

Real-time AI for particle physics measurements

Physics Colloquium Siegen, 16.07.2026

Torben FERBER (torben.ferber@kit.edu)
Institute of Experimental Particle Physics (ETP)
Karlsruhe Institute of Technology (KIT)





AI reconstruction team at KIT:

ongoing:

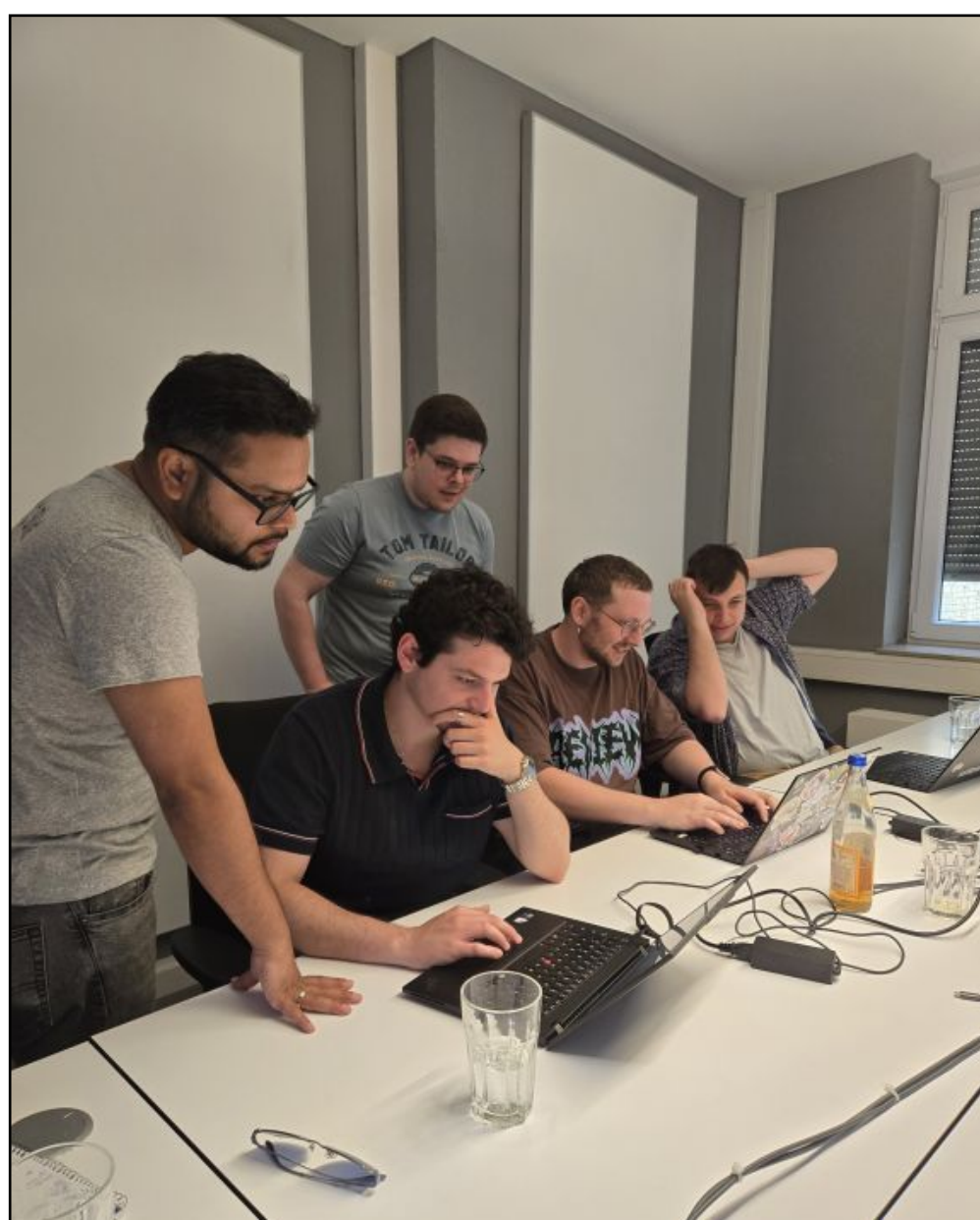
Dr. Giacomo De Pietro^{1,3}, Marc Neu², Kai Unger², Georgios Sotiropoulos², Tristan Brandes¹, Karlina Kröhnert¹, Marc Thomas¹, Marc Martini¹, Felix Herdtweck¹, Daniel Grossmann¹, Fabian Schmalenbach², Maximilian Korb², Christoph Boehrer⁴, Dr. Karim Guirguis⁴

¹ KIT, Institute of Experimental Particle Physics (ETP)

² KIT, Institute of Information Processing (ITIV)

³ KIT, Scientific Computing Center (SCC)

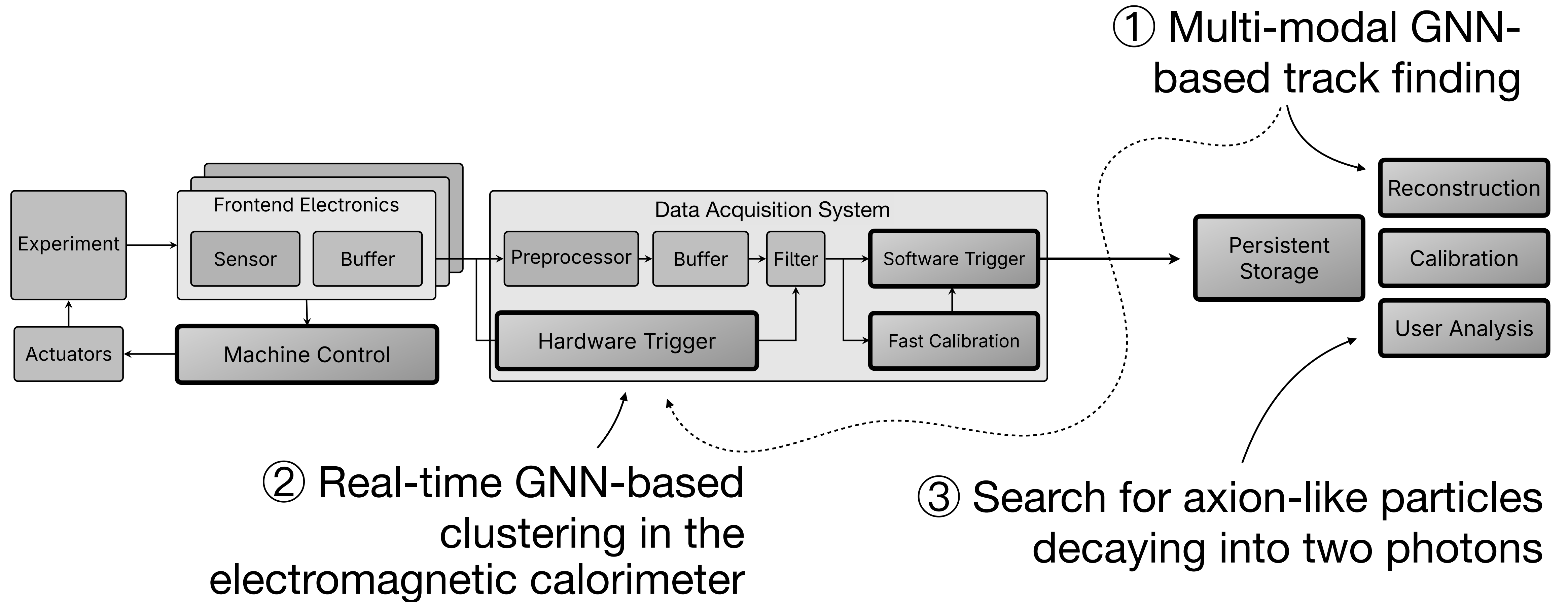
⁴ SiMa.ai



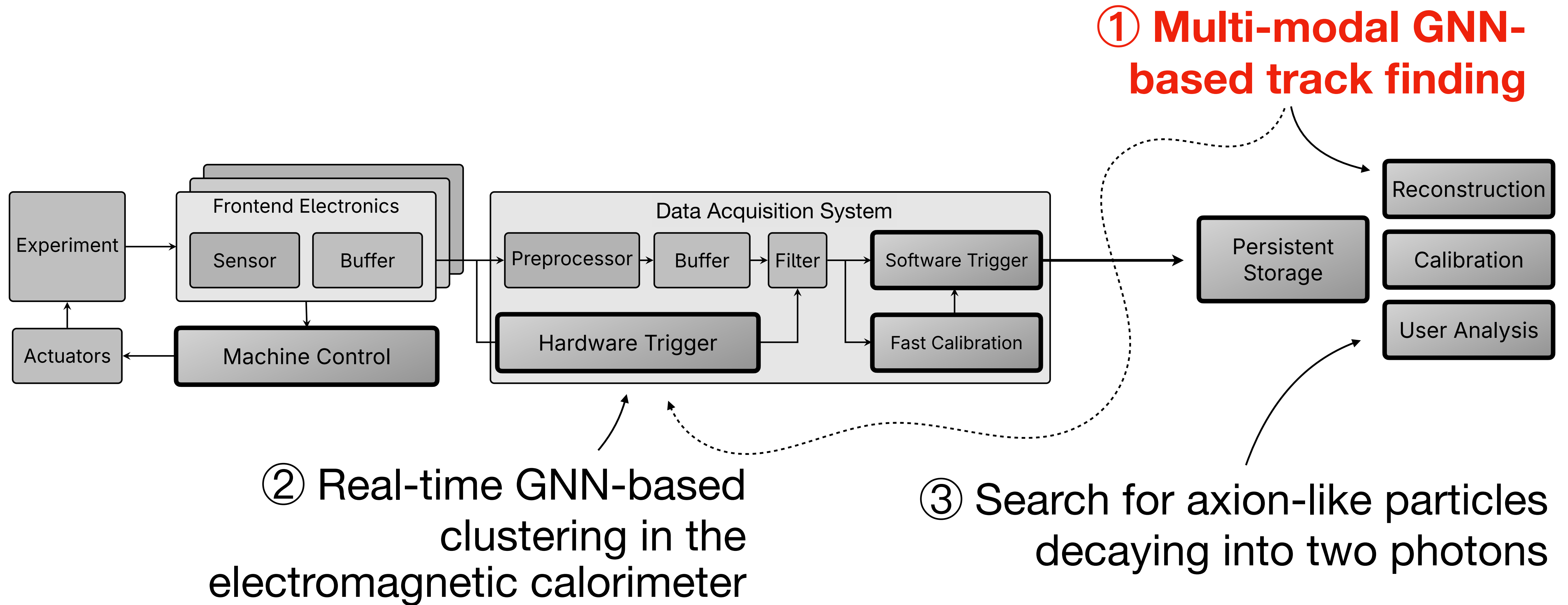
past:

Dr. Isabel Haide¹, Dr. Lea Reuter¹, Dr. Greta Heine¹, Thomas Lobmaier¹, Frank Baptist¹, Fabio Papagno², Valdrin Dajaku², Timo Justinger², Till Raedler²

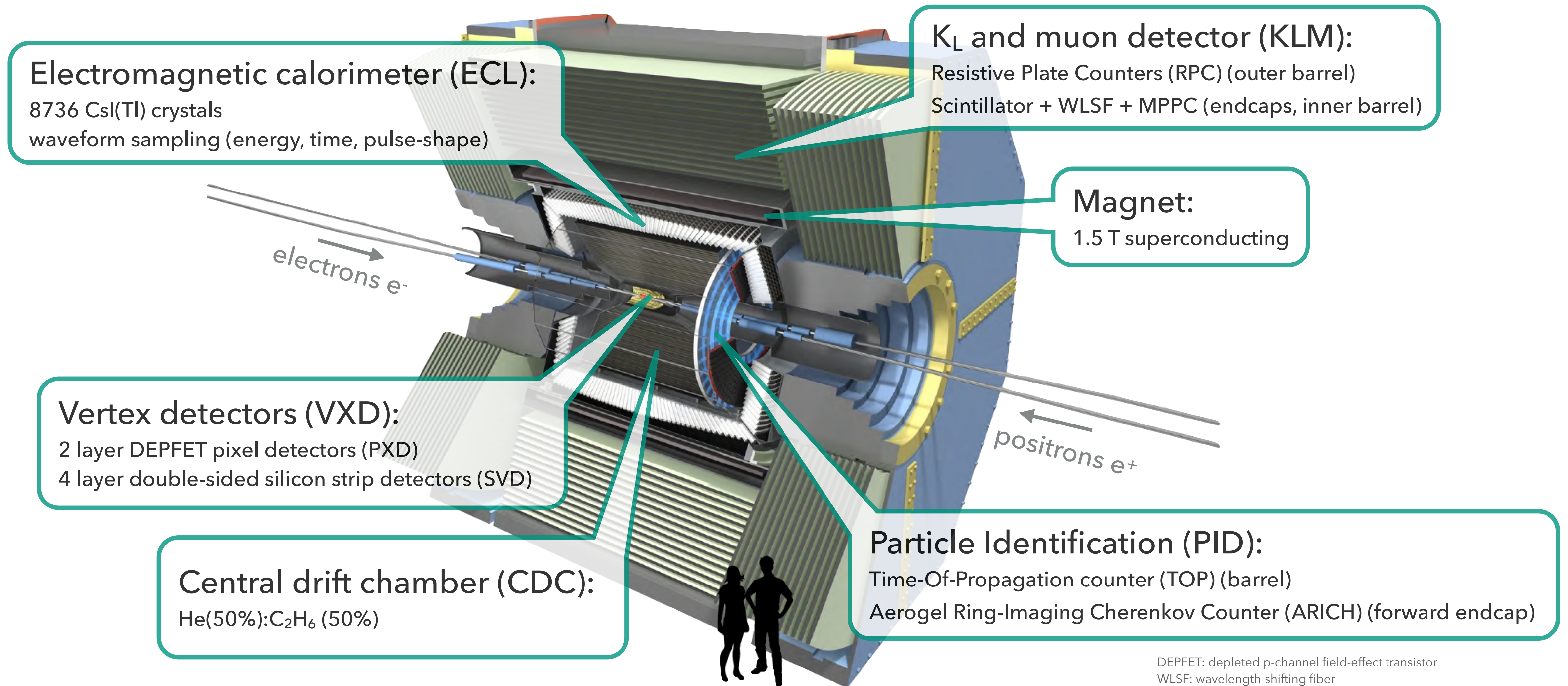
Outline



Outline

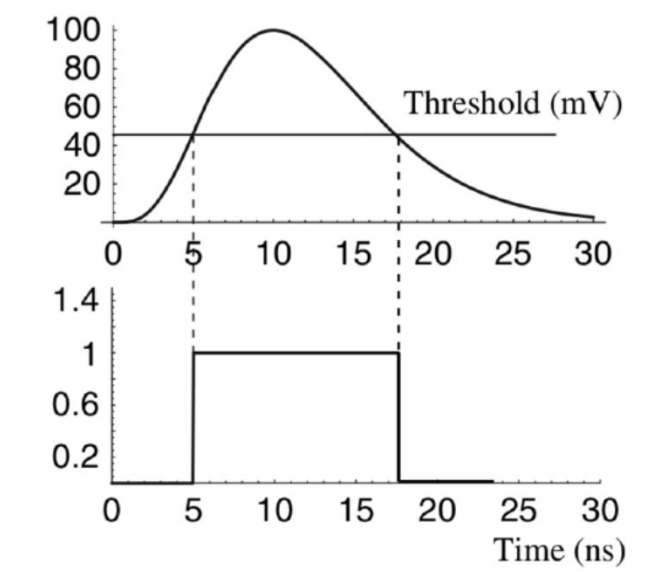
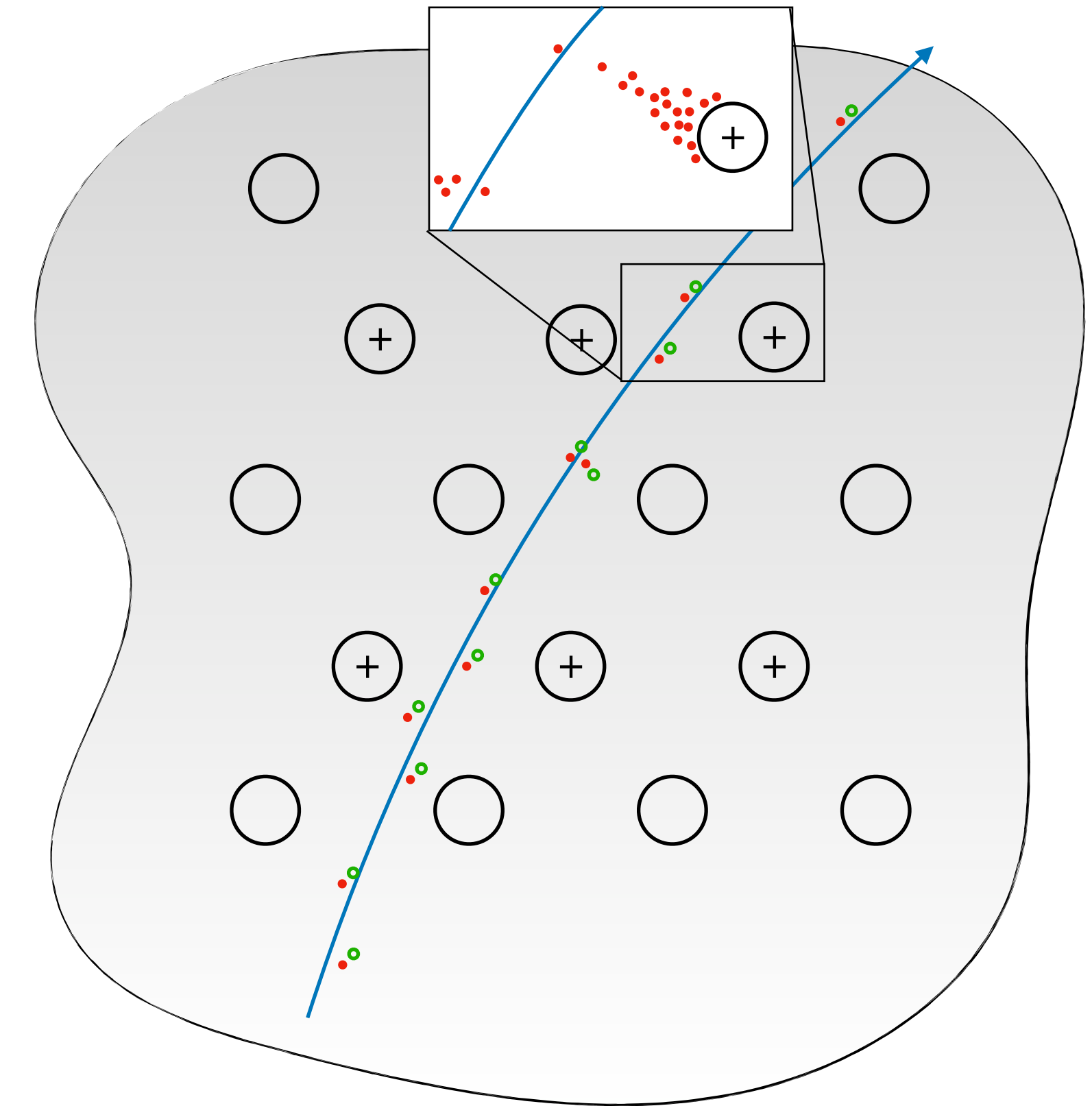
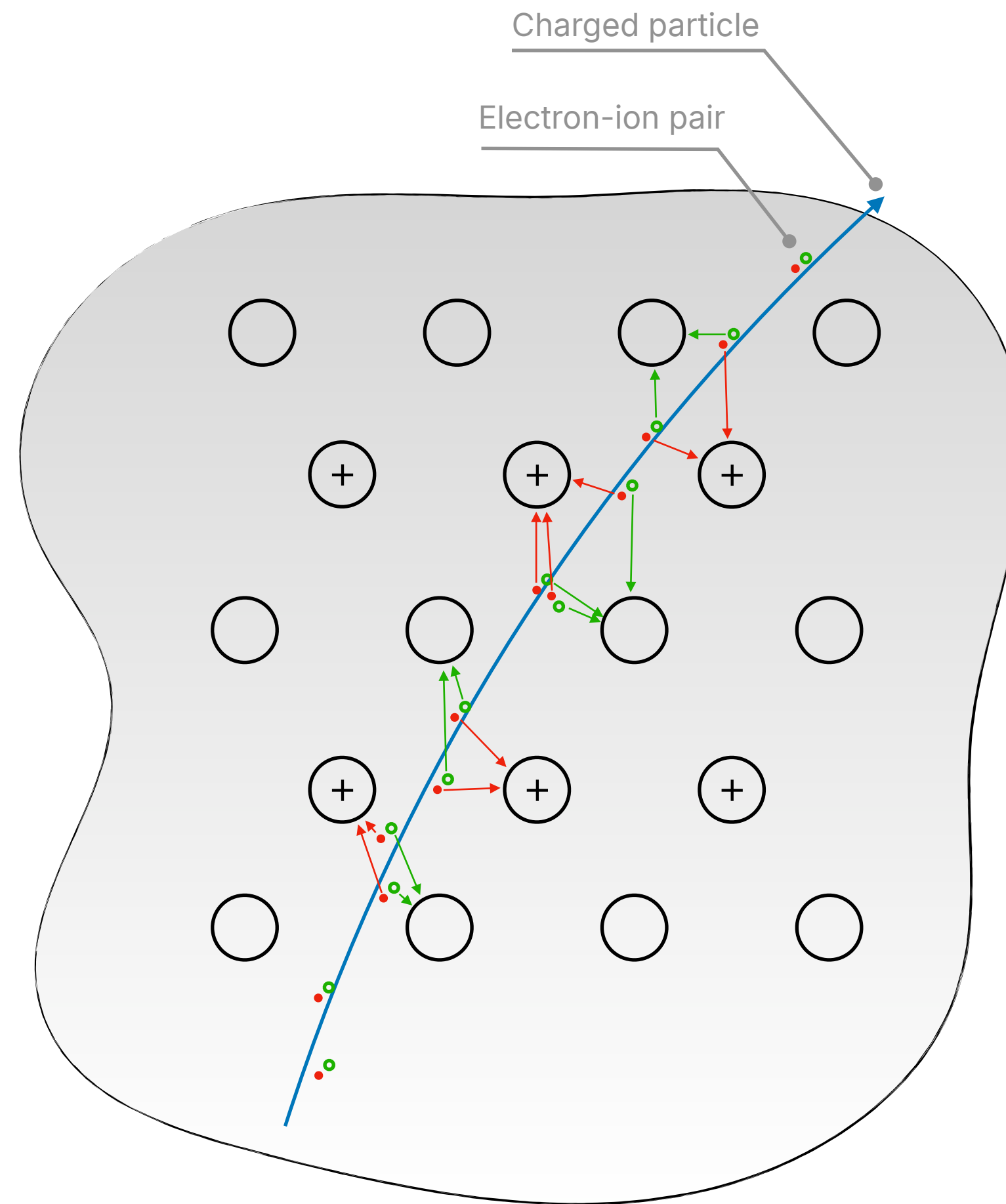
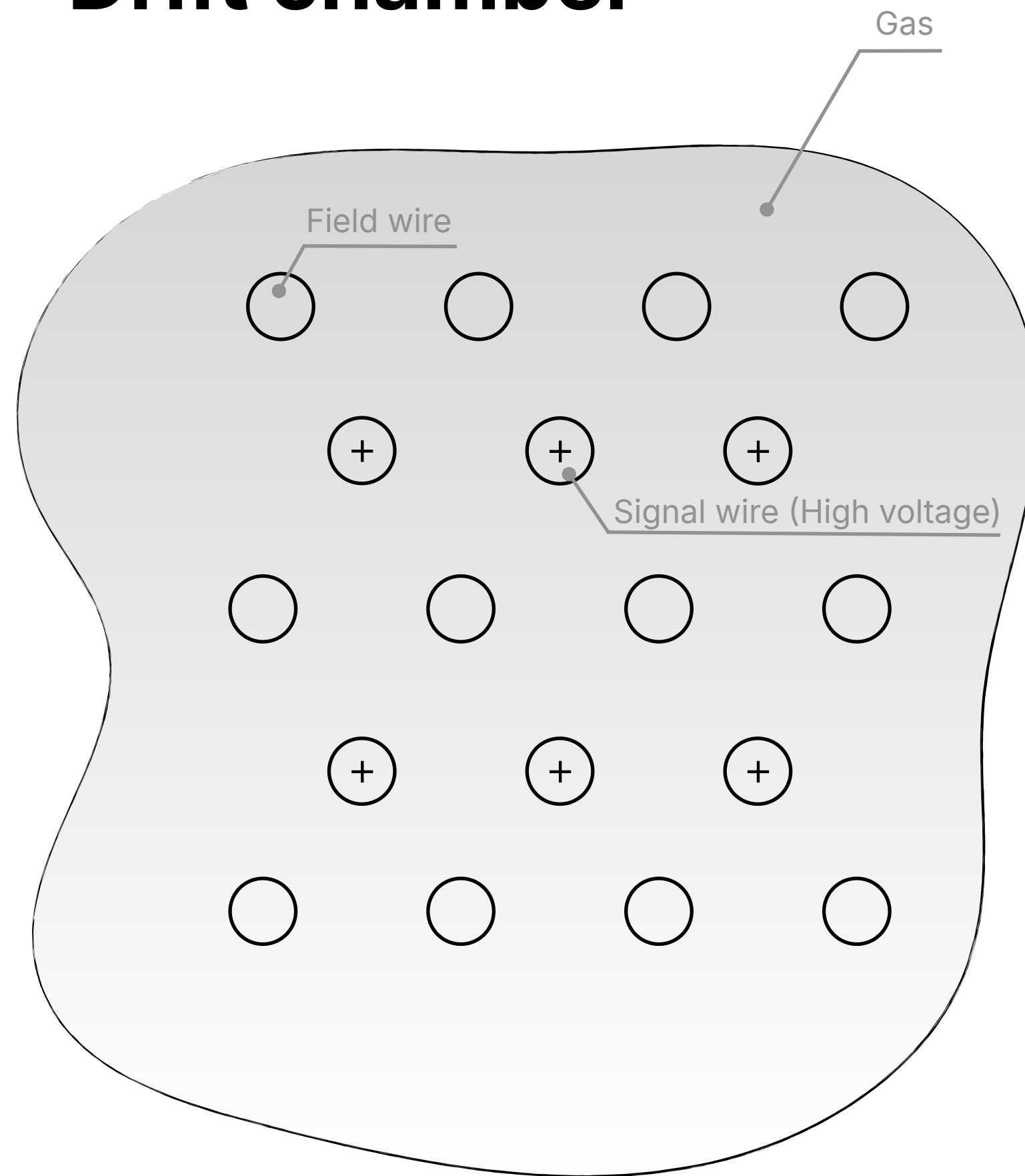


Belle II



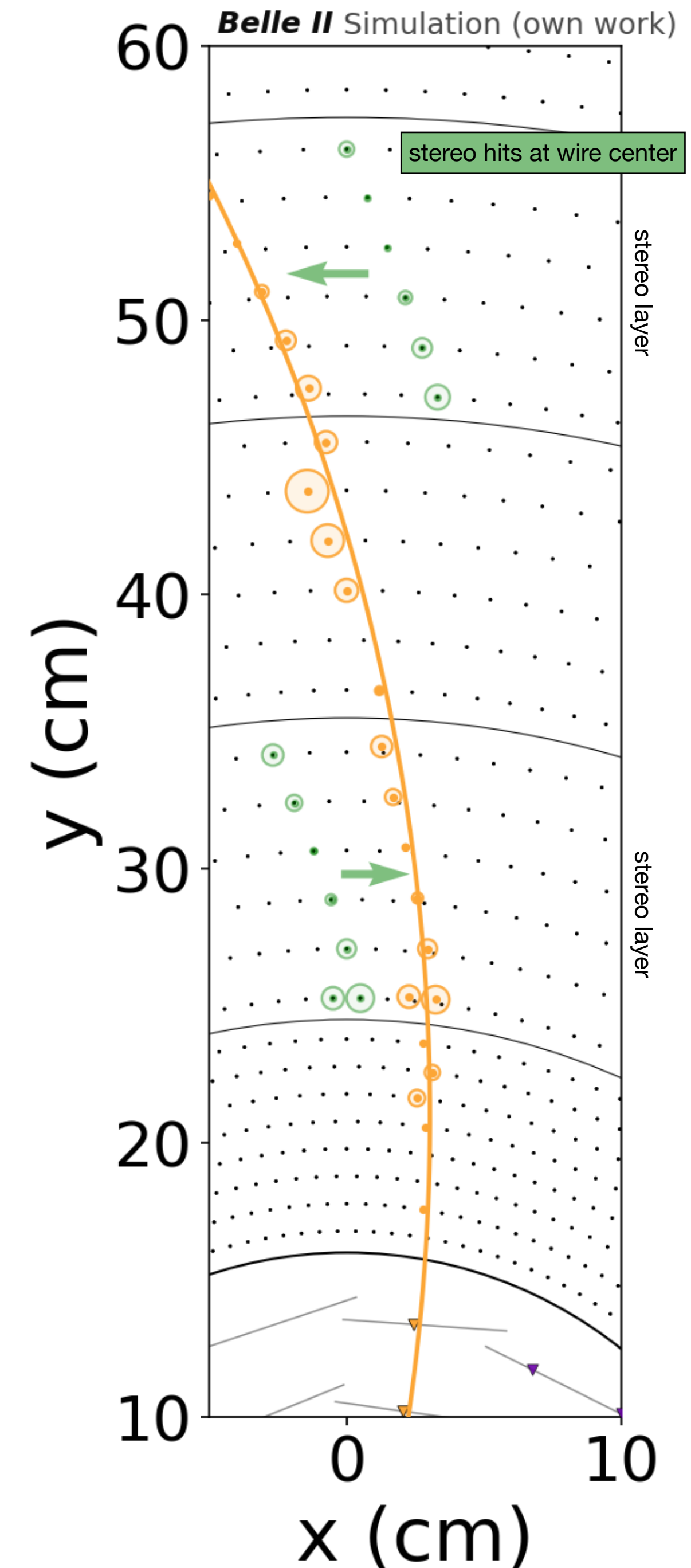
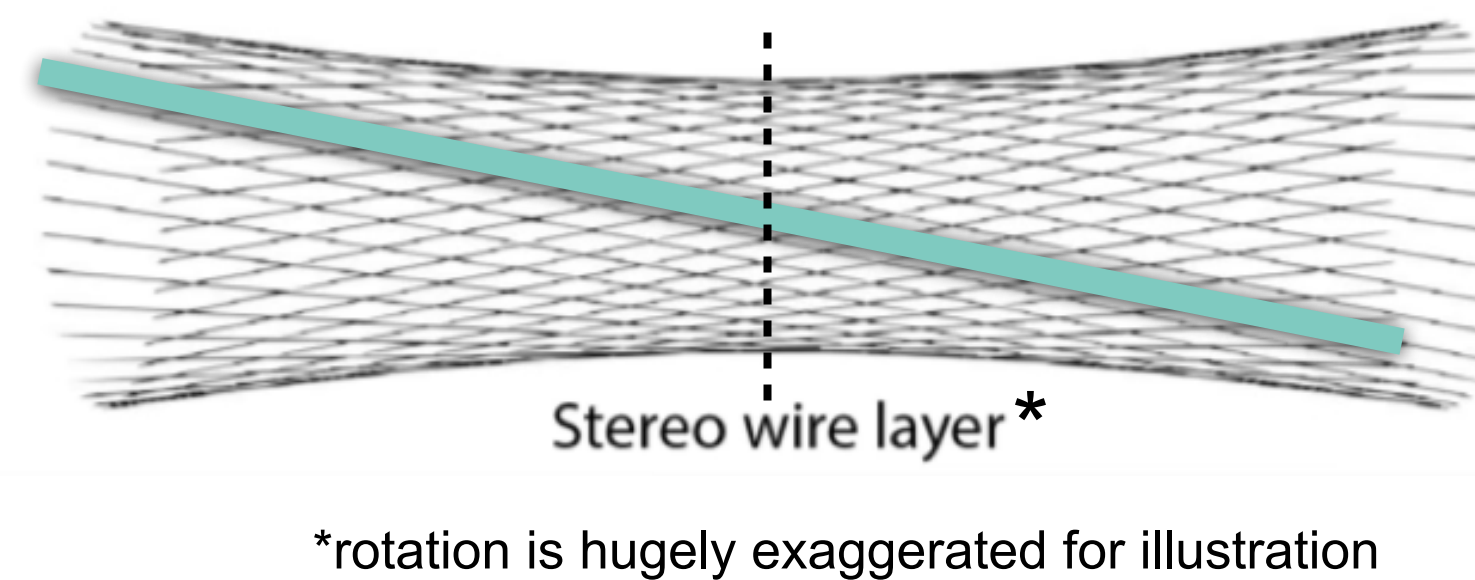
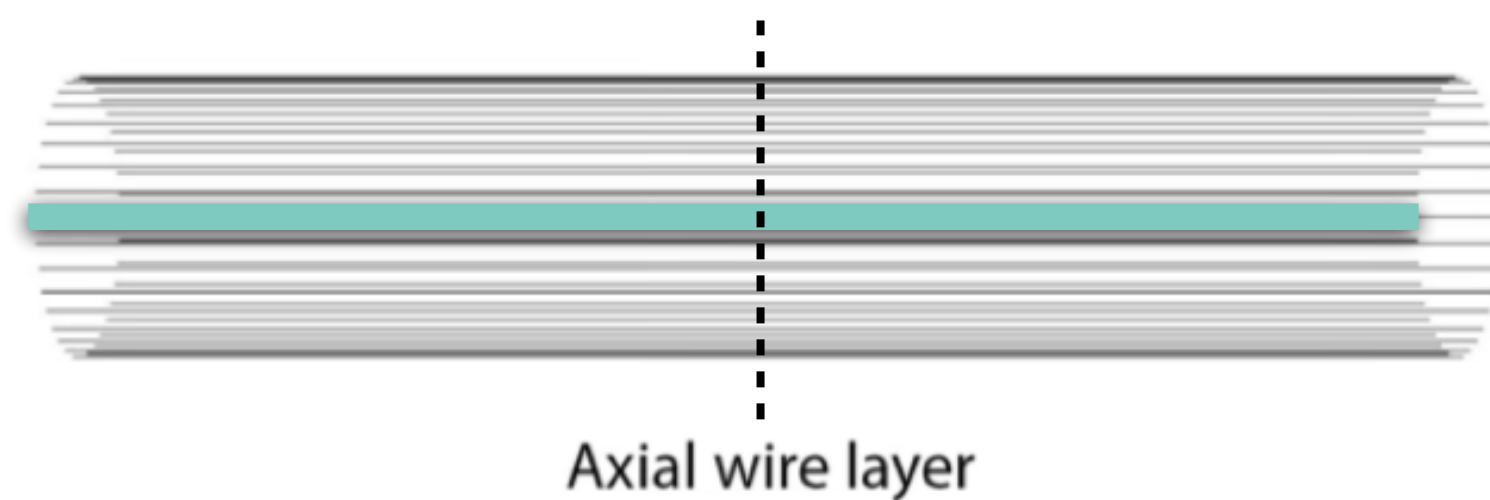
Charged particles in a gas detector

