DiT and quantum GAN based fast simulation for the CEPC long-bar crystal electromagnetic calorimeter

Zhihao Li[1,2], Tao Lin[2], Zhengyun You[3],

HaoZhi Yin[3], XiaoZhong Huang[2], Hideki Okawa[2], Weidong Li[1,2]

[1] University of Chinese Academy of Sciences, China

[2] Institute of High Energy Physics, Chinese Academy of Sciences, China

[3] Sun Yat-Sen University, China

Outline

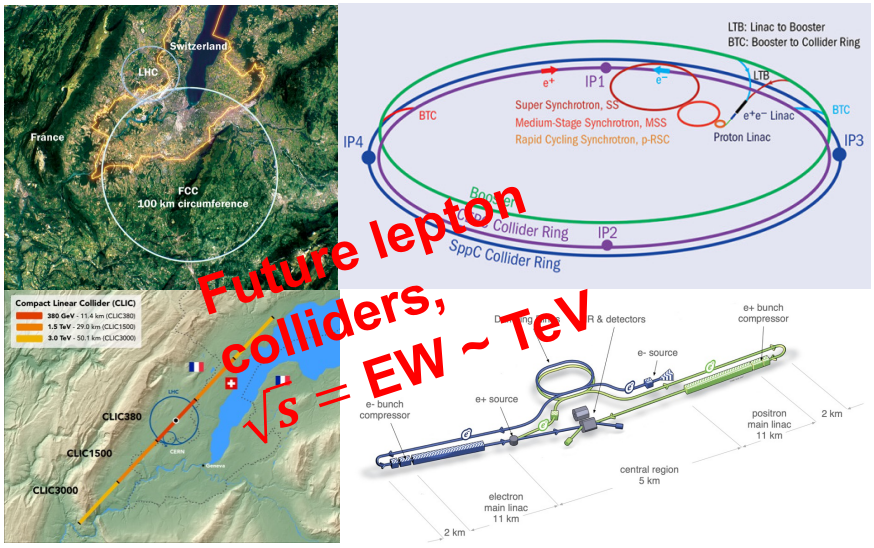
- Motivation
- DiT-based fast simulation method
- Quantum GAN
- Summary

Outline

- **Motivation**
- DiT-based fast simulation method
- Quantum GAN
- Summary

Future lepton collider

- Physics after Higgs discovery:
  - Precise measurement of Higgs, EW, top, flavor, QCD…
  - BSM physics (dark matter, EW phase transition, SUSY, LLP…)

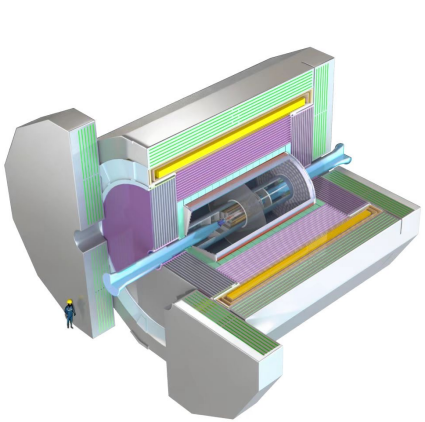


arXiv:2510.05260 arXiv:2312.14363

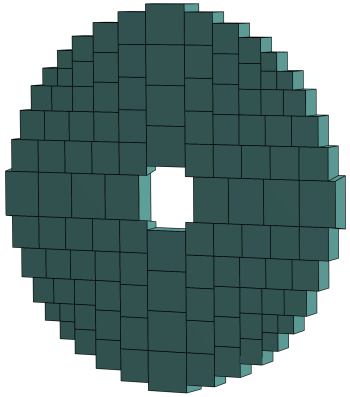
CEPC ECAL

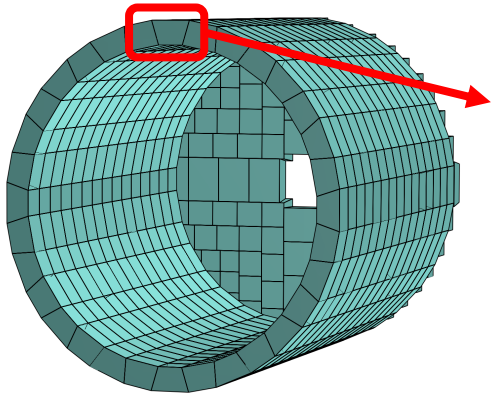
- **ECAL:** The long-bar crystal ECAL forms an interleaved three-dimensional mesh that delivers about 3% energy resolution with fine imaging.

$$\frac{\sigma_E}{E} \approx \frac{3\%}{\sqrt{E/GeV}}$$

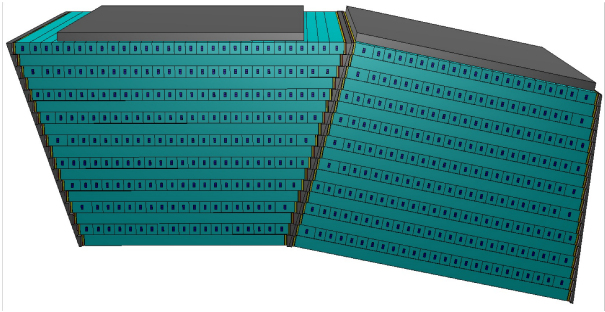


Reference Detector



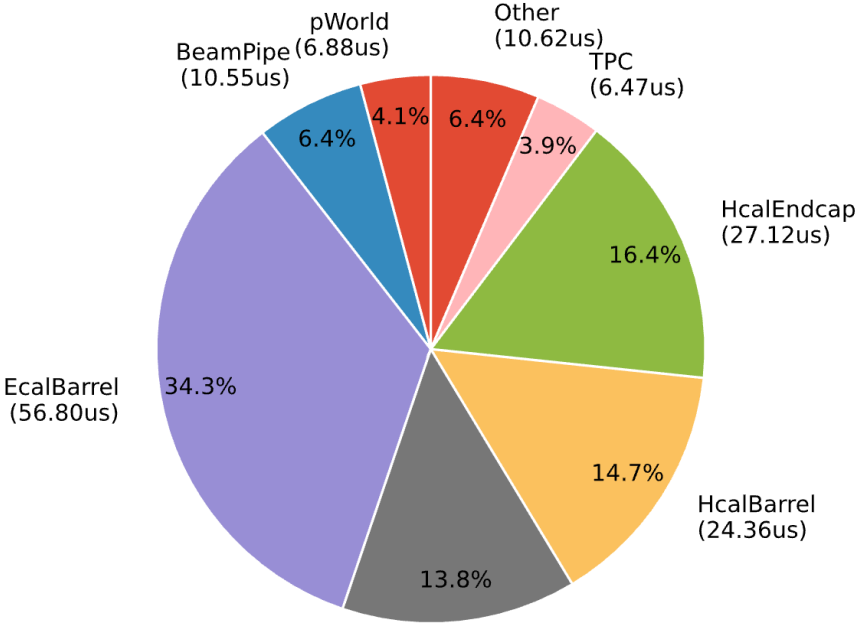


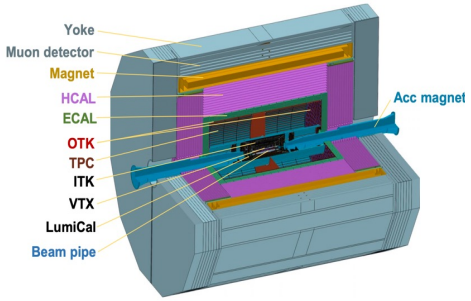
ECAL

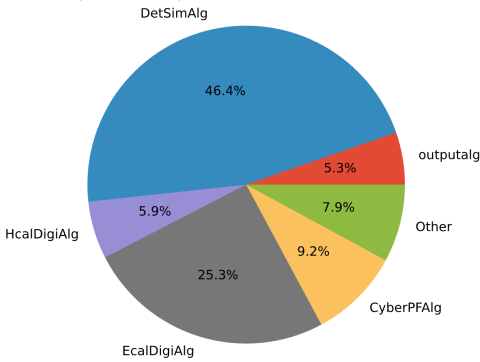


CEPCSW Computing Bottlenecks

- **Geant4 and ECAL Simulation Bottlenecks**: For $e^+e^- \rightarrow q\bar{q}$ at 240 GeV the GEANT4 step dominates 46.4% of the wall time; ECAL barrel and endcaps alone consume 34.4% and 13.8% of the total simulation budget, respectively.
  - Electromagnetic showers generate excessive secondary particles (high tracking multiplicity).
  - Targeted optimization for ECAL/HCAL only; keep standarc Geant4 elsewhere.







Outline

- Motivation
- **DiT-based fast simulation method**
- Quantum GAN
- Summary

DiT II: Diffusion Process

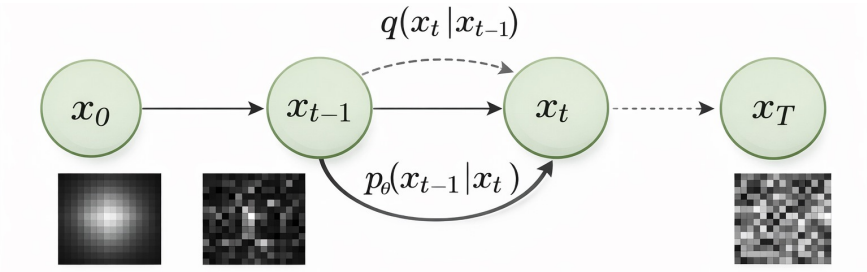
- **Forward diffusion process**
  - Starting from original data x_0, Gaussian noise is gradually added in steps to produce a sequence

    $$x_0 \rightarrow x_1 \rightarrow \cdots \rightarrow x_T$$

    where x_T approaches an isotropic Gaussian distribution as T.
- **Reverse diffusion process**
  - The model learns a parameterized denoising distribution

    $$p_\theta(x_{t-1} \mid x_t)$$

    which predicts how to iteratively remove noise from x_t to recover x_{t-1} effectively reversing the forward process.



