

2021

Searching Harder, Localizing Better, Classifying Faster: Optimizing Fast Radio Burst Detection And Analysis

Kshitij Aggarwal

West Virginia University, ka0064@mix.wvu.edu

Follow this and additional works at: <https://researchrepository.wvu.edu/etd>



Part of the [Data Science Commons](#), [External Galaxies Commons](#), [Other Astrophysics and Astronomy Commons](#), and the [Stars, Interstellar Medium and the Galaxy Commons](#)

Recommended Citation

Aggarwal, Kshitij, "Searching Harder, Localizing Better, Classifying Faster: Optimizing Fast Radio Burst Detection And Analysis" (2021). *Graduate Theses, Dissertations, and Problem Reports*. 10249.
<https://researchrepository.wvu.edu/etd/10249>

This Dissertation is protected by copyright and/or related rights. It has been brought to you by the The Research Repository @ WVU with permission from the rights-holder(s). You are free to use this Dissertation in any way that is permitted by the copyright and related rights legislation that applies to your use. For other uses you must obtain permission from the rights-holder(s) directly, unless additional rights are indicated by a Creative Commons license in the record and/ or on the work itself. This Dissertation has been accepted for inclusion in WVU Graduate Theses, Dissertations, and Problem Reports collection by an authorized administrator of The Research Repository @ WVU. For more information, please contact researchrepository@mail.wvu.edu.

2021

Searching Harder, Localizing Better, Classifying Faster: Optimizing Fast Radio Burst Detection And Analysis

Kshitij Aggarwal

Follow this and additional works at: <https://researchrepository.wvu.edu/etd>



Part of the [Data Science Commons](#), [External Galaxies Commons](#), [Other Astrophysics and Astronomy Commons](#), and the [Stars, Interstellar Medium and the Galaxy Commons](#)

Searching Harder, Localizing Better, Classifying Faster: Optimizing Fast Radio Burst Detection And Analysis

Kshitij Aggarwal

Dissertation Submitted to
The Eberly College of Arts and Sciences
at West Virginia University
in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy
in
Physics

Sarah Burke-Spolaor, Ph.D., Chair
Maura McLaughlin, Ph.D.
Duncan R. Lorimer, Ph.D.
Casey J. Law, Ph.D.

Morgantown, West Virginia, USA
2021

Keywords: Radio bursts, Astronomy data analysis, Extragalactic radio sources,
Radio transient sources
Copyright 2021 Kshitij Aggarwal

Abstract

Searching Harder, Localizing Better, Classifying Faster: Optimizing Fast Radio Burst Detection And Analysis

Kshitij Aggarwal

Fast Radio Bursts (or FRBs) are millisecond-duration transients of extragalactic origin. They exhibit dispersion caused by propagation through an ionized medium, and quantified by Dispersion Measure (DM). Around 800 FRBs (24 repeaters) have been discovered; so far, 24 FRBs have been confidently associated with a host galaxy. In this thesis, we discuss multiple new FRB search and analysis techniques and the corresponding tools that enable us to search for FRBs harder, localize them better, and classify candidates faster.

We discuss five open-source software suites that can be used in FRB analysis. These suites are used to distinguish between FRBs and radio frequency interference (RFI), model FRB properties, search for periodic activity, calculate the probability of an association between an FRB and the host galaxy, and unify data processing across multiple data formats.

We then present a robust comparative analysis of clustering algorithms to group candidates from REALFAST transient search system at the Karl G. Jansky Very Large Array. We design a performance metric that optimizes for a few pure clusters, i.e., clusters with either astrophysical or noise candidates. We show that using sky location along with DM/time improves clustering performance, and propose a strategy that can be used to decide which clustering algorithm is most fit for a particular application.

We present a dense sample of bursts from the repeating FRB 121102, discovered using our software. Using the Arecibo Telescope, we detected 133 bursts in 3 hours of data observed at 1.4 GHz. We determine the properties of the bursts using robust spectro-temporal modeling. We find that the bursts are band-limited, with a lack of emission below 1.3 GHz. We find the wait time distribution to be log-normal in form with a peak at 75 s. Poissonian and Weibull distributions do not describe the burst rate distribution well. The cumulative energy distribution can be described using a broken power-law model, with the break at $(2.3 \pm 0.2) \times 10^{37}$ ergs and a high-energy slope of -1.8 ± 0.2 .

Motivated by the banded nature of FRB 121102 bursts, we perform a simulation study to show that commonly used analyses of band-limited FRBs lead to observational biases. We show that all the observed shapes in the energy distributions of repeaters can be explained using these biases. We then recommend techniques to correct these biases: modeling burst spectra to robustly estimate the intrinsic properties, and using bursts that are within the observing band for energy distribution analyses.

Finally, we discuss the REALFAST search and analysis pipeline, compare it to the search pipelines on single-dish telescopes, and highlight the advantages of using an interferometer. Primarily, every detection with REALFAST comes with a precise localization that can be used to associate the FRB to a host galaxy. We then discuss five repeating FRBs that were localized using REALFAST.

To my parents,
Mukul and Bhawna Aggarwal
For your unconditional love and
support

Acknowledgements

Graduate school is like a wildlife safari. The goal is clear (explore the forest, find wild animals, and not get lost), but the path is not. It was my research family that enlightened me to forge this path towards the goal. This thesis was not possible without their constant support, guidance, and motivation. I am indebted to them for where I am today.

I am foremost thankful to my awesome supervisor Sarah Burke-Spolaor. You consistently provided me a way forward especially during the last two years. You always encouraged my ideas, motivated me to be independent, and supported my goals. You gave me the necessary wisdom when I was taken by emotions at various stages during my PhD. You taught me how to do research, how to write papers, and how to navigate various situations in life. I admire you a lot and try to incorporate your teachings into my life. You gave me opportunities for leadership and research when I doubted myself. You saw something in me, I still am trying to find, and brought out the best in it. I cannot thank you enough.

I am grateful to my thesis committee; Casey Law, Duncan Lorimer, and Maura McLaughlin. They not only guided the trajectory of my thesis but also helped me numerous times individually. Casey, you always patiently helped me with all the nagging questions I asked about the REALFAST pipeline, interferometry, and programming. Dunc, your knowledge, experience, and personality inspired me to be the best version of myself. Maura, your wisdom and attention to detail motivated me to go the extra mile in whatever I did. The constant support that all of you provided, gave me the confidence to tackle challenges I would otherwise consider too difficult.

The friend I am most thankful to is Devansh Agarwal, without whom this thesis was impossible. Working with you has been one of the best experiences of my graduate school. Our numerous discussions and arguments about wide-ranging concepts were more enriching than I could have ever imagined. Thank you for the amazing food, the crazy trips, your expert advice, everything! We answered so many research problems, taught each other complex concepts, wrote thousands of lines of code together, but we still could never answer the question “What to cook for dinner tonight?”.

Next, I would like to thank Shalabh Singh, my constant supporter, my bestie. You were always there to listen to me when times were tough and kept me humble when they were good. You were never afraid to convey the hard-to-swallow pills but also helped me navigate my way through and not get lost. You pique my intellectual self, made me into a mature thinker, and always showed me the broader picture. Your wisdom about life and Machine Learning was always ready to rescue me. Again, I cannot thank you enough for all the past years and look forward to a lifetime of friendship.

Joseph and Reshma, you both are responsible for my sanity today. The last two years were extremely hard on all of us, but you both made them just a little better, and that means the world to me. Those long walks kept me going on days when nothing else seemed to behave. Those weekend hiking trips in different parks energized me for another week. Those Friday night take-outs and movies gave me the much-needed respite after the chaos of writing and debugging. I sincerely thank you

both for that. Oh and also for always attending the FRB journal club! :P

Caitlin, thank you for always being there to support me, helping me navigate difficult conversations, and being there when I needed to vent out. Nihan and Brent, thanks for your great companionship and advice. Thanks, Haley for checking in on me at various times and for sending pictures of Fern. I would like to thank my friends and colleagues Amber, Belinda, Josh, Pranav, Tim, Will, Greg, Megan, Jessica, Evan, Harsha, Paul, Mitch. You all made coming to White Hall a pleasant experience. I will forever cherish those discussions and will miss all the gossip. and Paul Brook, I hope you find that attic someday! :)

Kevin, Dunc, Sarah, Loren, DJ, Maura, thanks for the amazing courses and discussions. I am also grateful to many people for their research advice and help: Nihan Pol, Liam Connor, Paul Demorest, Joe Lazio, Paul Baker, Michael Lam, Justin Linford, Harsha Blumer, Kevin Bandura. I also thank REALFAST collaboration for encouraging me and giving me opportunities to grow as a researcher.

Nate, each and every analysis in this thesis was enabled by you. You kept bowser (our computing cluster) running. You made sure that enough resources were available for everyone. You promptly fixed all the weird issues in GPUs and nodes, always found the missing library files, and somehow figured a way out of those software compatibility errors. I would also like to thank Joe Glaser and the high-performance computing staff at WVU for maintaining the computing facilities that I relied heavily on for my research. A general thanks to the staff and scientists at the Very Large Array, Green Bank Observatory, and Arecibo Observatory, for their help in observations, data transfer, and processing.

I would like to thank the office staff especially Viola Bryant for efficiently managing all the paperwork. A special thanks to Margaret Mattson for handling the travel paperwork and reimbursements (and thanks for patiently modifying my trip at the last minute, because I forgot to obtain the transit visa). A big thanks to the Caruth Center at WVU for helping me become a balanced person, cope with burnout, and form deeper relationships with people around me. Seeking help for mental health was not an easy step, but I am so glad I did.

I would like to express gratitude towards my undergraduate research mentors: Ravi Subrahmanyam and Yashwant Gupta. Ravi, you gave me, a sophomore electrical engineering student, a chance and introduced me to astronomy research. YG, you introduced me to the amazing world of fast radio transients.

I am grateful to Sushant and Rashmi for providing me with a family away from home. Thanks for all that amazing homemade food and for inviting me to all the festivals and celebrations. The homely environment and love you gave me all these years mean a lot to me.

I am also thankful to all the kids I interacted with in the past years. You recharged my brain and filled it with joy but also made me feel old. A big thanks to arXiv and Github for providing an open platform to access and share research papers and software.

And finally, I am forever grateful to my loving and caring family. Shrey, I thank you for being the greatest sibling one could ever ask for. Mom and dad, you have supported me in every decision I have taken, even when I decided to move to

the other side of the world for this PhD. You always fueled my passion for science and astronomy, motivated me to always dream big, inspired me to work hard, and supported me to rise up from my failures. What I am today would not be possible without your love, support, and care. I owe it all to you both and dedicate this thesis to you two.

Table of Contents

List of Tables	xii
List of Figures	xiii
1 Introduction to Fast Radio Transients	1
1.1 Basic terminology	2
1.2 Propagation effects	4
1.3 Fast radio bursts	5
1.3.1 Brief history of FRBs	5
1.4 What we observe	8
1.4.1 Dispersion measure	8
1.4.2 Scattering	9
1.4.3 Pulse width	9
1.4.4 Flux density and fluence	10
1.4.4.1 Radiometer equation	10
1.4.5 Repetition	11
1.4.6 Periodicity	12
1.4.7 Repeaters vs non-repeaters	14
1.5 What we derive	15
1.5.1 DM derived distance	15
1.5.2 Luminosity and energy	15
1.5.3 FRB rate and source counts	16
1.6 How to search for FRBs?	18
1.6.1 Data acquisition	18
1.6.2 Single-Pulse search	19
1.6.2.1 RFI mitigation	20
1.6.2.2 Dedispersion and averaging	22
1.6.2.3 Baseline estimation	23
1.6.2.4 Normalization	24
1.6.2.5 Matched filtering	25
1.6.3 Clustering	25
1.6.4 Classification	27
1.6.5 Sensitivity analysis of search	27
1.7 Localization of FRBs	28
1.7.1 Host galaxy population and local environment	29
1.7.2 Persistent radio emission	30
1.7.3 Multi-wavelength emission	31
1.8 FRB progenitor models	32
1.8.1 Black Hole models	33
1.8.2 Young magnetar models	33
1.9 Open questions	34
1.10 Thesis outline	35

2	Software	36
2.1	Fast Extragalactic Transient Candidate Hunter (FETCH)	36
2.2	Burstfit	37
2.3	FRB Periodicity Analysis (FRBPA)	38
2.4	Calculating Association Probability of FRBs (CASP)	38
2.5	Your Unified Reader (Your)	39
2.5.1	Statement of need	39
3	Robust Assessment of Clustering Methods for Fast Radio Transient Candidates	41
3.1	Chapter Overview	41
3.2	Introduction	42
3.3	Clustering	44
3.3.1	Expected number of candidates from a single astrophysical event	45
3.3.2	Unsupervised Clustering	46
3.3.3	Clustering RFI	47
3.4	Data	47
3.4.1	Feature Selection for Clustering	48
3.4.2	RFI database	48
3.4.3	Simulating and Injecting FRBs	49
3.4.4	Test Dataset	49
3.4.5	Pre-processing	50
3.5	Methods	51
3.5.1	Clustering Algorithms	51
3.5.2	Hyperparameter Tuning	51
3.5.3	Performance Metric	52
3.5.3.1	Homogeneity	53
3.5.3.2	Completeness	54
3.5.3.3	V-measure	54
3.5.3.4	Recall	55
3.5.3.5	Score	55
3.5.4	Advantages of this metric	55
3.6	Results	56
3.6.1	Optimal Hyperparameters	56
3.6.2	Effect of data processing	56
3.6.3	Evaluating performance on clean data	60
3.6.3.1	Completeness on Clean data	60
3.6.4	Benchmarking	62
3.7	Discussion	62
3.7.1	Feature Importance	62
3.7.2	What if I only use DM and time for clustering?	64
3.7.3	But which algorithm should I use?	66
3.8	Conclusions	68
3.9	Appendix	69
3.9.1	Clustering Algorithms	69
3.9.1.1	K-means	69

	3.9.1.2	Mean Shift	69
	3.9.1.3	Affinity Propagation	70
	3.9.1.4	Agglomerative clustering	71
	3.9.1.5	DBSCAN	71
	3.9.1.6	Optics	72
	3.9.1.7	HDBSCAN	73
	3.9.1.8	Birch	73
	3.9.2	Parameter Ranges for Hyperparameter tuning	74
4		Comprehensive analysis of a dense sample of FRB 121102 bursts	76
4.1		Chapter Overview	76
4.2		Introduction	77
4.3		Data	78
4.4		Methods	79
	4.4.1	The Petabyte Project	79
	4.4.2	Single-Pulse Search	80
	4.4.3	Completeness Limit	81
	4.4.4	Spectro-temporal Burst Modelling	82
	4.4.5	BURSTFIT	84
	4.4.5.1	Data Preparation	84
	4.4.5.2	Stage 1: Single-component Fitting	84
	4.4.5.3	Stage 2: Statistical Tests	85
	4.4.5.4	Stage 3: Multi-component Fitting	86
	4.4.5.5	Stage 4: MCMC	86
	4.4.5.6	Handling data saturation	87
	4.4.5.7	Caveats to our fitting analysis	88
4.5		Results	90
	4.5.1	Burst sample properties	91
4.6		Discussion	96
	4.6.1	Cumulative energy distribution	96
	4.6.1.1	Testing for a high-energy break	98
	4.6.2	Wait-time distribution	100
	4.6.3	Short-period periodicity search	101
	4.6.3.1	Difference Search	101
	4.6.3.2	Fast Folding Algorithm	103
	4.6.3.3	frbpa	104
	4.6.3.4	Lomb-Scargle	105
	4.6.4	Burst Rate	105
	4.6.5	Implications for progenitor models	107
	4.6.6	Comparison to previous work	107
	4.6.6.1	Threshold signal-to-noise ratio	110
	4.6.6.2	Single-pulse search software	111
	4.6.6.3	RFI mitigation and classification	112
	4.6.6.4	General comments	113
	4.6.7	Caveats	114

4.7	Conclusions	114
4.8	Appendix	115
5	Observational effects of banded repeating FRBs	121
5.1	Chapter Overview	121
5.2	Introduction	122
5.3	Methods	123
5.3.1	Generating a population of repeater bursts	123
5.3.2	Applying selection effects	124
5.3.3	Estimating Fluence and bandwidth	125
5.4	Results	126
5.4.1	Cumulative Energy Distribution	127
5.4.1.1	Challenges	129
5.4.1.2	Using in-band bursts	129
5.4.2	Energy Distribution	130
5.4.3	Distribution of spectral parameters	130
5.4.4	Calculating Energy	132
5.5	Discussion	135
5.5.1	Observing bandwidth and sub banded searches	135
5.5.2	Calculating fluences and bandwidths	136
5.5.3	Estimating intrinsic properties	136
5.5.4	Effects of power-law slope	137
5.6	Conclusions	137
6	Localizing FRBs with Realfast	140
6.1	Chapter Overview	140
6.2	Interferometers vs Single Dish Telescope	140
6.3	Transient search using interferometers	141
6.4	REALFAST pipeline	142
6.4.1	Correlator	144
6.4.2	Data Segment	144
6.4.3	Calibration	144
6.4.4	RFI Mitigation	145
6.4.5	Mean visibility subtraction	145
6.4.6	Dedisperse and Resample	145
6.4.7	Imaging: project baselines, grid visibilities, Fourier transform	146
6.4.8	Clustering	147
6.4.9	Visualizations and Classification	147
6.4.10	Data Cutout	147
6.4.11	Candidate Database and Portal	149
6.5	REALFAST Post-Processing	149
6.5.1	Refinement of properties	149
6.5.2	CASA calibration and Imaging	151
6.5.3	Deep imaging	151
6.5.4	Astrometry	152

6.6	REALFAST Localizations	152
6.6.1	FRB 190614D	154
6.6.2	FRB 190520	154
6.6.3	FRB 180301	155
6.6.4	FRB 20201124A	156
6.7	VLA/REALFAST detection of burst from FRB180916.J0158+65 and Tests for Periodic Activity	157
6.7.1	Realfast Detection of FRB 180916	157
6.7.2	Periodicity Analysis Techniques	158
6.7.3	Results and Discussion	159
7	Conclusion	161
7.1	Open source software	161
7.2	Clustering REALFAST candidates	162
7.3	Analysis of FRB 121102 bursts	163
7.4	Biases due to banded nature of repeater bursts	164
7.5	Localizing bursts using REALFAST	166
7.6	Past and future of FRBs	167

List of Tables

3.1	Parameter distributions of simulated FRBs	49
3.2	Optimal hyperparameters	58
3.3	Hyperparameter ranges for different clustering algorithms.	75
4.1	Priors used in the MCMC fitting.	88
4.2	Properties of the first 10 bursts	90
4.3	Results from the burst sample analysis	94
4.4	Fitted burst rate distributions.	107
4.5	Properties of all the bursts.	116
6.1	Properties of REALFAST FRBs	153

List of Figures

1.1	Example of a spectrogram and dedispersion	3
1.2	Pulse Profiles with the same fluence but different widths.	10
1.3	Dynamic spectra of FRB 121102 bursts	13
1.4	Flow diagram of a single-pulse search pipeline	18
1.5	Data acquisition system	19
1.6	Single-Pulse Search pipeline	20
1.7	Example of dedispersion at multiple DM trials.	24
1.8	Example of matched filtering	26
3.1	Number of candidates vs S/N	46
3.2	Maximum score at optimal hyperparameters	57
3.3	Violin plots of score vs. pre-processing cases	59
3.4	Completeness of different algorithms on clean data	61
3.5	Time taken to cluster for each algorithm	63
3.6	Importance of each feature	64
3.7	Score vs. algorithms for two feature clustering	66
4.1	Dynamic spectra of six FRB 121102 bursts	81
4.2	RFI fraction	82
4.3	Plots of spectra and fits in case of data saturation	89
4.4	Flowchart of fitting stages in BURSTFIT	89
4.5	Results of spectro-temporal fits on two bursts.	91
4.6	Corner plot of MCMC samples obtained after fitting B67.	92
4.7	Corner plot generated using MCMC samples for all the bursts	95
4.8	Fitted burst property versus the fitted burst time	96
4.9	Cumulative energy distribution of bursts	99
4.10	Wait time distribution	102
4.11	The cumulative wait time distributions	108
4.12	Burst Rate with respect to minimum wait time	108
4.13	Normalized histogram of burst properties.	110
5.1	Figure showing some simulated spectra.	124
5.2	Energy distributions of bursts detected at varying fluence thresholds .	128
5.3	Distribution of mean and standard deviation of burst spectra.	131
5.4	Energy distribution of FAST FRB 121102 bursts.	134
5.5	Energy distributions of repeater bursts.	134
6.1	REALFAST transient search pipeline	143
6.2	REALFAST plot of FRB 180916	148
6.3	REALFAST refinement plot of FRB 180916	150
6.4	Burst detection and observation histograms for FRB 180916	160

Chapter 1

Introduction to Fast Radio Transients

Fast radio transients, as the name suggests, refer to a class of astronomical events that satisfy the following two properties: they are “fast”, i.e., duration less than 1 second, and detectable at radio frequencies (Lorimer & Kramer, 2004). These can be further classified into several commonly discussed varieties, e.g. radio pulsars, rotating radio transients (RRATs), and fast radio bursts (FRBs). All of these transient classes share some similar observable characteristics, for instance their pulse durations and emission frequency. The origin of radio pulsars and RRATs is understood to be neutron stars (Lorimer & Kramer, 2004), while for FRBs that is still debatable (e.g. Petroff et al., 2019). Pulsars are rapidly rotating neutron stars that emit a radio pulse at every rotation. RRATs are a type of pulsars, but not all the pulses are detectable from them (Burke-Spolaor & Bailes, 2010). Therefore, it appears as if they emit single-pulses separated by minutes to hours (McLaughlin et al., 2006). Pulsars have been detected in the Milky Way, nearby globular clusters and Magellanic clouds. The spectral luminosity of pulsars and RRATs typically ranges between 10^{16} to $10^{21} \text{ erg s}^{-1} \text{ Hz}^{-1}$, making it impossible to detect these sources at cosmological distances with the current radio telescopes. FRBs, on the other hand, appear to be the brighter cousins of pulsars and RRATs, with spectral luminosities greater than $10^{27} \text{ erg s}^{-1} \text{ Hz}^{-1}$ (Petroff et al., 2021). But unlike pulsars and RRATs, (repeating) FRBs have not yet shown any underlying periodicity in consecutive bursts. All of these transients have been detected at radio frequencies between 100 MHz and 8 GHz (Lorimer & Kramer, 2004; Pastor-Marazuela et al., 2021; Gajjar et al., 2018; Petroff et al., 2021; Cordes & Chatterjee, 2019; CHIME/FRB Collaboration et al., 2019).

This chapter aims to introduce and summarise the very broad and rapidly evolving field of FRBs. We will touch on the basic concepts while also covering the broad width of the latest results. To limit the length of this chapter, we do not dwell on the depth of different results and instead provide relevant references that a curious reader can use for further reading.

1.1 Basic terminology

Before discussing the observable properties of radio transients and then diving deeper into a discussion of FRBs, it is helpful to define the standard terminology that will be repeatedly used throughout this chapter and in the thesis.

1. Frequency and time: Frequency refers to the radio frequency (or frequency channel) at which the observations were performed. Time refers to the time (or time sample) when the observations were performed.
2. Spectrogram: A two-dimensional data array, consisting of intensity (radio power) at each time sample and frequency channel. We show an example of this in the bottom two panels on the left in Figure 1.1 (see Section 1.6.1).
3. Time series and pulse profile: The frequency-averaged array of intensities as a function of time is referred to as a time series (top row in Figure 1.1). When this time series is made for data consisting of a pulsar or FRB signal, that pulse might also be seen in the time series. Suppose the effect of dispersion is first corrected (see Section 1.2) before the time series is generated. In that case, the pulse might be visible in the time series and is referred to as a pulse profile (top right plot in Figure 1.1).
4. Bandwidth, bandpass, and spectra: Bandwidth is the range of contiguous frequencies that are observed. The time-averaged array of intensities generated from the spectrogram is called the bandpass and it shows the response of the telescope receiver across the bandwidth. When this time-averaging is done on a dedispersed spectrogram with only the burst signal, then the pulse intensity

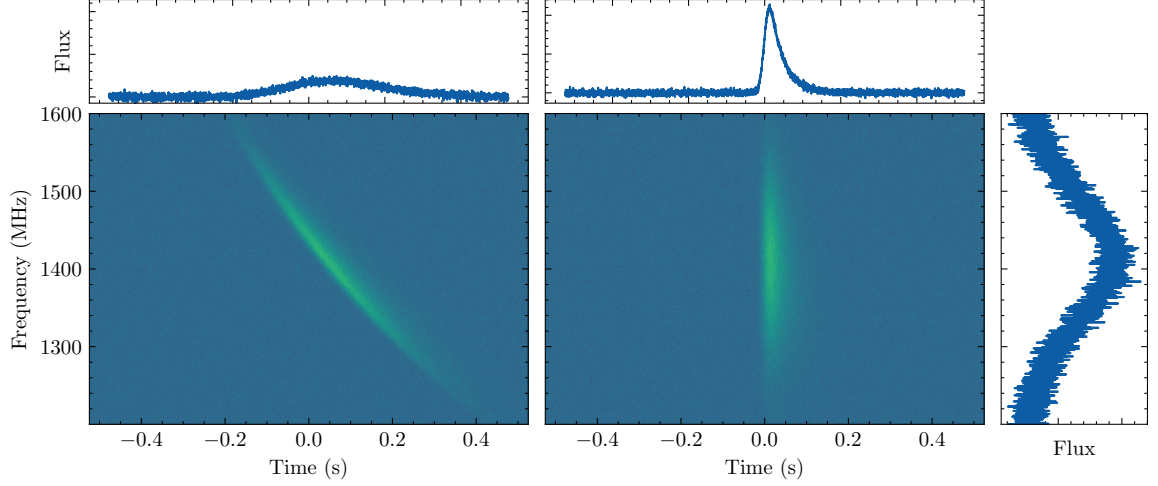


Figure 1.1: Figure showing a simple simulated spectrogram. The two-dimensional panels in the lower row represent the spectrogram of data, with frequency on the vertical axis and time on the horizontal axis. The dispersed signal can be seen in the left spectrogram, while the signal in the right spectrogram has been corrected for dispersion or dedispersed. The one-dimensional plots (or time-series) in the top row have been generated by averaging the spectrogram along the frequency axis and showing the burst profile. The one-dimensional plot on the right shows the spectra of the signal generated by time-averaging the dedispersed burst signal. Figure generated using BURSTFIT (Aggarwal et al., 2021a), also introduced in Chapter 4.

with respect to frequency can be observed. We refer to this as the spectrum (right-most plot in Figure 1.1).

5. Signal or burst: The astrophysical signal or intensities originating from a pulsar or FRB in the data.
6. Signal-to-noise ratio: It is defined as the ratio of the signal intensity to that of the standard deviation of data without the signal, typically abbreviated as S/N. It can be used to estimate the statistical significance of a signal. A signal with higher S/N has lesser chance probability of occurrence due to random noise, and would therefore have higher probability of being astrophysical.

1.2 Propagation effects

A transient signal, traveling through the cold ionized plasma of the interstellar medium (ISM) or intergalactic medium (IGM), undergoes four observable effects: dispersion, scattering, scintillation and Faraday rotation (Lorimer & Kramer, 2004). We discuss the first two effects here. FRB emission is dispersed in time, where the pulse arrival time varies as an inverse square relation with frequency. This introduces a frequency-dependent delay in the FRB signal, that can be written as:

$$\Delta t(\text{ms}) \approx 4.15 \times (\nu_{\text{lo}}^{-2} - \nu_{\text{hi}}^{-2}) \times \text{DM}, \quad (1.1)$$

where $\Delta t(\text{ms})$ is the time delay across the observing band with ν_{lo}^{-2} and ν_{hi}^{-2} as the low and high frequency of the band (in GHz). Here, DM is called the dispersion measure and can be defined as the integral of the electron density to the source at a distance (d), i.e.,:

$$\text{DM} = \int_0^d n_e \, ds, \quad (1.2)$$

where n_e is the free electron density along the path traveled by the signal. DM is often quoted using the units of $\text{cm}^{-3} \text{pc}$ and can be used to derive valuable insights about the distance and origin of the pulse. The left and middle plot in the bottom row of Figure 1.1 show the spectrogram with dispersion and after the data has been corrected for dispersion or “dedispersed”.

Fluctuations in the electron density in the ISM can lead to radio waves getting deflected. Due to this, waves originating from the same source can take paths of multiple lengths to reach the observer on Earth. Hence, some waves arrive later as they travel along a longer path. This is called scattering. In the simple case of a thin, infinite scattering screen (Scheuer, 1968), this effect can be approximated to convolving a one-sided exponential to the FRB pulse. In this case, the decay time of the exponential tail (τ) scales with the frequency (ν) as,

$$\tau \propto \nu^{-4}. \quad (1.3)$$

The exponential tail in the right pulse profile in Figure 1.1 shows the effect of scattering. The scattering index (-4) depends on the scale of turbulence in the ISM. In reality, the measured scattering indices can vary significantly from the fiducial (-4) index.

1.3 Fast radio bursts

Fast radio bursts (FRBs) are bright (few Janskys), millisecond-duration radio transients of extragalactic origin (Lorimer et al., 2007). We discuss the properties of FRBs in Section 1.4. Briefly, the dispersion of FRB signals is greater than what can be accounted for by the free electrons in the Milky Way disk and its halo, using the models of the Milky Way galaxy (Cordes & Lazio, 2002; Yao et al., 2017; Prochaska & Zheng, 2019). Therefore, FRBs are extragalactic in origin and the source of this residual dispersion is attributed to the intergalactic electrons (electrons between galaxies) and possibly to the host galaxy of the FRB (see Section 1.7).

1.3.1 Brief history of FRBs

FRBs were discovered in 2007, with the discovery of FRB 010724 or the “Lorimer Burst” during a single-pulse search of archival data from Parkes Radio Telescope observing the Small Magellanic Cloud (SMC; Lorimer et al., 2007). Expecting that the dispersion measure of sources in SMC would be higher than the DMs of Galactic pulsars, this search included DMs much higher than were typically used at that time (see Section 1.6 for details on single-pulse searches). FRB 010724 was serendipitously detected as an extremely bright pulse, with a luminosity much higher than seen from pulsars and with a DM much higher than that expected from Milky Way along that line of sight. This was soon followed by detection of more such bursts, but again with data from Parkes Telescope (Thornton et al., 2013), leading to speculations regarding the astrophysical nature of these sources. This was put to rest by the detection of FRB 121102 with Arecibo Telescope, again with large DM, millisecond duration, and large flux, characteristic of the bursts detected with Parkes Telescope (Spitler et al.,

2014).

Hundreds of hours of follow-up observations of these FRBs were performed with several radio and multi-wavelength telescopes, but no repeat burst or multi-wavelength emission was detected (Lorimer et al., 2007; Petroff et al., 2015). Then, Spitler et al. (2016) reported detection of additional bursts from FRB 121102 during follow-up observations, making it the first (of now many) repeating FRBs and one of the most studied FRBs to date. This provided the valuable opportunity to follow up this FRB and study it in greater detail. Using follow-up observations with the Jansky Very Large Array (VLA) and a novel millisecond imaging search technique (see Chapter 6 for more details), it was detected and localized to a dwarf galaxy at a redshift of 0.19, solidifying the extragalactic origin of FRBs. This localization also led to the detection of persistent radio emission co-located with this FRB (Chatterjee et al., 2017; Tendulkar et al., 2017; Marcote et al., 2017).

The following years witnessed the commissioning of commensal FRB search systems, i.e., systems that piggy-banked on the normal observations at the radio telescope and continuously searched for FRBs. Some such systems are: GREENBURST at GBT (Agarwal et al., 2020b), ASKAP (Shannon et al., 2018), CHIME/FRB (CHIME/FRB Collaboration et al., 2018), REALFAST at VLA (Law et al., 2018, 2015), ALFABURST at Arecibo (Foster et al., 2018), Deep Synoptic Array (DSA) (Hallinan et al., 2019), etc. This increased the pace of FRB discoveries, with the discovery of more repeating and non-repeating FRBs, driven primarily by large field-of-view telescopes like CHIME and ASKAP (Amiri et al., 2019; CHIME/FRB Collaboration et al., 2019; Fonseca et al., 2020; Shannon et al., 2018; Bannister et al., 2019). Then, two repeating sources, FRB 121102 and FRB 180916, were found to exhibit periodicity in their burst activity (CHIME/FRB Collaboration et al., 2020a; Rajwade et al., 2020b), which was shown to be frequency-dependent in the case of FRB 180916 (Pastor-Marazuela et al., 2021; Pleunis et al., 2021a; Aggarwal et al., 2020).

This was followed by a wave of FRB localizations by ASKAP, DSA, and VLA that started to build a population of host galaxies (see Section 1.7 and Chapter 6;

Ravi et al., 2019; Law et al., 2020; Bannister et al., 2019). But even after more than a dozen localized FRBs, no other FRB was found to be co-located with a persistent radio source. The observed DMs and redshifts of the localized FRBs were also used to constrain the baryonic content of the universe (Macquart et al., 2020). The FRB-like burst from a Galactic magnetar (the soft gamma-ray repeater SGR 1935+2154) in April 2020, accompanied by a hard X-ray burst, gave conclusive evidence that at least a sub-population of FRBs could originate from magnetars (Bochenek et al., 2020; CHIME/FRB Collaboration et al., 2020b; Kirsten et al., 2021a; Li et al., 2021; Ridnaia et al., 2021; Mereghetti et al., 2020). Finally, the large sample of FRBs detected at 400–800 MHz by CHIME/FRB provided a large catalog of FRB sources, facilitating studies of FRB population (The CHIME/FRB Collaboration et al., 2021a; Pleunis et al., 2021b). Currently, around 800 non-repeating FRBs and 20 repeating FRBs have been discovered. Finally, very recently, another repeating FRB has been found to be co-located with a persistent radio source, with FRB and its host properties very similar to that of the first repeater, FRB 121102 (Niu et al., 2021a).

It is worth noting here that FRBs were discovered because Lorimer et al. (2007) decided to search their data up to a higher DM value than was typically done at the time. This was made possible by the advances in computing infrastructure, development of single-pulse search algorithms, improved understanding of pulsars in Milky Way, and most importantly by a desire to detect pulsars outside our Milky Way galaxy. This search opened a gateway to the world of highly dispersed pulses that originated from far away galaxies, leading to this exciting field of FRBs. This story exemplifies the process of advancement of human knowledge: small incremental innovations when accumulated over a long time lead to a revolutionary discovery or invention i.e., it converts something that was an “unknown unknown” into something that is now a “known unknown”.

1.4 What we observe

A seemingly simple FRB pulse demonstrated in Figure 1.1 consists of a wealth of information. In this section, we discuss some observable properties that can be extracted from a typical FRB detection. As will be apparent by the end of this section, FRBs show a wide variety in all these properties, making it difficult to understand their origins and classify them into sub-populations.

1.4.1 Dispersion measure

As discussed in Section 1.2, the FRB signal undergoes a frequency-dependent delay as it propagates through the ionized plasma on its way from the unknown source to the radio telescope on Earth. The DM (see Eq 1.4) acts as a proxy for the distance to the source. The DM of FRBs is higher than that expected from the ionized plasma in that line of sight from the Milky Way disk and halo (Lorimer et al., 2007; Prochaska & Zheng, 2019; Cordes & Lazio, 2002). FRBs have shown a wide range of DM values, ranging from around 90 to 3000 pc cm⁻³ (Bhardwaj et al., 2021a; The CHIME/FRB Collaboration et al., 2021a). The observed dispersion (or DM_{obs}) of the FRB (from a source at a cosmological redshift z) consists of contributions from the medium it encounters between the FRB source and the Earth. This includes (starting from FRB source and moving to Earth): Local environment of the FRB (DM_{local}), host galaxy and halo (DM_{host}), intergalactic medium (DM_{IGM}), foreground galaxies (DM_{fg}), Milky way halo (DM_{MWhalo}) and Milky Way disk (DM_{MW}) so that

$$\text{DM}_{\text{obs}}(z) = \text{DM}_{\text{MW}} + \text{DM}_{\text{MWhalo}} + \text{DM}_{\text{fg}}(z) + \text{DM}_{\text{IGM}}(z) + \frac{\text{DM}_{\text{host}} + \text{DM}_{\text{local}}}{(1+z)}. \quad (1.4)$$

Of these, DM_{MW} can be estimated using the electron density models of the Milky Way (Cordes & Lazio, 2002; Yao et al., 2017). DM_{MWhalo} has been estimated to be between 50 and 80 pc cm⁻³ (Prochaska & Zheng, 2019). DM_{host} is typically assumed to be 50 pc cm⁻³, while DM_{foreground} has been recently estimated for some localized FRBs (Prochaska et al., 2019; Simha et al., 2021). Finally, DM_{IGM} can be related to

FRB redshift and is discussed in Section 1.5.1.

1.4.2 Scattering

FRBs show a wide variety of scattering timescales (discussed in Section 1.2), ranging from no scattering to tens of milliseconds of scattering at 1.4 GHz. More interestingly, unlike pulsars, the scattering timescales of FRBs are not related to the Galactic latitudes (Cordes & Chatterjee, 2019). This implies that the majority of scattering in FRBs originates from extragalactic gas, but it is to be noted that only 30% of the FRBs show scattering (Cordes & Chatterjee, 2019). The large catalog of FRBs recently published by CHIME/FRB (The CHIME/FRB Collaboration et al., 2021a) reported evidence in favor of a large population of FRBs with scattering time larger than 10 ms at 600 MHz.

1.4.3 Pulse width

The observed pulse width is typically considered to be a convolution of intrinsic width and instrumental broadening factors. For a Gaussian pulse, it is given by:

$$W = \sqrt{W_{\text{int}}^2 + t_{\text{samp}}^2 + \Delta t_{\text{DM}}^2 + \Delta t_{\text{DMerr}}^2 + \tau_s^2}, \quad (1.5)$$

where t_{samp} is the sampling time of observation, Δt_{DM} is the dispersive delay across an individual frequency channel, Δt_{DMerr} represents the dispersive delay due to de-dispersion at a slightly incorrect DM, τ_s is the scattering timescale and W_{int} is the intrinsic width of the pulse. Pulse width is usually measured at 50% of the peak (Lorimer & Kramer, 2004), but for complex pulse structure, it is common to quote the equivalent width of a top-hat pulse with the same integrated flux. Figure 1.2 shows two pulses with the same integrated flux but different peak fluxes and widths. A top-hat pulse with the same integrated flux is overlaid on both, showing the equivalent width.

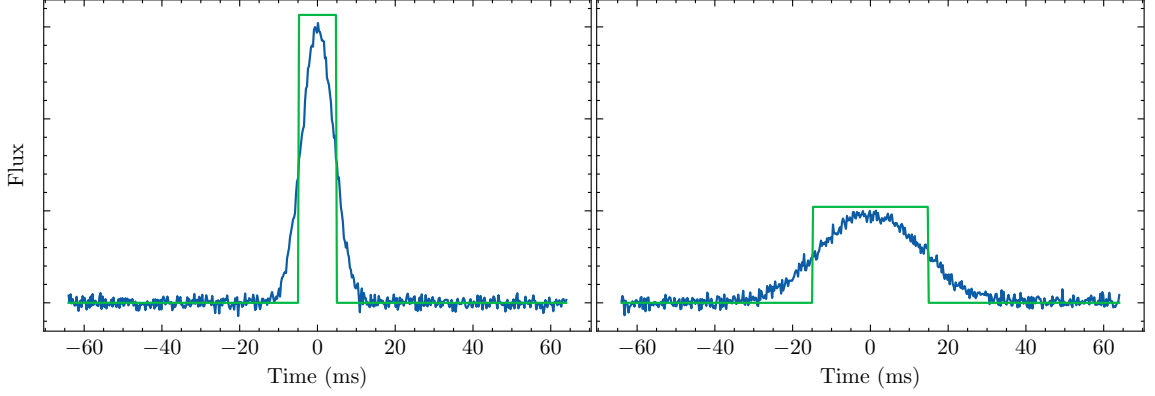


Figure 1.2: Two pulse profiles with the same integrated flux, i.e., fluence, but different pulse widths. Top-hat pulses with the same fluence are overlaid on both profiles, showing each real pulse’s equivalent width.

1.4.4 Flux density and fluence

Flux density is the spectral power received by a detector of unit projected area (Condon & Ransom, 2016). The commonly used units of flux density are Jansky, where,

$$1 \text{ Jy} = 10^{-26} \text{ W m}^{-2} \text{ Hz}^{-1}. \quad (1.6)$$

For pulsars and FRBs, flux and fluence are commonly reported properties. Fluence is defined as the total integrated flux, i.e

$$\mathcal{F} = S_{\text{peak}} W_{\text{eq}} = \int_{\text{pulse}} S(t) dt, \quad (1.7)$$

where, S_{peak} is the peak flux of the profile, W_{eq} is the equivalent width, and $S(t)$ is the pulse profile. Figure 1.2 shows two pulses with the same fluence but different peak fluxes and widths, along with top-hat pulses of the same fluence.

1.4.4.1 Radiometer equation

The radiometer equation can be used to estimate the expected root-mean-square fluctuation (σ_s) in the time series (Condon & Ransom, 2016; Lorimer & Kramer, 2004;

Petroff et al., 2019):

$$\sigma_s = \frac{T_{\text{sys}}}{G\sqrt{n_p\Delta\nu t_{\text{samp}}}}, \quad (1.8)$$

where T_{sys} is the system temperature (including sky temperature), G is the antenna gain, $\Delta\nu$ is the receiver bandwidth, n_p is the number of polarizations and t_{samp} is the data sampling interval.

The S/N of a pulse obtained from a dedispersed time series can then be converted into physically meaningful quantities (like flux density) by calibrating the time series. This is usually approximated by comparing the root-mean-square fluctuations of the time series to that obtained from radiometer equation.

1.4.5 Repetition

Most FRBs are found to be one-off events, i.e., even after hundreds of hours of follow-up search, no other repeat burst has been detected with properties similar to the original burst (Lorimer et al., 2007; Petroff et al., 2015; The CHIME/FRB Collaboration et al., 2021a). This limits the opportunities to study and follow-up these FRBs. But interestingly, so far, around 24 FRBs have been found to emit multiple bursts, i.e., repeat (Spitler et al., 2016; Petroff et al., 2021; CHIME/FRB Collaboration et al., 2019; The CHIME/FRB Collaboration et al., 2021a; Fonseca et al., 2020; Kumar et al., 2019). Repeaters, therefore, prove to be a valuable source to gain insight about FRBs, as radio and multi-wavelength follow-up detections can be accumulated with more observations. Hundreds of bursts from two of the most studied repeaters, FRB 121102 and FRB 180916, have been detected by more than a dozen different radio telescopes, and multi-wavelength upper limits have been placed using many other telescopes around the world (Law et al., 2017; Chatterjee et al., 2017; Cruces et al., 2020; Pastor-Marazuela et al., 2021; Li et al., 2021; CHIME/FRB Collaboration et al., 2020a; Scholz et al., 2020; Aggarwal et al., 2021a; Pearlman et al., 2020; Platts et al., 2021).

Follow-up observations have revealed that even repeaters show a variety of properties: periodicity in burst activity (see Section 1.4.6), bandlimited emission (see

Chapters 4 and 5; Pleunis et al., 2021b; Aggarwal et al., 2021a; Kumar et al., 2021c), extremely faint bursts (Kumar et al., 2019), a variety of host galaxies (Marcote et al., 2017; Marcote et al., 2020, also see Section 1.7.1), presence of a persistent radio counterpart (Chatterjee et al., 2017; Marcote et al., 2020; Law et al., 2021b), wide range of burst rates (CHIME/FRB Collaboration et al., 2019; Li et al., 2021; Pastor-Marazuela et al., 2021), burst detection from 100 MHz to 8 GHz (Pleunis et al., 2021a; Pastor-Marazuela et al., 2021; Gajjar et al., 2018; Zhang et al., 2018), non-Poissonian burst rate (Oppermann et al., 2018; Cruces et al., 2020; Aggarwal et al., 2021a), complex energy distribution (Aggarwal, 2021; Aggarwal et al., 2021a; Li et al., 2021; Pastor-Marazuela et al., 2021), frequency drift among components (or “sad trombone” effect; Hessels et al., 2019; Gajjar et al., 2018), wide range in observed and inferred host DMs (Tendulkar et al., 2017; Marcote et al., 2020; CHIME/FRB Collaboration et al., 2019), etc. Figure 1.3 shows the spectrogram of nine FRB 121102 bursts detected with Arecibo Telescope, adapted from Chapter 4. Even in this small sample, the variety in burst morphology can be easily observed.

In summary, although there is no clear consensus regarding the origin of repeating FRBs, repeaters are the ideal source for localization as discussed in Section 1.7 and Chapter 6. As a result, it is likely that the discovery and localization of more repeaters will contribute significantly to the understanding of the FRB population.

1.4.6 Periodicity

Motivated by the possibility of a neutron star origin of FRBs and drawing comparisons of FRB bursts to those from Galactic pulsars (especially giant pulses from young pulsars), periodicity searches are typically conducted on the bursts detected from repeating FRBs (Lorimer & Kramer, 2004; Lorimer et al., 2007; Petroff et al., 2019). Any underlying periodicity (millisecond to minutes duration) in the bursts would provide strong confidence in the neutron star models, at least for repeaters. So far, no repeater has shown any such millisecond to minutes scale periodicity in the bursts, even after extensive searches and a large number of burst detections (Li et al., 2021; Cruces et al., 2020; Pastor-Marazuela et al., 2021; Aggarwal et al., 2021a;

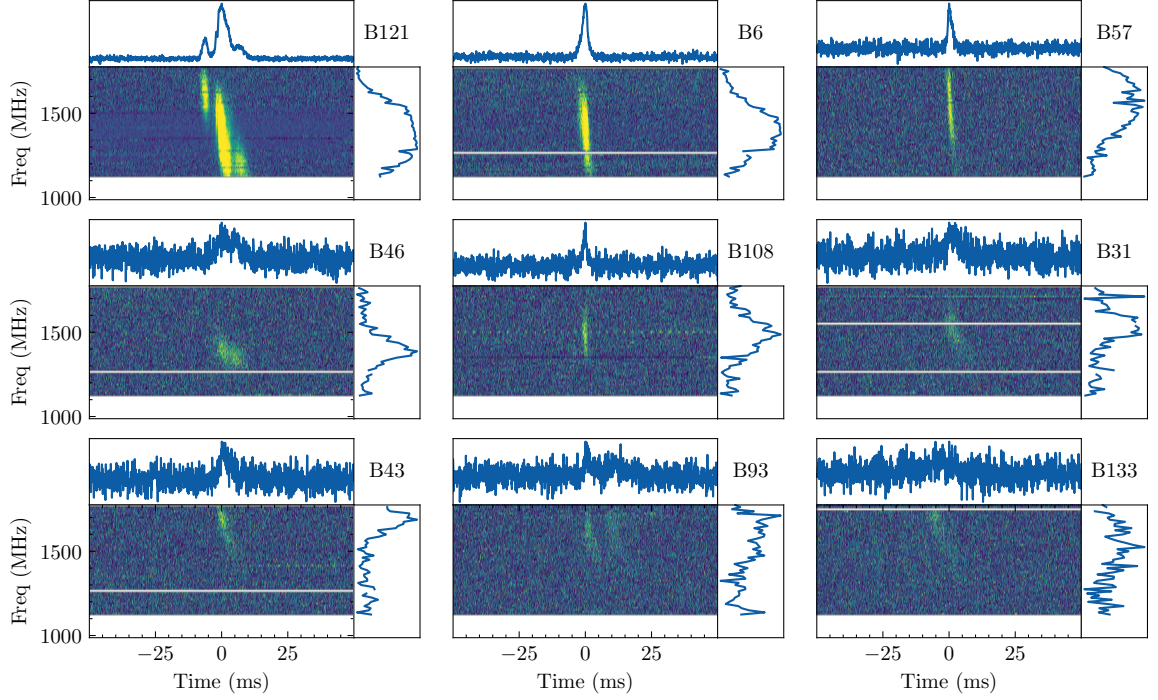


Figure 1.3: Dynamic spectra of nine bursts of FRB 121102, dedispersed at $DM = 560.5 \text{ pc cm}^{-3}$. For each burst, the top panel shows the burst profile obtained by averaging along the frequency axis, and the right panel shows the burst spectra obtained by averaging the burst data along the time axis. The white horizontal lines show the channels masked due to Radio Frequency Interference (RFI). The color scale of each spectrogram has been set from mean to 3 times the standard deviation of the off-pulse region. Figure adapted from Chapter 4.

Zhang et al., 2018). Recently, a multi-component FRB showed evidence of sub-second periodicity in its components (The CHIME/FRB Collaboration et al., 2021b). But it is yet to be seen if this FRB is a repeater and if other bursts from this FRB are consistent with this period.

In addition, progenitor models that favor a compact binary system (neutron star-neutron star, neutron star-black hole, neutron star-white dwarf, etc.) also predict a longer-term period in the FRB emission envelope. A similar long term period in burst activity has been observed for the two most observed repeaters: FRB 121102 and FRB 180916, with the former showing a period of 157 days (with a duty cycle of 56%; Rajwade et al., 2020b), while the latter with a period of 16 days (and a duty cycle of 20% at 400 MHz; CHIME/FRB Collaboration et al., 2020a). It has also been shown that the phase of peak burst rate and the duty cycle of FRB 180916 is frequency dependent, with higher frequency bursts arriving earlier in phase with narrower activity window, as compared to lower frequency bursts (Pastor-Marazuela et al., 2021; Pleunis et al., 2021a; Aggarwal et al., 2020). A clear consensus is yet to be reached on the origin of this periodicity, although many binary and slow precession models have been proposed to explain it.

1.4.7 Repeaters vs non-repeaters

The properties of repeating and non-repeating FRBs can be compared to understand whether these two populations have a common origin. It has been observed that repeater pulses tend to be temporally wider than non-repeaters. Repeaters also have, on average, narrower bandwidths than non-repeaters (Pleunis et al., 2021b). Also, repeaters usually show complex burst morphology comprised of multiple peaks, where the peak frequency of emission of individual components drifts downward in frequency as time progresses (Hessels et al., 2019). Non-repeaters have not demonstrated this downward-drifting so far. These differences could be due to beamed emission, where highly-beamed FRBs are less easily observed to repeat but are abundant enough to be detected as non-repeaters (Connor et al., 2020). It is also possible that this is due to a propagation effect: plasma lensing or circumbinary medium in the case of

a binary system (see Cordes et al., 2017; Main et al., 2018; Bilous et al., 2019), or it may be intrinsic to the two separate populations. A larger sample of repeater and non-repeater bursts, detected across multiple frequencies and with different telescopes, would be essential to probe these differences further. Moreover, Gardenier et al. (2021) used population synthesis analysis to show that some of the observed differences in these properties could be due to selection effects and argue that a single source population comprising of repeating FRBs can describe some early results of CHIME/FRB observations (CHIME/FRB Collaboration et al., 2019). A similar analysis applied on the first CHIME/FRB catalog might provide more constraining results (The CHIME/FRB Collaboration et al., 2021a).

1.5 What we derive

In this section, we discuss some FRB properties that can be inferred using the observed properties mentioned in the previous section.

1.5.1 DM derived distance

Dispersion measure can be used to constrain the distance to the FRB. The different sources of DM mentioned in Eq. 1.4, can be estimated, derived, or assumed using independent techniques, and the final component, i.e., DM_{IGM} can be related to the redshift of the source using “Macquart Relation” (Macquart et al., 2020; Inoue, 2004), by assuming standard cosmological parameters. A crude rule of thumb used to relate DM_{IGM} (in units of pc cm^{-3}) and z is: $z < DM_{\text{IGM}}/1200$ (Ioka, 2003; Inoue, 2004).

1.5.2 Luminosity and energy

The observed fluence and the estimated (or measured) distance to the FRB can be used to constrain the isotropic equivalent luminosity (and energy) of the burst. Depending on the spectrum of the burst and the simplifying assumptions made regarding it, there can be three primary ways to estimate the burst energy

from fluence (S) and distance (D_L). Assuming a broadband flat spectrum source, the energy contained in a burst can be referenced to the center frequency (ν_c) of the observation (Zhang, 2018; Li et al., 2021),

$$E = 4\pi 10^{-23} \left(\frac{D_L}{\text{cm}} \right)^2 \left(\frac{S}{\text{Jy s}} \right) \left(\frac{\nu_c}{\text{Hz}} \right) \text{erg}. \quad (1.9)$$

But, FRBs (especially repeaters) are generally bandlimited, with a Gaussian-like spectrum (Aggarwal et al., 2021a; Pastor-Marazuela et al., 2021; Pleunis et al., 2021b). In that case, they are neither broadband nor flat spectrum, and therefore the previous definition cannot be used. In such cases, the observable bandwidth of the FRB is typically used. But, as is shown in Chapter 5, this can lead to observational biases in the energy distribution analysis. Therefore, the third technique is to model the FRB spectra and use the fitted bandwidth (FWHM) instead of the visible bandwidth:

$$E = 4\pi 10^{-23} \left(\frac{D_L}{\text{cm}} \right)^2 \left(\frac{S}{\text{Jy s}} \right) \left(\frac{\text{FWHM}}{\text{Hz}} \right) \text{erg}. \quad (1.10)$$

The shape of the luminosity function of FRBs is currently not well understood but shapes like power-law, normal distribution, and Schechter function have been used (James et al., 2021; Luo et al., 2020a).

