Status of the SiD-Iowa PFA: New developments and plans

Garabed Halladjian, Remi Zaidan,

Ron Cassell, Mat Charles and Usha Mallik,

ALCPG11 Workshop, University of Oregon, Eugene

Outline

- **Overview of the baseline SiD-Iowa PFA:**
  - Building blocks of the PFA.
  - Diagnostic tools and baseline performance.
- **Latest developments:**
  - Clustering algorithm for clump finding.
  - Likelihood for linking.
- **Ongoing developments and plans:**
  - Shower building.

Data samples:

- 10,000 qqbar events split into 1000/9000 for training/analysis.

Overview
of the baseline SiD-Iowa PFA

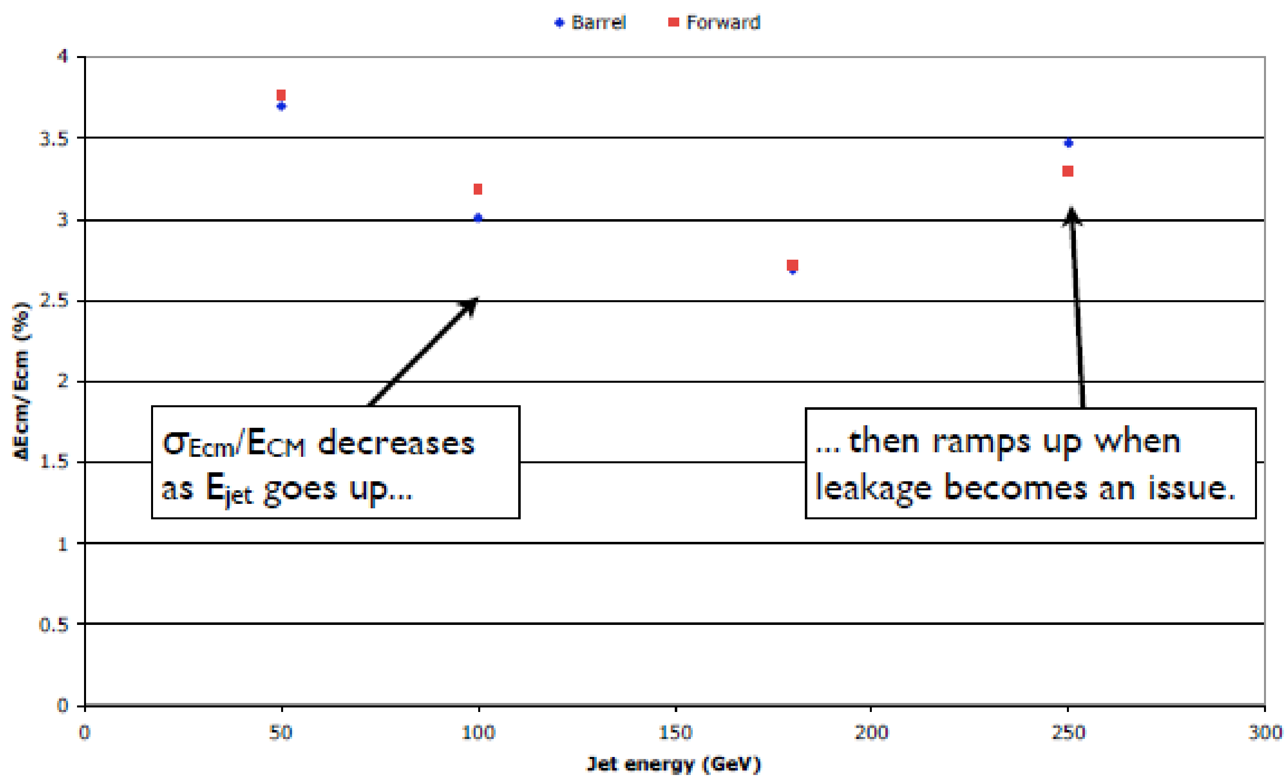
Algorithm flow of the SiD-Iowa PFA

- **Pre PFA:**
  - **MC hits** within 100 ns of the IP are digitized.
  - Hits belonging to **photons, muons, electrons and pre-shower MIPs** are removed from the hit list for clustering algorithm.
- **DTree sub-clustering:**
  - Next run a **Directed Tree Clustering** to form large clusters which are then broken up into sub-clusters classified into types like **MIPs**, **clumps**, **blocks** and **leftovers**.
  - Energy from the **leftovers** is shared among MIPs, clumps and blocks.
- **Shower building:**
  - **Tracks** are matched to **pre-shower MIPs** or tentatively otherwise to anything in the calorimeter to define **seeds**.
  - A **scoring** is used to link together **MIPs**, **clumps** and **blocks**.
  - Build **hadron showers** for one charged track at a time starting with **lowest momentum**.
  - Unused sub-clusters are then used to build **neutral hadron showers.**
- **Reconstructed particles:**
  - **Charged hadrons** are formed using **momentum from the tracks** and a **π^+ mass.**
  - **Neutral hadrons** are formed using **energy from the calorimeter** and a **K_L mass.**

Baseline performance (LOI 2009)

- **Event energy resolution ramps up at high energy:**
  - Partially due to leakage.
  - Algorithm performance also affected by overlapping showers.

The PFA was tuned up to 500 GeV CME for the LOI :
Use 500 GeV as a starting point for higher energy developments.

