Fast Simulation of Calorimetry showers (FastSim)

Mikołaj Piórczyński
ML4EP Meeting, 06.06.2024

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.

SAMSUNG

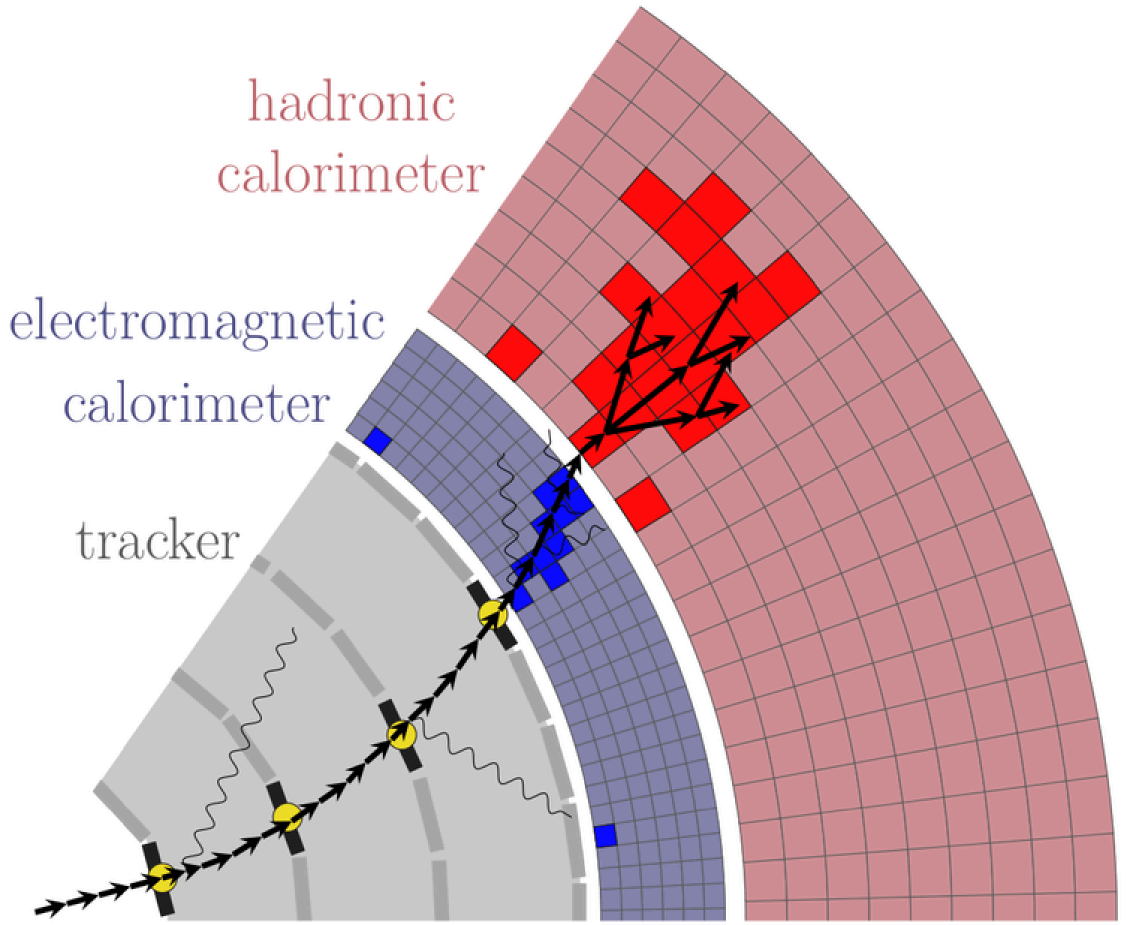
AI Clearing

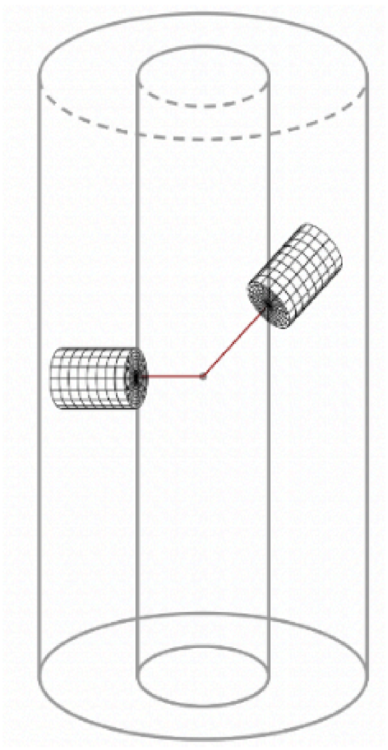
ML in PL ASSOCIATION

Faculty of Mathematics and Information Science

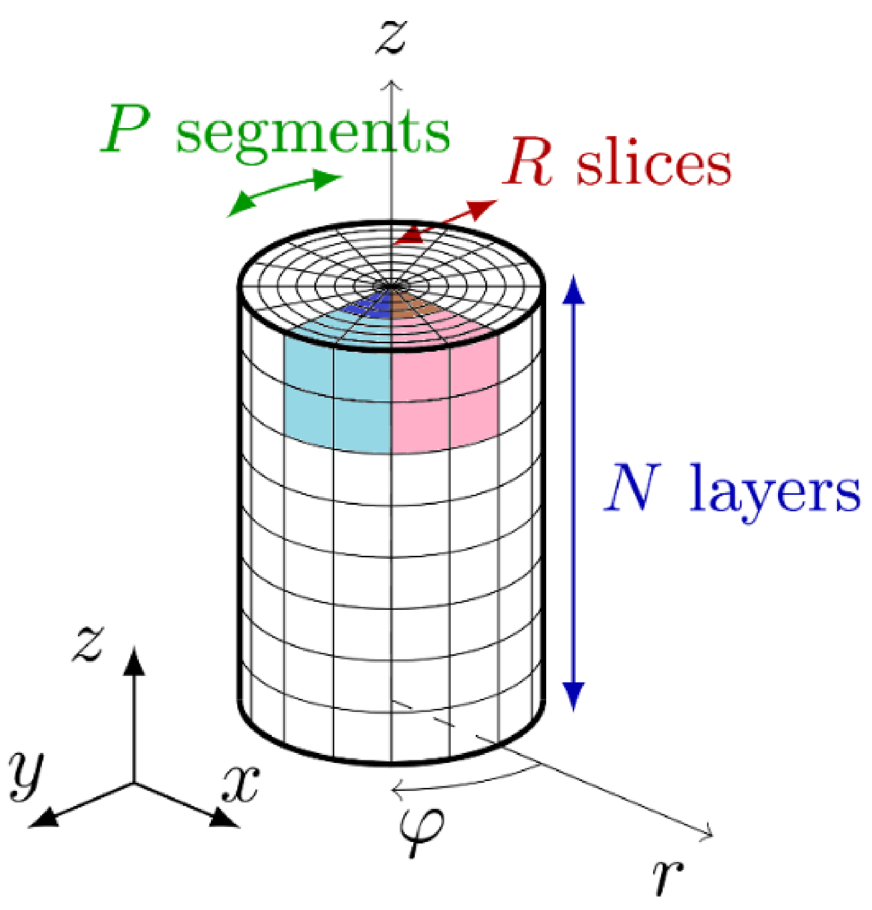
WARSAW UNIVERSITY OF TECHNOLOGY

Introduction



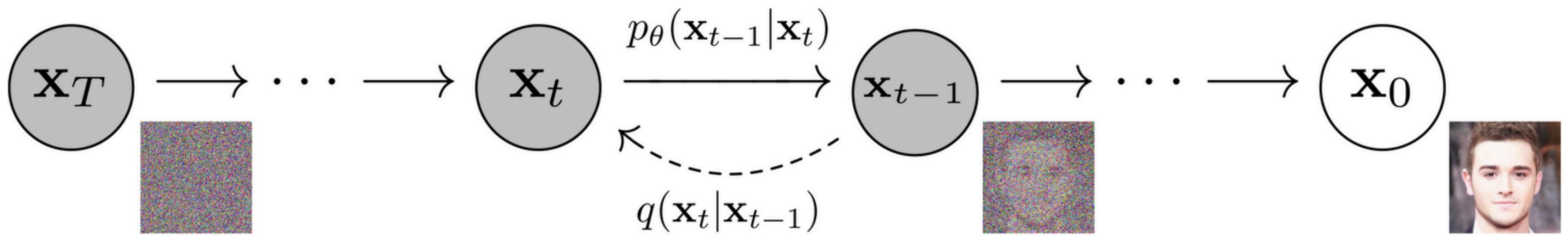


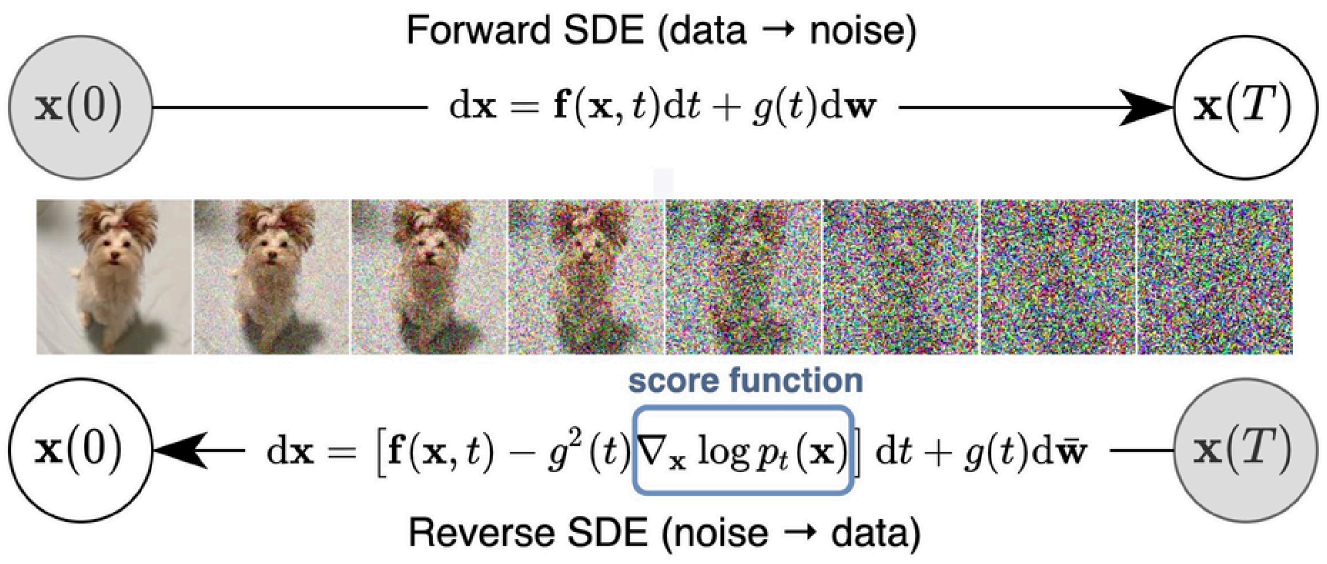
(a)



(b)

Introduction





How to make diffusion models faster?*

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

arXiv:2205.14135v2 [cs.LG] 23 Jun 2022

FlashAttention: Fast and Memory-Efficient Exact Attenti 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,
chrismre@cs.stanford.edu

June 24, 2022

Abstract

Transformers are slow and memory-hungry on long sequences, since the
of self-attention are quadratic in sequence length. Approximate attenti
to address this problem by trading off model quality to reduce the comp
not achieve wall-clock speedup. We argue that a missing principle is mak
aware—accounting for reads and writes between levels of GPU memory. W
an IO-aware exact attention algorithm that uses tiling to reduce the num
between GPU high bandwidth memory (HBM) and GPU on-chip SRAM. V
of FlashAttention, showing that it requires fewer HBM accesses thar
optimal for a range of SRAM sizes. We also extend FlashAttention to bl
an approximate attention algorithm that is faster than any existing ap
FlashAttention trains Transformers faster than existing baselines: 15% e
on BERT-large (seq. length 512) compared to the MLPerf 1.1 training s
GPT-2 (seq. length 1K), and 2.4× speedup on long-range arena (seq. lengt
formers, y