1.5.3 FRB rate and source counts

The all-sky rate of FRBs has been of interest since their discovery (Lorimer et al., 2007; Thornton et al., 2013; Shannon et al., 2018). This is because the volumetric rate of FRBs can be compared to the rate of other known transients (like a core-collapse supernova, gamma-ray bursts, etc.) to infer the source and intrinsic population of FRBs indirectly. FRB rates are typically calculated as an all-sky rate above a sensitivity limit, rather than a volumetric rate as the redshift distribution of FRBs is unknown (Petroff et al., 2019; Connor et al., 2020). In the simplest case, the all-sky rate of a telescope of sensitivity S_{min} , with instantaneous field-of-view¹ $\Omega \text{ deg}^2$,

¹Here, we have assumed that the whole telescope field-of-view has a constant sensitivity. This is generally not true as the shape of the beam is not flat. Beam shape can be accounted for while

that observes for T days and detects N FRBs would be,

$$R(> S_{\min}) = 4\pi \left(\frac{180}{\pi} \right)^2 \left(\frac{N}{\Omega T} \right) \text{sky}^{-1} \text{day}^{-1}. \quad (1.11)$$

Assuming FRBs as standard candles (i.e., with constant luminosity L) and distributed uniformly in a Euclidean universe, the observed flux (S) of FRBs at a distance r can be written as $S = L/(4\pi r^2)$. Further, assuming there are N FRBs detectable with a telescope of sensitivity (S_{\min}) within that volume, the constant density (ρ) of sources would be, $\rho = N(> S_{\min})/(4\pi r^3/3)$.

Using the above two equations, we get: $N(> S_{\min}) \propto S_{\min}^{-3/2}$. Using Eq 1.11, we can relate the rate with the minimum detectable flux: $R(> S_{\min}) \propto S_{\min}^{-3/2}$. This can be written more generally as,

$$R(> S_{\min}) \propto S_{\min}^{\gamma}, \quad (1.12)$$

where the source count index $\gamma = -1.5$ for Euclidean geometry (Petroff et al., 2019). Eq 1.12 can also be used to scale the rates for different telescopes with different detection thresholds to a common flux density,

$$R(> S_{\min}) = R_0 \left(\frac{S_{\min}}{S_0} \right)^{\gamma}, \quad (1.13)$$

where, R_0 and S_0 are the FRB rate and minimum detectable fluence for a given telescope. But, when combining rates from different telescopes, many other factors need to be accounted for: frequency of observation, beam pattern, bandpass, Radio Frequency Interference (RFI) environment, parameters of search pipeline, etc. As the FRB spectra are generally not independent of frequency, the spectral index² of the bursts also needs to be taken into account. Further, it has been shown that the source count index of FRBs can deviate from the Euclidean assumption (The CHIME/FRB

estimating the FRB rate, see Law et al. (2015); Deneva et al. (2009).

²Spectral index is defined as the slope of the power law relation between the observed flux and frequency of observation, i.e: $S_{\nu} \propto \nu^{\alpha}$. Here, S_{ν} is the flux of the source, ν is the frequency of observation and α is the spectral index.

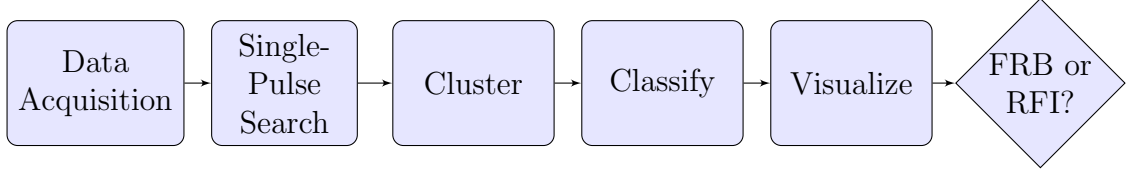


Figure 1.4: Flow diagram of a single-pulse search pipeline.

Collaboration et al., 2021a; James et al., 2021; Lawrence et al., 2017), and hence needs to be modeled carefully.

1.6 How to search for FRBs?

This section discusses some standard steps used to search for fast radio transients (FRBs or RRATs or single-pulses from pulsars). Given the wide range of properties of FRBs, it is necessary to make sure that the search pipeline is sensitive to detecting FRBs across the whole range of properties. Figure 1.4 shows the typical steps in an end-to-end workflow for single-pulse detection, starting from data recording to identifying if a candidate is an FRB or a result of RFI.

1.6.1 Data acquisition

Figure 1.5 shows the typical steps in the data acquisition system of any radio telescope. Before any search can be performed, the first step is to record the electromagnetic waves from a particular part of the sky. This is done using radio telescopes, which convert the electromagnetic waves into electrical signals and encodes the received information in voltages (Lorimer & Kramer, 2004; Condon & Ransom, 2016). The response of the telescope receiver (or feed) is measured over a range of frequencies, i.e., bandwidth. Therefore, for any observation, the intensity of signals at only the frequencies within the bandwidth is measured. These voltages are then digitized using an analog to digital converter (ADC), which samples the voltage amplitude at a rate that is twice the highest frequency of the bandwidth (i.e., Nyquist rate; Shannon, 1949). For an observation centered at 1.4 GHz, with a bandwidth of 800 MHz, the highest frequency would be 1.8 GHz. In this case, the ADC would have to sample



Figure 1.5: Flow diagram of a data acquisition system.

the analog voltages at 3.6 GHz, i.e., with a time resolution of 27 ns.

These digitized voltages are then fast Fourier transformed (FFT) to obtain the frequency-dependent intensities. The frequency resolution after the FFT is inversely proportional to the length of the FFT. Therefore, in the case of an N -point FFT, N time samples of digitized voltages are fast Fourier transformed into N frequency channels, followed by the FFT of the next set of N time samples (in case of non-overlapping FFTs), and so on. This results in a two-dimensional frequency-time spectrogram, with N frequency channels, each of resolution: bandwidth / N . The resulting time resolution would be N times the sampling time of the ADC. So, a 4096-point FFT of our previous example would result in a time resolution of $1.13\ \mu\text{s}$ and frequency resolution of $\sim 195\ \text{kHz}$. This is followed by a multiplier, where the complex voltages obtained after FFT are converted into intensities.

Such high time and frequency resolution observations are essential for studying transients like pulsars and FRBs that show variations at milliseconds and kHz scales. These data can then be saved to disk or downsampled by averaging adjacent time and frequency samples. This is then searched for dispersed pulses described in the next section.

1.6.2 Single-Pulse search

The following processing steps are applied on the spectrogram obtained from a radio telescope to search for transient signals with unknown dispersion measures, widths, and fluence. Figure 1.6 shows the typical sequence of these steps.

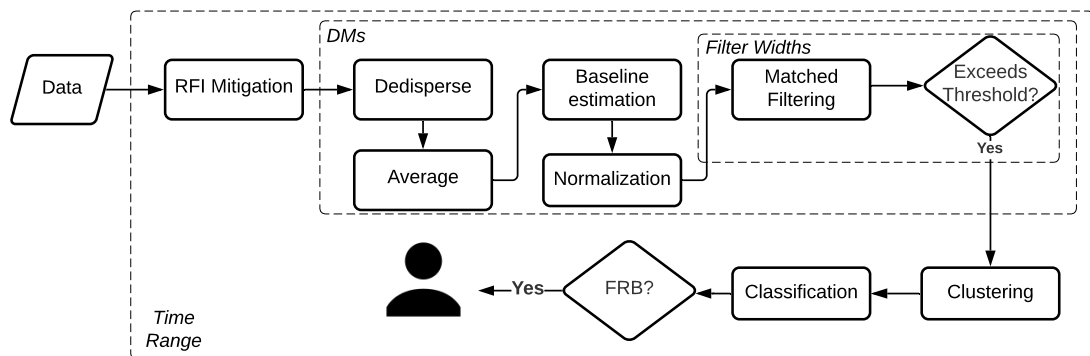


Figure 1.6: Flow diagram of a typical single-pulse search pipeline.

1.6.2.1 RFI mitigation

Artificially generated (human-made) radio signals that contaminate the observed data (spectrogram) are usually referred to as Radio Frequency Interference (or RFI). Multiple sources of RFI exist: mobile phones, power transmission lines, artificial satellites, radio towers, radar, etc., and are a ubiquitous problem for all the radio telescopes around the world. Some naturally occurring signals like lightning can be a source of contamination in the radio data. RFI can be grouped in two ways, based on its variability and extent. Based on its variability, it can be classified into time-varying or frequency-varying RFI and based on its frequency extent, it can be divided into narrowband (narrow frequency range) and broadband (present in all or most of the observing frequencies) RFI. RFI signals are usually much higher in intensity than the astrophysical signal (usually orders of magnitude stronger) and therefore reduce the sensitivity of search, if not mitigated. Time and frequency varying RFI can also mimic the dispersion of pulsars and FRBs and thus lead to false detections by automatic search and classification systems.

Naturally, radio astronomers have developed numerous techniques to mitigate RFI signals. The methods developed to improve the sensitivity of single-pulse searches can be grouped into two categories:

1. **Thresholding:** A constant threshold is applied to a section of the spectrogram, and any outlier lying outside the threshold range is considered to be RFI. This

is followed by using the same procedure on the next section of spectrogram and so on. Typically, a section of the spectrogram is chosen to consist of all the frequency channels within a range of time samples, similar to a non-overlapping running window in time. In some cases, this thresholding is applied multiple times with successively decreasing values of threshold on the same section of data to identify low-level RFI (CHIME/FRB Collaboration et al., 2019; Buch et al., 2016; Nita & Gary, 2010). There are several techniques to estimate this threshold value. The common steps of these techniques are:

- Assume that underlying data follow a certain distribution. Gaussian is the most common assumption here, but Chi-square and Gamma distributions have also been used depending upon the observing setup and data processing pipeline.
- Average along time to obtain a bandpass (i.e., intensities with respect to frequencies).
- For a range of frequency channels (or a window) estimate the mean and standard deviation.
- Set the threshold to a multiple of standard deviation plus the mean of the window.
- Compare the intensities in that window with this threshold and flag any data with intensity higher than the threshold.
- Shift the window to the next set of frequency channels in the bandpass, and repeat the above three steps.

To use a simple example, assuming the data follow a Gaussian distribution, the standard deviation can be robustly estimated using the Median Absolute Deviation (see, e.g., Leys et al., 2013), and then the threshold can be set to 3 or 5 times that value. Higher moments of data and other assumptions on the data distribution have also been used (Nita & Gary, 2010). This technique is primarily used to identify and mitigate narrowband and frequency variable

RFI.

2. **Zero DM subtraction:** As RFI signals originate from terrestrial sources and do not travel through the ionized plasma in the galaxy or the ISM, they do not show the frequency-dependent dispersion discussed earlier (see Section 1.2). This is not true for pulsars or FRBs, which show dispersion in their signal based on where in the galaxy they are or how far the source is. Therefore, one commonly used technique to mitigate broadband RFI is to generate a zero-DM time series (i.e., without de-dispersing the data and just averaging the frequency channels) and subtracting it from the original data. Broadband RFI will lead to peaks in the zero-DM time series and would be subtracted out. This was originally proposed by Eatough et al. (2009).

After identifying the RFI samples, the next challenge is to replace them with values that do not bias the statistics of the underlying data. This is typically done by replacing the data with mean, median, or some other estimator that provides a robust estimate of the average level of the underlying data.

1.6.2.2 Dedispersion and averaging

As mentioned, the signals from pulsars and FRBs are dispersed when they are observed on Earth (see Sections 1.2 and 1.4.1). The value of the DM of FRBs is not known *a priori*, and therefore needs to be determined. Moreover, this dispersion effect needs to be corrected during the search so that the signal present across different frequency channels can be combined and therefore detected at higher significance (see Figure 1.7). Therefore, during the single-pulse search, the original spectrogram is dedispersed at multiple values of DMs. The observed values of FRB DMs range from 100 pc cm^{-3} to 3000 pc cm^{-3} , so typically hundreds to thousands of values of DMs are searched between a range of 10 pc cm^{-3} to 10000 pc cm^{-3} . This is followed by frequency averaging, where a time series is generated at each DM trial and then searched for an outlier. Figure 1.7 shows this process for three trial DM values for a simulated pulse with $\text{DM}=500 \text{ pc cm}^{-3}$. Dedispersing the data at a DM close to the

intrinsic DM value of the pulse results in the highest possible S/N, with lower S/N when the pulse is dedispersed at DM values that are far from the intrinsic DM.

Dedispersion at each trial DM requires manipulating all the samples in the spectrogram. For each DM, the time delay at each time sample needs to be corrected at each frequency channel. This is, therefore, the most computationally expensive step in any single-pulse search pipeline. Due to the independent nature of dispersion correction at each sample, this process can be significantly improved from a parallelized GPU implementation that can dedisperse chunks of data in parallel. Several dedispersion algorithms have been developed: tree dedispersion (Taylor, 1974) and Fast Dispersion Measure Transform (Zackay & Ofek, 2017). They provide computational efficiency at the tradeoff of sensitivity, DM trial width, or require some prior transformations.

More DM trials translate to the detected DM being closer to the astrophysical DM, leading to a stronger detection. But more DM trials lead to more computations (scaling linearly with the number of DMs for brute force dedispersion implementation). Typically, DM trials are chosen based on a predefined fractional sensitivity loss between adjacent trials after considering observing factors like the width of the pulse, sampling time, frequency resolution. This process is described in detail in section 2.3 of Levin (2012).

1.6.2.3 Baseline estimation

Due to instrumental effects, sky background variance, RFI, or other factors, the mean level of the intensities in the data can vary with time (on seconds to minutes timescale). When such data are dedispersed, and frequency averaged to form time series, it can lead to baseline variations. This can make the search less sensitive by reducing the significance of the transient, making it difficult to identify the signal. Typically, this baseline is removed by estimating the mean (or its more robust counterpart, median) of a running window and subtracting it. Outliers in the time series can bias this baseline estimation, so sometimes samples above a specified threshold are ignored first (akin to RFI threshold) before the mean is estimated (Barsdell,

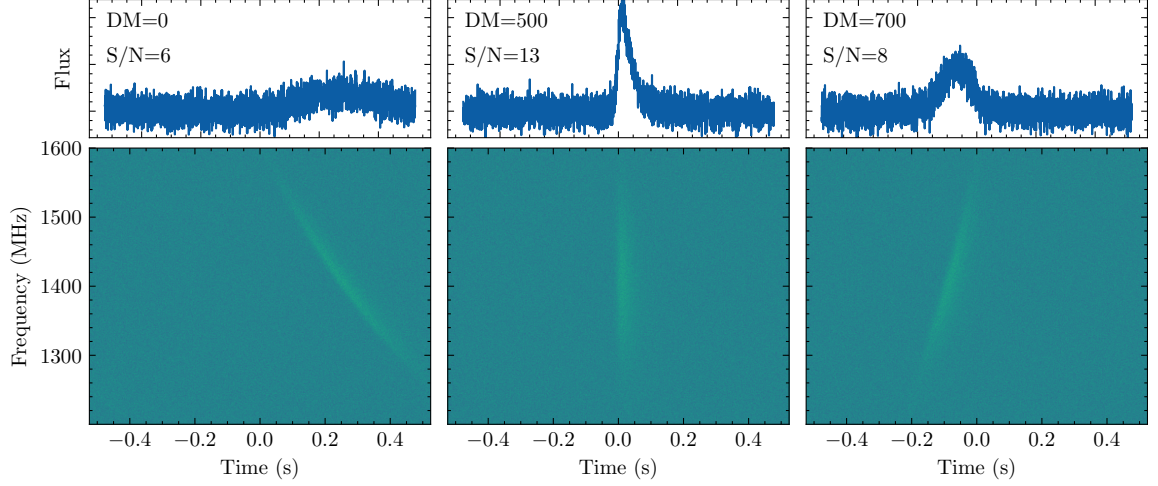


Figure 1.7: Figure showing the result of dedispersing a simulated pulse at multiple trial DM values. The intrinsic DM of the simulated burst was 500 pc cm^{-3} . The top row shows the frequency averaged time series of data dedispersed at DMs mentioned on the top left. The bottom rows show the spectrogram of the dedispersed data. S/N of each time series is also mentioned at the top left. We can see that dedispersing the pulse at the DM of 500 pc cm^{-3} results in the highest S/N. Figure generated using BURSTFIT.

2012). An unfortunate consequence of this is that it lessens FRB search sensitivities to longer bursts. This is because typical baseline windows are 1 to a few seconds long, so it would remove the signals as long as or longer than this scale. Therefore, longer baseline windows are usually preferred to avoid this.

1.6.2.4 Normalization

The baseline subtracted time series is usually normalized by dividing the data by its standard deviation. Again, a robust estimation of standard deviation is preferred. As mentioned during the discussion of RFI mitigation (Section 1.6.2.1), one of the commonly used methods is to calculate median absolute deviation and use that to estimate standard deviation (Leys et al., 2013). Following normalization, the signal-to-noise ratio can be measured directly from the amplitude of the time sample, as the standard deviation has been scaled to be one.

1.6.2.5 Matched filtering

The samples with a signal-to-noise ratio above a predefined threshold can be considered as potentially interesting candidates. But the detection significance of pulses wider than a single time sample can be improved by applying an additional processing step. This is called matched filtering or boxcar convolution search (Lorimer & Kramer, 2004; McLaughlin & Cordes, 2003). The time series obtained after the previous step is convolved with boxcar (or top-hat) filters of varying widths. For each case, samples in the time series that exceed the threshold are identified and reported as candidates. This is repeated for the dedispersed time series generated at all DM trials. Figure 1.8 shows S/N and the resulting time series, after a simulated pulse is convolved with boxcar filters of varying widths. As we can see, for this pulse with intrinsic width of 39 samples, the highest S/N is obtained for matched filters with nearby widths.

This process improves the S/N of wider pulses and makes them detectable. This can also be understood using the following example: consider that the dedispersed and normalized time series contains an astrophysical pulse of width W samples, with the amplitude of each sample as S (assuming a boxcar shaped pulse for simplicity). Here, if S is greater than the S/N threshold, then this pulse will be detected. But if S is less than this threshold, then following a convolution with a filter of width W , the S/N of the pulse will be higher at $S\sqrt{W}$, which might be greater than the S/N threshold, making it detectable. Realistically, the boxcar width would not match the width of the transient perfectly, but using convolution at multiple widths ensures that the pulse is detected at a boxcar width closer to the intrinsic value.

1.6.3 Clustering

Following the above search steps, the single-pulse search pipeline reports candidates with S/N greater than the threshold. For each candidate, it reports the trial dispersion measure, trial width, and time sample of the pulse. In the case of an interferometer, it might also report the sky location of the transient (see Chapter 3

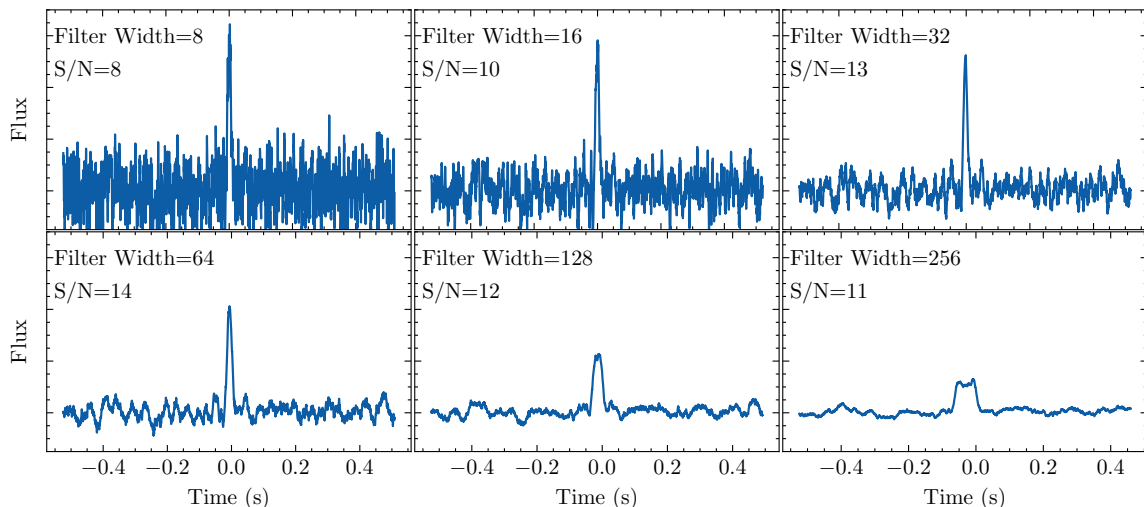


Figure 1.8: Figure showing the result of applying match filtering with boxcars of various widths to a simulated pulse. The signal to noise (S/N) and filter width (in samples) is provided on the top left of each plot. The intrinsic width of the pulse was 39 samples. Filter widths close to this pulse width result in the maximum S/N. Figure generated using BURSTFIT.

and Law et al., 2018). A bright astrophysical event might trigger the pipeline at multiple values of this trial space and can therefore lead to hundreds or thousands of candidates per true event. This is because nearby DM or width trials will detect the same transient, but with lower significance, that might still exceed the detection threshold³. Therefore, it is necessary to group all these redundant candidates together and only post-process the highest significant detection from the group. This is done using a clustering algorithm. For a single dish telescope, where only DM and time information is available, a friends-of-friends-based algorithm is typically used (Pang et al., 2018; Deneva et al., 2009; Barsdell, 2012). A simpler clustering technique could specify a maximum distance for each parameter, and candidates that lie within that specified distance of each other are considered part of a cluster. In the case of interferometers, the sky location of each candidate can also be used for clustering. A robust comparison of several clustering algorithms is provided in Chapter 3 and Aggarwal et al. (2021c).

³The following notebook provides a simple calculation to estimate the number of candidates expected for a single-pulse search pipeline: https://github.com/KshitijAggarwal/rfclustering/blob/main/notebooks/ncands_vs_snr.ipynb

1.6.4 Classification

Single-pulse search pipelines, as described above, can generate hundreds to thousands of candidates every day, depending upon the RFI environment, observing setup, trials searched, number of beams, etc. (Law et al., 2018; Agarwal et al., 2020b; Gupta et al., 2021). Each of these candidates exceeded the detection threshold of the pipeline, but that does not guarantee astrophysical origin. As discussed previously, RFI can also lead to false candidates, and even after applying extensive RFI mitigation techniques, more than 99% of all candidates generated from a search pipeline are still due to RFI. Therefore, each candidate has to be classified into RFI or astrophysical (FRB or pulsar). This has been traditionally done using human resources: i.e., the researchers working on the search system usually classify each candidate manually. This task becomes unsustainable for hundreds of candidates every day, and with the search pipelines running commensally. Therefore, machine learning techniques have been developed to classify candidates into FRB or RFI automatically. Some of these techniques are: Random Forests used in Foster et al. (2018) and Wagstaff et al. (2016); Convolutional Neural Networks used in Agarwal et al. (2020a) and Connor & van Leeuwen (2018), etc. Machine learning algorithms can be carefully designed to be highly accurate and robust. Moreover, with techniques like self-supervised learning, it is possible to retrain these algorithms with new data for improved performance with minimal human labeling. Finally, the candidates classified as astrophysical by the machine learning algorithm are visualized and manually verified (see Section 6.5 and Figures 6.2 and 6.3).

1.6.5 Sensitivity analysis of search

Until now Section 1.6 has discussed various processing steps involved in a single-pulse search. It is essential to understand the recovery rate or efficiency of the detection pipeline. A simple estimate of the theoretical sensitivity of the search can be made using the radiometer equation (see Section 1.4.4.1). In this case, S/N threshold times the root-mean-square fluctuation is used as the sensitivity limit. But only

after properly accounting for systematics due to the search, inferences about the underlying properties of the FRB population can be drawn. One way to do this is to create simulated models of all the steps in the pipeline and estimate the efficiencies directly from it. This is challenging as it requires perfect models of the instrumentation and software pipeline. A more commonly used technique is to inject simulated FRBs with a range of properties in the search pipeline and estimate the percentage of those that are correctly identified after each step and at the end (Gupta et al., 2021; Agarwal et al., 2020b). Simulated FRBs can be injected directly into the voltage data or added to the intensity data. They can have variations in DM, width, scattering timescale, fluence, spectral bandwidth, profile shape, etc., inspired by the expected variation in the transients that are being searched. Instrumental effects like bank drops and packet loss can also be simulated. At the end of the search step and then after the classification step, the percentage of injected FRBs that are recovered can be estimated. It can also be studied how this recovery rate varies with respect to any specific parameter (DM, width, etc.).

1.7 Localization of FRBs

As discussed in Section 1.4, the detected sample of FRBs show a wide variety of properties, and there is no clear consensus on their progenitors (see Section 1.8). To understand and constrain the progenitors of FRBs, it is necessary to study the host galaxies of these FRBs and their local environments. The localization region (or the possible region of sky from which an FRB could have originated) of an FRB detected by a single dish telescope is usually several arcminutes wide, depending on the size of the beam and S/N of the FRB. Given the estimated distance to the FRB (see Section 1.5.1), this region can contain hundreds of galaxies, detectable in sensitive follow-up optical observations (Eftekhari & Berger, 2017). Such a large localization region makes it impossible to associate an FRB to a particular host galaxy for almost the whole FRB sample currently known, except the lowest DM FRBs (Bhardwaj et al., 2021a,b). Even in the case of the lowest DM FRBs, precise localization is

critical to study the local environment of the FRB.

Sub-arcsecond localization enables the identification of the FRB host galaxy, which further enables the estimation of the redshift of the FRB (as the redshift of the galaxy can be determined using photometry or spectroscopy). This can then be used to estimate the luminosity of the FRB and derive meaningful insights into the DM contributions of host and ISM (see Section 1.4.1). It can also be used to understand the $DM_{\text{IGM}}-z$ relation (Pol et al., 2019; Hackstein et al., 2020; Zhu & Feng, 2021; Zhang et al., 2021; Hackstein et al., 2021; Batten et al., 2021), its variance (Macquart et al., 2020; Bhandari & Flynn, 2021), and study the halo of any foreground galaxy that intersects the line of sight of the FRB (Prochaska et al., 2019; Simha et al., 2021). A sample of host galaxies can then be compared against other known transients, as we discuss later.

Further, a better understanding of the local environment of FRBs is essential, as it can differ significantly from the broader properties of the FRB host galaxy. For instance, an FRB that lies in a star-forming galaxy might not lie in a star-forming region, or an FRB might lie near a region of high magnetic field (Tendulkar et al., 2017, 2021). The offset of FRB position itself from the center of the host galaxy can be compared against the offset distribution of other transients to draw conclusions about FRB sources (Heintz et al., 2020; Bhandari et al., 2021). The precise position of the FRB also enables multi-wavelength follow-up and archival search to identify any temporally and spatially co-located transient or persistent emission throughout the electromagnetic spectrum. We discuss this further in the following subsections.

Therefore, sub-arcsecond localizations, enabled by interferometers like VLA, ASKAP, and DSA, are key to better understanding the FRBs and their progenitors. We discuss the techniques to localize FRBs, and some FRBs localized using VLA in Chapter 6.

1.7.1 Host galaxy population and local environment

As of this writing, 24 FRBs have been localized or associated to a host galaxy, of which 8 are repeaters (Chatterjee et al., 2017; Bannister et al., 2019; Ravi et al.,

2019; Macquart et al., 2020; Bhandari et al., 2021; Law et al., 2020; Marcote et al., 2020; Ravi et al., 2021; Heintz et al., 2020; Prochaska et al., 2019). Bhandari et al. (2021) used a sample of 16 FRBs with secure host association (6 repeating and 10 non-repeating) and analyzed their host galaxy properties. They find that the FRB hosts are, in general, massive and moderately star-forming, and their masses are offset from the star-forming main-sequence. Further, FRB hosts do not track stellar mass and star formation rate in field galaxies in the local universe. Interestingly, hosts of all three localized repeating FRBs lie in the star-forming region. However, there appears to be no significant difference in the host properties of repeating and non-repeating FRBs. The global properties of FRB hosts are statistically indistinguishable from core-collapse supernovae and short gamma-ray bursts hosts and different from long gamma-ray bursts and super-luminous supernovae. The detection of FRB-like burst from Galactic magnetar and localization of an FRB to a globular cluster demonstrates that much is still to be learned about broader FRB host population (Bochenek et al., 2020; CHIME/FRB Collaboration et al., 2020b; Bhardwaj et al., 2021a).

Using sub-arcsecond localizations and high-resolution HST observations, the local environment of some FRBs has also been studied (Mannings et al., 2021; Tendulkar et al., 2017, 2021). FRBs lying in spiral galaxies appear on or near the spiral arms. Most FRBs do not lie in regions of elevated local star-formation rate and stellar mass surface densities in comparison to the mean global values of the host galaxy. It has been found that both FRB 180916 and FRB 121102 are offset from a nearby star-forming knot in their galaxy by around 200 pc (Tendulkar et al., 2021; Bassa et al., 2017). This is problematic for young magnetar models as such sources are found closer to their birth sites. For FRB 180916, the observations are compatible with a high-mass X-ray binary or gamma-ray binary with an O-type or B-type companion star (Tendulkar et al., 2021).

1.7.2 Persistent radio emission

The localization of the first repeating FRB, FRB 121102, also enabled its association with a co-located persistent radio source (PRS; Chatterjee et al., 2017). The

PRS is compact (projected size < 0.7 pc), persistent (non-bursty) with a luminosity of $L_{1.4\text{GHz}} = 2 \times 10^{29} \text{ ergs s}^{-1} \text{ Hz}^{-1}$, comparable to a low luminosity AGN. The spectrum of the PRS is relatively flat (spectral index ~ 0.07) to 8 GHz and declines above that cutoff (Chatterjee et al., 2017; Marcote et al., 2017; Resmi et al., 2021), and the emission is consistent with isotropic incoherent synchrotron radiation. Although the probability of chance coincidence of this PRS with that of FRB position is small, it is yet unknown if the presence of PRS has implications for FRB emission. Soon after the discovery of this PRS, theoretical models were proposed connecting the source of FRB121102 and PRS (Kashiyama & Murase, 2017; Beloborodov, 2017; Margalit & Metzger, 2018; Metzger et al., 2019). But the next dozen or so localized FRBs do not have any associated PRS, down to luminosities much lower than that of PRS of FRB121102. Therefore FRB 121102 was considered an outlier in this population.

Recently, the localization of FRB 190520 using VLA has found it to be associated with a PRS (see Chapter 6; Niu et al., 2021a). This implies that a sub-population of FRBs have an associated PRS, which can be revealed only in a large sample of localized FRBs and corresponding deep radio observations of those fields. Law et al. (2021b) use the population of localized FRBs, with meaningful constraints on possible PRS, to show that there is no strong evidence to show that repeating FRBs are more likely to be associated with a PRS than apparently non-repeating FRBs. Further, they show that the presence of PRS for an FRB can be predicted based on the active repetition rate and excess dispersion measure attributed to FRB local environment.

1.7.3 Multi-wavelength emission

Localization of FRBs also enables their multi-wavelength follow-up. This is useful as several progenitor models predict a transient or persistent multi-wavelength counterpart with the FRB. Therefore any detection of co-located emission would increase the confidence in such models (Burke-Spolaor, 2018). But thus far, extragalactic FRBs have not yielded any reliable coincident detection at other wavelengths (Scholz et al., 2016; Scholz et al., 2020; Pearlman et al., 2020). It is also possible that the FRB sources are too distant and below the detection threshold of the current

high-energy telescopes.

The only confirmed counterpart of any FRB is the hard X-ray and soft gamma-ray burst from the Galactic magnetar SGR 1935+2154, which was also temporally coincident with the FRB (Bochenek et al., 2020; CHIME/FRB Collaboration et al., 2020b; Kirsten et al., 2021a; Li et al., 2021; Ridnaia et al., 2021; Mereghetti et al., 2020). This unambiguously established that at least some FRBs originate from magnetars. Conversely, FRBs can be searched at the sites of other known transients. No FRB has been detected in association with GRBs, even after several attempts (Madison et al., 2019; Men et al., 2019; Palliyaguru et al., 2021; Hilmarsson et al., 2020; Guidorzi et al., 2020). Searches for FRBs from super-luminous supernovae have also not been associated with an FRB (Law et al., 2019).

1.8 FRB progenitor models

The vast range of observational properties for the currently known sample of FRBs makes it difficult for any one model to unify the FRBs as originating from a single source class. Therefore, it is likely that there are multiple populations of FRBs, so models can also be divided based on the properties they explain. Naturally, a large number of models have been proposed to explain FRBs (Platts et al., 2019), ranging from the relatively familiar, such as magnetic reconnection on magnetars (Lu & Kumar, 2018) or giant flares from young magnetars (Lyutikov et al., 2016), to the exotic, such as axion star phenomena (Iwazaki, 2014) and cusps from cosmic strings (Costa et al., 2018).

Based on time- and frequency-resolved FRB sub-structures, there is evidence for phenomenological divisions in FRBs, possibly pointing to *at least* two classes of FRB progenitor. A number of these populations may be associated with ambient or bursty X-ray emission (Burke-Spolaor, 2018). Two leading theories have emerged among the many progenitor models based on FRB energetics, pulse timescales, and occurrence rates. The theories point to either material around a large BH or a young magnetar as the FRB engine. Both repeating and non-repeating FRBs appear to

remain broadly consistent with these models, although we have limited information about FRB environments that are required to put the models to the test. Either way, any progenitor model would have to explain at least the mechanism for the high brightness temperatures ($> 10^{32}\text{K}$) and short (millisecond and in some cases microsecond) duration of the bursts, which indicate a coherent emission mechanism and compact emission region (Zhang, 2020).

1.8.1 Black Hole models

Almost all FRB-generating BH models have two things in common: the BH must be accreting at least at some minimum level, and its mass must be either intermediate or supermassive (e.g., $10^5 M_\odot$ or higher) to have a gravitational energy reservoir sufficient enough to produce the luminosities seen from FRBs.

To give a specific example, one model (Zhang, 2017) uses the interaction of gasflow from a $10^4 - 10^6 M_\odot$ BH with the magnetosphere of a foreground neutron star to generate FRBs. FRB121102’s pulses can be explained if the source of this gasflow is a low-luminosity accreting BH, provided it ejects a nearly isotropic disk wind outflow with a varying ram pressure (Zhang, 2018).

Other BH models have used similar mass requirements but with different physical mechanisms, showing that FRBs can be generated from coherent emission in strong plasma turbulence excited by a relativistic electron jet (Romero et al., 2016; Vieyro et al., 2017), or from the jet-confining regions in the accretion disk of an intermediate-mass BH (Katz, 2017). Here, the narrow collimated beam is similar to the jets emitted by an AGN, micro-quasars, or a GRB. Collisions of large superconducting dipoles around slowly accreting supermassive BH can also form FRBs (Thompson, 2017).

1.8.2 Young magnetar models

Pulsar-related phenomena are an attractive descriptor for FRB emission, as our galaxy contains many similar pulse emitters. To scale the neutron star energies up to

the level of FRBs, one needs a large magnetic field (i.e., magnetar) and large rotational spin-down energy (i.e., the magnetar must be young). FRB121102’s PRS may indicate a luminous pulsar wind nebula (Kashiyama & Murase, 2017; Beloborodov, 2017; Margalit & Metzger, 2018; Metzger et al., 2019; Lyubarsky, 2014). This expanding magnetized nebula (created by the same outburst which caused the FRB) can also explain all the observed properties (size, flux) of the persistent source associated with the FRB121102 (Margalit & Metzger, 2018). Another model explains the origin of FRBs from the synchrotron maser emission produced by flare ejecta from young magnetars (Metzger et al., 2019). The Galactic FRB may have been due to one such magnetar model (CHIME/FRB Collaboration et al., 2020b; Bochenek et al., 2020).

1.9 Open questions

We have described the enigmatic phenomenon of FRBs, their observational and derived properties, key results in our quest to understand them better, some progenitor models that try to explain them, and most importantly a detailed description of how an FRB search is performed, so we can detect more of these sources. Detecting and then localizing a large number of FRBs appears to be the most promising way of uncovering exotic FRBs that will challenge the proposed theoretical predictions and simultaneously provide a better understanding of this population. The CHIME/FRB instrument has been prolific in the detection aspect of FRBs (The CHIME/FRB Collaboration et al., 2021a), but larger samples at other frequencies will also provide complementary information. Large field-of-view telescopes like DSA will be promising in this regard. The second aspect is the localization of FRBs, which interferometers like VLA and ASKAP have pioneered, and this population will increase rapidly in the coming months and years.

We have come a long way from the discovery of the “Lorimer Burst”. We have answered many questions, and with every answer uncovered new questions. We know a lot more about FRBs than one could have imagined 15 years ago, but this journey is just getting started. Here are some of the puzzles still to be resolved:

1. Is there a single population of FRBs? Are repeaters and non-repeaters two separate populations?
2. Do all FRBs have the same progenitors? If not, do different progenitors explain the diverse populations, or can a single progenitor describe these seemingly different populations?
3. Is the presence of PRS an indicator of a particular sub-population of FRBs (e.g. FRB 121102-like FRBs)?
4. What observable traits do host galaxies of FRBs have? Are they similar to any other known transient?
5. What causes periodicity in the burst activity of some repeaters? And why are all repeaters not periodic?
6. What is the underlying energy distribution of repeating FRBs? What is that of the population of all FRBs?
7. Do FRBs have multi-wavelength emission associated with them? If so, why have we not detected it?

1.10 Thesis outline

In this thesis we will discuss several advancements that we made over the course of last four years to better search, localize, classify and analyse the FRBs. These are organized as follows. We discuss some open source software for FRB search and analysis in Chapter 2. We then talk about analysis to develop and compare clustering algorithms for REALFAST system at VLA in Chapter 3. This is followed by the search and analysis of a sample of bursts from FRB 121102 detected in data from Arecibo Telescope, in Chapter 4. Chapter 5 highlights some observational biases in the detection of bandlimited FRBs, and Chapter 6 discusses the REALFAST pipeline and some FRBs localized using the REALFAST system. Finally, we conclude and make predictions in Chapter 7.

Chapter 2

Software

In this chapter we discuss some open-source software that we have developed to assist in single pulse search and analysis of FRBs.

2.1 Fast Extragalactic Transient Candidate Hunter (FETCH)

FETCH (Agarwal et al., 2020a) is a set of eleven deep learning models that can be used to distinguish FRBs from RFI in real-time. All the models are combinations of two convolutional neural networks each with varying architectures. It was trained on data from telescopes at Green Bank Observatory but is designed to be robust to data from different telescopes, observing configurations (frequency, bandwidth, etc.) and RFI environments. This was achieved using a carefully designed training and pre-processing strategy. All the models in FETCH have a recall and precision above 99.5% on the test dataset, and it was successfully able to recover all real FRBs above an S/N of 10 in ASKAP and Parkes data. FETCH has since then been used to discover many new FRBs and re-detect pulses from many known repeating FRBs and pulsars (Agarwal et al., 2020b; Aggarwal et al., 2020; Law et al., 2020; Kumar et al., 2019; Aggarwal et al., 2021a; Bhandari et al., 2021; Pleunis et al., 2021a; Kumar et al., 2021c; Pearlman et al., 2020; Rajwade et al., 2020b; Majid et al., 2020; Kirsten et al., 2021b,a; Niu et al., 2021a; Ravi et al., 2021). FETCH is openly available on Github¹.

¹<https://github.com/devanshkv/fetch>

2.2 Burstfit

BURSTFIT is a python package to do robust spectro-temporal modeling of single pulses from FRBs or pulsars. It provides a framework to model any spectrogram consisting of any complex FRB or pulsar pulse using robust methods. It can easily incorporate any user-defined python function to model the profile, spectra, and spectrogram and is not limited to the available functions with the package. Modeling using BURSTFIT broadly consists of the following five steps:

- Data Preparation: Consists of extracting the relevant raw data around the burst time, normalizing it to zero mean and unit standard deviation and flagging the RFI channels.
- Single-component fitting: Fits the pulse profile followed by the spectra. Then using the fitted estimates, it tries to model the spectrogram. All the fitting at this stage uses `scipy.curve_fit`.
- Statistical Tests: Performs tests to compare the on-pulse residual data to off-pulse noise. If the two distributions are found to be statistically similar, then the fitting process is terminated, and more components are not fit; otherwise, more components are fit by repeating the previous step.
- Multi-component fitting: Fits all the components of the spectrogram together using `scipy.curve_fit` and the fitted values of parameters from the previous stages as the initial estimates.
- MCMC: Runs a full-scale Markov Chain Monte Carlo on all the parameters to obtain the posterior distribution of the parameters.

See Section 3.5 of Aggarwal et al. (2021a) for more details. BURSTFIT is openly available on Github².

²<https://github.com/thepetabyteproject/burstfit>

2.3 FRB Periodicity Analysis (FRBPA)

FRBPA (Aggarwal et al., 2020) is a python package to perform periodicity analysis on FRB bursts. It can be used to search for periodicity in the activity of repeating FRBs, as has been observed for FRB 121102 and FRB 180916 (see Section 1.4.6 and CHIME/FRB Collaboration et al., 2020a; Rajwade et al., 2020b). It provides three periodicity search techniques: Pearson Chi-square test (CHIME/FRB Collaboration et al., 2020a), Minimum fractional width of folded profile (Rajwade et al., 2020b) and Quadratic-Mutual-Information-based periodicity search technique (Huijse et al., 2018). FRBPA also consists of functions to visualize the periodicity and the exposure time for a given repeater. FRBPA is openly available on Github³.

2.4 Calculating Association Probability of FRBs (CASP)

CASP is a python package to calculate the chance association probability of FRBs to that of host galaxies. It implements the procedure described in sections 2 and 3 of Eftekhari & Berger (2017) to calculate the probability without and with using the estimated redshift of the FRB. The first approach assumes a Poisson distribution of radio sources across the sky. It calculates the chance coincidence probability using the number density of galaxies above a limiting r-band magnitude. In the second approach, the number density of galaxies at a given redshift is calculated by integrating the optical luminosity functions. It also implements the approach of Bloom et al. (2002, Sec. 6.1) to estimate the chance coincidence probability using r band magnitude and the expression given by Hogg et al. (1997). See Section 6.1 and Appendix of Aggarwal et al. (2021b) for more details. CASP is openly available on Github⁴.

³<https://github.com/KshitijAggarwal/frbpa>

⁴<https://github.com/KshitijAggarwal/casp>

2.5 Your Unified Reader (Your)

The data output from radio telescopes is in one of the two commonly used formats: `psrfits` (Hotan et al., 2004) and `Sigproc filterbank`⁵ (Lorimer, 2011). Software developed for transient searches often only works with one of these two formats, limiting their general applicability. Therefore, researchers have to write custom scripts to read/write the data in their format of choice before beginning any data analysis relevant to their research. This has led to the development of several python libraries to manage one or the other data format (like `pysigproc`⁶, `psrfits`⁷, `sigpyproc`⁸, etc). Still, no general tool exists which can work across data formats.

2.5.1 Statement of need

Your (Your Unified Reader) is a python-based library that unifies the data processing across multiple commonly used formats. **Your** was conceived initially to perform data ingestion for The Petabyte FRB search Project (TPP), which will uniformly search a large number of datasets from telescopes around the world for FRBs. As this project will process data in different formats from multiple telescopes worldwide, a unified reader was required to streamline the search pipeline. **Your** implements a user-friendly interface to read and write in the data format of choice. It also generates unified metadata corresponding to the input data file for a quick understanding of observation parameters and provides utilities to perform common data analysis operations. **Your** also provides several state-of-the-art radio frequency interference mitigation (RFI) algorithms (Agarwal et al., 2020b; Nita & Gary, 2010), which can now be used during any stage of data processing (reading, writing, etc.) to filter out artificial signals.

Published as Aggarwal et al. (2020)

Contributing authors: Devansh Agarwal, Joseph W Kania, William Fiore, Reshma Anna Thomas, Scott M. Ransom, Paul B. Demorest, Robert S. Wharton, Sarah Burke-Spolaor, Duncan R. Lorimer, Maura A. McLaughlin, Nathaniel Garver-Daniels

⁵<http://sigproc.sourceforge.net/>

⁶<https://github.com/demorest/pysigproc>

⁷<https://github.com/scottransom/presto/blob/master/python/presto/psrfits.py>

⁸<https://github.com/FRBs/sigpyproc3>

Your can be used at the data ingestion step of any transient search pipeline and can provide data and observation parameters in a format-independent manner. Generic tools can thus be used to perform the search and further data analysis. It also enables online processing like RFI flagging, decimation, subband search, etc.; functions for some of these are already available in **Your**. It can also be used to perform analysis of individual candidate events (using **Candidate** class): generate candidate data cutouts, create publication-ready visualizations, and perform GPU accelerated pre-processing for candidate classification (Agarwal et al., 2020a). It also consists of functions to run commonly used single-pulse search software **Heimdall**⁹ (Barsdell, 2012) on any input data format.

Your will benefit not only experienced researchers but also new undergraduate and graduate students who otherwise have to face a significant bottleneck to understand various data formats and develop custom tools to access the data before any analysis can be done on it. Moreover, **Your** is written purely in python, which is a commonly used language within Astronomy. It also comes with comprehensive documentation¹⁰ and example notebooks¹¹ to make it easier to get started.

Your uses the matplotlib library (Hunter, 2007) for plotting, and also makes use of various numpy (Harris et al., 2020), scipy (Virtanen et al., 2020), scikit-image (Pedregosa et al., 2011), numba (Lam et al., 2015) and Pandas (Reback et al., 2021; Wes McKinney, 2010) functions. **Your** also leverages several functions in the Astropy package (Astropy Collaboration et al., 2013; Price-Whelan et al., 2018): fits (astropy.io.fits), units (astropy.units), coordinates (astropy.coordinates) and time (astropy.time).

⁹<https://sourceforge.net/projects/heimdall-astro/>

¹⁰<https://thepetabyteproject.github.io/your/>

¹¹<https://github.com/thepetabyteproject/your/tree/master/examples>

Chapter 3

Robust Assessment of Clustering Methods for Fast Radio Transient Candidates

3.1 Chapter Overview

As discussed in Section 1.6, fast radio transient search algorithms identify signals of interest by iterating and applying a threshold on a set of matched filters. These filters are defined by properties of the transient such as time and dispersion. A real transient can trigger hundreds of search trials, each of which has to be post-processed for visualization and classification tasks. In this chapter, we have explored a range of unsupervised clustering algorithms to cluster these redundant candidate detections. We demonstrate this for REALFAST, the commensal fast transient search system at the Very Large Array. We use four features for clustering: sky position (l , m), time and dispersion measure (DM). We develop a custom performance metric that makes sure that the candidates are clustered into a *small* number of *pure* clusters, i.e, clusters with either astrophysical or noise candidates. We then use this performance metric to compare eight different clustering algorithms. We show that using sky location along with DM/time improves clustering performance by $\sim 10\%$ as compared to the traditional DM/time-based clustering. Therefore, positional information should be used during clustering if it can be made available.

We conduct several tests to compare the performance and generalisability of

Published as Aggarwal et al. (2021c)

Contributing authors: Sarah Burke-Spolaor, Casey J. Law, Geoffrey C. Bower, Bryan J. Butler, Paul B. Demorest, T. Joseph W. Lazio, Justin Linford, Jessica Sydnor, Reshma Anna-Thomas

clustering algorithms to other transient datasets and propose a strategy that can be used to choose an algorithm. Our performance metric and clustering strategy can be easily extended to different single-pulse search pipelines and other astronomy and non-astronomy-based applications.

3.2 Introduction

One of the significant difficulties when seeking fast-transient radio signals is the load of candidates that results from a transient search: it is common for a search algorithm to return millions to billions of candidates from a survey, only a few of which end up being genuine (the rest being thermal noise and radio-frequency interference—RFI). Even one bright event, whether astrophysical or artificial, can generate many hundreds of separate candidates. This is because search algorithms iterate over a set of matched filters and identify transients that exceed the detection threshold. Clustering algorithms to account for this effect are of dire importance to any radio transient search pipeline. A rigorous study of an effective clustering algorithm for fast-radio-transient searches is the primary purpose of the study reported here.

Here we briefly discuss the various steps in a single pulse search pipeline (see Section 1.6 for a detailed discussion). When searching for a radio transient, neither its dispersion measure nor width (i.e., duration) are known. Therefore, one must search over a range of DMs and widths to carry out a full-sensitivity search. DM values at which to search are chosen by considering the expected decline in signal-to-noise (S/N) ratio, due to pulse broadening, at the adjacent DMs (Cordes & McLaughlin, 2003; Levin, 2012). To summarise, a standard fast-transient-search pipeline dedisperses the data at various trial DMs, averaging all the frequencies to obtain a one-dimensional time series, followed by convolution using boxcar filters of various widths. Candidate pulses are identified by searching for peaks above a pre-decided threshold, with the S/N of a candidate determined from an estimate of the signal strength with respect to the standard deviation within the region defined by the boxcar filter width. Dedispersing at an incorrect DM, or using a boxcar filter of incorrect width, would reduce

the S/N of the pulse. As previously noted, any event can lead to multiple candidates being detected by the search pipeline if the S/N remains above the threshold at the incorrect DM or boxcar width.

This process can lead to a substantial number of redundant candidates caused by a single event. Clustering is performed on these candidates to automatically combine such events at the end of the search pipeline. Some algorithms that are currently in common use are friends of friends (FoF) and DBSCAN (Ester et al., 1996; Deneva et al., 2009; Barsdell, 2012). However, few clustering algorithms have been rigorously tested.

Throughout this chapter, our primary motivation was to identify the optimal clustering algorithm for single-pulse searches, in particular in searches for fast radio bursts (FRBs).

The REALFAST system at Very Large Array (VLA) is one such commensal fast transient search system (see Chapter 6). The REALFAST pipeline focuses on searching for transients at multiple DMs and trial widths, each of which is then post-processed. A total intensity (Stokes I) image is formed for each trial DM and width. Point sources in these images with S/N greater than a pre-set threshold trigger the detection pipeline. The data corresponding to each candidate are then saved to disk and is classified using a Deep Learning based classifier (Agarwal et al., 2020a). Visualizations that show the radio image, spectrogram, spectra, and profile of the candidate are then generated. These visualizations also consist of other relevant candidate parameters: signal to noise ratio (S/N), DM, width, relative sky position with respect to the pointing center, scan number, etc., and are used for follow-up inspection.

In this chapter, we use REALFAST data as a test-case to explore and compare candidate-clustering techniques. We also generalize our results to apply to single-dish telescopes, which do not have spatial (sky-location) information to use in a clustering algorithm. This chapter is laid out as follows: In §3.3 we provide a more detailed motivation for the need for clustering and a discussion of clustering methodologies. §3.4 describes the data used for testing the algorithm, followed by methods explained

in §3.5. The results of the analysis are presented in §3.6, followed by discussion and conclusion in §3.7 and §3.8, respectively.

3.3 Clustering

As mentioned previously, clustering is implemented between the search and the candidate processing steps of the pipeline. In the REALFAST system, after clustering, we choose the maximum S/N candidate from each cluster, and only those are analyzed in the candidate processing step. We also consider all the unclustered candidates as individual clusters of size one and pass them onward for processing.

In this section, we discuss the need to use clustering in the context of a single pulse search pipeline. Further, we use the following terminology throughout this chapter:

- Event: The actual physical occurrence of an astrophysical transient (e.g., FRB, pulsar) or Radio Frequency Interference (RFI).
- Candidate: A single detection reported by a search pipeline. It typically consists of a set of properties (sky location, DM, time, etc). Candidates may be triggered by random thermal noise or associated with an event. Multiple candidates can be associated to a single event.
- Observation: A set of candidates generated after the search pipeline is run on some data. It can be real or simulated and can have candidates associated with FRB or RFI or both.
- Dataset: A set of observations.
- Cluster: A group of candidates (or members) with the same labels assigned by a clustering algorithm.
- Member: Candidates within a cluster.
- True Labels: Each member of a cluster is associated with an event. We refer to this event as the true label of that member.

- Real/FRB/Transient: Event, cluster, or member associated with an astrophysical transient.
- RFI: Event, cluster, or member that is not astrophysical.

