

Clustering evaluation tools

Ewerton Belchior

January 30, 2025

Evaluation overview

Clustering workflow for SBND by Haiwang: <https://github.com/HaiwangYu/wcp-porting-img/tree/main/sbnd>

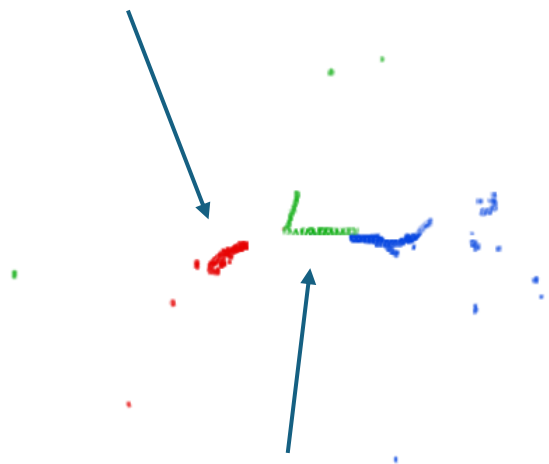
Goals:

- Create scripts/workflow to unpack clustering data
- Create a workflow to evaluate clustering performance (handscan + clustering metrics)
- Create new cluster IDs for truth energy deposits (originally truthDepos were clustered by APA) for matching. -> ***modified CellTree_Module.cc***
- Evaluate clustering (handscan + clustering metrics): completeness score and Adjusted Rand Index

BNB nu

<https://www.phy.bnl.gov/twister/bee/set/cd823059-dca1-48be-8cef-787b6cf2aa69/event/10/>

clustering



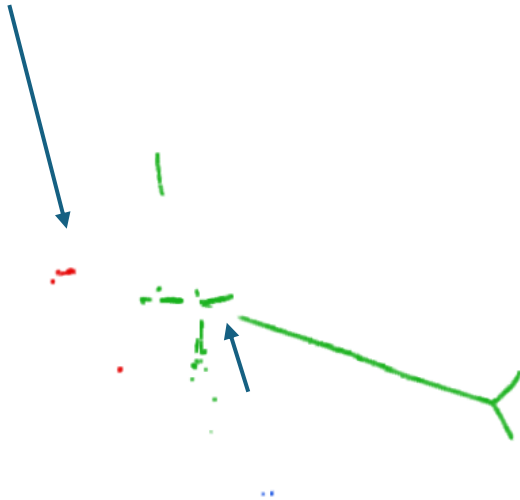
truthDepos



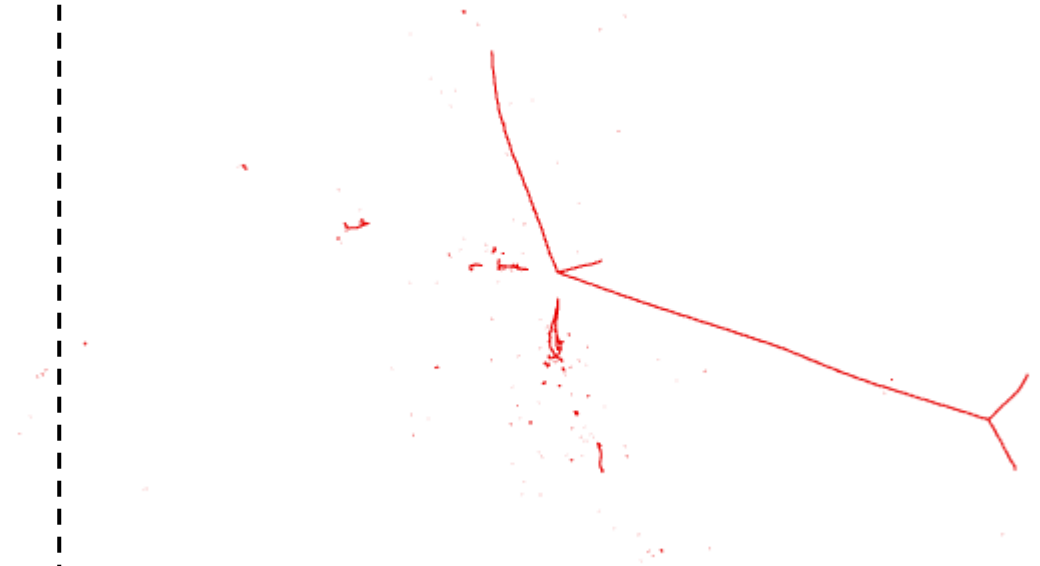
BNB nu

<https://www.phy.bnl.gov/twister/bee/set/cd823059-dca1-48be-8cef-787b6cf2aa69/event/4/>

clustering



truthDepos



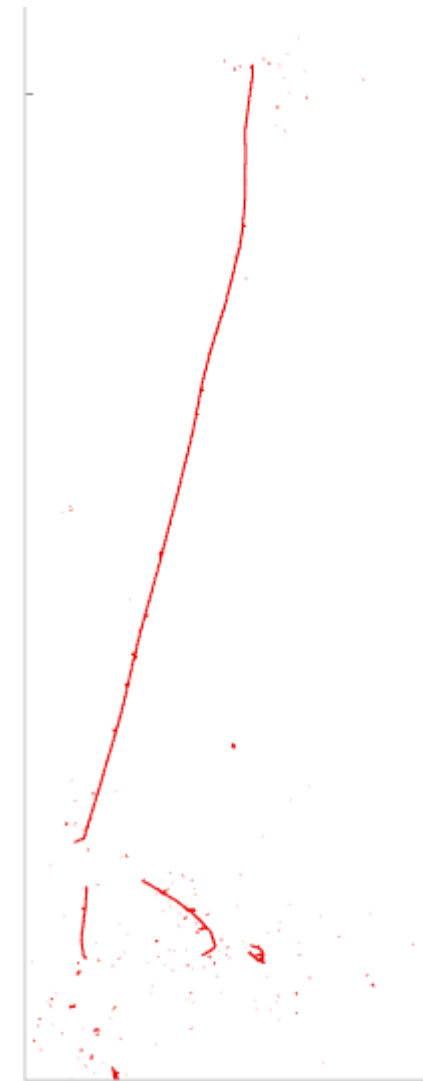
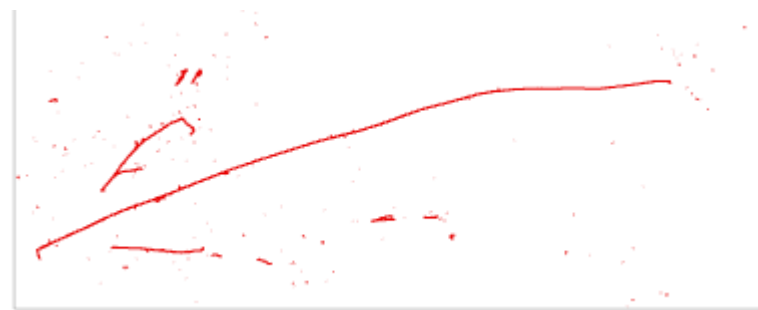
BNB nu

