Hashing and Metric Learning

for particle tracking

Rutherford Appleton Laboratory Seminar, 21st Oct 2020

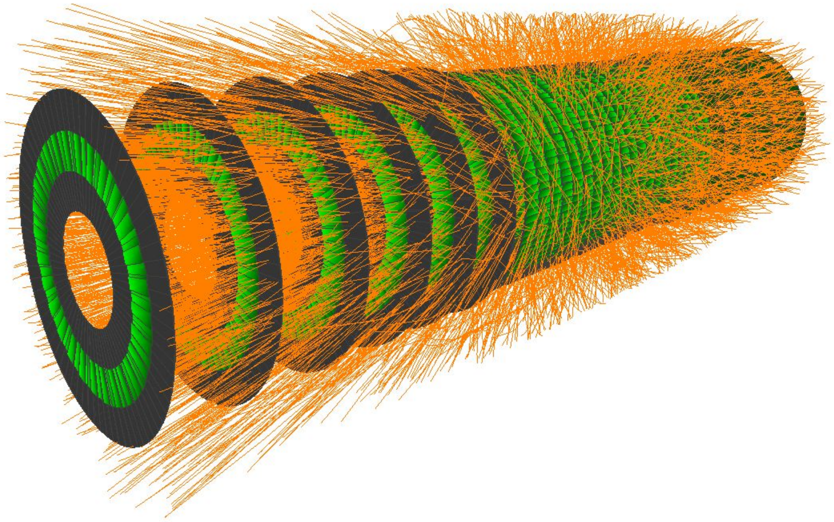
Sabrina Amrouche, T.Golling (UniGe)
A.Salzburger, N.Calace, M.Kiehn (CERN)
c.amrouche@cern.ch

The context

Combinatorial approach

↓

Try all combinations, > 90% are discarded



- High-luminosity LHC (and future FCC) will bring very high pile-up scenarios
- Optimization of combinatorics *limits* the physics

What makes it *interesting* !

TrackML : Solving The **Tracking** Challenge with Machine **Learning**

Accuracy Phase

Throughput Phase

May 2018

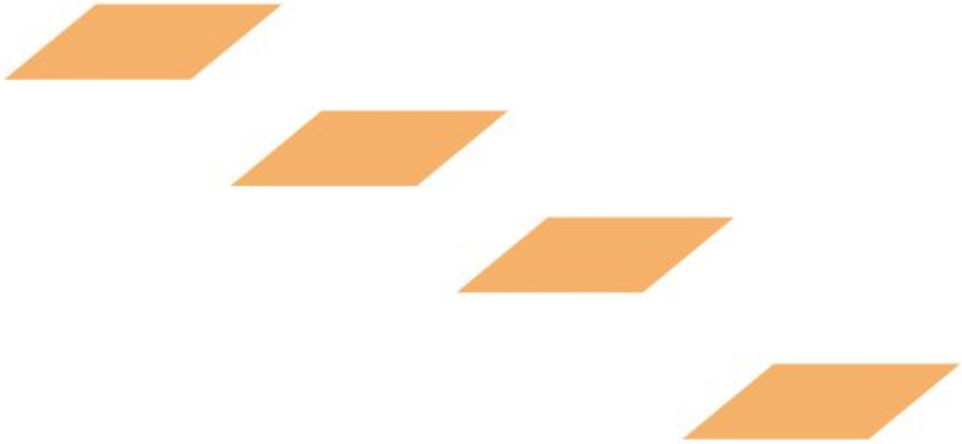
March 2019

Winner *Traditional* Tracking

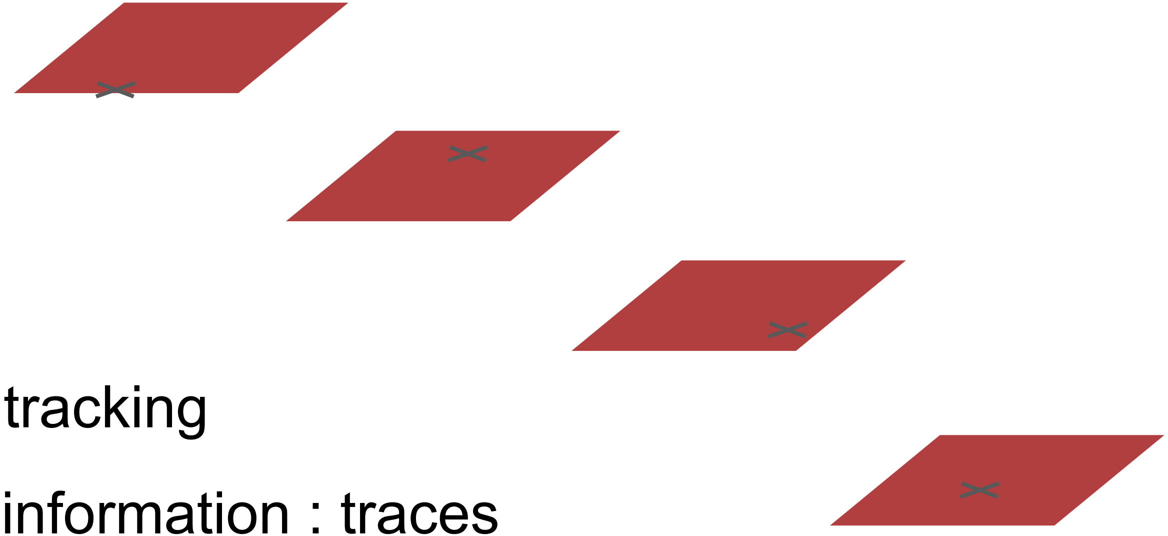
Winner *Traditional* Tracking

arXiv:1904.06778v2

Charged Particle

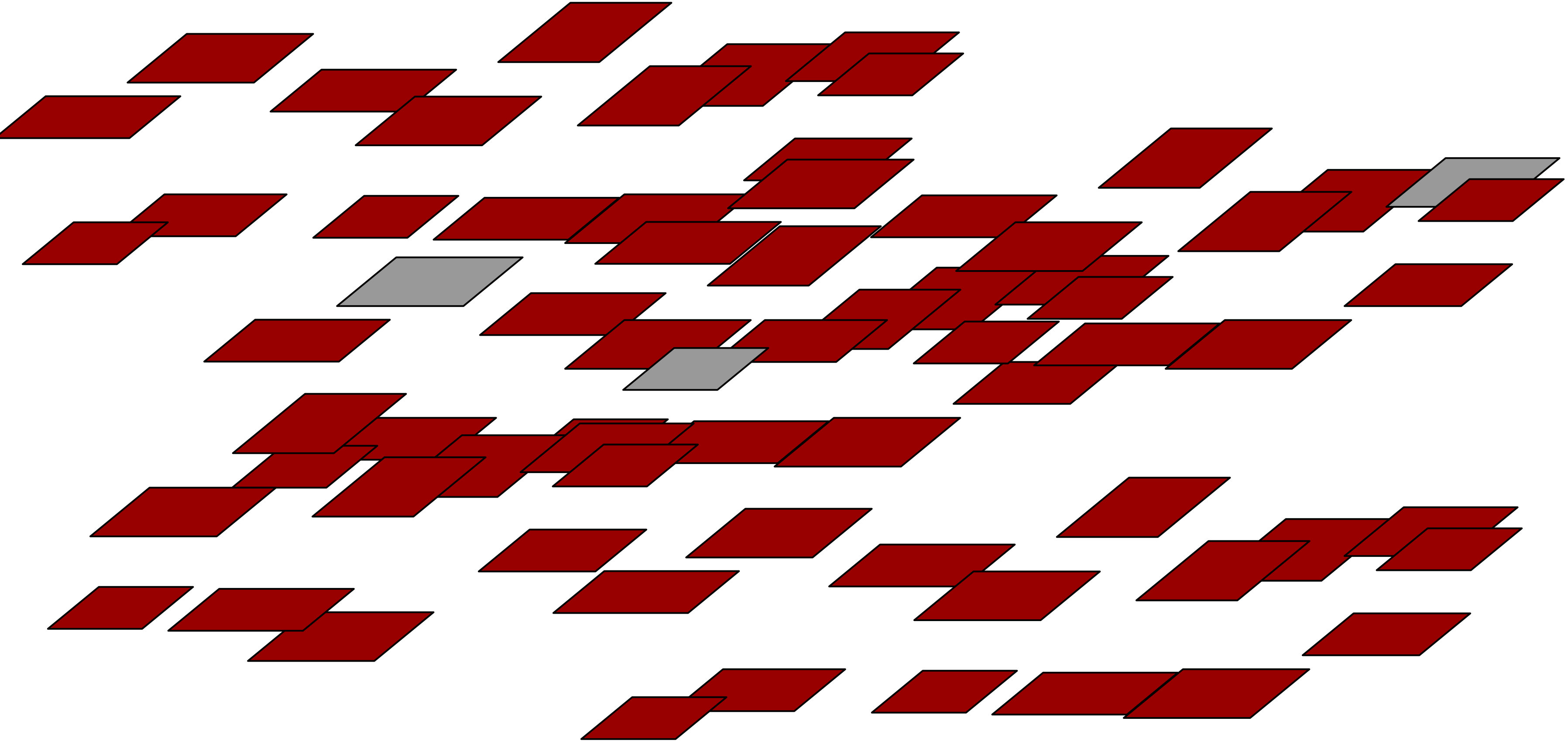


Reconstruction

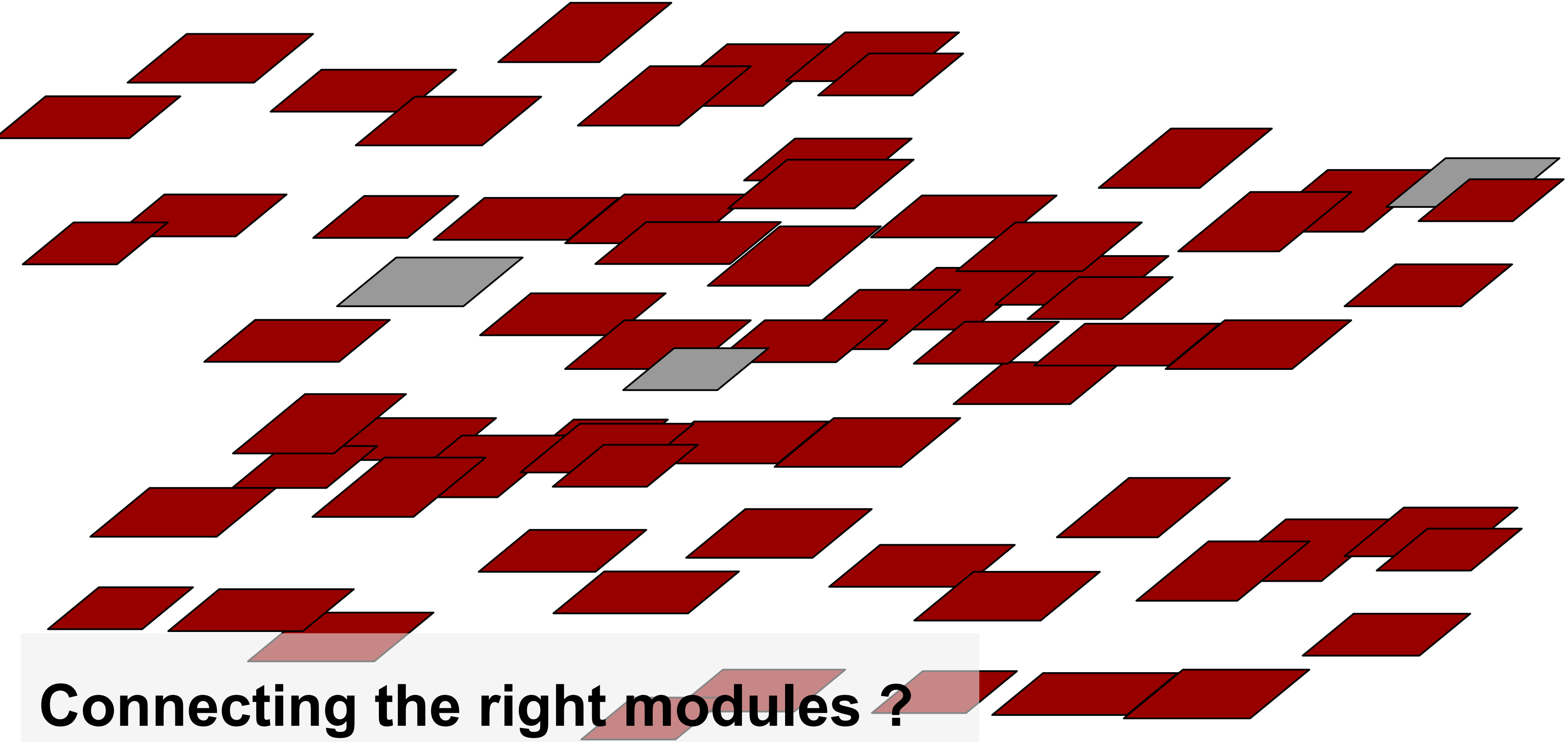


- Offline tracking
- Partial information : traces
- Connecting the right parts

What we get from the detector



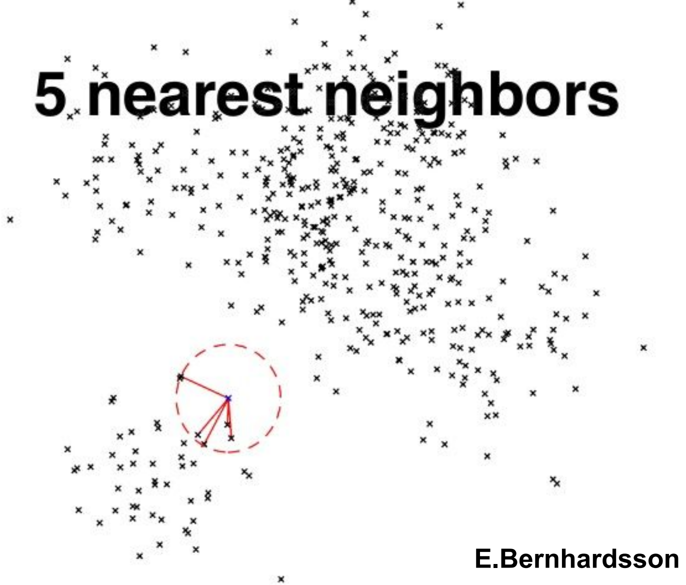
What we get from the detector



Connecting the right modules ?