DiT III: Dataset

Data Preprocess

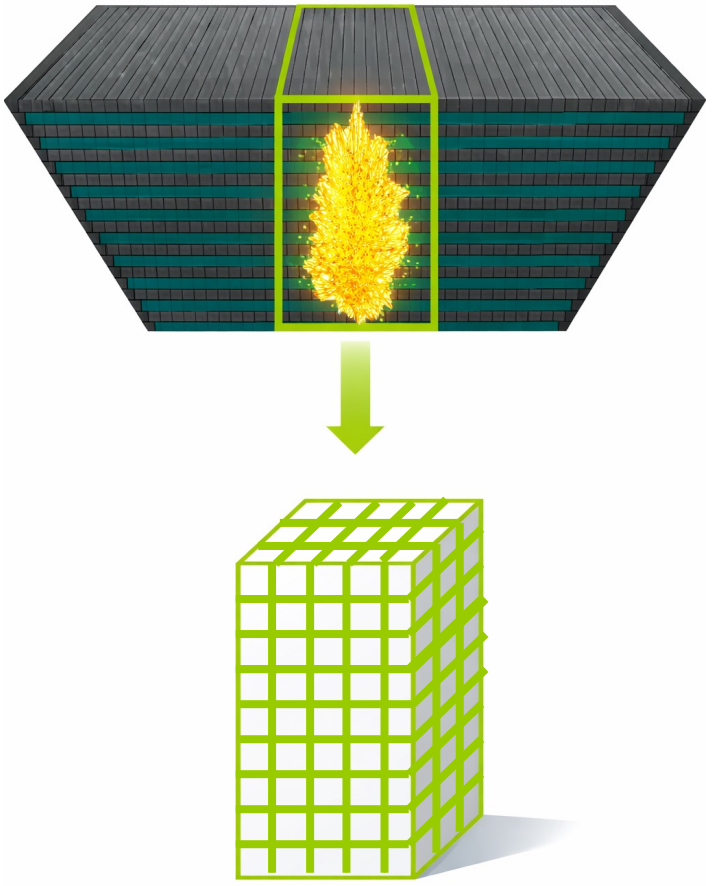
- Geant4 photon simulations based on CEPC Ref-TDR geometry.
- Segmented with DD4hep into a 15×15×18 voxel grid (15 mm cube size).
- Log-transformed and normalized voxel energies.

$$\hat{x_i} = \frac{\log(x_i + \epsilon) - \mu}{\sigma} \quad \text{where}$$

$$\mu = \mathbb{E}[\log(x_i + \epsilon)]$$

$$\sigma = \sqrt{\mathbb{E}[(\log(x_i + \epsilon) - \mu)^2]}, \quad \text{and}$$

$$\epsilon = 10^{-7}\ GeV$$

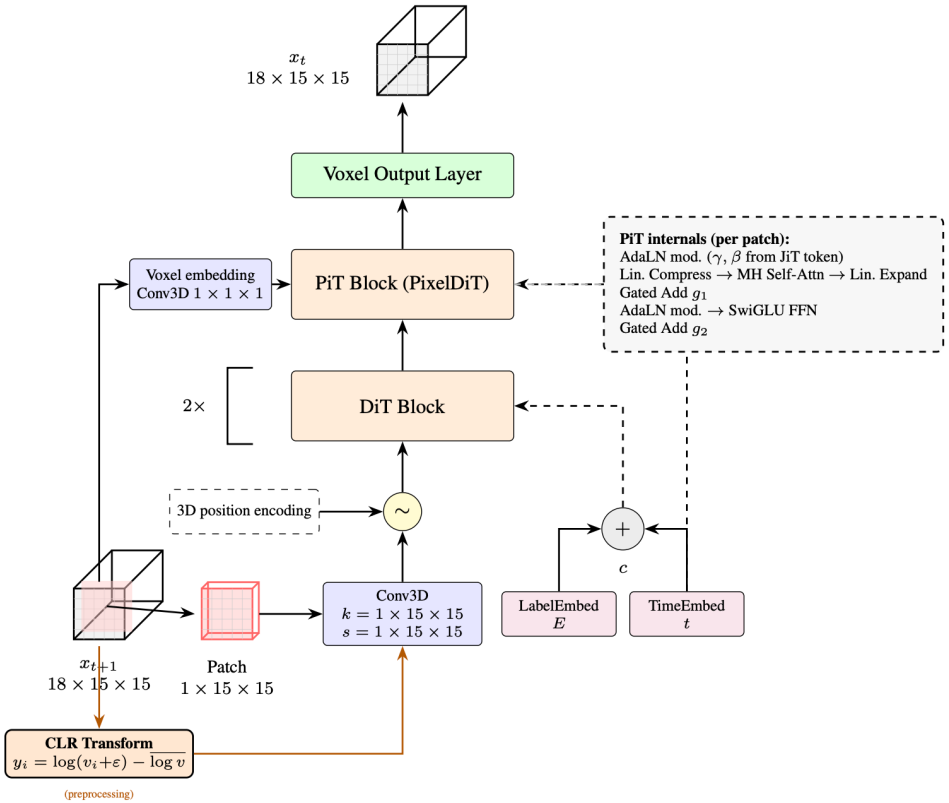


Area of interest selected as dataset

DiT IV: Model Architecture

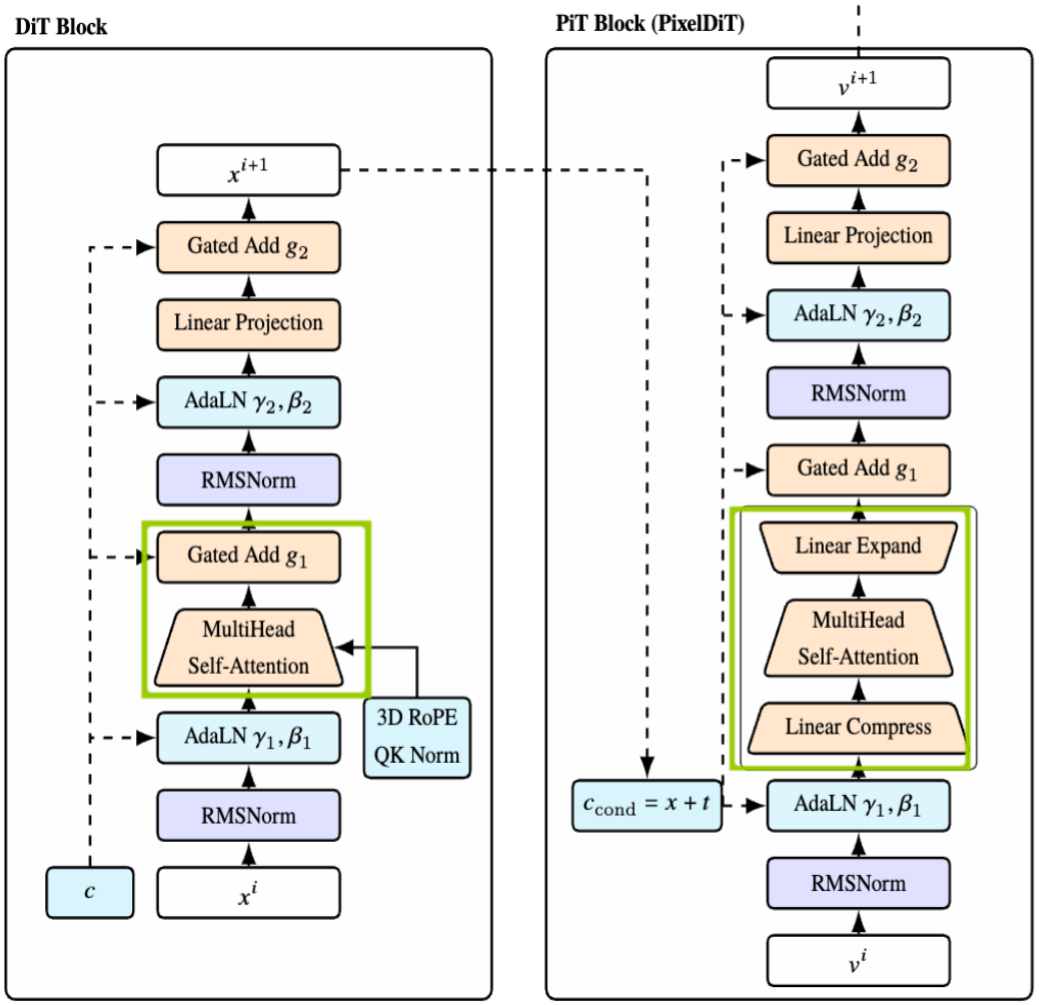
VoDiT4CAL Architecture:

- **DiT Blocks:** Transformer self-attention blocks exchange information acros tokens globally.
- **PiT Blocks**: Transformer self attention blocks exchange information within each token across voxels.
- **Voxel Output**: The final token representations are mapped back to energy space to produce voxel-leve predictions.



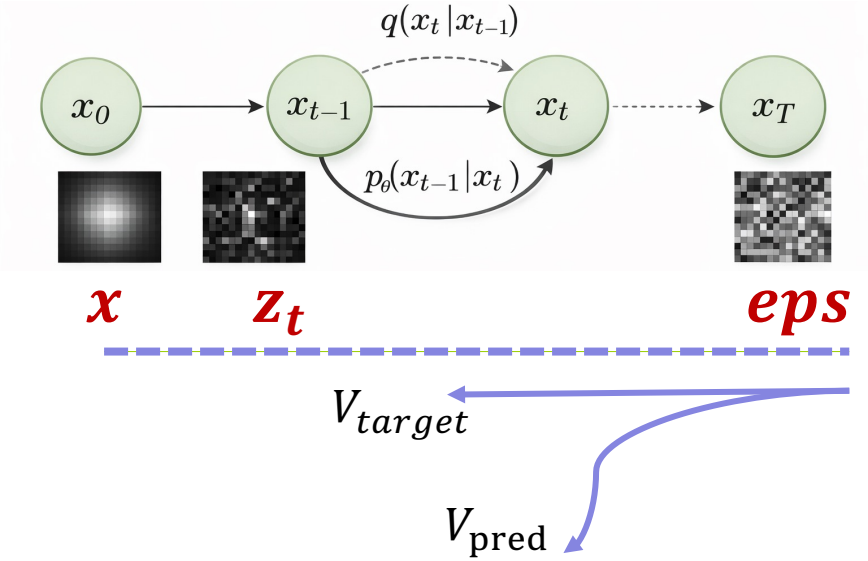
DiT IV: Model Architecture

- **DiT Block**:
  - Direct multi-head attention
  - AdaLN conditioning + 3D RoPE
- **PiT Block**:
  - Linear Compress → Attention → Expand
  - Same AdaLN/gating, optimized for high-res pixel-space



DiT V: Training

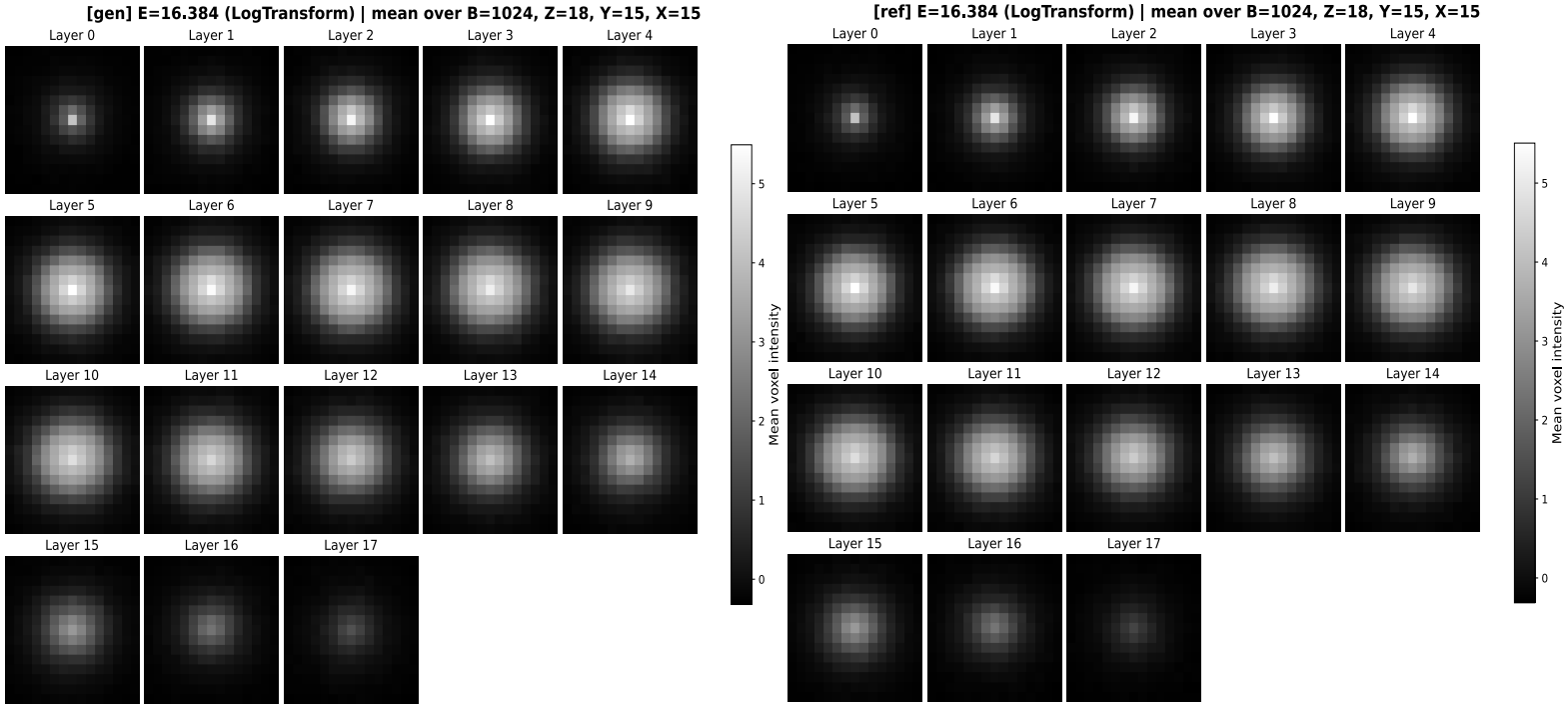
- **Training:**
  - **PyTorch** provides tensor computation and automatic differentiation, while **PyTorch Lightning** offers a high-level training framework.
  - **Training Time:** The total effective training time is **3h23m19s** on a single RTX 5090 GPU.



