SMASH
machine learning for science and humanities postdoctoral program

Jožef Stefan Institute

GPT-LIKE TRANSFORMER MODEL FOR SILICON TRACKING DETECTOR SIMULATION

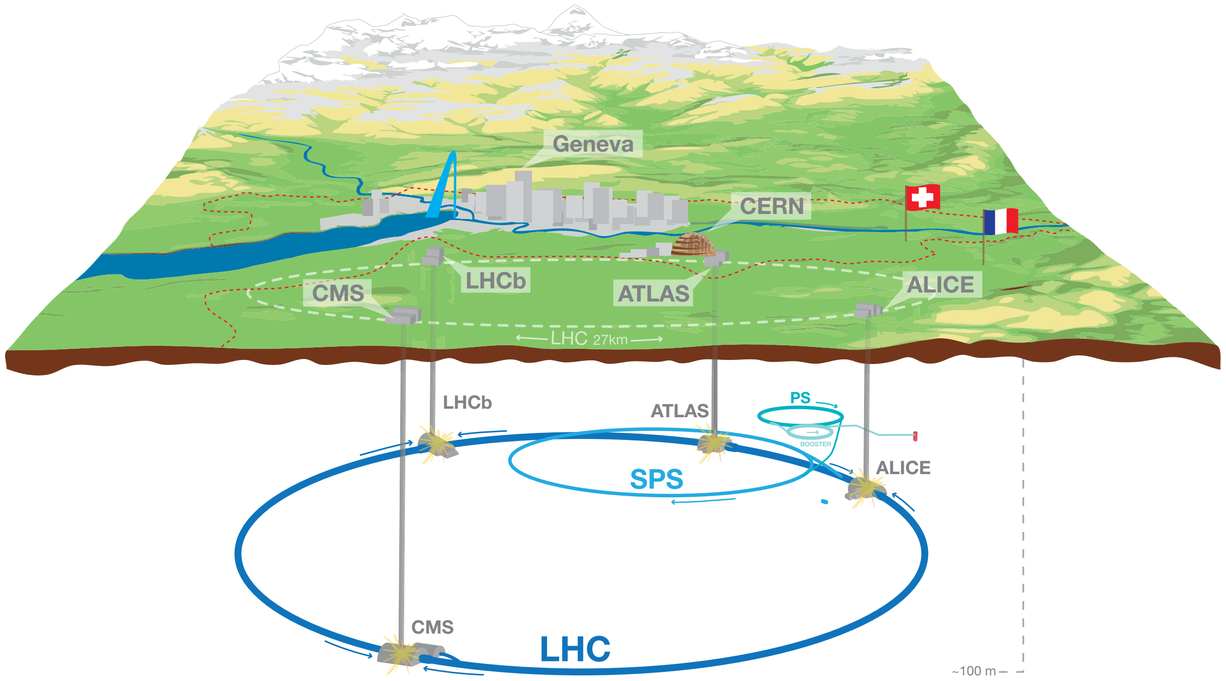
F9 Seminar
January 29, 2026

I FEEL SLOVENIA

Co-funded by the European Union

Tadej Novak
Jožef Stefan Institute

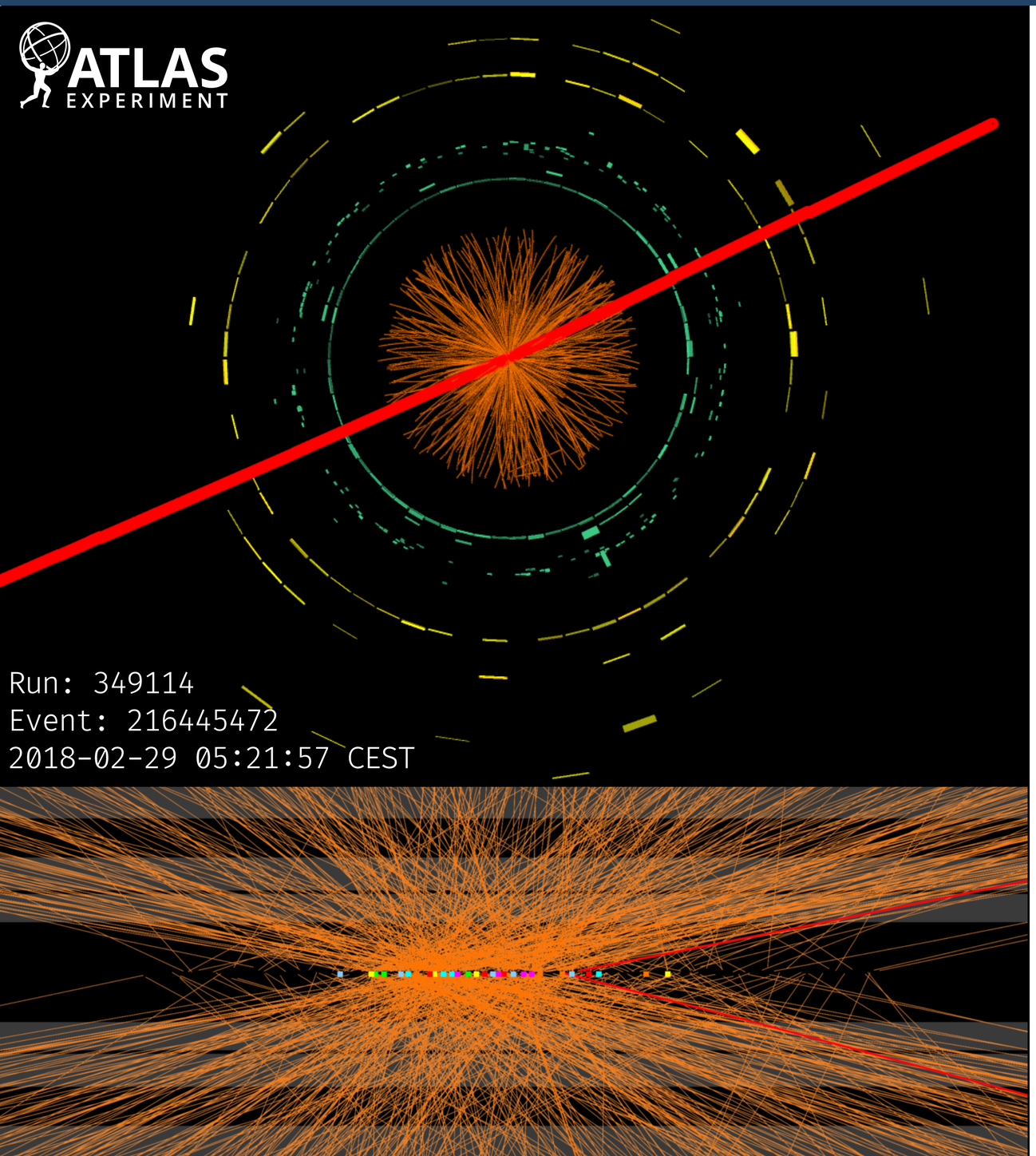
Large Hadron Collider & ATLAS Experiment



Source: CERN

- Largest particle collider — circumference of 27 km:
  - up to 40 million proton-proton collisions per second
- HL-LHC upgrade targeting 2030.
  - data rate 7-10 times greater
  - average number of collisions per bunch crossing rising to as much as 200, from 30-60 currently

Large Hadron Collider & ATLAS Experiment



- Largest particle collider — circumference of 27 km:
  - up to 40 million proton-proton collisions per second
- HL-LHC upgrade targeting 2030.
  - data rate 7-10 times greater
  - average number of collisions per bunch crossing rising to as much as 200, from 30-60 currently
- ATLAS detector a general purpose experiment.
  - Need to measure particle momentum and energy.

