ParticleCNNTransformer

Hierarchical CNN–Transformer for 3D Hadronic Shower Momentum Regression

Tamás Menyhárt, Róbert Lakatos, Dr. András Hajdu

Architecture

CNN + 3-Stage Transformer

Input

64^3 Voxel Energy Grid

Task

Proton Momentum Regression

Performance

MAE = 1.085 GeV

Mean Absolute Error

RMSE = 1.843 GeV

Root Mean Square Error

2.16 M

A few trainable Parameters

Calorimeter Reconstruction • Deep Learning • Particle Physics

Problem statement

Why 3 levels of self-attention and CNN are important?

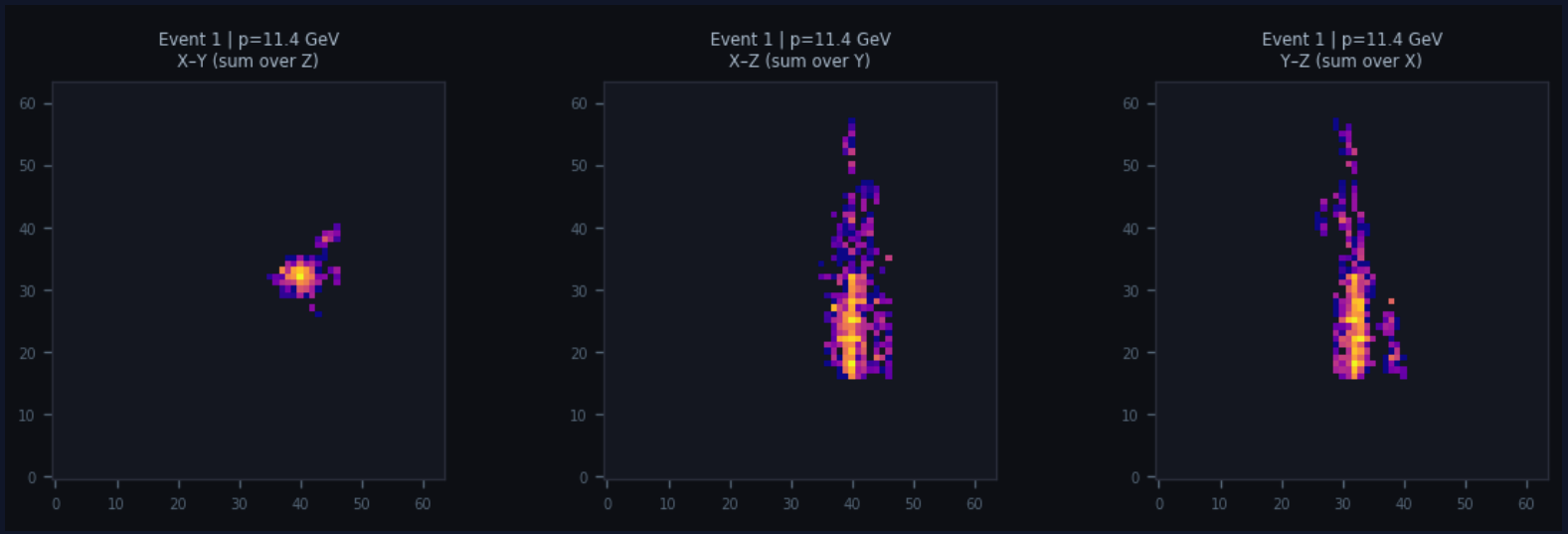
The Physics Problem

- Hadronic calorimeters measure particle showers deposited as 3D energy patterns
- Incoming proton momentum (1–20 GeV) must be inferred from the shower topology
- Traditional methods: calibration curves, linear weighting — limited for complex showers
- High-energy showers produce diffuse, multi-scale structures hard to characterise analytically
- The model lacks geometric awareness and the latent space is highly sparse.

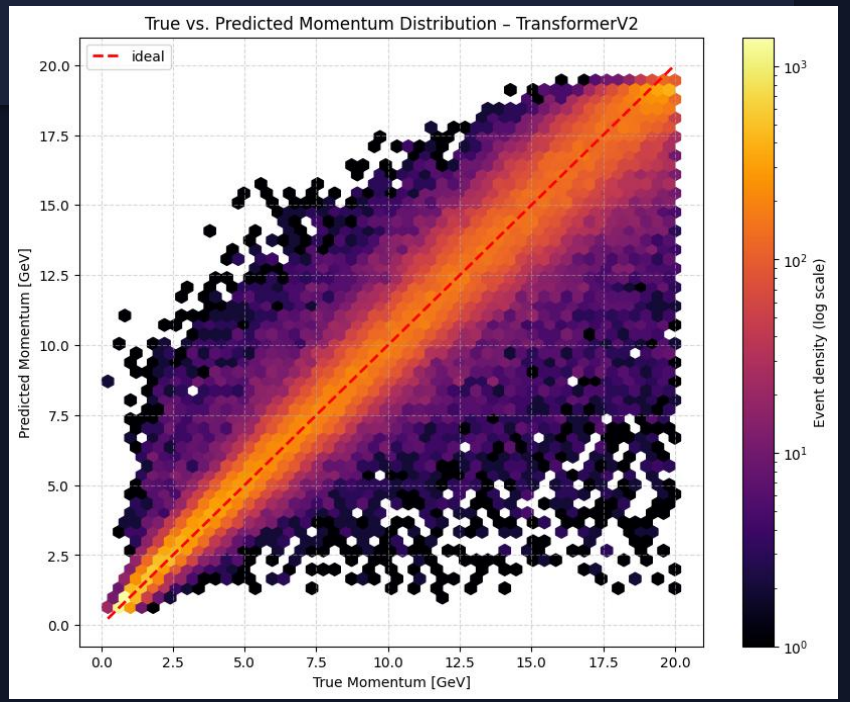
Why This Architecture?