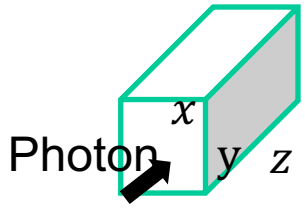

```
t = sample_t()
eps = randn_like(x) * noise_scale
z_t = (1 - t) * eps + t * x

v_target = (x - z_t) / (1 - t)
v_pred = model(z_t, t, E_cond)
loss = mse(v_pred, v_target)
```

Result I: Generation Quality

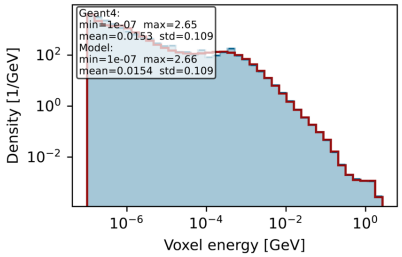
- Generate photon shower at **16 GeV** for example

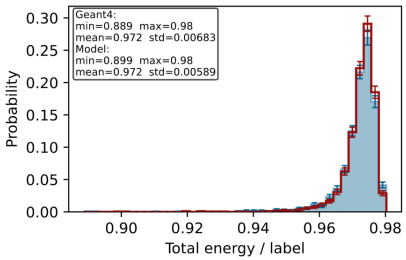


Result I: Generation Quality

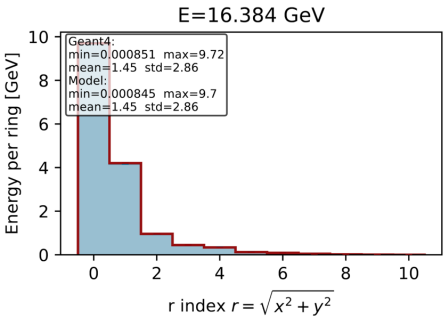


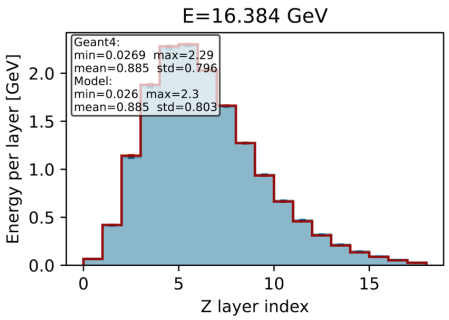
Overview



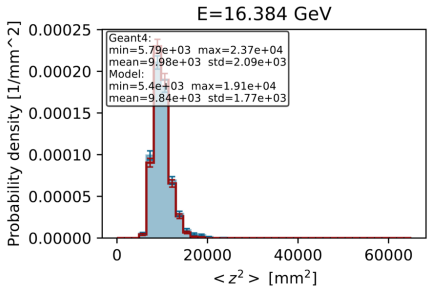


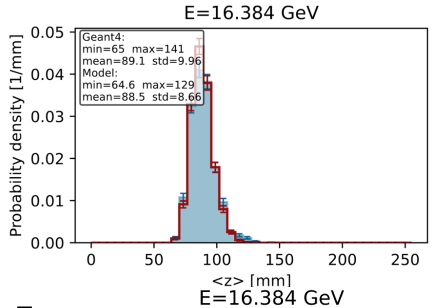
Geometry

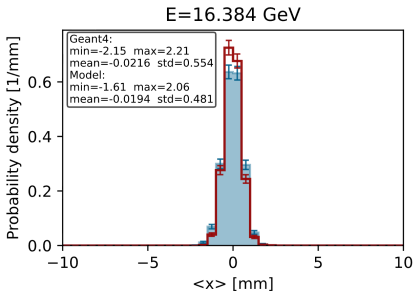


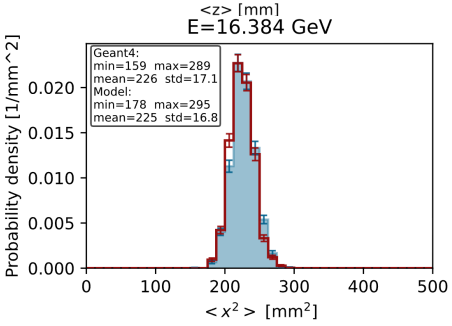


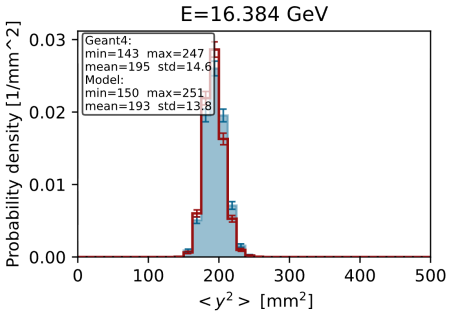
ZXY energy center

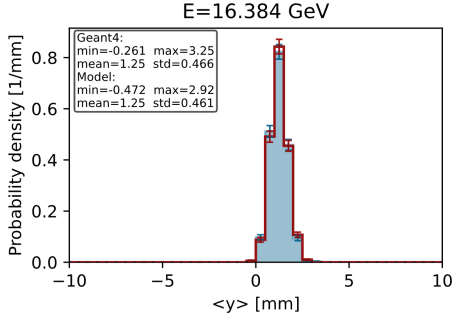






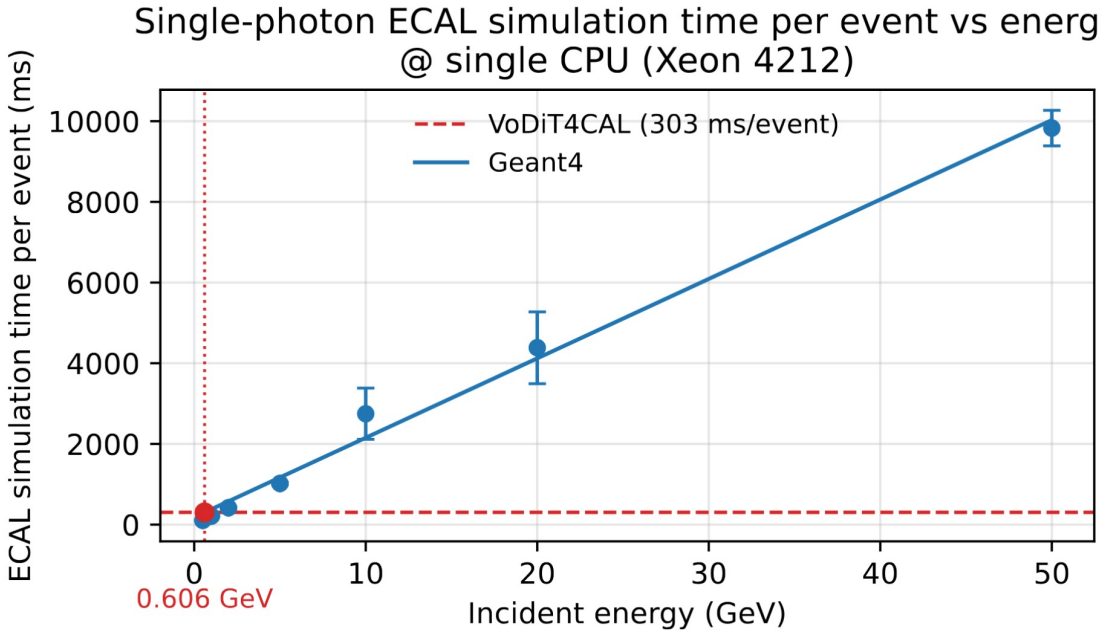






Result II: Generation Speed

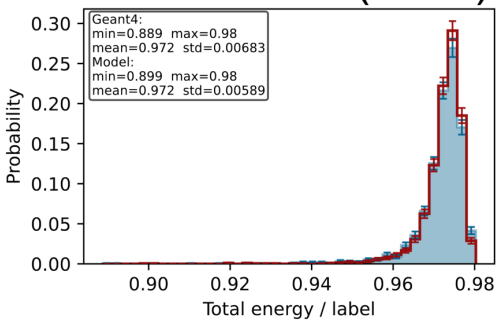
- Benchmark:
  - GPU (Nvidia RTX 8000): **3.19 ms /event**
  - CPU (Xeon 4214, single thread): **303.45 ms /event**
  - Geant4: generation time $T \sim E$ with particle energy E



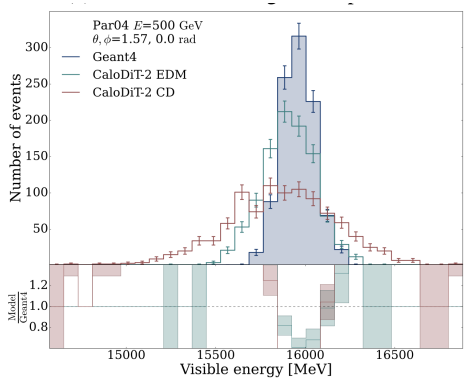
Result III: Compare with CaloDiT-2

- **Foundation**: Extends CaloDiT-2 thought to CEPC calorimetry
- **CaloDiT-2 Limitations & Our Solutions:**
  - Energy linearity issues: Improved via **CLR Data Transform**
  - Deep/slow architecture: Accelerated by PiT blocks
- **Scope Note:**
  - Direct comparison with original CaloDiT-2 not performed: dataset mismatch (CaloChallenge benchmarks vs. CEPC simulation Data)

VoDiT4CAL (Ours)



