

Revised classification of the CHIME fast radio bursts with machine learning*

Liang Liu (刘亮)¹ Hai-Nan Lin (林海南)^{2†} Li Tang (唐丽)¹ 

¹School of Physics and Electronic Information, Mianyang Teachers' College, Mianyang 621000, China

²Department of Physics, Chongqing University, Chongqing 401331, China

Abstract: Fast radio bursts (FRBs) are short-duration and energetic radio transients of unknown origin. Observationally, they are commonly categorized into repeaters and non-repeaters. However, this binary classification may be influenced by observational limitations such as sensitivity and time coverage of telescopes. In this study, we employ unsupervised machine learning techniques to re-examine the CHIME/FRB catalog, with the goal of identifying intrinsic groupings in the FRB population without relying on preassigned labels. Using t-distributed stochastic neighbor embedding (t-SNE) for dimensionality reduction and hierarchical density-based spatial clustering of applications with noise (HDBSCAN) for clustering, we find that the FRB sample naturally separates into two major clusters. One cluster contains nearly all known repeaters but is contaminated by some apparently non-repeaters, while the other cluster is dominated by non-repeaters. This suggests that certain FRBs previously labeled as non-repeaters may share intrinsic similarities with repeaters. Mutual information analysis reveals that rest-frame frequency width and peak frequency are the most informative features governing the clustering structure. Even when reducing the input space to just these two features, the classification remains robust.

Keywords: fast radio bursts, machine learning, unsupervised clustering

DOI: 10.1088/1674-1137/ae0725 **CSTR:** 32044.14.ChinesePhysicsC.50015102

I. INTRODUCTION

Fast radio bursts (FRBs) are millisecond-duration bursts of radio waves, first identified in 2007 from the archival data of the Parkes radio telescope [1]. To date, more than 1000 FRB sources have been detected by telescopes worldwide [2], with publicly available data repositories providing extensive catalogs¹⁾. These enigmatic transients have emerged as powerful tools for probing astrophysical and cosmological phenomena, yet their origins remain unresolved. The majority of FRBs exhibit high dispersion measures (DMs) that exceed the expected contribution from the Milky Way, indicating extragalactic or even cosmological distances [3, 4]. Their observed properties, along with polarization and spectral measurements, offer crucial insights into the physical mechanisms underlying their emission. In particular, the discovery of repeating FRBs, most notably FRB 121102 [5, 6], has provided key constraints on progenitor models, suggesting that at least a subset of FRBs arise from non-cataclysmic sources. However, despite the numerous theoretical models proposed to explain their emission mechanisms, a unifying theory capable of accounting for the full diversity of FRB phenomena remains elusive [7, 8].

FRBs have traditionally been categorized into two primary categories — repeating and non-repeating — based on the number of detected bursts. Distinct observational characteristics between these groups suggest the possibility of heterogeneous origins. Among repeating FRBs, only a small fraction exhibit high burst rates [9–11], while the majority display limited activity. It remains plausible that some FRBs classified as non-repeaters are in fact repeaters whose bursts have not been observed yet due to insufficient monitoring or sensitivity limitations, raising concerns about potential misclassification. Such uncertainties complicate efforts to constrain progenitor models and decipher the underlying emission mechanisms. Furthermore, emerging evidence suggests that FRBs may not be strictly dichotomous, with additional subpopulations potentially existing [12–16]. Therefore, a more refined classification scheme is essential for resolving their origins, improving theoretical modeling, and advancing our understanding of their astrophysical nature.

Although previous studies have reported significant differences in the parameter distributions of repeating and non-repeating FRBs, conventional comparative analyses have largely been confined to individual parameters or

Received 12 May 2025; Accepted 4 September 2025; Accepted manuscript online 5 September 2025

* Supported by the National Natural Science Fund of China (12275034, 12347101) and the Natural Science Fund of Chongqing (CSTB2022NSCQ-MSX0357)

† E-mail: linhn@cqu.edu.cn

1) www.wis-tns.org

©2026 Chinese Physical Society and the Institute of High Energy Physics of the Chinese Academy of Sciences and the Institute of Modern Physics of the Chinese Academy of Sciences and IOP Publishing Ltd. All rights, including for text and data mining, AI training, and similar technologies, are reserved.

low-dimensional projections [9, 17–20]. However, such approaches are inherently limited in their ability to capture the complex, potentially high-dimensional correlations that may underpin the observed diversity of FRB properties. To address these limitations, recent efforts have increasingly turned to machine learning [12–16, 21–23] — a branch of artificial intelligence well-suited to extracting hidden patterns from large, multi-dimensional datasets. Machine learning techniques have been widely applied in cosmological and astronomical research and can be broadly classified into two categories: supervised and unsupervised learning, distinguished by whether or not the data are labeled. In supervised learning, models are trained on labeled datasets to infer explicit mappings between inputs and target outputs, allowing prior knowledge to guide predictive accuracy. In contrast, unsupervised learning operates without labeled data, seeking to uncover intrinsic structures, patterns, or clusters within the data.

Machine learning techniques have demonstrated considerable promise in the classification of FRBs, although the resulting classifications can vary depending on the choice of methods and input parameters [12–16, 24]. Employing a range of supervised learning algorithms, Luo *et al.* [15] analyzed FRBs from the first Canadian Hydrogen Intensity Mapping Experiment Fast Radio Burst (CHIME/FRB) catalog, dividing the data into training and testing sets to evaluate model performance. Their approach enabled the identification of several candidate repeaters within the non-repeating FRB population. In contrast, Chen *et al.* [12] applied an unsupervised learning technique — Uniform Manifold Approximation and Projection (UMAP) — to the same dataset, achieving a repeater completeness of 95% and identifying 188 candidate repeaters from 474 non-repeating sources. Motivated by concerns that the sensitivity of Chen *et al.*'s method may be affected by the choice of the hyperparameter $n_neighbors$, Zhu-Ge *et al.* [14] revisited the CHIME/FRB catalog using UMAP alongside two additional dimensionality-reduction techniques: principal component analysis (PCA) and t-distributed stochastic neighbour embedding (t-SNE). Their analysis revealed the presence of not only the two conventional classes of repeating and non-repeating FRBs but also three additional subgroups, suggesting a more complex underlying classification structure. By applying an unsupervised decision tree algorithm to the first CHIME/FRB catalog, Luo *et al.* [24] observed a reduced distinction between repeaters and non-repeaters upon inclusion of the CHIME/FRB 2023 catalog.

