

李政道研究所
Tsung-Dao Lee Institute



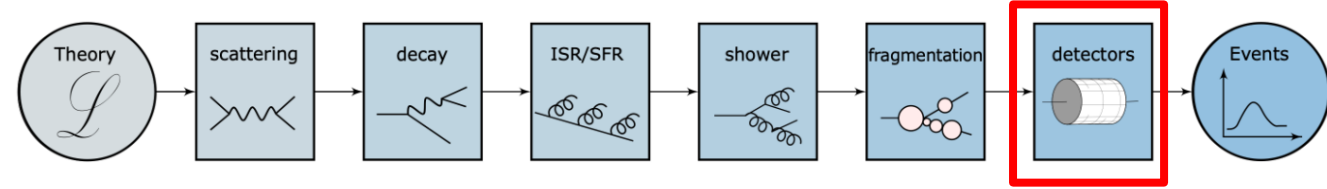
W

Yale

Latent Generative Model for Calorimeter Fast Simulation

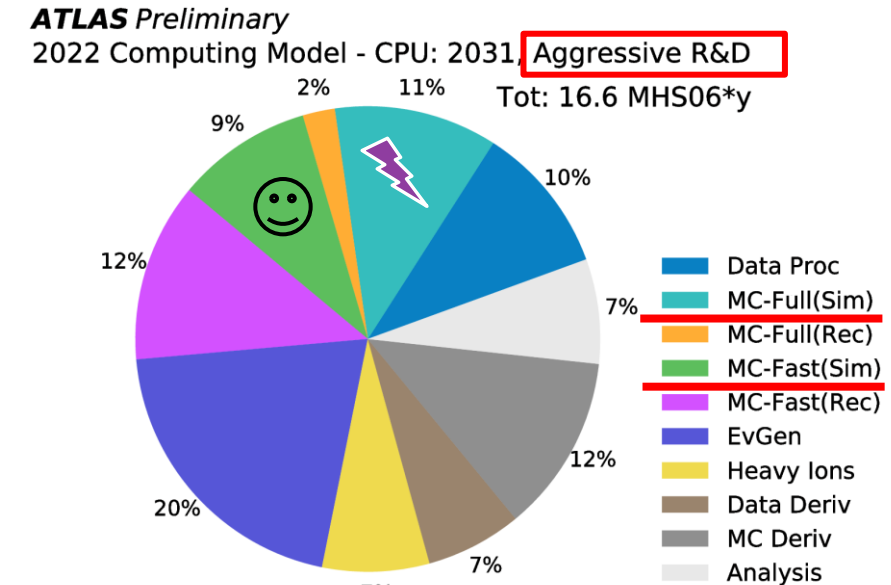
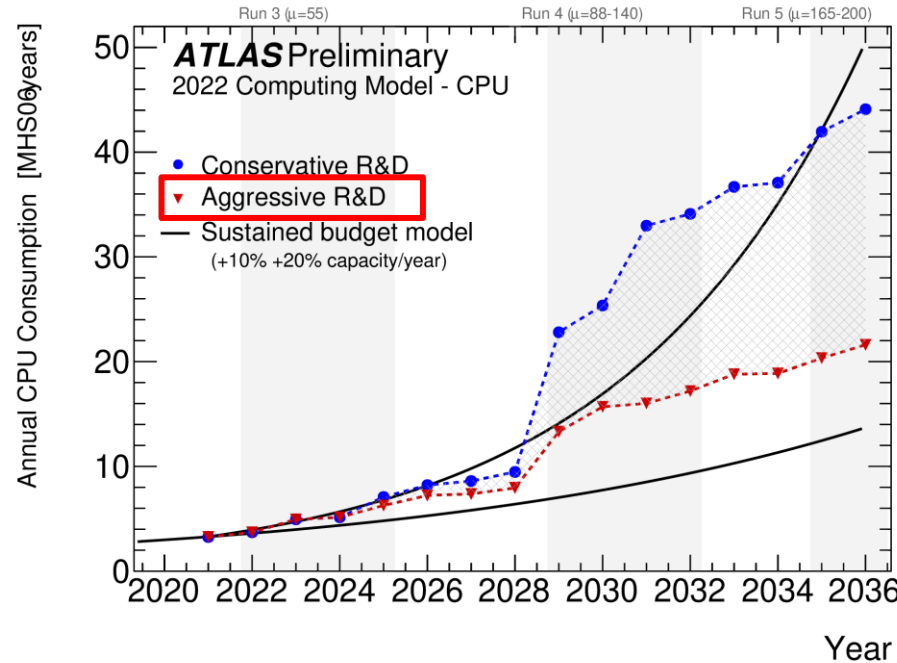
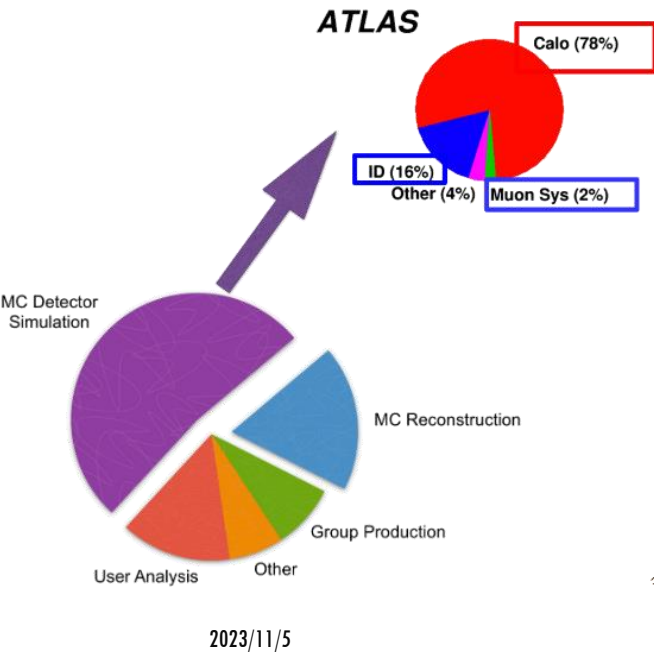
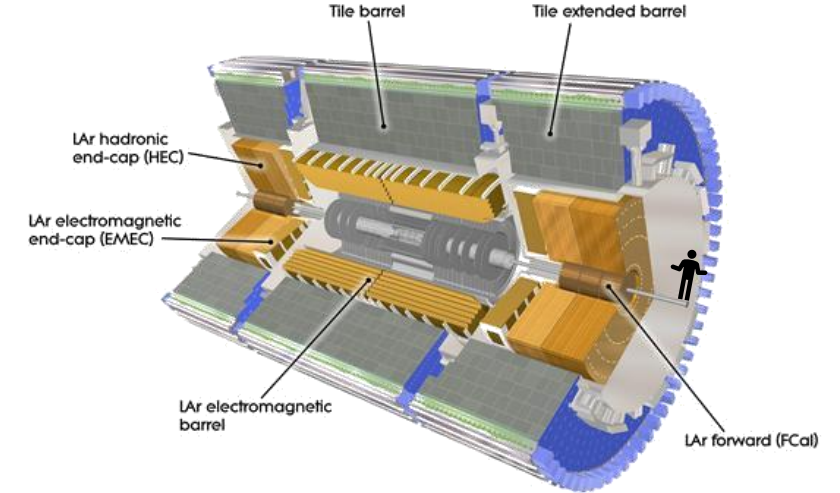
Qibin LIU, Chase Shimmin,
Xiulong LIU, Eli Shlizerman,
Shih-Chieh HSU, Shu LI

Introduction

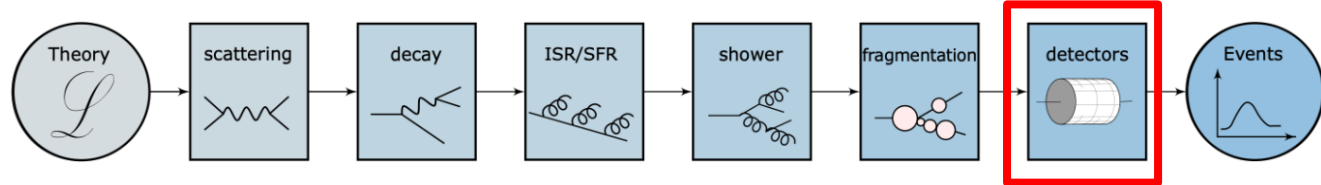


Machine Learning and LHC Event Generation, A. Butter et al. [2203.07460]

- Important step in the HEP workflow: Detector Simulation
- Calorimeter: “largest” part both in scale and computing cost
- Fast Simulation: most wanted and mandatory in the future
- Ultra-fast and scalable solution: latent generative model
- Implementations on CaloChallenge2022 datasets