ument cl
performa
63.1% a

hitectur
e grown
the self-
n impor
odels ac

G] 5 Oct 2022

Token Merging for Fast Stable Diffusion

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

30 Mar 2023

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

age
fits
and
28
el.

nd
of
ken
ave
ns-
op.
the
.g.,
in
ile
ney

Me

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

herine Crowson*[1] Stefan Andreas Baumann*[2] Alex Birch*[3] Tanishq Mathew Abraham[1]
Daniel Z. Kaplan[4] Enrico Shippole[5]

mples generated directly in RGB pixel space using our HDiT
d on FFHQ-1024² and ImageNet-256².

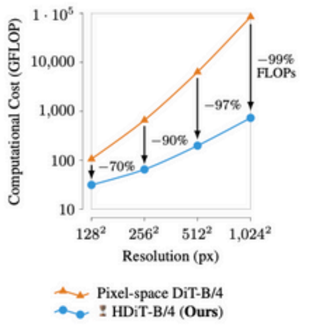


Figure 2: Scaling of computational cost w.r.t. target resolution of our HDiT-B/4 model vs. DiT-B/4 (Peebles & Xie, 2023a), both in pixel space. At megapixel resolutions, our model incurs less than 1% of the computational cost compared to the standard diffusion transformer DiT at a comparable size.

Abstract

nt the Hourglass Diffusion Transformer
an image generative model that exhibits
aling with pixel count, supporting train-
gh-resolution (e.g. 1024 × 1024) directly
space. Building on the Transformer ar-
e, which is known to scale to billions of
ers, it bridges the gap between the effi-
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.

*Equal contribution [1]Stability AI [2]LMU Munich [3]Birchlabs

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 10× to 50× faster in terms of wall-clock time compared to DDPMs, allow us to trade off computation
image interpolation directly
very low error.

Structural Pruning for Diffusion Models

Gongfan Fang Xinyin Ma Xinchao Wang*
National University of Singapore
gongfan@u.nus.edu, maxinyin@u.nus.edu, 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.

