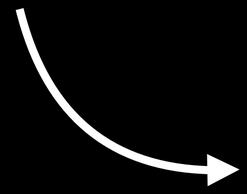
Towards foundation-style models for energy-frontier heterogeneous neutrino detectors via self-supervised pre-training

Fabio Cufino, Dr. Saúl Alonso-Monsalve, Dr. Umut Kose, Dr. Anna Mascellani, Prof. André Rubbia

Neutrino Physics and Machine Learning Conference, Irvine
15 June 2026



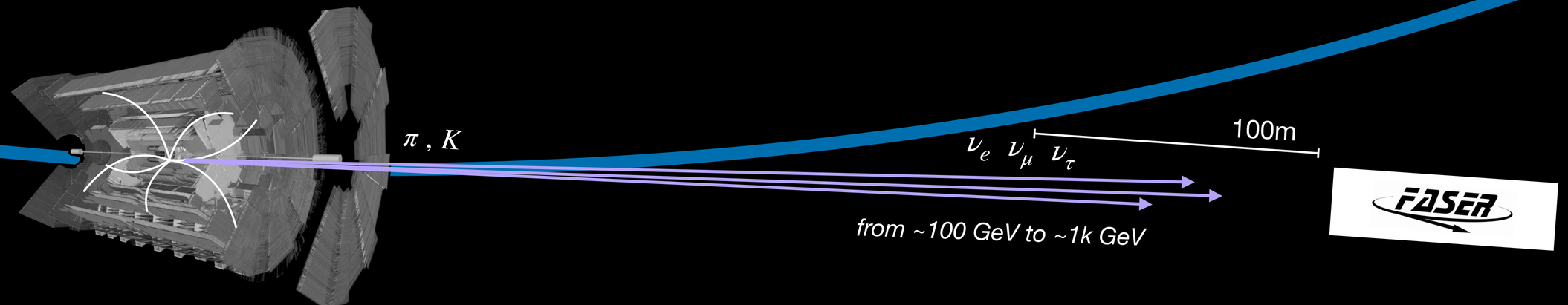
arXiv:2604.07037

ForwArd Search ExpeRiment

High Energy Neutrino Measurements at FASER

- **CERN LHC :** Production of unstable hadrons in forward direction of IP → decay → collimated neutrino beam → *FASER* detector.
  - **LHC Run 4**: Higher luminosity will pose challenges for the existing FASERν detector.
- **Proposed Solution: *FASERCal*** *(ETH Zurich)*
  - Fully electronic 3D Precision Calorimeter (inspired by SuperFGD detector at T2K) combined with calorimeters, and muon detector.

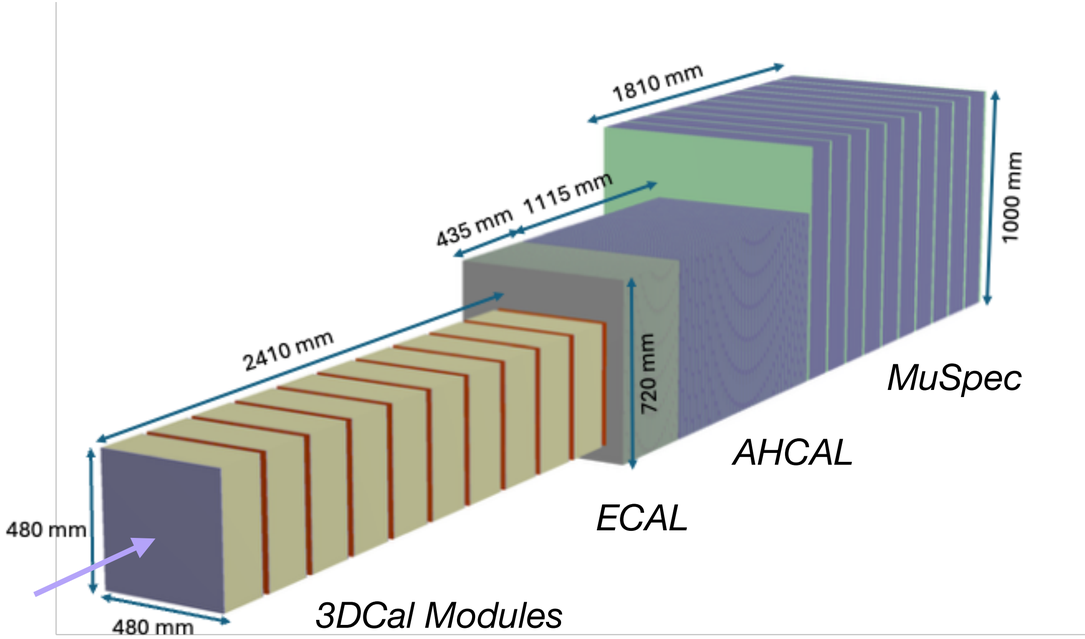
- Detector Goals:
  - Identify $\nu_e CC,\ \nu_\mu CC,\ \nu_\tau CC,$ and NC interactions.
  - Measure neutrino differential cross-sections and fluxes.



The detector

A heterogeneous detector for high-energy neutrinos

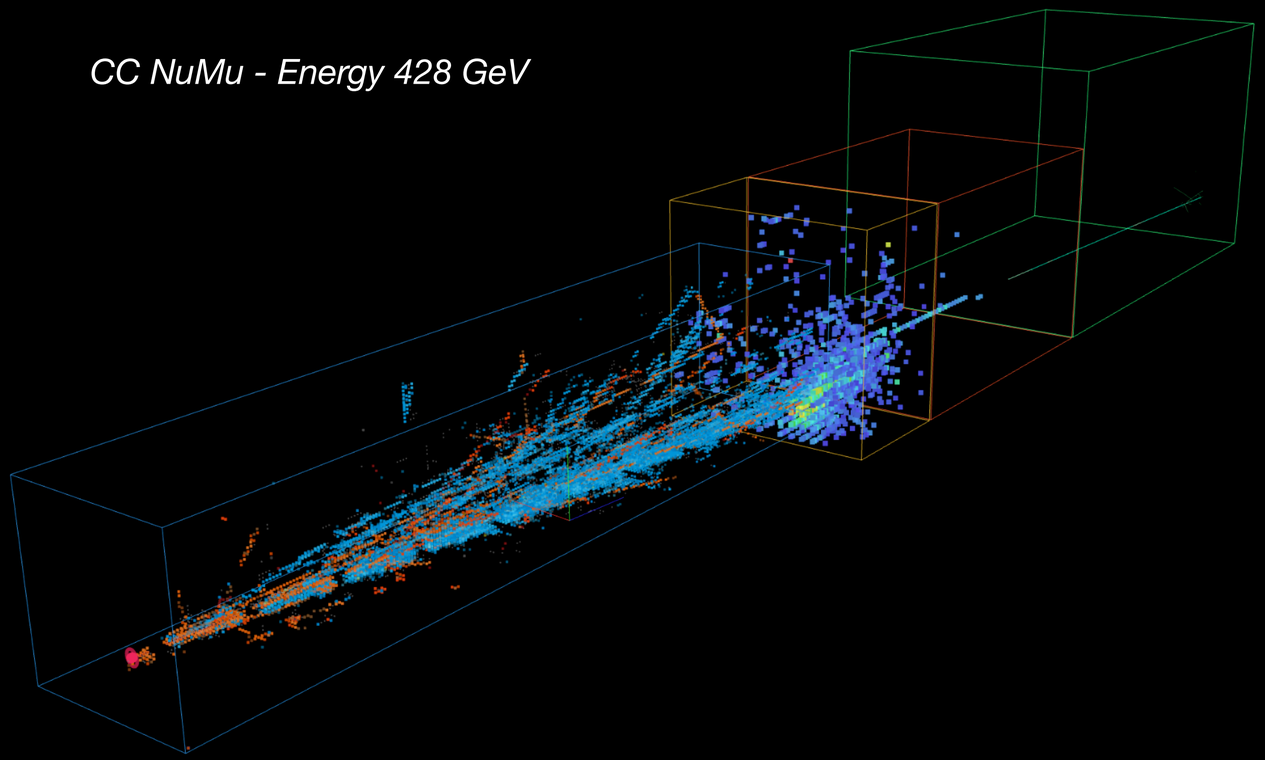
- **3DCal —** Fully-electronic 3D scintillator calorimeter (SuperFGD-inspired); 10 longitudinal modules; 10 module of 48 × 48 × 20 voxels of ~1 cm^3 → **>460,000 voxels.**
    - Provides sparse 3D voxel hits with charge.
- **ECAL** — electromagnetic calorimeter. A dense representation of the electromagnetic shower: a **5 × 5 energy matrix**.
- **AHCAL** — hadronic calorimeter. A second sparse calorimetric volume, at **coarser granularity** than the 3DCal → **(18x18x40)**
- **Muon spectrometer** — measures penetrating muons. Up to 10 hit-measuring planes per track.



