



Fast Simulation of Calorimetry showers (FastSim)

Mikołaj Piórczyński

FastSim Summer Students Meeting, 12.07.2024

▶ Who am I?



- Live in Warsaw, Poland.
- Defended bachelor's thesis with honors at Warsaw University of Technology. Thesis title: 'Efficient Inference in Transformer Models with Dense to Dynamic-k Mixture-of-Experts Model Conversion'.
- Worked 1.5 years as an intern in the Machine Translation Team at Samsung R&D Institute Poland and 0.5 years as an MLE intern at AI Clearing (AI-powered construction progress tracking based on drone-captured data).
- Co-organizing ML in PL Conference, one of the biggest ML-oriented conferences in Poland.
- Visited CERN during a high-school trip a few years ago.
- In free time: playing the violin and running

SAMSUNG



Faculty of Mathematics and Information Science

WARSAW UNIVERSITY OF TECHNOLOGY

How to make diffusion models faster?*

***while maintaining a high quality of samples**

FLASHATTENTION: Fast and Memory-Efficient Exact Attention with IO-Awareness

Tri Dao[†], Daniel Y. Fu[†], Stefano Ermon[†], Atri Rudra[‡], and Christopher Ré[†]

[†]Department of Computer Science, Stanford University

[‡]Department of Computer Science and Engineering, University at Buffalo, SUNY
{trid,danfu}@cs.stanford.edu, ermon@stanford.edu, atri@buffalo.edu, chrisrmre@cs.stanford.edu

June 24, 2022

Abstract

Transformers are slow and memory-hungry on long sequences, since the cost of self-attention is quadratic in sequence length. Approximate attention methods aim to address this problem by trading off model quality to reduce the computational cost. We argue that a missing principle is making attention IO-aware—accounting for reads and writes between levels of GPU memory. We propose an IO-aware exact attention algorithm that uses tiling to reduce the number of GPU high bandwidth memory (HBM) and GPU on-chip SRAM. The cost of FLASHATTENTION, showing that it requires fewer HBM accesses than existing methods. We also extend FLASHATTENTION to be an approximate attention algorithm that is faster than any existing approximate attention algorithm. FLASHATTENTION trains Transformers faster than existing baselines: 15% faster on BERT-large (seq. length 512) compared to the MLPerf 1.1 training suite, 2.4x speedup on long-range arena (seq. length 1K), and 2.4x speedup on long-range arena (seq. length 1K).

Transformers, y
ument cl
performa
63.1% a

Structural Pruning for Diffusion Models

Gongfan Fang Xinyin Ma Xinchao Wang*
National University of Singapore

{gongfan@nus.edu.sg, maxinyin@nus.edu.sg, xinchao@nus.edu.sg}

Abstract

Generative modeling has recently undergone remarkable advancements, primarily propelled by the transformative implications of Diffusion Probabilistic Models (DPMs). The impressive capability of these models, however, often entails significant computational overhead during both training and inference. To tackle this challenge, we present *Diff-Pruning*, an efficient compression method tailored for learning lightweight diffusion models from pre-existing ones, without the need for extensive re-training. The essence of Diff-Pruning is encapsulated in a Taylor expansion over *pruned timesteps*, a process that disregards non-contributory diffusion steps and ensembles informative gradients to identify important weights. Our empirical assessment, undertaken across several datasets highlights two primary benefits of our proposed method: 1) *Efficiency*: it enables approximately a 50% reduction in FLOPs at a mere 10% to 20% of the original training expenditure; 2) *Consistency*: the pruned diffusion models inherently preserve generative behavior congruent with their pre-trained models. Code is available at <https://github.com/VainF/Diff-Pruning>.

31 May 2023

Diffusion models have significantly advanced the fields of image, audio, and video generation, but they depend on an iterative sampling process that causes slow generation. To overcome this limitation, we propose *consistency models*, a new family of models that generate high quality samples by directly mapping noise to data. They support fast one-step generation by design, while still allowing multistep sampling to trade compute for sample quality. They also support zero-shot data-to-data generation, which is not possible with standard diffusion models.

Architecture
e grown
the self-
n impor
odels ad

GI 5 Oct 2022

Consistency Models

Pratulla Dhariwal¹ Mark Chen¹ Ilya Sutskever¹

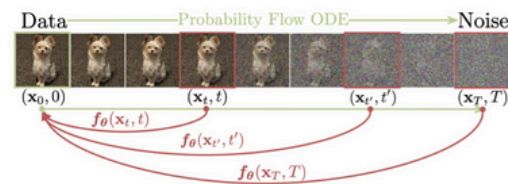


Figure 1: Given a Probability Flow (PF) ODE that smoothly converts data to noise, we learn to map any point (e.g., x_t , $x_{t'}$, and x_T) on the ODE trajectory to its origin (e.g., x_0) for generative modeling. Models of these mappings are called consistency models. They support multistep sampling by design, while still allowing fast one-step generation by design.

Dec 2023

Token Merging for Fast Stable Diffusion

Daniel Bolya Judy Hoffman
Georgia Tech
{dbolya, judy}@gatech.edu

Abstract

The landscape of image generation has been forever changed by open vocabulary diffusion models. However, at their core these models use transformers, which makes generation slow. Better implementations to increase the throughput of these transformers have emerged, but they still evaluate the entire model. In this paper, we instead

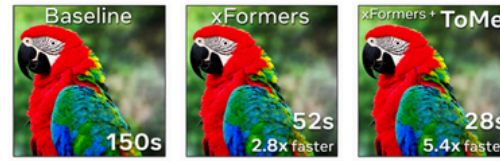


Figure 1. Token Merging for Stable Diffusion. When applied to the standard Stable Diffusion v1.5 model, our ToMe method achieves a 5.4x speedup over the baseline.

Published as a conference paper at ICLR 2021

DENOISING DIFFUSION IMPLICIT MODELS

Jiaming Song, Chenlin Meng & Stefano Ermon
Stanford University
{tsong, chenlin, ermon}@cs.stanford.edu

ABSTRACT

Denoising diffusion probabilistic models (DDPMs) have achieved high quality image generation without adversarial training, yet they require simulating a Markov chain for many steps in order to produce a sample. To accelerate sampling, we present denoising diffusion implicit models (DDIMs), a more efficient class of iterative implicit probabilistic models with the same training procedure as DDPMs. In DDPMs, the generative process is defined as the reverse of a particular Markovian diffusion process. We generalize DDPMs via a class of non-Markovian diffusion processes that lead to the same training objective. These non-Markovian processes can correspond to generative processes that are deterministic, giving rise to implicit models that produce high quality samples much faster. We empirically demonstrate that DDIMs can produce high quality samples 10x to 50x faster in terms of wall-clock time compared to DDPMs, allow us to trade off computation for image interpolation directly with very low error.

produce high quality samples in many terms of image generation, generatively exhibits higher sample quality than DDPMs (Kingma & Welling, 2013), autoregressive flows (Rezende & Mohamed, 2015); advances in optimization and architectures (Ho et al., 2017; Karras et al., 2018; Brock et al., 2018).

14), such as denoising diffusion probabilistic models (DDPMs) (Ho et al., 2017).

Scalable High-Resolution Pixel-Space Image Synthesis with Hourglass Diffusion Transformers

Benjamin Crowson^{*1} Stefan Andreas Baumann^{*2} Alex Birch^{*3} Tanishq Mathew Abraham¹
Daniel Z. Kaplan⁴ Enrico Shippole⁵

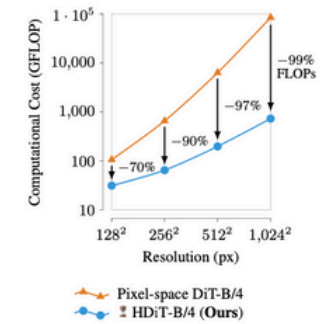


Figure 2: Scaling of computational cost w.r.t. target resolution of our HDiT-B/4 model vs. DiT-B/4 (Peebles & Xie, 2023a), both in pixel space. At megapixel resolutions, our model incurs less than 1% of the computational cost compared to the standard diffusion transformer DiT at a comparable size.

Examples generated directly in RGB pixel space using our HDiT-B/4 model at FFHQ-1024² and ImageNet-256².

Abstract

We present the Hourglass Diffusion Transformer (HDiT), an image generative model that exhibits competitive performance at high-resolution (e.g. 1024 x 1024) directly in pixel space. Building on the Transformer architecture, which is known to scale to billions of parameters, it bridges the gap between the efficiency of convolutional U-Nets and the scalability of Transformers.

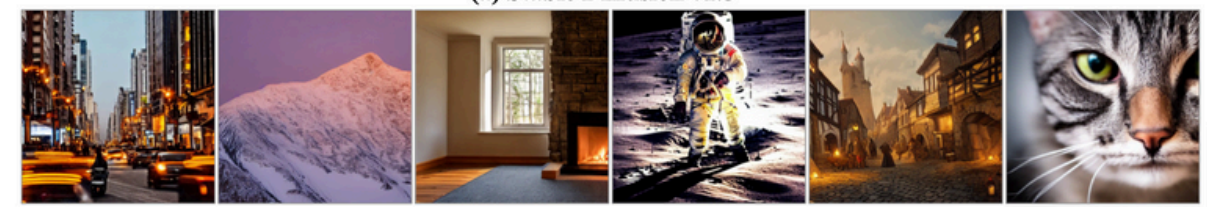
HDiT trains successfully without typical high-resolution training techniques such as multi-scale architectures, latent autoencoders or self-conditioning. We demonstrate that HDiT performs competitively with existing models on ImageNet 256², and sets a new state-of-the-art for diffusion models on FFHQ-1024².

Code and additional results are available on the project page: crowsonkb.github.io/hourglass-diffusion-transformers.

DeepCache: Accelerating Diffusion Models for Free

Xinyin Ma Gongfan Fang Xinchao Wang*
National University of Singapore
{maxinyin, gongfan}@nus.edu.sg, xinchao@nus.edu.sg

(a) Stable Diffusion v1.5



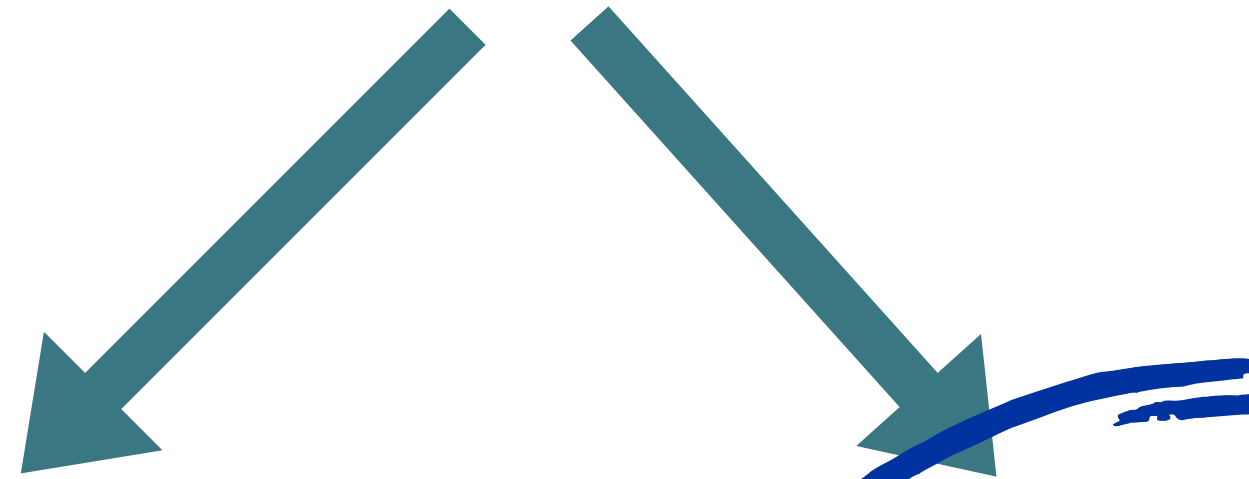
speeding-up diffusion models



model architecture

diffusion process

speeding-up diffusion models



model architecture

diffusion process

Project outline

1. DDIM sampling (<http://arxiv.org/abs/2010.02502>)
2. ~~Progressive Distillation (<https://arxiv.org/abs/2202.00512>)~~
3. EDM (<https://arxiv.org/abs/2206.00364>) + ODE solvers (Heun's, [DPM-Solver++](https://arxiv.org/abs/2211.01095) (<https://arxiv.org/abs/2211.01095>))
4. ~~Optional: Consistency Distillation (<https://arxiv.org/abs/2303.01469>)~~

