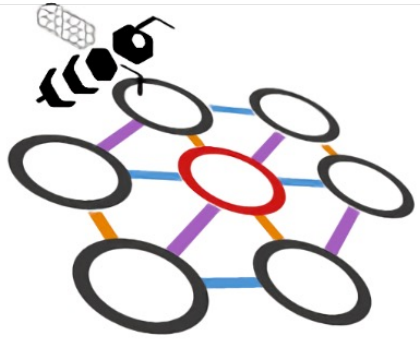
Brookhaven National Laboratory

Development of a Self-Supervised Foundation Model for Wire-Cell

Matteo Vicenzi (mvicenzi@bnl.gov) and Nitish Nayak

on behalf of the BNL Wire-Cell group

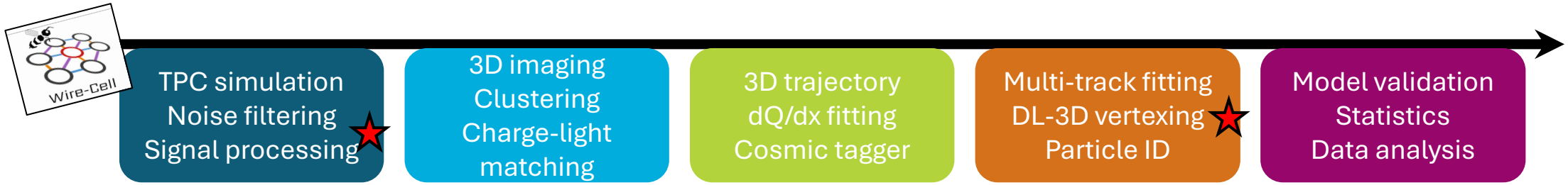
June 18th, 2026

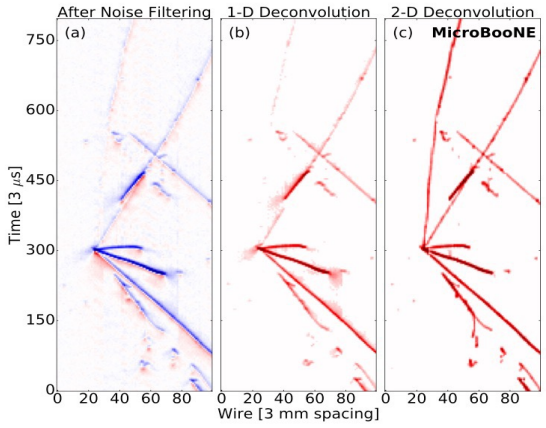


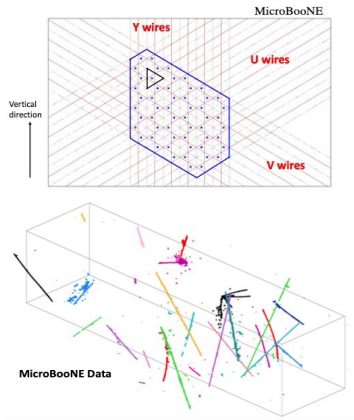
Wire-Cell

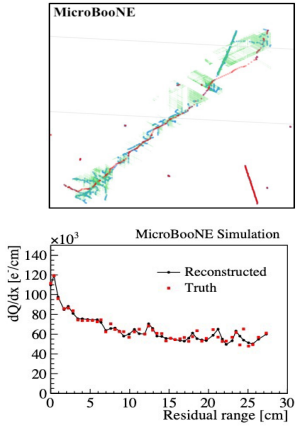
Wire-Cell is an **end-to-end reconstruction toolkit** for LArTPCs starting from raw signals up to full 3D reconstruction (particle flow).

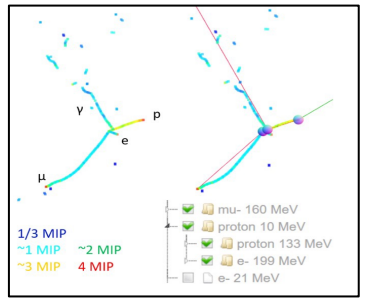
- Many success stories in uBooNE, now moving to (Proto)DUNE, SBN, …

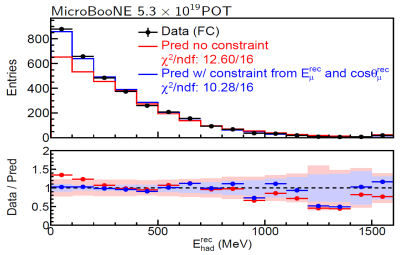












https://lar.bnl.gov/wire-cell/

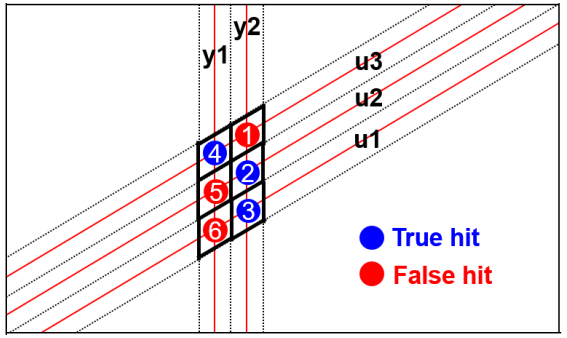
★ Already incorporates traditional and ML algorithms (DNN-ROI, vertexing)

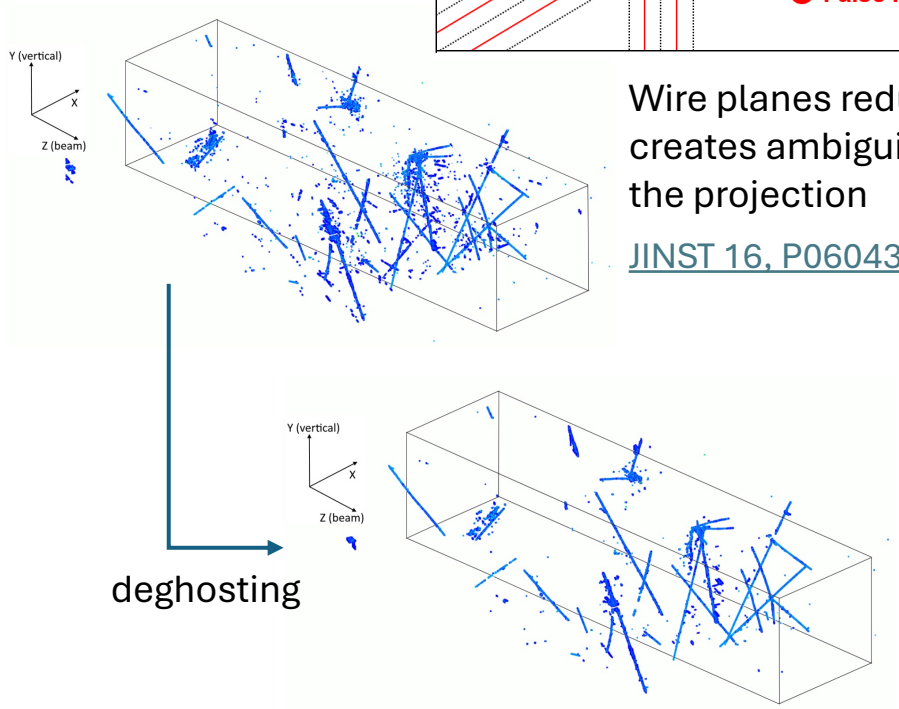
Why a Foundation Model?

