Draft version May 23, 2025
Typeset using LaTeX **preprint** style in AASTeX631

SETI@home: Data Analysis and Findings

David P. Anderson,[1] Eric J. Korpela,[1] Dan Werthimer,[1] Jeff Cobb,[1] and Bruce Allen[2]

[1] *Space Sciences Laboratory, University of California, Berkeley, CA 94720-7450*
[2] *Max Planck Institut für Gravitationsphysik (Albert Einstein Institut), Hanover, Germany*

(Dated: May 23, 2025)

ABSTRACT

SETI@home is a radio Search for Extraterrestrial Intelligence (SETI) project that looks for technosignatures in data recorded at the Arecibo Observatory. The data were collected over a period of 14 years and cover almost the entire sky visible to the telescope. The first stage of data analysis found billions of *detections*: brief excesses of continuous or pulsed narrowband power. The second stage removed detections that were likely radio frequency interference (RFI), then identified and ranked *signal candidates*: groups of detections, possibly spread over the 14 years, that plausibly originate from a single cosmic source. We manually examined the top-ranking signal candidates and selected a few hundred. In the third and final stage we are reobserving the corresponding sky locations and frequency ranges using the Five-hundred-meter Aperture Spherical Telescope (FAST) radio telescope. This paper covers SETI@home's second stage of data analysis. We describe the algorithms used to remove RFI and to identify and rank signal candidates. To guide the development of these algorithms, we used artificial *candidate birdies* that model persistent ET signals with a range of power, bandwidth, and planetary motion parameters. This approach also allowed us to estimate the sensitivity of our detection system to these signals.

Keywords: Radio Spectroscopy (1359), Search for Extraterrestrial Intelligence (SETI, 2127), Technosignatures (2128), Radio Frequency Interference (RFI), Radio Astronomy, Volunteer Computing, Distributed Computing

1. INTRODUCTION

1.1. *Background*

davea@berkeley.edu
korpela@berkeley.edu

The question of whether life exists in other parts of the universe is important and unanswered. The 1952 Muller-Urey experiment (Miller 1953; Miller & Urey 1959) demonstrated the possibility of abiotic production of the molecular components of living systems. The detection of amino acids in meteorites (Pearce & Pudritz 2015) and prebiotic molecules in interstellar space (Zeng et al. 2019; Rivilla et al. 2023) showed that such processes are possible even outside a planetary atmosphere.

The direct detection of living organisms outside the Solar System remains unlikely in the near future. A more likely scenario is an indirect detection, such as an atmospheric biosignature: a compound released into the atmosphere by biological processes. However, such compounds may also have an abiogenic source, so it is uncertain whether such a detection indicates life (Tokadjian et al. 2024; Court & Sephton 2012).

Detection of intelligence would provide more certain evidence of life. An extraterrestrial intelligence (ETI) could create artifacts, signals, or processes that are detectable at interstellar distances and have no natural counterpart. Such processes could be a form of radiation (electromagnetic, particle, or gravitational) or a physical artifact (a spacecraft or object passing through or remaining in the Solar System, a structure detectable at interstellar distance, or an atmospheric component that only has a technological means of production). Such items are collectively known as *technosignatures* (Haqq-Misra 2024).

Due to the relative ease of creating and detecting radio waves and the relative transparency of atmospheres and interstellar space to such waves, radio has been proposed as a means of detecting extraterrestrial intelligence (Cocconi & Morrison 1959). Two primary approaches have been used for such searches: *sky surveys* cover a large fraction of the solid angle of the entire sky, and *targeted searches* focus on individual stars or galaxies (Drake 1974).

Several targeted searches have been performed, including OZMA and OZMA II at Green Bank (Drake 1960; Sagan & Drake 1975; Drake 1986; Gray 2021), Phoenix at the Arecibo Observatory (Backus & Project Phoenix Team 2002) and at the Allen Telescope Array (ATA), and Breakthrough Listen projects at the Parkes and Green Bank observatories (Price et al. 2020; Enriquez et al. 2017). Recently, Breakthrough Listen has begun to observe targets at the Very Large Array (Tremblay et al. 2024) and MeerKAT (Czech et al. 2021). In addition, observations of multiple targets have been made at the FAST observatory in China (Luan et al. 2023) and the ATA (Tusay et al. 2024).

There have also been a number of sky surveys. Some have operated *commensally*, collecting data from a telescope while its pointing was being controlled by other projects. These include searches using various generations of the SERENDIP spectrometer at the Hat Creek and Green Bank observatories (Werthimer et al. 1988) and at the Arecibo observatory (Cobb et al. 2000; Bowyer et al. 2016). Other sky surveys used dedicated telescopes. These include the early Ohio State project and its "Wow!" signal (Kraus 1977), the "Fly's Eye" project (Siemion et al. 2012) and a brief survey of the anti-solar point (Hort et al. 2024) at the ATA.

