Toward a Foundation Model for Neutrino Physics

Self-distillation of Reusable Sensor-level Representations

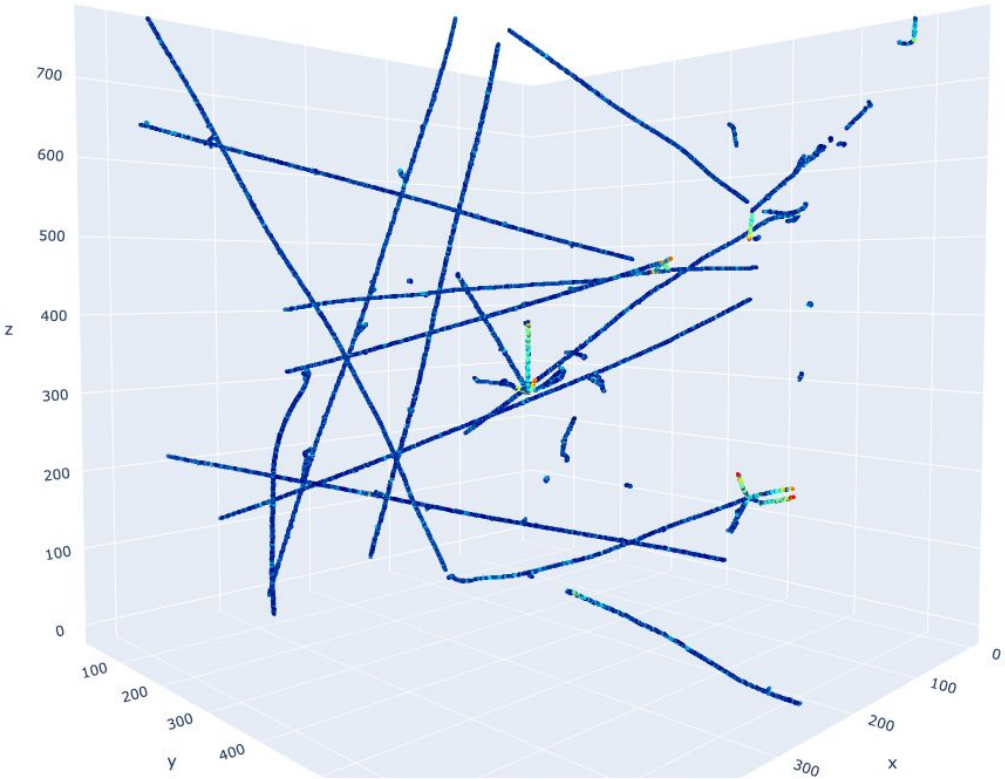
Samuel Young (Stanford)
Kazuhiro Terao (SLAC)
CHEP 2026

SLAC NATIONAL ACCELERATOR LABORATORY

Outline:

1. Our data and challenges
2. Foundation Models for sensory data: what & why
3. Early study results

Data Reconstruction/Analysis in Neutrino Experiments



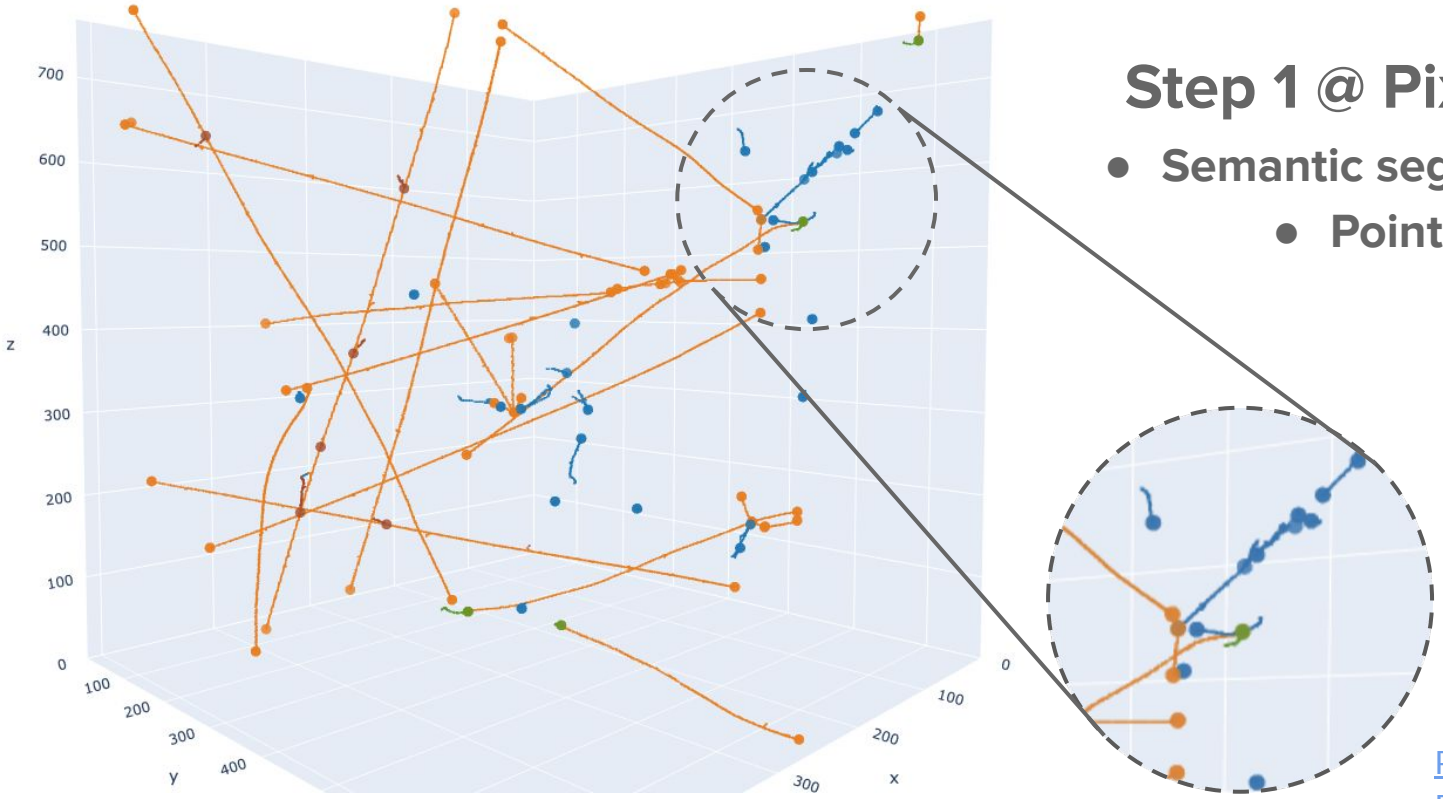
Goals:

- Identify neutrino interactions
- Infer flavor and momentum

Challenges:

- A large detector = large imaging area.
- Signal is sparse, but densely sampled.
- Signal (vertex) can happen anywhere
- Partial views (activities can occur outside)
- Large “noise” (unrelated particles)

Data Reconstruction/Analysis in Neutrino Experiments

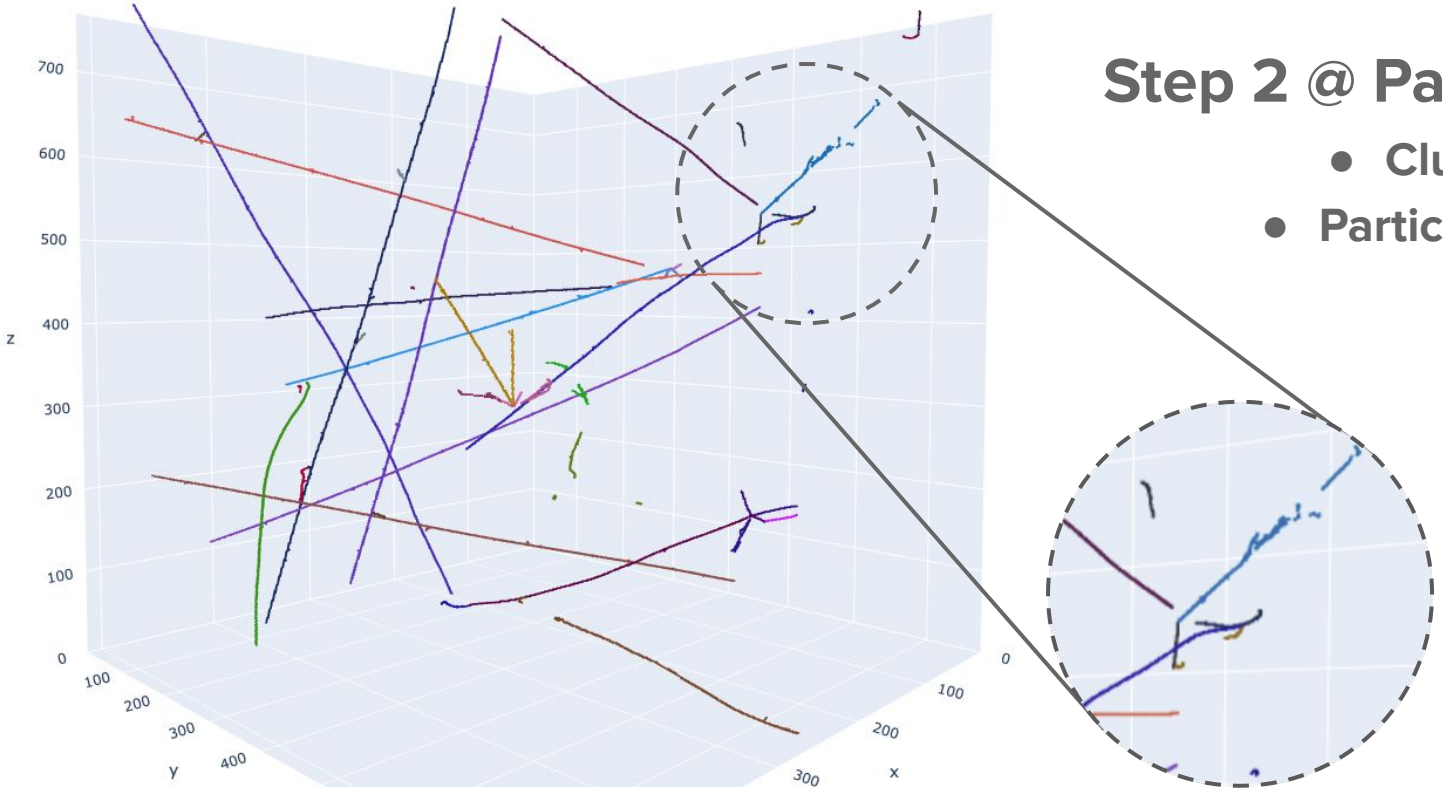


Step 1 @ Pixel-level

- Semantic segmentation
- Point prediction

Phys. Rev. D **104**, 032004
Phys. Rev. D **102**, 012005

Data Reconstruction/Analysis in Neutrino Experiments

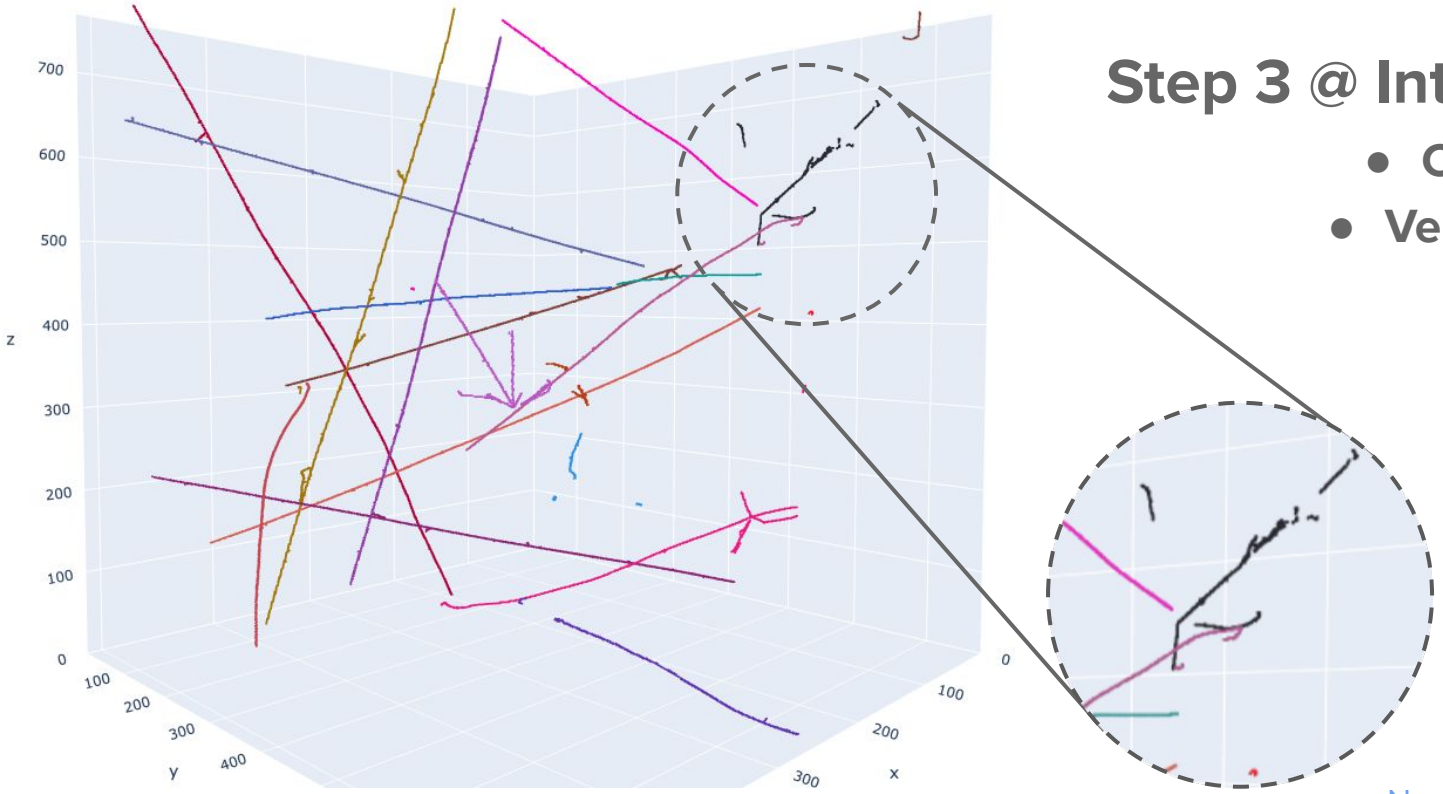


Step 2 @ Particle-level

- Clustering pixels
- Particle type, p, etc.