The Challenge

Forward LHC neutrinos produce compact but complex final states

CC NuMu - Energy 428 GeV



Energy-frontier neutrino interactions are not standard detector images, conventional reconstruction breaks down.

- TeV interactions → collimated topologies, high multiplicities, EM and hadronic activity strongly overlapping
- Only a fraction of the >460k cubes fire per event, yet *local configurations are deeply ambiguous*
- Hard for a Neural Network to capture: convolutional receptive fields are too local, and dense self-attention over hundreds of thousands of voxels is intractable — ***the global context the physics demands is the part that's most expensive to model***.

The idea

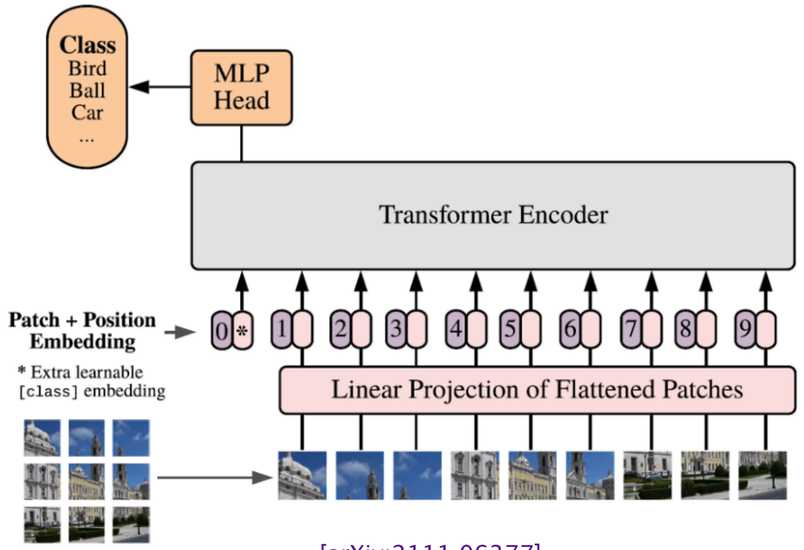
From task-specific reconstruction to representation learning

Complex events require a model to connect information across the whole detector and recognize event structure in a **learned embedding space.**

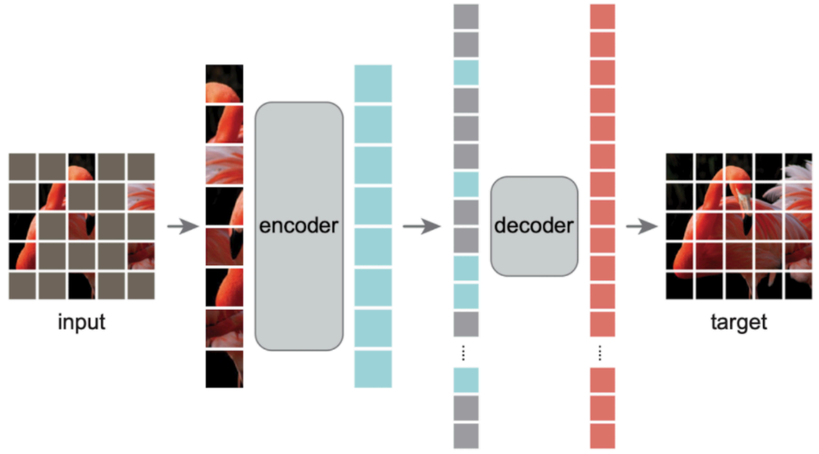
- Task-specific classification labels or regression targets only describe the final answer, not the full internal event topology.
- The model must disentangle the *local* event topology and learn *global* correlations across heterogeneous sub-detectors.

- **Core ingredients**

1. **A sparse heterogeneous encoder** — sparse-convolutional patch embeddings; processed with self-attention à la **ViT**, module-aware self-attention.
2. **Perceiver-IO** fusion across calorimetric and tracking streams.
3. **Multimodal self-supervised pre-training** — *Masked-prediction pretraining* — **MAE;** *physics-aware relational voxel-level targets* (ghost ID, interaction hierarchy, particle category).
4. **Reuse & transfer** — **Learn reusable representations once,** then adapt across downstream tasks**.**



[arXiv:2111.06377]



[arXiv:2010.11929v2]

Methodology

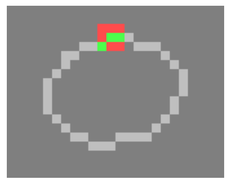
1. **Architecture:** From heterogeneous detector data to a compact latent representation
2. **Two-stage training procedure:** Self-supervised pretrainign followed by multi-task fine-tuning

Architecture

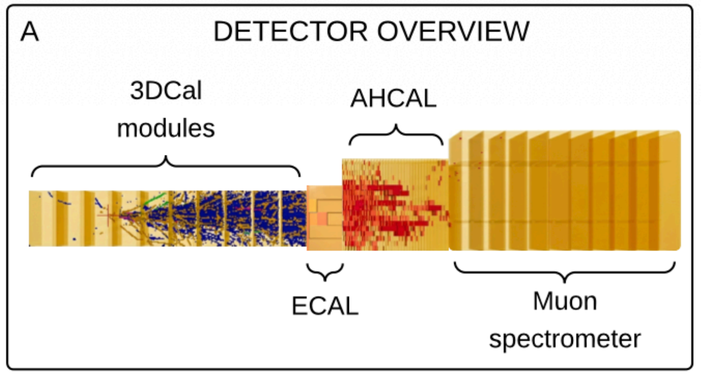
From heterogeneous detector to latents

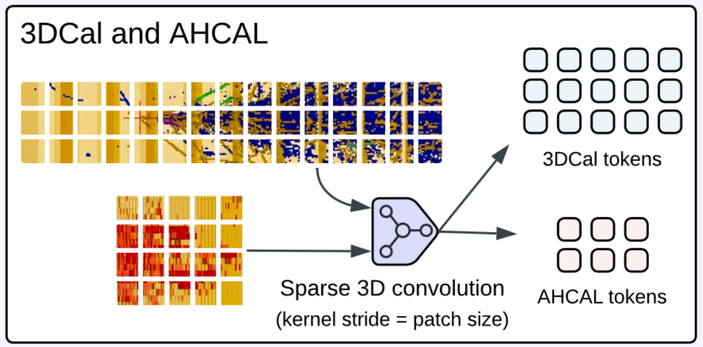
- **Input representation — tokenising each stream**
  - *3DCal & AHCAL* → sparse 3D convolution (SSC* SpConv, kernel & stride = patch size), acting only on occupied voxels → *patch embedder*
    - *3DCal*: 12 × 12 × 10-voxel patches → 4 × 4 × 20 grid, up to 320 tokens (only occupied kept)
    - *AHCAL*: 6 × 6 × 5-voxel patches → 3 × 3 × 8 grid, up to 72 tokens
  - *ECAL*: 5 × 5 energy matrix → fully-connected layer → one ECAL token
  - *Muon spectrometer:* hits → Kalman-filter track fit → attention block → one muon spectrometer token

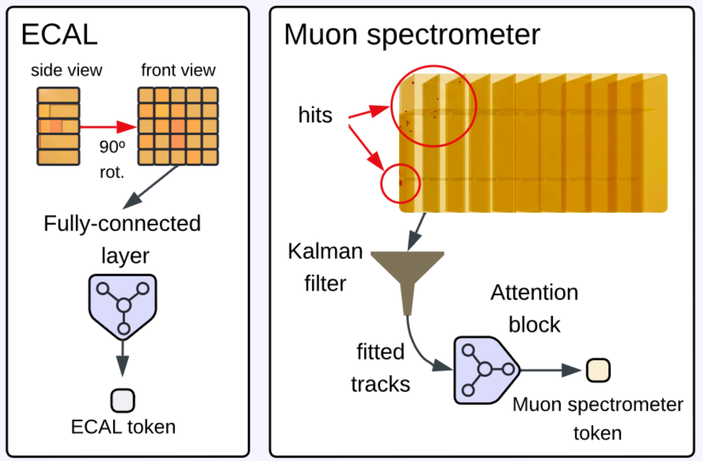
- **Submanifold Sparse Convolution Network:*
  - Convolves only on occupied voxels and keeps the active set fixed — so it never "fills in" empty space the way a dense CNN does.
  - Efficiency SCNN: (16 times faster than a CNN on a GPU) [Link]



[arXiv:1706.01307]







