Changes Since Last Generative Meeting

Architecture:

- Flash Attention $\rightarrow$ Flex Attention[1]
  - utilizes sparsity in the attention matrix
  - reduces memory usage
  - speeds up training
  - allows for larger batch sizes
- points attend to points with a distance of up to two layers

Dataset:

- remove cut on time
- relex cut in x and y to $[450, 450]$ mm

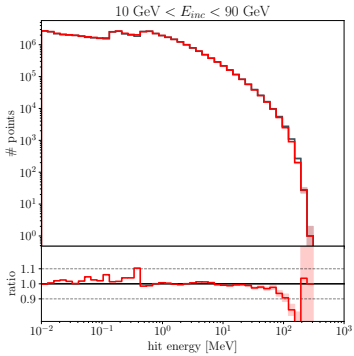
Timing:

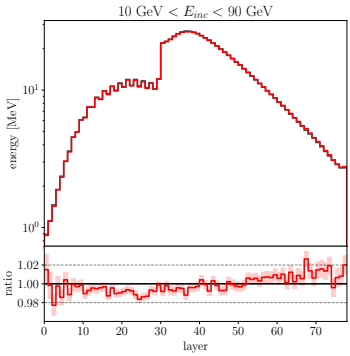
device	NFE	batch size	per shower
A100	200	1	222 ms
A100	200	2	112 ms
A100	200	4	52 ms
A100	200	16	50 ms
A100	200	64	44 ms

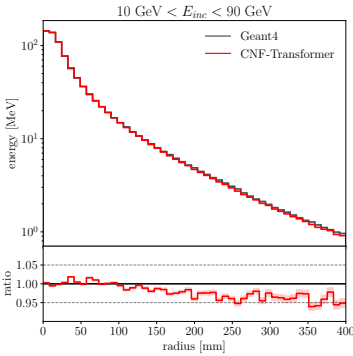
startup time is c.a. 30 s

[1] Juechu Dong et al. *Flex Attention: A Programming Model for Generating Optimized Attention Kernels.* 2024. arXiv: 2412.05496 [cs.LG]

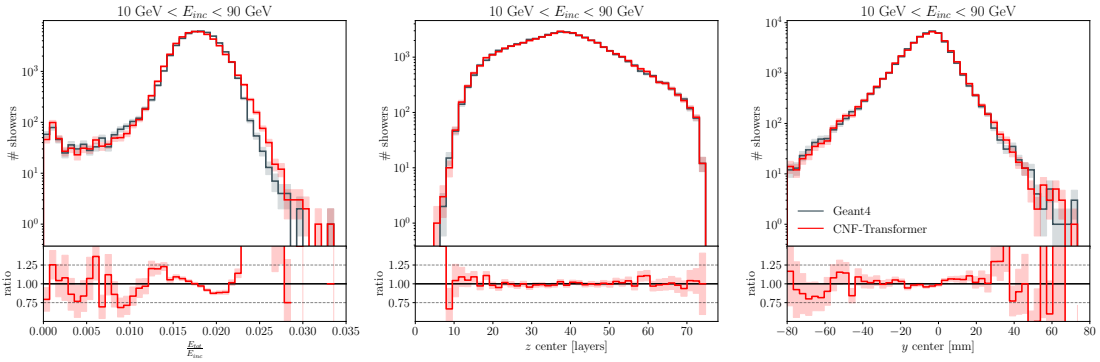
Hit Level Observables



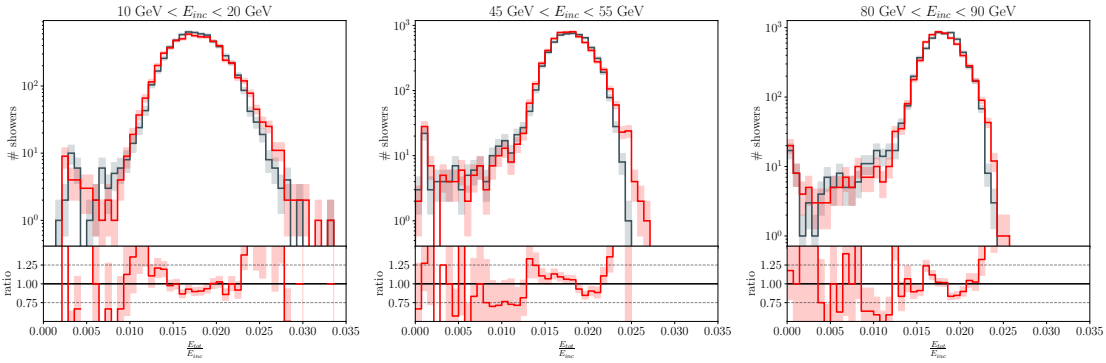




Shower Level Observables



Energy Slices



2D Histograms