arXiv:2007.03083
Phys. Rev. D **104**, 072004

Data Reconstruction/Analysis in Neutrino Experiments



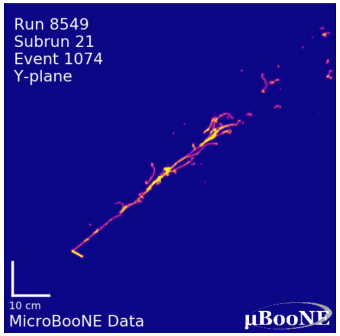
Step 3 @ Interaction-level

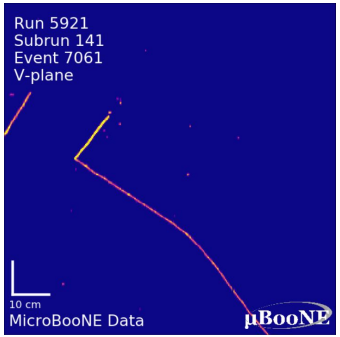
- Clustering particles
- Vertex, topology, etc.

Phys. Rev. D **104**, 072004
NeurIPS Phys. Sci. proceeding

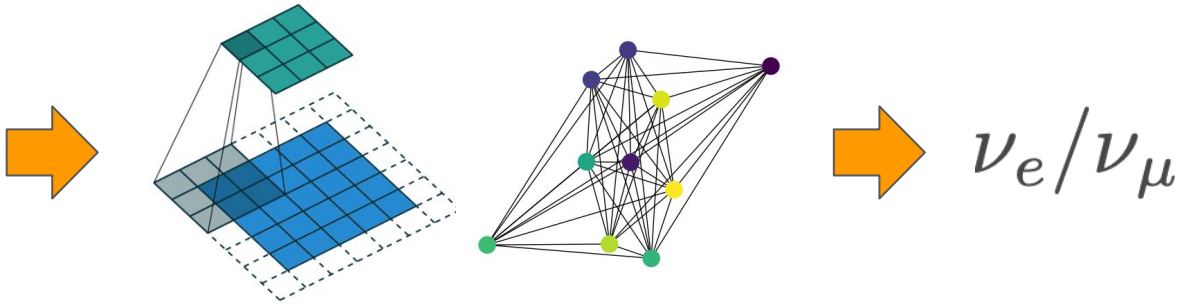
Deep learning in neutrino physics

Common approaches: deep neural network(s) trained using simulated datasets with "labels" via "supervised learning"





Deep Neural Nets



ν_e/ν_μ

"100 analyses, 100 AI models" (task-specific)

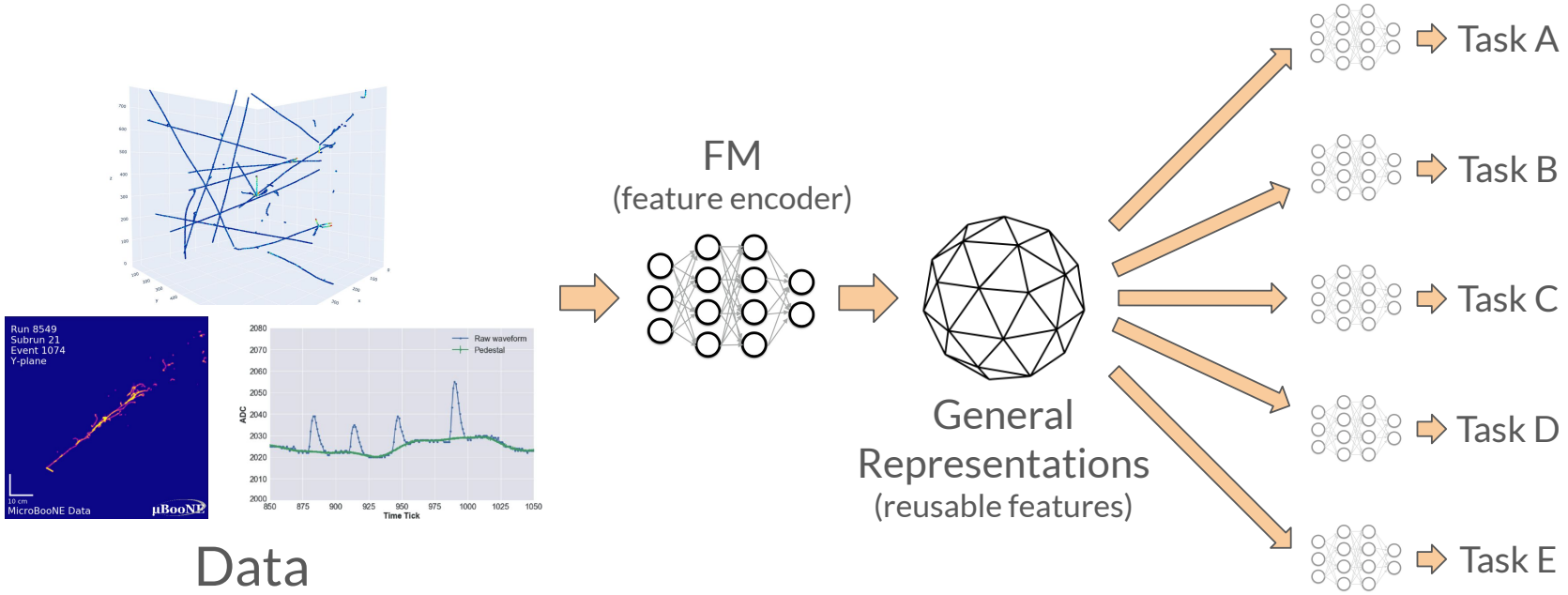
Poor reusability, cost for R&D/maintenance

Vulnerable against data-shift

High resolution + sensor-level ➔ hard to model.

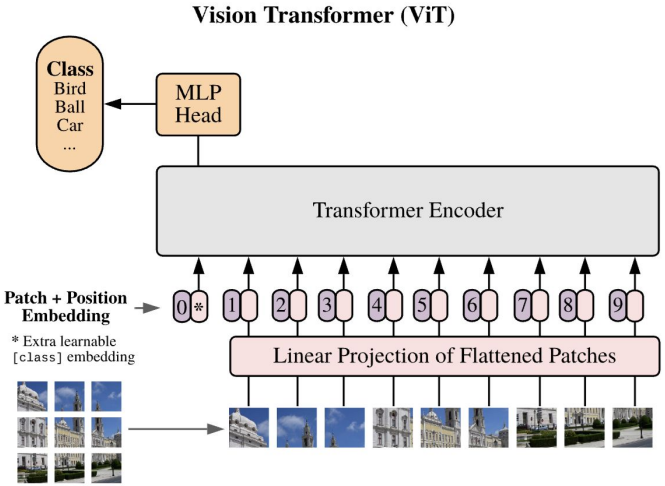
Foundation models = 2 training phases

1. **Pre-training** (Representation Learning) – computationally expensive
2. **Fine-tuning** (Adaptation) – computationally inexpensive (relatively speaking…)



SSL in Computer Vision

a cat and dog classifier will not learn anything about the difference between trees and flowers.



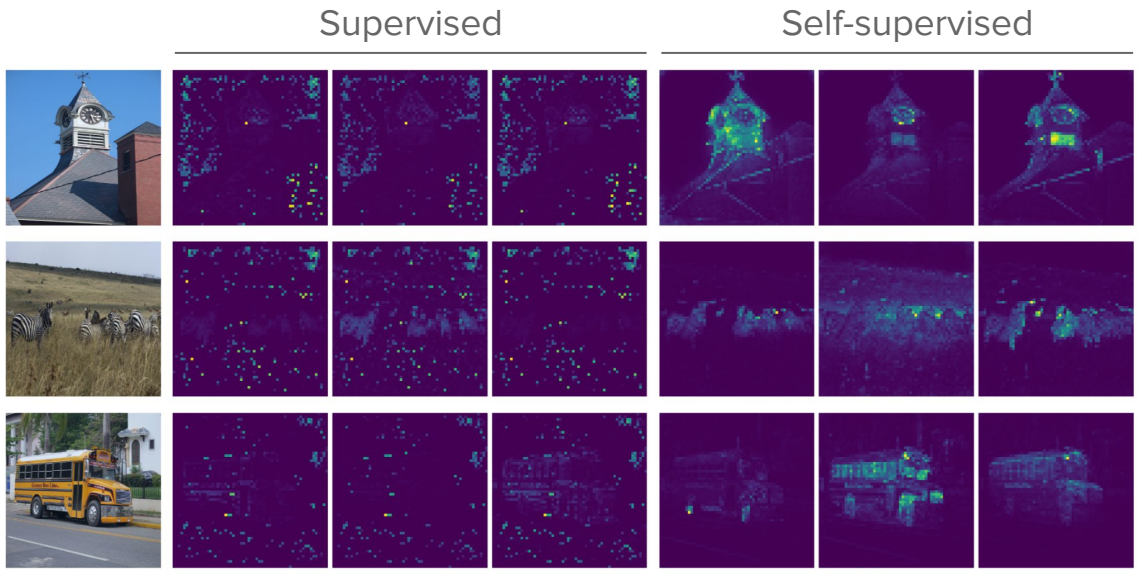
Attention maps

Attention maps from image classification in a vision transformer
DINO (2104.14294)

$$Attention(Q, K, V) = \underbrace{\text{softmax}(\frac{QK^T}{\sqrt{d_k}})}_{\text{scores}} V$$

SSL in Computer Vision

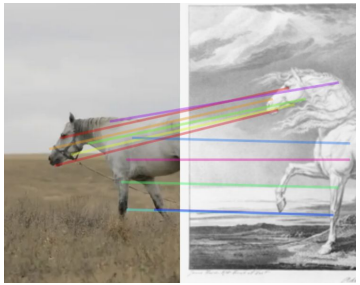
FM = learn more than the task requires so you can reuse it later



Attention maps from image classification and self-supervised tasks in a vision transformer
DINO (2104.14294)

Foundation models are generalists

Monocular depth estimation [1]

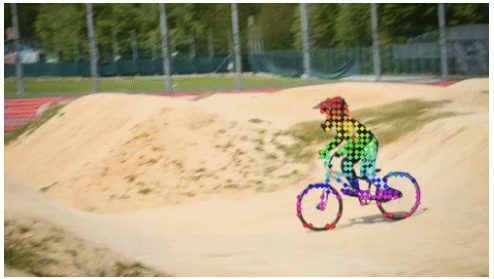


Point Correspondence [2]

self-supervised vision models are foundation models

DINO (SSL) [2]

Segmentation [3]

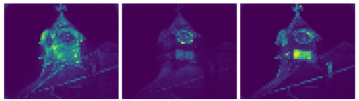


Video Tracking [4]

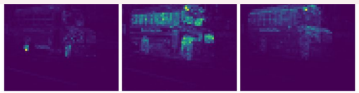
Foundation models are generalists

Monocular depth estimation [1]

self-supervised vision models are foundation models

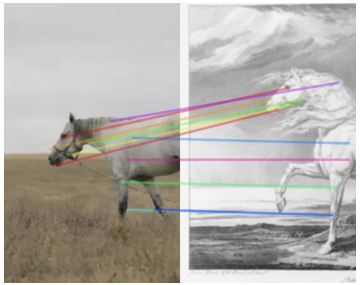


Let's apply to LArTPC data :)