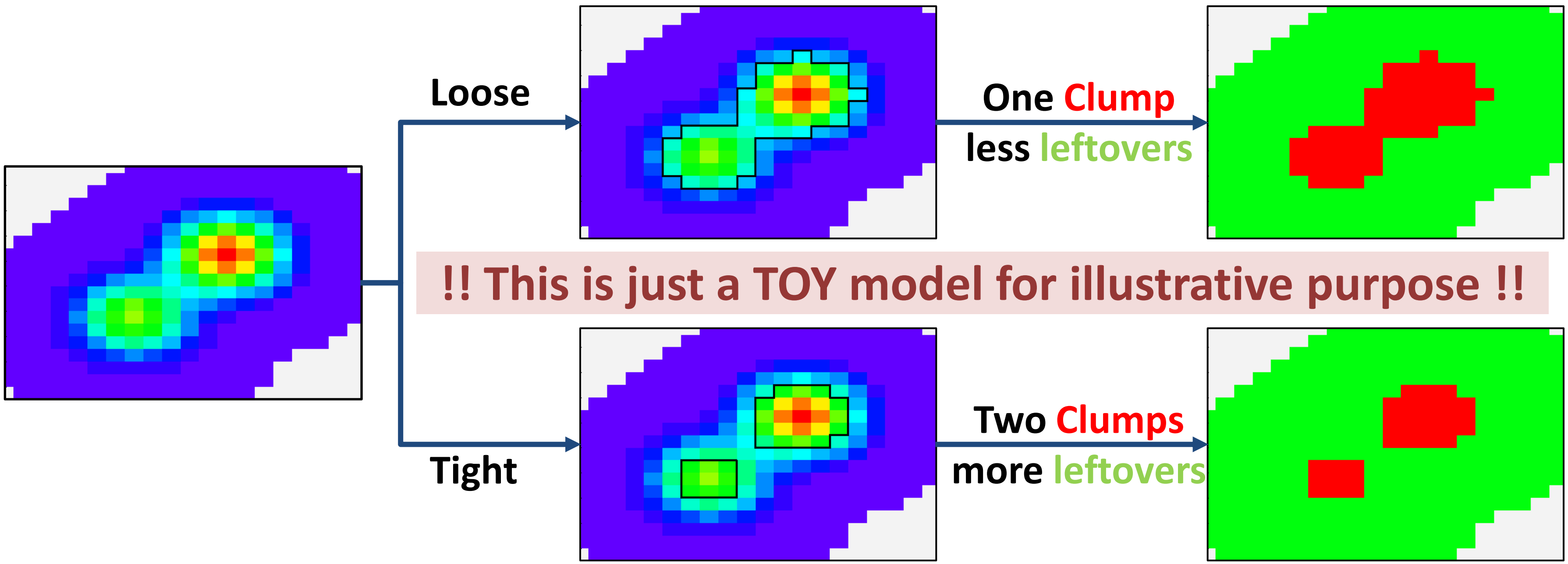


Diagnostic Tools

- Developed a set of tools to study the performance of the PFA at intermediate steps:
  - Looking at event/jet resolution not enough: need to break the PFA into its algorithmic pieces, optimize using a ground-up approach:
    - **Photon finding:**
      - An anti veto is in place which checks “photon-hits” for fakes and treats them as hadrons.
      - This is too sensitive to impurities and lead to real photons being treated as hadrons.
    - **DTree sub-clustering:**
      - Limit on purity (hit purity): 98%
      - MIP purity: 95% (satisfactory)
      - Clump purity: 83% (needs improvement)
    - **Scoring for linking:**
      - based on ad-hoc penalties and a likelihood not optimized for high energies.
    - **Shower building:**
      - Purity intrinsically limited by the clump purity.
      - Depends on track ordering: starts with lower momentum to limit undesired effects.
      - Handling of overlaps relies on imposing an energy/momentum balance.

Clump finding

How can we optimize the baseline clump finder?!



Optimization possibilities are limited:

- **We lose information in both loose and tight scenarios!**
- **The density information is only used for hit pre-selection.**

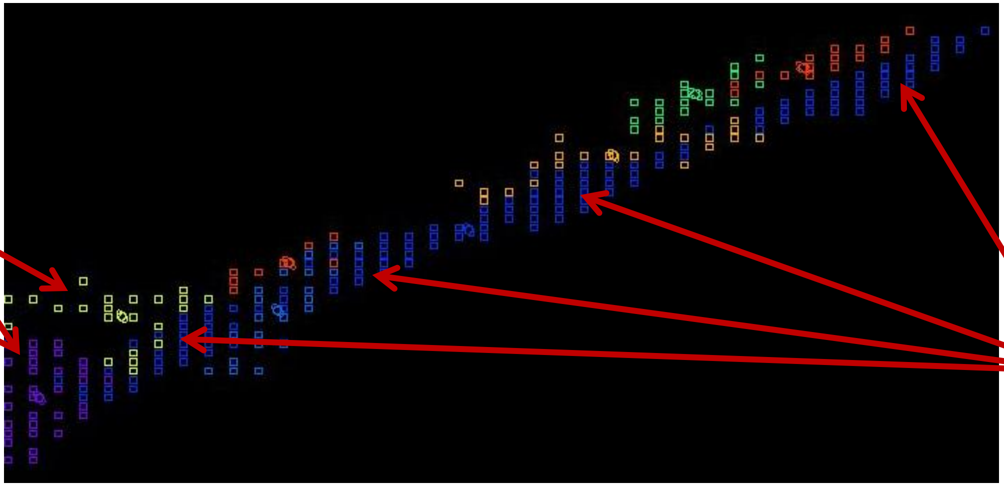
A simple Nearest Neighbour algorithm may not be suitable for overlap scenarios.

Alternative Clustering algorithm

- ***k*-means clustering algorithm:**
  - Define a **distance**: a metric that tells how likely a hit belongs to a cluster.
  - Find ***k* seeds**: initial set of clusters. The number of seeds determine the number of clumps.
  - **Loop on hits** and assign each hit to the “closest” seed.
- **Motivations:**
  - It’s simple.
  - It leaves the flexibility for physics input:
    - Trough the definition of the distance and the seeds.