Drift chamber



Charged particles in a drift chamber

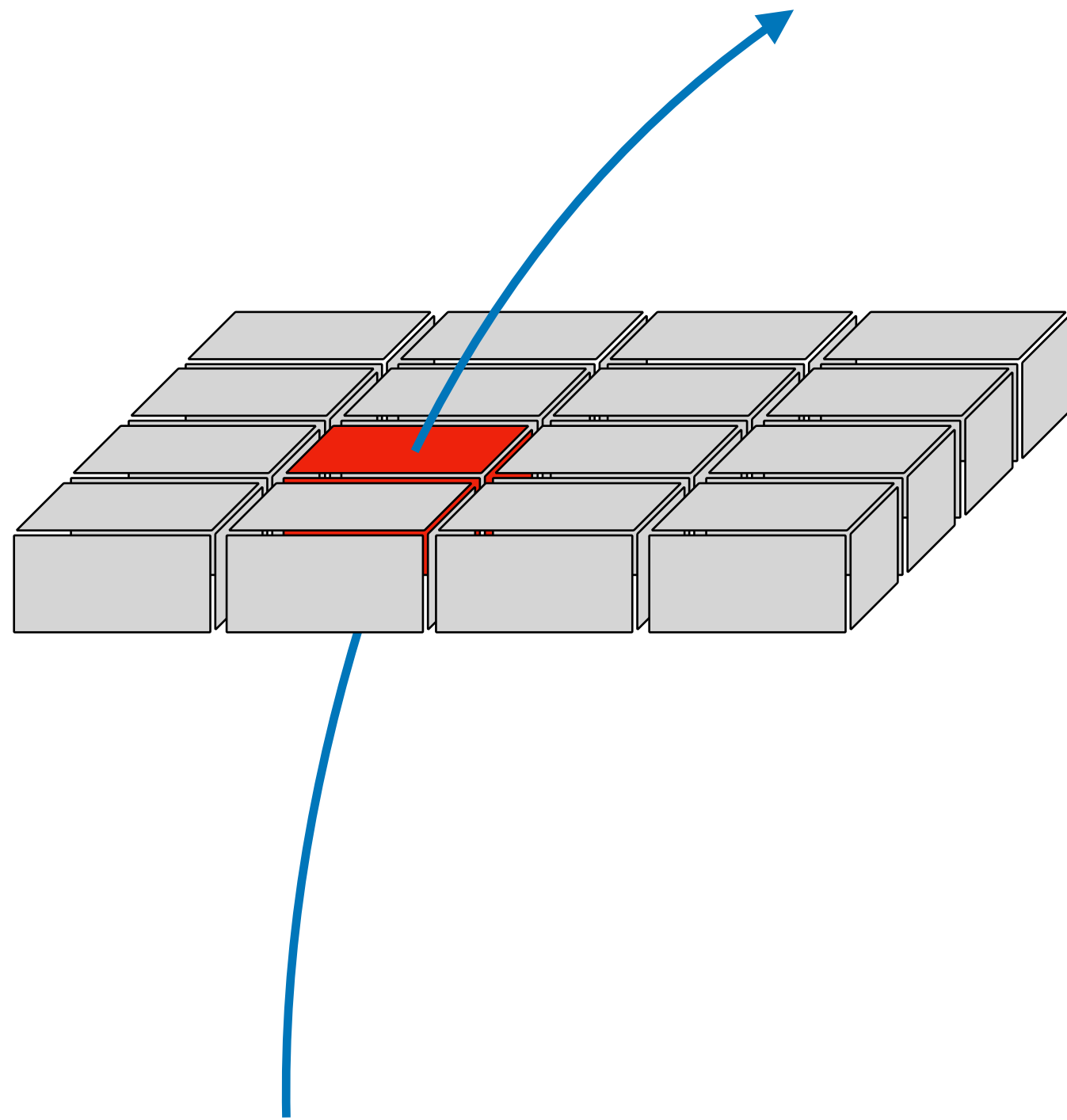
- Each fired wire gives a drift distance (drift time $\times$ drift velocity) = a **drift circle** around the wire: the track is tangent to these circles
- **Stereo wires** tilted by a small angle relative to axial wires: that tilt is what gives the z coordinate along the wire when combined to a track



Charged particles in silicon detectors

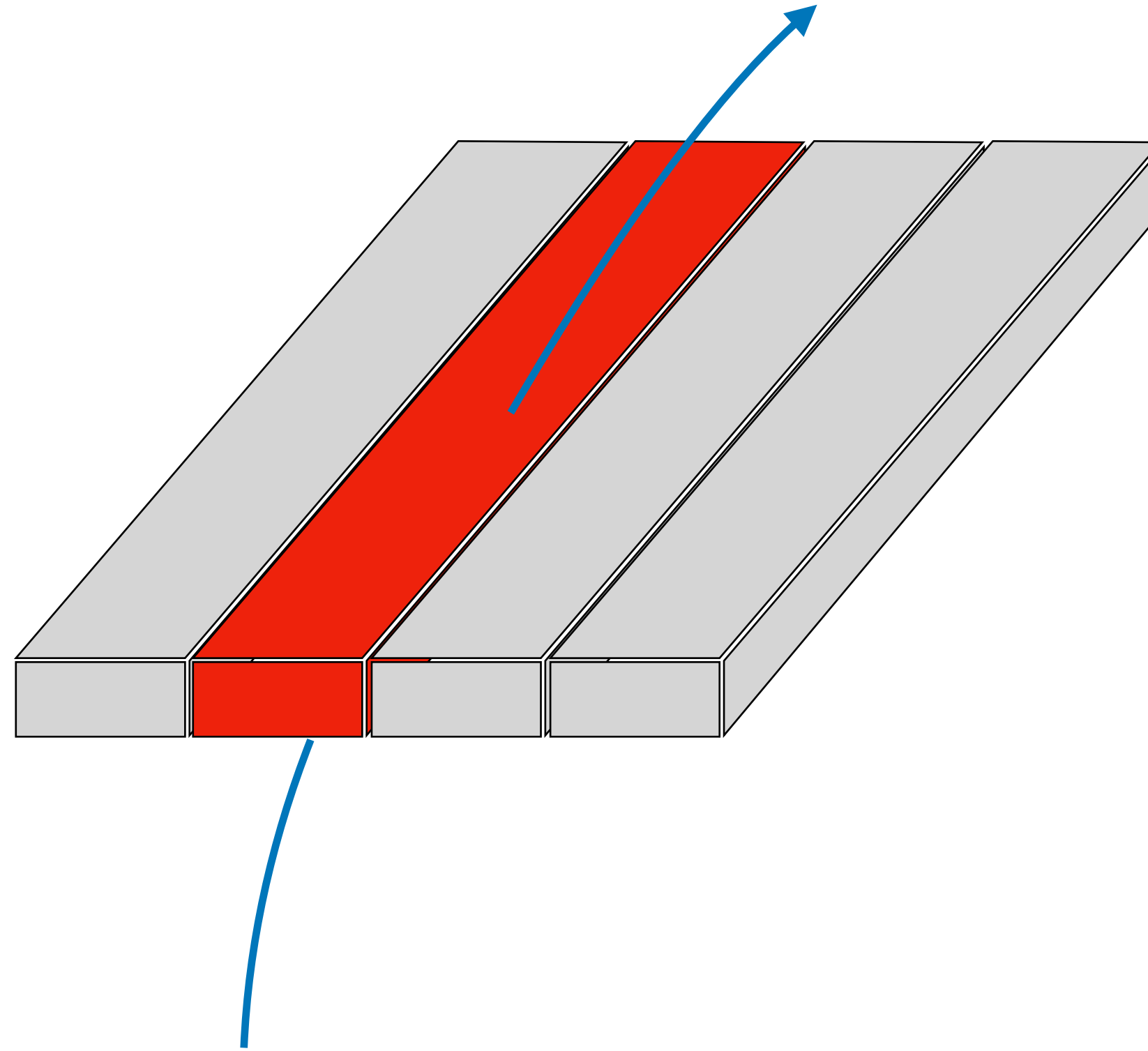
Pixels

segmented in 2D. One hit gives a precise 2D point directly (→ 3D with the layer position)

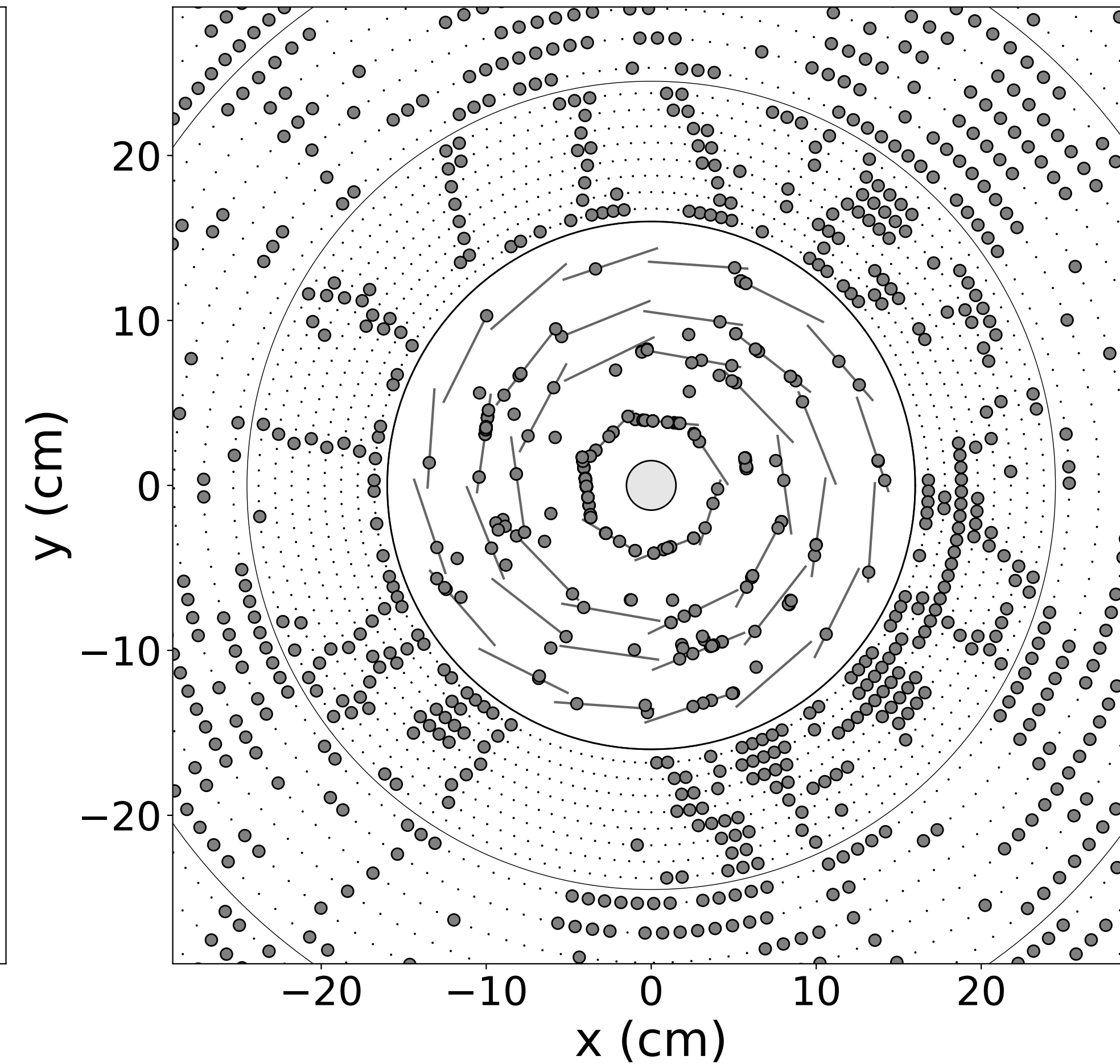
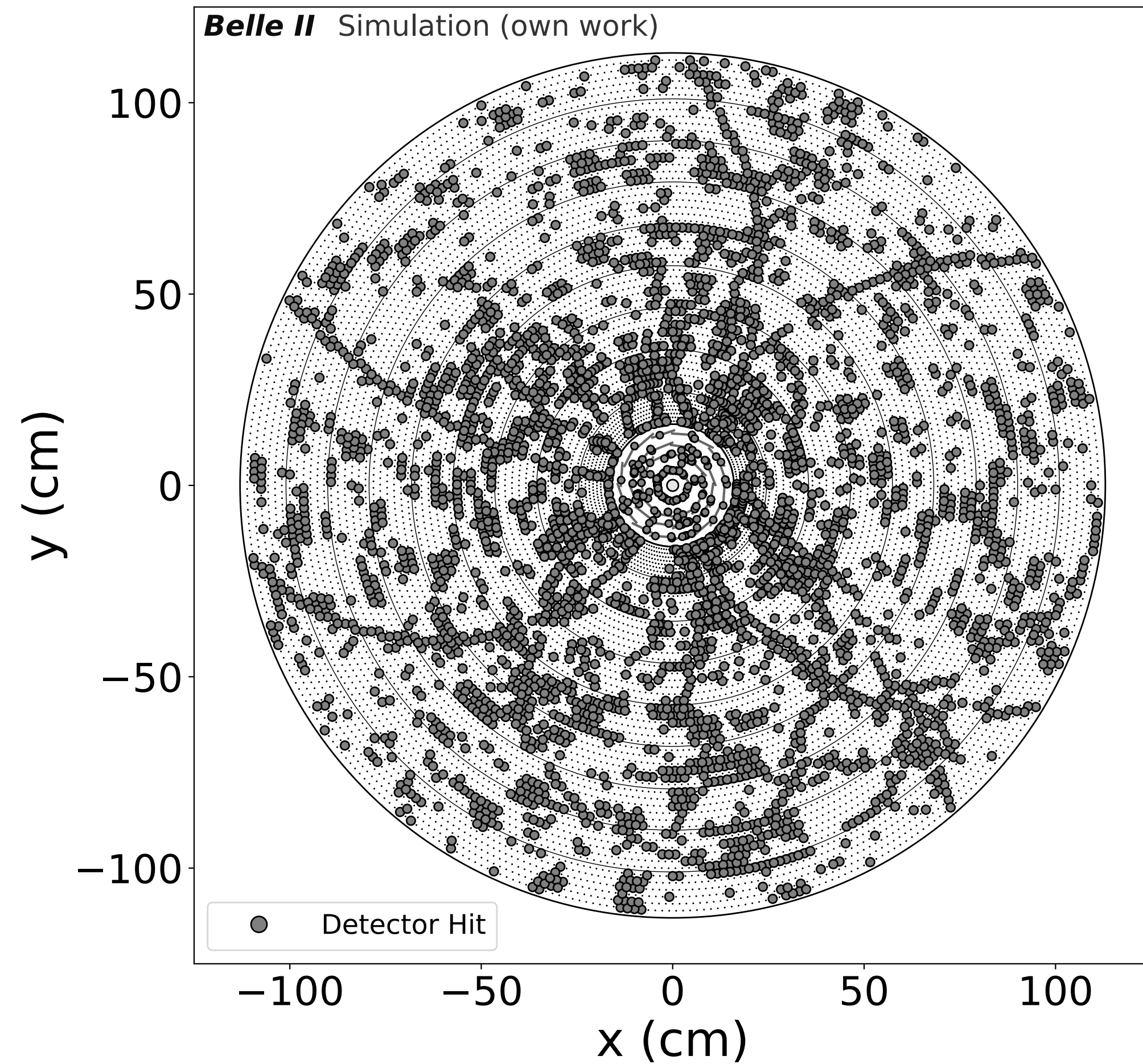


Strips

segmented in 1D. One hit gives one coordinate (→ the second coordinate needs a crossed strip layer)

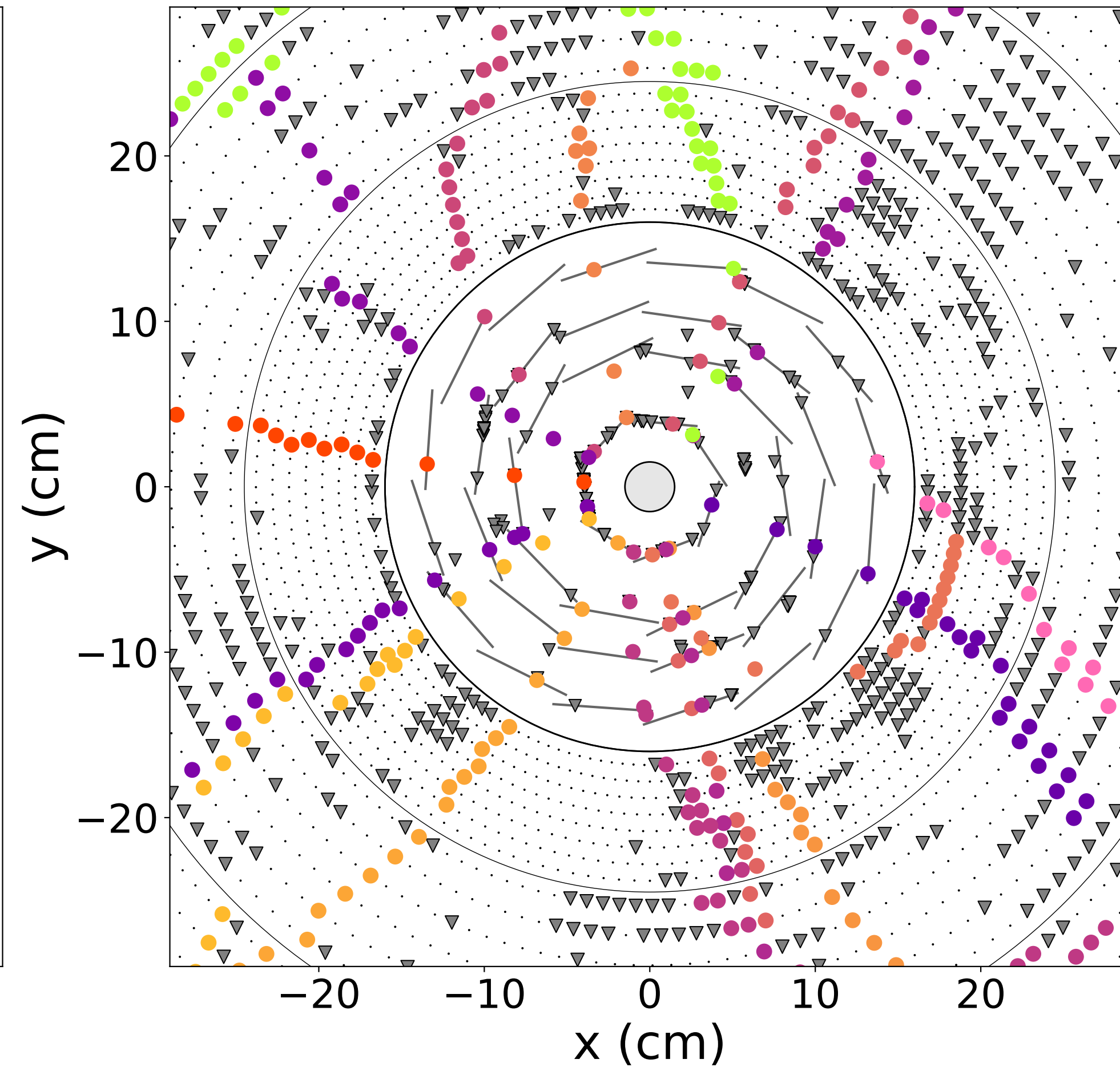
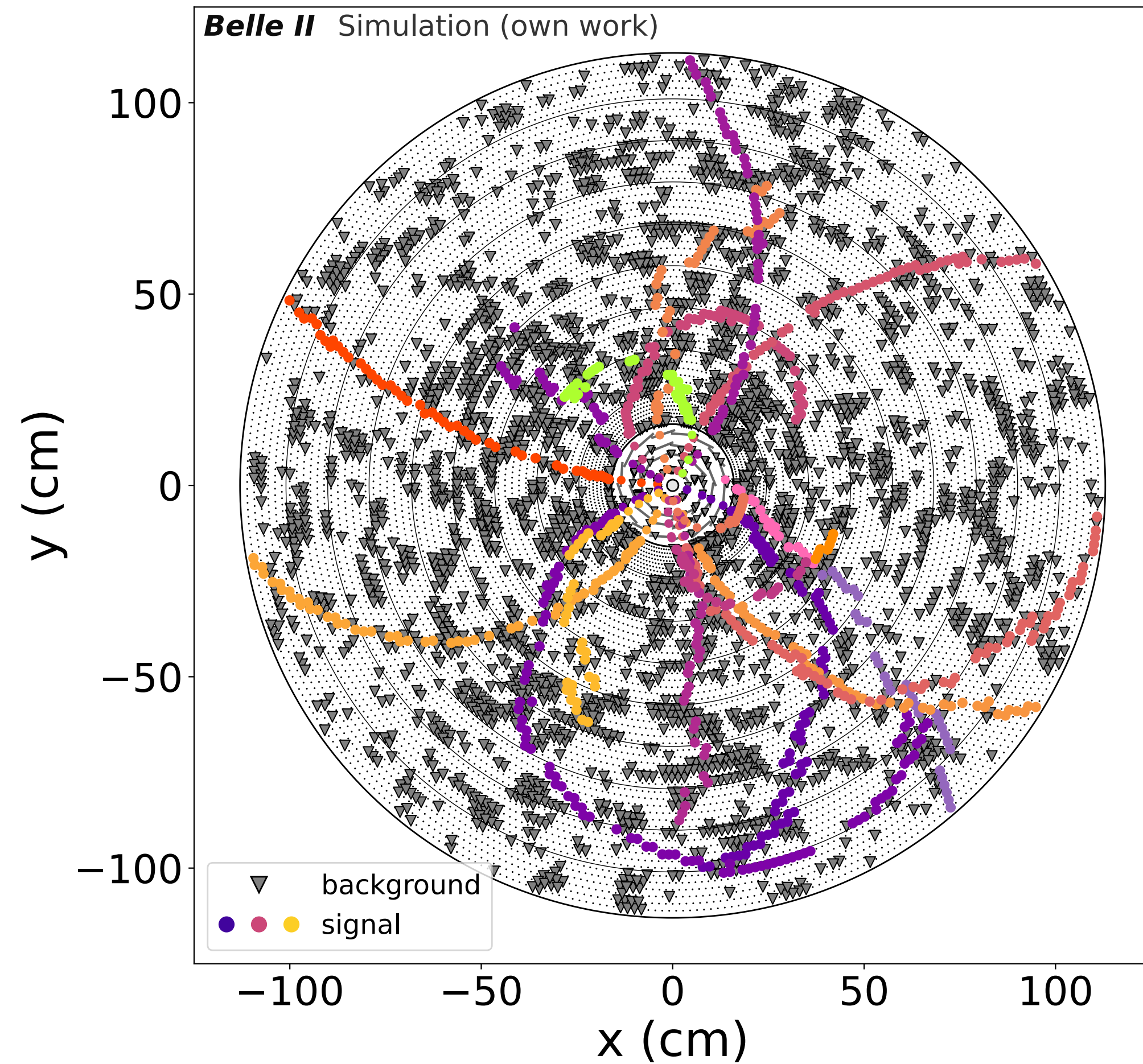


Raw hits in the Belle II detector

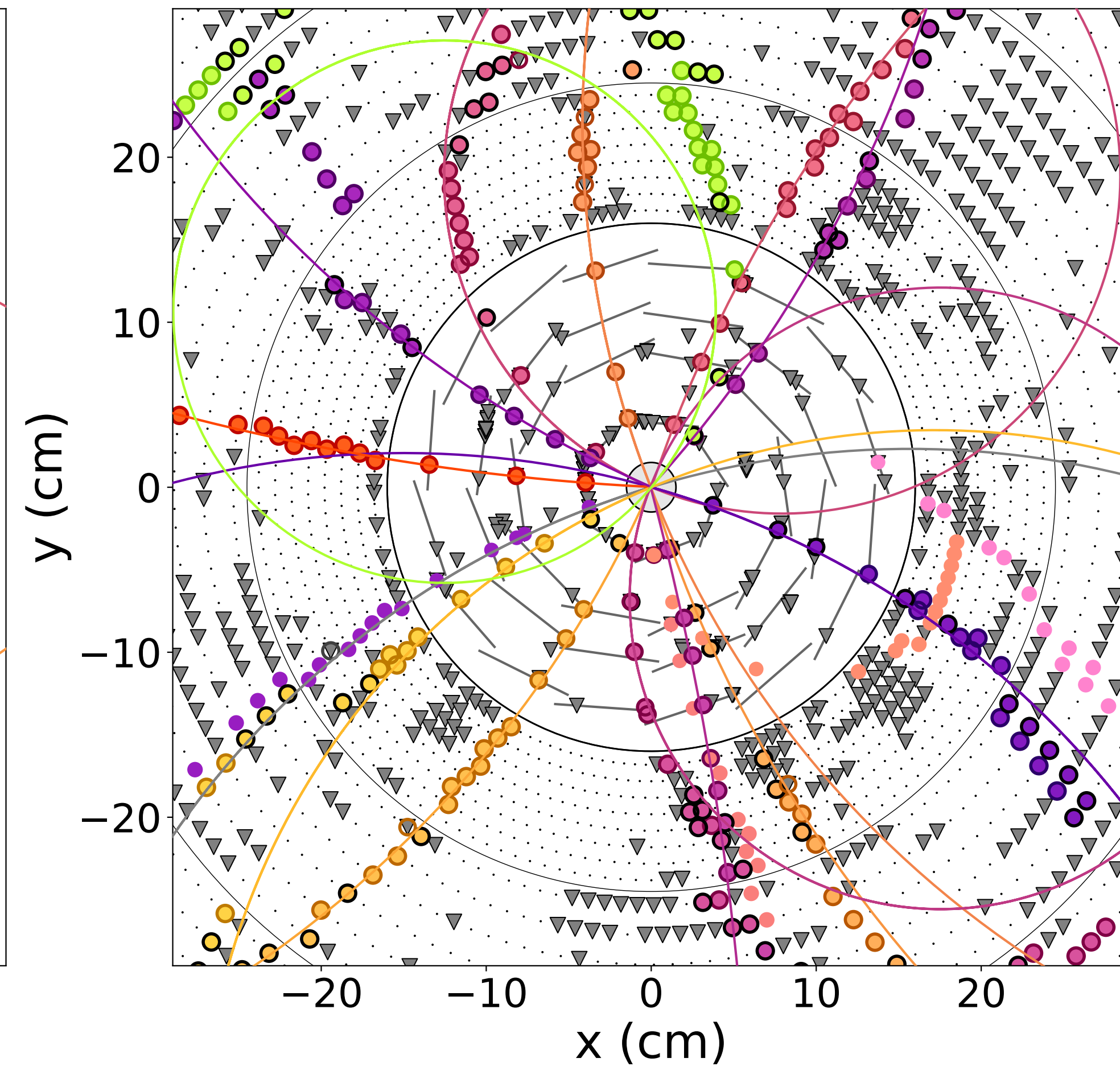
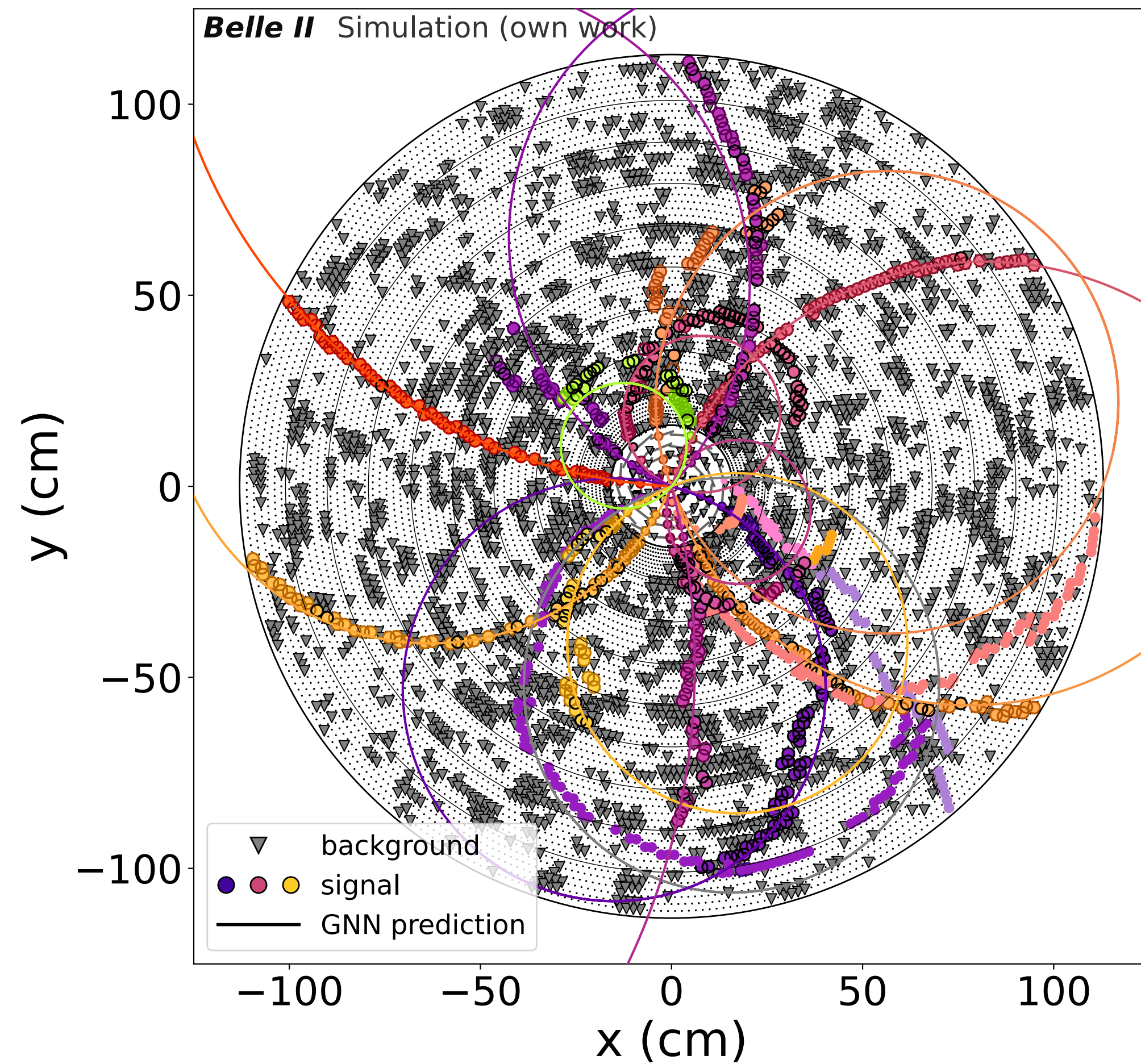


