INITIAL MACHINE LEARNING STUDY

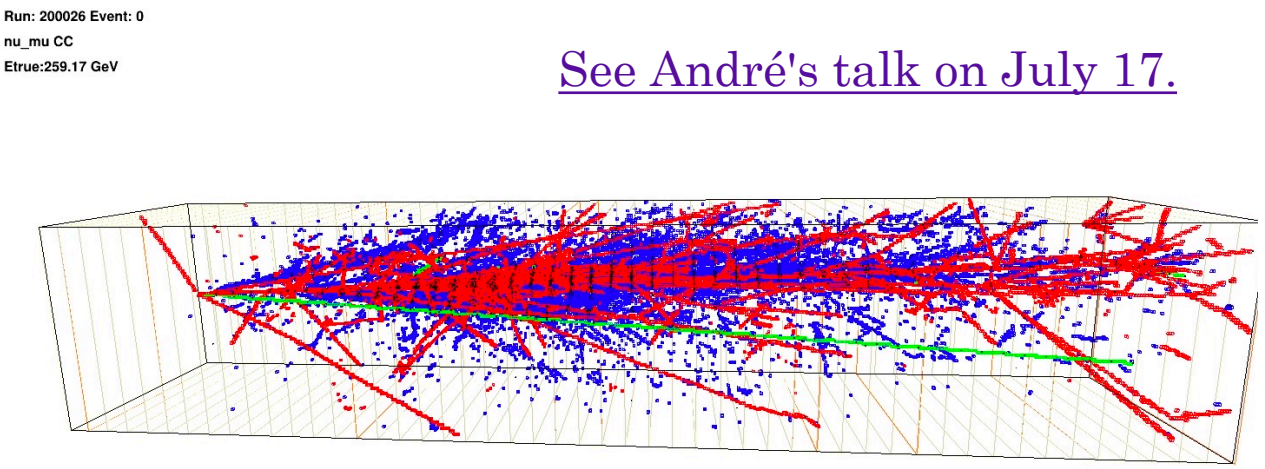
Saúl Alonso-Monsalve

ETHZ FASER meeting

7 August 2024

Dataset

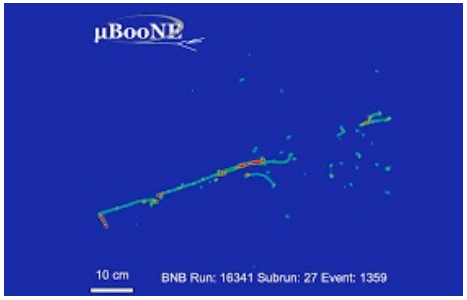
- André's simulation:
  - FASERCAL, code at: https://github.com/rubbiaa/FASER.
  - ~60K events (~100 GB).
- Wrote a Python script to retrieve the events and convert them into NumPy arrays:
  - Code and instructions: https://github.com/rubbiaa/FASER/tree/main/Python_io.
- **Goal**: use ML for the reconstruction (electromagnetic vs hadronic for now).
  - Implemented a simple net to test the feasibility of the method.
  - Keeping (per hit): track id, parent id , primary id , pdg, position (x, y, z), energy dep.
  - Plan to continue in this direction (together with Wissal).



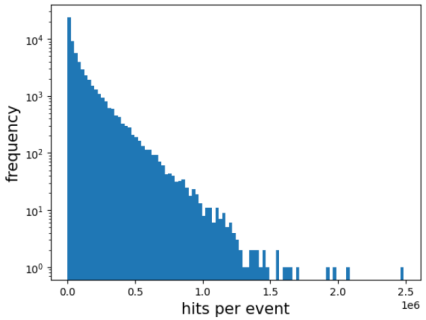
See André's talk on July 17.

Issues and considerations (I)

- Challenging ML problem:
  - **Sparse 3D images**: typical in neutrino physics.
    - If seen as "dense" images, most voxels are empty.

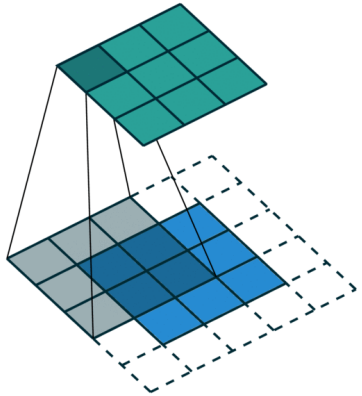


  - **Massive number of hits**.
    - Number of hits per event up to 2.5M!

