

Detecting FRB by DANCE: a method based on DEnsity ANalysis and Cluster Extraction

Mao Yuan ^{1,2,★}, Jiarui Niu ³, Yi Feng ⁴, Xu-ning Lv^{1,2}, Chenchen Miao ⁵, Lingqi Meng ³,
Bo Peng ³, Li Deng^{1,2}, Jingye Yan ^{1,2,†}, Weiwei Zhu ^{3,6,‡}

¹State Key Laboratory of Solar Activity and Space Weather, National Space Science Center, Chinese Academy of Sciences, Beijing 100190, China

²Radio Science and Technology Center (π Center), Chengdu 610041, China

³National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100101, China

⁴Zhejiang Lab, Hangzhou, Zhejiang 311121, China

⁵Qilu Normal University, College of Physics and Electronic Engineering, No. 2 Wenbo Road, Zhangqiu District, Jinan, China

⁶Institute for Frontiers in Astronomy and Astrophysics, Beijing Normal University, Beijing 102206, China

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT

Fast radio bursts (FRBs) are transient signals exhibiting diverse strengths and emission bandwidths. Traditional single-pulse search techniques are widely employed for FRB detection; yet weak, narrow-band bursts often remain undetectable due to low signal-to-noise ratios (SNR) in integrated profiles. We developed DANCE, a detection tool based on cluster analysis of the original spectrum. It is specifically designed to detect and isolate weak, narrow-band FRBs, providing direct visual identification of their emission properties. This method performs density clustering on reconstructed, RFI-cleaned observational data, enabling the extraction of targeted clusters in time-frequency domain that correspond to the genuine FRB emission range. Our simulations show that DANCE successfully extracts all true signals with $\text{SNR} > 5$ and achieves a detection precision exceeding 93%. Furthermore, through the practical detection of FRB 20201124A, DANCE has demonstrated a significant advantage in finding previously undetectable weak bursts, particularly those with distinct narrow-band features or occurring in proximity to stronger bursts.

Key words: methods: analytical – methods: data analysis – radio continuum: transients

1 INTRODUCTION

Fast radio bursts (FRBs) have captivated the astronomical community due to the unresolved questions about their origins and the complexities of their emission mechanisms (Zhang 2020; Petroff et al. 2022; Lyubarsky 2021; Zhang 2023). Although it is widely accepted that magnetars are the source of at least a subset of FRBs (Bochenek et al. 2020), alternative models—including those invoking non-neutron star progenitors—have also been proposed (i.e. see the review in Zhang 2023). Investigations into energetics, and burst morphology, particularly for repeating sources, provide crucial insight into the underlying physical processes (Gourdji et al. 2019; Beniamini et al. 2020; Rajwade et al. 2020; Chime/Frb Collaboration et al. 2020a; Cruces et al. 2021; Amiri et al. 2021; Li et al. 2021; Pleunis et al. 2021; Niu et al. 2022; Zhang et al. 2022; Zhou et al. 2022; Zhu et al. 2023). These studies benefit greatly from a larger sample of detected bursts; thus, detecting as many FRB bursts as possible is critical for revealing potential emission periodicities and constraining flux distributions.

A variety of search methods have been developed to detect both one-off and repeating FRBs. The most widely used pipelines are based on single-pulse search techniques (Cordes & McLaughlin

2003; Spitler et al. 2014; Champion et al. 2016; Zackay & Ofek 2017; Petroff et al. 2019; Men & Barr 2024). These methods involve de-dispersing data over various dispersion measures (DMs) and evaluating the signal-to-noise ratio (SNR) of integrated profiles for each DM trial. Two main strategies are used to identify candidates from de-dispersed signals: (1) selecting events with sufficiently high SNR—usually above a threshold of 5 (Zhang et al. 2018)—and (2) locating true pulses in DM-time diagrams. In the latter approach, a dispersed celestial burst produces prominent, centralized structures in a DM versus time scatter plot (Cordes & McLaughlin 2003) or a dedispersion transform space (Zackay & Ofek 2017).

The traditional SNR-based selection and DM search approaches, though widely adopted, often have limited detection sensitivity when striving to ensure true positive identification. The identification of genuine signals within DM-transformed data is frequently impeded by impulsive radio frequency interference (RFI), particularly for pulses with low dispersion measures (Eatough et al. 2009). To mitigate this limitation, some methods split the data into multiple subbands to enhance narrowband FRB detection (e.g. Sand et al. 2022; Zhou et al. 2022).

Machine learning techniques offer a promising alternative for identifying features in FRB signals (Pang et al. 2018; Zhang et al. 2018; Men et al. 2019; Aggarwal et al. 2021). These techniques are generally applied either to DM-transformed data or directly to the original spectral, exploiting the unique characteristics of dispersed pulses.

★ E-mail: yuanmao@nssc.ac.cn

† E-mail: yanjingye@nssc.ac.cn

‡ E-mail: zhuww@nao.cas.cn

In practical applications, such methods have demonstrated higher sensitivity compared to conventional search methods (Zhang et al. 2018). Nevertheless, despite these advances in automated detection, most pipelines—whether based on DM searches or feature-driven classification—still require visual inspection of pulse candidates. This secondary analysis in the time-frequency domain is essential for confirming low-confidence candidates and for conducting detailed investigations of the identified pulses.

We present an unsupervised machine learning tool, DANCE, designed to identify both dispersed and de-dispersed FRBs in the spectrum. The method exploits temporal and spectral information without requiring prior feature training. This method utilizes density analysis and clustering to generate visually interpretable candidates, directly marking their time-frequency locations in the spectrum. For clustering, we use DBSCAN (Density-Based Spatial Clustering of Applications, Ester et al. 1996) to distinguish astrophysical signals from background noise. Similar techniques have been used in FRB detection (Pang et al. 2018; Men & Barr 2024), where clusters are derived from secondary derived features. Because cluster methods are particularly effective for detecting discrete point samples in diagram spaces, some studies have employed clustering to assemble single pulse events within DM-transformed space at optimal DMs (Foster et al. 2018). In contrast, DANCE operates directly on the original spectral data, identifying clusters in the time-frequency spectrum rather than in a DM transformation diagram.

The structure of the paper is organized as follows. In Sec. 2, we introduce the methods, including descriptions of RFI mitigation, density analysis, and cluster extraction. Sec. 4 presents the performance of DANCE on both simulated FRB signals and real observations of FRB 20201124A. In Sec. 5 and Sec. 6, we discuss the extended application of DANCE and make a conclusion of our work.

2 METHODS

Our method exploits the fact that celestial signals, like FRBs, occupy localized regions in the time–frequency plane and exhibit power levels significantly higher than their surrounding noises. By filtering out lower-power samples, the target signal region becomes denser than the surrounding filtered data. Subsequent density analysis methods then enable us to distinguish the signal location from background noise within the spectrum.

The primary aim of DANCE is to project the time-frequency representation into a signal density space (SDS) and subsequently identify regions of higher density within this space. The SDS is represented as a binary matrix with dimensions identical to those of the original time–frequency representation. For each pixel $p_{ij} = (t_i, f_j)$ in the SDS, $p_{ij} = 1$ if the corresponding pixel in the spectrum exceeds a defined power threshold; otherwise $p_{ij} = 0$. The distribution of positive pixels thus characterizes the signal density derived from the spectrum, where genuine celestial signals, such as FRBs, are expected to produce a substantially higher concentration of positive pixels than random noise. The integrated flux in the signal’s frequency band consequently exhibits a distinct peak. DANCE is specifically designed to detect these high-density regions and to extract potential FRB candidates.

As illustrated in Fig. 1, the detection process comprises three principal stages: (1) RFI mitigation using a two-dimensional discrete wavelet transform (2D-DWT), (2) signal filtering via a running-mean algorithm, and (3) signal extraction through DBSCAN clustering.

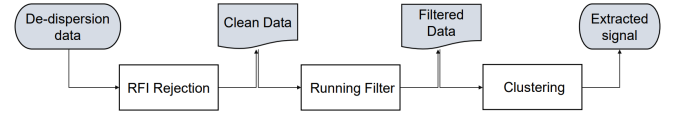


Figure 1. The detection process.

