



云南大学中国西南天文研究所

South-Western Institute For Astronomy Research, YNU

# Machine learning classification of CHIME fast radio bursts – II. Unsupervised methods

[arxiv.org/abs/2210.02471](https://arxiv.org/abs/2210.02471)

Reporter: 金奕澄

2025.05.30

# Method

- Data: CHIME/FRB catalogue (CHIME/FRB Collaboration, 2021)
- Feature extraction
- Dimensionality Reduction
  - PCA (Principal Component Analysis)
  - t-SNE (t-distributed Stochastic Neighbor Embedding)
  - UMAP (Uniform Manifold Approximation and Projection)
- Clustering
  - k-means (with PCA)
  - HDBSCAN (with t-SNE or UMAP)

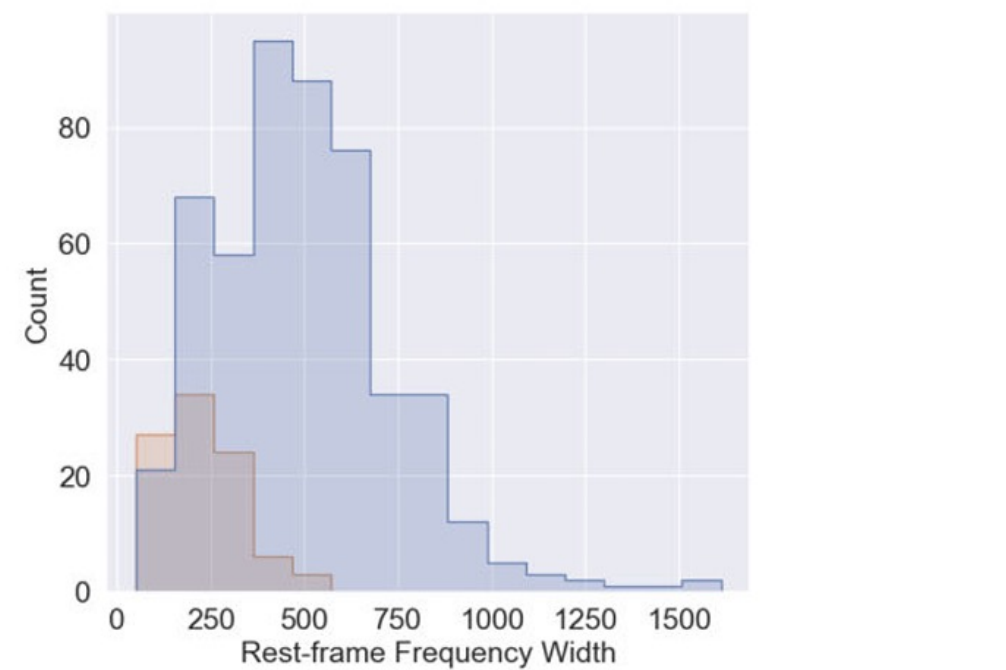
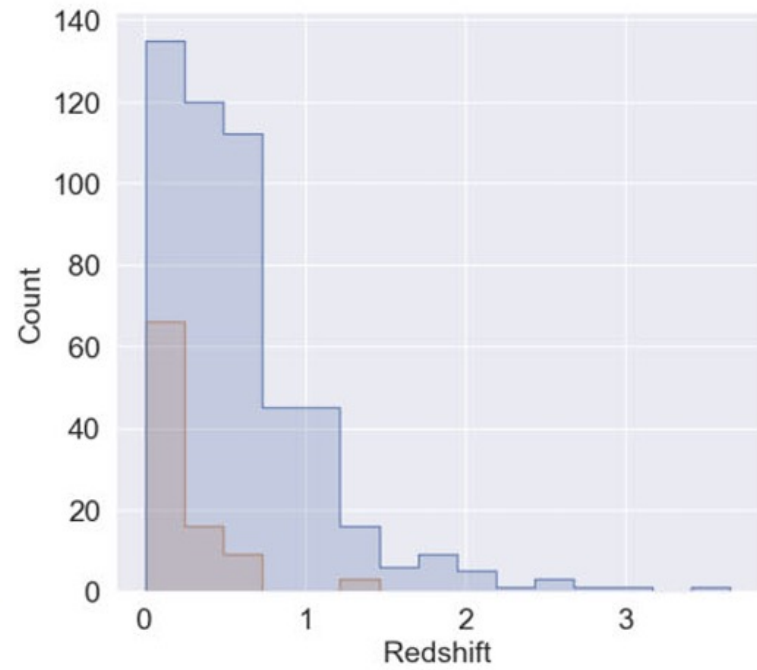
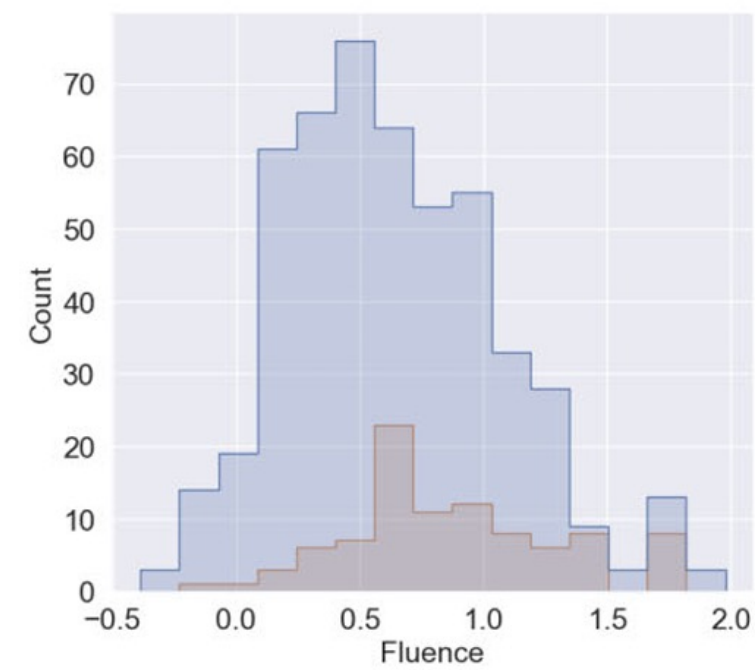
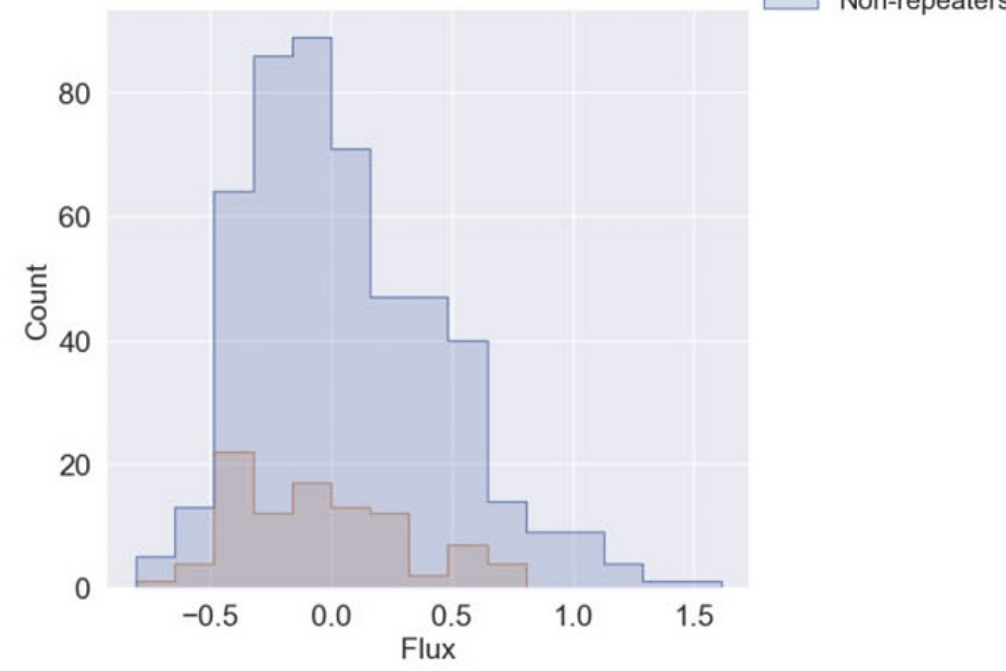
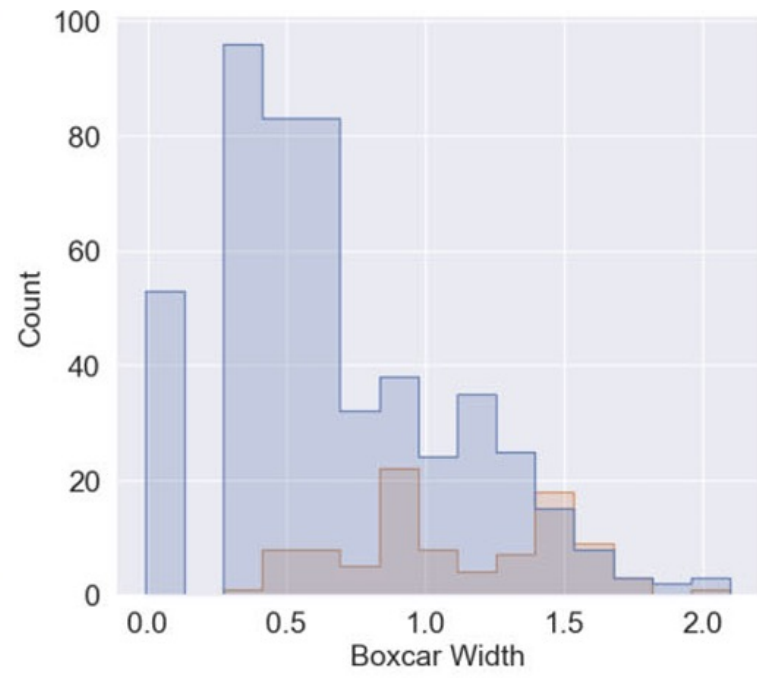
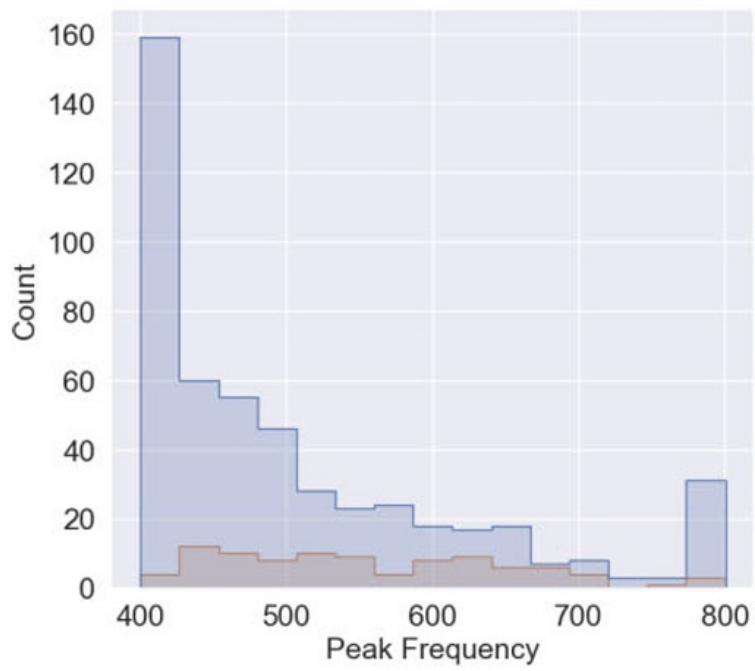
# Feature Extraction of FRBs

