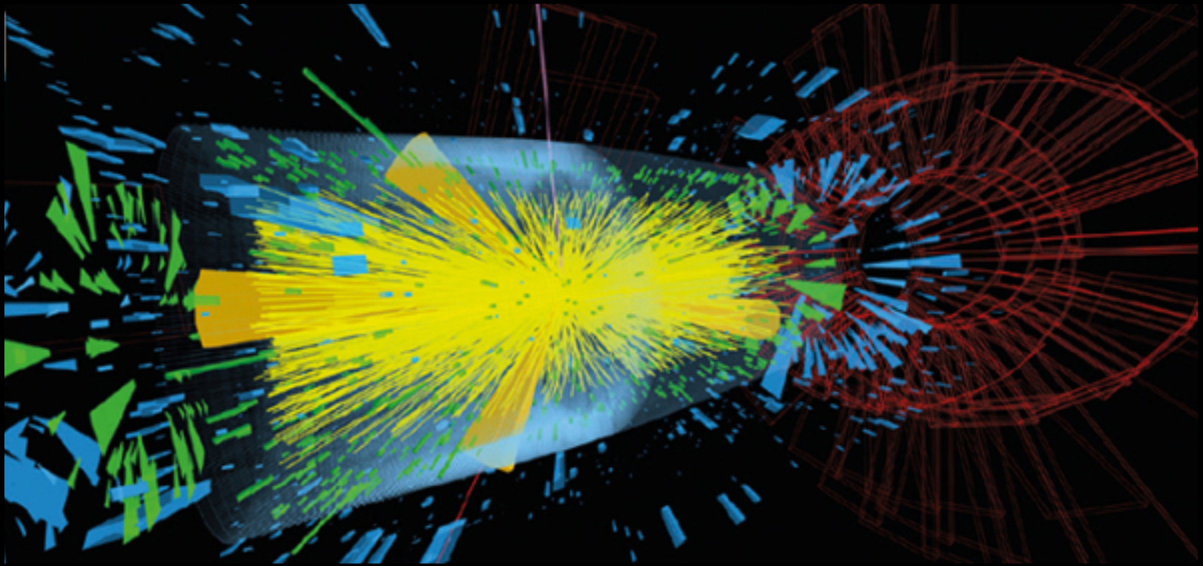
Fast & Accurate Calorimeter Simulation With Diffusion Models



