# Early use of ML for NR/ER discrimination from raw pixel data

Giuseppe Maria Oppedisano

CYGNO Experiment

# Intro

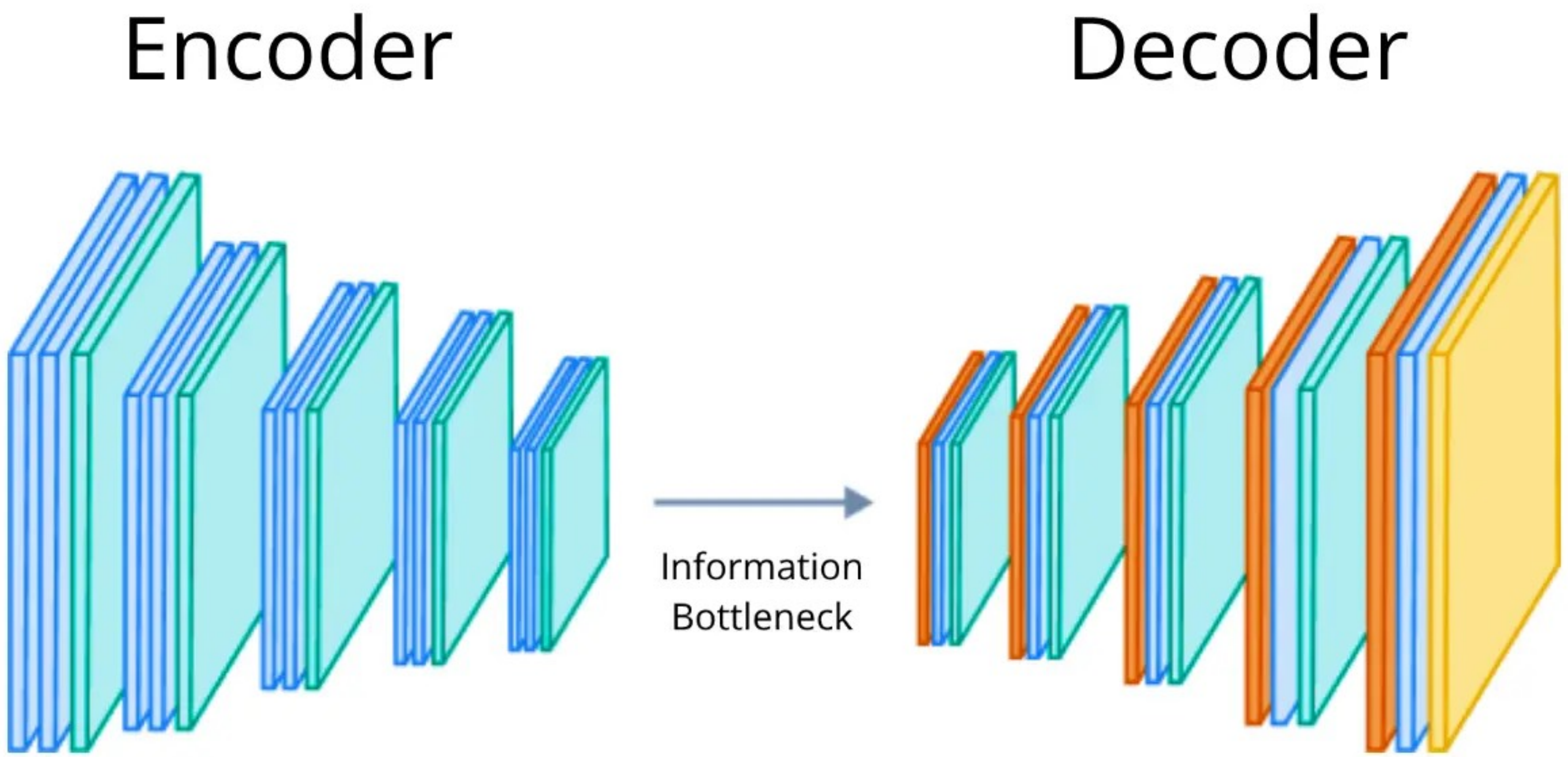
## Why ML

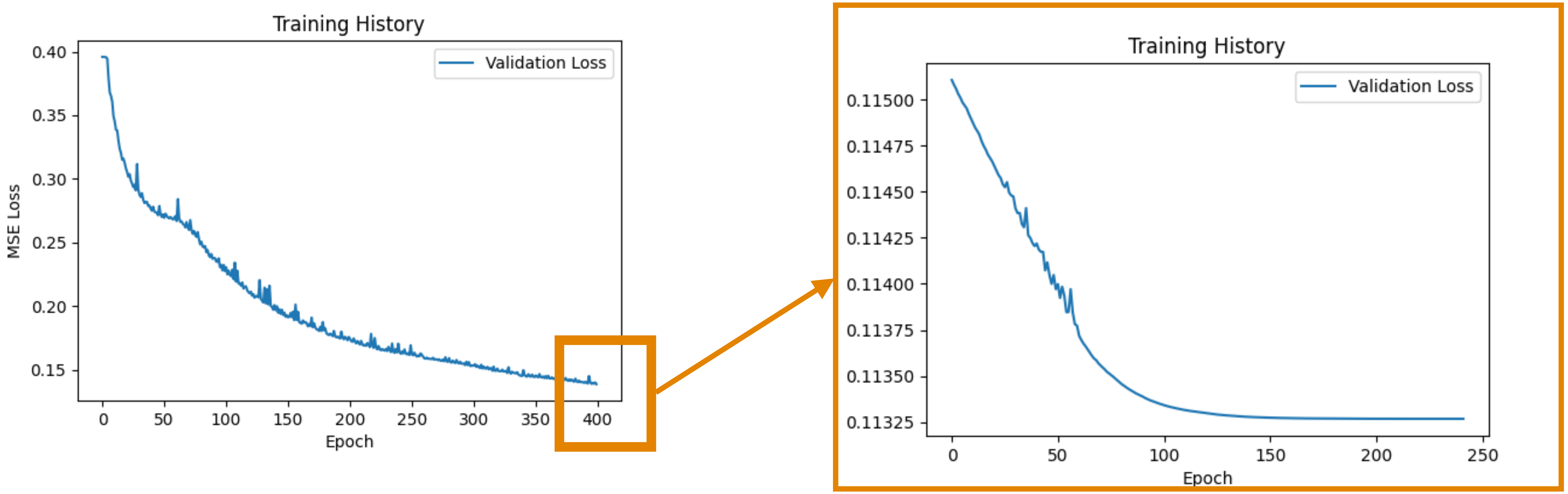
- We have lots of data at our disposal (unlabelled)
- Goal: Capturing (and later discriminating) different kinds of events
- Advantages: fast inference time, possibility to fine tune on different configurations of the detector
- Possibility to add montecarlo events to the dataset to increase performance

# The Idea

- Train an autoencoder on pedestal images (at the moment: 1 pedestal run)

- In test mode: send tracks with pictures
  ⟶ higher reconstruction error
  ⟶ anomaly





Train Time:
O(5hrs)

