SMASH
machine learning for science and humanities postdoctoral program

Jožef Stefan Institute

Generative Machine Learning for Fast Silicon Detector Simulation

CHEP 2024, Track 5 - Simulation and analysis tools
October 21, 2024

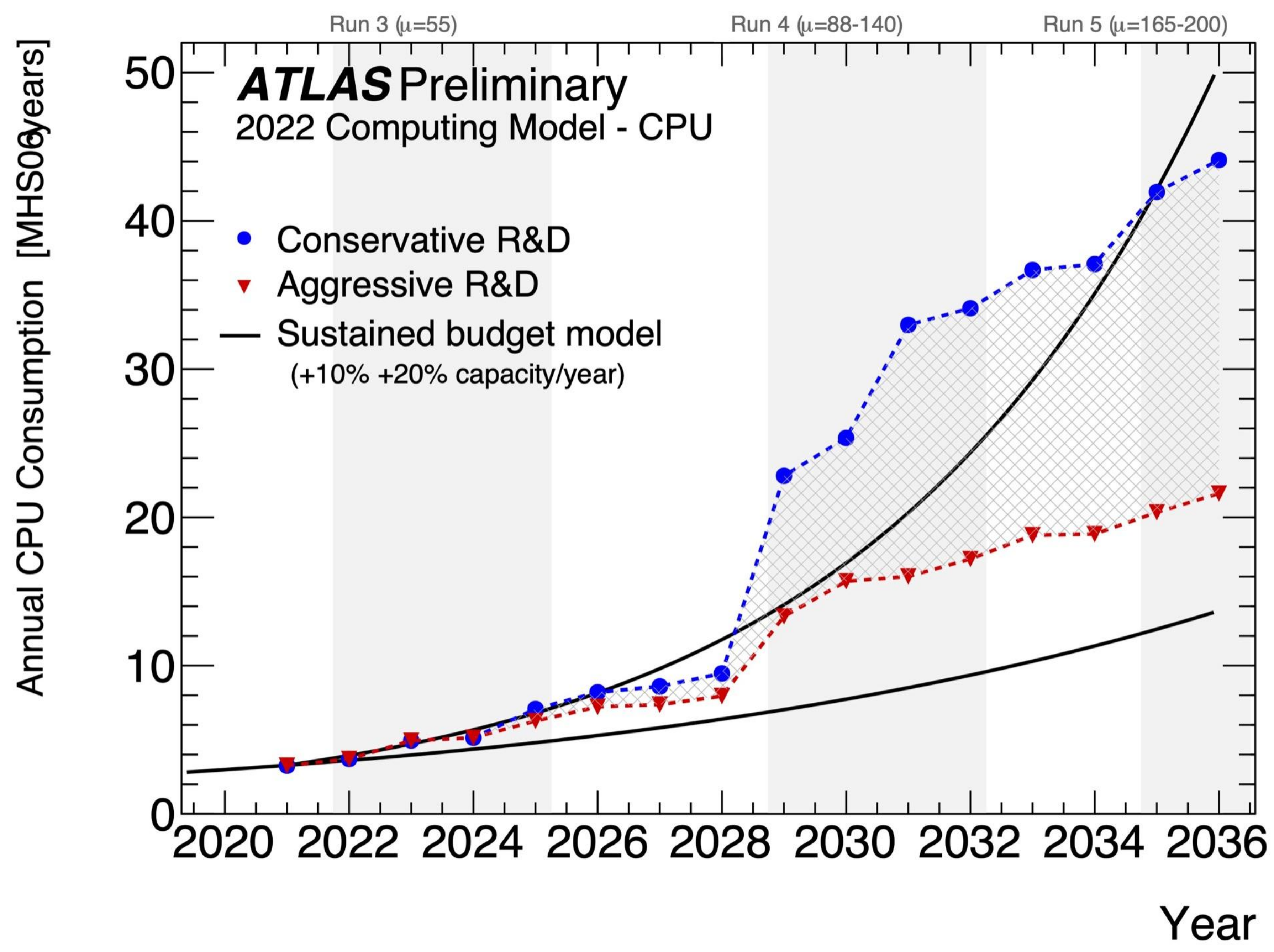
I FEEL SLOVENIA
Co-funded by the European Union

Tadej Novak
Jožef Stefan Institute

The Need for Monte Carlo Simulation

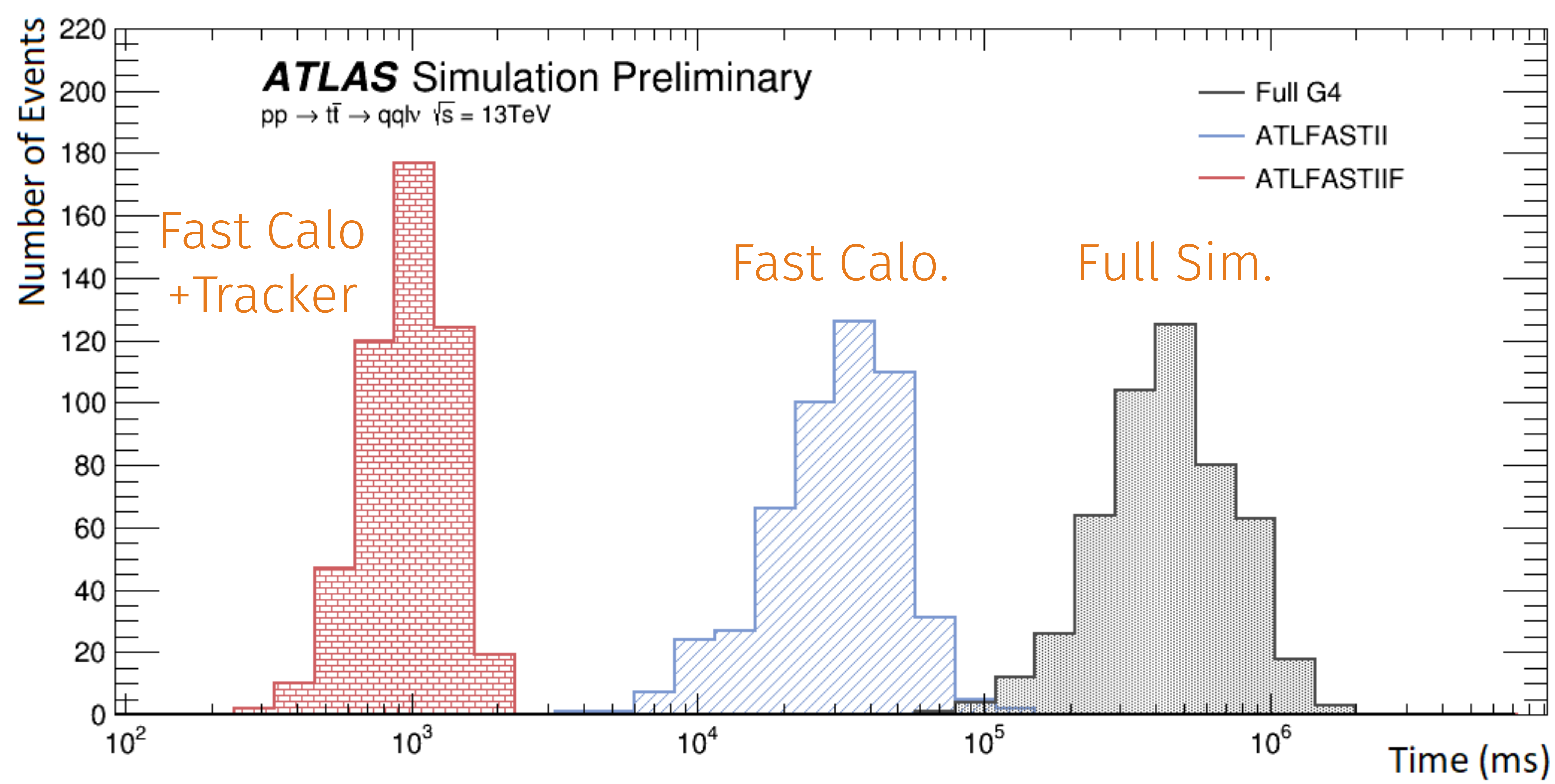
- A large part of the LHC physics programme relies on accurate Monte Carlo simulation of collision events.
  - every single particle needs to be simulated
  - detailed (full) detector response simulation most intensive
- Producing simulated samples → majority of experiments' CPU requirements
  - CMS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

