Generative transformers for learning point-cloud simulations

Joschka Birk, Frank Gaede, Anna Hallin, Gregor Kasieczka, Martina Mozzanica, **Henning Rose**
henning.rose@studium.uni-hamburg.de

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

CLUSTER OF EXCELLENCE
QUANTUM UNIVERSE

Glühwein Workshop
December 17, 2024

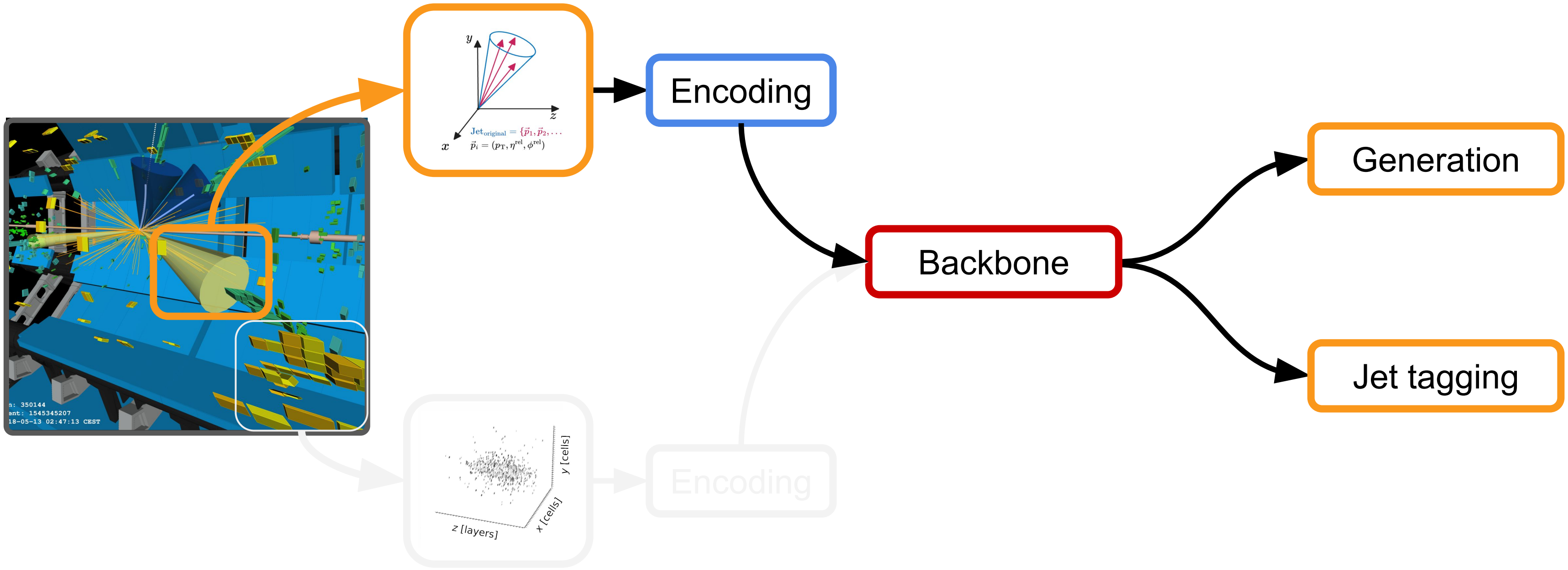
Motivation

OmniJet-α: The first cross-task foundation model for particle physics

Joschka Birk,[1,*] Anna Hallin,[1,†] and Gregor Kasieczka[1]
[1] *Institute for Experimental Physics, Universität Hamburg*
Luruper Chaussee 149, 22761 Hamburg, Germany

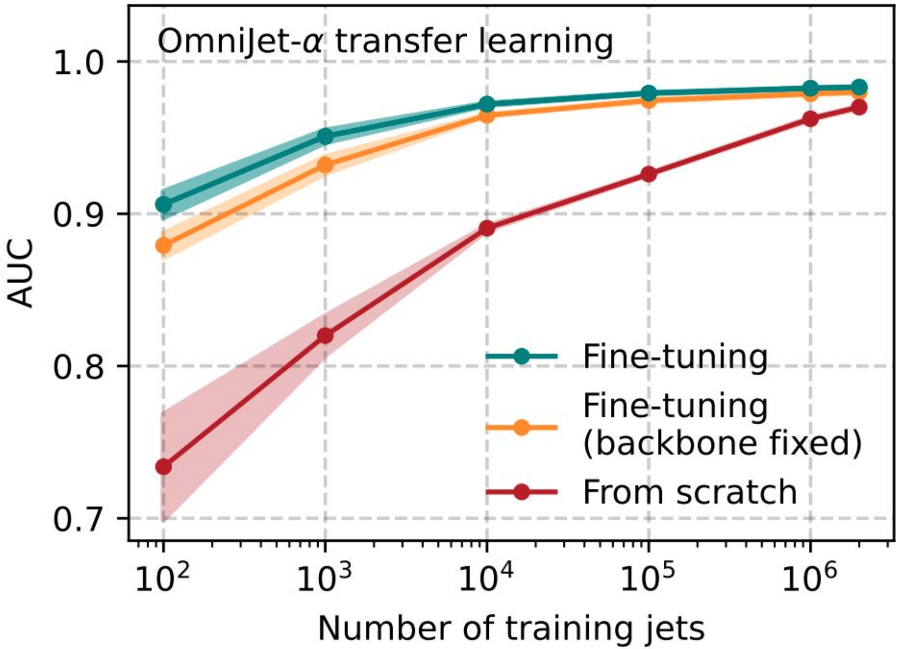
Foundation models are multi-dataset and multi-task machine learning methods that once pre-trained can be fine-tuned for a large variety of downstream applications. The successful development of such general-purpose models for physics data would be a major breakthrough as they could improve the achievable physics performance while at the same time drastically reduce the required amount of training time and data. We report significant progress on this challenge on several fronts. First, a comprehensive set of evaluation methods is introduced to judge the quality of an encoding from physics data into a representation suitable for the autoregressive generation of particle jets with
(the common backbone of foundation models). These measures motivate
lity tokenization compared to previous works. Finally, we demonstrate

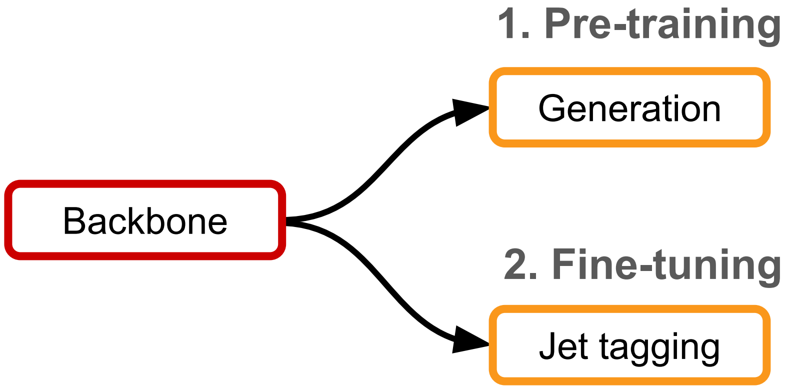
OmniJet-α



[1] Birk et al. *“OmniJet-α: The first cross-task foundation model for particle physics”* (arXiv:2403.05618) (2024)