Consistency Models

ng[1] Prafulla Dhariwal[1] Mark Chen[1] Ilya Sutskever[1]

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

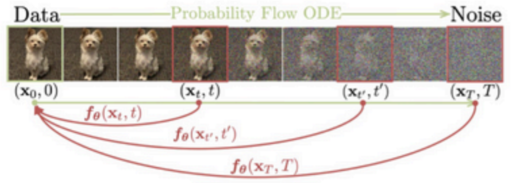


Figure 1: Given a Probability Flow (PF) ODE that smoothly converts data to noise, we learn to map any point (*e.g.*, $\mathbf{x}_t$, $\mathbf{x}_{t'}$, and $\mathbf{x}_T$) on the ODE trajectory to its origin (*e.g.*, $\mathbf{x}_0$) for generative modeling. Models of these mappings are called consistency models, as their outputs are trained to be

oduce high quality samples in many
n terms of image generation, genera-
rrently exhibits higher sample quality
rs (Kingma & Welling, 2013), autore-
flows (Rezende & Mohamed, 2015;
ices in optimization and architectures
t al., 2017; Karras et al., 2018; Brock
tion (Zhao et al., 2018).

14), such as denoising diffusion prob-

DeepCache: Accelerating Diffusion Models for Free

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

Dec 2023

(a) Stable Diffusion v1.5

2.3×

speeding-up diffusion models

model architecture

diffusion process

speeding-up diffusion models

model architecture

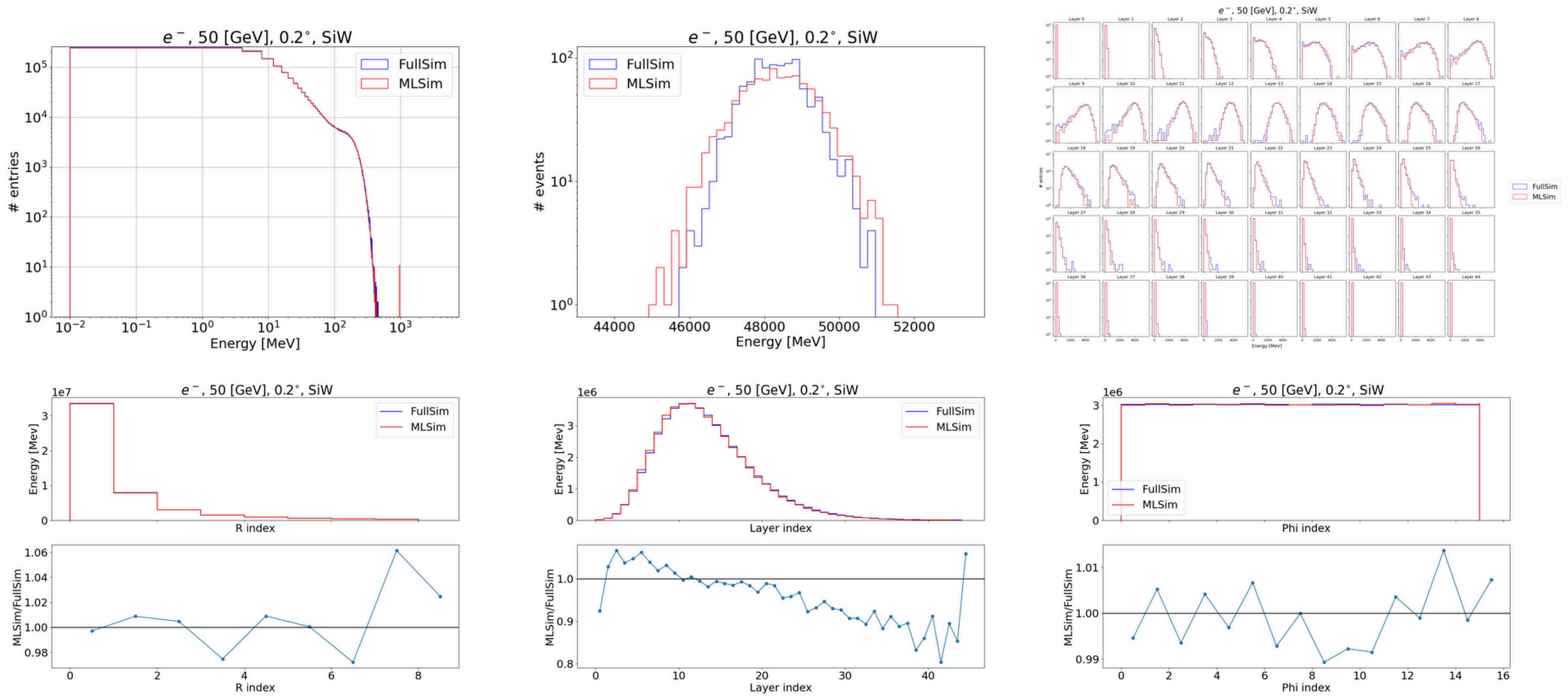
diffusion process

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++ (https://arxiv.org/abs/2211.01095))
4. Optional: Consistency Distillation (https://arxiv.org/abs/2303.01469)

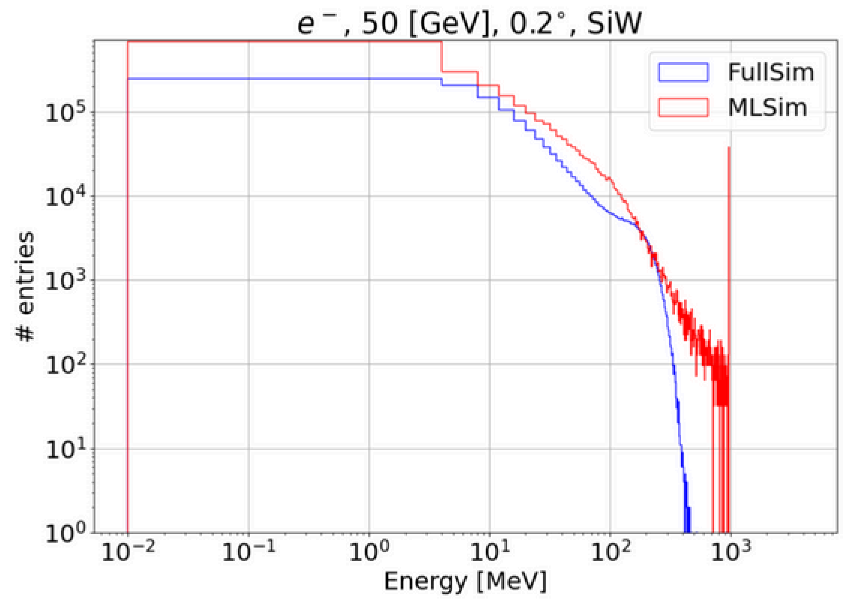
Code: https://gitlab.cern.ch/mpiorczy/diffusion4fastsim