3.3.1 Expected number of candidates from a single astrophysical event

As explained in the previous section, following search parameters are reported for each candidate detected by the REALFAST search pipeline: DM, time of occurrence of the candidate, relative sky position with respect to the pointing center (l, m), S/N, and width. Moreover, for each event, the pipeline can return multiple candidates at nearby (incorrect) values of DM, width, and time. The observed S/N of a candidate detected at a trial DM, width, sky location is given by (assuming no other losses, etc.)

$$S/N_{\text{observed}} = S/N_{\text{optimal}} \cdot F_{\text{widthloss}} \cdot F_{\text{beamloss}} \cdot F_{\text{DMloss}}, \quad (3.1)$$

where S/N_{optimal} is the S/N of the candidate when there is no loss. $F_{\text{widthloss}}$ is the loss due to incorrect boxcar width (Cordes & McLaughlin, 2003), F_{beamloss} is the loss due to position of candidate within the primary beam of the telescope, and F_{DMloss} is the loss due to incorrect DM value (Levin, 2012, §2.3).

Using this equation, we can calculate the number of candidates that will trigger a single pulse search system for an astrophysical event. For instance, we assume a VLA L-band (1–2 GHz) configuration with 256 frequency channels and a time resolution of 5 ms, and that REALFAST system is used to search for transients. We then search for DMs from 0 pc cm^{−3} to 3000 pc cm^{−3} and set $F_{\text{dmloss}} = 0.95$, i.e, up to a maximum of 5% loss in sensitivity between DM trials. Using this, and assuming an intrinsic pulse width of 30 ms, we can compute the DM array (Levin, 2012, §2.3). We also assume $t_{\text{scatt}} = 0$ as it is line-of-sight dependent and is typically small. We set our boxcar search widths to 5 ms, 10 ms, 20 ms, and 40 ms and S/N detection threshold at 8. Figure 3.1 shows the number of candidates detected with respect to input S/N of the transient for two different values of transient DMs and widths. Here, we have also assumed that the candidate is at the center of the beam, therefore the F_{beamloss}

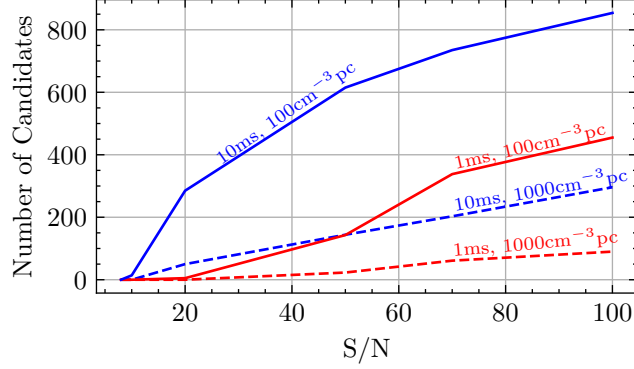


Figure 3.1: The number of candidates generated by a single pulse search pipeline for events with varying S/N. Different colors represent different intrinsic widths of the transient, and different line styles show different DMs. Observing and search configuration similar to that of REALFAST at L-band was chosen (Sampling time: 5 ms, Number of frequency channels: 256, Bandwidth: 1 GHz, DM range: 0 to 3,000 pc cm⁻³, Boxcar widths: [5 ms, 10 ms, 20 ms, 40 ms], S/N threshold: 8). See Section 3.3.1 for more details.

is 1. This figure clearly shows that the number of candidates detected by the pipeline can be large even for one event. This can overwhelm the real-time systems that are responsible for post-processing these candidates and writing their data to disk, hence motivating the use of clustering algorithms.

3.3.2 Unsupervised Clustering

In this chapter, we have taken the approach of unsupervised clustering. Here we briefly discuss unsupervised clustering and some of its caveats. Unsupervised clustering is the method used to find meaningful clusters from an unlabeled dataset, i.e., *a priori* information about number of clusters and the true cluster assigned for each candidate is not known (Jain et al., 1999). This is in contrast to supervised clustering for which this information is available. In this analysis, we do not know the clustering information of the candidates, therefore we opted for unsupervised clustering.

Unsupervised clustering is typically done by estimating the “distance” or “similarity” between different candidates, with the aim that candidates with low “distance” might be similar and belong to the same cluster. The clustering algorithms we dis-

cuss in this chapter use standard search pipeline features, without any expensive pre-processing, and can find reliable clusters in real-time. In some cases, physically meaningful relations between features could also be computed to enhance clustering performance, however we do not employ these here (Pang et al., 2018, §3.2).

The caveat to this approach is that unsupervised clustering techniques can be harder than supervised methods to tune for specific datasets. Also, due to the lack of true labels, it is difficult to evaluate the clustering performance. We discuss the above caveats further in Section 3.5.

3.3.3 Clustering RFI

Strong RFI events can also overwhelm the real-time pipelines by generating a large number of candidates, sometimes at all DM trials. Even though we use multiple filtration techniques to mitigate RFI, some signals still reach the pipeline’s clustering step. Narrow-band RFI can lead to many candidates at all DMs in the DM grid (and, because of the resulting time shift of the peak, corresponding time bins). In some cases, the RFI appears as a strong localized source in the radio image and hence is present as a dense cluster of points in the image plane (for instance, this can happen with a sufficiently high-altitude satellite). Therefore, we can leverage the clustering algorithm to cluster those thousands of triggers into one cluster, reducing the computational load by orders of magnitude. As it is not feasible to manually label RFI examples into multiple clusters, we cannot evaluate clustering algorithms’ performance on identifying separate RFI clusters. Instead, we only estimate clustering performance on FRB clusters. This is further discussed in Section 3.5.3.

3.4 Data

Here we describe the details of our dataset and the features we used for clustering. We used REALFAST data to generate a dataset containing representative RFI, and used simulated FRBs to generate a dataset with representative candidates. We then combined the two datasets and applied four different pre-processing techniques

(downsampling and normalization) on the features of the candidates to simulate 250 observations, consisting of candidates from both real and RFI events.

3.4.1 Feature Selection for Clustering

As mentioned in Section 3.2, the pipeline reports a set of measured parameters for each candidate that satisfies the S/N threshold criterion. For our clustering analysis, we cluster based on DM, time (as is the standard with most past FRB searches), and sky position (with relative direction cosines represented by l and m as angular distances from the observation’s pointing center) of the candidates. Candidates associated with an FRB event are expected to be densely located in l and m , as the FRB originates from a specific location in the sky. They would also show an expected S/N decrease in adjacent DM and time values (see Section 3.3.1) and would be closer for those parameters. On the other hand, RFI is randomly spread across this parameter space, but strong RFI can show a trend in DM and time if detected at multiple DMs.

3.4.2 RFI database

In this analysis, we used data from various commensal and commissioning observations of REALFAST system (project codes: 19A-242, 20A-330, 19B-223, 20A-163) in which the standard REALFAST pipeline detected only RFI candidates. This data spans a range of array configurations and other observing and search parameters (frequency, bandwidth, image pixel size, etc.). To create candidates representative of the real-time pipeline, we re-ran the REALFAST transient search on this data with the pipeline using default search parameters (Law et al., 2018, 2020). As clustering performance is expected to be sensitive to the RFI environment, we selectively chose datasets with a variety of RFI types. These datasets are representative of the broad range of RFI we have seen at VLA and therefore form a robust sample of RFI for our analysis. This procedure was used to generate RFI candidates from 13 observations, with a few to ~ 6000 candidates each. We manually verified that all these candidates were RFI and saved parameters relevant for clustering for each candidate (§3.4.1).

Table 3.1: Parameter distributions of simulated FRBs

Parameter	Distribution	Range/Values
S/N	Uniform	10, 40
DM (pc cm^{-3})	Uniform	10, <code>max_search_dm</code> ^a
Width (ms)	Uniform	1, 40 (ms)
Frequency	Uniform	L, S, C, X
Array configuration	Uniform	A, B, C, D ^b
Sky position (l, m)	Uniform	-fov/2, fov/2 ^c

^a `max_search_dm` is the maximum DM searched for a given configuration

^b Maximum baseline lengths for the four configurations (A, B, C, D) are 36.4 km, 11.1 km, 3.4 km, 1.03 km.

^c fov is the field of view at the randomly chosen frequency.

We will refer to this as the RFI dataset.

3.4.3 Simulating and Injecting FRBs

We also generated a dataset of “real” candidates, representing our signals of interest. This was done by generating simulated data (with standard radiometer noise for different array configurations, observing, and search parameters) and injecting simulated transients. The distribution of parameters used for injecting transients is given in Table 3.1. We searched this simulated data using REALFAST system with real-time search configuration to generate candidates. We saved the relevant parameters of all the candidates, and manually verified them to make sure that each candidate corresponds to the injected transient. We discarded any observation with less than four candidates. This procedure was used to generate real candidates from 114 simulated observations (with one transient injected in each observation). We will refer to this as the FRB dataset.

3.4.4 Test Dataset

To evaluate the performance of various clustering methods (described in Section 3.5.1), we generated a test dataset. We used this dataset to compare the performance of different pre-processing techniques, and also during hyperparameter tuning

(see Section 3.5.2).

The test dataset consists of multiple observations, each containing some RFI candidates and some FRB candidates. We enforce that each observation has one transient event, and so all FRB candidates in an observation would be associated to that single event. Therefore, a perfect clustering algorithm should form only one FRB cluster per observation. To generate such an observation, we randomly chose one observation each from the RFI and FRB dataset pool. We then randomly select $X\%$ of RFI candidates (where X is sampled from a uniform distribution between 20-100) from the RFI observation, all the FRB candidates from the FRB observation, and concatenate their features. We then randomise the order of the examples. This creates a set consisting of both RFI and FRB candidates. All observations with less than 10 total number of candidates were discarded. Using this process, we created 250 observations containing RFI and FRB candidates, which formed our test dataset.

3.4.5 Pre-processing

Pre-processing is the procedure that takes the event features and converts them into indexed parameter ranges such that all the parameters will be equally weighted in terms of their importance in the clustering.

As explained in Section 3.4.1, we use DM, time, l and m as features for clustering. Therefore, for each candidate in each observation of our database, we save these four parameters along with the image S/N of the candidate. After clustering, we use S/N to decide the representative candidate from each cluster. We convert the absolute value of DM to an index based on its index in the DM array for each candidate. Similarly, we also convert the time value (in seconds) to an index, based on the sample number corresponding to that time from the start of that processing segment. We also convert l and m (which is the offset of the candidate from primary beam center) to corresponding pixel values, using the synthesized beam size. Therefore, we convert all the features to corresponding indices. This is necessary as otherwise, different scales of different features might bias the distance estimates required in clustering.

The transient events we are interested in appear as a point source in the sky.

Therefore, all the candidates from that transient should constitute a small range of l and m index values. We downsampled the l and m values of all the candidates to increase the sky density of candidates, which might enhance the clustering performance. We tried downsampling factors of 1, 2 and 4 (henceforth referred to as **DS1**, **DS2** and **DS4** respectively). Although we have scaled all the features to their corresponding indices, we also evaluated clustering performance on standardized data (i.e, with zero mean and unit variance, hereafter referred to as **Norm**). Throughout the chapter, we report the performance of all the algorithms on all these different pre-processing techniques. We also try to determine the pre-processing technique which leads to the best clustering performance.

3.5 Methods

3.5.1 Clustering Algorithms

We compare eight algorithms to cluster our test dataset: K-means, Mean Shift, Affinity Propagation, Agglomerative Clustering, DBSCAN, Optics, HDBSCAN, and Birch. For all except HDBSCAN, we use the implementation of these algorithms in **scikit-learn** (Pedregosa et al., 2011; Buitinck et al., 2013). We use the python implementation of HDBSCAN by McInnes et al. (2017). We briefly discuss the details of these algorithms and their hyperparameters in Section 3.9.1. We refer the reader to the respective papers and **scikit-learn** documentation for more details.

3.5.2 Hyperparameter Tuning

Each clustering algorithm has several input parameters that can be used to control the algorithm’s clustering process and speed. These input parameters are called hyperparameters. Some algorithms are very sensitive to the choice of these hyperparameters while others are robust to a range of hyperparameters. Our aim is to find the hyperparameters for each algorithm, which leads to the best clustering performance (called optimal hyperparameters). The following three techniques are

typically used to obtain the optimal hyperparameters: brute force grid search, random sampling, and Bayesian optimization. Grid search involves generating a grid of points covering the whole parameter space uniformly. The performance metric is then calculated on all the grid points, and the hyperparameter combination with the maximum value of the metric is chosen. In random sampling, the hyperparameter combinations are randomly chosen from a distribution of parameters. Bayesian optimization uses Bayesian techniques to parse the parameter space and obtain the optimal hyperparameter combination.

Random sampling has been shown to be better than brute force grid search (Bergstra & Bengio, 2012). This is because, in most cases, only a few hyper-parameters really matter, the importance of which changes with different datasets. This makes grid search a poor choice for searching for hyper-parameters for different datasets. Therefore, we opted to use random sampling. Also, because our parameter space is not very large, we decided not to use Bayesian optimization. Section 3.9.2 shows the ranges and various possible values of different hyperparameters that were tried for each algorithm. Wherever necessary, we used a random state of 1996 in the algorithms for reproducibility.

3.5.3 Performance Metric

The general idea of using a performance metric is to have a common reference point to rank the effectiveness of clustering algorithms (and their hyperparameters). The one with the maximum value of the metric has the best general performance.

Critical to this idea is a clear statement of our goals. Our primary measurable goals with clustering are the following:

1. **Avoid missing a genuine event due to clustering.** This can happen due to over-aggressive clustering that identifies FRB candidates as false members of an RFI cluster. As only the highest S/N member from each cluster is processed further, assigning FRB members to RFI clusters will lead to FRB candidate not passing further in the pipeline. This can happen if there are RFI candidates

with S/N higher than that of the FRB candidates.

2. **Each event of interest should be singly identified.** All candidates from one FRB event should be clustered into one cluster, rather than many small separate clusters representing a single event of interest. This is to minimize the number of candidates that are passed to the pipeline for post-processing and classification.

To represent these goals, we have developed the following metric. A higher value of the metric is favourable. We use *homogeneity*, *completeness*, *v-measure* and *recall* to calculate the metric (hereafter referred to as *score*). In the following, an FRB cluster is defined as a cluster containing one or more FRB candidates, obtained after clustering. In the following discussion, we follow the terminology defined in Section 3.3.

3.5.3.1 Homogeneity

Homogeneity is the measure of purity of the clusters with respect to true labels, i.e it estimates if each cluster contains only members of a single class (i.e either FRB or RFI). We calculate homogeneity for each observation in the test dataset. As we are primarily interested in performance on FRB candidates, and as RFI can be clustered into multiple clusters (for which we don't have true information), we define homogeneity only for FRB clusters in the observation. For each FRB cluster, we calculate the ratio of the number of FRB candidates in that cluster to total number of candidates in that cluster. Homogeneity (h) is the weighted average of all these ratios, weighing them by the number of candidates in the cluster. Hence,

$$\begin{aligned} h &= \frac{1}{N_T} \sum_i \frac{n_{FRB}^i}{n_T^i} n_T^i \\ &= \frac{1}{N_T} \sum_i n_{FRB}^i, \end{aligned} \tag{3.2}$$

where i represents the i th FRB cluster, and the sum is over all the FRB clusters. n_{FRB}^i is the number of FRBs in the i th FRB cluster, n_T^i is the total number of candidates

in that cluster. $N_T = \sum_i n_T^i$ is the total number of candidates in all FRB clusters. h can be between 0 (when all FRB candidates are left unclustered) and 1 (when all FRB clusters contain only FRB candidates).

3.5.3.2 Completeness

Completeness is used to estimate if all members of a given class are assigned to the same cluster. We calculate completeness for each observation in the test dataset. We define completeness for FRB clusters, and a high completeness score will minimize the number of clusters the FRB candidates are clustered to. For each FRB cluster, we calculate the ratio of the number of FRB candidates in that cluster to the total number of FRB candidates. Completeness (c) is equal to the weighted average of all these ratios, weighing them by the number of candidates in the cluster¹. Hence,

$$c = \frac{1}{N_T} \sum_i \frac{n_{FRB}^i}{N_{FRB}} n_T^i, \quad (3.3)$$

where i represents i th FRB cluster, and the sum is over all the FRB clusters. n_{FRB}^i is the number of FRBs in i th FRB cluster, N_{FRB} is the total number of FRB candidates in that observation. $N_T = \sum_i n_T^i$ is the total number of candidates in all FRB clusters. c will be very small if all FRB candidates are assigned different clusters, and 1 when all FRB candidates are clustered into one cluster.

3.5.3.3 V-measure

V-measure is the harmonic mean between homogeneity and completeness. This is used as we want all the clusters to be pure and favor minimum number of clusters. Therefore, we want to maximize both homogeneity and completeness. V-measure (v) will be 1 when both h and c are 1 and will be 0 if either of those is 0. Hence,

$$v = \frac{2hc}{h+c}. \quad (3.4)$$

¹We include unclustered candidates as a single cluster in this case, while unclustered candidates were ignored while calculating homogeneity.

We calculate h , c , v for each observation in the test dataset and take a weighted average of all v 's (weighting by the total number of candidates in that observation) to get an estimate of V-measure for the whole dataset (V).

3.5.3.4 Recall

Recall is the fraction of FRBs that are recovered after clustering. After clustering, only the candidates with maximum S/N in each cluster are processed further in the pipeline. Therefore, if the clustering algorithm clusters FRB candidates together with high S/N RFI candidates, then the FRB will not be recovered from that cluster and might be missed. Therefore, recall (R) is defined as the ratio of the number of observations in the dataset for which FRB was recovered to the total number of observations in the dataset.

3.5.3.5 Score

Score is defined as the product of recall (R) and total V-measure (V).

$$Score = R \times V. \quad (3.5)$$

We use this score to compare the clustering performance of different algorithms and find the optimal hyperparameters for each clustering algorithm.

3.5.4 Advantages of this metric

We have defined the above metric with respect to FRB and RFI clusters, but it can be easily generalized to any application with goals generally similar to those laid out in Section 3.5.3. This metric ensures that the information in relevant clusters is not missed by overaggressive clustering while still minimizing the number of clusters formed. There are some advantages of this metric over other clustering metrics available in the literature (for a detailed comparison using a similar metric see Rosenberg & Hirschberg, 2007): 1) It is independent of the clustering algorithm, size

of the dataset, number of classes, and clusters. 2) It can appropriately use one (or more) base class (here FRB) to evaluate the clustering performance with respect to true labels, considering all data points of the relevant class. 3) Using homogeneity and completeness, V-measure gives importance to both pure clusters and minimum number of clusters. 4) By adequately weighting individual metrics, it is possible to concisely evaluate the clustering performance across multiple examples, in the form of a simple number (score).

3.6 Results

3.6.1 Optimal Hyperparameters

We show the maximum score obtained for each algorithm after hyperparameter search in Figure 3.2. Henceforth in this chapter, we will refer to these hyperparameters as “optimal hyperparameters”. This figure shows that Mean Shift has the maximum score, out of all the eight algorithms. Table 3.2 shows these optimal hyperparameters for each algorithm.

Figure 3.3 shows the distributions of scores at various hyperparameter values, for each combination of algorithm and pre-processing. Although not crucial to hyperparameter selection, the distribution of scores gives an insight into the robustness of the algorithm to the choice of input hyperparameters. In some cases, the scores vary between the full range of 0 and 1, while in others, the distribution is very narrow around high scores. The algorithms for which the score distributions peak around a high value should be more robust to the input hyperparameters than those for which the peak is at a middle or even low value of score. In the latter case, only a small range of hyperparameters would lead to a high score.

3.6.2 Effect of data processing

As mentioned in Section 3.4.5, we also repeated the above experiment after pre-processing the data in two ways: downsampling the l and m indices and data

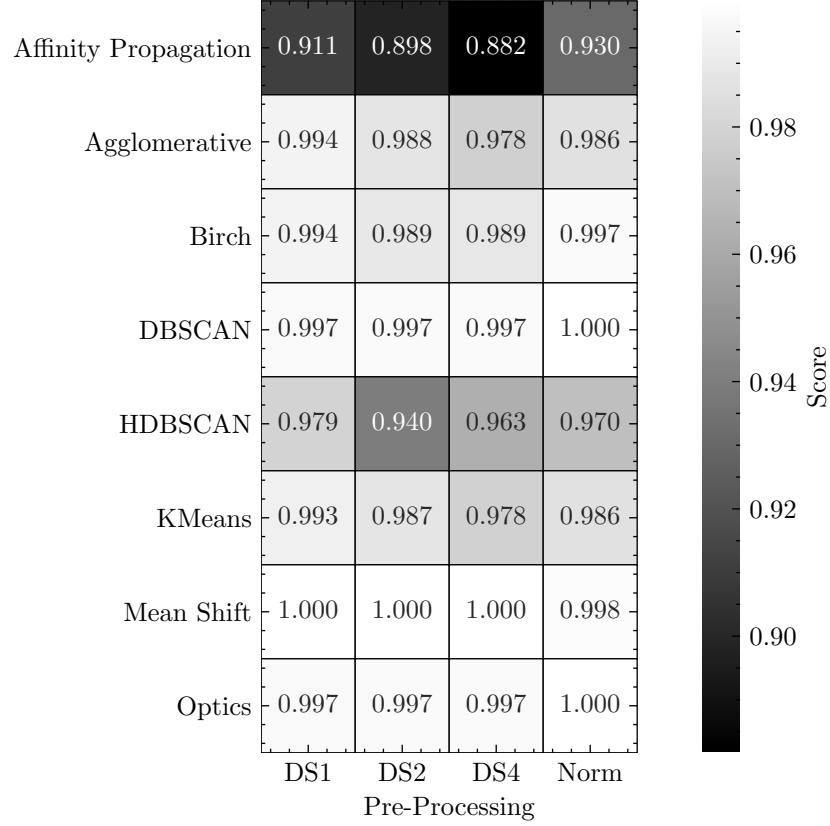


Figure 3.2: Maximum score obtained after random hyperparameter search (at the optimal hyperparameters) for each algorithm and pre-processing combination. Optimal hyperparameters for each case are given in Table 3.2. DS refers to downsampling applied to l and m indices. Norm refers to normalization of the four features (see Section 3.4.5)

Table 3.2: Optimal hyperparameters obtained for different algorithm and pre-processing combinations.

Algorithm	Hyperparameter	DS1	DS2	DS4	Norm
Affinity Propagation	affinity	euclidean	euclidean	euclidean	euclidean
	random_state	1996	1996	1996	1996
	damping	0.974	0.965	0.985	0.881
	preference	-884	-222	-219	-202
Agglomerative	n_clusters	5	7	6	3
	affinity	euclidean	manhattan	euclidean	euclidean
	compute_full_tree	auto	auto	auto	auto
Birch	linkage	ward	average	ward	ward
	n_clusters	5.000	7.000	6.000	10.000
	threshold	0.341	0.876	0.676	0.957
DBSCAN	branching_factor	13.000	56.000	85.000	77.000
	min_samples	2	2	2	2
	eps	14.163	14.726	14.615	1.082
	metric	chebyshev	chebyshev	chebyshev	cityblock
	algorithm	auto	auto	auto	auto
HDBSCAN	leaf_size	23	21	35	38
	min_samples	5	5	5	5
	metric	euclidean	euclidean	euclidean	cityblock
	min_cluster_size	2	3	2	9
	cluster_selection_method	eom	eom	eom	eom
KMeans	allow_single_cluster	True	True	False	True
	algorithm	full	elkan	full	auto
	n_clusters	5	6	6	3
	n_init	13	15	28	26
Mean Shift	random_state	1996	1996	1996	1996
	bandwidth	16.416	32.750	19.350	1.229
	bin_seeding	True	False	True	True
	cluster_all	True	True	True	True
Optics	min_samples	2	2	2	2
	eps	14.782	14.376	14.551	1.095
	metric	minkowski	chebyshev	minkowski	cityblock
	min_cluster_size	8	6	6	8
	p	14.672	-	11.009	-
	cluster_method	dbscan	dbscan	dbscan	dbscan
	xi	-	-	-	-

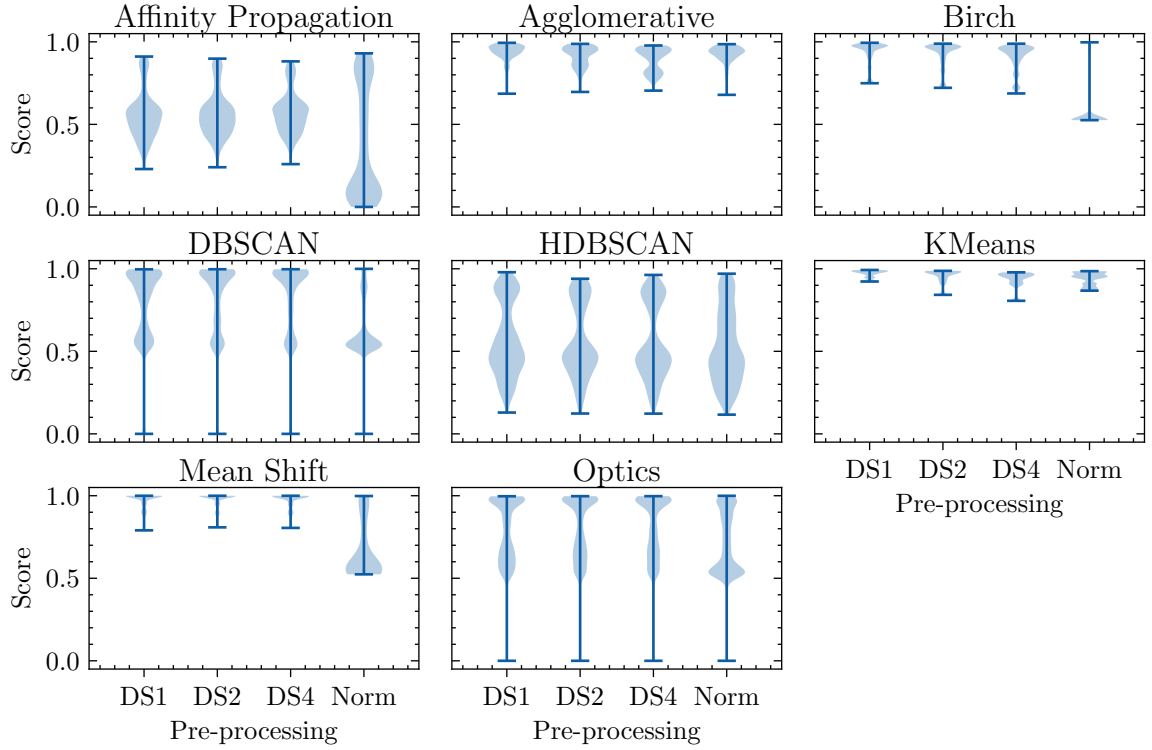


Figure 3.3: Violin plots of score vs. pre-processing for different clustering algorithms. Each violin plot shows the distribution of scores obtained at various hyperparameters evaluated during random hyperparameter search. Different sub-figures represent different algorithms (Sections 3.5.2, 3.6.1). DS refers to downsampling applied to l and m indices. Norm refers to normalization of the four features (see Section 3.4.5).

normalization. We show the results for this in Figures 3.2 and 3.3. As can be seen from Figure 3.2, there is no clear trend of clustering performance for different pre-processing cases. We also note that the shape of the score distribution remains the same across different downsampling factors for a given algorithm. This indicates that downsampling doesn't have a significant contribution to the clustering performance.

3.6.3 Evaluating performance on clean data

So far, we have evaluated the performance of clustering algorithms on data with real RFI candidates along with simulated FRB candidates. This, as stated earlier, was a reasonable approximation of the candidates from real observations. Usually, in the case of a real transient, the pipeline only gets triggered at candidates from real transient, and no RFI is seen (either because of the amplitude of real transient or because low-level RFI is flagged). Therefore, here we report the performance of these clustering algorithms (at the optimal hyperparameters) on a dataset containing candidates only from a real event and no RFI. This is done to test the generalisability of these clustering algorithms on data without RFI. This would also serve as an independent test on unseen datasets for which the hyperparameters of the algorithms weren't tuned.

3.6.3.1 Completeness on Clean data

We use the same procedure as described in Section 3.4.3, to generate a dataset of 100 observations with candidates from one simulated FRB each. We randomly chose the parameters of the simulated FRBs and observing configurations, as explained earlier and discarded any observation with less than ten candidates.

We use *completeness* (see Section 3.5.3) to report the clustering performance on this dataset. As there is no RFI candidate in this dataset, homogeneity would always be one and therefore is not a useful metric in this case. Here, a perfect clustering algorithm should generate just one cluster per observation for which completeness would be maximum, declining as the number of clusters increase. The overall completeness

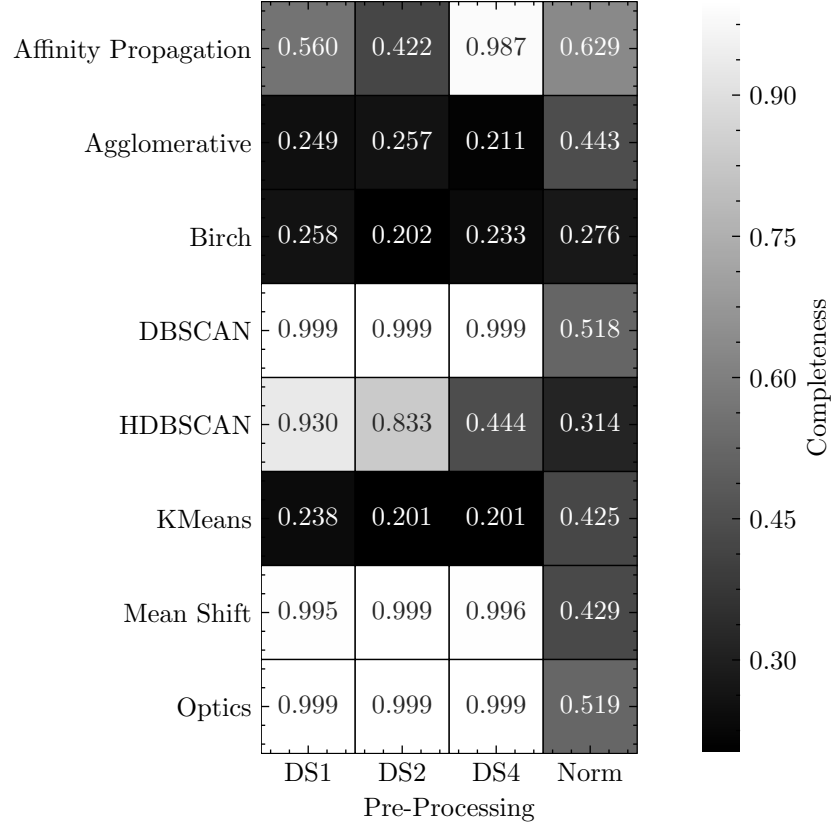


Figure 3.4: Completeness of different algorithms on clean data i.e without any RFI candidate (Section 3.6.3). High completeness score is better and would imply that the FRB candidates are clustered in a minimum number of clusters for each of the 100 observations. Each algorithm was evaluated at its optimal hyperparameters (Table 3.2). DS refers to downsampling applied to l and m indices. Norm refers to normalization of the four features (see Section 3.4.5).

for a dataset is the average of all the completeness values from 100 observations, each weighted by the number of candidates in the observation.

Figure 3.4 shows the overall completeness score of each algorithm. DBSCAN, HDBSCAN, Mean Shift and Optics have the highest completeness score. It is to be noted that the completeness score of these four algorithms was worse when the data was pre-processed to zero mean and unit standard deviation (i.e Norm). On the contrary, downsampling the image features did not show any significant effect (with a notable exception of DS4 for HDBSCAN).

3.6.4 Benchmarking

We evaluated the clustering speed of all clustering algorithms at their optimal hyperparameters. To do this, we generated an observation with a varying number of candidates consisting of random values for the four features. We then ran all the clustering algorithms on those observations and recorded the time taken for just the clustering step. We did this test with optimal hyperparameters obtained for all four pre-processing cases. As the clustering speed is primarily dependent on the number of candidates to be clustered, we didn't use real data for this test. We show the result of this test in Figure 3.5. The time taken did not vary significantly with parameters from different pre-processing cases, so we only show results using optimal hyperparameters for DS1 in this figure. DBSCAN and HDBSCAN are the fastest of these algorithms, while Affinity Propagation, Mean Shift, and Optics are the slowest, by at least an order of magnitude.

3.7 Discussion

3.7.1 Feature Importance

It is worth understanding the impact of feature selection on our outcome, as some features are expected to be more important than the others (Guyon et al., 2005; Dash & Liu, 2000). We use a Random Forest classifier (Breiman, 2001), implemented

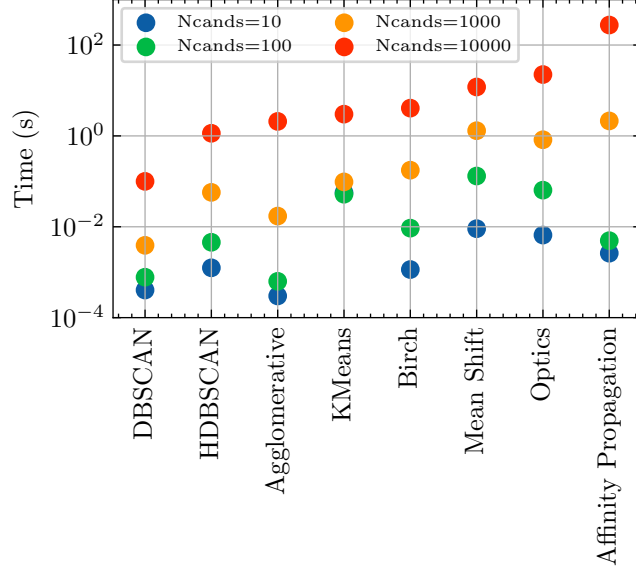


Figure 3.5: Time taken to cluster (in seconds) for each algorithm at their optimal hyperparameters. Different colors represent input data with different number of candidates. Results are shown only at optimal hyperparameters for DS1. DBSCAN and HDBSCAN are much faster than algorithms like Mean Shift and Affinity Propagation (Section 3.6.4).

in `scikit-learn`, to estimate the relative feature importance of the four features, (DM, time, l , m), in determining accurate clusters.

We used the test dataset (see Sec. 3.4.4) without any pre-processing, containing candidates from 250 observations (hereafter we refer to this data as DS1). We knew the true labels (RFI and FRB) for each candidate in those 250 observations. For each observation, we trained a Random Forest classifier (at the default input parameters) using all the candidates in that observation. From the trained classifier, we then used `feature_importances_` to obtain the relative feature importance of each of the four features. This attribute of the Random Forest classifier calculates Gini importance (Breiman, 2001) for each feature, which is representative of the importance of feature during classification. We repeated this for all the observations in our test dataset. To estimate the total feature importance, we averaged all the importance for each feature weighing each by the number of candidates in that observation. The total importance obtained is shown in Figure 3.6. As can be seen from this figure, l and m (sky position indices of the candidate) contribute much more towards classification

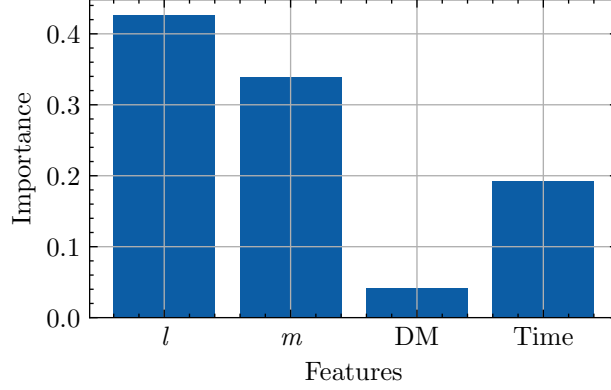


Figure 3.6: Importance of each feature, determined by training a Random Forest classifier to classify each observation into RFI and FRB. We trained the classifier individually on all observations in the test dataset and took a weighted average of the individual feature importance to obtain the above plot. l and m contribute much more towards classification than DM and Time (Section 3.7.1).

than the DM and Time indices of the candidates. A caveat to this simplistic analysis is that classification of all candidates into two classes, FRB and RFI, is not the same as clustering them into multiple clusters. The two cases would have been similar if all the RFI candidates could be assigned to a single cluster, which is not true. Therefore, even though this analysis shows that sky position contributes much more to classification, we suspect that the relative contribution of DM and Time for the clustering task would be higher than what is obtained here.

3.7.2 What if I only use DM and time for clustering?

In REALFAST system, we search for transients on the radio image. Therefore, for each candidate we get DM, Time, l and m information. But in many experiments, typically the ones using a single dish telescope or the ones not performing an image-based transient search, only DM and time information is available for each detected candidate. Therefore, in those cases, only DM and time can be used to cluster the candidates together.

We tested the clustering performance using only DM and time to cluster the observations in our test dataset. We used the optimal hyperparameters (listed in Table 3.2) on the test dataset to evaluate this for all pre-processing cases. We also tested

the clustering performance using only l and m indices to cluster our test dataset. As discussed in the previous section, the relative importance of sky positions is much higher than that of DM and time for a classification task. Therefore, clustering using only sky positions should give better scores than using just DM and time.

We show the results of these two tests in Figure 3.7 along with the scores when all four features are used for clustering. We only show scores for one pre-processing case (DS1), as results with other pre-processing techniques were also similar. As can be seen from this figure, scores obtained using just sky positions (red curve) or DM and Time (blue curve) follow each other closely. Using sky positions shows minor improvement in score for most of the algorithms. Using all four features, as expected, gives the highest score that is $\sim 10\%$ better than the other two cases.

This test highlights the importance of using sky positions along with the standard DM-time features to identify clusters of candidates originating from the same event. Therefore, if the sky position information is available for a candidate, it should also be used while clustering in the pipeline. With more and more interferometers (like ASKAP and DSA-110) implementing a REALFAST like search for transients on radio images in the future, it would be useful for them to incorporate sky position information to cluster candidates in their respective pipelines.

As careful readers would have noticed, a caveat to this test is that in clustering with two parameters, we didn't do a hyperparameter search to obtain the optimal hyperparameters that maximize the score using those two features. Instead, we used the hyperparameters that were optimal when four features were used. A full hyperparameter search using two parameters might lead to a different set of parameters that might improve the score further. But even with this simple test, it can be noted that only using sky positions for clustering gives an improvement in score in almost all cases.

We have demonstrated in this and the previous section that sky positions are overall more important for clustering than DM and Time. This could be because RFI candidates are more likely to span a wide range of time and DM values, which might overlap with those of FRB candidates, while they are still localized in the radio image.

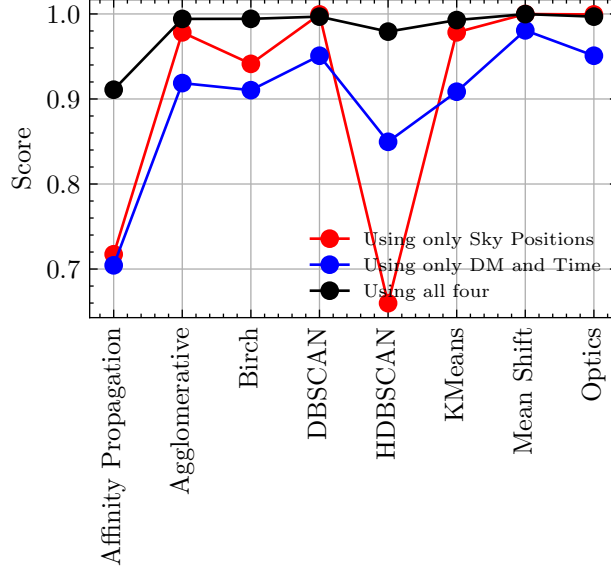


Figure 3.7: Score vs. algorithms for two feature clustering. Different colors represent different sets of features used to perform the clustering. We evaluated the scores on the test dataset. Results with DS1 pre-processing are shown here (Section 3.7.2).

Therefore, it is less likely (though still possible) for RFI to be very close to an FRB in the radio image. Similarly, RFI may be highly variable in frequency/time space, whereas in a radio image even unfocused (near-field) RFI will show up as contiguous streaks or other similarly structured patterns in images. Regardless of the reason for this, however, we have demonstrated here that when possible, sky positions should be used for clustering candidates.

3.7.3 But which algorithm should I use?

There are several considerations when deciding what algorithm to use based on the comparative analysis we have presented here.

- **Maximum Score:** As discussed in Section 3.5.3, we want the clustering algorithm to meet our application-specific goals; of not missing a genuine event and singly identifying FRB candidates. Our performance metric (called score) maximizes when these goals are met. Therefore, we could search for a set of optimal hyperparameters for each clustering algorithm, that gives the maximum score. All algorithms, except Affinity Propagation, have an optimal score above 0.95

(Figure 3.2).

- **Generalisable:** The clustering algorithm needs to generalize to various types of data it can encounter in the pipeline. By testing the algorithms and optimal hyperparameters obtained in the previous step on an independent dataset, one could quantify the algorithms’ generalisability. To be more application-specific, we tested this on a dataset with observations containing candidates only from a real event, without any RFI, and computed the completeness as the performance metric. Only four algorithms, DBSCAN, HDBSCAN, Mean Shift, and Optics gave completeness above 0.9 in this test (Figure 3.4).
- **Speed:** Finally, the clustering algorithm would only have a limited amount of time to cluster candidates. Therefore, even for a large number of candidates, it should not exceed the limited time constraint. In our specific application for REALFAST, clustering is performed on candidates generated from small segments of data that are tens of seconds long. Based on the other pipeline steps, clustering shouldn’t take longer than a few seconds. The number of candidates detected by the search step typically varies between a few to thousands of candidates for a segment. Based on these requirements DBSCAN, HDBSCAN, Agglomerative, and K-means can be used (Figure 3.5).

As an example using the REALFAST system, selecting the algorithms using the above three steps, we conclude that either DBSCAN or HDBSCAN can be used for clustering REALFAST data. Based on the results in Figures 3.4 and 3.7 we can further infer that DBSCAN is better than HDBSCAN. As reported earlier, we didn’t notice any improvement by using different pre-processing techniques, therefore no pre-processing is favored (Figure 3.2). Based on the results of this analysis, REALFAST system changed from HDBSCAN to DBSCAN in the real-time search pipeline. A similar procedure can also be used to choose the clustering algorithm for any other single-pulse search pipeline or even for a more general clustering application.

3.8 Conclusions

In this chapter, we have compared eight different unsupervised algorithms to cluster candidates generated by single-pulse search pipelines. We have also analyzed the effects of various pre-processing techniques on the data. We used real RFI from REALFAST system and simulated FRB candidates to test different algorithms. We have developed a performance metric to quantify clustering performance. This metric makes sure that FRBs are not missed due to overaggressive clustering while still minimizing the number of clusters formed. Using a random hyperparameter search, we obtained optimal hyperparameters, which maximizes this metric for different algorithms. We test all the algorithms with optimal hyperparameters on an independent dataset consisting of only FRB candidates to evaluate the generalisability of different algorithms. We also estimated the average clustering time for various algorithms on a dataset of varying sizes. Finally, we have proposed a strategy that can be used to choose a clustering algorithm, using various tests mentioned earlier. We apply this strategy to obtain a clustering algorithm appropriate for REALFAST system. This strategy can also be used at other single-pulse search systems to obtain the optimal clustering algorithm. Our strategy is generic enough to be used for other clustering applications. Our performance metric can also be used in other clustering applications where clustering information for only one cluster of interest is available, out of an unknown number of true clusters. We have also demonstrated that using spatial features for clustering improves the clustering performance compared to the traditional approach of just using DM and time features. All the scripts used in this analysis are openly available in a Github repository².

²<https://github.com/KshitijAggarwal/rfclustering>

3.9 Appendix

3.9.1 Clustering Algorithms

Here we give a brief overview of all the clustering algorithms used in this analysis and some details and potential advantages/disadvantages of each algorithm for our clustering application.

3.9.1.1 K-means

K-means (Macqueen, 1967) algorithm is one of the most widely used clustering algorithms. It is extremely easy to implement and computationally very efficient. Given an input number of clusters, the algorithm randomly initializes centroids for each cluster. Each example is then assigned a cluster based on the distance from that centroid. A new centroid is then computed for each cluster, and all examples are reassigned to the new centroids. This process is repeated till a convergence criterion is met. The main challenges with K-means are that it is not good at identifying non-spherical clusters and requires the number of clusters as input, both of which limit its ability to generalize on different datasets. As K-means starts with an initial random assignment of cluster centroids, the output clusters can also be sensitive to the initial choice of these centroids.

3.9.1.2 Mean Shift

Mean Shift (Comaniciu & Meer, 2002) is a centroid based algorithm, similar to K-means, but it doesn't require a number of clusters as input. The algorithm assumes that the data is drawn from an underlying probability density function and tries to estimate it using Kernel Density Estimation. Then, it calculates a centroid for each data point using the kernel and iteratively updates the centroid to be the mean of the points within a given region. Each update is defined by a mean shift vector, which always points towards the direction of maximum increase in density. At convergence, the centroid will be placed at the nearest highest density peak of the density function.

The same process is repeated for each data point, and the data points which lead to the same high-density peaks are then assigned to the same cluster.

The method of calculating the mean depends on the choice of the kernel. Hence, the key parameter here is the bandwidth of the kernel used. Therefore, Mean Shift only requires one hyperparameter, although the final clustering results are very sensitive its value. For example, a low bandwidth would form very narrow kernels, which might lead to a large number of small clusters, while a very large bandwidth might lead to just one big cluster. This algorithm automatically sets the number of clusters, based on the bandwidth, and can also model complex shaped clusters. But it is not highly scalable as it requires multiple nearest neighbor searches, and due to random initialisation it might not produce reproducible results.

3.9.1.3 Affinity Propagation

Affinity Propagation (Frey & Dueck, 2007) is based on the concept of “message passing” between data points. It tries to find “exemplars”, i.e, members that are representative of clusters. It uses the similarity between all pairs of data points, and simultaneously considers all data points as potential exemplars. It starts by calculating a similarity matrix, which can be defined as the negative squared distance between two data points. The diagonal of this matrix is set to a constant, called “preference”, which is an input hyperparameter. Preference determines how likely a particular data point would be to become an exemplar. The algorithm then calculates three matrices, called Responsibility, Availability, and Criterion Matrix. This is the “message passing” step.

The “responsibility” quantifies how well suited a point k is to be the exemplar for point i , considering all other potential exemplars for i . The “availability” quantifies how “appropriate” it would be for i to pick k as exemplar, considering other points preference for k . The details of the algorithms are given in (Frey & Dueck, 2007). Both these matrices are iteratively updated, till a convergence criterion is met. The Criterion Matrix, which is the sum of Responsibility and Availability matrices, is then used to assign clusters.

Affinity Propagation’s advantage is that it doesn’t require the number of clusters as input, but the algorithm is computationally complex and can be slow on large datasets.

3.9.1.4 Agglomerative clustering

Agglomerative clustering (Franti et al., 2006) is a type of hierarchical clustering. Hierarchical clustering algorithms start with each example being a different cluster and then merge the ones that are closer until there is only one cluster. Therefore, they can form a hierarchy of clusters (at various distances), which is represented as a tree. A linkage criterion (see Section 5.1 of Jain et al., 1999) is used to decide the merge strategy. The merge strategy minimizes or maximises a certain distance criterion between pairs of unmerged clusters, to determine when to merge clusters.

To determine the clusters from this cluster hierarchy, one has to choose a level or a cut in the tree. This can be done by providing the number of clusters as input hyperparameter. Although it is computationally expensive, fast implementations of this algorithm exist (Mullner, 2013), which we haven’t explored in this chapter. As was the case with K-means, the main challenge with this algorithm is to choose the number of clusters. In an application where the number of clusters itself can vary, this algorithm might not generalize well.

3.9.1.5 DBSCAN

Density Based Spatial Clustering of Applications with Noise, or DBSCAN (Ester et al., 1996), is a density-based clustering algorithm. It assumes that clusters lie in dense regions. It primarily requires two input hyperparameters: a density threshold (*MinPts*) of a core point and a radius (ϵ) of its neighborhood. A point that has at least *MinPts* adjacent points in its ϵ neighborhood is considered a core point. Core points and their neighborhood are considered dense regions that form clusters, and overlapping dense regions are merged into a single cluster. Any point that is neither a core point nor falls within the neighborhood of a core point is classified as noise.

Therefore, DBSCAN doesn't require the number of clusters as input. The sizes and shapes of clusters can vary based on the choice of ϵ and the metric used to compute the distances between points. A potential disadvantage of DBSCAN is the difficulty to tune the input parameters, which the clustering performance is very sensitive to. Also as DBSCAN assumes a constant density, it might not perform well when there are clusters of varying densities in the data.

3.9.1.6 Optics

Ordering Points To Identify the Clustering Structure or Optics (Ankerst et al., 1999), is a density-based clustering algorithm. Similar to DBSCAN, Optics requires two hyperparameters: ϵ and *MinPts*, although ϵ is not necessary (and is set to a large number by default). It uses the following distances: core distance (minimum radius required to classify a given point as core point) and reachability distance (calculated by comparing the distance between two core points and their core distances) to order points. The reachability distance for points in a cluster would be low. The Optics algorithm builds a reachability graph, which assigns each sample a reachability distance. A post-processing procedure is applied to the reachability plot to determine clusters. This procedure can be very sensitive to the input parameters. In a reachability graph, different clusters would appear as valleys. So, by identifying these valleys, it is possible to identify clusters. This can be done by looking for steep slopes in this graph, or by extracting all clusters at a fixed value of distance (which would result in DBSCAN like clusters for a given ϵ). Therefore, this algorithm processes several different distance parameters at the same time (in contrast to DBSCAN, which works at a fixed ϵ), after which an appropriate ϵ can be chosen. An advantage of Optics is that it can find clusters of varying density. Like other density-based algorithms, Optics doesn't require the number of clusters as input and can also find non-spherical clusters.

3.9.1.7 HDBSCAN

Hierarchical Density Based Spatial Clustering of Applications with Noise or HDBSCAN (Campello et al., 2015; McInnes et al., 2017; McInnes & Healy, 2017), is very similar to Optics, i.e., it takes the approach of DBSCAN but extends it by varying the values of ϵ . It does so by converting DBSCAN into a hierarchical clustering algorithm by effectively running DBSCAN at every possible value of ϵ , and forms a hierarchical tree that shows the clustering output. By parsing through the tree, going from one large cluster to many smaller clusters, HDBSCAN constructs a tree with persistent clusters based on its only hyperparameter: minimum cluster size. It then uses a stability criterion to extract the final clusters from the cluster tree.

In contrast with DBSCAN, HDBSCAN therefore uses multiple values of ϵ (or cuts in the hierarchical tree) and is thus able to identify clusters of varying densities. This overcomes the main limitations of DBSCAN, to find the optimal value of ϵ and the inability to form clusters of varying densities. As noted, HDBSCAN takes in only a minimum cluster size hyperparameter, eliminating the need for ϵ .

3.9.1.8 Birch

Balanced Iterative Reducing and Clustering using Hierarchies or Birch (Zhang et al., 1996) is a hierarchical clustering algorithm used typically on very large datasets. It is local, which means that the clustering decision is made without scanning all data points and existing clusters. In most cases, it requires only two scans of the whole database. It uses a Clustering Feature (or CF) which consists of summary of statistics for a given sub-cluster. Summary statistics in CF are used to calculate the distance between two sub-clusters, the radius (average distance of members from the centroid) and the diameter (average pairwise distance within a cluster) of a CF. It therefore doesn't require all the data in the sub-cluster to be saved to perform clustering.

Using these CFs it creates a CF-Tree (CFT). Each node of a CFT consists of multiple CFs. The Birch algorithm has two hyperparameters: branching factor and threshold; the former limits the number of CFs (or sub-clusters) in a node of CFT,

while the latter limits the distance for a new sample to be a part of an existing CF. A node of the CFT consists of a CF for each branch. The terminal nodes of a CFT are then clustered using another clustering algorithm (Agglomerative Clustering in our case) to obtain the final clusters. Depending on the algorithm used, it might require other hyperparameters like number of clusters.

3.9.2 Parameter Ranges for Hyperparameter tuning

Table 3.3: Hyperparameter ranges explored for different clustering algorithms using random sampling. Random uniform sampling was used to sample hyperparameters for all the parameter ranges/values.

