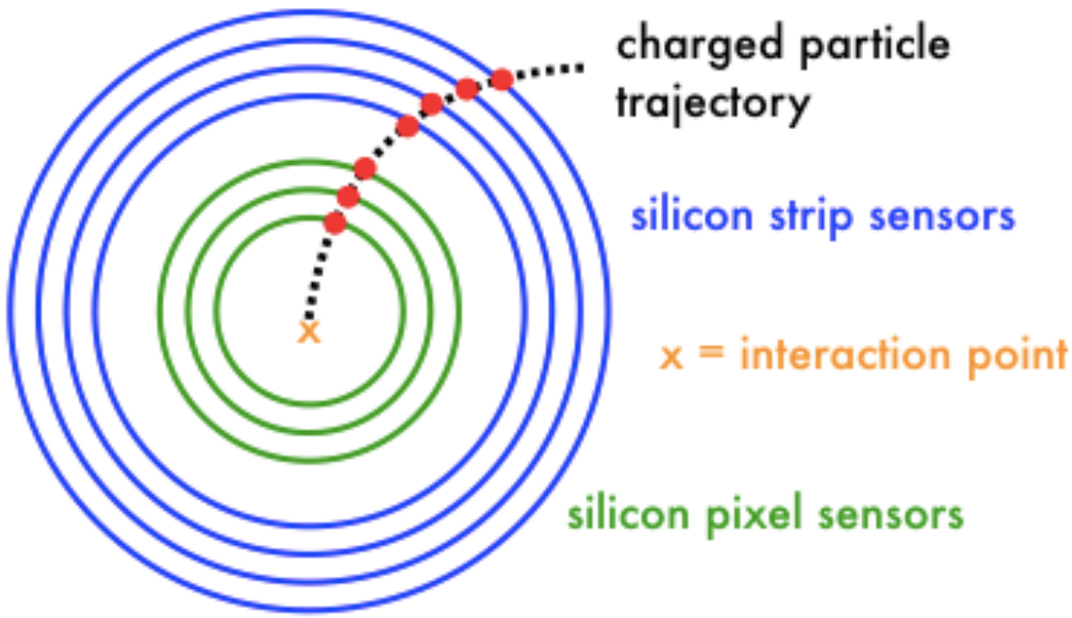
Graph Networks in CMS

Lindsey Gray

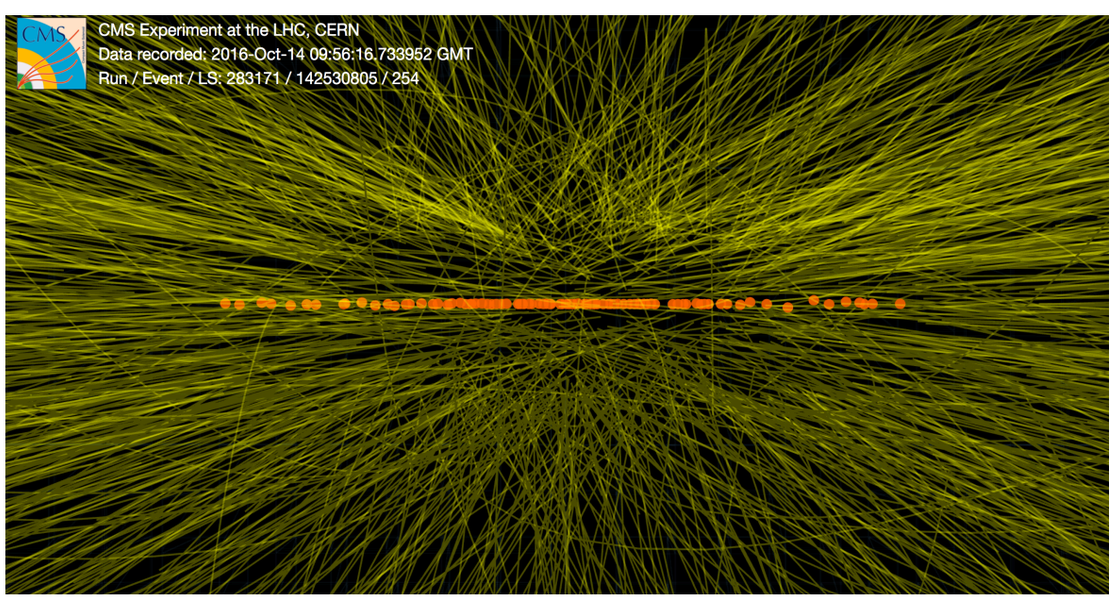
November 18, 2019

HL-LHC: Two Big Problems

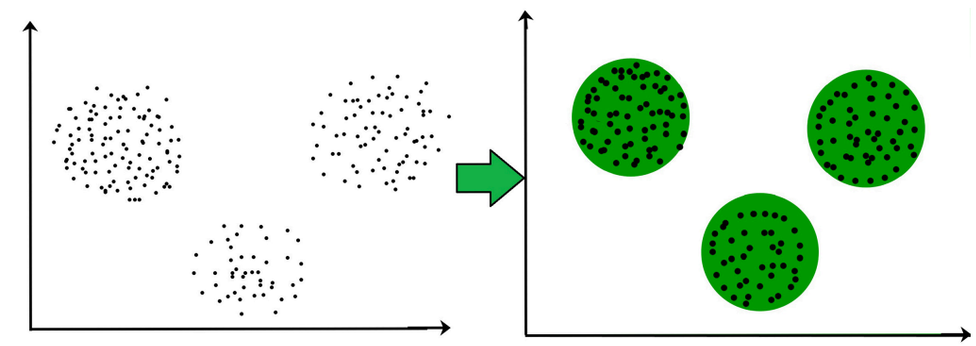
Tracking in 200 PU



1000s of tracks at once



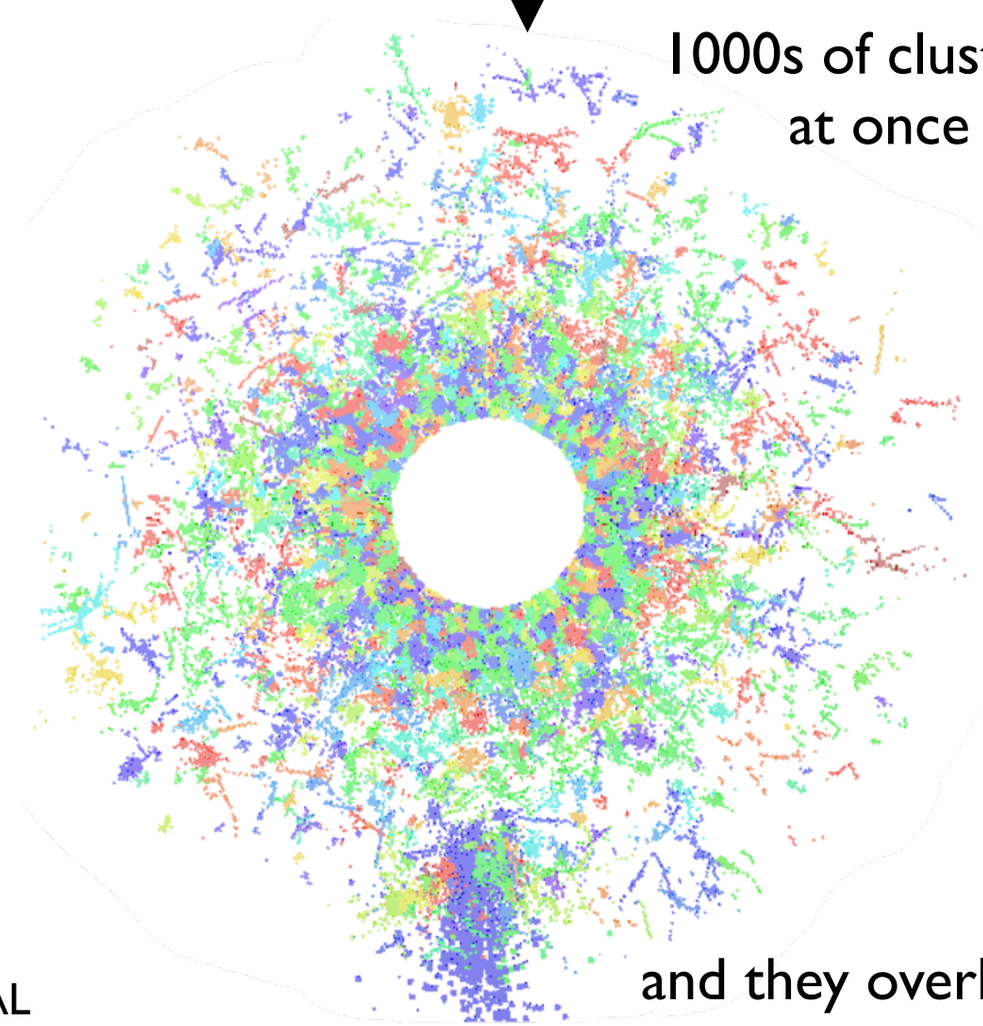
High-granularity calorimetry in 200PU



unlabelled hits

groups hits

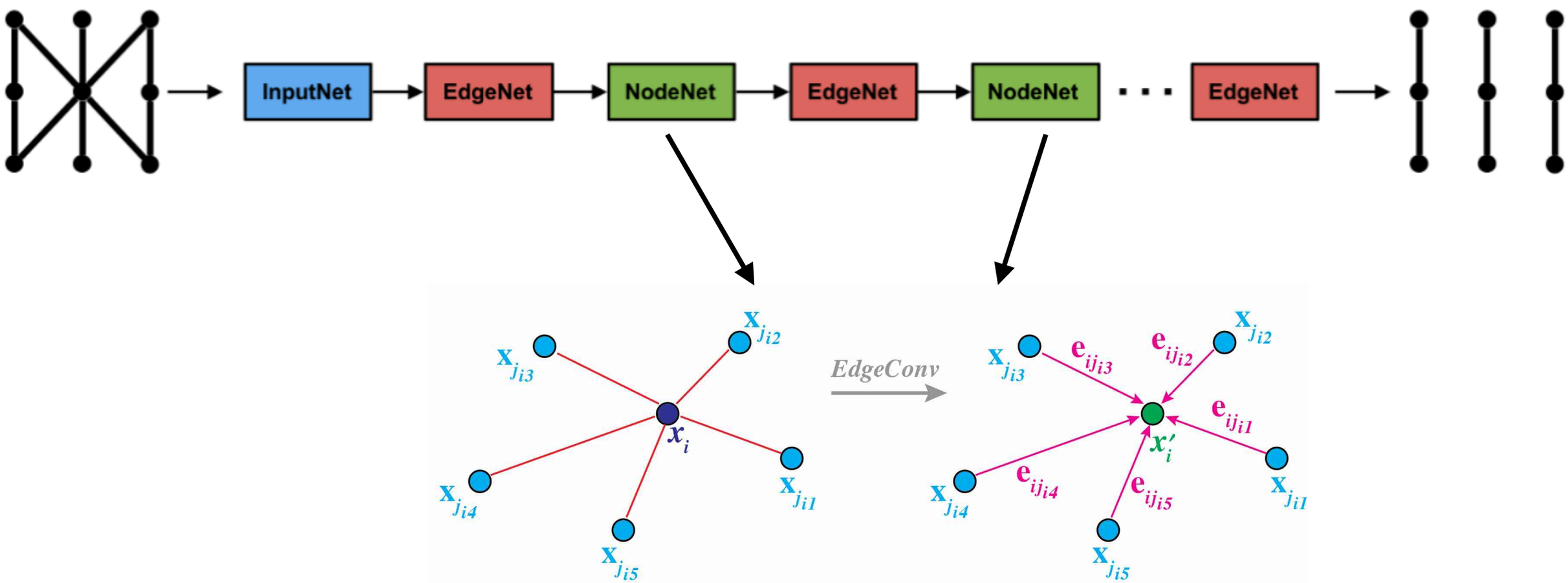
1000s of clusters at once



and they overlap :-)