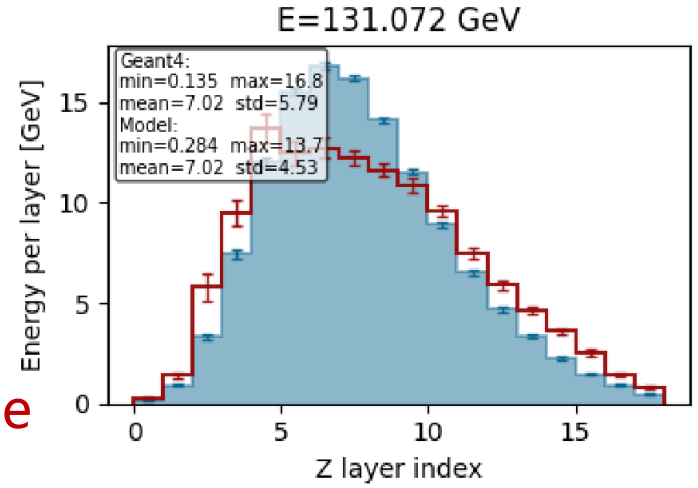
CaloDiT-2



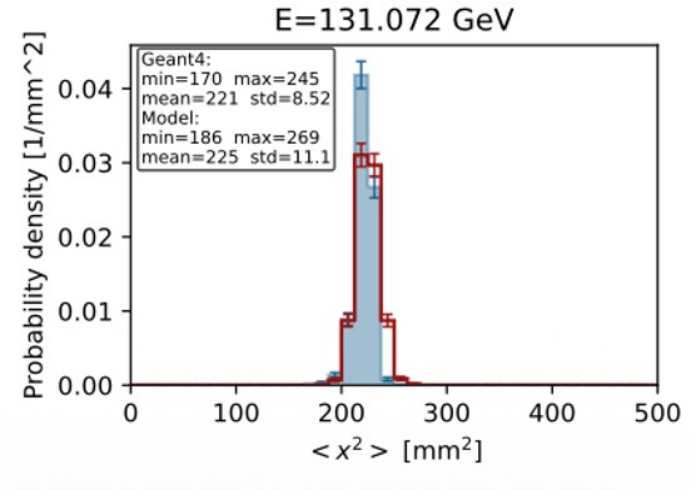
Models \ Timings	CPU (ms)	GPU (ms)
CaloDiT-2 EDM	6349.0± 7.6	171.8± 0.13
CaloDiT-2 CD	101.2± 0.04	2.9± 0.02
CaloDiT-1 DDPM [28]	17322.9± 33.9 24642± 1883*	639.96± 2.4 1036± 18*

Limitation & Open Challenges

- **Quality Degradation After Distillation**
  - Performance drops observed post-distillation
  - Exploring potential solutions: guidance loss



- **Variance Instability**
  - Fluctuating variance $< x^2 >$ across generations
  - Suspected cause: limited diversity in generated samples (mode collapse tendency)



Outline

- Motivation
- DiT-based fast simulation method
- **Quantum GAN**
- Summary

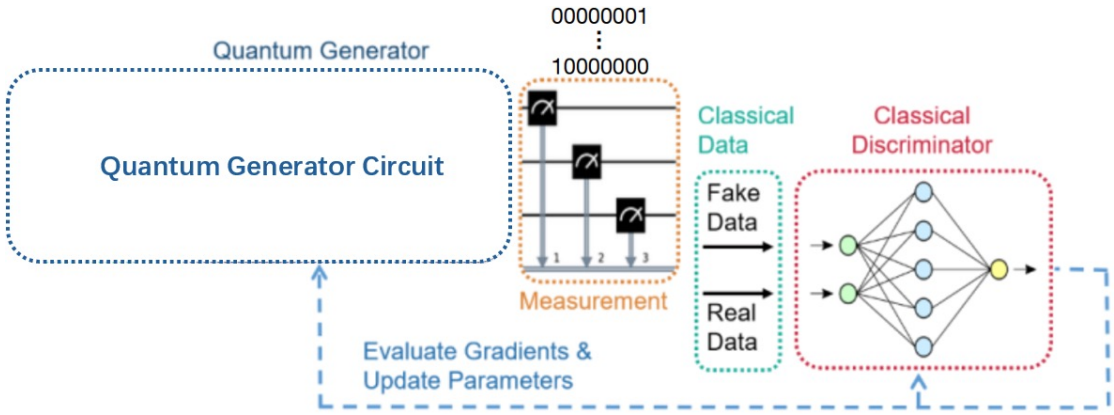
Quantum Motivation

- Our classical results have already shown that **DiT offers stronger stability, generation quality, and scalability than GAN.**
- the natural next question is: under the same diffusion/Transformer framework, **quantum modules can further improve global modeling?**
- **Early studies on quantum diffusion already suggest that this direction is promising**, providing initial evidence that quantum-enhanced diffusion models are worth exploring beyond qGAN.
- For example, **QGDM** reports faster convergence than QGAN and higher fidelity in mixed-state generation, while **Quantum Latent Diffusion Models** show better image-generation metrics than a comparable classical model and retain advantages in few-shot settings.

arXiv:2401.07039. arXiv:2501.11174

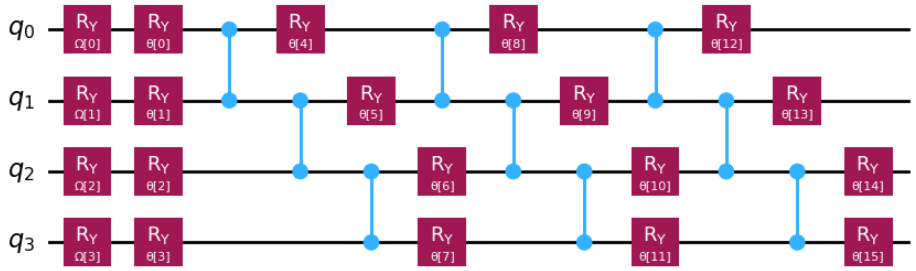
Quantum GAN for FastCaloSim

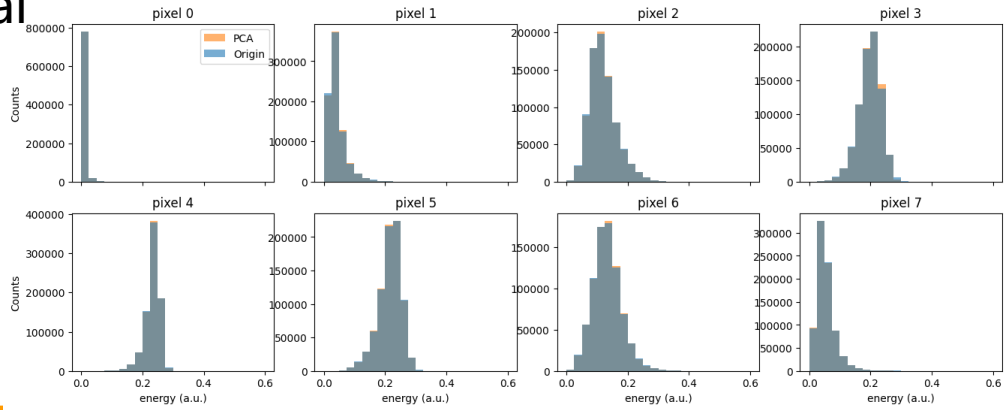
- Motivation: Quantum GANs in Calorimeter Simulation
  - Calorimeter simulation is one of the most CPU-intensive tasks in HEP.
  - Overcoming computational challenges through the integration of ML and QC.
- Quantum Generative Adversarial Network(GAN)
  - GAN is composed of two adversarial neural networks.
  - NISQ constraints necessitate hybrid classical-quantum algorithms.
  - The quantum GAN in this study adopts a quantum generator + classical discriminator.



Quantum GAN for FastCaloSim

- Quantum Generator Circuit
  - A generator circuit composed of 4 qubits.
  - RY gates contain trainable parameters.
- Training Data and Preprocessing
  - Simulated electron showers from Geant4 at a fixed incident angle.
  - Processed into one-dimensional 8-pixel distributions as training data for the GAN.
  - Input training data into the quantum circuit via PCA.