Source: ATLAS Software and Computing HL-LHC Roadmap



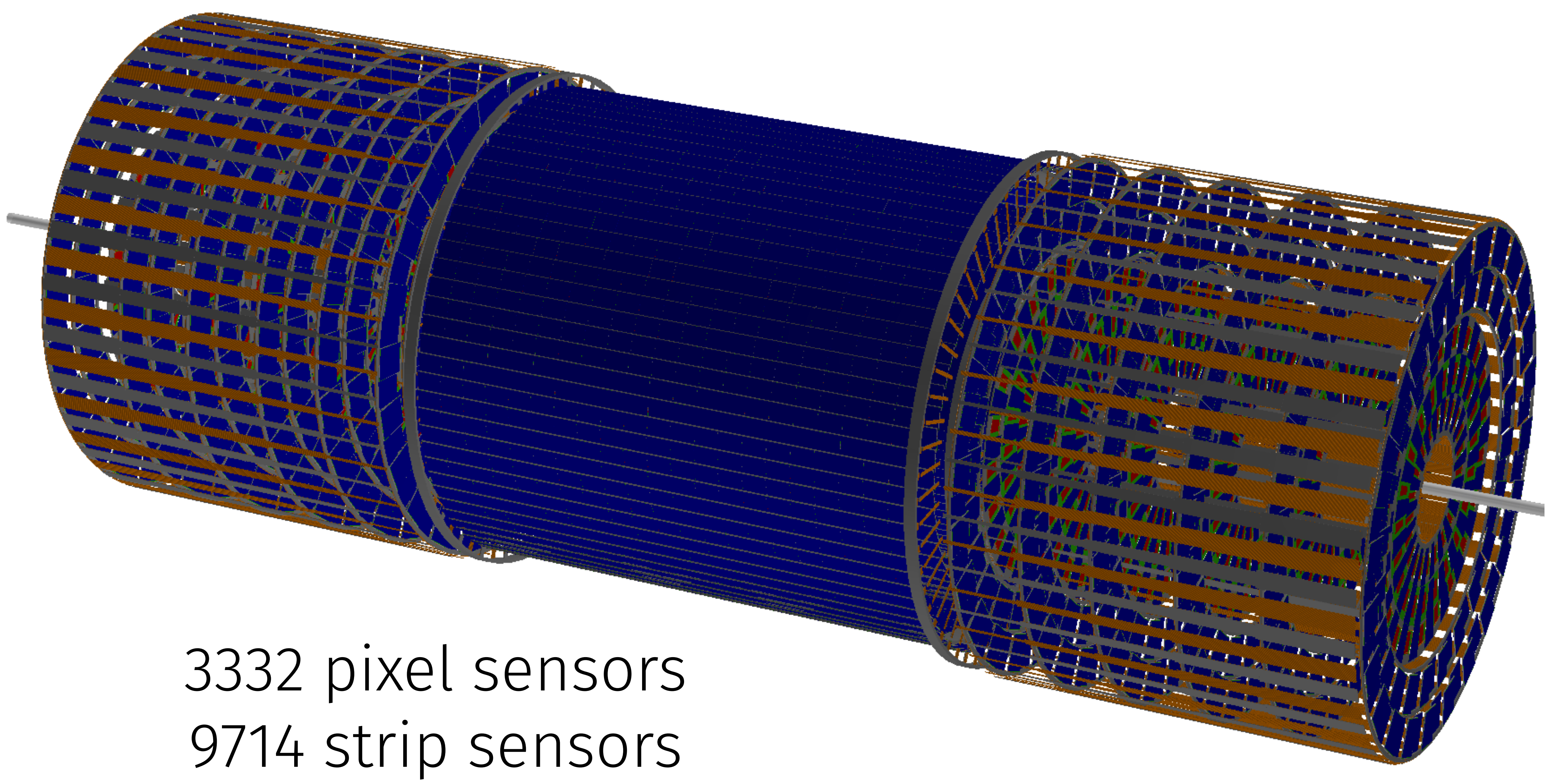
- Current methods do not scale with HL-LHC data rates and more aggressive R&D is needed.

- Large efforts to speed-up simulation — fast simulation.
  - Detector response to a particle is parameterised.
- Fast simulation for particle physics successfully applied at calorimeter level.
  - Generative neural networks also used.
  - Order of magnitude speed-up achieved.
- ATLAS tracking detectors fast simulation not production-ready yet.
  - Machine learning target of this project.



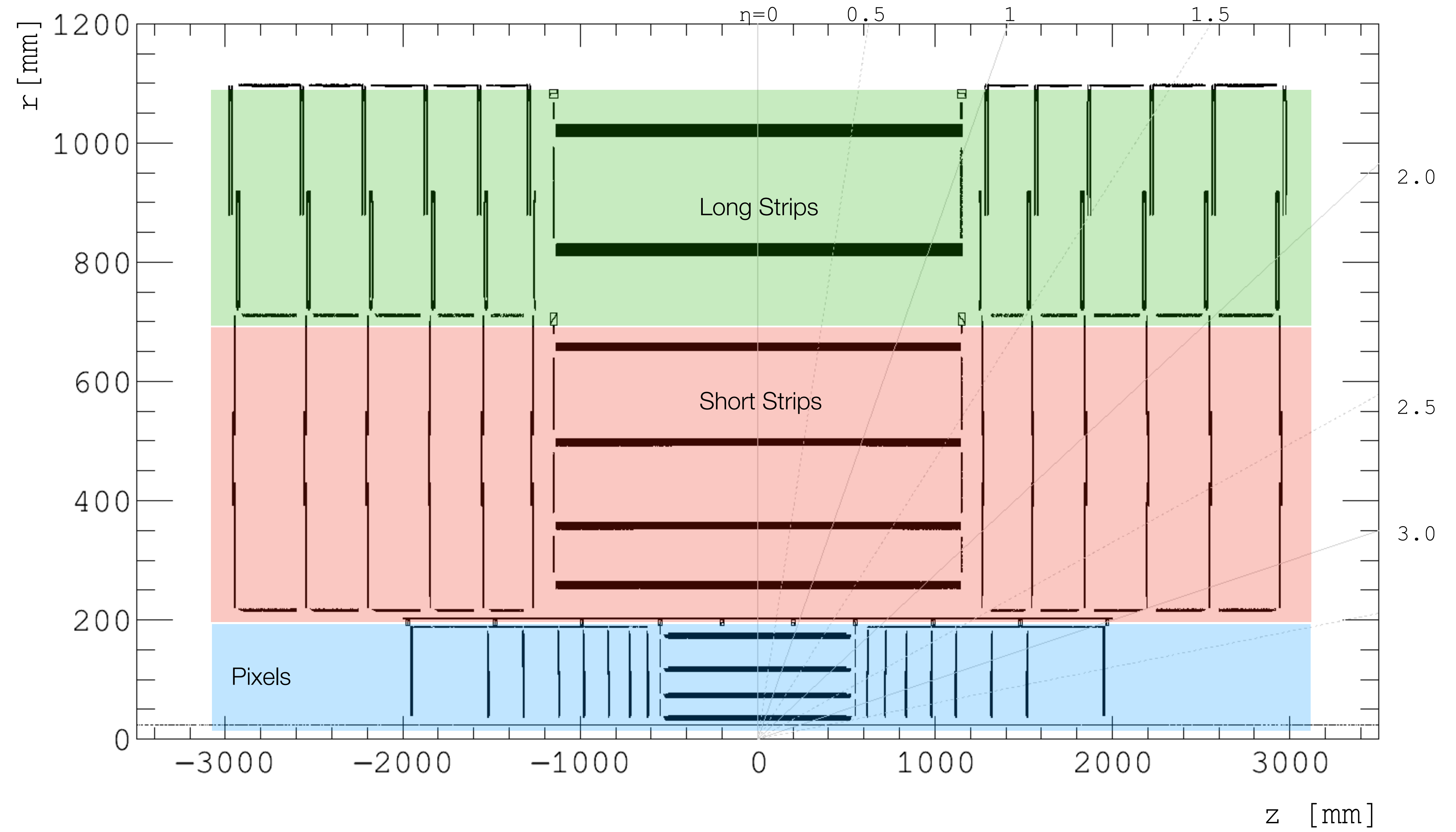
Source: ATLAS Simulation CPU Performance

- 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.