- Convolutional nets capture local features but struggle with long-range correlations
- Plain transformers on **$64^3 = 262{,}144$ tokens**: prohibitive **$O(N^2)$ attention cost**
- Multi-scale processing naturally matches the hierarchical structure of particle showers
- Attention mechanisms can learn which spatial regions drive the momentum estimate

Sample Input Events — 3-Axis Energy Projections

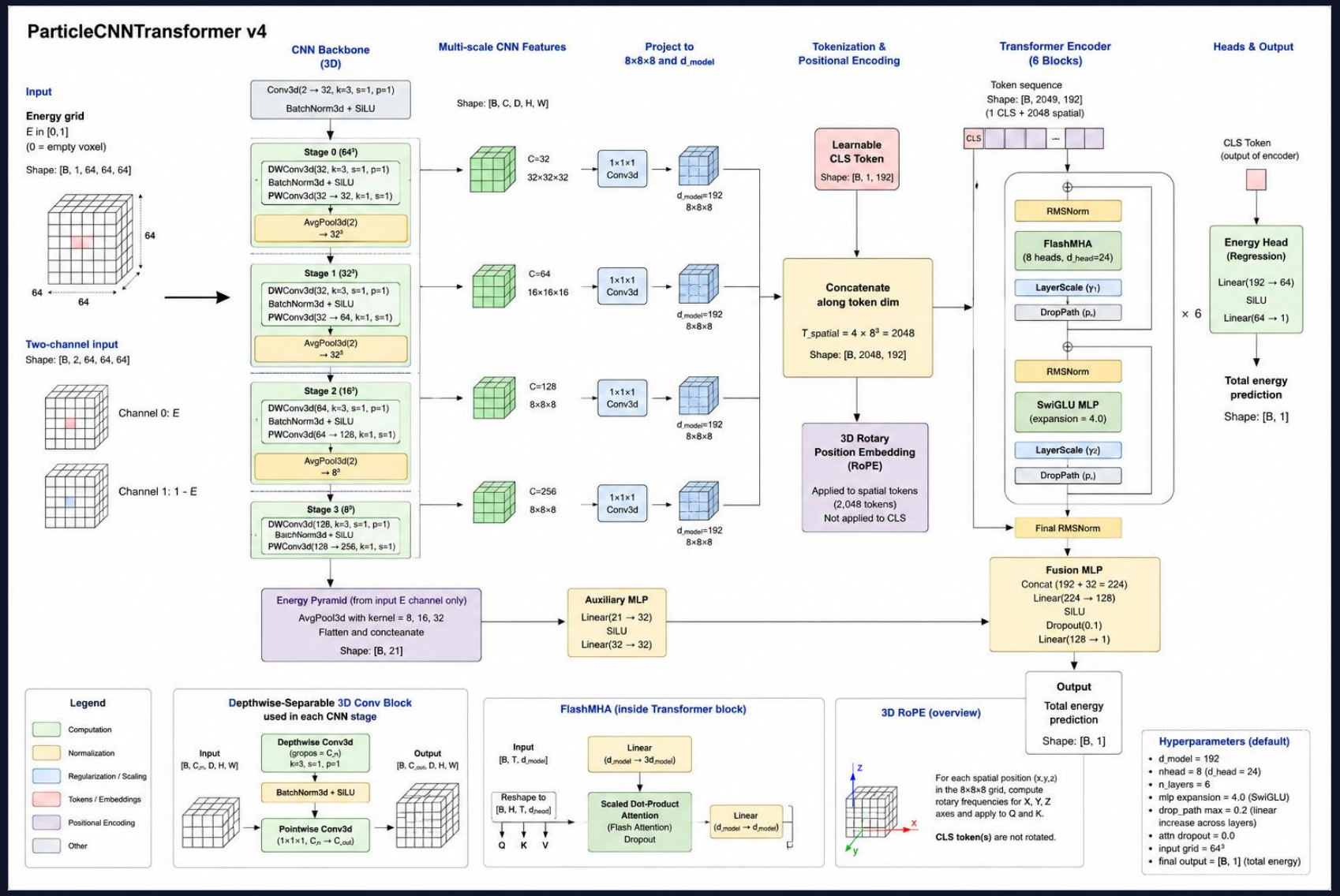


Density is LOGARITHMIC ! `MAE=1.0851|RMSE=1.8426`



Model Architecture

ParticleCNNTransformer — full pipeline overview



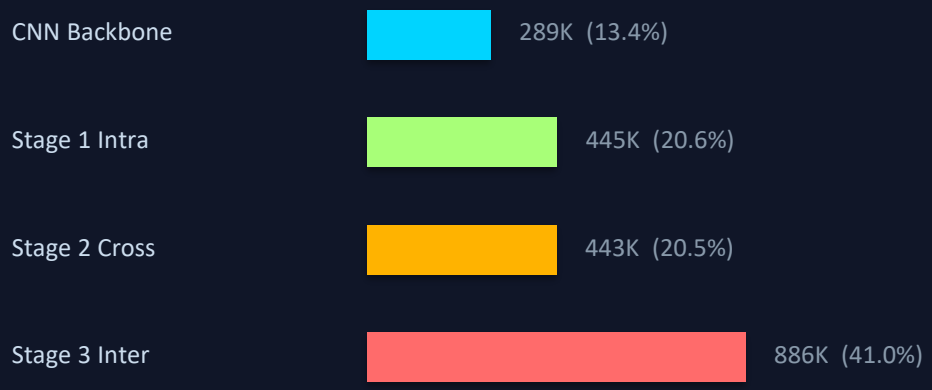
- **Input tensor construction**
  - Two-channel input: energy + complement (1 − energy)
  - Improves stability and representation richness
- **Hierarchical CNN levels (L0–L3)**
  - Progressive downsampling: $64^3 \rightarrow 32^3 \rightarrow 16^3 \rightarrow 8^3$
- **IntraSpatialAttn**
  - Splits each scale into spatial blocks (CLS token per block)
  - Capture local geometric patterns
- **CrossScaleAttn** (multi-scale fusion)
  - CLS token aggregates scale-wise information
- **InterBlockAttn** (global aggregation)
- **$O(N^2)$ attention cost** now reduced with 3-level aggregation.