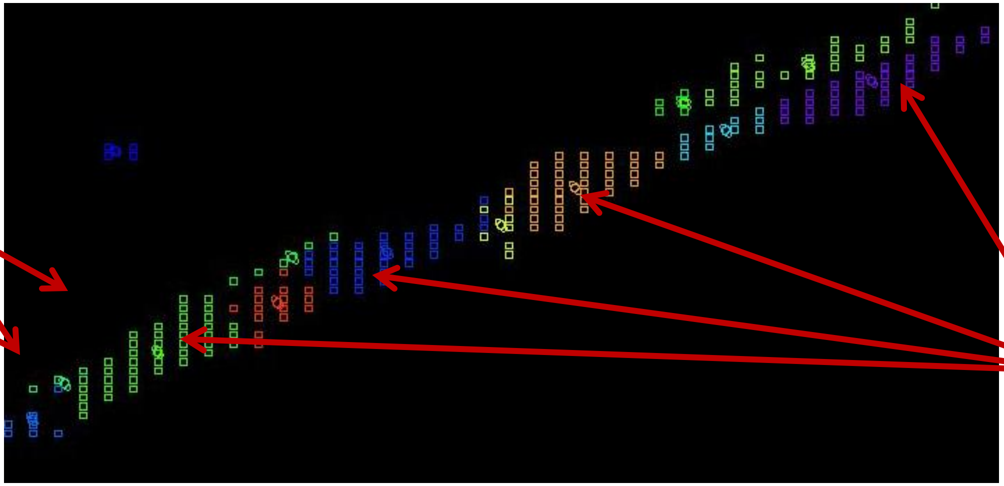
Performance in event displays

Baseline:
more energy was being reconstructed as small clumps with no shape



Baseline:
single clump

k-means:
losing some of the energy in low density regions



k-means:
clump broken

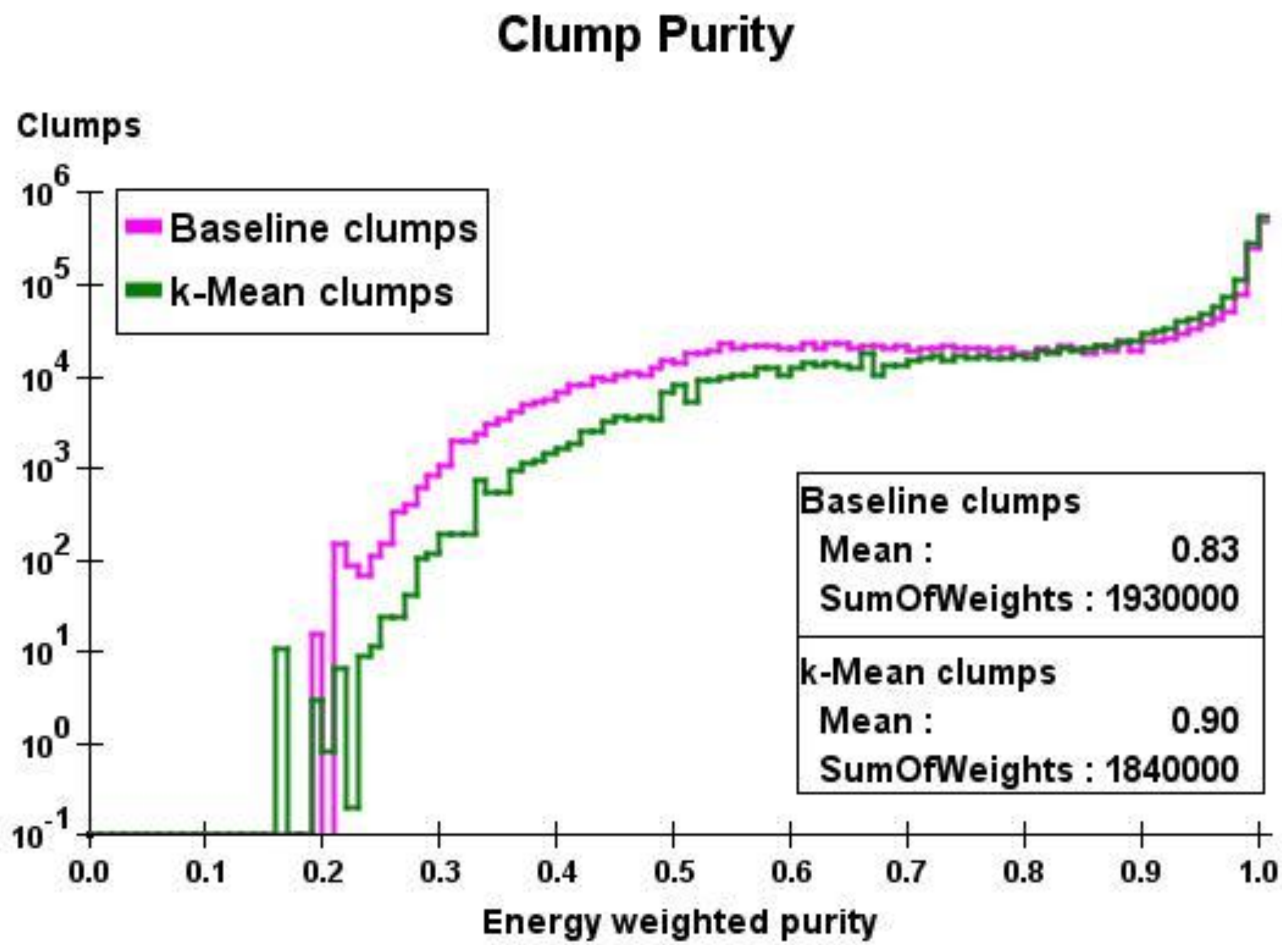
Performance in terms of purity

Improved purity

83% → 90%

Slightly less energy goes into clumps:

47.5% → 45.2% of the total event energy



Performance: ECal vs. HCal

Improvement is better in the HCal

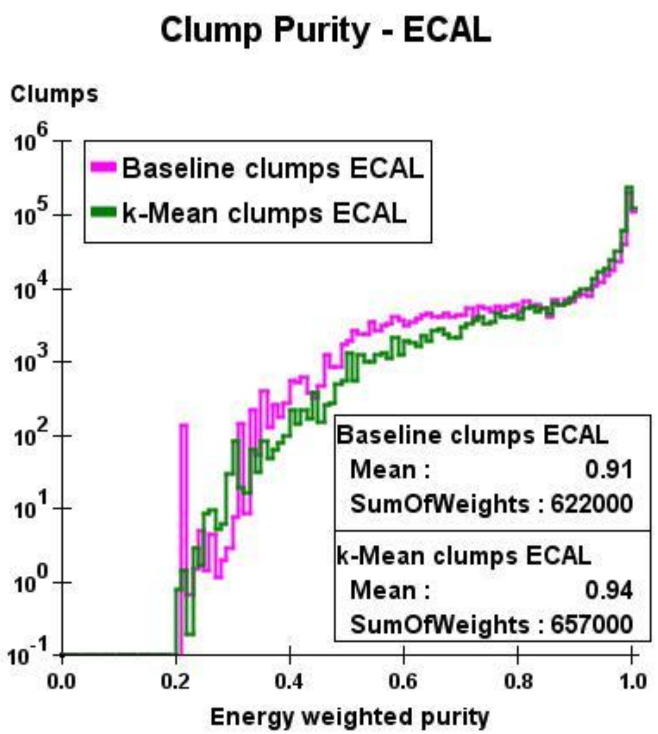
ECal: 91% → 94%

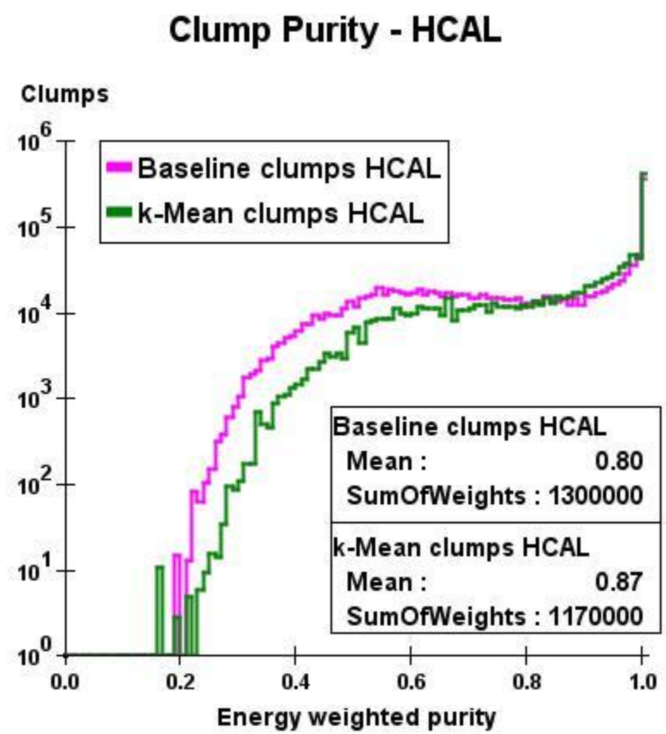
HCal: 80% → 87%

The energy loss is in the HCal

ECal: 15.3% →16.2%

HCal: 32.2% → 29%





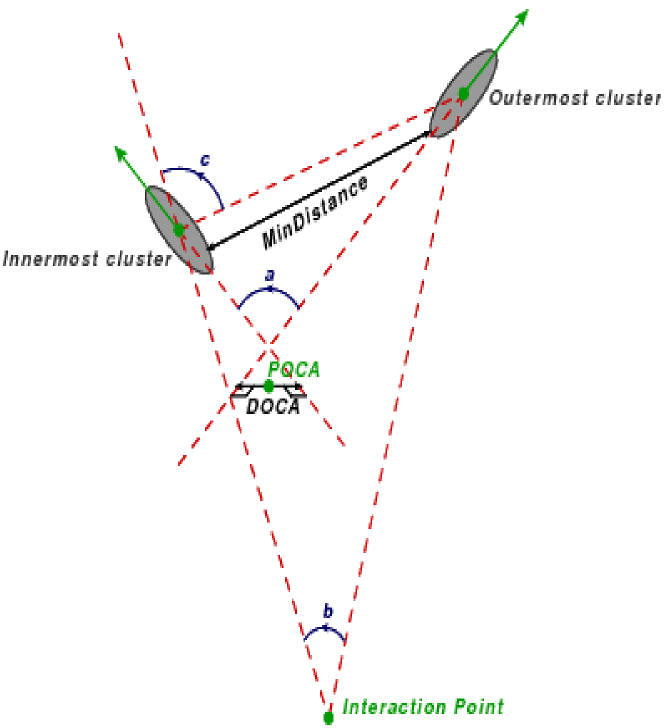
Likelihood for linking

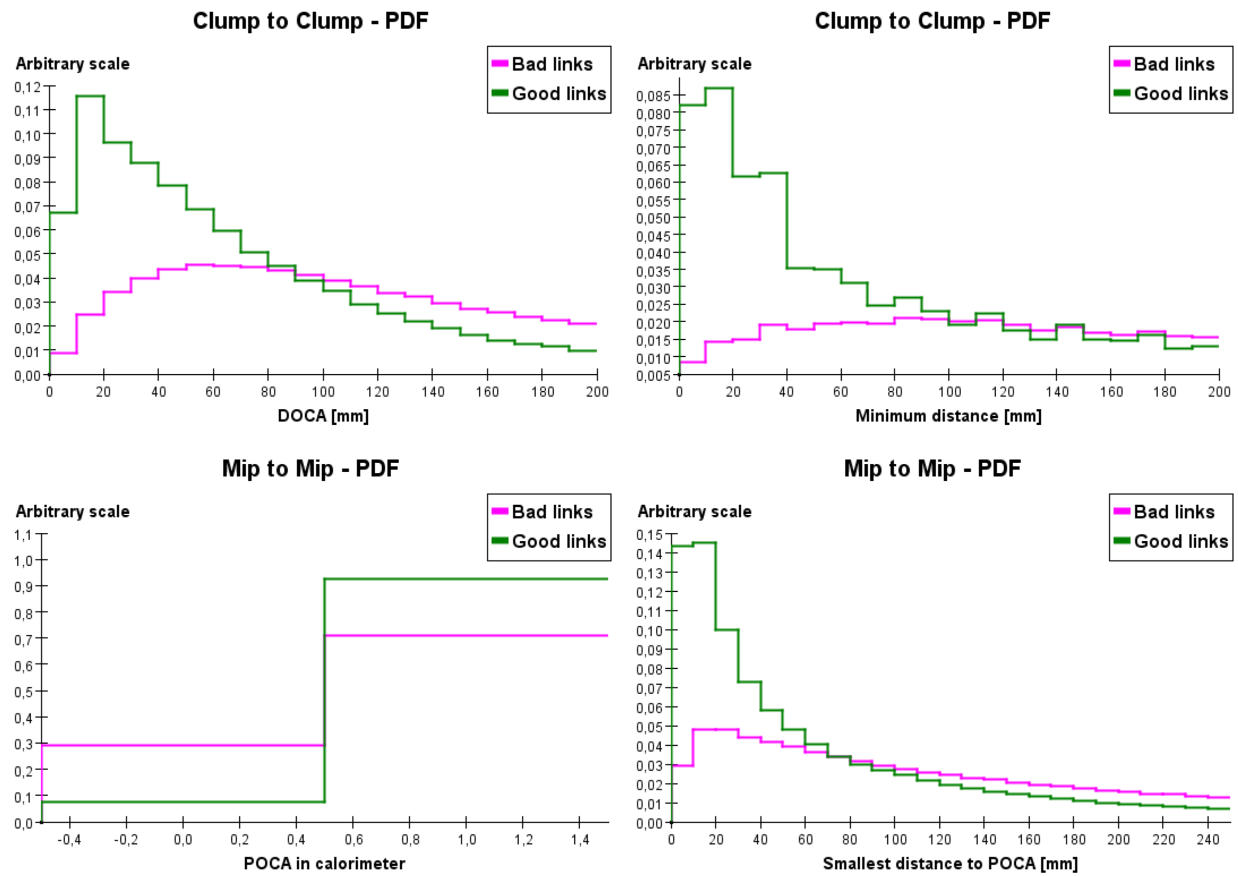
Likelihood for linking

- **Baseline likelihood:**
  - Training is done using a different clustering algorithm (not with DTree).
  - Training is done on events with lower jet energy spectrum (ZZ($qq\nu\nu$) at 500 GeV CME).
  - Uses a limited set of geometrical information.
- **New developments:**
  - Developed a training tool to use the correct clustering algorithm and jet energy.
  - Use new topological variables.

