Spectral clustering for jet physics

Speaker: Giorgio Cerro

Affiliation: University of Southampton, UK

Supervisor: Stefano Moretti

Journal: JHEP, 135 (2022)

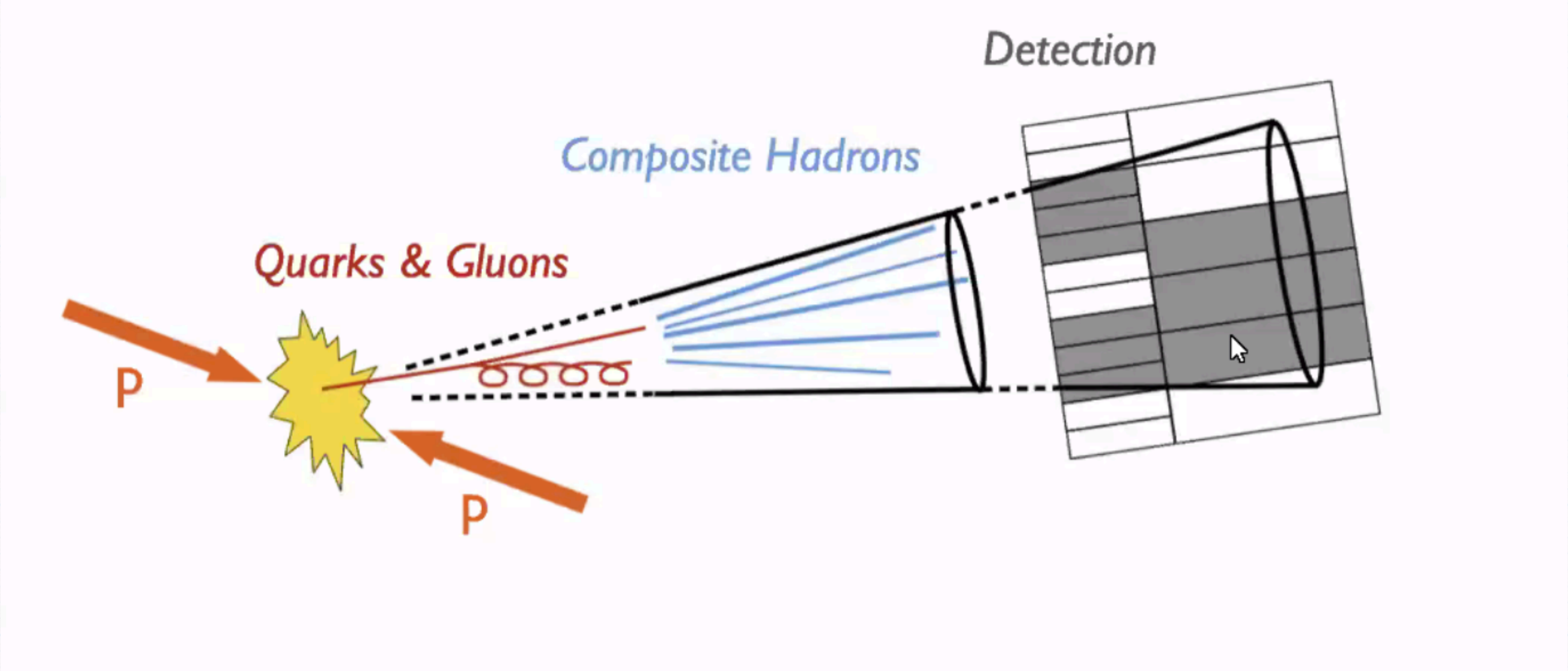
In collaboration with: H. Day-Hall, S. Dasmahapatra, C. H. Shepherd-Themistocleous