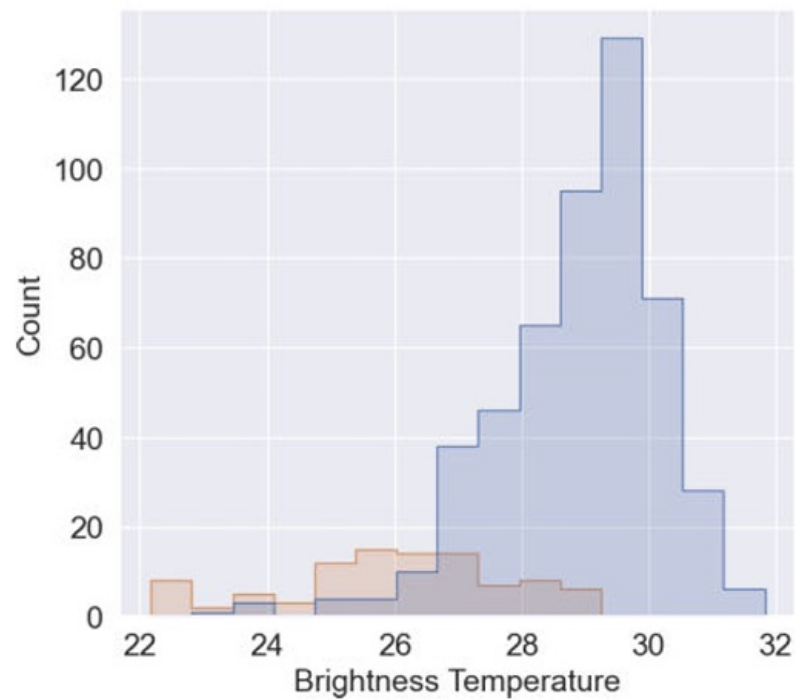
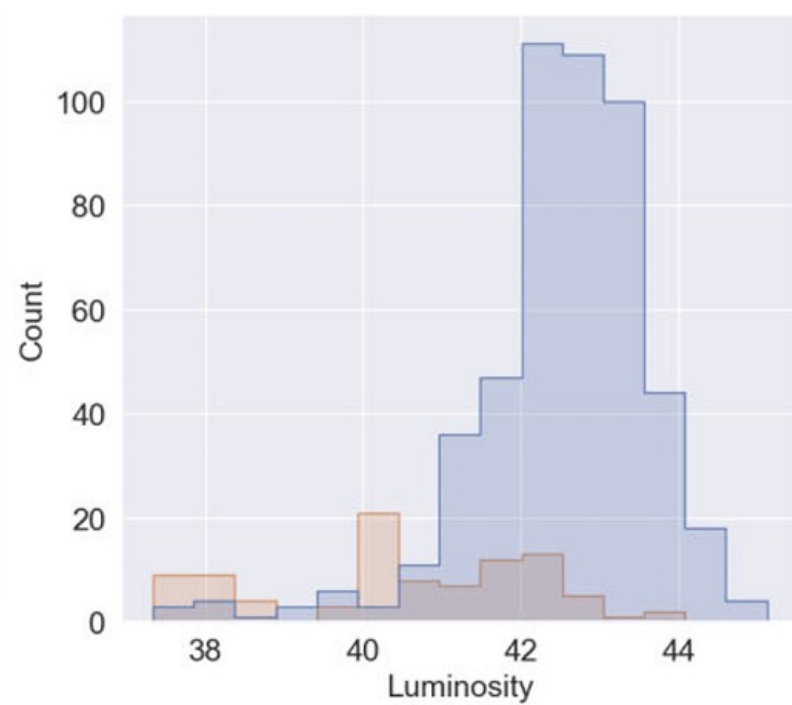
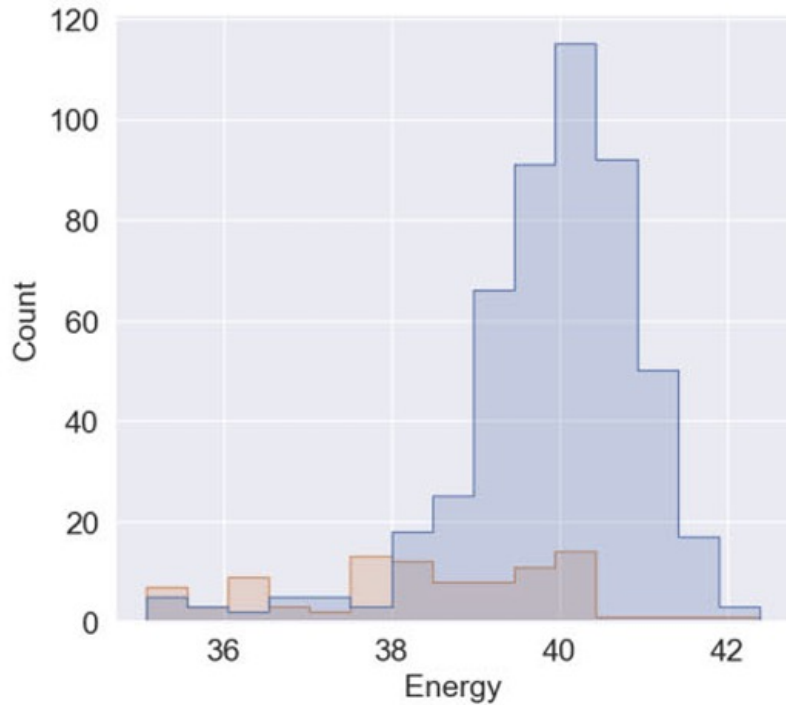
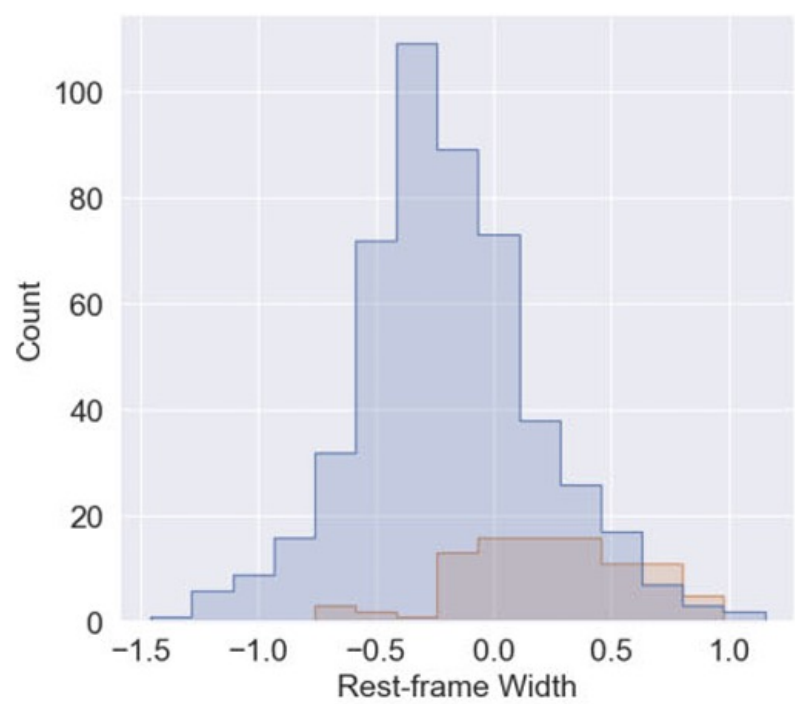
- Peak frequency  $\nu_c$  (MHz)
- Flux  $S_\nu$  (Jy) (peak flux of the band-average profile) (logarithmic)
- Fluence  $F_\nu$  (Jy ms) (logarithmic)
- Boxcar width  $\Delta t_{BC}$  (logarithmic)
- Redshift  $z$  (from  $DM_{IGM}$ )
- Rest-frame frequency width  $\Delta\nu$  (Broadband or narrowband?)
- Rest-frame pulse width  $\Delta t_r$  (logarithmic)
- Burst energy  $E$  (logarithmic)
- Peak luminosity  $L$  (logarithmic)
- Luminosity  $T_B$  (logarithmic)