Fast Similarity Search

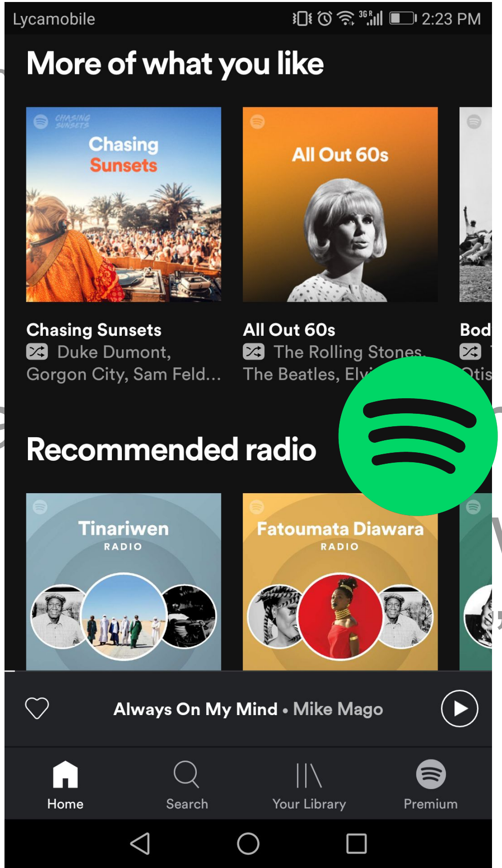
Approximate Nearest Neighbors or Hashing

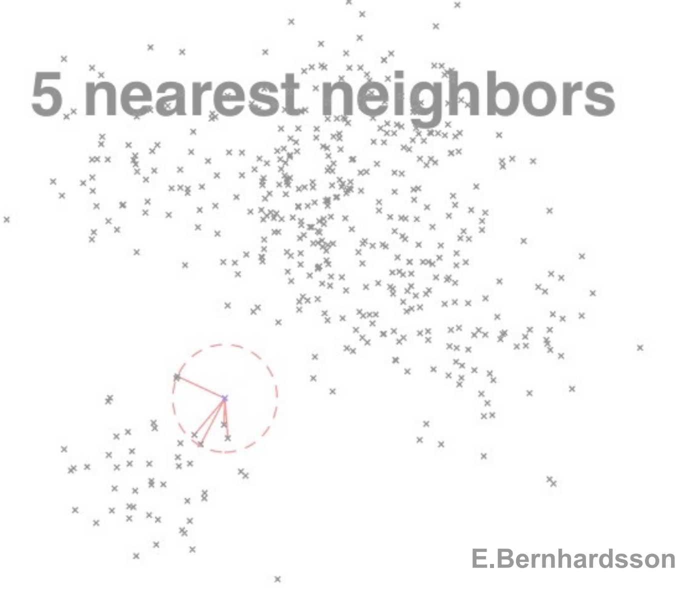


“ Given a query point in a large dataset, returns the set of points with the smallest distance”

Fast Similarity Search

Approximate Nearest Neigh





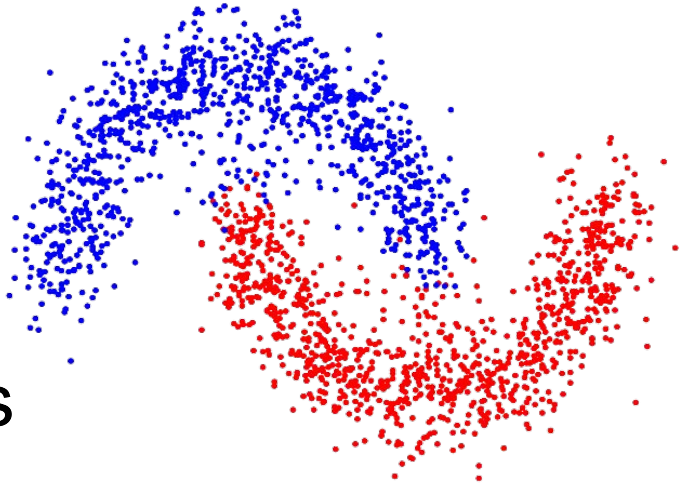
“ Given a … n a large dataset,

returns the … with the smallest

”

Spotify / **Annoy**

Star 5,597

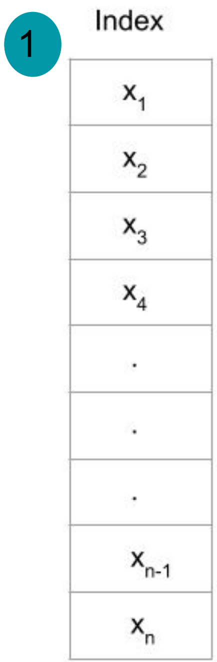


- “Many millions of songs”
- **< 0.1ms** to get **n** similar songs

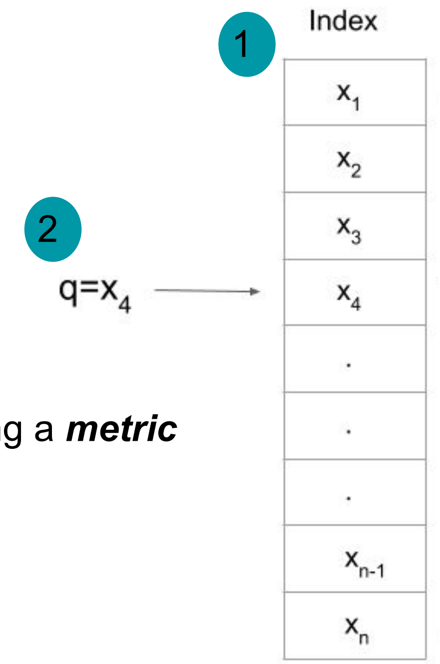
  [high-dimensional space]
- Unsupervised

ANN Strategy

1. Index building using a ***metric***

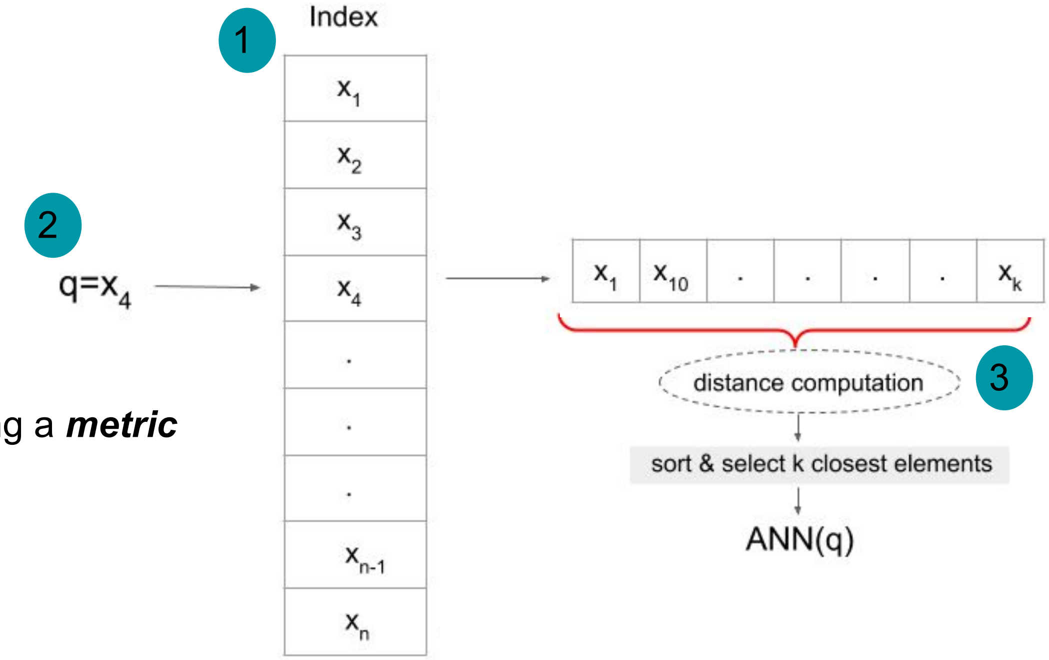


ANN Strategy



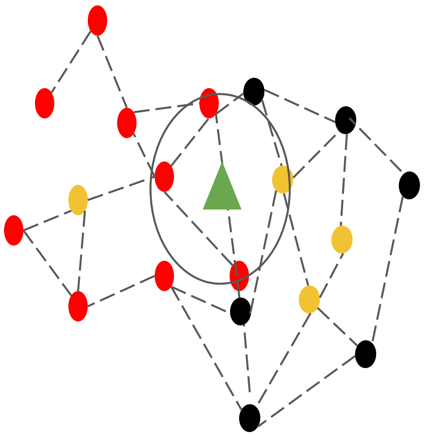
1. Index building using a ***metric***
2. ANN query

ANN Strategy

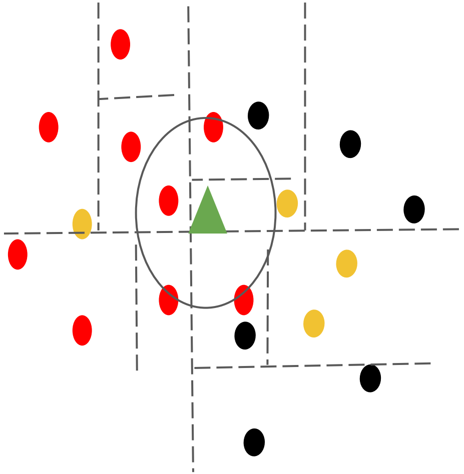


1. Index building using a ***metric***
2. ANN query
3. Get neighbors

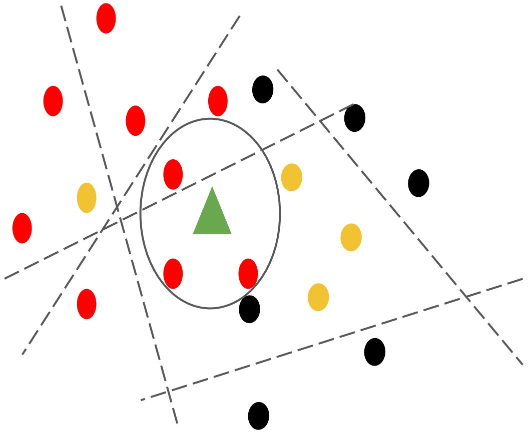
ANN Strategy ：Data structure



Graph



Tree

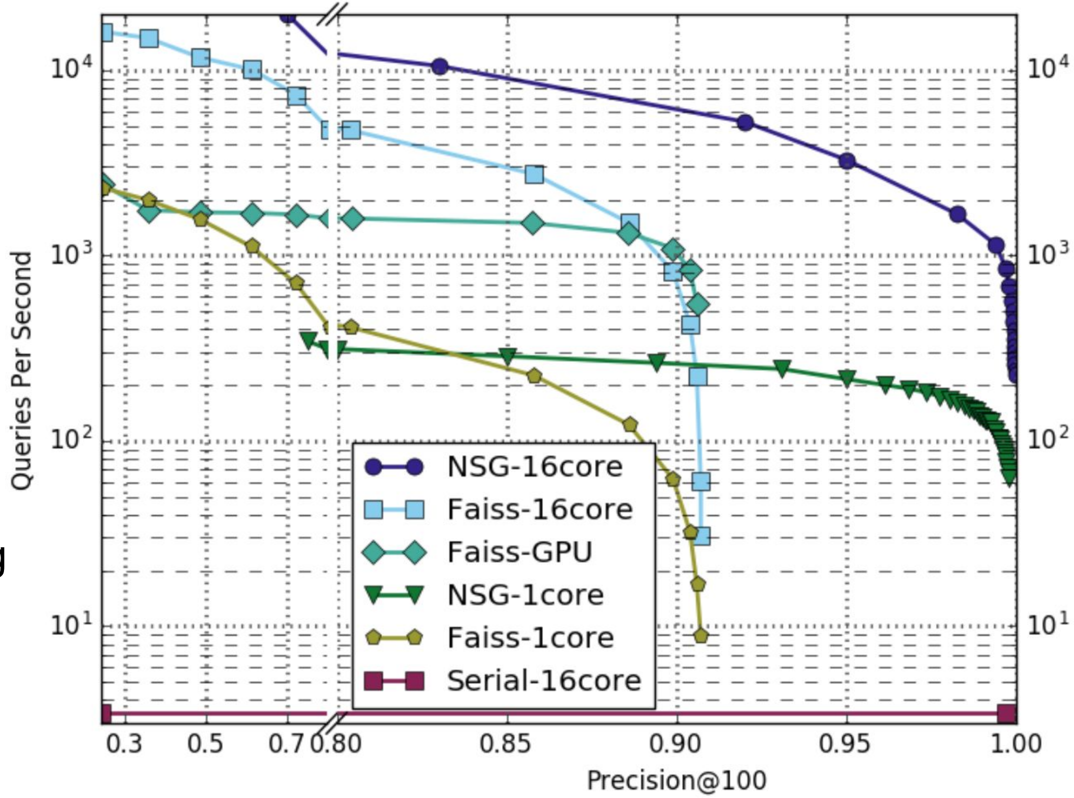


Random Projections

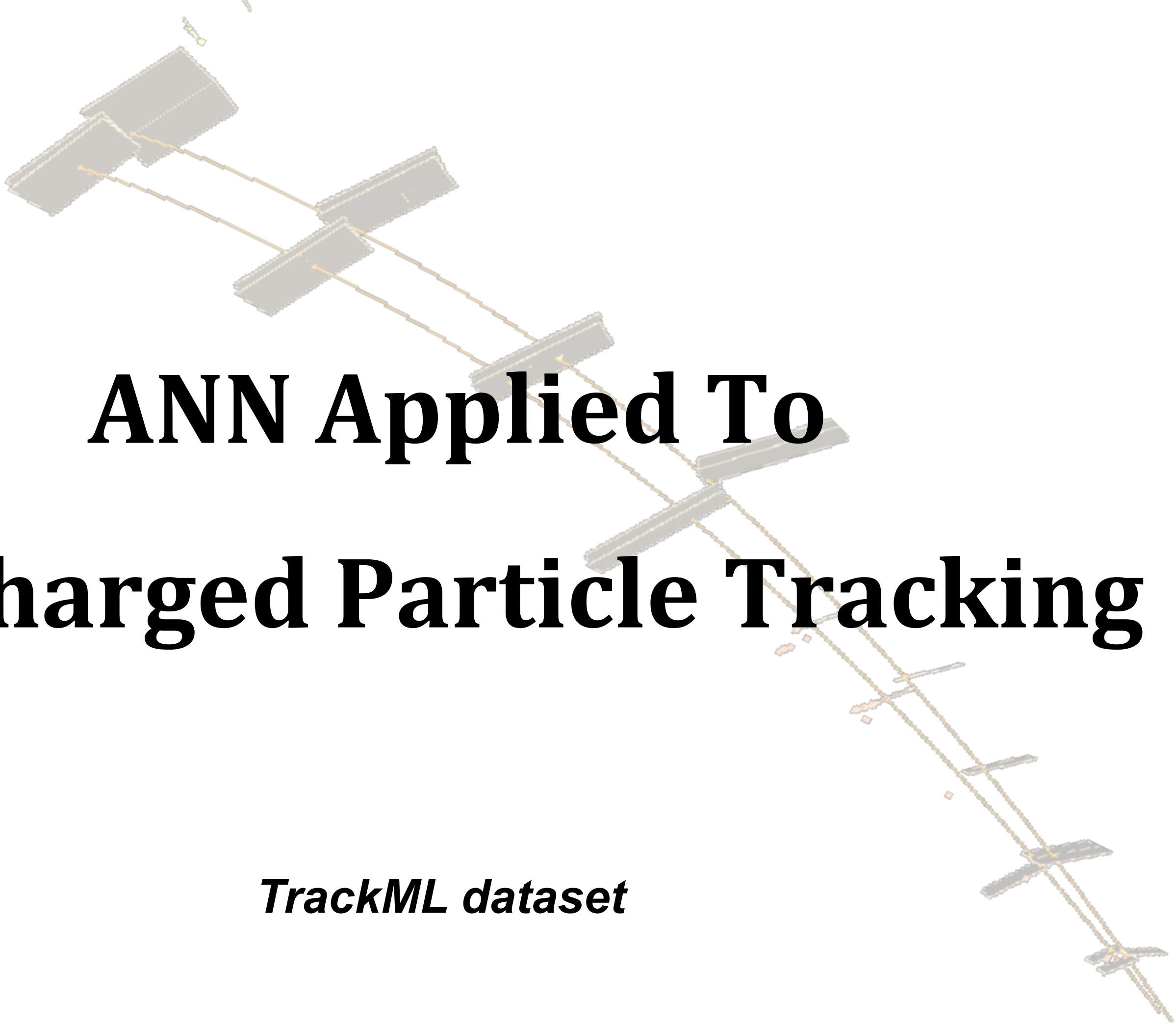
Fast Similarity Search : State of the Art

Benchmark

- **100 million** vector dataset
- **96** dimensions
- 1 core, 16 core and GPU
- Precision is irrelevant in tracking
- > 10^4 queries per second