Content

- Jet Clustering
- Spectral Clustering
- Results
- Strengths & Limits
- Summary

Jet clustering

In a laboratory



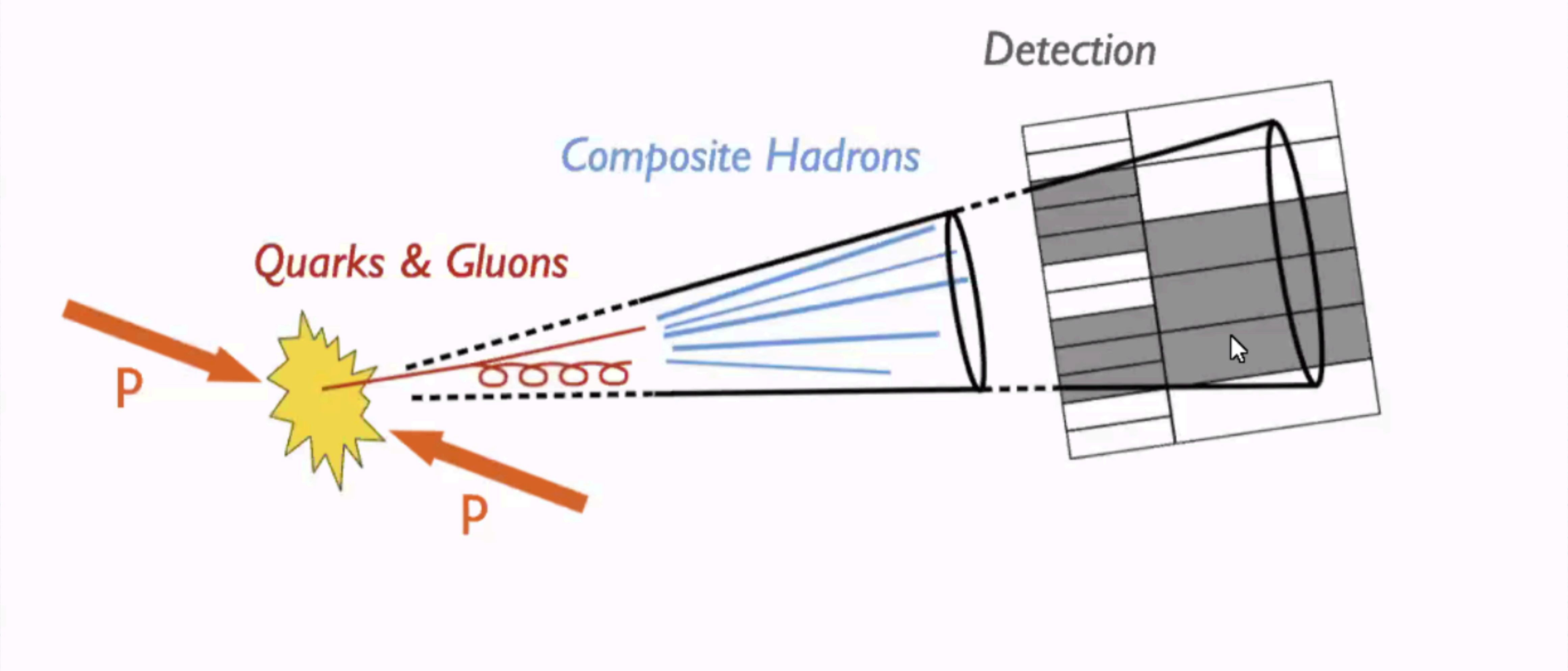
https://cms.cern/news/jets-cms-and-determination-their-energy-scale

Three events:

- $gg \to H_{125} \to h_{40}h_{40} \to b\bar{b}b\bar{b}$
- $gg \to H_{500} \to h_{125}h_{125} \to b\bar{b}b\bar{b}$
- $gg,qq \to t\bar{t} \to b\bar{b}W^{+}W^{-} \to b\bar{b}jjlv$

Jet clustering

In a laboratory



https://cms.cern/news/jets-cms-and-determination-their-energy-scale

Three events:

- $gg > H_{125} > h_{40}h_{40} > b\bar{b}b\bar{b}$
- $gg > H_{500} > h_{125}h_{125} > b\bar{b}b\bar{b}$
- $gg,qq > t\bar{t} > b\bar{b}W^{+}W^{-} > b\bar{b}jjl\nu$

Spectral clustering

- Compute Laplacian matrix
- Project particles into the eigenspace
- Merge the closest two
- Repeat until condition met