Previous classification efforts have typically treated each burst — whether a sub-burst from a non-repeating FRB or an individual burst from a repeating source — as an independent event. However, bursts originating from the same FRB source are likely to share intrinsic phys-

ical correlations. Ignoring these correlations and treating bursts as independent during model training risks introducing redundant information, which may bias the learning process towards overfitting and hinder the model's ability to capture the global properties of FRB sources. To mitigate these issues and improve the robustness of the classification, we employ unsupervised machine learning methods, including dimensionality reduction and clustering, to analyze the CHIME/FRB catalogs, retaining only a single representative burst from each source. Specifically, we only preserve the first detected burst from each repeating source and the first sub-burst from each non-repeating source while neglecting all subsequent (sub-)bursts from the same source. Our approach to repeating FRBs is aligned with that proposed by Zhong *et al.* [17].

The remainder of this paper is structured as follows. In Section II, we describe the datasets used in this study and outline the unsupervised machine learning techniques adopted for dimensionality reduction and clustering. In Section III, we present the main results of our analysis. Finally, discussions and conclusions are provided in Section IV.

II. DATA AND METHODOLOGY

A. Data selection

The dataset utilized in this study is primarily derived from the first CHIME/FRB catalog [20] and the CHIME/FRB 2023 catalog [25]. The first CHIME/FRB catalog contains 536 bursts detected between 400 and 800 MHz during the period from July 25, 2018 to July 1, 2019. Of those, 474 events originate from apparently non-repeating sources and 62 events are associated with 18 known repeating sources. Among the non-repeating bursts, 64 exhibit sub-burst structures. Six FRBs exhibiting zero values in both fluence and flux — FRB-20190307A, FRB20190307B, FRB20190329B, FRB20190329C, FRB20190531A, and FRB20190531B — are removed from the sample to ensure robustness in the subsequent analysis. Therefore, the first CHIME/FRB catalog contains $474+18-6=486$ available independent FRB sources. The CHIME/FRB 2023 repeater catalog comprises 25 repeating FRB sources detected between September 30, 2019 and May 1, 2021, including 6 sources that were previously classified as non-repeaters in the first CHIME/FRB catalog. To mitigate bias from intrinsic burst correlations within sources, we retain only the chronologically earliest burst per repeating source and first sub-burst per non-repeating source while excluding all subsequent (sub-)bursts from the same source. This ensures that exactly one event per source is utilized. The resultant dataset contains 505 independent bursts, among which 43 originate from repeating sources and 462 from non-repeating sources.

In our unsupervised machine learning framework, ten features are extracted either directly from the observed properties reported in the CHIME/FRB catalog or derived through standard calculations. These features include peak frequency, flux density, fluence, boxcar width, redshift, rest-frame frequency width, burst energy, luminosity, and brightness temperature. The distributions of these features are shown in Fig. 1. Detailed definitions and descriptions of each feature are provided below.

- Peak frequency ν_p (MHz). The peak frequency is defined as the frequency at which the sub-burst reaches its maximum flux density.

- Flux S_ν (Jy). The flux density reported in the catalog corresponds to the peak flux of the band-averaged profile and represents a lower-limit estimate. Logarithmic values are used in the analysis.

- Fluence F_ν (Jy ms). The fluence is defined as the time-integrated flux density of the burst, as provided in

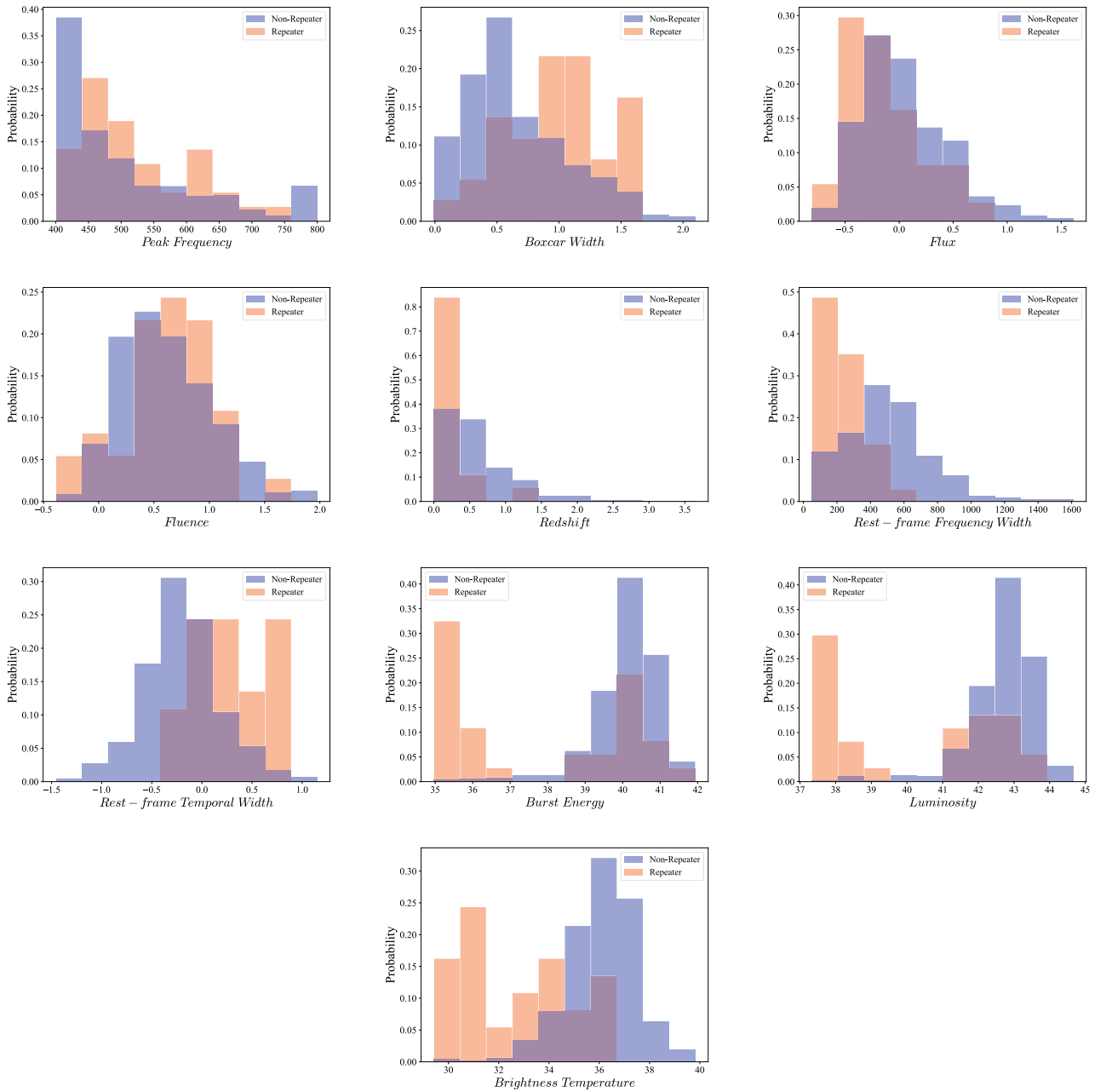


Fig. 1. (color online) Histograms of ten features that are used in the classification. The vertical axis is normalized such that the total probability is unity.

the CHIME/FRB catalog. Logarithmic values are adopted.