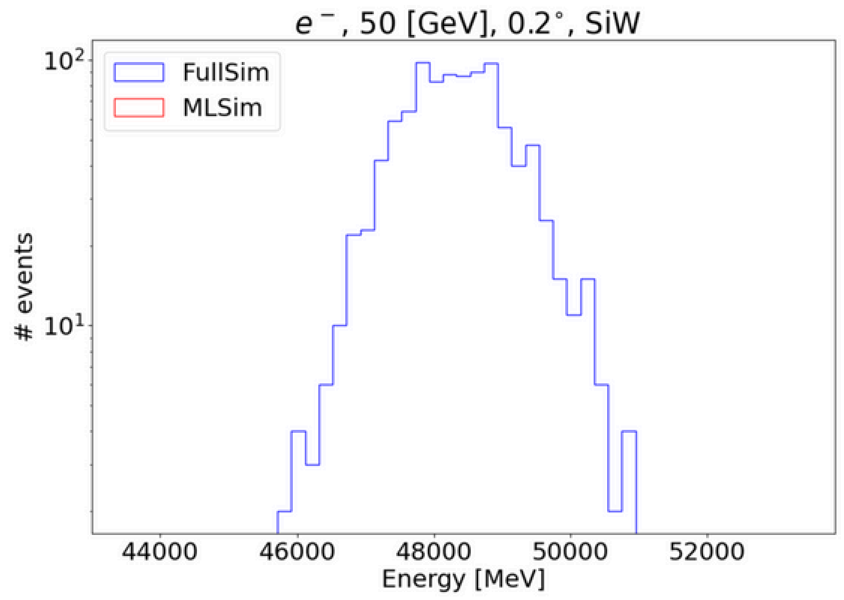
First results, DDPM

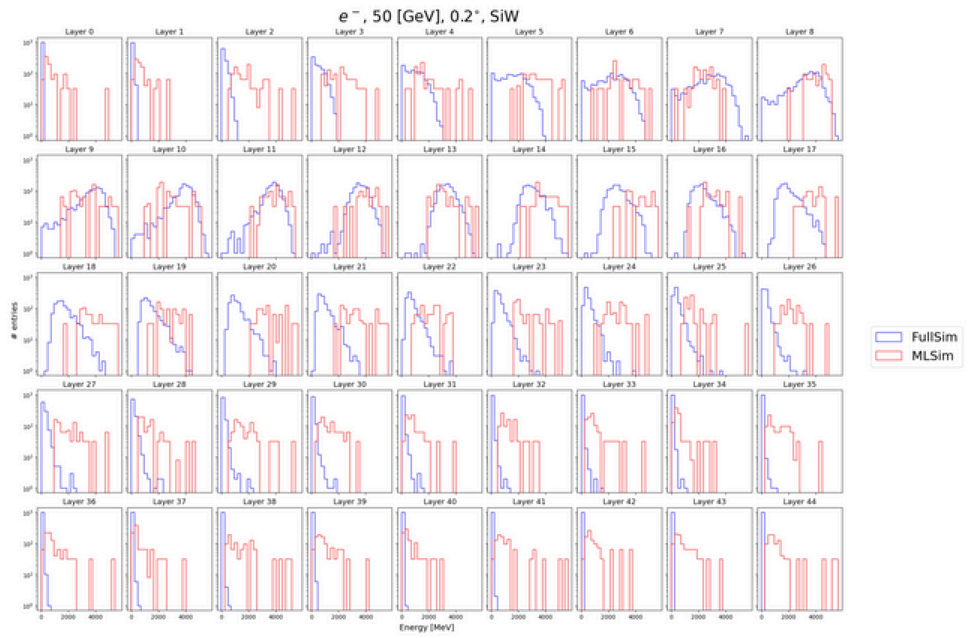


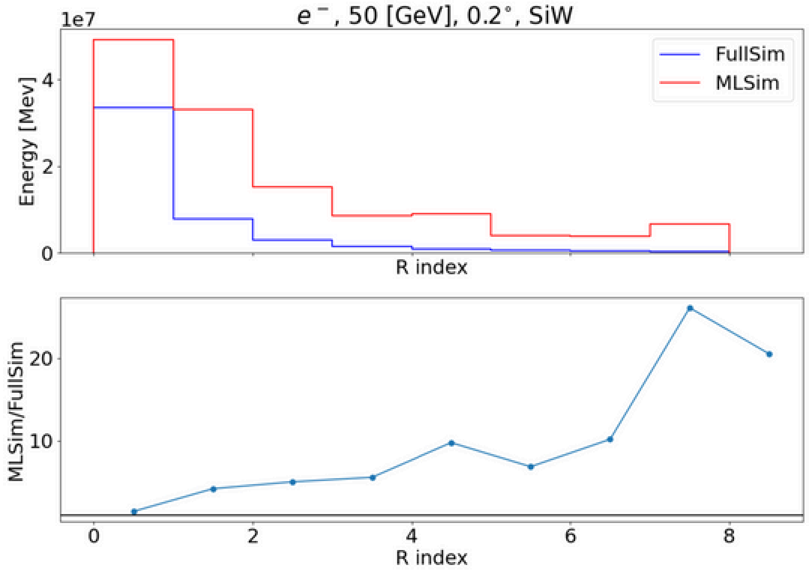
400 steps

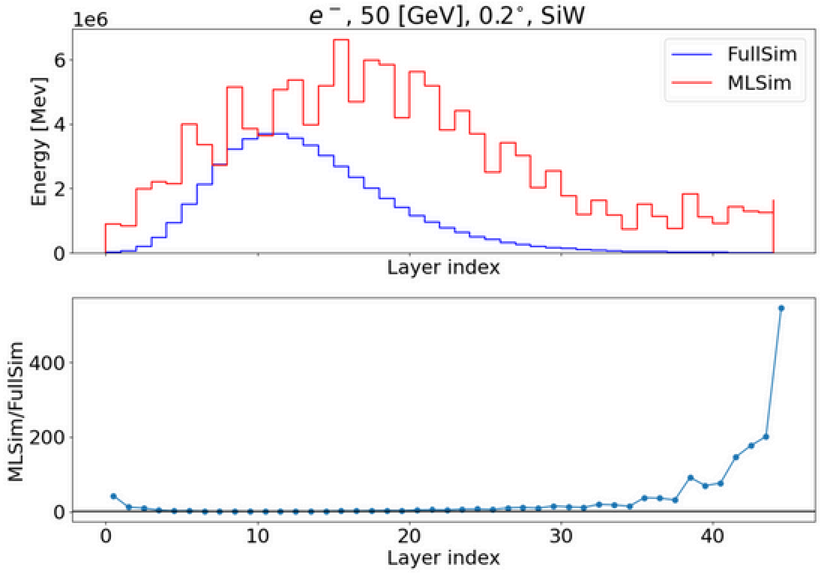
First results, DDIM (eta = 0.0)

