Unsupervised clustering for neutrino event exploration

Callum Wilkinson, LBNL

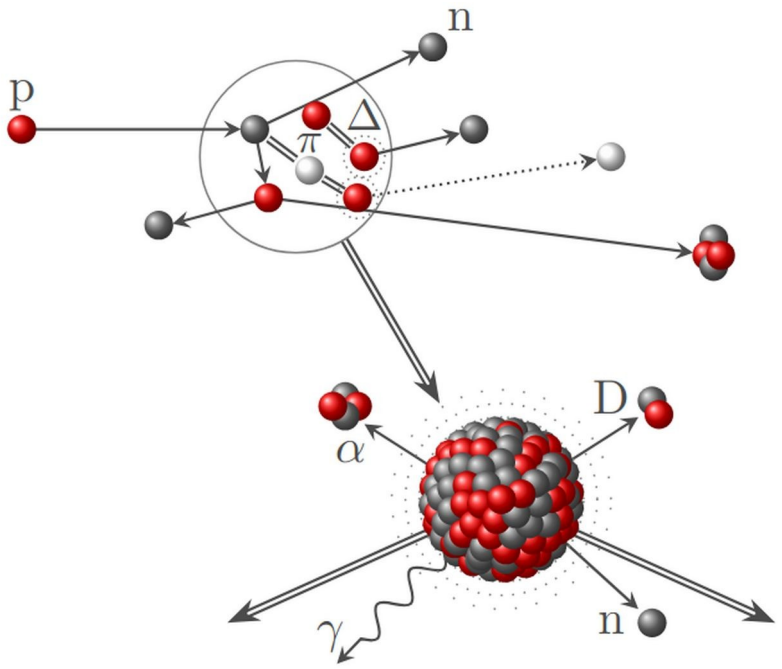
On behalf of the DUNE Collaboration

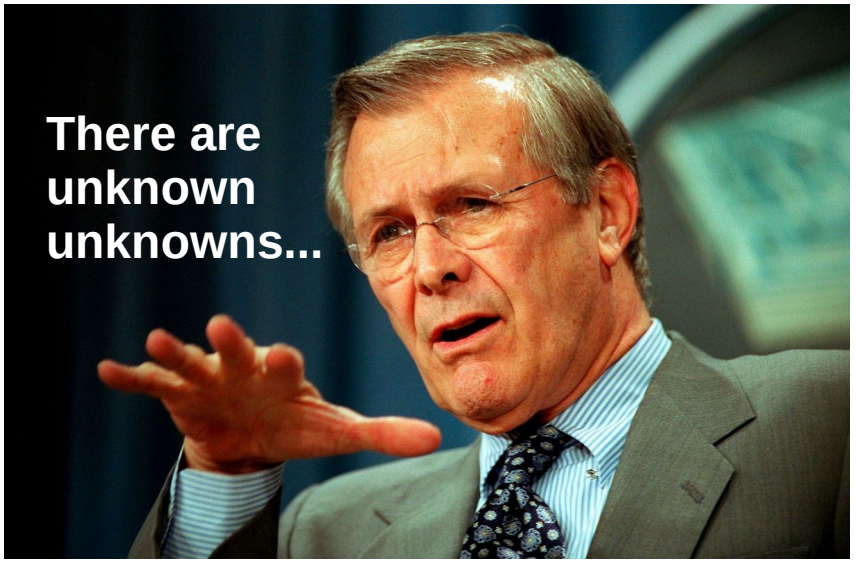
BERKELEY LAB

DUNE

What's the problem?

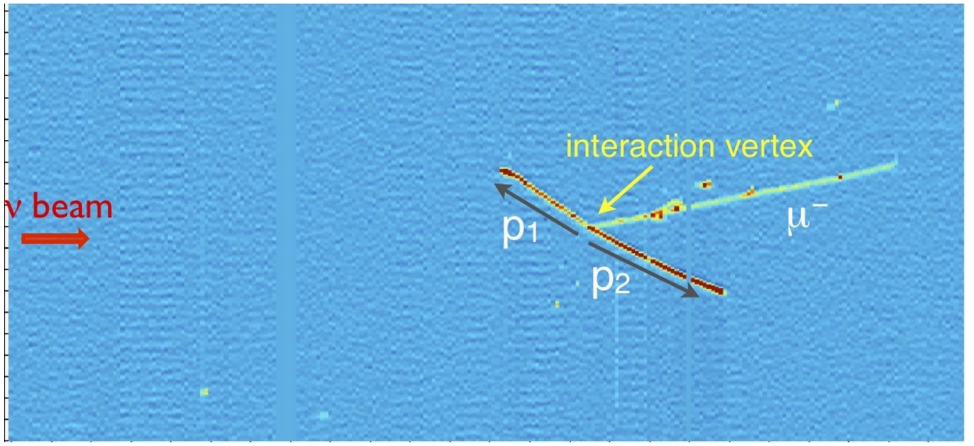
- ν-A interactions poorly understood, low-energy nuclear physics is hard
- Difficult to design analyses to measure signals we don't simulate *at all*… particularly if we also don't know where to look…





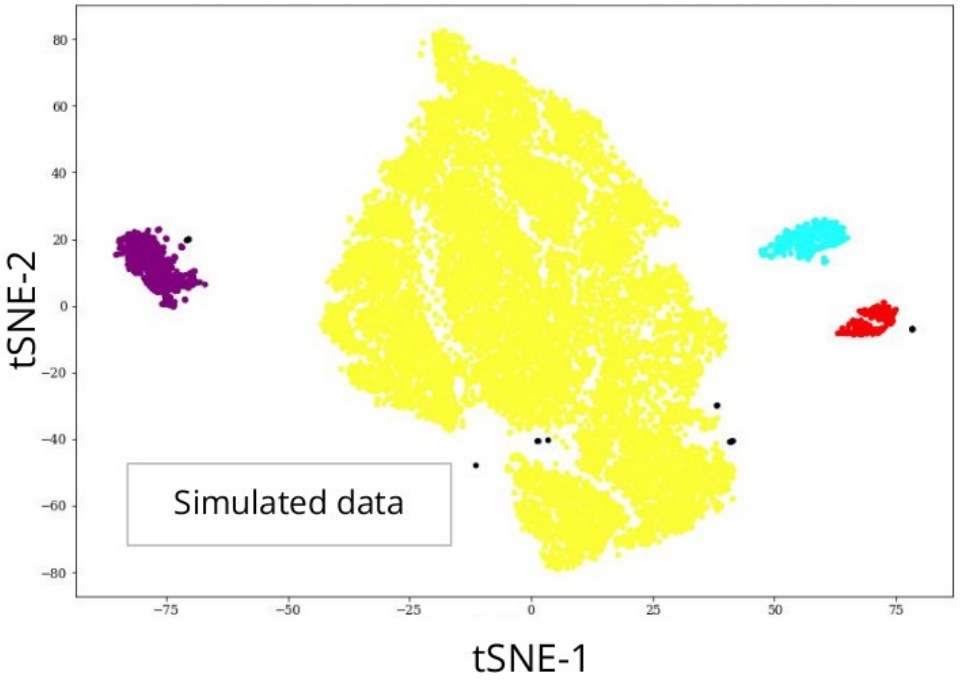
Sources of inspiration

- ArgoNeut hand-scanning: observed “back-to-back” protons
- Hypothesis: short range correlated n-p pairs in the nucleus
- Led to new theory and measurements