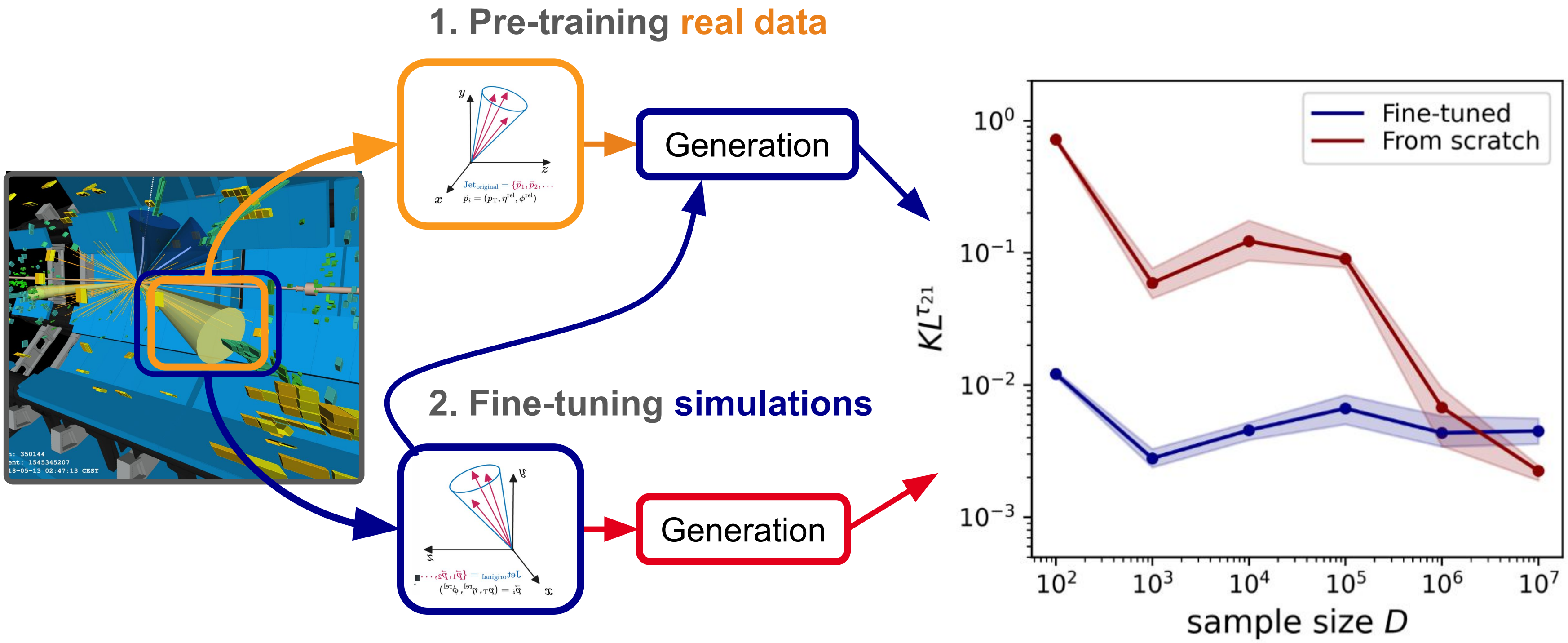
OmniJet-α cross-task





[1] Birk et al. *“OmniJet-α: The first cross-task foundation model for particle physics”* (arXiv:2403.05618) (2024)

OmniJet-α cross-data

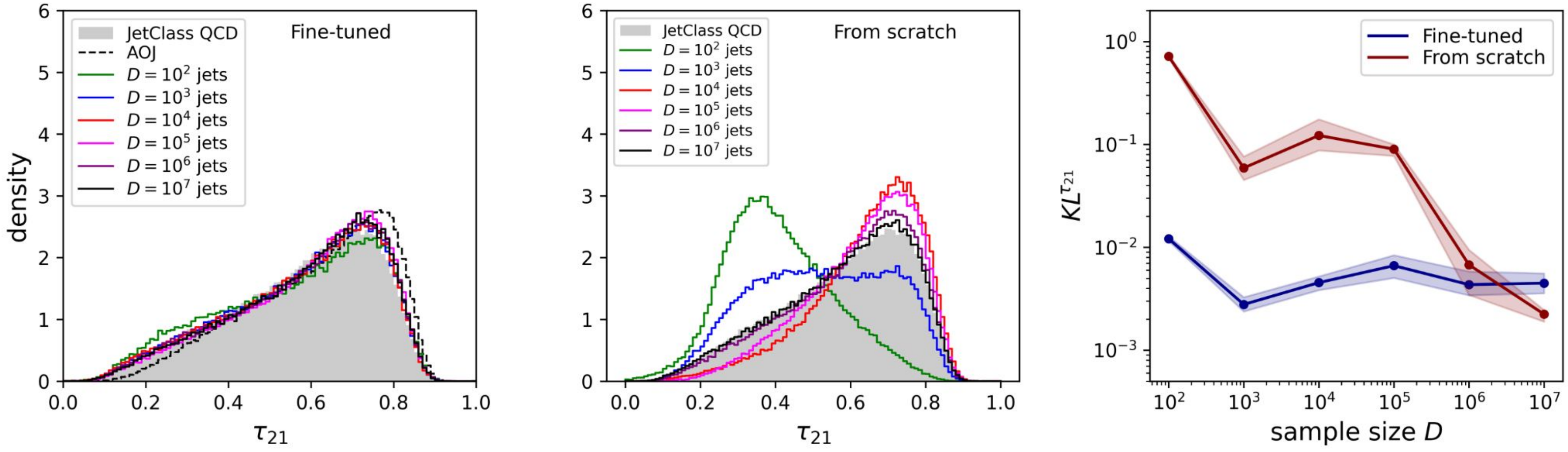


[1] Birk et al. *"OmniJet-α: The first cross-task foundation model for particle physics"* (arXiv:2403.05618) (2024)

OmniJet-α cross-data

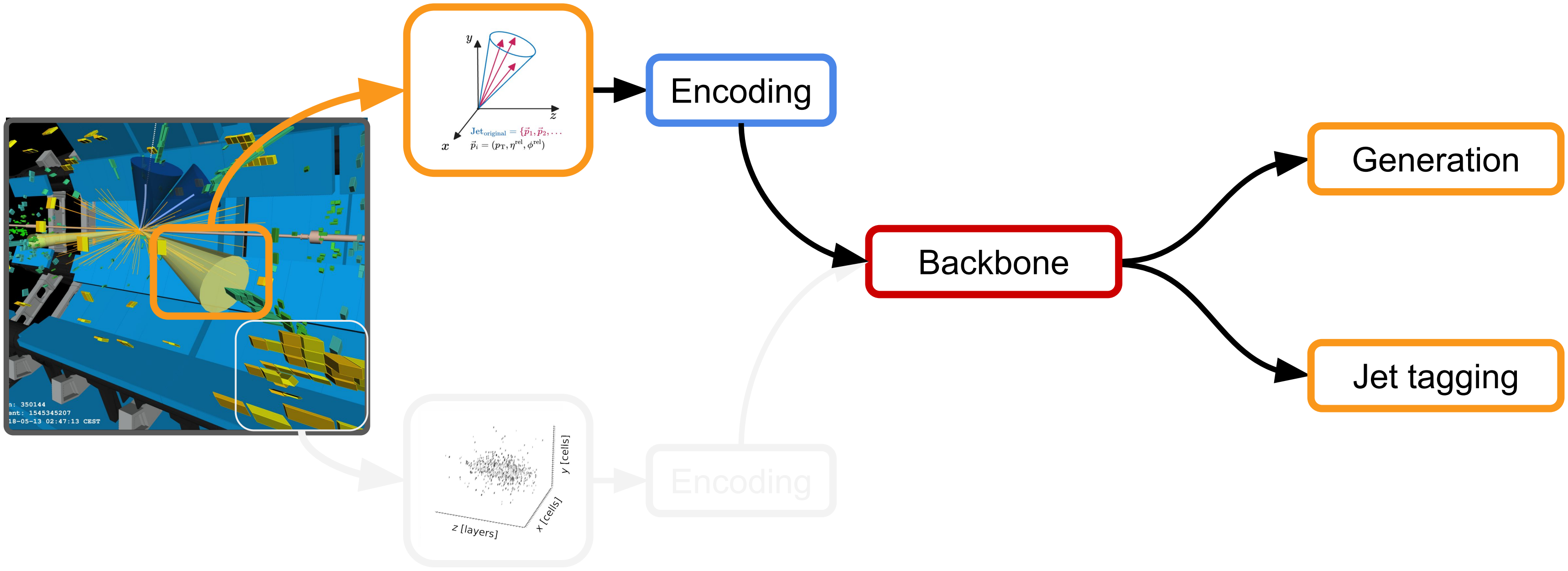
Aspen Open Jets: arxiv.2412.10504

- 180M ML-ready high p_T jets derived from CMS 2016 Open Data



[1] Birk et al. *"OmniJet-α: The first cross-task foundation model for particle physics"* (arXiv:2403.05618) (2024)