Track finding: Pattern recognition

Right answer:
14 tracks

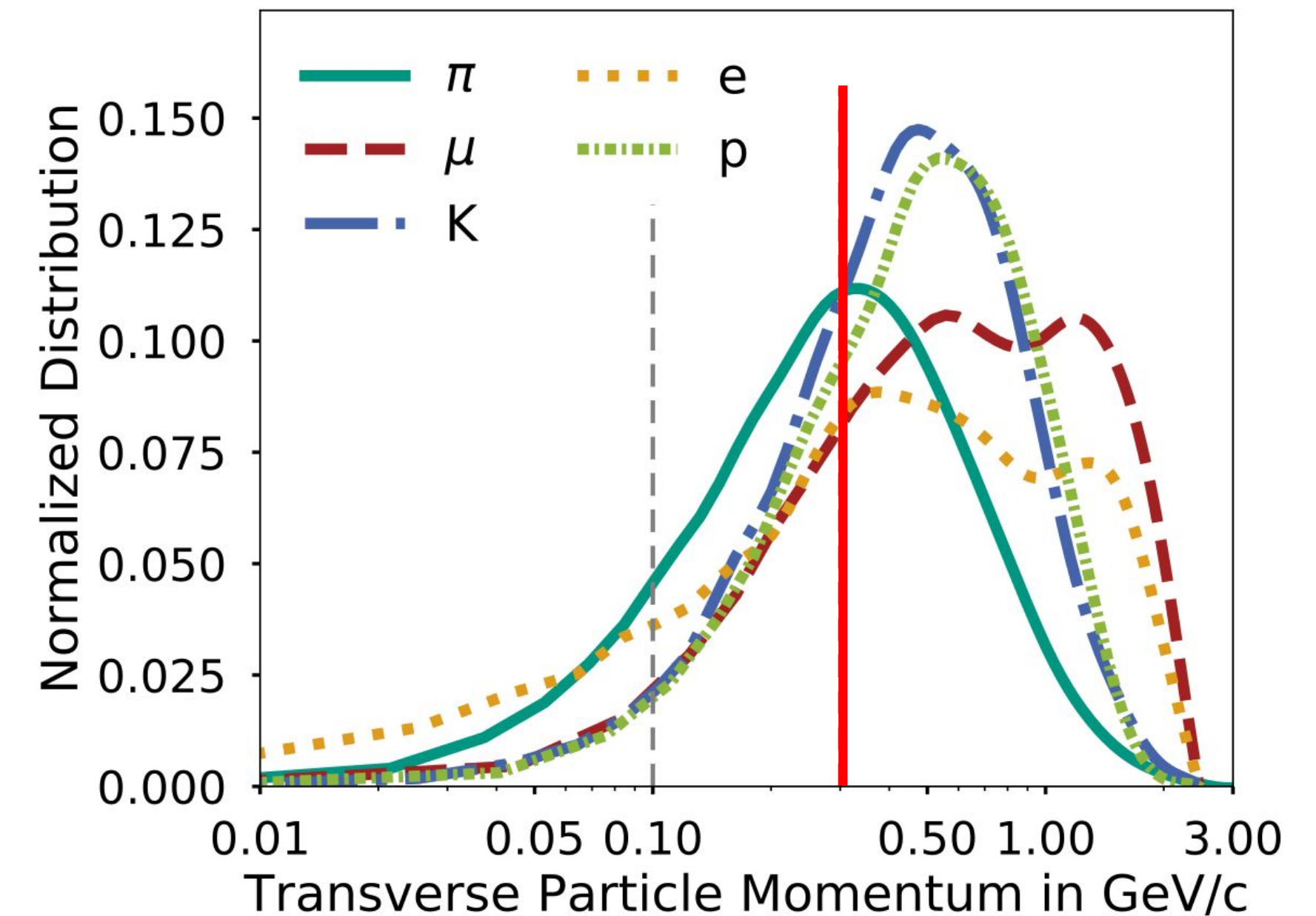


Track fitting: Minimisation of a helix fit



Charged particles at Belle II

- Particles with $p_t > 500$ MeV transverse the full tracking detector
 - On average around 55 CDC hits, 4 SVD hits, and 2 PXD hits per track
- Most charged particles are pions with momentum around 300 MeV range
 - Tracks curl back in the magnetic field
 - Multiple scattering contributes significantly to momentum resolution
- No pile-up, but high beam-induced backgrounds (photons, electrons, neutrons)

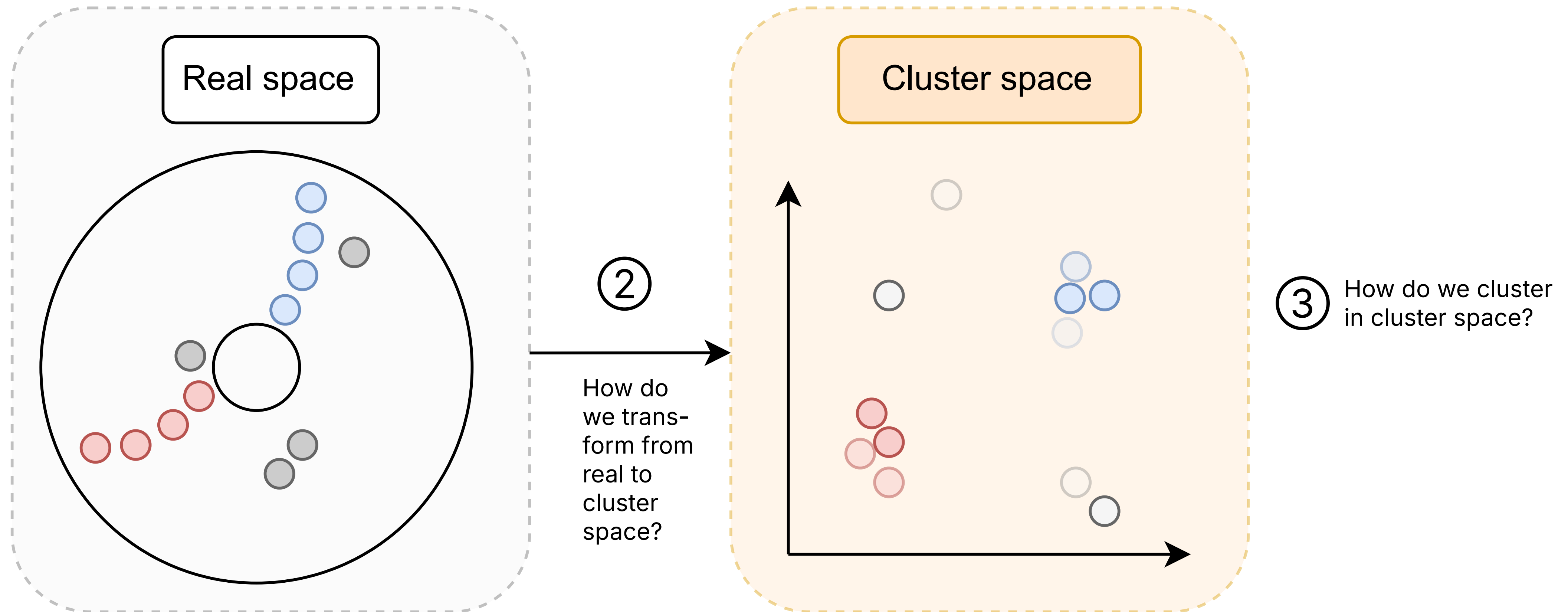


Algorithm requirements

- Identify the **number of tracks**
- Predict the **three-momentum, starting point**, and the **charge** of each track
- Assign the **associated hits** to each track
- Detect **overlapping tracks** without clear spatial boundaries
- Be sensitive to a large dynamic range of **track lengths** and **curvatures**
- Be **insensitive to the hit-ordering**
- Be **rotation** and **charge invariant**
- Be **fully differentiable, no iterations**
- Be **efficient for sparse inputs** (i.e. most inputs are zero for a given event)
- Be **robust against moderate changes** in detector configurations and backgrounds

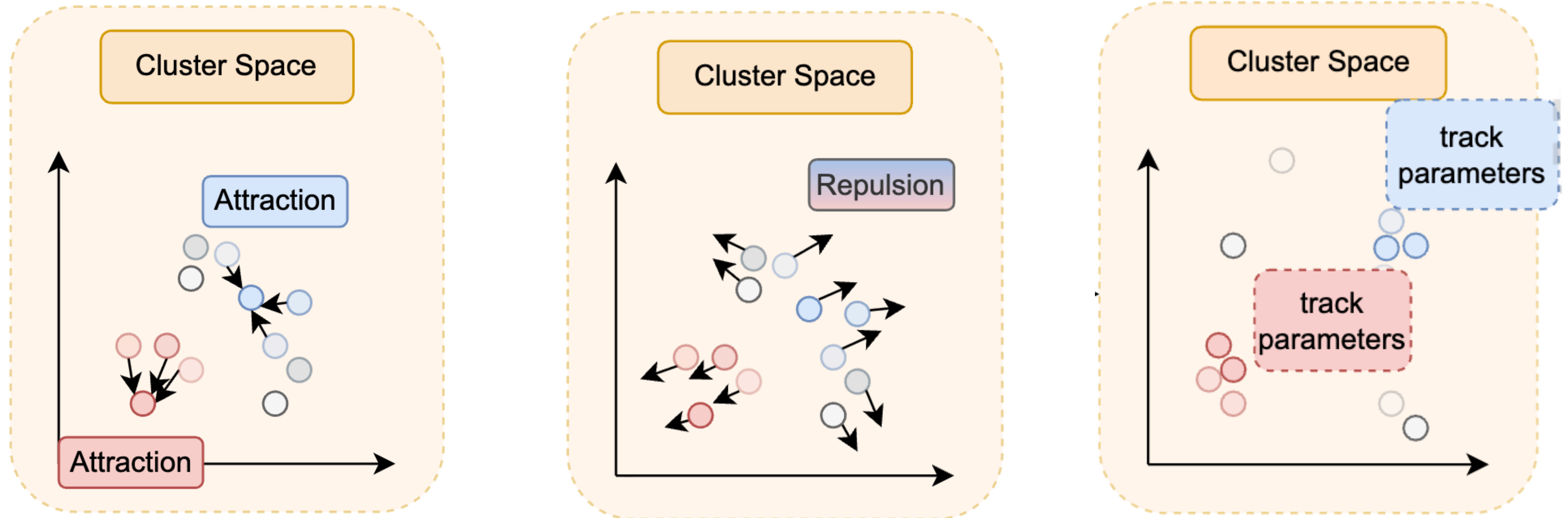
Trackfinding is a clustering problem in the wrong space

① What do we actually optimize?



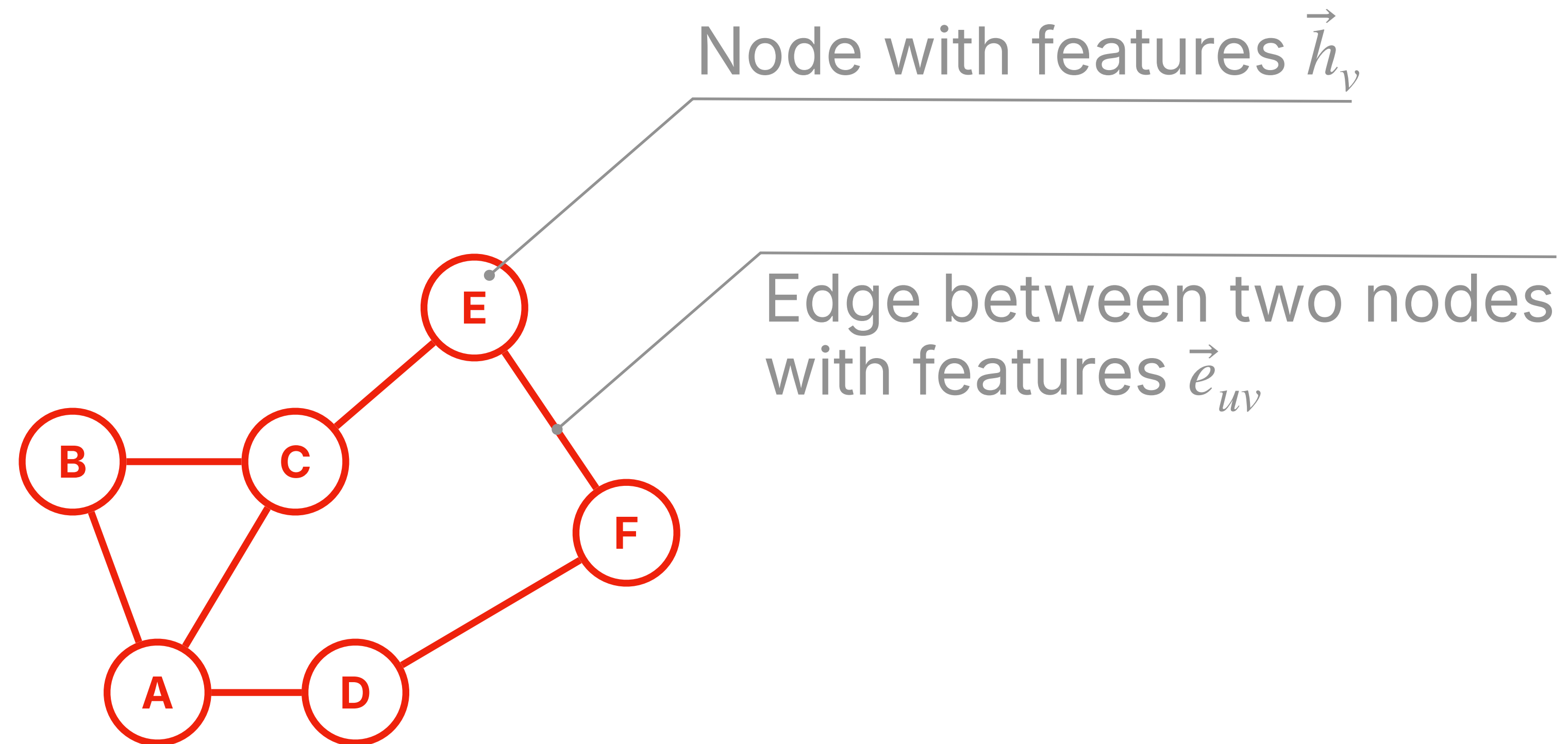
L. Reuter PhD thesis <https://publikationen.bibliothek.kit.edu/1000189859>

① What do we actually optimize: Object Condensation



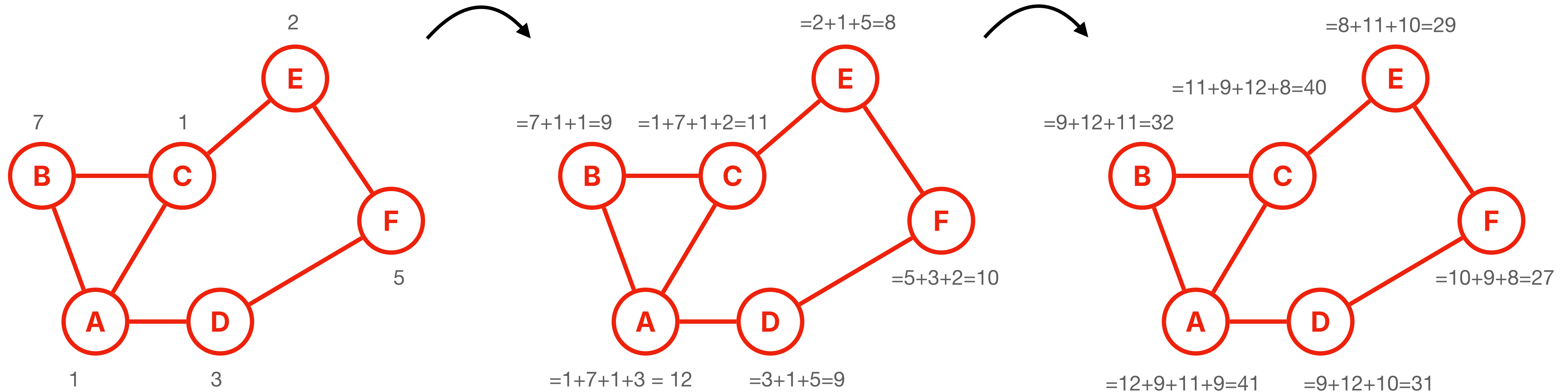
$$L = L_{\text{attractive}} + L_{\text{repulsive}} + L_{\beta} + L_{\text{suppress noise}} + L'_{\text{p}} + L'_{\text{v}} + L'_{\text{charge}}$$

Graph Neural Networks



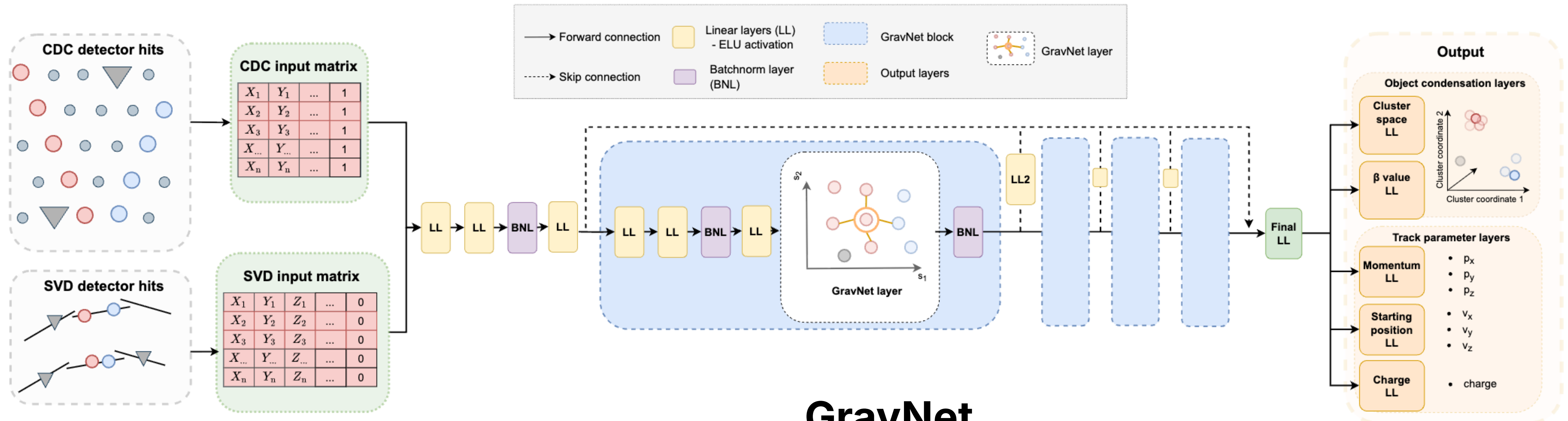
Graph: nodes are hits, edges are connections between nodes (no grid, no ordering required)

Graph Neural Networks: Message passing



Message passing: each node repeatedly updates itself from its neighbors' messages using an aggregation function, so information spreads across the graph (example: add-aggregation including self-node)

② How do we transform from real to cluster space: Deep GNNs

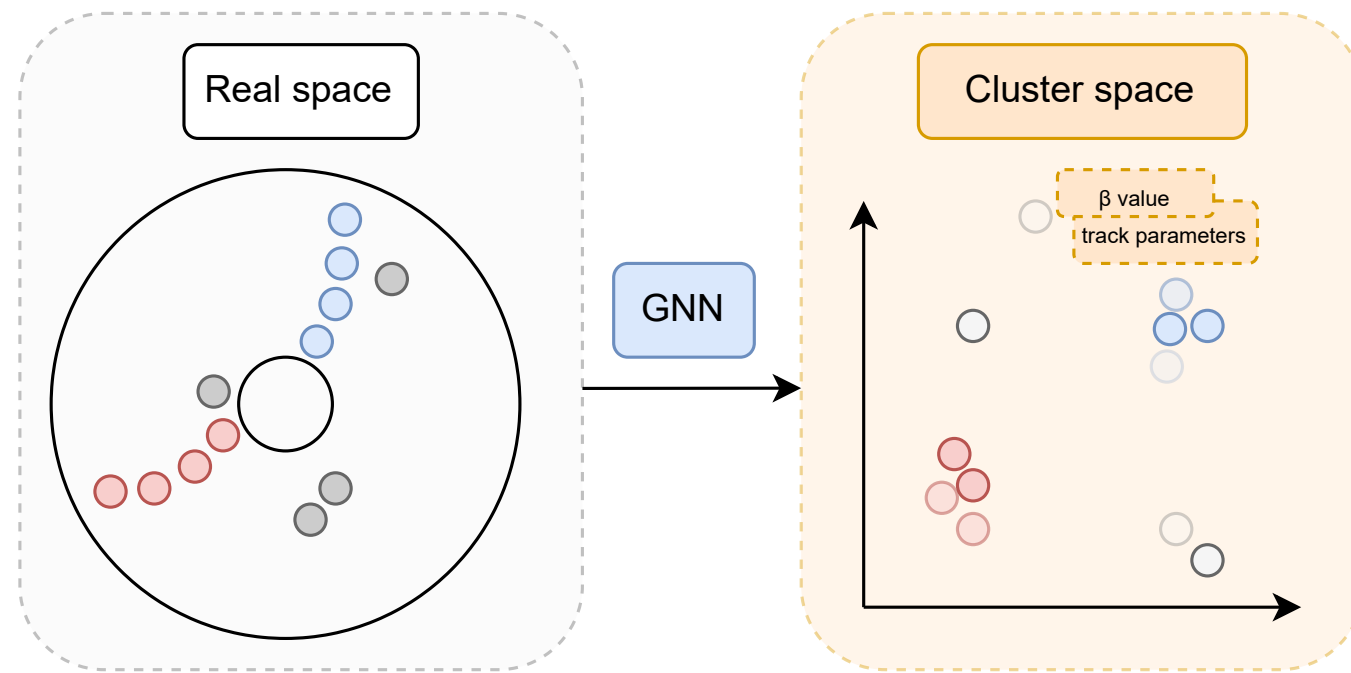


GravNet

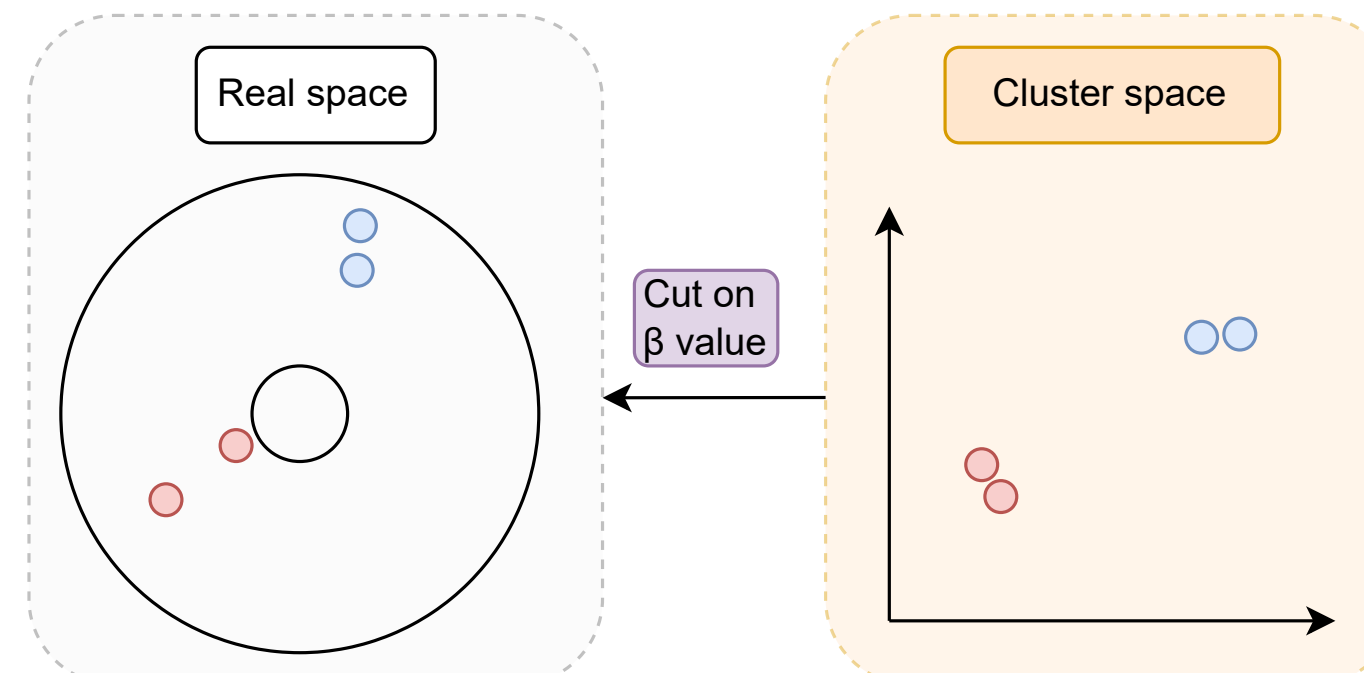
- Dynamic Graph Neural Network layer
- Learns an latent space representation of nodes and features
- Edges determined using k-NN neighbours in latent space (k=54)

L. Reuter et al. <https://arxiv.org/abs/2411.13596>
L. Reuter et al. <https://arxiv.org/abs/2602.10897>
S. R. Qasim et al. <https://arxiv.org/abs/1902.07987>

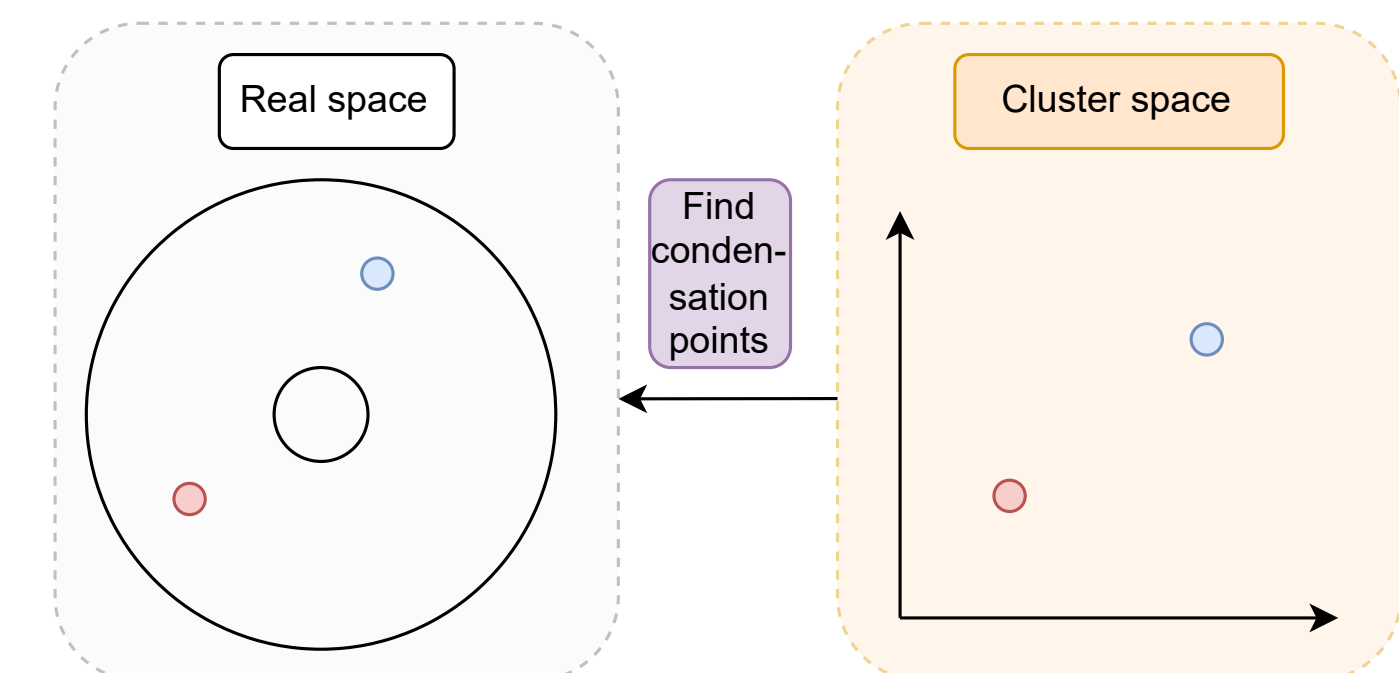
③ How do we cluster in cluster space?