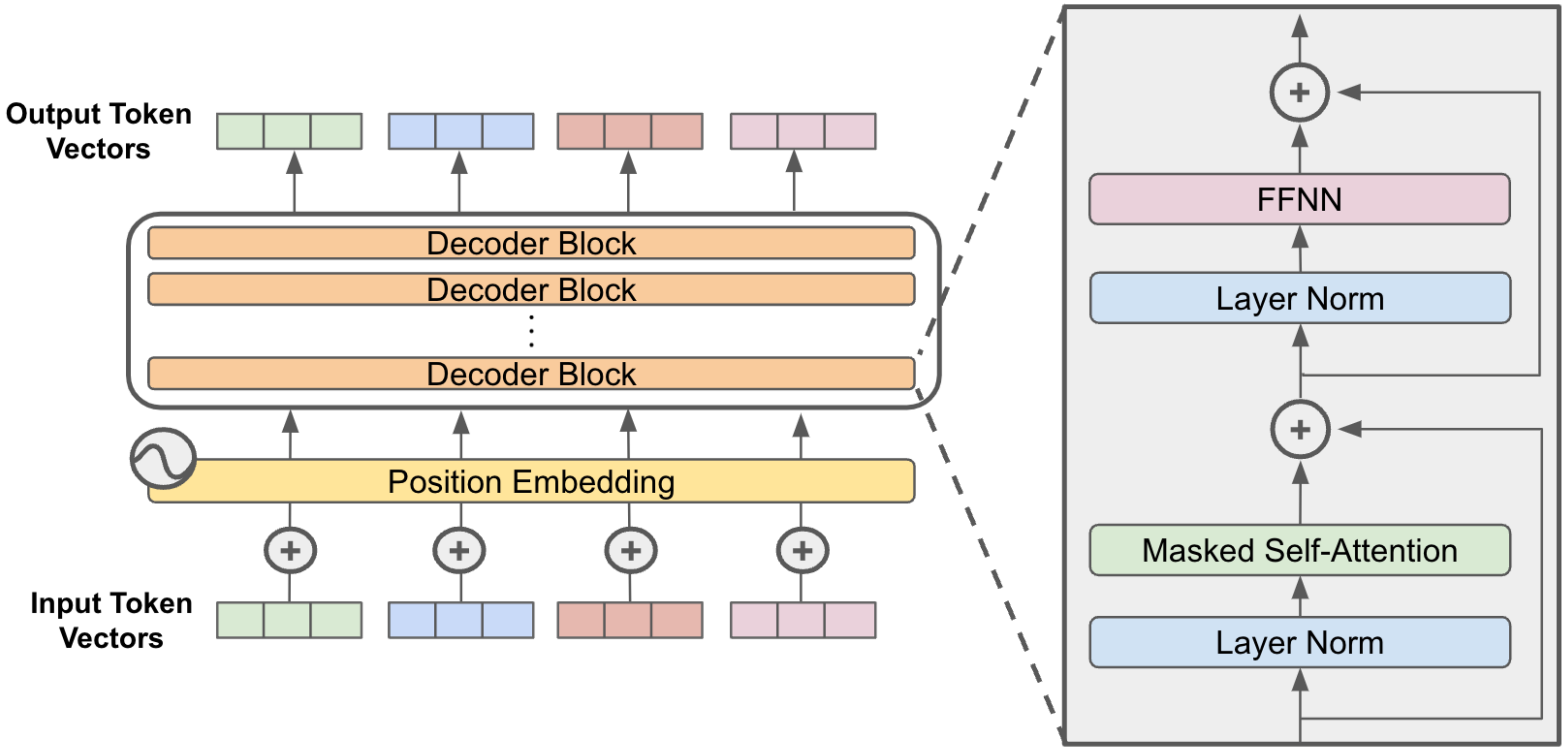


3332 pixel sensors
9714 strip sensors

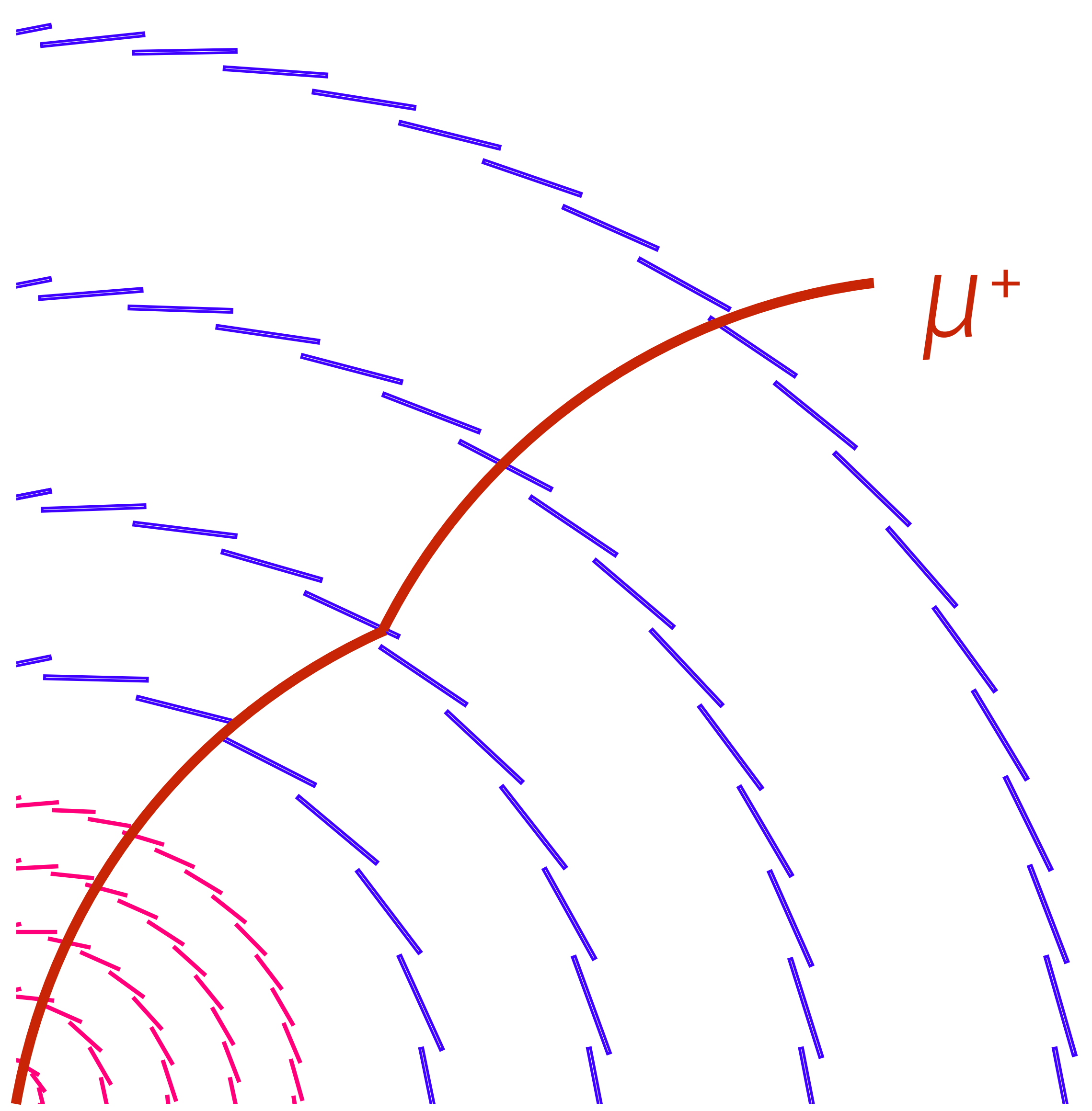
Source: The Open Data Detector Tracking System