2.1 Two-dimensional discrete wavelet transform

We employ the 2-D DWT to obtain RFI-mitigated data and identify contaminated regions in the time-frequency spectrum. The one-dimensional discrete wavelet transform (1-D DWT) decomposes a discrete signal series, $S[n]$, where $n = 0, 1, \dots, N - 1$, into approximation and detail at multi-resolution. The 2-D DWT extends these principles to a two-dimensional time-frequency representation, $S[m, n]$, enabling independent decomposition along the time (m) and frequency (n) axes.

For the 1-D DWT, a discrete signal, $S[n]$, is decomposed using a scaling function $\phi(t)$ and a wavelet function $\psi(t)$. The approximation and detail coefficients at resolution level j_0 and higher are defined as:

$$\begin{aligned}
 l(j_0, m) &= \sum_{n=0}^{N-1} S[n] \phi_{j_0, m}(n), \\
 h(j, m) &= \sum_{n=0}^{N-1} S[n] \psi_{j, m}(n) \quad (j \geq j_0),
 \end{aligned} \tag{1}$$

where $\phi_{j_0, m}(n) = 2^{j_0/2} \phi(2^{j_0} n - m)$ and $\psi_{j, m}(n) = 2^{j/2} \psi(2^j n - m)$ represent the scaled and translated scaling and wavelet functions, respectively. The scaling function acts as a low-pass filter, producing approximation coefficients $l(j_0, m)$ that capture the coarse signal structure, while the wavelet function acts as a high-pass filter, yielding detail coefficients $h(j, m)$ that represent high-frequency components (Mallat 1999). In practice, the 1-D DWT is implemented using a filter bank with low-pass and high-pass filters followed by downsampling, as described by the fast wavelet transform (Strang & Nguyen 1996).

The 2-D DWT extends the same framework to a time-frequency representation $S[m, n]$, where $m = 0, 1, \dots, M - 1$ and $n = 0, 1, \dots, N - 1$ denote time and frequency indices, respectively. The 2-D scaling and wavelet functions, $\phi(m, n)$ and $\psi(m, n)$, follow the same properties as the 1-D cases. In practice, users commonly choose separable functions to decompose the time-frequency representation (with time and frequency samplings m and n). Following this idea, the 2-D form contains one scaling function and three wavelet functions, they are

$$\begin{aligned}
 \phi_{j_0}^{ll}(m, n) &= \phi_{j_0}(m) \phi_{j_0}(n), \\
 \psi_j^{hl}(m, n) &= \psi_j(m) \phi_j(n), \\
 \psi_j^{lh}(m, n) &= \phi_j(m) \psi_j(n), \\
 \psi_j^{hh}(m, n) &= \psi_j(m) \psi_j(n).
 \end{aligned} \tag{2}$$

Then we get four groups of coefficients through 2-D DWT, they are

$$\begin{aligned}
 LL(j0, m, n) &= \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} S(m, n) \phi_{j0}^{ll}(m, n), \\
 HL(j0, m, n) &= \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} S(m, n) \psi_j^{hl}(m, n), \\
 LH(j0, m, n) &= \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} S(m, n) \psi_j^{lh}(m, n), \\
 HH(j0, m, n) &= \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} S(m, n) \psi_j^{hh}(m, n).
 \end{aligned} \tag{3}$$

Among these four groups of coefficients, The LL coefficients represent the approximation of $S[m, n]$, while the HL , LH , and HH coefficients capture horizontal, vertical, and diagonal details, respectively. The 2-D DWT is computed by applying 1-D DWT sequentially along the time and frequency axes, involving filtering and downsampling operations (Strang & Nguyen 1996).

In our application, a multi-resolution 2-D DWT decomposes the time-frequency representation into subbands at different scales, exposing RFI as localized high-amplitude coefficients in the detail subbands (HL , LH , and HH). RFI-mitigated data are obtained by applying thresholding techniques (e.g., soft or hard thresholding) to the wavelet coefficients and performing the inverse 2-D DWT (Ofir et al. 2010). The choice of wavelet basis (e.g., Daubechies or Haar) and decomposition levels depends on the RFI characteristics and is typically determined empirically.

2.2 Density analysis

We employ a density analysis to identify FRB signals directly within the time-frequency spectrum, rather than relying on the SNR of the integrated profile. This approach consists of two key procedures: signal filtering and density clustering.

Signal filtering is designed to identify “positive” signals in the time-frequency space. In this context, a “positive” signal refers to a sample whose power exceeds a predetermined threshold in the spectrum. Specifically, if the power of a sample surpasses this threshold, its corresponding time-frequency location is considered a true value; otherwise, it is classified as a false value. This method effectively transforms the power spectrum into a binary dataset, wherein the density of true signal positions in the vicinity of an FRB is expected to be significantly greater than in other areas. Nevertheless, if the FRB signal is weak (i.e., with a low SNR in the integrated profile), both desired signals and noise (or omitted RFI) can be marked as positive signals. We perform a second filter to the RFI-filtered data, the method we use is the running mean. The running mean operation averages the energy of isolated outliers within a limited window, helping to reduce noise and smooth outliers by sliding the window along the data series. Mathematically, the smoothed signal is

$$s_{\text{new}}(i) = \left\{ \frac{1}{k} \sum_{j=0}^{j=k-1} s(i+j), i = 0, 1, 2, \dots, n-k-1 \right\}, \tag{4}$$

Given that DANCE focuses on extracting only the spectra area of an FRB within the signal density space (SDS), we apply the running mean window exclusively within the bandpass for each temporal recording. We then filter out signals that fall below a specified threshold and flag the positions of positive signals in the time-frequency space. The result of this filtering is a binary dataset, where positive

signal locations are represented by 1 and false cases by 0, visually depicted as points in the time-frequency space.

After obtaining the density distribution of positive signals in the time-frequency space, our subsequent objective is to cluster the dense zones. We employ the DBSCAN method to identify potential dense clusters. DBSCAN is a density-based technique to discover clusters in a dataset with both designed signals and noise (Ester et al. 1996; Khan et al. 2014), it is a supervised machine learning method without designing the cluster numbers and types. The leading theories of DBSCAN are density-reachability and density-connectivity (Ester et al. 1996). Two key parameters dominate the clusters DBSCAN can find, the neighborhood radius of a point Eps and the minimum samples in the neighborhood radius, $MinPts$. For any sample p in a DBSCAN-identified cluster C , it must satisfy two conditions,

$$\begin{aligned}
 Num_{Eps}^p &\geq MinPts, \\
 dist_{p,q} &\leq Eps.
 \end{aligned} \tag{5}$$

The first relation indicates a density-reachable condition, while the second relation pertains to density-connected conditions. Among which, Num_{Eps}^p is the sample number of a circle centered at p with a radius is Eps , $dist_{p,q}$ is the distance between p and its any neighboring samples q . FRB-signal regions in the smoothed time-frequency space are dense and continuous in comparison with the noise. Therefore a density-reachable and density-connected clustering can find them well. In addition to the designed signals, noise can also be grouped into various discrete and small-size clusters. Setting large Eps and $dist_{p,q}$ prevents clustering the sparse noise, but that may also leave out some weak FRBs. We choose small Eps and $dist_{p,q}$ for DANCE.

2.3 Cluster extraction

Dense analysis can generate various clusters when we choose an easy density-connected condition. We need to gather the designated cluster samples and accurately identify the actual location of the FRB using the appropriate clusters. Cluster prominence is a valid criterion since only astronomical signals that have a significant range of frequencies and a wide pulse duration are capable of forming extensive and compact clusters. We can calculate the pixel number of a cluster sample to measure its prominence and extract the samples with substantial coverage in the SDS by applying a statistical criterion to all cluster sizes.

The extraction strategy is a Z-score transformation of the size statistic. We calculate the Z-score of all identified clusters and then filter the trivial samples. Mathematically, Z can be calculated like

$$Z = \frac{x - \mu}{\sigma}, \tag{6}$$

where x is the size statistic of all clusters, μ and σ is the mean and standard deviation of x . Prominent samples get high Z and it is easy to extract them by a simple threshold filtering.