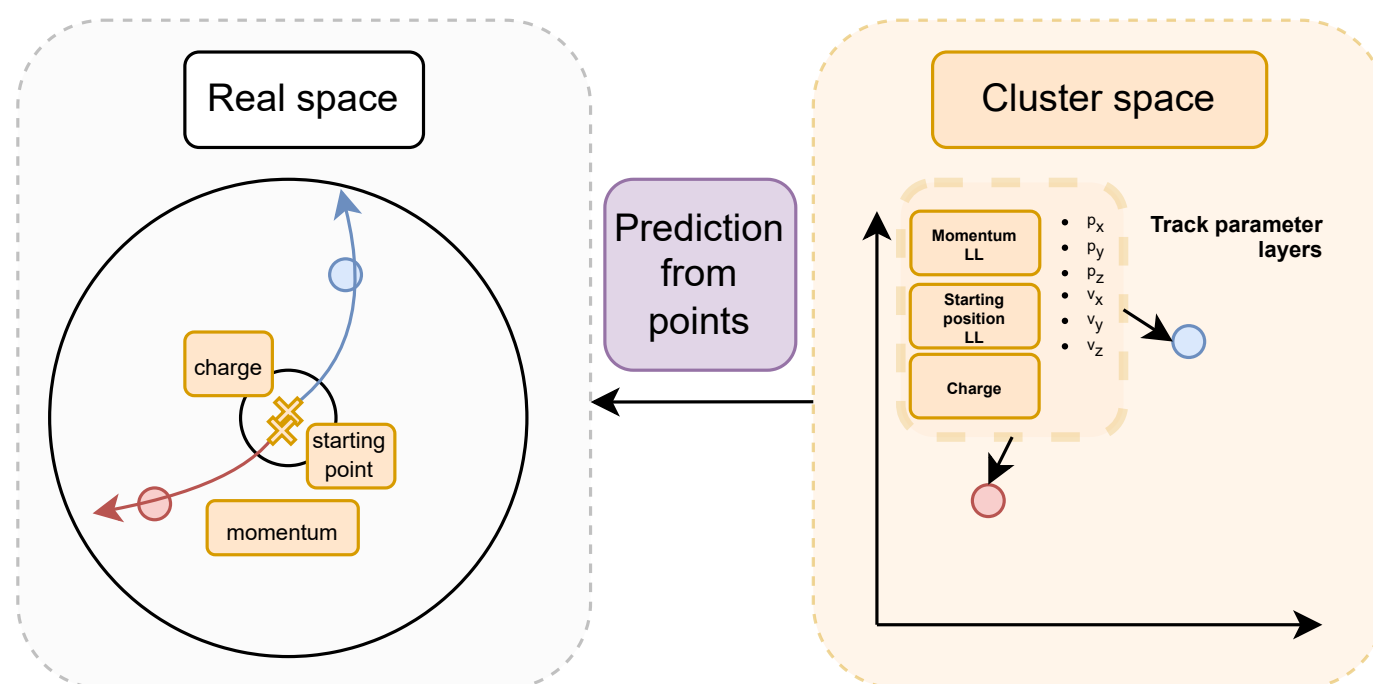
a) translation into cluster space



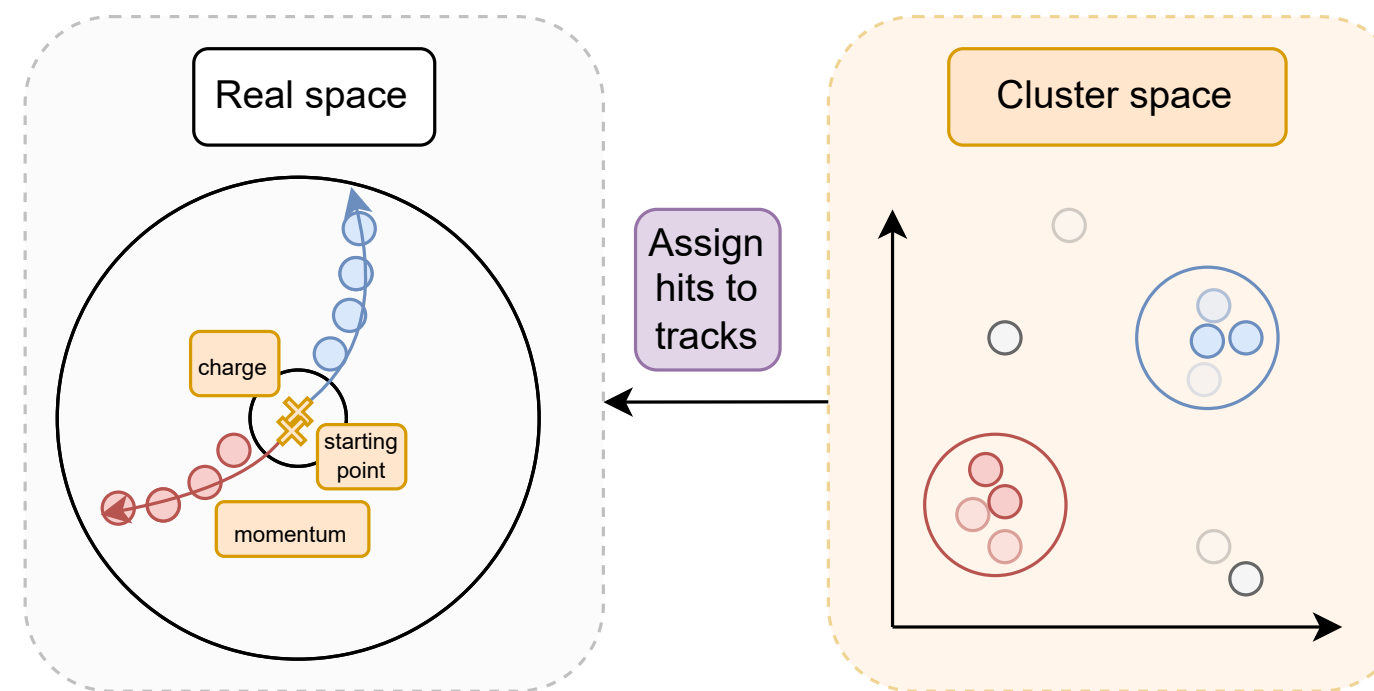
b) β threshold



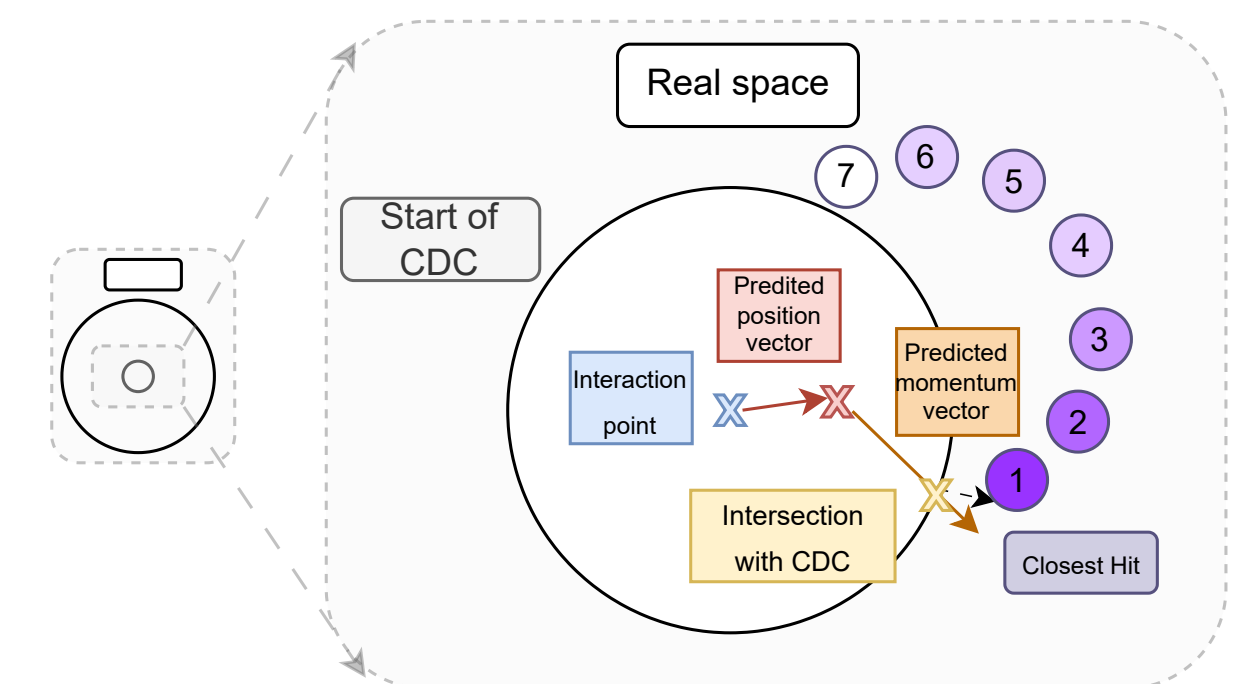
c) condensation point (CP) isolation



d) parameter extraction for CPs

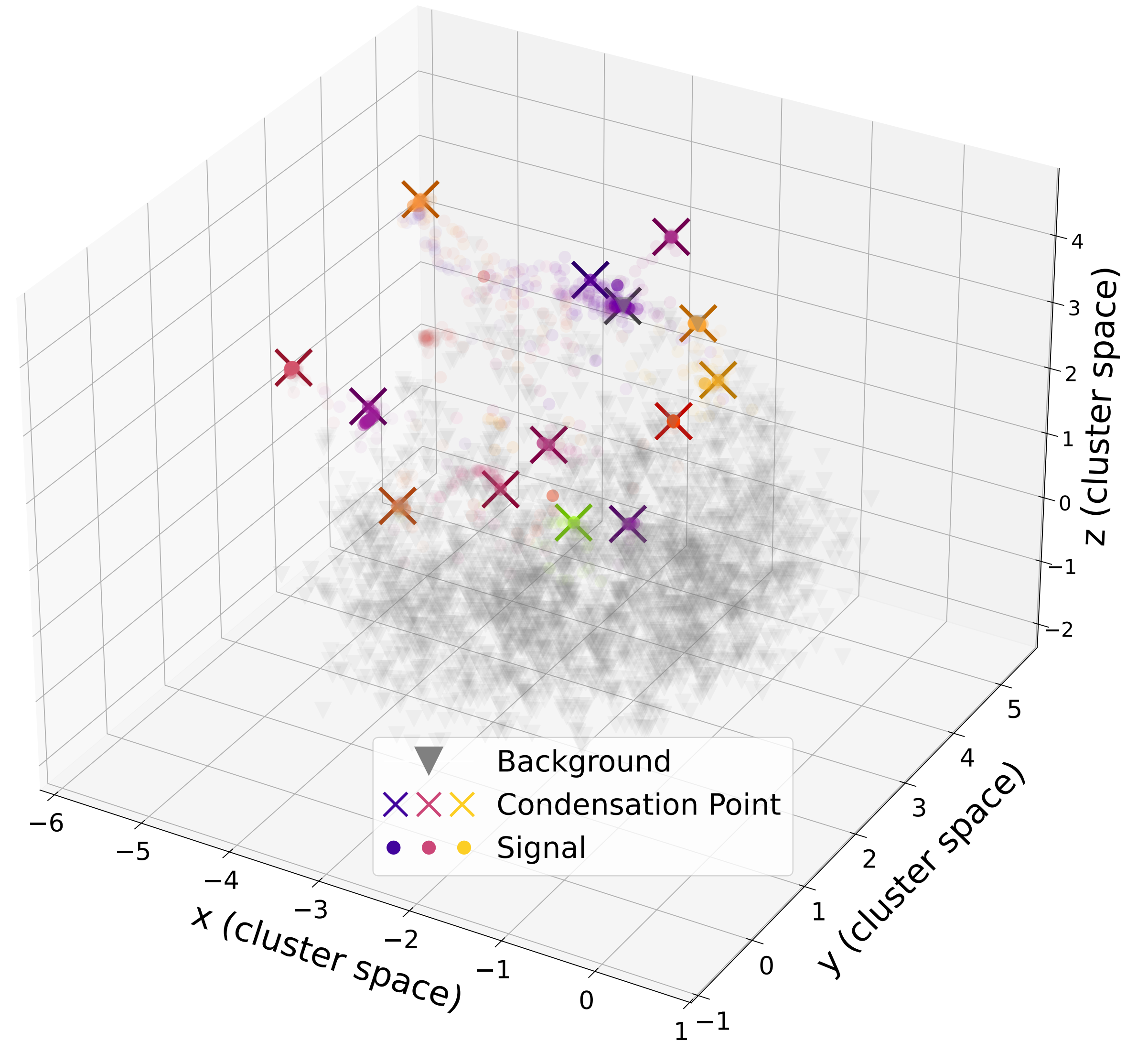
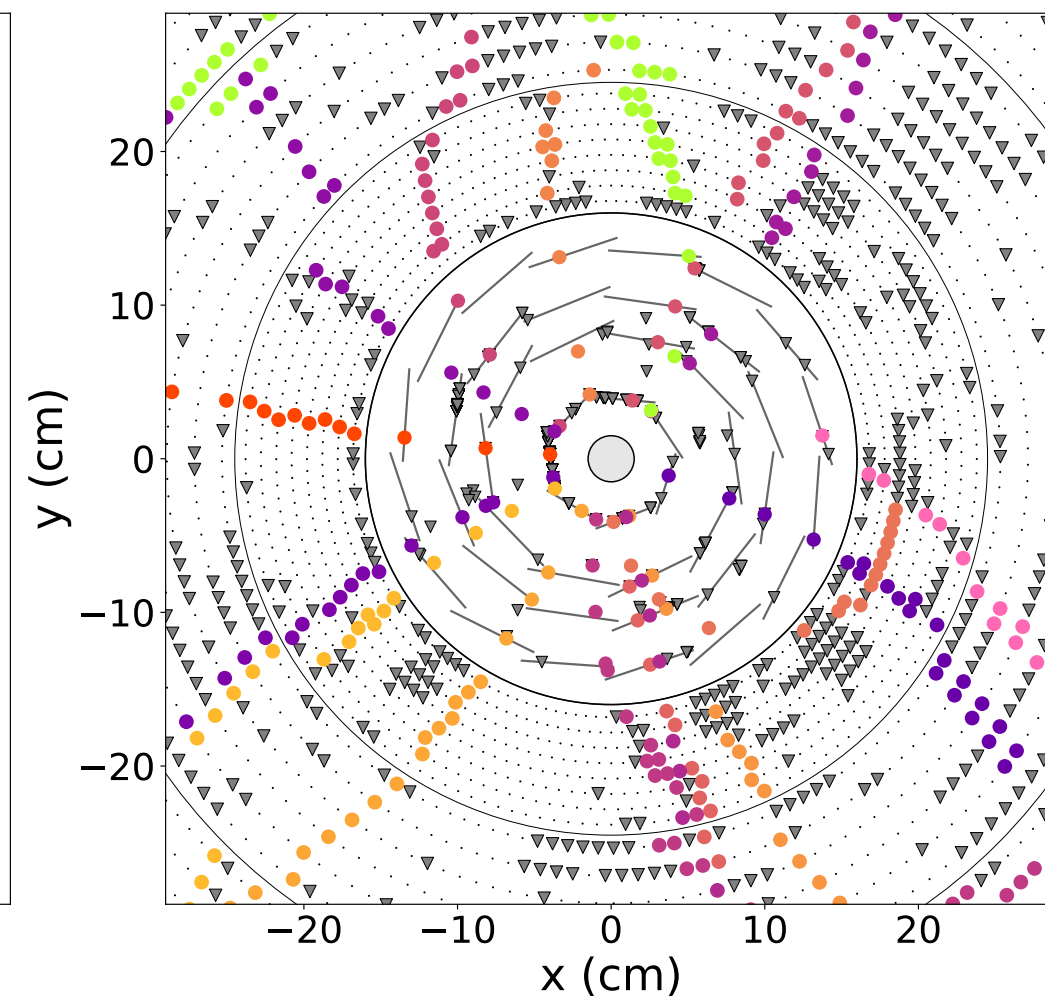
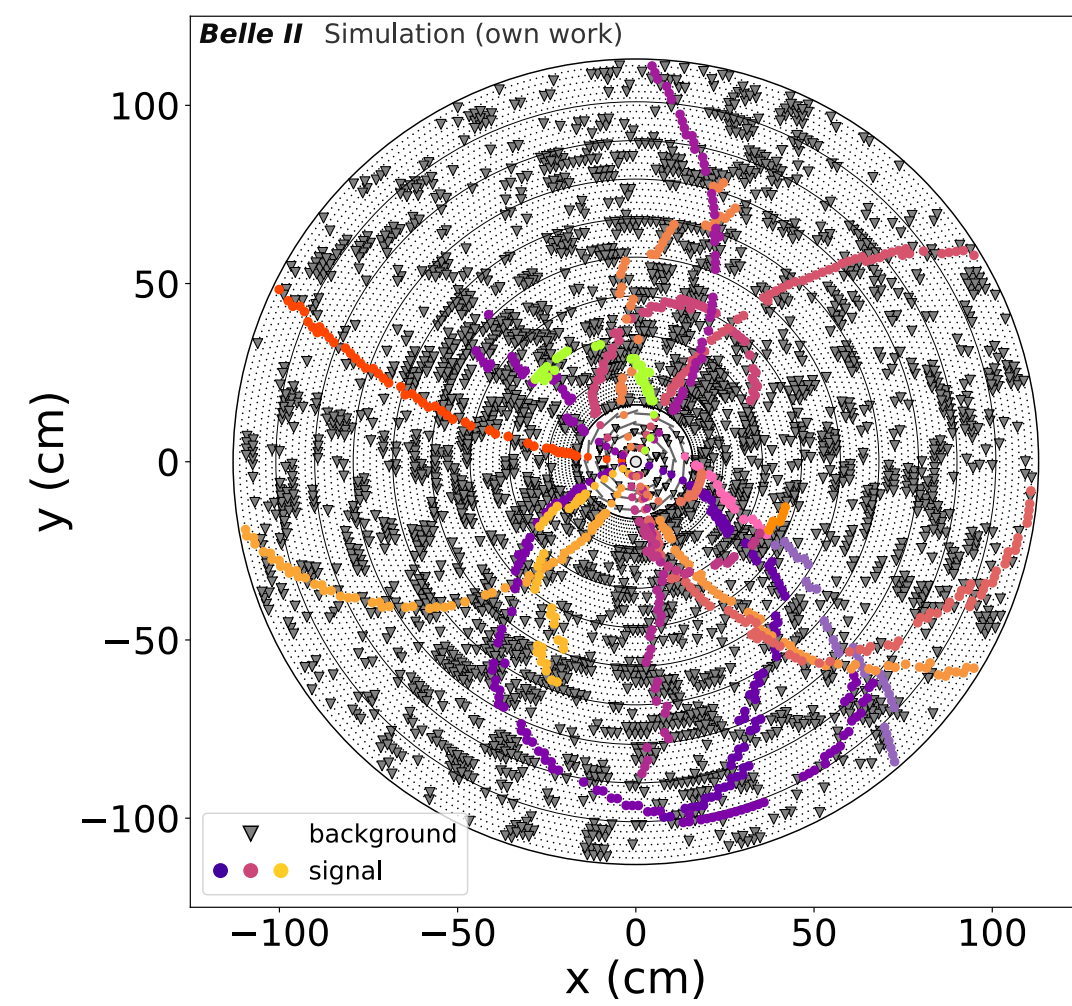


e) clustering around CPs



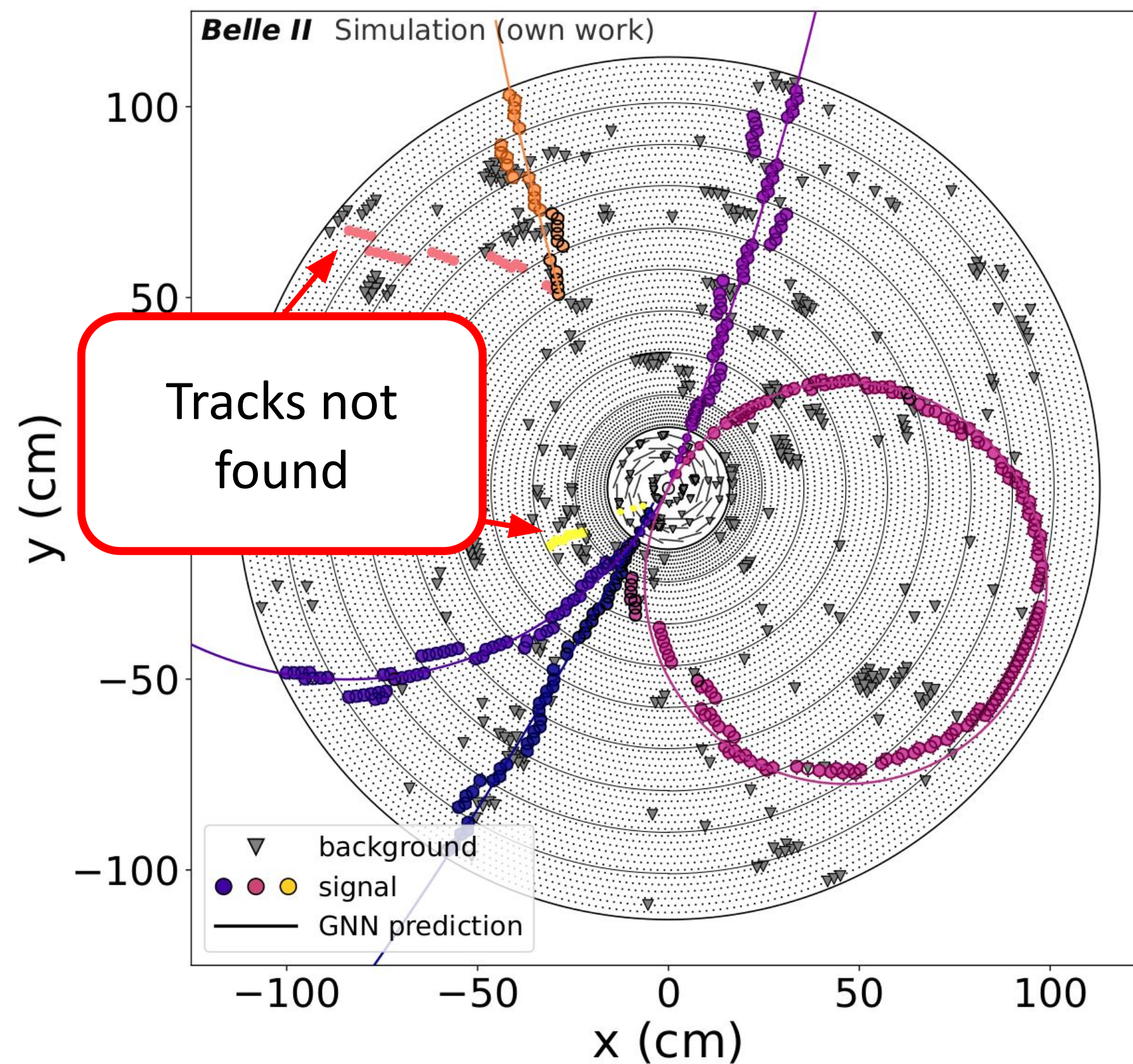
f) hit ordering

Learned cluster space

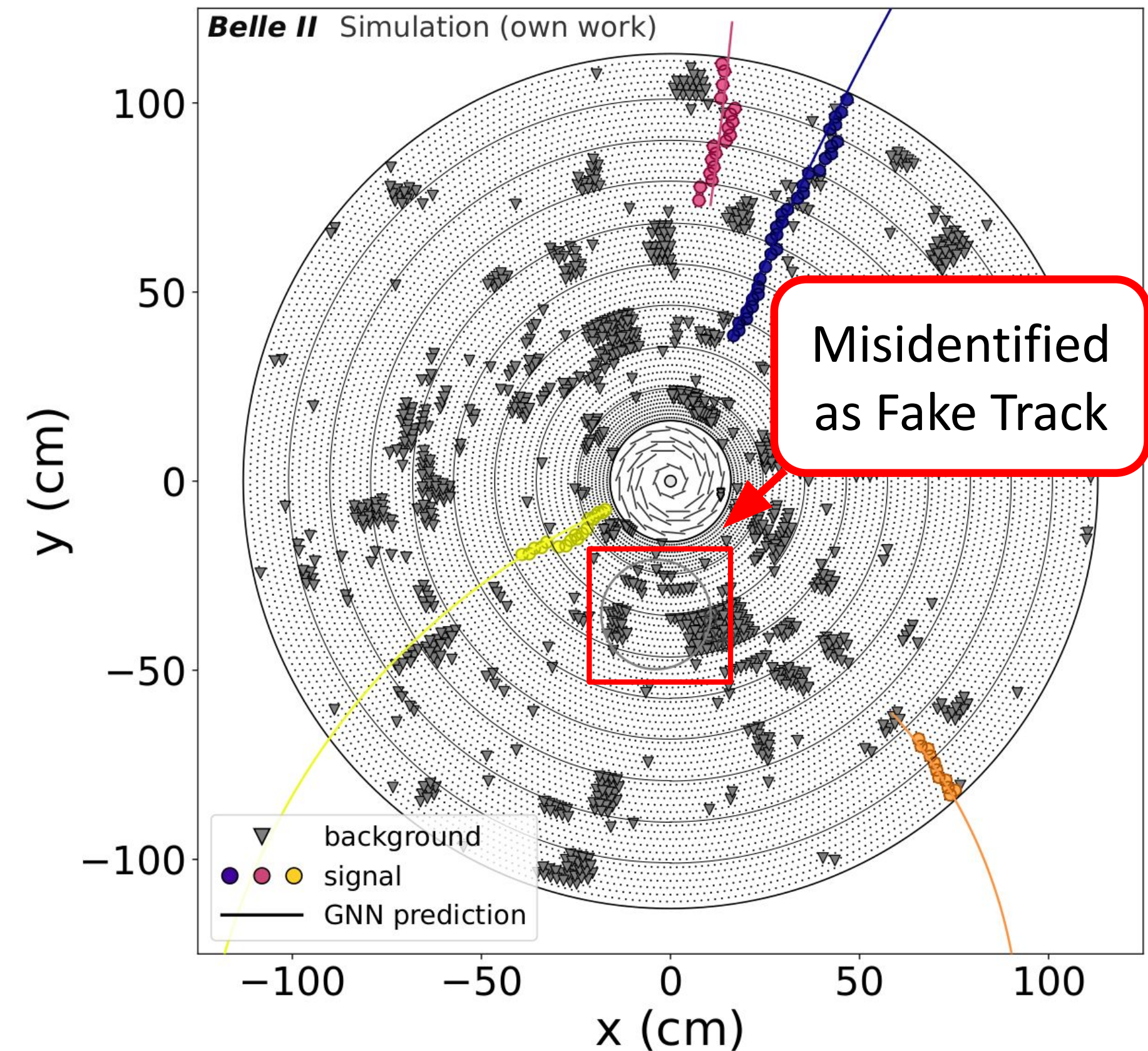


How good is tracking: Efficiency and fake tracks

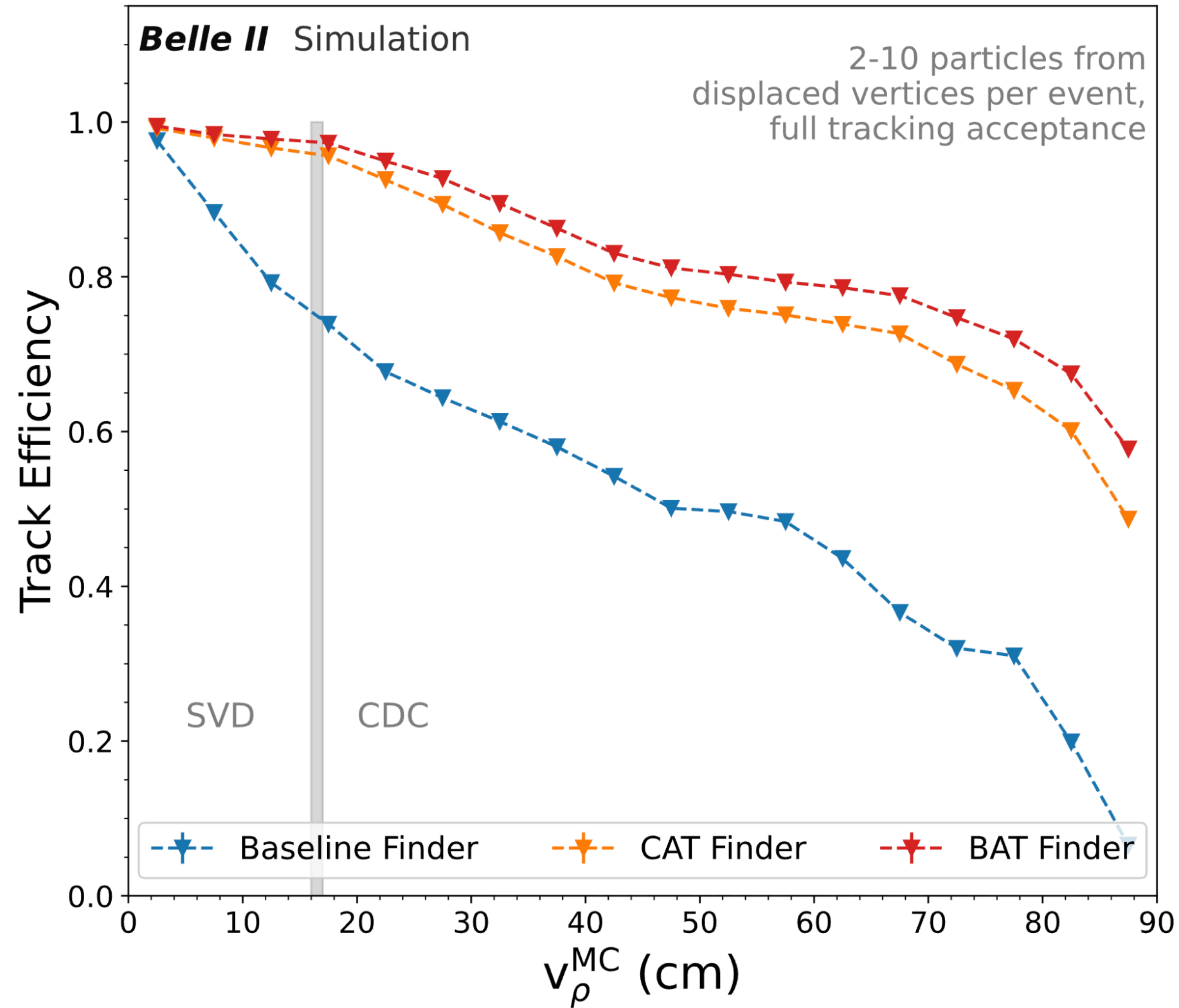
$$\text{Track Efficiency} = \frac{\text{true tracks}}{\text{all particles}} = \frac{5}{7}$$



$$\text{Track Purity} = \frac{\text{true tracks}}{\text{all tracks}} = \frac{4}{5}$$

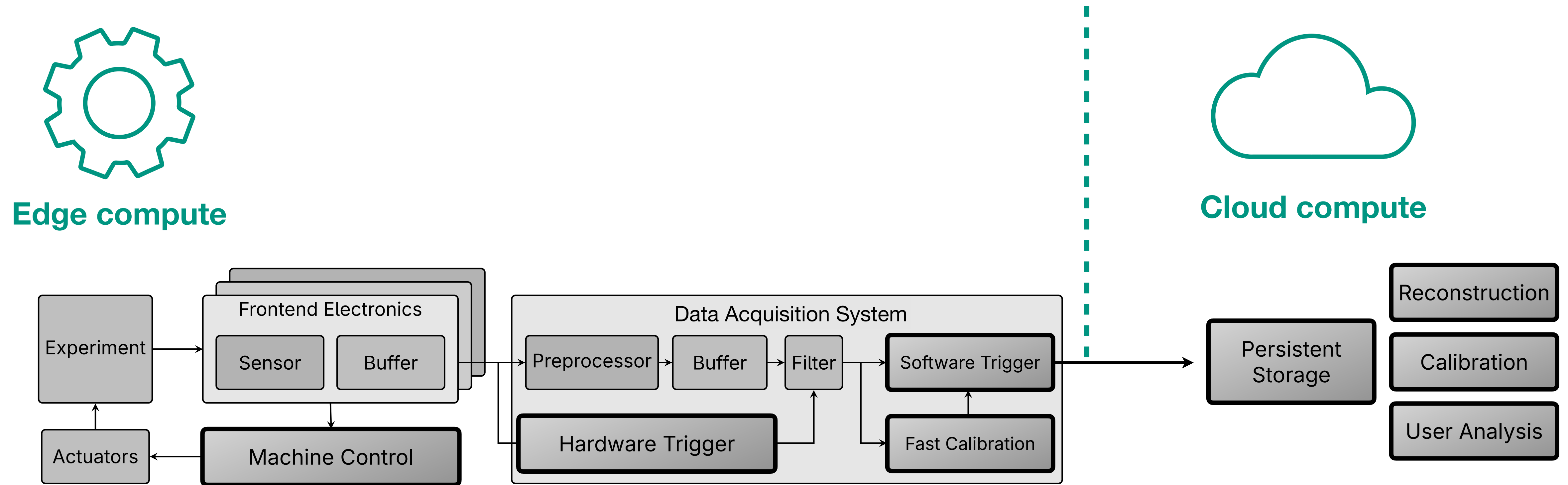


Results

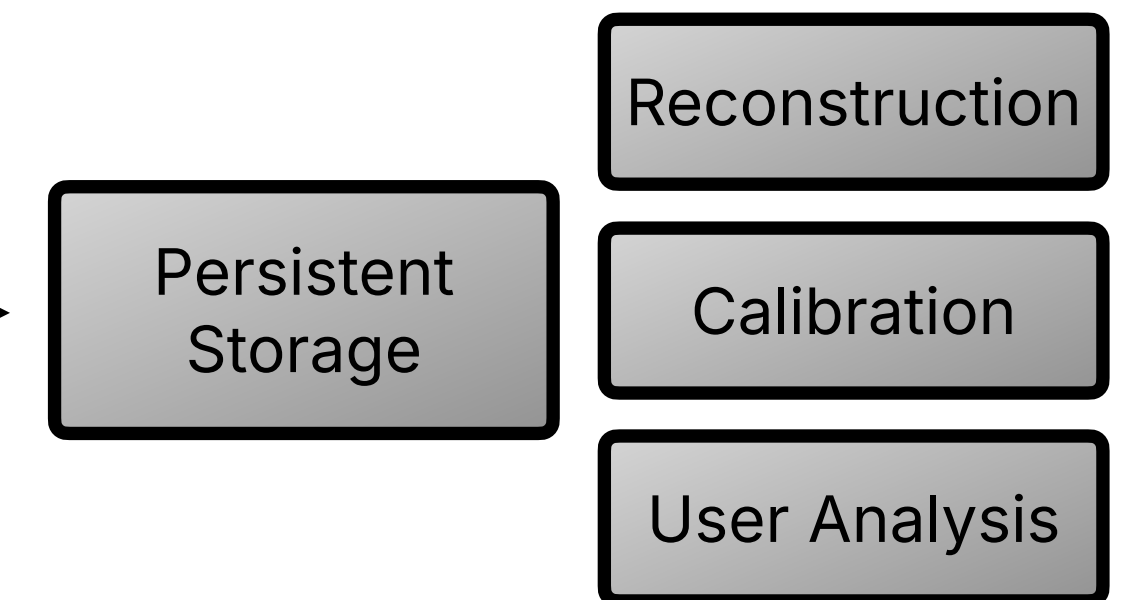
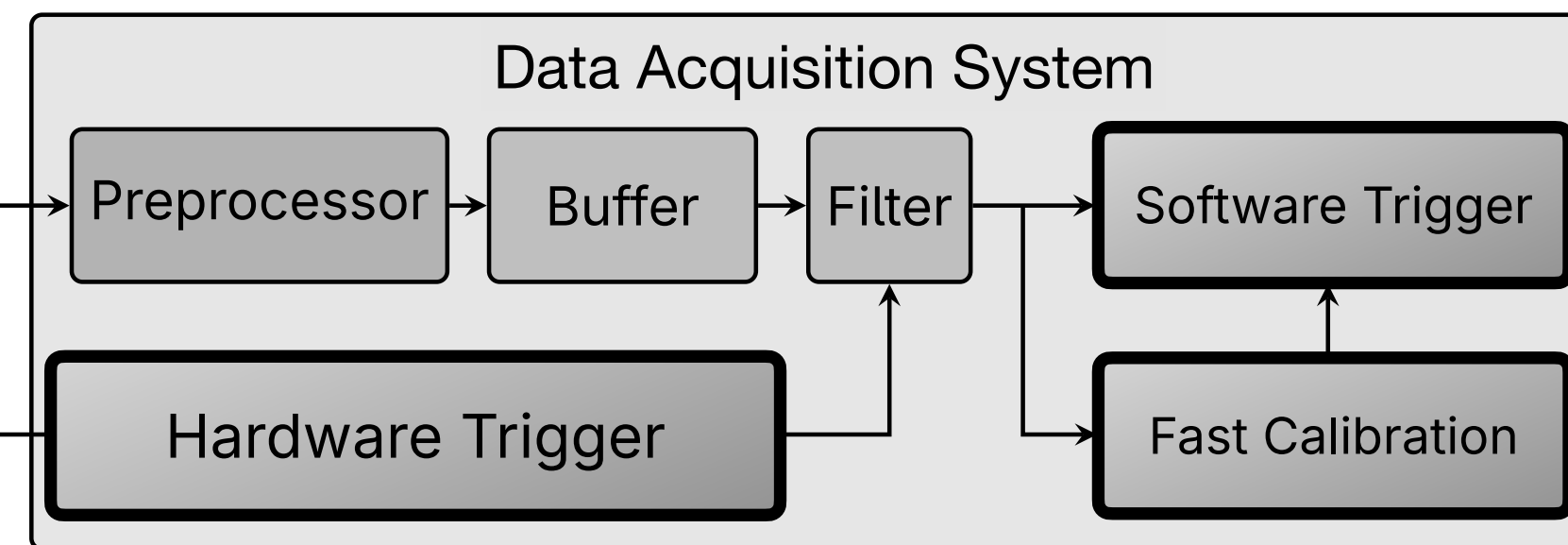
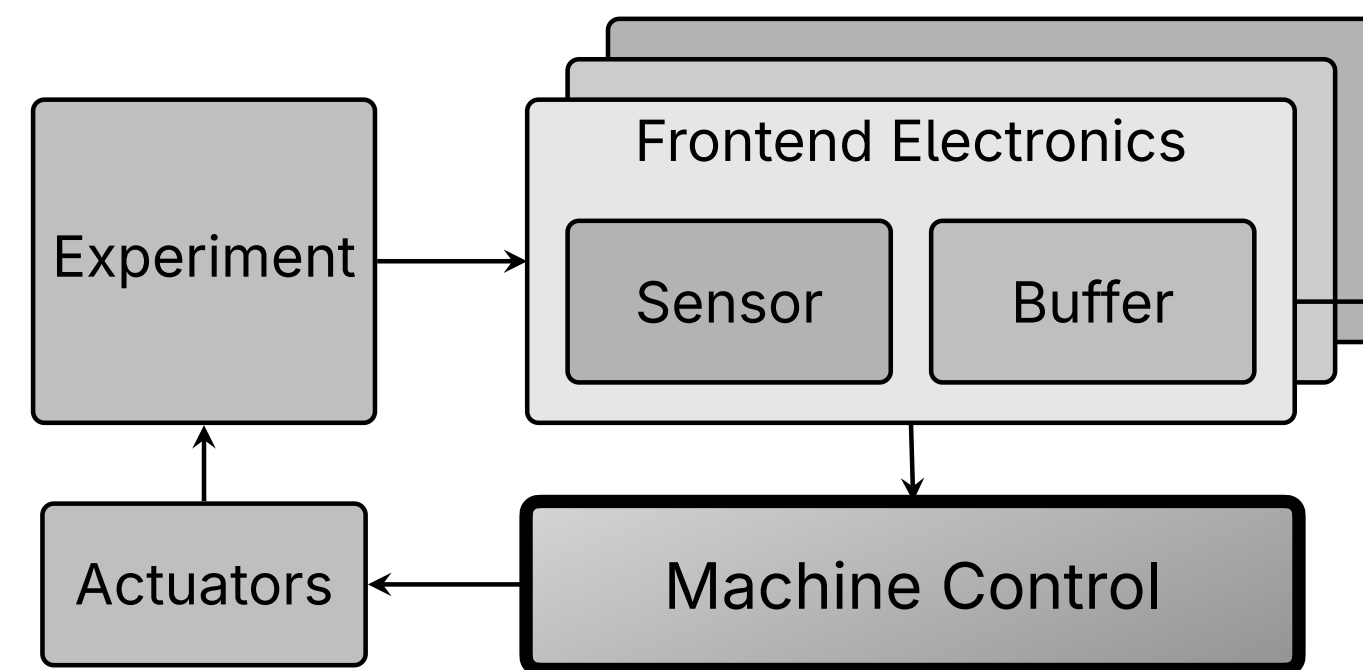
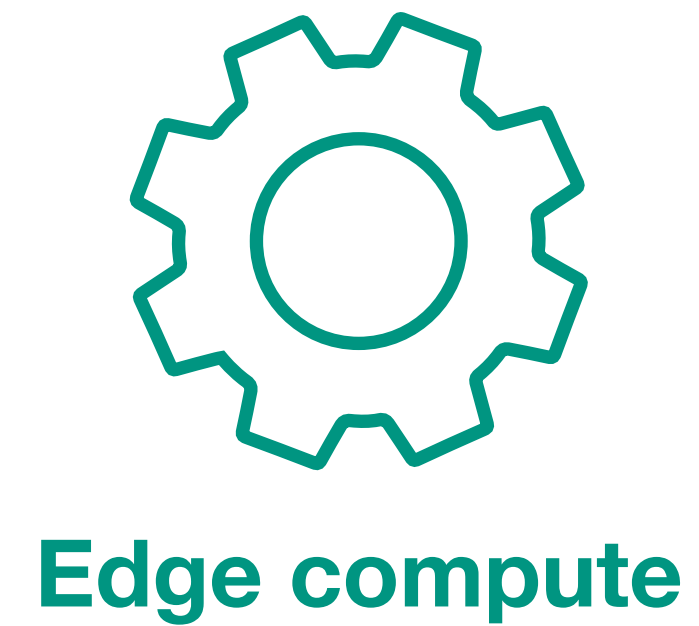


L. Reuter et al. <https://arxiv.org/abs/2602.10897>

Where are computations done?