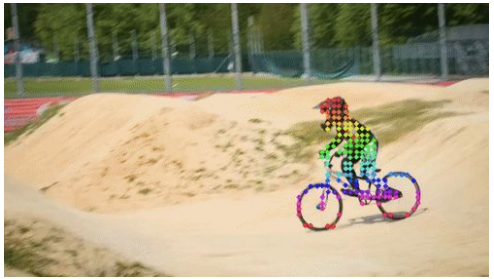


DINO (SSL) [2]

Segmentation [3]



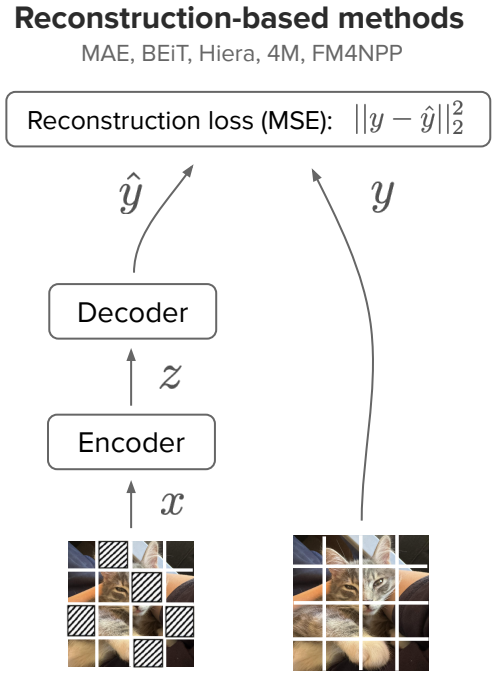
Point Correspondence [2]

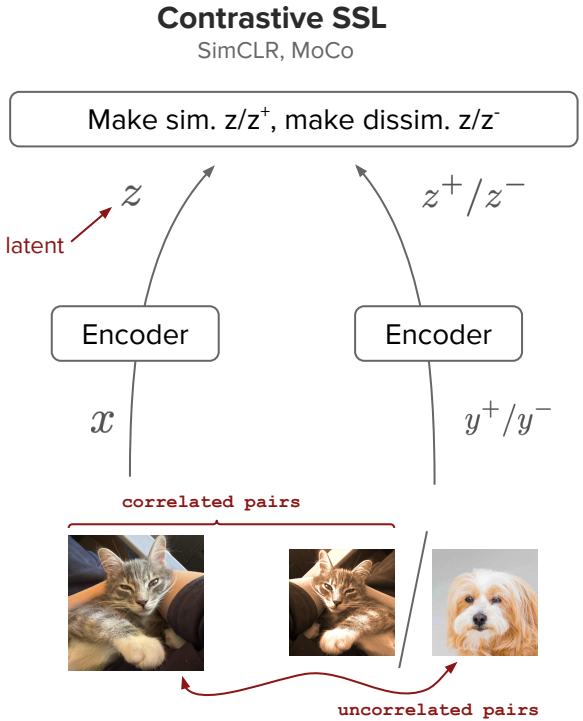


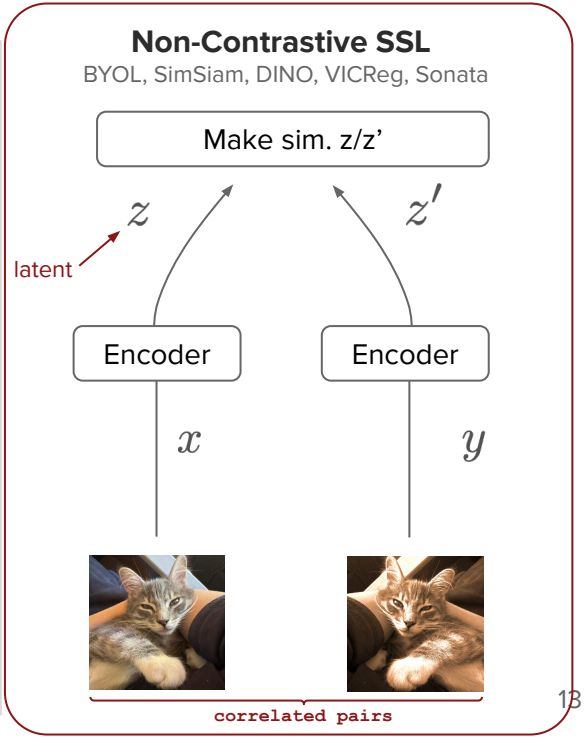
Video Tracking [4]

Techniques in self-supervised learning

learning data representations without label supervision by comparing altered/partial views of the same input.







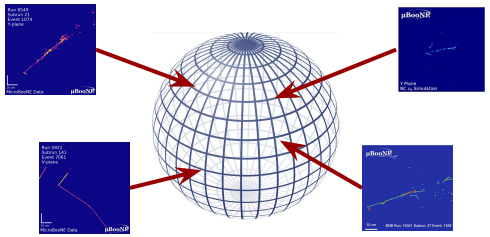
Panda: Self-distillation and hierarchy

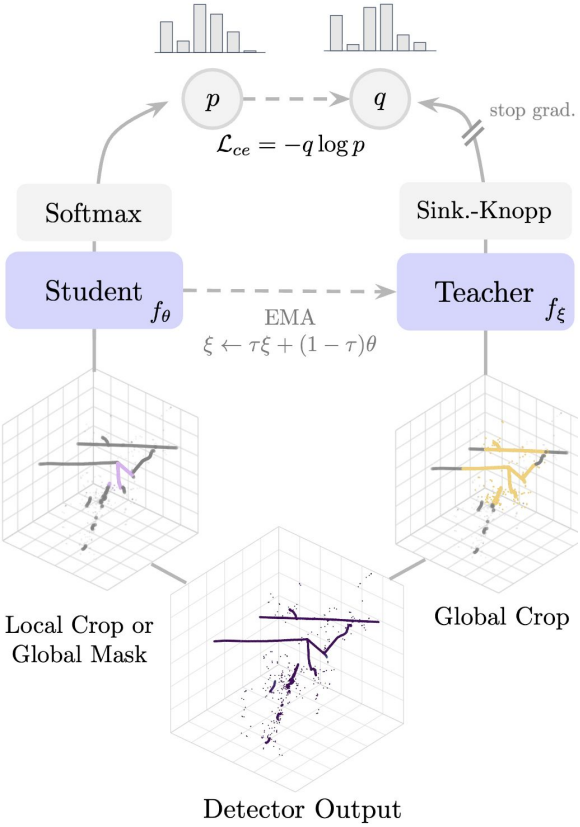
instead of **reconstructing masked portions of image directly**,

let's **predict where they would would end up on a unit sphere** (i.e., classify them),

by enforcing **consistency between global and local views** of the same image under **strong augmentations**.

this is **non-contrastive SSL.**



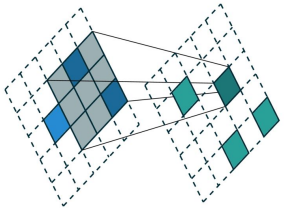


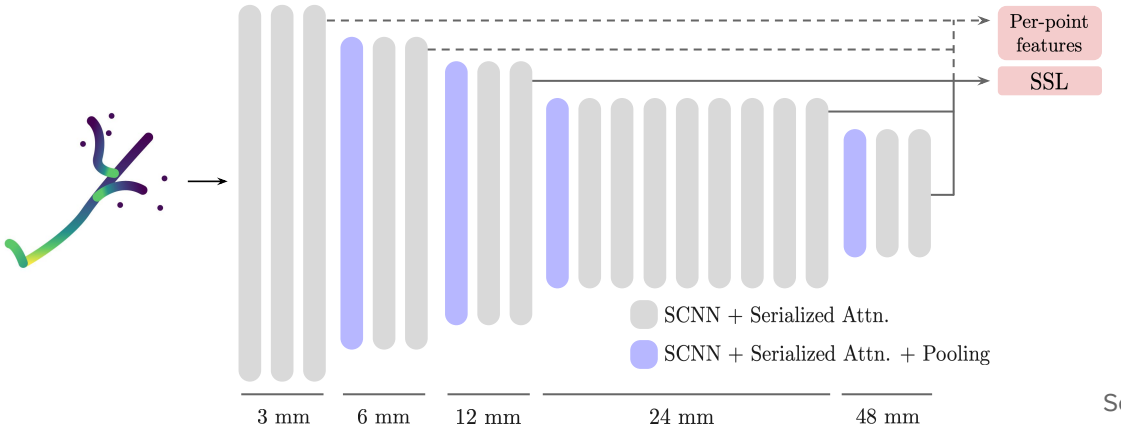
Panda: Self-distillation and hierarchy

Point Transformer V3 (CVPR ‘24)

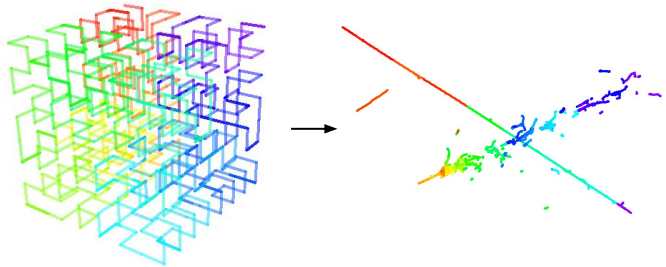
- Efficient serialization for approximate nearest neighbor (no kNN!)
- Local patch hierarchical attention
  - Pooling 5x, 48 to 512 features, patch size 256; 90M params
- Result: more expressive than Sparse UResNet, still scalable (compute scales linearly wrt # points)
- Dataset: PILArNet-M = 1M 3D point clouds (10k~100k/image)

SCNN: Sparse CNN





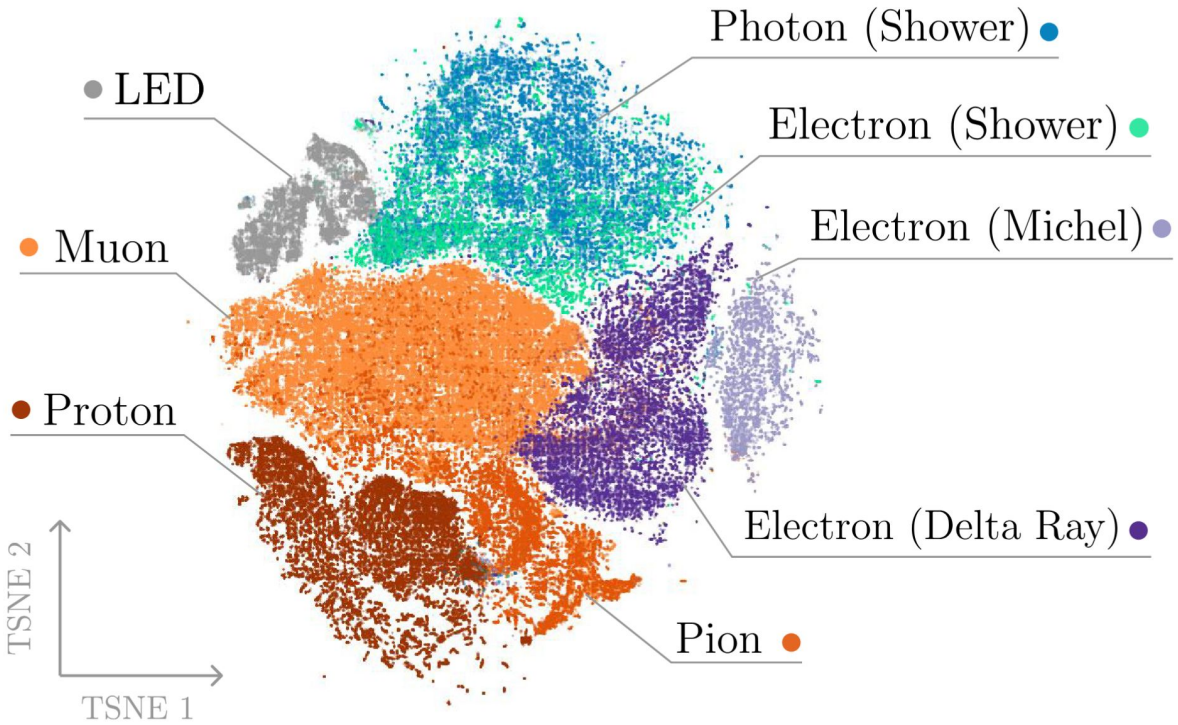
Approximate nearest neighbor via locality-sensitive hashing (Hilbert curve, Z-curve)