PRD90 (2014) 012008, arXiv:1405.4261

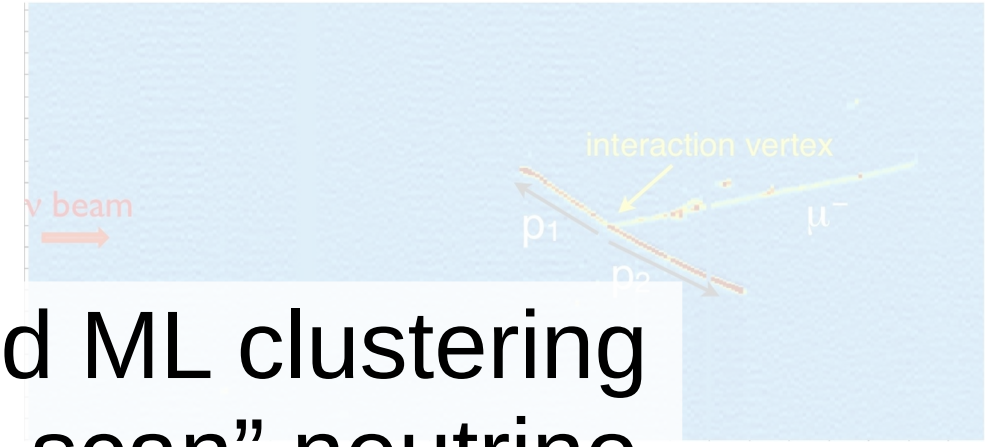
C. Amarasinghe, CoSSURF 2022 talk



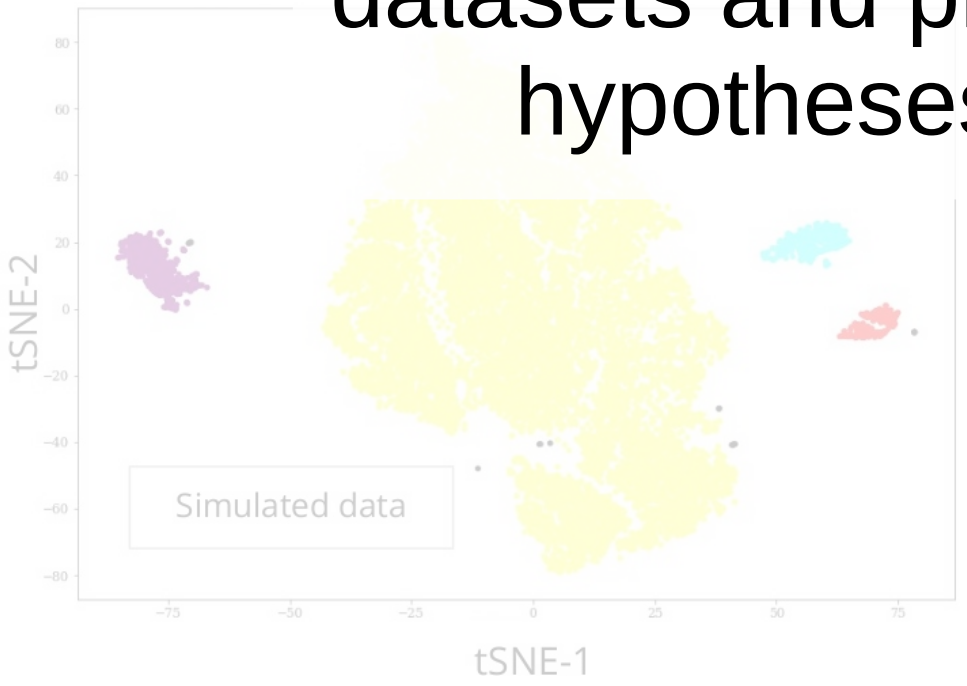
- LZ studies looking at both simulation and early calibration data
- Use dimensionality reduction (30 → 2) to look for unexplained clusters
- Example: cyan had an unexpected bug in the pulse shape model

Sources of inspiration

- ArgoNeut hand-scanning: observed “back-to-back” protons
- Hypothesis: short range correlated n-p pairs in the nucleus
- Led to new theory and measurements



Can unsupervised ML clustering algorithms “hand scan” neutrino datasets and provide interesting hypotheses for testing?



- Use dimensionality reduction (30 → 2) to look for unexplained clusters
- Example: cyan had an unexpected bug in the pulse shape model

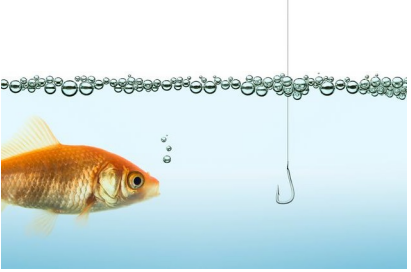
Unsupervised clustering for ~~neutrino~~ **cosmic** event exploration

Callum Wilkinson, LBNL

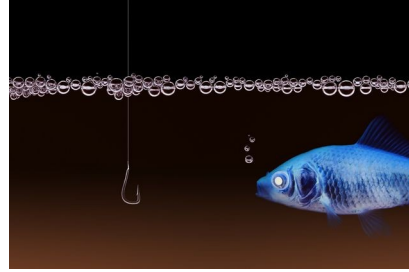
On behalf of the DUNE Collaboration

BERKELEY LAB

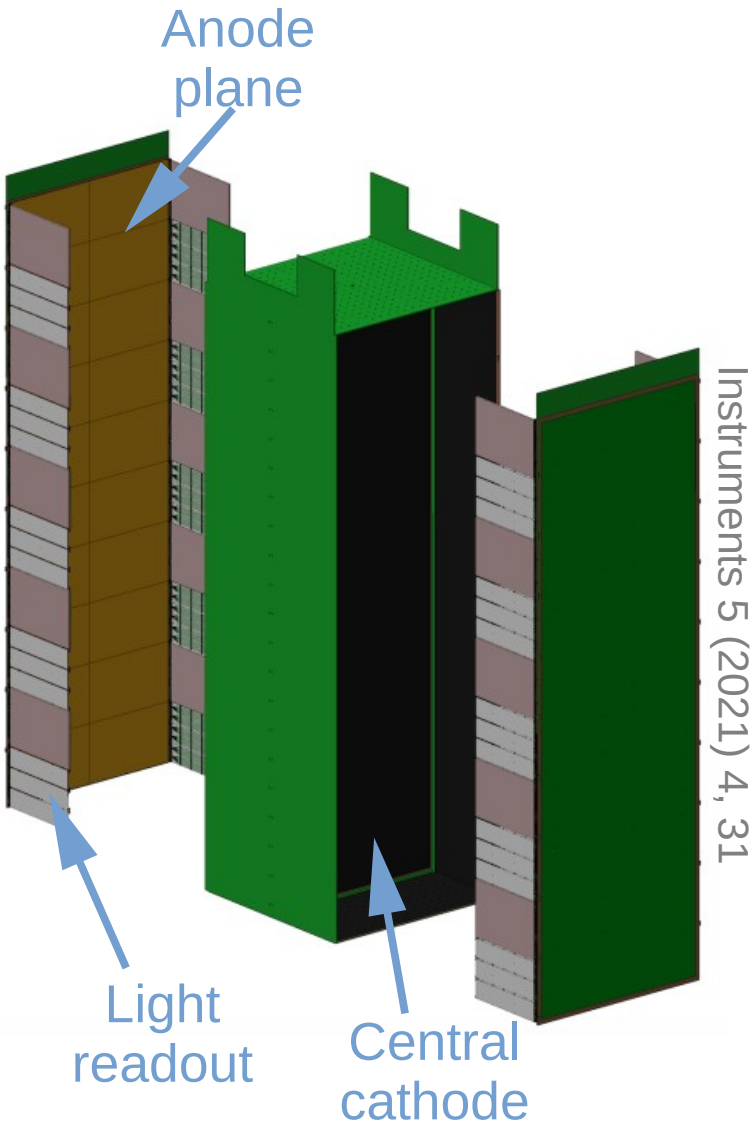
DUNE



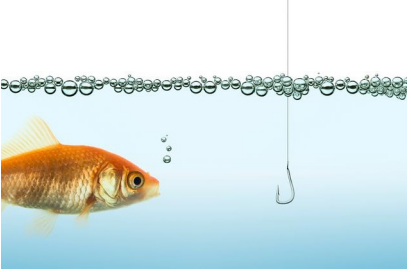
A pilot study with FSD data



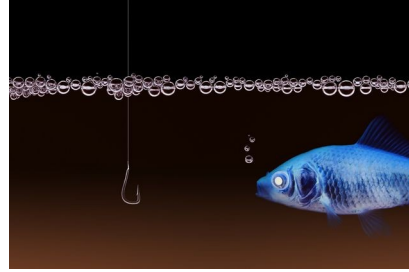
- Playing with neutrino data is *discouraged*, so this is a pilot study using <u>cosmic data</u>
- Data+simulation here is from the DUNE ND “full scale demonstrator” (FSD) prototype:
  - 3m x 1m x 1m LArTPC
  - ~4x4 mm pixel charge readout (anode)
  - Raw charge deposits projected onto the anode
  - Separated into events by timing
- **Task:** group visually or physically similar events to enable data-driven exploration



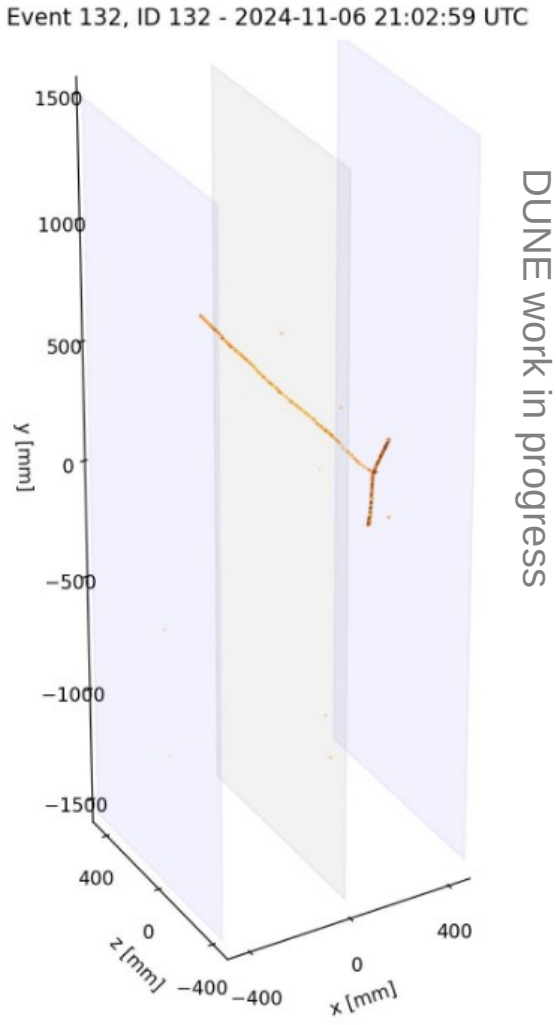
Instruments 5 (2021) 4, 31



A pilot study with FSD data



- Playing with neutrino data is *discouraged*, so this is a pilot study using cosmic data
- Data+simulation here is from the DUNE ND “full scale demonstrator” (FSD) prototype:
  - 3m x 1m x 1m LArTPC
  - ~4x4 mm pixel charge readout (anode)
  - Raw charge deposits projected onto the anode
  - Separated into events by timing
- **Task:** group visually or physically similar events to enable data-driven exploration