Wire-Cell imaging currently exploits: time/wire geometry + charge + local connectivity information.

Shortcomings:

- Charge gaps and **projection ambiguities** are not fully addressed.
- **Event topology** information is only partially leveraged by connectivity.
- Current ML in Wire-Cell is developed on simulation: biased by underlying **simulation-to-data differences**.



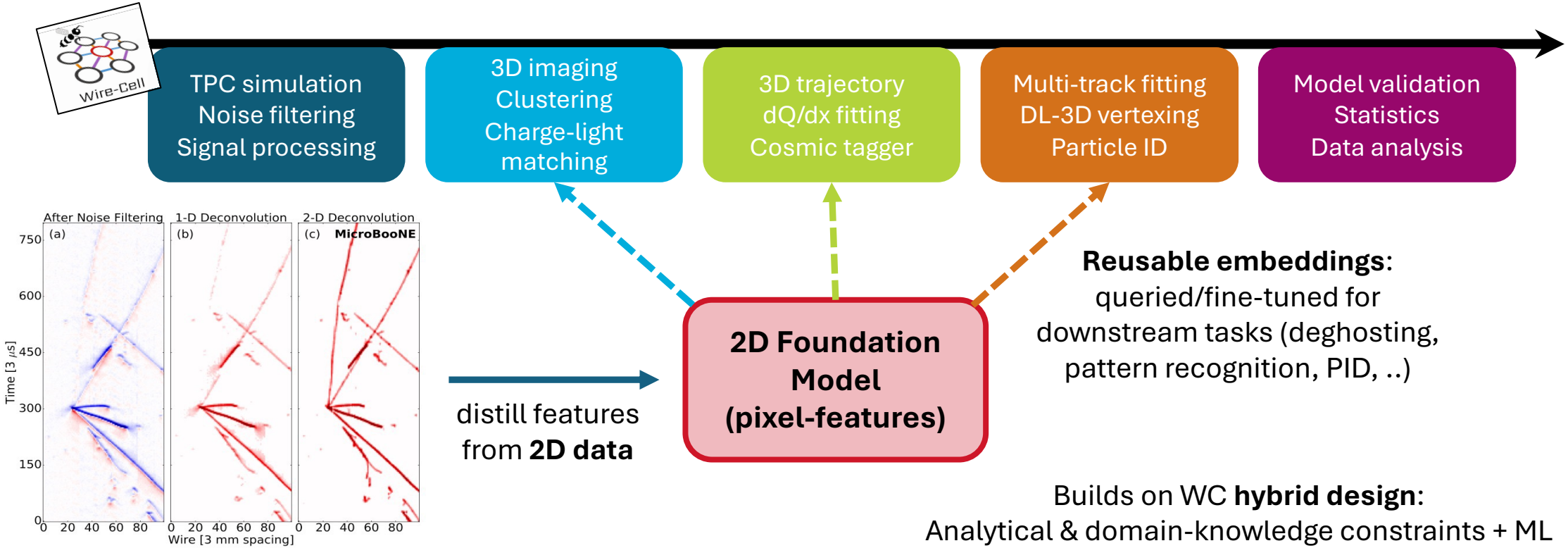


Wire planes redundancy creates ambiguities in the projection

JINST 16, P06043

Wire-Cell 2D FM

- Naturally closes the **MC/data gap** → trained directly on real data
- Encodes **global topology/structure** previously only partially leveraged



Designing the architecture

Powered by
WarpConvNet

Pixel-level features

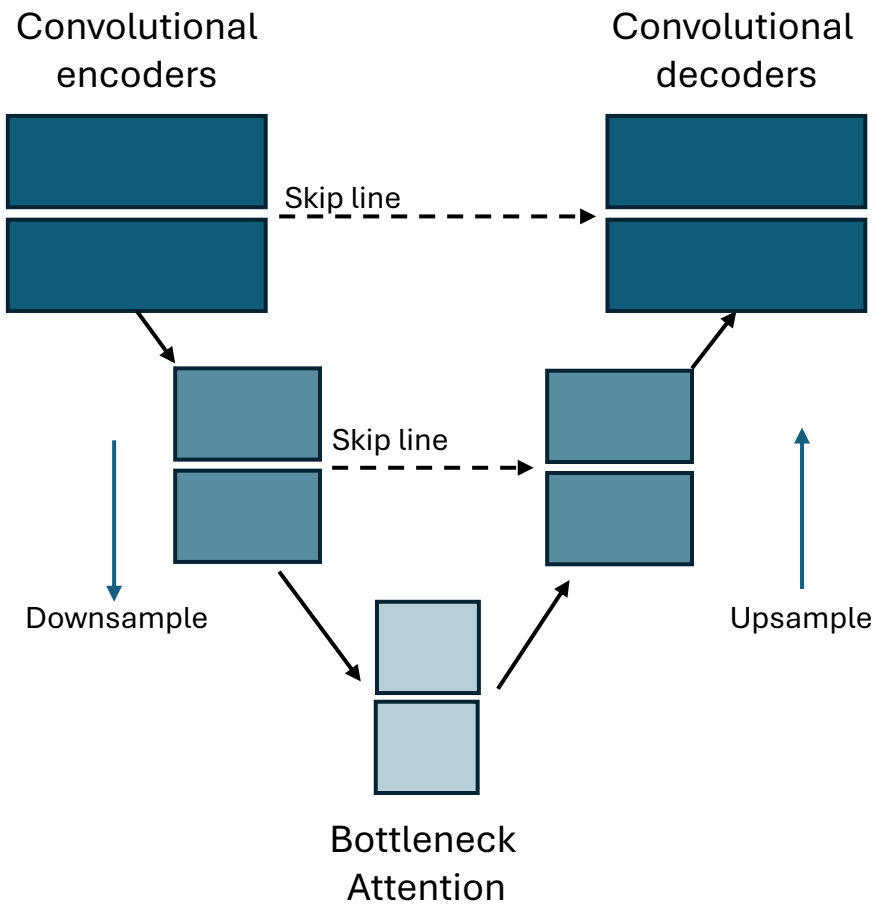
→ MinkUNet with attention bottleneck

- Need multi-scale context: **local *and* global** topology both matter!
- U-Net skip connections preserve fine-grained spatial details + bottleneck attention for global relationships