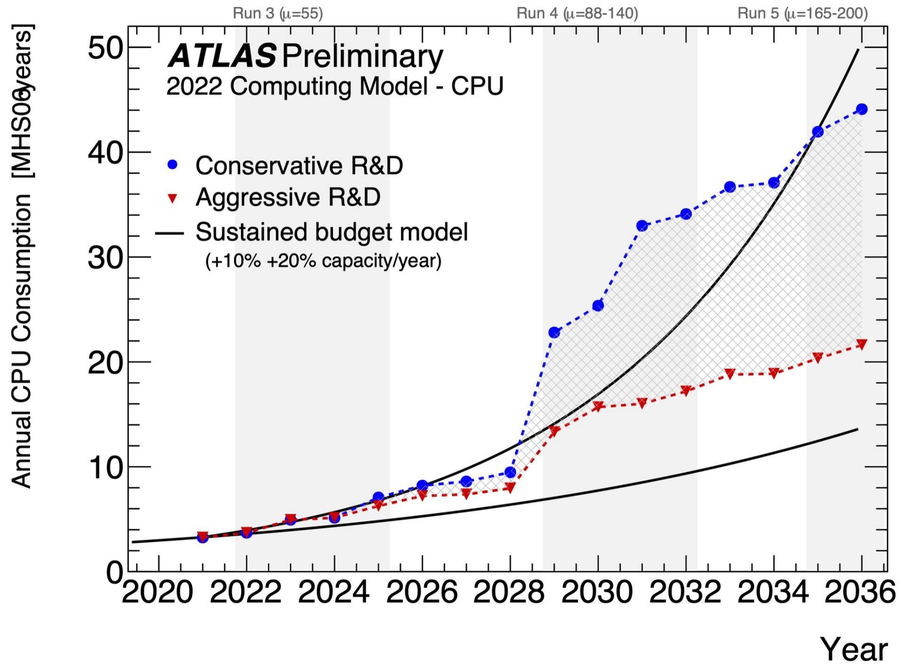
The Need for Monte Carlo Simulation

- A large part of the LHC physics programme relies on accurate Monte Carlo simulation of collision events and their interactions with the detectors.
  - every single particle needs to be simulated
  - detailed (full) detector response simulation using the Geant4 toolkit the most intensive
- Producing simulated samples → majority of experiments' CPU requirements
  - CMS used 85% CPU for Monte Carlo production during 2009-2016
  - half spent detector simulation

The Need for Monte Carlo Simulation

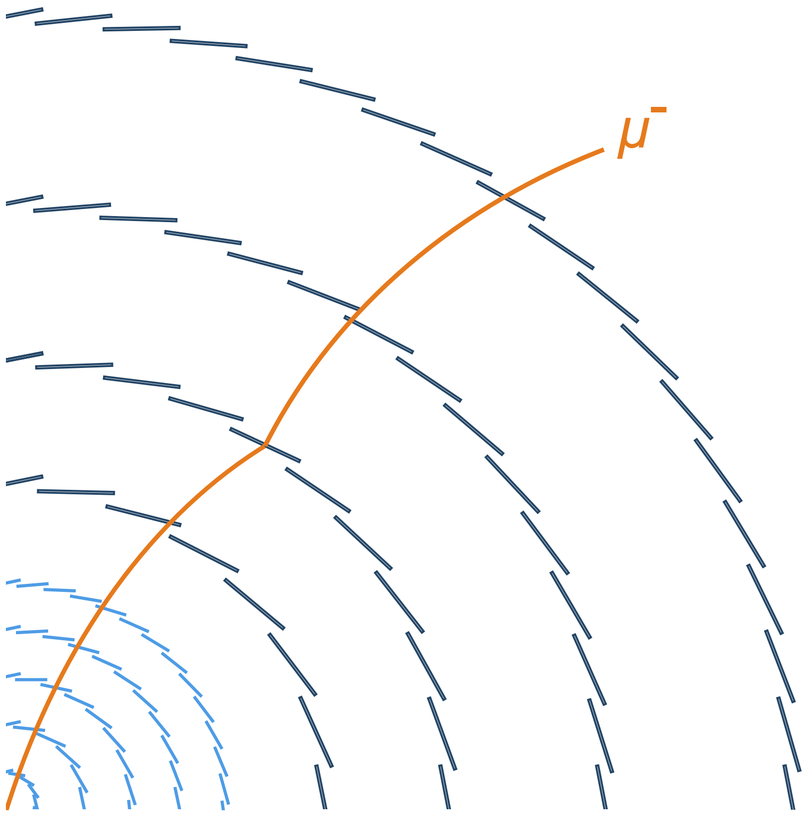
- A large part of the LHC physics programme relies on accurate Monte Carlo simulation of collision events and their interactions with the detectors.
  - every single particle needs to be simulated
    - detailed (full) detector response simulation using the Geant4 toolkit the most intensive
- Producing simulated samples → majority of experiments' CPU requirements
  - CMS used 85% CPU for Monte Carlo production during 2009-2016
  - half spent detector simulation

Source: ATLAS Software and Computing HL-LHC Roadmap



- Current methods do not scale with HL-LHC data rates and more aggressive R&D is needed.

Simulation Using Machine Learning



- Machine learning successfully applied to (fast) calorimeter simulation.
  - Calorimeters measure particle energies.
  - Data can be described by 2D images — many AI models inspired by industry.
  - Can achieve order of magnitude speed-up with physics performance sufficient for a large fraction of analyses.
- The first attempt of using ML for silicon tracking detectors simulation.
  - Data more sparse and sequential.
  - Using transformers.

Transformers

- Transformers commonly used with sequential data (most often LLMs).
- Using decoder-only architecture.
  - Input/output data are the same.
  - Target to predict the next element of the sequence.
    - The well known example are the GPT family of models.
- Specialised on discrete sequences which are tokenised (sequential integers).

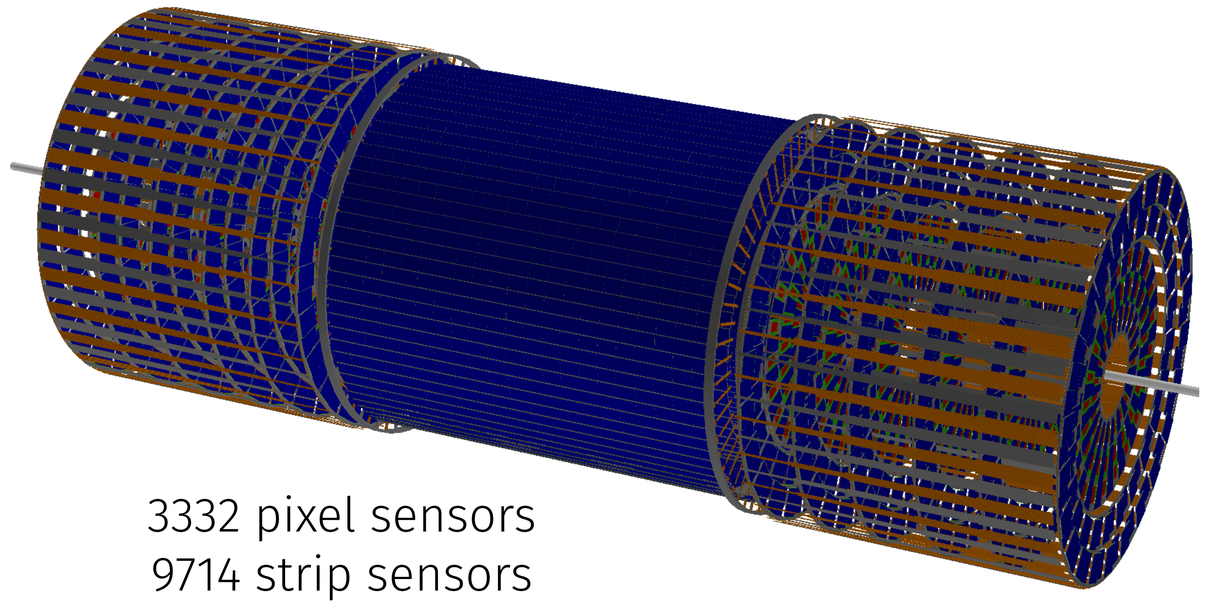
Transformers

- Transformers commonly used with sequential data (most often LLMs).
- Using decoder-only architecture.
  - Input/output data are the same.
  - Target to predict the next element of the sequence.
    - The well known example are the GPT family of models.
- Specialised on discrete sequences which are tokenised (sequential integers).