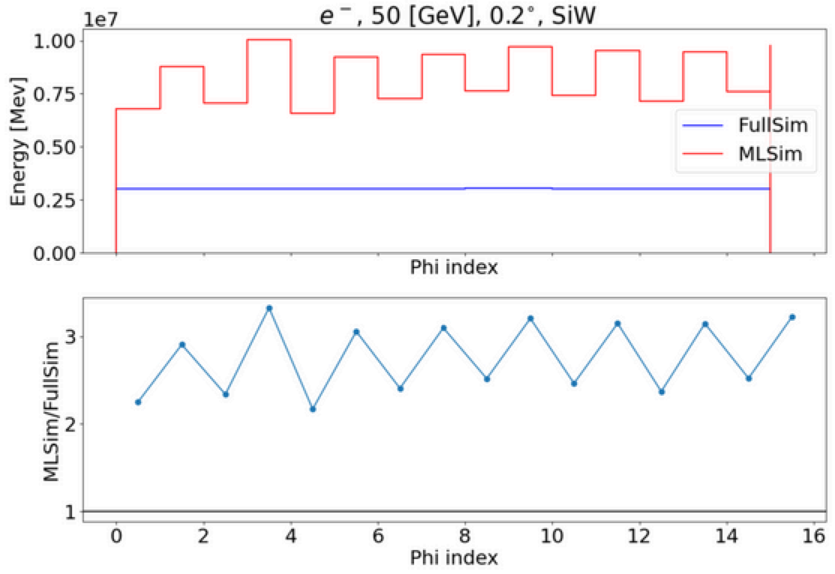






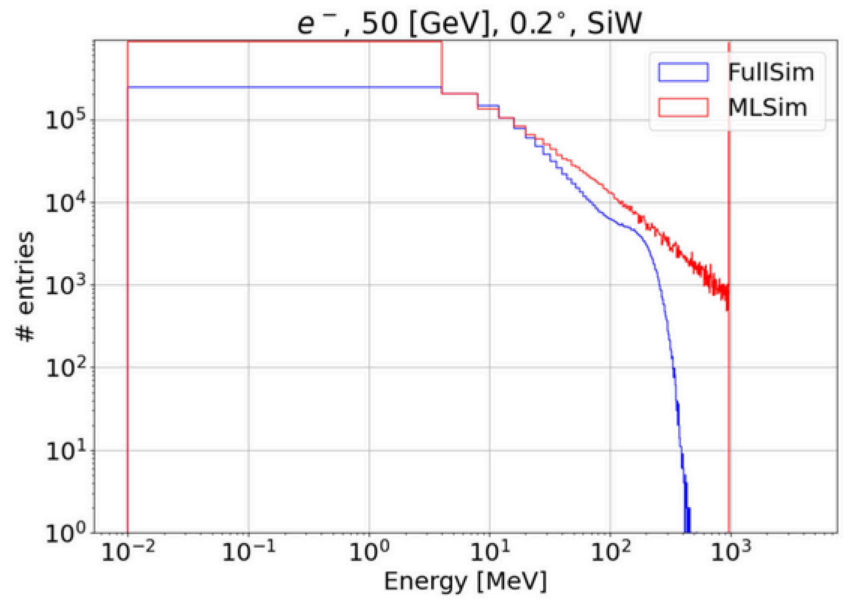


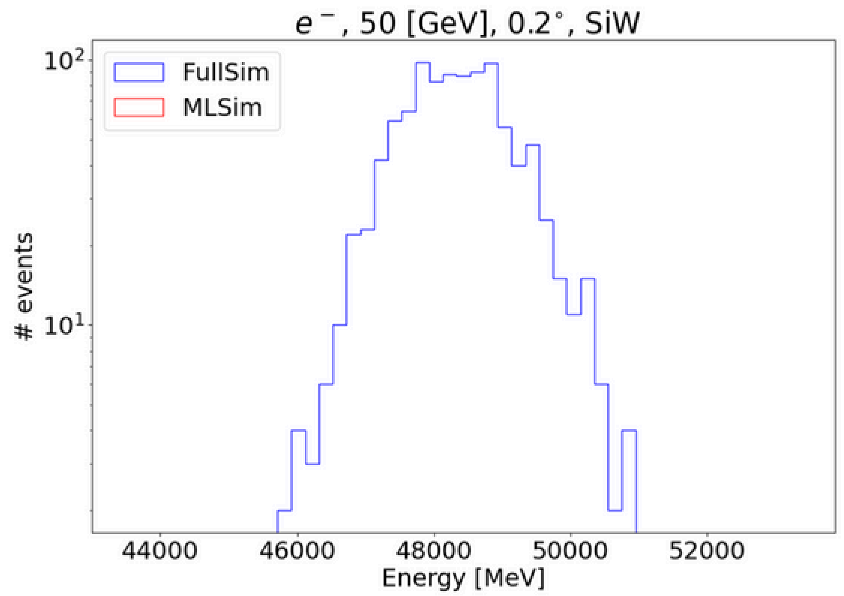


