Title

Novel Algorithmic and Astrophysical Methods in the Search for Radio Technosignatures

Author

Brzycki, Bryan Francis

Publication Date

2024

Peer reviewed|Thesis/dissertation

Novel Algorithmic and Astrophysical Methods in the Search for Radio Technosignatures

by

Bryan F. Brzycki

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Astrophysics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Imke de Pater, Chair
Dr. Andrew P. V. Siemion, Co-chair
Professor Aaron Parsons
Professor Steven Beckwith

Spring 2024

Novel Algorithmic and Astrophysical Methods in the Search for Radio Technosignatures

Abstract

Novel Algorithmic and Astrophysical Methods in the Search for Radio Technosignatures

by

Bryan F. Brzycki

Doctor of Philosophy in Astrophysics

University of California, Berkeley

Professor Imke de Pater, Chair

Dr. Andrew P. V. Siemion, Co-chair

Over the 60 years since the first published search for radio technosignatures, relatively established methods have come about for the detection and analysis of narrowband radio signals in the Search for Extraterrestrial Intelligence (SETI). Generally, this involves using position-switching to take multiple observations on and off the target of interest and detecting raw narrowband signals via a matched filter that linearly fits Doppler accelerations to each signal. High quality candidates are identified in those cases in which detected signals appear to persist through all "ON" observations and do not appear in any "OFF" observations, implying that the signal source is localized on the sky.

While these techniques are used commonly across the field, they are by no means perfect. Typical signal detection methods can struggle to detect all signals present when there are regions of time-frequency space that are densely populated, which means potential technosignatures may be missed. Furthermore, since we are fundamentally searching for a type of signal that has never been found before, it is difficult to quantify the accuracy of detection algorithms. Even the sky localization technique is not necessarily sufficient for distinguishing against radio frequency interference (RFI), which takes on many unknown morphologies and various intensity modulations.

In this thesis, we aim to push the bar forward for both signal detection and candidate identification (filtering). First, we develop a machine learning (ML) methodology for localizing narrowband signals in frequency and Doppler drift rate. Not only is this procedure faster over datasets than the standard tree detection algorithm, we train ML models to identify up to 2 signals within each stretch of 1024 frequency bins, whereas the standard algorithm can only identify 1 in the same stretch. From this work, we develop and independently present `setigen`, an open source Python library for the synthesis and injection of artificial narrowband signals into real observational data, both directly in the form of Stokes I intensities in

time-frequency space and in the form of raw complex voltages taken by baseband recorders. `setigen` can and has been used for creating large datasets used in ML training, validating detection algorithms using injection-recovery analysis, and developing new candidate filters.

Then, we propose a new candidate identification strategy based on plasma scattering from the interstellar medium (ISM). Theoretically, narrowband radio signals traveling through ionized plasma in our galaxy will exhibit strong intensity scintillations from multi-path scattering. As technosignature searches are typically tuned towards continuous narrowband signals, these scintillations should be imprinted on the received intensities and therefore detectable under the right observing parameters. Finally, we conduct a dedicated search for scintillated technosignatures towards the Galactic center and Galactic plane, for which the timescales of scintillation will be contained within individual observations. In addition to the specific scintillation analysis, we apply the sky localization filter to identify technosignature candidates. Though we do not find evidence of technosignatures, we set limits on their presence and comment about the feasibility of detection in the future.

Contents

List of Figures iv

List of Tables viii

1 Introduction 1
- 1.1 The Search for Extraterrestrial Intelligence . . . 1
- 1.2 Narrowband Signal Detection . . . 3
- 1.3 Technosignature Candidate Filters . . . 4
- 1.4 Radio Plasma Scattering . . . 5
- 1.5 Thesis Outline . . . 6

2 Narrow-Band Signal Localization for SETI on Noisy Synthetic Spectrogram Data 8
- 2.1 Introduction . . . 9
- 2.2 Data and Preprocessing . . . 11
  - 2.2.1 Noise Properties . . . 12
  - 2.2.2 Synthetic Signals . . . 14
  - 2.2.3 Dataset Construction . . . 15
  - 2.2.4 Preprocessing . . . 16
- 2.3 Methods . . . 17
  - 2.3.1 CNN Model Architectures . . . 18
  - 2.3.2 TurboSETI . . . 19
- 2.4 Results . . . 22
  - 2.4.1 Baseline vs. Final Model . . . 22
  - 2.4.2 TurboSETI vs. Final Model . . . 23
- 2.5 Discussion . . . 25
- 2.6 Summary . . . 27
- 2.7 Acknowledgements . . . 28

3 Setigen: Simulating Radio Technosignatures for the Search for Extraterrestrial Intelligence 29
- 3.1 Introduction . . . 29

3.2 Overview of Single Dish Signal Chains . . . 32
3.3 Code Methodology . . . 34
3.3.1 Spectrogram Module . . . 34
3.3.2 Raw Voltage Module . . . 40
3.4 Discussion . . . 49
3.4.1 Limitations . . . 49
3.4.2 Future Directions . . . 50
3.5 Summary . . . 51
3.6 Acknowledgements . . . 51

4 On Detecting Interstellar Scintillation in Narrowband Radio SETI 52
4.1 Introduction . . . 52
4.2 Scattering Theory and SETI . . . 55
4.2.1 Weak and Strong Scattering . . . 56
4.2.2 Effects of Strong Scintillation on Narrowband Signals . . . 57
4.3 Identifying Strong Scintillation in Detected Signals . . . 58
4.3.1 Diagnostic Statistics . . . 59
4.3.2 Constraints on Identifying Scintillation . . . 60
4.3.3 Synthesizing Scintillated Signals with Autoregressive-to-Anything (ARTA) . . . 62
4.4 Exploring the Parameter Space of ISM Scintillation with NE2001 . . . 66
4.5 Temporal analysis of detected narrowband RFI . . . 68
4.5.1 Finding and Characterizing Signals . . . 70
4.5.2 Observation Details . . . 71
4.5.3 Empirical Results . . . 73
4.6 Discussion . . . 74
4.6.1 Observational Recommendations for Scintillated Technosignature Searches . . . 74
4.6.2 The Impact of Models on Designing Observational Campaigns . . . 75
4.6.3 Building on the Analysis Pipeline . . . 76
4.6.4 Implications and Future Directions . . . 77
4.7 Acknowledgements . . . 78

5 The Breakthrough Listen Search for Intelligent Life: Galactic Center Search for Scintillated Technosignatures 79
5.1 Introduction . . . 79
5.2 Observations . . . 82
5.2.1 Scintillation Estimation with NE2001 . . . 82
5.2.2 Observing Strategy . . . 84
5.3 Methods . . . 88
5.3.1 DeDoppler Search . . . 88
5.3.2 Direction-of-Origin Filter . . . 89

5.3.3 Scintillation Diagnostic Statistics . . . 90
5.4 Results . . . 91
5.4.1 Signal Distributions . . . 91
5.4.2 RFI Analysis of NGP Observations . . . 95
5.4.3 Scintillation Likelihood Weighting . . . 96
5.4.4 Potential Candidates . . . 98
5.5 Discussion . . . 99
5.5.1 Search Sensitivity . . . 99
5.5.2 Figures of Merit . . . 100
5.5.3 Detectability of ISM Scintillation . . . 102
5.6 Conclusions . . . 103
5.7 Acknowledgements . . . 103
5.8 Appendix A: Observation Tables . . . 104
5.9 Appendix B: NGP Statistics . . . 106
5.10 Appendix C: Candidate Vetting . . . 107

6 Conclusions and Future Directions **108**

Bibliography **111**

List of Figures

1.1 Example of a radio spectrogram containing the carrier signal from Voyager 1 at X-band, showing Stokes I intensity on a logarithmic scale as a function of time and frequency. 2

2.1 Bandpass plot for Sgr B2 data over an integration time of 60 s. 12
2.2 Histogram of mean frame intensities over real GBT observation after trimming outliers, for a total of 126,419 samples. 13
2.3 Synthetic data frame with two signals, one "RFI" signal at 25 dB and zero drift, and one dimmer signal at 15 dB, normalized over the entire frame to mean 0 and variance 1. 16
2.4 Data frame containing the same data as Figure 2.3, instead normalized per frequency channel to mean 0 and variance 1. 17
2.5 Baseline model architecture for the two signal localization task. For each block, the layer type is shown, along with input and output shapes. Inputs have shape $32 \times 1024 \times 1$, normalized over all pixels. These are passed through an initial convolutional layer, followed by 3 pairs of convolutional and max pooling layers. This is followed by two fully connected layers and a dropout layer, before finally going into the output layer. For the single signal task, the last layer has 2 nodes instead of 4. 20
2.6 Final model architecture for the two signal localization task. Inputs have shape $32 \times 1024 \times 2$, combining the two normalizations described in Section 2.3.1. Residual connections are apparent between convolutional layers, followed by a batch normalization layer. This structure is repeated twice, followed by another convolution layer, two fully connected layers, and a dropout layer. For the single signal task, the output layer has 2 nodes instead of 4. 21
2.7 Box plot of RMSE in index/pixel units for the one signal dataset as a function of signal SNR. We compare the final and baseline models trained on both the full 0 – 25 dB dataset, as well as the truncated "bright" 10 – 25 dB dataset. 22
2.8 Box plot of RMSE in index/pixel units for the two signal dataset as a function of signal SNR. We compare the final and baseline models trained on both the full 0 – 25 dB dataset, as well as the truncated "bright" 10 – 25 dB dataset. 23
2.9 Box plot of RMSE in index/pixel units, comparing performance of our final model architecture with TurboSETI on the one signal dataset as a function of signal SNR. 24

2.10 Observational data frame containing an RFI signal, overlaid with localization predictions from our final CNN architecture (dashed) and TurboSETI (dotted). Although the real signal is more complex than those in our training data, the model produces reasonable predictions. 25

3.1 Radio spectrogram plots created from `setigen` frames. **A**: Frame with only synthetic chi-squared noise. **B**: Frame from panel A with an injected synthetic signal at SNR=30. **C**: "Real" GBT observation of Voyager I carrier signal at X-band. **D**: Frame from panel C with an injected synthetic signal at SNR=1000, with the same drift rate as the injected signal in panel B. 38
3.2 Basic layout of a voltage pipeline written using `setigen.voltage`. 41
3.3 Spectrogram derived from synthetic raw voltages, showing the edge of the coarse channel bandpass shape and a bright, slightly drifting cosine signal. The top panel shows an integrated profile, showing PFB scalloping loss towards the left and the synthetic signal towards the right. 47

4.1 Comparison of the Kolmogorov and square-law ACF models. Both functions are computed using a scintillation timescale of $\Delta t_d = 30$ s and a time resolution of $\Delta t = 4.65$ s. The $1/e$-height is shown as a dotted line. 59
4.2 Synthetic scintillated intensities ($N = 10^5$) generated using ARTA, using a sample interval of $\Delta t = 4.65$ s and scintillation timescale $\Delta t_d = 30$ s. **Top**: Synthetic intensity time series data, showing first 1000 samples. **Bottom left**: Histogram of intensities, showing the expected exponential distribution. **Bottom right**: Sample ACF plotted up to lag 64, with the target ACF Γ_{k} shown overlaid. . . . 61
4.3 Histograms of diagnostic statistics computed using $N = 1000$ ARTA-produced intensity time series realizations for representative scintillation timescales of 10, 30, and 100 s. Each time series is produced using $\Delta t = 4.65$ s and $\tau_{\mathrm{obs}} = 600$ s and does not include additive background noise. We plot histograms of the standard deviation, minimum, Kolmogorov-Smirnoff statistic, and least squares fit for the scintillation timescale, computed for each time series realization. 62
4.4 Comparison between methods for distance sampling, including uniformly, by stellar number density, and by stellar mass density. We use a line of sight of $(l, b) = (1, 0)$ out to a distance of 20 kpc. Bottom panel shows NE2001-produced scintillation timescales as a function of distance. 65
4.5 Set of Monte Carlo-sampled distributions of scintillation parameters at C-band, using $N = 10000$ realizations. We use a line of sight of $(l, b) = (1, 0)$ out to a distance of 20 kpc, and transverse velocities are uniformly sampled between 10 to 150 km/s. Dashed line shows median value, dotted lines show interquartile range (IQR). 66

4.6 Steps used in signal intensity analysis. **A**: Detected narrowband signal, in GBT data. **B**: De-drifted signal from panel A, with computed bounding frequencies in dashed white lines. **C**: Frame from panel B, normalized using the background noise along the frequency axis. **D**: Time series intensities computed by integrating power in panel C between the bounding frequencies and normalized to a mean intensity of 1. **E**: Sample ACF computed from panel D. 69
4.7 Histograms of diagnostic statistics for detected L-band signals with S/N≥25. For each statistic, the distribution from detected RFI is shown in black. Plotted for comparison are distributions from synthetic scintillated signals at S/N=25 with scintillation timescales of 10 s (blue), 30 s (orange), and 100 s (green). Across all diagnostic statistics, it would be difficult to distinguish a true scintillated signal from RFI given the L-band RFI distributions. 72
4.8 Histograms of diagnostic statistics for detected C-band signals with S/N≥25. For each statistic, the distribution from detected RFI is shown in black. Plotted for comparison are distributions from synthetic scintillated signals at S/N=25 with scintillation timescales of 10 s (blue), 30 s (orange), and 100 s (green). It could be possible to distinguish a true scintillated signal from RFI given the C-band RFI distributions. 72

5.1 Monte Carlo-sampled scintillation timescales for $(l, b) = (1°, 0°)$ at C-band with $N = 10^4$. The dashed line indicates the median timescale, and the dotted lines indicate the first and third quartiles. 83
5.2 Sky map of the first, second, and third quartiles for scintillation timescale Δt_d, with resolution $\Delta l = \Delta b = 0.25°$. Contours are plotted in each panel for timescales of 10, 30, 60, and 100 s. 85
5.3 Sky map of the fraction of the range 10 s $\leq \Delta t_d \leq$ 100 s covered by Monte Carlo-sampled scintillation timescales, with resolution $\Delta l = \Delta b = 0.25°$. Contours for 25%, 50%, and 75% coverage are shown. The dots show the Galactic plane targets for this survey. 86
5.4 Monte Carlo-sampled scintillation timescales for the North Galactic Pole, $(l, b) = (0°, 90°)$, at C-band with $N = 10^4$. The dashed line indicates the median timescale, and the dotted lines indicate the first and third quartiles. As expected, compared to a pointing near the Galactic center (Figure 5.1), the expected timescales are significantly longer. 86
5.5 Histograms of diagnostic statistics computed using $N = 1000$ realizations of synthetic scintillated intensity time series embedded in chi-squared radiometer noise. The synthetic observations had $\Delta f = 2.79$ Hz, $\Delta t = 2.5$ s, and $\tau = 600$ s, matching the observations taken in this study. The signals were generated with a bandwidth of 8 frequency bins (about 22 Hz) and S/N = 33. Noisy intensity time series were extracted from the synthetic observations and used to compute each diagnostic statistic. 87

5.6 Examples of signals found from the deDoppler search which passed the algorithmic event filter, but failed manual inspection. 92
5.7 Histograms of frequencies, drift rates, and S/N ratios of all detected hits in the Galactic center and plane survey. 93
5.8 Histogram of diagnostic statistics of detected hits and events throughout all Galactic center and plane observations. 94
5.9 Histogram of diagnostic statistics of detected hits and events throughout all Galactic center and plane observations, compared to the synthetic distributions. 94
5.10 Histogram of frequencies of detected hits in each NGP observation. 95
5.11 Histogram of diagnostic statistics of detected hits in each NGP observation. . . 95
5.12 Ranking estimates for all hits in detected events as a function of frequency. Synthetic ranking distributions are shown on the right panel for timescales of 10 s, 30 s, and 100 s. 97
5.13 Examples of signals found from the deDoppler search which passed the algorithmic event filter and ranked highly on the scintillation analysis. 97
5.14 Best candidates from the direction filter, which passed initial manual inspection. 99
5.15 Transmitter rate vs. EIRP for this study and previous radio technosignature searches. EIRP limits for this study at various distances for the GP targets and GC targets are plotted in blue and green, respectively. The EIRP limits for sources up to 8.5 kpc are marked in black. For each distance, we extend EIRP limits up to the right, corresponding to the maximum drift rate of 10 Hz/s searched in this study, for which $\beta \approx 0.11$. Note that while this study and Choza et al. 2023 use S/N = 33 to correct for the offset factor in `turboSETI`, we choose not to alter the results from any earlier studies which may have been affected by this. 101
5.16 Histograms of drift rates and S/N ratios of detected hits in each NGP observation. 106
5.17 Extended plots of adjacent cadences for the highest quality candidates that passed the directional filter. 107

List of Tables

2.1 Parameters for Sgr B2 data 12

4.1 Diagnostic statistics chosen to probe theoretical scintillation effects. For each statistic, we list the type of data used for computation, the theoretical behavior of that data type, and the asymptotic value of the statistic (in the absence of noise) as the observation length goes to infinity. 58

5.1 Survey Details 88
5.2 Survey Target List 104
5.2 Survey Target List 105

Acknowledgments

First, a huge thanks to my advisor over the last 6 years, Andrew Siemion. Beyond his intuition and genuine passion for SETI, I feel as though I've learned and grown so much more beyond science from Andrew. He made sure to be available at all times but at the same time allowed me to explore and pursue research avenues that I was genuinely interested in. His direction and support through various talk opportunities and the occasional media outlet query together make up the kind of non-technical experiences I'll be able to take with me wherever. I truly couldn't have asked for a better mentor.

Thank you to Imke de Pater for her guidance over the last few years. Our weekly meetings were truly important to me, not only for the science we discussed, but also for advice on navigating academia and progressing in my degree. I am grateful to the remaining members of my qual and thesis committees, Aaron Parsons and Steven Beckwith.

I would also like to thank Shelley Wright for inviting me to weekly meetings at UCSD for the last couple years and exposing to the optical side of SETI.

I thank the Breakthrough Listen Foundation for the funding and providing so many great opportunities to meet and collaborate with other SETI scientists, especially early-career researchers. Thank you to those involved in the group at various stages of my Ph.D.: Steve Croft, Matt Lebofsky, Howard Isaacson, Dave MacMahon, Vishal Gajjar, Danny Price, Brian Lacki, Dave DeBoer. Thanks also to Barbara Hoversten, Karen Aguilar, Amber Banayat, and Yasasha Ridel for all the help with the ins and outs on the administrative side of research.

A special shoutout to all the great early-career SETI scientists I've had the pleasure of meeting and spending time with, whether during Breakthrough events or otherwise: Sofia Sheikh, Karen Perez, Carmen Choza, Owen Johnson, Peter Ma, Evan Sneed, Luigi Cruz, Barbara Cabrales, and many others.

I would also like to thank my mentor, Aleks Navratil, during my data science internship back in 2019. While I obviously ended up staying to complete my degree, I appreciated your perspectives on academia and industry and navigating the two. More than anything, I enjoyed your mentoring style and had a great time working with you.

Thanks to my buddies from undergrad: Chris, Alan, and Emi. Even though for the last 6 years I've been 3 hours separated from everyone else, I'm so glad we were able to get together virtually so often. Whatever the future holds, I hope we can all keep it going.

I thank my family: Mom, Dad, Brendan, and Brett. Thanks to my parents for supporting me throughout college and into my Ph.D., especially since it seems like I've been bouncing from coast to coast for practically my whole life. Brendan, thanks for the gaming sessions and impromptu watch parties – talking sports and film with you has been such a great part of this chapter in my life. Brett, it's been fun chatting with you about the most random things, but mark my words, eventually I'll get you to care about sports too. Go Birds!

Lastly, but most importantly, I thank my partner Elizabeth for everything, really. I'm so glad we were able to navigate undergrad, COVID, and being apart for the first 4 years of my degree before I was able to work out of San Diego with you. Here's to the next chapter of our lives together!

Chapter 1

Introduction

1.1 The Search for Extraterrestrial Intelligence

Since antiquity, humans have wondered whether there is life out there, among the stars. It remains one of the most profound questions about our universe and our place in it.

The Search for Extraterrestrial Intelligence (SETI) seeks to answer this question by specifically searching for so-called "technosignatures," evidence of the existence of alien technology. Cocconi and Morrison 1959 first proposed searching for radio technosignatures around the 21 cm neutral hydrogen line, which they argued was a natural choice of wavelength for transmitting civilization that wanted to be detected. Drake 1961 carried out the first SETI search based on these ideas at the Green Bank Observatory in West Virginia, targeting narrowband signals within 400 kHz of the 21 cm line (1.42 GHz) using a single-channel receiver of bandwidth 100 Hz. Many decades later, technological developments in radio telescopes and computing hardware have enabled modern searches many magnitudes larger in scope (Werthimer et al. 1985; Tarter 2001; Siemion et al. 2013; Hickish et al. 2016; Worden et al. 2017; Lebofsky et al. 2019; Zhang et al. 2020). In this work, we focus on observations taken by single-dish telescopes, but in modern radio SETI, array telescopes and interferometers are being used in an ever greater capacity (Parsons et al. 2008; Tarter et al. 2011; Rampadarath et al. 2012; Harp et al. 2016a; Tingay et al. 2018a; Czech et al. 2021).

When a radio signal is received by an antenna, the wave induces a current in the antenna, which can be converted into a voltage and recorded digitally. Fundamentally, a radio telescope directs radio waves into a feed antenna at the focus of the dish, increasing the effective collecting area and angular resolution. At the Green Bank Telescope (GBT) and other single dish telescopes, the input voltages are commonly coarsely channelized by a polyphase filterbank (PFB) to obtain complex baseband voltages with a better spectral channel response than a simple Fast Fourier Transform (FFT; Bellanger et al. 1976; Parsons et al. 2006; Prestage et al. 2015; MacMahon et al. 2018; Price 2021). Many radio observatories perform this front-end processing on field-programmable gate array (FPGA) boards developed by the Collaboration for Astronomy Signal Processing and Electronics Research (CASPER; Hickish

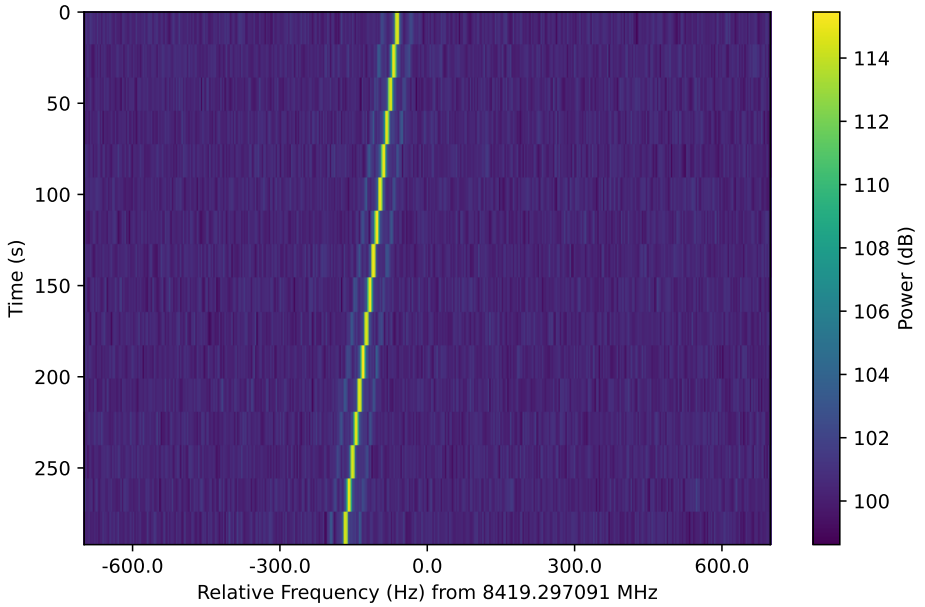


Figure 1.1: Example of a radio spectrogram containing the carrier signal from Voyager 1 at X-band, showing Stokes I intensity on a logarithmic scale as a function of time and frequency.

et al. 2016).

From the complex voltages, we can finely channelize further and reduce the data to create Stokes I intensities as a function of time and frequency. This data format has a few different names, including *radio spectrogram*, *dynamic spectra*, and *waterfall plot*. The intensity data is arranged in a 2-dimensional array, which can be thought of as a stack of consecutive spectra. Figure 1.1 shows an example spectrogram containing the carrier signal of Voyager 1, detected at X-band using the GBT. We visualize spectrogram data using a logarithmic colorbar and note that the spectrogram is essentially comprised of 16 consecutive spectra arranged vertically.

When a narrowband signal has a high duty cycle, it will appear as a line of persistent power with a vertical path through a spectrogram. If the signal has constant power and is transmitted without frequency modulation, we expect to observe a vertical line of constant intensity. However, due to the Doppler effect, a signal emitted from a source with relative velocity will appear at a different frequency. For a source with some relative *acceleration* compared to the receiver, such as from a planet's rotation or orbit, at each successive spectrum contained within the spectrogram, the signal will be Doppler shifted by a different amount, an effect called Doppler acceleration. If the observation length is short compared to the periodicity of acceleration, the signal will appear to be linearly drifting in frequency over time, as in Figure 1.1. When plotted, drifting signals will have an apparent slope corresponding to its *Doppler drift rate*.

In this thesis, we work mainly with observational data in spectrogram format. While the loss of phase information necessitates that signal detection must be done incoherently, most SETI searches use spectrogram data out of practicality, since raw voltage formats take up much more data storage and are correspondingly computationally expensive to analyze. In addition, most detection algorithms are a form of energy detection and do not require phases.

1.2 Narrowband Signal Detection

The first step of any radio technosignature search is the detection of raw signals. The parameter space of potential technosignatures is extremely vast, spanning properties such as the central frequency, Doppler acceleration, time of emission, bandwidth, and more (Wright et al. 2018). We do not have a priori knowledge of any of these parameters, so it is important to push the limits of sensitivity and collect as many signals as possible so that we do not miss a true technosignature, should it exist.

The standard detection method used in narrowband SETI is called the tree deDoppler algorithm, which efficiently implements an incoherent matched filter for linearly Doppler accelerated signals (Taylor 1974; Siemion et al. 2013). For every detected signal, the algorithm obtains the frequency, drift rate, and signal-to-noise (S/N) ratio. `TurboSETI` is a Python implementation used by Breakthrough Listen searches (Enriquez et al. 2017; Enriquez and Price 2019). While the deDoppler method is designed to specifically find high duty cycle (always "on") signals, since the underlying algorithm integrates along various trial drift rates, even complex signals with time and frequency modulations are detected and reported. As a result, this has been used in most modern SETI searches, since it is relatively robust to the unknown morphologies found in radio interference.

Other detection methods have been explored, though they typically do not have the same level of detail in characterizing signal properties. For example, algorithms have been proposed to find signals by detecting spikes in "energy" within radio spectrograms, such as by computing the spectral kurtosis and showing that it is inconsistent with pure radiometer noise (Nita et al. 2016). In a similar manner, machine learning (ML) algorithms have been designed to identify the presence of signals in noise (Zhang et al. 2018b). The model of choice for spectrogram analysis are convolutional neural networks (CNNs), since the 2-dimensional structure of spectrograms naturally lends itself to techniques from image analysis (Krizhevsky et al. 2012; He et al. 2016; Zhang et al. 2018a). On the one hand, these methods only report the existence of signals and not detailed parameters about the signals themselves, such as a fitted drift rate. On the other hand, if signals are detected in a snippet of spectrogram data, there is a rough localization implicit to the frequency range of the spectrogram itself, and a follow-up analysis to determine signal properties can then by conducted.

There are a few scenarios in which existing detection methods still struggle. For example, dim signals are all but eclipsed by bright signals close in frequency. This is especially dangerous, since one would expect emission from ETI, traveling pc to kpc, to have lower S/N levels

compared to terrestrial interference. In addition, it is computationally expensive to search for high drift rates. Sheikh et al. 2019 suggested that SETI searches should attempt to detect signals with drift rates of up to 200 nHz = (200 Hz/s)/1 GHz. Even drift rates of 10 Hz/s can smear signal power across adjacent frequency channels in high spectral resolution data, leading to an associated loss of detection power using naive integration-based S/N estimates (Gajjar et al. 2021). These are areas in which new algorithms and improvements to existing ones can make a real impact, and deep ML methods could play an important role in making detection pipelines efficient and robust.

1.3 Technosignature Candidate Filters

Radio emission comes in all shapes and sizes – in all spectral and temporal scales – and from natural and artificial sources alike. For narrowband SETI, we largely detect human-created radio frequency interference (RFI), since they have many of the same qualities we search for in technosignatures. Even though radio telescopes are generally in areas that attempt to limit terrestrial radio emission, such as the Green Bank Telescope (GBT) in the National Radio Quiet Zone, RFI still makes up the vast majority, if not all, of the detected signals (Maan et al. 2021). Therefore, filters are extremely important in order to remove RFI and focus on the most interesting detections. There are many low to high-level filters that are applied to search for bona fide technosignatures. The best, or most convincing, SETI candidates are those that pass many well-designed filters.

At a very high level, the selection criteria used to gather targets to comprise a survey provides an implicit filter. For example, a search of nearby stars is prudent because we should be able to detect dimmer signals at higher S/N ratios (Isaacson et al. 2017; Enriquez et al. 2017). Likewise, a search of stars that could detect Earth transiting the Sun is intuitive because ETI residing in or around planets orbiting those stars would know of Earth's existence and perhaps of its favorable conditions for life (Sheikh et al. 2020). Any signals detected in these kinds of surveys are naturally interesting by virtue of the higher-level target selection.

At the other extreme, a detected signal being narrowband (under 1 kHz) is itself a filter from natural sources, as far as we know, since astrophysical emission typically exhibits natural and thermal broadening (Cordes et al. 1997). Likewise, substantial non-zero Doppler drift rates are rare for anthropogenic emission, so filtering out signals with zero drift will remove those that are likely RFI.

The most common and perhaps singularly most trusted technique is called the direction-of-origin filter (or the ON-OFF filter). This method attempts to localize detected signals on the sky by taking observations via position switching. Representing unique targets as letters, where the primary target is A: observations are generally taken in cadences of ABAB, ABABAB, or ABACAD, depending on the survey parameters. If the other targets are separated by multiple full width at half maximum (FWHM) beamwidths, any signal that is truly originating from target A will go down by multiple orders of magnitude in the so-called "OFF" observations. In other words, if a signal is only detected in ON observations and not

detected in any OFF observations, we call that signal localized on the sky, and it becomes a compelling technosignature candidate.

Candidate filters used in modern SETI do not use actual signal morphology as a discriminant from RFI. There are a few reasons for this. First, narrowband SETI has typically focused on detecting ideal high duty cycle signals with no intrinsic time or frequency modulation, since such signals will be the most straightforward to detect and discriminate from RFI using direction-of-origin filters. Next, if an ETI civilization transmits in our direction, it is possible that information is encoded by some kind of intensity modulation in time-frequency space, but we do not know what that modulation might look like. Finally, and perhaps most importantly, we still do not have a detailed understanding of the empirical intensity modulation present in RFI, enough to accurately classify signals as types of RFI. Nevertheless, in this thesis, we explore the possibility of observing technosignatures that exhibit natural intensity modulations due to multipath scattering from ionized plasma encountered on their journey towards Earth from elsewhere in the Galaxy.

1.4 Radio Plasma Scattering

Radio waves are scattered by ionized plasma due to their interaction with free electrons. This effect has been observed in many pulsars and much of scattering theory has come about in order to characterize pulsar scattering from the ISM (Scheuer 1968; Roberts and Ables 1982; Narayan 1992).

When a radio wave enters a region with a different free electron number density, with a correspondingly different refractive index η, the wave is refracted and imparted with a phase shift $\Delta\phi \sim k\Delta\eta\Delta z$, where $k = 2\pi/\lambda$ is the spatial wavenumber, $\Delta\eta$ is the change in refractive index, and Δz is the thickness of the thin scattering screen (Thompson et al. 2017; Coles et al. 2010). As the wave travels through the turbulent ionized plasma, the phase shifts can be described by a power spectrum related to the scattering medium, commonly taken to be that of Kolmogorov turbulence. The distribution of phase fluctuations on the sky can be described by the phase structure function

$$D_\phi(\mathbf{s}) = \langle[\phi(\mathbf{r}+\mathbf{s}) - \phi(\mathbf{r})]^2\rangle_{\mathbf{r}}, \tag{1.1}$$

where $\mathbf{s}$ is a vector baseline in the scattering plane (Narayan 1992). For Kolmogorov turbulence, $D_\phi(\mathbf{s})$ follows a power law with exponent $\alpha = 5/3$ (Rickett 1990). The phase structure function gives the mean squared phase difference for a baseline $\mathbf{s}$, and therefore is useful in comparing length scales on the scattering screen.

For instance, the Fresnel scale $r_F \sim \sqrt{\lambda d}$ is the radius of the largest cross-section for which radio waves arrive coherently (path-induced phase delays below π) along the line-of-sight distance d. If $D_\phi(r_F) \ll 1$, we are in the weak scattering regime, since waves that would be coherent in free space have small phase fluctuations. If $D_\phi(r_F) \gg 1$, we are in the strong scattering regime, in which phase fluctuations within the Fresnel scale are large.

In this case, different propagation paths pick up uncorrelated stochastic phase changes, and the interference between them causes diffractive scattering.

The received power from the original radio wave will ultimately resemble a diffraction pattern at the observer plane, imposed by the scattering medium. In the case of strong scattering, multipath propagation will lead to nulls and peaks in accordance with Rayleigh-distributed amplitudes (Goodman 1975). As the source, scattering screen, and receiver move relative to each other in the transverse direction, the diffraction pattern will "sweep" across the receiver with velocity V_T. For a single dish telescope, this will result in apparent intensity scintillations over time.

Cordes and Lazio 1991 specifically analyzed the implications that this physical effect would have on the detectability of narrowband technosignatures in our Galaxy. They note that received scintillated intensities will be lower than the original intensity most of the time, but occasionally will be many times higher from constructive interference, so SETI searches should repeat targets at multiple observation epochs to maximize the likelihood of detection. In Chapter 4, we propose that this effect is resolvable on relatively short observational timescales and as such can be used as a discriminant from RFI. In Chapter 5, we apply our proposed techniques in a technosignature search towards the Galactic center and Galactic plane.

1.5 Thesis Outline

This thesis follows the chronological journey I took along the course of my research in narrowband signal detection and developing new analysis methods based on ISM scintillation.

While machine learning had been used previously in SETI to classify snippets of radio intensity spectrogram data (mainly artificial data), such methods do not provide details of detected signals nor account for the presence of multiple signals within a single frame of data. In Chapter 2, we use convolutional neural networks (CNNs) to localize narrowband signals in radio spectrograms. As a step for ultimately supporting the simultaneous localization of many signals within a data snippet, our CNN models attempt to localize up to two signals per snippet. Since there is no high-quality human-labeled narrowband dataset that provides localization parameters down to the specific frequency bins, we created an open-source Python library called `setigen` for the generation and injection of artificial narrowband signals into real observational data.

Initially, `setigen` created small snippets of synthetic data in a manner tailored for use in machine learning datasets. However, we quickly realized that this tool could be expanded and used in all facets of the Breakthrough Listen project, such as for validation and injection-recovery analysis of our detection algorithms and in the development of new analysis methods. In Chapter 3, we present `setigen` as a polished library for synthetic narrowband signal generation as both Stokes-I intensities and raw antenna voltages. We describe the inner workings of the library, including a detailed discussion of typical signal processing chains used for single dish radio telescopes, such as the GBT.

In Chapter 4, we present a new filter for technosignature candidates based on detecting ISM scintillation in narrowband signals. We demonstrate how one may estimate likely scintillation timescales towards targets of interest using the NE2001 free electron density model (Cordes and Lazio 2002), and offer methods for creating synthetic scintillated intensity time series to test detection algorithms. We identify summary statistics that probe the expected stochastic behavior of intensity scintillations and perform a limited analysis of RFI detected with the GBT at C-band in order to test the feasibility of this method. We find that there are regions of the parameter space in which scintillated candidates should be separable from detected signals in the RFI environment.

In Chapter 5, we conduct a dedicated search towards the Galactic center and Galactic plane for scintillated technosignatures. We use the timescale estimation procedure developed in the Chapter 4 to design an observing plan that targets the most detectable scintillation timescales under observation parameters close to those typically used in narrowband searches. We apply both the typical ON-OFF directional filter as well as our scintillation analysis methodology to identify candidate signals. Ultimately, we do not find evidence of technosignatures that stand up to manual inspection, scintillated or otherwise. We comment on the signal statistics of detected RFI at the C-band over 5 observational epochs and discuss limits on the presence of technosignatures towards the Galactic center.

Chapter 2

Narrow-Band Signal Localization for SETI on Noisy Synthetic Spectrogram Data

A version of this chapter was originally published as: Brzycki, B., Siemion, A.P., Croft, S., Czech, D., DeBoer, D., DeMarines, J., Drew, J., Gajjar, V., Isaacson, H., Lacki, B., Lebofsky, M., MacMahon, D.H.E., de Pater, I., Price, D.C., and Worden, S.P. 2020. Narrow-band signal localization for SETI on noisy synthetic spectrogram data. Publications of the Astronomical Society of the Pacific, 132(1017), p.114501.

As it stands today, the search for extraterrestrial intelligence (SETI) is highly dependent on our ability to detect interesting candidate signals, or technosignatures, in radio telescope observations and distinguish these from human radio frequency interference (RFI). Current signal search pipelines look for signals in spectrograms of intensity as a function of time and frequency (which can be thought of as images), but tend to do poorly in identifying multiple signals in a single data frame. This is especially apparent when there are dim signals in the same frame as bright, high signal-to-noise ratio (SNR) signals. In this work, we approach this problem using convolutional neural networks (CNN) as a computationally efficient method for localizing signals in synthetic observations resembling data collected by Breakthrough Listen using the Green Bank Telescope. We generate two synthetic datasets, the first with exactly one signal at various SNR levels and the second with exactly two signals, one of which represents RFI. We find that a residual CNN with strided convolutions and using multiple image normalizations as input outperforms a more basic CNN with max pooling trained on inputs with only one normalization. Training each model on a smaller subset of the training data at higher SNR levels results in a significant increase in model performance, reducing root mean square errors by at least a factor of 3 at an SNR of 25 dB. Although each model produces outliers with significant error, these results demonstrate that using CNNs to analyze signal location is promising, especially in image frames that are crowded with multiple signals.

2.1 Introduction

Many avenues in the search for extraterrestrial intelligence (SETI) are largely reliant on our ability to pick out interesting signals in a sea of optical and radio telescope data. Since the 1960s, radio searches for evidence of extraterrestrial intelligence (ETI) have increased in scope in tandem with our improving technology, covering larger instantaneous bandwidths and surveying more targets than before (Drake 1961; Werthimer et al. 1985; Horowitz et al. 1986; Korpela et al. 2001; Welch et al. 2009; Siemion et al. 2013; Wright et al. 2014; MacMahon et al. 2018; Price et al. 2018).

The Breakthrough Listen (BL) initiative is the most thorough SETI search effort, with access to top radio telescopes across the world specifically for use in SETI searches, including 20% of the telescope time on the Green Bank Telescope (GBT) in West Virginia, USA and 25% time on the CSIRO Parkes radio telescope in New South Wales, Australia (Worden et al. 2017; Isaacson et al. 2017; MacMahon et al. 2018; Price et al. 2018). In optical wavelengths, the search uses the Automated Planet Finder at the Lick Observatory in California, USA (Vogt et al. 2014). The BL search has expanded to include such facilities as the MeerKAT telescope in South Africa (Jonas 2009), the VERITAS Cherenkov Telescope at the Whipple Observatory in Arizona, USA (Weekes et al. 2002), the Murchison Widefield Array in Western Australia (Tingay et al. 2018b), and the FAST telescope in Guizhou Province, China (Zhang et al. 2020). Sifting through the sheer data volume collected, which can be on the order of hundreds of terabytes per day, is computationally expensive alone, but identifying interesting, anomalous signals is itself a tough open problem.

Most of the coherent radio signals that we observe in BL data are anthropogenic, termed radio frequency interference (RFI). Types of RFI include satellite telemetry, cellular mobile broadcasts, wireless internet, and a host of other artificial sources. These are all types of narrow-band signals, which means each signal has a small frequency bandwidth (generally of order less than 1 kHz). On the other hand, natural astrophysical phenomena usually produce broad-band signals. The challenge for technosignature searches is that if an intelligent civilization is producing signals at radio frequencies (technosignatures), either as directed transmissions or as by-products of advanced technology, these signals are also likely to be narrow-band and therefore appear similar to RFI. SETI searches to date have found mountains of RFI signals, but no conclusive evidence of technosignatures (Tarter 2001; Korpela et al. 2011; Siemion et al. 2013, 2014; Harp et al. 2016b; Enriquez et al. 2017; Gray and Mooley 2017; Tingay et al. 2018b; Wright et al. 2018; Price et al. 2020).

The science data we collect from radio telescopes are generally stored as arrays of detected intensity (Stokes-I) as a function of time and frequency. These can be visualized as dynamic spectra or "waterfall plots," with frequency on the x-axis, time on y-axis, and intensity as a color according to a colorscale. In other words, each pixel corresponds to an intensity value computed at that specific frequency and time. Narrow-band signals that are "on" for the duration of a short observation appear as lines across waterfall plots, which may be sloped due to the relative motion between the celestial source and the telescope, the so-called Doppler acceleration (Sheikh et al. 2019). If a signal is bright enough, it is easily

distinguishable by the human eye. However, it is simply impossible to visually inspect all the data we collect, which easily spans billions of frequency channels (Lebofsky et al. 2019).

Our standard narrow-band signal search method uses TurboSETI[1], an implementation of the "tree deDoppler" algorithm, which effectively averages along potential Doppler drift rates (slopes) in a spectrogram and searches for statistically high spikes in the resulting spectra (Taylor 1974; Siemion et al. 2013; Enriquez et al. 2017; Enriquez and Price 2019). If one picks the correct drift rate and there is a signal at that rate, one should get a detection, since averaging reduces the impact of random noise and preserves the signal. While the underlying tree-based algorithm is more efficient than a naive search over all drift rates, this approach requires many passes over the data and potentially misses fainter signals masked by bright RFI (Pinchuk et al. 2019).

A complementary parametric algorithm for localizing narrow-band signals is the Hough transform, an edge-detection technique that translates an image into another 2D representation whose features correspond to edges in the original image (Hough 1959; Barinova et al. 2012). Applying this transform to Stokes-I data and identifying bright features allows for the detection and localization of narrow-band signals, which manifest as "edges" in the data (Monari et al. 2006; Fridman 2011). This method also requires many passes over the data, however, and the features must still be extracted from the resulting transform (e.g. via thresholding).