Algorithm	Hyperparameter	Range/Values
Affinity Propagation	affinity	euclidean
	random_state	1996
	damping	0.5, 1
	preference	-1000, -200
Agglomerative	n_clusters	2, 10
	affinity	euclidean, manhattan, cosine
	compute_full_tree	auto
	linkage	complete, average, single, ward ^a
Birch	n_clusters	2, 10
	threshold	0.1, 20
	branching_factor	10, 100
DBSCAN	min_samples	2, 10
	eps	0.5, 15
	metric	euclidean, chebyshev, cityblock, manhattan, canberra, hamming ^b
	algorithm	auto
	leaf_size	20, 40
HDBSCAN	min_samples	2, 5
	metric	euclidean, chebyshev, cityblock, manhattan, canberra, hamming
	min_cluster_size	2, 10
	cluster_selection_method	eom, leaf
	allow_single_cluster	True, False
KMeans	algorithm	auto, full, elkan
	n_clusters	2, 10
	n_init	10, 30
	random_state	1996
Mean Shift	bandwidth	10, 40 ^c
	bin_seeding	True, False
	cluster_all	True, False
Optics	min_samples	2, 10
	eps	0.5, 15
	metric	minkowski, euclidean, chebyshev, canberra, cityblock, manhattan, hamming ^b
	min_cluster_size	2, 10
	p	1, 15 ^d
	cluster_method	dbscan, xi ^e
	xi	0, 1

^a ward only works with euclidean affinity

^b eps range of 0.1 to 1 was used with hamming metric, and a range of 0.1 to 4 was used with canberra metric

^c bandwidth of 1 to 10 was used for normalised pre-processing case

^d eps range of 0.1 to 1 was used with hamming metric, and a range of 0.1 to 4 was used with canberra metric

^e Value of eps was used only with dbscan method, and value of xi was used when xi was selected as cluster method

Chapter 4

Comprehensive analysis of a dense sample of FRB 121102 bursts

4.1 Chapter Overview

In this chapter, we present an analysis of a densely repeating sample of bursts from the first repeating fast radio burst, FRB 121102. We reanalysed the data used by Gourdji et al. (2019) and detected 93 additional bursts using our single-pulse search pipeline. In total, we detected 133 bursts in three hours of data at a center frequency of 1.4 GHz using the Arecibo telescope, and develop robust modeling strategies to constrain the spectro-temporal properties of all the bursts in the sample. Most of the burst profiles show a scattering tail, and burst spectra are well modeled by a Gaussian with a median width of 230 MHz. We find a lack of emission below 1300 MHz, consistent with previous studies of FRB 121102. We also find that the peak of the log-normal distribution of wait times decreases from 207 s to 75 s using our larger sample of bursts, as compared to that of Gourdji et al. (2019). Our observations do not favor either Poissonian or Weibull distributions for the burst rate distribution. We searched for periodicity in the bursts using multiple techniques, but did not detect any significant period. The cumulative burst energy distribution exhibits a broken power-law shape, with the lower and higher-energy slopes of -0.4 ± 0.1 and -1.8 ± 0.2 , with the break at $(2.3 \pm 0.2) \times 10^{37}$ ergs. We provide our burst fitting routines as a

Published as Aggarwal et al. (2021a)

Contributing authors: Devansh Agarwal, Evan F. Lewis, Reshma Anna-Thomas, Jacob Cardinal Tremblay, Sarah Burke-Spolaor, Maura A. McLaughlin, Duncan R. Lorimer

python package BURSTFIT¹ that can be used to model the spectrogram of any complex FRB or pulsar pulse using robust fitting techniques. All the other analysis scripts and results are publicly available².

4.2 Introduction

As mentioned in Chapter 1, both repeating and non-repeating FRBs show a variety of spectro-temporal features and polarization properties: frequency modulation, sub-millisecond structure, drifting sub-pulses, varying polarization position angle, etc (Shannon et al., 2018; Farah et al., 2019; Luo et al., 2020b). Detection of multiple bursts from repeaters (like FRB 121102) have facilitated detailed studies of their properties and their environment. However, even after extensive follow-up and detection of hundreds of bursts from FRB 121102, the intrinsic emission mechanism remains uncertain, and many progenitor models have been proposed to explain the observational results. As theoretical models lack robust predictions for the observed properties of bursts, several empirical techniques have been employed to model the observed properties of the bursts. Some of those are: (1) using Weibull and Poisson distributions to model the clustering of repeater bursts; (2) using a truncated and broken power-law to model the flux-density distribution; (3) using 2D Gaussians to model the burst spectro-temporal properties; (4) using Gaussians convolved with an exponential tail to model scattering; (5) using a statistical spectral index to compare burst rates at multiple frequencies; (6) using signal-to-noise and structure to maximize DM, etc. (for further details, see Li et al., 2021; Cruces et al., 2020; Pastor-Marazuela et al., 2021; Gourdji et al., 2019; Hessels et al., 2019; Houben et al., 2019; Gajjar et al., 2018).

In view of all these considerations, it is necessary to detect and carefully investigate a large number of bursts from the repeaters to improve the understanding of their emission mechanism. In this chapter, we reanalyze the observations for FRB 121102 previously reported by Gourdji et al. (2019) and present the detection of an additional

¹<https://github.com/thepetabyteproject/burstfit>

²<https://github.com/thepetabyteproject/FRB121102>

93 bursts, for a full sample of 133 bursts in these observations. We detail a thorough burst modeling procedure and report extreme frequency modulation in burst spectra and a dearth of burst emission below 1300 MHz. We present and compare the updated burst energy and wait-time distributions and demonstrate how these estimates change dramatically for an incomplete search. We also perform exhaustive short-period periodicity tests to detect any possible rotational period as predicted by neutron star-based progenitor models. We also discuss various differences between our single-pulse search pipeline and the one used by Gourdji et al. (2019) to explain the extra bursts detected in our search (see Section 4.6.6).

This chapter is laid out as follows: In §2 we briefly discuss the data used in this chapter and discuss the search and spectro-temporal modeling procedure in §3. We then present our modeling results in §4, followed by a discussion of those results and conclusions in §5 and §6.

4.3 Data

The data reported here were originally collected, searched for FRBs, and reported by Gourdji et al. (2019). Here we provide only a brief summary of the data used in this analysis, and refer the reader to Gourdji et al. (2019) for further details. The observations were carried out with the 305-m William E. Gordon Telescope at the Arecibo Observatory with the L-Wide receiver and recorded using the Puerto Rican Ultimate Pulsar Processing Instrument (PUPPI). FRB 121102 was observed with 800 MHz bandwidth at a center frequency of 1375 MHz on MJDs 57644 and 57645 for 5967 s and 5545 s respectively. The data were coherently dedispersed at a dispersion measure (DM) value of 557 pc cm^{-3} during the observations and were recorded with 1.56-MHz channel bandwidth and $10.25\text{-}\mu\text{s}$ sampling resolution. The data used for this study were further decimated to 12.5-MHz channel bandwidth with 64 total channels and $81.92\text{-}\mu\text{s}$ sampling interval.

4.4 Methods

4.4.1 The Petabyte Project

Characterizing the diversity and event rates of FRBs as a function of observing frequency is critical for understanding their nature, the extreme emission physics responsible for FRB and pulsar emission, and the relationship between these two classes of objects. Many surveys have sought comprehensive estimates of these values, all using different observing frequencies, telescopes, and search algorithms but without characterizing the completeness of their search. The Petabyte Project³ (TPP) aims to address these issues to provide robust event rate estimates and discoveries in several petabytes of new and archival radio data. TPP will perform a uniform search for FRBs in an unprecedented amount of archival data and better probe transients closer, farther, and at higher radio frequencies than previous searches. Our search will have a robust internal assessment of completeness, allowing us to confidently project the frequency-dependent rates of FRBs and other transients.

TPP will use YOUR (a recursive acronym for “your unified reader”; Aggarwal et al., 2020) to ingest the data and HEIMDALL (Barsdell, 2012)⁴ to search it for single pulses. The deep learning-based classifier FETCH (Agarwal et al., 2020a)⁵ is then used to classify the candidates identified by HEIMDALL. The data will be searched up to a DM of 5000 pc cm⁻³ (or more, if possible) and a pulse width of 32 ms. This pipeline can easily be modified to search for higher DMs and pulse widths, in specific cases. All the candidates above a signal-to-noise ratio (S/N) of 6 classified as astrophysical by FETCH will be manually verified. The maximum DM and the pulse width to be searched is governed by the observing frequency, data resolution, and GPU memory and hence would be dealt with on a case-by-case basis. The data and results presented in this chapter were analyzed under TPP, using an early version of the TPP pipeline. In the following subsection, we discuss the details of the single-pulse search pipeline used in this analysis.

³<https://thepetabyteproject.github.io>

⁴<https://sourceforge.net/projects/heimdall-astro>

⁵<https://github.com/devanshkv/fetch>

4.4.2 Single-Pulse Search

Within YOUR, we used `your_heimdall.py` which runs HEIMDALL on the PSR-FITS data files collected by PUPPI for the single-pulse search. We used two search strategies: (1) DM range between 450–650 pc cm⁻³ with a DM tolerance⁶ of 1% and (2) DM range between 10–5000 pc cm⁻³ with a tolerance of 25%, with a maximum pulse width of 84 ms in both cases (note that the widest FRB 121102 pulse reported at this frequency had a width of ~ 35 ms, see Cruces et al., 2020). The searches resulted in 1,428 and 11,276 candidates, respectively. For each candidate, we extracted a segment of the data (which we hereafter refer to as a “cutout”) centered at the arrival time (referenced to the top of the observing band) as reported by HEIMDALL with a time window equal to twice the dispersion delay using `your_candmaker.py`. We then used spectral kurtosis RFI mitigation with a 3σ threshold (Nita & Gary, 2010) to identify and excise frequency channels corrupted by RFI (see Figure 4.2 for fraction of bursts for which a frequency channel was flagged due to RFI) and used this cleaned data to create dedispersed frequency-time images where a factor of width/2 was used to decimate the time axis. We created the DM–time image by dedispersing the data from zero to twice the reported DM and simultaneously decimating the time axis as above (for more details on the candidate pre-processing, see Agarwal et al., 2020a). These cutouts are then used by FETCH to label FRBs and RFI, and were also manually verified. In total, we found 133 bursts with DMs consistent with that of FRB 121102 (i.e., between 550–580 pc cm⁻³) with 93 new bursts as compared to the previously published results. We highlight some important differences between our single-pulse search pipeline and the one used by Gourdji et al. (2019) in Section 4.6.6 to explain the new burst detections. We did not detect any bursts at other DMs. Our search missed one burst, B33, reported by Gourdji et al. (2019), probably because it was weak and narrowband (see Section 4.6.7 for caveats of our search). Figure 4.1 shows the dynamic spectra of some of the bursts. Candidate cutouts for all the bursts

⁶DM tolerance is the acceptable sensitivity loss between DM trials for a single-pulse search (for further details, see Aggarwal et al., 2021c; Levin, 2012)

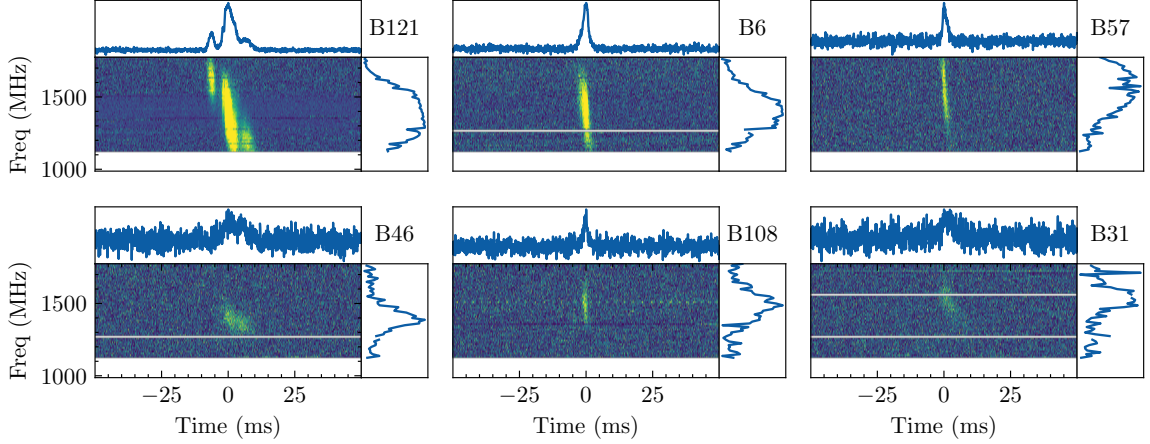


Figure 4.1: Dynamic spectra of six bursts, dedispersed at $DM = 560.5 \text{ pc cm}^{-3}$. For each burst, the top panel shows the burst profile obtained by averaging along the frequency axis, and the right panel shows the burst spectra obtained by averaging the burst data along the time axis. The white horizontal lines show the channels masked due to RFI. The color scale of each spectrogram has been set from mean to 3 times the standard deviation of the off-pulse region. The ranges of 1D plots are different for individual plots. Burst numbers are mentioned on the top right of each subplot.

are available on Github⁷.

4.4.3 Completeness Limit

We define the completeness limit as the pulse energy (also known as fluence) value above which any burst emitted during the observation would be detected. Determining the completeness limit of any single-pulse search is, therefore, essential to defining the sample of bursts to be used for statistical analyses. The most robust method of determining the completeness limit involves an exhaustive injection analysis. In such analyses, simulated transients (with varying properties) are injected on background data, and by analyzing the transients that were recovered (or missed), one can determine the completeness limit of a search (Li et al., 2021; Gupta et al., 2021; The CHIME/FRB Collaboration et al., 2021a; Agarwal et al., 2020b; Farah et al., 2019). Such an analysis requires access to a large amount of native-resolution data observed with the same telescope and observing configuration as the search data.

⁷<https://github.com/thepetabyteproject/FRB121102>

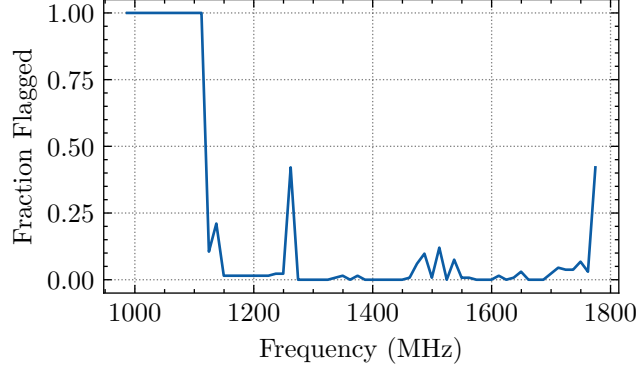


Figure 4.2: The fraction of bursts for which a frequency channel was flagged due to RFI.

As we had access to decimated data for just two observations, we could not do such an injection analysis. We, therefore, estimate the completeness limit from the radiometer equation. We use a conservative approach by including the effect of RFI mitigation, which reduces search sensitivity. We flagged more than 35% of data for many candidates due to RFI, leading to a usable bandwidth of 500 MHz. Using this smaller bandwidth and nominal pulse width of 1 ms, the fluence limit above an S/N of 8 is 0.0216 Jy ms. We use bursts with fitted fluence above this limit in the burst sample analysis.

4.4.4 Spectro-temporal Burst Modelling

To measure the properties of the bursts in our sample, we perform spectro-temporal modeling of all the detections. We model each component of the bursts' spectra with a Gaussian function and the profile using a Gaussian convolved with a one-sided exponential function to represent the Gaussian pulse and the scattering tail. Therefore, we model each component of the burst (at an observing frequency f and time t), using the following function,

$$\begin{aligned}
 \mathcal{F}(f, t; S, \mu_f, \sigma_f, \mu_{\text{DM}}, \sigma_t, \tau_{\text{sc}}) = & S \times 2.355\sigma_f \\
 & \times \mathcal{P}_f(t; \mu_{\text{DM}}, \sigma_t, \tau_{\text{sc}}) \\
 & \times \mathcal{G}(f; \mu_f, \sigma_f).
 \end{aligned} \tag{4.1}$$

Here, S is the fluence of the component, \mathcal{G} is the Gaussian function to model the spectra

$$\mathcal{G}(x; \mu_x, \sigma_x) = \frac{1}{\sigma_x \sqrt{2\pi}} \exp\left(-\frac{1}{2} \frac{(x - \mu_x)^2}{\sigma_x^2}\right), \quad (4.2)$$

and \mathcal{P} is used to model the profile. For bursts with scattering, we use the Gaussian convolved with a one-sided exponential function (McKinnon, 2014),

$$\begin{aligned} \mathcal{P}_f^{\text{scat}}(t; \mu_{\text{DM}}, \sigma_t, \tau_{\text{sc}}) = & \frac{1}{2\tau_{\text{sc}}} \left\{ 1 + \operatorname{erf} \left[\frac{t - (\mu_{\text{DM}} + \sigma_t^2/\tau_{\text{sc}})}{\sigma_t \sqrt{2}} \right] \right\} \\ & \times \exp\left(\frac{\sigma_t^2}{2\tau_{\text{sc}}^2}\right) \exp\left(\frac{t - \mu_{\text{DM}}}{\tau_{\text{sc}}}\right), \end{aligned} \quad (4.3)$$

where the mean of the Gaussian pulse after accounting for the dispersion delay at that respective frequency,

$$\mu_{\text{DM}} = \mu_t - 4.148808 \times 10^3 \text{ DM} \left(\frac{1}{f^2} - \frac{1}{f_{\text{top}}^2} \right) \text{ s}. \quad (4.4)$$

Here f_{top} is the highest frequency in the band (in MHz), f is the frequency of a channel (in MHz) and μ_t is the mean of the Gaussian pulse at f_{top} . At each frequency channel, the scattering time scale

$$\tau_{\text{sc}}(f) = \tau_{\text{sc}} \left(\frac{f}{f_{\text{ref}}} \right)^{-4}. \quad (4.5)$$

In this expression, f_{ref} is the reference frequency and is set to be 1 GHz. The exponent (-4) is assuming a normal distribution of plasma-density inhomogeneities. Finally, we defined $\mathcal{P}_f(t; \mu_{\text{DM}}, \sigma_t, \tau_{\text{sc}})$ to be a Gaussian $\mathcal{G}(t; \mu_{\text{DM}}, \sigma_t)$ for $\tau_{\text{sc}}/\sigma_t < 6$, and $\mathcal{P}_f^{\text{scat}}(t; \mu_{\text{DM}}, \sigma_t, \tau_{\text{sc}})$ for $\tau_{\text{sc}}/\sigma_t > 6$. We used this value for the cutoff in order to maintain numerical stability while calculating Equation 4.3.

It follows from the above discussion that our model \mathcal{F} is generated using seven parameters: $S, \mu_f, \sigma_f, \mu_t, \sigma_t, \tau_{\text{sc}}$ and DM. Using this model we fit the burst spectrograms, as described in the next section.

4.4.5 BURSTFIT

While fitting for complex FRB bursts is an arduous task, we scrupulously automate the entire procedure and present it as python package BURSTFIT⁸. BURSTFIT provides a framework to model any spectrogram consisting of any complex FRB or pulsar pulse using robust methods. It can easily incorporate any user-defined python function(s) to model the profile, spectra, and spectrogram and is not limited to the functions we have implemented for this current analysis. BURSTFIT primarily consists of the following five steps.

4.4.5.1 Data Preparation

First, we dedispersed the burst spectrogram at the DM obtained from the single-pulse search. This DM is usually accurate enough to correct for most of the dispersion. We cut out a time window of 200 ms encompassing the burst from this dedispersed spectrogram and normalized this data to zero mean and unit standard deviation using the off-pulse region. Both the cutout and normalized data were then used for fitting. We also masked all the channels flagged as RFI during the search and candidate pre-processing so as not to influence the fitting procedure.

4.4.5.2 Stage 1: Single-component Fitting

In this first stage of fitting, we used `scipy.curve_fit`⁹ to perform the fits and got an initial estimate of the fitted parameters. We created the time-series profile by summing along the frequency axis, and modelled it using $S \times \mathcal{P}_f(t; \mu_t, \sigma_t, \tau_{sc})$. We then used the fitted values of μ_t and σ_t to identify the time samples with the burst signal and average them to produce the burst spectra. We normalised this spectra to unit area, and modelled it using $\mathcal{G}(f; \mu_f, \sigma_f)$. Following this, we modelled the complete spectrogram. We first generated the model spectrogram by stacking N_f model pulse profiles ($S \times \mathcal{P}_f(t; \mu_{DM}, \sigma_t, \tau_{sc})$) together, where N_f is the number of

⁸<https://github.com/thepetabyteproject/burstfit>

⁹This routine is a part of the python-based `scipy` package. We used version 1.5.2 of this package in our analysis.

frequency channels. Note that the mean of each profile was already corrected for the dispersion delay at that respective frequency. This gave us a scattered and dispersed spectrogram at a given set of profile parameters and a DM.

We then multiplied this spectrogram with the model spectra to obtain the model spectrogram. Additionally, we clipped the model spectrogram at the estimated saturation level (see Section 4.4.5.6) and masked the RFI channels. Using this model we fit for all the seven parameters in $\mathcal{F}(f, t; S, \mu_f, \sigma_f, \mu_t, \sigma_t, \tau_{sc})$ along with DM by comparing the model with the dedispersed cutout spectrogram obtained in the previous step. Again, we used `scipy.curve_fit` for fitting, and used the estimates from individual profile and spectra fits as initial guesses for the parameters.

4.4.5.3 Stage 2: Statistical Tests

Following Stage 1, we obtained the residual spectrogram by subtracting the fitted model from the original spectrogram. Then we performed several statistical tests (see, e.g., Kramer et al., 1994) to compare the properties of on-pulse residual with respect to the off-pulse regions in the original spectrogram. We performed the following three comparisons: left off-pulse vs. right off-pulse, on-pulse vs. left off-pulse, and on-pulse vs. right off-pulse. We used the following four tests (all implemented within `scipy`): Kolmogorov-Smirnov test (for distribution comparison), F-test (for variance comparison), T-test (for mean comparison), and Kruskal test (for median comparison).

We considered the two distributions similar if at least two of the four tests had a p -value above 0.05 (i.e. we did not have significant support for the non-similarity of the distributions). Comparing left off-pulse with right off-pulse region gave us confidence in our choice of off-pulse region. We terminated the single component fitting procedure if either of the off-pulse vs. on-pulse comparisons demonstrates that the distributions are similar. If the distributions were different, we used the residual spectrogram and repeated Stage 1 and 2 to fit another component and compared the on-pulse residual with off-pulse data. We kept fitting for components until the statistical tests pass or until a maximum of five components is reached.

4.4.5.4 Stage 3: Multi-component Fitting

In cases where multiple components were found, we performed another stage of the fitting. Here, we generated a combined model consisting of all the components and fit for all components by comparing our model with the original spectrogram. This combined model (\mathcal{F}_{all}) was generated by summing together the individual component models (\mathcal{F}_i) for all N components,

$$\mathcal{F}_{\text{all}} = \sum_i^N \mathcal{F}_i. \quad (4.6)$$

Again, we used `scipy.curve_fit` for fitting. Here, the fit results of the individual component fits from previous stages were used as the initial guess for the parameters in `curve_fit`.

4.4.5.5 Stage 4: MCMC

While `scipy.curve_fit` is sufficient for fitting in many scenarios, in our testing we found that the estimates and the errors reported by `scipy.curve_fit` were not robust for our purposes. In many cases, the errors reported by `scipy.curve_fit` were possibly underestimated, and fitted results were highly susceptible to the choice of input parameter bounds. This was especially true for low-significance bursts and multiple-component bursts, where the least-squares-minimization technique struggles to find a good solution. Therefore, we added another stage to our fitting procedure and used a Markov Chain Monte Carlo (MCMC) to obtain the final fitting results. We used the results of previous stages (that used `scipy.curve_fit`) as initial estimates to determine the starting positions of the walkers for the MCMC. An advantage of the MCMC procedure is that it provides the full posterior distribution of all the fitted parameters, which we could then use to estimate the errors and further follow-up analysis of the burst sample. We used the Goodman and Weare affine invariant sampler (Goodman & Weare, 2010) as implemented in EMCEE (Foreman-Mackey et al., 2013). We used uniform priors for all the parameters, with the ranges of the

priors given in Table 4.1. We used the log-likelihood function

$$\ln \mathcal{L} = -0.5 \sum \left(\frac{\mathcal{S} - \mathcal{F}_{\text{all}}}{\sigma} \right)^2, \quad (4.7)$$

where \mathcal{S} refers to the original spectrogram, \mathcal{F}_{all} refers to the model, and σ is the off-pulse standard deviation of the measured spectrogram. The sum is over all the pixels in the two spectrograms. We used autocorrelation analysis to determine when the MCMC has converged¹⁰. We then estimated the burn-in¹¹ using the autocorrelation time and used the remaining samples to determine the fitting results. To decide if scattering was present in a burst, we used the percentage of samples with $\tau_{\text{sc}}/\sigma_t < 6$. If this percentage was greater than 50%, we concluded that scattering was not present (or was very small) in that burst. We do not report scattering timescales for such bursts.

We generated corner plots and fit-result plots to verify the quality of the fits, as shown in Figures 4.5 and 4.6. We provide all the results (output parameters, corner plots, fitting-result plots, etc.) from our analysis in a Github repository.¹²

4.4.5.6 Handling data saturation

The data we use in this analysis were recorded as 8-bit unsigned integers. Hence, the data range lies between 0–255, and any signal brighter than 255 is clipped at this value. We noticed data saturation for two bursts (B6 and B121), and hence this effect has been incorporated in our burst modeling. The spectrograms are subtracted by the off-pulse mean (μ_{off}) and divided by the off-pulse standard deviation (σ_{off}). While making the spectro-temporal model, we clip the values greater than $(255 - \mu_{\text{off}})/\sigma_{\text{off}}$. This effect is visible in Figure 4.3 for burst B121 where the red dot-dashed curve and green dotted curve show the fit to the burst spectra with and without clipping,

¹⁰See <https://emcee.readthedocs.io/en/stable/tutorials/autocorr> for details.

¹¹To avoid the phenomenon known as “burn-in”, where there is a high degree of correlation between neighboring samples in each MCMC chain, the initial values are typically discarded. This is especially important if the MCMC was initialized at a low probability region in the parameter space. Therefore, if the initial samples are not discarded, then those might bias the posterior distributions of MCMC samples. See section 7 of Hogg & Foreman-Mackey (2018) for more details.

¹²<https://github.com/thepetabyteproject/FRB121102>

Table 4.1: Priors used in the MCMC fitting.

Parameter	Minimum	Maximum
S	0	$500 \times \max(\text{time_series}) \times \sigma_t^{\text{fit}}$
μ_f	$-2 \times N_f$	$3 \times N_f$
σ_f	0	$5 \times N_f$
μ_t	$0.8 \times \mu_t^{\text{fit}}$	$1.2 \times \mu_t^{\text{fit}}$
σ_t	0	$1.2 \times (\sigma_t^{\text{fit}} + \tau_{\text{sc}}^{\text{fit}})$
τ_{sc}	0	$1.2 \times (\sigma_t^{\text{fit}} + \tau_{\text{sc}}^{\text{fit}})$
DM	$0.8 \times \text{DM}^{\text{fit}}$	$1.2 \times \text{DM}^{\text{fit}}$

Superscript fit refers to the values obtained using fits done in previous stages. N_f refers to the number of frequency channels. `time_series` refers to the 1-D array obtained by summing the dedispersed cutout spectrogram along the frequency axis. Subscripts t and f are used for profile and spectra parameters.

respectively. The fit performed without considering the saturation underestimates the burst’s spectral width, leading to an underestimated burst energy.

4.4.5.7 Caveats to our fitting analysis

There are some caveats to our fitting procedure that are worth noting here. First, as with any model-dependent fitting, our analysis and results are dependent on the choice of the functions we use to model the data. We described those functions and our motivation for using them in Section 4.4.4, but these are not the only proposed methods to model the spectrogram of an FRB. The spectrum of an FRB can also have power-law-like behavior (CHIME/FRB Collaboration et al., 2020b), and some other pulse broadening functions can also be used besides the exponential tail used to model the scattering effect (Bhat et al., 2004).

Second, the emission of some bursts in our sample was present only in the top part of the band. In many such cases, the emission appeared similar to the tail of a Gaussian function, with the mean lying outside our observing band. Although we allowed for such a mean value to be estimated with the MCMC procedure, our fitting results for such bursts would inevitably be unconstrained. We therefore mark such bursts with a † in Table 4.2 to highlight this.

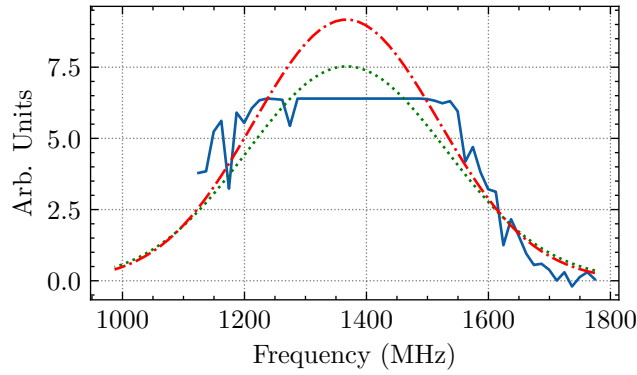


Figure 4.3: Plots of spectra and fits in case of data saturation. Blue solid line shows the spectrum of burst B121. The spectrum shows saturation between 1250 MHz and 1550 MHz, and any real structure in the spectrum is lost between those frequencies. The red dot-dashed line shows the model spectrum obtained when the fitting procedure incorporates the effect of saturation, while the dotted green line shows the spectrum obtained without considering saturation. The red curve better estimates the shape of the spectrum (assuming that the spectrum can be modeled using a Gaussian function), while the green curve underestimates the fluence and frequency width. See Section 4.4.5.6 for more details.

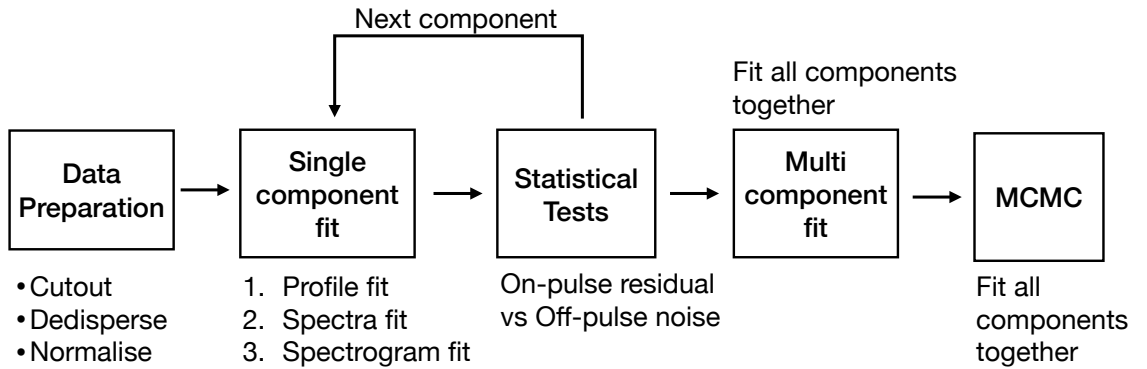


Figure 4.4: Flowchart showing the various stages of fitting in BURSTFIT. See Section 4.4.5 for details.

Table 4.2: Properties of the first 10 bursts. See Appendix for the full table.

Burst ^a ID	μ_f (MHz)	σ_f (MHz)	S (Jy ms)	μ_t^b (MJD)	σ_t (ms)	τ^c (ms)	DM (pc cm ⁻³)
B1*	1560 ⁺³⁰ ₋₃₀	210 ⁺⁴⁰ ₋₄₀	0.09 ^{+0.02} _{-0.02}	57644.408906976(1)	0.0 ^{+0.02} _{-0.02}	1.9 ^{+0.3} _{-0.3}	565.3 ^{+0.4} _{-0.4}
B2*	1200 ⁺¹⁰ ₋₁₀	50 ⁺¹⁰ ₋₁₀	0.043 ^{+0.007} _{-0.007}	57644.40956768(1)	1.35 ^{+0.05} _{-0.05}	-	562.4 ^{+0.8} _{-0.8}
B3.1†	2900 ⁺³⁰⁰ ₋₆₀₀	800 ⁺³⁰⁰ ₋₂₀₀	0.6 ^{+0.7} _{-0.3}	57644.409673699(3)	0.4 ^{+0.2} _{-0.2}	1.3 ^{+0.7} _{-0.7}	566.8 ^{+0.8} _{-0.9}
B3.2†	1100 ⁺³⁰⁰ ₋₁₄₀₀	1000 ⁺²⁰⁰⁰ ₋₁₀₀₀	0.09 ^{+0.13} _{-0.04}	57644.40967384(2)	0.3 ^{+1.4} _{-0.2}	0.3 ^{+0.7} _{-0.1}	564.7 ^{+1.5} _{-0.4}
B4†	3100 ⁺²⁰⁰ ₋₆₀₀	550 ⁺⁸⁰ ₋₁₁₀	2 ⁺⁴ ₋₂	57644.410072889(4)	1.1 ^{+0.2} _{-0.2}	0.3 ^{+0.2} _{-0.1}	564 ⁺¹ ₋₁
B5†	2100 ⁺⁹⁰⁰ ₋₁₆₀₀	2700 ⁺⁹⁰⁰ ₋₁₂₀₀	0.19 ^{+0.05} _{-0.04}	57644.410157834(4)	0.7 ^{+0.3} _{-0.3}	1.0 ^{+0.4} _{-0.3}	562.1 ^{+0.7} _{-0.6}
B6.1	1393 ⁺⁷ ₋₇	183 ⁺⁷ ₋₇	0.47 ^{+0.02} _{-0.03}	57644.411071954(1)	1.09 ^{+0.03} _{-0.04}	-	562.3 ^{+0.2} _{-0.1}
B6.2	1417 ⁺⁴ ₋₅	102 ⁺⁵ ₋₄	0.33 ^{+0.03} _{-0.02}	57644.4110719755(9)	0.57 ^{+0.03} _{-0.02}	-	560.9 ^{+0.2} _{-0.1}
B7.1†	3100 ⁺²⁰⁰ ₋₃₀₀	430 ⁺⁶⁰ ₋₈₀	10 ⁺²⁰ ₋₁₀	57644.412240214(5)	0.7 ^{+0.3} _{-0.4}	0.8 ^{+0.4} _{-0.4}	569 ⁺³ ₋₃
B7.2†	1460 ⁺²⁰ ₋₂₀	90 ⁺³⁰ ₋₂₀	0.09 ^{+0.01} _{-0.01}	57644.41224043(2)	1.9 ^{+0.9} _{-1.0}	1.2 ^{+0.6} _{-0.5}	569 ⁺³ ₋₂
B8†	3000 ⁺²⁰⁰ ₋₆₀₀	700 ⁺¹⁰⁰ ₋₁₀₀	1.1 ^{+1.5} _{-0.7}	57644.414123628(4)	1.0 ^{+0.3} _{-0.3}	0.8 ^{+0.4} _{-0.3}	567.5 ^{+0.9} _{-0.7}
B9*	1430 ⁺¹⁰ ₋₁₀	75 ⁺⁹ ₋₉	0.076 ^{+0.003} _{-0.003}	57644.41447161(2)	2.0 ^{+0.6} _{-0.6}	0.4 ^{+0.6} _{-0.6}	564 ⁺² ₋₂
B10	1630 ⁺¹⁰ ₋₁₀	82 ⁺⁸ ₋₈	0.1 ^{+0.01} _{-0.01}	57644.414475391(7)	1.4 ^{+0.3} _{-0.3}	0.5 ^{+0.3} _{-0.2}	562 ⁺³ ₋₃

1 σ errors on the fits are shown on superscript and subscript of each value in the table. For μ_t , the error on the last significant digit is shown in parenthesis.

^a Burst IDs are chronological. Individual component number (N) for multi-component bursts are appended to the burst IDs. Bursts modeled only using `curve_fit` are marked with *. Note that the errors on these bursts could be unreliable and may be either under or over-estimated. Bursts that extend beyond the observable bandwidth can also have unreliable estimates of spectra parameters and fluence (see Section 4.4.5.7). We mark those bursts with † to indicate that their fluence and spectra parameters could be unconstrained.

^b μ_t is the mean of the pulse profile in units of MJD. This can be considered as the arrival time of the pulse. It is referenced to the solar system barycenter, after correcting to infinite frequency using a DM of 560.5 pc cm⁻³.

^c τ is referred to 1 GHz.

Third, in some cases, the MCMC procedure was unable to find a robust solution. This was due to the presence of RFI, which could not be removed as the FRB signal coincided with the channels heavily corrupted by RFI. It also occurred when there was significant baseline variation in the data close to a weak FRB pulse. This could dominate the MCMC likelihood estimate, and therefore the procedure could not converge on a solution. For such cases, we only used `scipy.curve_fit` to perform the fits, and we modified the fitting bounds to obtain a visually good fit. We highlight these in Table 4.2 with *, and note that the values could be unreliable.

4.5 Results

As mentioned previously, we tripled the number of published bursts from these two observations to 133, by detecting 93 new bursts. Dynamic spectra for some of

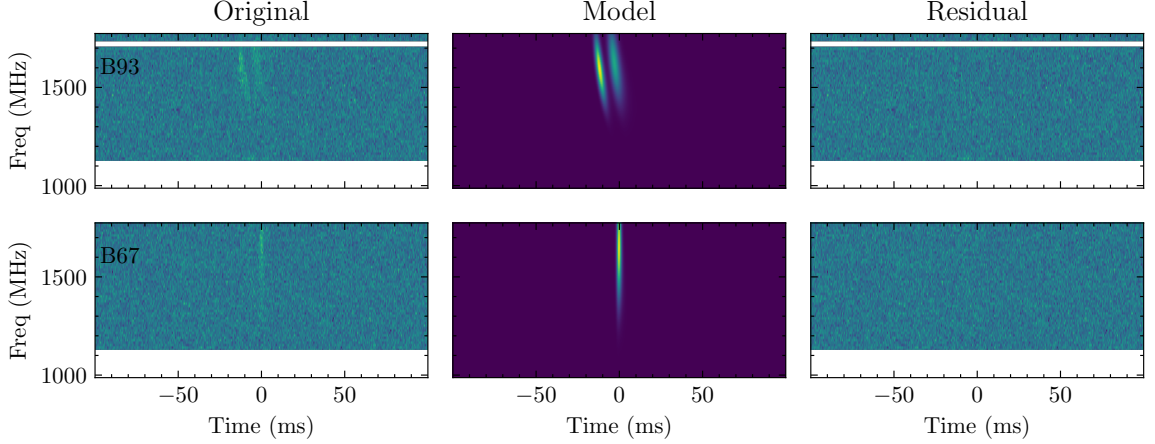


Figure 4.5: Results of spectro-temporal fits on two bursts (B93 and B67). The first column shows the original (normalized) dynamic spectra of the bursts. The burst can be seen in both cases towards the top of the band. B93 shows two components, separated by around 10 ms. The middle column shows the noise-free model spectrograms that best fit the original data. The last column shows the residual spectrogram obtained by subtracting the model from the original data. The residual spectrogram in both cases shows noise-like data with no remaining artifacts. See Section 4.4.5 for details of the fitting procedure.

the high-significance bursts are shown in Figure 4.1.

We used the burst modeling procedure described in the previous sections to estimate spectral and temporal properties for all the bursts. Figure 4.5 shows the fitting results for two bursts. Comparing the three columns, we can see that the modeled bursts look similar to the original burst signal and the residuals are noise-like, indicating that the models assumed for the burst spectrogram provide satisfactory fits. Figure 4.6 shows the posterior distribution of the burst properties obtained using MCMC for B67. The 1D plots show the parameter histogram, while the 2D plots show correlations between parameters. The properties of the bursts are given in Table 4.2.

4.5.1 Burst sample properties

We used the converged sample chains from the MCMC fitting for each burst to generate a cumulative corner plot with the whole burst sample properties. To do this, we randomly selected 1000 samples from the final 25% of the MCMC chains and

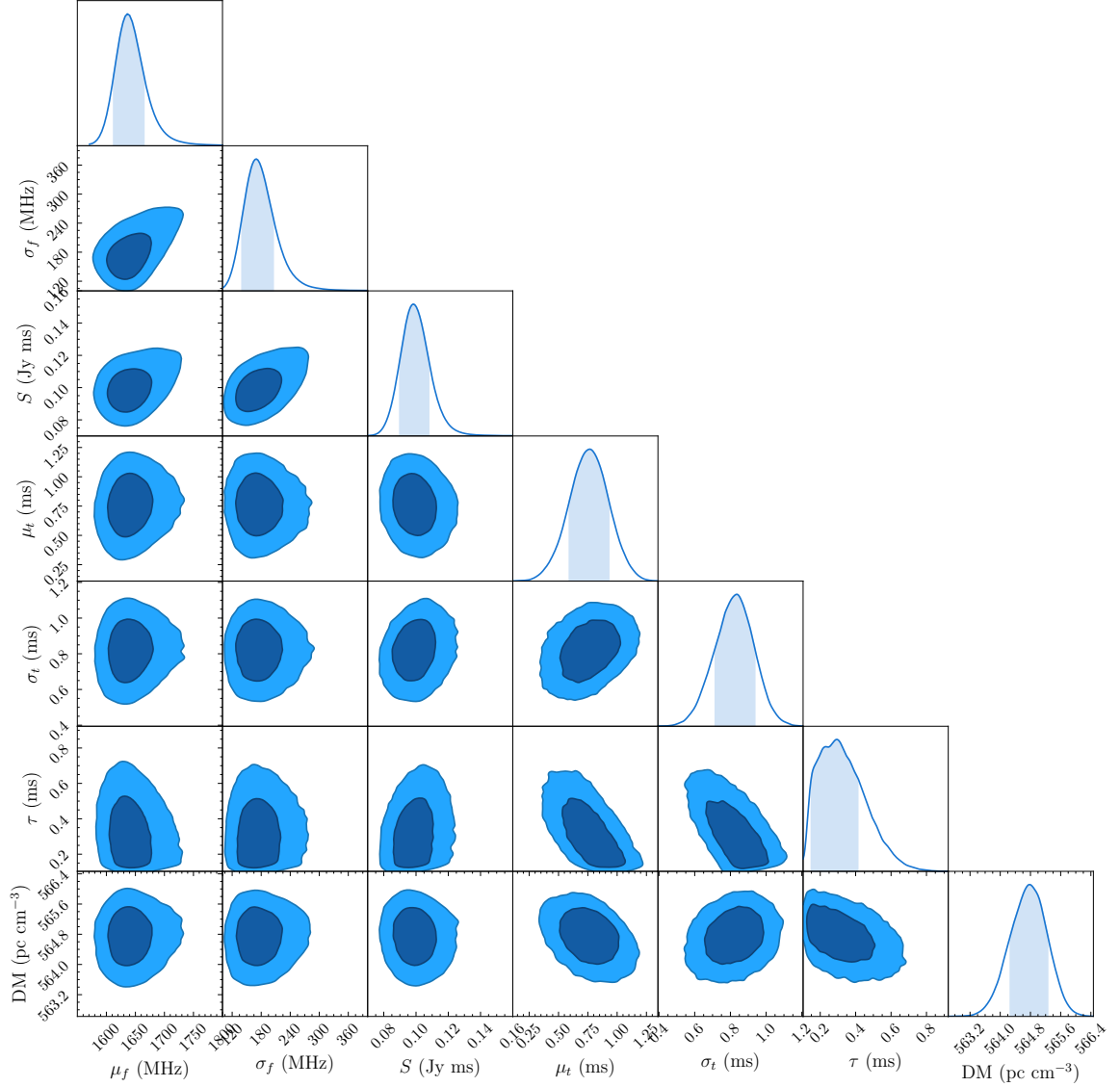


Figure 4.6: Corner plot generated using the MCMC samples obtained after fitting B67. The 1D histograms show the histogram of individual parameters, while the 2D plots show the correlations between parameters. The shaded and darkened region in 1D and 2D plots correspond to 1σ . μ_t is in units of milliseconds and is with respect to MJD=57645.41732241795. The value of μ_t is referenced to the observatory site (Arecibo Observatory) and at the top of the frequency band i.e. 1774.53125 MHz. See Sections 4.4.5 and 4.5 for more details.

then concatenated such samples from all the bursts. We then generated a corner plot using these samples, as shown in Figure 4.7¹³. We can now use Figure 4.7 to infer trends in various spectro-temporal properties of FRB 121102. Table 4.3 shows the summary statistics of all the bursts obtained from this analysis. We can see that the spectra of the bursts typically peak around 1650 MHz, and there is a dearth of burst emission below 1300 MHz. Most of the burst spectra peak within the top part of our observing band (i.e 1550–1780 MHz). This behavior possibly extends further to higher frequencies, as is evident from many burst spectra that increase towards the top part of the observing band with their spectral peak possibly lying outside our observing band. Interestingly, Platts et al. (2021) have recently reported complex bifurcating structures in some FRB 121102 bursts below 1250 MHz using higher-resolution data. It is, therefore, possible that the emission of FRB 121102 shows a different behavior below these frequencies, which might also vary with time.

As already noted by Gourdji et al. (2019), FRB 121102 shows a variety of spectral widths. Using our modeling, we observe that most bursts are frequency-modulated and have a typical frequency width of ~ 230 MHz. The bursts also show a wide range of intrinsic pulse widths, from 0.4 ms to 20 ms, and various scattering timescales, up to 3 ms. The median dispersion measure of the bursts we observe was 564 pc cm^{-3} , with a 1σ variation of $\sim 4 \text{ pc cm}^{-3}$. This variation in DM is also apparent in Figures 4.7 and 4.8. This value is consistent with the other published estimates (Li et al., 2021; Platts et al., 2021; Cruces et al., 2020; Gourdji et al., 2019). We also did not see any strong correlation between any two burst properties from Figure 4.7. Several of the bursts from this sample also show the characteristic sub-burst drift in frequency during the burst duration, sometimes referred to as the “sad-trombone” effect (Hessels et al., 2019). We did not detect any evidence of upward drifting as predicted by some FRB models (Cordes et al., 2017), and reported by Platts et al. (2021).

Many bursts in our sample show multiple components, and we estimated the

¹³Note that this corner plot is different from the one in Figure 4.6. Figure 4.6 shows the samples from MCMC fit on only one burst, while Figure 4.7 shows the cumulative samples obtained from MCMC fit on all the bursts.

Table 4.3: Results from the burst sample analysis. The values represent the median values with 1σ errors.

Parameter	Units	Value
μ_f	MHz	1608^{+100}_{-200}
σ_f	MHz	102^{+130}_{-30}
S	Jy ms	$0.13^{+0.13}_{-0.05}$
σ_t	ms	$1.1^{+0.9}_{-0.5}$
τ	ms	$0.7^{+0.9}_{-0.4}$
DM	pc cm ⁻³	564^{+5}_{-3}

properties of these components using our fitting procedure. Nine bursts in our sample show two components, while there is one burst with three components (see Table 4.2). We also note that it is difficult to differentiate between multiple closely spaced bursts and different components from single bursts. This is further complicated by the detection of a very wide (~ 35 ms) burst reported by Cruces et al. (2020). As there is no clear consensus on how to resolve this, we visually identified some bursts as components of a nearby burst and reported them as such in Table 4.2. We consider all the components as individual bursts for all the following analysis except the cumulative energy distribution analysis.

Figure 4.8 shows the scatter plots of various burst properties with respect to the burst time. The bursts from two observations are shown in different colors, and the time is referenced to the first burst of the respective observation. The burst properties do not show any temporal evolution at the seconds-to-minutes time scale. We also did not observe any distinction between the distribution of properties of bursts detected on two consecutive days.

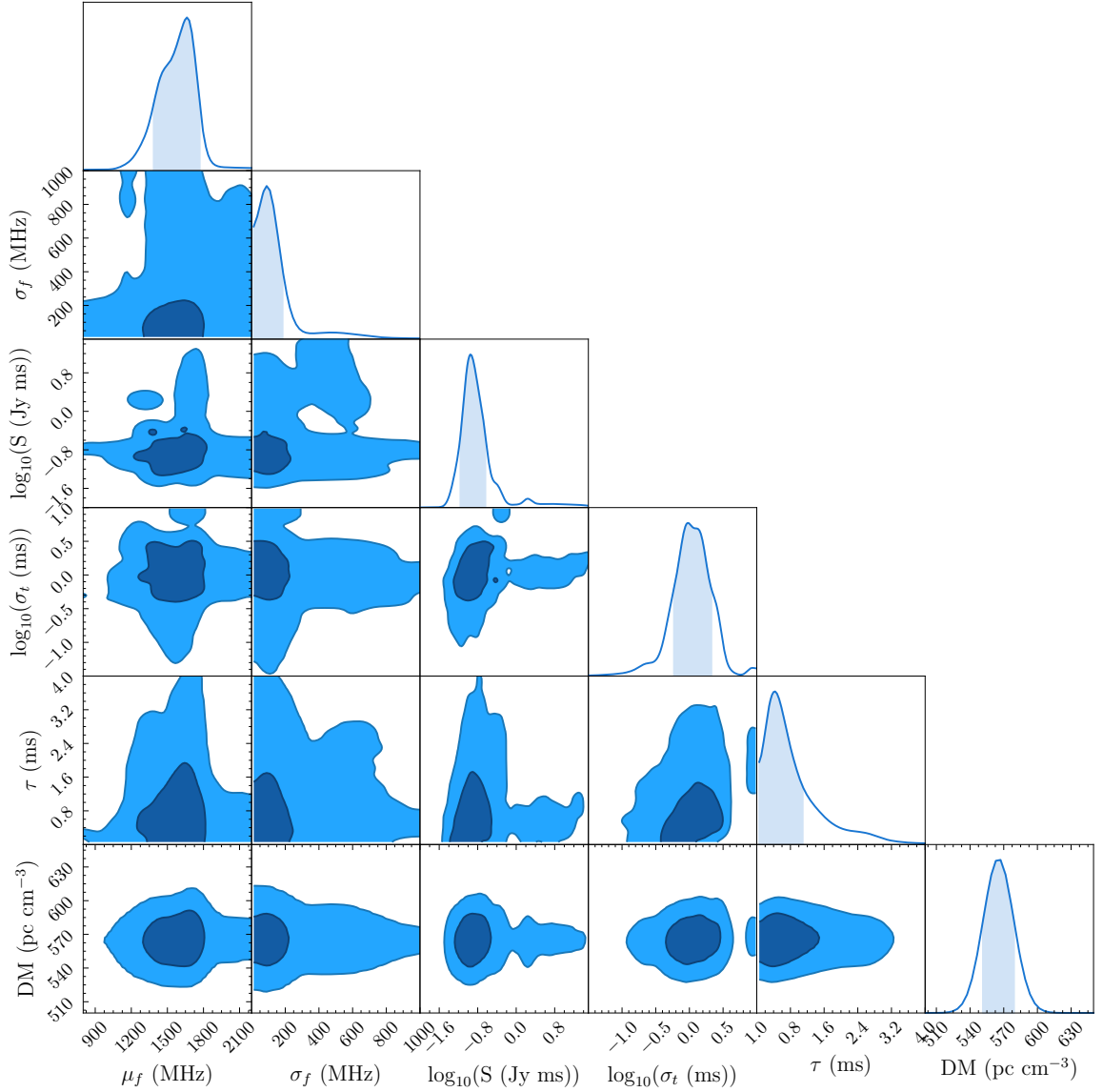


Figure 4.7: Corner plot generated using samples from the MCMC fit for all the bursts. The 1D histograms show the distribution of individual parameters for all the FRB 121102 bursts, while the 2D plots show the correlations for different parameters. The shaded and darker regions in 1D and 2D plots correspond to 1σ . See Section 4.5 for more details.

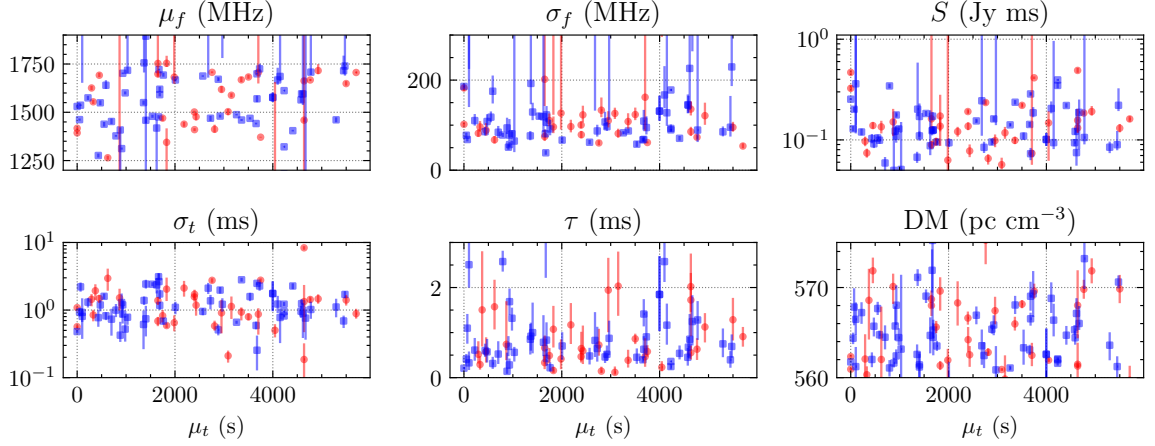


Figure 4.8: Scatter plots of fitted property of the bursts versus the fitted time of the burst. The data from the two observations are shown in different colors, with MJD 57644 in red and MJD 57645 in blue. μ_t is referred to the first detected burst in the observation. Only bursts fit using MCMC are shown here. The error bars represent 1σ errors on the fit. Errors on μ_t are very small, and hence are not visible. None of the six properties show any clear trend with respect to the burst time.

