Deep Learning Applications for Particle Physics in Tracking and Calorimetry

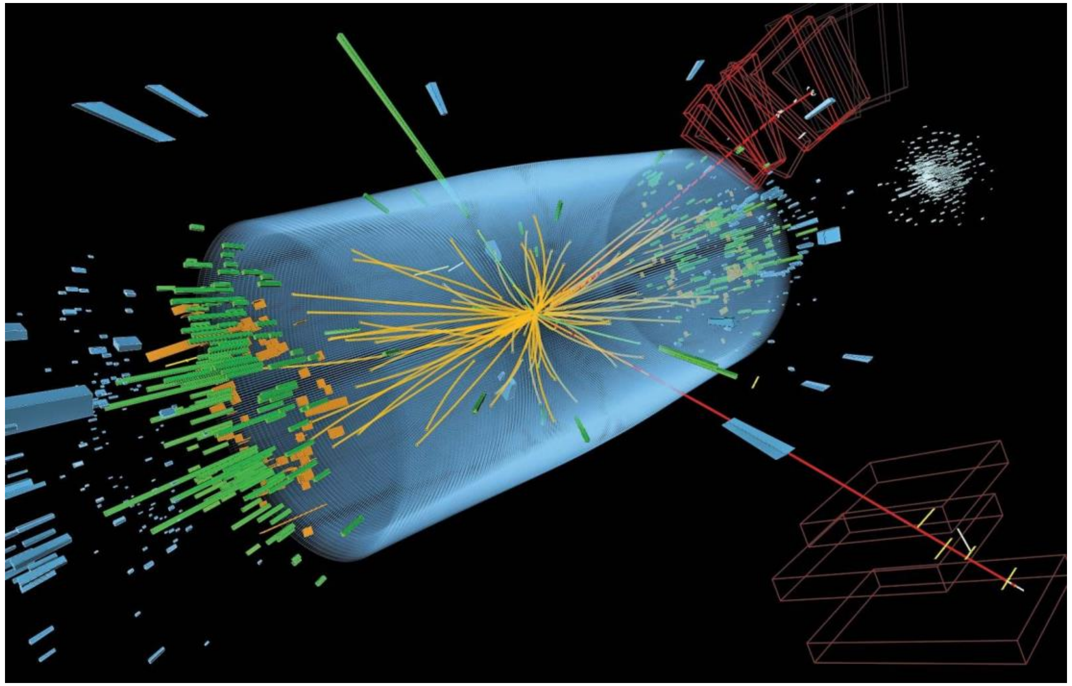
ALEX SCHUY

Introduction

- BACKGROUND AND IMPORTANCE
- MOTIVATION
- THESIS STATEMENT

Particle Physics

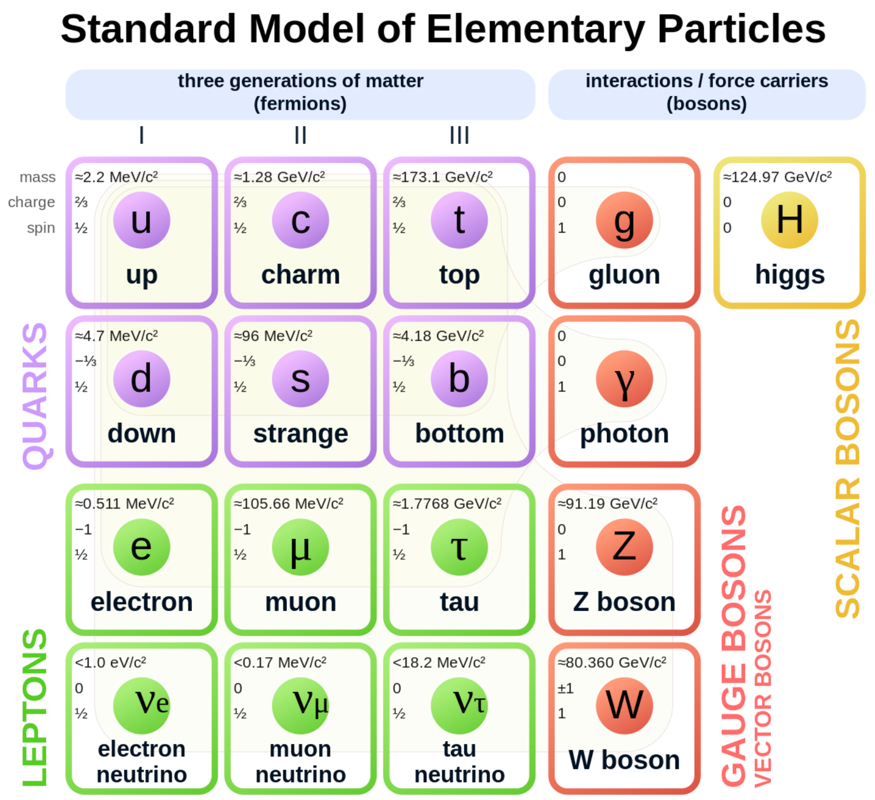
- The study of fundamental constituents of matter and their interactions.
- Rooted in centuries of scientific inquiry, culminating in the Standard Model.
- However, still many unresolved questions…



https://www.home.cern/science/accelerators/large-hadron-collider

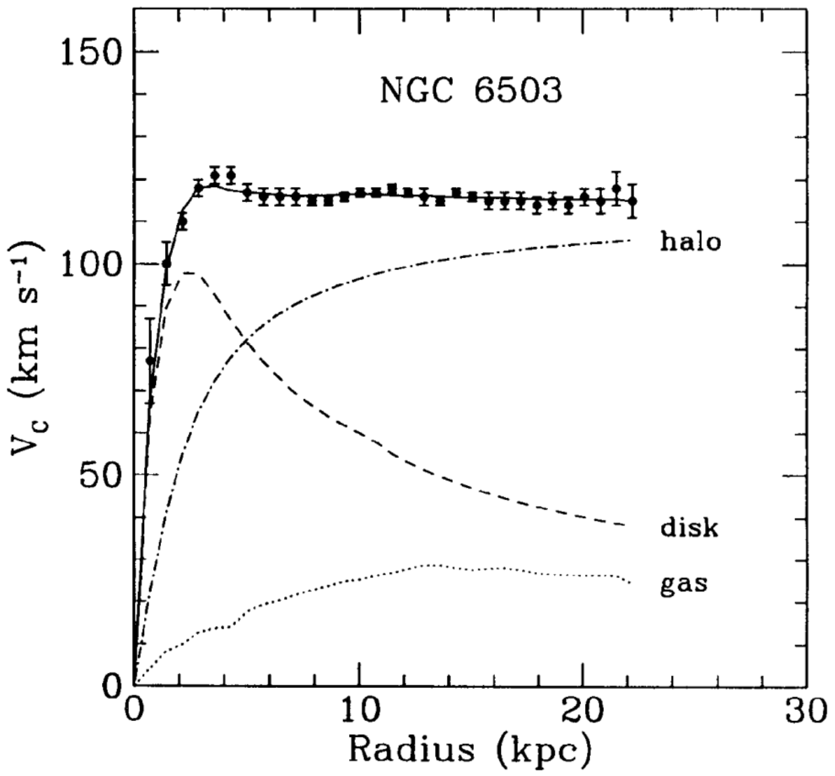
The Standard Model of Particle Physics

- Two types of particles:
  - Fermions – matter particles
  - Bosons – force carrying particles
- Describes three of the four fundamental interactions:
  - Electromagnetism
  - Weak force
  - Strong force



Dark Matter

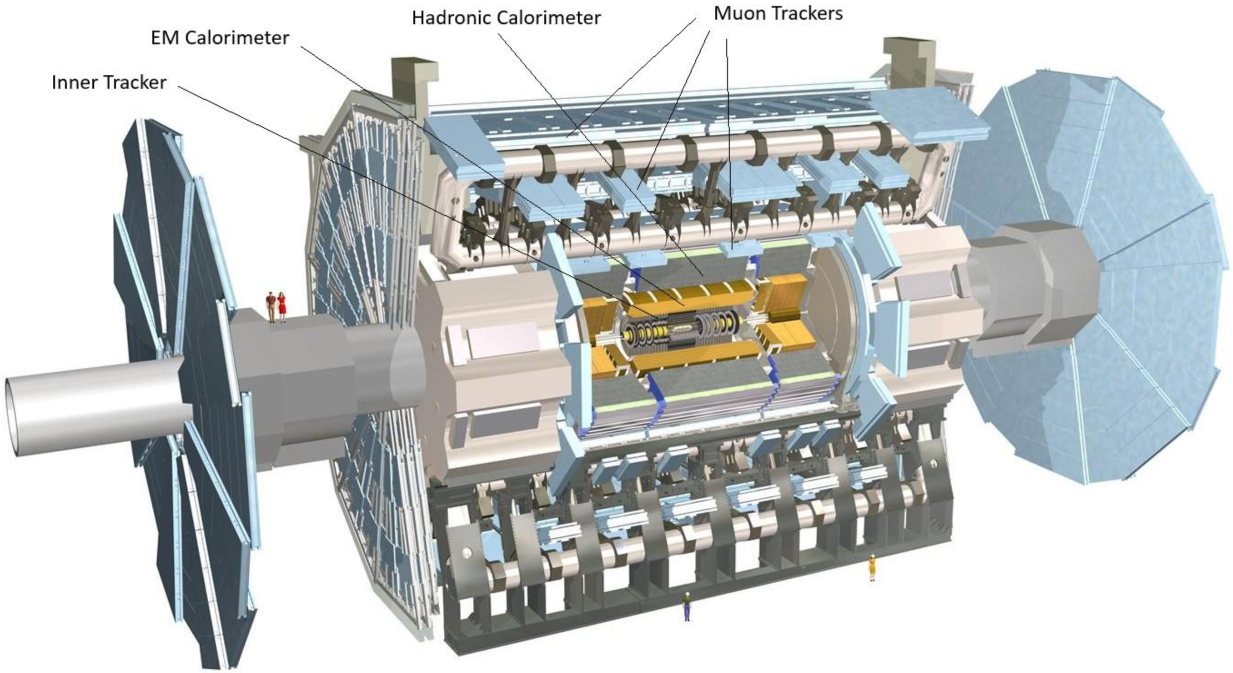
- Evidence
  - Galaxy rotation curves
  - Galaxy clusters
  - Gravitational lensing
  - Cosmic microwave background
  - Structure formation
  - …
- Theories
  - Weakly interacting massive particles (WIMPs)
  - Sterile neutrinos
  - Axions
  - …

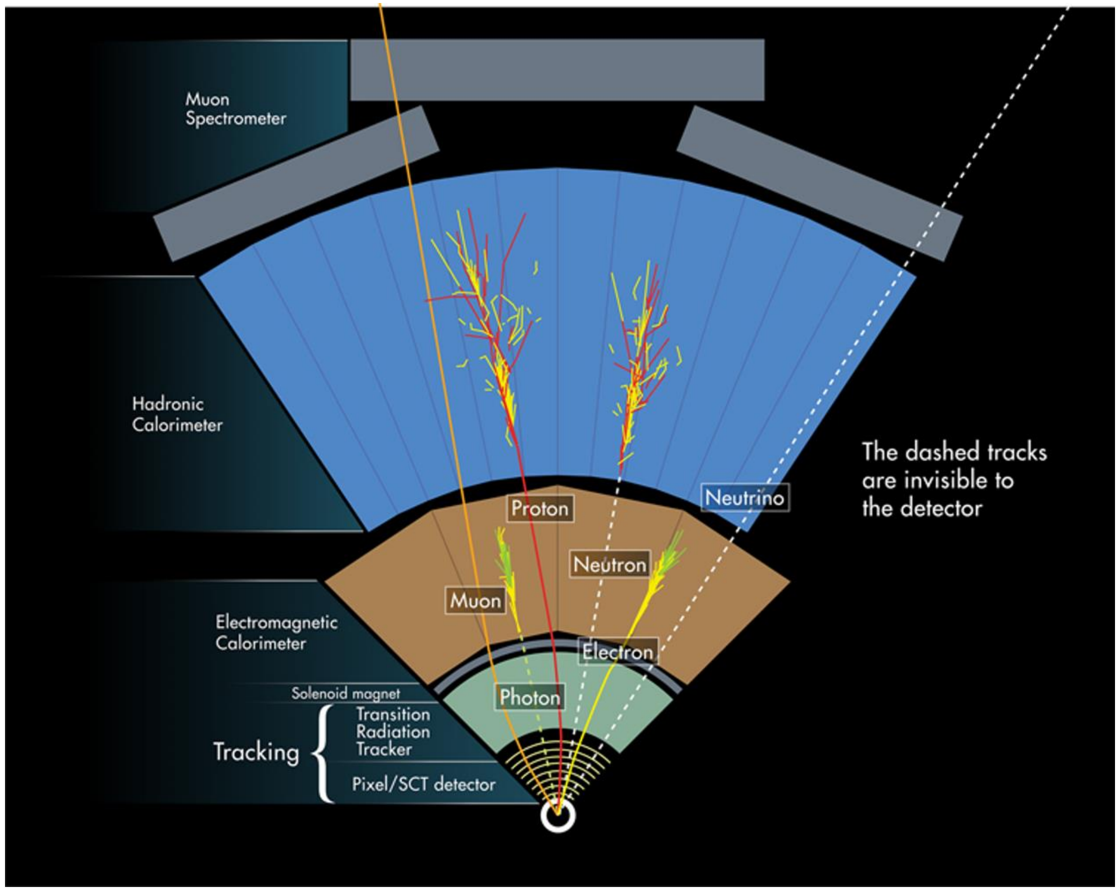


Galactic rotation curve for NGC 6503 showing disk and gas contribution plus the dark matter halo contribution needed to match the data. https://arxiv.org/pdf/1701.01840.pdf


LHCb
ATLAS
CERN
CERN
SUISSE
FRANCE
SPS 7 km
ALICE
CMS
LHC 27 km

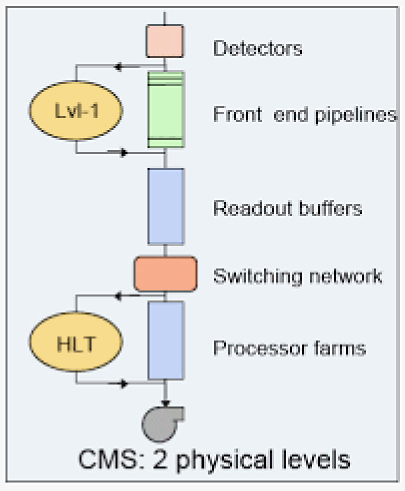
Particle Detectors and Their Role





Trigger System

- Event rate is too high to store everything
- Must decide which events to keep, which to throw out ("trigger system")
- Usually, a two-tier system
  - Level 1 trigger (L1) – ~μs latency
  - High-level trigger (HLT) – ~100 ms latency

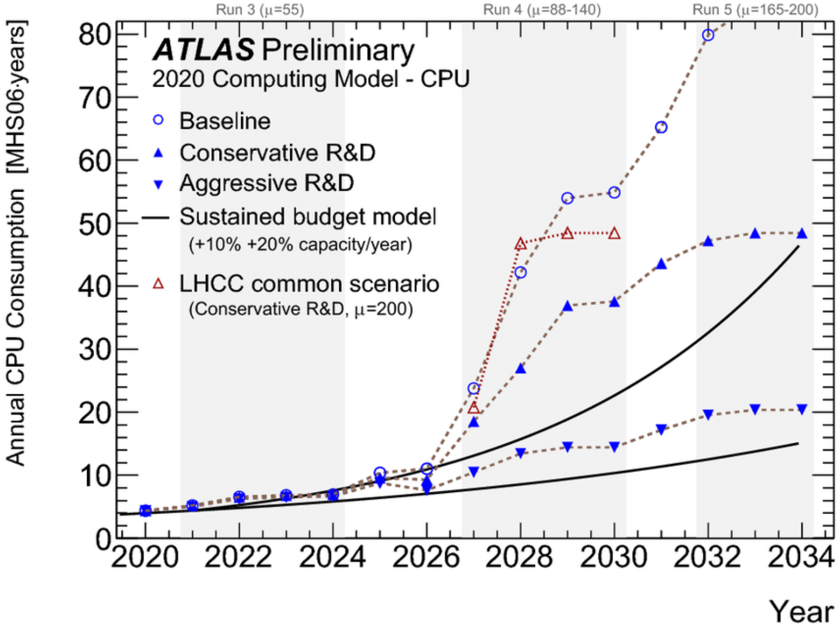


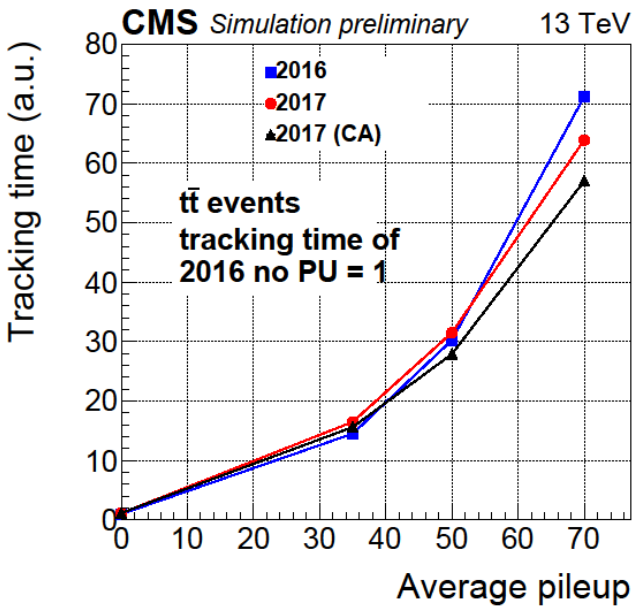
https://cds.cern.ch/record/2232067/files/arXiv:0810.4133.pdf