We also designed a double criterion to filter possible RFI or noise clusters. Because the SDS is RFI-clean and noise-smooth, omitted interference could be a baseline fluctuation or background noise. Hence signal pixels of them in SDS are widely distributed. An FRB pulse differs mainly in limited pulse width, as does its cluster. Therefore, we also calculate the characteristic width of previously extracted samples and exclude wide cases. The characteristic width is measured by the maximum range of the sample in the temporal dimension.

3 PROCESSING FLOW

This section describes the data processing and burst signal extraction steps used by DANCE. We show the visual results of the key flows using the actual observation data from FAST.

3.1 RFI mitigation and denoising

The strategy of most RFI mitigation techniques is flagging the RFI blocks and then zapping them (Agarwal et al. 2020; Rafiei-Ravandi & Smith 2023). Such an operation helps to enhance the SNR of detection results in dedispersed integrated signals. However, a zapped spectrum causes substantial signal density shifts that may bias cluster formation. As we mentioned in Sec.2, DBSCAN is a density-based connectivity approach that ceases clustering when encountering gaps in the data, potentially preventing the formation of a continuous dense cluster for the FRB signal.

To demonstrate the use of 2-D DWT for RFI reduction, we use a dataset that exhibits significant interference. The frequency range covered by this data, which was observed by FAST, is 1000–1500 MHz, and the duration time is 6.24 s. Both frequency and time samples in the dataset are 4096. We removed the baseline in the time dimension by deducting the mean power of each channel. Fig.2 shows the RFI mitigation procedures. We first decompose the data into multi-levels. In practice, the pixel resolution of the decomposed picture decreases in the form of 2^l where l is the level. For example, the original data size is 4096×4096 in time and frequency samples. When the decomposition level is $l = 5$, the size of the “LL” component should be 128×128 , while the other components (“LH”, “HL”, and “HH”) are $4096/2^i \times 4096/2^i$ which i is the output layer, defines from 1 to 5. We then set a threshold filtering to all the layers, in which the values beyond $\pm 1\sigma$ are replaced by 0. The blank data in panel (c) are the flagged outliers. Finally, we make a 2-D inverse DWT by using the filtered decomposition layers and derived reconstructed data, shown in panel (d).

The 2-D DWT method meets our density analysis requirements and offers several benefits. First, it allows RFI elimination without deleting data blocks. Second, only filtering a subset of the deconstructed components keeps the desired signals. Even if RFI signals are weaker than FRB signals, we can indicate them without erasing the planned signals. The bias only occurs with clean data or strong FRB signals.

3.2 Density analysis

The reconstructed spectrum is RFI-poor but is still not smooth due to some isolated outliers and noises, shown in Fig.2 and Fig.3 (left panel). As we introduced in Sec.2, outliers in the spectrum can contaminate the SDS.

We use a running filter to smooth the data and mitigate the fluctuations of noise. The middle panel in Fig.3 is the smoothed signal of the original spectrum (left panel), which shows that the new spectrum is smooth except for the high-density zone which indicates an FRB signal. The width of the running window in this case is 256 frequency channels (~ 30 MHz). It is important to note that the burst bandwidth could be broadened after a smooth operation. The extension effect is dependent on both the width of the running window and the signal strength. We have a discussion about it in the next section.

The SDS is a presentation of pixel location where the signal power is over an artificial strength threshold. To evaluate the density of SDS, we apply a filtering process based on the Interquartile Range (IQR), a measure of statistical dispersion. The IQR is the range between the

first quartile (25th percentile) and the third quartile (75th percentile) of a data set, effectively capturing the spread of the central 50% of the data. The right panel in Fig.3 is an SDS with a threshold at 1.5 IQR (beyond the 75th percentile of the signal power statistics). We perform a DBSCAN clustering analysis of the density distribution in SDS (Fig.3, right panel). As we introduced in Sec.2, Eps and $MinPts$ determine how the DBSCAN defines a density-connected set. The $MinPts$ is an easier-set parameter for DBSCAN. A relatively small $MinPts$ helps include all possible clusters. Usually, setting $MinPts$ to twice the dimensionality of the data, i.e., $MinPts = 4$, or near this value, is suitable for most density analysis cases (Sander et al. 1998; Schubert et al. 2017). The other parameter Eps is harder to set as it depends on the distance of the neighboring samples in the wanted clusters, which is a posterior measurement. Some researchers suggested adaptive mathematical approaches to calculate and get appropriate Eps before performing clustering (Ester et al. 1996; Starczewski et al. 2020). For FRB clustering, we determine the Eps and $MinPts$ according to the FRBs’ morphological characters. Being different from traditional point sets, the SDS are projection of a filtered spectrum. The majority of the points are locations of RFIs or noises. Clusters defined by them are random, sporadic, and small in size. On the contrary, the FRB clusters should be larger denser zones gathered with continuous points. Therefore we choose a larger $MinPts$, thus helping to pass most noise clusters. We define a dimensionless density parameter ρ to replace $MinPts$, i.e. $MinPts = \lceil \rho N_0 \rceil$, in which S (in pixel^2) is a clustering cell area determined by radius Eps , and ρ as the multiple of the average number of elements per pixel (N_0) in the SDS.

Considering that the narrow burst can exhibit a short duration within a timescale at tens of microseconds (or even narrower) (Farah et al. 2018; Chime/Frb Collaboration et al. 2020b), we set $Eps = 2$ which is twice the sampling timescale ($49.152 \mu\text{s}$) of the FAST. By filtering the spectrum using 1.5 times the IQR, we remove outliers and focus on the core clusters in SDS. This results in a background point density of 0.25 per pixel. Thus, the appropriate density parameter for DBSCAN is $1 \leq \rho \leq 4$.

A larger ρ helps to extract stronger signals, we suggested $\rho \sim 2$ is suitable for clustering signals from noises as well as possible. Fig.4 shows the result of FRB detection with DANCE. The left panel is the DBSCAN-identified clusters. Here the radius of the scan circle Eps is set as 2 pixels. The density-reachable threshold, $MinPts$ is set as 2 times larger than the sample density of SDS. We totally identified 25,023 clusters in this SDS, and we get two prominent samples, which are labeled as -1 and 13,180, shown in the middle panel. We use the number of samples a cluster contains to indicate its size. Among them, the cluster with the label 13,180 is the projection of an FRB while the most prominent cluster is the global background noise.

3.3 Cluster extraction

We extract prominent clusters based on their sizes. A statistic of the cluster size is represented in Fig.4, the middle panel. We calculated the statistical Z-score (Z) of these cluster sizes, and Z scores are shown in Fig.5, left panel. This case shows two prominent clusters with $Z > 3$. We then calculate the characteristic pulse width of each cluster, shown in the middle panel of Fig.5. As we have introduced in Sec.2, secondary filtering is necessary to remove some RFI- or background noise-clusters. We use the scale of temporal coverage of a cluster to define its pulse width. The middle panels indicate that the width of the background noise-cluster covers the whole observation subband (~ 200 ms).

