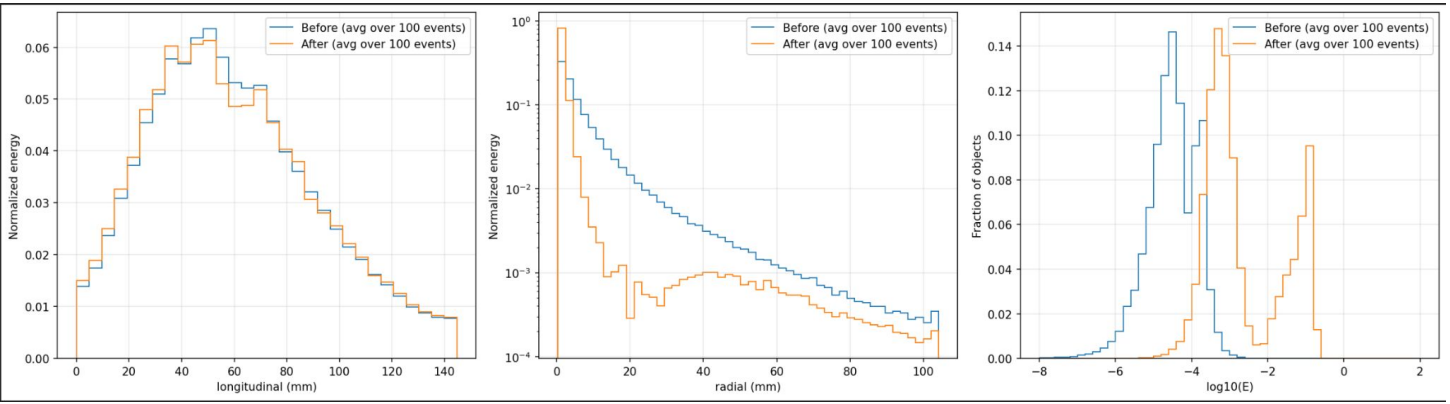
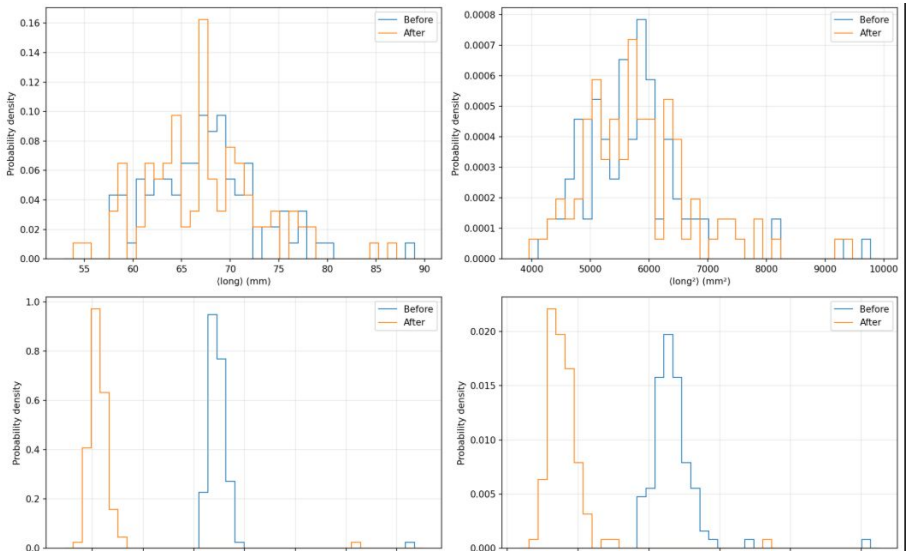
DBSCAN



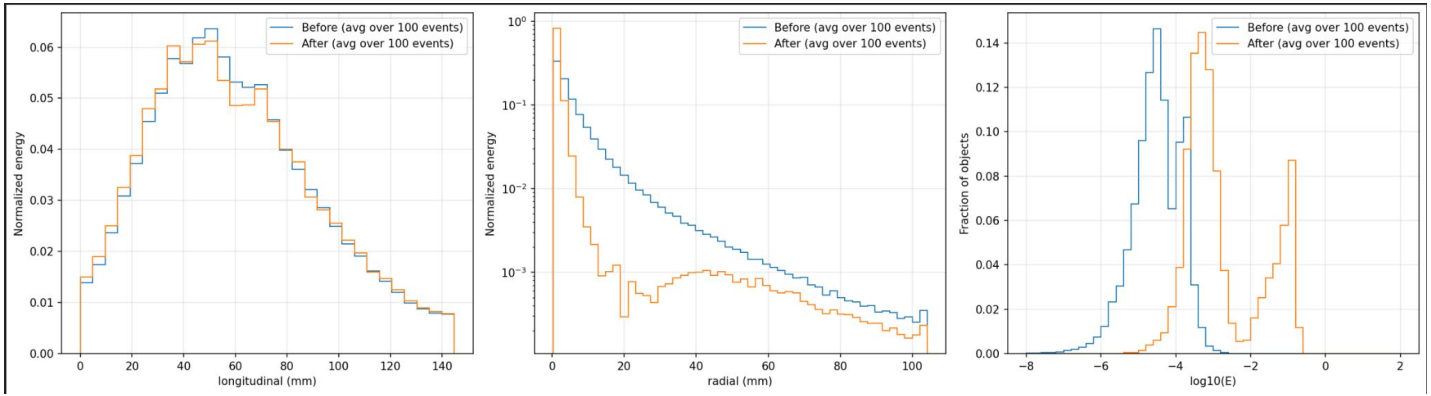
Parameters:

– eps_scaled = 1.5
– min_samples = 5
– n_events = 100
– adaptive = False



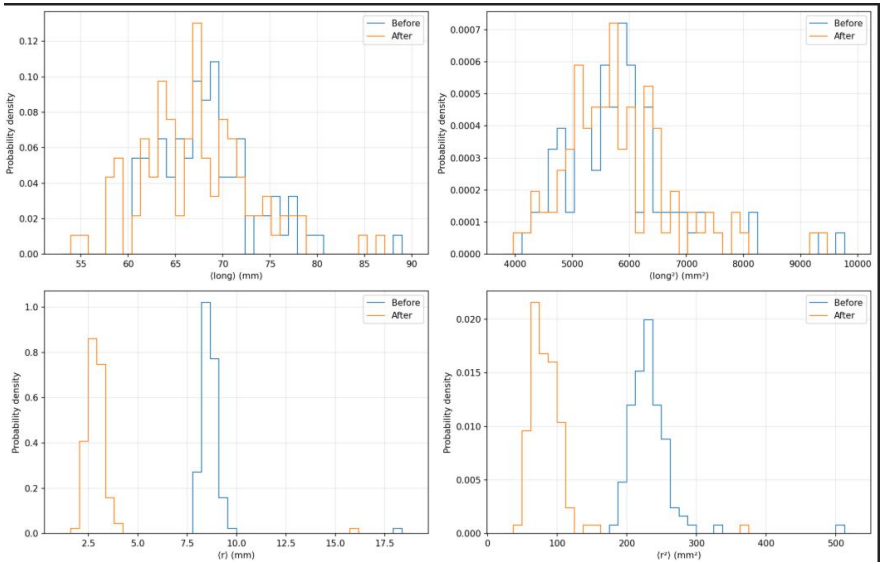

```
--- Object counts (across processed common events) ---
Processed events: 100
Raw steps BEFORE      : 3,588,883
Clusters AFTER        : 13,506
Reduction (AFTER vs BEFORE): -3,575,377 objects (99.62%)
-------------------------------------------------------
```

```
[Containment Δ medians] (After - Before)
  R50: Δmedian = -2.794 mm  (Before=4.152, After=1.358)
  R80: Δmedian = -9.625 mm  (Before=11.816, After=2.191)
  R90: Δmedian = -17.529 mm  (Before=20.465, After=2.935)
  w10: Δmedian = -1.840 mm  (Before=25.190, After=23.350)
  w50: Δmedian = -3.438 mm  (Before=65.300, After=61.862)
  w90: Δmedian = -0.375 mm  (Before=115.951, After=115.576)
```



