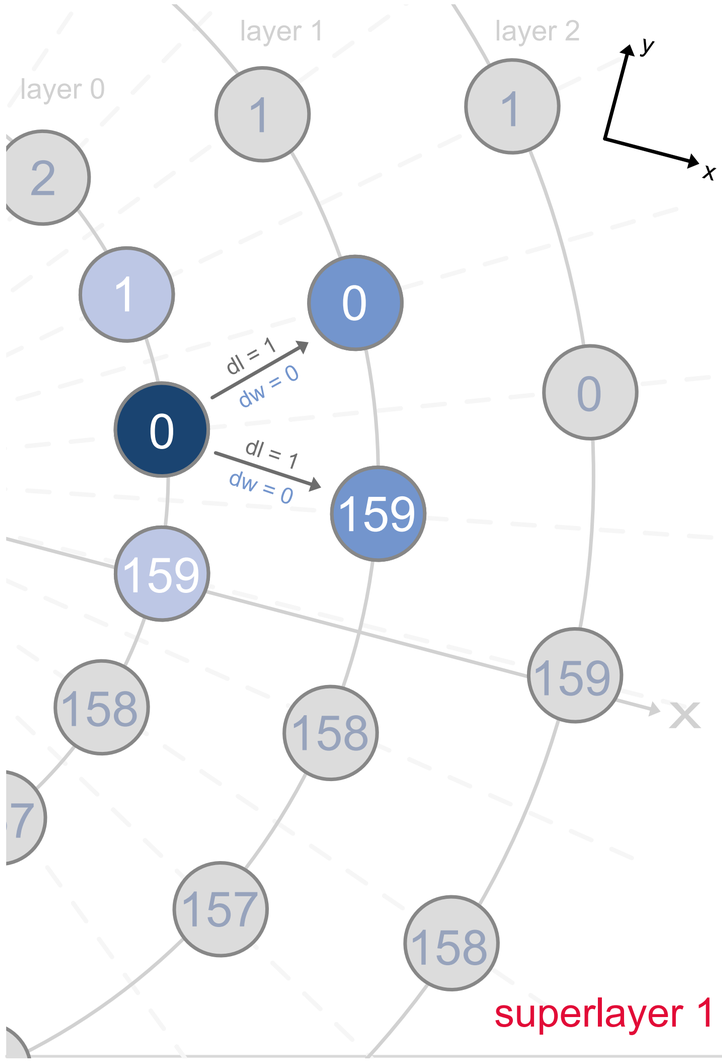


Graph Building



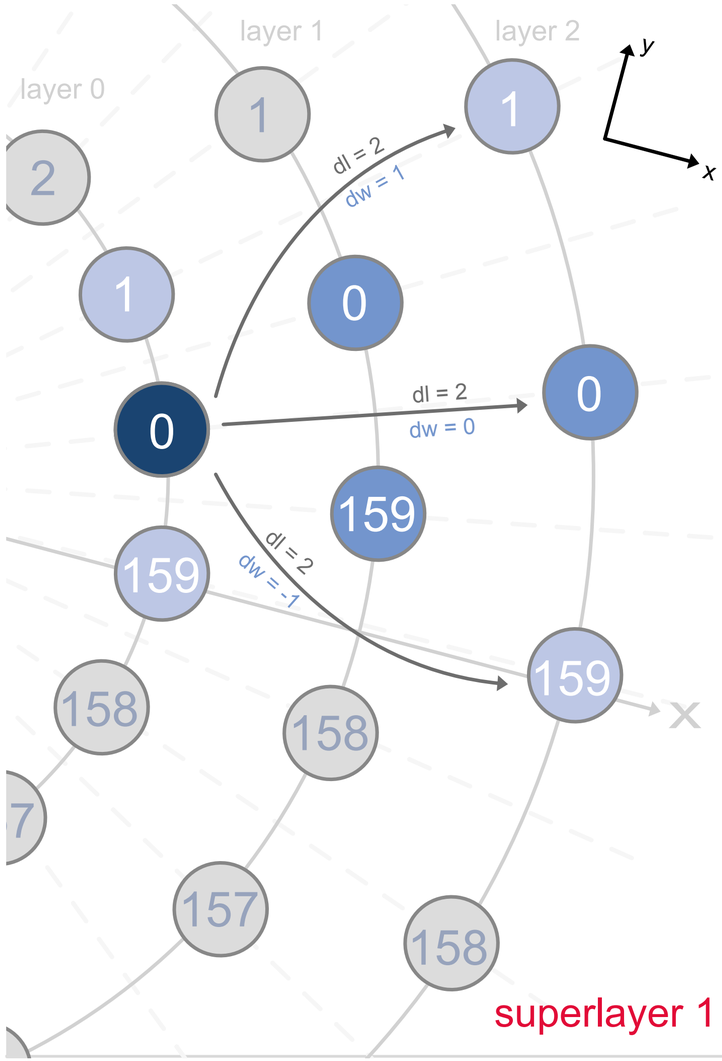
- same layer connections with 2 nearest neighbour wires (if hit)

Graph Building



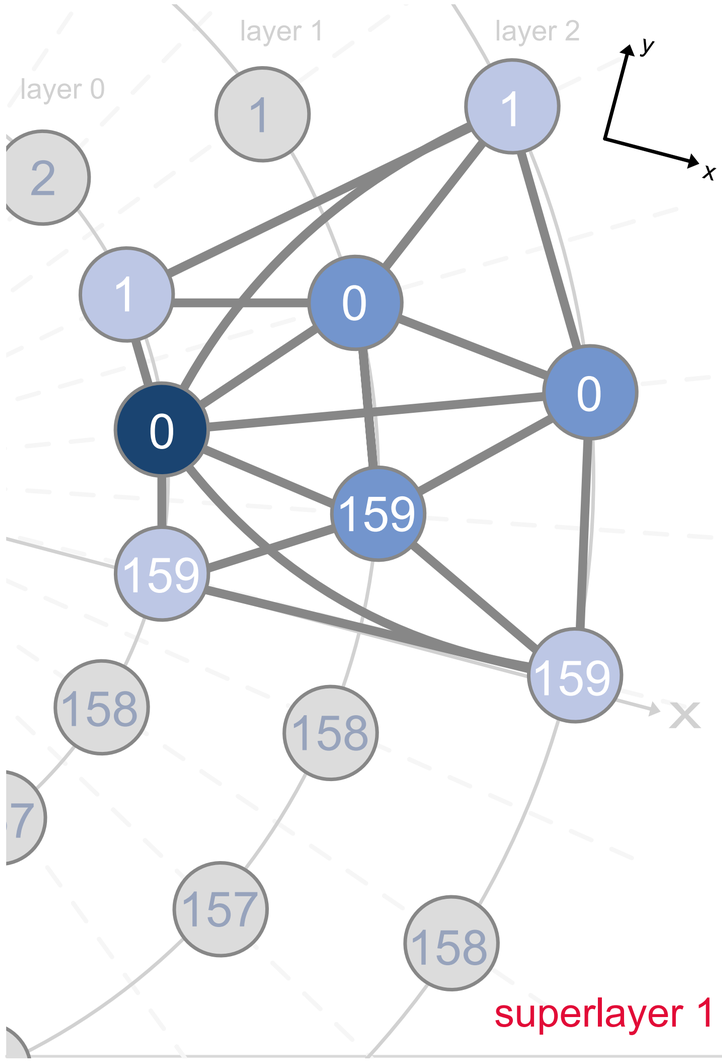
- same layer connections with 2 nearest neighbour wires (if hit)
- next layer connections with 2 nearest neighbours

Graph Building



- same layer connections with 2 nearest neighbours
- next layer connections with 2 nearest neighbours
- next to next layer connections with 3 nearest neighbours

Graph Building



- same layer connections with 2 nearest neighbours
- next layer connections with 2 nearest neighbours
- next to next layer connections with 3 nearest neighbours
- build all possible (allowed) connections

Graph Building: Features

node attributes

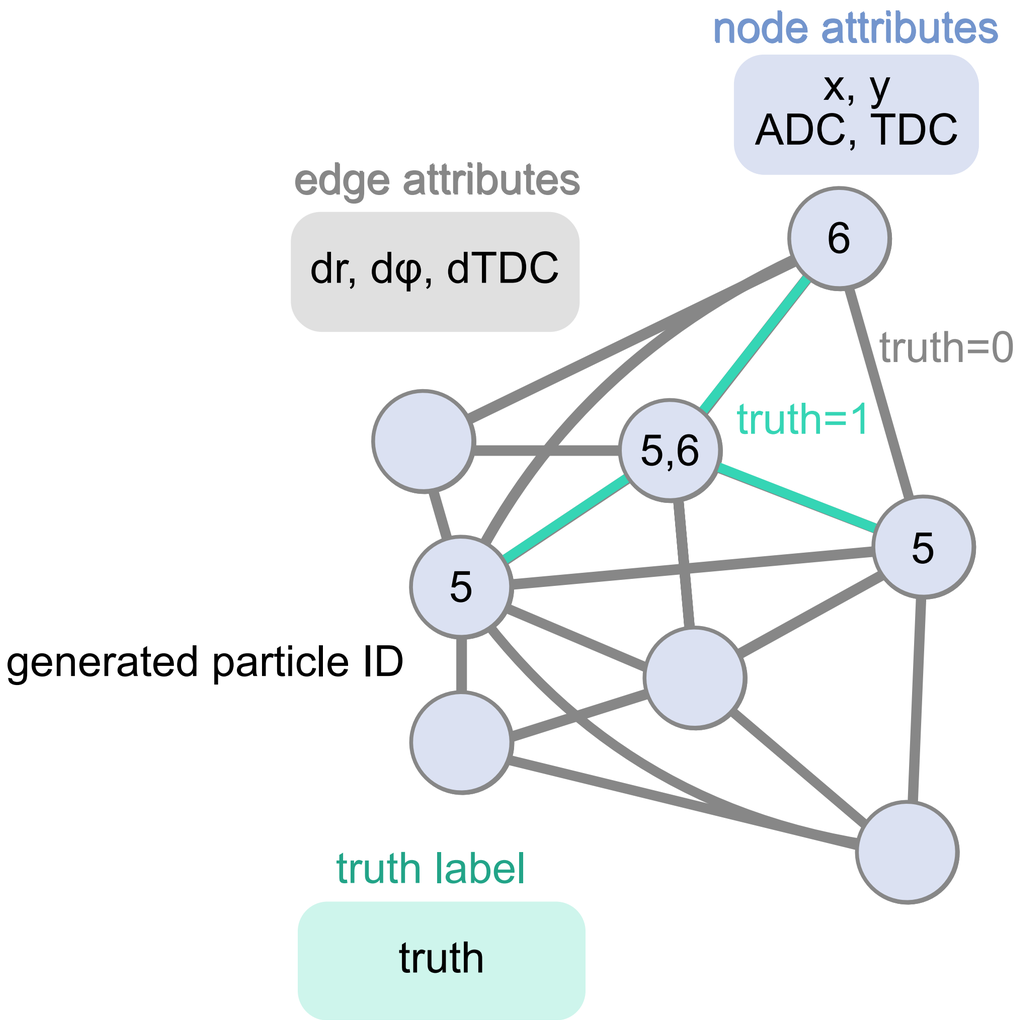
- wire coordinates (x, y)
- ADC (energy) and TDC (time)