Another method for detecting signals in radio data is to analyze the degree to which the data differs from an ideal statistical distribution, assuming only noise is present. For instance, higher order statistics such as the kurtosis can indicate that a portion of data differs significantly from an ideal Gaussian distribution. Applying this principle by calculating the kurtosis for time series voltage data or the spectral kurtosis for dynamic spectra can signal the presence of RFI (Ruf et al. 2006; Nita et al. 2016). Since these are relatively simple calculations, they can be done in real-time to flag or even mitigate RFI during observations.

While these approaches each have their own strengths, we would like to evaluate the effectiveness of machine learning (ML) methods in accurately identifying narrow-band signals, especially in the presence of bright RFI. Having a good signal localization and detection pipeline is crucial for identifying signals that are currently overlooked using conventional signal processing methods.

Advances in computer vision techniques, especially with convolutional neural networks (CNN), have proven quite effective in classification and object detection tasks (Krizhevsky et al. 2012; Simonyan and Zisserman 2014; Szegedy et al. 2015; Ren et al. 2015; He et al. 2016). For radio signal processing, CNNs have been used to classify radio transmissions based on their modulation schemes, using time series voltage data (O'Shea et al. 2018). In fact, Stokes-I data produced from radio observations is similar in structure to an image, so the data lends itself readily to computer vision techniques.

For instance, Zhang et al. 2018a used CNNs on Stokes-I data to detect new pulses from the fast radio burst FRB 121102. They created a dataset of synthetic FRB pulses and trained

[1] https://github.com/UCBerkeleySETI/turbo_seti

a classifier to detect whether or not a data frame contained a pulse. Using this model on real observations, they were able to find 72 new pulses within a 5 hour radio observation.

Similarly, Harp et al. 2019 created a synthetic dataset of radio spectrograms to resemble data taken by the Allen Telescope Array, inserting 6 different classes of narrow-band signals into an artificial noise background. They compared the performance of various deep CNN architectures in classifying the synthetic data frames, and found that their ML classifiers performed well for signals with relatively high signal-to-noise ratios (SNR).

Modern object detection methods such as You Only Look Once (YOLO; Redmon et al. 2016) use clever ways to quickly determine an arbitrary number of object bounding boxes in images. Even so, object detection and localization of long, thin objects remain particularly difficult. It is hard to draw meaningful bounding boxes around them, since such objects generally comprise only a small portion of bounding box areas, making it impractical to maximize the intersection over union measure with ground truth. In addition, since many radio signals can intersect at any position, it is harder to similarly split up an image frame into a coarse grid and only associate one signal with each grid cell, as in YOLO. This makes it especially difficult to detect an arbitrary number of signals in a frame. For this reason, we limit our present work to signal localization, in which we attempt to precisely predict the positions of a known number of signals in each image frame.

In this work, we investigate the effectiveness of machine learning signal localization on synthetic radio spectrogram data. We run experiments using CNN architectures and evaluate performance based on the root mean square error between true and predicted pixel locations as a function of signal intensity or SNR. We further compare these localization results with signal detections from TurboSETI. We conclude with future directions for improving signal localization and ultimately moving towards true object detection.

2.2 Data and Preprocessing

The SETI goal of looking for interesting signals in observations makes it difficult to get a large labeled dataset. To that end, it is an open question as to what sort of labels make the most sense – there are so many different forms and patterns in human RFI that results would be highly dependent on the number and nature of classes. Furthermore, manual inspection can be ineffective in identifying lower intensity signals (whereas averaging along various drift rates can increase the SNR and thus reveal dimmer signals).

To test the sensitivity and accuracy of signal search procedures, we generated a set of synthetic observations that resemble real data from the GBT. In general, the Breakthrough Listen instrument at the GBT takes data over a range of frequencies (over a large bandwidth of a few GHz) at the same time (MacMahon et al. 2018). Here, we focus on scientific data products that have a 1.4 Hz spectral resolution and a 1.4 second temporal resolution, at a frequency range of 4 – 8 GHz (C-band).

For this work, we analyzed image frames that are 32×1024 pixels – 32 time samples tall and 1024 frequency samples wide. This effectively spans a total range of about $32 \cdot 1.4$ s ≈ 45

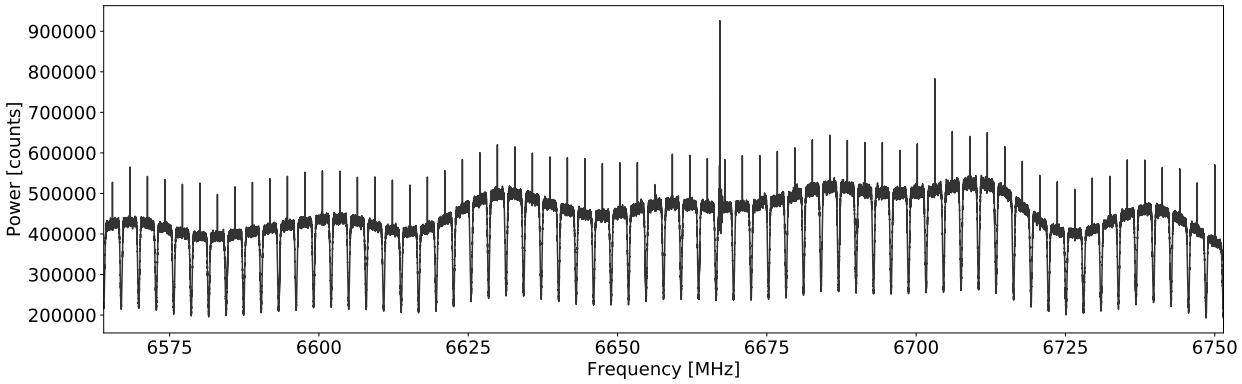


Figure 2.1: Bandpass plot for Sgr B2 data over an integration time of 60 s.

Parameter	Sgr B2
RA (J2000)	17^h 47^m 15.0^s
Dec (J2000)	-28° 22' 59.16"
Initial MJD	58465.71709
C-Band Frequency Coverage	6564 – 6752 MHz
Frequency Resolution	1.39698 Hz
Time Resolution	1.43166 s
Integration Time	60 s

Table 2.1: Parameters for Sgr B2 data

s and $1024 \cdot 1.4$ Hz ≈ 1430 Hz. Although our observations easily span billions of frequency channels, for practical reasons, we limit the number of frequency channels per frame to better facilitate the use of CNNs.

2.2.1 Noise Properties

Since the background radiometer noise in time-series voltage data closely follows a zero-mean Gaussian distribution, the noise in Stokes-I data ultimately follows a chi-squared distribution (McDonough and Whalen 1995; Nita et al. 2007; Thompson et al. 2017). However, due to instrumental effects such as coarse channel bandpass shapes and natural variations in detector sensitivity as a function of frequency, the raw intensity values we get from observations can vary appreciably.

To properly capture these intrinsic intensity variations in our synthetic data, we based their noise properties on actual data from the GBT. We used 4 – 8 GHz observations of

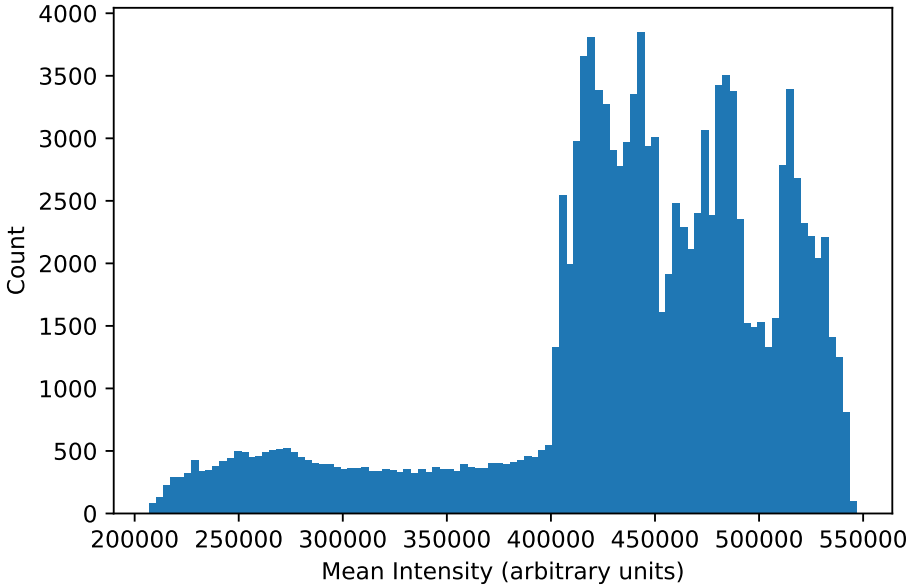


Figure 2.2: Histogram of mean frame intensities over real GBT observation after trimming outliers, for a total of 126,419 samples.

Sgr B2 taken on Dec 13, 2018 at 17:12:37 UTC. We found that using a smaller range of frequencies, 6564 – 6752 MHz, was sufficient for obtaining realistic background intensity values for the synthetic frames (Table 2.1).

At the GBT, the Versatile Green Bank Spectrometer (VEGAS; Prestage et al. 2015) digitizes and coarsely channelizes data using a polyphase filterbank. The data is then sent to the Breakthrough Listen data recorder system (MacMahon et al. 2018), which applies finer channelization to each coarse channel and records the resulting high spectral resolution data products.

Figure 2.1 shows the integrated bandpass plot of our observational data. Visible in the spectrum are the 64 coarse channels present in our data slice, which are about 3 MHz in width and characterized by intensity fall-offs on either edge. The spike at the center of each coarse channel is the so-called "DC bin," the sum of all samples within the channel, which arises from the Fourier Transform-based filterbank. The large spike at ∼6670 MHz is bright RFI. The overall bandpass shape reflects the inherent variation in sensitivity across the receiver.

We split this data into frames of size 32×1024 and calculate the mean intensity of each individual frame. Since some frames contain DC bins and bright RFI that bias these intensities toward higher values, we trim the resulting collection of mean intensities using sigma clipping with limits corresponding to 5σ. Figure 2.2 shows a histogram of these intensities after cutting out outliers, for a total of 126,419 remaining frames. Together with

Figure 2.1, we notice that the majority of mean intensities are concentrated between 4×10^5 and 5.5×10^5. The tail extending down to intensities of 2×10^5 is due to the lower sensitivities at the edges of each coarse channel. Note that each frame only covers ∼1.4 kHz in frequency, which is small in comparison to coarse channels (∼3 MHz). So, we make the assumption that larger scale systematic bandpass effects are not important within individual frames.

To initialize each synthetic data frame with noise, we randomly select from this empirical distribution to select a desired mean intensity μ_{noise}. To calculate the degrees of freedom for our chi-squared noise background, we note that our Stokes-I data uses two polarizations and that complex Fourier coefficients contribute both real and imaginary terms. These result in 4 degrees of freedom for every integration, so overall, the underlying distribution has a total of $k = 4 \cdot df \cdot dt$ degrees of freedom, where df and dt are the frequency and time resolutions of the data. Since the mean of a chi-squared distribution is k, we randomly sample values from this distribution to populate the empty data frame, and then scale every value up by a factor of μ_{noise}/k to match the desired mean intensity. Comparing the results to our observations, we find that this procedure indeed reproduces the noise distributions found in real data. The benefit in having a method for generating purely synthetic yet realistic background noise is that every data frame thus created is guaranteed to be free from signals of any kind, which helps in accurately evaluating signal search strategies.

2.2.2 Synthetic Signals

Narrow-band signals found in our radio frequency data come in many forms, with temporal and spectral structures of varying complexity. For example, modulation schemes in prolonged radio transmissions result in intensity variations over time, and some signals are emitted in short pulses in the first place (Sokolowski et al. 2015). Depending on their location in the galaxy and in the sky, narrow-band signals may also be subject to scintillation from the interstellar and interplanetary medium (Rickett 1977; Lotova et al. 1985; Cordes et al. 1997; Siemion et al. 2013; Price et al. 2019a). Even a constant amplitude sine wave signal in the time domain, in the presence of zero-mean Gaussian noise, follows a non-central chi-squared distribution of intensities in Stokes-I data (McDonough and Whalen 1995). Examples of various RFI morphologies are presented in Sheikh et al. 2020.

Despite the prevalence of complex and noisy narrow-band signals, for this work, we choose to create synthetic signals with constant intensity over time as heuristic models for real signals. Indeed, this makes the assumption that if an ML model can accurately localize these ideal signals, it will also properly localize more complex signals. Our intuition behind this assumption stems from the fact that search techniques such as TurboSETI will still find noisy signals, even though they are optimized for finding "simple" signals.

We developed a software package, Setigen[2], to facilitate the creation and injection of synthetic narrow-band signals into observational data frames. Based on the time and frequency resolutions of the Stokes-I data, Setigen can also calculate the corresponding idealized chi-

[2] https://github.com/bbrzycki/setigen

squared distribution for the background noise, into which synthetic signals can be added, as described in Section 2.2.1. This allows us to create large datasets of synthetic data frames for training and validating signal search pipelines.

We define the "start" of a signal as the index (or pixel) in the frequency direction where the center of the signal is during $t = 0$ in an image frame, and the "end" as the center position during the 32nd time sample. We randomly choose the starting and ending indices for each signal and the width of the signal in the frequency direction (limited to a narrow-band range). Starting and ending indices are always between 0 to 1023, inclusive. Accordingly, the maximum absolute drift rate, corresponding to starting and ending indices on opposite ends of the frequency range, would be about 31 Hz/s.

Because we would like to analyze the effectiveness of our machine learning algorithms on different SNR levels, we scale the intensity of each synthetic signal according to the desired SNR level, the background noise level, and the number of time samples:

$$\text{SNR} = \frac{I_{\text{signal}}}{\sigma_{\text{noise}}} \times \sqrt{n_t}, \tag{2.1}$$

where I_{signal} is the appropriate intensity of the injected constant signal at any given time sample, σ_{noise} is the standard deviation of the background noise, and n_t is the number of time samples (in this case, $n_t = 32$).

This definition is used so that the expected SNR matches the measured SNR if we had simply averaged through each time sample shifted at the correct drift rate, which is how current Doppler drift search pipelines, such as TurboSETI, work (Enriquez et al. 2017).

2.2.3 Dataset Construction

We generate two datasets to test signal localization, each with 120,000 training samples and 24,000 test samples. Since the signals we are interested in potentially span a large range of intensities, we specify SNR levels for our synthetic narrow-band signals using decibels, such that 0 dB $\rightarrow$ 1σ, 20 dB $\rightarrow$ 100σ, etc.

Our first dataset contains 32×1024 image frames with exactly one signal at SNR levels of 0, 5, 10, 15, 20, and 25 dB. So, for each SNR level, we generate 20,000 training frames and 4,000 test frames. For each frame, we also save the starting and ending indices (2 numbers) as labels.

Our second dataset contains frames with exactly two signals. One of the signals is 25 dB and at a zero drift rate, so that it is at a constant frequency at all time samples. This is meant to represent a typical RFI-like signal. The other signal is at SNR levels of 0 – 25 dB as in the first dataset. We save the starting and ending indices for both signals (4 numbers). Note that these labels *are* ordered, even though there is no preferred order in any given image frame.

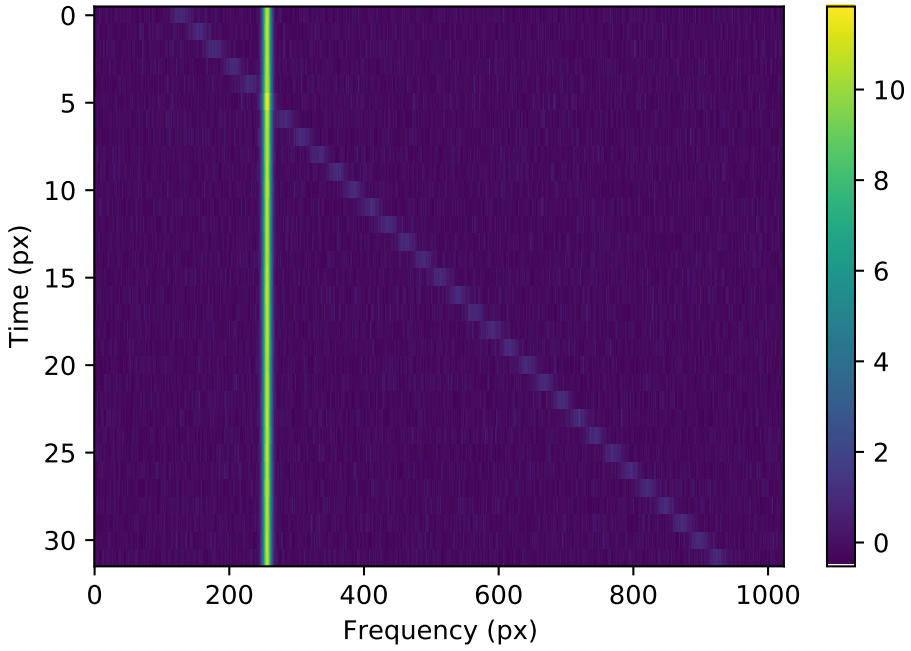


Figure 2.3: Synthetic data frame with two signals, one "RFI" signal at 25 dB and zero drift, and one dimmer signal at 15 dB, normalized over the entire frame to mean 0 and variance 1.

2.2.4 Preprocessing

By design, for each dataset, we generate 120,000 training frames and 24,000 test frames in total, the latter of which are used only for final evaluation. During training, we take a 80/20 random split of the training frames for training/validation.

The choices of normalization are very important for our data, which can exhibit regions of high contrast and varying instrument sensitivity. We choose to normalize our labels (indices between 0 and 1023 inclusive) to be between 0 and 1 by dividing out by 1024. Normalizing our input data frames is more interesting, and there are multiple potential ways to go about this.

The first would be to normalize over an entire frame by subtracting the mean and dividing by the variance over all pixels, so that our normalized frame has mean 0 and variance 1, as in Figure 2.3. Another method useful in astronomy is normalizing by frequency, where we subtract the mean and divide by the variance in the time direction for each frequency sample. This also yields mean 0 and variance 1, but serves to specifically normalize out differences in instrument sensitivity as a function of frequency. Normalizing over an entire frame preserves these sensitivity differences.

However, lots of detected narrow-band signals are RFI and thus moving with the Earth,

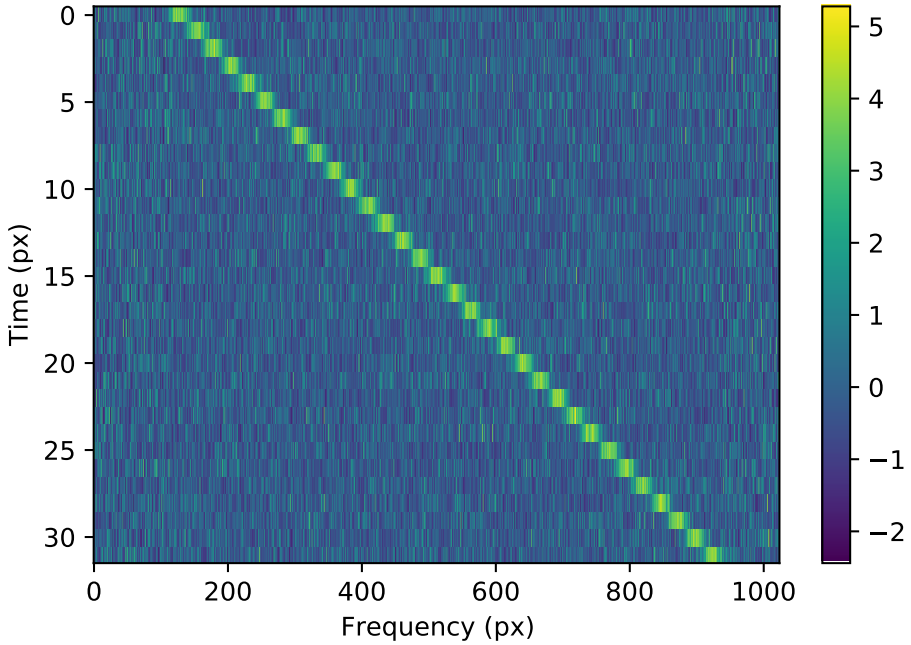


Figure 2.4: Data frame containing the same data as Figure 2.3, instead normalized per frequency channel to mean 0 and variance 1.

so they do not exhibit Doppler accelerations and appear as vertical lines in waterfall plots. So if we normalize by frequency, a constant vertical signal will disappear from the data, since we subtract out the average (constant) intensity, as in Figure 2.4. Since we are interested in localizing all signals and eventually comparing our machine learning methodology with standard search techniques, we certainly do not want to exclude this information via our normalization procedure. On the other hand, this can potentially strengthen our sensitivity towards dimmer, sloped signals.

Considering these idiosyncrasies in our data, we test both of these two normalization methods as inputs into our models.

2.3 Methods

We present the signal search methods used in this work. Namely, we discuss the CNN architectures explored for ML-based localization and briefly describe our standard deDoppler algorithm, TurboSETI.

2.3.1 CNN Model Architectures

In this work, we define both a "baseline" and a "final" model. We take our baseline architecture to be a simple CNN that is typically used on general image classification tasks. Our final model contains architectural improvements influenced by the nature of our data and training tasks. Specifically, we compare the baseline and final models and evaluate the extent to which the architectural changes improve localization accuracy.

For our two datasets, we use the same overall model architecture except for the final regression layer, which either has 2 or 4 nodes depending on whether we are predicting the position(s) of one or two signals. Since we would like to predict each signal position as best as possible, where the position is defined by its starting and ending indices, we seek to minimize the mean squared error between true and predicted indices.

Our models are implemented using the Keras functional API (Chollet et al. 2015). The source code for generating our datasets and training these models is available on GitHub.[3]

Baseline Model

For our baseline model, we choose a simple CNN with 4 convolutional layers, 3 max pooling layers, 2 fully connected layers (each with 64 nodes), and a dropout layer at 50%. Our input is a single 32×1024 data frame, normalized over the entire frame. This is a typical architecture for image classification tasks, making it a good baseline for comparison. We use rectified linear unit (ReLU) activations after each convolutional and fully connected layer. For the two signal task, this model has 1,073,988 trainable parameters. Figure 2.5 shows the baseline model architecture in detail.

Final Model

For our final model, we have 2 residual connections, 5 convolutional layers in total (using stride 2 instead of max pooling), 2 fully connected layers (both with 1024 nodes), and a dropout layer at 50%. We again use ReLU activations after each convolutional and fully connected layer, as well as batch normalization after summing frames in residual connections. Although our image frames were normalized to a mean of 0 and therefore contained negative numbers, we found that alternate activation functions to ReLU, such as tanh, did not improve localization accuracy.

Residual connections are marked by shortcuts between convolutional layers; in our case, we use element-wise addition between a given layer and a following convolutional layer (He et al. 2016). These connections reduce over-fitting and enhance accuracy by counteracting vanishing gradients in neural networks. Furthermore, since narrow-band signals are thin relative to our image frames, residual connections allow thinner features to propagate further into our models. We follow up these additive connections with batch normalization layers to

[3] github.com/bbrzycki/seti-nb-localization

ensure that the lower-order statistics of layer inputs at these positions in our model remain the same across batches of data (Ioffe and Szegedy 2015).

For our inputs, we express the data frames as "images" with two channels – one channel is the data frame normalized over the entire frame, and the other is the data frame normalized per frequency. Our rationale for using a two channel input is that when one normalizes over the entire frame, the model finds the brighter signals much more easily, and the dimmer signals could be washed out. However, most of time in radio data, the brightest signals are also at zero drift rate, since they originate from Earth. Therefore, normalizing by frequency could serve to remove these brightest signals and show more sensitivity to dimmer, drifted signals. Using both forms of image normalization as inputs into the same model could help better identify these different forms of signals appearing in our data.

For the two signal task, this model has 26,070,916 trainable parameters. Figure 2.6 shows the final model architecture in detail.

2.3.2 TurboSETI

To provide a standard for performance comparison, we also ran our one signal dataset through the TurboSETI suite. As mentioned before, TurboSETI uses a tree-based deDoppler algorithm to efficiently search for signals above a specified SNR threshold over a specified range of drift rates (Enriquez et al. 2017; Enriquez and Price 2019). For each detected signal, TurboSETI returns the starting index/frequency and drift rate, along with the calculated SNR.

Previous studies have used the detection threshold of an SNR of 10; going any lower results in an unacceptable number of false positive detections (Price et al. 2020; Sheikh et al. 2020). We likewise set a detection threshold at an SNR of 10, or 10 dB.

Selecting a maximum Doppler drift rate range to search presents a trade-off between potential detections and computational time. Sheikh et al. 2020 uses a maximum absolute drift rate of 20 Hz/s, which is the largest thus far for a deDoppler search strategy. As mentioned in Section 2.2.2, the largest possible absolute drift rate in our 32×1024 frames is 31 Hz/s, so we simply choose that as our maximum search drift rate.

Because TurboSETI works by integrating along straight line paths, it struggles to find dim signals that are close to or intersect brighter signals. We observe this with every frame in our two signal dataset, in which TurboSETI only finds the bright, non-drifted "RFI" signal. Accordingly, we only compare our TurboSETI search results with the ML predictions over the one signal dataset.

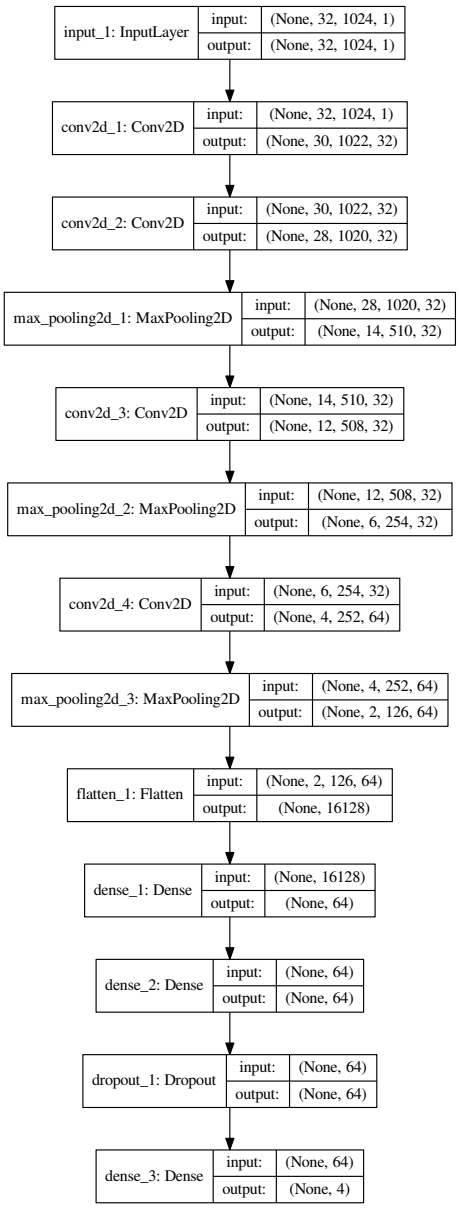


Figure 2.5: Baseline model architecture for the two signal localization task. For each block, the layer type is shown, along with input and output shapes. Inputs have shape $32 \times 1024 \times 1$, normalized over all pixels. These are passed through an initial convolutional layer, followed by 3 pairs of convolutional and max pooling layers. This is followed by two fully connected layers and a dropout layer, before finally going into the output layer. For the single signal task, the last layer has 2 nodes instead of 4.

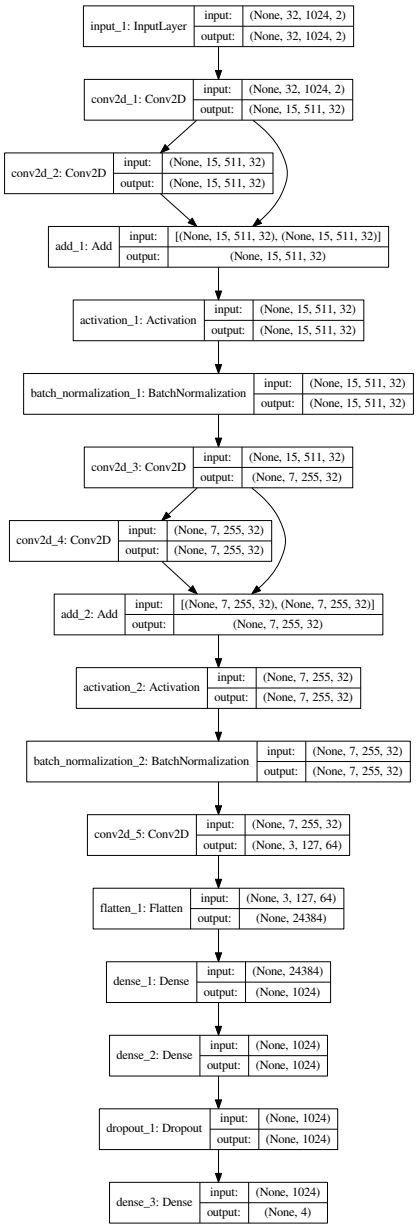


Figure 2.6: Final model architecture for the two signal localization task. Inputs have shape $32 \times 1024 \times 2$, combining the two normalizations described in Section 2.3.1. Residual connections are apparent between convolutional layers, followed by a batch normalization layer. This structure is repeated twice, followed by another convolution layer, two fully connected layers, and a dropout layer. For the single signal task, the output layer has 2 nodes instead of 4.

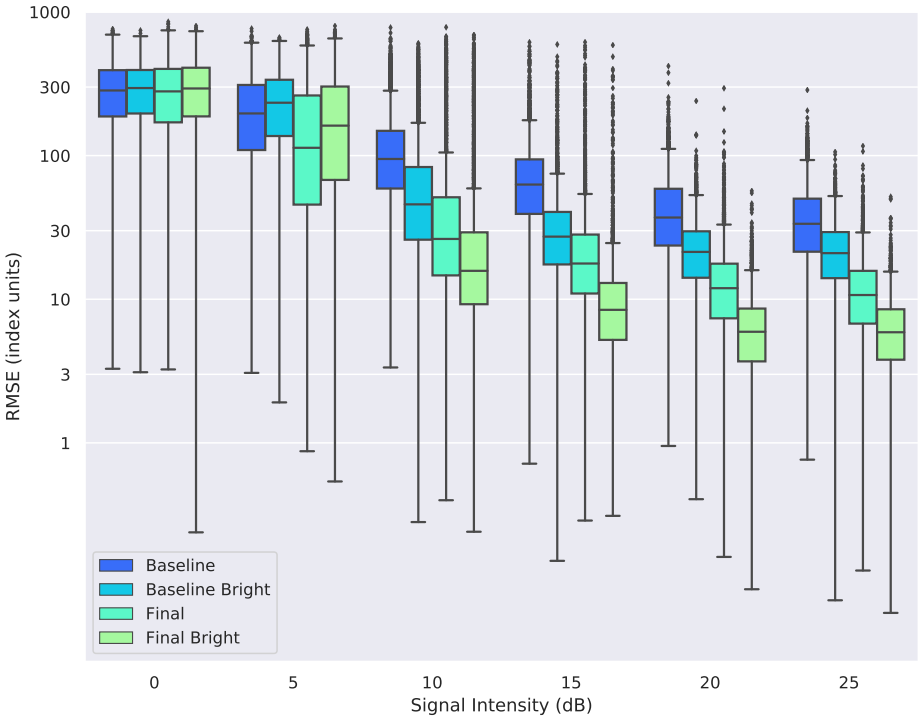


Figure 2.7: Box plot of RMSE in index/pixel units for the one signal dataset as a function of signal SNR. We compare the final and baseline models trained on both the full 0 – 25 dB dataset, as well as the truncated "bright" 10 – 25 dB dataset.

2.4 Results

2.4.1 Baseline vs. Final Model

For the baseline and final model architectures, we compare the error between true and predicted indices as a function of SNR. To get a more intuitive feel on model performance, we calculate 1024×RMSE, where RMSE is the root mean square error, to see the errors in units of pixels/indices:

$$\text{RMSE (index units)} = 1024 \times \sqrt{\frac{1}{n}\sum_{i}^{n}(y_i - \hat{y}_i)^2}, \tag{2.2}$$

where n is the number of indices in our labeled data (2 and 4 for one and two signal datasets), y_i are predicted indices, and $\hat{y}_i$ are true indices.

We first train our models on the full 0 – 25 dB SNR levels at each dataset. We also analyze the effect of only training on frames with a 10 – 25 dB signal, cutting out the 0

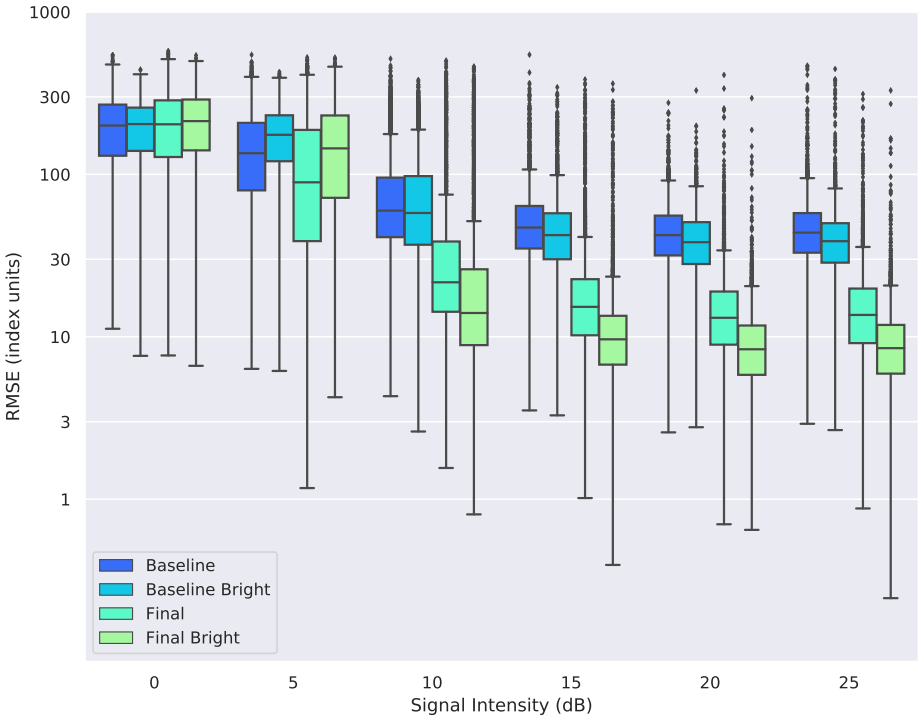


Figure 2.8: Box plot of RMSE in index/pixel units for the two signal dataset as a function of signal SNR. We compare the final and baseline models trained on both the full 0 – 25 dB dataset, as well as the truncated "bright" 10 – 25 dB dataset.

and 5 dB levels. Here, we again use a train-test split of 80/20 out of this restricted training set. However, we still evaluate these trained models on the full test set. Results from the restricted 10 – 25 dB dataset are labeled "bright" in Figures 2.7-2.8.

In these figures, we plot a box and whisker plot (with outliers above and below the median by 1.5 times the spread between 25% and 75% quartiles) for each SNR level and each training run for our baseline and final models on the full and bright datasets. For each case, we have outliers with high errors that would tend to bias our evaluations much higher if we only consider the mean RMSE.

2.4.2 TurboSETI vs. Final Model

For each frame in the one signal dataset, we use TurboSETI to get signal localization results, which we translate into starting and ending indices. This allows us to calculate the RMSE in index units, exactly as we did for the ML predictions. Figure 2.9 shows the results, again as a function of SNR level, compared to predictions from our final CNN model trained on "bright" frames. Note that we only compare predictions down to an intensity of 10 dB, since

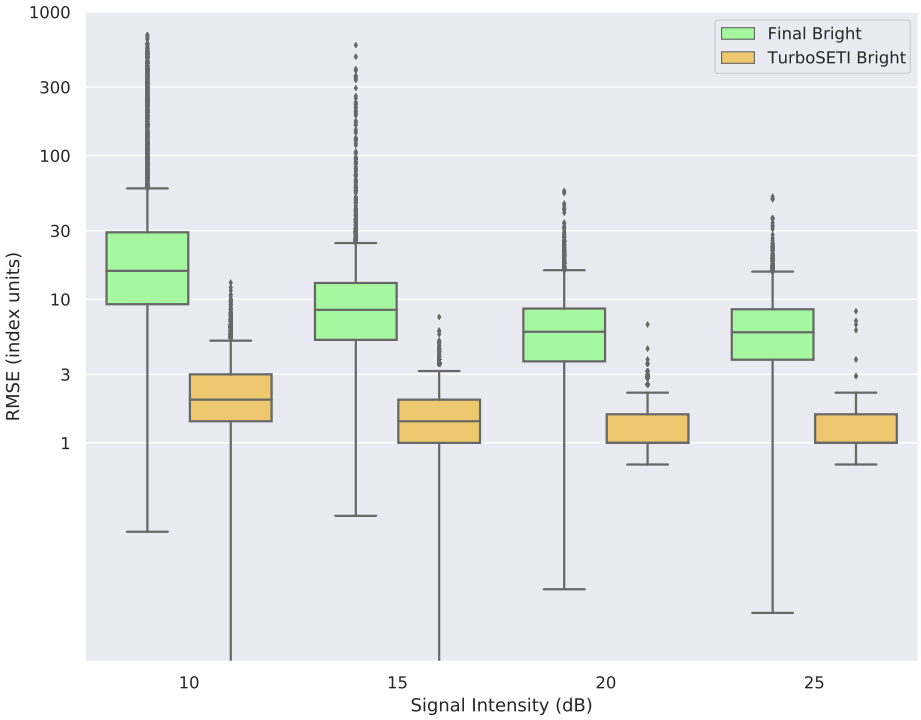


Figure 2.9: Box plot of RMSE in index/pixel units, comparing performance of our final model architecture with TurboSETI on the one signal dataset as a function of signal SNR.

that is our TurboSETI detection threshold.

We also compare the computational costs of each search method. For the final CNN model, generating predictions for all 24,000 frames in the test set takes about 70 seconds using an Nvidia Titan Xp GPU with a batch size of 1. Using a batch size of 32, generating all predictions takes about 34 seconds.

Although we run TurboSETI with a 10 dB threshold, for bench-marking purposes, we ran it on all 24,000 frames in the test set. This takes a total of 2.6 hours using a single CPU. In practice, however, TurboSETI is run on large data frames, on order of billions of frequency channels, as opposed to the comparatively small frames used in this work. This is because the tree-based search algorithm is most efficient when applied to a few large files, as opposed to many smaller ones. To make a fairer comparison, we ran TurboSETI on a large frame with the same amount of data as if all 24,000 test frames were concatenated along the frequency axis. Using the same search parameters of a 10 dB threshold and a maximum 31 Hz/s drift rate, TurboSETI takes about 20 minutes to finish searching this concatenated data frame. Making predictions using the CNN model is therefore more efficient, especially when using larger batch sizes.

Lastly, we tested both TurboSETI and the final CNN model on localizing a few known RFI

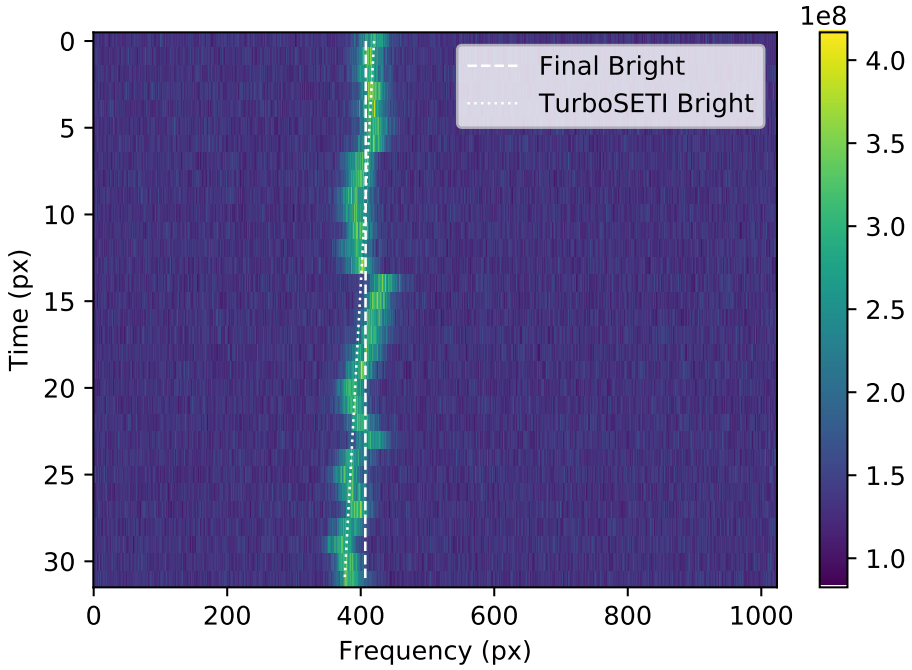


Figure 2.10: Observational data frame containing an RFI signal, overlaid with localization predictions from our final CNN architecture (dashed) and TurboSETI (dotted). Although the real signal is more complex than those in our training data, the model produces reasonable predictions.

signals in the C-band observation described in Section 2.2.1. Figure 2.10 shows an example of predicted localization paths from both methods for a complex RFI signal. As expected, TurboSETI generates the best-fit localization, since it integrates along each potential drift rate. Nevertheless, the final model gives a reasonable prediction, despite being trained on idealized signals with constant intensity and constant drift rate.

2.5 Discussion

For the full training datasets, the baseline and final models show a smooth progression of better median RMSE values from 0 dB to 25 dB. Our final model seems to outperform the baseline model consistently, particular for signals of at least 10 dB, on the order of a 3× reduction in error for both one and two signal datasets (Figures 2.7-2.8).

Unsurprisingly, for both one and two signal datasets, the signals at 0 dB (SNR=1; 1σ) do very poorly, with an typical error of 200 – 400 pixels off. As a check, note that the average expected distance between two randomly chosen points on a line segment of length L is $L/3$.

For 1024 total pixels, this means on average, selecting an index at random as a prediction should yield errors of about $1024/3 \approx 341$ pixels. So for 0 dB signals, our predictions are essentially random. In a way, this is perfectly acceptable, since in general we do not accept 1σ as a true detection of a signal in the first place.

When we compare these results with those from models only trained on $10 - 25$ dB frames, we see a few interesting things. As expected, the "bright" models then perform worse on the lower SNR signals, 0 and 5 dB. For 0 dB signals, we still get effectively random predictions, but the decrease in performance at 5 dB is notable. However, this restricted training set improves the performance at $10 - 25$ dB appreciably, in different ways. In the one signal case, the final model outperforms the baseline model by about $3\times$ for all SNR levels of at least 10 dB. In the two signal case, frames with at least 15 dB signals improved about $5\times$ between final and baseline models. We further observe that, in each case, restricting training to frames with at least a 10 dB signal improves overall model performance compared to training with the full dataset. Since both models seem to only be capable of nearly random predictions at the lowest SNRs, doing training steps on that data tends to hurt the performance at higher SNRs.