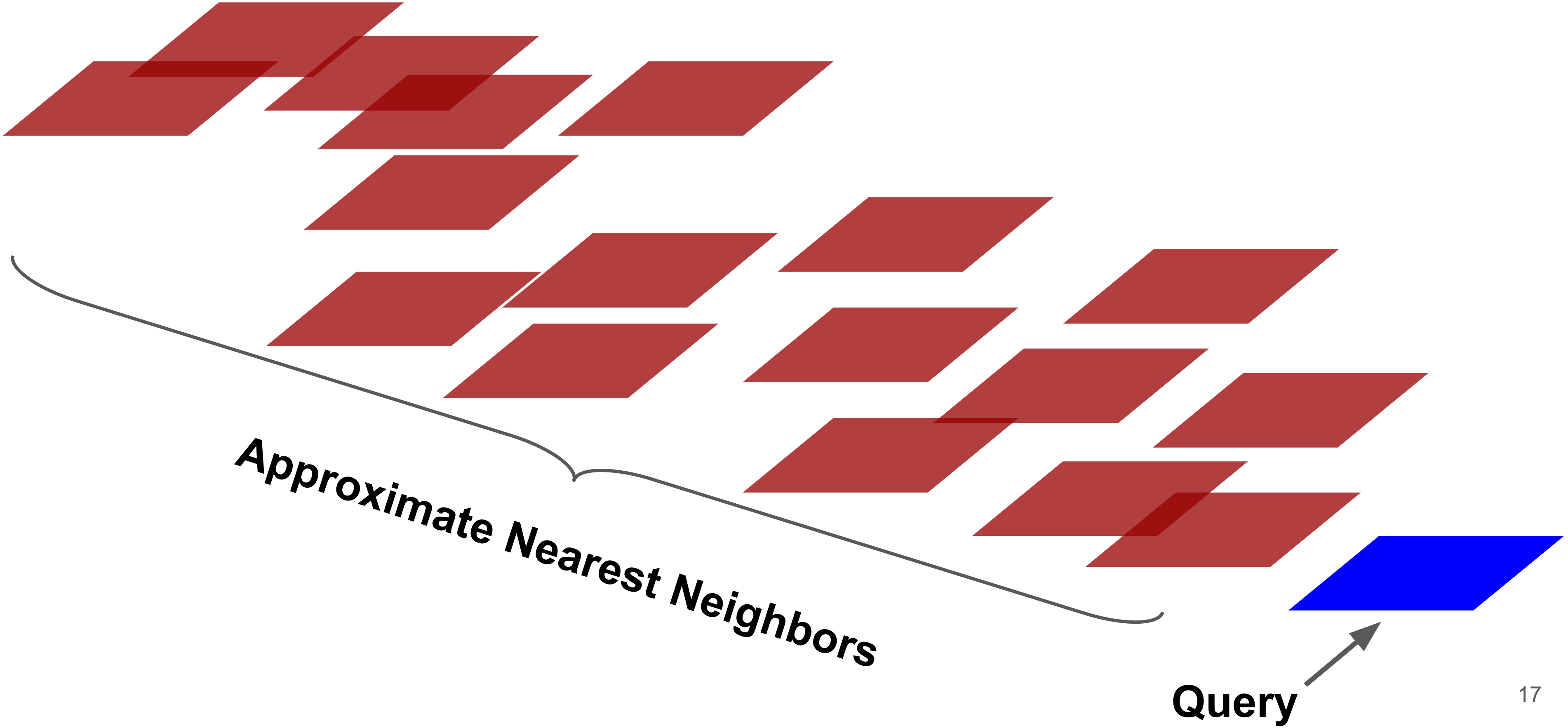
arXiv:1707.00143



ANN Applied To Charged Particle Tracking

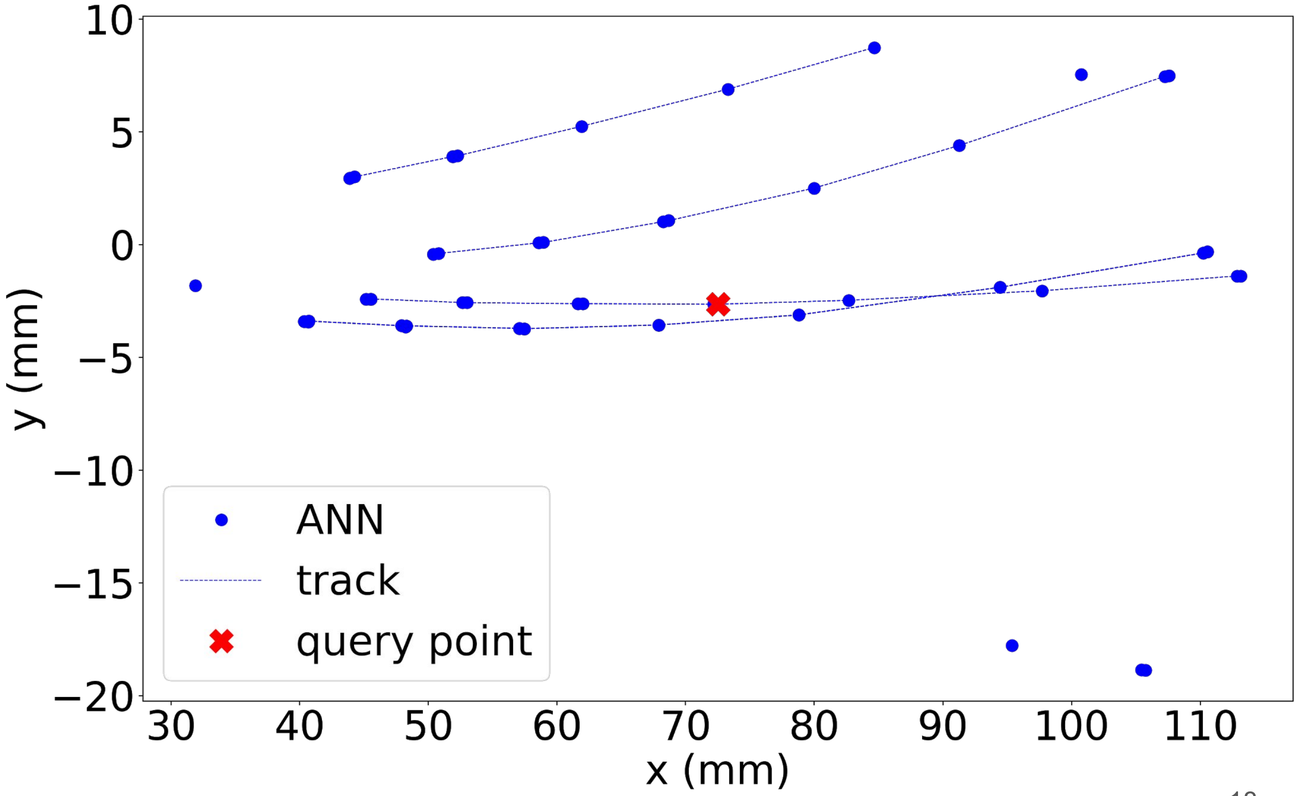
TrackML dataset

ANN Buckets

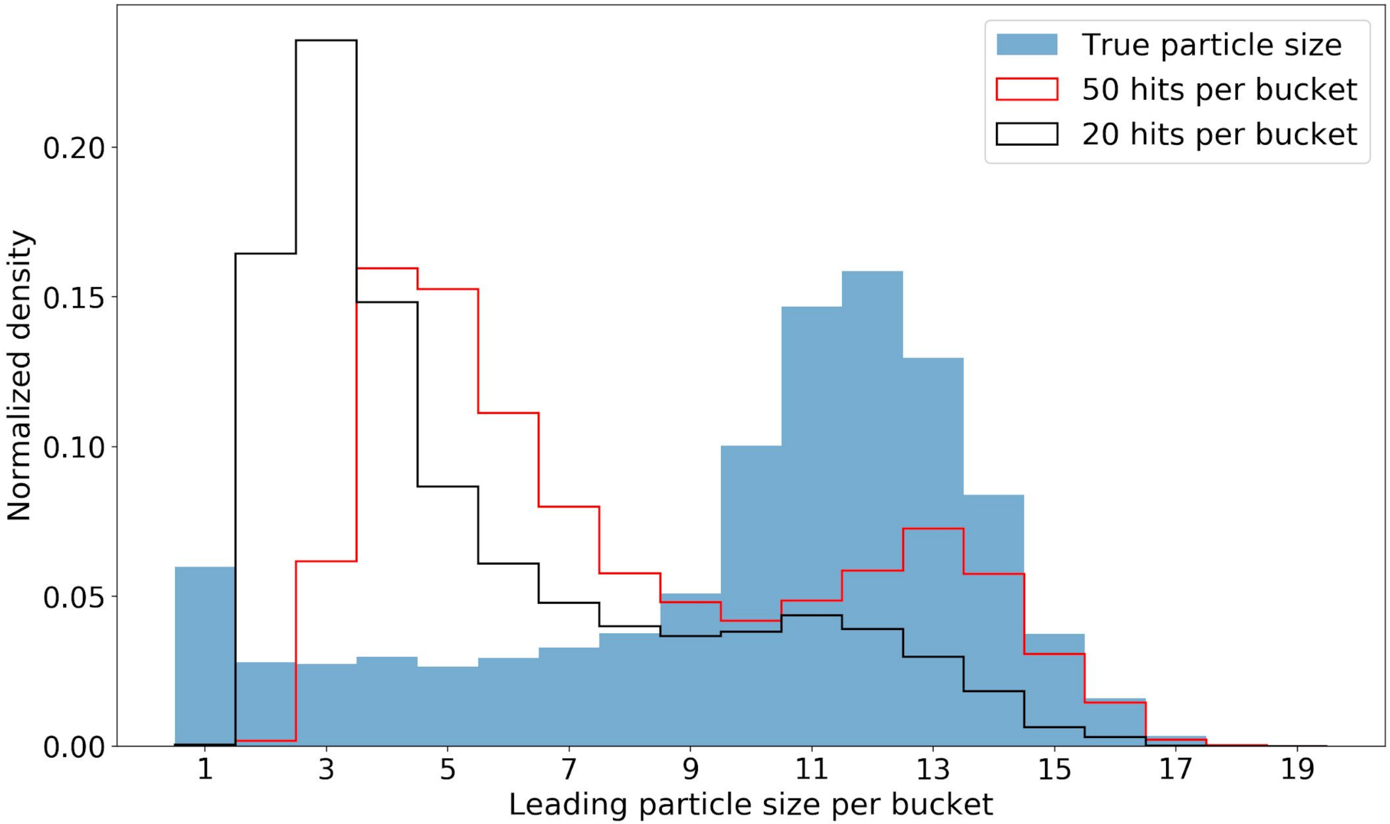


ANN Buckets

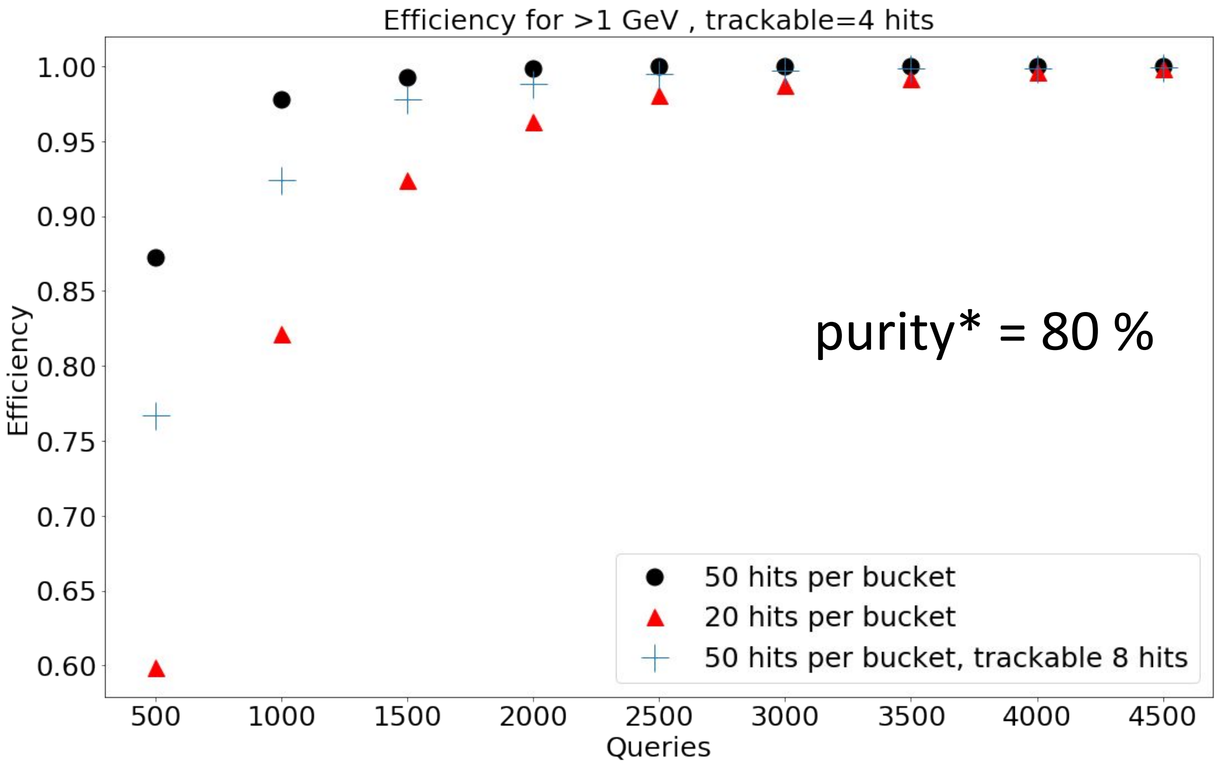
- **50** hits bucket
- Angular distance as metric
- Leading particle size **11**



Buckets Quality



ANN Efficiency

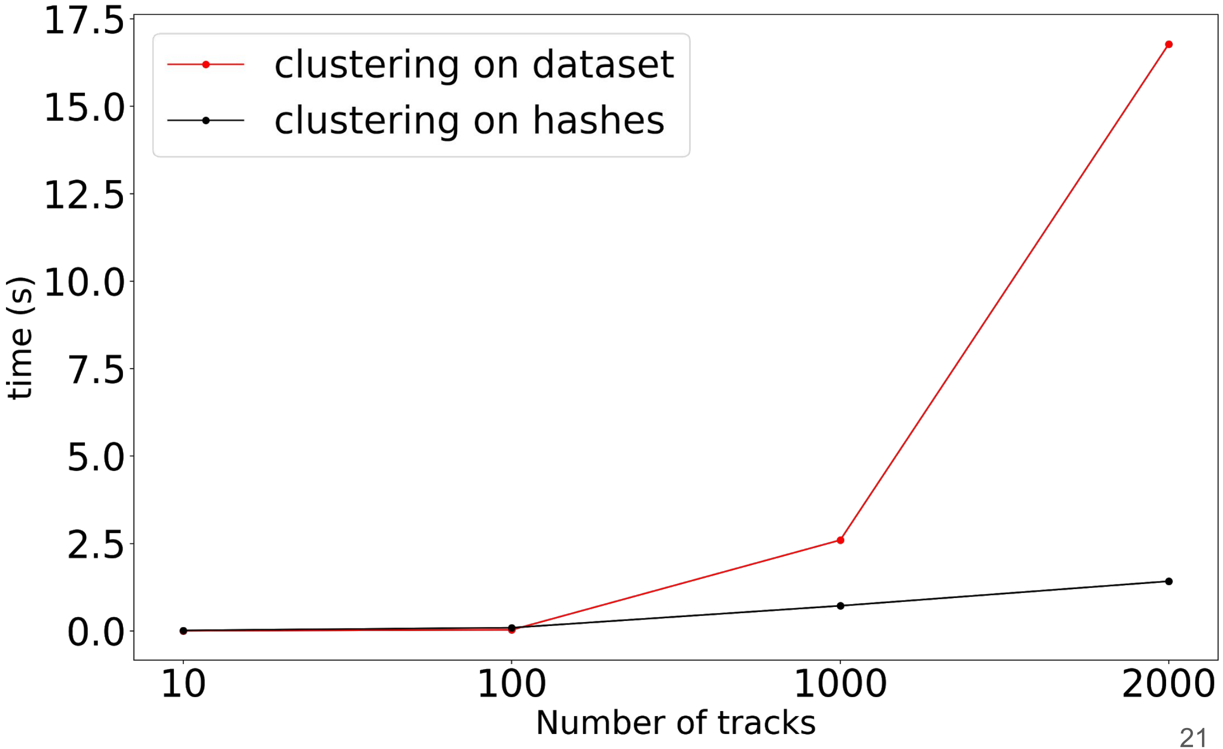


- Assuming perfect in-bucket tracking
- Particle ***found*** if ≥80% in bucket
- Trackable = Min particle size

* A track will be marked as reconstructed if 80 % of its hits are found inside the same bucket