Variables for likelihood

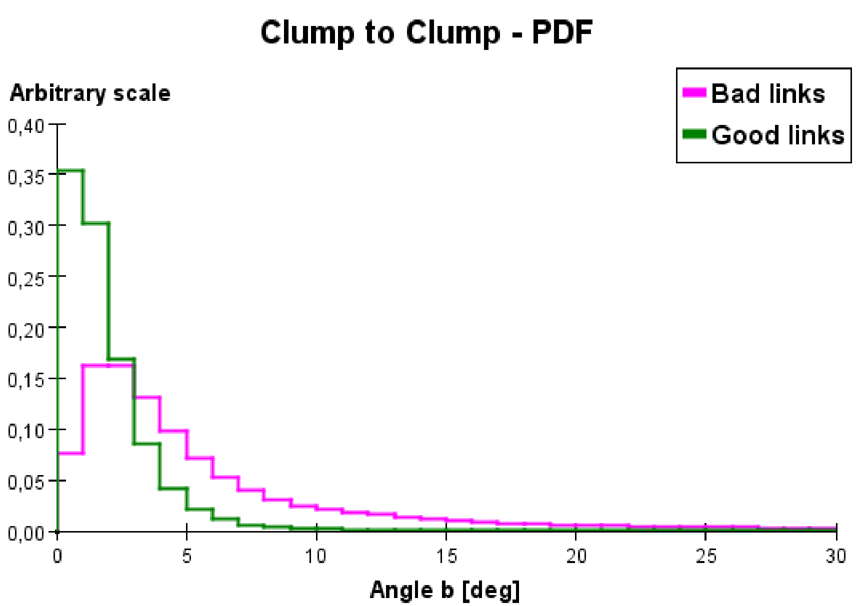
- Baseline variables:

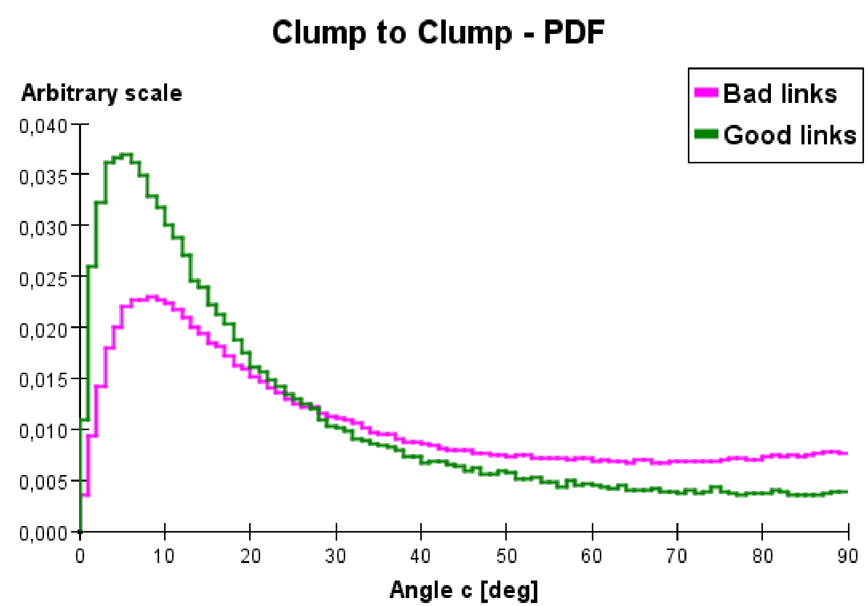




Extra variables for likelihood

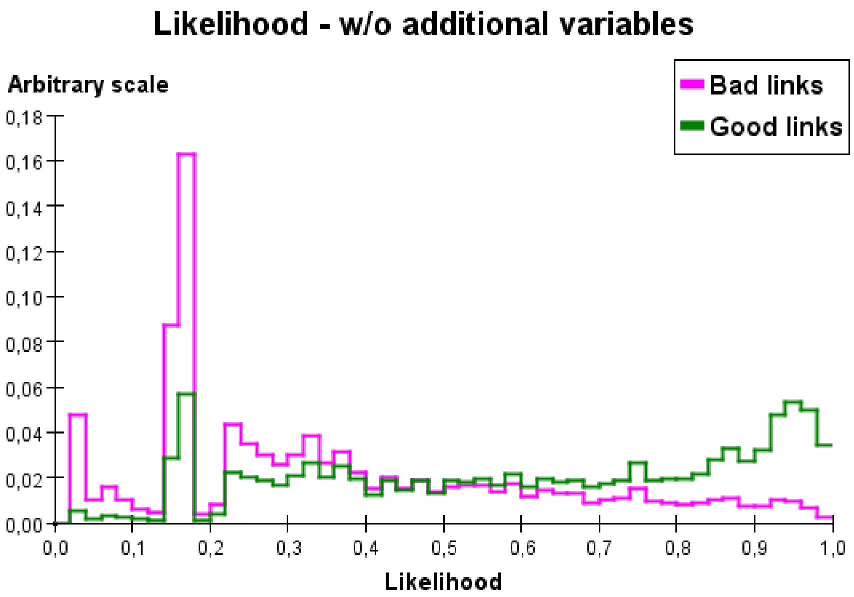
- Additional variables were tested and included into the likelihood:
  – Angular separation (angle b).
  – “Kink” angle (angle c).