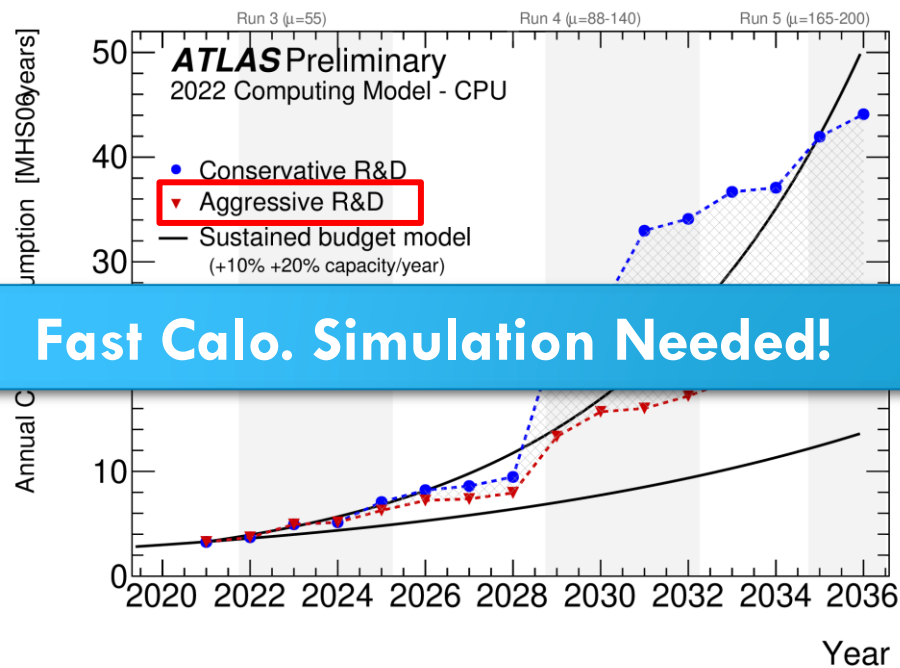
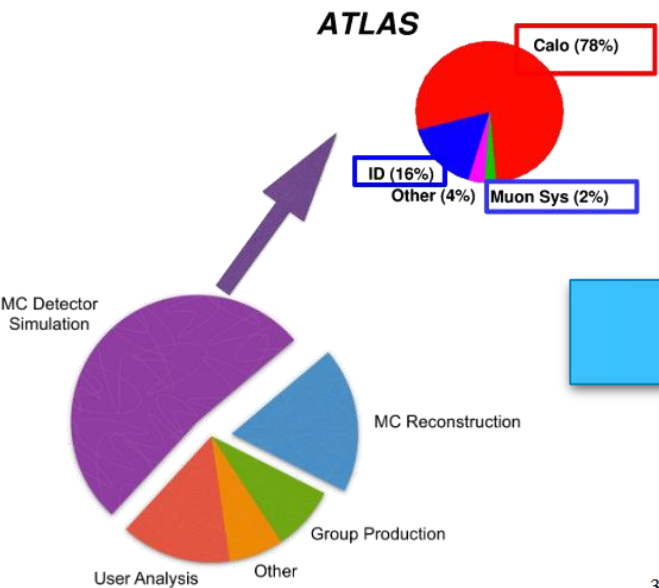
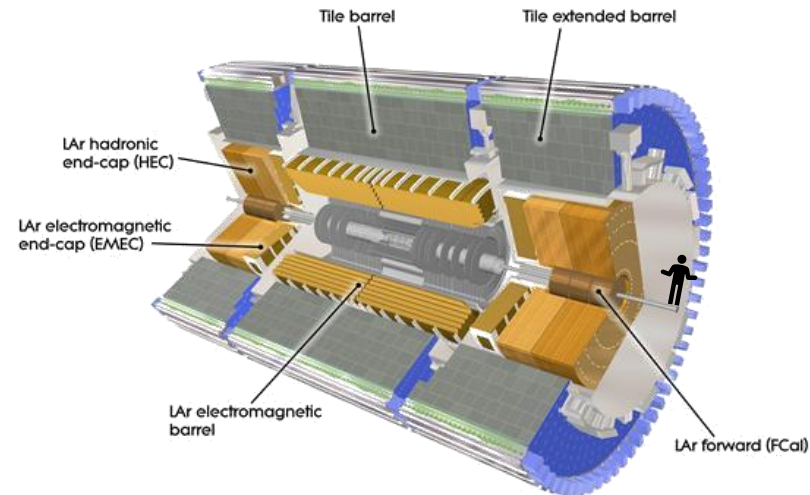


Introduction



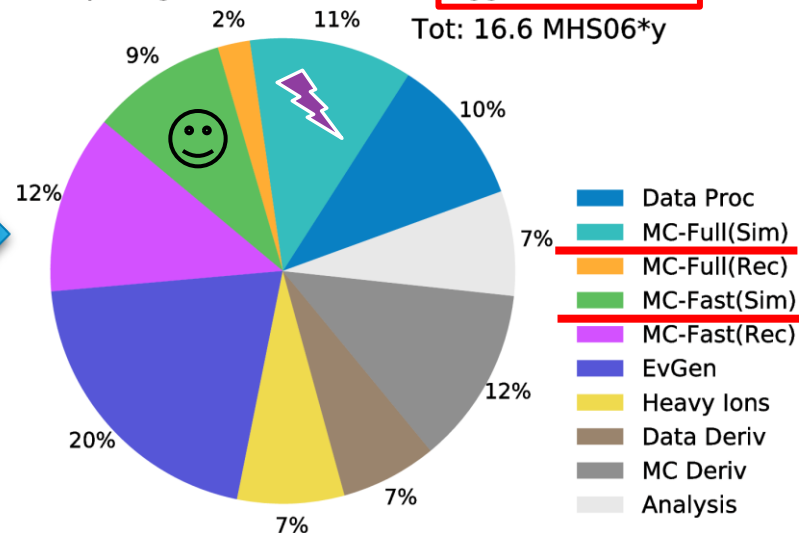
Machine Learning and LHC Event Generation, A. Butter et al. [2203.07460]

- Important step in the HEP workflow: Detector Simulation
- Calorimeter: “largest” part both in scale and computing cost
- Fast Simulation: most wanted and mandatory in the future
- Ultra-fast and scalable solution: latent generative model
- Implementations on CaloChallenge2022 datasets



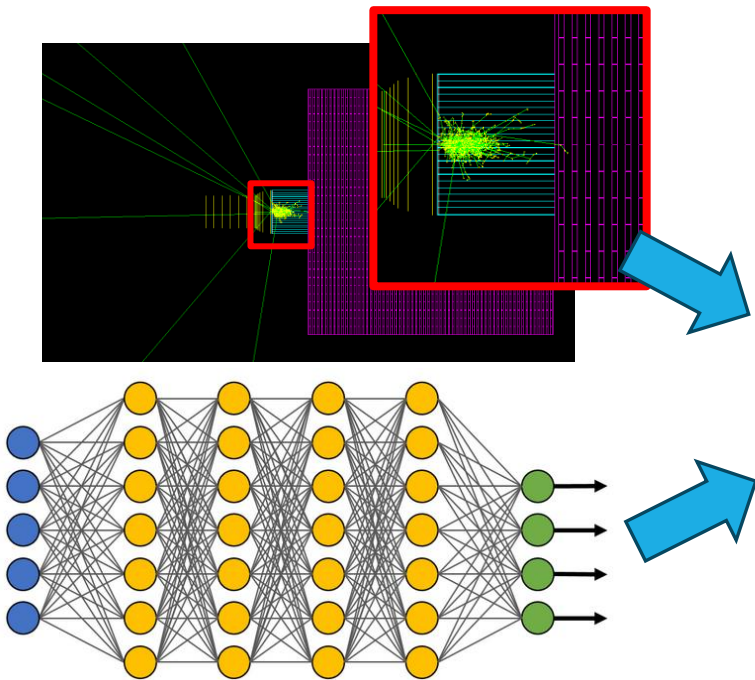
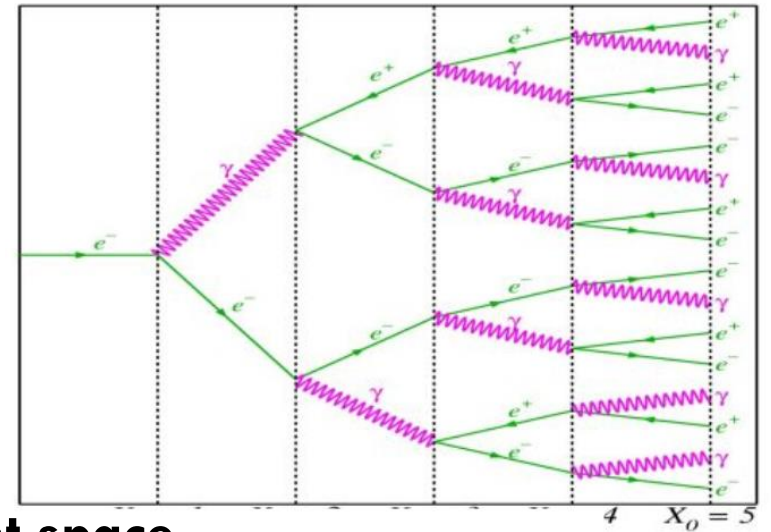
Fast Calo. Simulation Needed!