Model Efficiency & Parameters

2.16M parameters — transformer stages dominate; CNN backbone is lightweight

```
Parameter breakdown
-------------------------------------------------
CNN backbone                          288,822
Stage 1 intra-block                   444,672
Stage 2 cross-scale                   443,328
Stage 3 inter-block                   886,464
Pool + head                            97,153
-------------------------------------------------
TOTAL                               2,160,439
```

Parameter Distribution



Regularisation & Training Techniques

- LayerScale (init 1e-4): residual paths open gradually — prevents early instability
- DropPath / Stochastic Depth: drops entire residual paths per sample — stronger than Dropout
- RMSNorm instead of LayerNorm: faster, no mean subtraction, matches LLaMA/Gemma design
- AdamW with differential weight decay: LayerScale params exempt — prevents them being shrunk to zero
- Linear warmup + cosine decay learning rate schedule

CNN Backbone

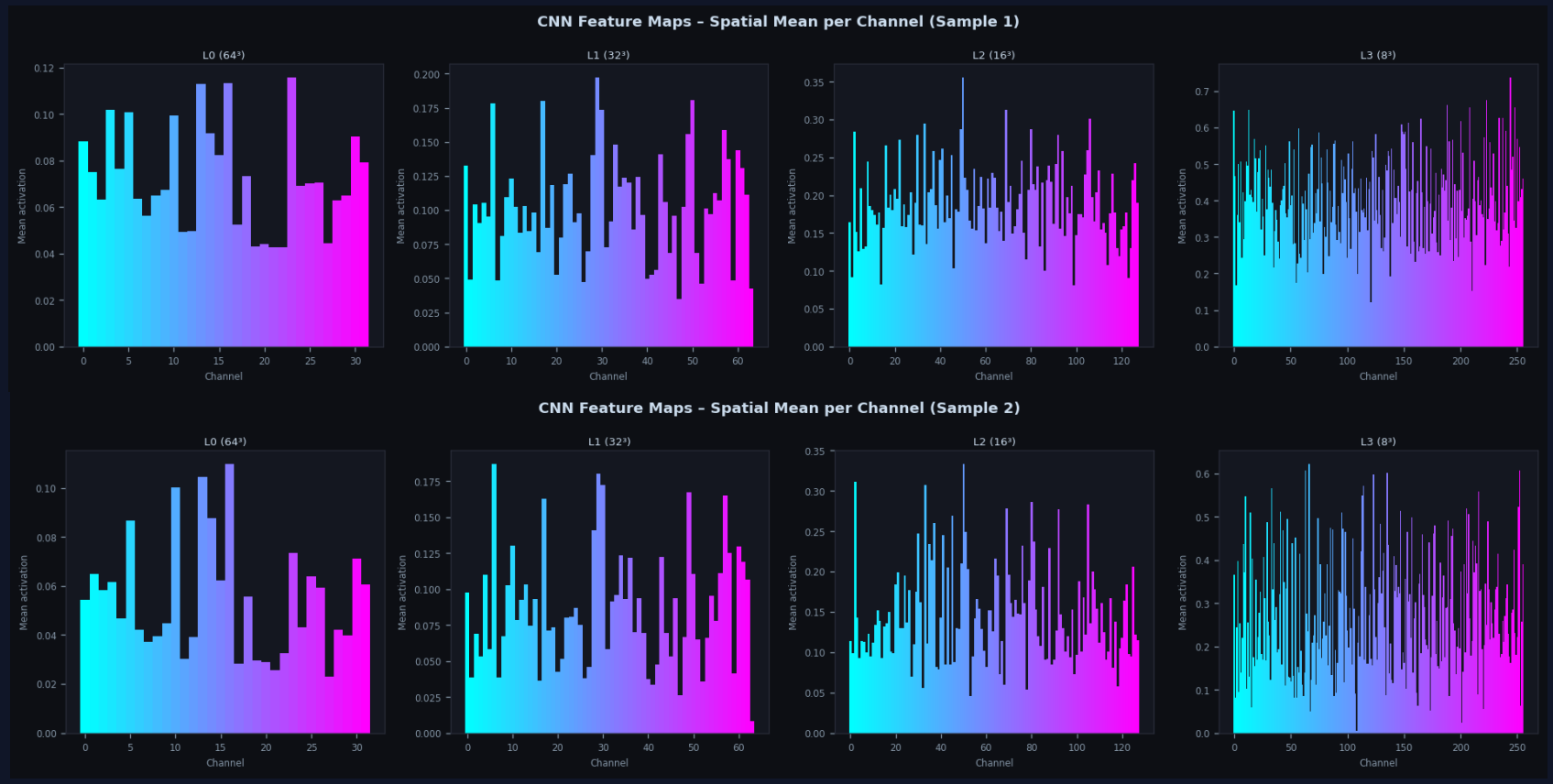
Depthwise-separable 3D convolutions across 4 resolution scales

L0 64^3 — 32 ch → **L1 32^3** — 64 ch → **L2 16^3** — 128 ch → **L3 8^3** — 256 ch

Highest activations

→ *All projected to $8^3 \times d\_model=192$ for transformer input*

CNN Feature Maps — Spatial Mean Activation per Channel (Sample 1 and 2). Highest values: 0.1, 0.175, 0.35, 0.6



CNN Backbone

Depthwise-separable 3D convolutions across 4 resolution scales

L0 64^3

32 ch

→

L1 32^3

64 ch

→

L2 16^3

128 ch

→

L3 8^3

256 ch

→ All projected to $8^3 \times d\_model=192$ for transformer input

Activation grows with depth — L3 captures most abstract, high-energy features. Features that describe the whole figure are more important than small details. (Possible noise?)

Three-Stage Transformer Encoder

Hierarchical attention from local to global — $O(N^2)$ cost reduced ×7.8

Stage 1

IntraSpatialAttn

- 8 spatial blocks, each 4^3=64 tokens
- Full self-attention per block + CLS token
- 3D RoPE positional encoding (relative distances)
- Output: 8 block CLS tokens per scale
- Flat attention: 512^2 → 8×65^2 (×7.8 faster)