# Redshift

- FRB DMs from Deng & Zhang 2014; Gao, Li & Zhang 2014; James et al. 2022
- $DM_{\text{MWS}}$  from NE2001 (Cordes & Lazio 2002)
- $\lambda$ CDM model from Planck Collaboration VI 2020
  - $H_0 = 67.4 \text{ km s}^{-1} \text{ Mpc}^{-1}, \Omega_m = 0.315$
- They assumed both H and He are fully ionized, i.e.  $\chi(z) \sim 7/8$ .
- $f_{\text{IGM}} \sim 0.83$  (Hogan & Peebles 1998)
- $DM_{\text{halo}}: 30 \text{ pc cm}^{-3}, DM_{\text{host}}: 70 \text{ pc cm}^{-3}$  (Dolag et al. 2015 ; Xu & Han 2015 ; Arcus et al. 2020 ; Hashimoto et al. 2020 ; Yamasaki & Totani 2020)
- Minimum redshift 0.002248 (10Mpc) (to avoid zero or negative values)

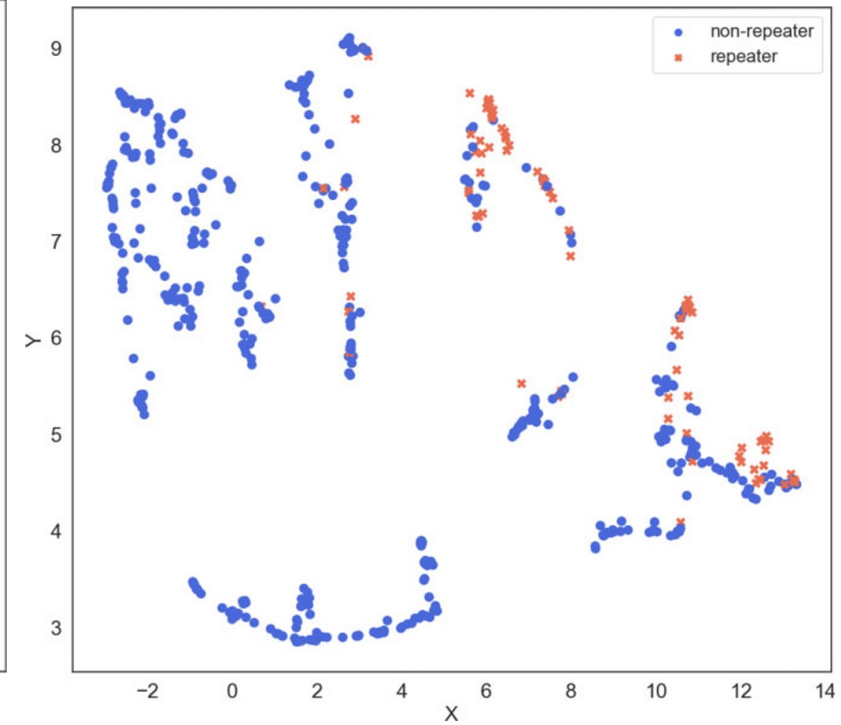
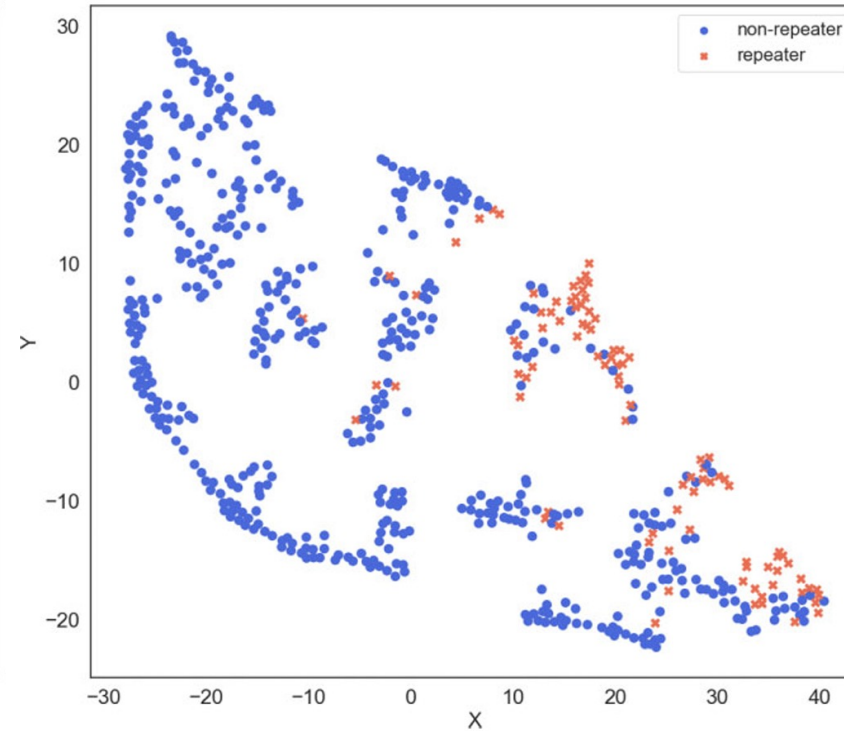
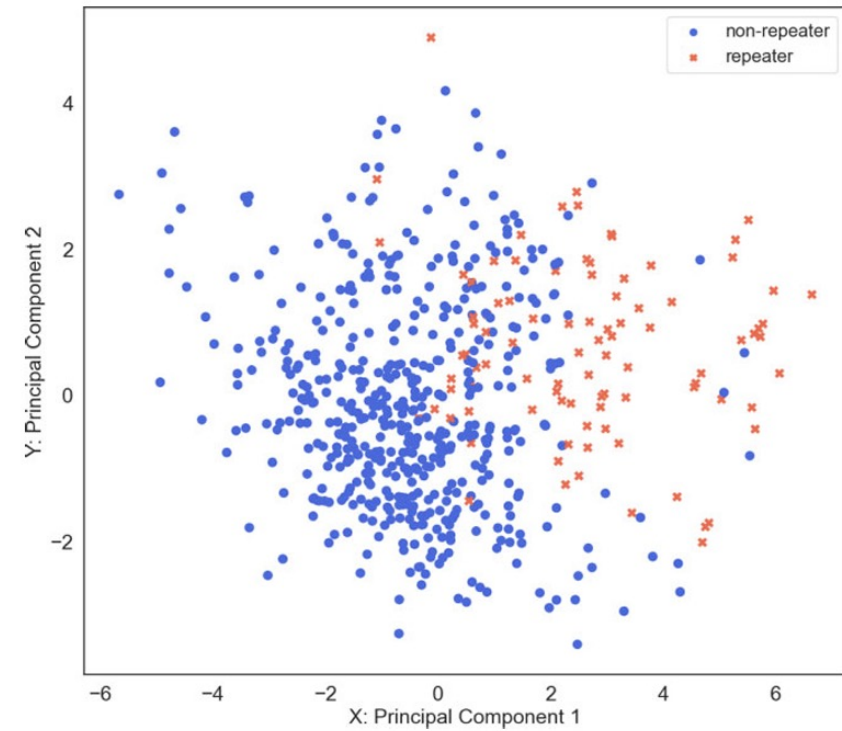
Repeaters  
Non-repeaters





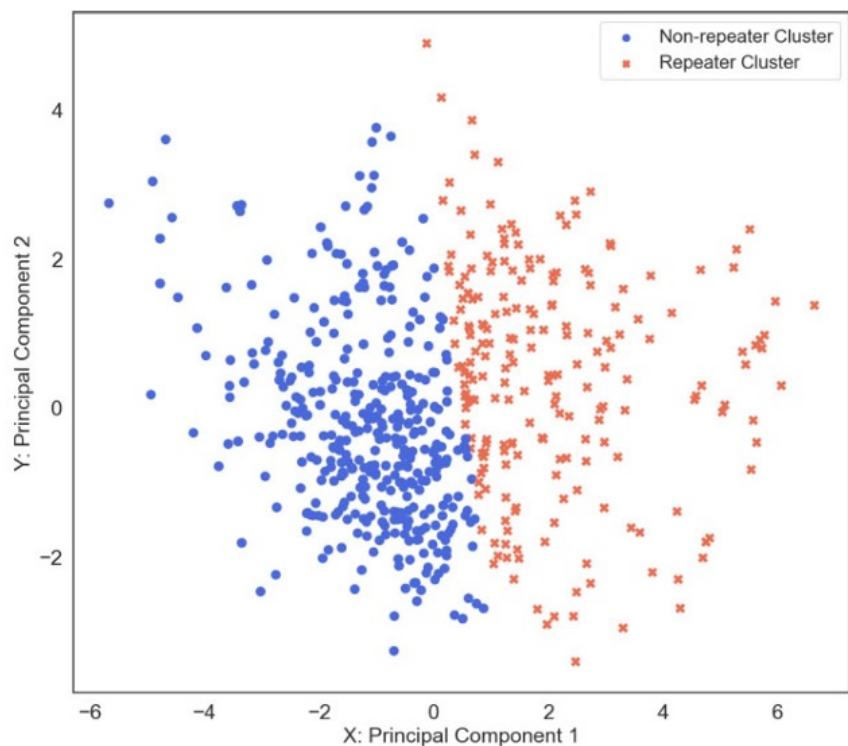
Repeaters  
Non-repeaters

# Dimensionality Reduction

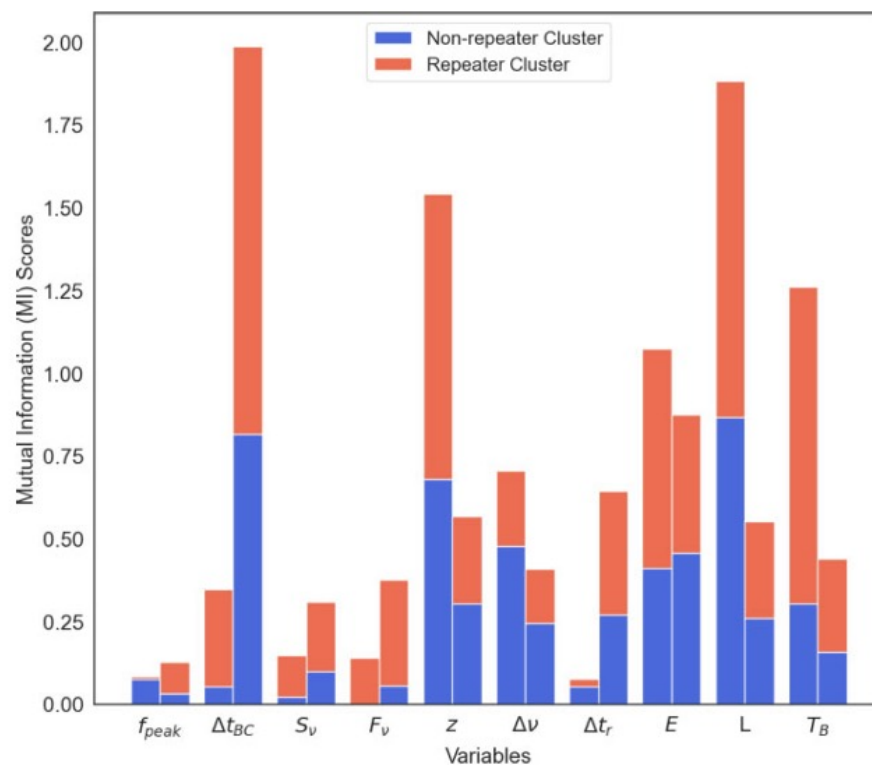


(repeater or non-repeater is based on observation.)