ATLAS Preliminary 2022 Computing Model - CPU: 2031 Aggressive R&D
Tot: 16.6 MHS06*y

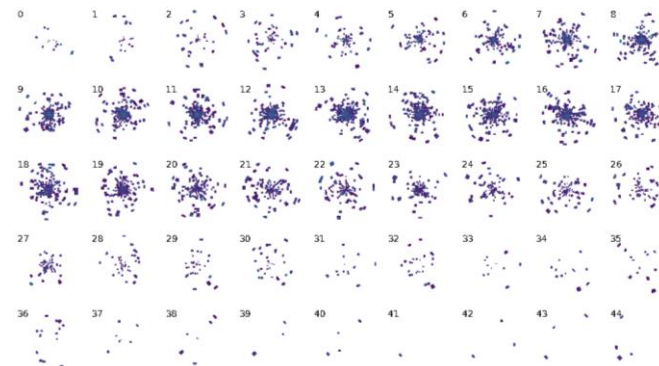


Simulation of Calorimeter

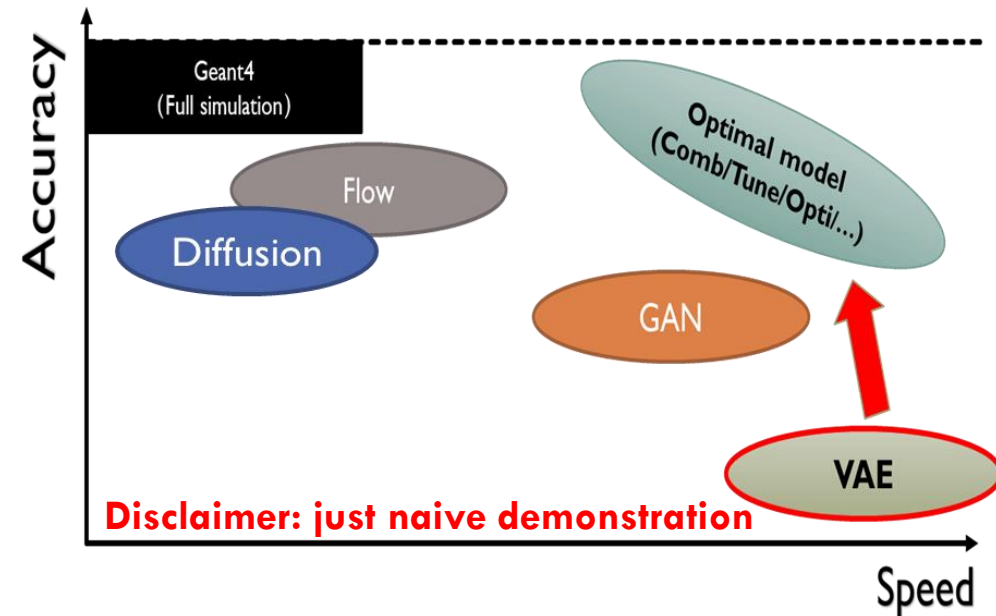
- Simulate hits \vec{E} corresponding to each calo. readout channel
- Full simulation(GEANT4): Tracing of every secondary particle
- Fast simulation: generate response in one/few pass(es)
- **VAE- based generative model: fast and well-controlled latent space**



$$P(\vec{E}_i | E_{inc}, type, \dots)$$

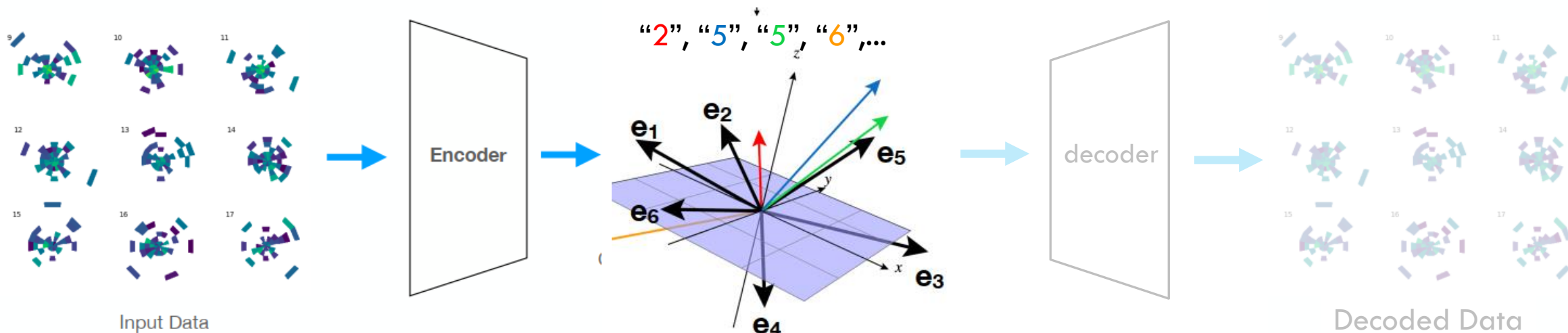


1 Shower on 45 layers calo.



Two-Step Generative Model: *Encoding*

- Compress and encode the calo. data into latent space: **“Auto-Encoder”**
- Quantized the latent space into code with **Vector Quantization** [\[1711.00937\]](#)
- $\mathbb{R}^D \rightarrow \mathbb{Z}^L$: Large compression ratio ($D \gg L$) and more descriptive
- Well defined objective and good scaling in general

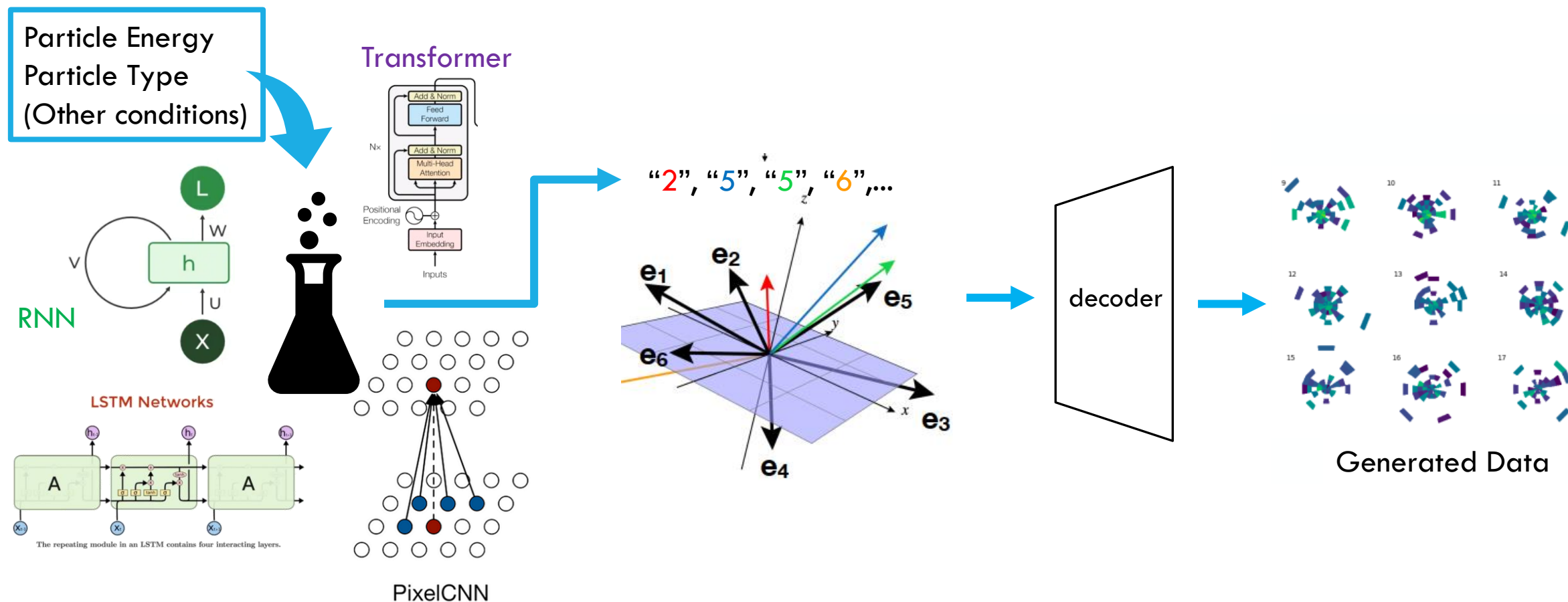


$$\left| \text{sg}[z_{enc}(x)] - e \right|^2 + \beta \left| z_{enc}(x) - \text{sg}[e] \right|^2$$

Vector Quantization	Commitment Loss
Dictionary learning // Update codebook	Tries to keep encoder predictions close to codebook values.