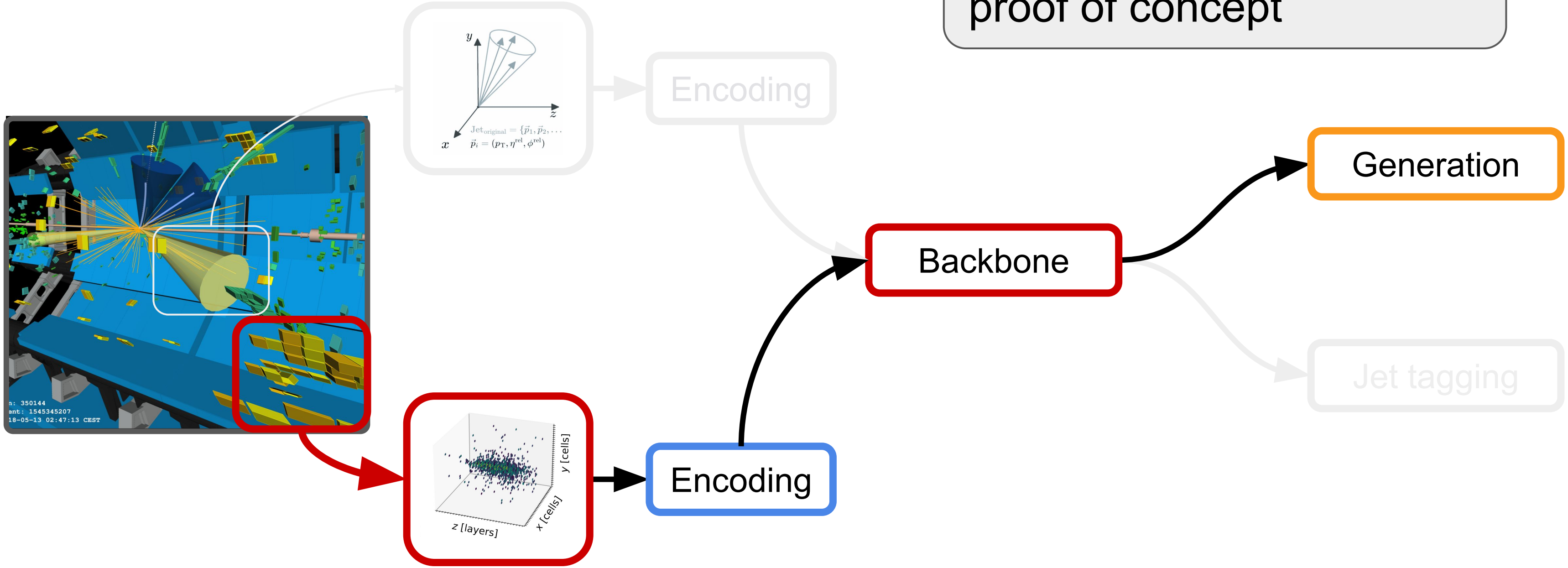
OmniJet-α



[1] Birk et al. *"OmniJet-α: The first cross-task foundation model for particle physics"* (arXiv:2403.05618) (2024)

OmniJet-α Calorimeter

! not a foundation model !
proof of concept



[1] Birk et al. *"OmniJet-α: The first cross-task foundation model for particle physics"* (arXiv:2403.05618) (2024)

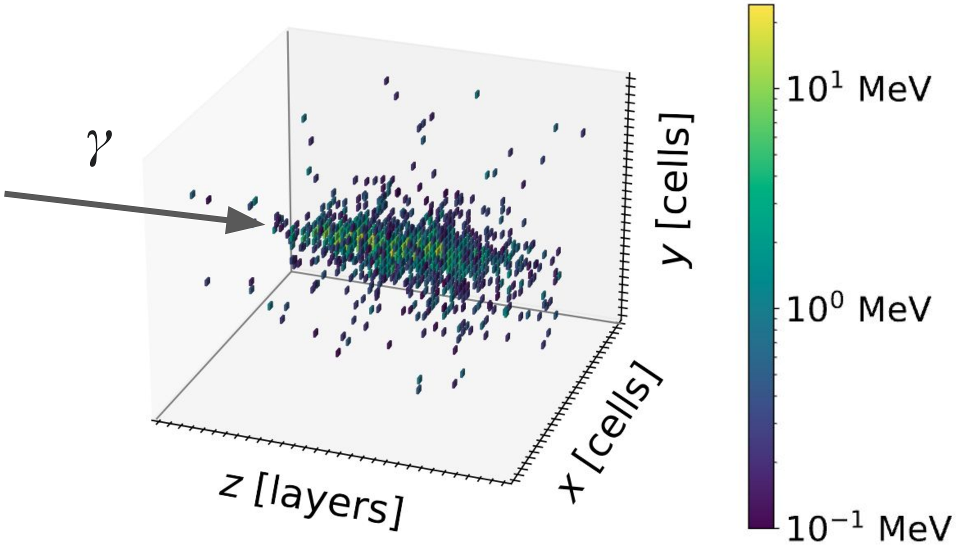
Data

Based on the proposed design of the International Large Detector (ILD)[2]:

- γ-energy 10-100 GeV
- x, y, z $\in$ [0,29]
- voxel energy 0-13 MeV
- 100-1700 hits per shower

[3]

20 times more hits than particles in a jet!



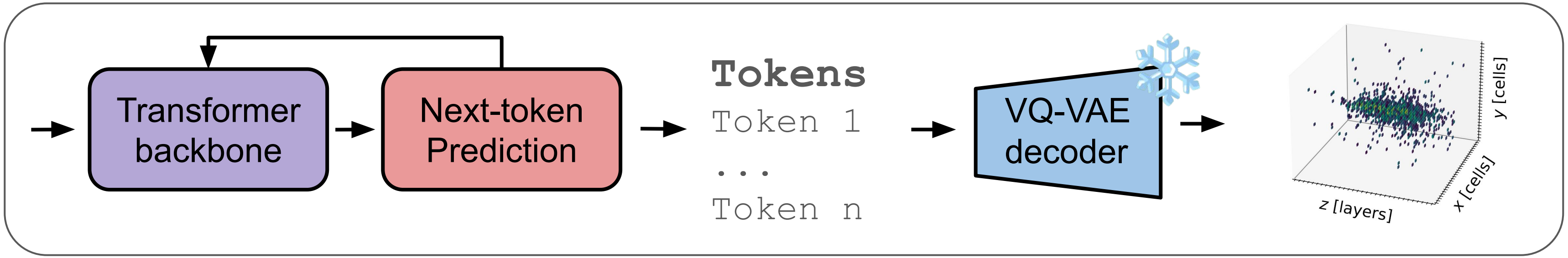
[2] ILD Collaboration *"International Large Detector: Interim Design Report"* (arXiv:2003.01116) (2020)