Stage 2

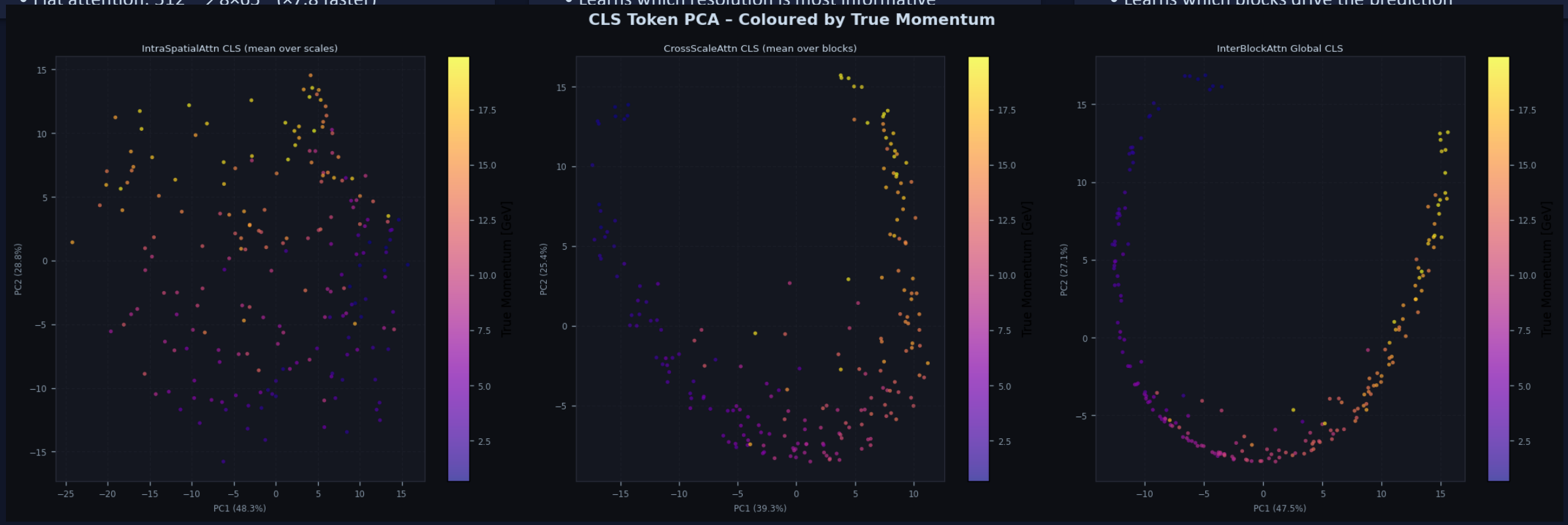
CrossScaleAttn

- Per block: attends across 4 scale CLS tokens
- Sequence length: only 5 tokens (S+1)
- No RoPE — scale index is not spatial
- Output: 8 fused block CLS tokens
- Learns which resolution is most informative

Stage 3

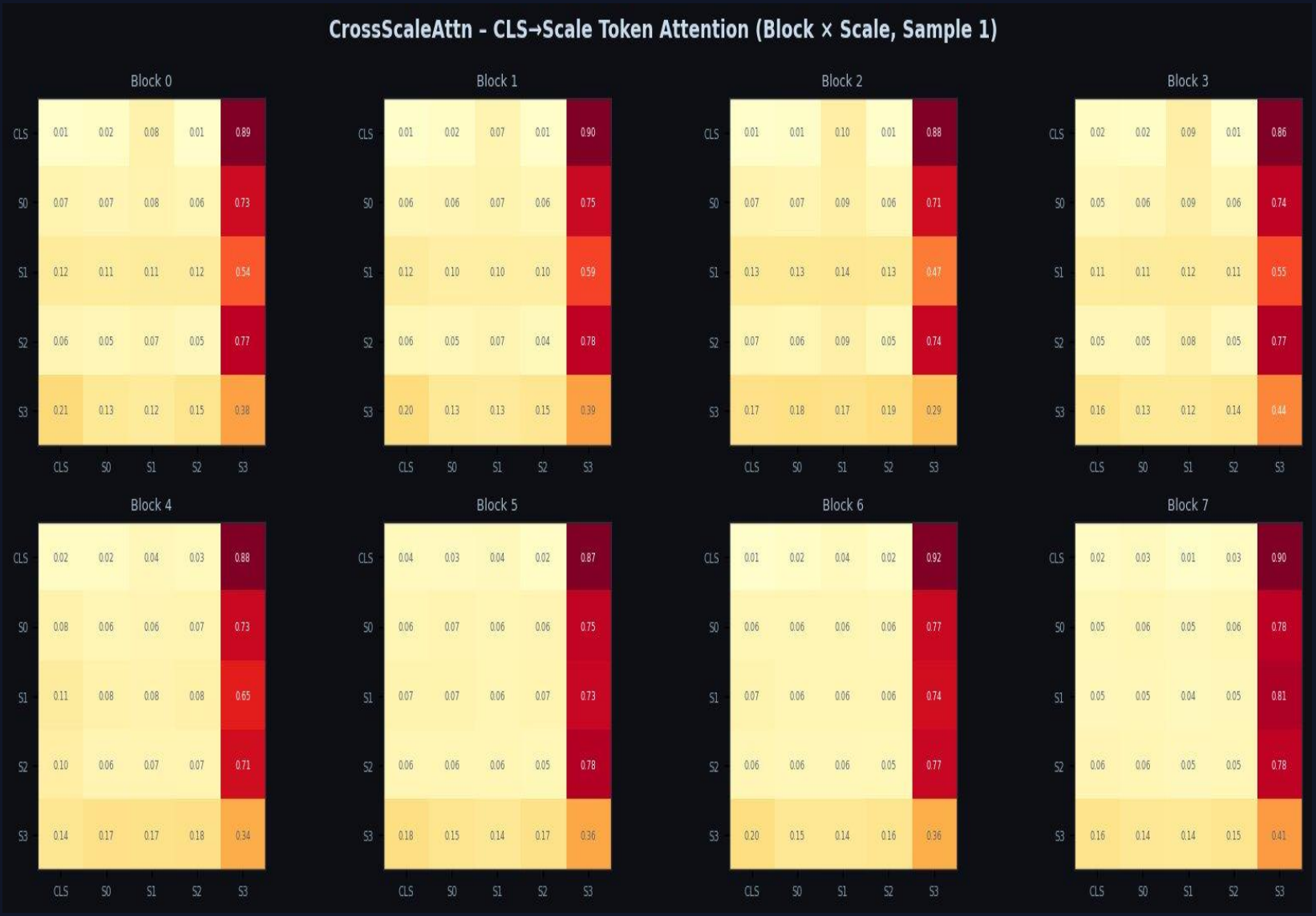
InterBlockAttn

- Single global CLS token + 8 block tokens
- 9 tokens total — negligible cost
- 2 transformer layers for global reasoning
- Output: 1 global representation vector
- Learns which blocks drive the prediction