Two-Step Generative Model: *Sampling*

- Latent (codes) sampled with token model: RNN/PixelCNN/Transformer...

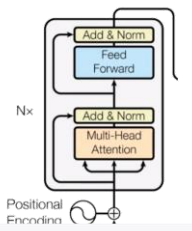


Two-Step Generative Model: *Sampling*

- Latent (codes) sampled with token model: RNN/PixelCNN/Transformer...
- **Bridge to modern rapidly developed AI model: GPT (imp. minGPT)**

Particle Energy
Particle Type
(Other conditions)

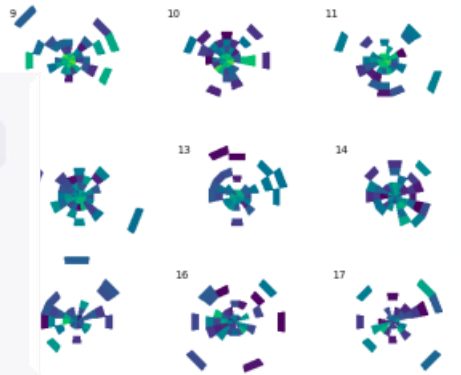
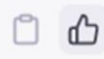
Transformer



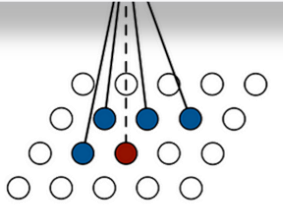
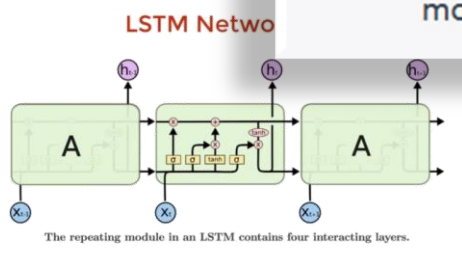
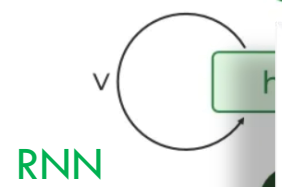
"2", "5", "5", "6", ...



The choice of the best model for generating discrete tokens depends on the specific task and requirements. However, one commonly used model for generating discrete tokens is the GPT (Generative Pre-trained Transformer) model, which is a transformer-based language model.



Generated Data

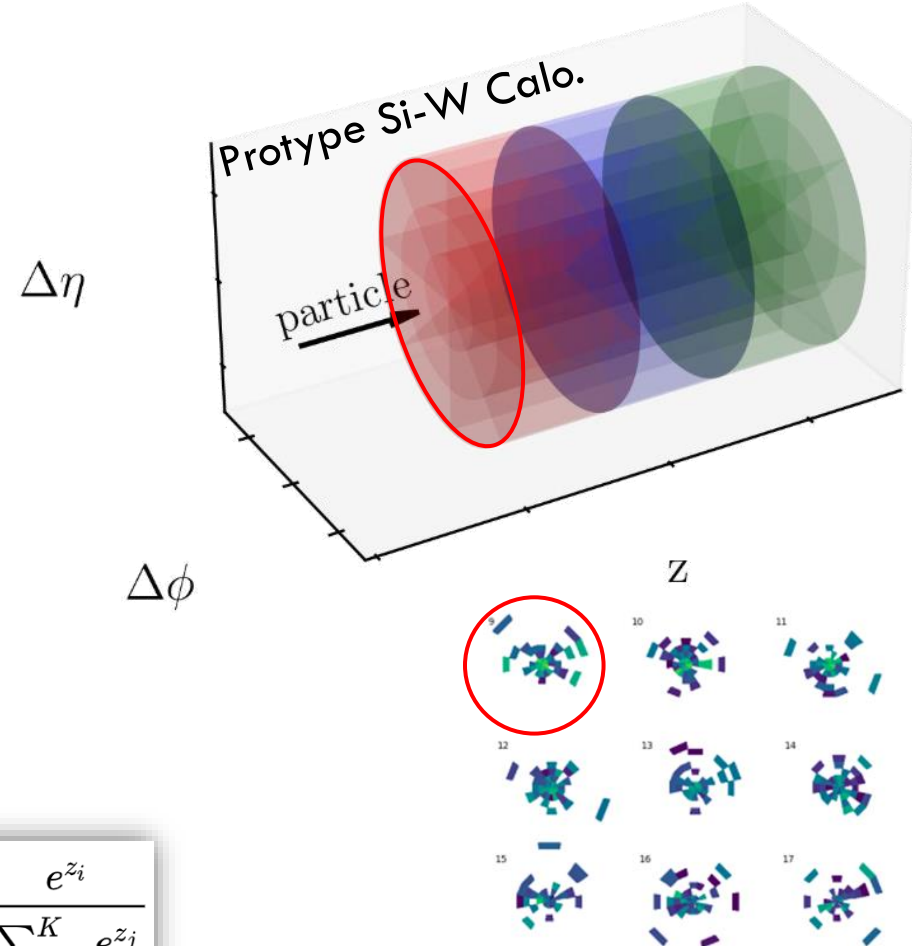


PixelCNN

e4

Dataset and Preprocessing

- Common calo. dataset: [CaloChallenge2022](#)
- Cylinder with 384~40500 channels
- Particle incident energy E_c : GeV ~ TeV (γ, π, e)
- Large dynamic range: KeV ~ GeV for each channel E_i
- High sparsity: most channels empty and compressible
- Preprocessing: Normalization & Log \Leftrightarrow SoftMax & Exp



$$\sigma(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}}$$

$E_c = 20$ [GeV]

Normalized by $\sum E_i$

Logarithm + Scaling

Soft-max Output

Exponential

Recover by $\sum E_i$

1	2	1
1	8	2
1	1	1

0.06	0.11	0.06
0.06	0.44	0.11
0.06	0.06	0.06

0.56	0.63	0.56
0.56	0.78	0.63
0.56	0.56	0.56

-1.26	-1.26	-1.26
-1.26	-0.41	-0.78
-1.26	-0.95	-1.26

0.06	0.06	0.06
0.06	0.39	0.17
0.06	0.11	0.06

1	1	1
1	7	3
1	2	1

$\sum E_i = 18$ [GeV]

$R \equiv \sum E_i / E_c = 0.9$

Encoder Pass

Decoder Pass

$\sum E_i = R * E_c = 18$ [GeV]

Sampled by Latent Model

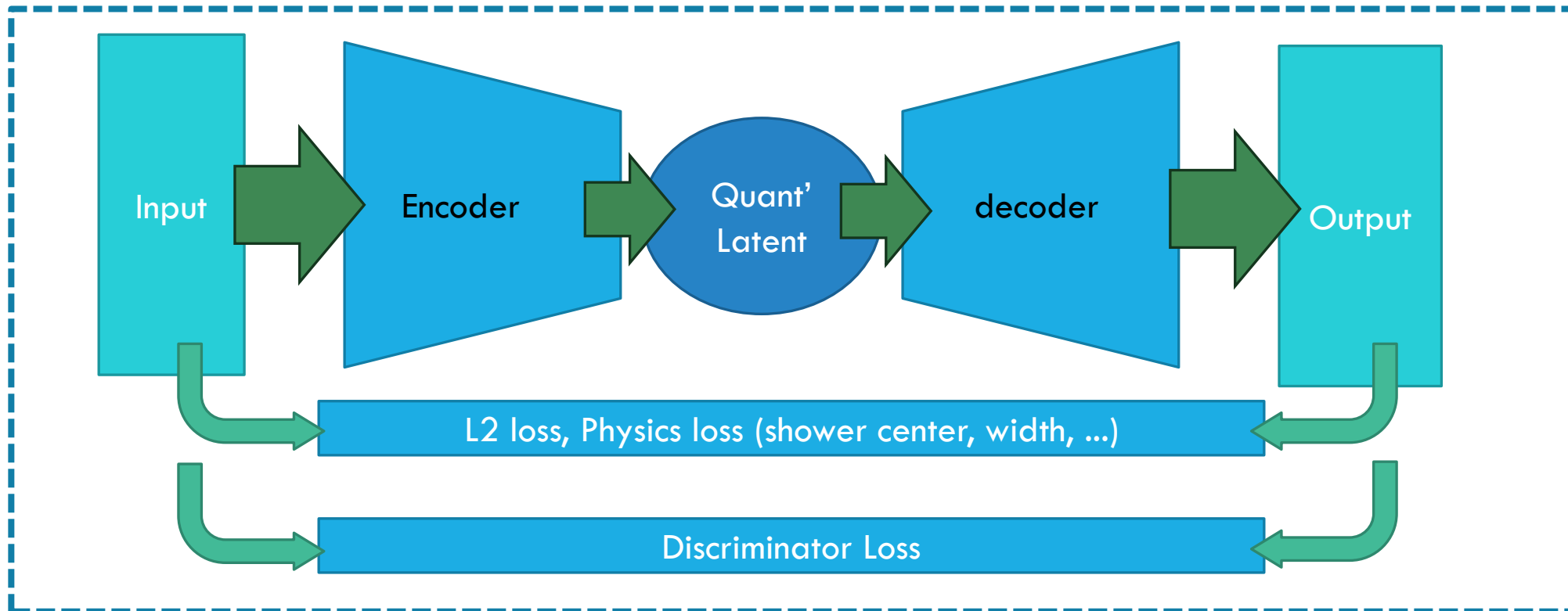
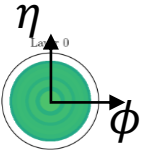
Step1 Implementation: VQGAN

$$Q^* = \arg \min_{E, G, Z} \max_D \mathbb{E}_{x \sim p(x)} \left[\mathcal{L}_{VQ}(E, G, Z) + \lambda \mathcal{L}_{GAN}(\{E, G, Z\}, D) \right]$$