Sparse data

→ WarpConvNet framework

- LArTPC images: only ~0.2% of pixels carry signal (~185x memory reduction vs. dense)
- WarpConvNet is a recent **CUDA-optimized** library for **sparse convolutions** (C. Choy et al., NVLabs, 2025)



Self-distillation

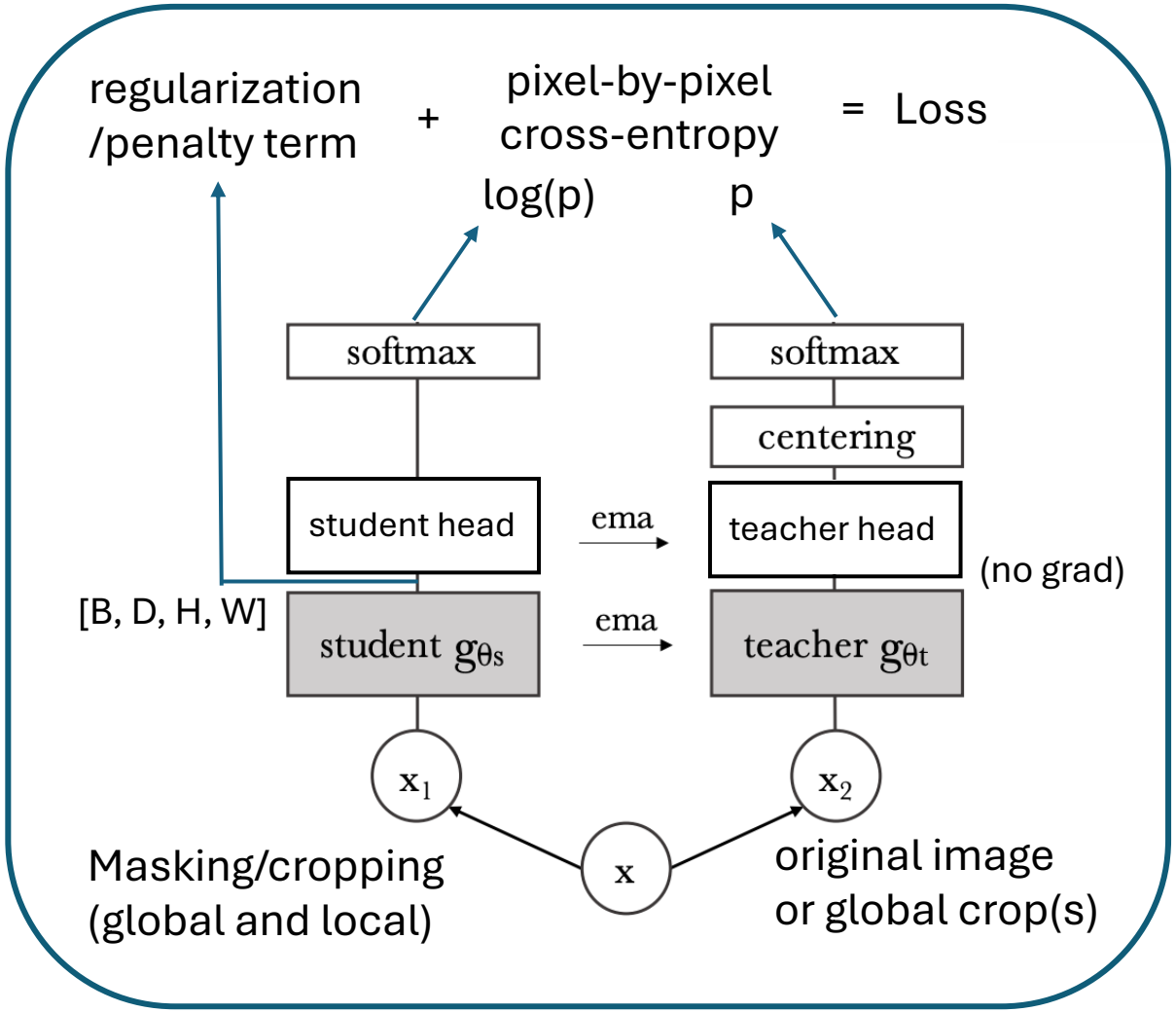
Training strategy based on **DINO:** self-supervised, no labels.

Two network, same backbone architecture, different purpose:

- Student:
  - sees **local + global** augmentations
  - tries to **predict** the teacher's output: pixel-by-pixel cross-entropy loss
- Teacher:
  - sees **only global** context
  - provides **~stable but moving target**
  - slowly updated from student (~1% student influence per step)