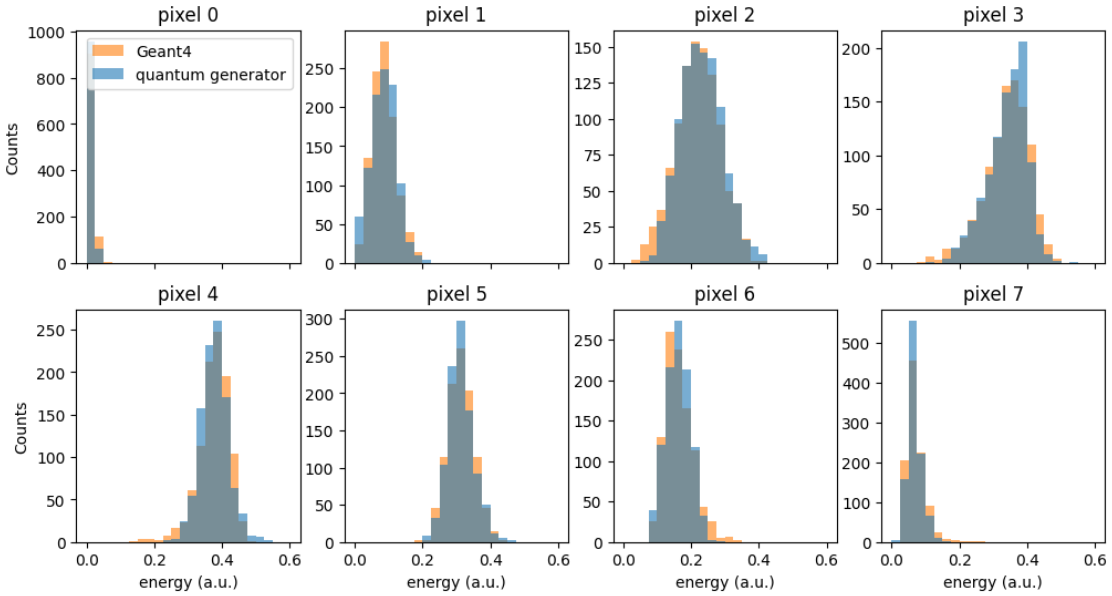


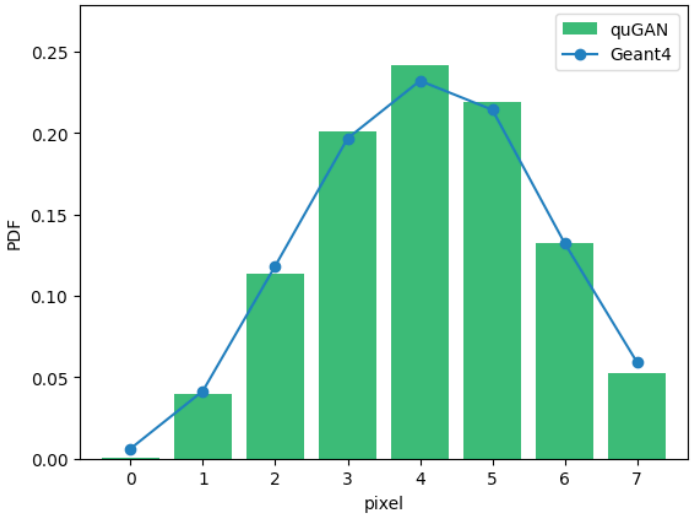


Quantum GAN for FastCaloSim

Quantum GAN Training Results

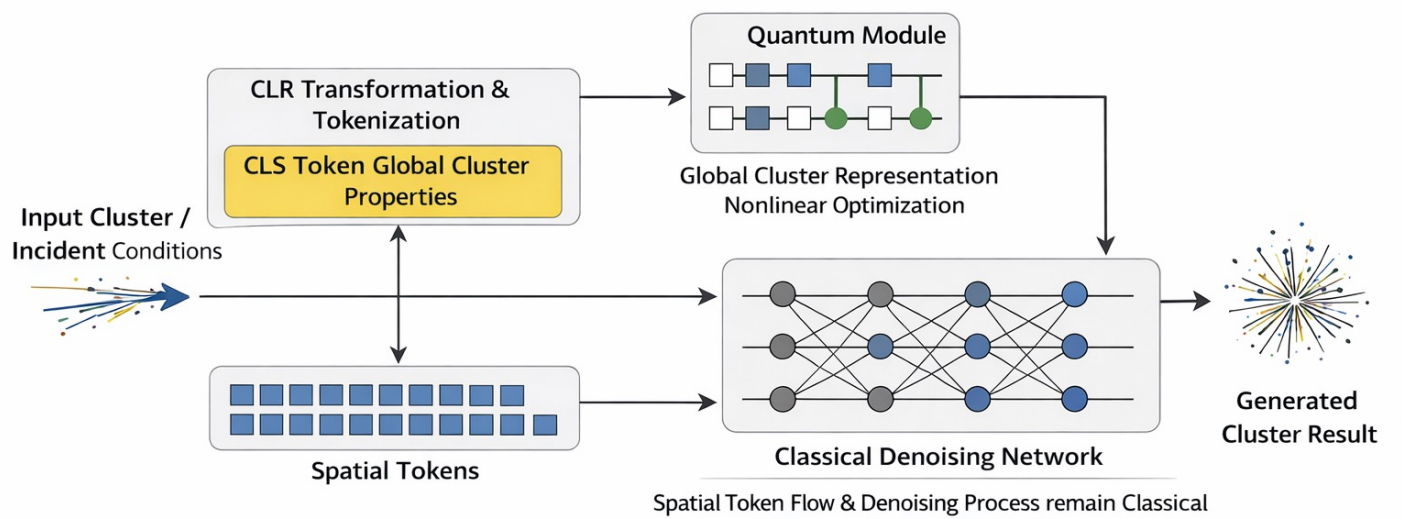
- Generated data is consistent with Geant4





Ref XiaoZhong Huang, Quantum GAN for fast shower simulation

Future Plan: Quantum-DiT hybrid model



- Current DiT model uses **CLR transform**; a dedicated **CLS token** captures global shower properties.
- The quantum module operates **exclusively on the CLS token**, performing nonlinear renormalization of the global shower representation.
- Spatial token flow and backbone denoising remain **fully classical** — minimizing quantum overhead while preserving generation fidelity
- CLS and spatial tokens are **jointly trained end-to-end**, enabling the quantum module to meaningfully influence the denoising process

Outline

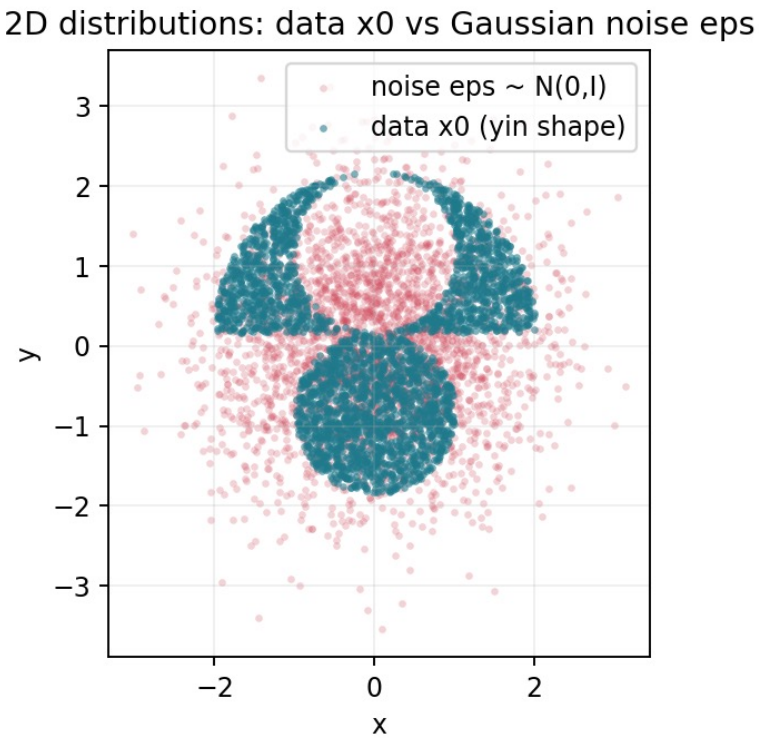
- Motivation
- DiT-based fast simulation method
- Quantum GAN
- **Summary**

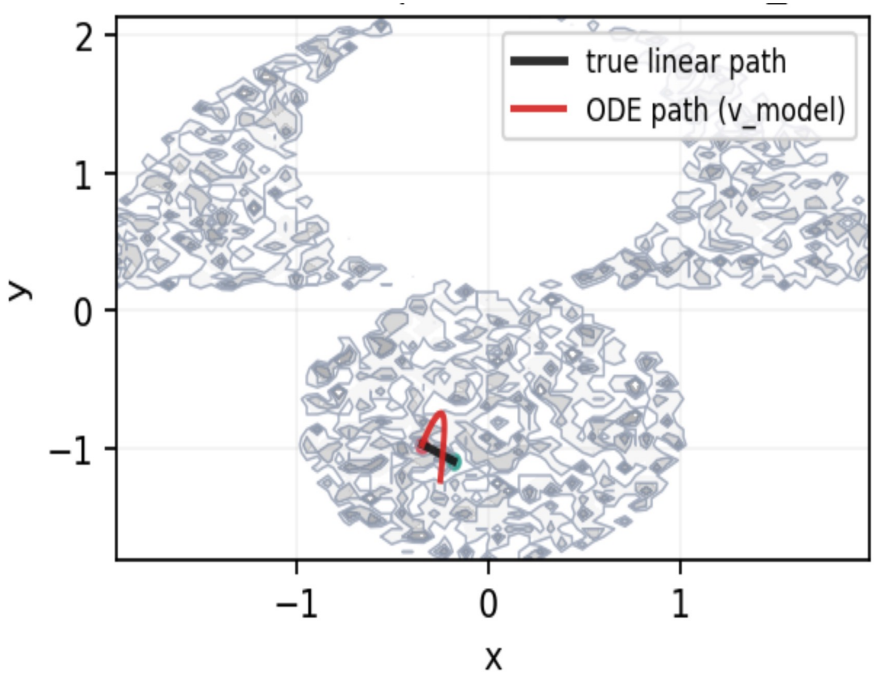
Summary

- VoDiT4CAL achieves high-precision fast simulation of the CEPC crystal ECAL at **~100ms** level.
  - While preserving the quality of showers and key physical observables, it delivers approximately a **100× speedup** for 50 GeV events on a single CPU core.
  - This provides a practical solution to **mitigate computing bottlenecks** and paves the way for larger-scale physics studies.
  - Will apply distillation techniques to **reduce inference time** while maintaining generation quality; Will extend to **multi-particle** and **more complex incident** conditions
  - Will try to explore **Quantum Transformer**
  - Will ultimately **integrate it into CEPCSW framework** to effectively replace the Geant4 ECAL simulation module

Thank you!

TOY Data Demo





CLR Transform

- For every sample x_0:

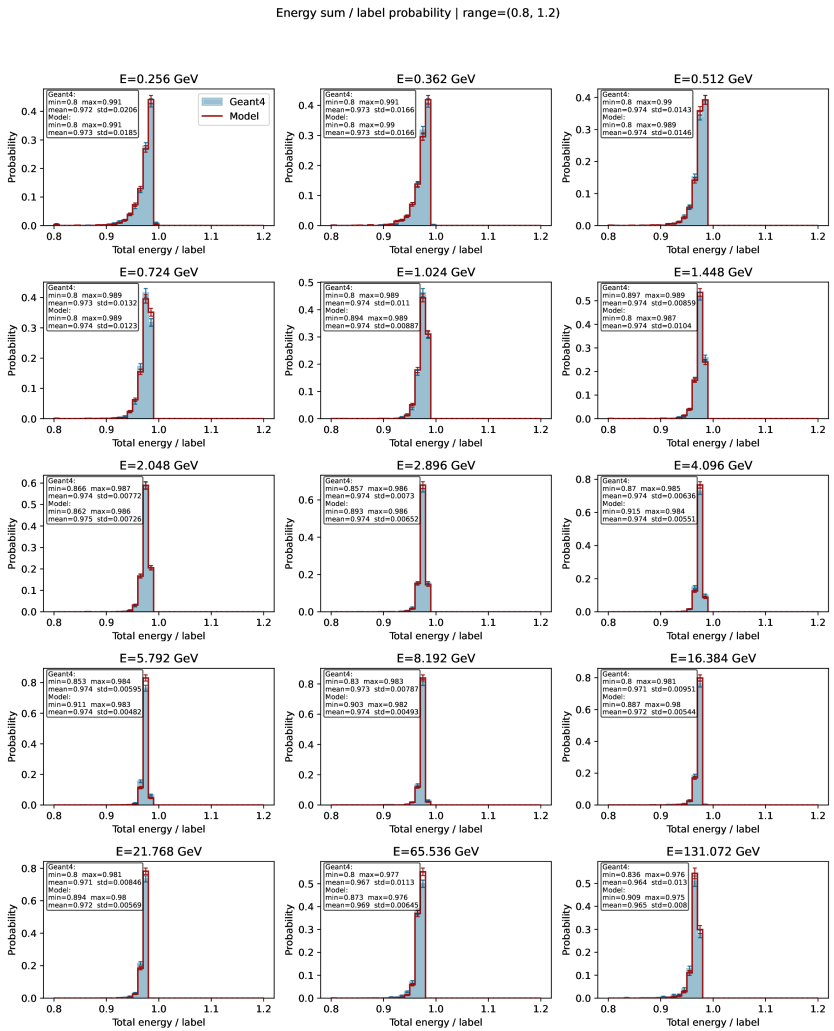
$$g = \exp\left(\frac{1}{N}\sum_{j=1}^{N}\log(x_{0,j}+\epsilon)\right)$$