edge attributes

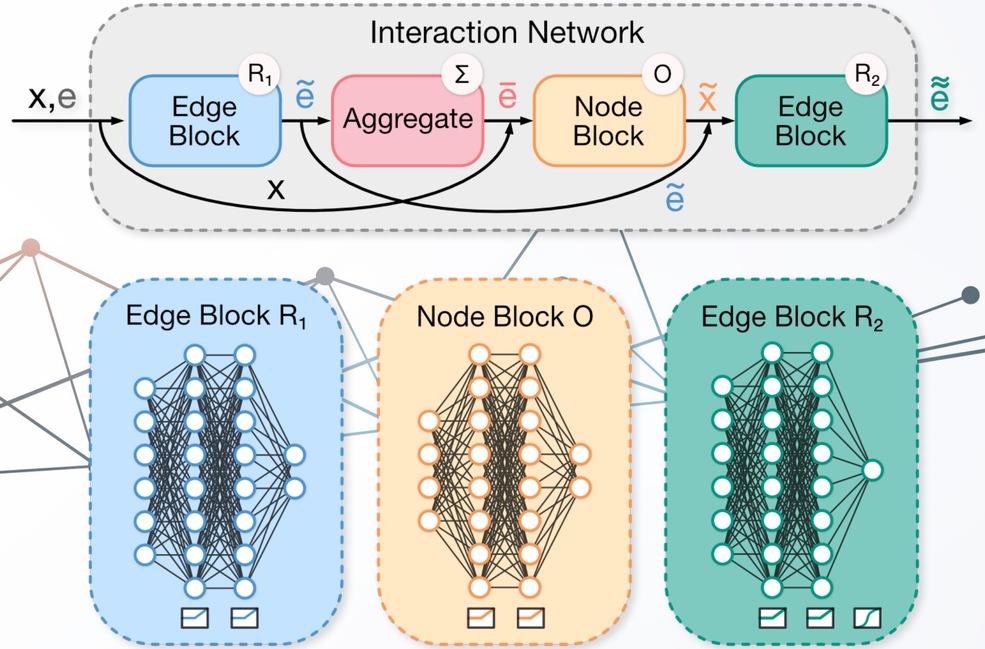
- dr, dφ, dTDC

truth labels

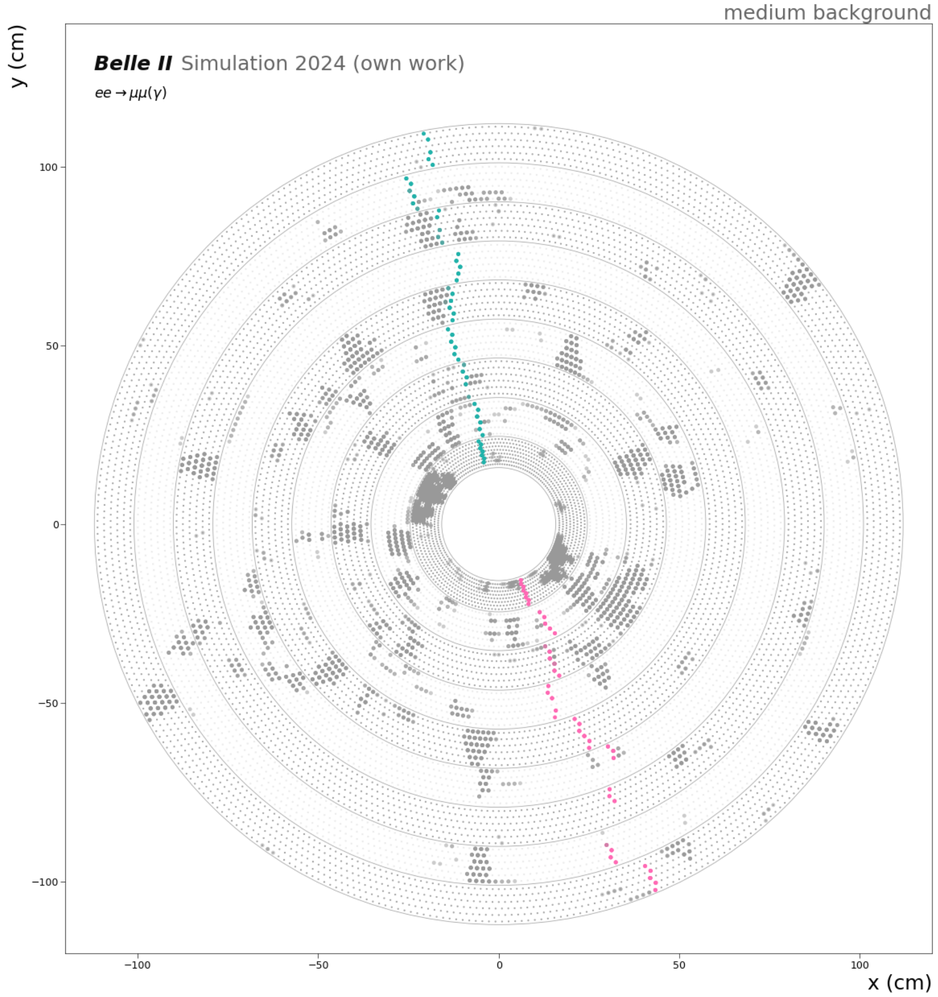
- truth

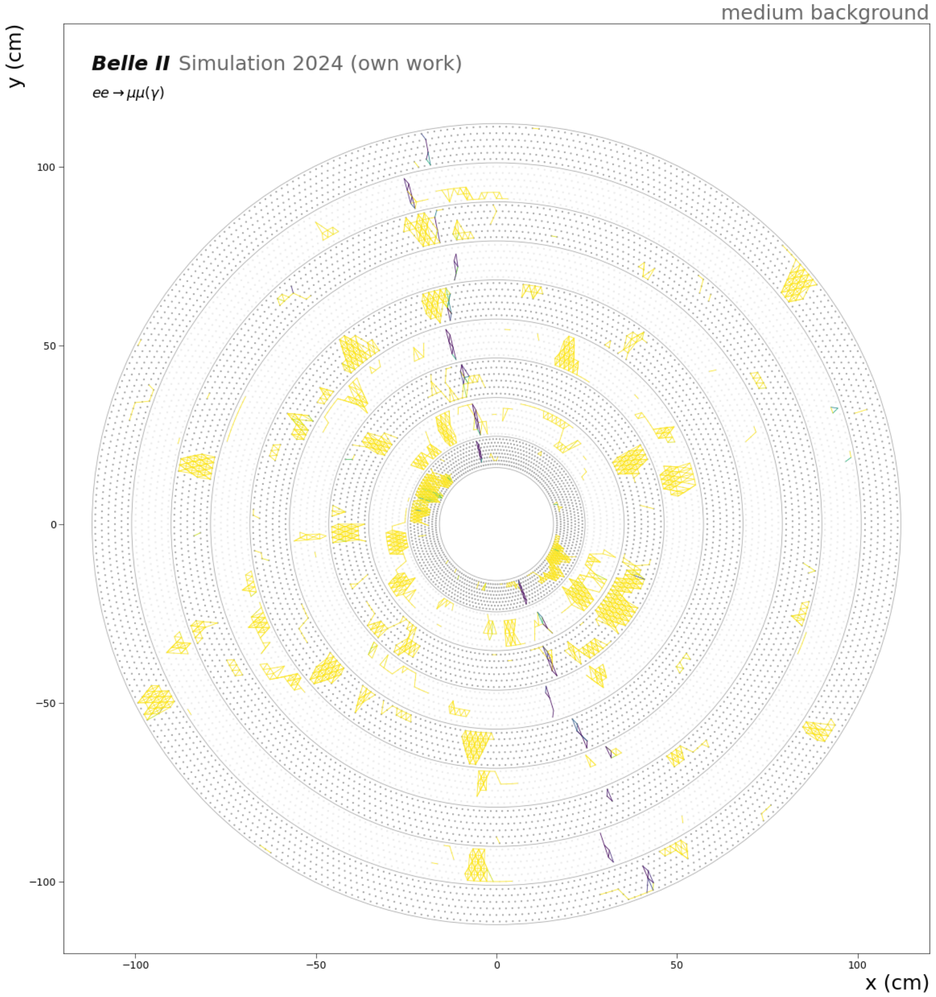


Hit Cleanup

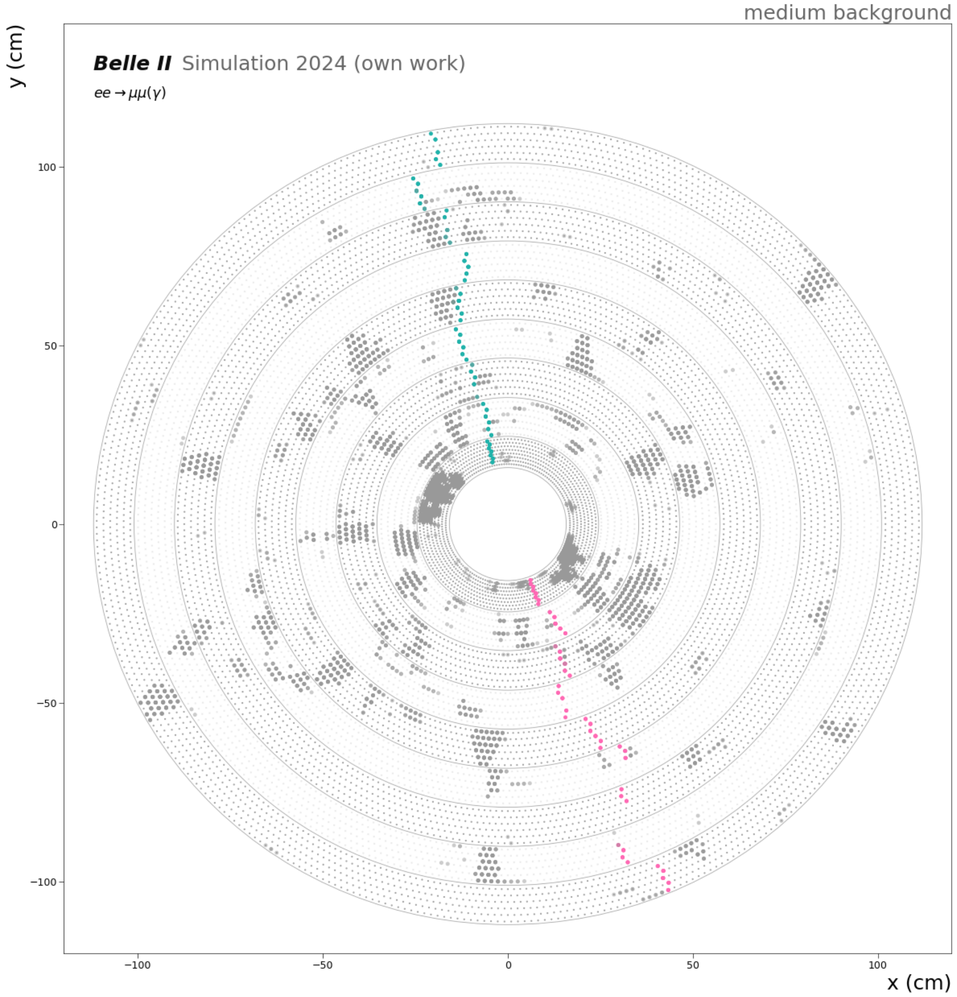

Interaction Network
x,e
Edge Block
R1
ẽ
Aggregate
Σ
ē
Node Block
O
x̃
Edge Block
R2
ẽ
x
ẽ
Edge Block R1
Node Block O
Edge Block R2