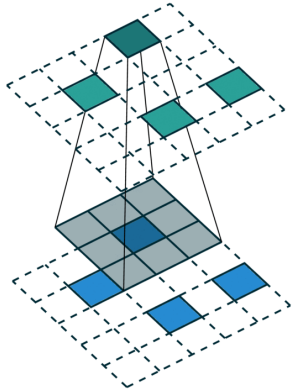


  - Solution: **Sparse Submanifold Convolutions**.
    - Efficient for sparse images (both memory and time).
    - Ignore empty voxels!
    - MinkowskiEngine implementation from NVIDIA.

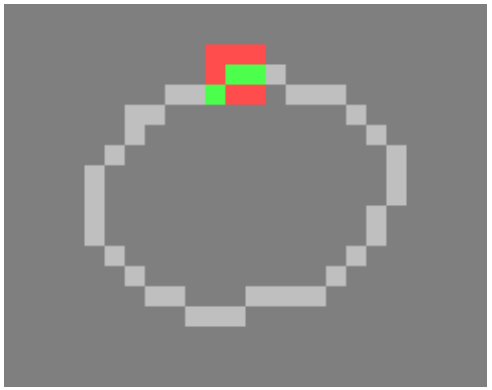
Dense convolution



Sparse convolution

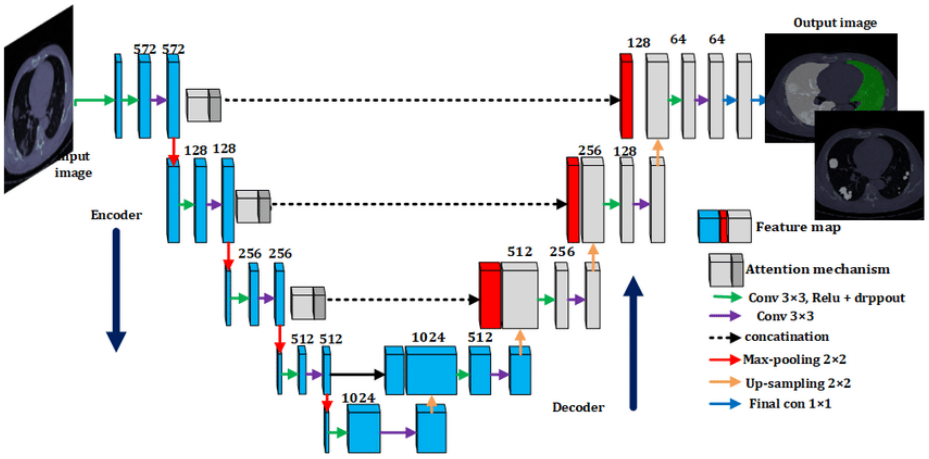


Sparse convolution



Issues and considerations (II)

- Architecture: **Sparse U-Net**.
  - U-Nets are the **state-of-the-art** for **semantic segmentation** tasks.
  - Initially proposed for biomedical image segmentation (arxiv.org/pdf/1505.04597).



- Model parameters:
  - 5 encoder layers.
  - 5 decoder layers.
  - 36M parameters.

- 60% of the events used for training.
  - 10% for validation.
  - 30% for test (results in next slides).

Issues and considerations (III)

- For each hit (voxel), the net tries to predict one of the following labels:
  - **0: muonic** (PDGs: 13, -13, 14, -14).
  - **1: electromagnetic** (PDGs: 11, -11, 22).
  - **2: hadronic** (any other PDG).

- Before running the network, events must be **voxelised**.
  - For simplicity, I chose a voxel size of 1 mm^3.
  - Some hits (with different PDG/ParentID) **share the same coordinates**.
    - For now, I'm ignoring those voxels.
    - I guess the solution is to sum their energy depositions and choose as true label the PDG from the particle that contributes the most to that hit.

- Made a cut on energy deposition to be >0.5 per hit.
  - Same as in André's code.
  - Unit?

