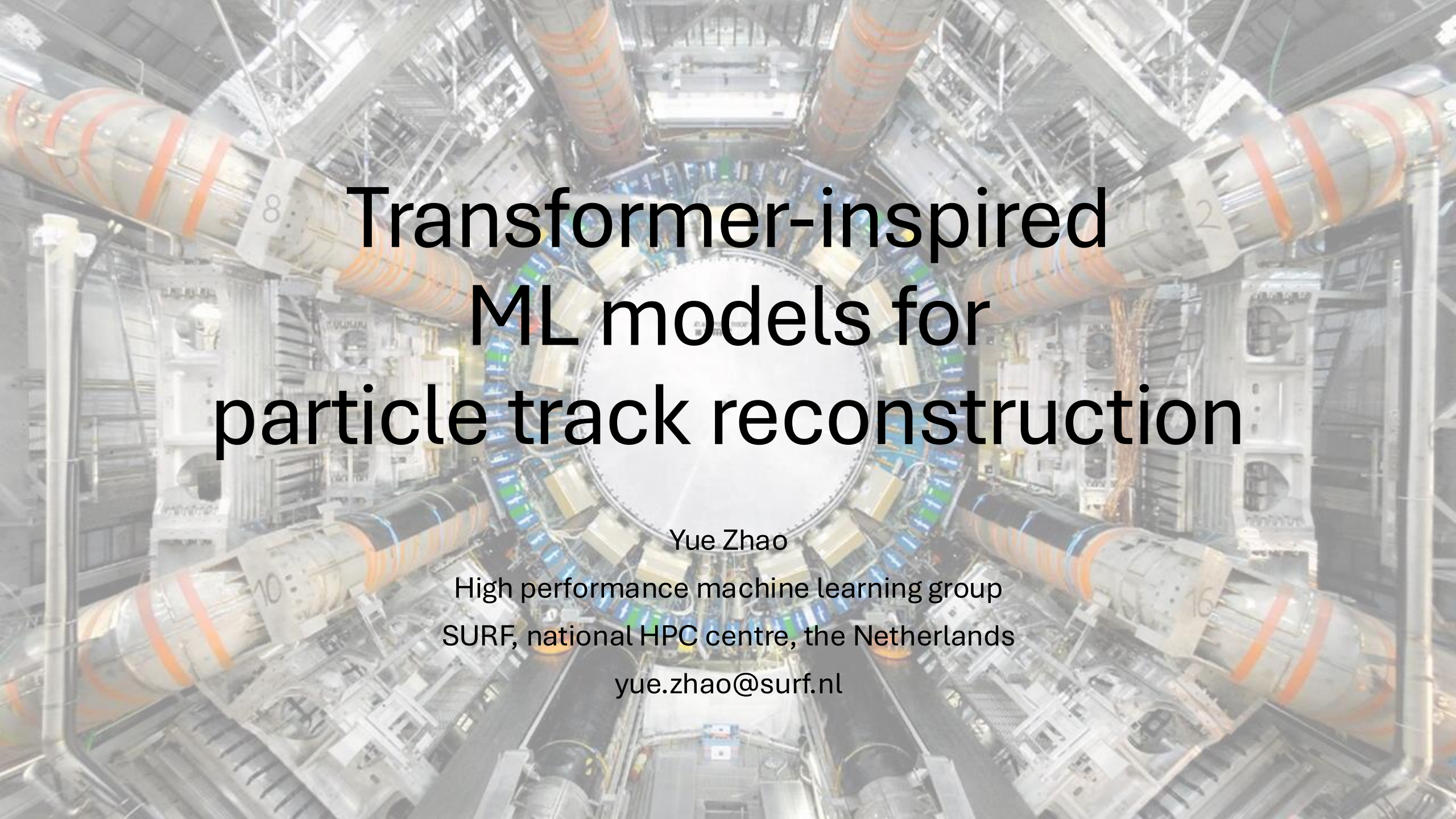


# Transformer-inspired ML models for particle track reconstruction

Yue Zhao

High performance machine learning group

SURF, national HPC centre, the Netherlands

yue.zhao@surf.nl

# The collaboration





- Sascha Caron — High-Energy Physics, Radboud University, The Netherlands; National Institute for Subatomic Physics (Nikhef)
- Nadezhda Dobreva — Radboud University, The Netherlands
- José Martín-Guerrero — Intelligent Data Analysis Laboratory (IDAL), University of Valencia, Spain; ValgrAI, Spain
- Uraz Odyurt — University of Twente, The Netherlands; Nikhef, The Netherlands
- Slav Pshenov — Nikhef, The Netherlands
- Roberto Ruiz de Austri Bazan — Instituto de Física Corpuscular, University of Valencia
- Evgeniy Shalugin — Radboud University
- Zef Wolffs — Institute of Physics, University of Amsterdam; Nikhef, the Netherlands
- Yue Zhao — SURF (national HPC centre), the Netherlands

# The collaboration

Present at the conference!

| | |
|---|---|
| • Sascha Caron | High-Energy Physics, Radboud University, The Netherlands<br>National Institute for Subatomic Physics (Nikhef) |
| • Nadezhda Dobreva | Radboud University, The Netherlands |
| • José Martín-Guerrero | Intelligent Data Analysis Laboratory (IDAL)<br>University of Valencia, Spain<br>ValgrAI, Spain |
| • Uraz Odyurt | University of Twente, The Netherlands<br>Nikhef, The Netherlands |
| • Slav Pshenov | Nikhef, The Netherlands |
| • Roberto Ruiz de Austri Bazan | Instituto de Física Corpuscular, University of Valencia |
| • Evgeniy Shalugin | Radboud University |
| • Zef Wolffs | Institute of Physics, University of Amsterdam<br>Nikhef, the Netherlands |
| • Yue Zhao | SURF (national HPC centre), the Netherlands |