Oz Amram

In collaboration with Kevin Pedro

Fermilab

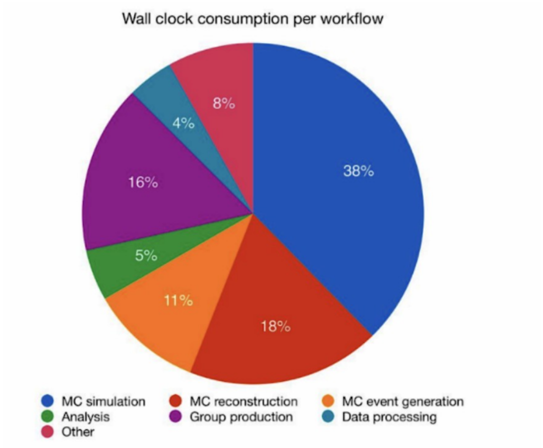
FERMILAB-SLIDES-23-052-PPD

Norfolk, Virginia, USA • **May 8-12, 2023**

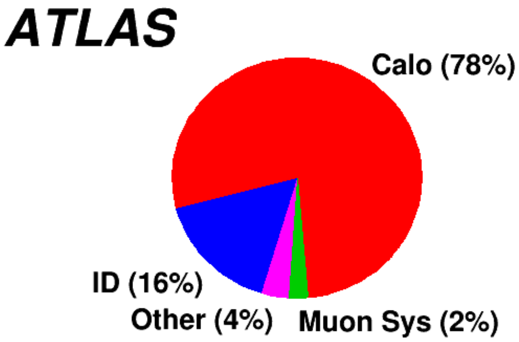
CHEP 2023

Computing in High Energy & Nuclear Physics

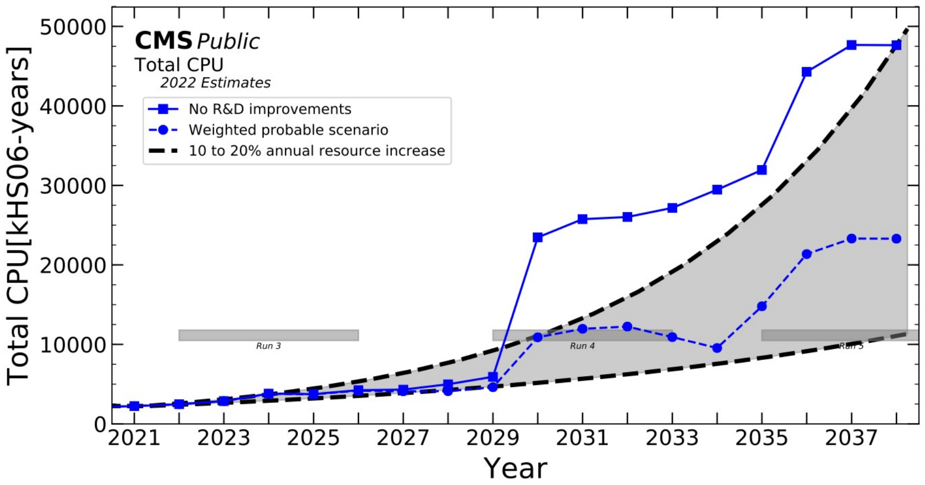
The Need for Fast Simulation



ATLAS CPU hours used by various activities in 2018

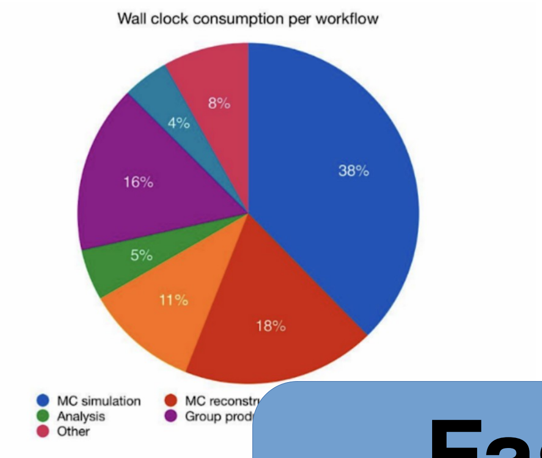


Subdetector CPU fraction for 50 ttbar events
MC16 Candidate Release

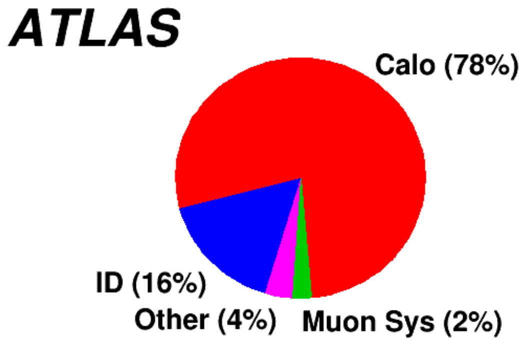


- Geant4 calo simulation is a significant part of ATLAS computing budget
  - CMS will face similar needs with HGCAL in HL-LHC
- For HL-LHC, computing simulation more crunched
  - Reconstruction usage will scale ~linearly with pileup → less resources for sim.

The Need for Fast Simulation

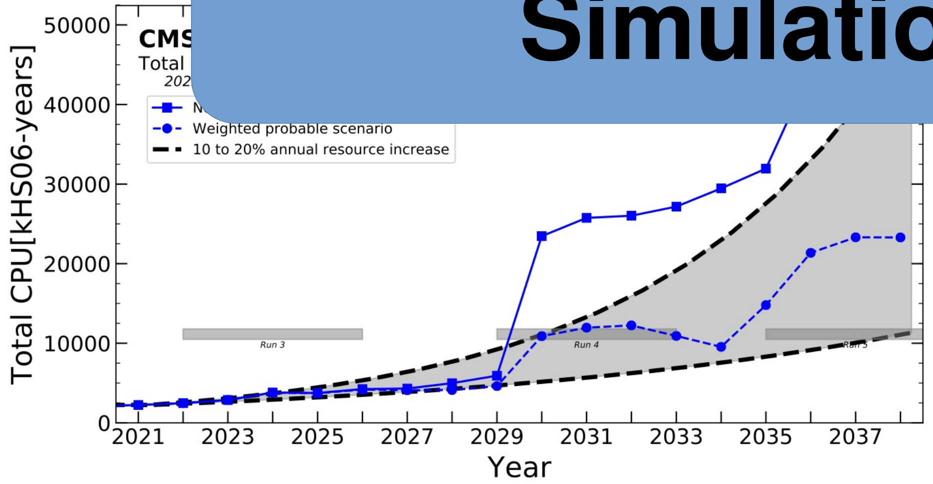


ATLAS CPU hours us



- Geant4 calo simulation is a significant part of ATLAS computing budget
  - CMS will face similar needs

Fast & Accurate Calorimeter Simulation is Needed!



ing simulation more crunched

  - Reconstruction usage will scale ~linearly with pileup → less resources for sim.

Diffusion Models

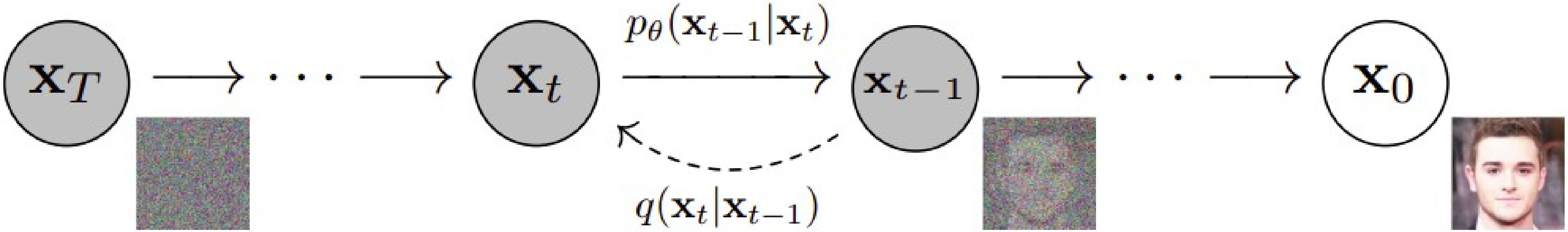
- Diffusion has become the dominant paradigm for ML image generation
  - Dalle-2, Midjourney, Stable Diffusion, etc
- Easy training, high quality results, reasonable computation times

"AI aiding physicists at LHC to analyze data and discover new particles"

Diffusion Models : Technical Details

2006.11239

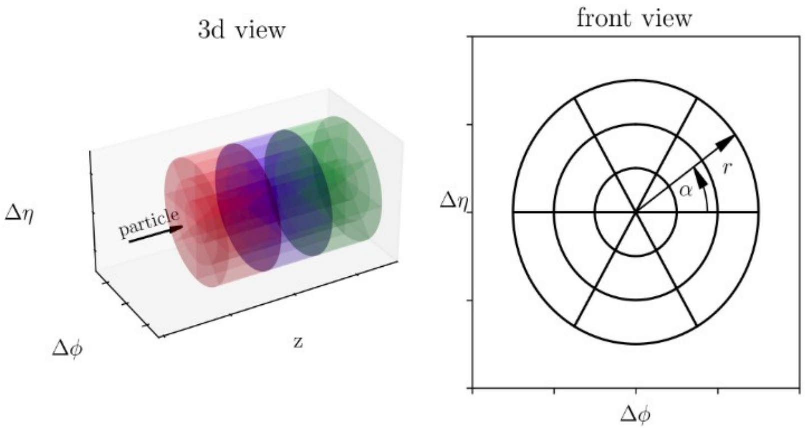
NSteps typically ~few hundred

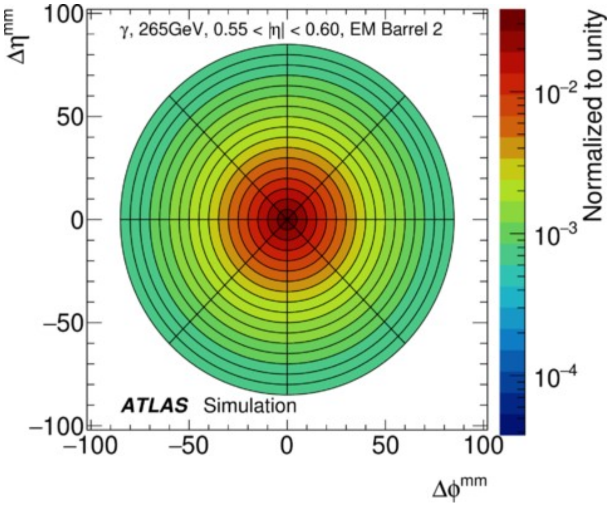


- Diffusion process: Starting with some image, **iteratively add Gaussian noise**, eventually reaching pure noise
- Train a model to **invert the diffusion process**
- Generate by starting from noise image, **iteratively denoise** using trained model
- Can condition on additional input information
  - Eg. text prompt or incident particle energy