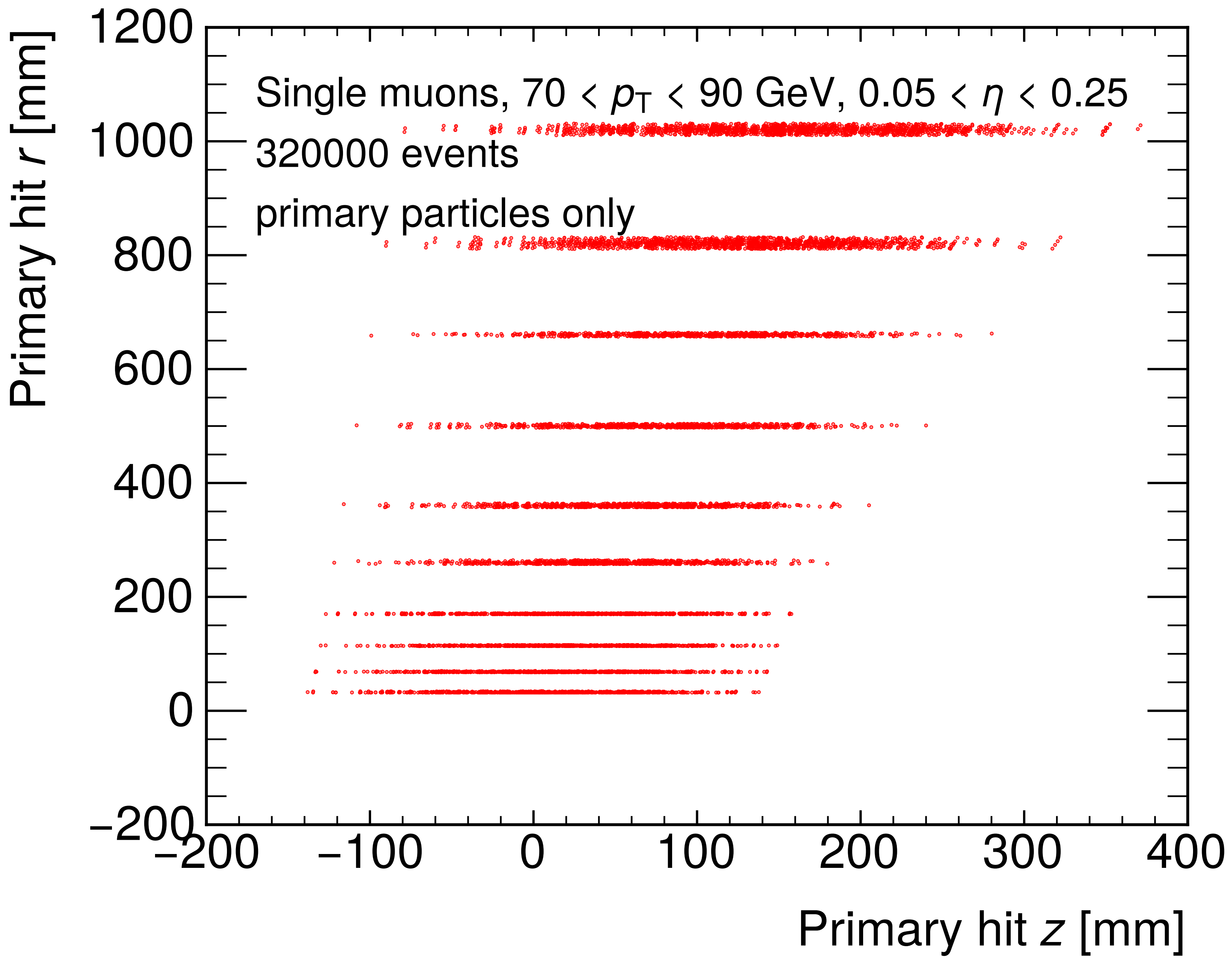


- Transformers commonly used with sequential data (most commonly LLMs), see 1706.03762.
- Using decoder-only architecture.
  - Input/output data are the same.
  - Target to predict the next element of the sequence.
  - The well known example are the GPT family of models.
- Specialised on discrete sequences which are tokenised (sequential integers).
  - Can be anything e.g. words, detector modules, ...
- For this application all continuous data is discretised (rounded to two decimal points) and each feature is tokenised separately.



Source: Cameron R. Wolfe





- A sequence of detector hits.
    - With additional start and end “virtual hit” to describe input and output state with the same data structure.
- 7 features per hit:
    - particle ID + geometry ID
    - particle momentum (after the hit)
    - hit position on the sensitive detector (local)
- Each hit is an element of a sequence, each particle has its own sequence.
- Local coordinates taken to constrain hits on the sensitive parts and prevent them happening in the vacuum.

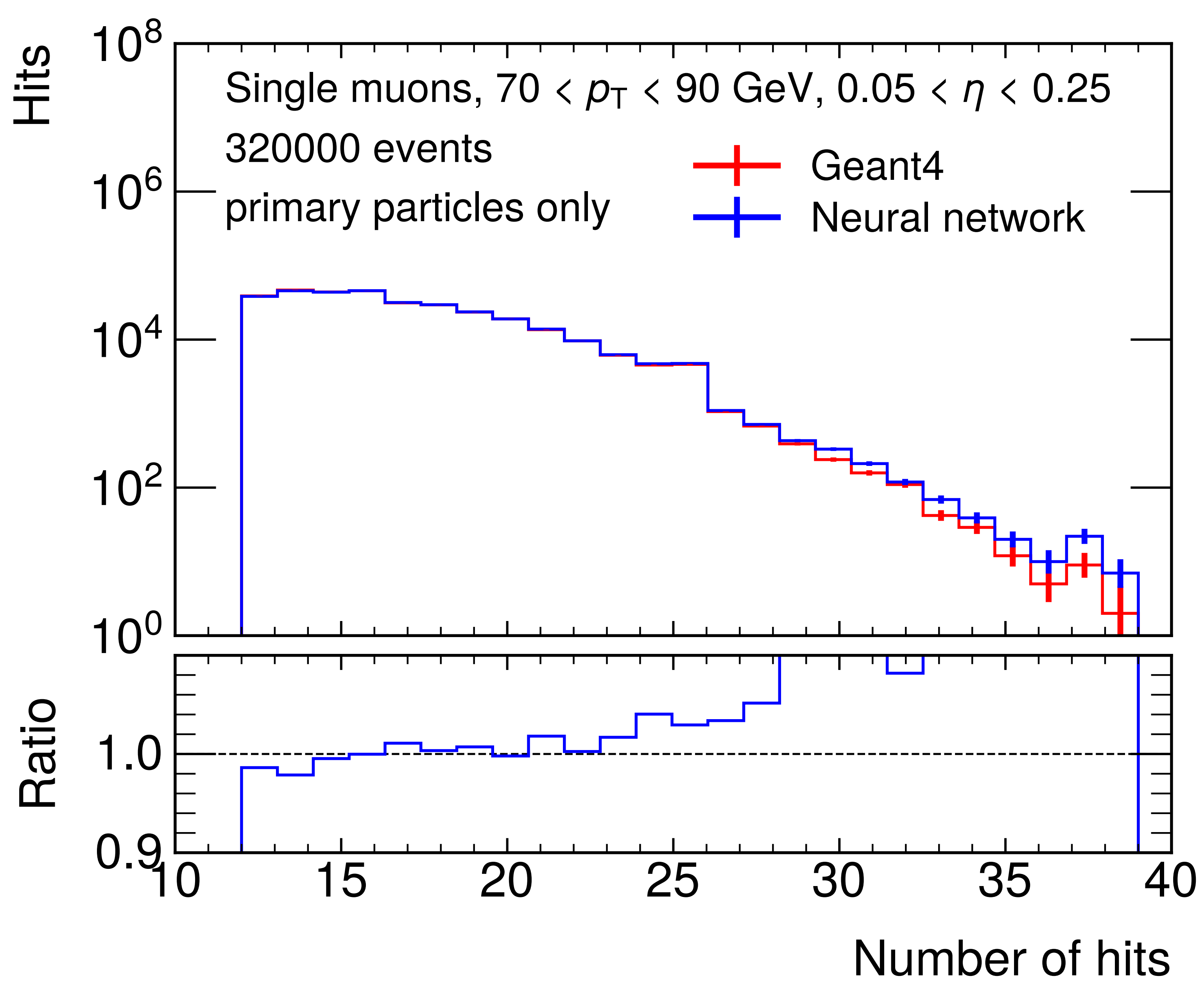
Training & Inference Setup

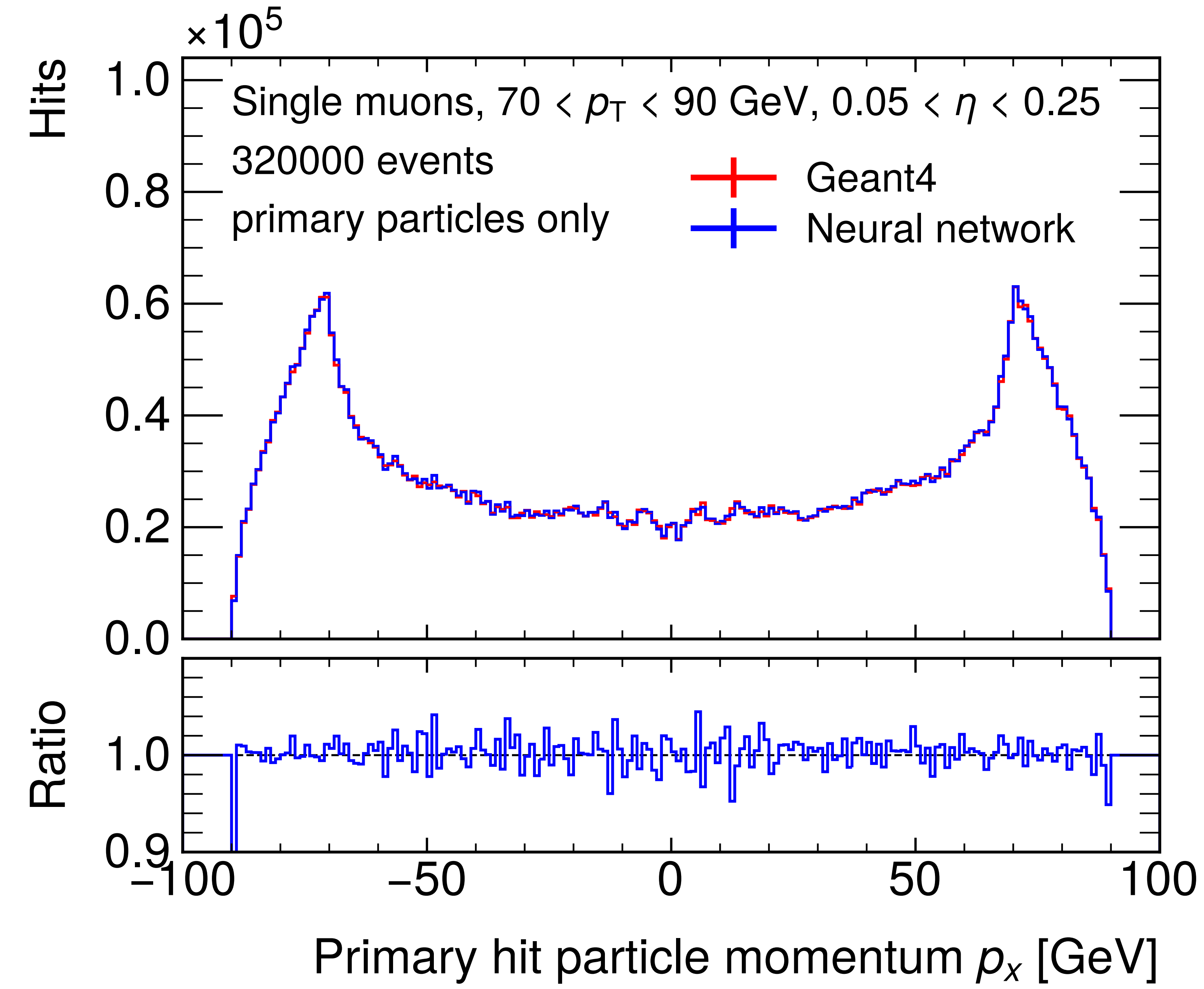
- Sample details:
  - single muons, $70 < p_T < 90$ GeV, $0.05 < \eta < 0.25$
  - 320000 events
  - training : validation : test = 2 : 1 : 1
  - augmented with random numbers between 1 and 10000
- Training performed on the Vega HPC using 4x NVIDIA GeForce A100 40GB GPU.
  - model size: 30.4 M parameters
  - duration: ~6 days
- Learning rate variation using cosine annealing with warm restarts with a period of one epoch and fixed amplitude.
- Inference: most probable next sequence element
  - ~8 s / 10k particles on a single A100