To date, no repeatable detections of interstellar technosignatures have been made.

1.2. *SETI@home*

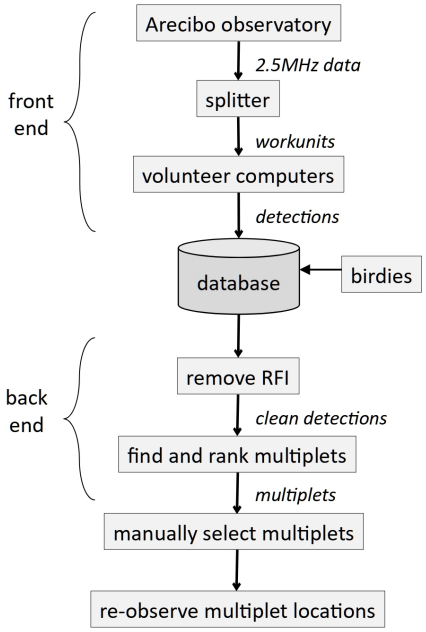

Figure 1. The SETI@home data analysis process.

The SETI@home project, described here, is a commensal sky survey. We recorded and analyzed data from the Arecibo radio observatory from 2006 through 2020, looking for radio technosignatures.

As shown in Figure 1, the *front end* of SETI@home digitizes and records the radio signal, then processes these data to find *detections*: brief excesses of continuous or pulsed narrowband power. This processing is compute-intensive; it was done using the Berkeley Open Infrastructure for Network Computing (BOINC) volunteer computing system (Anderson 2020) with a pool of several hundred thousand home computers. The front end of SETI@home is described in detail in Korpela et al. (2025).

The *back end* of SETI@home takes these detections (about 12 billion of them) as input, identifies and removes radio frequency interference (RFI), and finds signal candidates, or *multiplets*: groups of detections whose properties are consistent with those of an ET *target signal*. These detections are selected from the entire duration of the project. Each multiplet is assigned *scores* inversely proportional to the probability that it is the result of random noise.

The output of the back end is a set of about 20 million multiplets, ranked by score. These were examined, in decreasing score order, by experts trained to distinguish possible ET signals from noise and RFI. This process is labor intensive, and we were able to examine only ~ 1000 multiplets. Thus, we say that we have *uncovered* a target signal if a multiplet that includes detections from the signal appears in the 1000 top-scoring multiplets.

The manual inspection produced a set of roughly 200 multiplets. Each of these will be reobserved: its sky position and frequency range will be examined with greater sensitivity than was originally available. In addition, we will control the telescope pointing, so that each location can be observed longer, and we can use techniques such as on/off observing.

The back end involves many heuristic algorithms. We want to ensure that these algorithms work as intended: in other words, if our data contained a target ET signal, then a) the detections resulting from the signal would not be rejected as RFI, b) the back end would find a multiplet composed of these detections, and c) the multiplet would have sufficiently high scores, compared with noise multiplets, that it would be uncovered in the above sense. For this purpose, we inject artificial *birdies* into our data, modeling ET signals with a range of parameters (e.g. flux, position, planetary motion of the transmitter). If the back end uncovers these birdies, we have some confidence that it would uncover a similar ET signal.

The probability that SETI@home uncovers a target ET signal depends on several factors: most notably the flux of the signal, its bandwidth, and the length of our observations of its sky position. The birdie signal injection mechanism gives us a way to estimate this probability as a function of these parameters: we can insert many birdies with various parameters, and see how many we uncover – i.e., what fraction of birdies produce multiplets whose scores rank in the top 1000 non-birdie multiplets.

This paper describes the SETI@home back end and presents its results. §2 describes the types of signals we are looking for. §3 summarizes the SETI@home front end. §4 describes the sky coverage of our data. The birdie mechanism is described in §5. Sections 6 and 7.10 describe the algorithms for RFI removal and multiplet detection. §8 describes the implementation of the back end. §9 presents our results. §10 discusses related work and the contributions of SETI@home, and §11 proposes future work.

2. TARGET SIGNALS

SETI@home looks for a range of *target signals* with characteristics typical of technological origin, but not known to occur naturally. Specifically, we look for a) continuous narrowband signals whose power is concentrated in a small frequency band (< 1221 Hz), b) pulsed narrowband signals, also with bandwidths up to about 1221 Hz, that turn on and off with a constant but unknown period and duty cycle, and c) signals having a repeating structure with periods up to several seconds; see Harp et al. (2016).