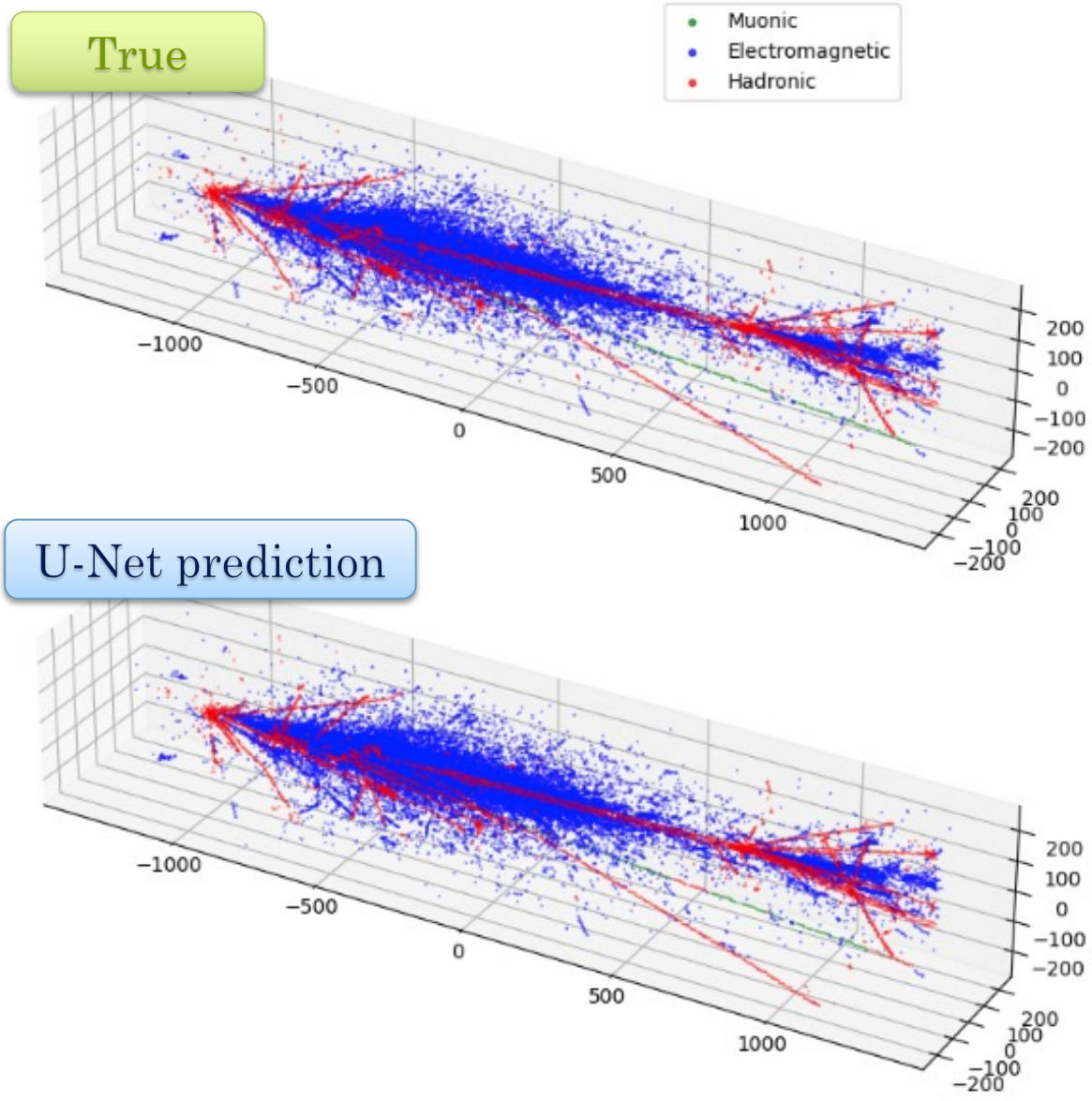
Results

- Confusion matrix:

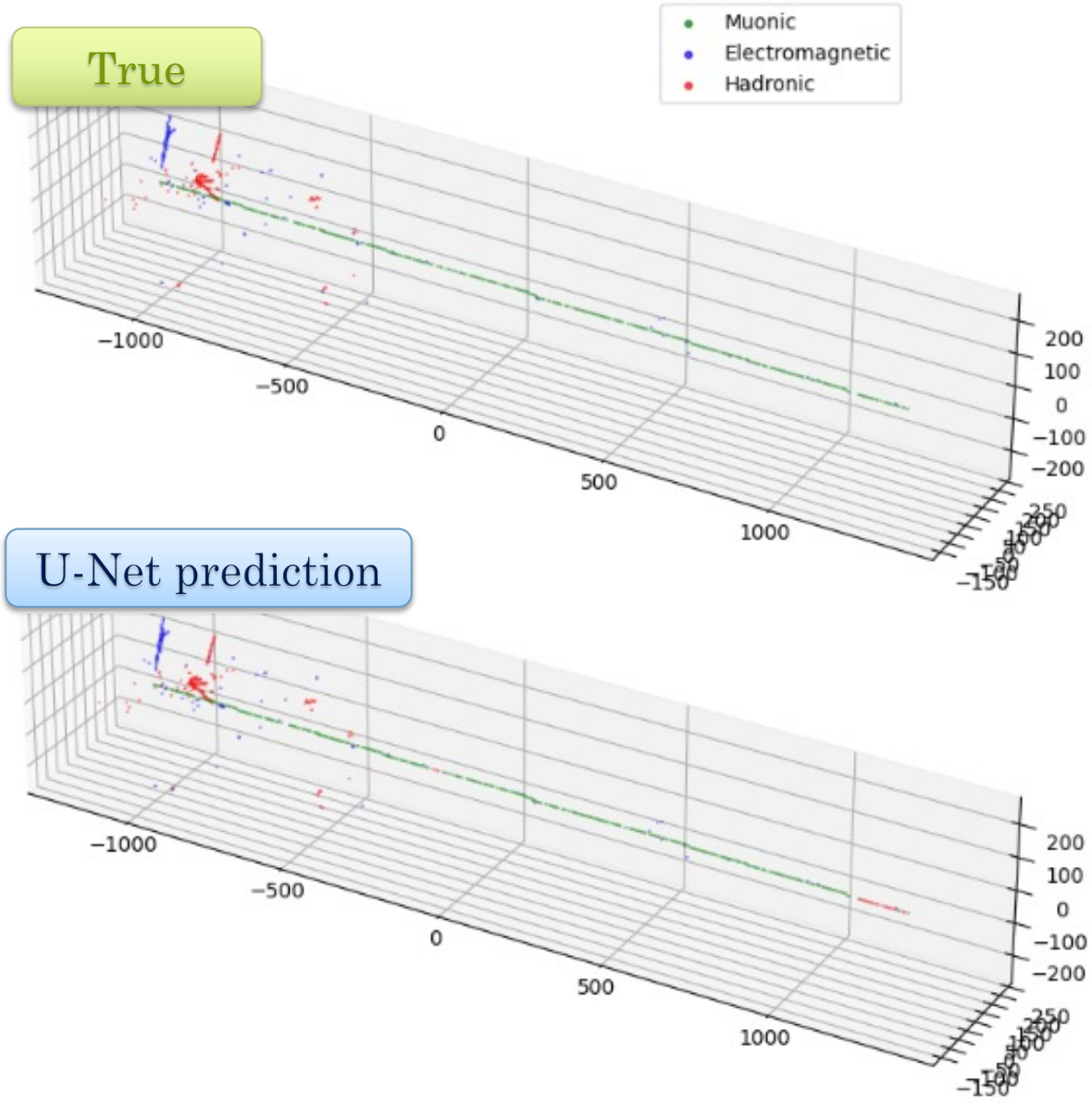
	Pred muonic	Pred electromagnetic	Pred hadronic
True muonic	**1,487,655**	133,486	1,697,436
True electromagnetic	19,424	**632,768,573**	5,758,154
True hadronic	353,480	13,661,374	**96,716,912**

- Precision (purity, cols in table):
  - Muonic: 80%.
  - Electromagnetic: 98%.
  - Hadronic: 93%.
- Recall (efficiency, rows in table):
  - Muonic: 45%.
  - Electromagnetic: 99%.
  - Hadronic: 87%.