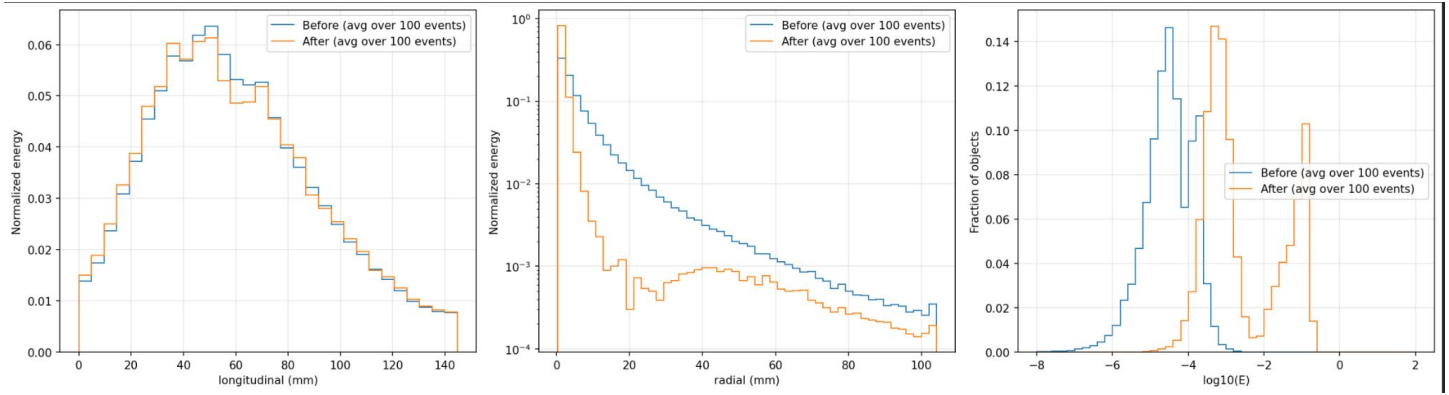
Parameters:

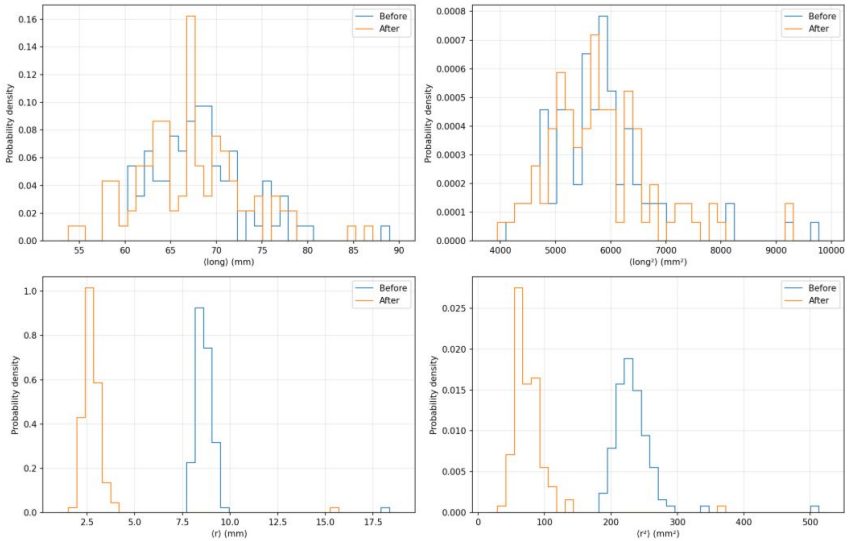
– eps_scaled = 1.5
– min_samples = 3
– n_events = 100
– adaptive = False




```
--- Object counts (across processed common events) ---
Processed events: 100
Raw steps BEFORE      : 3,588,883
Clusters AFTER        : 15,056
Reduction (AFTER vs BEFORE): -3,573,827 objects (99.58%)
-----------------------------------------------------
```

```
[Containment Δ medians] (After - Before)
  R50: Δmedian = -2.796 mm  (Before=4.152, After=1.356)
  R80: Δmedian = -9.638 mm  (Before=11.816, After=2.178)
  R90: Δmedian = -17.531 mm  (Before=20.465, After=2.934)
  w10: Δmedian = -1.605 mm  (Before=25.190, After=23.586)
  w50: Δmedian = -4.497 mm  (Before=65.300, After=60.803)
  w90: Δmedian = -0.260 mm  (Before=115.951, After=115.691)
```



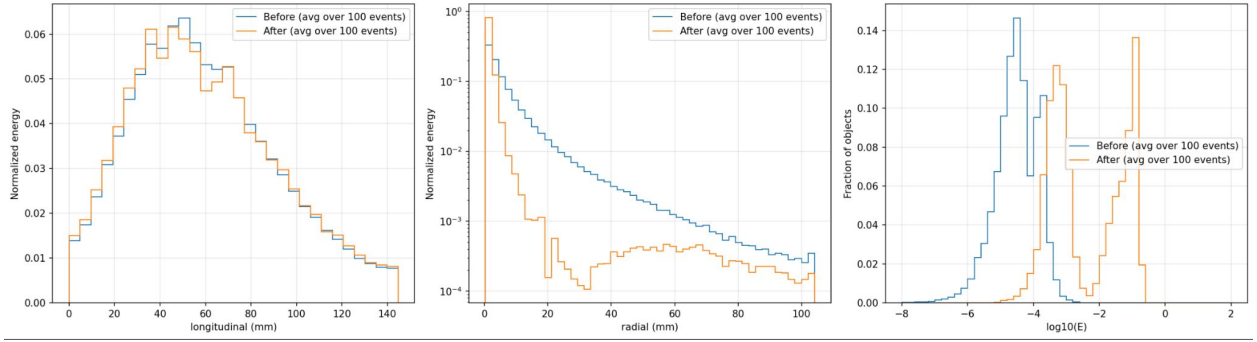


Parameters:

– eps_scaled = 1.5
– min_samples = 7
– n_events = 100
– adaptive = False

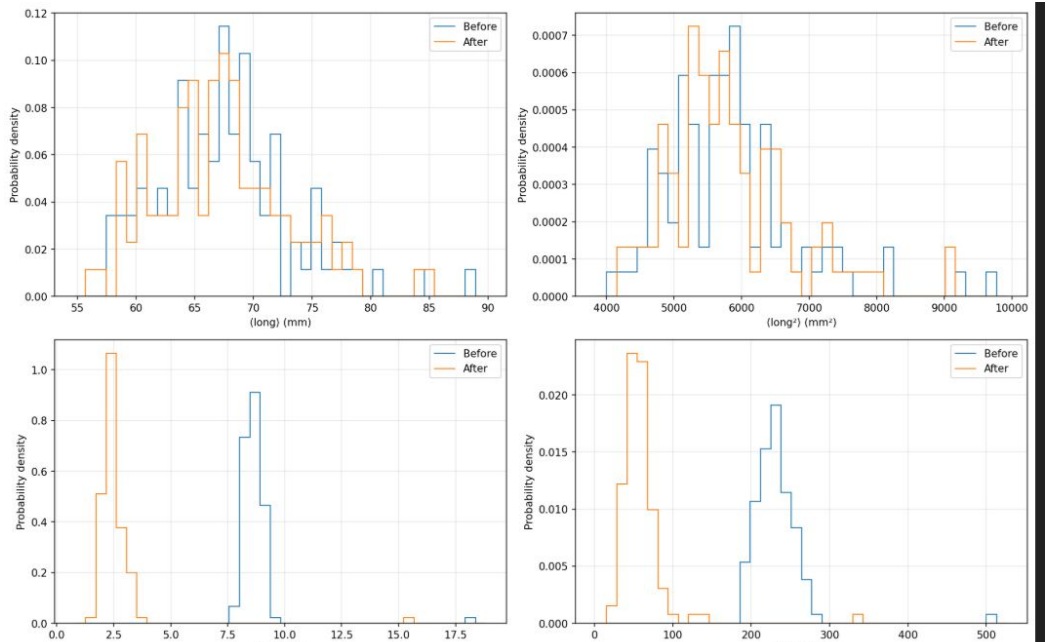
```
[Containment Δ medians] (After − Before)
  R50: Δmedian = -2.798 mm  (Before=4.152, After=1.353)
  R80: Δmedian = -9.627 mm  (Before=11.816, After=2.189)
  R90: Δmedian = -17.561 mm  (Before=20.465, After=2.904)
  w10: Δmedian = -1.888 mm  (Before=25.190, After=23.302)
  w50: Δmedian = -3.449 mm  (Before=65.300, After=61.851)
  w90: Δmedian = -0.636 mm  (Before=115.951, After=115.315)
```

```
--- Object counts (across processed common events) ---
Processed events: 100
Raw steps BEFORE      : 3,588,883
Clusters AFTER        : 12,377
Reduction (AFTER vs BEFORE): -3,576,506 objects (99.66%)
-------------------------------------------------
```