- VQVAE combined with adversary trained discriminator (VQGAN)
- Pixel-wise loss: L2 (MSE) loss comparing input and decoded
- Physics-aware loss: shower center and width difference,...
- Shower information compressed into latent space as codes

$$\|E' - E\|^2$$

$$|\eta \cdot (E' - E)|$$



Performance on Small Dataset(π)

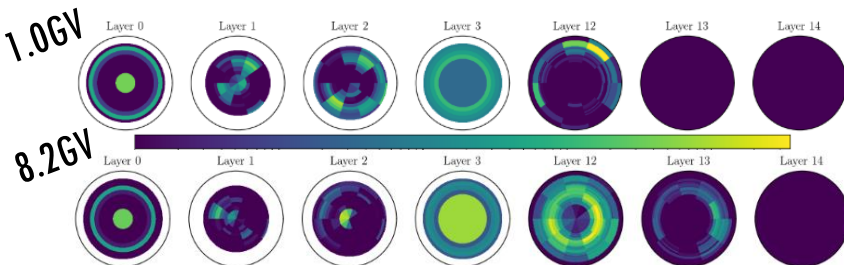
- Small and irregular geometry: fully connected layers utilized
- Average shower: matches ground truth for all calo. layers
- Energy response: good agreement in wide energy range
- Dist. of physics variables: $\langle S^2 \rangle$ reaching 0.01 level

Separation Power

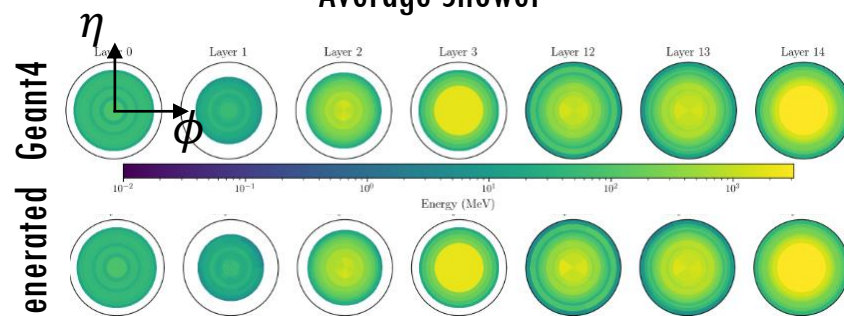
$$\langle S^2 \rangle = \frac{1}{2} \sum_{i=1}^{n_{\text{bins}}} \frac{(h_{1,i} - h_{2,i})^2}{h_{1,i} + h_{2,i}}$$

[2009.03796]

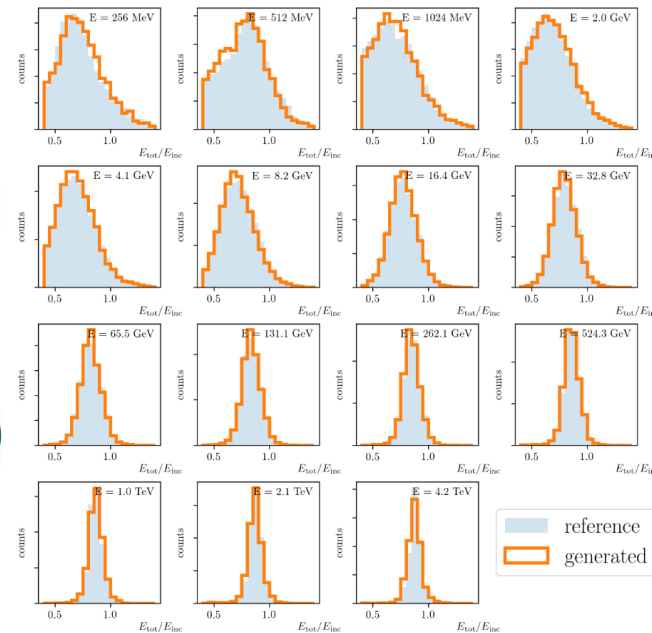
Arbitrary Generated Shower



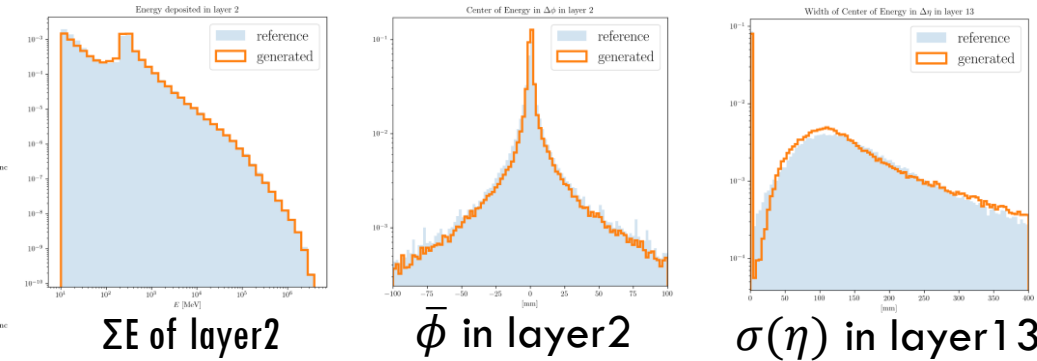
Average Shower



Total Energy Response ($\Sigma E/E_c$)


 $\langle S^2 \rangle = 0.001 \sim 0.01$

Selections of well-modelled distributions

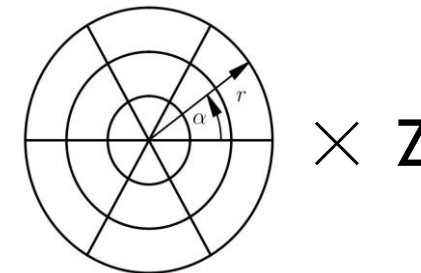


Metrics Per Layer	Mean	Best	Worst
ΣE	0.003	0.001	0.006
Shower Center	0.014	0.009	0.019
Shower Width	0.017	0.006	0.037