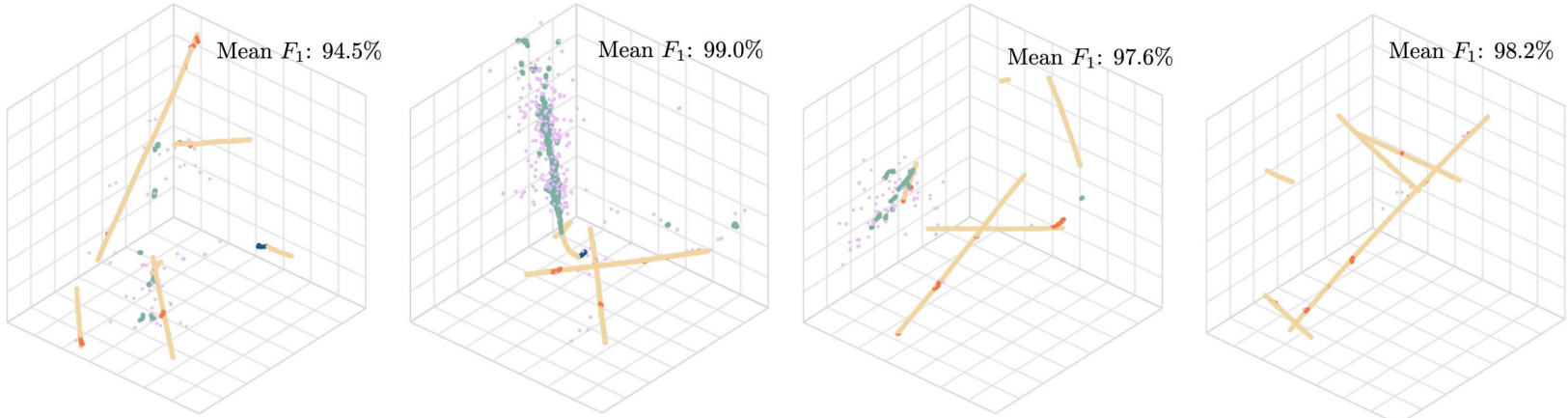
Serialize ➙ cut into non-overlapping patches ➙ windowed attention

(Note: muon length O(1 m) ⇒ ~20 coarse voxels / muon)

Learned representations

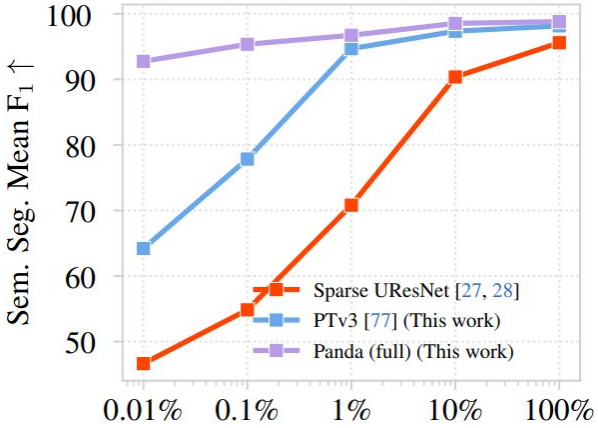


Task A: Semantic Segmentation

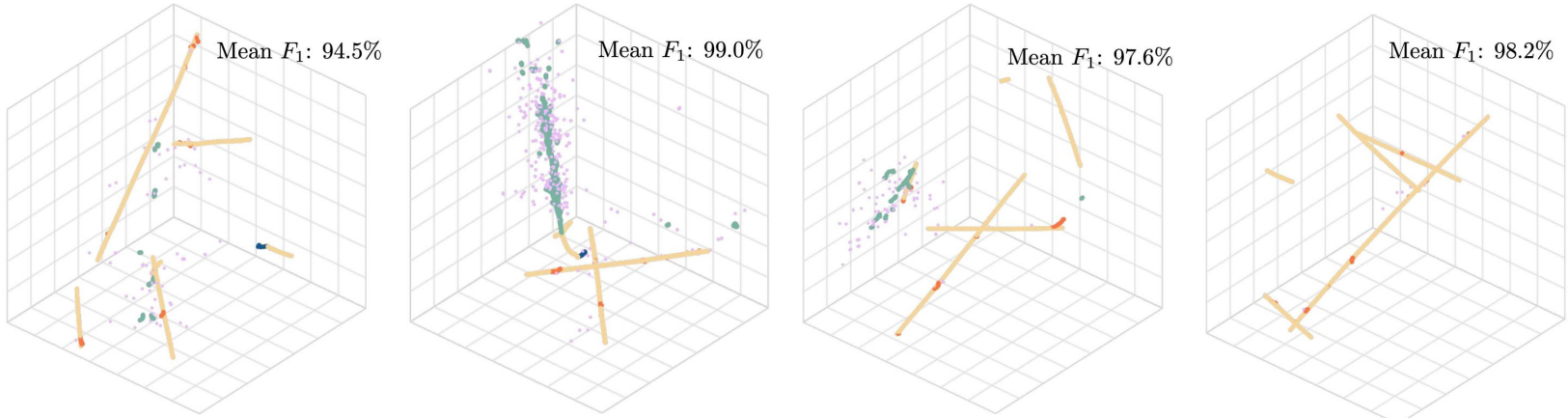


Semantic Segmentation	$PT_{K\%} + FT_{K\%}$					$PT_{100\%} + FT_{K\%}$				
Method / $K\%$	0.01%	0.1%	1%	10%	100%	0.01%	0.1%	1%	10%	100%
Supervised										
○ UResNet [27, 28]	46.6	54.8	70.8	90.4	95.6	46.6	54.8	70.8	90.4	95.6
● PTv3 [77]	64.2	77.8	94.7	97.3	98.2	64.2	77.8	94.7	97.3	98.2
Self-supervised										
● Panda (lin.)	79.1	90.1	93.2	93.7	93.9	92.2	93.6	93.8	93.9	93.9
● Panda (dec.)	83.5	92.8	95.9	97.3	98.2	92.6	95.2	96.2	97.8	98.2
● Panda (full)	**85.2**	**93.7**	**96.4**	**97.7**	**98.8**	**92.8**	**95.4**	**96.7**	**98.5**	**98.8**

○ Prev. SOTA ● This work

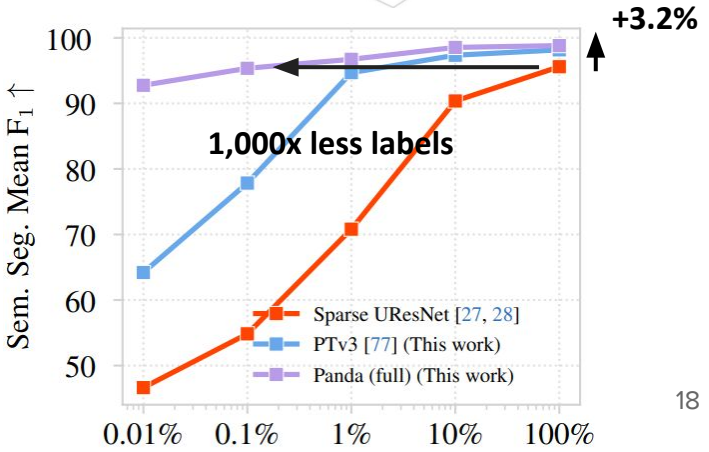


Task A: Semantic Segmentation

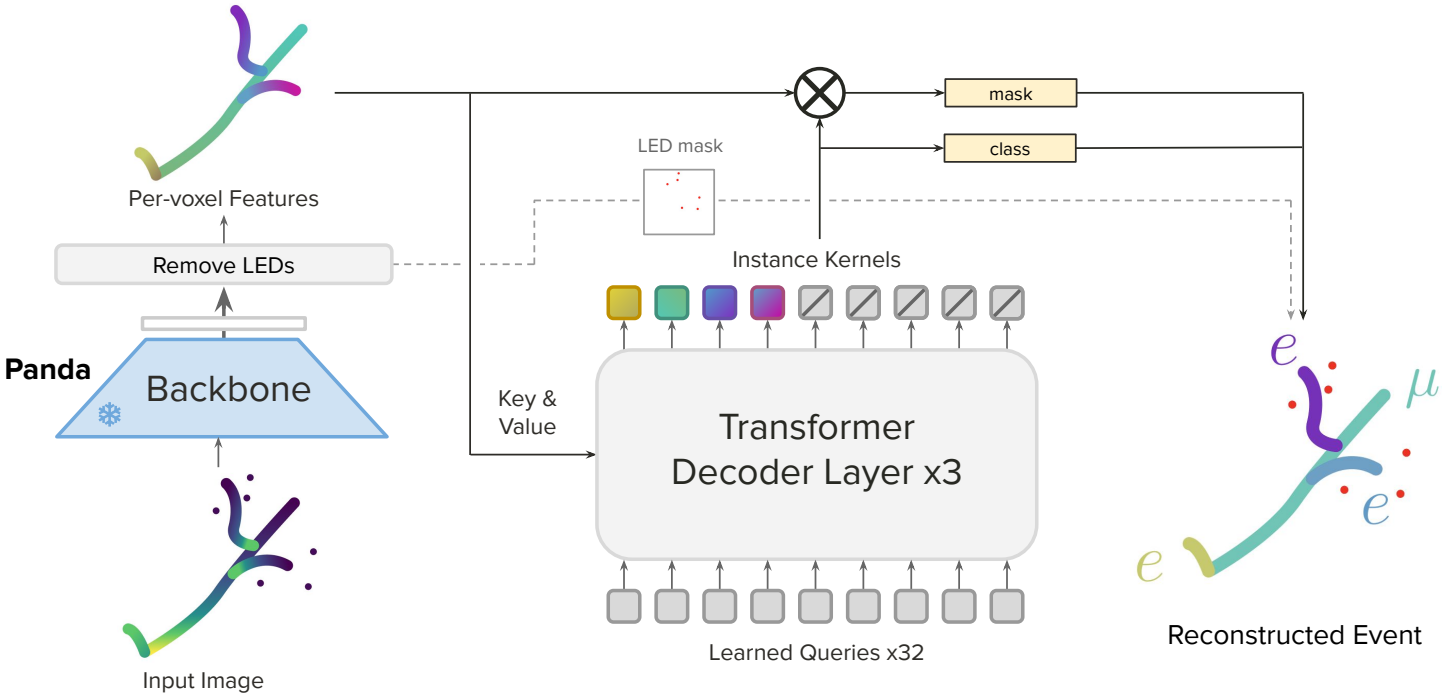


Semantic Segmentation	$PT_{K\%} + FT_{K\%}$					$PT_{100\%} + FT_{K\%}$				
Method / $K\%$	0.01%	0.1%	1%	10%	100%	0.01%	0.1%	1%	10%	100%
Supervised										
○ UResNet [27, 28]	46.6	54.8	70.8	90.4	95.6	46.6	54.8	70.8	90.4	95.6
● PTv3 [77]	64.2	77.8	94.7	97.3	98.2	64.2	77.8	94.7	97.3	98.2
Self-supervised										
● Panda (lin.)	79.1	90.1	93.2	93.7	93.9	92.2	93.6	93.8	93.9	93.9
● Panda (dec.)	83.5	92.8	95.9	97.3	98.2	92.6	95.2	96.2	97.8	98.2
● Panda (full)	**85.2**	**93.7**	**96.4**	**97.7**	**98.8**	**92.8**	**95.4**	**96.7**	**98.5**	**98.8**

