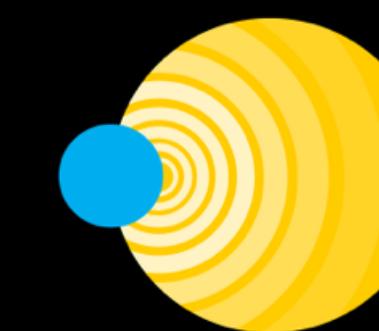


# BREAKTHROUGH LISTEN

## Algorithmic and Astrophysical Methods in the Search for Radio Technosignatures

**BRYAN BRZYCKI**  
**UNIVERSITY OF CALIFORNIA BERKELEY**  
**REU SEMINAR, JULY 3, 2024**

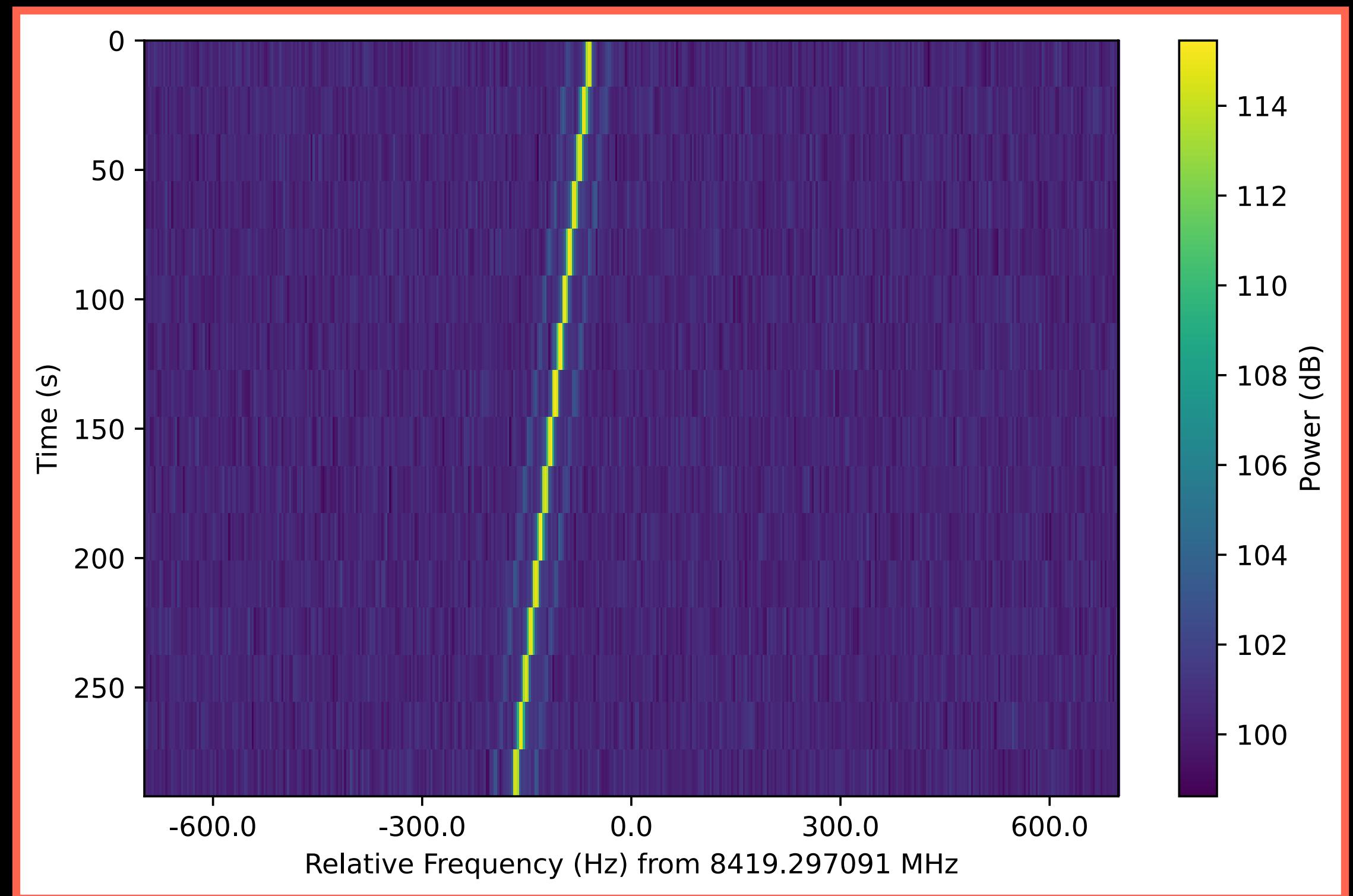


**BERKELEY SETI**  
RESEARCH CENTER



# Anatomy of a narrowband radio transmission

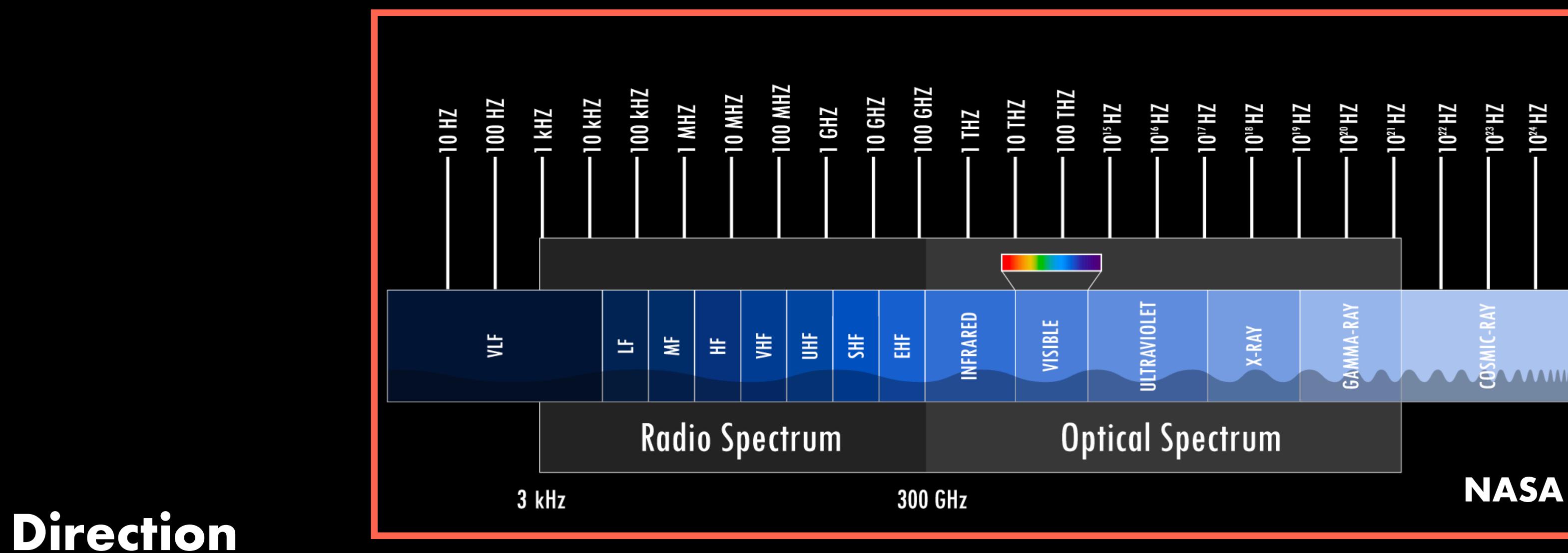
- Time-frequency spectrogram, or dynamic spectra
- High duty cycle, or continuous wave
- Doppler drift: changing Doppler shift from relative acceleration



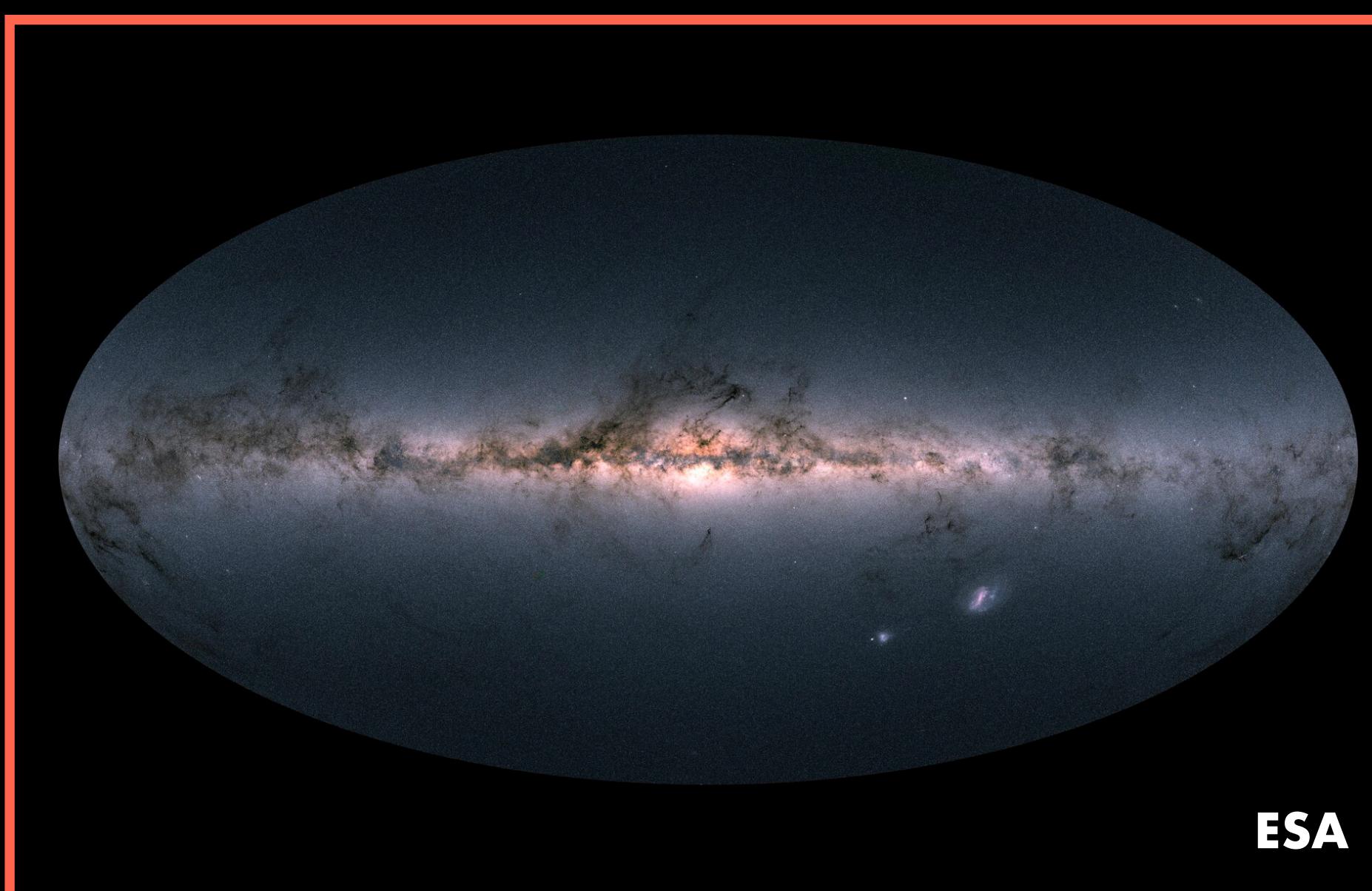
**Detected signal from Voyager 1**

# Where should we look?

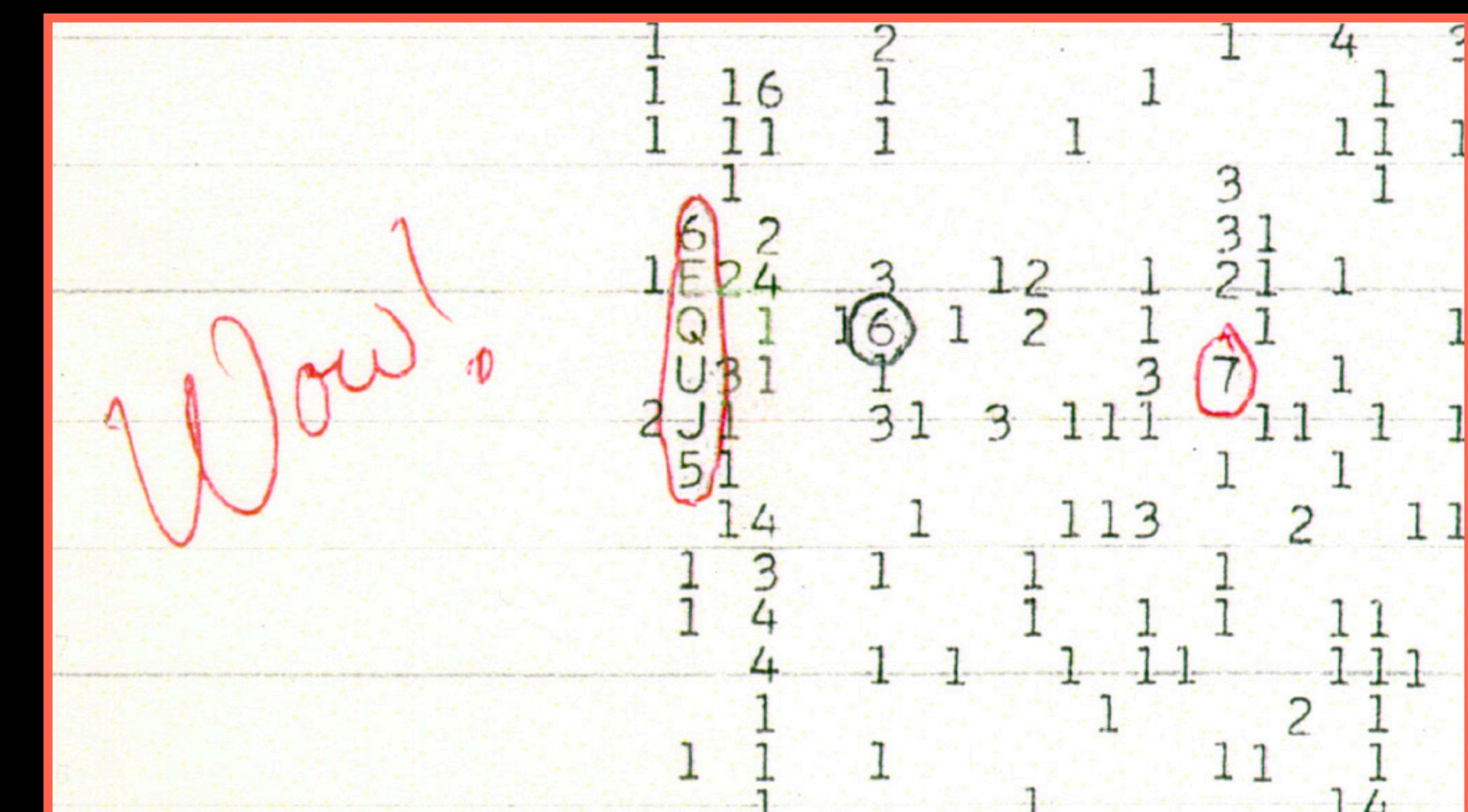
# Frequency



# Direction



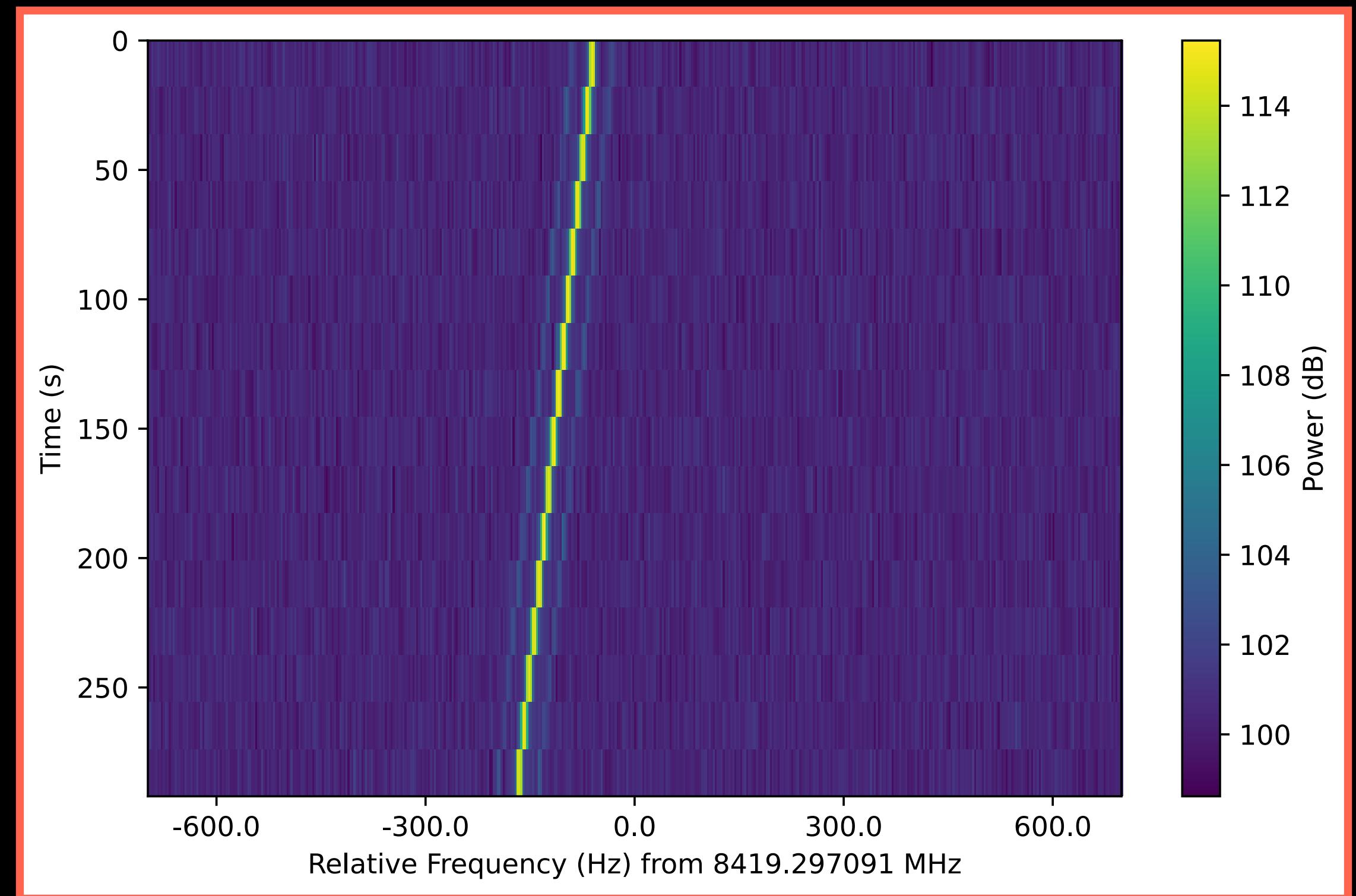
ESA



# Big Ear Radio

# Modern radio SETI in practice

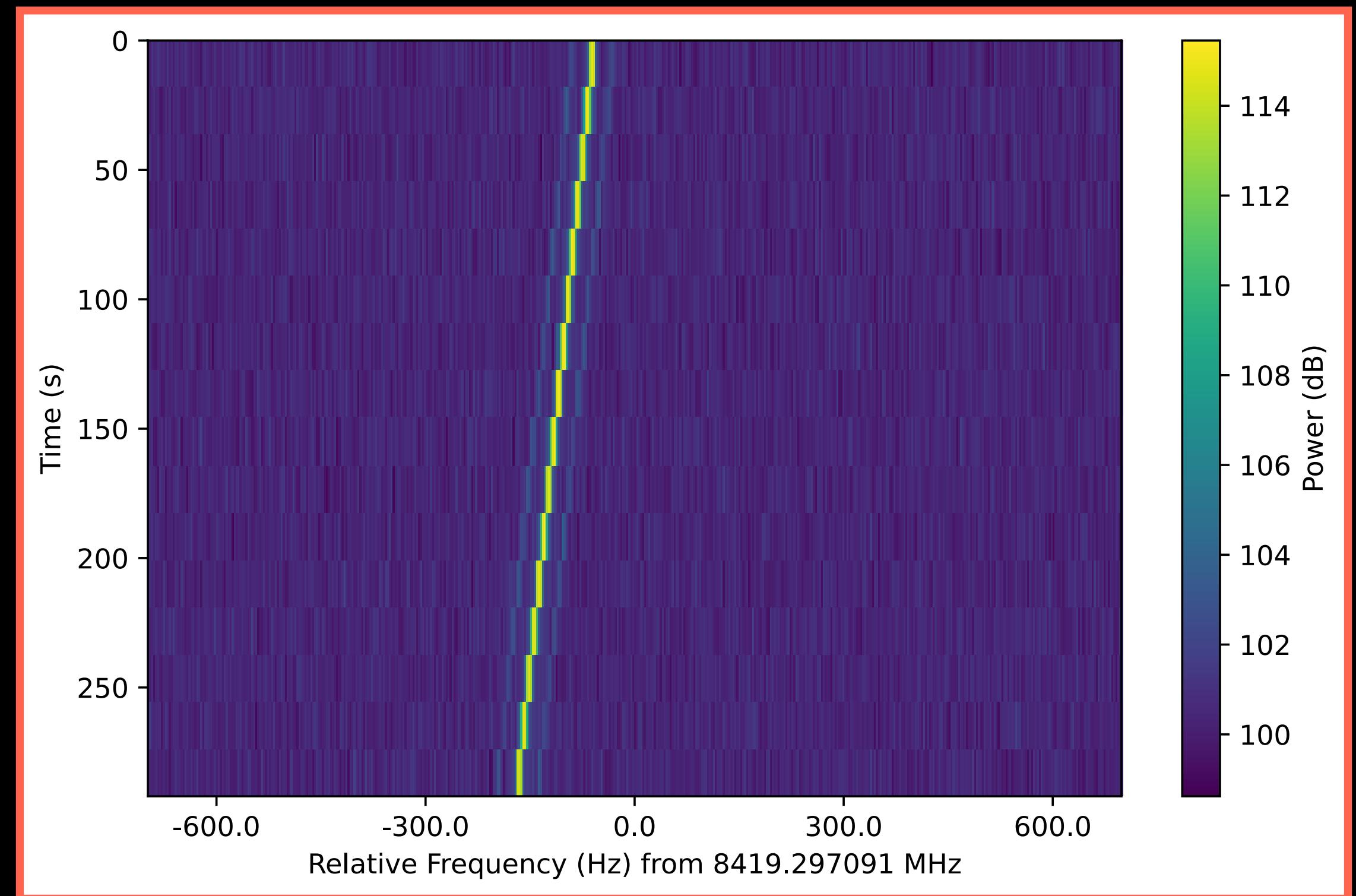
- Basic signal detection
  - Gather all signals possible above signal-to-noise threshold
- Candidate identification and differentiation
  - Primarily against human-created radio frequency interference (RFI)
- Lots of manual inspection



**Detected signal from Voyager 1**

# Modern radio SETI in practice

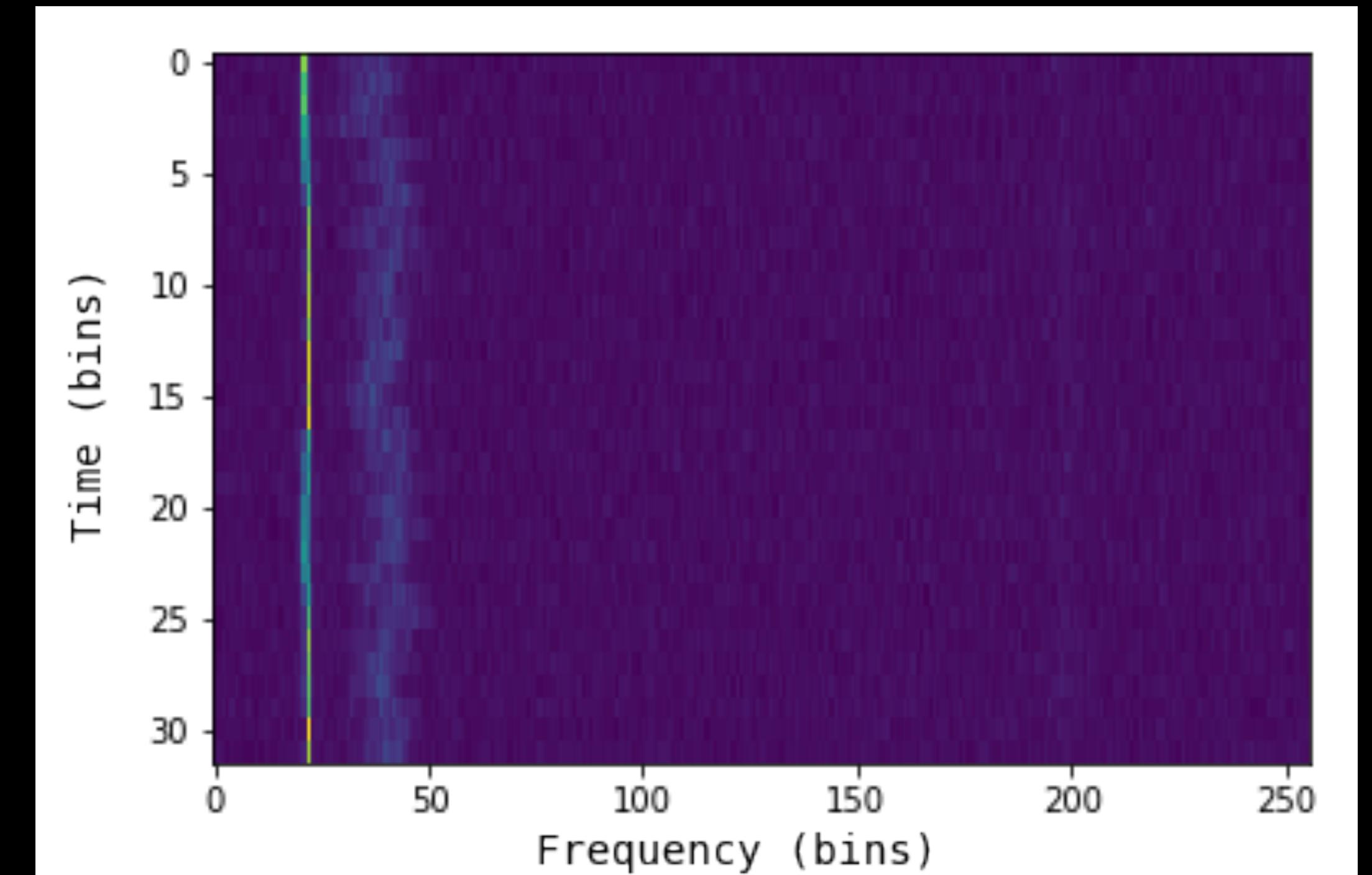
- Basic signal detection
  - Gather all signals possible above signal-to-noise threshold
- Candidate identification and differentiation
  - Primarily against human-created radio frequency interference (RFI)
- Lots of manual inspection



**Detected signal from Voyager 1**

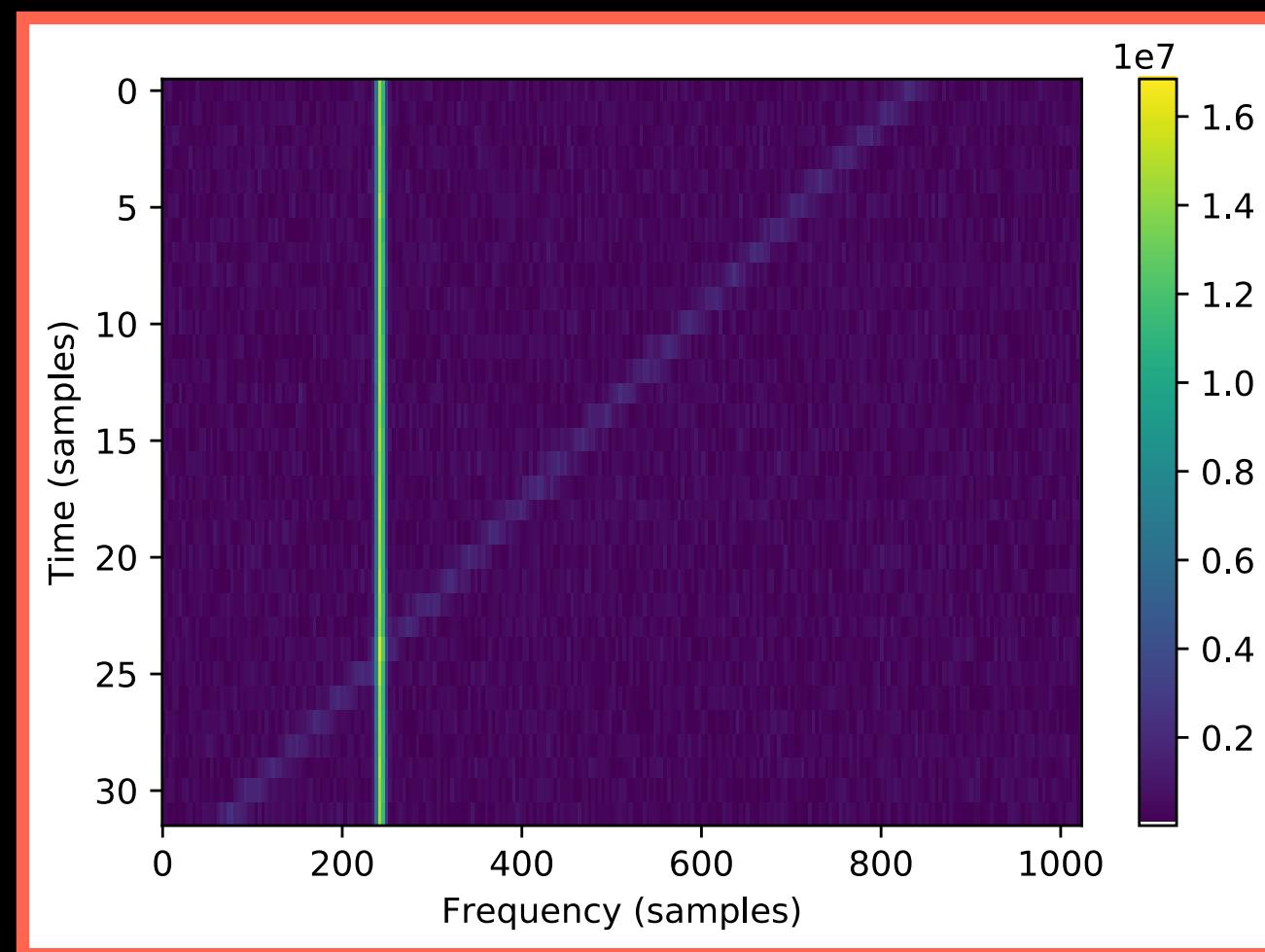
# Signal detection / localization with machine learning

- Standard integration-based pipeline:
  - Dim signals concealed by nearby bright signals
  - Computationally expensive to search high drift rates

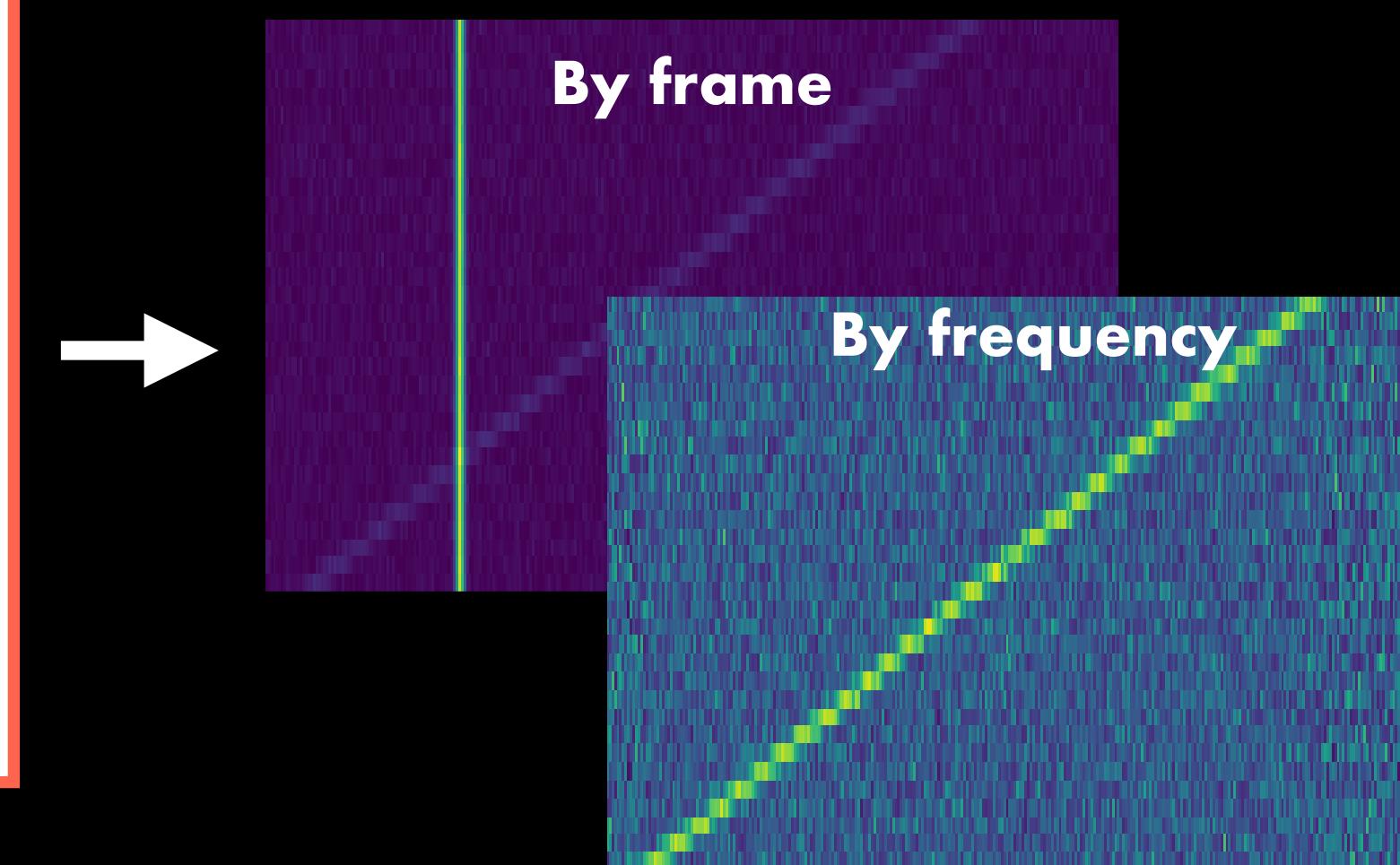


**Snippet of GBT data at C-band**

## Synthetic training data

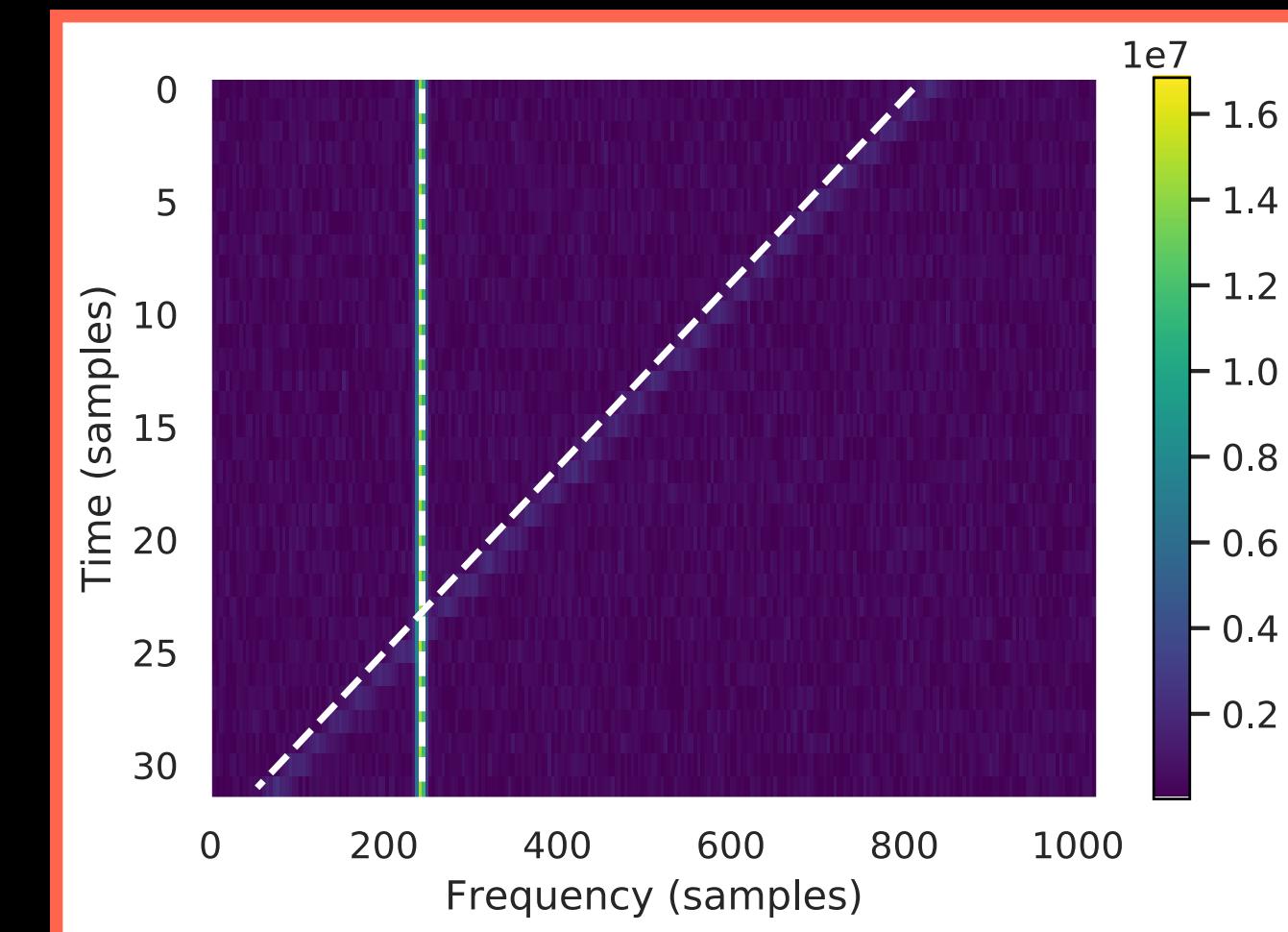


## Normalization



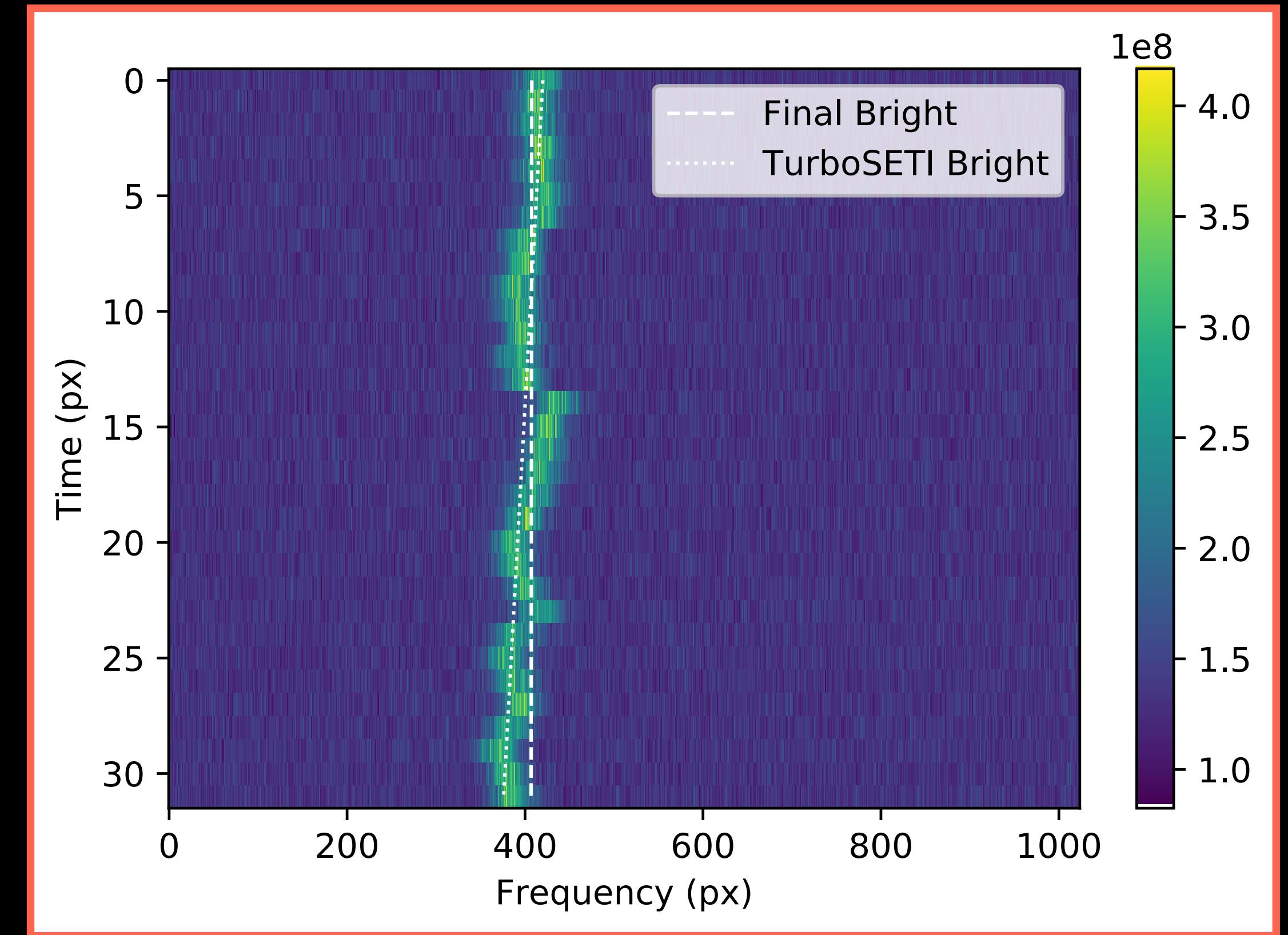
**Convolutional  
Neural Network**

## Predicted locations



# Takeaways

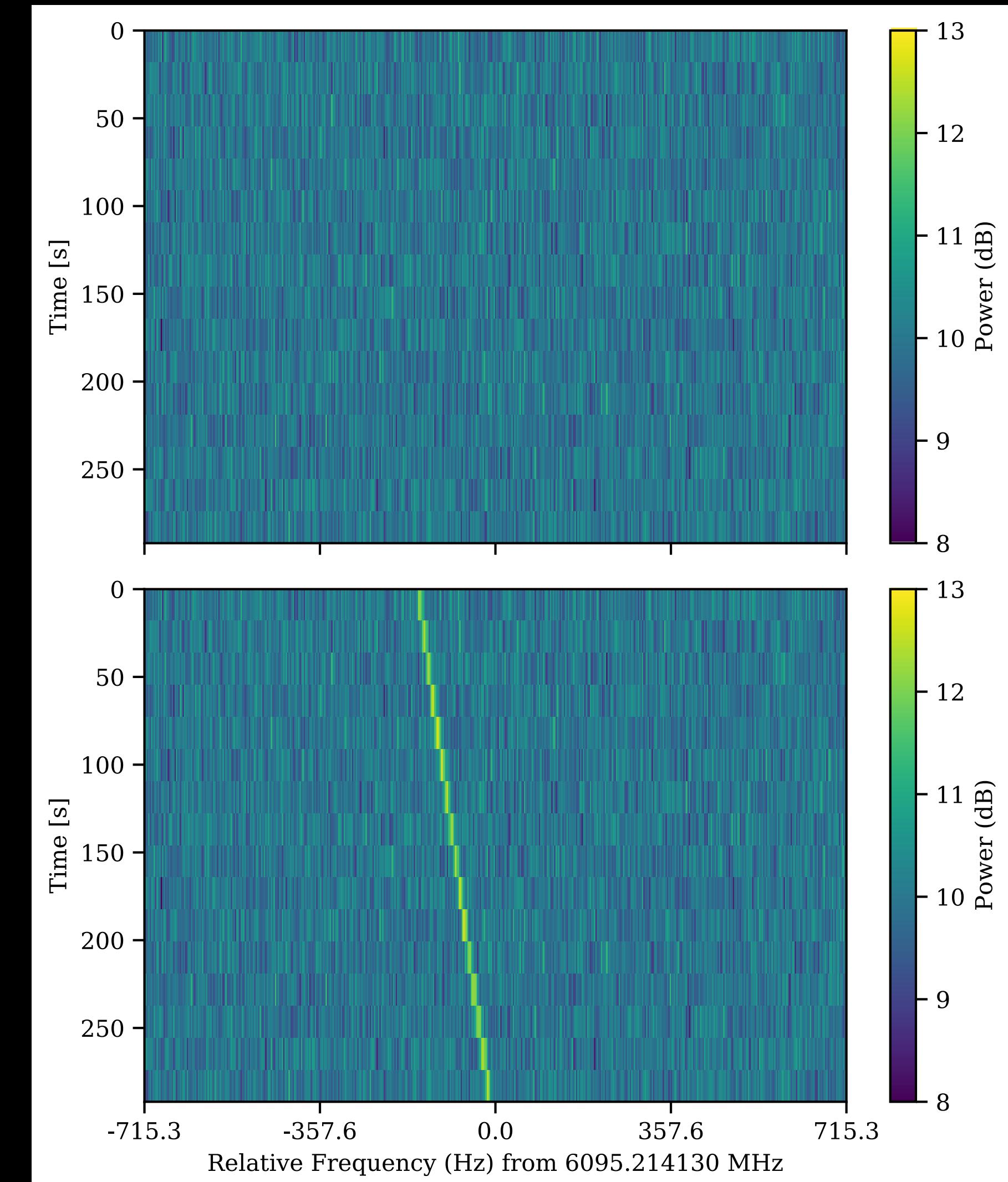
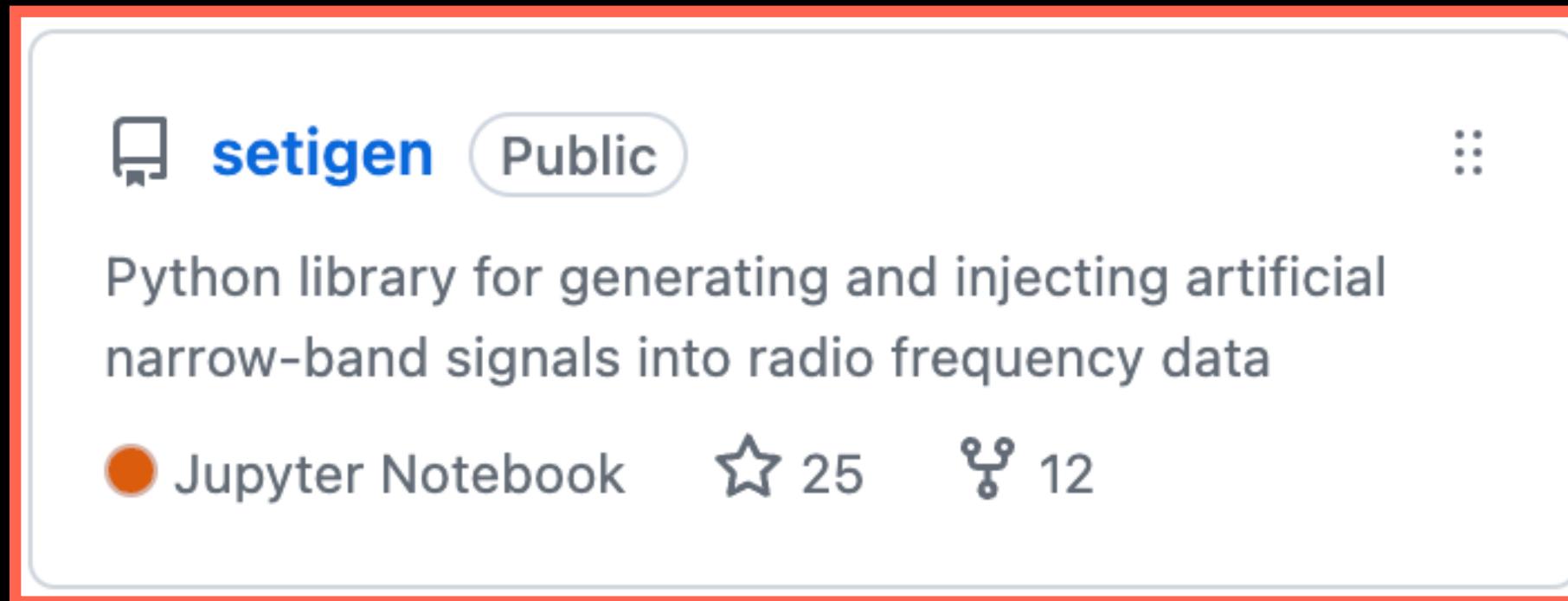
- Less accurate than deDoppler methods, but generally 20-40x faster
- Trained on ideal signals but still relatively robust