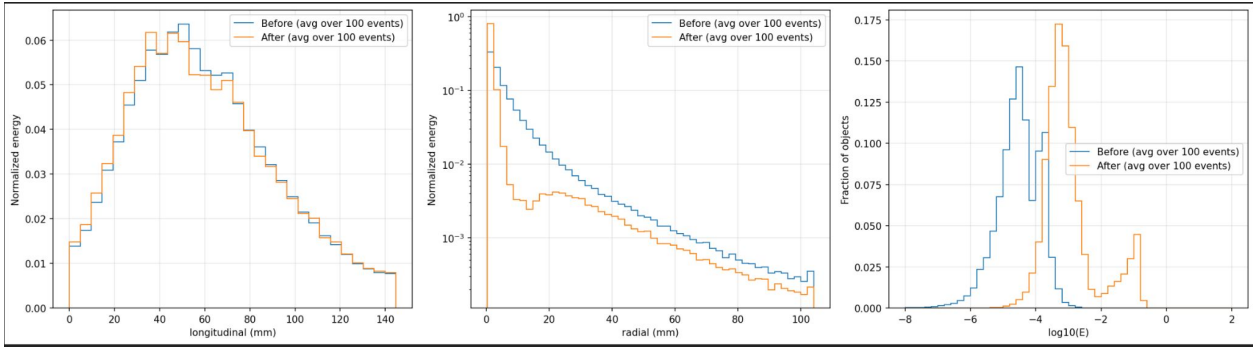
Parameters:

– eps_scaled = 2.5
– min_samples = 5
– n_events = 100
– adaptive = False



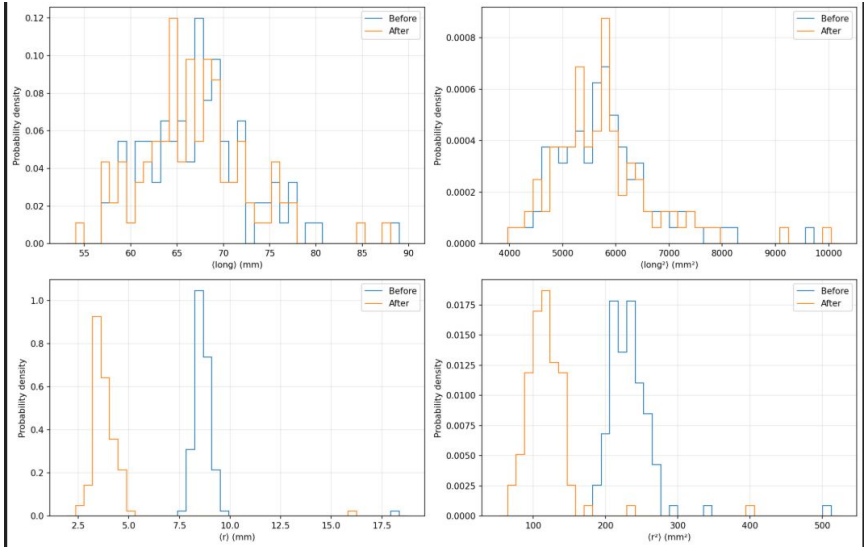

```
[Containment Δ medians] (After − Before)
  R50: Δmedian = -2.784 mm  (Before=4.152, After=1.368)
  R80: Δmedian = -9.578 mm  (Before=11.816, After=2.238)
  R90: Δmedian = -17.506 mm  (Before=20.465, After=2.959)
  w10: Δmedian = -2.222 mm  (Before=25.190, After=22.968)
  w50: Δmedian = -4.616 mm  (Before=65.300, After=60.684)
  w90: Δmedian = -1.331 mm  (Before=115.951, After=114.620)
```

```
--- Object counts (across processed common events) ---
Processed events: 100
Raw steps BEFORE      : 3,588,883
Clusters AFTER        : 10,211
Reduction (AFTER vs BEFORE): -3,578,672 objects (99.72%)
-------------------------------------------------
```



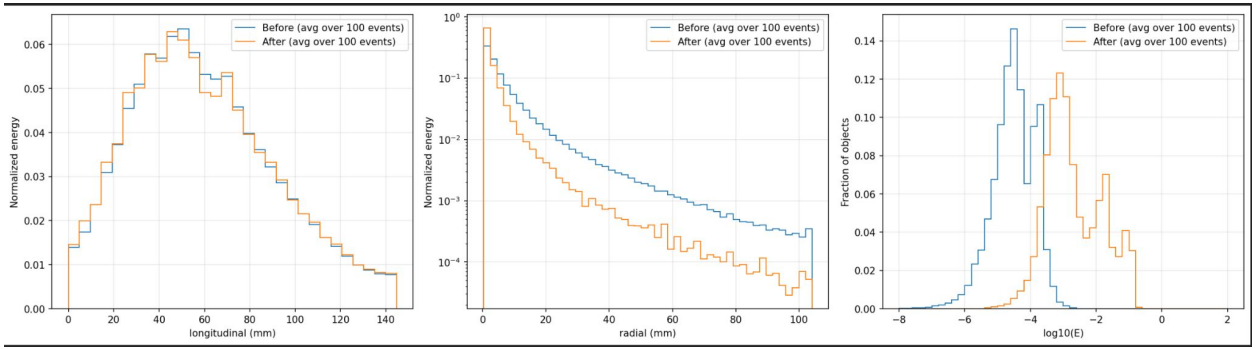
Parameters:

– eps_scaled = 0.5
– min_samples = 5
– n_events = 100
– adaptive = False



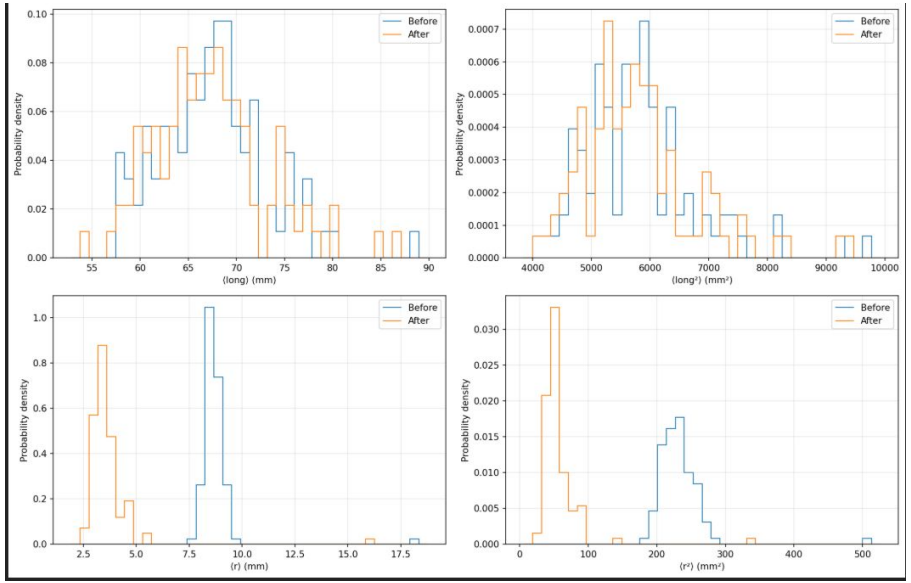

```
--- Object counts (across processed common events) ---
Processed events: 100
Raw steps BEFORE      : 3,588,883
Clusters AFTER        : 25,114
Reduction (AFTER vs BEFORE): -3,563,769 objects (99.30%)
-------------------------------------------------------
```

```
[Containment Δ medians] (After - Before)
  R50: Δmedian = -2.924 mm  (Before=4.152, After=1.227)
  R80: Δmedian = -9.614 mm  (Before=11.816, After=2.202)
  R90: Δmedian = -17.021 mm  (Before=20.465, After=3.444)
  w10: Δmedian = -0.741 mm  (Before=25.190, After=24.450)
  w50: Δmedian = -1.711 mm  (Before=65.300, After=63.589)
  w90: Δmedian = -0.203 mm  (Before=115.951, After=115.749)
```