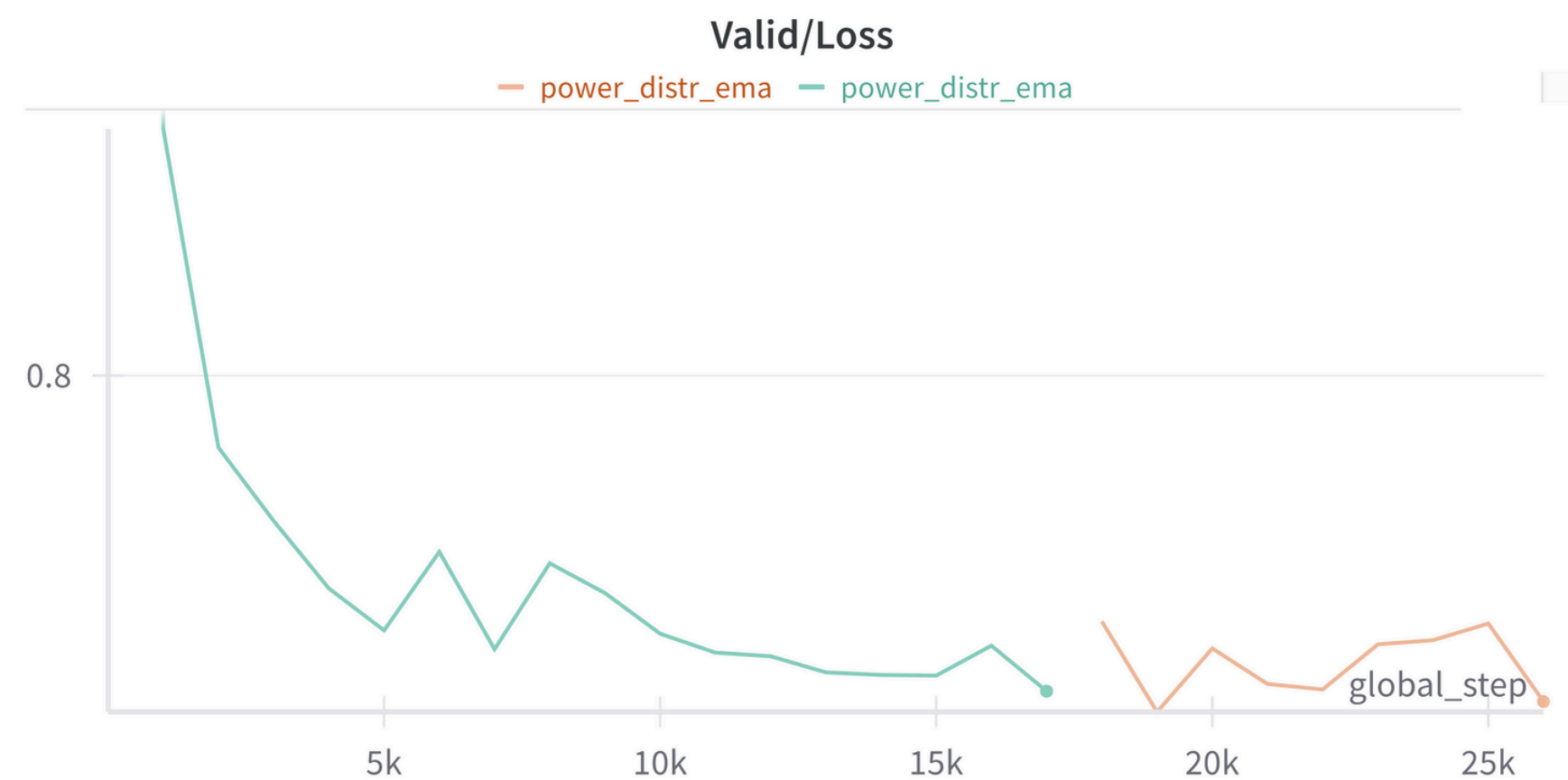
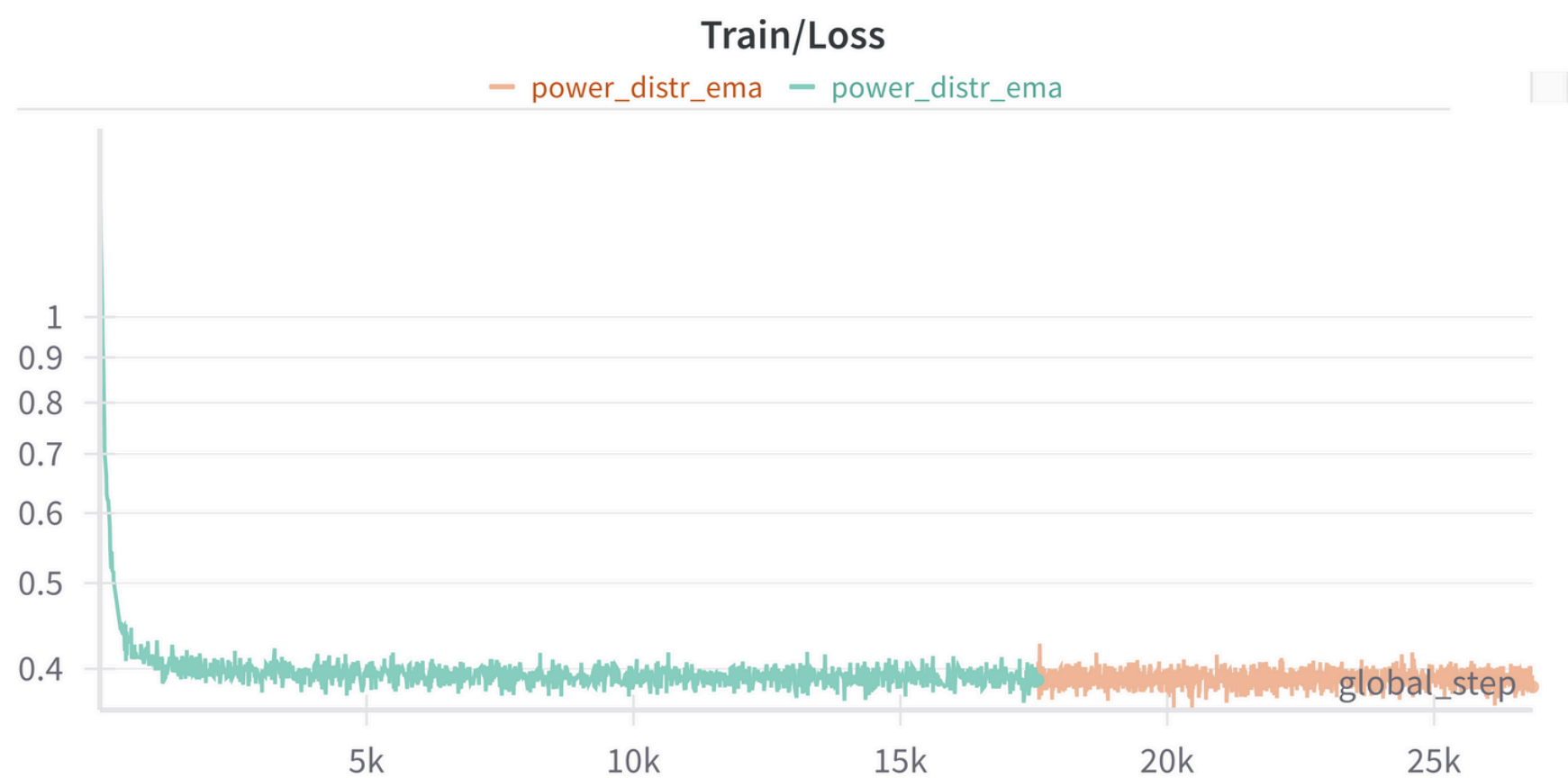


Code: <https://gitlab.cern.ch/mpiorczy/diffusion4fastsim>

DDIM analysis: <https://quiver-cornucopia-fdc.notion.site/DDIM-sampling-b35f60d08b5f4b61b2b075d32bfb97d8>