A Path Forward with Hep.TrkX

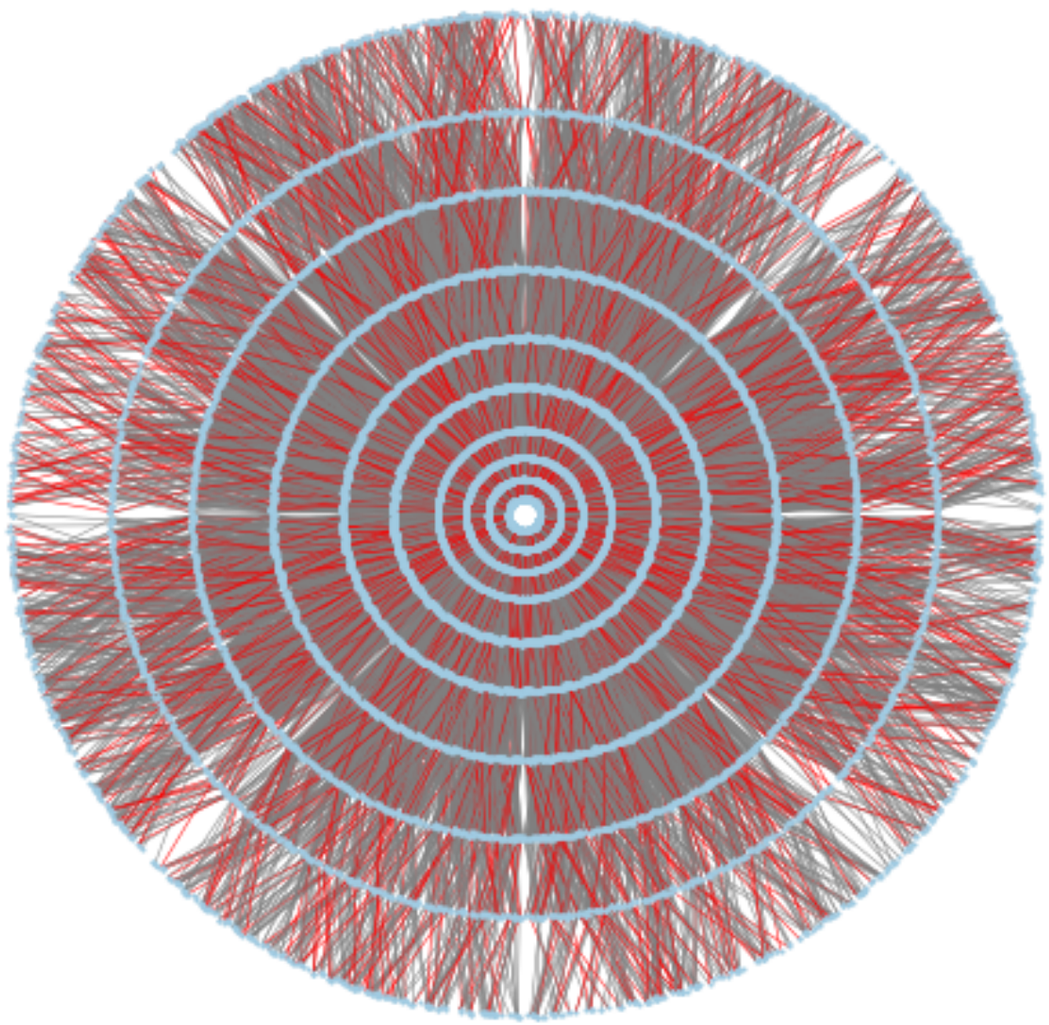
- Aim of project was to discover NN track finders
  - Very promising solution in graph neural networks
  - Particular flavor: graph convolutional edge classifiers

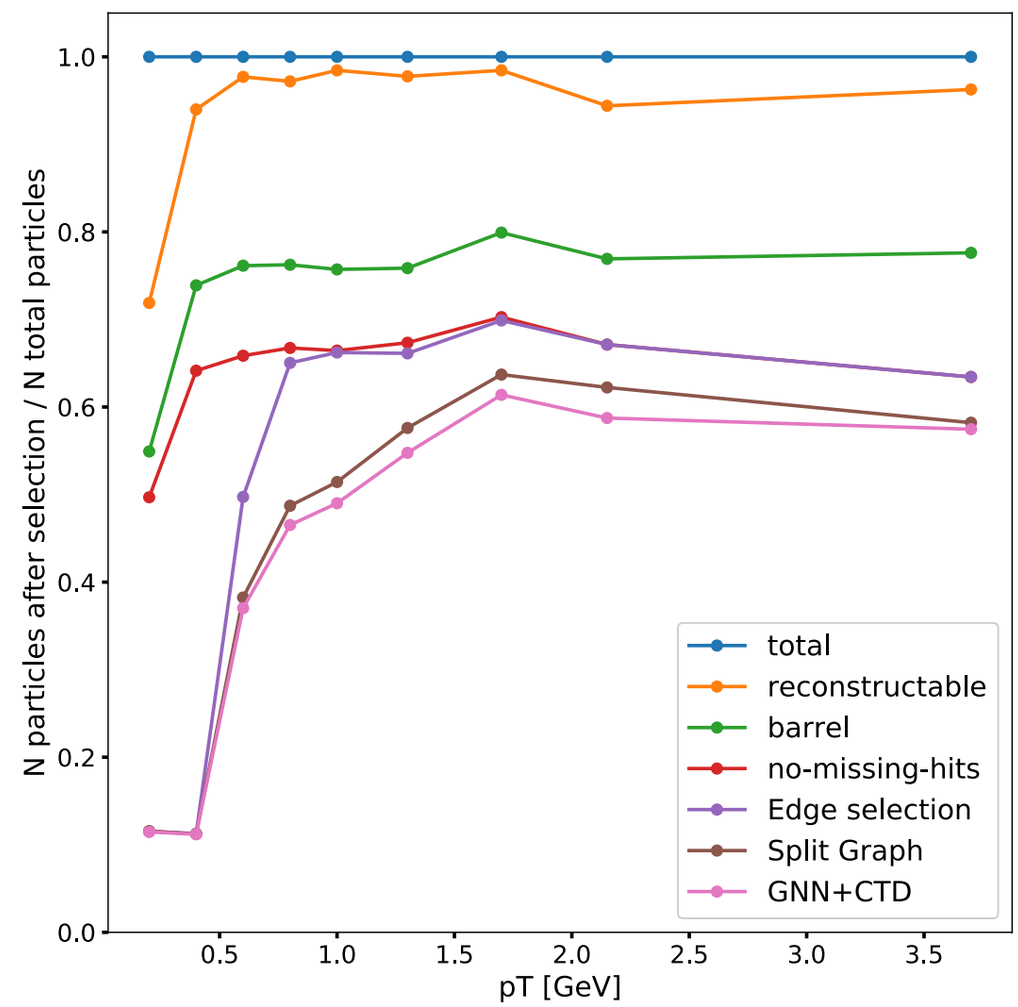


 https://arxiv.org/abs/1810.06111

First Performance and Promise

- Many selections applied to yield training set
  - Important: sectorization and no missing hits
  - These are "easy" tracks but this also early days for the these kinds of network in HEP
- Applying GNN, assembling tracks -> 97% efficient relative to preselection
  - Track-segment selection GNN executes significantly faster than Kalman filter

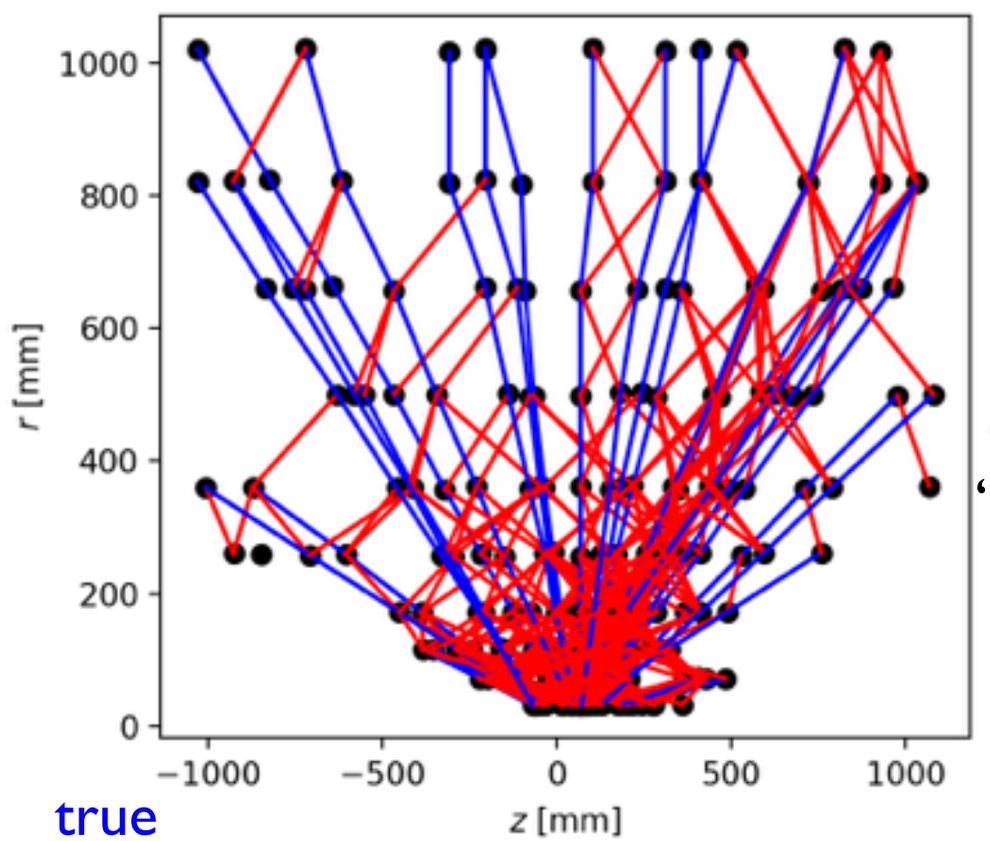




X. Ju @ CTD 2019

From Tracks to Clusters

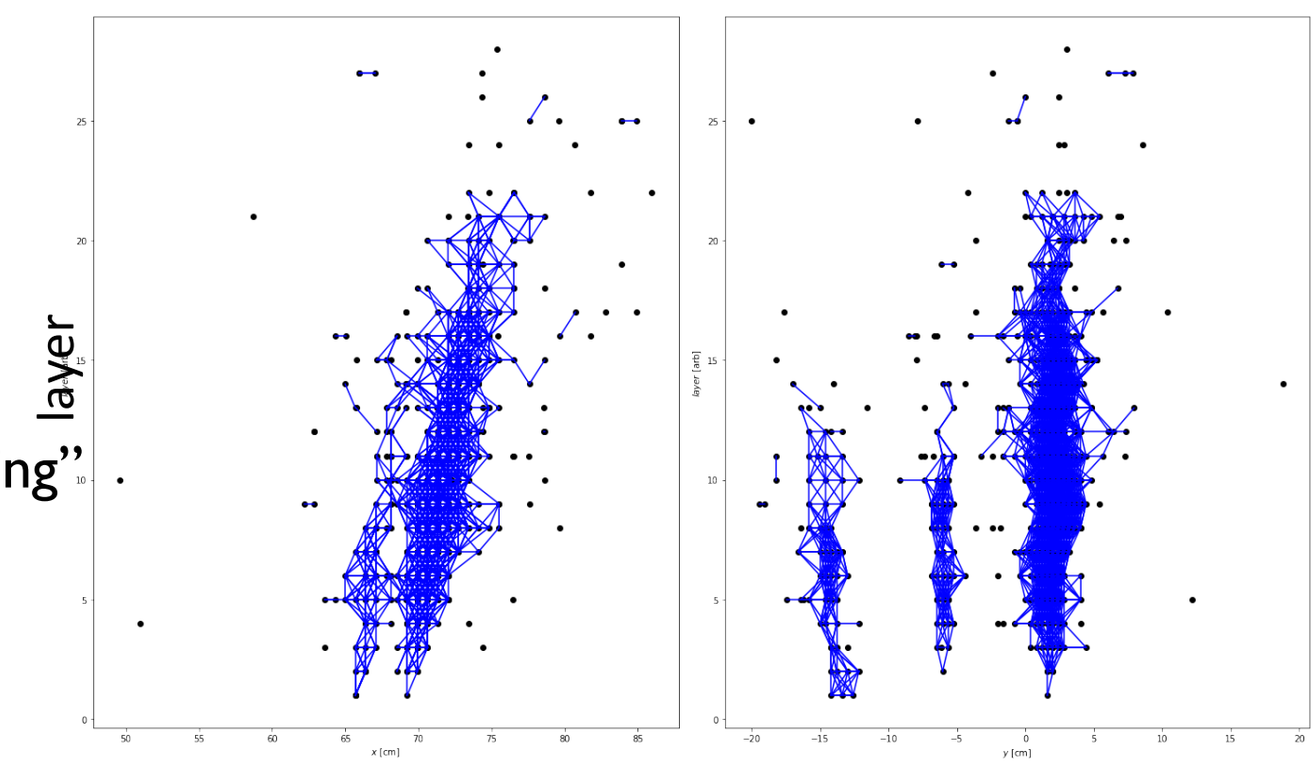
- Imagine a calorimeter cluster as a clump of edges
  - In terms of abstract of relationships between hits this is the same object as a track



true

false

"same thing"