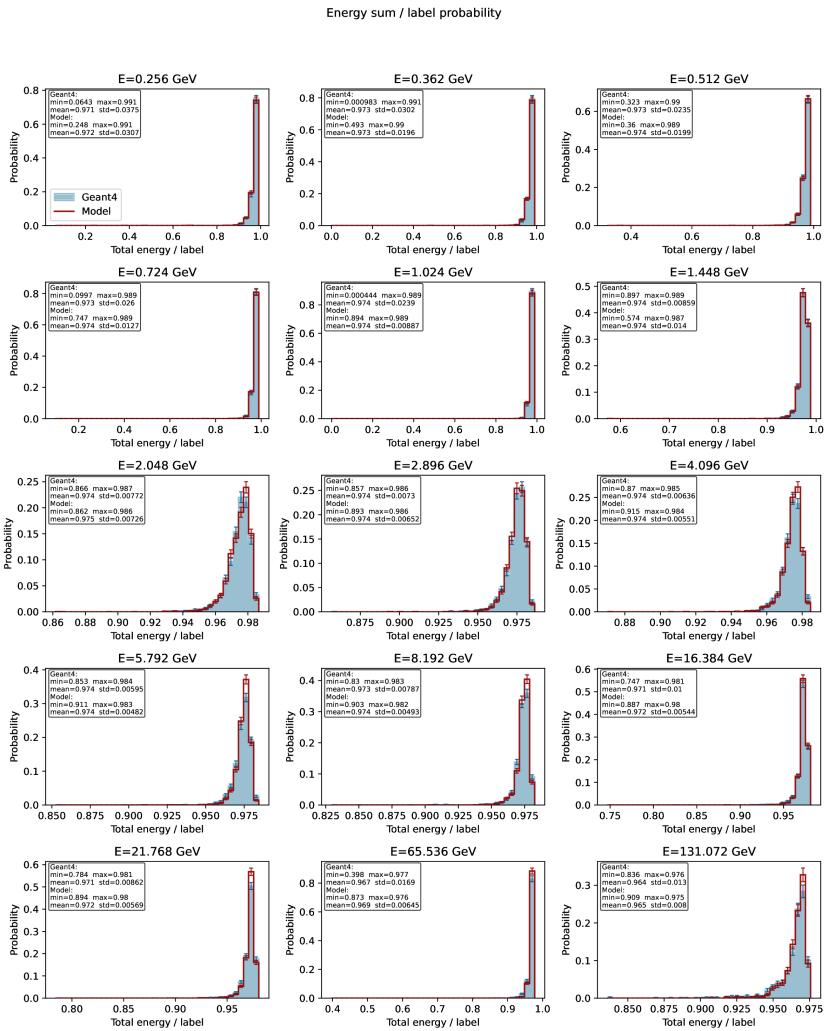
$$y_i = \log\left(\frac{x_{0,i}+\epsilon}{g}\right) = \log(x_{0,i}+\epsilon) - \frac{1}{\mathrm{N}}\sum_{j=1}^{N}\log(x_{0,j}+\epsilon)$$

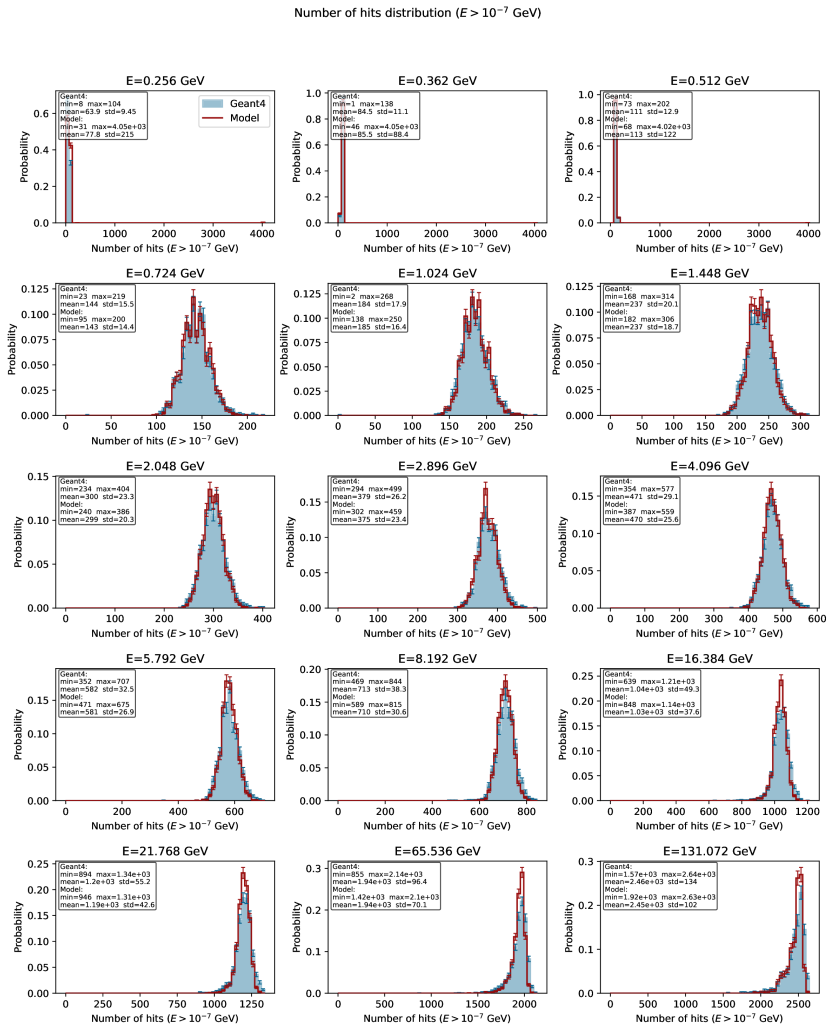
- y In the zero-sum subspace (composition-preserving)

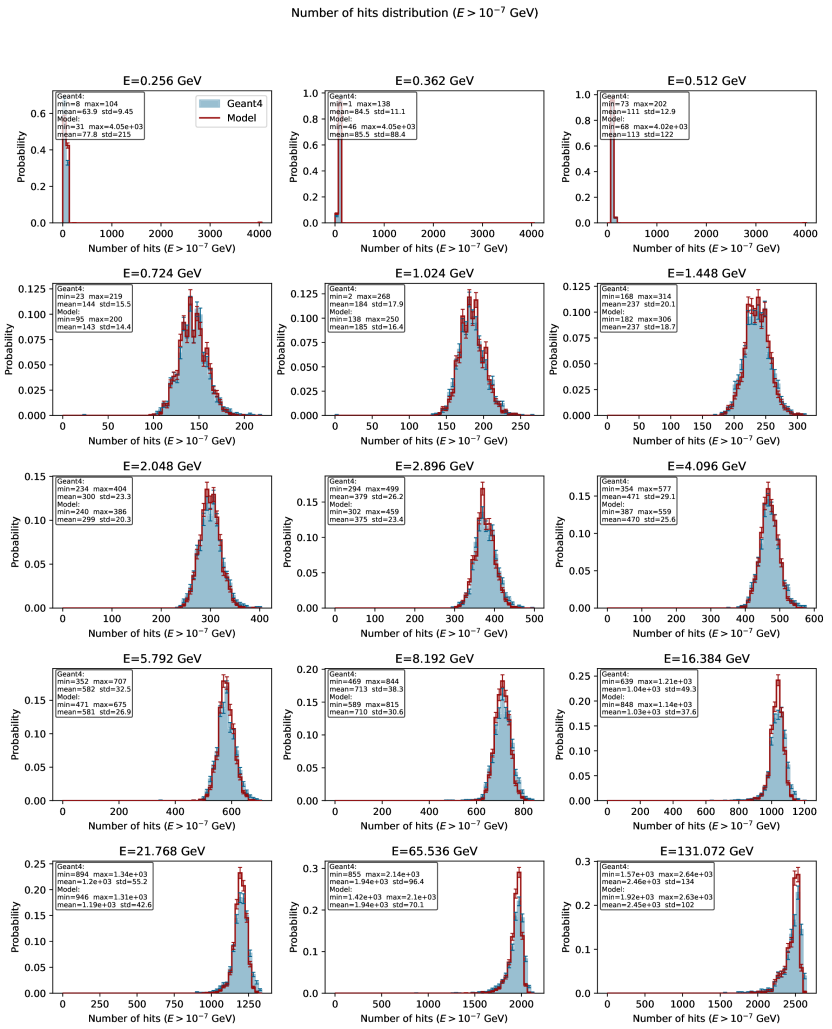
$$\sum_i y_i = 0$$

SUM

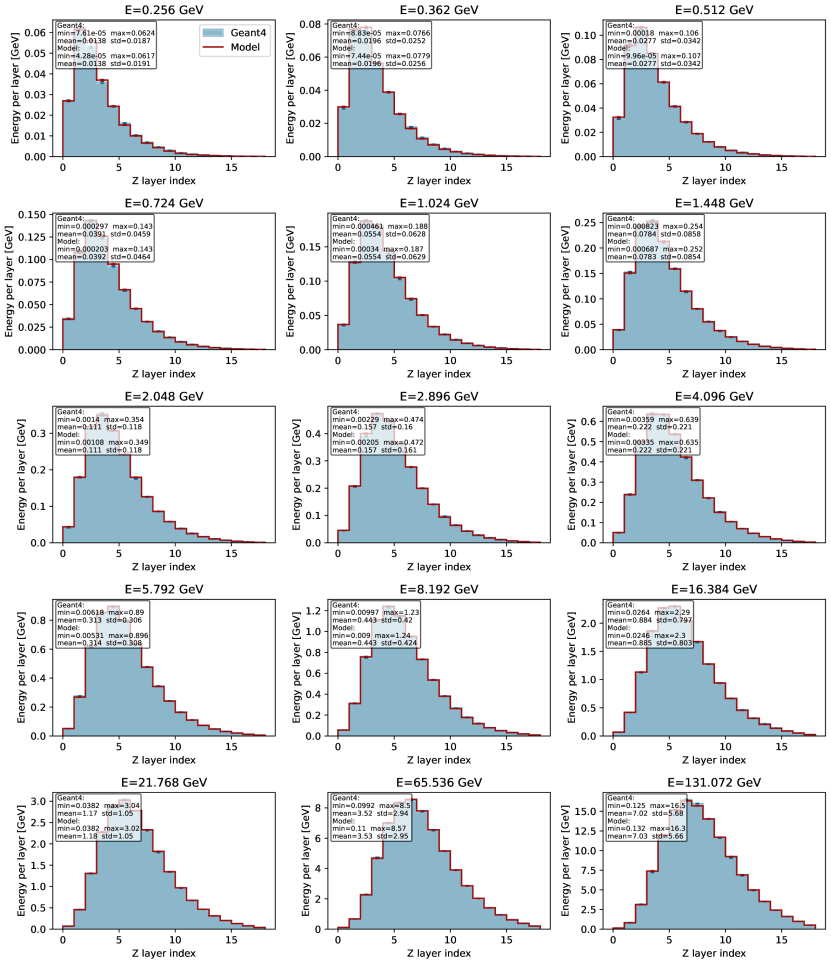

Energy sum / label probability | range=(0.8, 1.2)
E=0.256 GeV
E=0.362 GeV
E=0.512 GeV
E=0.724 GeV
E=1.024 GeV
E=1.448 GeV
E=2.048 GeV
E=2.896 GeV
E=4.096 GeV
E=5.792 GeV
E=8.192 GeV
E=16.384 GeV
E=21.768 GeV
E=65.536 GeV
E=131.072 GeV
Geant4
Model
Probability
Total energy / label


Energy sum / label probability
E=0.256 GeV
E=0.362 GeV
E=0.512 GeV
E=0.724 GeV
E=1.024 GeV
E=1.448 GeV
E=2.048 GeV
E=2.896 GeV
E=4.096 GeV
E=5.792 GeV
E=8.192 GeV
E=16.384 GeV
E=21.768 GeV
E=65.536 GeV
E=131.072 GeV
Geant4
Model
Probability
Total energy / label