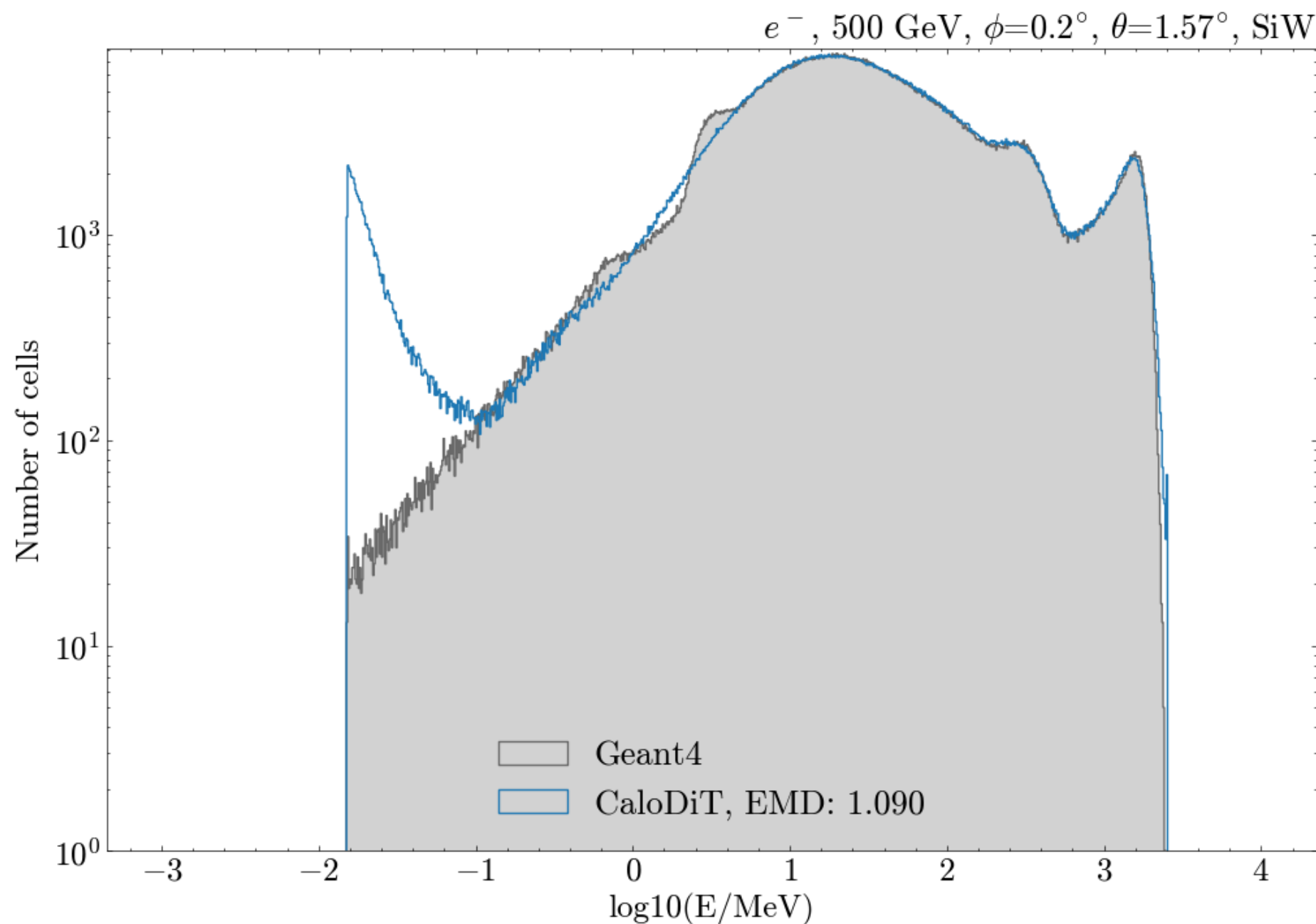


Updates: EDM w/ power energy distribution

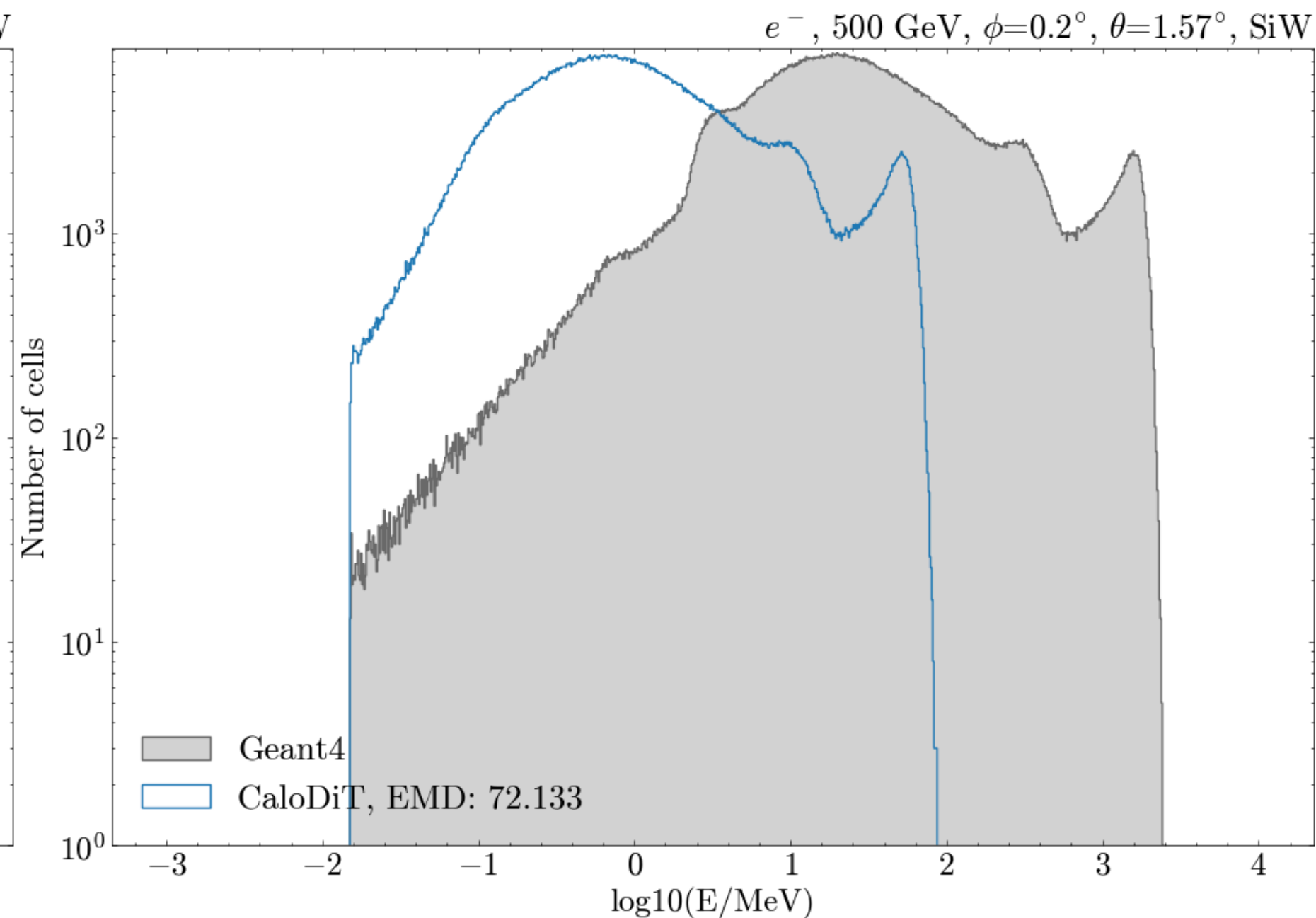




Updates: EDM w/ uniform energy distribution



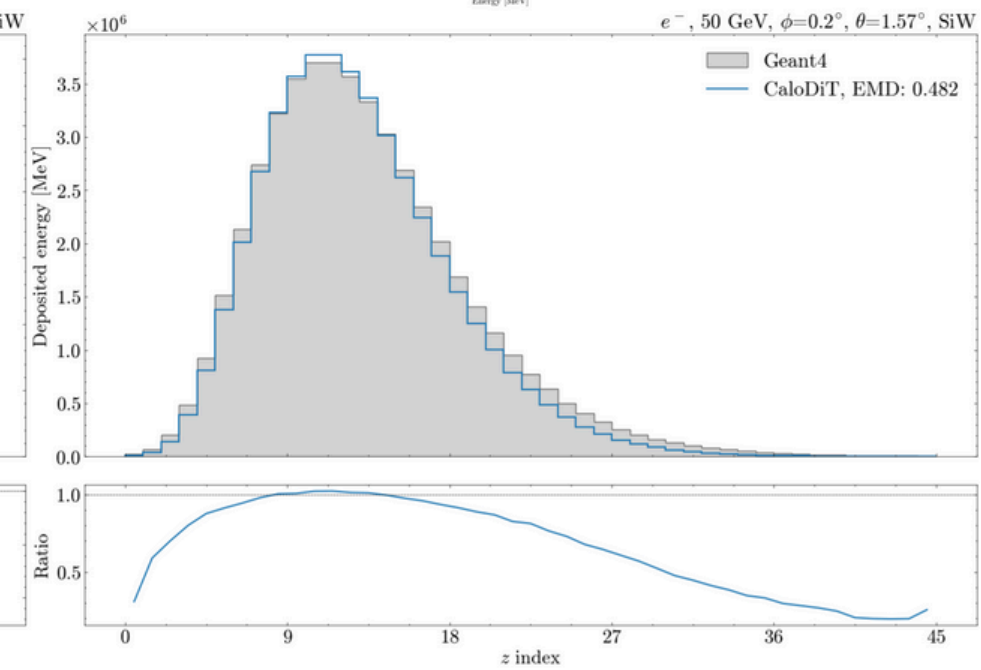
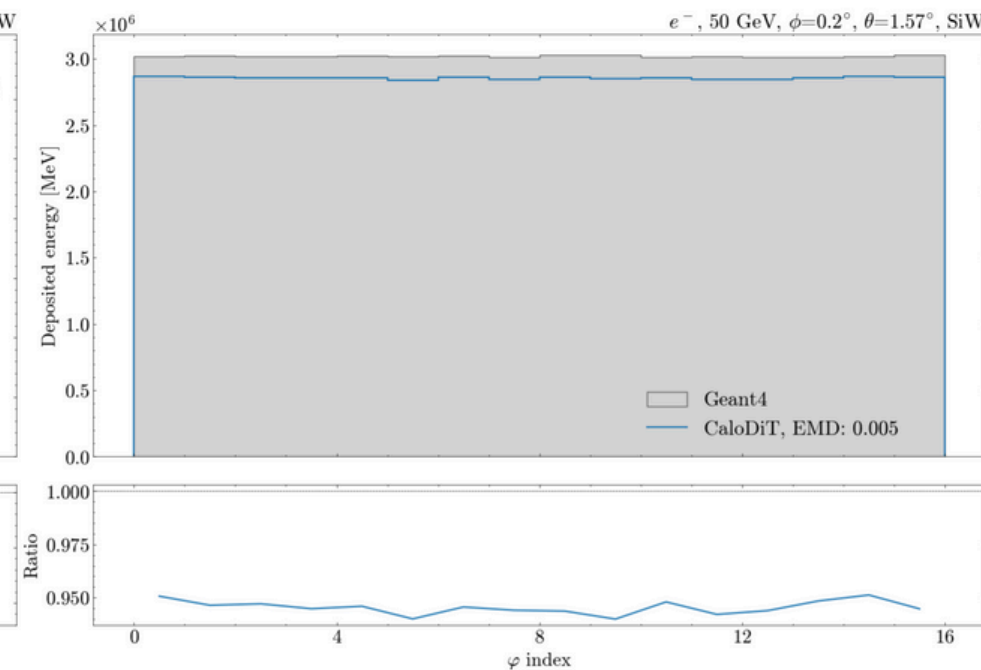
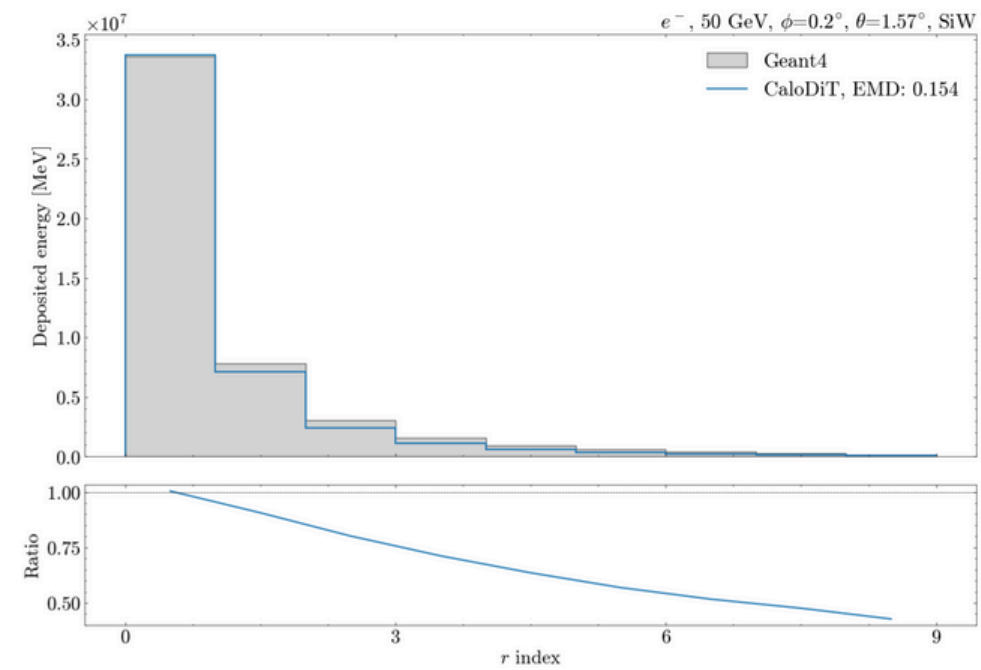
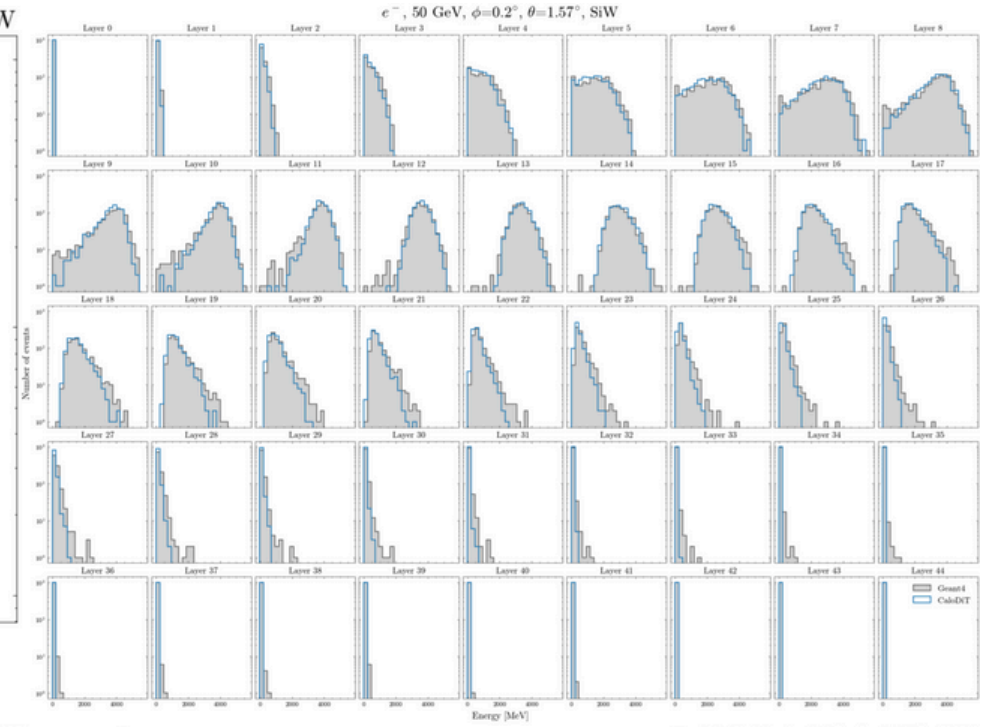
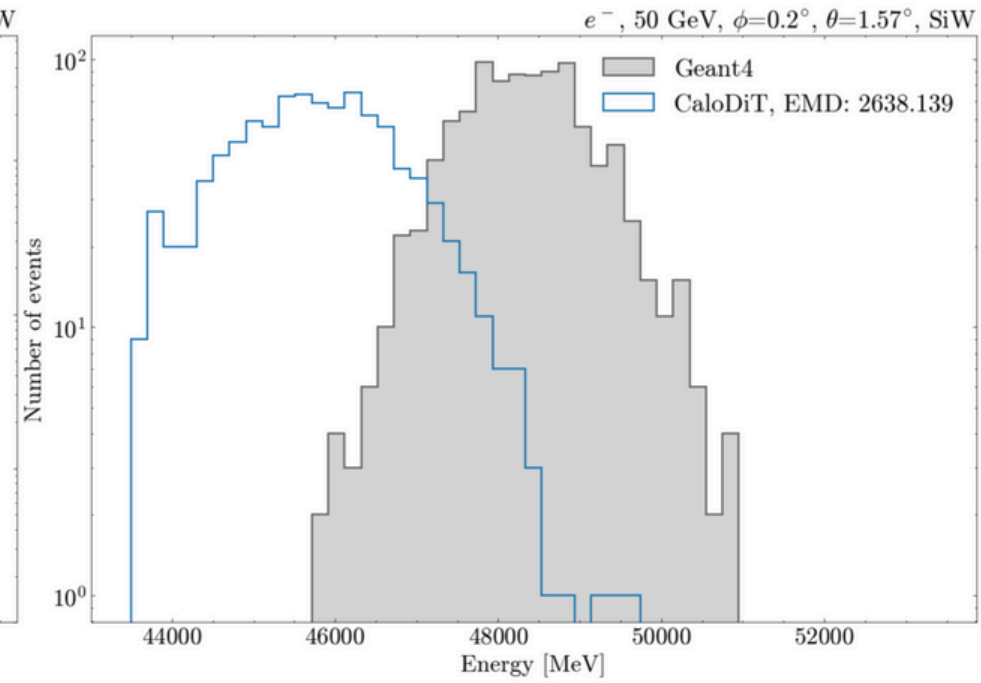
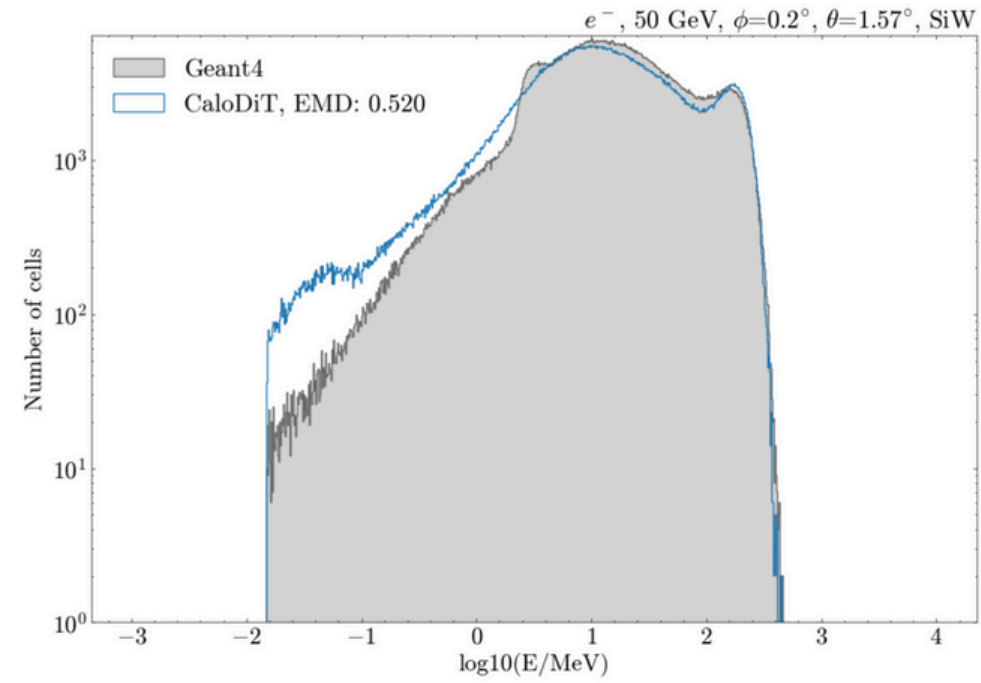
power spectrum