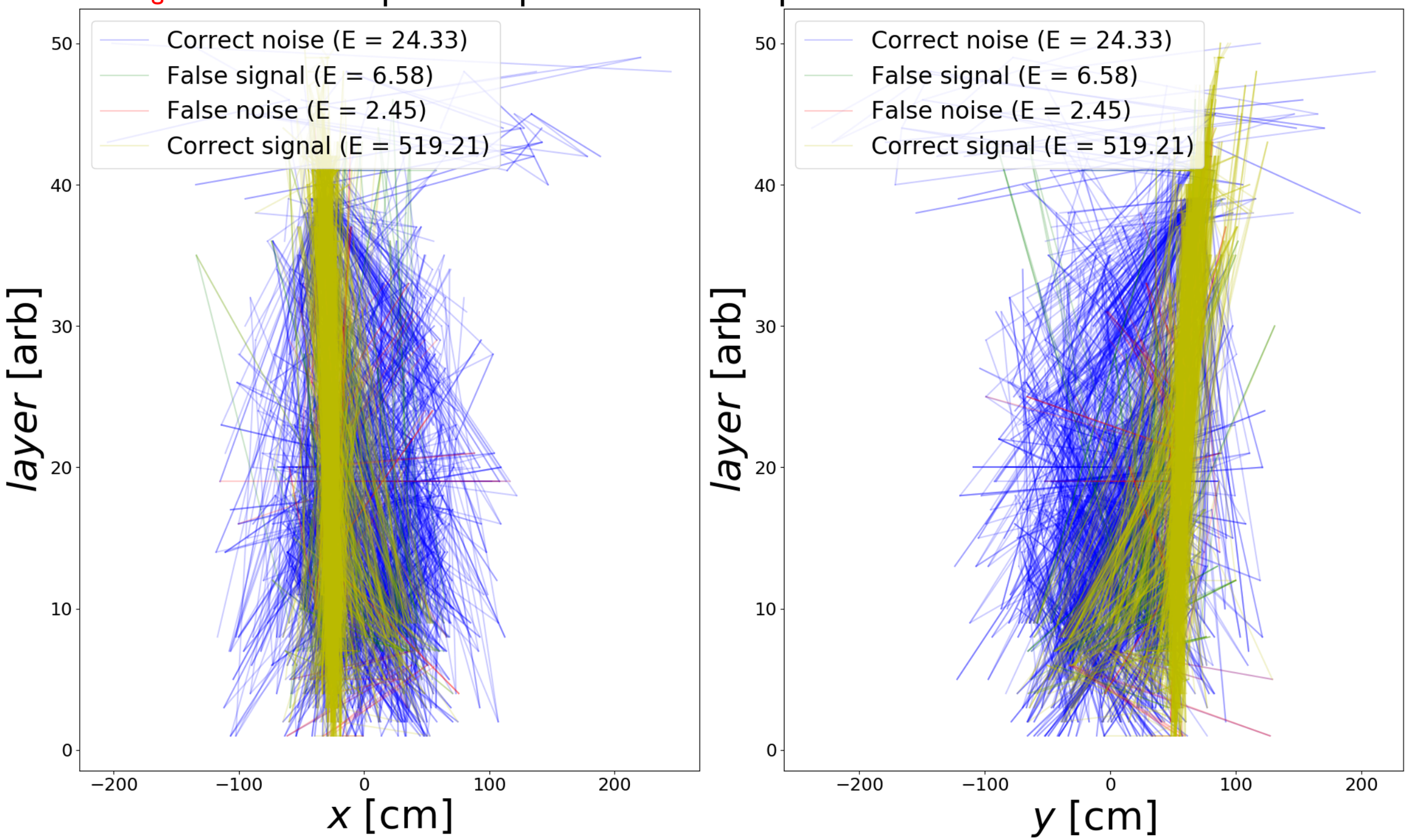
false edges not shown (too many)

Pion Classification in HGCal

true negatives
true positives
false positives
false negatives

Pion shower identified in a sea of noise

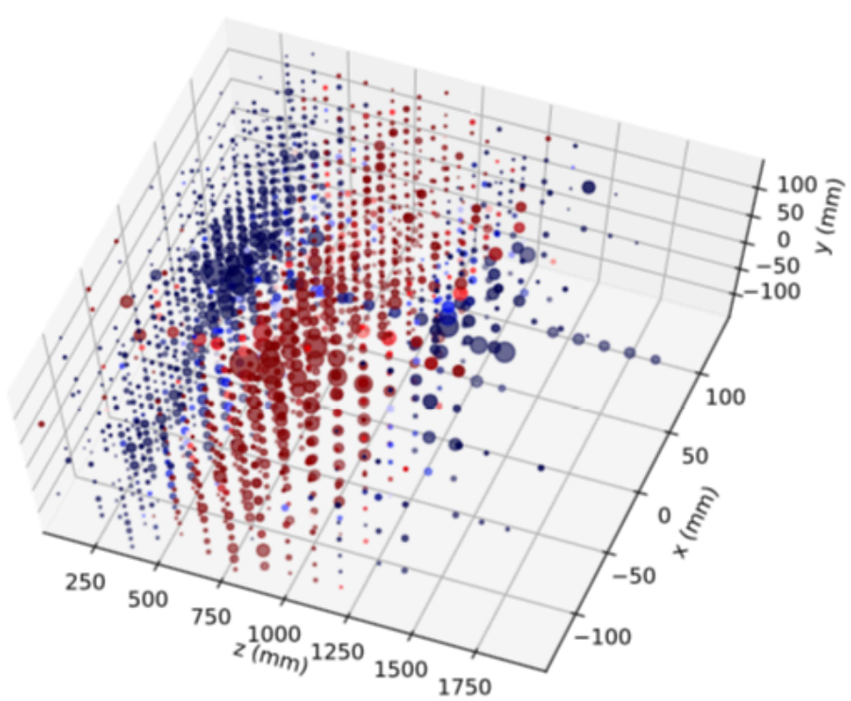
In depth example for more complex scenarios in a few slides…



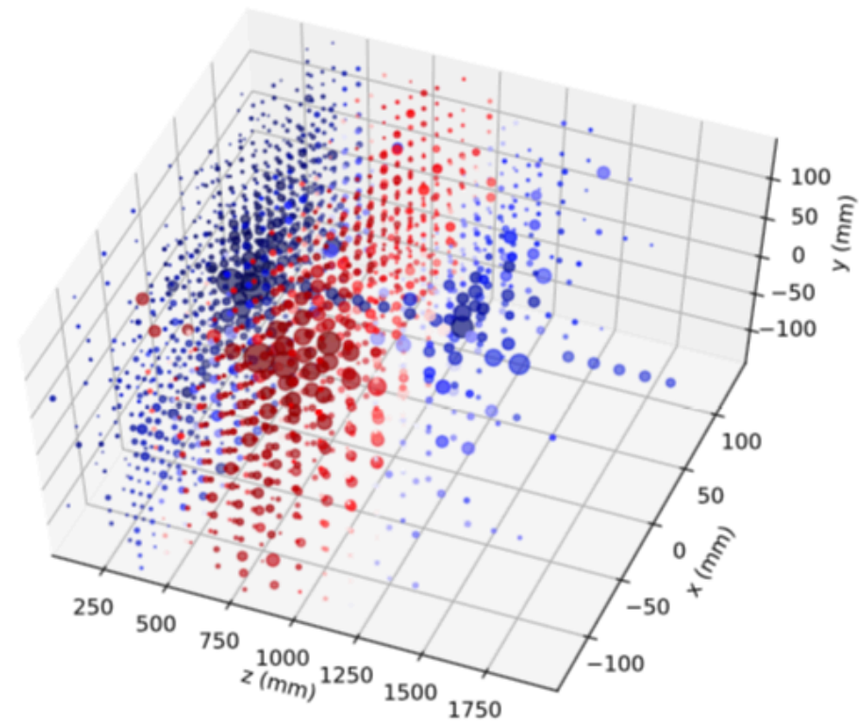
S. An / LG / T. Klijnsma / K. Pedro

Finally: Towards Actual Clustering

- Finding correct edges in a static graph is a well characterized learning problem
  - And you can use simple algorithms to keep your data processing quick!
- Clustering is better characterized as "finding the correct graph" rather than finding the correct subgraph
  - It's possible to learn the the latent space which best clusters the data
  - But! Right now, the number of clusters / categories needs to be known beforehand



(a) Truth



(b) Reconstructed

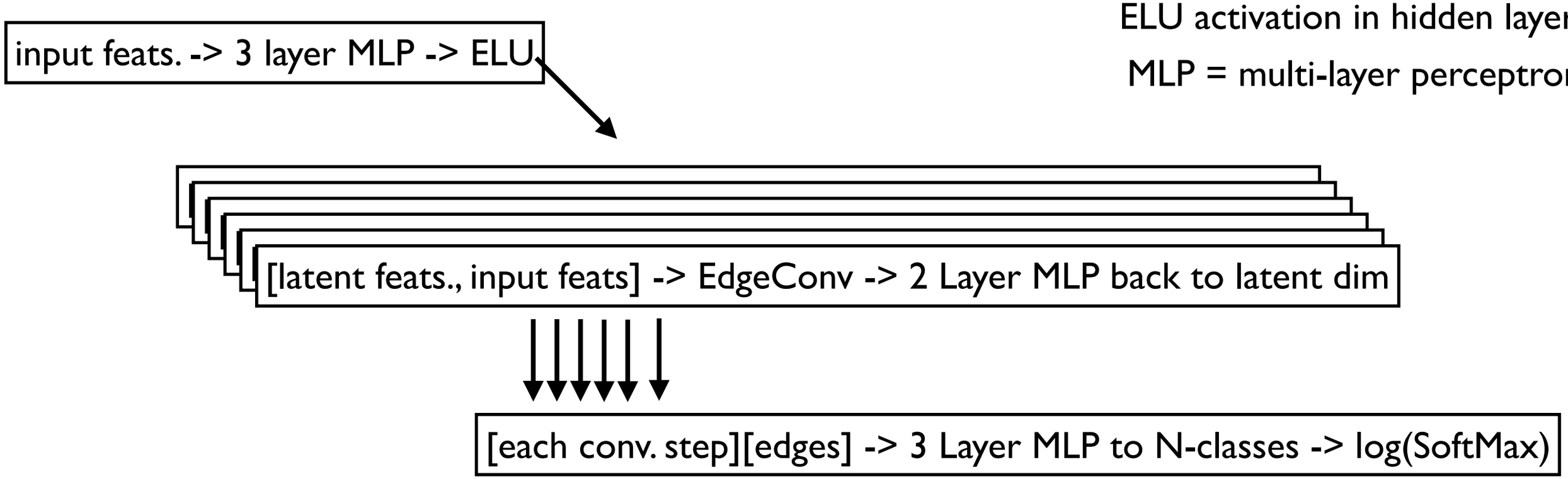
https://arxiv.org/abs/1902.07987
https://arxiv.org/abs/1801.07829

Edge Classification in Practice

- Given edges between HGCal hits assign class
  - classes are: [noise, HAD, EM, MIP]
  - noise is defined as (no sim matched energy | sim-matched but different clusters)
    - We may want a fifth class of "can't decide" edges
  - Particle classes are defined via sim truth, EM is electrons, photons, pi0s, MIPs are muons, hadrons are everything else
- Network is a fixed-graph edge-convolutional network, classifying output edges
  - 64 dim latent space, 6 graph iterations, log softmax output