GPT & Text	Physics
paragraph	track
sentence	hit
word	hit feature

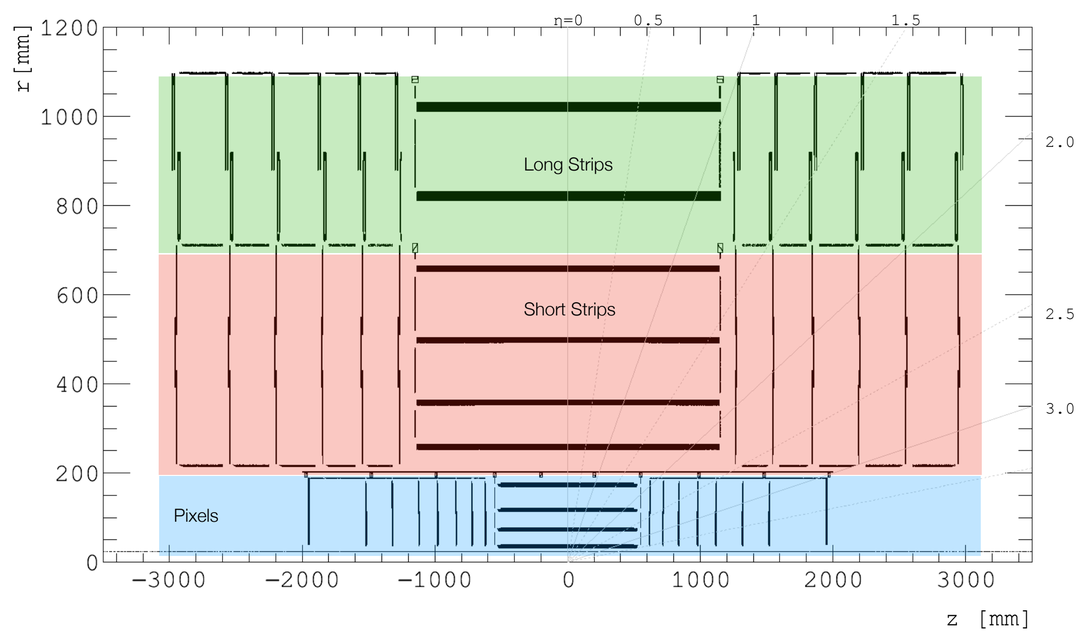
Open Data Detector

- A generic, HL-LHC style tracking detector.
- Each sensor split into multiple readout channels.
  - Can be described as a 2D surface.
- Goal to be reasonably close to a real-world detector.
  - Loosely modelled after the ATLAS ITk (58700 sensors, ~5 billion electronic channels).
- Ensures the ability to generalise R&D projects for silicon tracking detectors.