uniform spectrum

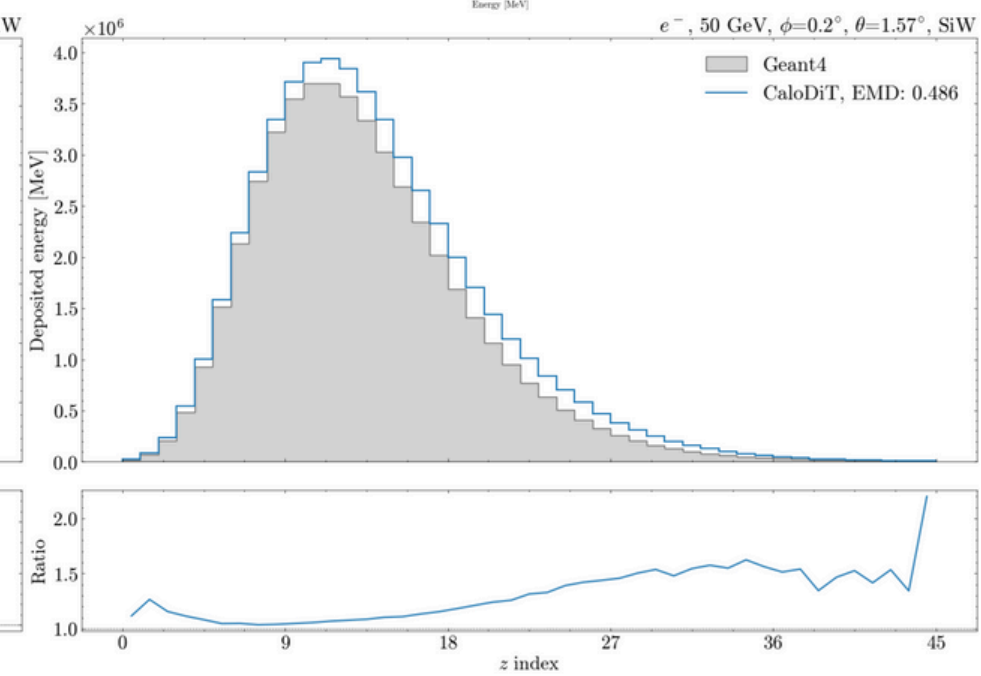
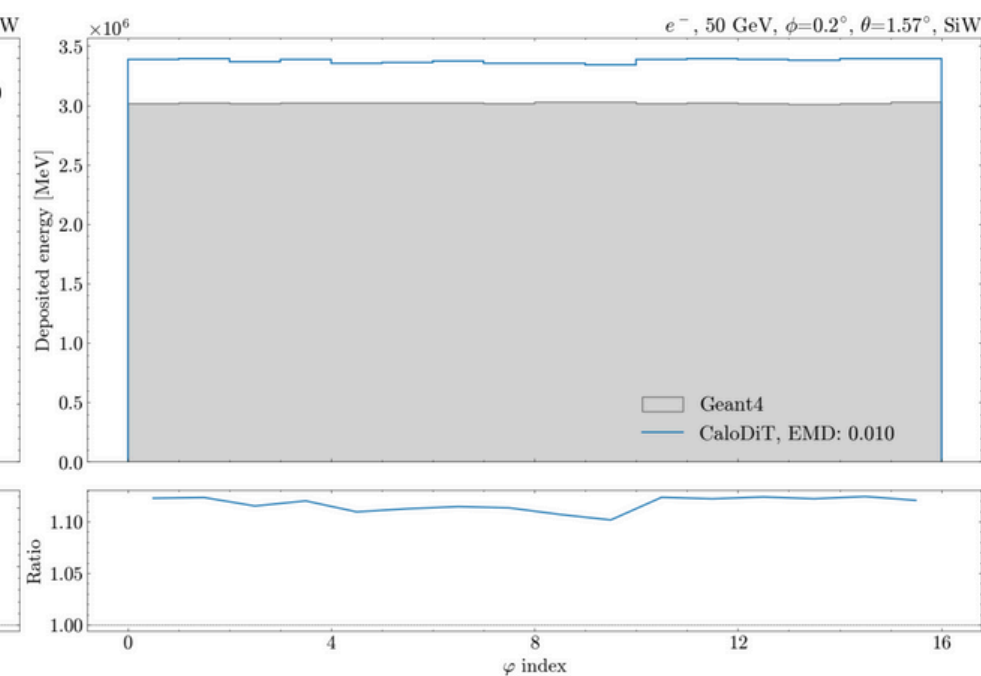
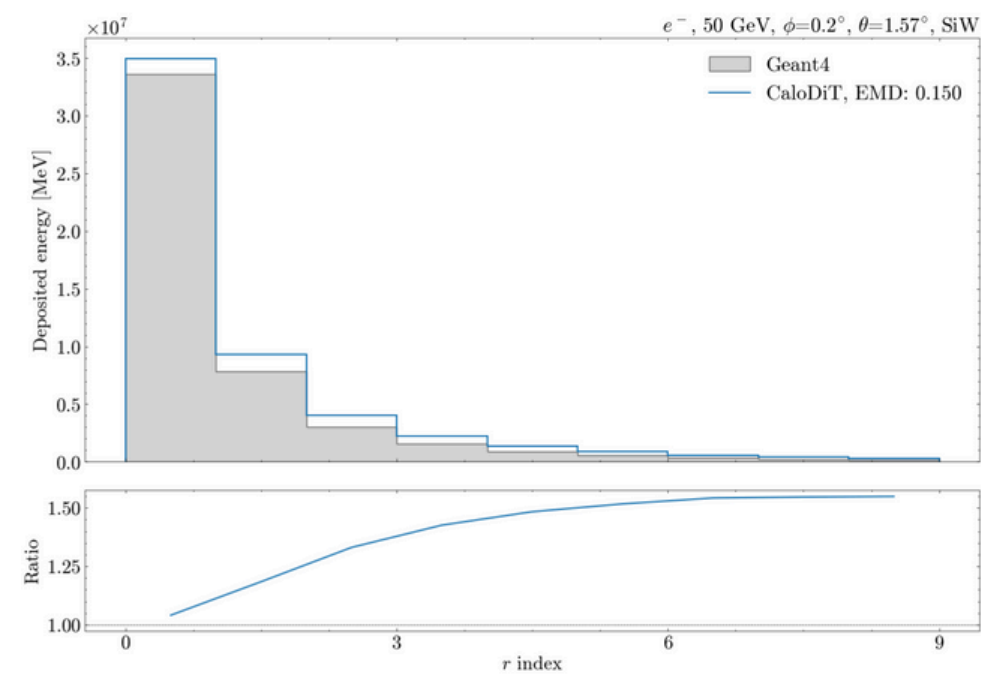
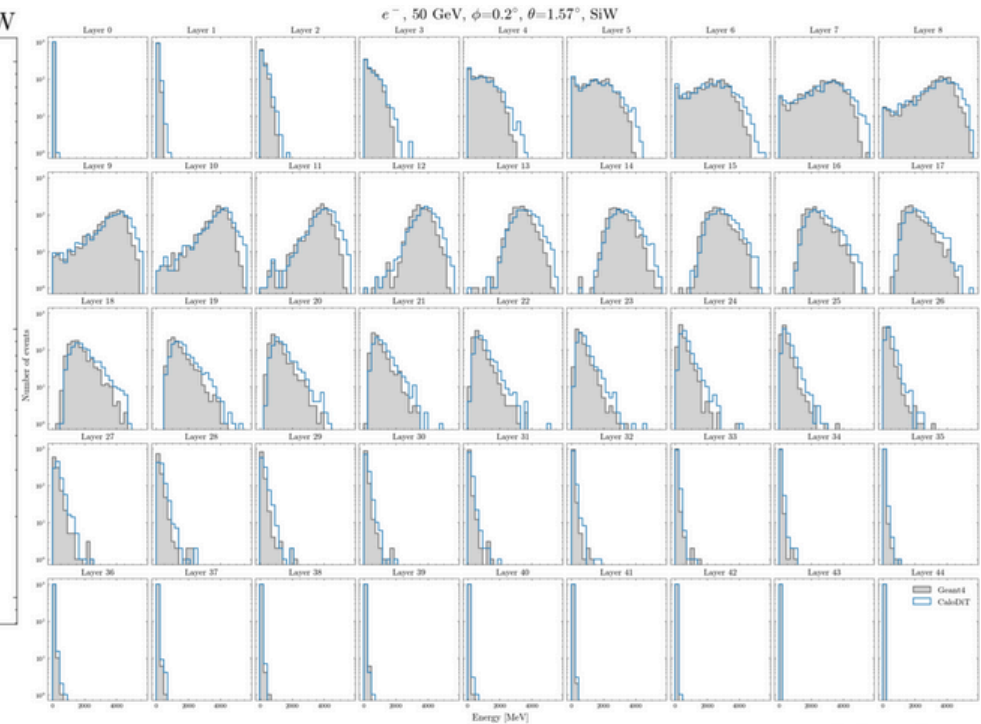
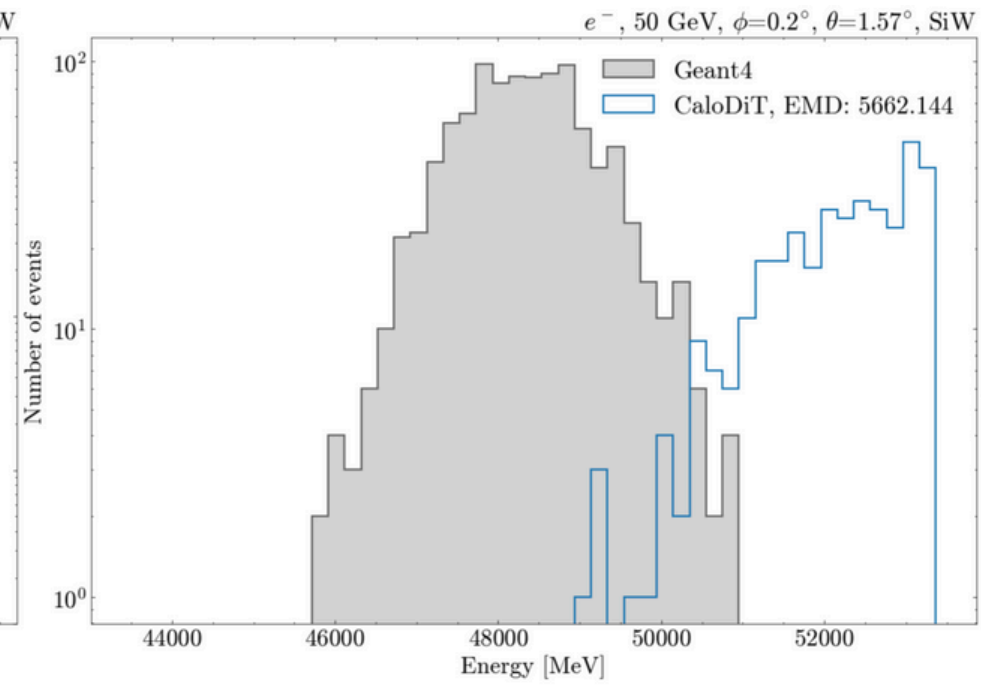
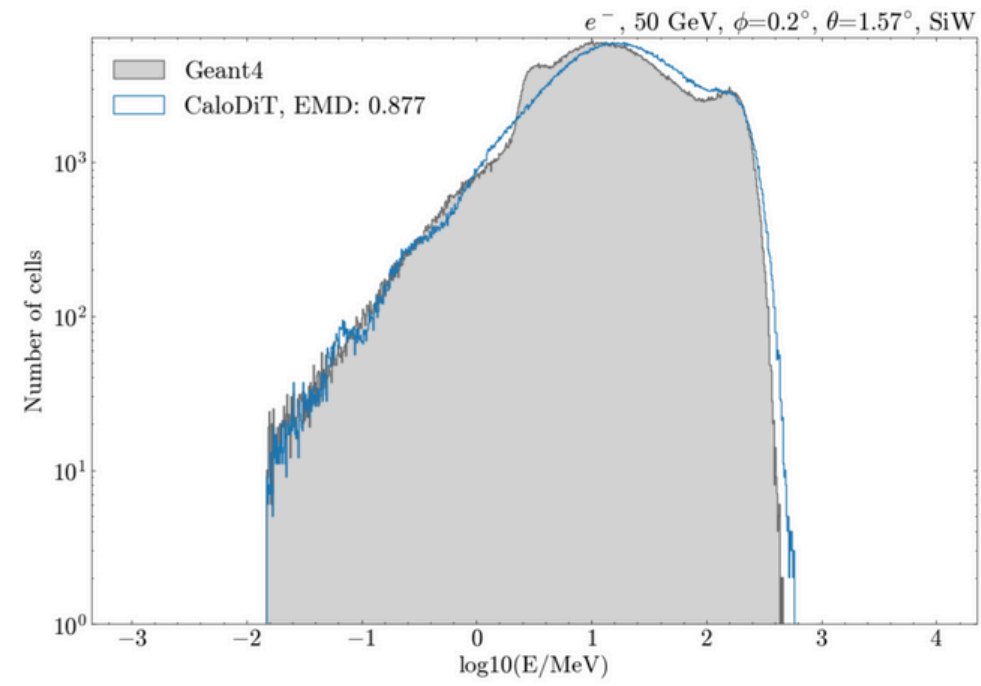