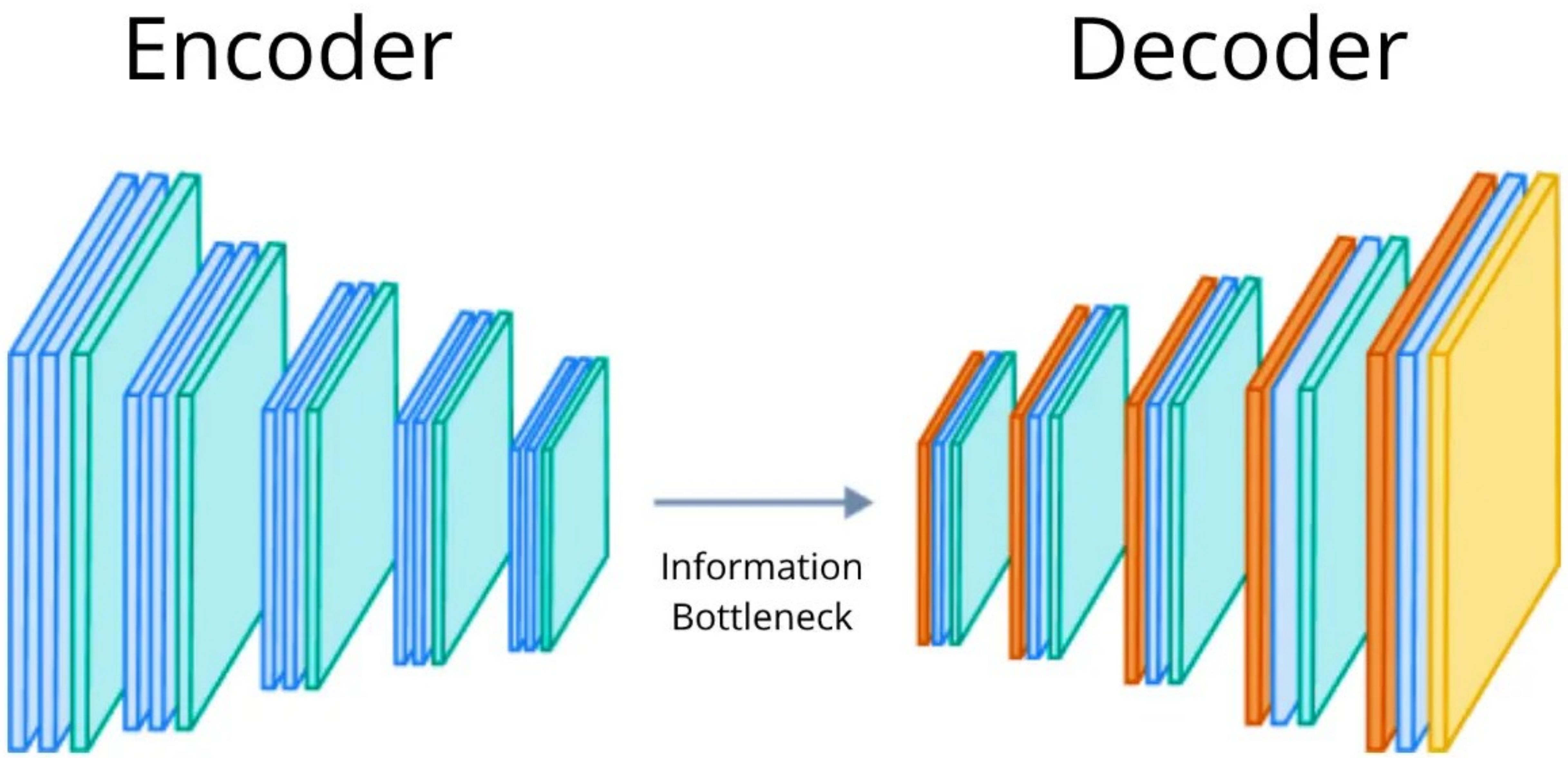
Inference time:
O(20ms)/image

GPU used:
M1 Pro 16Gb

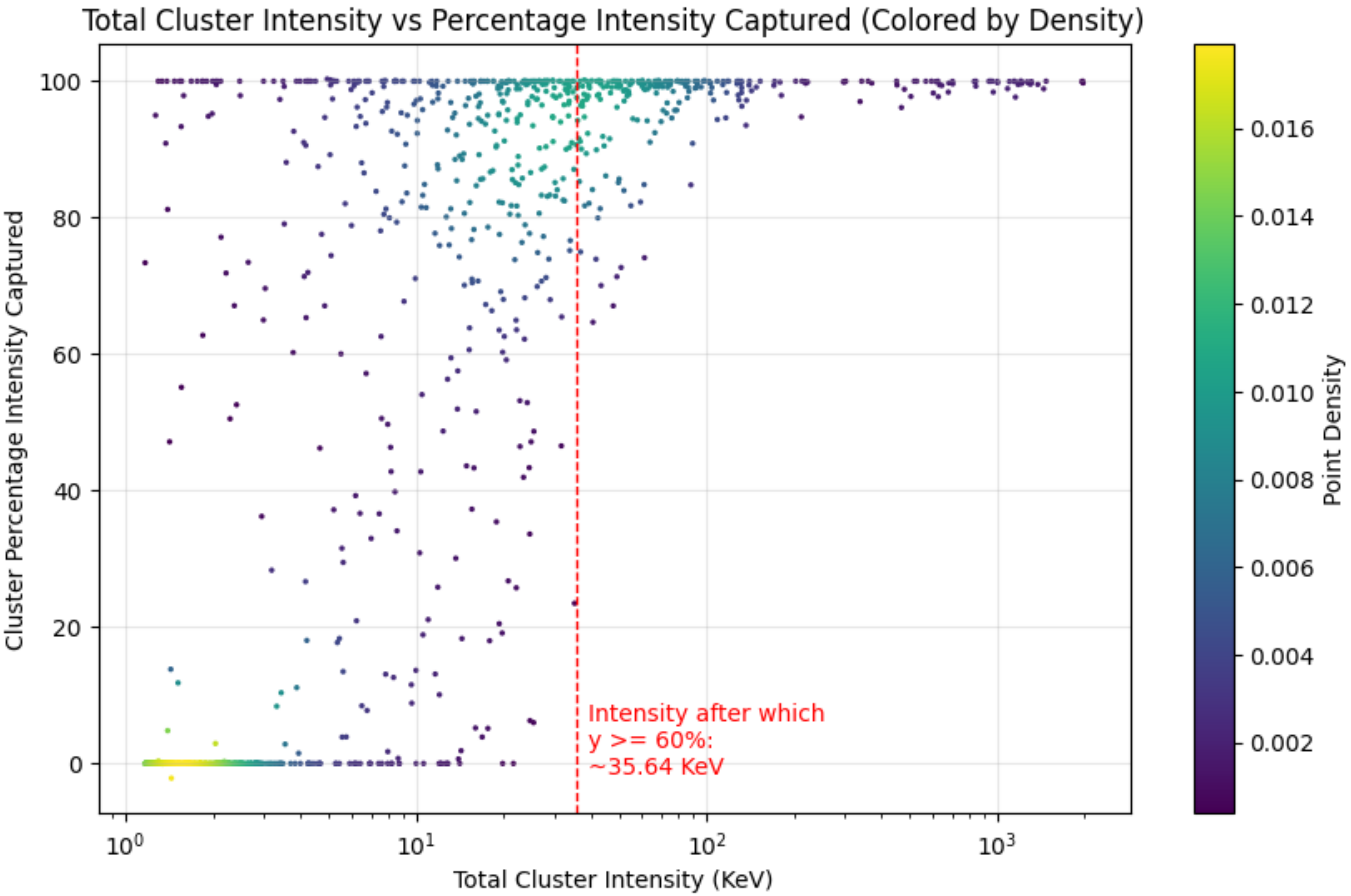
# The Idea

- Train an autoencoder on pedestal images (at the moment: 1 pedestal run)