The best performing models were the final model trained on the bright dataset. At an SNR level of 25 dB, the one and two signal cases reached a median of about 6 and 9 pixels of RMSE, respectively. We expect the one signal models to do better than the two signal models, since there is only a single bright signal to try to localize. Nevertheless, it was surprising that even our best models did not consistently localize either of these cases to extremely high precision, i.e. median RMSE of about 1 pixel. On the one hand, localizing a bright signal to a precision of 6 pixels out of 1024 is decent (corresponding to about 8.4 Hz), but on the other, we expected to do even better, since we have synthetic datasets and therefore know *precisely* where the center of each signal lies. Instead, as we observe in Figure 2.9, TurboSETI very accurately localizes these signals, to a median RMSE of about $1 - 2$ pixels for each SNR.

For the ML predictions, there are still large outliers, even reaching levels of randomness ($300 - 500$ pixels off) in the two signal case. One potential explanation and limitation in our model is that our labels are ordered. For two signals, at the highest SNR, both signals are at 25 dB. The model could have a harder time differentiating one signal for another (where the only distinguishing factor is zero vs. non-zero drift rate) and produce bad predictions. Our test set has 4,000 images for each SNR level, so even having a small fraction of these show up as outliers at the highest SNRs would be compounded if we tried using this model on real data. Recall that a 32×1024 px image at 1.4 s and 1.4 Hz resolution is about 45 s by 1430 Hz in total. For a typical 5 minute C-band observation ($4 - 8$ GHz), this is equivalent to about 18.65 million data frames. Before it makes sense to use our CNN pipeline in searches on real data, we need to find ways of very precisely localizing these signals so that we are not swamped with false positives and inaccurate positions.

Nevertheless, inspecting ML predictions in the two signal case revealed that the models appear to learn that the first two labels (i.e. corresponding to the starting and ending index of the zero drift RFI signal) should be the same index, since it predicted essentially the same

value between [0, 1) up to a few significant figures (at least to differences of 1/1024).

Despite being much more accurate over the one signal dataset, TurboSETI takes much longer to produce localization predictions than our CNN models for the same amount of data, on the order of 20 minutes vs. 34 seconds, as discussed in Section 2.4.2. Of course, this is not necessarily surprising, since our ML predictions take advantage of GPU-accelerated calculations, especially when we batch together multiple data frame inputs at once.

For every frame in the two signal dataset, we also observe that TurboSETI struggles to find multiple signals that are close together or intersecting, and instead only detects the brighter one. For the same dataset, our final CNN model can generally localize such signals to an RMSE of about 10 – 20 pixels for signals at 10 dB and above. Besides being more computationally efficient, CNN-based pipelines may therefore be better at finding elusive signals that standard search techniques tend to miss.

In addition, it is encouraging that when we take into account the multiple possible normalizations, our model performance improves, especially in the two signal case with a model RFI signal. We believe that this could generalize well to frames with over two signals, as long as we can find a good way of matching labels and predictions, perhaps without necessarily enforcing an ordering.

2.6 Summary

Accurately identifying the presence and positions of signals in radio data is important for finding candidate technosignatures and ensuring that we do not miss interesting signals in the presence of bright RFI. Computer vision techniques allow us to ingest complex image frame data and distill them into relevant information, such as signal locations.

We found that our final model outperformed our baseline model for all SNRs in both datasets, and training each model on datasets limited to 10 – 25 dB results in significant increases in model performance. Our best results had a median RMSE of about 6 pixels in the one signal case and about 9 pixels in the two signal case. Since we used simple signals embedded in ideal chi-squared noise, we expected our localization models to perform even better, especially in the one signal case. Nevertheless, while these errors are higher than expected and come with a host of outliers that perform much more poorly, these results are promising for future work in localizing narrow-band signals in images.

We also did an analysis using TurboSETI to detect signals in our synthetic datasets, and compared the results with our ML predictions. We found that while TurboSETI produces more accurate localizations for the one signal dataset, our ML pipelines are much faster and able to produce meaningful results even for more complex tasks, such as localizing both signals in our two signal dataset.

Overall, object detection and localization of long, thin objects is difficult since they do not match the typically rectangular shape of many other objects, and so it is harder to maximize intersection over union measures with ground truth. Although on the one hand detecting lines seems simple intuitively, the relative lack of information compared to broader,

extended objects makes it more difficult, especially within a noisy background. Nevertheless, with a few key assumptions that are specific to the radio data we collect for SETI, we can make more progress in precisely localizing radio frequency signals.

A future direction for this work is to investigate the effectiveness of treating multiple normalizations as independent inputs to the same model, instead of combining them as a single two channel input. Each normalization could have a few convolutional layers to itself, and would be added to each other to learn features with contributions from both normalizations. Indeed, this approach could scale better and benefit from additional data preprocessing techniques beyond the two normalizations discussed in this paper.

We can also easily use this CNN architecture to classify signals, or to both classify and localize simultaneously depending on how we choose our labels and loss functions. We are interested to see how well this method extends to more than two signals in a single image frame, and eventually, we would like to develop a pipeline for signal *detection* of an arbitrary number of signals in a given image frame.

2.7 Acknowledgements

Breakthrough Listen is managed by the Breakthrough Initiatives, sponsored by the Breakthrough Prize Foundation. The Green Bank Observatory is a facility of the National Science Foundation, operated under cooperative agreement by Associated Universities, Inc. We thank the staff at the Green Bank Observatory for their operational support.

Chapter 3

Setigen: Simulating Radio Technosignatures for the Search for Extraterrestrial Intelligence

A version of this chapter was originally published as: Brzycki, B., Siemion, A.P., de Pater, I., Croft, S., Hoang, J., Ng, C., Price, D.C., Sheikh, S., and Zheng, Z., 2022. Setigen: Simulating Radio Technosignatures for the Search for Extraterrestrial Intelligence. The Astronomical Journal, 163(5), p.222.

The goal of the search for extraterrestrial intelligence (SETI) is the detection of non-human technosignatures, such as technology-produced emission in radio observations. While many have speculated about the character of such technosignatures, radio SETI fundamentally involves searching for signals that not only have never been detected, but also have a vast range of potential morphologies. Given that we have not yet detected a radio SETI signal, we must make assumptions about their form to develop search algorithms. The lack of positive detections also makes it difficult to test these algorithms' inherent efficacy. To address these challenges, we present `setigen`, a Python-based, open-source library for heuristic-based signal synthesis and injection for both spectrograms (dynamic spectra) and raw voltage data. `setigen` facilitates the production of synthetic radio observations, interfaces with standard data products used extensively by the Breakthrough Listen project (BL), and focuses on providing a physically-motivated synthesis framework compatible with real observational data and associated search methods. We discuss the core routines of `setigen` and present existing and future use cases in the development and evaluation of SETI search algorithms.

3.1 Introduction

Since the inception of radio SETI in the 1960s, technosignature searches have greatly expanded to cover more sky area, wider frequency ranges, and a larger variety of signal mor-

phologies (Drake 1961; Werthimer et al. 1985; Tarter 2001; Siemion et al. 2013; Wright et al. 2014; MacMahon et al. 2018; Price et al. 2018; Gajjar et al. 2021). Arguably the most developed branch of radio SETI is the search for narrow-band technosignatures, with signal bandwidths under 1 kHz, for which search algorithms are constantly being produced and improved (Siemion et al. 2013; Enriquez et al. 2017; Pinchuk et al. 2019; Margot et al. 2021). These algorithms operate on either voltage time series data or time-frequency spectrogram data (i.e., dynamic spectra, waterfall plots).

The incoherent tree deDoppler method is the primary search strategy for Doppler-accelerated narrow-band signals in radio spectrograms (Taylor 1974; Siemion et al. 2013; Enriquez et al. 2017; Margot et al. 2021). An ideal sinusoidal emitter will appear to exhibit a frequency drift over time due to relative acceleration between the emitter and receiving telescope (Sheikh et al. 2019). Under a constant relative acceleration, such a signal will have a linear drift or slope in a spectrogram of Stokes I intensities. The tree deDoppler algorithm efficiently integrates spectra over potential drift rates and identifies signals above a threshold signal-to-noise ratio (SNR). Breakthrough Listen, the most comprehensive SETI search program to date (Worden et al. 2017), developed `turboSETI`[1], an open-source implementation of the deDoppler algorithm that serves as the backbone of many technosignature searches (Enriquez et al. 2017; Enriquez and Price 2019; Price et al. 2020; Sheikh et al. 2020; Gajjar et al. 2021).

This method works well for signals with high duty cycles and linear drift rates, but it can struggle to properly detect more complex signals (Pinchuk et al. 2019). This is particularly problematic given the increasingly complex radio frequency interference (RFI) environment within which these searches are conducted. Moreover, the lack of robust, labeled, narrow-band signal datasets can make it difficult to quantify a given implementation's detection accuracy, especially in light of RFI and variable bandpass responses.

For more complex signal morphologies, machine learning (ML) algorithms have been proposed that use computer vision techniques to classify image-like spectrograms. However, the same lack of labeled, narrow-band signal data makes creating supervised ML models difficult. Zhang et al. 2018b used a self-supervised approach in which spectrogram data was divided in time into two halves, for which the ML task was to predict the second half given the first. For an ML-based direction-of-origin filter, Pinchuk and Margot 2022 used a separate non-ML method to detect signals and create an algorithmically-labeled spectrogram dataset. In most cases, however, supervised approaches have relied on generating synthetic signals of various classes in order to guarantee correct labels (Harp et al. 2019; Brzycki et al. 2020; Margot et al. 2021).

To address these issues, we present `setigen`, an open-source Python library[2] that facilitates the creation of synthetic narrow-band signals and supports injection into observational data. `setigen` is meant to provide a general-use heuristic framework for creating mock radio SETI data. A primary design aspect is ensuring that the synthesis process is grounded as

[1] https://github.com/UCBerkeleySETI/turbo_seti
[2] https://github.com/bbrzycki/setigen

much as possible in physical quantities to better interface with real observations and search algorithms. `setigen` makes heavy use of `NumPy`[3] for efficient matrix operations (Oliphant 2006; Harris et al. 2020) and `blimpy`[4] for interfacing with data products routinely used by BL (Price et al. 2019b).

There are two main modules in `setigen`, "spectrogram" and "voltage," dedicated to the most common data formats used in radio SETI. The spectrogram module works with Stokes I (intensity) data stored as time-frequency arrays and is designed to be flexible and heuristic-based. It can be used to generate many small snippets of data containing synthetic signals for quick algorithm test cases or for full labeled datasets. The voltage module creates synthetic antenna voltages, follows these voltages through a software-based signal processing chain that models a standard single dish signal pipeline, including quantization and a polyphase filterbank, and saves the final complex voltages. This requires a lot more computational power, so voltage `setigen` routines can be optionally GPU-accelerated via `CuPy`[5] (Okuta et al. 2017). Since the voltage module models the signal processing chain, it can be used to produce more "realistic" signals, test complex voltage processing software, and evaluate how each signal processing element affects the final signal sensitivity.

Radio SETI searches typically operate on data in spectrogram format, since it compresses data and enables visualization and analysis of broader signal morphology in time-frequency space (Enriquez et al. 2017; Margot et al. 2018; Pinchuk et al. 2019; Price et al. 2020; Sheikh et al. 2020). As such, `setigen` was initially written to create large datasets of radio spectrograms for use in supervised ML search experiments. The library was later expanded to support synthesizing raw voltage-level data to complement existing use cases.

`setigen` has already been used in a variety of applications, such as the development and testing of search algorithms. It has been used to create synthetic datasets with position labels for ML localization tasks in single observations (Brzycki et al. 2020). `setigen` has also been used to inject synthetic signals within ON-OFF cadences, each comprised of 6 consecutive observations and used as a direction-of-origin filter for SETI. Ma et al. (submitted) injected signals into ON-OFF cadences taken with the Robert C. Byrd Green Bank Telescope (GBT; MacMahon et al. 2018) to train a sophisticated variational autoencoder model that can classify cadences as potential SETI candidates. Similarly, `setigen` was used extensively to produce training and test data in BL's first Kaggle ML competition[6], in which contestants were tasked with classifying synthetic technosignature candidates in ON-OFF cadences.

Outside of ML, synthetic `setigen` data is used in injection-recovery testing for `turboSETI` as well as for a new search code, `hyperseti`[7]. The voltage module has been used to test and upgrade parts of the Allen Telescope Array's (Welch et al. 2009) software signal processing pipeline. Furthermore, `setigen` has been used to test RFI rejection and detection techniques for the Parkes Multibeam Galactic Plane Survey SETI search, helping to discriminate ter-

[3] https://numpy.org/
[4] https://github.com/UCBerkeleySETI/blimpy
[5] https://cupy.dev/
[6] https://www.kaggle.com/c/seti-breakthrough-listen
[7] https://github.com/UCBerkeleySETI/hyperseti

restrial signals from different regions in the sky as SETI surveys with multiple antennas or beams become more popular (Perez et al., in prep).

This paper is organized as follows. Section 3.2 outlines the standard signal chain and processing pipeline used in single dish radio SETI observations to motivate details behind `setigen`'s synthesis methods. Section 3.3 presents the code methodology: Section 3.3.1 describes the spectrogram module for producing and working with synthetic Stokes I time-frequency data, while Section 3.3.2 describes the voltage synthesis module in detail, connecting components of typical radio signal chains to software analogues used in `setigen`. In Section 3.4, we discuss current limitations of the library and future directions for signal synthesis for SETI.

3.2 Overview of Single Dish Signal Chains

To motivate the capabilities of `setigen`, we first give a broad overview of the standard single dish data recording pipeline, as well as some details pertinent to the Breakthrough Listen digital recorder (BL DR) system at the GBT (MacMahon et al. 2018).

In a single-dish radio telescope, incoming radiation is reflected off the dish surface toward a feed horn at the focus. The feed couples incident free-space electromagnetic radiation to voltages within the telescope's receiver system.

These voltages are passed to an analog down-conversion system containing a heterodyne mixer, which shifts the signal from the target RF range into an intermediate frequency (IF) range near baseband more suitable for receiver hardware. The resulting voltages are then digitized by analog-digital converters (ADC) to a specified number of bits $N_{\text{bits,d}}$ at a given sampling rate f_s. The BL DR system digitizes voltages to 8-bit at a sampling rate of $f_s = 3$ GHz for each linear polarization (MacMahon et al. 2018).

Radio telescope pipelines commonly use polyphase filterbanks (PFB; Bellanger et al. 1976; Harris and Haines 2011; Price 2021) to help partition the usable band and improve the spectral channel response of the system. For example, the BL DR system uses an 8-tap PFB to divide the 1.5 GHz Nyquist range into $N_{\text{coarse}} = 512$ "coarse" spectral channels, which in turn are divided among 8 compute nodes (MacMahon et al. 2018). This procedure performs a Fast Fourier Transform (FFT) with a length of $P = 2N_{\text{coarse}} = 1024$. For receivers with wide bandwidths, such as C-band at 3.95–8.00 GHz, multiple copies of these elements, starting from the analog mixer, are employed to cover the full band (NRAO 2019).

The digital processing components of the BL DR system are done on custom signal processing boards using field-programmable gate arrays (FPGAs), provided by the Collaboration for Astronomy Signal Processing and Electronics Research (CASPER; Hickish et al. 2016). These boards use fixed point arithmetic and increase numerical bit size when doing computations (MacMahon et al. 2018). Accordingly, both real and imaginary components of the resulting complex voltages must be requantized (e.g. to $N_{\text{bits,r}}$) before they are written to disk. The BL DR system records these as 8-bit signed integers in GUPPI (Green Bank

Ultimate Pulsar Processing Instrument; DuPlain et al. 2008) raw format, based on FITS (Pence et al. 2010) and stored as `.raw` files (Lebofsky et al. 2019).

Since raw voltage data comes at the highest resolution possible given the ADC sampling rate, data volumes are large, especially during standard BL observing campaigns. Therefore, we finely channelize or "reduce" raw data into spectrograms (also known as dynamic spectra or "waterfall plots"), 2D arrays of intensity (Stokes I) as a function of time and frequency (Lebofsky et al. 2019). Multiple versions with different resolutions can be created from the same set of raw data by varying the FFT length N_{fine} and integration factor N_{int}.

During fine channelization, an FFT of length N_{fine} is performed on complex raw voltages within individual coarse channels, resulting in N_{fine} fine channels each. So, we can express the full Nyquist bandwidth as

$$f_N = \frac{f_s}{2} = N_{\text{coarse}} N_{\text{fine}} \Delta f. \tag{3.1}$$

This gives us an expression for the spectrogram's frequency resolution:

$$\Delta f = \frac{f_s/2}{N_{\text{coarse}} N_{\text{fine}}}. \tag{3.2}$$

If the total observation length is τ and the number of time channels (pixels) in the final spectrogram is N_t, then

$$N_t = \frac{\tau}{\Delta t}, \tag{3.3}$$

assuming that τ is a multiple of the spectrogram's time resolution Δt. In practice, extraneous samples are truncated when necessary to satisfy this requirement.

The integration factor N_{int} is the number of spectra integrated in the time direction. To get an expression for Δt, we can think in terms of the total number of samples collected (for a single linear polarization):

$$N_s = \tau f_s. \tag{3.4}$$

The pipeline takes in N_s real samples in time and, via a P-point FFT, transforms the data into a complex 2D array in time-frequency space, with non-integrated dimensions $N_t N_{\text{int}} \times P N_{\text{fine}}$.

$$N_s = N_t N_{\text{int}} P N_{\text{fine}} = 2 N_t N_{\text{int}} N_{\text{coarse}} N_{\text{fine}}. \tag{3.5}$$

Note that since the FFT is performed on real voltages, the unique frequency extent is ultimately halved per the Nyquist range.

Combining Eqs. 3.2–3.5, we get

$$N_s = \tau f_s = 2 N_t N_{\text{int}} N_{\text{coarse}} N_{\text{fine}} \tag{3.6}$$

$$= 2 \frac{\tau}{\Delta t} N_{\text{int}} N_{\text{coarse}} N_{\text{fine}} \tag{3.7}$$

$$\Delta t = \frac{2 N_{\text{int}} N_{\text{coarse}} N_{\text{fine}}}{f_s} = \frac{N_{\text{int}}}{\Delta f}. \tag{3.8}$$

Although N_{fine} and N_{int} must both be integers, we otherwise have fine control over Δf and Δt through Eqs. 3.2 and 3.8.

3.3 Code Methodology

As object-oriented software, `setigen` has a set of important classes and routines that are described below. For more technical details and examples of the API, see the full documentation[8].

3.3.1 Spectrogram Module

The spectrogram module provides an interface for synthesizing Stokes I (waterfall) data in a format common to radio SETI and is oriented around the `Frame` class. A `Frame` object contains a 2D data array of intensities as a function of time and frequency, as well as accompanying metadata, such as starting frequency and time-frequency resolutions.

Data frames can be initialized from either saved observational data or frame parameters. Frames can extract Stokes I data and observational metadata from filterbank (`.fil`) or HDF5 files (`.h5`). The most important metadata for `setigen` are the physical parameters of the underlying intensity data: resolutions and ranges in both time and frequency. Empty frames can therefore be created simply by specifying these parameters along with desired data array dimensions.

Noise Synthesis

In most SETI applications, we search for statistically-significant signals embedded in noise. Since voltage noise in the absence of RFI approximately follows a zero-mean normal distribution (Thompson et al. 2017), the radiometer noise in spectrogram data follows a chi-squared distribution (McDonough and Whalen 1995; Nita et al. 2007). When the time and frequency resolutions are coarse enough, the spectrogram noise approaches a normal distribution by the central limit theorem.

Specifically, suppose we have a sequence of input voltages $\{x_n\}$ following a Gaussian distribution with zero mean. During the coarse channelization process, the polyphase filterbank applies, at its core, an FFT to bring the voltages into frequency space:

$$X_k = \sum_{n=0}^{N-1} w_n x_n e^{-2\pi i k n/N}, \ k = 0, \ldots, N-1, \tag{3.9}$$

where N is the number of frequency bins and $\{w_n\}$ are coefficients of a windowing function applied to improve the spectral response (Price 2021).

More specifically, the filterbank sums over M rows of P samples before a P-point FFT, so that the response of the rth row of P samples is:

$$X_{k,r} = \sum_{p=0}^{P-1} \left[\sum_{m=0}^{M-1} w_{n'} x_{n''} \right] e^{-2\pi i k p/P}, \tag{3.10}$$

[8] https://setigen.readthedocs.io/

where $n' = mP + p$ and $n'' = (r - M + m)P + p$ are indices of the windowing coefficients and voltages samples in terms of m and p. Here, we assume that the MP windowing coefficients are symmetric about the midpoint, so that $w_n = w_{MP-n-1}$.

Ignoring quantization for the moment, we store the complex components of the resulting FFT voltages, $\mathrm{Re}(X_k)$ and $\mathrm{Im}(X_k)$, as raw voltage data. Since these are linear combinations of independent zero-mean Gaussian variables (i.e. x_n), they both follow zero-mean Gaussian distributions.

In the absence of a windowing function ($w_n = 1$), for each channel besides the real-valued DC and Nyquist bins, the variances of the real and imaginary components are equal (σ^2; McDonough and Whalen 1995). When a windowing function is used, the underlying statistics can change such that the variances of the complex components differ as a function of spectral bin (Nita et al. 2007). However, for commonly chosen symmetrical windows (e.g. Hamming), this effect is negligible in most spectral bins.

For a single linear polarization, the power is given by

$$I_{x,k} = |X_k|^2 = \mathrm{Re}(X_k)^2 + \mathrm{Im}(X_k)^2 \tag{3.11}$$

Assuming both complex components have the same variance σ^2, the power follows a chi-squared distribution with two degrees of freedom:

$$I_{x,k} \sim \sigma^2 \chi^2(2) \tag{3.12}$$

During the fine channelization step, we integrate N_{int} spectra in the time direction and combine power from N_{pol} polarizations. Therefore, in the final Stokes I spectrogram, the total number of chi-squared degrees of freedom is given by:

$$\mathrm{DOF} = 2N_{\mathrm{pol}}N_{\mathrm{int}} = 2N_{\mathrm{pol}}\Delta f \Delta t \tag{3.13}$$

$$I_k \sim \sigma^2 \chi^2(2N_{\mathrm{pol}}\Delta f \Delta t), \tag{3.14}$$

using Eq. 3.8. For dual-polarization Stokes I data, $\mathrm{DOF} = 4\Delta f \Delta t$. This allows us to generate synthetic chi-squared noise with the correct number of degrees of freedom just from frame resolutions, which are either directly specified or inferred from observations. Since non-calibrated intensity values are arbitrarily scaled, we can simply scale the magnitudes of synthetic chi-squared noise to match empirical observational noise distributions.

The main function for noise synthesis across a frame is `add_noise`, which adds random noise to every pixel in the data array. By default, it generates chi-squared noise with a user-specified mean intensity μ. Since the mean of a chi-squared distribution equals the number of degrees of freedom, for dual-polarization data, we have

$$I_k \sim \left(\frac{\mu}{4\Delta f \Delta t}\right) \chi^2(4\Delta f \Delta t) \tag{3.15}$$

$$\langle I_k \rangle = \left(\frac{\mu}{4\Delta f \Delta t}\right) \cdot 4\Delta f \Delta t = \mu \tag{3.16}$$

$$\mathrm{Var}(I_k) = \left(\frac{\mu}{4\Delta f \Delta t}\right)^2 \cdot 2 \cdot 4\Delta f \Delta t = \frac{\mu^2}{2\Delta f \Delta t}. \tag{3.17}$$

In addition to chi-squared noise, `add_noise` can also generate Gaussian noise. By the central limit theorem, as the degrees of freedom increase, a chi-squared distribution approaches a normal distribution. For example, $N_{\text{int}} = 51$ for BL's standard high spectral resolution data product, so DOF = 204 and the resulting background noise is close to Gaussian. Directly synthesizing Gaussian-distributed noise can save normalization steps in data processing, but should be used carefully when comparing with real observational data.

A useful extension of the noise synthesis function is `add_noise_from_obs`, which draws from archived observational statistics to set realistic intensity values. The observations were taken using the GBT at C-band and reduced to (1.4 s, 1.4 Hz) resolution. For example, for chi-squared noise, the function randomly selects an archived mean intensity, scales it to the appropriate frame resolution, and populates noise per Eq. 3.15. An implementation detail of BL's fine channelization software, `rawspec`[9], is that as part of the FFT, intensity values are scaled up by a factor of the FFT length N_{fine}. So, for observations going through the BL data pipeline (i.e. the same digitization and coarse channelization hardware):

$$\mu \propto N_{\text{fine}} N_{\text{int}} \tag{3.18}$$
$$\propto N_{\text{fine}} \Delta f \Delta t \tag{3.19}$$
$$\propto \Delta t. \tag{3.20}$$

Alternatively, the function also accepts user-provided arrays of background noise intensity statistics from which to sample instead. This can be used for synthesizing data with intensity ranges from other telescopes (e.g. Parkes) or even GBT data at different frequency bands or sensitivities.

After noise synthesis, the frame will update class attributes storing the estimated mean μ_b and standard deviation σ_b of the background noise. For an empty frame, the first noise synthesis function will set these properties directly. For pre-loaded observational data and further noise injection, the frame estimates the background noise through iterative sigma clipping at the 3σ level to exclude outliers. For frames small enough that noise statistics do not change over the frequency bandwidth, this enables signal injection at desired SNR levels.

Signal Synthesis

For narrow-band signal synthesis, the `add_signal` function creates heuristic, user-defined signals in spectrogram data. Our convention is that the spectrogram data has time on the y-axis and frequency on the x-axis.

In spectrogram `setigen`, narrow-band signals have a "central" frequency at each timestep and a unique spectral profile centered at that frequency. As such, there are four main heuristic descriptors for a narrow-band signal in `setigen`:

1. `path` – $I_p(t)$: Central signal frequencies as a function of time, e.g. linear (constant) drift rate, quadratic drift rate

[9] https://github.com/UCBerkeleySETI/rawspec

2. `t_profile` – $I_t(t)$: Signal intensity as a function of time, e.g. constant intensity, Gaussian pulses

3. `f_profile` – $I_f(f, f_0)$: Spectral profile as a function of frequency (offset from central frequency), e.g. sinc^2 profile, Gaussian profile

4. `bp_profile` – $I_{bp}(f)$: Bandpass profile as a function of absolute frequency

These descriptors are parameters for `add_signal` and are Python functions by type. A set of common functions are provided with `setigen`, and others can be custom-written. The simplest and most ideal kind of narrow-band signal has a constant intensity and drift rate; such signals can be created straightforwardly through the wrapper function `add_constant_signal`.

For a pixel at (t, f) in the time-frequency spectrogram, the intensity of a synthetic signal is calculated as

$$I(t, f) = I_t(t) I_f(f, I_p(t)) I_{bp}(f). \tag{3.21}$$

As such, Eq. 3.21 is computed for every pixel in the spectrogram, since there is no robust way to constrain arbitrary intensity profiles. For example, even an ideal Gaussian function is non-zero at all distances and defining a suitable range depends on the experiment. For large spectrograms, it can be inefficient to calculate intensities for pixels far from the main signal, so users can provide a custom frequency range to limit the signal calculation.

The signal calculation is fully heuristic, in that the calculation is completely user-specified and does not take other effects into account, such as FFT leakage or spectral responses. Since intensity is treated as a function of time and frequency, this process can overlook how intensities are integrated in reality. As a partial solution, `add_signal` provides the option to separately sub-integrate within each pixel in time and frequency directions.

In a similar vein, a difficult effect to handle robustly is Doppler smearing, in which a highly drifting signal will have its power spread into multiple frequency channels within the same time channel (Sheikh et al. 2019). While an analytical form exists for the spectral profile of a linearly drifting cosine signal, the smearing effect will naturally apply to more complex signals. Variable spectral profiles are not yet supported in `setigen`, but from a user standpoint, it would be tedious to manually construct custom smearing profiles that change at each timestep. Using a similar process to numerical integration, `add_signal` has the option to approximate Doppler smearing by computing and averaging a given number of copies of the signal, spaced evenly between signal center frequencies in adjacent timesteps. For instance, for the ith time channel at $t = t_i$, copies of the signal centered at even spacings between $I_p(t_i)$ and $I_p(t_{i+1})$ are averaged together to get the ith spectral profile. This is done for all time channels, so that channels with smaller signal drifts will be brighter than those with larger signal drifts by the correct ratio, as long as the number of copies gives enough coverage over the channel with the largest signal drift.

Sometimes it can be difficult or unwieldy to wrap up a desired signal property into a separate function, or perhaps there is existing external code that produces such properties.

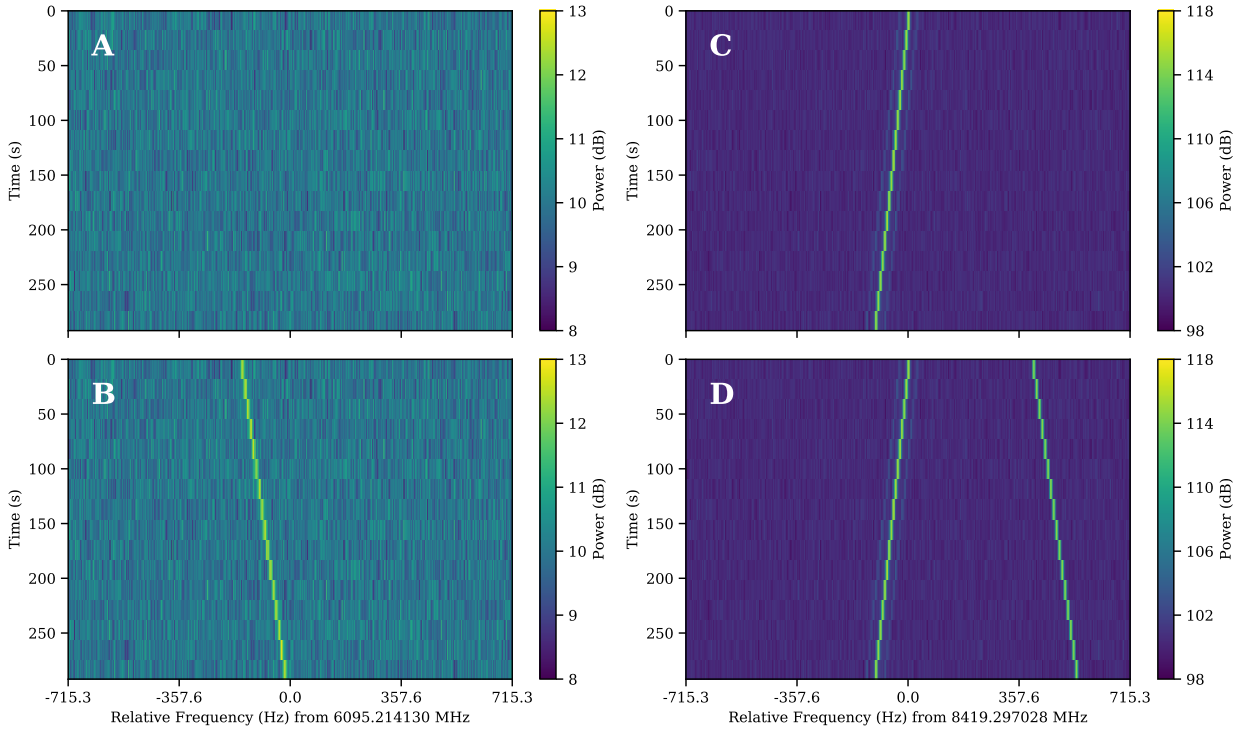


Figure 3.1: Radio spectrogram plots created from `setigen` frames. **A**: Frame with only synthetic chi-squared noise. **B**: Frame from panel A with an injected synthetic signal at SNR=30. **C**: "Real" GBT observation of Voyager I carrier signal at X-band. **D**: Frame from panel C with an injected synthetic signal at SNR=1000, with the same drift rate as the injected signal in panel B.

In these cases, we can instead use NumPy arrays to describe these signals, rather than functions. As of now, the `path`, `t_profile`, and `bp_profile` arguments can be arrays.

Common Frame Operations

Besides supporting noise and narrow-band signal injection, `setigen` comes with a set of tools for radio spectrogram analysis. These range from convenience functions for parameter calculations to frame-level data transformations.

For instance, estimating the SNR of a signal in an integrated spectrum is a common step in radio analysis. This can be done through a frame's `integrate` function, which can also be used along the frequency axis to produce an intensity time series array.

To inject a signal at a desired SNR, the `get_intensity` function calculates the requisite

signal level as

$$I_t = \text{SNR} \cdot \frac{\sigma_b}{N_t^{1/2}}, \tag{3.22}$$

assuming that the frame has background noise with standard deviation σ_b and that the SNR is measured by dividing the integrated signal maximum by the integrated noise deviation. As discussed in Section 3.3.1, each frame tracks an estimate of σ_b calculated using iterative sigma clipping and updates it when synthetic noise is injected.

It can be convenient to define signals in terms of the pixels they traverse rather than the frequencies. To convert between these for a given frame, one can use the `get_frequency` and `get_index` functions. We define the *unit drift rate* for a given spectrogram resolution to be the drift rate given by

$$\dot{\nu}_1 = \frac{\Delta f}{\Delta t}, \tag{3.23}$$

which can be accessed with the `unit_drift_rate` attribute. For a linearly-drifting signal passing through the top and bottom of the frame, the corresponding drift rate can be calculated using the `get_drift_rate` function.

Given a frame with a linearly-drifting signal, we can "de-drift" the frame using `setigen.dedrift`. This shifts each spectrum an appropriate amount along the frequency direction so that such a signal would, on average, appear to have zero frequency drift, making it simpler to calculate the SNR. In practice, empirical drift rates are not generally multiples of the unit drift rate, so de-drifted signals will not be perfectly aligned.

We can create a "slice" of a frame by specifying left and right frequency indices, analogous to NumPy array slicing, by using the frame's `get_slice` function. This results in a new frame with a truncated range, which can be helpful for isolating signals in time-frequency space for further analysis.

If one is interfacing with other BL or astronomy codebases, outputting `setigen` frames to filterbank or HDF5 format can be very useful. These are done via the `save_fil` and `save_hdf5` functions. Frame objects can also be written and loaded with `pickle`, a convenient serialization method that can keep data and user-provided metadata together.

Demonstration: Spectrogram Module

We present a minimal working example of creating a data frame with synthetic noise and a drifting signal. First, we construct an empty frame with the desired resolution; here, we use parameters that match those of BL's high frequency resolution data product:

```
from astropy import units as u
import setigen as stg

frame = stg.Frame(fchans=256,
                  tchans=16,
                  df=2.7939677238464355*u.Hz,
```

```
                dt=18.253611008*u.s,
                fch1=6095.214842353016*u.MHz)
```

Then, we add chi-squared noise with a desired mean, such as 10:

```
frame.add_noise(x_mean=10, noise_type='chi2')
```

Finally, we add a simple drifting signal through our frame at SNR=30 and plot the result in decibels (dB). The inputs to `add_signal` shown below are pre-written library functions that themselves return the functions described in Section 3.3.1. Since they are indeed Python functions by type, the signal parameters allow for much more flexibility beyond this basic example.

```
frame.add_signal(
    stg.constant_path(
        f_start=frame.get_frequency(index=100),
        drift_rate=2*u.Hz/u.s
    ),
    stg.constant_t_profile(
        level=frame.get_intensity(snr=30)
    ),
    stg.gaussian_f_profile(width=10*u.Hz),
    stg.constant_bp_profile(level=1)
)

frame.plot()
```

The frames after adding noise and after adding the signal are shown in Figures 3.1A and 3.1B.

We also show an example with a signal detected from Voyager I in an X-band observation using the GBT, in Figure 3.1C. Injecting a signal into the Voyager frame with the same drift rate as in the example (Figure 3.1B), now at SNR=1000, we get Figure 3.1D.

3.3.2 Raw Voltage Module

The raw voltage module is designed for synthesizing complex voltage data, providing a set of classes that models the signal processing pipeline described in Section 3.2. Instead of directly synthesizing spectrogram data, we can produce real voltages, pass them through a virtual pipeline, and record to file in GUPPI raw format. As this process models actual hardware used by BL for recording raw voltages, this enables lower level testing and experimentation.

The basic signal flow is shown in Figure 3.2. At the lowest level, a `DataStream` can accept noise and signal sources (as Python functions) and generate real voltages on demand. An `Antenna` models an antenna or dish used in radio telescopes and has one or two `DataStream`

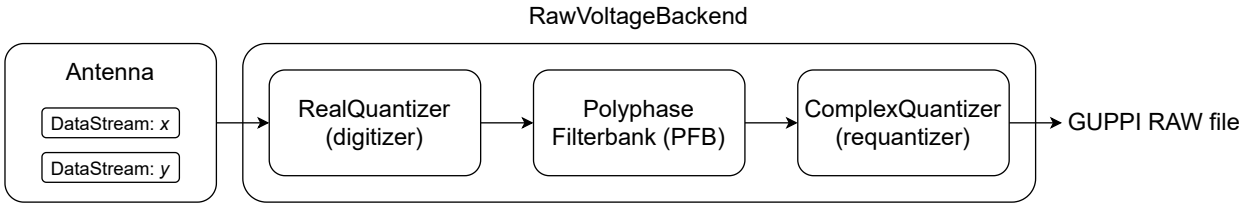


Figure 3.2: Basic layout of a voltage pipeline written using `setigen.voltage`.

objects, corresponding to linear polarizations that are unique and not necessarily correlated. As described in Section 3.2, the sampled real voltages are passed to a processing pipeline which consists, at its core, of a digitizer, a polyphase filterbank (PFB), and a requantizer. In hardware, processing is done in fixed point arithmetic on an FPGA, but for simplicity, we use floating point. The digitizer quantizes input voltages to a specified number of bits and a target full width at half maximum (FWHM) in the quantized voltage space. The filterbank implements a software PFB, coarsely channelizing input voltages. The requantizer takes the resulting complex voltages and quantizes each component to either 8 or 4 bits, suitable for saving into GUPPI raw format.

The `RawVoltageBackend` object wraps around these elements and connects the full pipeline together. Given an observation length in seconds or a number of data recording "blocks," the main function `record` retrieves real voltage samples as needed and passes them through each backend element, finally saving the quantized complex voltages out to disk.

Since voltage data is taken with very high sample rates, e.g. Gigasamples/sec (Gsps), the voltage module is much more computationally expensive than the spectrogram module. To increase efficiency, most of the data manipulations are done with matrix operations, allowing for GPU acceleration with `CuPy` (Okuta et al. 2017).

Antennas and DataStreams

The `DataStream` class represents a stream of real voltage data for a single polarization and antenna. A data stream has an associated sample rate f_s, such as 3 GHz for the BL DR. As of now, the voltage module does not implement heterodyne mixing or bandpass filtering. Instead, data streams use a reference frequency `fch1` and frequency sign (ascending or descending from `fch1`) for voltage calculations.

The `Antenna` class is similarly defined by a sample rate, reference frequency, and frequency sign. For two linear polarizations, an `Antenna`'s data streams are available via the `x` and `y` attributes. For one polarization, only the former is available. For convenience, the `streams` attribute gets the list of available data streams for an antenna. One can add noise and signal sources to these individual data streams.

Real voltage noise is modeled as ideal Gaussian noise and added through the `add_noise` function. Note that this actually stores a Python function to the data stream that is only evaluated when `get_samples` is called. It also updates the data stream's `noise_std` attribute, which keeps track of the standard deviation of the voltages in that data stream. This is useful for injecting signals at target spectrogram SNRs.

Drifting cosine signals can be added to a data stream using `add_constant_signal`. For more complex signals, one can write custom voltage functions to add using `add_signal`. Voltage signal sources are Python functions that accept an array of timestamps and output a corresponding sequence of real voltages. Here is a simple example that adds a non-drifting cosine signal with frequency `f_start`:

```
def cosine_signal(ts):
    delta_f = f_start - antenna.x.fch1
    return np.cos(2 * np.pi * delta_f * ts)

antenna.x.add_signal(cosine_signal)
```

As custom signals are added, the `noise_std` parameter may no longer accurately reflect the background noise. In these cases, one can run the data stream's `update_noise` function to estimate noise empirically. This is not done by default to save computation, especially when there are multiple well-behaved voltage sources (e.g. Gaussian noise, cosine signals).

Quantization

The quantization process takes a continuous input voltage distribution and scales it to a target distribution that can be described by N_{bits} bits. Since real voltage noise can be modeled by a Gaussian process, we can define this scaling in terms of the standard deviation or FWHM.

For real voltages $\{v\}$, target bit size N_{bits}, target mean μ_q (ideally 0), and target standard deviation σ_q, the quantized voltages v_q are given by:

$$v_s = \left\lfloor \frac{\sigma_q}{\sigma_v}(v - \langle v \rangle) + \mu_q \right\rfloor \tag{3.24}$$

$$v_q = \min(\max(-2^{N_{\text{bits}}-1}, v_s), 2^{N_{\text{bits}}-1} - 1) \tag{3.25}$$

We can define quantizers in terms of a target FWHM w_q, in which case $\sigma_q = \frac{w_q}{2\sqrt{2 \ln 2}}$.

The digitizer quantizes real voltages, while the requantizer receives complex voltages and quantizes per complex component. Quantization is run per polarization and antenna, and background statistics can be cached to save computation in subsequent calls. This is facilitated by the `RealQuantizer` and `ComplexQuantizer` classes.

Polyphase Filterbank

The `PolyphaseFilterbank` class implements and applies a PFB to quantized input voltages. Instead of directly applying a P-point FFT, a PFB first splits incoming voltages between P branches and lets M samples accumulate in each branch (Price 2021). A windowing function is applied over the $M \times P$ samples, the samples are summed over the M so-called polyphase taps, and finally a P-point FFT is taken of the result to get complex raw voltages in $N_{\text{coarse}} = P/2$ coarse channels. Further samples are read in groups of P and split between the PFB branches; accumulated samples step forward to the next tap to make room. PFBs have a better channel response than standard FFTs, especially as M increases, and are common in high spectral resolution radio backends (Price 2021).

The two main parameters for a `PolyphaseFilterbank` are the number of taps M and the number of branches P. Since the PFB works on MP samples at once, the object continuously caches samples for on-demand computation. The PFB also accepts a symmetric windowing function as an argument (Hamming, by default) and generates MP coefficients up front (Blackman and Tukey 1958).

Combining Components and Recording Data

The `RawVoltageBackend` class contains the full machinery to collect, process, and write complex voltage data to GUPPI raw files, as in the standard pipeline shown in Figure 3.2. Nevertheless, since the individual signal processing components are all exposed as part of the voltage module, custom pipelines can be written by chaining them in different ways.

A `RawVoltageBackend` takes in components external to the data recording process as parameters, such as the antenna, digitizer, PFB, and requantizer. All other parameters and functions are specific to data recording and actually obtaining data from the external components.