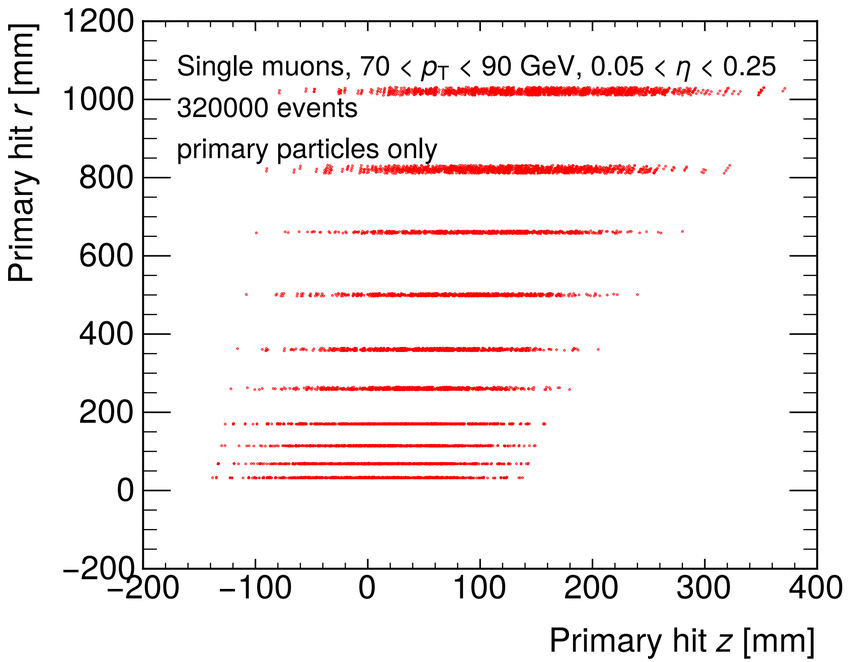


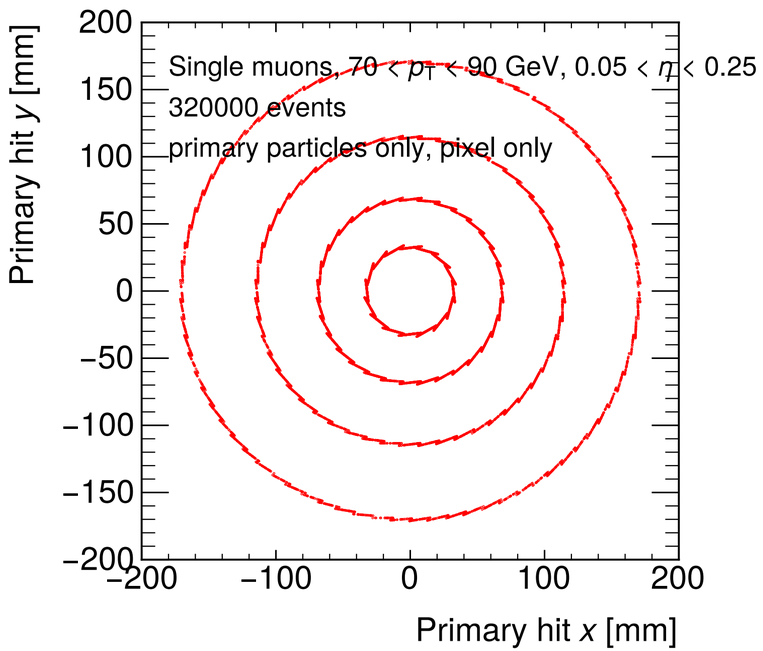
3332 pixel sensors
9714 strip sensors

Source: The Open Data Detector Tracking System



Reference & Training Data

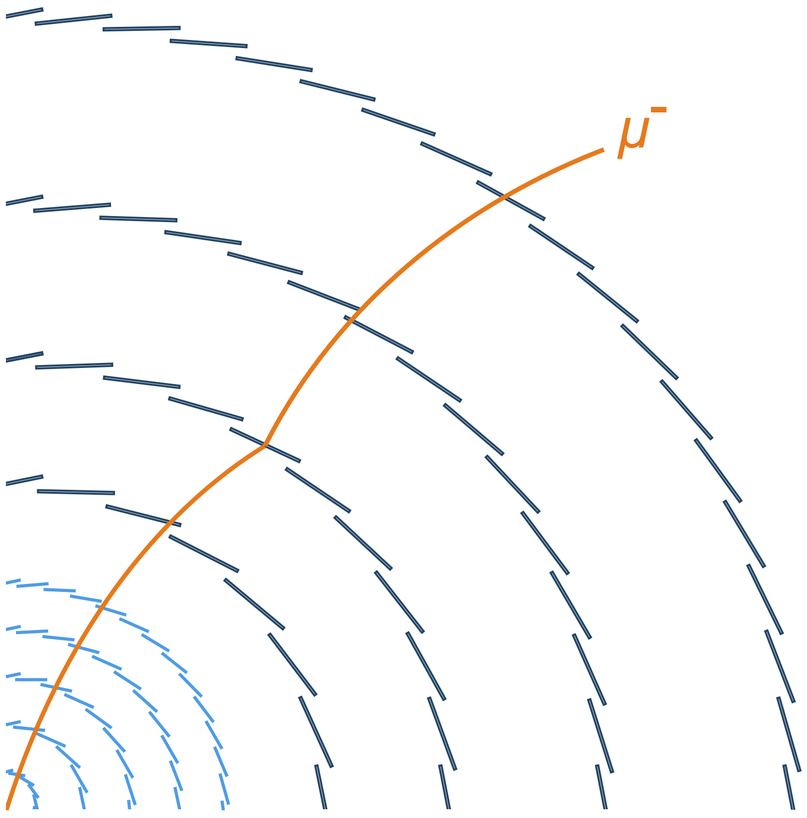




- Single-particle datasets generated and processed using adapted ACTS tracking software validation chain.
  - Full Geant4 simulation.
  - Only primary particles considered — effectively only modelling multiple scattering.

Dataset	p_T [GeV]	η	ϕ	Events
single μ^-	80-85	0.05-0.1	0-0.1	10^6
single $\mu^\pm$	70-90	0.05-0.25	incl.	10^8
single e^-	80-85	0.05-0.1	0-0.1	10^6
single π^+	80-85	0.05-0.1	0-0.1	10^6

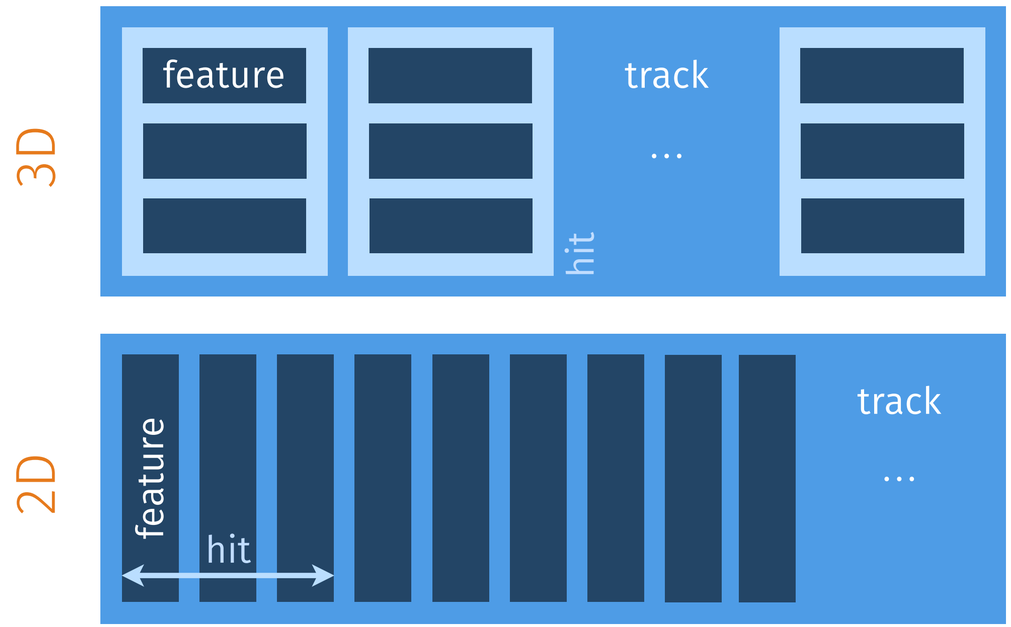
Data Representation of Silicon Detector Hits (1)



- Track = a sequence of detector hits.
  - With additional start and end “virtual hit” to describe input and output state with the same data structure.
- 8 features per hit:
  - hit index (auxiliary feature)
  - particle ID + geometry ID
  - particle momentum
  - local hit position on the sensitive detector
- Local coordinates taken to constrain hits on the sensitive parts and prevent them happening in the vacuum.

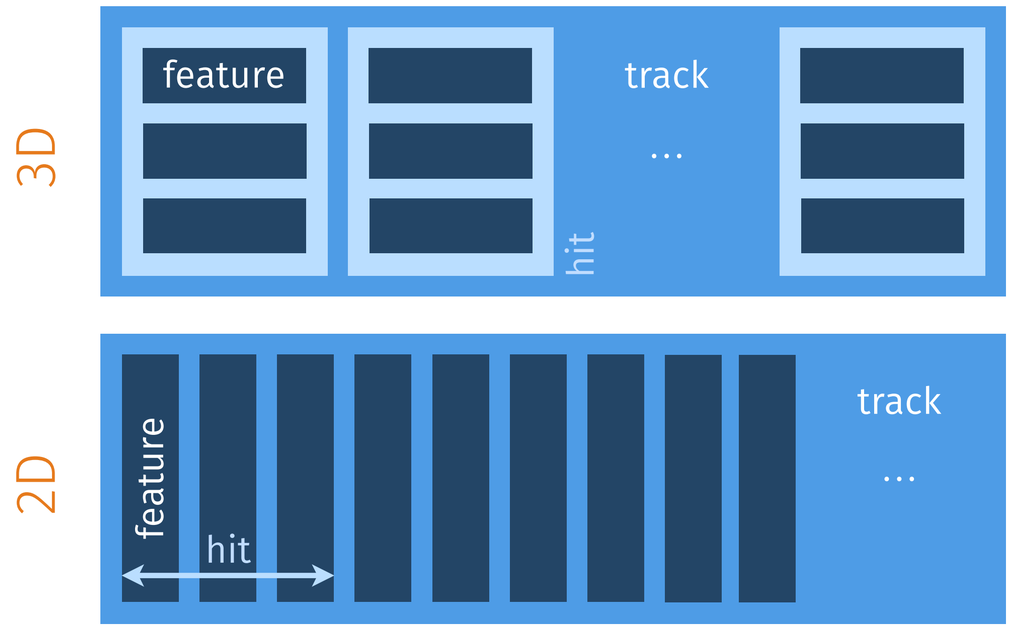
Data Representation of Silicon Detector Hits (2)