A pilot study with FSD data

- Playing with ne
  so this is a pilo
- Data+simulatio
  “full scale dem
  - 3m x 1m x 1m
  - ~4x4 mm pix
  - Raw charge
  - Separated in
- **Task:** group vis
  events to enable data-dr

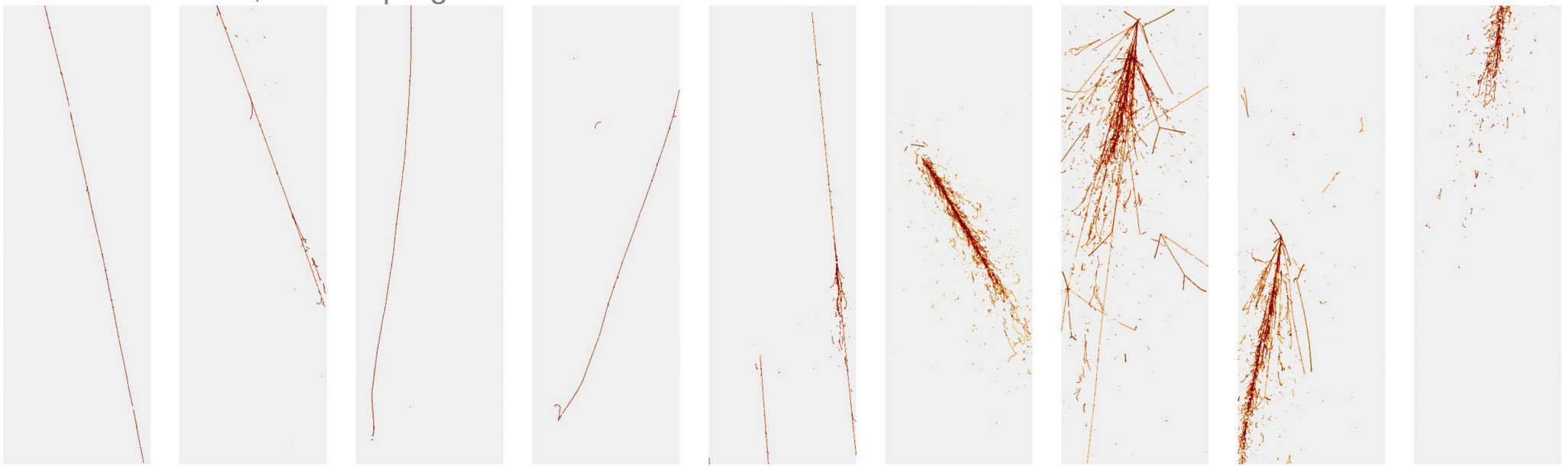
This work relies entirely on the collective work of lots of people!

Detector, data, simulation and software...

Example FSD cosmic simulation

- Example simulated cosmic events, with *very rough* truth labels
- I'm going to use the simulation later to see whether the data-driven clustering is doing "sensible" things
- But, these labels are not used in training at all

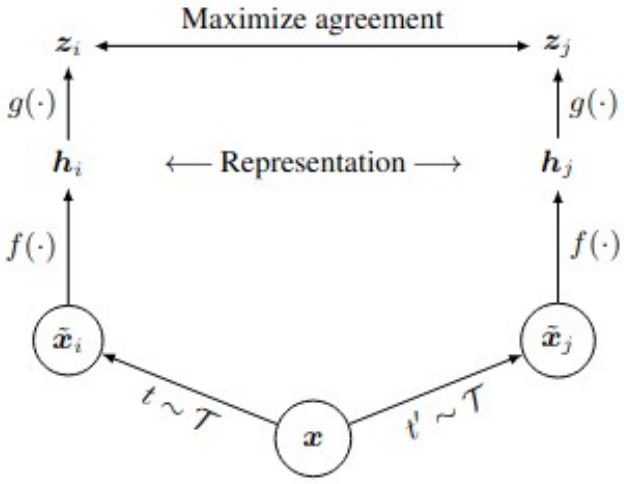
DUNE simulation, work in progress



Methods 1: SimCLR (arXiv:2002.05709)

General idea: train a model to distinguish between "similar" and "dissimilar" data without labels through augmentation invariance

1) Take a batch of N images
2) Augment each image twice → **positive pair**
3) Pass each image through an encoder
4) Project encoded features to a latent space with MLP
5) Use NT-Xent loss to pull **positive pairs** together, and use other images in the batch as **negative pairs**



$$\ell_{i,j} = -\log \frac{\exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}$$

Positive pair

Negative pairs

$$\mathrm{sim}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u}^{\mathrm{T}}\mathbf{v}}{\|\mathbf{u}\| \, \|\mathbf{v}\|}$$, τ is a "temperature" hyperparameter

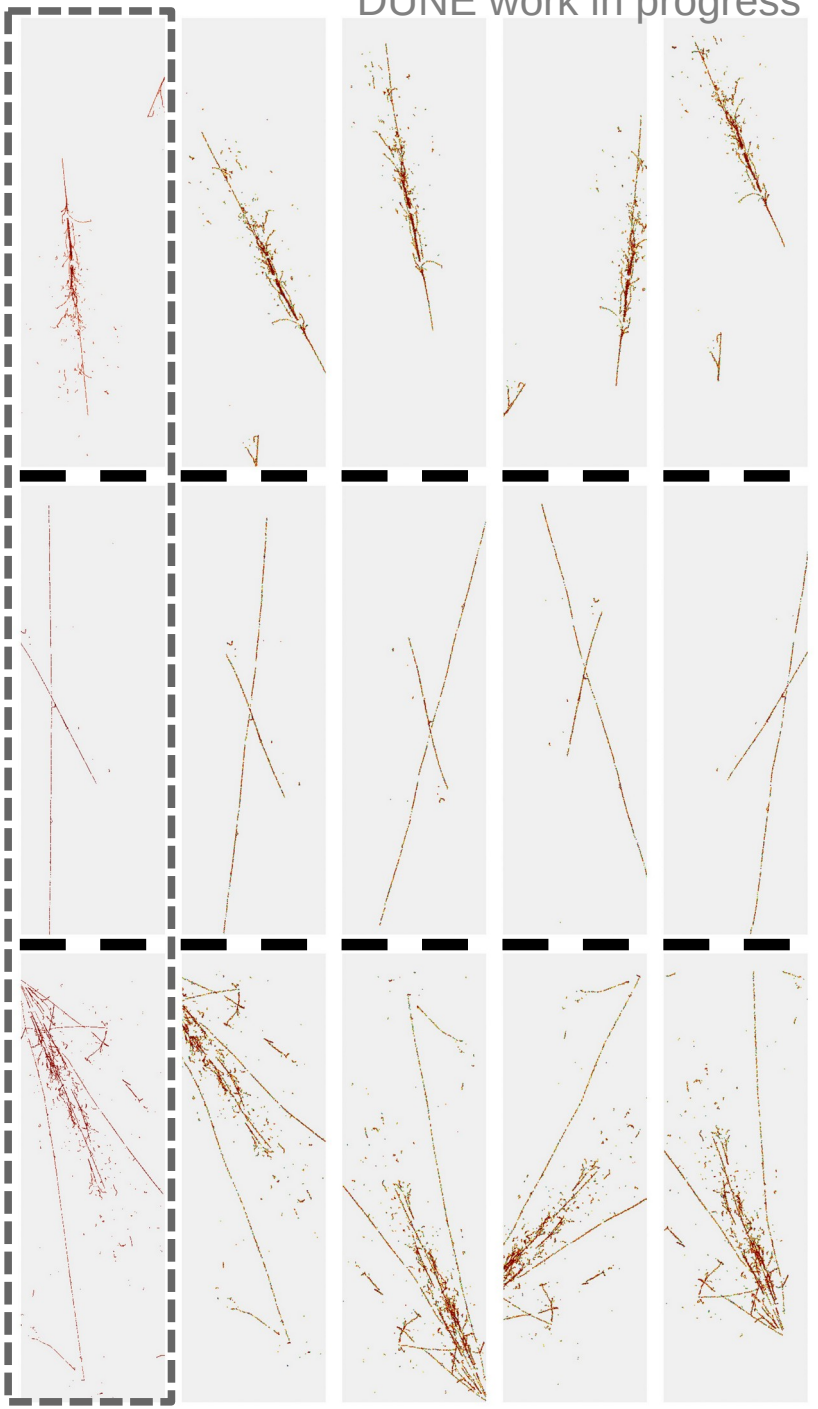
Augmentations

Aim: prevent the model from learning to use certain features, *e.g., rotational invariance through random rotations*