How fast are computations done?



**Hard latency
constraint**

Miss the deadline and the system fails

**Soft latency
constraint**

Miss the deadline
occasionally and
you can recover

**No latency
constraint**

You can always recover

When must you compute at the edge?

- **Buffer or control-loop limits:** Not enough time to round-trip (transfer data to cloud → process → transfer back)
- **Security concerns:** cloud-connection offers attack vector
- **Safety concerns:** loss of cloud-connection result in system failure
- **No high-bandwidth link available**



Brain-computer interfaces



Radar



Self-driving cars



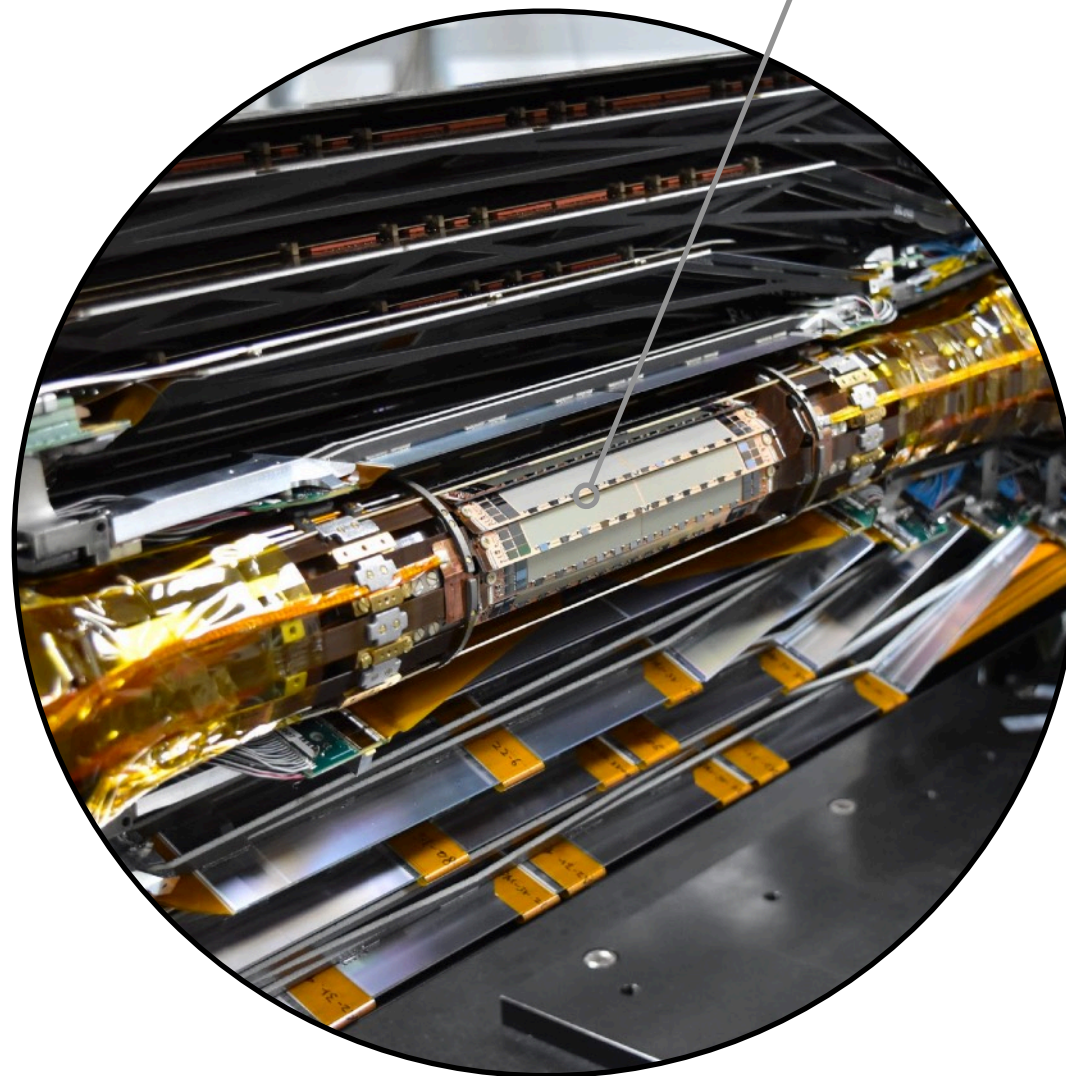
5G communication



Finance



Drones



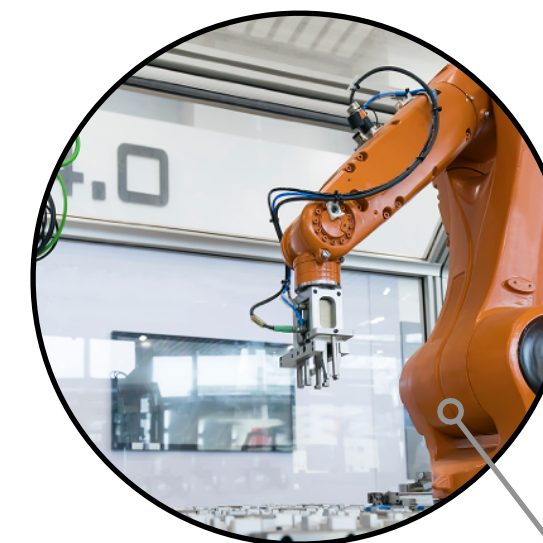
Particle physics experiments



Medical imaging



Accelerators



Industry 4.0

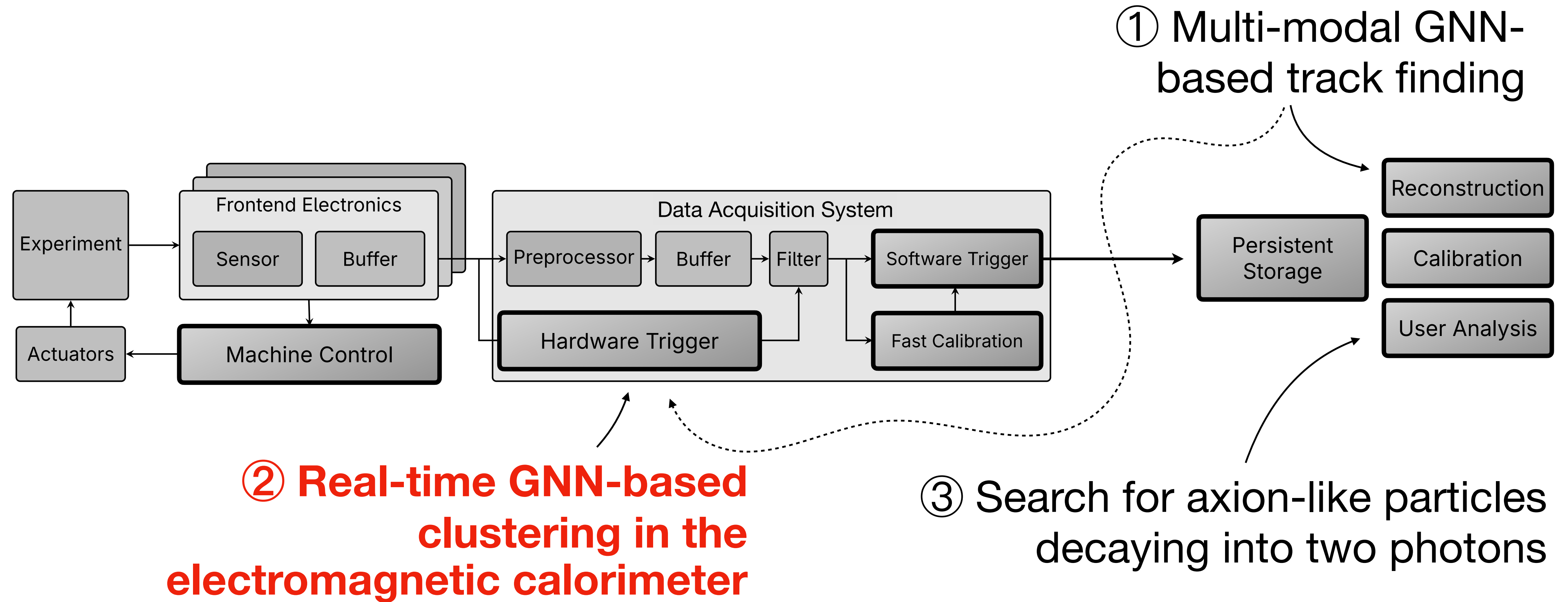


Encryption and Decryption

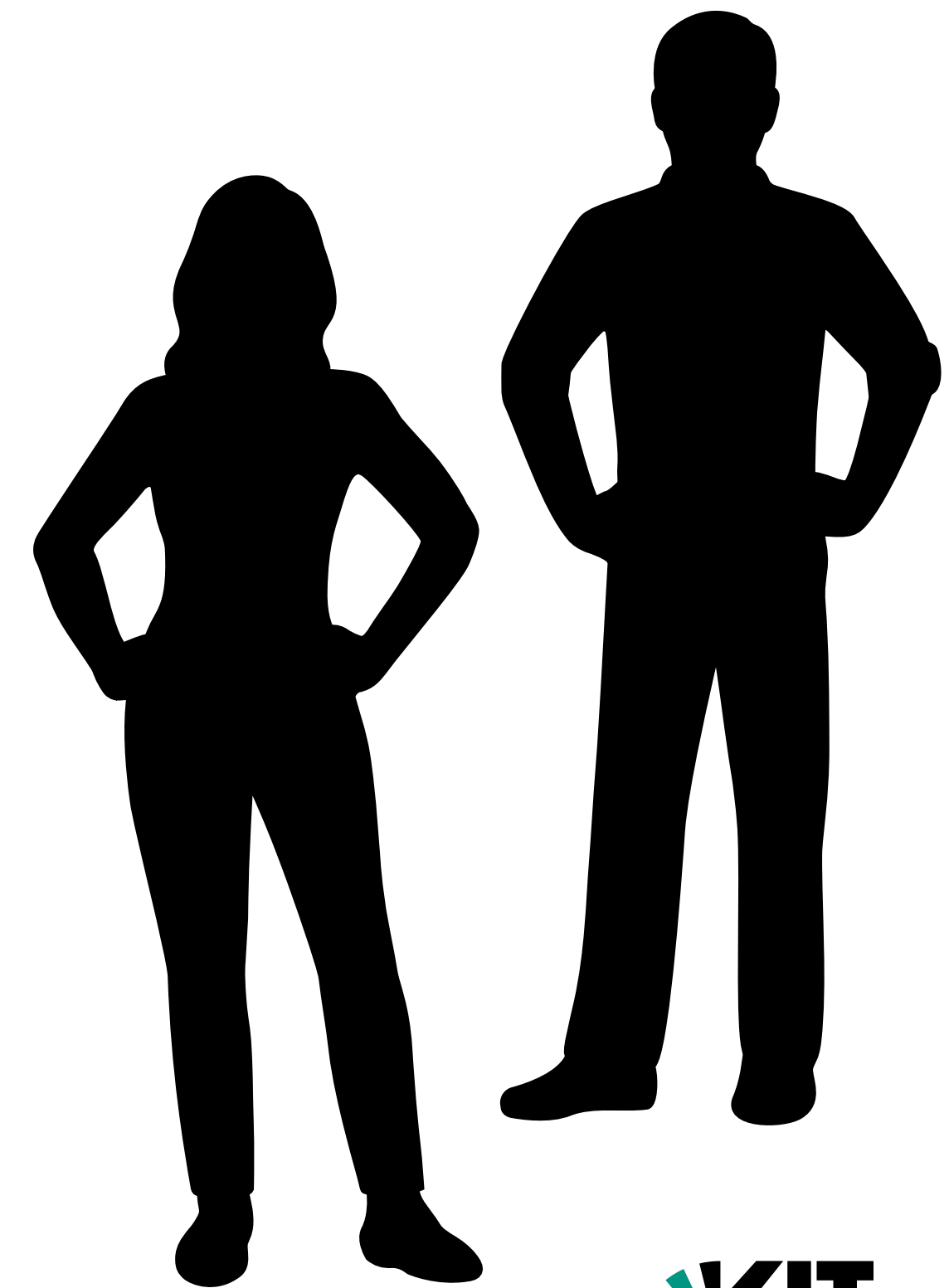
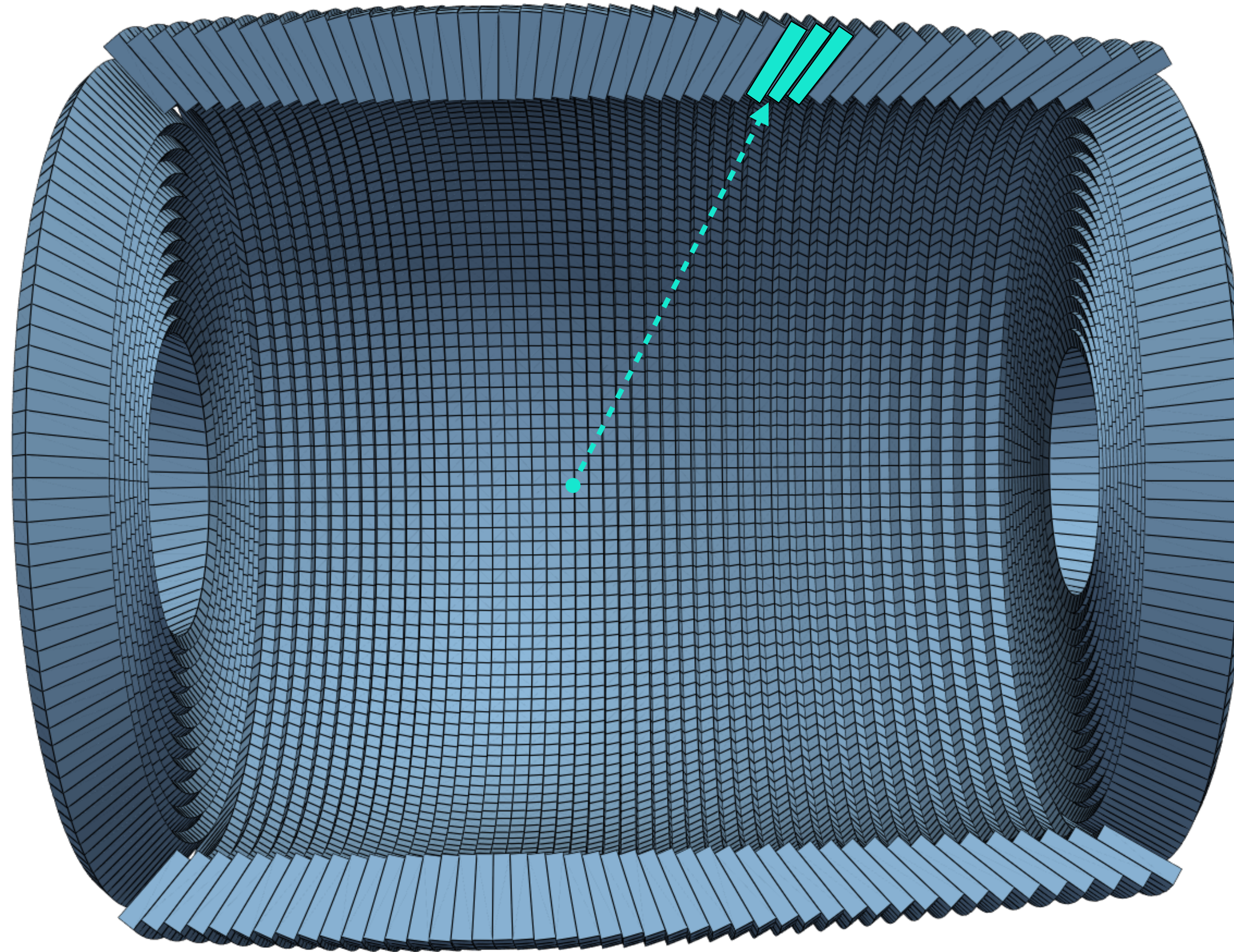


Robots

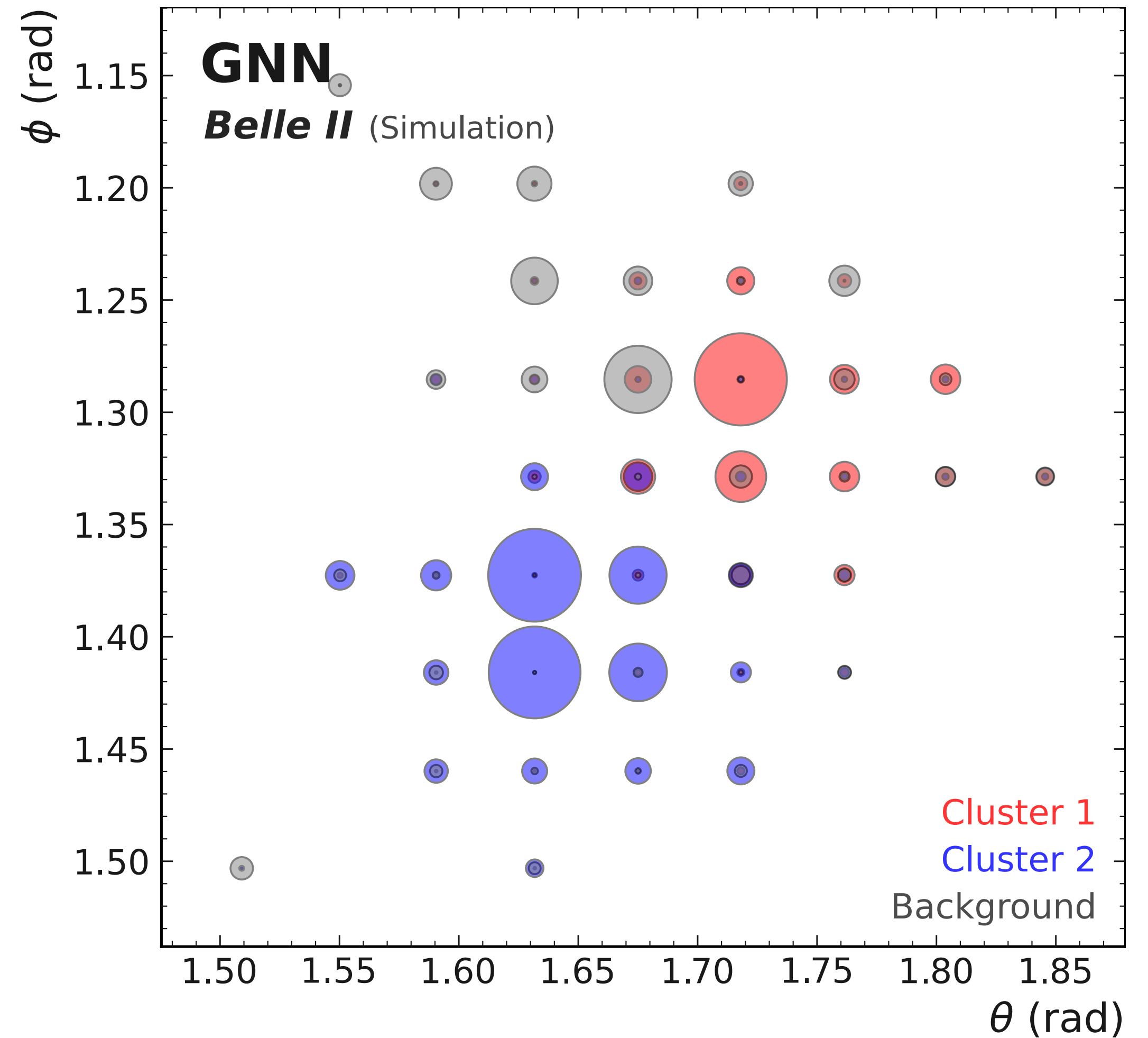
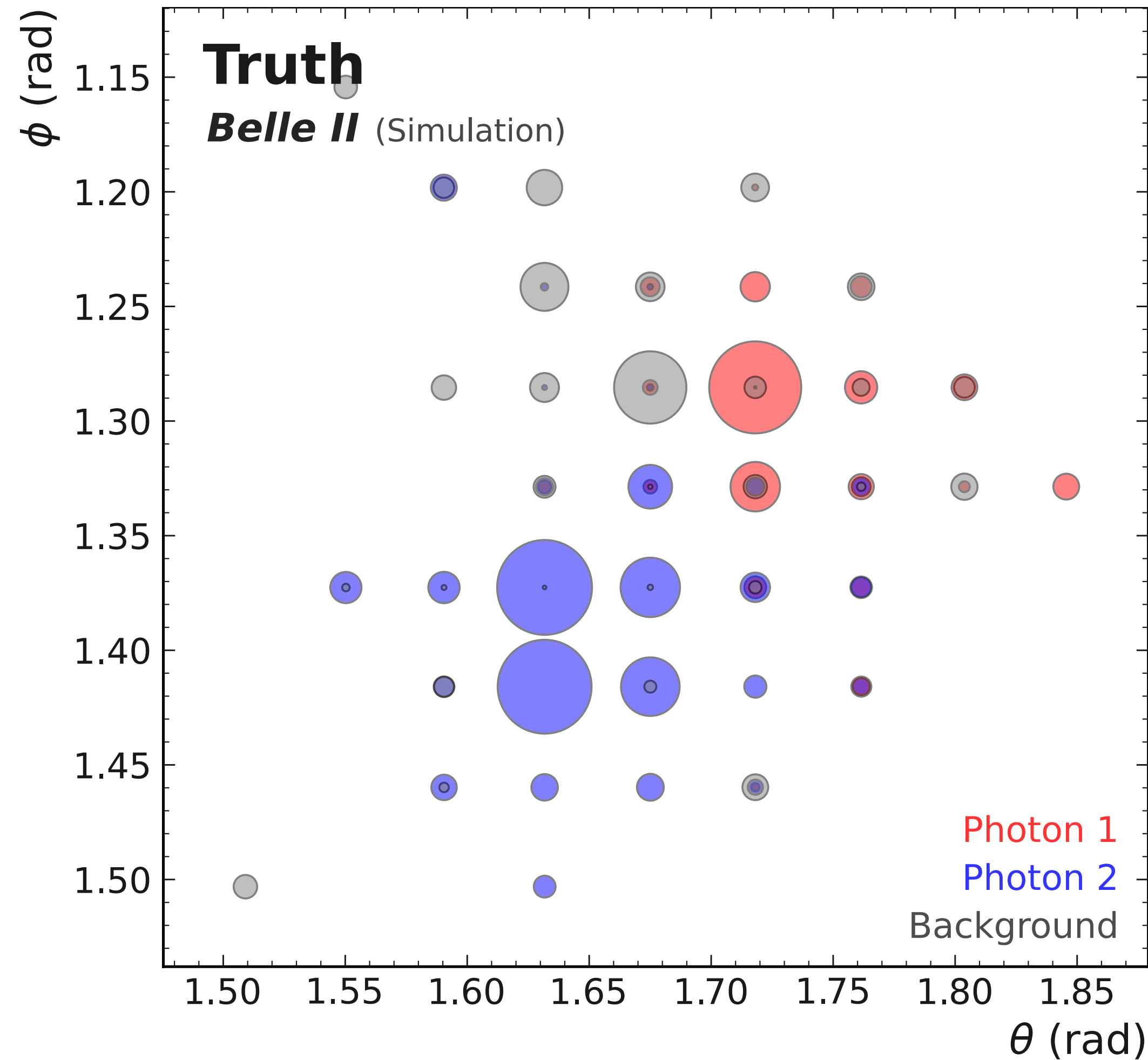
Outline



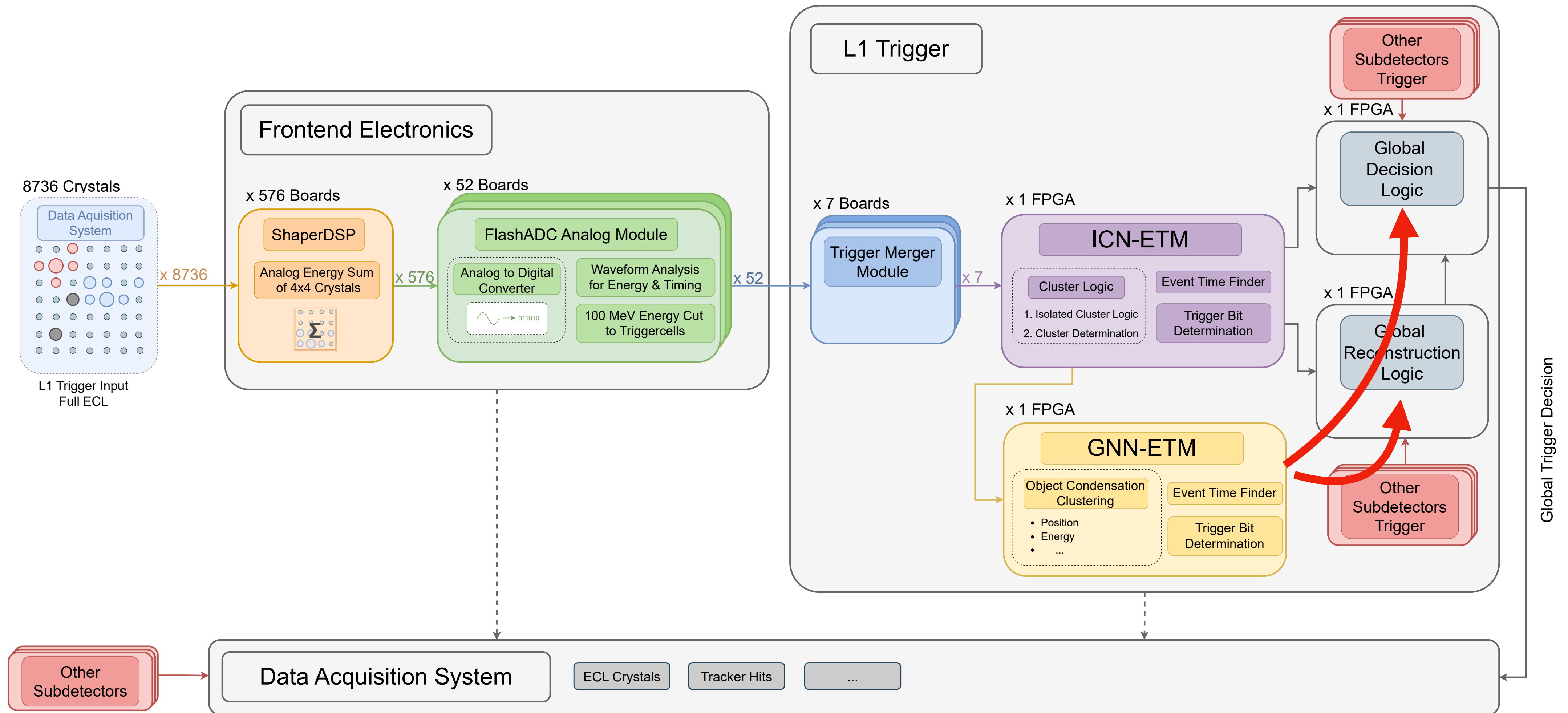
Neutral particles in a calorimeter



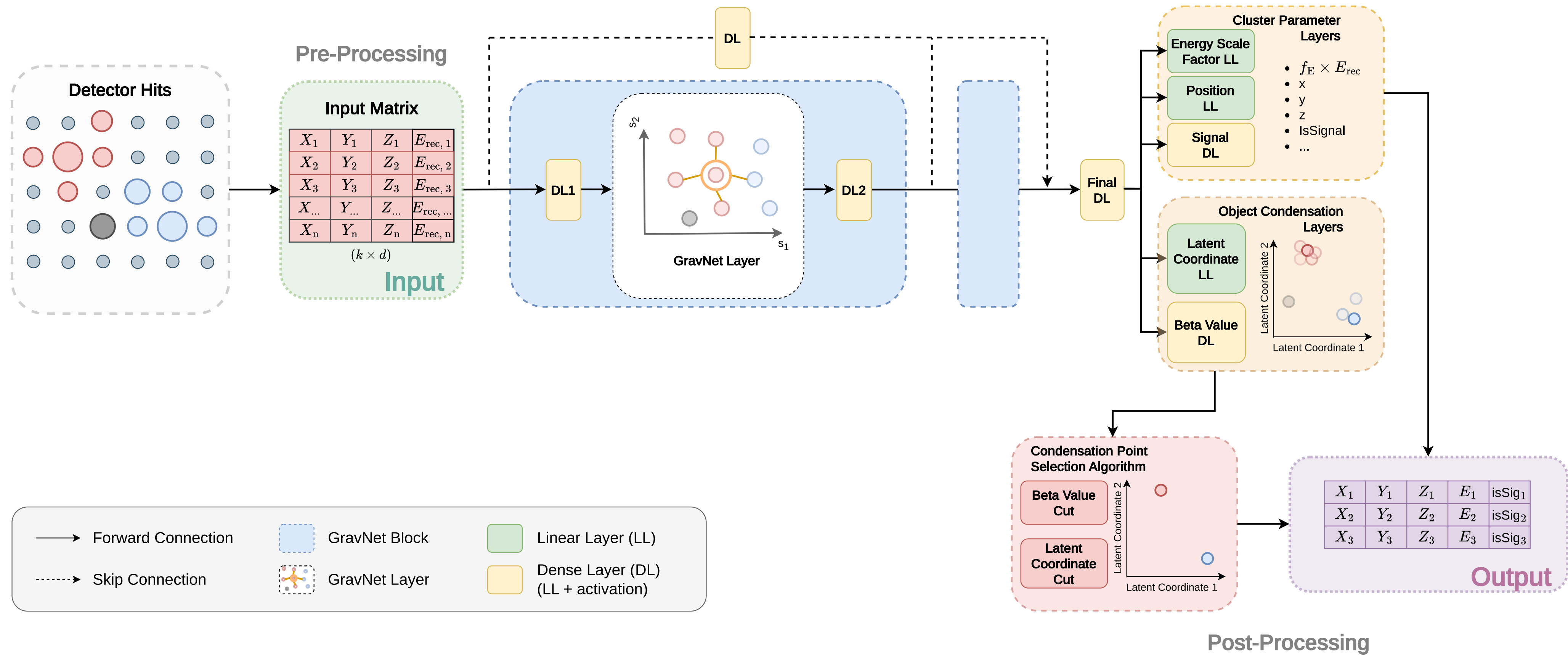
Neutral particle reconstruction: Clustering