Challenges in Event Reconstruction

To discover **new physics**, need **higher luminosity & energy**…

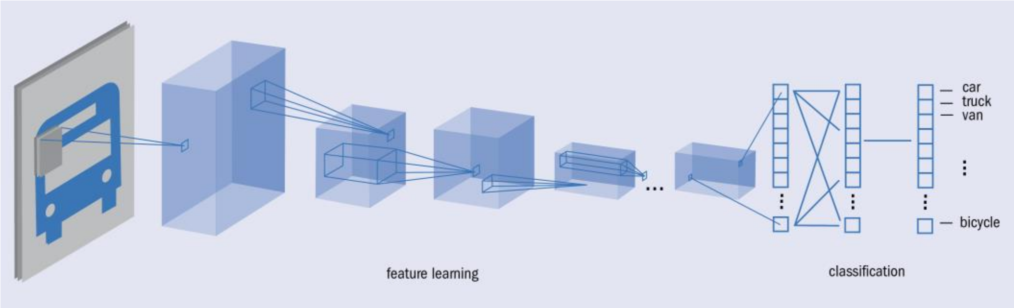
…which leads to more complex events, **increasing computing demand** and **decreasing performance**



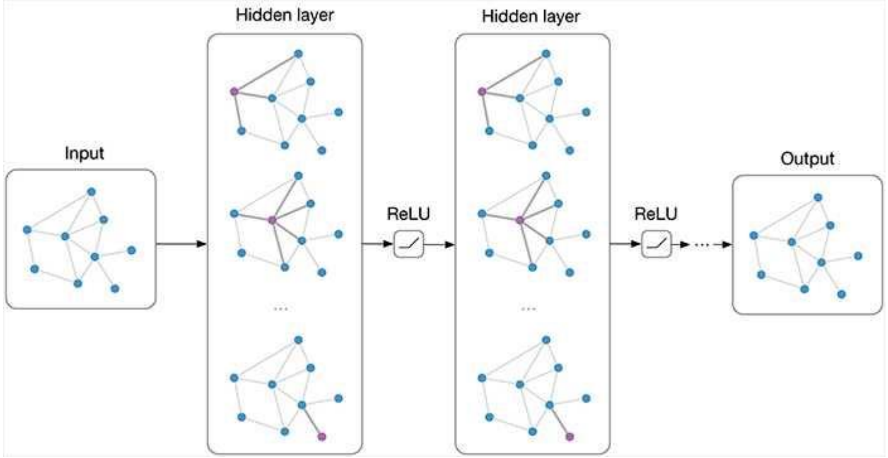


Deep Learning

- An evolving subfield of machine learning (ML) and artificial intelligence (AI), with applications in particle physics.
- Physics-relevant techniques include classification, tagging, noise reduction, event reconstruction, event simulation, anomaly detection…



https://cerncourier.com/a/the-rise-of-deep-learning/



https://theaisummer.com/Graph_Neural_Networks/

Study 1: Performance of a Geometric Deep Learning Pipeline for HL-LHC Particle Tracking

- FUNDAMENTALS
- METHODOLOGY
- RESULTS

Particle Tracking

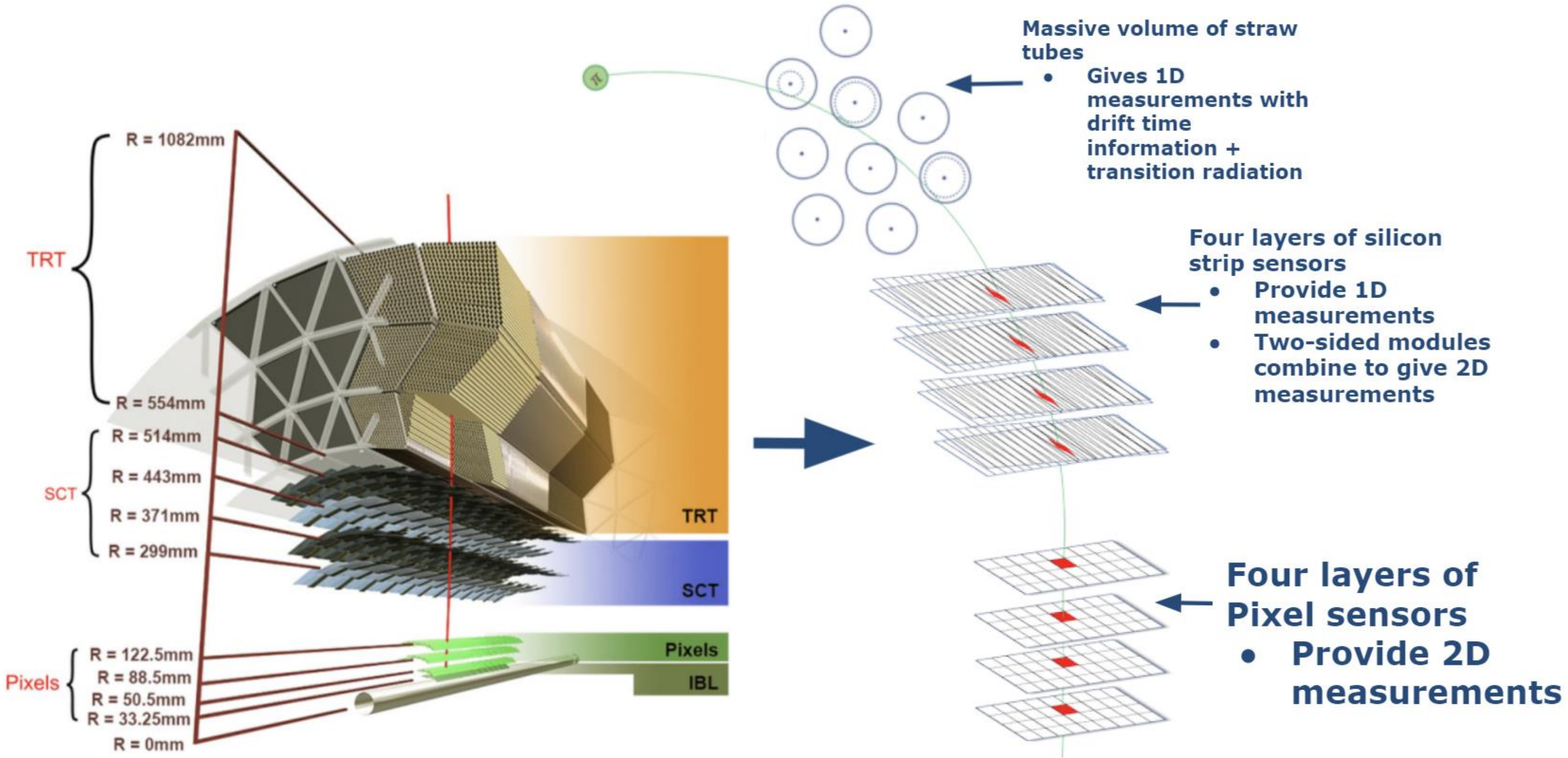
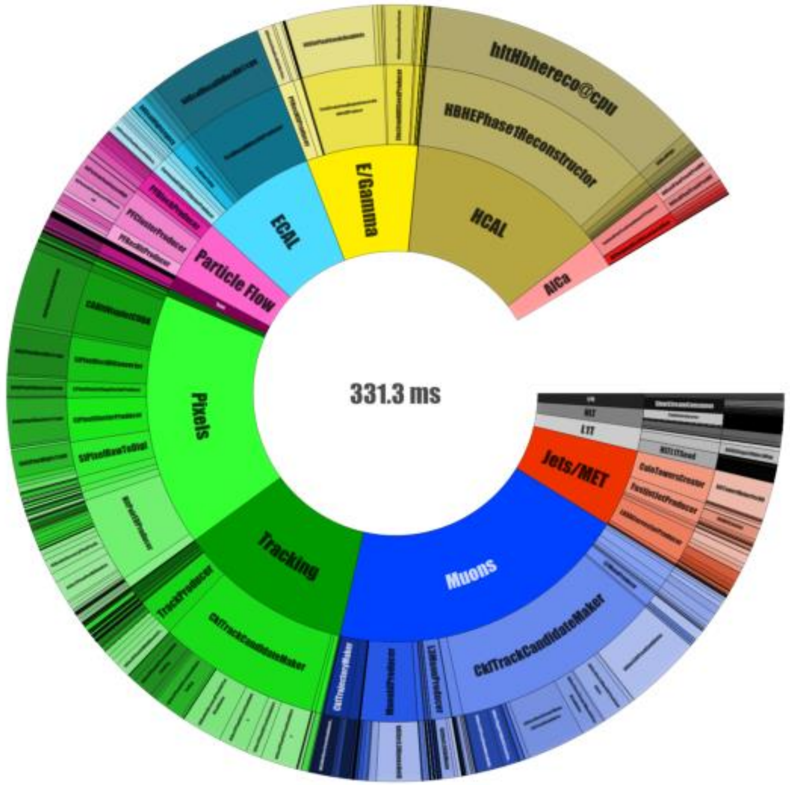
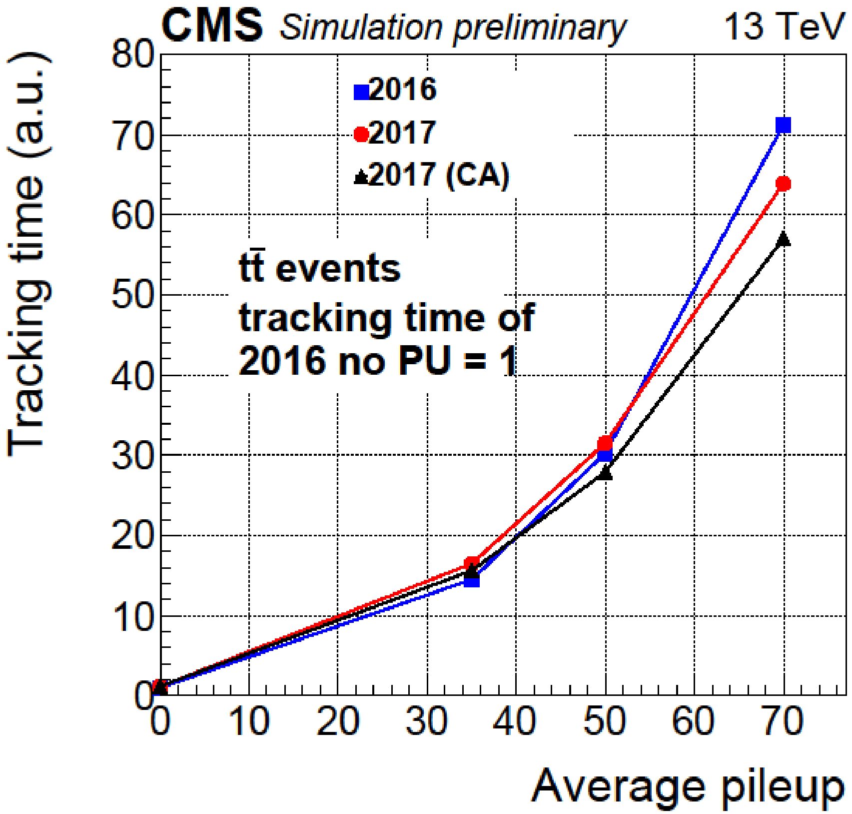


Image: https://atlassoftwaredocs.web.cern.ch/trackingTutorial/idoverview/

Particle Tracking Challenge

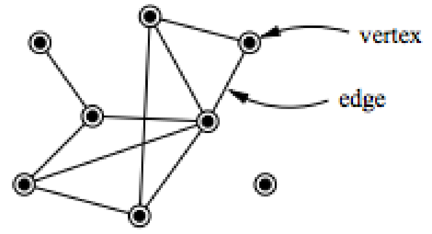




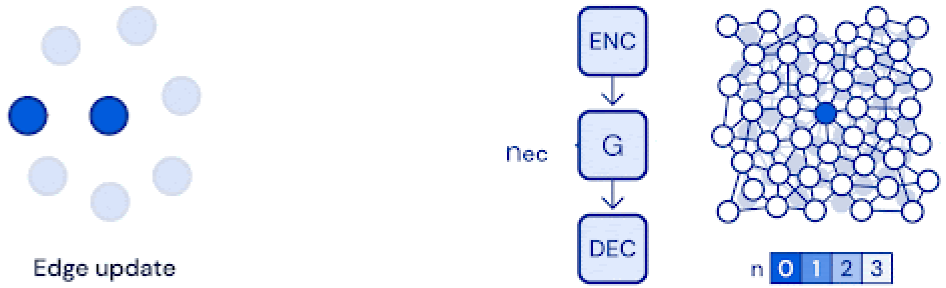
https://cds.cern.ch/record/2792313/files/DP2021_013.pdf

Graph Neural Networks

- Graphs excel at representing relationships.
- GNNs are tailored for graph-structured data.
- GNNs use message-passing to update graph information.



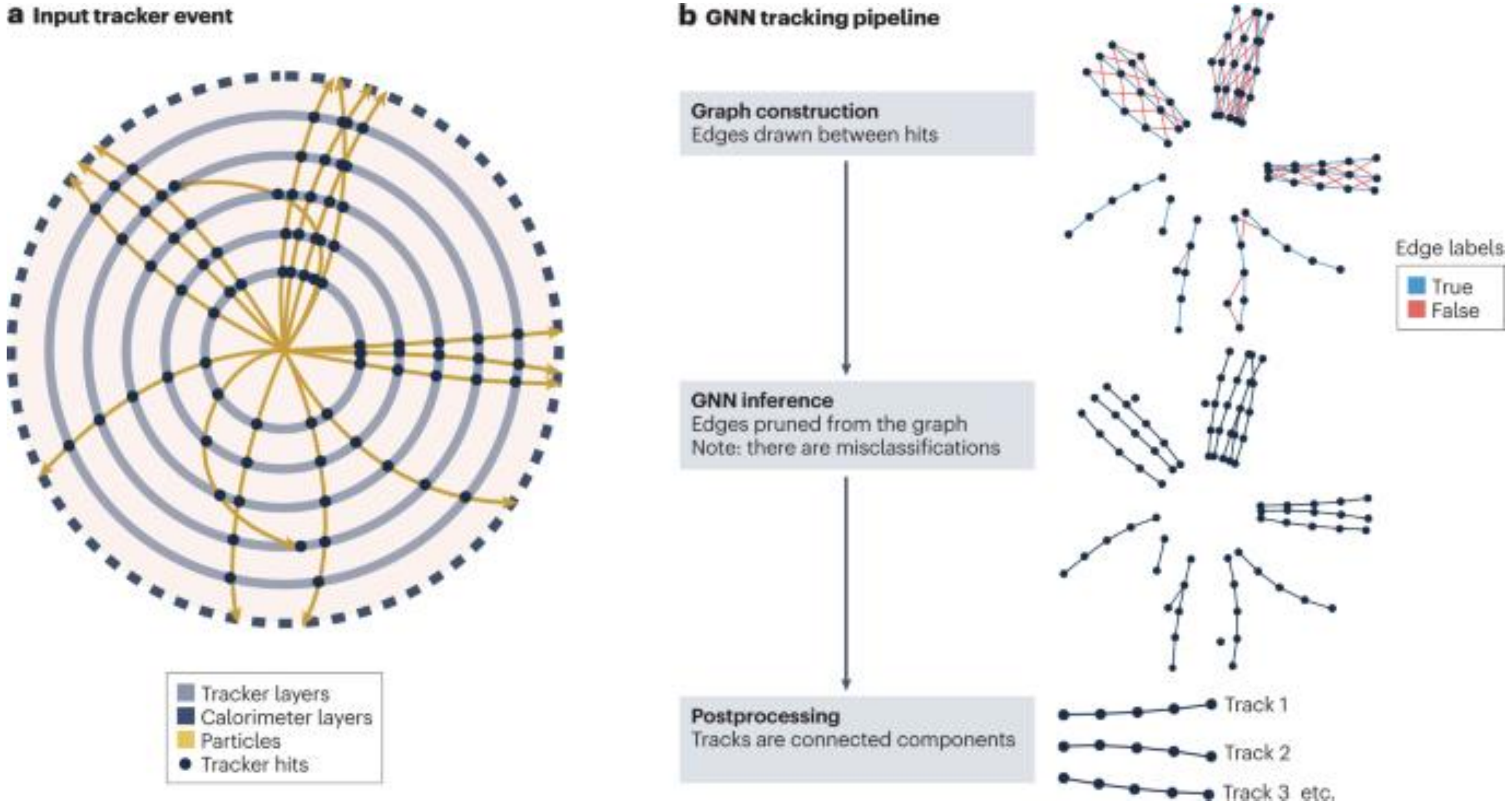
https://en.wikipedia.org/wiki/Vertex_(graph_theory)



Example of a message passing GNN.
Left: a single message passing update.
Right: illustration of receptive field after n passes.