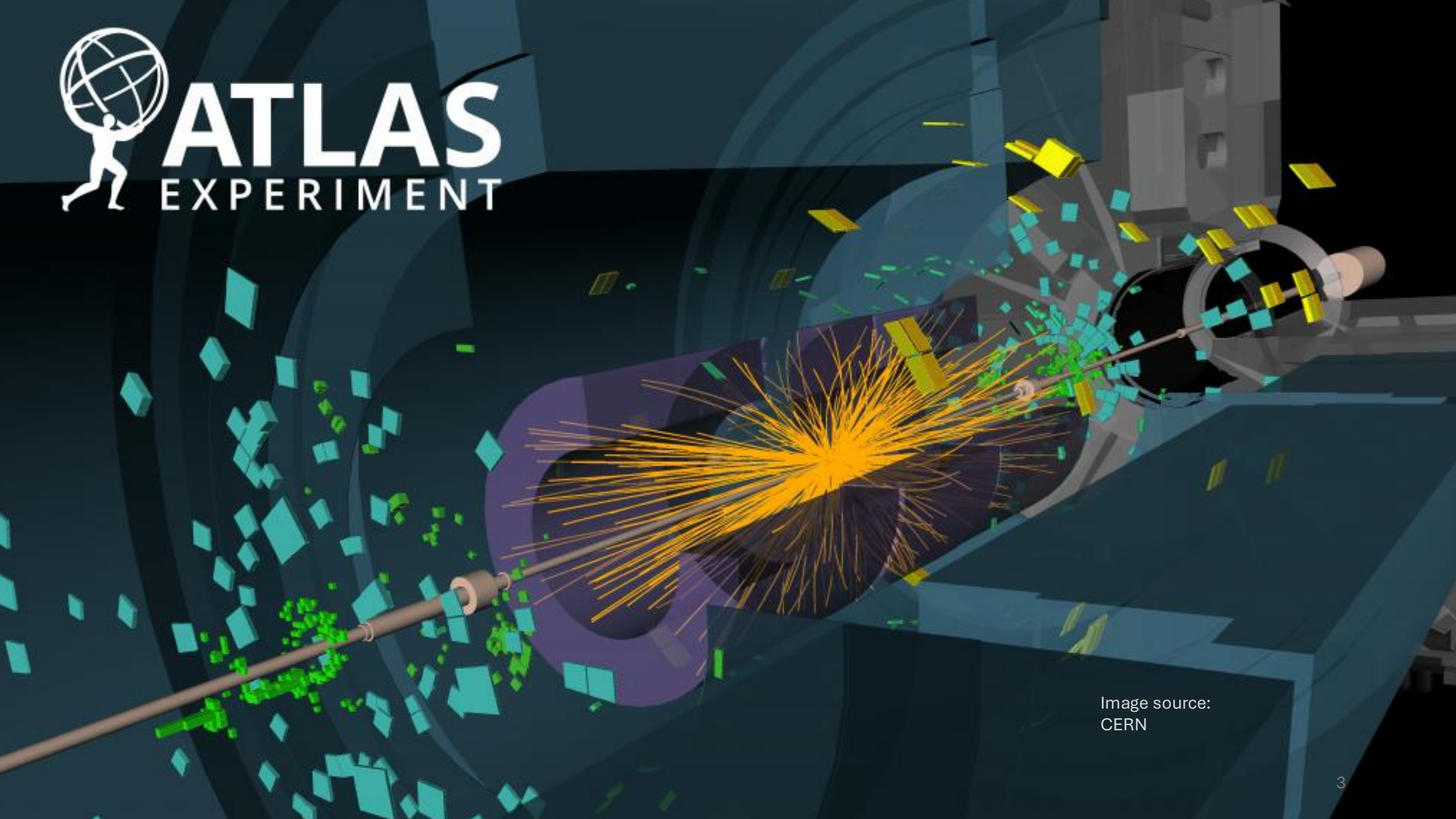



Image source:
CERN

# The data: 3D point cloud of hits

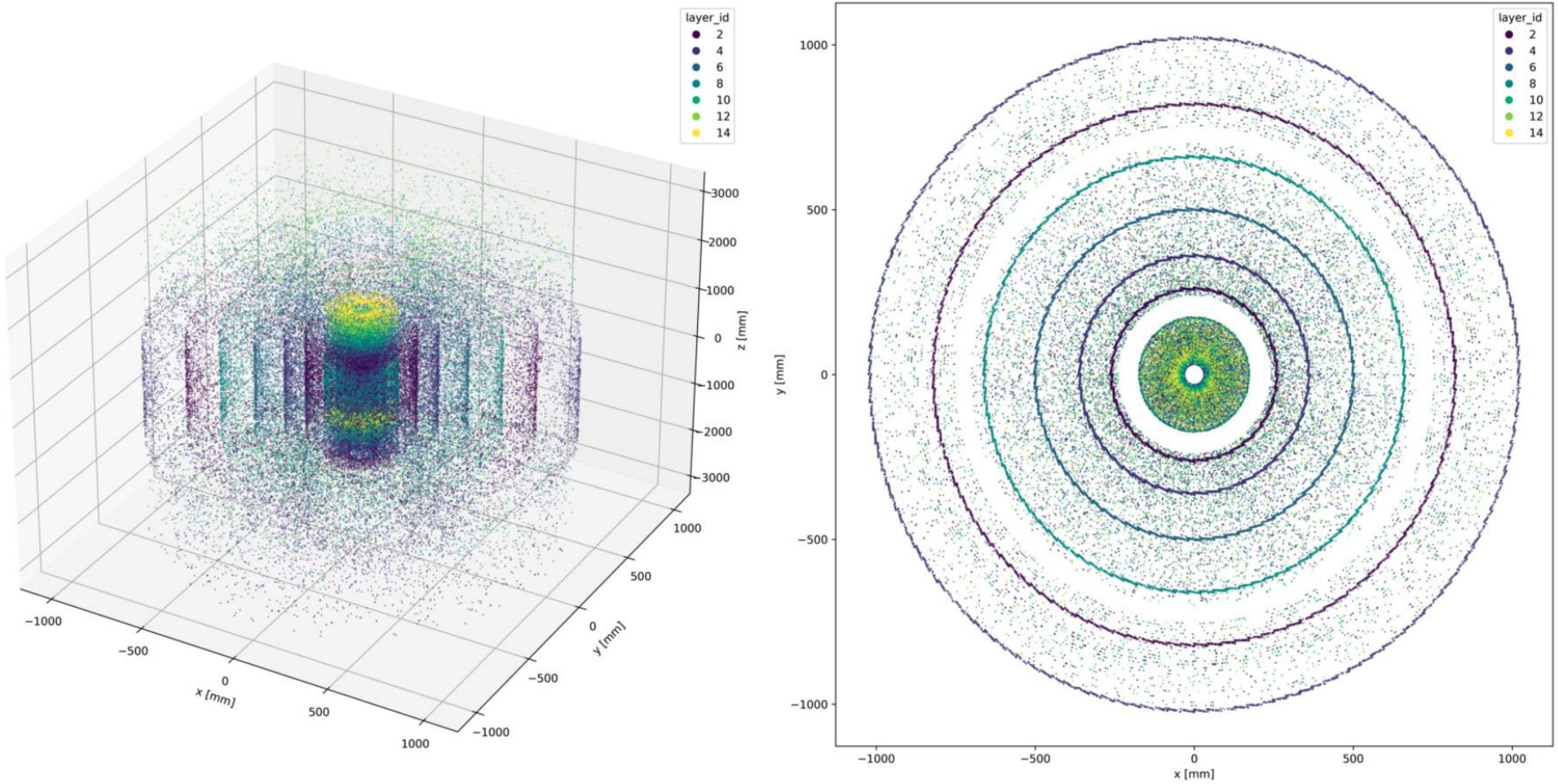



Image source: Schoonheid, L. (2023). Fit-assisted DNN model for HEP Particle Tracking. [BSc thesis]. UvA, VU Amsterdam.

# The task:

Reconstruct tracks from hits

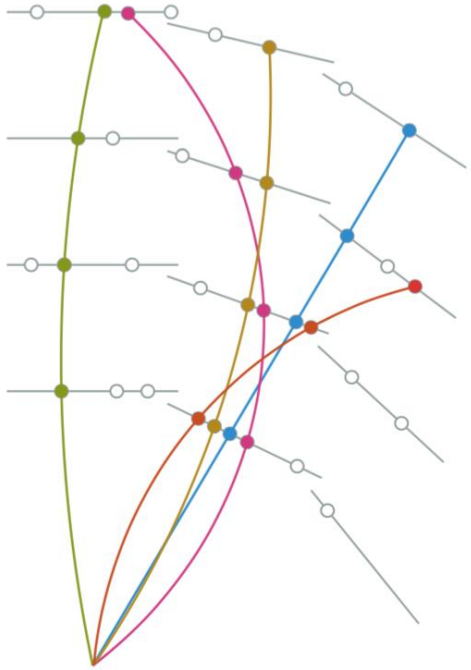
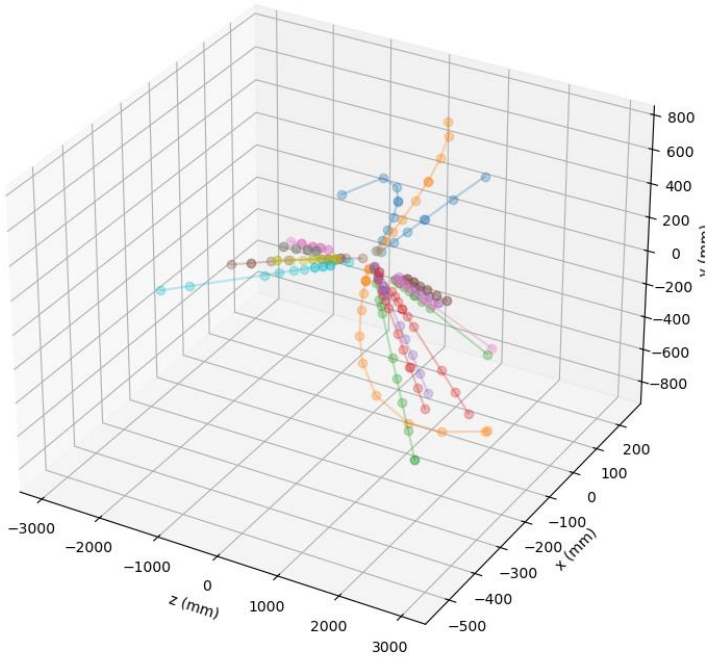


Image source: arXiv:1904.06778 [hep-ex]