**C-band RFI signal, with ML prediction dashed and TurboSETI localization dotted.**

# Setigen

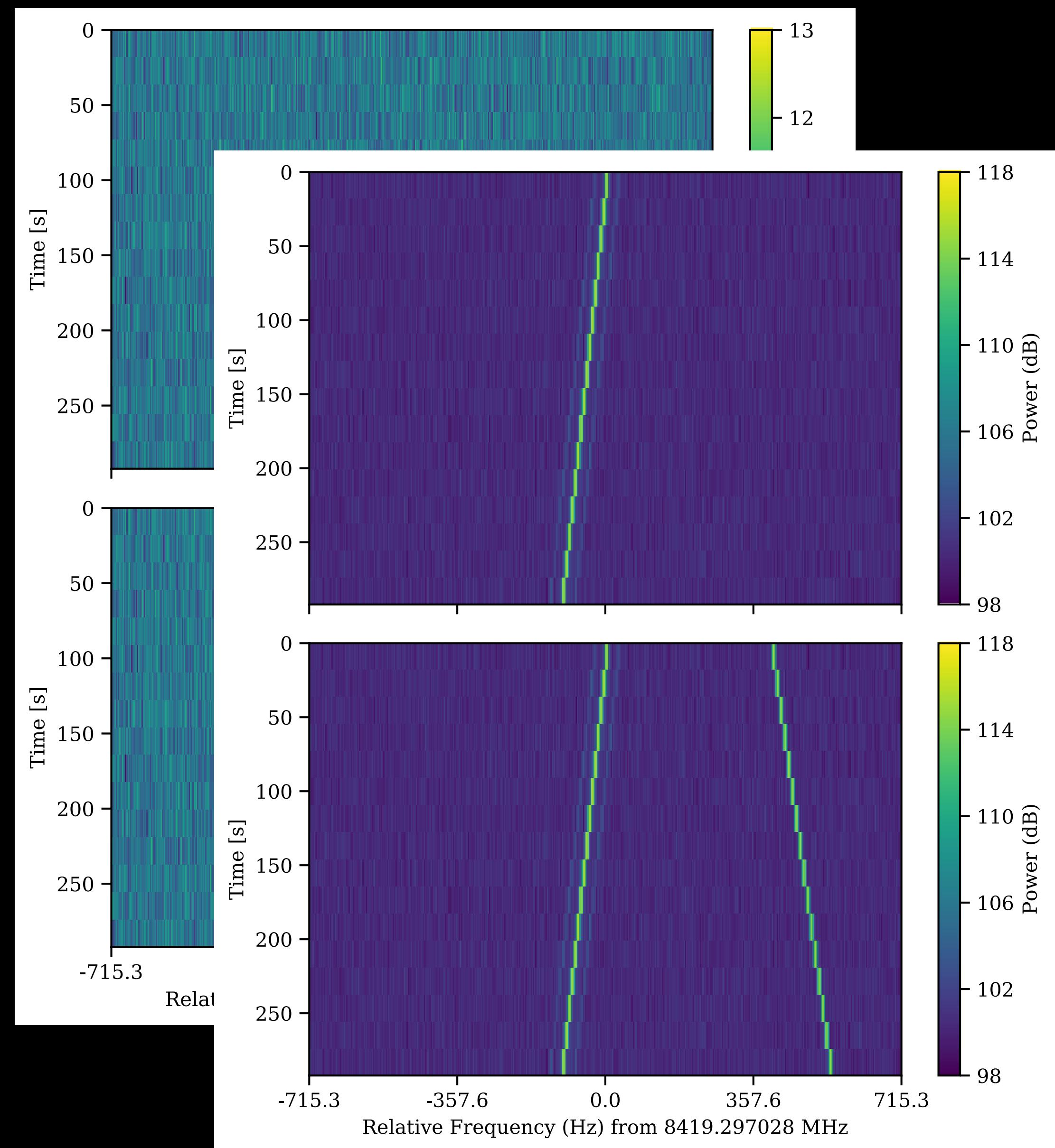
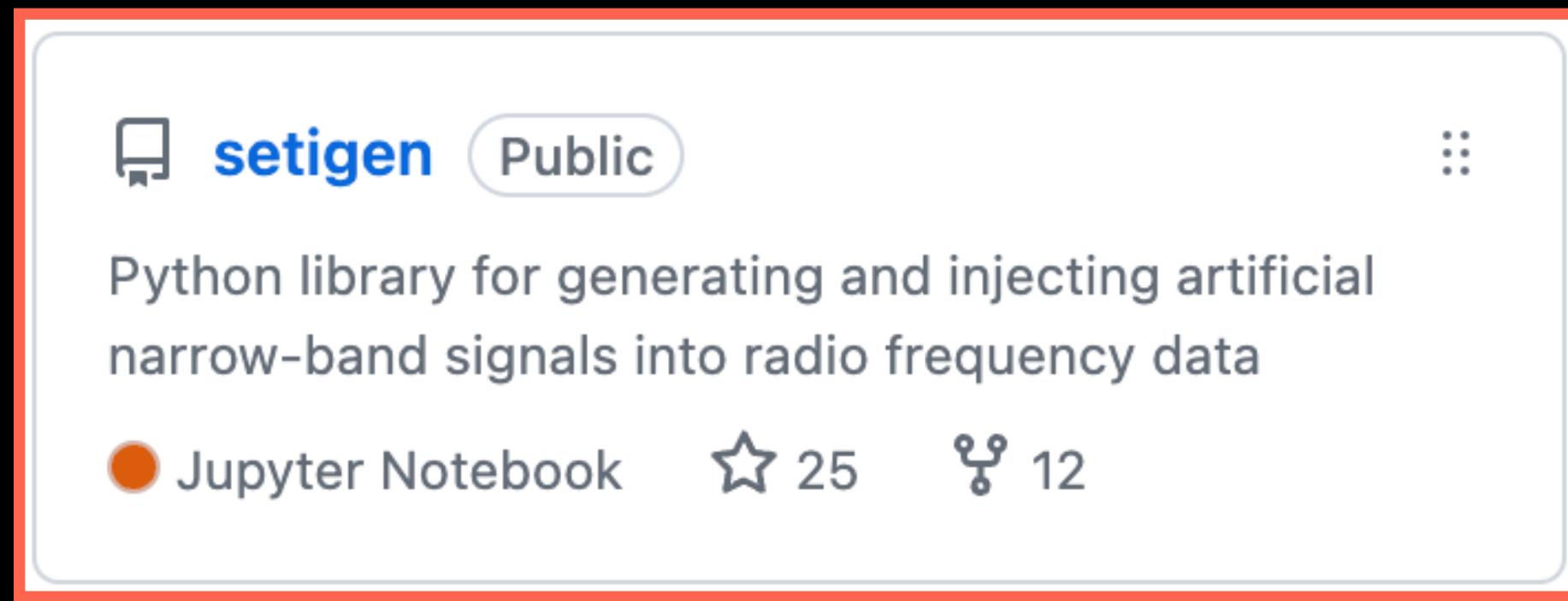
- Open-source Python library for creating synthetic spectrogram and voltage data
- Specific focus on narrowband signal generation and injection



Brzycki et al. 2022, AJ

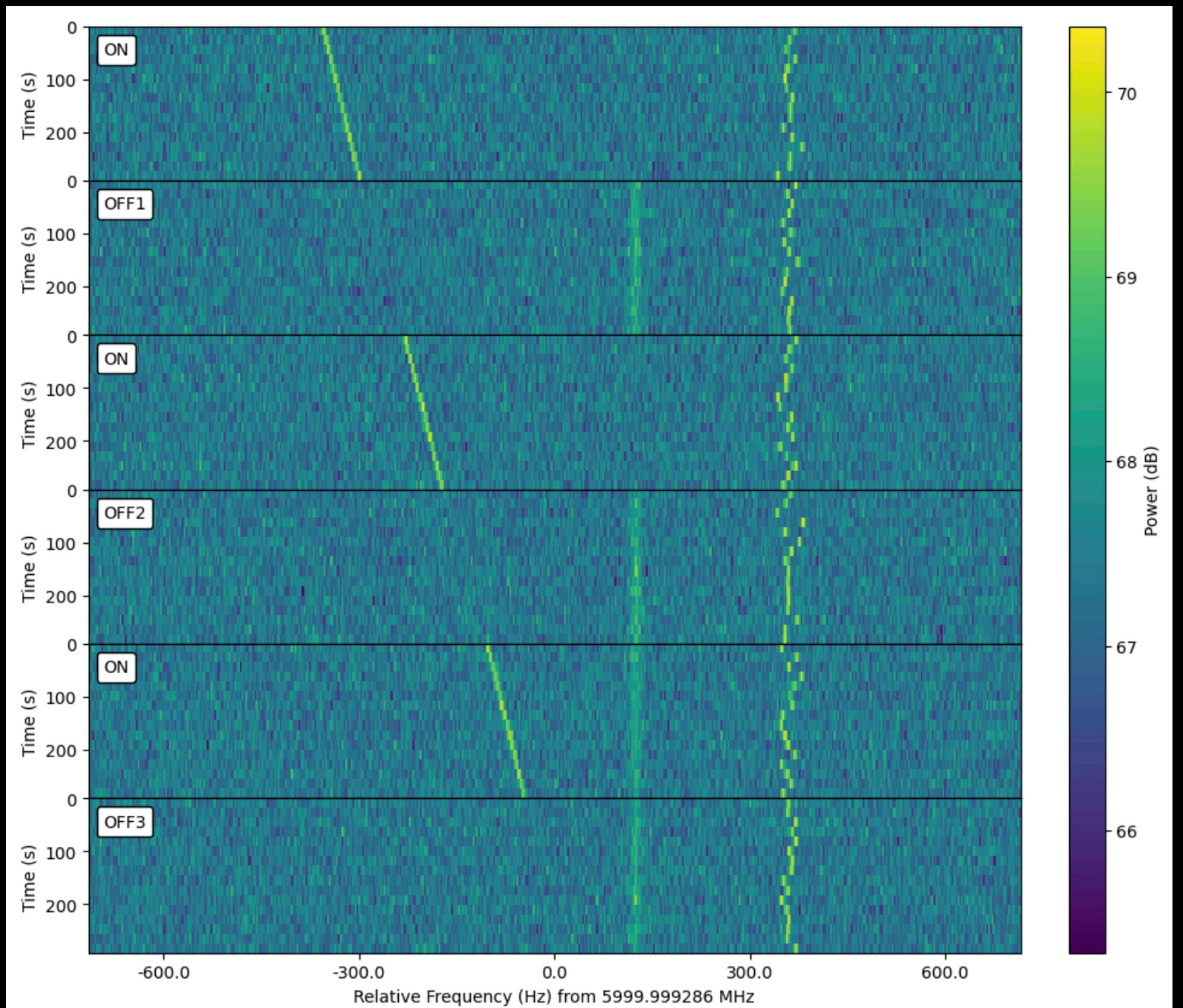
# Setigen

- Open-source Python library for creating synthetic spectrogram and voltage data
- Specific focus on narrowband signal generation and injection



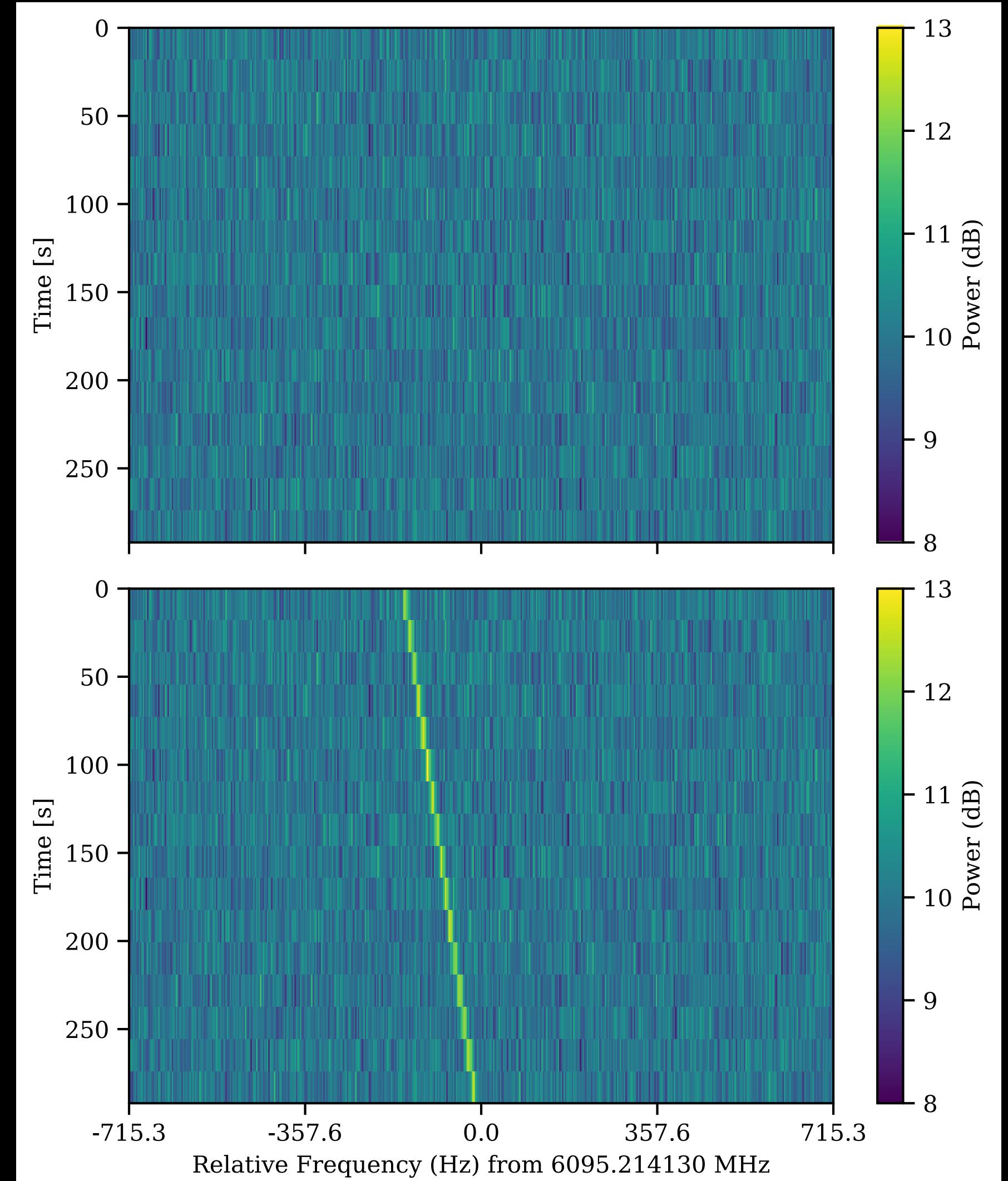
# Setigen

- Open-source = always something to contribute and improve!
- Project has expanded massively since inception:
  - Support for full observational cadences
  - Publication-ready figures
  - More advanced operations such as deDrifting, slicing, and integration-based Doppler-smearing
  - Voltage synthesis and basic frontend simulation (.raw complex voltages)



# Setigen's basic design philosophy

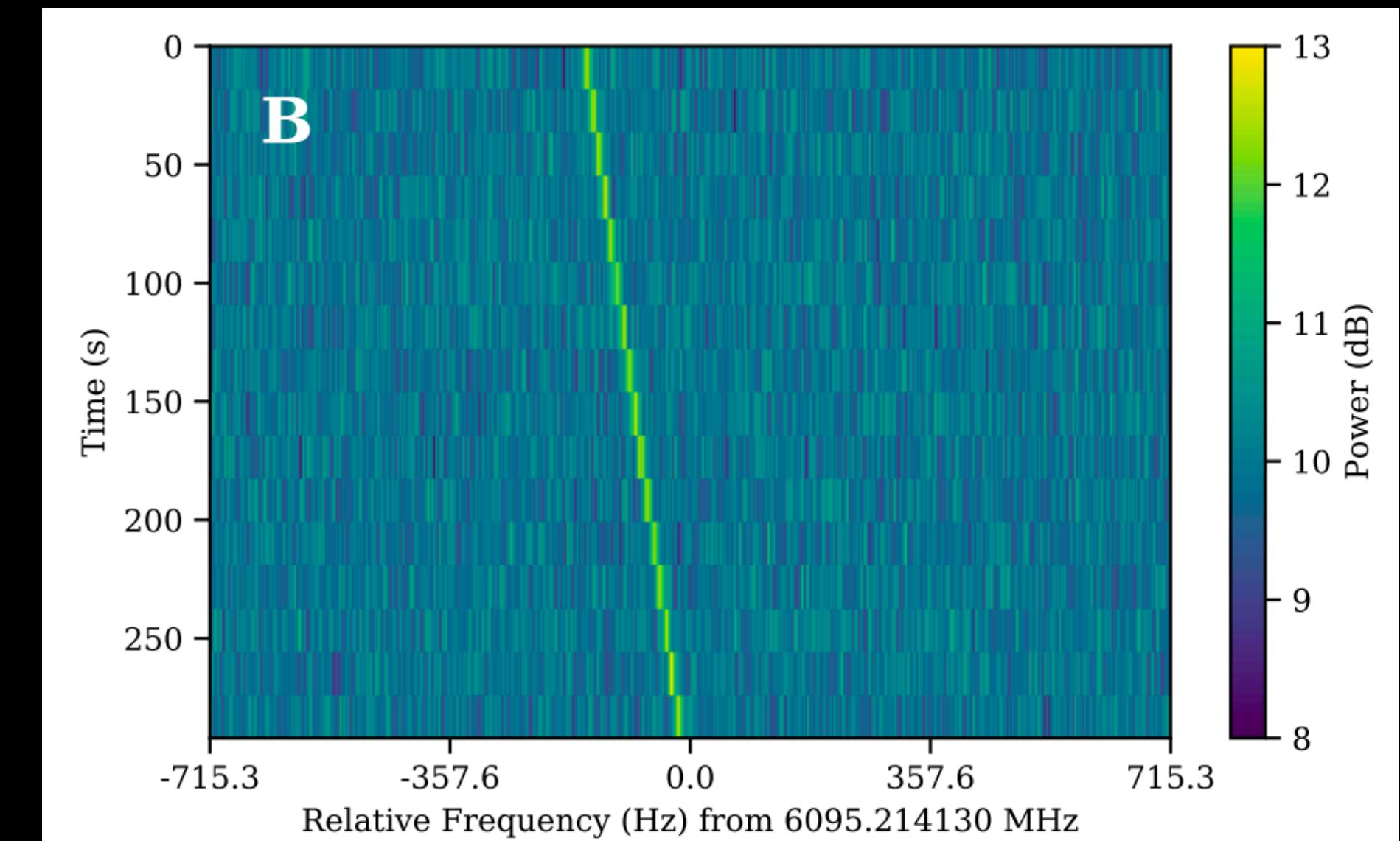
- The Frame class is used to represent a basic time-frequency intensity spectrogram
- Can load in .fil or .h5 data directly into Frames, or generate synthetic chi-squared noise
- Signal injection methods adds power on top of existing Stokes I data – this is not technically “correct”!
  - $(x^2 + y^2)$  vs.  $(x + y)^2$



# Setigen's basic spectrogram signal calculation

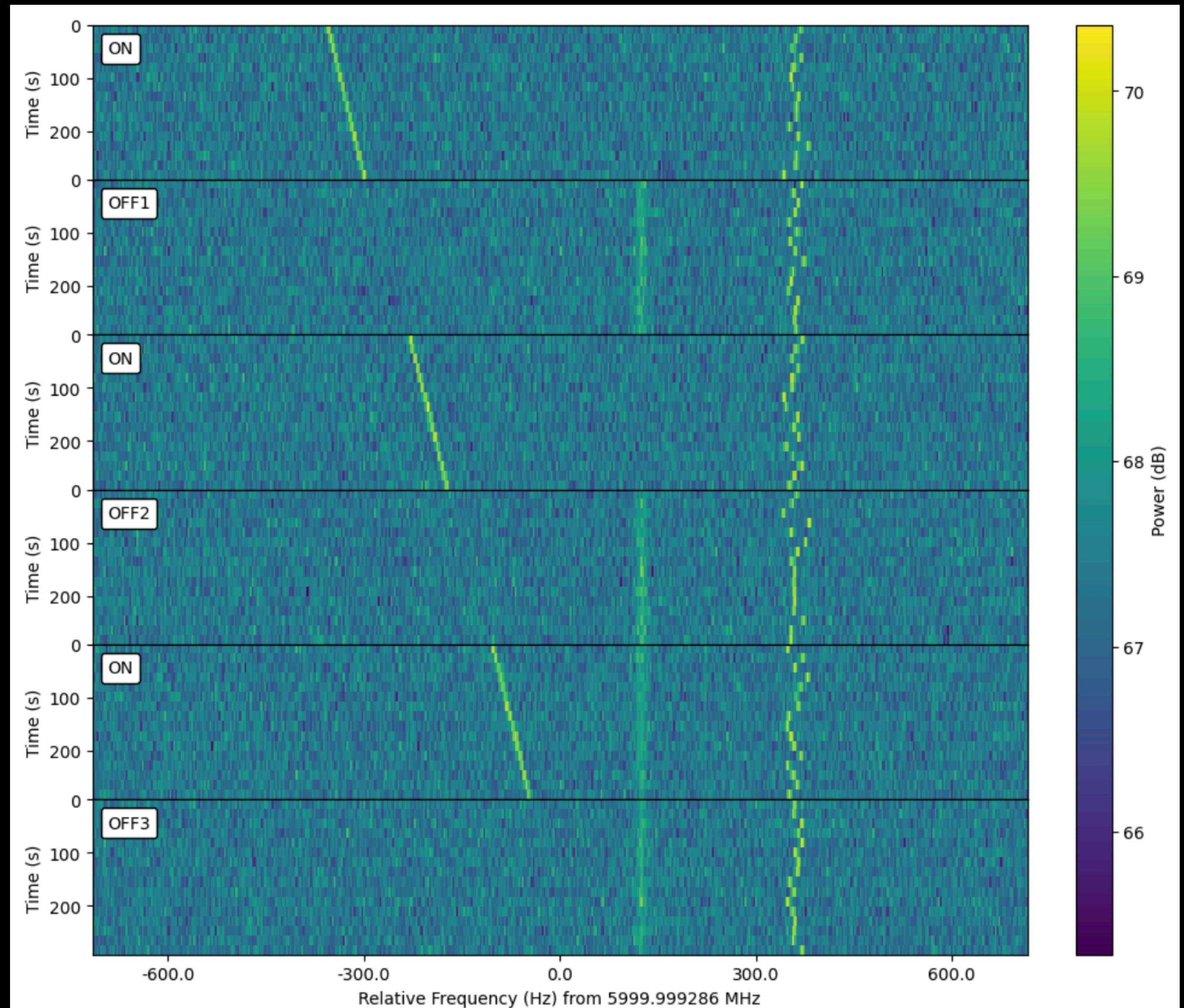
1. path— $I_p(t)$ : Central signal frequencies as a function of time, e.g., linear (constant) drift rate, quadratic drift rate
2. t\_profile— $I_t(t)$ : Signal intensity as a function of time, e.g., constant intensity, Gaussian pulses
3. f\_profile— $I_f(f, f_0)$ : Spectral profile as a function of frequency (offset from central frequency), e.g.,  $\text{sinc}^2$  profile, Gaussian profile
4. bp\_profile— $I_{bp}(f)$ : Bandpass profile as a function of absolute frequency

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



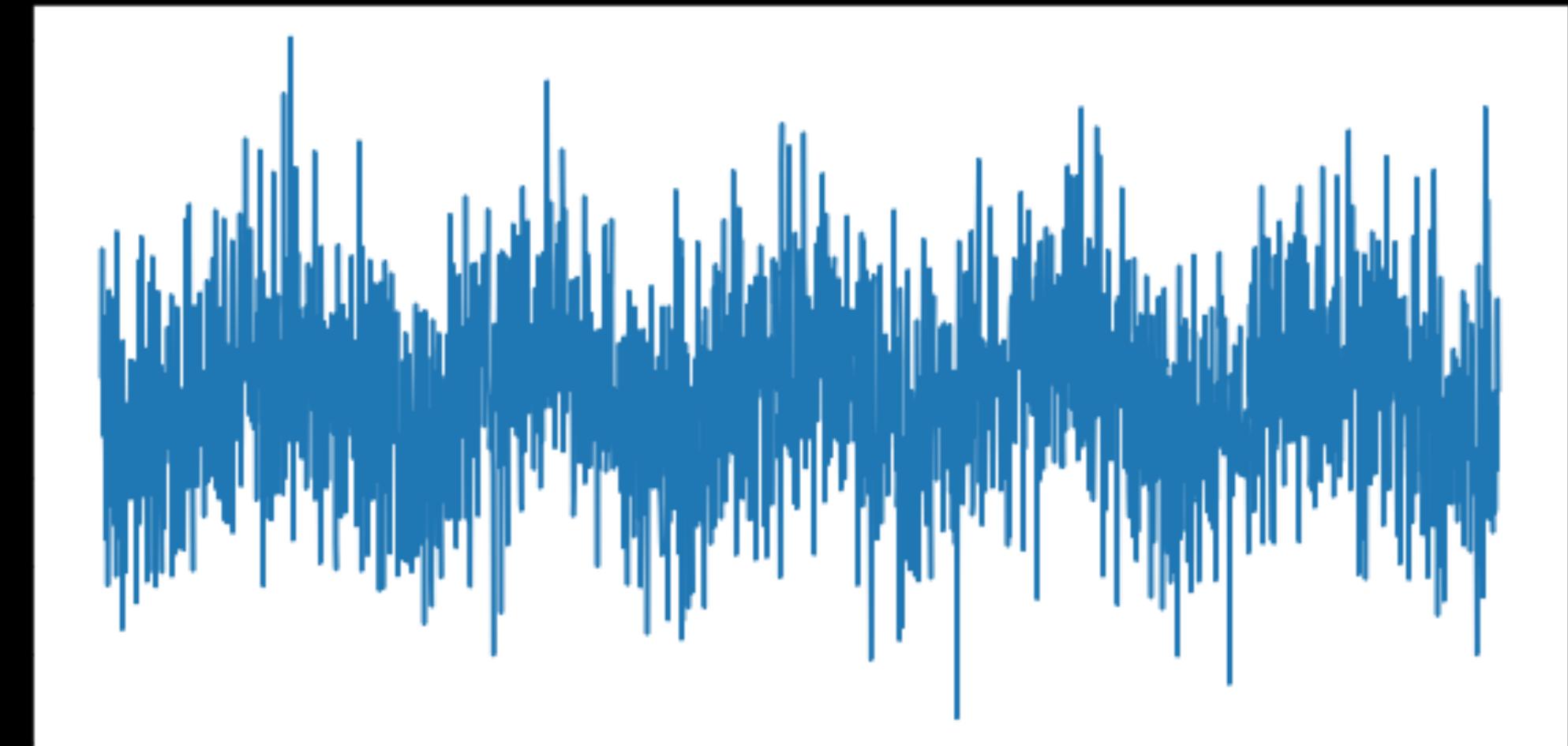
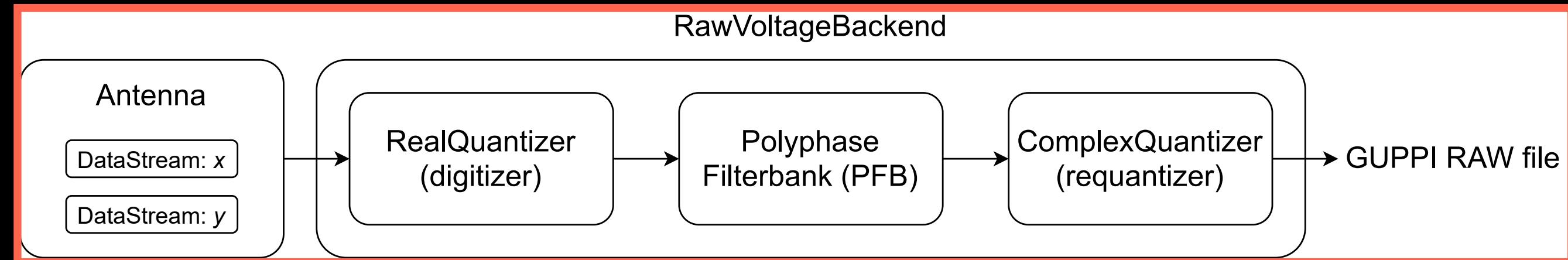
# Setigen's basic design philosophy

- The Cadence object can be thought of as a souped-up array of Frames
- Can use array slicing, like in NumPy, to select which Frames to inject signals
- Can specify the cadence type, e.g. ABACAD, and filter by letter
- Mainly convenience so you don't need to manually track relative time for signal calculations



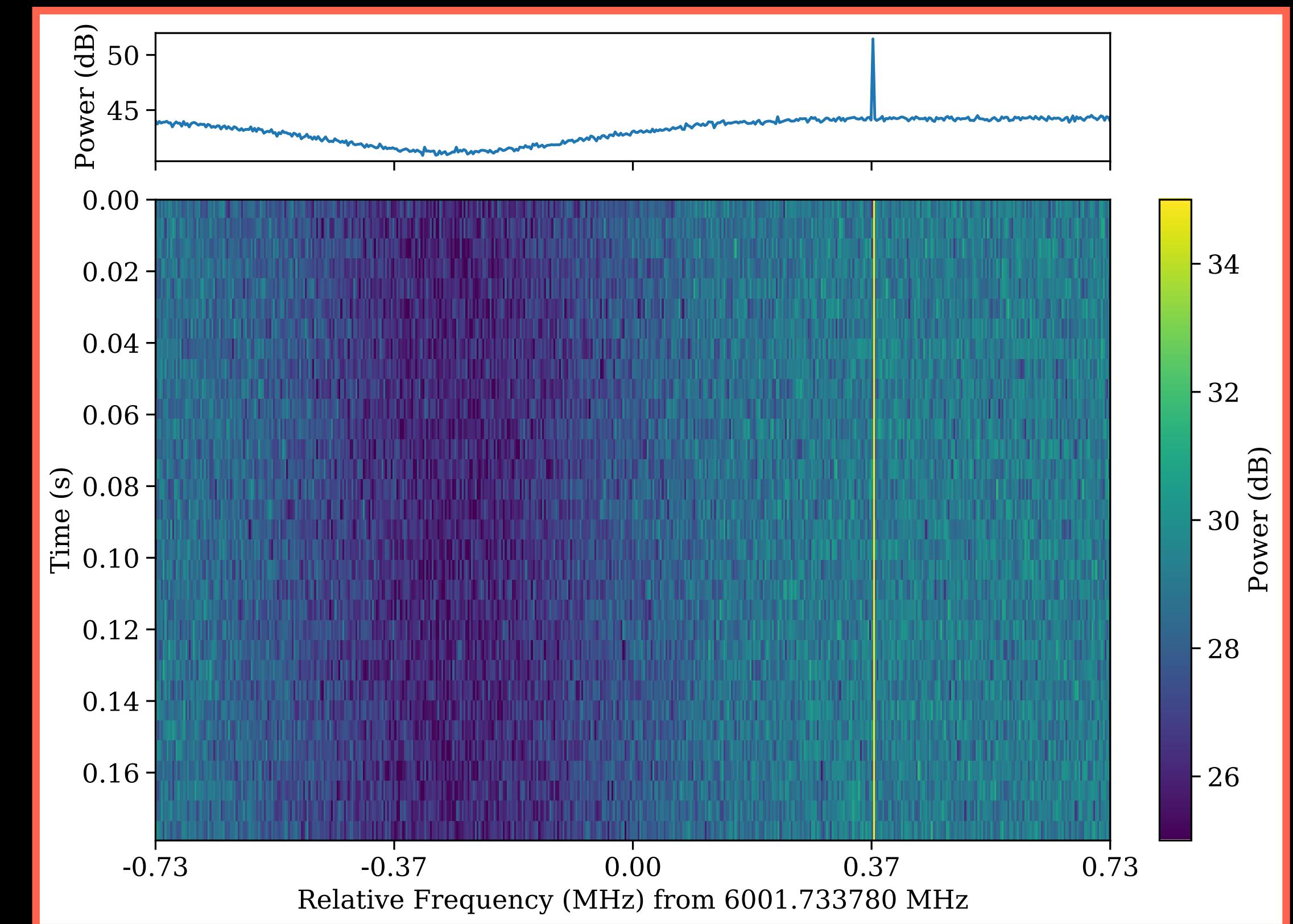
# Setigen's voltage module design philosophy

- Synthetic complex voltage data
- Noise and signal injection is done at the real voltage level, as an antenna would receive power – you would put a sine wave for a perfect non-drifting monotone signal
- Simple (but effective) models of real pipeline components, such as digitizers and a polyphase filterbank



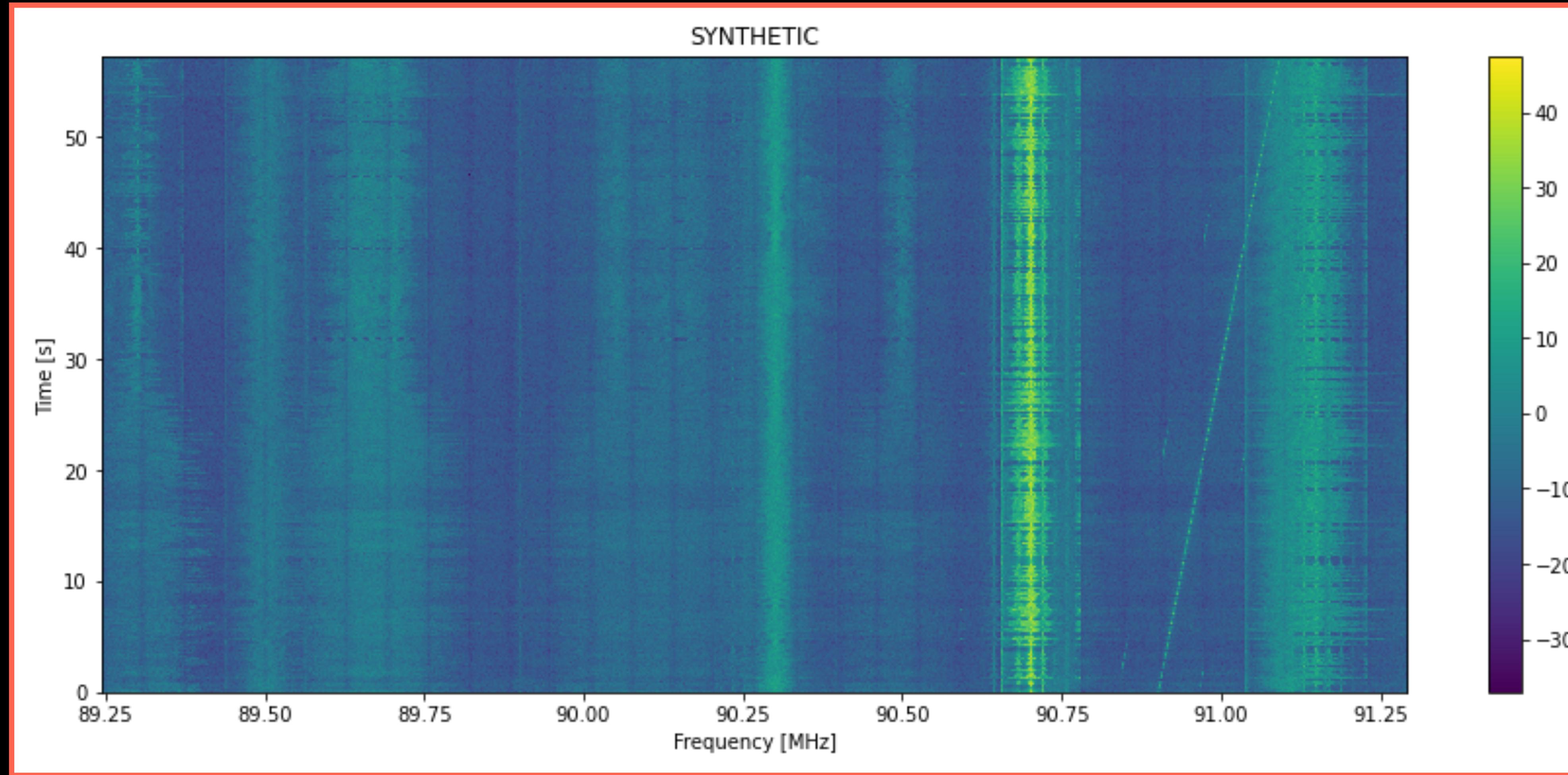
# Setigen's voltage module design philosophy

- Data formats and steps are based on Breakthrough Listen conventions, e.g. GUPPI RAW format
  - Used to test end-to-end detection pipeline (reduction, etc.)
- Wayyy more expensive compared to spectrogram injection, but is the most “correct” way to do things. Can use GPU to accelerate, which helps a bit.



# Silly RTL-SDR demo injection

[github.com/bbrzycki/rtlsdr-to-setigen](https://github.com/bbrzycki/rtlsdr-to-setigen)

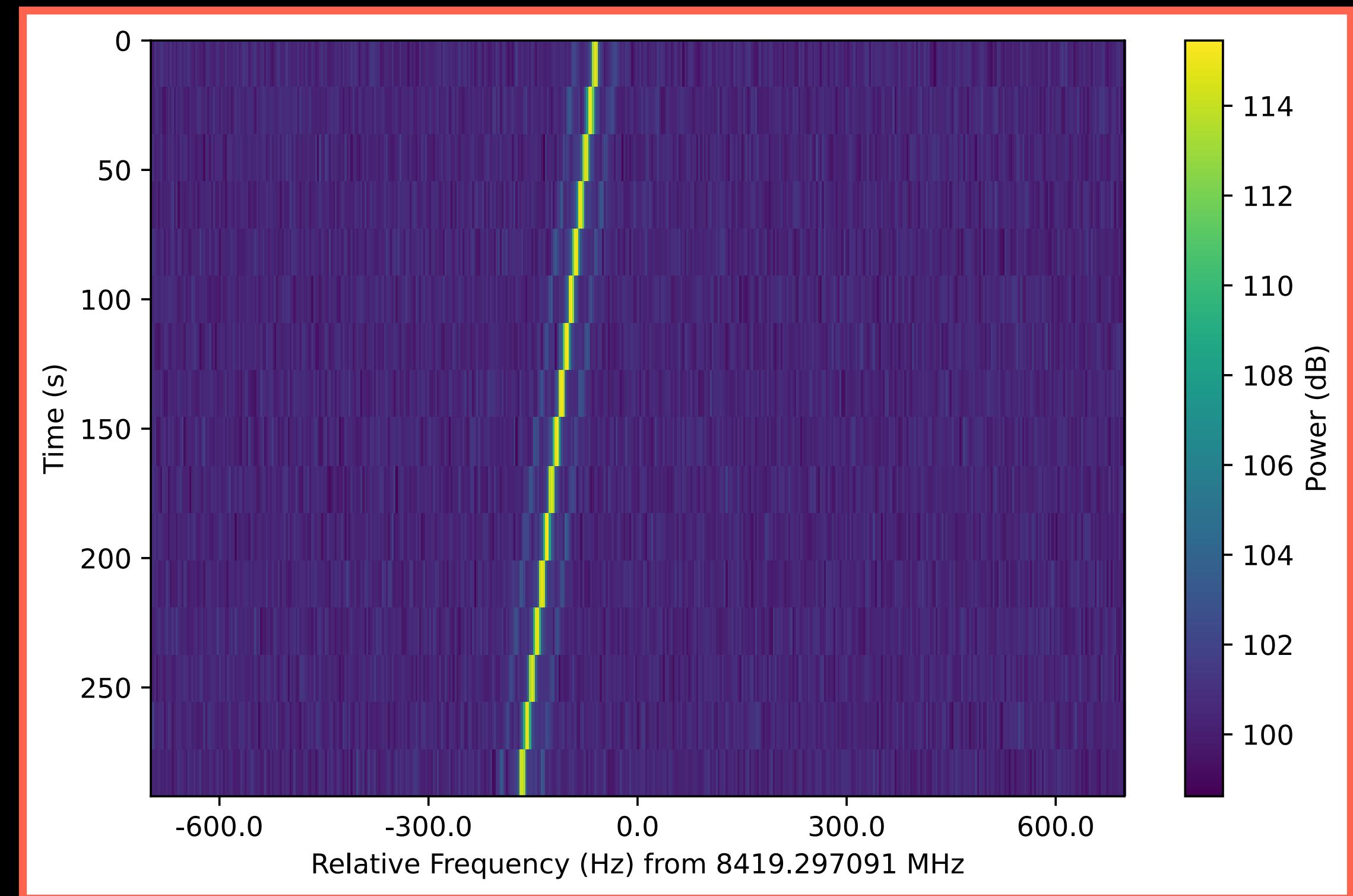


# **Links, if any of this sounds helpful for your project!**

- [github.com/bbrzycki/setigen](https://github.com/bbrzycki/setigen)
- [setigen.readthedocs.io](https://setigen.readthedocs.io)
- Brzycki et al. 2022, AJ
  - In particular, includes a discussion of the essential single-dish signal chain with some useful equations

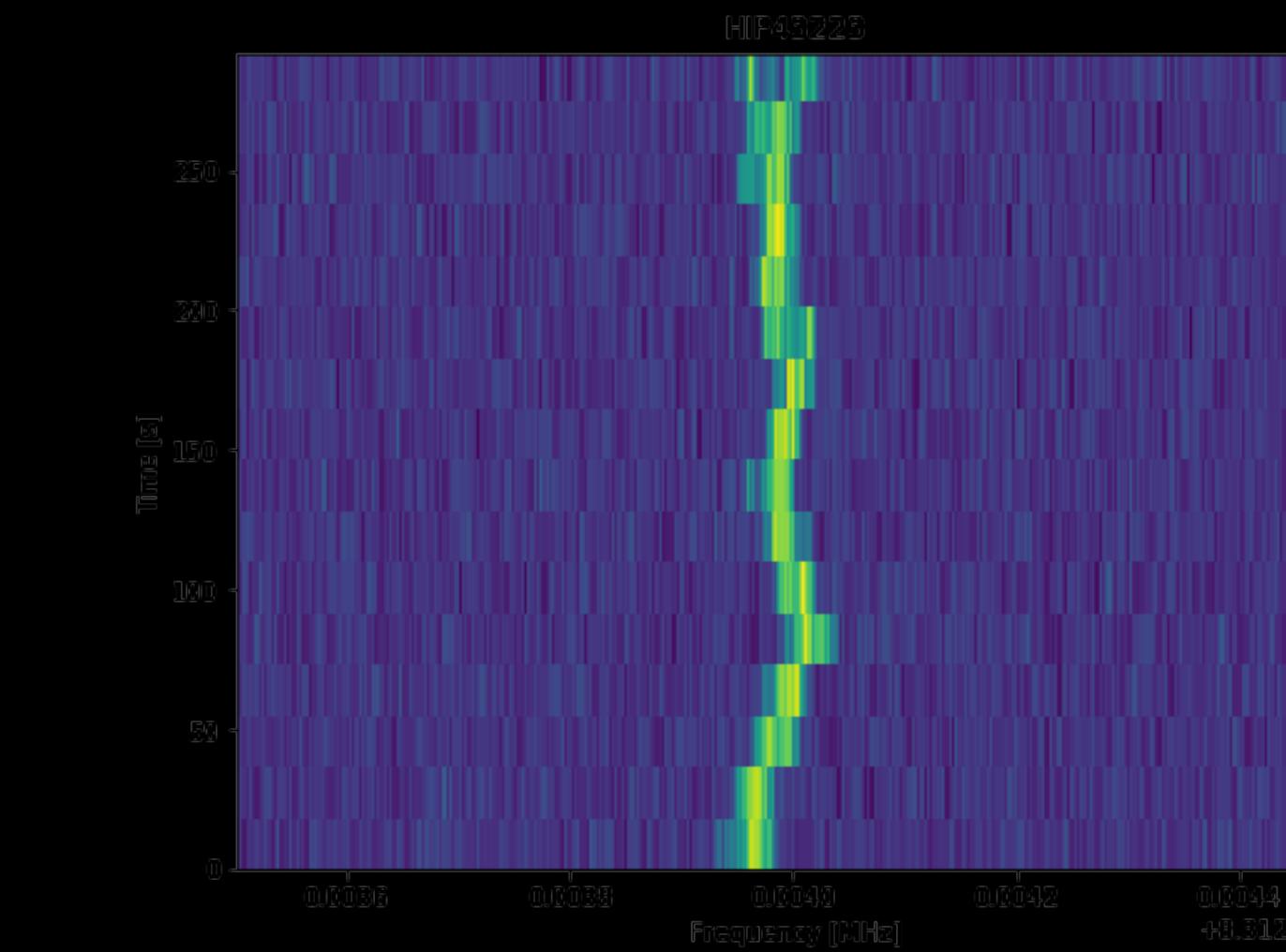
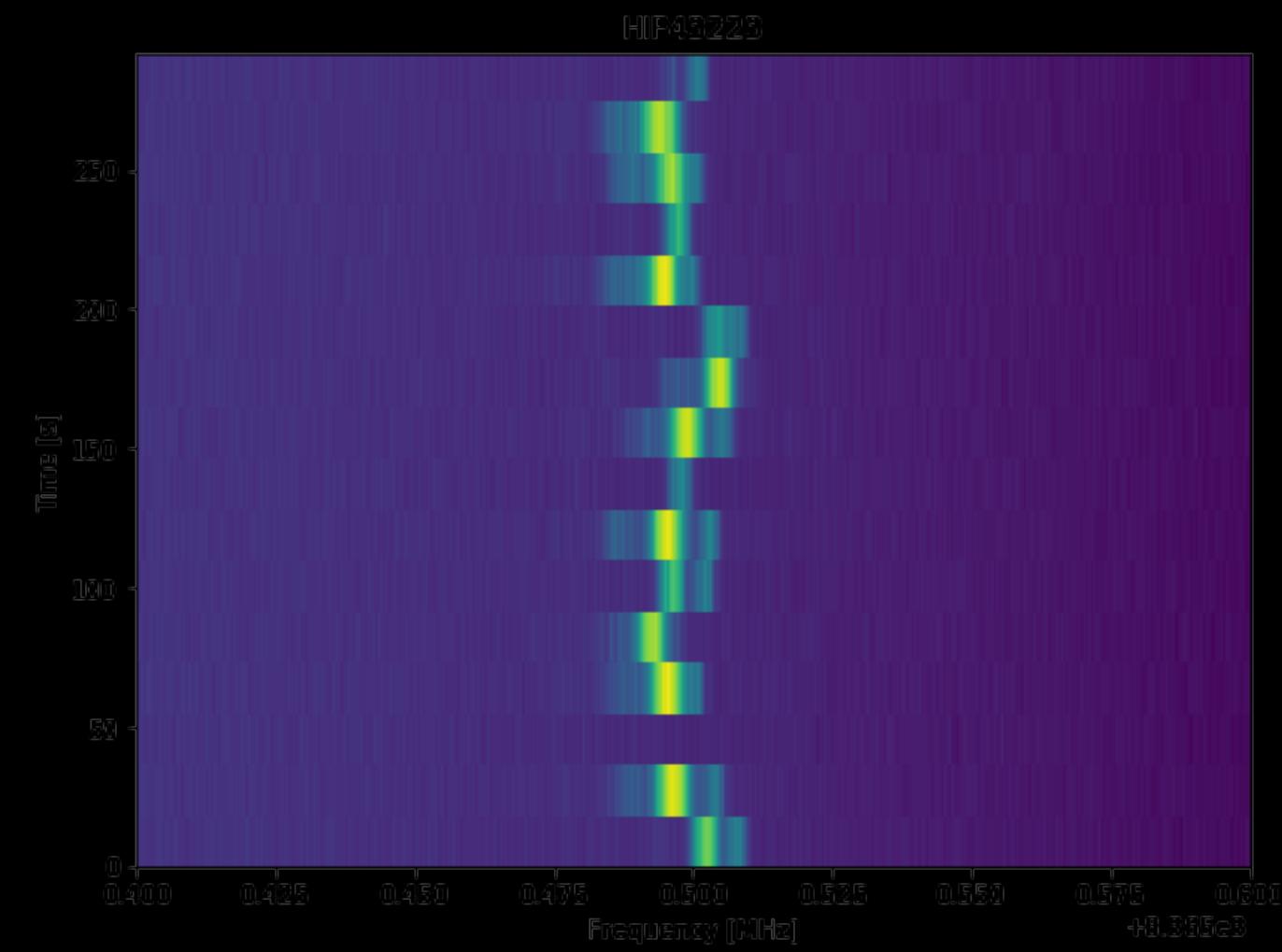
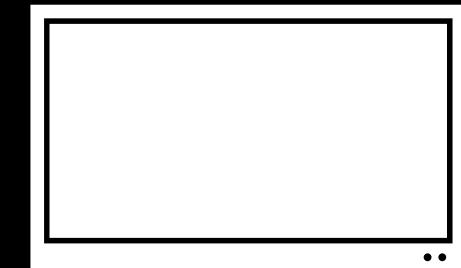
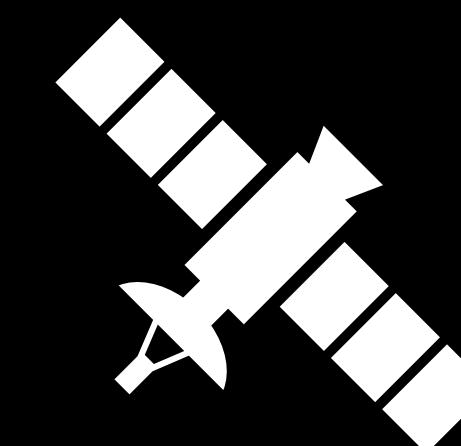
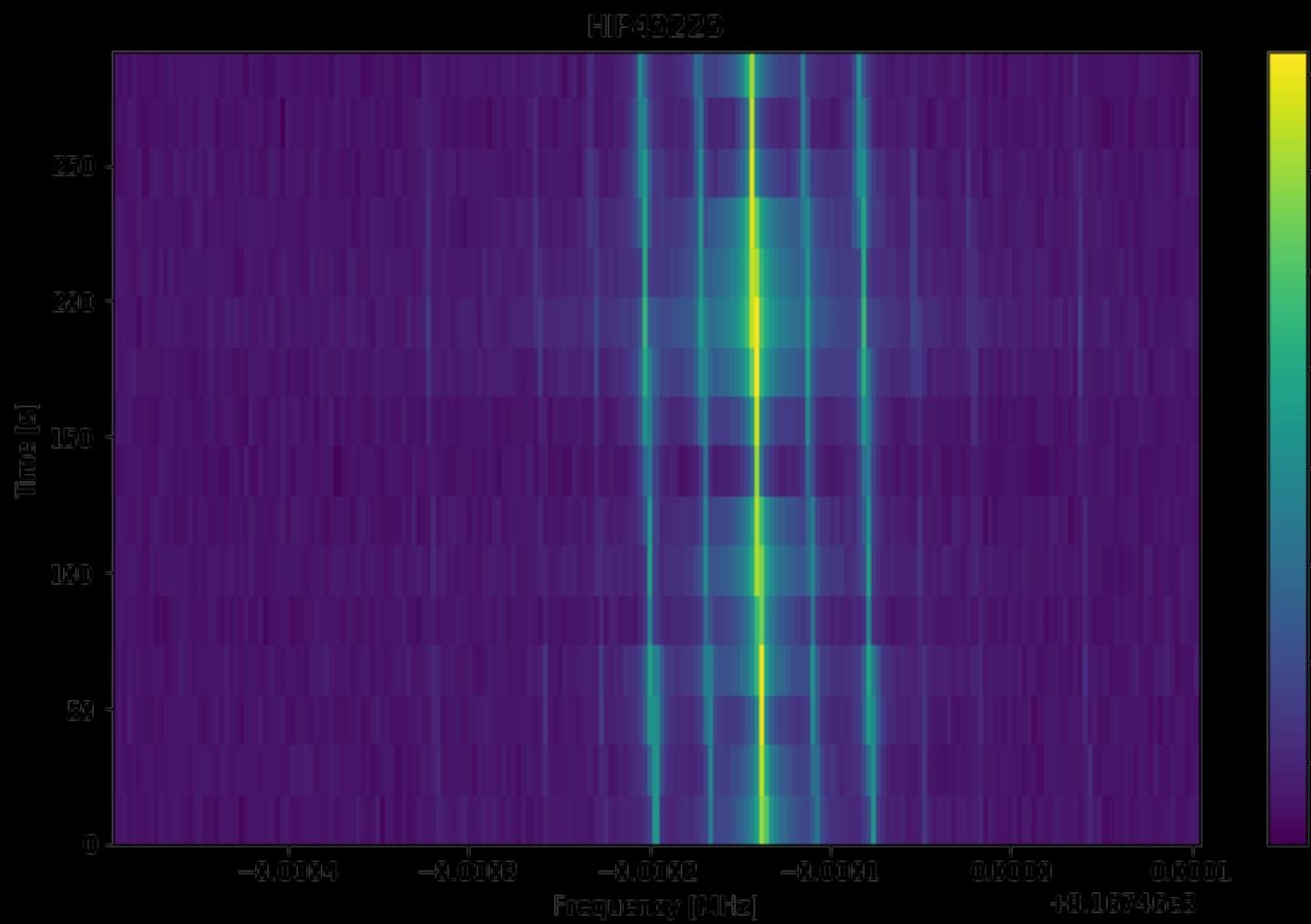
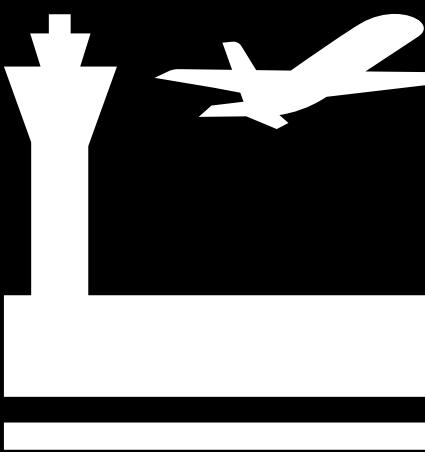
# Modern radio SETI in practice