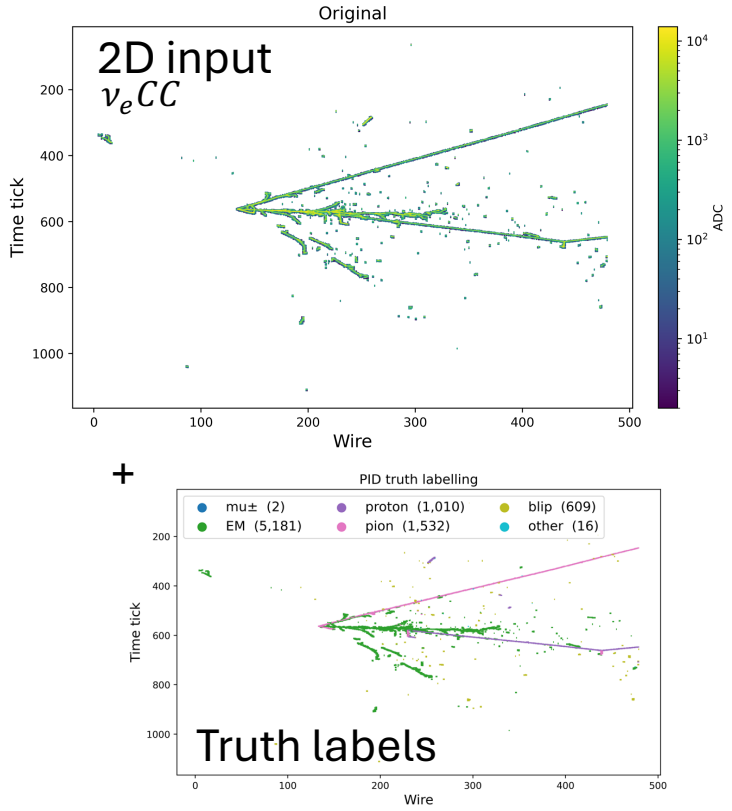
DINO = DIstillation with **NO** labels

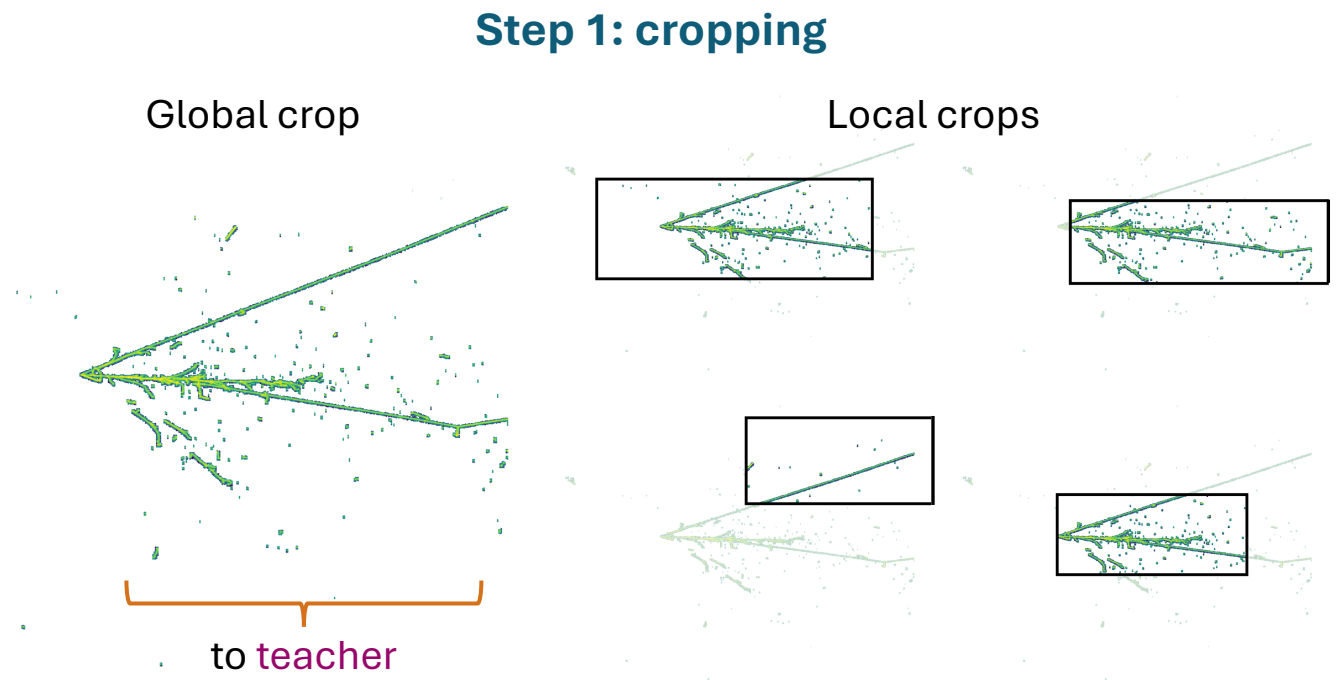
arXiv:2104.14294, arXiv:2304.07193



Data augmentation

For model development: **simulated ν interactions** in Wire-Cell with DUNE-like geometry (single APA, collection view) with **pixel-level truth** labelling.