<https://www.phy.bnl.gov/twister/bee/set/cd823059-dca1-48be-8cef-787b6cf2aa69/event/19/>

clustering



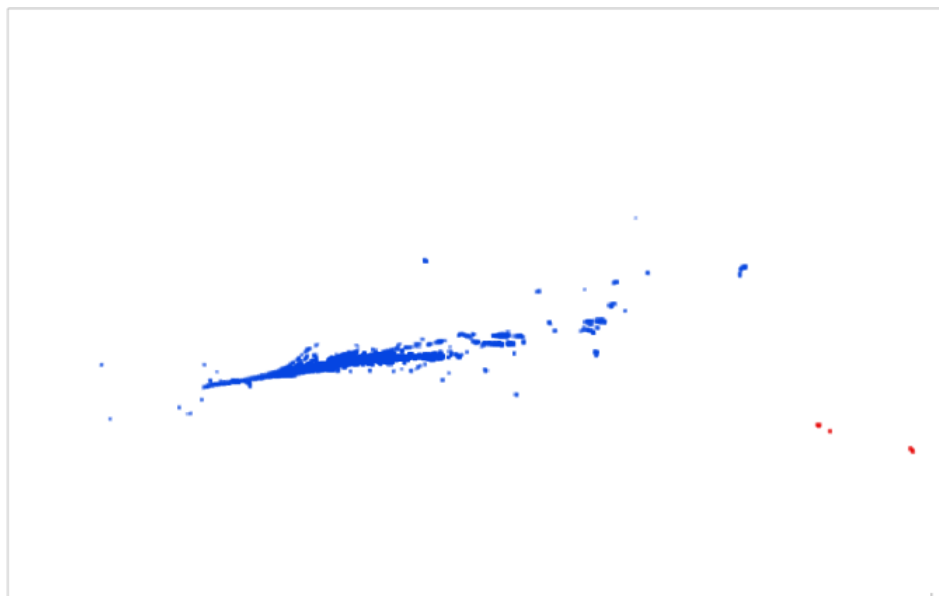
truthDepos



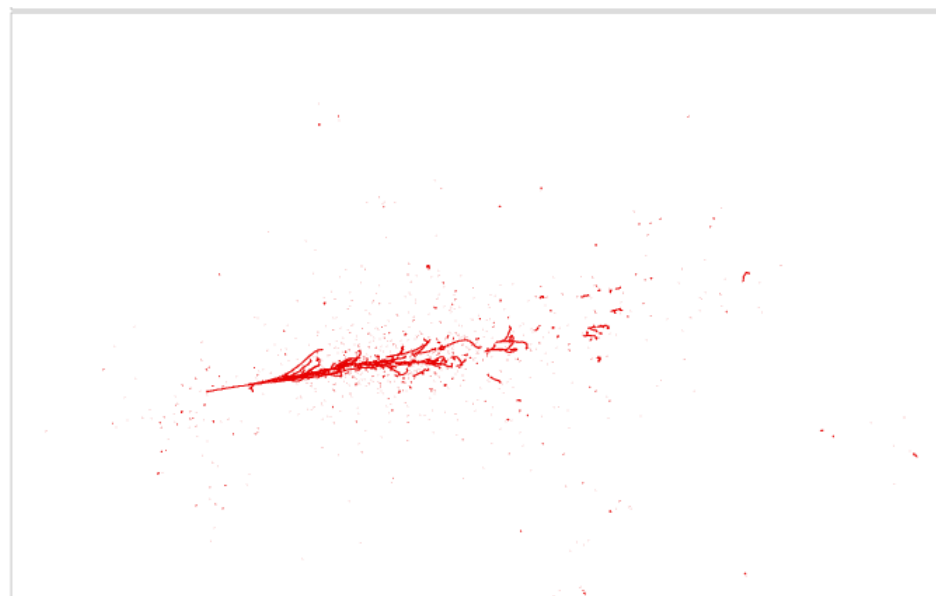
BNB nuE

<https://www.phy.bnl.gov/twister/bee/set/11e67291-4462-4c06-b0d0-2dc701ddeba7/event/13/>

clustering



truthDepos



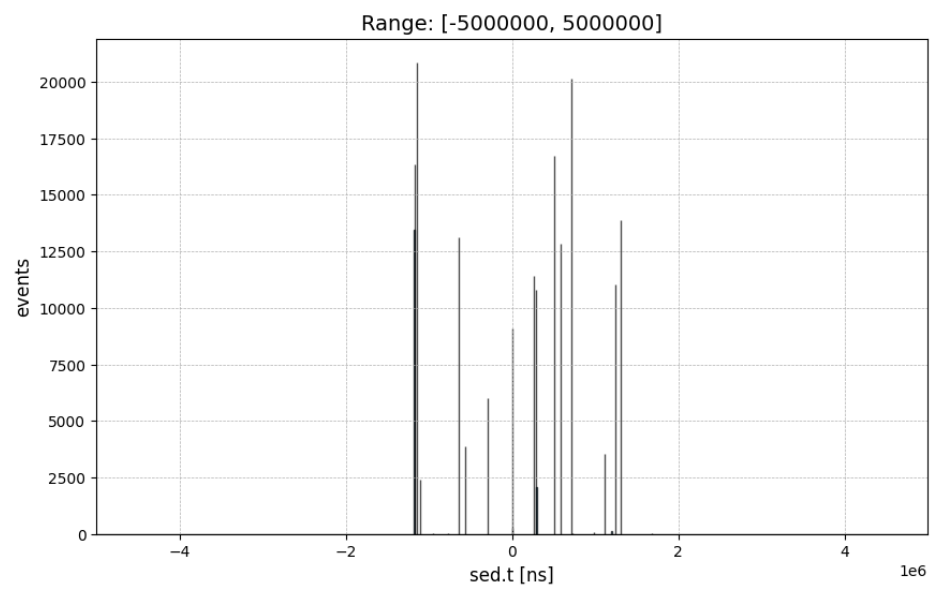
New clustering for truth energy deposits (CellTree_Module.cc)

- Matching developments can use truth cluster ID! (e.g. DL clustering)
- Use time information to cluster truth depos and create a new cluster ID.
- -> **will made available soon!**

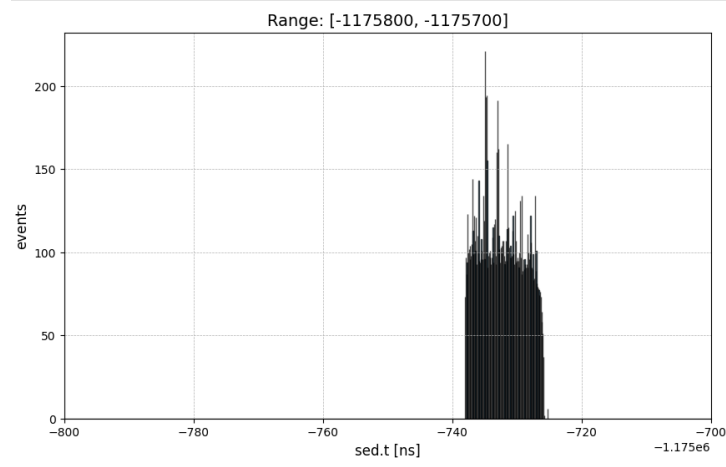
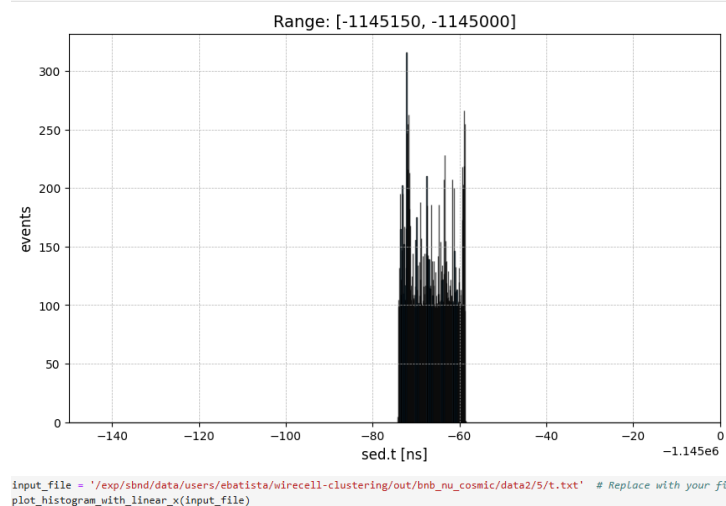
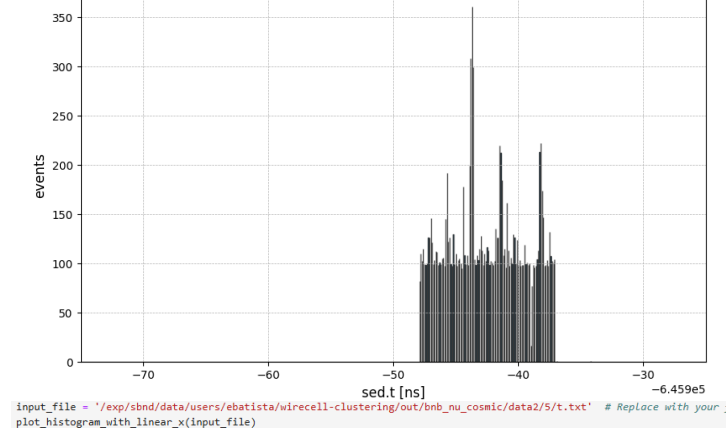
<https://www.phy.bnl.gov/twister/bee/set/3101d076-91b4-40bb-a690-b884a583f51a/event/5/>