- Basic signal detection
  - Gather all signals possible above signal-to-noise threshold
- Candidate identification and differentiation
  - Primarily against human-created radio frequency interference (RFI)
- Lots of manual inspection

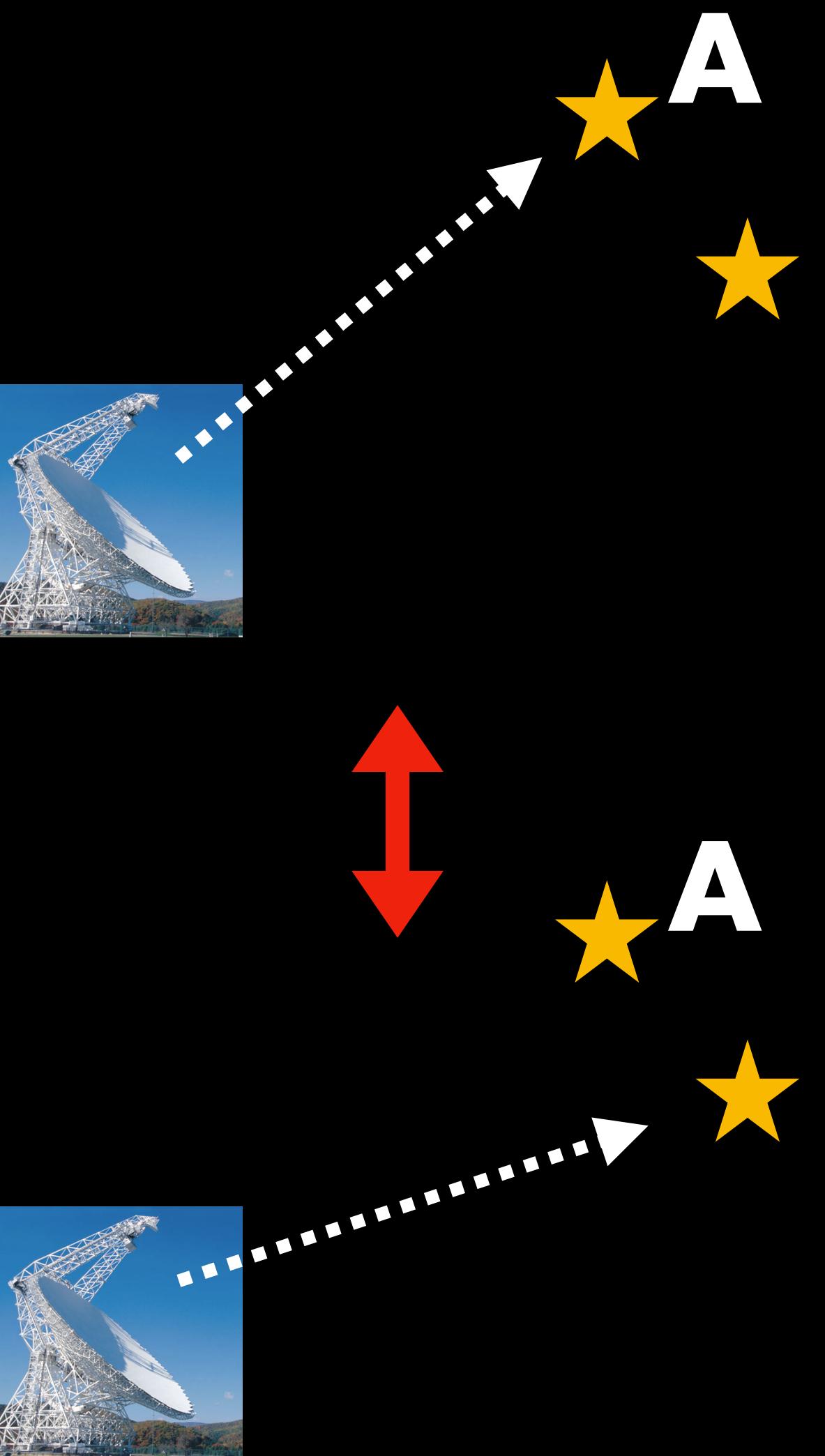


**Detected signal from Voyager 1**

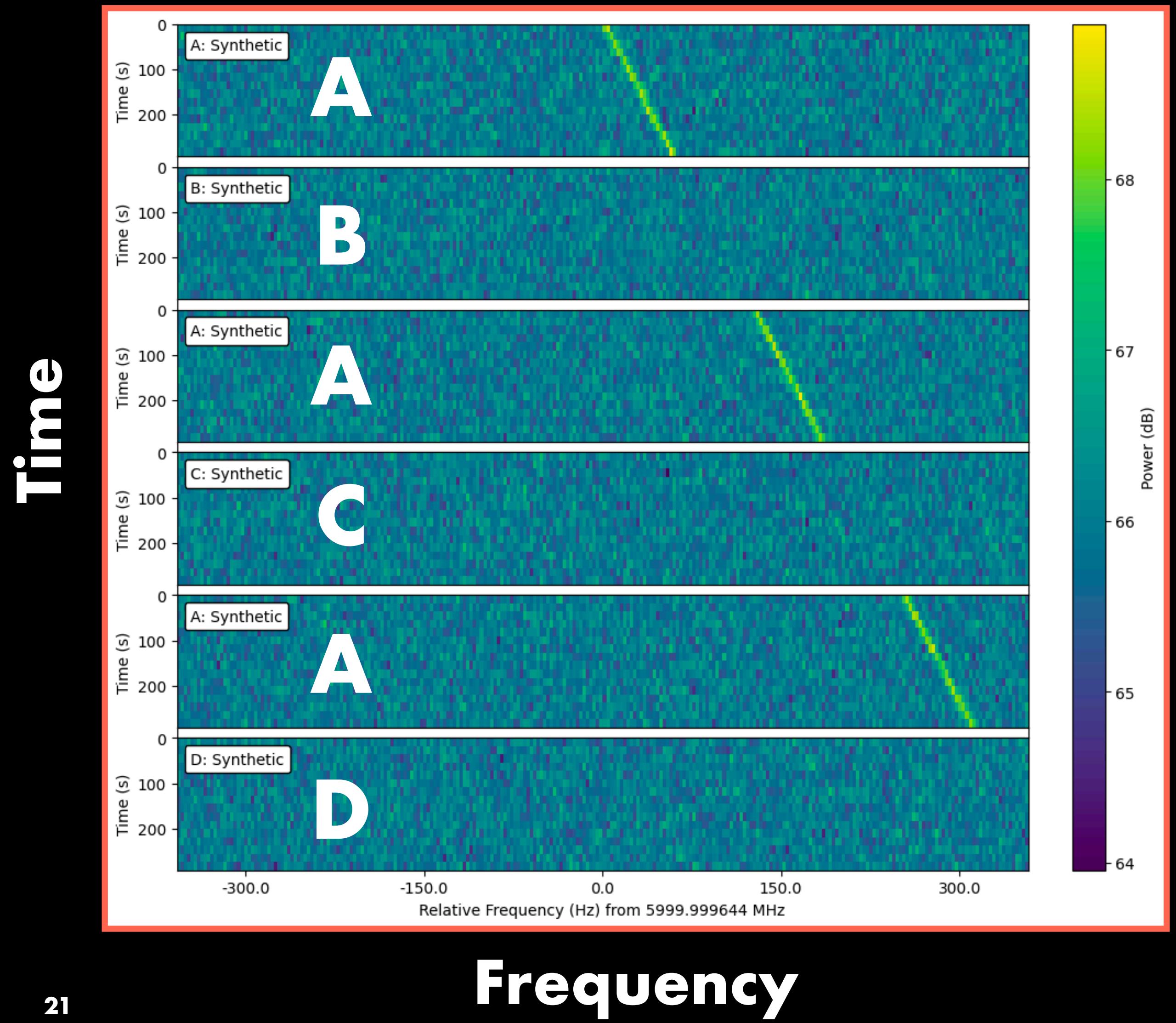
# Differentiation against RFI



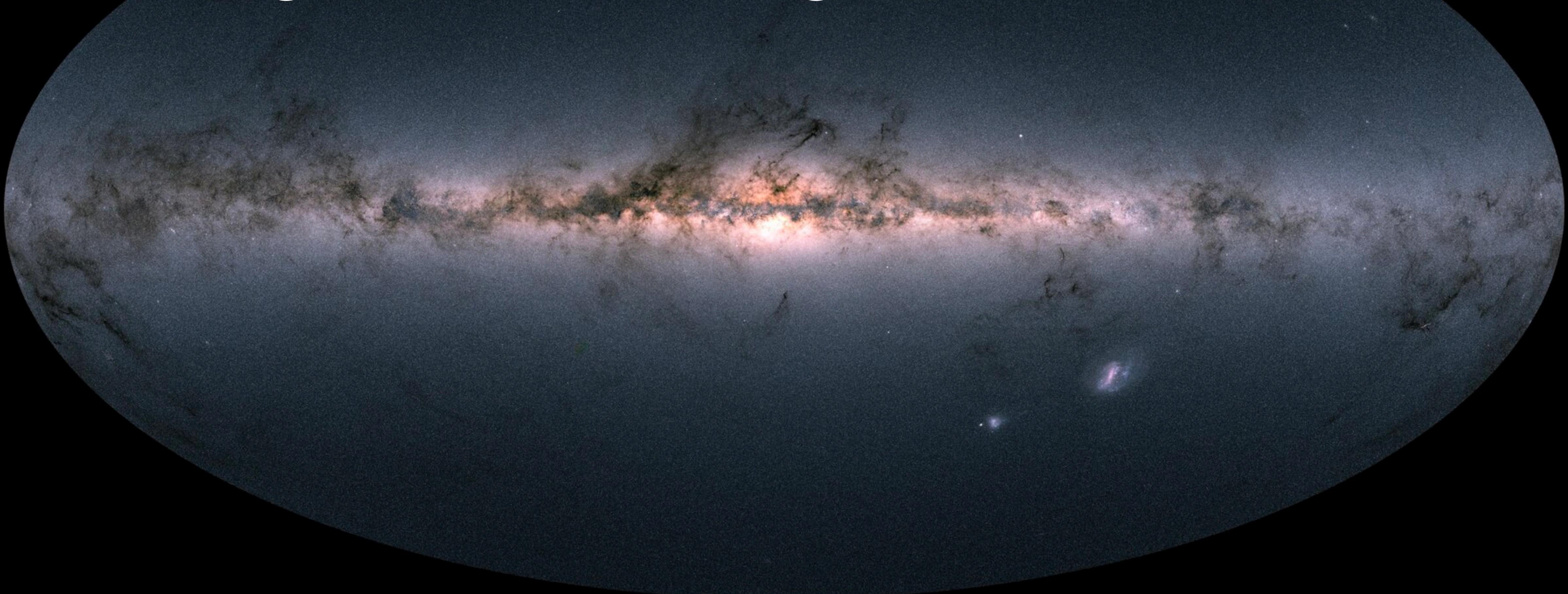
# Sky localization (ON-OFF) filter



NRAO

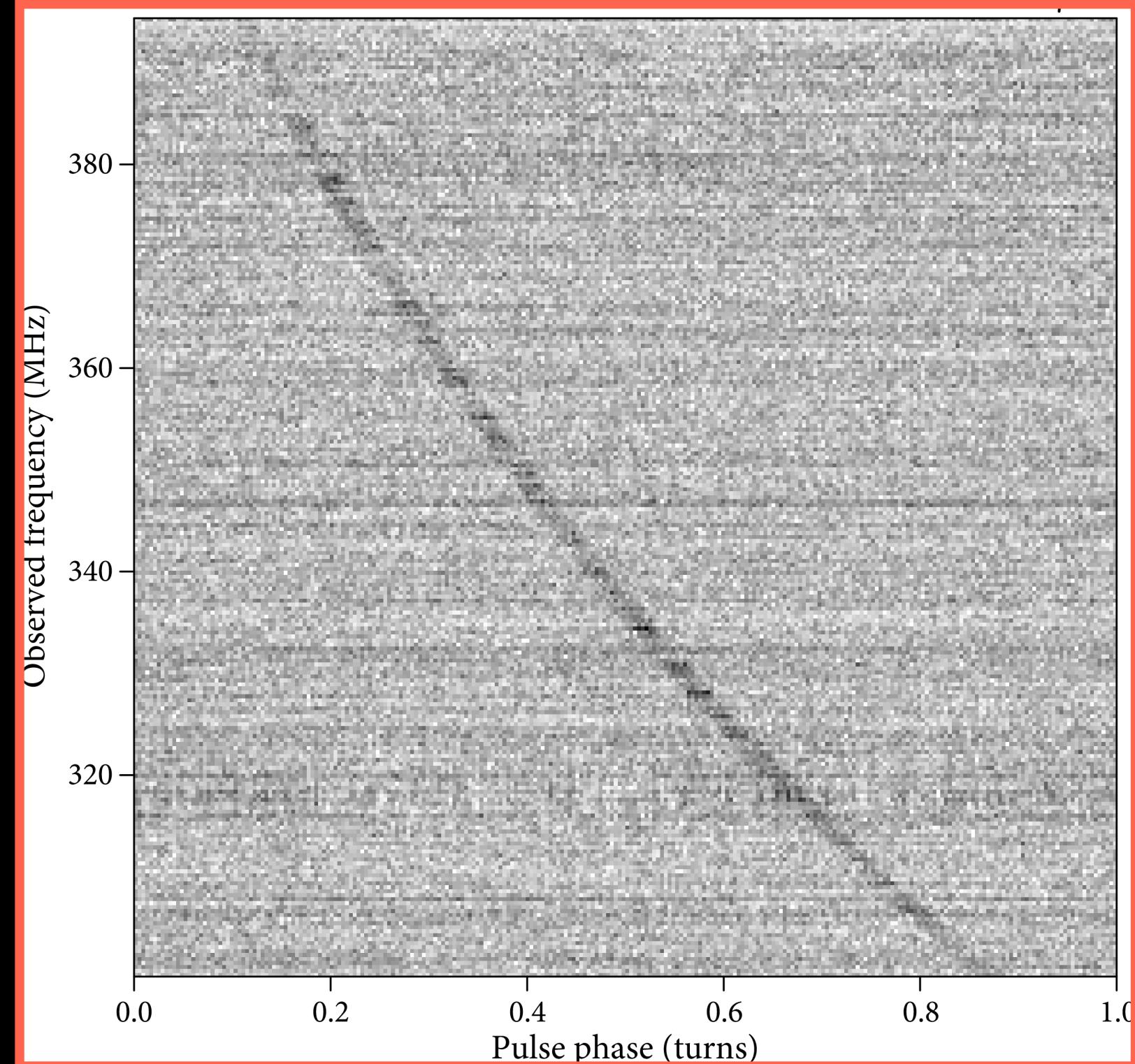


# Can we use astrophysical phenomena to distinguish technosignatures from RFI?



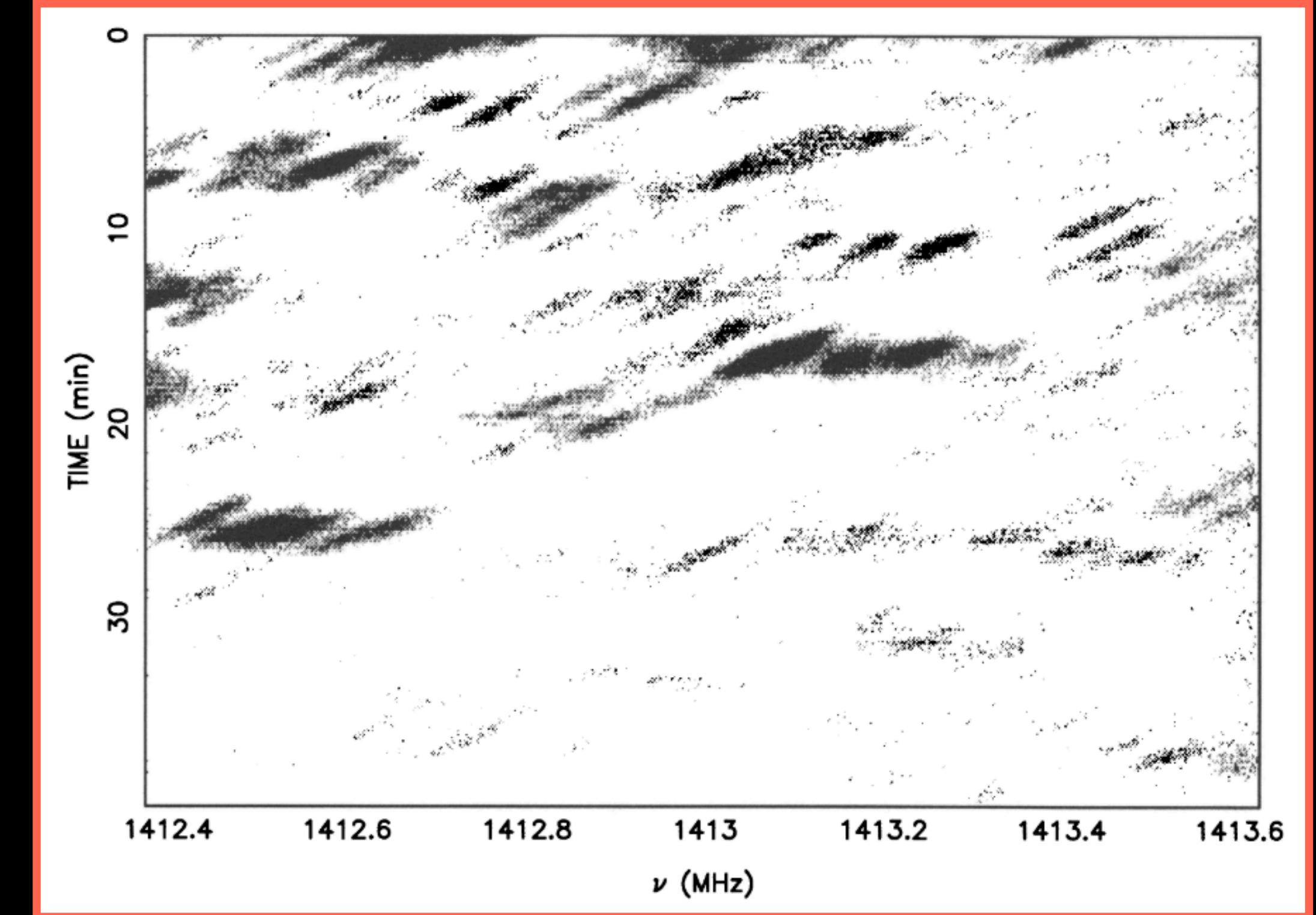
# Pulsar observations probe radio ISM plasma effects

**Dispersion**



Condon & Ransom 2016

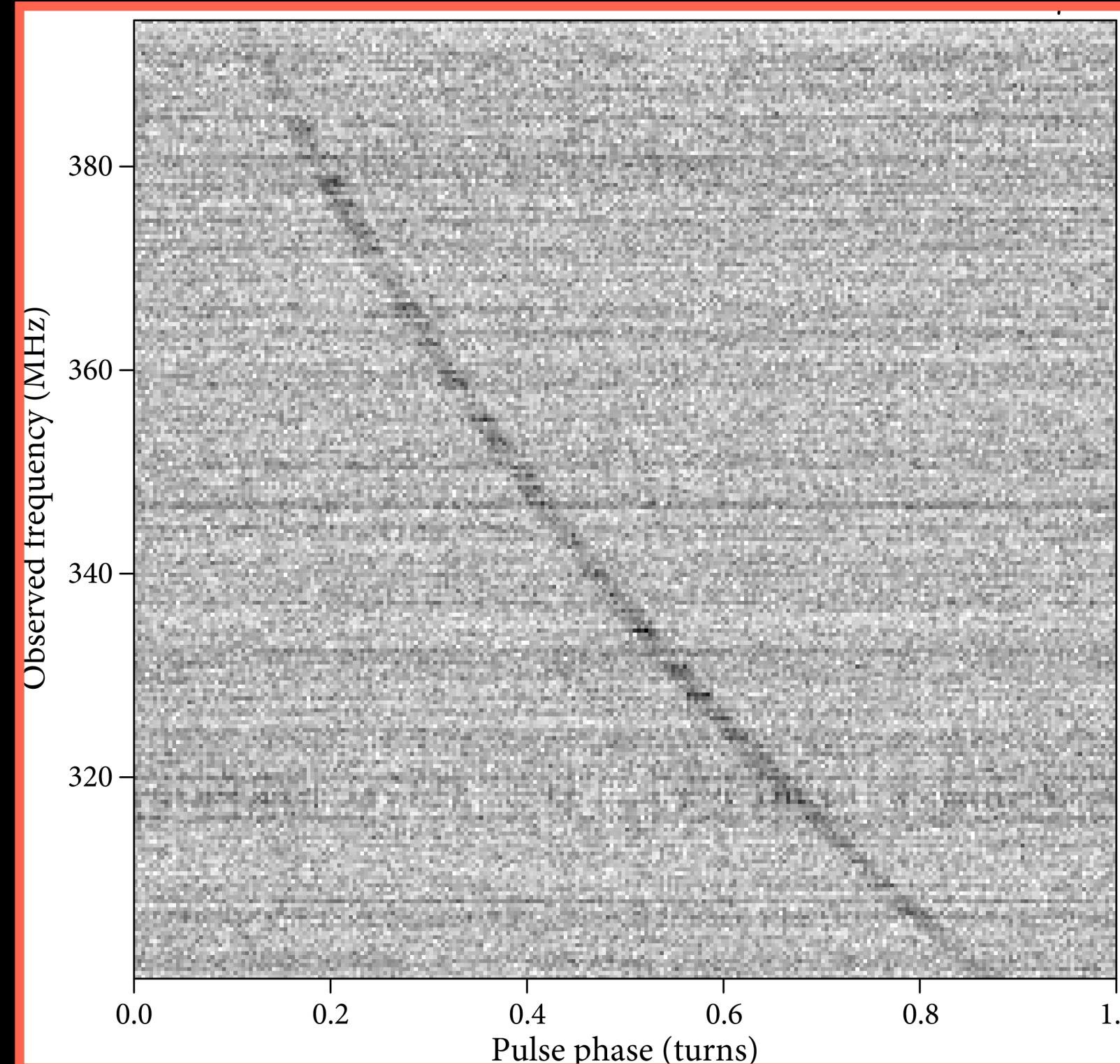
**Scattering**



Cordes & Lazio 1991

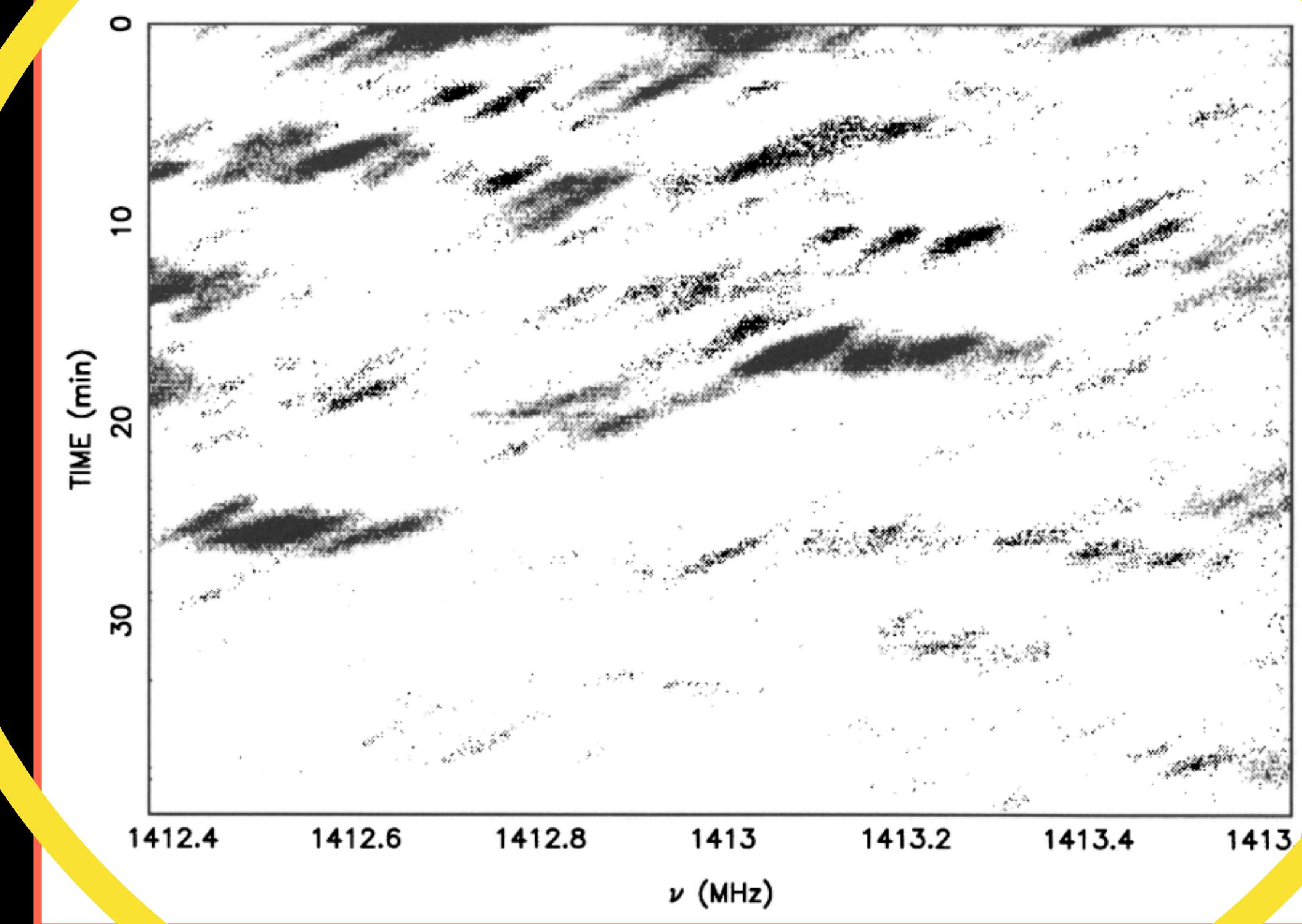
# Pulsar observations probe radio ISM plasma effects

Dispersion



Condon & Ransom 2016

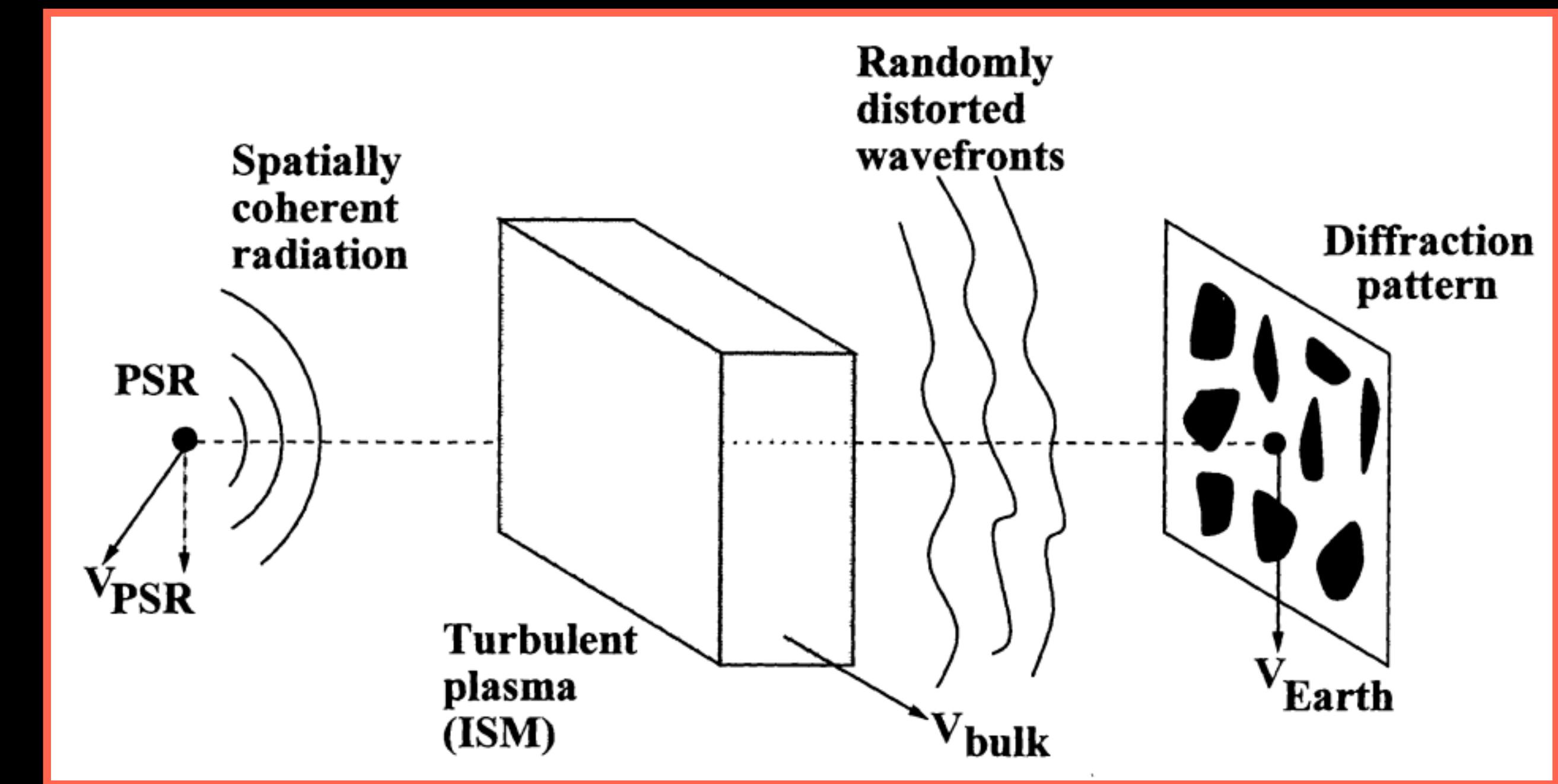
Scattering



Cordes & Lazio 1991

# Diffractive scintillation in the ISM

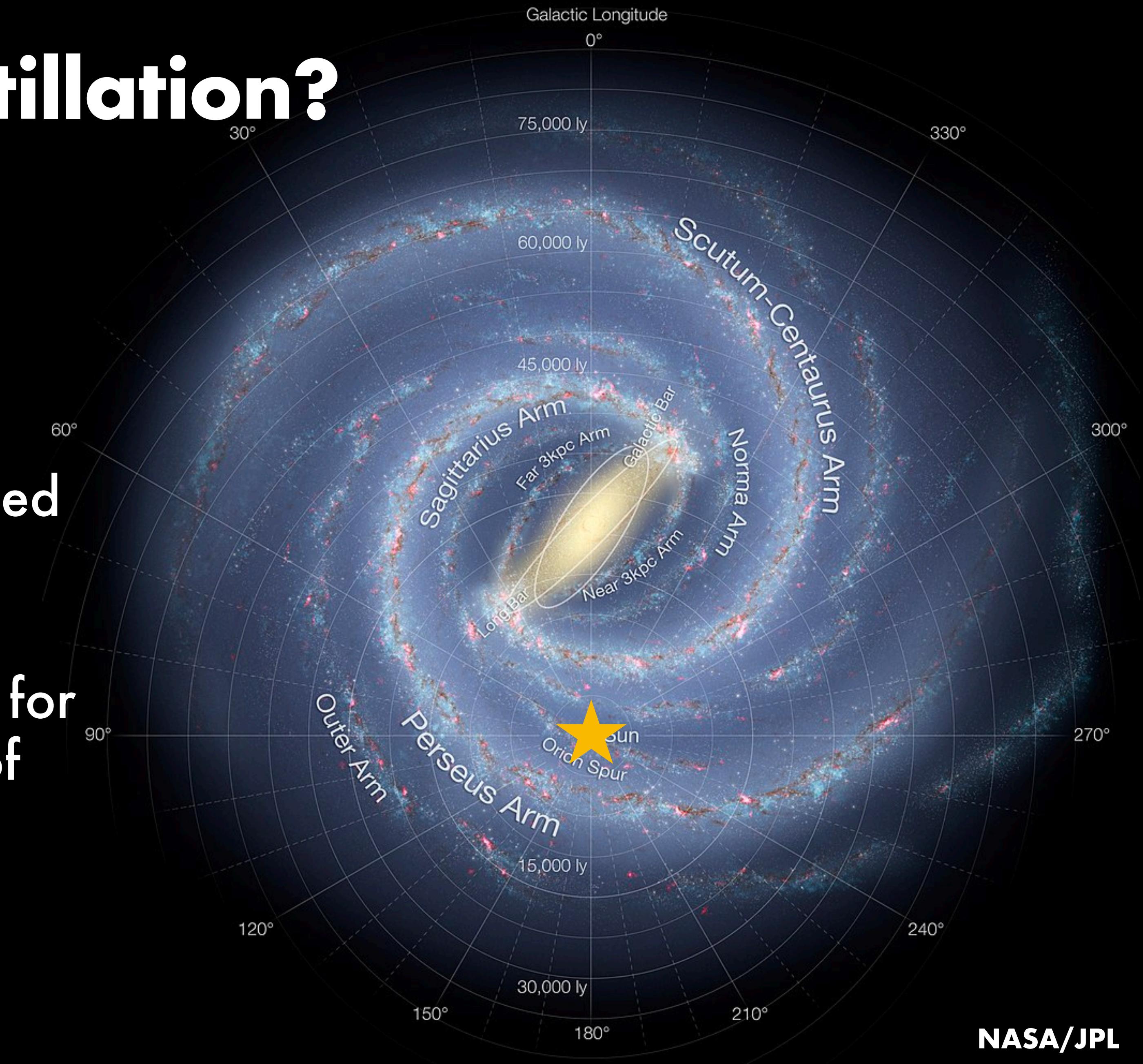
- Electron density fluctuations in ionized plasma → interference pattern
- Can lead to 100% intensity modulation, especially towards the Galactic center, with characteristic scintillation timescale  $\Delta t_d$



Cordes 2002

# Why search for scintillation?

- A filter that directly implies extra-solar origin
- Well-suited for continuous or pulsed narrowband signals
- One of the best places to search for scintillation corresponds to one of the best places to look for ETI – the Galactic Center



# **Can we detect scintillated narrowband technosignatures?**

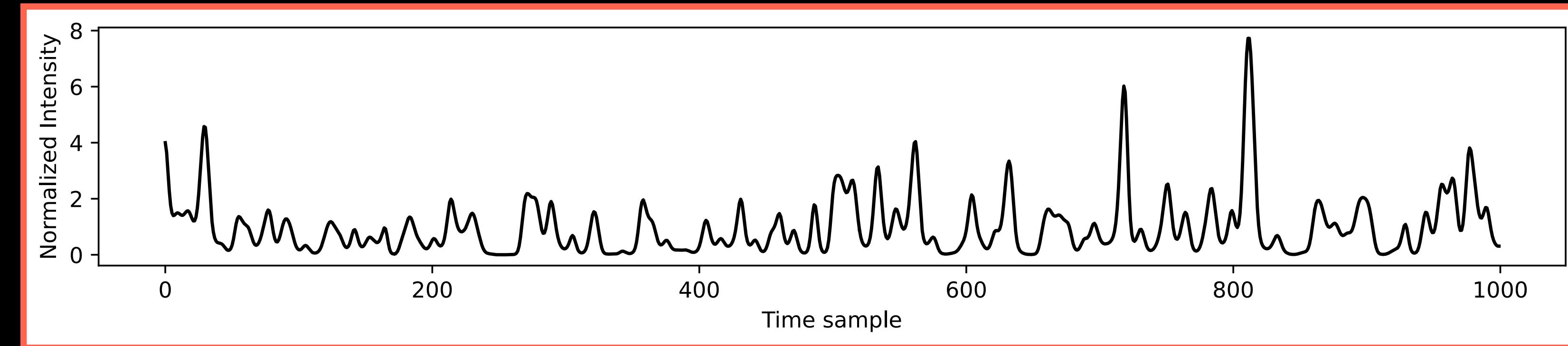
- 1. Given a signal exhibiting scintillations, can we identify it as such?**
- 2. Can we differentiate scintillated signals from existing RFI (in terms of intensity modulation)?**

# What would strongly scintillated signals look like?

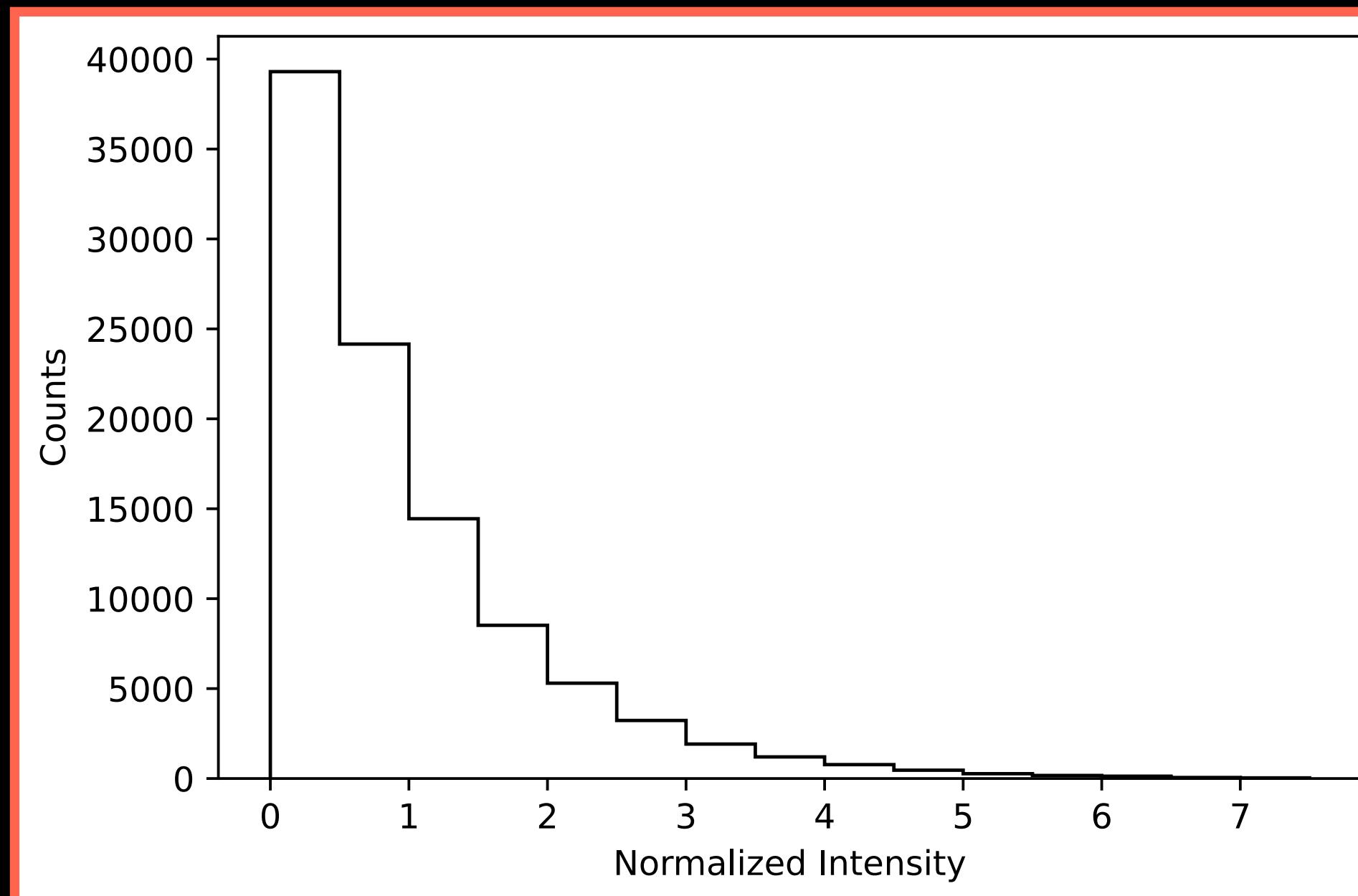
Cordes & Lazio 1991

Cordes, Lazio, & Sagan 1997

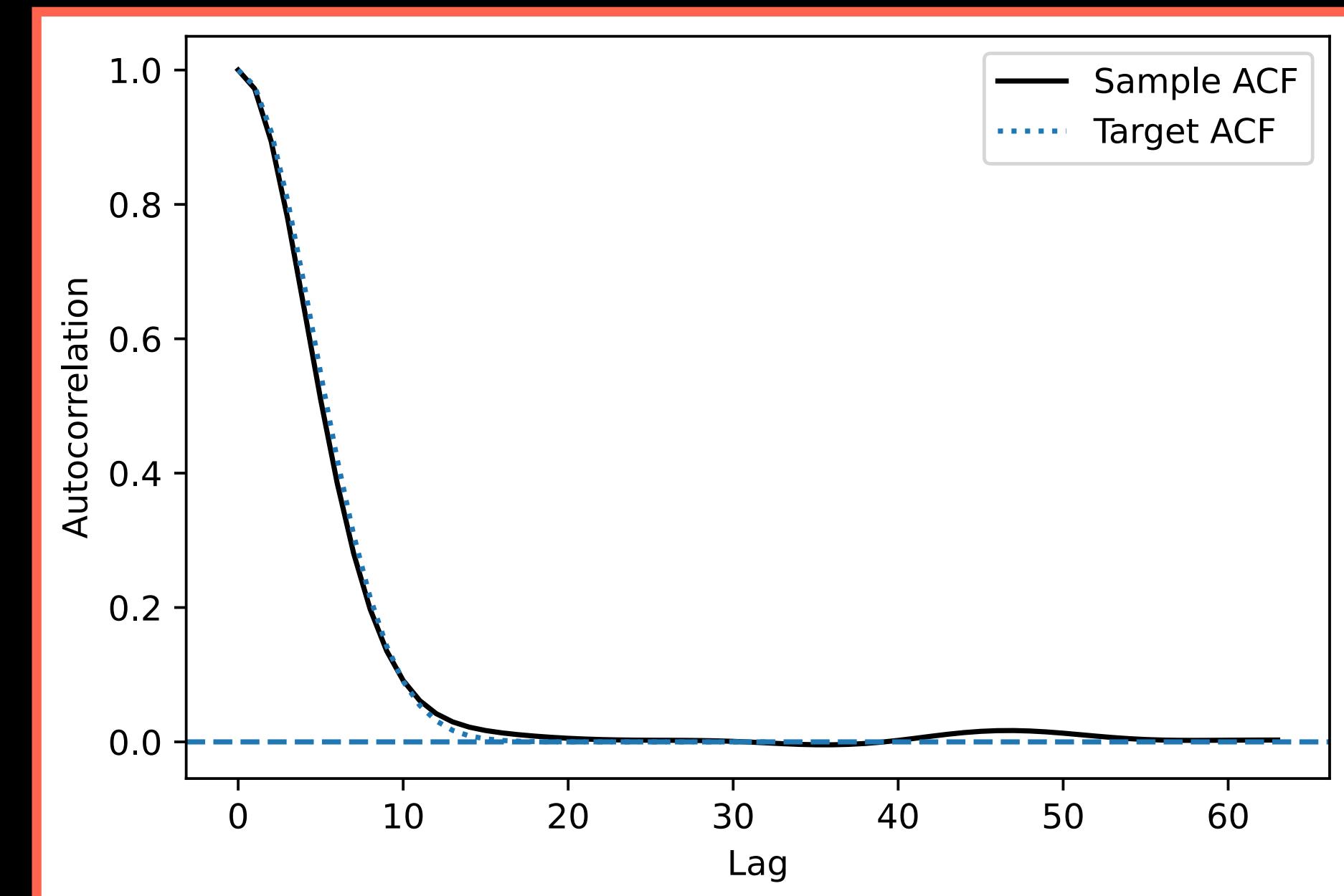
## Synthetic time series



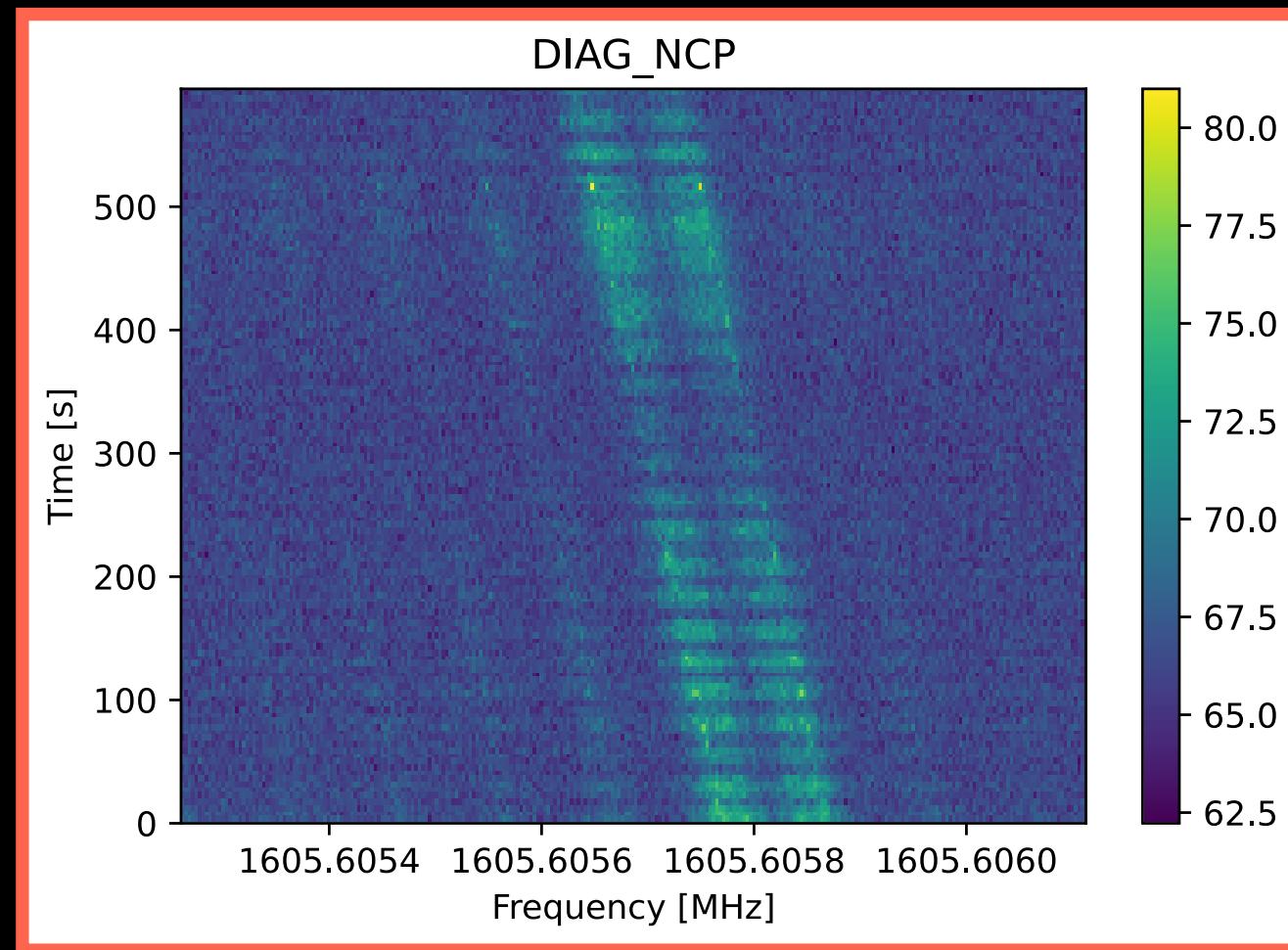
## Exponential intensity distribution



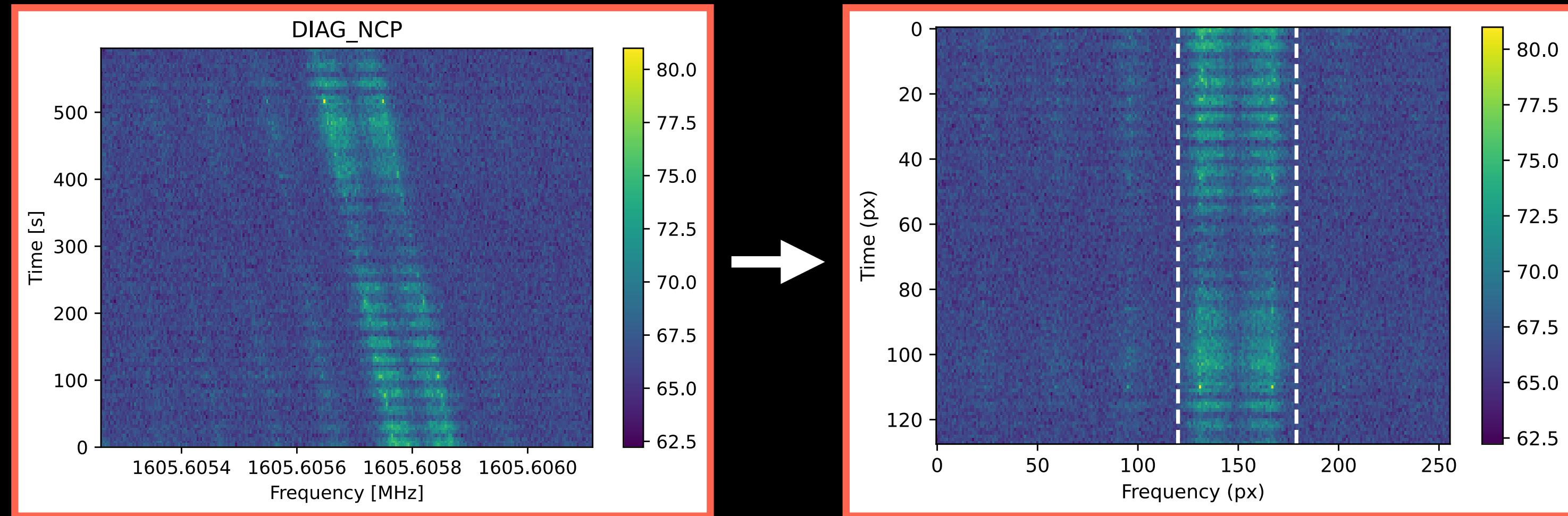
## Near-Gaussian autocorrelation



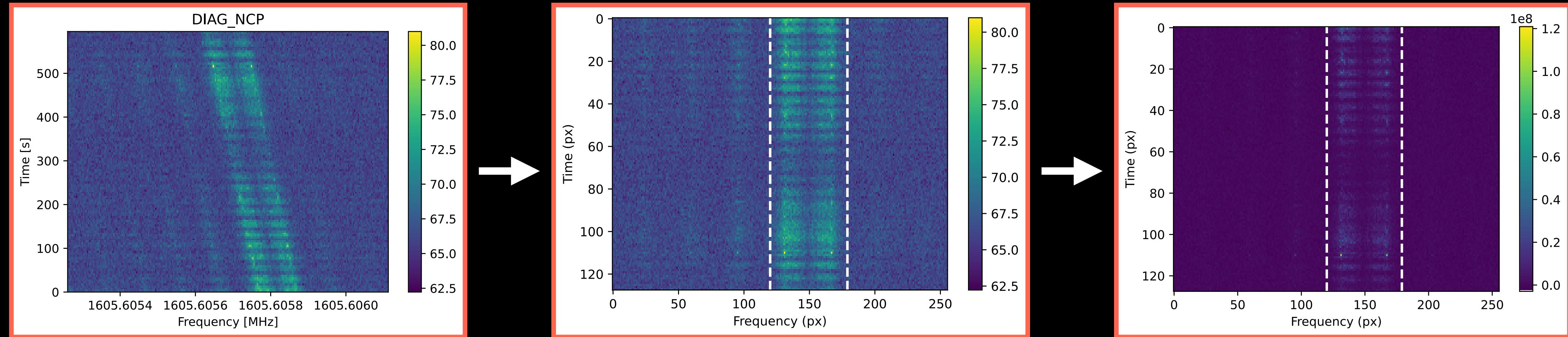
# Analysis steps for a detected narrowband signal