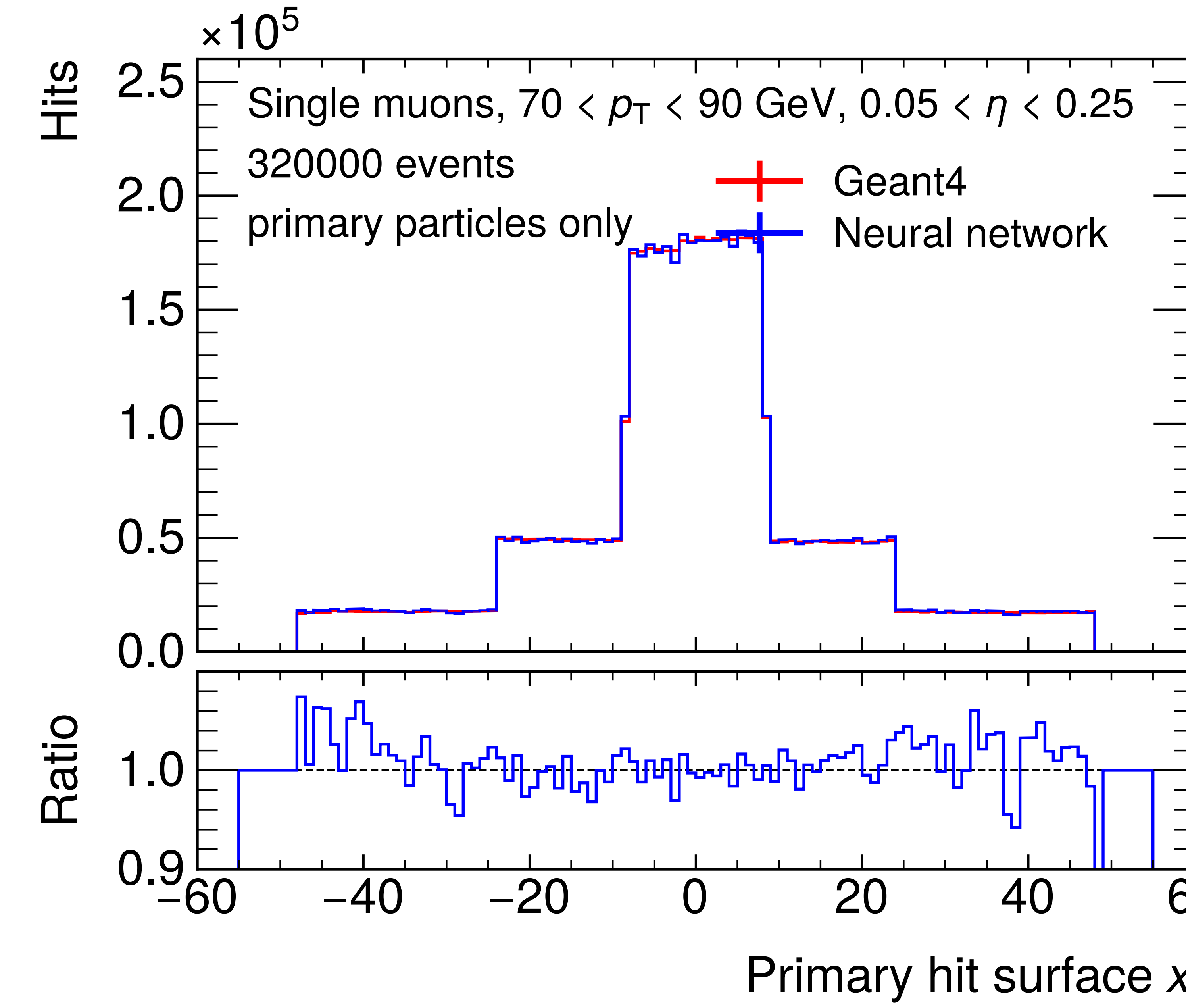
Model Parameter	Value
input dimension	256
layers	3
heads	4
feedforward dim.	1024
activation	GELU
dropout	0.1

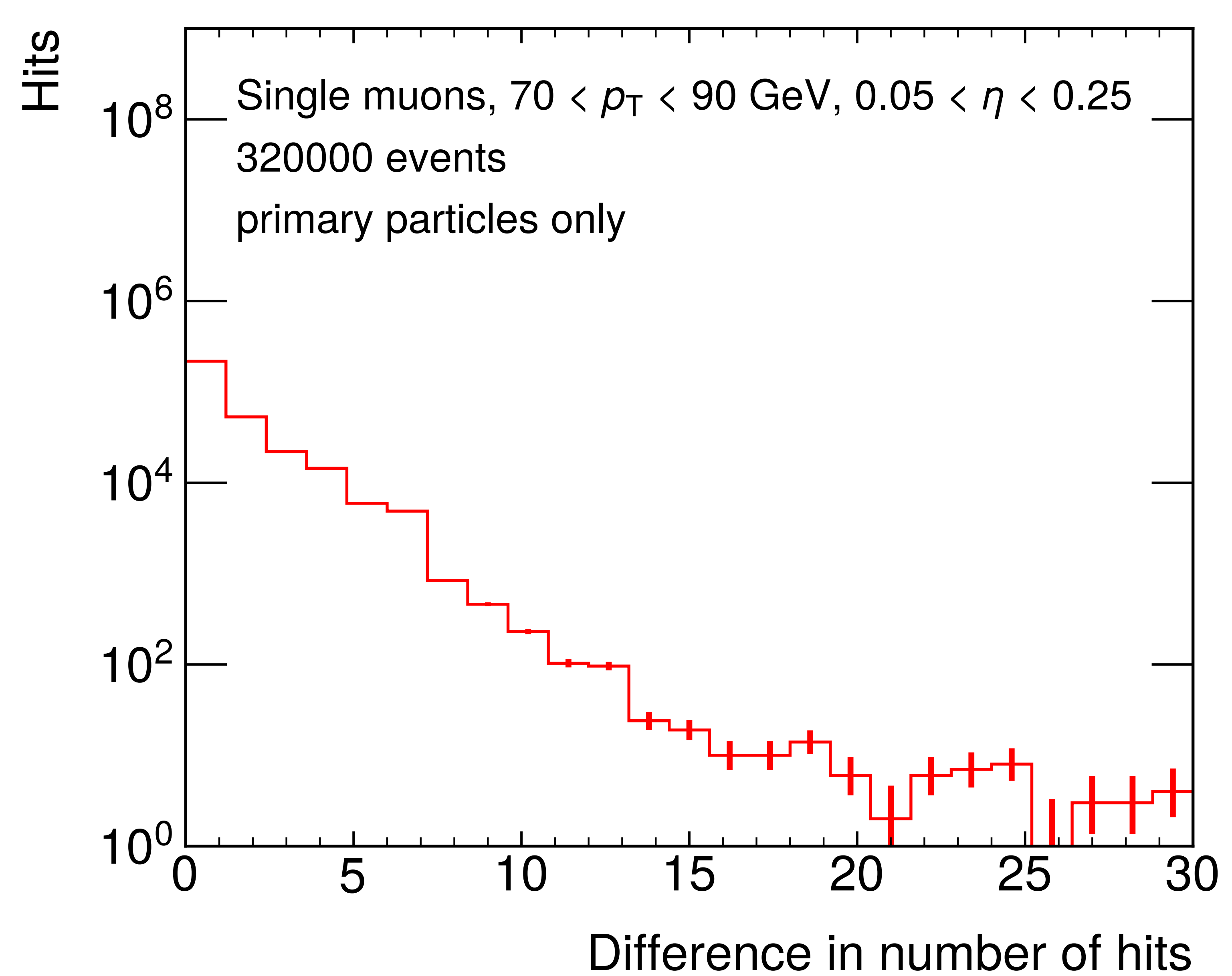
Training Parameter	Value
epochs	15000
optimizer	AdamW
learning rate	0.001
weight decay	0.01
gradient clipping	5.0
batch size	2400