ELU activation in hidden layers

MLP = multi-layer perceptron



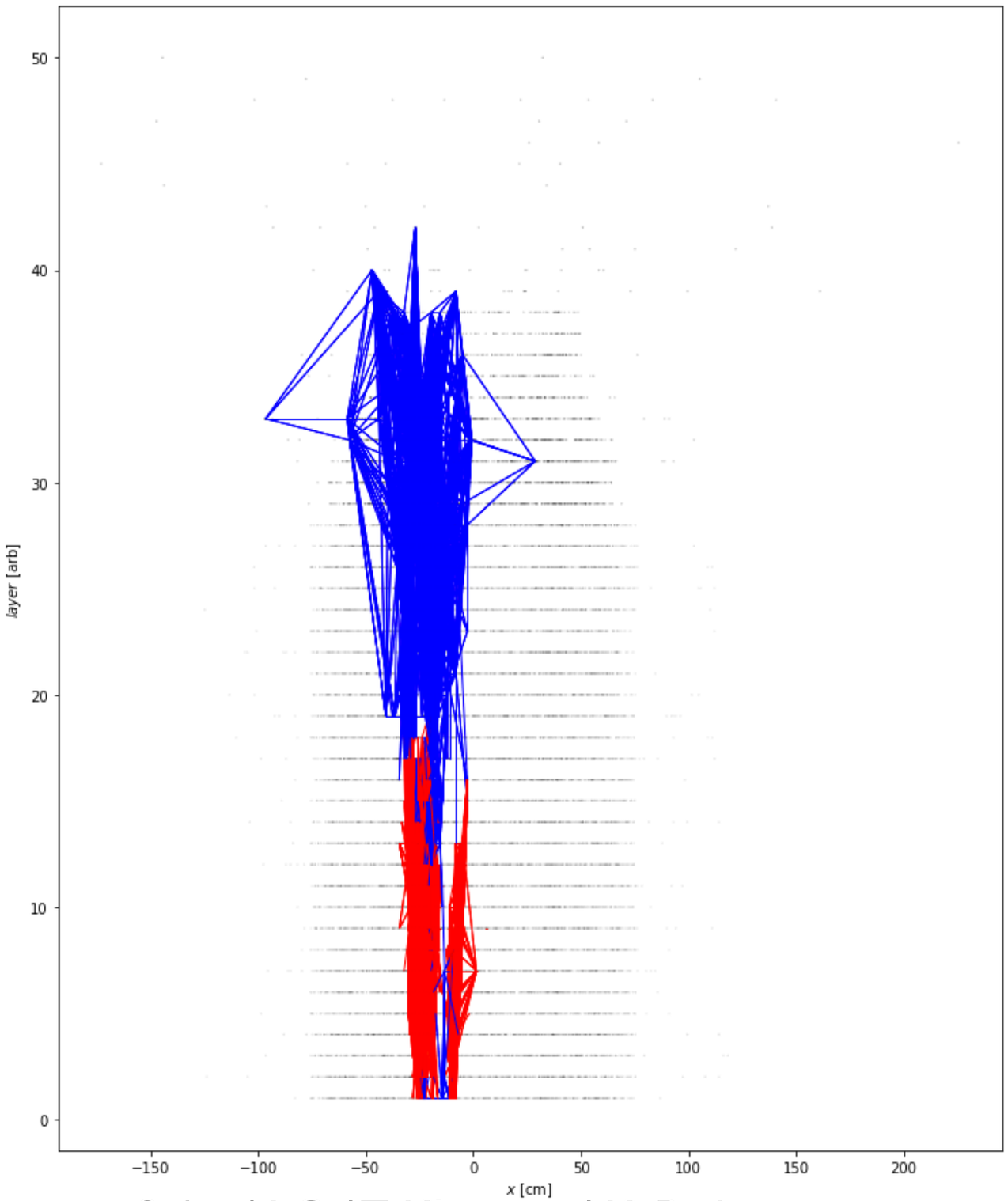
Confusion Matrix after 100 Epochs

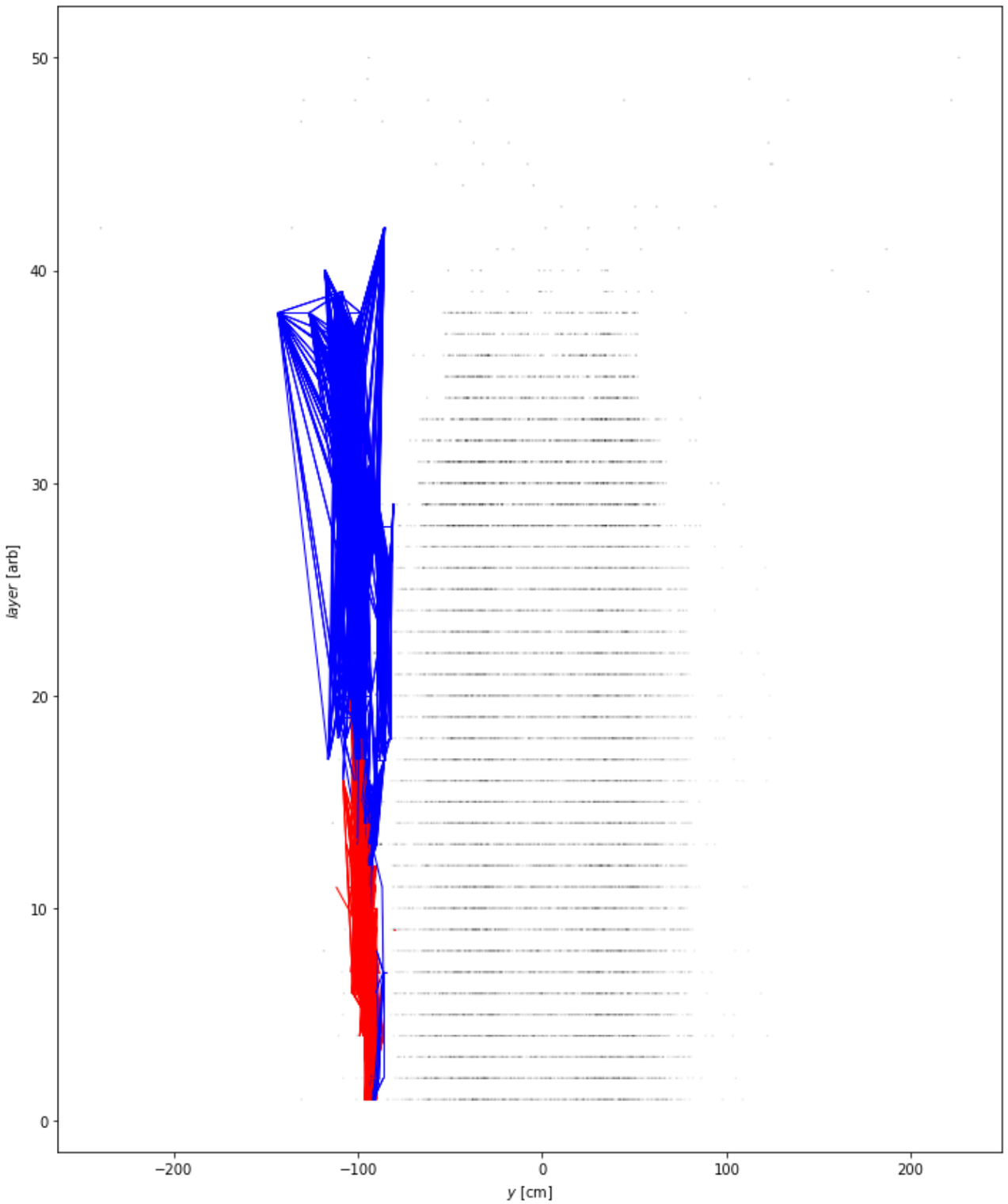
- Large amount of noise edges in training and validation data
  - More false edges than correct edges, given no noising cleaning (next will try pointnet to clean up)
  - Last time -> After 39 epochs only ~65% in each category
  - ~10% confusion EM to Hadron, and MIP -> Hadron + EM
  - Large fraction of noise edges in all categories still
    - Best guess so far is that this is coming from mixed clusters or completely overlapping EM clusters, etc.

	Noise	Hadrons	EM	MIP
as Noise	[[9.9308336e-01	9.3118154e-02	1.7514184e-01	3.0126752e-02]
as Hadrons	[3.4312138e-03	8.3171761e-01	7.4755028e-02	1.0359010e-01]
as EM	[3.4758809e-03	7.3548414e-02	7.4969733e-01	4.6642639e-02]
as MIP	[9.0124477e-06	1.6159752e-03	4.0577480e-04	8.1964052e-01]]

Efficiencies:	[0.99308336	0.8317176	0.7496973	0.8196405]
Efficiencies any cat:	[	0.9068819	0.8244523	0.9698732]

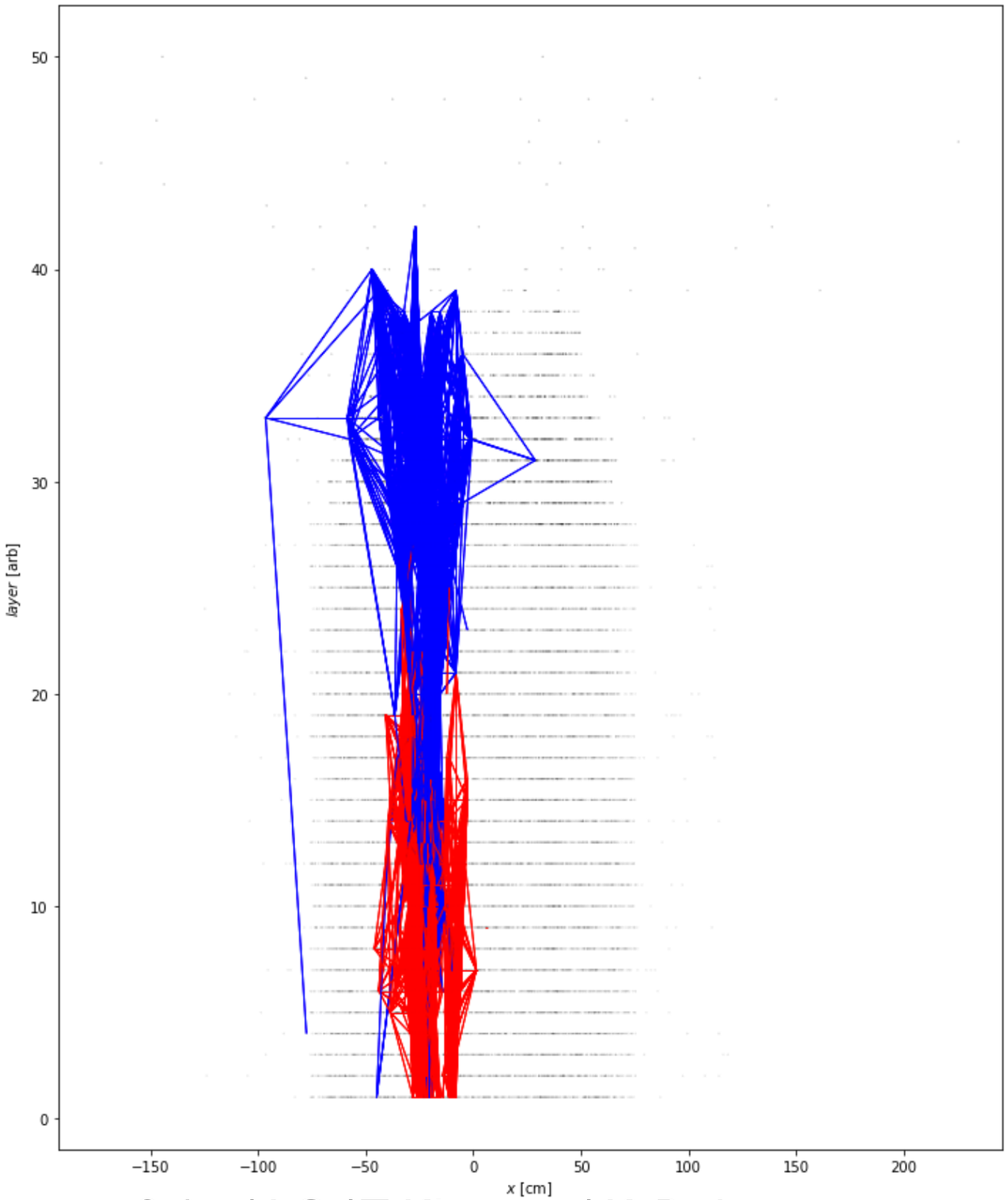
Example pi + photon Tau : **Prediction**