- Boxcar width Δt_{BC} (ms). The burst duration is characterized by the width of the boxcar function after convolution, as defined in the catalog. Logarithmic values are used in the analysis.

- Redshift z . Most FRBs in our sample lack direct redshift measurements, with the exception of two sources — FRB20121102A ($z = 0.19273$) [26] and FRB2018-0916B ($z = 0.0337$) [27] — for which spectroscopic redshifts are available via host galaxy associations. For the remaining FRBs, redshifts are estimated using the empirical relation between dispersion measure (DM) and redshift, which will be described in detail below.

- Rest-frame frequency width $\Delta\nu$ (MHz). The rest-frame frequency width characterizes the intrinsic spectral extent of each burst and is calculated as the redshift-corrected difference between the maximum and minimum observed frequencies,

$$\Delta\nu = (\nu_{\text{max}} - \nu_{\text{min}})(1 + z). \quad (1)$$

- Rest-frame temporal width Δt_r (ms). The rest-frame temporal width of each sub-burst is determined using fit-burst Δt [20], corrected for cosmological time dilation,

$$\Delta t_r = \frac{\Delta t}{1 + z}. \quad (2)$$

Logarithmic values of these rest-frame temporal widths are used in the analysis.

- Burst energy E (erg). The burst energy of each FRB is estimated as

$$E = \frac{4\pi D_L^2}{1 + z} F_\nu \nu_p, \quad (3)$$

where F_ν is the specific fluence, D_L is the luminosity distance, and ν_p is the observed peak frequency of the burst. Logarithmic values of the inferred energies are adopted for subsequent analysis.

- Luminosity L (erg s⁻¹). The luminosity of FRBs is estimated as

$$L = 4\pi D_L^2 S_\nu \nu_p, \quad (4)$$

where S_ν denotes the specific peak flux density. Logarithmic values of the derived luminosities are adopted in

the analysis.

- Brightness temperature T_B (K). The brightness of a source is characterized by its brightness temperature, defined as the temperature of a blackbody emitting the same specific intensity. Accounting for cosmological effects, T_B is calculated as [15]

$$T_B = \frac{S_\nu D_A^2}{2\pi k_B (\nu_p \Delta t)^2} (1 + z)^3, \quad (5)$$

where k_B is the Boltzmann constant, Δt is the observed duration, and D_A is the angular diameter distance. Logarithmic values are used in the analysis.

To infer the redshift of unlocalized FRBs from the DM, we decompose the observed DM of an extragalactic FRB into four components, as is convention [28–30]:

$$\text{DM} = \text{DM}_{\text{MW}} + \text{DM}_{\text{halo}} + \text{DM}_{\text{IGM}} + \frac{\text{DM}_{\text{host}}}{1 + z}, \quad (6)$$

where the four terms on the right-hand-side denote the contributions from the Milky Way interstellar medium, galactic halo, intergalactic medium, and host galaxy, respectively. The Milky Way contribution, DM_{MW} , is estimated using the NE2001 electron density model based on pulsar observations [31]. Following the work of Zhu-Ge et al. [14], the galactic halo and host galaxy terms are fixed to $\text{DM}_{\text{halo}} = 30 \text{ pc cm}^{-3}$ and $\text{DM}_{\text{host}} = 70 \text{ pc cm}^{-3}$, respectively. The contribution from the intergalactic medium is computed within a flat Λ CDM cosmology as

$$\text{DM}_{\text{IGM}}(z) = \frac{21cH_0\Omega_b f_{\text{IGM}}}{64\pi G m_p} \int_0^z \frac{1 + z}{\sqrt{\Omega_m(1 + z)^3 + \Omega_\Lambda}} dz, \quad (7)$$

where c is the speed of light, G is Newton's gravitational constant, m_p is the proton mass, and $f_{\text{IGM}} = 0.83$ is the baryon fraction of the IGM [32]. The cosmological parameters are set to the Planck 2018 values: $H_0 = 67.4 \text{ km s}^{-1} \text{ Mpc}^{-1}$, $\Omega_m = 0.315$, $\Omega_\Lambda = 0.685$, and $\Omega_b = 0.0493$ [33]. Redshifts are then inferred by inverting the relation above given the observed DM. For sources located near the Milky Way, the resulting redshift can approach zero or become formally negative. In such cases, a minimum redshift cutoff of $z = 0.002248$ — corresponding to a luminosity distance of 10 Mpc — is adopted.

B. Dimensionality reduction

While high-precision observations of FRBs offer unprecedented opportunities to probe their enigmatic nature, the inherently high dimensionality of these datasets presents considerable challenges for both classification and physical interpretation. To address this, dimensionality reduction — a process that projects data from a high-

dimensional space onto a more tractable low-dimensional manifold — has become essential within the field of machine learning [34]. Dimensionality reduction provides a reasonable visualization in a low-dimensional plane to analyze the intrinsic connections of the data. However, conventional linear techniques often fail to preserve the proximity of similar data points in the reduced space, a limitation particularly problematic when the data reside on non-linear, low-dimensional manifolds embedded in higher dimensions. To overcome this, a suite of non-linear dimensionality reduction methods has been developed, with an emphasis on preserving local structure. Among these, Stochastic Neighbor Embedding (SNE) has gained prominence for its capacity to generate meaningful low-dimensional representations [35]. SNE achieves this by converting high-dimensional Euclidean distances into conditional probabilities that quantify pairwise similarity under a Gaussian kernel and subsequently optimizing a cost function — formulated as the Kullback–Leibler divergence between the high- and low-dimensional distributions.

However, like many manifold learning algorithms, SNE suffers from the so-called "crowding problem." In high-dimensional spaces, a data point may have numerous equidistant neighbours; yet, when projected into a low-dimensional space, there is insufficient room to accommodate all of these neighbours without artificially compressing them or causing overlaps. This spatial constraint distorts the true pairwise relationships, often resulting in a congested central region in the visualization. To mitigate this issue, the t-distributed stochastic neighbour embedding (t-SNE) algorithm replaces the Gaussian kernel used in the low-dimensional space with a student's t-distribution [36–37]. Owing to its heavier tails, the t-distribution allows for moderate distances in the high-dimensional space to be mapped to larger separations in the low-dimensional embedding. This adjustment effectively alleviates the crowding effect, preserving the relative topology of the data more faithfully.