4.6 Discussion

4.6.1 Cumulative energy distribution

Energy distributions can provide useful intuition into the emission mechanism of the source. Regular pulsar emission typically shows a log-normal distribution, whereas giant pulses show power-law cumulative distributions (Burke-Spolaor et al., 2012). Crab giant pulses have how evidence that the index depends on pulse width and energy, with flatter indices for weaker and shorter pulses (Karuppusamy et al., 2010; Popov & Stappers, 2007; Bera & Chengalur, 2019). High-energy magnetar emission has been described by power-law distributions with γ ranging from roughly -1.6 to -1.8 (Cheng et al., 2020). Previous studies of FRB 121102 energy distributions have used a single power-law fit ($N(> E) \propto E^\gamma$) to model the cumulative distribution and have obtained different values of γ ranging from -0.7 in Law et al. (2017), -1.1 in Cruces et al. (2020), -1.7 in Oostrum et al. (2020) and -1.8 in Gourdjji et al. (2019). Another well-studied repeating source, FRB 180916, shows $\gamma = -1.3$ at 400 MHz, although recent observations have reported a flattening of the power-law

at lower energies (Chawla et al., 2020; Pastor-Marazuela et al., 2021). We calculate the isotropic energy (E) of a burst as,

$$E = 4\pi 10^{-23} \left(\frac{D_L}{\text{cm}} \right)^2 \left(\frac{S}{\text{Jy s}} \right) \left(\frac{\sigma_f}{\text{Hz}} \right) \text{erg}. \quad (4.8)$$

Here, D_L is the luminosity distance to FRB 121102, 972 Mpc, as reported by Tendulkar et al. (2017). S and $2.355 \sigma_f$ are the fitted fluence and FWHM of the Gaussian spectra.

To make the cumulative energy distribution, we choose only bursts for which the $\pm 1\sigma$ bounds on the spectral peak fell within our observing band. This was done as our fluence estimates obtained from fitting are reliable for bursts within our band and because we are incomplete to the population of bursts that are partially outside our band (Aggarwal, 2021). Therefore, from a total of 133 bursts, we obtained 60 bursts that satisfied this criteria. For each of the 60 such bursts, we used the posterior distribution of bandwidths and fluences from the MCMC based fitting analysis to calculate the distribution of energies (using Eq. 4.8). We then randomly sample one energy from the burst energy distributions of each of the 60 bursts and generate a cumulative energy distribution using those 60 energies. We repeated this process 1000 times and thereby generated 1000 cumulative energy distributions.

The previous studies of the cumulative energy distribution of FRB 121102 have reported a break in power-law with a flattening towards low energies (Gourdji et al., 2019; Cruces et al., 2020). Also, it was visually evident in our data that the cumulative distribution flattened towards low energies. Therefore, a single power-law would not have been sufficient to accommodate the burst energy distribution. Therefore, we used `scipy.curve_fit` to fit each of these 1000 energy distributions with a broken power-law of the form,

$$N(\geq E) = \begin{cases} E_{\text{scale}} \left(\frac{E}{E_{\text{break}}} \right)^\alpha, & \text{if } E < E_{\text{break}} \\ E_{\text{scale}} \left(\frac{E}{E_{\text{break}}} \right)^\beta, & \text{if } E > E_{\text{break}}. \end{cases} \quad (4.9)$$

Here, α and β are the two power-law indices, E_{break} is the break energy, and E_{scale}

is the energy scaling. Figure 4.9 shows the cumulative energy distributions above an estimated completeness of 5.8×10^{36} ergs, calculated from the aforementioned completeness limit on the fluence, 0.0216 Jy ms (Section 4.4.3), and median bandwidth ($2.355 \sigma_f$) of the bursts i.e. 240 MHz. It also shows (in red) the power-law fit to each cumulative energy distribution.

The median of the distribution of fitted power-law indices and break energy (with 1σ errors) are given by $\alpha = -0.4 \pm 0.1$, $\beta = -1.8 \pm 0.2$ and $E_{\text{break}} = (2.28 \pm 0.19) \times 10^{37}$ ergs. This break at E_{break} could indicate the actual completeness energy limit of our observations, and we might therefore be incomplete to the bursts with energies below E_{break} . This could be due to the incompleteness of our observations to the weak, band-limited bursts. The bursts above this energy are well fitted by the power-law of index $\beta = -1.8 \pm 0.2$. The break in energy distribution has also been reported for other repeating FRBs (Pastor-Marazuela et al., 2021). It is worth noting that our higher-energy power-law index β is also consistent with the power-law index estimated by Gourdji et al. (2019) above a completeness threshold of 2×10^{37} ergs. In the context of pulse-energy distributions, the similarity of the power-law indices of FRB 121102 with both those of Crab giant pulses and magnetar pulses might also imply a common origin (Lyu et al., 2021).

4.6.1.1 Testing for a high-energy break

Figure 4.9 shows that two high-energy bursts deviate from the power-law fits. This has also been seen in Crab giant pulses, where this behavior was speculated to be due to supergiant pulses (Mickaliger et al., 2012). We therefore tested the presence of a high-energy break in the power-law (between $5 - 9 \times 10^{37}$ ergs). We assumed that the break energy estimated above to be the completeness threshold, and only used the energies greater than that value. Then we repeated the bootstrapping method to fit the cumulative distribution of the remaining bursts using: a single power-law and a broken power-law. Note that in this test we were only fitting the bursts with energy greater than 2.3×10^{37} ergs.

The fitted slope obtained for the single power-law fit was $-1.8_{-0.2}^{+0.1}$. The fitted

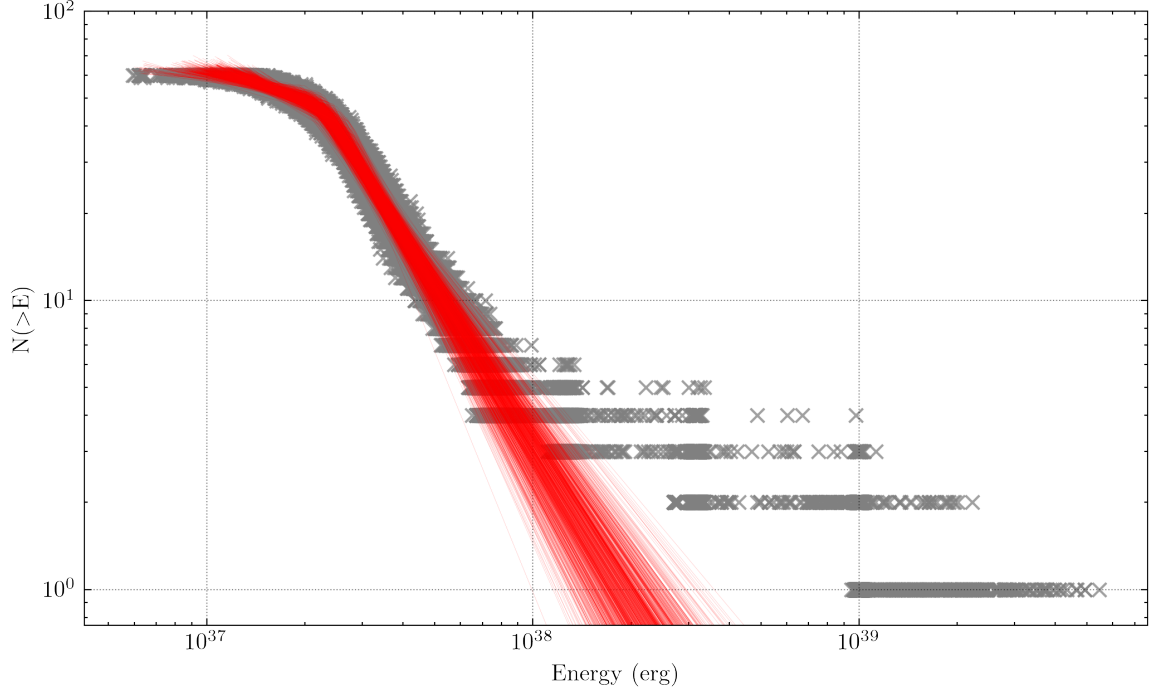


Figure 4.9: Cumulative energy distribution of bursts above our completeness limit of 5.81×10^{36} ergs. The grey points show the cumulative energy distributions created by bootstrapping the burst energies. The red lines show the broken power-law fit to each of these cumulative energy distributions. See Section 4.6.1 for more details.

slopes (below and above the break energy) for the double power-law fit were $-1.8^{+0.2}_{-0.2}$ and $-0.5^{+0.1}_{-0.3}$ with fitted break energy of $7.8^{+1.5}_{-1.9} \times 10^{37}$ ergs. The power-law slope obtained in the single power-law fit (and the lower energy slope in case of double power-law fit) was consistent with the higher energy slope reported earlier ($\beta = -1.8 \pm 0.2$). We also found that the reduced chi-square value for the single power-law case was $1.2^{+2.0}_{-0.6}$, while that for the double power-law fit was $0.08^{+0.08}_{-0.03}$. This indicates that the double power-law fit model over-fitted the data, and so a single power-law is sufficient. This test gives further confidence that above the energy of 2.3×10^{37} ergs, bursts from FRB 121102 follow a single power-law with slope $\beta = -1.8 \pm 0.2$, and that there is no evidence for a higher energy break in the energy distribution.

4.6.2 Wait-time distribution

The left panel in Figure 4.10 shows the distribution of wait times between bursts, which follows a bi-modal distribution as also seen in previous studies (Li et al., 2019a; Gourджи et al., 2019). On wait times greater than 1 s, we use `scipy.curve_fit` to fit a log-normal function to the main distribution, finding a peak at 74.8 ± 0.1 s. The peak of our wait-time distribution is significantly lower than the 207 ± 1 s peak found in this data by Gourджи et al. (2019), due to our increased sample of bursts filling in the wait-time gaps. As more bursts are discovered in an observation of constant length, the average time between bursts decreases, lowering the peak of the wait time distribution. Our findings most closely match the wait-time distribution of Zhang (2018), which peaks at ~ 67 s; while the original paper did not report an exact peak, we used their publicly available data¹⁴ to perform our wait-time analysis. These similarities suggest that careful single-pulse searches using machine-learning algorithms allow us to obtain a robust sample of bursts that accurately reflects the burst population.

As in previous studies (e.g., Li et al., 2019a; Katz, 2018; Gourджи et al., 2019), we find a smaller population of bursts with sub-second separations. However, unlike the previously reported distributions, which cluster around tens of milliseconds, our sub-second burst separations span the range of tens of milliseconds up to nearly one second without as clear of a break between the two distributions. For this analysis, we assume that each closely spaced pair of bursts is composed of two separate bursts instead of components of a single broader burst, leading to the larger distribution of short wait times compared to other papers. This assumption will not drastically alter the fitted log-normal distribution since the sub-second wait-time population is small, and their removal will not significantly alter the shape of the main distribution. As in Gourджи et al. (2019) and Li et al. (2019a), we find that the wait time between bursts and their relative fluences are not correlated.

To quantify the change in the wait-time peak as a function of the number of bursts in a sample, we perform the same wait-time analysis for a random selection

¹⁴Table 2, accessible at <https://doi.org/10.3847/1538-4357/aadf31>

of our bursts over a range of burst numbers, as seen in the right-hand panel of Figure 4.10. Each point represents a fit to 300 random selections of that number of bursts, with the error bars representing the standard deviation of all of the fitted peaks. Our findings show the expected effect that as more bursts are included in a sample from a constant-length observation, the average time between bursts will decrease along with the fitted wait-time distribution peak. We find that the distribution of fitted wait times peaks exponentially decays with added bursts with a timescale of ~ 29 seconds. We observe the same effect in the Zhang (2018) dataset, which shows a timescale of ~ 25.5 seconds. The peak obtained by Gourdji et al. (2019) using this data is shown in the figure and matches with our fitted exponential curve. We perform a similar analysis by filtering out the lowest fluence bursts from our sample and find that the wait time peak increases as the minimum fluence limit is increased, and weaker bursts are excluded. This serves to explain the higher wait-time peaks calculated by previous papers with fewer bursts and higher fluence limits.

4.6.3 Short-period periodicity search

In this section we discuss the results of various periodicity searches tested on this burst sample to search for a short-period periodicity (millisecond to hundreds of seconds).

4.6.3.1 Difference Search

We first perform a periodicity search on the burst times (μ_t in Table 4.2) by calculating the differences between consecutive pulse times and searching over a range of trial periods to determine how many differences are evenly divisible, within some tolerance, by this period. We searched over trial periods starting with the minimum difference between pulses, in integer divisors down to the minimum difference divided by 256, after removing all differences less than 50 ms, in both the full set of pulses and 100 trials where only three-quarters of the bursts were randomly selected (to sample a more complete range of minimum differences). Differences less than 50 ms were

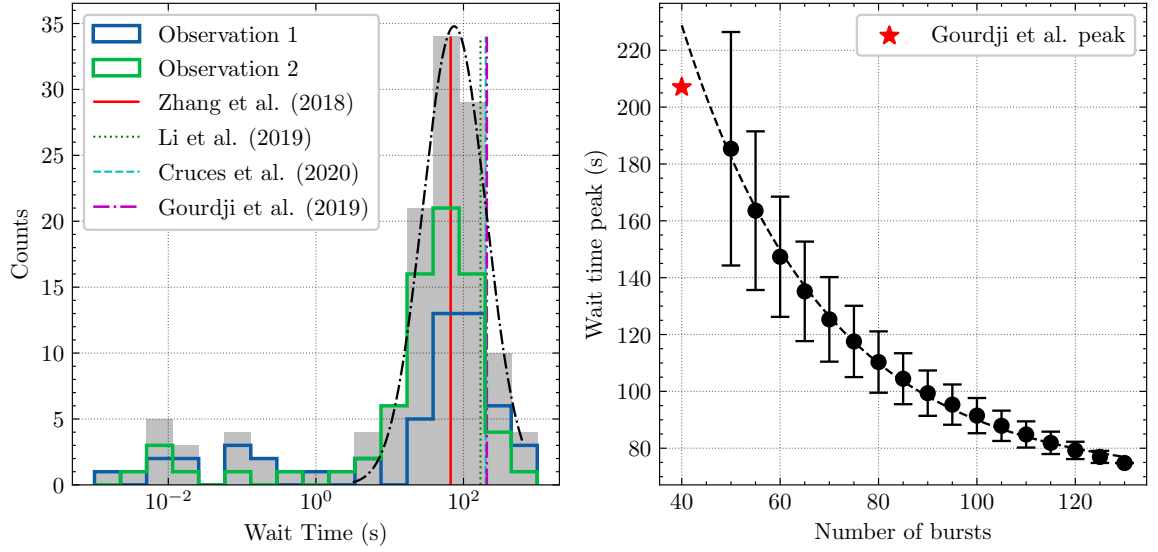


Figure 4.10: Left: the wait time distribution for both observations is plotted in gray, with the fitted log-normal function peaking at 74.8 ± 0.1 s. The individual observations are plotted in green and blue, respectively. The fitted wait-time peaks from previous works are notated with vertical lines. Right: the average fitted wait-time peaks as a function of the number of bursts randomly selected from the full sample. The fitted exponential distribution is shown with a black dashed line, and the error bars represent one standard deviation of the fitted peaks. The fitted peak from Gourdj et al. (2019) is represented with a red star.

removed, since potential single bursts with widths greater than 30 ms have been reported from FRB 121102 (e.g. Katz, 2018), and the distinction between whether these bursts are single multi-component bursts or separate bursts is unclear. Furthermore, shorter trial periods are much more likely to return false positive results. To allow for a variety of possible emission mechanisms for FRB 121102, including a broad or multi-component pulse profile, we searched phase tolerances ranging from 1% to 50%. At any given phase tolerance, a trial period is considered to fit a difference between two pulses if the difference is an integer multiple of the trial period, within an error equal to the phase tolerance.

We also searched 1000 simulated time series of identical length, with the same number of pulses distributed randomly, using the same methodology in order to gauge the significance of any detected periodicities. By searching for periods in a set of bursts with no underlying period, we can evaluate whether our period search finds a real periodicity in the data, or if it is a coincidence. Above a 50% tolerance, we get many more pulse matches in all of the random timeseries than we do with the real data, likely due to the FRB pulse distribution not following a random distribution (see Figure 4.10 for the distribution of pulse arrival time differences).

The most significant period found was 658.838 milliseconds, which fit nine pulses at a tolerance of 3%, with a false alarm probability (FAP) of 0.3% for random trials at that tolerance. However, considering all 50 tolerance values searched over two observations, the effective FAP is 30%, and we therefore conclude that no periodicity can be detected through this differencing method.

4.6.3.2 Fast Folding Algorithm

We ran a fast folding algorithm (FFA) on each observation using `riptide` (Morello et al., 2020). Unlike the periodicity search in the previous section which uses the calculated pulse arrival times, `riptide` searches for periodic signals in the entire dedispersed time series. This allows us to efficiently search over a greater range of trial periods, and will not be affected by issues such as missing bursts, or considering closely spaced individual bursts as a single multi-component burst. However,

while it is more sensitive to weak, time-averaged periodic emission, it is less sensitive to periodicities only found in the detected single pulses. The FFA folds each dedispersed time series over a range of trial periods to create an integrated pulse profile. For each observation, we searched time series with DMs ranging from 550 pc cm^{-3} to 580 pc cm^{-3} and at periods greater than 500 ms (the approximate period at which folding algorithms are more sensitive than Fourier techniques; see Parent et al., 2018) and less than 20 s (to ensure a sufficient number of pulses across the observation for a pulsed detection). We used 1024 output bins, with boxcar filters providing sensitivity to pulses with widths ranging from 1 to 300 ms.

First, candidates due to RFI, such as periods at exact integers and known RFI frequencies, were removed. Of the remaining candidates, a signal-to-noise cutoff of 10σ was applied for a total of 1,250 candidate periods between the two observations.

We then folded the relevant dedispersed time series for each candidate using the `prepfold` command from the `PRESTO` (Ransom, 2011) package at the candidate period and DM identified by `riptide`. We allowed `prepfold` to search in DM in order to determine which candidates had signal-to-noise ratios which peaked for DMs inside of the searched range. We expanded our acceptable DM range from 520 to 610 pc cm^{-3} to allow leeway in `prepfold`’s search, leaving 91 periodic candidates in total. We visually inspected these profiles and found that all were consistent with RFI or noise, with no evidence of emission at the same phase over time or frequency. We compared the candidate periods found by the algorithm on both observations, and found that the only common periods were caused by RFI.

4.6.3.3 `frbpa`

We used `frbpa` (Aggarwal et al., 2020) to search for a short-period periodicity using the burst times (MJDs). We used two methods: search to find the period that minimizes the fractional width of folded profile (Rajwade et al., 2020b) and a Quadratic-Mutual-Information-based periodicity search technique (Huijse et al., 2018). In the first method, we phase-coherently folded the burst times between trial periods and generated a set of profiles consisting of the source activity with respect

to the trial period phase. We then measured the width of the source activity in each folded profile. Low width would indicate that source activity is concentrated in a small set of contiguous phase windows, indicating the presence of periodic activity (Rajwade et al., 2020b). The second method uses quasi-mutual-information to estimate the period. It has been shown to be robust to noise and works well on sparsely sampled data as well (Huijse et al., 2018). We searched for periods between 1–1000 seconds on bursts from the two days individually and did not recover any significant period.

4.6.3.4 Lomb-Scargle

We also used `timeseries.LombScargle` from the `astropy` library (Astropy Collaboration et al., 2013; Price-Whelan et al., 2018) to search for periods ranging from 100 ms to 1000 seconds on the bursts from the two observations separately. Note that sensitivity to 100-ms periods requires sampling frequencies higher than the traditional Nyquist limit. This is possible because the effective Nyquist frequency for unevenly sampled data set is much smaller than the traditional limit (VanderPlas, 2018). The most significant periods are approximately 118 and 179 milliseconds. However, false alarm probabilities of around 3% and 26%, respectively, indicate that the detected periods are unlikely to be real. We therefore conclude that we did not detect any significant periodicity in the bursts using the Lomb-Scargle method.

4.6.4 Burst Rate

Previous observations of FRB 121102 have found significant evidence for pulse clustering on short time scales, where the burst separations deviate from a Poissonian distribution (Oppermann et al., 2018; Oostrum et al., 2020; Cruces et al., 2020). The Weibull distribution, as described in Oppermann et al. (2018), is a modification of the Poisson distribution, with a shape parameter k describing the degree of pulse clustering. Clustering is present for $k < 1$ with lower values corresponding to more clustering, while $k = 1$ reduces to the Poissonian case; a value of $k \gg 1$ causes the distribution to peak more sharply at the event rate and indicates a periodic signal.

A better understanding of the burst statistics may help us understand the progenitor of FRB 121102 and help strategize the timing of future observations.

Figure 4.11 shows the cumulative probability density of the wait times between consecutive bursts, fitted to the Weibull and Poisson cumulative density functions as defined by Oppermann et al. (2018). The fitted values are given in Table 4.4, as well as the reduced chi-squared statistic and the coefficient of determination, r^2 , which ranges from zero to one with a value of one representing a perfect fit. In Figure 4.11, we also plot the values of the reduced chi-squared statistic for the Poisson and Weibull distributions when fitted to only the wait times longer than a range of chosen minimum wait times. We observe that both the Poisson and Weibull distributions fit the main population of longer wait times much better than the entire set of bursts. The Weibull distribution’s ability to account for clustering allows it to have a significantly better fit when including shorter wait times, as its reduced chi-squared statistic is a factor of 18 smaller than that of the corresponding Poisson distribution (see Table 4.4). However, the Weibull distribution is only slightly favored over the Poisson distribution when fitting only longer wait times. We find that the reduced chi-squared statistic is equal to 0.368 for the Weibull fit to wait times greater than one second, indicating that the distribution is overfitted. These findings may indicate that the main distribution of bursts with longer wait times roughly follows a Poissonian distribution, while the entire burst rate distribution cannot be accurately described with solely a Poisson or Weibull distribution. This may result from our decision to consider each burst as a separate burst rather than a sub-component of a broad burst.

In addition to their observations, Cruces et al. (2020) used the original dataset from Gourdji et al. (2019) to study the burst rate statistics and found that the addition or removal of the sub-second wait time population in each dataset significantly impacts the extrapolated burst rate behavior. In each case, removing these short wait times led to the fitted Weibull shape parameter k increasing towards one, further indicating that the main distribution of pulses may follow a Poissonian distribution while the shorter distributions do not. However, the sample used in Cruces et al. (2020)

Table 4.4: Fitted burst rate distributions.

	Rate (hour ⁻¹)	k	χ^2	r^2
Poisson (all)	65 ± 8.4	\dots	495	0.953
Weibull (all)	42 ± 9	0.63 ± 0.07	27.5	0.970
Poisson ($\delta t > 1\text{s}$)	41 ± 1.6	\dots	1.076	0.994
Weibull ($\delta t > 1\text{s}$)	46 ± 1.5	1.16 ± 0.04	0.368	0.997

The posterior values for the Poisson and Weibull distributions as well as the reduced chi-squared statistic and r^2 value, fit both to the entire set of wait times as well as only wait times greater than one second. The errors represent 1σ uncertainties.

only had two sub-second wait times; our more extensive sample of short wait times allows us to confirm this behavior with more statistical significance. In Figure 4.12, we plot the fitted burst rates for the Weibull and Poisson distributions and the Weibull shape parameter k as a function of minimum wait time. As the minimum wait time increases, both burst rates converge to a rate of roughly 45 bursts per hour. We also find that the fitted value of k increases with the minimum wait time, reaching a value of $k = 1$ at a minimum wait time cutoff of roughly 0.1 s.

4.6.5 Implications for progenitor models

Based on our results, any progenitor model proposed for FRB 121102 would have to explain the following observations: (1) band-limited emission; (2) varying peak emission of the spectra and its lack below 1300MHz; (3) median scattering timescale of 0.7 ms, with a maximum value of around 2 ms; (4) rapid variability of these three properties at second timescales. Further, some of these observations have also been reported for other FRBs (Kumar et al., 2021c; Shannon et al., 2018; Pastor-Marazuela et al., 2021).

4.6.6 Comparison to previous work

In this chapter, we have presented 93 additional bursts detected on reprocessing the data presented in Gourdji et al. (2019). Figure 4.13 shows the distribution of properties of new bursts detected with our pipeline as compared to the ones already

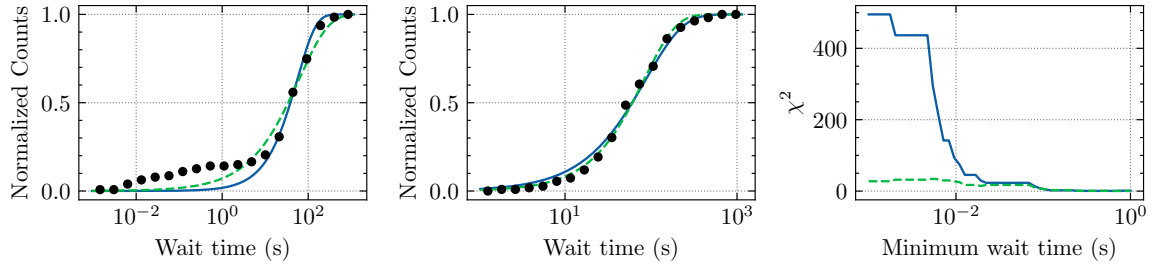


Figure 4.11: The cumulative wait time distributions fitted by the Poisson (blue solid lines) and Weibull (green dashed lines) distributions for the entire sample of wait times (left) and for the population of wait times greater than one second (middle). The fitted values for the Poisson distribution for the left and middle plots are $r = 65 \text{ hr}^{-1}$ and $r = 41 \text{ hr}^{-1}$ respectively. The fitted values of the Weibull distribution are $k = 0.634$, $r = 42 \text{ hr}^{-1}$ and $k = 1.162$, $r = 46 \text{ hr}^{-1}$ respectively. Right: The reduced chi-squared statistic of the Poisson and Weibull fits, when fitted to wait times greater than the minimum wait time. While the middle and right panels show that both the Weibull and Poisson distributions fit the longer wait times well, the left panel shows that neither distribution accurately fits for the larger sample of sub-second wait times. The right panel shows that pulse clustering starts to impact the fits once wait times shorter than $\sim 50 \text{ ms}$ are considered.

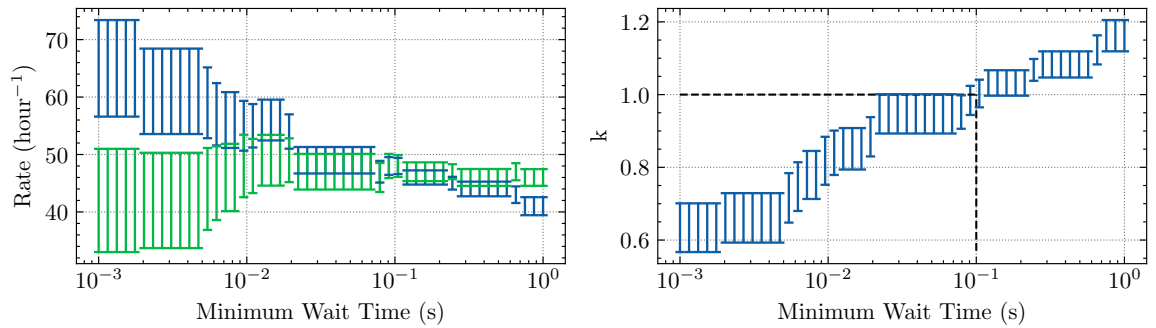


Figure 4.12: Left: the fitted burst rate per hour for the Poisson (blue) and Weibull (green) distributions as a function of the minimum wait time used in the fitting. As the minimum wait time increases and we begin to sample only the main distribution of wait times, the burst rates converge to roughly 45 bursts per hour. Right: The Weibull shape parameter k as a function of minimum wait time. The fitted value increases with the minimum wait time, and black dotted lines have been added to indicate that the Weibull parameter increases to a value of 1.0 at a minimum wait time of 0.1 seconds.

published by Gourdji et al. (2019). We performed Kolmogorov-Smirnov (KS) tests to compare the distributions of fitted parameters of old versus the new bursts. The distribution of S , σ_t and DM are similar, while those for μ_f , σ_f and τ are different. This indicates that the searches carried out by Gourdji et al. (2019) missed bursts that span the entire range of these parameter values, rather than just the weaker bursts.

Recently, Li et al. (2021) reported a large sample of bursts from FRB 121102 detected using the Five-hundred-meter Aperture Spherical radio Telescope. The mean of the wait-time distribution estimated from that larger sample of bursts (70 ± 12 s) is consistent with what we report in Section 4.6.2. They also did not detect any short-term periodicity, similar to our findings. Further, they report a bimodal energy distribution for FRB 121102. Aggarwal (2021) highlighted that this bimodality disappears when burst bandwidth, instead of center frequency of the observing band, is used to calculate the energy. Moreover, the burst bandwidths reported in Li et al. (2021) were estimated visually, and not using a fitting procedure. This can lead to observational biases that will make the interpretation of intrinsic energy distributions difficult (Aggarwal, 2021).

All the tools and software used in our pipeline are independent of the ones used in the original work (Gourdji et al., 2019). This brings the critical question of understanding why our pipeline detected more bursts or why the original work missed the bursts. While an exhaustive comparison of the two pipelines using a standard dataset consisting of simulated FRBs is beyond the scope of this chapter, here, we try to investigate the reasons for different results based on our understanding of the software used. We discuss three reasons that might contribute to very different recovery rates¹⁵ of the two pipelines.

¹⁵By recovery rate, we mean the percentage of transients (above the completeness limit) correctly identified by the software. A perfect pipeline would have a recovery rate of 100%, indicating that it can detect all transients present in the data.

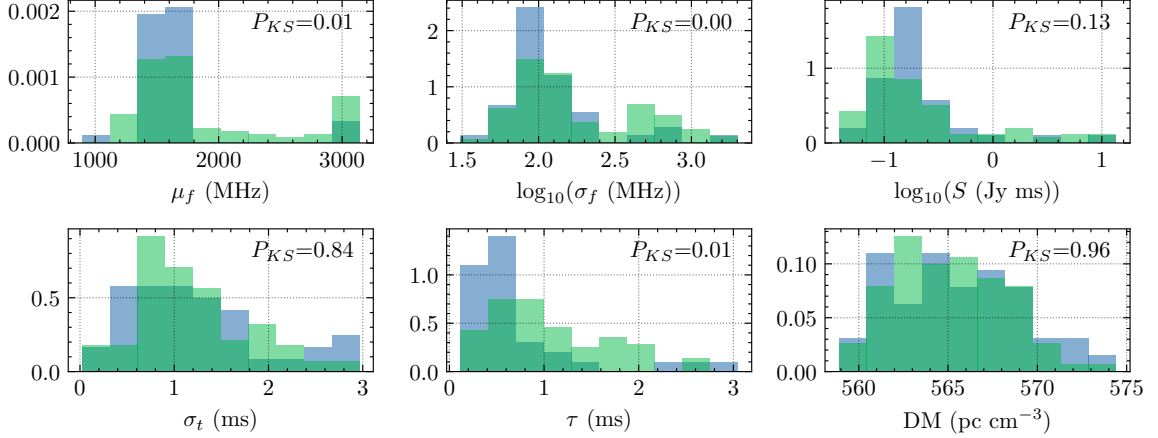


Figure 4.13: Normalized histogram of burst properties. New bursts detected in this analysis are shown in green, while the ones published by Gourdji et al. (2019) are shown in blue. For all the bursts, the properties shown here were estimated using the fitting procedure described in Section 4.4.5. We did a KS-test to compare the old and new burst distributions, for all the properties, with p-values obtained from these tests given on the top right of each plot.

4.6.6.1 Threshold signal-to-noise ratio

Although Gourdji et al. (2019) used a S/N threshold of 6 for the search, they discarded candidate groups with a maximum S/N less than 8. We used a S/N threshold of 6 in our search. Therefore, they would have missed low S/N candidates.

Assuming that the higher energy power-law slope of -1.8 (estimated in Section 4.6.1) is intrinsic to FRB, we can estimate the expected ratio of number of bursts with S/N greater than 6 to that above a S/N of 8. This is given by $(6/8)^{-1.8} = 1.7$. The observed ratio of the number of bursts above S/N of 6 ($N = 133$) to that above S/N of 8 ($N = 70$) is: $133/70 = 1.9$. This implies that we detected more bursts between S/N of 6 and 8, than expected from the power-law distribution of energies. But, it is important to note that this simple estimate relies on the following assumptions:

1. Our observations are complete to bursts with $S/N < 8$. But the flatter energy distribution at low energies indicates that this might not be true.
2. The fluence and energy distributions are similar. This is true only when there is no incompleteness due to banded nature of the burst spectra (Aggarwal, 2021).

3. The burst energy distribution can be modeled by a single power-law, even at low energies.

Further, we now estimate the number of purely noise candidates we expect above a S/N greater than 6 in our search. The chance probability of a purely noise candidate with S/N greater than 6 is $P = 9.9 \times 10^{-10}$. The number of trials in our search can be calculated by:

$$N_{\text{trial}} = N_{\text{DM}} \sum_{i=0}^{10} \frac{N_{\text{time}}}{2^i}, \quad (4.10)$$

where, N_{trial} is the total number of trials in our search, N_{DM} is the number of trial DMs (65, between 450–650 pc cm⁻³ at a tolerance of 1%) and N_{time} is the total number of time samples in the data (~ 3 hrs at a time resolution of 81.92- μ s). The sum is over the boxcar widths searched (2^0 to 2^{10} , doubling at each step). Therefore, we expect $P \times N_{\text{trial}} \sim 20$ purely noise events, with a S/N greater than 6, for these number of trials. But number of events we detected with $S/N > 6$ in our single pulse search were 1,427 (with 133 FRB and 1,294 RFI candidates).

This shows that we detected much more candidates above a S/N of 6 than expected from pure Gaussian noise, implying that data is non-Gaussian. Most of the candidates we obtained were due to RFI, which is expected. It is possible that some weak events are still misidentified, however this will not influence the results of our analysis.

4.6.6.2 Single-pulse search software

Gourdji et al. (2019) used `single_pulse_search.py` (Ransom, 2011) to search for the bursts, and manually verified the candidates. In contrast, we used HEIMDALL for the single-pulse search, and FETCH for classification. Keane & Petroff (2015) highlight several steps at which a single-pulse search pipeline might miss a transient. A few such steps are sizes of boxcar convolutional kernel, spacing between DM trials, the position of the boxcar convolution with respect to the phase of the pulse, clustering

of redundant candidates, etc. Although recovery rates for both these search softwares have been shown to be $>90\%$ in their respective pipelines (Patel et al., 2018; Agarwal et al., 2020b; Gupta et al., 2021), a thorough comparison of the two search strategies has so far not been done.

4.6.6.3 RFI mitigation and classification

Gourdji et al. (2019) used a different RFI mitigation strategy than our pipeline. They used the classifier (SPS) presented in Michilli et al. (2018) to filter the RFI candidates and then manually verified the remaining 125 candidates to search for real pulses. SPS was designed specifically to search for Galactic single pulses in a LOFAR survey. Features in LOFAR data would be very different from the data used in this study. This is primarily due to different observing bands (1.4GHz for this study, compared to 100MHz for LOFAR), RFI environment, telescopes, and observing backends. Moreover, the dispersion in FRB pulses is typically much larger than that seen in Galactic transients. Due to these differences, it is not possible to translate the performance of SPS on LOFAR data to the data used in this chapter. Michilli et al. (2018) mention the use of specially designed filters for such datasets on which SPS was not trained, but they did not report the performance of these filters on any such data. Agarwal et al. (2020a) and Connor & van Leeuwen (2018) also highlight that it is non-trivial to generalize a machine learning algorithm to unseen data without rigorous pre-processing and injection tests. It is therefore possible that SPS missed to correctly identify real pulses that were correctly detected by `single_pulse_search.py`.

As mentioned previously, we used spectral kurtosis for RFI mitigation (Nita & Gary, 2010) and used FETCH for classification. It is known that very strict RFI flagging can lead to a reduced recovery rate, as the RFI algorithm might flag real signal as RFI (Rajwade et al., 2020a). Spectral kurtosis is robust to astrophysical signals and performs better than the simple median-based RFI thresholds (Nita & Gary, 2010). FETCH was developed to be robust to different telescopes, observing configurations and RFI environments (Agarwal et al., 2020a). We accomplished this by a carefully designed pre-processing and training strategy. We also showed that

the performance of FETCH remains consistent on unseen data. This has been further established by the new FRBs discovered using FETCH on different telescopes (Kumar et al., 2019; Law et al., 2020; Aggarwal et al., 2020; Rajwade et al., 2020b; Kumar et al., 2021c; Pleunis et al., 2021b; Kirsten et al., 2021b).

4.6.6.4 General comments

As mentioned in Section 4.4.3, a robust analysis of any single-pulse search software and pipeline requires exhaustive injection analysis. Such an analysis can highlight the percentage of transients recovered with respect to various physical parameters of interest, like DM, width, phase, etc. These metrics are essential in understanding possible inefficiencies and estimating the completeness of any single-pulse search. Such analysis done for specific pipelines that use HEIMDALL has reported a recovery rate of $> 90\%$ (Agarwal et al., 2020b; Gupta et al., 2021). While our pipeline is not identical, it is very similar to the one reported by Agarwal et al. (2020b). On a less diverse dataset, the recovery rates for `single_pulse_search.py` were also reported to be $> 90\%$ (Patel et al., 2018), but a more rigorous analysis would better assess the robustness of this search software.

Additionally, if a machine learning classifier is deployed in a single-pulse search pipeline, then it is necessary to estimate the recovery rate for the classifier itself. Such rigorous analyses have been performed for some classifiers (Agarwal et al., 2020a; Gupta et al., 2021), and provide insights into their recovery rate estimates. The robustness of a classifier to unseen data should also be carefully investigated (or verified by manual inspection), before reporting the completeness of any search. While human recovery rates for visual classification of thousands of candidates have not been estimated for FRBs, similar estimates are present in computer vision literature. Here, the neural networks (a top-5 error rate of 3.5%) routinely outperform humans (a top-5 error rate of 5.1%) in image classification tasks (He et al., 2015; Russakovsky et al., 2014).

Further, as most of the bursts detected in this sample are narrow-band (i.e., present only in a part of the frequency band), traditional searches might still miss

bursts. Sub-banded searches would be more sensitive to detect such band-limited bursts, especially at wide-band systems (R. Anna-Thomas et al. 2021, in preparation; Kumar et al., 2021c; Gourdji et al., 2019).

4.6.7 Caveats

Finally, it is appropriate to highlight and reiterate four main caveats to the analysis presented in this chapter. (1) The data used in this analysis was downsampled to 64 frequency channels. We did not have access to the native resolution data; therefore, all the search and analysis was performed on downsampled data. The sensitivity of single-pulse search would be higher on native resolution data; therefore, our pipeline may have missed some pulses. (2) We only performed a search on the data that averaged over the full bandwidth to create timeseries. Given the band-limited nature of many bursts, a sub-band search on this data might reveal weaker and narrower bursts. (3) As mentioned previously, our estimate of completeness limit is not robust, as such a robust estimate requires injection analysis that was not possible with the available data. (4) The reported properties of the bursts depend on the assumption that burst spectra and profile follow the assumed functional forms used for fitting.

4.7 Conclusions

This chapter presents a dense sample of FRB 121102 bursts detected at L-band using Arecibo Observatory, analyzed as a part of The Petabyte Project. More importantly, we report 93 new bursts detected with our single-pulse search pipeline, as compared to the published results (Gourdji et al., 2019), making a total of 133 burst detections in 3 hours of data. We have developed a robust burst fitting procedure to model the spectro-temporal properties of FRBs and provide it as a user-friendly python package BURSTFIT¹⁶. We use the MCMC procedure implemented in BURSTFIT to estimate the properties of all the bursts in our sample. We find that the burst

¹⁶<https://github.com/thepetabyteproject/burstfit>

spectra can be well modeled using a Gaussian function, with a median width of 230 MHz and a median peak at 1608 MHz. Most of the burst emission is present in the top of our band, and there is a lack of emission below 1300 MHz, consistent with other published results (Gourdji et al., 2019; Platts et al., 2021). Many bursts also show a scattering tail, with a median timescale of 0.7 ms. Some bursts show complex structures like multiple components and frequency drift. The wait time distribution of the bursts shows two distributions, at millisecond and second timescales. The latter of the two follows a log-normal distribution, with the peak at 74.8 s, consistent with other published results (Zhang, 2018). We further note that the peak of the wait time changes significantly based on the number of bursts in an observation. We find that both Poisson and Weibull distributions fit the burst rate distributions at long wait times (> 1 second) equally well, and neither accurately describes the whole burst rate distribution. We did not detect any short-period periodicity in the bursts. The cumulative burst energy distribution is well modeled by a double power-law with a break. We find the value of low and high-energy slopes to be -0.4 ± 0.1 and -1.8 ± 0.2 with a break at $(2.3 \pm 0.2) \times 10^{37}$ ergs. Our analysis reveals that only the bursts that are completely within the band should be used for energy distribution analysis. We discuss some possible differences between our single-pulse search pipeline and the one used by Gourdji et al. (2019), to explain the different results obtained using the two approaches. All the analysis scripts and results presented in this chapter are openly available in a Github repository¹⁷ for the community to use in their repeater burst analyses.

4.8 Appendix

Table 4.5 shows the fitted properties (with 1σ errors) of all the bursts reported in this analysis.

¹⁷<https://github.com/thepetabyteproject/FRB121102>

Table 4.5: Properties of all the bursts presented in this analysis.

Burst ^a	μ_f	σ_f	S	μ_t^b	σ_t	τ^c	DM
ID	(MHz)	(MHz)	(Jy ms)	(MJD)	(ms)	(ms)	(pc cm ⁻³)
B1*	1560 ⁺³⁰ ₋₃₀	210 ⁺⁴⁰ ₋₄₀	0.09 ^{+0.02} _{-0.02}	57644.408906976(1)	0.0 ^{+0.02} _{-0.02}	1.9 ^{+0.3} _{-0.3}	565.3 ^{+0.4} _{-0.4}
B2*	1200 ⁺¹⁰ ₋₁₀	50 ⁺¹⁰ ₋₁₀	0.043 ^{+0.007} _{-0.007}	57644.40956768(1)	1.35 ^{+0.05} _{-0.05}	-	562.4 ^{+0.8} _{-0.8}
B3.1†	2900 ⁺³⁰⁰ ₋₆₀₀	800 ⁺³⁰⁰ ₋₂₀₀	0.6 ^{+0.7} _{-0.3}	57644.409673699(3)	0.4 ^{+0.2} _{-0.2}	1.3 ^{+0.7} _{-0.7}	566.8 ^{+0.8} _{-0.9}
B3.2†	1100 ⁺³⁰⁰ ₋₁₄₀₀	1000 ⁺²⁰⁰⁰ ₋₁₀₀₀	0.09 ^{+0.13} _{-0.04}	57644.40967384(2)	0.3 ^{+1.4} _{-0.2}	0.3 ^{+0.7} _{-0.1}	564.7 ^{+1.5} _{-0.4}
B4†	3100 ⁺²⁰⁰ ₋₆₀₀	550 ⁺⁸⁰ ₋₁₁₀	2 ⁺⁴ ₋₂	57644.410072889(4)	1.1 ^{+0.2} _{-0.2}	0.3 ^{+0.2} _{-0.1}	564 ⁺¹ ₋₁
B5†	2100 ⁺⁹⁰⁰ ₋₁₆₀₀	2700 ⁺⁹⁰⁰ ₋₁₂₀₀	0.19 ^{+0.05} _{-0.04}	57644.410157834(4)	0.7 ^{+0.3} _{-0.3}	1.0 ^{+0.4} _{-0.3}	562.1 ^{+0.7} _{-0.6}
B6.1	1393 ⁺⁷ ₋₇	183 ⁺⁷ ₋₇	0.47 ^{+0.02} _{-0.03}	57644.411071954(1)	1.09 ^{+0.03} _{-0.04}	-	562.3 ^{+0.2} _{-0.1}
B6.2	1417 ⁺⁴ ₋₅	102 ⁺⁵ ₋₄	0.33 ^{+0.03} _{-0.02}	57644.4110719755(9)	0.57 ^{+0.03} _{-0.02}	-	560.9 ^{+0.2} _{-0.1}
B7.1†	3100 ⁺²⁰⁰ ₋₃₀₀	430 ⁺⁶⁰ ₋₈₀	10 ⁺²⁰ ₋₁₀	57644.412240214(5)	0.7 ^{+0.3} _{-0.4}	0.8 ^{+0.4} _{-0.4}	569 ⁺³ ₋₃
B7.2†	1460 ⁺²⁰ ₋₂₀	90 ⁺³⁰ ₋₂₀	0.09 ^{+0.01} _{-0.01}	57644.41224043(2)	1.9 ^{+0.9} _{-1.0}	1.2 ^{+0.6} _{-0.5}	569 ⁺³ ₋₂
B8†	3000 ⁺²⁰⁰ ₋₆₀₀	700 ⁺¹⁰⁰ ₋₁₀₀	1.1 ^{+1.5} _{-0.7}	57644.414123628(4)	1.0 ^{+0.3} _{-0.3}	0.8 ^{+0.4} _{-0.3}	567.5 ^{+0.9} _{-0.7}
B9*	1430 ⁺¹⁰ ₋₁₀	75 ⁺⁹ ₋₉	0.076 ^{+0.003} _{-0.003}	57644.41447161(2)	2.0 ^{+0.6} _{-0.6}	0.4 ^{+0.6} _{-0.6}	564 ⁺² ₋₂
B10	1630 ⁺¹⁰ ₋₁₀	82 ⁺⁸ ₋₈	0.1 ^{+0.01} _{-0.01}	57644.414475391(7)	1.4 ^{+0.3} _{-0.3}	0.5 ^{+0.3} _{-0.2}	562 ⁺³ ₋₃
B11	1550 ⁺¹⁰ ₋₁₀	100 ⁺¹⁰ ₋₁₀	0.074 ^{+0.008} _{-0.007}	57644.414878803(4)	0.8 ^{+0.2} _{-0.2}	0.3 ^{+0.2} _{-0.1}	560.3 ^{+0.9} _{-0.9}
B12	3100 ⁺²⁰⁰ ₋₃₀₀	450 ⁺⁶⁰ ₋₈₀	20 ⁺²⁰ ₋₁₀	57644.41537809(1)	1.9 ^{+0.5} _{-0.5}	1.5 ^{+1.3} _{-0.8}	569 ⁺³ ₋₄
B13	1692 ⁺¹³ ₋₉	90 ⁺²⁰ ₋₁₀	0.14 ^{+0.01} _{-0.01}	57644.416314736(5)	1.5 ^{+0.2} _{-0.2}	0.6 ^{+0.3} _{-0.2}	572 ⁺¹ ₋₁
B14	1260 ⁺¹⁰ ₋₁₀	67 ⁺⁹ ₋₈	0.13 ^{+0.02} _{-0.02}	57644.41830674(8)	3 ⁺¹ ₋₁	1.6 ^{+0.6} _{-0.5}	562 ⁺⁶ ₋₅
B15.1†	1550 ⁺¹⁰ ₋₁₀	80 ⁺¹⁰ ₋₁₀	0.07 ^{+0.01} _{-0.01}	57644.418309206(7)	0.9 ^{+0.3} _{-0.3}	0.8 ^{+0.5} _{-0.4}	565 ⁺² ₋₁
B15.2†	2000 ⁺¹⁰⁰⁰ ₋₁₀₀₀	3000 ⁺¹⁰⁰⁰ ₋₁₀₀₀	0.12 ^{+0.03} _{-0.04}	57644.418309273(2)	0.2 ^{+0.2} _{-0.1}	0.7 ^{+0.2} _{-0.2}	569.4 ^{+0.4} _{-0.5}
B16†	3100 ⁺²⁰⁰ ₋₆₀₀	510 ⁺⁶⁰ ₋₁₁₀	6 ⁺⁹ ₋₅	57644.420508580(4)	1.2 ^{+0.2} _{-0.2}	0.5 ^{+0.3} _{-0.2}	570 ⁺² ₋₂
B17	2000 ⁺¹⁰⁰⁰ ₋₂₀₀₀	3000 ⁺¹⁰⁰⁰ ₋₁₀₀₀	0.15 ^{+0.05} _{-0.04}	57644.42110545(1)	1.5 ^{+0.6} _{-0.5}	0.7 ^{+1.0} _{-0.4}	570 ⁺¹ ₋₂
B18*	1530 ⁺³⁰ ₋₃₀	140 ⁺³⁰ ₋₃₀	0.046 ^{+0.008} _{-0.008}	57644.426303862(4)	1.5 ^{+0.1} _{-0.1}	-	564.8 ^{+0.8} _{-0.8}
B19*	1330 ⁺³⁰ ₋₃₀	160 ⁺²⁰ ₋₂₀	0.11 ^{+0.02} _{-0.02}	57644.42721003(2)	1.9 ^{+0.8} _{-0.8}	1.6 ^{+0.5} _{-0.5}	564 ⁺¹ ₋₁
B20†	2900 ⁺⁴⁰⁰ ₋₁₀₀₀	1000 ⁺²⁰⁰⁰ ₋₄₀₀	0.23 ^{+0.47} _{-0.08}	57644.427376859(4)	0.7 ^{+0.3} _{-0.3}	0.6 ^{+0.5} _{-0.3}	565.6 ^{+0.8} _{-0.6}
B21†	3100 ⁺²⁰⁰ ₋₄₀₀	600 ⁺¹⁰⁰ ₋₁₀₀	3 ⁺⁷ ₋₂	57644.42794036(2)	4.2 ^{+0.9} _{-1.0}	1.8 ^{+1.7} _{-0.8}	576 ⁺⁴ ₋₅
B22†	3000 ⁺³⁰⁰ ₋₇₀₀	700 ⁺¹⁰⁰ ₋₂₀₀	1.3 ^{+1.7} _{-0.9}	57644.428593784(8)	1.9 ^{+0.4} _{-0.4}	1.5 ^{+0.6} _{-0.6}	568 ⁺² ₋₂
B23	1699 ⁺¹⁴ ₋₉	70 ⁺²⁰ ₋₁₀	0.089 ^{+0.009} _{-0.008}	57644.430170170(2)	0.7 ^{+0.1} _{-0.1}	0.3 ^{+0.2} _{-0.1}	569.0 ^{+1.0} _{-1.0}
B24*	1670 ⁺²⁰ ₋₂₀	80 ⁺²⁰ ₋₂₀	0.0384 ^{+0.0007} _{-0.0007}	57644.430170303(1)	0.02 ^{+0.05} _{-0.05}	1.2 ^{+0.3} _{-0.3}	563.0 ^{+0.5} _{-0.5}
B25	1750 ⁺¹³⁸⁰ ₋₇₀	200 ⁺³⁹⁰ ₋₆₀	0.16 ^{+3.67} _{-0.03}	57644.430171419(4)	1.4 ^{+0.2} _{-0.2}	0.7 ^{+0.3} _{-0.3}	568 ⁺¹ ₋₁
B26	1470 ⁺¹⁰ ₋₁₀	110 ⁺²⁰ ₋₁₀	0.096 ^{+0.009} _{-0.009}	57644.431295361(4)	0.8 ^{+0.2} _{-0.2}	0.5 ^{+0.2} _{-0.2}	565.6 ^{+0.7} _{-0.7}