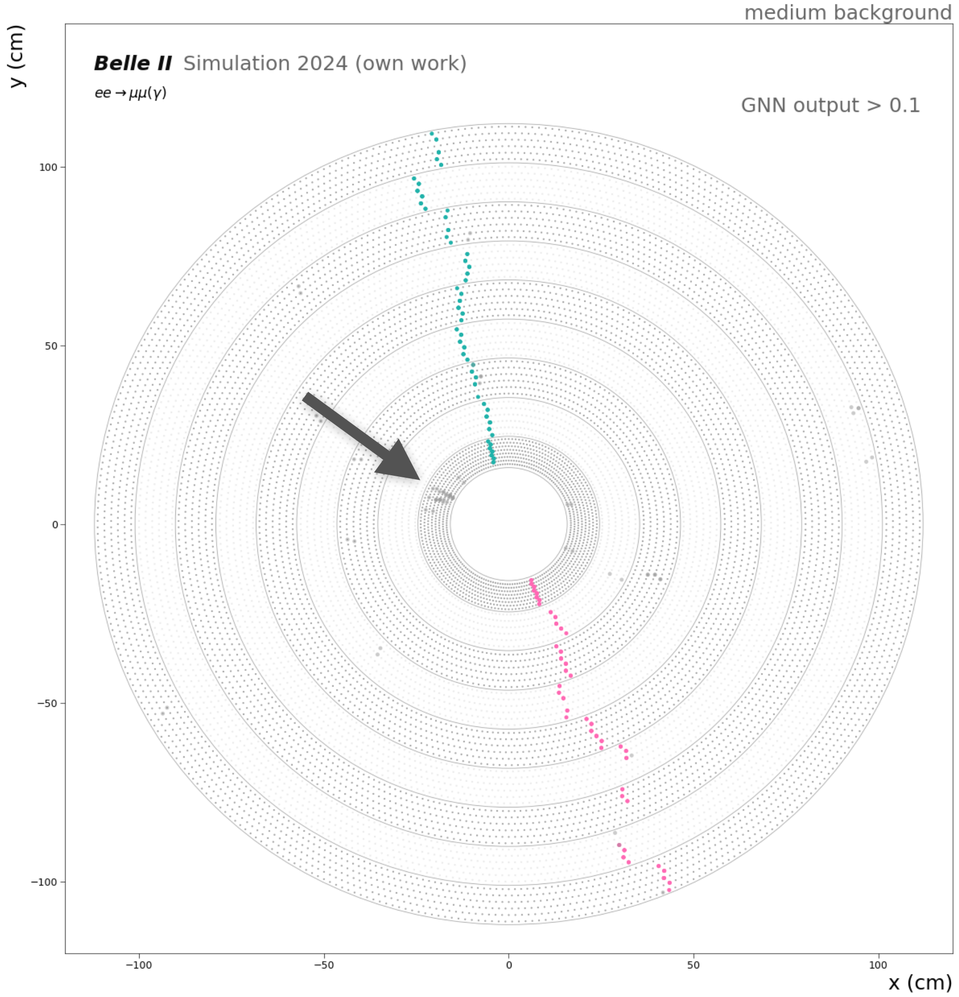
Graph Neural Network Output



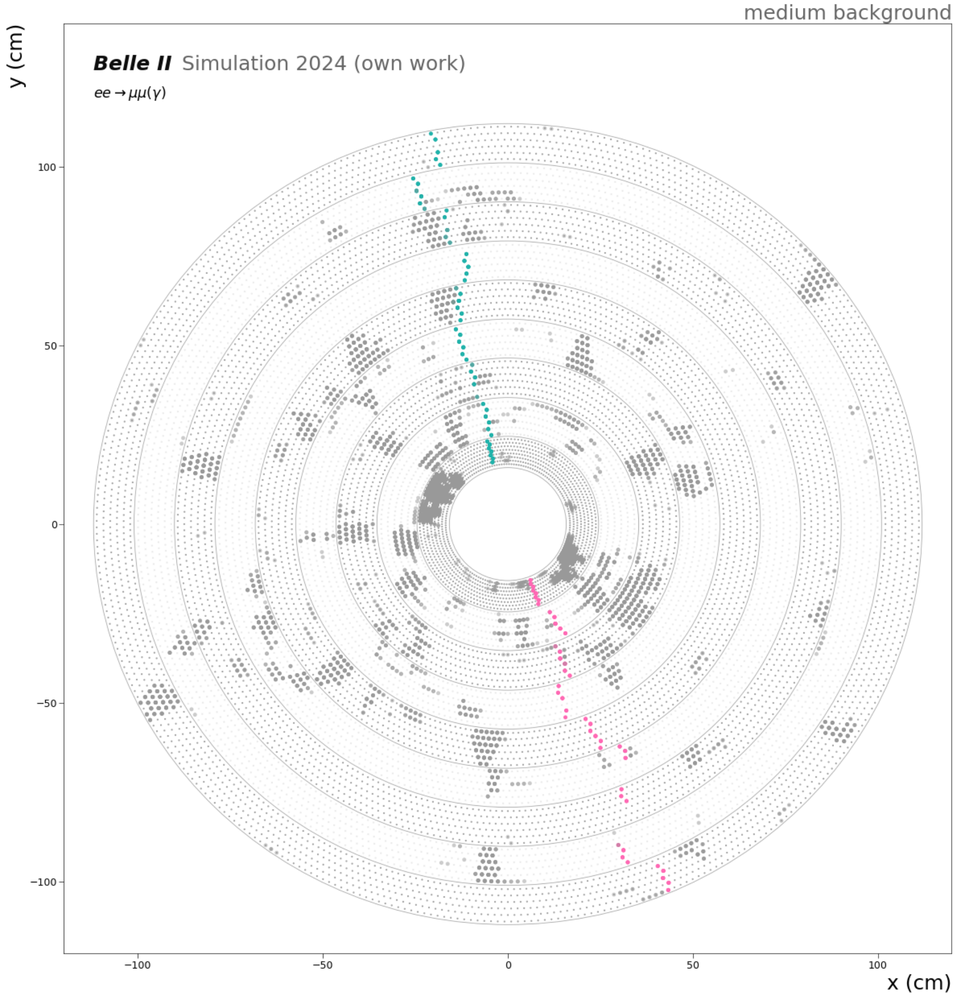


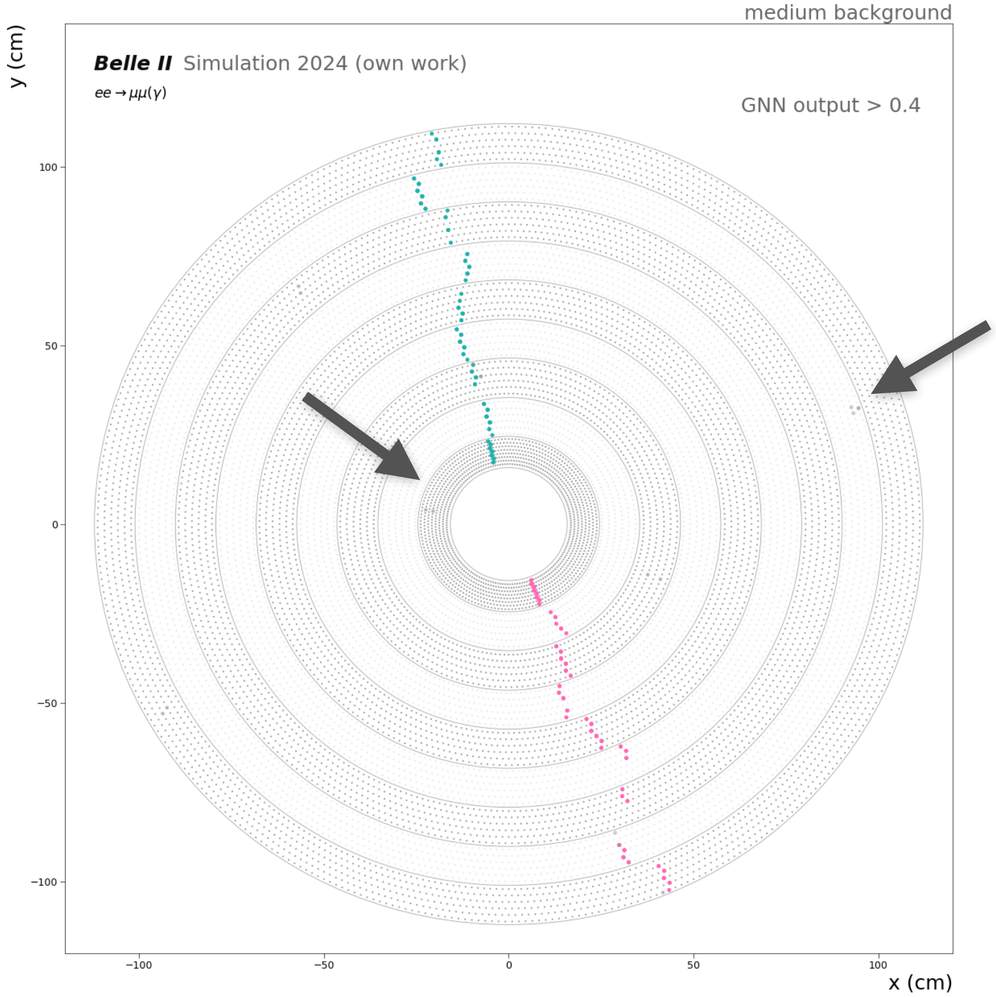
Graph Neural Network Output



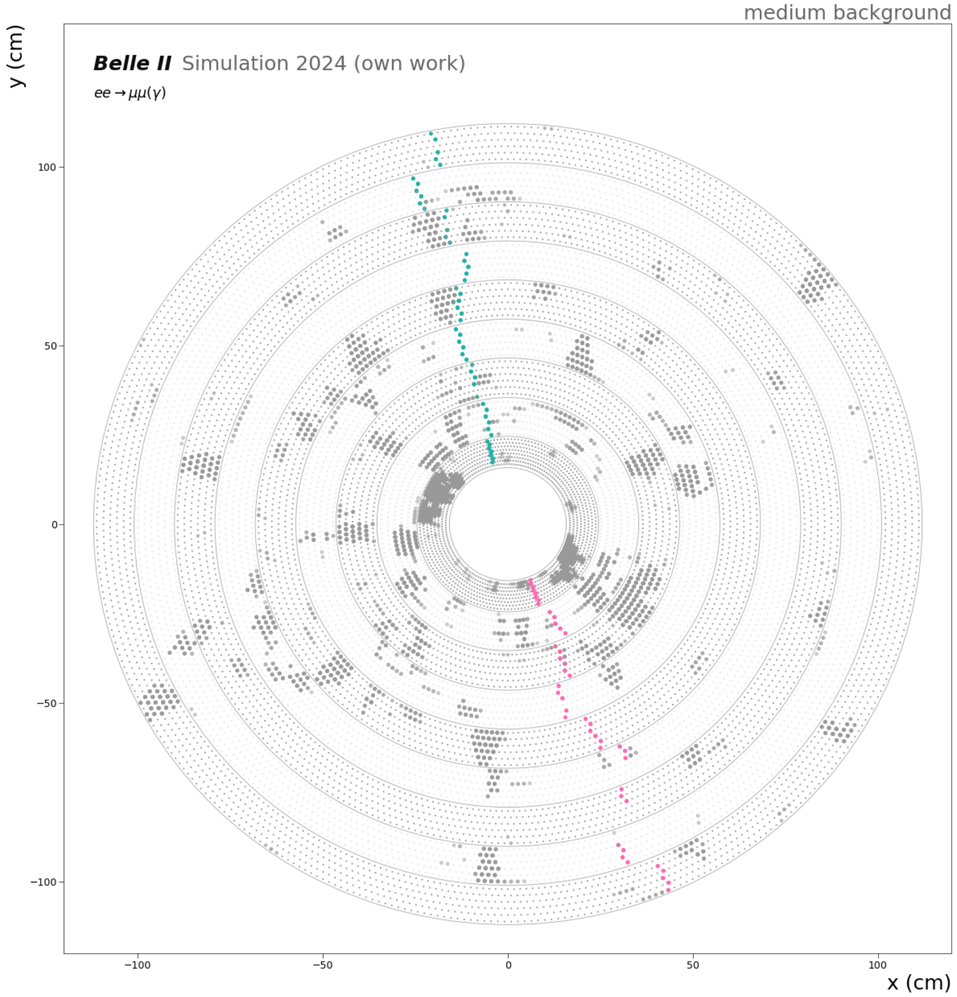


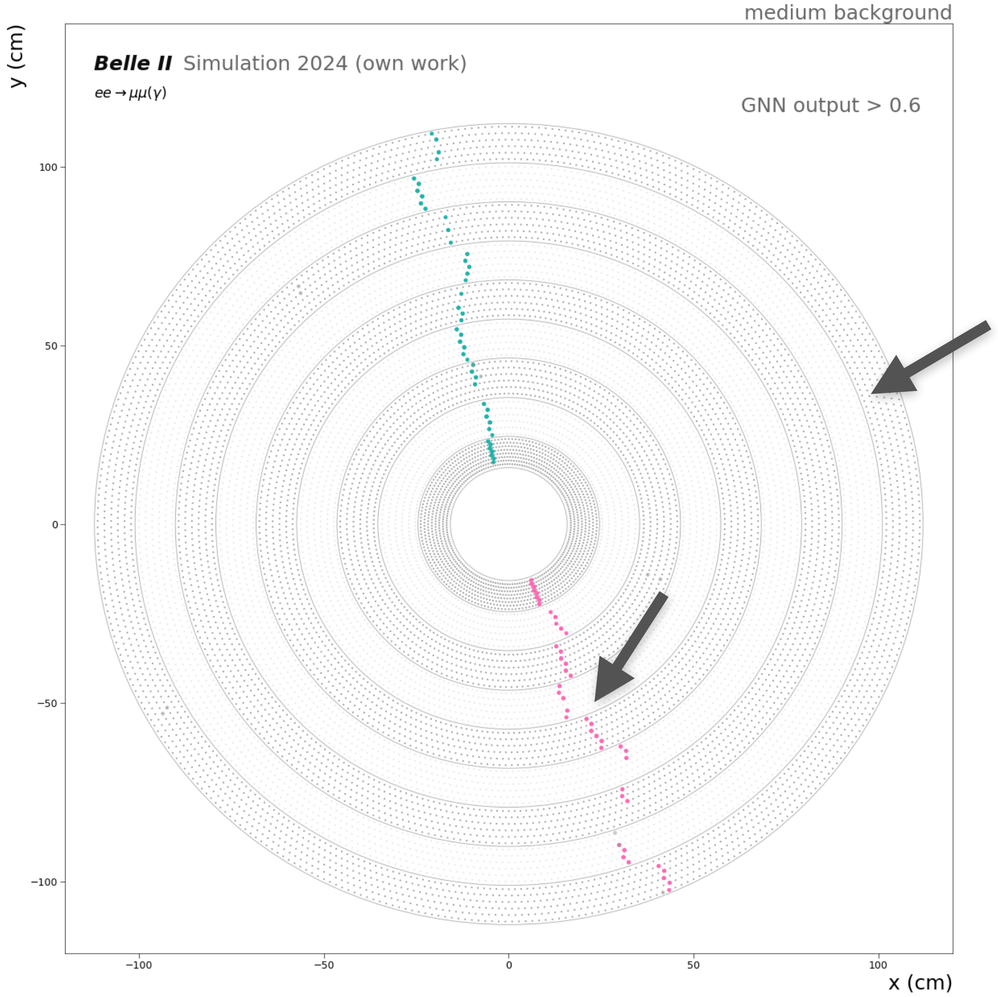
Graph Neural Network Output



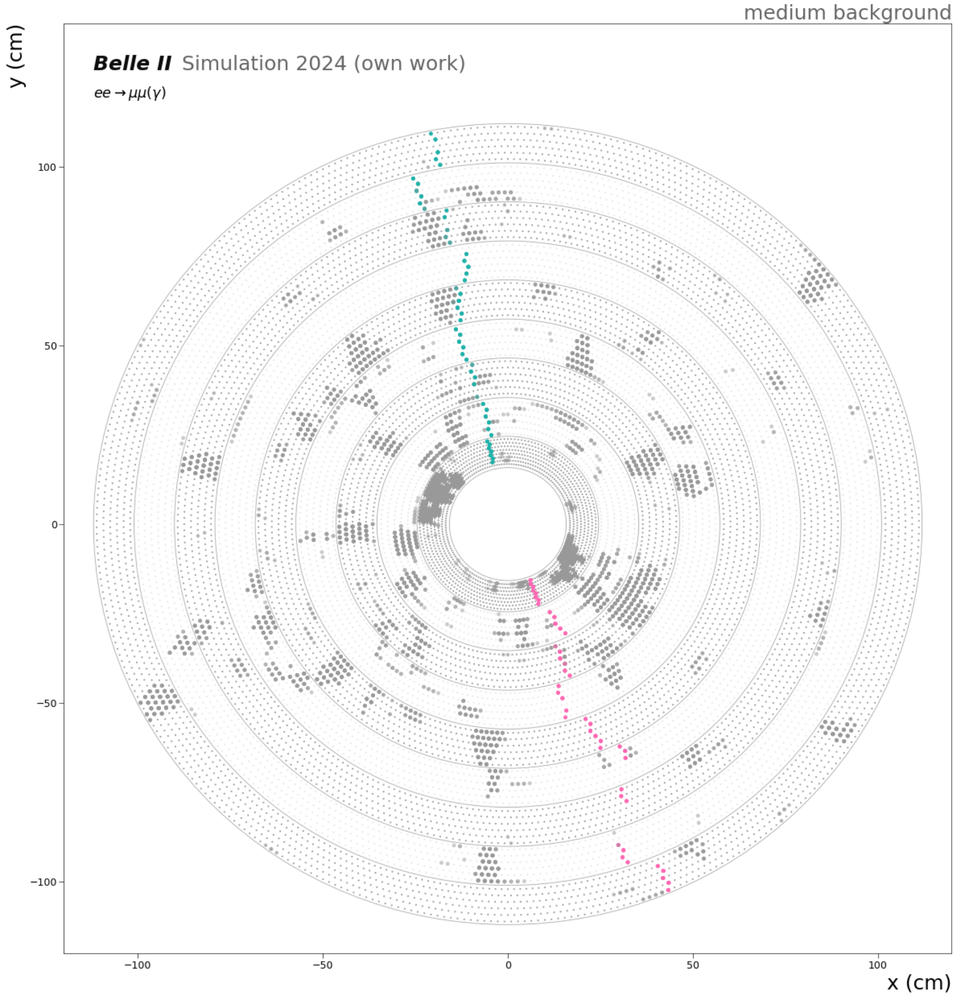