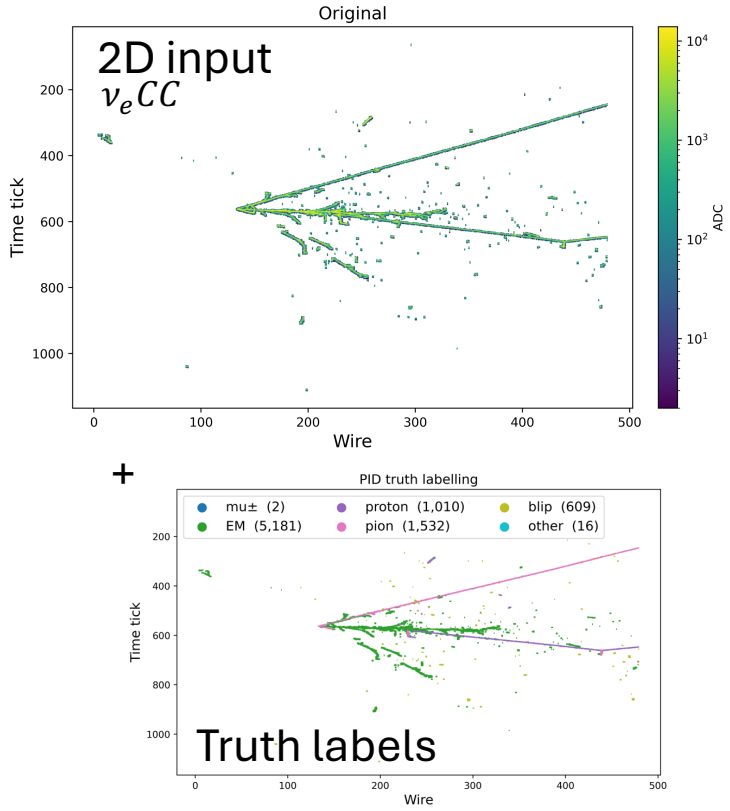


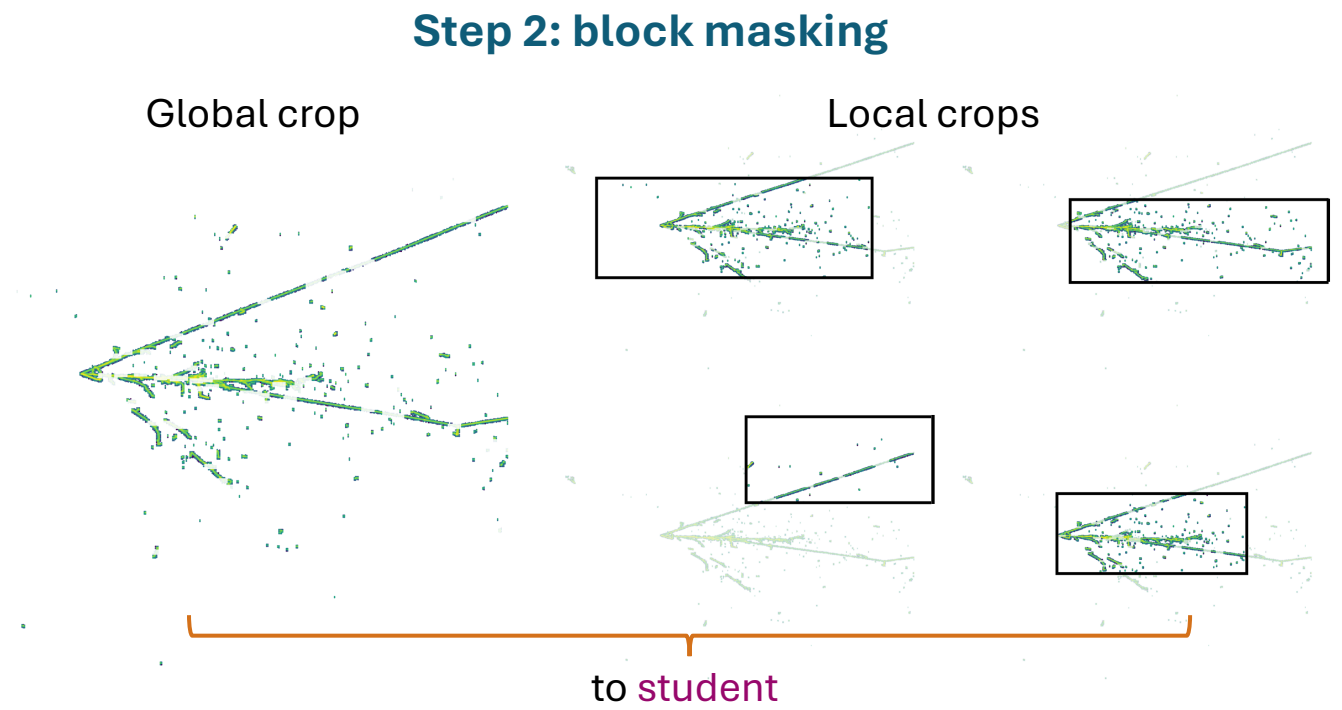


Seeded by Gaussian-blurred activity (**anchor sampling**), then sized via scale ranges. Activity-aware but currently geometry-agnostic → aiming to replace it with **physics-motivated** crops (e.g., vertices).

Data augmentation

For model development: **simulated ν interactions** in Wire-Cell with DUNE-like geometry (single APA, collection view) with **pixel-level truth** labelling.

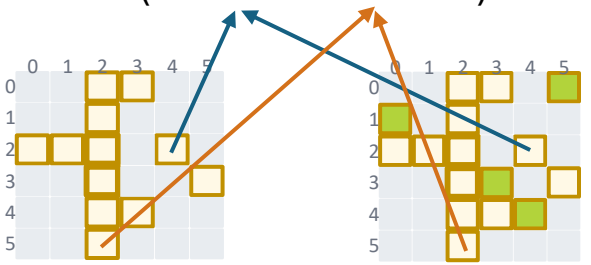




K block centers sampled randomly from active voxels (until mask ratio is reached), masked block is 10x10 window

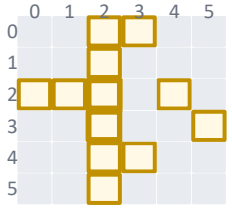
Loss objective

DINO objective:
Pixel-by-pixel cross-entropy loss
(where both active)