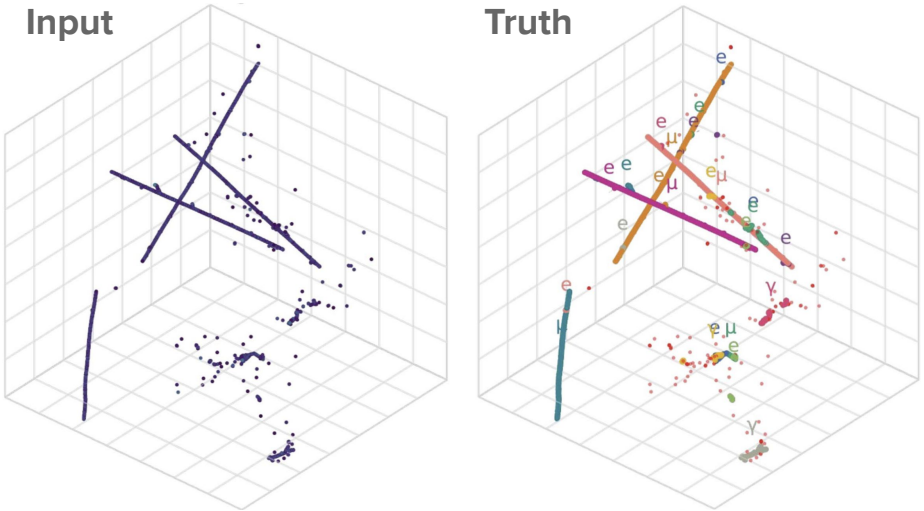
○ Prev. SOTA ● This work

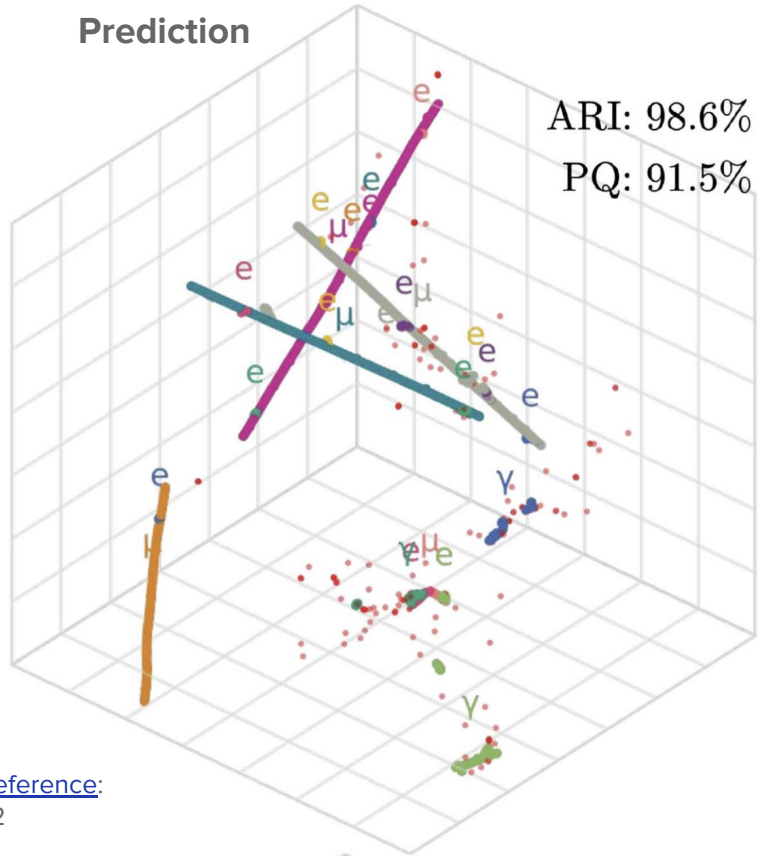


Instance Segmentation: separating particles from one another



Task B: Clustering Pixels Into A Particle + Type ID



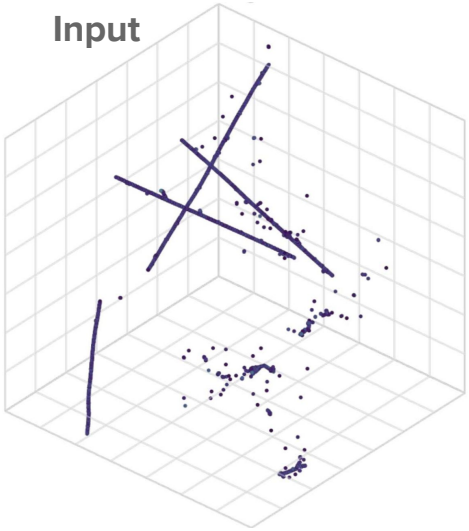


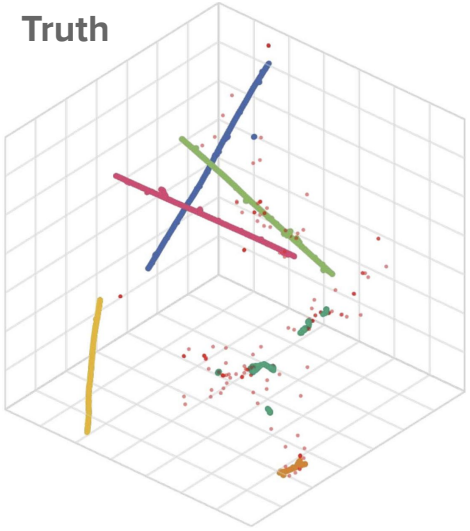
Instance Seg.	Param.	Interaction Id.		Particle Identification						
Method		PQ	ARI	PQ	ARI	γ	e	μ	π	p
Supervised										
• PTv3 [77]	95.1 M	96.3	93.7	86.9	96.6	93.5	94.9	98.6	94.2	98.1
Self-supervised										
• Panda (dec.)	4.4 M	96.6	94.5	89.5	97.3	96.2	96.3	99.1	95.7	98.4
• Panda (full)	95.1 M	**97.6**	**94.9**	**92.5**	**98.0**	**98.4**	**97.2**	**99.3**	**96.0**	**98.6**

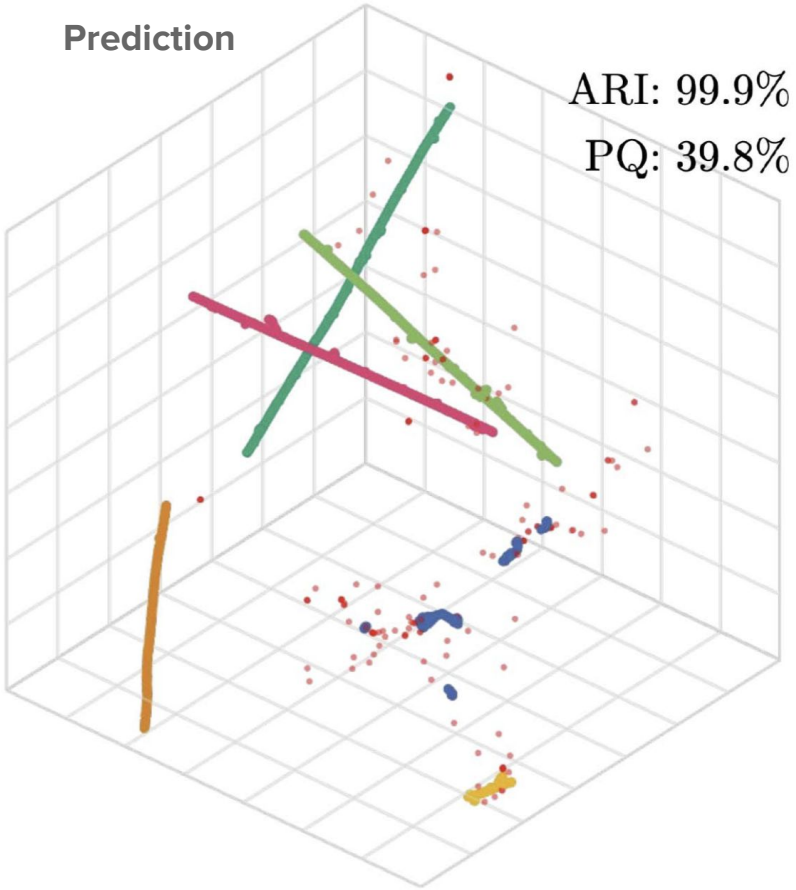
• This work

SPINE reference:
ARI 98.2

Task C: Clustering Particles Into an Interaction







Instance Seg.	Param.	Interaction Id.		Particle Identification						
Method		PQ	ARI	PQ	ARI	γ	e	μ	π	p
Supervised										
• PTv3 [77]	95.1 M	96.3	93.7	86.9	96.6	93.5	94.9	98.6	94.2	98.1
Self-supervised										
• Panda (dec.)	4.4 M	96.6	94.5	89.5	97.3	96.2	96.3	99.1	95.7	98.4
• Panda (full)	95.1 M	**97.6**	**94.9**	**92.5**	**98.0**	**98.4**	**97.2**	**99.3**	**96.0**	**98.6**

• This work

Summary and outlook

Project site: youngsm.com/panda
Code: github.com/deeplearnphysics/panda

Summary:

- Popular methods of label-free self-supervised learning in regular computer vision can be applied successfully to HEP pattern recognition problems.
- The field is moving towards larger models, more data, and more compute.
  - High upfront cost of pretraining is largely amortized across downstream tasks.
- Efforts currently focus on efficient transfer learning across tasks, but…

Outlook:

- There is current interest to experiment with cross-experimental FMs (e.g., FM for DUNE ND+FD).
- Multi-modal pre-training being investigated (e.g., charge + light)
- AI-ready 200 TB dataset with O(10M) LArTPC + Water Cherenkov events targeted for Q4 2026.
- Common benchmarks for community to optimize for being created for shared AI/ML R&D efforts in nu phys (DUNE, SuperK, IceCube, …); to be presented at NPML 2026.

Extras

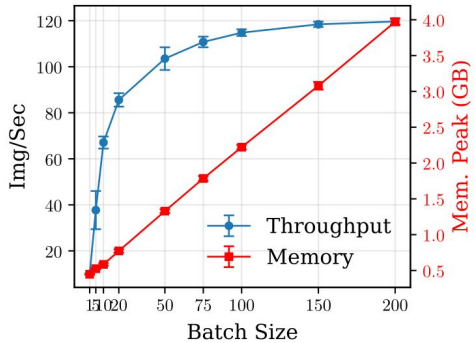
Needs!

Resources

- Compute: larger models, more compute, more data = better (FM scaling law)
  - Vision models are dominated largely by data memory (as opposed to a model)

Panda (A100)

Pre-training 10 epochs for 1M image dataset (~ 10 GPU-day)



Config.	Batch	Params	Wall-time [s/iter]	Data Xfer [s/iter]	VRAM [GB]
Pre-train	12	104M	0.895	0.005	34.3
SemSeg (dec)	12	16.3M	0.210	0.001	3.4
PanSeg (vtx-dec)	6	4.3M	0.292	0.005	5.3
PanSeg (pid-dec)	6	4.3M	0.321	0.004	12.7

Needs!

Resources

- Compute: larger models, more compute, more data = better (FM scaling law)
  - Vision models are dominated largely by data memory (as opposed to a model)
  - A tiny R&D models 5-10 A100-day, for a large model x10-x100
  - Fine-tuning (task-specific models) ~ 1 A100-day per task
- Storage: large data size
  - PILArNet-1M (3D point cloud) ~167GB
  - PILArNet-10M (+ TPC/Optical waveforms) ~200TB
  - Publishing “encoded, AI-ready dataset” … somewhere in between?

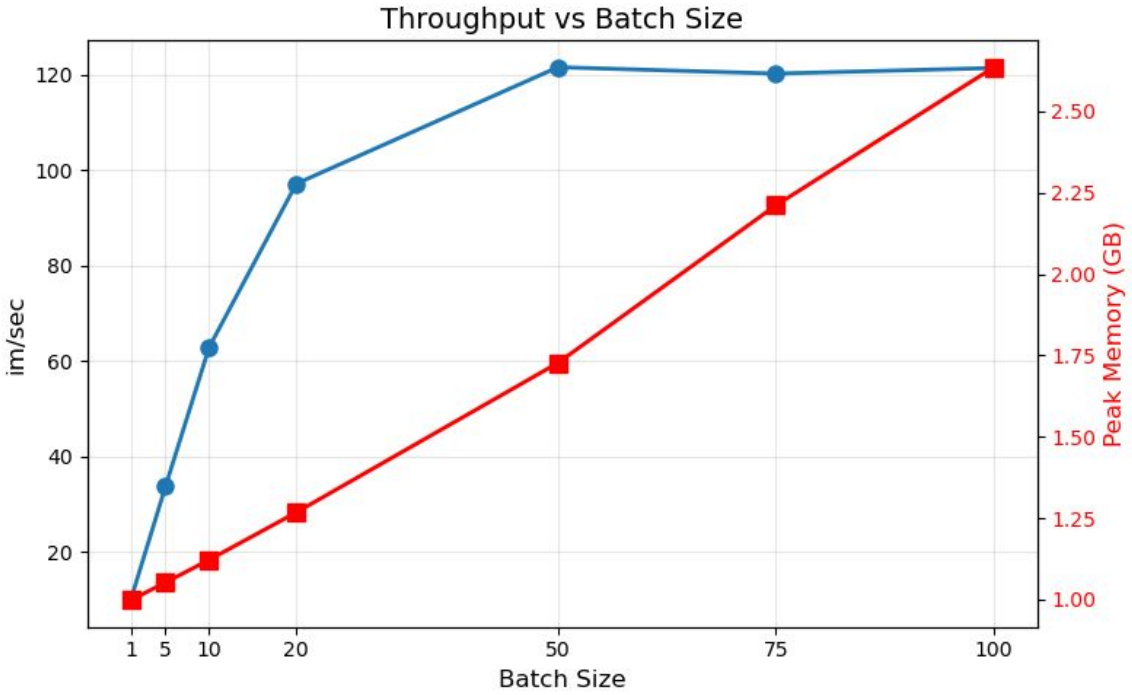
Needs!