In this study, we project the input data — FRBs characterized by ten features — onto a two-dimensional manifold and accordingly set the `n_components` hyperparameter of the t-SNE algorithm to 2. Among the various tunable parameters in t-SNE, the most critical is `perplexity`, which acts as a smooth measure of the effective number of local neighbours considered during the estimation of similarity probabilities in the high-dimensional space. Conceptually, `perplexity` governs the trade-off between capturing local structure and preserving global relationships: lower values accentuate fine-grained clustering, while higher values incorporate broader context at the potential cost of merging distinct group boundaries. Following the heuristic proposed by Oskolkov [38], `perplexity` is typically chosen to scale with the square root of the dataset size ($N = 505$ in our

case). Therefore, we set `perplexity` = $\sqrt{N} = 22$. A summary of all hyperparameter settings employed in the t-SNE dimensionality reduction is given in Table 1.

C. Clustering method

Following dimensionality reduction, we apply clustering techniques to analyze the resulting low-dimensional representations. Clustering, a core class of unsupervised learning methods, enables the identification of intrinsic structure in unlabeled data by partitioning it into distinct groups, or clusters, such that intra-cluster similarity is maximized while inter-cluster similarity is minimized. Unsupervised clustering automatically classifies the data without the need to preset the number of classes. Among the variety of clustering algorithms, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) offers distinct advantages in handling complex and noisy datasets [39]. DBSCAN defines clusters as contiguous regions of high point density, separated by regions of lower density, and is therefore capable of discovering clusters of arbitrary shape. As a result, DBSCAN is particularly well-suited for analyzing the potentially heterogeneous and irregular structures found in FRB parameter space.

The efficacy of DBSCAN in uncovering meaningful structure within complex datasets hinges on two key hyperparameters: neighbourhood radius, ϵ , and minimum number of points, `minPts`, required to define a dense region. The parameter ϵ sets the spatial scale at which local density is evaluated, determining the maximum distance within which two points are considered neighbours. The parameter `minPts` specifies the minimum number of

Table 1. List of t-SNE hyperparameters.

Name	Value
<code>n_components</code>	2
<code>perplexity</code>	22
<code>early_exaggeration</code>	2
<code>learning_rate</code>	'auto'
<code>n_iter</code>	1000
<code>n_iter_without_progress</code>	300
<code>min_grad_norm</code>	1e-7
<code>metric</code>	'euclidean'
<code>metric_params</code>	None
<code>init</code>	'random'
<code>verbose</code>	0
<code>random_state</code>	22
<code>method</code>	'barnes hut'
<code>angle</code>	0.5
<code>n_jobs</code>	None
<code>square_distances</code>	'deprecated'

data points, including the point itself, that must reside within the ϵ -neighbourhood for a region to be considered dense. Together, these parameters define the local density criteria that underpin both cluster formation and the identification of noise. Based on these criteria, each data point is classified into one of three categories:

- **Core point:** A point that has at least `minPts` points (including itself) within its ϵ -neighbourhood. Core points reside in high-density regions and form the structural backbone of clusters.
- **Border point:** A point that falls within the ϵ -neighbourhood of a core point but does not itself meet the `minPts` threshold to be considered a core point. These points lie at the periphery of dense regions.
- **Noise point (or outlier):** A point that is neither a core point nor a border point. Such points do not fall within the ϵ -neighbourhood of any core point and lie in low-density, unclustered regions.

Owing to DBSCAN's sensitivity to variations in local density, its applicability can be limited in real-world scenarios, where data are often heterogeneous and contaminated by noise. To address these challenges, Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) extends DBSCAN by integrating hierarchical density estimation with a robust cluster extraction framework [40–42]. Unlike DBSCAN, which relies on a fixed global density threshold, HDBSCAN constructs a condensed cluster tree from the mutual reachability graph of the data, capturing a nested hierarchy of clusters across varying density levels. This multi-scale approach enables the identification of meaningful structures in data with non-uniform density. In light of these advantages, we adopt HDBSCAN to analyze the population structure of FRBs. To determine the optimal HDBSCAN hyperparameters, we employ a grid search approach to maximize classification performance. The hyperparameter settings used in our study are summarized in Table 2.

III. RESULTS AND ANALYSIS

A. Results of the full features

We first apply t-SNE to a ten-dimensional parameter space encompassing both observed and derived properties of repeating and non-repeating sources. The resulting two-dimensional projection is shown in the left panel of Fig. 2, where non-repeaters (orange) and repeaters (green) occupy distinct regions of the embedded space. Notably, 25 newly identified repeaters from the CHIME/FRB 2023 repeater catalog (including 6 FRBs

Table 2. List of HDBSCAN hyperparameters.

Name	Value
<code>min_cluster_size</code>	150
<code>min_samples</code>	2
<code>cluster_selection_epsilon</code>	0
<code>max_cluster_size</code>	0
<code>metric</code>	'euclidean'
<code>alpha</code>	1
<code>p</code>	None
<code>algorithm</code>	'best'
<code>leaf_size</code>	40
<code>memory</code>	Memory(cachedir = None, verbose = 0)
<code>approx_min_span_tree</code>	True
<code>gen_min_span_tree</code>	False
<code>core_dist_n_jobs</code>	4
<code>cluster_selection_method</code>	'eom'
<code>allow_single_cluster</code>	False
<code>prediction_data</code>	False
<code>match_reference_implementation</code>	False

misclassified as non-repeaters in the first CHIME/FRB catalog) are highlighted as red stars. Most known repeaters cluster in the lower quadrant of the embedding, with the exception of FRB20181017A and FRB20180910A, while non-repeating bursts form several well-defined and spatially separated groupings. The presence of non-repeaters within or near repeaters suggests possible misclassification.