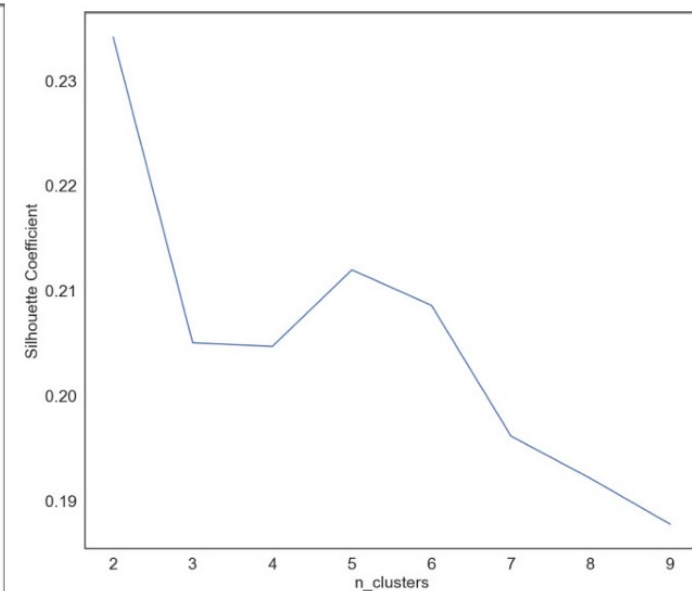
# PCA + K-means Clustering



**Figure 6.** The  $k$ -means clustering result in PCA space. The right cluster, containing a higher ratio of repeaters, is identified as the repeater cluster, while the left one is identified as the non-repeater cluster. The two clusters are consistent with the distribution of observed repeaters and non-repeaters in Fig. 2.



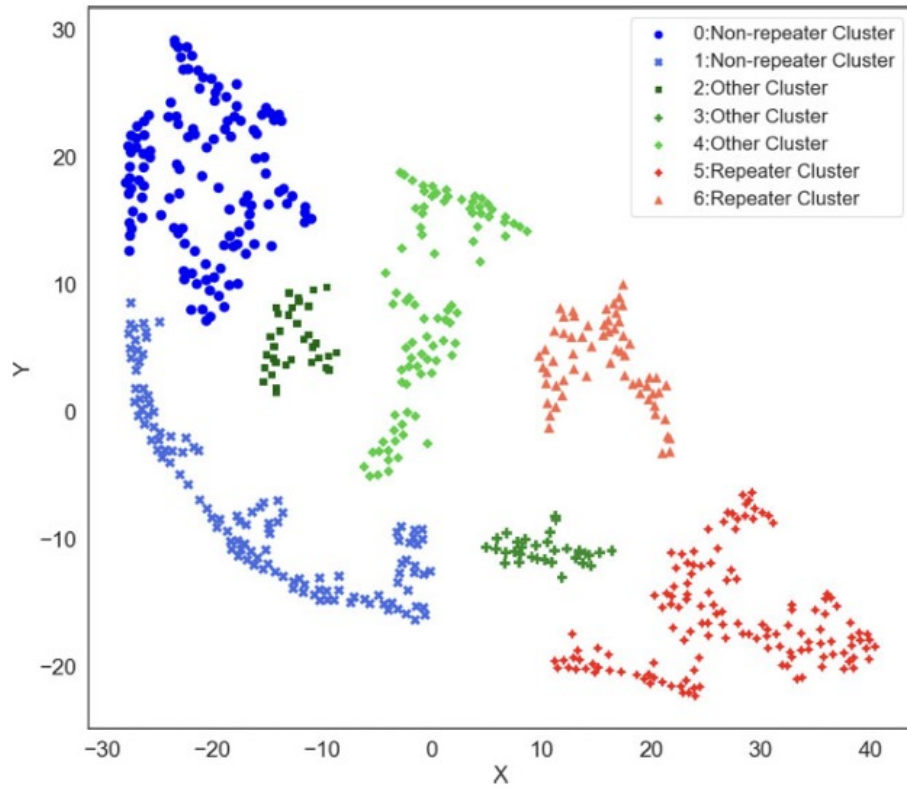
**Figure 9.** Feature correlation of PCA +  $k$ -means. The left bar correlates with the x-axis, while the right bar is related to the y-axis. Boxcar width, redshift, energy, luminosity, and brightness temperature are the most important features in PCA.



**Figure 14.** Silhouette coefficients of  $k$ -means with respect to different number of clusters, directly in the raw space without PCA. In addition to the maximum value at two clusters, the plateau corresponding to five and six clusters implies possible subcategories with lower significance.