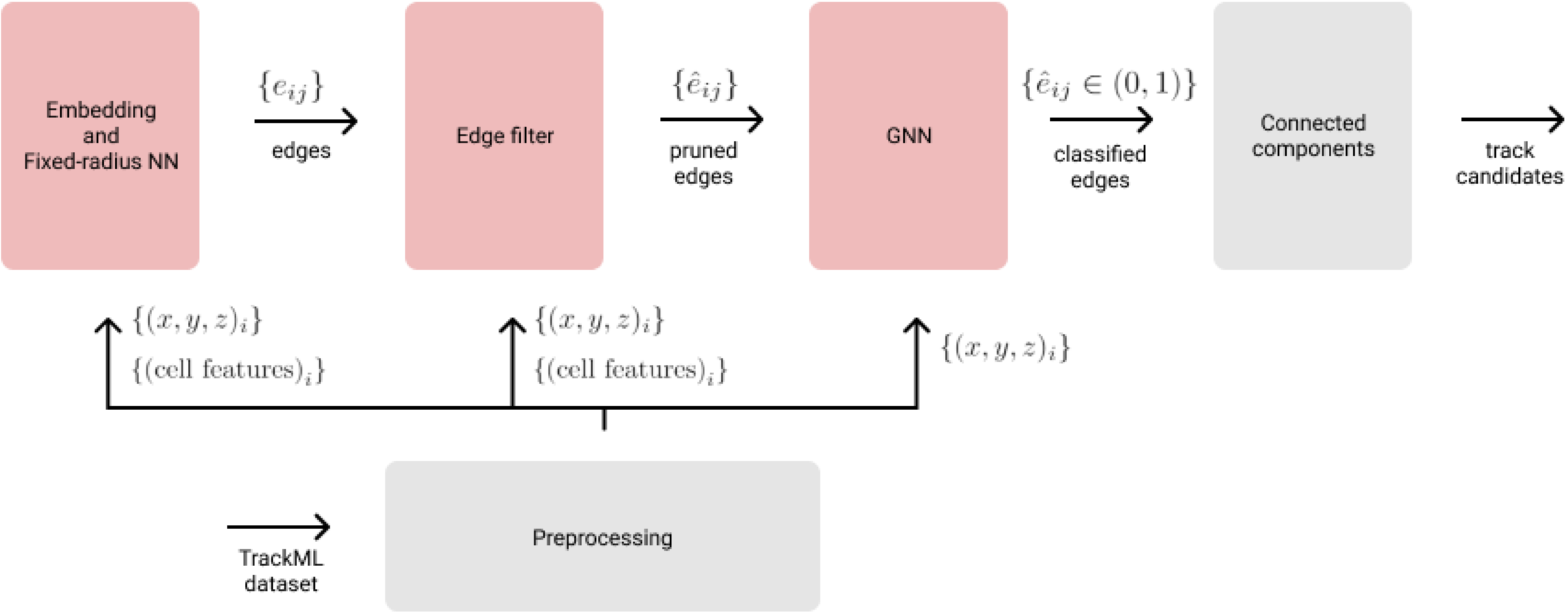
https://deepmind.google/discover/blog/towards-understanding-glasses-with-graph-neural-networks/

Graphs for Tracking

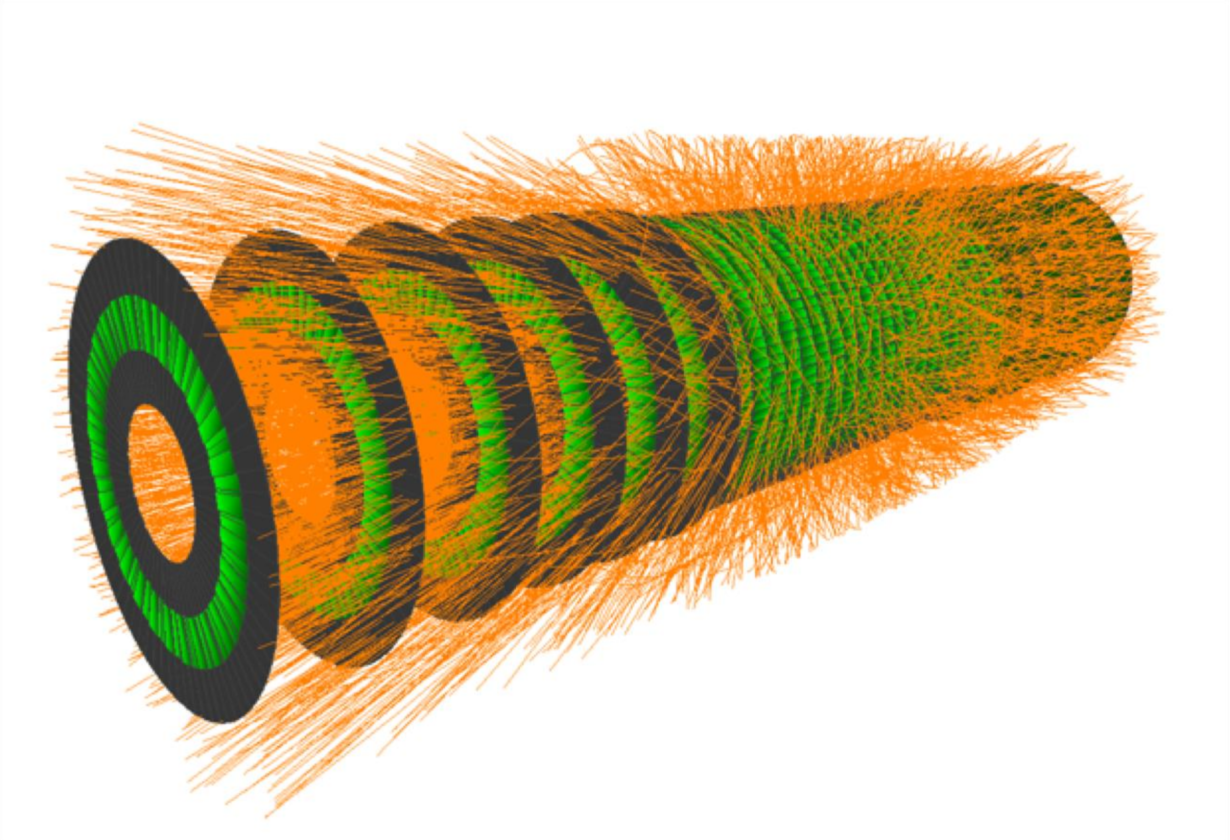


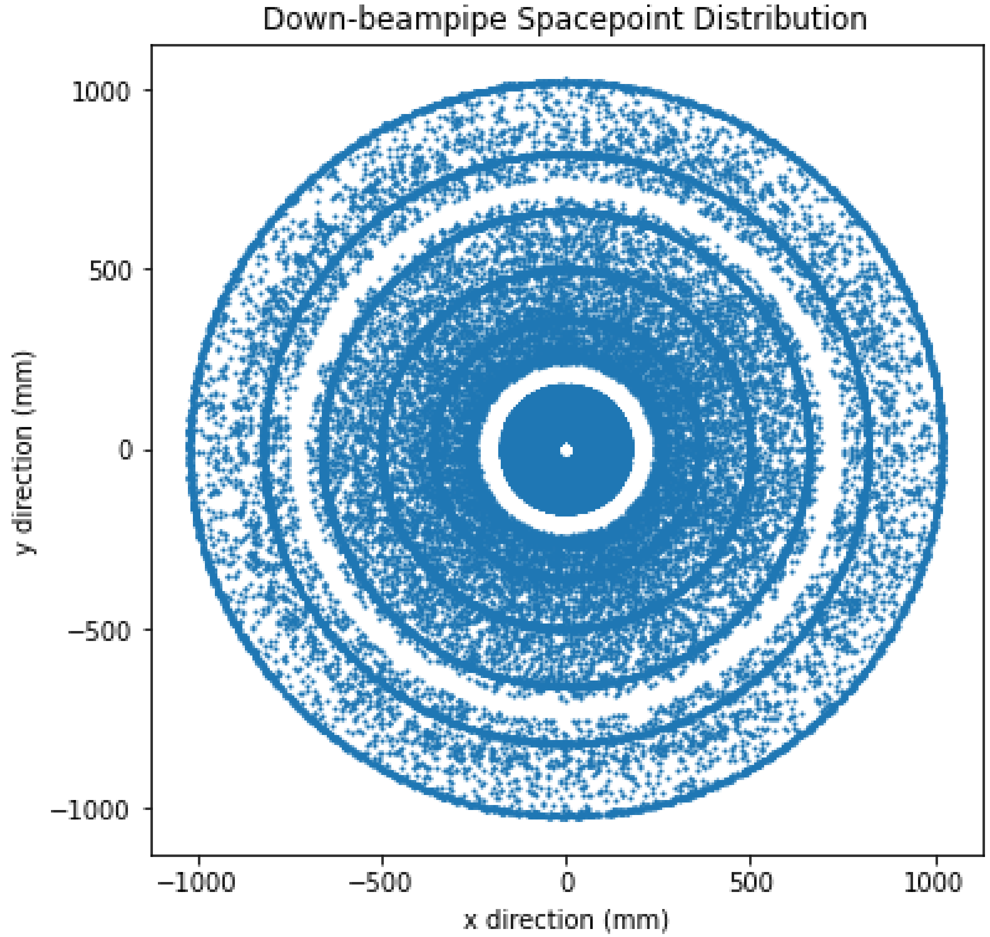
https://www.nature.com/articles/s42254-023-00569-0

Exa.TrkX Pipeline



Data



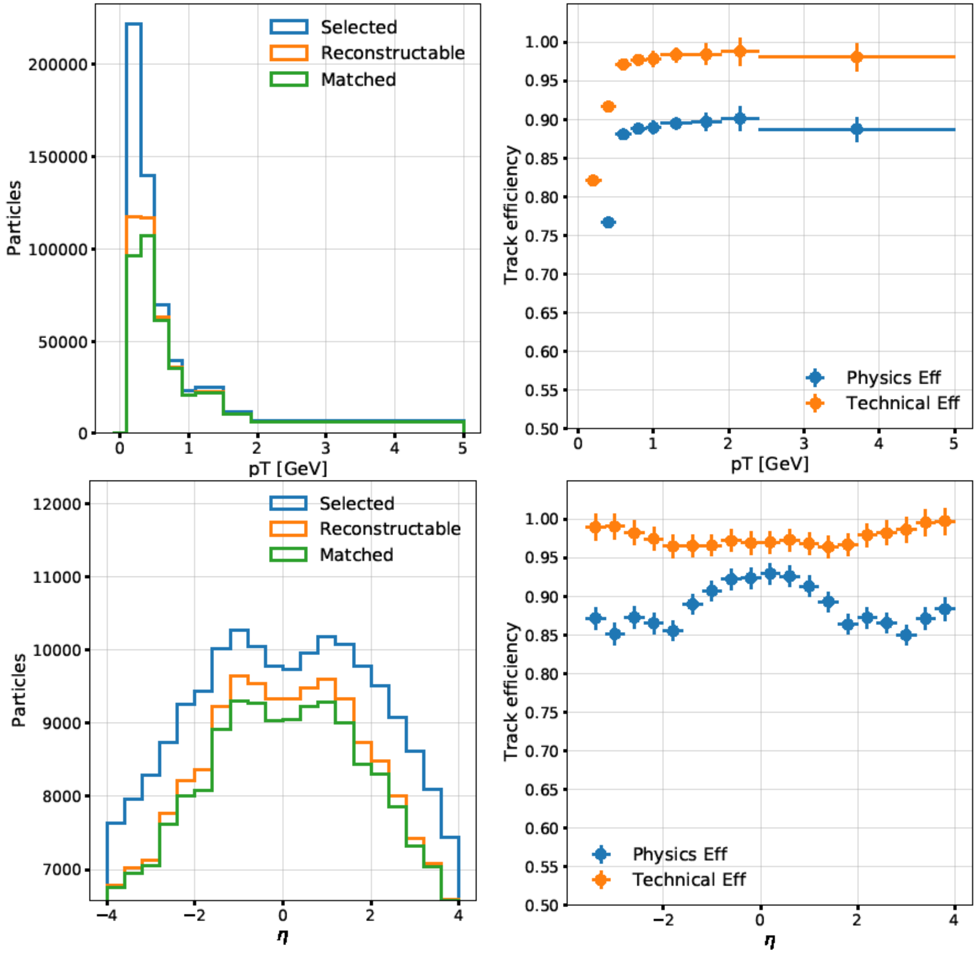

Down-beampipe Spacepoint Distribution
1000
500
0
−500
−1000
y direction (mm)
−1000
−500
0
500
1000
x direction (mm)

Performance and Results

$$\epsilon_{\text{phys}} = \frac{N_{particles}(\text{selected, matched})}{N_{particles}(\text{selected})}$$

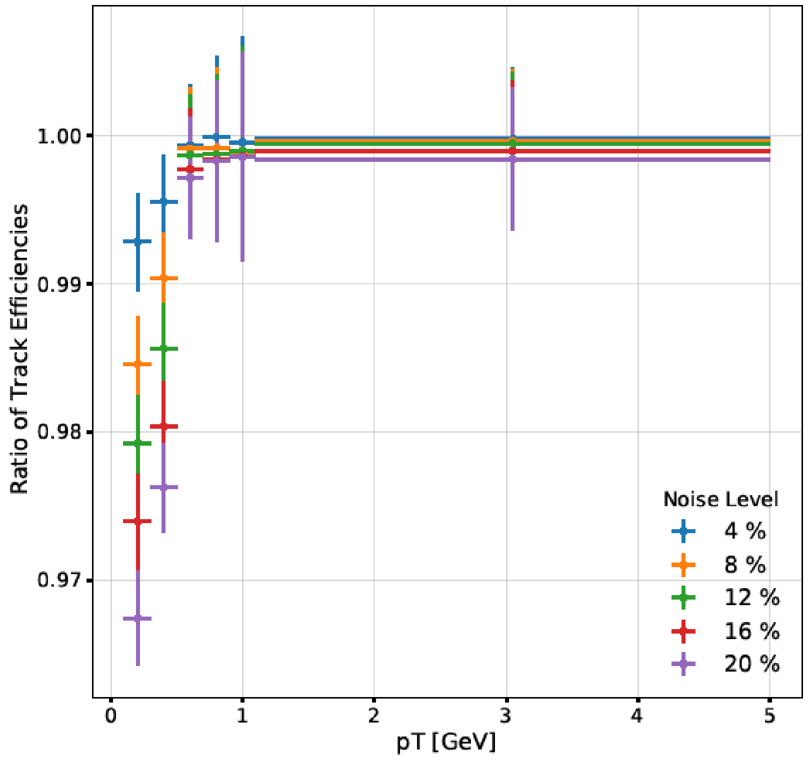
$$\epsilon_{\text{tech}} = \frac{N_{particles}(\text{selected, reconstructable, matched})}{N_{particles}(\text{selected, reconstructable})}$$

$$\text{Purity} = \frac{N_{tracks}(\text{selected, matched})}{N_{tracks}(\text{selected})}$$



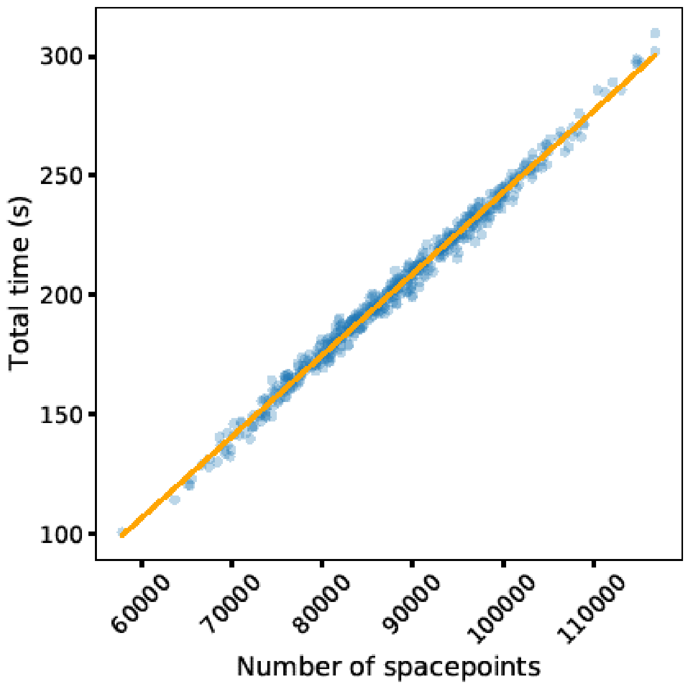
Noise Study

Noise (%)	ϵ_{tech}	Purity
0	91.5	59.3
4	91.5	59.3
8	91.1	58.0
12	90.9	56.8
16	92.2	54.8
20	89.9	53.9

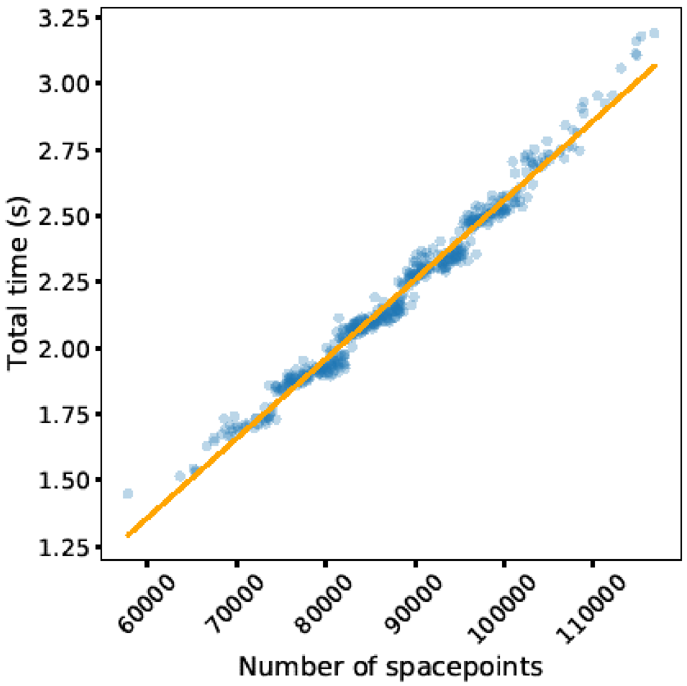


Timing Study

CPU



GPU



Conclusion

- Showed that a deep learning approach to tracking can achieve linear scaling.
- Latency reduced by 100x using GPU, but still too slow for now.
- Robust to detector noise and pile-up.
- Need further studies:
  - Detector-specific (verify performance)
  - Algorithmic / hardware (reduce latency)