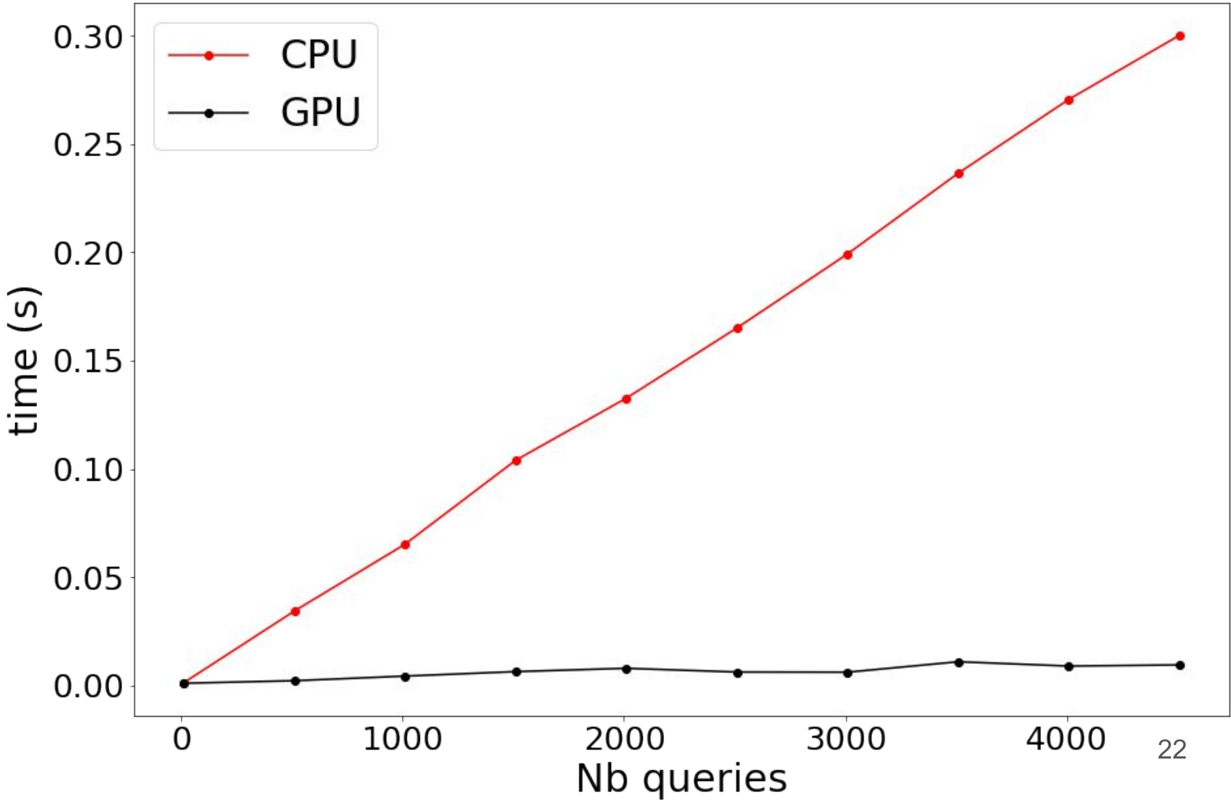
Scaling Of ~~Tracking~~ Clustering : Full Event vs Buckets

- Agglomerative Clustering (AC) as proxy for standard tracking
- AC is $O(n^2)$, it computes ~all pairwise distances



ANN batch mode for GPU - CPU

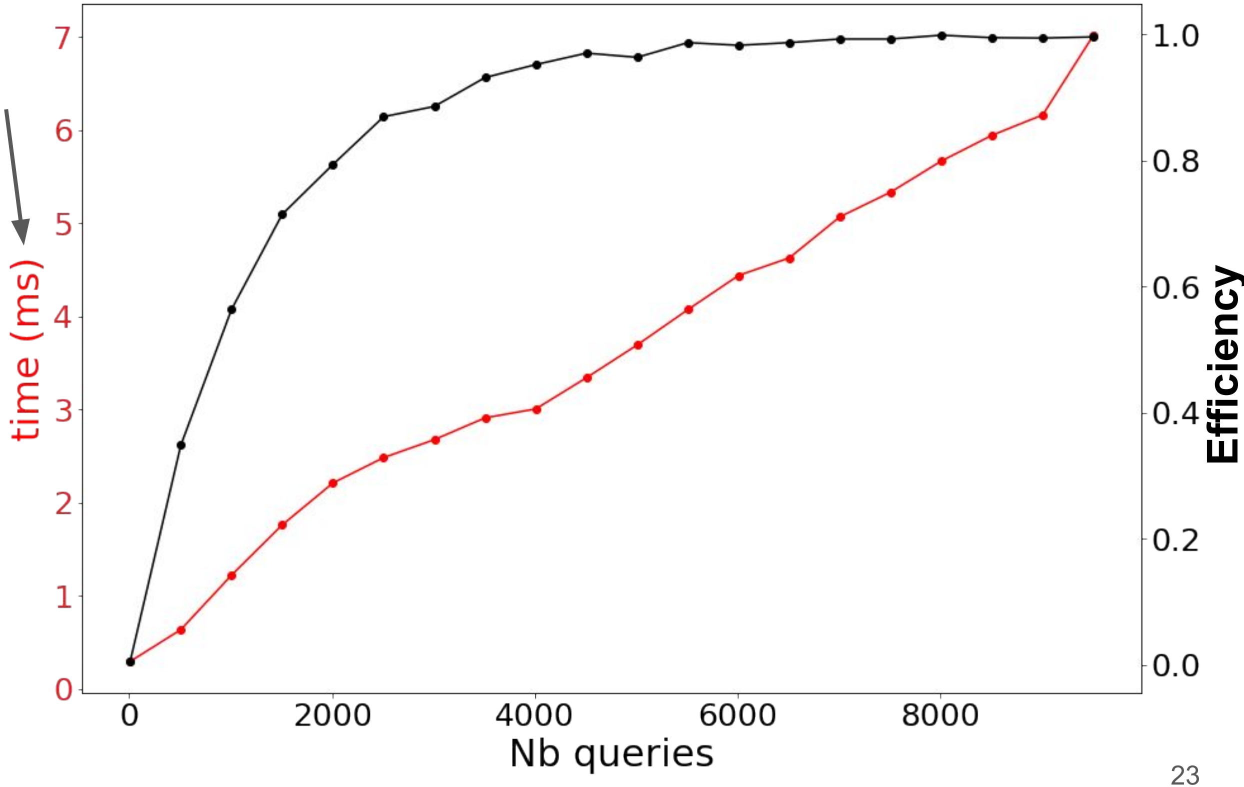
- Bucketing (only) scaling on CPU vs GPU.
- Implementation in Python.
- Hardware: NVIDIA Tesla K40m, 12 GB RAM, 2880 CUDA core.

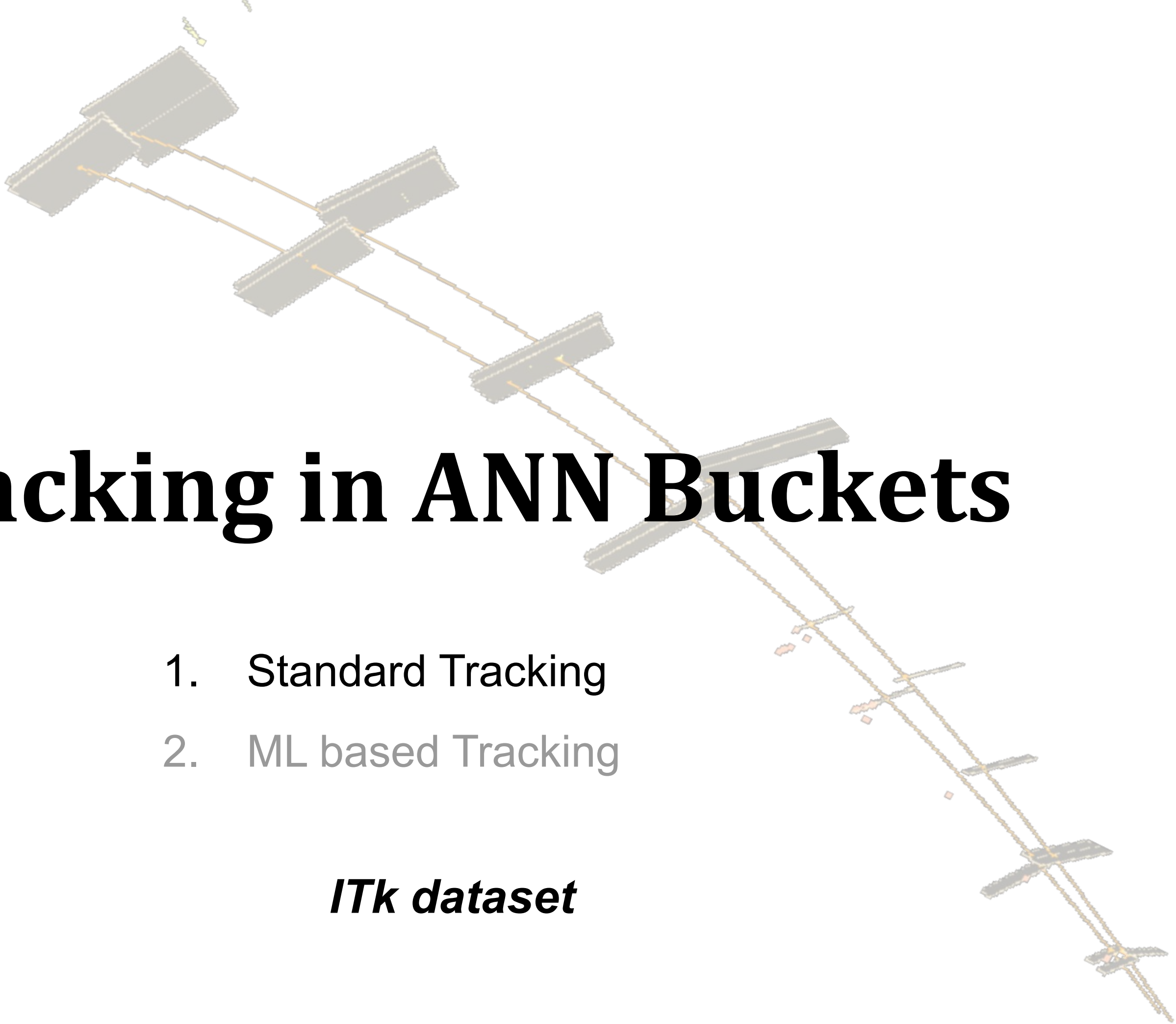


ANN batch mode for GPU

- Bucketing (only) scaling on **GPU**.
- Implementation in Python.
- Hardware: NVIDIA Tesla K40m, 12 GB RAM, 2880 CUDA core.
- Assuming perfect in-bucket

  tracking. Purity 80%.