# The task:

Reconstruct tracks from hits




~300 hits per event

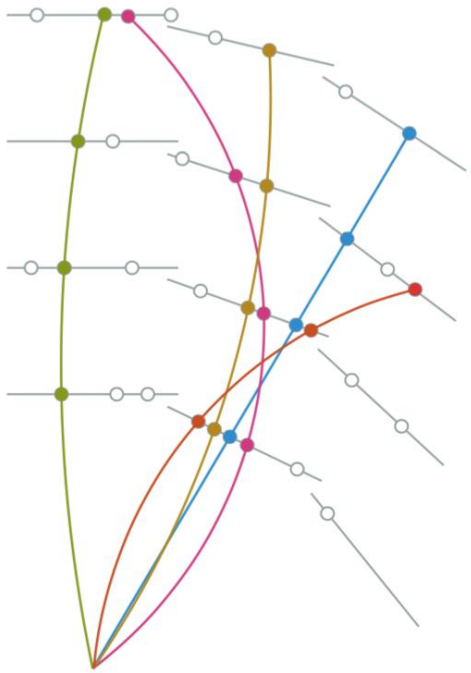
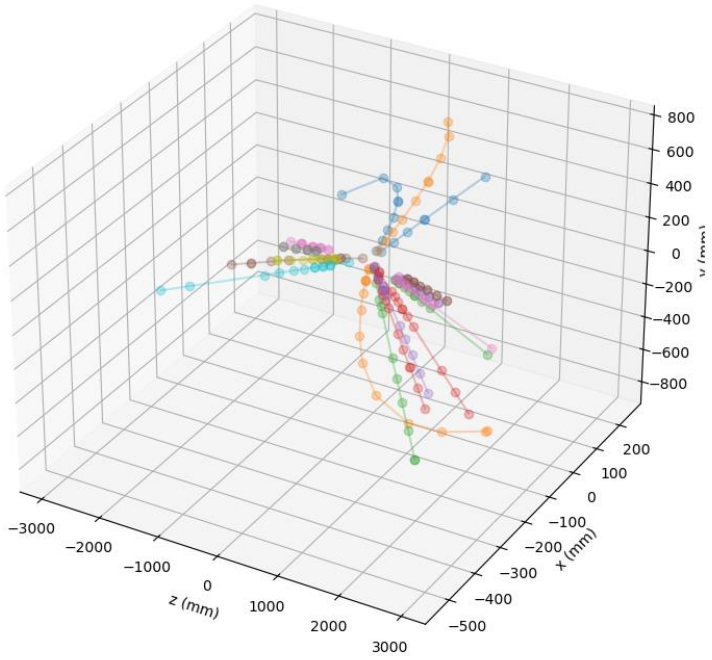


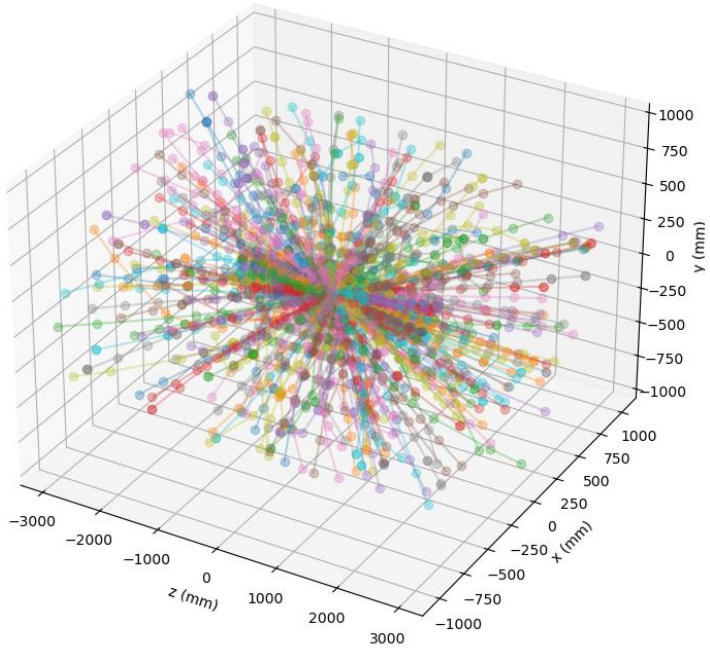
Image source: arXiv:1904.06778 [hep-ex]

# The task:

## Reconstruct tracks from hits




~300 hits per event




~3500 hits per event

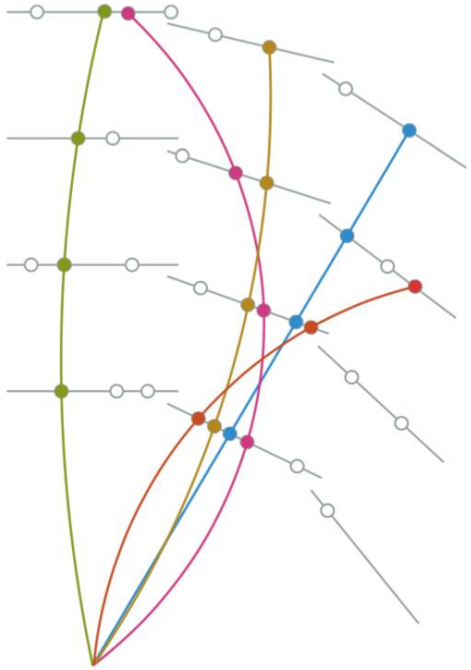


Image source: arXiv:1904.06778 [hep-ex]

# The data challenge

- 40 million collision events per second at LHC
- Hundreds of terabytes of data per second (online)
- Tens of petabytes saved per year (offline)

- High Luminosity LHC conditions with $<\mu>$ ~ 200 creates a substantial CPU time challenge for current reconstruction techniques due to combinatorics

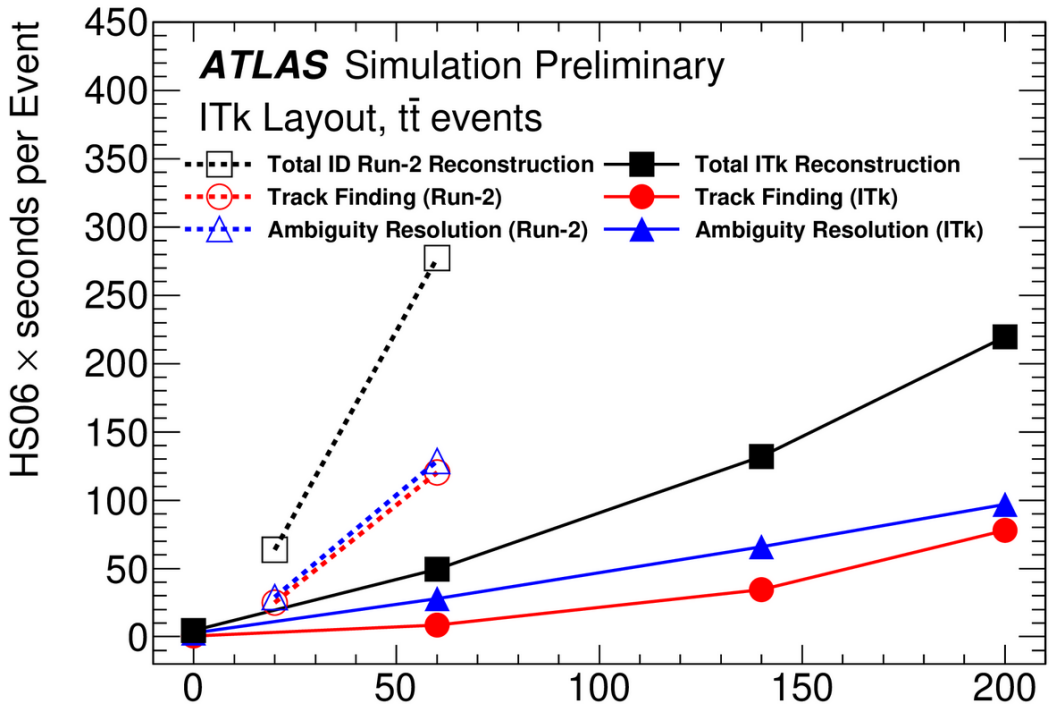
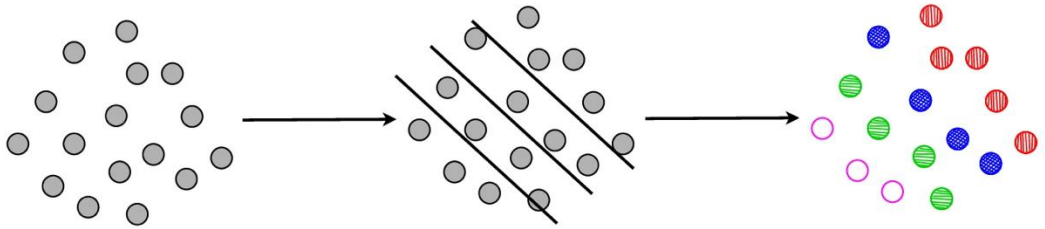