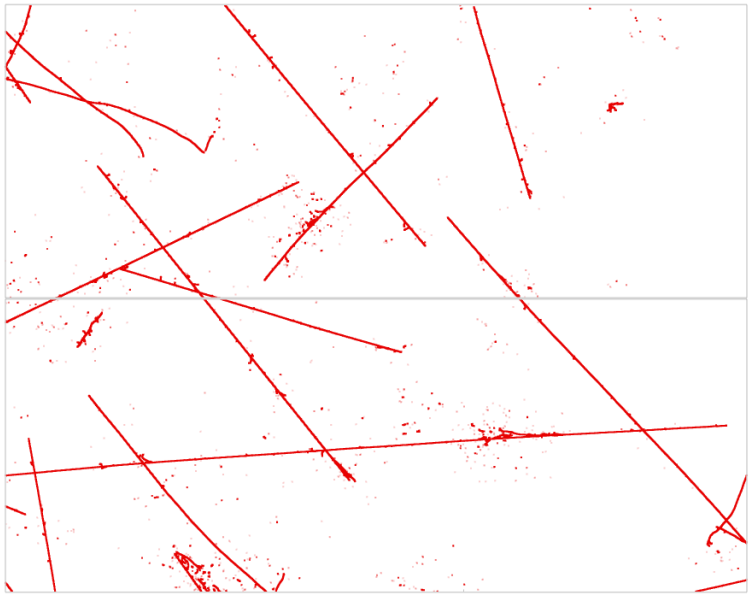
```
input_file = '/exp/sbnd/data/users/ebatista/wirecell-clustering/out/bnb_nu_cosmic/data2/5/t.txt' # Replace with your file
plot_histogram_with_linear_x(input_file)
```



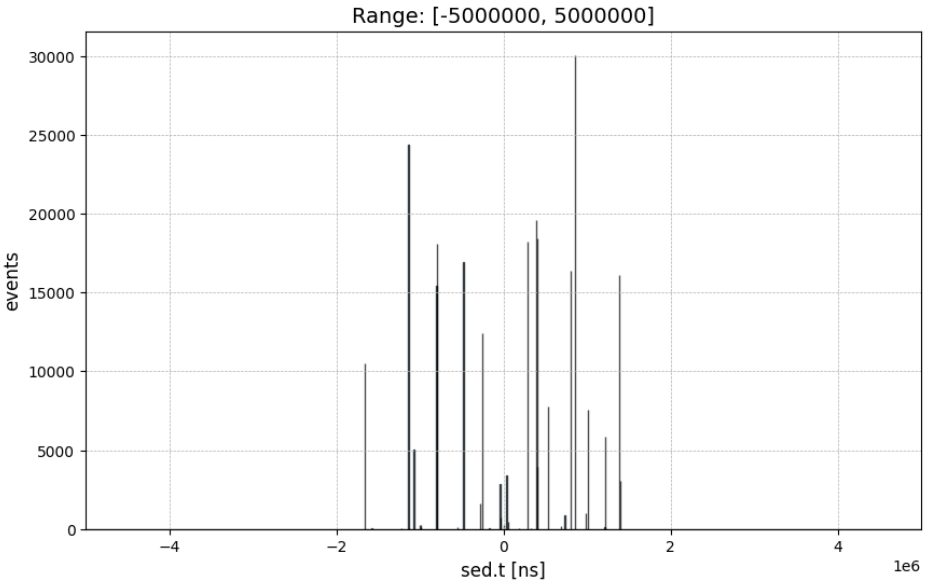
$\Delta t \sim 20$ ns



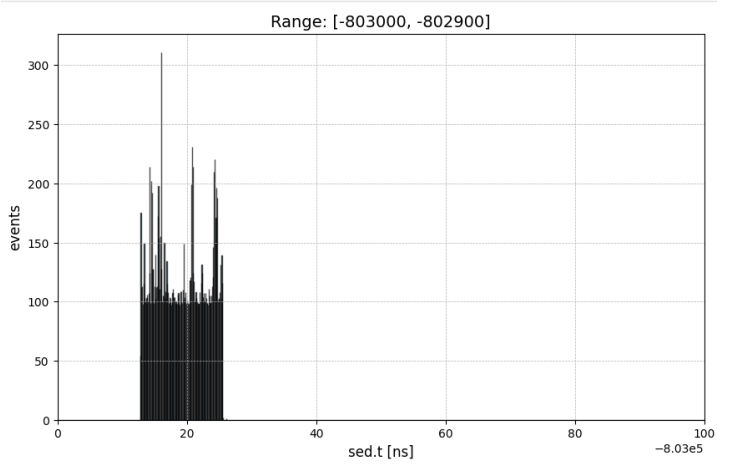
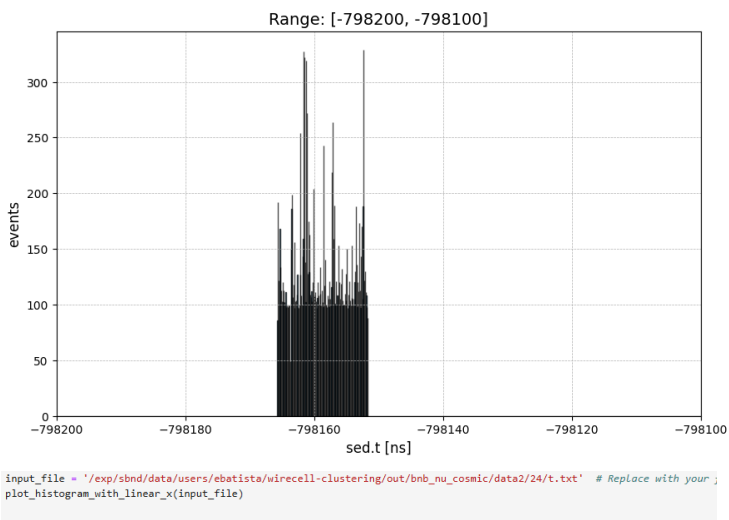
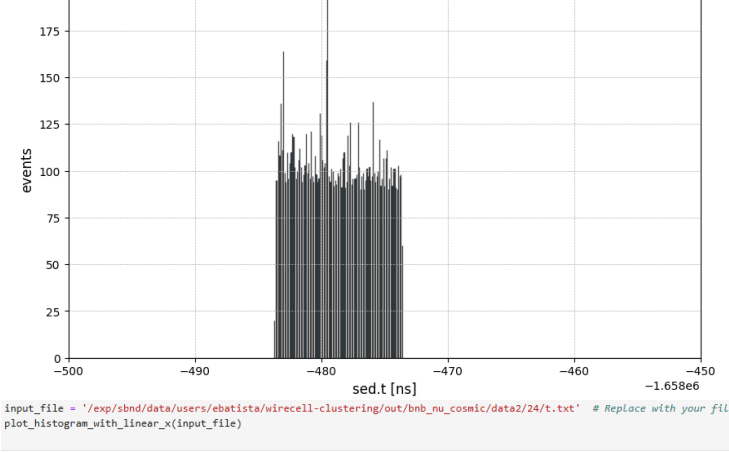
<https://www.phy.bnl.gov/twister/bee/set/3101d076-91b4-40bb-a690-b884a583f51a/event/24/>