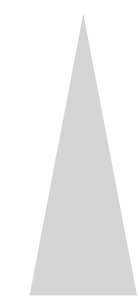


The Belle II Electromagnetic calorimeter trigger

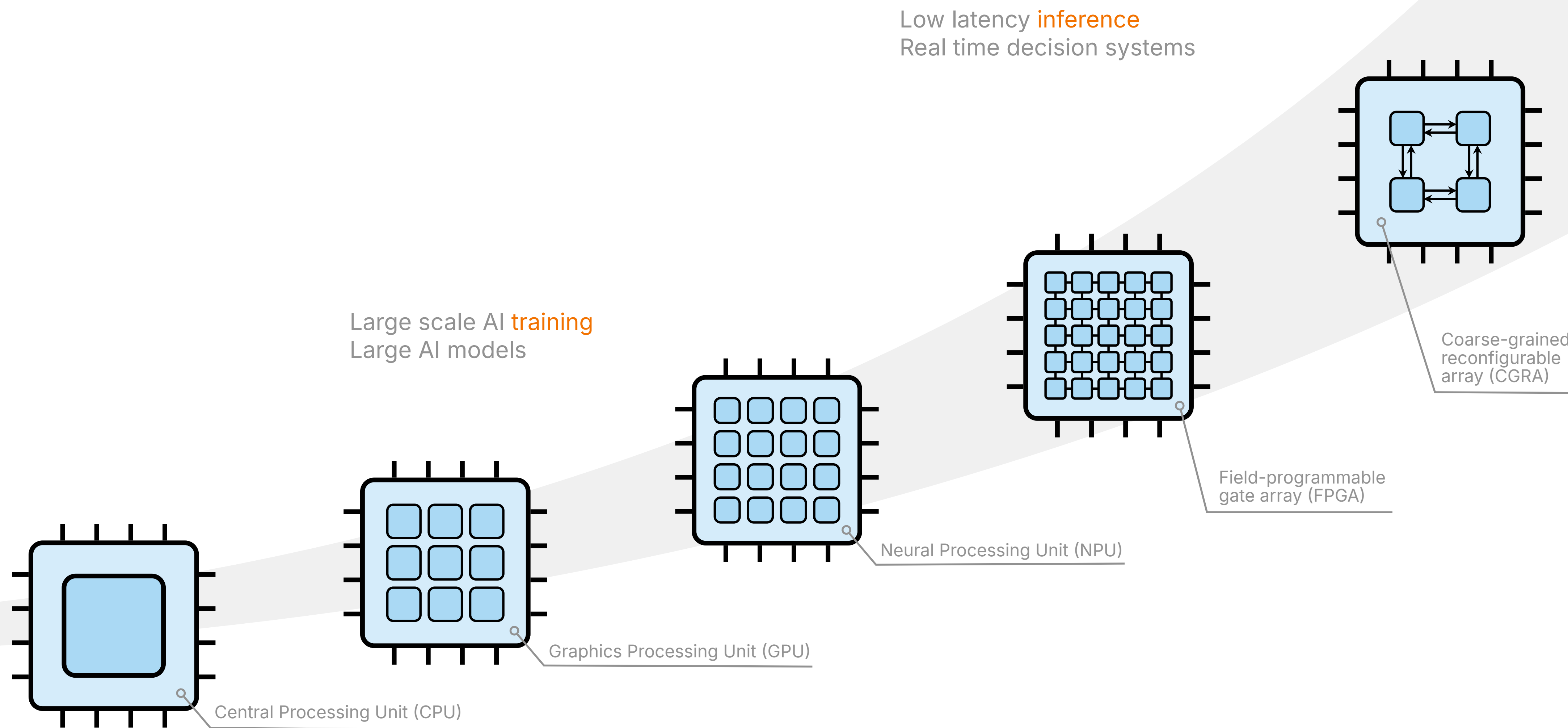


Realtime GNN-based clustering: Armadillo





Real-time performance

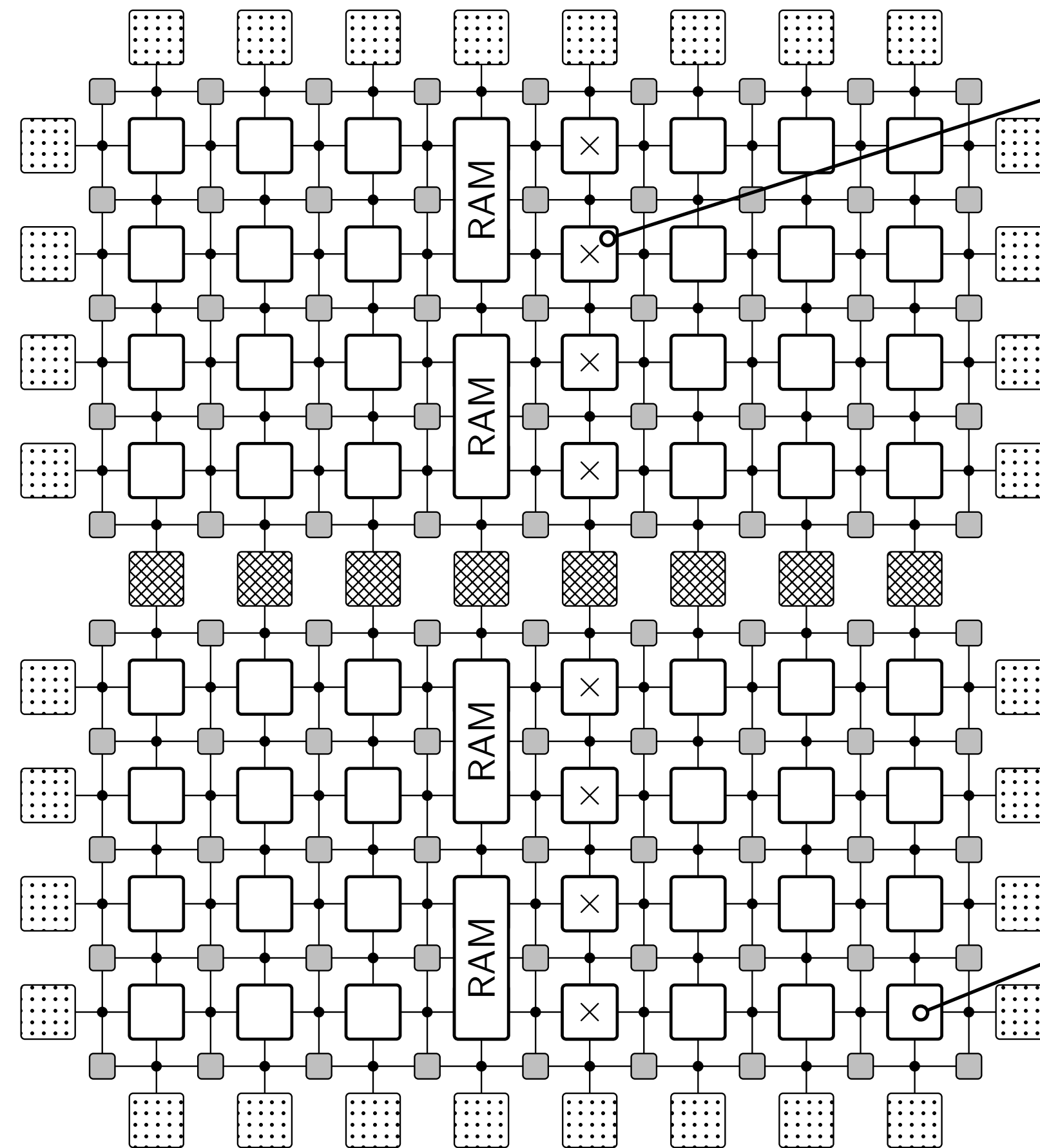


General-purpose computing



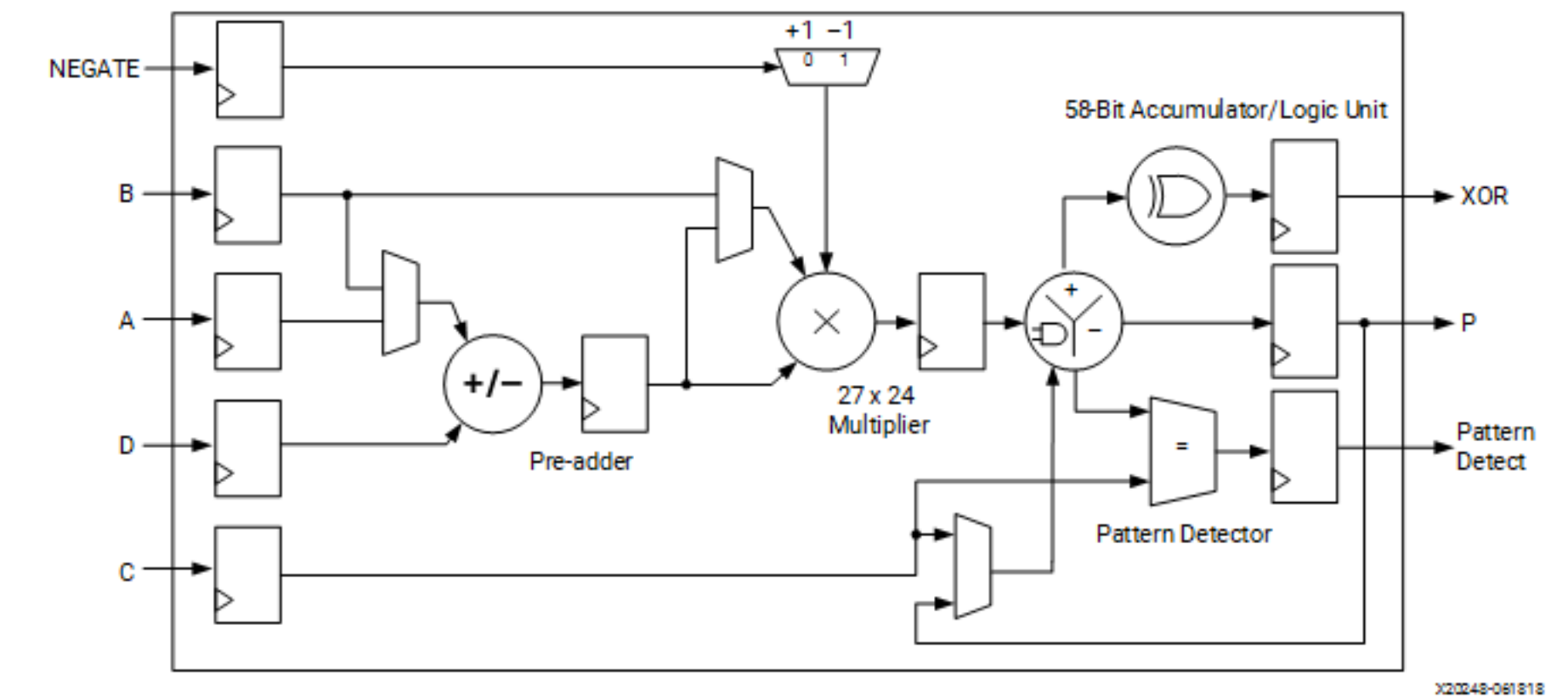
Specialised hardware

How does an FPGA work? (simplified)



DSP (digital signal processor)

$$P = C \text{ (or } P) + (A \pm B) \times D$$

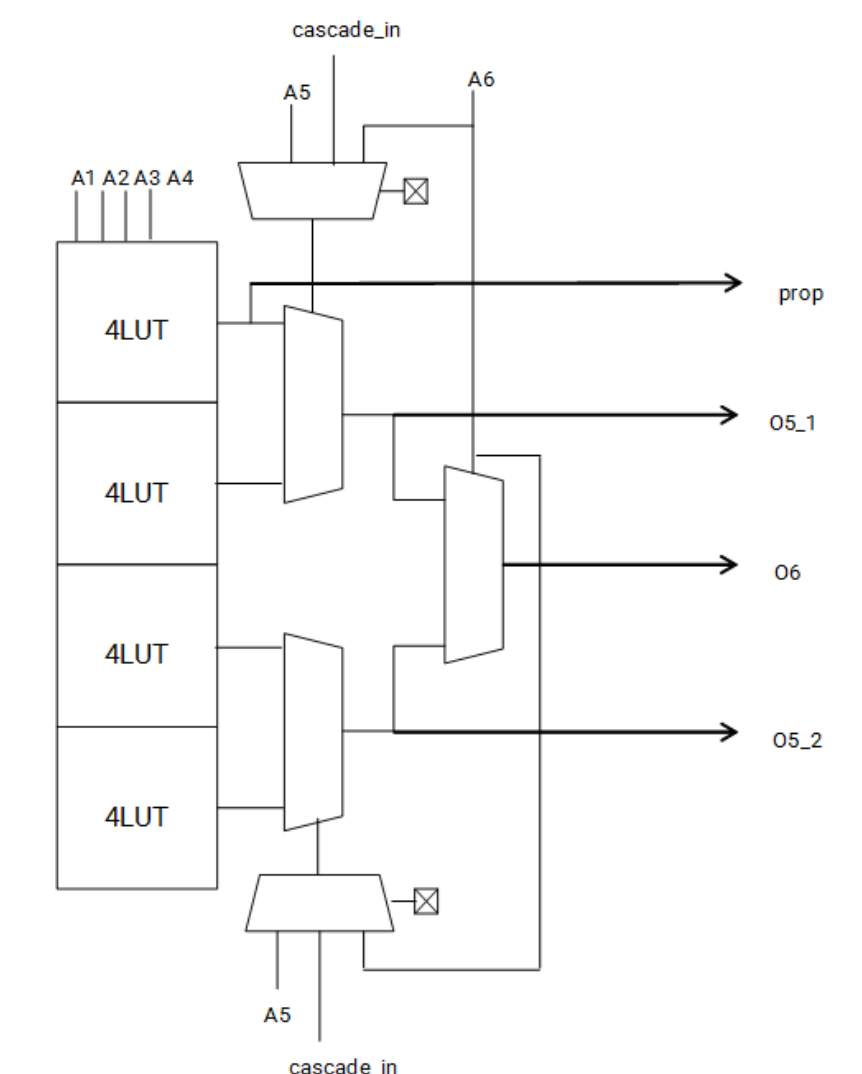


LUT (lookup table)

$$O6 = f(I_0 \dots I_5), \text{ any Boolean } f$$

Example (2bit)

I0	I1	O
0	0	1
0	1	1
1	0	0
1	1	1



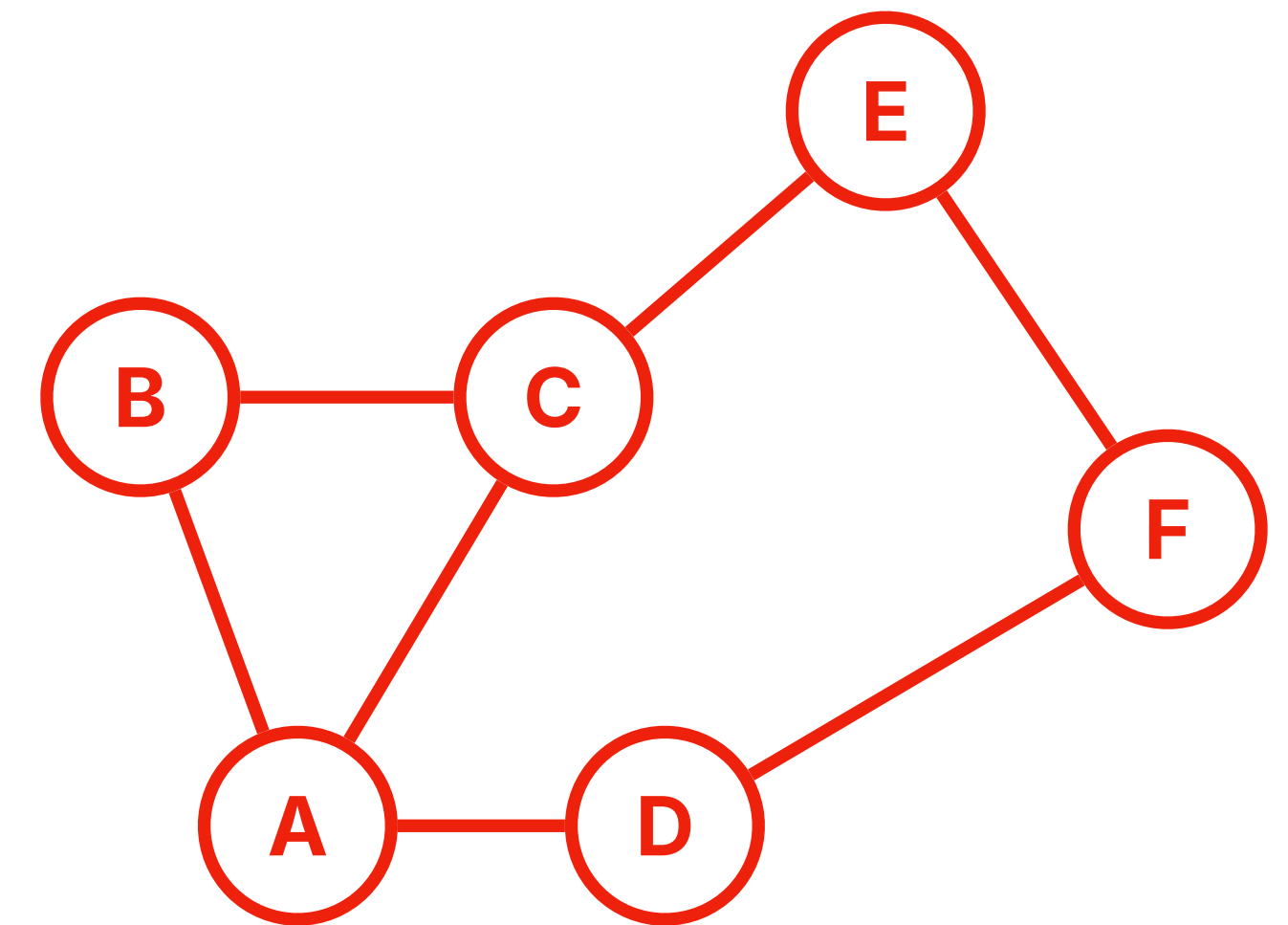
M. Neu PhD thesis

AMD <https://docs.amd.com/r/en-US/am005-versal-clb/IMUX-Register>

AMD <https://docs.amd.com/r/en-US/am004-versal-dsp-engine/DSP58-Architecture>

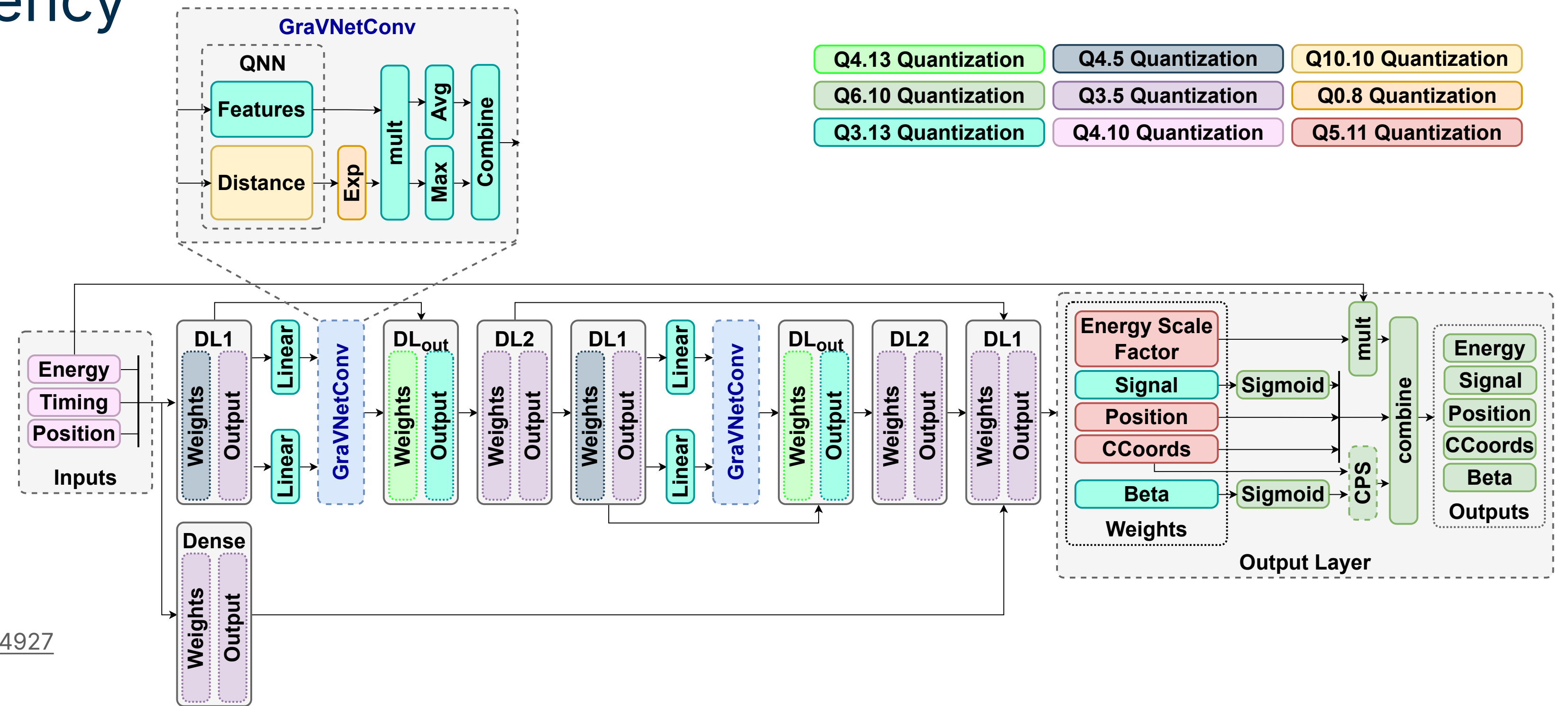
Why are dynamic GNNs so difficult on FPGAs?

- GNNs work by exchanging information between nodes along edges
- **A** *gathers* information from its neighbours **B**, **C**, and **D**, and performs an operation (e.g. summing) them and stores that result as new value of **A**
→ In **dynamic** GNNs neighbours of **A** are only known at *runtime*
- Either need to support a large number of concurrent random memory accesses or need to improve regularity in the memory accesses for efficient execution



Armadillo on an FPGA

- 32 non-zero inputs
- 8 Million events per second
- **3.168 μ s** end-to-end latency
- 100% uptime



I. Haide PhD thesis <https://publikationen.bibliothek.kit.edu/1000184927>

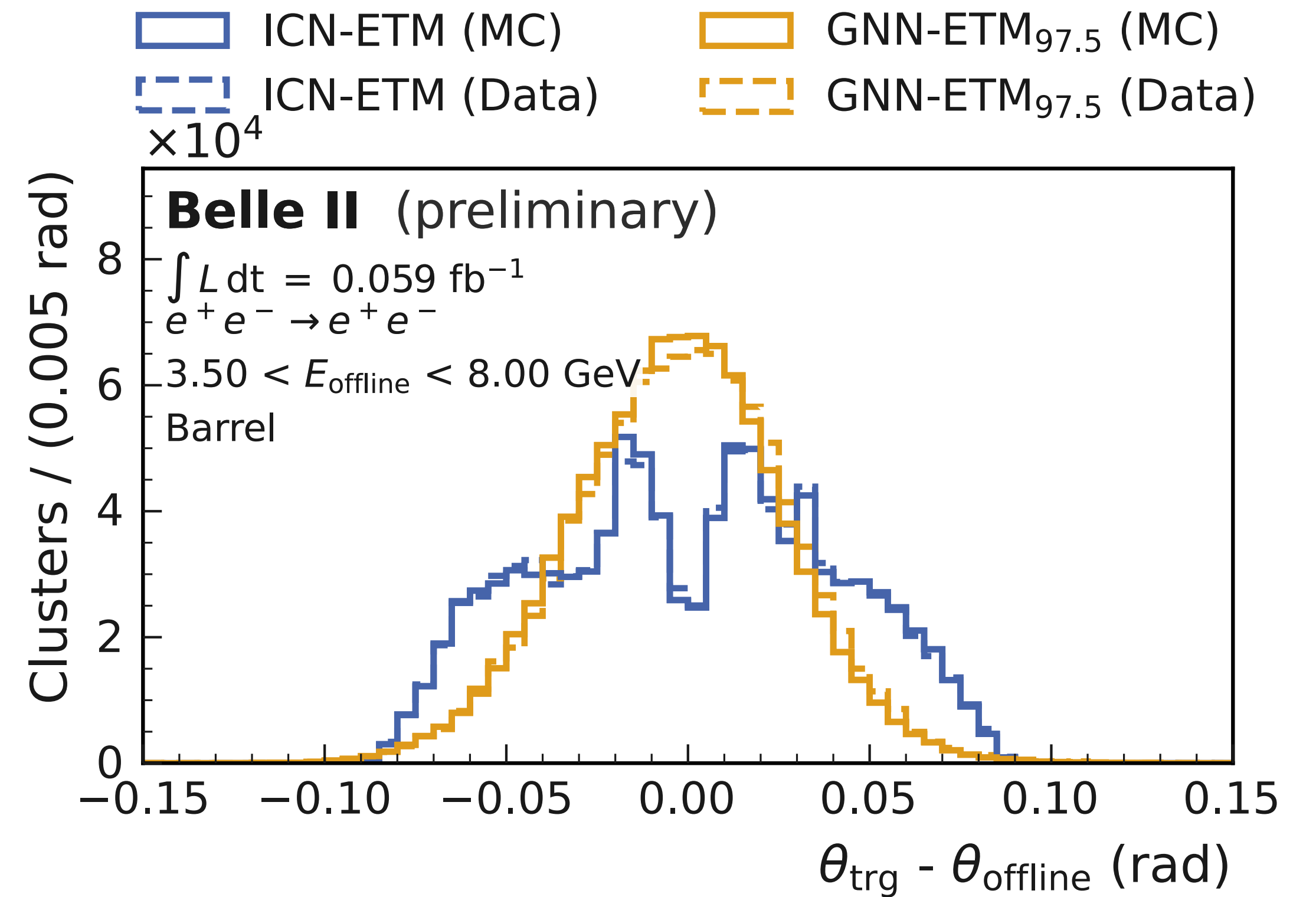
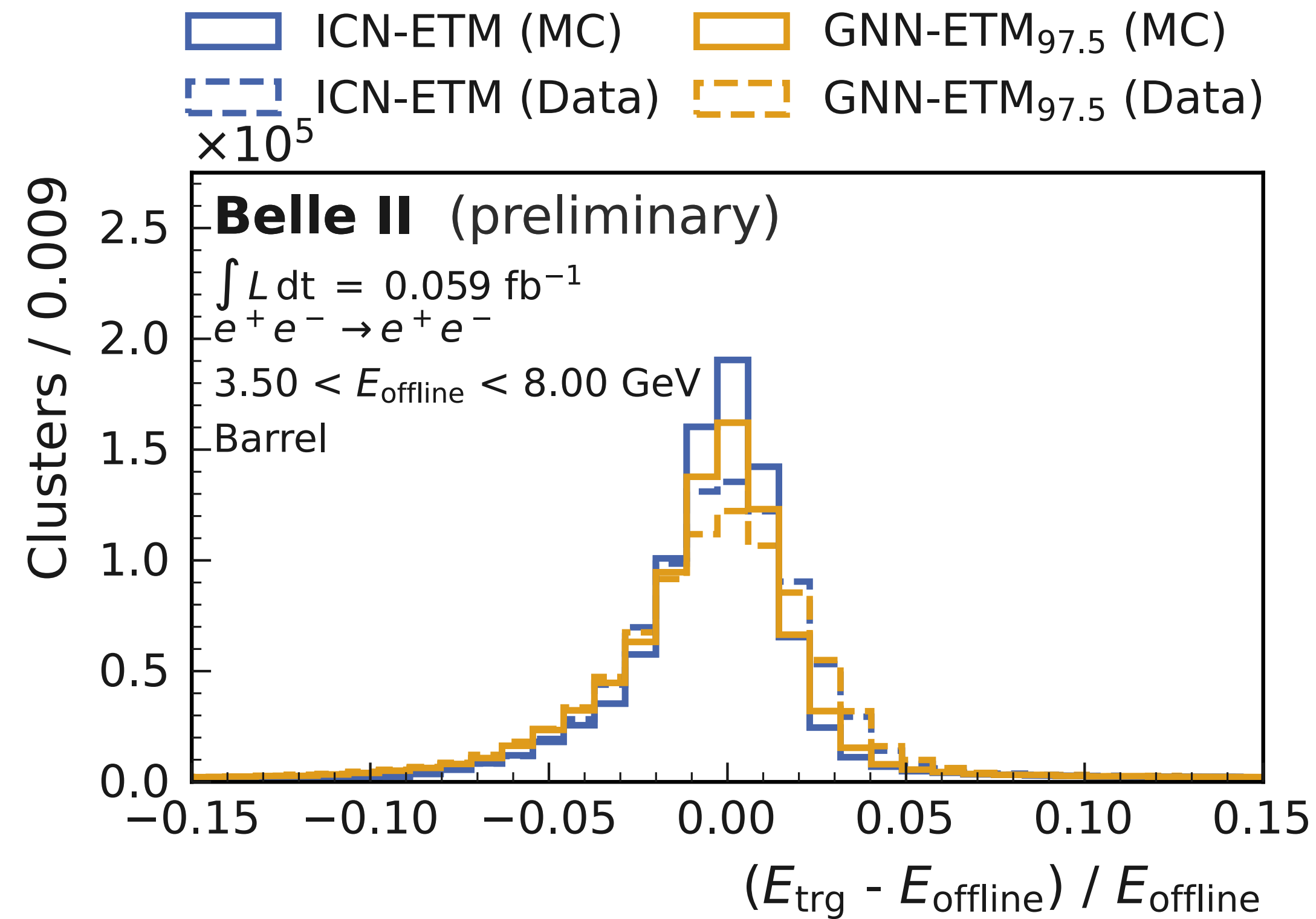
M. Neu PhD thesis