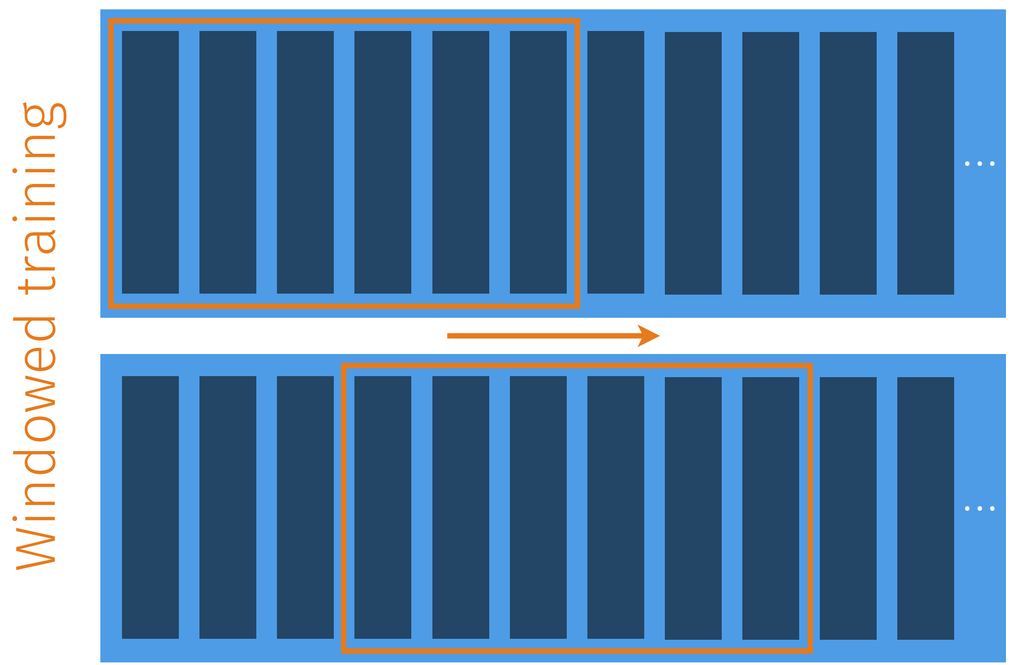
- Flat sequence, 2D “GPT-like” information.
  - One token per feature.
    - Total of up to 19125 tokens (of which 2222 different detector modules).
  - Every k^{th} sequence element represents k^{th} feature.
  - Numerical features rounded to two decimal points.



Data Representation of Silicon Detector Hits (2)

- Flat sequence, 2D “GPT-like” information.
  - One token per feature.
    - Total of up to 19125 tokens (of which 2222 different detector modules).
  - Every k^{th} sequence element represents k^{th} feature.
  - Numerical features rounded to two decimal points.
- Training on windows of 4 hits.
  - Predicting the 4^{th} hit based on the 3 input ones.
  - Maximum sequence length 34 tokens.
  - Inference iterative per feature — full correlations taken into account.

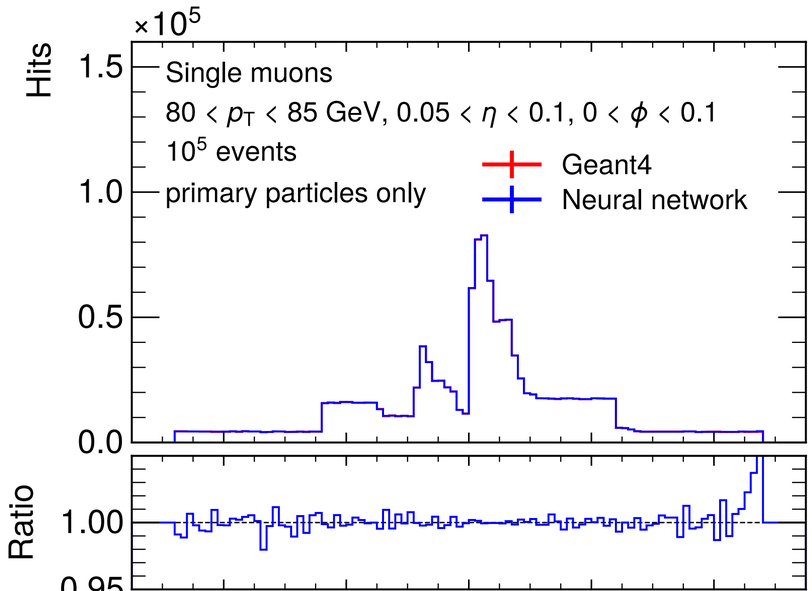


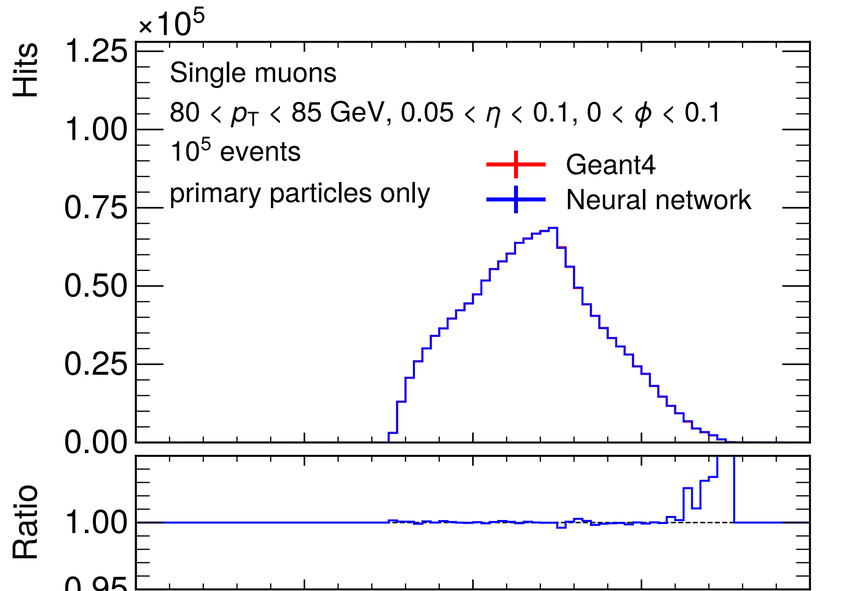


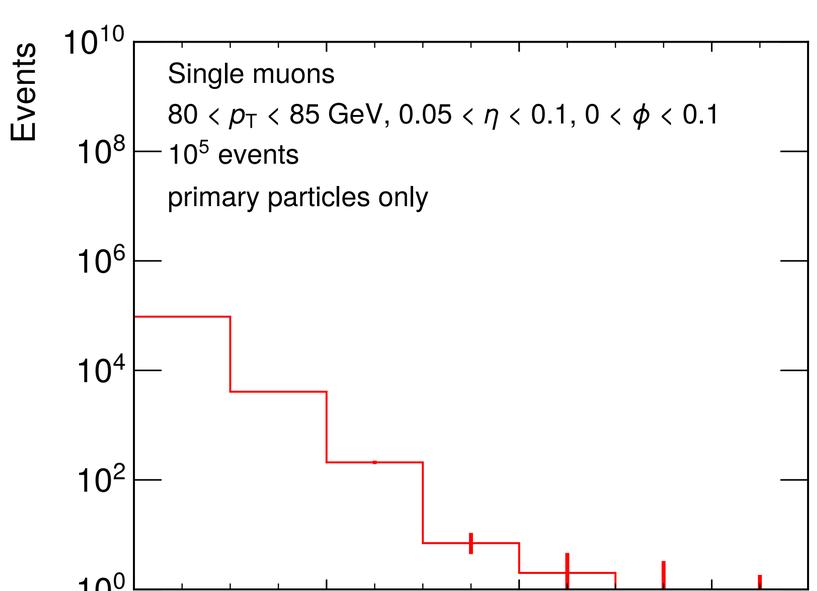
Training Setup

- Using off-the-shelf nanoGPT implementation.
  - Powered by PyTorch.
  - GPT2-like model.
- Two model sizes with different neural network layer dimensions.
  - 11.2M and 35.0M parameters for muons
    (size also slightly depends on the token dictionary size).
- Training on Vega and Arnes HPCs using Nvidia A100 and H100 GPUs.
  - duration: ~5 days on 2xH100 GPUs
  - large models — benefiting from large memory and multi-GPU support
- Tested also on the UL FRI research cluster FRIDA.
  - Universal code, able to run also on AMD GPUs.