Image source: Fast tracking for the HL-LHC ATLAS detector

# 1st phase of our project

Caron, S. *et al*. Trackformers: in search of transformer-based particle tracking for the high-luminosity LHC era. *Eur. Phys. J. C* **85**, 460 (2025). https://doi.org/10.1140/epjc/s10052-025-14156-3

EncCla



Encoder-only

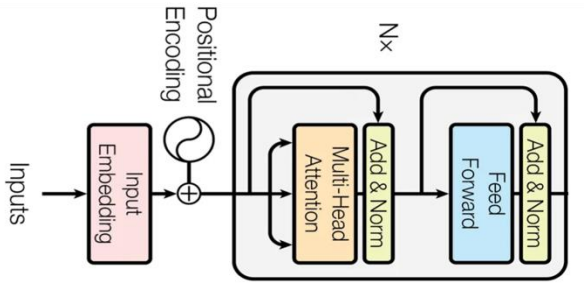
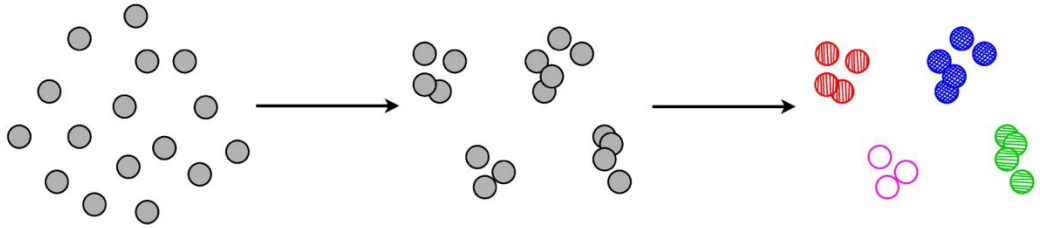


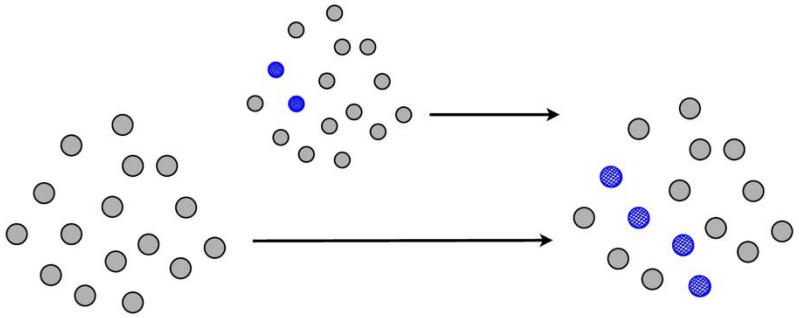

Image source: arXiv:1706.03762

EncReg

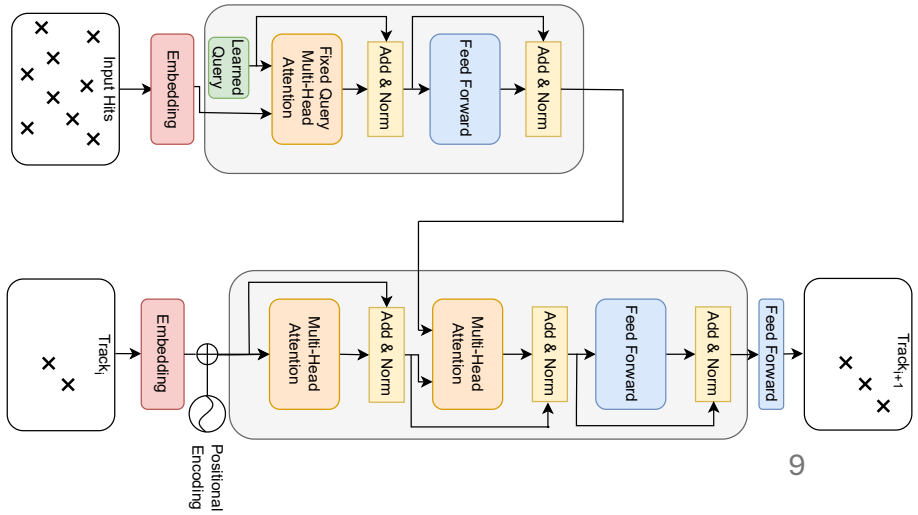


Encoder-only

EncDec



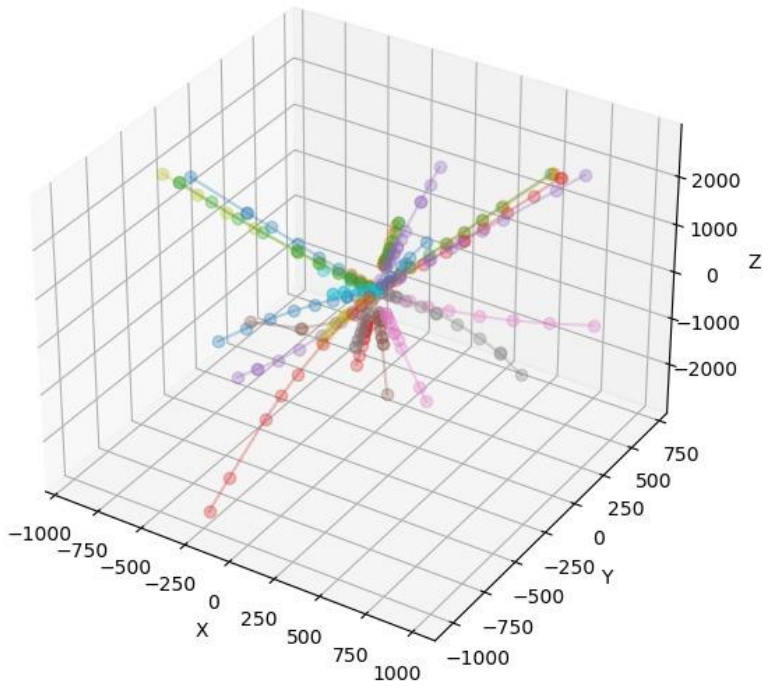
Encoder-Decoder



# Results of 1st phase

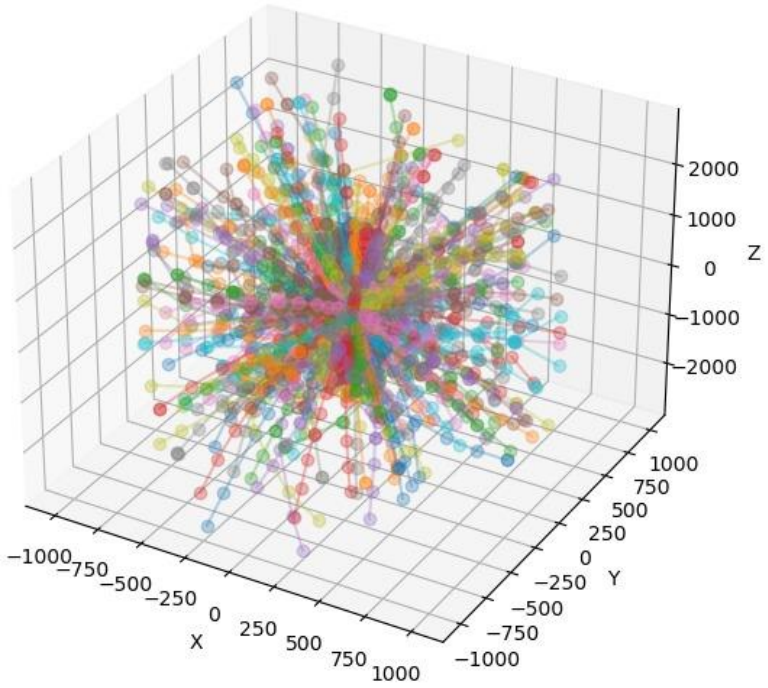
Kaggle score: double majority
- >50% hits from one particle
- <50% of its hits outside the track