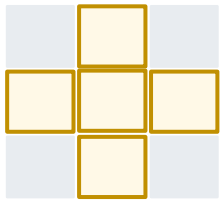


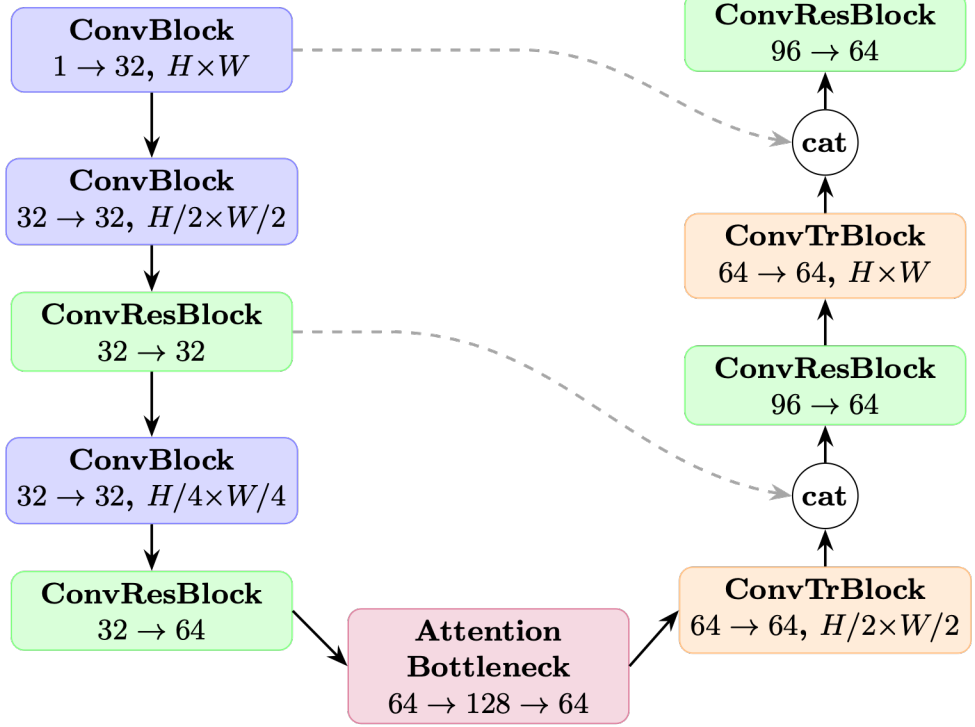
Teacher sees masked positions as well

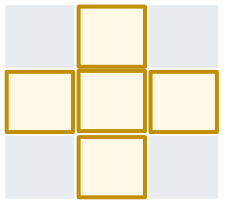
Active pixel in patch post-masking



Down-sampling lowers resolution



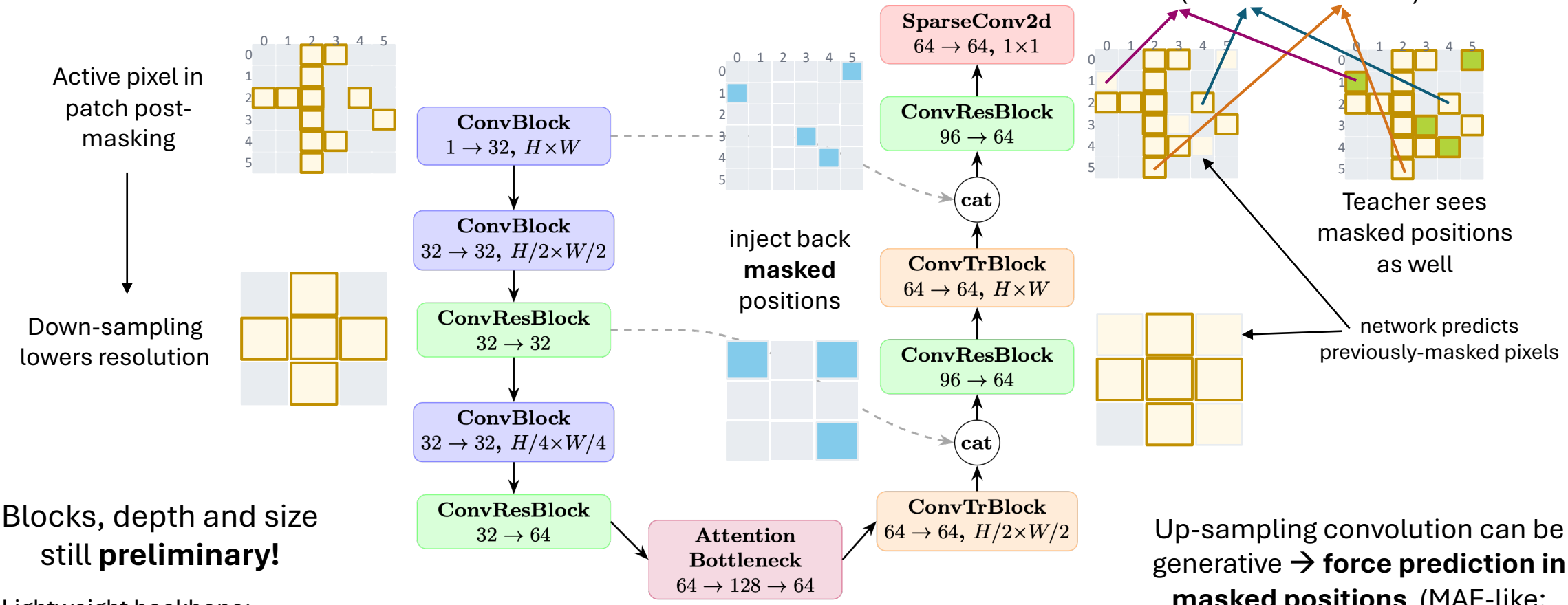




Blocks, depth and size still **preliminary!**

Lightweight backbone:
477k params (~1.8 MB)

Loss objective



Blocks, depth and size still **preliminary!**

Lightweight backbone:
477k params (~1.8 MB)

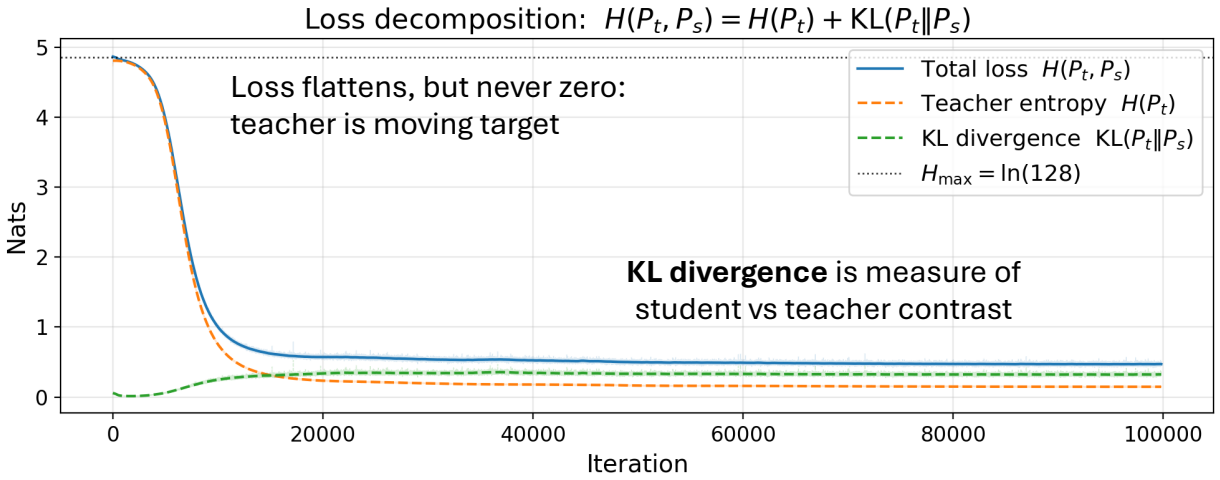
Up-sampling convolution can be generative → **force prediction in masked positions** (MAE-like: mask autoencoder)

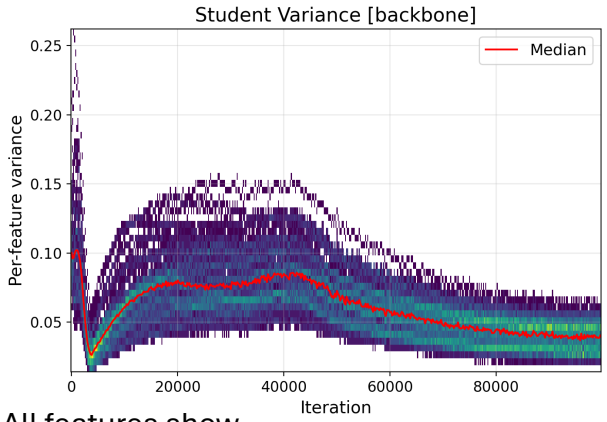
Training stability

Performed **small-scale** test trainings: ~100k images, O(50) batch size, O(100) epochs

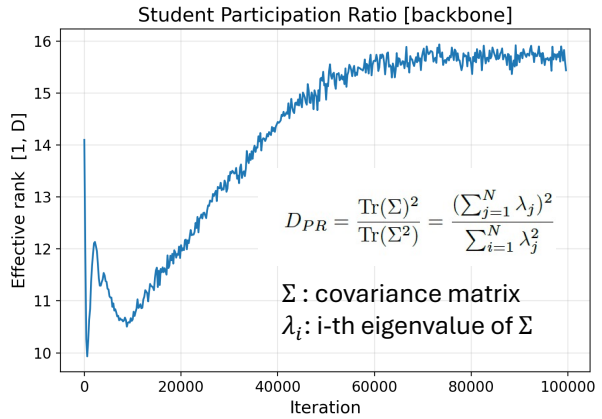
Preliminary validation only – no production scale yet!