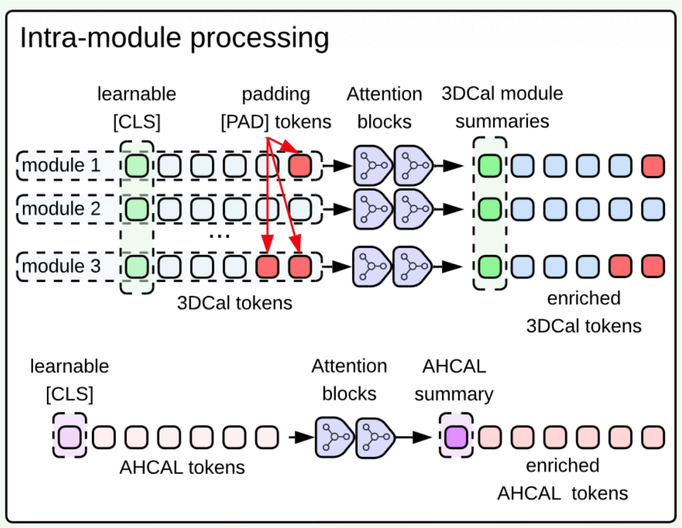
Architecture

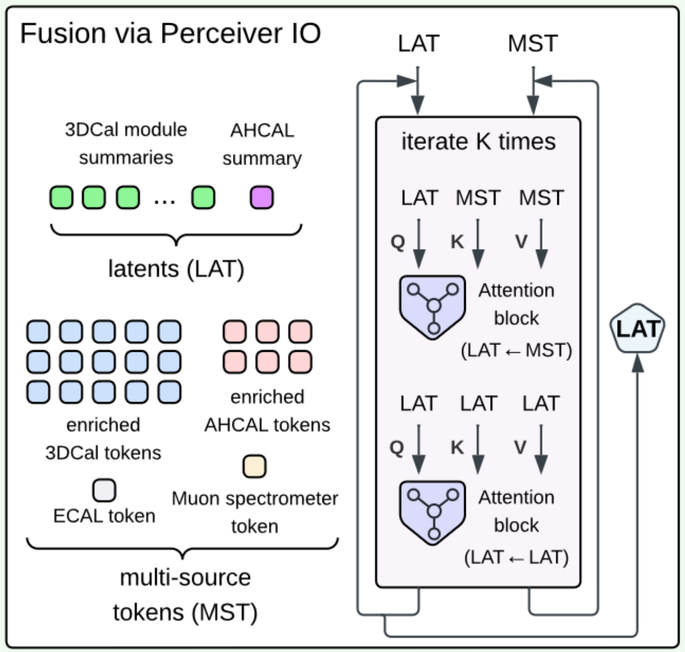
From heterogeneous detector to latents

Hierarchical encoder — two stages

- *Intra-module processing:*
  - 3DCal tokens are grouped by detector module.
  - Each module is processed with module-level self-attention using learned positional embeddings.
  - A learned module-level **[CLS] token** summarizes the information in each module.
  - AHCAL runs a parallel stack with its own **[CLS]**.
    - → module & AHCAL summaries + enriched tokens.

- *Fusion via Perceiver-IO:*
  - The event is represented by a small set of latent tokens initialized from:
    - 3DCal module summaries, AHCAL summary
  - These latents **cross-attend** to all detector streams:
    - enriched 3DCal tokens, enriched AHCAL tokens, ECAL token, muon spectrometer token
  - Latents then **self-attend**, allowing information to propagate globally across the full event.





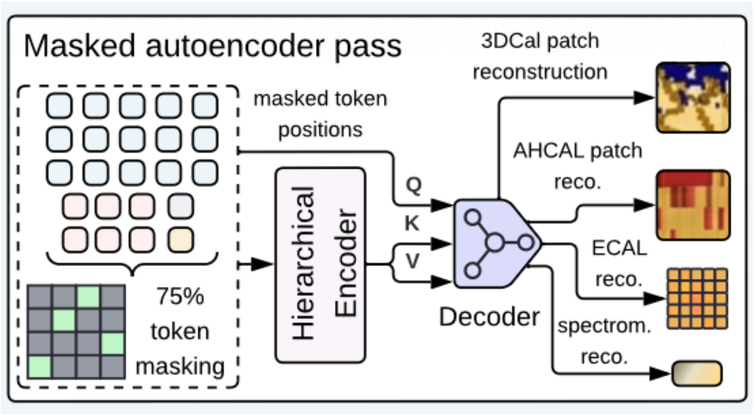
Compact Event Representation

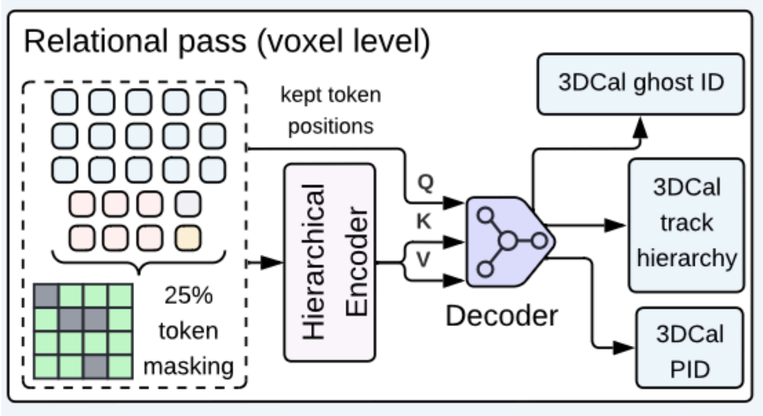
Stage 1: Pre-training

Two complementary self-supervised objectives

Phase 1 — Masked reconstruction (MAE), 400 epochs

- 75% of occupied calorimeter patches masked, encoder only sees the remaining visible patches
- Lightweight decoder predicts voxel occupancy & charge in missing regions
- Forces the encoder to infer non-local spatial correlations
- Captures global shower geometry & cross-detector context
  - → defines the *MAE encoder*





Phase 2 — Relational voxel-level pass (+100 epochs, mask ratio 0.25)

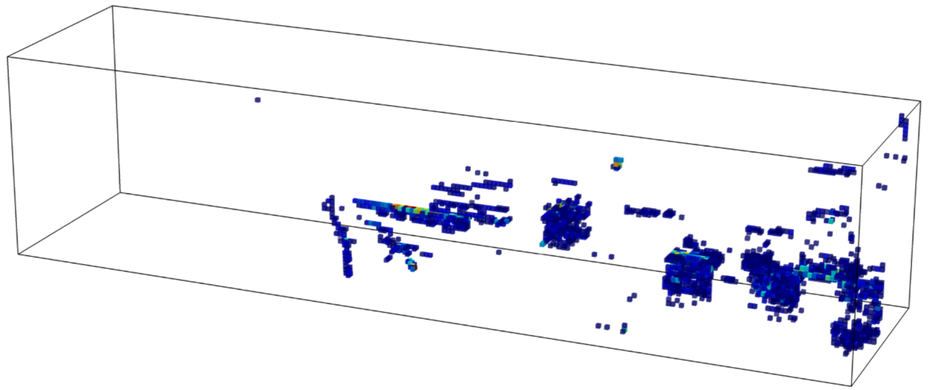
On kept patches the encoder predicts three physics-aware targets from simulation truth:

- *Ghost ID* — reconstructed voxels with no matched true particle (binary)
- *Hierarchy* — background / primary / secondary (soft per-voxel labels)
- *Particle category* — EM / muonic / hadronic (soft; can co-exist in a voxel)
  - → defines the *MAE+Rel encoder*