Degree matrix

$$L(t) = Z(t)^{-\frac{1}{2}}(B(t) - A(t))Z(t)^{-\frac{1}{2}}.$$

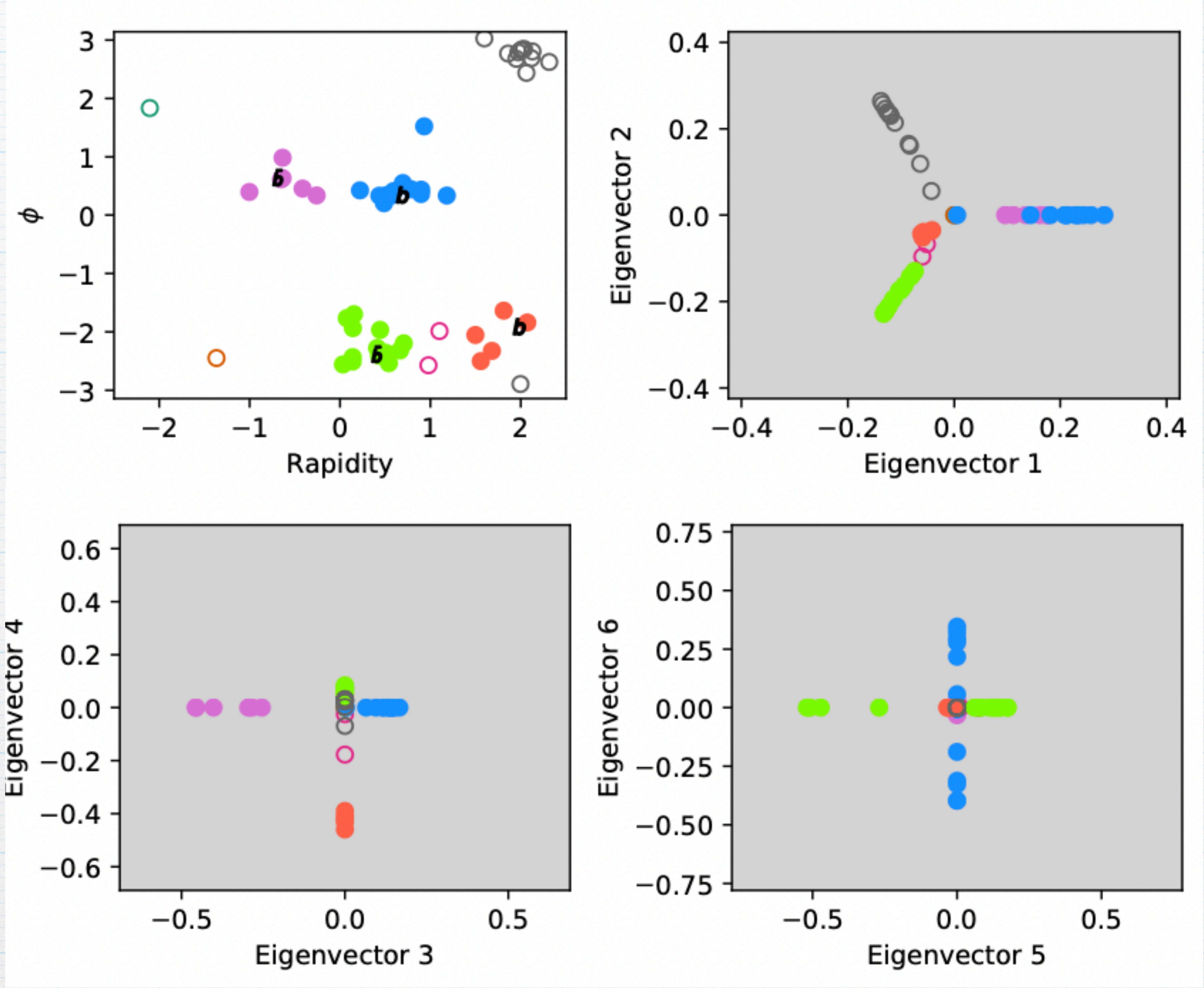
Weight matrix

Affinity matrix

$$a(t)_{i,j} = \exp(-d(t)^{\alpha}_{i,j}/\sigma_v)$$

Spectral clustering

- Compute Laplacian matrix
- Project particles into the eigenspace
- Merge the closest two
- Repeat until condition met