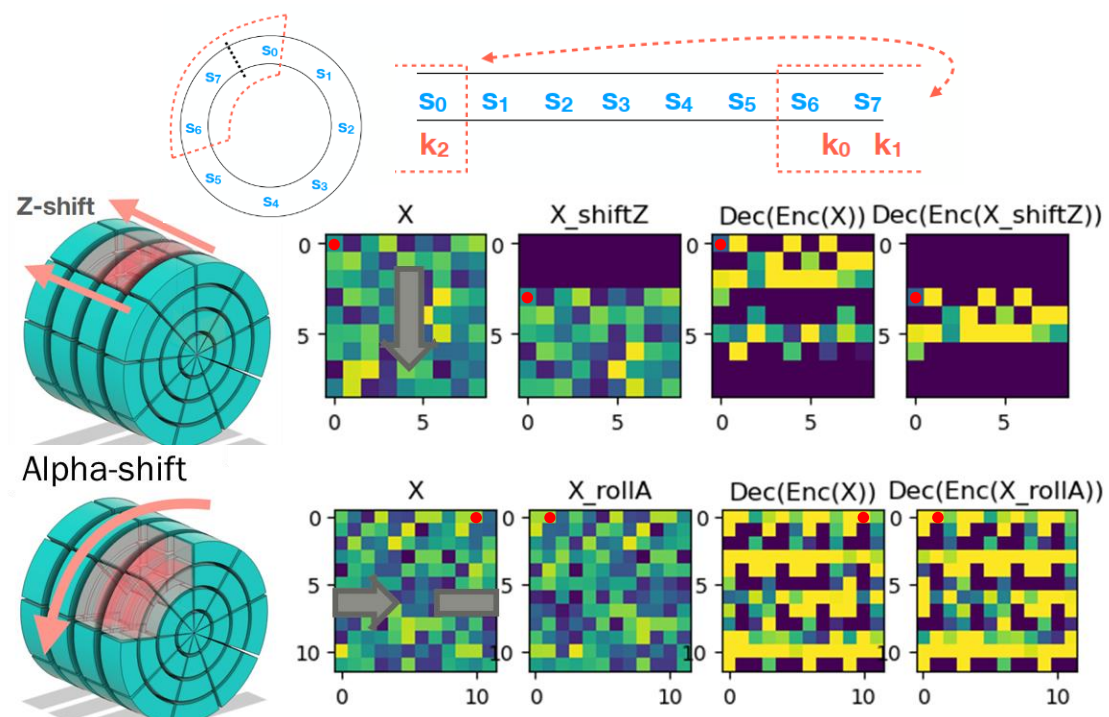
 $\langle S^2 \rangle$ of All Variables

Scale to Large Dataset

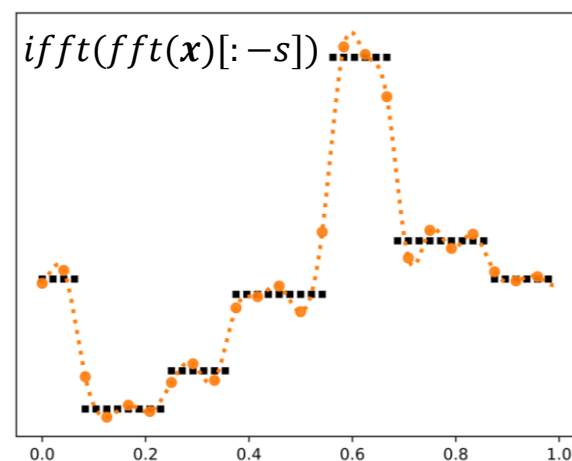
- Orthogonal segmentation: cylindrical convolution operator
- Equivariant down-/up- sampling: FFT resampling
- Residual and Attention: capture of long-range information
- Layer-wise normalization: ΣE layer encoded into latent codes
- Tricks of training: HPO, adaptive weight, ... and *patience*



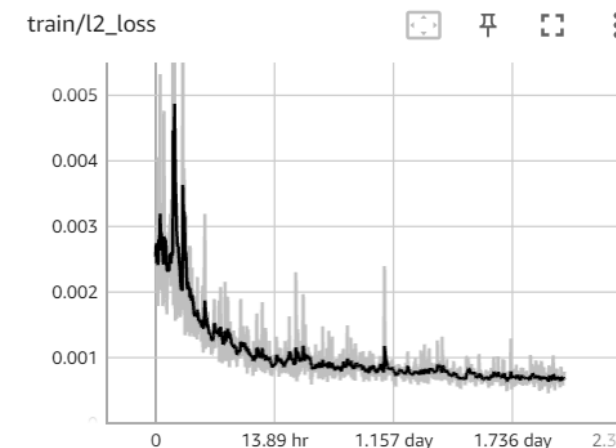
Datasets	Ch(Z)	Ch(α)	Ch(r)
“Easy”	5/7	(Irregular)	
“Medium”	45	16	9
“Hard”	45	50	18



FFT down-sampling

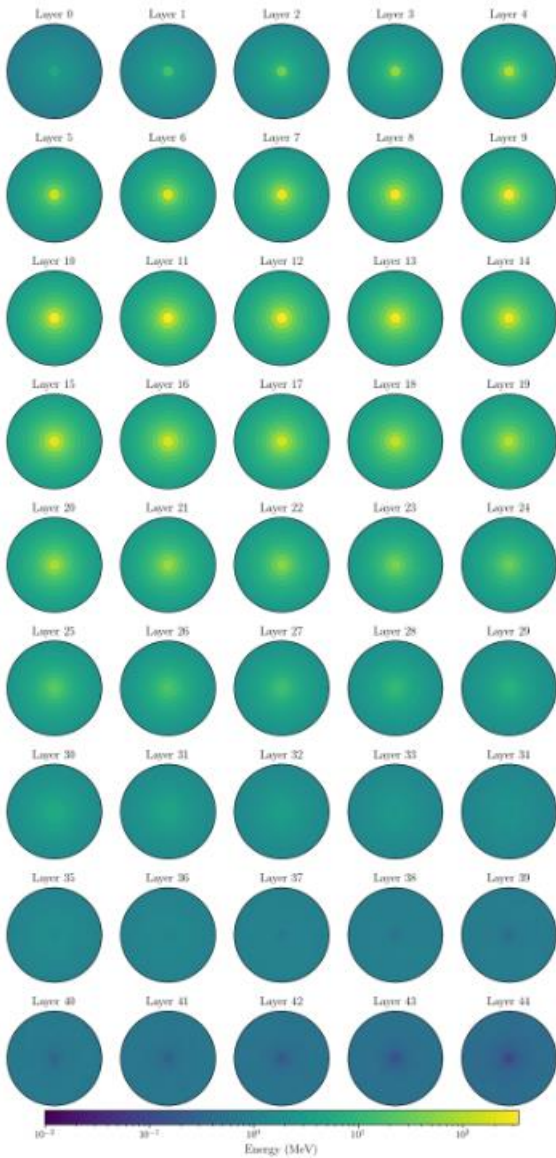


Typical Loss Curve

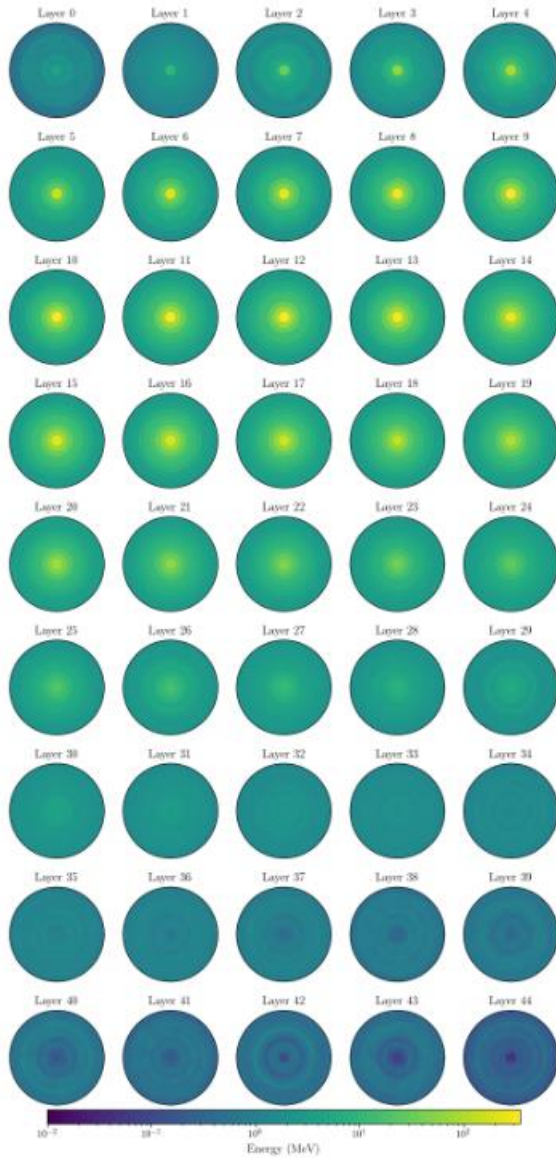


Scale to Large Dataset

Geant4

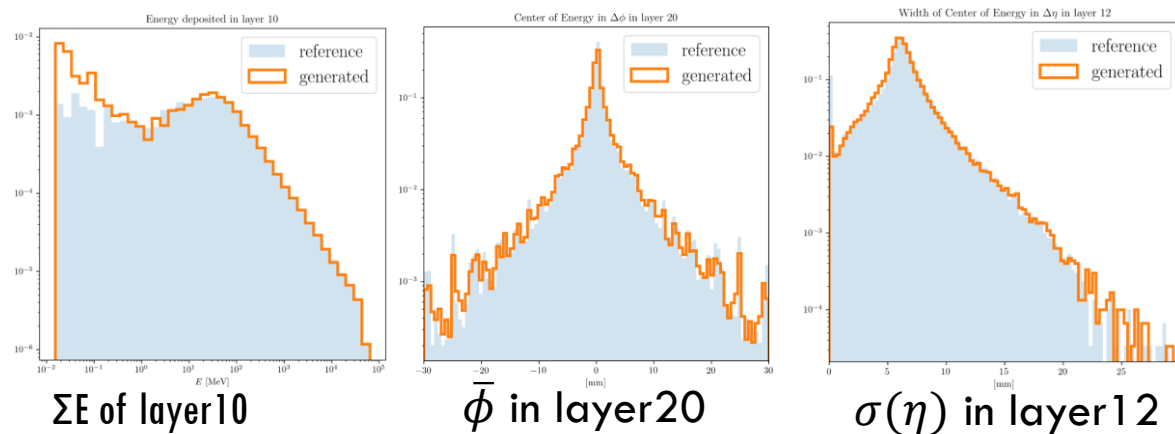


Generated



- In general good agreement with input
 - Not perfect at first and last several calo. layers
- Higher sparsity, larger dynamic range, lower stats