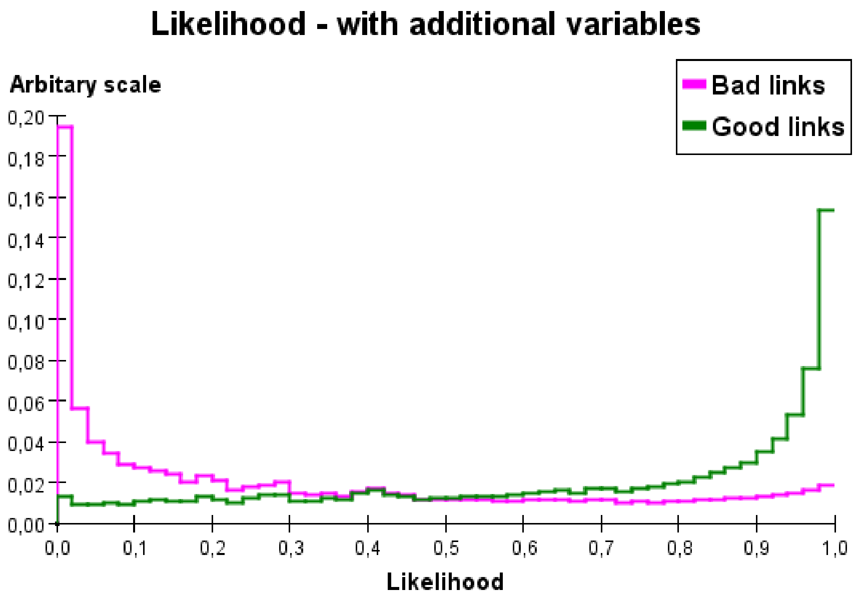




Performance

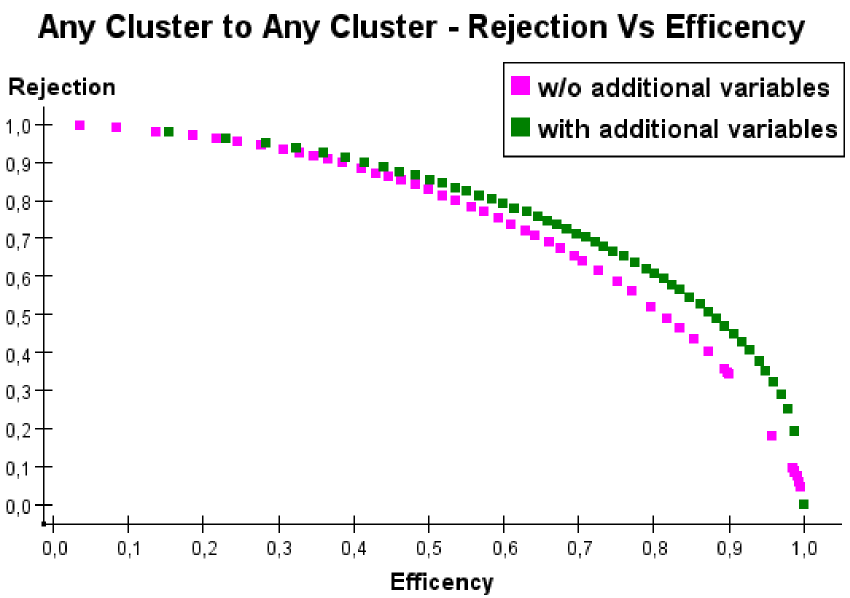
- Improvement is visible in the likelihood distributions; clearer discrimination after adding the new variables.

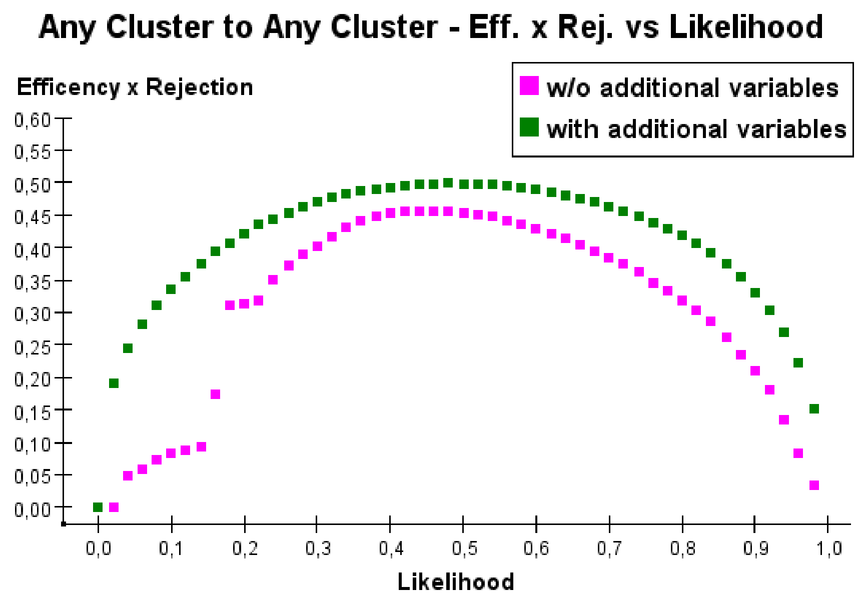




Performance

- Use efficiency and rejection factor to measure likelihood discrimination:
  - Efficiency: fraction of good links above a given cut.
  - Rejection: fraction of bad links below a given cut.





Correlations

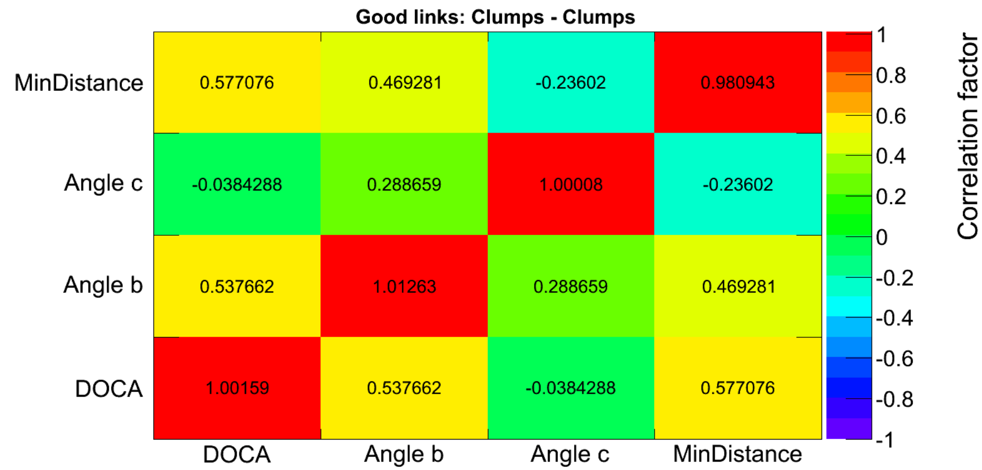
- **Correlation factor:**
  - **Defined as:**

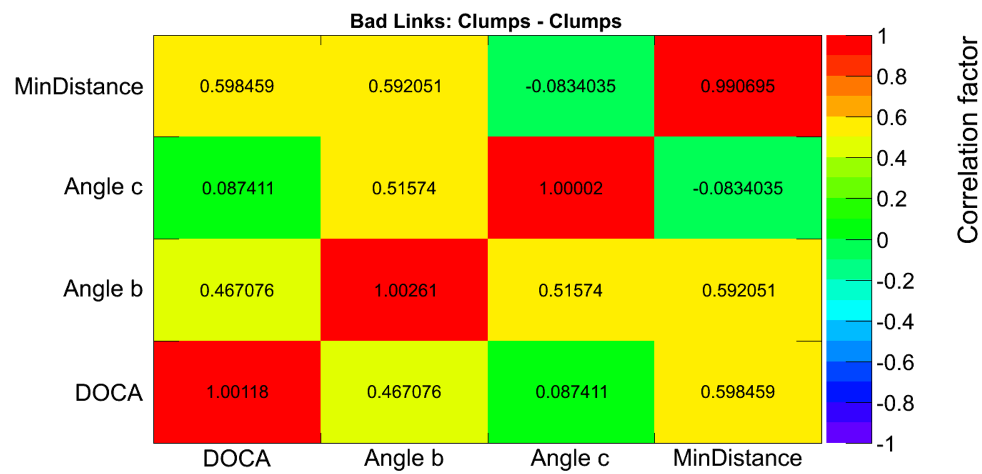
$$\rho_{x,y} = \frac{\left\langle \left(x-\bar{x}\right)\times\left(y-\bar{y}\right)\right\rangle}{\sqrt{\left\langle \left(x-\bar{x}\right)^2\right\rangle \times \left\langle \left(y-\bar{y}\right)^2\right\rangle}} = \frac{\mathbf{cov}_{x,y}}{\sigma_x \times \sigma_y}$$