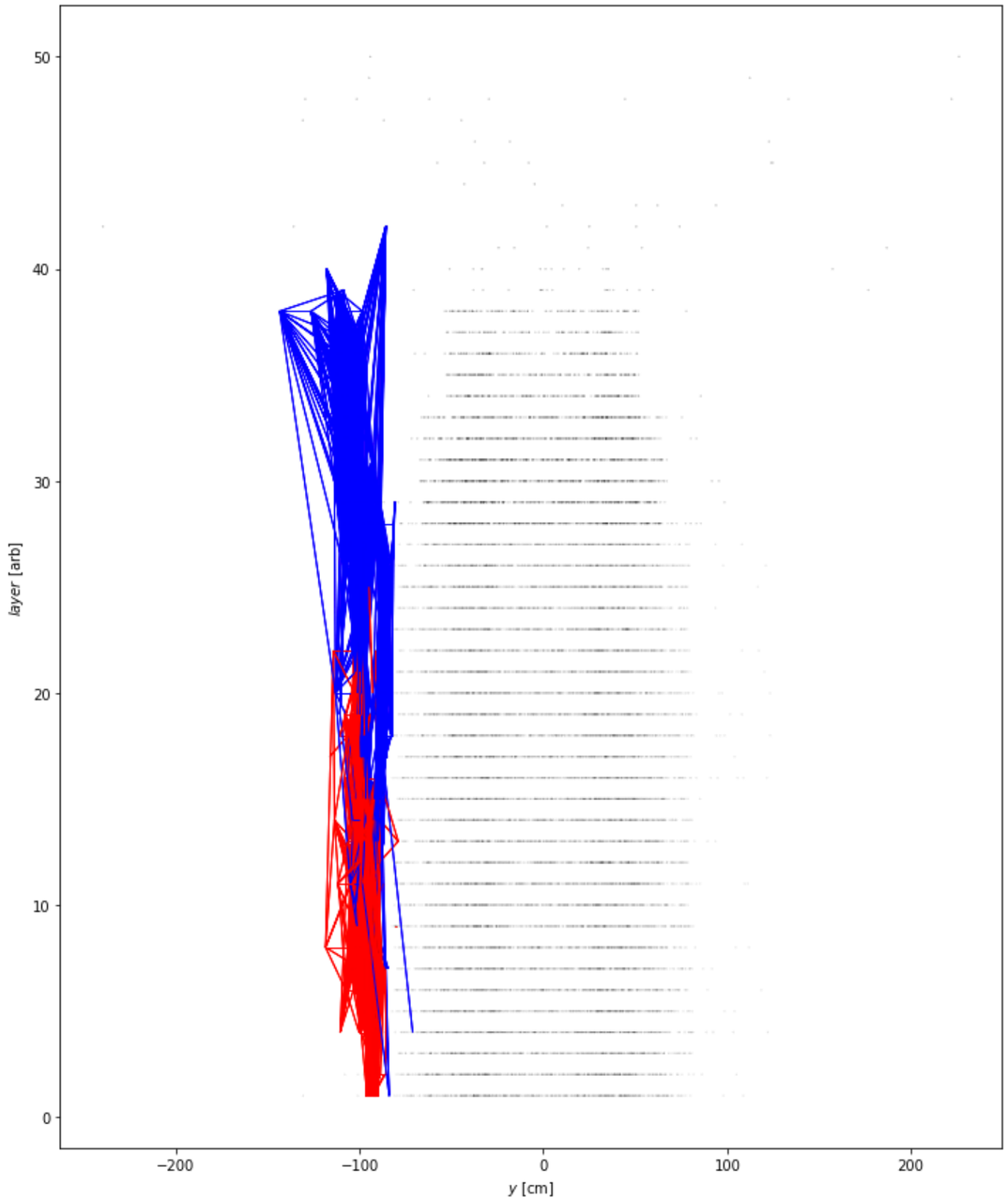




S. An / LG / T. Klijnsma / K. Pedro

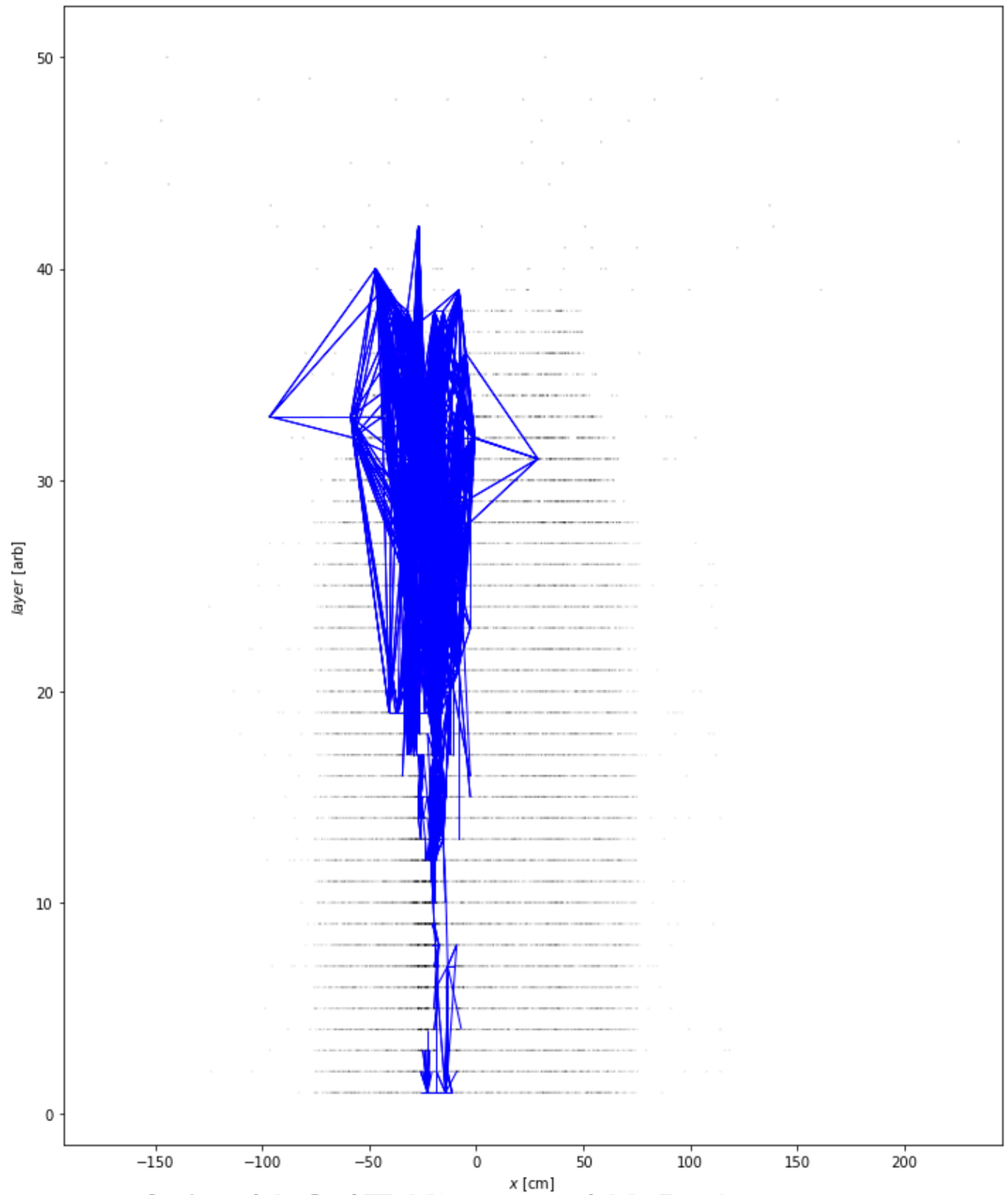
Example pi + photon Tau : **Truth**

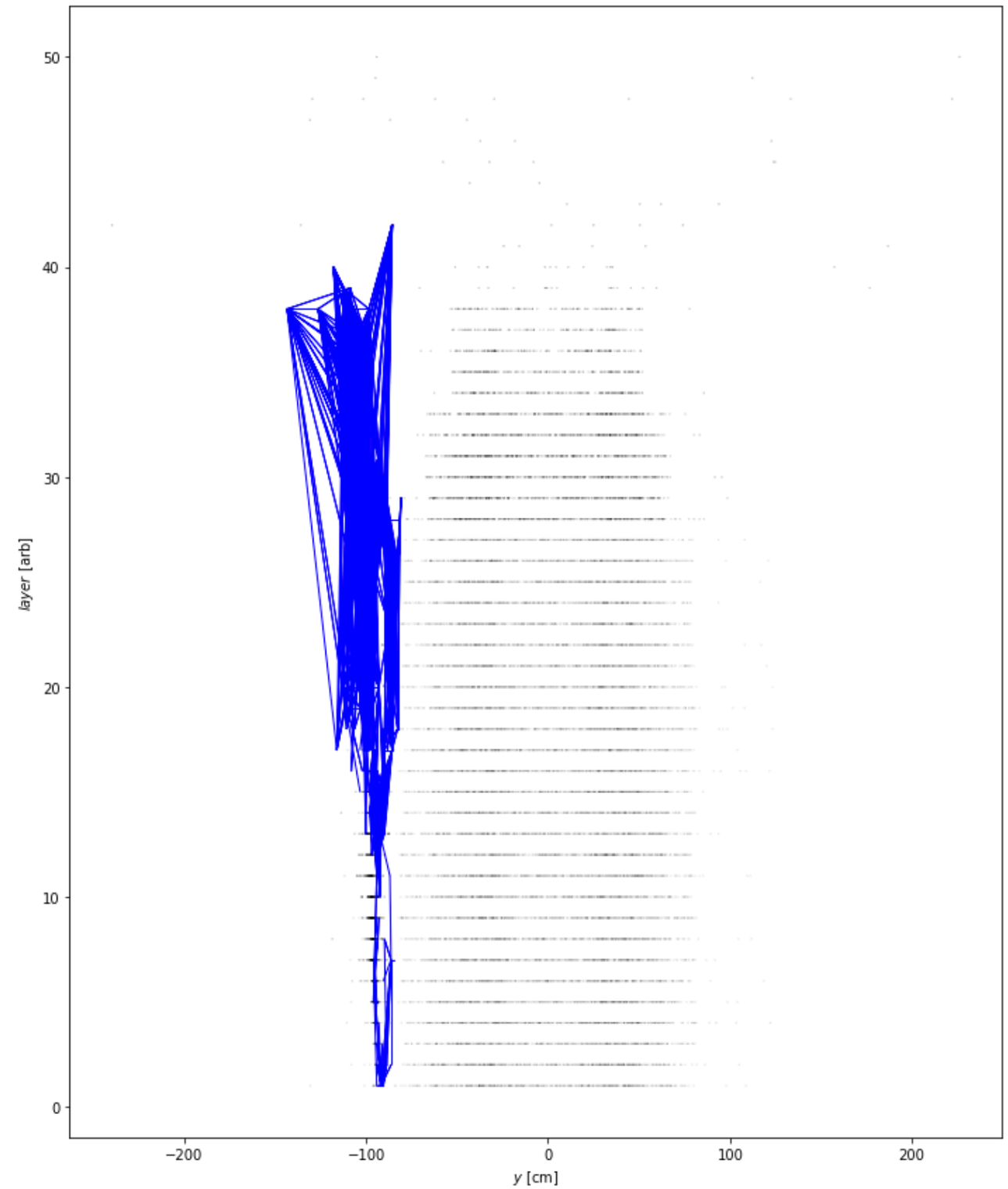




S. An / LG / T. Klijnsma / K. Pedro

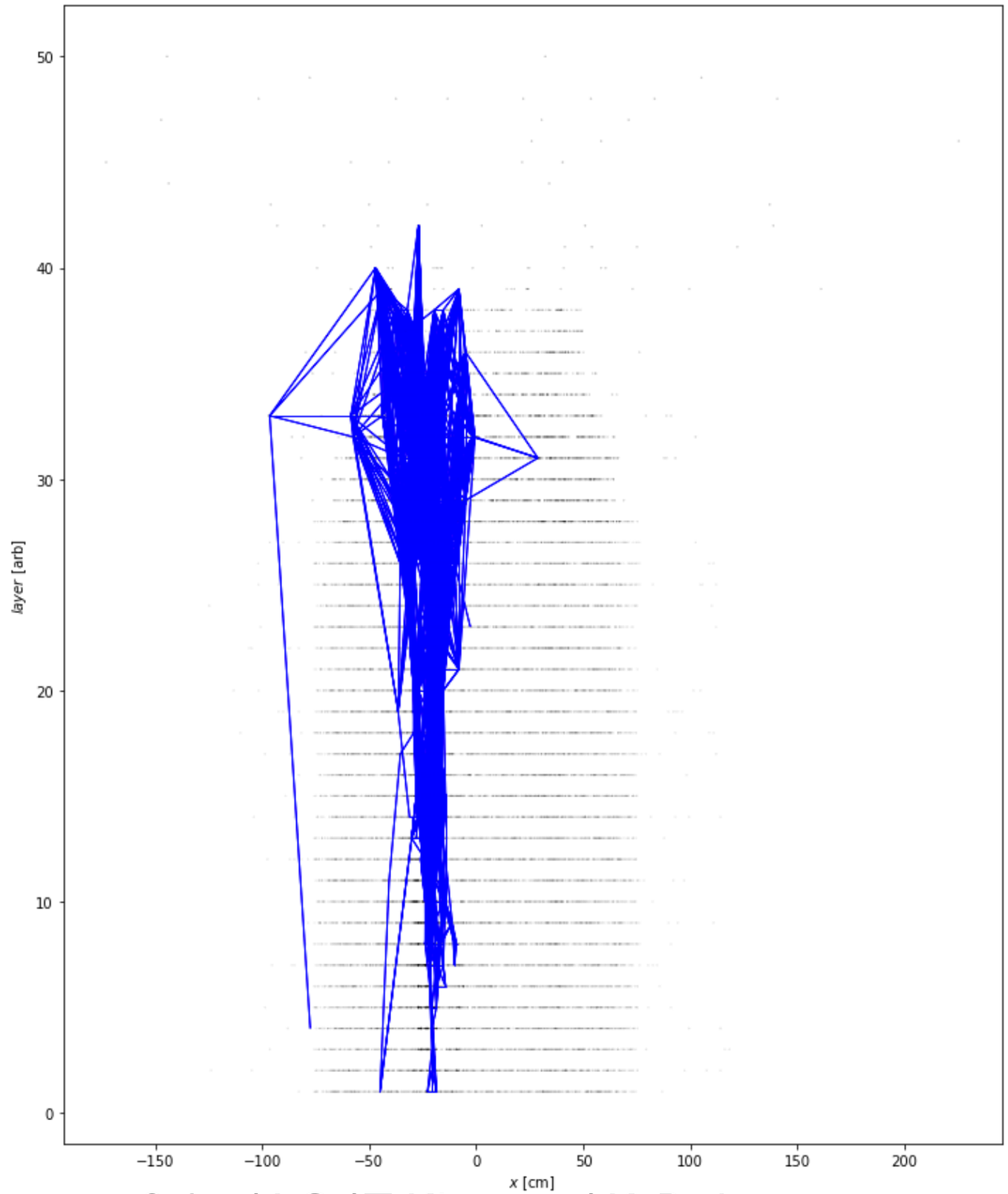
Hadron Projection : **Prediction**

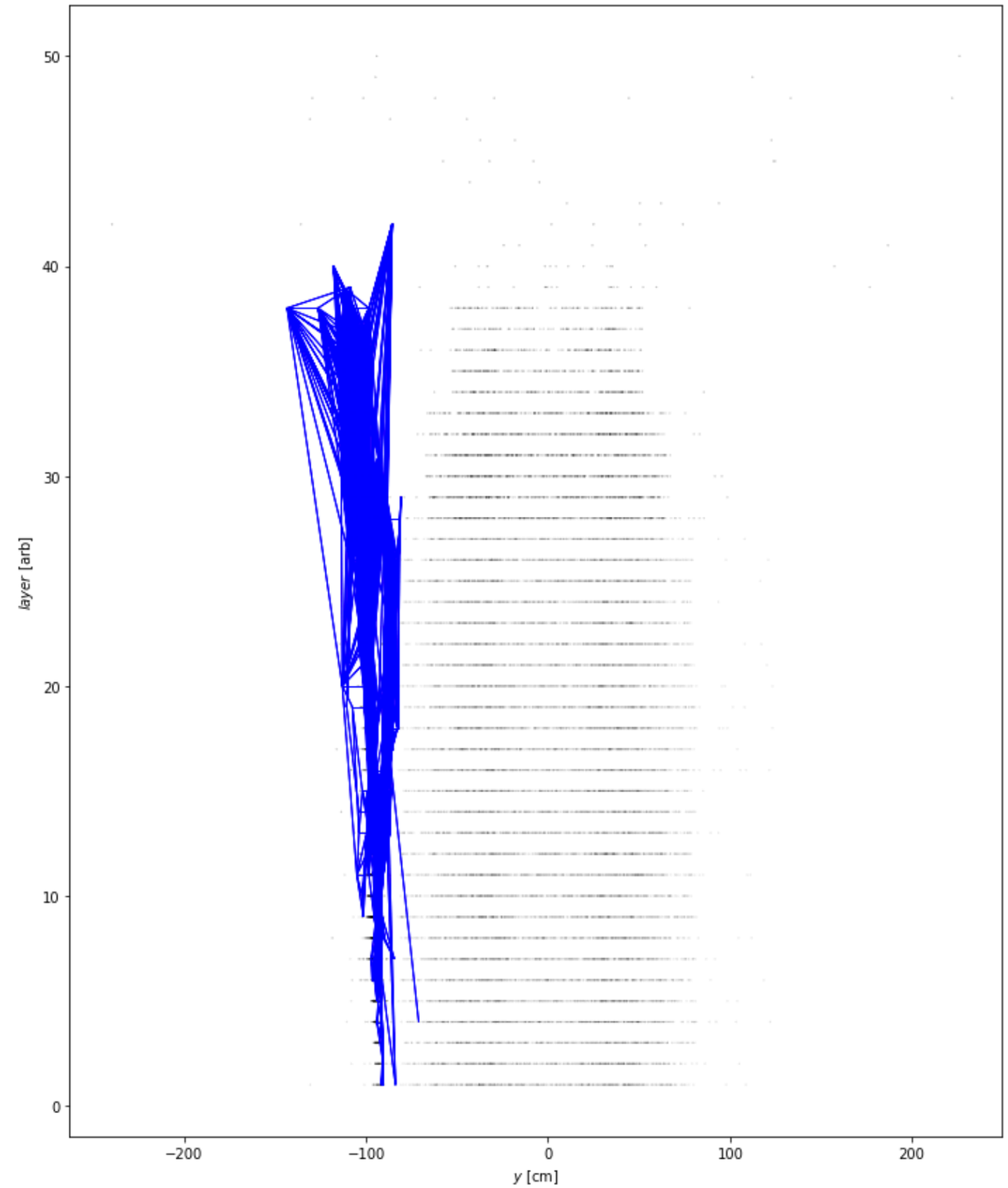