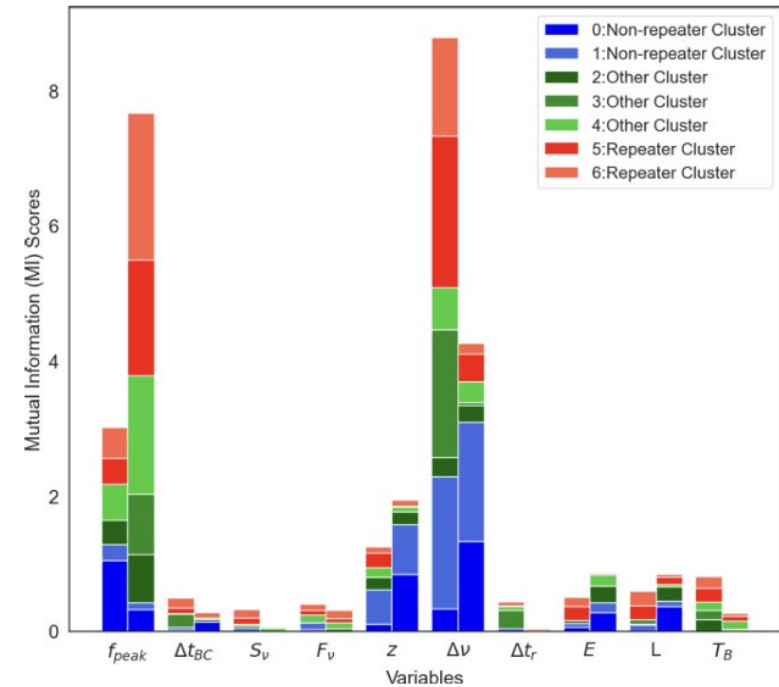


# t-SNE + HDBSCAN Clustering



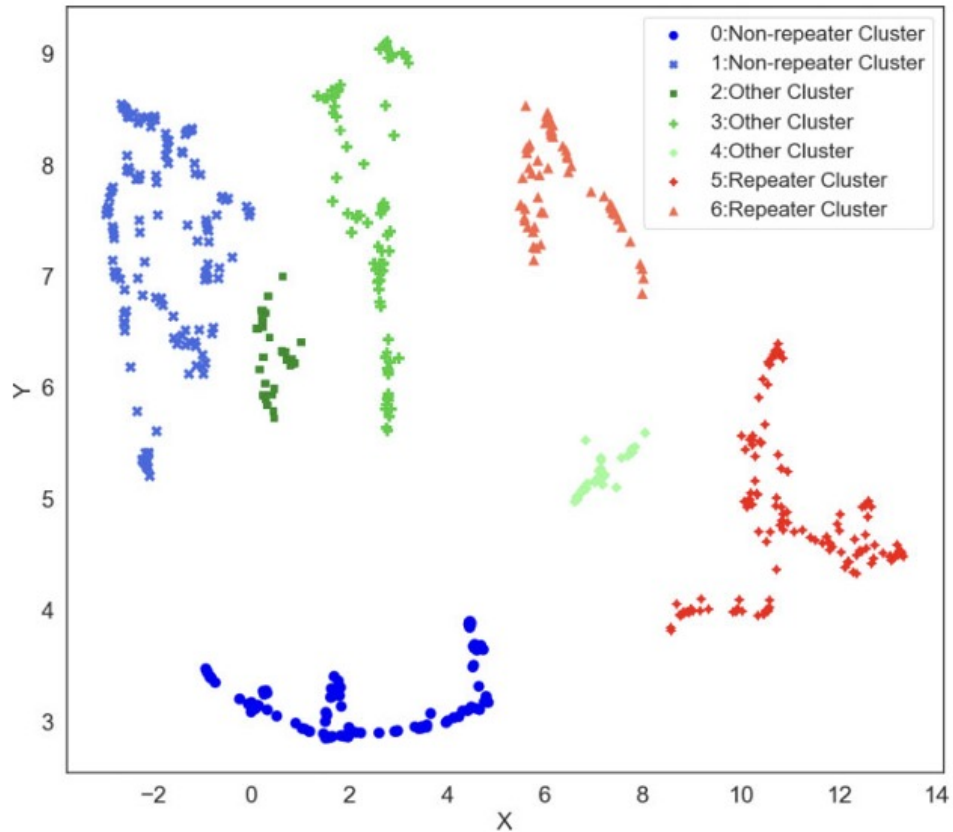
**Figure 7.** Clustering results of HDBSCAN in the t-SNE plane. Non-repeater clusters are marked in blue, while repeater clusters are marked in red. Other clusters are marked in green.

**Figure 9.** Feature correlation of PCA +  $k$ -means. The left bar correlates with the  $x$ -axis, while the right bar is related to the  $y$ -axis. Boxcar width, redshift, energy, luminosity, and brightness temperature are the most important features in PCA.

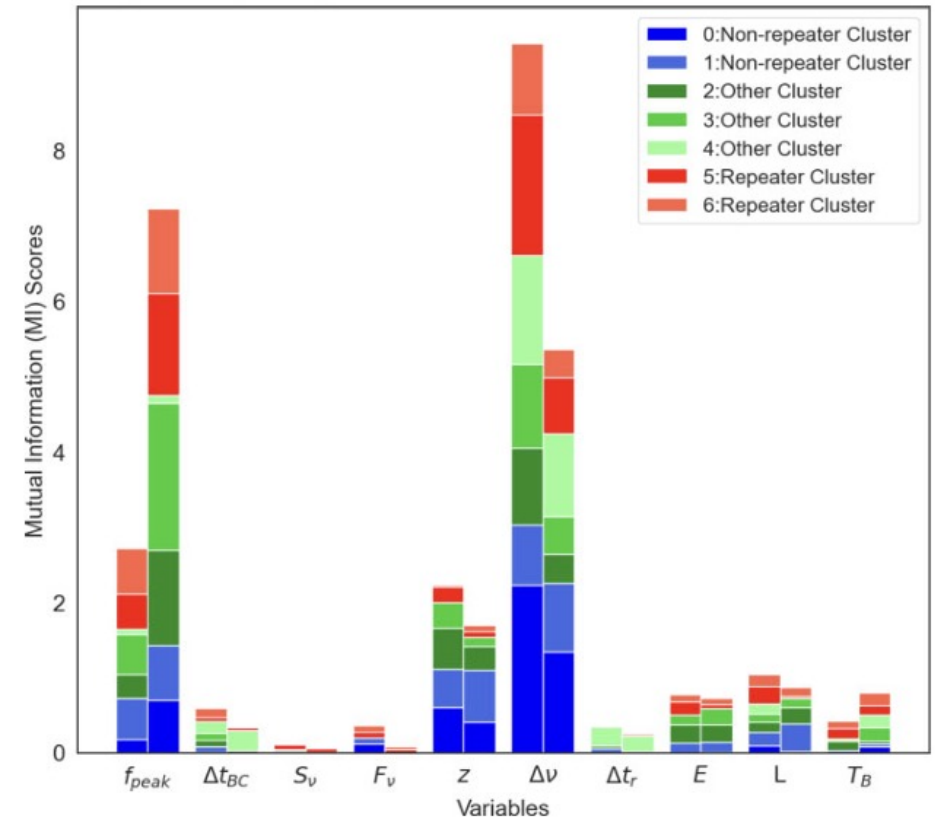


**Figure 10.** Feature correlation of t-SNE + HDBSCAN. For non-repeater clusters in blue, peak frequency, redshift, and rest-frame frequency width are the most dominant features, while redshift is not as significant in the repeater clusters coloured red. On the whole, peak frequency, redshift, and rest-frame frequency width are the most important features.

# UMAP + HDBSCAN Clustering

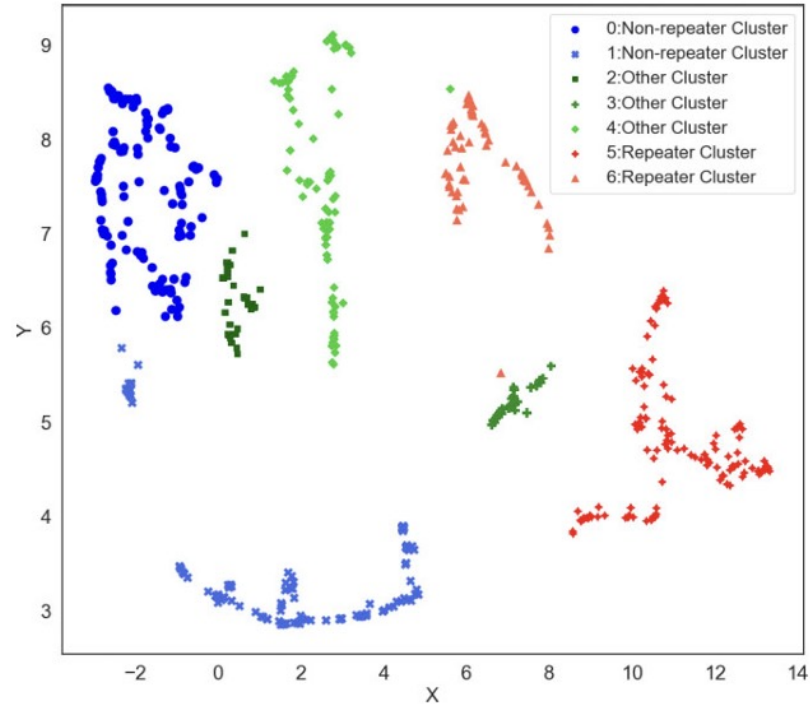


**Figure 8.** Clustering results of HDBSCAN in the UMAP plane. Non-repeater clusters are marked in blue, while repeater clusters are marked in red. Other clusters are marked in green.