Spectral clustering

- Compute Laplacian matrix
- Project particles into the eigenspace
- Merge the closest two
- Repeat until condition met

Distance in the new space:

$$d'(t)_{i,j} = s(t)_{i,j} \arccos\left(\frac{m(t)_i \cdot m(t)_j}{\|m(t)_i\|\|m(t)_j\|}\right)$$

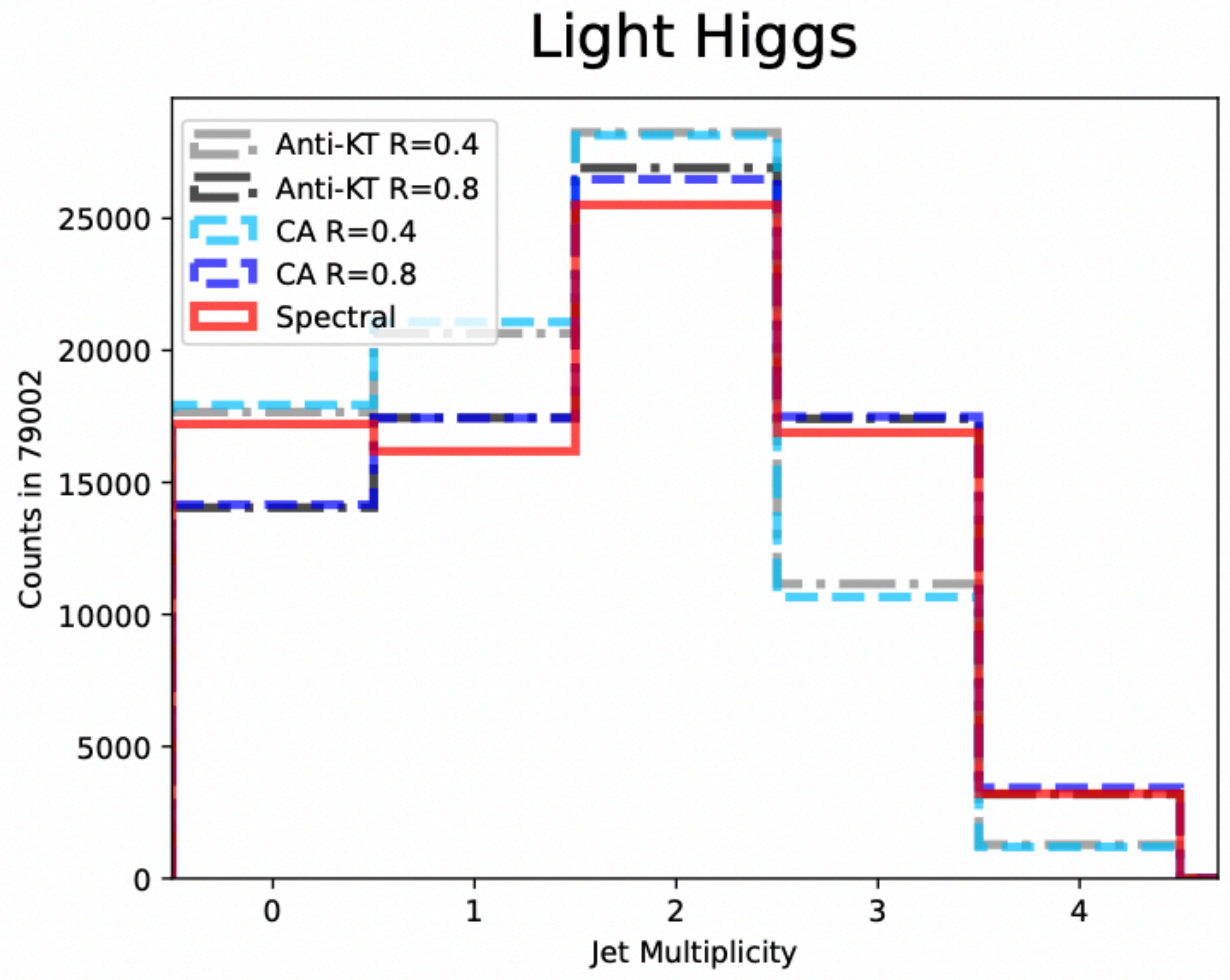
Spectral clustering

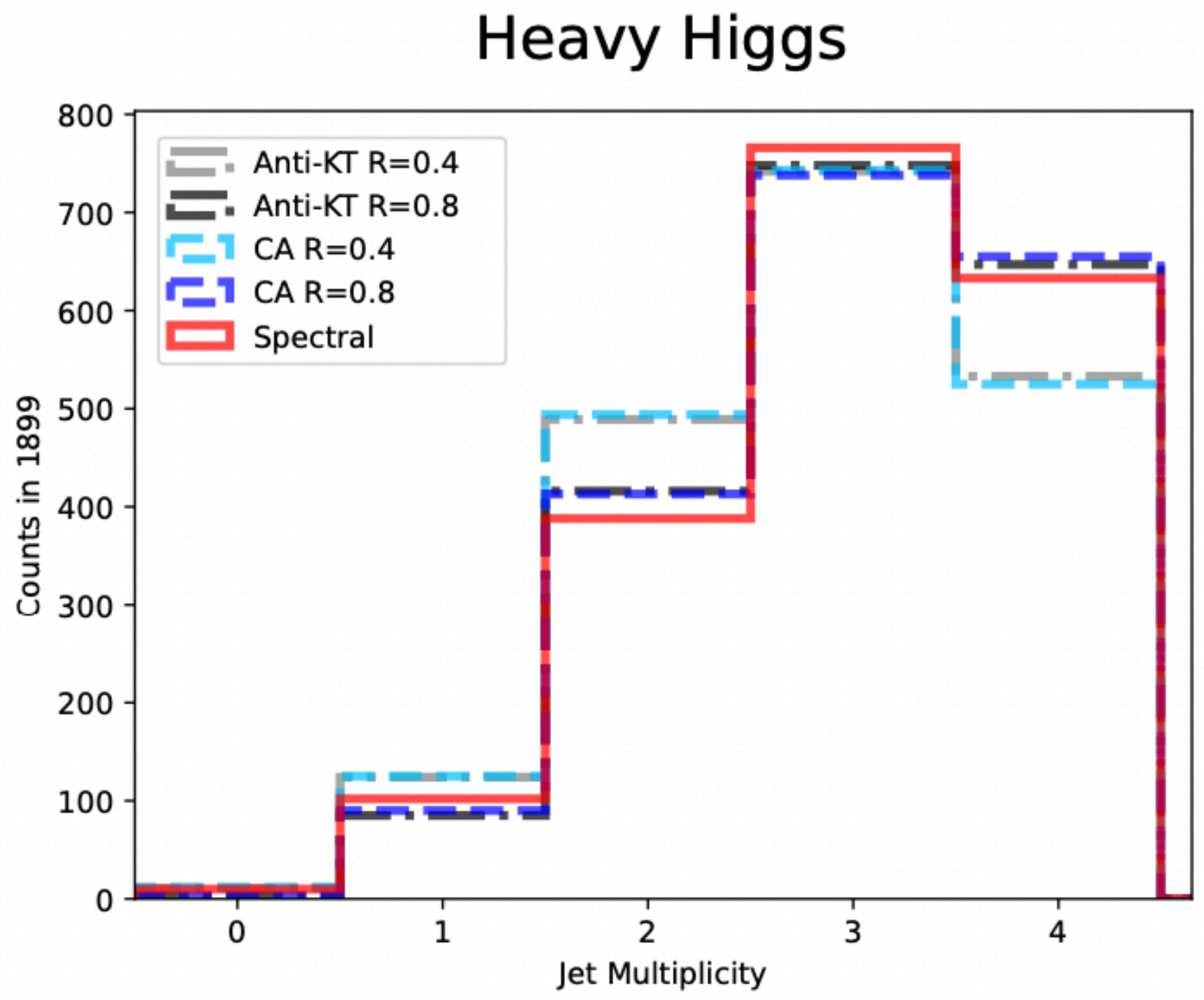
- Compute Laplacian matrix
- Project particles into the eigenspace
- Merge the closest two
- **Repeat until condition met**