Selections of well-modelled distributions

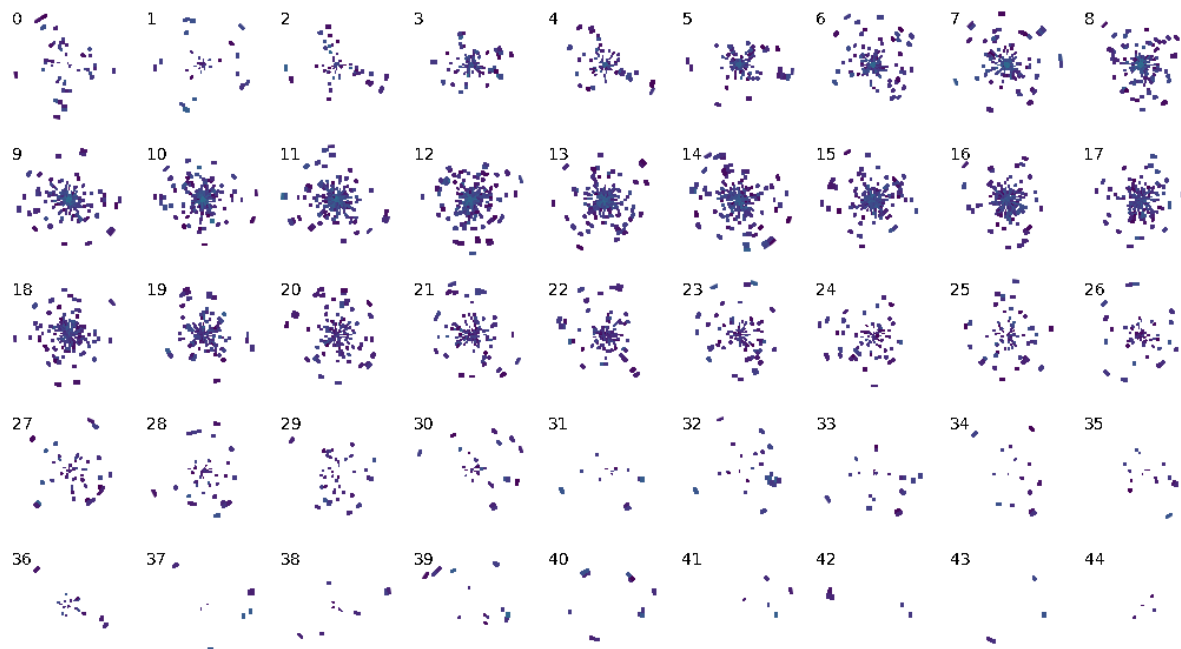


Metrics Per Layer	Mean	Best	Worst
ΣE	0.009	0.001	0.024
Shower Center	0.012	0.002	0.033
Shower Width	0.020	0.003	0.057

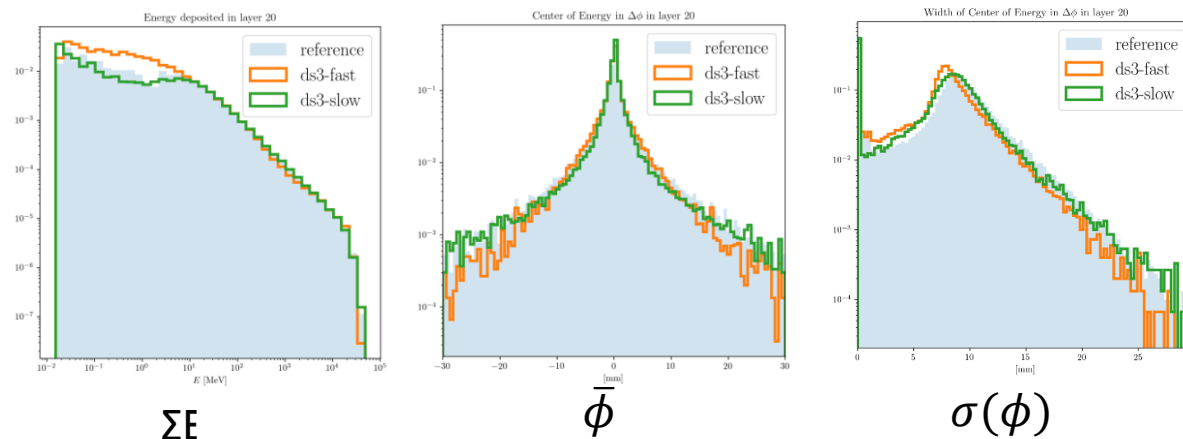
Scale to Large Dataset

- 2 models with different N#pars. and complexity
- $\langle S^2 \rangle$ reaching 0.01 for energy response and 0.02 for shower shape
- Generally good modelling of inner calo. layers

Arbitrary Generated Shower (162.28GeV)



Distribution at Layer 20



ds3-fast	Mean	Best	Worst
ΣE	0.021	0.002	0.130
Shower Center	0.024	0.003	0.076
Shower Width	0.044	0.004	0.133
ds3-slow	Mean	Best	Worst
ΣE	0.004	0.001	0.011
Shower Center	0.014	0.007	0.045
Shower Width	0.032	0.006	0.095

Performance Summary

- Sampling time tested on 1xV100 GPU with 512 showers/batch
- Step1 (en/decoder) forward time at same level regardless of geometry
- Step2 (transformer) dominated for the total sampling time
- $\langle S^2 \rangle$ measured all the energy and shape variables of different calo. layers:
- Best performed layers reach 0.001 and worst at 0.1 level

Model	Chan. (D)	S1 time/ms	S2 time/ms	Total time/ms	Latent Size (L)	Best $\langle S^2 \rangle$	Worst $\langle S^2 \rangle$
ds1-photon	368	0.02	0.23	0.25	42	0.001	0.023
ds1-pion	533	0.02	0.26	0.28	46	0.001	0.037
ds2	6480	0.17	0.46	0.63	70	0.001	0.057
ds3-fast	40500	0.35	0.79	1.14	184	0.002	0.133
ds3-slow	40500	1.7	34.4	36.1	274	0.001	0.095

Thinking of VAE model in Calorimeter Fast Simulation

Promising way to compress the high dimensional calo. data or scale other model:

- However high demanding of engineering effort

Well-controlled and general latent space:

- Useful for downstream tasks, e.g. reconstruction
- Interplay with [foundation model](#)

Information bottleneck or “limitation”:

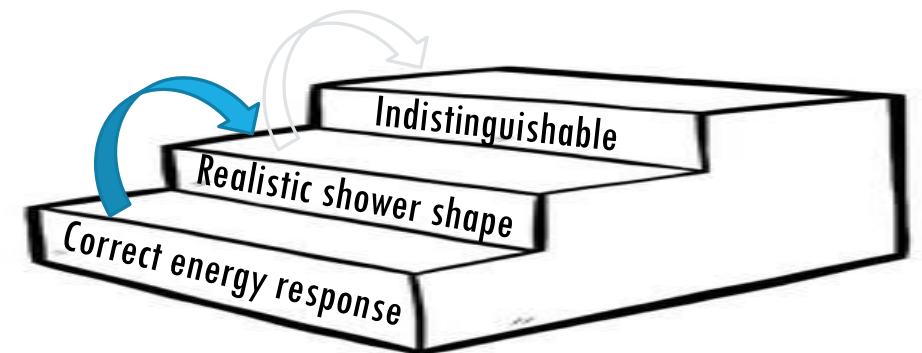
- More or less reduced the randomness and shorter “period”
- Higher compression, worse classifier score (easier to be caught)

Potential usage of a super-fast latent-labeled simulation model:

- On-the-fly testing of a real-time reconstruction system?
- “Deterministic” fast simulation (like a labeled image)?
- More ideas?