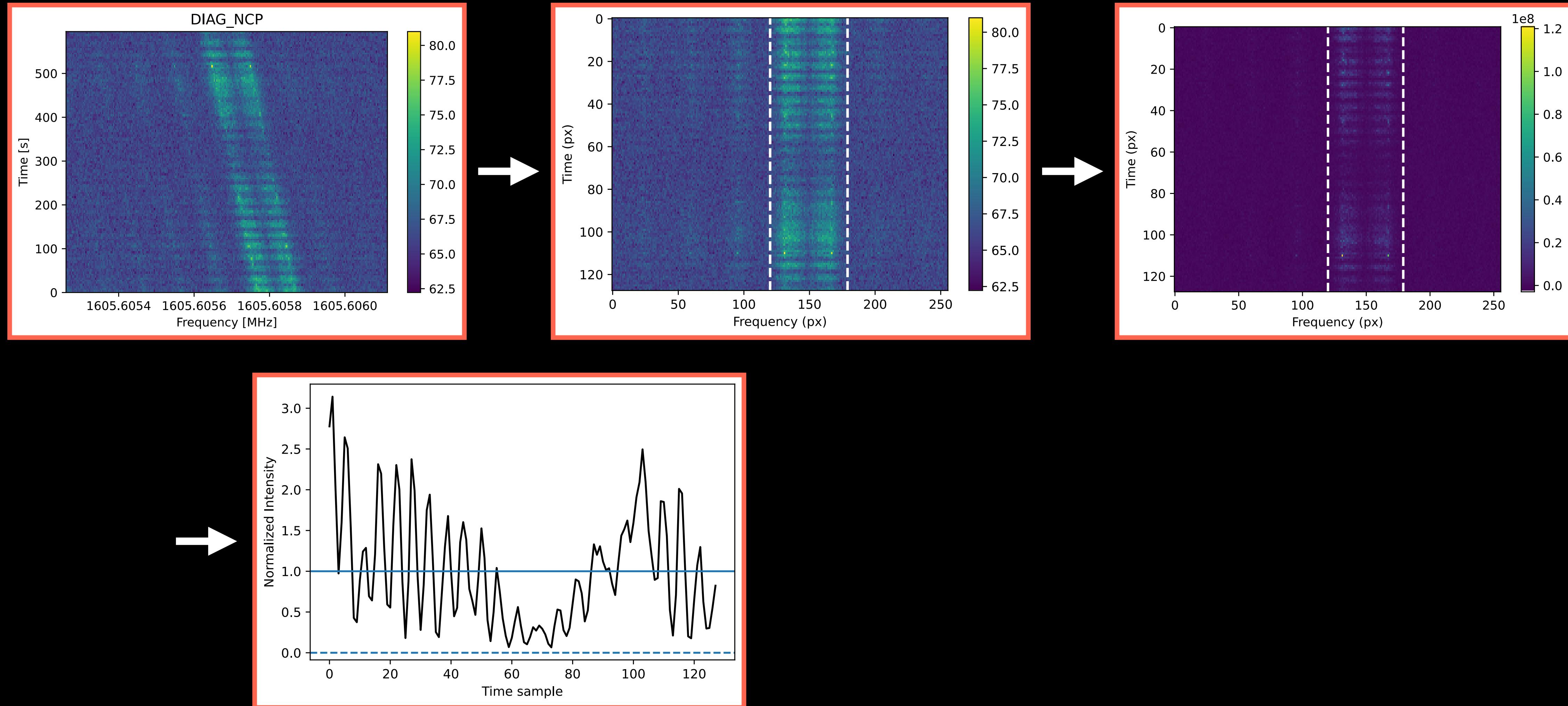
# Analysis steps for a detected narrowband signal



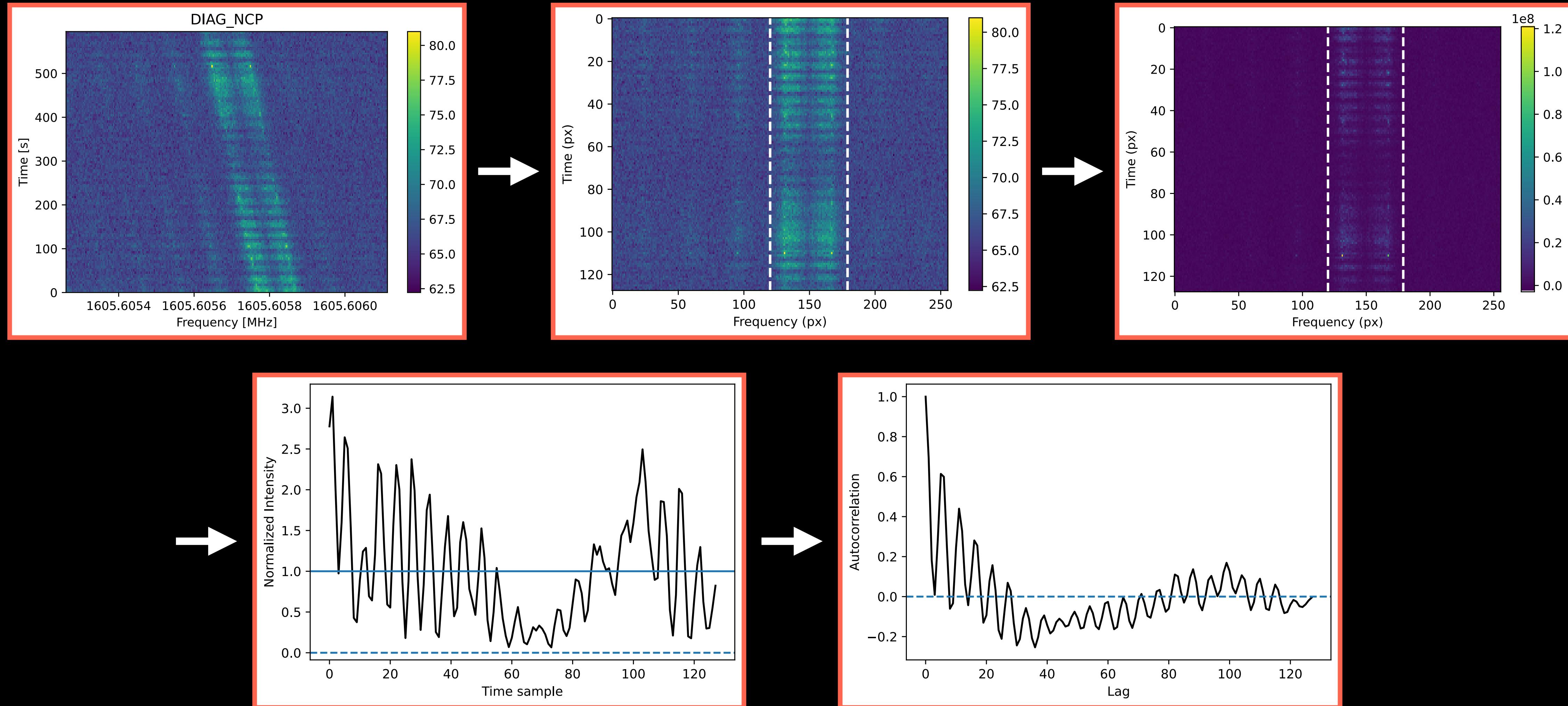
# Analysis steps for a detected narrowband signal



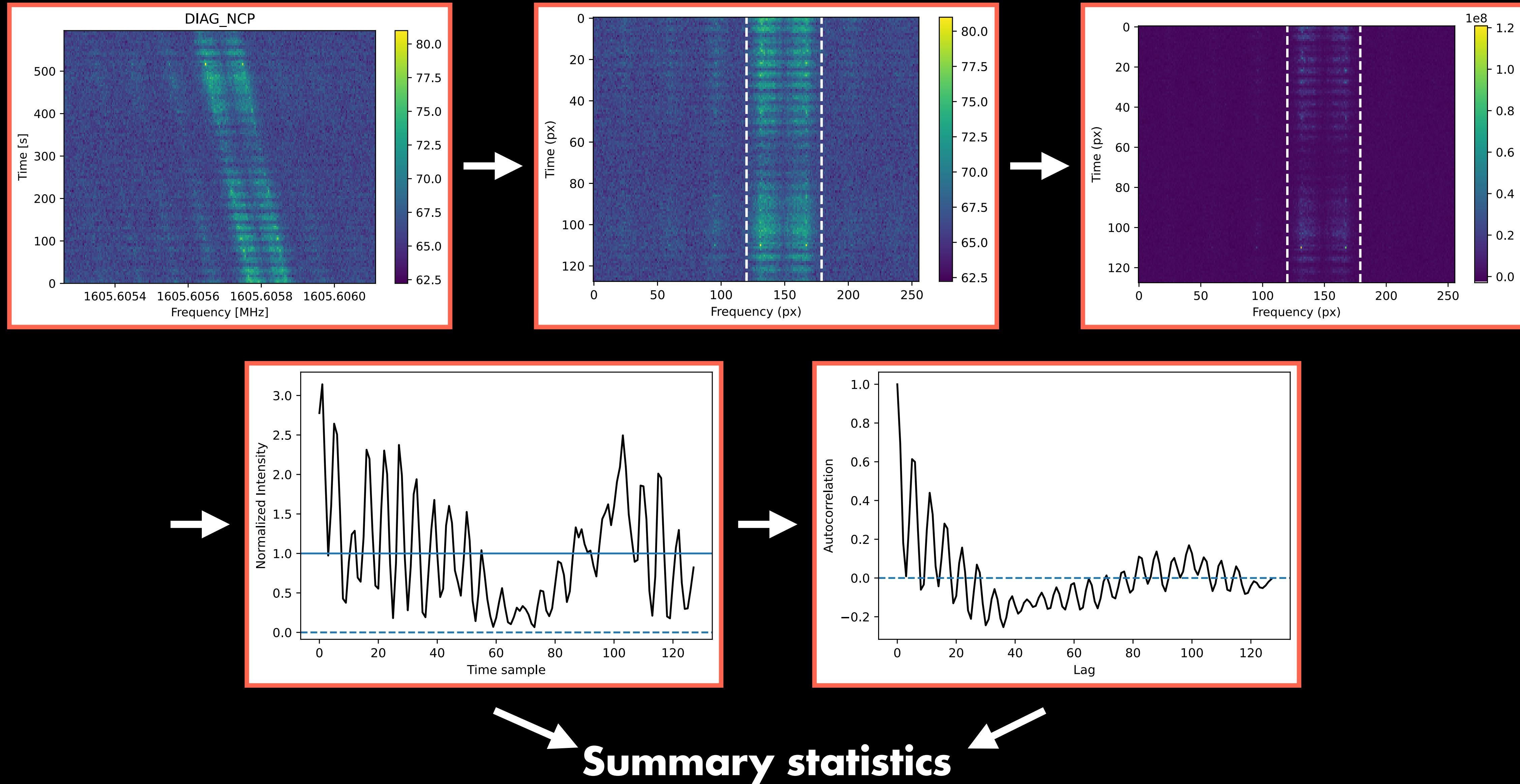
# Analysis steps for a detected narrowband signal



# Analysis steps for a detected narrowband signal



# Analysis steps for a detected narrowband signal



# Set of summary statistics

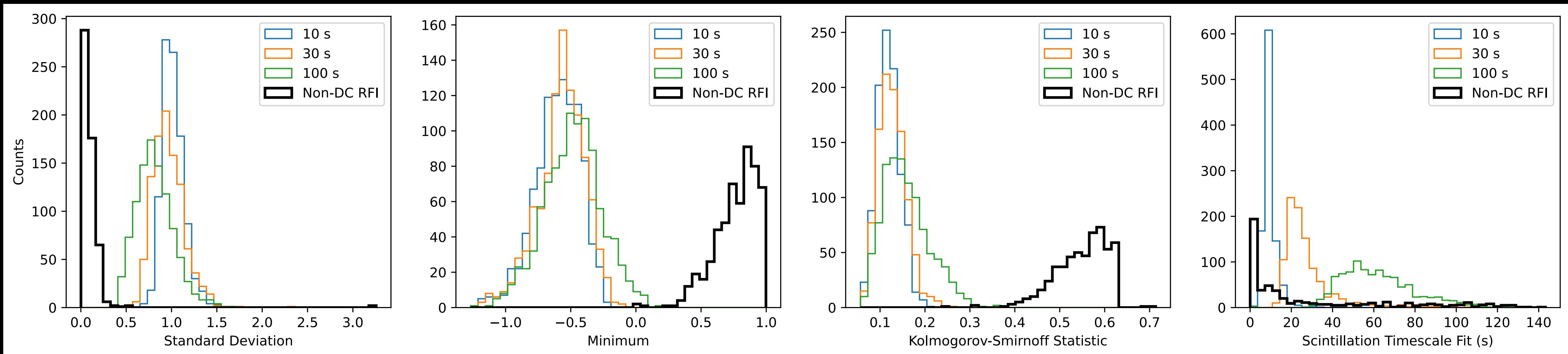
Statistic	Asymptotic Value (with no background noise)	Target
Standard Deviation (RMS)	1	Exponential distribution
Minimum	0	Exponential distribution
Kolmogorov-Smirnoff statistic	0	Exponential distribution
Scintillation Timescale Fit with Least Squares	Variable	Near-Gaussian autocorrelation

# GBT RFI vs. injected synthetic scintillated signals



**S/N > 25**

**C band (4 – 8 GHz)**



**Standard Deviation**

**Minimum**

**Kolmogorov-Smirnov Statistic**

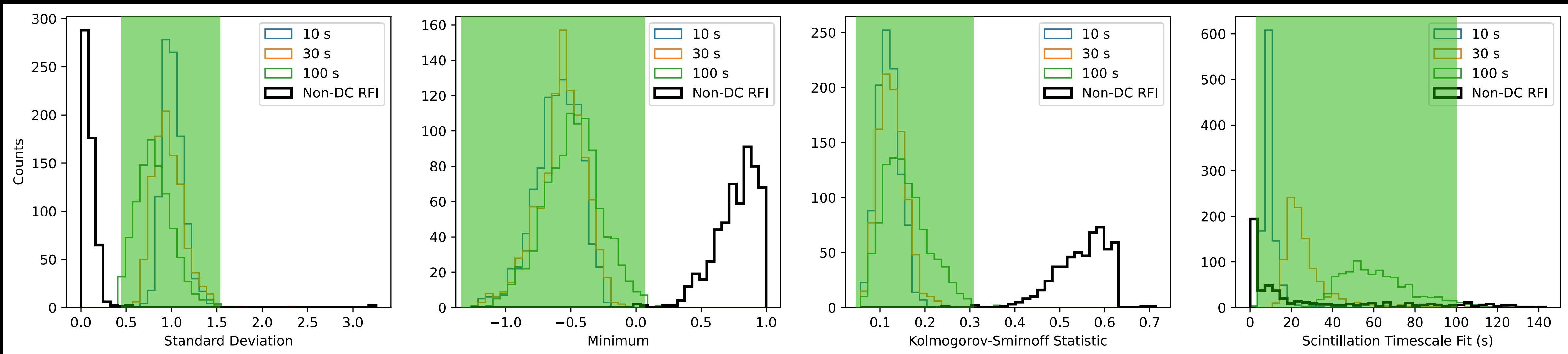
**Scintillation Timescale Fit**

# GBT RFI vs. injected synthetic scintillated signals



**S/N > 25**

**C band (4 – 8 GHz)**



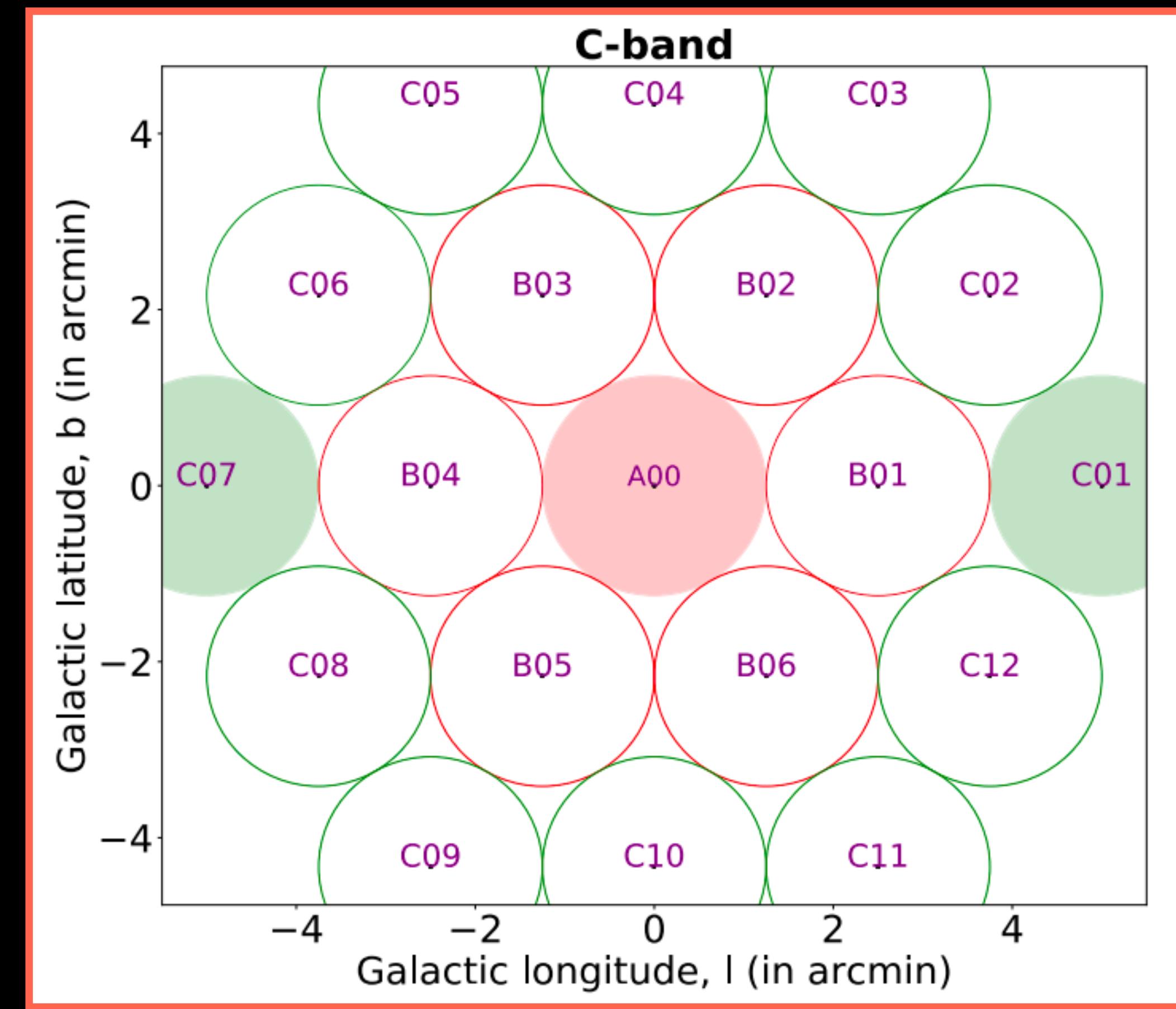
**Standard Deviation**

**Minimum**

**Kolmogorov-Smirnov Statistic**

**Scintillation Timescale Fit**

# Planning Galactic center / plane observations



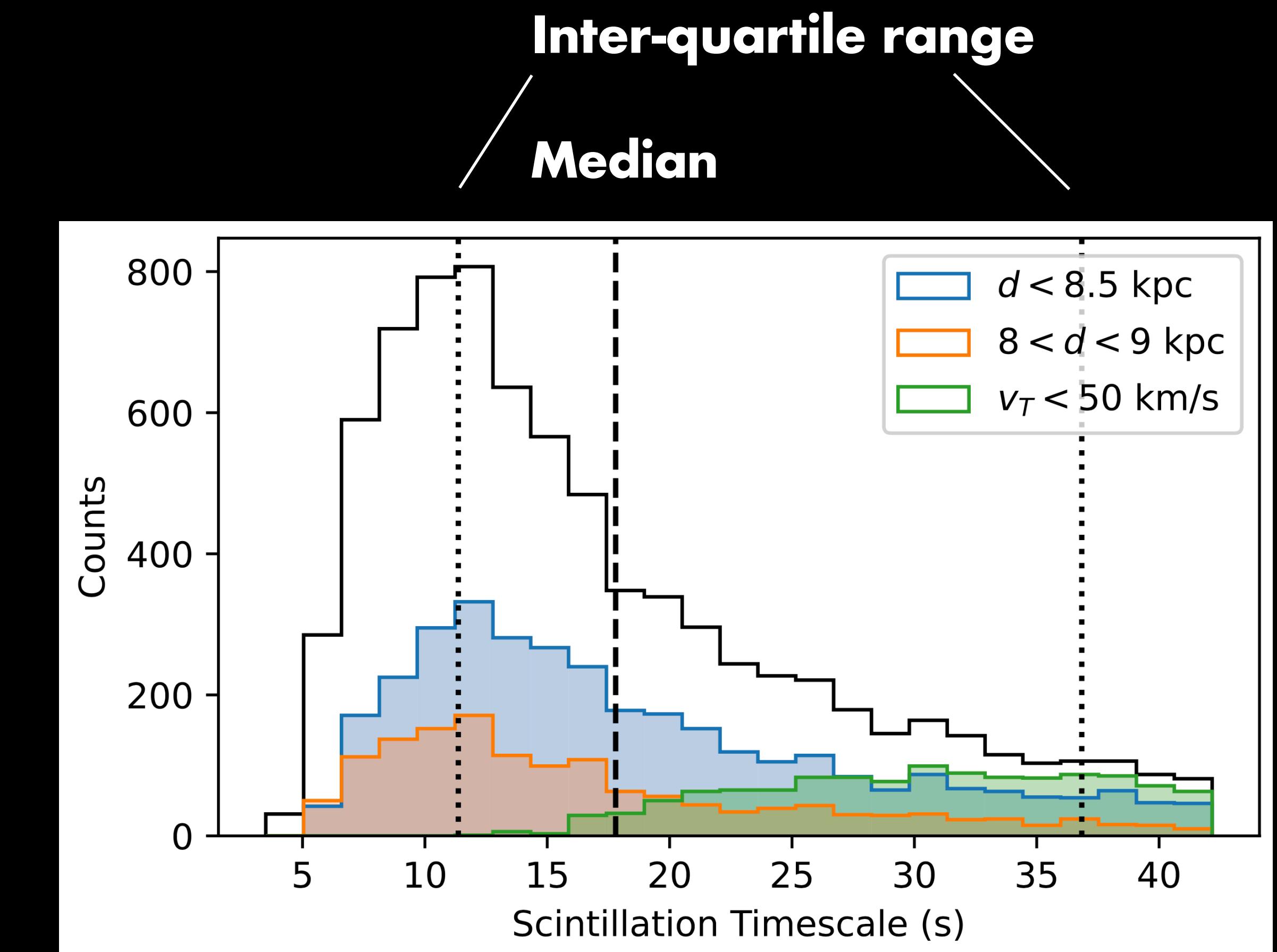
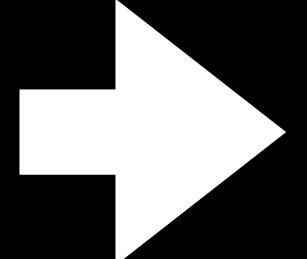
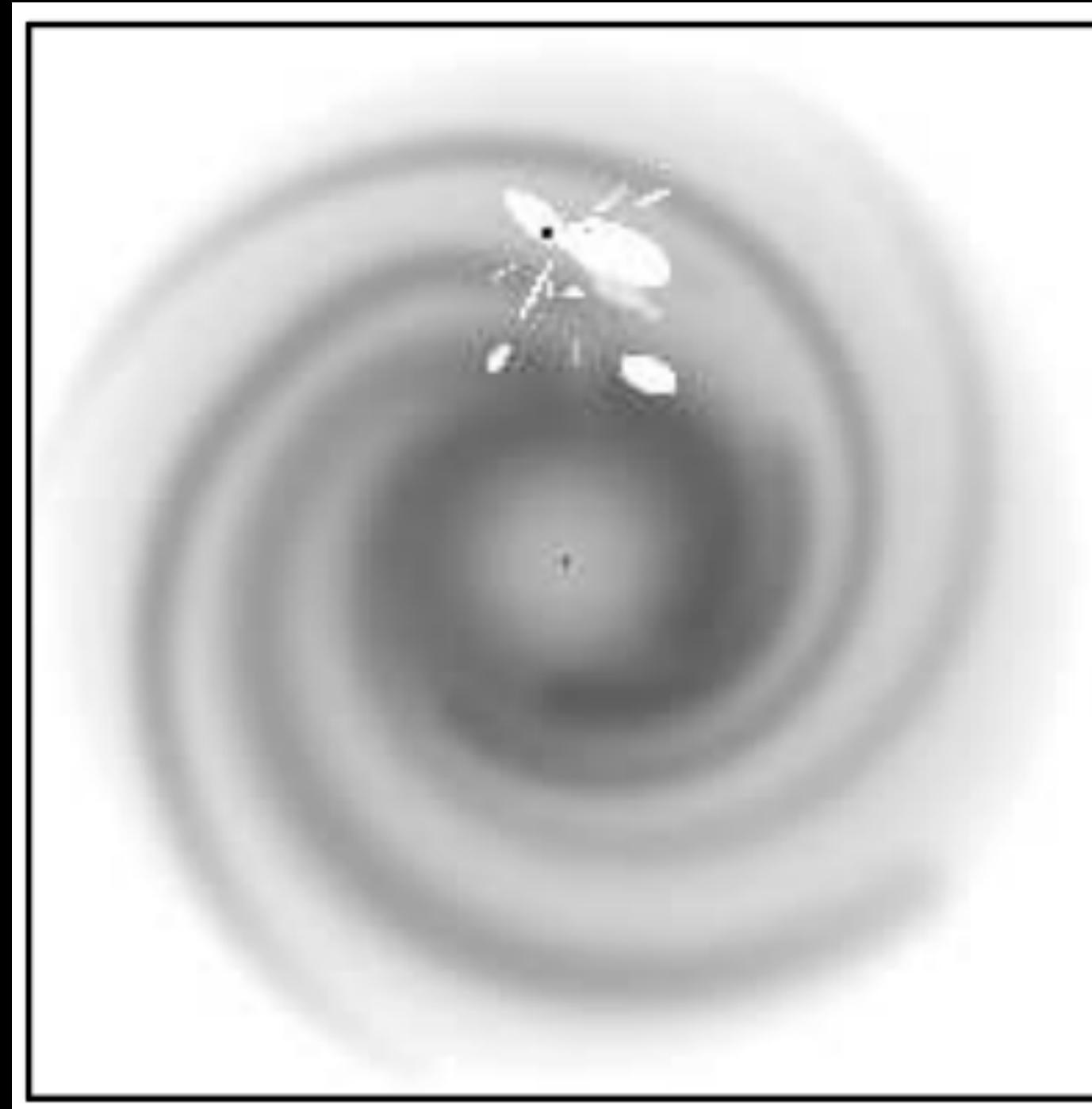
Gajjar et al. 2021

# Planning Galactic center / plane observations

NE2001. I. A NEW MODEL FOR THE GALACTIC DISTRIBUTION  
OF FREE ELECTRONS AND ITS FLUCTUATIONS

J. M. CORDES  
Astronomy Department and NAIC, Cornell University, Ithaca, NY 14853  
cordes@spacenet.tn.cornell.edu

T. JOSEPH W. LAZIO  
Naval Research Lab, Code 7213, Washington, D.C. 20375-5351  
Joseph.Lazio@nrl.navy.mil

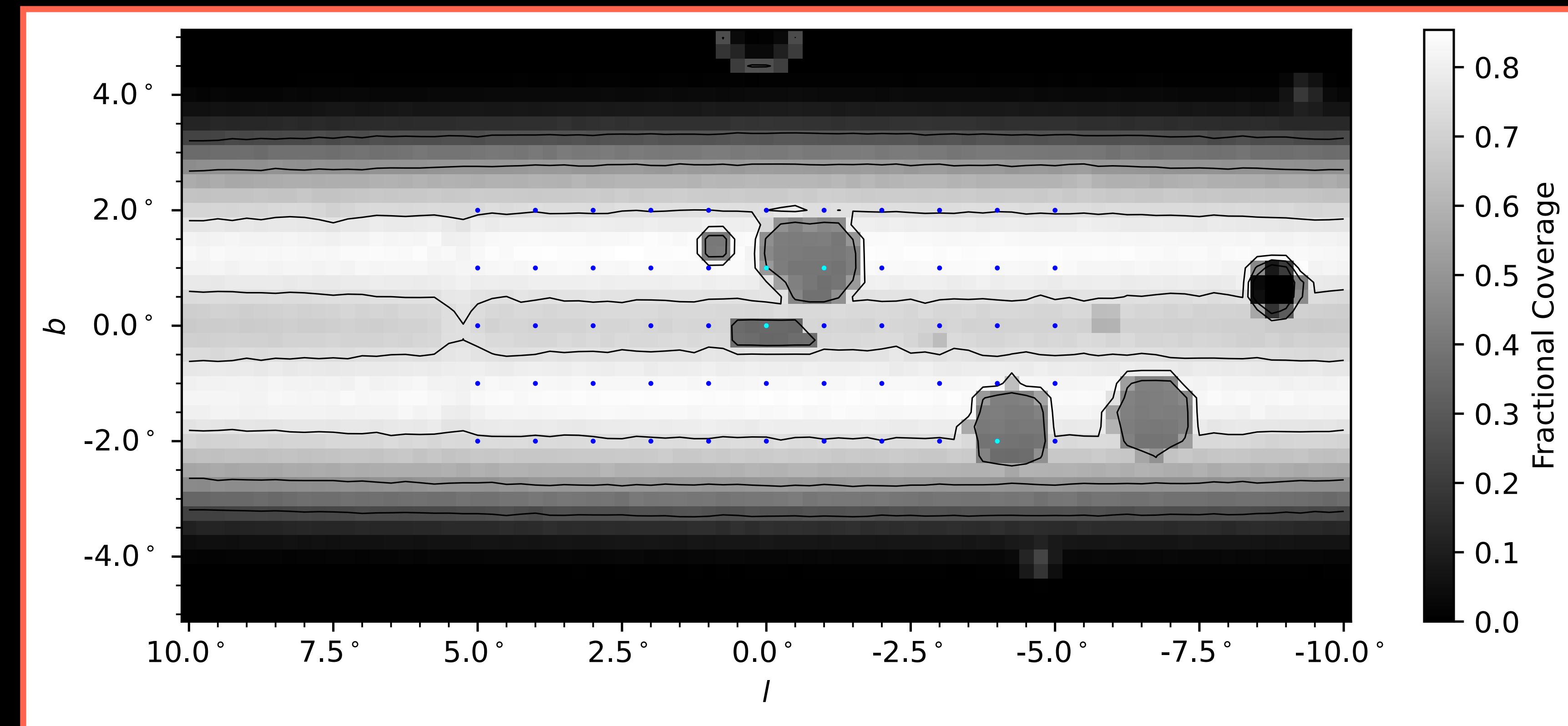


$(l, b) = (5, 0)$  at C-band

$$\Delta t_d \propto \nu^{6/5} v_T^{-1}$$

# Planning Galactic center / plane observations

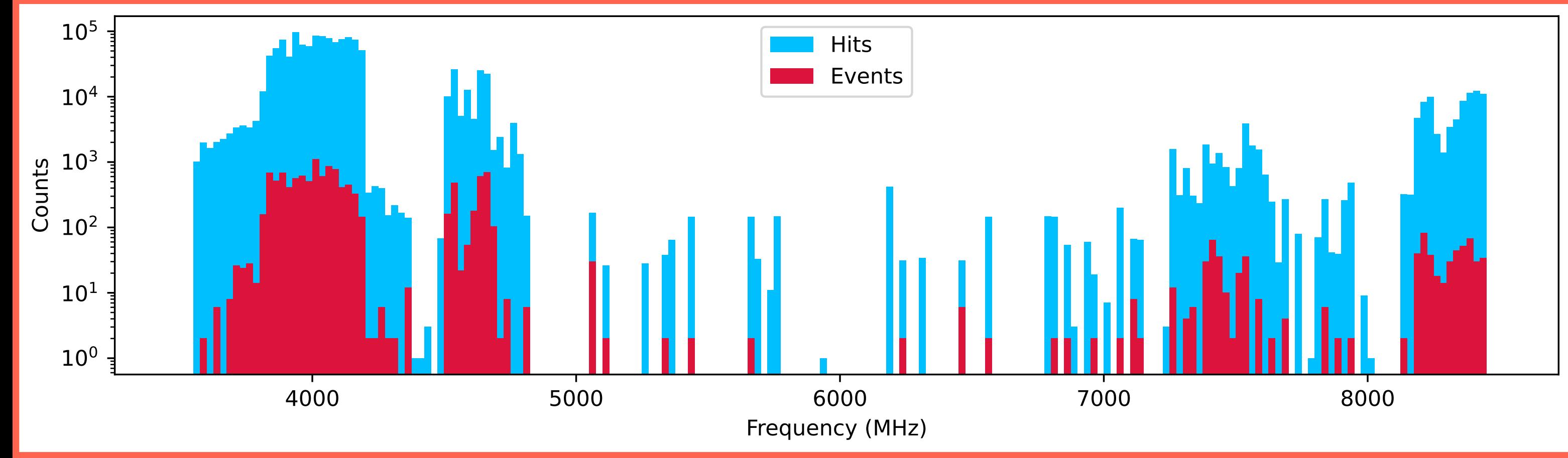
Fractional coverage for scintillation  
timescales between 10 s and 100 s

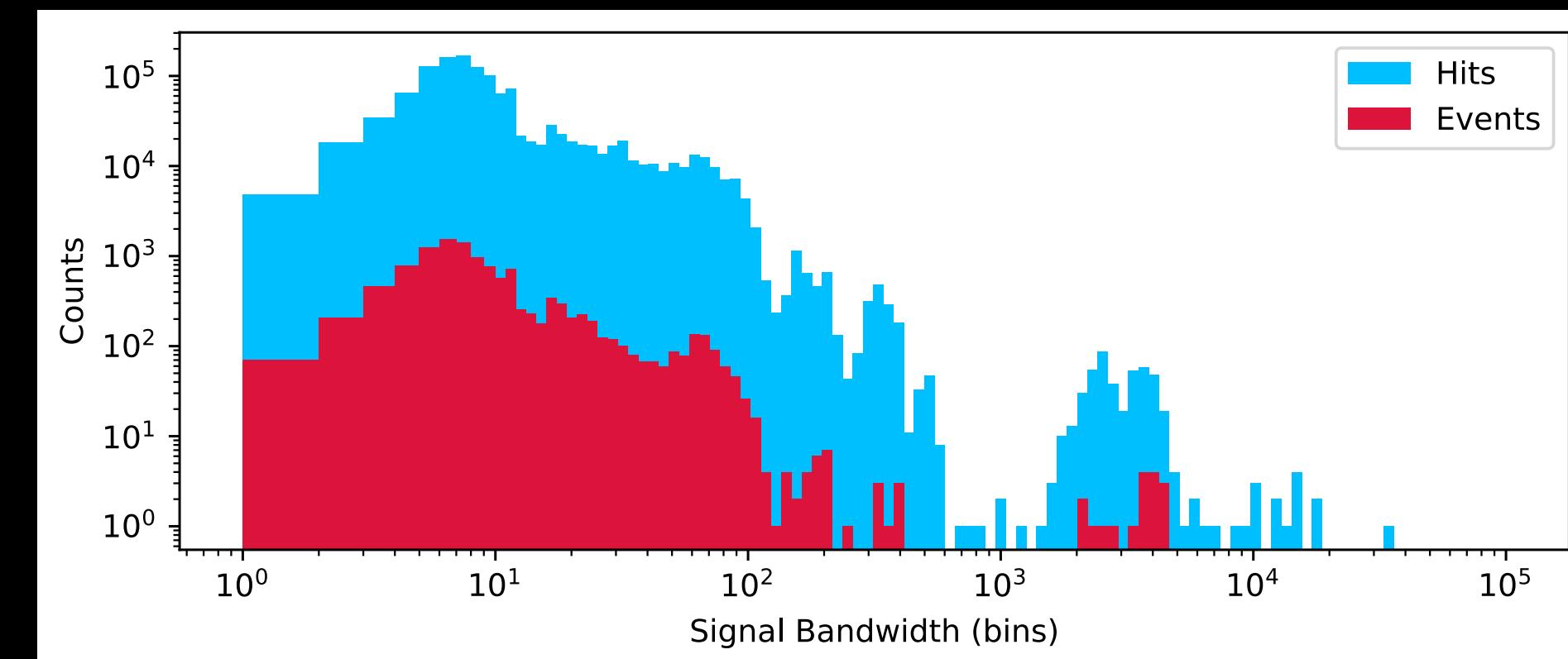
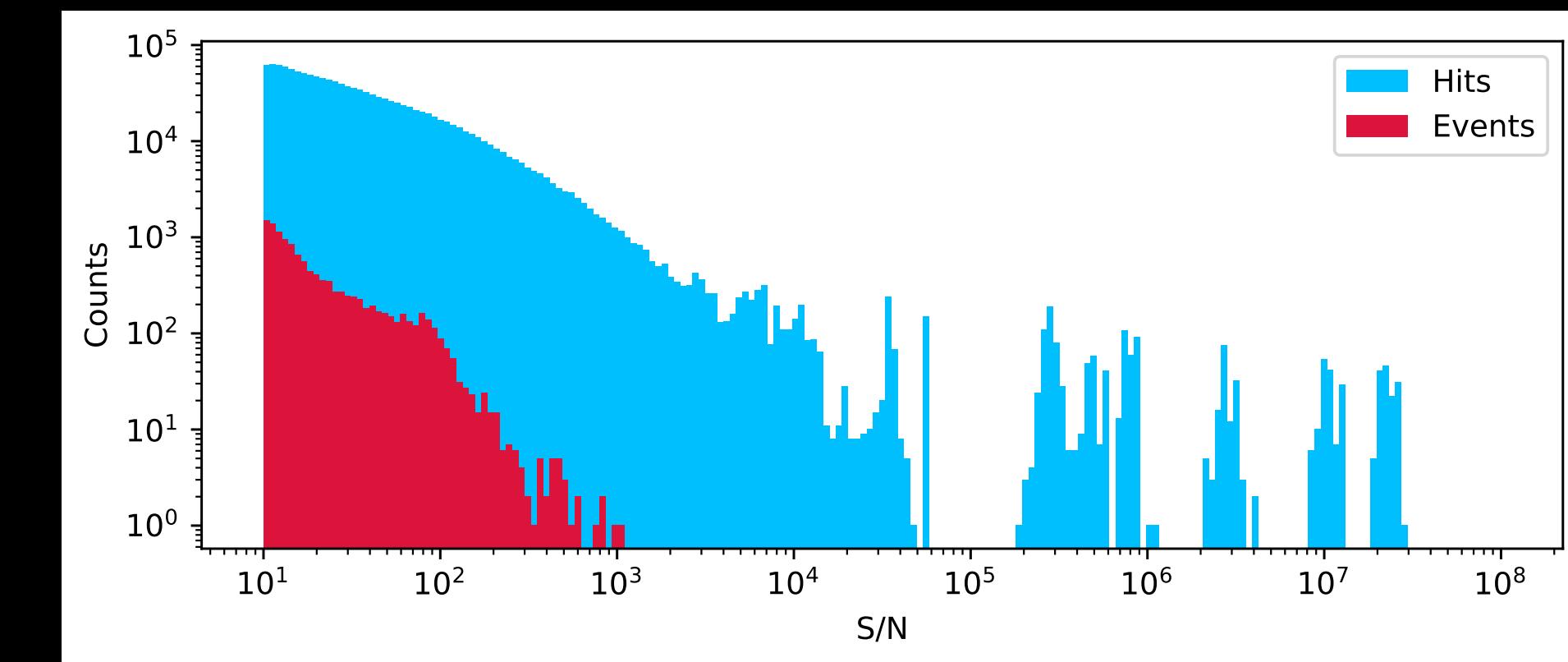
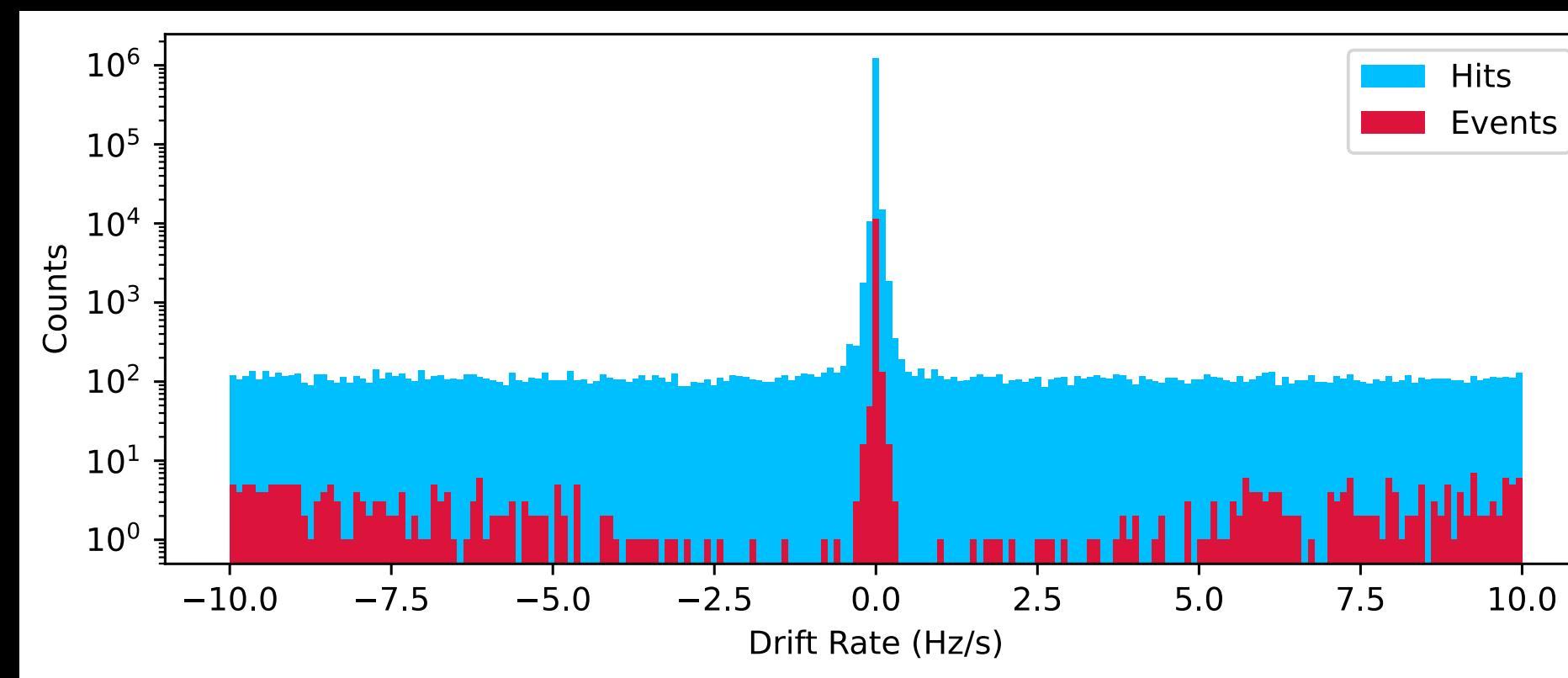
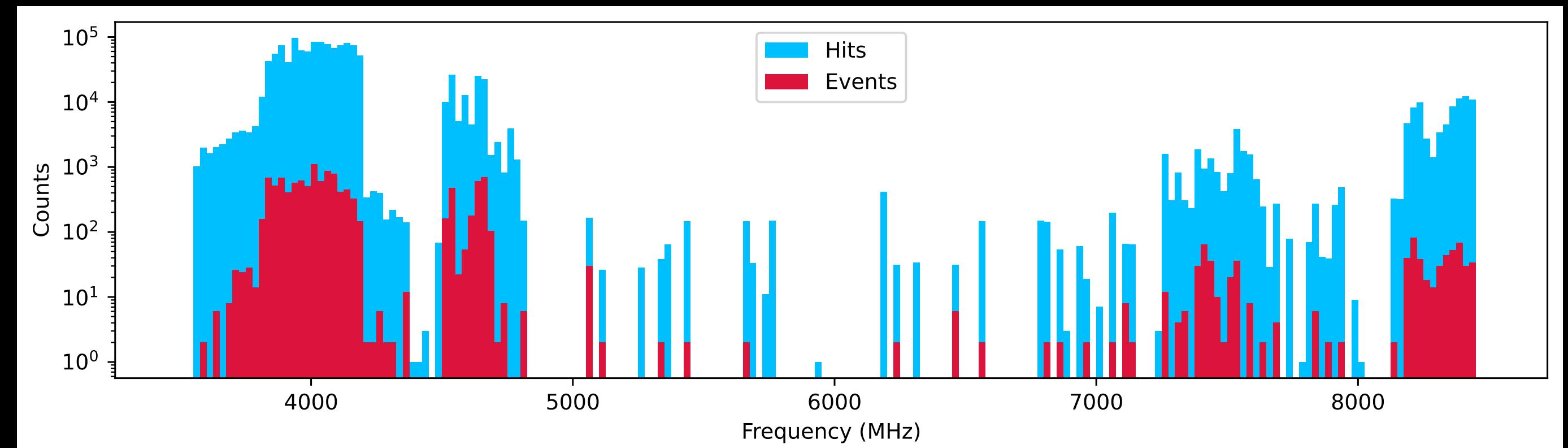