I. Haide et al. <https://arxiv.org/abs/2602.15118>

M. Neu et al. <https://arxiv.org/abs/2607.09347>

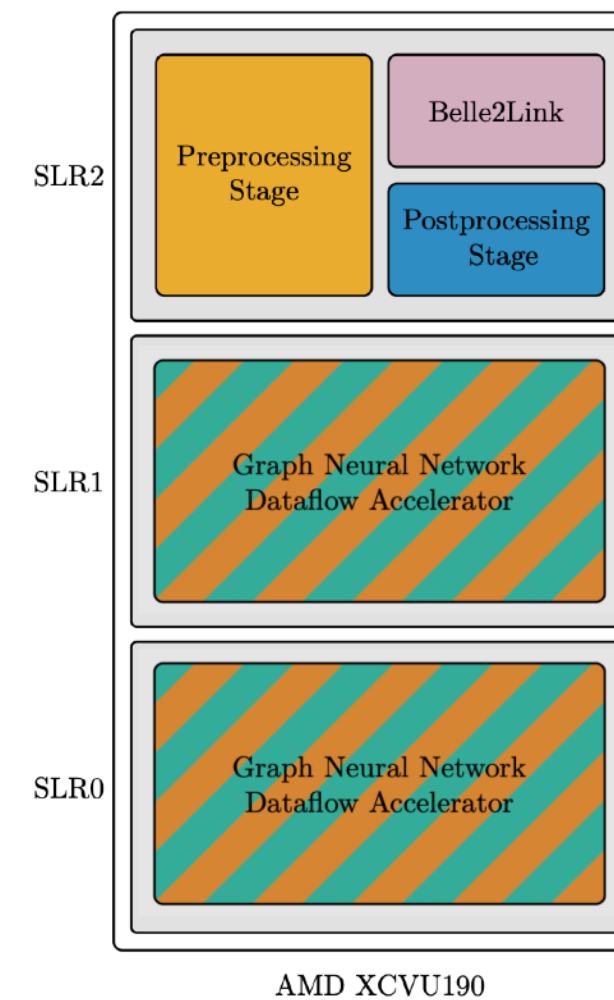
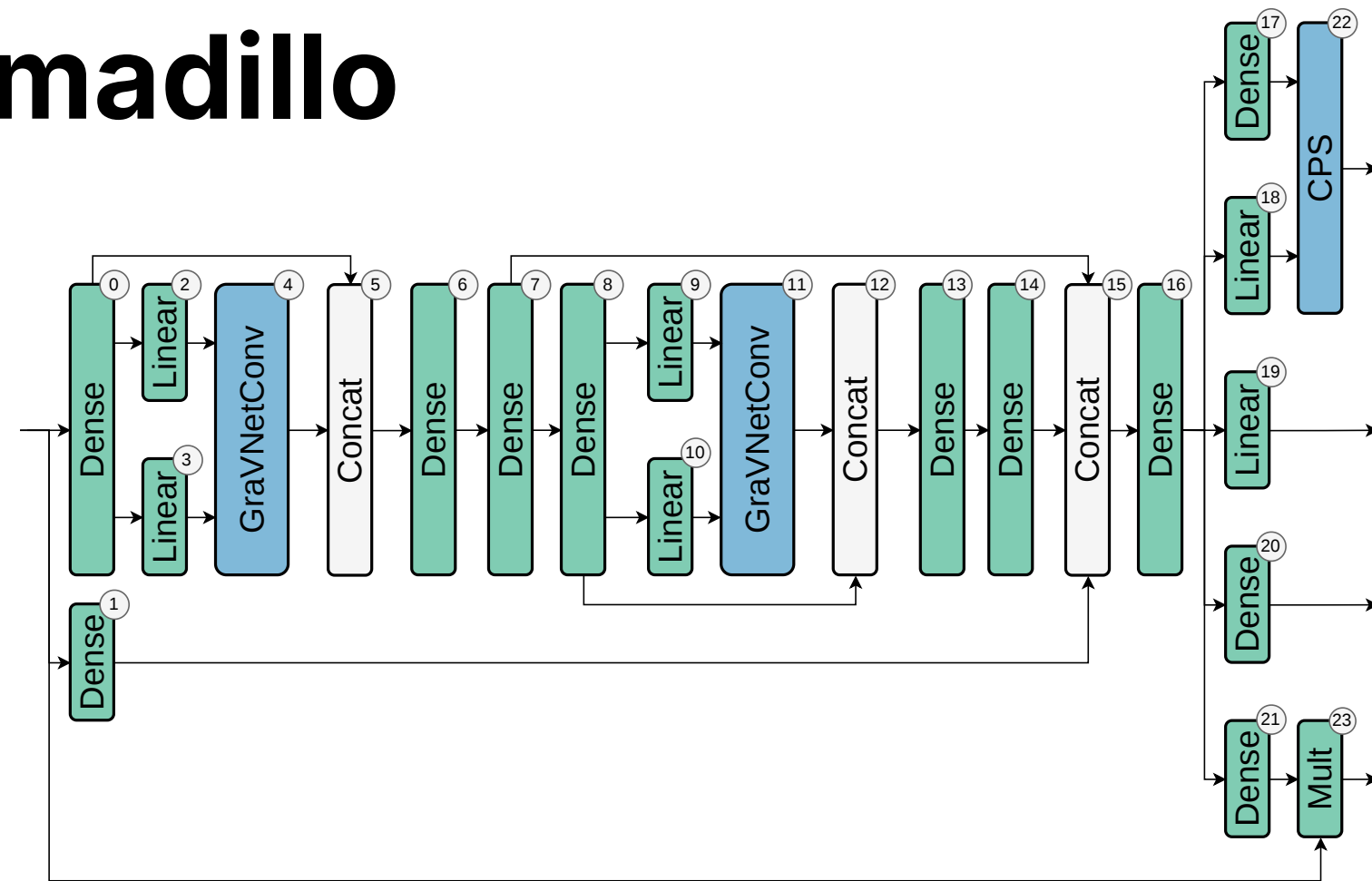
M. Neu et al. <https://arxiv.org/abs/2507.05099>

Armadillo during collisions



From Armadillo to Sunset (simplified)

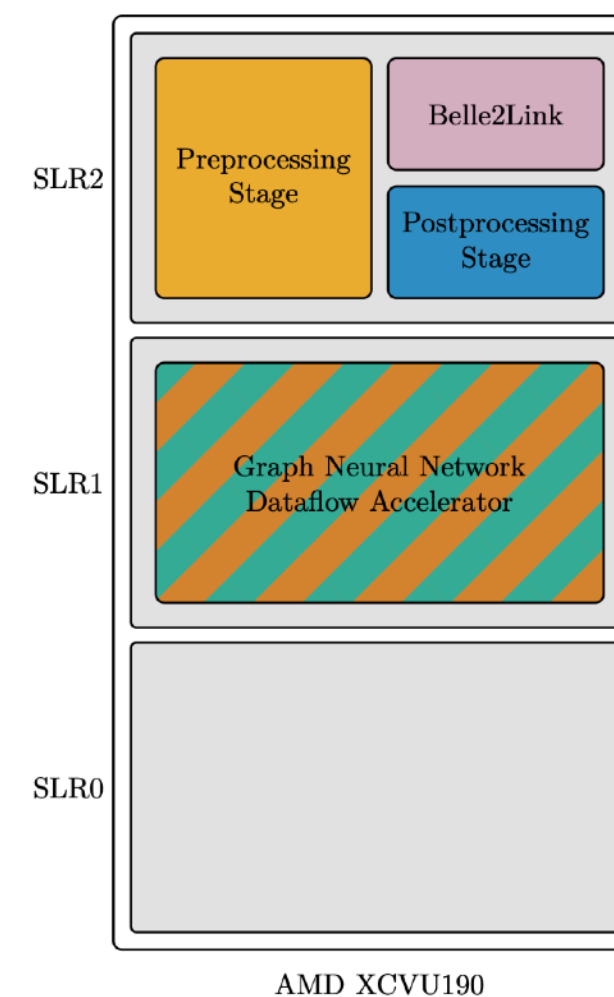
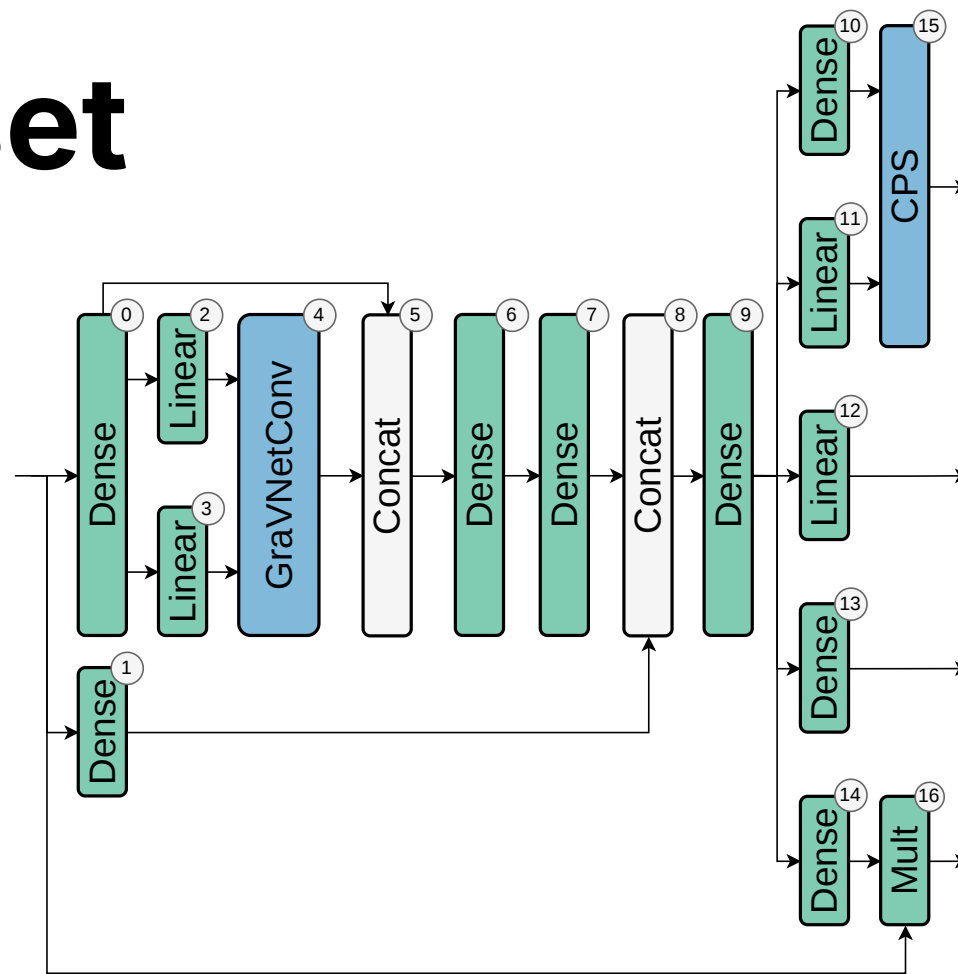
Armadillo



- Internal 16-bit quantization

latency: 3.168 μ s

Sunset



- No DSP usage
- Internal 8-bit quantization

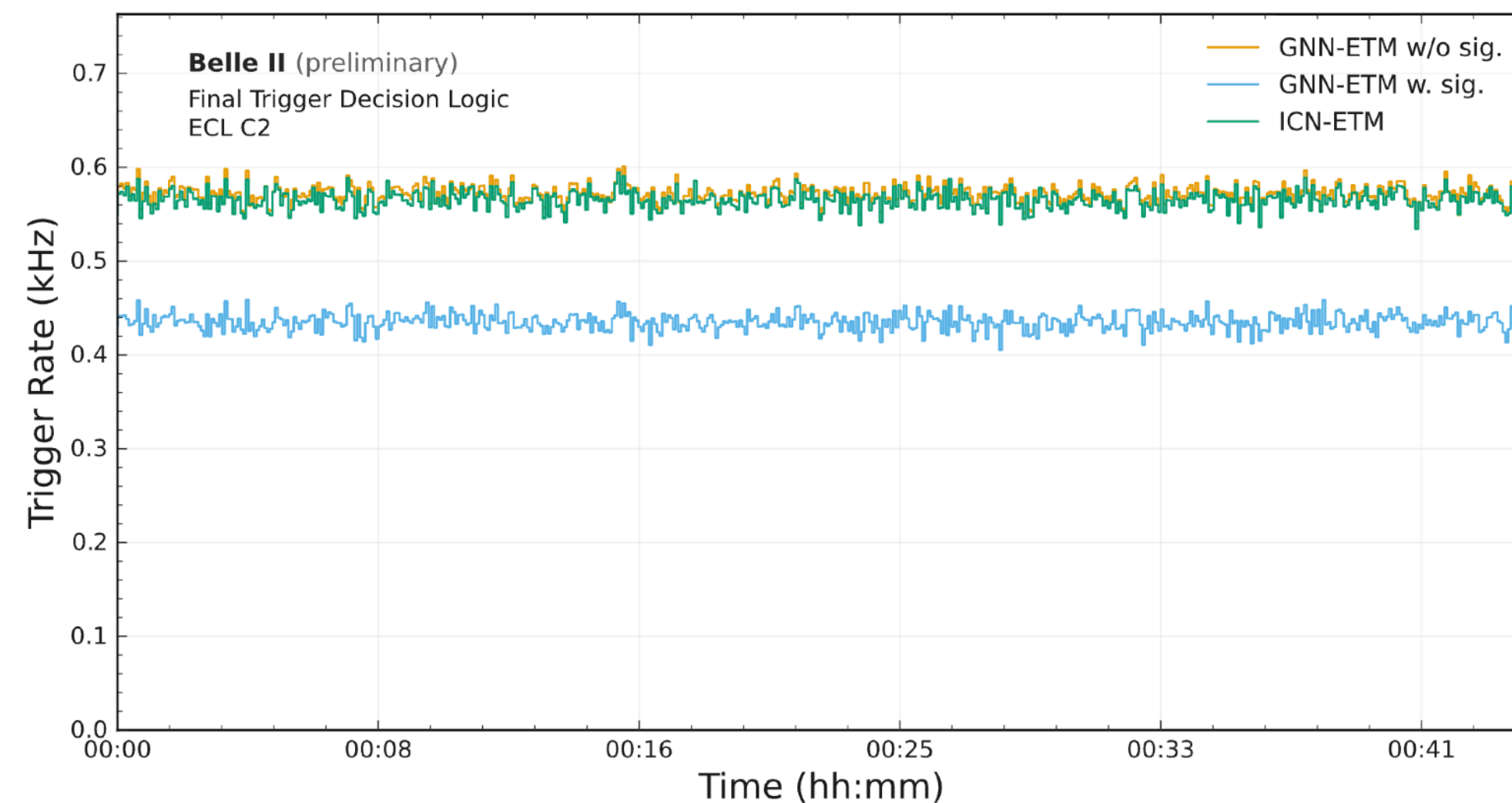
latency: 1.053 μ s

I. Haide PhD thesis <https://publikationen.bibliothek.kit.edu/1000184927>
M. Neu PhD thesis
I. Haide et al. <https://arxiv.org/abs/2602.15118>
M. Neu et al. <https://arxiv.org/abs/2607.09347>
M. Neu et al. <https://arxiv.org/abs/2507.05099>

The first hard realtime GNN on the critical path: Sunset

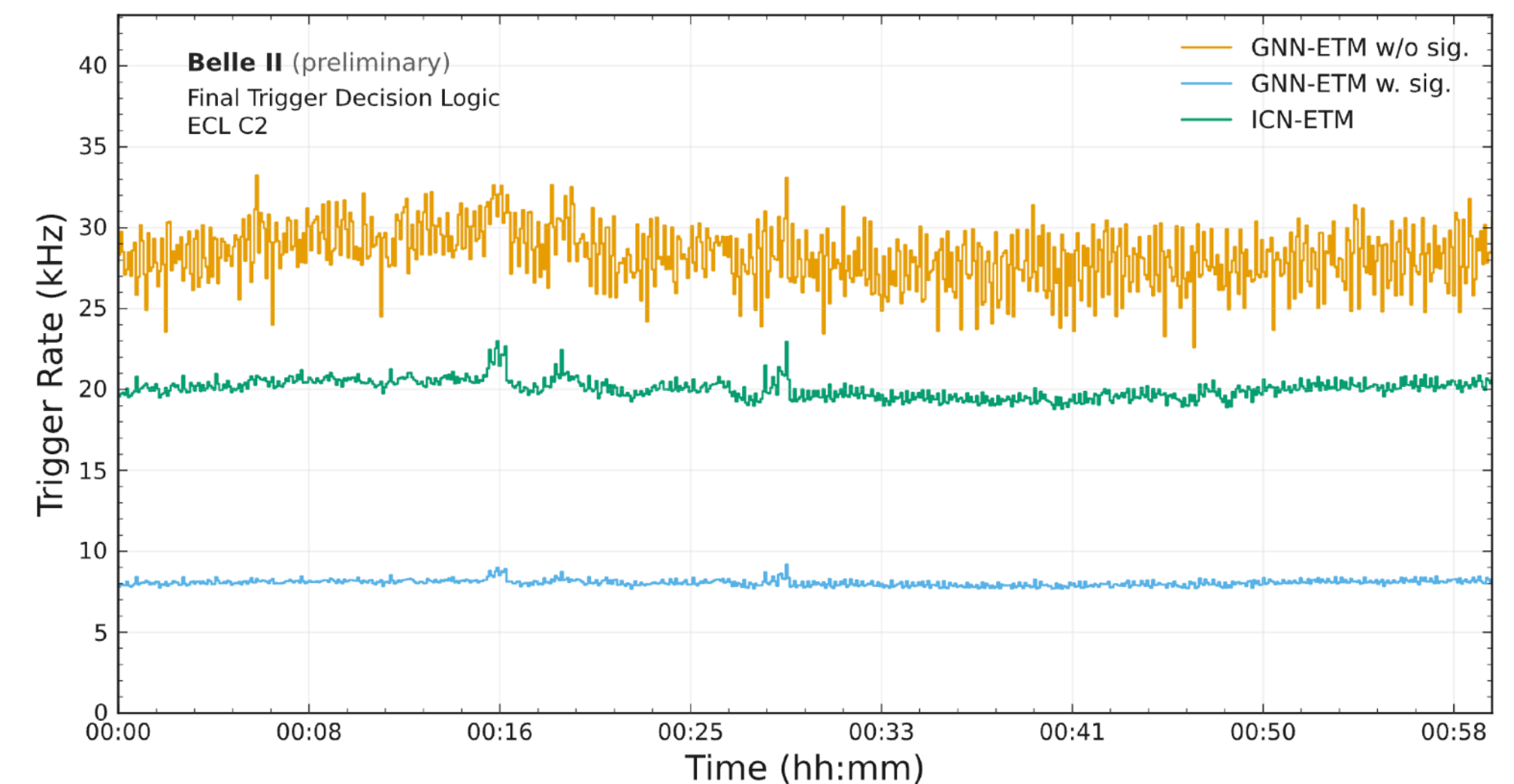
Cosmic data

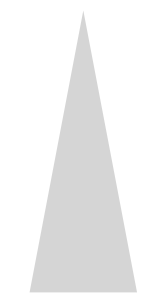
June 2026



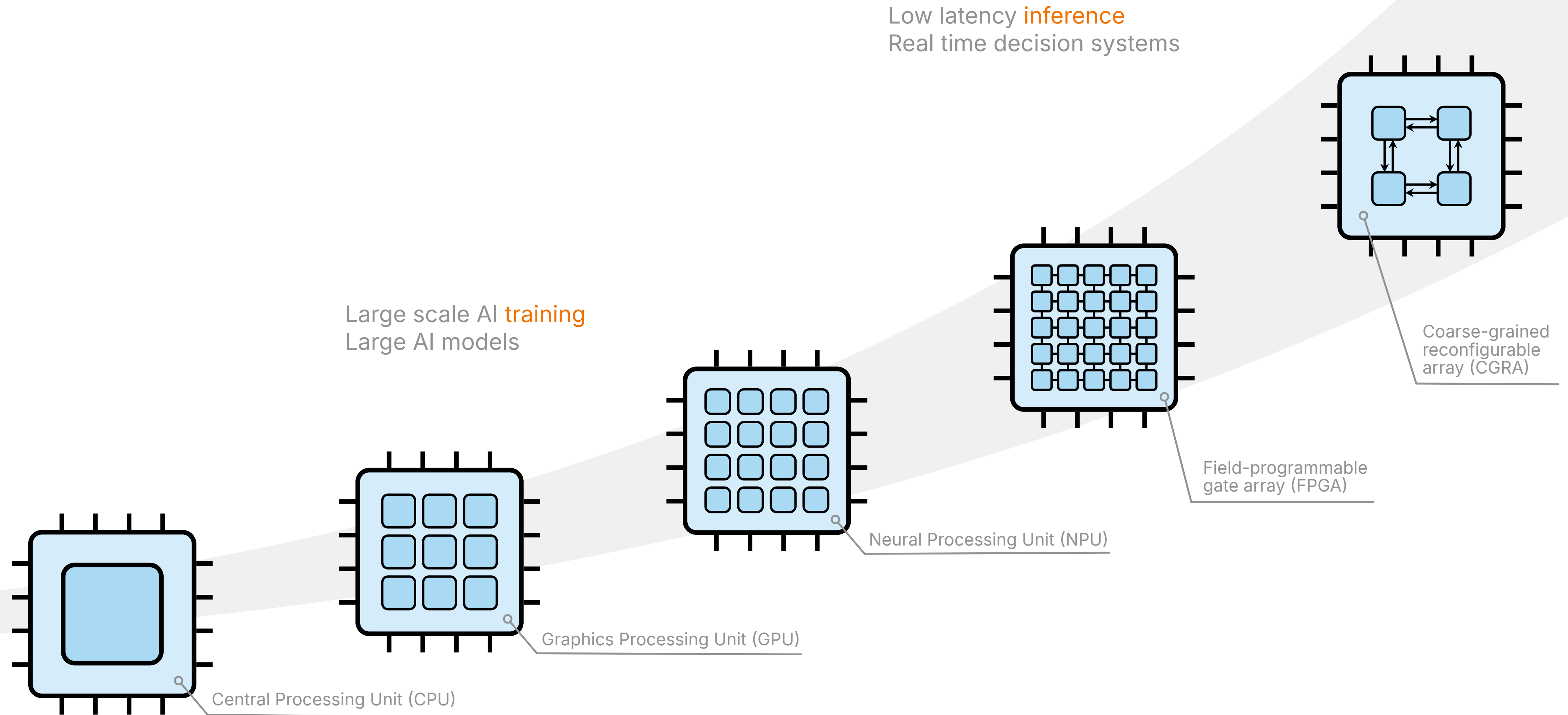
Collision data (250 MHz bunch crossing)

May 2026





Real-time performance

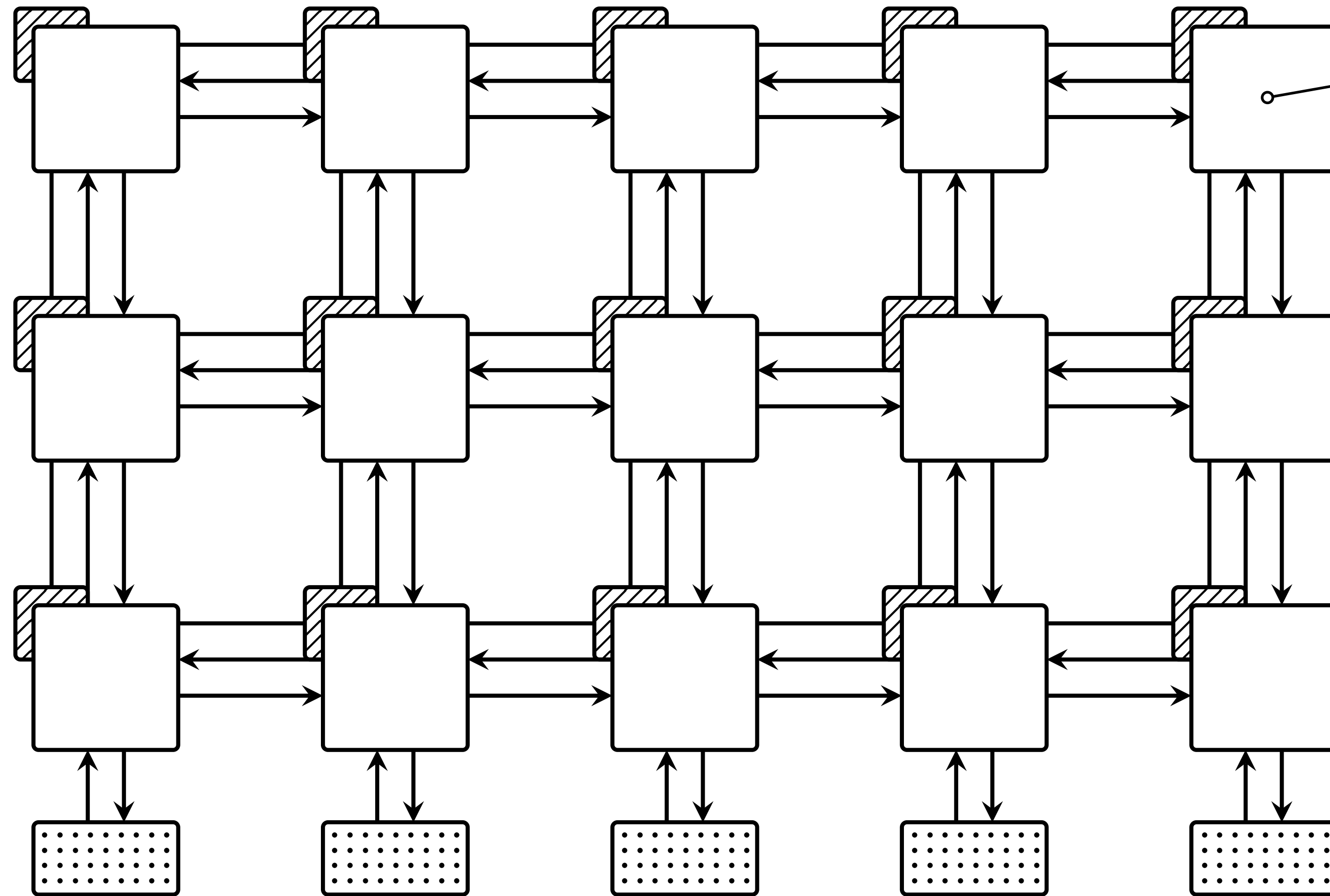


General-purpose computing



Specialised hardware

How does a Coarse-Grain Reconfigurable Array (CGRA) work?



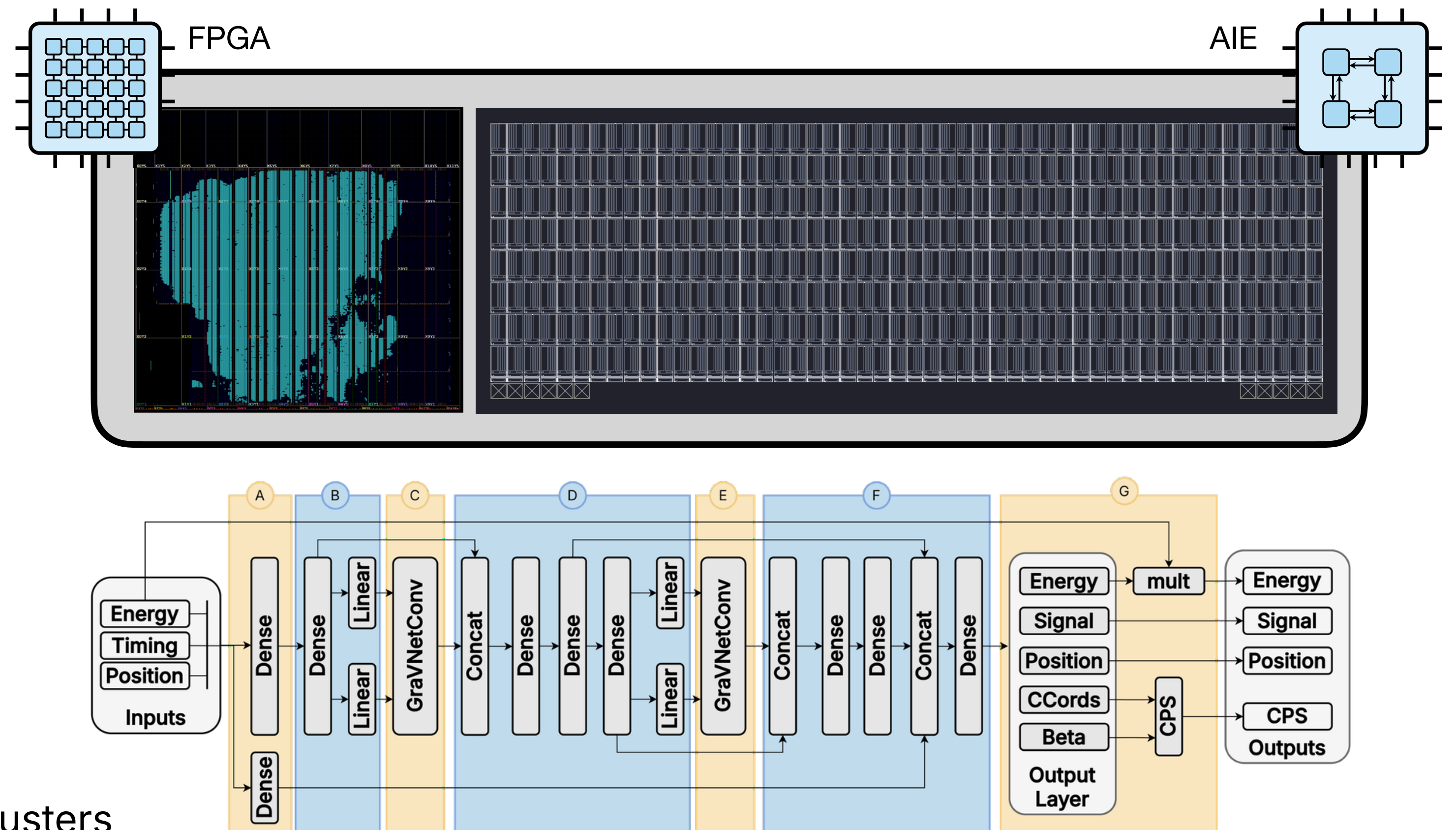
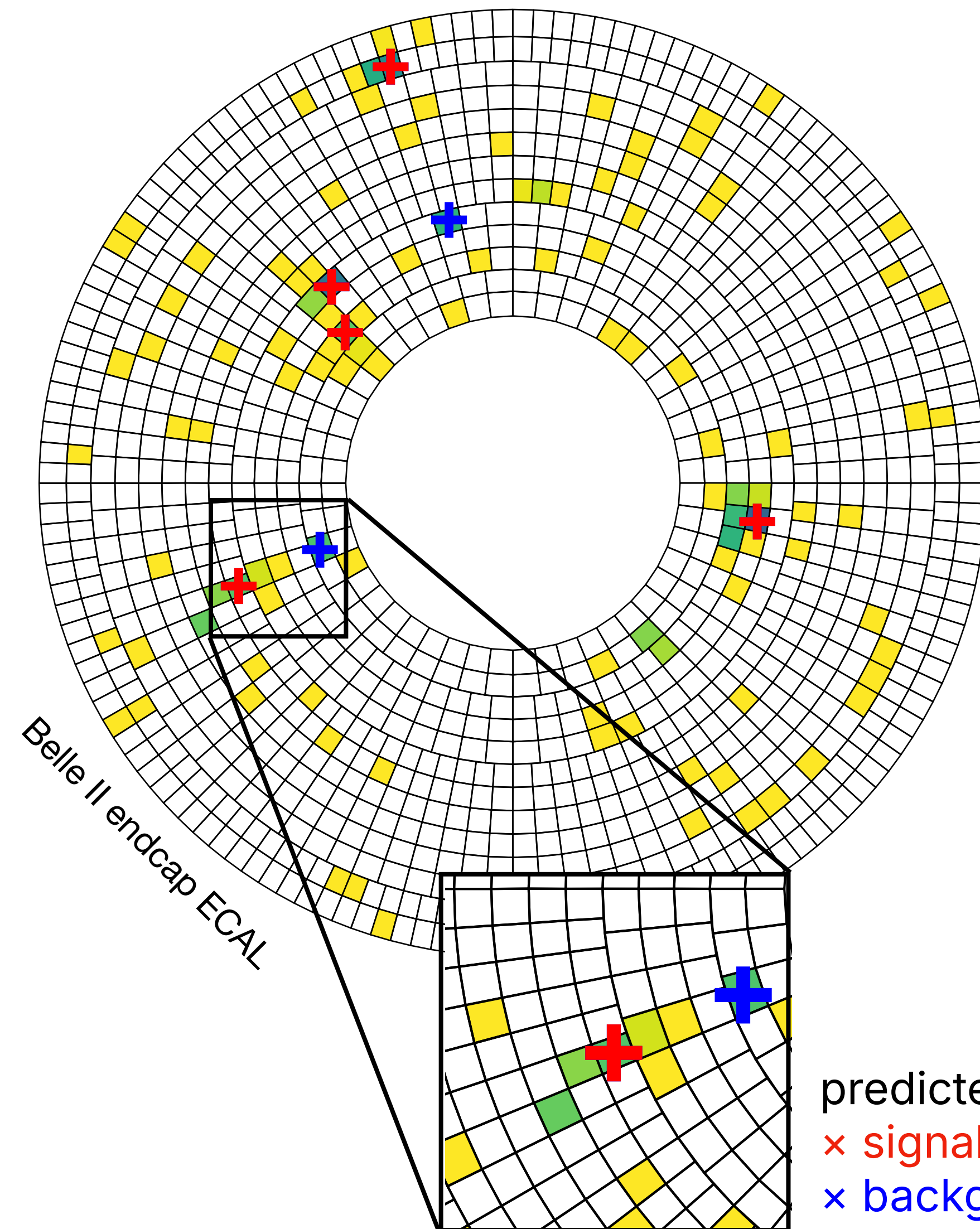
Very Long Instruction Word Processors

Parallel execution of multiple multiply/accumulate, load, and store operations per cycle