Challenges: FSD aspect ratio and the large extent of many images

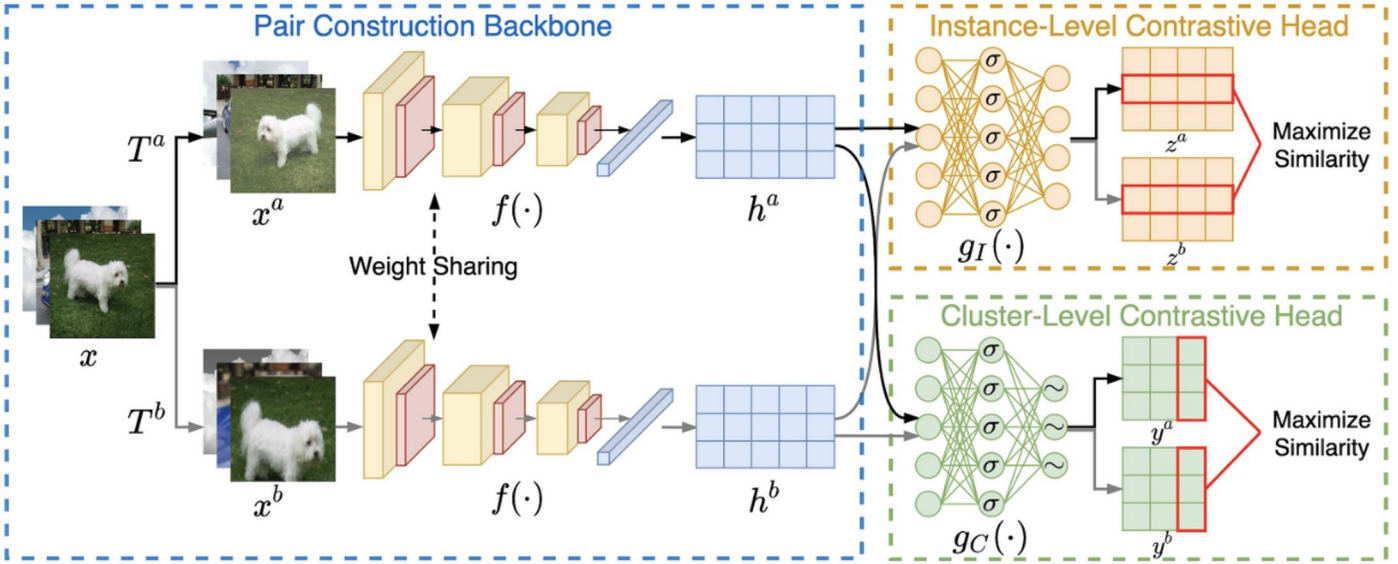
Use a variety of operations including skews, flipping, rotations, distortions, dropping blocks of pixels, …

Full list in backups, I'm hiding a ***lot*** of trial and error here



Original

Methods 2: Contrastive *Clustering* (arXiv:2009.09687)



arXiv:2009.09687

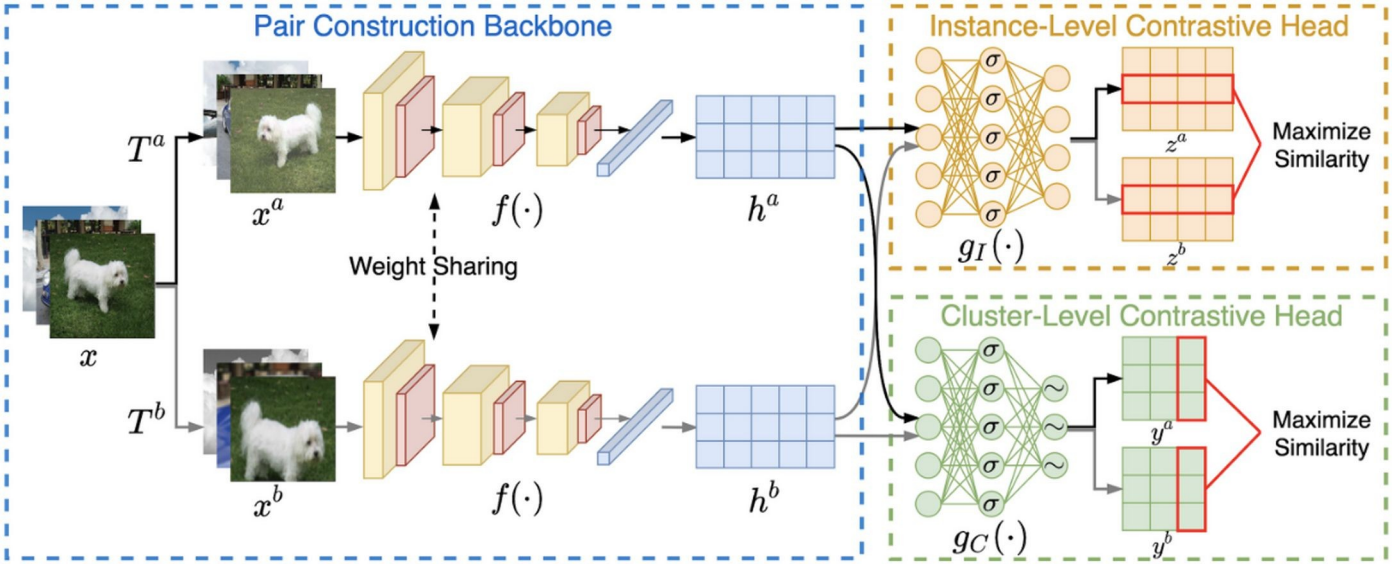
So far so SimCLR!

Extends SimCLR with a *clustering* head which produces soft *cluster* assignments, + compares those across the two augmented batches

(Two-layer) MLP that takes output from the CNN
→ 64D latent space

NT-Xent term in the loss

Input two augmented batches of images



arXiv:2009.09687

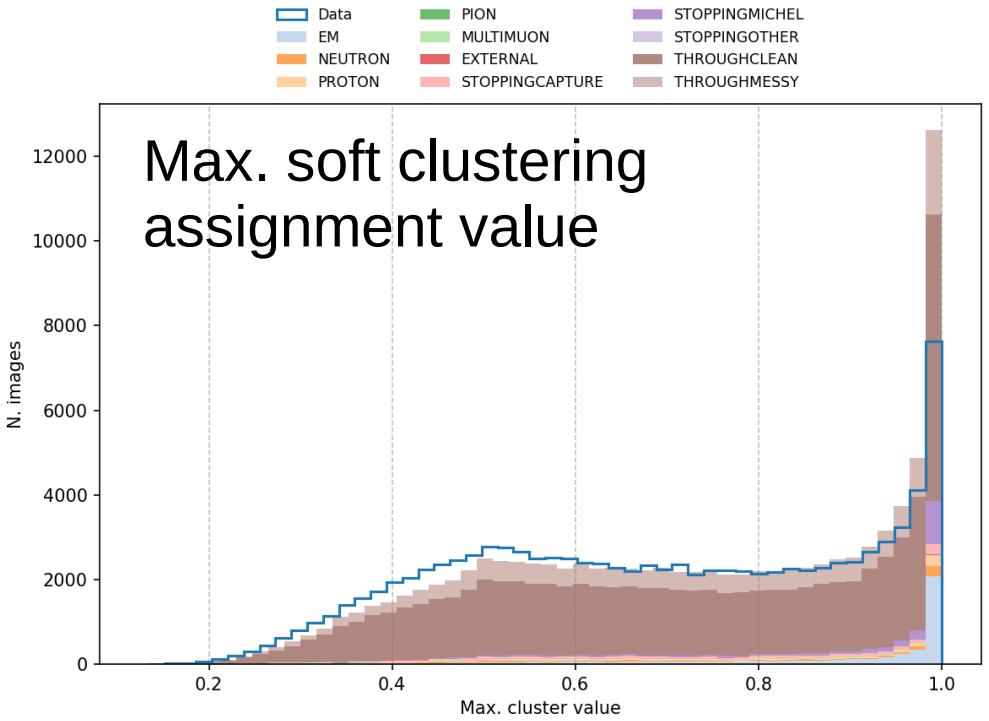
Sparse CNN with (6x) 3x3 downsampling and fully convolutional layers
→ Global pooling for output

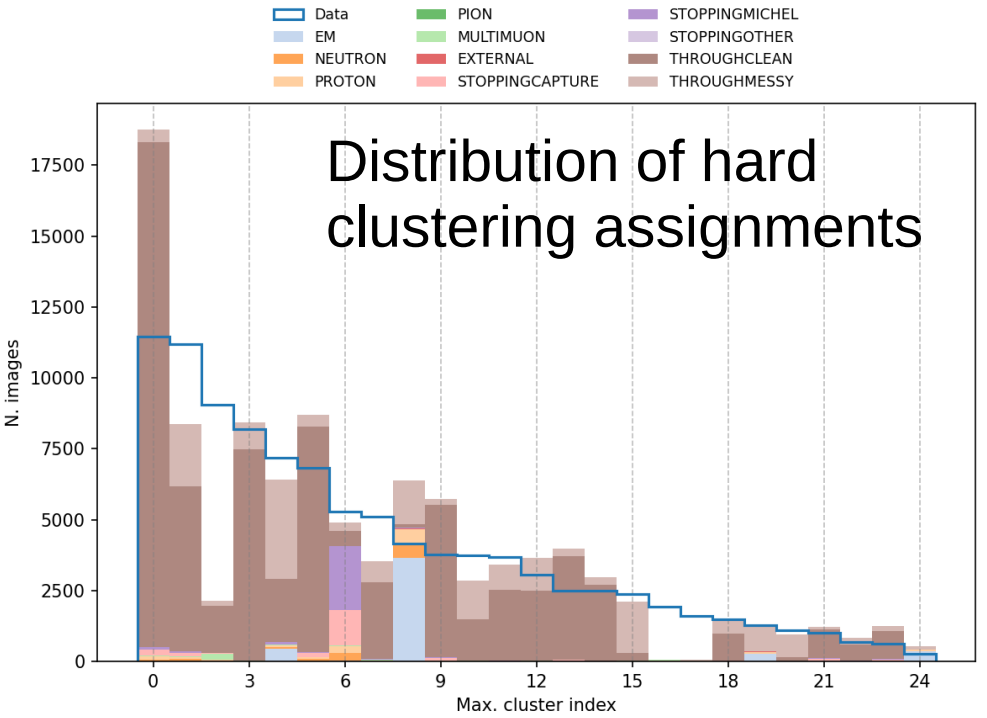
(Probably deeper than it needs to be)