Parameters:

– eps_scaled = adaptive
– min_samples = 5
– n_events = 100
– adaptive = True




```
--- Object counts (across processed common events) ---
Processed events: 100
Raw steps BEFORE      : 3,588,883
Clusters AFTER        : 22,169
Reduction (AFTER vs BEFORE): -3,566,714 objects (99.38%)
-------------------------------------------------------
```

```
[Containment Δ medians] (After − Before)
  R50: Δmedian = -2.513 mm  (Before=4.152, After=1.638)
  R80: Δmedian = -7.770 mm  (Before=11.816, After=4.047)
  R90: Δmedian = -13.605 mm  (Before=20.465, After=6.859)
  w10: Δmedian = -0.711 mm  (Before=25.190, After=24.480)
  w50: Δmedian = -4.484 mm  (Before=65.300, After=60.817)
  w90: Δmedian = -0.014 mm  (Before=115.951, After=115.937)
```

Report:
I wanted to first evaluate how well a density based clustering approach would cluster these showers so i ran some tests with the parameters (eps= 0.5/1.5/2.5 and min_samples = 3/5/7) as long as a final adaptive epsilon method.

What i found:

Eps: distance between two steps to be considered neighbours

- Increasing eps: less total clusters/more steps gets merged → point cloud size decreases but shower profiles degrade due to over merging of steps
- Decreasing eps: more total clusters → point cloud size increases but better preservation of shower tails

Min_samples: Number of neighbours required to be considered a cluster

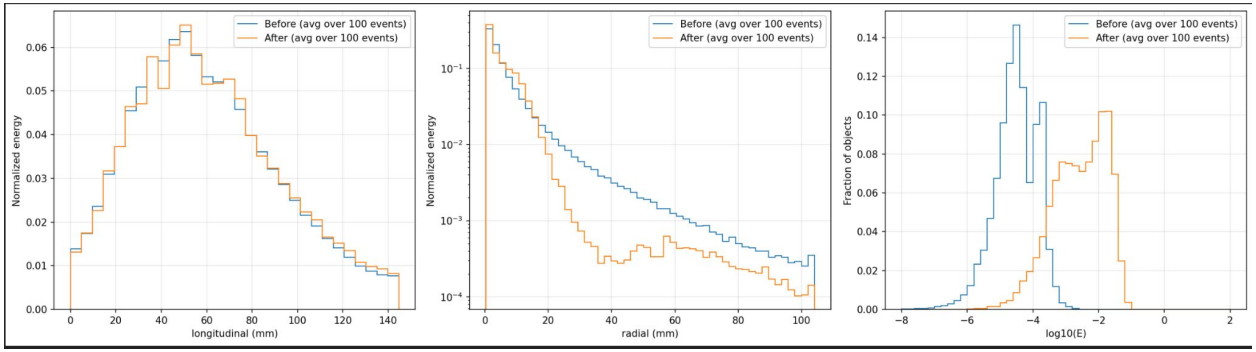
- Increased min_samples: needs more point to form a cluster so tail regions aren't preserved as well and can become noise
- Decreasing min_samples: need less point to form a cluster and hence the tail regions are slightly better preserved with less noise points overall

Overall DBSCAN seems to significantly reduce the point cloud dimensionality at the cost of accurate shower preservation. This approach is not detector agnostic so we still rely on cellID to get layer information and since we are clustering layer by layer, the longitudinal profile seems to be roughly preserved with the largest shift in the core of the longitudinal profile of 4.84mm.This is expected since we dont cluster between layers so the overall longitudinal development is still preserved.The radial profile on the other hand significantly deteriorates where radial containment worsens after clustering even in the core.This is also expected as dbscan uses a single radius in each layer causing the steps in the dense cores to connect earlier and drag neighbours inwards. This explains why when decreasing eps we saw better radial preservation as less points get dragged into the core leading to the overall radial profile being better preserved

I also implemented an adaptive epsilon method.A simple two zone adaptive scheme (larger eps in the core and smaller eps in the tail) hoping to improve the compression-physics fidelity trade off.When trying an adaptive epsilon approach we saw that the core could be appropriately clustered whilst limiting over merging of steps in the tail regions. This lead to slightly better longitudinal profiles preservation compared to when using a global epsilon value and even better radial profile preservation but still with significant room to improve.
In terms of performance, DBSCAN was able to cluster one event per second on average.

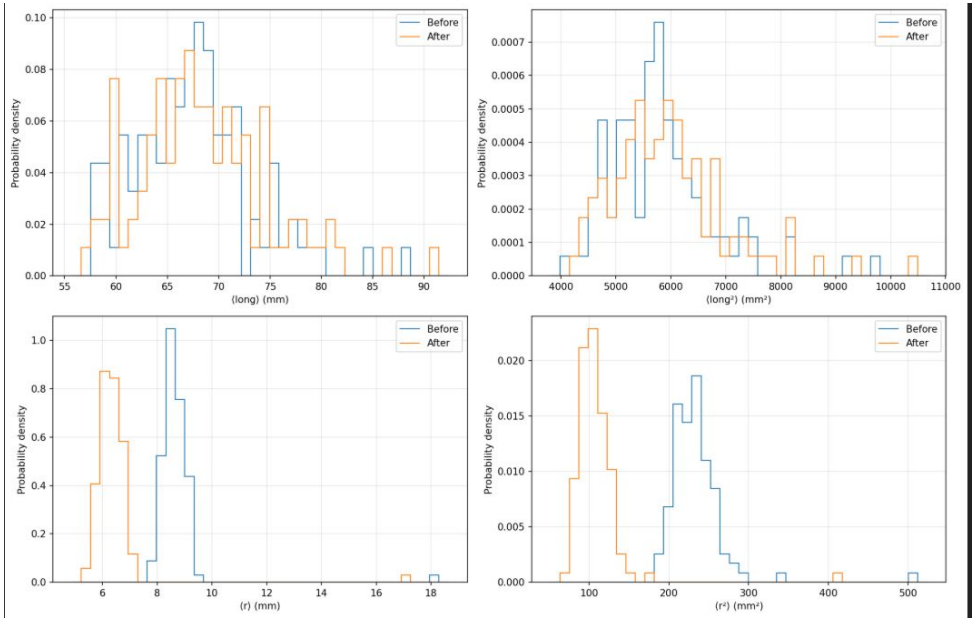
CLUE

K4Clue study:https://cds.cern.ch/record/2882302/files/Publication.pdf



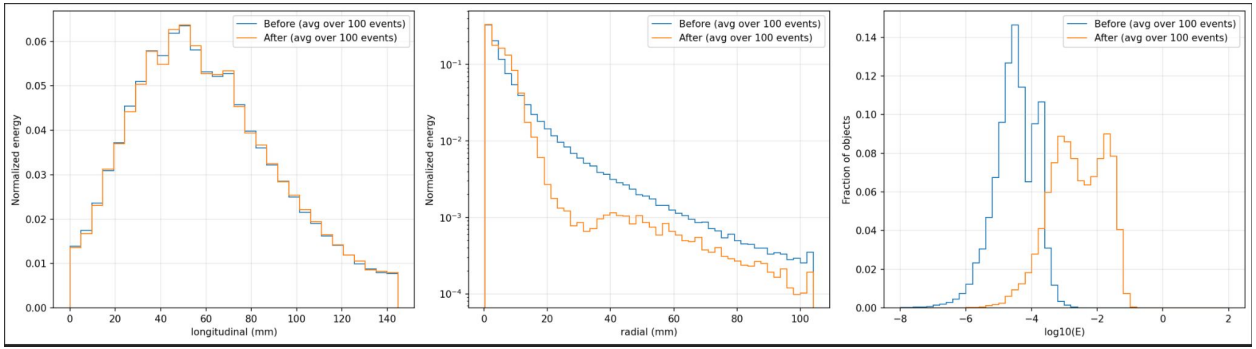
Parameters:

– dc = 15.3
– deltac = None
– rhoc = 5
– outlier_factor = 3



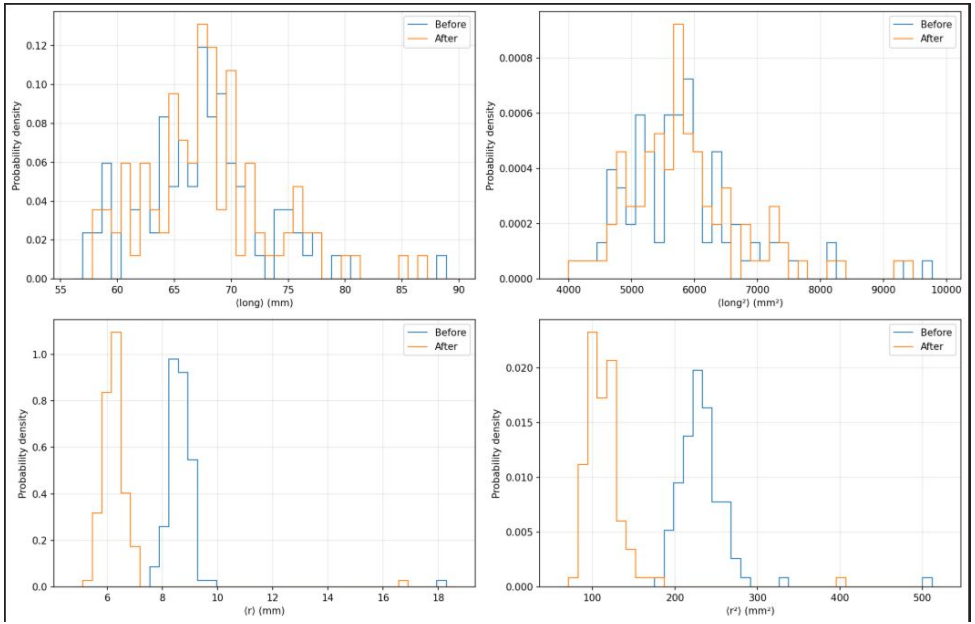

```
--- Object counts (across processed common events) ---
Processed events: 100
Raw steps BEFORE      : 3,588,883
Clusters AFTER        : 28,449
Reduction (AFTER vs BEFORE): -3,560,434 objects (99.21%)
-------------------------------------------------------
```

```
[Containment Δ medians] (After − Before)
  R50: Δmedian = +0.363 mm  (Before=4.152, After=4.514)
  R80: Δmedian = -3.134 mm  (Before=11.816, After=8.682)
  R90: Δmedian = -9.396 mm  (Before=20.465, After=11.068)
  w10: Δmedian = -0.106 mm  (Before=25.190, After=25.084)
  w50: Δmedian = -0.435 mm  (Before=65.300, After=64.865)
  w90: Δmedian = +0.063 mm  (Before=115.951, After=116.015)
```



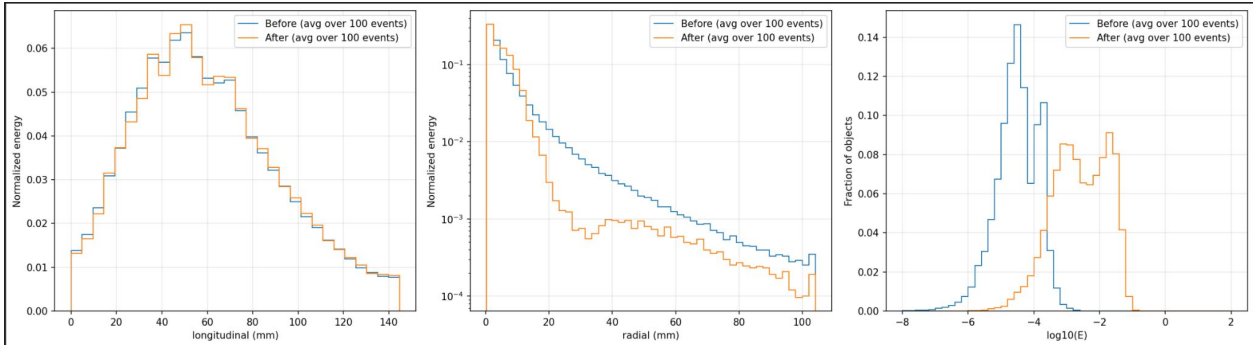
Parameters:

– dc = 10
– deltac = None
– rhoc = 5
– outlier_factor = None



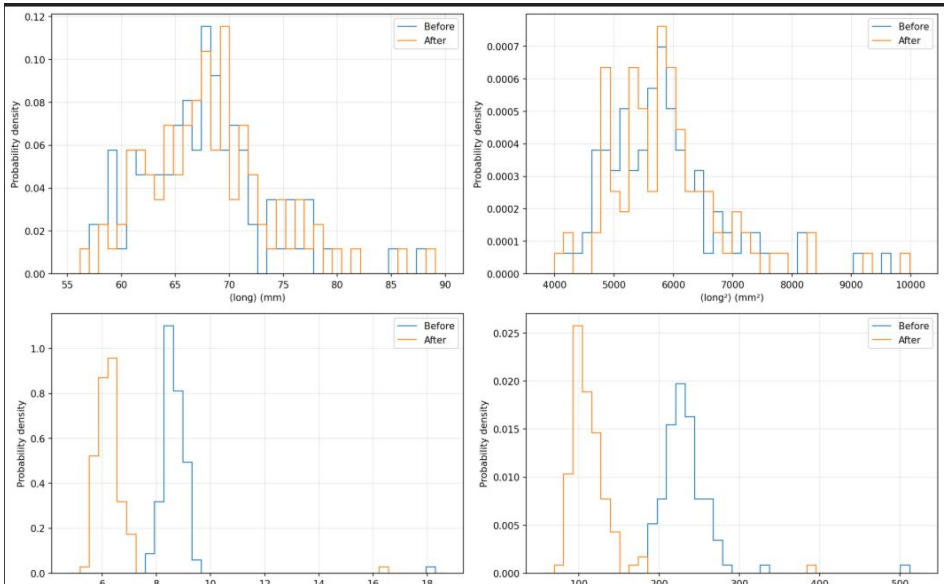

```
--- Object counts (across processed common events) ---
Processed events: 100
Raw steps BEFORE      : 3,588,883
Clusters AFTER        : 30,994
Reduction (AFTER vs BEFORE): -3,557,889 objects (99.14%)
-------------------------------------------------------
```

```
[Containment Δ medians] (After - Before)
  R50: Δmedian = +0.369 mm  (Before=4.152, After=4.521)
  R80: Δmedian = -2.994 mm  (Before=11.816, After=8.822)
  R90: Δmedian = -9.313 mm  (Before=20.465, After=11.152)
  w10: Δmedian = -0.057 mm  (Before=25.190, After=25.133)
  w50: Δmedian = -1.062 mm  (Before=65.300, After=64.238)
  w90: Δmedian = -0.024 mm  (Before=115.951, After=115.928)
```



Parameters:

– dc = 10
– deltac = 13
– rhoc = 5
– outlier_factor = None




```
--- Object counts (across processed common events) ---
Processed events: 100
Raw steps BEFORE      : 3,588,883
Clusters AFTER        : 29,737
Reduction (AFTER vs BEFORE): -3,559,146 objects (99.17%)
-------------------------------------------------------
```

```
[Containment Δ medians] (After − Before)
  R50: Δmedian = -0.127 mm  (Before=4.152, After=4.024)
  R80: Δmedian = -1.867 mm  (Before=11.816, After=9.949)
  R90: Δmedian = -7.307 mm  (Before=20.465, After=13.157)
  w10: Δmedian = -0.006 mm  (Before=25.190, After=25.185)
  w50: Δmedian = -0.795 mm  (Before=65.300, After=64.505)
  w90: Δmedian = +0.421 mm  (Before=115.951, After=116.373)
```

Report:
Next I wanted to test how well an existing approach would work. CLUE is designed for reconstructed hits with energy weighting, whereas here we operate on raw steps where energies are very small. I therefore reproduced the parameters in the K4Clue paper and then varied them to study sensitivity—adapting CLUE to a hit-count density version (removing energy weights) while keeping the rest of the logic intact: dc (density radius), rhoc (density threshold), optional deltac (peak-separation threshold), and outlier_factor (tail rejection).

What i found:
Longitudinal profiles were the most stable across configurations; the main effect is a small core shift ≈ 0.795 mm, consistent with our layer-by-layer clustering choice (similar to DBSCAN's longitudinal behavior).