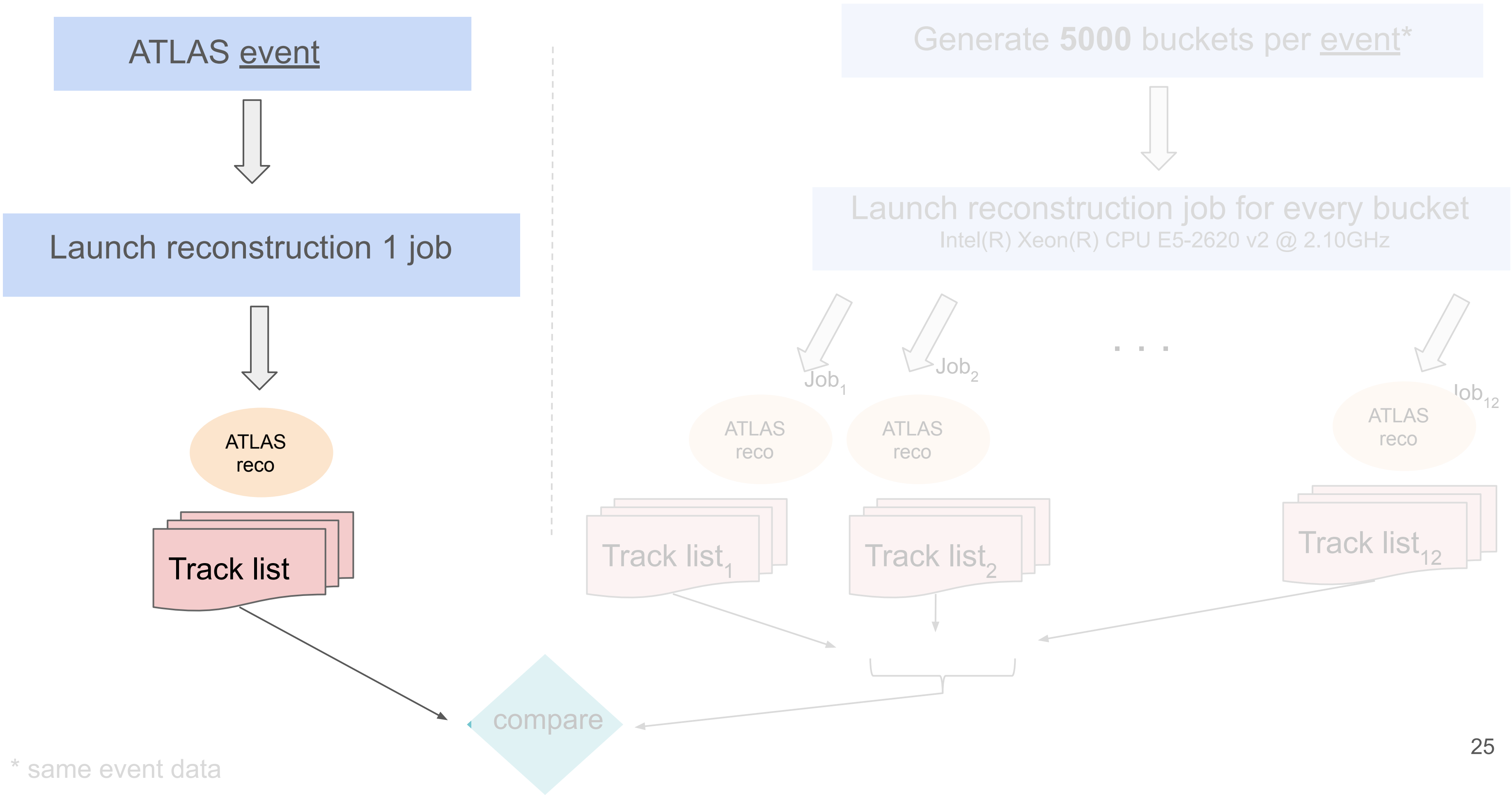
Tracking in ANN Buckets

1. Standard Tracking
2. ML based Tracking

ITk dataset

Tracking in buckets

(1) Standard ATLAS Phase-2 reconstruction

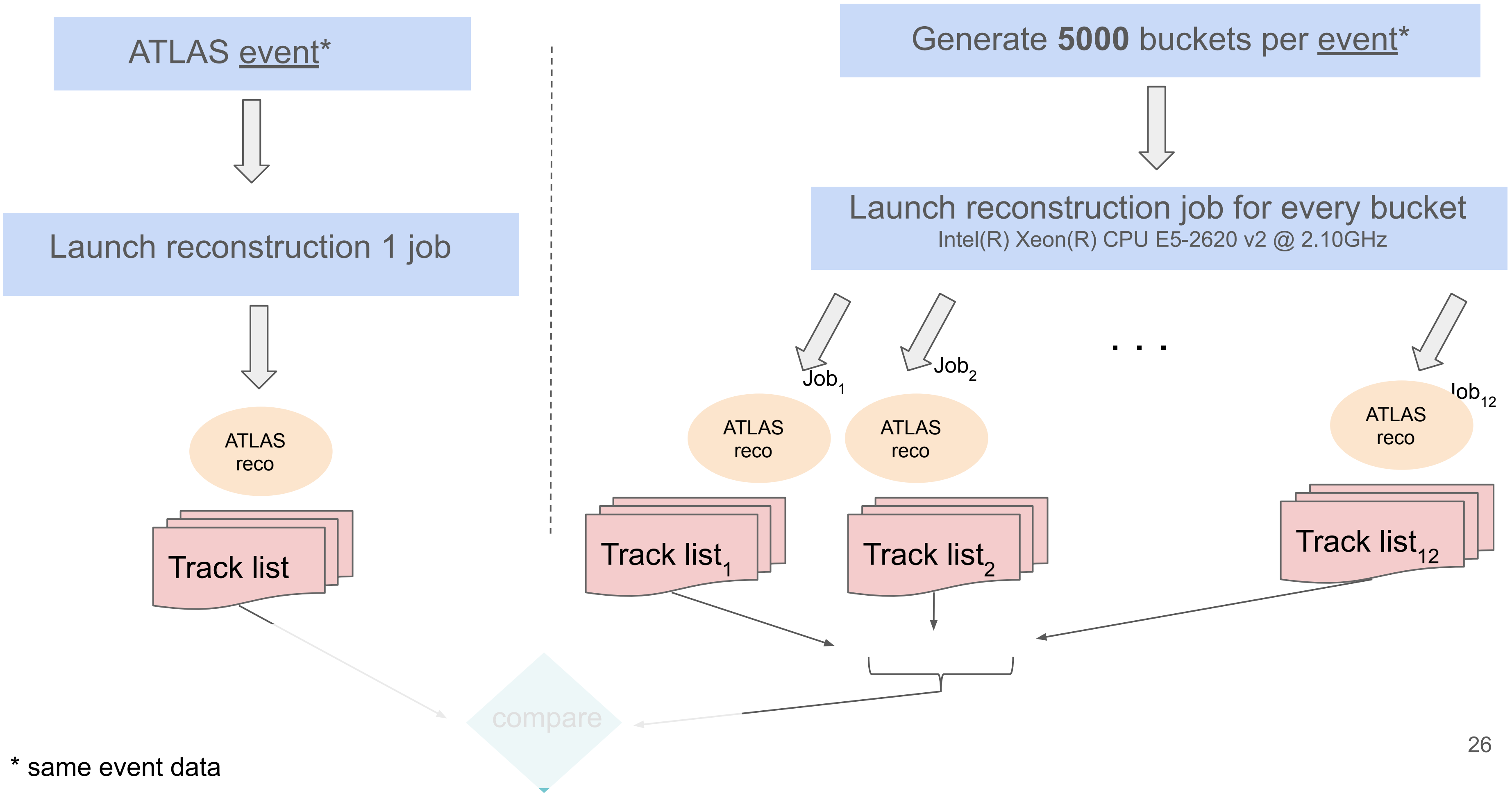


* same event data

Tracking in buckets

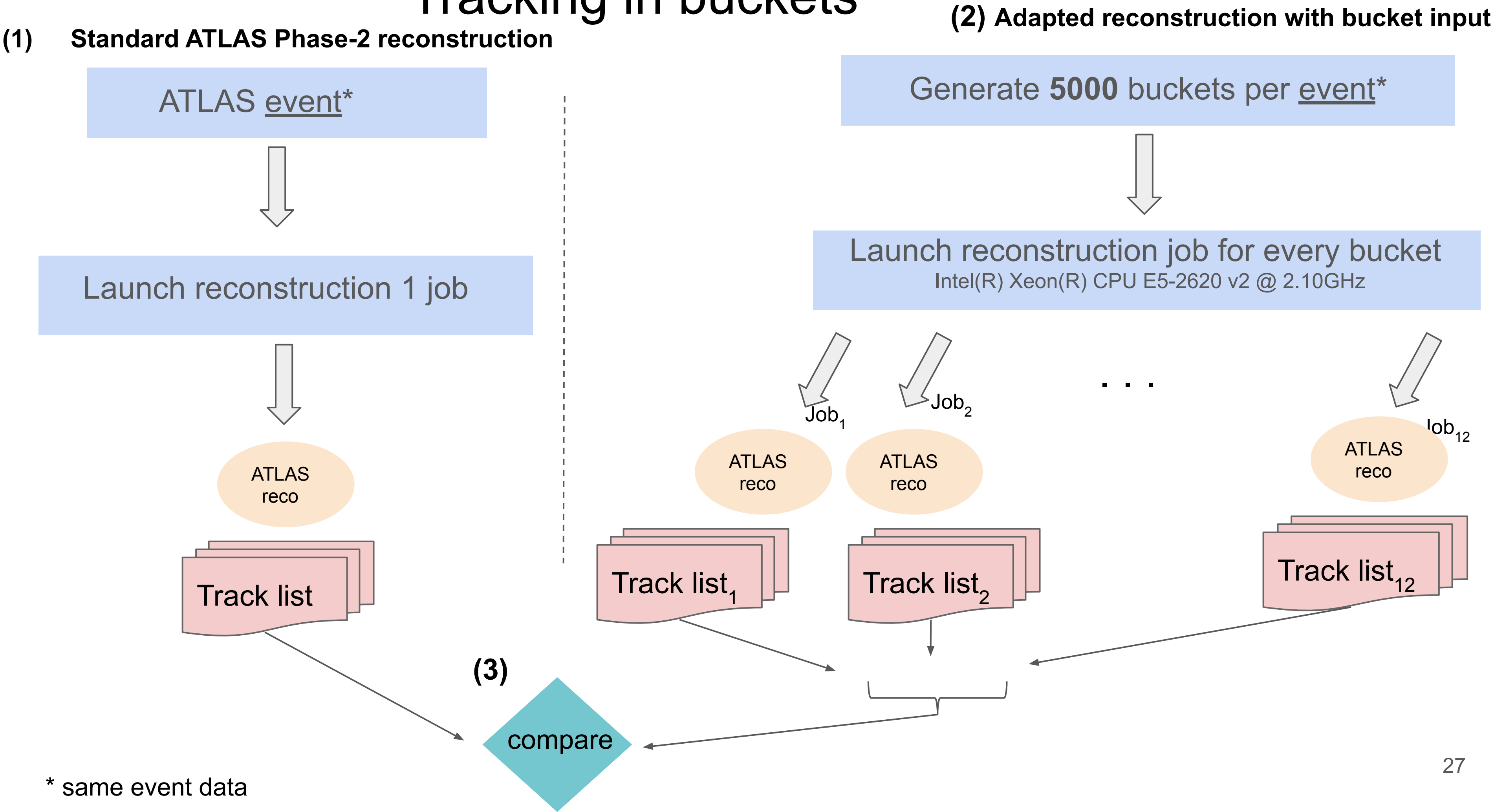
(1) Standard ATLAS Phase-2 reconstruction

(2) Adapted reconstruction with bucket input



* same event data

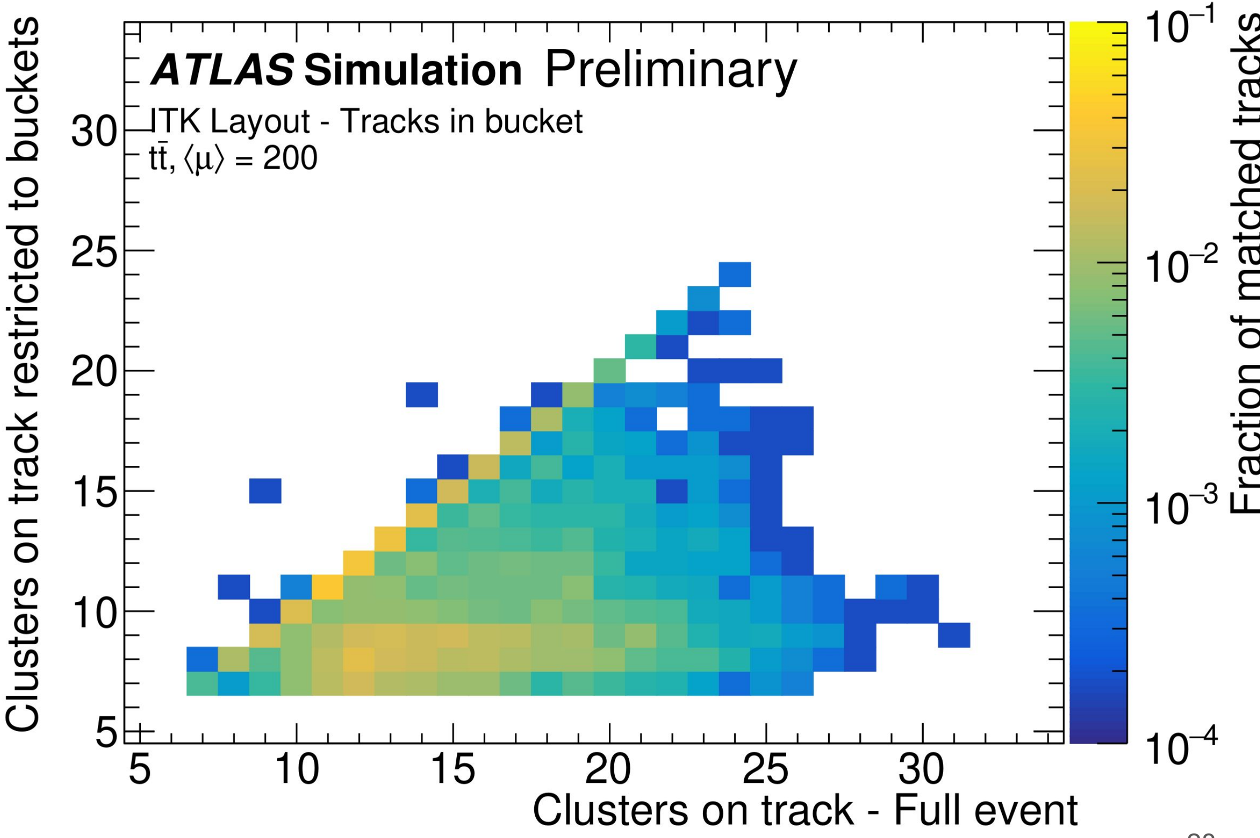
Tracking in buckets



* same event data

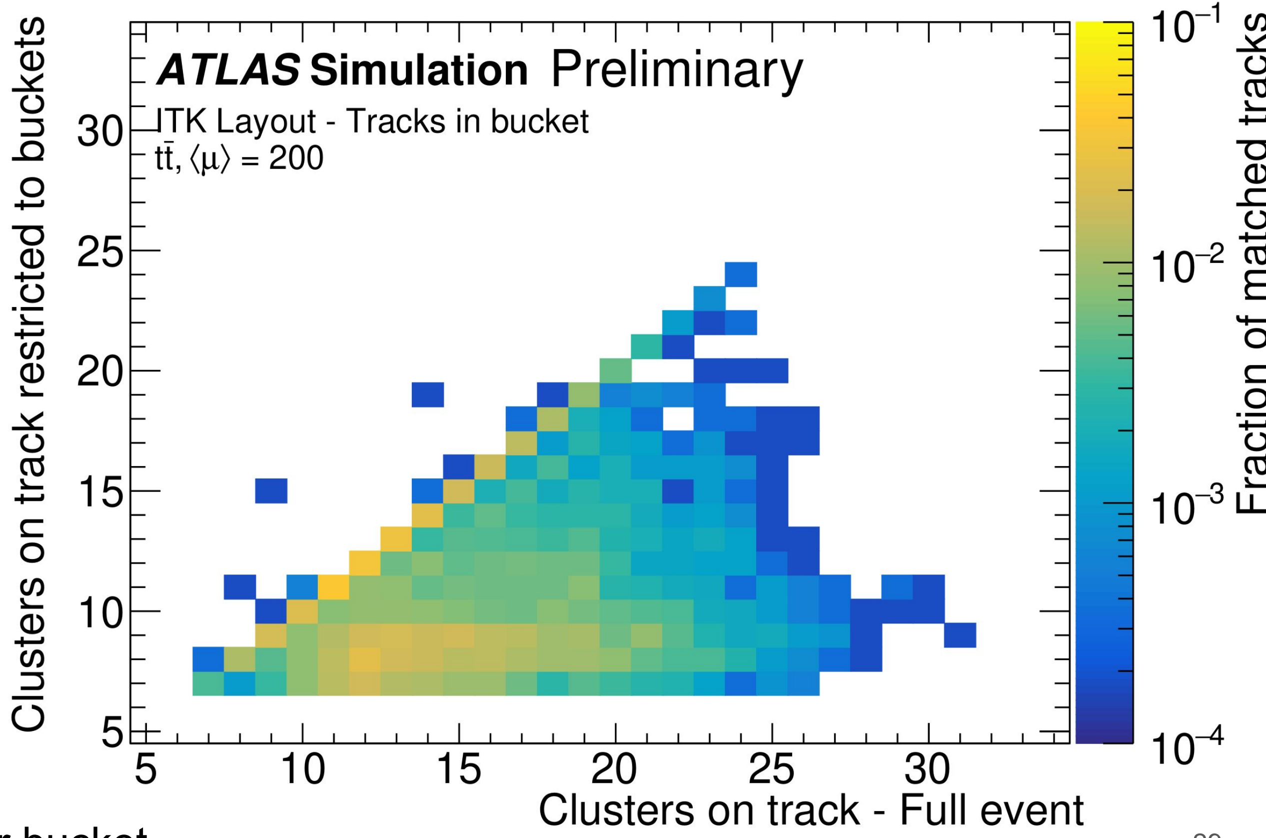
ATLAS Reconstruction in bucket

Number of clusters on tracks running the full event reconstruction versus restricting the track reconstruction algorithms to a bucket of 50 hits.



ATLAS Reconstruction in bucket

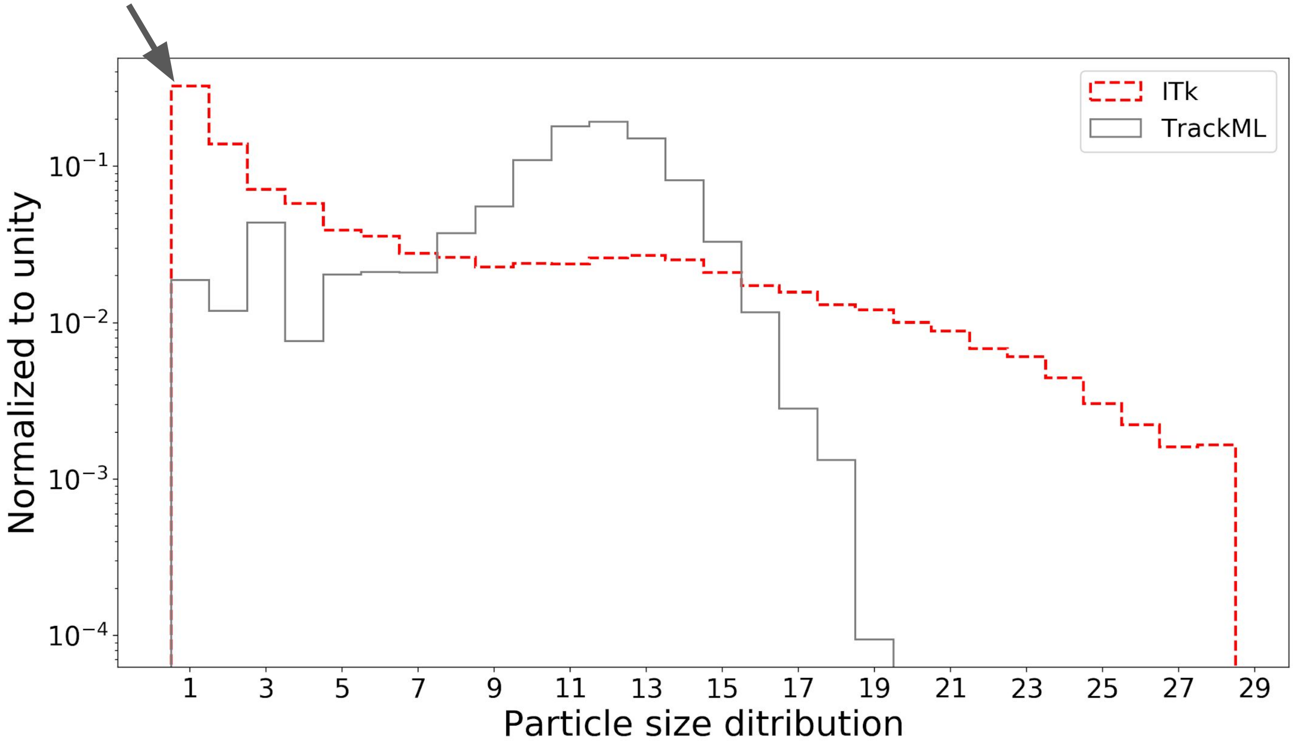
Number of clusters on tracks running the full event reconstruction versus restricting the track reconstruction algorithms to a bucket of 50 hits.