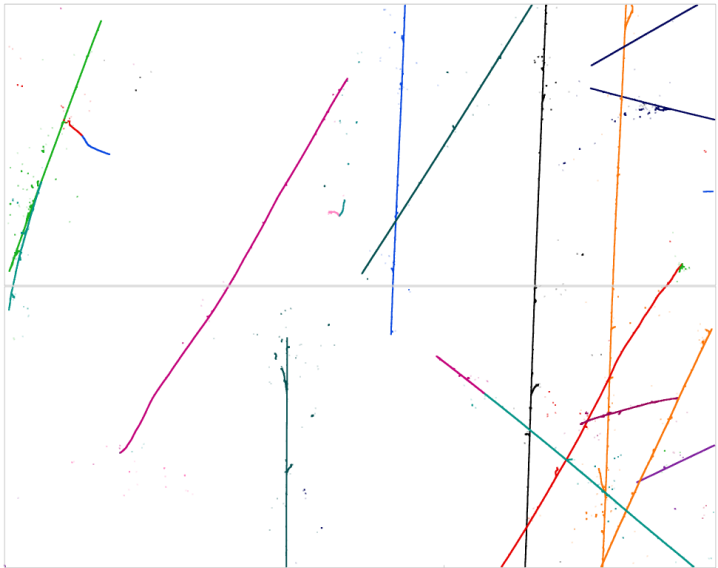
```
input_file = '/exp/sbnd/data/users/ebatista/wirecell-clustering/out/bnb_nu_cosmic/data2/24/t.txt' # Replace with your file
plot_histogram_with_linear_x(input_file)
```



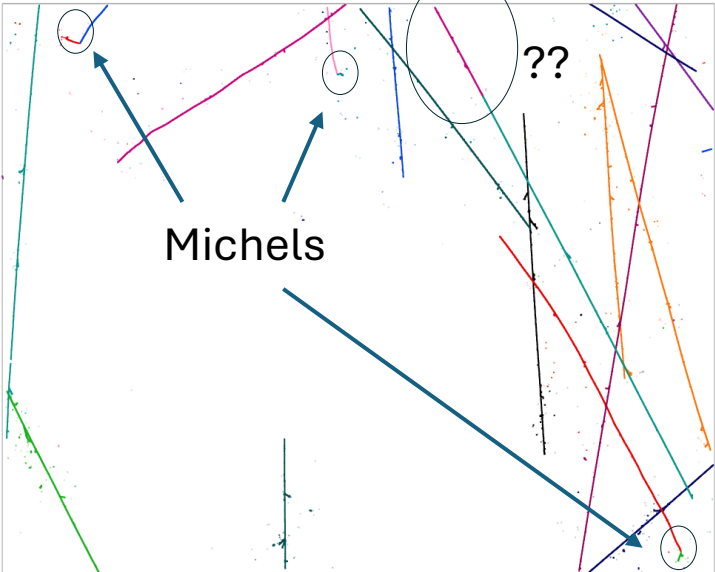
$\Delta t \sim 20 \text{ ns}$



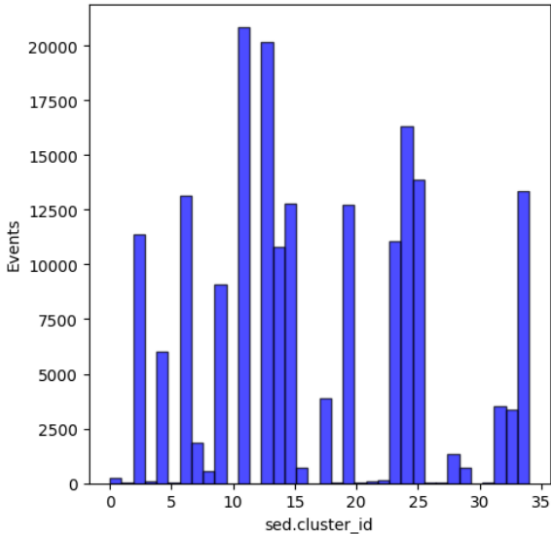
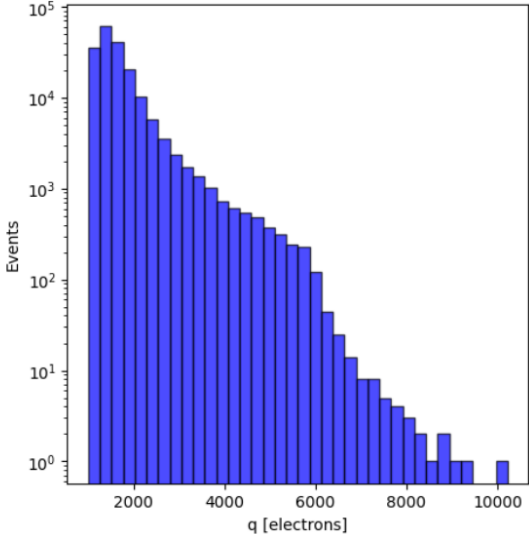
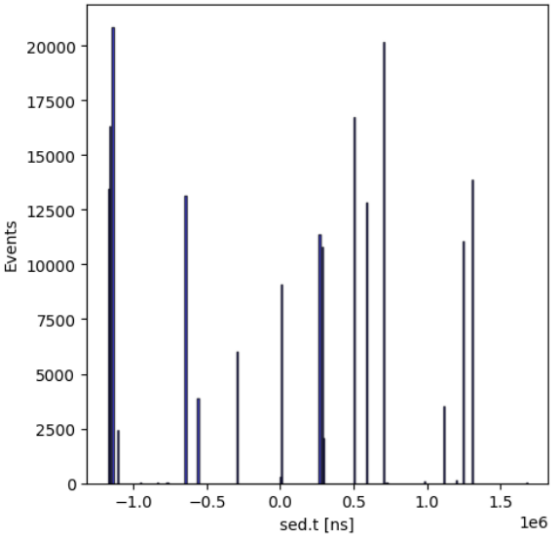
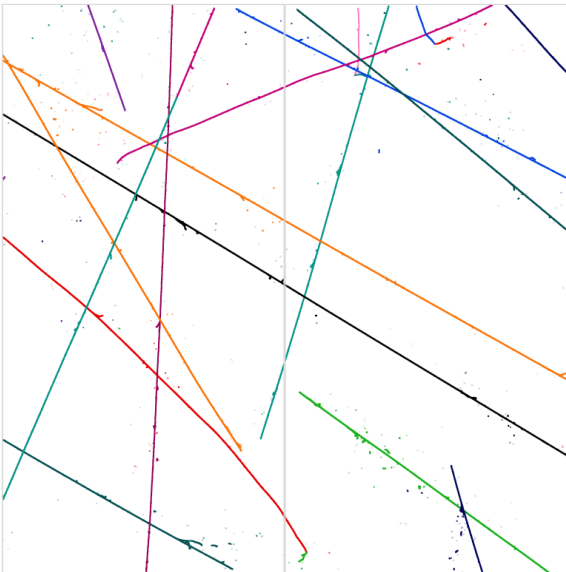
truthDepos ($\Delta t = 15$ ns , “Y-view”)



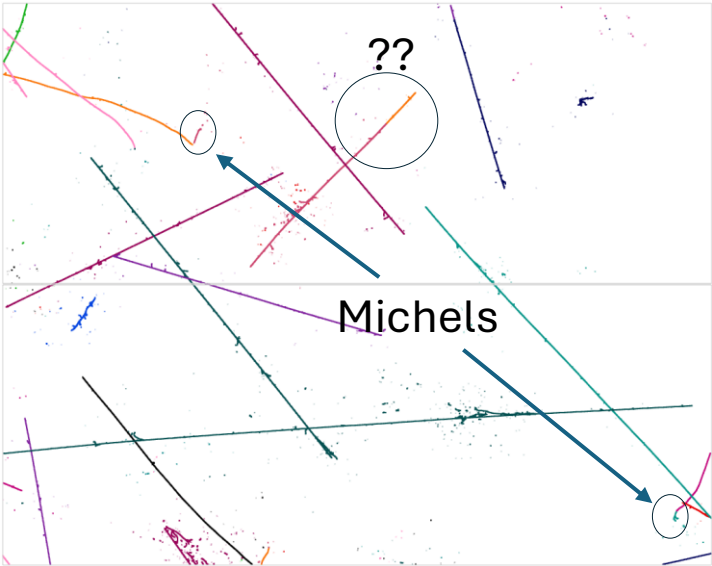
truthDepos ($\Delta t = 15$ ns, “X-view”)