(Two-layer) MLP + softmax, outputs soft cluster assignments

Adds NT-Xent + small entropy term to the loss

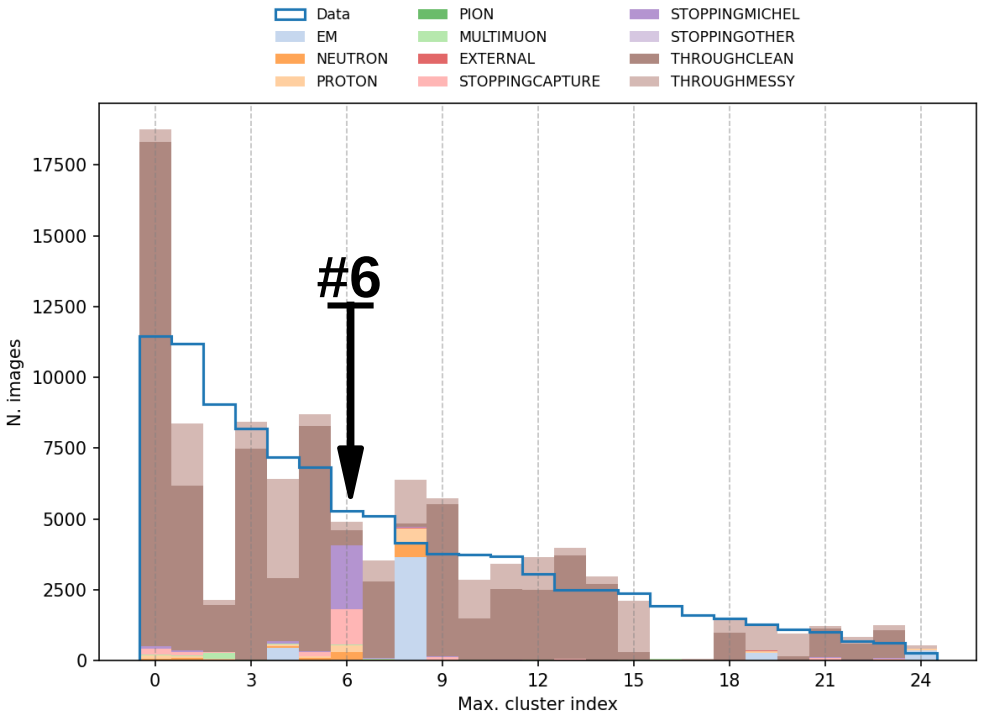
Training results (N_{cluster} = 25)



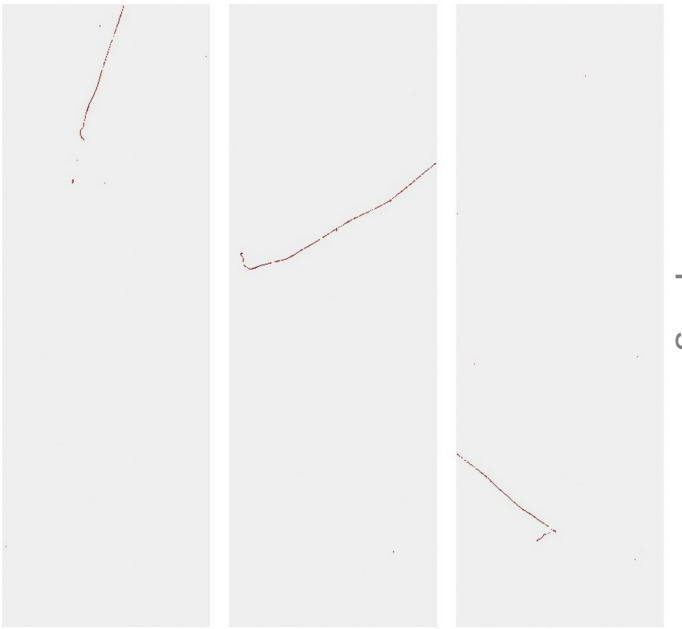


- Trained with 5M ***data*** images (possibly overkill)
- After training, pass 100k new data and 100k simulation images through the encoder + clustering head
- Hard cluster index = max soft clustering assigment

Clustering results: the good



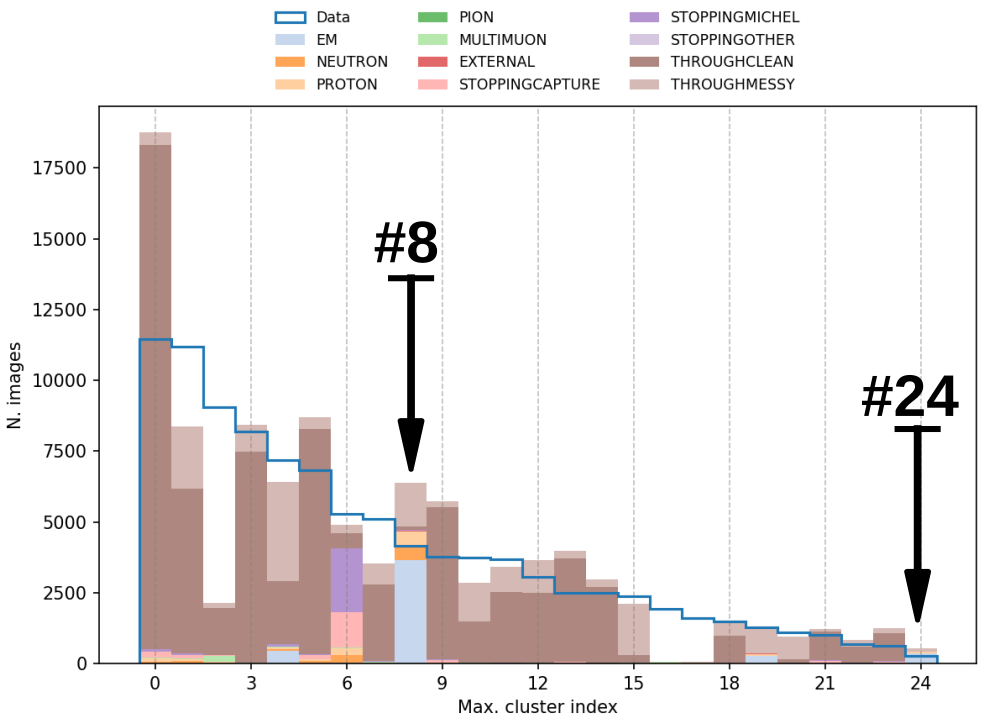
#6 Data events



DUNE work in progress

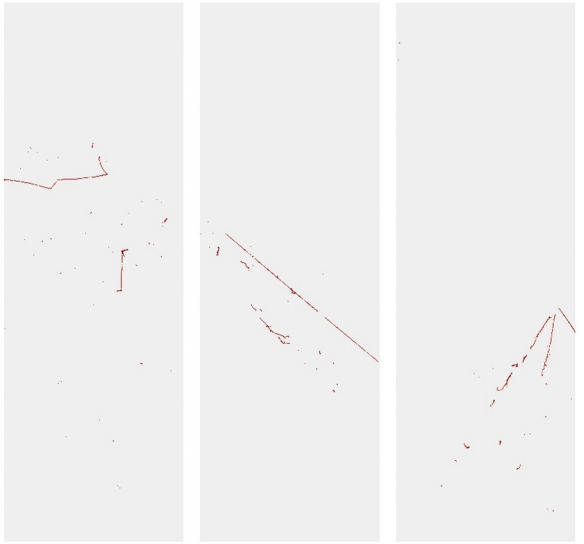
- **Data** cluster 6 *appears* to correspond to stopping muons in simulation
- Randomly chosen examples from data look consistent with that interpretation

Clustering results: the good (2)

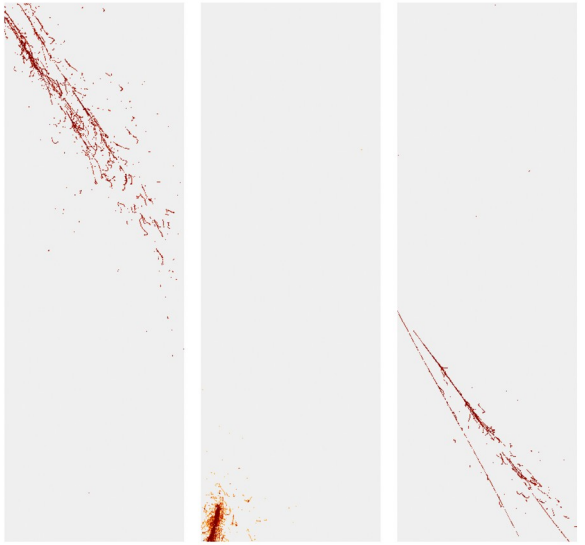


- *Most* non-muon induced events in simulation appear to correspond to two data clusters
- Those clusters also appear to have meaningful differences