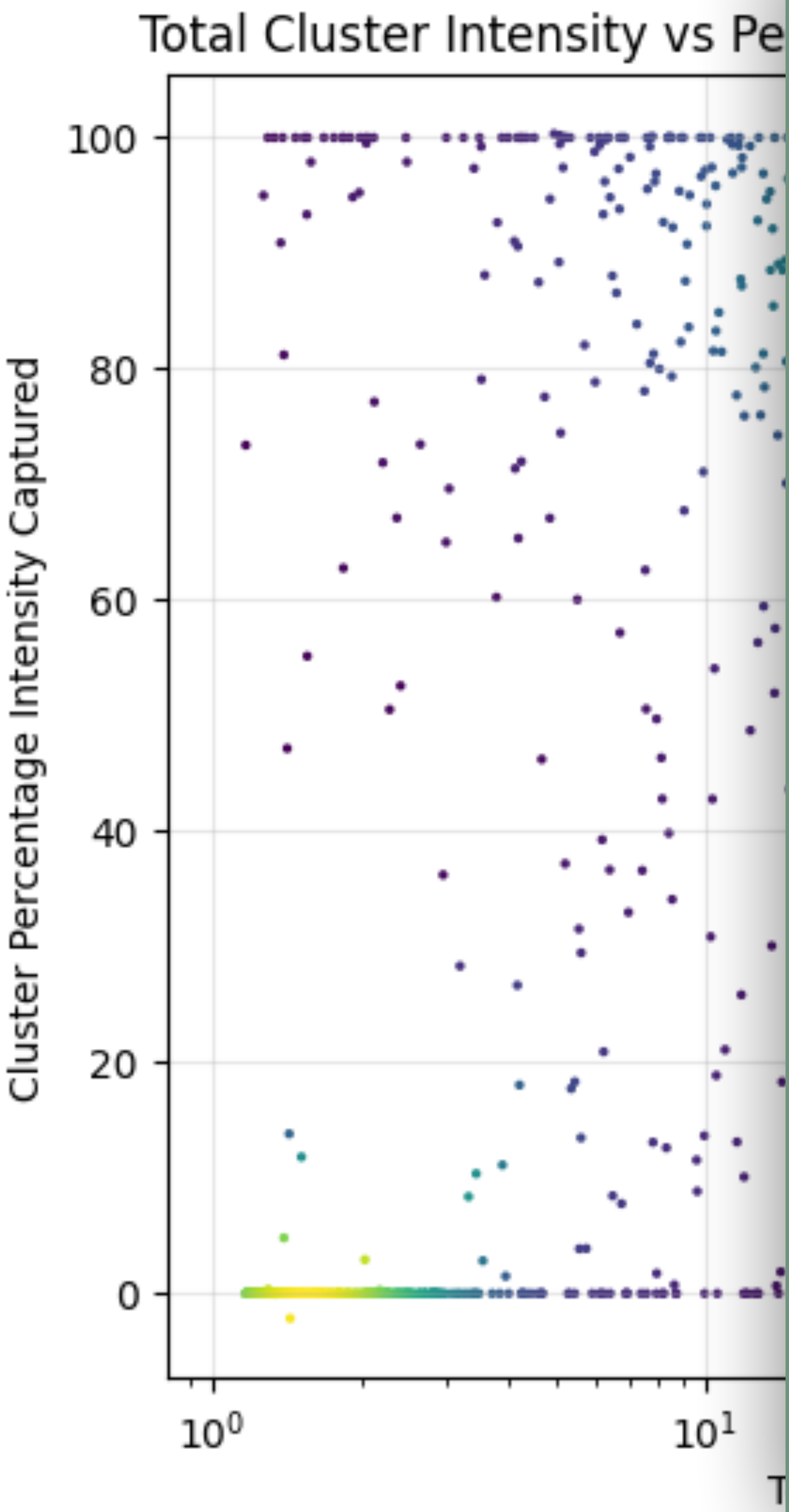
- In test mode: send tracks with pictures
  ⟶ higher reconstruction error
  ⟶ anomaly