As described by Lebofsky et al. 2019, the block size $N_{\text{blocksize}}$ refers to the number of bytes in a single block of data in GUPPI format. Each data block has an associated header with observing metadata, such as target and frequency information. The number of blocks per file also must be specified to size individual raw files; multiple raw files may be associated with a single pointing. For standard 5 minute GBT observations, BL DR uses $N_{\text{blocksize}} = 134217728$ with 128 blocks per file.

To specify the coarse channels that are actually recorded to disk, we can set the starting index and the number of consecutive channels N_{chan} to ultimately save. Purely for computational efficiency, we always perform a full FFT and truncate to obtain the desired coarse channels, instead of directly doing the transform operation on the subset of coarse channels. Especially when using a GPU to accelerate synthesis, this can fill up memory rather quickly, potentially to the point of overflow. Therefore, the `RawVoltageBackend` has an additional option to divide individual data blocks into a given number of sub-blocks, such that each sub-block will fully fit in memory.

For a single antenna, the number of bytes $N_{\text{blocksize}}$ in a block can be related to the number of time channels $N_{t,\text{block}}$ corresponding to a single block in (non-integrated) spectrogram format as

$$N_{\text{blocksize}} = 2N_{\text{pol}} \left(\frac{N_{\text{bits,r}}}{8} \right) N_{\text{chan}} N_{t,\text{block}} \tag{3.26}$$

$$= \frac{1}{4} N_{\text{pol}} N_{\text{bits,r}} N_{\text{chan}} N_{t,\text{block}}, \tag{3.27}$$

based on the structure of raw files as described by Lebofsky et al. 2019.

Multi-Antenna Support

To simulate voltage data for interferometric pipelines, it can be useful to synthesize raw voltage data from multiple antennas. `setigen` supports synthesizing multi-antenna output through the `MultiAntennaArray` class, which creates a list of $N_{\texttt{ant}}$ antennas each with an associated integer delay (in time samples). In addition to the individual data streams that allow the user to add noise and signals to each antenna, there are "background" data streams `bg_x` and `bg_y` in `MultiAntennaArray`, representing correlated noise or RFI that is detected at each antenna, subject to the (relative) delays. Signals and noise can therefore be added to the background across all array elements as well as to individual antennas.

The only difference in the pipeline is instead of supplying a `Antenna` as input to a `RawVoltageBackend`, one would supply a `MultiAntennaArray`. Then, the output is saved as a multi-antenna extension of the GUPPI raw format.

Creating Signals at a Target Spectrogram SNR

During the course of the full signal processing pipeline, an injected cosine signal passes through multiple quantization and FFT steps. In many SETI experiments, a signal's SNR in spectrogram data is used for thresholding and analysis, so it is important to be able to estimate this SNR given pipeline parameters.

Suppose that we have a cosine signal with amplitude A at a frequency corresponding to the center of a fine spectral channel, and that this signal is injected onto a background of Gaussian noise $\mathcal{N}(0, \sigma_v^2)$. Since the voltage data is real-valued, the signal magnitude becomes $A/2$ in frequency space. As the voltages pass through the coarse and fine channelization steps, the signal magnitude picks up factors of P and N_{fine}, respectively, compared to the background noise.

The background noise will follow a chi-squared distribution with DOF $= 2N_{\text{pol}}N_{\text{int}}$ (Section 3.3.1), scaled by multiplicative factors arising from quantization and FFT calculations. Since the input voltage noise has variance σ_v^2, the standard deviation of the noise power σ_b will be proportional to the standard deviation $\sigma_{b,0}$ of a chi-squared distribution with mean σ_v^2. The time integration step to get the SNR will reduce this noise by a factor of $N_t^{1/2}$.

To get an expression for N_t given observation parameters, suppose our synthetic observation has N_{block} total blocks and that the time covered by a single block is τ_{block}. Then, we have the following equations:

$$\Delta t = \frac{N_{\text{int}}}{\Delta f} = \frac{P}{f_s} N_{\text{fine}} N_{\text{int}} \tag{3.28}$$

$$\tau_{\text{block}} = N_{t,\text{block}} \Delta t \tag{3.29}$$

$$N_t = \frac{N_{\text{block}} \tau_{\text{block}}}{N_{\text{int}} \Delta t} = \frac{N_{\text{block}} N_{t,\text{block}}}{N_{\text{int}}}. \tag{3.30}$$

Combining all of these factors, we can express the final SNR of the signal as the ratio between the integrated (mean) signal power and the integrated background noise standard deviation as

$$\sigma_{b,0} = \sigma_v^2 \left(\frac{2}{\text{DOF}} \right)^{1/2} \tag{3.31}$$

$$\text{SNR} = \frac{I}{\sigma_b} = \frac{(A/2)^2 P N_{\text{fine}}}{\sigma_{b,0} / N_t^{1/2}}. \tag{3.32}$$

This yields the amplitude or signal level in terms of the target SNR:

$$A = \left(\text{SNR} \cdot \frac{4 \sigma_{b,0}}{P N_{\text{fine}} N_t^{1/2}} \right)^{1/2} \tag{3.33}$$

Notice that A has a linear dependence on the standard deviation σ_v of the real voltage noise in a data stream, which can arise from multiple sources, especially in a multi-antenna array. Given pipeline parameters, the `get_level` function can be used to calculate A/σ_v.

For a non-drifting cosine signal, we can also approximate the effect of spectral leakage between fine channels by comparing the signal frequency to the nearest channel central frequency. A signal with amplitude A centered at a frequency δf away from the center of the closest fine spectral channel will have its power I attenuated by[10]

$$\frac{I'}{I} = \text{sinc}^2 \left(\frac{|\delta f|}{\Delta f} \right). \tag{3.34}$$

Since intensity goes as voltage squared, we provide a function `get_leakage_factor` to calculate an amplitude adjustment factor f_l to easily scale from A to a new amplitude A' that corresponds to the non-attenuated intensity:

$$f_l = \frac{1}{\text{sinc} \left(\frac{|\delta f|}{\Delta f} \right)} \tag{3.35}$$

$$A' = f_l A. \tag{3.36}$$

[10] $\text{sinc}\ x = \sin x / x$

Finally, for a linearly-drifting cosine signal, if the drift rate $\dot{\nu}$ exceeds the unit drift rate $\dot{\nu}_1$, signal power will be smeared across multiple frequency bins in spectrogram data. This is a linear effect in spectrogram data, so cosine amplitudes should be increased by a factor of $(\dot{\nu}/\dot{\nu}_1)^{1/2}$ to counter-act the apparent loss in power.

Injecting Synthetic Signals into Raw Voltage Data

In addition to creating fully synthetic complex voltage data from scratch, the `RawVoltageBackend` supports injecting or adding synthetic data to existing observational GUPPI raw data. The pipeline remains mostly the same, except for a few important differences that we detail below.

In order to get meaningful results, we must know and match details about the specific signal processing pipeline that produced the existing raw data. `setigen` provides a helper function called `get_raw_params` to extract header information from the raw data file, but other information must be provided separately by the user, such as the sampling rate and PFB parameters.

Since recorded voltage data has already gone through multiple quantization steps, we cannot directly add time series voltages together (i.e. at the original ADC sampling rate). Instead, we choose to synthesize complex voltage data separately, add it to the recorded voltage data, and apply a final quantization step to match the initial distribution as best as possible.

However, this process requires that we create and process signals that are not necessarily embedded in noise. In typical narrow-band signal injection scenarios, we wish to synthesize and inject signals whose distributions are non-Gaussian (e.g. a cosine signal). Since the quantization steps assume that the input and output voltage distributions are both Gaussian, attempting to quantize bare narrow-band signals will cause distortion and introduce clipping artifacts. Furthermore, without a reference noise distribution, quantization can scale the magnitude of processed signals in undesired ways, making SNR estimation difficult.

To address these issues, we approach the quantization steps differently. If there is already a synthetic noise source, we proceed normally through all steps in the pipeline. Otherwise, we skip the initial digitization step before the PFB, and instead treat the input voltages as if they followed a zero-mean Gaussian distribution with variance 1. Using a reference distribution allows us to set signal magnitudes with the `get_level` function to achieve target SNR levels. We then estimate the post-PFB mean and standard deviation of the reference Gaussian voltages and quantize the synthetic voltages based on these values instead of those from the "real" synthetic distribution. This way, if the synthesized voltages were actually embedded in $\mathcal{N}(0, 1)$ noise, the resulting signal quantization would be very similar.

For each data block in the recorded raw file, the `RawVoltageBackend` will set requantizer statistics (target mean μ_q and target standard deviation σ_q) calculated from the existing data for each combination of antenna, polarization, and complex component. The synthetic voltages are requantized to the corresponding standard deviations in each complex component, but instead of centering to the target mean, they are centered to zero mean. This is

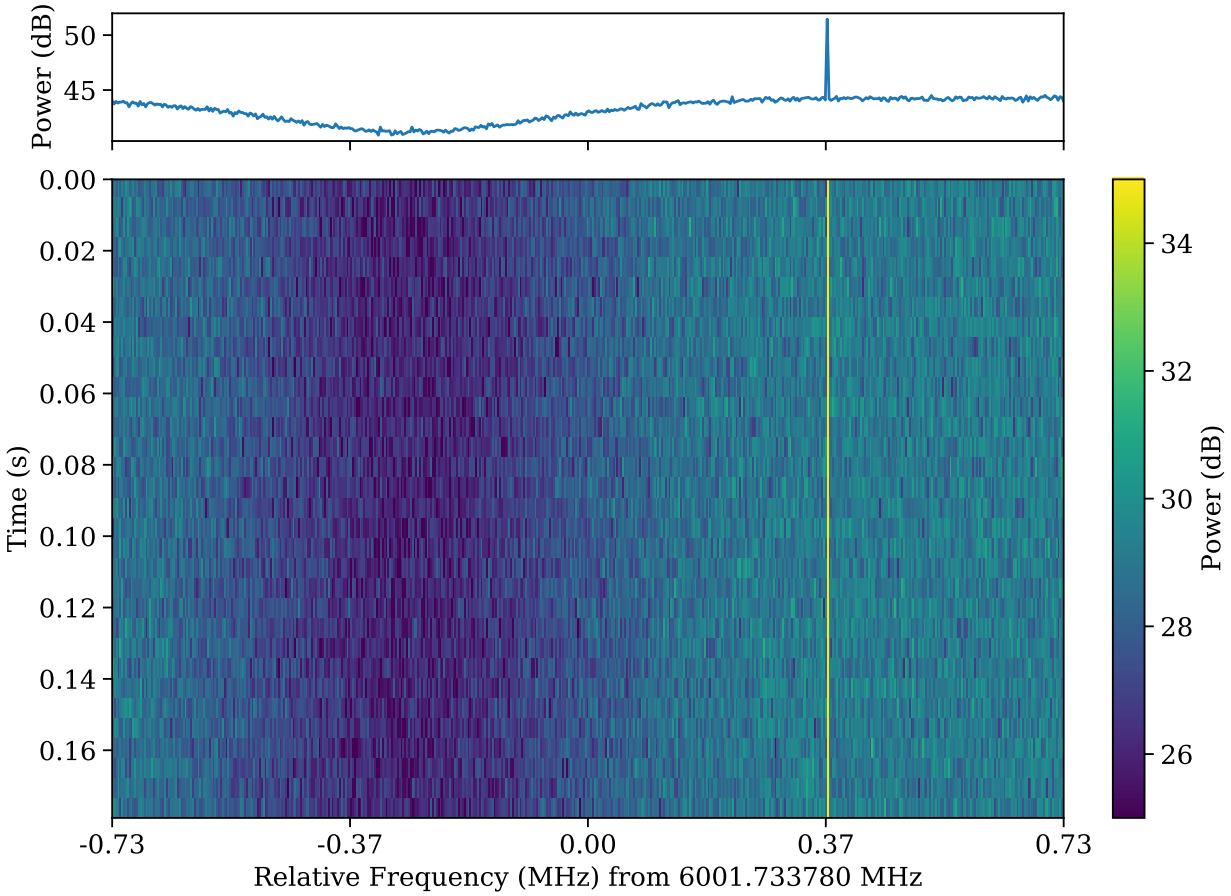


Figure 3.3: Spectrogram derived from synthetic raw voltages, showing the edge of the coarse channel bandpass shape and a bright, slightly drifting cosine signal. The top panel shows an integrated profile, showing PFB scalloping loss towards the left and the synthetic signal towards the right.

so that when we add the quantized synthetic data to the existing data, we do not change the overage voltage mean. After these are added together, we finally requantize once more to the target mean and target standard deviation to match the existing data statistics and magnitudes as best as possible.

Demonstration: Voltage Module

Here, we present a simple pipeline created with the raw voltage module to inject a drifting cosine signal in Gaussian noise. First, we create the signal processing elements:

```
from astropy import units as u
from setigen.voltage import *

d = RealQuantizer(target_fwhm=32,
                  num_bits=8)

f = PolyphaseFilterbank(num_taps=8,
                        num_branches=1024)

r = ComplexQuantizer(target_fwhm=32,
                     num_bits=8)
```

Then, we create the antenna, setting the sampling rate and reference frequency. With two polarizations, we can add Gaussian noise and a constant amplitude, Doppler drifting cosine signal to both data streams:

```
a = Antenna(sample_rate=3*u.GHz,
            fch1=6000*u.MHz,
            ascending=True,
            num_pols=2)

for s in a.streams:
    s.add_noise(v_mean=0,
                v_std=1)

    s.add_constant_signal(f_start=6002.1*u.MHz,
                          drift_rate=-2*u.Hz/u.s,
                          level=0.004)
```

We connect these components through the recording backend, defining the dimensions and size of the final raw voltage data product, and record a block of data to file.

```
rvb = RawVoltageBackend(a,
                        digitizer=d,
                        filterbank=f,
                        requantizer=r,
                        start_chan=0,
                        num_chans=64,
                        block_size=134217728,
                        blocks_per_file=128,
                        num_subblocks=32)
```

```
rvb.record(output_file_stem='example_1block',
           num_blocks=1,
           length_mode='num_blocks',
           header_dict={'TELESCOP': 'GBT'},
           verbose=True)
```

After saving the raw voltages to disk, we reduce using `rawspec` with $N_{\text{fine}} = 1024$ and $N_{\text{int}} = 4$. A snippet of the resulting spectrogram output is shown in Figure 3.3, where intensities are plotted on a decibel scale. The signal is readily apparent, as is the frequency bandpass shape arising from the PFB.

3.4 Discussion

3.4.1 Limitations

While `setigen` is a flexible library that enables quick narrow-band dataset generation, it is important to discuss the limitations when using it for science.

First and foremost, `setigen` relies on heuristic, user-defined signals, rather than simulations from first principles. The search for technosignatures is necessarily informed by human bias, specifically applied via our assumptions about a technosignature's potential characteristics and morphology. It is possible that radiation from an extraterrestrial intelligence will exist in a form that we have not considered or designed searches for. Even when we consider only the problem of excision of anthropogenic RFI, we have to be careful when applying algorithms developed using the simplest of narrow-band signals. Although there might never be a way of fully emulating the breadth and variety of the RFI environment, `setigen` can still be used to generate labeled, complex signals to test the efficacy of new and existing algorithms.

In a similar vein, the spectrogram module enables users to quickly generate signals that "look" like the narrow-band signals we see in observations. However, since spectrogram signal injection does not have access to phase information, it is impossible to replicate the "correct" intensity statistics when adding a signal to integrated Stokes I noise. For example, adding a perfect cosine signal to zero-mean Gaussian noise in the voltage domain results in a non-central chi-squared intensity distribution in Stokes I data, but adding a signal with constant intensity directly to chi-squared noise in a spectrogram does not result in the same distribution (over the pixels occupied by that signal; McDonough and Whalen 1995). While this effect is negligible for high SNR signals, algorithms developed to target low SNR signals may suffer from intrinsic inaccuracies in the intensity statistics.

Signal injection in the complex voltage domain also has limitations since we are not able, in software, to directly add signals in the real (analog) voltage stage. Raw data is quantized multiple times in hardware, so the injection step has to take place using complex voltages that are quantized in a similar way. While fundamental steps in the pipeline are linear, such

as PFB operations (Eq. 3.10), quantization inherently breaks this linearity. Because of this, summing real and synthetic voltages that are independently processed can lead to artifacts and intensity discrepancies that would not arise if we could inject at the start of the signal processing pipeline.

3.4.2 Future Directions

`setigen` is written and developed with the needs of SETI researchers in mind, so new functionality and improvements are constantly being added. Here, we describe some potential enhancements that may be added in the near future.

As it stands, the spectrogram module is especially targeted at producing small frames with synthetic signals rather than injecting into large, broadband observations. While this suffices in many cases, it may be useful to inject within large data files in which frequency bandpass shapes significantly change the background intensities. For instance, for use in SNR estimation, `setigen` calculates background noise statistics over an entire frame rather than localized around the target signal injection frequency. For a large enough frame, this is both an inefficient and inaccurate calculation due to variable bandpass shapes. An improvement would be to localize the noise calculation to a window around the target injection site, as well as to similarly localize the signal injection calculation to prevent unnecessary computation.

The spectrogram module is also currently designed expressly to synthesize narrow-band signals. There are many similarities in both signal processing and experimental design between technosignature searches and searches for time-varying phenomena such as pulsars and fast radio bursts (FRBs); `setigen` could thus be expanded to include broadband signal injection (Zhang et al. 2018a; Gajjar et al. 2021).

An exciting potential addition is to use parameterized ML methods to create labeled, realistic signals. By taking ideas from style transfer, a synthetic RFI signal could be created by specifying heuristic parameters and having an ML model generate such a signal with RFI-like properties (Gatys et al. 2016; Dai et al. 2017). While generative adversarial networks (GANs) have been used before to create radio spectrograms (Zhang et al. 2018b), conditional GANs that accept input parameters might help produce more specific, labeled signals, which can be better for certain SETI experiments. Furthermore, better RFI modeling could help improve ML-based searches for astrophysical phenomena like FRBs in the presence of different classes of RFI.

Some of these enhancements may use a lot more computational power than the current synthesis process, so the option to GPU-accelerate the standard spectrogram module would be critical. Some of these enhancements may require a more careful look at file input/output methods when reading and writing large observational data files to avoid unnecessary or slow operations.

The raw voltage module can also be expanded to support alternate radio telescope configurations and backends, such as those behind interferometers like MeerKAT (Jonas 2009). While `setigen` already has basic multi-antenna functionality, it could be helpful to build on this with general-use utilities, such as routines that predict how a given injected signal

would appear across multiple antennas or beams. The voltage module could also support additional requantization and recording modes, such as 2 and 16-bit. As interferometer usage in modern radio SETI continues to grow, `setigen` capabilities can be extended to help test signal detection in commensal and beam-formed observations (Czech et al. 2021).

3.5 Summary

In this paper, we presented `setigen`, an open-source Python library for the creation and injection of synthetic narrow-band radio signals. `setigen` can produce both finely channelized spectrogram data and coarsely channelized complex voltage data. The spectrogram module is designed to be intuitive and quick to use to facilitate the construction of synthetic datasets for SETI experiments and testing. While the voltage module is more complex and computationally intensive, it enables analysis of signals that pass through a software-defined pipeline, which can be helpful in understanding the implications of the instrumentation pipeline itself in SETI searches.

`setigen` is constantly being improved with the needs of SETI research in mind. As open-source software, the library is freely available, and we encourage the SETI community to use and contribute to it.

3.6 Acknowledgements

Breakthrough Listen is managed by the Breakthrough Initiatives, sponsored by the Breakthrough Prize Foundation. The Green Bank Observatory is a facility of the National Science Foundation, operated under cooperative agreement by Associated Universities, Inc. We thank the staff at the Green Bank Observatory for their operational support.

Chapter 4

On Detecting Interstellar Scintillation in Narrowband Radio SETI

A version of this chapter was originally published as: Brzycki, B., Siemion, A.P., de Pater, I., Cordes, J.M., Gajjar, V., Lacki, B., and Sheikh, S., 2023. On Detecting Interstellar Scintillation in Narrowband Radio SETI. The Astrophysical Journal, 952(1), p.46.

To date, the search for radio technosignatures has focused on sky location as a primary discriminant between technosignature candidates and anthropogenic radio frequency interference (RFI). In this work, we investigate the possibility of searching for technosignatures by identifying the presence and nature of intensity scintillations arising from the turbulent, ionized plasma of the interstellar medium (ISM). Past works have detailed how interstellar scattering can both enhance and diminish the detectability of narrowband radio signals. We use the NE2001 Galactic free electron density model to estimate scintillation timescales to which narrowband signal searches would be sensitive, and discuss ways in which we might practically detect strong intensity scintillations in detected signals. We further analyze the RFI environment of the Robert C. Byrd Green Bank Telescope (GBT) with the proposed methodology and comment on the feasibility of using scintillation as a filter for technosignature candidates.

4.1 Introduction

The Search for Extraterrestrial Intelligence (SETI) aims to answer one of the most important scientific questions: are we alone in the universe? Complementing other subfields of astrobiology in the attempt to detect life outside our planet, radio SETI strives to detect and constrain the existence of technosignatures, signals that betray the presence of intelligent extraterrestrial civilizations.

Radio and microwave astronomy has played an important role in modern SETI since the initial suggestion by Cocconi and Morrison 1959 to search near the neutral hydrogen line at

1.42 GHz for continuous narrowband emission. Out of the whole electromagnetic spectrum, radio frequencies are a strong candidate for searches since such emission is expected to arise from advanced civilizations for a portion of their technological activity[1], radio photons are efficient to produce, and radio waves travel relatively unimpeded by the atmosphere, dust, and the ISM (Oliver and Billingham 1971; Siemion et al. 2014). Narrowband emission is particularly tantalizing as a discriminant from natural astrophysical radio phenomena, whose emission bandwidth is usually, at minimum, hundreds of Hz at microwave frequencies due to broadening effects (Tarter 2001). From the relative ease at which our own civilization produces continuous, Hz-width signals, we anticipate that extraterrestrial civilizations will similarly emit narrowband signals.

From the first dedicated radio search for technosignatures by Drake 1961, SETI experiments have vastly expanded along multiple axes to cover larger frequency bandwidths, higher resolutions, and additional signal types (Werthimer et al. 1985; Tarter 2001; Siemion et al. 2013; Wright et al. 2014; MacMahon et al. 2018; Price et al. 2018; Gajjar et al. 2021). The Breakthrough Listen (BL) initiative began in 2016 as the most comprehensive SETI search program to date, observing with large instantaneous bandwidths at facilities across the world, including the Robert C. Byrd Green Bank Telescope (GBT) in West Virginia, USA and the CSIRO Parkes telescope in New South Wales, Australia (Worden et al. 2017; MacMahon et al. 2018; Price et al. 2018).

While the technology used in radio SETI has developed and improved throughout the decades, the requirements for a theoretical technosignature detection have not changed significantly. Narrowband signals are assumed to be non-natural in origin, but there is yet an ever-present background of human-made radio interference (RFI), comprised of both ground and space-based transmissions. Having a robust way of differentiating technosignature candidates from RFI is paramount if we are to ever have a convincing detection (Horowitz and Sagan 1993).

The primary strategy for RFI rejection in radio SETI is sky localization. If a signal is detected in multiple telescope directions, it is considered RFI, since a bona fide extra-solar technosignature should originate from a single location on the sky. To this end, BL uses ON-OFF observations, in which different pointings on the sky are observed in a cadence according to a ABABAB or ABACAD pattern (Enriquez et al. 2017; Price et al. 2020). To further tighten the directional filter, we require that a signal must appear in all 3 ON (A) observations to be considered a candidate.

For a directional filter to properly work, signals must be continuous throughout the observational cadence. Ideally, a candidate would be detected in repeat observations localized in the sky, requiring even longer signal durations. However, as in terrestrial emissions, extra-solar narrowband signals could appear pulsed and otherwise have low duty-cycles. In such cases, signals could appear in only one or two ON observations in a cadence and for a subsection of those observations, causing them to be missed by current filters.

[1]Judging from the technological development of our own civilization, we expect intelligent civilizations to emit radio waves as intentional transmissions or as unintentional leakage from normal activity.

On the other hand, RFI can also appear in only ON observations. For example, RFI signals could exhibit intensity modulations that follow the observational cadence of 5 minutes a pointing, a false positive that would pass the directional filter. While we observe false positives like this in practice, having directional requirements still serves as an interpretable basis for determining candidates, which would induce follow-up observations for potential re-detection.

This begs the question: can we differentiate narrowband signals as RFI based on morphology alone? Since ETI signals must travel to us through interstellar space, are there effects that would be observable and sufficiently unique compared to RFI modulations?

One possibility is that radio frequency scattering effects, such as diffractive scintillation and spectra broadening, could imprint on extra-solar narrowband signals, altering them enough to be resolved and distinguished from terrestrial RFI. A signal filter based on astrophysical properties would be an important tool, when applicable, for evaluating candidate technosignatures. For signals that fail the directional filter, a scattering-based filter might preserve missed candidates; for those that pass, it would amplify the likelihood of a true detection.

Radio wave scattering has been studied extensively since the onset of radio astronomy. Weak scattering from the ionosphere and solar wind or interplanetary medium (IPM) was observed to scintillate radio emission from stars (Smith 1950; Hewish et al. 1964). Pulsars themselves were discovered during one such study, and subsequent pulsar observations revealed strong scattering from the ISM (Hewish et al. 1968; Scheuer 1968; Roberts and Ables 1982). Since then, much of our understanding of ISM scattering has come about by observing pulsars, especially by analyzing pulse broadening and intensity fluctuations in time-frequency space (Narayan 1992). This observational work has led to models describing the stochastic nature of scintillation and broadening.

Plasma effects on narrowband signals have been analyzed by Cordes and Lazio 1991 and Cordes et al. 1997. Spectral broadening from the IPM has been observed in the transmissions of artificial probes and studied extensively (Goldstein 1969; Woo and Armstrong 1979; Harmon and Coles 1983; Woo 2007). For the ISM, scintillation has been historically interesting to SETI as a factor that changes the detectability of a technosignature. Most of the time, the signal intensity is reduced, but occasionally the intensity will spike as a result of constructive interference. Cordes and Lazio 1991 recommend multiple observations spaced in time to maximize the chance of catching at least one detection.

In this work, we investigate the parameter space of scattering relevant to narrowband radio SETI and investigate whether resolved scattering effects can be used to flag technosignature candidates in the proverbial haystack of RFI. In Section 4.2, we review scattering theory relevant to narrowband signals. In Section 4.3, we introduce methods for identifying the presence of scintillation in radio spectrogram data and for producing synthetic scintillated intensity time series. In Section 4.4, we present an approach for estimating likely scattering properties as a function of observation parameters using the NE2001 model. In addition to examining theoretical properties of scintillated narrowband signals, in Section 4.5, we perform a statistical analysis on detected narrowband signals in multiple radio bands

using the GBT. We compare properties of real RFI signals with those of theoretical scintillated ETI signals to determine the conditions under which scattering effects can be used as effective SETI filters. Finally, we summarize our results, discuss limitations, and give recommendations on potential scintillation-based technosignature searches in Section 4.6.

While examples in this paper use certain values for observational parameters, such as observation length and time resolution, the methods developed in this work are meant to be broadly applicable to various radio observations. As such, we provide a Python library `blscint`[2] that implements many of the key components of our scintillation search methodology.

4.2 Scattering Theory and SETI

Observational and theoretical work on radio scattering have been done to characterize both the bulk power spectrum of electron density fluctuations as well as the effect of localized ionized scattering structures along the line of sight (Rickett 2007). In this work, we limit our considerations to the wavenumber spectrum of ISM plasma fluctuations as a first order approximation of scattering along any line of sight.

The dominant effect causing radio scattering in ionized plasma is refraction due to variations in electron density. The changes in refractive index give rise to changes in phase when a plane radio wave is passing through the scattering layer. These phase variations, along with path-induced phase delays, are propagated to the observer's plane, creating an interference pattern.

Since ionized plasma is a complex, stochastic medium, it is most useful to describe the power spectrum of turbulent scales. In practice, it is common to use the phase structure function:

$$D_\phi(x, y) = \langle [\phi(x + x', y + y') - \phi(x, y)]^2 \rangle_{x',y'}, \tag{4.1}$$

where x, y are coordinates in the scattering plane. This equation can also be expressed in terms of a vector baseline $\boldsymbol{r} = \langle x, y \rangle$, which is useful when describing interferometer measurements. For single dish measurements, this "baseline" is set by the relative transverse velocity V_T of the diffraction pattern during an observation of length τ, so that $r = V_T \tau$. Here, we assume that the pattern is effectively "frozen," in that V_T dominates the intrinsic random motion of material in the scattering medium. The structure function is usually taken to be a power law in wavenumber (length scale), so that

$$D_\phi(r) \propto r^\alpha \tag{4.2}$$

for some power α (Rickett 1990; Narayan 1992).

The phase spectrum of the scattering medium determines the type of diffraction pattern seen by the observer, so it is important to constrain this at a high level. A common assumption is that ionized scattering media are isotropic and follow Kolmogorov turbulence, such

[2] https://github.com/bbrzycki/blscint

that energy cascades from large turbulent structures with an outer length scale down to an inner length scale. Long-term pulsar observations show evidence that ISM scattering exhibits a Kolmogorov spectrum over many orders of magnitude (Ramachandran et al. 2006). Kolmogorov turbulence is described by $\alpha = 5/3$ in Equation 4.2.

Another important case of turbulence is the square-law regime, for which $\alpha = 2$. This typically applies when the spatial wavenumber probed by the observation (i.e. $r = V_T\tau$) is smaller than the inner scale. This regime yields nice analytical expressions for scattering behavior, such as the spectral broadening function being a Gaussian. Some ISM scattering studies have accordingly used Gaussian models derived using $\alpha = 2$ as approximations for the Kolmogorov case ($\alpha = 5/3$; Roberts and Ables 1982; Cordes 1986; Gupta et al. 1994).

4.2.1 Weak and Strong Scattering

Since turbulence and scattering are inherently stochastic processes, it helps to compare characteristic scales to describe the underlying physics.

The so-called diffractive length scale r_{diff} is defined as the characteristic transverse distance over which the root mean square phase difference is 1 rad. This can be compared with the Fresnel radius r_F, which describes the size of the largest cross-section along the observer-source path for which waves arrive coherently in free space, with path-induced phase delays less than π.

If $r_{\text{diff}} \gg r_F$, we are in the weak scattering regime, in which refractive phase changes are small compared to path-induced phase differences and the characteristic size of a coherent emission patch on the sky is r_F (Narayan 1992). If $r_{\text{diff}} \ll r_F$, we are instead in the strong scattering regime, in which the characteristic coherent patch size becomes r_{diff}, and plasma-induced phase changes span many radians over the Fresnel radius. The strength of scattering depends on a variety of factors, such as the free electron number density, the strength of turbulence, the emission frequency, and the distance of the source. Along a given line of sight, the scattering strength increases and eventually transitions from weak to strong (Cordes and Lazio 1991). The transition distance, for which $r_{\text{diff}} \sim r_F$, depends on the emission frequency.

In the strong scattering regime, there are two types of scintillation. Diffractive scintillation is relatively fast (on order minutes to hours) and requires a compact source, such as a pulsar, while refractive scintillation is weaker and slower (on order days to years) (Narayan 1992). Diffractive scintillation arises from multi-path propagation from emission across the scattering medium, while refractive scintillation is a larger-scale geometric effect that can itself modulate diffractive scintillation effects. Since potential narrowband ETI emission would have a compact source, we focus on strong diffractive scintillation in this paper.

The "modulation index" m_d is the root mean square of the fractional flux variation due to scintillation. In weak scattering, $m_d \ll 1$, whereas in strong scattering, $m_d \sim 1$.

4.2.2 Effects of Strong Scintillation on Narrowband Signals

Pulsar observations are effective probes of intensity scintillations in time and frequency given their persistent, broadband signals. On the other hand, since narrowband signals are by definition restricted in spectral extent, we are mostly limited to studying temporal effects. To guide the discussion, we can write a basic model for the intensity of a scintillated narrowband signal:

$$I_{\text{scint}}(t) = g(t)S + N(t), \tag{4.3}$$

where $g(t)$ is the scintillation gain, S is the fixed intensity of the original signal, and $N(t)$ is the background noise.

One observable effect is that for independent observations, the detected signal intensity will follow an exponential probability density function (PDF):

$$f_g(g) = \exp(-g)H(g), \tag{4.4}$$

where H is the Heaviside step function (Cordes and Lazio 1991; Cordes et al. 1997). If we assume a continuous-wave (CW) transmitter and think of radio waves as complex phasors, we start with signals of constant amplitude modulus. As the signal refracts at different points across the scattering medium, it picks up random phase changes. Due to multi-path propagation, many independent de-phased versions of the signal are summed together at the observing plane. The asymptotic result is that an ISM scintillated signal can be modeled as a random complex Gaussian variable, whose amplitude follows a Rayleigh distribution and whose intensity therefore follows an exponential distribution (Goodman 1975).

Another effect arising from the statistical power density spectrum of plasma turbulence is that the diffraction pattern at the observing plane has a spatial autocorrelation function (ACF) with a characteristic spatial scale r_{diff}. Though this work limits discussion to the effects on narrowband signals, strong diffractive scintillations also have a spectral ACF with a characteristic scintillation bandwidth (also known as the decorrelation bandwidth).

For a single dish telescope taking a long radio observation, the diffraction pattern will sweep across the telescope at a relative transverse velocity, so that observations display a temporal ACF in diffracted intensity. In terms of the phase structure function, the temporal ACF of g is given by

$$\Gamma_I(\tau) = |\Gamma_E(\tau)|^2 = \exp\left[-D_\phi(V_T\tau)\right] \tag{4.5}$$

in the Rayleigh limit (Cordes and Lazio 1991; Coles et al. 2010). Note that in this work, we use the normalized autocorrelation.

The ACF thus has a representative timescale $\Delta t_d = r_{\text{diff}}/V_T$ over which scintillation occurs. By convention, Δt_d is measured as the half–width at $1/e$–height of the ACF, which has been historically estimated to be a Gaussian function. In other words,

$$\Gamma_{\text{sq}}(\tau) = \exp\left[-\left(\frac{\tau}{\Delta t_d}\right)^2\right]. \tag{4.6}$$

Statistic	Data Type	Theoretical Behavior	Asymptotic Value
Standard Deviation (RMS)	Intensity	Exponential	1
Minimum	Intensity	Exponential	0
Kolmogorov-Smirnoff Statistic	Intensity	Exponential	0
Autocorrelation Function ACF(τ)	Autocorrelation	Near-Gaussian	$\Gamma_I(\tau)$
Least Squares Fit for Δt_d	Autocorrelation	Near-Gaussian	Δt_d

Table 4.1: Diagnostic statistics chosen to probe theoretical scintillation effects. For each statistic, we list the type of data used for computation, the theoretical behavior of that data type, and the asymptotic value of the statistic (in the absence of noise) as the observation length goes to infinity.

However, under the Kolmogorov assumption, it is more precise to use

$$\Gamma_{\mathrm{k}}(\tau) = \exp\left[-\left|\frac{\tau}{\Delta t_d}\right|^{5/3}\right]. \tag{4.7}$$

The Kolmogorov form is near-Gaussian, as shown in Figure 4.1. In this work, we use the Kolmogorov form Γ_{k} throughout, but all methods can be performed with the square-law form as well.

We note that an additional scattering effect on narrowband signals is spectral broadening. This causes power at a single frequency to spread over a bandwidth

$$\Delta\nu_{sb} = C_2/(2\pi\Delta t_d), \tag{4.8}$$

where C_2 is a constant of order unity that depends on the scattering medium; $C_2 = 2.02$ is used in Cordes and Lazio 1991. However, at microwave frequencies, spectral broadening is typically smaller than commonly used frequency resolutions in SETI, so this effect would be difficult to observe except in lines of sight with extreme scattering.

4.3 Identifying Strong Scintillation in Detected Signals

Since scintillation is inherently stochastic, we have to use statistical indicators to identify its presence in a detected narrowband signal. Accordingly, we extract time series intensity data from signals in radio Stokes I spectrograms and identify several "diagnostic statistics" that probe the theoretical asymptotic behavior described in Section 4.2.2. For our scintillation analysis, we think of each signal detected within an observation of length τ_{obs} and spectrogram time resolution Δt as a sequence of $N_t = \tau_{\mathrm{obs}}/\Delta t$ statistically dependent random intensity samples drawn from the asymptotic distributions.

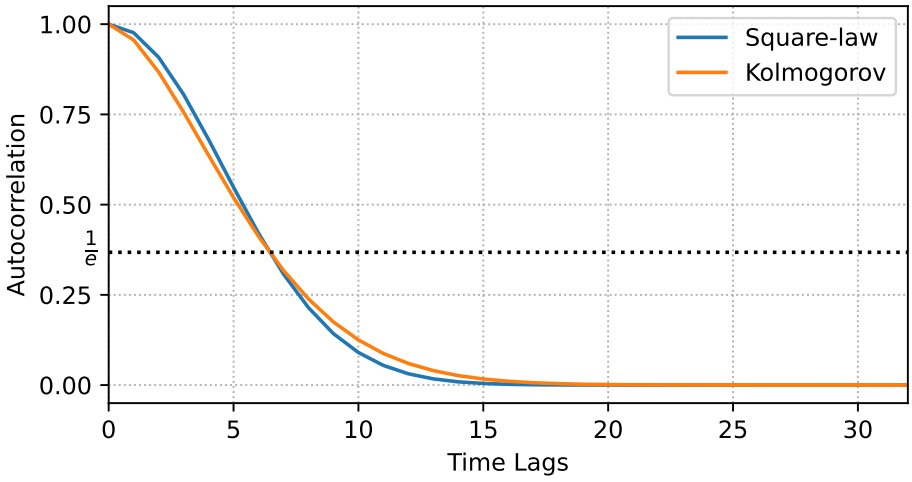


Figure 4.1: Comparison of the Kolmogorov and square-law ACF models. Both functions are computed using a scintillation timescale of $\Delta t_d = 30$ s and a time resolution of $\Delta t = 4.65$ s. The $1/e$-height is shown as a dotted line.

4.3.1 Diagnostic Statistics

Given time series intensity data for a detected narrowband signal, we can compute *diagnostic statistics* for the expected asymptotic behavior of a scintillated signal. This process is analogous to feature engineering in machine learning, where these statistical "features" are designed to have a physical basis behind them. The closer a given diagnostic statistic is to the expected asymptotic value, the higher likelihood the original signal is scintillated. As such, we can create thresholds using these statistics to function as filters for interesting candidate signals.

In this paper, we offer a few examples of useful diagnostic statistics, but note that the list is in no way exhaustive and that there may be other interesting statistical features that help determine whether a given signal may be exhibiting scintillations. These can be found in Table 4.1, as well as asymptotic values in the absence of noise.

First, we want statistics that can probe the expected exponential distribution of intensities. For this discussion, assume that the time series for an idealized scintillated signal is normalized to mean 1. The standard deviation of intensity samples lends itself naturally to evaluating the degree of scintillation and tends to 1 for a normalized exponential distribution. In other words, $m_d = (\langle g(t)^2 \rangle / \langle g(t) \rangle^2 - 1)^{1/2} \sim 1$ for strong diffractive scintillation.

For a strongly scintillated signal, we expect to see complete destructive interference, leading to a minimum intensity near 0. In reality, signals are embedded in random voltage noise, so that during periods of destructive interference, measured intensities can actually be below the mean noise level. As a necessary pre-processing step to help isolate signal

intensities (Section 4.5.1), we subtract the noise mean from data spectrograms, which can result in minimum signal "intensities" that are negative.

Another statistical measure that addresses this directly is the Kolmogorov-Smirnoff (K-S) statistic, which is used to compare a sample distribution to a target ideal distribution using the empirical cumulative distribution function (ECDF). In this case, we compute the K-S statistic against an ideal exponential distribution with rate $\lambda = 1$, keeping in mind that our time series have an assumed mean of 1. In practice, we do not know the actual mean intensities of our signals, so we can only estimate a sample mean as we normalize the time series to mean 1. So, instead of using established tables of statistic values to determine p-values, we use the statistic itself to set thresholds. The lower the K-S statistic for an intensity time series, the closer the intensities are to being exponentially distributed.

We must note that the assumption of an unmodulated CW signal, or at least a high-duty cycle signal, is important for these statistics. For example, radio transmissions on Earth are usually modulated, so for such signals, the exponential intensity distribution arising from scintillation would be convolved with the distribution of the modulation. If the modulation is faster than the spectrogram time resolution Δt, then the modulation averages out within time bins, essentially giving us a CW signal. However, if the timescale of modulation is in between Δt and $\tau_{\rm obs}$, it is likely that the intensities of the scintillated modulated signal would no longer be exponential at the observer.

A scintillated signal will yield a flux time series with a characteristic ACF width equal to Δt_d. From time series signal data, we can compute the ACF at all lags k, normalized to 1 at lag 0. We can then compare the empirical ACF with the theoretical model $\Gamma_{\rm k}$ by using raw values or by fitting with least squares. In the presence of noise, the ACF spikes at lag 0 compared to non-zero lags, since the random fluctuations add in quadrature. This is especially significant for low intensity signals. Instead of only using raw (normalized) ACF values, it is therefore more reliable to fit $\Gamma_{\rm k}$ and the noise spike in one shot using least squares and to derive the corresponding scintillation timescale Δt_d. Following the treatment in (Reardon et al. 2019), we fit the following expression to the empirical ACF:

$$\Gamma_{\rm k,n}(\tau) = A\Gamma_{\rm k}(\tau)\Lambda(\tau, \tau_{\rm obs}) + W\delta(\tau), \tag{4.9}$$

where A, W are multiplicative factors, δ is the Kronecker delta or discrete unit impulse function, and Λ is the triangle function with zeros at $\pm\tau_{\rm obs}$ used to model the sample autocorrelation. The least squares fit gives values for A, W, and Δt_d within $\Gamma_{\rm k}$. This process yields consistent results as if we first excluded lag 0 from the fit, which is also commonly done (Rickett et al. 2014). Since detected signals may be RFI and have complex ACFs, having values for A and W can help us identify and exclude poor fits (i.e. if A is close to 0, it is unlikely that the signal's ACF truly matches $\Gamma_{\rm k}$).