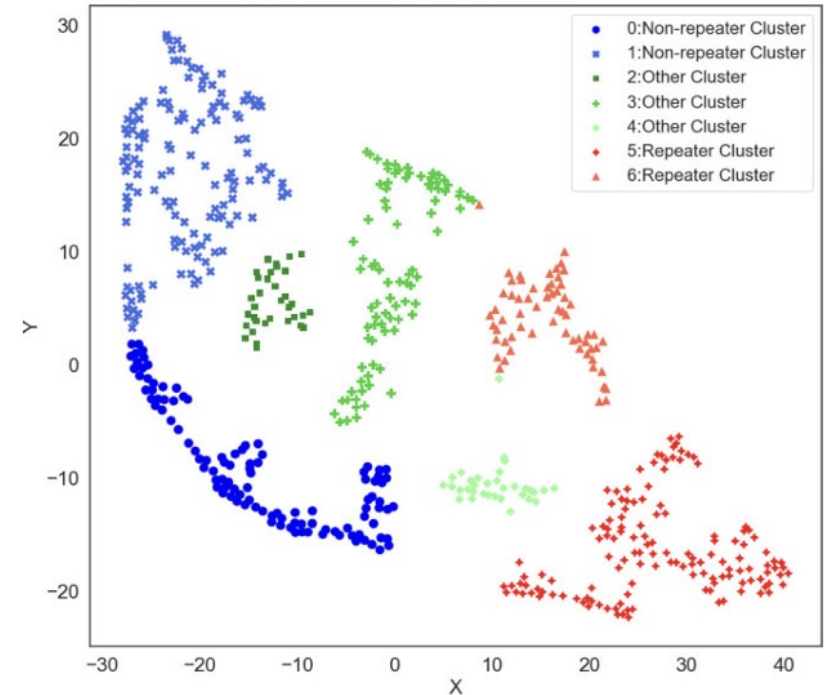
**Figure 11.** Features correlation of UMAP + HDBSCAN. For non-repeater clusters in blue, peak frequency, redshift, and rest-frame frequency width are the most dominant features, while redshift is not as significant in the repeater clusters coloured red. On the whole, peak frequency, redshift, and rest-frame frequency width are the most important features.

# Comparison between different Dimensionality Reduction before Clustering



**Figure 12.** Data points in UMAP plane coloured with the cluster labels identified by t-SNE and HDBSCAN. Only two points are in the different series of clusters.

Coordinates: UMAP Dimensionality Reduction  
Color: t-SNE DR + HDBSCAN Clustering



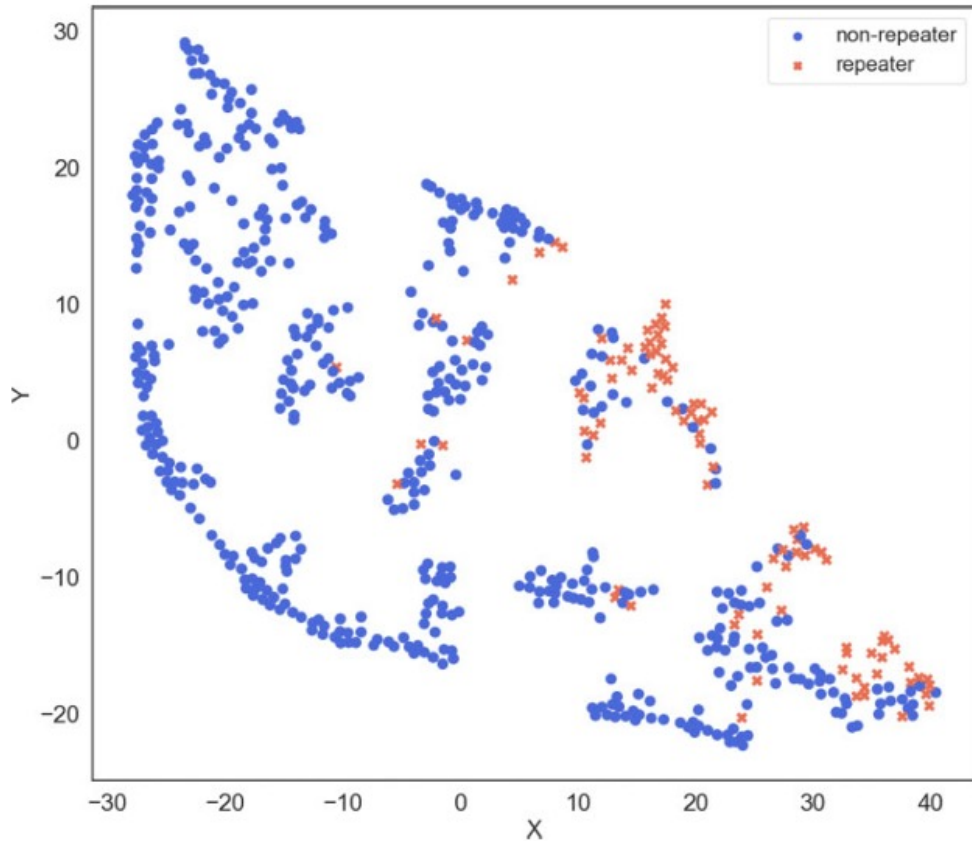
**Figure 13.** Data points in t-SNE plane coloured with the cluster labels identified by UMAP and HDBSCAN. Only two points are in the different series of clusters.

Coordinates: t-SNE DR  
Color: UMAP DR + HDBSCAN Clustering

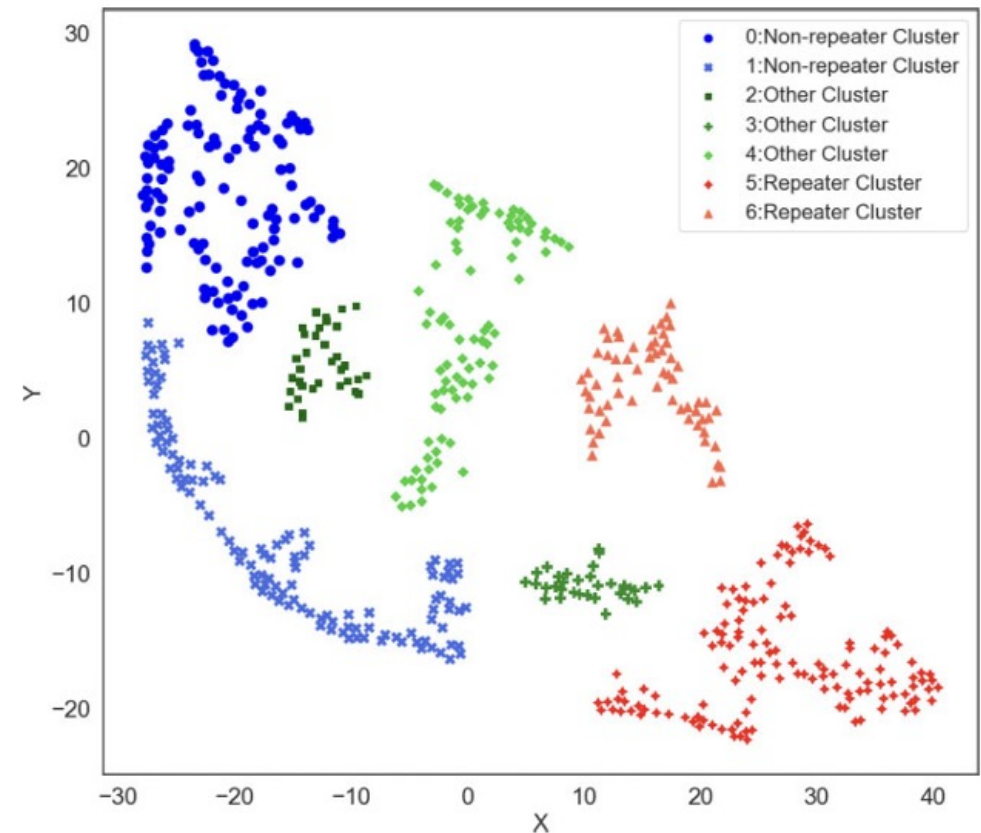
# Repeater Clusters

- The clusters that only include non-repeaters are recognized as 'non-repeater clusters'
- The clusters containing more than a number (they defined 15% as the criterion) of repeaters are classified as 'repeater clusters. Non-repeaters in repeater clusters can then be treated as hidden repeaters.
- Clusters that only have a few repeaters and the ratio is not enough for 15%, are labelled as 'other clusters.