Results: Simulation (1)

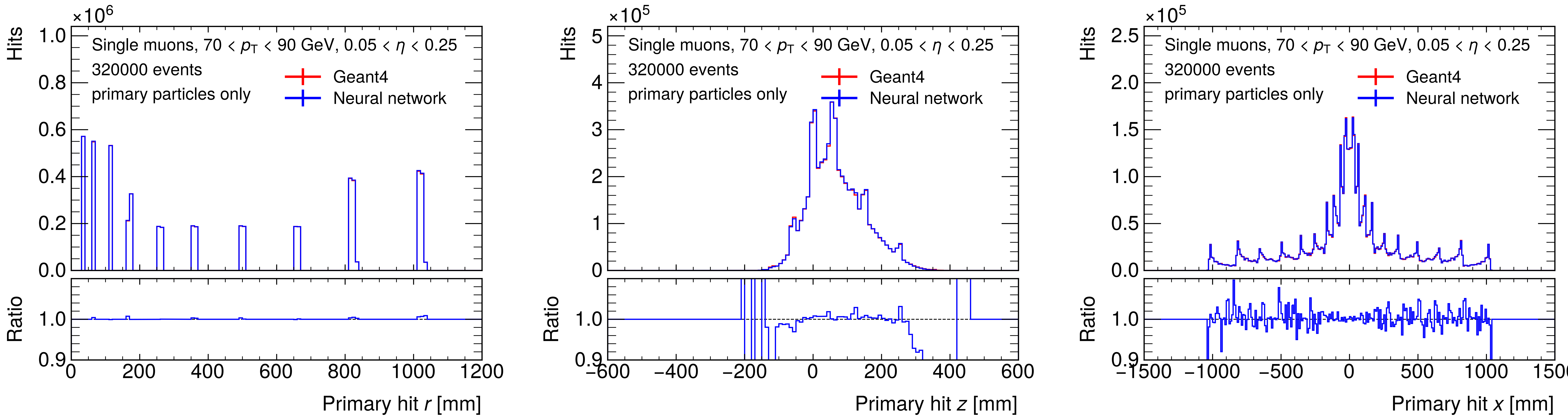






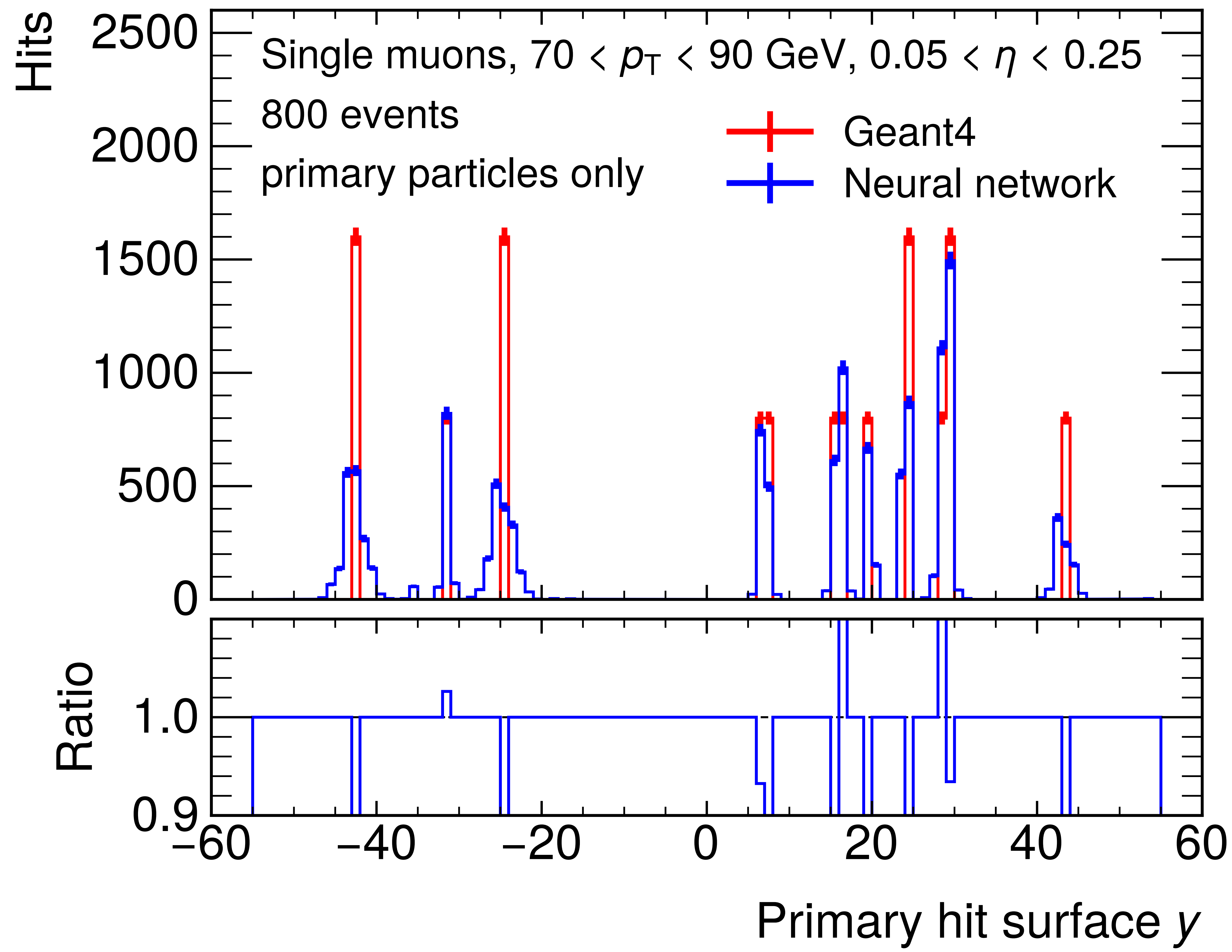


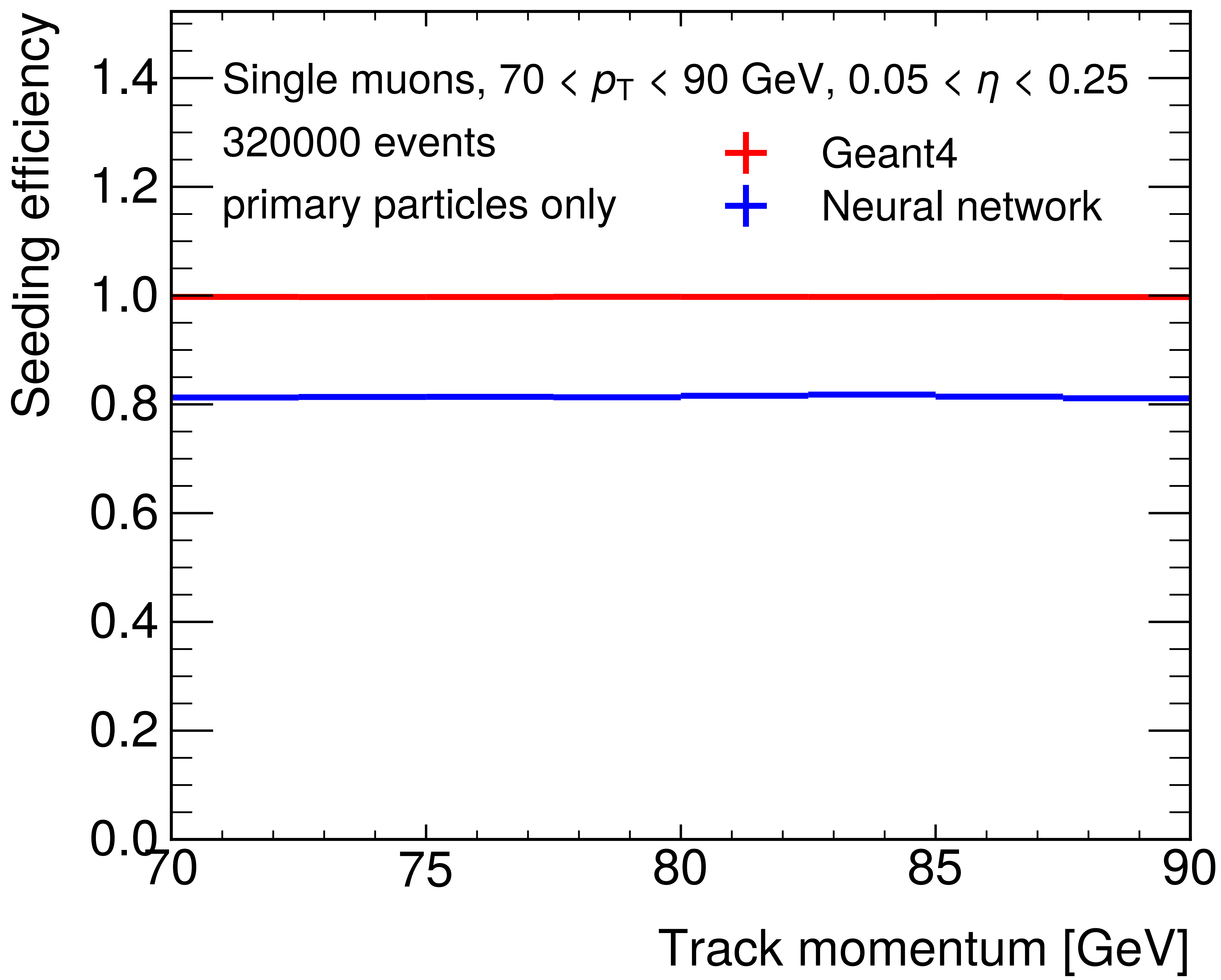
- Good agreement with full simulation.
  - Coordinates fluctuating a bit up to ±10 %.
- Number of hits accurately reproduced.
  - Some difference seen but random numbers included in inference.



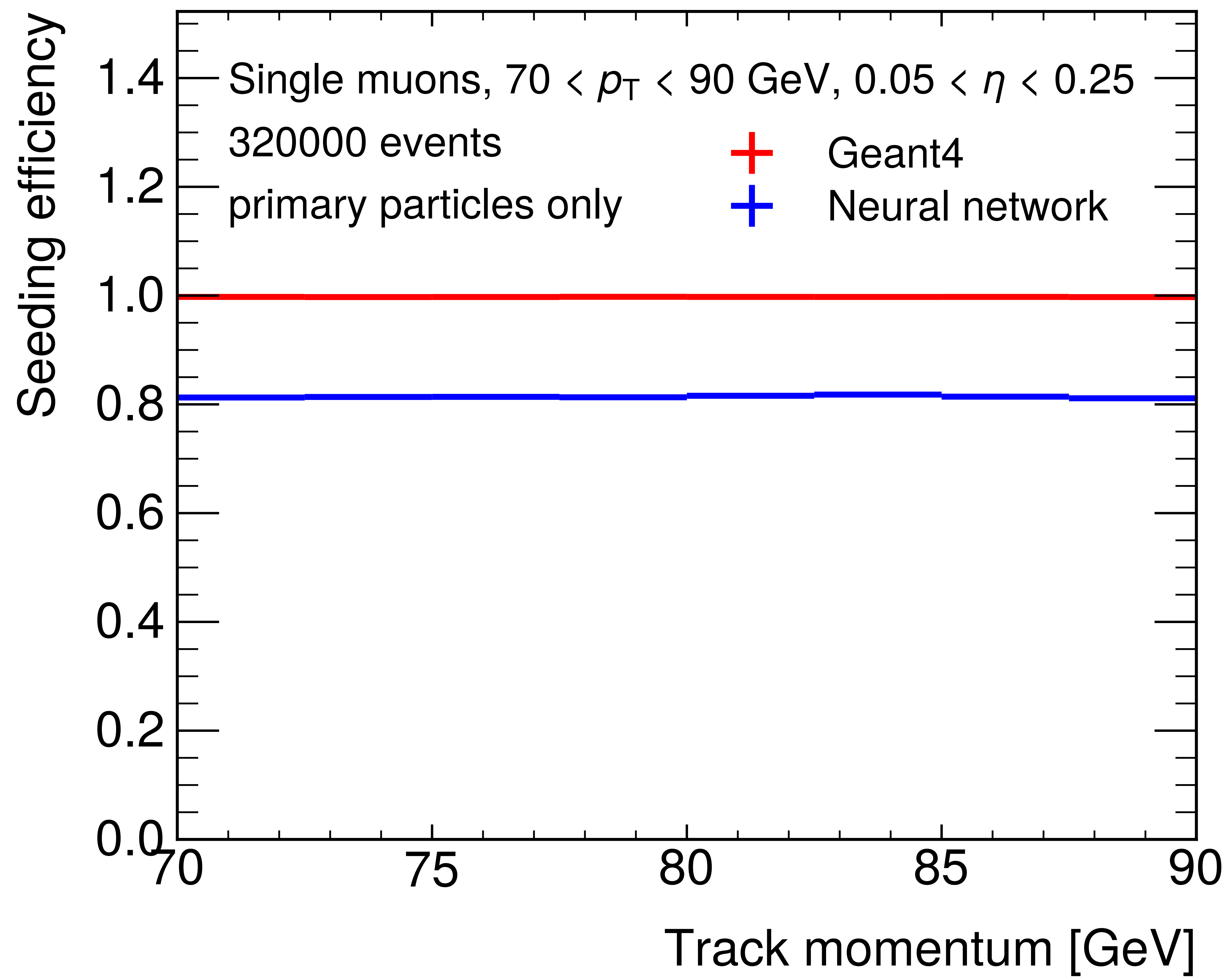