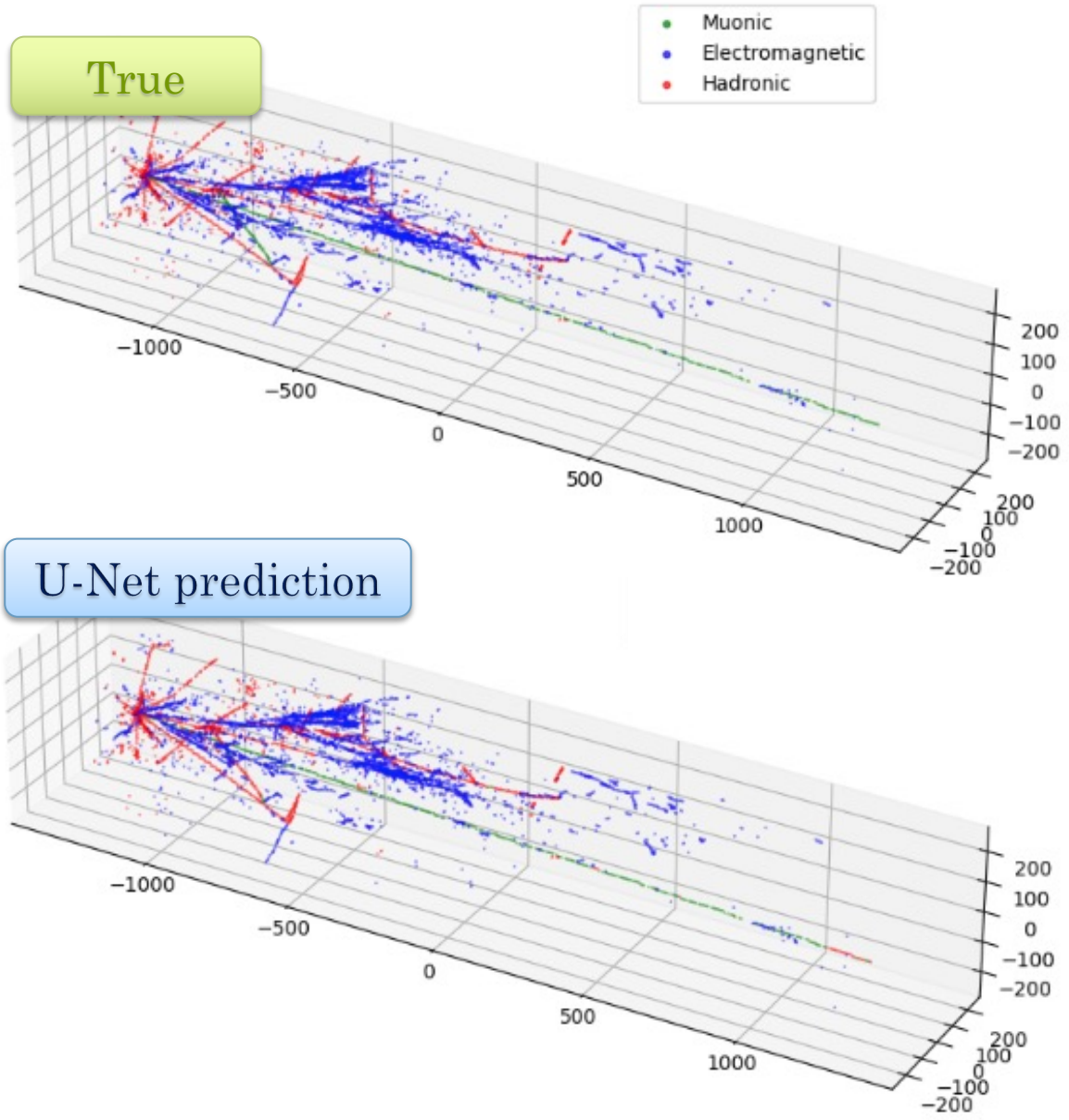
Event display (I)