The HDBSCAN algorithm is applied to the two-dimensional embedding to examine the latent structure uncovered by the t-SNE projection. The clustering results, presented in the right panel of Fig. 2, reveal a robust bifurcation of the FRB sample into two principal groups. The first cluster, shown in orange, is dominated by non-repeating FRBs and is hereafter referred to as the "non-repeater cluster." Notably, this group also includes the previously known repeater FRB20181017A and recently confirmed repeater FRB20180910A, both of which are misclassified within the non-repeater population — an anomaly that will be addressed in detail in Section III.B. The second cluster, shown in green, encompasses all remaining repeaters in the sample, and is accordingly designated as the "repeater cluster." Intriguingly, five out of six newly identified repeating sources (except for FRB20180910A), which are originally labeled as non-repeaters in the CHIME/FRB catalog, coincide spatially with this repeater cluster. This spatial congruence provides compelling evidence that certain sources previously considered non-repeating may in fact possess intrinsic properties akin to known repeaters. The "repeater

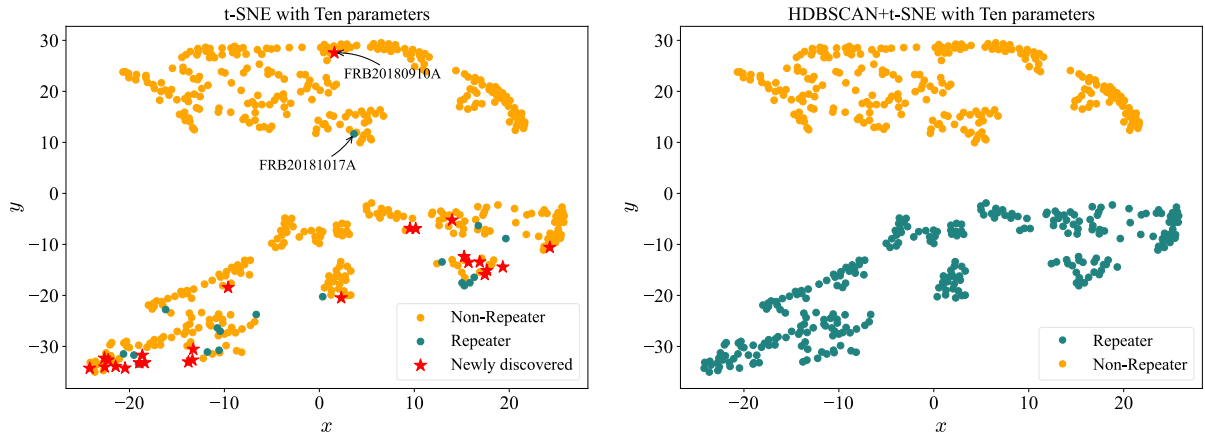


Fig. 2. (color online) Dimensionality reduction and clustering results with ten features of FRBs.

cluster" contains 41 confirmed repeaters and 230 candidate repeaters, which implies that more than half of FRB sources may be intrinsic repeaters. These results support the hypothesis that repetition may be a function of observational limitations rather than a fundamental dichotomy in progenitor types. As such, non-repeating FRBs situated within the repeater cluster can be treated as repeater candidates, warranting further targeted follow-up.

To identify the key physical parameters underpinning the unsupervised clustering of FRBs, we quantify the dependency between each of the ten features and the cluster structure derived from the t-SNE embedding followed by HDBSCAN segmentation. This is achieved through the computation of mutual information (MI) [43–44], a non-negative measure of statistical dependence between two variables. By definition, MI vanishes when variables are independent and increases with stronger dependency, making it a robust and model-free metric for feature relevance. We employ the MI regression implementation in the scikit-learn library to estimate the mutual information between each feature and the two-dimensional t-SNE coordinates. For each identified cluster, MI scores are computed separately with respect to the two projection axes. The resulting MI distributions are shown in Fig. 3, where each feature is represented by a pair of bars corresponding to its relevance along the x - and y -axes of the embedding. For the non-repeater cluster, peak frequency, rest-frame frequency width, and redshift emerge as the most informative features. In contrast, the repeater cluster exhibits comparatively lower dependence on redshift. Overall, the MI analysis indicates that peak frequency and rest-frame frequency width are the two most influential features governing the overall geometry of the low-dimensional representation.

B. Results of the intrinsic features

The MI scores calculated in the ten-dimensional parameter space demonstrate that not all features exhibit significant contributions to the clustering structure of FRB

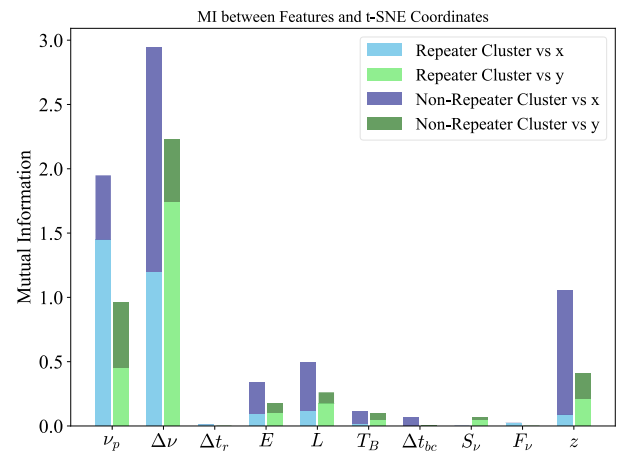


Fig. 3. (color online) Feature correlation of t-SNE + HDBSCAN with ten features.

populations. Notably, some features are non-intrinsic properties: redshift is a manifestation of the distance of FRB source, flux and fluence are inversely proportional to the square of distance, and boxcar width is an instrumental representation of burst duration that correlates with rest-frame temporal width. To optimize the feature space, we exclude these four non-intrinsic parameters exhibiting consistently low MI scores, thereby retaining six intrinsic features: peak frequency, rest-frame frequency width, rest-frame temporal width, burst energy, luminosity, and brightness temperature. This refined feature set is subsequently employed to recalculate the t-SNE dimensionality reduction and implement HDBSCAN clustering analysis, with all hyperparameters kept the same as in the ten-feature case. The updated results based on the six intrinsic features are shown in Fig. 4.

The two-dimensional projection of the refined six-dimensional parameter space preserves the overall topological separation identified in the original analysis. Repeaters (green) remain predominantly confined to the lower region of the embedding, whereas non-repeaters (orange) continue to cluster in the upper region. The

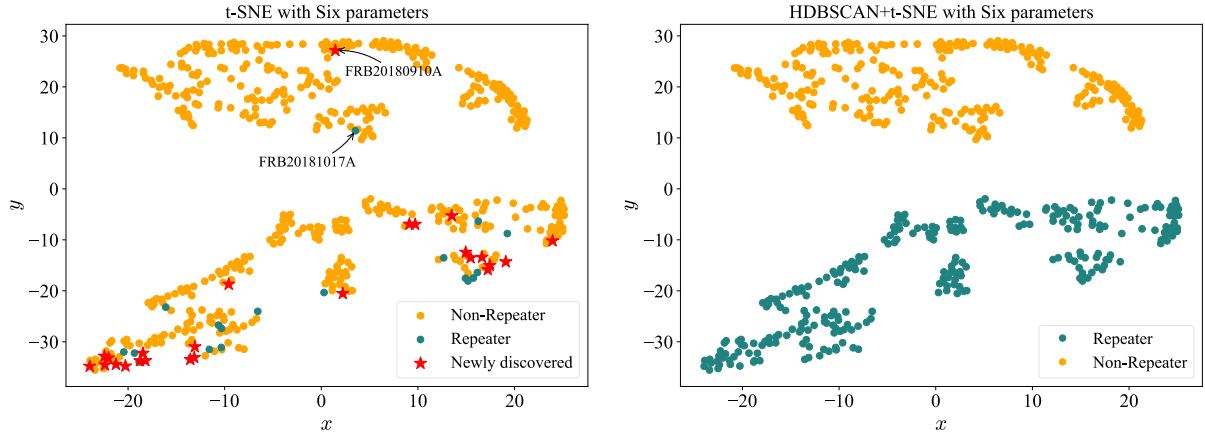


Fig. 4. (color online) Dimensionality reduction and clustering results with six features of FRBs.