Dataset: Calo Challenge

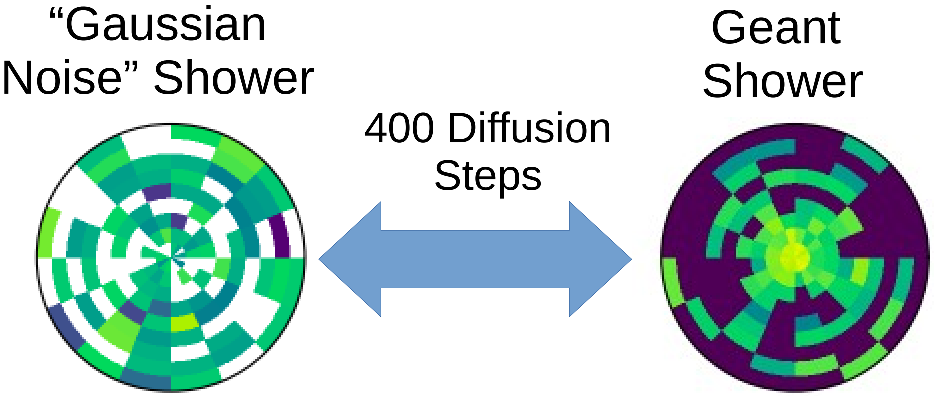
- Community challenge to compare generative models for Calorimeter simulation
- Standard datasets to allow comparison
  - Dataset1: ATLAS-like geometry, 5 layer cylinder with **irregular binning**, 368 voxels
  - Dataset2: 45 layers, 6480 total voxels
  - Dataset3: 45 layers, 40,500 total voxels

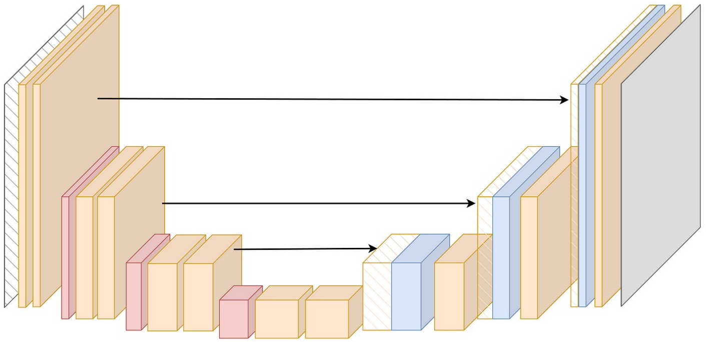




‘CaloDiffusion’

- We train diffusion models to generate synthetic calorimeter showers based on Geant simulations
- We use **400 steps** to interpolate from real shower to Gaussian noise
- Denoising network is has ‘U-net’ architecture based on 3D convolutions
  - Primary input: Noisy shower
  - Conditioning inputs: incident particle energy & diffusion step
- Training objective normalized noise component of the shower
  - Denoising → subtracting noise off
- Several novel optimizations utilized



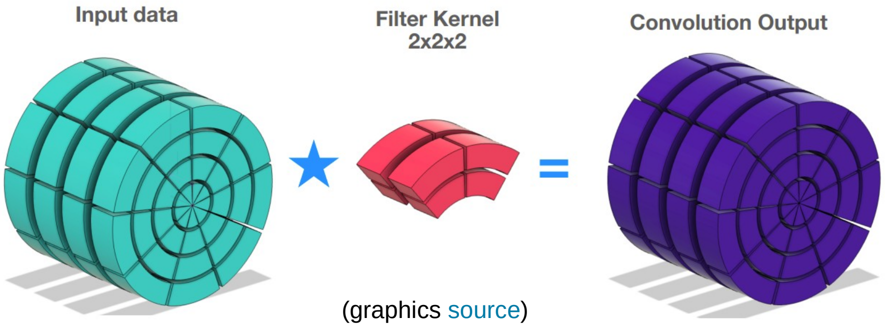


U-nets compress to a smaller dim space but also include skip connections

Optimizing for Cylindrical Data

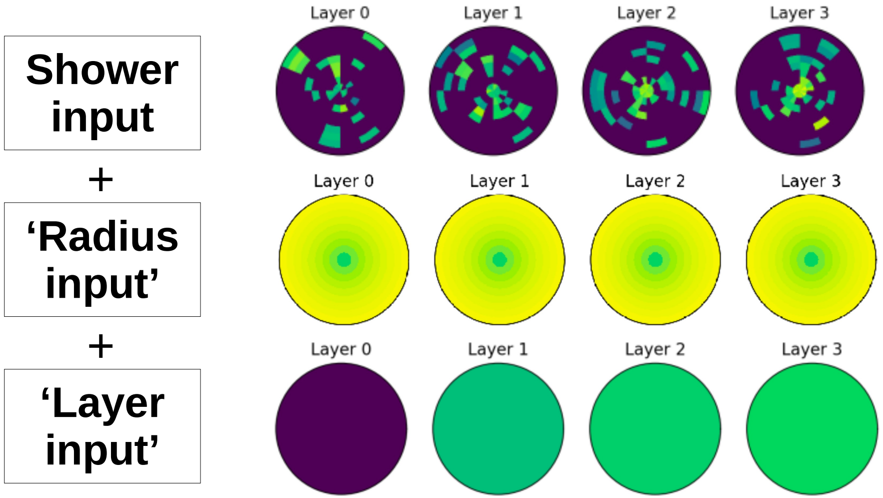
- Regular convolutions assume pure translation symmetry
- Our data : phi is **periodic**, and R & Z **not translation invariant**