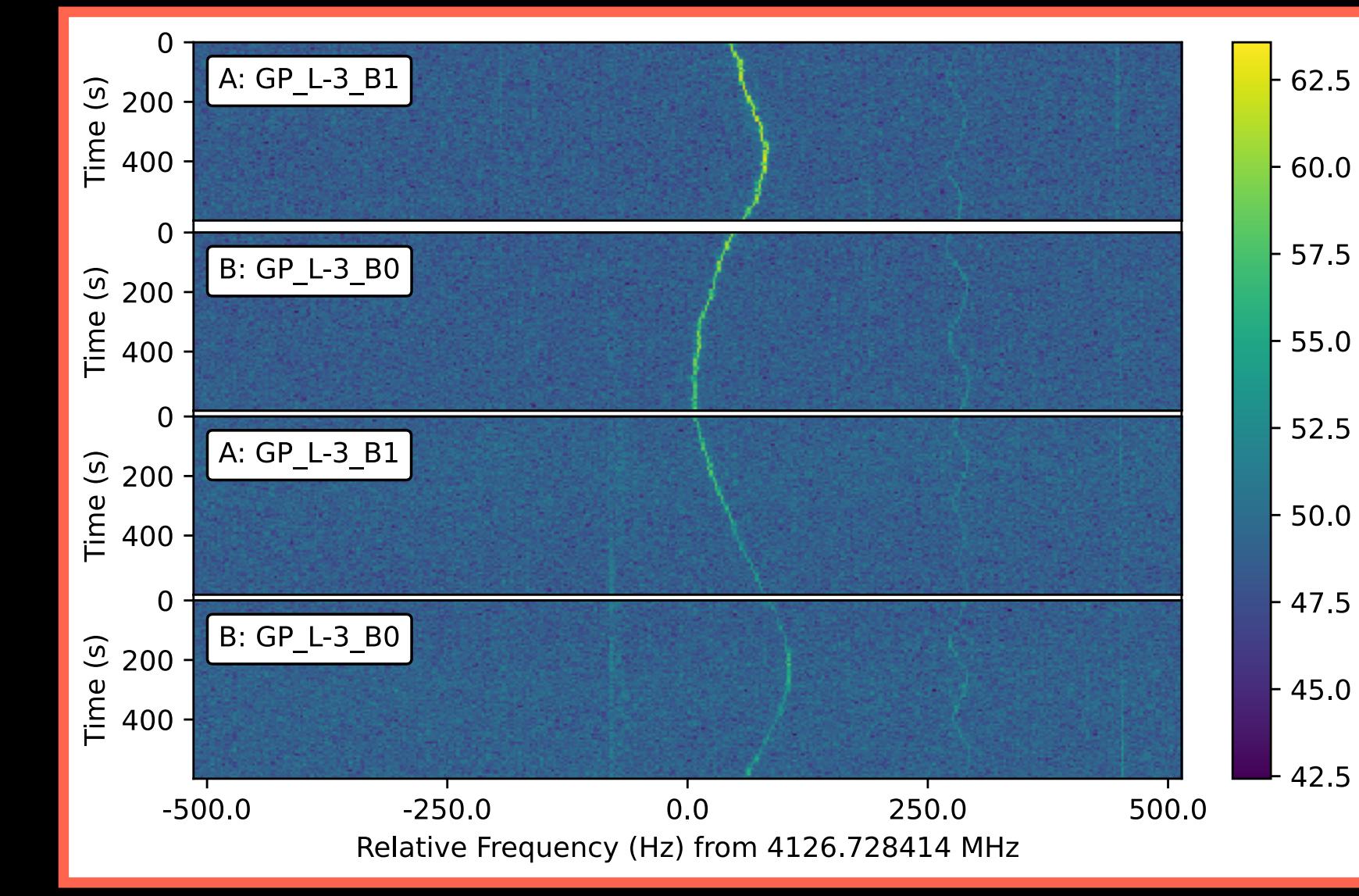
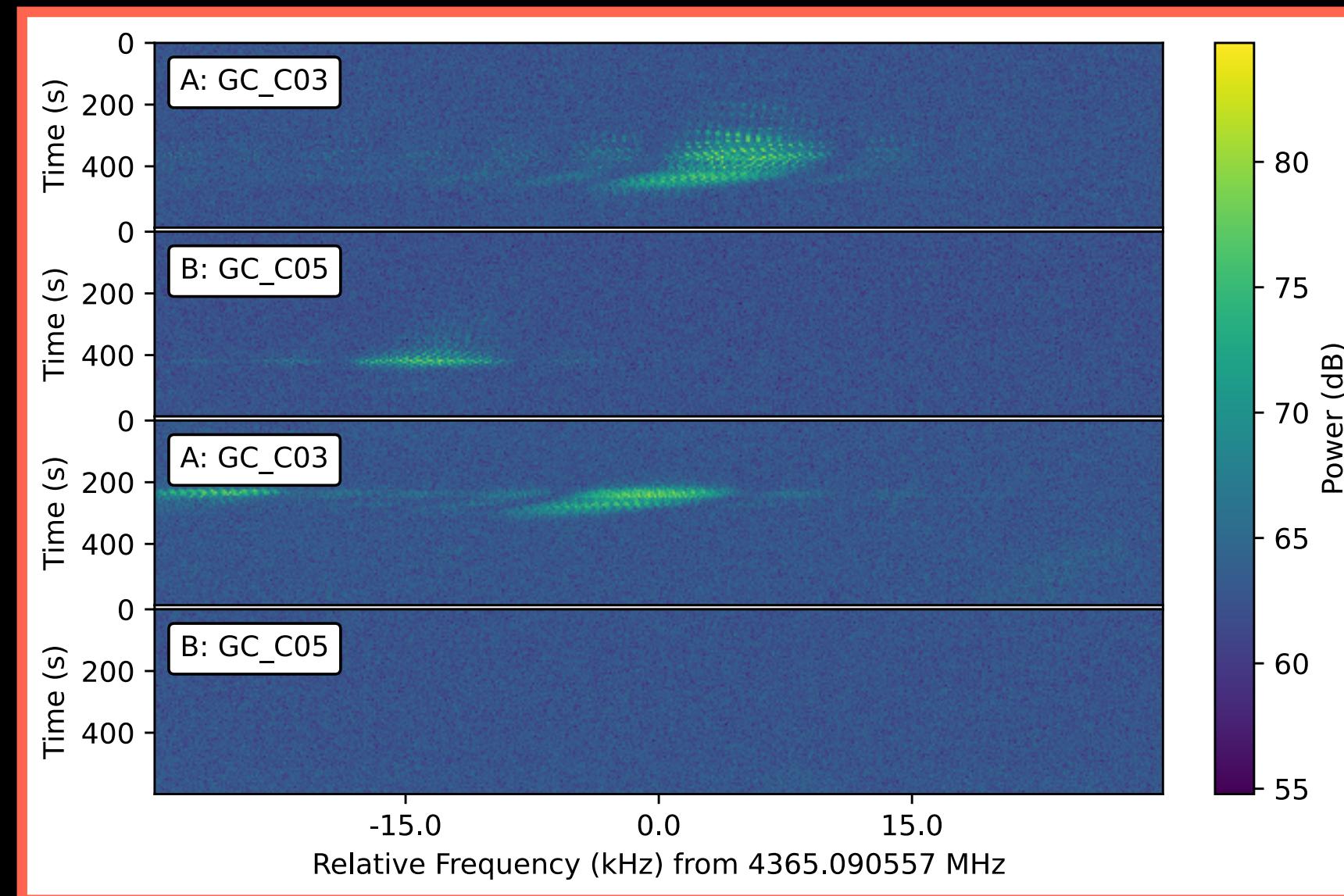
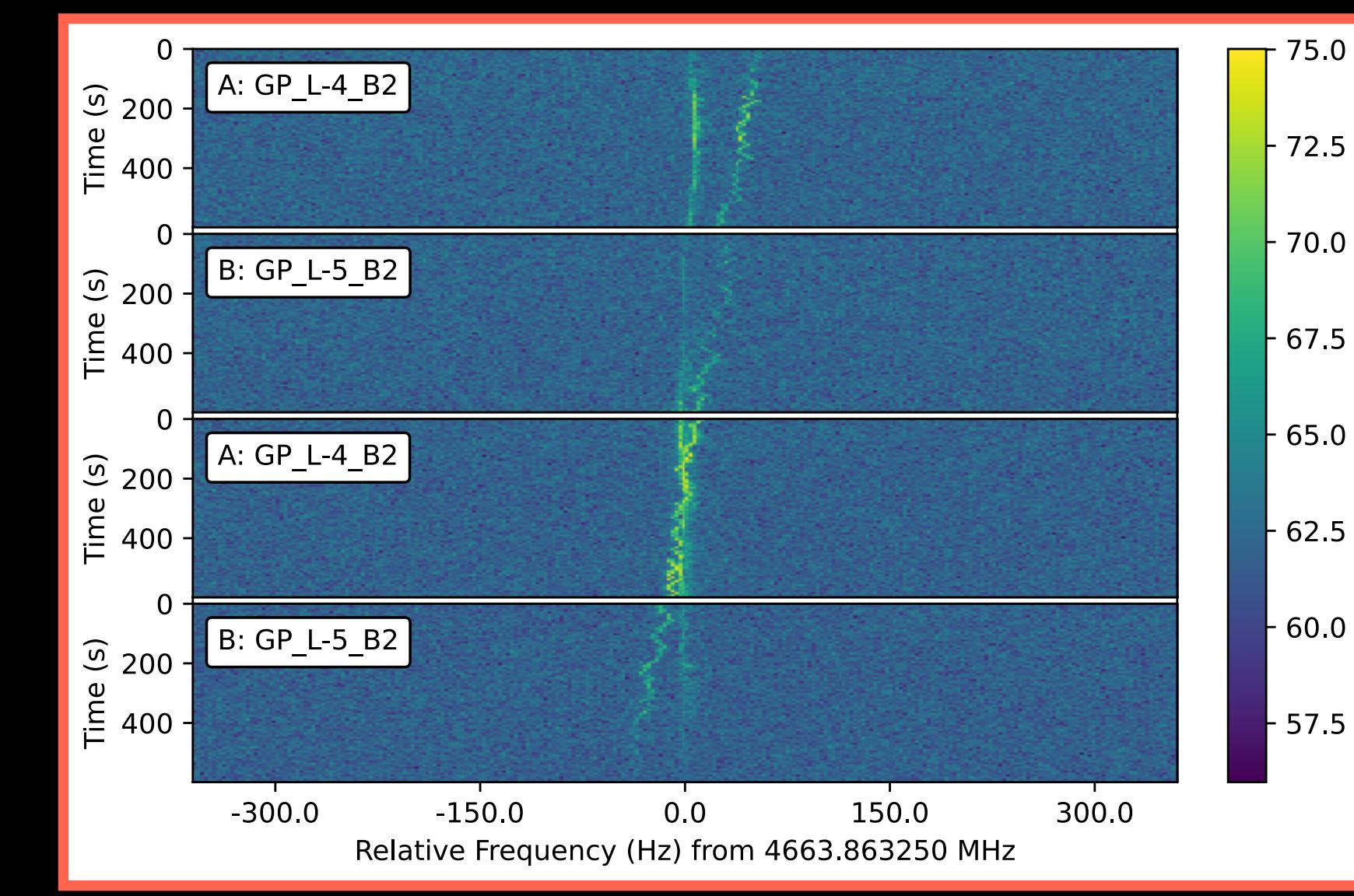
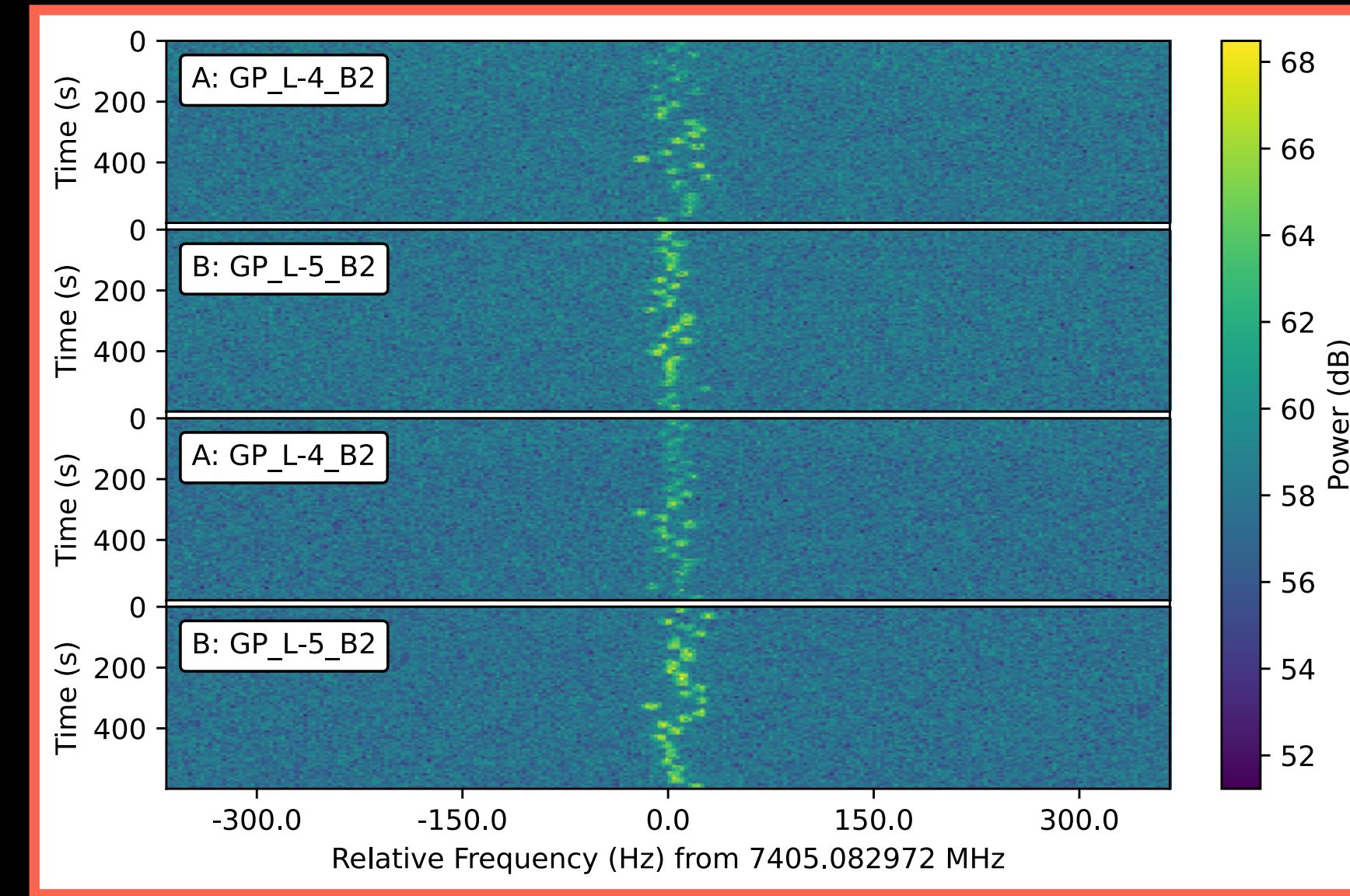
Brzycki et al. (in prep)

# Scintillation survey towards the Galactic center / plane

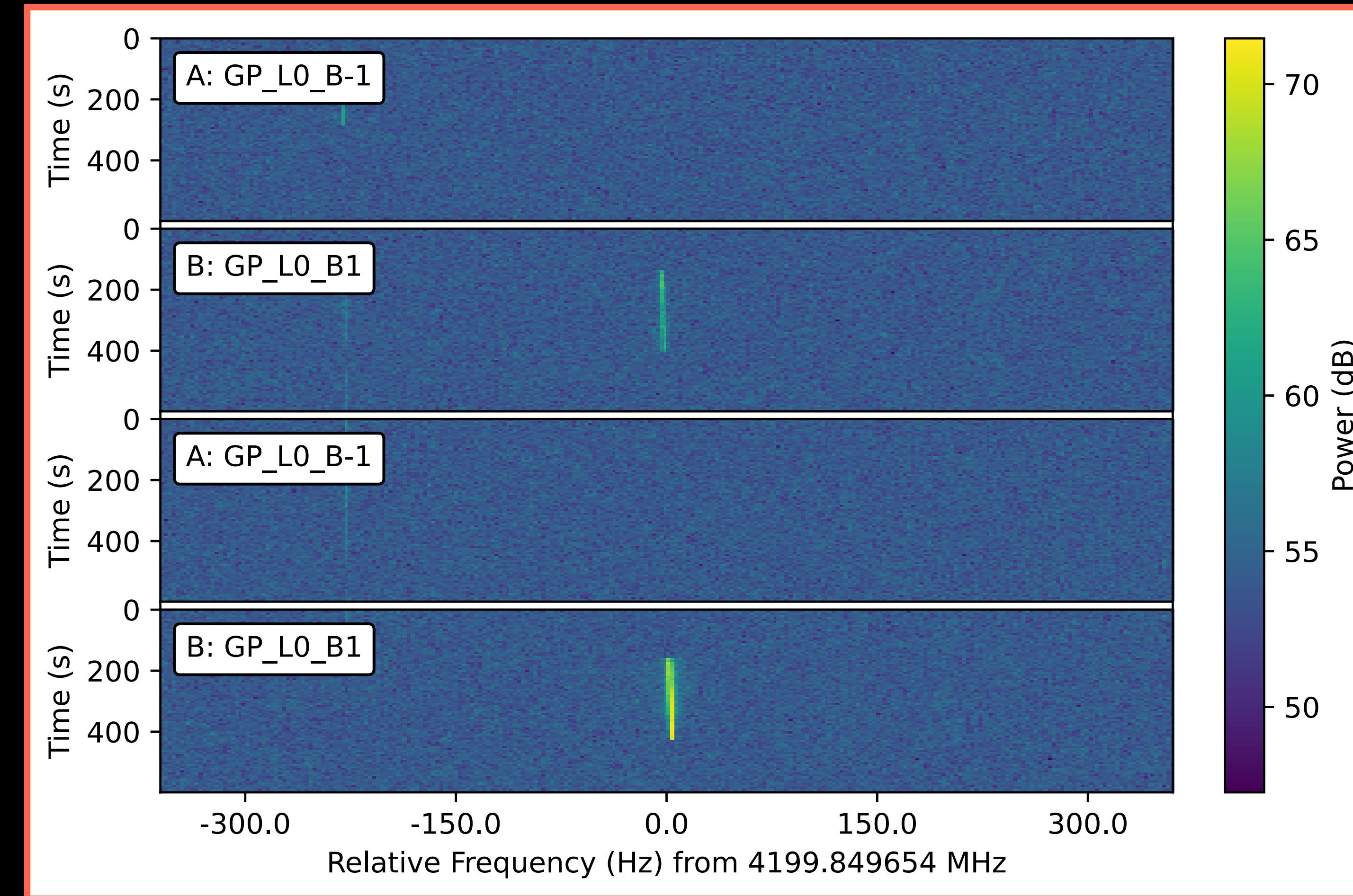
- ABAB cadences
- 10 minutes per observation
- 2.5 s, 2.8 Hz resolution
- 73 pointings in total, for about 24.3 hr of observations
- 1.3 million detected signals, 6018 “events” passed ON-OFF filtering

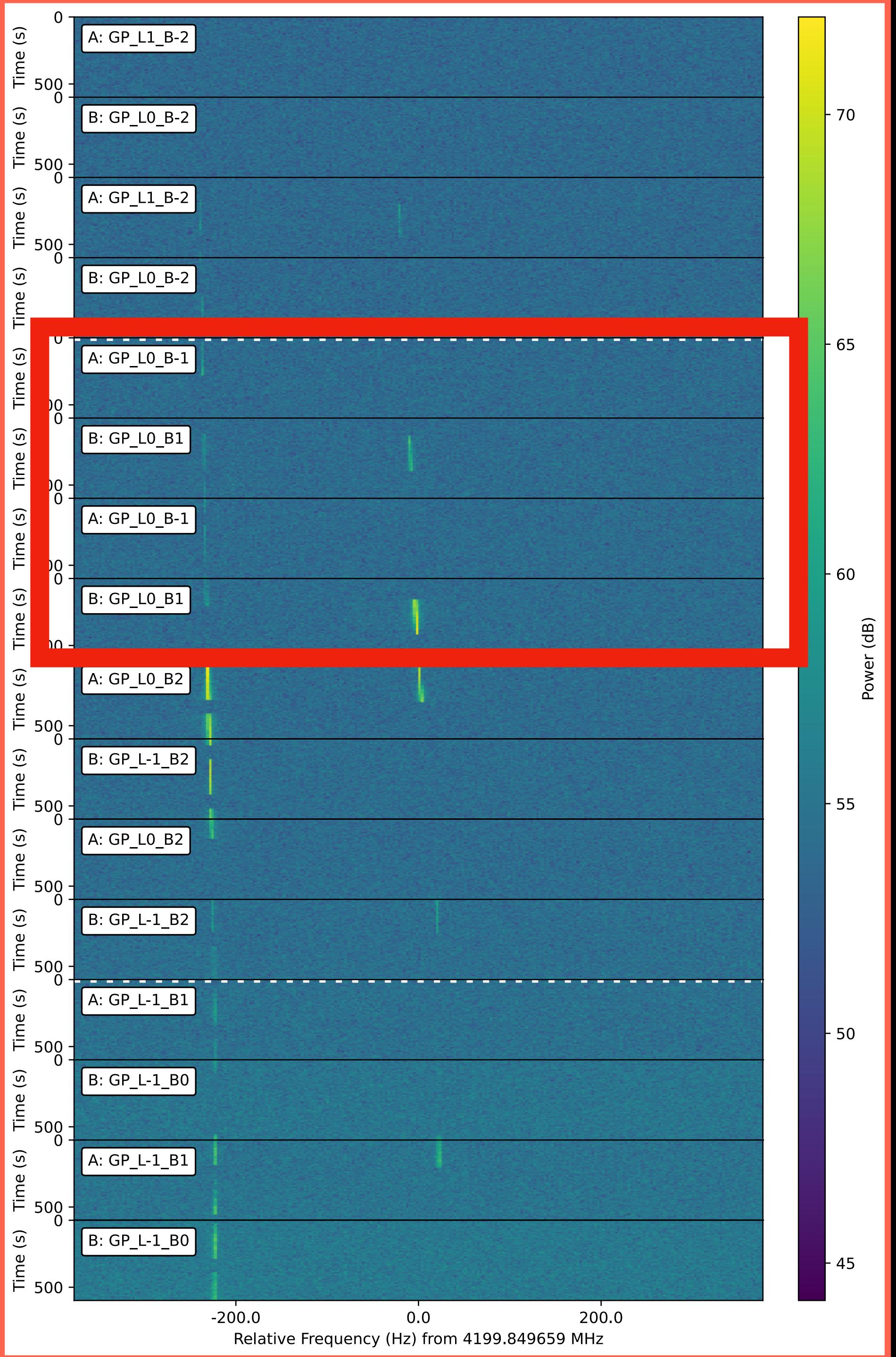




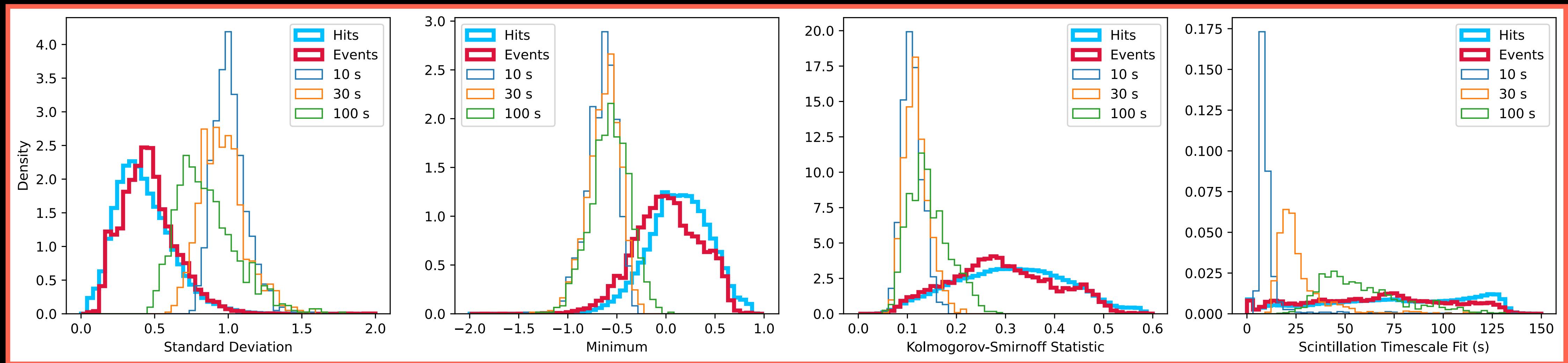


# Best candidate event we found





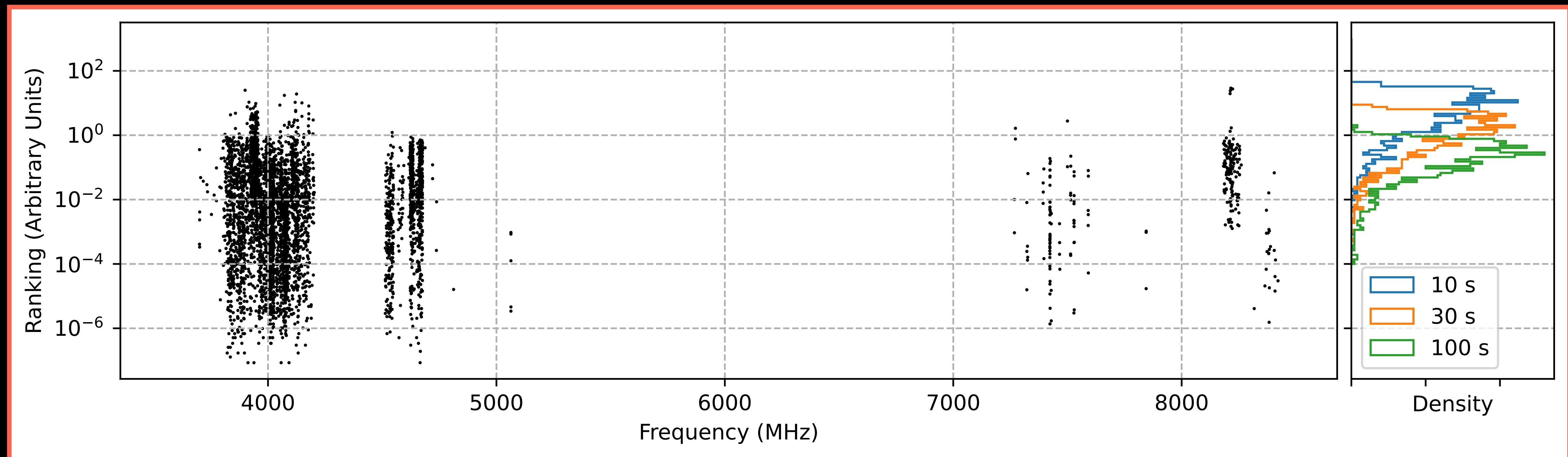
# Manual inspection turned up no true candidates... so can RFI look like scintillation?



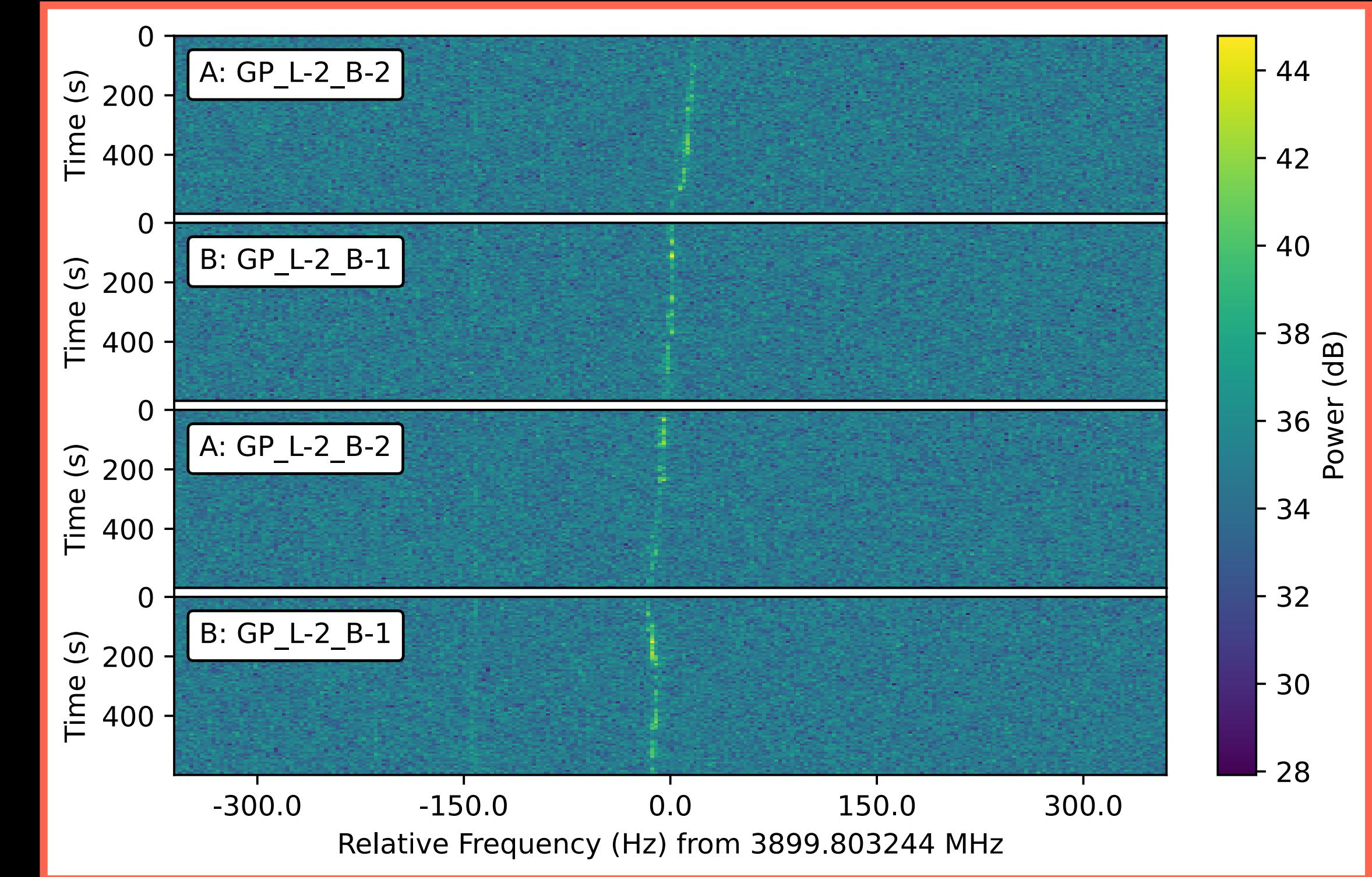
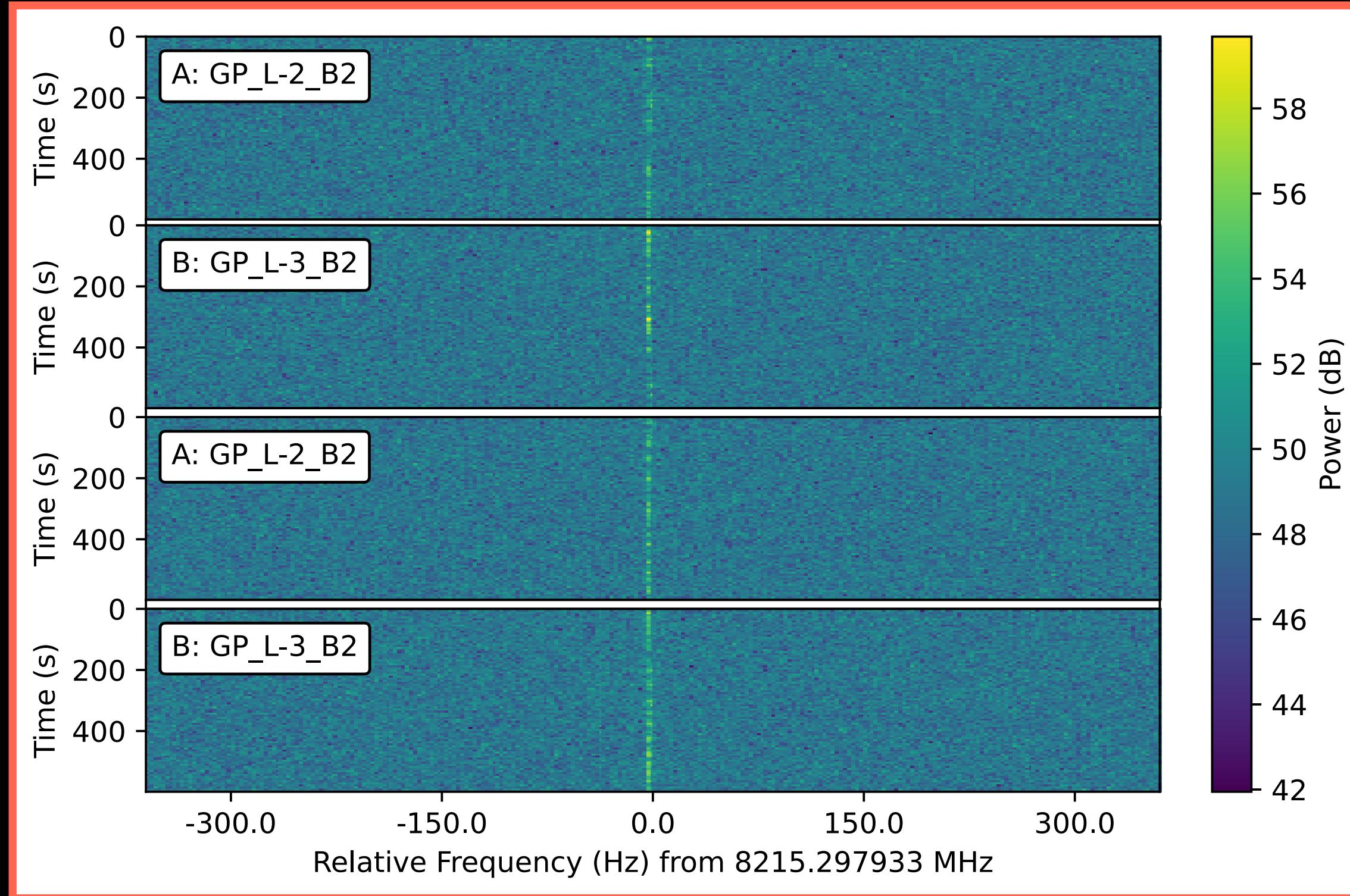
Empirical PDFs for detected signals vs. injected synthetic scintillated signals with timescales of 10, 30, 100 s

# Simple score for ranking signals

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



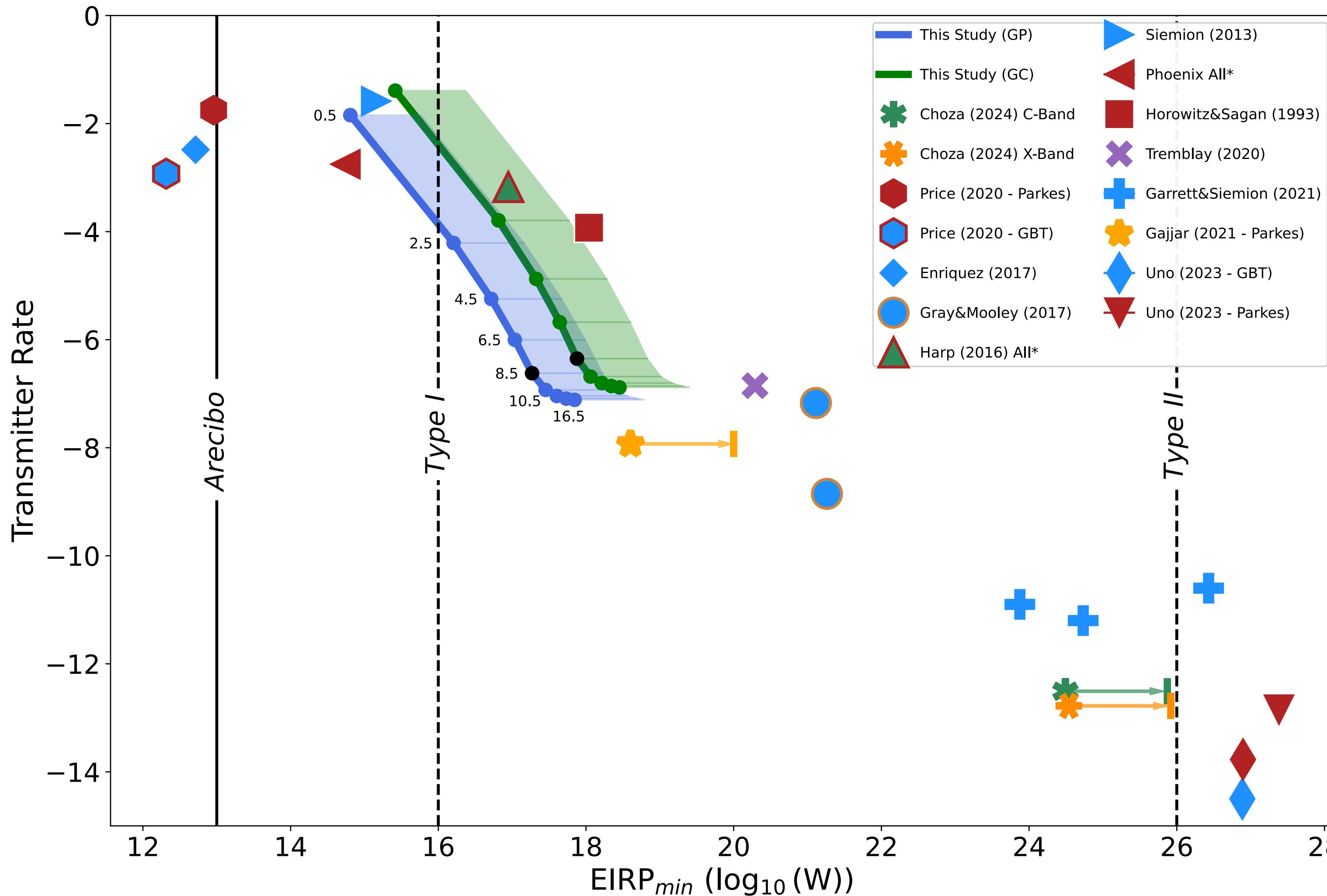
# Example events with high scintillation scores



# Some takeaways

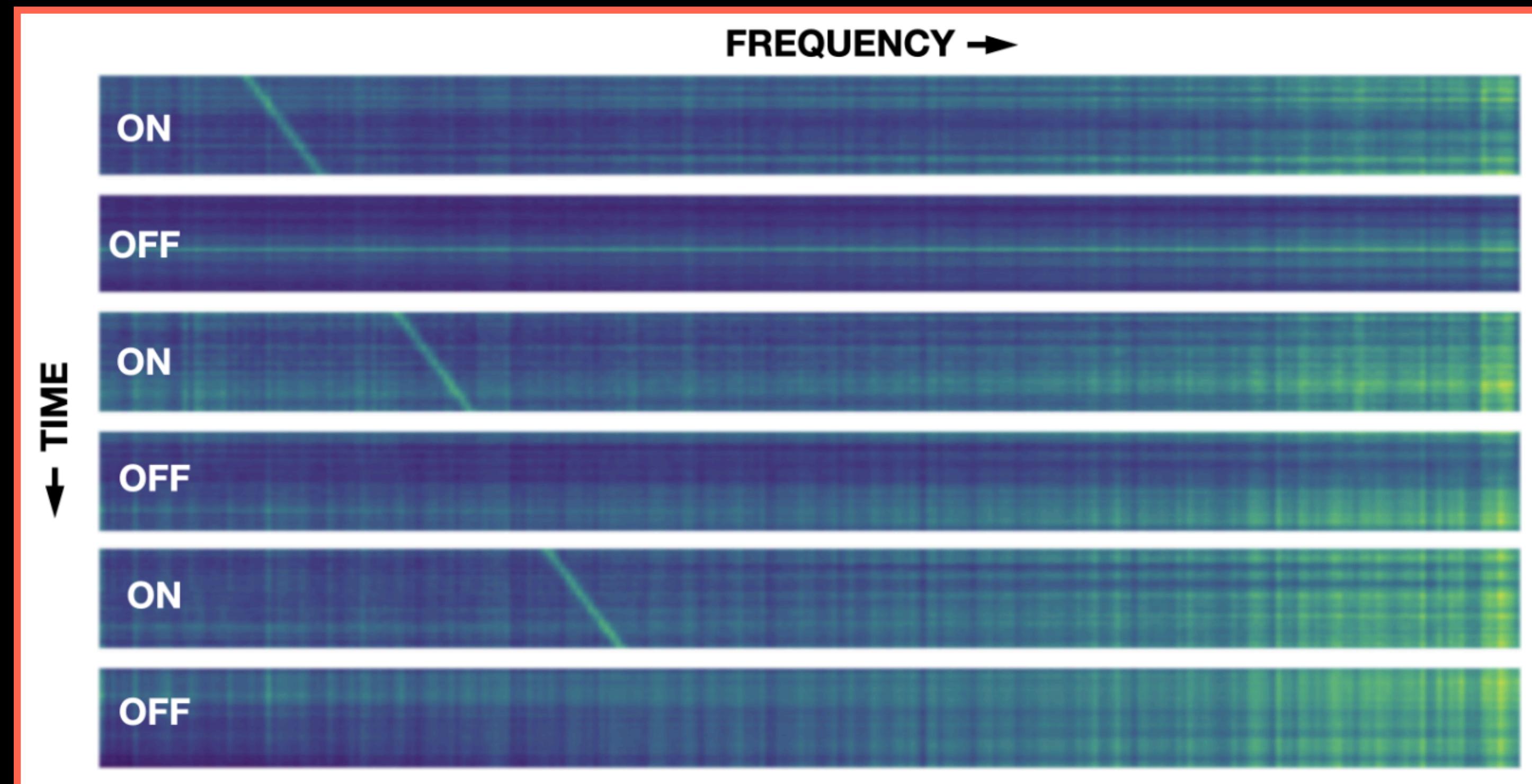
- Unfortunately, no candidates, scintillated or otherwise, passed all filters in our survey of the Galactic center and Galactic plane
- For sources up to a distance of 8.5 kpc, we set a limit for the equivalent isotropic radiated power of  $\sim 2e17$  W (1e4 times larger than Arecibo)
- Until we get a much deeper understanding of the RFI environment around our telescopes, the ON-OFF filter will likely remain the most reliable indicator

# Extra Slides



# Other people actually use Setigen!

- Injection – recovery testing
- Designing multibeam search surveys
- Development of end-to-end software for the Allen Telescope Array
- ML dataset production (e.g. Kaggle competition)

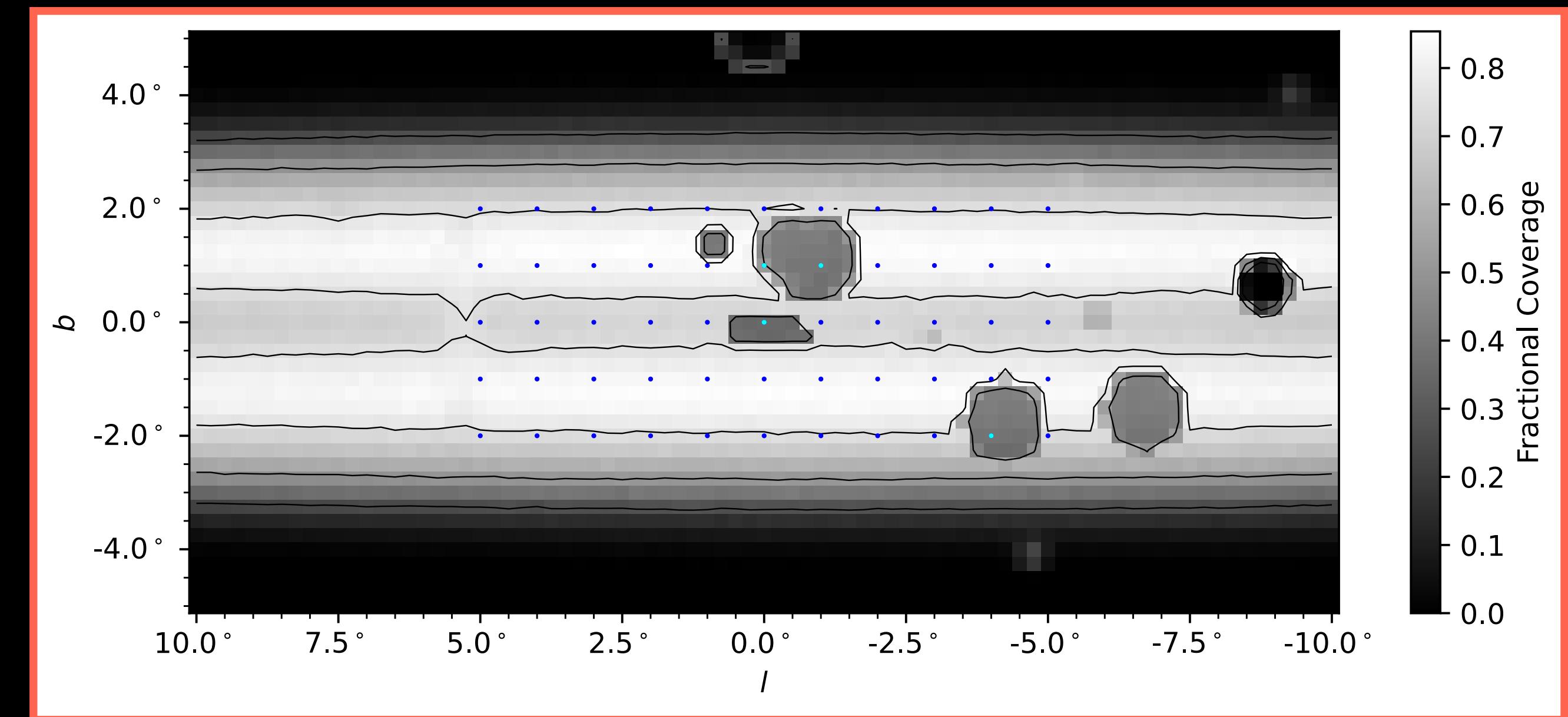


Breakthrough Listen x Kaggle 2021

# Planning Galactic center / plane observations

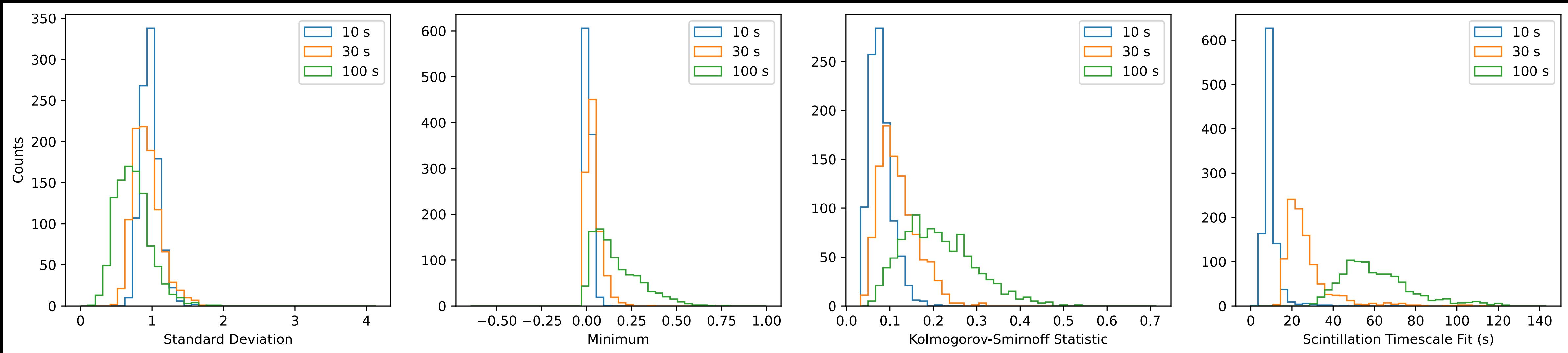
- Estimate most likely scintillation timescales using NE2001 model (Cordes & Lazio 2002) and scale by Monte Carlo sampling different possible physical configurations