ATLAS Tracking takes **~ 2ms** per bucket

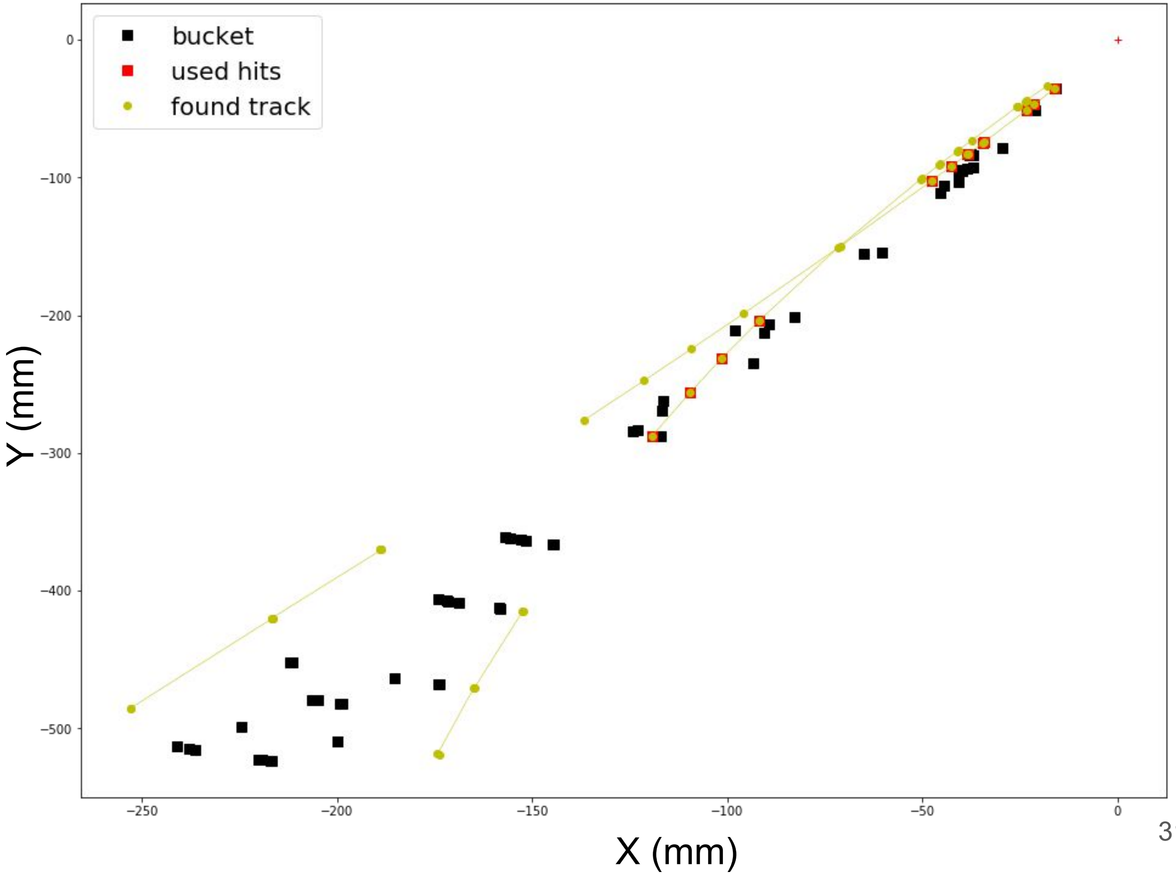
Tracks in Buckets : TrackML vs ITk

- Noise hits are hits associated to non reconstructable particles (<4 hits).
- In TrackML ***noise*** is kept to a minimum of **~10%**.
- ITk has **~50%** noise hits per event.



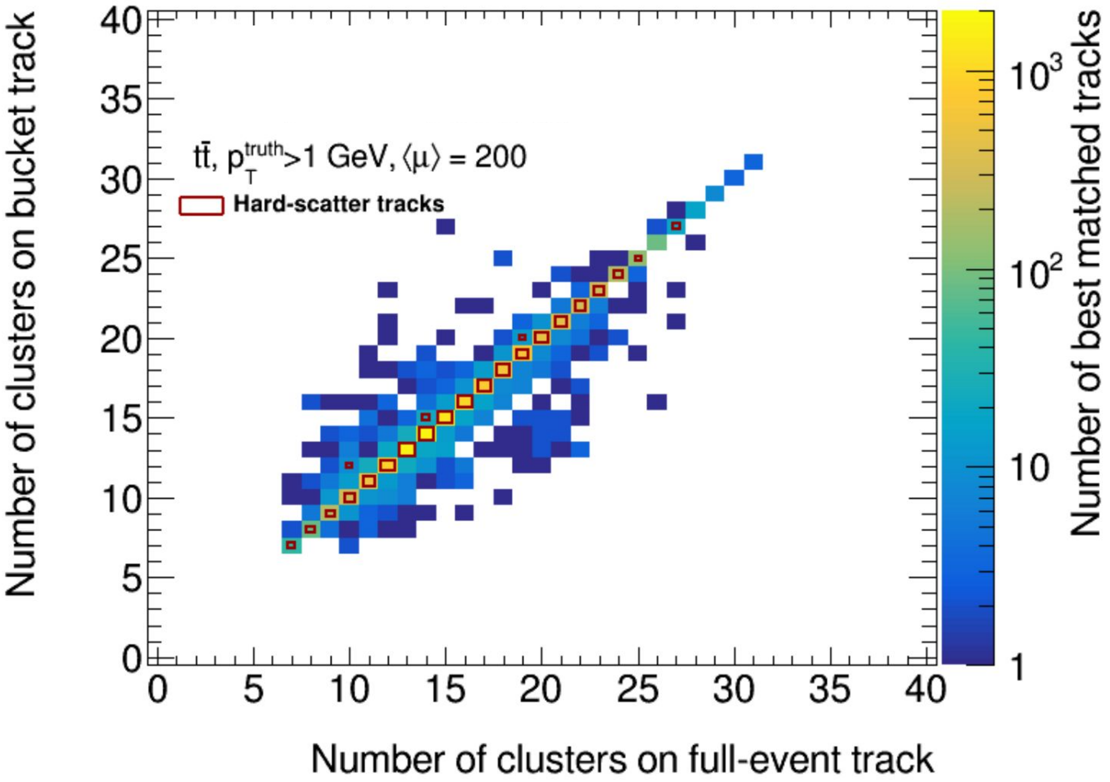
Seeding in Buckets - Standard Tracking

- Only pixel seeds built from the bucket.
- The Track Finder completes the trajectory with access to the full event.
- **75K** buckets (filtering mechanism can help)



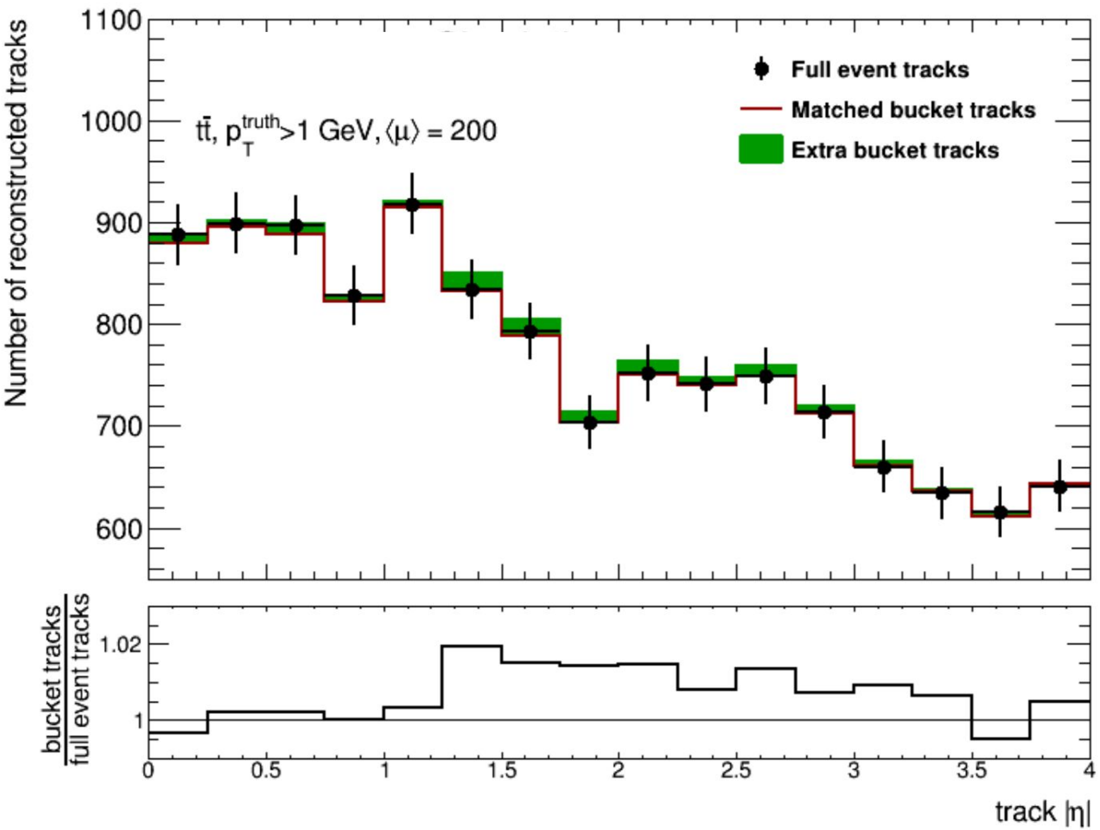
Seeding in Buckets - Standard Tracking

Number of clusters on tracks running the full event reconstruction versus restricting the **seeding** to a bucket of **70 hits**.



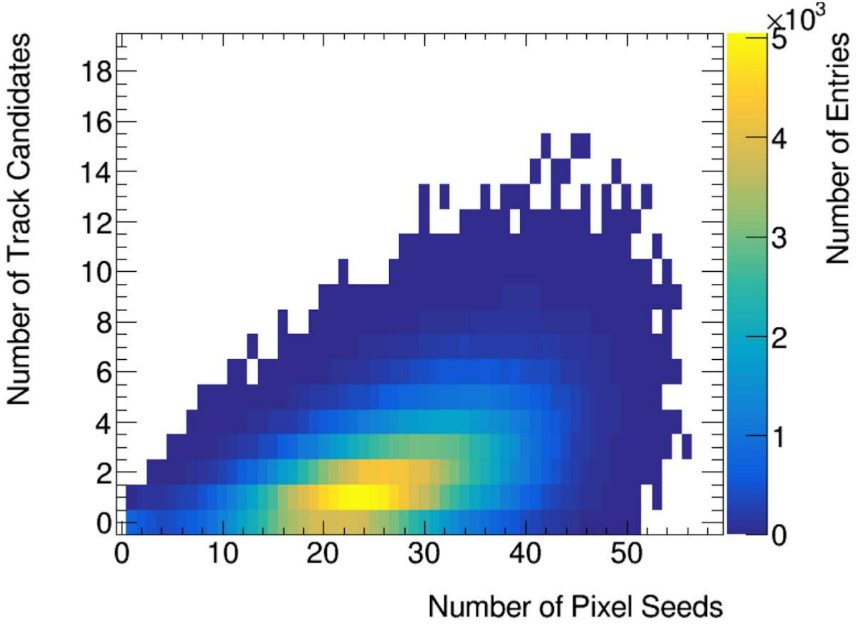
Seeding in Buckets - Standard Tracking

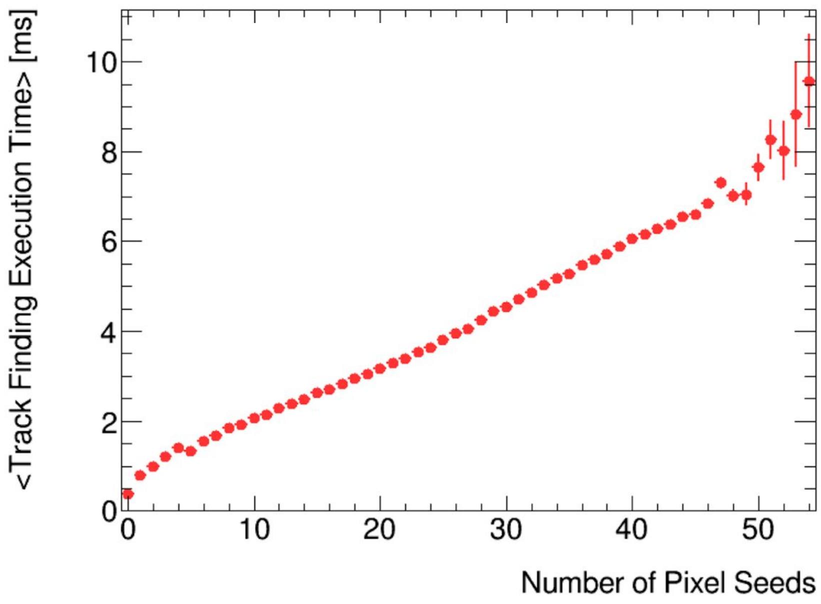
- Seeding cut in P_T reduced compared to Standard Tracking (900MeV → 400MeV).
- Extra tracks as a result of cuts loosening and *more pure* seeding environment.
- Mostly in the low P_T spectrum.