Updates: Euler, 32 steps

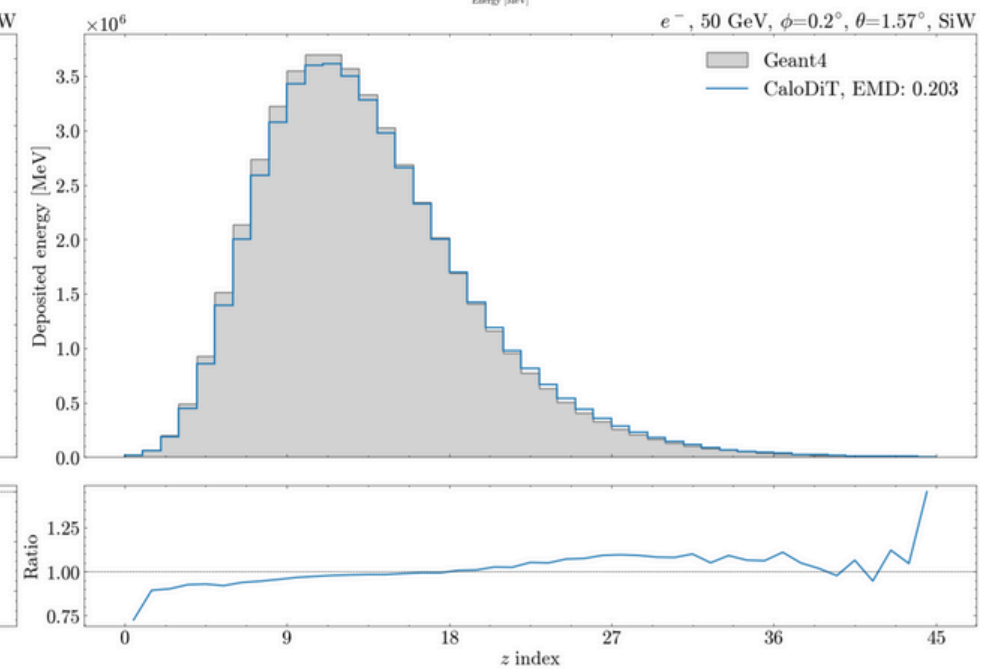
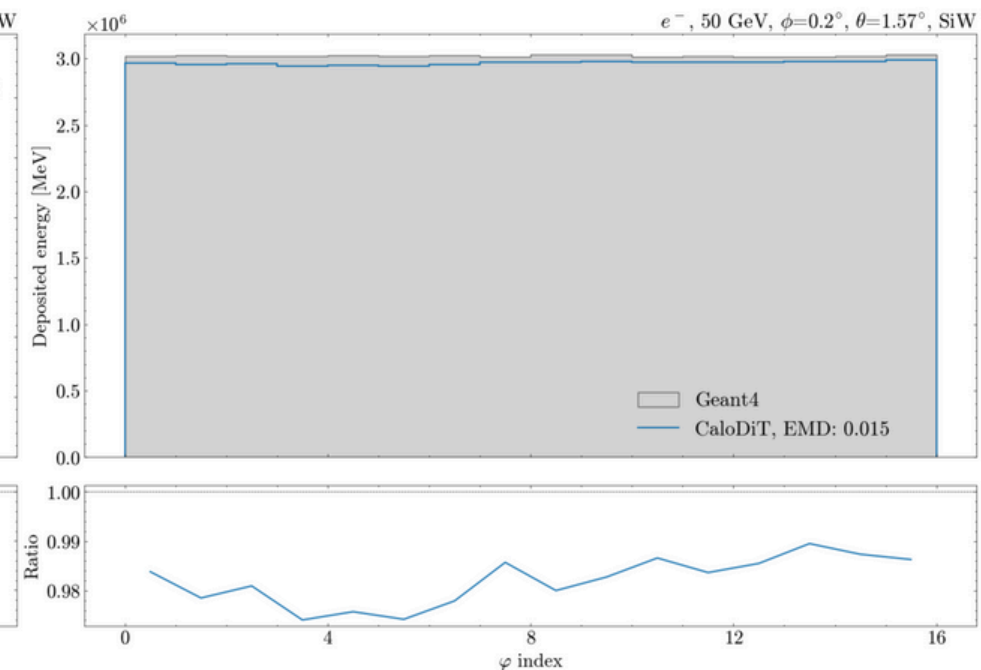
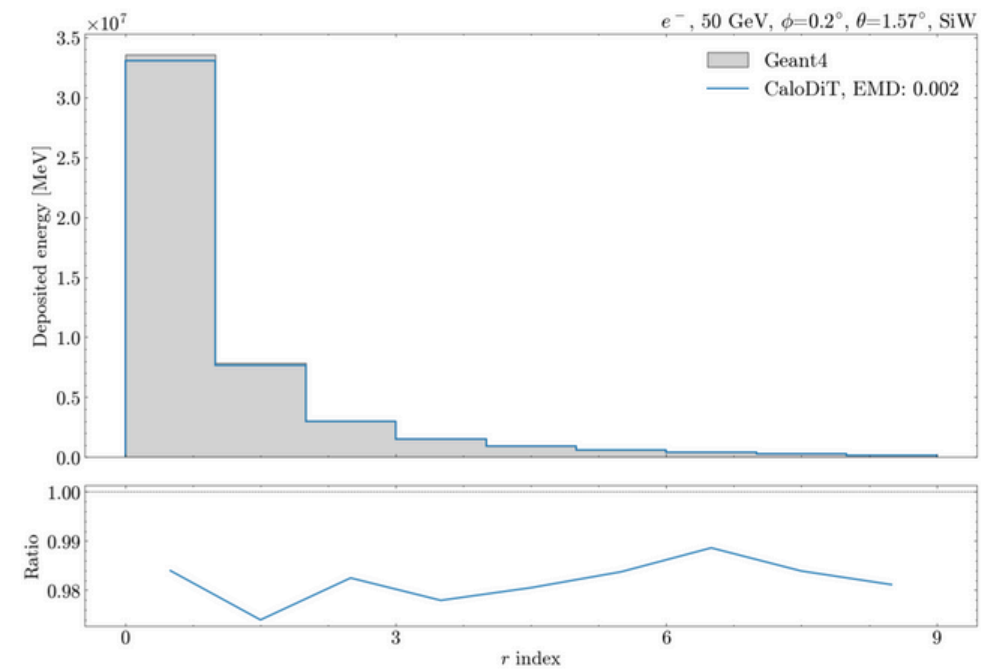
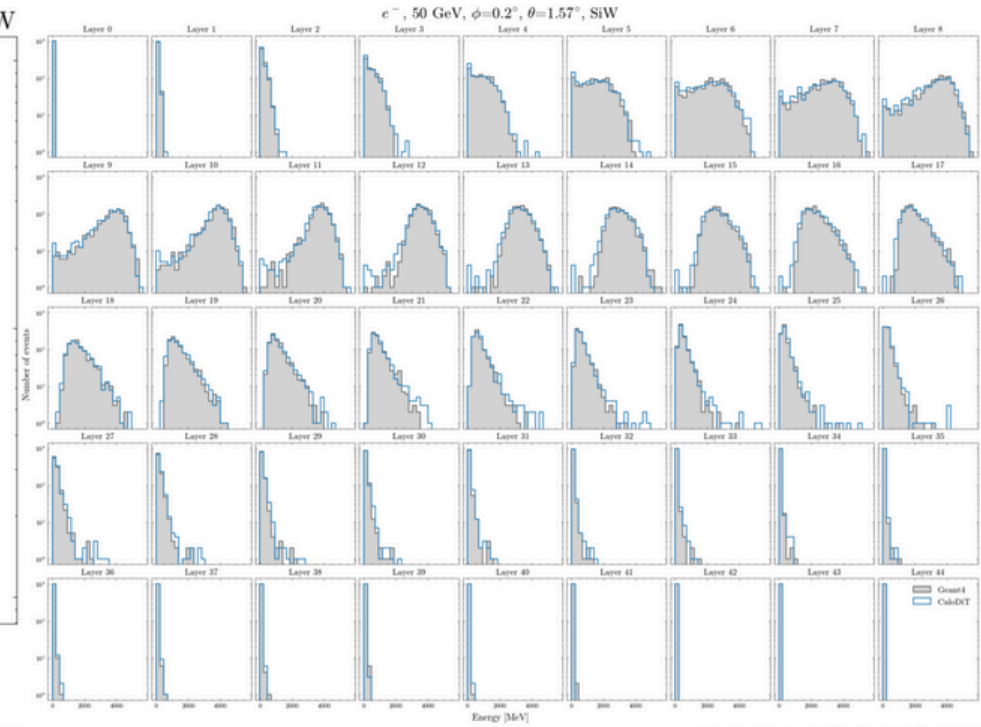
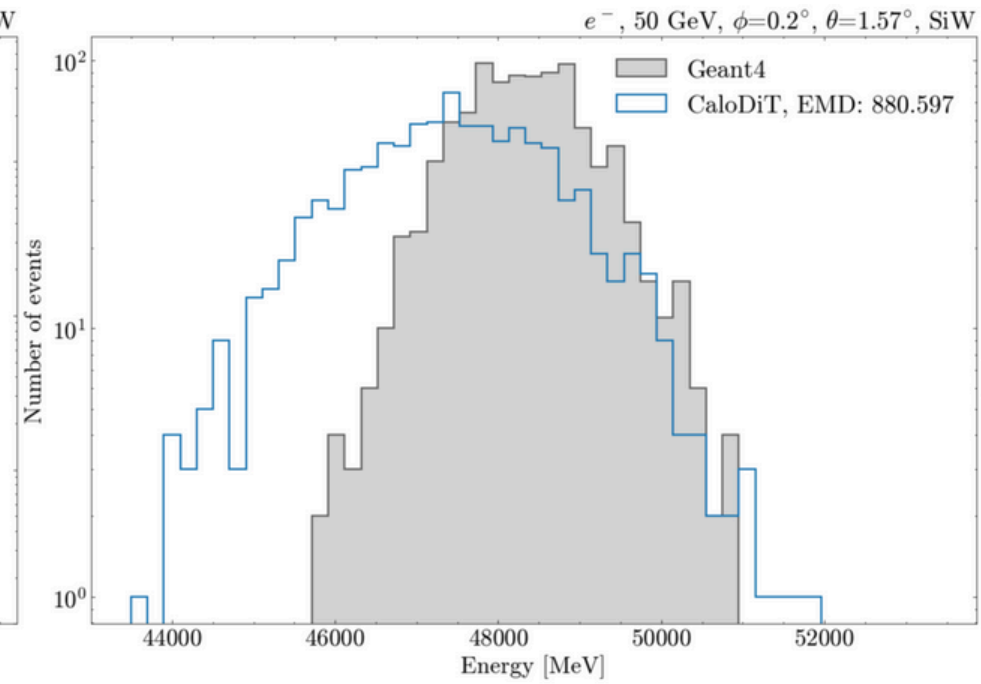
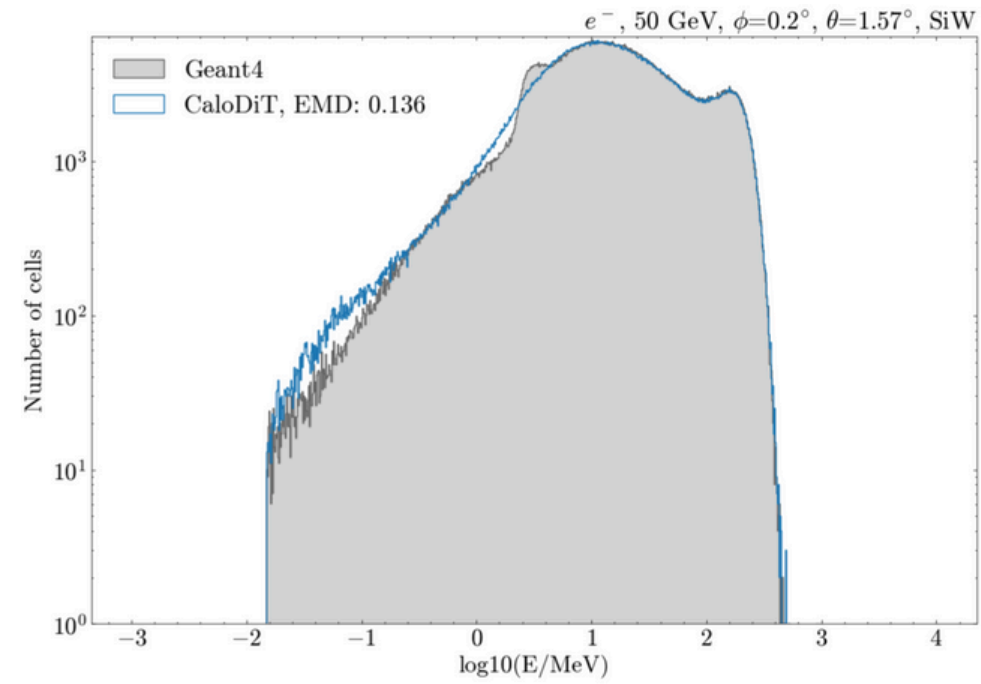