Implement **cylindrical convolutions** to respect periodic boundary of phi



‘Circularly’ pad phi dimension before 3D conv

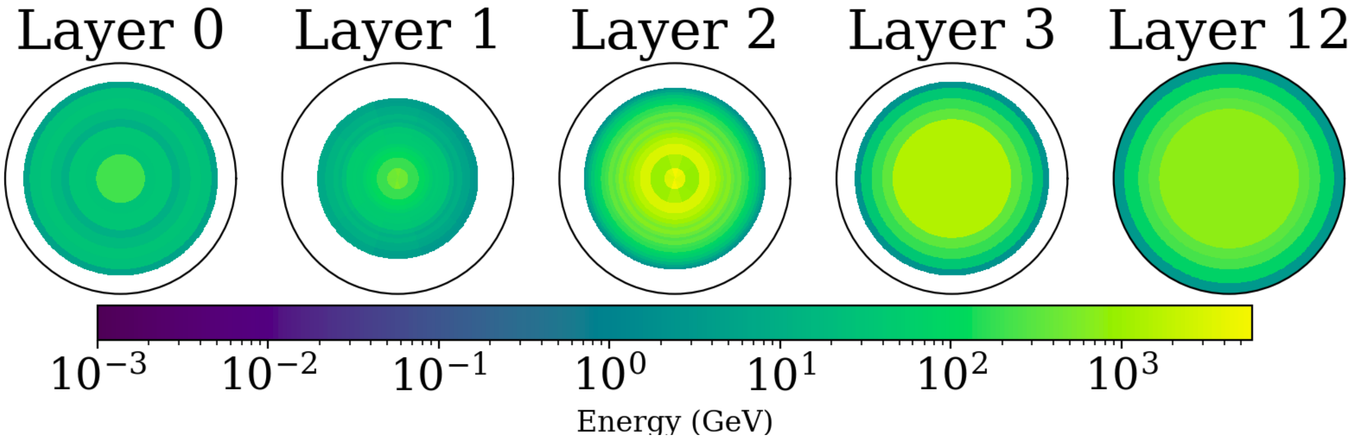
Allow convolutions to be **conditional on R & Z** by using additional channels