# Comparison between Observation and Clustering Results (t-SNE + HDBSCAN)

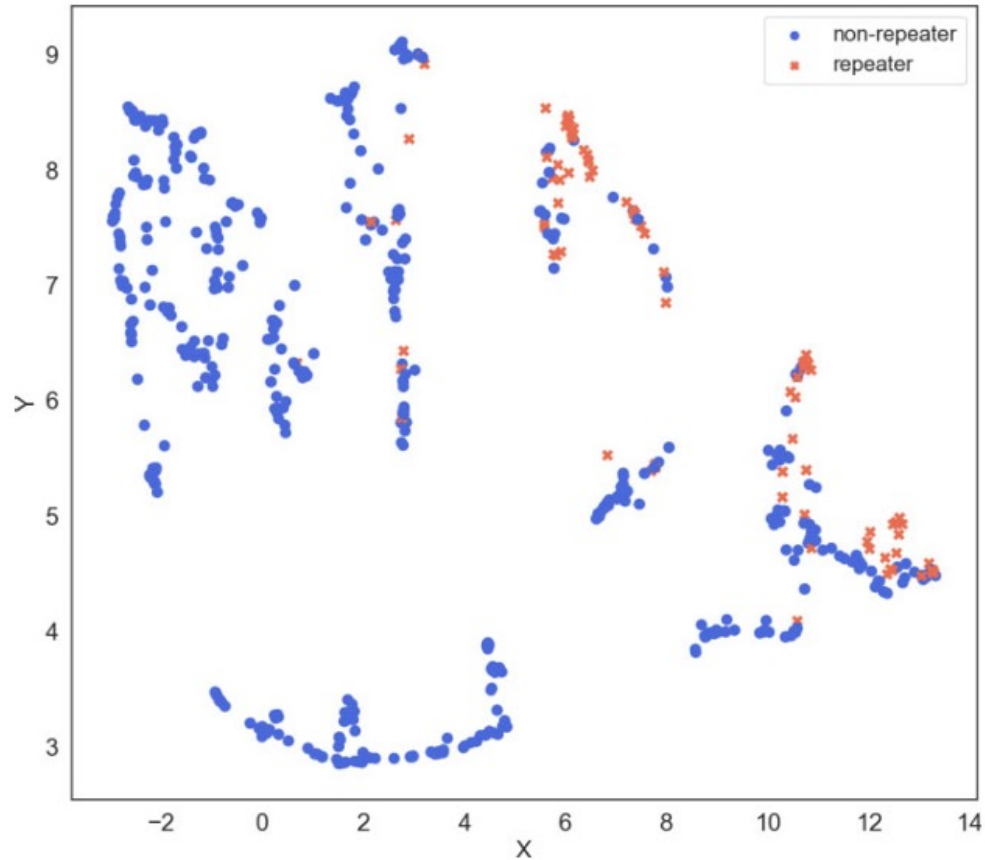


**Figure 3.** Dimensionality reduction results from t-SNE. Most repeating FRBs are on the right-hand side and generally reside in the same clusters.

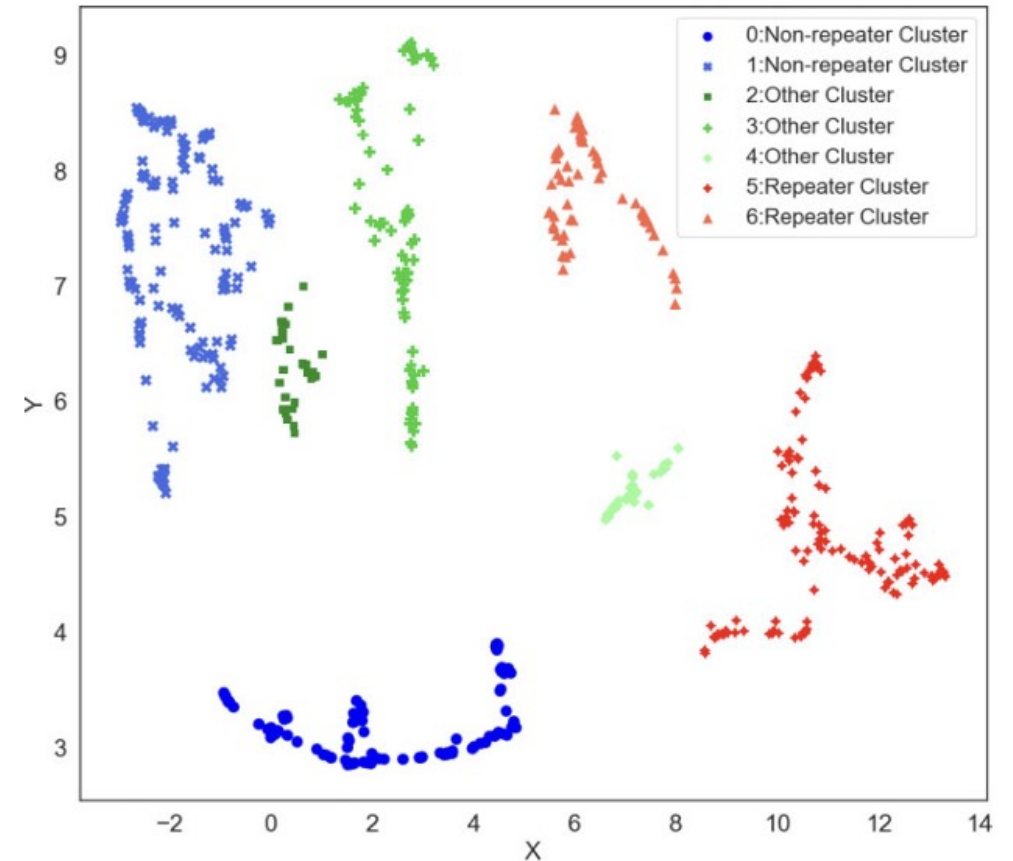


**Figure 7.** Clustering results of HDBSCAN in the t-SNE plane. Non-repeater clusters are marked in blue, while repeater clusters are marked in red. Other clusters are marked in green.

# Comparison between Observation and Clustering Results (UMAP + HDBSCAN)



**Figure 4.** Dimensionality reduction results from UMAP. Similar to t-SNE, most repeating FRBs are on the right-hand side and generally reside in the same clusters.



**Figure 8.** Clustering results of HDBSCAN in the UMAP plane. Non-repeater clusters are marked in blue, while repeater clusters are marked in red. Other clusters are marked in green.

# Conclusion

- They utilized unsupervised machine learning to learn the features of the FRBs in the first CHIME/FRB catalogue and attempt to reveal their hidden properties.
- Repeaters and non-repeaters have different distributions in many features.
- Unsupervised machine learning algorithms, without the input of the observed repeatability of FRBs, can classify FRBs into clusters that have high correspondence with repeaters and non-repeaters. This suggests that repeaters and non-repeaters indeed belong to different categories.

# Conclusion

- In addition to spotting the two most significant categories of repeaters and non-repeaters, unsupervised learning methods also imply that there may exist five to seven subspecies based on their traits.
- Learning from multiple parameters, unsupervised machine learning identifies some hidden repeaters from the apparent non-repeaters.



Thank you