[3] Buhmann et al. *"Getting High: High Fidelity Simulation of High Granularity Calorimeters with High Speed"* (arXiv:2005.05334) (2020)

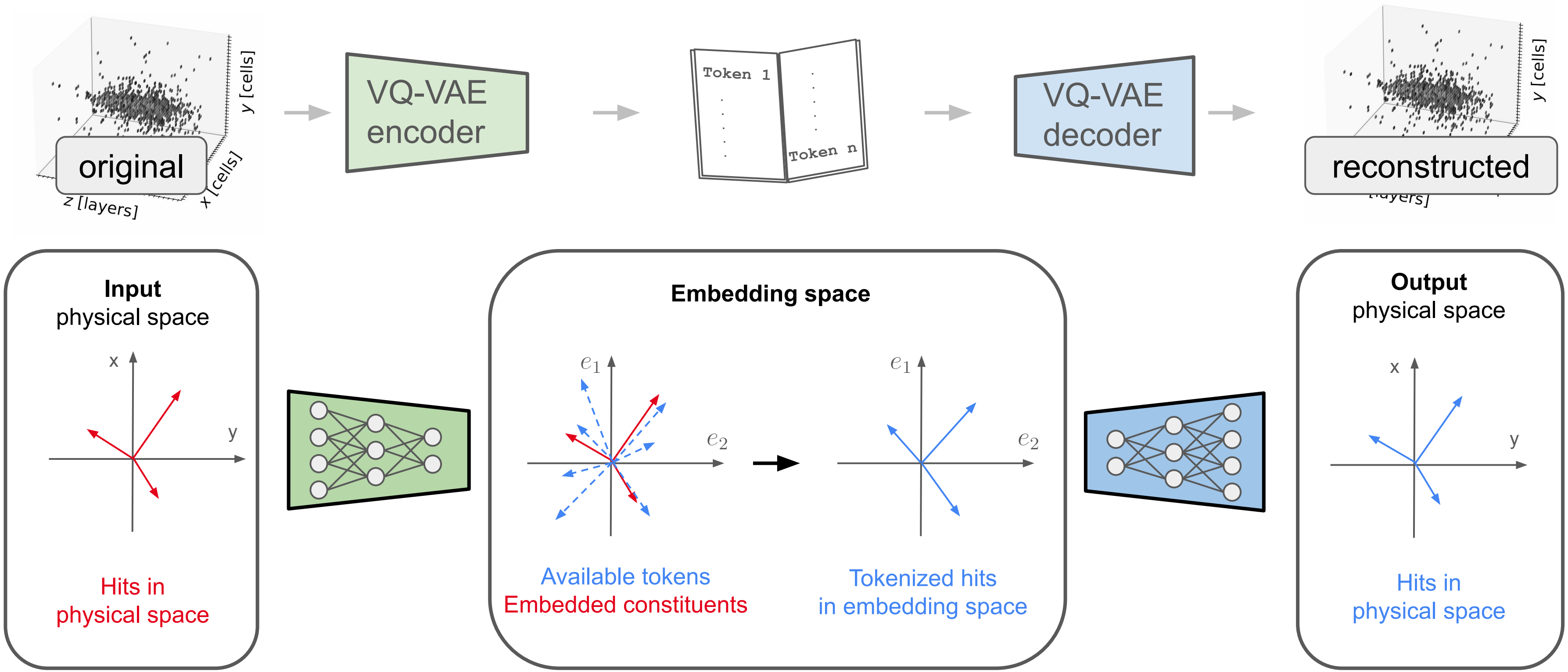
Overview

1. Tokenization / Reconstruction

2. Generation



Tokenization

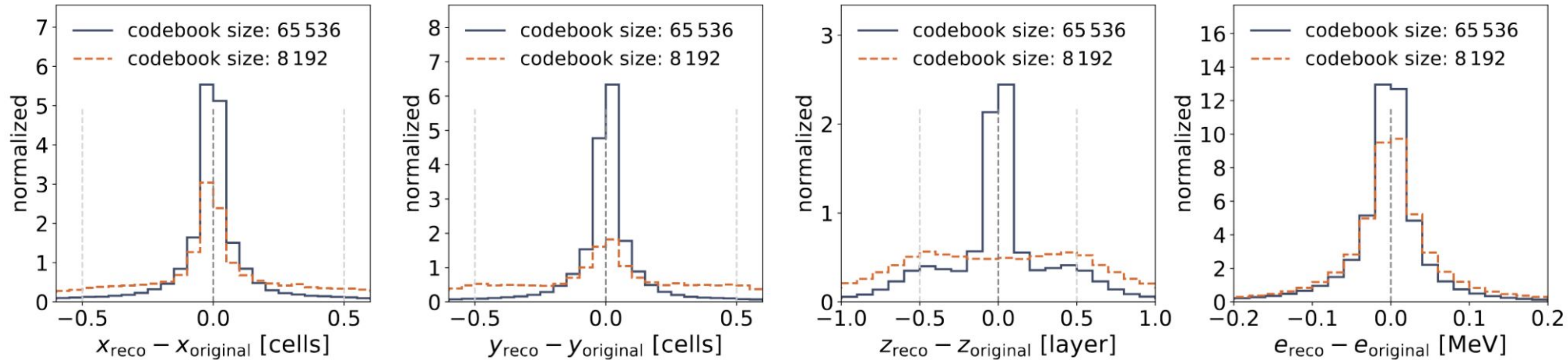


Token quality

30³ possible voxels
= **27.000**

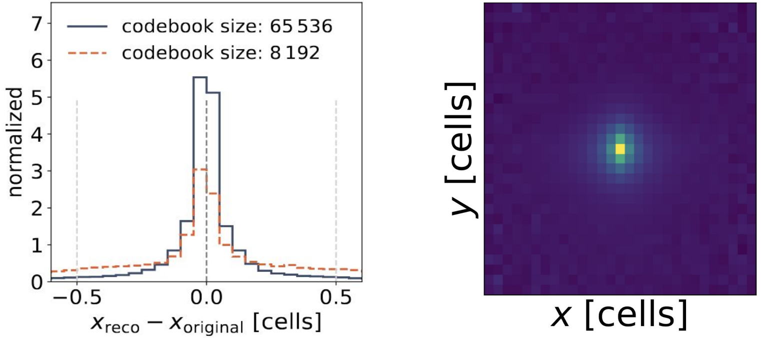
discrete

continuous

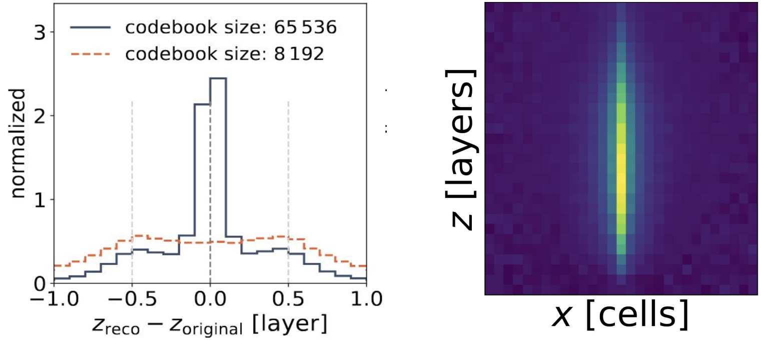


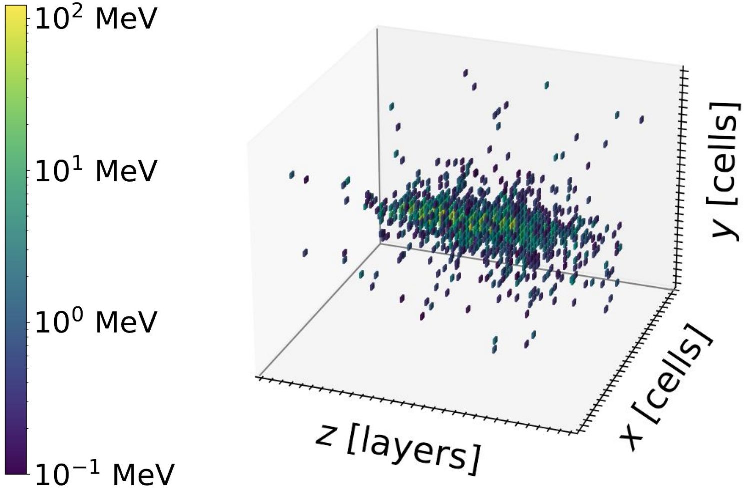
Token quality

x:

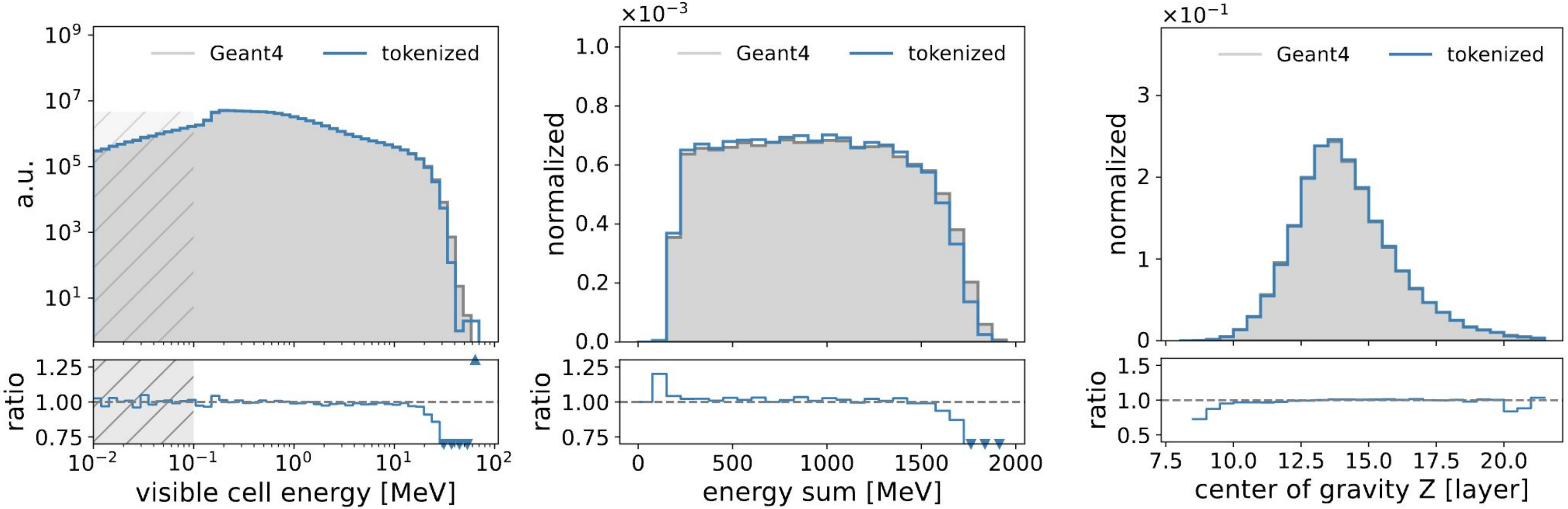


z:



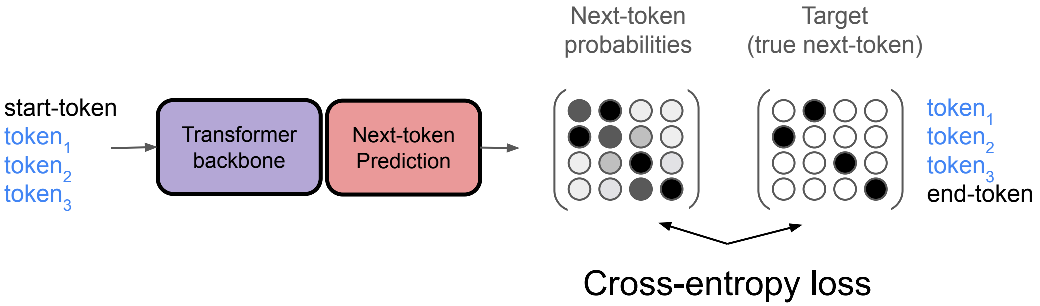