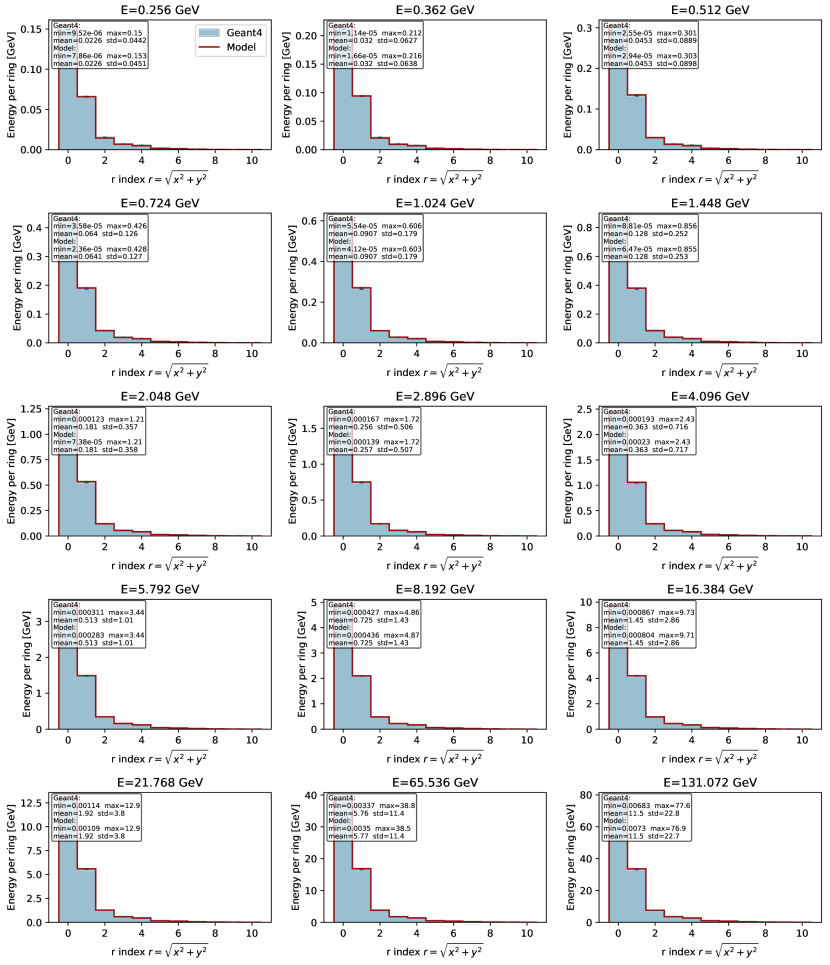

Number of hits distribution ($E > 10^{-7}$ GeV)
E=0.256 GeV
E=0.362 GeV
E=0.512 GeV
E=0.724 GeV
E=1.024 GeV
E=1.448 GeV
E=2.048 GeV
E=2.896 GeV
E=4.096 GeV
E=5.792 GeV
E=8.192 GeV
E=16.384 GeV
E=21.768 GeV
E=65.536 GeV
E=131.072 GeV
Geant4
Model
Probability
Number of hits ($E > 10^{-7}$ GeV)

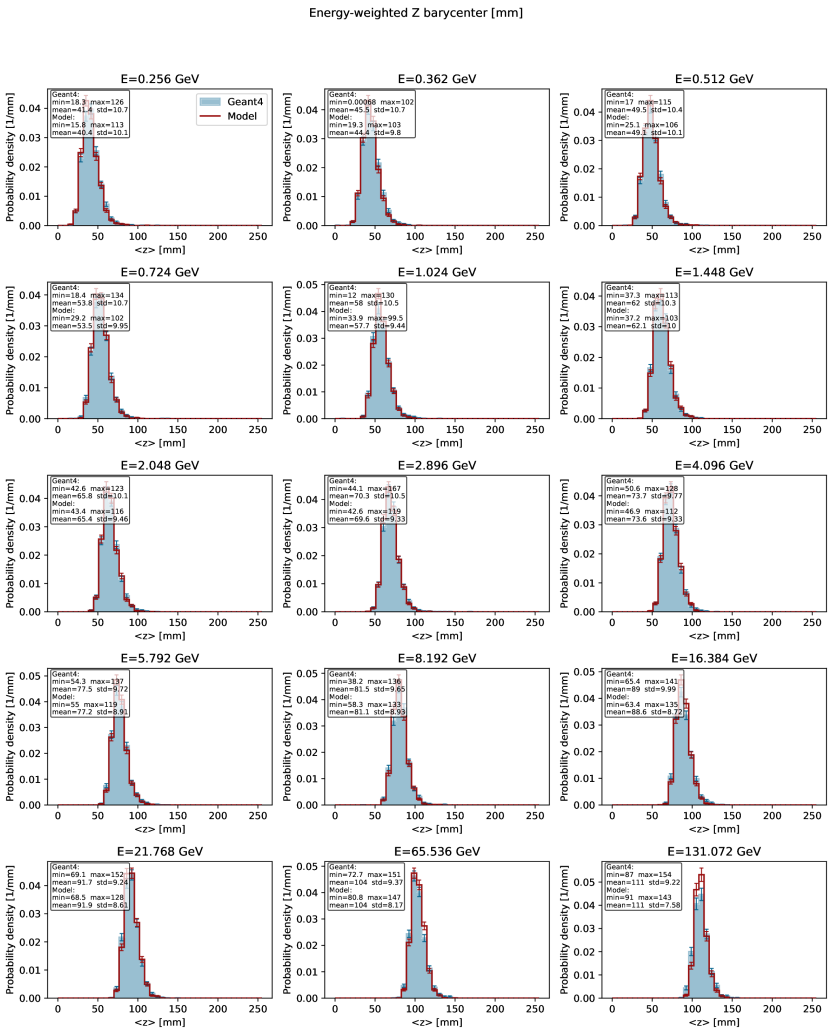

Number of hits distribution ($E > 10^{-7}$ GeV)
E=0.256 GeV
E=0.362 GeV
E=0.512 GeV
E=0.724 GeV
E=1.024 GeV
E=1.448 GeV
E=2.048 GeV
E=2.896 GeV
E=4.096 GeV
E=5.792 GeV
E=8.192 GeV
E=16.384 GeV
E=21.768 GeV
E=65.536 GeV
E=131.072 GeV
Geant4
Model
Probability
Number of hits ($E > 10^{-7}$ GeV)

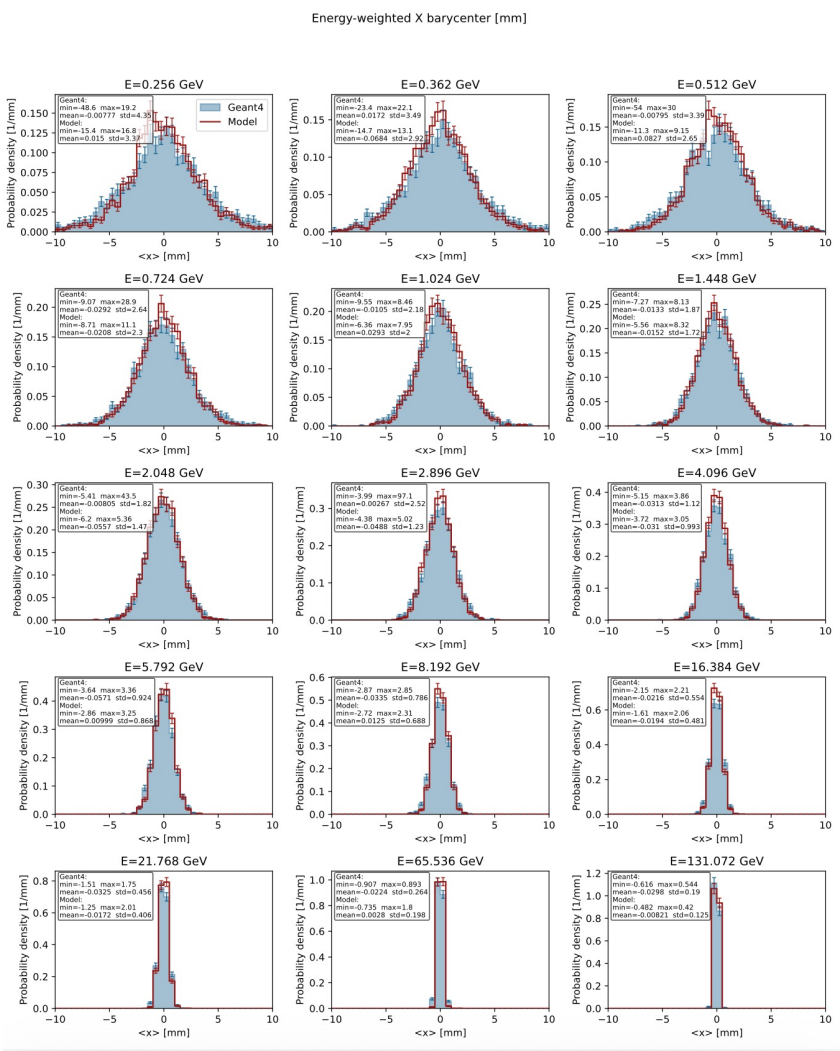
Mean energy per Z layer [GeV]



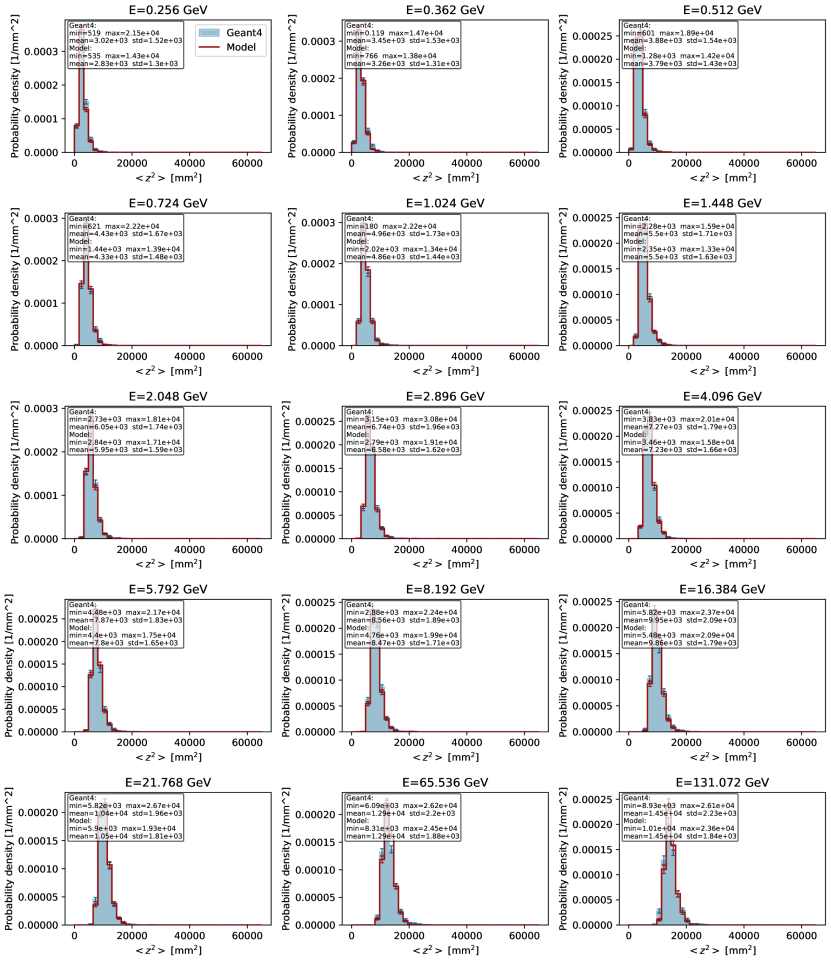
Mean energy vs radial index [GeV]