M. Neu PhD thesis

Large real-time GNNs on FPGA

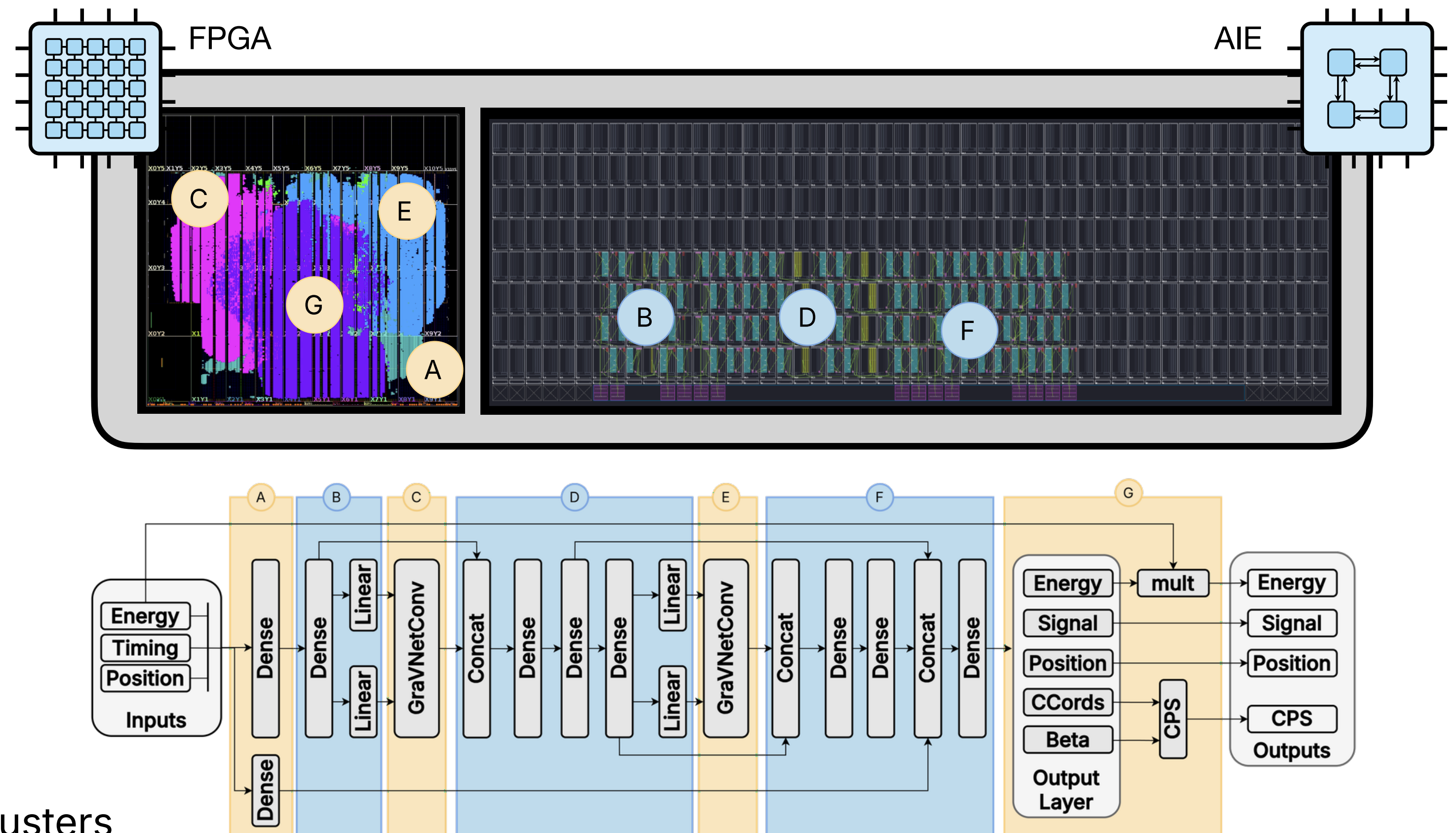
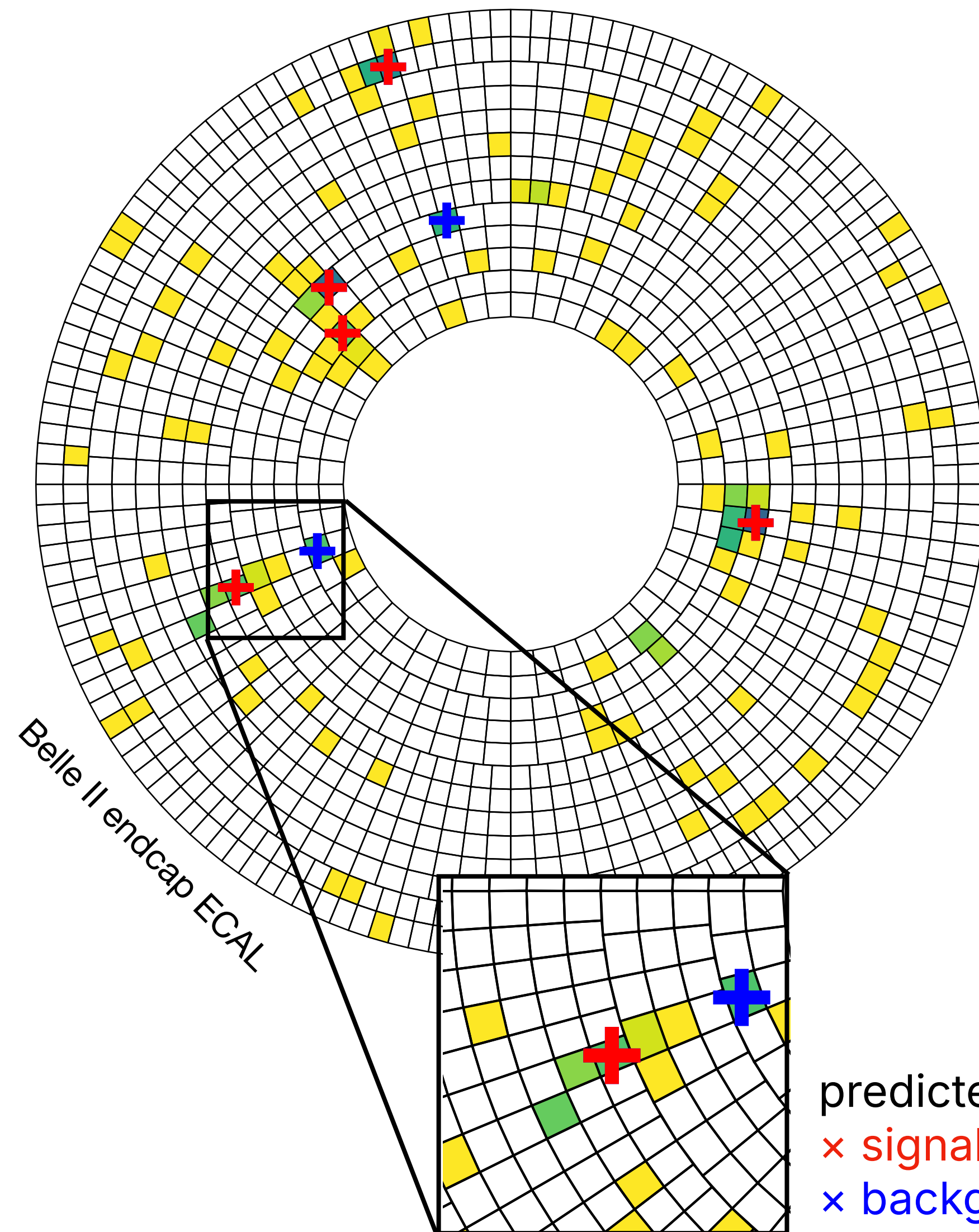
128 non-zero inputs
throughput: 1.86 million events per second
latency 6.06 μ s per event



M. Neu et al. <https://arxiv.org/abs/2605.10612>

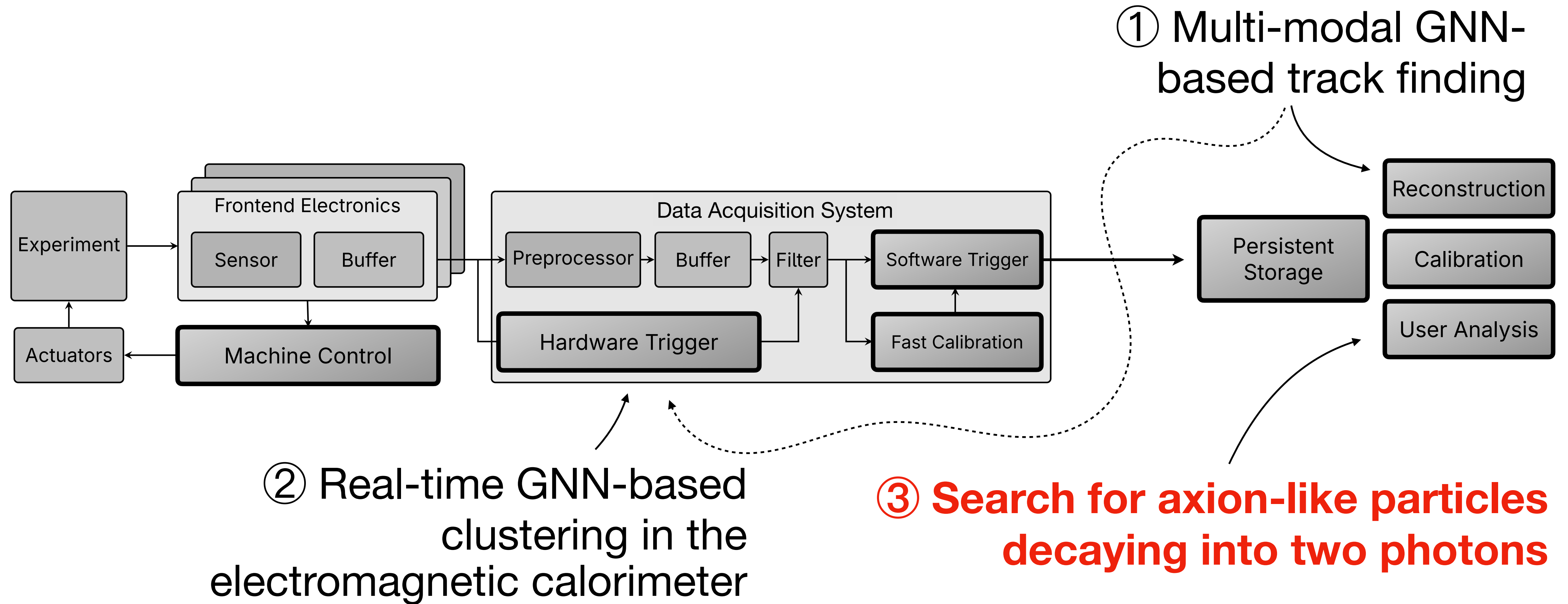
Large real-time GNNs on FPGA and AIEs

128 non-zero inputs
throughput: 2.94 million events per second **(+ 53%)**
latency 7.15 μ s per event **(+18%)**



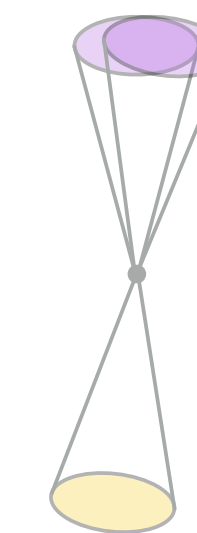
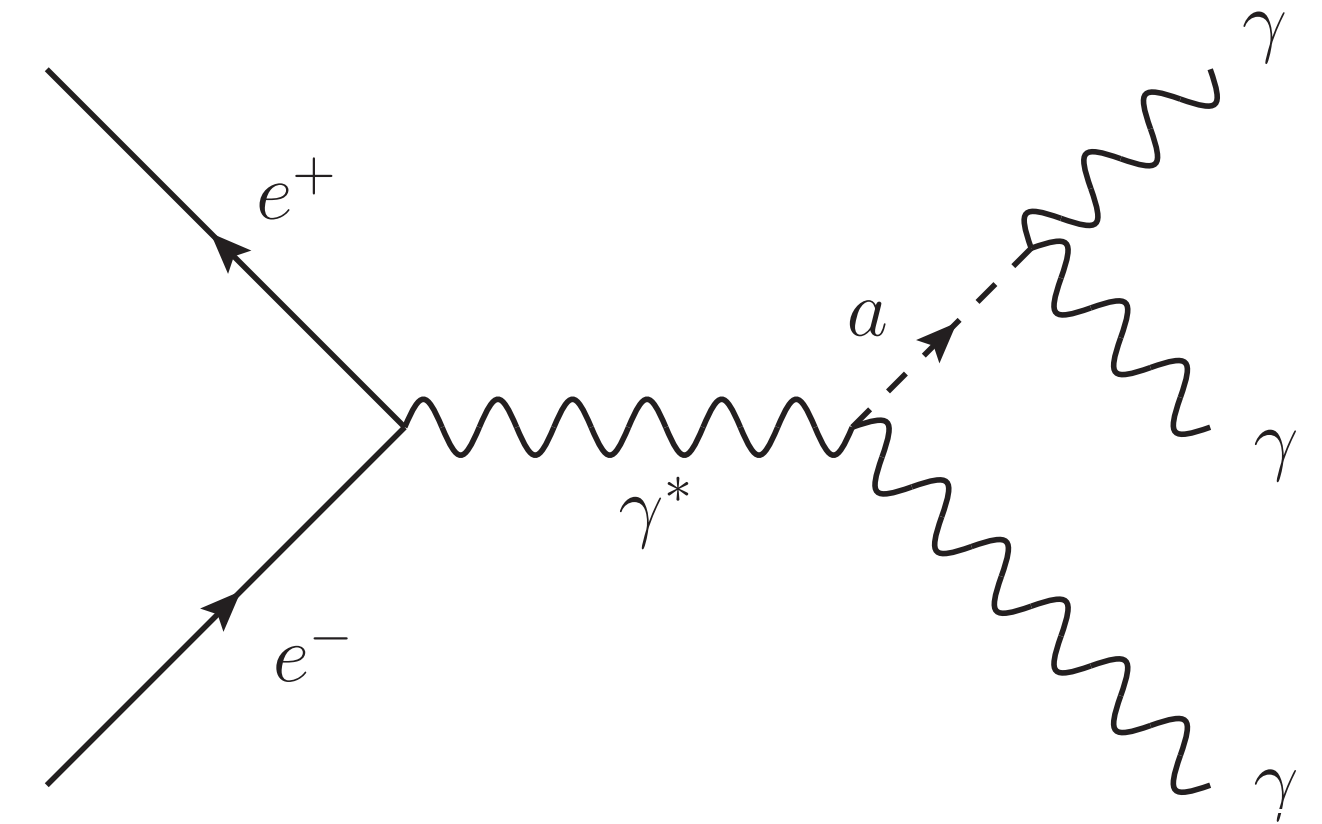
M. Neu et al. <https://arxiv.org/abs/2605.10612>

Outline

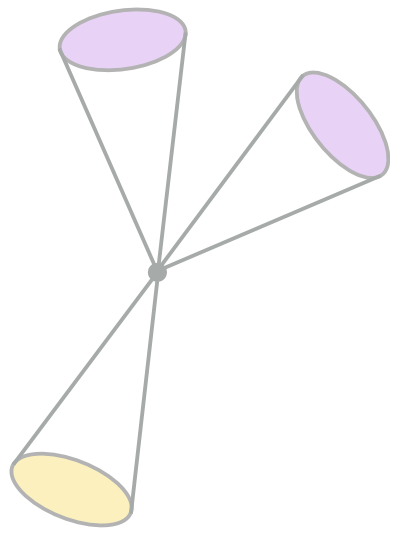
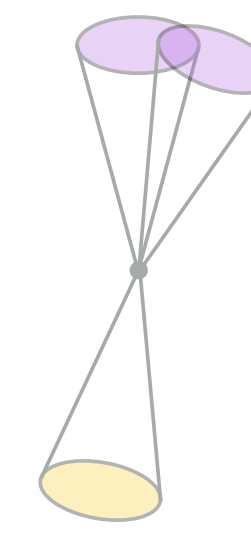


Search for axion-like particles in Belle II

- 408 fb⁻¹ at 10.58 GeV and below (about half the available dataset)
- Search peak in the diphoton invariant mass $M_{\gamma\gamma}$
- Selection:
 - 3 high-energy photons and nothing else (requires high efficiency tracking to reject $e^+e^- \rightarrow e^+e^-\gamma$)
 - $M_{\gamma\gamma\gamma}$ close to known beam energy
 - Neural network on kinematic properties to suppress $e^+e^- \rightarrow \gamma\gamma\gamma$
 - Kinematic fit of three photons to known beam energy to improve mass resolution (0.2% for highest ALP masses)



$m_a \rightarrow 0$

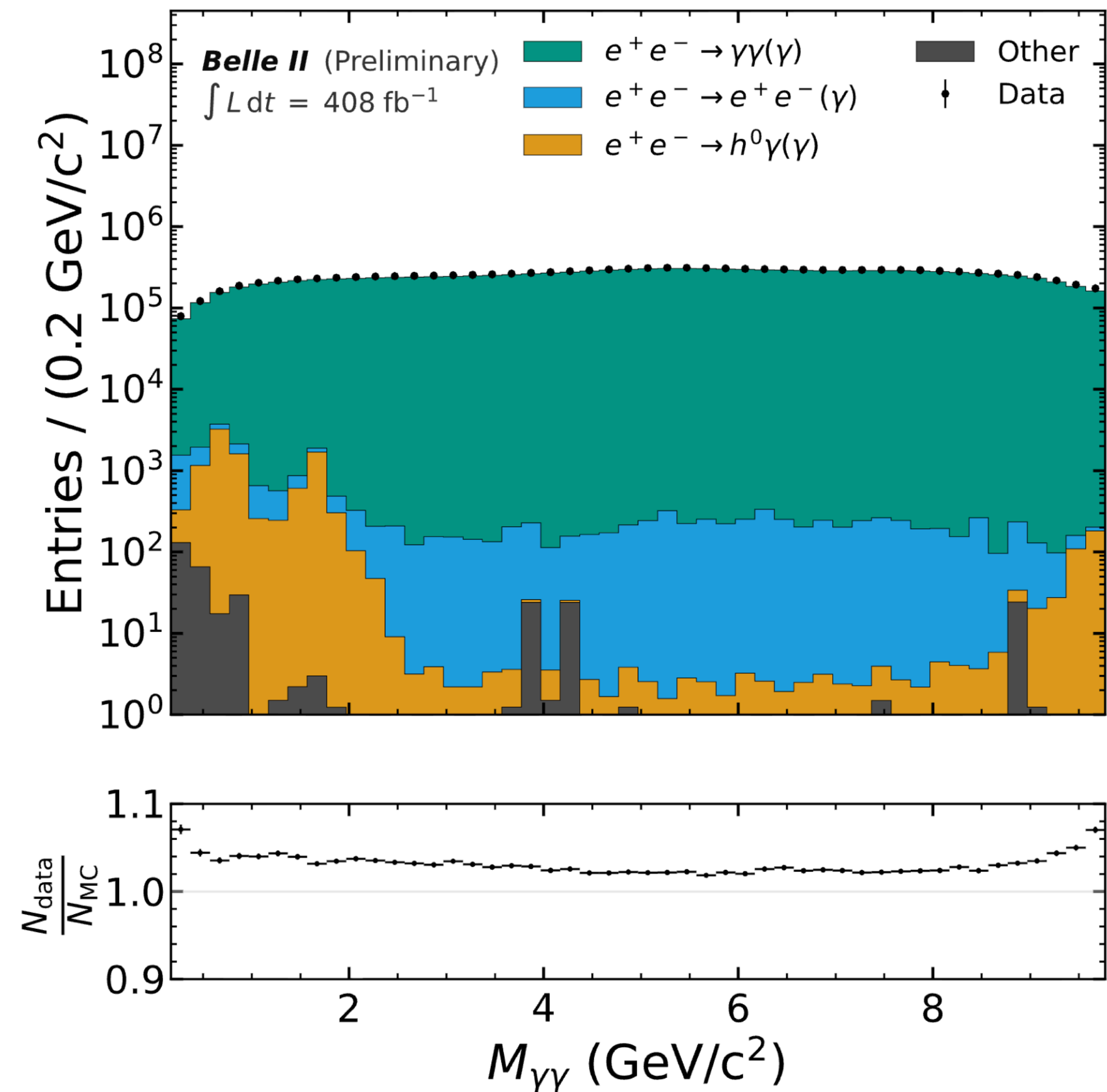
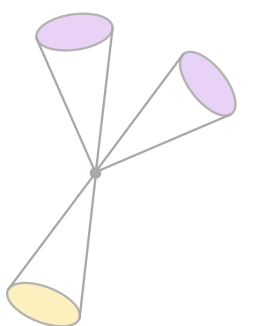
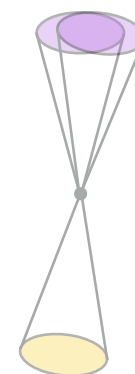
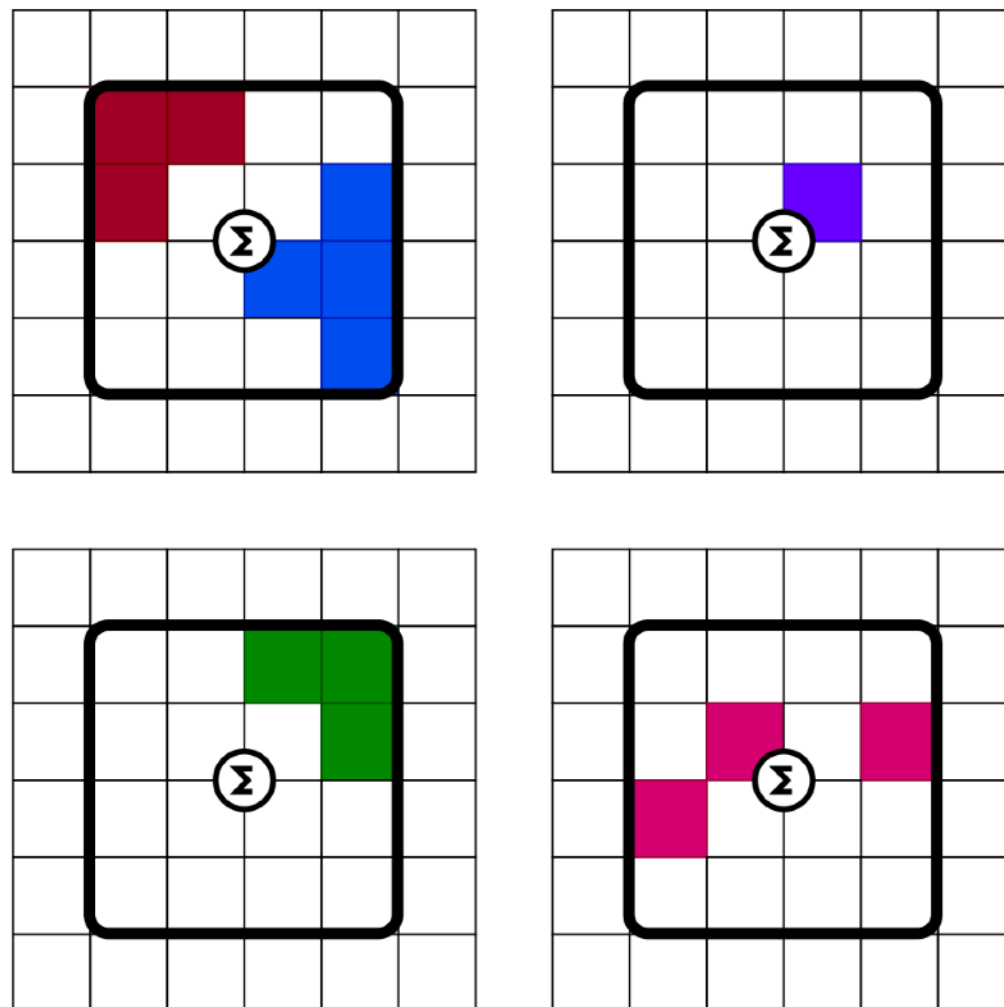


$m_a \rightarrow E_{\text{beam}}$

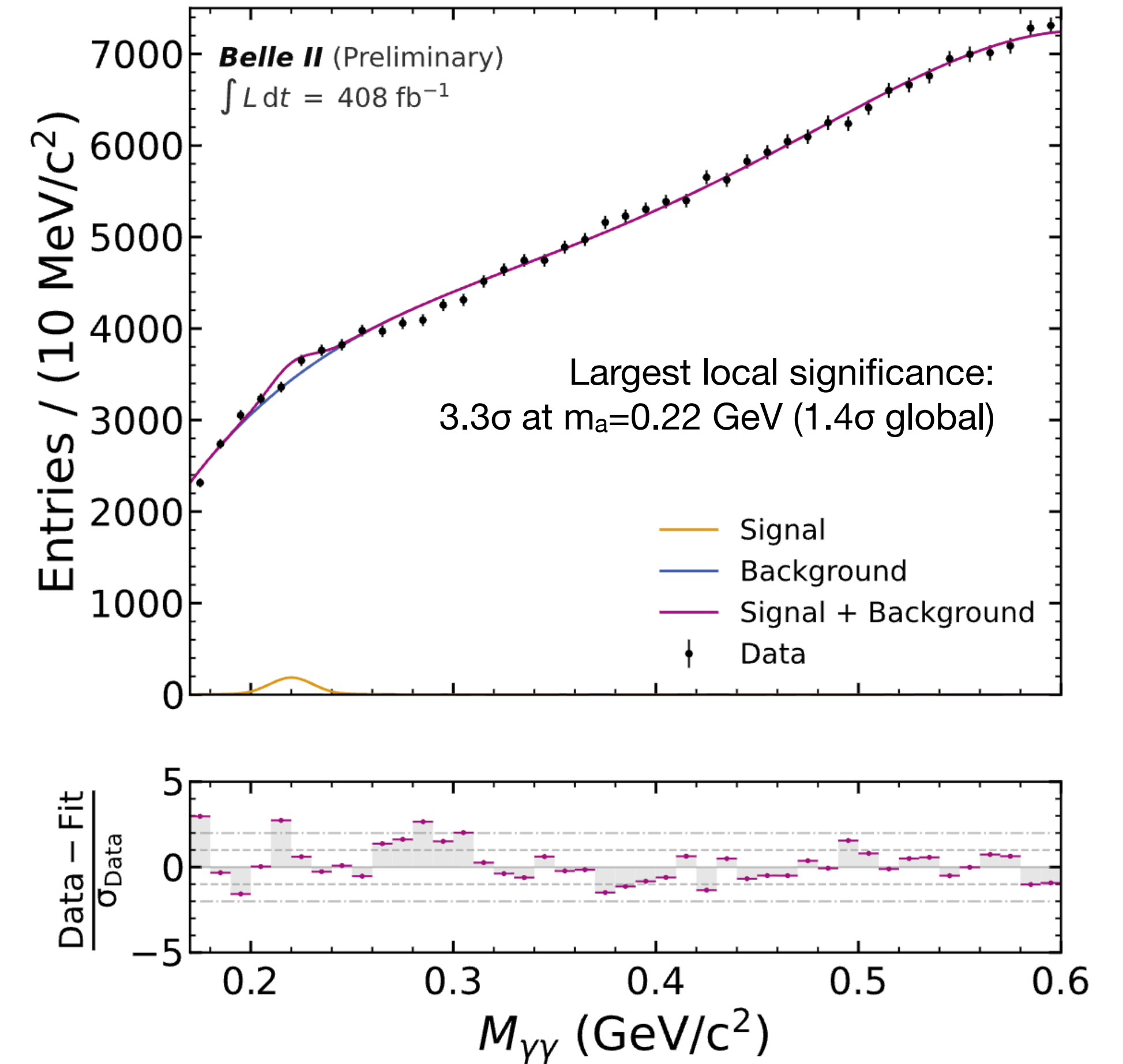
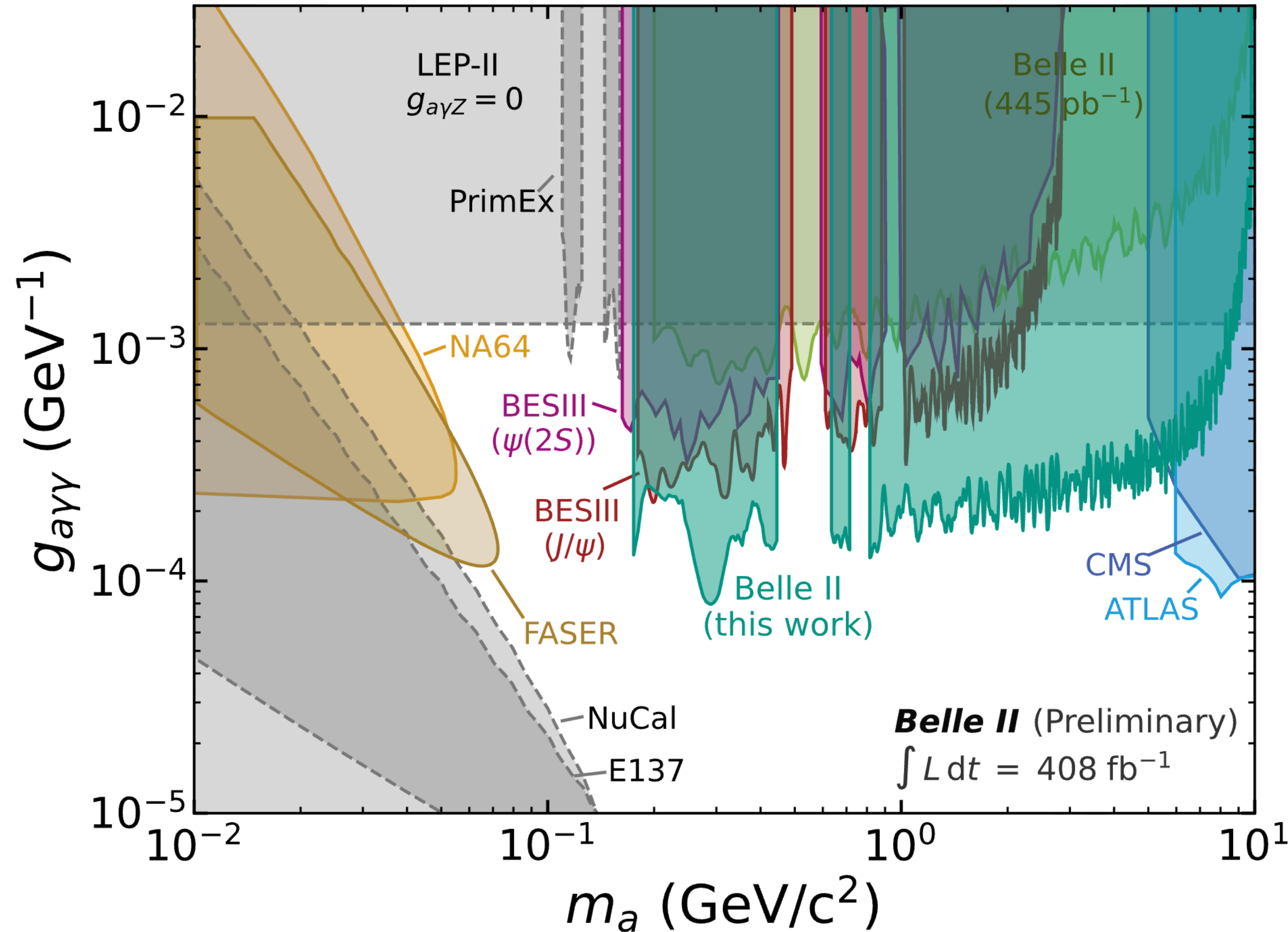
Search for axion-like particles in Belle II

$e^+e^- \rightarrow a(\rightarrow \gamma\gamma)\gamma$ looks like very high rate Standard Model $e^+e^- \rightarrow \gamma\gamma$ if the two decay photons are close together

Especially for the existing L1 trigger.

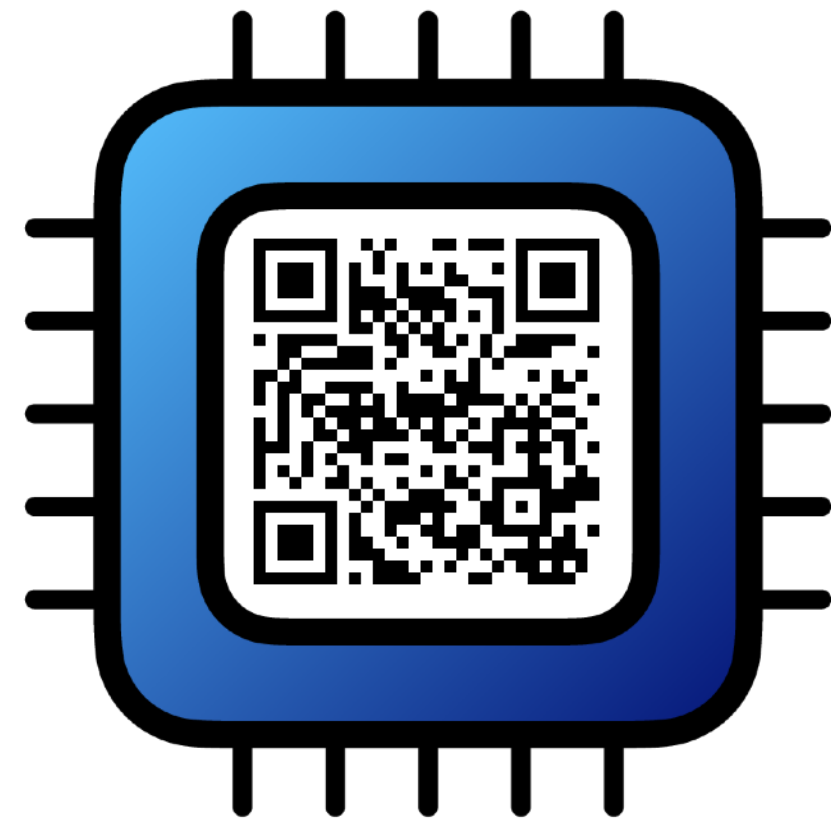


Search for axion-like particles in Belle II



Conclusions

- **Belle II** sets world-leading limits on **axion-like particles**, but this search will become trigger-limited
- Offline, **GNN-based tracking** already outperforms conventional methods
- Similar GNN algorithm operates as a **hard real-time L1 trigger** within 1.053 μs
- **DEEP** develops **FPGA+AI** tooling for triggers on AMD Versal hardware



Want to know more?
www.erumdata-deep.de