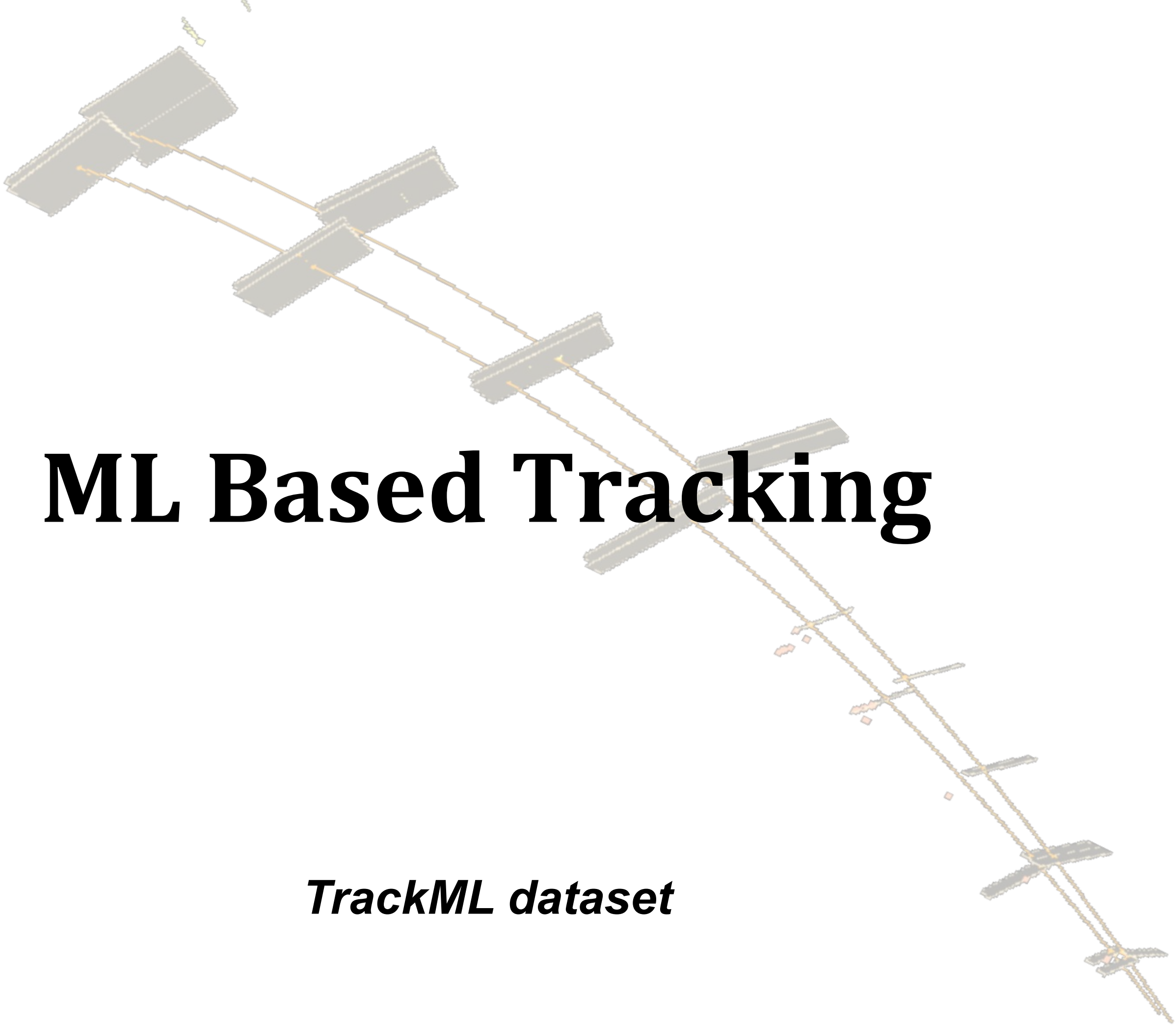


Seeding in Buckets - Standard Tracking

Pixel Seeds to Track Candidates





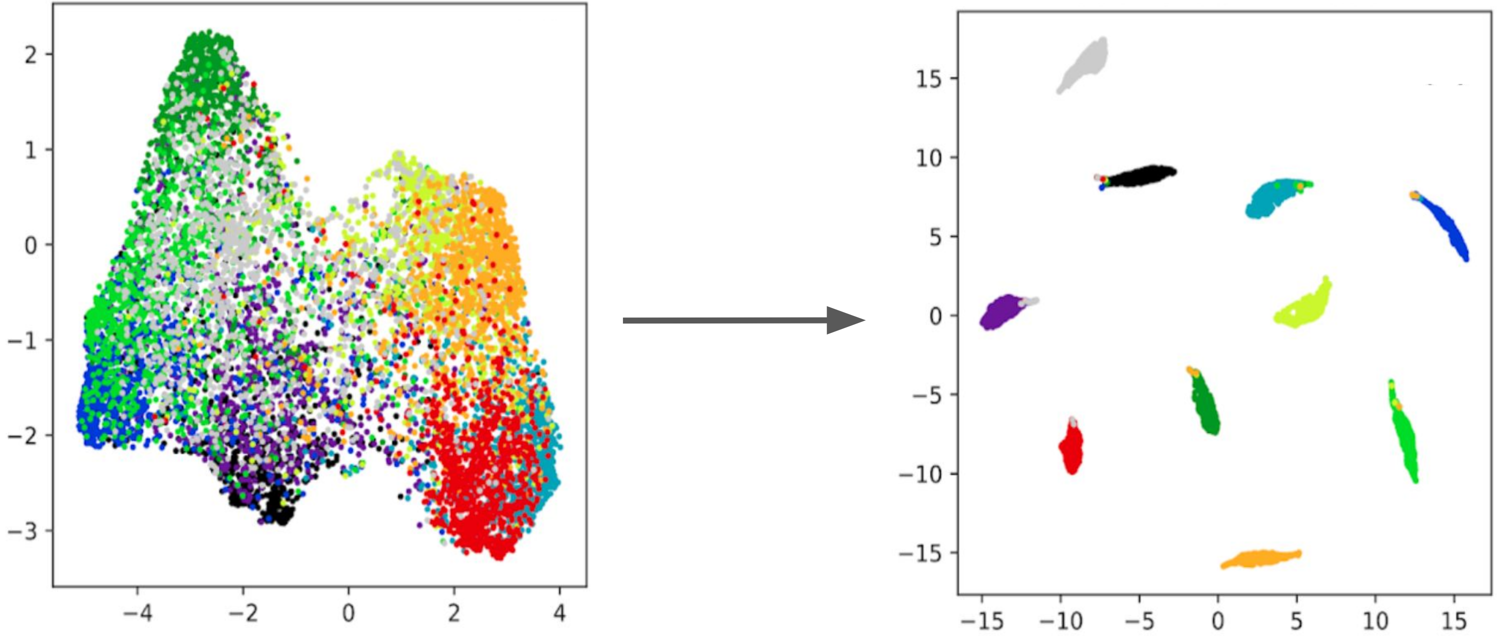


ML Based Tracking

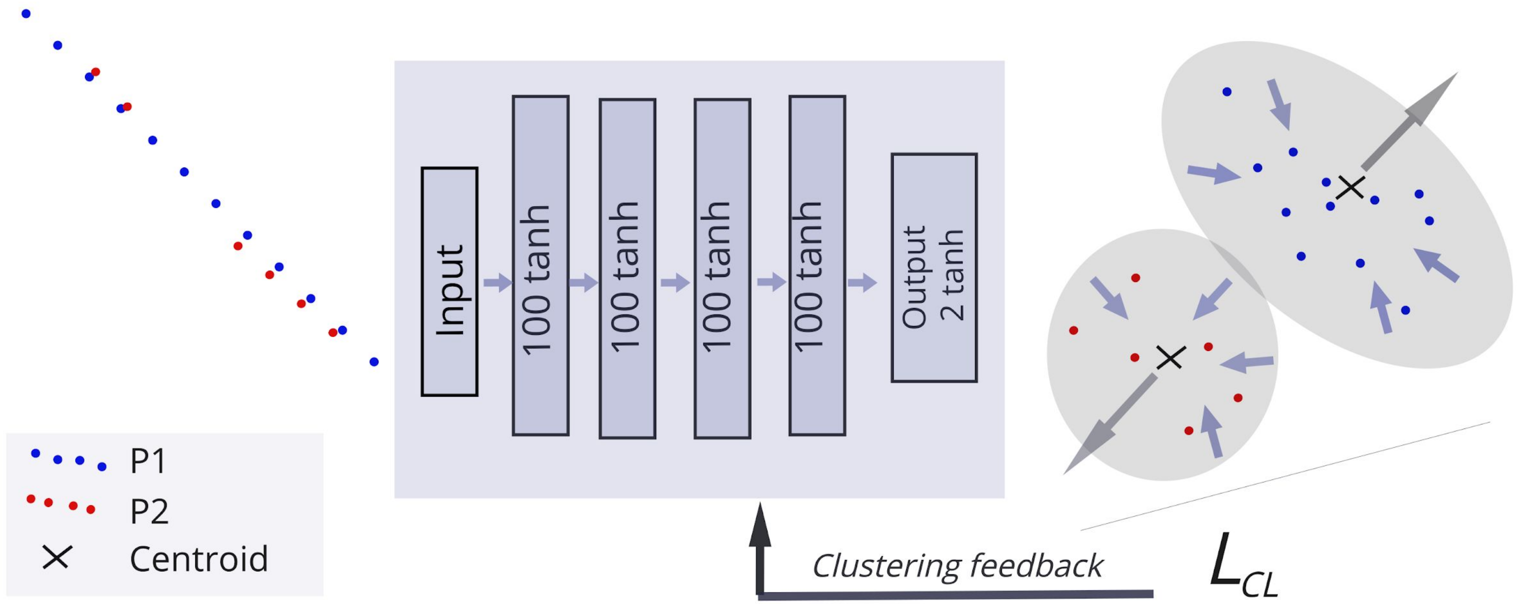
TrackML dataset

Metric (similarity) Learning

Knowing the truth association from simulation, we can **learn** the patterns to map **hits to particles**.



TrackNet : Tracking aware ML



TrackNet : Tracking aware ML

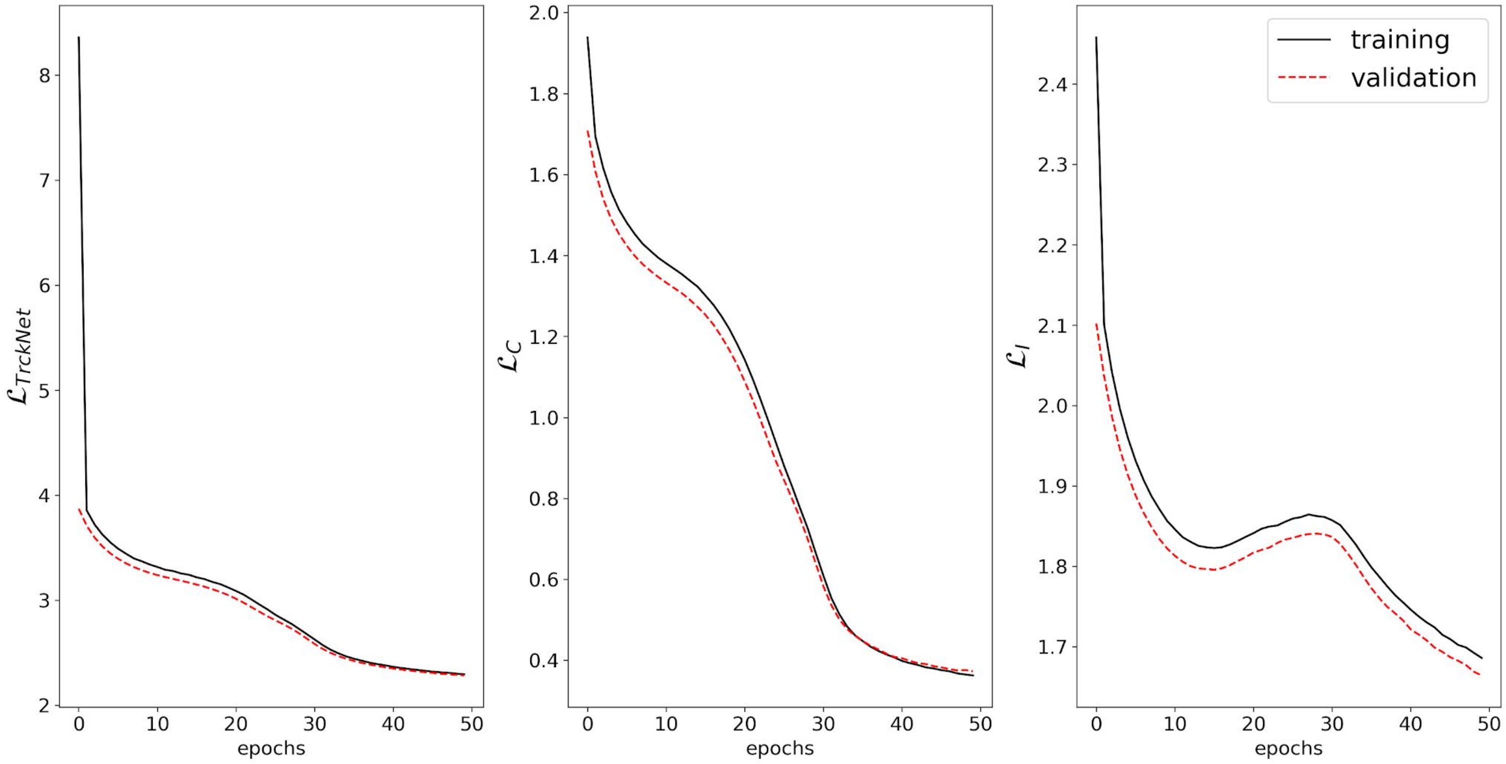
The Loss Function Design : How to **communicate** our **preferences** to the model.

$$\mathcal{L}_{TrackNet} = (\alpha\mathcal{L}_C + \beta\mathcal{L}_I + \gamma\mathcal{L}_{CL})^\zeta$$

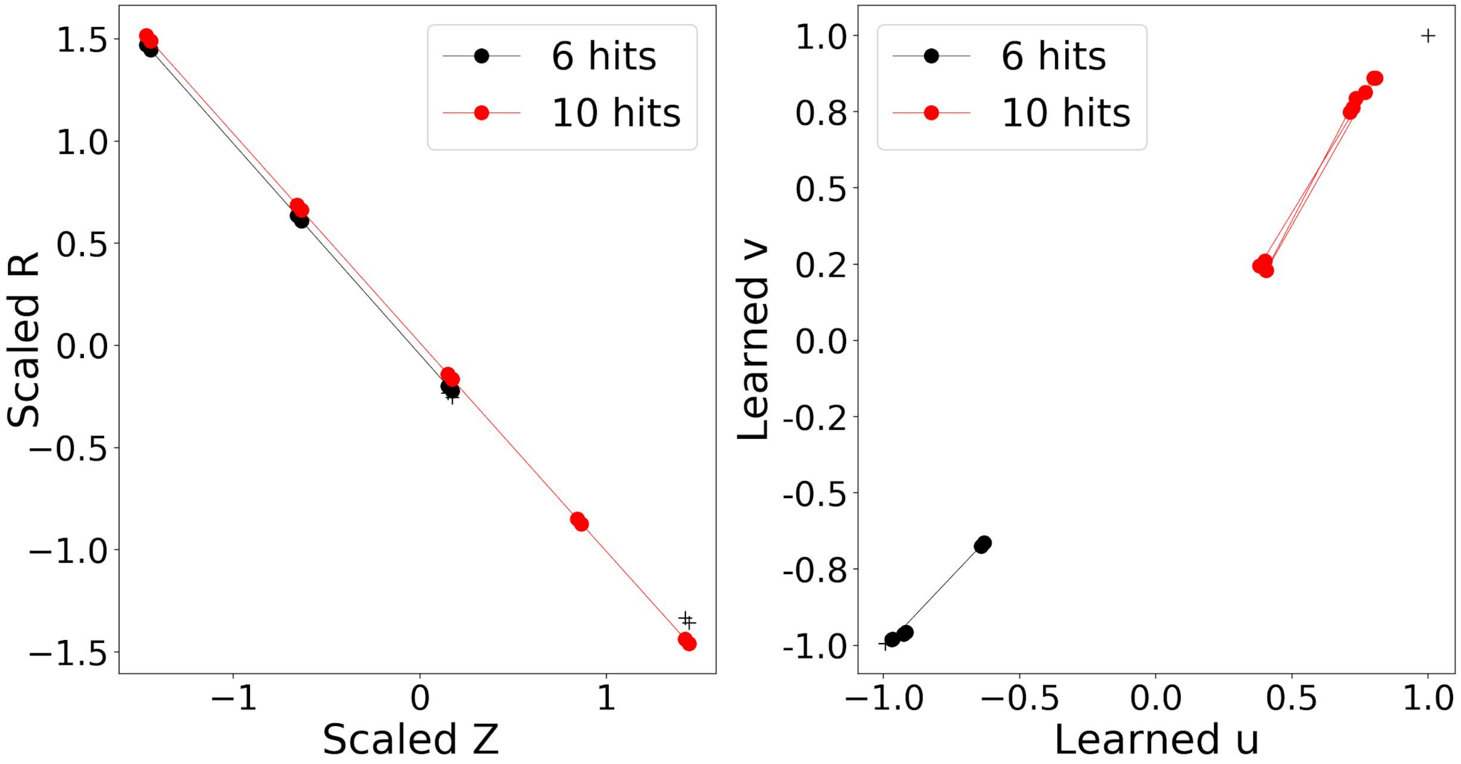
$$\mathcal{L}_C = \sum_{i=0}^{K} S(c_i) \;\; ; \;\; \mathcal{L}_I = \frac{1}{S(\mu_{[1..k]})} \;\; ; \;\; \mathcal{L}_{CL} = \sum \frac{b-a}{\max(a,b)}$$

Compactness **Isolation** **Compact Isolated Clusters**

Loss Evolution Over the Epochs

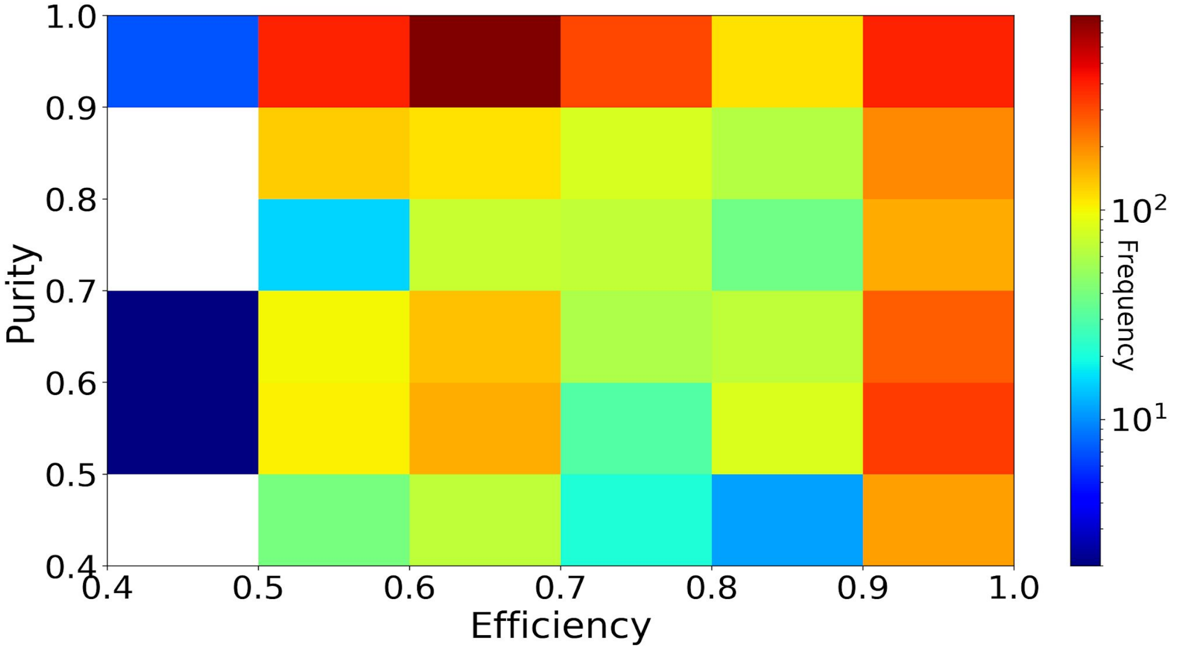


Example on a 20 hits Bucket



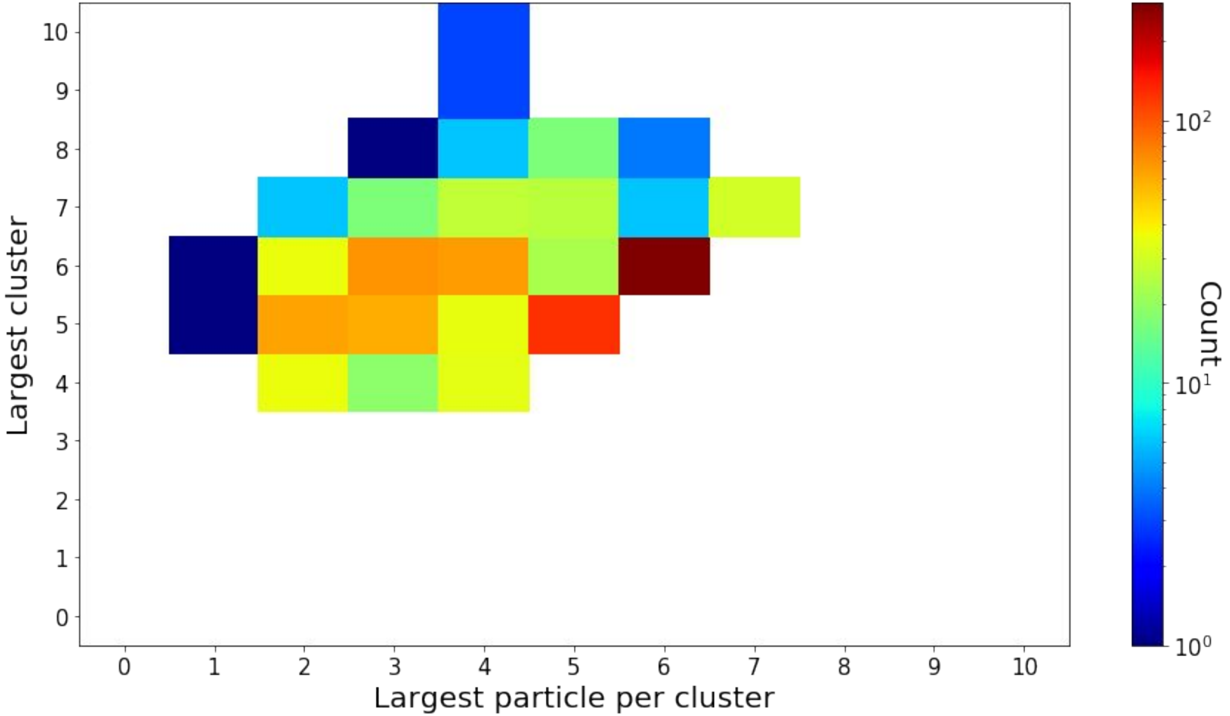
Cluster Efficiency and Purity

- Cluster Efficiency : how many hits of the particle are contained in the cluster.
- Cluster Purity : how many hits of the cluster belong to the same particle.