4.3.2 Constraints on Identifying Scintillation

There are various factors at play that affect the possibility of detecting scintillation. The first is that the time resolution must be high enough to sufficiently resolve scintles (scintillation

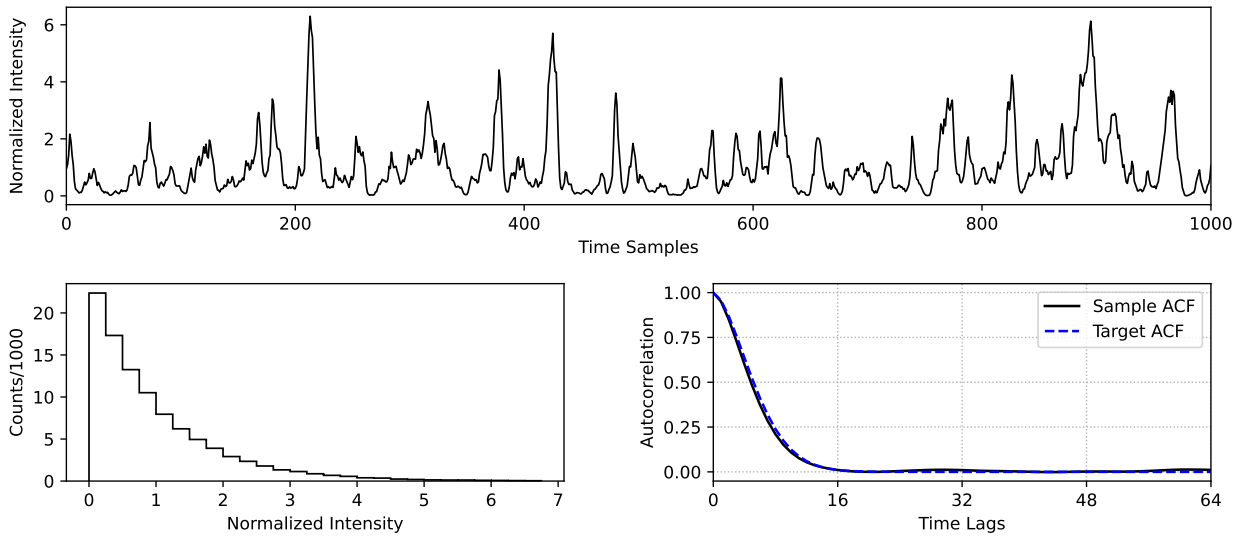


Figure 4.2: Synthetic scintillated intensities ($N = 10^5$) generated using ARTA, using a sample interval of $\Delta t = 4.65$ s and scintillation timescale $\Delta t_d = 30$ s. **Top**: Synthetic intensity time series data, showing first 1000 samples. **Bottom left**: Histogram of intensities, showing the expected exponential distribution. **Bottom right**: Sample ACF plotted up to lag 64, with the target ACF Γ_{k} shown overlaid.

maxima). Similarly, the integration time per observation has to be long enough to collect enough scintles for better convergence to the theoretical ACF.

However, the observation length should be short enough that the receiver gain is stable. Gain fluctuations would change the underlying noise as well as the detected signal intensities over time. While this is an effect that can theoretically be corrected for using data at signal-free frequencies, for practical purposes, it is simpler to limit the observation length such that we can assume gain stability. This further avoids the potential problem of basing calculations on a "signal-free" region in time-frequency space that in actuality is occupied by dim RFI that escaped detection.

The detected narrowband signal must be bright enough to compute accurate statistics while embedded in noise. Noise fluctuations in the time series representation of a scintillated signal's intensity will move the empirical distribution away from exponential and mask the ACF structure. Note that since the ACF of white noise is an impulse at lag 0 and that the ACF operation is linear for uncorrelated functions, we can still fit a scaled version of the ideal profile Γ_{k} for a scintillated signal's ACF, adding an additional term to fit for the noise impulse. However, for signals with low signal-to-noise ratios (S/N), the impulse will be the overwhelming part of the extracted ACF, which can make it harder to make an accurate fit.

As one might expect in radio SETI, the RFI environment is a significant obstacle for

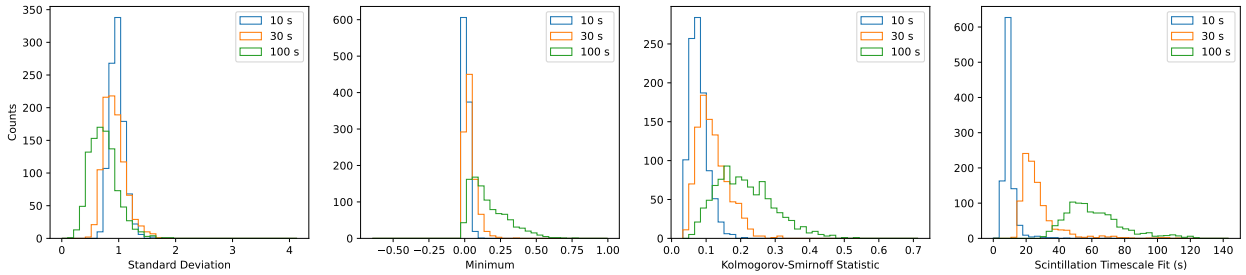


Figure 4.3: Histograms of diagnostic statistics computed using $N = 1000$ ARTA-produced intensity time series realizations for representative scintillation timescales of 10, 30, and 100 s. Each time series is produced using $\Delta t = 4.65$ s and $\tau_{\text{obs}} = 600$ s and does not include additive background noise. We plot histograms of the standard deviation, minimum, Kolmogorov-Smirnoff statistic, and least squares fit for the scintillation timescale, computed for each time series realization.

detection. Our present tools for detecting narrowband signals make simplifying assumptions as to the kinds of signals that we hope to be sensitive to. Broadband RFI can be modulated at different frequencies, so sometimes a bright enough broadband signal passes our S/N thresholds and is falsely flagged as a "narrowband" detection. Broadband RFI can also overlap real narrowband signals, majorly distorting the extracted intensity time series data. It is also possible that certain modulation schemes in narrowband RFI present confounding factors for scintillation detection; perhaps some forms of RFI already appear to be scintillated (at least according to the theoretical properties identified). In Section 4.5, we perform an initial analysis of the narrowband RFI environment at the GBT, computing the various diagnostic statistics and comparing them with those predicted for scintillated signals.

4.3.3 Synthesizing Scintillated Signals with Autoregressive-to-Anything (ARTA)

Since observations are necessarily limited in time, we have a finite number of samples per target. Furthermore, we work with large search parameter spaces for which there is a trade-off between the length of time per target and the number of targets searched. Unless a specific pointing is otherwise scientifically interesting, it may be more useful to spend a shorter integration time on a larger number of pointings. Taken together, in most cases, we will be working with a low number of time samples per observation, which implicitly adds measurement error to each diagnostic statistic.

We would like to better understand the relationship between observation parameters, the scintillation timescale, and the expected natural error in our diagnostic statistics. Since there are a number of factors involved, it is difficult to quantify the expected errors analytically.

Instead, we designed a method to create synthetic scintillated time series data, allowing us to compute the empirical distribution for each diagnostic statistic and observe the corresponding spread from the asymptotic values.

Theoretical studies have created models of scintillation phase screens and simulated light waves passing through each screen as a function of space and frequency, such as Coles and Filice 1984, Hamidouche and Lestrade 2007, Coles et al. 2010, and Ravi and Deshpande 2018. While this gives the best physical intuition for a given set of parameters, for our work, we need to be able to quickly produce a large quantity of synthetic scintillated narrowband signals over different scintillation and observation parameters. Since we are specifically interested in asking when scintillation might be detectable for SETI, we choose to rely on predictions from established theory to more efficiently create synthetic data rather than to generate our own rigorous simulations, although this may be a valuable direction for the future.

One method to produce synthetic scintillated data is to first compute the power spectrum S of scintillations using a Fast Fourier Transform (FFT) of the target autocorrelation (in the voltage domain, $\Gamma_{\mathrm{k}}^{1/2}$). One may then produce a complex voltage time series by taking the inverse FFT of complex Gaussian noise multiplied by $S^{1/2}$. Finally, taking the squared magnitude of the voltage series yields an intensity time series following an exponential distribution and ACF of Γ_{k}. While this method is relatively straightforward and satisfies asymptotic scintillation properties, we would like to present an alternative synthesis technique that may have broader uses in SETI for future applications.

Synthetic time series data following overarching statistical distributions can be produced using autoregressive models. Cario and Nelson 1996 developed a model called the "autoregressive to anything" (ARTA) process for generating time series data with arbitrary marginal distribution and autocorrelation structure (up to a specified number of lags). While this work focuses on the effects of scintillation on CW narrowband signals, having the ability to match arbitrary target distributions for first and second-order statistics could be useful for SETI applications that aim to model other astrophysical effects or even certain types of RFI.

In our case, the target marginal distribution is exponential and the autocorrelation structure is the near-Gaussian curve Γ_{k}. We construct ARTA processes to model the noise-free scintillation gain $g(t)$ of a 100% modulated narrowband signal over time. In the style of Equation 4.3, we can produce synthetic intensities with $I(t) = g(t)S$, for any choice of signal intensity S. Figure 4.2 shows an example of synthetic scintillated intensities generated in this way with $S = 1$, along with a histogram and ACF plot demonstrating the asymptotic behavior.

To construct an ARTA process Y_t, we provide a marginal distribution with cumulative distribution function (CDF) F_Y and an autocorrelation structure $\rho_Y = (\mathrm{Corr}[Y_t, Y_{t+1}], \ldots, \mathrm{Corr}[Y_t, Y_{t+p}])$, where p is the number of lags specified (Cario and Nelson 1996). Since the model is computed numerically, ρ_Y is finite, and the model will only attempt to match the ACF up to lag p. The computation involves solving the Yule-Walker equations for a $p \times 1$ vector of autoregressive process parameters, which in turn requires inverting a $p \times p$ matrix. This limits the number of lags out to which we can effectively compute, but

for scintillation analysis, this will rarely be an issue.

While this procedure results in an ARTA process with correlations close to ρ_Y, Cario and Nelson 1996 describe methods to improve convergence to the target correlations. By perturbing the input correlations to the model and doing a grid search in the parameter space, we can arrive numerically at final correlations that have higher accuracy. In this work and in `blscint` routines, we choose to forego this additional step, since it increases computational time significantly without much reward. Since using a finite observation length means that, by definition, we are performing small sample experiments, any marginal increase in the asymptotic correlation accuracy is quickly overshadowed by intrinsic sampling error.

With this tool, for any set of parameters $(\Delta t, \tau_{\text{obs}}, \Delta t_d)$, we can create datasets with many time series realizations to analyze the measurement error implicit in our limited-length observations. Note that we control the observational parameters, such as Δt and τ_{obs}, but not the scintillation timescale Δt_d. This implies that we should choose observational parameters in such a way that we minimize our measurement error with respect to the most likely scintillation timescales. So to make this process most useful, we should attempt to estimate the most likely or most detectable scintillation timescales; this is addressed in more detail in Section 4.4.

The parameter spaces involved are vast, but we can focus on representative values close to those commonly used in radio SETI today. In other words, we try to only make slight perturbations to observational parameters used by modern spectrogram searches and similarly limit the range of scintillation timescales to practically consider. Ideally, it will be possible to directly analyze SETI observations taken for other purposes for evidence of scintillation using the methods developed in this paper.

For example, suppose we want to evaluate our sensitivity to scintillation timescales in the range of 10–100 s. The high spectral resolution data format used by BL has 2.79 Hz and 18.3 s resolution for 5 minutes, resulting in 16 time samples per observation. If we instead take observations for 10 minutes at 4.65 s resolution, yielding 128 time samples, our diagnostic statistics are more accurate and sensitive to a larger range of scintillation timescales. With these parameters, we create synthetic noise-free time series observations with ARTA, compute the diagnostic statistics, and plot histograms of each as a function of scintillation timescale as shown in Figure 4.3.

The different scintillation timescales yield observable differences in the empirical probability density function for each diagnostic statistic. Panels 1–3 all show diagnostic statistics that target the asymptotic exponential distribution of intensities. As the scintillation timescale decreases and approaches the time resolution, each scintle will generally be covered by individual time samples. As $\Delta t_d \sim \Delta t$, the ACF structure becomes irrelevant and the observed intensity samples better match the theoretical intensity distribution. In each of Panels 1–3, the 10 s histogram is the tightest around the asymptotic statistic value, whereas the 100 s histogram has the largest spread and general deviation from the asymptotic value. As the scintillation timescale increases relative to the time resolution, more samples cover individual scintles, and so the ACF structure reduces the apparent exponentiality of the intensities within a single observation or time series realization. Panel 4 shows the least squares fit

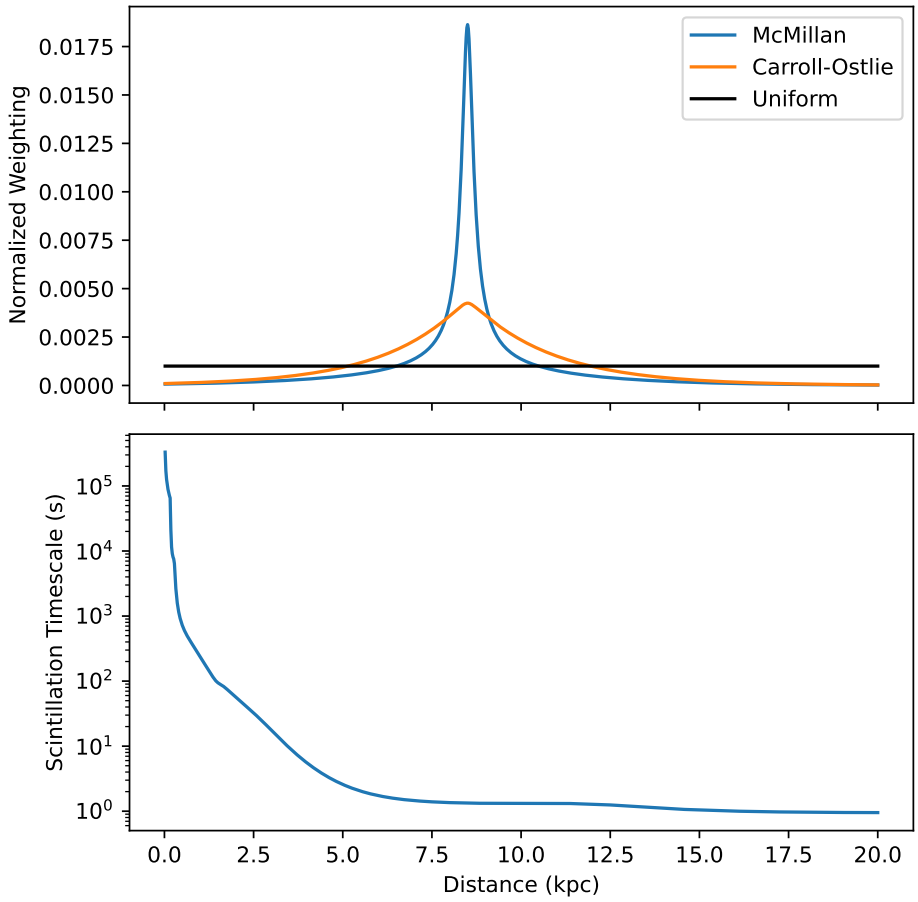


Figure 4.4: Comparison between methods for distance sampling, including uniformly, by stellar number density, and by stellar mass density. We use a line of sight of $(l, b) = (1, 0)$ out to a distance of 20 kpc. Bottom panel shows NE2001-produced scintillation timescales as a function of distance.

for the scintillation timescale; this similarly has the largest error for the largest scintillation timescales, since there are fewer scintles during the same observation length. Once again, note that here, the diagnostic statistics are calculated for time series intensities with no additive background noise to observe how a low sample count effects the measurement error.

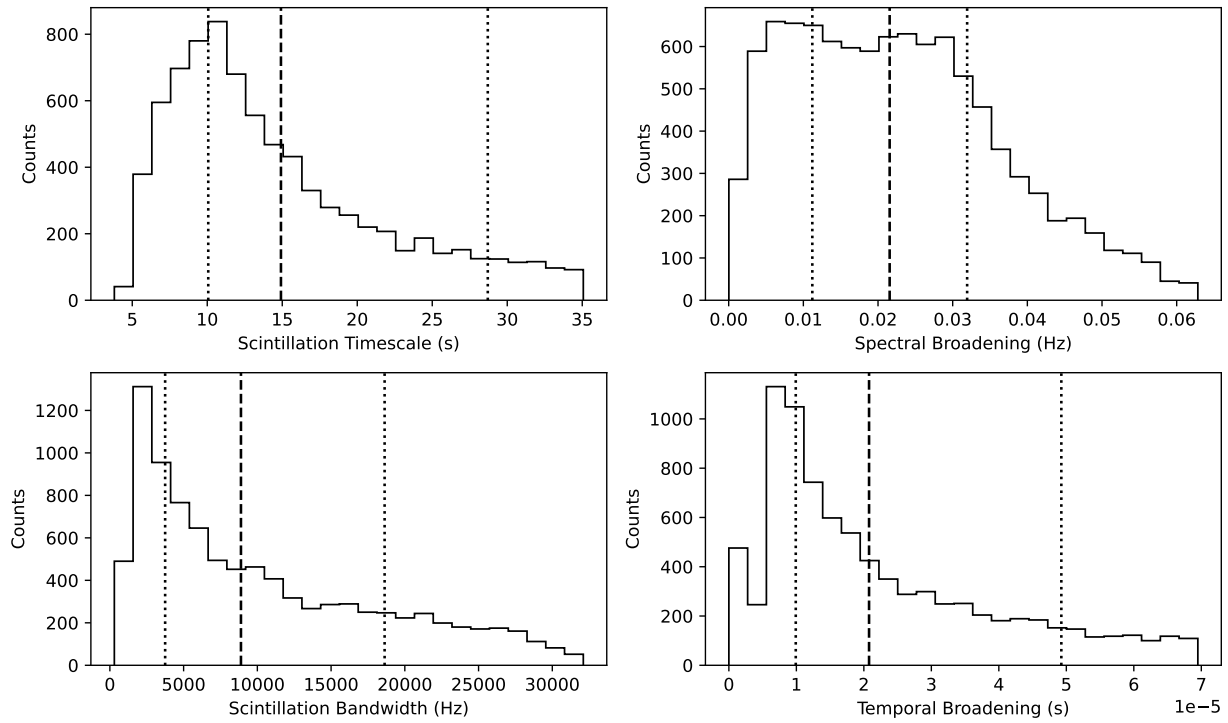


Figure 4.5: Set of Monte Carlo-sampled distributions of scintillation parameters at C-band, using $N = 10000$ realizations. We use a line of sight of $(l, b) = (1, 0)$ out to a distance of 20 kpc, and transverse velocities are uniformly sampled between 10 to 150 km/s. Dashed line shows median value, dotted lines show interquartile range (IQR).

4.4 Exploring the Parameter Space of ISM Scintillation with NE2001

The likelihood of detecting scintillation depends heavily on our physical location in our Galaxy and the lines of sight at which we observe. To determine the best targets for detecting scintillation, we need to estimate the quantitative effects of scintillation on narrowband signals in various directions on the sky. This depends on the plasma free electron number density and strength of turbulence along the line of sight.

Cordes and Lazio 2002 developed the NE2001 free electron density model for our Galaxy, based on pulsar observations and scattering studies. NE2001 models various Galactic features and estimates the dispersion measure (DM) and characteristic scattering scales to distance d along any given line of sight through the Galaxy. The scattering scales computed include the scintillation timescale, spectral broadening, scintillation bandwidth, and temporal broadening. This allows us to uniquely estimate the asymptotic statistical properties of

scintillation, which can help decide promising targets for scintillation analysis.

Given a distance d and Galactic coordinates (l, b), the publicly-available code for NE2001 model estimates the expected scintillation timescale and bandwidth at frequency $\nu = 1$ GHz and transverse velocity $V_T = 100$ km/s. From this point, we have the scaling relation:

$$\Delta t_d \propto \nu^{2/\alpha} V_T^{-1}, \tag{4.10}$$

where $\alpha = 5/3$ for Kolmogorov turbulence and $\alpha = 2$ for square-law turbulence (Cordes et al. 1997; Coles et al. 2010). With Equation 4.10, we can scale raw NE2001 values to estimate scintillation properties for specific observational setups.

We would like to narrow the parameter space of possible observing configurations and scintillation timescales to those that are most amenable to detection with current facilities. With the NE2001 model, we can estimate scintillation properties for a given set of input parameters, including the sky direction, distance, frequency, and transverse velocity. However, these inputs constitute an enormous parameter space, with no clear *a priori* preference from a SETI perspective. Even with bounds for each individual parameter, it would be prohibitively computationally expensive to calculate properties across each combination of potential parameters. Instead, we choose to use Monte Carlo sampling over the parameter space, using enough samples to sufficiently capture the core statistics of the distribution of scintillation properties.

For sampling, we fix a sky direction (l, b) and a target radio frequency band. We then sample the frequency ν uniformly within that band (as a narrowband signal could be found anywhere in the band). In this paper, we will refer to common radio bands used with the GBT, including L (1.15–1.73 GHz), S (1.73–2.6 GHz), C (3.95–8.0 GHz), and X (8.0–11.6 GHz) (GBT Support Staff 2017; MacMahon et al. 2018).

For the distance d, we have to specify a maximum distance $d_{\max}$, but the minimum distance $d_{\rm tr}$ is that at which weak scattering transitions to strong scattering. We can sample uniformly from $[d_{\rm tr}, d_{\max}]$, but we can also attempt to match the potential distribution of distances that ETI would actually occur. For example, we can sample distances based on the expected distribution of stellar number densities along the line of sight through the Galaxy. For this, we use model parameters from Gowanlock et al. 2011, who adapted a model from Carroll and Ostlie 2007 that matches the observed density in the solar neighborhood. To see the effects on our sampling, we can also sample by stellar mass density, though this is less precise, since we typically expect ETI to reside around less massive stars. We use the model provided in McMillan 2016 to compute stellar mass density along a line of sight. In Figure 4.4, we compare these models as a function of distance along Galactic coordinates $(l, b) = (1, 0)$, showing them alongside NE2001-generated scintillation timescales. As expected, the mass density profile is significantly sharper than the number density, but both more heavily weight the Galactic center region compared to uniform distance sampling.

Finally, the transverse velocity V_T is perhaps the hardest to constrain in general. For scintillation, V_T depends on the relative transverse velocities of the source, observer, and scattering screen, each of which is difficult to predict. A representative transverse velocity

for Galactic pulsars is about 100 km/s (Cordes 1986). The transverse velocity for an ETI source, especially in our solar neighborhood, might be on order 10 km/s instead (Cordes and Lazio 1991; Cordes and Rickett 1998). Depending on the line of sight, for sources far across the Galaxy (i.e. 10 kpc or so), differential Galactic rotation can add components to the transverse velocity on order of 100 km/s as well. An emitter's orbital velocity and spin velocity can also contribute. Since all of these independent effects are non-trivial and stochastic, we can at best set heuristic transverse velocity ranges and sample uniformly between them, understanding that even the limits themselves are only useful to an order of magnitude.

Taking all these parameters together, we can create sampled distributions for each scintillation scale. Figure 4.5 shows a realization of Monte Carlo simulations for C-band in the $(1, 0)$ direction with $N = 10000$ realizations, using a number density-based weighting on distance samples. We use a maximum distance of 20 kpc and a transverse velocity range of 10 to 150 km/s. It is readily apparently that the resultant distributions are significantly skewed. For example, short distances from the observer will lead to long scintillation timescales. Since the goal of the parameter space analysis is to evaluate the observational setup that gives us the best likelihood for detecting scintillation in narrowband signals, we focus on the central statistics. For skewed distributions, we choose to calculate the median and interquartile ranges (IQR) as representative values for each scale.

From Figure 4.5, we conclude that signals at C-band in the direction $(1, 0)$ are likely to have scintillation timescales ranging between 10–28 s. Indeed, since this is the IQR, only half of the sampled timescales lie in that range, and there is an implicit bias towards the lower end of that range and below. What this really tells us is that if we are searching in that sky direction and at that frequency, we should make sure to choose observational parameters so that we are sensitive to scintillation timescales between 10–28 s. Also, note that spectral broadening is on order 0.01 Hz, which is negligible compared to typical spectral resolutions used in modern radio SETI.

With this tool, we can estimate which range of scintillation timescales to target for a given sky direction and frequency band.

4.5 Temporal analysis of detected narrowband RFI

To evaluate whether it is viable to detect scattering effects like scintillation in detected narrowband signals, we must characterize the standard RFI environment within which SETI observations are taken. The majority of narrowband RFI is generated from communication applications, therefore it is common for RFI to show intensity modulation in frequency or time. Depending on the nature of this modulation and the free electron column density along a line of sight, RFI could confound the detection of actual scintillated extra-solar signals. We must therefore analyze the RFI environment, regardless of sky direction, with respect to temporal statistics that can be used to identify the presence of ISM scintillation. In this

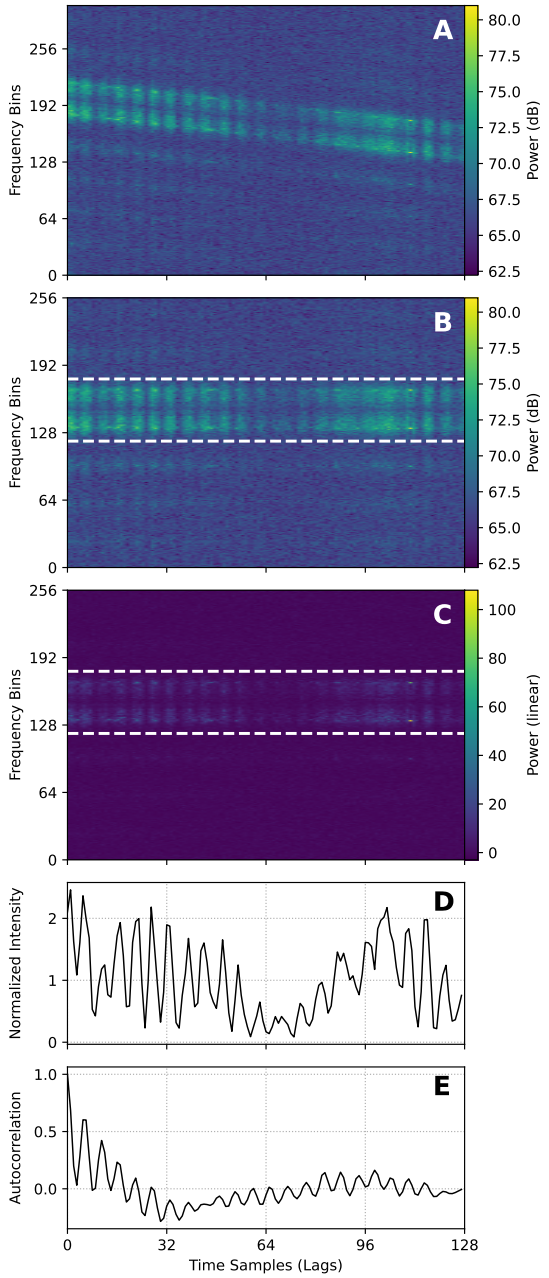


Figure 4.6: Steps used in signal intensity analysis. **A**: Detected narrowband signal, in GBT data. **B**: De-drifted signal from panel A, with computed bounding frequencies in dashed white lines. **C**: Frame from panel B, normalized using the background noise along the frequency axis. **D**: Time series intensities computed by integrating power in panel C between the bounding frequencies and normalized to a mean intensity of 1. **E**: Sample ACF computed from panel D.

paper, we focus on RFI present in GBT observations, which comprise a significant fraction of BL data.

We must note that it is technically possible that any given detected signal in this "RFI" analysis is actually a technosignature. However, we can confidently say that the overwhelming majority of signals encountered will be anthropogenic in origin. Furthermore, in this analysis, we take observations in a direction where Δt_d is long compared to τ_{obs}. This way, detected signals will not be modulated by ISM scintillation within a single observation, so whether or not a given signal is a technosignature is irrelevant to our analysis.

4.5.1 Finding and Characterizing Signals

In this section, we outline the general process for detecting signals and extracting intensity time series data, from which we can compute diagnostic statistics and run our scintillation analysis. Figure 4.6 demonstrates the step-by-step process on a real GBT RFI signal.

The first step in analyzing the RFI environment is curating a dataset of detected signals. We need some form of energy detection to pinpoint the frequencies and preferably the drift rates of narrowband signals. The most common method for detection used by BL is the tree deDoppler code `turboSETI`[3], which efficiently implements a matched filter for linearly drifting narrowband signals (Enriquez et al. 2017; Enriquez and Price 2019). `turboSETI` gives us the signal frequency at the beginning of the observation and the best-fit drift rate. However, to extract intensity data for scintillation analysis, we additionally need the frequency bandwidth that the signal occupies.

Ultimately, we aim to construct a "bounding box" of sorts around each narrowband signal. Since narrowband signals can have an overarching Doppler drift rate, these bounding boxes are defined by a starting central frequency, a drift rate, and a signal bandwidth. In time-frequency space, these become bounding parallelograms, since we take the signal bandwidth to follow the extracted drift rate at each time step. Given a fit for the drift rate, we can de-drift a spectrogram containing the signal by shifting each individual spectrum accordingly, reducing the problem to finding the frequency bandwidth that overwhelmingly captures the signal's power.

There is no singular correct way to bound radio signals found in spectrogram data. There are many morphologies of narrowband signals, such as those with unstable oscillator frequencies or varying intrinsic bandwidths. Signal leakage also affects bright signals and spreads the power into neighboring spectral bins. Background noise and nearby spurious signals can additionally complicate the bandwidth calculation.

Signal bound estimation has been done before in radio astronomy. For pulsars, Straten et al. 2012 measures the size of individual pulses as the width at a user-specified fraction of the peak intensity. In one of the rare instances of bandwidth estimation in narrowband SETI, Pinchuk et al. 2019 calculates signal bounds at the 5σ–level, regardless of the detected signal's peak S/N.

[3] https://github.com/UCBerkeleySETI/turbo_seti

Our goal is to find the tightest frequency bounds that do not exclude a significant amount of signal power, so that we can accurately represent the intensity behavior over time. If our bounds are too tight, we risk excluding and distorting information; if they are too loose, noise fluctuations can take over and wash out the signal.

In this work, we choose to bound signals at 1% of their maximal intensity. First, we de-drift and integrate a spectrogram along the time axis to get a spectrum centered on the signal. To make a fit of the noise background, we first exclude most of the bright data points with sigma clipping up to 3σ. Then, we fit a straight line to the remaining points and obtain the final corrected spectrum by subtracting this fit from the original spectrum. The signal bounds are calculated as the frequency bins on the left and right of the signal center whose intensities dip below 1% of the maximum intensity in the corrected spectrum. This method is balanced, capturing most of the power from signals that have apparent bandwidths ranging from a few Hz to a kHz. Figure 4.6B shows an example of such a fit.

To analyze the properties of a signal's intensity over time, we need to isolate the signal as best as possible from the noise background. To estimate the noise background, we use sigma clipping along the frequency axis to calculate the mean and standard deviation of noise at each timestep. We then normalize the de-drifted spectrogram at every sub-spectrum by subtracting the according noise mean and dividing by the according noise standard deviation. Theoretically, this standardizes the instrument response over the course of the observation and centers the background intensity to 0. It also serves as a crude way of filtering out simple broadband interference. Figure 4.6C shows the resulting spectrogram.

To get the intensity time series for a signal, we integrate the normalized spectrogram along the frequency axis between the computed frequency bounds, resulting in a 1D array of length N_t. To standardize the analysis, we additionally normalize this time series to have a mean of 1, as shown in Figure 4.6D. From the normalized time series, we compute the ACF (Figure 4.6E). With these two together, we can calculate all the diagnostic statistics to compare with theoretical scintillation properties.

It is important to note that since we attempt to normalize the noise background of the spectrogram to a mean of 0 via subtraction, we may end up with negative values in our final extracted time series. Since we cannot remove the noise fluctuation entirely, the time series intensities will always be affected by noise in this way. Normalizing the time series to a mean of 1 can have the additional effect of making the negative "intensities" even more negative. Nevertheless, we choose to compute diagnostic statistics using the normalized time series.

4.5.2 Observation Details

In this exploration of RFI properties, we are investigating the distribution of diagnostic statistics in real, detected RFI signals to evaluate whether these statistics can be used to identify the presence of scintillation. We must therefore ensure that our observations are unlikely to contain any scintillated signals.

For this reason, and for additional convenience, we choose to observe towards the north celestial pole (NCP). We verified with NE2001 that the expected scintillation timescales are

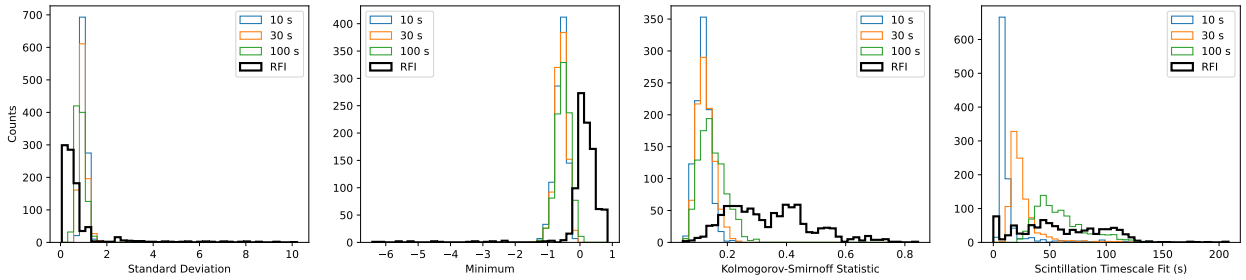


Figure 4.7: Histograms of diagnostic statistics for detected L-band signals with S/N≥25. For each statistic, the distribution from detected RFI is shown in black. Plotted for comparison are distributions from synthetic scintillated signals at S/N=25 with scintillation timescales of 10 s (blue), 30 s (orange), and 100 s (green). Across all diagnostic statistics, it would be difficult to distinguish a true scintillated signal from RFI given the L-band RFI distributions.

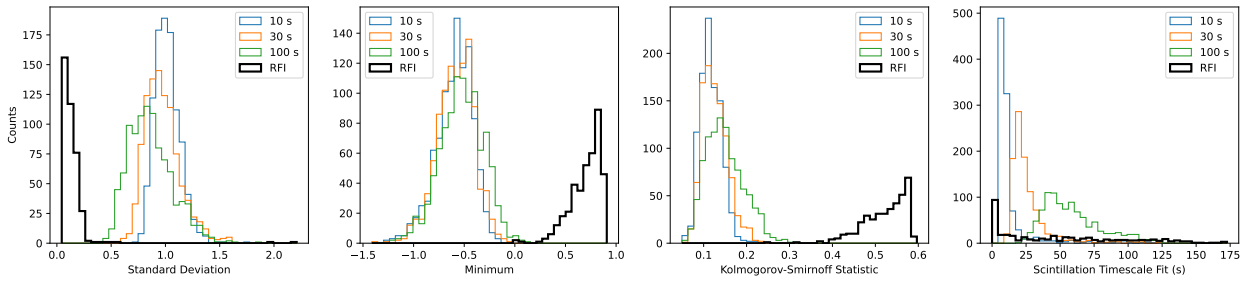


Figure 4.8: Histograms of diagnostic statistics for detected C-band signals with S/N≥25. For each statistic, the distribution from detected RFI is shown in black. Plotted for comparison are distributions from synthetic scintillated signals at S/N=25 with scintillation timescales of 10 s (blue), 30 s (orange), and 100 s (green). It could be possible to distinguish a true scintillated signal from RFI given the C-band RFI distributions.

long compared to desired observation parameters. For instance, at 1 GHz (L-band), a signal at 1 kpc with $V_T = 100$ km/s would show a scintillation timescale of 702 s. The other bands we use at the GBT (S, C, and X) correspond to even longer expected timescales due to the frequency scaling.

The process for identifying scintillation can be performed over many observational timescales. In our case, we focus our analysis on data resolutions close to those used typically by BL. BL normally runs analysis on 5 minute integrations at a frequency resolution of 2.79 Hz and a time resolution of 18.2 s, for 16 pixels or time samples per observation. We use the same frequency resolution, but extend the data by taking 10 minute integrations at a resolution of 4.65 s, so that we get 128 samples per observation. Having more time samples leads to better diagnostic statistics and better time resolution but requires significantly more data storage.

For this work, we used the GBT to take 10 minute observations of the NCP each at L and C-band on 2022 May 16. To find narrowband signals, we use `turboSETI` with a detection threshold of S/N=10 to search up to maximum drift rates of ±5 Hz/s. As an additional step, we exclude detections of the so-called "DC bin" in each coarse node, a vertical artifact of the FFT performed during fine channelization.

4.5.3 Empirical Results

Using the procedure described in Section 4.5.1, we compute diagnostic statistics for detected signals in GBT observations taken at L and C-bands. For convenience, in this discussion, we will refer to detected GBT signals as "RFI". While these observations are very unlikely to contain scintillated signals, we cannot necessarily rule out the presence of technosignatures in our data. Nevertheless, we can comfortably say that the vast majority of signals are human-created interference.

To best compare with our expectations for scintillated narrowband signals, we create synthetic GBT observations with scintillated signals produced using the methods in Section 4.3.3 and run them through the same analysis pipeline. For the synthetic signals, we construct separate datasets using $\Delta t_d = 10$, 30, and 100 s, as in Figure 4.3.

The synthesis process described in Section 4.3.3 does not take noise into consideration. In this work, we treat narrowband signals as additional power that is present on top of the noise background. As such, we assume that the effects of ISM scintillation are imprinted on the signal independently from the noise background. To construct a synthetic observation, we compute a realization of a scintillated signal's intensity over time using ARTA and inject a signal with those intensities onto a radio spectrogram with a realistic noise background, following Equation 4.3. We use the Python package `setigen`[4] to inject artificial signals and compare directly with real GBT observations (Brzycki et al. 2022). For each scintillation timescale, we generate $N = 1000$ signals with zero drift rate and the same S/N that matches

[4] https://github.com/bbrzycki/setigen

our `turboSETI` detection threshold. We calculate diagnostic statistics for the artificial signals in the same way that we do for detected RFI.

The histogram comparisons for each diagnostic statistic at L and C-bands are shown in Figures 4.7 and 4.8. The bold, black histograms show the non-DC RFI samples in the respective frequency band, whereas the thinner histograms represent the synthetic signal datasets. The less the RFI distributions intersect with the scintillated signal distributions, the better our methodology can distinguish a true scintillated signal.

At a glance, C-band RFI has better separation than L-band RFI from the scintillated signal distributions, across all diagnostic statistics. In particular, for C-band, the statistics pertinent to the exponential distribution of scintillated intensities (standard deviation, minimum, K-S statistic) have relatively well-defined separations. These can be used to set thresholds (or target ranges) for each statistic, which can be combined to help filter detected signals for scintillation candidates. While the fitted scintillation timescale distributions intersect appreciably, in practice, thresholds can still be set using synthetic signal distributions and used as filters.

Comparatively, a significant portion of the L-band RFI occupies the same ranges of statistics as the synthetic signals. This means that existing RFI would confound the detection of real scintillated signals with these methods. From our observations, we observe that lower frequencies (such as L and S bands) have a relatively higher density of RFI with many morphologies, and this could be causing the distributions of statistics looking broader and more irregular than those for C-band RFI.

4.6 Discussion

4.6.1 Observational Recommendations for Scintillated Technosignature Searches

The empirical RFI distributions suggest that at the GBT, higher frequencies will be better for creating statistics-based thresholds.[5] The RFI environment at C and X-bands is less dense and less diverse than that at L and S-bands. However, scintillation effects decrease inversely with increasing frequency, lengthening the scintillation timescales (Equation 4.10). There is also a trade-off in choosing which frequencies to search: higher frequencies have more favorable RFI properties but require either longer observations or pointings with more scattering.

For each observing band, the RFI environment sets unavoidable statistics thresholds. At L-band, for instance, it is possible that there is no sky direction and no target scintillation timescale amenable for a scintillated technosignature search. While the properties of the local RFI environment certainly vary as a function of time and location, our observations suggest that lower frequencies may always be difficult to use. Specifically, the empirical

[5]For other telescope sites, a similar RFI analysis would need to be conducted in order to draw similar insights about RFI vs. frequency.

L-band RFI distributions covered the ideal asymptotic value for each diagnostic statistic, implying that no variation of observational parameters could unambiguously distinguish an appreciable fraction of RFI from real scintillated signals.

On the other hand, for C-band and above, we must tend towards longer observing lengths or point towards regions of higher scattering, such as the Galactic center, in order to capture enough scintles. As discussed by Gajjar et al. 2021, there are a multitude of reasons that an ETI detection might be most likely towards the Galactic center, making this an attractive option for a scintillated technosignature search.

As the field of radio SETI grows and as new technosignature candidates are found, more work is being done in signal verification and follow-up analysis (Sheikh et al. 2021; Tao et al. 2022). To this end, beyond dedicated searches for scintillation, the methods introduced in this paper may also be used as supplementary analysis for other radio SETI searches. For example, given an interesting narrowband detection that passes some SETI filters, one might ask additionally whether the signal is ISM-scintillated. Following the steps in this work and using `blscint`, one could estimate likely scintillation timescales along the observation's line of sight at the detected signal frequency. Then, one could generate synthetic ARTA datasets to set diagnostic statistic thresholds and compare how the statistics for the detected signal measure up. Assuming the signal was still compelling after these steps, it would be prudent to do a similar detected RFI analysis using the same telescope, frequency band, observation length, and time resolution to check for RFI with confounding modulation. While emission from distant sources along the Galactic plane has the best chance of exhibiting detectable scintillation within individual observations, these methods constitute a concrete framework for evaluating the likelihood of scintillation in signals from any observational radio SETI campaign.

4.6.2 The Impact of Models on Designing Observational Campaigns

The effectiveness of a designated search for scintillated technosignatures will depend on how well we can estimate the most likely values for Δt_d as a function of sky direction and frequency.

The fewer unknown degrees of freedom in our Monte Carlo sampling procedure (Section 4.4), the better the timescale estimates will be. For example, if we wanted to estimate what timescales are possible for emission near a particular known star, we would already begin with the location (l, b) and distance d. The only major parameters left would be the target frequency range (which we can control) and the effective transverse velocity. By constraining sampling parameters, one can get tighter bounds for scintillation timescales and tune observation parameters accordingly.

Our Monte Carlo procedure for scattering strength estimates relies on the NE2001 electron density model. While NE2001 remains a popular choice, the YMW16 model from Yao et al. 2017 has emerged as another prominent Galactic electron density model. There have

been studies comparing both, such as Deller et al. 2019 and Price et al. 2021, particularly with regards to DM and distance estimation applied to new pulsar datasets. While YMW16 benefits from more recent data, when compared to independent pulsar measurements, both models have their own systematic estimation biases that depend on the location in the Galaxy (Price et al. 2021).

The key difference for this work is that NE2001 uses scattering measurements in its fit and estimates scattering properties throughout the Galaxy (Cordes and Lazio 2002). YMW16 specifically avoids using scattering measurements, arguing that the majority of scattering arises from relatively thin features along the line of sight and therefore cannot be used to appropriately describe the large-scale distribution of scattering (Yao et al. 2017). However, the YMW16 model still attempts to estimate pulse broadening timescales by using an empirical τ–DM relation simplistically, resulting in unreliable scattering values, especially for fast radio bursts (Ocker et al. 2021).

