The OmniJet-α Foundation Model

Extending to Calorimeter Showers

8th Round table on Deep Learning
November 14th 2025

Joschka Birk, Frank Gaede, Anna Hallin, Gregor Kasieczka, Martina Mozzanica,
Henning Rose

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

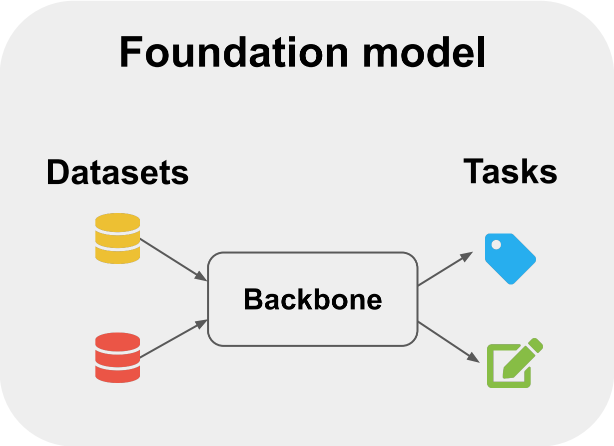
CLUSTER OF EXCELLENCE
QUANTUM UNIVERSE

What is OmniJet-α?

- First cross-task **Foundation** model for jet-physics[1]

- But we aim to build a foundation model for all particle physics

- Applied on Calorimeter showers to show **Adaptability** of the architecture[2]

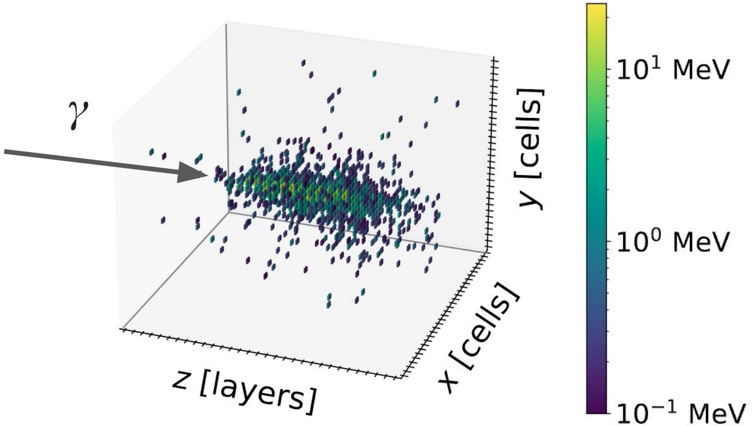


1 Birk et al. *"OmniJet-α: The first cross-task foundation model for particle physics"* 2403.05618 (2024)
2 Birk et al. *"OmniJet-α_C: Learning point cloud calorimeter simulations using generative transformers"* 2501.05534 (2025)

Why Calorimeter Showers?

- γ-showers in the ECAL in the proposed International Large Detector (ILD) [4]
  = **long sequences**, sparse spatial structure

- Perfect stress test for **architecture scaling**



Shower Properties

- γ-energy 10-100 GeV
- x, y, z in 3D grid (30^3 voxels)
- hit energy 0-13 MeV
- 100-1700 hits per shower

[5]

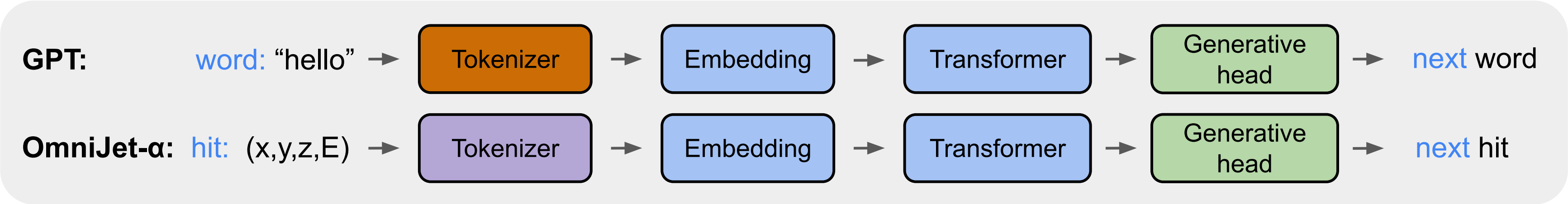
~30 times longer than jets

4 ILD Collaboration *"International Large Detector: Interim Design Report"* 2003.01116 (2020)
5 Buhmann et al. *"Getting High: High Fidelity Simulation of High Granularity Calorimeters with High Speed"* 2005.05334 (2020)

VQ-VAE = Vector Quantised-Variational AutoEncoder

OmniJet-α architecture

- Autoregressive decoder-only Transformer (GPT-1 style)[3]
- Predicts the next hit-token, one at a time, forming showers sequentially
- GPTs use **Byte-Level Tokenizers** (codebooksize ~40k)
- **VQ-VAE** acts as a **Tokenizer**: it turns continuous calorimeter data into discrete hit-tokens