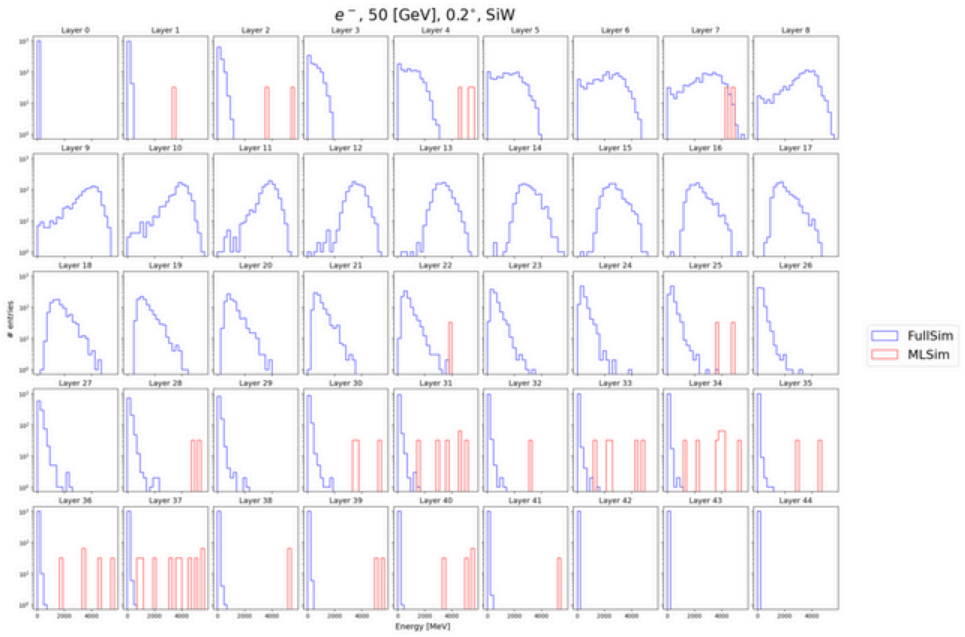


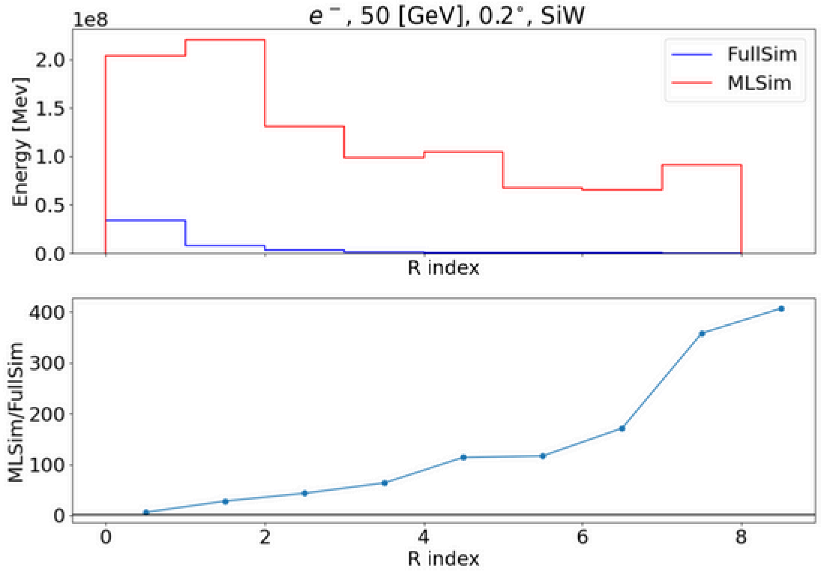
400 steps

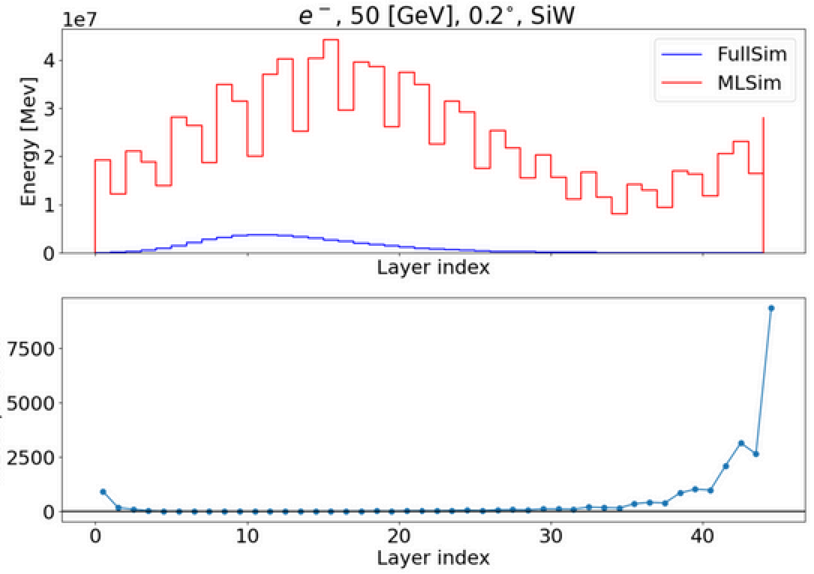
First results, DDIM (eta = 0.0)