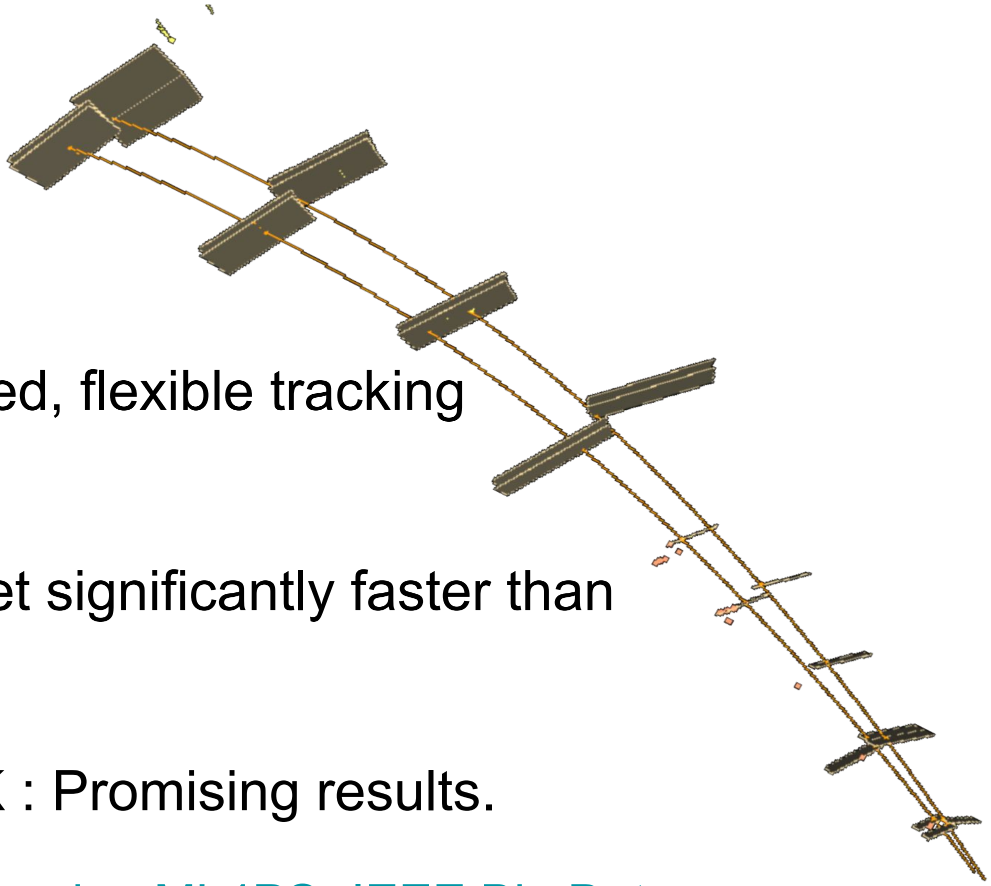


Cluster Size and Particle Size

- Current developments to filter small clusters (particles).
- 6 hits clusters allow good track parameter estimates.



Summary

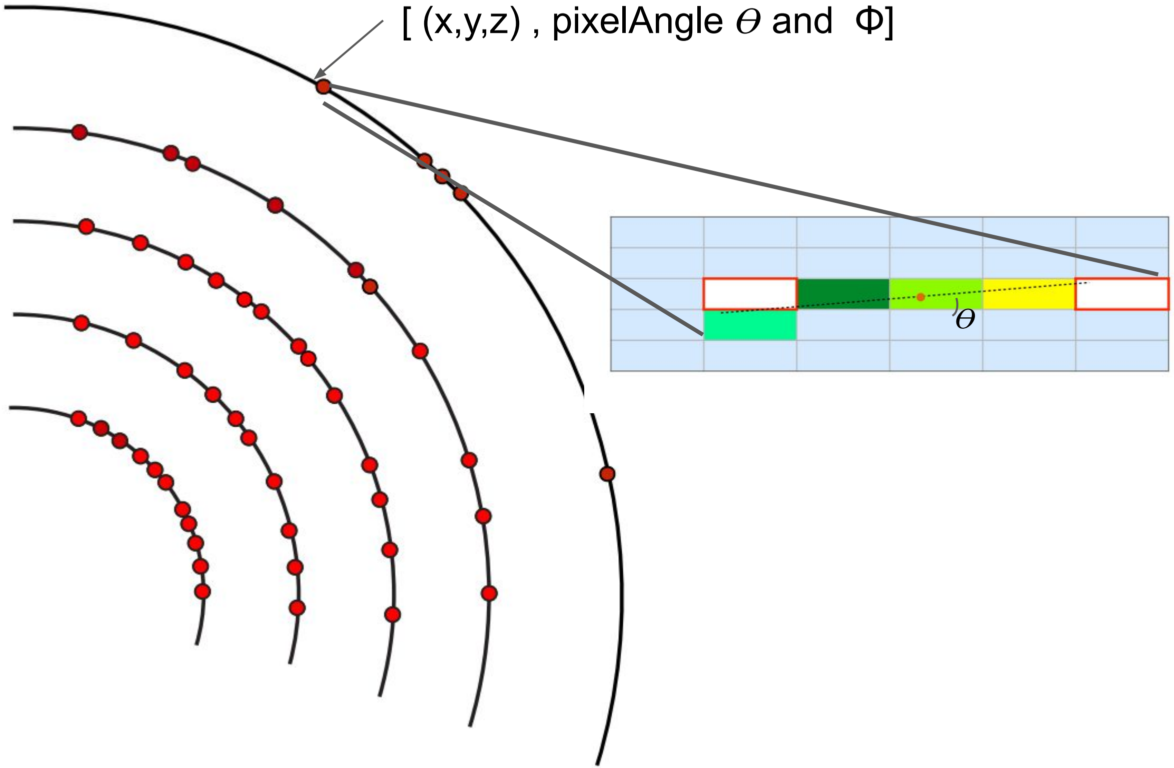


- ANNs : Data-driven, unsupervised, flexible tracking
- Significant speed-up potential.
- Full event mapping with TrackNet significantly faster than combinatorics.
- Current tests of TrackNet on ITK : Promising results.
- Material on ANNs for Tracking Neurips-ML4PS, IEEE Big Data

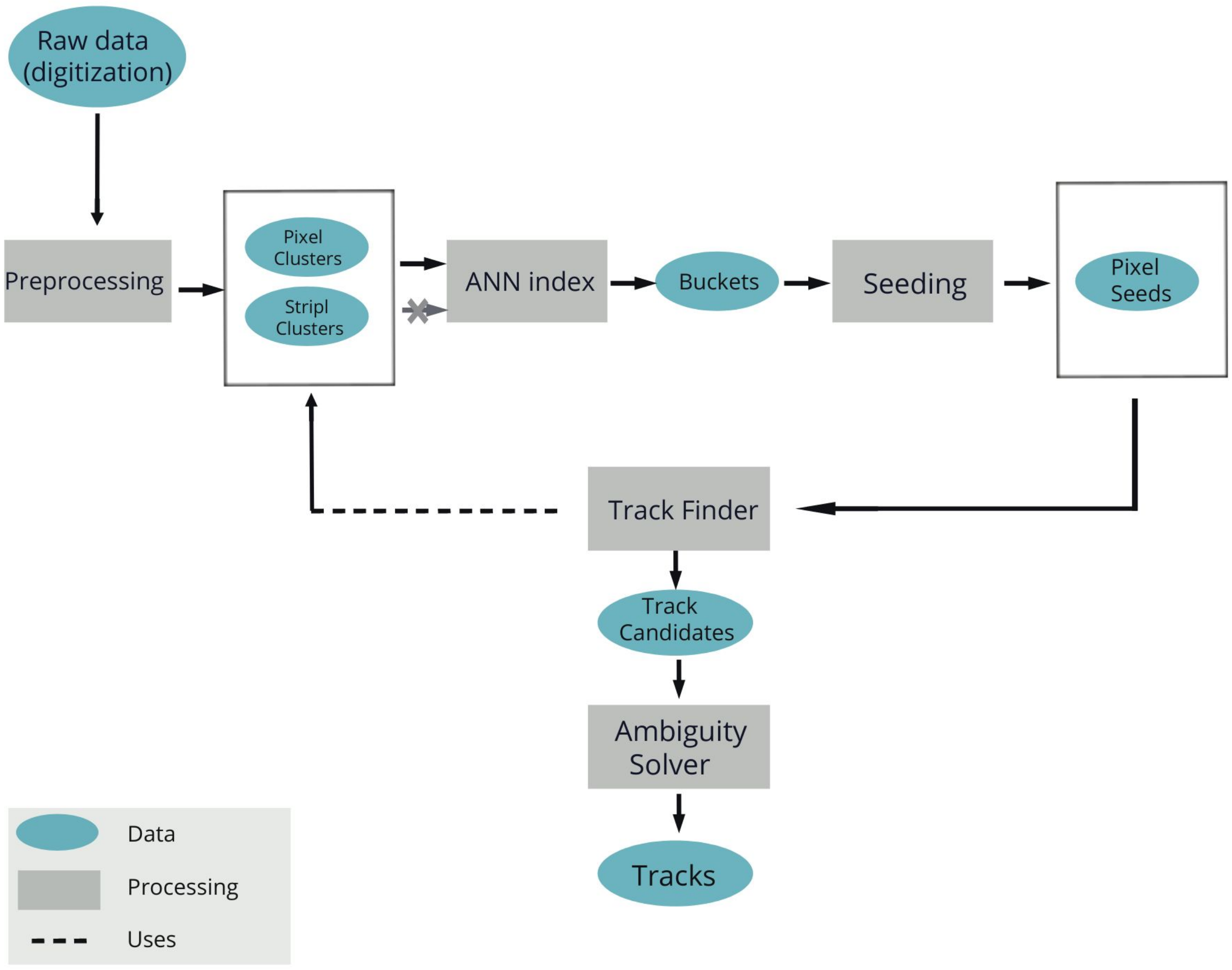
Backup

Dataset : TrackML

- 10k particles, 100k points
- **5** features : global x, y, z and inner angles.
- Hashing library used : Annoy (spotify)

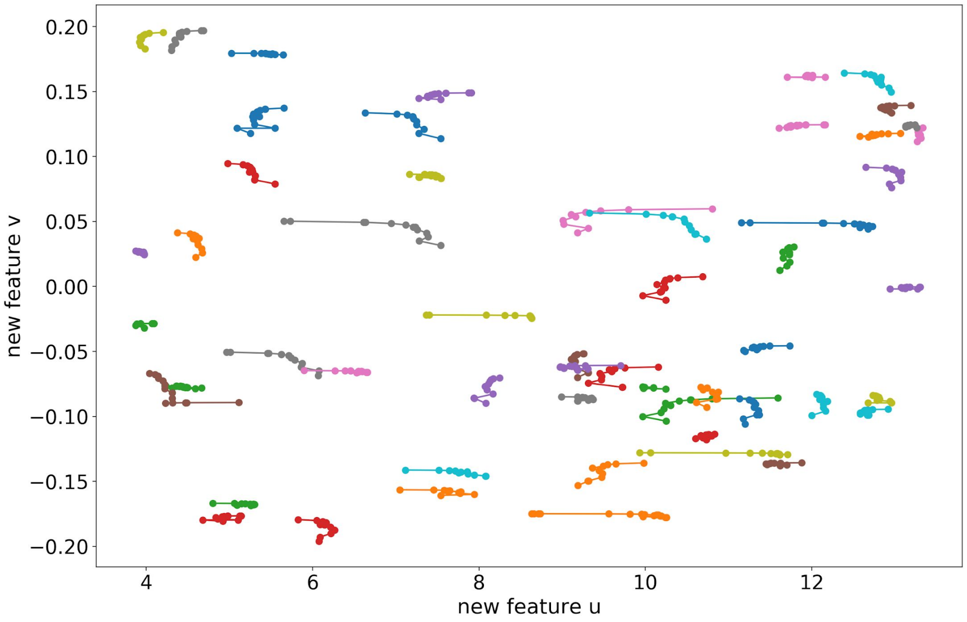


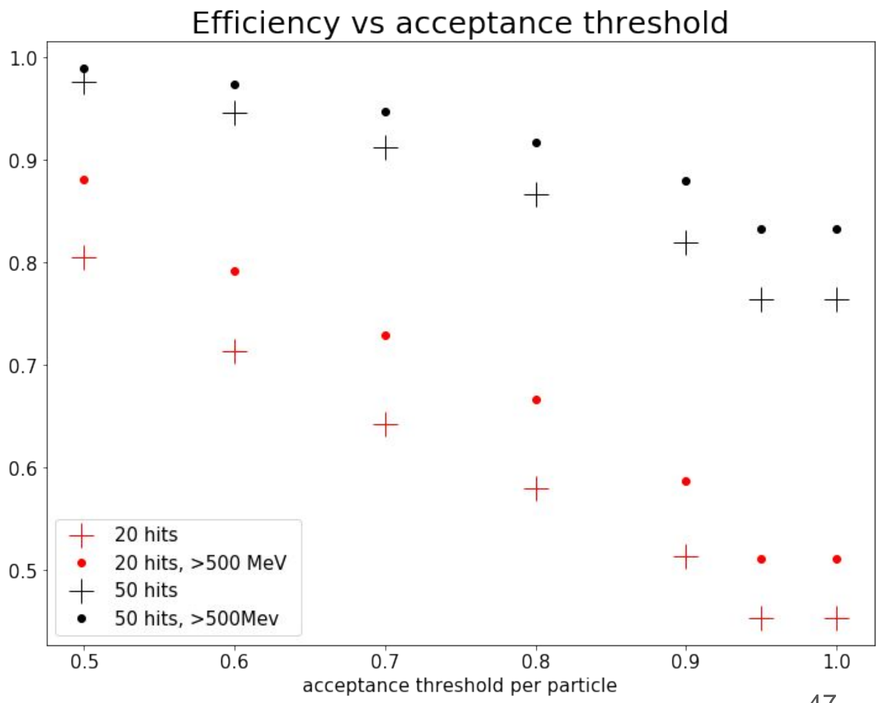
Simulated with ACTS, Ttbar event, mu 200


Raw data (digitization)
Preprocessing
Pixel Clusters
Stripl Clusters
ANN index
Buckets
Seeding
Pixel Seeds
Track Finder
Track Candidates
Ambiguity Solver
Tracks
Data
Processing
Uses

Metric Learning : LFDA on TrackML

Local Fisher Discriminant Analysis





Buckets Filtering- TrackML - 10 events