Hit-level Observables Modelling: Small Single-muon Dataset

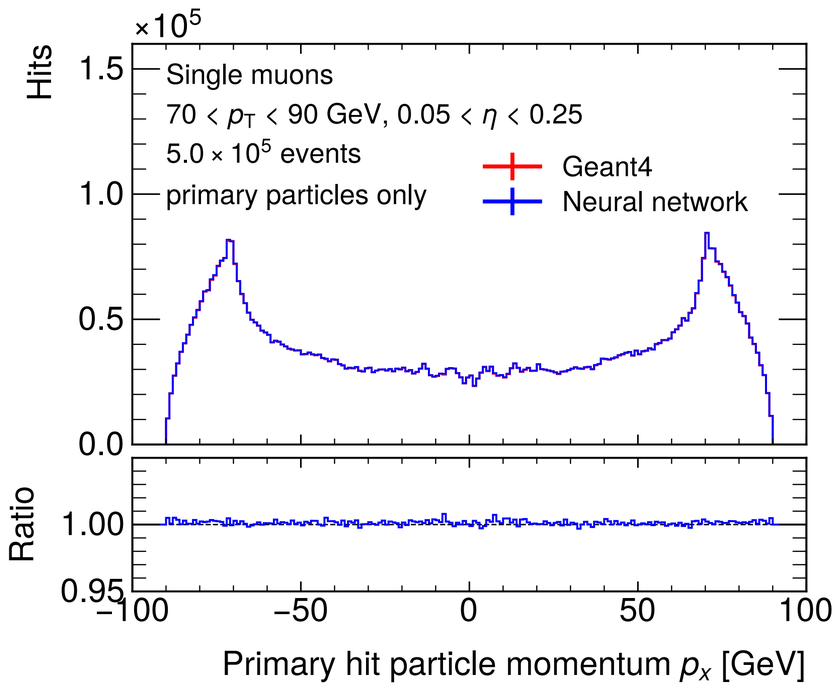


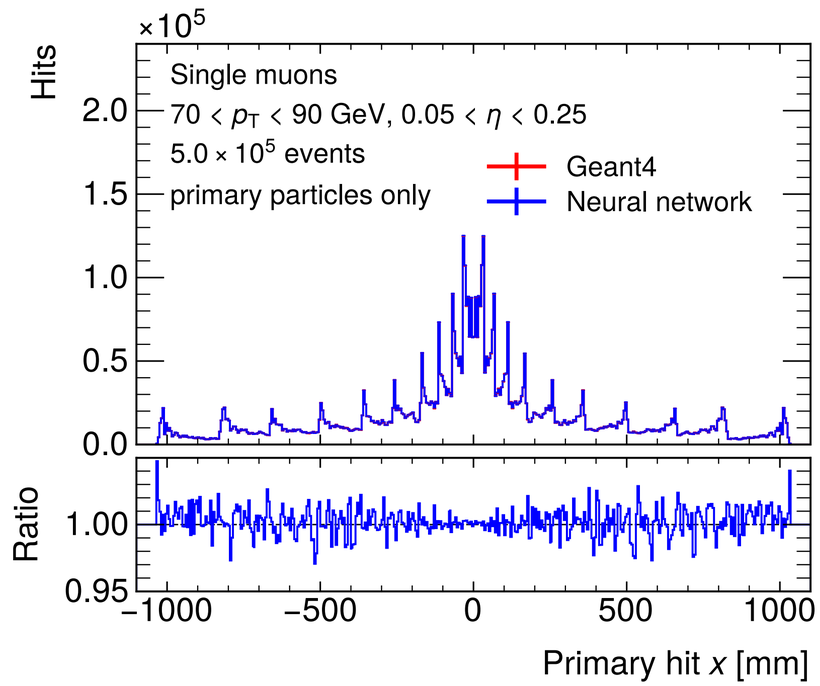




- Hit-level observables match well between Geant4 and the neural network.
  - Both raw trained observables (surface x coordinate) and derived quantities (global z coordinate).
- Difference in total number of hits small.
  - Mostly happen at module edges were overlap between modules may be present.

Hit-level Observables Modelling: Large Single-muon Dataset





- Similar level of agreement as when using smaller model or input dataset.
  - Larger fluctuations of global coordinates but all within 5 %.
- Reminder: The model is generative so fluctuations are expected.

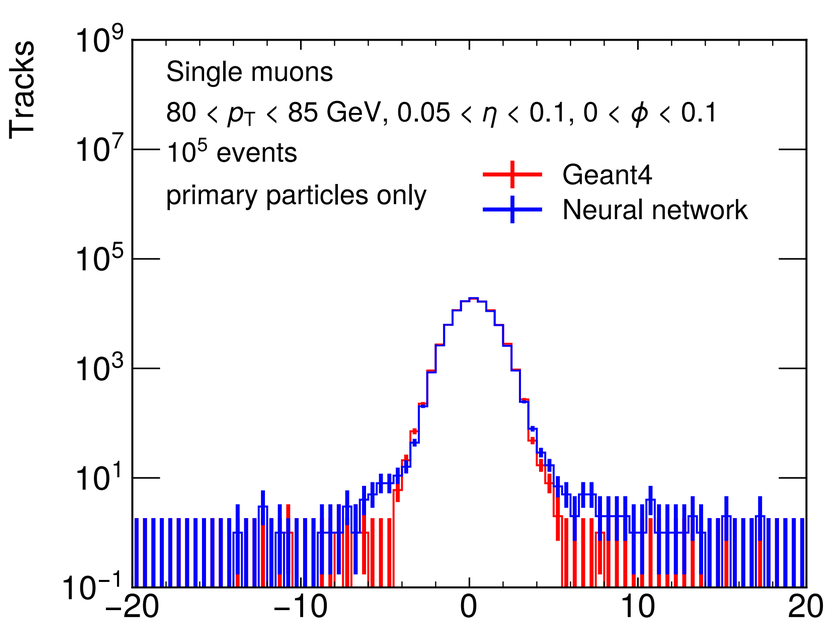
Higher-Level Validation

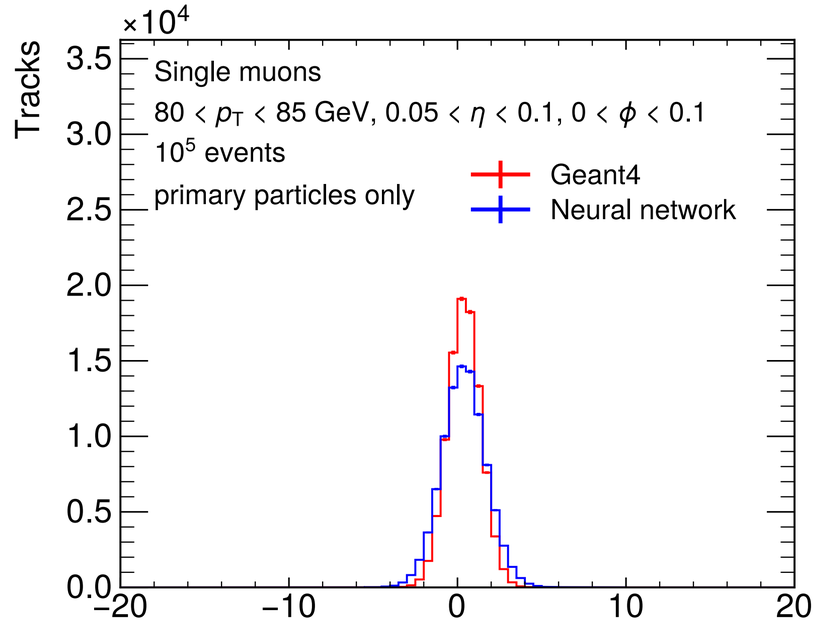
- Reconstructing the output into tracks and looking at higher level physics observables.
  - track seeding — finding triplets of hits that would produce a physically feasible track.
  - track fitting — matching track seeds with the remaining hits and fitting the track
- Rounding already reduces the efficiency — better input preparation is needed.