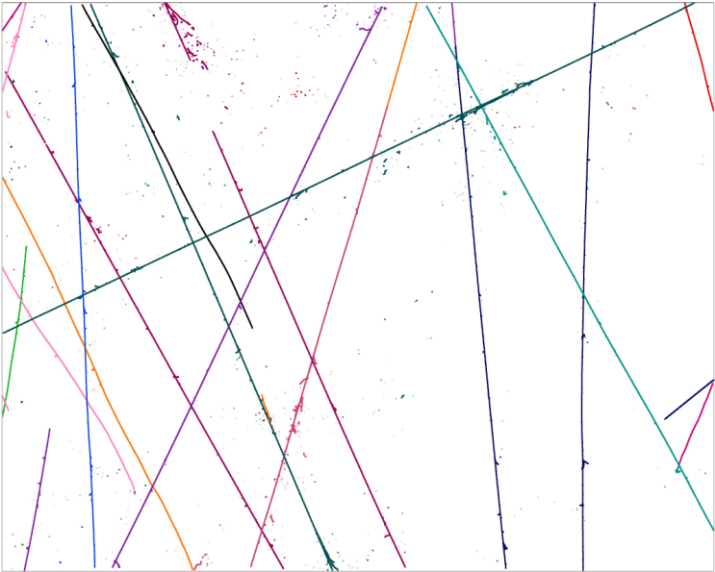
truthDepos ($\Delta t = 15$ ns, “Z-view”)



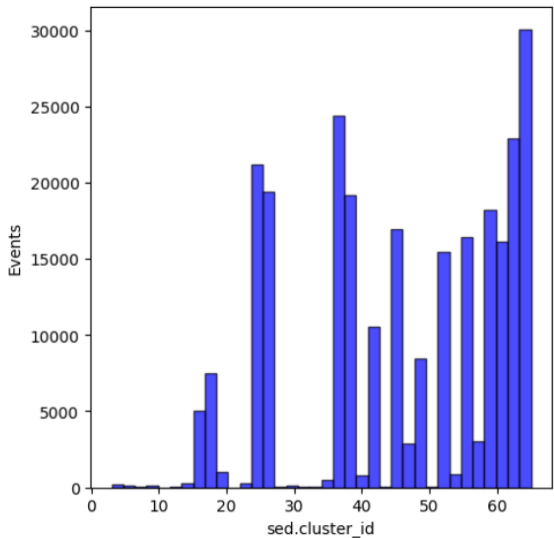
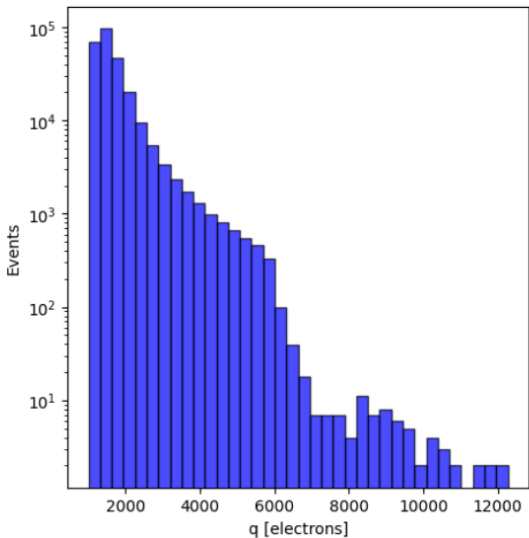
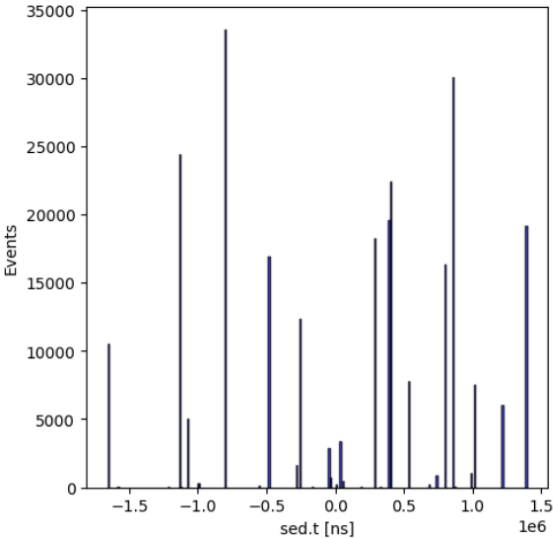
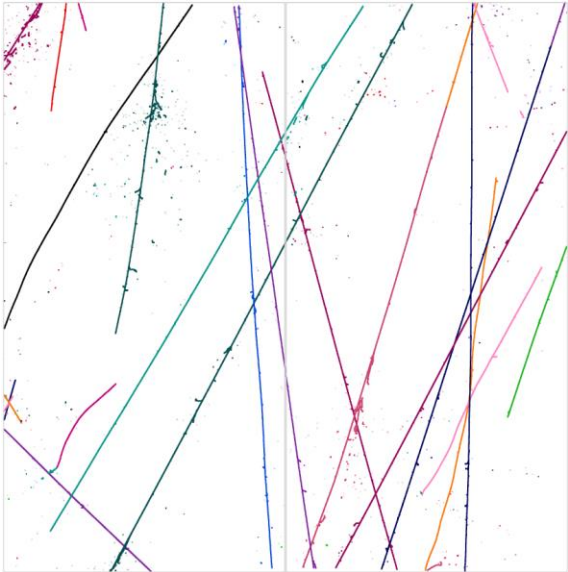
truthDepos ($\Delta t = 15$ ns , “Y-view”)



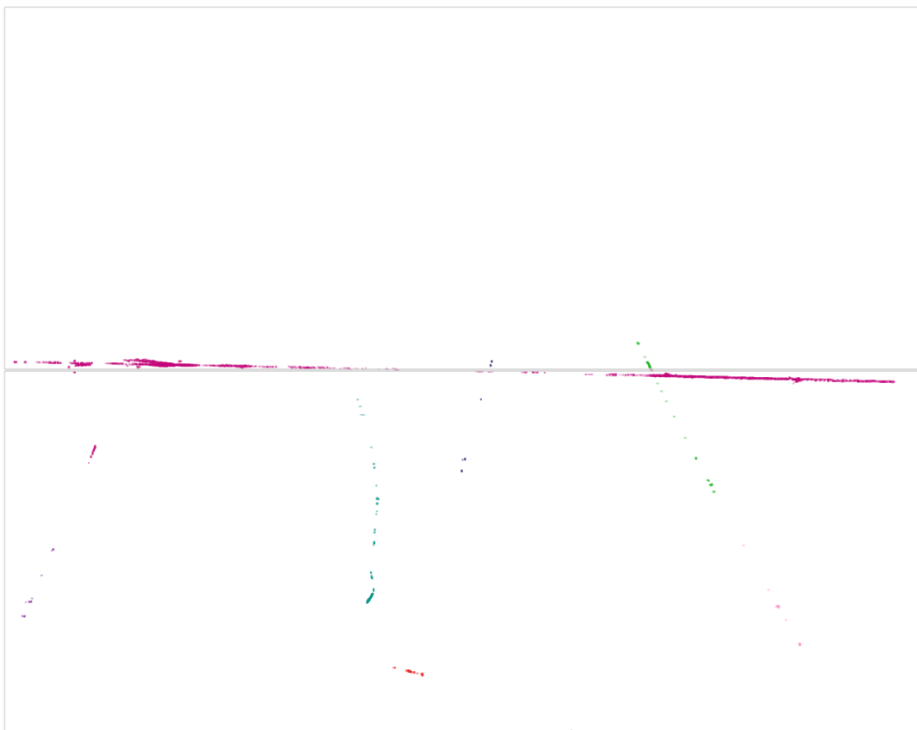
truthDepos ($\Delta t = 15$ ns, “X-view”)



truthDepos ($\Delta t = 15$ ns, “Z-view”)