**Fractional coverage for scintillation timescales between 10 s and 100 s**



# Statistics using synthetic scintillated intensities (no noise)

**600 s “observation”, 4.65 s resolution**



**Standard Deviation**

**Minimum**

**Kolmogorov-Smirnov Statistic**

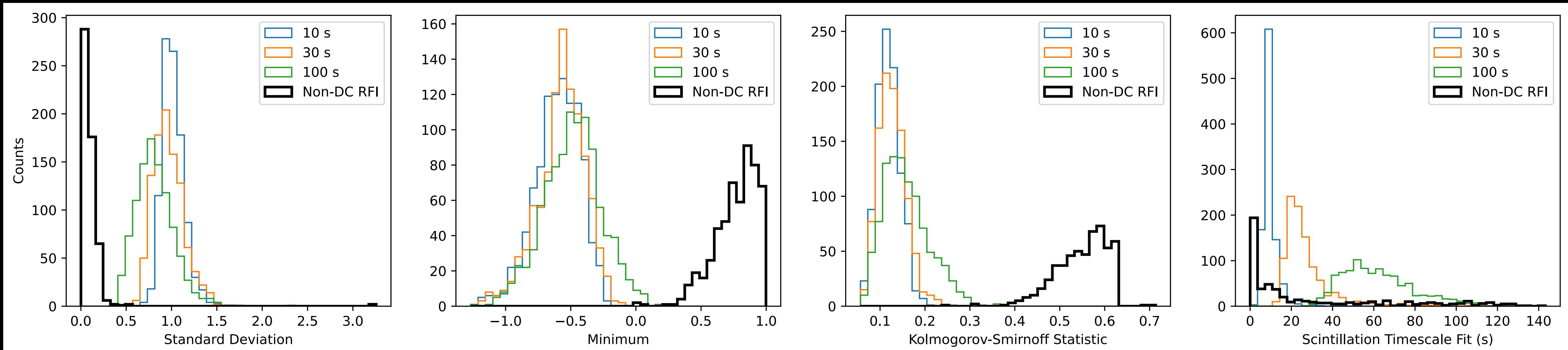
**Scintillation Timescale Fit (s)**

# GBT RFI vs. injected synthetic scintillated signals



**S/N > 25**

**C band (4 – 8 GHz)**



**Standard Deviation**

**Minimum**

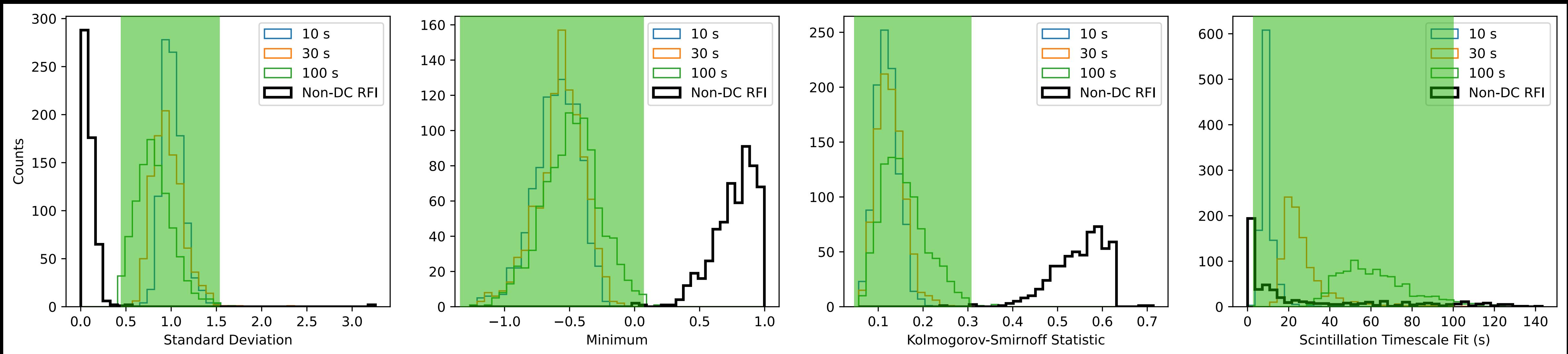
**Kolmogorov-Smirnov Statistic**

**Scintillation Timescale Fit**

# GBT RFI vs. injected synthetic scintillated signals

C band (4–8 GHz)

S/N > 25



Standard Deviation

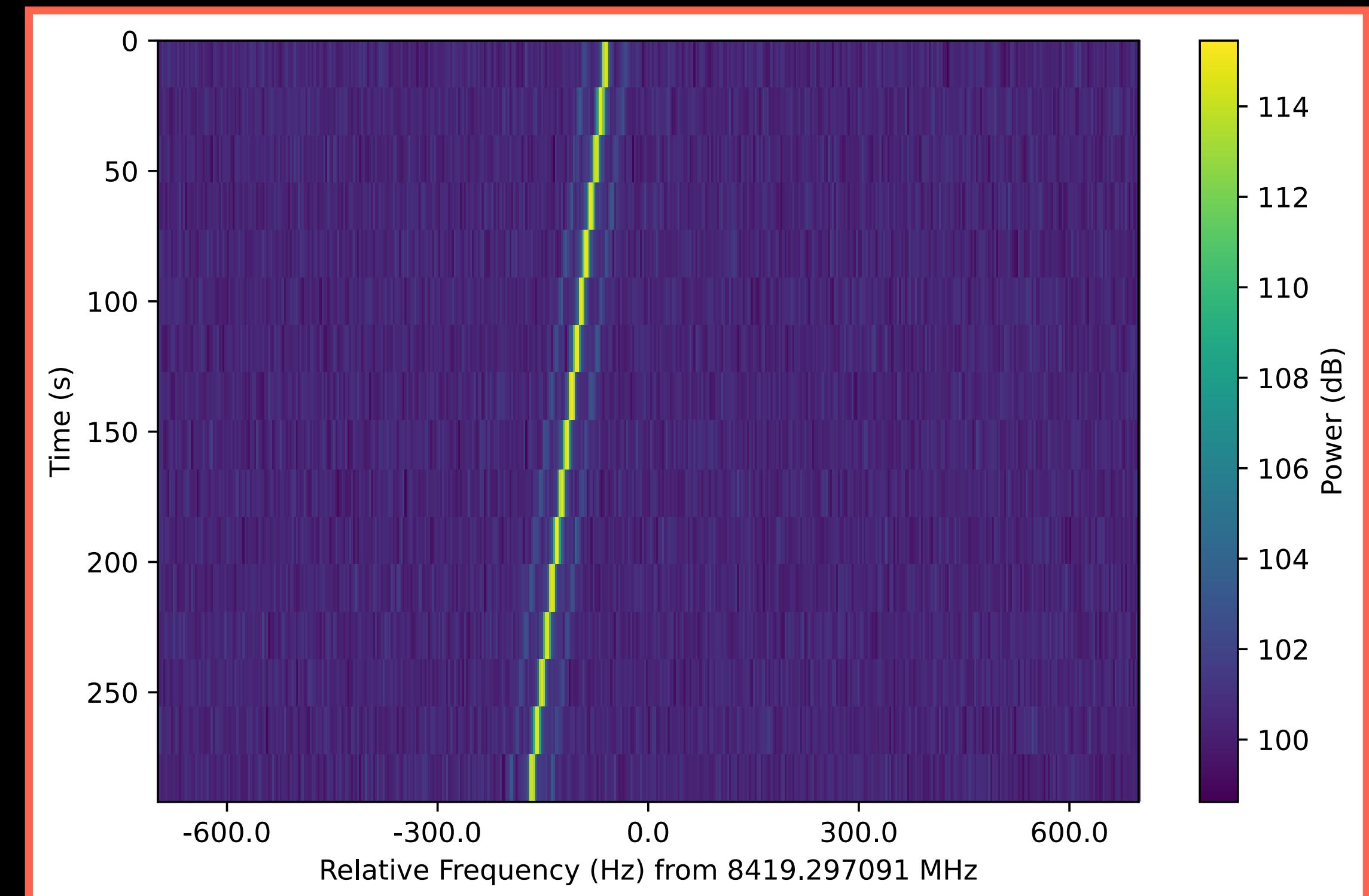
Minimum

Kolmogorov-Smirnoff Statistic

Scintillation Timescale Fit

# Basic signal detection

- Incoherent deDoppler (TurboSETI)
- Energy detection
- Machine learning (ML)

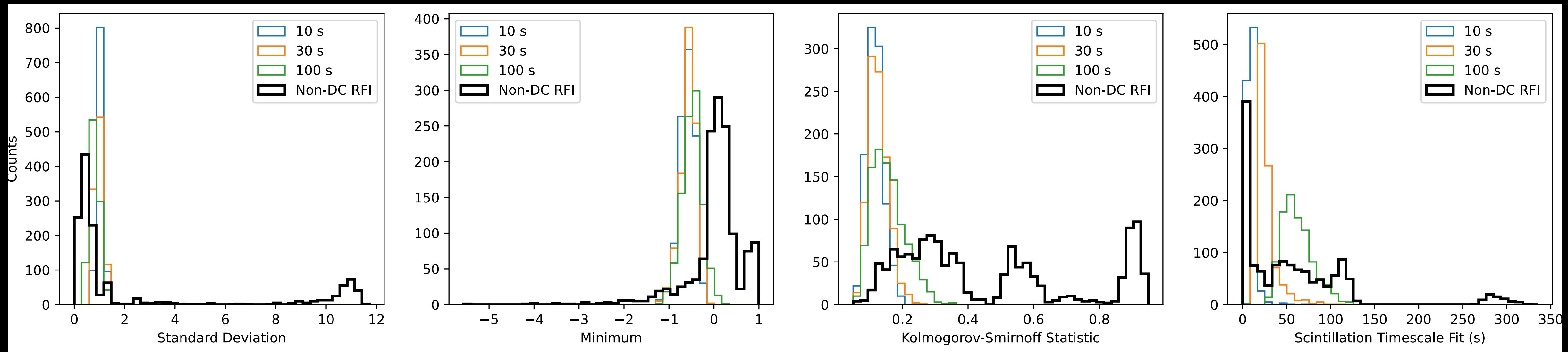


**Detected signal from Voyager 1**

# GBT RFI vs. injected synthetic scintillated signals

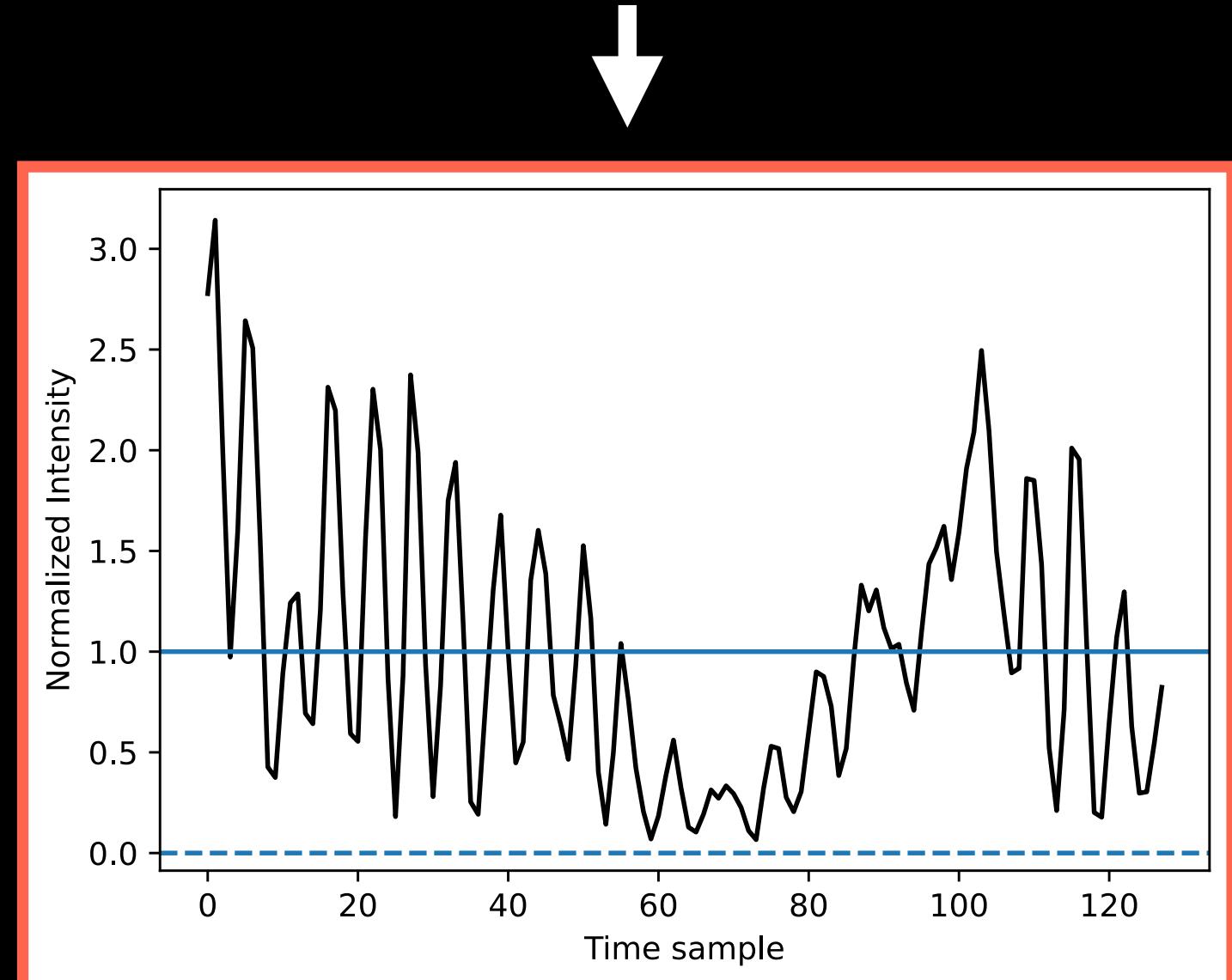
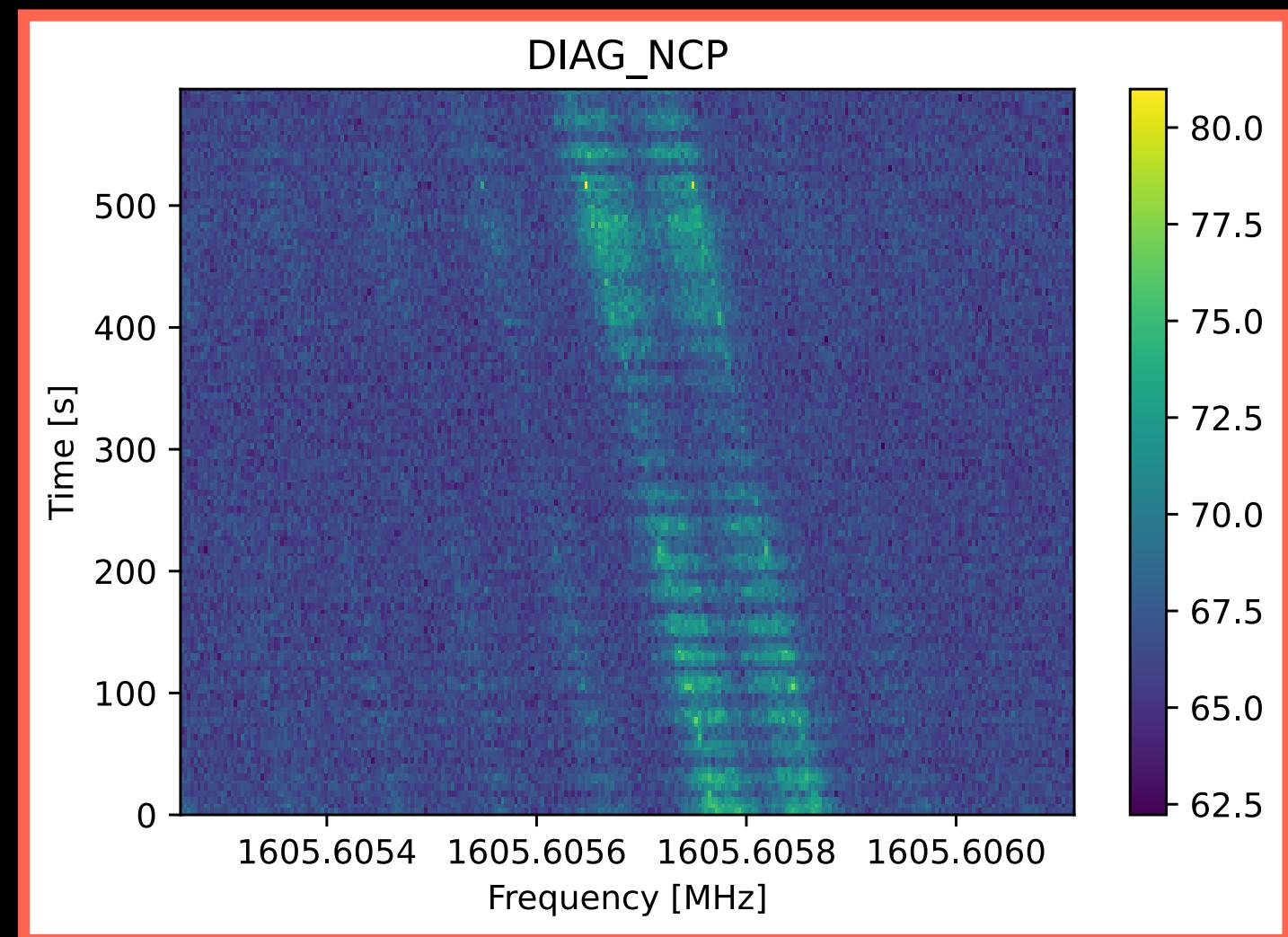
L band (1–2 GHz)

S/N > 25



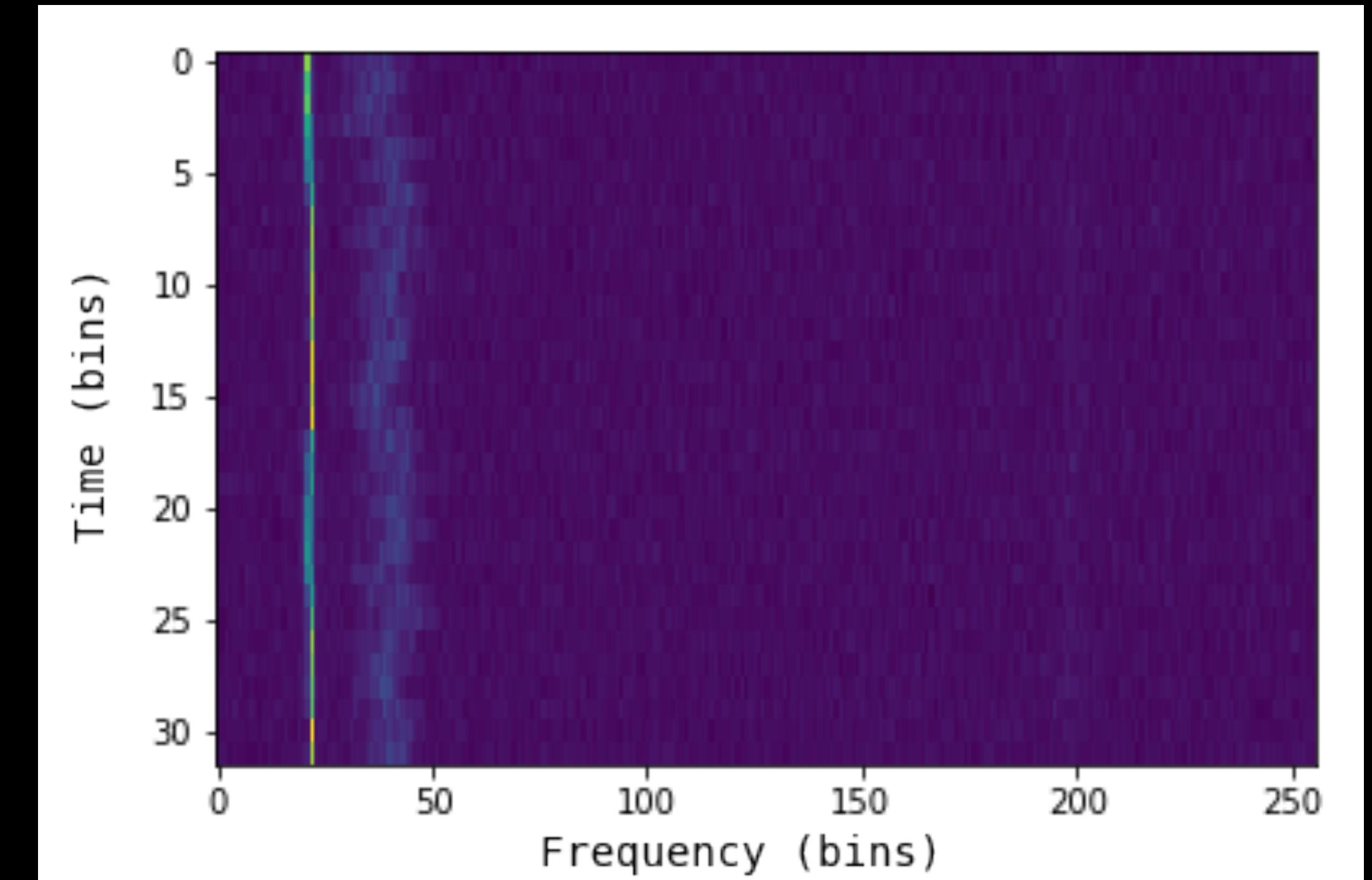
# How might we detect scintillations?

- Estimate intensity time series from signals detected with deDoppler methods
- Since scintillation is stochastic, identify summary statistics that probe asymptotic behavior
- Would existing RFI modulation confound real scintillation?
  - Create synthetic scintillated time series
  - Compare statistics of detected signals with those of synthetic scintillated signals



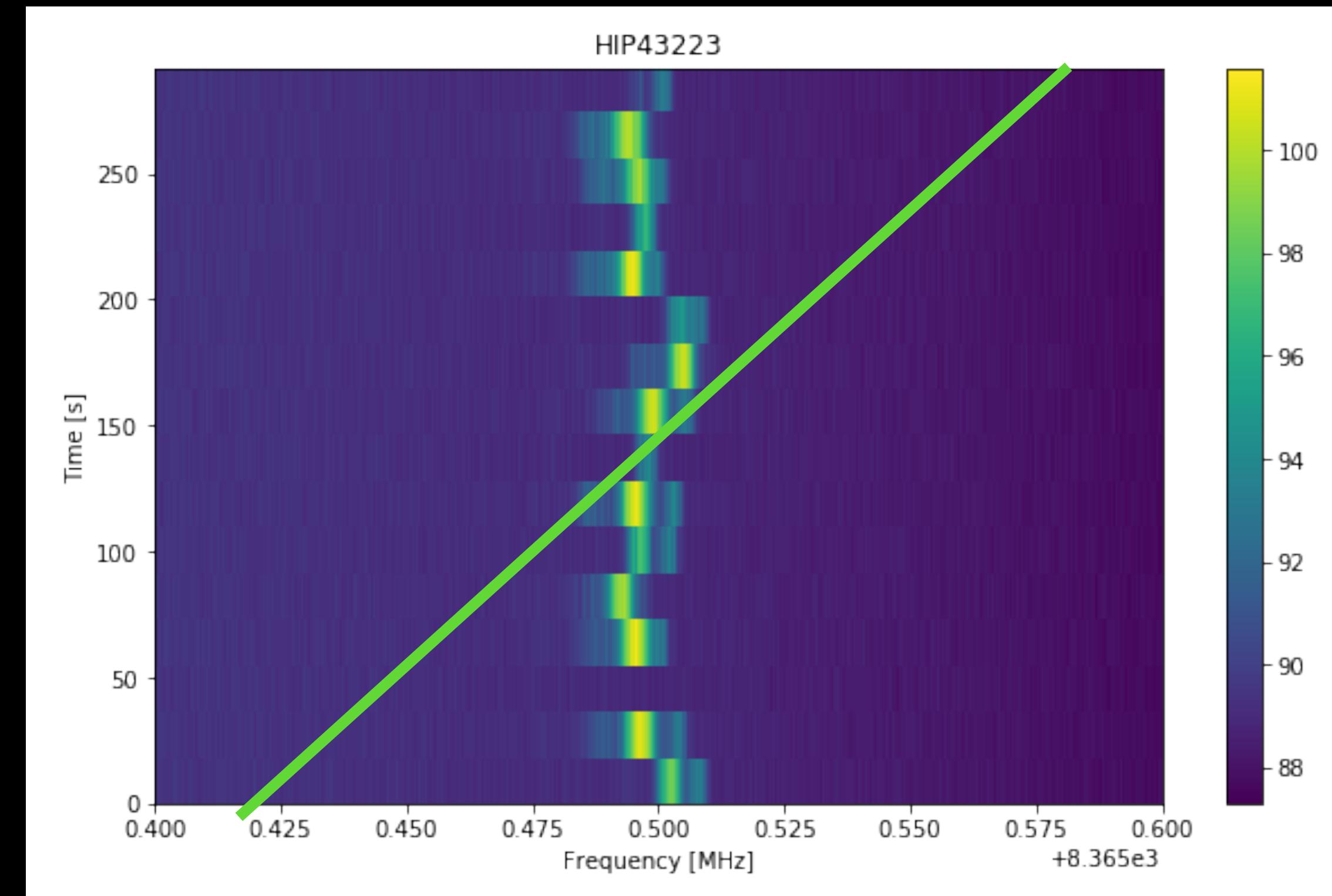
# Narrowband signal localization with machine learning

- Standard deDoppler pipeline:
  - Dim signals concealed by nearby bright signals
  - Computationally expensive to search high drift rates

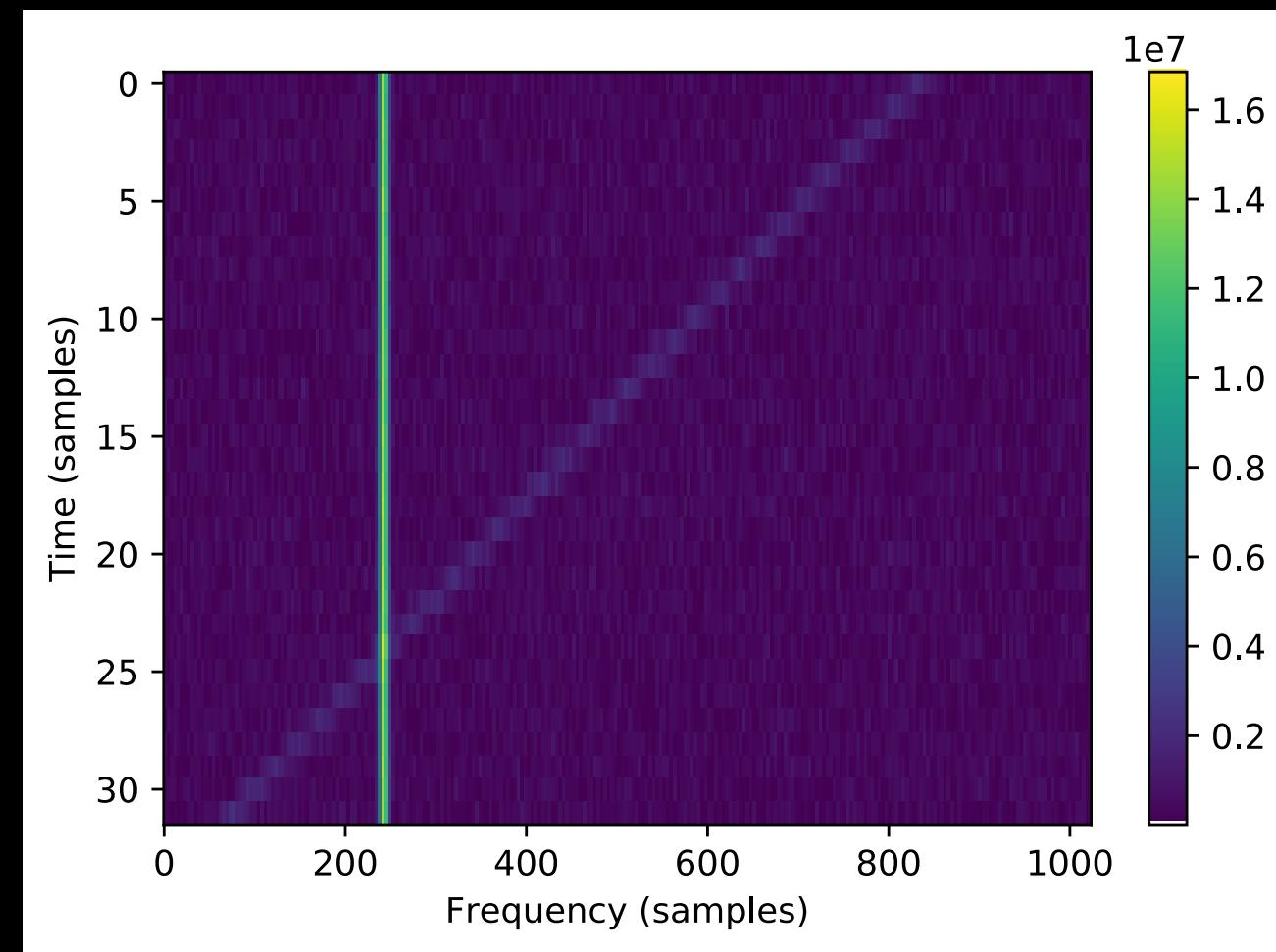


Small snippet of GBT data at C-band

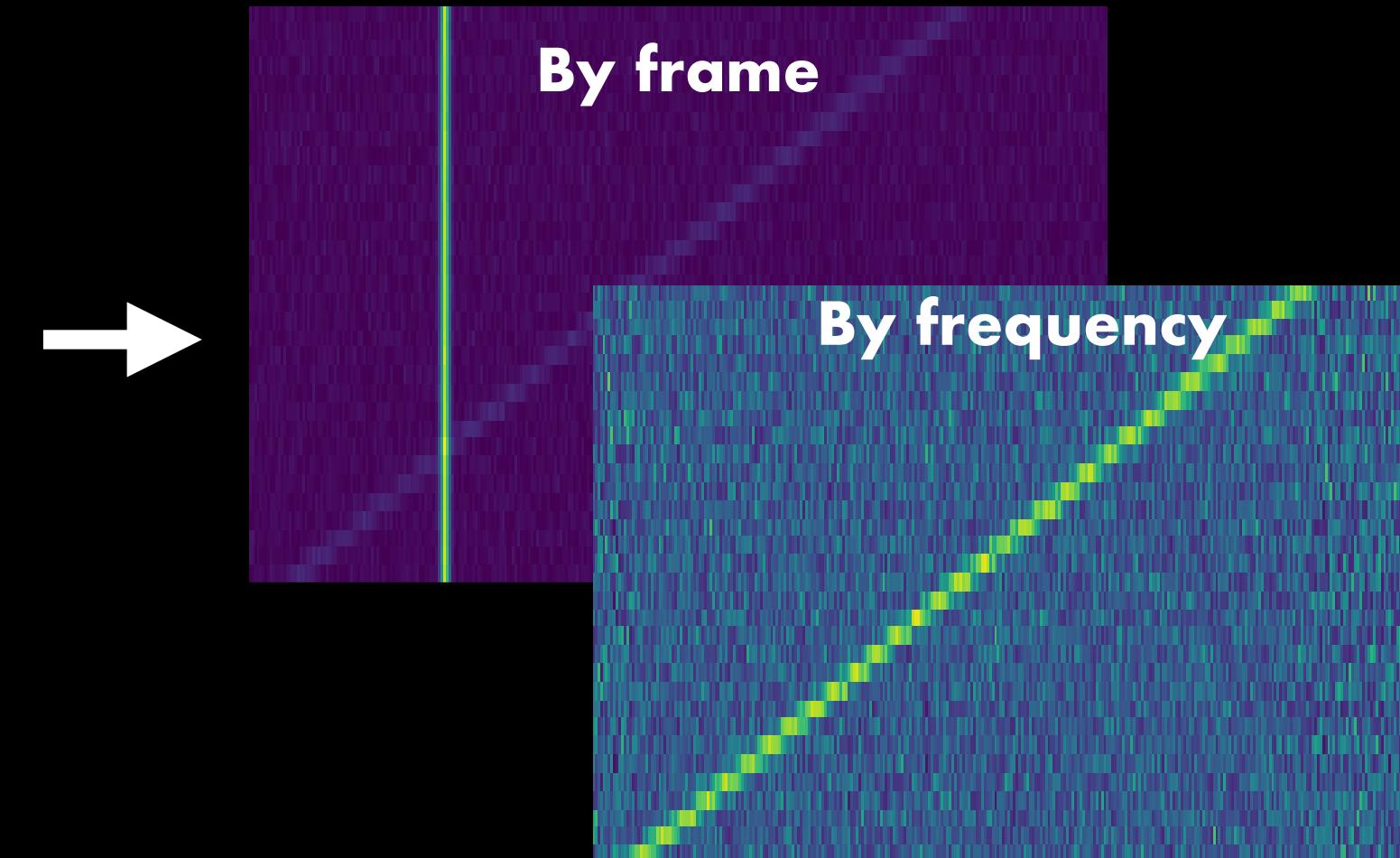
- **Masking?**



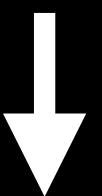
## Synthetic training data



## Normalization



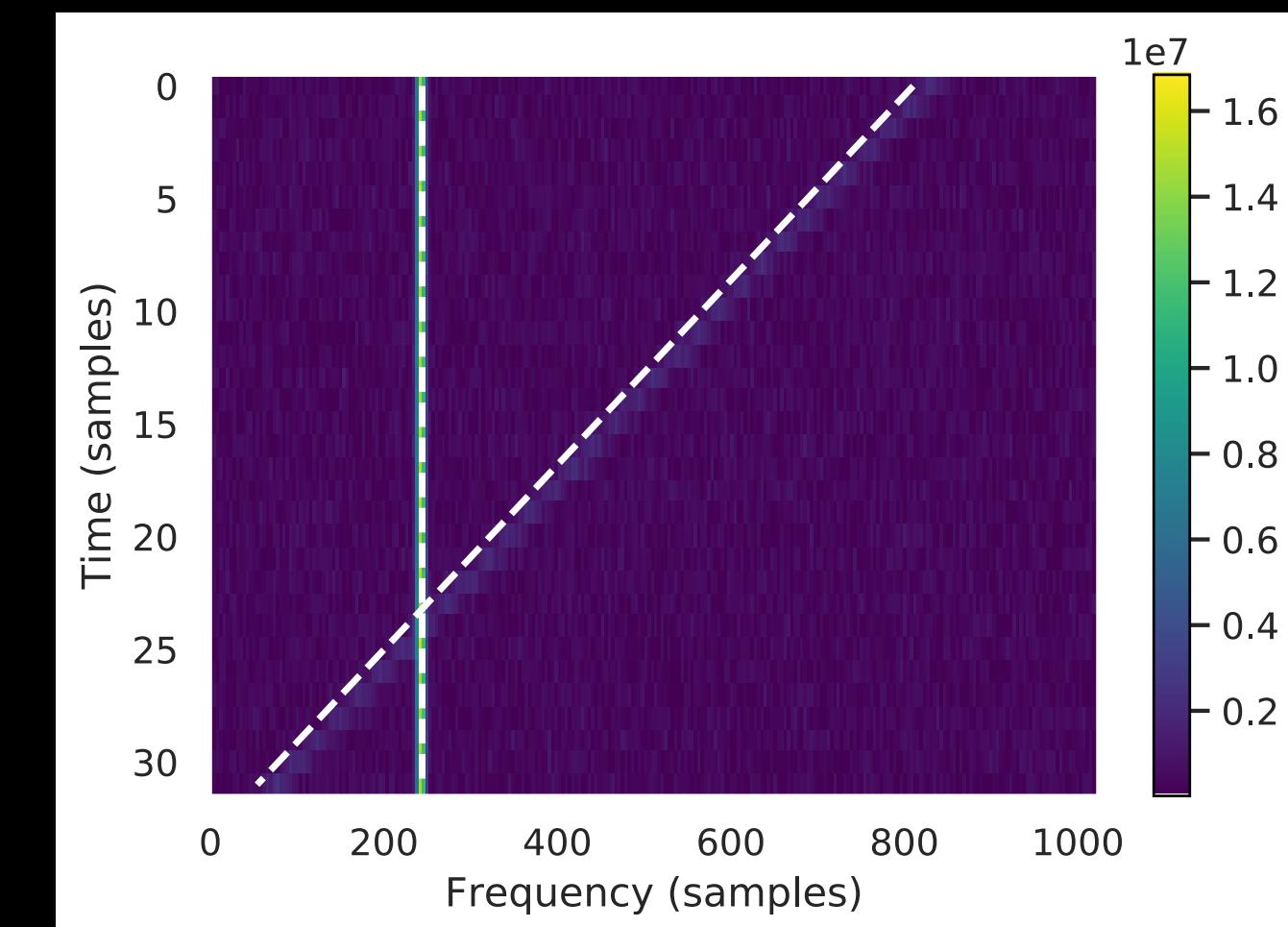
- 1 or 2 signals per input



**Neural Network**

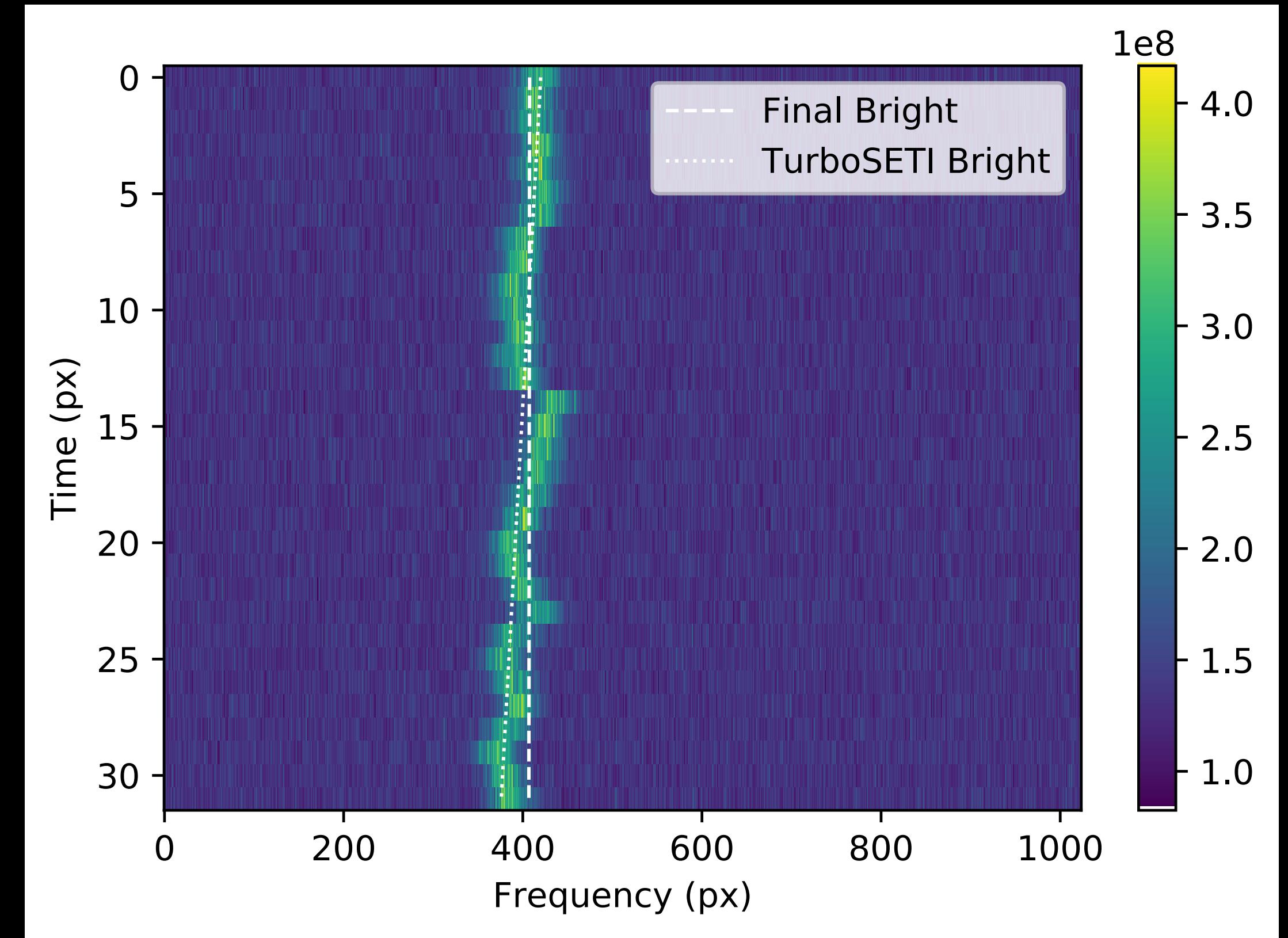


## Predicted locations



# Takeaways

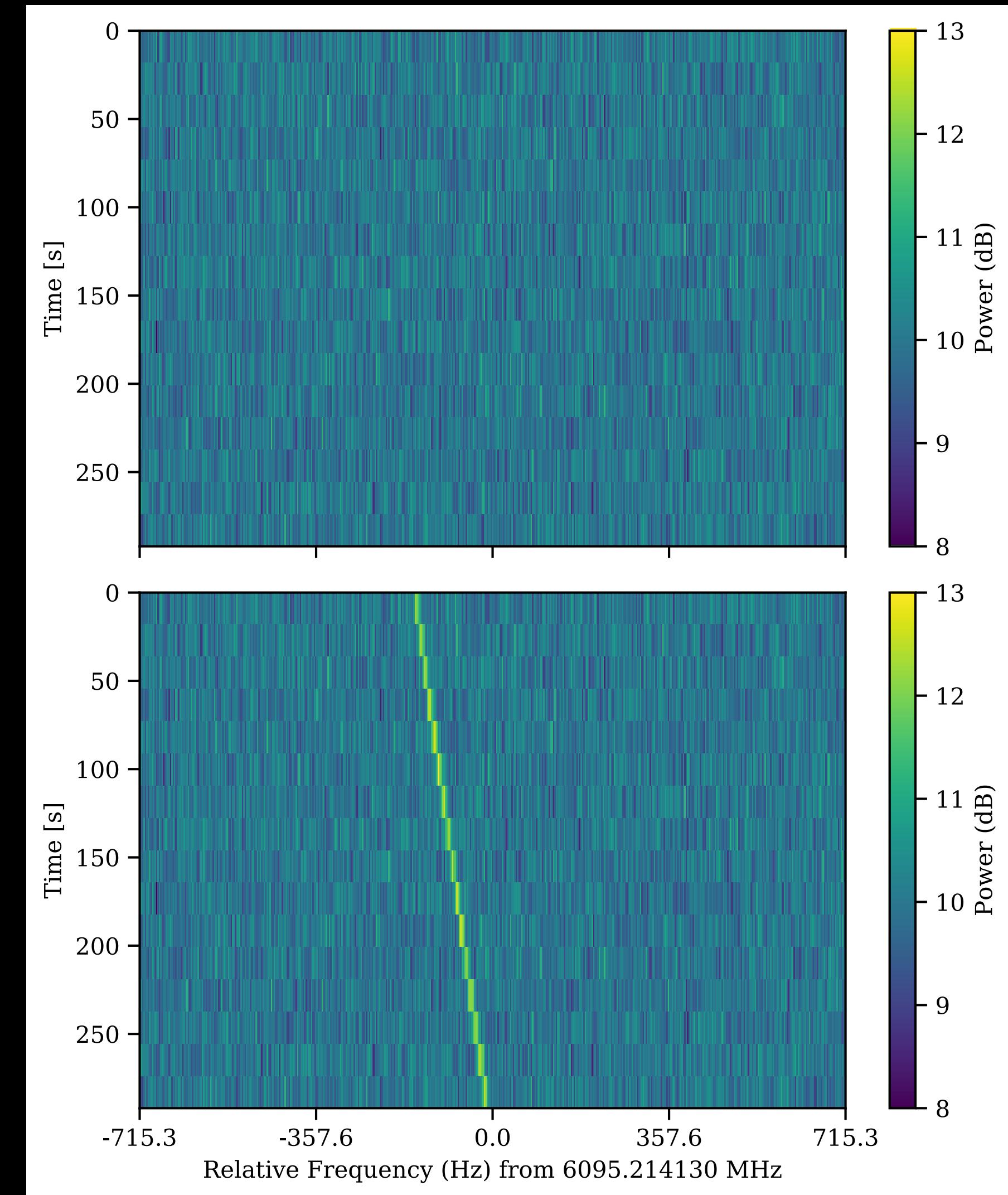
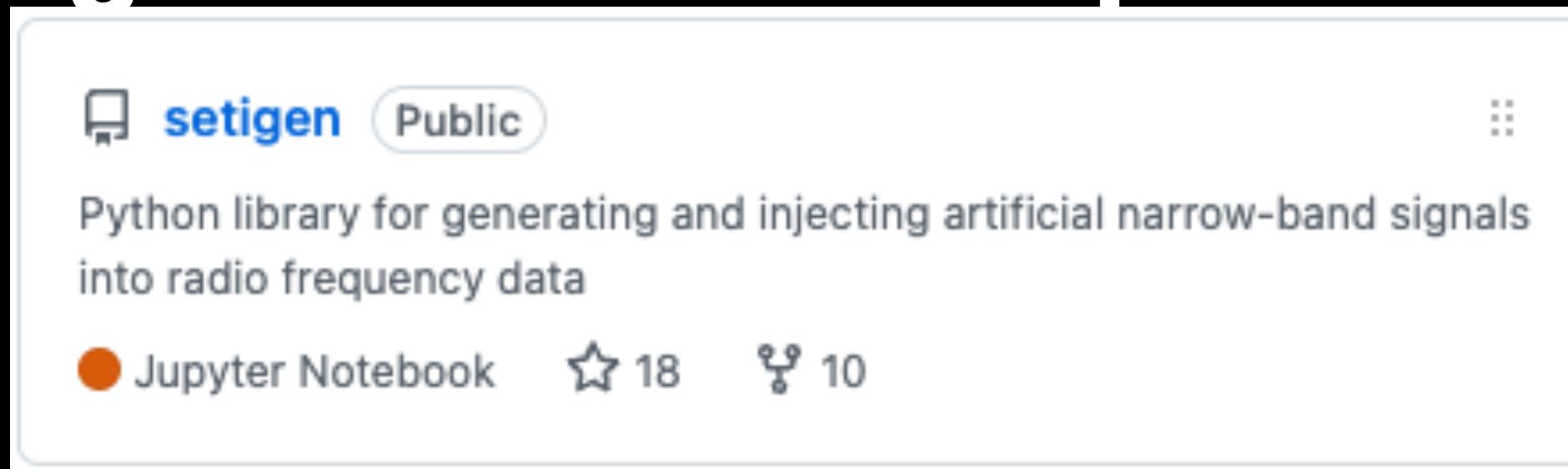
- Less accurate than deDoppler methods, but generally 20-40x faster
- Trained on ideal signals but still relatively robust
- For production use, would need to extend to variable number of signals



C-band RFI signal, with ML prediction dashed and TurboSETI localization dotted.