CrossScaleAttn — Multi-Resolution Fusion

Each spatial block independently learns which CNN scale to trust most



Key Observations

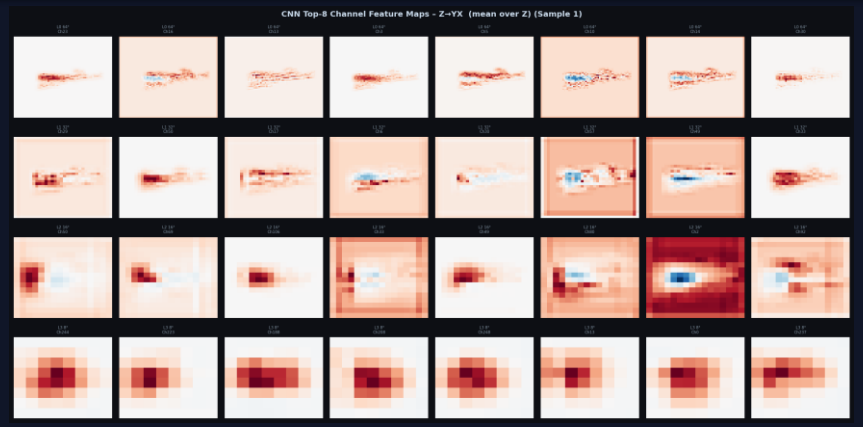
- S3 (8^3 native resolution) consistently dominates attention weight across all 8 blocks
- CLS row shows highest attention to S3 (~0.88–0.92) — coarse scale is most informative
- S1 (32^3) receives least weight, suggesting intermediate resolution adds least value
- Weight pattern is consistent across blocks, indicating stable learned scale preference
- S3 dominance aligns with physics: coarse shower topology drives momentum more than fine detail

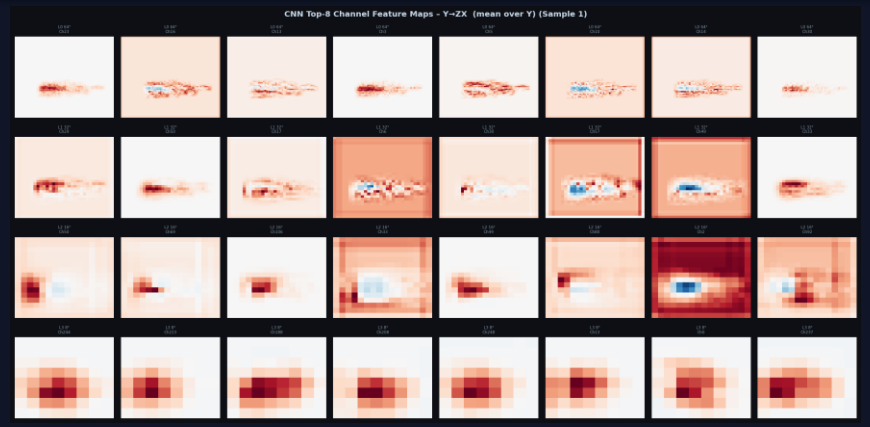
Attention Matrix Structure

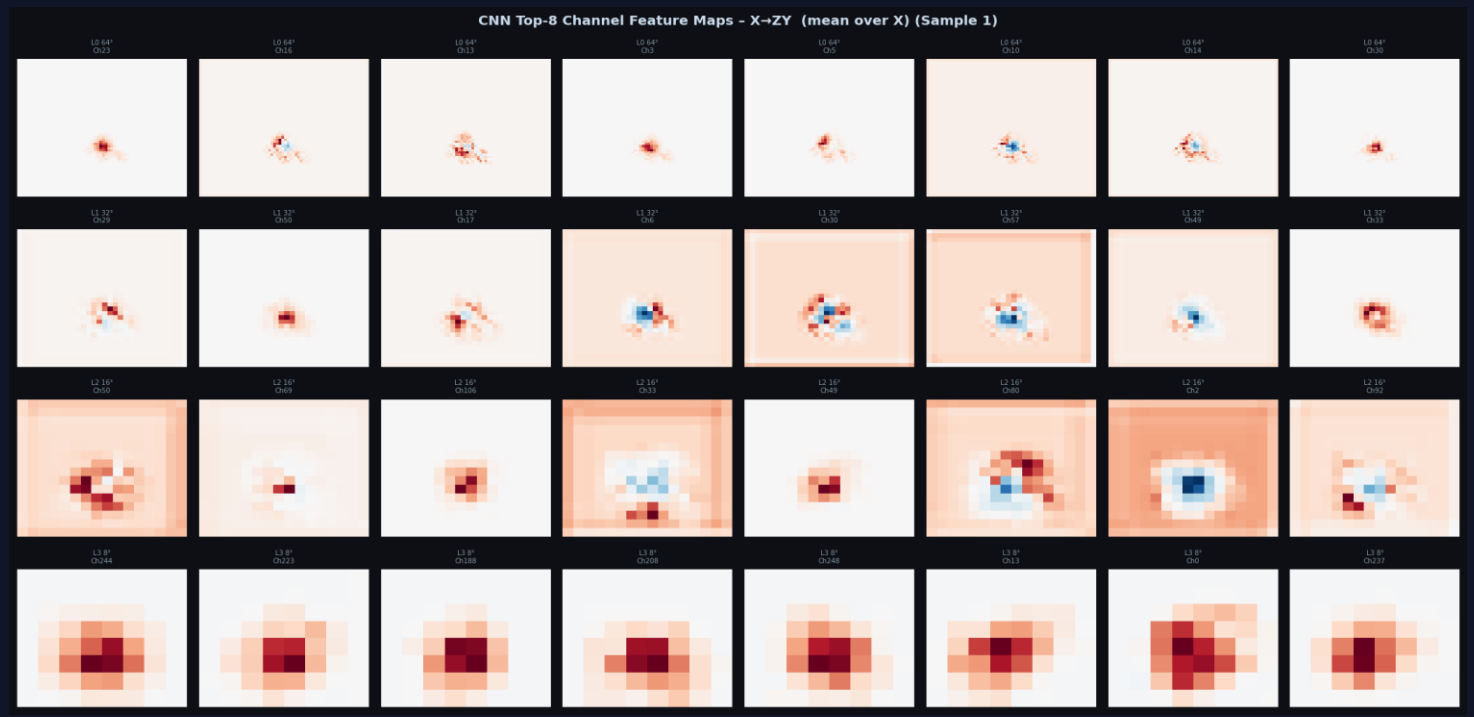
Row = query token
Column = key token
Value = attention weight
(head-averaged softmax)