Resources

- Compute: larger models, more compute, more data = better (FM scaling law)
  - Vision models are dominated largely by data memory (as opposed to a model)
  - A tiny R&D models 5-10 A100-day, for a large model x10-x100
  - Fine-tuning (task-specific models) ~ 1 A100-day per task
- Storage: large data size
  - PILArNet-1M (3D point cloud) ~167GB
  - PILArNet-10M (+ TPC/Optical waveforms) ~200TB
  - Publishing “encoded, AI-ready dataset” … somewhere in between?
- Software ecosystem - (always containers)
  - The core: pytorch, spconv, flash-attn, torch-scatter, pytorch-lightning, omegaconf (hydra), pybind11, hugging-face
  - Auxiliary: numba, cupy, jax, WandB, warpconvnet

Scalability

Semantic Segmentation (measured on single A100)

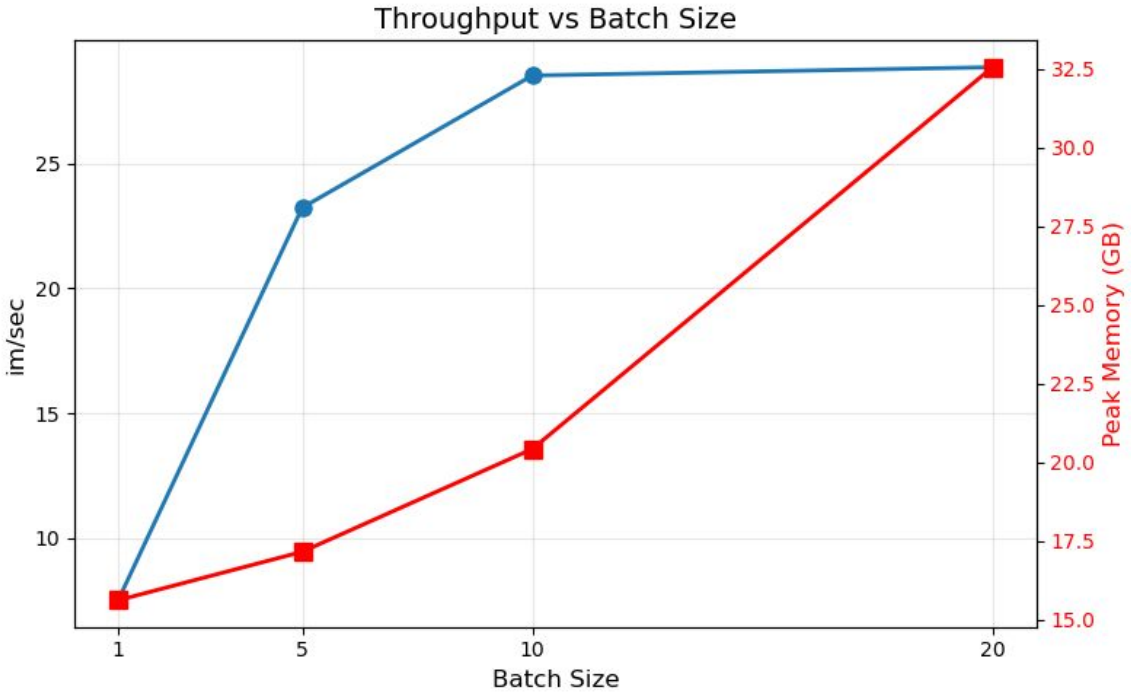


Peak throughput:
8.3 ms/image
120 img/sec

Mem@peak:
~1.75 GB

Scalability

Instance/Panoptic Segmentation (model forward + NMS post-processing)



Peak throughput:

34.5 ms / image

29 img/sec

Mem@peak:

~20 GB

NMS post-processing is serial, but parallelized via multiprocessing

Sharpening + Centering

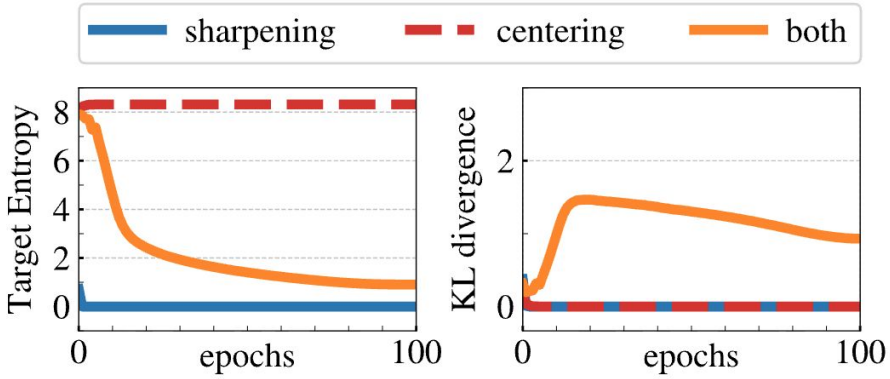


Figure 7: **Collapse study. (left)**: evolution of the teacher's target entropy along training epochs; **(right)**: evolution of KL divergence between teacher and student outputs.

There are two forms of collapse: regardless of the input, the model output is uniform along all the dimensions or dominated by one dimension. The centering avoids the collapse induced by a dominant dimension, but encourages an uniform output. Sharpening induces the opposite effect. We show this complementarity by decomposing the cross-entropy H into an entropy h and the Kullback-Leibler divergence ("KL") D_{KL}:

$$H(P_t, P_s) = h(P_t) + D_{KL}(P_t|P_s). \quad (5)$$

A KL equal to zero indicates a constant output, and hence a collapse.

Sharpening:

$$P_s(x)^{(i)} = \frac{\exp(g_{\theta_s}(x)^{(i)}/\tau_s)}{\sum_{k=1}^{K} \exp(g_{\theta_s}(x)^{(k)}/\tau_s)}, \quad (1)$$

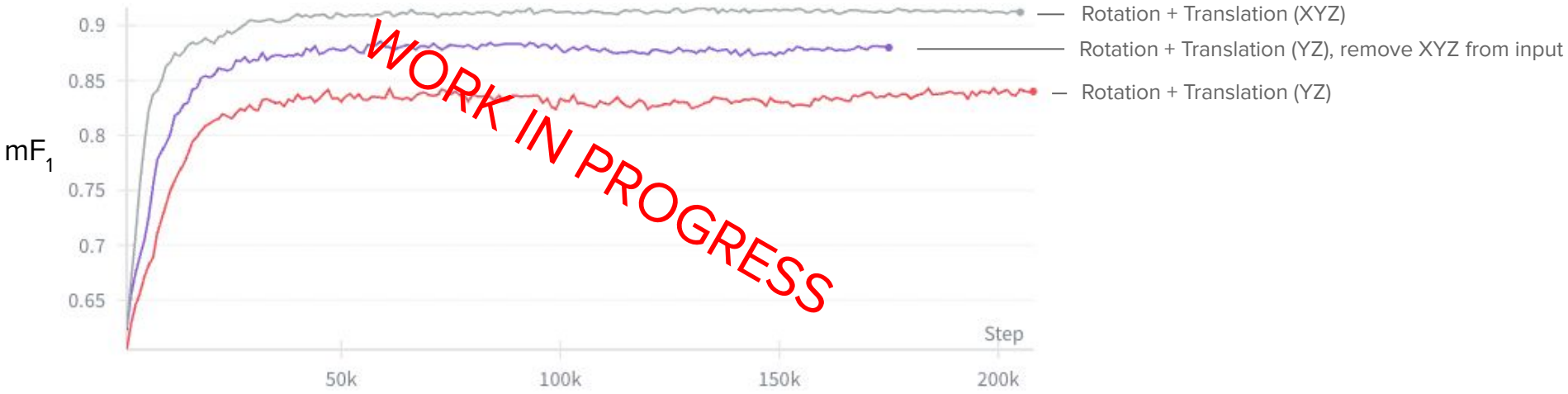
with $\tau_s > 0$ a temperature parameter

From *Emerging Properties in Self-Supervised Vision Transformers* (arXiv:2104.14294)

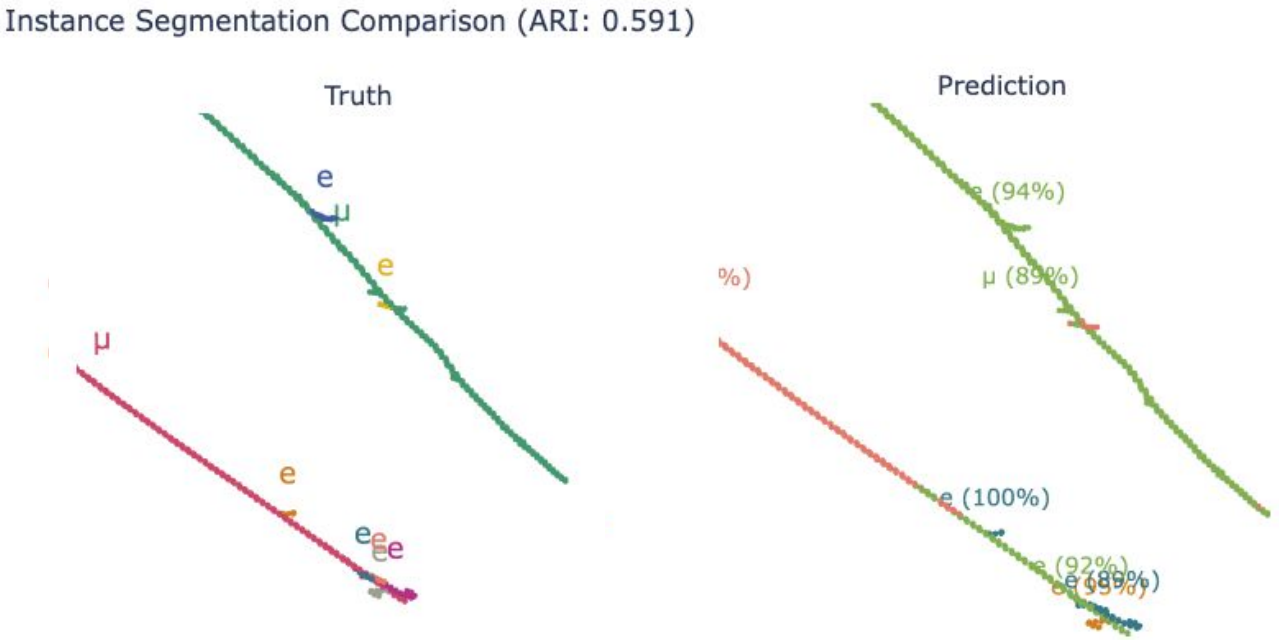
Augmentations

What about diffusion/attenuation?

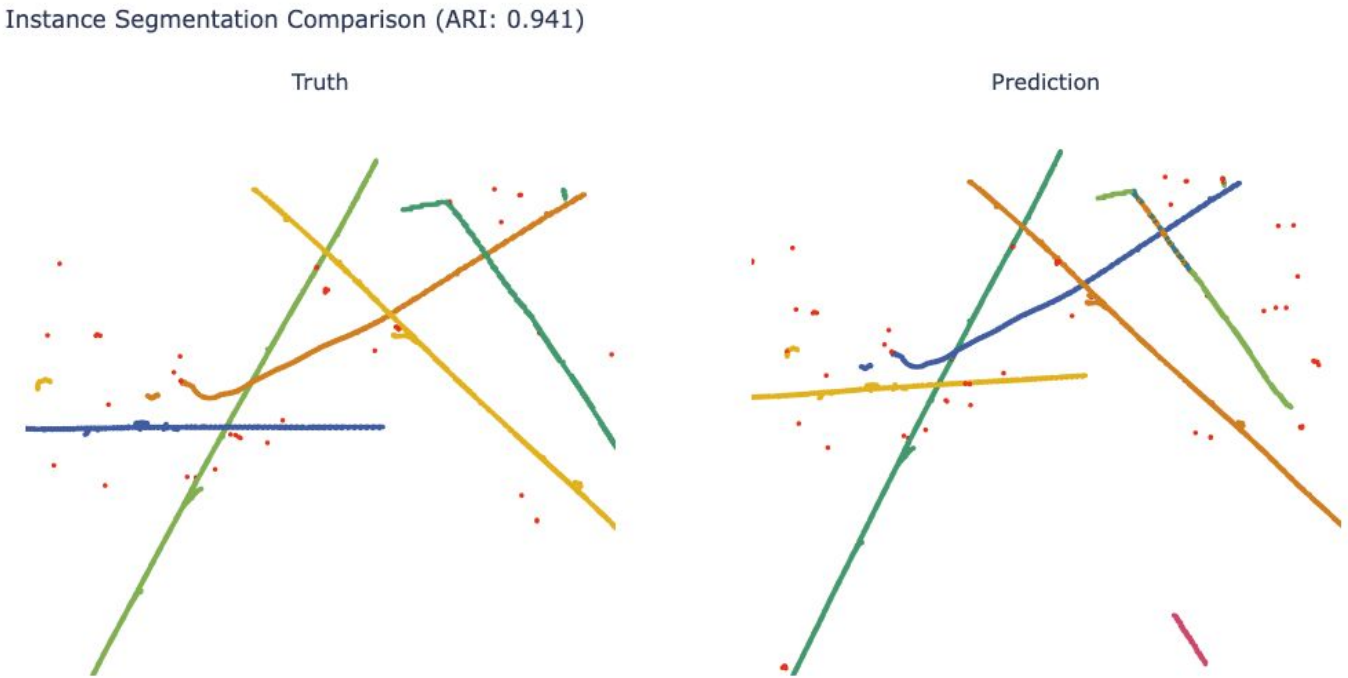
Linear Evaluation on pre-training – 1M events