#8 Data events



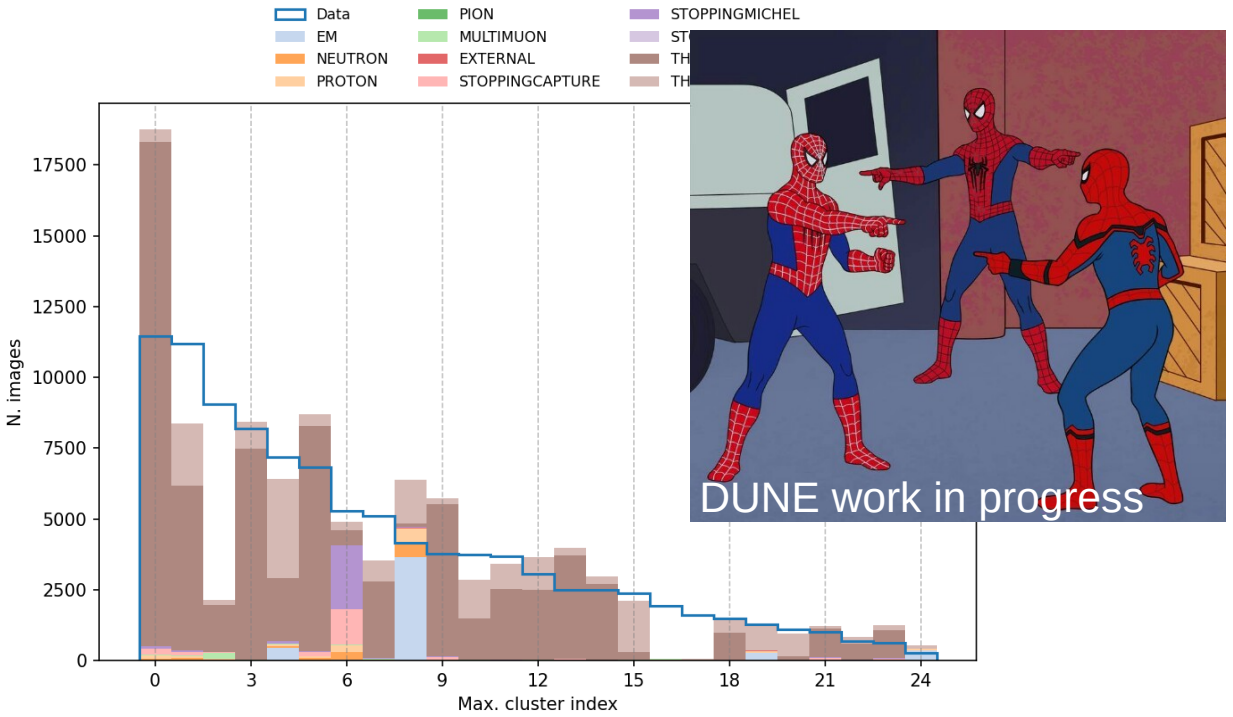
#24 Data events



DUNE work in progress

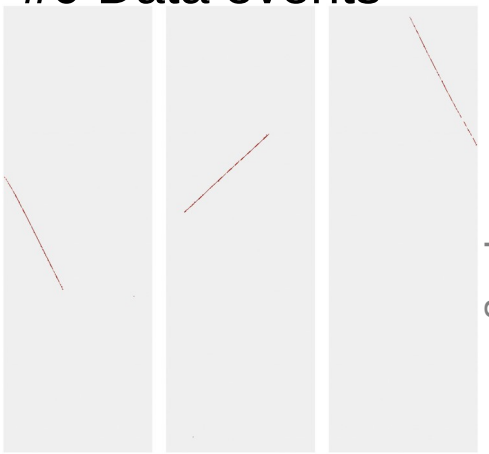
DUNE work in progress

Clustering results: the bad

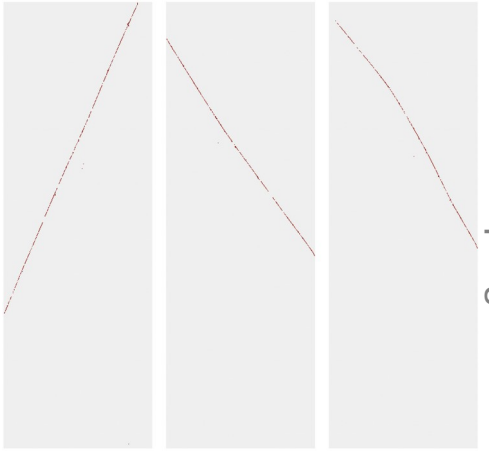


- “Clean” through-going simulation tracks correspond to multiple data clusters
- Generally these clusters have less confident assignments
- My guess is that it’s due to the imbalance between event types

#0 Data events

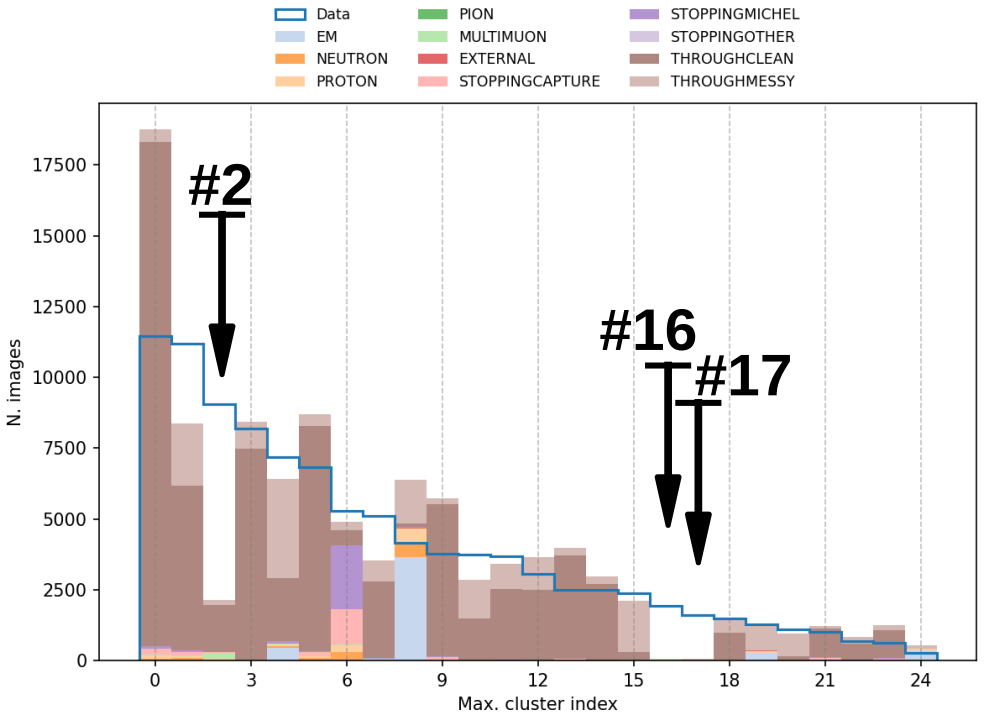


#7 Data events



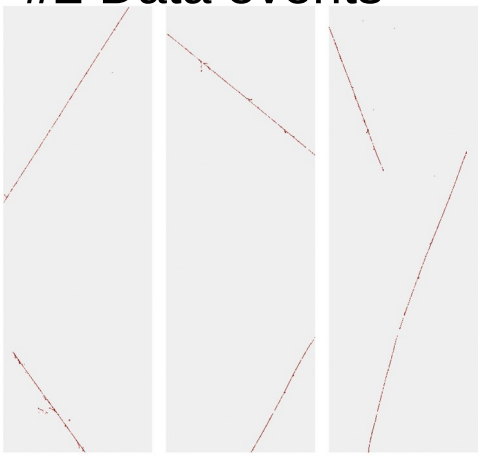
#13 Data events

Clustering results: the interesting



- Some data clusters don’t seem to correspond to simulation labels
- Likely caused by event separation failures in data, to be investigated…
- But a fun observation from this method