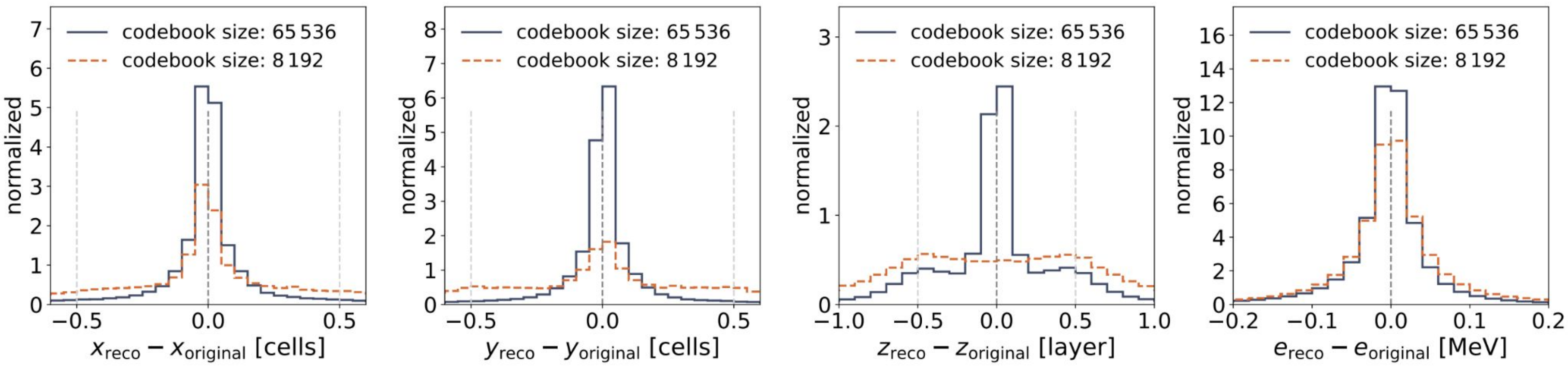


! **Problem:** Input data is continuous → VQ-VAE introduces **loss of information**

3 Radford et al, "*Improving language understanding by generative pre-training*" (2018)

Loss of information

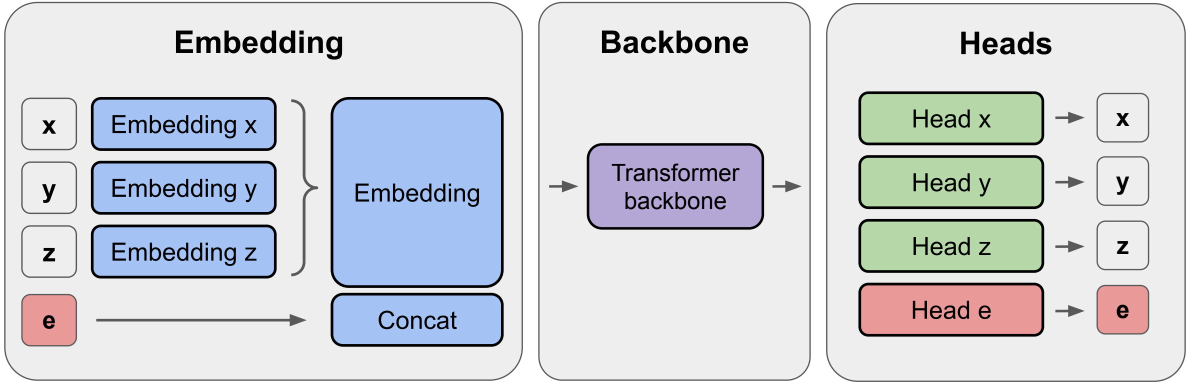
- Going from shower to token-space and back, introduces an error



Upgrades to OmniJet-α

Direct Spatial Tokenization:

- Instead of compressing hits to tokens, encode coordinates directly in the embedding
- Full spatial precision preserved
- Simpler pipeline → closer to a universal architecture