S. An / LG / T. Klijnsma / K. Pedro

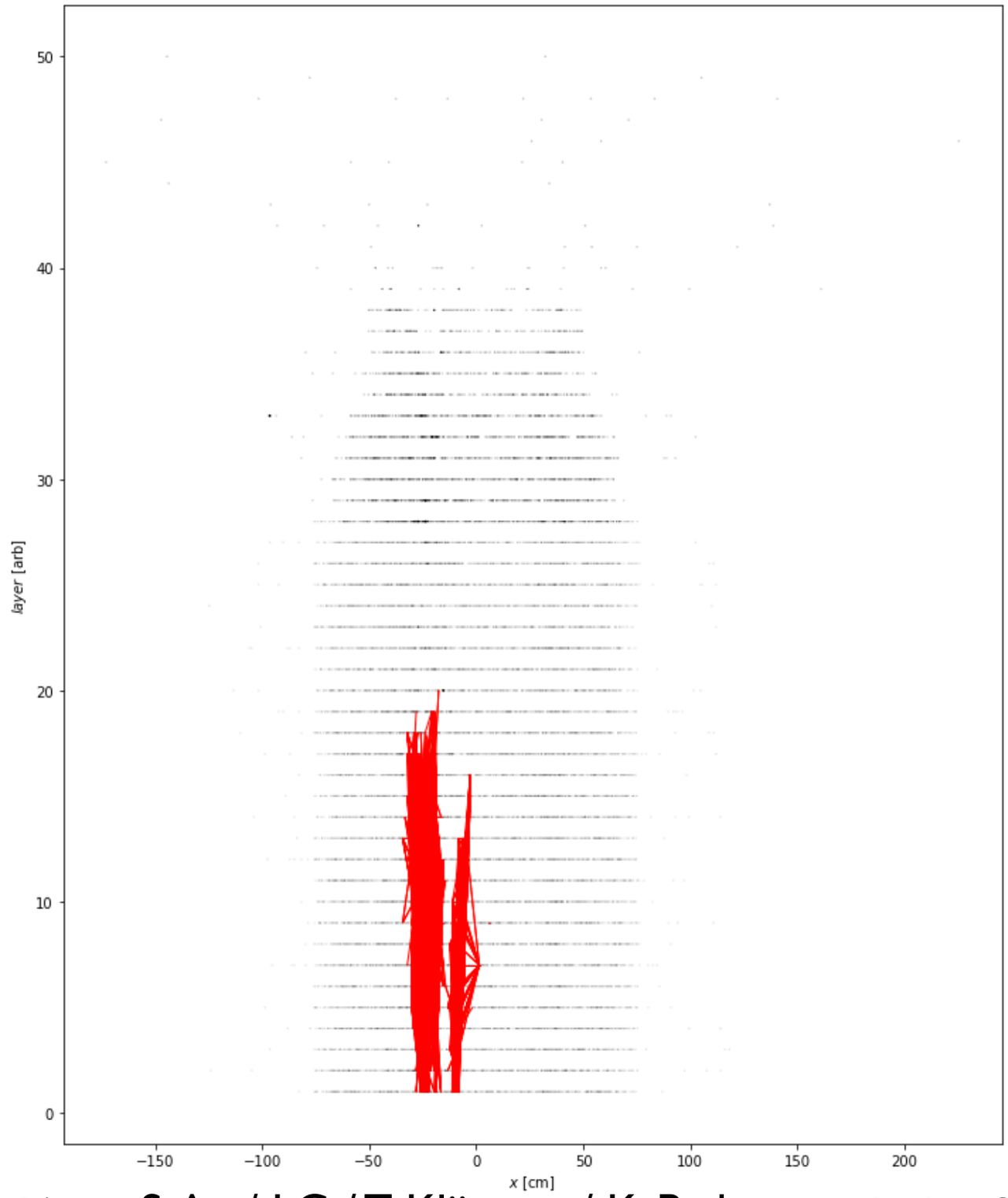
Hadron Projection : **Truth**

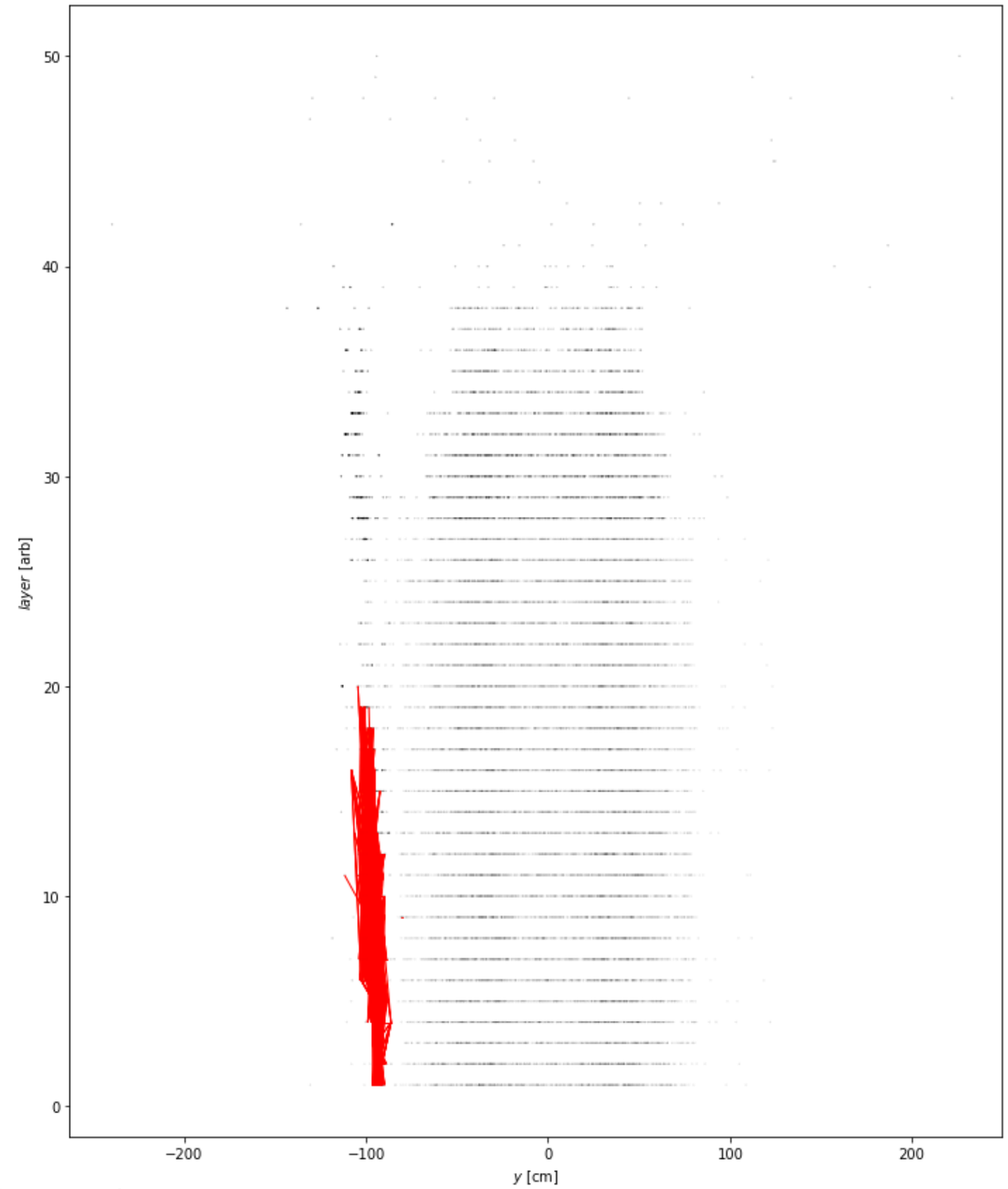




S. An / LG / T. Klijnsma / K. Pedro

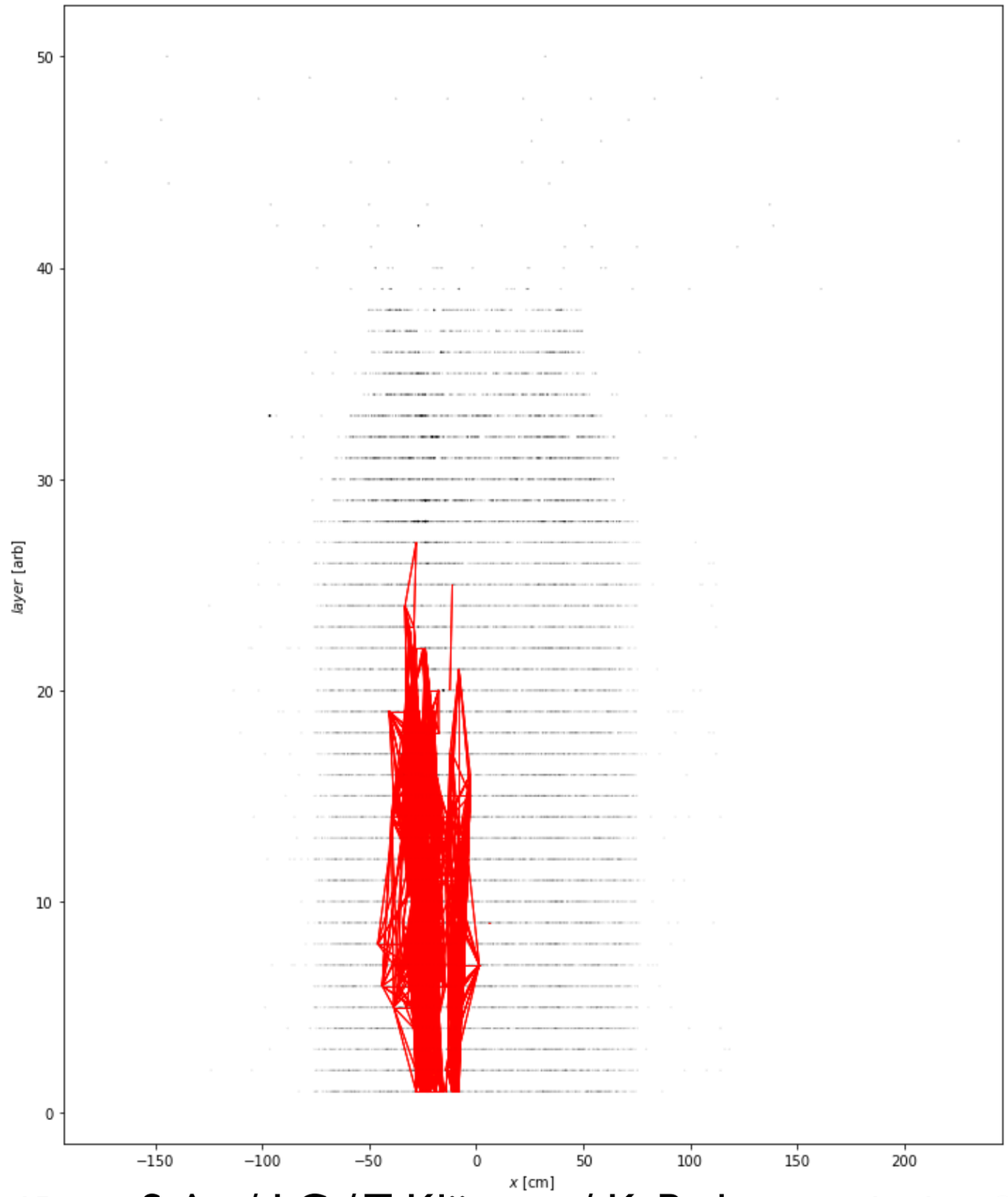
EM Projection : **Prediction**

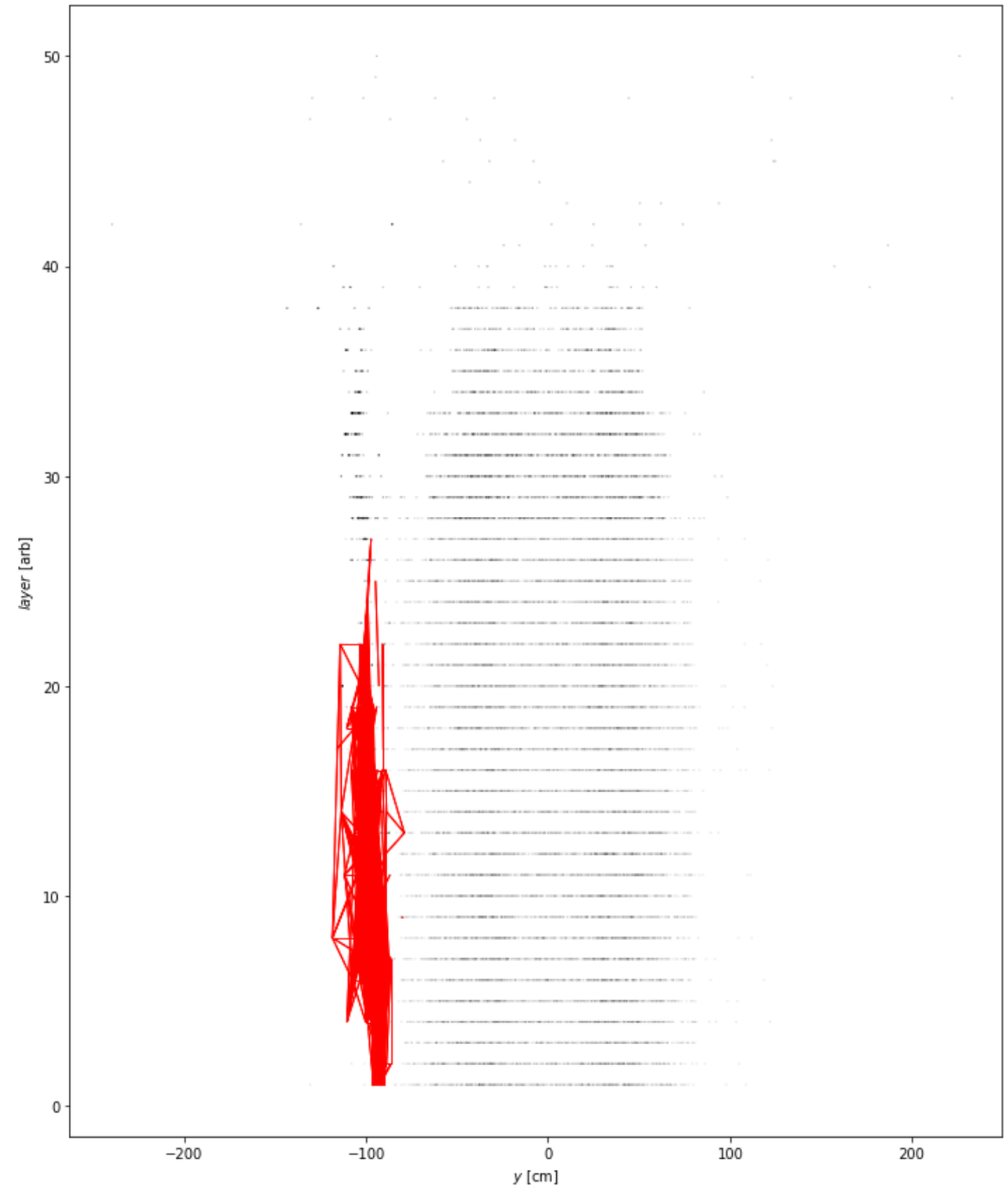




S. An / LG / T. Klijnsma / K. Pedro

EM Projection : **Truth**