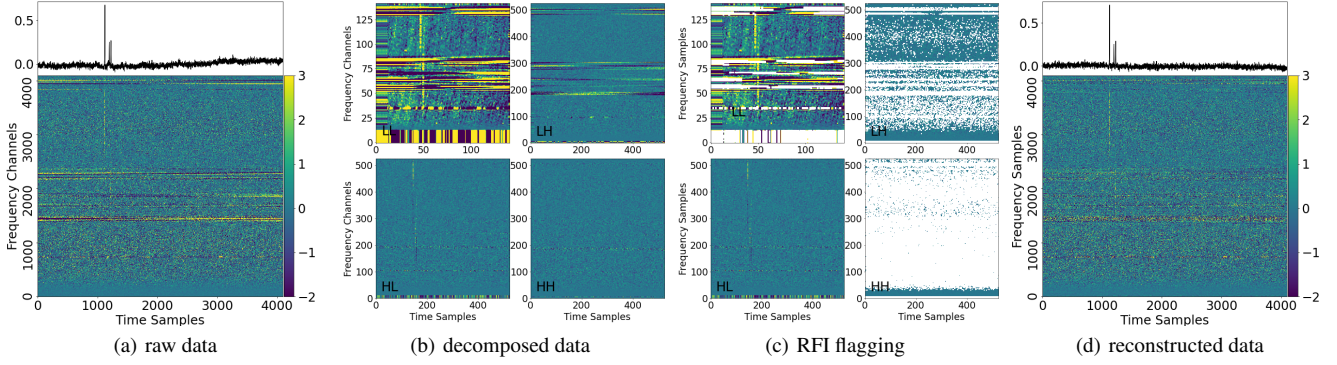


Figure 2. RFI mitigation using the 2-D DWT for strong FRB signals. (a) The original data containing three FRB bursts, with the top panel showing the integrated signal. (b) The wavelet decomposition at the third level, displaying the four subcomponents “LL, LH, HL, and HH”, respectively. Among these, “LL” represents the low-frequency approximation of the raw data, “HL, LH” extract horizontal and vertical features of the data, respectively, and “HH” captures fine-scale details. The decomposition is performed over five levels, with the third-level components shown here as representative examples. (c) RFI identification, where blanked regions indicate flagged interference at a 1σ threshold. “HL” component is retained unflagged, as the desired FRB signals are primarily represented in this channel. (d) The reconstructed data obtained after RFI removal from the decomposed components. All panels share an identical color scale for direct comparison.

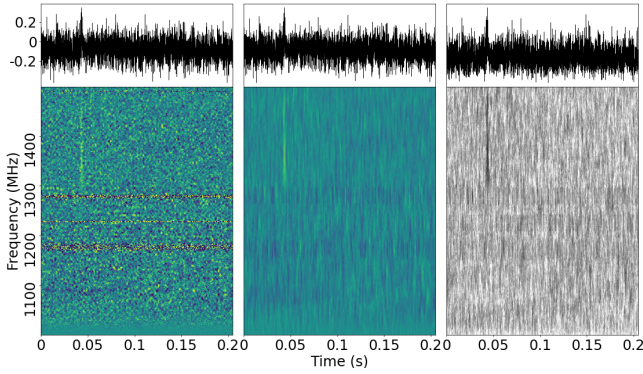


Figure 3. Transformation from the reconstructed spectrum to the signal density space (SDS). The left panel shows the reconstructed data; the middle panel displays the smoothed data; and the right panel presents the corresponding signal density representation. A weak FRB signal is visible within the region at 0.04-0.045 s in time and approximately 1300-1500 MHz in frequency.

Though vast sporadic noises have comparable widths with FRB signals (e.g. 1-10 ms), their sizes are trivial. We extract the FRB cluster by setting thresholds to cluster sizes and widths. In this case, we choose cluster samples with $Z > 3$ and $W < 100$ (ms). FRB signals are the only sample that has both a prominent size and reasonable pulse width, as shown in the right panel. By using DANCE we can extract the weak FRB with more convincing visible detection results than the common integrated profile method.

4 RESULTS

4.1 Detection on simulated FRBs

We evaluate the capability of DANCE by detecting simulated FRBs in random noise. The experiment strategy is to produce stimulative FRB signals with various strengths in background data, and then use DANCE to extract the targets. By counting false and missing identification cases, we finally approximate the detection limits of DANCE.

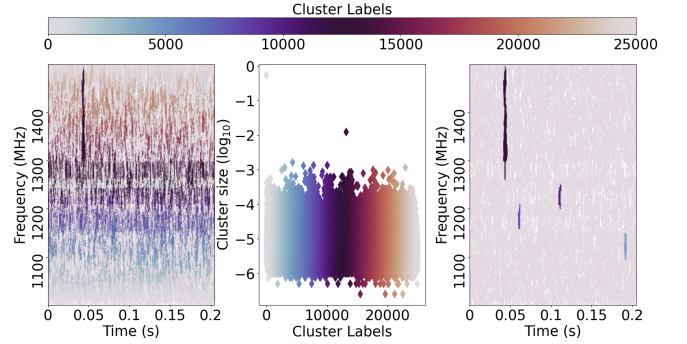


Figure 4. DBSCAN-identified clusters. The top color bar indicates different clusters. Left: all separated cluster samples in the SDS of Fig. 3, with each color representing a distinct cluster. Middle: the size of each cluster sample, measured by the proportion of pixels each cluster occupies. Two notable clusters are labeled -1 and 13180 in the color bar, with sizes of 54% ($\log_{10} = -0.64$) and 1% ($\log_{10} = -2$), respectively. Right: the five largest cluster samples, including the background noise (label -1), an FRB signal (label 13180), and three other noise clusters.

For the background data, we generate a matrix with a dimension is 4096×4096 . The element values of the data block follow the Gaussian distribution with a mean value of 0 and a standard deviation of 1. We take the noise data as a spectrum background, in which the integration time and frequency bands are assumed at 200 ms and 500 MHz. For the simulated burst, we use a two-dimensional Gaussian function to generate FRB signals. Assuming the signal strength is $F(f, t)$ at f MHz and t ms, it has the form like

$$F(f, t) = Ae^{-\left(\frac{(C_t-t)^2}{2\sigma_t^2} + \frac{(C_f-f)^2}{2\sigma_f^2}\right)} \quad (7)$$

Here A dominates the strength, i.e. the SNR, of the integrated pulse. C_t and C_f indicate the center location of the signal at time and frequency, respectively. During the simulation, our progress randomly generated three parameters A , σ_t , and σ_f , which control the pulse strength, width, and bandwidth, respectively.

Fig. 6 presents combined data (left panel) and the simulation data

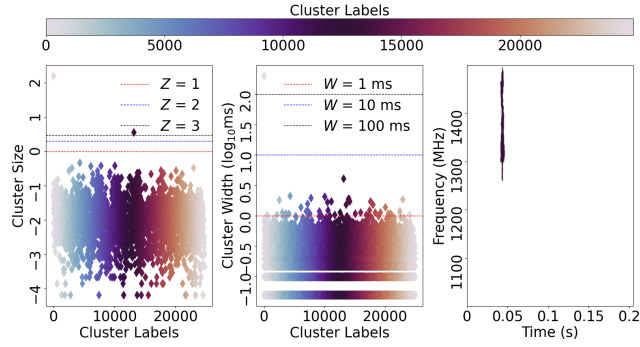


Figure 5. Cluster extraction based on Z-score and pulse width. The color bars at the top indicate the cluster labels. Left: Z-score of all cluster sizes in the SDS of Fig.3, each color indicates a cluster. Middle: the pulse width of each cluster sample. Right: the filtered FRB cluster with a Z-score of 3.6 and a pulse width of 4.1 ms, respectively. This cluster has been filtered out based on a criteria of $Z > 3$ and $W < 100$.

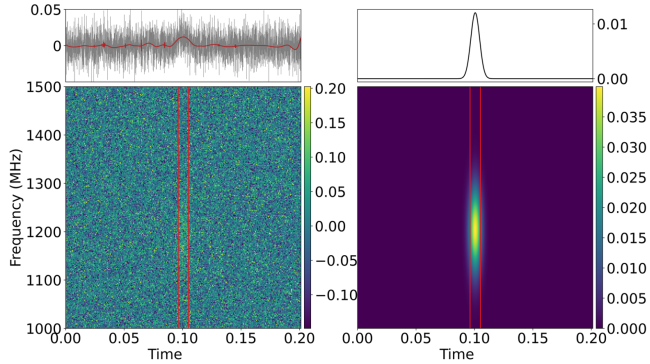


Figure 6. A presentation of a simulated FRB. The left panel is the data mixed with noise and analog FRB signal. The right panel is the simulated FRB signal. The pulse width is 9 ms, which is denoted by the vertical red lines. The bandwidth is 140 MHz and the SNR of the combined signal is ~ 8.3 .