10-50 tracks (up to 500 hits) per event



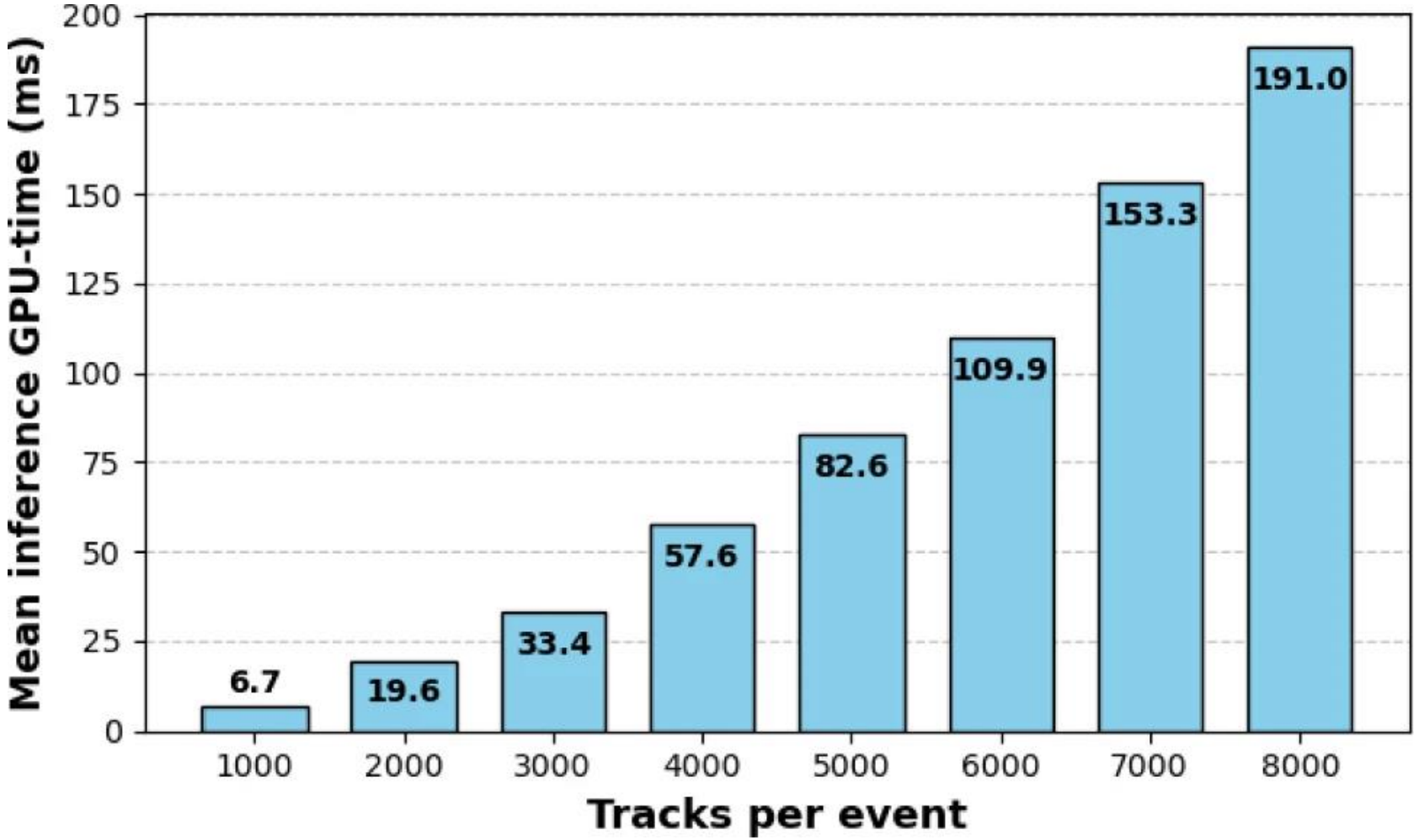

Kaggle score: 94%
Inference speed: ~2 miliseconds

200-500 tracks (up to 5k hits) per event



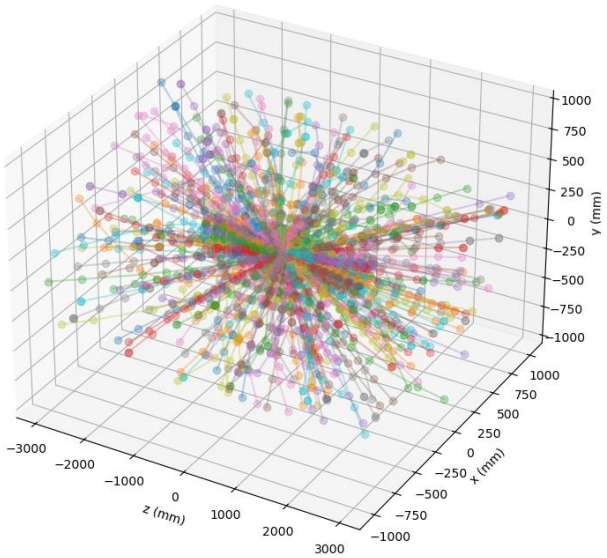

Kaggle score: 60-70%
Inference speed: ~50 miliseconds

# Results of 1st phase: inference speed of EncReg

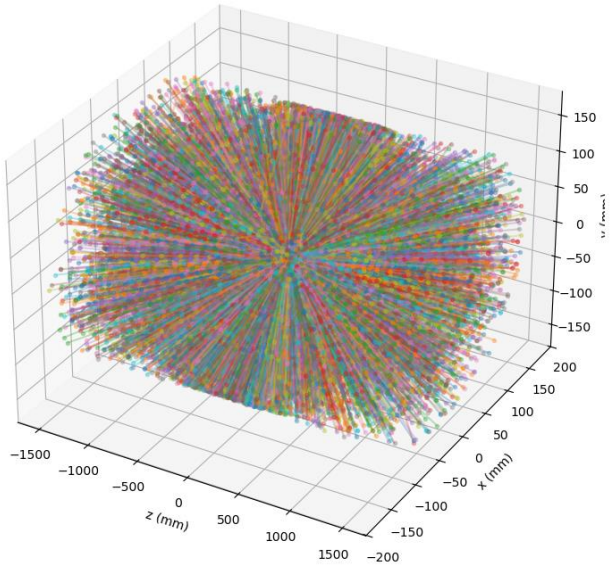


# Current efforts in our project (2nd phase)

- To address memory issue
  - FlexAttention and masking




~3500 hits per event




~60.000 hits per event

$attention\ matrix \sim (\#\ hits)^2$

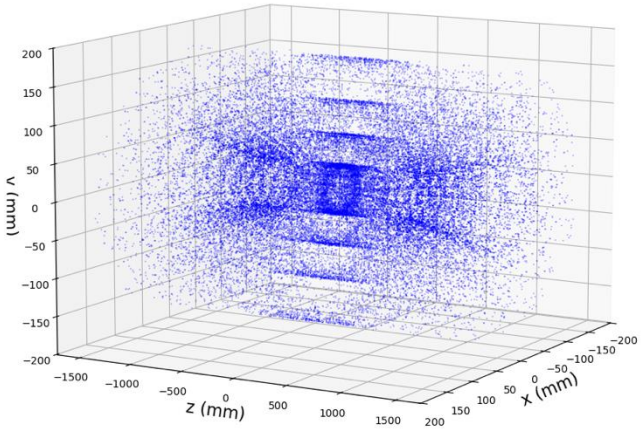
| | Hit 1 | Hit 2 | … | Hit 60k |
|---|---|---|---|---|
| Hit 1 | | | | |
| Hit 2 | | | | |
| … | | | | |
| Hit 60k | | | | |

# Current efforts in our project (2nd phase)

- To address memory issue
  - FlexAttention and masking
  - Domain decomposition or convolutions over $\phi$?
- To improve accuracy
  - Projections of hits to surfaces → Clustering
  - Clustering in latent space of EncCla/EncReg model

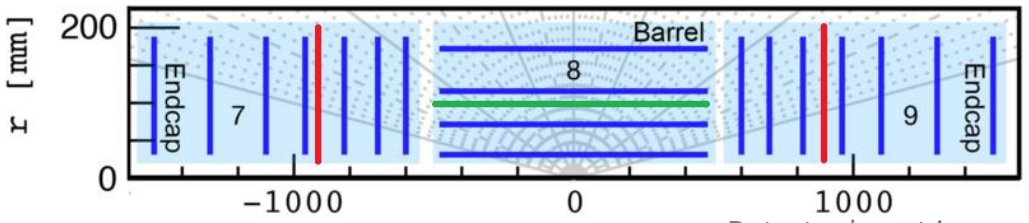
# Projections of hits to surfaces → Clustering

- Tracks of interest with $p_t > 0.9\ GeV$ are local in space
- Project pixel detector hits onto 3 surfaces (straight line to the origin)
- Choose surfaces to minimize hits spread