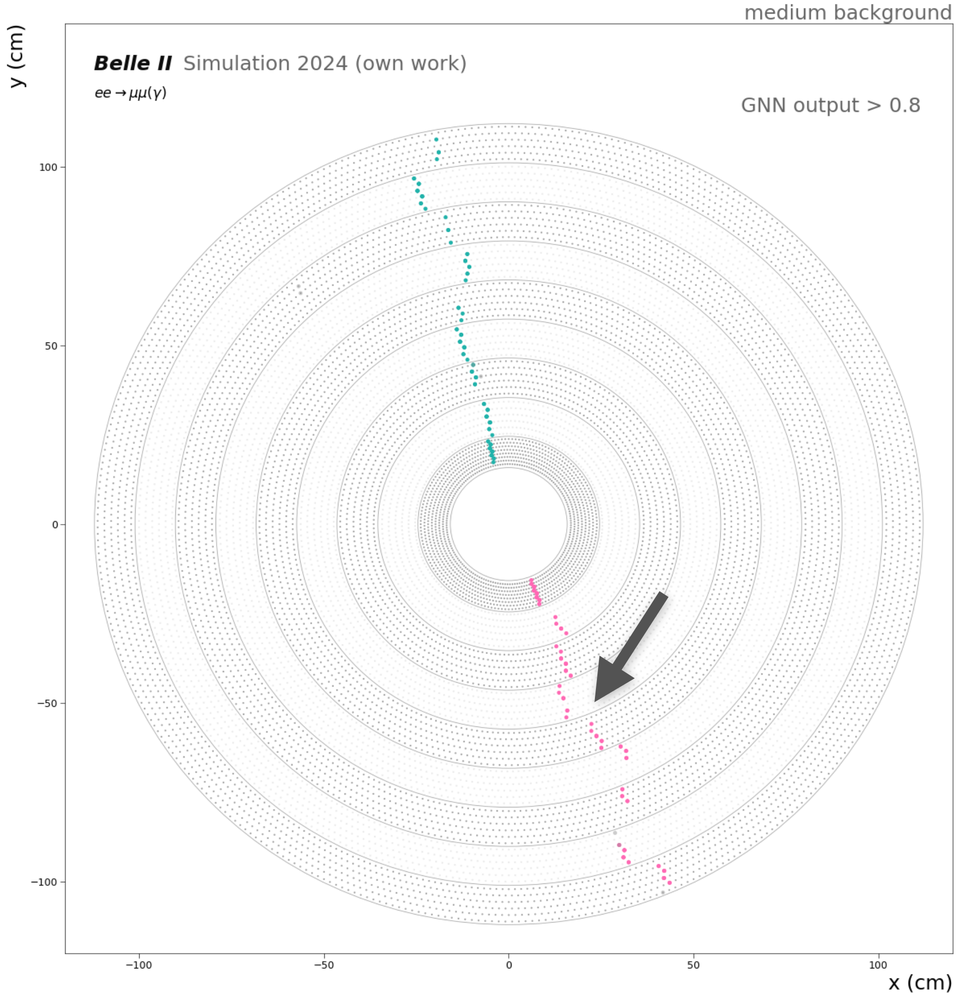
Graph Neural Network Output



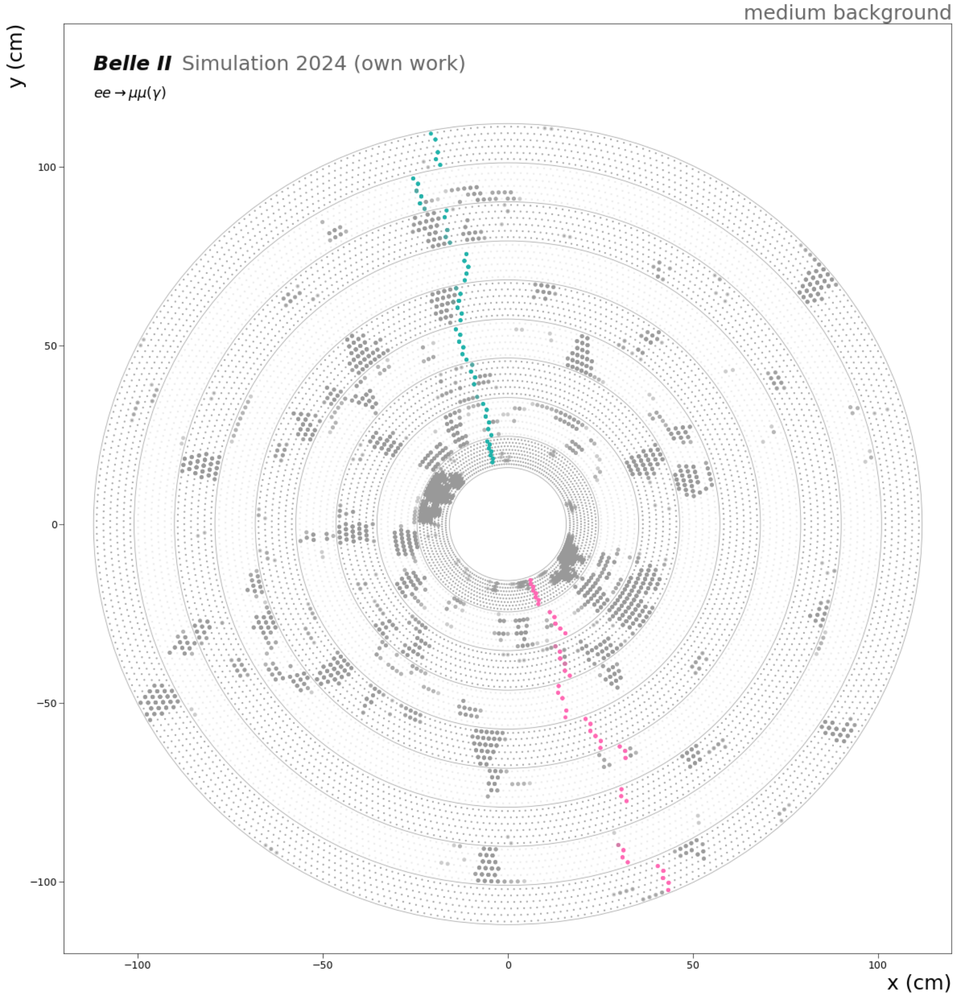


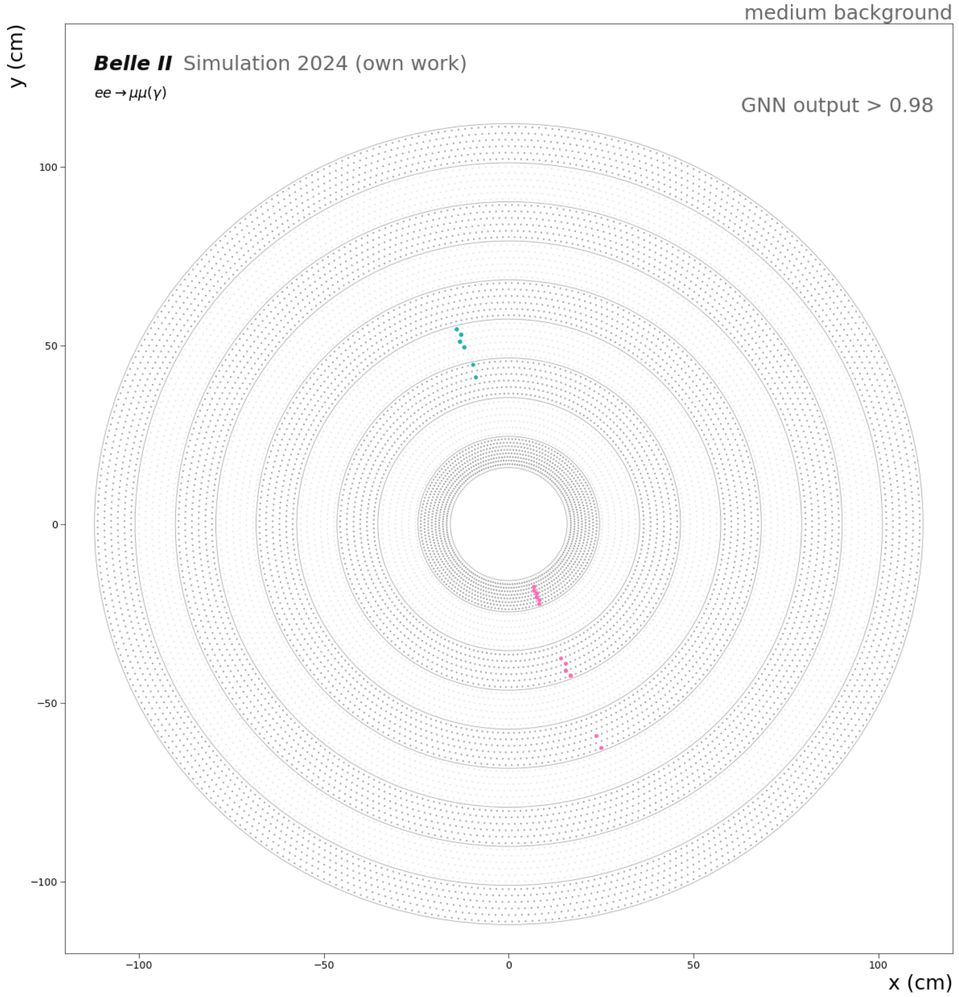
Graph Neural Network Output



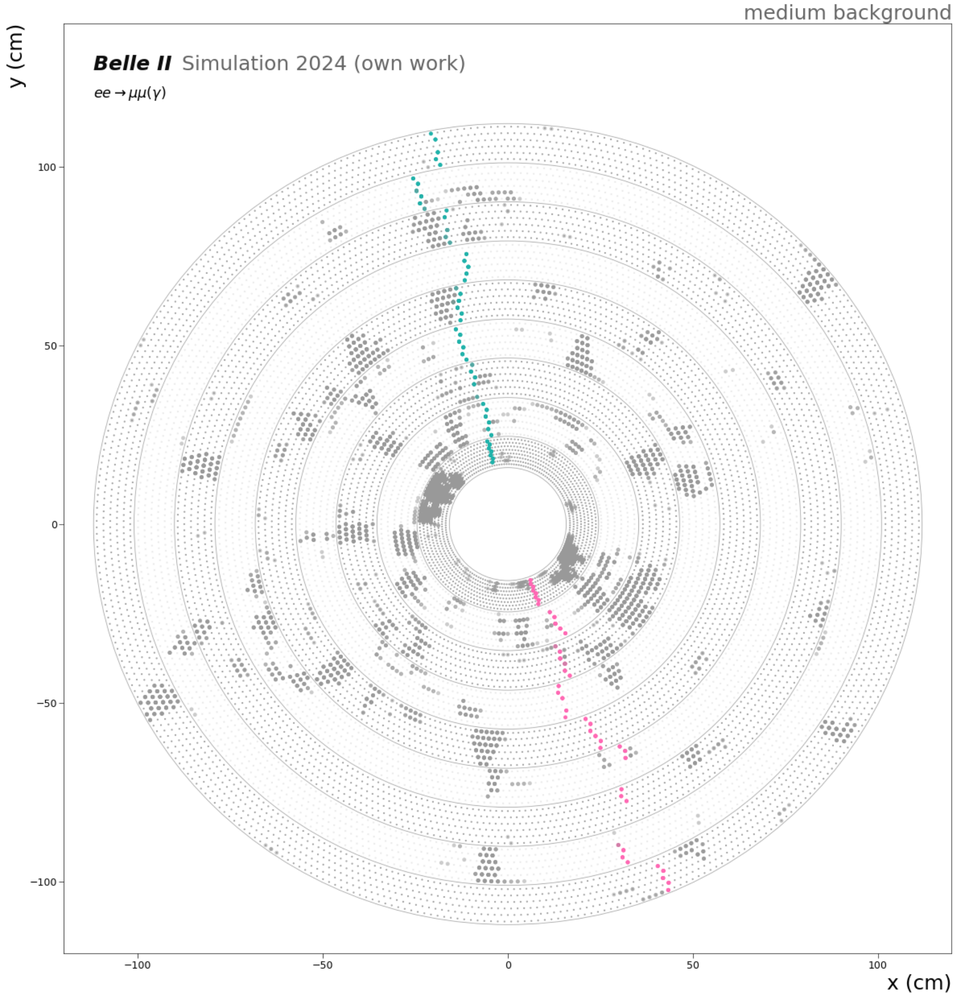


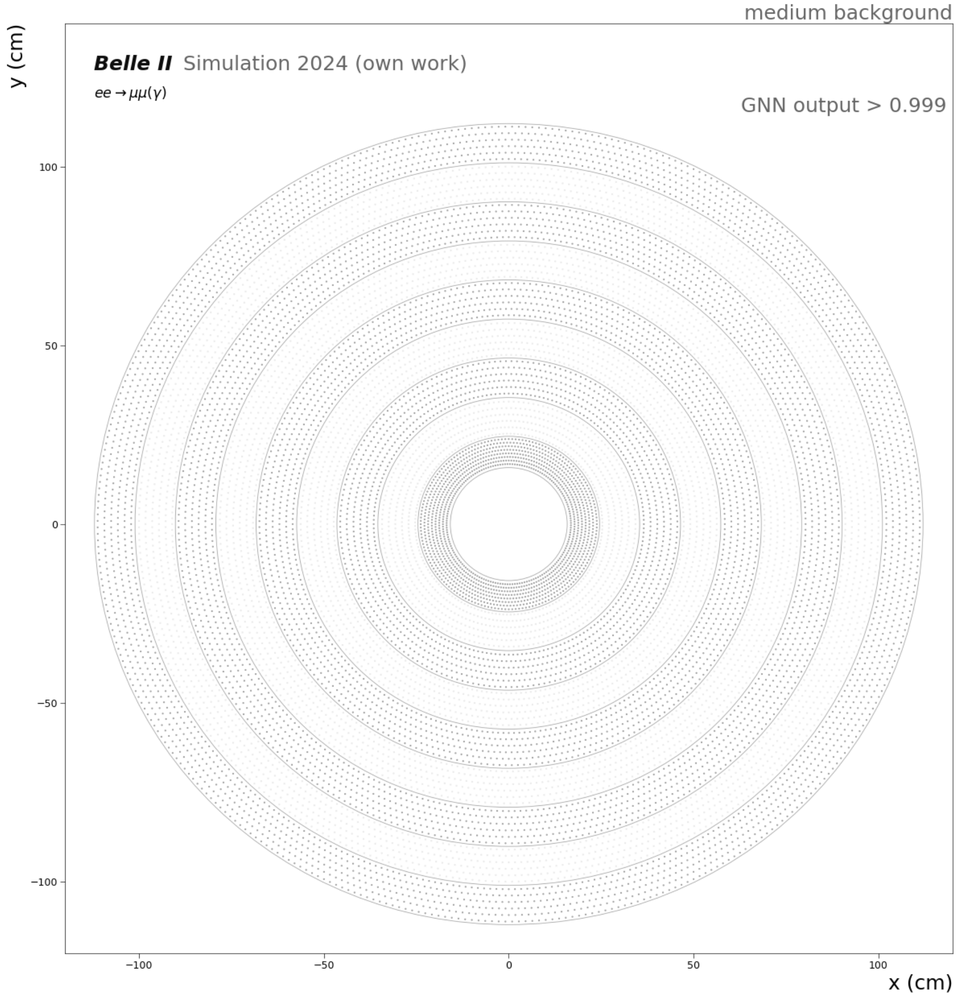
Graph Neural Network Output



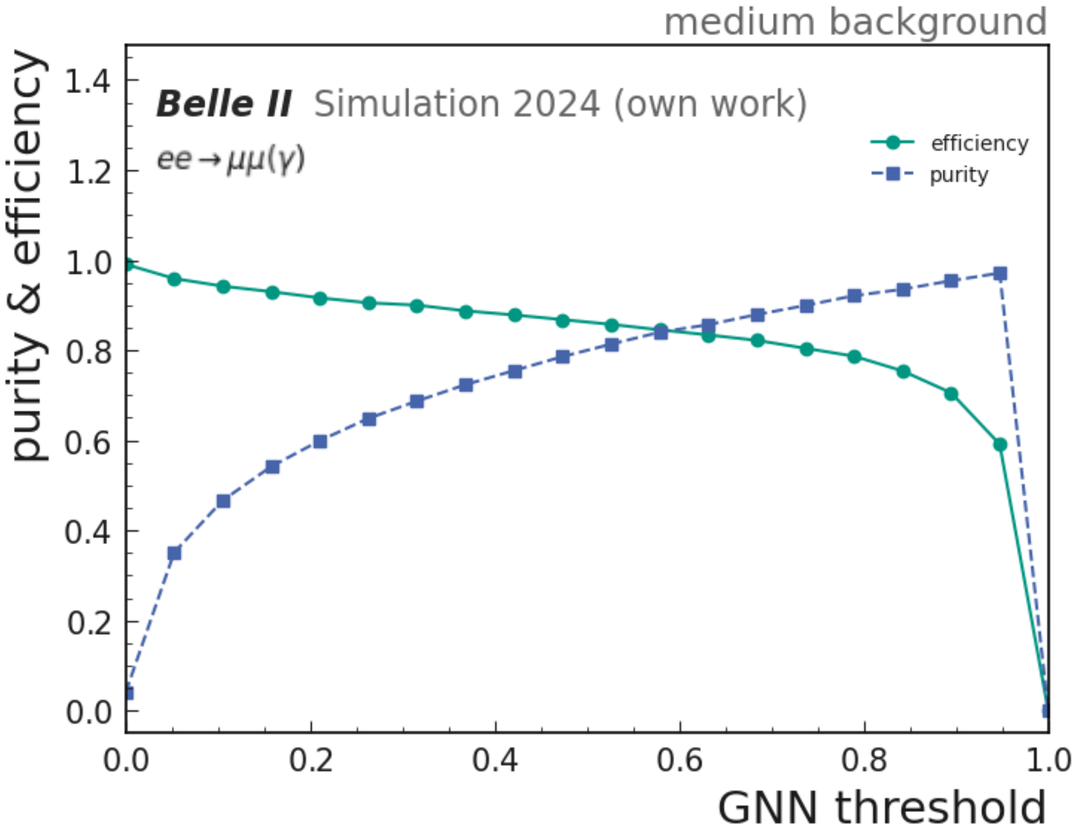