CNN Feature Map Analysis

Top-8 channel activations across L0–L3, projected along Z axis







Observations

- L0 (64^3): fine-grained elongated shower core clearly resolved — RdBu diverging scale shows gradient structure
- L1–L2 (32^3–16^3): intermediate structure, increasing abstraction, shower boundaries visible
- L3 (8^3): all channels show symmetric blob — coarse energy centroid captured
- Blue regions indicate inhibitory responses (boundary detection / contrast enhancement)
- Channel specialisation increases with depth — more diverse filter responses at L3

Regression Performance

RMSE and MAPE across the full 1–20 GeV momentum range

MAE = 1.085 GeV

Global mean absolute error

RMSE = 1.843 GeV

Global root mean square error

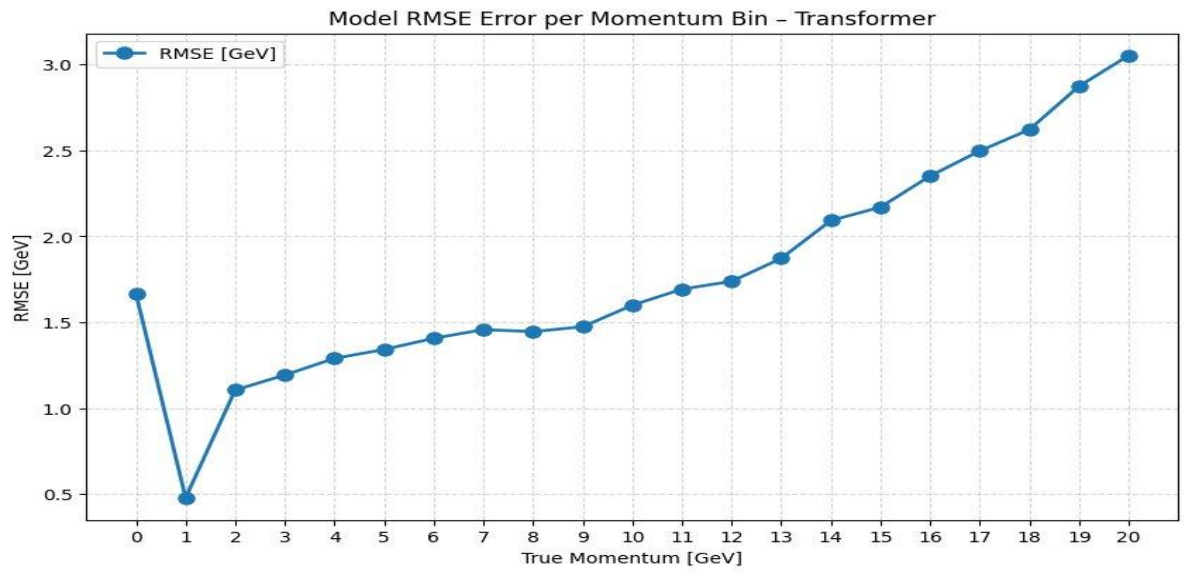
~10% MAPE

At high momentum (10–20 GeV)

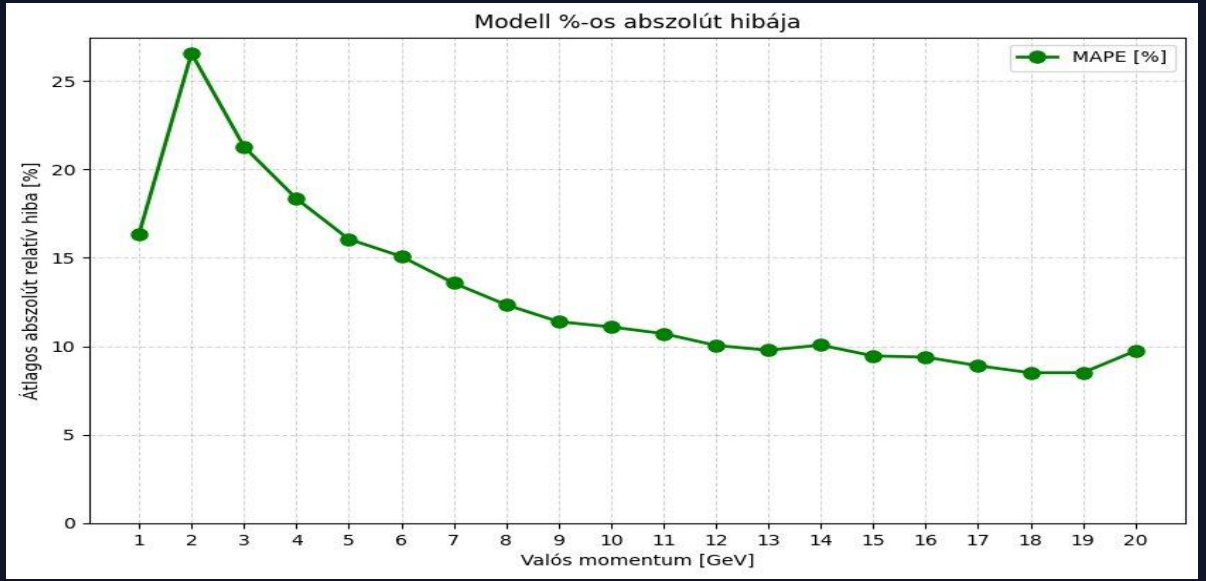
2.16 M

Trainable parameters

RMSE per Momentum Bin



MAPE per Momentum Bin



Key Insight:

RMSE grows roughly linearly with true momentum — consistent with calorimeter stochastic resolution ($\sigma/E \propto 1/\sqrt{E}$). MAPE converges to ~10% above 8 GeV, indicating stable relative performance at high energies. The elevated error at 1–2 GeV likely reflects shower containment effects at low energies.