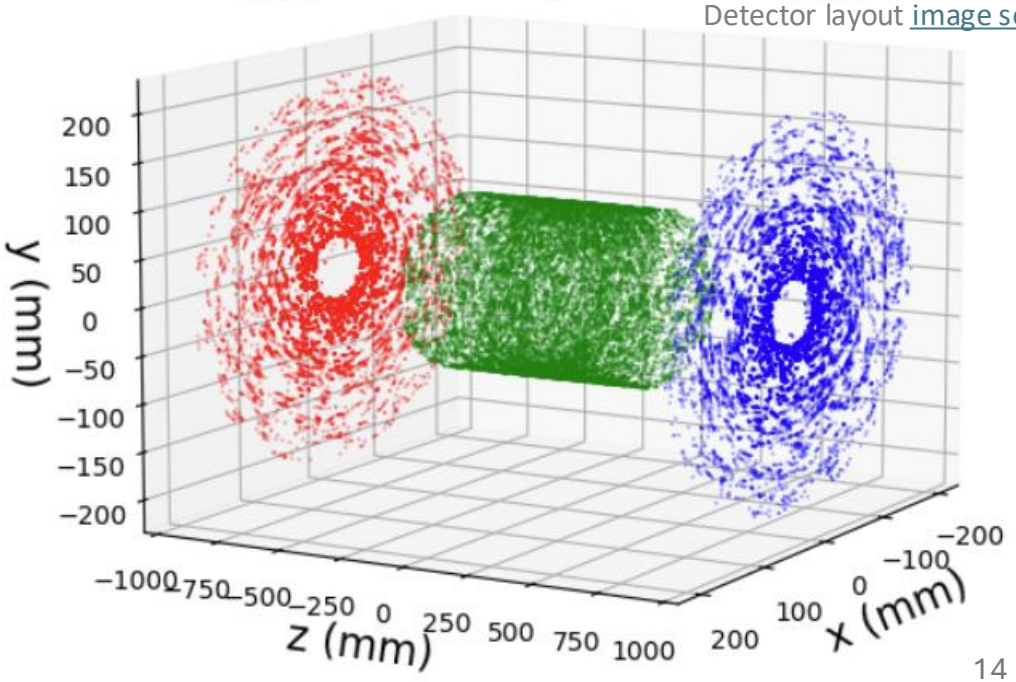



Pixel detector layout

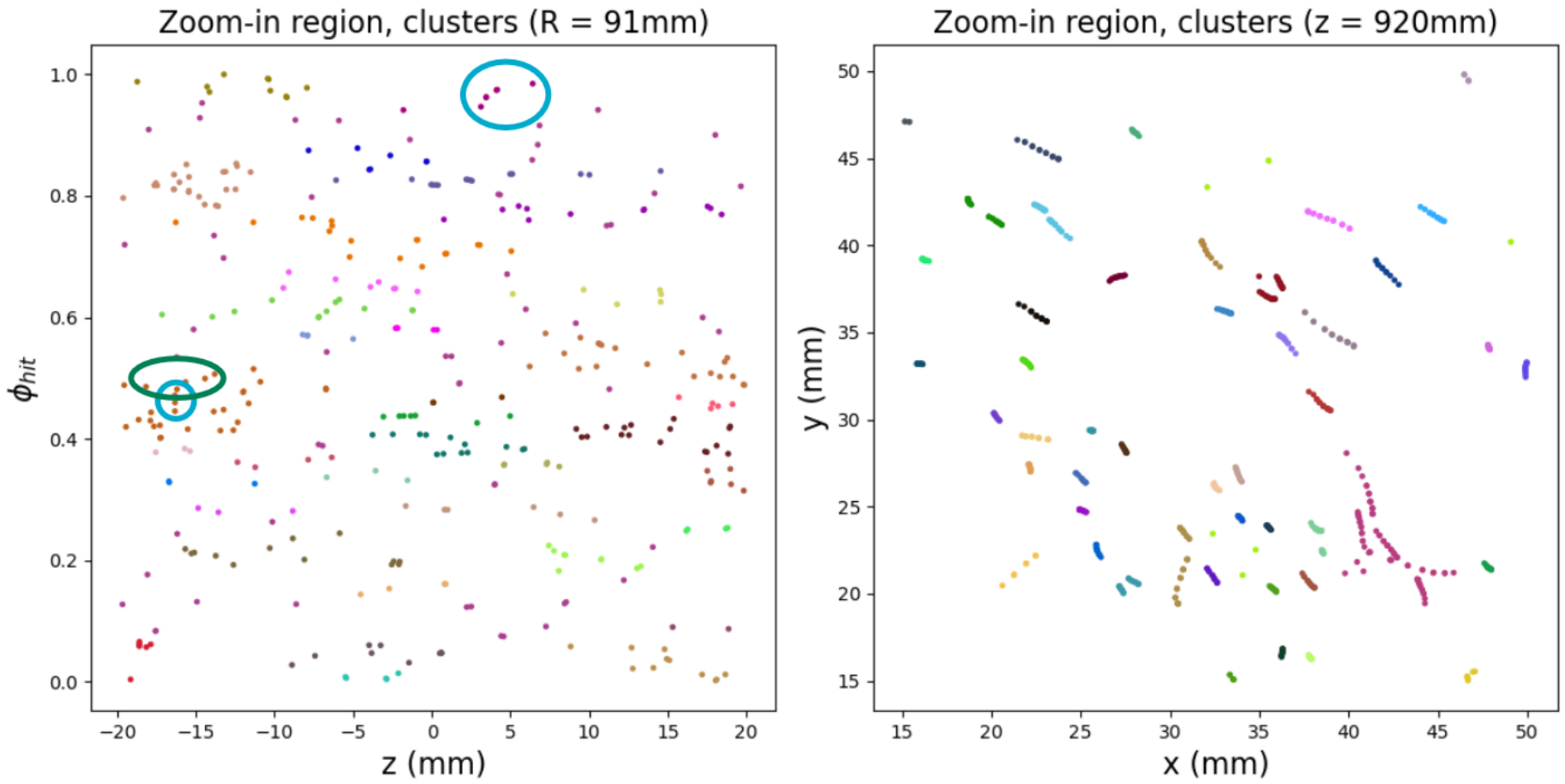
Barrel cylinder proj. surface
Endcap projection surface
Detector layers




Detector layout image source



# Projections of hits to surfaces → Clustering



# Projections of hits to surfaces → Clustering

**BARREL**

| | Recon. Ratio | Track DM eff. | # clusters | # tracks | Attn. matrix reduction |
|---|---|---|---|---|---|
| **DBSCAN** | 0.987 | 0.14 | 1119 | 779 | 0.009138 |
| **Iterative window** | 0.955 | 0.004 | 499 | 779 | 0.000285 |
| **Density Peaks** | 0.982 | 0.024 | 684 | 779 | 0.002237 |
| **Grid window** | 0.908 | 0.01 | 1176 | 779 | 0.000998 |

**ENDCAP**

| | Recon. Ratio | Track DM eff. | # clusters | # tracks | Attn. matrix reduction |
|---|---|---|---|---|---|
| **DBSCAN** | 0.985 | 0.832 | 756 | 267 | 0.00052 |
| **Iterative window** | 0.993 | 0.468 | 871 | 267 | 0.002063 |
| **Density Peaks** | 0.993 | 0.831 | 1684 | 267 | 0.000653 |
| **Grid Window** | 0.989 | 0.577 | 1330 | 267 | 0.001099 |

# Current efforts in our project (2nd phase)

- To address memory issue
  - FlexAttention and masking
  - Domain decomposition or convolutions over ϕ?
- To improve accuracy
  - Projections of hits to surfaces → Clustering
  - Clustering in latent space of EncCla/EncReg model
- Simulations for more training data: 40k events – collaborate?