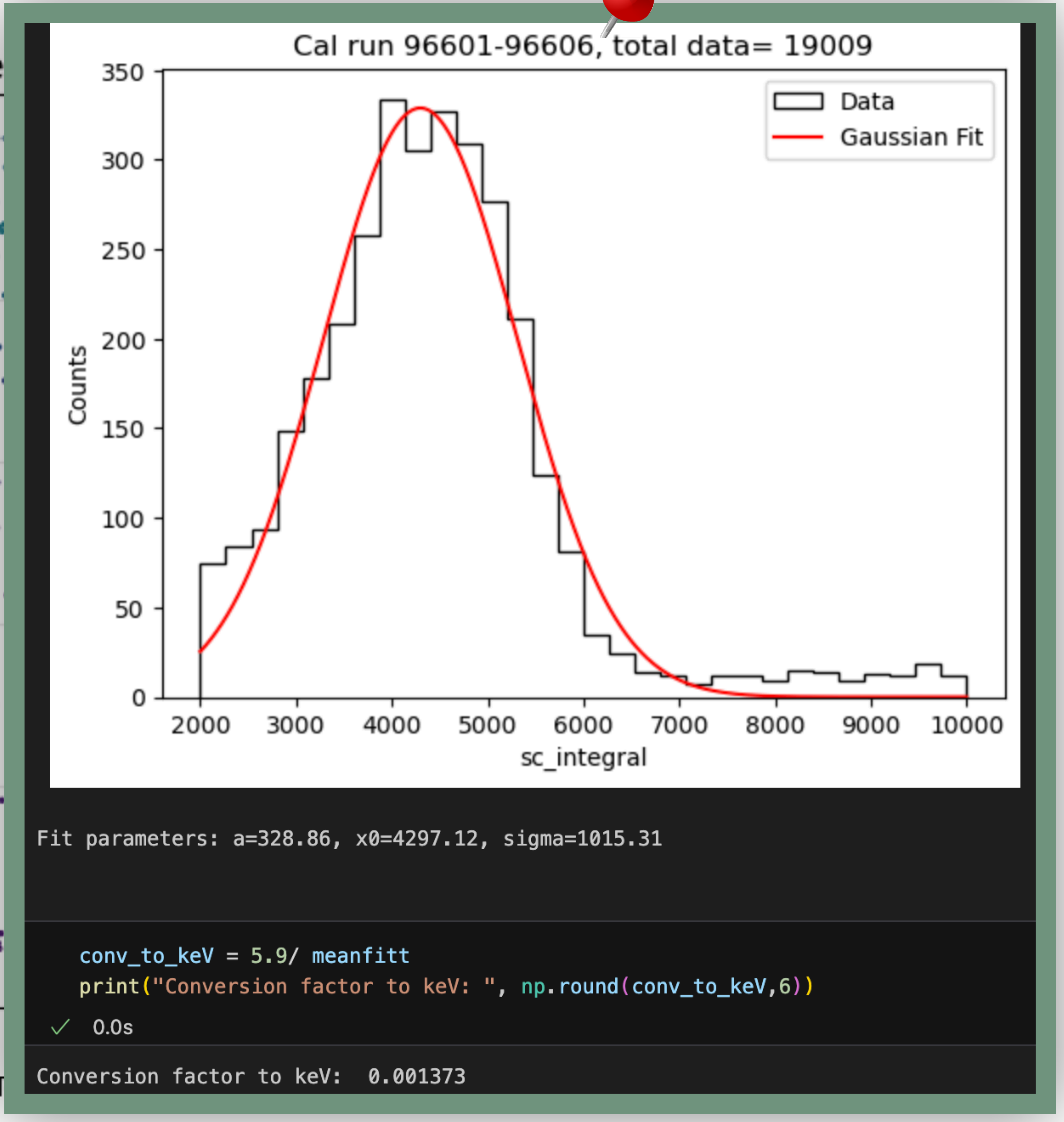


# Redpix Comparison



# Redpix Comparison





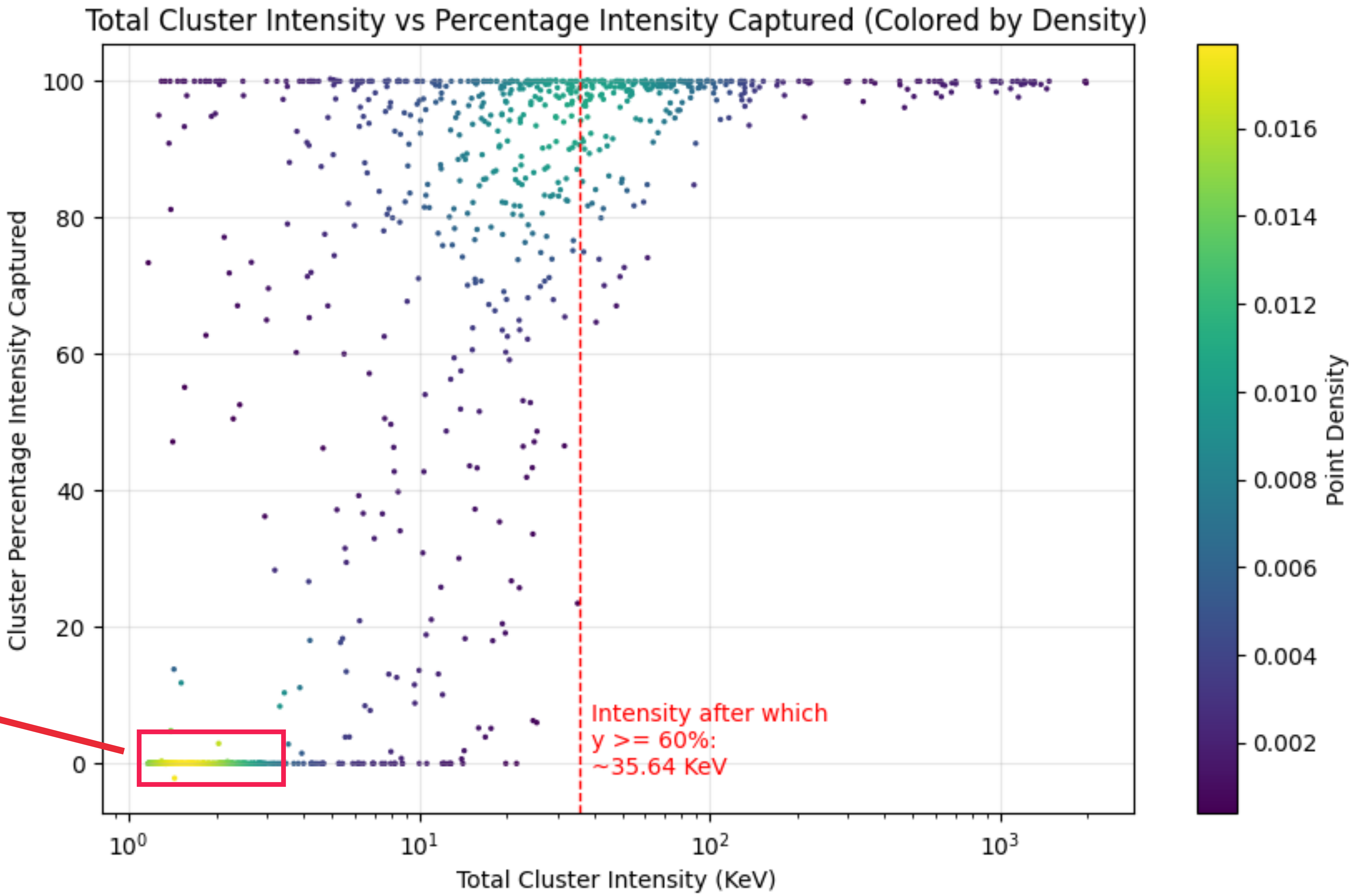
```
Fit parameters: a=328.86, x0=4297.12, sigma=1015.31
```

```
conv_to_keV = 5.9/ meanfitt
print("Conversion factor to keV: ", np.round(conv_to_keV,6))
```