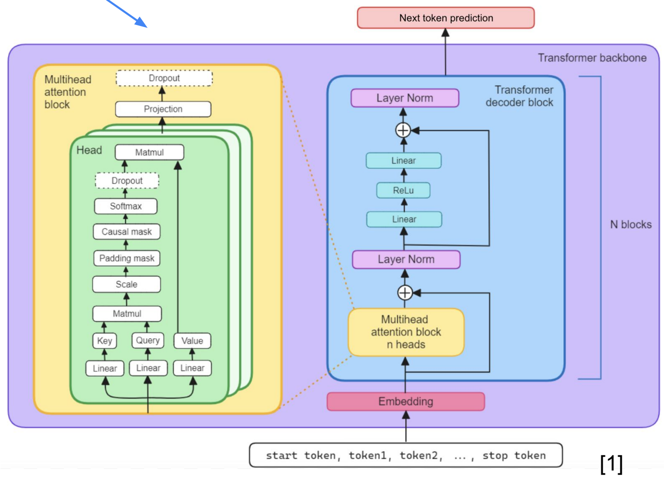
Token quality



Generative training

- Transformer architecture of OmniJet-α [1] (adapted from the original GPT-1 architecture [4])
- Works like a language model, but generates hit-tokens instead of word-tokens





[1]

[1] Birk et al. *"OmniJet-α: The first cross-task foundation model for particle physics"* (arXiv:2403.05618) (2024)
[4] Radford et al, "*Improving language understanding by generative pre-training*" (2018)

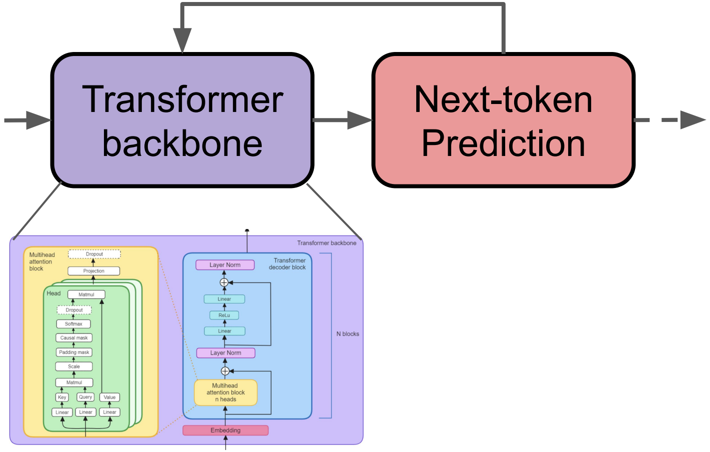
Generative prediction

Shower = {start-token, token_1, ..., token_n, end-token}

token_i = integer value ∈ [1,..., 65,536]