Continued on next page

Table 4.5 – continued from previous page

Burst ^a	μ_f	σ_f	S	μ_t^b	σ_t	τ^c	DM
ID	(MHz)	(MHz)	(Jy ms)	(MJD)	(ms)	(ms)	(pc cm ⁻³)
B27	1340 ⁺⁷⁰ ₋₁₄₀₀	1000 ⁺²⁰⁰⁰ ₋₁₀₀₀	0.17 ^{+0.11} _{-0.09}	57644.43223490(2)	2.0 ^{+1.0} _{-0.7}	1.1 ^{+0.5} _{-0.4}	570 ⁺² ₋₂
B28	1750 ⁺⁴⁰ ₋₂₀	100 ⁺²⁰ ₋₁₀	0.14 ^{+0.03} _{-0.02}	57644.432242722(1)	0.59 ^{+0.05} _{-0.05}	0.16 ^{+0.09} _{-0.05}	561.3 ^{+0.5} _{-0.5}
B29	1680 ⁺¹⁴⁴⁰ ₋₃₀	130 ⁺⁴⁸⁰ ₋₄₀	0.06 ^{+1.5} _{-0.01}	57644.434045197(4)	0.6 ^{+0.2} _{-0.2}	0.4 ^{+0.4} _{-0.2}	562 ⁺¹ ₋₃
B30	1440 ⁺¹⁰ ₋₁₀	100 ⁺²⁰ ₋₁₀	0.12 ^{+0.01} _{-0.01}	57644.43636575(2)	2.1 ^{+0.6} _{-0.5}	1.2 ^{+0.5} _{-0.4}	568 ⁺² ₋₃
B31	1502 ⁺⁶ ₋₆	100 ⁺⁸ ₋₇	0.19 ^{+0.01} _{-0.01}	57644.438794966(4)	1.6 ^{+0.1} _{-0.2}	0.5 ^{+0.1} _{-0.1}	566.6 ^{+0.7} _{-0.7}
B32	1410 ⁺⁸ ₋₈	79 ⁺⁹ ₋₈	0.14 ^{+0.01} _{-0.01}	57644.43884518(1)	1.7 ^{+0.3} _{-0.4}	0.6 ^{+0.4} _{-0.3}	564 ⁺¹ ₋₁
B33	1480 ⁺²⁰ ₋₂₀	120 ⁺²⁰ ₋₂₀	0.08 ^{+0.01} _{-0.01}	57644.439212855(8)	1.2 ^{+0.3} _{-0.4}	0.6 ^{+0.6} _{-0.3}	563.0 ^{+1.0} _{-1.0}
B34†	2600 ⁺⁶⁰⁰ ₋₁₄₀₀	2000 ⁺¹⁰⁰⁰ ₋₁₀₀₀	0.27 ^{+0.1} _{-0.06}	57644.440688613(5)	1.3 ^{+0.4} _{-0.4}	1.0 ^{+0.4} _{-0.3}	567.0 ^{+0.8} _{-0.9}
B35	1707 ⁺⁵ ₋₅	61 ⁺⁶ ₋₅	0.24 ^{+0.01} _{-0.01}	57644.442997729(6)	2.8 ^{+0.2} _{-0.2}	0.7 ^{+0.4} _{-0.2}	575 ⁺³ ₋₃
B36†	3100 ⁺²⁰⁰ ₋₄₀₀	470 ⁺⁷⁰ ₋₉₀	10 ⁺²⁰ ₋₁₀	57644.44358918(2)	3 ⁺² ₋₁	2 ⁺² ₋₁	574 ⁺¹¹ ₋₆
B37	1410 ⁺²⁰ ₋₂₀	130 ⁺³⁰ ₋₂₀	0.065 ^{+0.007} _{-0.008}	57644.443590029(3)	0.58 ^{+0.11} _{-0.09}	0.14 ^{+0.1} _{-0.04}	562.8 ^{+0.5} _{-0.4}
B38	1620 ⁺²⁰ ₋₂₀	110 ⁺²⁰ ₋₂₀	0.1 ^{+0.01} _{-0.01}	57644.445225058(7)	0.9 ^{+0.5} _{-0.5}	1.9 ^{+0.7} _{-0.9}	567 ⁺² ₋₂
B39	1500 ⁺¹⁰ ₋₂₀	130 ⁺¹⁰ ₋₁₀	0.057 ^{+0.01} _{-0.006}	57644.446788124(1)	0.21 ^{+0.04} _{-0.04}	0.12 ^{+0.13} _{-0.05}	560.9 ^{+0.1} _{-0.2}
B40	1590 ⁺¹⁰ ₋₁₀	81 ⁺¹¹ ₋₉	0.12 ^{+0.02} _{-0.02}	57644.447567822(8)	1.1 ^{+0.5} _{-0.7}	2.0 ^{+0.8} _{-0.7}	560 ⁺³ ₋₃
B41*	1379 ⁺¹ ₋₁	31 ⁺¹ ₋₁	0.4 ^{+0.02} _{-0.02}	57644.44772750(2)	0.8 ^{+0.2} _{-0.2}	3.0 ^{+0.2} _{-0.2}	559 ⁺² ₋₂
B42	1670 ⁺¹⁰ ₋₁₀	110 ⁺²⁰ ₋₂₀	0.099 ^{+0.009} _{-0.008}	57644.449915568(3)	0.9 ^{+0.2} _{-0.2}	0.4 ^{+0.2} _{-0.2}	566.0 ^{+0.8} _{-0.7}
B43	1683 ⁺¹² ₋₉	120 ⁺²⁰ ₋₁₀	0.22 ^{+0.01} _{-0.01}	57644.451605444(1)	0.79 ^{+0.1} _{-0.09}	0.9 ^{+0.1} _{-0.1}	568.2 ^{+0.5} _{-0.5}
B44†	3100 ⁺²⁰⁰ ₋₄₀₀	520 ⁺⁵⁰ ₋₇₀	11 ⁺¹³ ₋₇	57644.452389712(6)	1.8 ^{+0.3} _{-0.3}	1.1 ^{+0.5} _{-0.5}	569 ⁺² ₋₄
B45	1710 ⁺¹⁴⁹⁰ ₋₆₀	160 ⁺⁴⁷⁰ ₋₇₀	0.07 ^{+1.81} _{-0.02}	57644.453937473(4)	0.9 ^{+0.2} _{-0.4}	0.5 ^{+0.4} _{-0.2}	562 ⁺¹ ₋₁
B46	1371 ⁺² ₋₂	61 ⁺² ₋₂	0.41 ^{+0.01} _{-0.01}	57644.454477404(8)	2.8 ^{+0.2} _{-0.2}	0.6 ^{+0.1} _{-0.1}	569.6 ^{+0.8} _{-0.8}
B47	900 ⁺⁸⁰⁰ ₋₁₁₀₀	2000 ⁺¹⁰⁰⁰ ₋₂₀₀₀	0.15 ^{+0.04} _{-0.09}	57644.457883227(3)	0.5 ^{+0.1} _{-0.1}	0.23 ^{+0.14} _{-0.09}	561.9 ^{+0.3} _{-0.2}
B48.1†	2900 ⁺³⁰⁰ ₋₁₂₀₀	600 ⁺²⁵⁰⁰ ₋₁₀₀	1 ⁺¹⁵ ₋₁	57644.464507488(7)	0.4 ^{+0.4} _{-0.3}	0.7 ^{+0.6} _{-0.3}	565 ⁺⁴ ₋₂
B48.2†	0 ⁺¹³⁰⁰ ₋₅₀₀	500 ⁺⁵⁰⁰ ₋₄₀₀	0.7 ^{+26.2} _{-0.6}	57644.4645075(2)	0.9 ^{+3.2} _{-0.6}	0.7 ^{+3.6} _{-0.4}	600 ⁺¹⁰⁰ ₋₂₀₀
B49.1	1660 ⁺¹⁰ ₋₁₀	140 ⁺²⁰ ₋₁₀	0.49 ^{+0.03} _{-0.03}	57644.46475893(1)	8.3 ^{+0.5} _{-0.5}	2.0 ^{+0.7} _{-0.6}	568 ⁺⁴ ₋₄
B49.2	2400 ⁺⁷⁰⁰ ₋₁₄₀₀	2000 ⁺¹⁰⁰⁰ ₋₁₀₀₀	0.16 ^{+0.05} _{-0.04}	57644.464759903(1)	0.2 ^{+0.1} _{-0.1}	0.6 ^{+0.2} _{-0.1}	561.3 ^{+0.3} _{-0.2}
B50	1460 ⁺¹⁰ ₋₁₀	90 ⁺¹⁰ ₋₁₀	0.13 ^{+0.02} _{-0.01}	57644.46476275(1)	1.3 ^{+1.0} _{-0.9}	1.7 ^{+0.6} _{-0.8}	561 ⁺² ₋₂
B51†	2500 ⁺⁶⁰⁰ ₋₉₀₀	1000 ⁺¹⁰⁰⁰ ₋₇₀₀	0.3 ^{+0.2} _{-0.2}	57644.465729923(7)	1.4 ^{+0.6} _{-0.4}	1.2 ^{+0.5} _{-0.4}	569 ⁺¹ ₋₁
B52	1667 ⁺⁵ ₋₄	73 ⁺⁶ ₋₅	0.185 ^{+0.009} _{-0.009}	57644.466222289(4)	1.4 ^{+0.1} _{-0.2}	0.6 ^{+0.2} _{-0.2}	570 ⁺¹ ₋₁
B53	1720 ⁺³⁰ ₋₂₀	120 ⁺³⁰ ₋₂₀	0.19 ^{+0.03} _{-0.02}	57644.468095365(4)	1.5 ^{+0.2} _{-0.2}	1.1 ^{+0.3} _{-0.3}	572 ⁺¹ ₋₁
B54*	1600 ⁺⁶⁰⁰ ₋₆₀₀	100 ⁺¹⁰⁰ ₋₁₀₀	0.0 ^{+0.1} _{-0.1}	57644.47117767(7)	0 ⁺¹⁰ ₋₁₀	0.9 ^{+0.2} _{-0.2}	560 ⁺⁴⁰ ₋₄₀
B55	1650 ⁺¹⁰ ₋₁₀	100 ⁺¹⁰ ₋₁₀	0.13 ^{+0.01} _{-0.01}	57644.474717918(6)	1.4 ^{+0.3} _{-0.2}	1.3 ^{+0.5} _{-0.5}	570 ⁺² ₋₂

Continued on next page

Table 4.5 – continued from previous page

Burst ^a	μ_f	σ_f	S	μ_t^b	σ_t	τ^c	DM
ID	(MHz)	(MHz)	(Jy ms)	(MJD)	(ms)	(ms)	(pc cm ⁻³)
B56	1706 ⁺⁴ ₋₄	54 ⁺⁶ ₋₅	0.16 ^{+0.01} _{-0.01}	57644.477082041(3)	0.9 ^{+0.2} _{-0.1}	0.9 ^{+0.2} _{-0.2}	560 ⁺² ₋₂
B57	1530 ⁺⁶ ₋₆	186 ⁺⁶ ₋₆	0.254 ^{+0.006} _{-0.005}	57645.4110889611(4)	0.48 ^{+0.03} _{-0.02}	0.21 ^{+0.02} _{-0.02}	561.75 ^{+0.06} _{-0.06}
B58	1462 ⁺⁶ ₋₆	78 ⁺⁶ ₋₅	0.129 ^{+0.008} _{-0.008}	57645.411651653(4)	0.9 ^{+0.1} _{-0.1}	0.4 ^{+0.1} _{-0.1}	564.4 ^{+0.6} _{-0.7}
B59	1537 ⁺⁵ ₋₄	68 ⁺⁴ ₋₄	0.2 ^{+0.01} _{-0.01}	57645.411889086(9)	2.2 ^{+0.4} _{-0.3}	1.1 ^{+0.3} _{-0.3}	568 ⁺² ₋₂
B60.1	3000 ⁺³⁰⁰ ₋₁₀₀₀	440 ⁺⁸⁰ ₋₂₂₀	4 ⁺⁶ ₋₄	57645.412281872(4)	0.8 ^{+0.2} _{-0.1}	0.3 ^{+0.3} _{-0.2}	566 ⁺² ₋₂
B60.2	2300 ⁺⁹⁰⁰ ₋₆₀₀	500 ⁺³⁰⁰ ₋₃₀₀	0.4 ^{+1.6} _{-0.2}	57645.412281990(6)	0.9 ^{+0.6} _{-0.5}	2.5 ^{+0.6} _{-0.5}	561 ⁺² ₋₂
B61†	1000 ⁺¹⁰⁰⁰ ₋₁₀₀₀	2800 ⁺⁹⁰⁰ ₋₁₁₀₀	0.25 ^{+0.08} _{-0.07}	57645.41286869(2)	2.0 ^{+2.0} _{-1.0}	2 ⁺² ₋₁	580 ⁺³ ₋₁₃
B62	1570 ⁺¹⁰ ₋₁₀	110 ⁺¹⁰ ₋₁₀	0.12 ^{+0.01} _{-0.009}	57645.413644740(5)	1.3 ^{+0.2} _{-0.2}	0.6 ^{+0.3} _{-0.2}	567 ⁺¹ ₋₁
B63	1280 ⁺¹⁰ ₋₁₀	86 ⁺¹¹ ₋₉	0.1 ^{+0.01} _{-0.01}	57645.41609534(1)	0.8 ^{+0.2} _{-0.2}	0.6 ^{+0.2} _{-0.1}	562.0 ^{+0.8} _{-0.9}
B64†	2000 ⁺¹⁰⁰⁰ ₋₁₀₀₀	3000 ⁺¹⁰⁰⁰ ₋₁₀₀₀	0.32 ^{+0.09} _{-0.09}	57645.41639521(1)	1.4 ^{+0.5} _{-0.5}	3 ⁺¹ ₋₁	564 ⁺³ ₋₂
B65	1550 ⁺¹⁰ ₋₁₀	110 ⁺²⁰ ₋₁₀	0.091 ^{+0.009} _{-0.009}	57645.416564818(6)	1.2 ^{+0.2} _{-0.2}	0.6 ^{+0.3} _{-0.2}	566.0 ^{+1.0} _{-1.0}
B66	1440 ⁺¹⁰ ₋₁₀	120 ⁺¹⁰ ₋₁₀	0.105 ^{+0.009} _{-0.008}	57645.417467306(6)	1.6 ^{+0.2} _{-0.2}	-	567.7 ^{+0.7} _{-0.7}
B67	1640 ⁺³⁰ ₋₂₀	180 ⁺³⁰ ₋₃₀	0.099 ^{+0.009} _{-0.008}	57645.417897463(2)	0.8 ^{+0.1} _{-0.1}	0.3 ^{+0.1} _{-0.1}	564.8 ^{+0.5} _{-0.5}
B68†	2000 ⁺¹⁰⁰⁰ ₋₁₀₀₀	2000 ⁺¹⁰⁰⁰ ₋₁₀₀₀	0.2 ^{+0.05} _{-0.06}	57645.419005902(8)	1.0 ^{+0.6} _{-0.5}	0.9 ^{+0.4} _{-0.4}	567.9 ^{+0.9} _{-1.3}
B69	1450 ⁺²⁰ ₋₁₀	80 ⁺¹⁰ ₋₁₀	0.059 ^{+0.009} _{-0.009}	57645.41920745(1)	0.7 ^{+0.3} _{-0.3}	0.5 ^{+0.3} _{-0.2}	564 ⁺¹ ₋₁
B70†	3100 ⁺²⁰⁰ ₋₅₀₀	600 ⁺¹⁰⁰ ₋₁₀₀	0.9 ^{+1.7} _{-0.6}	57645.419896226(4)	0.8 ^{+0.3} _{-0.3}	0.8 ^{+0.4} _{-0.3}	564 ⁺¹ ₋₁
B71	1378 ⁺⁵ ₋₄	78 ⁺⁵ ₋₄	0.34 ^{+0.02} _{-0.02}	57645.420265931(9)	1.0 ^{+0.2} _{-0.2}	2.6 ^{+0.2} _{-0.2}	561.1 ^{+0.9} _{-1.0}
B72†	3100 ⁺²⁰⁰ ₋₅₀₀	700 ⁺²⁰⁰ ₋₁₀₀	1.2 ^{+3.4} _{-0.8}	57645.420600439(9)	1.2 ^{+0.7} _{-0.8}	2 ⁺¹ ₋₁	559 ⁺² ₋₂
B73†	3100 ⁺²⁰⁰ ₋₅₀₀	600 ⁺¹⁰⁰ ₋₁₀₀	2 ⁺³ ₋₁	57645.420679524(7)	1.5 ^{+0.6} _{-0.8}	1.7 ^{+0.8} _{-0.7}	575 ⁺³ ₋₃
B74†	3100 ⁺²⁰⁰ ₋₄₀₀	520 ⁺¹⁰⁰ ₋₉₀	5 ⁺¹² ₋₄	57645.420752363(6)	1.2 ^{+0.8} _{-0.6}	1.9 ^{+0.7} _{-0.9}	551 ⁺¹¹ ₋₂
B75†	2200 ⁺⁸⁰⁰ ₋₁₄₀₀	2700 ⁺⁹⁰⁰ ₋₁₂₀₀	0.39 ^{+0.07} _{-0.08}	57645.421284056(4)	0.9 ^{+0.3} _{-0.3}	1.8 ^{+0.3} _{-0.3}	567.6 ^{+0.9} _{-0.7}
B76	1410 ⁺¹⁰ ₋₃₅₀	90 ⁺²⁰ ₋₁₀	0.05 ^{+0.111} _{-0.006}	57645.421290699(5)	0.42 ^{+0.25} _{-0.09}	0.14 ^{+0.42} _{-0.05}	561.6 ^{+0.6} _{-0.5}
B77	1312 ⁺⁵ ₋₅	51 ⁺⁶ ₋₆	0.12 ^{+0.01} _{-0.01}	57645.42161300(2)	1.3 ^{+0.5} _{-0.5}	0.7 ^{+0.2} _{-0.2}	566 ⁺² ₋₂
B78	1170 ⁺²⁰ ₋₄₀	80 ⁺⁴⁰ ₋₁₀	0.11 ^{+0.03} _{-0.02}	57645.42184846(1)	0.9 ^{+0.3} _{-0.2}	0.28 ^{+0.14} _{-0.08}	562.4 ^{+0.7} _{-0.9}
B79	1702 ⁺⁴ ₋₄	57 ⁺⁶ ₋₅	0.18 ^{+0.01} _{-0.01}	57645.421872443(3)	1.2 ^{+0.2} _{-0.2}	1.7 ^{+0.3} _{-0.3}	568 ⁺² ₋₂
B80	1601 ⁺⁷ ₋₈	81 ⁺⁸ ₋₇	0.13 ^{+0.01} _{-0.01}	57645.422454976(3)	0.5 ^{+0.3} _{-0.2}	1.3 ^{+0.2} _{-0.3}	564.4 ^{+1.0} _{-0.8}
B81	1720 ⁺³⁹⁰ ₋₁₀	60 ⁺³¹⁰ ₋₂₀	0.05 ^{+0.1} _{-0.01}	57645.423009490(6)	0.78 ^{+0.08} _{-0.18}	0.6 ^{+0.1} _{-0.2}	563 ⁺⁸ ₋₅
B82†	3100 ⁺²⁰⁰ ₋₄₀₀	580 ⁺⁶⁰ ₋₇₀	4 ⁺⁴ ₋₂	57645.424145870(2)	1.3 ^{+0.1} _{-0.1}	-	569.2 ^{+0.7} _{-0.7}
B83†	2400 ⁺⁷⁰⁰ ₋₁₃₀₀	2000 ⁺²⁰⁰⁰ ₋₁₀₀₀	0.15 ^{+0.07} _{-0.04}	57645.424617988(6)	0.7 ^{+0.3} _{-0.3}	0.4 ^{+0.2} _{-0.2}	564.7 ^{+0.7} _{-0.5}
B84†	3000 ⁺³⁰⁰ ₋₇₀₀	600 ⁺¹⁰⁰ ₋₂₀₀	1 ⁺³ ₋₁	57645.425478263(8)	1.0 ^{+0.8} _{-0.6}	1.7 ^{+0.9} _{-0.9}	563 ⁺² ₋₃
B85*	1310 ⁺²⁰ ₋₂₀	130 ⁺²⁰ ₋₂₀	0.11 ^{+0.02} _{-0.02}	57645.426746546(2)	0.01 ^{+0.06} _{-0.06}	2.0 ^{+0.4} _{-0.4}	566.1 ^{+0.2} _{-0.2}

Continued on next page

Table 4.5 – continued from previous page

Burst ^a	μ_f	σ_f	S	μ_t^b	σ_t	τ^c	DM
ID	(MHz)	(MHz)	(Jy ms)	(MJD)	(ms)	(ms)	(pc cm ⁻³)
B86	1550 ⁺¹⁰ ₋₁₀	80 ⁺¹⁰ ₋₁₀	0.067 ^{+0.009} _{-0.008}	57645.426792398(7)	0.6 ^{+0.4} _{-0.3}	0.9 ^{+0.3} _{-0.3}	563 ⁺¹ ₋₁
B87	1760 ⁺⁴⁴⁰ ₋₅₀	190 ⁺²⁰⁰ ₋₅₀	0.2 ^{+0.38} _{-0.03}	57645.426922532(4)	1.4 ^{+0.2} _{-0.2}	0.9 ^{+0.3} _{-0.3}	571.1 ^{+0.9} _{-0.9}
B88.1	2000 ⁺¹⁰⁰⁰ ₋₁₀₀₀	2000 ⁺¹⁰⁰⁰ ₋₂₀₀₀	0.17 ^{+0.06} _{-0.06}	57645.427338932(7)	0.7 ^{+0.4} _{-0.3}	1.0 ^{+0.7} _{-0.5}	561 ⁺² ₋₄
B88.2	1460 ⁺²⁰ ₋₁₀	130 ⁺²⁰ ₋₂₀	0.11 ^{+0.01} _{-0.01}	57645.427339178(9)	2.4 ^{+0.4} _{-0.3}	-	561 ⁺¹ ₋₂
B89	1479 ⁺⁸ ₋₈	120 ⁺⁸ ₋₇	0.18 ^{+0.01} _{-0.01}	57645.428904793(6)	2.4 ^{+0.2} _{-0.2}	-	569.0 ^{+0.7} _{-0.8}
B90	1420 ⁺⁴⁰ ₋₅₀₀	120 ⁺²⁷⁹⁰ ₋₃₀	0.09 ^{+0.11} _{-0.02}	57645.42990044(2)	2.4 ^{+0.6} _{-0.6}	0.8 ^{+0.6} _{-0.3}	565 ⁺² ₋₂
B91	1476 ⁺⁹ ₋₉	81 ⁺¹⁰ ₋₈	0.12 ^{+0.01} _{-0.01}	57645.43034452(2)	3.1 ^{+0.4} _{-0.4}	-	572 ⁺² ₋₂
B92	1721 ⁺⁴ ₋₄	39 ⁺⁶ ₋₄	0.17 ^{+0.02} _{-0.02}	57645.43044763(1)	2.6 ^{+0.8} _{-0.7}	3 ⁺¹ ₋₁	580 ⁺¹⁰ ₋₁₀
B93.1	1591 ⁺⁹ ₋₉	106 ⁺¹⁰ ₋₉	0.126 ^{+0.008} _{-0.008}	57645.430622487(3)	1.2 ^{+0.1} _{-0.2}	0.4 ^{+0.2} _{-0.2}	565.3 ^{+0.8} _{-0.9}
B93.2	1620 ⁺¹⁰ ₋₁₀	120 ⁺²⁰ ₋₂₀	0.13 ^{+0.01} _{-0.01}	57645.430622591(7)	1.7 ^{+0.3} _{-0.3}	0.8 ^{+0.4} _{-0.3}	564 ⁺² ₋₂
B94†	3100 ⁺²⁰⁰ ₋₆₀₀	500 ⁺⁶⁰ ₋₁₂₀	7 ⁺¹⁰ ₋₆	57645.431306274(4)	1.1 ^{+0.2} _{-0.2}	0.6 ^{+0.4} _{-0.3}	565 ⁺² ₋₄
B95	1693 ⁺¹¹ ₋₈	90 ⁺²⁰ ₋₁₀	0.095 ^{+0.007} _{-0.007}	57645.431478373(2)	0.69 ^{+0.08} _{-0.08}	0.22 ^{+0.12} _{-0.08}	563.5 ^{+0.6} _{-0.5}
B96†	2000 ⁺¹⁰⁰⁰ ₋₂₀₀₀	2800 ⁺⁸⁰⁰ ₋₁₁₀₀	0.2 ^{+0.06} _{-0.05}	57645.43254531(2)	2.1 ^{+0.8} _{-0.9}	0.9 ^{+0.6} _{-0.3}	565 ⁺² ₋₁
B97†	2900 ⁺⁴⁰⁰ ₋₈₀₀	800 ⁺⁵⁰⁰ ₋₂₀₀	0.3 ^{+0.4} _{-0.1}	57645.433086695(3)	0.6 ^{+0.4} _{-0.4}	1.0 ^{+0.4} _{-0.4}	562.4 ^{+0.9} _{-0.5}
B98	1667 ⁺⁷ ₋₇	67 ⁺⁸ ₋₇	0.094 ^{+0.01} _{-0.009}	57645.434330453(5)	1.1 ^{+0.2} _{-0.2}	0.7 ^{+0.3} _{-0.3}	559 ⁺² ₋₂
B99†	2000 ⁺¹⁰⁰⁰ ₋₂₀₀₀	2800 ⁺⁸⁰⁰ ₋₁₁₀₀	0.13 ^{+0.03} _{-0.03}	57645.440066847(4)	0.8 ^{+0.2} _{-0.2}	0.3 ^{+0.2} _{-0.1}	565.8 ^{+0.4} _{-0.4}
B100	1688 ⁺⁴ ₋₄	62 ⁺⁵ ₋₅	0.154 ^{+0.008} _{-0.007}	57645.440814641(2)	1.0 ^{+0.1} _{-0.1}	0.5 ^{+0.2} _{-0.2}	567.0 ^{+1.0} _{-1.0}
B101*	1510 ⁺²⁰ ₋₂₀	130 ⁺²⁰ ₋₂₀	0.042 ^{+0.006} _{-0.006}	57645.441999925(3)	0.87 ^{+0.02} _{-0.02}	-	568.6 ^{+0.4} _{-0.4}
B102	2000 ⁺¹²⁰⁰ ₋₃₀₀	300 ⁺³⁰⁰ ₋₂₀₀	0.2 ^{+2.7} _{-0.1}	57645.442100164(8)	2.1 ^{+0.4} _{-0.4}	0.8 ^{+0.6} _{-0.4}	564 ⁺² ₋₂
B103	1470 ⁺¹⁰ ₋₁₀	90 ⁺¹⁰ ₋₁₀	0.08 ^{+0.01} _{-0.01}	57645.44263325(1)	1.4 ^{+0.3} _{-0.3}	0.6 ^{+0.5} _{-0.3}	562 ⁺¹ ₋₁
B104*	1460 ⁺⁶⁰ ₋₆₀	90 ⁺²⁰ ₋₂₀	0.07 ^{+0.09} _{-0.09}	57645.4427413(1)	0 ⁺² ₋₂	2.1 ^{+0.5} _{-0.5}	569 ⁺⁸ ₋₈
B105	1670 ⁺¹⁰ ₋₁₀	110 ⁺²⁰ ₋₂₀	0.095 ^{+0.008} _{-0.007}	57645.444480955(2)	0.51 ^{+0.1} _{-0.09}	0.5 ^{+0.1} _{-0.1}	565.4 ^{+0.5} _{-0.5}
B106	1480 ⁺⁷ ₋₈	97 ⁺¹⁰ ₋₉	0.16 ^{+0.01} _{-0.01}	57645.444919472(7)	2.0 ^{+0.2} _{-0.2}	0.5 ^{+0.2} _{-0.1}	566 ⁺¹ ₋₁
B107	2800 ⁺³⁰⁰ ₋₁₀₀₀	370 ⁺⁷⁰ ₋₂₆₀	10 ⁺¹⁰ ₋₁₀	57645.445443070(5)	1.2 ^{+0.3} _{-0.3}	0.9 ^{+0.4} _{-0.4}	561 ⁺³ ₋₂
B108	1488 ⁺⁴ ₋₄	83 ⁺⁴ ₋₄	0.141 ^{+0.006} _{-0.006}	57645.448802883(2)	0.66 ^{+0.03} _{-0.03}	-	561.1 ^{+0.3} _{-0.3}
B109	1460 ⁺⁵ ₋₅	87 ⁺⁵ ₋₅	0.24 ^{+0.01} _{-0.01}	57645.449987035(8)	2.8 ^{+0.2} _{-0.2}	-	568 ⁺¹ ₋₁
B110*	1620 ⁺¹⁰ ₋₁₀	100 ⁺¹⁰ ₋₁₀	0.1 ^{+0.01} _{-0.01}	57645.451201066(8)	1.7 ^{+0.4} _{-0.4}	1.3 ^{+0.5} _{-0.5}	561 ⁺² ₋₂
B111	1469 ⁺⁶ ₋₆	58 ⁺⁷ ₋₆	0.109 ^{+0.01} _{-0.009}	57645.45198993(1)	1.6 ^{+0.2} _{-0.2}	0.5 ^{+0.2} _{-0.1}	563 ⁺² ₋₂
B112	1442 ⁺⁴ ₋₅	68 ⁺⁴ ₋₄	0.142 ^{+0.008} _{-0.008}	57645.453426198(4)	0.59 ^{+0.08} _{-0.07}	0.4 ^{+0.07} _{-0.06}	563.1 ^{+0.6} _{-0.6}
B113	1699 ⁺⁴ ₋₄	68 ⁺⁶ ₋₅	0.29 ^{+0.01} _{-0.01}	57645.453639067(4)	2.5 ^{+0.1} _{-0.2}	0.6 ^{+0.3} _{-0.2}	569 ⁺² ₋₂
B114	1570 ⁺²⁰ ₋₂₀	110 ⁺²⁰ ₋₂₀	0.07 ^{+0.01} _{-0.01}	57645.453640216(3)	0.3 ^{+0.2} _{-0.1}	1.4 ^{+0.3} _{-0.3}	566.5 ^{+0.7} _{-0.8}

Continued on next page

Table 4.5 – continued from previous page

Burst ^a	μ_f	σ_f	S	μ_t^b	σ_t	τ^c	DM
ID	(MHz)	(MHz)	(Jy ms)	(MJD)	(ms)	(ms)	(pc cm ⁻³)
B115	1730 ⁺³⁰ ₋₁₀	90 ⁺²⁰ ₋₂₀	0.19 ^{+0.03} _{-0.02}	57645.454258102(7)	2.3 ^{+0.3} _{-0.3}	0.9 ^{+0.6} _{-0.4}	569 ⁺³ ₋₃
B116*	1480 ⁺⁴⁰ ₋₄₀	260 ⁺⁶⁰ ₋₆₀	0.08 ^{+0.02} _{-0.02}	57645.4544929400(9)	0.05 ^{+0.07} _{-0.07}	1.4 ^{+0.3} _{-0.3}	565.7 ^{+0.2} _{-0.2}
B117	1580 ⁺²⁰ ₋₂₀	130 ⁺³⁰ ₋₂₀	0.1 ^{+0.02} _{-0.01}	57645.45736147(1)	1.8 ^{+0.9} _{-0.8}	1.8 ^{+0.8} _{-0.8}	563 ⁺³ ₋₃
B118	1580 ⁺³⁰ ₋₂₀	130 ⁺³⁰ ₋₂₀	0.1 ^{+0.02} _{-0.01}	57645.45736147(1)	1.7 ^{+0.9} _{-0.8}	1.8 ^{+0.9} _{-0.8}	563 ⁺³ ₋₃
B119	1670 ⁺⁶⁰ ₋₂₀	170 ⁺⁶⁰ ₋₃₀	0.21 ^{+0.04} _{-0.02}	57645.458536220(6)	1.2 ^{+0.4} _{-0.3}	2.6 ^{+0.5} _{-0.5}	569 ⁺² ₋₂
B120	1690 ⁺¹⁰⁴⁰ ₋₃₀	130 ⁺³⁸⁰ ₋₃₀	0.09 ^{+0.87} _{-0.01}	57645.459103357(6)	0.8 ^{+0.3} _{-0.5}	1.2 ^{+0.5} _{-0.4}	566 ⁺² ₋₂
B121.1	1612 ⁺² ₋₂	90 ⁺² ₋₂	0.394 ^{+0.007} _{-0.007}	57645.460053270(1)	0.82 ^{+0.04} _{-0.04}	0.25 ^{+0.06} _{-0.07}	562.1 ^{+0.2} _{-0.2}
B121.2	1321 ⁺² ₋₃	178 ⁺² ₋₂	1.73 ^{+0.02} _{-0.02}	57645.4600533283(4)	0.95 ^{+0.02} _{-0.02}	0.36 ^{+0.02} _{-0.01}	561.98 ^{+0.03} _{-0.03}
B121.3	1181 ⁺⁹ ₋₁₁	93 ⁺⁹ ₋₈	0.35 ^{+0.03} _{-0.02}	57645.460053409(4)	1.22 ^{+0.09} _{-0.07}	-	561.6 ^{+0.2} _{-0.3}
B122	1405 ⁺⁴ ₋₄	71 ⁺⁵ ₋₅	0.22 ^{+0.01} _{-0.01}	57645.462106655(9)	2.3 ^{+0.2} _{-0.2}	0.5 ^{+0.2} _{-0.1}	565 ⁺¹ ₋₁
B123*	1690 ⁺²⁰ ₋₂₀	90 ⁺²⁰ ₋₂₀	0.09 ^{+0.02} _{-0.02}	57645.462556118(6)	1.1 ^{+0.4} _{-0.4}	2.0 ^{+0.6} _{-0.6}	567 ⁺³ ₋₃
B124	1560 ⁺¹⁰ ₋₂₀	150 ⁺²⁰ ₋₂₀	0.094 ^{+0.008} _{-0.008}	57645.464052738(3)	1.0 ^{+0.2} _{-0.2}	0.5 ^{+0.2} _{-0.2}	565.3 ^{+0.5} _{-0.6}
B125	1600 ⁺¹⁰ ₋₁₀	140 ⁺²⁰ ₋₂₀	0.121 ^{+0.01} _{-0.009}	57645.464188555(4)	1.2 ^{+0.2} _{-0.2}	0.7 ^{+0.2} _{-0.2}	564.4 ^{+0.8} _{-0.8}
B126	1600 ⁺¹¹⁹⁰ ₋₆₀	230 ⁺²³⁸⁰ ₋₉₀	0.07 ^{+0.13} _{-0.02}	57645.464503855(6)	0.9 ^{+0.4} _{-0.5}	0.9 ^{+0.6} _{-0.4}	567 ⁺² ₋₂
B127	2000 ⁺¹⁰⁰⁰ ₋₁₀₀₀	2000 ⁺¹⁰⁰⁰ ₋₂₀₀₀	0.19 ^{+0.06} _{-0.11}	57645.465133243(7)	0.8 ^{+0.4} _{-0.4}	1.2 ^{+0.6} _{-0.5}	566.0 ^{+1.9} _{-0.7}
B128†	1000 ⁺¹⁰⁰⁰ ₋₁₀₀₀	2700 ⁺⁹⁰⁰ ₋₁₅₀₀	0.25 ^{+0.05} _{-0.09}	57645.466100959(6)	1.1 ^{+0.2} _{-0.2}	0.9 ^{+0.3} _{-0.3}	567.0 ^{+0.7} _{-0.9}
B129	1710 ⁺¹³⁸⁰ ₋₂₀	90 ⁺⁴⁵⁰ ₋₃₀	0.11 ^{+4.56} _{-0.02}	57645.466379997(5)	1.0 ^{+0.3} _{-0.3}	1.3 ^{+0.4} _{-0.4}	573 ⁺² ₋₃
B130†	2000 ⁺¹⁰⁰⁰ ₋₁₀₀₀	3000 ⁺¹⁰⁰⁰ ₋₁₀₀₀	0.18 ^{+0.05} _{-0.05}	57645.46833947(1)	1.4 ^{+0.4} _{-0.4}	0.6 ^{+0.4} _{-0.2}	565.5 ^{+1.0} _{-0.8}
B131	1460 ⁺¹⁰ ₋₁₀	90 ⁺¹⁰ ₋₁₀	0.08 ^{+0.01} _{-0.01}	57645.47236212(1)	1.1 ^{+0.4} _{-0.3}	0.7 ^{+0.3} _{-0.3}	564 ⁺¹ ₋₂
B132	1720 ⁺⁸⁰ ₋₂₀	100 ⁺⁷⁰ ₋₂₀	0.09 ^{+0.03} _{-0.01}	57645.474138540(3)	0.7 ^{+0.1} _{-0.1}	0.4 ^{+0.2} _{-0.2}	561 ⁺¹ ₋₁
B133	1740 ⁺¹⁸⁰ ₋₅₀	230 ⁺¹¹⁰ ₋₄₀	0.22 ^{+0.1} _{-0.03}	57645.474448229(4)	1.7 ^{+0.2} _{-0.2}	0.7 ^{+0.3} _{-0.3}	570.6 ^{+0.7} _{-0.7}

1 σ errors on the fits are shown on superscript and subscript of each value in the table.

For μ_t , the error on the last significant digit is shown in parenthesis.

^a Burst IDs are chronological. Individual component number (N) for multi-component bursts are appended to the burst IDs. Bursts modeled only using `curve_fit` are marked with *. Note that the errors on these bursts could be unreliable and may be either under or over-estimated. Bursts that extend beyond the observable bandwidth can also have unreliable estimates of spectra parameters and fluence (see Section 4.4.5.7). We mark those bursts with † to indicate that their fluence and spectra parameters could be unconstrained.

^b μ_t is the mean of the pulse profile in units of MJD. This can be considered as the arrival time of the pulse. It is referenced to the solar system barycenter, after correcting to infinite frequency using a DM of 560.5 pc cm⁻³.

^c τ is referred to 1 GHz.

Chapter 5

Observational effects of banded repeating FRBs

5.1 Chapter Overview

Recent observations have shown that repeating Fast Radio Bursts (FRBs) exhibit band-limited emission, whose frequency-dependent amplitude can be modeled using a Gaussian function. In this chapter, we show that banded emission of FRBs can lead to incompleteness across the observing band. This biases the detected sample of bursts and can explain the various shapes of cumulative energy distributions seen for repeating FRBs. We assume a Gaussian shape of the burst spectra and used simulations to demonstrate the above bias using an FRB 121102-like example. We recovered energy distributions that showed a break in power-law and flattening of power-law at low energies, based on the fluence threshold of the observations. We provide recommendations for single-pulse searches and analysis of repeating FRBs to account for this incompleteness. Primarily, we recommend that burst spectra should be modeled to estimate the intrinsic fluence and bandwidth of the burst robustly. Also, bursts that lie mainly within the observing band should be used for analyses of energy distributions. We show that the bimodality reported in the distribution of energies of FRB 121102 by Li et al. (2021) disappears when burst bandwidth, instead of the center frequency of the observation, is used to estimate energy. Sub-banded searches will also aid in detecting band-limited bursts. All the analysis scripts used

Published as Aggarwal (2021)

in this chapter are available in a Github repository¹.

5.2 Introduction

Fast radio bursts (FRBs) show a wide variety of properties (see Chapter 1). It is an ongoing effort to disentangle the properties arising from the FRB source itself (i.e., intrinsic) to those introduced due to various selection effects. Two properties associated primarily with repeating FRBs are sub-burst drifting and band-limited emission (Hessels et al., 2019; Law et al., 2017). In this chapter, the effect we focus on is the band-limited emission of repeaters and discuss various observational biases caused by it.

The source of the band-limited nature of FRB emission is currently not understood. However, there exist multiple progenitor and propagation models that try to explain the band-limited nature of FRBs (Simard & Ravi, 2020; Beniamini & Kumar, 2020; Metzger et al., 2019; Lu & Kumar, 2018; Cordes et al., 2017). Further, it has been reported that the emission of the first repeating FRB, FRB 121102 favors 1600 MHz, and there is a lack of emission observed below 1200 MHz (Aggarwal et al., 2021a; Platts et al., 2021). It is unclear if the emission behavior changes below that frequency and is below the detection threshold or is just not present (Platts et al., 2021). Similar banded² emission has also been reported for other repeating FRBs (Pastor-Marazuela et al., 2021; Li et al., 2021; Aggarwal et al., 2021a). Further, it has been reported that the peak emission frequency appears to be random. Also, ultra-wideband observations have shown that there is no evidence for oscillations in the spectra, i.e the burst emission is present only within the narrow envelop and is not simultaneously present at any other frequency (Kumar et al., 2021c; Law et al., 2017). The periodicity in burst activity and its frequency dependence further complicates the interpretation of this effect (Pastor-Marazuela et al., 2021; Aggarwal et al., 2020).

¹https://github.com/KshitijAggarwal/banded_repeater_analysis

²We use “band-limited” and “banded” emission interchangeably across this chapter. Both of these refer to the finite bandwidth emission of the bursts.

Energy distribution of repeating FRBs also show a variety of shapes and features like: a simple power-law, broken power-law, smooth flattening of power-law at low energies, etc (Cruces et al., 2020; Li et al., 2021; Aggarwal et al., 2021a; Pastor-Marazuela et al., 2021). The shape of the energy distribution can provide important information about the emission mechanism of the FRB source. Giant pulses from neutron stars and high energy emission from magnetars has been described using a power-law distribution (Bera & Chengalur, 2019; Cheng et al., 2020). A similar power-law index for FRBs with that seen from neutron stars might imply a common origin. In some studies of repeaters, the deviation of energy distribution from a simple power-law has been attributed to the intrinsic emission process of the FRB. But, in this chapter, we demonstrate that many of the observed shapes of the energy distributions, can be attributed to biases due to the band-limited nature of emission of repeaters. It is therefore necessary to account for these biases, before making conclusions about the intrinsic nature of FRB emission.

We start by discussing our modeling methods in §2, followed by results demonstrated using an example in §3, and discuss the implications in §4. We then conclude with some recommendations in §5.

5.3 Methods

This section briefly discusses the methods we used to model the FRB spectra and various observational effects. We follow a two-step approach: 1) Simulate a population of FRBs with spectral properties described by predefined functions, 2) Estimate the detectable FRBs from this sample based on a sensitivity threshold.

5.3.1 Generating a population of repeater bursts

Previous studies have shown that the spectra of repeating FRBs can be modeled using a Gaussian function (Pleunis et al., 2021b; Aggarwal et al., 2021a; Kumar et al., 2021c; Law et al., 2017). Therefore, we assume that the repeater burst’s spectra follow a Gaussian function, parameterized by a mean (μ_f) and a standard deviation

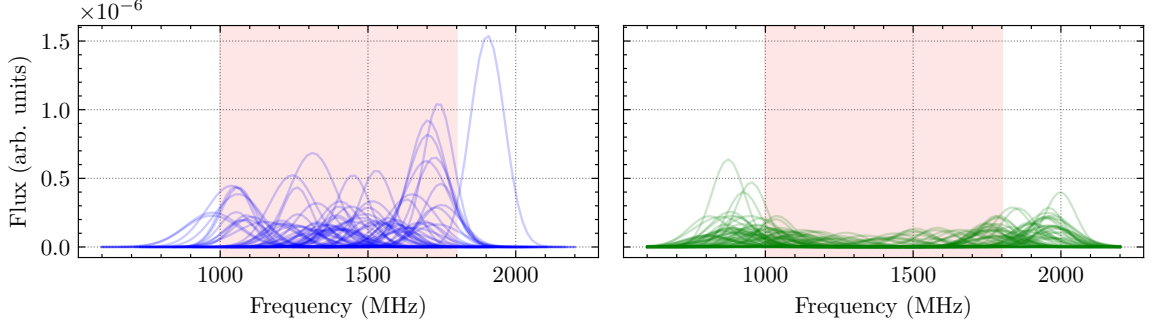


Figure 5.1: Figure showing some simulated spectra modeled using a Gaussian function. The mean and standard deviation has been sampled from a Uniform distribution. The energy distribution is assumed to be a power-law with a slope of -1.8. The left panel shows the detected spectra, while the right panel shows the ones that weren't detected, using a constant fluence threshold. The observing band is shown in red. Spectra that had enough energy within the band were detected, while the ones without enough signal within the observing band were not detected. See Section 5.3.2 for more details.

(σ_f). We also assume that μ_f and σ_f itself follow a predefined distribution (Gaussian or Uniform). Finally, we assume that the cumulative energy distribution follows a power-law with a slope (α). We then draw 50,000 samples from the above three distributions each to represent 50,000 bursts from a repeater. We convert the energy into an intrinsic fluence using its spectral bandwidth and assuming a nominal distance to the source.

5.3.2 Applying selection effects

Next, we aim to determine the bursts that a given observational system will detect. To simplify the analysis, we ignore the effect of any signal lost due to non-idealized dispersion measure (DM) and boxcar search step (see Aggarwal et al., 2021a,c; Agarwal et al., 2020b; Keane & Petroff, 2015, on details of those effects), and only consider two selection effects: observational bandwidth and fluence limit. Due to the limited observing bandwidth of most instruments and the distribution of burst spectra across a wide frequency range (see Figure 5.1), the observed fluence of the burst will depend on the fraction of the burst spectra that lies within the observational band. Therefore, for each simulated burst, we estimate its observed fluence

by integrating the spectra (using the bursts' μ_f and σ_f) within the limits of observing bandwidth. Therefore, assuming Gaussian spectra of the form,

$$\mathcal{G}(\nu; \mu_\nu, \sigma_\nu) = \frac{1}{\sigma_\nu \sqrt{2\pi}} \exp\left(-\frac{1}{2} \frac{(\nu - \mu_\nu)^2}{\sigma_\nu^2}\right), \quad (5.1)$$

where ν is the observing frequency. The observed fluence is given by:

$$S_{obs} = S_{int} \int_{\nu_{start}}^{\nu_{end}} \mathcal{G}(\nu; \mu_\nu, \sigma_\nu) d\nu, \quad (5.2)$$

where S_{int} and S_{obs} are the intrinsic and observed burst fluence, ν_{start} and ν_{end} are the start and end frequency of the observing band. If the burst spectra lie primarily within the bandwidth, then the integral in the above equation will be close to one, and so the observed fluence will be very similar to the intrinsic fluence. As the fraction of burst signal outside the band increases, the observed fluence will get lesser as compared to intrinsic fluence.

If this observed fluence is greater than the fluence threshold of the search pipeline, then the burst will be detected. Therefore, out of the population of bursts that do not lie primarily within our observing band, we are sensitive to detect only the bright ones (as illustrated in Figure 5.1). This will introduce strong selection effects in the properties of the detected bursts.

5.3.3 Estimating Fluence and bandwidth

Typically, the detected bursts' fluence (and then energy) is determined only from the signal visible in the observing band. This is done using the signal-to-noise (S/N) ratio obtained from the de-dispersed and frequency averaged time-series profile of the burst. However, this procedure underestimates the intrinsic fluence of the burst, as it is sensitive to signal present only in the observing band. On the other hand, if we model the observed spectra using a Gaussian function, then the total fluence (including the signal not visible in the band) can be estimated and determined (see section 3.5 of Aggarwal et al., 2021a). Using this burst modeling, it is possible to

obtain a more robust estimate of the intrinsic fluence of the burst (i.e., S_{int}).

Similarly, the bandwidth of the bursts is typically determined by manually identifying the range of frequency channels in which the burst signal is visible. For band-limited bursts that lie on the edge of the band, this would lead to underestimation of the burst bandwidths. Again, suppose we model the burst spectra using a Gaussian model. In that case, we can accurately determine the full width at half max (FWHM) of all the bursts robustly, even if only a small fraction of burst spectra is within the observing band. We use the following equation to estimate the isotropic energy of the bursts from the fluence (F), bandwidth (FWHM), and distance (D_L) of the bursts

$$E = 4\pi 10^{-23} \left(\frac{D_L}{\text{cm}} \right)^2 \left(\frac{F}{\text{Jy s}} \right) \left(\frac{\text{FWHM}}{\text{Hz}} \right) \text{erg.} \quad (5.3)$$

Henceforth, we refer to the S/N derived fluence as $F_{S/N}$, and energies estimated using $F_{S/N}$ and manually identified bandwidth as $E_{S/N}$. We will use F_{fit} and E_{fit} when the fluence and energy are determined using fitting.

The above three steps give us a sample of simulated and detected repeater bursts that can now be compared to infer the observational biases and incompleteness, with some assumptions on the intrinsic (μ_f , σ_f , α) and observational (fluence threshold and bandwidth) parameters. This is explored in the next section.

5.4 Results

Here we consider a simple example to report some of the observed effects of the band-limited nature of burst spectra. We consider an FRB 121102-like repeater observed at varying detection thresholds. We assume the following intrinsic properties for this repeater: cumulative energy distribution follows a single power-law with a fixed slope ($\alpha=-1.5$), a distance of 972 Mpc, normal distribution of μ_f with mean 1650 MHz and standard deviation of 250 MHz, and normal distribution of σ_f with mean of 300 MHz and standard deviation of 250 MHz. We also assume that our observing bandwidth is 800 MHz with a center frequency of 1375 MHz. The choice

of these values is inspired by the observed properties of FRB 121102 reported by Aggarwal et al. (2021a). We then observe this repeater at three different fluence limits (i.e. sensitivity limit of the observing system) of: 0.02 Jy ms, 0.1 Jy ms and 0.4 Jy ms. As discussed previously, some of these bursts will not be detected by the search system, as not enough burst signal is present within the band. In the following sub-sections, we discuss three unique observational effects observed for this simulated repeater. We also try to explain the various observed properties of the two most studied repeaters so far: FRB 121102 and FRB 180916, in the context of the incompleteness due to the banded nature of their bursts.

5.4.1 Cumulative Energy Distribution

Many different shapes and slopes have been reported for the cumulative energy distribution of repeater burst energies (Aggarwal et al., 2021a; Pastor-Marazuela et al., 2021; Cruces et al., 2020; Law et al., 2017). Here we discuss how telescope sensitivity plays a crucial role in determining the observed shape of the cumulative energy distribution by influencing the bursts that are detected by the system.

The bottom row of Figure 5.2 shows the cumulative energy distribution of bursts detected with two different fluence thresholds. Red pluses show the intrinsic power-law, green crosses show the $E_{S/N}$ for the detected bursts, and blue triangles represent the E_{fit} of the detected bursts. The trend of blue triangles and green crosses points deviate significantly from a simple intrinsic power law due to the absence of weak bursts that are not detected. Notably, the shape of observed cumulative energy distribution changes with sensitivity thresholds. As the sensitivity of the observations decrease (left to right in Figure 5.2), the observed distributions deviate from a single power-law to a broken power-law, which appears as a smooth turn at even lower thresholds. The broken power-law is similar to what has been seen for FRB 121102 (Cruces et al., 2020; Aggarwal et al., 2021a) while the smooth turn in power-law was reported for FRB 180916 (Pastor-Marazuela et al., 2021). Here, we have shown that both these effects can occur due to the band-limited nature of repeater bursts.

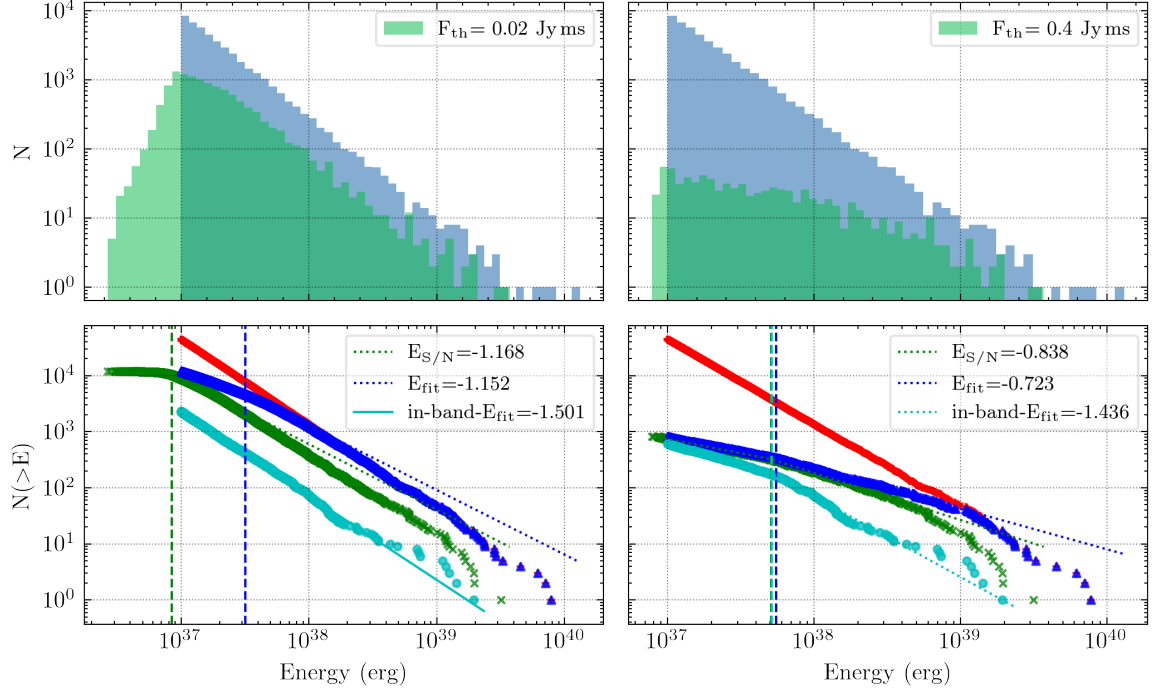


Figure 5.2: Energy distributions of bursts detected at varying fluence thresholds. Different columns show different fluence thresholds - left: 0.02 Jy ms, right: 0.4 Jy ms. Top: Histogram of burst energies. Intrinsic energies of the bursts are shown in blue and show the single power-law, while those of detected bursts are shown in green. In this case, the burst energies are estimated using $F_{S/N}$ (i.e using fluence derived from signal to noise of the burst; see Sections 5.4.1 and 5.4.2). Bottom: Cumulative energy distribution of bursts. Red pluses show the intrinsic energy distribution that follows a power-law with a slope of -1.5. The other three colors show detected bursts. Green crosses represents energies estimated using $F_{S/N}$. Blue triangles and cyan circles are with energies estimated using fitting. Blue triangles shows all the detected bursts, while cyan circles only shows bursts that were primarily within the observing band. Solid lines show a single power-law fit, and dotted lines show a broken power-law fit. The vertical dashed lines shows the break energy for the broken power-law fit. Values in legends report the fitted slope for single power-law fit, and high energy slope for broken power-law fit. As we can see, due to inaccurate energy estimation and missed bursts, the energy distributions can deviate significantly from the intrinsic distribution. See Section 5.4.2 for more details.