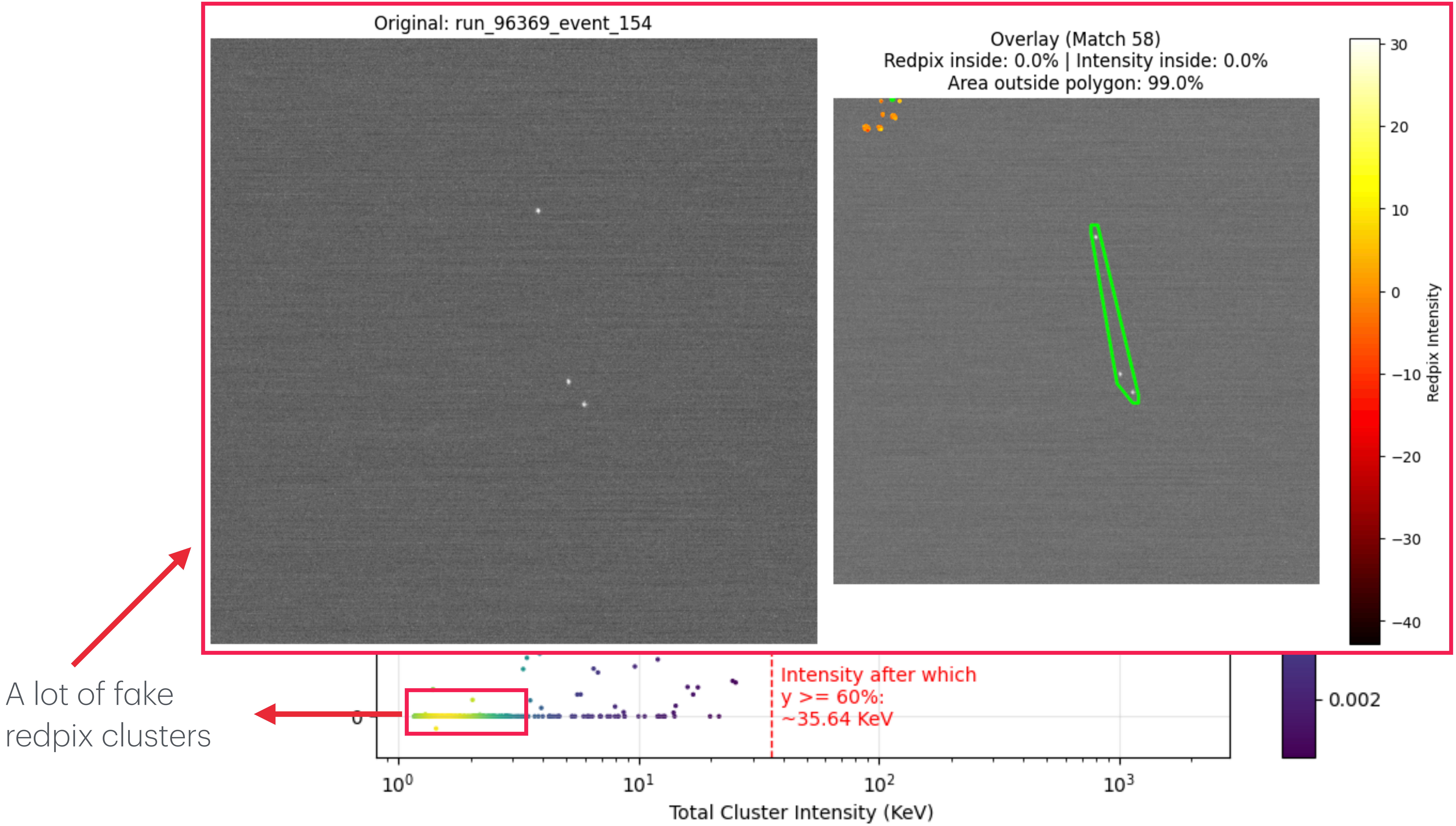
✓ 0.0s

```
Conversion factor to keV:  0.001373
```