Tracking efficiency

Sample	Track seeding efficiency	Track fitting efficiency
G4 Reference	99.9%	99.9%
G4 Reference (rounded)	99.9%	98.1%
Generated (11.2M par.)	99.4%	94.9%
Generated (35.0M par.)	99.7%	96.3%

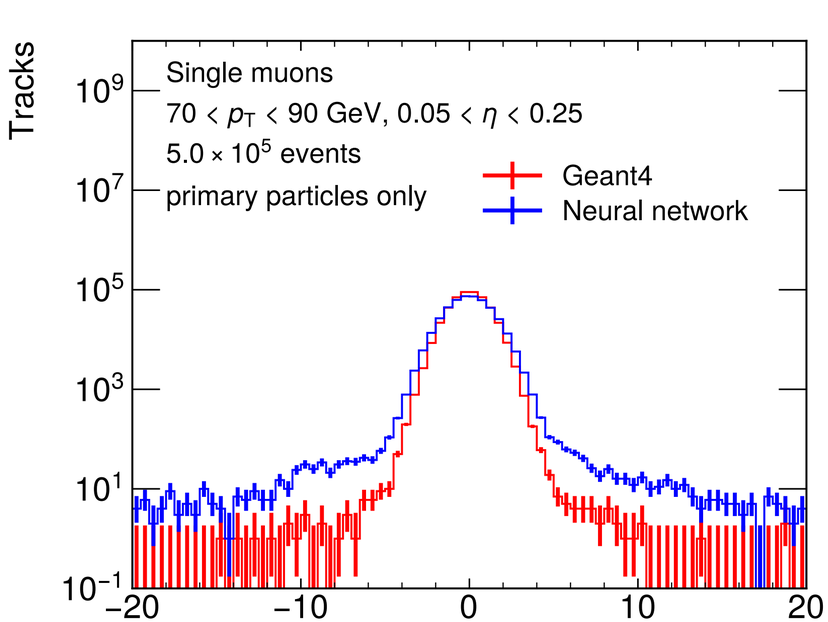
Track Properties: Small Single-muon Dataset

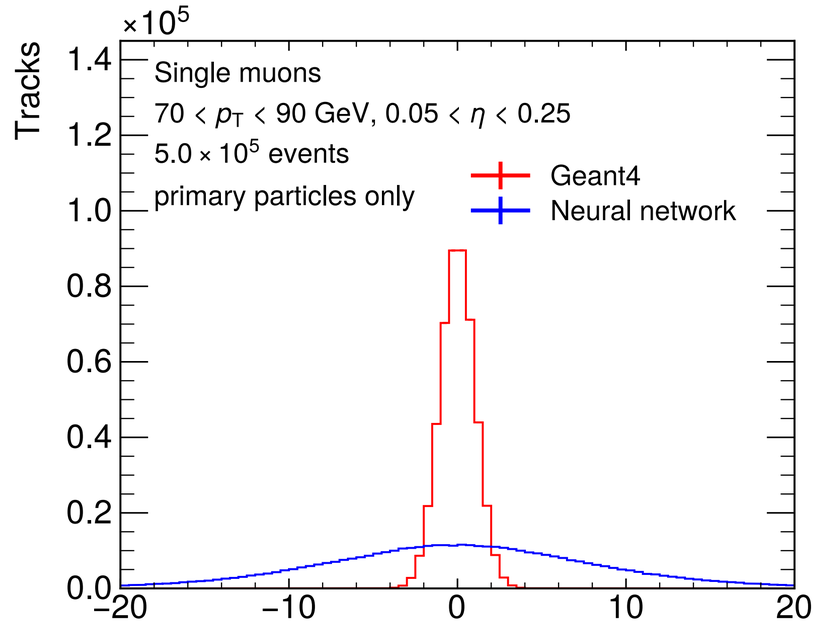




- Fairly good modelling of track properties (taking into account inefficiencies)
  - Depends on the property, but both reconstructed properties and their resolutions agree well.
- Neural network produces wider distributions with longer tails.

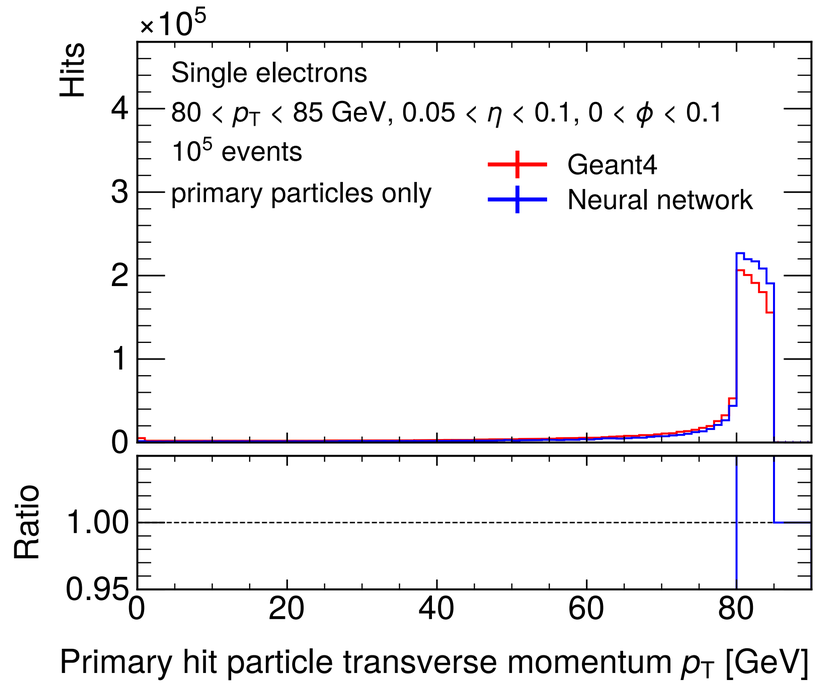
Track Properties: Large Single-muon Dataset

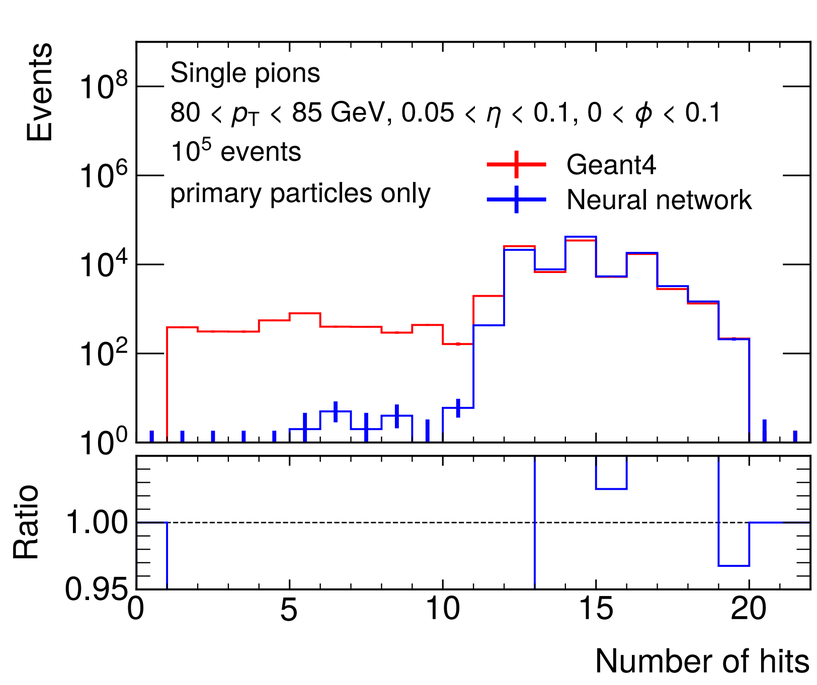




- Going towards a larger phase space a significant quality reduction is observed.
  - Distributions are wider, including the resolution.
  - The ϕ coordinate is especially poorly modelled with the resolution distribution about 5x wider.
  - The average values are learned well.