Autoregressive prediction

start-token



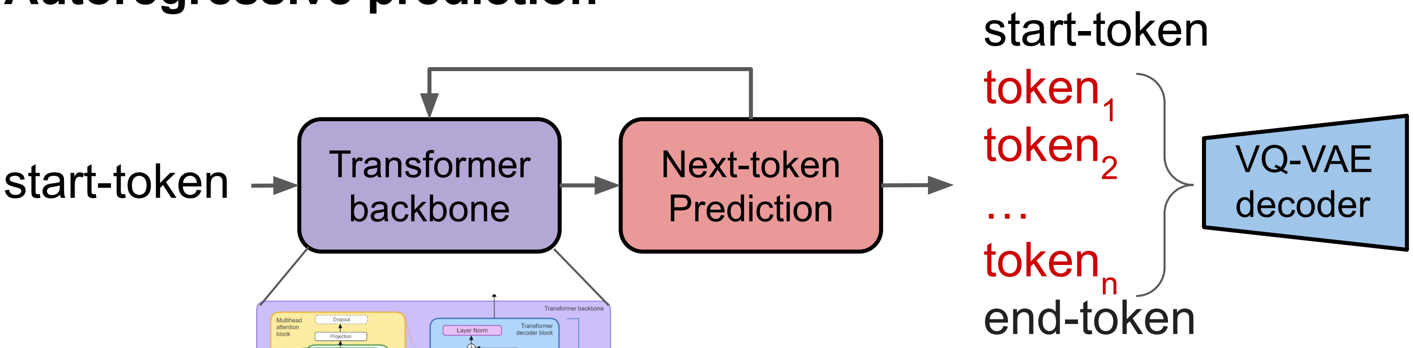
start-token
token_1
token_2
…
token_n
end-token

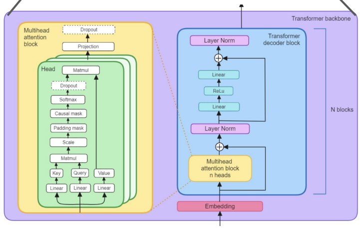
Generative prediction

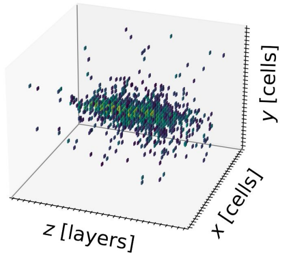
Shower = {start-token, token_1, ..., token_n, end-token}

token_i = integer value $\in$ [1,..., 65,536]

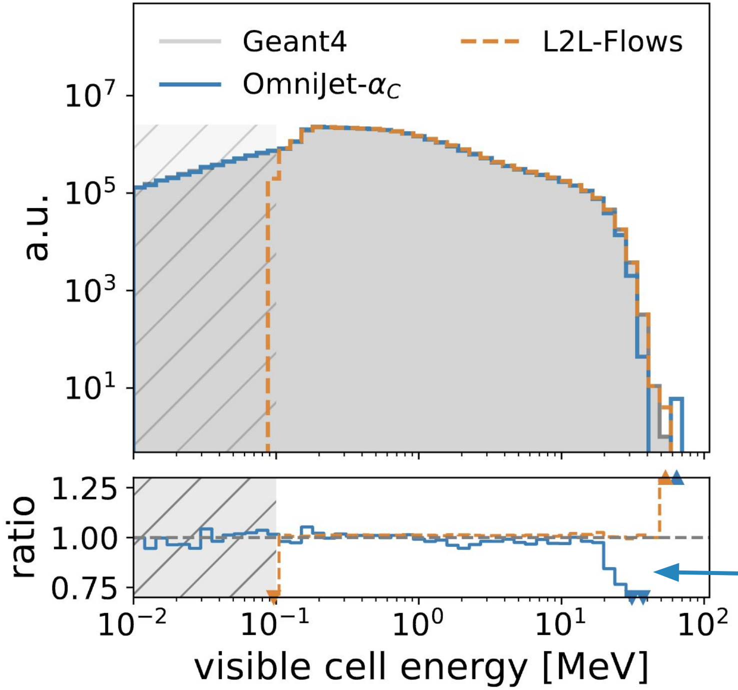
Autoregressive prediction







Results - visible cell energy



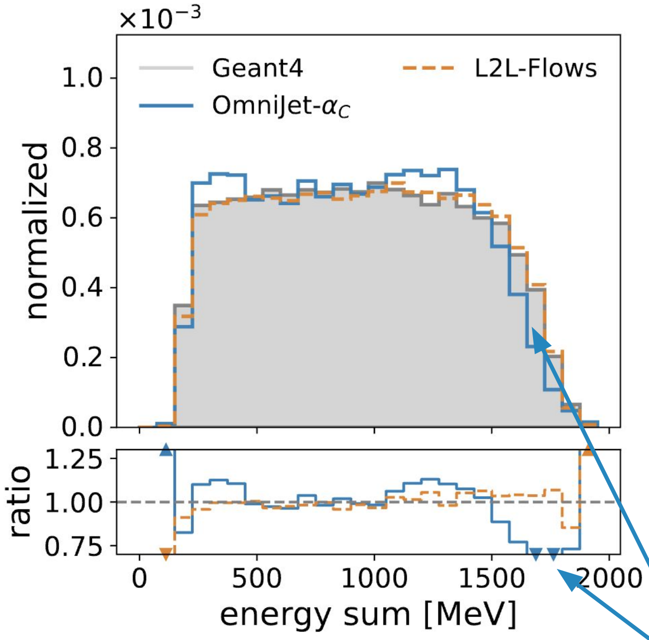
- Comparison of 40k showers
- Only considering MIP hits for analysis
- **Geant4 [3]**
  - Simulated Showers
- **L2LFlows [5]**
  - State-of-the-art generative model
  - Post-processed and calibrated
- **OmniJet-α_C**
  - Post-processed not calibrated

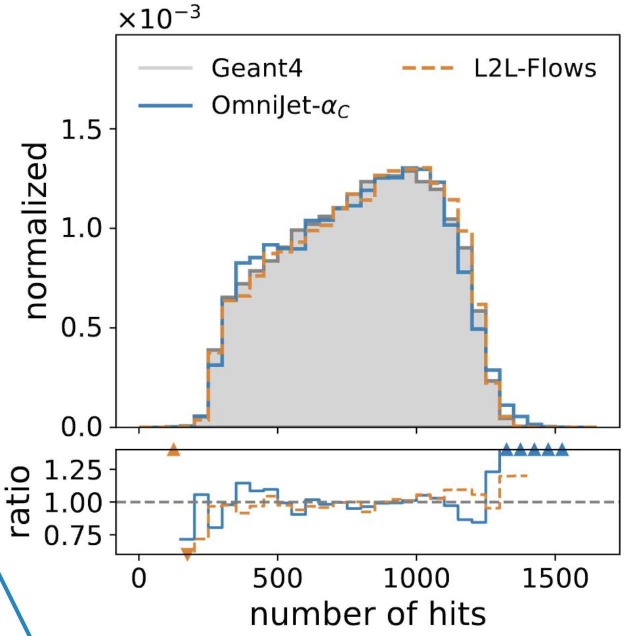
To few high energy hits

[3] Buhmann et al. *"Getting High: High Fidelity Simulation of High Granularity Calorimeters with High Speed"* (arXiv:2005.05334) (2020)

[5] Buss et al. *"Convolutional L2LFlows: Generating Accurate Showers in Highly Granular Calorimeters Using Convolutional Normalizing Flows"* (arXiv:2405.20407) (2024)