- Loss converges cleanly
- Decomposition shows healthy student-teacher (KL) divergence
- **No dimensional collapse,** non-trivial for DINO-like training with pixel-level loss





All features show non-zero variances



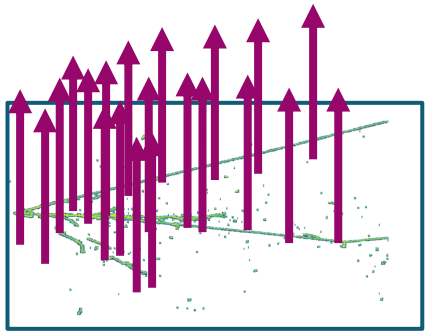
Participation ratio, ie how many feature-space dimensions are active

Performance evaluation

Validation: are the learned features **meaningful**?

- Simple approach is **k-NN clustering** (nearest neighbor) → probes feature space directly without requiring additional supervised training.

Run images through frozen network



per-pixel tokens
+ truth class (PID)

.. or pooling all pixel tokens in the image:

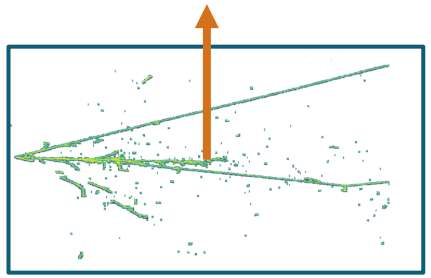
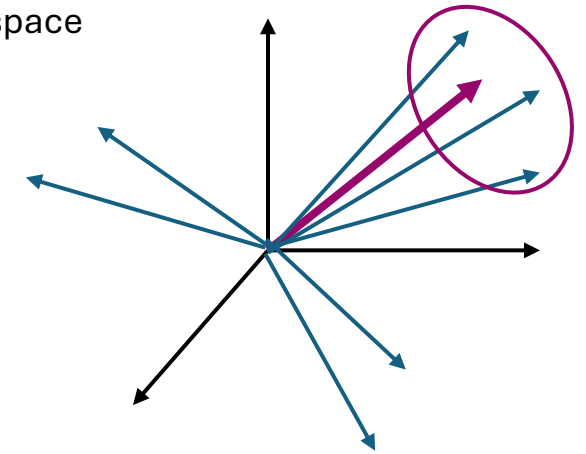


image-level token
+ truth class (event ID)

Check neighbors in feature space

D-dim space



cluster of k-closest neighbors (cosine similarity)

Use cluster for **prediction**: predicted class is most-common class in cluster → check against truth.

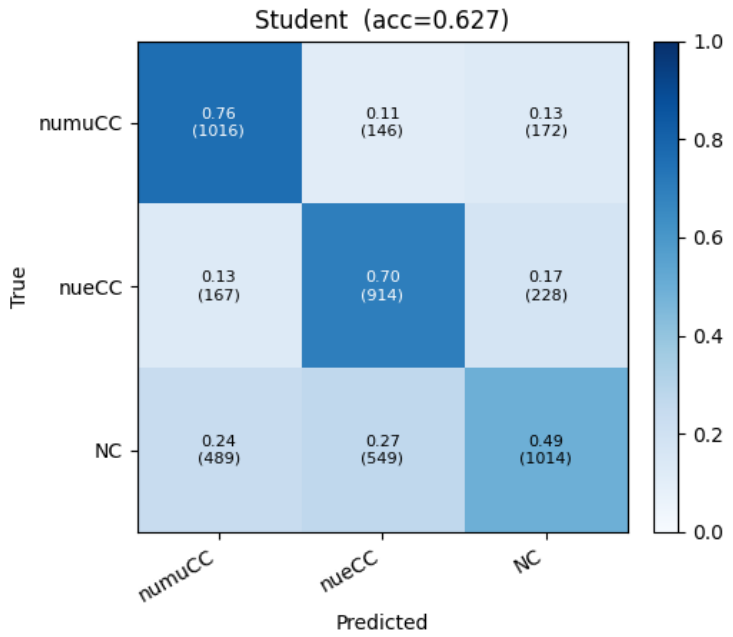
k-NN classification

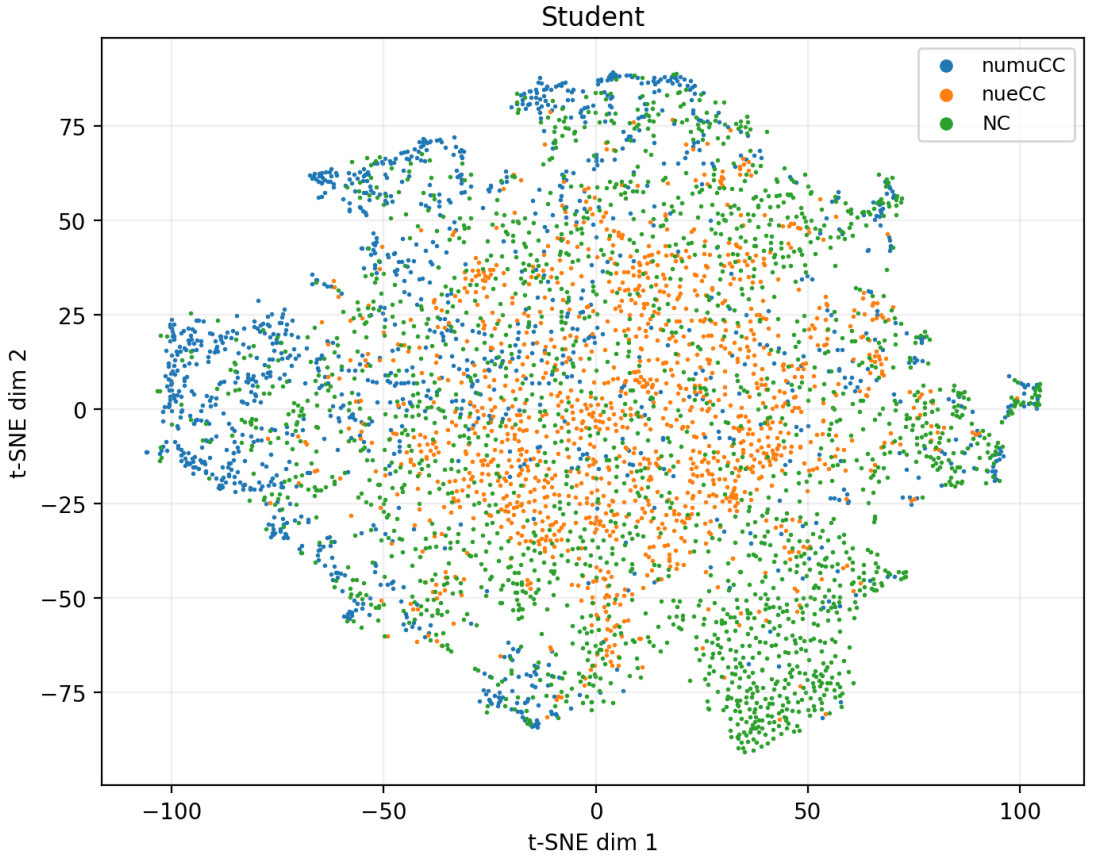
Preliminary results from small-scale trainings