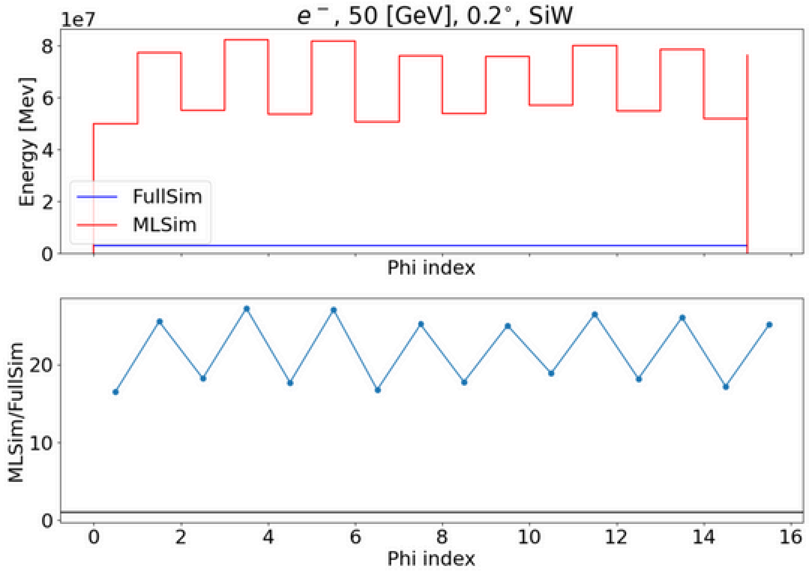






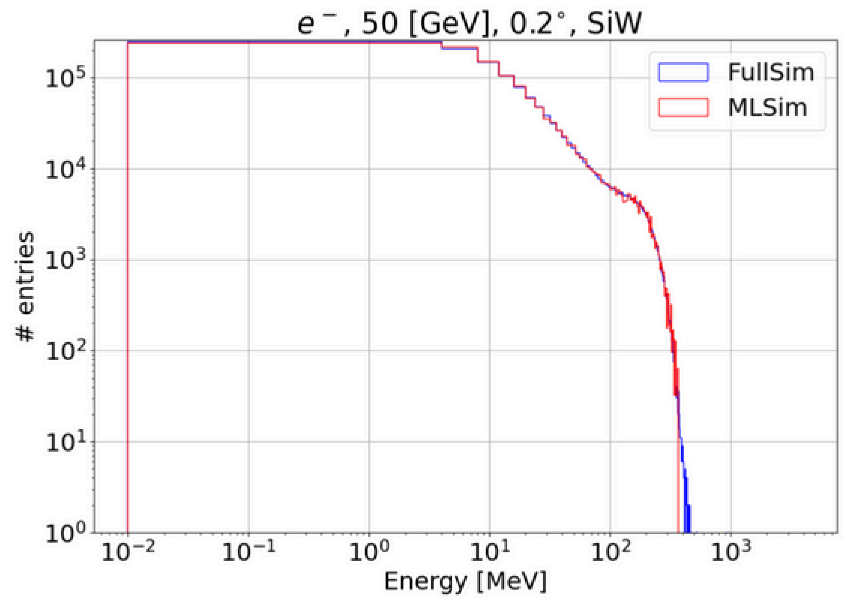


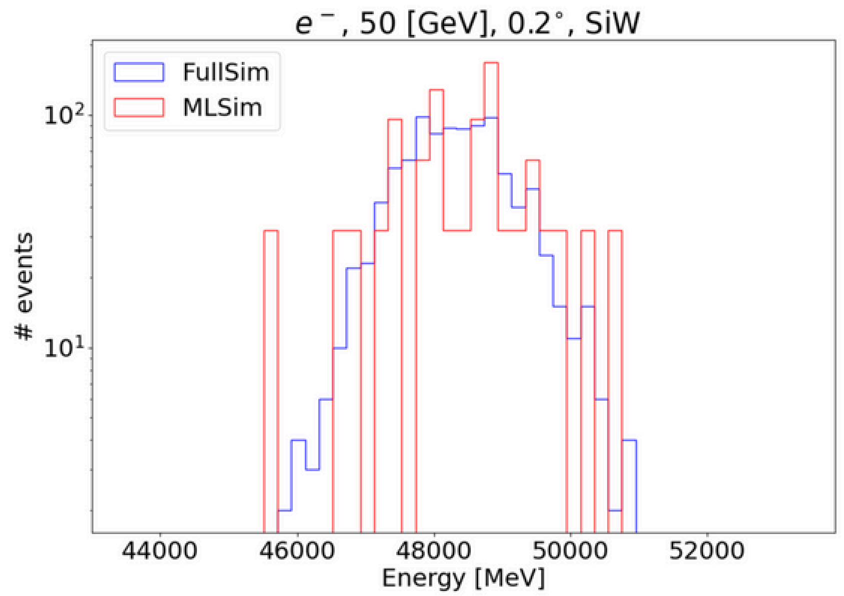


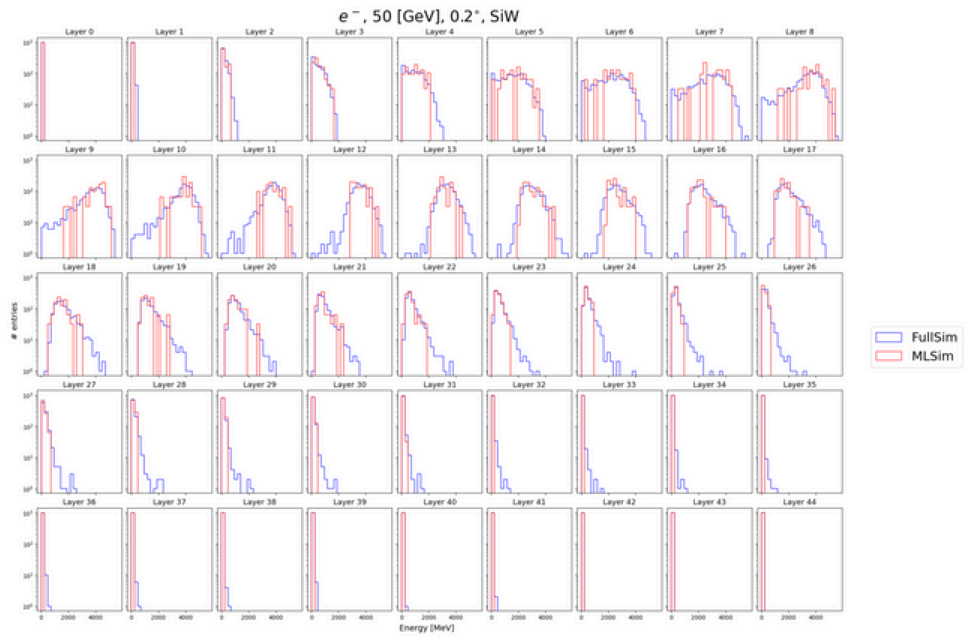


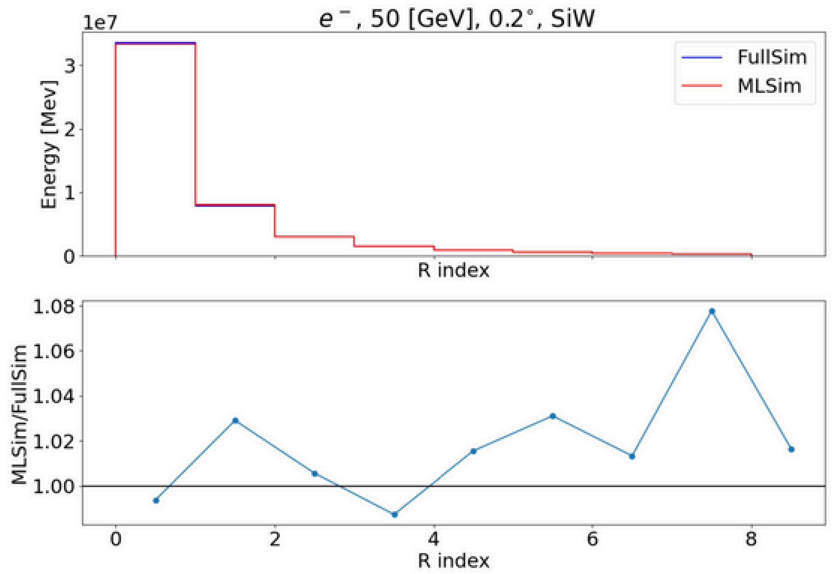
200 steps

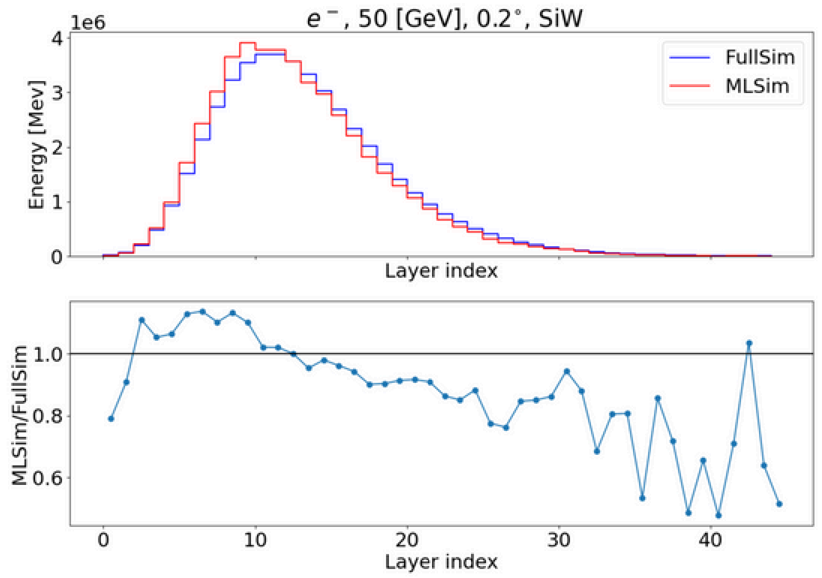
First results, DDIM (eta = 1.0)