Various Step1 En/Decoder Models for Dataset3

Latent size L	AUC (50k cls-low)
140	0.9998
184	0.9962
274	0.9426
900	0.7876



Concluding

Calorimeter simulation is vital in HEP but computing-intensive

Machine learning methods show great potential for fast-calorimetry

Two steps model proposed based on VQVAE architecture

- **Vector Quantization enabling well controlled compression and flexible latent space**
- **GPT model adapted to do the conditional sample in the latent space**

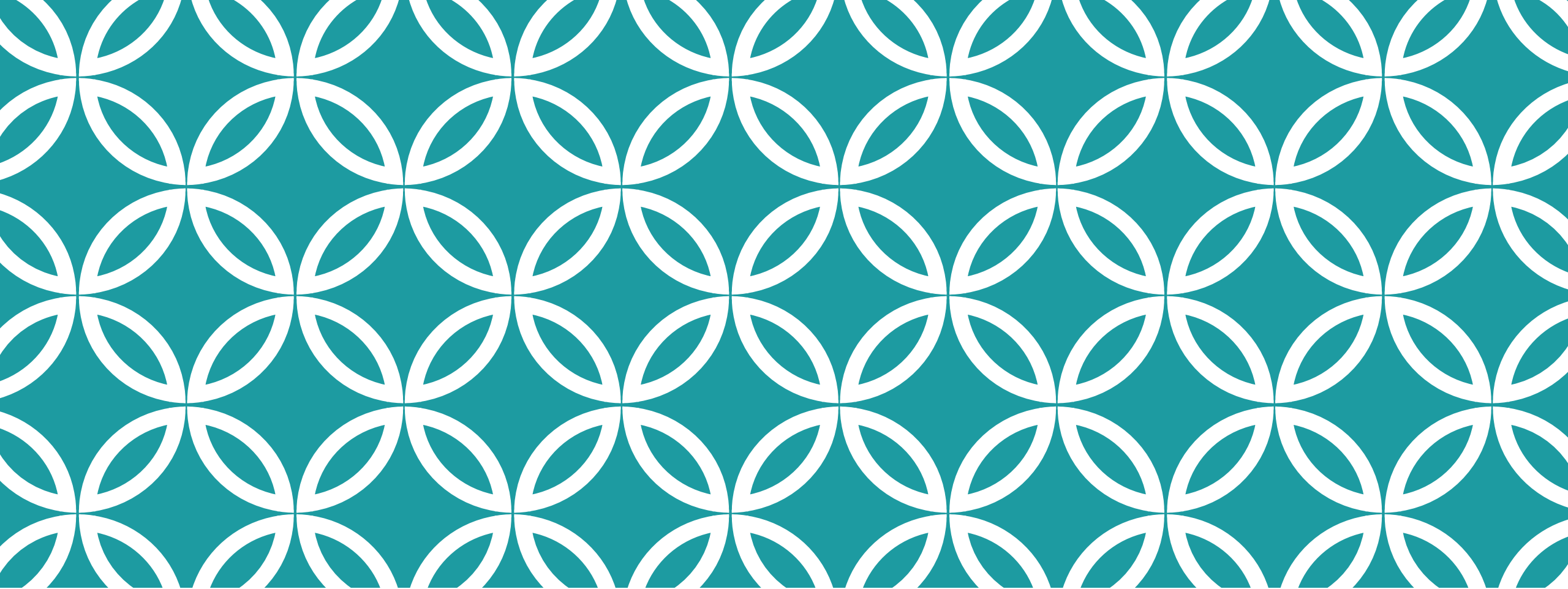
Methods designed for calorimeter data:

- Soft-max normalization, FFT resampling, cylindrical convolution

Performance on CaloChallenge datasets presented

- Promising performance on averaging shower and distribution of key variables
- Ultra-fast generation and scaling dominated by latent model
- Quality of generated detail features not perfect: more study ongoing

Potential application of latent based ultra-fast calorimetry simulation



Thanks for your Attention! |

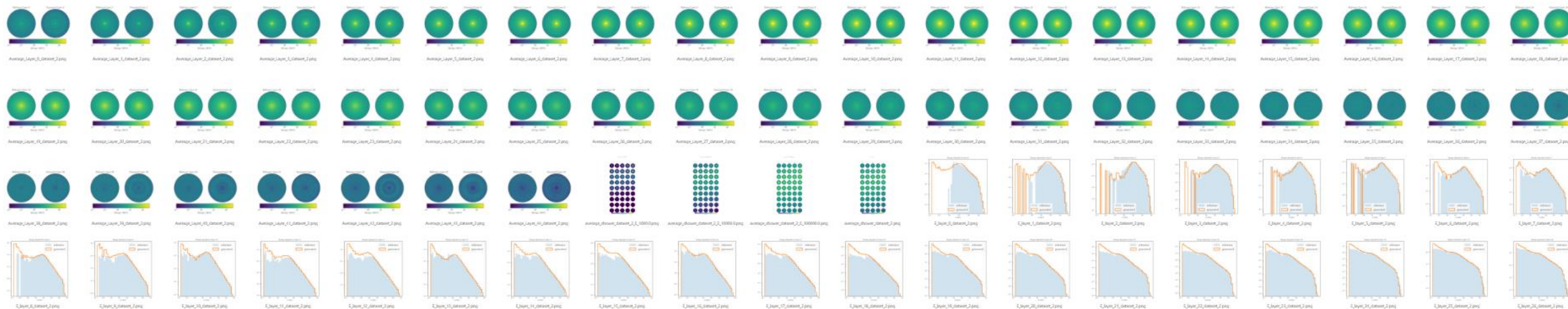
Full Evaluations for Photon Dataset

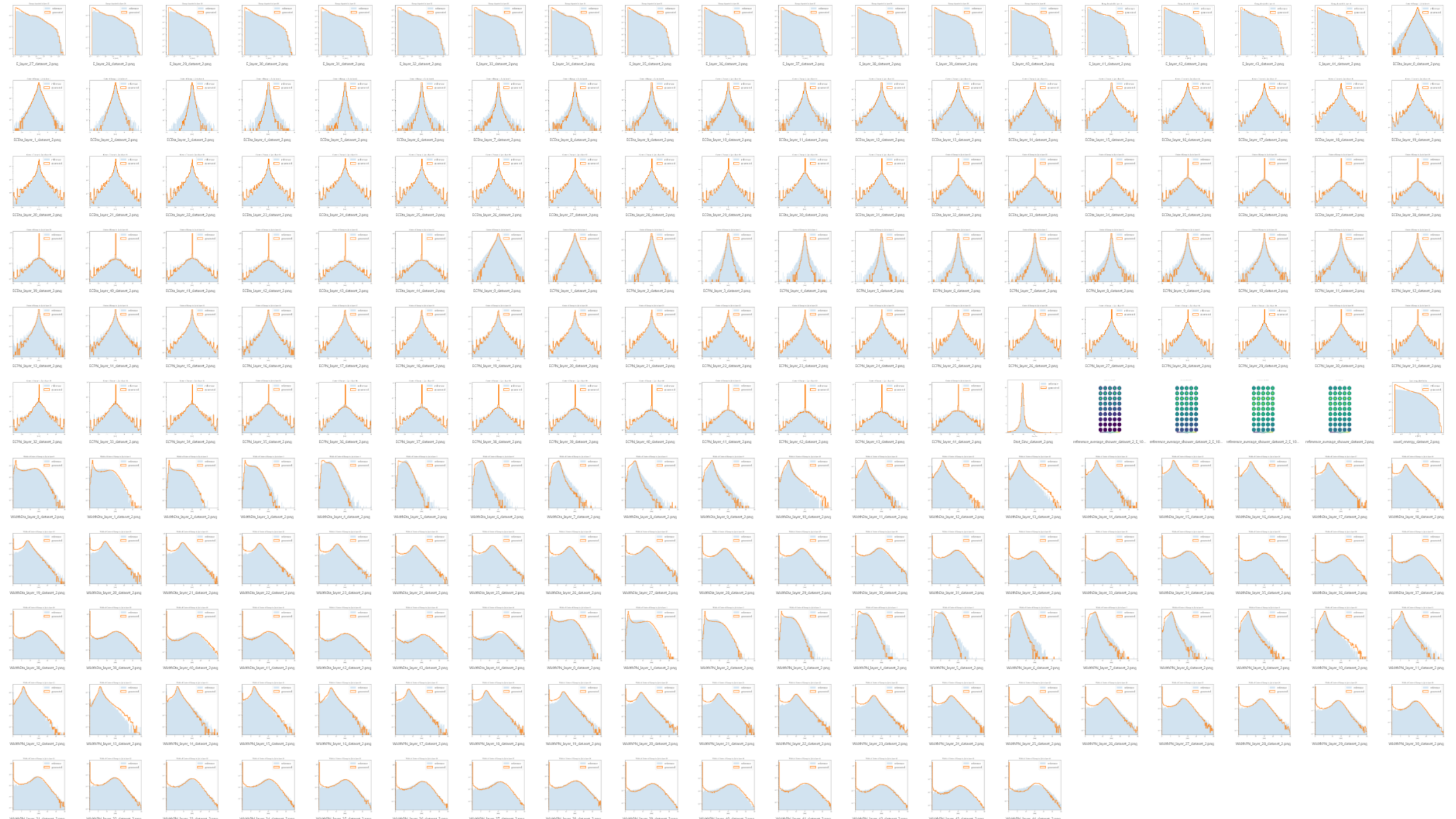


Full Evaluations for Pion Dataset



Full Evaluations for Dataset2





Full Evaluations for Dataset3 (slow model)