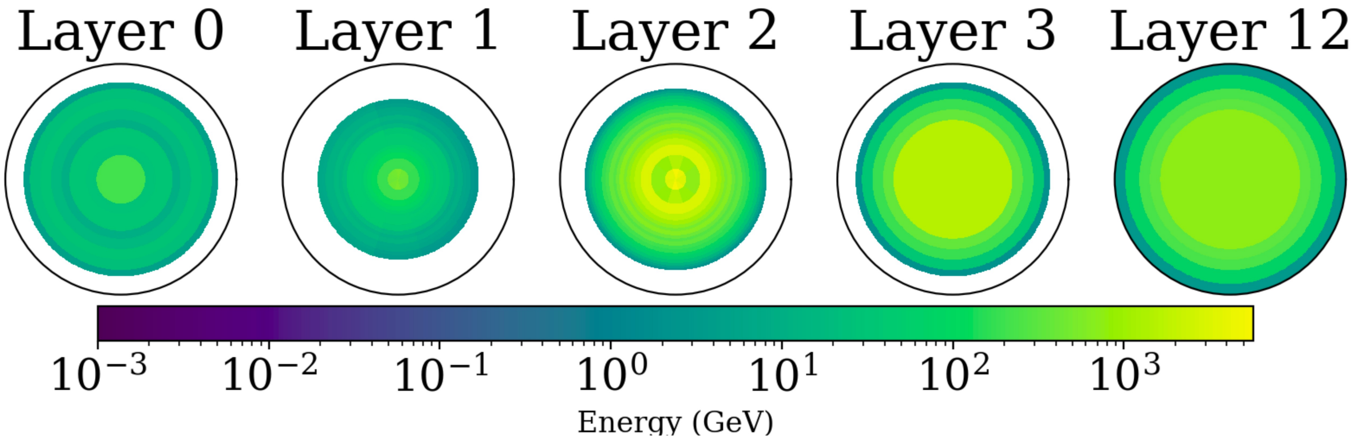
Additional input channels

Average Showers

Geant

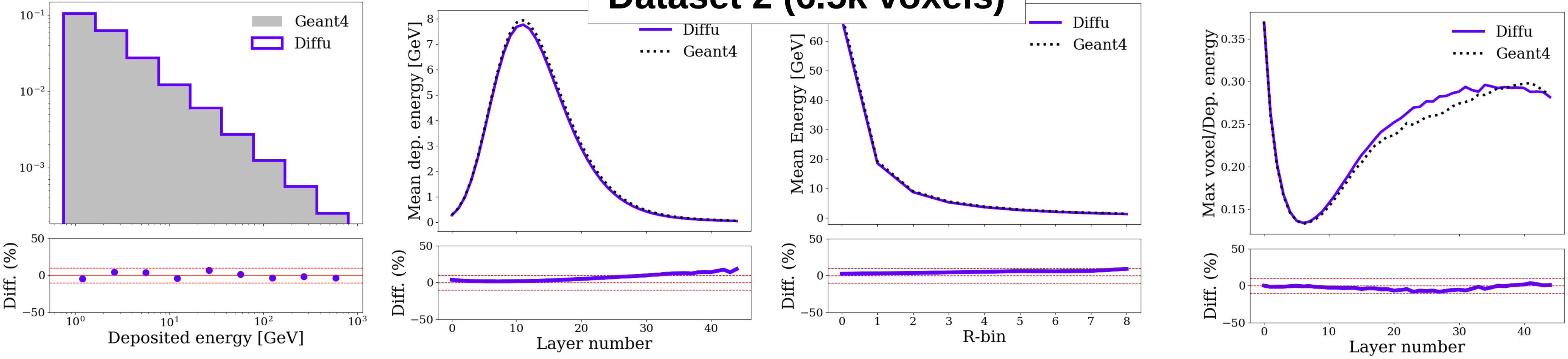


Diffusion

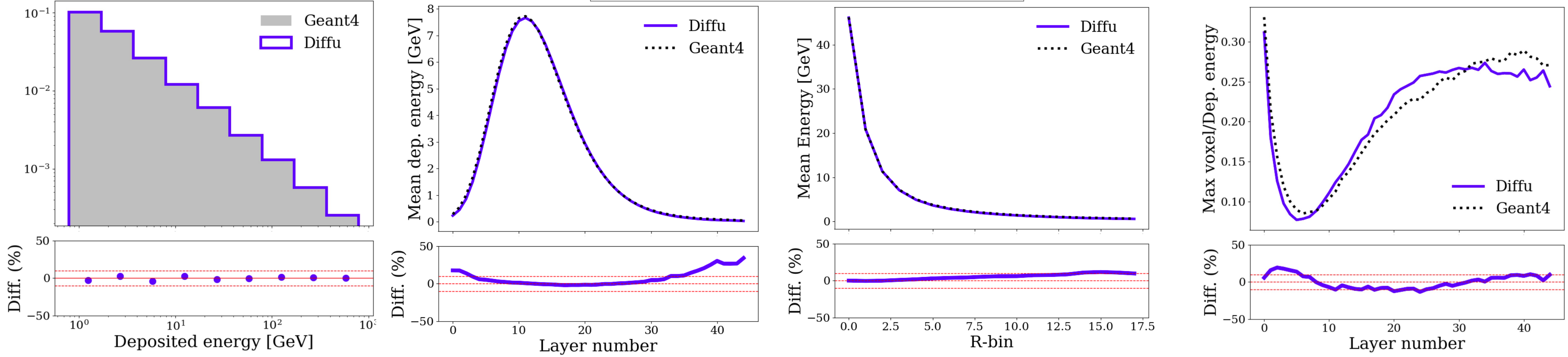


Results: Datasets 2 & 3

Dataset 2 (6.5k voxels)

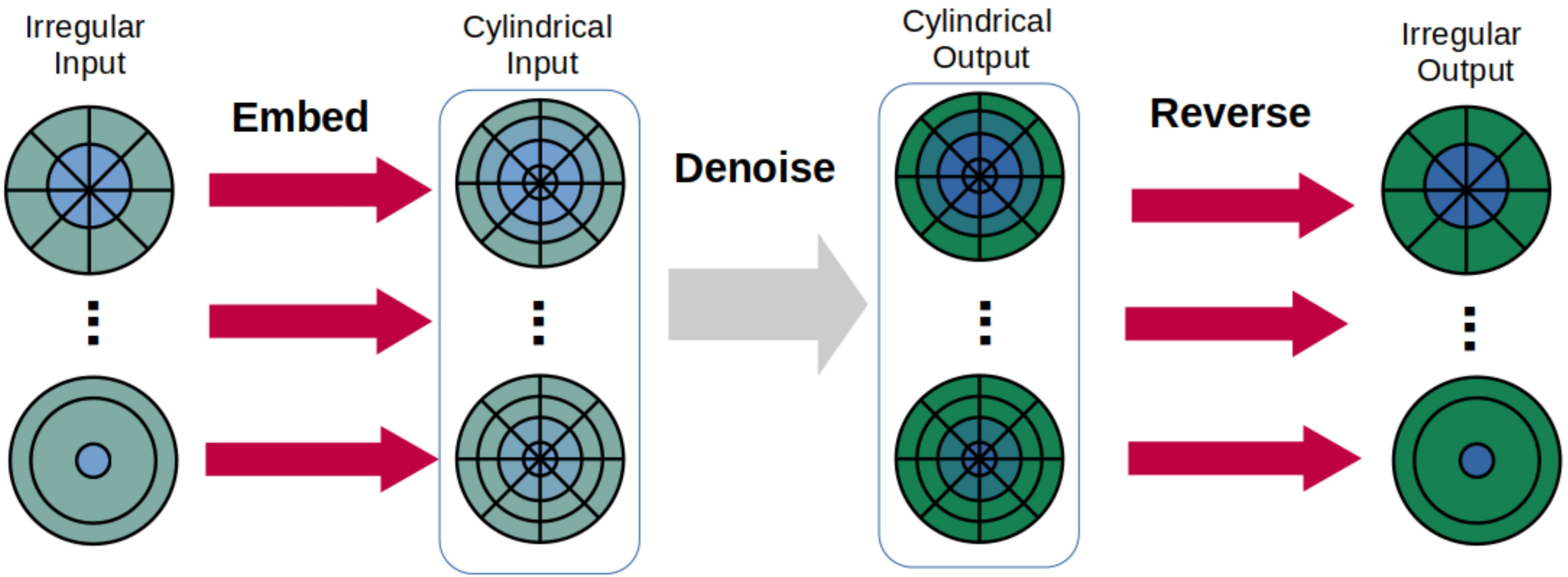


Dataset 3 (40k voxels)