Results - energy sum / number of hits

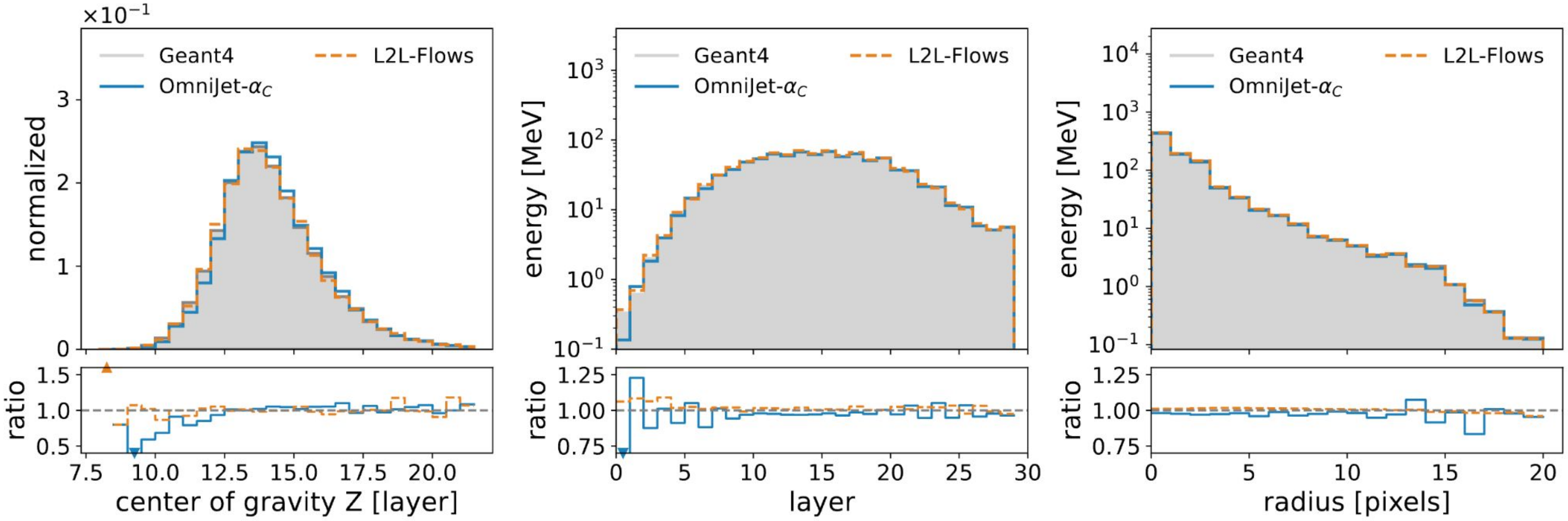




To few high energy showers

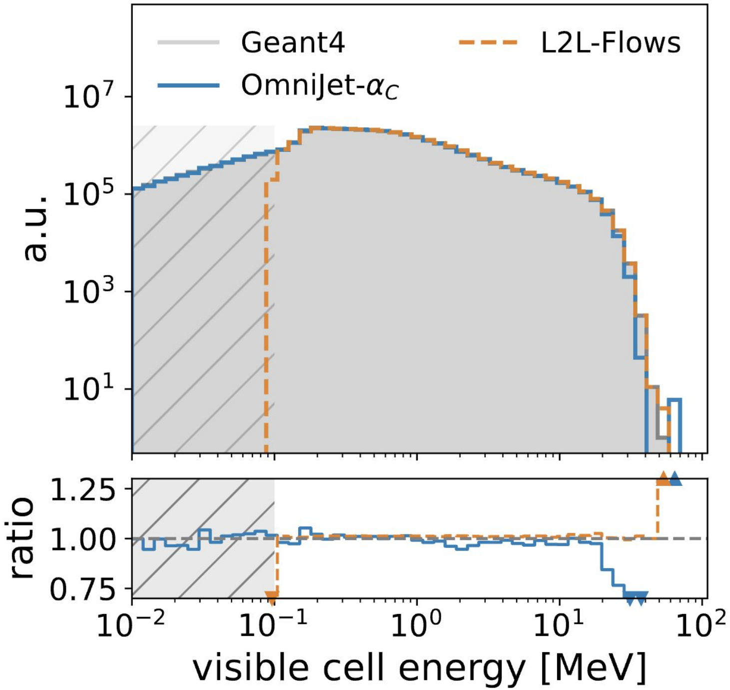
Results - more histograms

Good performance on COG profile, mean energy per layer and energy per radius



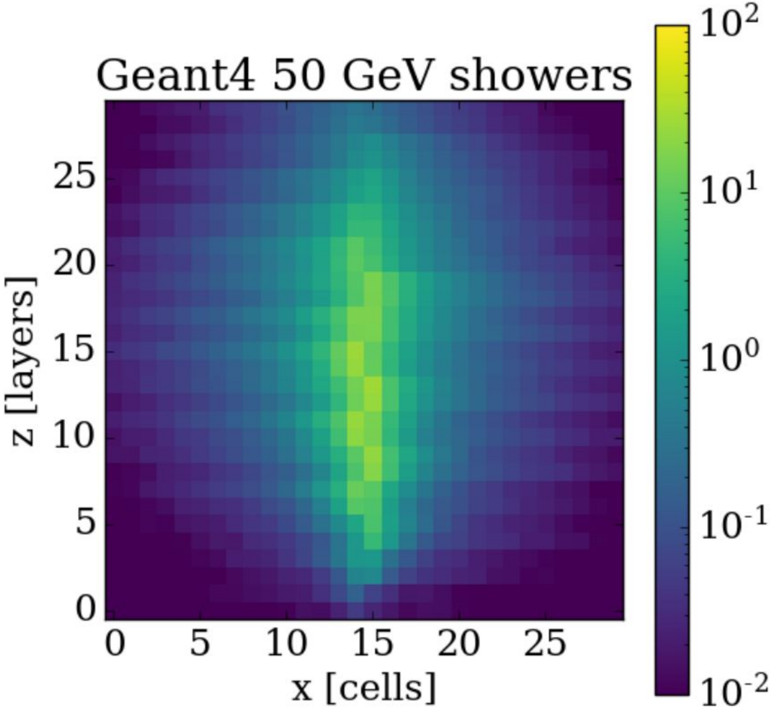
Summary

- Geometry independent
- No conditioning needed
- Generates showers with **good agreement** on shower-level and hit-level
- **Proof of concept** for the versatility of **OmniJet-α**, applying it to a completely different subdomain