5.4.1.1 Challenges

Various possible shapes of the observed cumulative distribution make it challenging to interpret and infer intrinsic FRB properties. A break in the power-law, if present, might indicate the real completeness limit of the system. This can be observed by the green crosses in the bottom-left plot of Figure 5.2. The higher energy slope appears similar to the intrinsic one. Therefore, a break in observed power-law might indicate that the higher energy power-law follows the intrinsic shape and may be used to draw inferences about the intrinsic properties of the FRB. A smooth turnover in cumulative distribution (right panel of Figure 5.2), on the other hand, renders a completeness limit derived from such an analysis inaccurate, and even the higher energy slope might not represent the intrinsic distribution of energies. Even the distribution of E_{fit} might not reflect the real power-law, as the detection threshold would also bias it.

5.4.1.2 Using in-band bursts

When using E_{fit} , although we are using the correct fluences of the bursts, we still would have missed a population of weak bursts that were not primarily within our observing band (see Section 5.3.2). Therefore, we would detect a larger sample of bright bursts, which will make the energy distribution flatter. We can account for this by analyzing only the in-band bursts, i.e., the bursts whose spectra lie primarily within our observing band³. These bursts are labeled as in-band- E_{fit} and shown as cyan circles in Figure 5.2. These in-band bursts provide a more robust and reliable estimate of the distribution of burst energies, even when the observations are not very sensitive. This is because our observations are complete to the in-band bursts, and by using fitting to determine burst properties, we have mitigated both selection effects listed earlier (see Section 5.3.2). The slope of the in-band burst energies follows the intrinsic distribution. But, if the distribution shows a turnover or flattening, then even this method cannot be used to estimate the intrinsic power-law slope of the

³A similar condition was also used by Aggarwal et al. (2021a) to select the bursts for cumulative energy distribution analysis.

bursts reliably.

5.4.2 Energy Distribution

Continuing the previous example, the top row in Figure 5.2 shows the histogram of $E_{S/N}$ for two fluence thresholds. Again, the intrinsic cumulative distribution of burst energies follows a power-law. The detected energy distribution again changes significantly with the fluence threshold, and even in this simple case, it shows a variety of shapes. All of these shapes can be attributed to the missed bursts and inaccurate estimates of burst energies. Using E_{fit} leads to distributions that are more similar to the intrinsic one, and using E_{fit} for only in-band bursts would be an even more accurate representation of intrinsic distribution.

5.4.3 Distribution of spectral parameters

We consider two choices of intrinsic distributions for both μ_f and σ_f : Uniform and Normal, and observed the respective distributions of the detected bursts (see Figure 5.3). The μ_f distribution of detected bursts follows a normal-like distribution for both intrinsic distributions. μ_f distribution for FRB 121102 has been reported to be an asymmetric normal distribution, with a negative skew, i.e., with a tail towards lower frequencies being drawn out (Aggarwal et al., 2021a). The FRB emission also prefers higher frequencies in the 1.4 GHz observations. This cannot be recovered using a uniform intrinsic distribution of μ_f , as in this case, the recovered distribution would peak at the center of the observing band (bottom left panel in Figure 5.3). On the other hand, if the intrinsic μ_f distribution is normal, with its mean present towards the top of the center frequency of the observing band, then an asymmetric normal distribution is recovered. The recovered μ_f distribution of our simple example also shows a negatively skewed distribution (top left panel in Figure 5.3). This is because fewer bursts will be detected towards the higher part of the band. In this case, the peak of the recovered distribution will lie close to the peak of the intrinsic distribution. Hence, based on results presented in Aggarwal et al. (2021a), we can

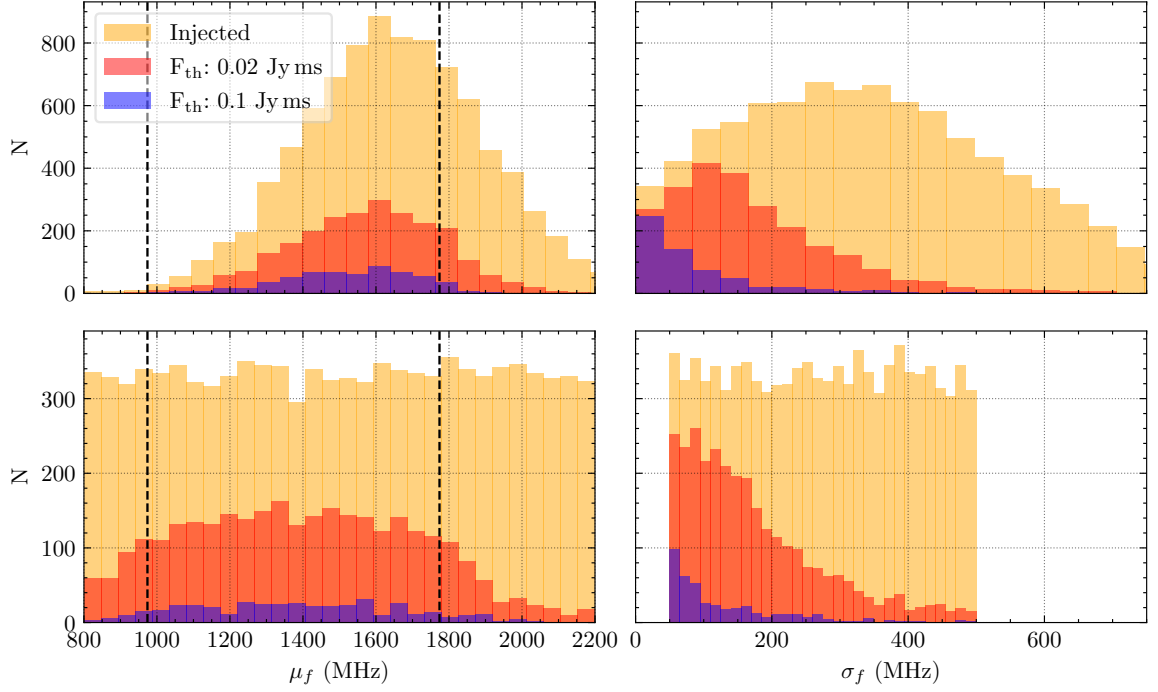


Figure 5.3: Distribution of mean (μ_f) and standard deviation (σ_f) of burst spectra. The intrinsic distribution is shown in yellow, and the distribution of bursts detected at various fluence thresholds is shown in red and blue. The top panels consider a normal intrinsic distribution of μ_f and σ_f , while bottom panels assume a uniform intrinsic distribution. Vertical dotted lines mark the observing band. See Section 5.4.3 for more details.

infer that the intrinsic μ_f distribution of FRB 121102 bursts could be normal, with a mean of ~ 1600 MHz.

The observed normal distribution of σ_f reported by Aggarwal et al. (2021a) can also be explained only by an intrinsic normal distribution of σ_f and is not recovered by a uniform distribution of σ_f . This can also be seen in the right panels in Figure 5.3.

5.4.4 Calculating Energy

As mentioned previously, we used Eq. 5.3 to estimate the energy of the burst. This equation uses the burst bandwidth along with the fluence to estimate the energy. This formalism is the way to estimate the burst energy when the emission is not broad-band. Under the assumption that emission is broad-band, a standard technique to estimate energy from fluence uses center frequency of the band, instead of burst bandwidth (Zhang, 2018). This is given by,

$$E = 4\pi 10^{-23} \left(\frac{D_L}{\text{cm}} \right)^2 \left(\frac{F}{\text{Jy s}} \right) \left(\frac{\nu_c}{\text{Hz}} \right) \text{erg}, \quad (5.4)$$

where ν_c is the center frequency of the observing band. Recently, Li et al. (2021) used this latter definition of energy and found that the burst energy of FRB 121102 follows a bimodal distribution, using a large sample of bursts. We re-calculated the burst energies using the burst bandwidths and fluences reported in their Supplementary Table 1 (Li et al., 2021) and compared them with the energies used by Li et al. (2021).

Figure 5.4 shows the distribution of energies calculated using these two techniques. The distribution of energies does not show any bimodality when burst bandwidths, instead of center frequency, are used to estimate energies. Moreover, the resulting distribution of energies is similar to the ones shown in Figure 5.2, implying that this result is likely affected by band incompleteness. The method of using center frequency to estimate burst energy makes two key assumptions: 1) Emission is broadband, 2) Spectral index is zero, i.e the emission does not depend on frequency.

But, the emission from repeaters is characteristically band-limited and Gaussian-like. Therefore, none of these two assumptions are valid for repeating FRBs and so burst bandwidths should be used to estimate energy.

Figure 5.5 shows the distribution of energies (derived using burst bandwidths) of FRB 121102 bursts detected with FAST and Arecibo (Aggarwal et al., 2021a; Li et al., 2021), along with the recovered distribution of detected bursts using some fiducial values for intrinsic properties similar to the simulation example discussed earlier. Therefore, our analysis shows that the observed shapes in the energy distributions of FRB 121102 can occur due to observational effects caused by bandedness along with a normal distribution of the peak and width of the burst spectra. Notably, both these effects were reported for the bursts presented in Li et al. (2021) and Aggarwal et al. (2021a).

We performed power-law fits on the cumulative distribution of energies (derived using burst bandwidths) from FRB 121102 bursts detected by FAST. We only used the bursts above the energy of 1.2×10^{37} ergs, estimated from the 95% completeness limit (0.06 Jy ms) of FAST observations reported by Li et al. (2021) and mean burst bandwidth of FAST bursts (200 MHz). We fitted these bursts using a single power-law and a broken power-law. The fitted slope for single power-law fit was -0.716 ± 0.002 . The slopes (below and above the break energy) for the broken power-law fit were -0.693 ± 0.001 and -1.12 ± 0.02 with a fitted break energy of $(1.05 \pm 0.02) \times 10^{38}$ ergs. The high energy slope is consistent with the results of Cruces et al. (2020), while they are inconsistent to those reported by Aggarwal et al. (2021a). It should be noted that the energies of FAST bursts were not corrected for the incompleteness due to banded nature of bursts. Further, the fluences were derived using S/N and not using fitting. As discussed, both these effects could lead to the incorrect estimation of energy distribution for this sample.

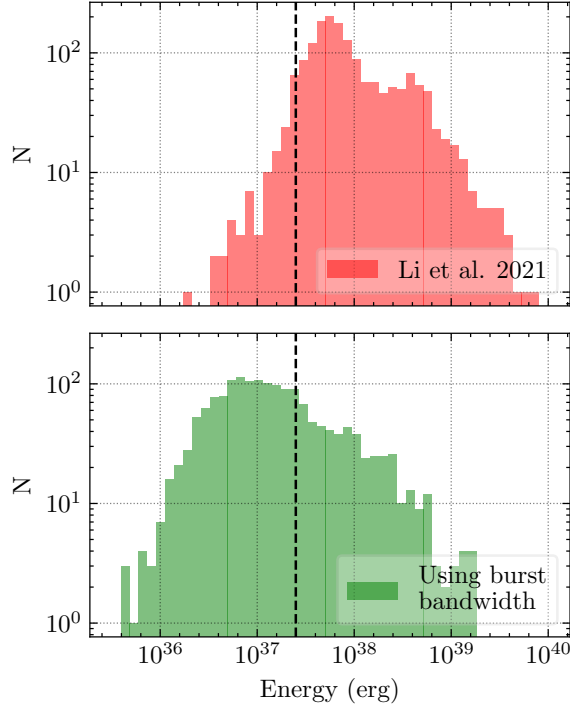


Figure 5.4: Energy distribution of FRB 121102 bursts reported by Li et al. (2021). The top panel shows energies calculated using the center frequency of the band (i.e., 1.25 GHz), while the bottom panel shows energies calculated using the bandwidths of the bursts. Vertical dashed line represents the 90% completeness limit of FAST observations estimated by Li et al. (2021). The bottom distribution does not show the bimodality seen in the top figure.

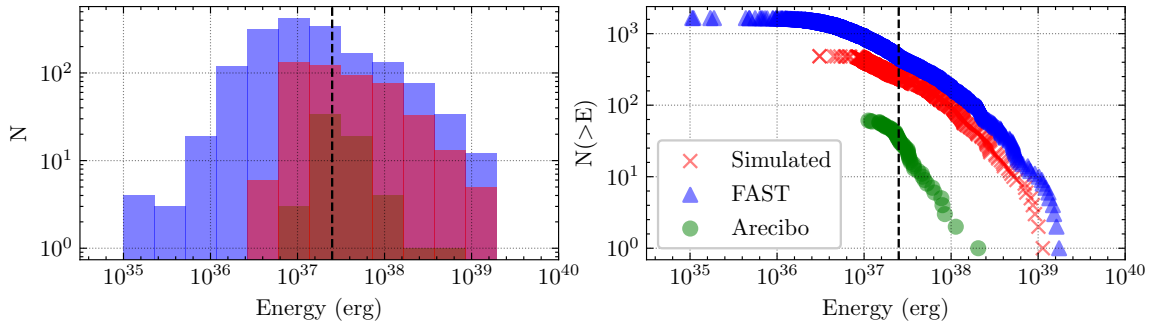


Figure 5.5: Energy distributions of repeater bursts. The left panel shows the histogram of energies, and the right panel shows the cumulative distribution of those energies. Blue triangles represent FRB 121102 bursts reported by Li et al. (2021), where we used burst bandwidth to estimate the burst energy. Green circles show the in-band FRB 121102 bursts reported by Aggarwal et al. (2021a). Red crosses show the bursts that were detected using the simulation example discussed in Section 5.4. Black dashed lines shows the 90% completeness limit of FAST observations (Li et al., 2021). The distribution of the observed FRB 121102 energies is very similar to those recovered using simulations. See Section 5.4.4 for more details.

5.5 Discussion

The results presented in the previous section caution against estimating completeness limits without accounting for the variety of spectral properties of the FRB, especially repeaters. It further complicates the interpretation of energy distributions and properties intrinsic to the FRB. Two main challenges stand out due to the band-limited nature of repeater bursts: 1) Detecting the band-limited bursts, 2) Robust estimation of fluence and bandwidth of bursts that are only partially within the observing band. In the following subsections, we discuss these two challenges.

5.5.1 Observing bandwidth and sub banded searches

One possible technique to mitigate the first challenge listed above is to have a large observing bandwidth. A larger observing bandwidth would make it more likely for the burst spectra to fall within the observing band and aid in detecting more bursts.

Another way is to perform subband searches instead of searching full observational bandwidth (R. Anna-Thomas et al. 2021, in prep; Kumar et al., 2021c). The observing band is divided into multiple subbands, on which a single-pulse search is then performed. This strategy is more sensitive to weak band-limited single pulses that will not exceed the detection threshold with a more traditional full-band single-pulse search.

For the example presented in the previous section, we estimated the increase in the number of detected bursts using a subband search as compared to a full band search. We divided the observing band into four subbands of 200 MHz each and then performed the single-pulse search. We detected more bursts for all the three fluence threshold cases (an increase of 5, 16, and 41%). More bursts, as expected, reduced the observational biases and resulted in recovered properties that were more representative of the intrinsic distributions.

5.5.2 Calculating fluences and bandwidths

The second challenge listed earlier was a robust estimation of fluence and bandwidth of detected bursts. There is no agreed-upon method to estimate the fluence of the bursts as mentioned previously (see Section 5.3.2); signal within the observing band is typically used to estimate burst fluence. In some cases, this signal has been modeled using a power-law, a Gaussian, or a running power-law (The CHIME/FRB Collaboration et al., 2021a; Pleunis et al., 2021b; Aggarwal et al., 2021a) to estimate the burst fluence and other properties. This is because different FRBs show different spectral properties. In this analysis, we have assumed repeater spectra to be Gaussian, as has been reported by observational campaigns on the most studied repeaters (Li et al., 2021; Aggarwal et al., 2021a; Platts et al., 2021; Pastor-Marazuela et al., 2021; Law et al., 2017). The advantage of assuming a functional form for the spectra is that it enables the estimation of fluence and bandwidth of the burst signal, even if the whole burst is not visible in the band (Aggarwal et al., 2021a). This can provide a robust estimate of the intrinsic fluence and bandwidth of the burst if the assumed functional form is correct.

5.5.3 Estimating intrinsic properties

It is now possible to establish a hierarchical framework to infer the intrinsic properties of the repeater bursts. Such a framework would require the following ingredients: 1) observed (or preferably fitted) fluences of the detected bursts from a repeater, 2) completeness limit of the observing system estimated using rigorous injection analysis (Agarwal et al., 2020b; Gupta et al., 2021), 3) DM grid (or DM tolerance) used in single pulse search, 4) Boxcar widths searched. It would need to assume a distribution for burst energies, a spectral shape, and optionally an intrinsic DM and width distribution.

5.5.4 Effects of power-law slope

We also tested the effect of power-law slope on the energy and spectral parameter distributions of the detected bursts. We re-analyzed the example listed above with two more power-law slopes: -1.2 and -1.8 , and observed the recovered distributions. We did not detect any significant difference in the results for these two power-law slopes with respect to the ones presented earlier, for a slope of -1.5 . The cumulative distribution of energy showed a similar flattening and turnover with decreasing sensitivity. The in-band bursts still provided a more robust and accurate estimate of the intrinsic slope. The burst energy distributions also showed similar shapes as reported in Section 5.4.2.

5.6 Conclusions

In this chapter, we have discussed various observational effects that can arise due to the banded nature of the spectra of repeating FRBs. Primarily, the banded nature of burst spectra leads to a non-uniform completeness limit across the observing band. This is because many bursts that lie primarily outside the observing band will not be detected. Therefore, contrary to what is typically understood, the search pipeline is not complete to all the fluences above the sensitivity limit determined using traditional injection analysis (Agarwal et al., 2020b; Gupta et al., 2021; Pastor-Marazuela et al., 2021). This incompleteness must be accounted in the analysis that is used to determine the intrinsic properties of the FRB. Also, it is challenging to estimate the intrinsic fluence and bandwidth of bursts that lie on the edge of the observing band. We assumed the spectra to have a Gaussian shape and simulated bursts from an FRB 121102-like source. We then showed that the energy distribution of the detected bursts looks substantially different from the intrinsic distribution and might show peculiar shapes. We also showed that modeling the burst spectra using a Gaussian to determine the intrinsic fluence and bandwidth provides more robust results than traditional approaches.

A normal intrinsic distribution of μ_f and σ_f can explain the observed distribu-

tion seen for FRB 121102 (Aggarwal et al., 2021a), if the peak of the μ_f distribution lies towards the higher frequency end of the observing band. We point out that burst bandwidths, instead of center frequency, should be used to estimate the energy of the banded repeaters from fluence. We also showed that the bimodality in the energy distribution of FRB 121102 bursts reported by Li et al. (2021) disappears when the energy is estimated using burst bandwidths instead of the center frequency of the band. Based on our tests, we can make the following recommendations for single pulse search and analysis of repeater bursts:

1. Fluence and energies derived using fitting should be preferred over the ones estimated from signal to noise of the burst. Moreover, only the bursts whose peak and bulk of the emission lie primarily within the observing band (i.e., the in-band bursts) should be used to make inferences about the intrinsic distribution of energies and other properties of the FRB.
2. If the cumulative energy distribution shows a break in the power-law, then the higher energy power-law could follow the intrinsic distribution. A smooth turnover in the power-law will probably not represent the intrinsic distribution in the absence of a break. Even still, if the observations are not very sensitive, it might be impossible to recover the intrinsic properties of the repeater (see Section 5.4.1). Moreover, these effects depend on the observed fluences from the source. An energetic repeater, which is also close to us, might be easier to interpret than one that is further away.
3. Subbanded searches are more sensitive to such band-limited bursts and will aid in resolving some of the observational biases listed in this chapter.
4. Analysis to determine the search pipeline completeness should incorporate band-limited spectra of FRBs in the simulated FRB injections.

We note that these conclusions apply only to band-limited transient emission, i.e. they might not necessarily apply to situations where the intrinsic burst bandwidth is

much greater than the bandwidth of the observing hardware in use. All the analysis scripts and notebooks used in this work are provided in a Github repository⁴.

⁴https://github.com/KshitijAggarwal/banded_repeater_analysis

Chapter 6

Localizing FRBs with Realfast

6.1 Chapter Overview

In Chapter 1, we discussed several advantages of localizing FRBs. Briefly, localization region for an FRB can be approximated to the beam shape on the sky. Localizing an FRB enables the identification of its host galaxy, determination of the redshift of the FRB, the study of the local environment, accurate estimation of luminosity, calculation of various dispersion contributors along the line of sight of the FRB, improvement in $DM_{\text{IGM}}-z$ relation, multiwavelength follow-up, etc. This makes localization a key ingredient to uncover the still-hidden processes that give rise to FRBs. Currently, several localization systems are in place around the world that can search and localize FRBs to sub-arcsecond precision: DSA, ASKAP, V-FASTR, and VLA (Hallinan et al., 2019; Bannister et al., 2019; Law et al., 2018, 2015; Wayth et al., 2011). This chapter starts with a brief comparison of interferometers and single-dish radio telescopes, followed by an FRB search using interferometers. It then describes the REALFAST system on the Jansky Very Large Array. We also discuss some FRBs that the REALFAST system has localized.

6.2 Interferometers vs Single Dish Telescope

The beam size (θ_{HPBW}) of a single-dish telescope observing at a center frequency (with corresponding wavelength λ) can be approximated to be: $\theta_{\text{HPBW}} \sim \lambda/D$, where D is the diameter of the telescope (Condon & Ransom, 2016). On the other hand, the synthesized beam size of an interferometer is $\theta_{\text{HPBW}} \sim \lambda/B$, where B is the largest

separation between two antennas of the interferometer. Typically, the diameter of radio telescopes is a few to tens of meters, while the baselines of interferometers can be several kilometers in length. Therefore, the synthesized beam size of an interferometer can be several hundred times smaller than that of the single-dish telescope, enabling precise localization.

Survey speed with an interferometer, i.e speed at which an interferometer can cover a certain amount of sky at a certain sensitivity, scales as: $N_b(A_e N_a / T_{sys})^2 (\lambda / D)^2$. Here, N_b is the number of beams, A_e is the effective area of an antenna element, N_a is the number of antennas in the interferometer, T_{sys} is the system temperature, λ is the wavelength of observation and D is the diameter of an individual antenna. This speed scales roughly with the square of the product of the number of dishes and their diameter. Therefore, VLA surveys are almost 40 times faster than Green Bank Telescope, assuming all other factors are equal (Cordes, 2008; Law et al., 2011). Instead of forming a beam, interferometers measure the visibilities by correlating signals from different antennas in the array. This has the added advantage of better interference rejection, which is not possible for single-dish radio telescopes.

6.3 Transient search using interferometers

There are primarily two approaches to perform transient searches using interferometers. One approach is *beamforming*, where different *synthesized beams* are formed at different positions in the sky, each providing a stream of data. In this case, the data stream from the individual beam can be processed independently, similar to the processing applied on the output of a single dish telescope. To fully cover the primary beam (of HPBW $\sim \lambda/D$) using synthesized beams (of HPBW $\sim \lambda/B$), where D is the diameter of an individual dish and B is the maximum baseline of the interferometer, $N \propto (B/D)^2$ beams or pixels would be needed. Considering B to be 25 m (diameter of a VLA antenna) and D to be 1000 m would imply that 1600 beams would be needed. For larger baselines (i.e., $D > 10,000$ m), hundreds of thousands of beams would be required to cover the entire field of view. Each beam would then

need to be processed independently. Hence, the total computational complexity is $\mathcal{O}[N_t N_f N_{DM} N_{pix}]$, where N_t , N_f , N_{DM} and N_{pix} are number of time samples, number of frequency channels, number of DM trials searched and number of pixels/beams formed. For a large number of beams this makes it highly compute-intensive.

The second, and more commonly used approach, is to compute “visibilities,” i.e., the correlation of electromagnetic waves between pairs of antennas. The data rate of visibility data scales as the number of baselines (pairs of antennas), i.e $N_{bl} = N_{ant}(N_{ant} - 1)/2$, while still consisting of same information as the beamformed data. Therefore, the visibility data rate is independent of the number of beams or pixels in the image. The visibilities can then be dedispersed and Fourier transformed to form images. At this stage, the data rate for beamforming approach and imaging approach would be same, but the computational complexity for imaging approach is $\mathcal{O}[N_t N_{DM}(N_f N_{bl} + N_{pix}^2 \log_2(N_{pix}^2))]$. The first term here is to dedisperse the visibilities and the second term is from 2d FFT for imaging. For large baseline interferometers like VLA, this approach would be orders of magnitude less computationally expensive as the beamforming approach.

6.4 REALFAST pipeline

REALFAST is the commensal, fast-transient search system at the Jansky Very Large Array. It leverages the approach of fast interferometric imaging to search for fast transients in data from VLA, in observations at frequencies 1–10 GHz. Interferometric imaging has been shown to be computationally efficient and maintains ideal sensitivity (Law et al., 2011; Law & Bower, 2012; Law et al., 2015, 2018). In this section we briefly describe the REALFAST search pipeline (see Figure 6.1), and draw comparisons to a typical single-pulse search pipeline described in Section 1.6 (see Law et al., 2018, for more details).

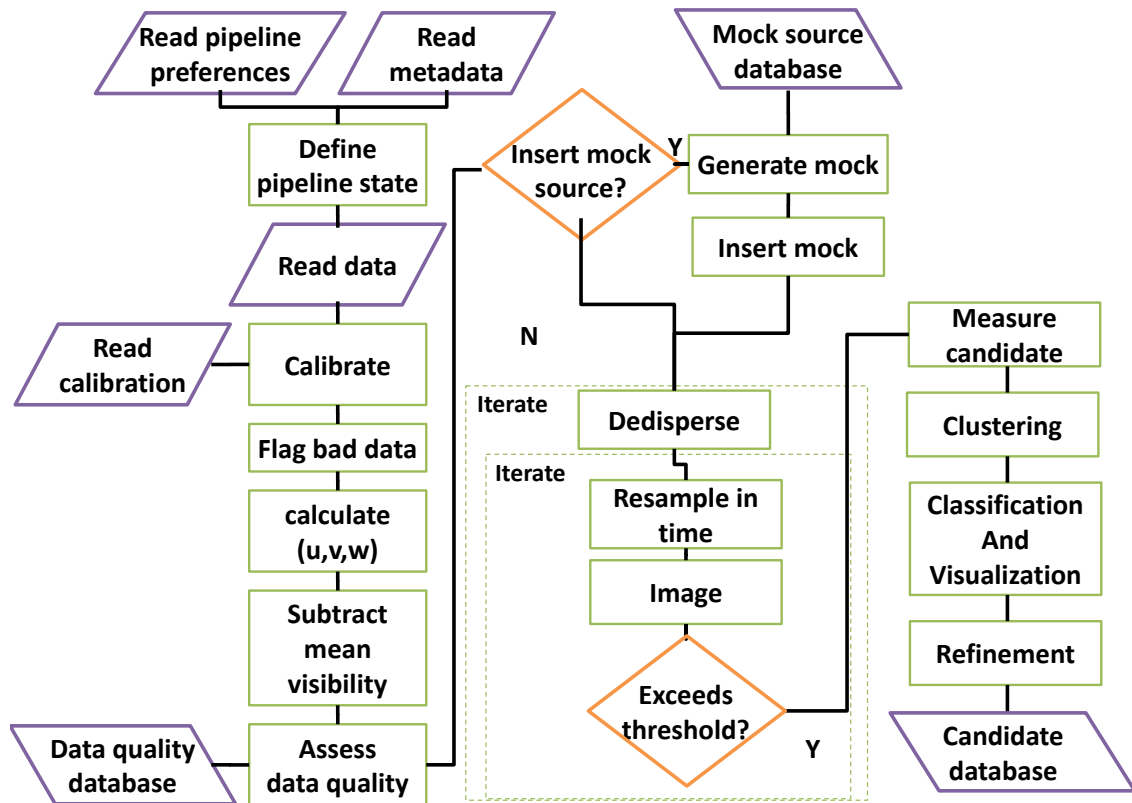


Figure 6.1: REALFAST transient search pipeline. The green boxes denote different processing steps and orange diamonds denote decision points. The purple boxes slanted to the right denote sources of data and those slanted to the left denote output data products. The dashed lines show iterations over trail search values of dispersion measure and sampling time. Figure adapted from Law et al. (2018).

6.4.1 Correlator

The first step is the correlation of voltage data streams from the VLA antennas. This is performed in two stages by a field-programmable gate array system called WIDAR, and a compute cluster known as correlator backend (Perley et al., 2011). Typically, the correlated data is averaged to around 1–5 s, in the correlator. REALFAST reconfigures the correlator to preserve the fast-sampled correlated data product (with a sampling time of 5 or 10 ms), which is then transferred over to the REALFAST computing cluster. This distribution of visibilities is done using *vys* and *vysmaw*, which is a protocol to distribute and consume VLA data (Pokorný & Law, 2017).

6.4.2 Data Segment

REALFAST reads and processes data in short time segments. Consecutive segments overlap to keep a uniform sensitivity for dispersed signals that might lie at the edge of a segment. The segment size is decided based on the largest duration on which mean visibility subtraction is accurate, i.e., the inverse of the fringe rate (of the order of seconds). Each segment is therefore independent and can be processed in parallel. It is useful to note here that this data has one extra dimension of baseline, in addition to the three dimensions with time, frequency channel, and polarization, for a single dish telescope.

6.4.3 Calibration

Calibration solutions generated by the online calibration system TelCal are then applied to the data. Antennas and polarizations with very low gain amplitudes are also flagged here. In the case of potential astrophysical candidates, CASA is used to apply the calibration products. This is discussed in Section 6.5.

6.4.4 RFI Mitigation

Multiple RFI mitigation strategies are applied to this data. Some of these are similar to what was discussed in Section 1.6.2.1, i.e., large deviations in frequency channels and time integrations are identified and flagged. This is applied for each baseline and polarisation. Next, large deviations in complex visibilities across all baselines for each frequency channel, integration and polarization, are filtered out and flagged. This filter is unique to interferometers. Interference local to a subset of antennas will affect only some baselines (pairs of antennas) and can be removed.

6.4.5 Mean visibility subtraction

Millisecond integrated images are much less sensitive than the minutes to hour-long integrated images that are usually generated with interferometers (~ 10 mJy in ~ 10 ms; see Law et al., 2018). Therefore, at this sensitivity, roughly two persistent sources are detectable per square degree (Becker et al., 1995). By subtracting the mean of the visibilities in time (for each segment), constant sources in the images can be removed, eliminating the need for source catalogs. This leaves only the transients with millisecond-scale variability in the images. It is those non-persistent emissions which we then seek to identify through the transient search procedure below.

6.4.6 Dedisperse and Resample

Now, the visibilities are dedispersed for the specified values of DMs in the DM grid. Similar to Section 1.6.2.2, the DM array is generated by defining a sensitivity loss between adjacent DM trials. In `REALFAST`, this loss is set to a default value of 5%. To search for wider candidates, the visibilities are resampled in time by the corresponding width. The widths used in the `REALFAST` are 1, 2, 4, and 8 samples, respectively.

Therefore, data in each segment is independently searched for each DM in the DM list and resample factor in the width list. The visibilities are resampled in time at the specified value of width and then dedispersed at the given value of the DM

trial. Both these operations are performed for time-frequency data present for each baseline and polarization. These functions are written as kernels with numba¹ (so that they can be run on multi-core CPUs), and their GPU implementation (RFGPU²) is used in the realtime system now.

6.4.7 Imaging: project baselines, grid visibilities, Fourier transform

The next step is to generate the image using visibilities. This reconstructs the sky brightness and is needed to search for and localize the FRB. The sky brightness can be obtained by Fourier transforming the visibilities. This is done using Fast Fourier Transform (FFT) as it has lower computational requirements than traditional Fourier Transform implementations. But FFT requires data to be present in a uniformly spaced grid. Therefore, the visibilities are first transformed and gridded to a uniform grid.

Given a time of observation, the position of the target source and observation frequencies, the baselines of antennas are projected to the uvw co-ordinates. The uvw co-ordinates are chosen such that the w -axis is along the source direction, with uv plane orthogonal to it and v axis points to the North Celestial Pole (Thompson et al., 2017; Anantharamaiah et al., 1989). Then the visibilities are gridded using a simple pillbox gridding algorithm at each time step.

Then a 2d inverse FFT is performed on these gridded visibilities to form an image, for each DM and width trial, at each time step (or integration). No primary beam correction is applied, and images are not deconvolved to remove the effect of the point-spread function. The signal to noise of each image is calculated by using the amplitude of the peak pixel in the image and the standard deviation of the image. S/N values that are larger than the threshold are then reported as candidates, and their metadata is returned for further processing.

¹<https://numba.pydata.org/>

²<https://github.com/realfastvla/rfgpu>

6.4.8 Clustering

For each candidate above the S/N threshold, the pipeline reports the following parameters: S/N, DM, width trial, time of candidate, location in the image (l, m) . As discussed in detail in Chapter 3, each real event can trigger the pipeline to return hundreds to thousands of candidates, and therefore all these are clustered using DBSCAN (Ester et al., 1996). This process is similar to what was discussed in Section 1.6.3 and has been discussed in detail for the REALFAST system in Chapter 3. After clustering, the candidates with maximum S/N in each cluster are returned and processed further for visualization and classification.

6.4.9 Visualizations and Classification

At this stage, each candidate metadata is used to reprocess the visibilities to generate a candidate visualization plot. Figure 6.2 shows example of one such plot for FRB180916 (Aggarwal et al., 2020). It shows the candidate parameters, radio image, spectrogram with the candidate’s dedispersed pulse, spectra, and time series. SNR (im/k) refers to image S/N and Kalman score of the FRB (Zackay, in prep). Time-frequency spectrogram extracted for the maximum pixel in the radio image, for each candidate, is then pre-processed and provided to FETCH, a deep learning-based classifier (Agarwal et al., 2020a), to classify the candidate into RFI or FRB.

6.4.10 Data Cutout

For each candidate obtained above, raw visibilities for a time window around the candidate are temporarily saved. This data could be several GB per candidate. Therefore, if, on further investigation, the candidate is found to be RFI, then the data corresponding to the candidate is deleted. In contrast, those that are potentially astrophysical are saved for further analysis and refinement (see Section 6.5).

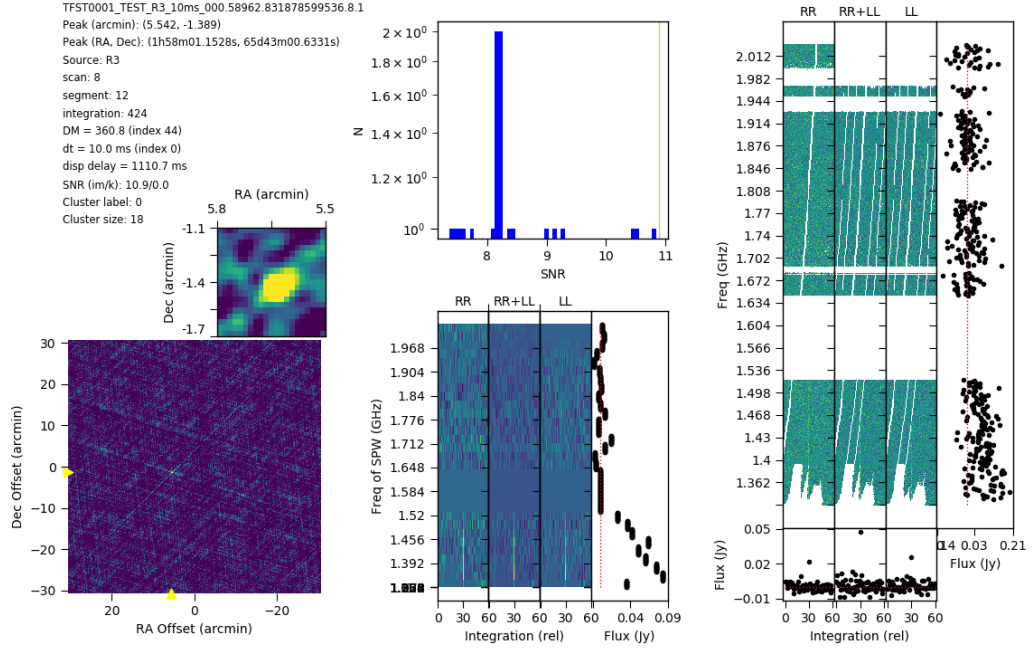


Figure 6.2: REALFAST realtime candidate detection plot of FRB 180916. The meta-data corresponding to the candidate and the observation is shown in the top left. The left plots show the radio image, formed by dedispersing and resampling the visibilities at the trial DM and width value. The middle and right panels show the spectrograms along with the time series and spectra of the data around the burst. The top panel in the middle column shows the histogram of signal-to-noise ratios in that observation. RR and LL represent the two polarisations.

6.4.11 Candidate Database and Portal

All the metadata of candidates, along with the search parameters, and information of VLA observation on which REALFAST was run, is saved in a central elastic-search³ based database. It also contains the statistical properties of each candidate. The candidate visualizations and metadata is displayed on the REALFAST portal⁴, where the REALFAST team members can verify the candidates and label them for further analysis. Candidate information for some candidates is also made publicly available on the portal.

6.5 REALFAST Post-Processing

Candidates marked for further refinement on the REALFAST portal go through a series of offline analyses to refine the parameters and improve the detection significance. Notebooks to perform this analysis are available on Github⁵.

6.5.1 Refinement of properties

The first step is the automatic refinement triggered when a REALFAST team member marks a candidate as interesting or astrophysical on the portal. This process reruns the image search on the candidate data cutout, but with a finer DM grid (with a fractional sensitivity loss of 1% as compared to 5% for the real-time search), larger image size (if possible), and only searches DMs around the detected DM of the candidate. The highest S/N candidate recovered in this search is then visualized using a refinement plot. Figure 6.3 shows one such refinement plot for the FRB 180916 candidate presented in previous section. This plot shows visualizations complementary to the original candidate plot and therefore assists the REALFAST team to decide if further analysis is required on the candidate.

Candidates that are found interesting are then refined manually. Several methods are again tried on the candidate data cutout: finer DM grid, larger image size,

³<https://www.elastic.co/elasticsearch/>

⁴search.realfast.io/

⁵https://github.com/KshitijAggarwal/realfast_analysis/tree/master/refinoloc

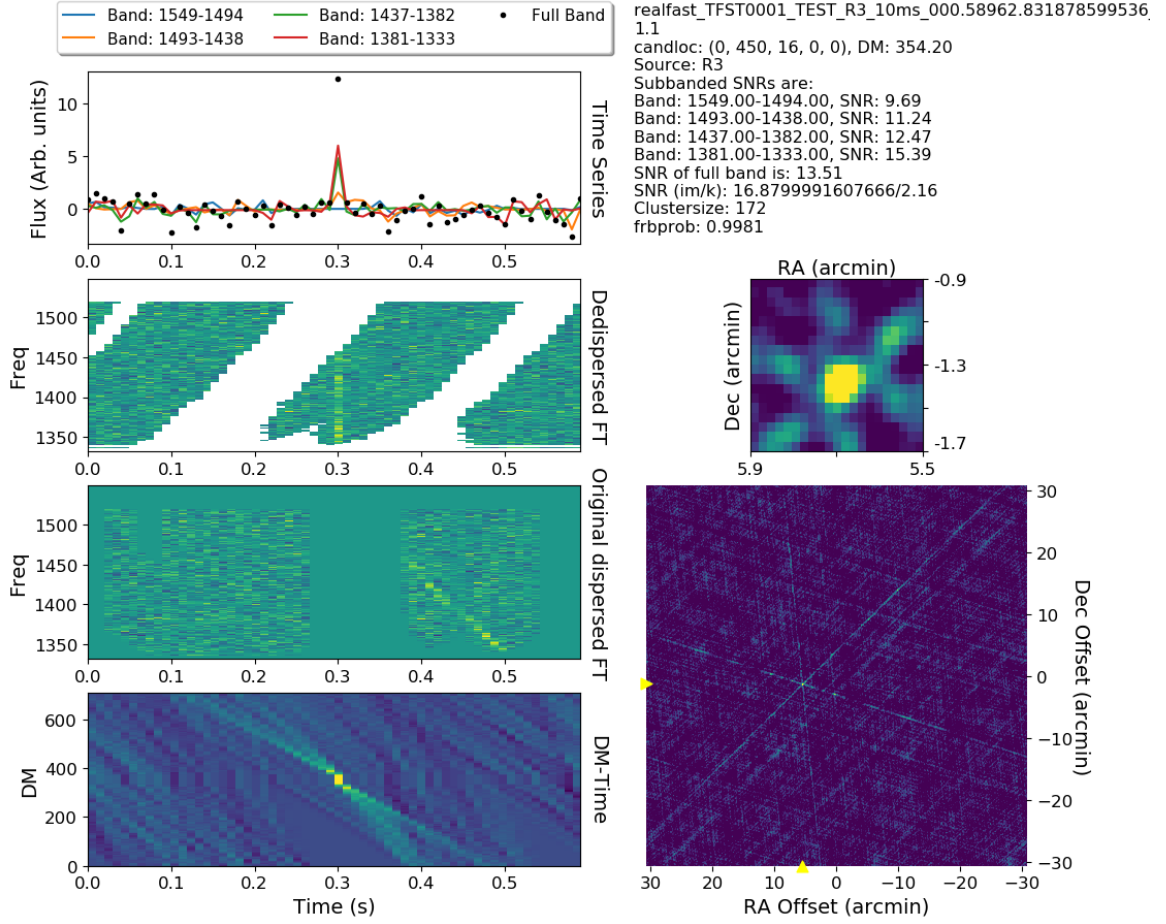


Figure 6.3: REALFAST refinement plot of FRB 180916. Some relevant metadata is shown in the top right. Radio images are shown on the right column. Left column shows (top to bottom): dedispersed time series, dedispersed spectrogram, original spectrogram (not dedispersed), and DM-time of the candidate. S/N of various subbands is also shown, along with the respective time series.

varying flagging thresholds, sub-band search, etc. (Law et al., 2020). The significance of an astrophysical transient should improve when dedispersed at a DM closer to the DM of the candidate (see Chapters 1 and 3). Further, the noise-like events or RFI will not be reproduced when flagging, or image gridding parameters are changed. If the transient signal is present only in a subband, then a sub-band search on only the relevant frequencies will improve the detection significance of the candidate. Candidates that show improvements on refinement are then considered astrophysical and then imaged using CASA.

6.5.2 CASA calibration and Imaging

As mentioned, calibration and imaging done during the real-time search make several assumptions to save time and maintain the real-time processing ability. Further, the point spread function of the interferometer is not deconvolved from the image, which makes the images more difficult to visually interpret and model.

We dedisperse the raw data cutout of the astrophysical candidates and identify the time segments corresponding to the burst signal. These time segments are then saved to a separate file. These two steps are done using SDMPY⁶. We then convert this file to a measurement set (MS) to be then read by Common Astronomy Software Application (CASA; McMullin et al., 2007).

We download the VLA calibration pipeline solutions corresponding to the observation with the candidate and apply those calibration and flagging tables to the burst cutout data using CASA task `applycal`. After calibration, we use CASA task `tclean` to generate the image and estimate the signal-to-noise ratio of the image. As the FRB signal could be in a subband, we search different spectral window ranges to find the spectral window range that generates the image with the highest S/N. We then use CASA task `imfit` around the rough position of the candidate to fit an ellipse to the source in the radio image and measure its centroid location, peak flux density, and 1σ image plane uncertainties. This provides us with the burst position and its statistical error.

6.5.3 Deep imaging

As mentioned previously, the VLA correlator also saves visibilities with a coarser time resolution (1 to 5 seconds) for the whole observation, in parallel to the high time resolution (10 ms) data around the burst. These slow-sampled visibilities can then be calibrated, flagged, and imaged to generate deep radio images to search for any persistent radio emission co-located with the FRB position (Law et al., 2020; Bhandari et al., 2021; Niu et al., 2021a). Using the same calibration tables for deep imaging

⁶<https://github.com/demorest/sdmpy>

and for fast-transients ensures linked systematics for both timescale images.

6.5.4 Astrometry

To estimate the systematic errors on the burst (and any persistent radio source) position, we use the deep images and compare the sources detected in the deep radio image to those listed in the optical PanSTARRS survey DR1 (Chambers et al., 2016). We use the PYSE source finding package (Spreeuw et al., 2018) on the deep radio image formed by concatenating all the available VLA observations of this field. We identify the sources above a significance of 10σ in the radio image. We then search for these sources in PanSTARRS survey DR1 using PSQUERY⁷ (and a search radius of a few arcseconds) and calculate the offset between the optical positions and positions from our deep radio images. We can then estimate the systematic uncertainties on our radio positions from the median and 1σ offset of these cross-matched sources.

If there is no short-timescale phase variation after calibration, these systematic uncertainties can also be used for burst position. This should be true if the phase-reference switching is performed frequently enough and is consistent with the phase-coherence timescales for the VLA.

6.6 REALFAST Localizations

The REALFAST system has been running commensally on all L, S, C and X band observations since early 2020, and has so far searched for a total of 100 days (49 at L, 21 at S, 18 at C and 12 at X band). It is currently the only real-time coherent-imaging interferometric search system in the world. The prototype REALFAST system was used for the first localization of an FRB, FRB 121102 (Chatterjee et al., 2017).

The REALFAST team has collaborated with the CHIME/FRB, ASKAP, and FAST teams to follow up and localize the repeating FRBs discovered by these instruments. Although the CHIME/FRB instrument has the highest discovery rate of any FRB instrument currently in operation, it has poor localization ability, with local-

⁷<https://github.com/realfastvla/psquery>

ization regions that are typically a few of arcminutes in radius (The CHIME/FRB Collaboration et al., 2021a). Similarly, FAST can detect very faint FRBs but again cannot localize them to arcsecond precision (Niu et al., 2021b). Therefore, VLA can be used to follow-up repeating FRBs discovered by these telescopes, where the REALFAST system can search for (and localize) the bursts, and slow-sampled visibilities can be used to image the sky and search for persistent radio emission simultaneously. This section describes some FRBs localized using REALFAST, along with the unique properties of these FRBs. Table 6.1 lists the properties of all these FRBs and their host galaxies.

Table 6.1: Properties of FRBs localized using REALFAST.

Properties	FRB 190520	FRB 180301	FRB 20201124A	FRB 190614D
DM (pc cm ⁻³)	1210.3±0.8	536 ⁺⁸ ₋₁₃	420±10	959.2±5
DM _{ISM} NE2001 (pc cm ⁻³)	60	152	76	83.5
RA (J2000)	16h02m04.266s	06h12m54.42s	05h08m03.50s	4h20m18.13s
Dec (J2000)	−11d17′17.33″	+04d40′15.9″	26d03′37.71″	+73d42′24.3″
Position error (″)	0.1	0.6	1.9	0.8
Fluence (Jy ms)	0.03–0.33	4.9 ^{+0.5} _{-0.4}	2.4±0.1	0.62±0.07
Pulse width (ms)	13.5±1.2	7 ⁺² ₋₃	< 10	5
Host Galaxy				
Redshift	0.241	0.3304	0.098	0.63±0.12
M _* (10 ⁹ M _⊙)	~0.6	2.3±0.6	30 ⁺² ₋₃	~3.9
SFR (M _⊙ /yr)	0.41	1.93±0.58	1.7±0.56	-
PRS				
Luminosity (W Hz ⁻¹)	3.0 × 10 ²²	< 2.0 × 10 ²²	< 3.0 × 10 ²¹	< 3.0 × 10 ²²
Lum. Freq.	3.0 GHz	1.5 GHz	1.4 GHz	1.4 GHz
RA (J2000)	16h02m04.259s	-	-	-
Dec (J2000)	−11d17′17.38″	-	-	-
Reference	Niu et al. (2021a) Bhandari et al. (2021) Ravi et al. (2021) Law et al. (2020)			

Position error refers to the error on the position of the FRB. M_{*} and SFR refer to the stellar mass and star-formation rate of the host galaxy. Lum. Freq. refers to the frequency at which the luminosity of the PRS (or an upper limit) has been reported.

6.6.1 FRB 190614D

FRB 190614D was detected serendipitously on 14 June 2019, during the observation of a CHIME repeater FRB20180814B (Law et al., 2020). This FRB had a DM of 959 pc cm^{-3} , which was much larger than the DM of the source being followed up (189 pc cm^{-3}). It was very faint with a fluence of 0.64 Jy ms , and has the highest DM of any localized FRB so far. Repeat bursts from the source were not detected in 15 hours of VLA and 153 hours of the CHIME/FRB observations. Follow-up observations with Keck and Gemini revealed a pair of galaxies potentially associated with the FRB. Both the galaxies are at a photometric redshift of 0.6 with r -band magnitude of $\sim 23 \text{ mag}$, but with different color and stellar mass. The two galaxies are broadly consistent with Milky Way like stellar mass and star formation rates. No persistent radio emission was seen associated with this FRB down to a luminosity of $3 \times 10^{22} \text{ W Hz}^{-1}$ at 1.4 GHz.

6.6.2 FRB 190520

FRB 190520 was discovered with the FAST as part of the Commensal Radio Astronomy FAST Survey (Li et al., 2019b) at 1.05-1.45 GHz in May 2019 (Niu et al., 2021a). Follow-up observations of this FRB with FAST between April and September 2020 led to a detection of 75 bursts in 18.5 hrs with a mean DM of $1210.3 \pm 0.8 \text{ pc cm}^{-3}$, indicating that it is an active repeater. We observed this FRB with VLA during the second half of 2020 and detected nine bursts (3, 5, and 1 at L, S, and C band, respectively). Deep radio images generated using slow sampled data from VLA revealed the presence of a persistent radio source co-located with this FRB. The flux of this PRS at the S-band was measured to be $202 \pm 8 \mu \text{ Jy}$. This makes FRB 190520, the second FRB, after FRB 121102, with a co-located persistent radio source.

Optical follow-up of this position using CFHT/MegaCam found a galaxy (J160204.31–111718.5) at the location of the FRB. Subaru/MOIRCS observations show that the FRB and PRS lie on the periphery of this galaxy. Spectroscopic observations with Palomar 200-inch Hale Telescope and the Low-Resolution Imaging Spectrome-

ter (LRIS) at the Keck I Telescope show that the redshift of this galaxy is 0.241, with a star-formation rate of $\sim 0.41 \text{ M}_{\odot} \text{ yr}^{-1}$ and a mass of $\sim 6 \times 10^8 \text{ M}_{\odot}$, making it a dwarf galaxy with high star-formation rate.

Considering the dispersion contributions from the Milky Way (Cordes & Lazio, 2002) and from IGM (for this redshift) implies a very large host DM of $\text{DM}_{\text{host}} \sim 900 \text{ pc cm}^{-3}$. This demonstrates that the distribution of host DMs of FRBs can have a long tail, which might add considerable variance to the estimates for the DM_{IGM} . Further, the co-located PRS, star-forming dwarf galaxy, high repetition rate, and high host DM make FRB 190520 a clear analog of FRB 121102, implying that some correlations might exist between these properties.

6.6.3 FRB 180301

FRB 180301 was discovered during Breakthrough Listen observations with Parkes Telescope (Price et al., 2019) at a DM of 512 pc cm^{-3} . Then, in sensitive follow-up observations with FAST, it was found to emit repeat pulses (Luo et al., 2020b). FAST observations revealed that this FRB shows a diversity of polarization position angle swings.

We performed follow up observations of this FRB with VLA, in which the REALFAST system detected one burst at L band, with VLA in C-configuration. Follow-up optical observations with Nordic Optical Telescope and Gemini Telescopes and archival Pan-STARRS images at the localized position from VLA revealed a faint galaxy PSO J093.2268+04.6703 as the host of this FRB, with a high association probability. The redshift of this galaxy is 0.3304, and the FRB lies at the outskirts of this galaxy (projected offset 11 kpc). It is a star-forming galaxy with a star formation rate of $\sim 1.93 \text{ M}_{\odot} \text{ yr}^{-1}$ and a mass of $\sim 2.3 \times 10^9 \text{ M}_{\odot}$ (Bhandari et al., 2021). No persistent radio emission was detected co-located with the FRB down to a luminosity of $2 \times 10^{22} \text{ W Hz}^{-1}$ at 1.5GHz.

6.6.4 FRB 20201124A

In March 2021, the repeating source FRB 20201124A was reported to be in an active state (CHIME/FRB Collaboration, 2021; Lanman et al., 2021). Over the next few months, repeat bursts were detected from this FRB by multiple telescopes between frequencies 700 MHz and 2 GHz (Kumar et al., 2021b,a; Xu et al., 2021; Pearlman et al., 2021; Law et al., 2021a; Wharton et al., 2021; Spitler & Hilmarsson, 2021; Herrmann, 2021; Main et al., 2021; Kirsten et al., 2021c), followed by localization reported from several interferometers (Law et al., 2021a; Wharton et al., 2021; Marcote et al., 2021; Day et al., 2021). We performed VLA observations of this source in April 2021 at L and C band and detected one burst with a signal to noise ratio of 26, DM of 420 ± 10 pc cm⁻³, consistent with the published values of around 413 pc cm⁻³. We measured the burst fluence to be 2.4 ± 0.1 Jy ms, and localized it with an uncertainty of around 2'' (dominated by systematics). The burst was found to be associated with the galaxy SDSS J050803.48+260338.0 (hereafter J0508+2603), with a high association probability (Ravi et al., 2021).