- Global coordinates show good agreement describing complex detector structure.
- Larger deviations in tails of the z-coordinate due to lower statistics.

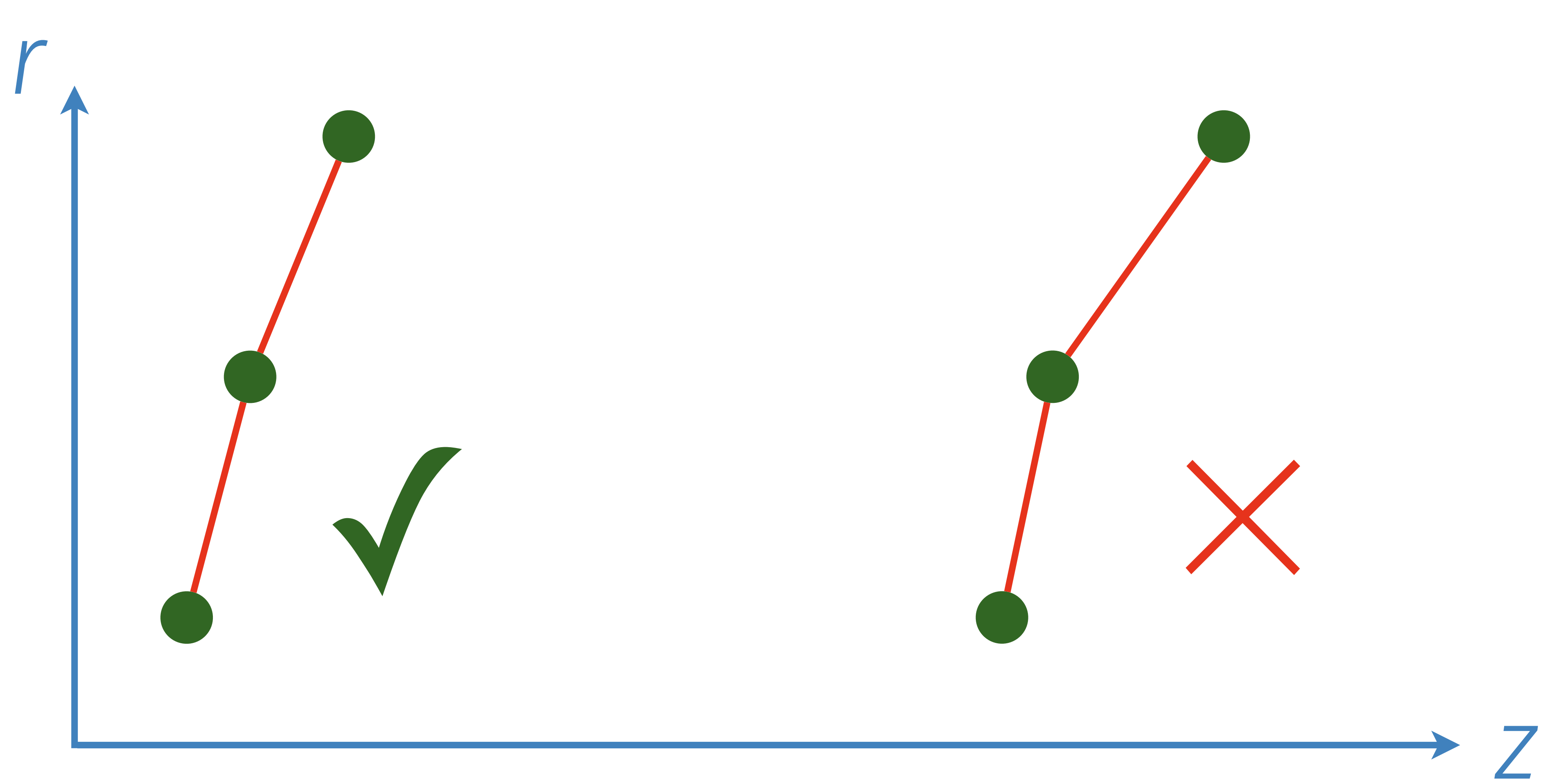
- Each event gets assigned a random integer between 1 and 10000, both at training and inference.
- Coordinates smeared around the true value — generative nature of the model is achieved.