  - **Measures the degree of correlation:**

$$-1 \le \rho_{x,y} \le 1$$

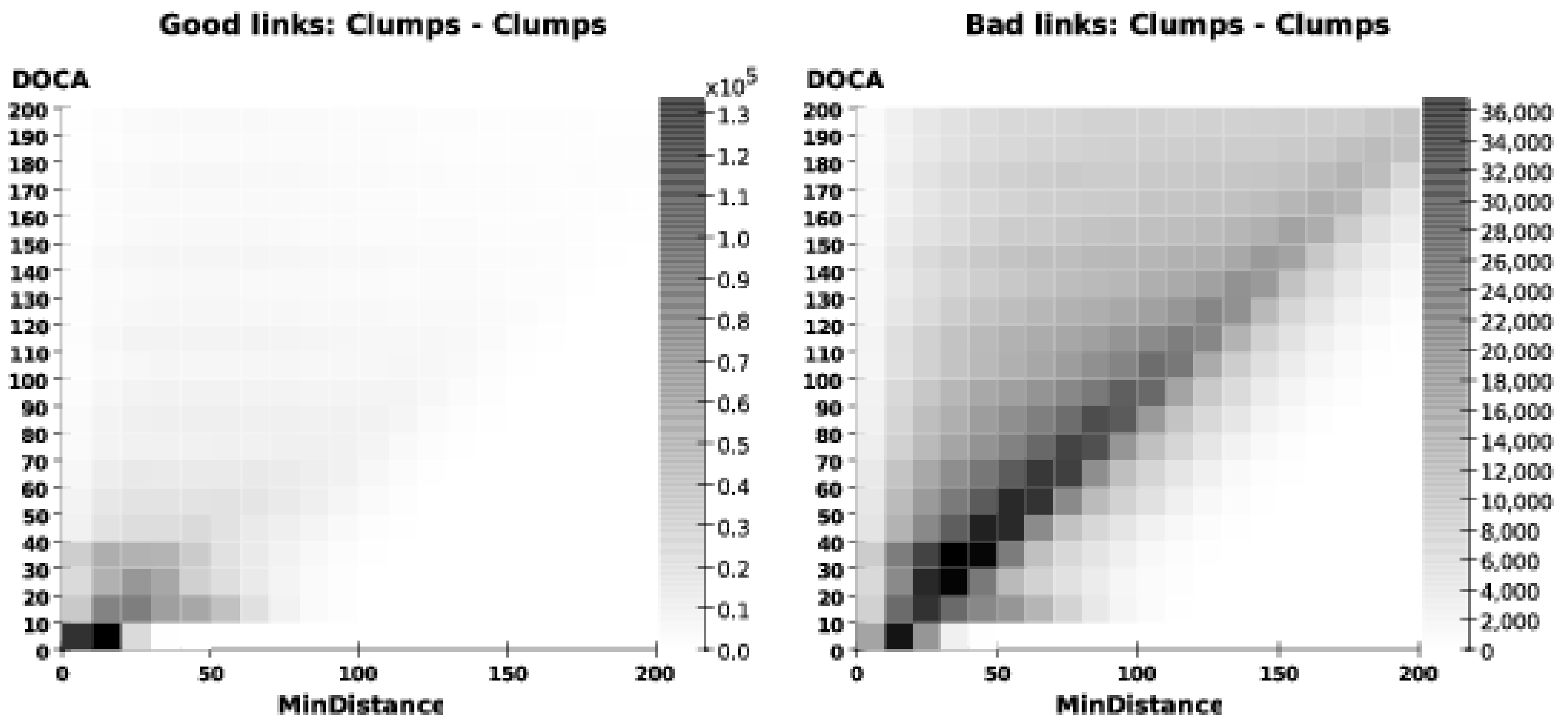
- **Likelihood function assumes independent variables:**
  - **Correlations between variables may cause peaks in the likelihood distribution for background in the signal region and vice versa.**





Correlations

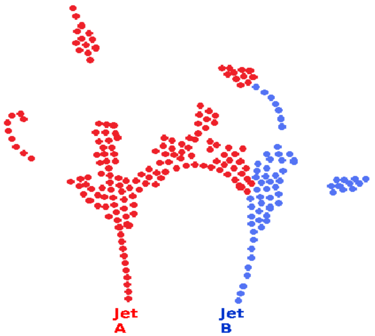
- Using a 2-dimentional distributions for correlated variables should enhance likelihood performance with respect to using independent 1-dimentional distributions.