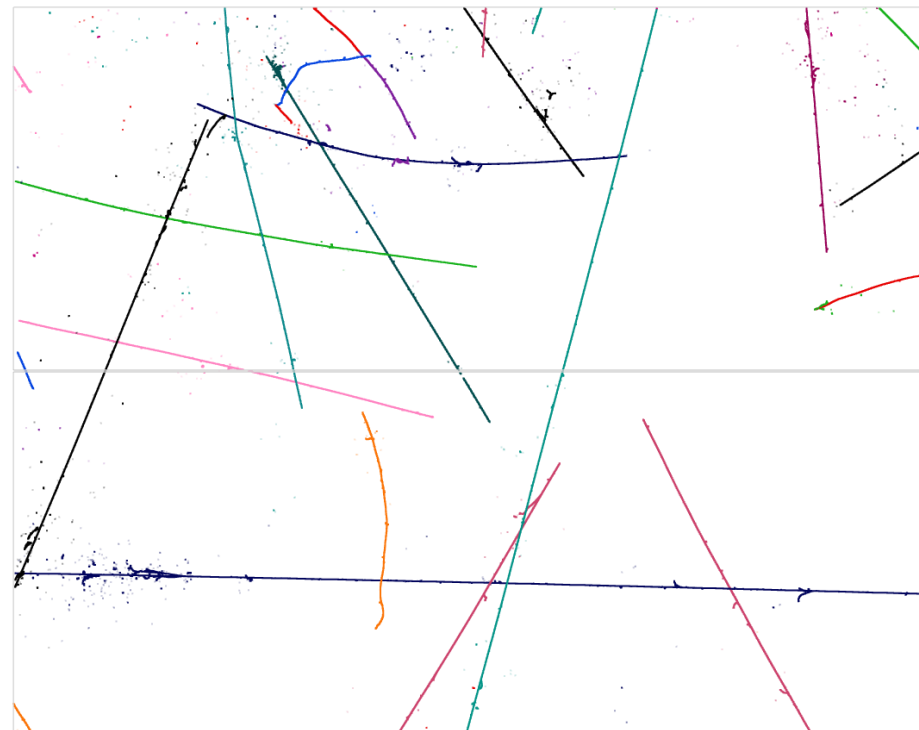


Clustering



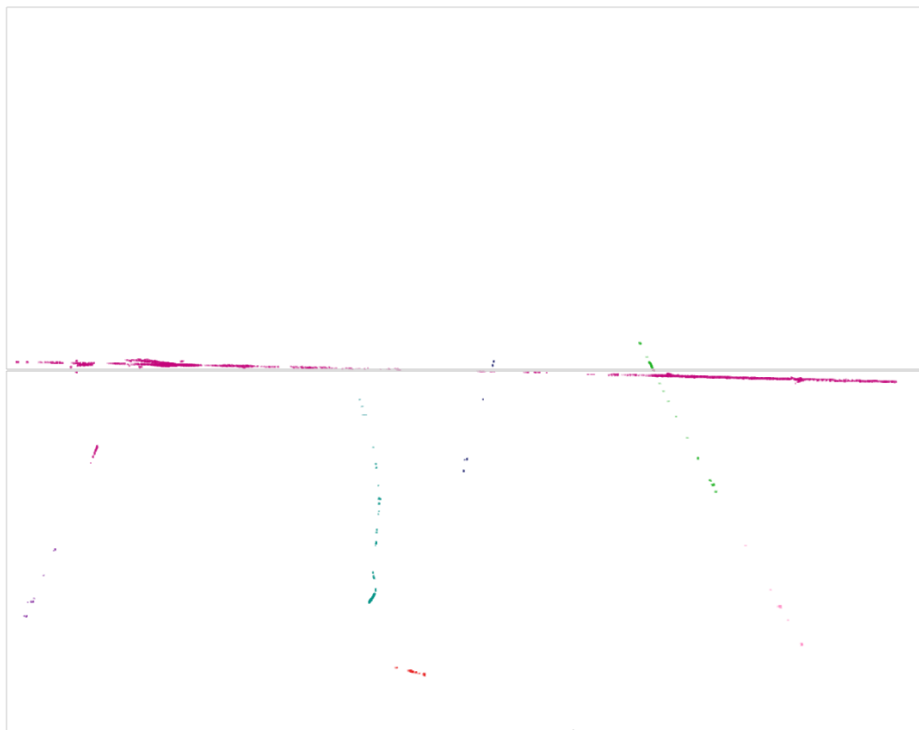
```
drift_speed = 1.56 * wc.mm / wc.us, // 1.56 for sbnd  
time_offset = -200 * wc.us,
```

truthDepos ($\Delta t = 15$ ns, "Y-view")



```
physics.analyzers.wirecell.drift_speed : 1.56 # mm/us  
physics.analyzers.wirecell.t0_corrected : true  
physics.analyzers.wirecell.saveJSON : true  
physics.analyzers.wirecell.SpacePointLabels : ["truthDepo"]  
physics.analyzers.wirecell.SimEnergyDepositLabel : "ionandscint:"
```

Clustering



```
drift_speed = 1.56 * wc.mm / wc.us, // 1.56 for sbnd  
time_offset = -200 * wc.us,
```

truthDepos ($\Delta t = 15$ ns, "Y-view")



```
physics.analyzers.wirecell.drift_speed : 1.56 # mm/us  
physics.analyzers.wirecell.t0_corrected : false  
physics.analyzers.wirecell.saveJSON : true  
physics.analyzers.wirecell.SpacePointLabels : ["truthDepo"]  
physics.analyzers.wirecell.SimEnergyDepositLabel : "ionandscint:"
```

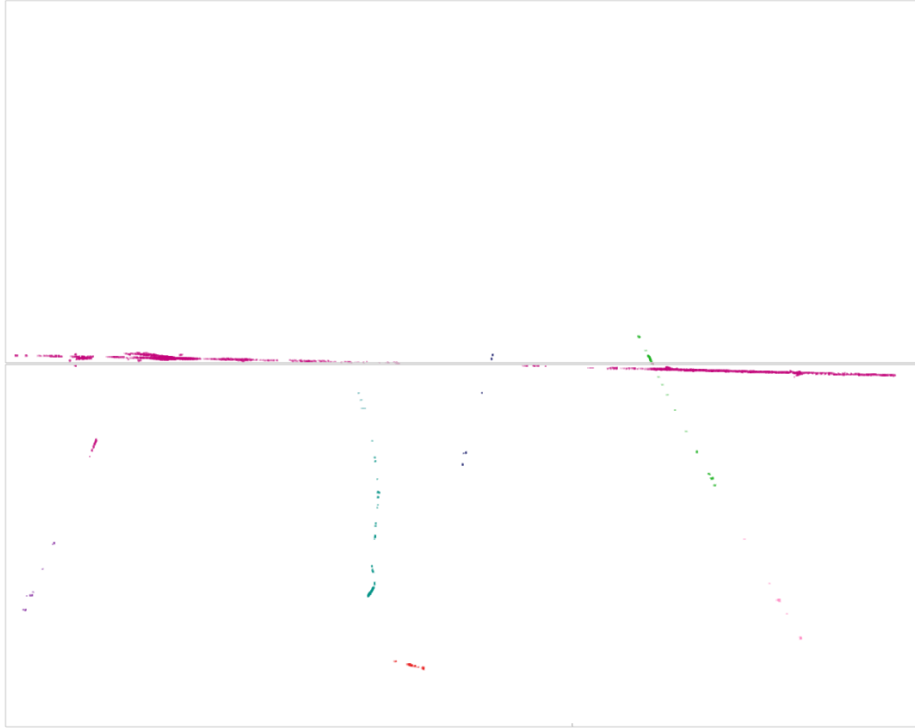
Evaluation steps overview

1. The idea is that each true point has a label (clusterid). The same for reco cluster points.
2. Match each reco cluster (predicted point) to true point using KDTree (query method with $k=1$).
3. Use their matched labels to calculate a completeness score and ARI score (based on their labels).

Completeness score: Calculate how well all points of the same true class are assigned to the same predicted cluster using the completeness score. Score between 0.0 and 1.0. 1.0 stands for perfectly complete labeling.

Adjusted Rand Index (ARI): Measures similarity between predicted clusters and true labels, adjusted for chance. **Range**: $[-1, 1]$, higher is better.