While it may be difficult to develop a model that robustly constrains the effects of scattering along any line-of-sight in the Galaxy, doing so to even an order-of-magnitude would be crucial for designing scintillation search strategies for SETI, as well as for evaluating whether existing narrowband detections could benefit from scintillation analysis. As new pulsars are discovered and new Galactic electron density models are produced, we suggest that attention should still be given to scattering measurements and predictions.

4.6.3 Building on the Analysis Pipeline

While it involves many steps, the method for search and intensity extraction described in this paper is relatively straightforward. We rely on standard deDoppler search methods (e.g. `turboSETI`) to both find and characterize signal paths in one shot. Since we are searching for a stochastic effect, keeping the processing simple is not necessarily a detriment. However, our pipeline will still flag bright broadband signals that are able to exceed our S/N threshold. The philosophical question on whether a broadband impulse that contains sharp spectral features could be considered narrowband notwithstanding, using additional pre-processing to detect broadband signal features could better standardize the types of signals passing through the intensity extraction pipeline.

Machine learning (ML) could be used to aid scintillated searches, such as for creating initial classifications of signal type and eventually even for doing final candidate analysis. In particular, deep learning techniques, such as convolutional neural networks (CNNs) have been used effectively on a variety of tasks using radio spectrograms (Zhang et al. 2018a; Harp et al. 2019; Brzycki et al. 2020; Pinchuk and Margot 2022; Ma et al. 2023). CNNs could be used to filter out spectrograms with clear broadband emission and would be relatively straightforward to integrate into the pipeline. There is certainly an avenue for complementing domain-based statistical features with computer vision methods, as is done in time-domain SETI (Giles and Walkowicz 2019).

ML techniques could also be applied to the extracted time series or even to the raw signal spectrogram to directly classify likely scintillation candidates. From the standpoint of

interpretability, having a set of diagnostic statistics with direct links to the expected theoretical behavior of scintillated narrowband signals provides us with intuitive filter thresholds, whereas a direct ML approach might not. However, used in tandem with our methods for producing synthetic scintillated signals, supervised ML algorithms such as random forest classifiers could be used to rank each of our diagnostic statistics in their importance towards correctly distinguishing scintillated signals from RFI (Breiman 2001). This could be a valuable future direction for scintillation-based searches and may very well be a function of each observatory's unique RFI environment.

4.6.4 Implications and Future Directions

In this work, we only focus on searching for strong scintillation on high duty-cycle narrowband signals. Since the ionosphere and IPM will tend to vary intensity relatively slightly in most cases, we identified strong scintillation from the ISM as detectable from 100% intensity modulations. Analysis of the RFI environment at the GBT suggests that weakly scintillated extra-solar signals would be difficult to distinguish from existing interference, while strongly scintillated signals can be separated along multiple diagnostic statistics.

A common procedure during signal verification of an interesting candidate is to search for other signals close in frequency that are similar in morphology (Sheikh et al. 2021). Along these lines, the possibility of simultaneous ETI signals at multiple frequencies is interesting from the perspective of a scintillation analysis. For signals separated by less than the scintillation bandwidth, we should see the same intensity modulation over time. However, for signals separated by more than the scintillation bandwidth, we would receive different intensity time series that still have the same overall scintillation timescale. With our tool to estimate scintillation timescales and bandwidths, if we were to detect multiple spectrally-nearby scintillation candidates within the same observation, we would have yet another way to contextualize the detected signals and determine whether they might actually be technosignatures.

We limit our search methodology to high duty-cycle signals, so that any fluctuations in intensity is purely due to scintillation. If an ETI transmitter is attempting to send information, the initial signal will already be modulated. This could also confound the presence of scintillation. However, we argue that along the lines of sight and distances for which we would expect narrowband signals to be scintillated, the identification of scintillation is itself a message. An ETI civilization advanced enough to transmit a message through interstellar space should understand the effects of plasma on radio emission, since it would distort the initial transmission and hinder communication. With this in mind, an ETI beacon might instead transmit a pure, unmodulated signal, expecting that other civilizations could detect the presence of scintillation in an artificial, narrowband signal. Instead of explicitly encoding a message in the narrowband signal, the mere presence of scintillation would communicate the message: "we are here."

Radio scattering from ionized plasma presents in other ways, such as broadband modulation and dispersion. While broadband SETI searches are relatively less common, as we

explore new regions of the potential SETI signal parameter space, scintillation could be searched along the frequency axis analogously to our search along the time axis. The scintillation bandwidth, the spectral analogue of the scintillation timescale, does not vary as a function of transverse velocity, so parameter estimation may be less uncertain (Cordes and Lazio 1991). Broadband signal searches are also able to use coarser frequency resolutions than narrowband searches, though they would likely have to use much finer time resolutions.

We hope that this work will lead to more discussion and theoretical work on other ways in which the actual radio emission that we receive can be used to identify the extra-solar origin of technosignatures. Beyond scattering, there are still properties of radio emission, such as polarization, that are only beginning to be considered in depth from a SETI perspective (Tao et al. 2022). Whether it is because certain effects are stochastic or because human radio emission exploits every facet of radio light possible for communication, extracting non-trivial information from a radio signal's detailed morphology has been and will remain difficult. We may need to push the limits of detectability along hitherto unexplored axes to discover the first technosignature.

4.7 Acknowledgements

Breakthrough Listen is managed by the Breakthrough Initiatives, sponsored by the Breakthrough Prize Foundation. The Green Bank Observatory is a facility of the National Science Foundation, operated under cooperative agreement by Associated Universities, Inc. We thank the staff at the Green Bank Observatory for their operational support. S.Z.S. acknowledges that this material is based upon work supported by the National Science Foundation MPS-Ascend Postdoctoral Research Fellowship under Grant No. 2138147.

Chapter 5

The Breakthrough Listen Search for Intelligent Life: Galactic Center Search for Scintillated Technosignatures

The search for extraterrestrial intelligence (SETI) in radio frequencies has focused on spatial filtering as a primary discriminant from terrestrial interference (RFI). Individual search campaigns further choose targets or frequencies based on criteria that theoretically maximize the likelihood of detection, serving as high level filters for interesting targets. Most filters for technosignatures do not rely on intrinsic signal properties, as the RFI environment is difficult to characterize. In Brzycki et al. 2023, we propose that the effects of interstellar medium (ISM) scintillation on narrowband technosignatures may be detectable under certain conditions. In this work, we perform a dedicated survey for scintillated technosignatures towards the Galactic center and Galactic plane at C-band (3.95–8.0 GHz) using the Robert C. Byrd Green Bank Telescope (GBT) as part of the Breakthrough Listen (BL) program. We conduct a Doppler drift search and directional filter to identify potential candidates and analyze results for evidence of scintillation. We characterize the C-band RFI environment at the GBT across multiple observing sessions. We do not find evidence of putative narrowband transmitters with drift rates between ± 10 Hz/s towards the Galactic center, scintillated or otherwise, above an EIRP of 1.9×10^{17} W up to 8.5 kpc.

5.1 Introduction

The Search for Extraterrestrial Intelligence (SETI) is the organized effort to detect technosignatures, signatures that would unequivocally indicate the existence of alien technology. Over the past few decades, SETI has expanded massively in scope, due to technological innovation and an influx of resources. Radio SETI in particular now regularly uses radio

telescopes across the Earth, including large antenna arrays, large swaths of instantaneous bandwidth, and high time and frequency resolutions supported by advancements in data storage and pipelining (Tarter 2001; Siemion et al. 2013; Hickish et al. 2016; Price et al. 2020; Margot et al. 2021; Gajjar et al. 2021). Radio SETI is complemented by optical SETI, which has sought to detect lasers and Cherenkov radiation from cosmic rays (Stone et al. 2005; Lipman et al. 2019; Acharyya et al. 2023). SETI as a field is constantly evolving and becoming increasingly effective in quantifying the search for technosignatures (Tarter 2001; Wright et al. 2018).

The Breakthrough Listen (BL) Initiative is the largest concentrated effort in modern SETI to search for technosignatures (Worden et al. 2017; Isaacson et al. 2017). Beginning in 2017, BL has been an instrumental part in the development and proliferation of modern SETI efforts, from optical to radio. BL has commissioned time on the Robert C. Byrd Green Bank Telescope (GBT) in West Virginia and the CSIRO Parkes telescope in Australia for radio searches (MacMahon et al. 2018; Price et al. 2018) and time on the Automated Planet Finder (APF) in California for optical searches (Radovan et al. 2014; Lipman et al. 2019).

Radio SETI has historically focused on the detection of narrowband high duty cycle signals, signals that are generally always "on" (also called continuous wave signals). This largely stems from the assumptions that narrowband signals are not produced from natural sources, and the practicality that high duty signals are always present and therefore can be isolated in the sky and re-detected. Recent searches have expanded to target additional morphologies, such as broadband signals and pulsed signals (Gajjar et al. 2021, 2022; Suresh et al. 2023).

There are a few fundamental difficulties underlying narrowband technosignature searches. For instance, we do not know definitively the kinds of technosignature that may exist; even if they are narrowband radio signals, we do not know the central frequencies, frequency and time modulation patterns (i.e. for communication or information sharing), or even emission arrival times. To address this, modern radio searches typically use wide instantaneous bandwidths, cover many targets with multiple observations, and use big data analysis techniques to detect signals and sort them by interest.

Another major issue is the presence of radio frequency interference (RFI), anthropogenic emissions that are regularly picked up in the sidelobes of radio telescopes. Even the GBT, which is located in a federally-mandated "Radio Quiet Zone," picks up a large share of RFI, which have a large diversity in morphology in their own right (Price et al. 2020; Maan et al. 2021). RFI can originate from cell phones, television, GPS services (satellite), and virtually any digital device. As such, RFI can be both narrowband and broadband, confounding technosignature searches of all kinds. Most of the algorithmic analysis in modern SETI is concerned with differentiating detected signals between plausible technosignatures and terrestrial RFI. The most common methods are the identification of non-zero Doppler accelerations, since that would imply a non-terrestrial frame of reference, and localization on the sky using location-switching observations.

However, we may also be able to use the information present in the intrinsic power detected in radio signals as a discriminant. Cordes and Lazio 1991 described how ionized

plasma in the interstellar medium (ISM) could scatter narrowband radio technosignatures in the same way as it scatters pulsar emission, an effect that has been readily observed and analyzed in pulsar observations to probe the properties of intervening ISM plasma. Multi-path propagation through the turbulent ISM in the so-called strong scattering regime can result in 100% intensity modulations in narrowband signals. At times, constructive interference brings the overall signal intensity many times higher than the scattering-free intensity, which is beneficial for SETI in that this would bring an otherwise undetectable technosignature above a search's signal-to-noise (S/N) threshold.

Brzycki et al. 2023 suggested that ISM scintillations may imprint on the signature of high duty cycle narrowband signals within the duration of individual observations, resulting in detectable intensity modulations that follow predicted theoretical distributions and could be differentiated from RFI. A radio signal whose intensity fluctuations are consistent with ISM scattering would be a very strong candidate for a bona fide technosignature, since the physical nature of the fluctuations would necessarily imply an extra-solar origin. To resolve such fluctuations on the timescale of typical radio observations, we must observe through an appreciable column of free electrons. The best direction for this is therefore towards the Galactic center (GC), for which the column of free electrons in ionized plasma is the greatest.

Beyond scattering effects, the GC is particularly intriguing to target for SETI. Just as the plasma density is the highest, the stellar number density and therefore propensity for life to originate increases towards the GC. From a game theoretic point of view, a plausible common direction of interest for all ETI in the Galaxy should be towards its center (Gajjar et al. 2021). ETI civilizations willing and capable of sending strong transmitted radio signals as beacons might set up such transmitters at the GC or send targeted signals in the direction of the GC. ETI capable of receiving radio signals might point their radio antennae towards the GC to detect either targeted transmissions or even leakage radiation from normal technological activity.

For observable intensity scintillations, narrowband radio signals must travel through enough plasma in order to hit the strong scattering regime. However, to actually detect these scintillations from background noise, we simultaneously need the detected signals to have a high integrated S/N. Of course, we can only theorize about the energy budget and technological capabilities of ETI for transmitting sufficiently bright beacons. A common framework for discussing such capabilities is the Kardashev scale, which classifies theoretical civilizations based on their available energy budget (Kardashev 1964). Kardashev Type I civilizations are able to utilize the energy available on their planet (through solar radiation or other means), while Type II civilizations can directly use the full energy provided by their host star (Gray 2020). Representative powers for these classifications are about 10^{16} W for Type I and 10^{26} W for Type II. Using BL hardware and observational parameters, Gajjar et al. 2021 set limits for the equivalent isotropic radiated power (EIRP) of technosignatures up to a distance of 8.5 kpc from Earth towards the GC as above 5×10^{17} W, just an order of magnitude above the definition of a Type I civilization.

We expect strong scintillation to manifest in sources with distances on the kpc-scale, so detectable sources anywhere from us towards the GC will therefore be consistent with

the energy budget of a Type I civilization. In fact, since these limits assume an isotropic emitter, the requisite energy budget is even more favorable for a targeted transmitter. For instance, when it was operational, the Arecibo Planetary Radar's S-band transmitter had a directional gain of about 10^7 (Siemion et al. 2013). An antenna with this gain at the GC would only need to be powered by a 5×10^{10} W transmitter to match the EIRP limit set in Gajjar et al. 2021. While that is still a very large amount for our civilization to produce and transmit continuously, it is plausible that a more technologically-advanced ETI could.

In the present work, we conduct a radio narrowband search at C-band of the GC and nearby directions through the Galactic plane (GP) in order to search for ISM scintillated narrowband signals. We apply an ON-OFF directional filter as performed in prior SETI searches. In addition, we perform a scintillation analysis on detected signals towards targets in the survey. We also perform this analysis on control pointings towards the North Galactic Pole, in order to comment on RFI properties at the Green Bank Telescope (GBT) and to determine how similar RFI appears to scintillation at C-band. In Section 5.2, we detail our methodology for estimating likely scintillation timescales and present the observing strategy. In Section 5.3, we describe the directional and scintillation analysis performed across our observations. In Section 5.4, we present signal distributions and show examples of signals that passed the initial directional filter. In Section 5.5, we discuss the results and implications of our analysis, both for the survey targets and for the NGP observations. Finally, in Section 5.6, we contextualize this survey with regard to prior SETI efforts and suggest future directions.

5.2 Observations

5.2.1 Scintillation Estimation with NE2001

Intuitively, we can maximize the likelihood of detecting scintillated technosignatures by observing along the Galactic plane, especially towards the Galactic center. However, scintillation timescales can vary by orders of magnitude depending on distance, frequency, and direction on the sky. It is therefore useful to estimate expected timescales so that we can set appropriate observation parameters, such as observation length, time resolution, and frequency resolution.

We derive scintillation scale estimates based on the NE2001 model (Cordes and Lazio 2002), following the procedure detailed in Brzycki et al. 2023. Namely, for each sky direction (l, b) and distance d, we can use the NE2001 model to first estimate the scintillation timescale Δt_d at 1 GHz and 100 km/s transverse velocity. We can then scale this estimate for any frequency and transverse velocity (V_T), since $\Delta t_d \propto \nu^{1.2} V_T^{-1}$ for strong scattering. We can create a distribution of plausible timescales by Monte Carlo sampling, in which we random sample the distances, frequencies, and transverse velocities from prior distributions.

Signals of interest could occur at any frequency in the band, so for C-band, we simply use a uniform prior distribution from 3.95–8.0 GHz. Estimating possible diffraction pattern

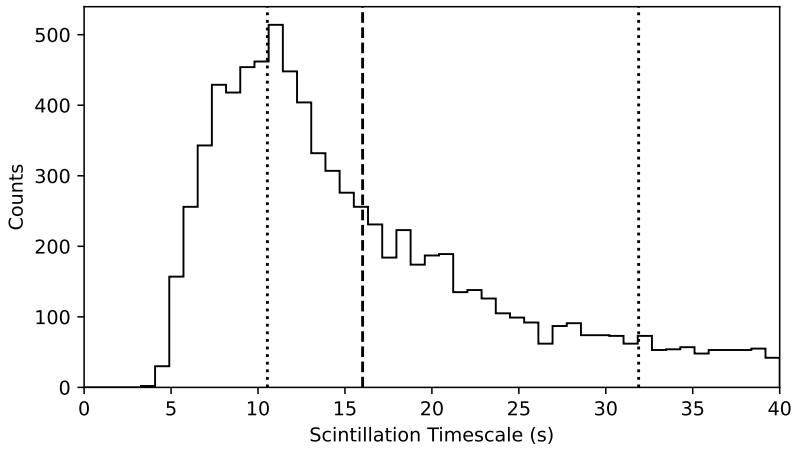


Figure 5.1: Monte Carlo-sampled scintillation timescales for $(l, b) = (1°, 0°)$ at C-band with $N = 10^4$. The dashed line indicates the median timescale, and the dotted lines indicate the first and third quartiles.

transverse velocities is very complex and dependent on many independent variables, so for simplicity, we also choose a uniform prior from 10–150 km/s, following Brzycki et al. 2023. Along a line of sight, however, locations closer to the Galactic center will generally have a higher concentration of stars, so they should be weighted higher. We can create a normalized probability density function for distance along the line of sight by using a model for the stellar number density of the Galaxy.

There are a few ways to model the Galactic distribution of stars which may contain radio transmitters. For example, we can convert from models of the Galaxy's stellar mass density with respect to radius and distance from the midplane by making a simplifying assumption of 1 $M_\odot$ per star (Uno et al. 2023; Brzycki et al. 2023). In this work, we follow (Gajjar et al. 2021) and instead use a more direct model of the Galaxy's stellar number density suggested by Carroll and Ostlie 2007 of the form:

$$n_*(z, R) = n_0(e^{-z/z_{\text{thin}}} + 0.085e^{-z/z_{\text{thick}}})e^{-R/h_R}, \tag{5.1}$$

where z is the height from the midplane and R is the radius from the Galactic center. Gowanlock et al. 2011 analyzed a set of Galactic number density models and found that $n_*(z, R)$ gave the closest simultaneous fit to both the total stellar disk mass and the observed stellar density in the local neighborhood, where the normalization factor $n_0 = 5.502$ stars pc^{-3}, the thin-disk scale height $z_{\text{thin}} = 350$ pc, the thick-disk scale height $z_{\text{thick}} = 1000$ pc, and the radial scale length $h_R = 2.25$ kpc.

For a line of sight along Galactic coordinates (l, b), we compute weights proportional to the stellar number density starting from a distance d_{tr}, the minimum distance for which the strong scattering regime applies to all frequencies in the band, to a specified maximum

distance $d_{\max}$, in increments of distance Δd. In this work, we take $d_{\max} = 20$ kpc and $\Delta d = 0.1$ kpc for the timescale estimates.

For an input direction (l, b), we conduct Monte Carlo sampling with $N = 10^4$ using these independent variable priors to create a theoretical distribution of potential scintillation timescales. These distributions tend to be significantly skewed with long tails, so we use quartiles to characterize the best timescale ranges to target.

To help choose target directions for our survey, we repeat this process along a grid near the Galactic center and capture summary statistics for each sky direction's Δt_d distribution. Figure 5.2 shows each quartile (25%, median, and 75%) of the sampled timescale distribution for $-10° \leq l \leq 10°$, $-5° \leq b \leq 5°$, and $\Delta l = \Delta b = 0.25°$. Contours are shown for timescales of 10, 30, 60, and 100 s. For any particular sky direction, if our search procedure is designed to detect the range of timescales from the lower quartile to the upper quartile, we are by definition sensitive to 50% of the expected timescales from that direction. While this alone is not particularly comprehensive, it at least provides a point of reference for constraining the long-tailed Δt_d distributions.

5.2.2 Observing Strategy

We design the survey first by deciding which timescales might be best detected with the BL pipeline at the GBT. However, deciding the range of timescales to which our search should be sensitive is somewhat arbitrary, since it is fully continuous. Absent of any practical constraints, sensitivity to scintillated intensity fluctuations intuitively increases with higher time resolution Δt and longer observation times τ. While this would broaden the range of detectable timescales, this comes with a price in terms of data storage and telescope time. This cost is magnified for a survey, where there are many targets, none of which are known a priori to be more likely to harbor technosignatures. The survey must balance covering a wide enough berth around the Galactic center while taking detailed observations for each individual target.

While the GC itself is interesting for SETI, the scattering properties towards the GC are still uncertain and potentially at odds with NE2001 predictions (Yao et al. 2017; Suresh et al. 2021). So, we would also like to survey the broader Galactic plane, thereby probing other scattering screens. However, the GBT beam size is about 2.2' at the center of C-band (1.6–3.2' across the band). Arranging observations adjacently is only viable in a cluster near the GC, whereas a broader Galactic plane map will necessarily be coarser.

Based on Figure 5.2 and the Monte Carlo simulations that went into it, we choose to target timescales from 10–100 s. In Figure 5.3, we calculate the fraction of sampled timescales in each sky direction covered by this interval, along with contours for 25%, 50%, and 75% coverage. For about $-2° \leq b \leq 2°$, over half of the sampled timescales are within this range. In fact, the coverage peaks near $\pm 1°$ and dips towards the Galactic plane. This is expected; directions along the Galactic plane should have shorter scintillation timescales and dip below the lower edge of the interval.

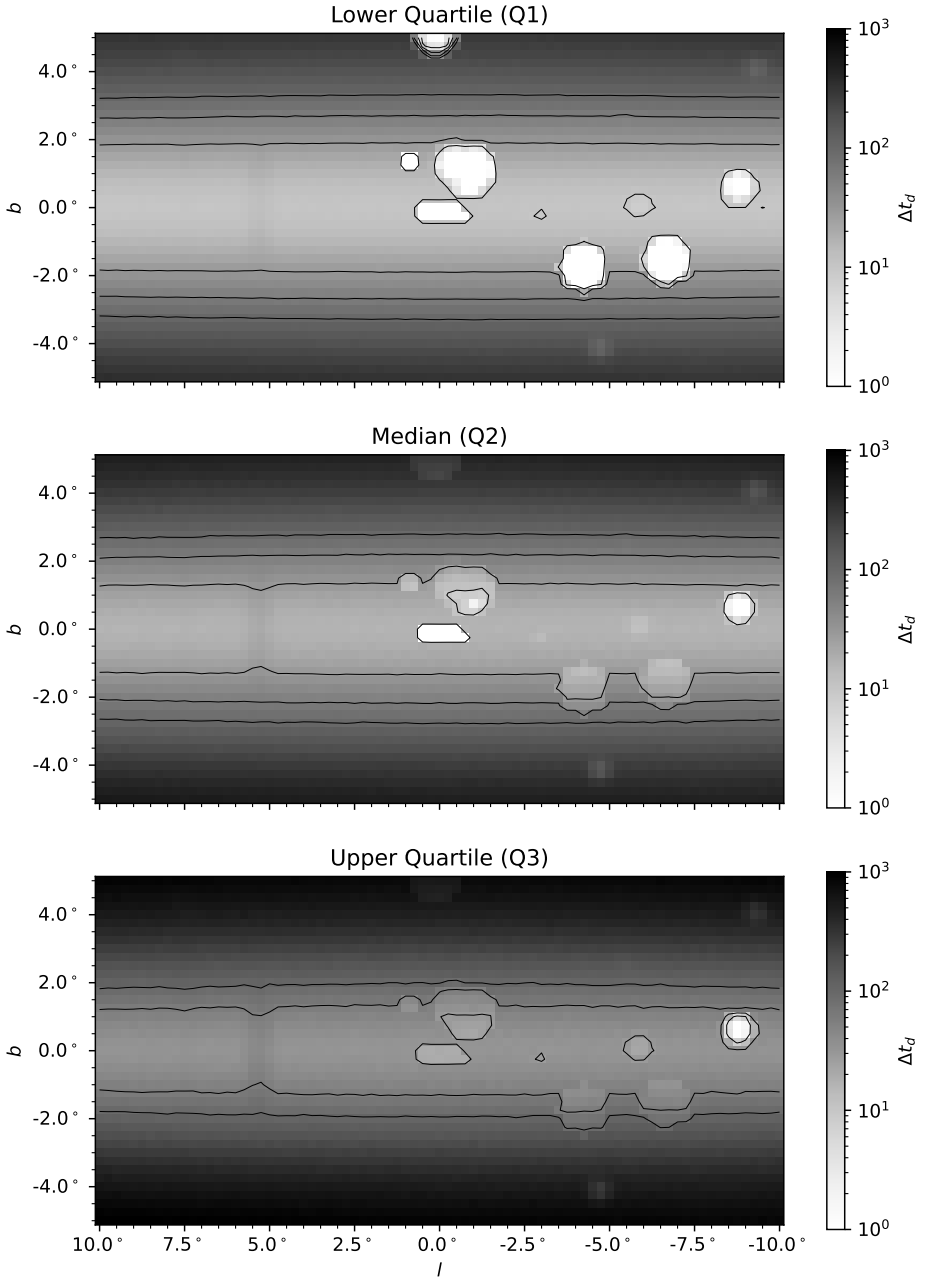


Figure 5.2: Sky map of the first, second, and third quartiles for scintillation timescale Δt_d, with resolution $\Delta l = \Delta b = 0.25°$. Contours are plotted in each panel for timescales of 10, 30, 60, and 100 s.

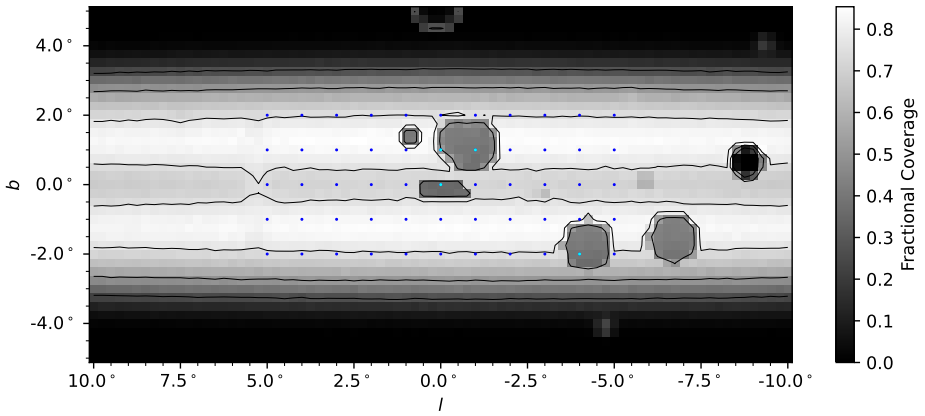


Figure 5.3: Sky map of the fraction of the range 10 s $\leq \Delta t_d \leq$ 100 s covered by Monte Carlo-sampled scintillation timescales, with resolution $\Delta l = \Delta b = 0.25°$. Contours for 25%, 50%, and 75% coverage are shown. The dots show the Galactic plane targets for this survey.

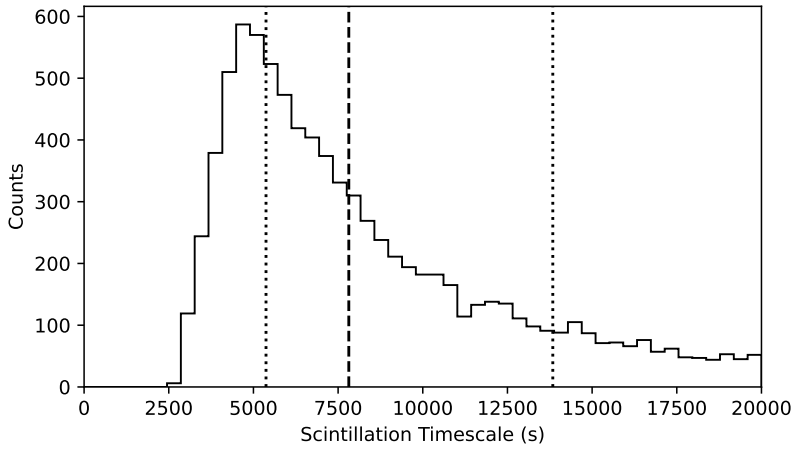


Figure 5.4: Monte Carlo-sampled scintillation timescales for the North Galactic Pole, $(l, b) = (0°, 90°)$, at C-band with $N = 10^4$. The dashed line indicates the median timescale, and the dotted lines indicate the first and third quartiles. As expected, compared to a pointing near the Galactic center (Figure 5.1), the expected timescales are significantly longer.

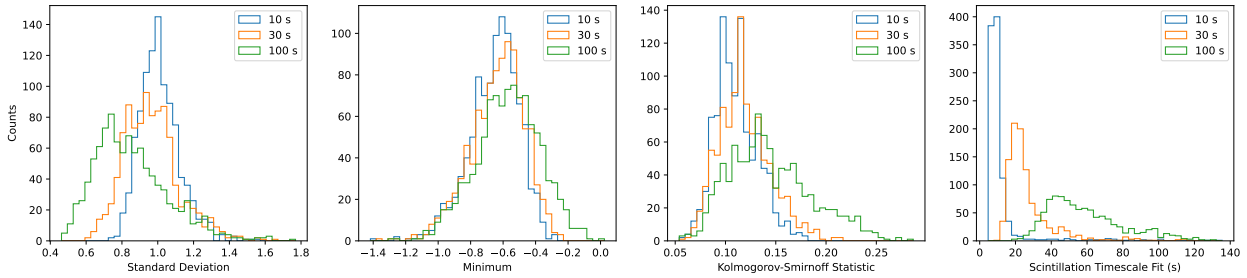


Figure 5.5: Histograms of diagnostic statistics computed using $N = 1000$ realizations of synthetic scintillated intensity time series embedded in chi-squared radiometer noise. The synthetic observations had $\Delta f = 2.79$ Hz, $\Delta t = 2.5$ s, and $\tau = 600$ s, matching the observations taken in this study. The signals were generated with a bandwidth of 8 frequency bins (about 22 Hz) and S/N = 33. Noisy intensity time series were extracted from the synthetic observations and used to compute each diagnostic statistic.

To capture 10 s $\leq \Delta t_d \leq$ 100 s, we base our observation parameters heuristically on these limits. To resolve individual scintles, the time resolution must be a factor smaller than the minimum timescale. Here, we choose a factor of 4, so that $\Delta t \lesssim \Delta t_{d,\min}/4$. Similarly, to capture enough scintles, we choose to require a factor of 5 larger than the maximum timescale, so that $\tau \gtrsim 5\Delta t_{d,\max}$.

Typically, individual BL observations are 5 minutes long, and for narrowband searches, the data is reduced into fine spectral resolution spectrograms with time and frequency resolutions of 18.25 s and 2.79 Hz, respectively. These data products are about 100 GB in total for each C-band observation. For narrowband scintillation observations, we keep the same frequency resolution, but extend the observations to $\tau = 10$ min and reduce the data to a finer $\Delta t = 2.5$ s. This increases the data size per observation by a factor of 14.6, for a total of 1.6 TB per pointing.

For the Galactic center (GC) map, we choose to match the 19 GBT pointings used in Gajjar et al. 2021 for C-band, observing in the same order as to maximize separation between ON and OFF pointings. This naturally gives a standard of comparison for any interesting detections in our observations. This totals 6 hr 20 min of observations, excluding slew time. For the Galactic plane (GP) map, we take observations spaced by $\Delta l = \Delta b = 1^\circ$ on an 11×5 grid with $-5^\circ \leq l \leq 5^\circ$ and $-2^\circ \leq b \leq 2^\circ$. Excluding the GC observation already taken, this amounts to 54 new pointings, for a total of 18 hours. These targets are outlined in detail in Table 5.2 (Appendix 5.8) and are shown overlaid in Figure 5.3.

For all of these targets, we would like to not only conduct a scintillation analysis, but also apply an ON-OFF directional filter to identify robust candidates. So, we group each target into pairs and take two interleaved 10 min integrations, comprising an ABAB cadence. In this way, every target gets a total of 20 min, and signals found within the two 10 min

Table 5.1: Survey Details

Start Target	Start Date (UTC)	Start MJD	No. Targets	Session Target Ranges
NGP0	2023-05-25 03:52:06	60089.16118056	14	GP_L5_B2, ..., GP_L3_B-1
NGP1	2023-06-08 02:53:41	60103.12061343	18	GP_L3_B-2, ..., GP_L-1_B0
NGP2	2023-07-20 00:14:19	60145.00994213	16	GP_L-1_B-1, ..., GP_L-4_B1
NGP3	2023-11-27 15:19:53	60275.63880787	19	GP_L-4_B2, ..., GC_C05
NGP4	2024-02-12 10:49:39	60352.45114583	6	GC_C08, ..., GC_C12

Note. Session targets follow the initial NGP observation within the same observing session. Observations were taken in ABAB cadences, with targets drawn from Table 5.2 (Appendix 5.8) in chronological order. The session ranges list the starting and ending targets within a single observing session.

integrations can be compared in terms of their intensity statistics. Note that since there are an odd number of GC pointings, to include the true Galactic center pointing A00, we observe in the order A00–C01–C07–A00–C01–C07; for all other targets, we observe in alternating pairs.

In addition to these targets of interest, we take an additional observation of the North Galactic Pole (NGP), $(l, b) = (0°, 90°)$, at the beginning of every observing session. Figure 5.4 shows a histogram of Monte Carlo simulations of scintillation timescales towards the NGP. These are significantly longer than our observation times, so we do not expect to detect ISM scintillation towards the NGP under our observation and data parameters. Note that this does not necessarily mean that strong scintillation is not present in this direction on the sky, but rather that our particular study will not be sensitive to the physical timescales in question. These observations provide a useful control for the RFI environment for each observing session and further allow comparison of RFI over time, from session to session.

The GBT observations are processed and digitized as complex voltages by the Breakthrough Listen Digital Recorder, which uses field-programmable gate array (FPGA) boards developed by the Collaboration for Astronomy Signal Processing and Electronics Research (CASPER; Parsons et al. 2006; Hickish et al. 2016; MacMahon et al. 2018). These complex voltages are reduced to the final Stokes-I intensity spectrogram data products at the target fine resolutions using `rawspec`[1] (Lebofsky et al. 2019; Brzycki et al. 2022). The survey details, including the start date of the initial NGP observation for each observing session, are summarized in Table 5.1.

5.3 Methods

5.3.1 DeDoppler Search

In Stokes I spectrogram data, narrowband signals with constant intensity appear as thin, continuous features in the time direction. Relative acceleration between the source and

[1] https://github.com/UCBerkeleySETI/rawspec

receiver causes changes in the Doppler shift of a signal over time, an effect commonly referred to as Doppler drift. Even if the acceleration is cyclic (e.g. part of a planet's orbit or rotation), if the observation time is small compared to the periodicity, the Doppler drift rate will be approximately constant and the signal path will be linear in time-frequency space.

To find narrowband signals in spectrogram data, we use the deDoppler code `turboSETI`[2], which efficiently implements the tree deDoppler algorithm (Taylor 1974; Siemion et al. 2013; Enriquez et al. 2017; Enriquez and Price 2019). Each signal in the data has an unknown drift rate, so the algorithm integrates spectrograms along a series of trial drift rates in order to find the drift rate that maximizes the detected signal-to-noise (S/N) ratio. We call each detection in an observation a "hit." The essential output from running `turboSETI` are the detected hits' starting central frequencies, drift rates, and S/N ratios. We exclude "detections" of the DC bin from each coarse channel, a systematic artifact of the discrete Fourier transform operation used in the voltage reduction process. The output format of `turboSETI` is essentially a plain text table, which can be easily read into Python using the `Pandas` package for further analysis (McKinney et al. 2011).

Even though the time resolution and observing length are different from previous BL searches, we aim to stay consistent by running `turboSETI` with a purported detection threshold of S/N = 10. However, Choza et al. 2023 found that `turboSETI` systemically overestimates its own sensitivity by a factor of about 3.3, so the true detection threshold of this and prior studies is S/N = 33. In fact, brighter detected signals allow for better separation from noise and therefore yield more accurate scintillation statistics, so we elect to continue using these inputs for the detection pipeline in this study.

We run our detection pipeline up to drift rate limits of ± 10 Hz/s. The minimum drift rate step used is $\Delta f/\tau \approx 0.004$ Hz/s, so about 5000 Doppler drift trials per signal are used to find the best matched fit. We run `turboSETI` in parallel over spectrogram data distributed across 40 BL compute nodes at the GBT, utilizing the GPU on each compute node for additional computational efficiency.

5.3.2 Direction-of-Origin Filter

To ensure that signals of interest are localized in the sky, we apply a direction-of-origin (or ON-OFF) filter as a discriminant against local RFI. Narrowband signals that appear as hits in each ON observation and in no OFF observations are considered "events" worthy of manual follow-up inspection.

In this work, we observed our targets in ABAB cadences. So, we perform the ON-OFF analysis on both A and B targets, taking each as the ON direction and filtering for events.

Determining which hits correspond to the same signal can happen in a few ways. To stay consistent with past BL searches, we exclude hits from consideration in which there is a hit in an OFF observation with frequency ν_{off} with

$$\nu_0 - |\dot{\nu}| \cdot 2\tau \leq \nu_{\text{off}} \leq \nu_0 + |\dot{\nu}| \cdot 2\tau. \tag{5.2}$$

[2] https://github.com/UCBerkeleySETI/turbo_seti

Similarly, we exclude hits if they do not have a corresponding hit within the appropriate frequency range in the other ON observation.

We note that this type of criteria for ON-OFF filtering has been criticized as unnecessarily broad, since it excludes a relatively large bandwidth which may be occupied by unrelated signals. Margot et al. 2021 instead tightened the frequency bounds for related hits and imposed an additional requirement that drift rates must be close across hits to comprise an event (of course requiring that these must not also be satisfied for any hits in OFF pointings). While this approach is significantly more precise for identifying events for ideal linearly-drifting narrowband signals, it is not uncommon to observe continuous signals whose component hits vary in drift rate from pointing to pointing. Since these signals often appear in adjacent ON-OFF pairs, they are classified as RFI, but we cannot assume that technosignatures could not exhibit this kind of drift rate variance over similar observational timescales. Whether it is a physical acceleration or an unstable oscillator creating this apparent variation, it may be prudent to relax the assumption that hits belonging to the same signal must have similar *fitted* drift rates. Narrowband signals in practice can be quite complex, so developing a robust and nuanced approach for identifying technosignature events could be an important avenue for future investigation.

5.3.3 Scintillation Diagnostic Statistics

To determine whether a signal is ISM scintillated, we must extract and analyze the intensity time series from noisy spectrograms. Given a signal's starting frequency ν_0 and drift rate $\dot{\nu}$ from the deDoppler analysis, we follow the procedure described in Brzycki et al. 2023 to extract a normalized intensity time series:

1. Select spectrogram frame with a truncated frequency band centered around detected signal. The frequency bandwidth is $N_f \Delta f + \dot{\nu}\tau$, with $N_f = 256$ pixels.
2. De-drift signal by shifting each spectra in the frame so that the signal is aligned in the frequency direction.
3. If possible, bound the signal by integrating in the time direction and identifying edges of the signal at 1% of the maximum integrated intensity. If the signal bandwidth is too large with respect to the spectrogram size, return to Step 1 with $N_f \leftarrow 2N_f$.
4. Normalize frame over the frequency direction, using the off-signal background.
5. Truncate frame using edges from Step 3 and integrate intensities along frequency direction.
6. Normalize resulting time series to a mean intensity of 1.

We then analyze the normalized time series to determine whether the signal intensities are consistent with strong ISM scintillation. Brzycki et al. 2023 identified a set of *diagnostic*

statistics that can help identify the presence of scintillation, including the standard deviation, the minimum, the Kolmogorov-Smirnoff (K-S) statistic, and the best fit timescale to the autocorrelation function (ACF).

5.4 Results

5.4.1 Signal Distributions

We conduct the deDoppler hit search, apply the direction-of-origin filter to find events, and compute scintillation diagnostic statistics over all cadences in our survey. The 54 GP targets make up 27 cadences, and the 19 GC targets are divided into 10 cadences, in which the triplet A00–C01–C07 is divided into two pseudo ABAB cadences, A00–C01 and C01–C07. Across all observations of the 73 targets (24.3 hr), we detect a total of 1.28M hits and 6018 events. Figure 5.6 shows examples of detected hits, arranged in ABAB cadences.

We plot the distributions of hit properties such as the detected frequency, drift rate, S/N, and bandwidth in Figure 5.7. Note that the bandwidths were only estimated as a by-product of the scintillation analysis, and as such, this is the first time in a narrowband search that we can analyze the bandwidth distribution for large number of detected hits.

We can readily identify a few primary groups of radio emission by frequency; the largest is centered at about 4 GHz, followed by a couple centered at about 4.6 GHz and 8.3 GHz, and perhaps another group at about 7.5 GHz. Note that the histogram counts are plotted on a logarithmic scale. These groups coincide with known regions populated by satellite emissions (Choza et al. 2023), so it is very likely that signals found at these frequencies are attributable to RFI.

When we inspect the bandwidth plots in Figure 5.7, we notice a population of hits over about 6 kHz in detected bandwidth. On inspection, these signals seem to have both narrowband and broadband features; in many cases, the narrowband components are repeated in a comb-like structure. These are detected by the deDoppler pipeline but do not match the morphology that we expressly search for, high duty cycle narrowband signals. Panel (f) in Figure 5.6 is an example with a relatively large measured bandwidth. In these cases, the hit detection reduces to something closer to an energy detection algorithm, and it is still very useful to detect these signals for use in ON-OFF filtering. After all, we do not necessarily know the true morphologies of technosignatures. While previous studies have undoubtedly detected many similar signals in this way, this is the first study to our knowledge that expressly reports on the signal bandwidths as a fundamental step in the analysis.