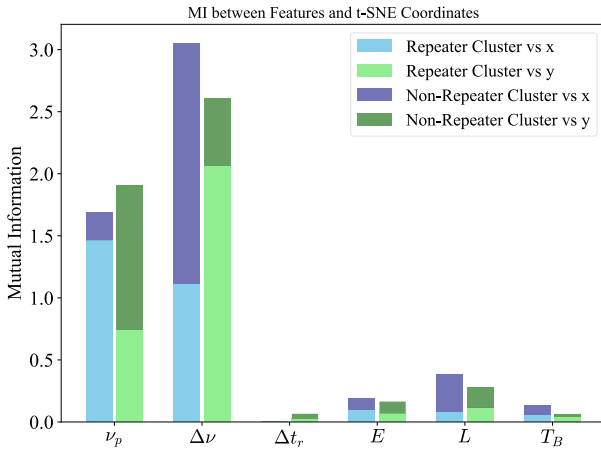


Fig. 5. (color online) Feature correlation of t-SNE + HDBSCAN with six features.

newly confirmed repeater FRB 20180910A and previously known repeater FRB 20181017A are again located within the non-repeater-dominated region, suggesting that these sources remain outliers irrespective of feature dimensionality. Application of HDBSCAN to the updated t-SNE embedding reproduces the previous segmentation, yielding two principal clusters with comparable spatial boundaries. This outcome confirms that the excluded features — redshift, flux, fluence, and boxcar width — play a minimal role in the latent separability of FRB populations. We further recompute the MI between feature values and t-SNE coordinates for both clusters in Fig. 5. It is found that the rest-frame frequency width and peak frequency continue to dominate, while the MI scores for the other features remain comparatively low.

Given the consistently highest MI scores for peak frequency and rest-frame frequency width, we restrict the input feature space to these two parameters. t-SNE and HDBSCAN are reapplied to this reduced set, with the resulting embeddings shown in Fig. 6. Remarkably, the fundamental separation between repeating and non-repeat-

ing FRBs is maintained: repeaters cluster in the lower quadrant, while non-repeaters dominate the upper. Despite the sharp dimensionality reduction, the core structural division observed in the ten- and six-feature analyses persists, suggesting robustness against feature pruning. Likewise, HDBSCAN continues to recover two principal populations, confirming the strong discriminatory power of peak frequency and rest-frame frequency width alone. A comparison of Figs. 2, 4, and 6 reveals that classification performance remains largely unchanged even when the input feature dimensionality is sharply reduced. This consistency stems from the dominant role of two key features: ν_p and $\Delta\nu$, as evidenced by their exceptionally high MI scores in Figs. 3 and 5. Note that Figs. 2, 4, and 6 are not identical, though the differences are minimal. The slight shift of each data point across the three feature sets (10, 6, and 2) is visually indistinguishable because of the large scales of the x and y axes in the embedding plane.

To facilitate direct interpretation of the classification boundary within the original feature space, we analyze the scatter plot of ν_p versus $\Delta\nu$ shown in Fig. 7, where a discernible separation between repeating and non-repeating FRBs emerges. The support vector machine (SVM) algorithm is implemented to construct the maximum-margin hyperplane that optimally separates the two classes through margin maximization. The resulting decision boundary, shown as a solid black line in Fig. 7, is given by $\Delta\nu = 0.95\nu_p + 1.30$. Support vectors — samples located on the dashed margin lines — govern the position of this boundary. It is worth mentioning that, in the ideal linearly separable scenario, all training samples are expected to lie outside the margin boundaries, such that the perpendicular distance from each sample to the decision hyperplane is at least one. That is, points residing within the margin region violate the optimality constraints: although some may remain correctly classified, they lack sufficient separation from the decision boundary, while others are misclassified. For instance, the previously known repeater FRB20181017A is misclassified by the

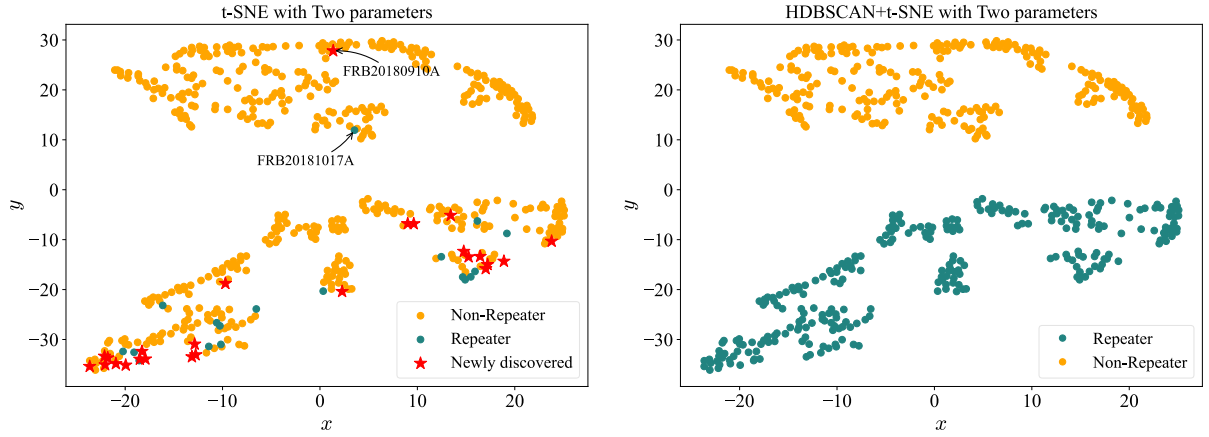


Fig. 6. (color online) Dimensionality reduction and clustering results with two features of FRBs.