Electron & Pion Performance





- Electron momentum loss not properly modelled.
- Relatively low probability of pions decaying/stopping in the tracker is not properly captured.
- These are low-statistics tails that do not get captured by the model.
- Tracking efficiency and properties comparable (taking into account the bias).

Inference Speed: Comparison with CPU

Inference wall time per 10 000 tracks

Compute type	11.2M par.	35.0M par.
24 CPU cores: AMD Zen 2	35 min	80 min
24 CPU cores: AMD Zen 4	17 min	37 min
GPU: NVIDIA A100 PCIe	16.4 s	36.0 s
GPU: NVIDIA H100 PCIe	11.4 s	21.4 s
GPU: NVIDIA H100 SXM	8.0 s	13.9 s

Geant4 on 24 AMD Zen 2 cores: 35 s
24 AMD Zen 4 cores: 17 s

- Transformers too large/slow to be useable on CPUs.
- On GPUs fast memory throughput needed (e.g. PCIe vs SXM).
- Same-generation CPU-GPU pairs comparable in speed between Geant4 and NVIDIA GPUs.
- Reducing model precision to half 16-bit precision (bf16) improves throughput for 33 % with no impact on physics performance.

Conclusions & Outlook

- Good ensemble agreement can be achieved for detector hits.
- Training fairly long, inference speed significantly depends on sequence length (due to slow attention mechanism).
- Large token space requires a large model, can not reach Geant4 physics performance when looking at a large phase-space.
  - Electrons and muons suffer from tail effects that need specialised handling during neural network training.
- Optimal computing infrastructure are HPCs with significant fraction of GPUs (assuming physics performance can be made comparable with Geant4).
- Results published in arXiv:2512.24254
- Further work is needed to optimise the model and to be able to apply it to one of the actual experiments.

I FEEL SLOVENIA

Co-funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 101081355.

The operation (SMASH project) is co-funded by the Republic of Slovenia and the European Union from the European Regional Development Fund.

Decoder-only Transformer (with numbers from my 3D model)

Source: Cameron R. Wolfe

output linear layer
d = n_tokens

Output Token Vectors

layers
n = 8

Decoder Block
Decoder Block
⋮
Decoder Block

positional encoding
sin/cos

Position Embedding

Input Token Vectors

token embedding layer
d = 512
(model dim)

feed-forward NN
d = 2048
linear model dim → ff dim
and back with dropout 0.2

FFNN

Layer Norm

Masked Self-Attention

Layer Norm

multi-head attention
heads = 8

masking ensures that only previous elements in a sequence are used

Muon Training Setup

- Using off-the-shelf nanoGPT implementation.
- Learning rate scheduling with linear rise over a few epochs and then cosine decay over ~4000 of them.
- Training on Vega and Arnes HPCs using NVIDIA A100 and H100 GPUs.
  - duration: ~5 days on 2xH100 GPUs

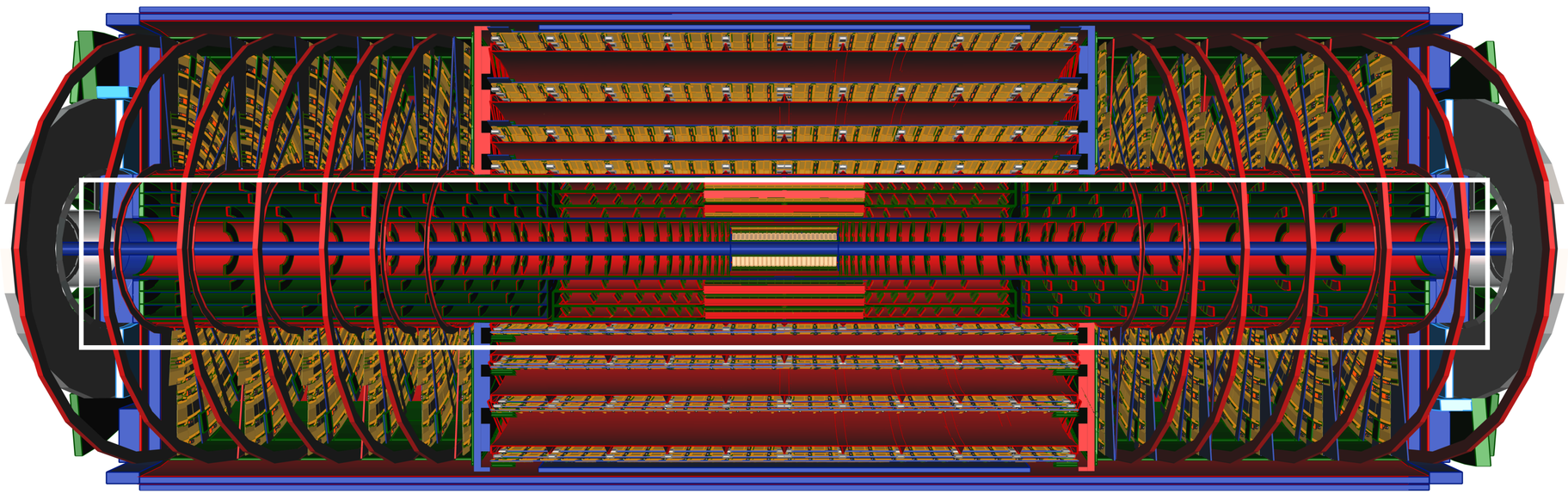
Model Parameter	11.2M par.	35.0M par.
input dimension	256	512
layers	8	8
heads	8	8
feedforward dim.	1024	2048
activation	GELU	GELU
dropout	0.2	0.2

Training Parameter	11.2M par.	35.0M par.
optimiser	AdamW	AdamW
base learning rate	0.004	0.002
minimum learning rate	0.0004	0.0002
weight decay	0.01	0.01
gradient clipping	5.0	5.0
best epoch	5200	5200

Inference Speed: Nvidia GPU Comparison

Precision:	Training [s / epoch]		Inference [s / 10 000 tracks]	
	`fp32`	`bf16`	`fp32`	`bf16`
A100 40GB SXM	75	50	13.0	9.34
A100 80GB PCIe	77	52	15.1	11.1
H100 80GB SXM	38	26	6.78	5.21
GH200 96GB	34	23	6.23	5.68
B200 180GB SXM	31	21	4.54	4.28

Structure of the ITk



Source: ATL-PHYS-PUB-2021-024

Pixel detectors

- 2D silicon detectors
- 5 barrel, 9 endcap layers
- 9164 modules
- up to 614400 readout channels per module

Strip detectors

- 1D silicon detectors
  - double-modules with 90° rotation to gain 2D detection
- 4 barrel, 6 endcap layers
- 49536 modules
- up to 1536 readout channels per module