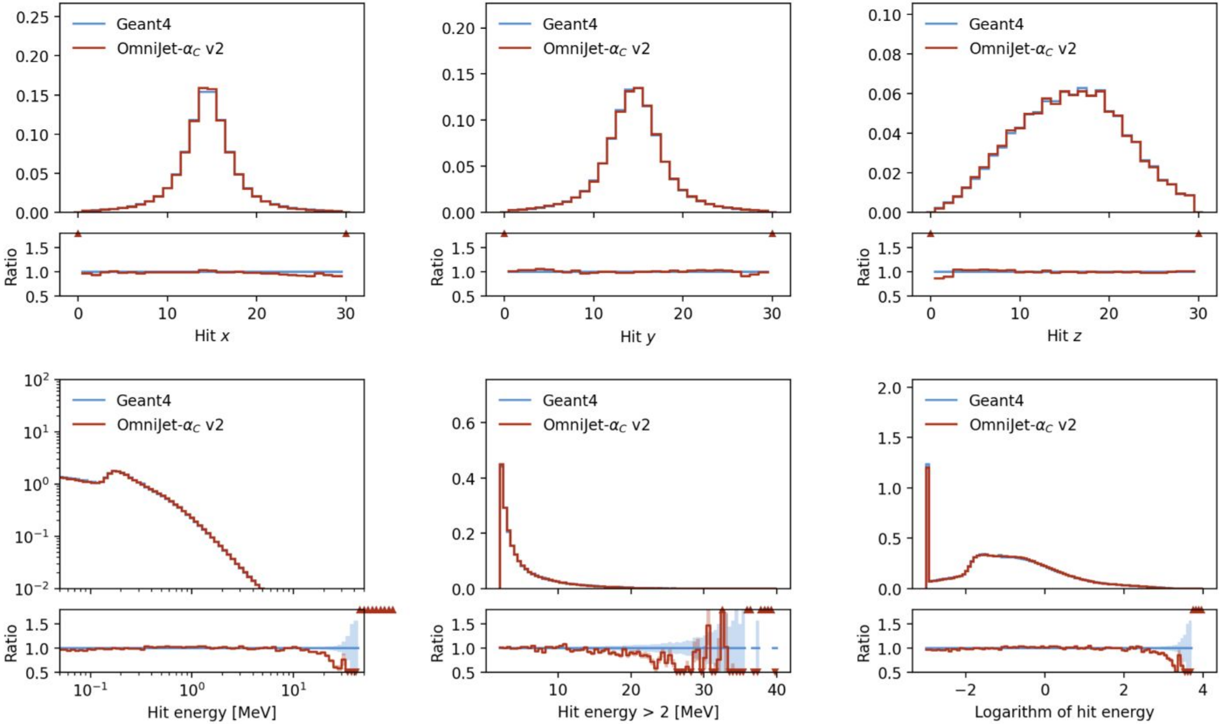


Use **Mixture of Gaussians** to sample continuous **energy values**

Results

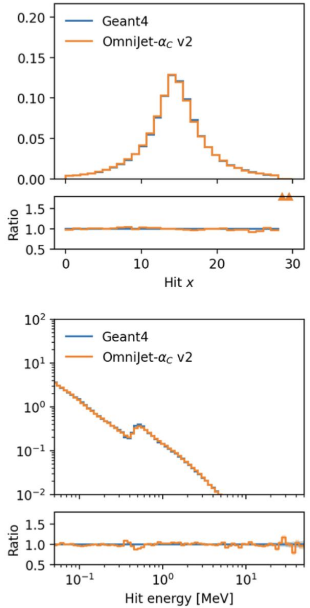
Hit-Level Distributions:

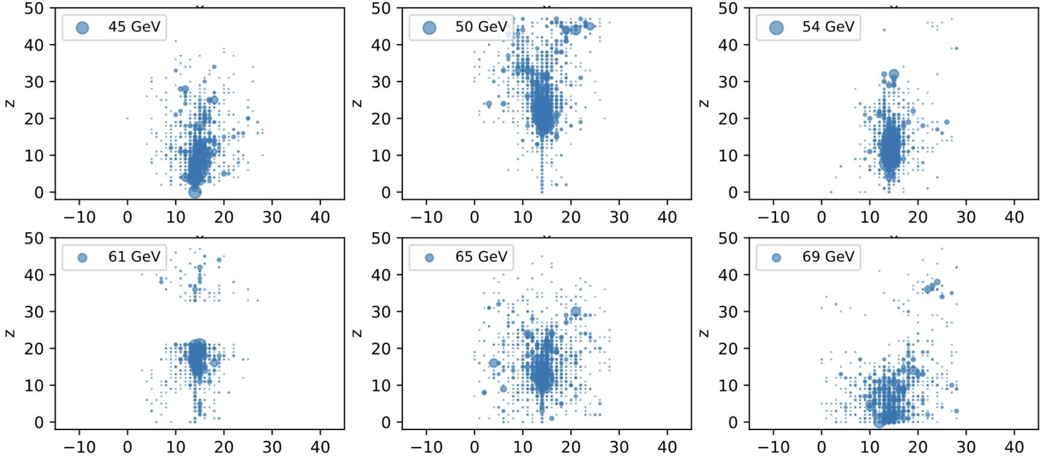
- Spatial distributions are well reproduced
- Hit-energy spectrum matches Geant4 over 3 orders of magnitude



Results - HCAL

- OmniJet-α can also be trained to model pion showers in an HCAL





Plots produced by Martina Mozzanica

Conclusion

- Direct spatial tokenization **removes Tokenization** bottlenecks
  - Full spatial precision, simpler pipeline
- **High-fidelity generation** across ECAL & HCAL
  - Spatial and energy distributions match Geant4
- Architecture remains **flexible for future extensions**
  - Same framework can incorporate additional detectors without major changes

Results

Global-Level Distributions:

- Global distributions are well reproduced