Updates: Heun, 31 steps

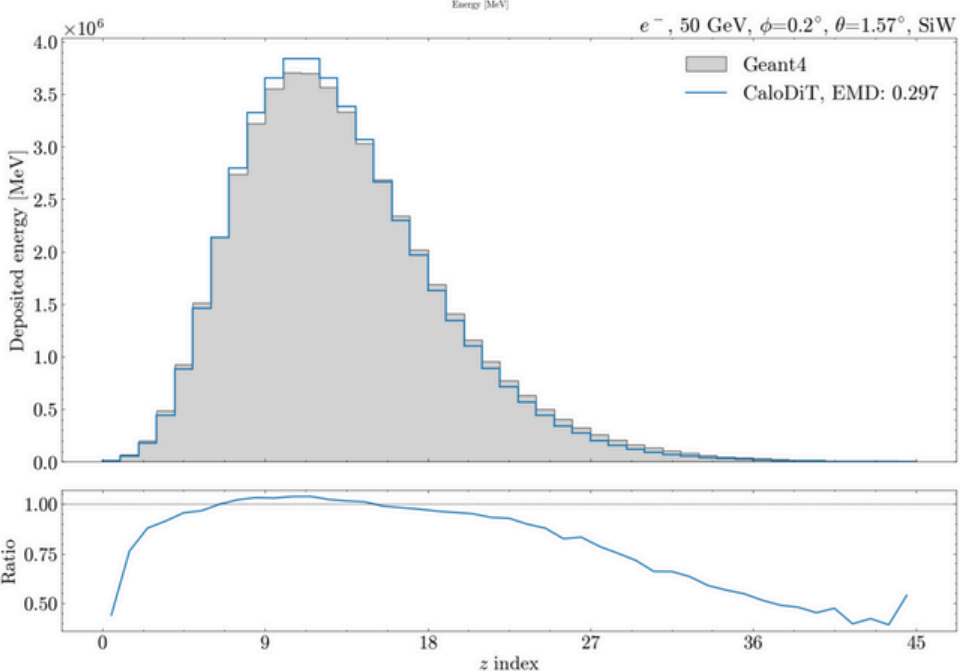
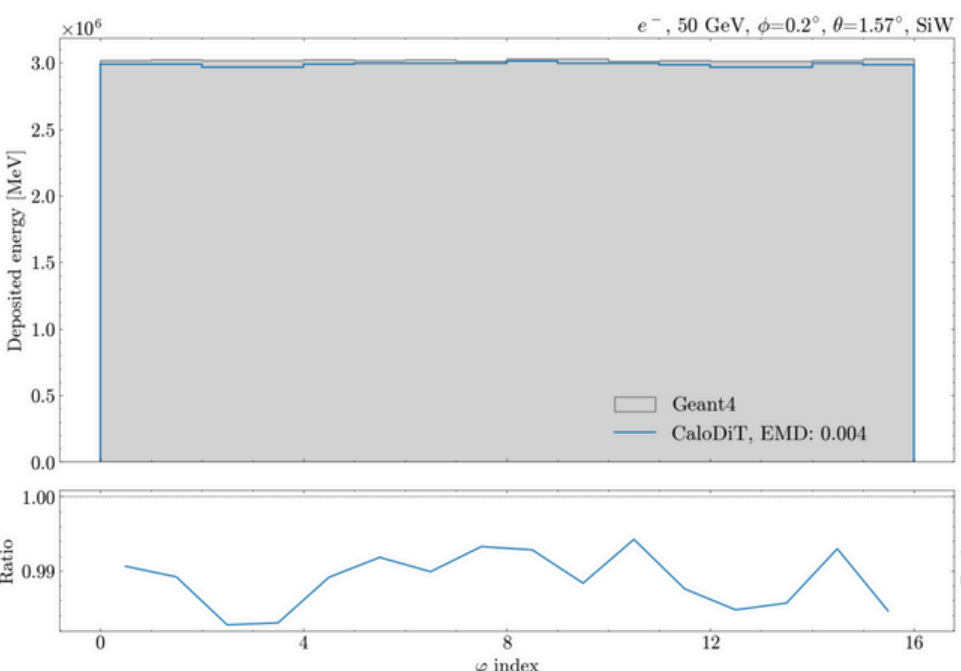
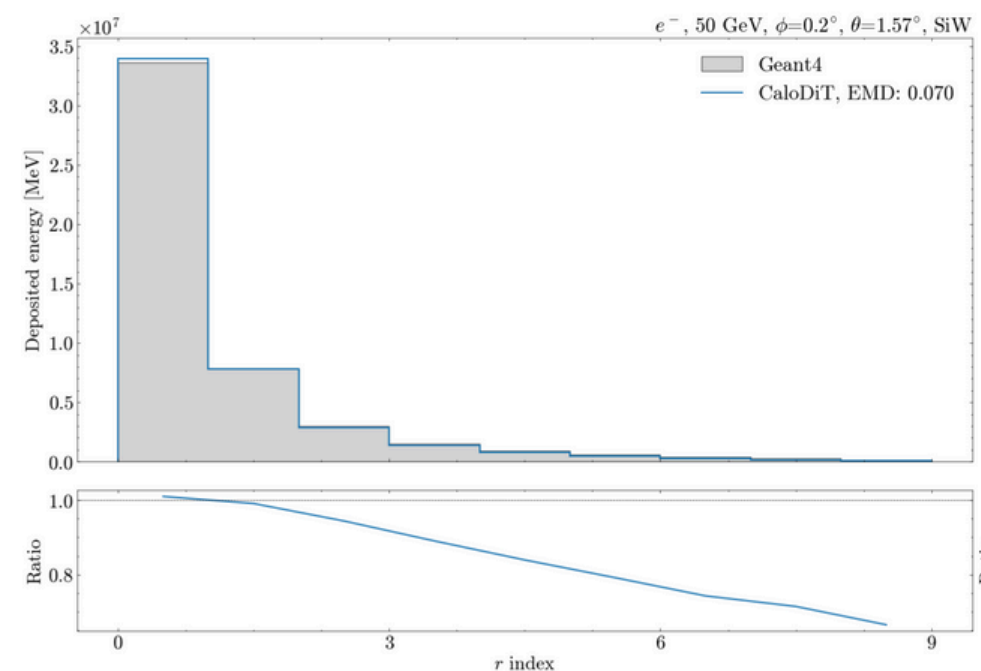
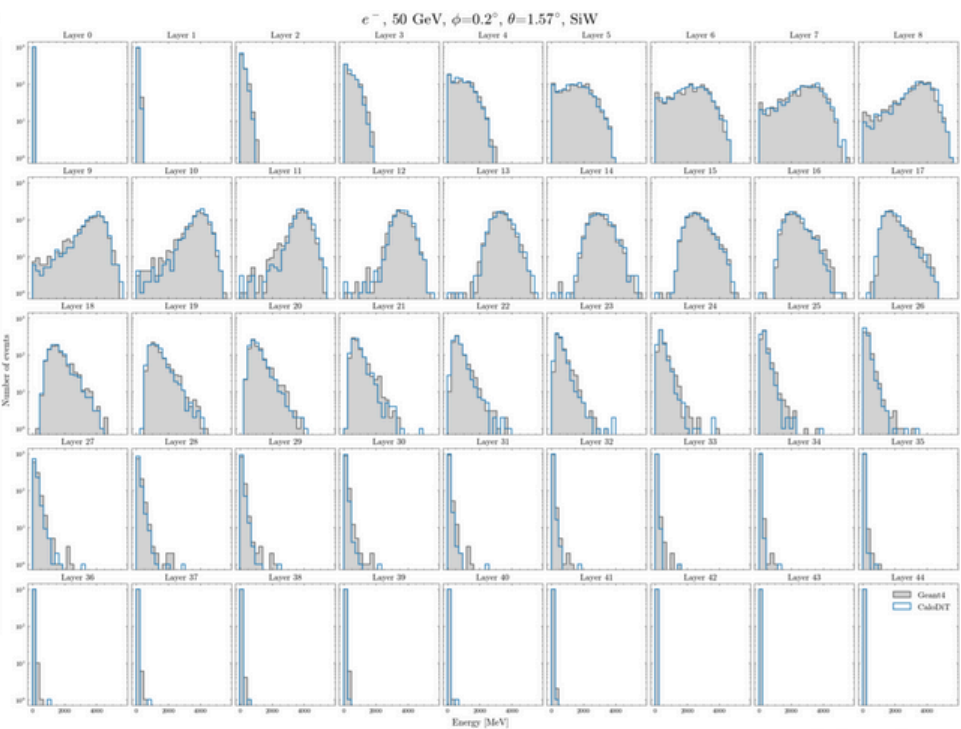
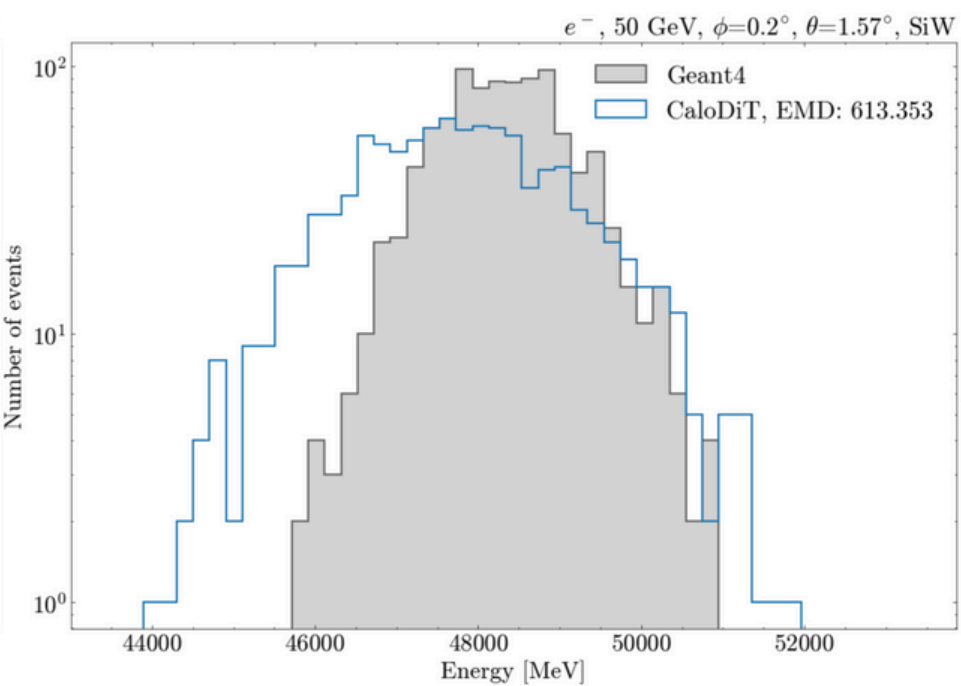
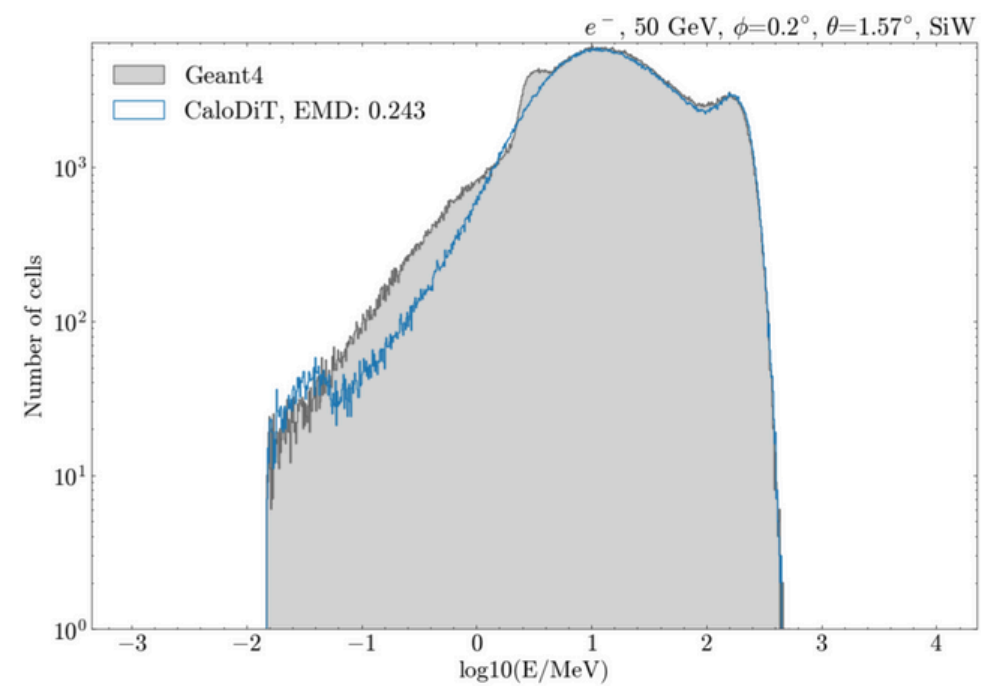