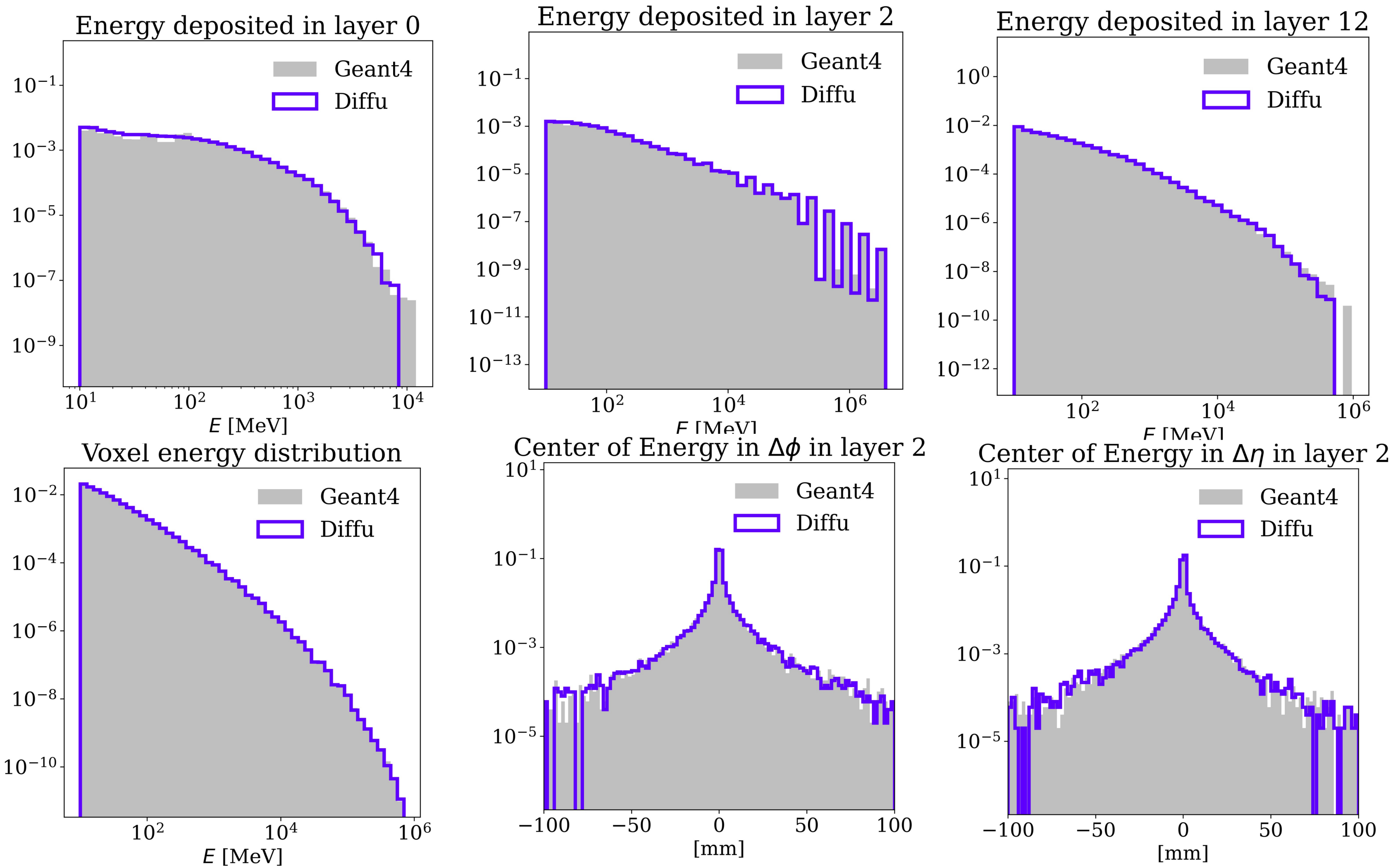


Embedding Irregular Geometries

- Dataset 1 (ATLAS detector) is cylindrical but has **irregular structure** in layers
  - Different radial / angular bins in each layer → can't apply cylindrical convolutions
  - Previous approaches have used fully connect networks or very large 1D CNN's
- Learn an **embedding** that maps input into **regular cylindrical structure**



Dataset 1 Results



Quantifying Performance

- Train a NN classifier to distinguish between Geant showers and CaloDiffusion showers
- Quantify sample quality based on AUC on holdout set)

	Dataset 1 (ATLAS-like)	Dataset 2	Dataset 3
Classifier AUC *	**~0.65**	**~0.6**	**~0.7**

AUC much less than 1 → Very similar showers!

Additional metrics in backup

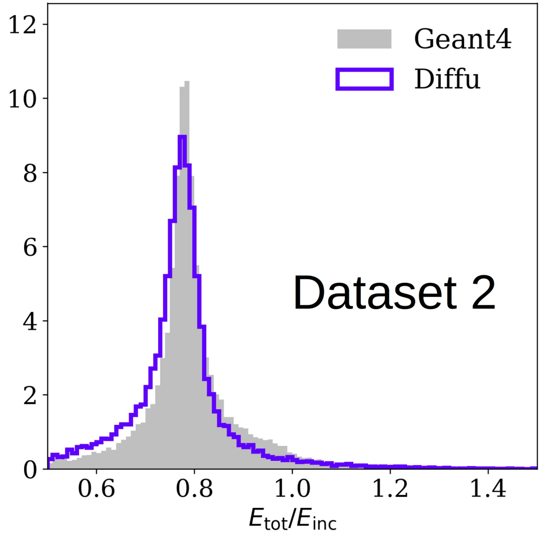
*Preliminary numbers, somewhat dependent on exact classifier training setup

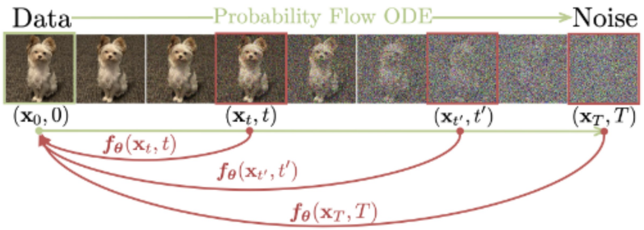
Future Work

- Some "global" properties (ie total shower energy), can still be improved
  - Hard to specifically optimize in diffusion training
  - Will try batch-level MMD loss

- Generation time is slower than other ML approaches b/c of iterated generation (still faster than Geant)
  - Can be improved with different sampling algos, compression, or **distillation methods**
  - Or start generation from **approximate shower** instead of pure noise ("Cold Diffusion", 2208.09392)

- Extend to more complicated geometries e.g. **CMS HGCal**





"Consistency Models" distill diffusion model to allow ~few step generation

Outlook

- CaloDiffusion able to generate very high quality showers
- Utilized several optimizations for **cylindrical** calorimeter geometries & new **embedding** approach for irregular shapes
- Classifiers struggle to distinguish between Geant & CaloDiffusion showers
- Future work: continue to optimize training, improve generation time, more complicated geometries

Acknowledgements

- Co-author : Kevin Pedro
- This work was performed with support of the U.S. CMS Software and Computing Operations Program under the U.S. CMS HL-LHC R&D Initiative.
- Additionla support from the Fermi National Accelerator Laboratory, managed and operated by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the U.S. Department of Energy
- Thank you to the CaloChallenge organizers for the datasets and evaluation code
  - Michele Faucci Giannelli, Gregor Kasieczka, Claudius Krause, Ben Nachman, Dalila Salamani, David Shih and Anna Zaborowska

Thanks!