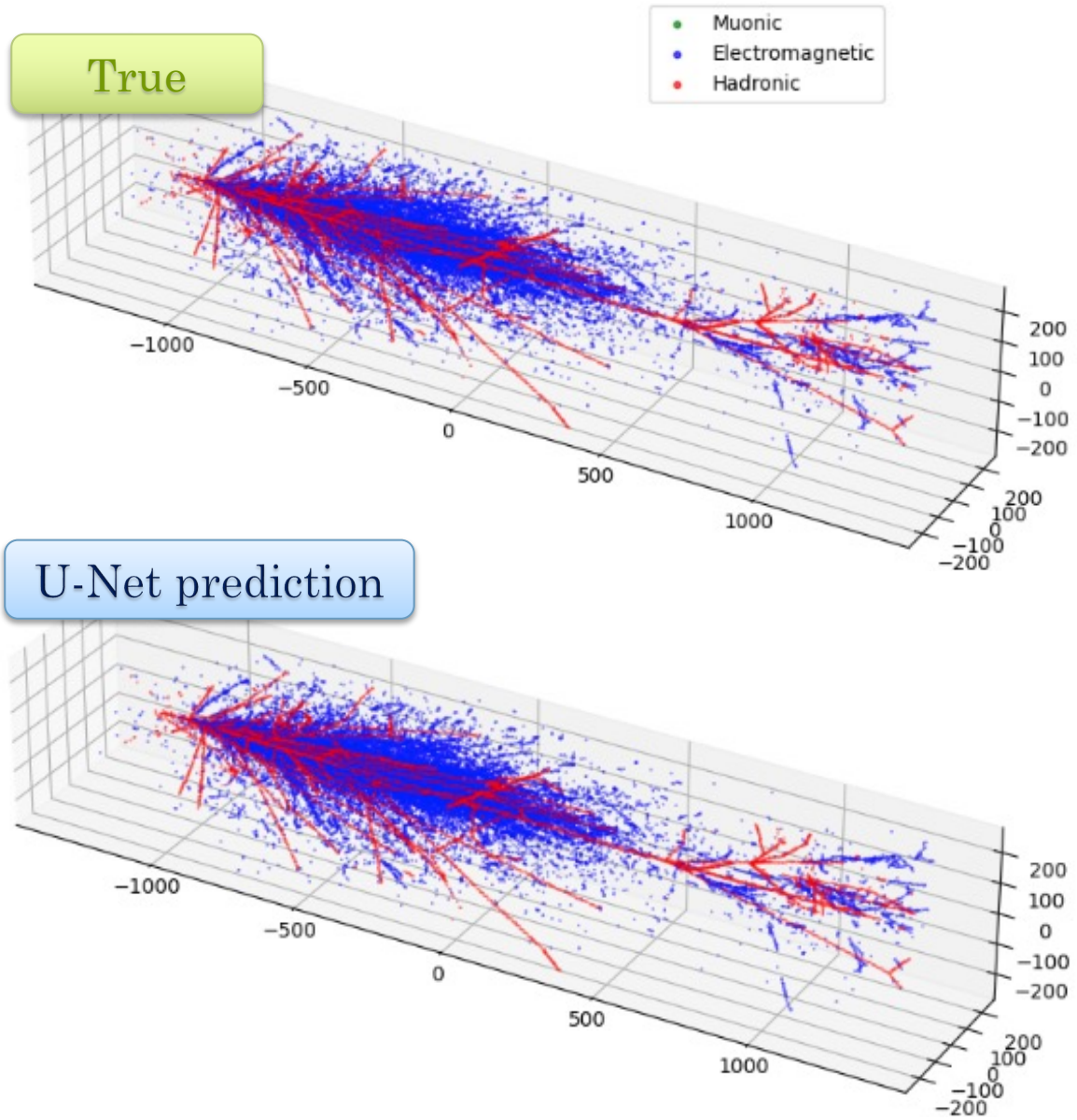
Event display (II)



Event display (III)



Event display (IV)



Summary

- First attempt to use ML in FASERCAL simulation.
- Very promising results!
  - Room for improvement.
  - Should compare to simple energy deposition cut.
- Decide where to keep the ML code.
- Next steps?

INITIAL MACHINE LEARNING STUDY

Saúl Alonso-Monsalve

ETH Zurich

ETHZ FASER meeting

7 August 2024