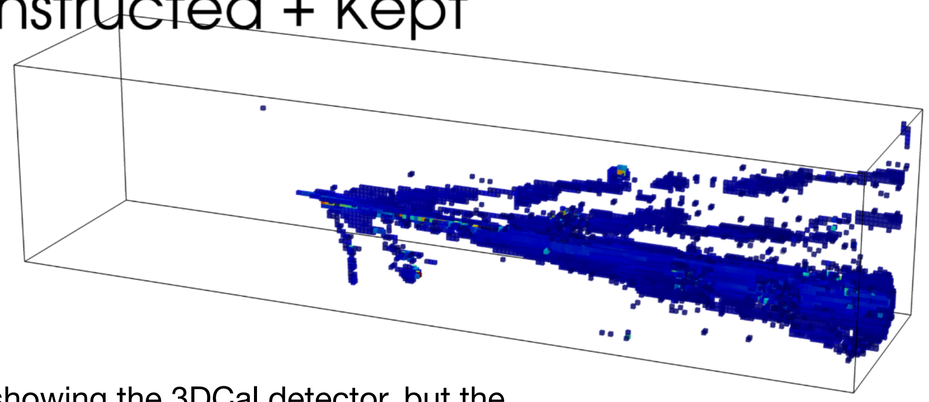
Stage 1: Pre-training

Phase 1: Masked Reconstruction Examples

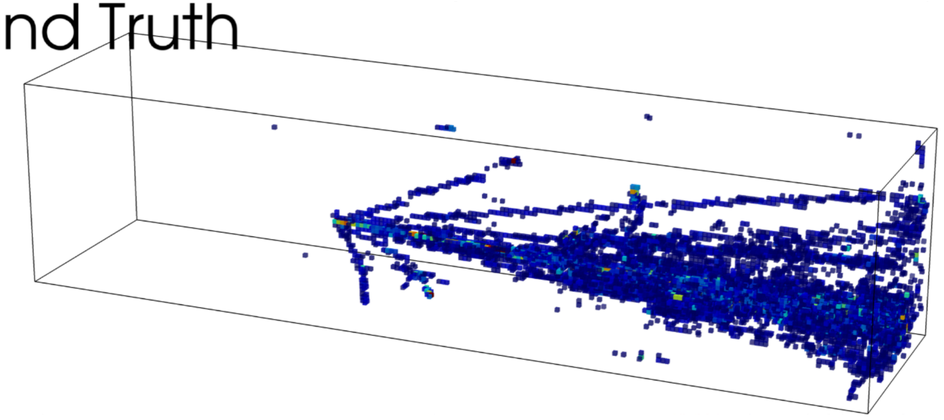
Kept



Reconstructed + Kept



Ground Truth



- The reconstruction is not exact at the voxel level, especially in dense shower cores and fine secondary structures.
- But the model learns the physically important large-scale structure:
  - shower direction;
  - longitudinal development;
  - compact high-energy regions;
  - extended secondary tracks;
  - downstream continuation of the event.

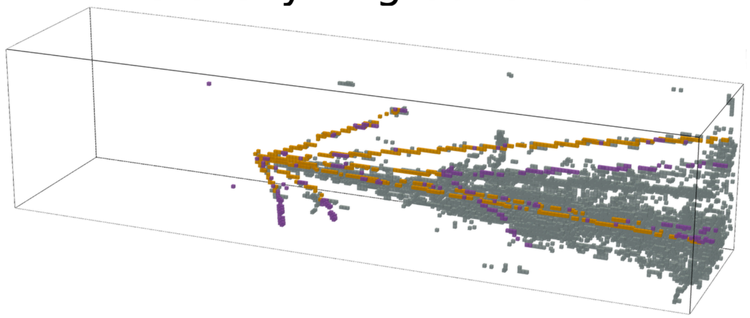
Achtung: Only showing the 3DCal detector, but the model learns to correlate all subdetectors

Stage 1: Pre-training

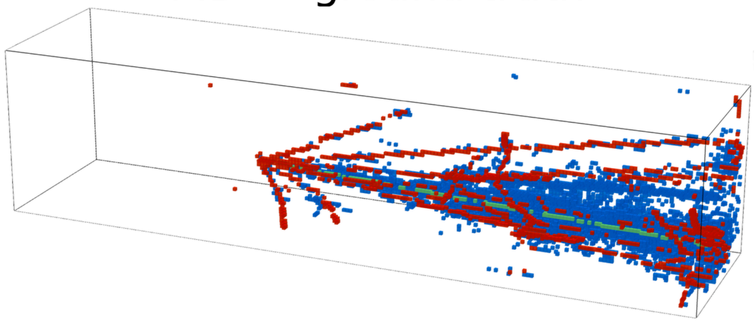
Phase 2: Relational voxel-level pass

*The labels are soft because dense shower regions can mix contributions from several true particles in the same reconstructed voxel

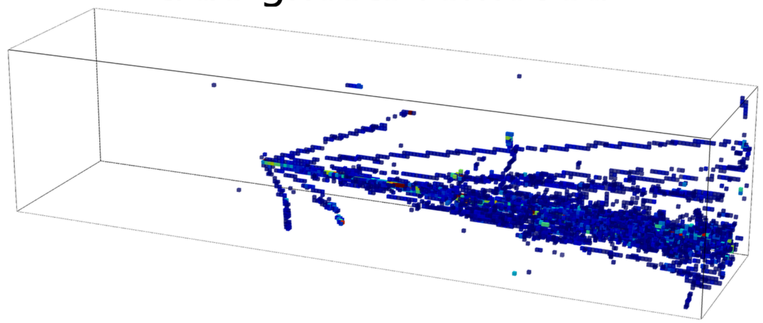
hierarchy — ground truth



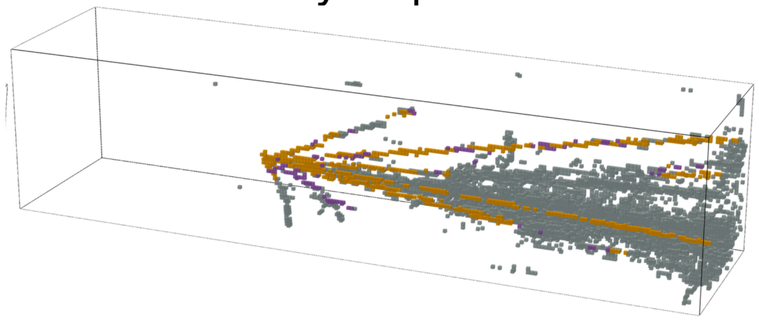
PID — ground truth



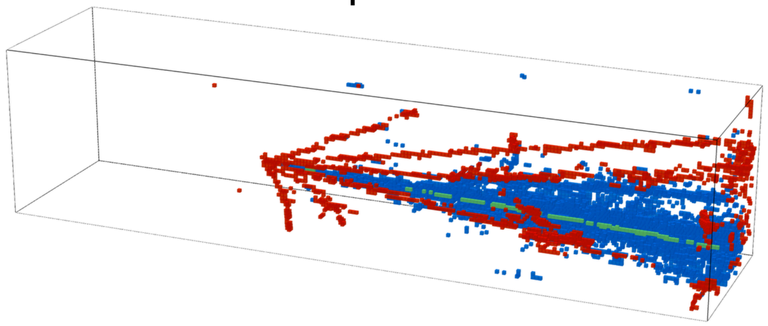
true ghosts removed



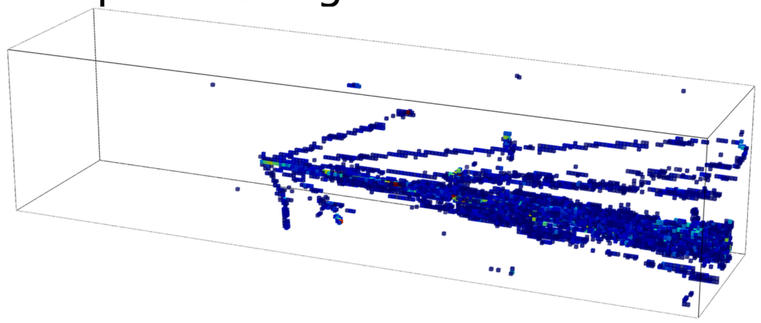
hierarchy — prediction



PID — prediction



predicted ghosts removed



background primary secondary

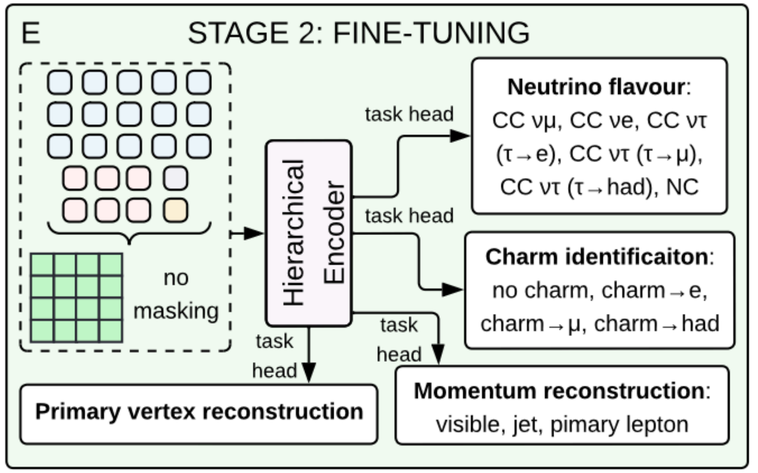
EM muon hadronic

Stage 2: Fine-Tuning

Joint multi-task fine-tuning

Stage 2 — joint multi-task fine-tuning (no masking):

- *Decoder discarded; shared encoder kept*; lightweight task heads read the latent. Four tasks at once:
  - Neutrino flavour (6-classes): CC νμ, CC νe, CC ντ→e, CC ντ→μ, CC ντ→had, NC
  - Charm ID (4-classes): no charm, charm→e, charm→μ, charm→had
  - Momentum reconstruction: visible, jet, primary lepton momenta
  - Primary-vertex reconstruction
- Multi-task because the observables are physically coupled (one event interpretation).