For the diagnostic scintillation statistics, it is important to compare the results with theoretical distributions as a point of reference. While there are ideal asymptotic values if the statistics were computed over infinite time series, since real observations are finite duration, this introduces natural variance of the estimated statistics around the asymptotic values. So, it is helpful to simulate scintillated signals embedded in realistic radiometer noise to observe the empirical spread in these statistics as a function of Δt_d. Figure 5.5 shows histograms of

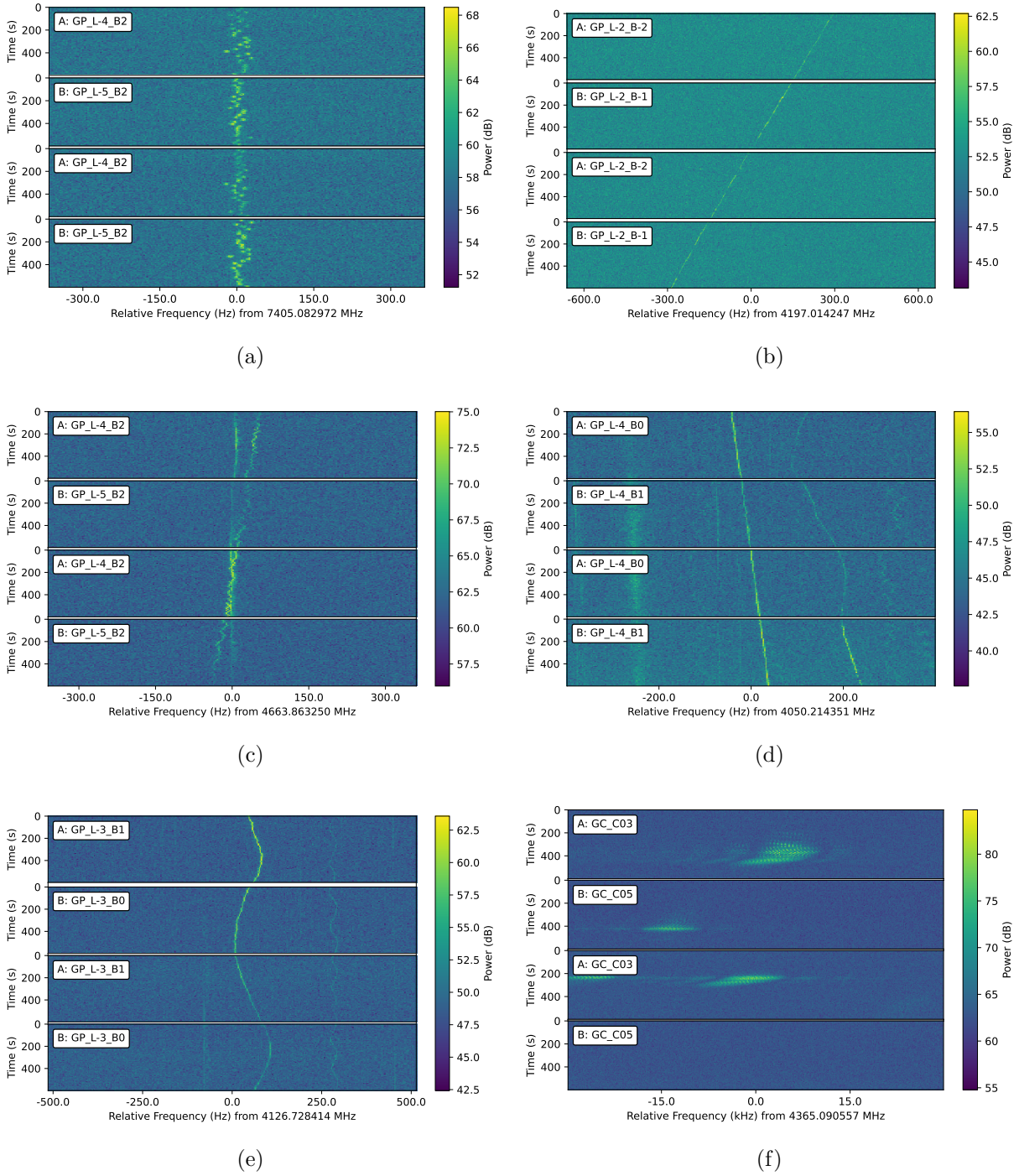
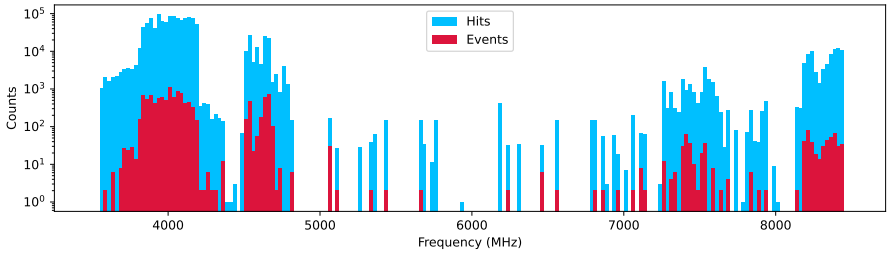
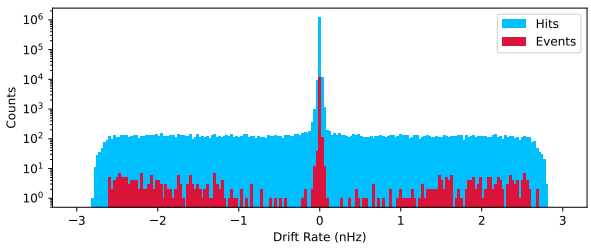


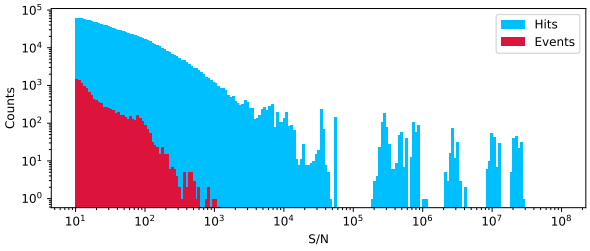
Figure 5.6: Examples of signals found from the deDoppler search which passed the algorithmic event filter, but failed manual inspection.



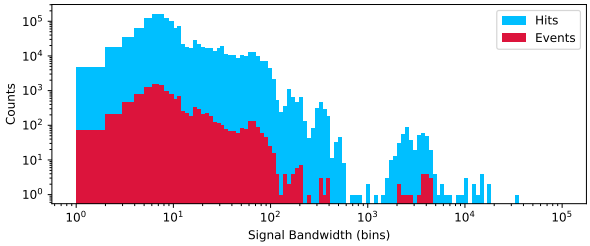
(a) Detected hit and event frequencies



(b) Detected hit and event drift rates



(c) Detected hit and event S/N ratios



(d) Detected hit and event signal bandwidths

Figure 5.7: Histograms of frequencies, drift rates, and S/N ratios of all detected hits in the Galactic center and plane survey.

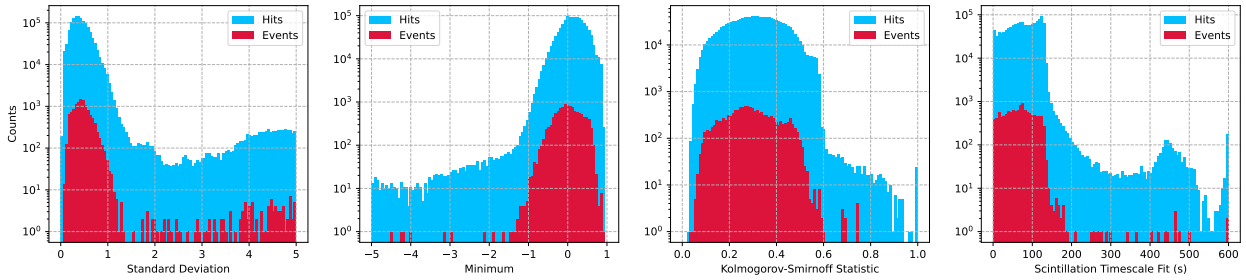


Figure 5.8: Histogram of diagnostic statistics of detected hits and events throughout all Galactic center and plane observations.

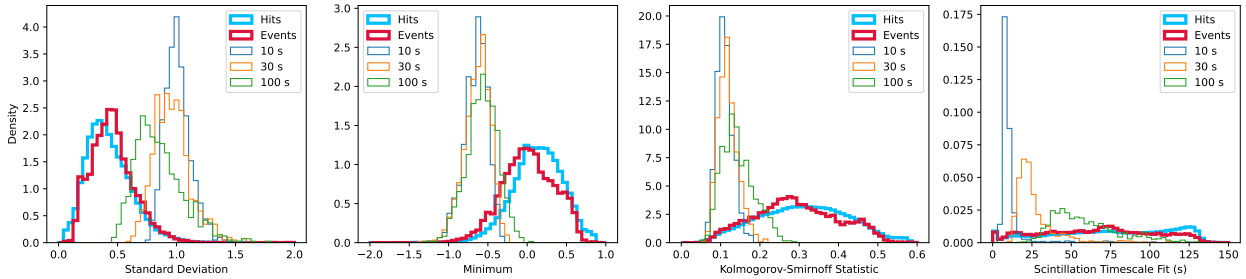


Figure 5.9: Histogram of diagnostic statistics of detected hits and events throughout all Galactic center and plane observations, compared to the synthetic distributions.

the diagnostic statistics computed over three datasets of synthetic observations corresponding to $\Delta t_d = 10$ s, 30 s, and 100 s, each with $N = 1000$ realizations. The scintillated intensities were generated using `blscint`[3] and injected in synthetic noise with `setigen` at an S/N ratio of 33, matching the sensitivity of the deDoppler search, with the same time resolution and observation length as our GBT data (Brzycki et al. 2022).

The peak (mode) of the bandwidth histogram in Figure 5.7 occurs at about 6–7 frequency bins, corresponding to about 20 Hz. So, for our synthetic scintillation distributions, we create the artificial signals with a bandwidth of 8 frequency bins as a representative match for a large proportion of the detected hits. In the future, it may be useful to match the empirical bandwidth distribution for the creation of synthetic signals, but here we keep the frequency profile consistent across the entire dataset.

In Figure 5.8, we plot histograms of the diagnostic statistics for all detected hits and events. To directly compare these with the synthetic datasets, we normalize these histograms into probability density functions by dividing by the bin size and total counts, so that each

[3] https://github.com/bbrzycki/blscint

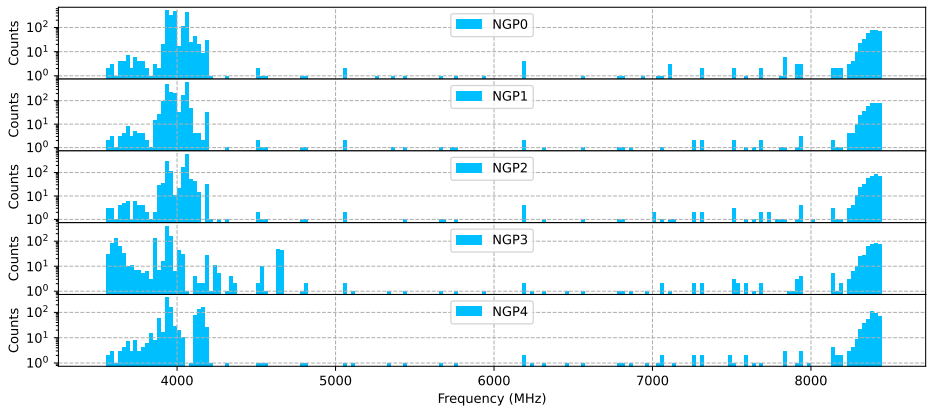


Figure 5.10: Histogram of frequencies of detected hits in each NGP observation.

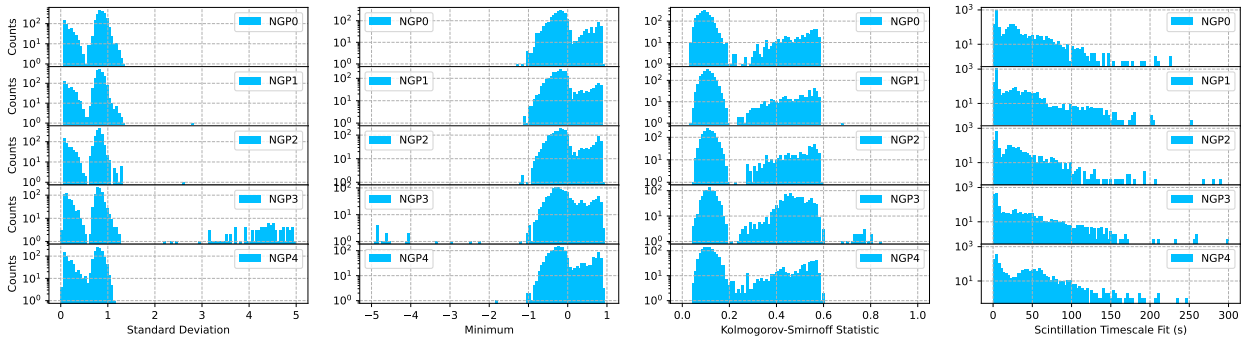


Figure 5.11: Histogram of diagnostic statistics of detected hits in each NGP observation.

distribution integrates to 1. We arrive at Figure 5.9, noting that the density axis is plotted on a linear scale. The event distribution is quite similar to the hit distribution, with only slight excesses towards the parameter space occupied by our theoretical scintillated signals.

5.4.2 RFI Analysis of NGP Observations

The survey was taken over the course of 5 observing sessions, so we have 5 separate 10 minute observations of the NGP (Table 5.1). As shown in Figure 5.4, the expected scintillation timescales are quite high compared to our observing length of 600 s. We therefore do not expect to observe real physical scintillations. However, it is possible that RFI intensity modulations, for communication purposes or otherwise, appear similar to ISM scintillations and therefore confound our directed search. In addition, since we have multiple observations of the same target spaced over months, we can observe any changes in the summary statistics

of RFI at each epoch.

Figure 5.10 stacks the frequency histograms of detected hits in each NGP observation in chronological order. At a glance, we notice the largest changes in frequency occupation in NGP3 versus the other observations, with an excess near 3.6 GHz and a spattering of signals across 4–5 GHz. NGP4 has a noticeably clump of signals near 4.1 GHz. Otherwise, much of the same structure remains over the course of the survey. However, we can also compare these distributions to the frequency distribution of hits detected towards our targets of interest in Figure 5.7. While there are many more signals plotted for our main survey, there is a noticeable absence of signals between 7.3–8 GHz towards the NGP. In Figure 5.7, the group of signals at 4 GHz has a broader peak and there appears to be a spike of signals near 8.2 GHz that is not present towards the NGP.

We similarly plot histograms of extracted diagnostic statistics in Figure 5.11. Once again, NGP3 seems the most unique, with a spread of signals towards higher standard deviation, lower minima, and high K-S statistics. Here there are significant differences in distributions between the NGP pointings and the survey targets in Figure 5.8. First, there is a peak in the NGP standard deviations at about 0.9 that is not obvious for the targets. This is a heavily confounding factor, since the ideal asymptotic value for a scintillated signal is 1. Ironically, this confounding peak only appears in a direction for which we do not expect to observe physical scintillations and not in the directions in which we hope to. Otherwise, the general shapes of statistics distributions are quite different from those for the NGP and are comparatively smoother. Of course, Figure 5.8 contains the results from 73 different locations on the sky, so this blending is expected.

Overall, NGP3 seems to reflect the most distinct RFI environment for a series of statistics. We note from Table 5.1 that NGP3 is the most separated from the other observing sessions in time, namely by multiple months in either direction. It is also possible that the time of year is a significant variable for these differences, if RFI at the GBT varies seasonally, but it is difficult to make that claim without additional observations closer in epoch. A longitudinal study on narrowband RFI towards specific regions on the sky, taken with regularly-spaced observations, could help explain the nature of variations in RFI as a function of direction and time and help determine whether they are predictable.

5.4.3 Scintillation Likelihood Weighting

Beyond manually inspecting signal statistics, we can combine diagnostic statistics with respect to the expectations obtained from synthetic datasets and develop a rudimentary ranking metric for all hits in our search. In other words, we seek to turn the comparisons possible in Figure 5.9 into a representative weight for each hit.

For simplicity, we use the normalized synthetic distributions in Figure 5.9 as empirical probability density functions (PDFs) for each diagnostic statistic, represented as $p(x, \Delta t_d)$. We can think of each hit as a vector of diagnostic statistics. For a given timescale, we define

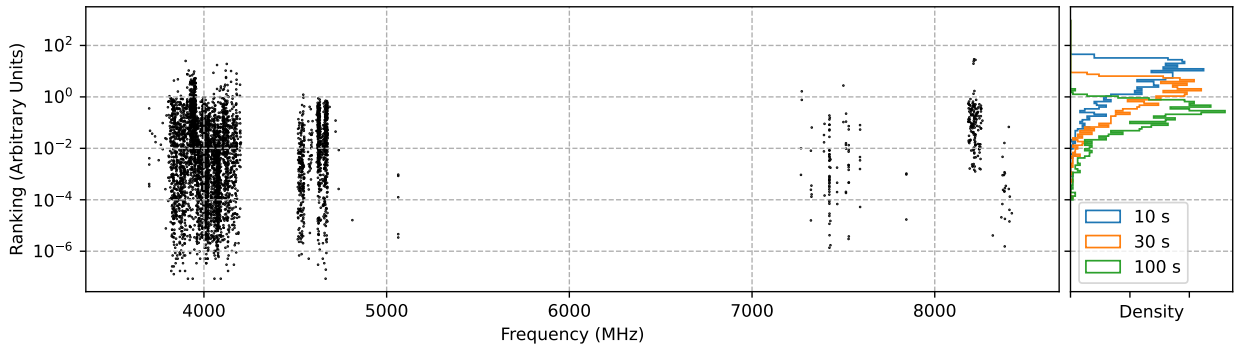
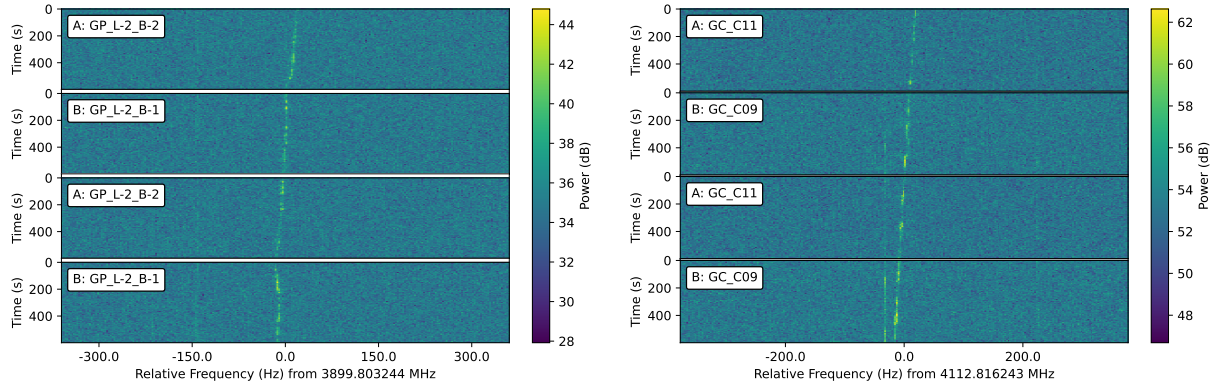


Figure 5.12: Ranking estimates for all hits in detected events as a function of frequency. Synthetic ranking distributions are shown on the right panel for timescales of 10 s, 30 s, and 100 s.



(a) (b)

Figure 5.13: Examples of signals found from the deDoppler search which passed the algorithmic event filter and ranked highly on the scintillation analysis.

the weight

$$w(\mathbf{h}, \Delta t_d) = p_{\text{std}}(h_{\text{std}}, \Delta t_d) \times p_{\text{min}}(h_{\text{min}}, \Delta t_d) \times p_{\text{KS}}(h_{\text{KS}}, \Delta t_d) \times p_{\text{fit}}(h_{\text{fit}}, \Delta t_d). \tag{5.3}$$

To arrive at the final weight, we take

$$w(\mathbf{h}) = \max_{\Delta t_d \in [10\text{ s}, 30\text{ s}, 100\text{ s}]} w(\mathbf{h}, \Delta t_d), \tag{5.4}$$

with the rationale that if a hit is truly scintillated with some timescale Δt_d, the weight $w(\mathbf{h}, \Delta t_d)$ should be maximized for the most appropriate Δt_d among 10 s, 30 s, and 100 s. If a hit is clearly not scintillated and its diagnostic statistics are far from the empirical PDFs, the weight w will typically be 0.

We compute $w(\mathbf{h})$ over all detected hits that were part of events and plot each against frequency in Figure 5.12. For reference, we compute rankings for each signal in the synthetic datasets as well and plot those distributions to the right. We readily note the apparent frequency groups of emitters that we identified in Section 5.4.1. Within each group, hit rankings span multiple orders of magnitude.

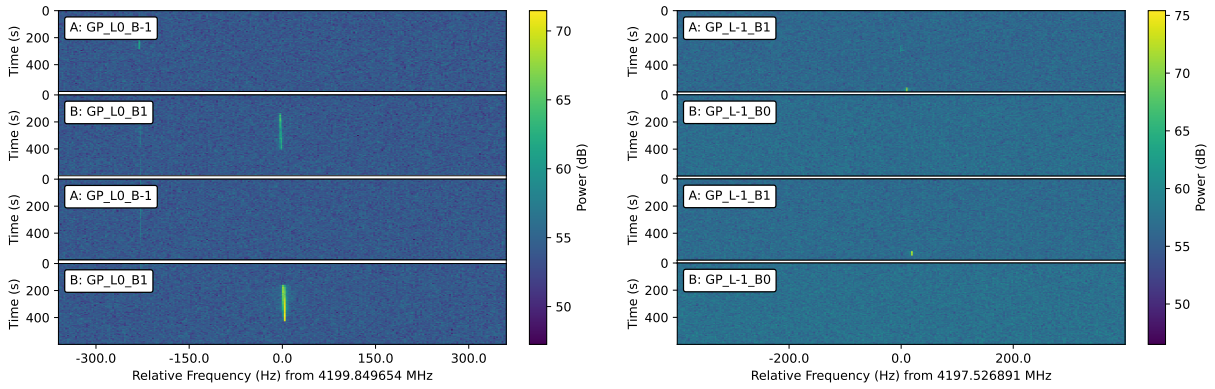
In Figure 5.13, we show examples of detected events that contain a highly ranked hit relative to the total collection of hits. In both examples, the highest ranking hit was in the first B pointing (second spectrogram), and the ranking values were 24.9 and 2.5, respectively. Visually, these signals appear to exhibit desired scintillation properties, but they appear in both ON and OFF pointings. Though it is simplistic, this heuristic allows us to order events for manual inspection based on which are most likely to be scintillated.

5.4.4 Potential Candidates

We manually inspect all 6018 events that passed our sky localization algorithm. We find that the signal is clearly still present in at least one OFF pointing, as is the case for each signal in Figure 5.6, in all but 2 events. Those candidate events are shown in Figure 5.14.

It is quite clear that for both events, the detected signal is pulsed, since they do not persist throughout their entire observations. As such, it is possible that these signals are RFI that happened to only emit during the A or B targets of a cadence. While we only took 2 observations per target, we can check adjacent cadences taken during the same observing session. Since our survey is concentrated relatively close to the GC, if these are RFI signals that were caught in the telescope sidelobes, we would still expect to see them in nearby cadences.

In Figure 5.17 (Appendix 5.10), we plot 4 cadences each, all from the same observing session. For event (a), the cadences shown were taken consecutively, and the original detection was the second cadence from top to bottom. It is clear that the pulsed signal is present before and after the cadence of detection, in observations of different targets in the sky, so it must be RFI. Likewise, for event (b), the cadences shown were not taken consecutively, but



(a) (b)

Figure 5.14: Best candidates from the direction filter, which passed initial manual inspection.

were all from the same observing session. The original detection is the bottom cadence, but we can see short pulses with the same duration at nearby frequencies in each of the other cadences. Since this signal was also detected in other pointings, we conclude that it is most likely attributable to RFI as well.

5.5 Discussion

5.5.1 Search Sensitivity

For Doppler drifting narrowband signals, the minimum detectable flux is given by

$$F_{\min} = \mathrm{S/N}_{\min} \frac{S_{\mathrm{sys}}}{\beta} \sqrt{\frac{\Delta f}{N_{\mathrm{pol}} \tau}}, \tag{5.5}$$

where $\mathrm{S/N}_{\min}$ is the detection threshold, S_{sys} is the system equivalent flux density (SEFD), N_{pol} is the number of polarizations, and β is the dechirping efficiency, the factor by which our detected S/N falls off as a result of Doppler drift (Gajjar et al. 2021). For drift rates $\dot{\nu}$ that are greater than the unit drift rate $\dot{\nu}_1 = \Delta f / \Delta t$, signal power will smear across adjacent frequency channels within each individual spectrum, resulting in an apparent loss of sensitivity following $\beta \sim \dot{\nu}_1 / \dot{\nu}$ (Brzycki et al. 2022; Choza et al. 2023).

To estimate the SEFD for GP targets, we use the measured values provided by the GBT for C-band, where $S_{\mathrm{sys}} = 2kT_{\mathrm{sys}}/A_{\mathrm{eff}}$, where k is the Boltzmann constant and T_{sys} is the system temperature (GBT Support Staff 2017). However, GC pointings additionally capture radio continuum emission from the GC background, which must be accounted for in the noise power as $S_{\mathrm{sys}} = 2k(T_{\mathrm{sys}} + T_{\mathrm{GC}})/A_{\mathrm{eff}}$, where T_{GC} is the brightness temperature

corresponding to the GC background. Following Gajjar et al. 2021, we use the approximation from Rajwade et al. 2017 that $T_{\rm GC} \approx (568\ {\rm K})/\nu_{\rm GHz}^{1.13}$.

We can then estimate the minimum detectable EIRP from the inverse square law

$$\mathrm{EIRP}_{\rm min} = 4\pi d^2 F_{\rm min}, \tag{5.6}$$

for source distance d. For transmitters located at a distance of 8.5 kpc towards GP pointings, we obtain a limit of $\mathrm{EIRP}_{\rm min} = 1.9 \times 10^{17}$ W. For the maximum drift rates (10 Hz/s) searched in this study, we will get a larger minimum EIRP by a factor of about $1/\beta \sim 9$, yielding $\mathrm{EIRP}_{\rm min} = 1.7 \times 10^{18}$ W. For targets at 8.5 kpc towards GC pointings, we obtain $\mathrm{EIRP}_{\rm min} = 7.5 \times 10^{17}$ W for $\beta = 1$, and $\mathrm{EIRP}_{\rm min} = 6.7 \times 10^{18}$ W at maximum drift rates.

5.5.2 Figures of Merit

Since there are so many axes of analysis and such a large parameter space intrinsic to SETI, it can be difficult to directly and meaningfully compare technosignature searches. One common way is to use so-called "figures of merit," which assign a representative value towards each study based on their observational search parameters.

The popular Drake Figure of Merit is given by

$$\mathrm{DFM} = \frac{\Delta\nu_{\rm tot}\Omega}{F_{\rm min}^{3/2}}, \tag{5.7}$$

where $\Delta\nu_{\rm tot}$ is the total observing bandwidth and Ω is the total sky coverage (Drake 1984). By design, the larger this value, the more comprehensive the SETI search. For our study, the DFM is about 1×10^{33} for $\dot{\nu} \leq 0.004$ Hz/s, down to 4×10^{31} for $|\dot{\nu}| = 10$ Hz/s.

However, the DFM only measures a few aspects of a search and treats each direction on the sky equally, which is especially misleading for our study because the distribution of potential transmitters is likely heavily biased towards the GC. To address these limitations, Enriquez et al. 2017 developed the Continuous Waveform Transmitter Rate Figure of Merit, defined as

$$\mathrm{CWTFM} = \zeta_{\rm AO}\frac{\mathrm{EIRP}_{\rm min}}{N_*}\frac{\nu_c}{\Delta\nu_{\rm tot}}, \tag{5.8}$$

where N_* is the number of observed stars, ν_c is the central frequency, and $\zeta_{\rm AO}$ is a normalization factor such that CWTFM = 1 when the EIRP matches that of Arecibo's planetary radar. The smaller the CWTFM, the more sensitive the study. The portion $\nu_c/(N_*\Delta\nu_{\rm tot})$ is referred to as the "transmitter rate," which encodes information about the portion of frequency space searched and the breadth of targets searched. We estimate the number of stars observed in the survey by numerically integrating Equation 5.1 along the full-width half-maximum of our beams. Up to 8.5 kpc away, we estimate a total of about 6.5 million and 3.5 million stars towards the GP and GC pointings, respectively, yielding CWTFMs of 2.2 and 17.

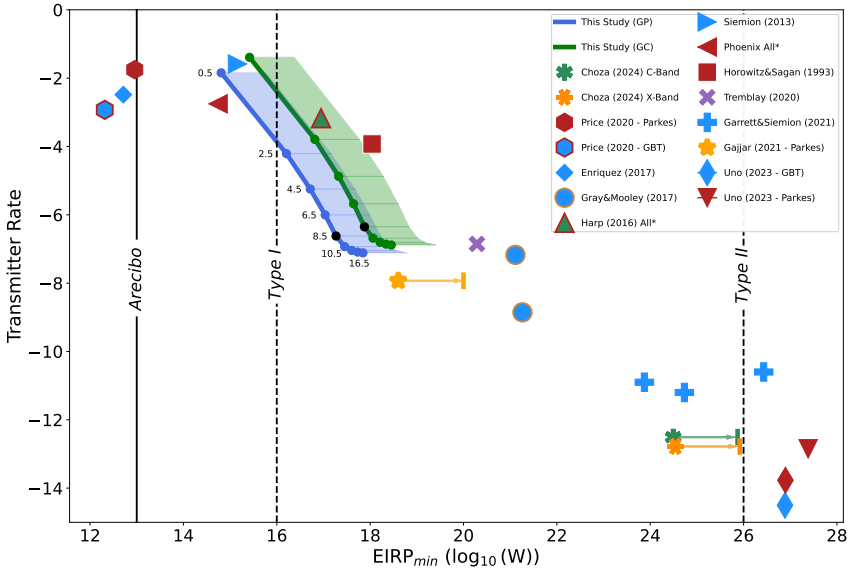


Figure 5.15: Transmitter rate vs. EIRP for this study and previous radio technosignature searches. EIRP limits for this study at various distances for the GP targets and GC targets are plotted in blue and green, respectively. The EIRP limits for sources up to 8.5 kpc are marked in black. For each distance, we extend EIRP limits up to the right, corresponding to the maximum drift rate of 10 Hz/s searched in this study, for which $\beta \approx 0.11$. Note that while this study and Choza et al. 2023 use S/N = 33 to correct for the offset factor in `turboSETI`, we choose not to alter the results from any earlier studies which may have been affected by this.

We plot the transmitter rate against maximum EIRP for this study in Figure 5.15, along with the results from prior technosignature searches. Since we have a method for estimating the number of stars within our telescope beams and since there is technically no hard distance cutoff, we can actually compute multiple CWTFM ratios as a function of distance through the Galaxy. We start at 0.5 kpc and extend the calculation by steps of 2 kpc until we get to 16.5 kpc, plotting each step as a separate point. We mark the point closest to the GC ($d = 8.5$ kpc) in black for contrast. Also, note that our EIRP limits are computed assuming $\beta = 1$, so for the highest absolute drift rates in the search (10 Hz/s), we will get a larger minimum EIRP by a factor of about $1/\beta \sim 9$, indicated by extending the limits to the right at each distance. For the groups of GP and GC pointings (in blue and green, respectively), this effectively builds up a shaded region in the plot, interpolating between EIRP estimates for each distance.

The transmitter rate plot for prior SETI searches has a general power law relationship with the minimum EIRP, which parallels a common assumption that ETI transmitters throughout the Galaxy might follow a natural power law (Drake 1984; Shostak 2000; Enriquez et al. 2017). Our 8.5 kpc limit marks the lowest transmitter rate and $\mathrm{EIRP}_{\mathrm{min}}$ limits of SETI searches near the Kardashev I scale, so this work does push the implied power law deeper.

The path of the values plotted depends on the model for stellar number density in our Galaxy, and seems to have a knee at about 10.5 kpc. Since 8.5 kpc gets the closest to the GC (and contains it), we focus on those EIRP limits as representative for this study, but it is interesting the note the seemingly diminishing returns we get in transmitter rate as we increase in distance and minimum EIRP. Of course, the number density of stars falls off as the distance from the GC increases, but even the search volume expanding with distance does not make up for this.

5.5.3 Detectability of ISM Scintillation

This search was designed to maximize the likelihood of detecting ISM scintillated technosignatures within relatively short observing times with BL hardware. Though scintillation is a stochastic effect, for the right Δt_d, multiple scintles should be resolved within the course of a single observation. However, we do not know the true values of Δt_d any more than we know what frequency ETI will transmit at, even towards the GC, so we can only estimate using theoretical models for the free electron density and for the distribution of stars in our Galaxy. In the future, we can get higher confidence in our estimates by updating these models, especially towards the GC (Ocker and Cordes 2024).

Furthermore, the RFI environment around GBT is uncertain and dynamic, as we saw in Section 5.4.2. There seem to be populations of RFI that are more similar to scintillated signals than others, and we seemed to detect a greater proportion towards the NGP. In fact, Brzycki et al. 2023 conducted an RFI analysis of the North Celestial Pole (NCP) at C-band, and found very few signals whose diagnostic statistics coincided with the distributions of synthetic scintillation signals. RFI is generally picked up through the sidelobes of the antenna pattern of radio telescopes, so it makes sense that we will observe some interference (especially from geosynchronous satellites) in certain directions and not in others. Unfortunately, it is difficult to pinpoint the exact origin for any given group of RFI. Since the detected RFI environment seems to systemically vary with observing direction, taking observations of separate sky locations to serve as controls for RFI detection or mitigation may simply be unhelpful or inapplicable. In other words, RFI occupancy analysis, such as that done by Choza et al. 2023, may necessarily have to directly use data from observations of targets of interest, even though they may contain technosignatures and therefore bias against true positive detections. We must be increasingly careful not to excise technosignature candidates, at the risk of preserving many false positives for manual inspection.

One exciting possibility behind searching for ISM scintillations in narrowband signals was the idea that it would give us a way to classify one-off signals as high quality technosignature

candidates. By convention, SETI searches have required that true candidates both pass the direction-of-origin filter and are detectable on re-observation (Sheikh et al. 2021). It is possible that a source transmits for a period of time but drops off; for instance, if an ETI is transmitting towards a set of targets one after another. However, in our analysis, we have detected individual hits in single observations towards the GC and GP that have high scintillation rankings $w(\mathbf{h})$. If these were not part of ON-OFF cadences, in which we were able to verify that the signals persisted throughout, we might otherwise consider these signals as potentially modulated by ISM scattering. It seems clear that unless we obtain a detailed understanding of the various intensity modulations present in RFI, we will continue to require sky localization filters. That being said, if a candidate signal observed within the GP passed sky localization filters and had a large scintillation ranking consistent with the expected timescales in that direction, it would be very compelling as a potential technosignature.

5.6 Conclusions

After observing 73 targets towards the GC and GP at C-band, we do not report any detections of narrowband radio signals that are inconsistent with anthropogenic RFI. Specifically, towards pointings in the GP, we find no evidence of putative radio transmitters with an EIRP above 1.9×10^{17} W up to 8.5 kpc, covering an estimated 6.5 million stars. Likewise, we find no evidence of putative ETI transmitters towards the GC with an EIRP above 7.5×10^{17} W up to 8.5 kpc, covering an estimated 3.5 million stars. We also find no evidence of scintillated signals towards the GC at the same EIRP limit, but find that the radio frequency environment at the GBT has populations of confounding RFI with scintillation-like intensity modulations. This interestingly depends on the direction in the sky, since observations of the North Galactic Pole, the direction in which we expect the least to observe true physical scintillation, revealed the highest concentrations of confounding RFI. Getting a better idea of which RFI types are semi-localized on the sky may help future SETI searches and inform which directions are most fruitful for observing potential scintillation. Nevertheless, RFI modulation types vary and are present in high enough quantities that scintillation analysis may not be enough to classify a one-off detection as a legitimate technosignature candidate; we will continue to require that signals both pass sky localization filters and are repeatable for the sake of additional confidence.

5.7 Acknowledgements

Breakthrough Listen is managed by the Breakthrough Initiatives, sponsored by the Breakthrough Prize Foundation. The Green Bank Observatory is a facility of the National Science Foundation, operated under cooperative agreement by Associated Universities, Inc. We thank the staff at the Green Bank Observatory for their operational support. B.B. would also like to thank Shelley Wright and Paul Horowitz for helpful discussions.

5.8 Appendix A: Observation Tables

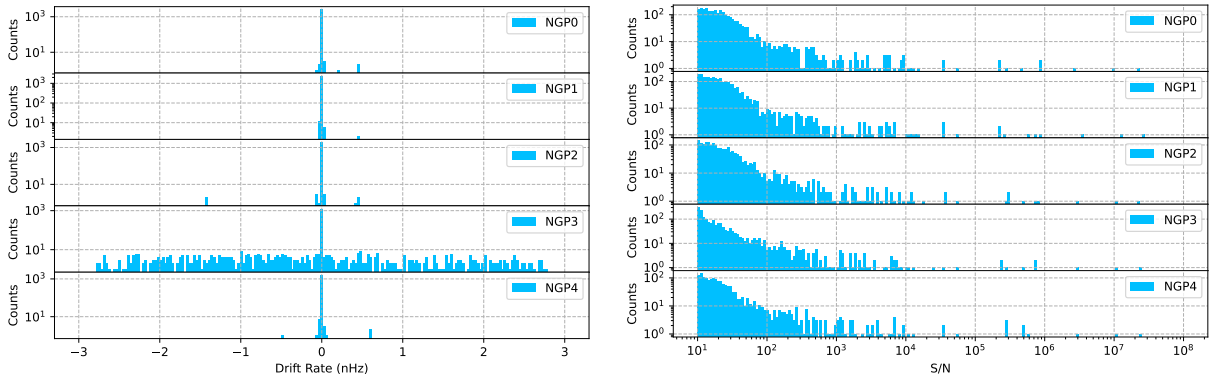
Table 5.2: Survey Target List

Target	l (deg)	b (deg)	R.A. (h:m:s)	Decl. (d:m:s)	Star Count	Δt_d Predictions (s)		
						Q1	Median (Q2)	Q3
GP_L-5_B2	-5.000	2.000	17:25:15.86	-32:03:46.66	8.3×10^4	32.6	52	106
GP_L-4_B2	-4.000	2.000	17:27:52.95	-31:13:59.85	8.6×10^4	32.6	51.3	102
GP_L-3_B2	-3.000	2.000	17:30:27.15	-30:24:00.12	8.9×10^4	32.4	50.7	103
GP_L-2_B2	-2.000	2.000	17:32:58.59	-29:33:48.06	9.1×10^4	32.6	50.6	103
GP_L-1_B2	-1.000	2.000	17:35:27.41	-28:43:24.26	9.3×10^4	32.6	51	104
GP_L0_B2	0.000	2.000	17:37:53.70	-27:52:49.27	9.4×10^4	32.4	50.2	101
GP_L1_B2	1.000	2.000	17:40:17.60	-27:02:03.61	9.3×10^4	32.7	51.1	102
GP_L2_B2	2.000	2.000	17:42:39.20	-26:11:07.80	9.1×10^4	32	50.6	103
GP_L3_B2	3.000	2.000	17:44:58.61	-25:20:02.30	8.9×10^4	33	51.6	106
GP_L4_B2	4.000	2.000	17:47:15.94	-24:28:47.58	8.6×10^4	32.6	51.3	107
GP_L5_B2	5.000	2.000	17:49:31.27	-23:37:24.09	8.3×10^4	32.9	51.5	107
GP_L-5_B1	-5.000	1.000	17:29:11.89	-32:37:09.93	1.2×10^5	15	23.6	48.3
GP_L-4_B1	-4.000	1.000	17:31:47.85	-31:47:03.71	1.2×10^5	15	23.5	48.5
GP_L-3_B1	-3.000	1.000	17:34:20.95	-30:56:45.23	1.2×10^5	14.9	23.1	48.2
GP_L-2_B1	-2.000	1.000	17:36:51.31	-30:06:15.08	1.3×10^5	15.1	23.6	49
GP_L-1_B1	-1.000	1.000	17:39:19.05	-29:15:33.83	1.3×10^5	0.201	8.49	29.1
GP_L0_B1	0.000	1.000	17:41:44.30	-28:24:42.01	1.3×10^5	4.73	13.8	31.6
GP_L1_B1	1.000	1.000	17:44:07.16	-27:33:40.14	1.3×10^5	15.2	23.6	47.7
GP_L2_B1	2.000	1.000	17:46:27.76	-26:42:28.71	1.3×10^5	15	23.7	47
GP_L3_B1	3.000	1.000	17:48:46.19	-25:51:08.18	1.2×10^5	15.3	23.9	47.6
GP_L4_B1	4.000	1.000	17:51:02.56	-24:59:39.01	1.2×10^5	15.3	24	49.1
GP_L5_B1	5.000	1.000	17:53:16.97	-24:08:01.62	1.2×10^5	16.3	25.7	55.2
GP_L-5_B0	-5.000	0.000	17:33:10.84	-33:10:05.27	1.6×10^5	10.3	16.5	33.4
GP_L-4_B0	-4.000	0.000	17:35:45.56	-32:19:40.30	1.7×10^5	10.5	16.3	31.9
GP_L-3_B0	-3.000	0.000	17:38:17.42	-31:29:03.75	1.7×10^5	10.4	15.8	31.5
GP_L-2_B0	-2.000	0.000	17:40:46.56	-30:38:16.19	1.8×10^5	10.3	16.1	32.2
GP_L-1_B0	-1.000	0.000	17:43:13.12	-29:47:18.18	1.8×10^5	10.4	16.4	32.5
GP_L1_B0	1.000	0.000	17:47:58.92	-28:04:52.87	1.8×10^5	10.5	16	31.9
GP_L2_B0	2.000	0.000	17:50:18.41	-27:13:26.54	1.8×10^5	10.4	16.1	31.8
GP_L3_B0	3.000	0.000	17:52:35.76	-26:21:51.70	1.7×10^5	10.3	16.2	32.7
GP_L4_B0	4.000	0.000	17:54:51.08	-25:30:08.81	1.7×10^5	10.4	16.1	32.2
GP_L5_B0	5.000	0.000	17:57:04.47	-24:38:18.26	1.6×10^5	11.5	18.1	37.6
GP_L-5_B-1	-5.000	-1.000	17:37:12.79	-33:42:31.73	1.2×10^5	15.1	24.1	51
GP_L-4_B-1	-4.000	-1.000	17:39:46.11	-32:51:48.70	1.2×10^5	15.2	24	48.8
GP_L-3_B-1	-3.000	-1.000	17:42:16.61	-32:00:54.79	1.3×10^5	15.1	23.8	49.3
GP_L-2_B-1	-2.000	-1.000	17:44:44.41	-31:09:50.54	1.3×10^5	15.1	23.7	47
GP_L-1_B-1	-1.000	-1.000	17:47:09.65	-30:18:36.49	1.3×10^5	15.2	23.8	48.5
GP_L0_B-1	0.000	-1.000	17:49:32.45	-29:27:13.16	1.3×10^5	15.1	24.3	49.7
GP_L1_B-1	1.000	-1.000	17:51:52.93	-28:35:41.02	1.3×10^5	15.1	23.5	48.8
GP_L2_B-1	2.000	-1.000	17:54:11.19	-27:44:00.54	1.3×10^5	15.1	23.6	48.1
GP_L3_B-1	3.000	-1.000	17:56:27.36	-26:52:12.15	1.3×10^5	15.1	23.7	47.7
GP_L4_B-1	4.000	-1.000	17:58:41.53	-26:00:16.28	1.2×10^5	15.1	23.6	47.4
GP_L5_B-1	5.000	-1.000	18:00:53.81	-25:08:13.32	1.2×10^5	16.3	26	57.4
GP_L-5_B-2	-5.000	-2.000	17:41:17.77	-34:14:28.34	8.5×10^4	33.3	52.9	109
GP_L-4_B-2	-4.000	-2.000	17:43:49.57	-33:23:27.99	8.8×10^4	0.122	28	70.1
GP_L-3_B-2	-3.000	-2.000	17:46:18.56	-32:32:17.44	9.0×10^4	32.6	50.7	106
GP_L-2_B-2	-2.000	-2.000	17:48:44.90	-31:40:57.25	9.3×10^4	32.3	50.5	105
GP_L-1_B-2	-1.000	-2.000	17:51:08.70	-30:49:27.93	9.4×10^4	32.3	50.8	104