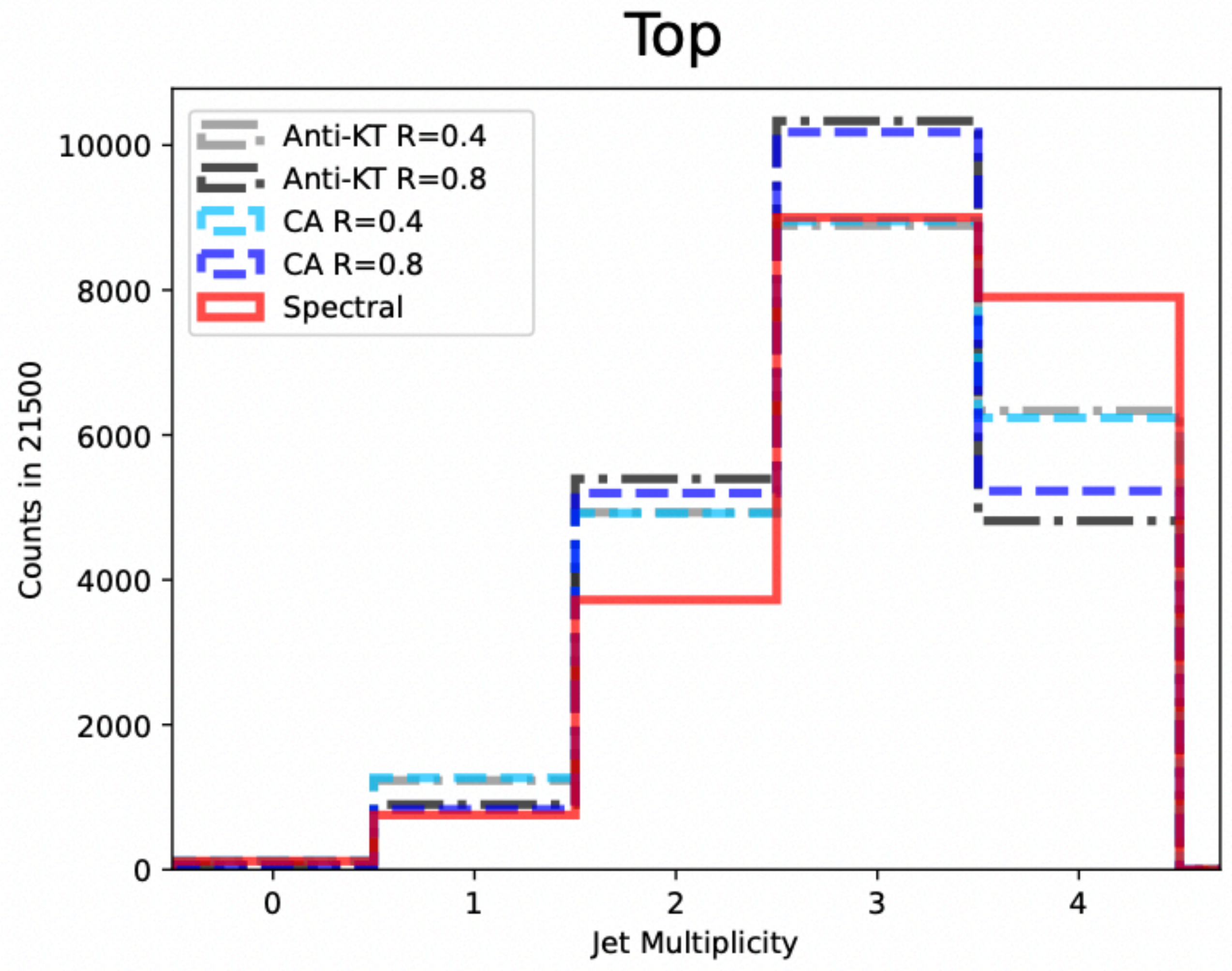
The weight is updated and the laplacian is computed again until