Clustering

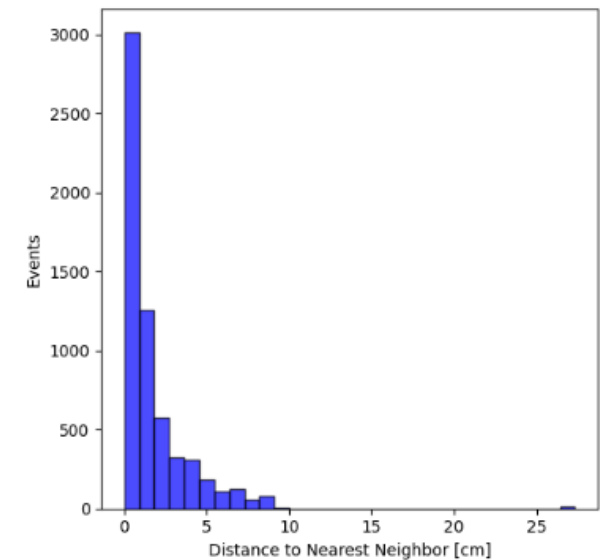


truthDepos

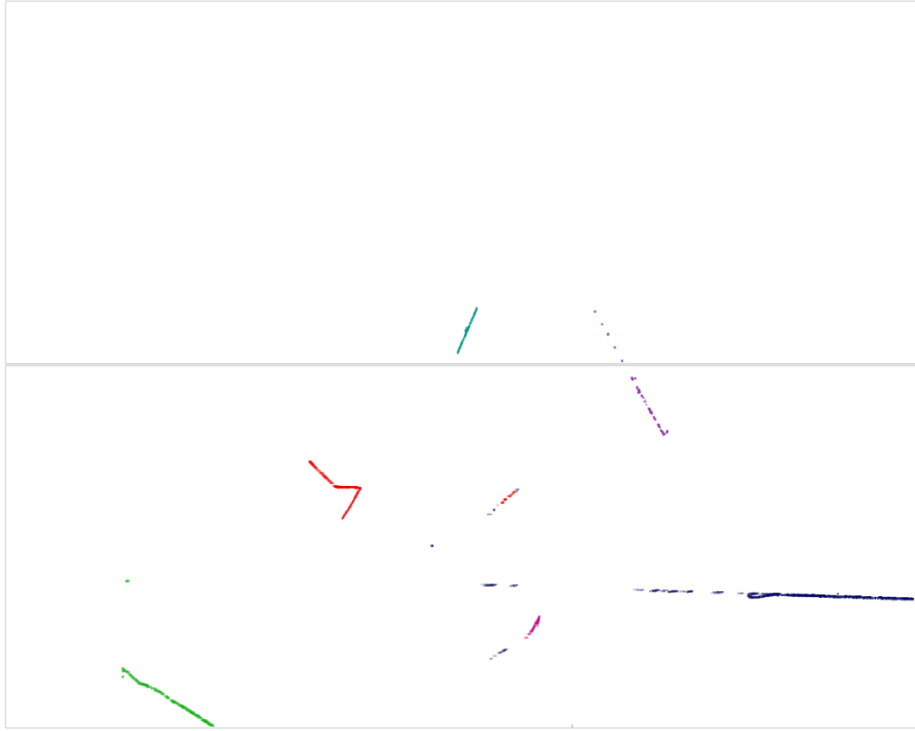


Completeness: 0.927
Adjusted Rand Index: 0.941

<https://www.phy.bnl.gov/twister/bee/set/671c3aff-29d9-44d2-af88-330927a22af1/event/1/>



Clustering

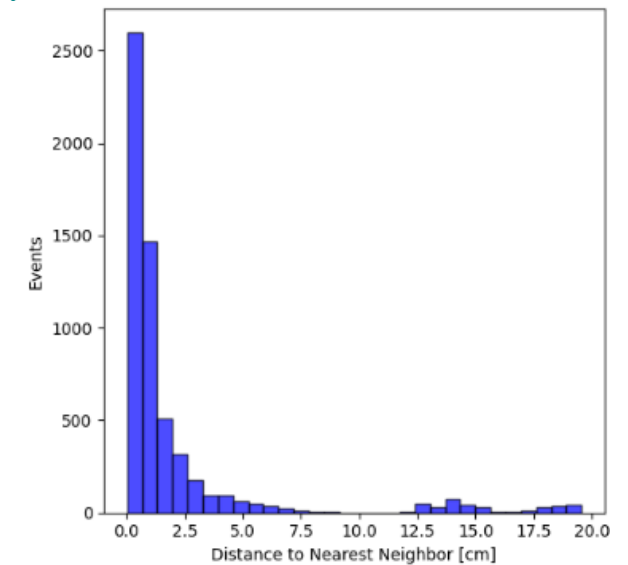


truthDepos

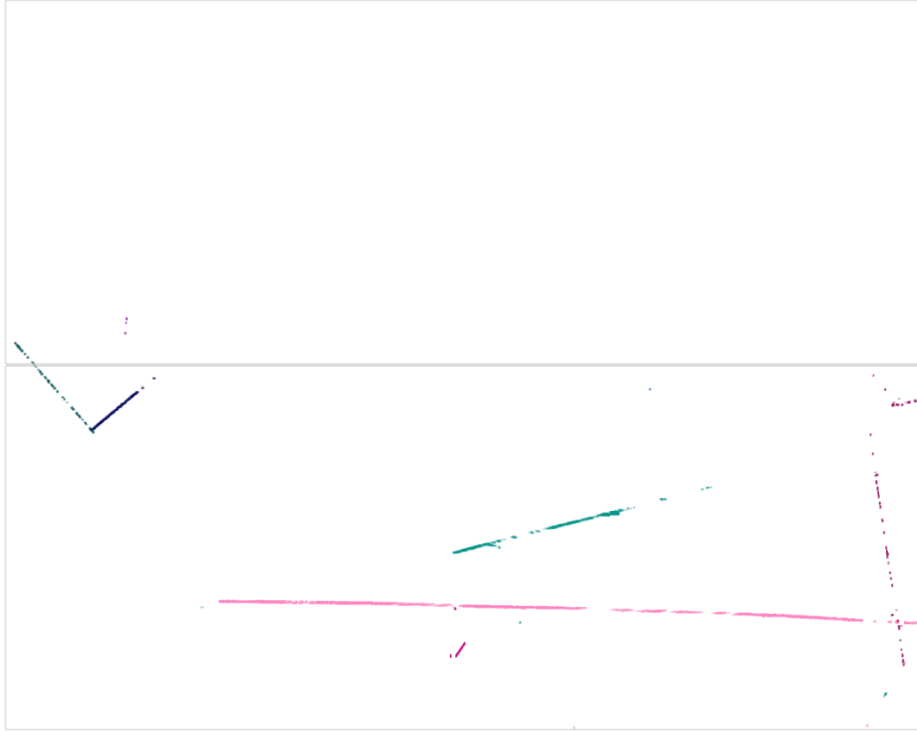


Completeness: 0.960
Adjusted Rand Index: 0.939

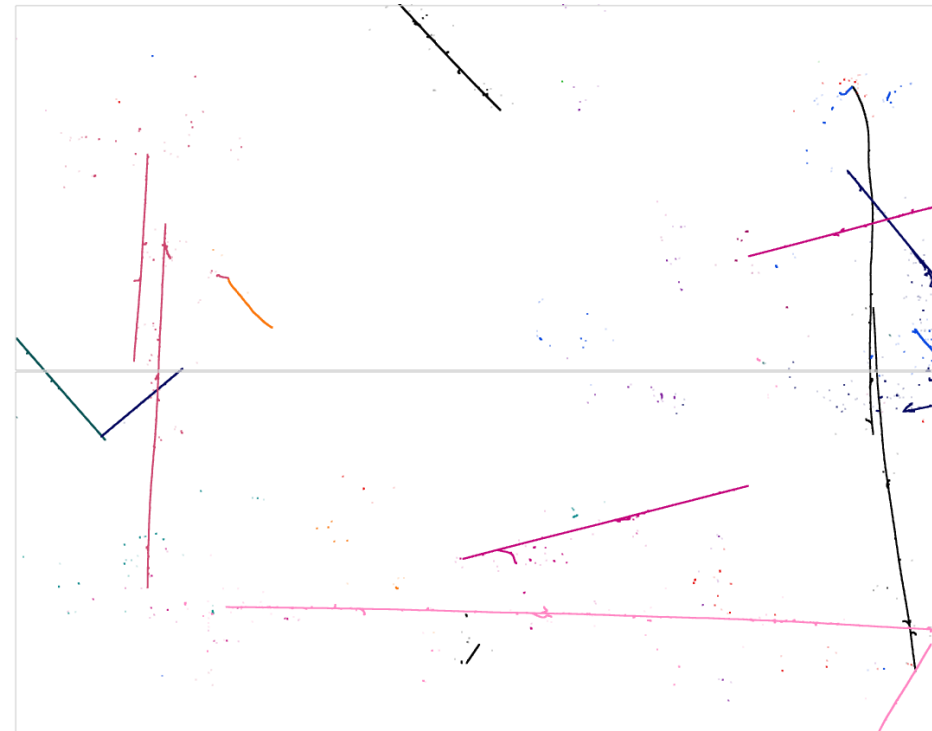
<https://www.phy.bnl.gov/twister/bee/set/671c3aff-29d9-44d2-af88-330927a22af1/event/2/>