Updates: LMS (Adams-Bashforth), 32 steps



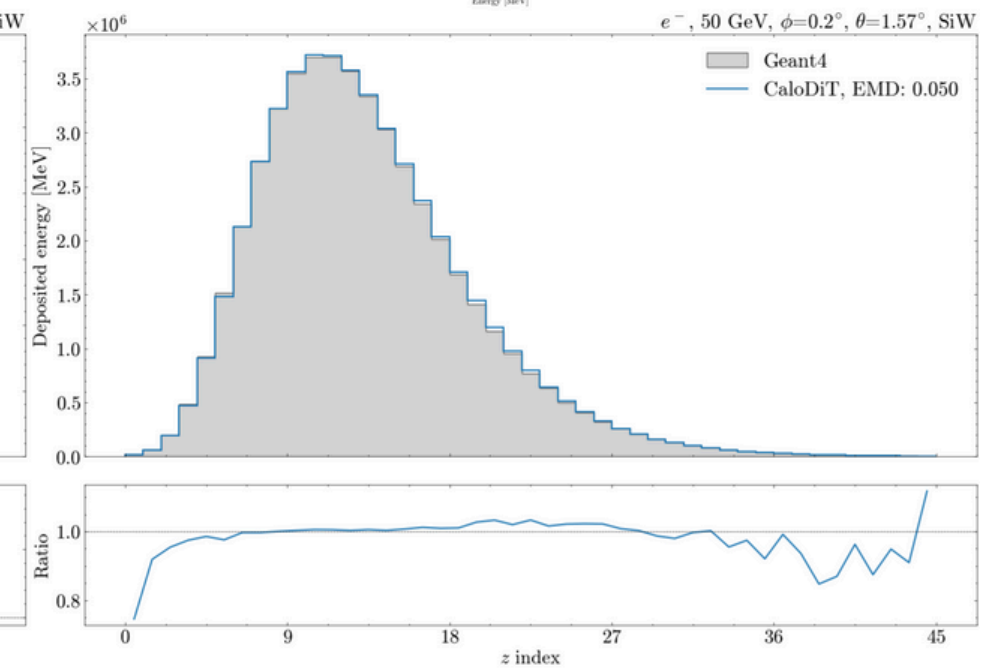
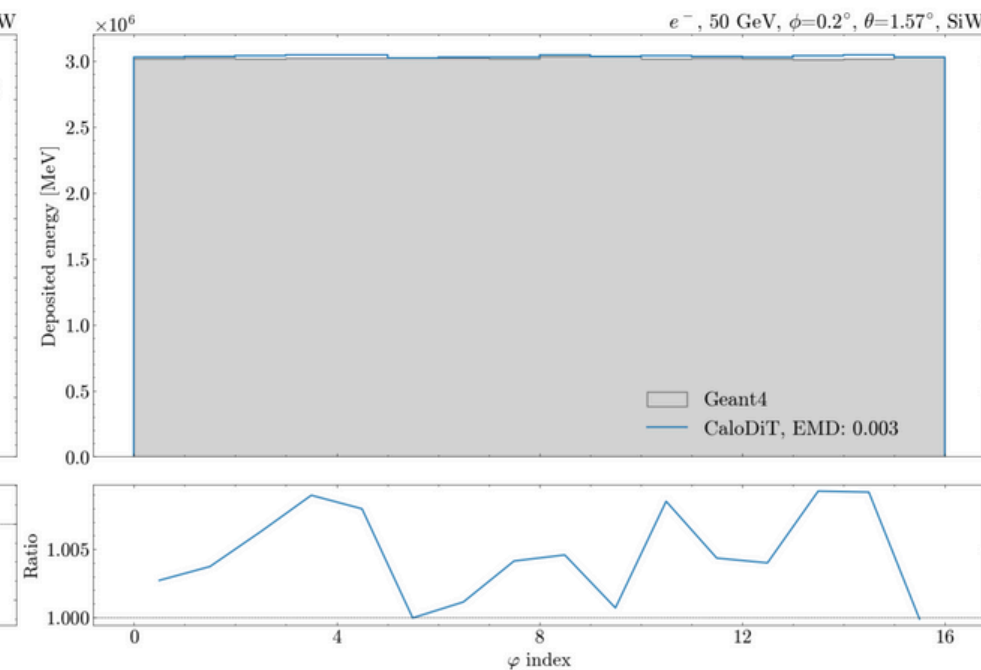
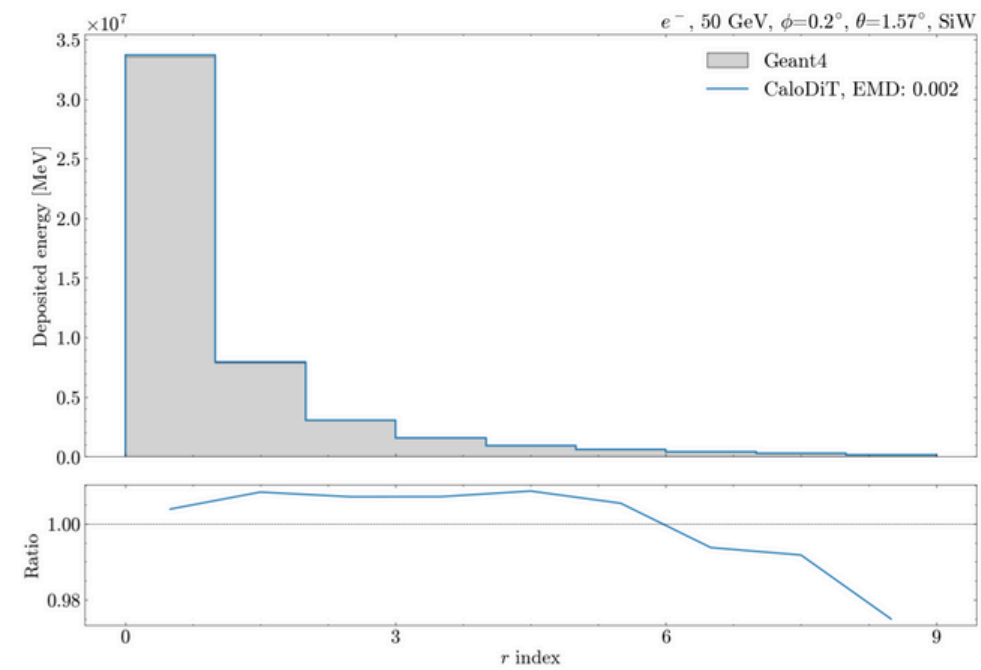
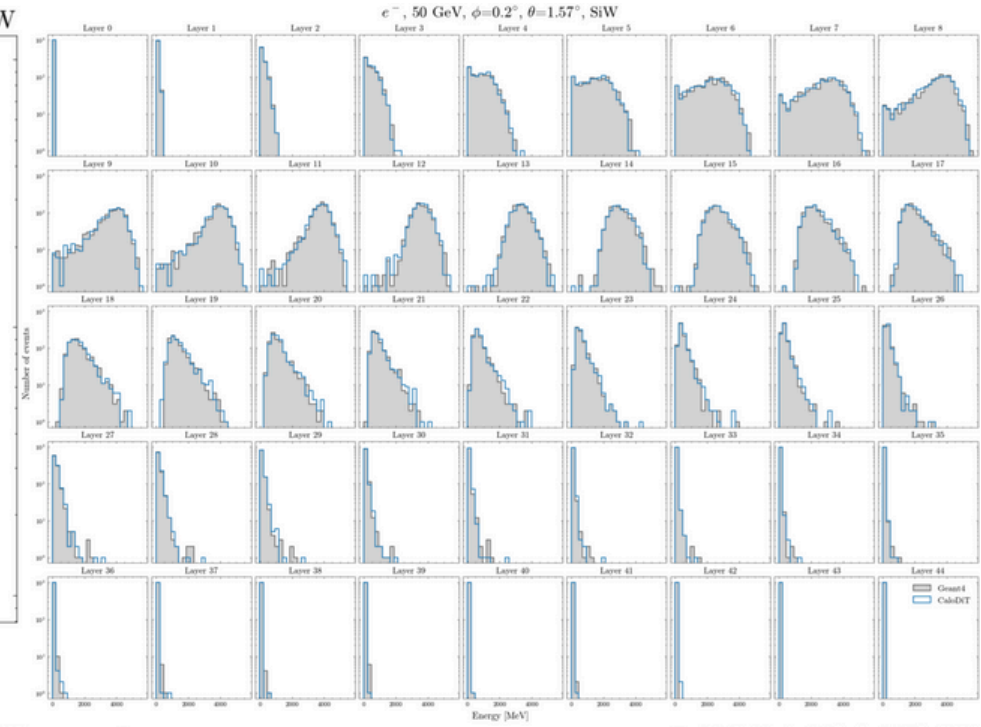
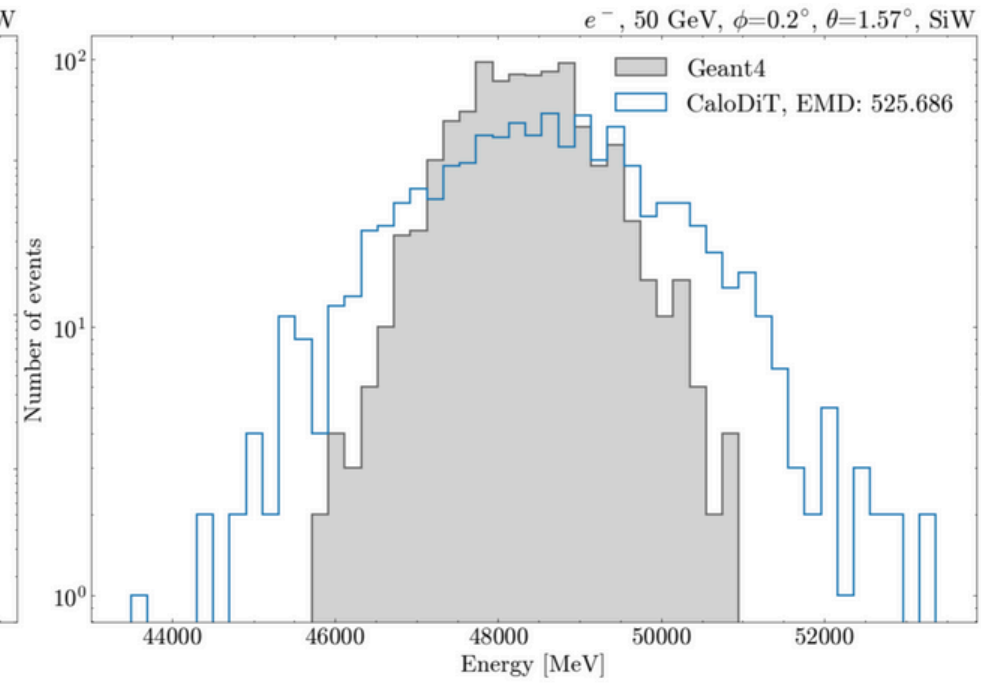
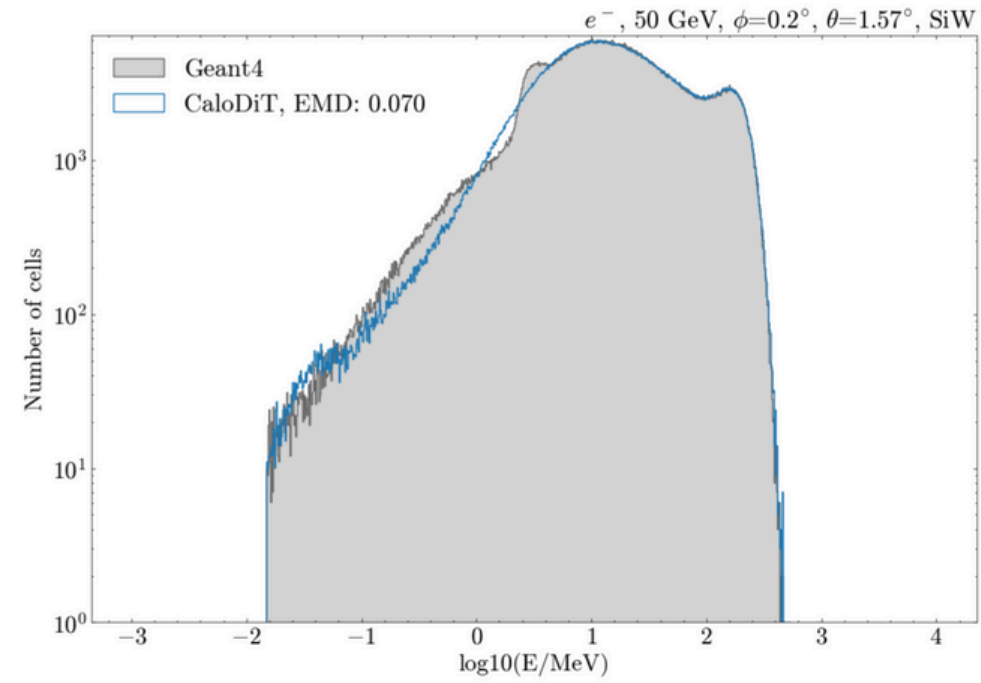