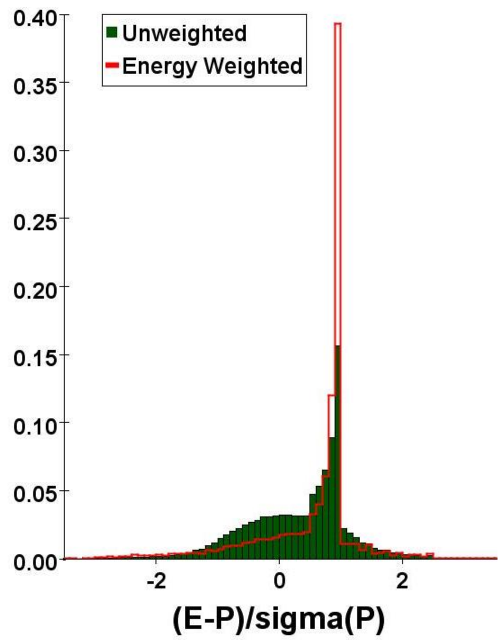


Ongoing developments: shower building

Shower building: baseline

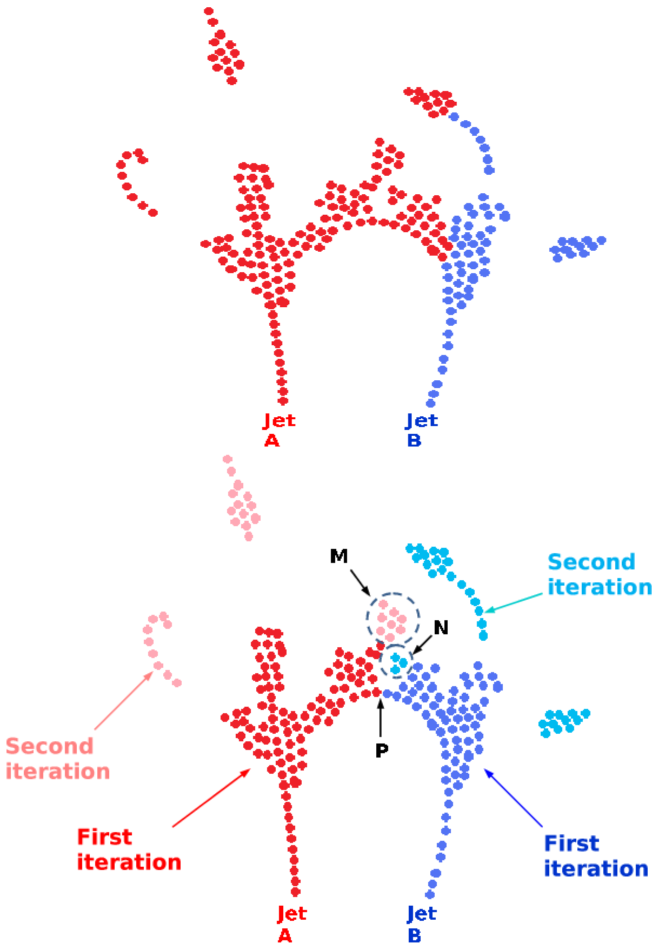
- Starts with the lower momentum tracks:
  - The motivation is that a high momentum track can easily “eat” the shower of a nearby low momentum track.
- Implements a E-P balance criteria to determine when a shower should stop expanding:
  - Needed as a sanity check.
  - Important in dense environments at higher energies.





Shower building: ongoing developments

- Implementing a 2-iterations strategy:
  - First iteration:
    - Build a robust shower core with high purity:
      - Track order independent: leave ambiguities to be resolved in a second iteration.
      - Strong links: leave secondary neutrals to the second iteration.
      - Outgoing shower building: leave back-scattering cases to the second iteration.
  - Second iteration:
    - Has the task of getting the maximum efficiency by solving “difficult” cases:
      - the output of the first iteration can be a very useful information.



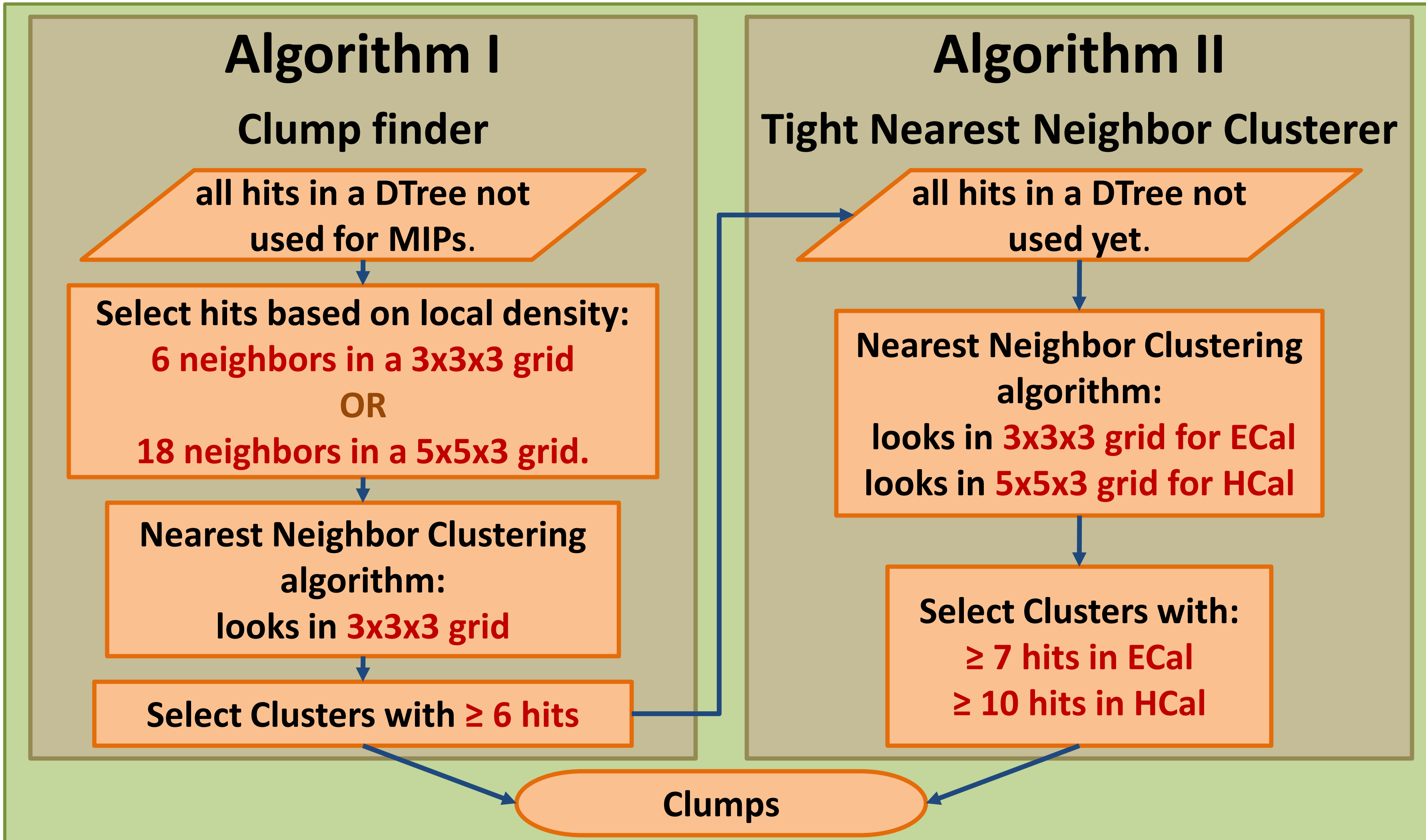
Conclusion

- Detailed diagnostics have been performed on intermediate algorithmic pieces:
  - several areas where improvement can be made were identified.
- Defined an optimization strategy that follows a ground-up approach:
  - Observed significant improvements in:
    - Clump finding.
    - Likelihood for linking.
  - Currently working on improving the shower building:
    - implementing a two steps algorithm: first build a high purity shower core, then come back and deal with "special cases".

Backup

Baseline clump finding algorithms

The University of Iowa



Some details about k-means clump finder

- **Hit pre-selection**:
  - Local hit density cut:
    - **3 neighbors in a 3x3x3 grid** OR **9 neighbors in a 5x5x3 grid**.
- **Local analogue density definition**:
  - **Average hit energy per cell in a 3x3x3 grid.**
  - **In HCal this is identical to digital density.**
- **Seed finding**:
  - Search for **local maxima** of analogue density:
    - A hit where all neighbours have lower or equal density.
  - Select the **hits on local maxima** and all their **neighbours in a 3x3x3 grid**.
  - Run a **Nearest neighbour** clustering algorithm.
- **Distance**:
  - **Geometrical distance** between a hit and the **nearest** hit in the seed.

Phase space for linking

- Decided to cut on angular separation:
  - reduce the combinatorics.
  - eliminate physics dependence on likelihood training