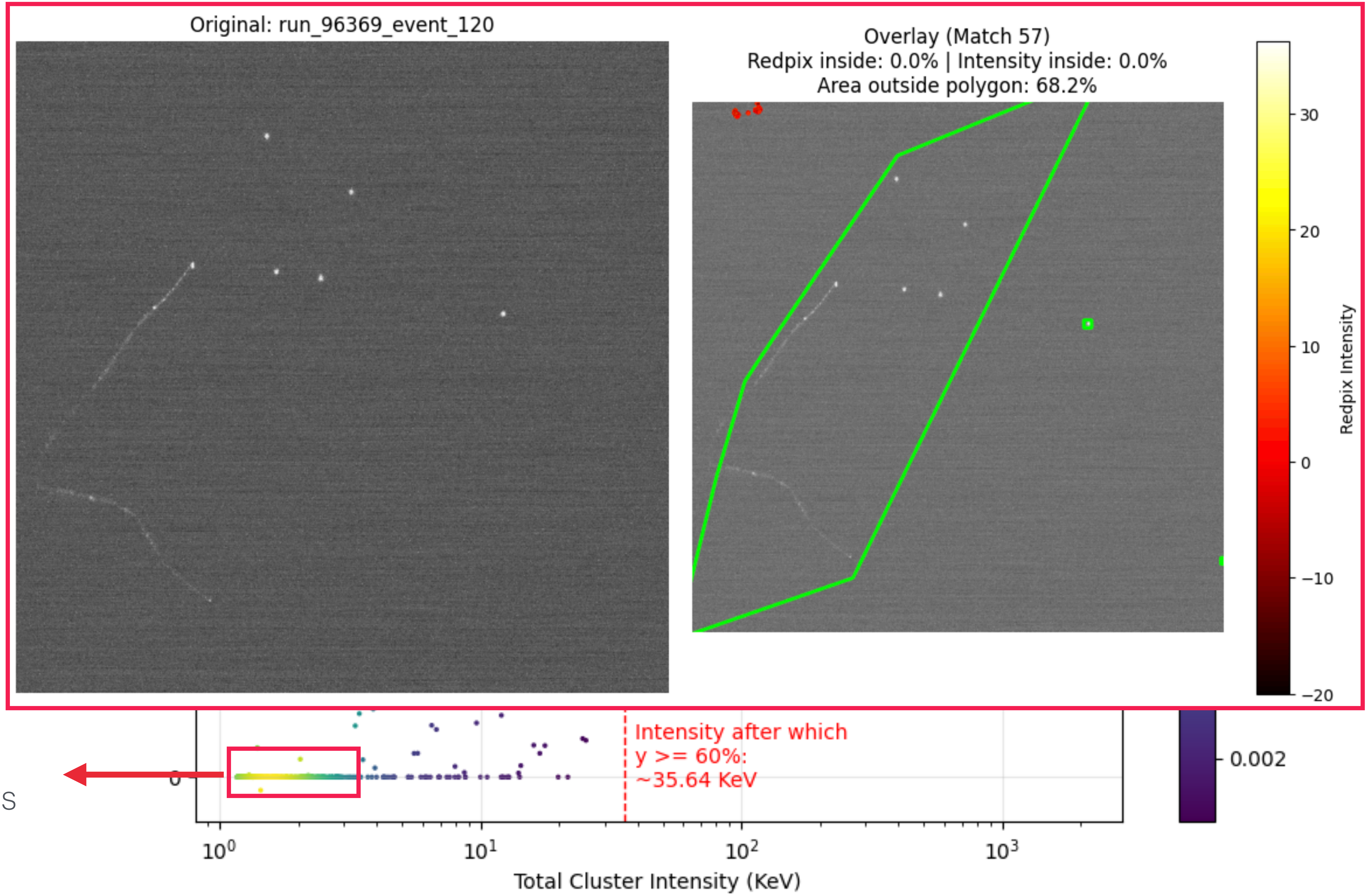
# Redpix Comparison



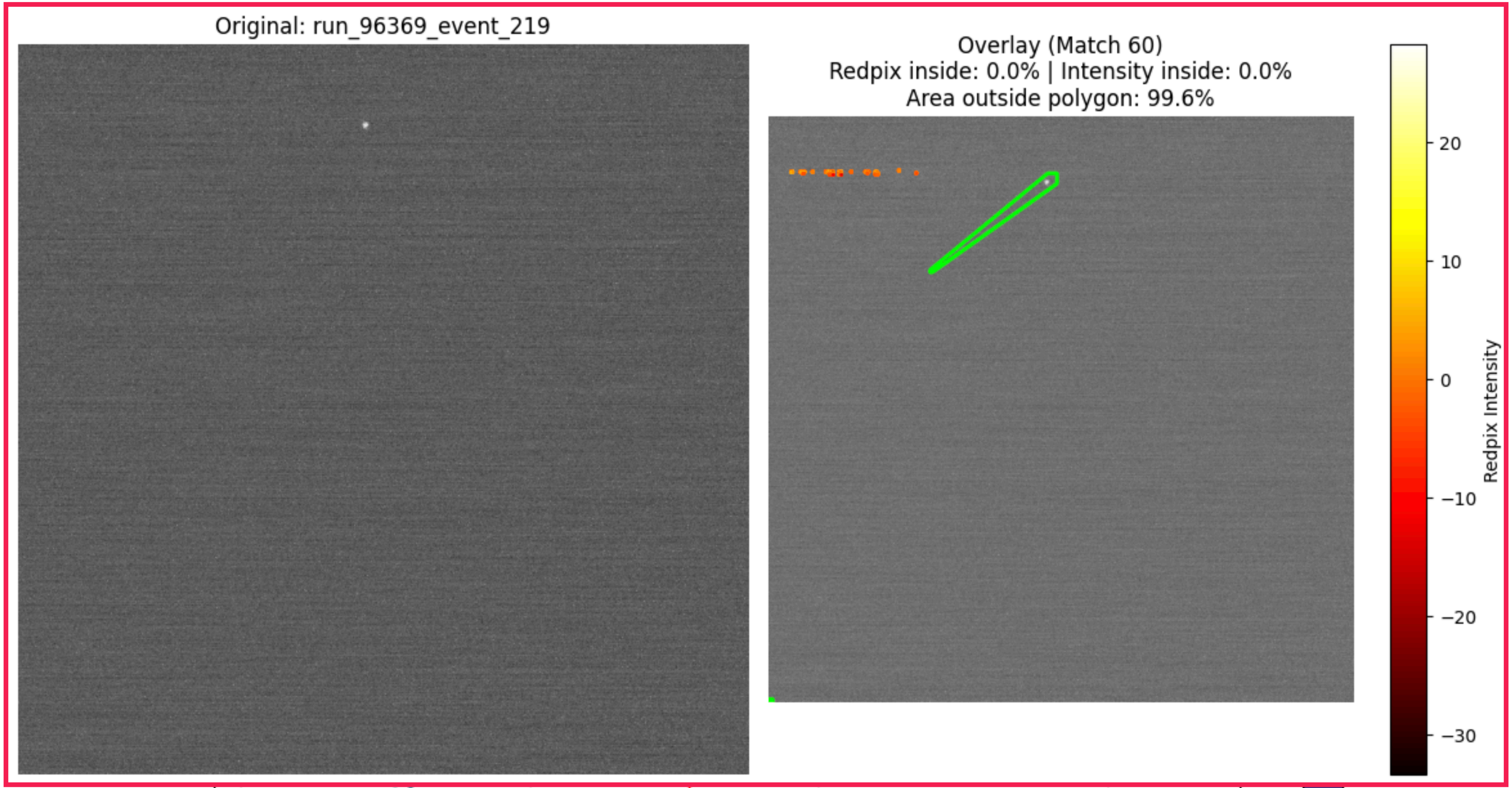
A lot of fake redpix clusters

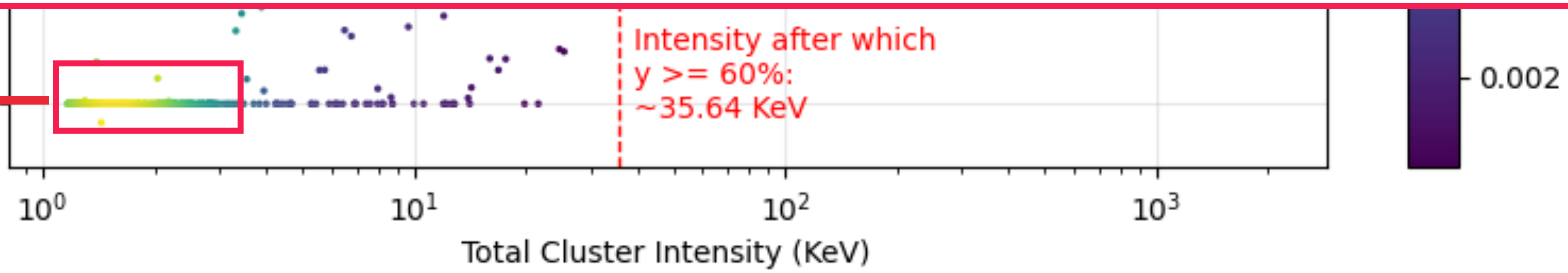


A lot of fake redpix clusters

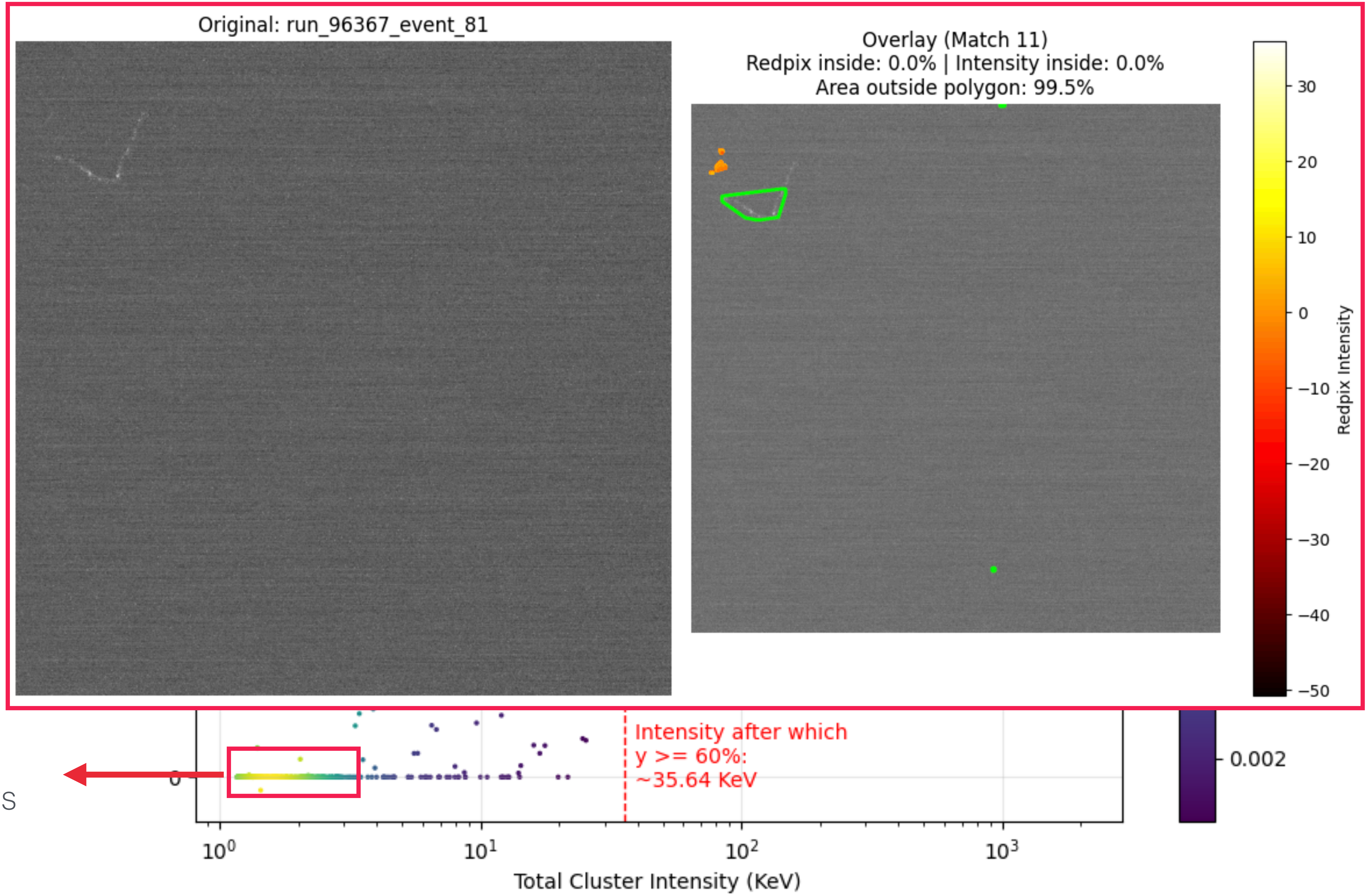


A lot of fake redpix clusters





A lot of fake redpix clusters



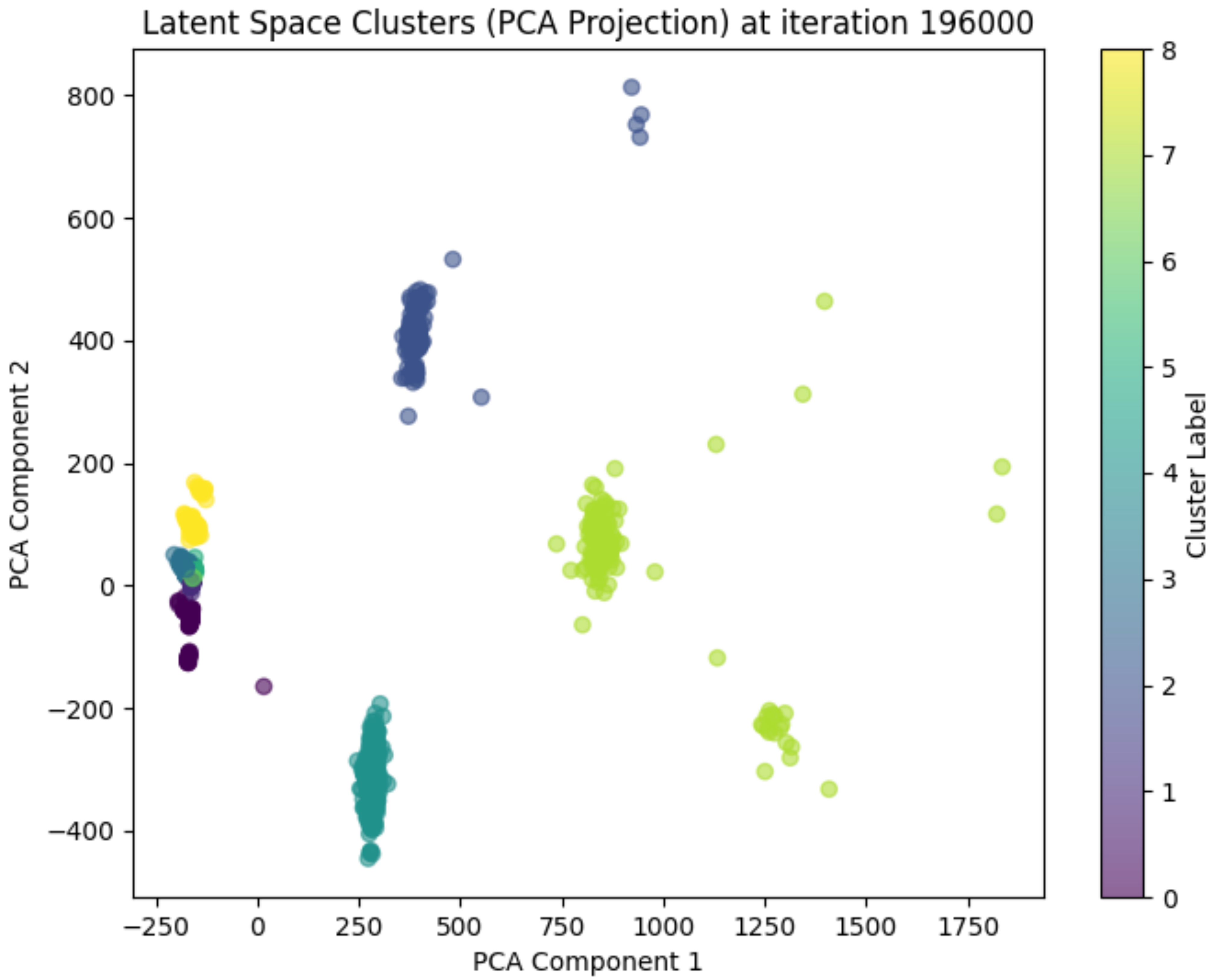
A lot of fake redpix clusters

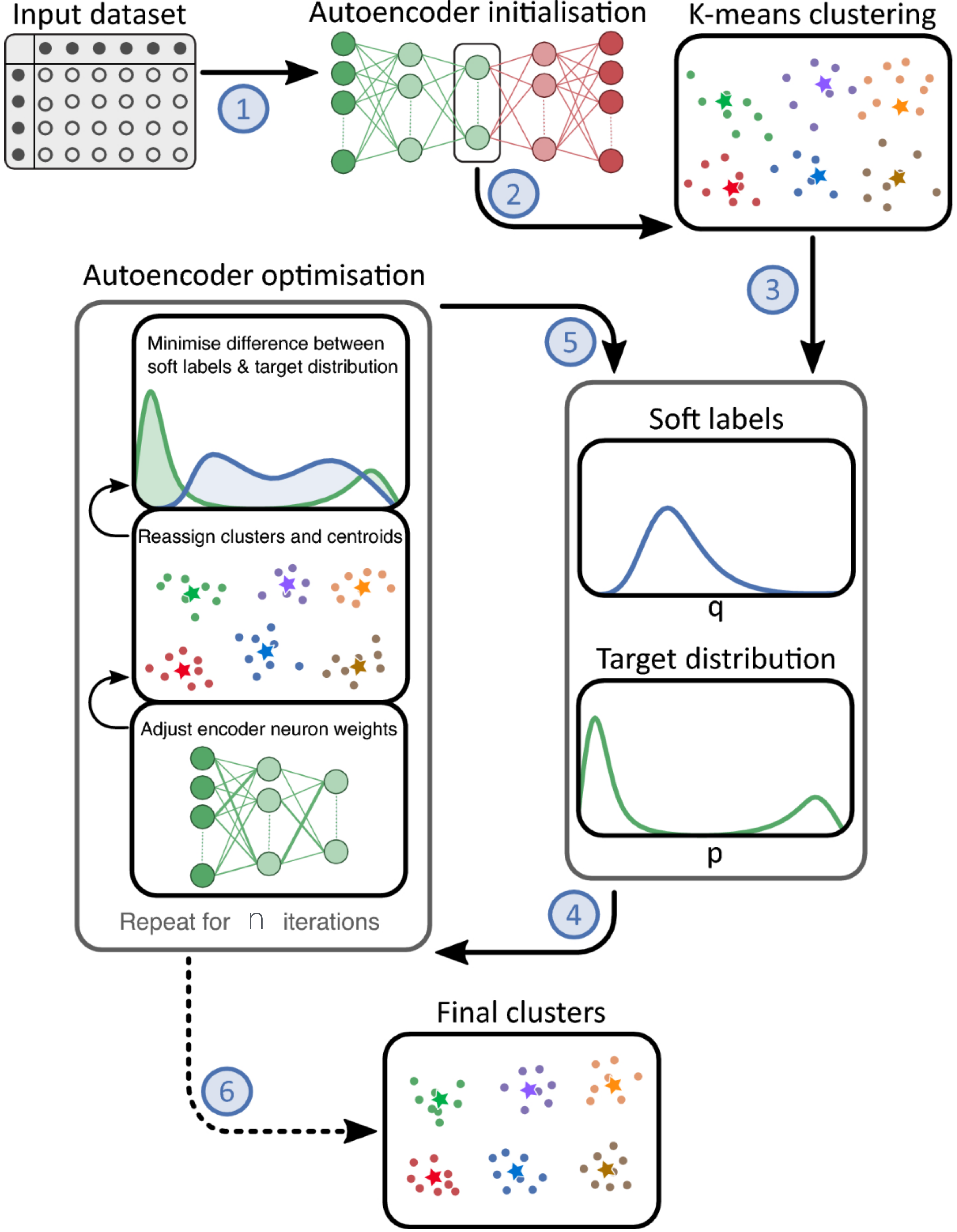
# Unsupervised Classification

## Idea: Deep Embedded Clustering (DEC)

- Assign labels to che clusters based on the latent space representation

- Input: 100x100 patches of the image





Input dataset
1
Autoencoder initialisation
2
K-means clustering
3
Autoencoder optimisation
5
Minimise difference between soft labels & target distribution
Reassign clusters and centroids
Adjust encoder neuron weights
Repeat for n iterations
Soft labels
q
Target distribution
p
4
6
Final clusters