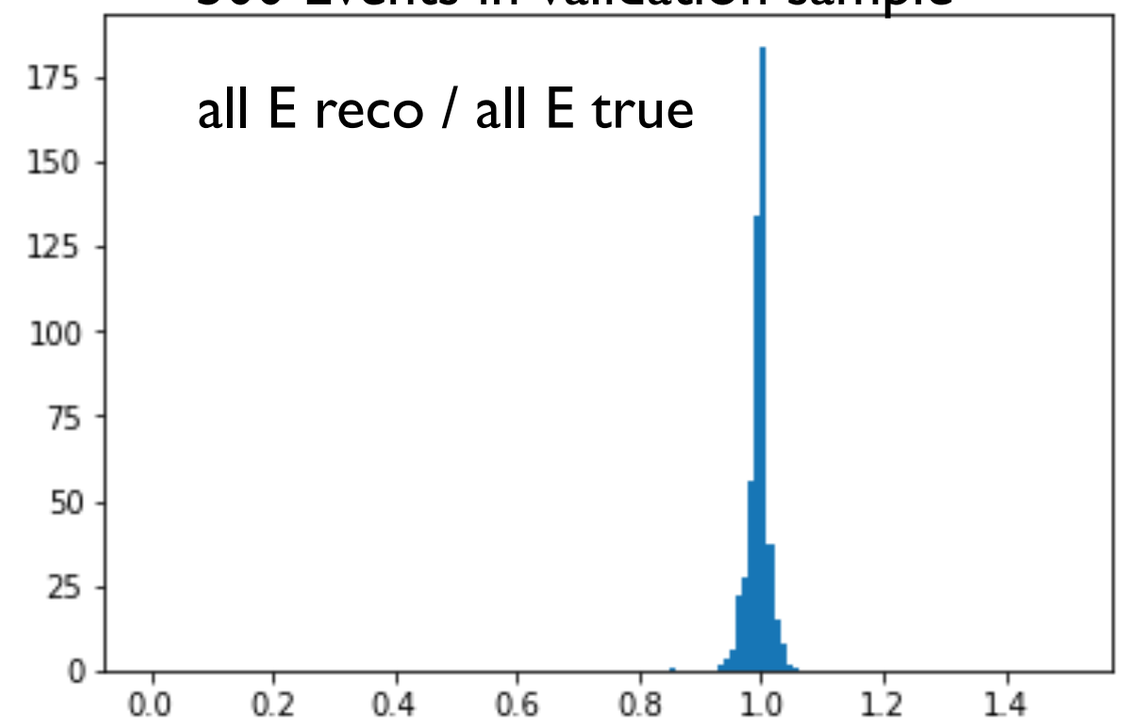


S. An / LG / T. Klijnsma / K. Pedro

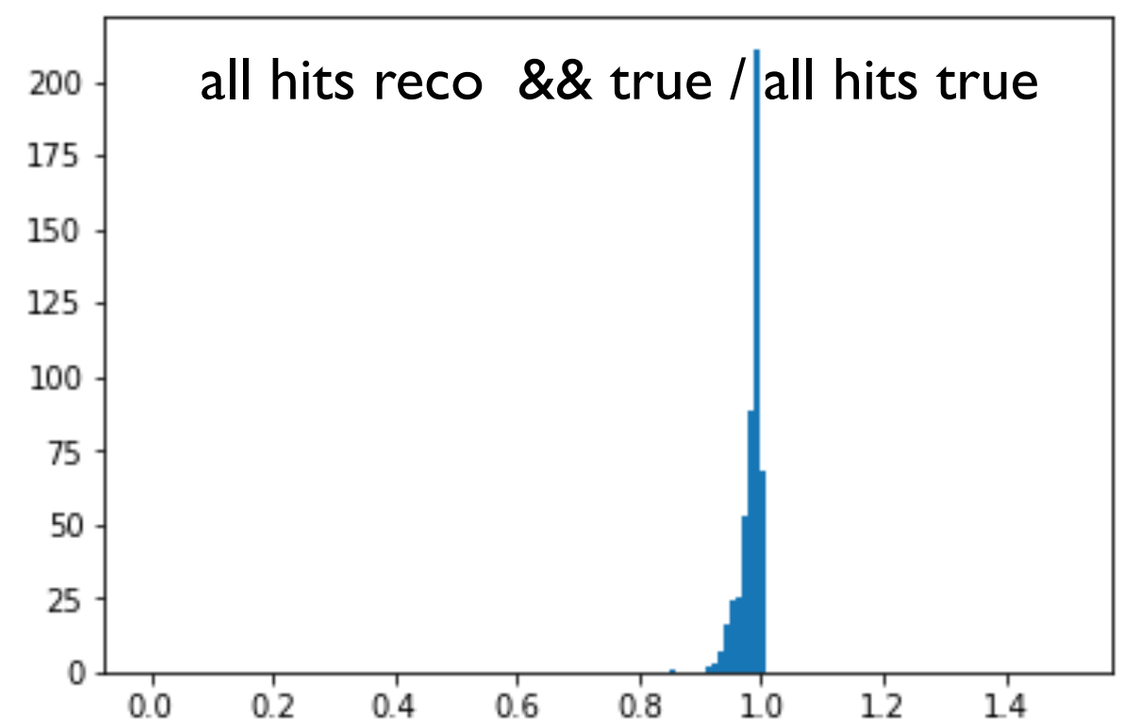
Energy Collection Efficiency

- Checking with statistics
  - Energy collection is very good
  - Some over-collection of noise
  - Most of hits connected together are truth hits
- This should allow for high quality reconstruction
  - We're now focusing on getting this into CMSSW