Radial profiles still degraded—though less than DBSCAN. With the paper-aligned parameters (larger dc, rhoc=5, outlier_factor=3) the radial tails worsened by >7 mm, indicating over-merging in dense cores. Reducing dc localizes the density estimate, which visibly reduces core over-merging and preserves tails better, at the cost of more, smaller clusters (fragmentation); overall radial fidelity remains imperfect.

Adding a deltac separation further marginally improves radial and longitudinal cores by enforcing a minimum distance between clusters, mitigating merging between close clusters that were being merged before.

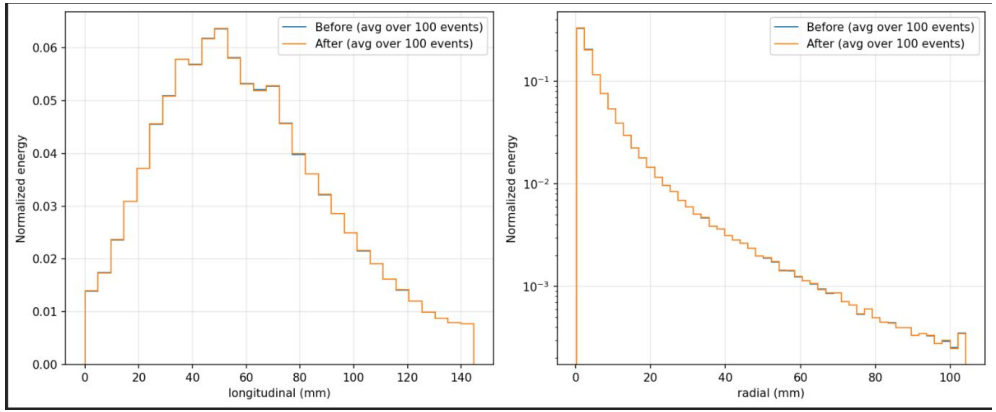
With outlier_factor = 3 I observed a loss of total energy as low-density tail steps are flagged as outliers and dropped. In a reconstructed-hit workflow (with energies) this would preferentially remove truly low-energy hits; in our step-count variant it suppresses legitimate tail structure.

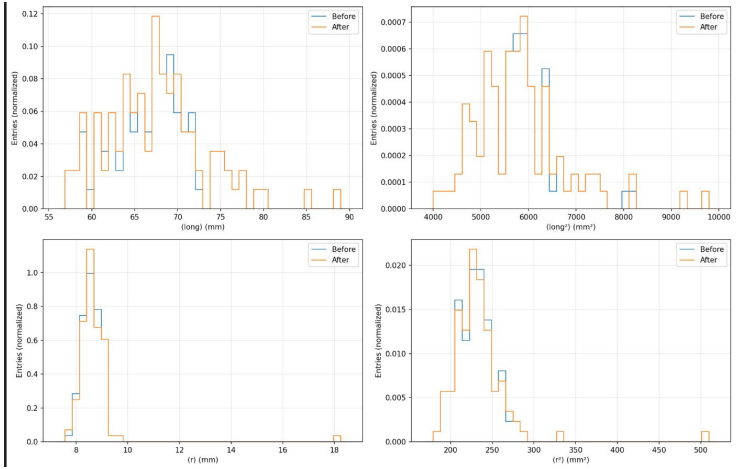
Overall. Compression was comparable to DBSCAN, but CLUE preserved physics observables slightly better—notably a smaller longitudinal-core shift and somewhat better radial containment—while still achieving large point-count reduction. That said, radial preservation still leaves room for improvement under step-count density.

Performance. While not the focus, CLUE ran slower in my tests (≈ 4 s/event vs ≈ 1 s/event for DBSCAN on the same setup). This gap should be addressable with parallelization (CUDA) and further vectorization.

Grid Clustering

CaloClouds study: https://arxiv.org/pdf/2305.04847v2



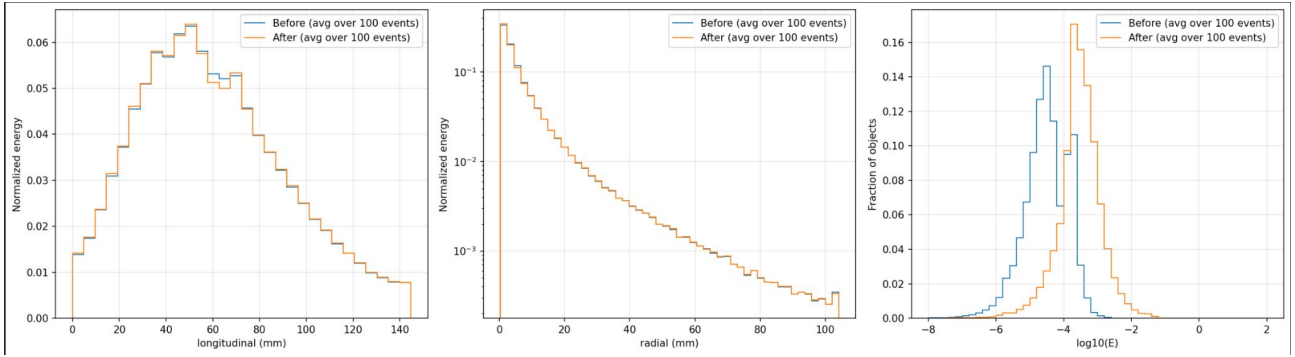


Parameters:

– grid_trans_mm = 0.85
– second_stage = None

```
--- Object counts (across processed common events) ---
Processed events: 100
Raw steps BEFORE      : 3,588,883
Clusters AFTER        : 685,536
Reduction (AFTER vs BEFORE): -2,903,347 objects (80.90%)
-----------------------------------------------------
```

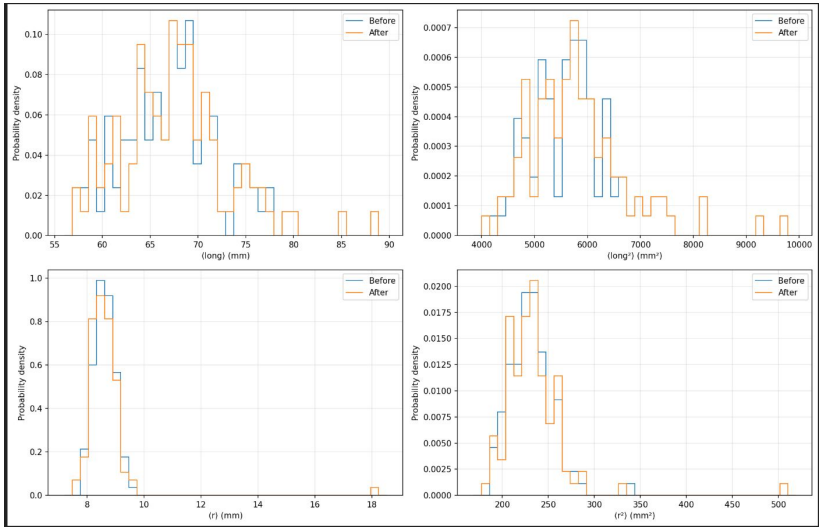
```
[Containment Δ medians] (After - Before)
  R50: Δmedian = -0.018 mm  (Before=4.152, After=4.134)
  R80: Δmedian = +0.025 mm  (Before=11.816, After=11.842)
  R90: Δmedian = +0.067 mm  (Before=20.465, After=20.531)
  w10: Δmedian = +0.000 mm  (Before=25.190, After=25.191)
  w50: Δmedian = +0.034 mm  (Before=65.300, After=65.334)
  w90: Δmedian = +0.004 mm  (Before=115.951, After=115.955)
```



Parameters:

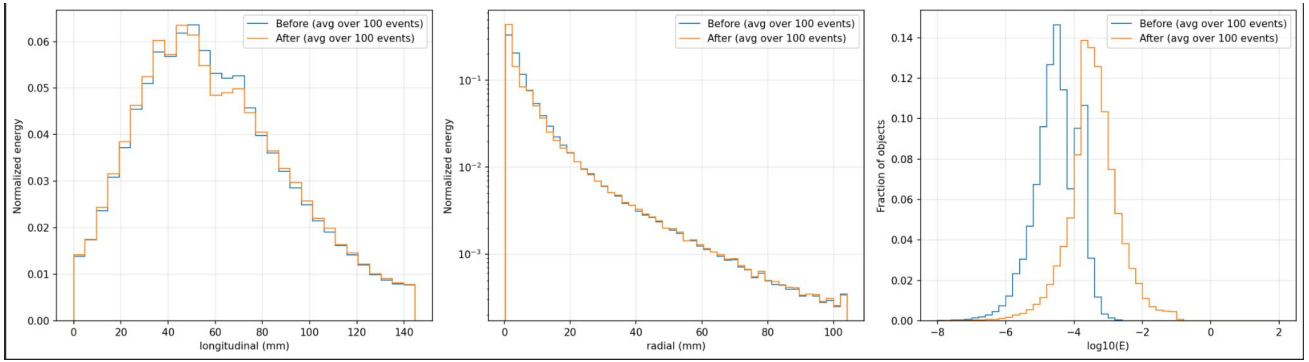
– grid_trans_mm = 3.0

– second_stage = None




```
--- Object counts (across processed common events) ---
Processed events: 100
Raw steps BEFORE      : 3,588,883
Clusters AFTER        : 307,916
Reduction (AFTER vs BEFORE): -3,280,967 objects (91.42%)
-------------------------------------------------------
```

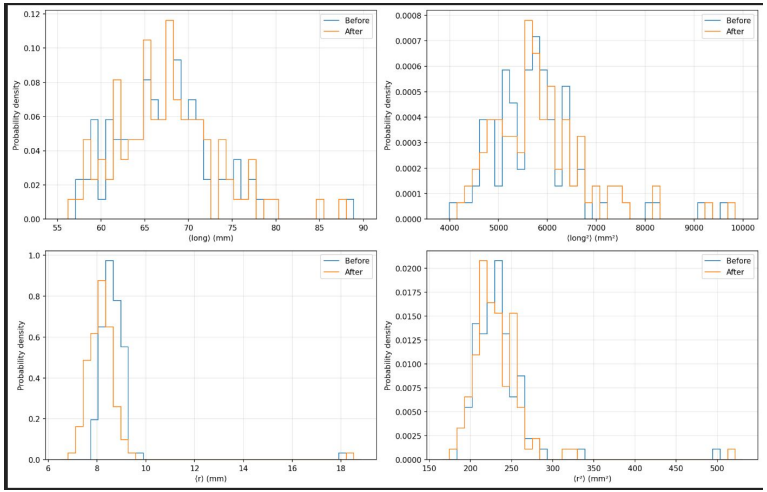
```
[Containment Δ medians] (After - Before)
  R50: Δmedian = -0.134 mm  (Before=4.152, After=4.017)
  R80: Δmedian = +0.031 mm  (Before=11.816, After=11.847)
  R90: Δmedian = -0.005 mm  (Before=20.465, After=20.460)
  w10: Δmedian = -0.022 mm  (Before=25.190, After=25.169)
  w50: Δmedian = -0.113 mm  (Before=65.300, After=65.187)
  w90: Δmedian = +0.025 mm  (Before=115.951, After=115.976)
```



Parameters:

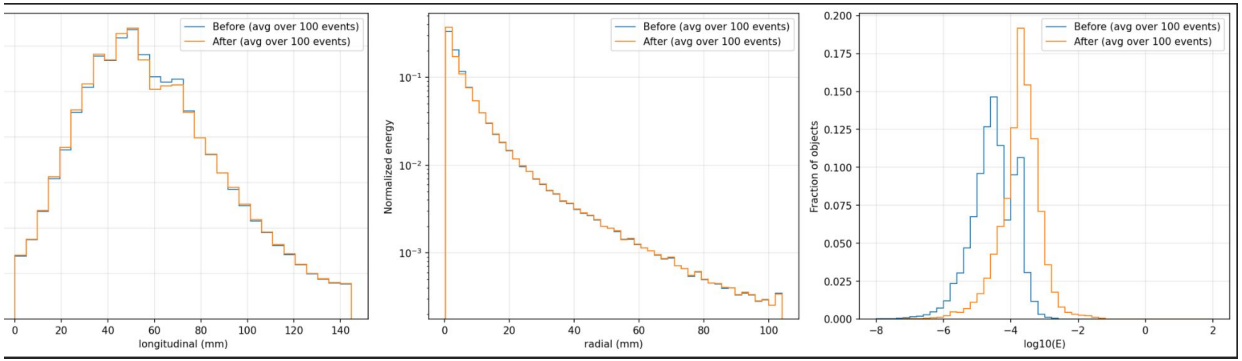
– grid_trans_mm = 10.0

– second_stage = None




```
--- Object counts (across processed common events) ---
Processed events: 100
Raw steps BEFORE      : 3,588,883
Clusters AFTER        : 144,028
Reduction (AFTER vs BEFORE): -3,444,855 objects (95.99%)
-------------------------------------------------------
```

```
[Containment Δ medians] (After − Before)
  R50: Δmedian = -1.151 mm  (Before=4.152, After=3.001)
  R80: Δmedian = -0.482 mm  (Before=11.816, After=11.334)
  R90: Δmedian = +0.091 mm  (Before=20.465, After=20.556)
  w10: Δmedian = -0.103 mm  (Before=25.190, After=25.088)
  w50: Δmedian = -4.093 mm  (Before=65.300, After=61.207)
  w90: Δmedian = +0.156 mm  (Before=115.951, After=116.108)
```



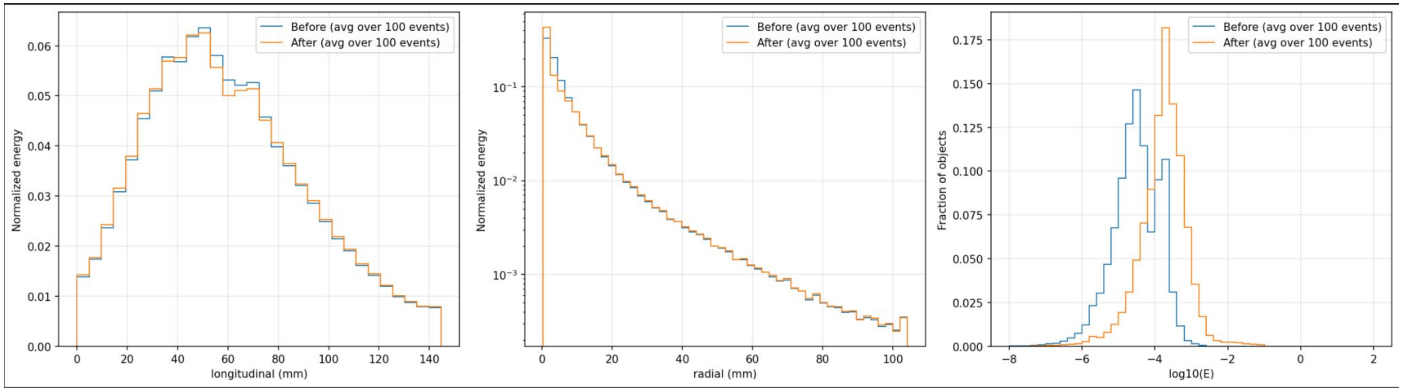
Parameters:

– grid_trans_mm = 0.85

– second_stage = CCL4

```
--- Object counts (across processed common events) ---
Processed events: 100
Raw steps BEFORE      : 3,588,883
Clusters AFTER        : 490,406
Reduction (AFTER vs BEFORE): -3,098,477 objects (86.34%)
-----------------------------------------------------
```

```
[Containment Δ medians] (After - Before)
  R50: Δmedian = -0.160 mm  (Before=4.152, After=3.992)
  R80: Δmedian = +0.122 mm  (Before=11.816, After=11.938)
  R90: Δmedian = +0.159 mm  (Before=20.465, After=20.624)
  w10: Δmedian = +0.008 mm  (Before=25.190, After=25.198)
  w50: Δmedian = -2.152 mm  (Before=65.300, After=63.148)
  w90: Δmedian = +0.058 mm  (Before=115.951, After=116.010)
```