Evaluation setup

Same architecture, three encoder initializations

Everything is held fixed — architecture, task heads, head init, fine-tuning data and schedule — **except the encoder's starting weights**.

- **Scratch:** standard random initialization; baseline model.
- **MAE**: encoder initialized from masked-reconstruction pre-training; learns from energy reconstruction.
- **MAE+Rel:** encoder initialized from the full pre-training objective; combines masked energy reconstruction with the relational voxel-level pass.

Any difference downstream is therefore attributable to the *representation*, not to the model or the training budget.

- Because flavour, charm, energy flow and vertex are physically coupled (one event interpretation), joint multi-task fine-tuning tests whether the representation is a genuine *shared basis* — not a per-task trick

Metrics: classification — one-vs-rest AUC + confusion matrix; regression — residual median / IQR (vertex in mm, fractional σ_MAD for momenta).

Results

Classification and Regression

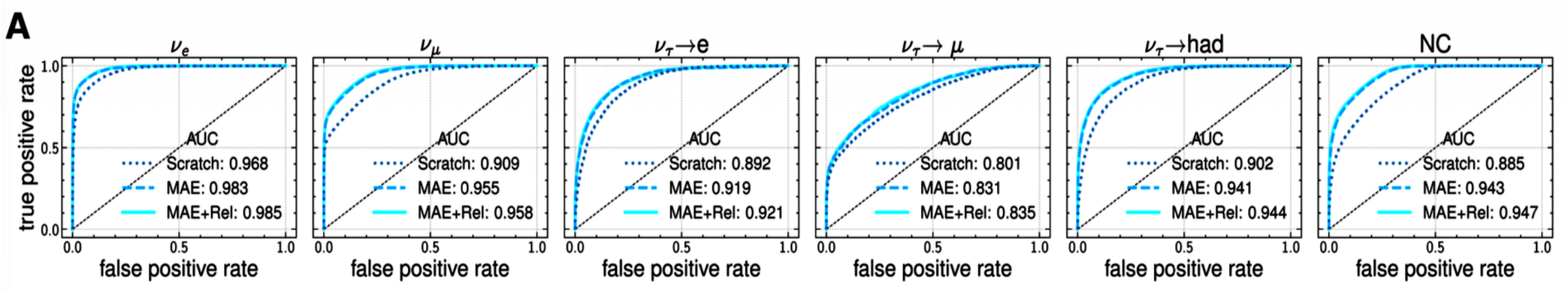
Flavor classification

Flavor identification: pre-training improves every channel

- One-vs-rest AUC (Scratch → MAE → MAE+Rel):
  - **νe CC:** 0.968 → 0.983 → 0.985
  - **νμ CC:** 0.909 → 0.955 → 0.958
  - **NC**: 0.885 → 0.943 → 0.947
- Cleaner confusion matrix: diagonal performance improves for— νe CC 0.71 → 0.85, NC 0.50 → 0.71.

The largest gains land in the hardest channels (CC NuTau) — dense overlap, secondary activity make tau channels the most ambiguous.

- One-vs-rest AUC (Scratch → MAE+Rel):
  - **ντ→had:** AUC 0.902 → 0.944 (+ 0.042)
  - **ντ→e:** AUC 0.892 → 0.921 (+ 0.029)
  - **ντ→μ:** AUC 0.801 → 0.835 (+ 0.034)



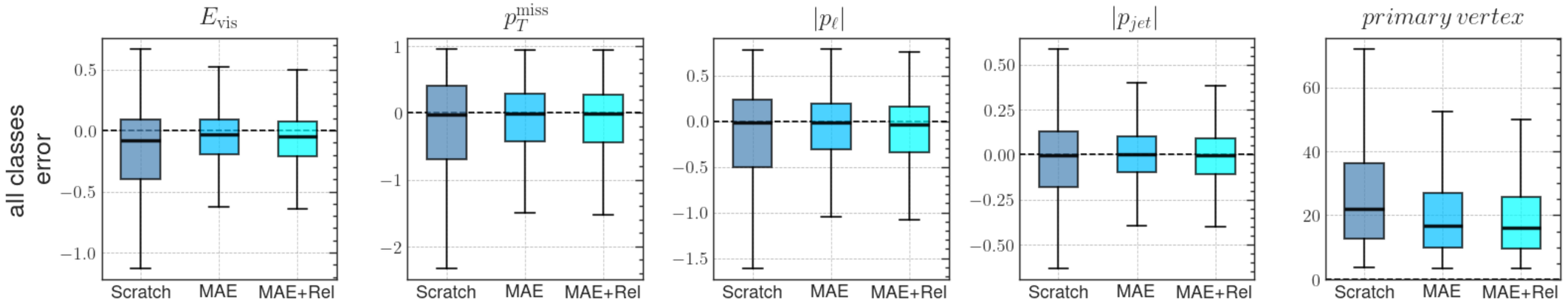
Regression results

Shared representation improves

- **Primary-vertex reconstruction** (clearest, most uniform gain):
  - The 3D vertex displacement dPV = ‖x_true – x_recoll decreases for MAE, but more in MAE+Rel; interquartile ranges reduced significantly.
  - Improvement is visible in both the common νe/νμ CC channels and in the difficult tau and NC samples → reflects a better shared latent, not one easy topology.

- **Kinematic targets:**
  - *Evis and |pjet|:* medians move closer to zero; spreads become smaller; strongest gains in charged-current channels and ντ → had
  - *Missing pT and |pℓ|:* improvements are visible but less uniform; tau categories remain broad because the selected samples are intrinsically difficult.

***The encoder is not only improving classification. It improves the shared representation used for both discrete and continuous physics targets.*

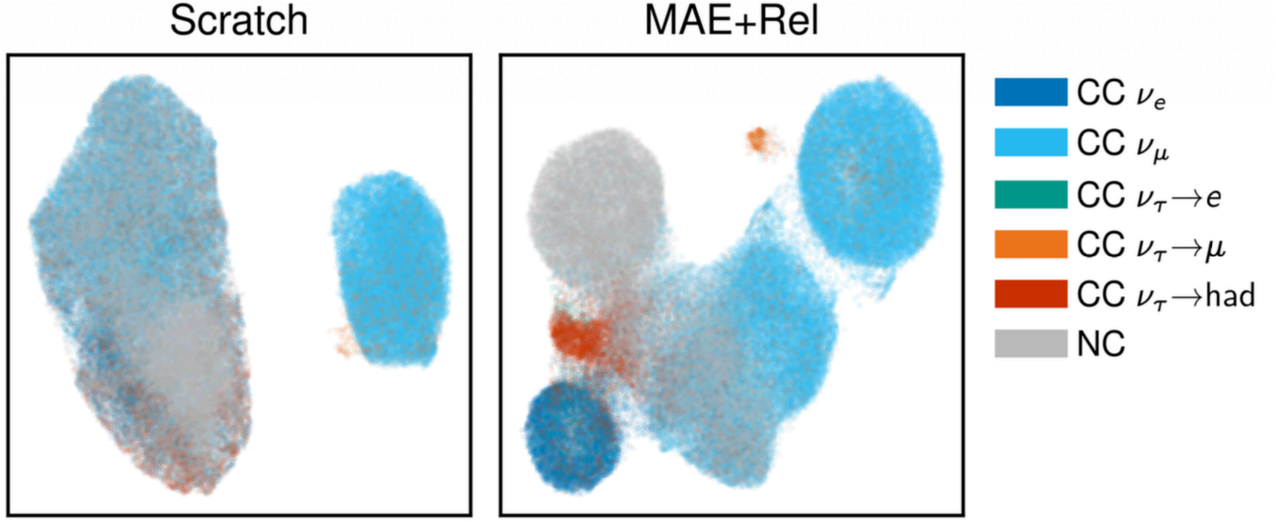


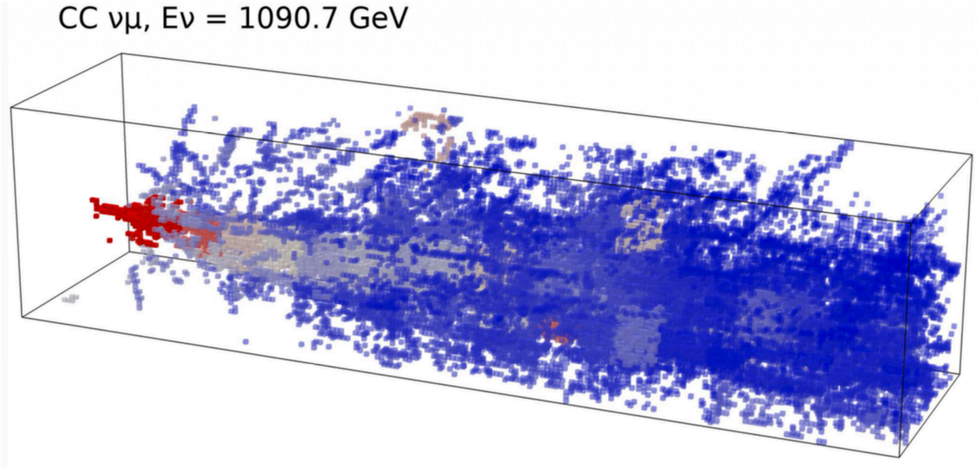
Interpretability: structure of the latent space

Pre-training yields a more structured latent space

Latent-space structure

- UMAP projections show that MAE+Rel produces a cleaner low dimensional geometry than Scratch
- Flavour harder groups are better separated.