Graph Neural Network Output





Hit Cleanup Results



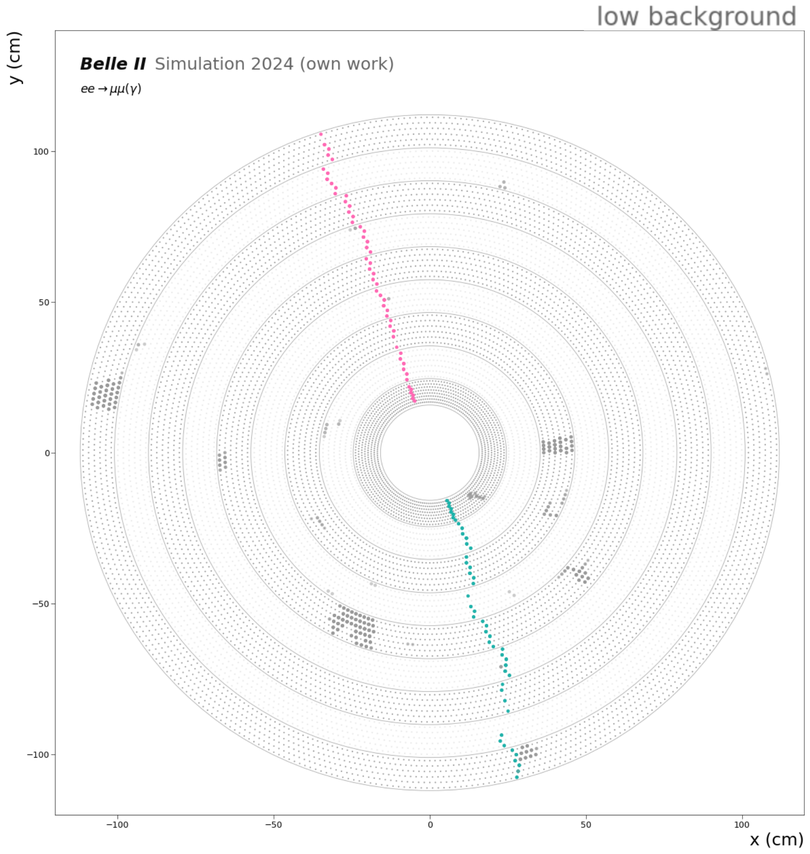
$$\text{efficiency} = \frac{\text{kept signal hits}}{\text{all signal hits}}$$

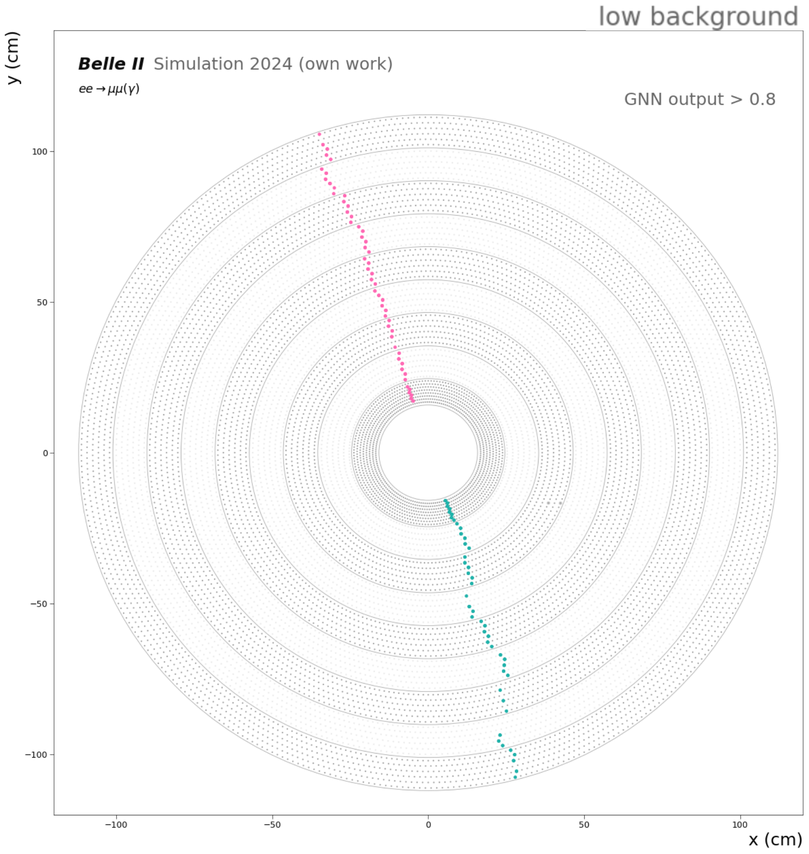
„how much of signal hits kept“

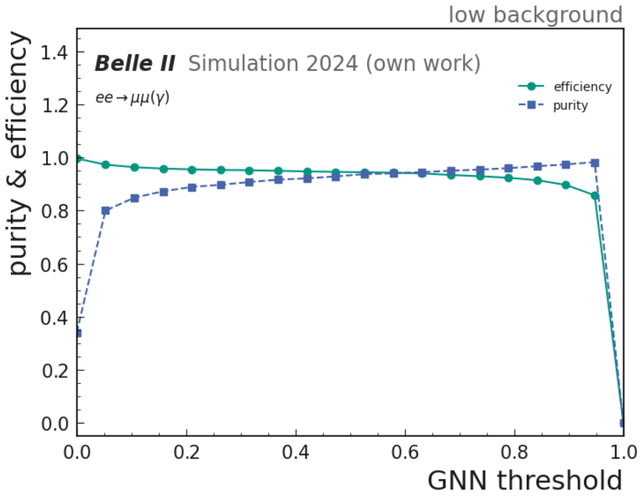
$$\text{purity} = \frac{\text{kept signal hits}}{\text{all kept hits}}$$