- Evaluating performance using the ACTS (A Common Tracking Software) framework.
  - Default test setup for the Open Data Detector.
- Seeding efficiency only ~82 % compared to 99 % for full simulation.
  - Rounding has no significant effect on the reference sample.
  - Hit displacement from the estimated helix is too large.
    - Threshold defined by the maximum allowed multiple-scattering effect.





- Evaluating performance using the ACTS (A Common Tracking Software) framework.
  - Default test setup for the Open Data Detector.
- Seeding efficiency only ~82 % compared to 99 % for full simulation.
  - Rounding has no significant effect on the reference sample.
  - Hit displacement from the estimated helix is too large.
    - Threshold defined by the maximum allowed multiple-scattering effect.

- Transformers can describe a sequence of physics data very well.
    - But the results are too random at the moment, needs optimisation.
- Training relatively long, but inference is fast.
- Future plans:
    - Optimise the current setup for better tracking performance.
    - Try to describe continuous features with floating point numbers.
    - Try proper generative sampling of a transformer.

I FEEL SLOVENIA

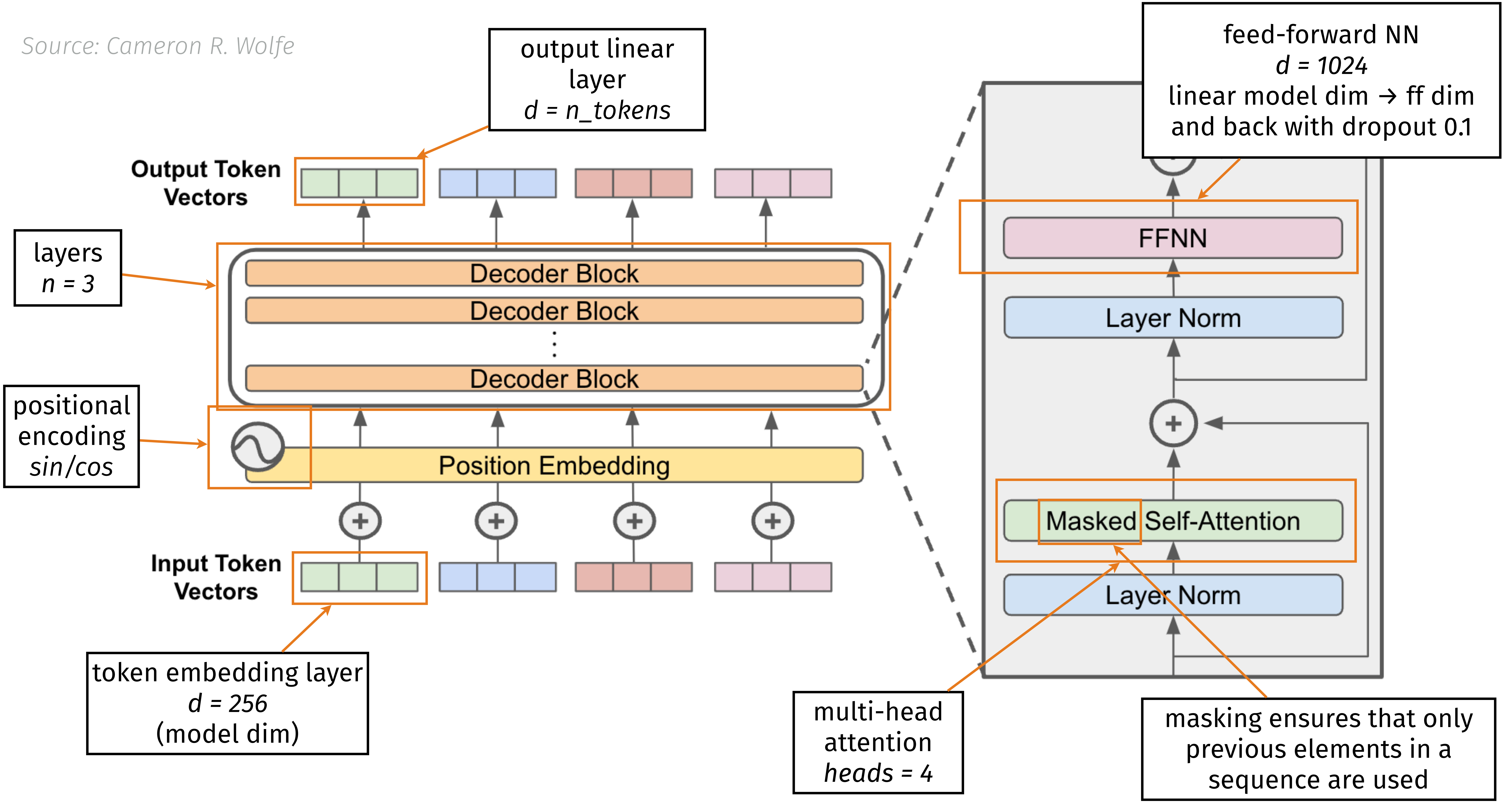
Co-funded by the European Union

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.

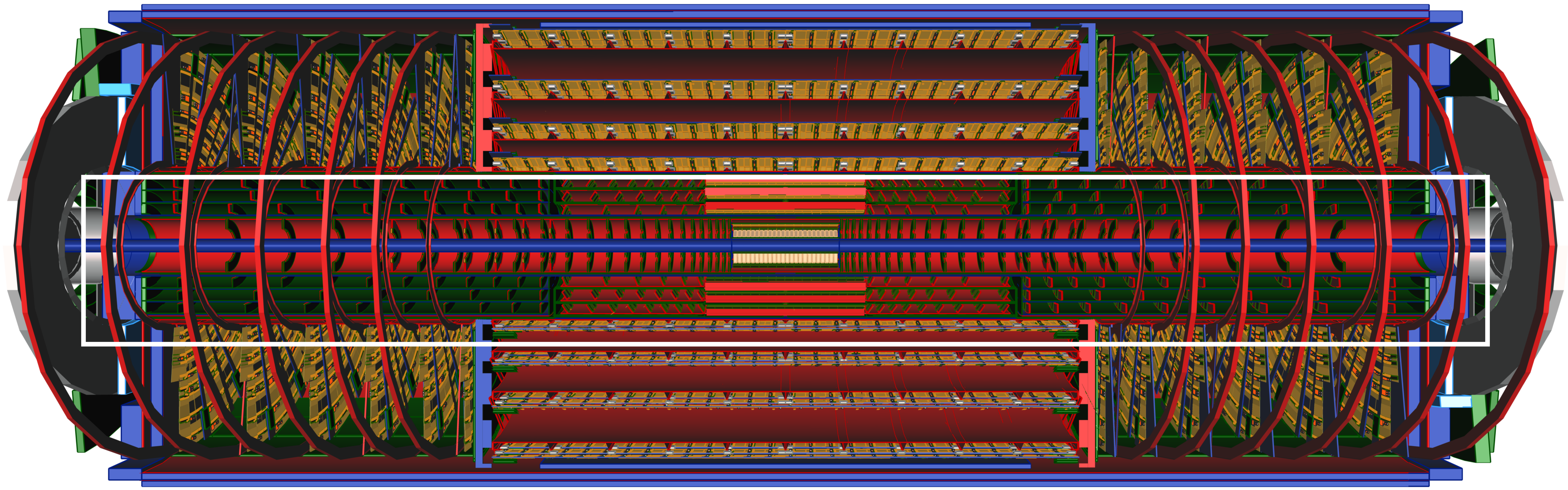
The operation (SMASH project) is co-funded by the Republic of Slovenia and the European Union from the European Regional Development Fund.

Decoder-only Transformer

Source: Cameron R. Wolfe



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