Poor instance reconstruction – particle

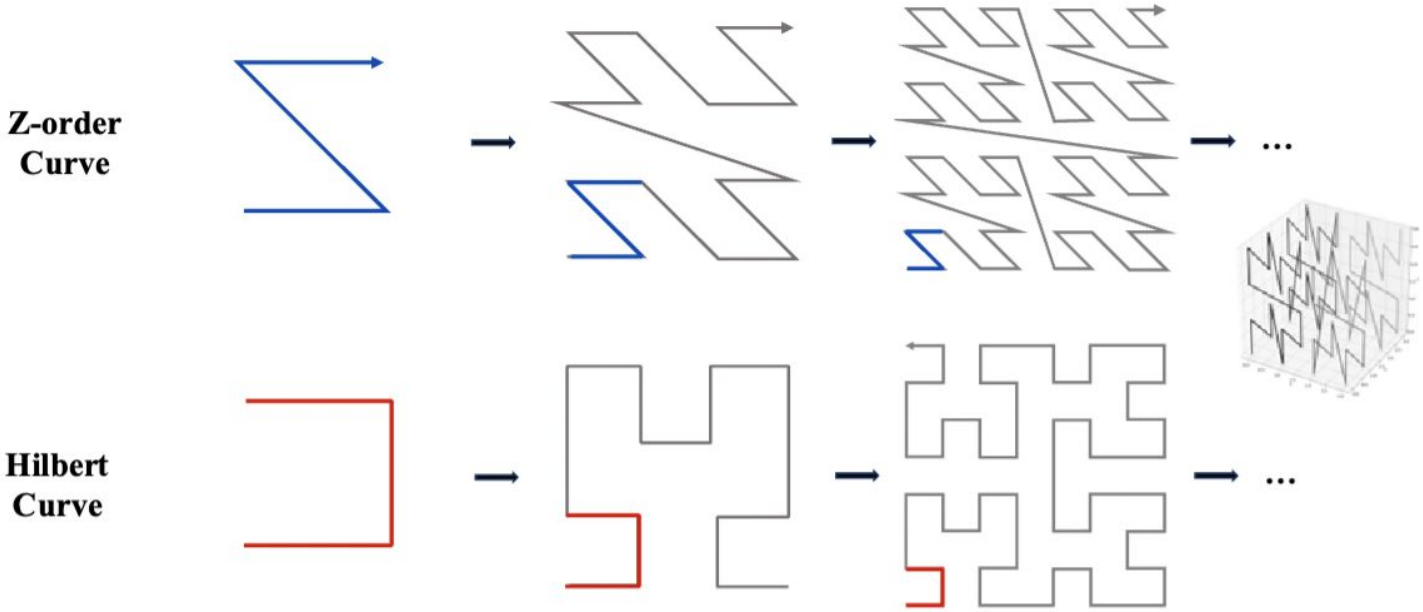


Poor instance reconstruction - interaction



Serialized Attention

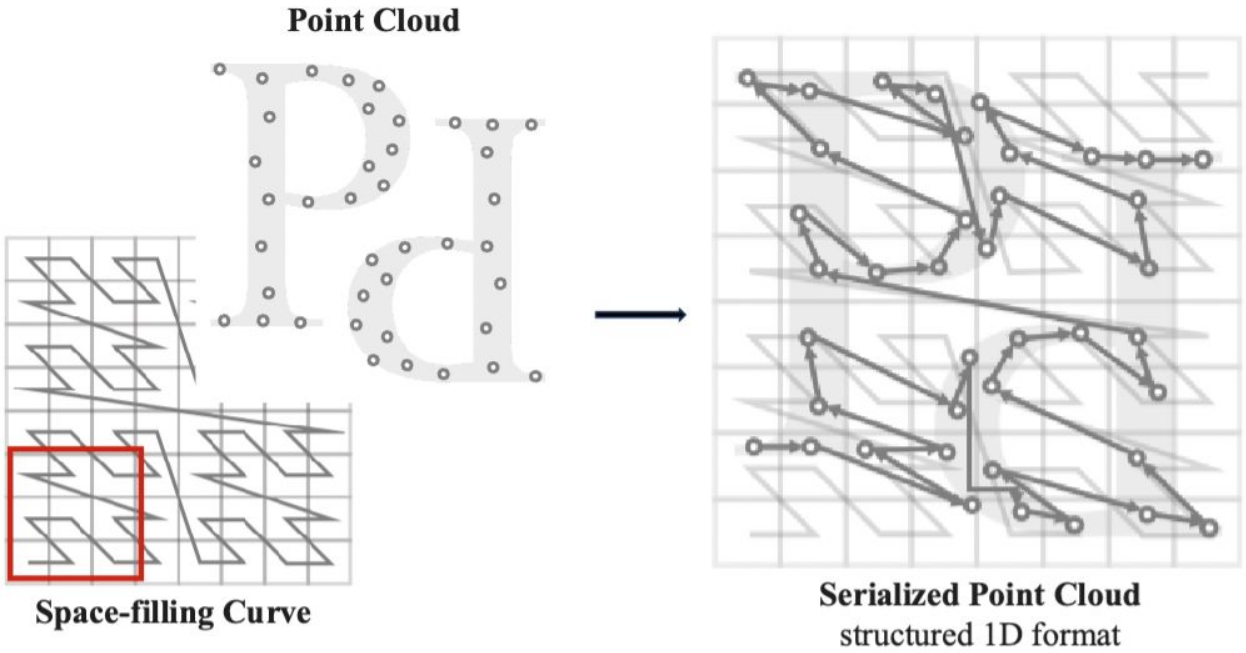
See Point Transformer V3 paper (arXiv:2312.10035) for more detail



From https://xywu.me/talk/20250609_Stanford_Spatial_Representations.pdf

Serialized Attention

See Point Transformer V3 paper (arXiv:2312.10035) for more detail

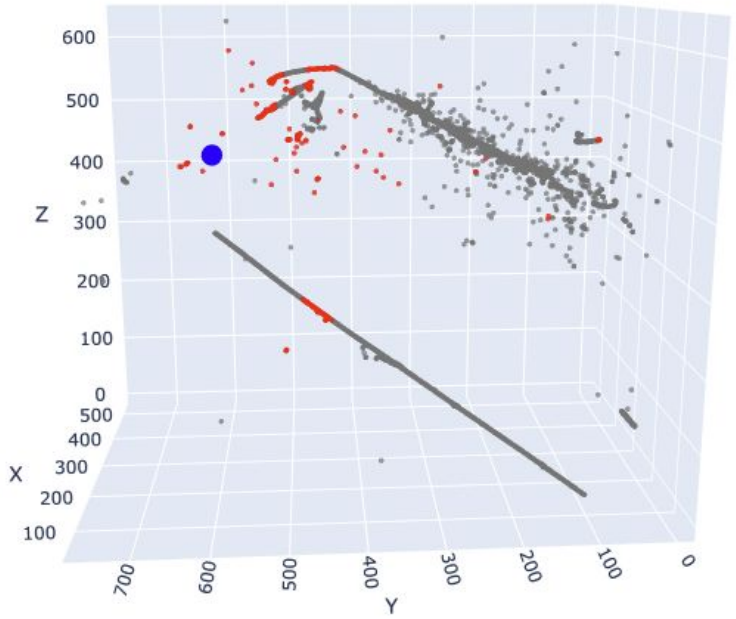


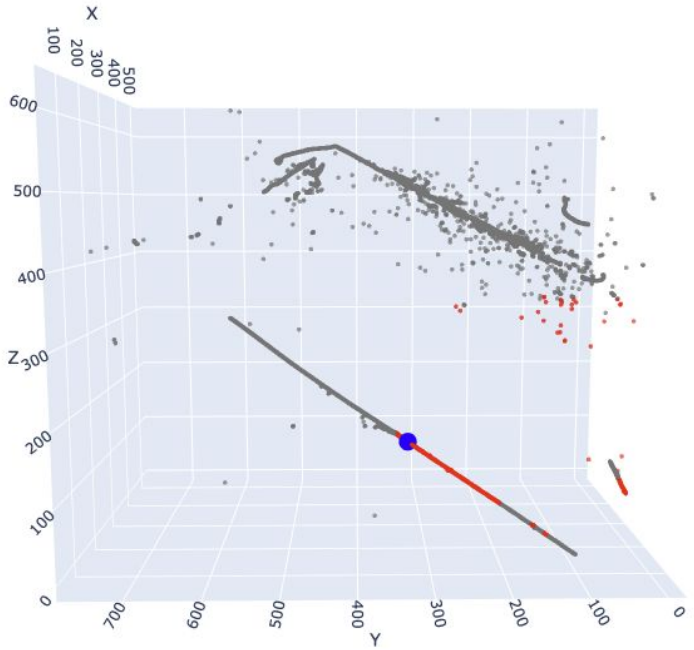
From https://xywu.me/talk/20250609_Stanford_Spatial_Representations.pdf

Serialized Attention – 256 voxel patches

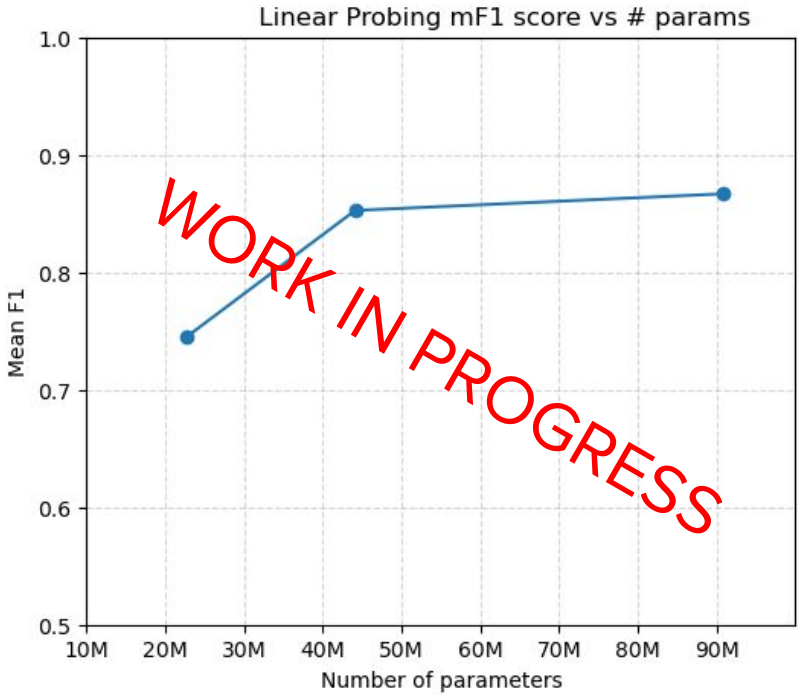
Receptive field at a single point at stage 0

- not perfect, but good enough with enough depth.





Scaling Model Params



Pretraining techniques

- Self-supervised learning
- Supervised learning

INVARIANCE & RECONSTRUCTION

- **Masked reconstruction** 1D · 2D · 3D · 1D+2D+3D
  Recover masked pixels, tokens, or local geometry from partial observations.
  MAE, BEiT, iBOT, Point-BERT, Point-MAE, PoLAr-MAE, Hiera / Point-M2AE, 4M
- **Non-contrastive alignment** 1D · 2D · 3D
  Teacher-student or twin-view agreement, usually without explicit negatives.
  BYOL, SimSiam, DINO + DINOv2, Sonata, Panda
- **Predictive latent modeling** 2D
  Predict semantic latent targets rather than reconstructing raw pixels or prototype distributions
  JEPA / LeJEPA / VICReg, data2vec, BYOL, point2vec
- **Contrastive alignment** 2D · 3D
  Bring two views of the same sample together and push different samples apart.
  SimCLR, MoCo, PointContrast, "Contrastive Learning for Robust Representations of Neutrino Data"

Research challenges

- Our data is large (1E4 to 1E8 entities)
- Different in nature (e.g. density, manifold)