Regression Performance

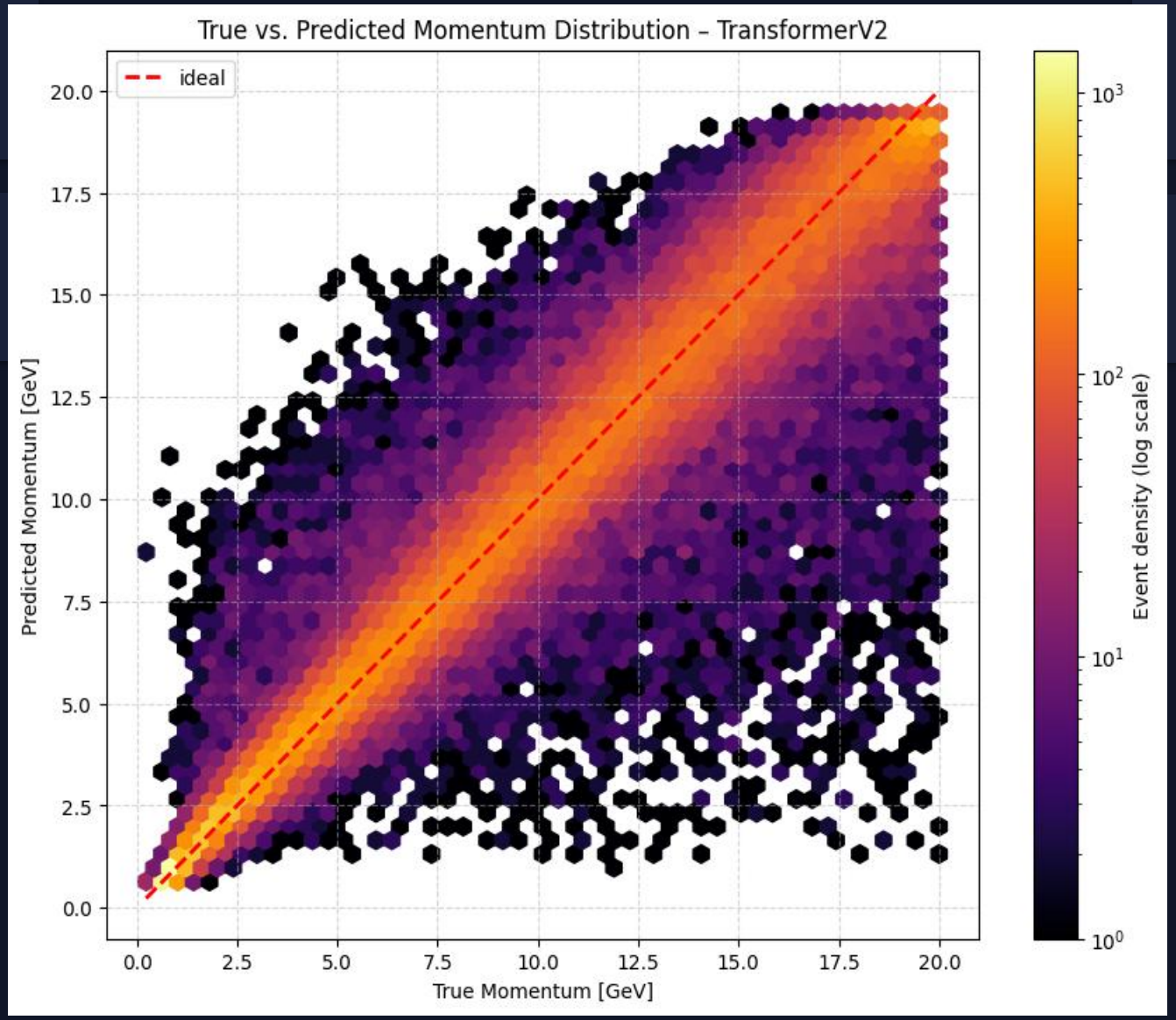
RMSE and MAPE across the full 1–20 GeV momentum range

MAE = 1.085 GeV

Global mean absolute error

RMSE = 1.843 GeV

Global root mean square error



2.16 M

Trainable parameters

~10% MAPE

At high momentum (10–20 GeV)

Conclusions & Outlook

Hierarchical CNN–Transformer proves effective for 3D shower momentum regression

Key Conclusions

- 3-stage hierarchical attention reduces $O(N^2)$ complexity by ×7.8 vs flat transformer
- Multi-scale CNN features + CLS token aggregation captures local and global shower structure
- PCA of final CLS tokens shows a clean 1D momentum manifold — model internalised the physics
- CrossScaleAttn consistently prioritises the coarsest scale (S3), matching physical intuition
- MAE ~1.1 GeV and MAPE ~10% above 8 GeV — competitive for a 2M parameter model

Future Directions

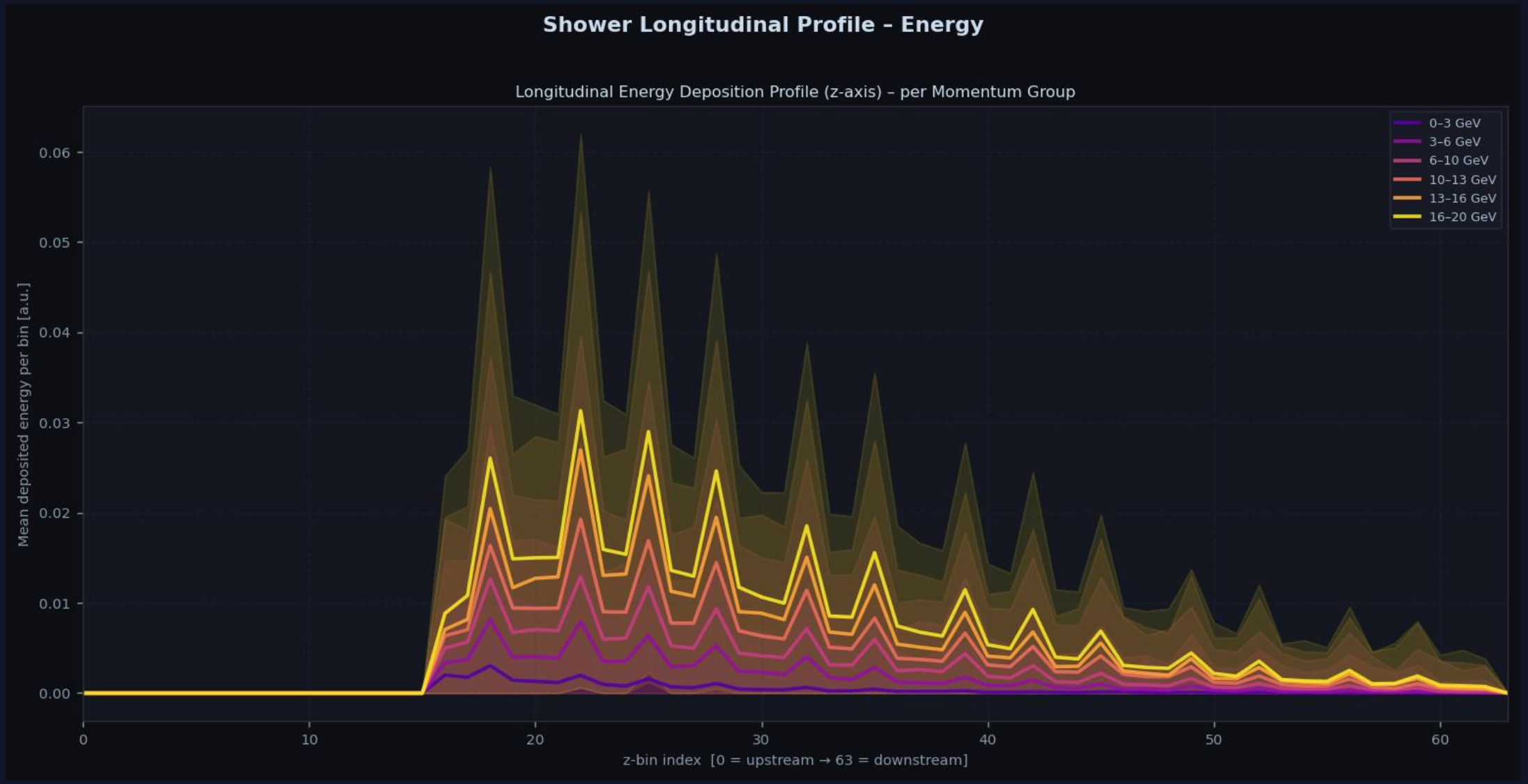
- Extend to multi-particle events and mixed hadron species (pion)
- Compare against graph neural networks (GNN) on the same dataset
- Investigate attention rollout for shower core localisation
- Scale to full calorimeter geometry (non-cubic, irregular segmentation)
- Explore uncertainty quantification for physics analyses