The host galaxy of this FRB is young and extremely massive with a mass of $10^{10.5}$ M_⊙, and is at a redshift of 0.098. Interestingly, it is more massive than the host galaxy of any other repeating FRB. A deep radio image formed using C band observations showed a compact radio source with a peak flux density of 221 ± 15 μJy, co-located with the FRB position. Further analysis by Ravi et al. (2021) revealed that persistent emission is extended on scales greater than 50 mas and found its properties to be consistent with the star-forming rate of around 1.7 M_⊙ yr⁻¹ in the galaxy. Similar conclusions were reached using ASKAP observations of this source by Fong et al. (2021).

6.7 VLA/REALFAST detection of burst from FRB180916.J0158+65 and Tests for Periodic Activity

In this section, we report on the detection of a burst from FRB 180916 by REALFAST/VLA and present software for interpreting FRB periodicity. We demonstrate a range of periodicity analyses with bursts from FRB 180916, FRB 121102 and FRB 180814. Our results for FRB 180916 and FRB 121102 are consistent with published results. For FRB 180814, we did not detect any significant periodic episodes. The REALFAST-detected and other high-frequency bursts for FRB 180916 tend to lie at the beginning of the activity window, indicating a possible phase-frequency relation. The python package `frbpa` can be used to reproduce and expand on this analysis to test models for repeating FRBs.

6.7.1 Realfast Detection of FRB 180916

One of the repeating fast radio bursts (FRBs), FRB 180916, has been found to have periodic episodes of higher activity, with a period of 16.35 days (CHIME/FRB Collaboration et al. 2020a, hereafter PR3). Its bursts are clustered in a 4-day phase window, with some cycles showing no bursts, while others show multiple bursts.

On April 23 at 20:11 UTC, we used the REALFAST commensal fast transient search system at the Karl G. Jansky Very Large Array (VLA) to observe FRB 180916. We detected a burst with an S/N of 13 at a DM of 349.8 pc cm^{-3} (Aggarwal & Realfast Collaboration 2020, hereafter rfATel). The REALFAST system localized the burst in real-time to a location of J2000 R.A.= $01^{\text{h}}58^{\text{m}}00^{\text{s}}.634$, Decl.= $65^{\circ}43'00''.6331$. The position of this burst is consistent with the reported localization by EVN (Marcote et al., 2020), given the VLA localization precision of $0.8''$. The details of the burst were reported in rfATel.

Published as Aggarwal et al. (2020)

Contributing authors: Casey J. Law, Sarah Burke-Spolaor, Geoffrey Bower, Bryan J. Butler, Paul Demorest, Justin Linford, and T. J. W. Lazio

6.7.2 Periodicity Analysis Techniques

Since the discovery of periodicity in the activity of FRB 180916, many observatories have reported detections of bursts from this FRB (Chawla et al., 2020; Marcote et al., 2020; Pilia et al., 2020; Scholz et al., 2020; Sand et al., 2020). In total, 19 bursts have been detected at telescopes other than CHIME from FRB 180916. This, along with the bursts reported by PR3, leads to a sample of 51 bursts that have been used in the analysis reported here.

Following the procedure in PR3 and considering a period of 16.35 days, we generated the phase histogram of all published bursts from FRB 180916 (Fig. 6.4, top panel). Most of the bursts lie within a 4 day (or $4/16 = 0.25$ phase) phase window from phase 0.4 to 0.6. Our VLA detection lies at a phase of 0.3, which is the earliest phase at which a burst has been detected so far. The addition of bursts from telescopes other than CHIME makes the phase distribution more symmetric. We also generated the phase histograms at other periods within the period error reported by PR3, which resulted in a similar conclusion.

We used three tests to search for episode periodicity in this burst sample. First, we used the Pearson Chi-square test done by PR3. Second, we followed the approach of Rajwade et al. (2020a) to search for the period with a folded profile of minimum fractional width. We also use the Quadratic-Mutual-Information-based periodicity search technique (Huijse et al., 2018) implemented in P4J⁸ to search for a period in these bursts. All three search techniques were used on all 51 bursts and also on 32 CHIME bursts. Following PR3, we also searched for a periodicity after binning the data to obtain just the “activity days”.

All the scripts developed for periodicity search and phase analysis reported here are available as a python package `frbpa`⁹. `frbpa` has various functions that can be used to search for periodicity in the activity of repeating FRBs. It can also be used to visualize the dependence of the burst MJDs and observations on phases.

⁸<https://github.com/phuijse/P4J>

⁹<https://github.com/KshitijAggarwal/frbpa>

6.7.3 Results and Discussion

For all the methods listed above, we recovered a period that was consistent with the results of PR3, using all the bursts and using just the “activity days”. We also extended the analysis of aliasing in PR3 to include bursts from all other telescopes. As the periodicity at all the frequencies is expected to be the same, the standard deviation of the burst phases should be minimum at the optimal period (or its alias), which was found to be the case at the published period¹⁰.

We also used `frbpa` to search for periodicity in two other active repeaters: FRB 121102 (R1), and FRB 180814 (R2). Rajwade et al. (2020a) reported a period of 157 days in R1 using a sample of 235 bursts detected over a time span of 7 years. We used P4J to search for periodicity on this sample and recovered a period consistent with their results. For R2, we used 21 bursts detected by CHIME¹¹. We did not detect any significant period, using all the three techniques for periodicity search on these bursts. Moderately significant detections at activity windows of 33, 45, 90, and 138 days were observed. Further burst detections are needed to verify the periods reported for this FRB.

We also note that for FRB 180916, the bursts detected at high frequency (i.e., above 600 MHz; detected using VLA and Effelsberg) are at a lower phase value than most of the low-frequency bursts (Fig. 6.4, top panel). This indicates that there may be a correlation between frequency and phase, where high-frequency emission is suppressed at higher phase values. Although these high-frequency observations have good coverage within the activity phases (Fig. 6.4, bottom panel) with good sensitivity, the detections have only occurred at phase < 0.4 . This cannot be explained by the models which invoke an interacting neutron-star-binary system to explain the periodicity. These models predict a wider activity window at higher frequencies, as high-frequency photons are generally transmittable (Ioka & Zhang, 2020).

Therefore, more high-frequency observations across different activity phases

¹⁰We also followed these methods after including the openly available but unpublished CHIME burst sample from <https://www.chime-frb.ca>, which lead to similar conclusions

¹¹<https://www.chime-frb.ca>

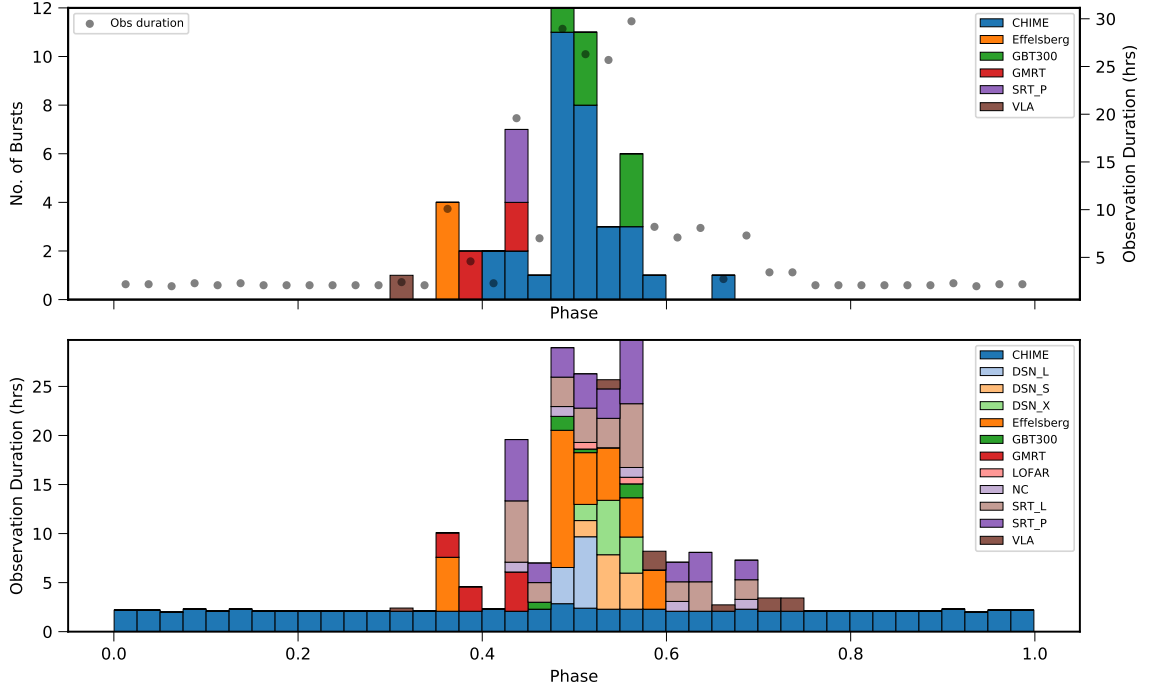


Figure 6.4: Burst detection and observation histograms for FRB 180916, assuming a period of 16.35 days. Different colors represent different observatories. *Top*: Stacked histogram of the detected bursts relative to phases for all published FRB 180916 bursts. The black dots show the total observation duration for each phase, summed for all the telescopes in the bottom panel. *Bottom*: Stacked histogram of observation duration with respect to phase.

would be imperative to comment on the periodicity at high frequency. Moreover, with the detection of more bursts from FRB 180916 and other repeaters in the future, a more confident analysis of plausible periodicity (and phase-dependent burst rates) in their activity would be possible.

Chapter 7

Conclusion

We have discussed several novel techniques and software to optimize the search and analysis of FRBs. We summarize these, provide some context on how these tools have impacted the research in the past years, and update the progress that has been made since these projects were published. We also make some predictions regarding the field of FRBs and the exciting science that lies ahead.

7.1 Open source software

In Chapter 2, we discussed five open-source software packages that we developed to aid in FRB analysis. Each package is provided as a well-documented and tested library that is easy to install and use for any researcher. The motivation behind all of them was essentially the same. We wanted to perform some FRB analysis for a project and could not find any standard tool. So we decided to write our software, use it for the project and make it available for the community to use.

FETCH is a set of deep learning classifiers that can be used to classify FRBs and RFI, in real-time, with high accuracy. It has already been used to discover new FRBs and detect hundreds of bursts from known repeating FRBs. BURSTFIT was initially developed to perform spectro-temporal modeling on FRB 121102 bursts detected using Arecibo Telescope, but can be used to model any FRB (or pulsar) robustly, using any user-defined or standard functions. Since then, it has also been used to model FRBs detected by REALFAST and GBT. FRBPA can be used to search for periodicity in FRB activity and visualize it. It was used in the analysis to demonstrate frequency dependence in the periodic activity in FRB 180916 bursts and measure the period in

bursts detected at 100 MHz with LOFAR and at 1.4 GHz with Apertif. CASP provides the code to estimate the chance coincidence probability of FRBs with the host galaxy. Bayesian frameworks have now been developed that will provide a more robust estimate of association probability, especially in the case of populated fields with multiple nearby galaxies of varying brightness. YOUR (Your Unified Reader) provides an extensive framework to unify data processing across commonly used data formats for pulsar and FRB analysis. It has tools to visualize data and candidates, perform a single-pulse search, RFI mitigation, convert between data formats, pre-process candidates, etc. It is being used in the analysis of several currently unpublished FRBs for a range of interesting applications.

7.2 Clustering REALFAST candidates

In Chapter 3, we discussed a robust assessment of eight unsupervised clustering algorithms to cluster the candidates generated by the REALFAST search pipeline. As mentioned in that chapter and Section 1.6, any single pulse search pipeline searches for transients in data by applying a set of matched filters based on FRB properties like dispersion and width. This can lead to thousands of triggers from an event (RFI or astrophysical) that can overwhelm the post-processing system. Therefore, clustering these redundant candidates is essential for any search pipeline. We used real RFI from REALFAST system and simulated FRB candidates to test different algorithms. For each candidate, we used four features: DM, time, sky location provided by the REALFAST pipeline.

A challenge with clustering candidates from a single pulse search is that candidates generated due to RFI cannot be easily assigned to individual clusters, so true clusters for RFI candidates are not available. In this chapter, we developed a custom performance metric that can use the information about FRB clusters to evaluate the overall performance of clustering algorithms. It is based on the criterion that candidates should be found in only a small number of clusters, and each cluster should contain only FRB or RFI candidates, i.e., we optimized to form a small number of

pure clusters.

We then used this clustering metric to perform a random hyperparameter search and comparison on these eight clustering algorithms. We tested the generalisability of these algorithms by testing them on an independent dataset with only FRB candidates. Finally, we compared the average clustering time of each algorithm on datasets of varying sizes. Using these tests, we concluded that DBSCAN and HDBSCAN could be used for clustering candidates generated by REALFAST system. We subsequently deployed the DBSCAN algorithm on the commensal REALFAST system. We also showed that incorporating spatial features (unique to image-based single-pulse search systems on interferometers) improves the clustering performance compared to the traditional approach of just using DM and time features.

We also proposed a generic strategy that can be used to choose a clustering algorithm for single pulse search systems. It can also be used in other clustering applications. Our performance metric can especially prove useful in applications where the clustering information for only one cluster of interest is available out of an unknown number of true clusters.

7.3 Analysis of FRB 121102 bursts

In Chapter 4, we presented a dense sample of 133 bursts from the repeating source, FRB 121102, detected in 3 hours of observations at L-band using Arecibo Observatory. Importantly, this sample includes 93 additional bursts as compared to the analysis presented by Gourdji et al. (2019) using the same dataset. These data were searched using the pipeline being developed for The Petabyte Project (Aggarwal et al., 2021a). This pipeline uses YOUR, HEIMDALL and FETCH to read the data, search for transients and classify the candidates automatically. We developed BURSTFIT to analyze the properties of all the bursts presented in this work and provided it as a user-friendly python package (also see Chapter 2). We use the MCMC procedure implemented in BURSTFIT to obtain the burst properties.

The burst spectra were well modeled using a Gaussian, with a median band-

width of 230 MHz and a median peak of 1608 MHz. The burst emission favored the top of our observing band, with a lack of emission below 1300 MHz. Many bursts also showed a clear scattering tail, with a median timescale of 0.7 ms. The wait time distribution of bursts showed two peaks: one around 10 ms and the other near 70 s. The latter distribution follows a log-normal distribution, with a peak at 74.8 s, consistent with other observations that analyzed a dense sample of FRB 121102 bursts. We also found that the burst rate distribution of bursts with wait times longer than one second was fit equally well by Poisson and Weibull distributions. We searched the burst times and data for a millisecond to minute timescale period, using many techniques, but did not detect any significant period. We show that the bursts primarily within the observing band should be used for energy distribution analysis. The cumulative burst energy distribution was well fit using a double power-law, with a break. The low and high energy slopes were -0.4 ± 0.1 and -1.8 ± 0.2 with a break at $(2.3 \pm 0.2) \times 10^{37}$ ergs.

Finally, we compared our single-pulse search pipeline to that used by Gourdji et al. (2019) and argued that the following reasons could have led to different results: threshold signal-to-noise ratio, single-pulse search software, RFI mitigation, and classification. We further discussed that recovery rates of the software used in our analysis had been studied in detail and shown to be $> 90\%$ (Agarwal et al., 2020b; Gupta et al., 2021), while the ones used in Gourdji et al. (2019) have not been tested as rigorously.

7.4 Biases due to banded nature of repeater bursts

The previous section and Chapter 4 showed that the bursts of FRB 121102 are band-limited and can be well modeled by a Gaussian function. In Chapter 5, we highlighted the observational biases that arise due to this banded nature of FRB 121102, or any other repeater bursts. These biases lead to non-uniform completeness of observations across the observing band and can significantly impact the observed energy distribution of detected bursts. This is because the bursts that do not lie primarily

within the observing band will not be detected. Therefore, a single completeness limit value typically derived using injection analysis is insufficient in the case of banded repeaters. Further, a second bias can occur when estimating the energies of the detected bursts. A significant fraction of the emission envelope for many bursts might lie outside the observing band. In this case, just using the fluence and bandwidths observable within the band would lead to incorrect estimation of the burst energies.

We used an FRB 121102-like simulation example and demonstrated the above biases using it. We also showed that the observed distribution of mean and standard deviation of FRB 121102 spectra shown in Chapter 4 could be recovered using a normal intrinsic distribution of these parameters. We highlight that burst bandwidths instead of the center frequency of observation should be used to convert fluences to energies, as repeater bursts are band-limited and characteristically do not have a flat spectrum. We then showed that the bimodality reported in the energy distribution of FRB 121102 bursts detected by Li et al. (2021) disappears when burst bandwidths are used to estimate energies.

We showed that the observed shapes in the energy distribution of FRB 121102 and FRB 180916 can be explained using these biases. Based on our tests, we made the following recommendations for single pulse search and analysis of repeater bursts to overcome these biases. Primarily, the fluence and bandwidth of bursts should be derived using the fitting. The fitting procedure can incorporate the burst emission that lies outside the observational band based on the functional form of the spectra. Secondly, only the bursts that lie primarily within the observing band should be used for energy distribution analysis. We showed that observations are more likely to be complete to this sample of in-band bursts than the larger detected sample of bursts. Subband searches will be more sensitive to band-limited bursts, and the analysis that determines search pipeline completeness should incorporate this banded nature of burst spectra.

7.5 Localizing bursts using REALFAST

In Chapter 6, we discussed the FRB search and localization procedure for REALFAST. Sub-arcsecond localizations are crucial to understanding the progenitors, host galaxies, and environments of FRB and are possible only with interferometers like VLA, DSA, and ASKAP. Interferometric fast imaging is feasible for long-baseline interferometers like VLA, as compared to beamforming, where tens of thousands of beams would be required. We described the REALFAST search pipeline, compared it to a typical single-pulse search pipeline (discussed in Section 1.6), and highlighted many steps unique to interferometers. Interferometers enable better RFI removal, as interference local to a set of antennas can be removed by finding outliers across the baselines.

By performing mean subtraction in time (over seconds timescale), constant radio sources can be removed, and a simple threshold can be used to detect the fast transients in the images. REALFAST receives correlated visibilities at a time resolution of 5 ms or 10 ms. It then calibrates the data, flags the RFI, and performs mean subtraction on it. This is followed by dedispersion and resampling of the data based on the trial DM and width values. It then generates the radio images and searches for outliers above the threshold on these images. Clustering (see Chapter 3) and classification (see Chapter 2) is then performed on the candidates, which are then visualized on the REALFAST portal for human verification.

Interesting candidates are refined (automatically and manually) by changing and fine-tuning the search parameters to better classify them into RFI or astrophysical. Finally, astrophysical candidates are calibrated and imaged using CASA, and their precise position is determined. Slow sampled (1-5 s resolution) data from the correlator are used to generate deep radio images to search for any persistent radio emission co-located with the FRB, and systematic uncertainties are estimated by comparing the positions of sources in deep radio images to their corresponding positions from optical surveys. We then discussed five repeaters that have been localized using REALFAST.

FRB 190614D was discovered and localized by the REALFAST system to a pair of galaxies at a photometric redshift of 0.6. The two galaxies are broadly consistent with Milky Way like stellar mass and star formation rates. Repeat bursts from this source were not detected follow-up observations with VLA and CHIME/FRB. No persistent radio emission was seen associated with this FRB.

FRB 190520, discovered with FAST, was localized to the outskirts of a dwarf star-forming galaxy at a redshift of 0.24. It is only the second FRB, after FRB 121102, to have a co-located persistent radio source. Moreover, the implied host DM of this FRB is extremely high ($\sim 900 \text{ pc cm}^{-3}$), challenging the notion of using simple approximations for host DMs to infer relationships between intergalactic DM and redshift of FRBs.

FRB 180301, discovered with Parkes telescope, was localized to the outskirts of a star-forming galaxy at a redshift of 0.33. No persistent radio source was found with this FRB.

FRB 20201124A, discovered with CHIME/FRB, was localized to an extremely massive galaxy at a redshift of 0.098. A compact radio source found co-located with this FRB was consistent with the star-formation of this galaxy and is therefore not similar to the PRS associated with FRB 121102 and FRB 190520.

Finally, FRB 180916, discovered and shown to have periodic burst activity with CHIME/FRB, was localized to a position consistent with the published results. We performed several periodicity tests by compiling all the bursts from this FRB available at the time and provided it in a python package FRBPA. We reported that high-frequency bursts lie earlier in phase as compared to the bursts detected at lower frequencies. This was later shown to be true in extensive observational campaigns using Apertif, LOFAR, and CHIME.

7.6 Past and future of FRBs

A little over four years ago, when I started my graduate studies, the FRBs field looked very different from what it is today. Just over 20 FRBs were known

at that time, of which only one was a repeater. FRB 121102 had become the first FRB to be localized a few months before. Moreover, it introduced another FRB property, a co-located persistent radio source, that researchers promptly added to their list of FRB observables. At the time of writing, around 800 FRBs have been detected, of which 24 are repeaters and 24 FRBs (8 repeaters) have been associated to a host galaxy. Any predictions made in 2017 about the future of FRBs would have considerably underestimated the wealth of results that followed, primarily driven by the commissioning of new instruments. Moreover, it is difficult to predict the impact of new instruments on a young research field like that of FRBs, where every discovery can significantly impact the fundamental understanding. Many of the five-year predictions made in 2019 in an FRB review (Petroff et al., 2019) were realized, and the predictions were revised in 2021 (Petroff et al., 2021).

Our imaginations limit us from predicting something exotic and wildly out-of-the-box as we tend to extrapolate from the known. Therefore, many of the predictions seem overly conservative in hindsight. This leads to “surprising” results. Some FRB results that can be categorized as surprising are: detection of repeating pulses from an FRB (FRB 121102), FRB associated with PRS (FRB 121102, FRB 190520), detection of an FRB-like burst within our Galaxy (FRB 200428), sub-second periodicity in burst components (FRB 20191221A), an FRB localized to a Globular Cluster (FRB 20200120E), an FRB with extremely high host DM (FRB 190520) and FRB with periodic burst activity (FRB 121102, FRB 180916). All of these results were made possible when a previously unexplored part of transient space was probed: high sensitivity (FAST), MJy bursts (STARE2), large field-of-view (ASKAP and CHIME), interferometers (localization using VLA), novel algorithms (search for periodic emission), etc. Framed this way, it sounds remarkably similar to the unexplored space of high DMs (and single-pulse transients) probed by Lorimer et al. (2007), which led to the discovery of the first FRB! It also serves as a reminder to actively seek out avenues that are still untapped, as those carry the potential for remarkable discoveries.

In the future, the number of discovered and localized FRBs will continue to increase rapidly, leading to the classification of the population into several sub-classes

based on common properties (like repetition, morphology, luminosity, host galaxy, local environment, width, periodicity, etc.). More localizations (with VLA, ASKAP, DSA, and CHIME-outiggers) will play a key role in understanding the progenitors of FRBs, while more discoveries (with GBT, FAST, CHIME, Apertif) will improve the understanding of the FRB population. Wideband searches and high-resolution searches will probe deeper into the FRB emission and reveal the yet-unseen frequency and temporal structures. We will discover more exotic sources, perhaps with extremely high (or low) local DM, unusual hosts, unexpected (or delayed) multiwavelength counterparts, weird pulse profiles, etc.

With the rapid advancement in computing hardware and software and rapid development of machine learning (ML), it is reasonable to predict that ML will support a lot of FRB discoveries in the future. Improved algorithms for FRB searches (dedispersion, matched filtering, RFI removal, etc.) implemented on cutting-edge hardware will further accelerate this pace.

In line with the title of this thesis, upcoming instruments like CHORD, DSA-2000, SKA, and CRACO, and the FRB community as a whole will search *harder* for FRBs, localize them *better*, classify them *faster* and make our understanding *stronger* in the coming years. As the number of sources (and host galaxies) accumulate, so can the number of exceptions to the traditionally understood properties of FRBs. A unified theory might emerge that beautifully connects the range of observed properties, but there is no guarantee. In either case, embracing the exceptions, building on the known concepts, and sometimes letting go of the traditional theories may prove the only path forward, similar to what was done after “Lorimer Burst” was detected.

Bibliography

- Agarwal, D., Aggarwal, K., Burke-Spolaor, S., Lorimer, D. R., & Garver-Daniels, N. 2020a, MNRAS, 497, 1661
- Agarwal, D., Lorimer, D. R., Surnis, M. P., et al. 2020b, MNRAS, 497, 352
- Aggarwal, K. 2021, ApJL, 920, L18
- Aggarwal, K., Agarwal, D., Lewis, E. F., et al. 2021a, ApJ, 922, 115
- Aggarwal, K., Burke-Spolaor, S., Tejos, N., et al. 2021b, ApJ, 913, 78
- Aggarwal, K., Law, C. J., Burke-Spolaor, S., et al. 2020, Research Notes of the American Astronomical Society, 4, 94
- Aggarwal, K., & Realfast Collaboration. 2020, The Astronomer’s Telegram, 13664, 1
- Aggarwal, K., Agarwal, D., Kania, J. W., et al. 2020, Journal of Open Source Software, 5, 2750. <https://doi.org/10.21105/joss.02750>
- Aggarwal, K., Burke-Spolaor, S., Law, C. J., et al. 2021c, ApJ, 914, 53
- Amiri, M., Bandura, K., Bhardwaj, M., et al. 2019, Nature, doi:10.1038/s41586-018-0867-7. <https://doi.org/10.1038/s41586-018-0867-7>
- Anantharamaiah, K. R., Cornwell, T. J., & Narayan, R. 1989, Synthesis imaging in radio astronomy, 6, 415
- Ankerst, M., Breunig, M. M., Kriegel, H.-P., & Sander, J. 1999, in Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data, SIGMOD ’99 (New York, NY, USA: Association for Computing Machinery), 49–60. <https://doi.org/10.1145/304182.304187>
- Astropy Collaboration, Robitaille, T. P., Tollerud, E. J., et al. 2013, Astronomy & Astrophysics, 558, A33

- Bannister, K. W., Deller, A. T., Phillips, C., et al. 2019, *Science*, doi:10.1126/science.aaw5903. <https://science.sciencemag.org/content/early/2019/06/26/science.aaw5903>
- Barsdell, B. R. 2012, PhD thesis, Swinburne University of Technology
- Bassa, C. G., Tendulkar, S. P., Adams, E. A. K., et al. 2017, *The Astrophysical Journal*, 843, L8. <https://doi.org/10.3847/2041-8213/aa7a0c>
- Batten, A. J., Duffy, A. R., Wijers, N. A., et al. 2021, *MNRAS*, 505, 5356
- Becker, R. H., White, R. L., & Helfand, D. J. 1995, *ApJ*, 450, 559
- Beloborodov, A. M. 2017, *ApJ*, 843, L26
- Beniamini, P., & Kumar, P. 2020, *MNRAS*, 498, 651
- Bera, A., & Chengalur, J. N. 2019, *MNRAS*, 490, L12
- Bergstra, J., & Bengio, Y. 2012, *J. Mach. Learn. Res.*, 13, 281
- Bhandari, S., & Flynn, C. 2021, *Universe*, 7, 85
- Bhandari, S., Heintz, K. E., Aggarwal, K., et al. 2021, arXiv e-prints, arXiv:2108.01282
- Bhardwaj, M., Gaensler, B. M., Kaspi, V. M., et al. 2021a, *ApJL*, 910, L18
- Bhardwaj, M., Kirichenko, A. Y., Michilli, D., et al. 2021b, *ApJL*, 919, L24
- Bhat, N. D. R., Cordes, J. M., Camilo, F., Nice, D. J., & Lorimer, D. R. 2004, *ApJ*, 605, 759
- Bilous, A. V., Ransom, S. M., & Demorest, P. 2019, *The Astrophysical Journal*, 877, 125. <https://doi.org/10.3847/1538-4357/ab16dd>
- Bloom, J. S., Kulkarni, S. R., & Djorgovski, S. G. 2002, *AJ*, 123, 1111

- Bochenek, C. D., Ravi, V., Belov, K. V., et al. 2020, *Nature*, 587, 59. <http://dx.doi.org/10.1038/s41586-020-2872-x>
- Breiman, L. 2001, *Mach. Learn.*, 45, 5. <https://doi.org/10.1023/A:1010933404324>
- Buch, K. D., Bhatporia, S., Gupta, Y., et al. 2016, *Journal of Astronomical Instrumentation*, 5, 1641018
- Buitinck, L., Louppe, G., Blondel, M., et al. 2013, in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 108–122
- Burke-Spolaor, S. 2018, *Nature Astronomy*, 2, 845
- Burke-Spolaor, S., & Bailes, M. 2010, *MNRAS*, 402, 855
- Burke-Spolaor, S., Johnston, S., Bailes, M., et al. 2012, *MNRAS*, 423, 1351
- Campello, R. J. G. B., Moulavi, D., Zimek, A., & Sander, J. 2015, *ACM Trans. Knowl. Discov. Data*, 10, doi:10.1145/2733381. <https://doi.org/10.1145/2733381>
- Chambers, K. C., Magnier, E. A., Metcalfe, N., et al. 2016, *arXiv e-prints*, arXiv:1612.05560
- Chatterjee, S., Law, C. J., Wharton, R. S., et al. 2017, *Nature*, 541, 58. <https://doi.org/10.1038/nature20797>
- Chawla, P., Andersen, B. C., Bhardwaj, M., et al. 2020, *ApJL*, 896, L41
- Cheng, Y., Zhang, G. Q., & Wang, F. Y. 2020, *MNRAS*, 491, 1498
- CHIME/FRB Collaboration. 2021, *The Astronomer’s Telegram*, 14497, 1
- CHIME/FRB Collaboration, Amiri, M., Bandura, K., et al. 2018, *ApJ*, 863, 48
- CHIME/FRB Collaboration, Andersen, B. C., Bandura, K., et al. 2019, *ApJL*, 885, L24

- CHIME/FRB Collaboration, Amiri, M., Andersen, B. C., et al. 2020a, *Nature*, 582, 351
- CHIME/FRB Collaboration, Andersen, B. C., Bandura, K. M., et al. 2020b, *Nature*, 587, 54
- Comaniciu, D., & Meer, P. 2002, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24, 603
- Condon, J. J., & Ransom, S. M. 2016, *Essential Radio Astronomy*
- Connor, L., Miller, M. C., & Gardenier, D. W. 2020, *MNRAS*, 497, 3076
- Connor, L., & van Leeuwen, J. 2018, *AJ*, 156, 256
- Cordes, J. M. 2008, in *Astronomical Society of the Pacific Conference Series*, Vol. 395, *Frontiers of Astrophysics: A Celebration of NRAO's 50th Anniversary*, ed. A. H. Bridle, J. J. Condon, & G. C. Hunt, 225
- Cordes, J. M., & Chatterjee, S. 2019, *ARA&A*, 57, 417
- Cordes, J. M., & Lazio, T. J. W. 2002, *arXiv Astrophysics e-prints*, astro-ph/0207156
- Cordes, J. M., & McLaughlin, M. A. 2003, *ApJ*, 596, 1142
- Cordes, J. M., Wasserman, I., Hessels, J. W. T., et al. 2017, *ApJ*, 842, 35
- Costa, R., Gordin, J. E. B., & Weltman, A. 2018, *arXiv:1807.01976*, *arXiv:1807.01976*
- Cruces, M., Spitler, L. G., Scholz, P., et al. 2020, *Monthly Notices of the Royal Astronomical Society*, 500, 448. <https://doi.org/10.1093/mnras/staa3223>
- Dash, M., & Liu, H. 2000, in *Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Current Issues and New Applications*, PADKK '00 (Berlin, Heidelberg: Springer-Verlag), 110–121
- Day, C. K., Bhandari, S., Deller, A. T., Shannon, R. M., & Moss, V. A. 2021, *The Astronomer's Telegram*, 14515, 1

- Deneva, J. S., Cordes, J. M., McLaughlin, M. A., et al. 2009, *ApJ*, 703, 2259
- Eatough, R. P., Keane, E. F., & Lyne, A. G. 2009, *MNRAS*, 395, 410
- Eftekhari, T., & Berger, E. 2017, *ApJ*, 849, 162
- Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. 1996, in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD'96 (AAAI Press)*, 226–231
- Farah, W., Flynn, C., Bailes, M., et al. 2019, *MNRAS*, 488, 2989
- Fong, W.-f., Dong, Y., Leja, J., et al. 2021, *ApJL*, 919, L23
- Fonseca, E., Andersen, B. C., Bhardwaj, M., et al. 2020, *ApJL*, 891, L6
- Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, *PASP*, 125, 306
- Foster, G., Karastergiou, A., Golpayegani, G., et al. 2018, *MNRAS*, 474, 3847
- Franti, P., Virmajoki, O., & Hautamaki, V. 2006, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28, 1875
- Frey, B. J., & Dueck, D. 2007, *Science*, 315, 972. <https://science.sciencemag.org/content/315/5814/972>
- Gajjar, V., Siemion, A. P. V., Price, D. C., et al. 2018, *ApJ*, 863, 2
- Gardenier, D. W., Connor, L., van Leeuwen, J., Oostrum, L. C., & Petroff, E. 2021, *A&A*, 647, A30
- Goodman, J., & Weare, J. 2010, *Communications in Applied Mathematics and Computational Science*, 5, 65
- Gourdji, K., Michilli, D., Spitler, L. G., et al. 2019, *The Astrophysical Journal*, 877, L19. <https://doi.org/10.3847/2041-8213/ab1f8a>
- Guidorzi, C., Marongiu, M., Martone, R., et al. 2020, *A&A*, 637, A69

- Gupta, V., Flynn, C., Farah, W., et al. 2021, MNRAS, 501, 2316
- Guyon, I., Gunn, S., Ben-Hur, A., & Dror, G. 2005, in Advances in Neural Information Processing Systems, ed. L. Saul, Y. Weiss, & L. Bottou, Vol. 17 (MIT Press), 545–552. <https://proceedings.neurips.cc/paper/2004/file/5e751896e527c862bf67251a474b3819-Paper.pdf>
- Hackstein, S., Brüggen, M., & Vazza, F. 2021, MNRAS, 501, 3825
- Hackstein, S., Brüggen, M., Vazza, F., & Rodrigues, L. F. S. 2020, MNRAS, 498, 4811
- Hallinan, G., Ravi, V., Weinreb, S., et al. 2019, in Bulletin of the American Astronomical Society, Vol. 51, 255
- Harris, C. R., Millman, K. J., van der Walt, S. J., et al. 2020, Nature, 585, 357. <https://doi.org/10.1038/s41586-020-2649-2>
- He, K., Zhang, X., Ren, S., & Sun, J. 2015, arXiv e-prints, arXiv:1512.03385
- Heintz, K. E., Prochaska, J. X., Simha, S., et al. 2020, ApJ, 903, 152
- Herrmann, W. 2021, The Astronomer’s Telegram, 14556, 1
- Hessels, J. W. T., Spitler, L. G., Seymour, A. D., et al. 2019, ApJL, 876, L23
- Hilmarsson, G. H., Spitler, L. G., Keane, E. F., et al. 2020, MNRAS, 493, 5170
- Hogg, D. W., & Foreman-Mackey, D. 2018, ApJS, 236, 11
- Hogg, D. W., Pahre, M. A., McCarthy, J. K., et al. 1997, MNRAS, 288, 404
- Hotan, A. W., van Straten, W., & Manchester, R. N. 2004, PASA, 21, 302
- Houben, L. J. M., Spitler, L. G., ter Veen, S., et al. 2019, A&A, 623, A42
- Huijse, P., Estévez, P. A., Förster, F., et al. 2018, ApJS, 236, 12
- Hunter, J. D. 2007, Computing in Science & Engineering, 9, 90

- Inoue, S. 2004, MNRAS, 348, 999
- Ioka, K. 2003, ApJL, 598, L79
- Ioka, K., & Zhang, B. 2020, ApJL, 893, L26
- Iwazaki, A. 2014, arXiv:1412.7825, arXiv:1412.7825
- Jain, A. K., Murty, M. N., & Flynn, P. J. 1999, ACM Comput. Surv., 31, 264.
<https://doi.org/10.1145/331499.331504>
- James, C. W., Prochaska, J. X., Macquart, J. P., et al. 2021, MNRAS, arXiv:2101.08005
- Karuppusamy, R., Stappers, B. W., & van Straten, W. 2010, A&A, 515, A36
- Kashiyama, K., & Murase, K. 2017, ApJ, 839, L3
- Katz, J. 2018, Progr. in Particle and Nuclear Phys., 103, 1
- Katz, J. I. 2017, MNRAS, 471, L92
- Keane, E. F., & Petroff, E. 2015, MNRAS, 447, 2852
- Kirsten, F., Snelders, M. P., Jenkins, M., et al. 2021a, Nature Astronomy, 5, 414
- Kirsten, F., Marcote, B., Nimmo, K., et al. 2021b, arXiv e-prints, arXiv:2105.11445
- Kirsten, F., Ould-Boukattine, O. S., Nimmo, K., et al. 2021c, The Astronomer's Telegram, 14605, 1
- Kramer, M., Wielebinski, R., Jessner, A., Gil, J. A., & Seiradakis, J. H. 1994, A&AS, 107, 515
- Kumar, P., Shannon, R. M., Keane, E., Moss, V. A., & Askap-Craft Survey Science Project. 2021a, The Astronomer's Telegram, 14508, 1
- Kumar, P., Shannon, R. M., Moss, V., Qiu, H., & Bhandari, S. 2021b, The Astronomer's Telegram, 14502, 1

- Kumar, P., Shannon, R. M., Osłowski, S., et al. 2019, *ApJL*, 887, L30
- Kumar, P., Shannon, R. M., Flynn, C., et al. 2021c, *MNRAS*, 500, 2525
- Lam, S. K., Pitrou, A., & Seibert, S. 2015, in *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*, 1–6
- Lanman, A. E., Andersen, B. C., Chawla, P., et al. 2021, arXiv e-prints, arXiv:2109.09254
- Law, C., Tendulkar, S., Clarke, T., Aggarwal, K., & Bethapudy, S. 2021a, *The Astronomer’s Telegram*, 14526, 1
- Law, C. J., & Bower, G. C. 2012, *ApJ*, 749, 143
- Law, C. J., Connor, L., & Aggarwal, K. 2021b, arXiv e-prints, arXiv:2110.15323
- Law, C. J., Jones, G., Backer, D. C., et al. 2011, *ApJ*, 742, 12
- Law, C. J., Bower, G. C., Burke-Spolaor, S., et al. 2015, *ApJ*, 807, 16
- Law, C. J., Abruzzo, M. W., Bassa, C. G., et al. 2017, *ApJ*, 850, 76
- Law, C. J., Bower, G. C., Burke-Spolaor, S., et al. 2018, *ApJS*, 236, 8
- Law, C. J., Omand, C. M. B., Kashiyama, K., et al. 2019, *ApJ*, 886, 24
- Law, C. J., Butler, B. J., Prochaska, J. X., et al. 2020, *ApJ*, 899, 161
- Lawrence, E., Vander Wiel, S., Law, C., Burke Spolaor, S., & Bower, G. C. 2017, *AJ*, 154, 117
- Levin, L. 2012, PhD thesis, Swinburne University of Technology
- Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. 2013, *Journal of Experimental Social Psychology*, 49, 764. <https://www.sciencedirect.com/science/article/pii/S0022103113000668>

- Li, B., Li, L.-B., Zhang, Z.-B., et al. 2019a, *International Journal of Cosmology, Astronomy and Astrophysics*, 1, 22. <https://doi.org/10.18689/ijcaa-1000108>
- Li, C. K., Lin, L., Xiong, S. L., et al. 2021, *Nature Astronomy*, 5, 378
- Li, D., Dickey, J. M., & Liu, S. 2019b, *Research in Astronomy and Astrophysics*, 19, 016. <https://doi.org/10.1088/1674-4527/19/2/16>
- Li, D., Wang, P., Zhu, W. W., et al. 2021, *Nature*, 598, 267. <https://doi.org/10.1038/s41586-021-03878-5>
- Lorimer, D. R. 2011, *SIGPROC: Pulsar Signal Processing Programs*, , , ascl:1107.016
- Lorimer, D. R., Bailes, M., McLaughlin, M. A., Narkevic, D. J., & Crawford, F. 2007, *Science*, 318, 777
- Lorimer, D. R., & Kramer, M. 2004, *Handbook of Pulsar Astronomy*, Vol. 4
- Lu, W., & Kumar, P. 2018, *MNRAS*, 477, 2470
- Luo, R., Men, Y., Lee, K., et al. 2020a, *MNRAS*, 494, 665
- Luo, R., Wang, B. J., Men, Y. P., et al. 2020b, *Nature*, 586, 693
- Lyu, F., Meng, Y.-Z., Tang, Z.-F., et al. 2021, *Frontiers of Physics*, 16, 24503
- Lyubarsky, Y. 2014, *MNRAS*, 442, L9
- Lyutikov, M., Burzawa, L., & Popov, S. B. 2016, *MNRAS*, 462, 941
- Macquart, J. P., Prochaska, J. X., McQuinn, M., et al. 2020, *Nature*, 581, 391
- Macqueen, J. 1967, in *In 5-th Berkeley Symposium on Mathematical Statistics and Probability*, 281–297
- Madison, D. R., Agarwal, D., Aggarwal, K., et al. 2019, *ApJ*, 887, 252
- Main, R., Bethapudi, S., & Marthi, V. 2021, *The Astronomer’s Telegram*, 14933, 1

- Main, R., Yang, I. S., Chan, V., et al. 2018, *Nature*, 557, 522
- Majid, W. A., Pearlman, A. B., Nimmo, K., et al. 2020, *ApJL*, 897, L4
- Mannings, A. G., Fong, W.-f., Simha, S., et al. 2021, *ApJ*, 917, 75
- Marcote, B., Paragi, Z., Hessels, J. W. T., et al. 2017, *The Astrophysical Journal*, 834, L8. <https://doi.org/10.3847/2041-8213/834/2/18>
- Marcote, B., Nimmo, K., Hessels, J. W. T., et al. 2020, *Nature*, 577, 190
- Marcote, B., Kirsten, F., Hessels, J. W. T., et al. 2021, *The Astronomer’s Telegram*, 14603, 1
- Margalit, B., & Metzger, B. D. 2018, *ApJ*, 868, L4
- McInnes, L., & Healy, J. 2017, in *Data Mining Workshops (ICDMW)*, 2017 IEEE International Conference on, IEEE, 33–42
- McInnes, L., Healy, J., & Astels, S. 2017, *The Journal of Open Source Software*, 2, 205
- McKinnon, M. M. 2014, *PASP*, 126, 476
- McLaughlin, M. A., & Cordes, J. M. 2003, *ApJ*, 596, 982
- McLaughlin, M. A., Lyne, A. G., Lorimer, D. R., et al. 2006, *Nature*, 439, 817
- McMullin, J. P., Waters, B., Schiebel, D., Young, W., & Golap, K. 2007, in *Astronomical Society of the Pacific Conference Series*, Vol. 376, *Astronomical Data Analysis Software and Systems XVI*, ed. R. A. Shaw, F. Hill, & D. J. Bell, 127
- Men, Y., Aggarwal, K., Li, Y., et al. 2019, *MNRAS*, 489, 3643
- Mereghetti, S., Savchenko, V., Ferrigno, C., et al. 2020, *ApJL*, 898, L29
- Metzger, B. D., Margalit, B., & Sironi, L. 2019, *MNRAS*, 485, 4091
- Michilli, D., Hessels, J. W. T., Lyon, R. J., et al. 2018, *MNRAS*, 480, 3457

- Mickaliger, M. B., McLaughlin, M. A., Lorimer, D. R., et al. 2012, *ApJ*, 760, 64
- Morello, V., Barr, E. D., Stappers, B. W., Keane, E. F., & Lyne, A. G. 2020, *MNRAS*, 497, 4654
- Mullner, D. 2013, *Journal of Statistical Software*, Articles, 53, 1. <https://www.jstatsoft.org/v053/i09>
- Nita, G. M., & Gary, D. E. 2010, *Monthly Notices of the Royal Astronomical Society*, 406, L60
- Niu, C. H., Aggarwal, K., Li, D., et al. 2021a, arXiv e-prints, arXiv:2110.07418
- Niu, C.-H., Li, D., Luo, R., et al. 2021b, *ApJL*, 909, L8
- Oostrum, L. C., Maan, Y., van Leeuwen, J., et al. 2020, *Astronomy & Astrophysics*, 635, A61. <https://doi.org/10.1051/0004-6361/201937422>
- Oppermann, N., Yu, H.-R., & Pen, U.-L. 2018, *Monthly Notices of the Royal Astronomical Society*, 475, 5109. <https://doi.org/10.1093/mnras/sty004>
- Palliyaguru, N. T., Agarwal, D., Golpayegani, G., et al. 2021, *MNRAS*, 501, 541
- Pang, D., Goseva-Popstojanova, K., Devine, T., & McLaughlin, M. 2018, *MNRAS*, 480, 3302
- Parent, E., Kaspi, V. M., Ransom, S. M., et al. 2018, *ApJ*, 861, 44
- Pastor-Marazuela, I., Connor, L., van Leeuwen, J., et al. 2021, *Nature*, 596, 505
- Patel, C., Agarwal, D., Bhardwaj, M., et al. 2018, *ApJ*, 869, 181
- Pearlman, A. B., Majid, W. A., Prince, T. A., et al. 2021, *The Astronomer’s Telegram*, 14519, 1
- . 2020, *ApJL*, 905, L27
- Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, *Journal of Machine Learning Research*, 12, 2825

- Perley, R. A., Chandler, C. J., Butler, B. J., & Wrobel, J. M. 2011, *ApJL*, 739, L1
- Petroff, E., Hessels, J. W. T., & Lorimer, D. R. 2019, *A&A Rev.*, 27, 4
- . 2021, arXiv e-prints, arXiv:2107.10113
- Petroff, E., Johnston, S., Keane, E. F., et al. 2015, *MNRAS*, 454, 457
- Pilia, M., Burgay, M., Possenti, A., et al. 2020, *ApJL*, 896, L40
- Platts, E., Weltman, A., Walters, A., et al. 2019, *Phys. Rept.*, 821, 1
- Platts, E., Caleb, M., Stappers, B. W., et al. 2021, *MNRAS*, 505, 3041
- Pleunis, Z., Michilli, D., Bassa, C. G., et al. 2021a, *ApJL*, 911, L3
- Pleunis, Z., Good, D. C., Kaspi, V. M., et al. 2021b, arXiv e-prints, arXiv:2106.04356
- Pokorny, M., & Law, C. J. 2017, *vysmaw: Fast visibility stream muncher*, , , ascl:1710.001
- Pol, N., Lam, M. T., McLaughlin, M. A., Lazio, T. J. W., & Cordes, J. M. 2019, *ApJ*, 886, 135
- Popov, M. V., & Stappers, B. 2007, *A&A*, 470, 1003
- Price, D. C., Foster, G., Geyer, M., et al. 2019, *MNRAS*, 486, 3636
- Price-Whelan, A. M., Sipőcz, B. M., Günther, H. M., et al. 2018, *The Astronomical Journal*, 156, 123
- Prochaska, J. X., & Zheng, Y. 2019, *MNRAS*, 485, 648
- Prochaska, J. X., Macquart, J.-P., McQuinn, M., et al. 2019, *Science*, 366, 231
- Rajwade, K. M., Mickaliger, M. B., Stappers, B. W., et al. 2020a, *MNRAS*, 493, 4418
- . 2020b, *MNRAS*, 495, 3551
- Ransom, S. 2011, *PRESTO: PulsAR Exploration and Search TOOLkit*, , , ascl:1107.017

- Ravi, V., Catha, M., Addario, L. D., et al. 2019, *Nature*, doi:10.1038/s41586-019-1389-7. <https://doi.org/10.1038/s41586-019-1389-7>
- Ravi, V., Law, C. J., Li, D., et al. 2021, arXiv e-prints, arXiv:2106.09710
- Reback, J., McKinney, W., Jbrockmendel, et al. 2021, pandas-dev/pandas: Pandas 1.2.1, Zenodo, doi:10.5281/ZENODO.3509134. <https://zenodo.org/record/3509134>
- Resmi, L., Vink, J., & Ishwara-Chandra, C. H. 2021, *A&A*, 655, A102
- Ridnaia, A., Svinkin, D., Frederiks, D., et al. 2021, *Nature Astronomy*, 5, 372
- Romero, G. E., del Valle, M. V., & Vieyro, F. L. 2016, *Phs. Rev. D*, 93, 023001
- Rosenberg, A., & Hirschberg, J. 2007, in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)* (Prague, Czech Republic: Association for Computational Linguistics), 410–420. <https://www.aclweb.org/anthology/D07-1043>
- Russakovsky, O., Deng, J., Su, H., et al. 2014, arXiv e-prints, arXiv:1409.0575
- Sand, K. R., Gajjar, V., Pilia, M., et al. 2020, *The Astronomer’s Telegram*, 13781, 1
- Scheuer, P. A. G. 1968, *Nature*, 218, 920
- Scholz, P., Spitler, L. G., Hessels, J. W. T., et al. 2016, *The Astrophysical Journal*, 833, 177. <https://doi.org/10.3847/1538-4357/833/2/177>
- Scholz, P., Cook, A., Cruces, M., et al. 2020, *ApJ*, 901, 165
- Shannon, C. 1949, *Proceedings of the IRE*, 37, 10. <https://doi.org/10.1109/jrproc.1949.232969>
- Shannon, R. M., Macquart, J. P., Bannister, K. W., et al. 2018, *Nature*, 562, 386
- Simard, D., & Ravi, V. 2020, *ApJL*, 899, L21

- Simha, S., Tejos, N., Prochaska, J. X., et al. 2021, ApJ, 921, 134
- Spitler, L., & Hilmarsson, H. 2021, The Astronomer’s Telegram, 14537, 1
- Spitler, L. G., Cordes, J. M., Hessels, J. W. T., et al. 2014, The Astrophysical Journal, 790, 101. <https://doi.org/10.1088/0004-637x/790/2/101>
- Spitler, L. G., Scholz, P., Hessels, J. W. T., et al. 2016, Nature, 531, 202. <https://doi.org/10.1038/nature17168>
- Spreeuw, H., Swinbank, J., Molenaar, G., et al. 2018, PySE: Python Source Extractor for radio astronomical images, , , ascl:1805.026
- Taylor, J. H. 1974, A&AS, 15, 367
- Tendulkar, S. P., Bassa, C. G., Cordes, J. M., et al. 2017, ApJ Letters, 834, L7
- Tendulkar, S. P., Gil de Paz, A., Kirichenko, A. Y., et al. 2021, ApJL, 908, L12
- The CHIME/FRB Collaboration, :, Amiri, M., et al. 2021a, arXiv e-prints, arXiv:2106.04352
- The CHIME/FRB Collaboration, Andersen, B. C., Bandura, K., et al. 2021b, arXiv e-prints, arXiv:2107.08463
- Thompson, A. R., Moran, J. M., & Swenson, George W., J. 2017, Interferometry and Synthesis in Radio Astronomy, 3rd Edition, doi:10.1007/978-3-319-44431-4
- Thompson, C. 2017, ApJ, 844, 65
- Thornton, D., Stappers, B., Bailes, M., et al. 2013, Science, 341, 53
- VanderPlas, J. T. 2018, The Astrophysical Journal Supplement Series, 236, 16. <http://dx.doi.org/10.3847/1538-4365/aab766>
- Vieyro, F. L., Romero, G. E., Bosch-Ramon, V., Marcote, B., & del Valle, M. V. 2017, A&A, 602, A64

- Virtanen, P., Gommers, R., Oliphant, T. E., et al. 2020, *Nature Methods*, 17, 261
- Wagstaff, K. L., Tang, B., Thompson, D. R., et al. 2016, *PASP*, 128, 084503
- Wayth, R. B., Briske, W. F., Deller, A. T., et al. 2011, *ApJ*, 735, 97
- Wes McKinney. 2010, in *Proceedings of the 9th Python in Science Conference*, ed. Stéfan van der Walt & Jarrod Millman, 56 – 61
- Wharton, R., Bethapudi, S., Gautam, T., et al. 2021, *The Astronomer’s Telegram*, 14529, 1
- Xu, H., Niu, J., Lee, K., et al. 2021, *The Astronomer’s Telegram*, 14518, 1
- Yao, J. M., Manchester, R. N., & Wang, N. 2017, *ApJ*, 835, 29
- Zackay, B., & Ofek, E. O. 2017, *ApJ*, 835, 11
- Zhang, B. 2017, *ApJ*, 836, L32
- . 2018, *ApJL*, 867, L21
- . 2020, *Nature*, 587, 45
- Zhang, T., Ramakrishnan, R., & Livny, M. 1996, in *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, SIGMOD ’96* (New York, NY, USA: Association for Computing Machinery), 103–114. <https://doi.org/10.1145/233269.233324>
- Zhang, Y. G., Gajjar, V., Foster, G., et al. 2018, *ApJ*, 866, 149
- Zhang, Z. J., Yan, K., Li, C. M., Zhang, G. Q., & Wang, F. Y. 2021, *ApJ*, 906, 49
- Zhu, W., & Feng, L.-L. 2021, *ApJ*, 906, 95