$$\sum_{i \neq j} \sqrt{d'(t)_{i,j}} < R$$

Results

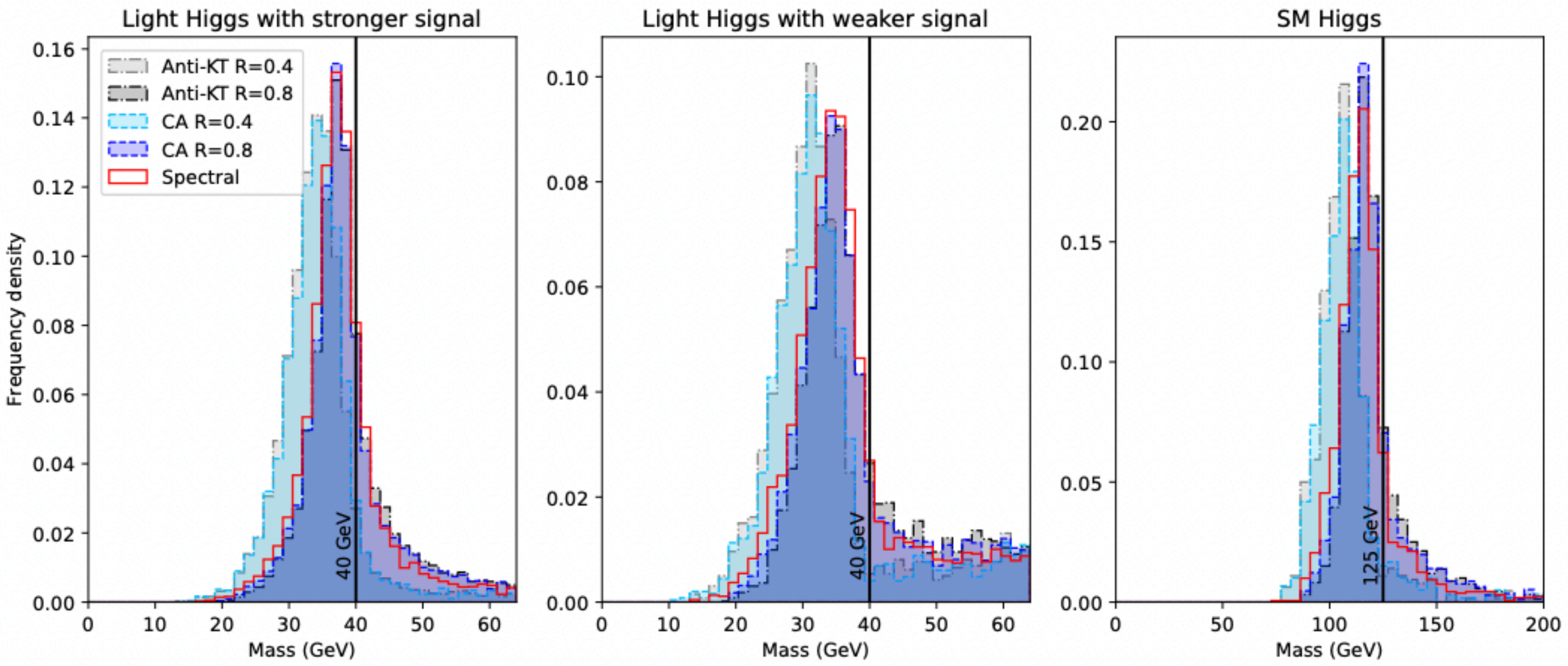






Jet multiplicities for the anti-k_T and CA (for two R_{k_T} choices) and spectral clustering algorithms on the Light Higgs, Heavy Higgs and Top MC samples. For all such datasets, the hard scattering produces 4 partons in the final state, so maximising a multiplicity of 4 jets indicates good performance.

Results



Three mass selections are plotted for the Light Higgs dataset. From left to right we show: the invariant mass of the 4b-jet system, of the 2b-jet system with heaviest invariant mass and of the 2b-jet system with lightest invariant mass (as defined in the text). Three jet clustering combinations are plotted as detailed in the legend. The spectral clustering algorithm is consistently the best performer in terms of the narrowest peaks being reconstructed and comparable to anti-k_T/CA with $R_{k_T} = 0.8$ in terms of their shift from the true Higgs mass values, with anti-k_T/CA with $R_{k_T} = 0.4$ being the outlier.