Architecture Pipeline

64^3 Energy Grid → CNN (4 scales) → IntraSpatialAttn → CrossScaleAttn → InterBlockAttn → Regression Head

Longitudinal Energy Deposition Profile

Hadronic showers initiate at z≈18 and exhibit characteristic fluctuating profiles. Higher momentum groups deposit more energy but show similar longitudinal shape, consistent with nuclear cascade dynamics.

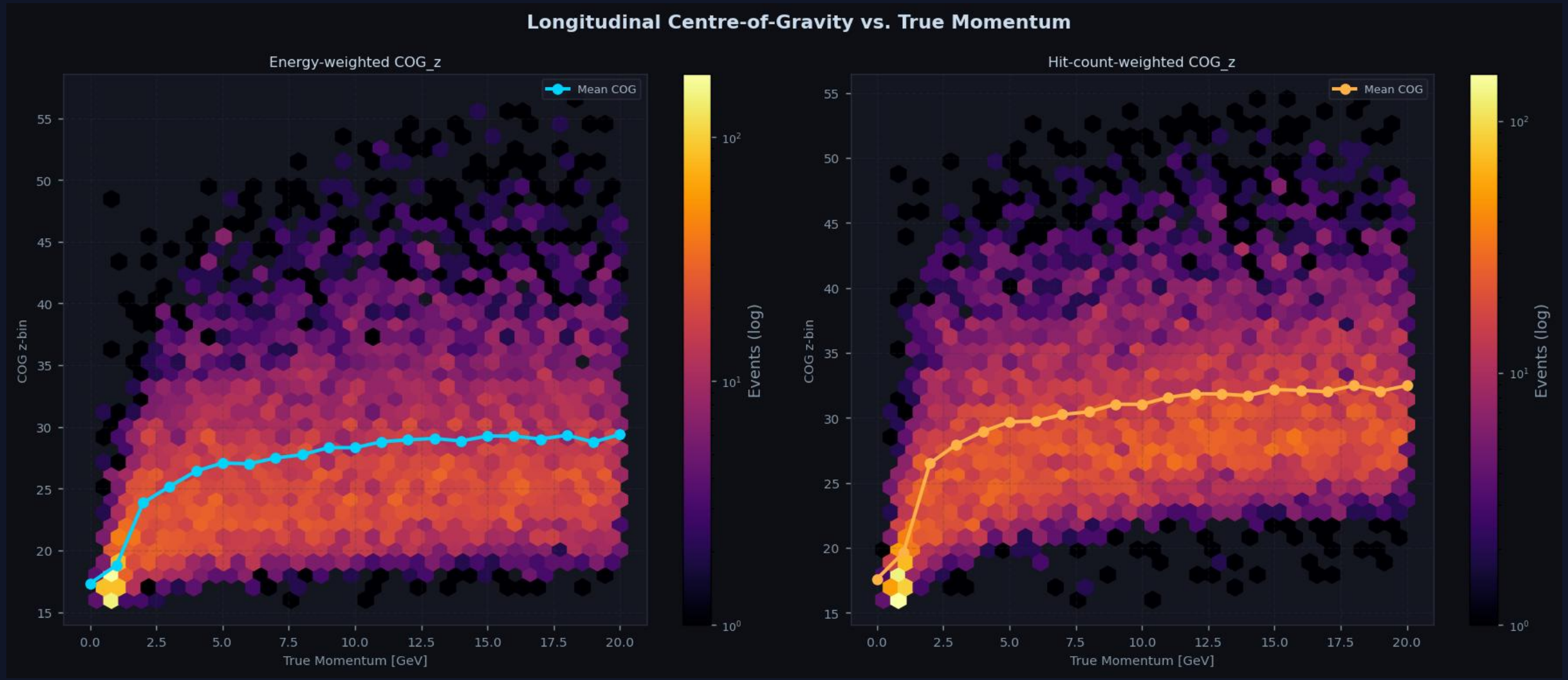


Shower Containment

90% of shower energy is contained within z≈41–44 across all momentum groups. Higher-energy showers extend slightly deeper, but the difference is small (~3 bins), confirming adequate longitudinal detector coverage.

Longitudinal Centre-of-Gravity

Energy-weighted COG saturates at z≈29, while hit-count-weighted COG sits ~3 bins deeper, revealing a low-energy secondary particle tail in the shower's downstream region. Both metrics plateau above ~5 GeV.



Hit Density Heatmap

Vertical bands reflect the sandwich calorimeter's absorber-active layer structure. The shower maximum (dashed line) shifts from z≈18 at low momenta to z≈26 above 5 GeV, then remains stable — validating the longitudinal shower model.