Detector-subsystem ablation

- Saliency concentrates near the interaction region & the main downstream shower;
- It is not spread diffusely over all occupied voxels. This suggests that the model uses physically meaningful event skeletons, not only global occupancy.

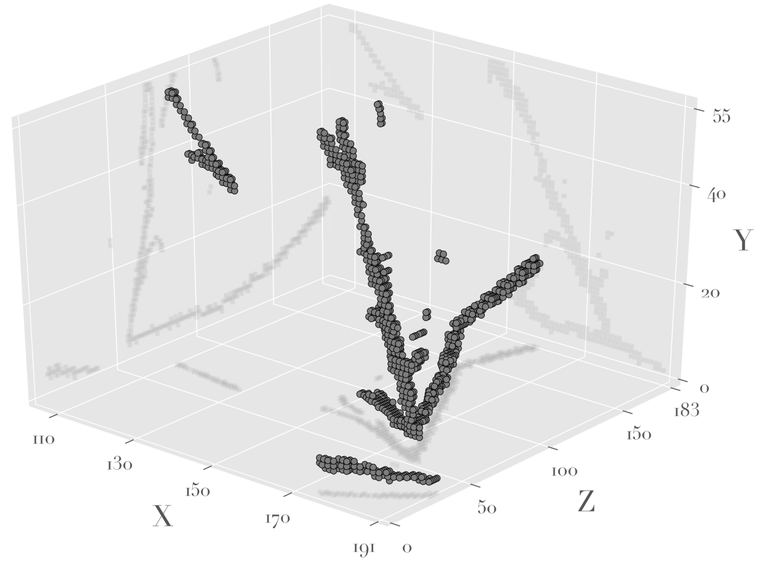
Transfer learning

Does the encoder generalize beyond FASERCal?

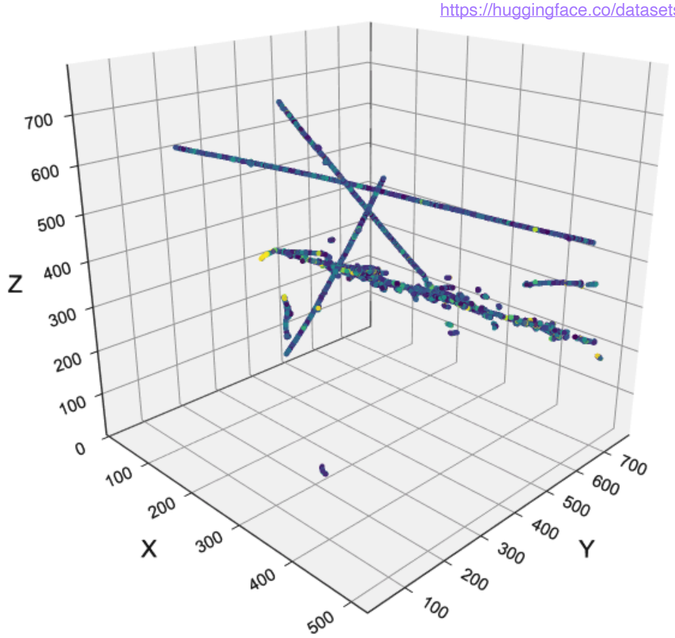
In-domain gains are encouraging — but the *foundation-style* claim holds if the *representation transfers to other detectors, tasks and energy scales.*

- We test two target domains at increasing distance from FASERCal.
  - *SuperFGD*: Plastic scintillator — close technology, GeV single particles
  - *PILArNet LArTPC* — different technology, energy regime and a different particle-ID benchmark.
- *The key question is whether the FASERCal pre-trained encoder has learned reusable priors.*
  - **What we reuse:** the transferable core of the source encoder — attention blocks, latent cross-/self-attention, normalisation layers, global query token.
  - **What we re-initialize:** detector-specific patch embeddings, positional encodings and the task heads.

https://journals.aps.org/prd/pdf/10.1103/PhysRevD.103.032005



https://huggingface.co/datasets/DeepLearnPhysics/PILArNet-M



Target 1 — *SuperFGD*

Fine-grained plastic scintillator

- *It is still a non-trivial shift:* From TeV-scale neutrino interactions to GeV-scale single-particle events.
- Per-class accuracy (confusion-matrix diagonal), Best published baseline → MAE+Rel:
  - Protons: 0.907 → 0.943
  - Charged pions: 0.643 → 0.609
  - Muons: 0.595 → 0.748
  - Electrons: 0.772 → 0.787
- **Beats the strongest published baseline** for protons, muons & electrons; narrows the gap for pions.
- *Takeaway:* our encoder retains useful priors: shower shape, track-like versus shower-like topology, energy deposition patterns

True class	Method	Pred. p	Pred. $\pi^{\pm}$	Pred. $\mu^{\pm}$	Pred. $e^{\pm}$
p	GBDT-Transformer [38]	0.907	0.067	0.007	0.019
	GBDT-RNN [38]	0.896	0.073	0.006	0.025
	GBDT-SIR-PF [38]	0.891	0.077	0.008	0.024
	Scratch (ours)	0.919	0.060	0.003	0.018
	MAE+Rel (ours)	**0.943**	0.040	0.006	0.011
$\pi^{\pm}$	GBDT-Transformer [38]	0.057	**0.643**	0.041	0.259
	GBDT-RNN [38]	0.080	0.623	0.036	0.261
	GBDT-SIR-PF [38]	0.080	0.606	0.042	0.272
	Scratch (ours)	0.062	0.532	0.267	0.139
	MAE+Rel (ours)	0.053	0.609	0.227	0.111
$\mu^{\pm}$	GBDT-Transformer [38]	0.071	0.190	0.595	0.144
	GBDT-RNN [38]	0.089	0.233	0.506	0.172
	GBDT-SIR-PF [38]	0.126	0.236	0.517	0.121
	Scratch (ours)	0.020	0.173	0.661	0.146
	MAE+Rel (ours)	0.019	0.079	**0.748**	0.155
$e^{\pm}$	GBDT-Transformer [38]	0.020	0.199	0.009	0.772
	GBDT-RNN [38]	0.027	0.200	0.007	0.766
	GBDT-SIR-PF [38]	0.017	0.237	0.006	0.740
	Scratch (ours)	0.019	0.179	0.091	0.712
	MAE+Rel (ours)	0.012	0.084	0.117	**0.787**

[38] S. Alonso-Monsalve et al: Artificial intelligence for improved fitting of trajectories of elementary particles in dense ma terials immersed in a magnetic field, Communications Physics 6, 119 (2023).

*The single-particle comparison is not strictly like-for-like because the reference uses a different 1024^3 setup, but the multi-particle benchmark is the clean apples-to-apples comparison.

Target 2 — PILArNet

Different detector technology, task and energy regime

- *PILArNet is a much stronger transfer test than SuperFGD:*
  - detector technology, spatial morphology, energy scale etc
- **Multi-particle benchmark**:
  - Accuracy: Best 0.9644 → MAE+Rel 0.9662
  - AUROC: Best 0.942 → MAE+Rel 0.951
- **Single-particle benchmark*:**
  - Accuracy: Best 0.9014 → MAE+Rel 0.9154
  - AUROC: Best 0.842 → MAE+Rel 0.891
- Encoder pre-trained on TeV neutrino interactions for FASERCal **can adapt** to PILArNet and match or exceed specialized target-trained baselines.

	Single-particle classification*		Multi-particle classification	
Method	Accuracy	AUROC	Accuracy	AUROC
Deterministic [40]	0.8656	0.753	0.9604	0.938
Naive Ensembles [40]	0.8844	0.827	0.9640	0.944
Bootstrap Ensembles [40]	0.9014	0.842	0.9644	0.942
MC Dropout [40]	0.8734	0.795	–	–
EDL-MLL [40]	0.8622	0.762	0.9604	0.935
EDL-BR [40]	0.8253	0.701	0.9223	0.900
EDL-Brier [40]	0.8751	0.748	0.9596	0.911
Scratch (ours)	0.8798	0.834	0.9333	0.922
MAE+Rel (ours)	**0.9154**	**0.891**	**0.9662**	**0.951**

[40] D. H. Koh, A. Mishra, and K. Terao, Deep neural network uncertainty quantification for LArTPC reconstruction, Journal of Instrumentation 18 (12), P12013.