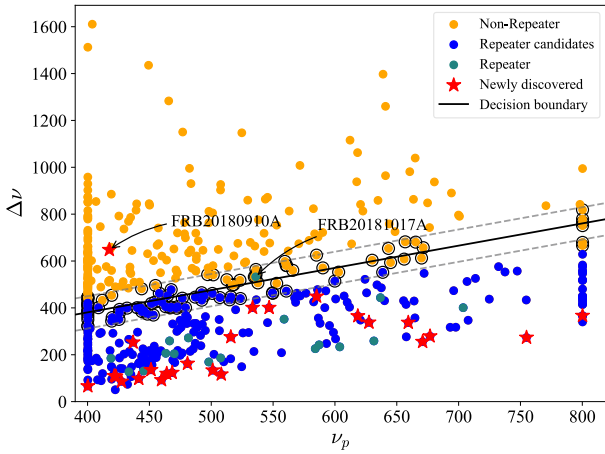


Fig. 7. (color online) 2D plane of peak frequency ν_p and rest-frame frequency width $\Delta\nu$.

model as a non-repeater, so it lies within the margin. Surprisingly, the newly confirmed repeater FRB20180910A, which shows a high $\Delta\nu$, falls outside the expected region and is also misclassified as a non-repeater. Compared with higher-dimensional embeddings using six or ten features, this two-feature model provides a more interpretable, physically motivated classification boundary with no appreciable loss in performance. These findings highlight peak frequency and rest-frame frequency width as the two key parameters encoding the physical differences between FRB populations.

C. Evaluating the model performance

In machine learning, precision and recall are two key metrics commonly used to evaluate classifier performance. Precision measures the proportion of predicted positive instances that are correctly identified, corresponding to the accuracy of positive predictions. Recall, by contrast, quantifies the fraction of true positive instances correctly identified among all actual positives and is also referred to as the sensitivity or true positive rate. In our study, the positive is repeaters, while the negative is non-

repeaters. Given that the classification of non-repeating FRBs may be affected by limited follow-up observations and the incomplete detection of repetition, maximizing recall is particularly important to ensure that potential repeaters are not inadvertently excluded. Precision, in this context, would overly penalize the model for predicting repeaters among sources yet to be confirmed, thereby introducing bias against true but observationally incomplete repeaters. Recall thus offers a more meaningful measure of classifier effectiveness for FRB population studies.

Recall is defined as the ratio of true positive (TP) to the sum of true positive (TP) and false negative (FN),

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad (8)$$

where TP denotes the number of correctly classified repeaters, and FN represents repeaters incorrectly assigned to the non-repeater category. As the classification outcome in the two-feature space remains consistent with that derived from the higher-dimensional feature sets (ten or six features), we compute recall based on the two-feature results. Our sample contains 43 repeaters (18 repeaters in the first CHIME/FRB catalog, and 25 newly discovered repeaters), two of which are misclassified as non-repeaters (FRB20180910A and FRB20181017A). The recall value is therefore $41/(41+2) = 0.95$, indicating that even with only two input features, the model successfully recovers the majority of known repeaters. We have confirmed that reducing the number of features from ten to two does not reduce recall. This performance underscores that the combination of peak frequency and rest-frame frequency width captures the essential physical distinctions between repeating and non-repeating FRBs.

IV. DISCUSSION AND CONCLUSIONS

Despite the increasing detection rate of FRBs, their

physical origins remain poorly understood. Current phenomenological classifications categorize FRBs into repeating and non-repeating populations based on burst recurrence. However, such distinctions are inherently limited by observational selection effects, notably instrumental sensitivity thresholds and finite monitoring durations, which may lead to misclassification. Addressing this challenge is crucial for advancing progenitor mechanism studies and developing accurate theoretical models. Traditional analytical methods struggle to handle the high-dimensional and heterogeneous nature of FRB datasets, prompting recent adoption of machine learning techniques, particularly unsupervised approaches, to uncover latent patterns in these complex data. A critical limitation of prior studies lies in their treatment of individual bursts (whether from repeaters or non-repeaters) as independent events, an approach that introduces data redundancy and risks biasing classification outcomes.

In this study, we reanalyzed the first CHIME/FRB catalog and the CHIME/FRB 2023 repeater catalog through an unsupervised machine learning framework optimized for capturing intrinsic FRB population characteristics. To mitigate bias from intrinsic burst correlations within sources, only one burst from each source was utilized. First, the t-SNE algorithm was employed to reduce the dimensionality of the FRB parameter space onto a two-dimensional manifold. Subsequently, HDBSCAN was applied to the low-dimensional embedding to identify natural groupings without prior assumptions regarding the number of clusters. One cluster encompasses all known repeaters, except for FRB20181017A and FRB-20180910A, and is henceforth referred to as the "repeater cluster." The second cluster is predominantly composed of non-repeating sources. Notably, five out of six FRBs initially classified as non-repeating sources, but recently confirmed as repeaters, are naturally embedded within the repeater cluster. A notable exception is FRB20180910A, which our analysis categorizes as a non-repeater despite its recent confirmation as a repeating source. This source is particularly distinctive because its ratio of $\Delta\nu$ to ν_p is significantly larger than that of other detected or predicted repeating FRBs. This outlier property separates FRB20180910A from the rest of the sample and may suggest a distinct emission mechanism or physical environment for this source. For the other outlier, FRB-20181017A, its misclassification can be primarily attributed to its location near the SVM decision boundary in

the $\Delta\nu$ - ν_p parameter space, as illustrated in Fig. 7. Importantly, we demonstrated that the clustering structure remains stable even when the dimensionality of the feature space is reduced from ten to two, indicating the robustness of the identified separation.

MI analysis revealed that the peak frequency ν_p and rest-frame frequency width $\Delta\nu$ are the most informative parameters shaping the low-dimensional representation, in agreement with earlier findings [14]. Scatter plots in the $\Delta\nu$ - ν_p plane reveal that the two clusters can be separated by a straight line $\Delta\nu = 0.95\nu_p + 1.30$, with repeaters tending to exhibit systematically narrower $\Delta\nu$ compared to non-repeaters. The outlier sources FRB20181017A and FRB20180910A are misclassified primarily due to their $\Delta\nu$ values exceeding the empirical decision boundary determined via an SVM model. Overall, our findings suggest that a simplified classification scheme based solely on $\Delta\nu$ and ν_p can effectively distinguish between repeating and non-repeating FRBs. This approach offers a promising framework for the future analysis of expanding FRB catalogs.

Unsupervised machine learning has previously been employed to classify CHIME/FRBs, as demonstrated by Zhu-Ge *et al.* [14] and Qiang *et al.* [21]. Although the method used in our work bears similarity to these prior studies, it incorporates several novelties. First, our dataset selection criteria differ. Previous analyses typically utilized the full CHIME/FRB catalog, including all bursts from repeating sources and all sub-bursts from apparently non-repeating sources. By contrast, our study employs only one burst per source. Second, our feature selection approach diverges. While we initially adopt the same ten features used in earlier work, we demonstrate that the number of features can be reduced to two while maintaining virtually unchanged model performance. This simplification not only streamlines classification but also reveals the key physical properties governing FRB categorization. Third, our results exhibit significant differences. Unlike prior analyses that partitioned FRBs into multiple categories, we show that with appropriate hyperparameter tuning, our method robustly segregates the sample into two distinct clusters. Finally, we report a critical new finding: repeaters and non-repeaters exhibit linear separability in the $\Delta\nu$ - ν_p plane. This provides a method for FRB classification without dimensionality reduction, suggesting a fundamental distinction between the two populations.