(right panel). The artificial FRB in this case located at the center of the spectrum, with a pulse width of 9 ms and a bandwidth of 140 MHz. Here we define the pulse width W_p (in bins) as the full-width half maximum (FWHM) of the peak of the integrated simulated pulse. The SNR of the combined integrated signal is ~ 8.3 , which was calculated by

$$SNR = \frac{1}{\sqrt{W_p} \sigma} \sum_{i=1}^{n_{bins}} p_{simu}, \quad (8)$$

where σ is the standard deviation of the integrated time series of the background data, and $\sum_{i=1}^{n_{bins}} p_{simu}$ denotes the summed power of the simulated gaussian pulse within the FWHM.

Fig.7 is the detection result of the simulated signal above. The right bottom panel shows that DANCE finds this burst correctly. We use DANCE to process the simulated data, the total sample number (SN) is 2,550. Among these samples, the pulse width ranges from 1 ms to more than 10 ms; and the bandwidth ranges from 100 MHz to 500 MHz. The SNR varies from less than 2 to greater than 8. We counted the true positive samples and false positive samples and recorded the number of the two samples as TN and FN . We use the concept of recall and precision to test the performance of DANCE on

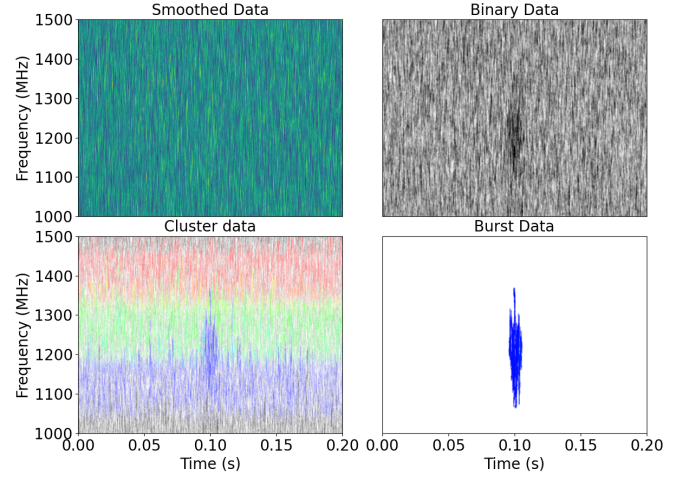


Figure 7. A presentation of the detection result of DANCE to the data shown in Fig.6. The progress DANCE first smooths the data and then filters the smoothed data according to the strength values of each pixel. The cluster data is derived by performing the DBSCAN on the binary data, and the final burst data is filtered by the size and width of the clusters.

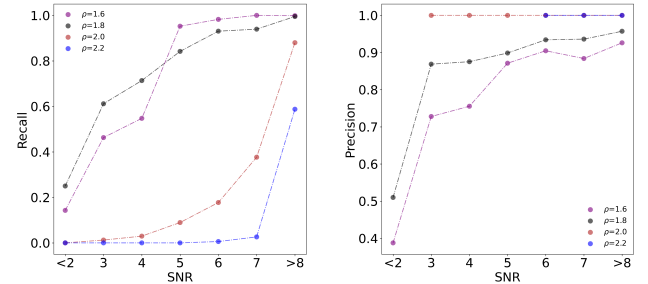


Figure 8. Recall and precision of DANCE based on simulated FRBs.

simulated FRBs. Because we do not set negative samples (i.e. a data set without simulating FRB signals), the recall and precision can be calculated as

$$Recall = \frac{TP}{SN}; \quad (9)$$

$$Precision = \frac{TP}{TP + FP}.$$

We conducted a series of simulations to evaluate the performance of DANCE. The simulated bursts span an SNR range from below 2 to above 8. Two parameter sets were tested, with $Eps = 2$ and $Eps = 4$, respectively. For each Eps value, we explored various density thresholds, ρ . Fig.8 presents the recall and precision achieved with $Eps = 4$ and $\rho = 1.6, 1.8, 2.0$, and 2.2 . Since the detection sensitivity with $Eps = 4$ is excessively high for bursts with $SNR \gtrsim 3$, we focus primarily on the results obtained with $Eps = 2$. The results indicate that an appropriate density parameter should be smaller than 2. Considering bursts with $SNR \geq 5$ as representative of convincing FRB signals, the recall and precision reach 0.98/0.90 for $\rho = 1.6$ and 0.93/0.94 for $\rho = 1.8$.

It should be noted that the recall and precision of signal detection based on simulation data are usually biased from real data, especially

when the real data contains unstable RFI (Yuan et al. 2022). We will present the performance of DANCE on real FRBs in the next section.

4.2 Detection on FRB20201124A

We test the performance of DANCE with a real observation case of FRB 20201124A. This is a highly active repeater that was first discovered by the Canadian Hydrogen Intensity Mapping Experiment Fast Radio Burst (CHIME/FRB) (Chime/Frb Collaboration 2021). The data we examined was observed by FAST on September 29th, 2021, at a period of heightened activity for this source. (Zhou et al. 2022; Zhang et al. 2022; Jiang et al. 2022; Niu et al. 2022).

The observation time is 1 hour and the frequency band covers from 1000 to 1500 MHz. The time and frequency resolutions are 49.152 μ s and 0.122 MHz. We de-disperse at first and then split the data into subsets without any time or frequency down-sampling. Every subset remains the same dimension (4096 $\times$ 4096) with the simulated data. The parameters we adapt also remain the same with the experiment on simulated bursts.

We totally processed 17,263 data sets (0.2 s for each one), and the positive samples are 4000. The parameters for cluster forming and extraction are $Eps = 2$; $MinPts = 6$ and $Z = 3$. We down-sampled every positive data set to a 256 $\times$ 256 spectrum and then identified the real burst by eye. We confirmed 652 true bursts from the positive triggers at last. As a comparison, the number of reported bursts in this data set is 500 (Zhang et al. 2022). Among the signals we certified, there are many extremely weak signals whose SNR is far below 3σ . Fig. 9 and 10 show parts of the weak bursts. Interestingly, these weak bursts are not only isolated within a short timescale (0.2 s) but also closely leading or following strong bursts. We defined these two types of weak bursts as isolated and associated weak bursts. Such pre- or post-cursor weak burst components are also mentioned in other repeaters, like FRB 2020121102 (Caleb et al. 2020).

The successful capture of weak FRBs at the noise level is the advantage of DANCE over approaches based on integrated profile strength. Nevertheless, we also notice that the precision is not comparable with the experiments on simulated data. Misidentifications could be an unavoidable cost for weak FRB extraction as the RFI in real data is more complex than simulated Gaussian noise. Indeed, a high parameter level will definitely improve the precision, but it will also lead to an under-reporting of parts of weak FRBs. We suggest that it is worthwhile to sacrifice precision to improve the detection ability of weak signals, especially since the DANCE provides a visual advantage that greatly reduces the manual confirmation work.

5 DISCUSSION

DANCE shows high sensitivity in identifying weak FRBs that remain indiscernible in conventional profiles. However, these target signals are de-dispersed, and the detection performance depends critically on the choice of parameters. In this section, we discuss the trials of blind searching of FRBs under different parameter settings.

The two key parameters, Eps and ρ (i.e. the $MinPts$), determine the minimum spatial extent and density threshold required for a cluster to be identified. For a randomly distributed SDS, assuming that the point density is ρ_0 per pixel ($\rho_0 = 0.25$ when the spectrum is filtered at a 1.5 IQR level), and given clustering parameters Eps and ρ , then the probability of forming a cluster in the area is

$$P_{S,\rho} = C_{[S]}^{[S\rho\rho_0]} \rho_0^{[S\rho\rho_0]}, \quad (10)$$

where $S \sim \pi Eps^2$. Theoretically, increasing either Eps or ρ reduces

the probability of noise-induced clustering, thereby improving the reliability of true signal identification. Meanwhile, overly large parameter values may also suppress weak or small-scale signals.

In practice, adopting multiple parameter combinations can be advantageous when dealing with complex burst morphologies. Fig. 11 illustrates clustering results for several bursts with distinct bandwidths, durations, and intensities. A suggested detection approach, using $Eps = 2$, $\rho = 1.9$ successfully identifies all isolated structures, while a second configuration with $Eps = 6$ and $\rho = 2.5$ detects all bursts as well, and producing clusters confined to more precise time–frequency regions.