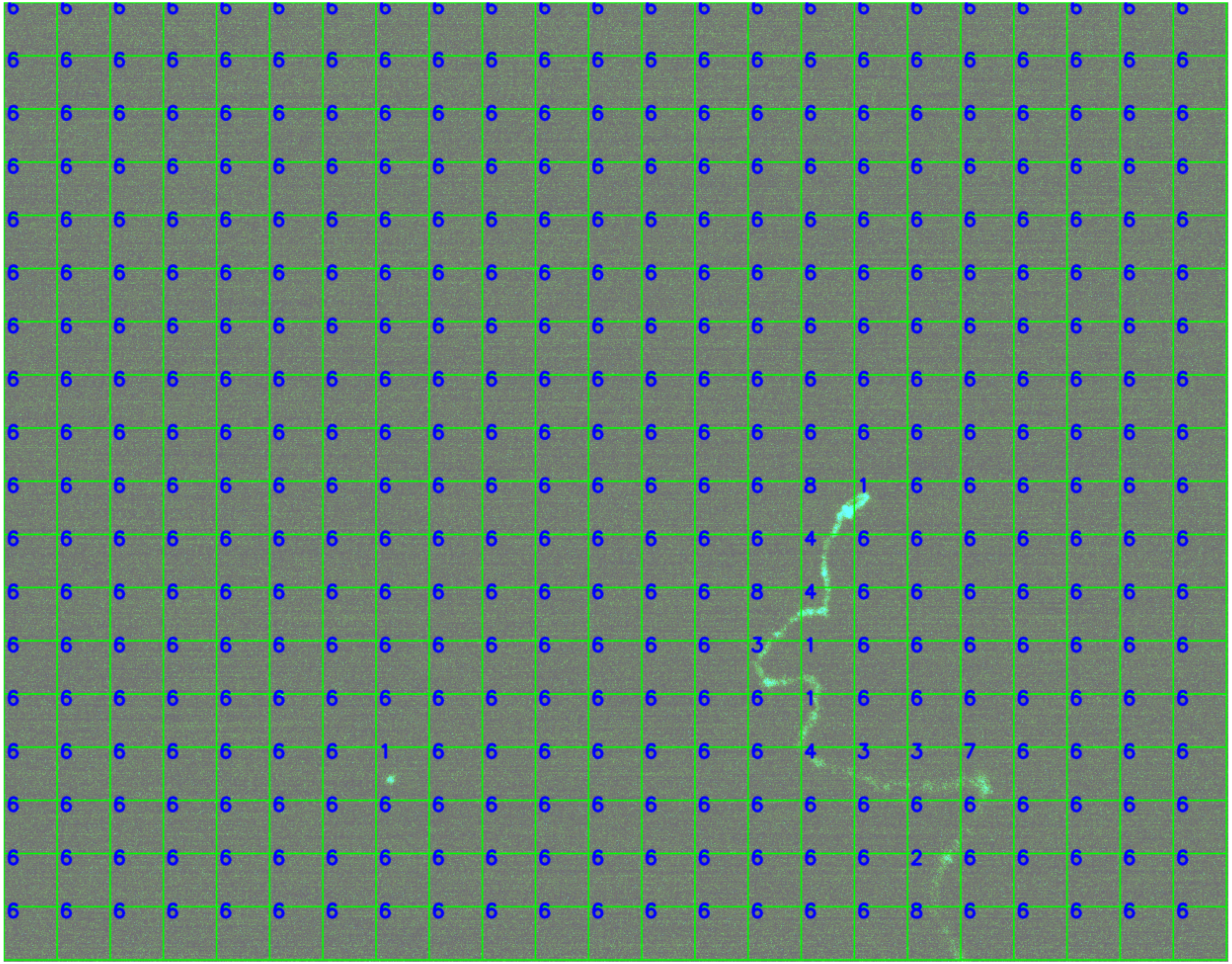
*Sci Rep* **14**, 1045 (2024). https://doi.org/10.1038/s41598-024-51699-z

# Results

## In development

- At the moment the model was trained with a small subset of images
- In order to achieve reasonable results:
- Bigger and more balanced dataset

Original Image with Cluster Labels Overlaid



| | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 |
| 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 |
| 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 |
| 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 |
| 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 |
| 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 |
| 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 |
| 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 |
| 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 |
| 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 8 | 1 | 6 | 6 | 6 | 6 | 6 | 6 |
| 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 4 | 6 | 6 | 6 | 6 | 6 | 6 | 6 |
| 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 8 | 4 | 6 | 6 | 6 | 6 | 6 | 6 | 6 |
| 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 3 | 1 | 6 | 6 | 6 | 6 | 6 | 6 | 6 |
| 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 1 | 6 | 6 | 6 | 6 | 6 | 6 | 6 |
| 6 | 6 | 6 | 6 | 6 | 6 | 6 | 1 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 4 | 3 | 3 | 7 | 6 | 6 | 6 | 6 |
| 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 |
| 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 2 | 6 | 6 | 6 | 6 | 6 |
| 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 8 | 6 | 6 | 6 | 6 | 6 |

# What's next

- The model is able to identify relevant portions of the images, but redpixes are not a good benchmark

- Defining a benchmark ⟶ montecarlo?

- Develop the dataset for the DEC

Thank you for your attention