Study 2: DeepCalo

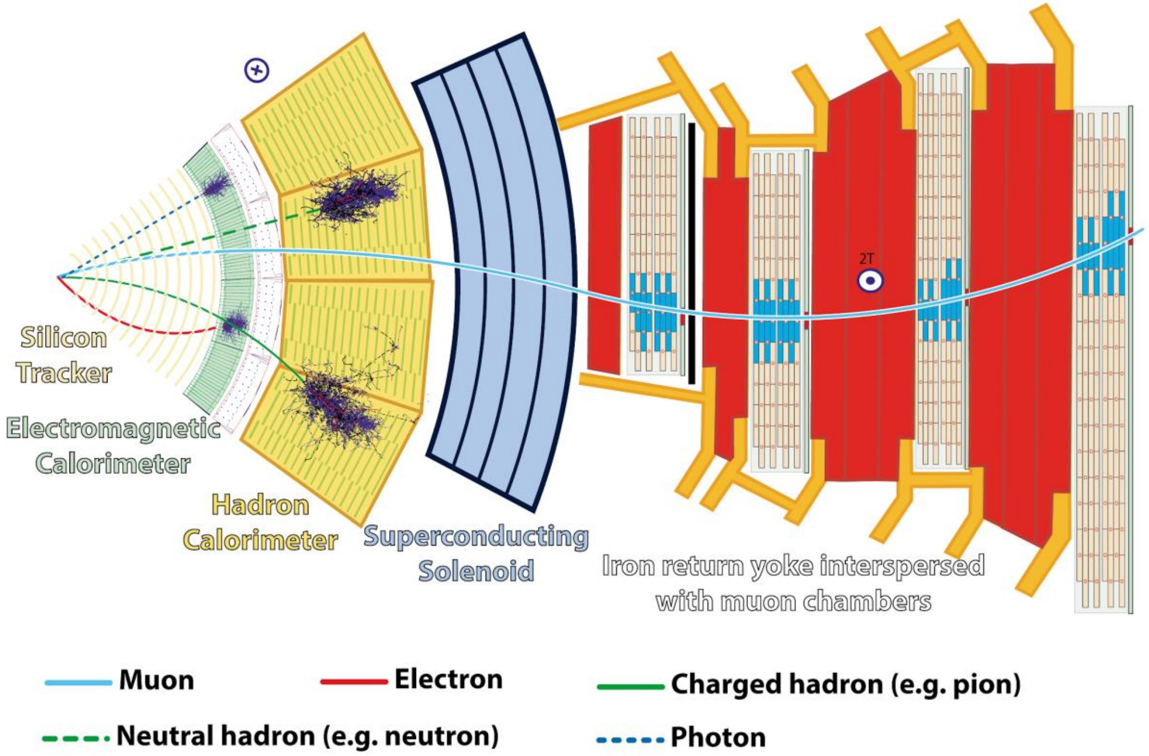
- FUNDAMENTALS
- METHODOLOGY
- RESULTS

Particle Energy Reconstruction

Comes at the end of a complicated chain

- Tracking (previous study)
- Calorimeter clustering (see next study)
- Particle flow

Objective: estimate energy of particle given lower-level information

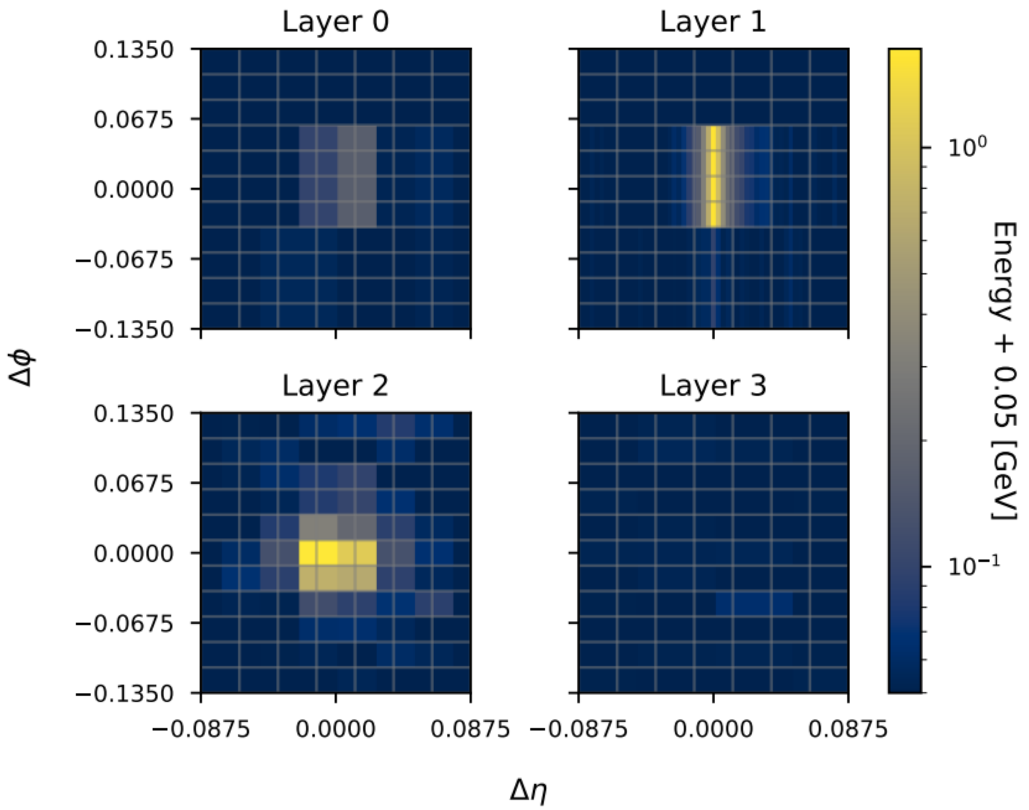


Data

$Z \rightarrow ee$ decays in the forward region $|\eta| > 3.1$ are used.

Many inputs:

- ECAL images (energy, time, noise, gain)
- Scalar variables
- Tracks

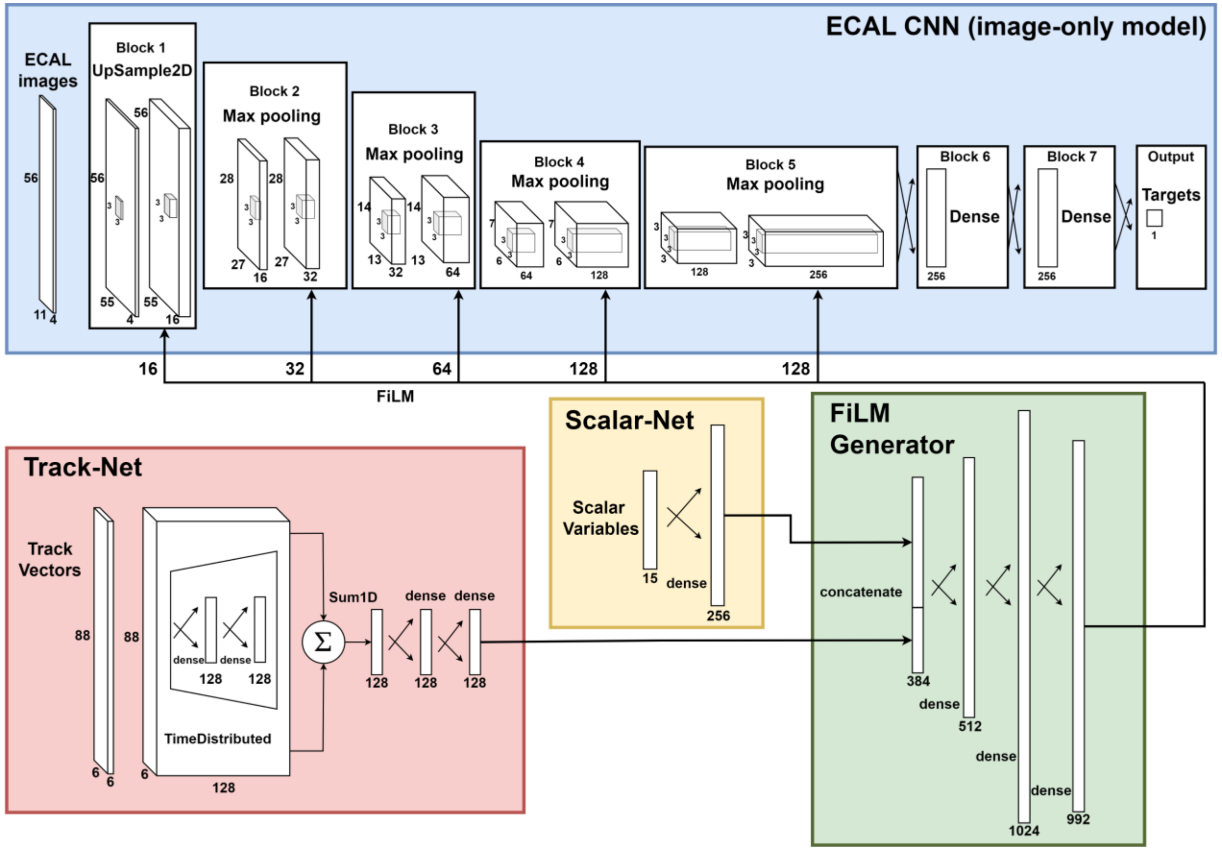


DeepCalo

Deep learning model built to improve electron/photon energy regression

Two models:

- Full model (shown on right)
- Image-only (CNN only)



Field Programmable Gate Arrays (FPGAs)

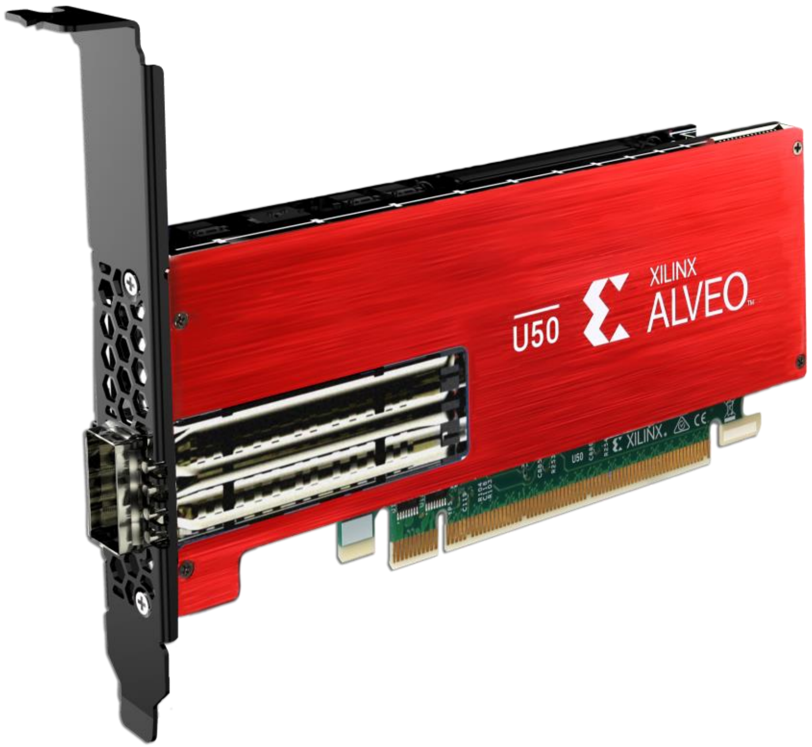
Digital integrated circuits that are **configurable** after manufacture.

Consist of:

- Basic configurable logic blocks (CLBs)
- Programmable interconnects
- RAM
- DSP blocks

More flexible than GPUs, allowing for **higher efficiency** and **lower latency**.

Designed using HLS4ML, which is a high-level tool to implement ML on FPGAs.



Quantization

FLOATING POINT

Can represent wide range of magnitudes and precisions.

Common in CPUs and GPUs.

$$\underbrace{(1.100011)_2}_{\text{Significand}} \times 2^{\overbrace{3}^{\text{Exponent}}} = 12.375$$

FIXED POINT

Less flexible, but generally simplifies design, leading to higher efficiency and reduced cost.

Common in embedded systems (e.g., FPGAs).

$$\underbrace{1100}_{\text{Integer bits}}.\underbrace{0110}_{\text{Fractional bits}} = 12.375$$

Quantization

How should we convert a floating-point model to a fixed-point model with lower precision?

- Post-training quantization (PTQ): approximate each weight/bias with closest fixed-point equivalent.
- Quantization-aware training (QAT): simulate quantization during the training process.

Tuning Precision

To optimize the precision for PTQ and QAT, we used the following two-step approach:

1. Scan bit widths from 32 to 2 bits, with integer bits varying for PTQ.
2. Scan the same bit widths, but with fixed integer/fractional bit ratio based on (1.) for QAT.

PTQ vs QAT

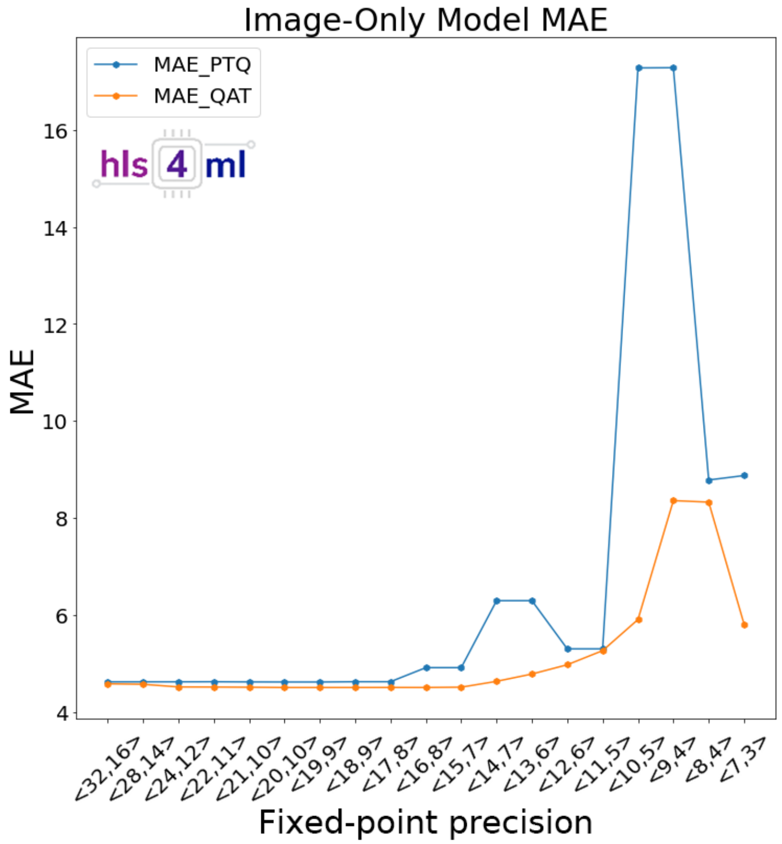
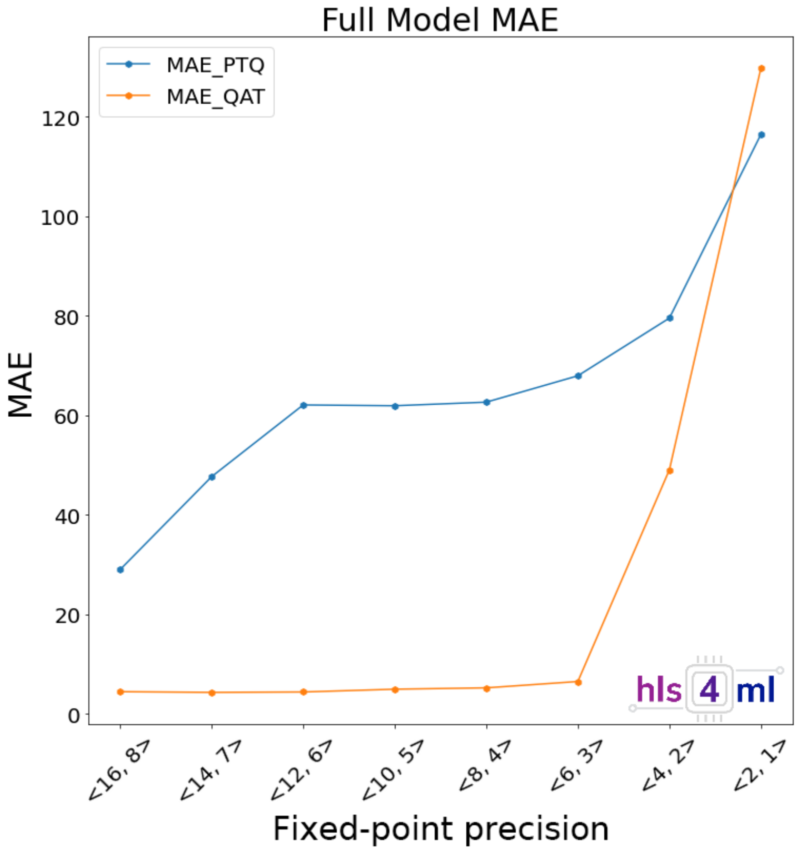
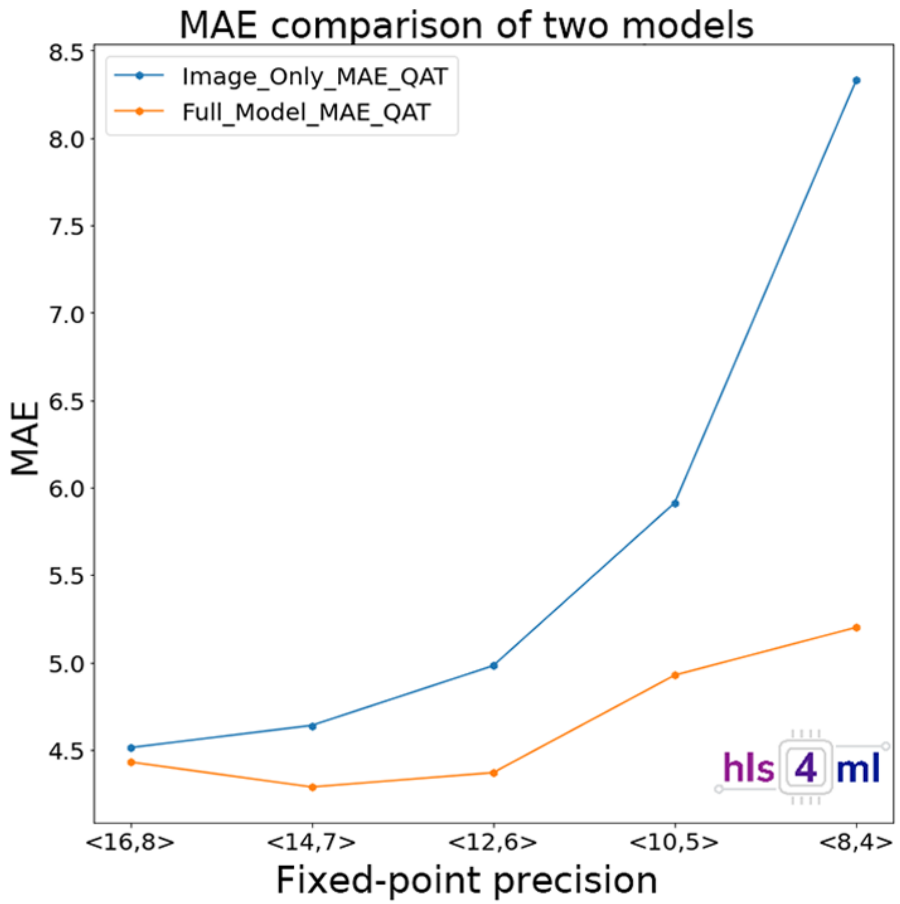