We use DANCE to extract a non-dispersed FRB under different cluster parameters, shown in Fig. 12. The left top panel shows the FRB signal and the other panels are clusters extracted by DANCE. Among the cluster panels, the three in the right column are the whole clusters of DBSCAN. We present all the clusters because DANCE detected no positive samples under the corresponding parameters.

This burst is the sample shown in the top left panel of Fig. 10 (MJD=59485.78380). We downsample the spectrum to a 4096 $\times$ 4096 matrix, the time resolution is 0.2 s. Thus the original triple components burst was compressed together. The Eps and $MinPts$ set for DBSCAN were labeled in each panel. The cluster extraction threshold Z is 3. For each group of Eps , low density causes sporadic noise or RFI clusters (left panels), and high density leads to discrete FRB clusters, which have small sizes and can not pass the extraction threshold.

The clustering to non-dispersed FRB demonstrates the capability of DANCE for FRB searches. Using this method, one can detect a pulse without performing de-dispersion. That is, we can use DANCE search FRB or other transit pulses without much computational resources and providing immediate visual assessment, and provide a rapid initial check for the presence of transient bursts in an observation. Nevertheless, there are limitations of DANCE in blind searches. First, it is parameter-dependent: FRBs with different DMs have varying distributions in the time–frequency domain, so DANCE may be triggered under different density-reachability and connectivity conditions. Second, only strong pulses might be clustered in relatively larger scale and density parameters. Third, DANCE does not provide DM information.

This limitation arises from the intrinsic behavior of the underlying DBSCAN algorithm, whose effectiveness strongly depends on the geometric morphology of the target signal (Ester et al. 1996). DBSCAN defines clusters as regions of high point density in a metric space, i.e. points shared a same cluster are neighbored in Euclidean distance, which defines as $dist_{p,q} = \sqrt{(x_p - x_q)^2 + (y_p - y_q)^2}$. Thus a neighbouring clustered samples approximately forms an isotropic circular or elliptical zone in the feature space. In the case of non-dedispersed signals, the intrinsic geometry of the burst is highly anisotropic, appearing as a long, tilted track across the time–frequency plane due to dispersion. This mismatch between the isotropic distance metric and the anisotropic signal structure can lead to cluster fragmentation or reduced recall, as distant parts of the same burst fall outside the defined neighborhood radius Eps .

Furthermore, in non-dedispersed data, the effective density of signal points projected along the dispersion track is substantially lower than that in dedispersed data. Consequently, achieving the $MinPts$ threshold for core point identification becomes more difficult. Increasing the neighborhood size Eps can partially compensate for this effect, but it also increases the inclusion of noise, thereby reducing precision. Consequently, DBSCAN—and thus DANCE—has limited sensitivity to weak, dispersed bursts; typically, only strong initial pulses can be clustered with a positive response of DANCE.

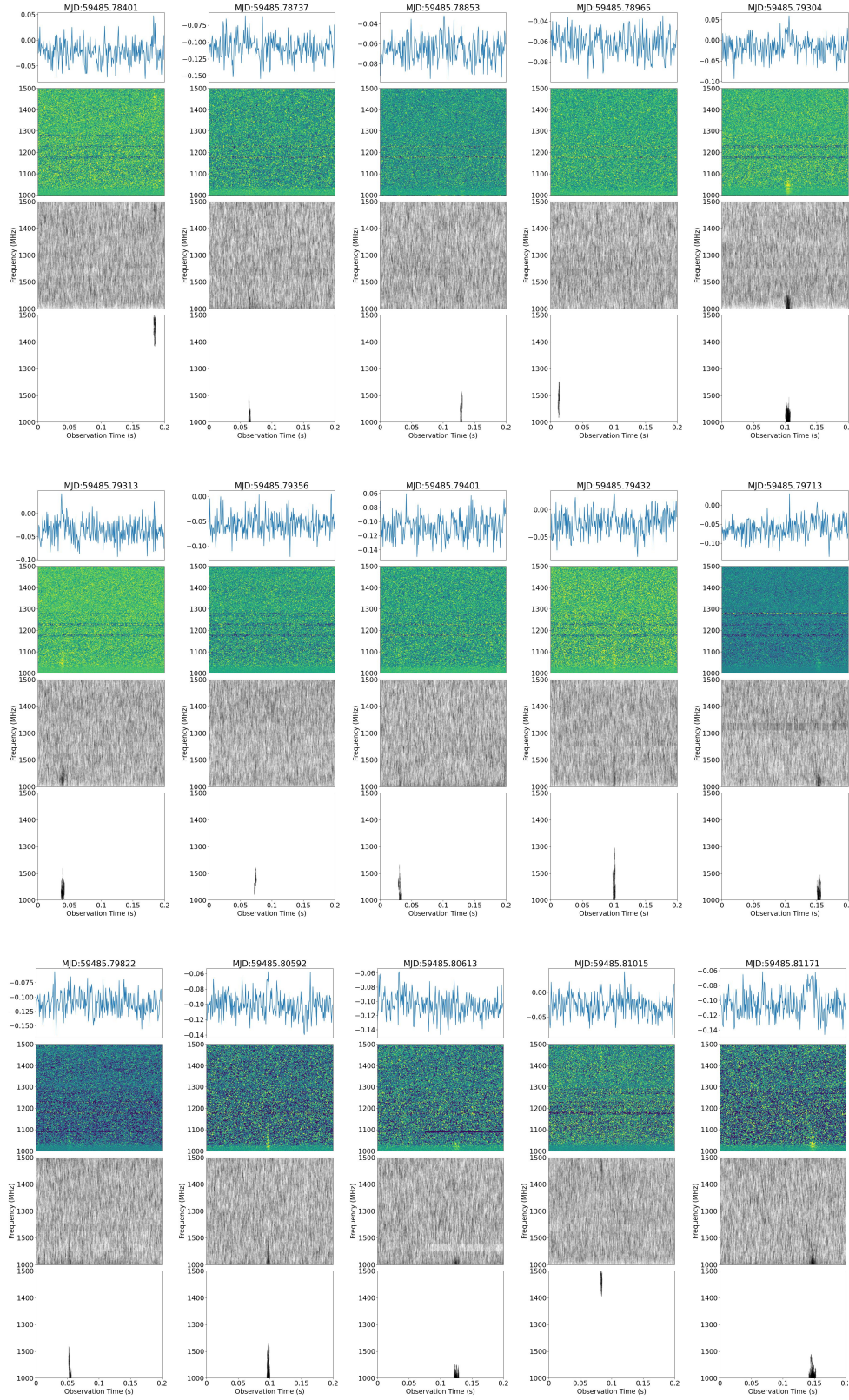


Figure 9. True FRB signals — isolated weak bursts. In all of these subfigures, the top panels are integrated profiles of the RFI-cleaned data, which have been 16-downsampled in both the frequency and time domains to provide a better representation of the bursts shown below the profiles. The bottom two panels show the signal density representations and the extracted burst labels. Because these bursts appear in isolation within 0.2 s, we named these bursts as isolated weak bursts.