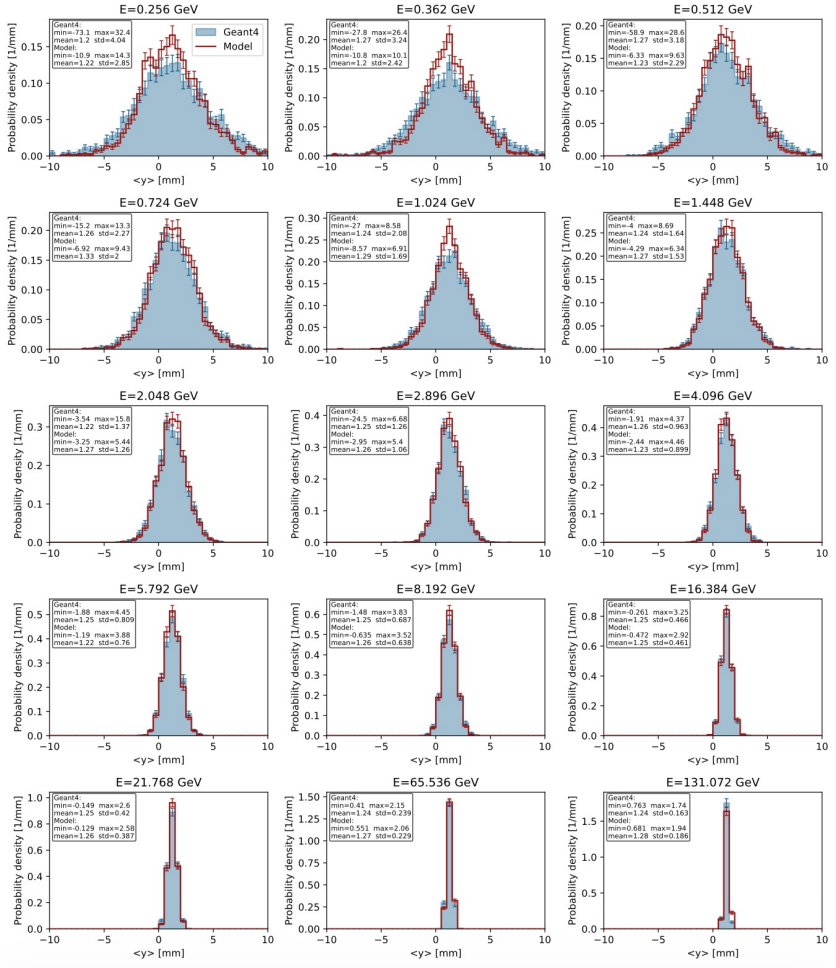

Energy-weighted Z barycenter [mm]
E=0.256 GeV
E=0.362 GeV
E=0.512 GeV
E=0.724 GeV
E=1.024 GeV
E=1.448 GeV
E=2.048 GeV
E=2.896 GeV
E=4.096 GeV
E=5.792 GeV
E=8.192 GeV
E=16.384 GeV
E=21.768 GeV
E=65.536 GeV
E=131.072 GeV
Geant4
Model
Probability density [1/mm]
<z> [mm]


Energy-weighted X barycenter [mm]
E=0.256 GeV
E=0.362 GeV
E=0.512 GeV
E=0.724 GeV
E=1.024 GeV
E=1.448 GeV
E=2.048 GeV
E=2.896 GeV
E=4.096 GeV
E=5.792 GeV
E=8.192 GeV
E=16.384 GeV
E=21.768 GeV
E=65.536 GeV
E=131.072 GeV
Geant4
Model
Probability density [1/mm]
<x> [mm]

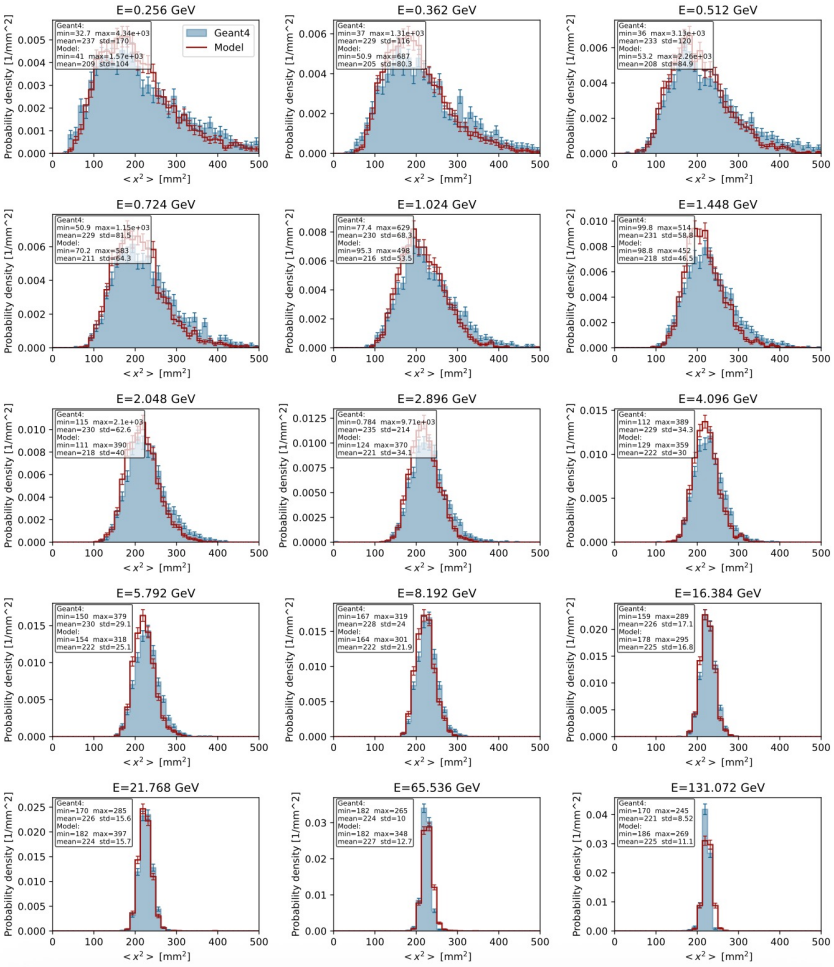
Energy-weighted second moment along Z [mm^2]



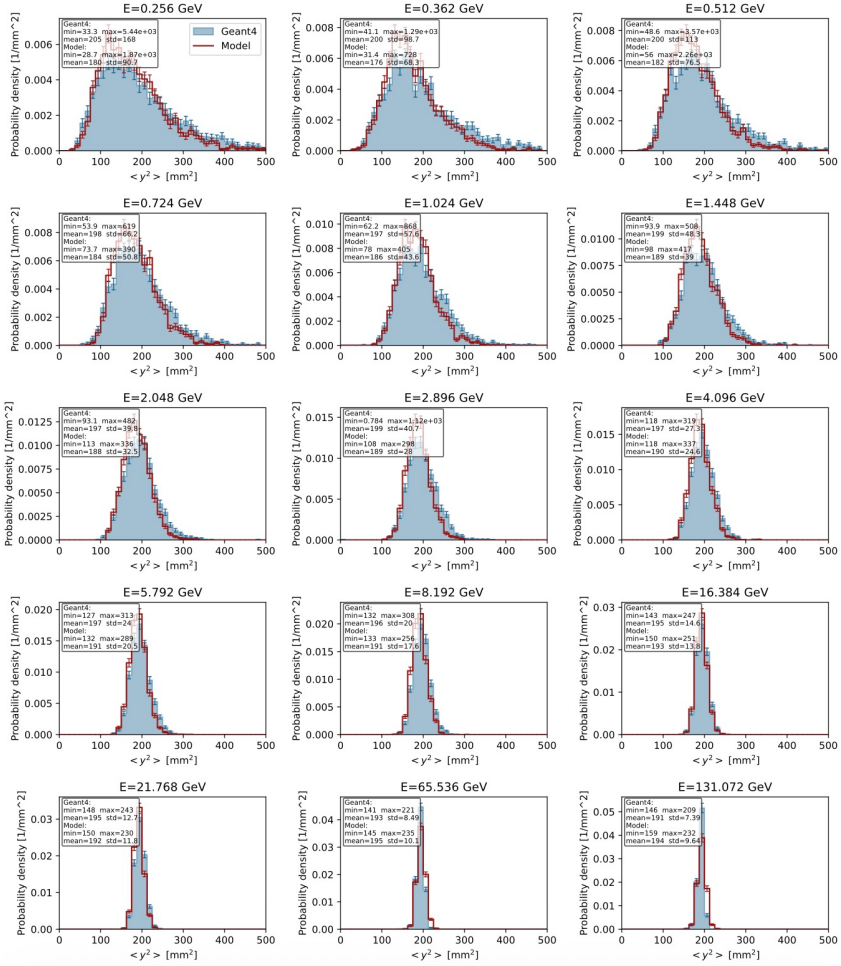
Energy-weighted Y barycenter [mm]



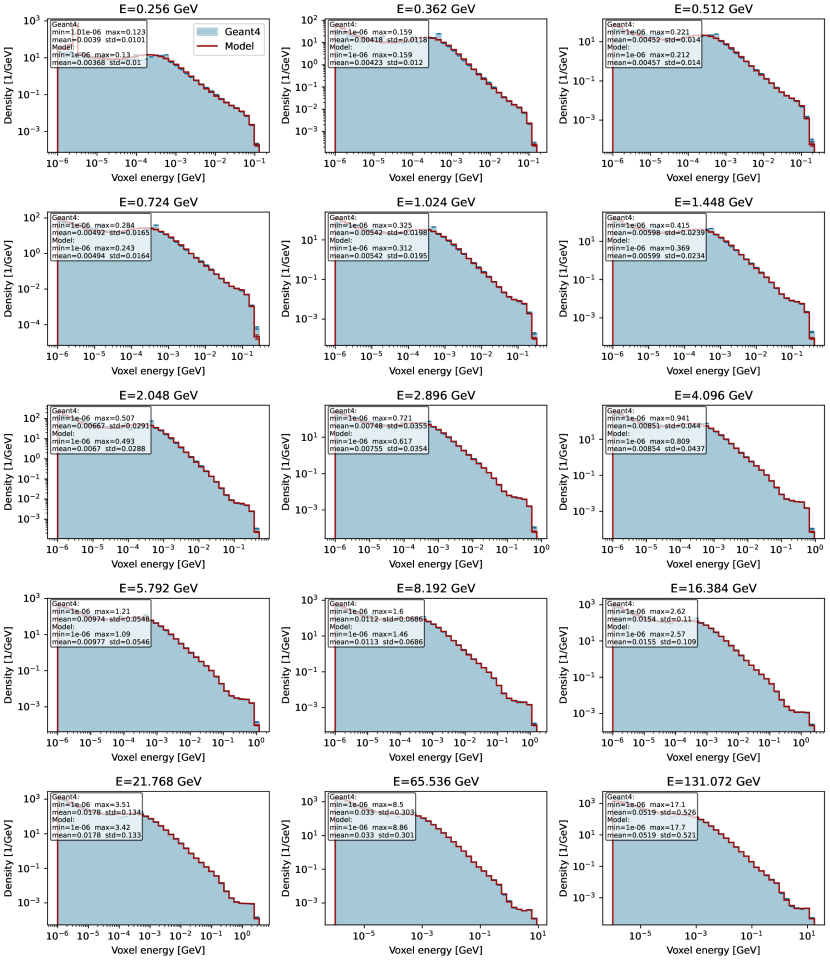
Energy-weighted second moment along X [mm^2]



Energy-weighted second moment along Y [mm^2]



Voxel energy density | combined range | log y | log x



Voxel energy density (ALRTransform) | combined range | log y