# Current efforts in our project (2nd phase)

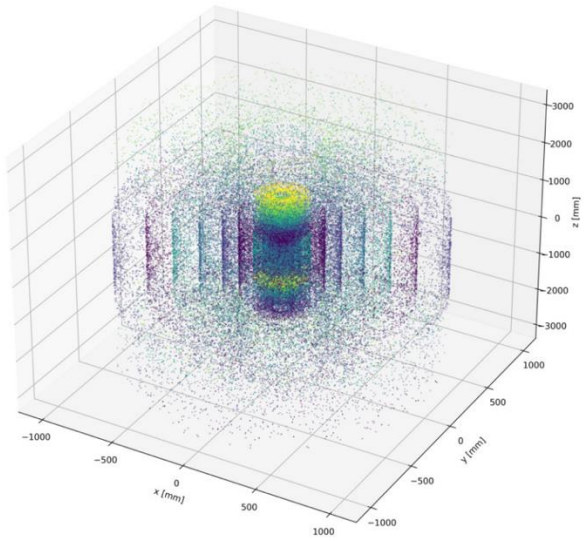
- To address memory issue
  - FlexAttention and masking
  - Domain decomposition or convolutions over $\phi$?
- To improve accuracy
  - Projections of hits to surfaces → Clustering
  - Clustering in latent space of EncCla/EncReg model
- Simulations for more training data: 40k events – collaborate?
- Tokenization

# From Hits to Tracks: Tokens in Physics

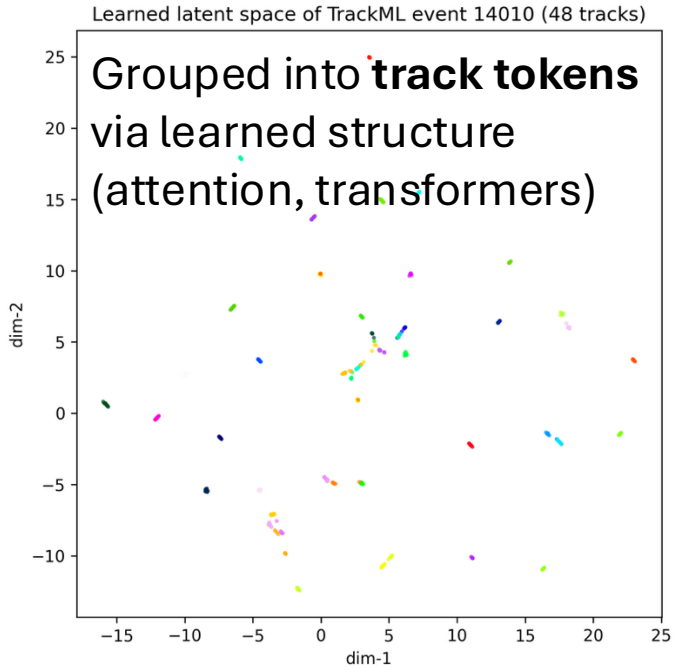
10^5

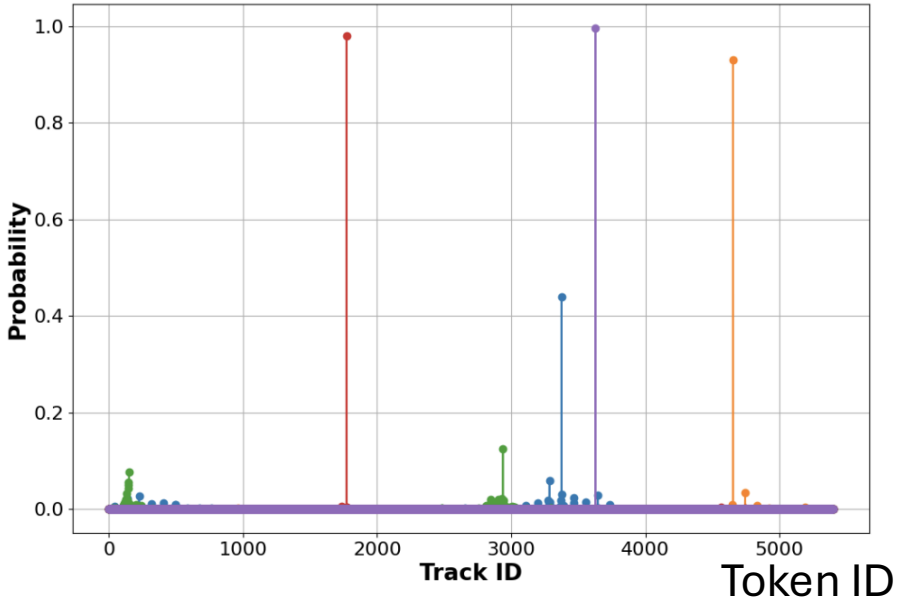
Hits → Encoder → Latent Vector (representation) → Decoder → Track Token

10^3



O(10000) hits (no pre-processing)




SoftMax outputs of the EncCla model for the first five hits for one event for all track classes

- ”Trackformers: ”, *Eur.Phys.J.C* 85 (2025) 4, 460 , e-Print: 2407.07179 [hep-ex]

# From Hits to Tracks: Tokens in Physics

- Goal is to convert a 4-vector to integer
- VQ-VAE replaces the continuous latent sampling of a standard VAE with a discrete codebook, trained end-to-end

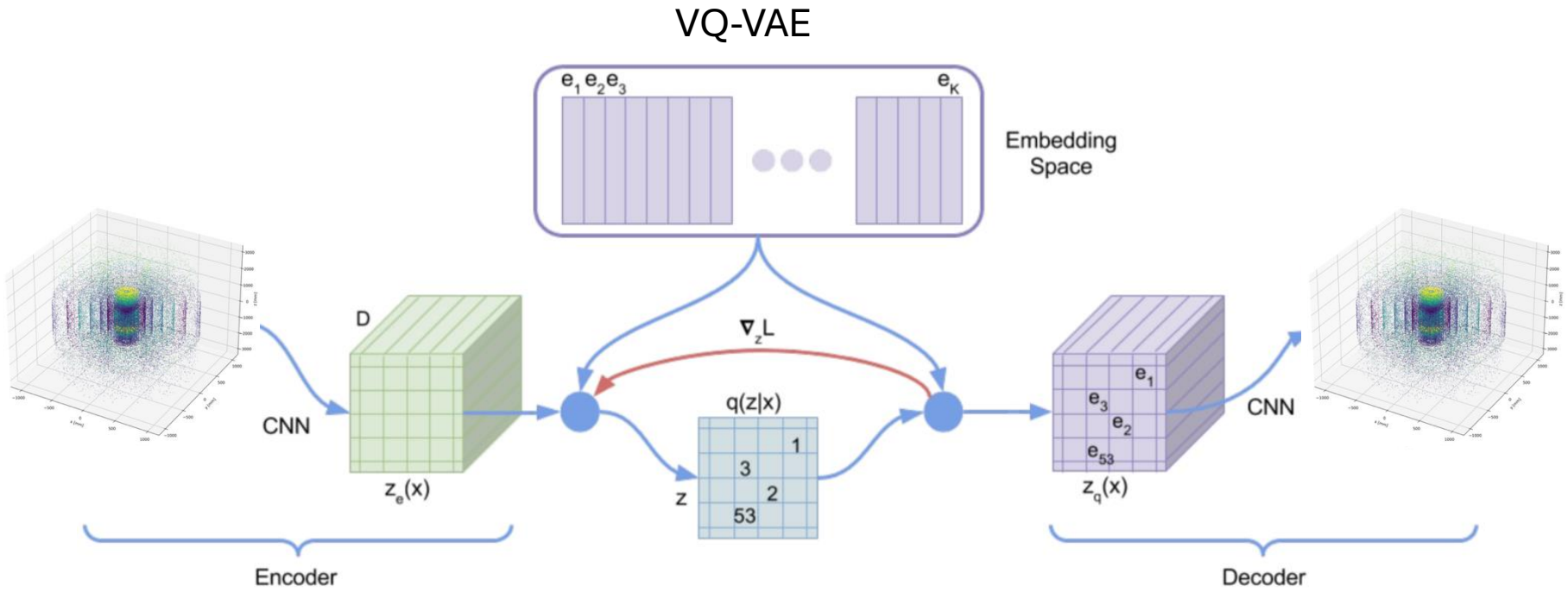



Image modified from : arXiv:1711.00937

# Key messages

- Transformer-based architectures are a promising direction for tracking
- FlexAttention and masking help address quadratic scaling of attention matrix
- Projecting the data to ‘barrels’ and ‘endcaps’ improves accuracy
- New TrackML-like simulations provide more training data
- If we see tracks as tokens, can we use VQ-VAE to tackle tracking?