#2 Data events

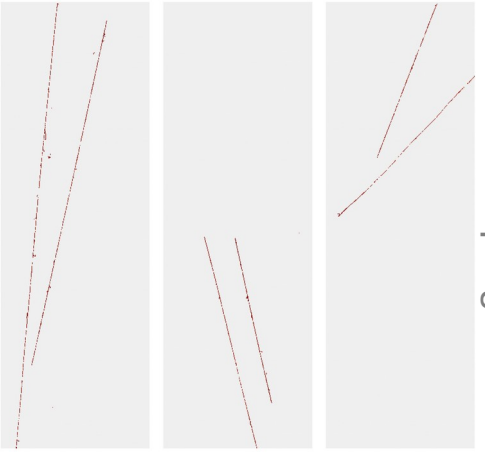


DUNE work in progress

#16 Data events

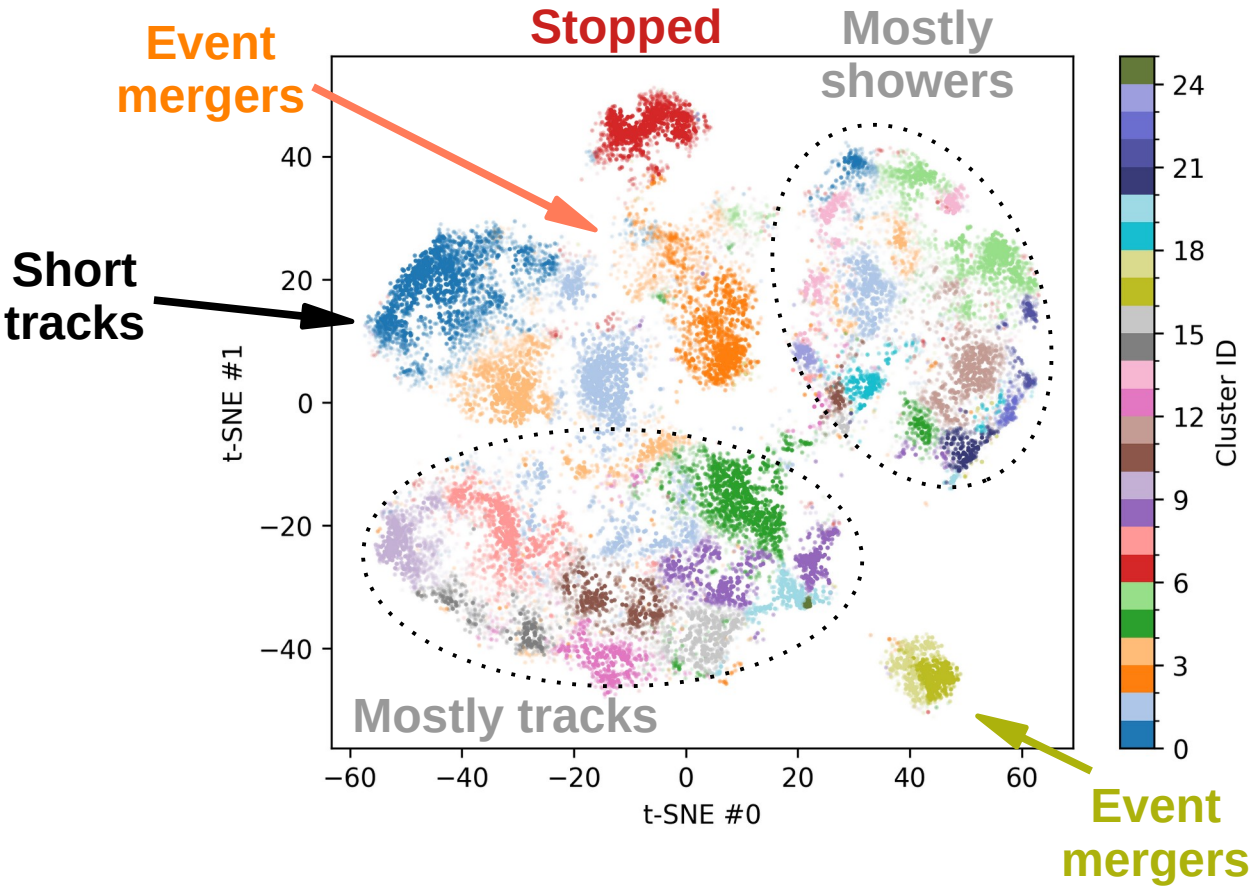
DUNE work in progress

#17 Data events



DUNE work in progress

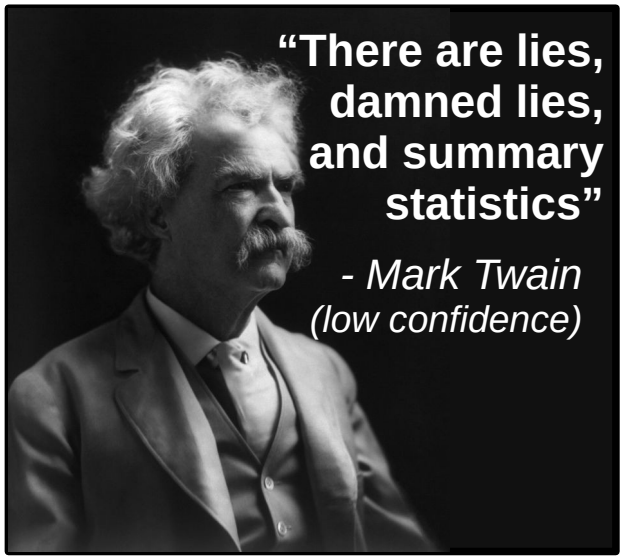
T-SNE for comparison



- 64D instance contrastive head output → 2D
- Colourscale from clustering head hard assignments

Outlook

- Investigated using unsupervised clustering for *event exporation*
- Qualitative assessment:
  - Clear clusters of different “event types”
  - Including some interesting surprises
  - All generally sensible
- But, quantitative metrics are a challenge, lots of trial and error involved



- **Method improvements:** improved clustering techniques exist, refined architecture/augmentations, 3D images...
- **Physics improvements:** move to neutrino-scattering simulation
- Suggestions very welcome!

Questions?


Corporate needs you to find the differences between this picture and this picture.
They're the same picture.

Missing data: nuclear de-excitation

- Liege intranuclear cascade (INCL) +ABLA nuclear de-excitation model
- Significant fraction of E_ν goes to very low energy cluster production
- Energy loss to clusters increases with nuclear size (only ^{16}O shown)

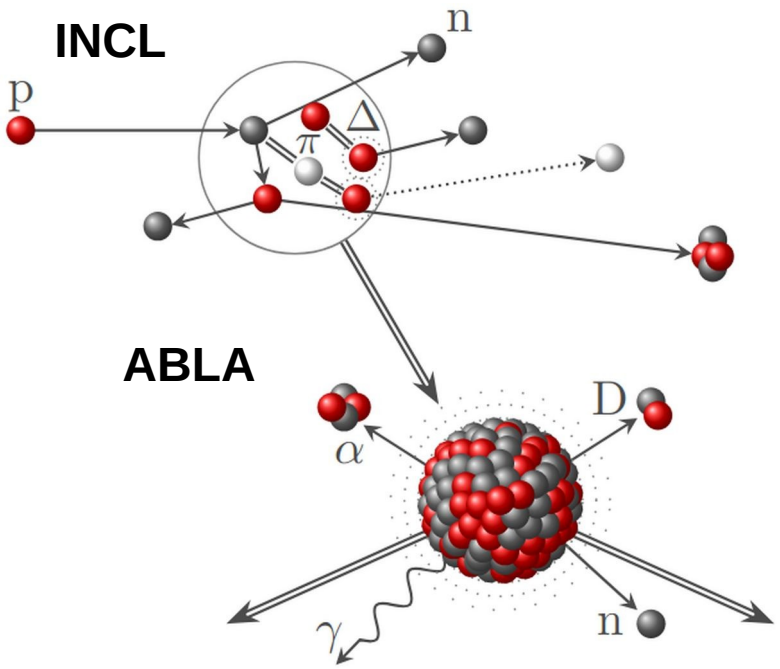
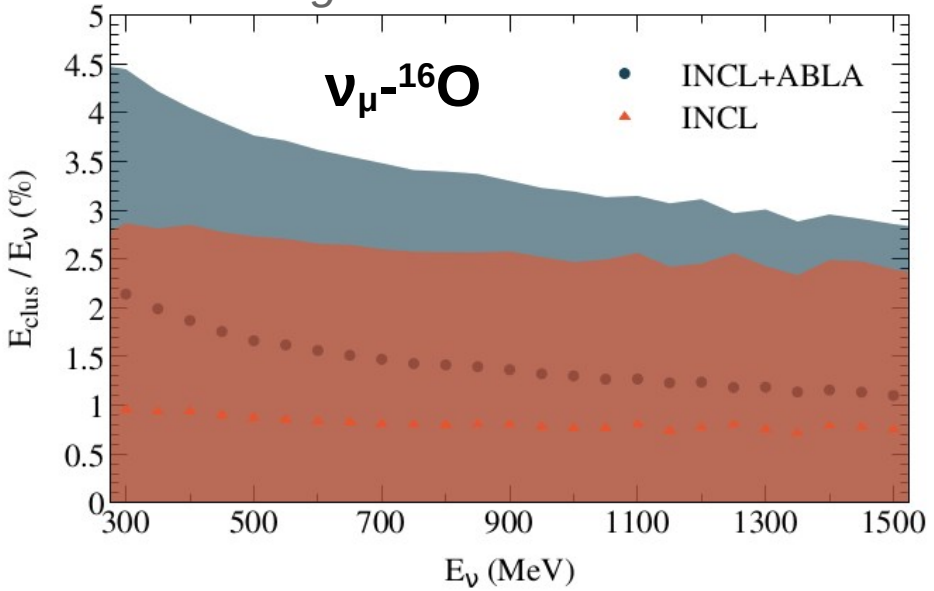


Figure from A. Ershova

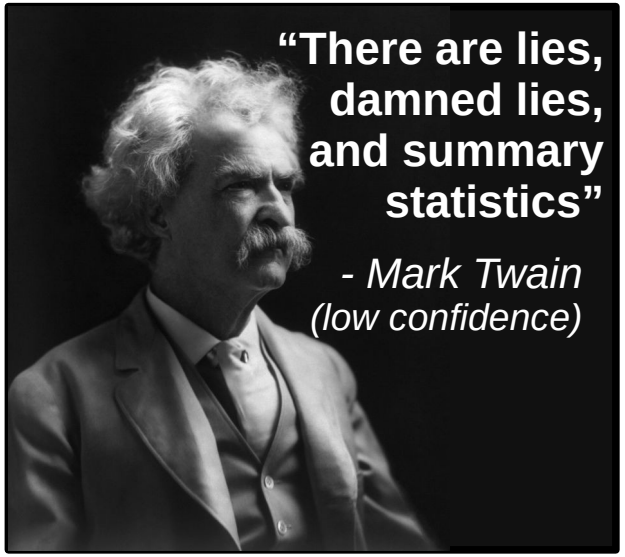


- Bias to both Hyper-K and DUNE energy reconstruction…
- No available data constraints, limited theory inputs, but a major impact