k-NN clustering of **image-level** token → predict event ID

This is not the target of the foundation model → **sanity check**

k=5, epoch=10





t-SNE projection of the 64-D feature space shows hints of separation

k-NN classification

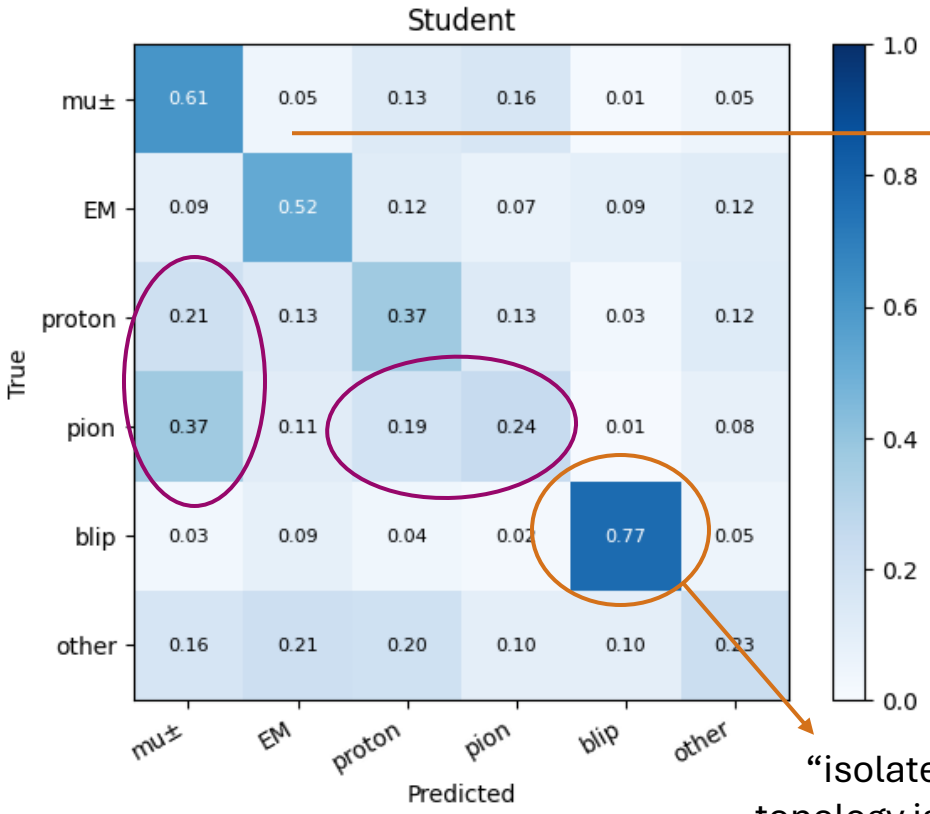
Preliminary results from small-scale trainings

Training dataset:
100k images, DUNE-like
single APA, 1050 (w) x 1500 (t)

k-NN clustering of **per-pixel** tokens → predict pixel-level PID

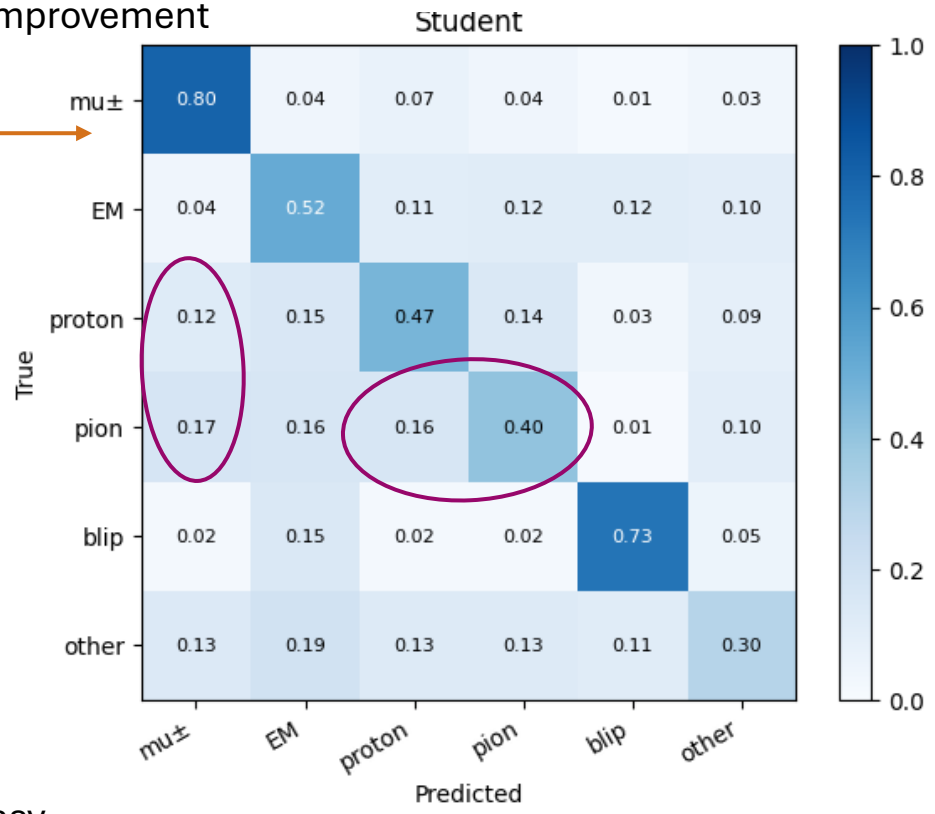
Pixel-level truth labels still being **improved**

k=5, epoch=10



strong improvement

k=5, epoch=100



“isolated” topology is easy

Summary

We are developing a **2D Foundation Model** for Wire-Cell:

- No MC dependence: **DINO self-distillation** trains directly on data, naturally closing the simulation-to-data gap
- **The machinery works:** training dynamics are stable and healthy.
- **Very early but promising**: extracted pixel-level features encouraging from small-scale tests.

→ **Model still evolving**: further optimizations, full-scale training, physics-motivated augmentations, and downstream tasks ahead!

Thank you!