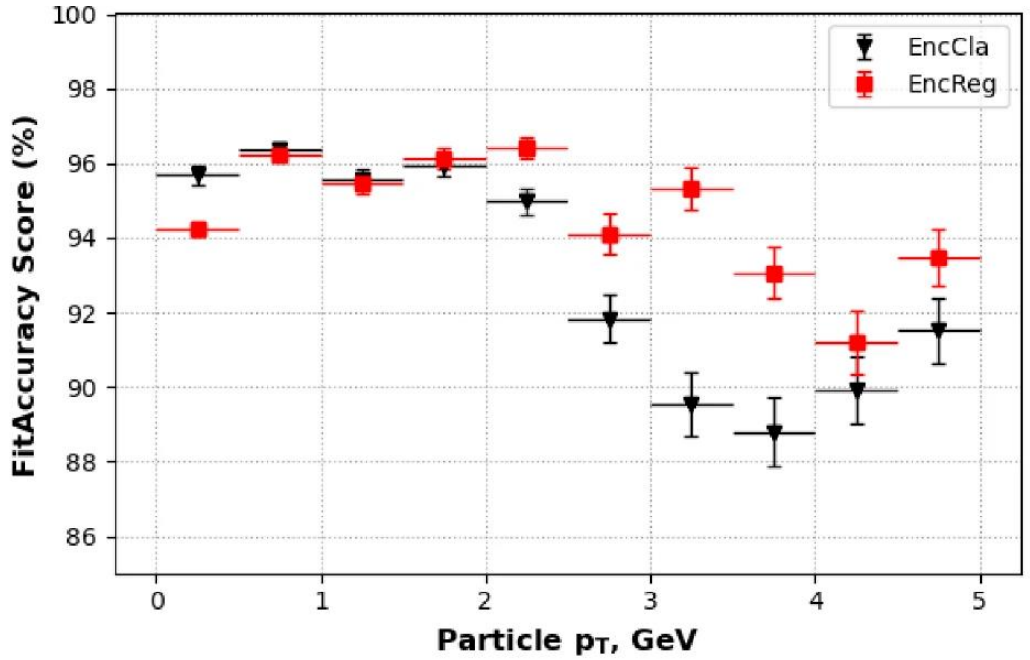
# The collaboration

Present at the conference!

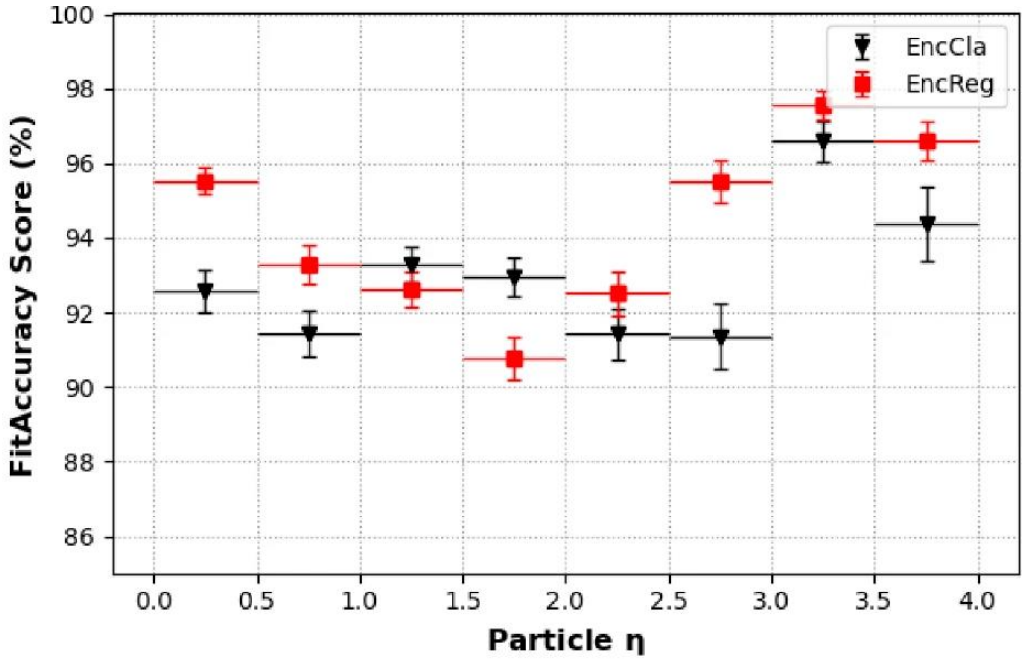
| | |
|---|---|
| • Sascha Caron | High-Energy Physics, Radboud University, The Netherlands<br>National Institute for Subatomic Physics (Nikhef) |
| • Nadezhda Dobreva | Radboud University, The Netherlands |
| • José Martín-Guerrero | Intelligent Data Analysis Laboratory (IDAL)<br>University of Valencia, Spain<br>ValgrAI, Spain |
| • Uraz Odyurt | University of Twente, The Netherlands<br>Nikhef, The Netherlands |
| • Slav Pshenov | Nikhef, The Netherlands |
| • Roberto Ruiz de Austri Bazan | Instituto de Física Corpuscular, University of Valencia |
| • Evgeniy Shalugin | Radboud University |
| • Zef Wolffs | Institute of Physics, University of Amsterdam<br>Nikhef, the Netherlands |
| • Yue Zhao | SURF (national HPC centre), the Netherlands |

# Additional info
# Results of 1st phase: 10-50 tracks

Kaggle score: double majority
- >50% hits from one particle
- <50% of its hits outside the track



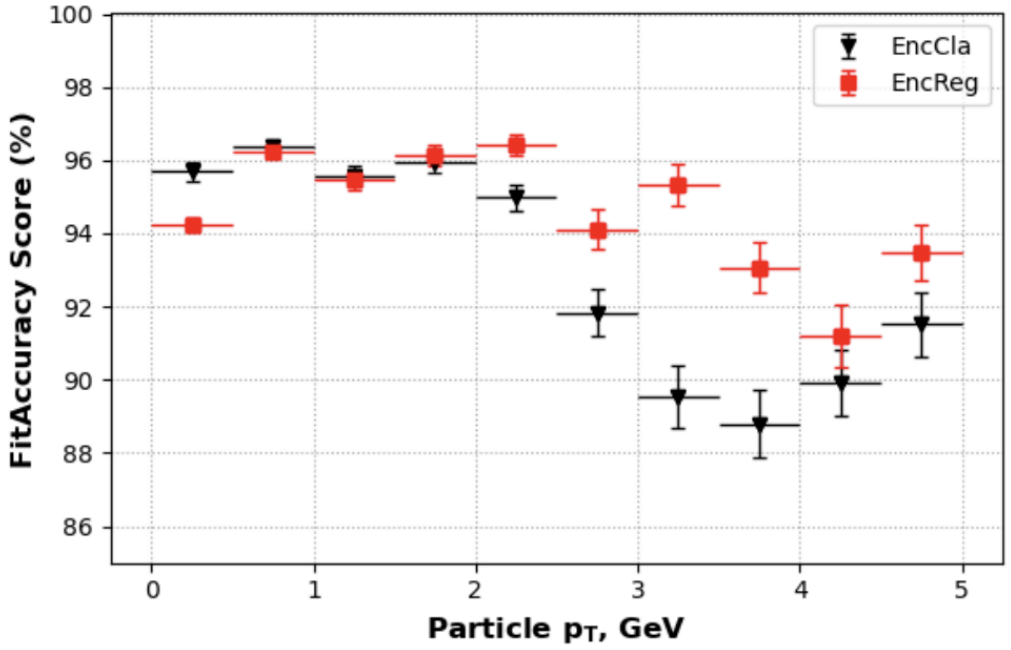

(a) FitAccuracy versus the transverse momentum $p_T$.



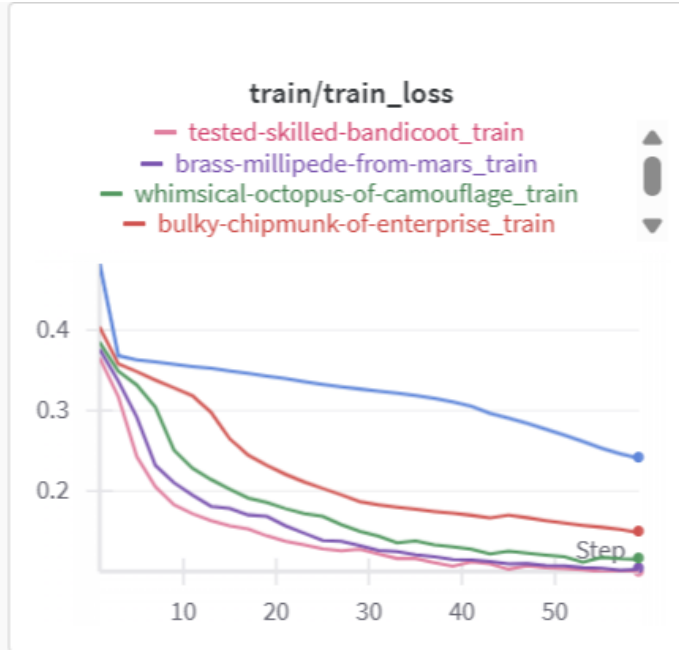

(b) FitAccuracy versus pseudorapidity $\eta$.

FitAccuracy = Kaggle score without weights for hits

# Model performance → Not enough training data in trackML challenge ! (model can still improve)








(a) FitAccuracy versus the transverse momentum $p_T$.