Conclusions

Summary and Future Prospects

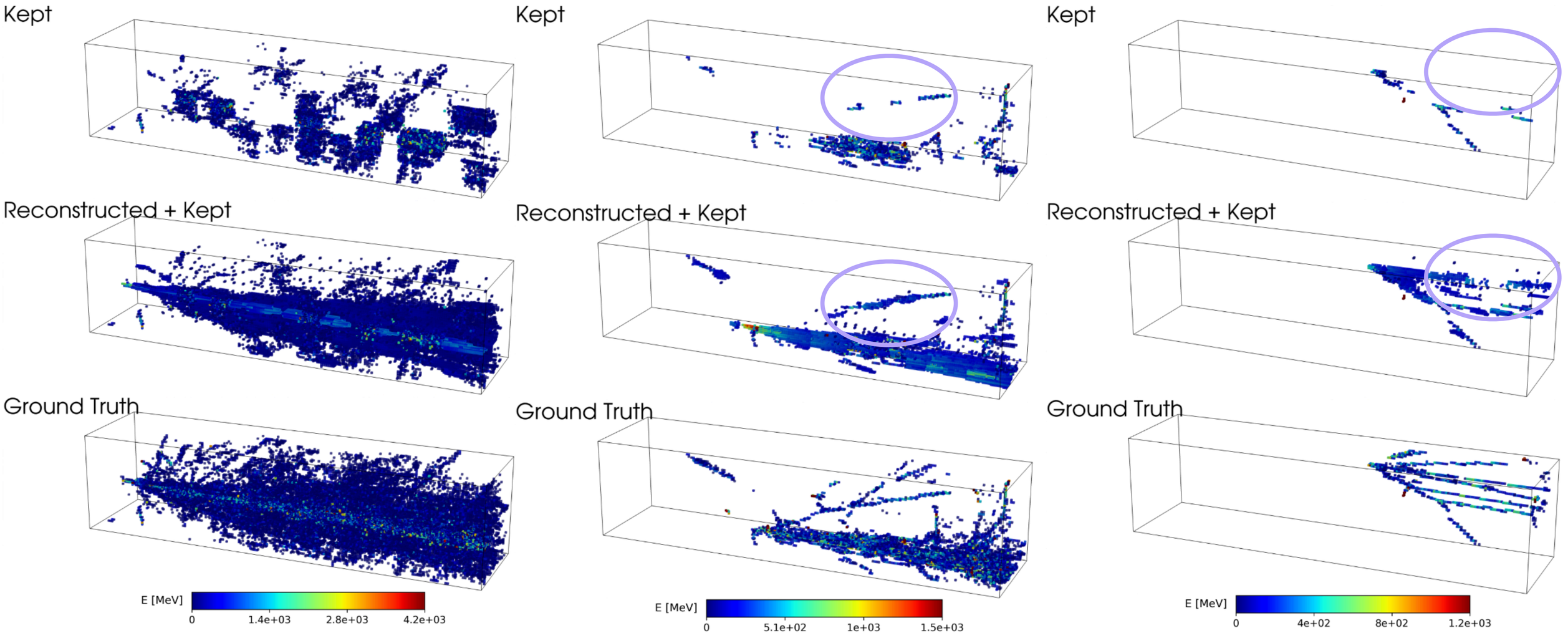
- First steps towards a foundation-style model for energy-frontier neutrino detectors
- Self-supervised pre-training is needed! — The largest gains appear in ambiguous, high-value channels such as tau and heavy-flavour-related topologies.
- The encoder transfers to SuperFGD and PILArNet, suggesting that part of the representation is reusable across detector geometries, technologies, tasks, and energy scales.
- To be clear, this is not yet a finished universal foundation model for neutrino physics.
  - But it shows that the key ingredients can coexist in one detector-aware encoder!
- **Thanks!**

Backup

Stage 1: Pre-training

These examples show how the MAE objective behaves across different event topologies.

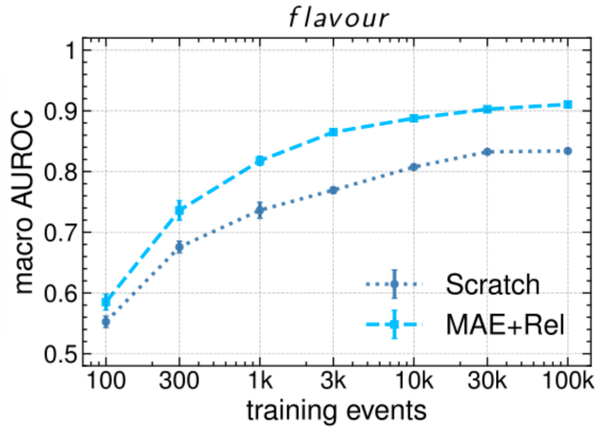
More Masked Reconstruction Examples

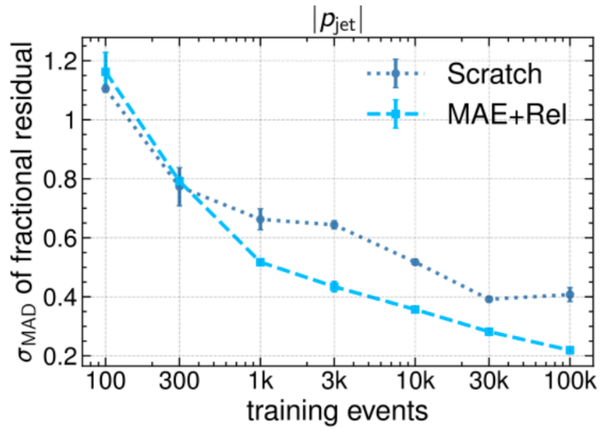


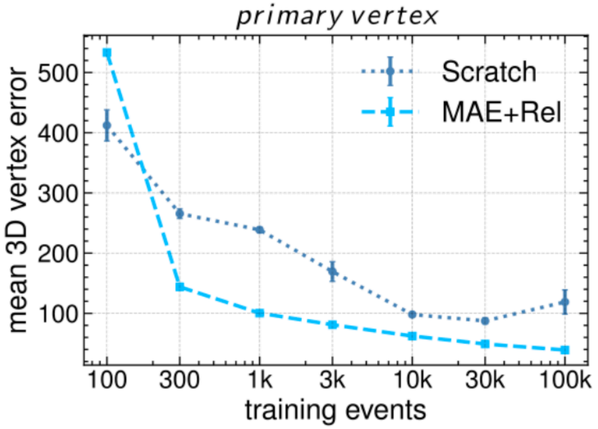
Data efficiency

Pre-training buys an order of magnitude in labelled data

- ≈10× fewer labels for matched flavour accuracy: at ~10^3 events MAE+Rel reaches what Scratch needs ~10^4 to match
- Flavour macro-AUROC at ~10^3 events: ~0.74 (Scratch) → ~0.82 (MAE+Rel)
- Vertex error at 10^3 events: 240 mm → 100 mm
- Jet-momentum σ(MAD) at 10^3 events: 0.66 → 0.52
- Gap persists at 10^5 events: flavour ~0.84 vs ~0.91; charm ~0.67 vs ~0.80