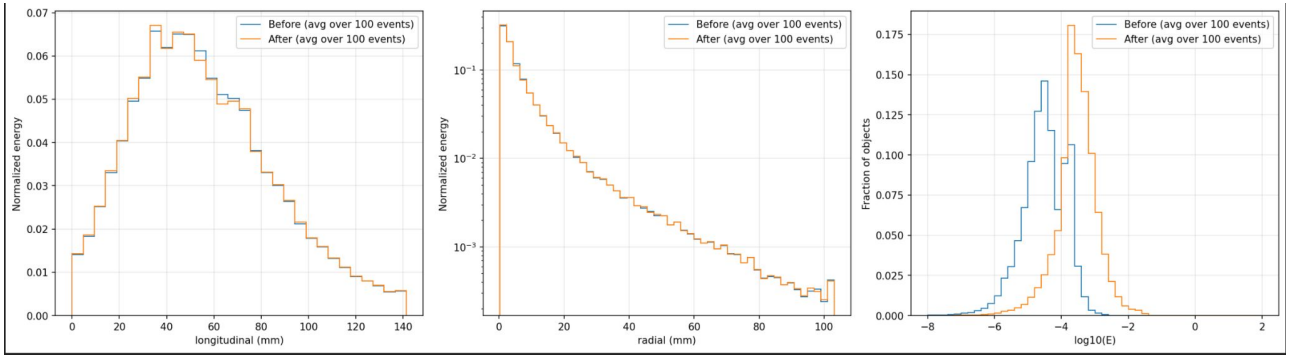
Parameters:

– grid_trans_mm = 0.85

– second_stage = CCL8

```
--- Object counts (across processed common events) ---
Processed events: 100
Raw steps BEFORE      : 3,588,883
Clusters AFTER        : 431,135
Reduction (AFTER vs BEFORE): -3,157,748 objects (87.99%)
-----------------------------------------------------
```

```
[Containment Δ medians] (After - Before)
  R50: Δmedian = -0.778 mm  (Before=4.152, After=3.374)
  R80: Δmedian = +0.157 mm  (Before=11.816, After=11.973)
  R90: Δmedian = +0.246 mm  (Before=20.465, After=20.711)
  w10: Δmedian = -0.018 mm  (Before=25.190, After=25.173)
  w50: Δmedian = -0.114 mm  (Before=65.300, After=65.186)
  w90: Δmedian = +0.012 mm  (Before=115.951, After=115.963)
```

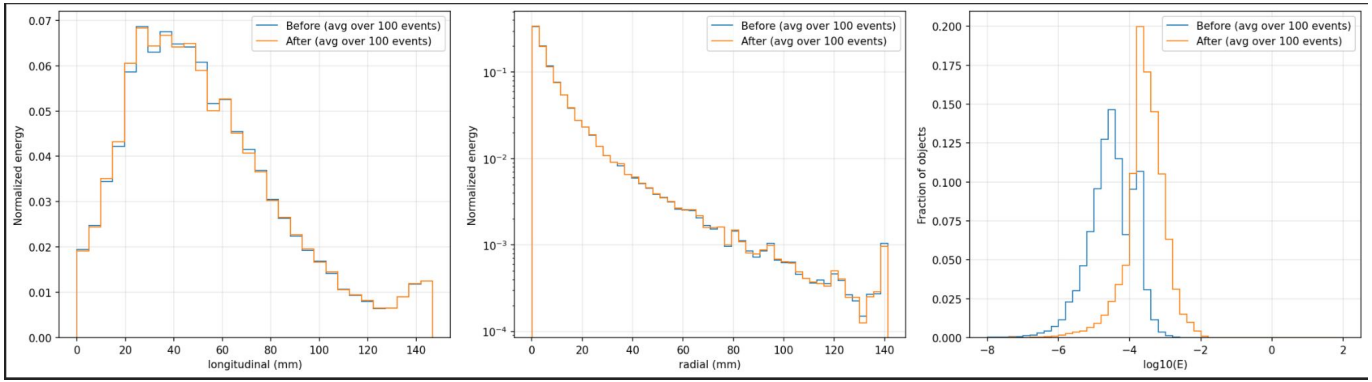


Parameters:

– grid_trans_mm = 3.0
– second_stage = None
– Energy: 50 GeV

```
--- Object counts (across processed common events) ---
Processed events: 100
Raw steps BEFORE      : 1,814,551
Clusters AFTER        : 184,544
Reduction (AFTER vs BEFORE): -1,630,007 objects (89.83%)
-----------------------------------------------------
```

```
[Containment Δ medians] (After - Before)
  R50: Δmedian = -0.144 mm  (Before=4.216, After=4.072)
  R80: Δmedian = +0.052 mm  (Before=11.932, After=11.983)
  R90: Δmedian = +0.048 mm  (Before=20.692, After=20.741)
  w10: Δmedian = -0.185 mm  (Before=24.945, After=24.760)
  w50: Δmedian = -0.056 mm  (Before=60.305, After=60.249)
  w90: Δmedian = +0.078 mm  (Before=111.081, After=111.158)
```



Parameters:

– grid_trans_mm = 3.0
– second_stage = None
– Energy: 10 Gev

```
--- Object counts (across processed common events) ---
Processed events: 100
Raw steps BEFORE      : 365,437
Clusters AFTER        : 49,417
Reduction (AFTER vs BEFORE): -316,020 objects (86.48%)
-------------------------------------------------------
```

```
[Containment Δ medians] (After − Before)
  R50: Δmedian = -0.084 mm  (Before=4.295, After=4.211)
  R80: Δmedian = +0.042 mm  (Before=11.749, After=11.791)
  R90: Δmedian = -0.160 mm  (Before=20.448, After=20.288)
  w10: Δmedian = -0.062 mm  (Before=19.966, After=19.904)
  w50: Δmedian = -0.049 mm  (Before=50.581, After=50.532)
  w90: Δmedian = +0.044 mm  (Before=101.621, After=101.665)
```

Report:

Finally I focused on another existing approach—the pre-processing stage from the CaloCloud generative pipeline—to test how well layer-wise transverse gridding can cluster steps while preserving physics observables.

What i found:

Grid_trans_mm:

- smaller → smaller transverse cells per layer; finer quantization, better fidelity of shower shapes, but less step-count compression.
- larger → larger cells; coarser quantization, more merging and higher compression, with increasing risk of profile distortion.

Second stage (CCL4/CCL8):

- CCL4 → conservative connected-component merge (4-neighborhood); tends to remove obvious single-cell fragments while limiting over-merge.
- CCL8 → more aggressive merge (8-neighborhood); yields higher compression at a greater risk of over-merging adjacent cells, especially along diagonals.

With a low grid_trans_mm (0.85 mm), longitudinal and radial profiles were almost perfectly preserved: projecting a fixed grid onto each layer prevents the inward drag that we saw in radius-based clustering—steps are only merged with local neighbors that fall within the same cell. Compression is lower than DBSCAN or CLUE, but both longitudinal and radial observables are preserved significantly better. Increasing grid_trans_mm behaves as expected: more compression with minor, then noticeable divergence in radial and (to a lesser extent) longitudinal profiles as cells enlarge. At 10 mm, compression improves further but degradation becomes clear, including more visible longitudinal core shifts. Importantly, these deviations are measured at the step level and are sub-cell relative to CLD's ~5 mm readout cell size; at the readout level (what the detector sees), much of this may be below granularity and thus less consequential.

I also added a second stage on top of 0.85 mm gridding to reduce step count further. With CCL4, adjacent occupied cells are merged without aggressive diagonal growth: I observed extra compression compared to bare 0.85 mm gridding with minimal additional loss in the before/after physics profiles. With CCL8, compression increases again, but core worsening in the longitudinal profile and tail loss in the radial profile are more apparent (though not catastrophic). Given the small compression delta between CCL4 and CCL8, CCL4 is the safer default when physics fidelity is the priority. Finally, I verified this scheme across different energies (50 GeV vs 10 GeV) and observed the same favorable pattern: good compression with excellent preservation of the key observables, especially at finer grids.

In terms of performance, while not the primary metric, this approach was the fastest in my tests ~3–4 events/s—versus ~1 event/s for DBSCAN and ~0.25 events/s for CLUE under the same conditions. This further supports gridding as a strong baseline to develop: it's simple, naturally parallelizable, and offers an attractive compression–fidelity–throughput balance.

```
ECAL clustering (grid, axis): 100%|                                                                                | 100/100 [00:28<00:00,  3.57it/s]
```