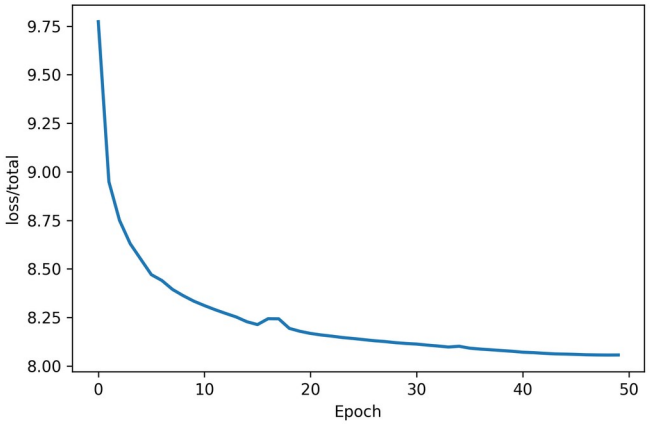
Phys. Rev. D108, 112008
Phys. Rev. D106, 032009

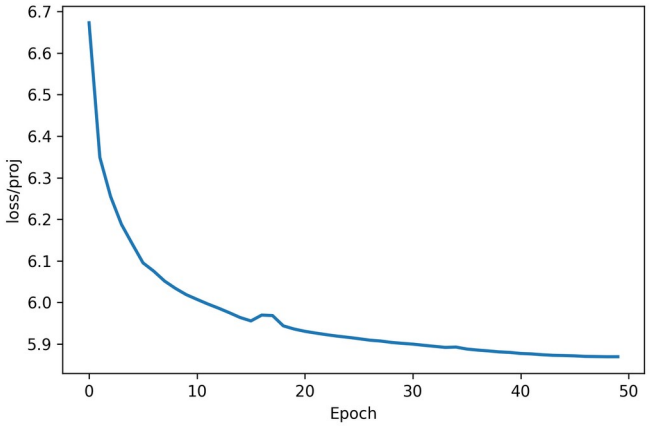
Assessing clustering performance

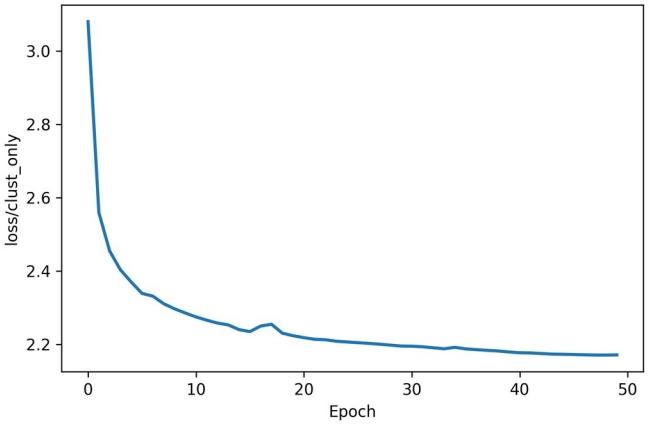
- There are many tunable hyper parameters, the most important are:
  - The augmentations used
  - The number of clusters
  - Strength of entropy term
  - Temperatures in the two NT-Xent contributions to the loss
- But, *quantitative* tuning of these is difficult. Unlabeled cluster quality metrics exist, but can give misleading results:
  - Weak augmentations improve all metrics… because the model learns angular information
  - Small N. clusters perform better on all cluster separation metrics → but qualitatively fail completely
- Ideas welcome...

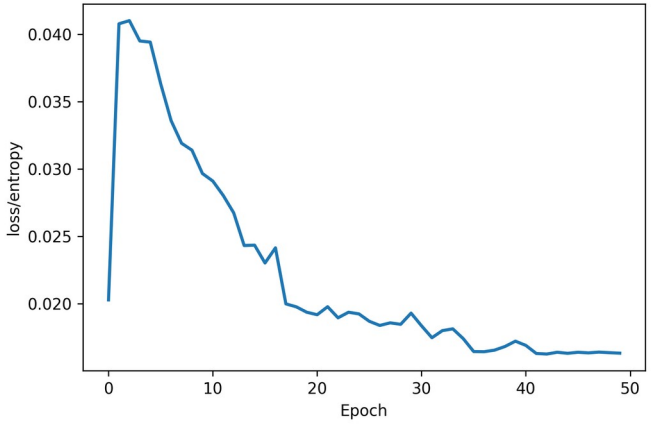


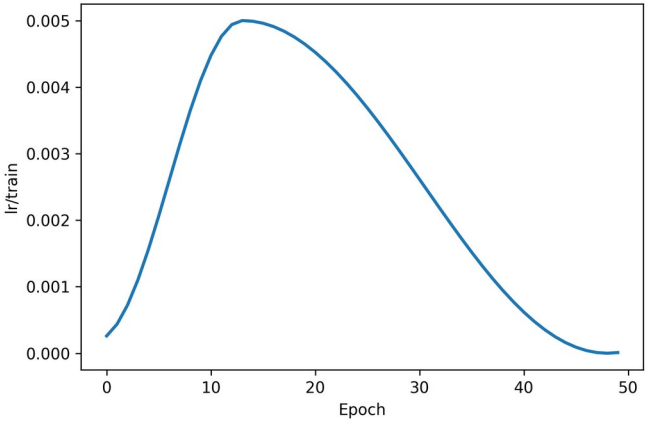
Training plots











5 million events, ~10 hours on 4 A100s

What is the CNN encoder structure?

To paraphrase: “ChatGPT please make this code dump into a nice table showing the model structure”:

Visual / Table summary

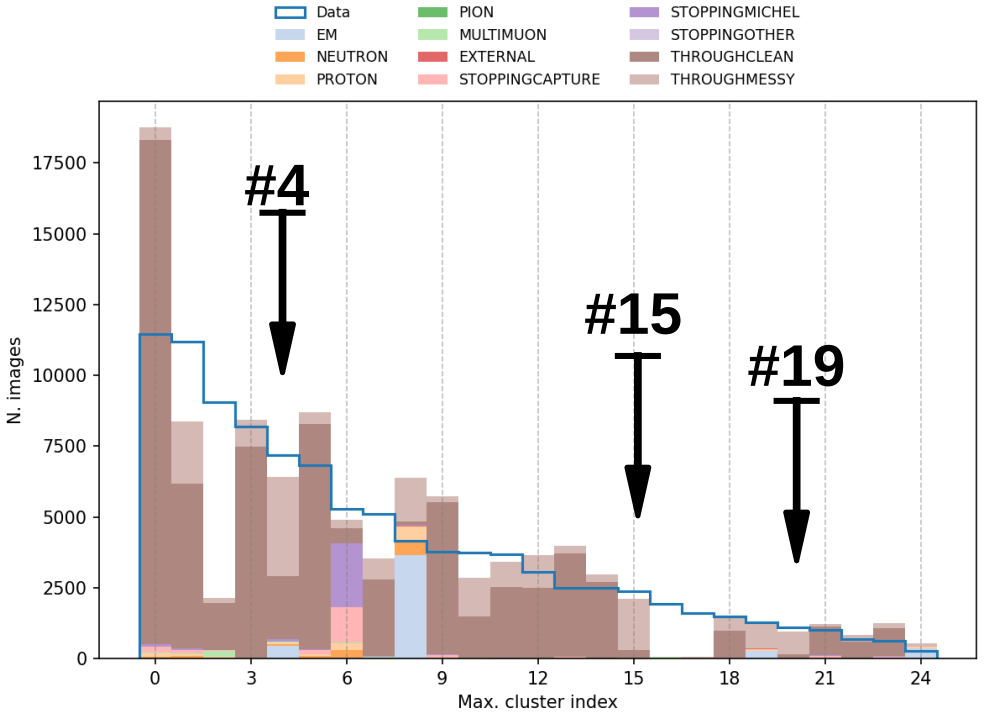
Stage	Operation	Channels	Spatial size (HxW)	Notes
Input	Sparse tensor	1	768×256	—
Block 1	Conv3×3 (stride 2), 2×Conv3×3	48	384×128	Initial downsample
Block 2	Conv3×3 (stride 2), 2×Conv3×3	96	192×64	—
Block 3	Conv3×3 (stride 2), 2×Conv3×3 + BN	192	96×32	—
Block 4	Conv3×3 (stride 2), 2×Conv3×3 + BN	192	48×16	—
Block 5	Conv3×3 (stride 2), 2×Conv3×3 + BN	192	24×8	—
Block 6	Conv3×3 (stride 2), 2×Conv3×3 + BN	192	12×4	Output features

Augmentation procedure

Series of operations which randomly:

- Zero out blocks of varying size:
  - 0–50 blocks of 5–10 x 5–10 pixels
  - 500–2000 blocks of 1–3 x 1–3 pixels
- Mirror image in x/y both in-place and globally
- Image shear in x and y (Gaus, $\sigma = 0.1$)
- Rotation (Gaus, $\sigma = 6°$)
- Grid distortion:
- Scale image (Gaus, $\sigma = 10\%$)
- Scale all + individual charges:
  - +/-2% on the raw charge (Q)
  - +/-2% after taking $\log_{10}(1+Q)$
- Splat onto regular grid
- *Semi*-random crop to 768x256

Clustering results: the good again



- “Messy” through-going tracks are mostly sensibly grouped together
- The additional separation in data seems sensible, simulation labels very coarse