Clustering

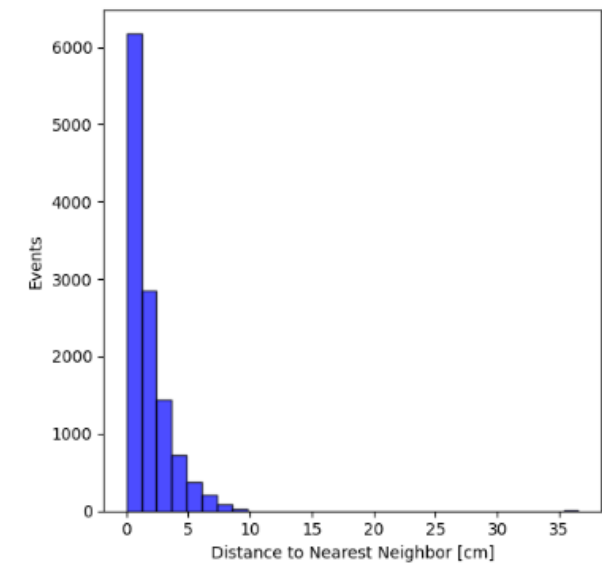


truthDepos

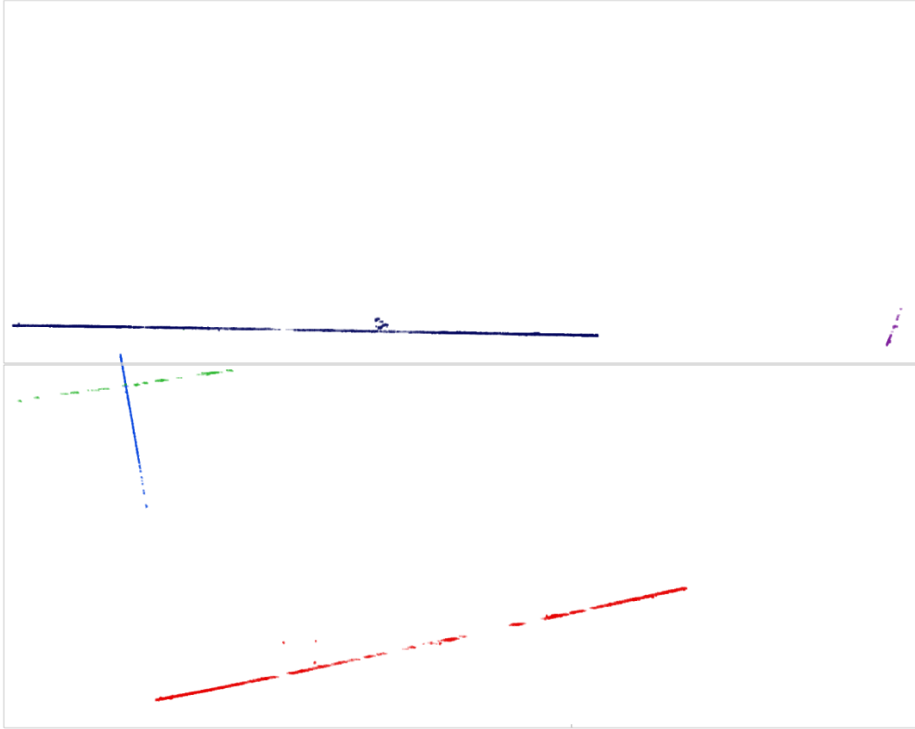


Completeness: 0.995
Adjusted Rand Index: 0.994

<https://www.phy.bnl.gov/twister/bee/set/671c3aff-29d9-44d2-af88-330927a22af1/event/3/>

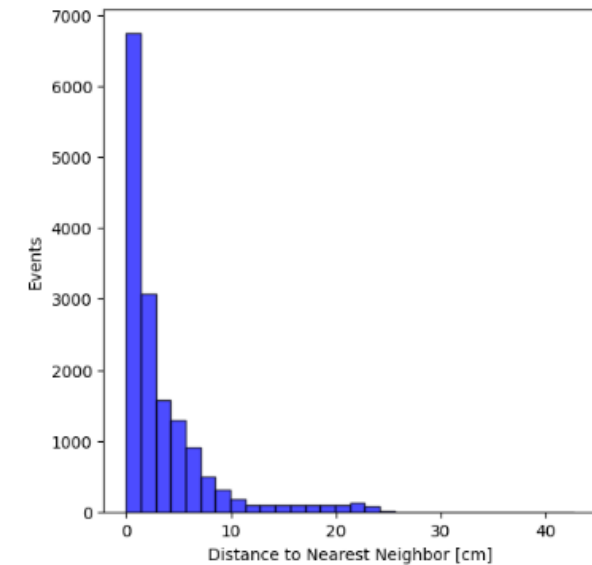
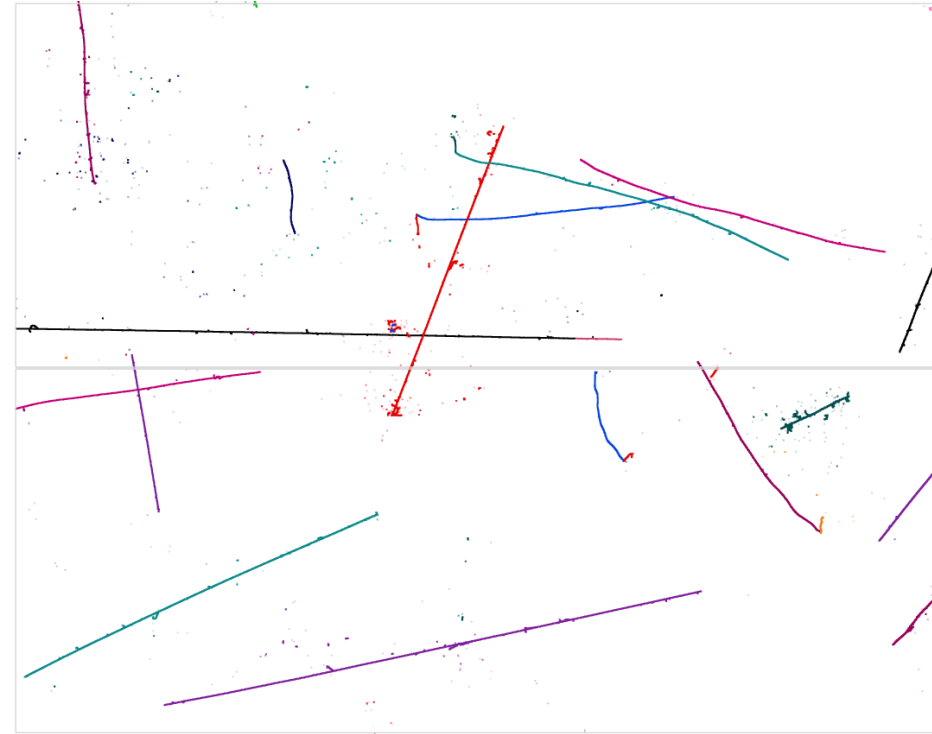


Clustering



Completeness: 0.999
Adjusted Rand Index: 0.856

truthDepos



<https://www.phy.bnl.gov/twister/bee/set/671c3aff-29d9-44d2-af88-330927a22af1/event/4/>