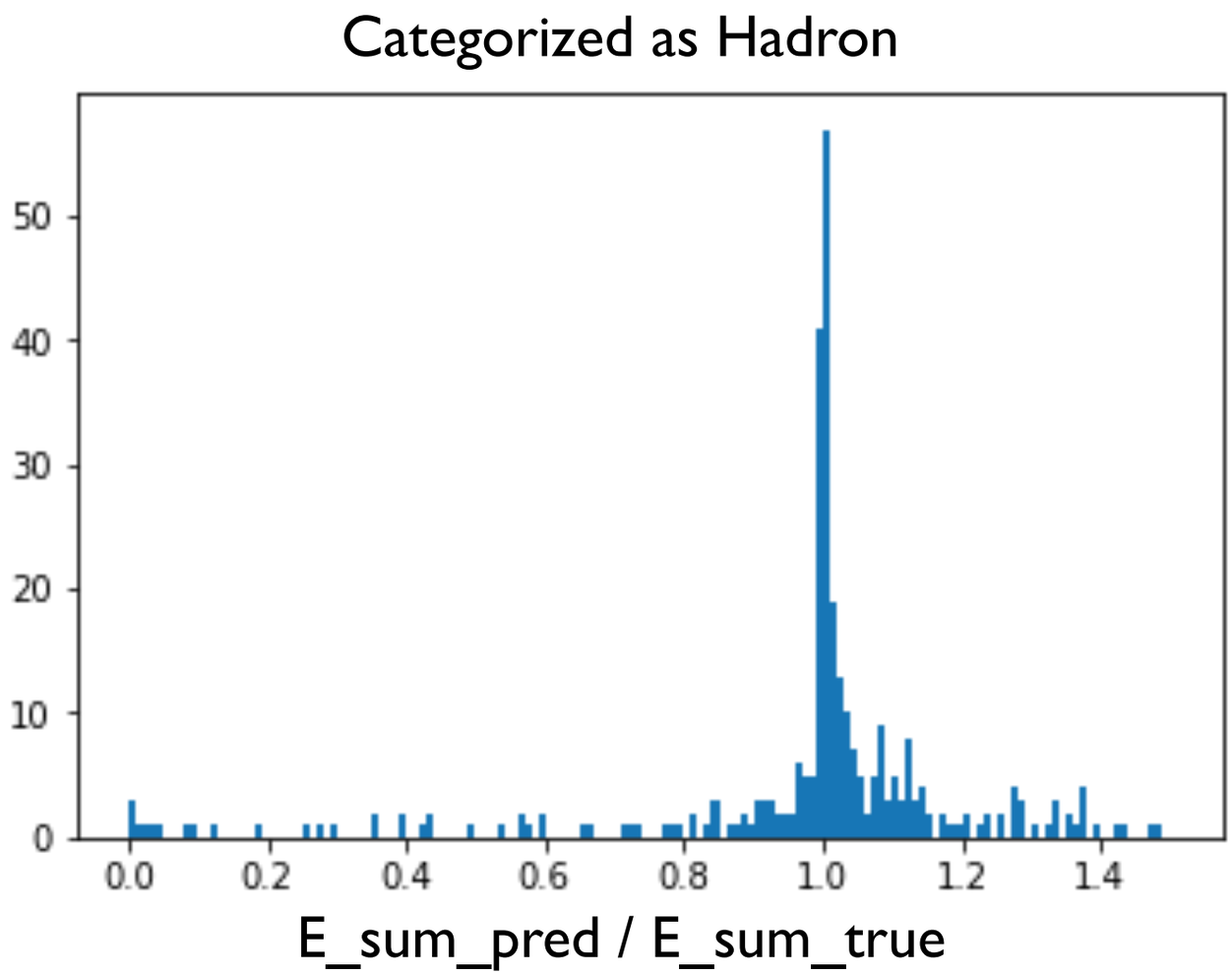
500 Events in validation sample

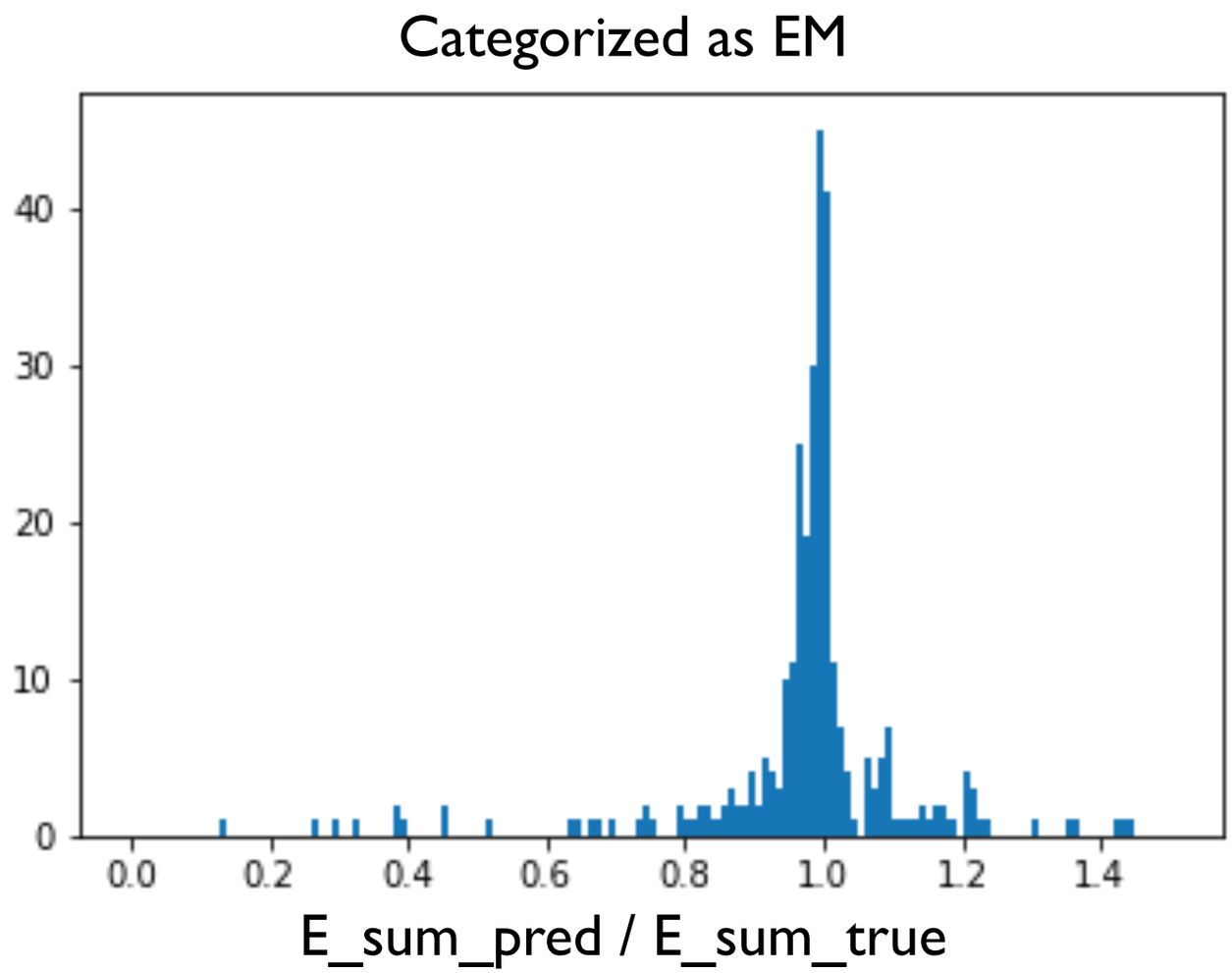


E_sumhits_pred / E_sumhits_true



Energy Categorization

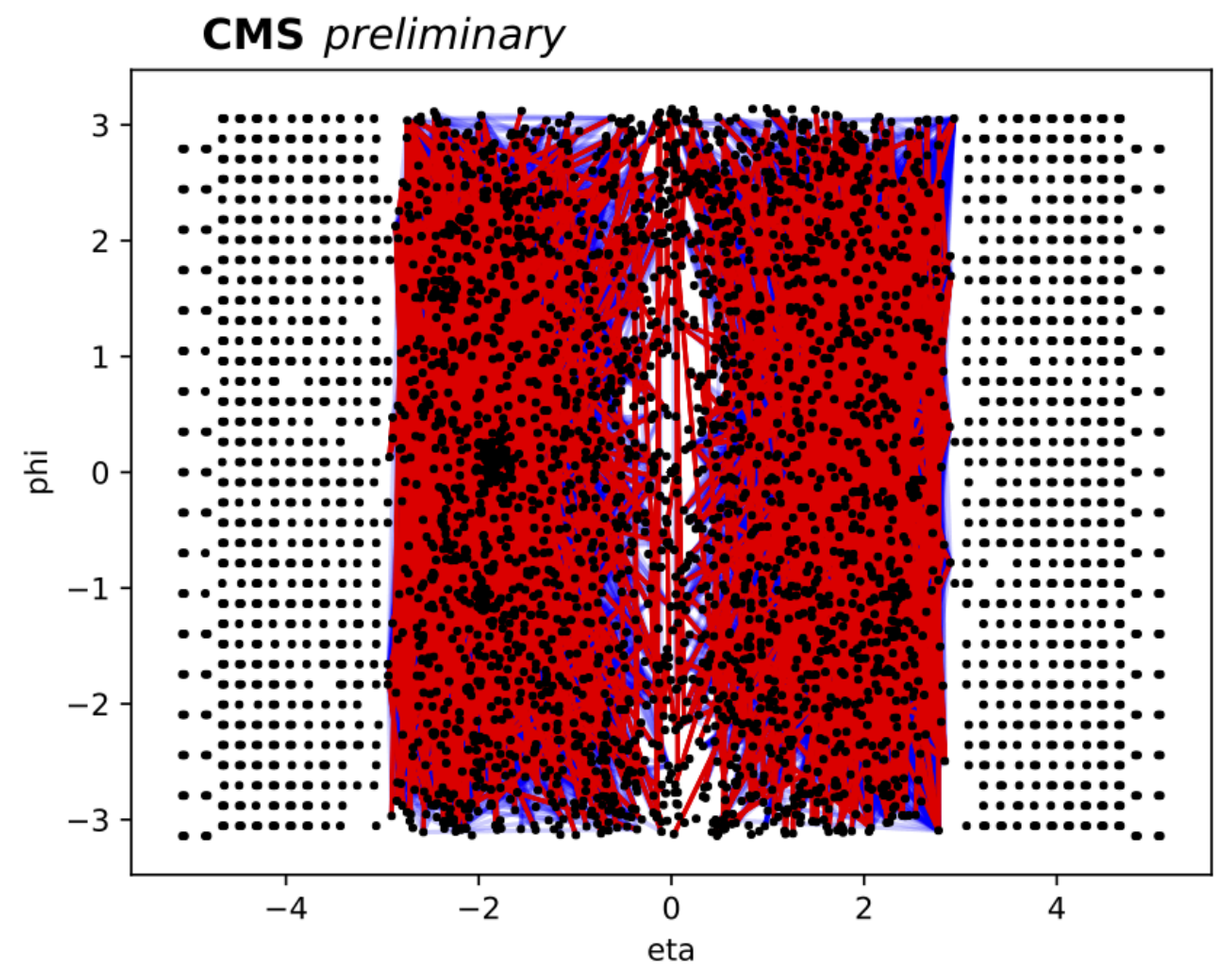




This includes low energy pions and photons, easy to misclassify.
It is important only that these plots are reasonably strongly peaked.

Recent Developments: Particle Flow

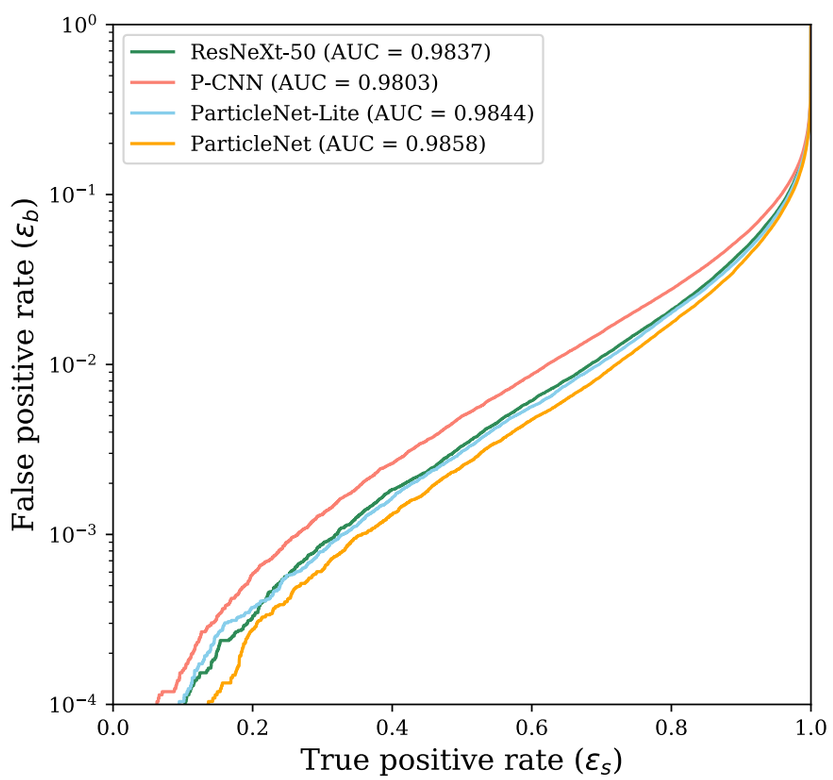
- Particle flow is yet another clustering problem
  - Cluster tracks, calorimeter energy deposits, other detector information into things that behave conceptually like gem-particles
- First explorations here are using edge classifiers
  - Other architectures to be considered



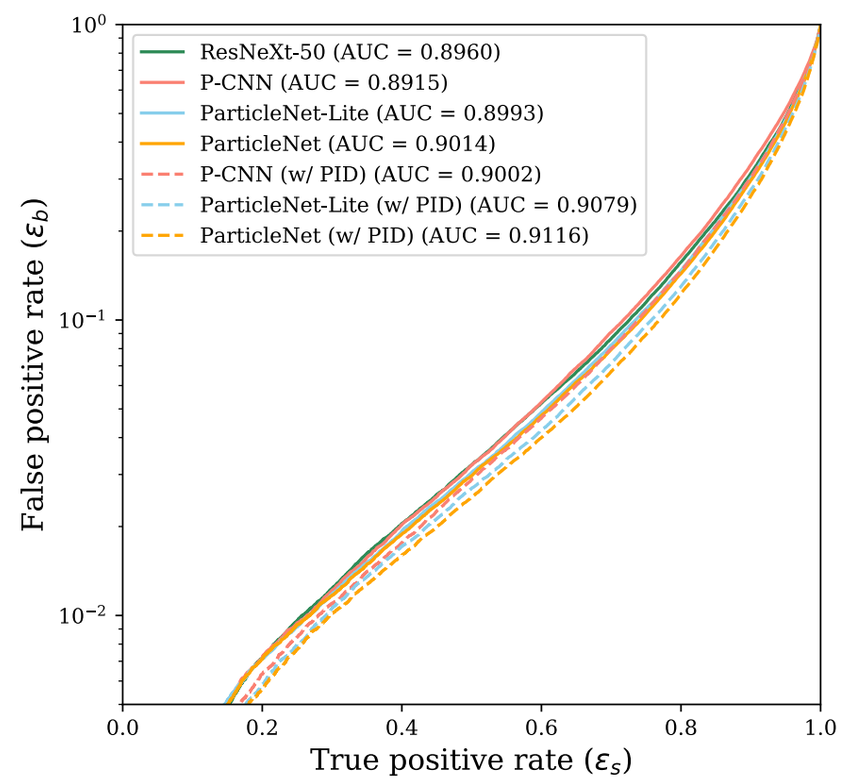
Work in progress: ~80% edge efficiency first try

Not Just for Reconstruction: Jet Tagging!

- Uses a dynamic-graph convolutional network to associate discriminating sets of particles in a jet
  - Output is a multiclassifier, softmax probabilities of jet being certain variety
  - Improved performance compared to jet images and other standard convolutional networks



Top-Tagging



Quark/Gluon Discrimnation

 https://arxiv.org/pdf/1902.08570.pdf

Outlook

- Graph networks are powerful tools
  - Highly flexible with respect to input data
  - Feature engineering is abstract and part of training
    - i.e. dimensionality of latent space for encoding
  - Generally inputs are basic quantities
    - four momenta, coordinates, times, energies
  - Much more 'natural' way of dealing with data in ML
- Implementations in tracking and clustering are all very recent
  - It will be some time before these are adopted as actual reconstruction algorithms
  - The path forward is compelling and there are multiple proofs of concept in a variety of domains
    - The tau result in 0 PU we went through in detail already surpasses human written algorithms