Image-Only Model MAE
MAE_PTQ
MAE_QAT
hls 4 ml
MAE
4
6
8
10
12
14
16
<32,16>
<28,14>
<24,12>
<22,11>
<21,10>
<20,10>
<19,9>
<18,9>
<17,8>
<16,8>
<15,7>
<14,7>
<13,6>
<12,6>
<11,5>
<10,5>
<9,4>
<8,4>
<7,3>
Fixed-point precision


Full Model MAE
MAE_PTQ
MAE_QAT
hls 4 ml
MAE
0
20
40
60
80
100
120
<16, 8>
<14, 7>
<12, 6>
<10, 5>
<8, 4>
<6, 3>
<4, 2>
<2, 1>
Fixed-point precision

Full Model vs Image-only Model



Latency

Coprocessor	CPU			GPU			FPGA	
Type	Ryzen 7 3700X	Ryzen 5 5600H	AMD EPYC 7262	RTX 2070 Super	Tesla V100	RTX 2080 Ti	single-stream	mixed-type
Batch=1								
Latency	7.52ms	8.75ms	5.865ms	8.47ms	4.8ms	8.2ms	1.106ms	0.898ms
Speedup	1.164×	**1**×	1.492×	1.033×	1.823×	1.067×	7.911×	**9.744**×
Power	53.73W	29.13W	42.65W	49.77W	60.11W	64.54W	19.76W	20.75W
Energy	404.05mJ	254.888mJ	250.142mJ	421.552mJ	288.528mJ	529.228mJ	21.855mJ	18.634mJ
Batch=5								
Latency	11.5ms	13.45ms	10.545ms	9.75ms	5.1ms	7ms	2.695ms	1.485ms
Speedup	1.17×	**1**×	1.275×	1.379×	2.637×	1.921×	4.991×	9.057×
Power	62.44W	37.67W	48.94W	51.83W	61.73W	84.18W	21W	23.775W
Energy	718.06mJ	506.66mJ	516.07mJ	505.345mJ	314.825mJ	589.26mJ	56.595mJ	35.305mJ

Conclusion

Deploying deep learning models to FPGAs can further reduce latency while preserving accuracy through appropriate optimization and design (including use of QAT).

- Image-only: 14.1x (9.7x) speedup compared to CPU (GPU)
- Full model: 7.9x (5.3x) speedup

Study 3: SPVCNN for Hadronic Calorimetry Clustering

- FUNDAMENTALS
- METHODOLOGY
- RESULTS

Jets

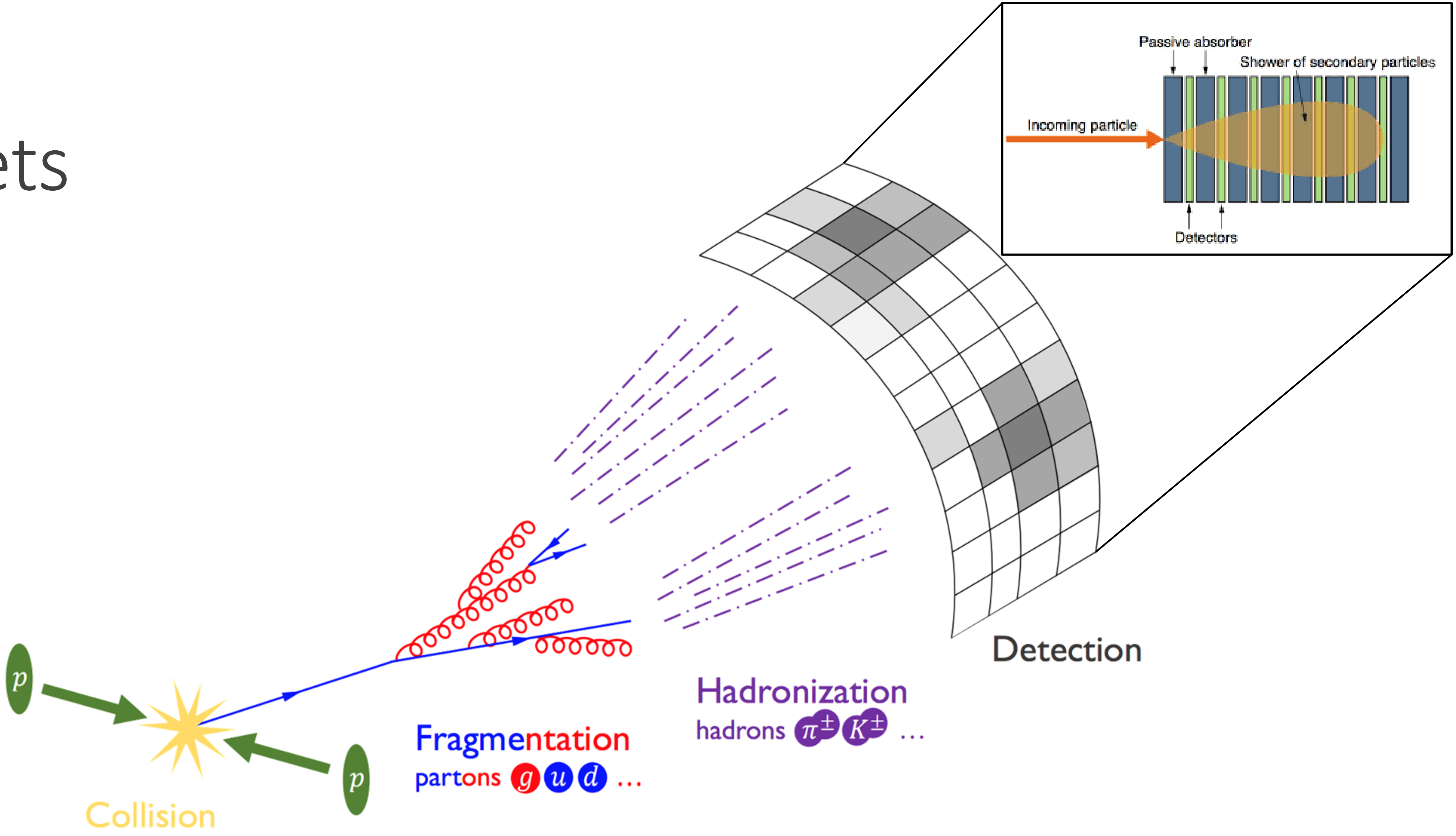
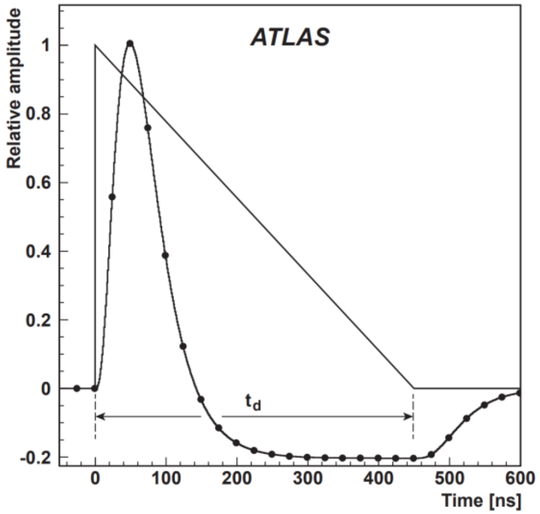
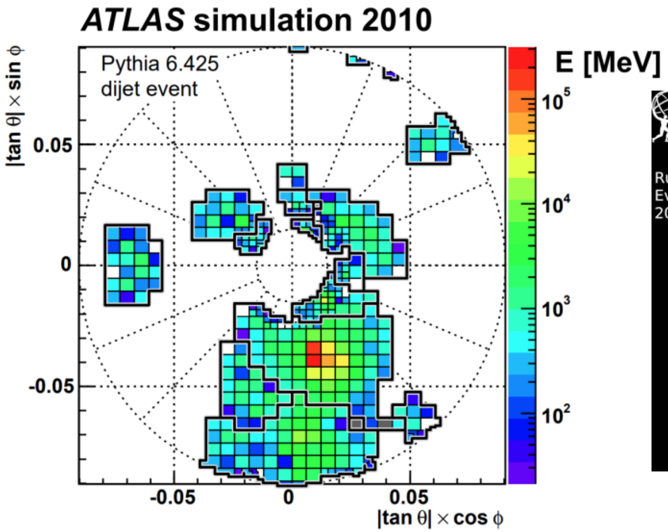
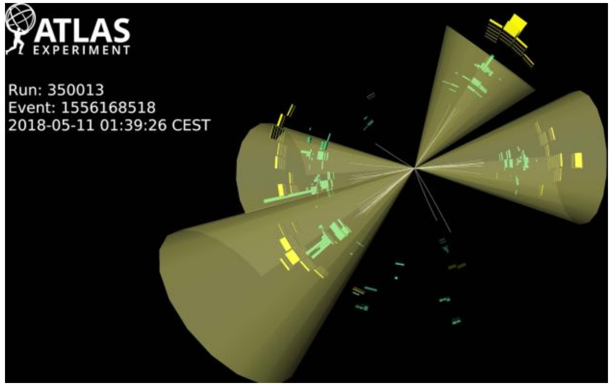


Image: https://www.ericmetodiev.com/post/jetformation/



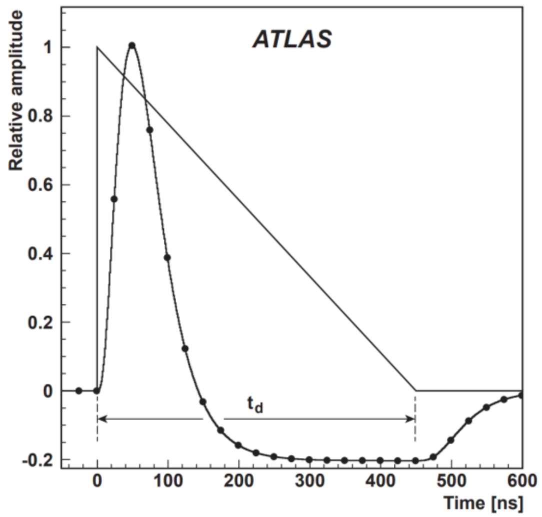


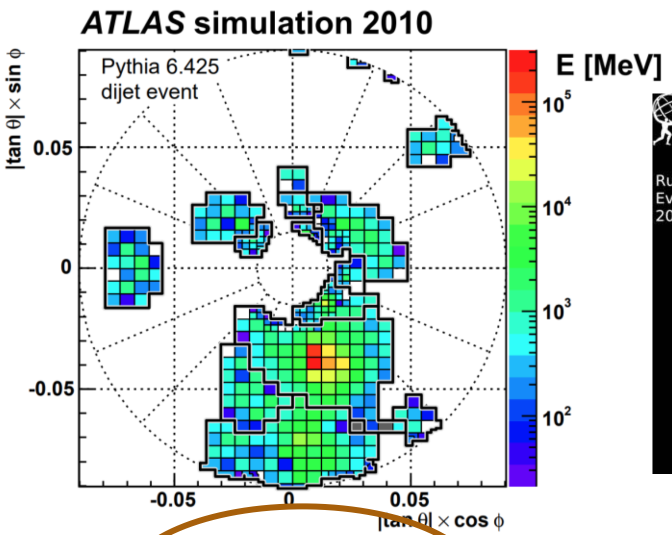


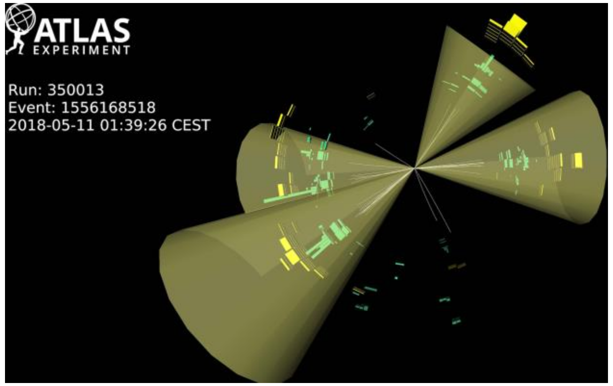
Digitization → **Clusters** → **Particles / Jets**

Images (left, middle): https://arxiv.org/pdf/1603.02934.pdf

Image (right): https://atlas.cern/updates/briefing/double-Higgs-to-bottoms







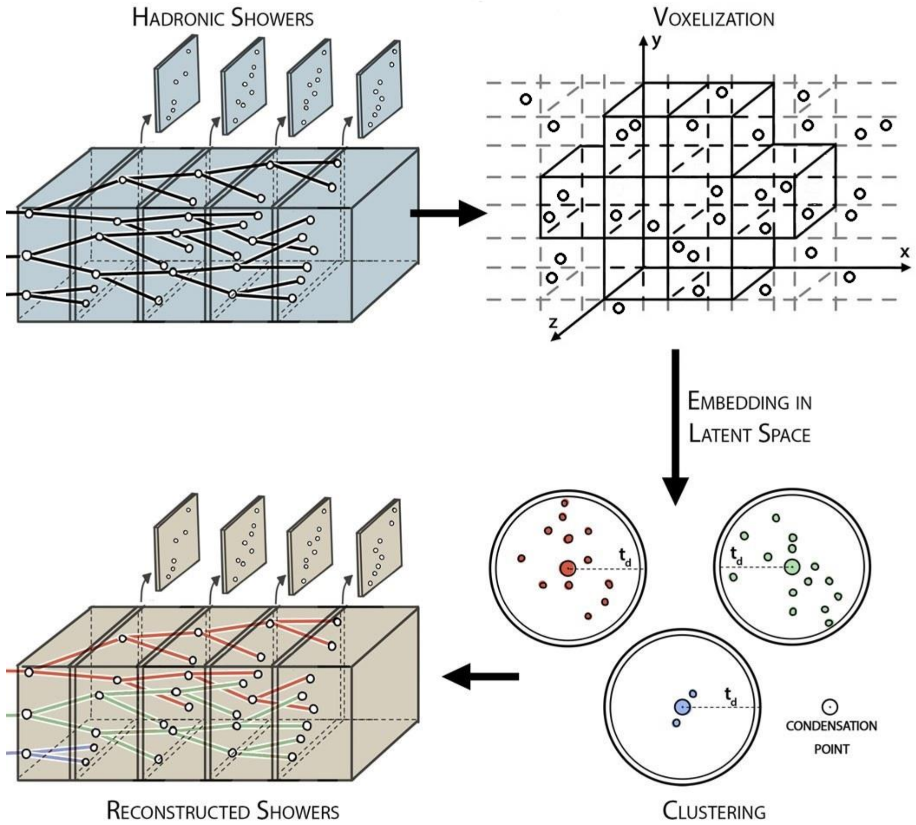
Digitization → **Clusters** → **Particles / Jets**

Images (left, middle): https://arxiv.org/pdf/1603.02934.pdf

Image (right): https://atlas.cern/updates/briefing/double-Higgs-to-bottoms

Clustering Methodology With SPVCNN

- **Hadronic Showers**: incident particles shower upon interaction with passive material, possibly producing several hits in 3D space.
- **Voxelization**: Hits are mapped to a regular grid for convolutional processing.
- **Clustering**: NN maps hits to a 5+1D embedded space. A bounded nearest-neighbor method clusters the hits.
- **Reconstructed Showers**: Clustering assignments reconstruct incident particle showers. This information is passed to downstream algorithms which perform energy regression, jet clustering, etc.



SPVCNN Motivation

Achieved first place on SemanticKITTI leaderboard

Designed for **3D tasks** that require:
- **Low latency**
- **High computational efficiency**
- **High accuracy**

Original motivating problem was driverless cars.

Reconstruction in particle physics shares many of the same requirements.