References

- [1] D. R. Lorimer, M. Bailes, M. A. McLaughlin *et al.*, *Science* **318**, 777 (2007)
- [2] J. Xu, Y. Feng, D. Li *et al.*, *Universe* **9**(7), 330 (2023)
- [3] E. F. Keane, S. Johnston, S. Bhandari *et al.*, *Nature* **530**, 453 (2016)
- [4] S. Chatterjee, C. J. Law, R. S. Wharton *et al.*, *Nature* **541**, 58 (2017)
- [5] P. Scholz, L. G. Spitler, J. W. T. Hessels *et al.*, *Astrophys. J.* **833**(2), 177 (2016)
- [6] L. G. Spitler, P. Scholz, J. W. T. Hessels *et al.*, *Nature* **531**,

- 202 (2016)
- [7] E. Platts, A. Weltman, A. Walters *et al.*, *Phys. Rept.* **821**, 1 (2019)
- [8] B. Zhang, *Nature* **587**, 45 (2020)
- [9] D. Li, P. Wang, W. W. Zhu *et al.*, *Nature* **598**(7880), 267 (2021)
- [10] H. Xu, J. R. Niu, P. Chen *et al.*, *Nature* **609**(7928), 685 (2022)
- [11] C. H. Niu, K. Aggarwal, D. Li *et al.*, *Nature* **606**(7916), 873 (2022) [Erratum: *Nature* 611 (7936), E10 (2022)]
- [12] B. H. Chen, T. Hashimoto, T. Goto *et al.*, *Mon. Not. Roy. Astron. Soc.* **509**(1), 1227 (2021)
- [13] X. Yang, S. B. Zhang, J. S. Wang *et al.*, *Mon. Not. Roy. Astron. Soc.* **522**(3), 4342 (2023)
- [14] W. P. Sun, J. G. Zhang, Y. Li *et al.*, *Astrophys. J.* **980**(2), 185 (2025)
- [15] J. W. Luo, J. M. Zhu Ge, and B. Zhang, *Mon. Not. Roy. Astron. Soc.* **518**(2), 1629 (2022)
- [16] J. M. Zhu-Ge, J. W. Luo, and B. Zhang, *Mon. Not. Roy. Astron. Soc.* **519**(2), 1823 (2022)
- [17] B. C. Andersen *et al.* (CHIME/FRB Collaboration), *Astrophys. J. Lett.* **885**(1), L24 (2019)
- [18] E. Fonseca, B. C. Andersen, M. Bhardwaj *et al.*, *Astrophys. J. Lett.* **891**(1), L6 (2020)
- [19] M. Amiri *et al.* (CHIME/FRB Collaboration), *Astrophys. J. Supp.* **257**(2), 59 (2021)
- [20] S. Q. Zhong, W. J. Xie, C. M. Deng *et al.*, *Astrophys. J.* **926**(2), 206 (2022)
- [21] B. J. R. Raquel, T. Hashimoto, T. Goto *et al.*, *Mon. Not. Roy. Astron. Soc.* **524**(2), 1668 (2023)
- [22] C. R. García, D. F. Torres, J. M. Zhu-Ge *et al.*, *Astrophys. J.* **977**(2), 273 (2024)
- [23] D. C. Qiang, J. Zheng, Z. Q. You *et al.*, *Astrophys. J.* **982**(1), 16 (2025)
- [24] J. W. Luo, J. R. Niu, W. Y. Wang *et al.*, *Astrophys. J.* **988**(1), 62 (2025)
- [25] B. C. Andersen *et al.* (CHIME/FRB Collaboration), *Astrophys. J.* **947**(2), 83 (2023)
- [26] S. P. Tendulkar, C. Bassa, J. M. Cordes *et al.*, *Astrophys. J. Lett.* **834**(2), L7 (2017)
- [27] B. Marcote, K. Nimmo, J. W. T. Hessels *et al.*, *Nature* **577**(7789), 190 (2020)
- [28] J. P. Macquart, J. X. Prochaska, M. McQuinn *et al.*, *Nature* **581**(7809), 391 (2020)
- [29] W. Deng and B. Zhang, *Astrophys. J. Lett.* **783**, L35 (2014)
- [30] H. Gao, Z. Li, and B. Zhang, *Astrophys. J.* **788**, 189 (2014)
- [31] J. M. Cordes and T. J. W. Lazio, arXiv: astro-ph/0207156
- [32] M. Fukugita, C. J. Hogan, and P. J. E. Peebles, *Astrophys. J.* **503**, 518 (1998)
- [33] N. Aghanim *et al.* (Planck Collaboration), *Astron. Astrophys.* **641**, A6 (2020) [Erratum: *Astron. Astrophys.* 652, C4 (2021)]
- [34] L. Cayton, Univ. of California at San Diego Tech. Rep **12**(1-17), 1 (2005)
- [35] G. Hinton and S. Roweis. In *Proceedings of the 16th International Conference on Neural Information Processing Systems* (Cambridge, MA, USA: MIT Press, 2002), p. 857
- [36] L. van der Maaten and G. E. Hinton, *Journal of Machine Learning Research* **9**, 2579 (2008)
- [37] L. van der Maaten, *J. Mach. Learn. Res.* **15**, 3221 (2014)
- [38] N. Oskolkov. *How to tune hyperparameters of tSNE*. Towards Data Science, US (2019).
- [39] M. Ester, H. P. Kriegel, J. Sander *et al.* In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, (Portland, Oregon: AAAI Press, 1996), p.226
- [40] L. McInnes and J. Healy. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, (New Orleans, LA, USA: IEEE Press, 2017), p.33
- [41] R. J. G. B. Campello, D. Moulavi, and J. Sander. In J. Pei, V. S. Tseng, L. Cao *et al.*, editors, *Advances in Knowledge Discovery and Data Mining*, (Berlin, Heidelberg: Springer Berlin Heidelberg, 2013), p.160
- [42] L. McInnes, J. Healy, and S. Astels, *JOSS* **2**(11), 205 (2017)
- [43] R. Battiti, *IEEE Trans. Neural Netw. Learn. Syst.* **5**(4), 537 (1994)
- [44] M. Rovira, K. Engvall, and C. Duwig, *Chem. Eng. J.* **438**, 135250 (2022)