Table 5.2: Survey Target List

Target	l (deg)	b (deg)	R.A. (h:m:s)	Decl. (d:m:s)	Star Count	Δt_d Predictions (s)		
						Q1	Median (Q2)	Q3
GP_L0_B-2	0.000	-2.000	17:53:30.10	-29:57:49.97	9.5×10^4	32.6	50.8	102
GP_L1_B-2	1.000	-2.000	17:55:49.21	-29:06:03.83	9.4×10^4	32.7	51.8	105
GP_L2_B-2	2.000	-2.000	17:58:06.15	-28:14:09.96	9.3×10^4	32.4	50.3	104
GP_L3_B-2	3.000	-2.000	18:00:21.03	-27:22:08.79	9.0×10^4	33.1	51.7	109
GP_L4_B-2	4.000	-2.000	18:02:33.95	-26:30:00.73	8.8×10^4	33.2	52.1	107
GP_L5_B-2	5.000	-2.000	18:04:45.01	-25:37:46.15	8.5×10^4	33.2	52	107
GC_A00	-0.056	-0.046	17:45:40.04	-29:00:28.10	1.8×10^5	0.00693	0.144	18.5
GC_C01	0.028	-0.046	17:45:51.95	-28:56:11.99	1.8×10^5	0.00692	0.0604	18.5
GC_C07	-0.139	-0.046	17:45:28.12	-29:04:44.14	1.8×10^5	0.00696	0.0716	18.8
GC_B01	-0.014	-0.046	17:45:45.99	-28:58:20.05	1.8×10^5	0.00699	0.0836	18.7
GC_B04	-0.097	-0.046	17:45:34.08	-29:02:36.13	1.8×10^5	0.00687	0.0649	18.3
GC_B02	-0.035	-0.010	17:45:34.57	-28:58:16.40	1.8×10^5	0.00738	0.0601	19
GC_B05	-0.077	-0.082	17:45:45.52	-29:02:39.78	1.8×10^5	0.00625	0.0519	18.2
GC_B03	-0.077	-0.010	17:45:28.61	-29:00:24.42	1.8×10^5	0.00781	0.117	18.9
GC_B06	-0.035	-0.082	17:45:51.47	-29:00:31.72	1.8×10^5	0.00652	0.0952	19.2
GC_C02	0.007	-0.010	17:45:40.52	-28:56:08.37	1.8×10^5	0.00749	0.0587	18.5
GC_C04	-0.056	0.026	17:45:23.14	-28:58:12.69	1.8×10^5	0.0104	0.106	19.4
GC_C03	-0.014	0.026	17:45:29.10	-28:56:04.69	1.8×10^5	0.0101	0.0921	19.1
GC_C05	-0.097	0.026	17:45:17.18	-29:00:20.67	1.8×10^5	0.0102	0.0925	18.6
GC_C08	-0.118	-0.082	17:45:39.56	-29:04:47.83	1.8×10^5	0.00652	0.0763	19.2
GC_C06	-0.118	-0.010	17:45:22.65	-29:02:32.42	1.8×10^5	0.00768	0.0849	18.7
GC_C11	-0.014	-0.118	17:46:02.90	-29:00:35.29	1.8×10^5	0.00652	0.142	18.7
GC_C09	-0.097	-0.118	17:45:51.00	-29:04:51.46	1.8×10^5	0.00664	0.101	19.2
GC_C10	-0.056	-0.118	17:45:56.95	-29:02:43.38	1.8×10^5	0.00638	0.0988	19.1
GC_C12	0.007	-0.082	17:45:57.42	-28:58:23.65	1.8×10^5	0.00681	0.0929	18.9

5.9 Appendix B: NGP Statistics

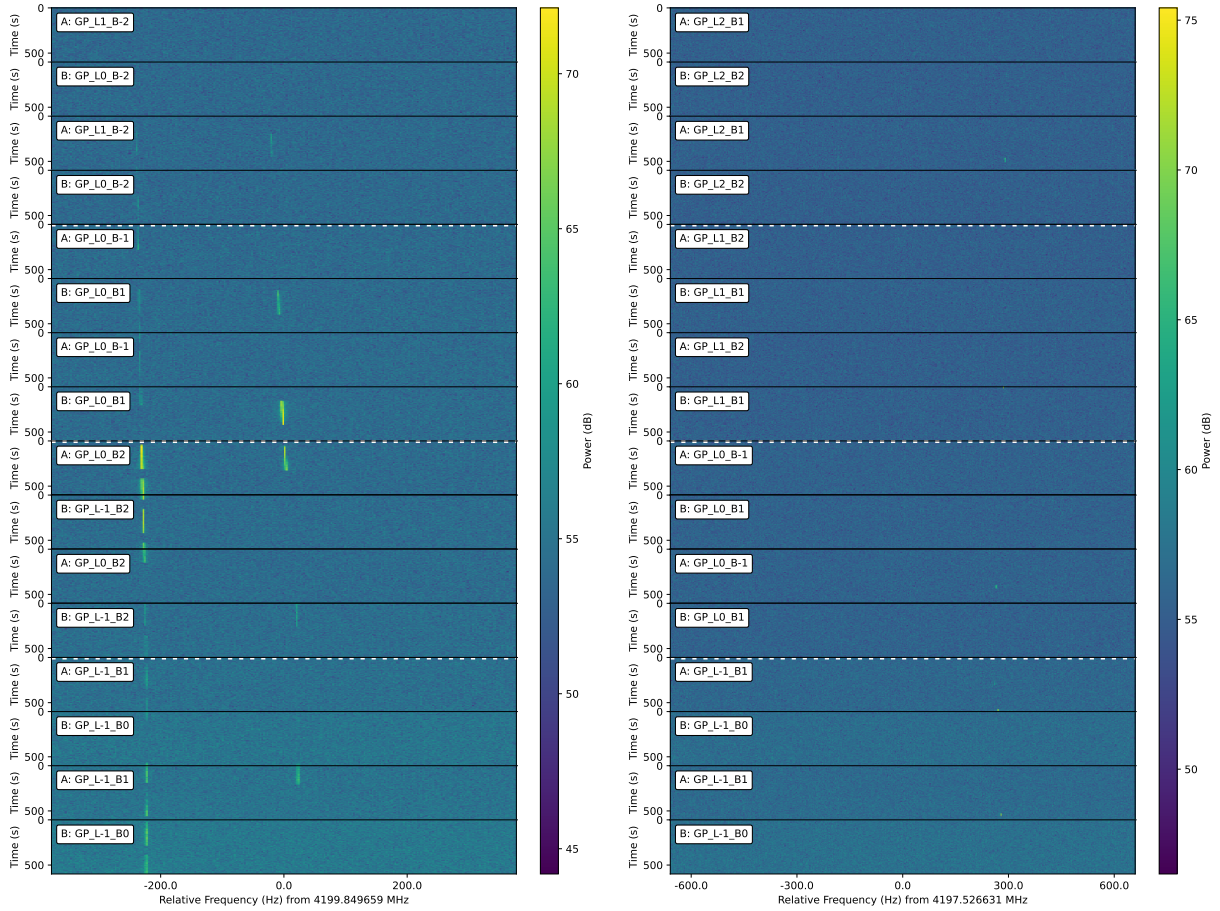


(a) Distribution of hit and event drift rates

(b) Distribution of hit and event S/N ratios

Figure 5.16: Histograms of drift rates and S/N ratios of detected hits in each NGP observation.

5.10 Appendix C: Candidate Vetting



(a) (b)

Figure 5.17: Extended plots of adjacent cadences for the highest quality candidates that passed the directional filter.

Chapter 6

Conclusions and Future Directions

The technological explosion enjoyed by SETI over the past few decades must be accompanied by a similar explosion in algorithmic development. We now have access to quite literally petabytes of radio spectrogram data and even that represents a tiny fraction of the potential parameter space actively considered across SETI. It is equally important to decide *where* to look in time, frequency, and physical space, as it is to decide *how* to look within the data that we can already take. In the case of ISM scintillation, as considered in Chapters 4–5, these two aspects of SETI are intertwined. Of course, new theoretical methods are only useful if they thrive in the wild – in other words, when handling the RFI environment. Regardless, the development of new signal or candidate filters may open up the possibility of detecting new types of signals, such as non-repeated one-offs. Having new filtering techniques at our disposal can also result in the design of new targeted surveys that might hitherto be unjustifiable. For example, conducting a survey across the Galactic plane near the Galactic center as we did in Chapter 5 might be unusual if we did not identify the region as particularly amenable to C-band searches for ISM-scintillated technosignatures.

In this work, we divided the "how" into two main categories: raw signal detection and candidate identification. Both include the use of various filters built on different signal attributes. Improvements in signal detection pipelines are important so that we do not miss technosignatures, even in the midst of dense RFI. A variety of new programs and algorithms are being used, from matched filters to neural networks, throughout the field of radio SETI. New implementations, such as `hyperseti`[1] and `seticore`[2], aim to standardize established search procedures. Others, such as that used in Margot et al. 2021, find many more signals within a given frequency range by carefully masking the bandwidth of already-detected signals. Machine learning techniques attack this from multiple angles, from localization (Chapter 2) to energy detection (Ma et al. 2023). As new methods come about, it is important to measure their effectiveness on verified datasets. In the absence of high-quality human-labeled data, we can produce controlled datasets with injected synthetic signals using

[1] https://github.com/UCBerkeleySETI/hyperseti
[2] https://github.com/lacker/seticore

`setigen` (Chapter 3), allowing us to evaluate the efficacy of our algorithms before we use them on real observational data.

As for candidate filtering in radio SETI, the main challenge is contending with RFI contamination within radio observations. We do not know for sure how technosignatures may be artificially modulated at the source, if at all. At the same time, however, we as researchers are not necessarily privy to the details of the artificial modulation amongst RFI, even though they originate from humans. While common general modulation schemes are public knowledge, originating sources and the details of their morphologies are not (at least not universally). At least in radio SETI, studies have typically avoided engaging with any specifics regarding actual signal morphology, instead choosing to bypass detailed morphology or rely on neural networks to search for outliers by modeling signals without much domain-specific input. It stands to reason that having a better understanding of our immediate radio environment should result in the development of better candidate filters. Most studies analyzing RFI have approached it from the angle of "spectral occupancy," in which we identify the sheer density of detected signals and note that candidates within those regions are most likely attributable to interference. However, there is so much more information intrinsic to these signals – after all, sending information is generally the reason for the existence of most of these detected narrowband signals in the first place – than their location on the frequency spectrum. As part of our scintillation search in Chapter 5, we conducted a brief analysis on RFI detected at the GBT across observation epochs spread over a couple of months, but dedicated longitudinal studies focusing on anthropogenic RFI could be useful for both narrowband SETI and other radio astronomy studies.

There have been recent and ongoing efforts to categorize detected RFI based on morphology, largely using unsupervised machine learning to decide which signals look similar (or distinct) from one another. However, these efforts are largely self-contained and have not led to the development of general algorithms or accessible lookup databases for SETI researchers to use. The vagueness and sheer parameter space of ground truth RFI modulation is the root cause of this. While it is a difficult and open-ended problem, it is worth directly addressing this with efforts to gain intimate understanding of the RFI environments around our telescopes (as a function of frequency, time, and sky direction). In addition, we can expand `setigen` to support the generation of synthetic signals with known modulations described in the communications literature, as a reference point for signals detected in real observations.

Nevertheless, based on our own study in Chapter 5, it is possible that any search that includes an analysis of signal intensity characteristics likely cannot afford to forego additionally using the standard ON-OFF filter that is already ubiquitous in radio SETI. In other words, if there is enough of an overlap in parameter space between theoretical target technosignatures and existing RFI, we cannot have confidence that candidates are truly technosignatures on alternate filters alone. The ON-OFF directional filter is simple in construction, but remains one of the most robust filters that we currently have. However, it is understood throughout the SETI field that a bona fide technosignature detection must pass multiple high-quality filters to garner widespread legitimate confidence and excitement.

Studies that employ the directional filter still rely on manual inspection, and for millions of raw hits, there are typically thousands of events that warrant manual inspection, most of which obviously should not have passed the algorithmic filters. So, there is still plenty of runway for developing better algorithms to identify candidates with limited false positives, which would reduce the human capital required to manually vet obvious false flags and allow us to push the boundaries of detection towards lower S/N ratios.

Although this thesis focused on narrowband technosignatures, it is entirely possible that ETI are sending or passively emitting broadband signals. Since such signals occupy a wide swath of bandwidth, the emission is typically relatively short in time, whether they are pulsed or not, which makes detection more difficult than for continuous-wave narrowband signals. Despite this, broadband represents an entire parameter space of signals that have not been typically targeted in technosignature searches. Furthermore, in our narrowband searches, we regularly pick up broadband signals with narrowband features. In those cases, they are false positives, so by modeling broadband signals or otherwise developing a way to reliably classify them, we can simultaneously improve the precision of narrowband searches. Once again, `setigen` can be extended to support the injection of broadband signals to facilitate the development of such classification techniques.

Radio SETI strives to answer a fundamental question about our existence in the universe, a question that very well may be unknowable. The field is centered on attempting to quantize that which is extremely difficult to quantize and rigorize that which is extremely difficult to rigorize. However, in the face of such a profound question, the only way to truly know is to try. Although SETI searches are done for the sake of SETI, the science that we do begets progress in other areas of astrophysics, such as pulsars and fast radio bursts. Likewise, innovation in radio SETI requires the understanding and adaptation of ideas throughout both signal processing and astrophysics. Every year, radio technosignature searches push boundaries in signal processing, detection, and classification. While we have still barely scratched the surface of the search parameter space, it is no meaningless platitude to say that we are closer than ever before.

Bibliography

Acharyya, A., Adams, C., Archer, A., et al. (2023). "A VERITAS/Breakthrough Listen Search for Optical Technosignatures". *The Astronomical Journal* 166.3, p. 84.

Barinova, O., Lempitsky, V., and Kholi, P. (2012). "On detection of multiple object instances using hough transforms". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34.9, pp. 1773–1784.

Bellanger, M., Bonnerot, G., and Coudreuse, M. (1976). "Digital filtering by polyphase network: Application to sample-rate alteration and filter banks". *IEEE Transactions on Acoustics, Speech, and Signal Processing* 24.2, pp. 109–114.

Blackman, R. B. and Tukey, J. W. (1958). "The measurement of power spectra from the point of view of communications engineering—Part II". *Bell System Technical Journal* 37.2, pp. 485–569.

Breiman, L. (2001). "Random forests". *Machine learning* 45, pp. 5–32.

Brzycki, B., Siemion, A. P., Croft, S., et al. (2020). "Narrow-band signal localization for SETI on noisy synthetic spectrogram data". *Publications of the Astronomical Society of the Pacific* 132.1017, p. 114501.

Brzycki, B., Siemion, A. P., Pater, I. de, et al. (2022). "Setigen: Simulating Radio Technosignatures for the Search for Extraterrestrial Intelligence". *The Astronomical Journal* 163.5, p. 222.

Brzycki, B., Siemion, A. P., Pater, I. de, et al. (2023). "On Detecting Interstellar Scintillation in Narrowband Radio SETI". *The Astrophysical Journal* 952.1, p. 46.

Cario, M. C. and Nelson, B. L. (1996). "Autoregressive to anything: Time-series input processes for simulation". *Operations Research Letters* 19.2, pp. 51–58.

Carroll, B. and Ostlie, D. (2007). *An Introduction to Modern aAtrophysics*. Addison-Wesley Reading, MA, USA.

Chollet, F. et al. (2015). *Keras*. https://keras.io.

Choza, C., Bautista, D., Croft, S., et al. (2023). "The Breakthrough Listen Search for Intelligent Life: Technosignature Search of 97 Nearby Galaxies". *The Astronomical Journal* 167.1, p. 10.

Cocconi, G. and Morrison, P. (1959). "Searching for Interstellar Communications". *Nature* 184.4, pp. 844–846. DOI: 10.1038/184844a0.

Coles, W. and Filice, J. (1984). "Dynamic spectra of interplanetary scintillations". *Nature* 312.5991, pp. 251–254.

Coles, W. A., Rickett, B. J., Gao, J., et al. (2010). "Scattering of pulsar radio emission by the interstellar plasma". *The Astrophysical Journal* 717.2, p. 1206.

Cordes, J. M. and Lazio, T. J. (1991). "Interstellar scattering effects on the detection of narrow-band signals". *The Astrophysical Journal* 376, pp. 123–133.

Cordes, J. M. and Lazio, T. J. W. (2002). "NE2001. I. A new model for the galactic distribution of free electrons and its fluctuations". *arXiv preprint astro-ph/0207156*.

Cordes, J. M., Lazio, T. J. W., and Sagan, C. (1997). "Scintillation-induced Intermittency in SETI". *The Astrophysical Journal* 487.2, p. 782.

Cordes, J. (1986). "Space velocities of radio pulsars from interstellar scintillations". *The Astrophysical Journal* 311, pp. 183–196.

Cordes, J. and Rickett, B. (1998). "Diffractive interstellar scintillation timescales and velocities". *The Astrophysical Journal* 507.2, p. 846.

Czech, D., Isaacson, H., Pearce, L., et al. (2021). "The Breakthrough Listen Search for Intelligent Life: MeerKAT Target Selection". *PASP* 133.1024, 064502, p. 064502. DOI: 10.1088/1538-3873/abf329.

Dai, B., Fidler, S., Urtasun, R., et al. (2017). "Towards diverse and natural image descriptions via a conditional gan". *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2970–2979.

Deller, A., Goss, W., Brisken, W., et al. (2019). "Microarcsecond VLBI pulsar astrometry with PSRπ II. Parallax distances for 57 pulsars". *The Astrophysical Journal* 875.2, p. 100.

Drake, F. D. (1961). "Project Ozma". *Physics Today* 14, pp. 40–46.

Drake, F. (1984). *SETI Science Working Group Report.*

DuPlain, R., Ransom, S., Demorest, P., et al. (2008). "Launching guppi: the green bank ultimate pulsar processing instrument". *Advanced Software and Control for Astronomy II*. Vol. 7019. International Society for Optics and Photonics, p. 70191D.

Enriquez, E. and Price, D. (2019). "turboSETI: Python-based SETI search algorithm". *ascl*, ascl–1906.

Enriquez, J. E., Siemion, A., Foster, G., et al. (2017). "The Breakthrough Listen Search for Intelligent Life: 1.1-1.9 GHz Observations of 692 Nearby Stars". *ApJ* 849, 104, p. 104. DOI: 10.3847/1538-4357/aa8d1b.

Fridman, P. (2011). "SETI: The transmission rate of radio communication and the signal's detection". *Acta Astronautica* 69.9-10, pp. 777–787.

Gajjar, V., LeDuc, D., Chen, J., et al. (2022). "Searching for broadband pulsed beacons from 1883 stars using neural networks". *The Astrophysical Journal* 932.2, p. 81.

Gajjar, V., Perez, K. I., Siemion, A. P. V., et al. (2021). "The Breakthrough Listen Search For Intelligent Life Near the Galactic Center. I." *AJ* 162.1, 33, p. 33. DOI: 10.3847/1538-3881/abfd36.

Gatys, L. A., Ecker, A. S., and Bethge, M. (2016). "Image style transfer using convolutional neural networks". *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2414–2423.

GBT Support Staff (2017). *The Proposer's Guide for the Green Bank Telescope.*

Giles, D. and Walkowicz, L. (2019). "Systematic serendipity: a test of unsupervised machine learning as a method for anomaly detection". *Monthly Notices of the Royal Astronomical Society* 484.1, pp. 834–849.

Goldstein, R. (1969). "Superior conjunction of Pioneer 6". *Science* 166.3905, pp. 598–601.

Goodman, J. W. (1975). "Laser speckle and related phenomena". *Statistical Properties of Laser Speckle Patterns* 9, pp. 9–75.

Gowanlock, M. G., Patton, D. R., and McConnell, S. M. (2011). "A model of habitability within the milky way galaxy". *Astrobiology* 11.9, pp. 855–873.

Gray, R. H. (2020). "The extended Kardashev scale". *The Astronomical Journal* 159.5, p. 228.

Gray, R. H. and Mooley, K. (2017). "A VLA search for radio signals from m31 and m33". *The Astronomical Journal* 153.3, p. 110.

Gupta, Y., Rickett, B. J., and Lyne, A. G. (1994). "Refractive interstellar scintillation in pulsar dynamic spectra". *Monthly Notices of the Royal Astronomical Society* 269.4, pp. 1035–1068.

Hamidouche, M. and Lestrade, J.-F. (2007). "Simulation of the interstellar scintillation and the extreme scattering events of pulsars". *Astronomy & Astrophysics* 468.1, pp. 193–203.

Harmon, J. and Coles, W. (1983). "Spectral broadening of planetary radar signals by the solar wind". *The Astrophysical Journal* 270, pp. 748–757.

Harp, G. R., Richards, J., Tarter, J. C., et al. (2016a). "SETI Observations of Exoplanets with the Allen Telescope Array". *AJ* 152, 181, p. 181. DOI: 10.3847/0004-6256/152/6/181.

Harp, G., Richards, J., Shostak, S., et al. (2019). "Machine Vision and Deep Learning for Classification of Radio SETI Signals". *arXiv preprint arXiv:1902.02426*.

Harp, G., Richards, J., Tarter, J. C., et al. (2016b). "SETI observations of exoplanets with the Allen Telescope Array". *arXiv preprint arXiv:1607.04207*.

Harris, C. R., Millman, K. J., Walt, S. J. van der, et al. (2020). "Array programming with NumPy". *Nature* 585.7825, pp. 357–362.

Harris, C. and Haines, K. (2011). "A mathematical review of polyphase filterbank implementations for radio astronomy". *Publications of the Astronomical Society of Australia* 28.4, pp. 317–322.

He, K., Zhang, X., Ren, S., et al. (2016). "Deep residual learning for image recognition". *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

Hewish, A., Bell, S., Pilkington, J., et al. (1968). "Observation of a Rapidly Pulsating Radio Source". *Nature* 217, pp. 709–713.

Hewish, A., Scott, P., and Wills, D. (1964). "Interplanetary scintillation of small diameter radio sources". *Nature* 203, pp. 1214–1217.

Hickish, J., Abdurashidova, Z., Ali, Z., et al. (2016). "A decade of developing radio-astronomy instrumentation using CASPER open-source technology". *Journal of Astronomical Instrumentation* 5.04, p. 1641001.

Horowitz, P., Matthews, B. S., Forster, J., et al. (1986). "Ultranarrowband searches for extraterrestrial intelligence with dedicated signal-processing hardware". *Icarus* 67.3, pp. 525–539.

Horowitz, P. and Sagan, C. (1993). "Five years of project META-an all-sky narrow-band radio search for extraterrestrial signals". *The Astrophysical Journal* 415, pp. 218–235.

Hough, P. V. (1959). "Machine analysis of bubble chamber pictures". *Proc. of the International Conference on High Energy Accelerators and Instrumentation, Sept. 1959*, pp. 554–556.

Ioffe, S. and Szegedy, C. (2015). "Batch normalization: Accelerating deep network training by reducing internal covariate shift". *arXiv preprint arXiv:1502.03167*.

Isaacson, H., Siemion, A. P. V., Marcy, G. W., et al. (2017). "The Breakthrough Listen Search for Intelligent Life: Target Selection of Nearby Stars and Galaxies". *PASA* 129.5, p. 054501. DOI: 10.1088/1538-3873/aa5800.

Jonas, J. L. (2009). "MeerKAT—The South African array with composite dishes and wideband single pixel feeds". *Proceedings of the IEEE* 97.8, pp. 1522–1530.

Kardashev, N. S. (1964). "Transmission of Information by Extraterrestrial Civilizations." *Soviet Astronomy, Vol. 8, p. 217* 8, p. 217.

Korpela, E., Werthimer, D., Anderson, D., et al. (2001). "SETI@ home—massively distributed computing for SETI". *Computing in science & engineering* 3.1, pp. 78–83.

Korpela, E. J., Anderson, D. P., Bankay, R., et al. (2011). "Status of the UC-Berkeley SETI efforts". *Instruments, Methods, and Missions for Astrobiology XIV*. Vol. 8152. International Society for Optics and Photonics, p. 815212.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "Imagenet classification with deep convolutional neural networks". *Advances in neural information processing systems*, pp. 1097–1105.

Lebofsky, M., Croft, S., Siemion, A. P. V., et al. (2019). "The Breakthrough Listen Search for Intelligent Life: Public Data, Formats, Reduction and Archiving". *arXiv e-prints*, arXiv:1906.07391, arXiv:1906.07391.

Lipman, D., Isaacson, H., Siemion, A. P. V., et al. (2019). "The Breakthrough Listen Search for Intelligent Life: Searching Boyajian's Star for Laser Line Emission". *Publications of the Astronomical Society of the Pacific* 131, p. 034202. DOI: 10.1088/1538-3873/aafe86.

Lotova, N., Blums, D., and Vladimirskii, K. (1985). "Interplanetary scintillation and the structure of the solar wind transonic region". *Astronomy and Astrophysics* 150, pp. 266–272.

Ma, P. X., Ng, C., Rizk, L., et al. (2023). "A deep-learning search for technosignatures from 820 nearby stars". *Nature Astronomy*, pp. 1–11.

Maan, Y., Leeuwen, J. van, and Vohl, D. (2021). "Fourier domain excision of periodic radio frequency interference". *Astronomy & Astrophysics* 650, A80.

MacMahon, D. H. E., Price, D. C., Lebofsky, M., et al. (2018). "The Breakthrough Listen Search for Intelligent Life: A Wideband Data Recorder System for the Robert C. Byrd Green Bank Telescope". *Publications of the Astronomical Society of the Pacific* 130.986, p. 044502. DOI: 10.1088/1538-3873/aa80d2.

Margot, J.-L., Greenberg, A. H., Pinchuk, P., et al. (2018). "A Search for Technosignatures from 14 Planetary Systems in the Kepler Field with the Green Bank Telescope at 1.15-1.73 GHz". *AJ* 155, 209, p. 209. DOI: 10.3847/1538-3881/aabb03.

Margot, J.-L., Pinchuk, P., Geil, R., et al. (2021). "A Search for Technosignatures around 31 Sun-like Stars with the Green Bank Telescope at 1.15–1.73 GHz". *The Astronomical Journal* 161.2, p. 55.

McDonough, R. N. and Whalen, A. D. (1995). *Detection of signals in noise*. Academic Press.

McKinney, W. et al. (2011). "pandas: a foundational Python library for data analysis and statistics". *Python for high performance and scientific computing* 14.9, pp. 1–9.

McMillan, P. J. (2016). "The mass distribution and gravitational potential of the Milky Way". *Monthly Notices of the Royal Astronomical Society*, stw2759.

Monari, J., Montebugnoli, S., Orlati, A., et al. (2006). "Generalized Hough transform: A useful algorithm for signal path detection". *Acta Astronautica* 58.4, pp. 230–235.

Narayan, R. (1992). "The physics of pulsar scintillation". *Philosophical Transactions of the Royal Society of London. Series A: Physical and Engineering Sciences* 341.1660, pp. 151–165.

Nita, G. M., Gary, D. E., Liu, Z., et al. (2007). "Radio frequency interference excision using spectral-domain statistics". *Publications of the Astronomical Society of the Pacific* 119.857, p. 805.

Nita, G. M., Hickish, J., MacMahon, D., et al. (2016). "EOVSA Implementation of a Spectral Kurtosis Correlator for Transient Detection and Classification". *Journal of Astronomical Instrumentation* 5.04, p. 1641009.

NRAO (2019). *GBT Proposer's Guide*.

O'Shea, T. J., Roy, T., and Clancy, T. C. (2018). "Over-the-air deep learning based radio signal classification". *IEEE Journal of Selected Topics in Signal Processing* 12.1, pp. 168–179.

Ocker, S. K. and Cordes, J. M. (2024). "NE2001p: A Native Python Implementation of the NE2001 Galactic Electron Density Model". *Research Notes of the AAS* 8.1, p. 17.

Ocker, S. K., Cordes, J. M., and Chatterjee, S. (2021). "Constraining Galaxy Halos from the Dispersion and Scattering of Fast Radio Bursts and Pulsars". *The Astrophysical Journal* 911.2, p. 102.

Okuta, R., Unno, Y., Nishino, D., et al. (2017). "CuPy: A NumPy-Compatible Library for NVIDIA GPU Calculations". *Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Thirty-first Annual Conference on Neural Information Processing Systems (NIPS)*.

Oliphant, T. E. (2006). *A guide to NumPy*. Vol. 1. Trelgol Publishing USA.

Oliver, B. M. and Billingham, J. (1971). "Project Cyclops: A design study of a system for detecting extraterrestrial intelligent life". *The 1971 NASA/ASEE Summer Fac. Fellowship Program (NASA-CR-114445*.

Parsons, A., Backer, D., Chang, C., et al. (2006). "PetaOp/Second FPGA Signal Processing for SETI and Radio Astronomy". *2006 Fortieth Asilomar Conference on Signals, Systems and Computers*, pp. 2031–2035. DOI: 10.1109/ACSSC.2006.355123.

Parsons, A., Backer, D., Siemion, A., et al. (2008). "A scalable correlator architecture based on modular FPGA hardware, reuseable gateware, and data packetization". *Publications of the Astronomical Society of the Pacific* 120.873, p. 1207.

Pence, W. D., Chiappetti, L., Page, C. G., et al. (2010). "Definition of the flexible image transport system (fits), version 3.0". *Astronomy & Astrophysics* 524, A42.

Pinchuk, P. and Margot, J.-L. (2022). "A machine learning–based direction-of-origin filter for the identification of radio frequency interference in the search for technosignatures". *The Astronomical Journal* 163.2, p. 76.

Pinchuk, P., Margot, J.-L., Greenberg, A. H., et al. (2019). "A search for technosignatures from TRAPPIST-1, LHS 1140, and 10 planetary systems in the Kepler field with the Green Bank Telescope at 1.15–1.73 GHz". *The Astronomical Journal* 157.3, p. 122.

Prestage, R. M., Bloss, M., Brandt, J., et al. (2015). "The versatile GBT astronomical spectrometer (VEGAS): Current status and future plans". *2015 USNC-URSI Radio Science Meeting (Joint with AP-S Symposium)*. IEEE, pp. 294–294. DOI: 10.1109/USNC-URSI.2015.7303578.

Price, D. C., Foster, G., Geyer, M., et al. (2019a). "A fast radio burst with frequency-dependent polarization detected during Breakthrough Listen observations". *MNRAS* 486.3, pp. 3636–3646. DOI: 10.1093/mnras/stz958.

Price, D., Enriquez, J., Chen, Y., et al. (2019b). "Blimpy: Breakthrough Listen I/O Methods for Python". *The Journal of Open Source Software* 4.42, 1554, p. 1554. DOI: 10.21105/joss.01554.

Price, D. C. (2021). "Spectrometers and polyphase filterbanks in radio astronomy". *The WSPC Handbook of Astronomical Instrumentation: Volume 1: Radio Astronomical Instrumentation*. World Scientific, pp. 159–179.

Price, D. C., Flynn, C., and Deller, A. (2021). "A comparison of Galactic electron density models using PyGEDM". *Publications of the Astronomical Society of Australia* 38, e038.

Price, D. C., Enriquez, J. E., Brzycki, B., et al. (2020). "The Breakthrough Listen Search for Intelligent Life: Observations of 1327 Nearby Stars Over 1.10–3.45 GHz". *AJ* 159.3, 86, p. 86. DOI: 10.3847/1538-3881/ab65f1.

Price, D. C., MacMahon, D. H. E., Lebofsky, M., et al. (2018). "The Breakthrough Listen search for intelligent life: Wide-bandwidth digital instrumentation for the CSIRO Parkes 64-m telescope". *Publications of the Astronomical Society of Australia* 35, p. 41. DOI: 10.1017/pasa.2018.36.

Radovan, M. V., Lanclos, K., Holden, B. P., et al. (2014). "The automated planet finder at Lick Observatory". *Ground-based and Airborne Telescopes V*. Vol. 9145. Proc. SPIE, 91452B, 91452B. DOI: 10.1117/12.2057310.

Rajwade, K., Lorimer, D., and Anderson, L. (2017). "Detecting pulsars in the Galactic Centre". *Monthly Notices of the Royal Astronomical Society* 471.1, pp. 730–739.

Ramachandran, R., Demorest, P., Backer, D., et al. (2006). "Interstellar plasma weather effects in long-term multifrequency timing of pulsar B1937+ 21". *The Astrophysical Journal* 645.1, p. 303.

Rampadarath, H., Morgan, J. S., Tingay, S. J., et al. (2012). "The First Very Long Baseline Interferometric SETI Experiment". *The Astronomical Journal* 144.2, p. 38. DOI: 10.1088/0004-6256/144/2/38.

Ravi, K. and Deshpande, A. A. (2018). "Scintillation-based Search for Off-pulse Radio Emission from Pulsars". *The Astrophysical Journal* 859.1, p. 22.

Reardon, D., Coles, W., Hobbs, G., et al. (2019). "Modelling annual and orbital variations in the scintillation of the relativistic binary PSR J1141- 6545". *Monthly Notices of the Royal Astronomical Society* 485.3, pp. 4389–4403.

Redmon, J., Divvala, S., Girshick, R., et al. (2016). "You only look once: Unified, real-time object detection". *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788.

Ren, S., He, K., Girshick, R., et al. (2015). "Faster r-cnn: Towards real-time object detection with region proposal networks". *Advances in neural information processing systems*, pp. 91–99.

Rickett, B. J. (1977). "Interstellar scattering and scintillation of radio waves". *Annual review of astronomy and astrophysics* 15.1, pp. 479–504.

Rickett, B. (1990). "Radio propagation through the turbulent interstellar plasma". *Annual review of astronomy and astrophysics* 28.1, pp. 561–605.

— (2007). "What do Scintillations tell us about the Ionized ISM?" *SINS-Small Ionized and Neutral Structures in the Diffuse Interstellar Medium*. Vol. 365, p. 207.

Rickett, B., Coles, W., Nava, C., et al. (2014). "Interstellar Scintillation of the Double Pulsar J0737- 3039". *The Astrophysical Journal* 787.2, p. 161.

Roberts, J. and Ables, J. (1982). "Dynamic spectra of pulsar scintillations at frequencies near 0.34, 0.41, 0.63, 1.4, 1.7, 3.2 and 5.0 GHz". *Monthly Notices of the Royal Astronomical Society* 201.4, pp. 1119–1138.

Ruf, C. S., Gross, S. M., and Misra, S. (2006). "RFI detection and mitigation for microwave radiometry with an agile digital detector". *IEEE transactions on geoscience and remote sensing* 44.3, pp. 694–706.

Scheuer, P. (1968). "Amplitude variations in pulsed radio sources". *Nature* 218.5145, pp. 920–922.

Sheikh, S. Z., Siemion, A., Enriquez, J. E., et al. (2020). "The Breakthrough Listen search for intelligent life: a 3.95–8.00 GHz search for radio technosignatures in the restricted earth transit zone". *The Astronomical Journal* 160.1, p. 29.

Sheikh, S. Z., Smith, S., Price, D. C., et al. (2021). "Analysis of the Breakthrough Listen signal of interest blc1 with a technosignature verification framework". *Nature Astronomy* 5.11, pp. 1153–1162.

Sheikh, S. Z., Wright, J. T., Siemion, A., et al. (2019). "Choosing a maximum drift rate in a SETI search: astrophysical considerations". *The Astrophysical Journal* 884.1, p. 14.

Shostak, S. (2000). "SETI merit and the galactic plane". *Acta Astronautica* 46.10-12, pp. 649–654.

Siemion, A. P., Benford, J., Cheng-Jin, J., et al. (2014). "Searching for extraterrestrial intelligence with the Square Kilometre Array". *arXiv preprint arXiv:1412.4867*.

Siemion, A. P., Demorest, P., Korpela, E., et al. (2013). "A 1.1-1.9 GHz SETI survey of the Kepler field. I. A Search for narrow-band emission from select targets". *The Astrophysical Journal* 767.1, p. 94.

Simonyan, K. and Zisserman, A. (2014). "Very deep convolutional networks for large-scale image recognition". *arXiv preprint arXiv:1409.1556*.

Smith, F. (1950). "Origin of the Fluctuations in the Intensity of Radio Waves from Galactic Sources: Cambridge Observations". *Nature* 165.4194, pp. 422–423.

Sokolowski, M., Wayth, R. B., and Lewis, M. (2015). "The statistics of low frequency radio interference at the Murchison Radio-astronomy Observatory". *2015 IEEE Global Electromagnetic Compatibility Conference (GEMCCON)*. IEEE, pp. 1–6.

Stone, R. P. S., Wright, S. A., Drake, F., et al. (2005). "Lick Observatory Optical SETI: Targeted Search and New Directions". *Astrobiology* 5, pp. 604–611. DOI: 10.1089/ast.2005.5.604.

Straten, W. van, Demorest, P., and Osłowski, S. (2012). "Pulsar data analysis with PSRCHIVE". *arXiv preprint arXiv:1205.6276*.

Suresh, A., Cordes, J. M., Chatterjee, S., et al. (2021). "4–8 GHz Spectrotemporal Emission from the Galactic Center Magnetar PSR J1745–2900". *The Astrophysical Journal* 921.2, p. 101.

Suresh, A., Gajjar, V., Nagarajan, P., et al. (2023). "A 4–8 GHz Galactic Center Search for Periodic Technosignatures". *The Astronomical Journal* 165.6, p. 255.

Szegedy, C., Liu, W., Jia, Y., et al. (2015). "Going deeper with convolutions". *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9.

Tao, Z.-Z., Zhao, H.-C., Zhang, T.-J., et al. (2022). "Sensitive Multibeam Targeted SETI Observations toward 33 Exoplanet Systems with FAST". *The Astronomical Journal* 164.4, p. 160.

Tarter, J., Ackermann, R., Barott, W., et al. (2011). "The first SETI observations with the Allen telescope array". *Acta Astronautica* 68, pp. 340–346.

Tarter, J. (2001). "The search for extraterrestrial intelligence (SETI)". *Annual Review of Astronomy and Astrophysics* 39.1, pp. 511–548.

Taylor, J. H. (1974). "A Sensitive Method for Detecting Dispersed Radio Emission". *AAPS* 15, p. 367.

Thompson, A. R., Moran, J. M., and Swenson George W., J. (2017). *Interferometry and Synthesis in Radio Astronomy, 3rd Edition*. DOI: 10.1007/978-3-319-44431-4.

Tingay, S. J., Kaplan, D. L., Lenc, E., et al. (2018a). "A Serendipitous MWA Search for Narrowband Signals from 'Oumuamua". *ApJ* 857, 11, p. 11. DOI: 10.3847/1538-4357/aab359.

Tingay, S., Tremblay, C., and Croft, S. (2018b). "A Search for Extraterrestrial Intelligence (SETI) toward the Galactic Anticenter with the Murchison Widefield Array". *The Astrophysical Journal* 856.1, p. 31.

Uno, Y., Hashimoto, T., Goto, T., et al. (2023). "Upper limits on transmitter rate of extragalactic civilizations placed by Breakthrough Listen observations". *Monthly Notices of the Royal Astronomical Society* 522.3, pp. 4649–4653.

Vogt, S. S., Radovan, M., Kibrick, R., et al. (2014). "APF–The Lick Observatory Automated Planet Finder". *PASP* 126, p. 359. DOI: 10.1086/676120.

Weekes, T., Badran, H., Biller, S., et al. (2002). "VERITAS: the very energetic radiation imaging telescope array system". *Astroparticle Physics* 17.2, pp. 221–243.

Welch, J., Backer, D., Blitz, L., et al. (2009). "The Allen Telescope Array: The first widefield, panchromatic, snapshot radio camera for radio astronomy and SETI". *Proceedings of the IEEE* 97.8, pp. 1438–1447.

Werthimer, D., Tarter, J., and Bowyer, S. (1985). "The Serendip II Design". *Symposium-International Astronomical Union*. Vol. 112. Cambridge University Press, pp. 421–424.

Woo, R. (2007). "Space weather and deep space communications". *Space Weather* 5.9.

Woo, R. and Armstrong, J. (1979). "Spacecraft radio scattering observations of the power spectrum of electron density fluctuations in the solar wind". *Journal of Geophysical Research: Space Physics* 84.A12, pp. 7288–7296.

Worden, S. P., Drew, J., Siemion, A., et al. (2017). "Breakthrough Listen: A new search for life in the universe". *Acta Astronautica* 139.Supplement C, pp. 98–101. DOI: https://doi.org/10.1016/j.actaastro.2017.06.008.

Wright, J. T., Kanodia, S., and Lubar, E. (2018). "How much SETI has been done? Finding needles in the N-dimensional cosmic haystack". *The Astronomical Journal* 156.6, p. 260.

Wright, S. A., Werthimer, D., Treffers, R. R., et al. (2014). "A near-infrared SETI experiment: instrument overview". *Ground-based and Airborne Instrumentation for Astronomy V*. Vol. 9147. International Society for Optics and Photonics, 91470J.

Yao, J., Manchester, R., and Wang, N. (2017). "A new electron-density model for estimation of pulsar and FRB distances". *The Astrophysical Journal* 835.1, p. 29.

Zhang, Y. G., Gajjar, V., Foster, G., et al. (2018a). "Fast radio burst 121102 pulse detection and periodicity: a machine learning approach". *The Astrophysical Journal* 866.2, p. 149.

Zhang, Y. G., Won, K. H., Son, S. W., et al. (2018b). "Self-supervised Anomaly Detection for Narrowband SETI". *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, pp. 1114–1118.

Zhang, Z.-S., Werthimer, D., Zhang, T.-J., et al. (2020). "First SETI Observations with China's Five-hundred-meter Aperture Spherical Radio Telescope (FAST)". *The Astrophysical Journal* 891.2, p. 174.