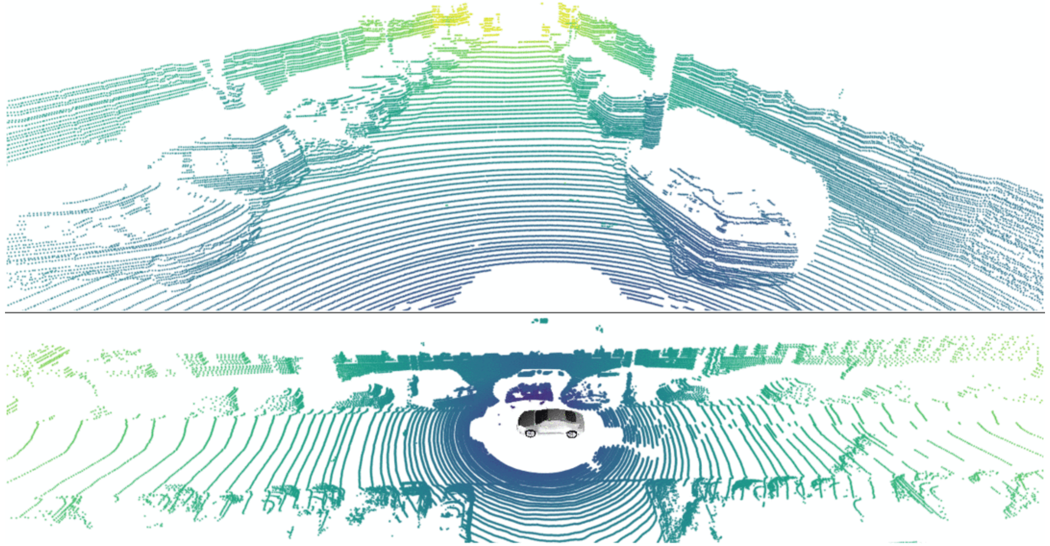
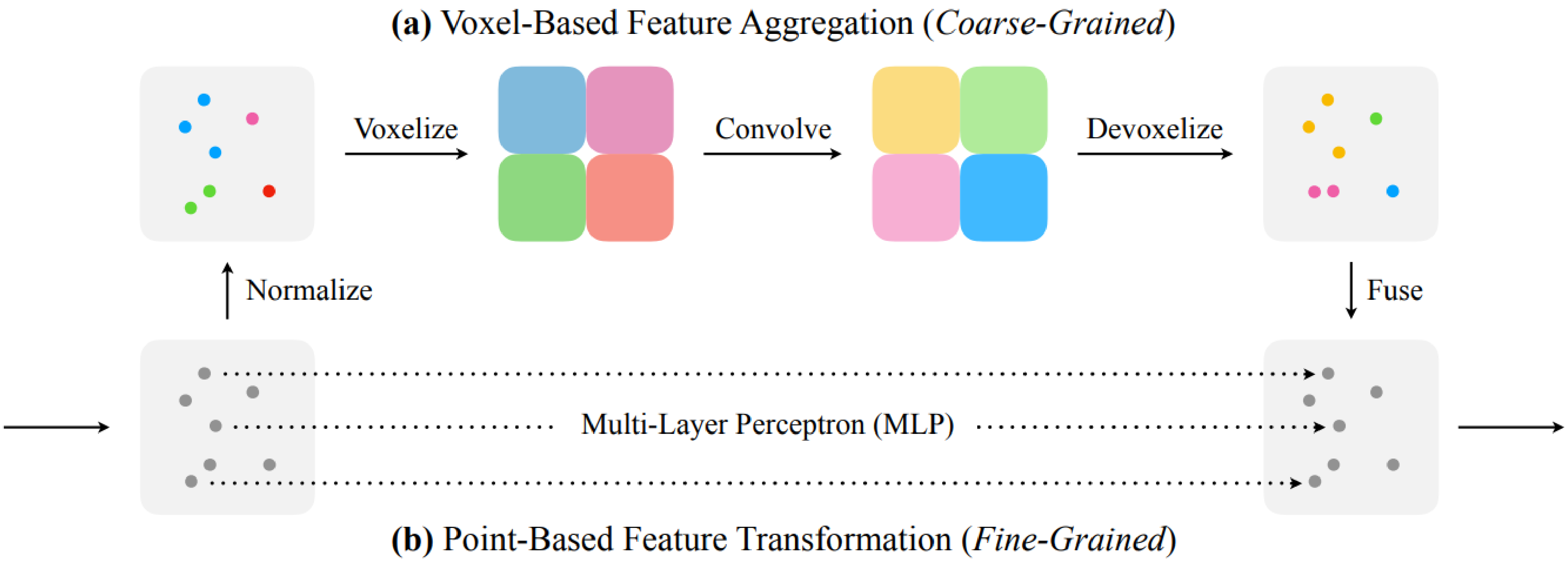
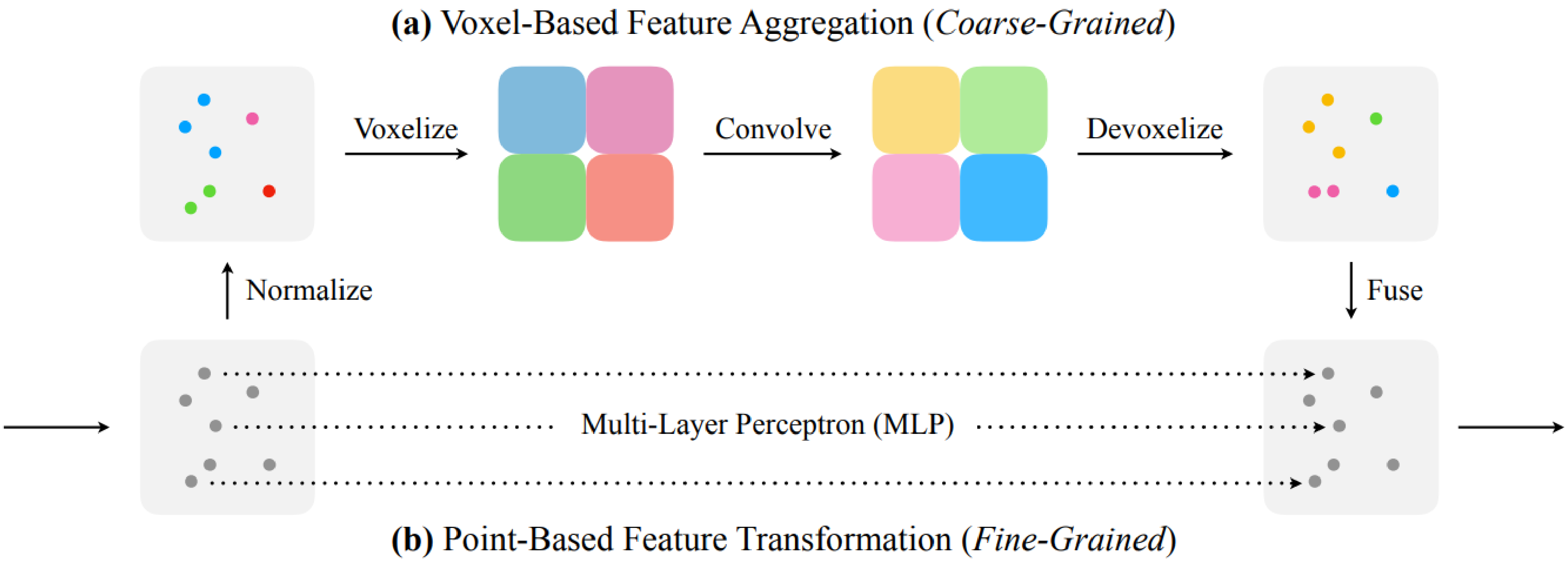


Image: http://lidar-panoptic.cs.uni-freiburg.de/

Point-Voxel Convolution (PVConv)

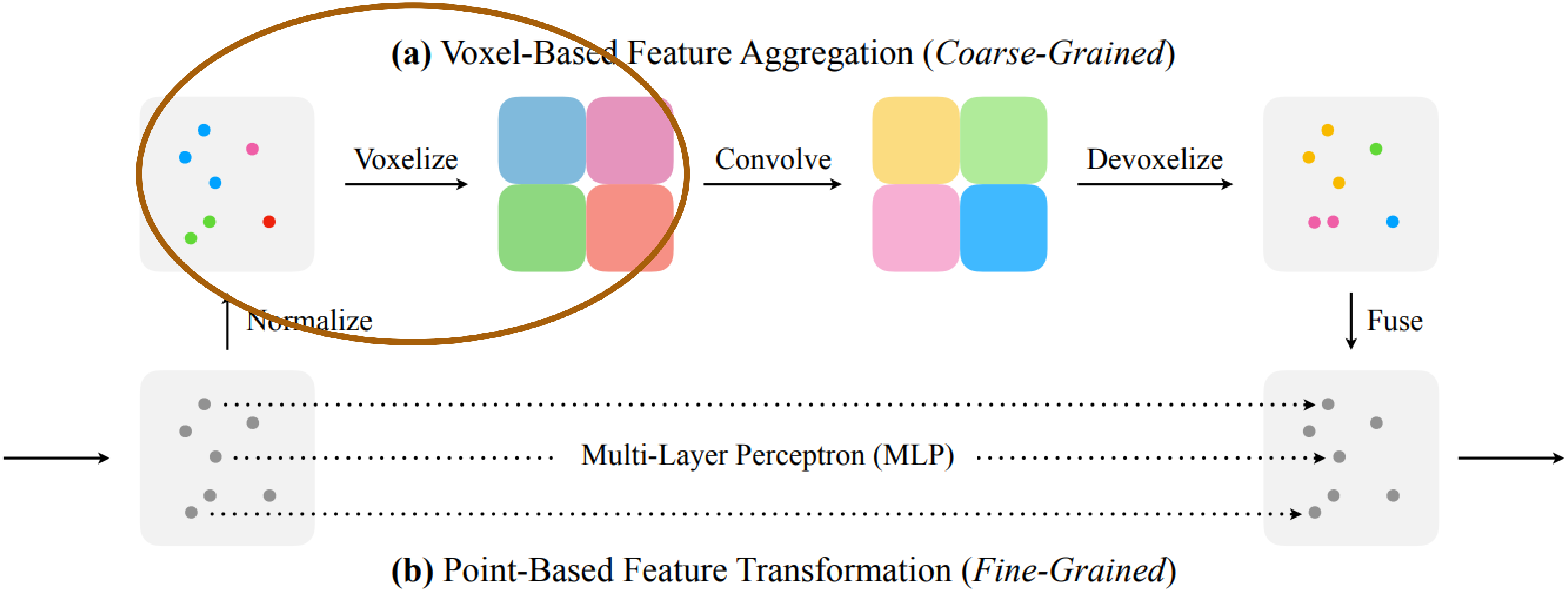


Sparse Point-Voxel Convolution (SPVConv)



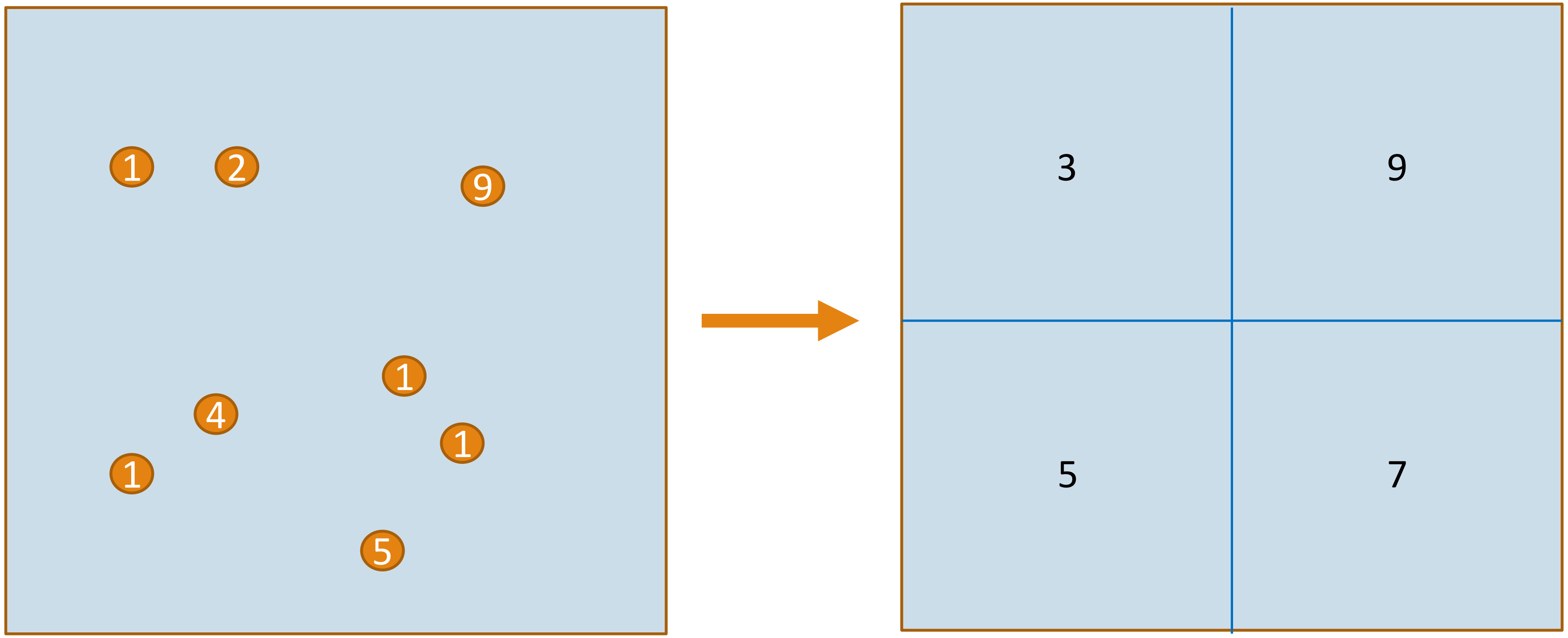
- Simply replaces upper branch with sparse convolution.
- Some details with normalization/voxelization and devoxelization/fusion:
  - Hashing, trilinear interpolation

Sparse Point-Voxel Convolution (SPVConv)

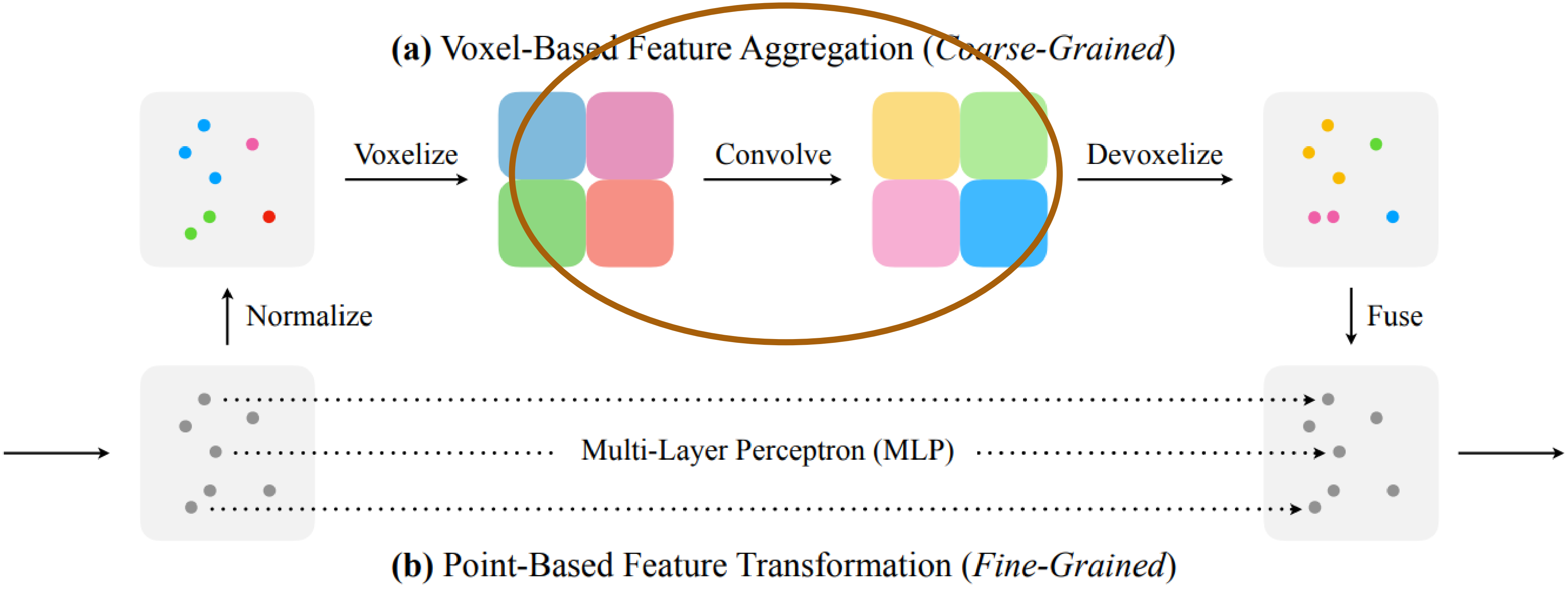


- Simply replaces upper branch with sparse convolution.
- Some details with normalization/voxelization and devoxelization/fusion:
  - Hashing, trilinear interpolation

Voxelization

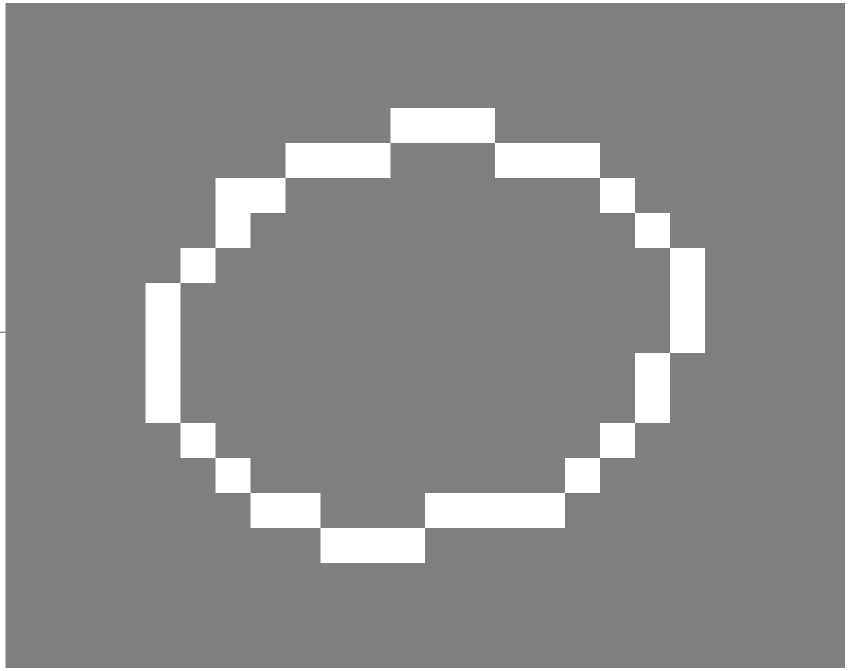


Sparse Point-Voxel Convolution (SPVConv)