Updates: Euler, 256 steps

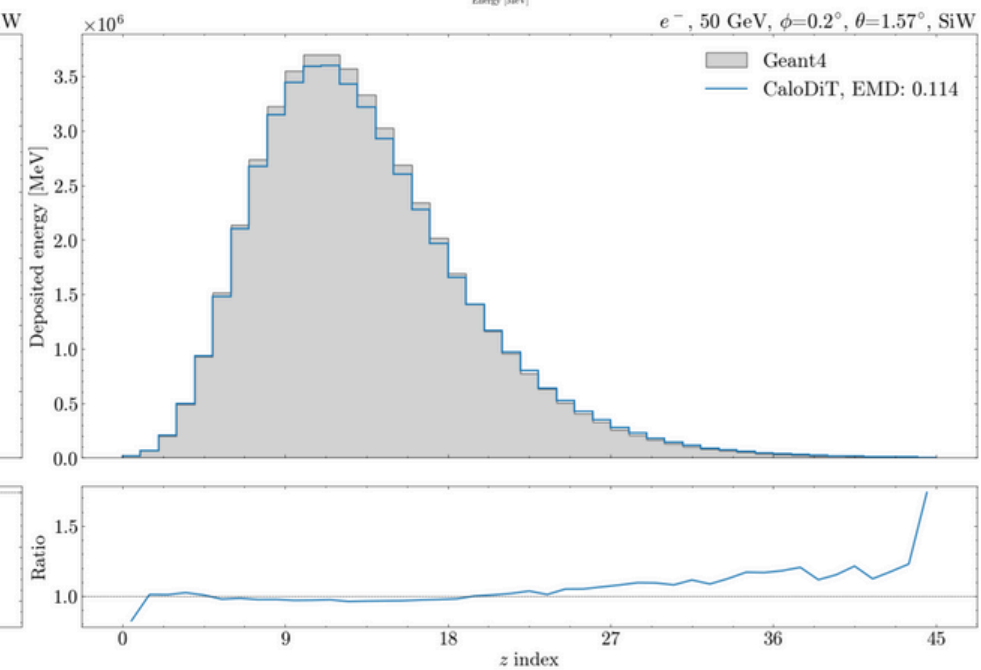
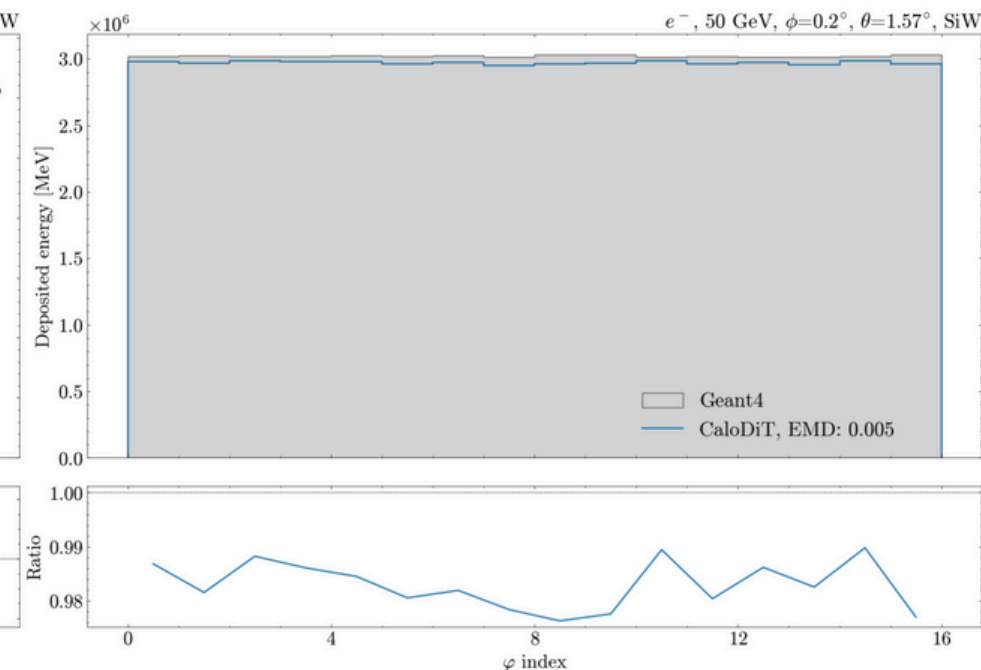
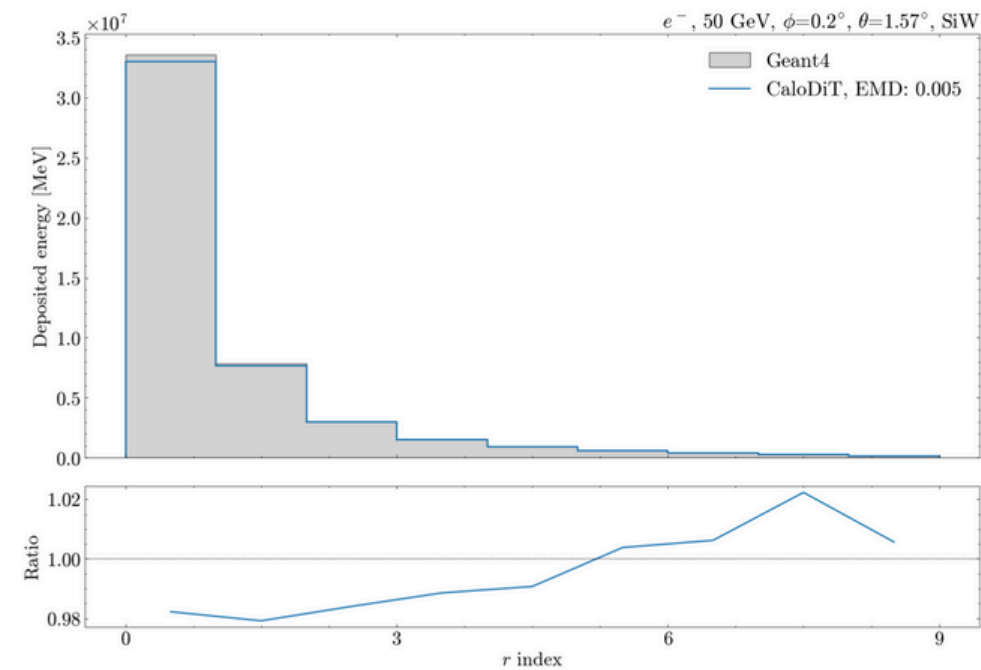
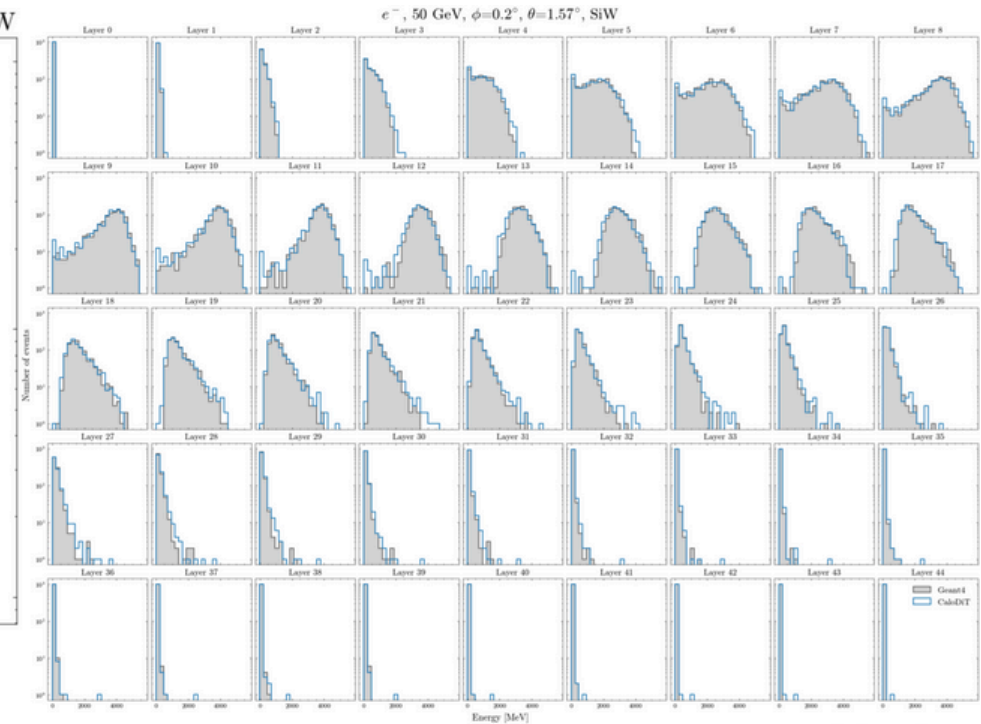
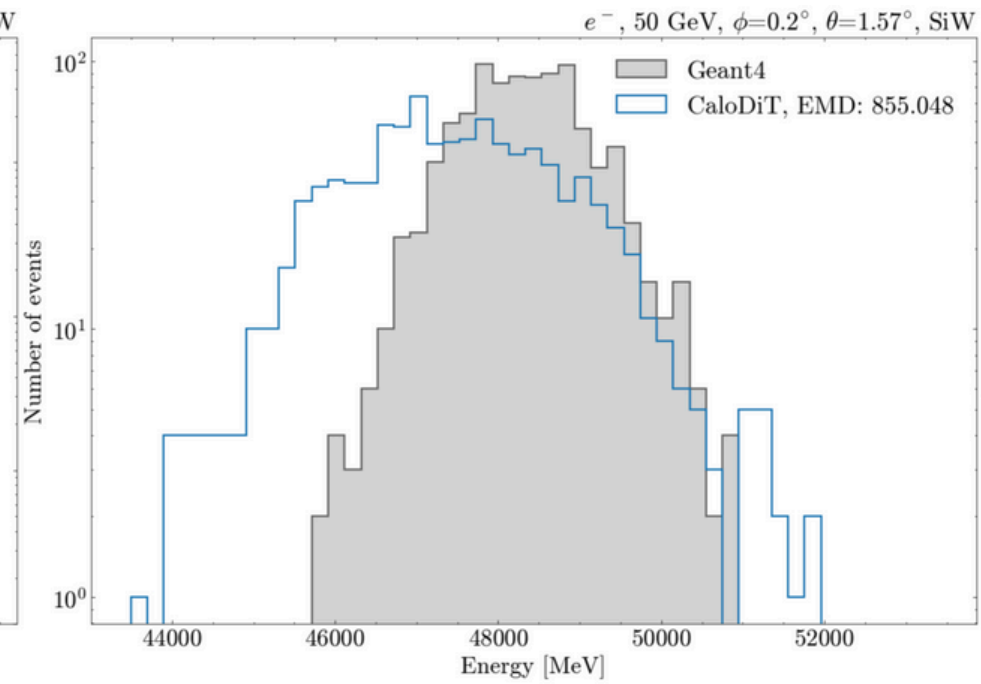
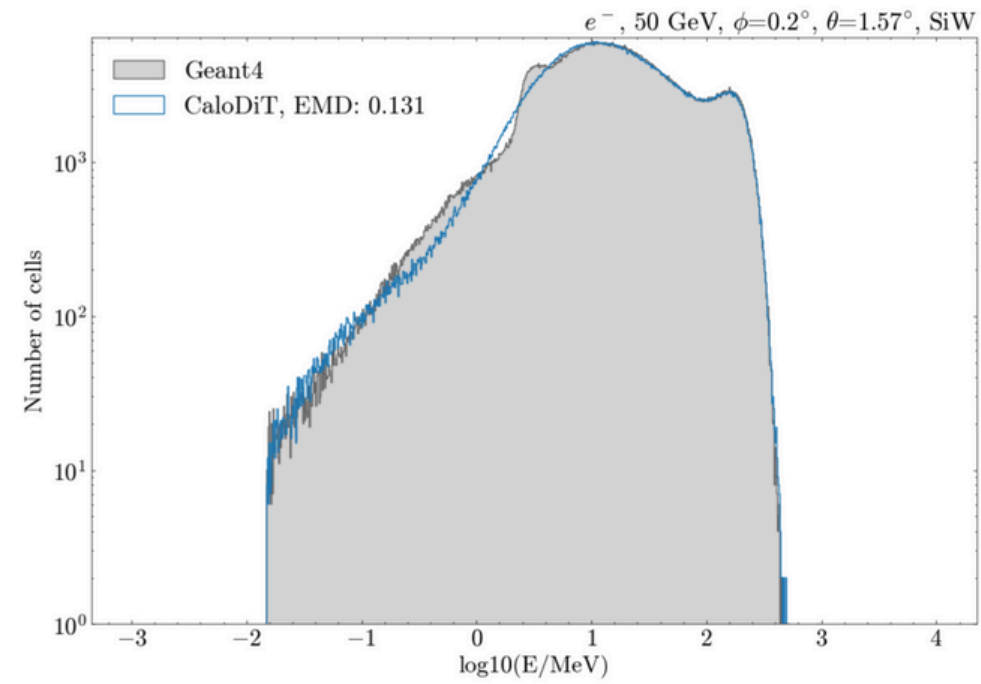




Updates: Heun, 255 steps



Updates: LMS (Adams-Bashforth), 256 steps





Next steps

1. Investigate why the model is underfitting (total event energy distribution, cell log-energy distribution)
2. Fix training with uniform energy distribution
3. Finish the implementation of DPM-Solver++