„how much of kept hits are signal“

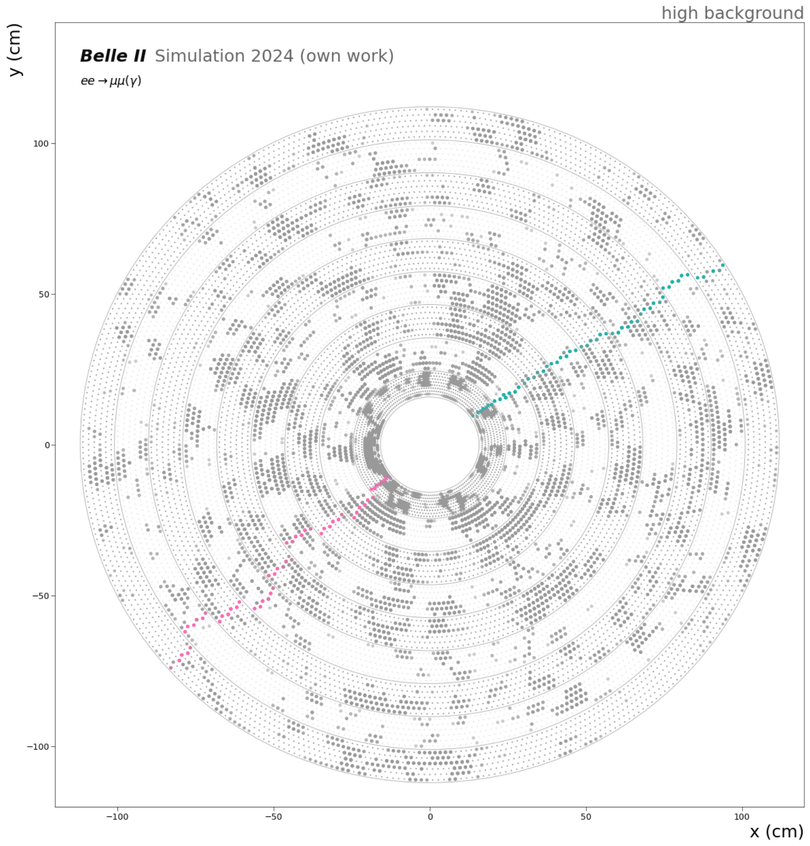
Hit Cleanup Example Low Background

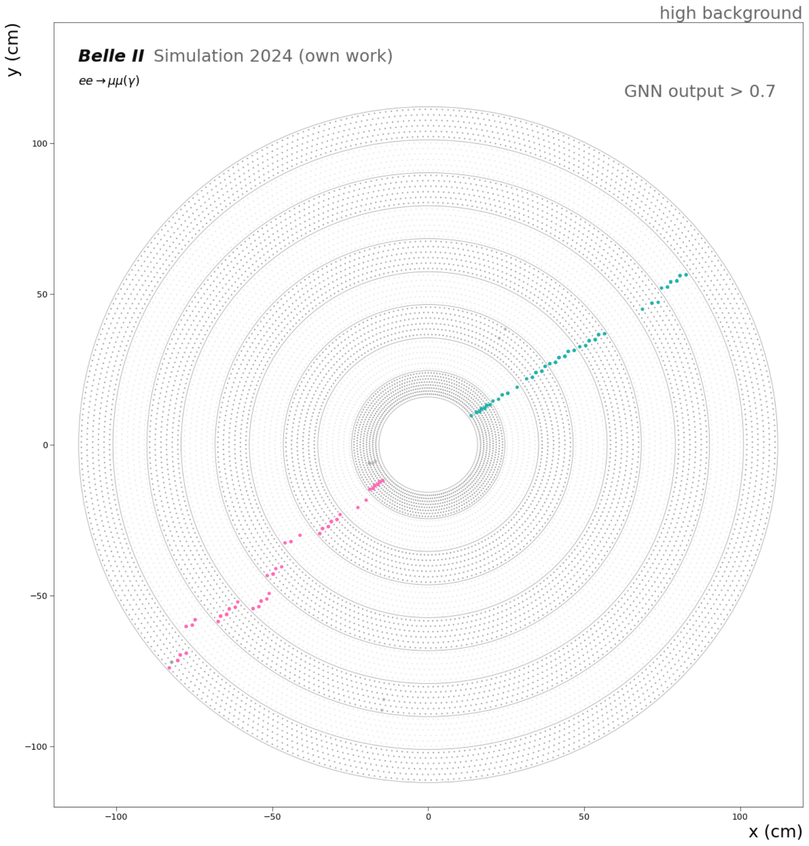


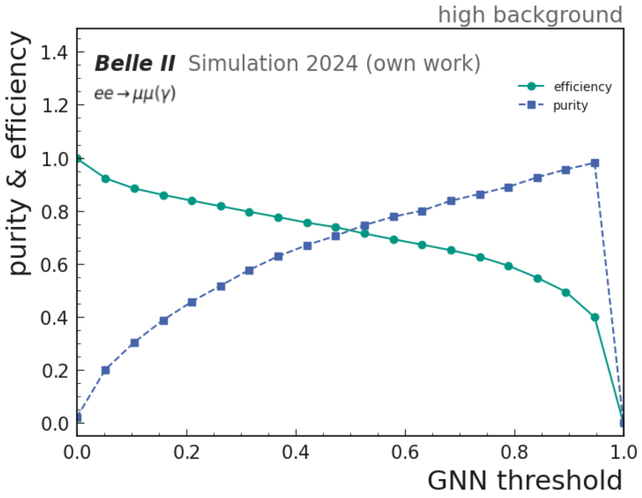




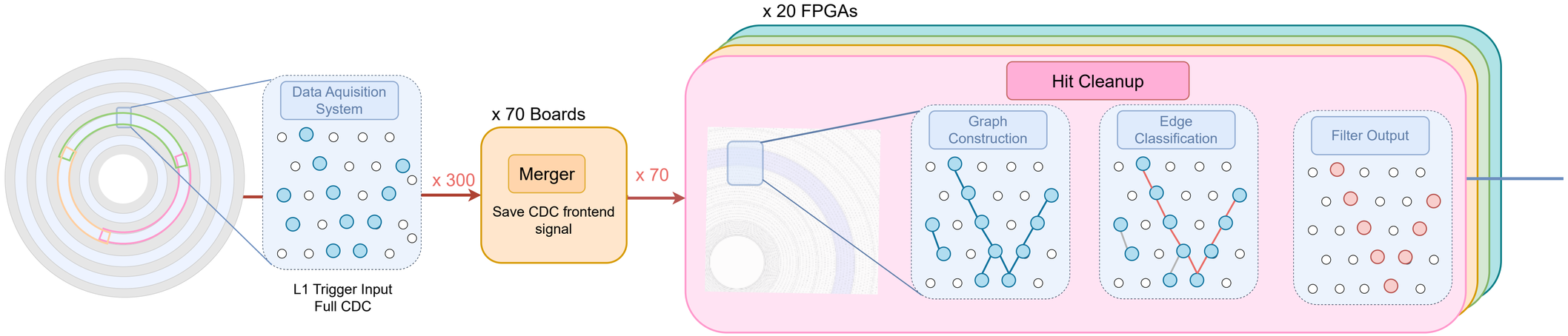
Hit Cleanup Example High Background







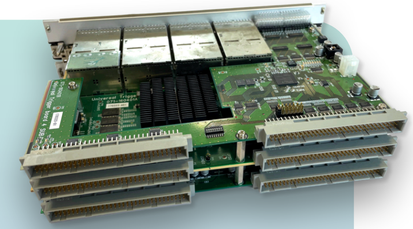
Next steps towards implementation



CDC segmentation

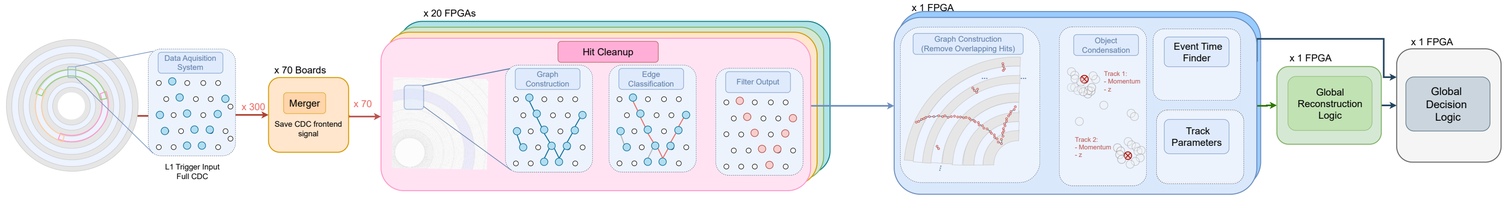
- segmentation by superlayer & angle on 20 FPGA boards
- handle overlaps
- handle merging

FPGA implementation

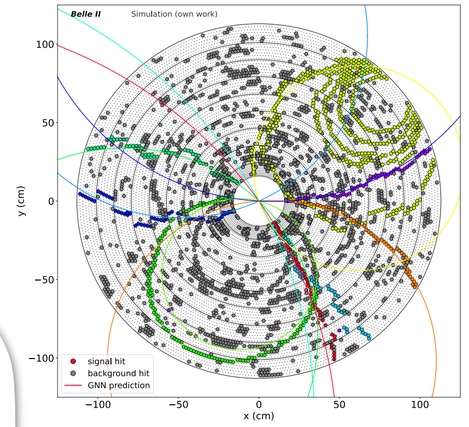


- size reduction studies (quantization, graph size compression, pruning, etc.)
- implementation (hls4ml)

Summary

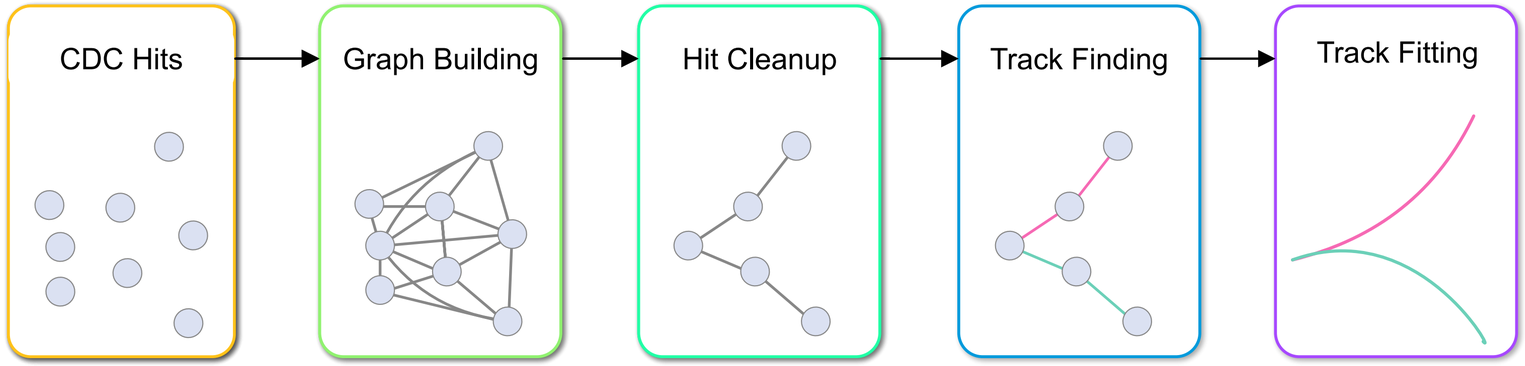


- tracking/ trigger is challenging for higher backgrounds
- hit clean up with GNNs shows promising results