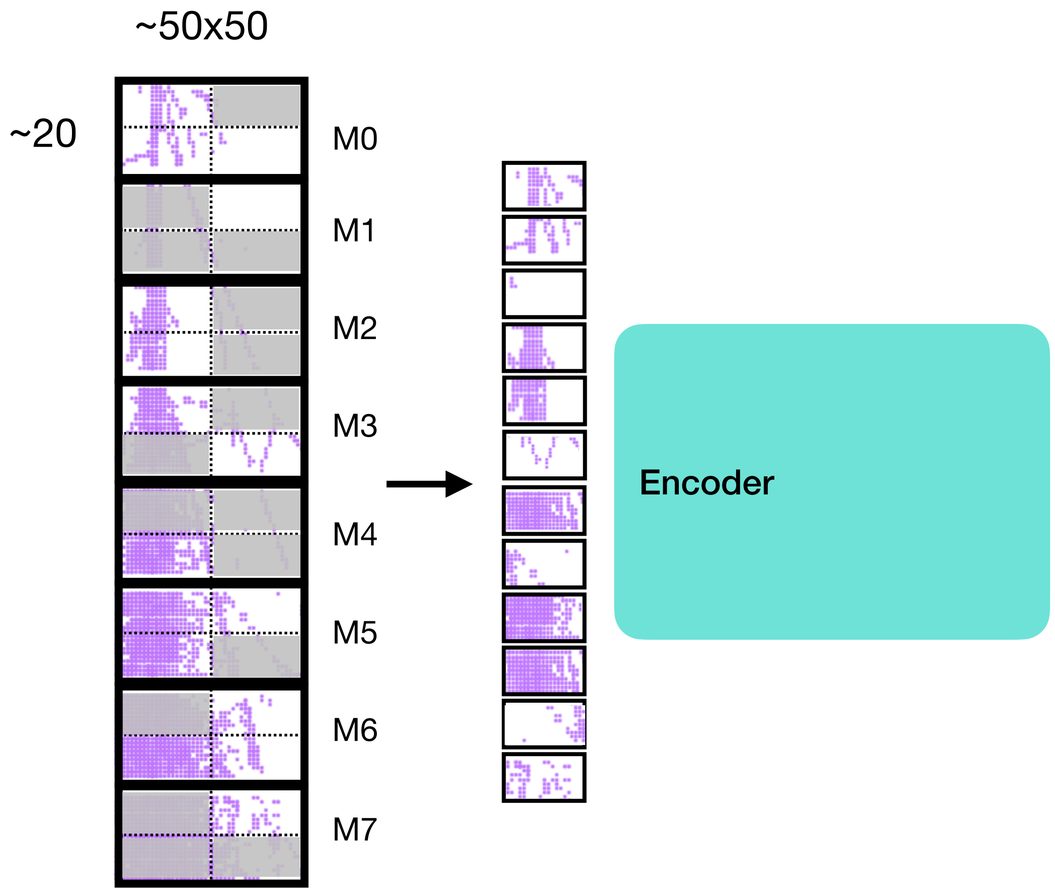




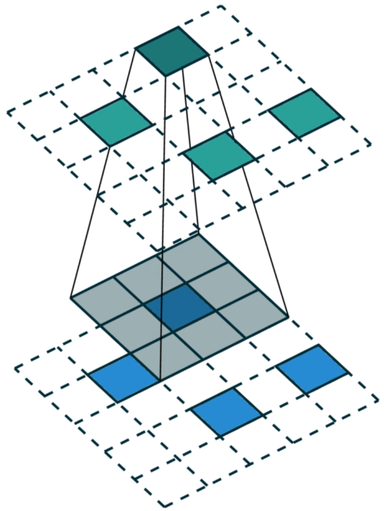


Pre-Training

Sparse Submanifold Neural Network

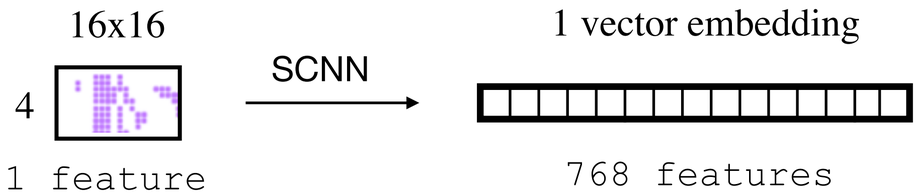


SSCN: Creates a vector embedding for every patch in the input



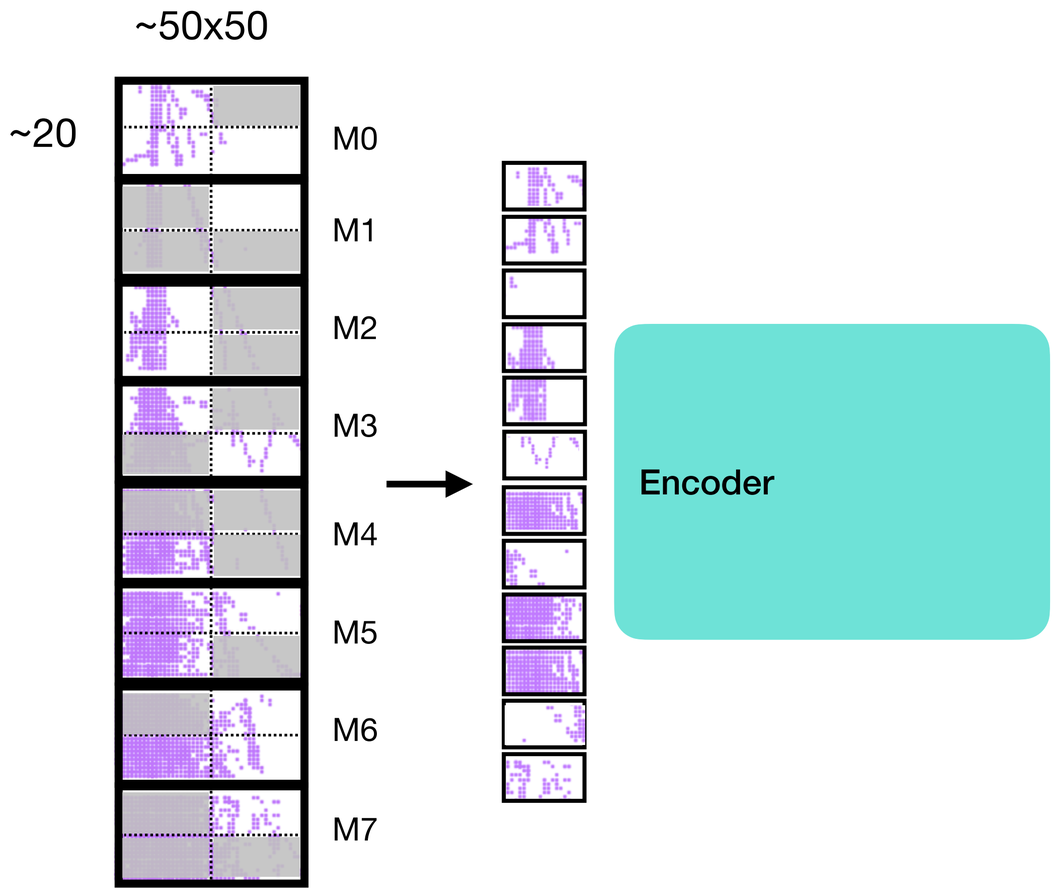
SCNN reduces the spatial coordinates while increasing the feature dimensions.

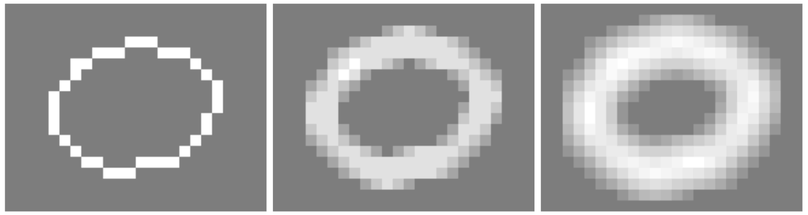
It performs convolution only on active voxels, effectively ignoring the vast empty regions typical in detector data.



Pre-Training

Sparse Submanifold Neural Network





Multiple regular 3 × 3 convolution with weights 1/9

- Sparsity on the grid rapidly disappear
- ***SCNN:*** the set of active output sites is restricted to be identical to the set of active input sites.
  - Computes convolutions ***only*** at sites that were ***already active***
- Preservation of sparsity pattern
- How is it possible to do patching?
  - *SCNN with* `Kernel size = stride` *behaves like CNN but only on active patches*

How transfer is done in practice

Details on the patches/windows

- The target detector is converted into sparse patch tokens, then the compatible pre-trained encoder core is reused and fine-tuned.

- **Target crop → sparse patch grid**
  - SuperFGD
    - Input per particle: **120 × 120 × 120** voxel crop
    - Patch size: **12 × 12 × 10**
    - Maximum patch grid:
      120/12 × 120/12 × 120/10 = 10 × 10 × 12 = 1200
      possible patch positions
    - Only occupied patches are kept → actual token count is event-dependent and much smaller.
  - PILArNet
    - Input per particle: **168 × 168 × 180** centred LArTPC crop
    - Patch size: **12 × 12 × 10**
    - Maximum patch grid:
      168/12 × 168/12 × 180/10 = 14 × 14 × 18 = 3528
      possible patch positions
    - Again, only occupied sparse patches become tokens.

- **FASERCal modules → target local windows**
  - In the original FASERCal model: 3DCal tokens are grouped by physical detector module
    For transfer, the detector has no FASERCal modules, so we replace this with local windows in patch space.
  - **SuperFGD**
    - **Patch grid: 10 × 10 × 12**
    - Window size: **2 × 2 × 2 patches**
    - Window grid: **5 × 5 × 6 = 150 possible windows**
    - Up to **2 CLS tokens per active window**
    - Maximum summaries: **150 × 2 = 300 window-summary tokens**
- **PILArNet**
  - **Patch grid: 14 × 14 × 18**
  - Window size: **2 × 2 × 3 patches**
  - Window grid: **7 × 7 × 6 = 294 possible windows**
  - Up to **2 CLS tokens per active window**
  - Maximum summaries: **294 × 2 = 588 window-summary tokens**