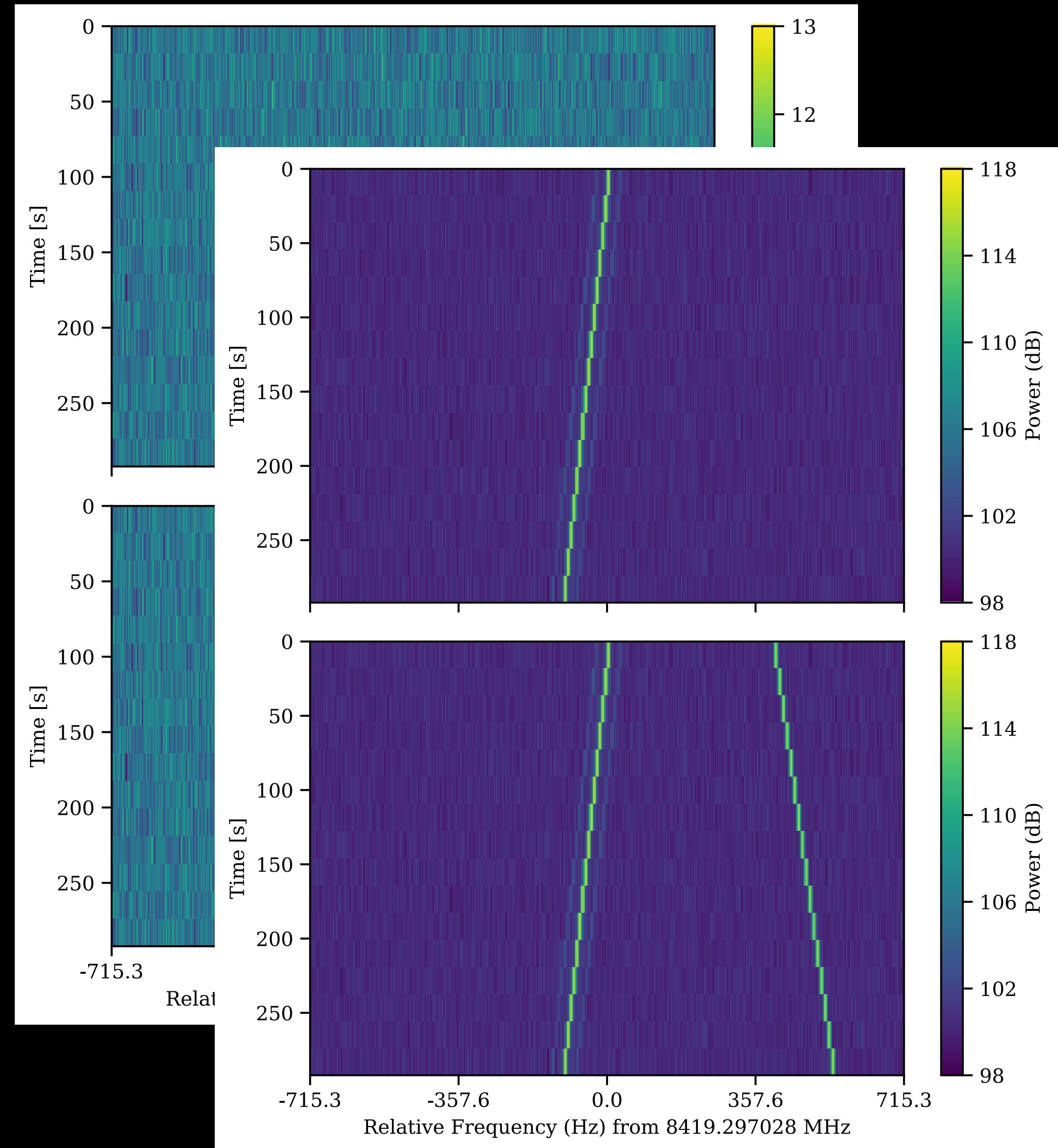
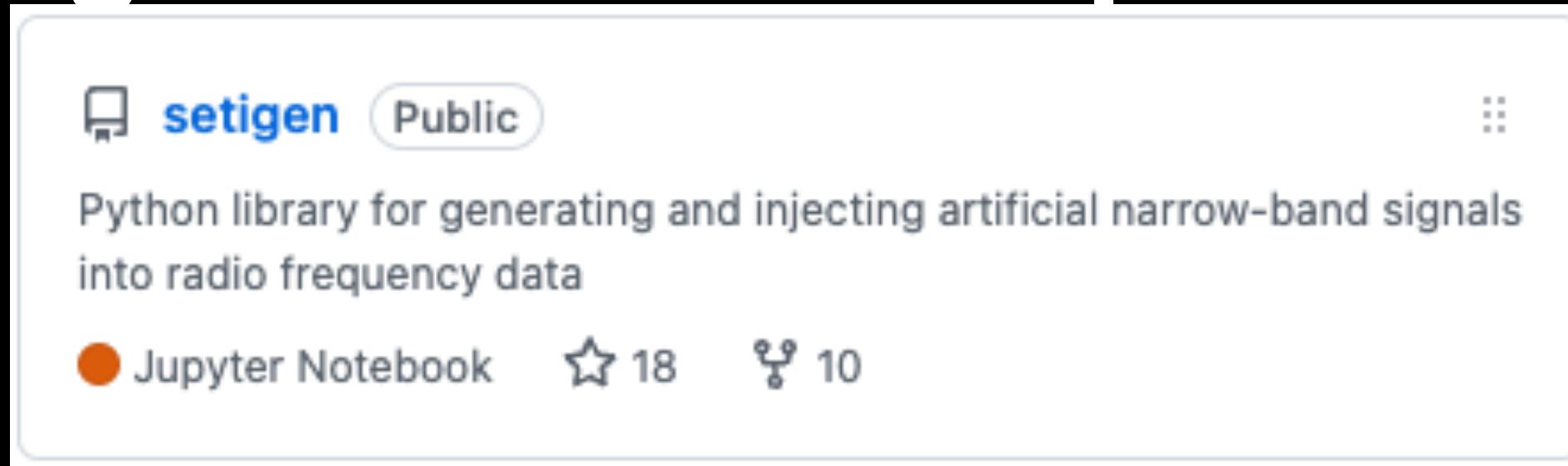
# Setigen

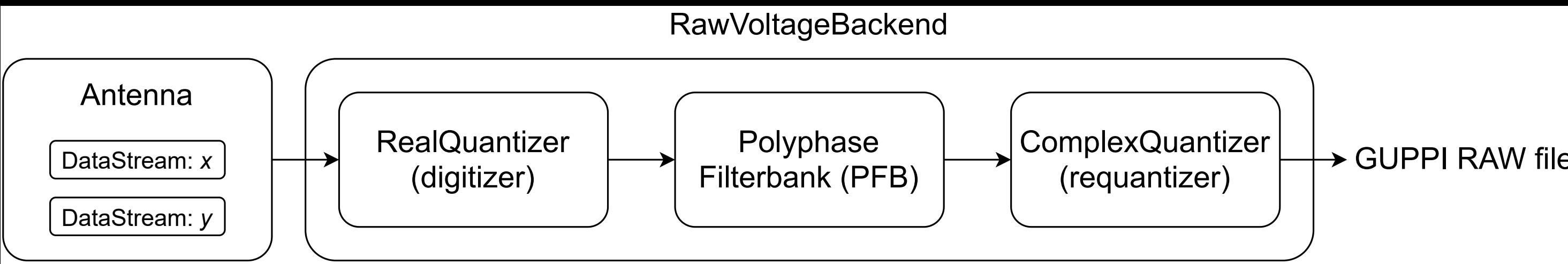
- **Python library for synthetic spectrogram and voltage data**
- **Specific focus on narrowband signal generation and injection**



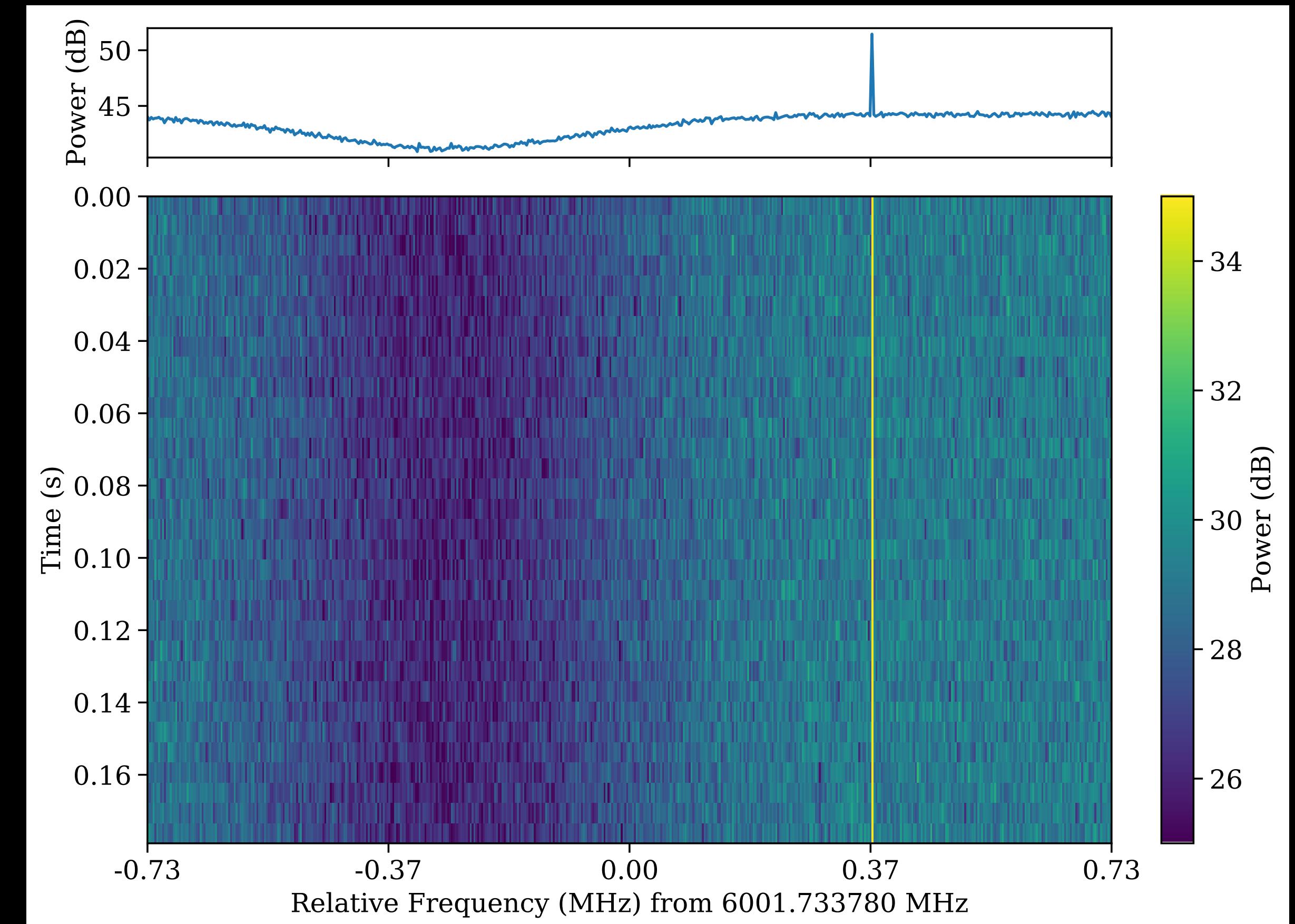
# Setigen

- **Python library for synthetic spectrogram and voltage data**
- **Specific focus on narrowband signal generation and injection**



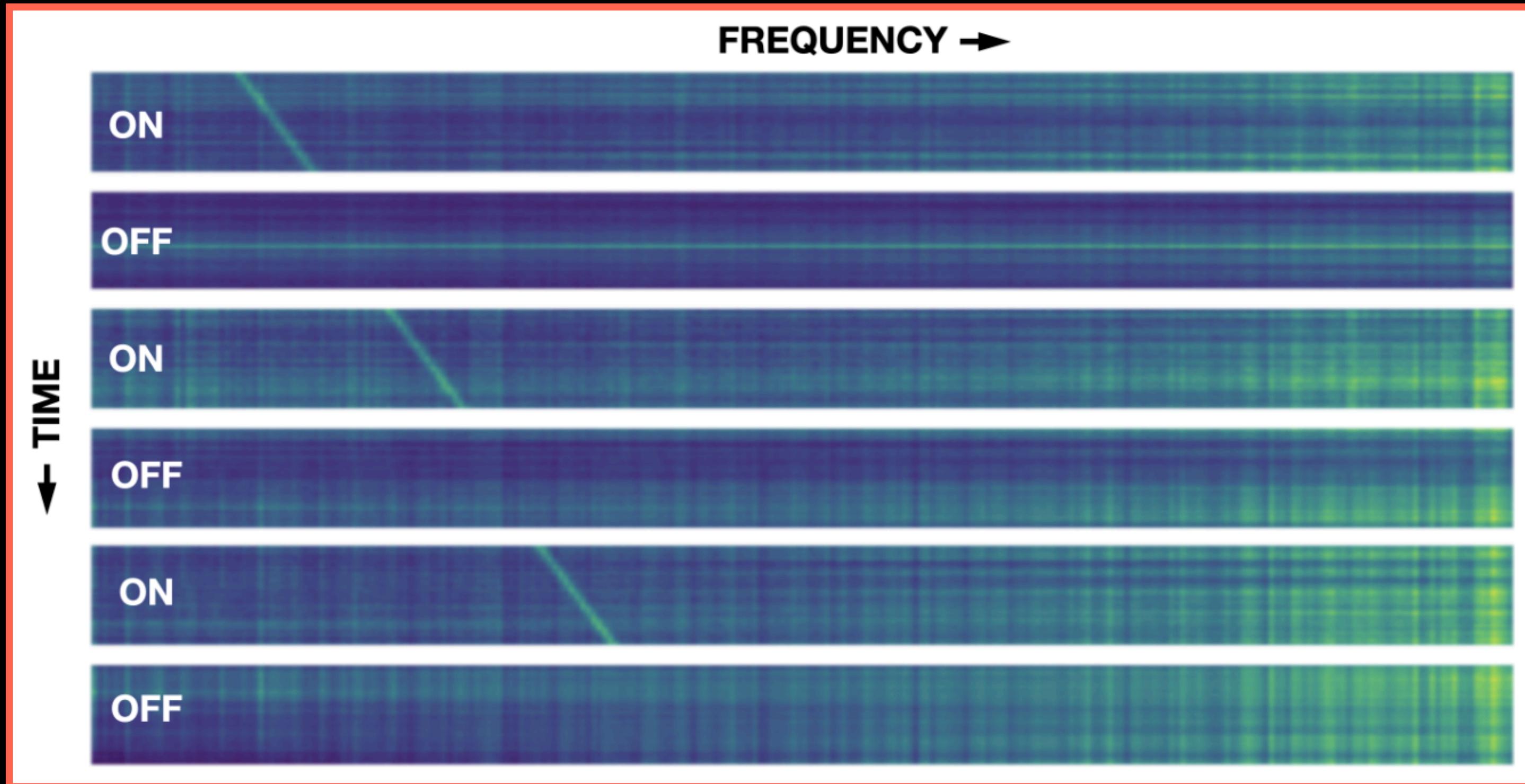


- **Synthetic complex voltage data**
- **Simple models of backend components, such as a polyphase filterbank**



# Applications of Setigen beyond my research

- **Injection — recovery testing**
- **ML dataset production (e.g. Kaggle)**
- **Multibeam search surveys**
- **Development of software for the Allen Telescope Array**



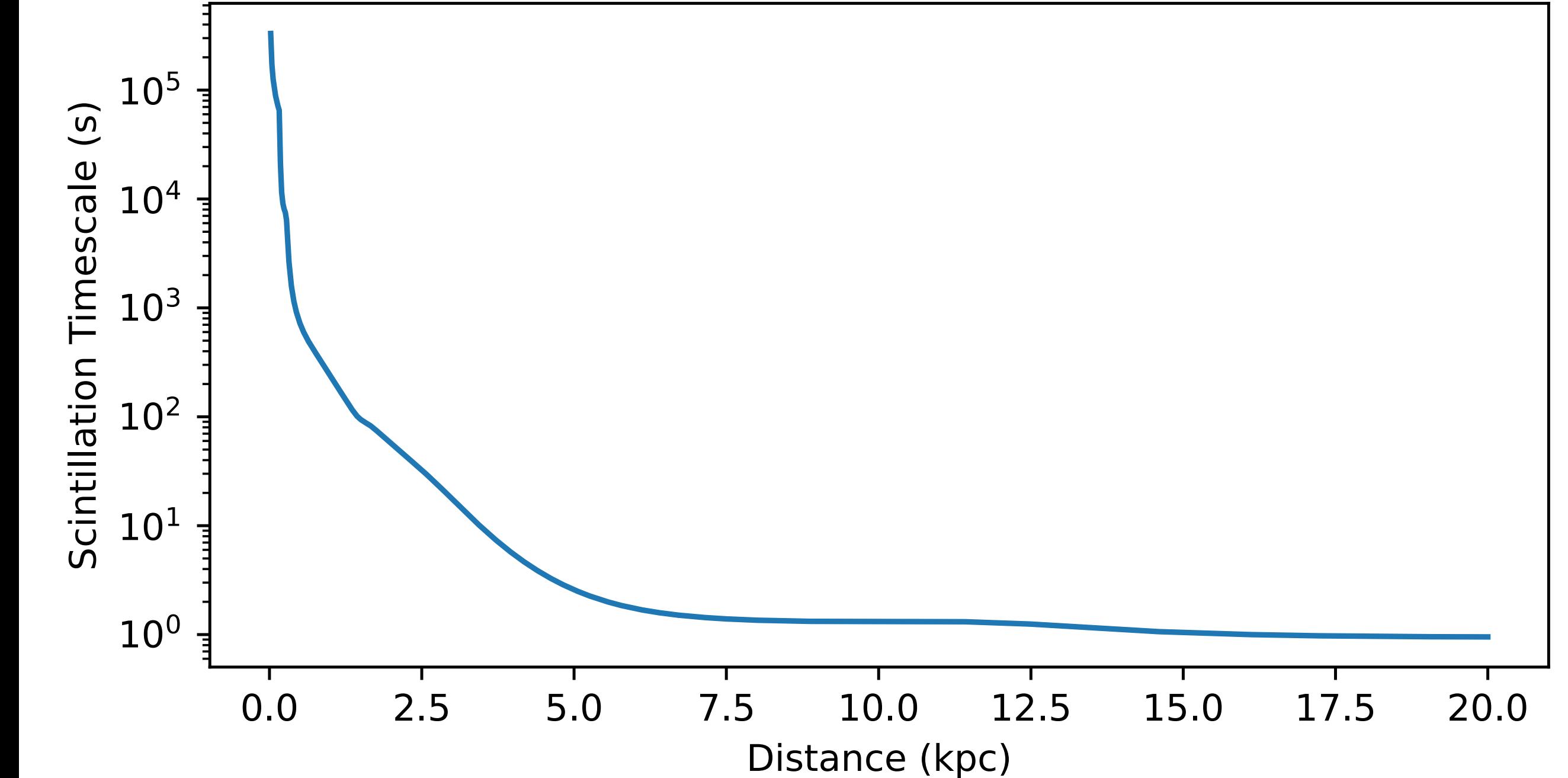
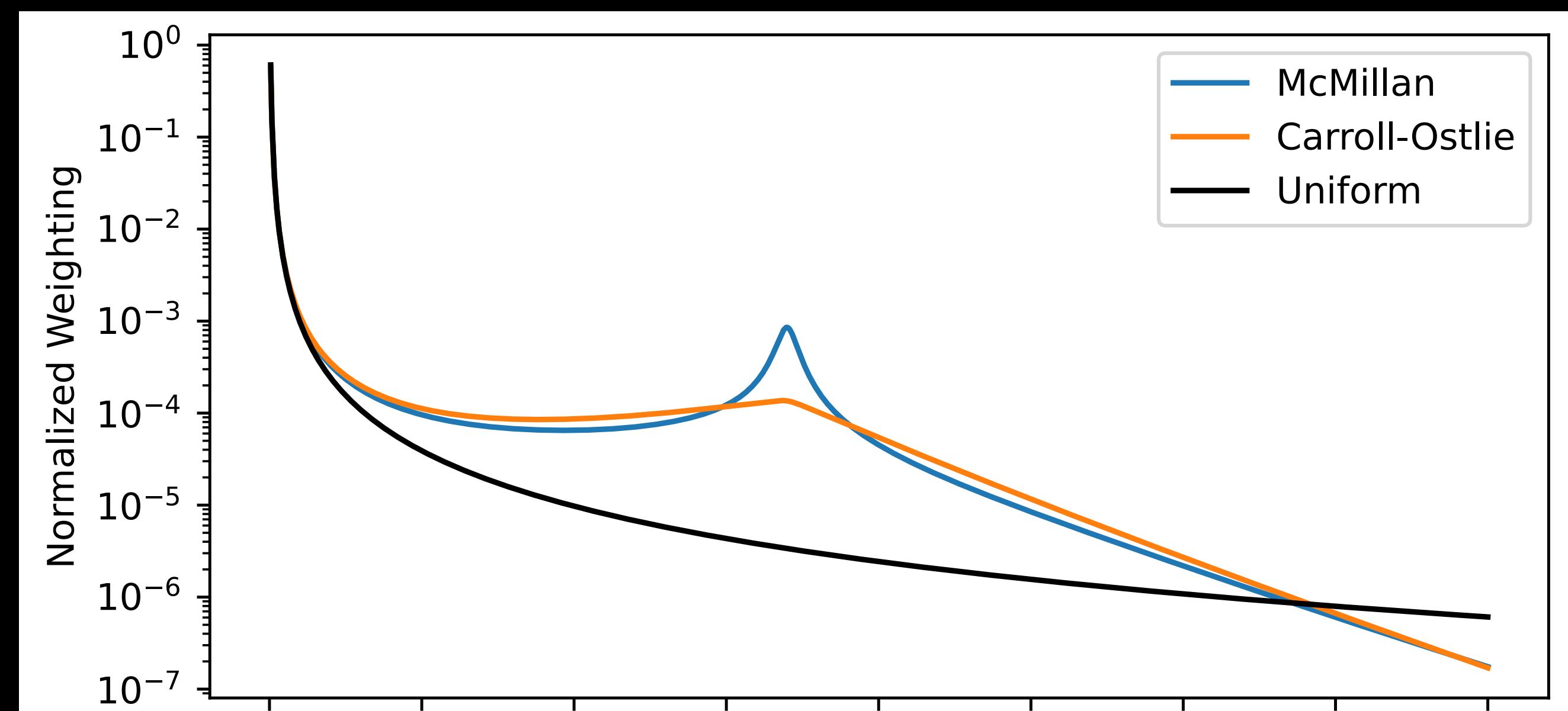
Breakthrough Listen x Kaggle

$$(l, b) = (1, 0)$$

# Density-based sampling

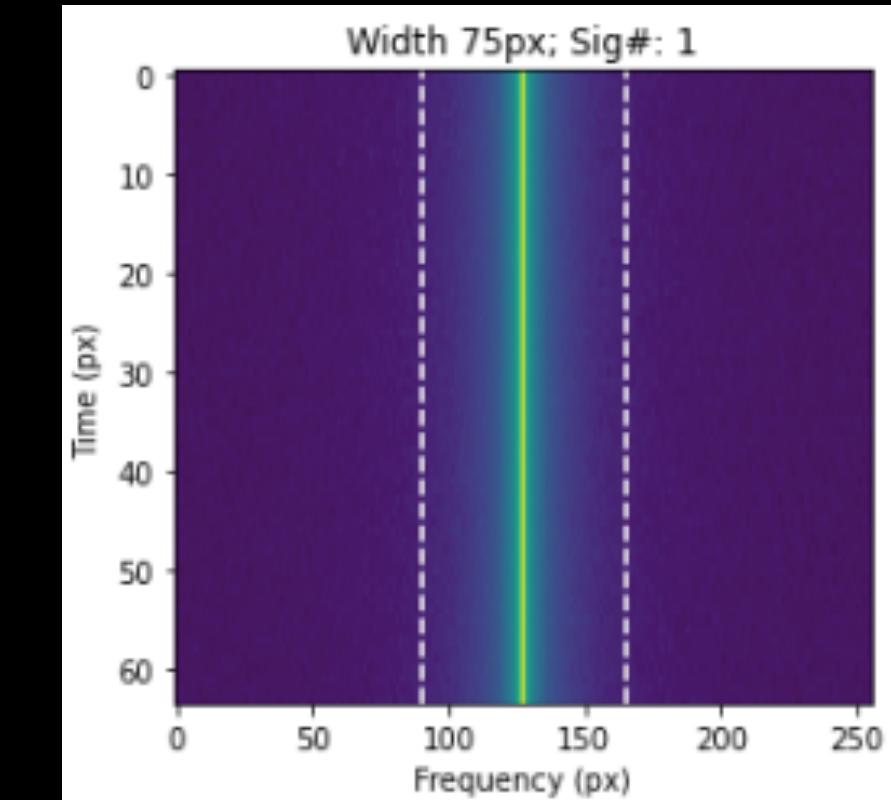
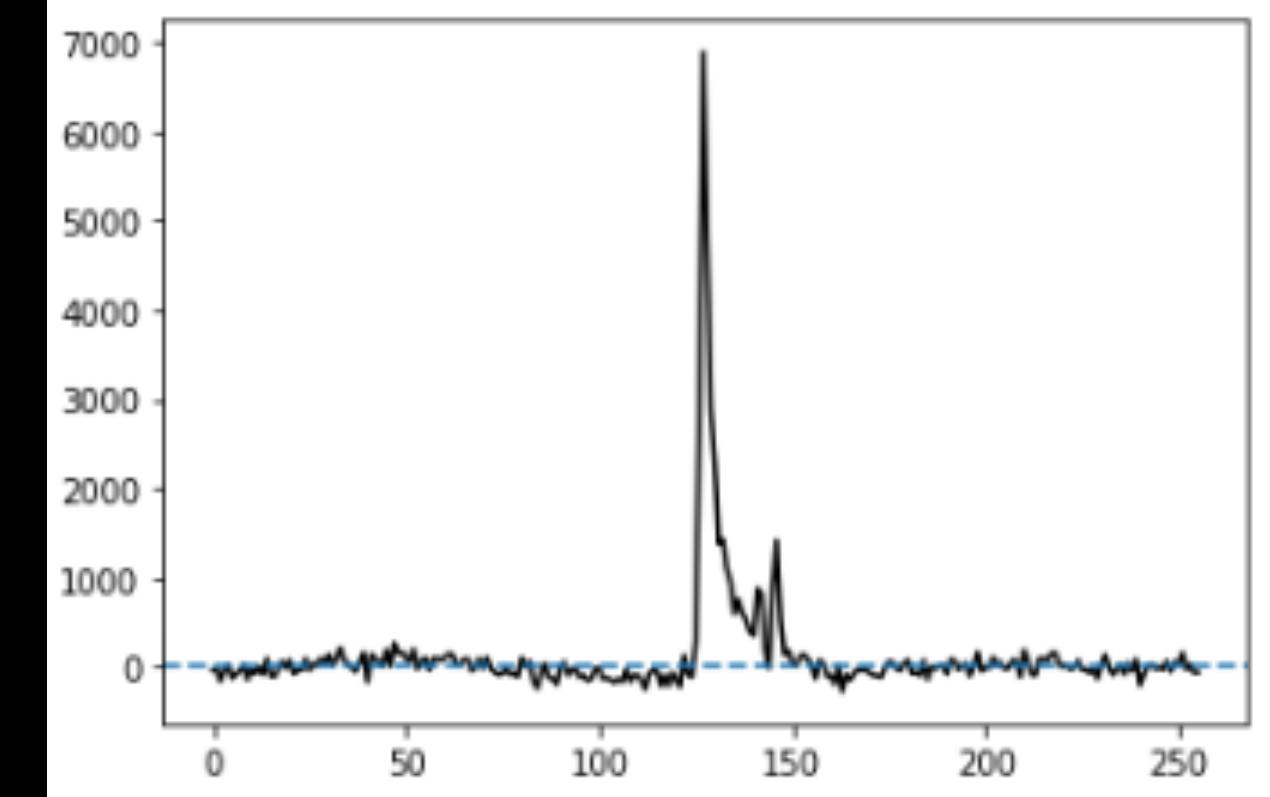
**Modulating by the inverse square-law for detectability:**

**Depends on the assumptions made about transmission power and resources.**

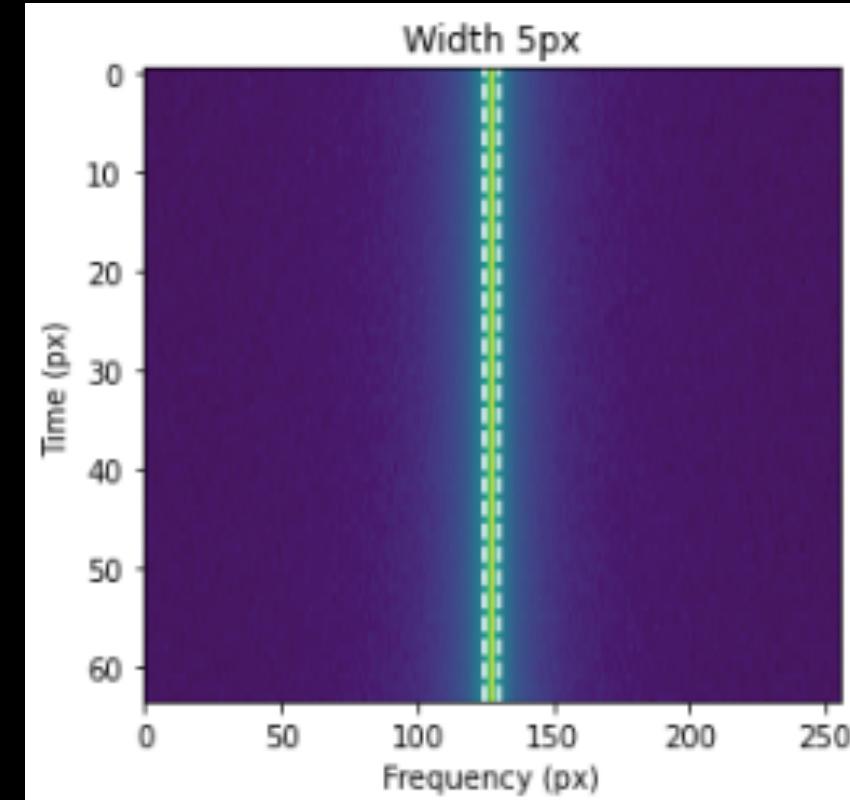


# Selecting bounding boxes

- After experimentation with various methods, the final pipeline uses a combination of baseline fitting and peak detection to calculate the right size of frame to use
- The final bounds are created using a thresholding method, similar to PSRCHIVE
- Take the final bounded signal and integrate in the frequency direction to derive our raw time series — then we normalize to mean of 1 before calculating our scattering statistics



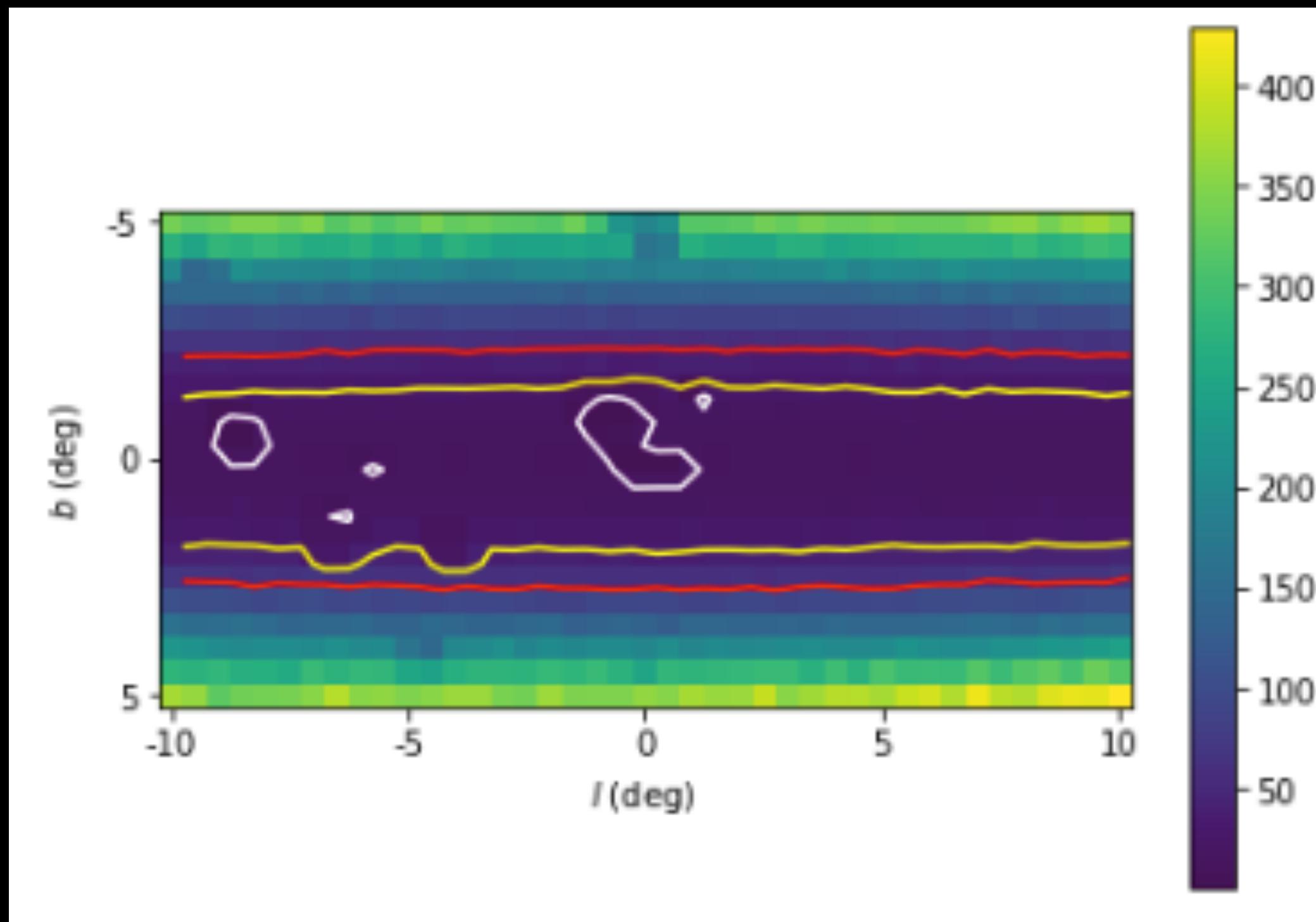
Polynomial fit



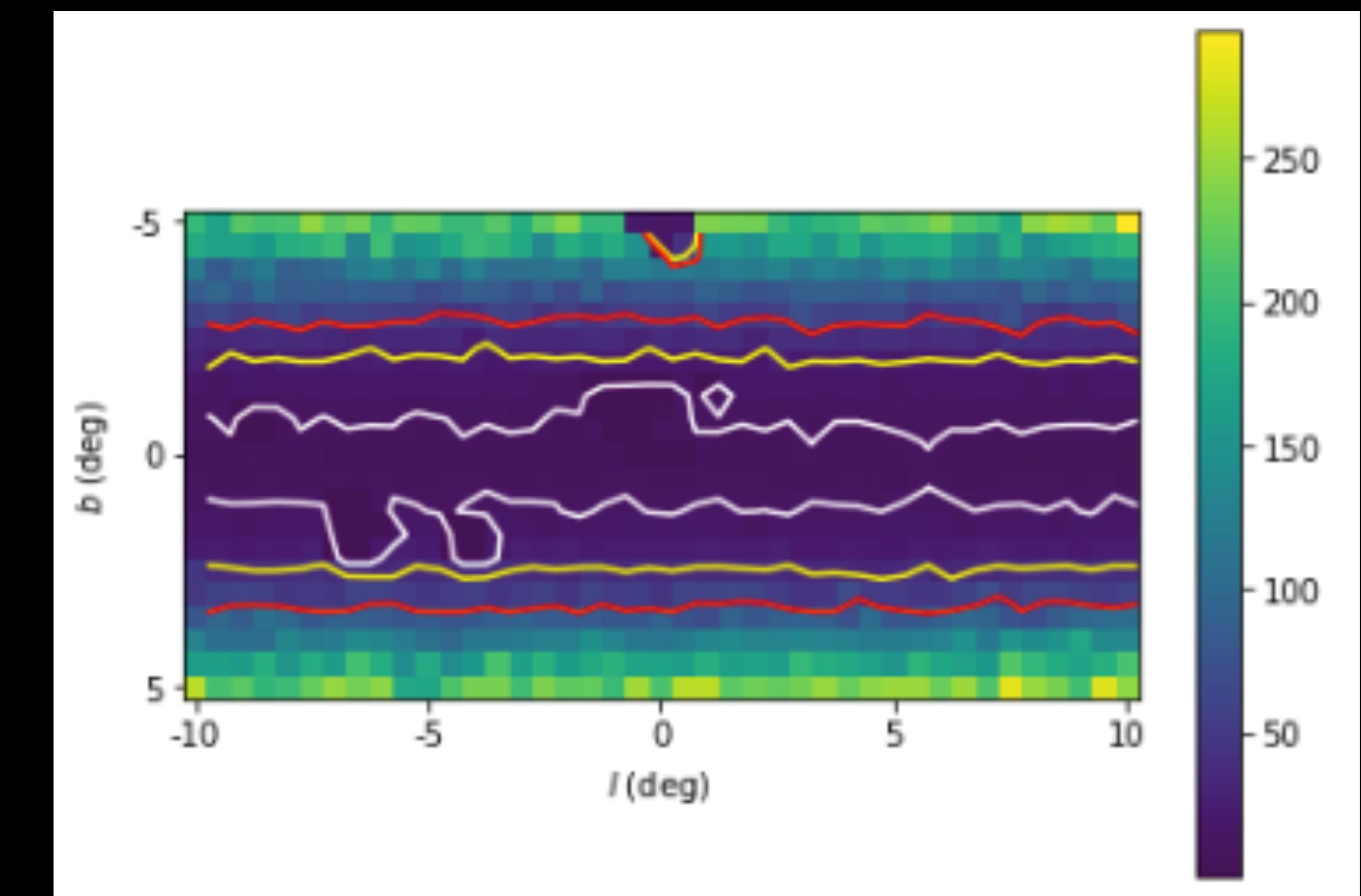
Threshold fit

# Scintillation maps around the GC at C-band

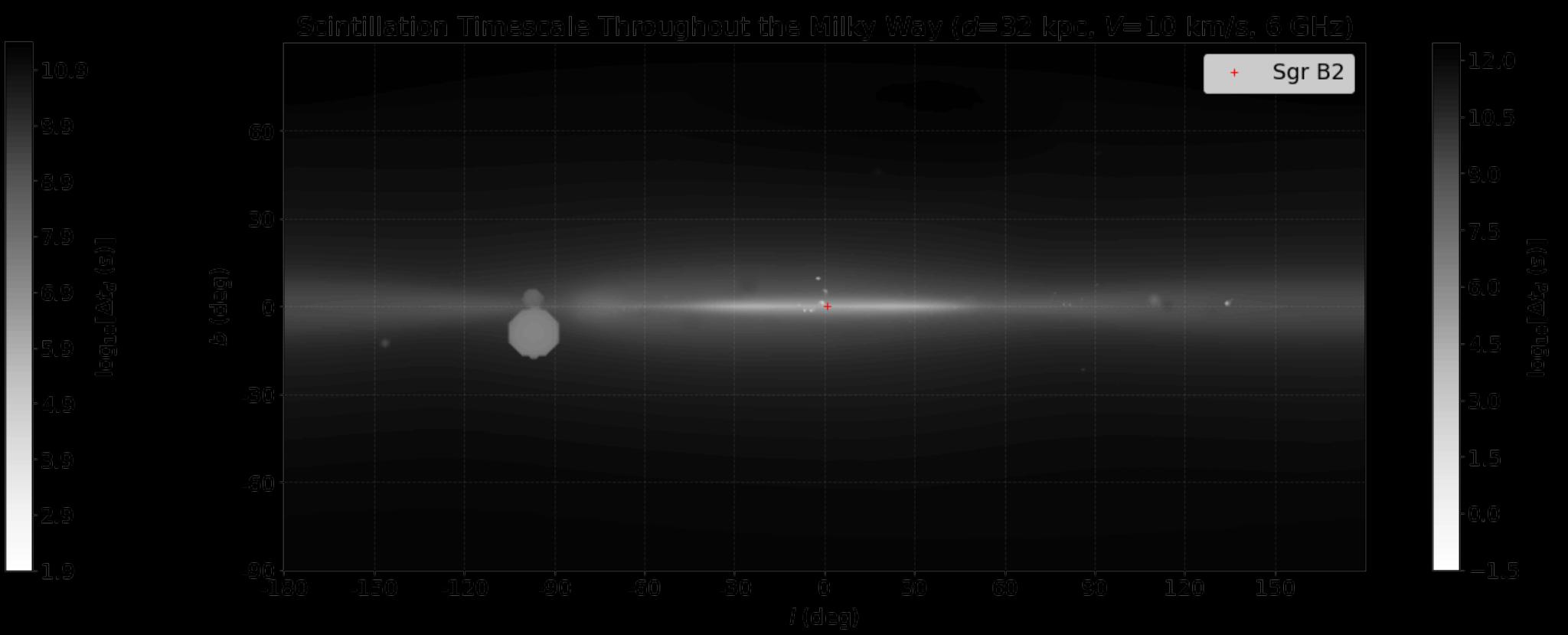
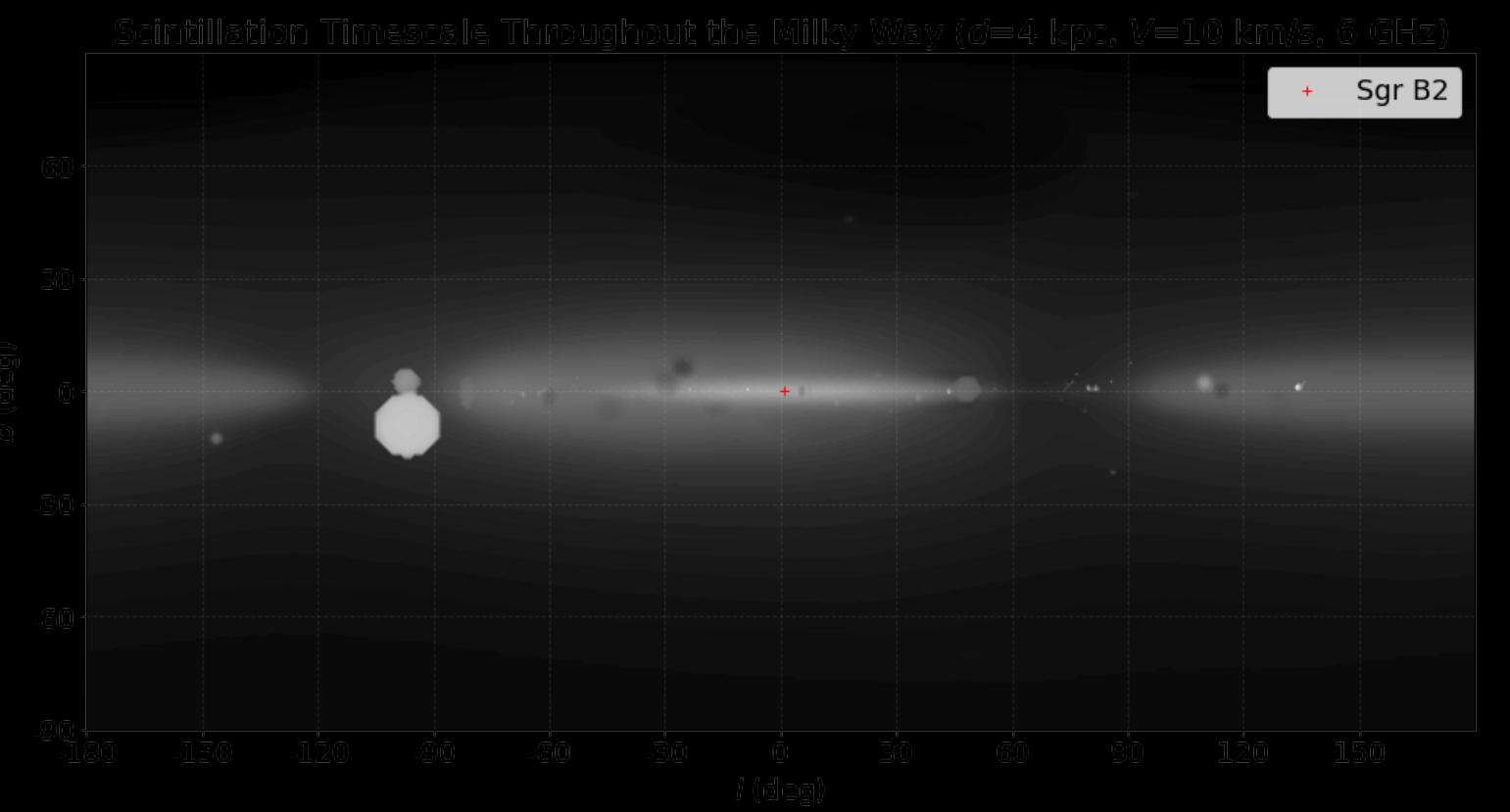
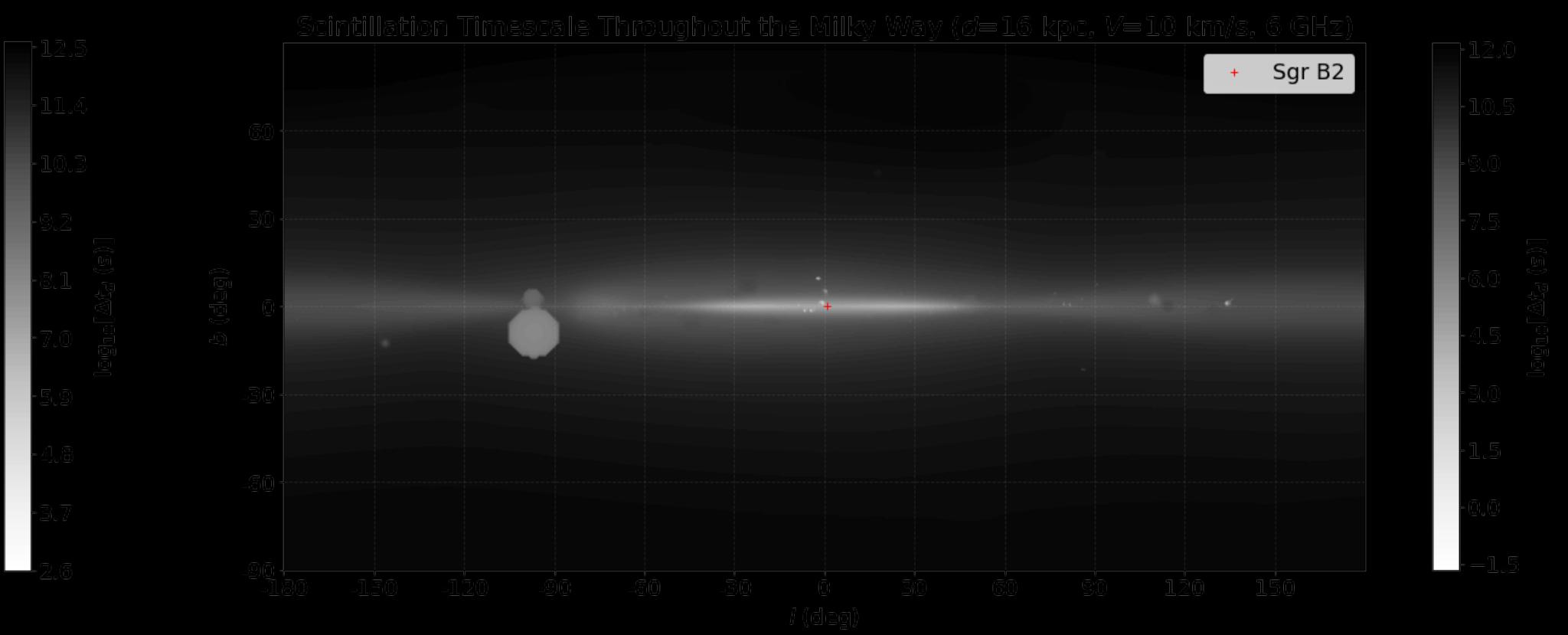
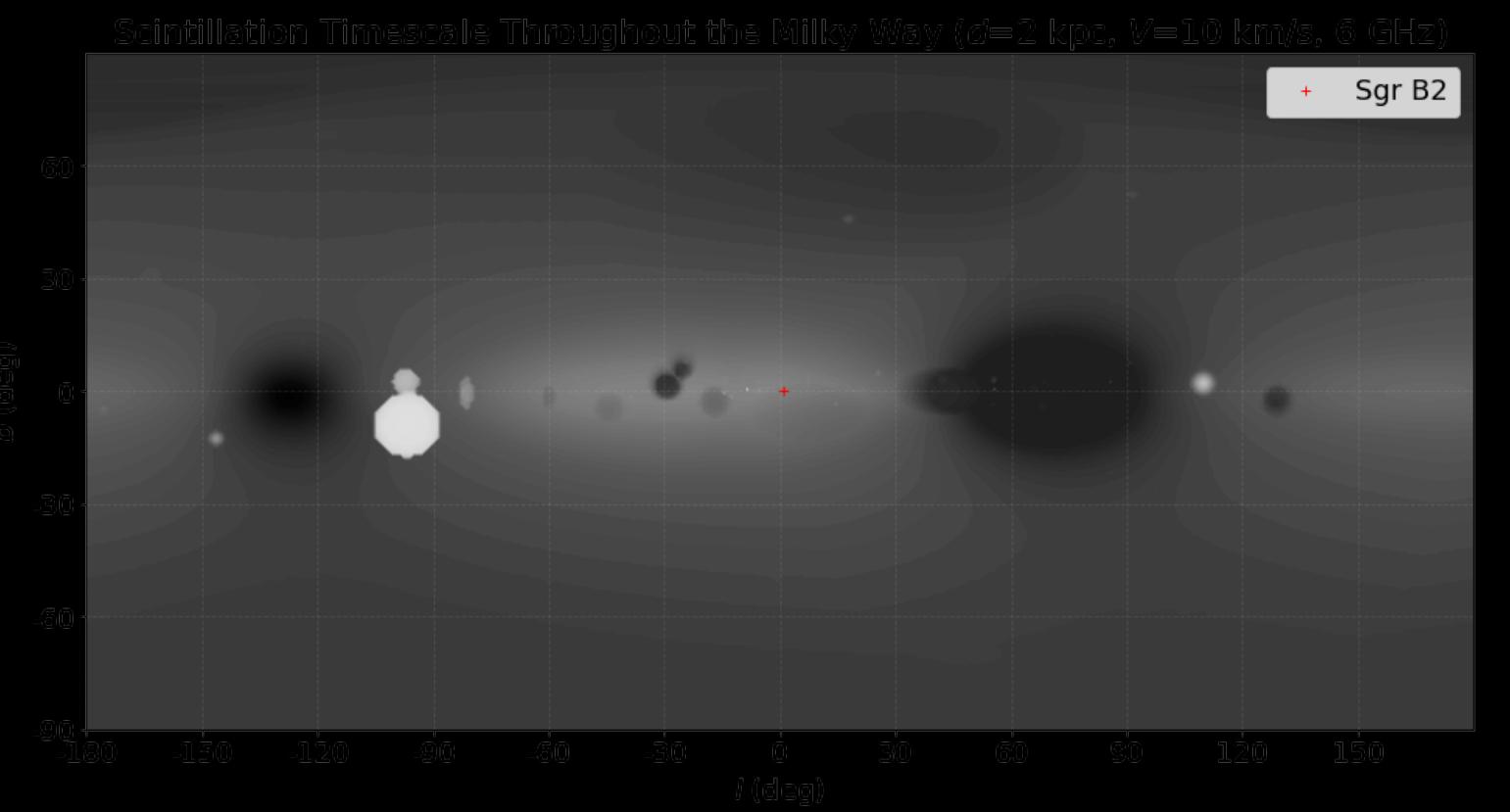
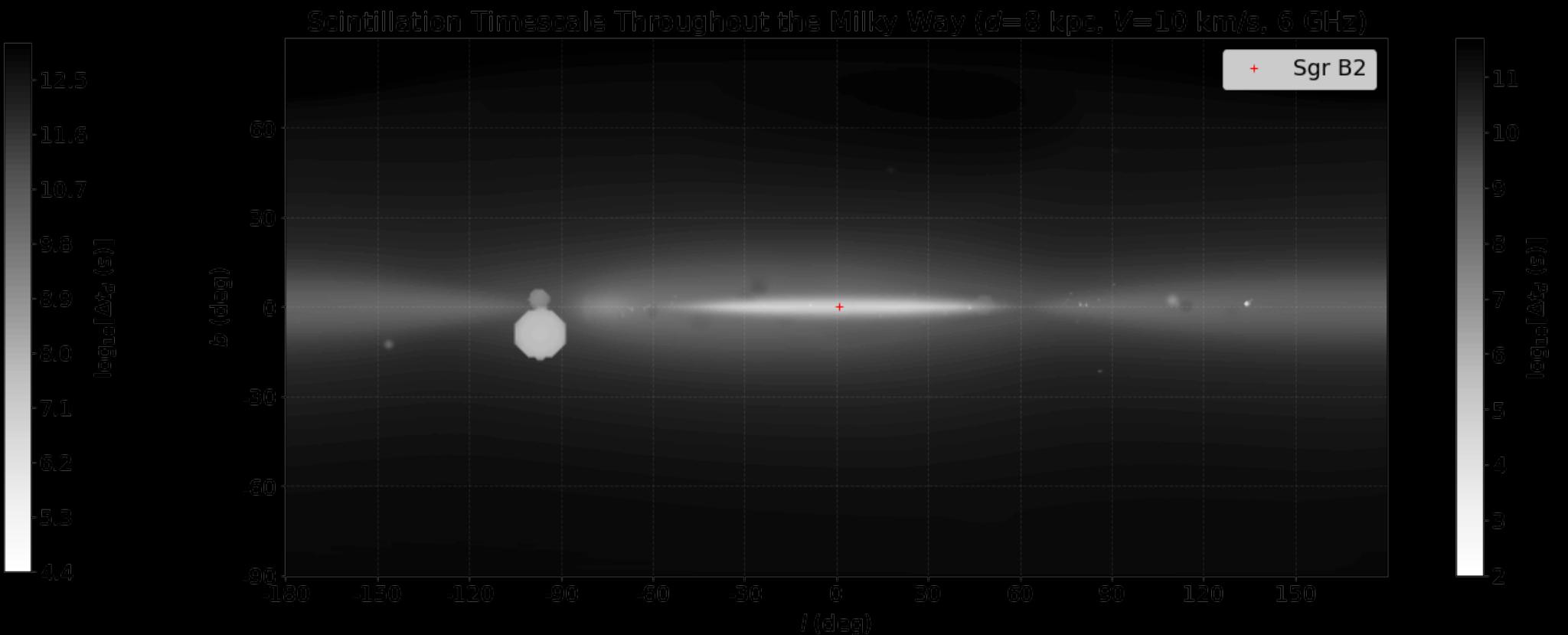
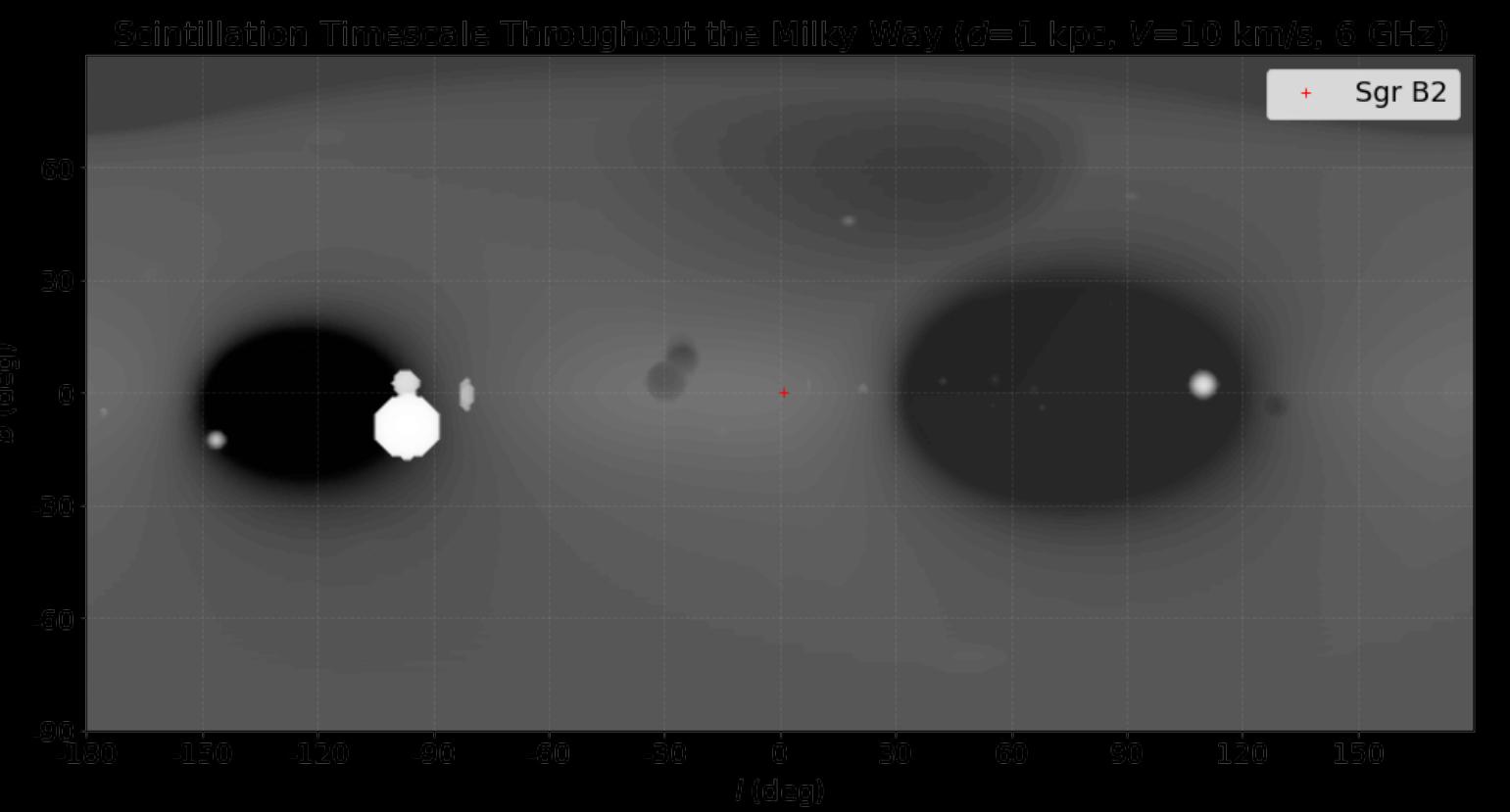
Median