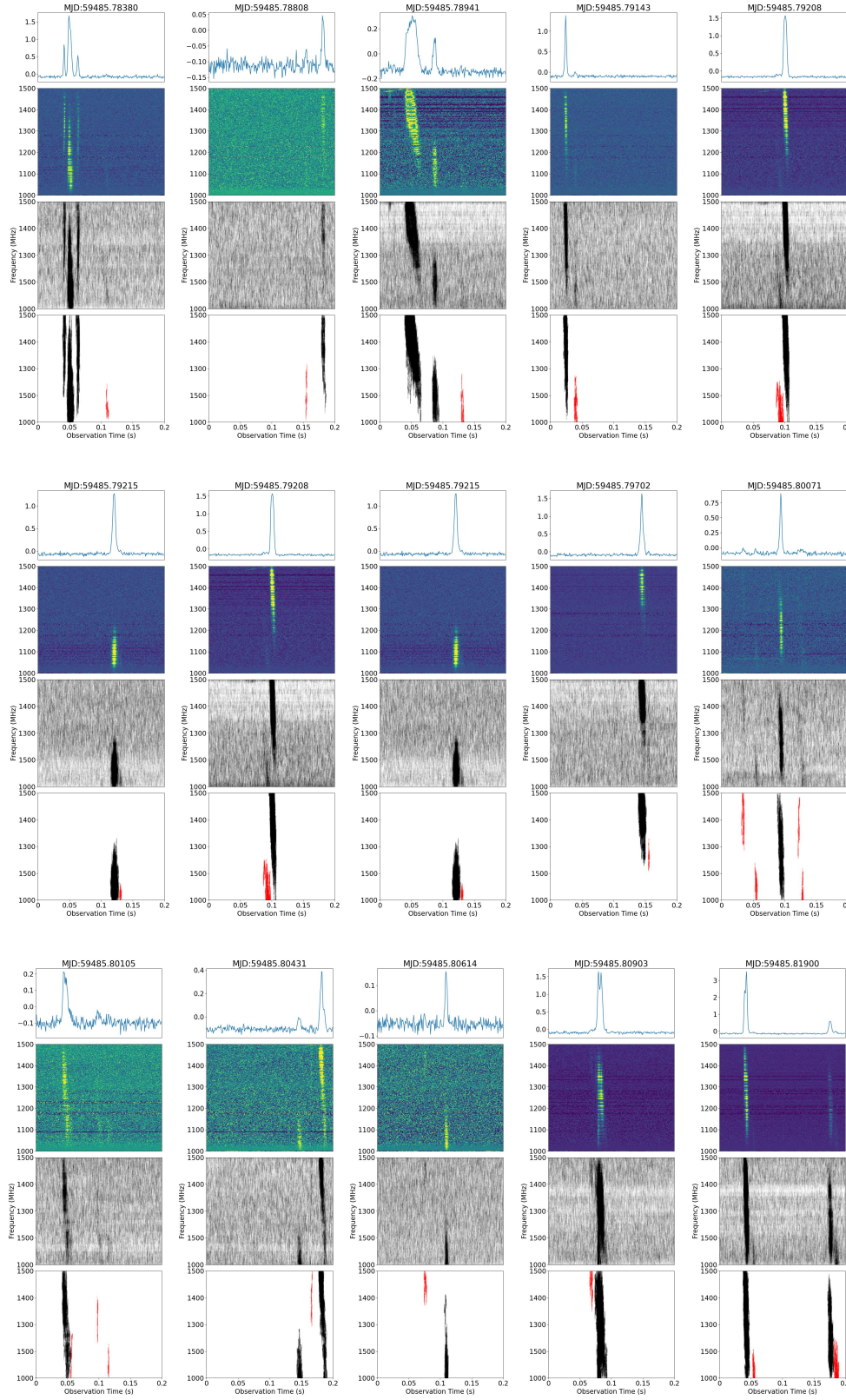


Figure 10. True FRB signals — associated weak bursts. The difference between these panels and the figures in Fig.9 is that weak bursts here (red clusters) are accompanied by relatively strong bursts (black clusters) within a character timescale like 0.2 s. We defined such bursts as associated weak bursts.

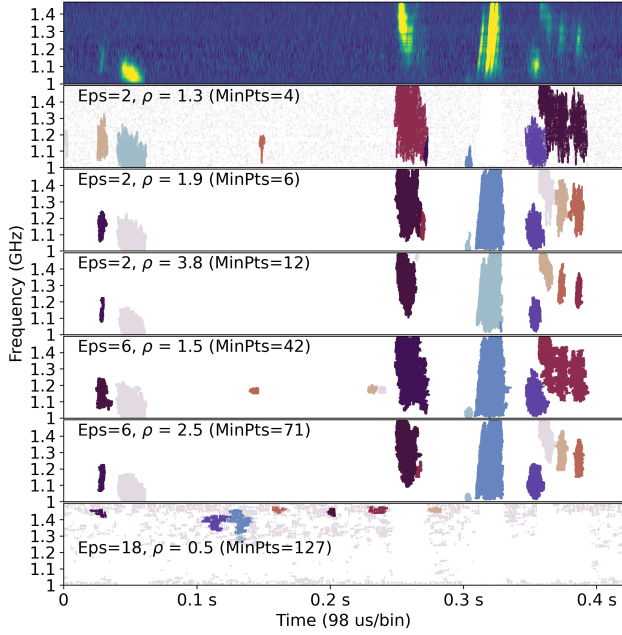


Figure 11. Clustering of FRBs under different parameter settings. The top panel shows a smoothed spectrum containing several bursts, while the six bottom panels display the ten most significant clusters identified under varying (Eps , ρ) combinations, as indicated in each panel.

Moreover, since different dispersion slopes correspond to distinct track geometries, the detection sensitivity may vary even for bursts of comparable intrinsic strength. It prevents a unified quantitative assessment of DANCE’s performance on non-dedispersed data.

Despite these limitations, DANCE remains a useful tool for preliminary detection in raw, non-dedispersed data. It can efficiently highlight strong transient candidates without the need for computationally expensive dedispersion trials. Once such candidates are triggered, a subsequent dedispersed, post-processed search can be conducted with optimized parameters to achieve higher sensitivity and completeness. This two-step strategy effectively combines DANCE’s robustness with improved performance after dispersion correction.

6 CONCLUSION

We developed an unsupervised method, DANCE, to detect and isolate FRB signals in the original spectrum, providing visual identification of their precise time-frequency locations and detailed morphology. DANCE demonstrates reliable accuracy in detecting weak FRBs; simulations indicate that for FRBs with $SNR > 5$, the detection precision reaches 93%. The practical performance on real observation shows that DANCE has uncovered numerous previously undetected weak, narrow-band FRBs in historical datasets. This method shows promise as a powerful tool for identifying and distinguishing weak FRBs, even in cases where they are located very close to stronger signals.

ACKNOWLEDGEMENTS

This work made use of data from the FAST, a Chinese national mega-science facility, built and operated by the National Astro-

nomical Observatories, Chinese Academy of Sciences. This work is supported by the independently funded research program of National Space Science Center (“2025 Young Scientist Innovation Program”, No. E5PD40004S). This work is supported by the National Natural Science Foundation of China (NSFC, No. 12503055) and the Postdoctoral Fellowship Program of CPSF under Grant Number GZB20250737. This work is also supported by the National SKA Program of China (Nos. 2020SKA0120200 and 2020SKA0120100).

DATA AVAILABILITY

The data supporting this article will be shared when a reasonable request is made to the corresponding author. The raw data and can be accessed from FAST data center <https://fast.bao.ac.cn/>. The code used in this article is publicly available at: https://github.com/Yuanao/DANCE_v-test.