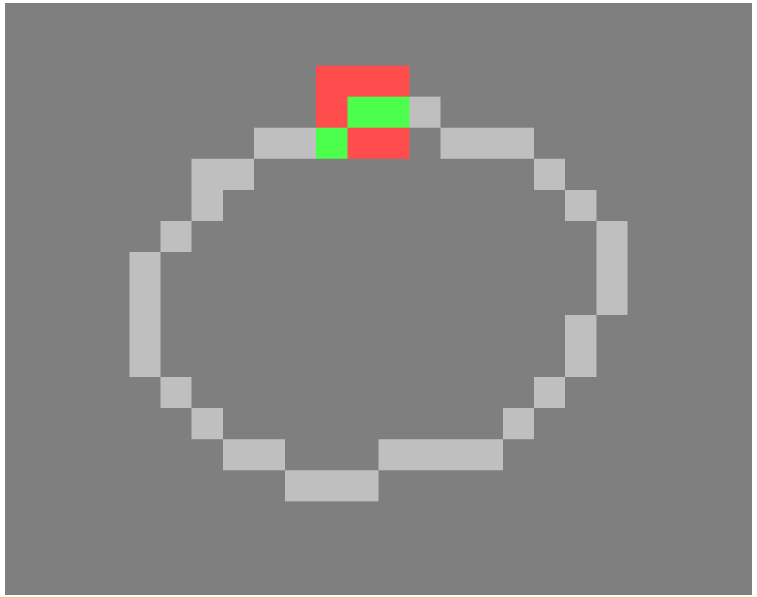
- Simply replaces upper branch with sparse convolution.
- Some details with normalization/voxelization and devoxelization/fusion:
    - Hashing, trilinear interpolation

Generalized Sparse Convolution

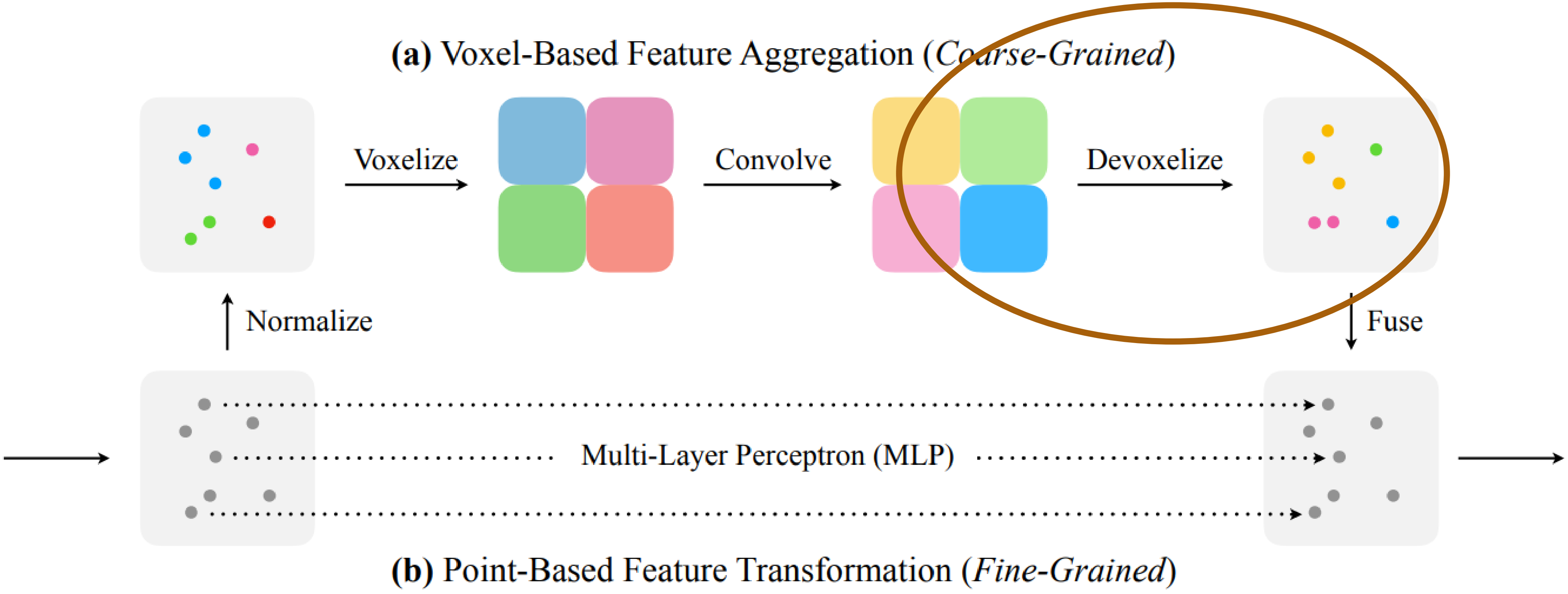


- Sparse convolutions operate **directly on sparse tensors**.
- **Avoids wasted computation** and allows for **higher resolution**.

- Naïve implementations (top) would quickly reduce sparsity.
- Modern implementations (bottom) allow for arbitrary input (c_{in}) and output (c_{out}) coordinates. The example shown is a 'submanifold sparse convolution', which sets $c_{in} = c_{out}$, thus preserving sparsity. This is (almost) used in SPVCNN.



Devoxelization



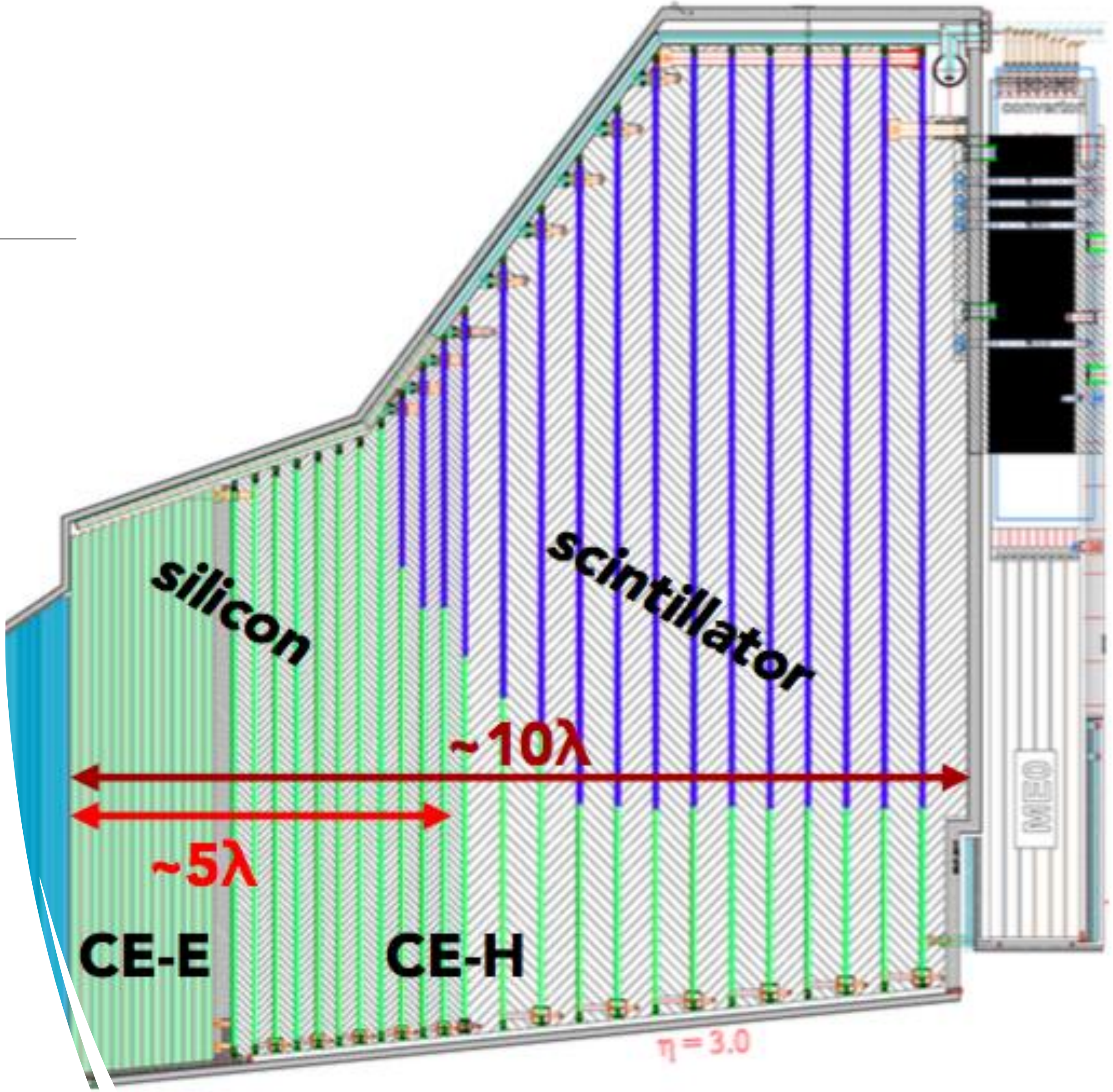
- Simply replaces upper branch with sparse convolution.
- Some details with normalization/voxelization and devoxelization/fusion:
    - Hashing, trilinear interpolation

CMS High-Granularity Calorimeter (HGCAL)

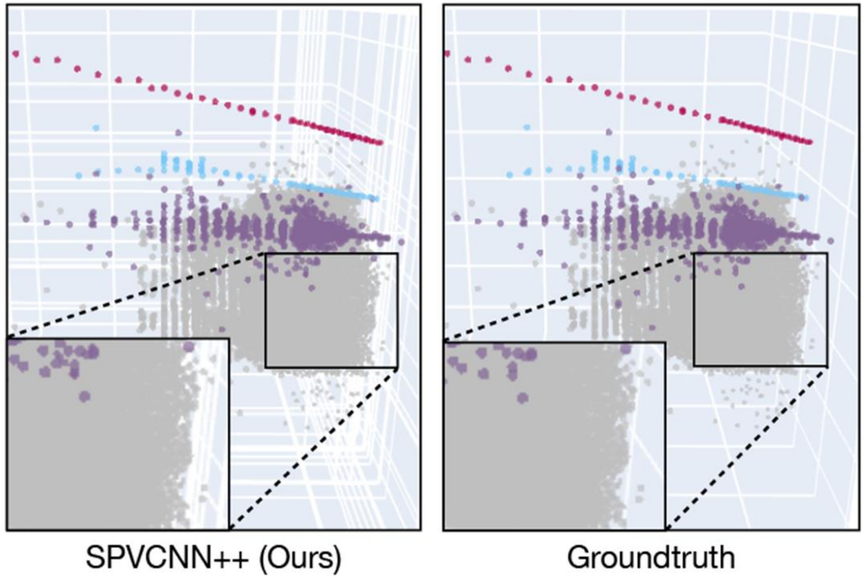
Major upgrade for HL-LHC: **6.5M channels**, **50 layers.**

Finer granularity, timing resolution → greater benefit from 3D deep learning.

Despite increased data volume, cannot sacrifice latency.



HGCAL Results



Left – predicted clusters from SPVCNN.
Right – event display from HGCAL.

Each point represents an energy deposit in the calorimeter. Each color corresponds to a cluster.

HGCAL Results

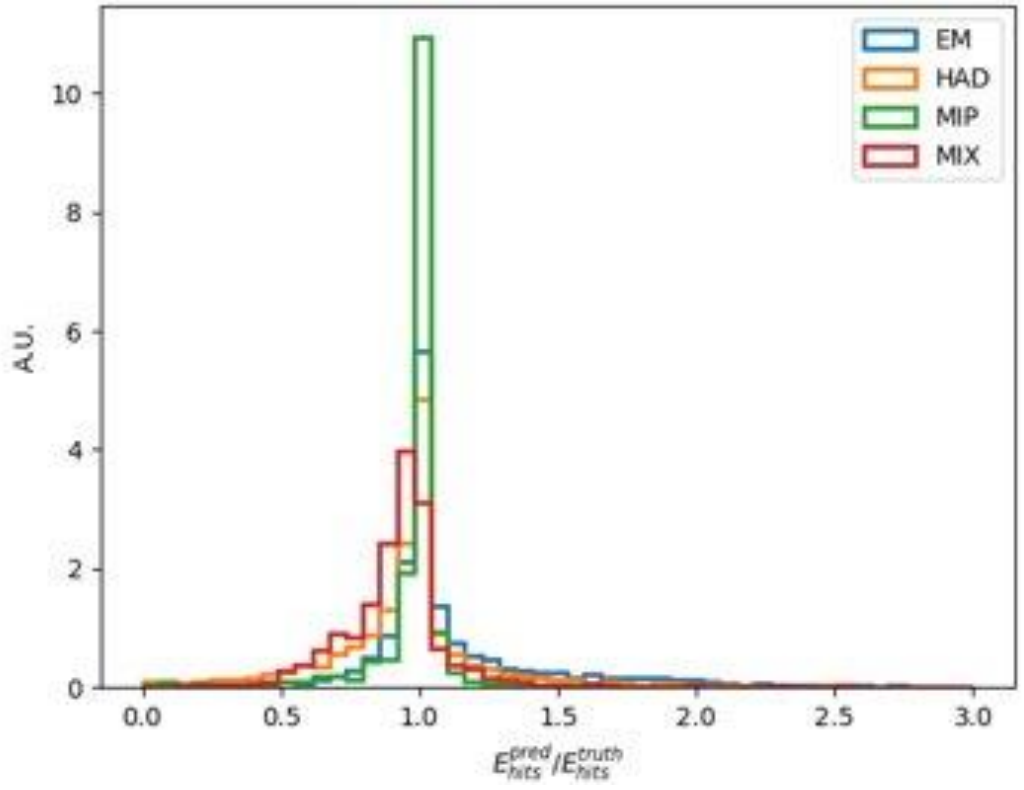
	mIoU	SQ	RQ	PQ
GravNet	0.93	0.89	0.74	0.69
GravNet (optimized)	0.93	0.90	0.83	0.76
SPVCNN	0.98	0.92	0.85	0.80

IoU – measure of overlap between predicted and true classes (signal and noise).
SQ – average overlap between predicted and true clusters for each semantic class.
RQ – fraction of clusters for each semantic class that were matched.
PQ – product of SQ and RQ.

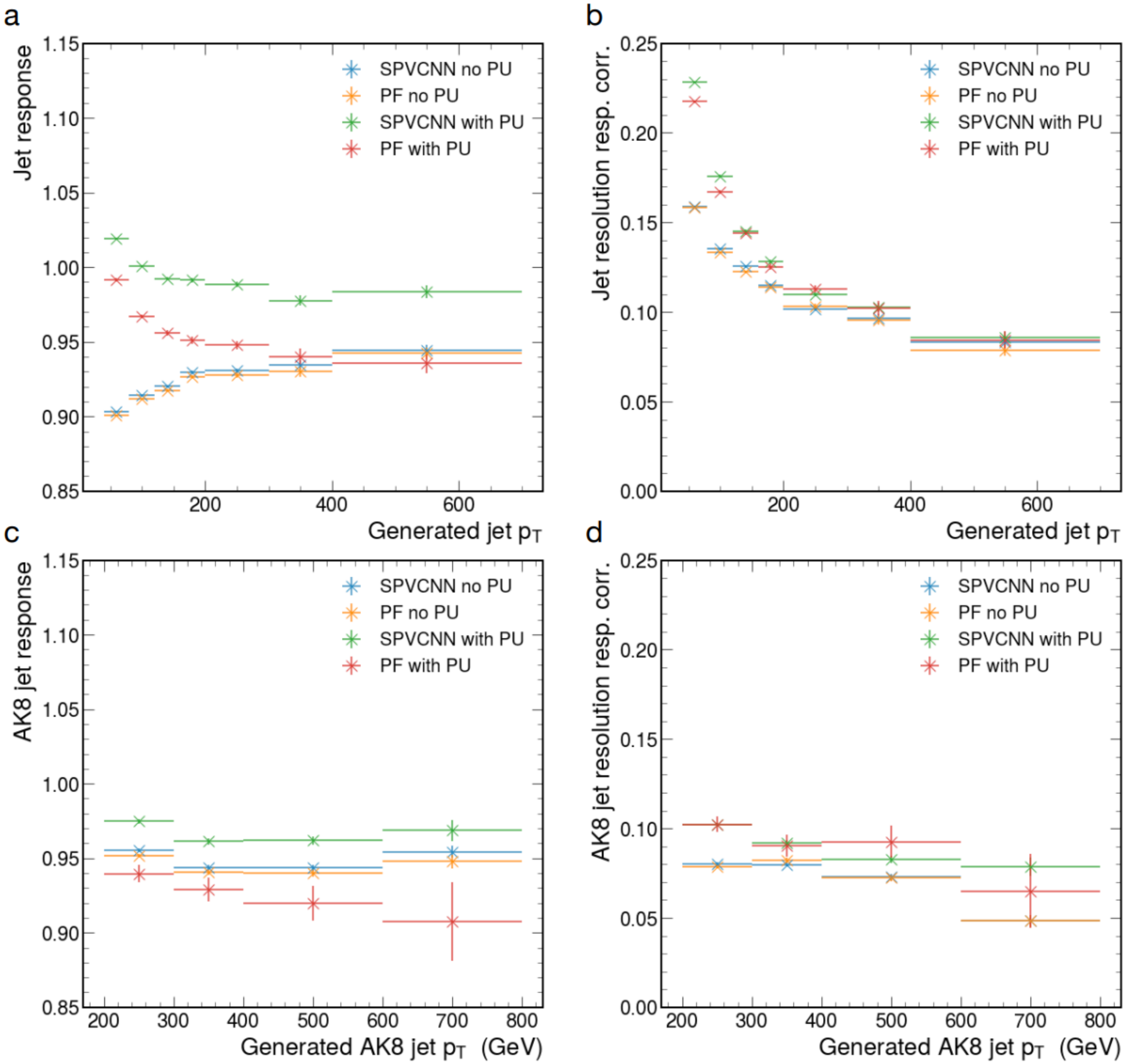
HGCAL Results

Right: **ratio of predicted to true energy** for each predicted cluster, split into four types:

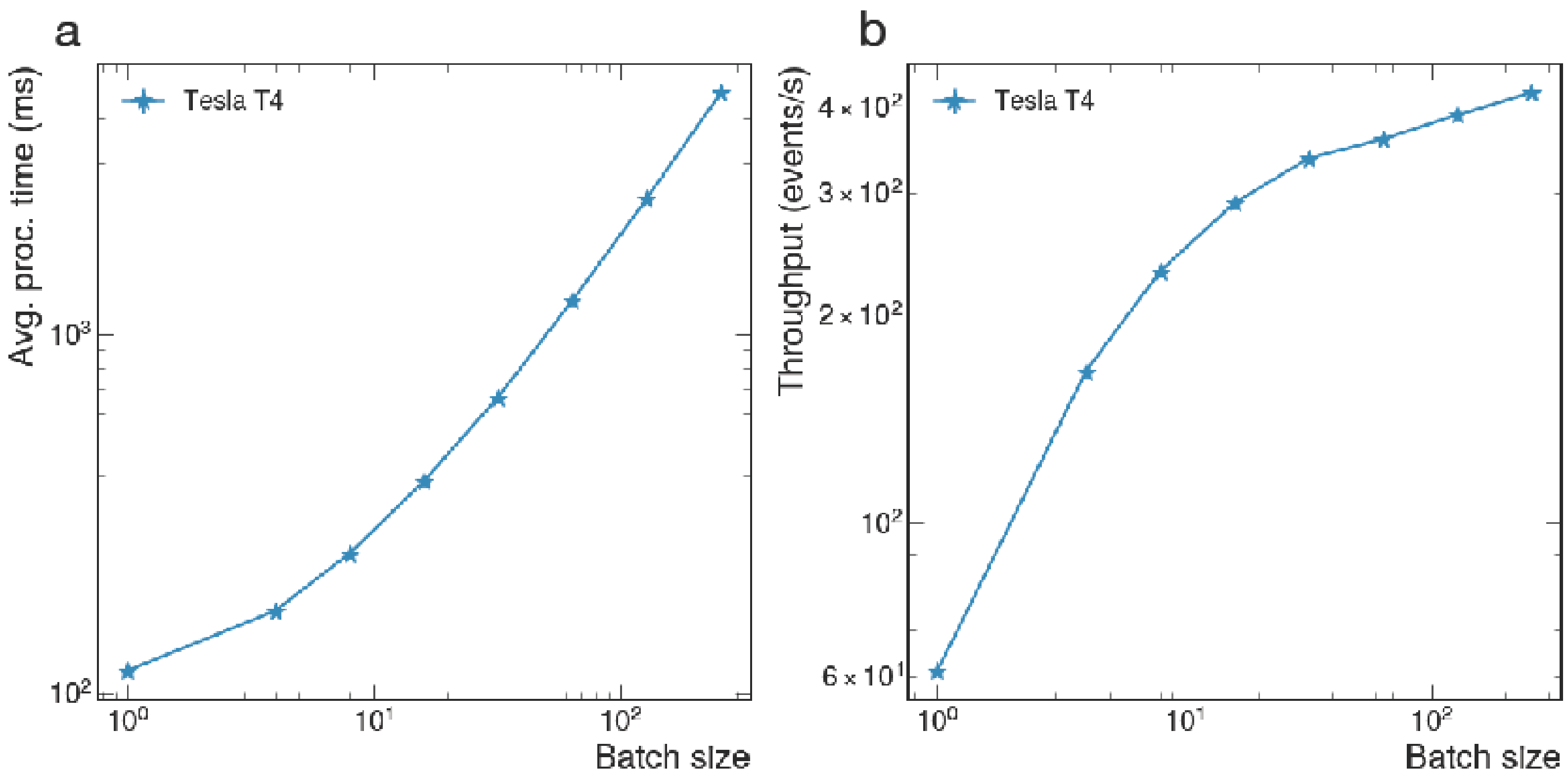
- Electromagnetic (EM) particles
- Hadronic (HAD) particles
- Minimum-ionizing particles (MIP)
- A mixture of the above (MIX)



HCAL Results



Latency / Throughput



Future Implications

Modern convolutional approaches that exploit tricks for **efficient computation** are **competitive** with **current clustering methods** and other proposed **ML methods** for the HL-LHC**.**

Latency at the level needed for the HLT **(~ms)** is **achievable with GPU accelerators**. Beyond this level, further innovations are probably required, e.g., exploiting FPGAs and ASICs.

Conclusion

Wrap-up

Advancements in particle detector technology (e.g., the **HL-LHC)** have the potential to **address outstanding problems in particle physics**.

Deep learning methods on GPUs or FPGAs can **overcome challenges** in performance at higher energies and luminosities predicted with conventional approaches.

This thesis showcased several such examples in **tracking, calorimetry clustering, and energy regression**, which are crucial steps in event reconstruction.

Further development along these lines will likely significantly enhance the effectiveness of the HL-LHC and future detectors.

Work

Papers

- An, F., … Schuy, A., Hsu, S. C., ... et al. Precision Higgs physics at the CEPC. Chinese Physics C, 43(4), 043002, 2019.
- Kiuchi, R., … Schuy, A., Hsu, S. C., ... et al. Physics potential for the $H \rightarrow ZZ^*$ decay at the CEPC, 2021. The European Physical Journal C, 81, 1-9.
- **Ju, X., Murnane, … Schuy, A., Chauhan, A., Hsu, S. C., … et al. Performance of a geometric deep learning pipeline for HL-LHC particle tracking. Eur. Phys. J. C 81, 876 (2021). ***
- Belloni, A., ... Schuy, A., Khoda, E., … et al. Report of the Topical Group on Electroweak Precision Physics and Constraining New Physics for Snowmass 2021, 2022. arXiv:2209.08078.
- Abbott, B., … Schuy, A., Khoda, E., Hsu, S. C., … et al. Anomalous quartic gauge couplings at a muon collider, 2022. arXiv: 2203.08135.
- **Chen, C., … Schuy, A., Hauck, S., Hsu, S. C., … et al. Accelerating CNNs for Particle Energy Reconstruction on FPGAs. Under review. ***
- Abbott, B., … Schuy, A., Khoda, E., Hsu, S. C., … et al. Anomalous production of massive gauge boson pairs at muon colliders.

Talks

- Schuy, A., … Hsu, S. C., … et al, "Extending RECAST for Truth-Level Reinterpretations", DPF 2019. arXiv: https://arxiv.org/abs/1910.10289
- Schuy, A., … Hsu, S. C., … et al, "RECAST for Mono-S(bb) with ATLAS", DM@LHC 2019.
- **Schuy, A., … Hauck, S., Hsu, S. C., … et al, "Low-latency Calorimetry Clustering at the LHC with SPVCNN", Fast Machine Learning for Science Workshop, 2022. ***

Soon-to-be published

- **Schuy, A., … Zhao, H., Hsu, S. C., Hauck, S., … et al. Accelerating Hadronic Calorimetry with Sparse Point-Voxel Convolutional Neural Networks. ***
- Roberts, N., …, Schuy, A., Hsu, S.C., Lin, L. FAIR modeling for Perovskite Solar Cells: An Open-Source Machine Learning Pipeline.

Outreach

- Engineering Discovery Days – April 2019
- QuarkNet Masterclass – 2019 & 2022
- Mentor Interlake high school interns – 2023

NSF grants:
- PHY-2110963
- OAC-2117997
- DMR-2019444

IMOD

* Included in thesis

Acknowledgements

Research Advisor: **Shih-Chieh Hsu**

Mentor: **Scott Hauck**

Post-docs: **Elham Khoda & Ke Li**

Collaborators (too many to list, but thanks in particular to): **Haoran Zhao, Zhijian Liu, Jeff Krupa, Aram Apyan, Phil Harris, Dennis Yin, Thomas Klijnsma, Lukas Heinrich**, and many others…

Special thanks to…

Gordon Watts for encouragement as an undergrad

Catherine Provost for her continued support as the graduate counselor

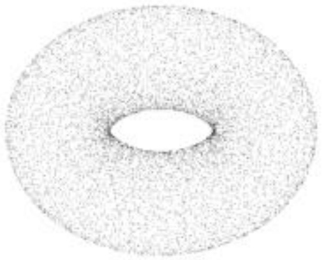
Finally, thanks to the **staff and faculty** of the Physics department and the UW for making my education & research possible

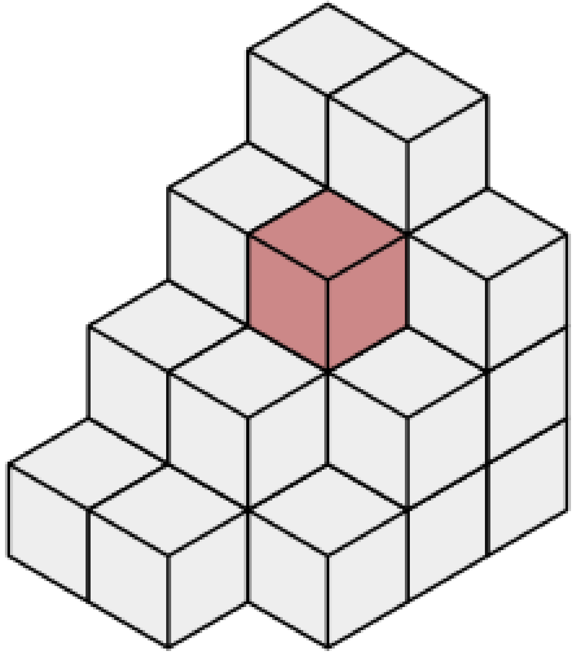
Backup

Previous Approaches

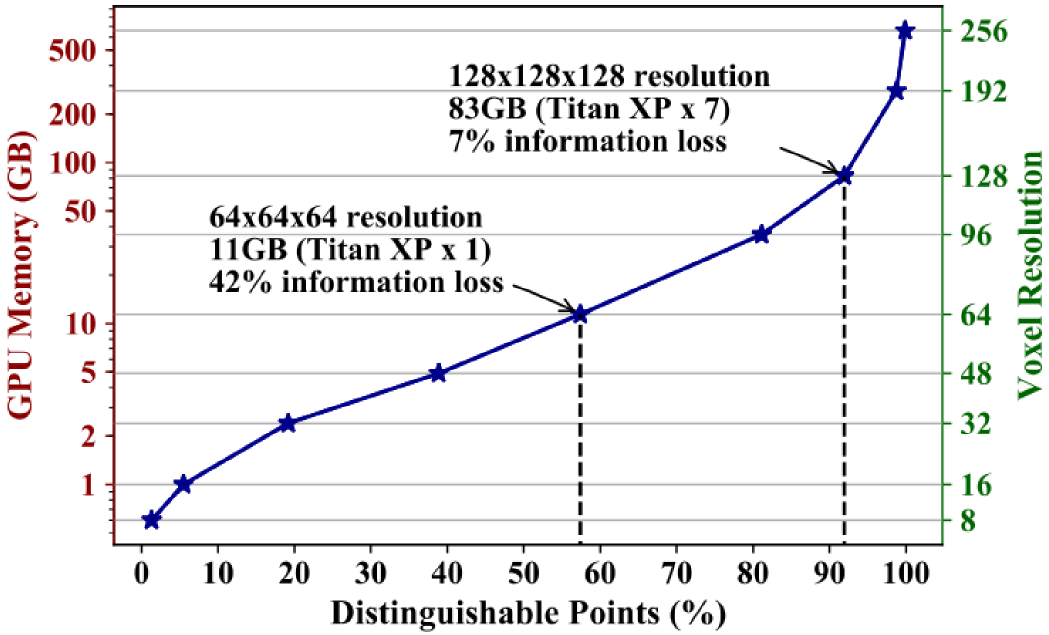
Fall into two categories:

- **Point cloud models**
- **Voxel models**

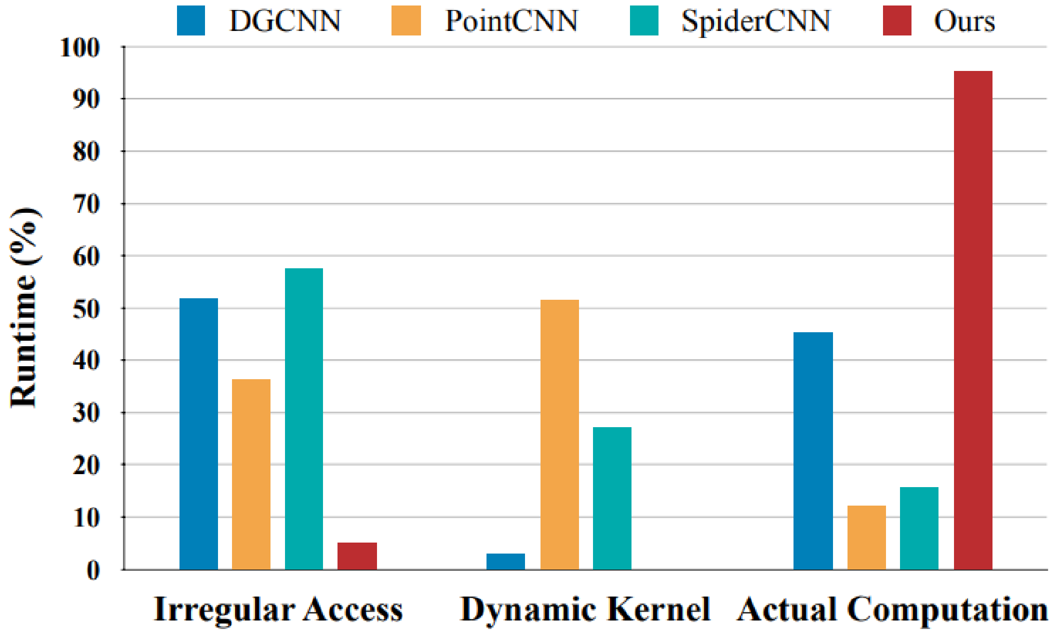




Limitations of Previous Approaches



(a) Voxel-based: memory grows cubically



(b) Point-based: large memory/computation overheads

HGCAL Samples

Zero pileup, double-tau dataset.

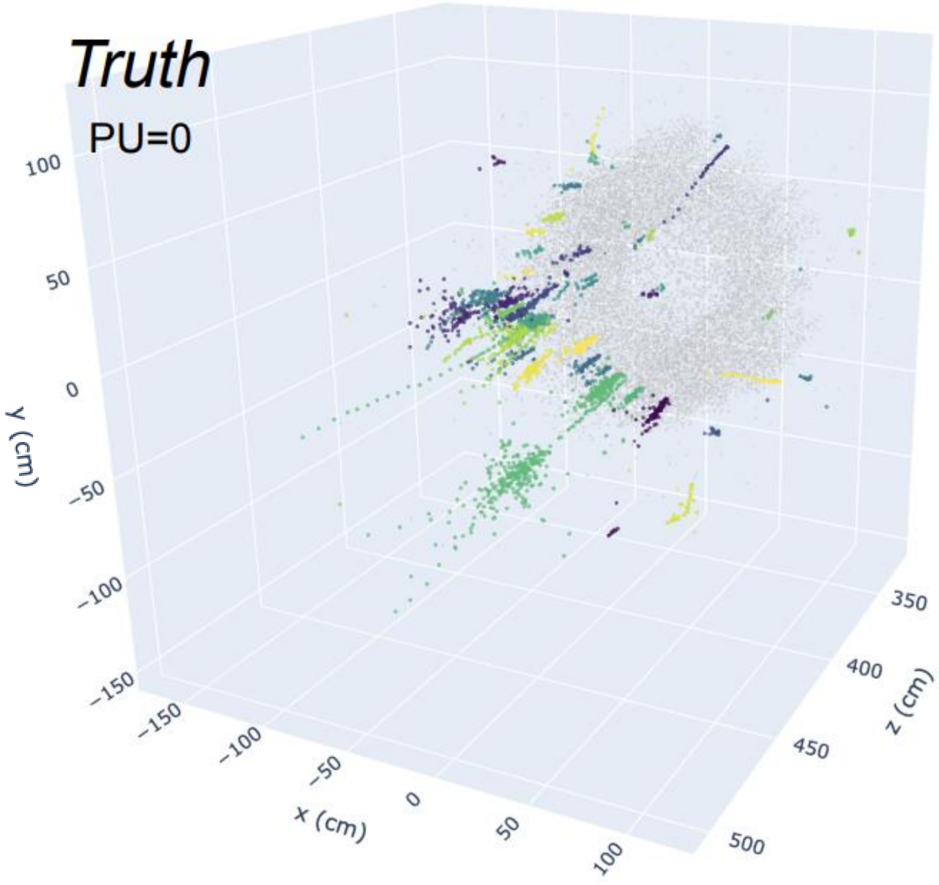
CMS detector simulation with GEANT4.

Simulation-level energy deposits are mapped onto reconstructed energy deposits to form the truth definition.

Inseparable showers (due to overlap) are merged.

Each event has ~20K hits.

See CR2022_033.pdf (cern.ch) for detailed description of samples.



HCAL Samples

- Zero pileup, ttbar dataset.
- CMS detector simulation with GEANT4.
- Simulation-level energy deposits are mapped onto reconstructed energy deposits to form the truth definition.
- The details of truth matching are a bit different than for HGCAL – won't go into it here.