Strengths

- Number of hyper parameters (6)
- Generality
- Control of the algorithm

Limits

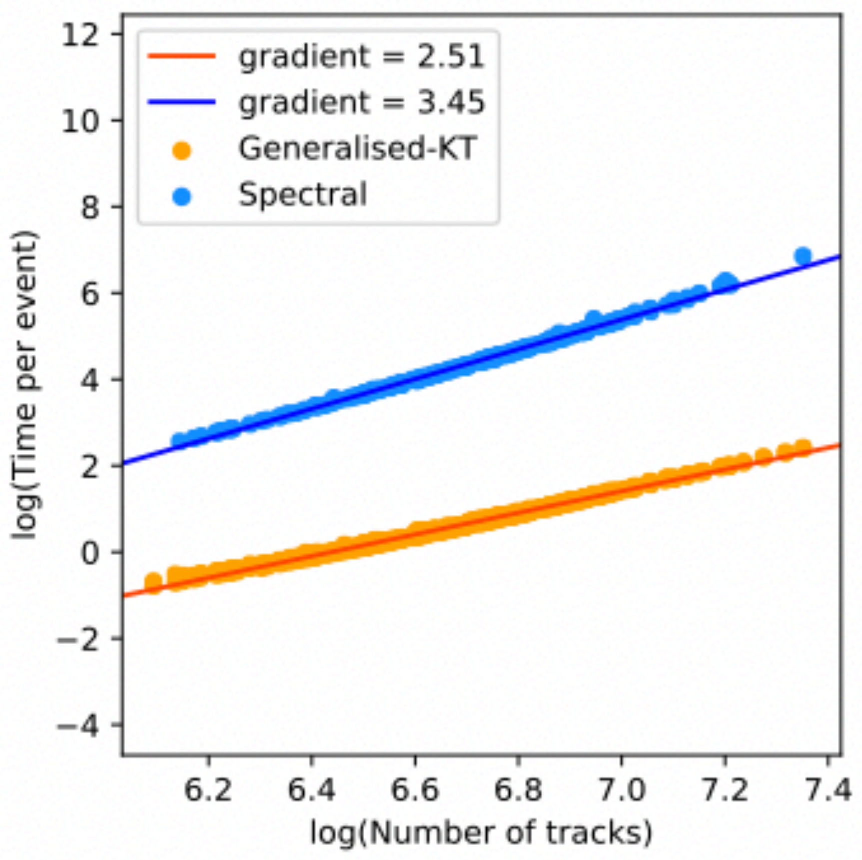
- Speed
- Pair merging

Strengths

- Number of hyper parameters (6)
- Generality
- Control of the algorithm

Limits

- Speed
- Pair merging



The run time of spectral clustering compared to a naïve implementation of generalised k_T (without the performance refinements in [31]), on datasets of varying size. Simple fits are shown for each dataset, in both linear and logarithmic scale. This shows that spectral clustering runs in just over $\mathcal{O}(n^3)$.

Summary

FastJet API: https://github.com/HenryDayHall/fastjet_spectraljet

- Spectral clustering: novelty for jet reconstruction
- Results:jet multiplicities and mass peaks
- Strengths and limits
- Coming next:
    - Make it faster
    - Explore the eigenspace

Reference: https://link.springer.com/article/10.1007/JHEP02(2022)165

Infrared & collinear safety

Distance in the new space:

$$d'(t)_{i,j} = s(t)_{i,j} \arccos\left(\frac{m(t)_i \cdot m(t)_j}{\|m(t)_i\|\|m(t)_j\|}\right)$$

↓

$$s_{i,j}(t) = 1 - \frac{\kappa}{\kappa + \min(p_T(t-1)_i, p_T(t-1)_j) d(t-1)_{i,j}},$$

Hyper parameters

- σ: scale parameter in physical space
- α: shape of the distribution
- k_{NN}: minimum number of neighbours
- λlimit: threshold for number of eigenvectors
- β: quality of information of eigenvectors
- R: stopping condition