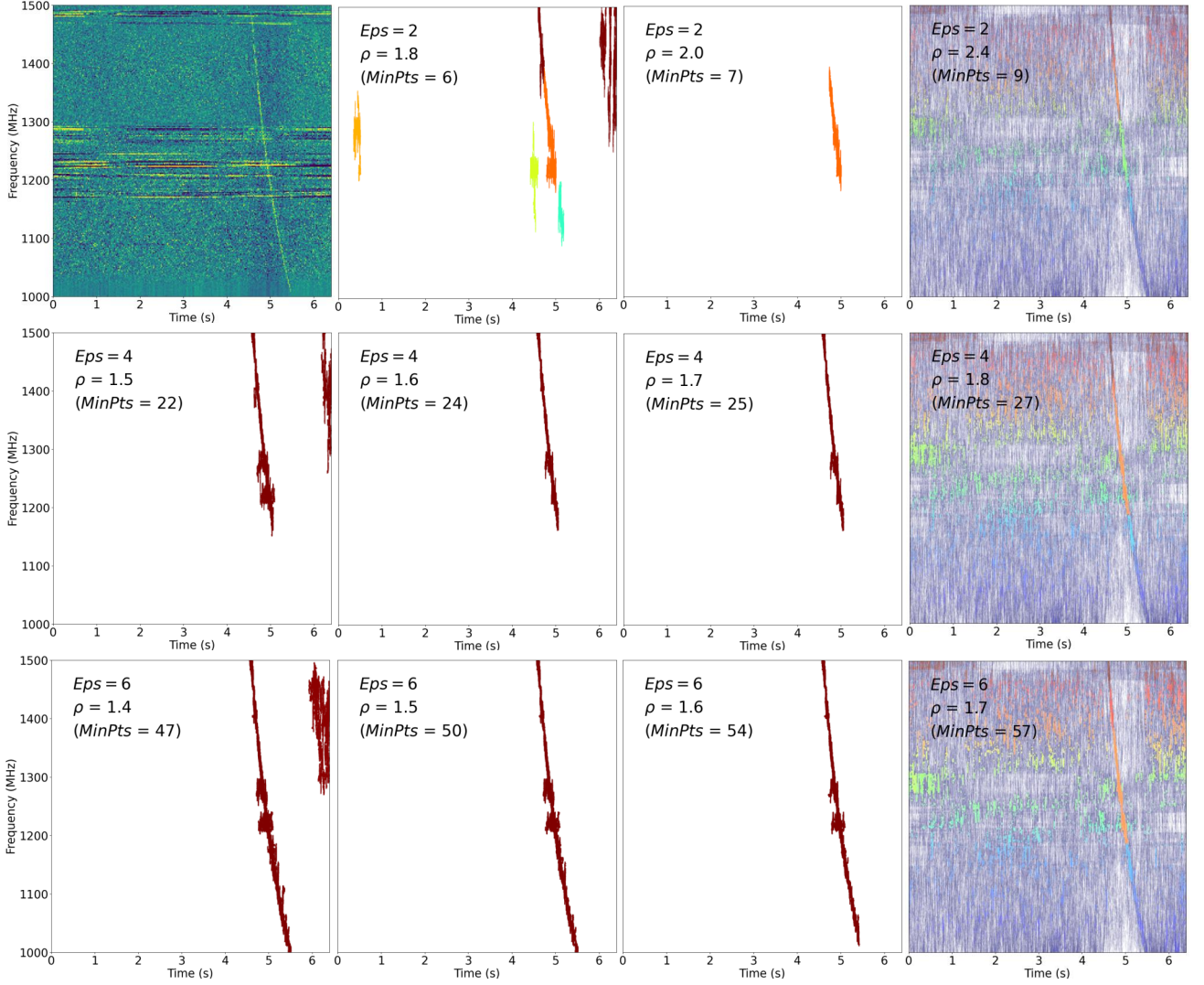


Figure 12. DANCE performance on non-dispersed FRB signals. The top-left sub-panel displays the original spectrum. The three sub-panels in the right column present all identified clusters of the spectrum using different clustering parameters before final extraction. The remaining sub-panels present the filtered clusters as the final results.

REFERENCES

- Agarwal D., et al., 2020, *Monthly Notices of the Royal Astronomical Society*, 497, 352
- Aggarwal K., et al., 2021, *The Astrophysical Journal*, 914, 53
- Amiri M., et al., 2021, *The Astrophysical Journal Supplement Series*, 257, 59
- Beniamini P., Wadiasingh Z., Metzger B. D., 2020, *Monthly Notices of the Royal Astronomical Society*, 496, 3390
- Bochenek C. D., Ravi V., Belov K. V., Hallinan G., Kocz J., Kulkarni S. R., McKenna D. L., 2020, *Nature*, 587, 59
- Caleb M., et al., 2020, *Monthly Notices of the Royal Astronomical Society*, 496, 4565
- Champion D., et al., 2016, *Monthly Notices of the Royal Astronomical Society: Letters*, 460, L30
- Chime/Frb Collaboration et al., 2020a, *Nature*, 582, 351
- Chime/Frb Collaboration et al., 2020b, *Nature*, 582, 351
- Chime/Frb Collaboration 2021, *The Astronomer’s Telegram*, 14497, 1
- Cordes J., McLaughlin M. A., 2003, *The Astrophysical Journal*, 596, 1142
- Cruces M., et al., 2021, *Monthly Notices of the Royal Astronomical Society*, 500, 448
- Eatough R. P., Keane E. F., Lyne A. G., 2009, *MNRAS*, 395, 410
- Ester M., Kriegel H.-P., Sander J., Xu X., et al., 1996, in *kdd*. pp 226–231
- Farah W., et al., 2018, *Monthly Notices of the Royal Astronomical Society*, 478, 1209
- Foster G., et al., 2018, *Monthly Notices of the Royal Astronomical Society*, 481, 2612
- Gourdji K., Michilli D., Spitler L., Hessels J., Seymour A., Cordes J., Chatterjee S., 2019, *The Astrophysical Journal Letters*, 877, L19
- Jiang J.-C., et al., 2022, *Research in Astronomy and Astrophysics*, 22, 124003
- Khan K., Rehman S. U., Aziz K., Fong S., Sarasvady S., 2014, in *The fifth international conference on the applications of digital information and web technologies (ICADIWT 2014)*. pp 232–238
- Li D., et al., 2021, *Nature*, 598, 267
- Lyubarsky Y., 2021, *Universe*, 7, 56
- Mallat S., 1999, *A wavelet tour of signal processing*. Elsevier
- Men Y., Barr E., 2024, *Astronomy & Astrophysics*, 683, A183
- Men Y., et al., 2019, *Monthly Notices of the Royal Astronomical Society*, 488, 3957
- Niu J.-R., et al., 2022, *Research in Astronomy and Astrophysics*, 22, 124004
- Offringa A., De Bruyn A., Biehl M., Zaroubi S., Bernardi G., Pandey V., 2010, *Monthly Notices of the Royal Astronomical Society*, 405, 155
- Pang D., Goseva-Popstojanova K., Devine T., McLaughlin M., 2018, *Monthly Notices of the Royal Astronomical Society*, 480, 3302

- Petroff E., Hessels J., Lorimer D., 2019, *The Astronomy and Astrophysics Review*, 27, 1
- Petroff E., Hessels J., Lorimer D., 2022, *The Astronomy and Astrophysics Review*, 30, 2
- Pleunis Z., et al., 2021, *The Astrophysical Journal*, 923, 1
- Rafiei-Ravandi M., Smith K. M., 2023, *The Astrophysical Journal Supplement Series*, 265, 62
- Rajwade K., et al., 2020, *Monthly Notices of the Royal Astronomical Society*, 495, 3551
- Sand K. R., et al., 2022, *The Astrophysical Journal*, 932, 98
- Sander J., Ester M., Kriegel H.-P., Xu X., 1998, *Data mining and knowledge discovery*, 2, 169
- Schubert E., Sander J., Ester M., Kriegel H. P., Xu X., 2017, *ACM Transactions on Database Systems (TODS)*, 42, 1
- Spitler L., et al., 2014, *The Astrophysical Journal*, 790, 101
- Starczewski A., Goetzen P., Er M. J., 2020, *Journal of Artificial Intelligence and Soft Computing Research*, 10
- Strang G., Nguyen T., 1996, *Wavelets and filter banks*. SIAM
- Yuan M., et al., 2022, *Monthly Notices of the Royal Astronomical Society*, 513, 4787
- Zackay B., Ofek E. O., 2017, *The Astrophysical Journal*, 835, 11
- Zhang B., 2020, *Nature*, 587, 45
- Zhang B., 2023, *Reviews of Modern Physics*, 95, 035005
- Zhang Y. G., Gajjar V., Foster G., Siemion A., Cordes J., Law C., Wang Y., 2018, *The Astrophysical Journal*, 866, 149
- Zhang Y.-K., et al., 2022, *Research in Astronomy and Astrophysics*, 22, 124002
- Zhou D., et al., 2022, *Research in Astronomy and Astrophysics*, 22, 124001
- Zhu W., et al., 2023, *Science Advances*, 9, eadf6198

This paper has been typeset from a $\text{\LaTeX}$ file prepared by the author.