Mode

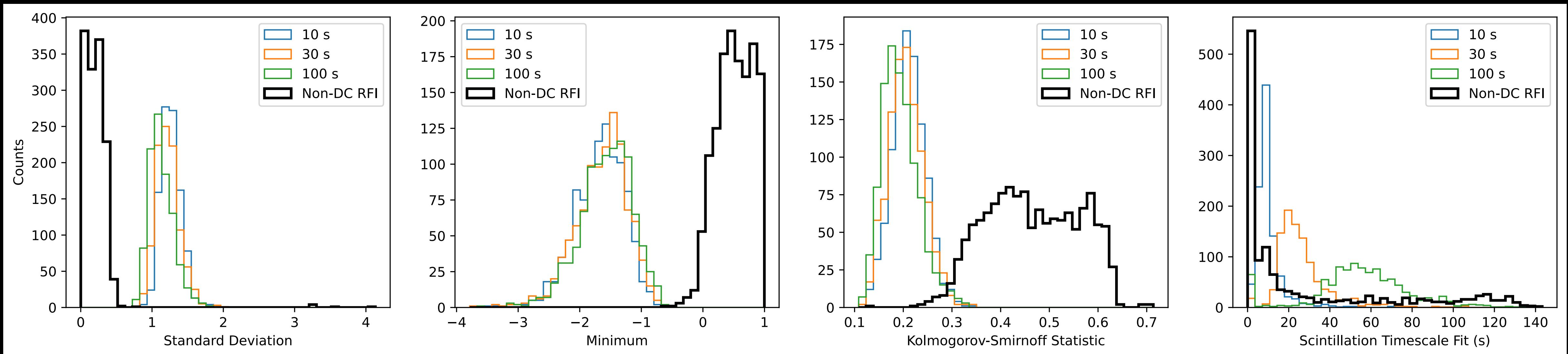


10 s, 30 s, 60 s



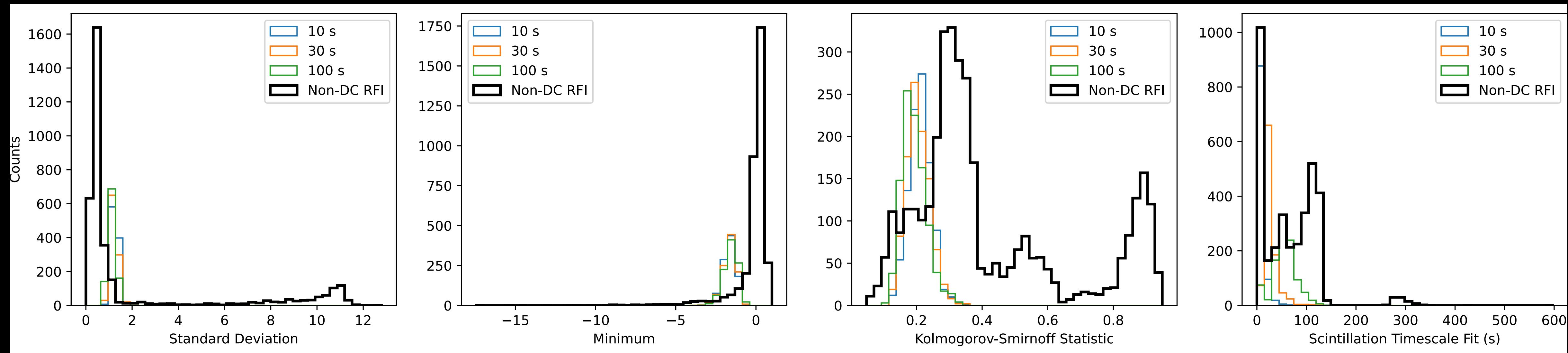
**C band**

**S/N = 10**



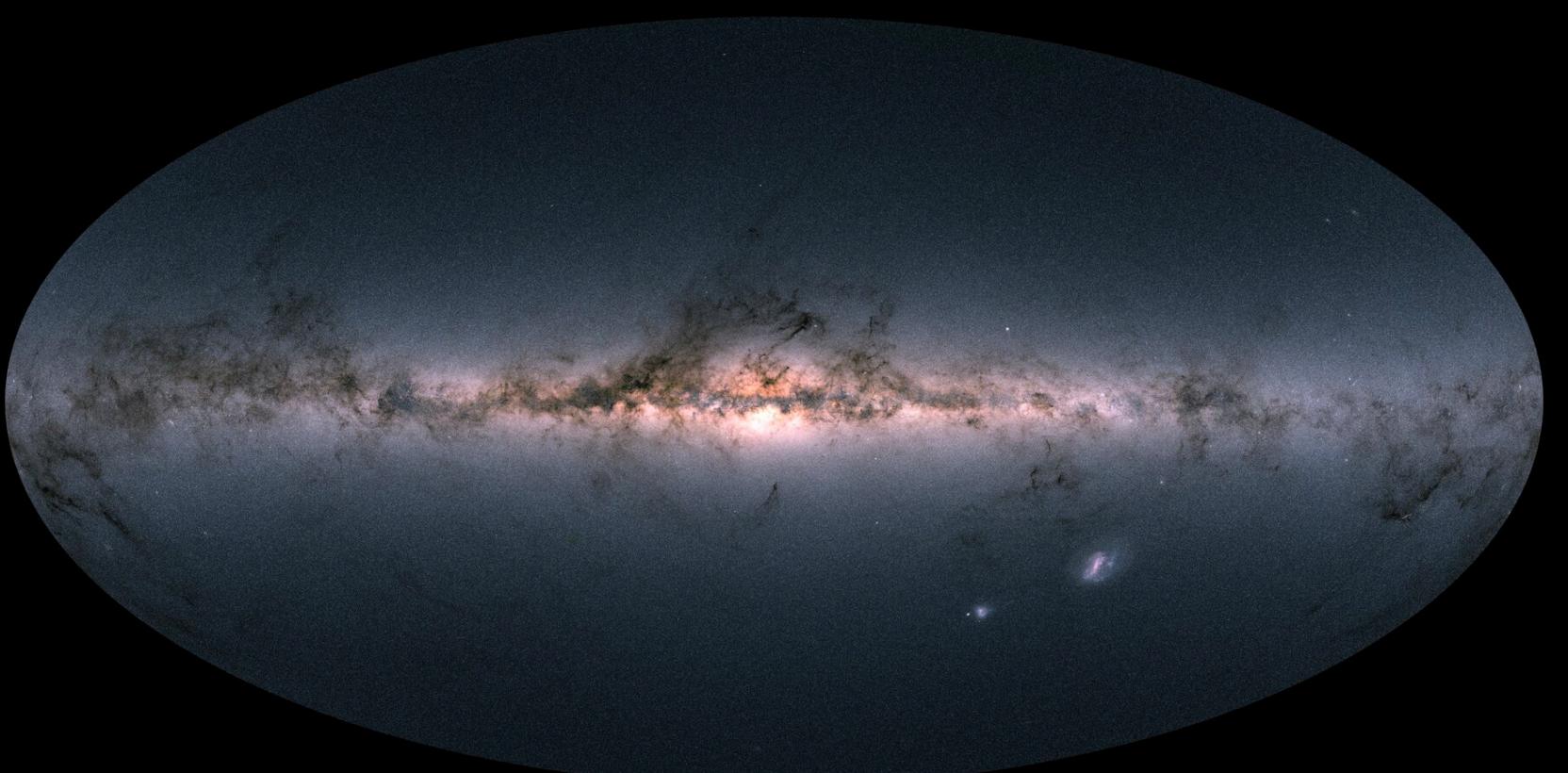
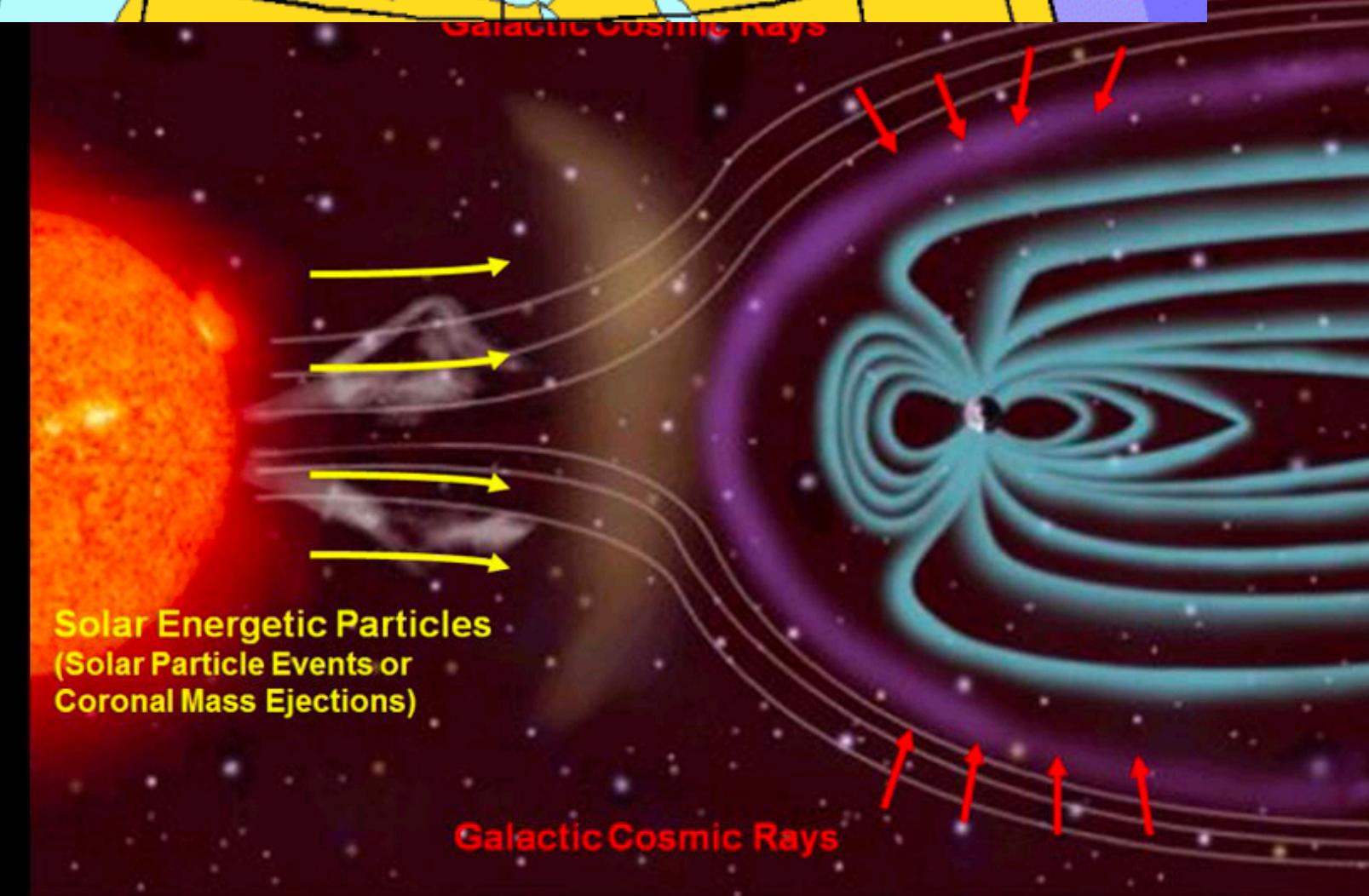
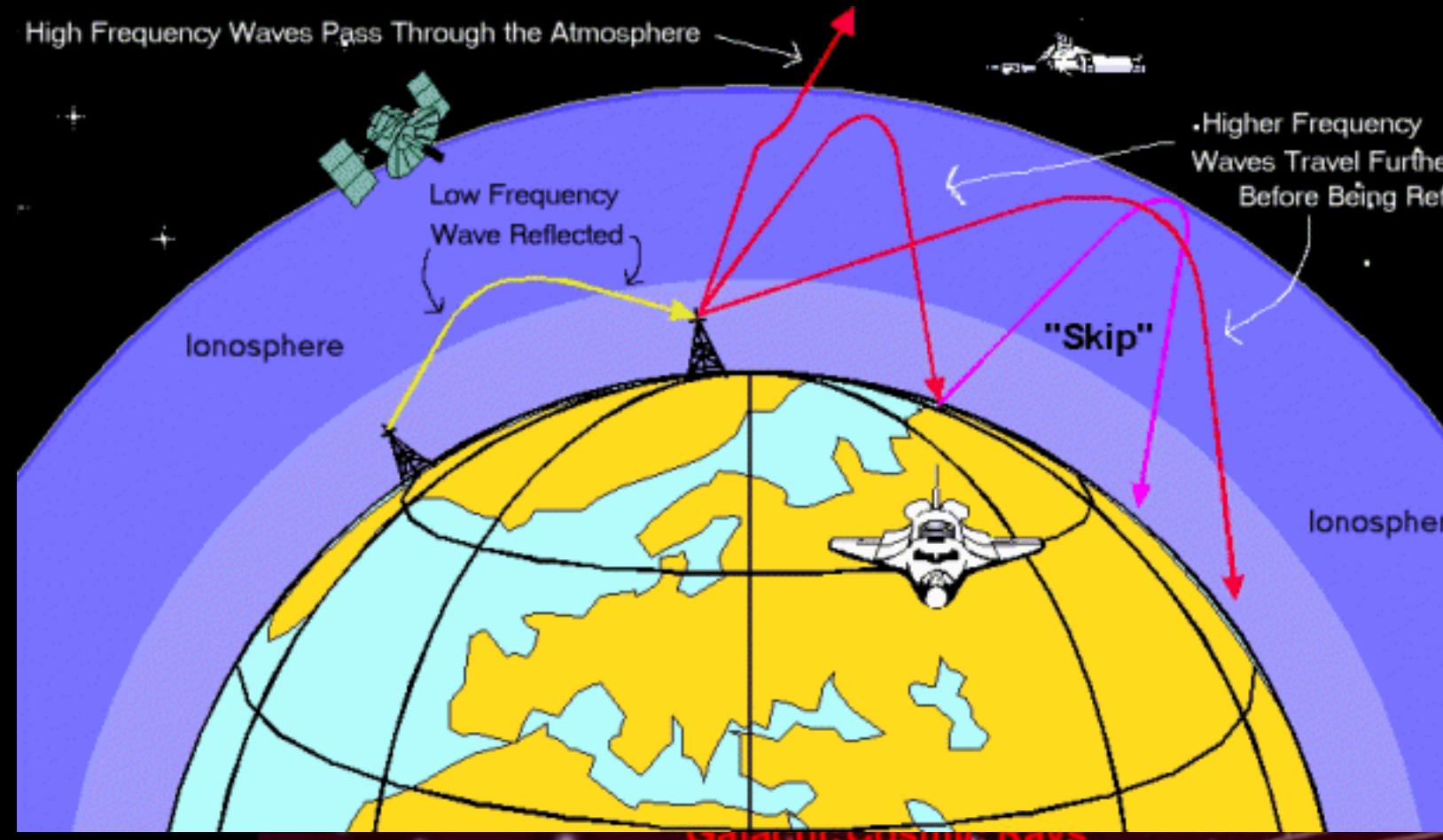
**L band**

**S/N = 10**



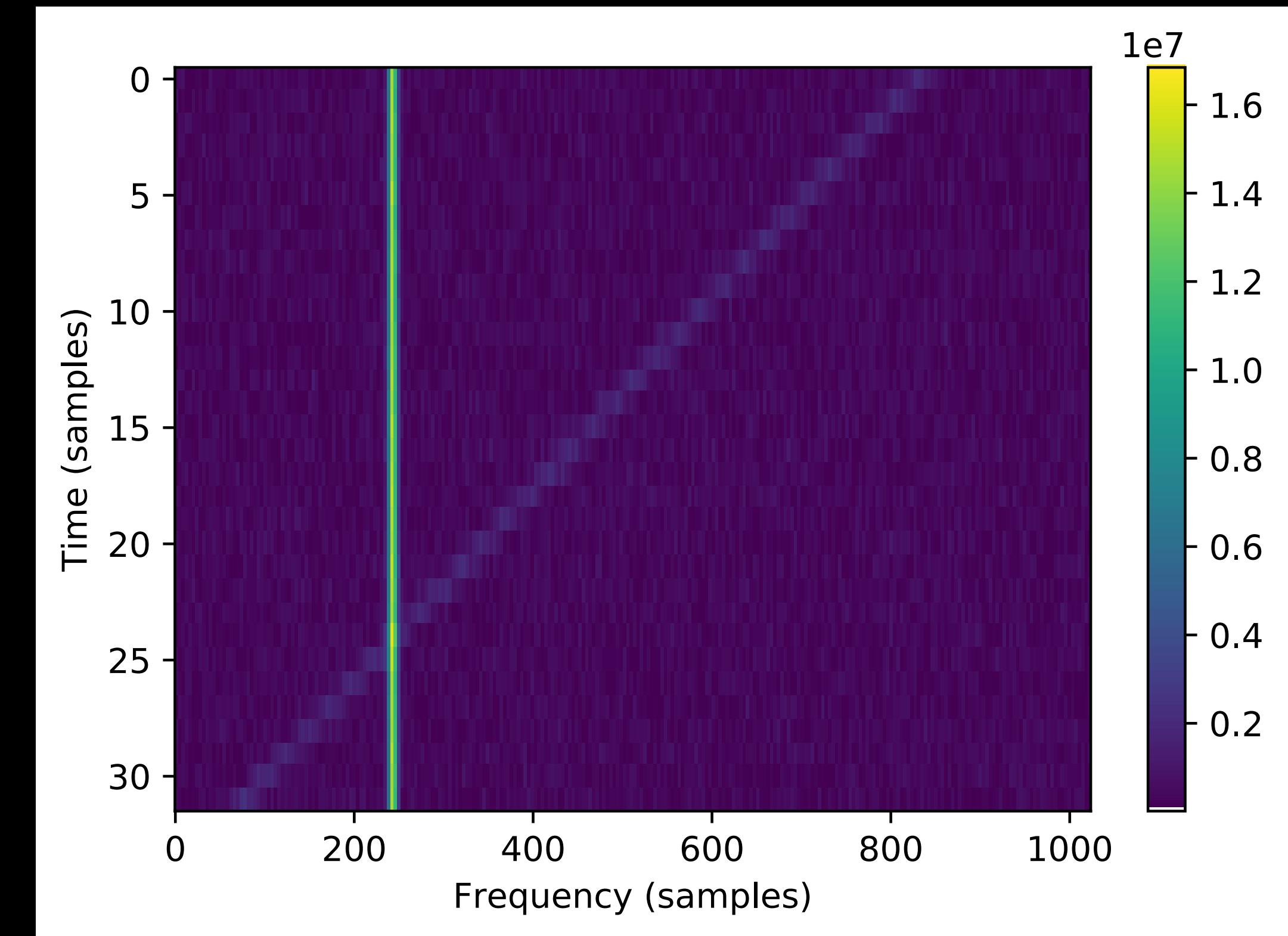
# Scattering intensity

- **Ionosphere — weak**     $m_d \ll 1$
- **IPM — mostly weak**
- **ISM — can be strong!**     $m_d \approx 1$



# NARROW-BAND SIGNAL LOCALIZATION (BRZYCKI ET AL. 2020)

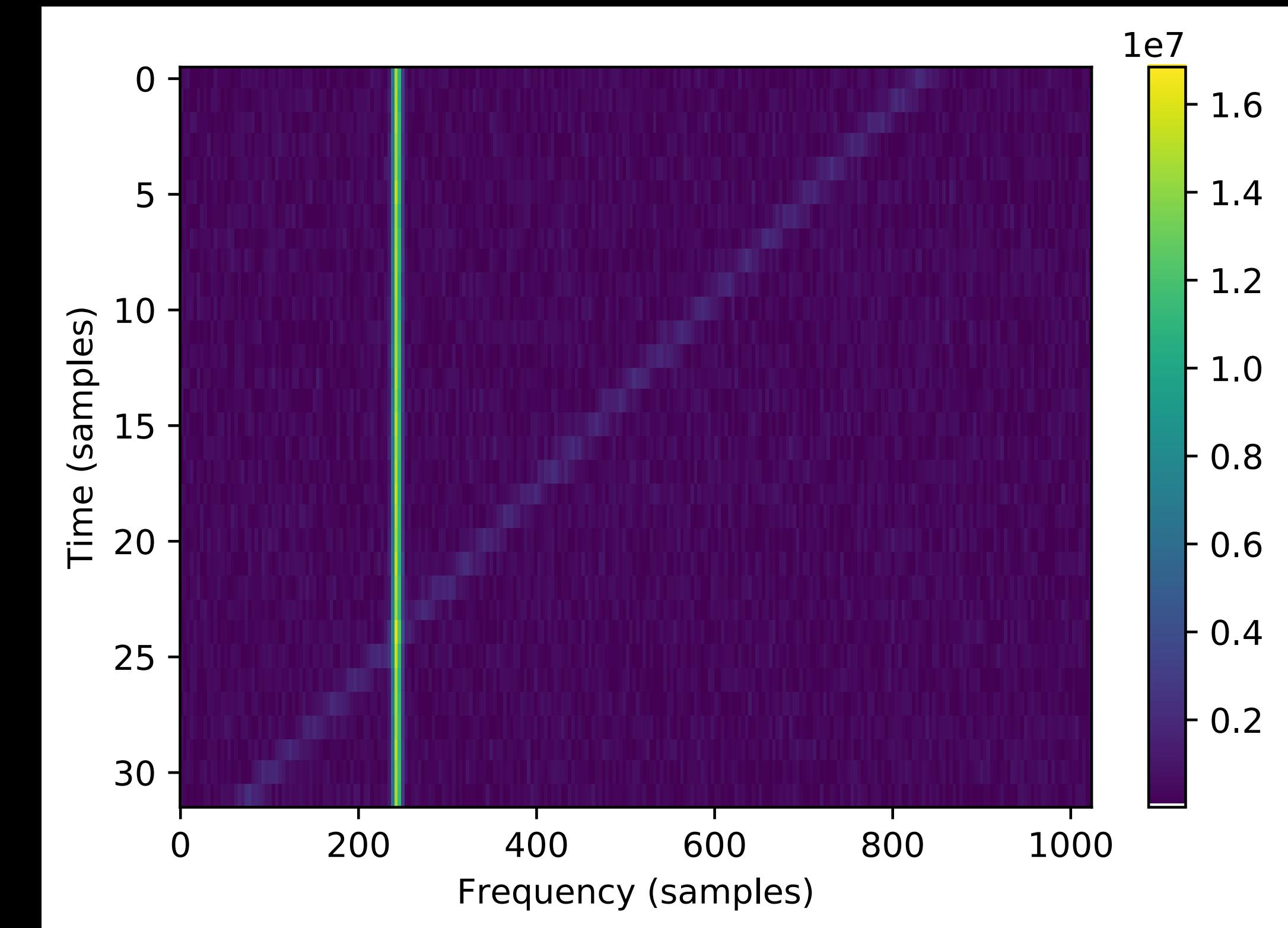
- With a means of simulating observational data, we can produce frames that would be hard to find “organically” using TurboSETI
- Importantly, we can generate lots of relevant labeled synthetic data, and train a CNN to find these signals
- Localization of narrow-band signals is a good initial ML problem because it’s a relatively simple task; predict 2 numbers per signal



Example of a frame with 2 synthetic signals, at 25 and 15 dB.

# NARROW-BAND SIGNAL LOCALIZATION (BRZYCKI ET AL. 2020)

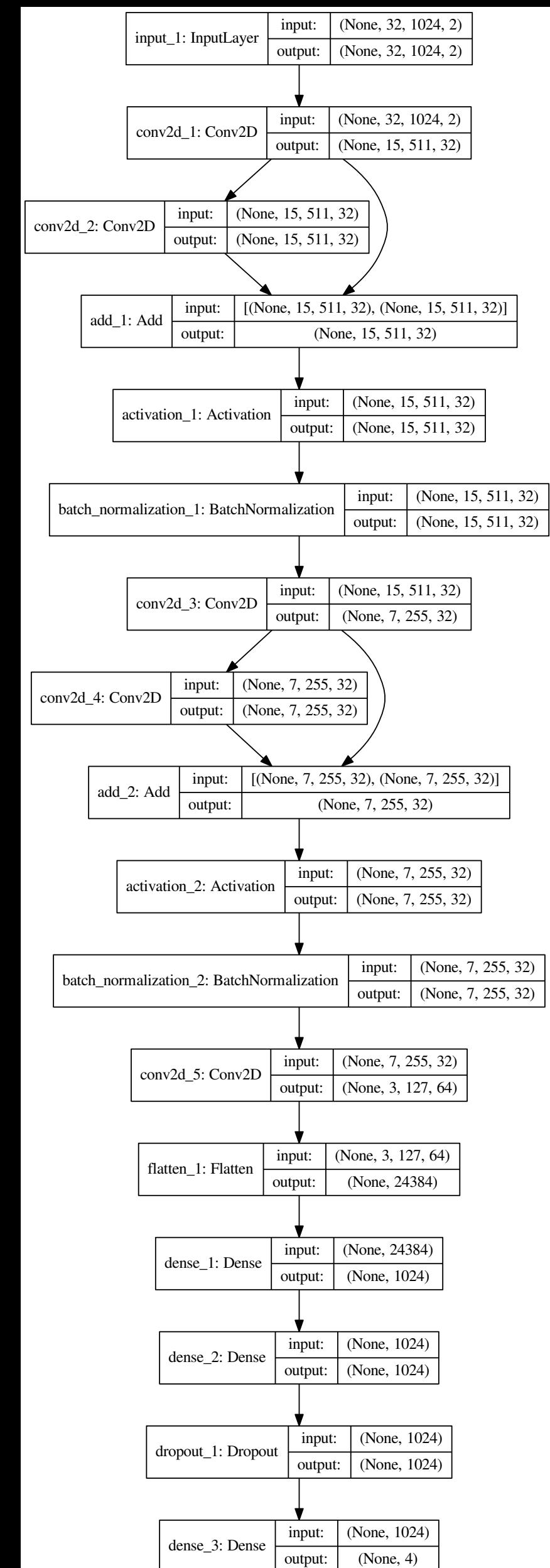
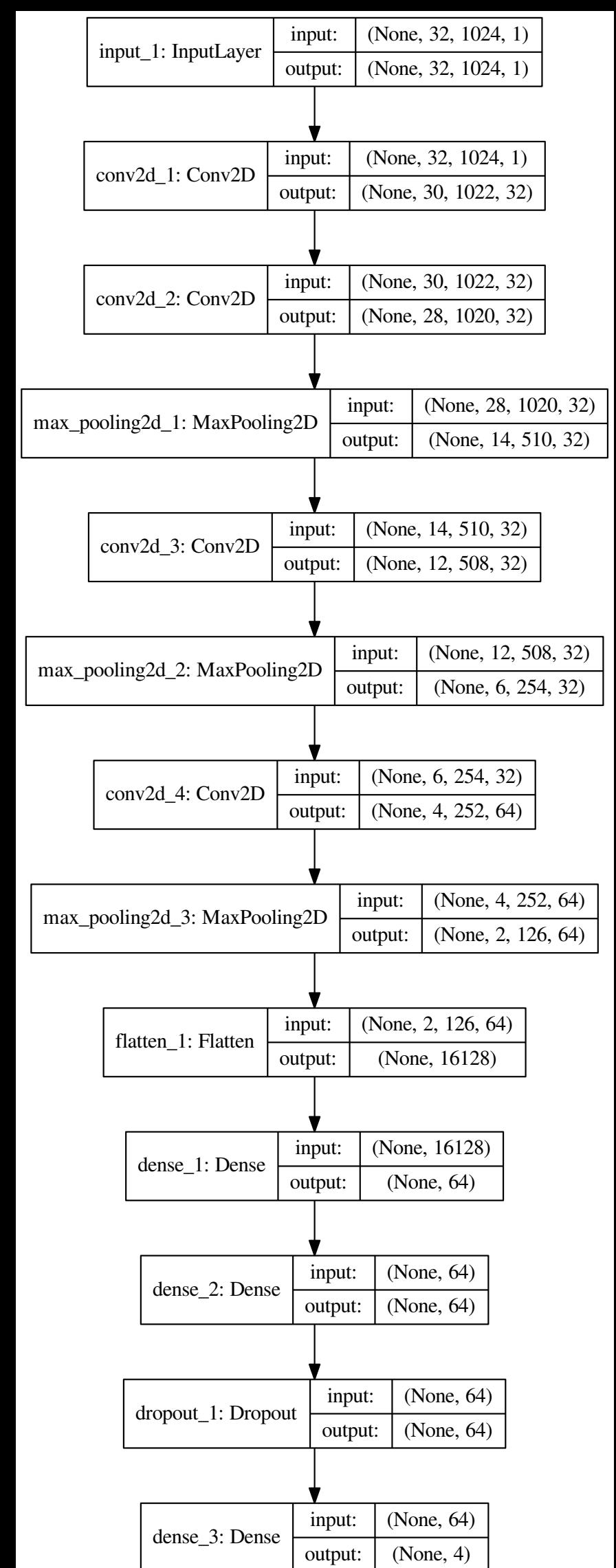
- Created two main datasets, both with 120,000 training samples and 24,000 test samples of size 32x1024 px
  - One signal, at 0, 5, ..., 25 dB with random drift rate
  - Two signals, one at 0, 5, ..., 25 dB with random drift rate, and the other at 25 dB with 0 drift rate (meant to simulate “bright” RFI)
- The one signal dataset allows for direct comparison with TurboSETI; the two signal dataset tests the effectiveness of localizing multiple signals simultaneously



Example of a frame with 2 synthetic signals.

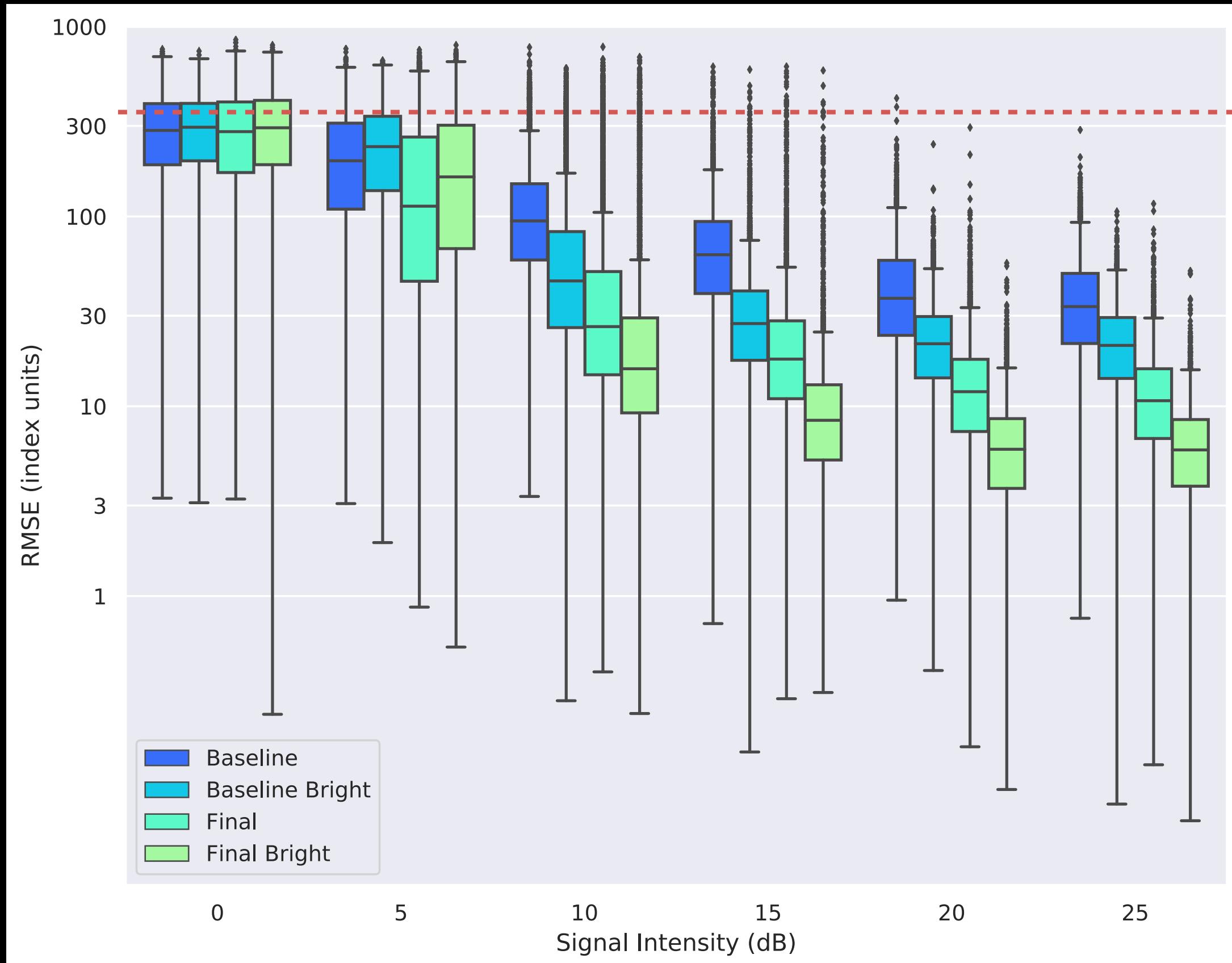
# MODEL ARCHITECTURES

- Used convolutional neural networks, especially suited for image input data
- Created a “baseline” and a “final” model, to compare performance:
  - Baseline model uses convolutional layers, max pooling, and fully connected layers
  - Final model includes residual connections, stride 2 convolutions instead of max pooling, and batch normalization
- In addition to training these models over all input training data, we did alternate training over only 10 - 25 dB signal frames, labeling these as “bright” models

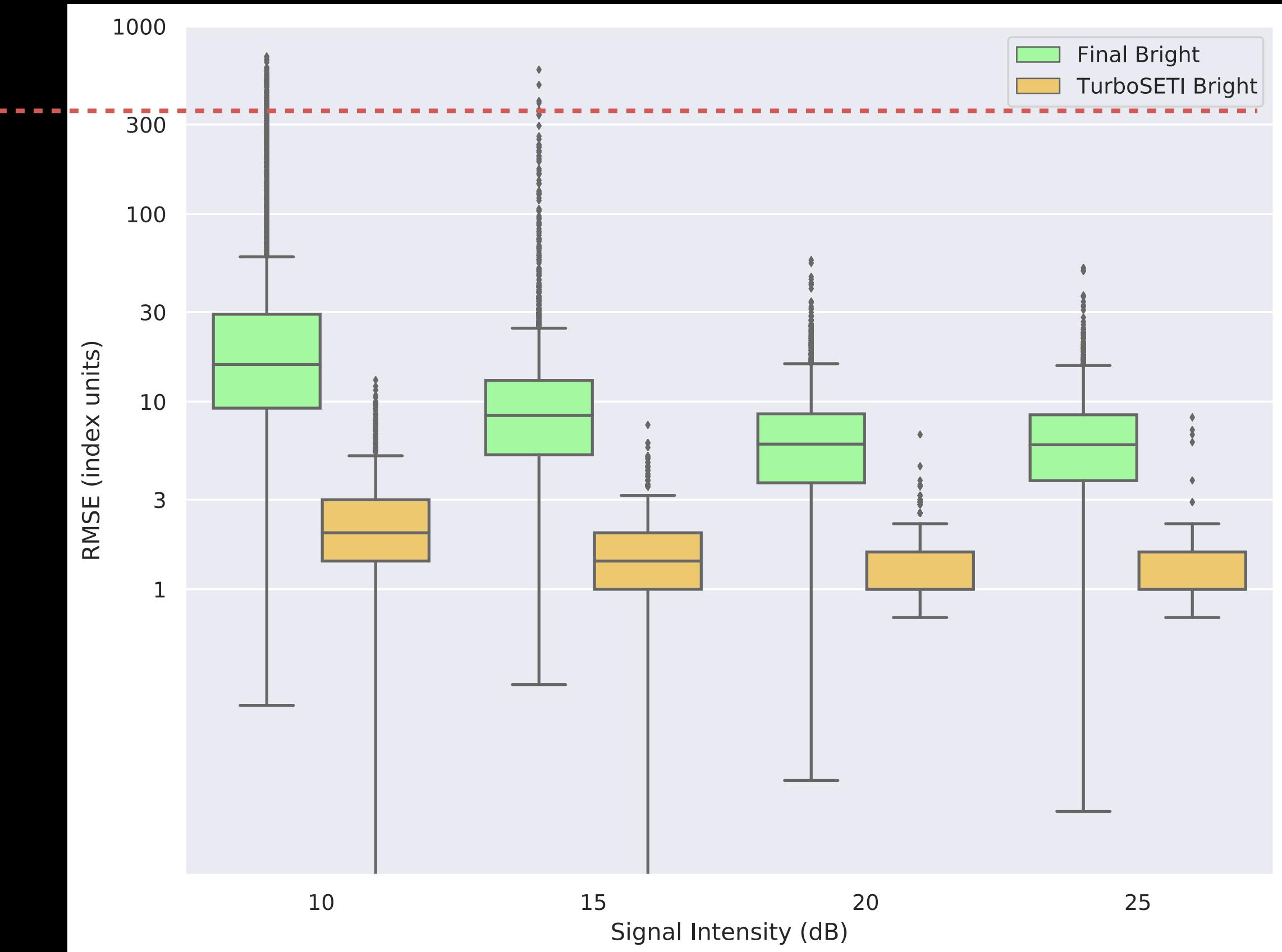


$$\text{RMSE (index units)} = 1024 \times \sqrt{\frac{1}{n} \sum_i^n (y_i - \hat{y}_i)^2}$$

# ONE SIGNAL RESULTS ON TEST DATA



**Root mean squared error (in pixels) across different models**

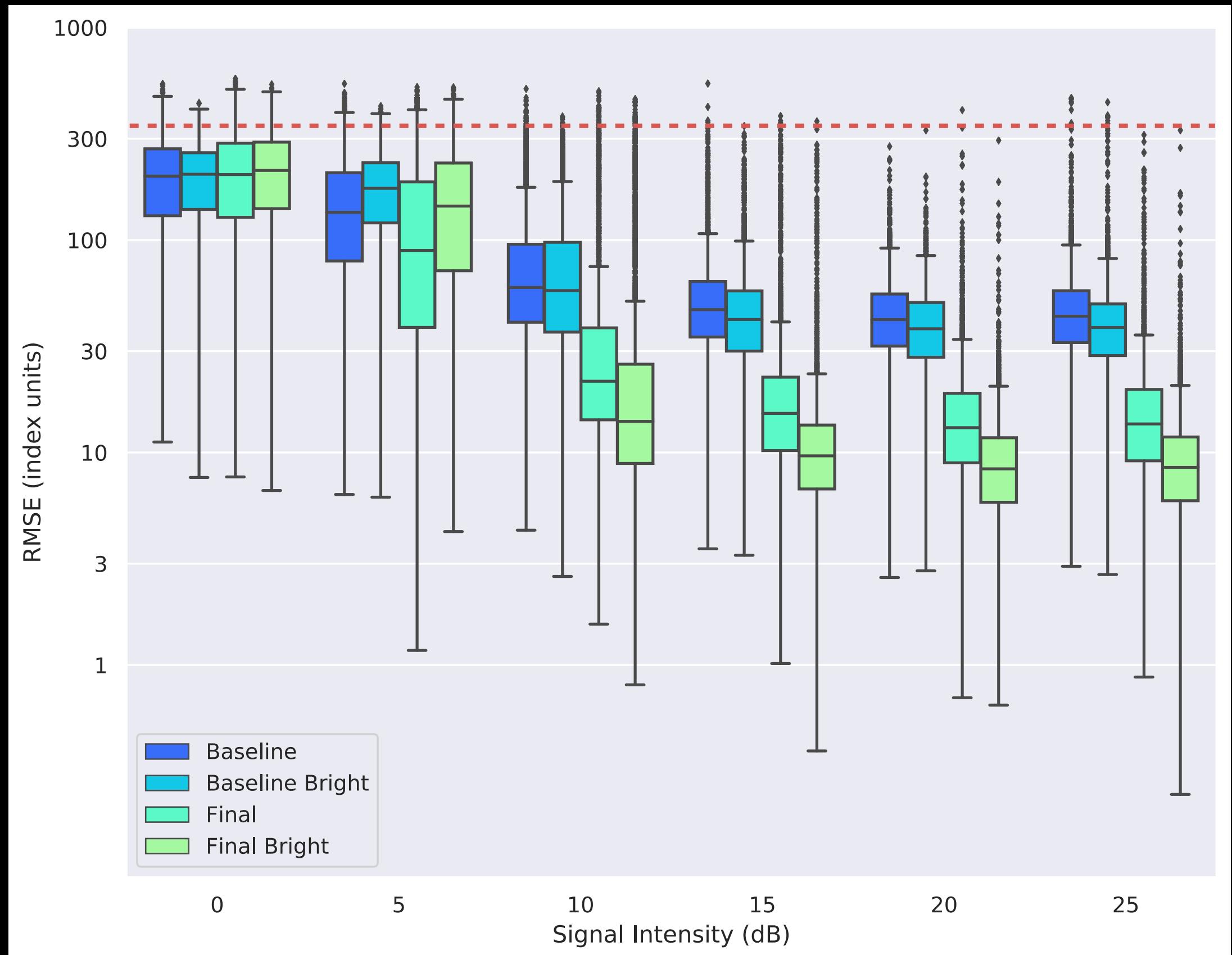


**Root mean squared error (in pixels) for best ML model vs. TurboSETI. Only calculated for SNR > 10.**

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

# TWO SIGNAL RESULTS ON TEST DATA

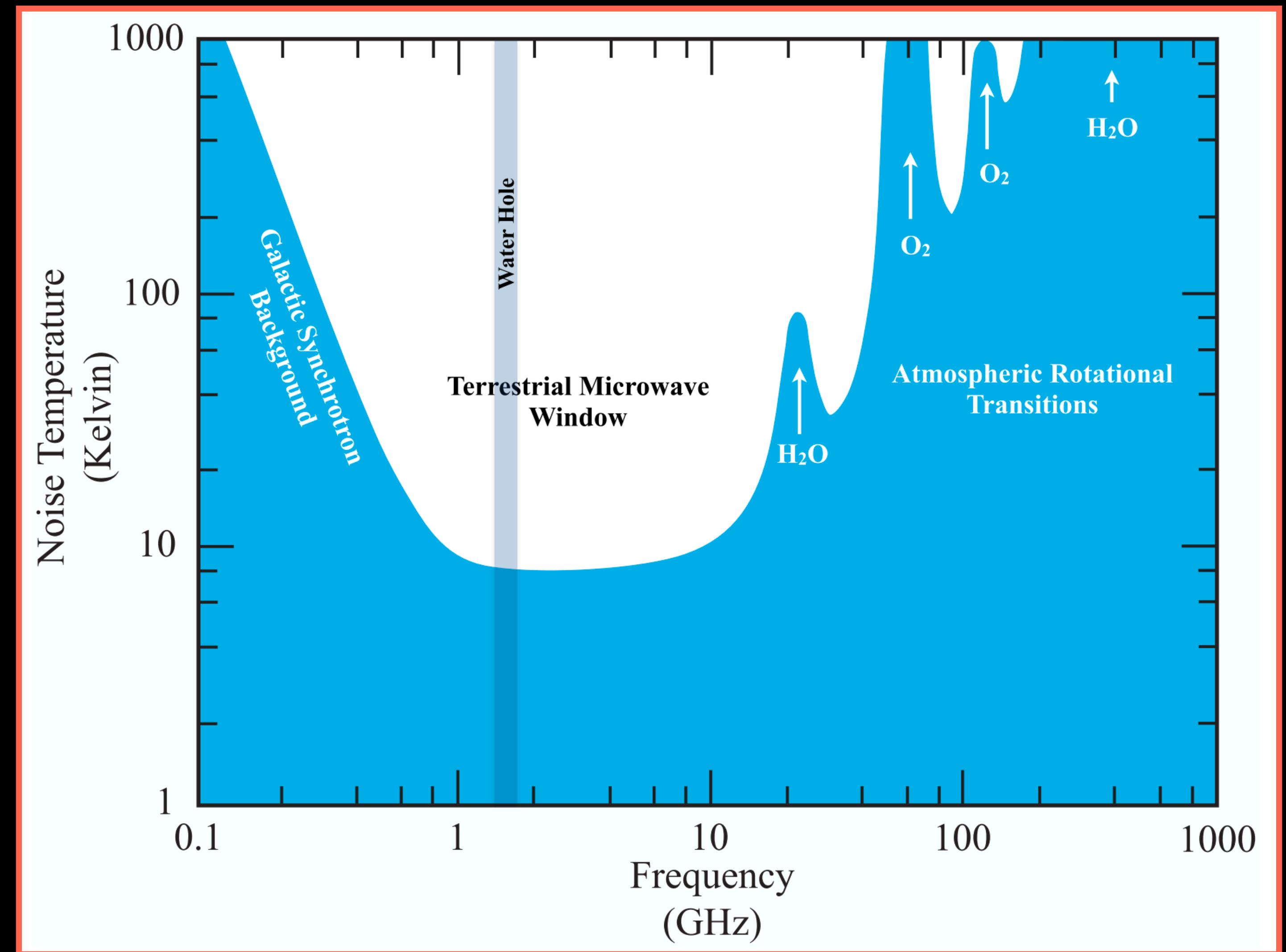
- **Performance over two signal case is slightly worse than in the one signal case**
- **Even though we used ideal synthetic signals, our best models failed to localize to extremely high precision**
- **Nevertheless, our two signal model architectures were able to localize the dimmer signal better than random, so these results are still encouraging**



**Root mean squared error across different signal intensities, in pixels, for various neural network architectures in the 2 signal case.**

# Why radio?

- Low energy to produce
- Low attenuation
- Produced by technology!



Siemion et al.

# Prior SETI research on scintillation

- Cordes & Lazio 1991 recommended multi-epoch observing campaigns
- Many studies acknowledge the possibility of scintillation
- Generally, SETI techniques aren't sensitive to detailed morphology
- Stochastic effects are hard to describe

INTERSTELLAR SCATTERING EFFECTS ON THE DETECTION OF NARROW-BAND SIGNALS  
JAMES M. CORDES AND T. JOSEPH LAZIO  
National Astronomy and Ionosphere Center and Department of Astronomy, Cornell University, Ithaca, NY 14853  
*Received 1990 October 4; accepted 1991 January 15*