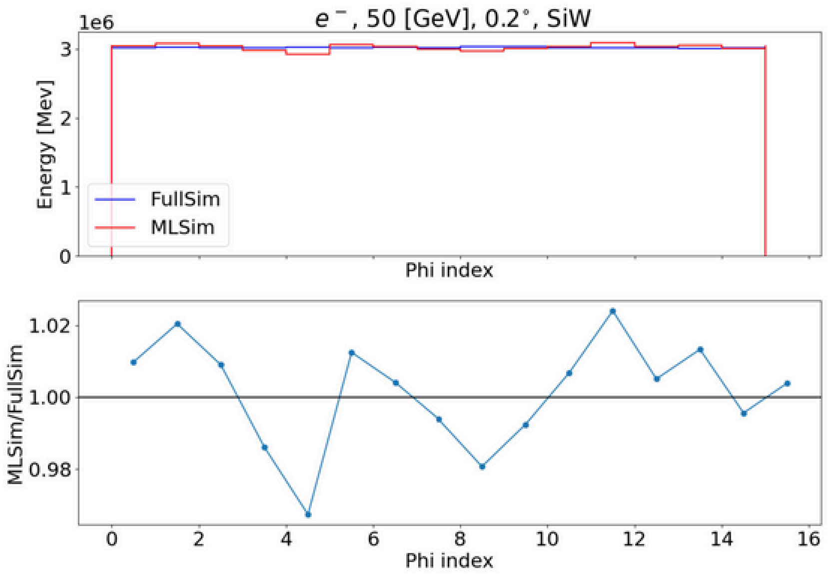






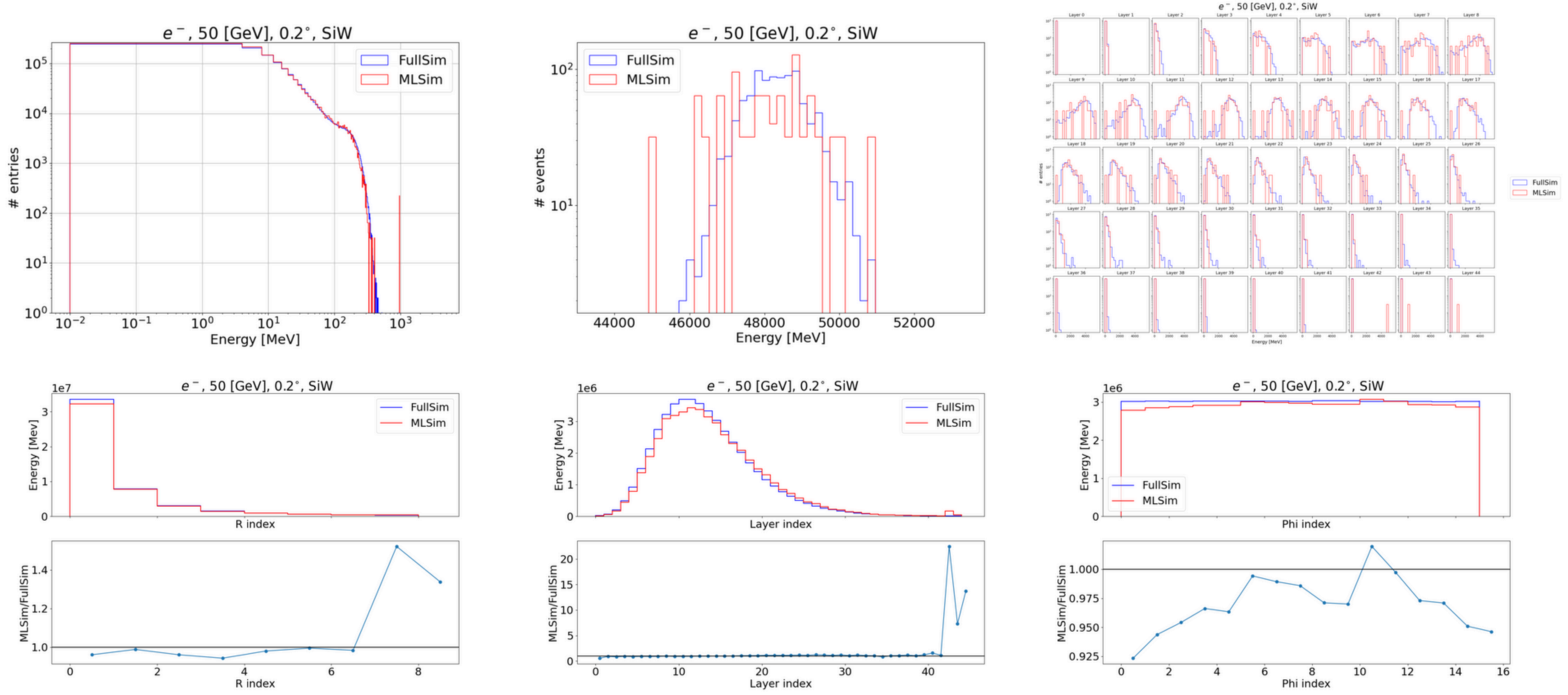






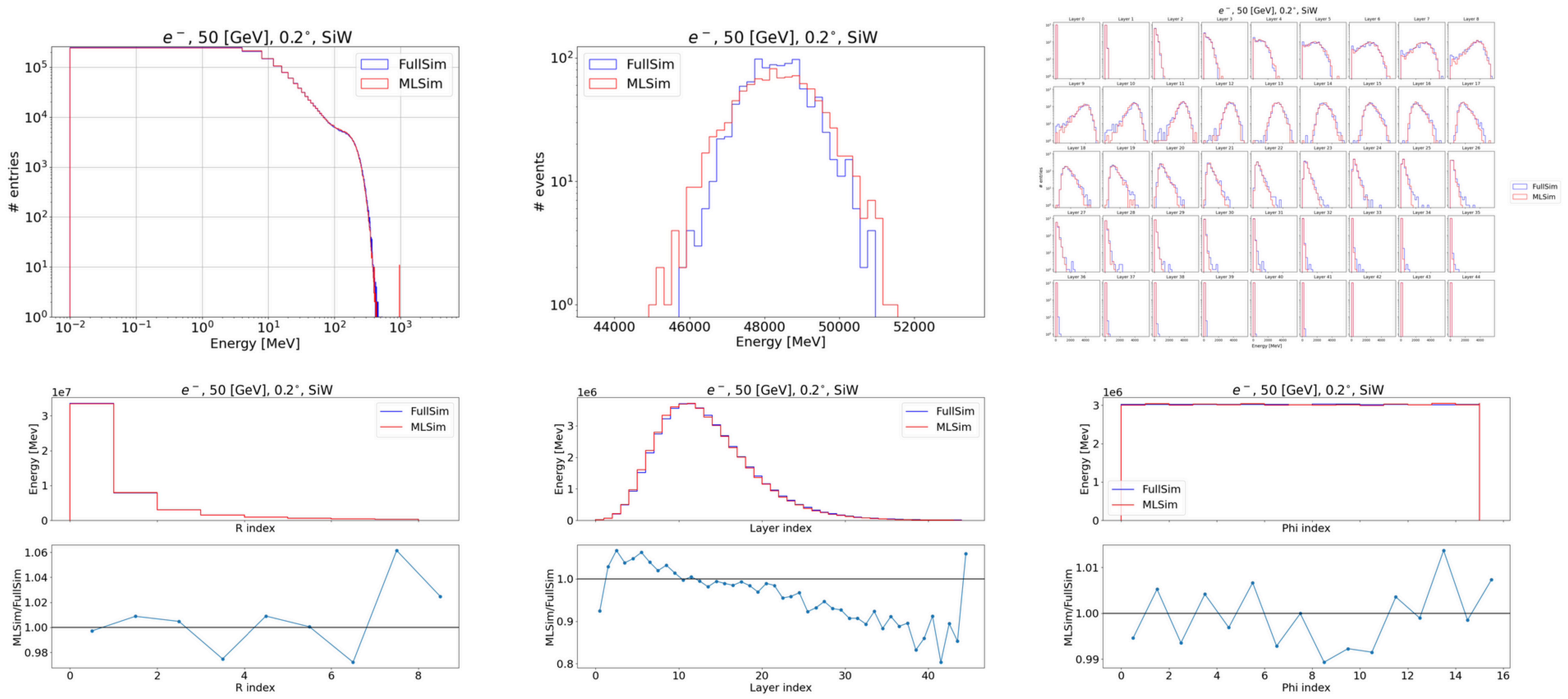
400 steps

First results, DDIM (eta = 1.0)



200 steps

First results, DDPM



400 steps

Next steps

1. Broader evaluation of DDIM
2. Strided sampling with DDPM (https://arxiv.org/abs/2102.09672)
3. (Maybe) Investigate if it’s not beneficial to train the model with a higher number of diffusion steps during the training and sample with a similar number of steps during the inference. I.e. if T = 1000/4000, S = 200 better than T = 400, S = 200?
4. Progressive Distillation