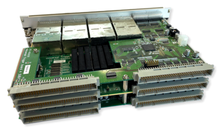


plot from Lea Reuter

GNN-based Tracking Pipeline



FPGA *



arXiv:2307.07289

offline solution exists, but with very large networks

Next talk: Marc Neu

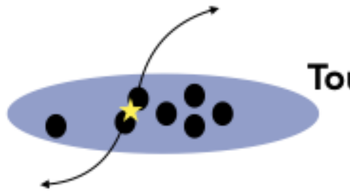
* Collaboration with ITIV (Department of Electrical Engineering and Information Technology at KIT)

Backup

Belle II backgrounds

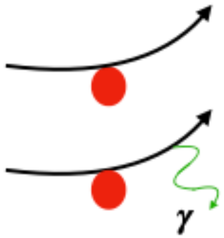
Background Processes

Particle Scattering



Touschek Scattering

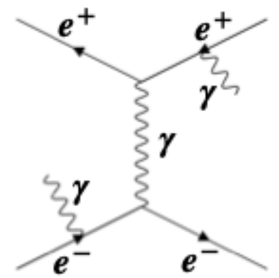
$$\propto N_{particles} \times \rho \propto I \times \frac{I}{n_b \sigma_b}$$



Coulomb Scattering + Bremsstrahlung

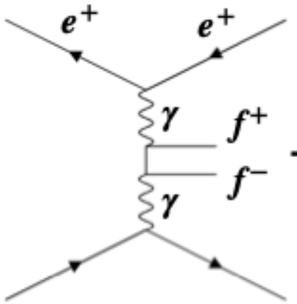
$$\propto N_{particles} \times N_{gas\ molecules} \propto P \times I \times Z_{eff}^2$$

Luminosity



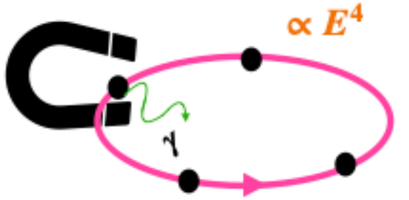
Radiative Bhabha

$\propto \mathscr{L}$



Two-photon

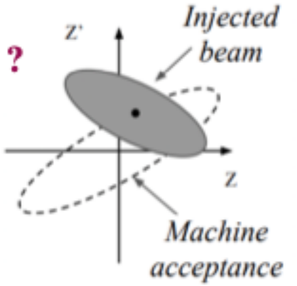
Synchrotron rad.



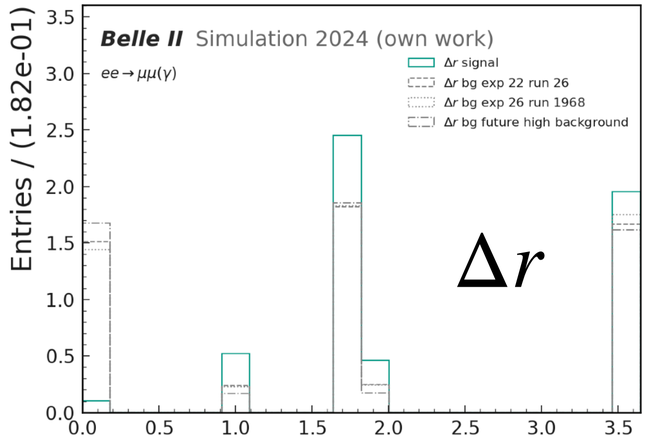
$\propto E^4$

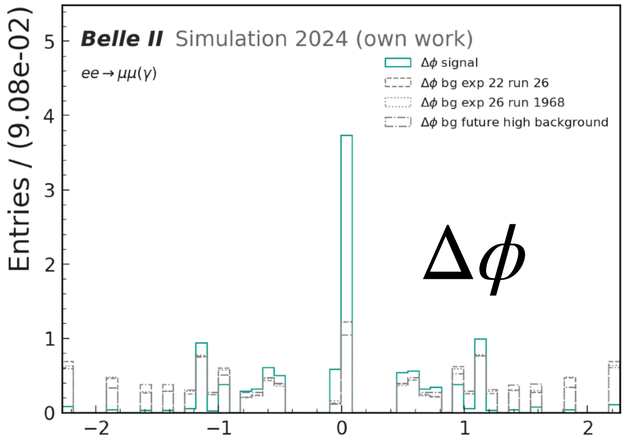
Injection bkg.

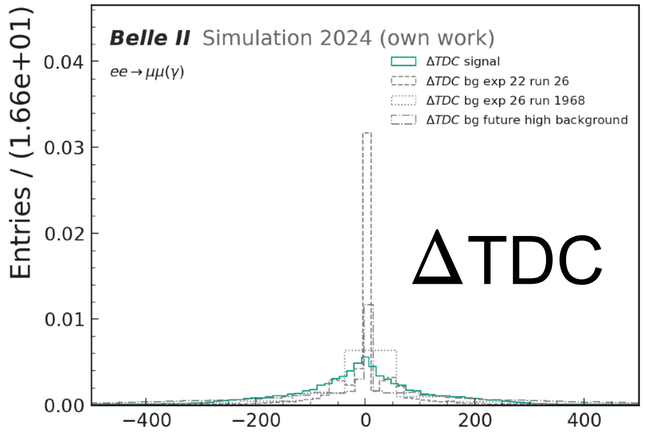
$\propto$?

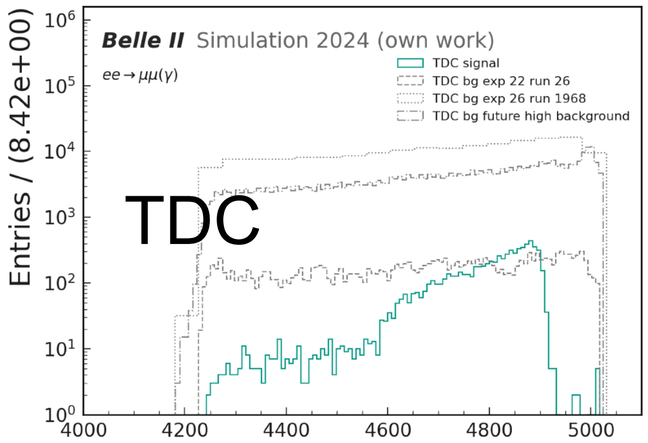


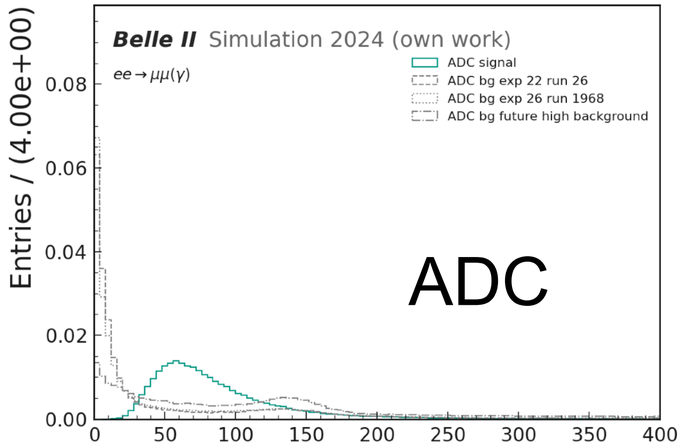
Feature Distributions



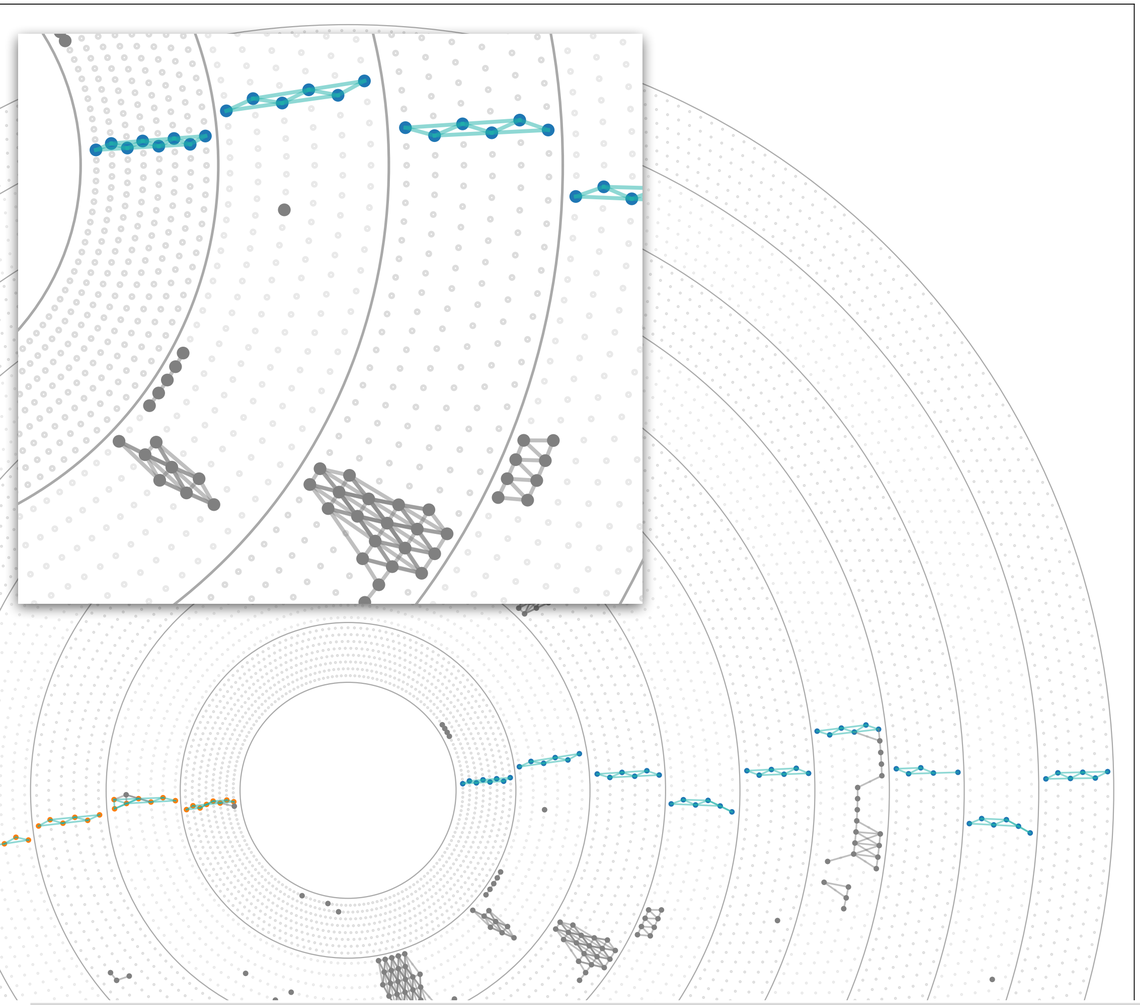








Graph Building: Results



	low bg	medium bg	high bg
% Signal Hits	29.1%	3.7%	1.6%
# Hits	250	1800	3900
# Edges	500	4200	8200