Backup

Technical Details

- ‘logit’ transformation of voxel energies and then standard scale to zero mean and unit variance
  - Correct preprocessing important for diffusion process, related to scale of added noise
- Denoising network uses ‘U-net’ architecture with cylindrical convolutions
  - Two conditional inputs : shower energy and diffusion step
  - ~400k params for dataset1 and 2, 1.1M for dataset3
- 400 diffusion steps, ‘cosine’ noise schedule (2102.09672)
- Choices for training objective:
  - Datasets 1 and 2 : Network is trained to predict noise component of image
  - Dataset 3 : Network trained to predict weighted average of noise component and un-noised image,
    - More stable, recommended by 2206.00364
- Sampling uses DDPM algorithm (2006.11239)

Additional Metrics

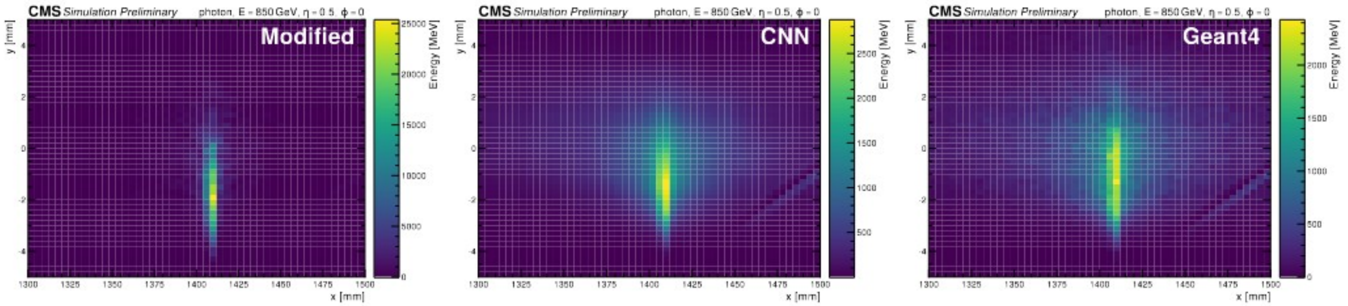
- Distance metrics:
  - Frechet Particle Distance and Kernel Particle Distance (proposed in 2211.10295)
    - Use implementation proposed for CaloChallenge, based on high level shower features
  - We find that the computation of FPD is slightly biased, ie non-zero values even comparing different random samples of Geant to each other
  - Compare scores for Diffu-Geant (D-G) vs Geant-Geant (G-G)

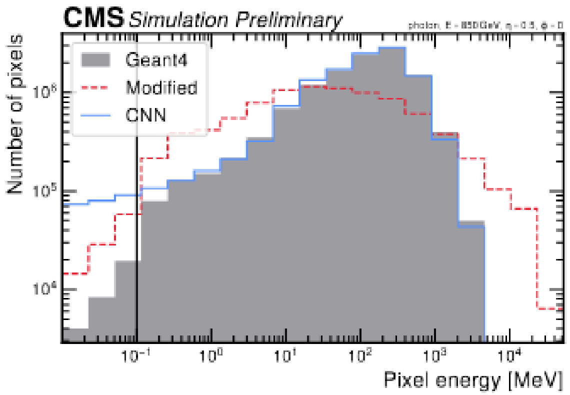
	Dataset 1 (ATLAS-like)	Dataset 2	Dataset 3
FPD (D-G / G-G)	0.035 / 0.008	0.095 / 0.008	0.275 / 0.011
KPD (D-G / G-G)	0.007 / 0	0.0001 / 0	0.0007 / 0

Embedding Details

- First find superset of all radial/angular bins → embedding space
- For each layer, embedding in radial dimension is an M_i x M_* matrix
  - M_i (M_*) is number of radial bins in layer i (embedding space)
  - Initialize weights be proportional to area overlap of bins + 10^-3 * Gaussian noise
- Reverse matrix is M_* x M_i, initialized to pseudo-inverse of embedding matrix
- For now, enforcing phi symmetry, energy is split evenly among phi bins (not learnable)
- Found small benefits of conditioning on phi in addition to R & Z
  - There is slight non-uniformity in phi in the energy distributions of dataset1

Previous Work (arXiv:2202.05320)



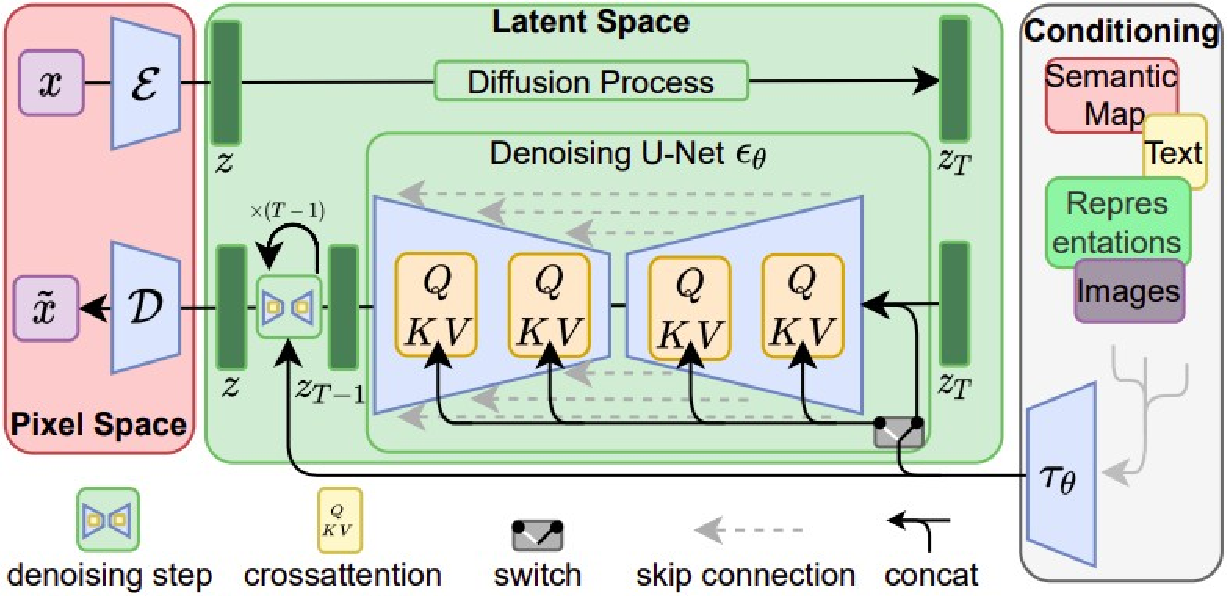


- Generated calorimeter showers with regular & ‘fast’ version of Geant4
- Use a CNN network to ‘denoise’ fast-sim shower image to match high granularity one
- Decent performance in a relatively simple setup
  - Studies showed adding more info to the network beyond ‘energy image’ only moderately improved performance
    - Tried multiplicity, time of energy deposit, other Geant info

Latent Diffusion Models

- Key advantage is that costly diffusion steps done in smaller latent space
- Relies on encoder not losing any important info
  - ‘perceptual loss’ supposed to reduce blurriness
  - Small regularization of latent space (std. normal KL or vector quantization) during AE training
- Conditioning setup very flexible
  - Text prompts using some language model
  - Image conditioning
  - …