Backup

Geometry independence



Overlay of 2000 photon showers along y

Hyperparameters

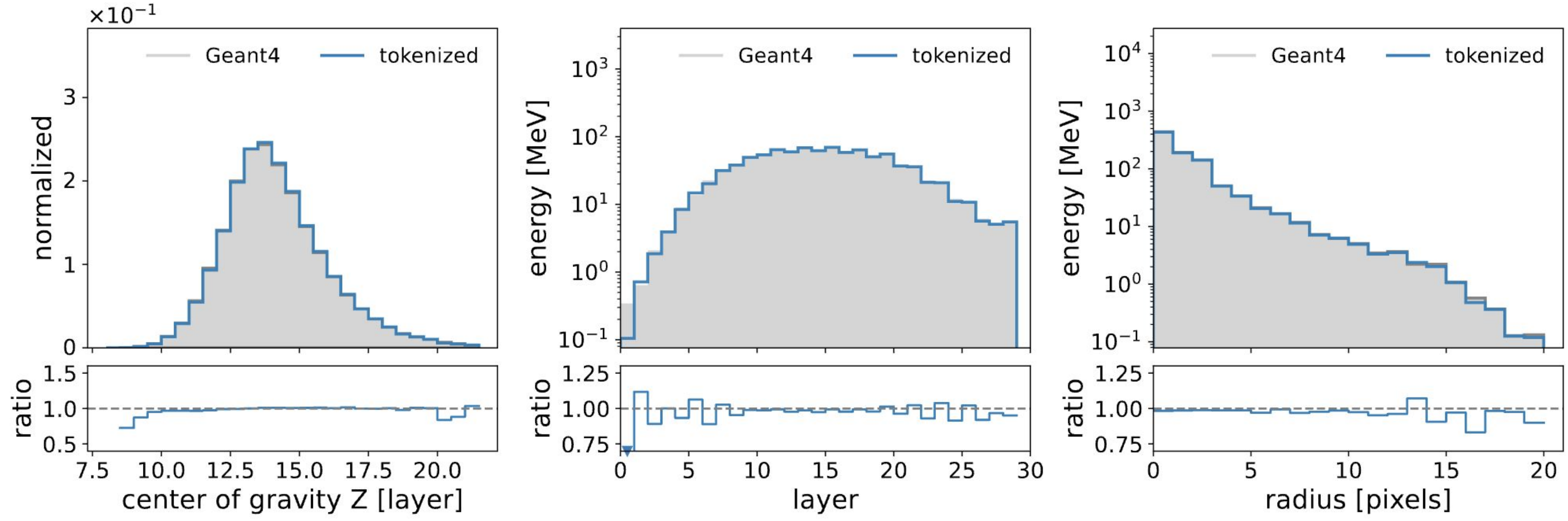
VQ-VAE

Hyperparameter	Value
Learning rate	0.001
Optimizer	Ranger
Batch size	152
Batches per epoch	1000
Number of epochs	588
Hidden dimension	128
Codebook size	65 536
β	0.8
α	10
Replace frequency	100

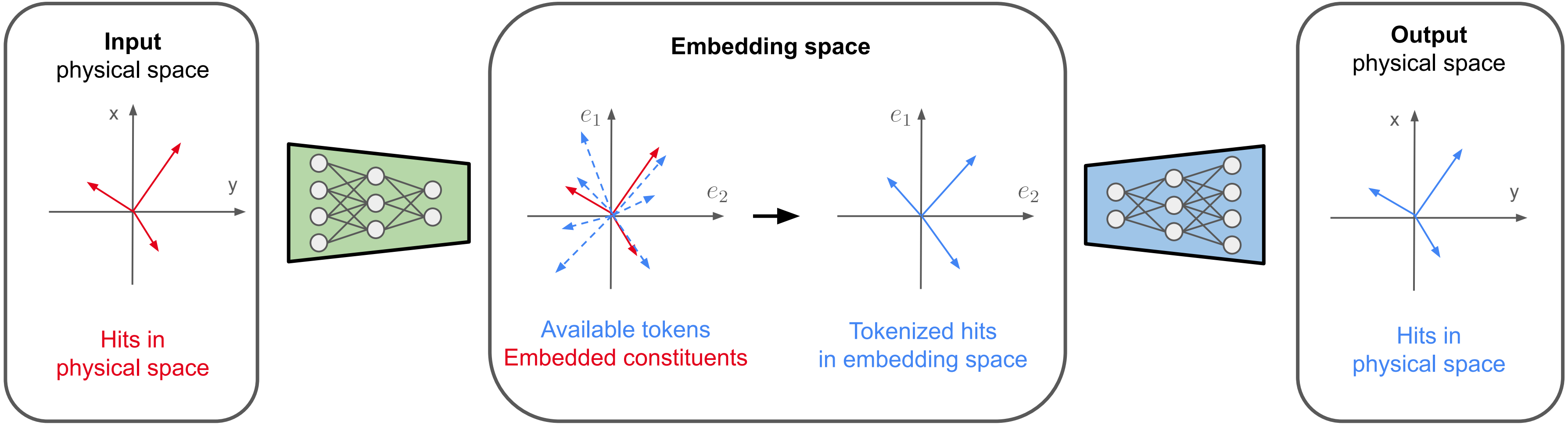
Generative model

Hyperparameter	Value
Learning rate	0.001
Optimizer	Ranger
Batch size	72
Batches per epoch	6000
Number of epochs	106
Embedding dimension	256
Number of heads	8
Number of GPT blocks	3

Token quality



Tokenization



$$\mathcal{L} = \mathcal{L}_{\text{reconstruction}} + \alpha \cdot \mathcal{L}_{\text{commitment}}$$

physical space

embedding space

Important Hyperparameters:
latent dimension: 8
alpha: 10
codebook size: 65.536

Training samples: 855.000 showers

VQ-VAE losses

Reconstruction loss: $\mathcal{L}_{\text{reconstruction}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - x_i')^2$ MSE-loss

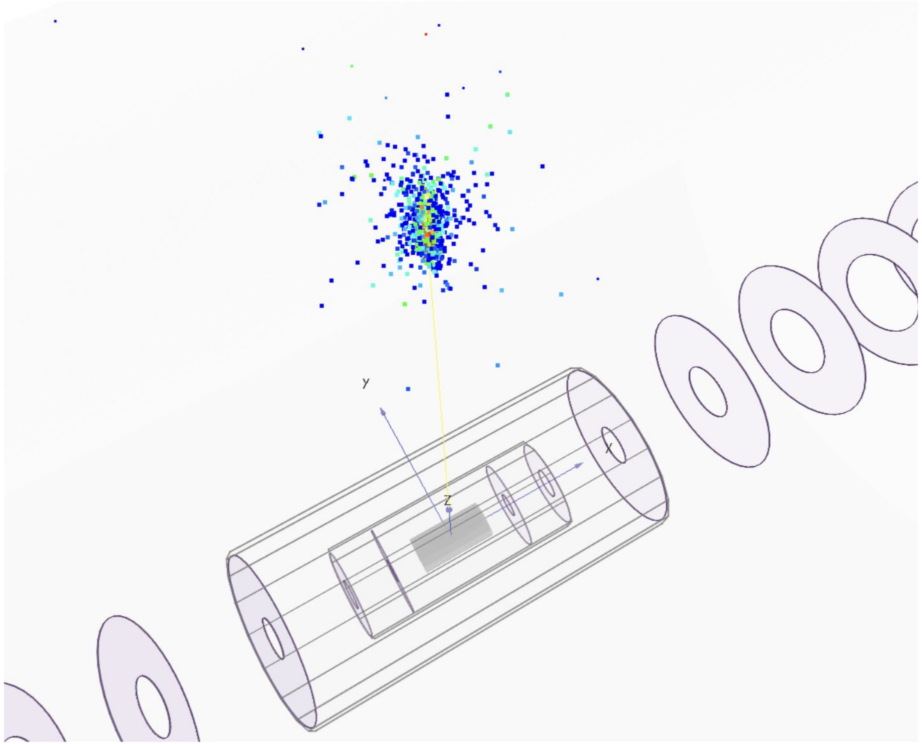
Commitment loss: $\mathcal{L}_{\text{commitment}} = \beta \cdot \|\text{sg}[z_e] - z_q\|^2 + (1-\beta) \cdot \|z_e - \text{sg}[z_q]\|^2$

Complete loss: $\mathcal{L} = \mathcal{L}_{\text{reconstruction}} + \alpha \cdot \mathcal{L}_{\text{commitment}}$

sg represents stop gradient operator meaning no gradient

straight-through estimator (STE) is used to pass the gradients straight through the quantization operation - to ensure the **STE** is accurate, the codebook and the encoder representations are pulled together using the **sg**

Data - What is a point cloud?



Simulated shower in the electromagnetic calorimeter of the envisioned International Large Detector (ILD)

(2005.05334) "Getting High" generates geometry-independent calorimeter showers as point clouds.