Graph Building Models

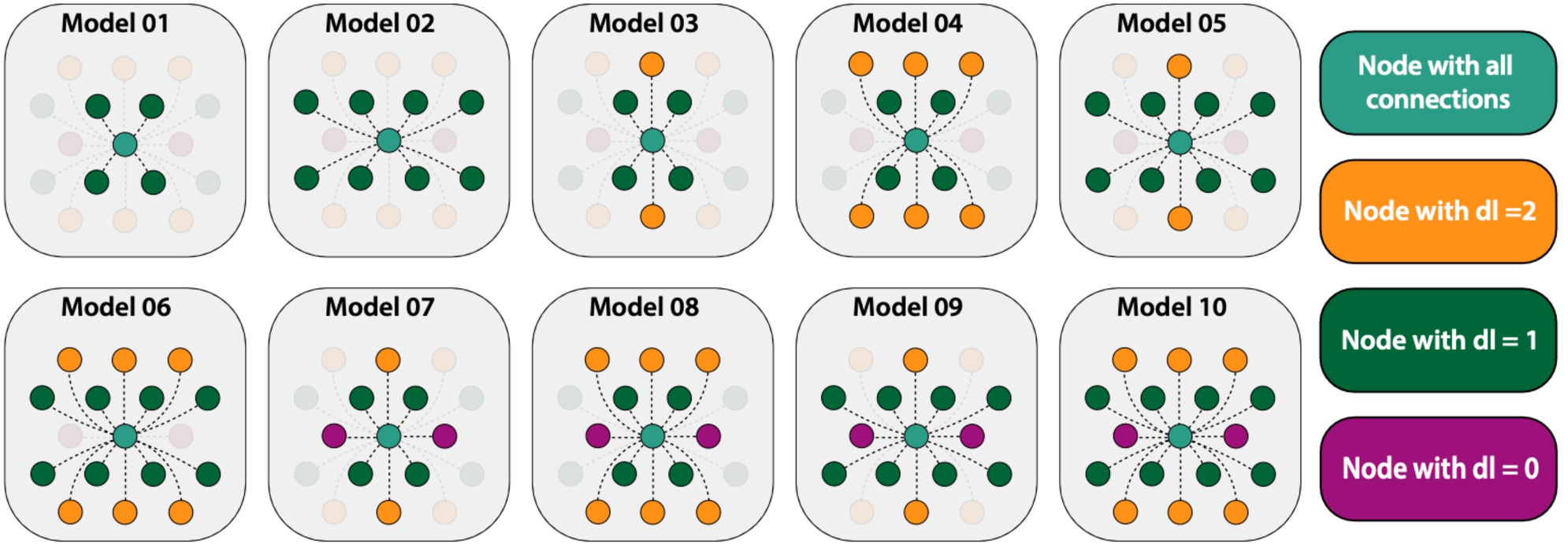
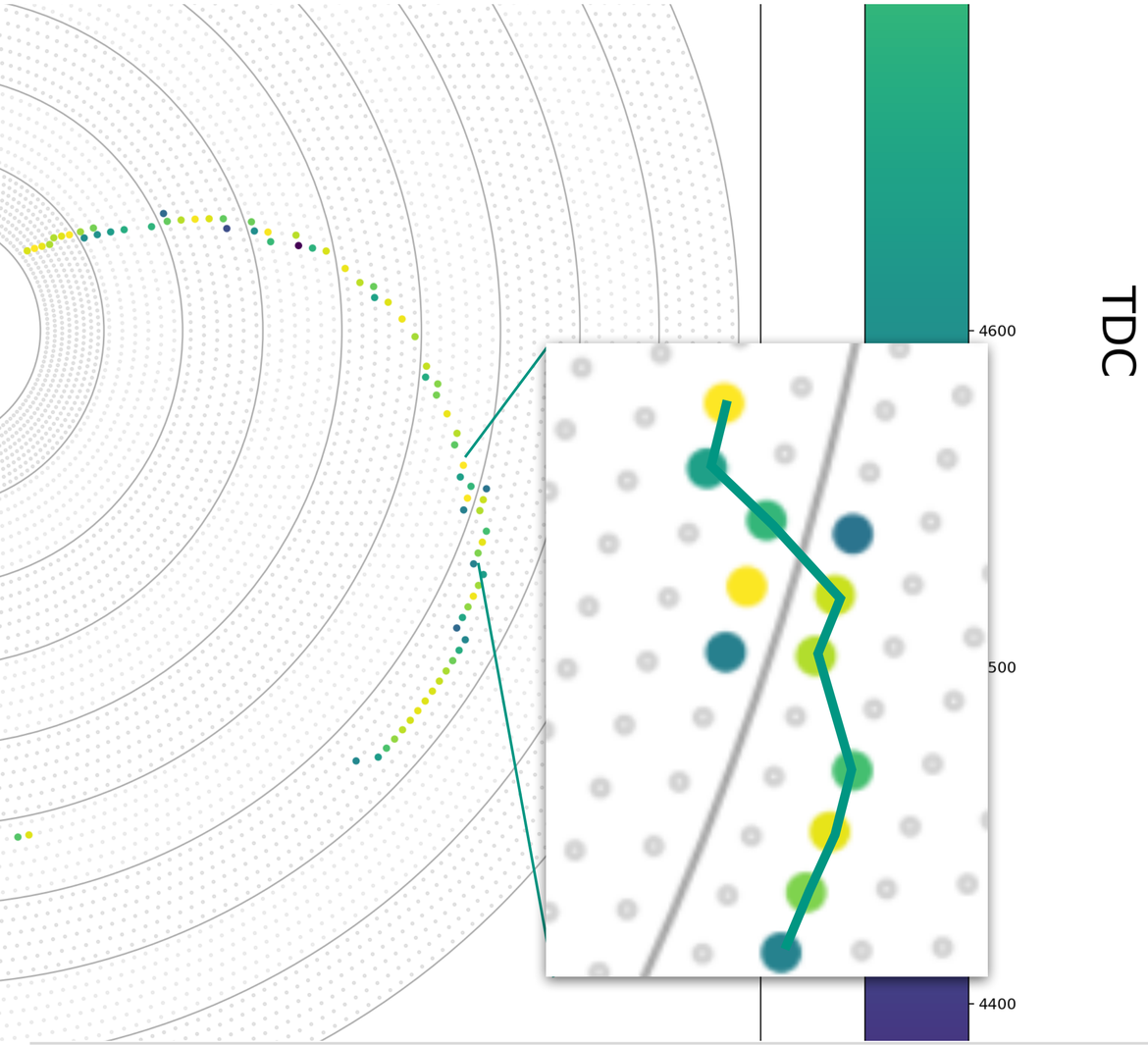


Figure 7.9.: Ten different graph-building patterns analyzed in this thesis, showing possible connections for a node in the Central Drift Chamber.

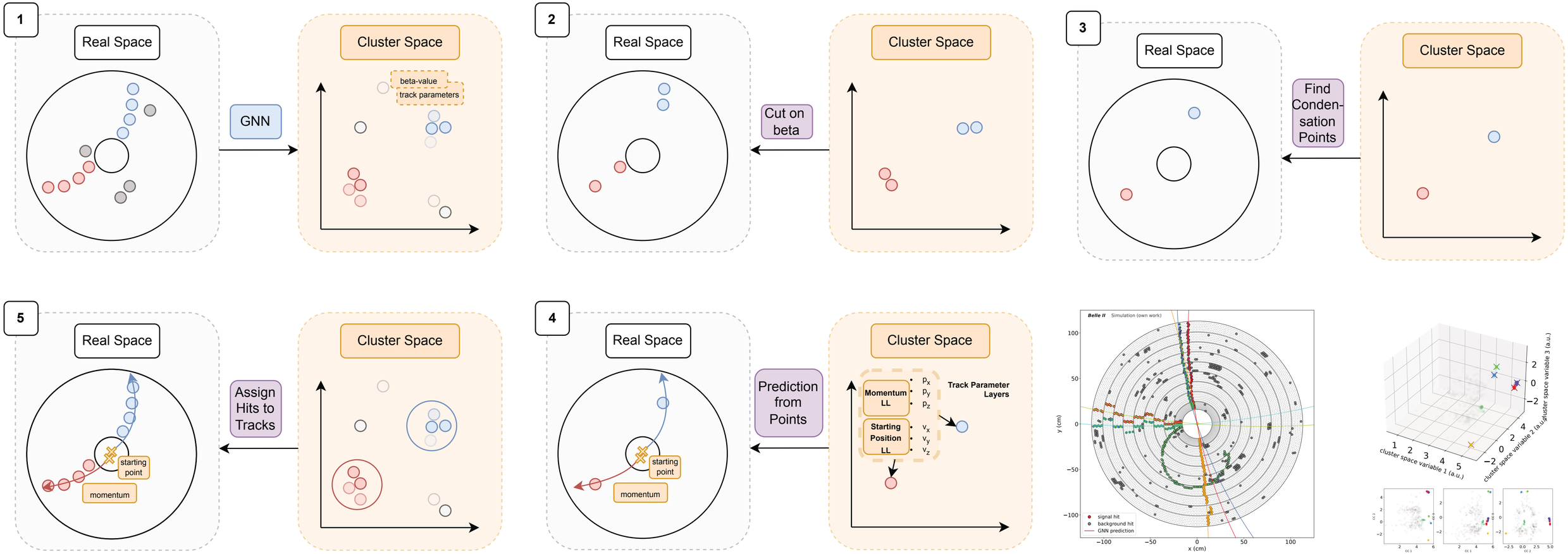
Masterthesis „Graph-Building and Input Feature Analysis for Edge Classification in the Central Drift Chamber at Belle II“ of Philipp Dorwarth, May 2023

True Hit Definition

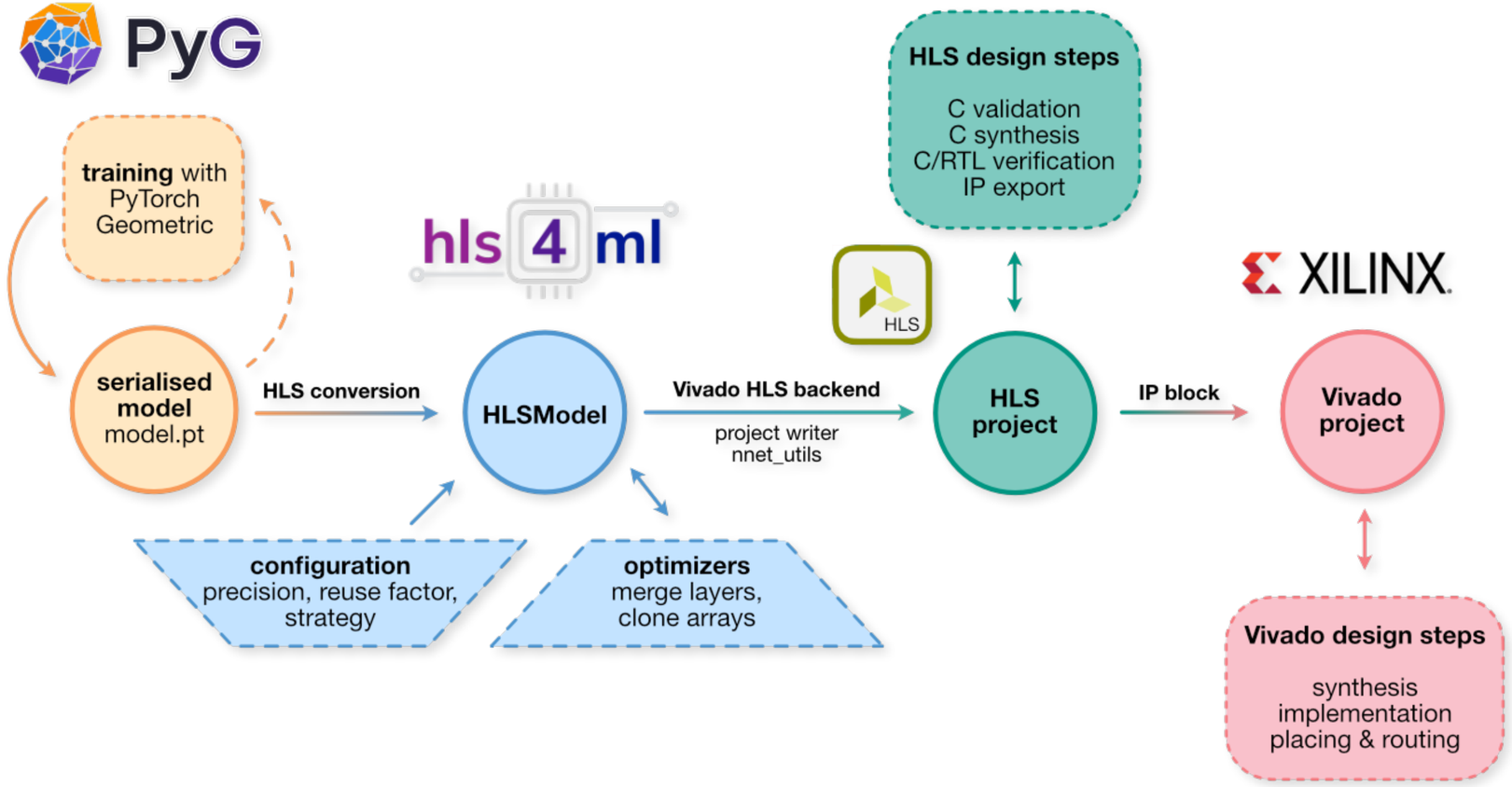


- What hits do we want to keep?
  1) all hits from signal particles?
  2) or only some?

Object Condensation Track Finding (offline)



Highlevel Synthesis for Machine Learning (hls4ml)



- automatic translation of ML models to hardware level
- based on Vivado HLS
- fast prototyping by automated workflow
- several configuration parameters

Design Optimisation

Quantization

- reduction of number of bits used representing the NN model (precision)
- data representation by arbitrary precision fixed-point data format *ap_fixed<W,I>*
- bit widths directly affect resource usage

Compression

- reduction of model parameter number
- number of model parameters strongly depends on the number of hidden nodes/ neurons

Pipelining

- key advantage of FPGAs: throughput acceleration by parallelisation and pipelining
- total latency = iteration latency + II · (number of functions -1)
- pipelining includes task parallelism, pipelining within a run, of runs or within a task