Stable Diffusion (aka Latent Diffusion)



2112.10752

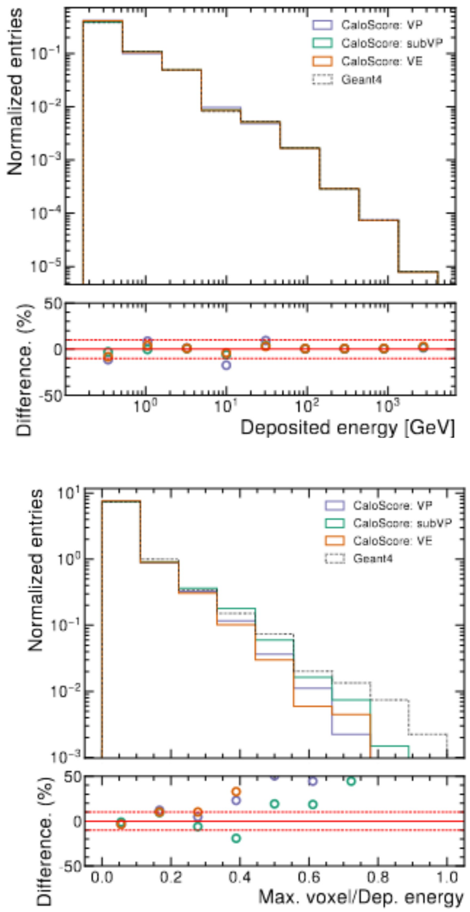
- First encode your image with an autoencoder to a smaller latent space
  - They used a factor of 4 or 8 for each dimm.
- Transform your conditioning data into a latent rep
- Denoising performed on the latent representation of your image, using conditioned data
  - Conditioning done using an attention mechanism
- Decode back into pixel space

Existing work: CaloScore (2206.11898)

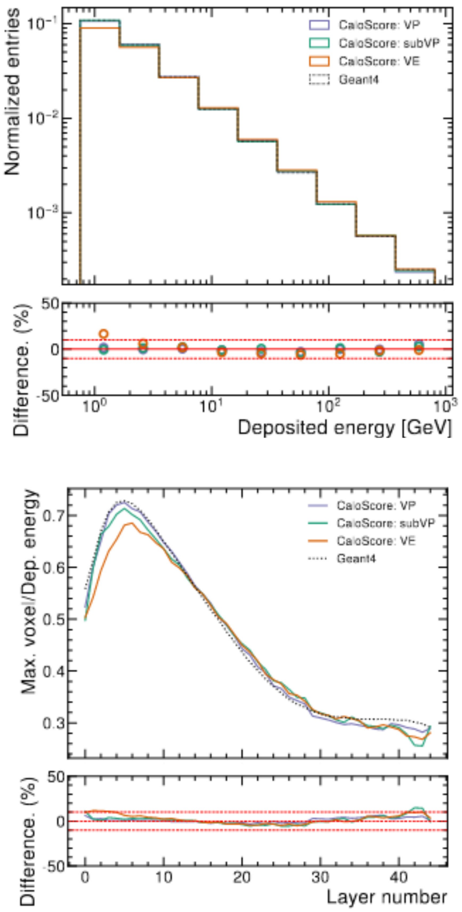
- *Score based* (instead of denoising) diffusion model for calorimeter generation
  - Instead of Gaussian noise, more complicated Markov chain
  - Learn score of the data ($\nabla_x \log(p(x))$ at each iter → can invert process
  - Converted to cartesian geometry (with some loss of information) → no 1 to 1 comparison possible with our work
- Some ML literature showing score based and denoising diffusion are connected
  - See eg Appendix B3 of arXiv:2206.00364

CaloScore Plots

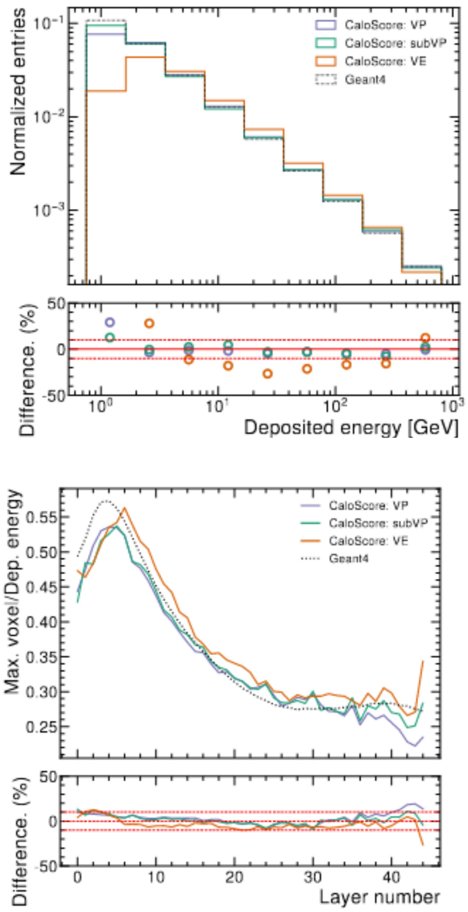
D-1



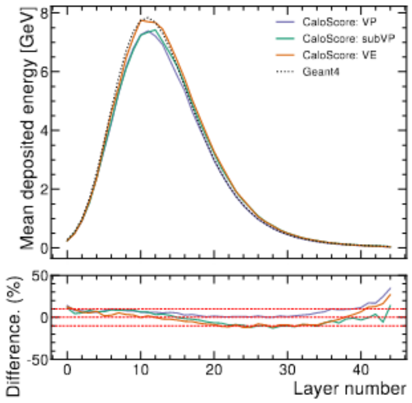
D-2



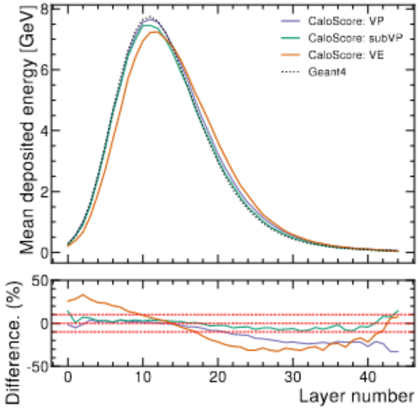
D-3



D-2



D-3



Geometry Diagram