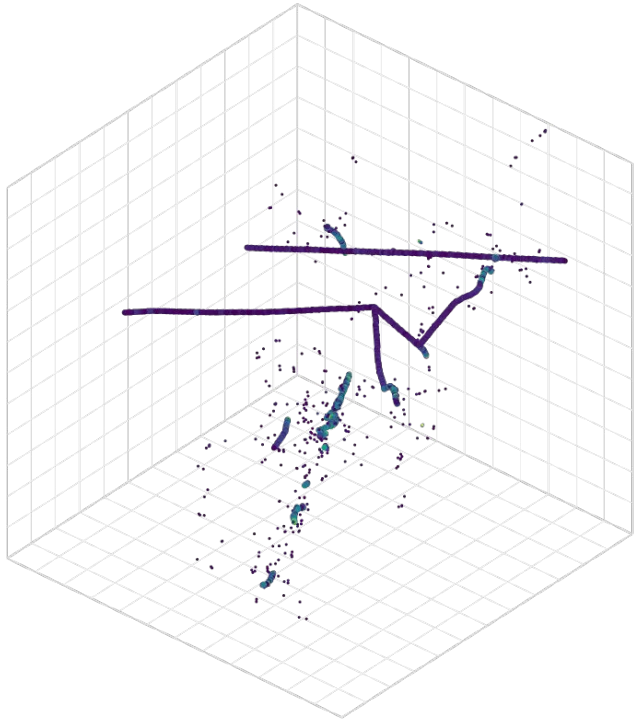
GENERATIVE, GEOMETRIC & CROSS-MODAL

- **Autoregressive token modeling** 2D
  Learn by next-token prediction over image tokens or rasterized sequences.
  PointGPT, PointMamba, FM4NPP, NEPA
- **Generative: denoising / flow** 2D · multimodal
  Learn representations through denoising, diffusion, or flow matching objectives.
  Diffusion-based repr. learning, Self-Flow, VAE, "Score-based Diffusion Models for Generating LArTPC Images"
- **Geometry & multi-view 3D** 2D→3D · 3D
  Use view consistency or scene geometry as the supervisory signal.
  VGGT, Dust3r/Mast3r
- **Cross-modal alignment** any modality (+text)
  Align representations across language, images, and point clouds.
  CLIP, Astro-CLIP, Perception Encoder, SigLIP, Concerto/Utonia
- **Labeled + multi-task supervision** supervised
  General-purpose encoders trained on broad label spaces, segmentation masks, or task mixtures.
  BiT / JFT, SAM, task mixtures, OmniLearn, L-GAtr

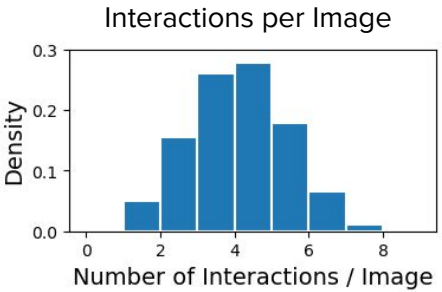
Dataset: PILArNet-Medium

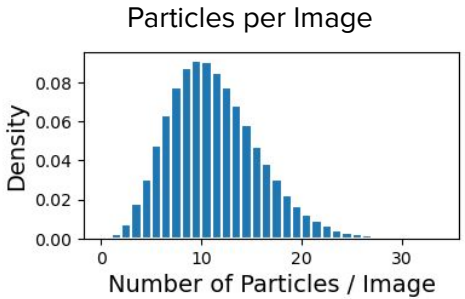
Open data!

- Simulated dataset of 1.2M 3D events
- $(2.3\ \text{m})^3$ cube $(768\ \text{px})^3$. ~5B non-zero voxels.
- +1M events on top of previous open dataset, PILArNet (2020).
- Simply 3D energy depositions, equivalent to “digital hits” from a LArTPC (e.g., DUNE Near Detector)
- 1024 - 30,000 voxels/event



Raw Depositions

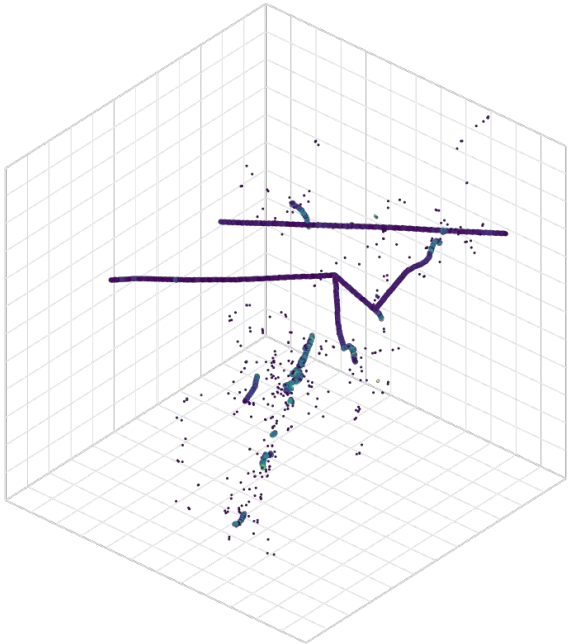




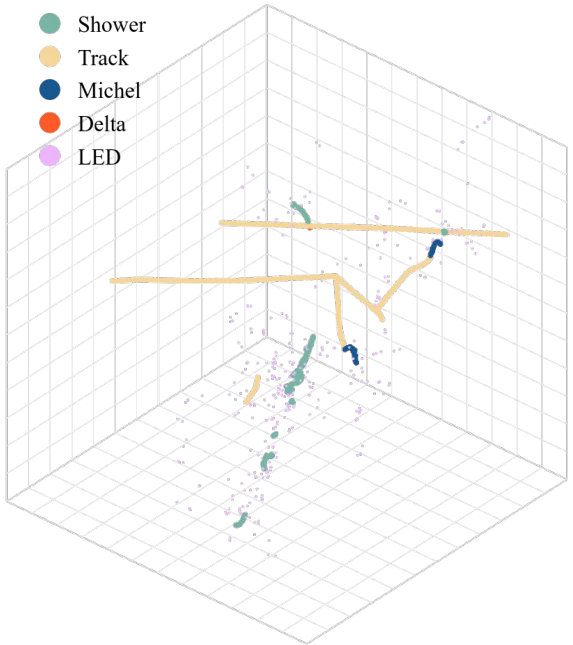
Young, S. et. al, *Particle Trajectory Representation Learning with Masked Point Modeling* (2025)

Semantic Segmentation Labels

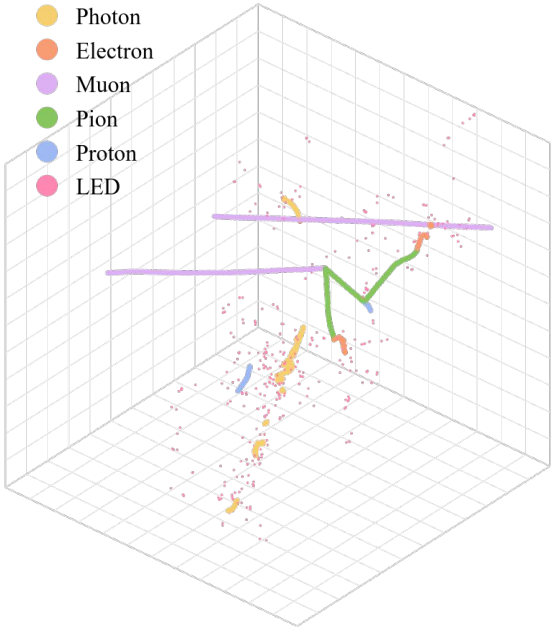
Open data!



Raw Depositions



Semantics Reconstruction

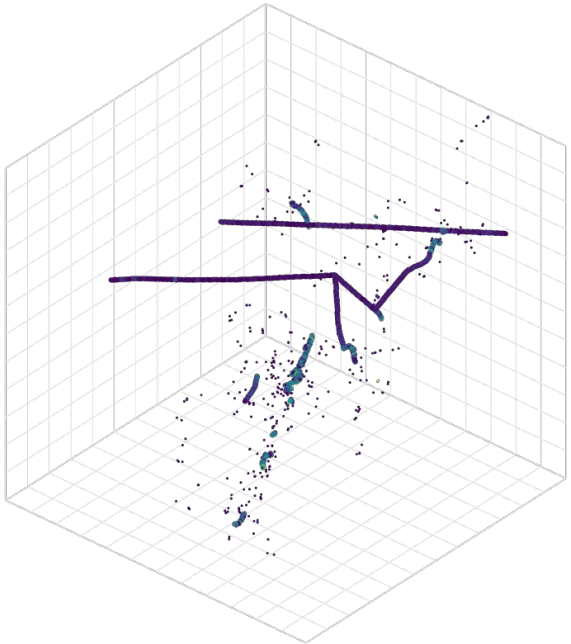


Particle ID Reconstruction

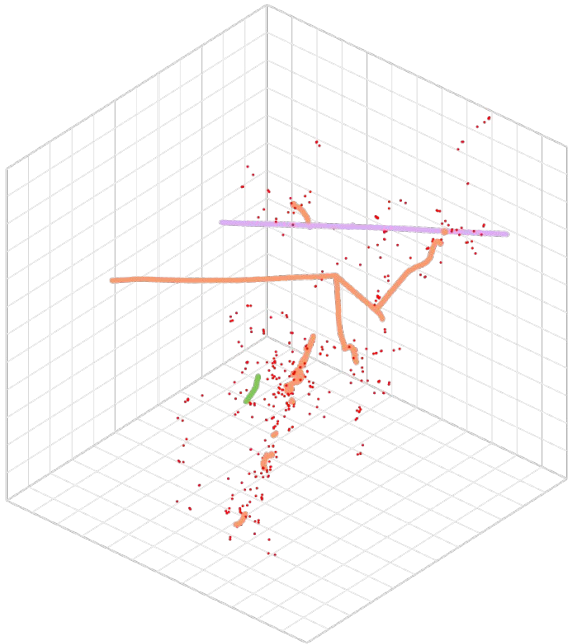
Young, S. et. al, *Particle Trajectory Representation Learning with Masked Point Modeling* (2025)

Instance Labels

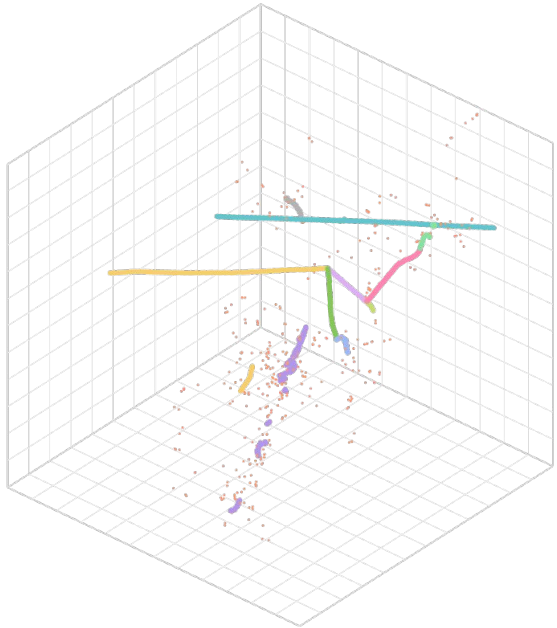
Open data!



Raw Depositions

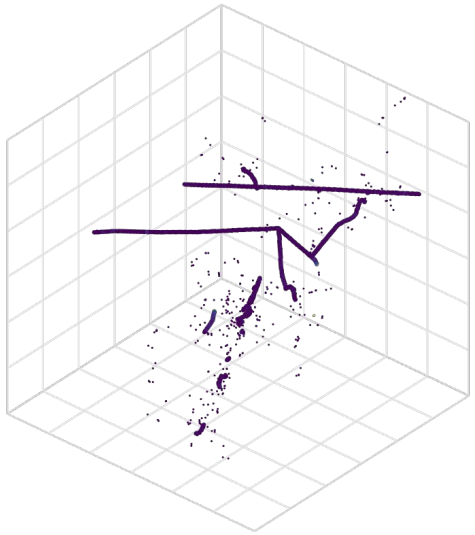


Interaction-level



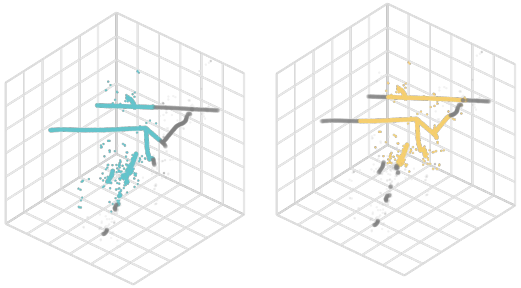
Particle-level

Young, S. et. al, *Particle Trajectory Representation Learning with Masked Point Modeling* (2025)

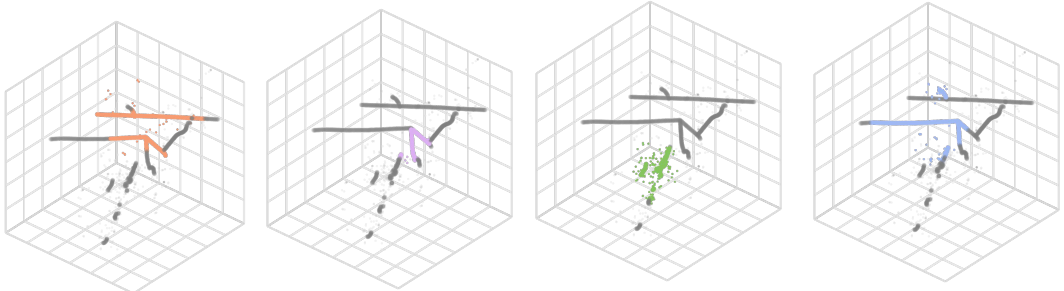


Full Image

Global Views



Local Views



Masked Views