Natural sources of such signals are not known. The narrowest known natural source of radio emissions (OH masers, Cohen et al. 1987) is about 500 Hz in bandwidth, but such masers are not found in the band observed by SETI@home. The narrowest known emissions in the SETI@home band are about 1 km/s (4750 Hz) in width (Peek et al. 2010, 2011). Similarly, there are no known natural sources of narrowband pulses. Pulsars generate broadband pulses on the timescales covered by SETI@home, but these pulses are removed by a filter that eliminates broadband features from the SETI@home analysis (Korpela et al. 2025).

We look for signals that are transmitted over a long period (ideally our entire observation period) and with a high duty cycle. It is unlikely that we would detect a transient signal (one lasting a few seconds or minutes), because the probability that we would be observing its location during that period is small.

We assume that the signal transmitter is in a nearly inertial frame (or Doppler corrected to appear as such), or is an uncorrected transmitter on the surface of a rotating planet orbiting a star, in orbit around a planet orbiting a star, or orbiting a star.

We assume that the movement of the transmitter relative to Earth is small enough that its sky position as viewed from Earth is nearly constant during the 14 years observing period. This would not be true for a transmitter located within the Solar System.

2.1. *Doppler shift*

The frequency at which Arecibo receives a signal is Doppler-shifted by the velocities of both transmitter and receiver in the direction of the signal path at the times of transmission and receipt, respectively. The shift due to receiver motion, and the time derivative of this shift, are known. They are a result of the accelerations of the Arecibo receiver's reference frame due to Earth rotation and Earth orbit.

The sender could apply a correction to the transmission frequency to cancel the changing shift due to the transmitter's acceleration. If a signal is intended as a beacon and is directional, it presumably would be adjusted in this way. We call such a signal *barycentric*. After the correction for receiver motion is applied, the signal would be received at a nearly constant frequency.

Other potential ET signals (leakage or omnidirectional beacons) would not be corrected in this way and would have a Doppler shift corresponding to the changing velocity of the transmitter. We call such signals *nonbarycentric*. After receiver correction, they would appear to drift in frequency at a varying rate. The ranges of the sender shift and its derivative depend on the movements of the transmitter. We look for target signals for which these ranges are consistent with the motions of habitable-zone planets orbiting F and G type stars. These assumptions are detailed in §9.

3. THE SETI@HOME FRONT END

The SETI@home front end is described in a companion paper Korpela et al. (2025). We summarize it here.

3.1. *The Arecibo radio telescope*

SETI@home used the radio telescope at the Arecibo Observatory (AO), a fixed dish with 305-meter aperture. Its maximum usable viewing angle was 20° away from the zenith. Given its location at latitude 18° north, this resulted in a visible area consisting of a band from -2° to 38° Dec, covering about 25% of the sky. SETI@home can detect signals only within this area (Korpela et al. 2011).

Various receivers could be positioned in the focal plane. Our observations began in 1998 using a single-polarization line-feed. (Korpela et al. 2001; Korpela et al. 2002) In 2006, observations continued using the 7-beam dual polarization Arecibo L-band Feed Array (ALFA) receiver (Cordes et al. 2006). The ALFA receiver provided significantly better sensitivity (>10×) and a wider field of view; it also enabled multibeam RFI rejection (§6.3). Hence, the observations using ALFA superseded the earlier observations.

ALFA consisted of 14 separate receivers, grouped in polarization pairs, or *beams*. The seven beams were arranged in a hexagonal pattern. The sensitivity of a given receiver can be approximated by a Gaussian function of the angle from its center. The half-power width of a beam was 0.°05; we denote this θ_{beam}.

The beams pointed in slightly different directions, so they are subject to different Doppler shifts. We denote the shift for beam b at time t by $\Delta\nu_{\rm AO}(b,t)$, and its time derivative by $\frac{d(\Delta\nu_{\rm AO})}{dt}(b,t)$. The largest component of $\Delta\nu_{\rm AO}$ is typically due to the orbital velocity of Earth. $\frac{d(\Delta\nu_{\rm AO})}{dt}$ is dominated by acceleration towards Earth's rotational axis and is therefore negative, with a value of ~ -0.16 Hz s^{-1}.

3.2. *Data recording and splitting*

SETI@home covers a frequency band of 2.5 MHz centered at 1.42 GHz, near the fine structure transition of the ground state of HI. We chose the hydrogen line because it is considered to be a likely frequency for deliberate beacon transmissions. ETI who are aware of the HI transition are likely to use it to study the structure of the galaxy, so there are potentially a large number of observers at this frequency.

The SETI@home band is limited to 2.5 MHz because of performance limitations in recording data and distributing it over the Internet. This band is sufficiently large to contain signals near 1.42 GHz, even after Doppler shifting by receiver motion and by the target range of transmitter motion.

The analog feed from each of ALFA's 14 receivers (7 feeds with dual linear polarizations) goes into a *data recorder* computer, where it is down-converted from 1.42 GHz to quadrature baseband. A low-pass filter eliminates signals outside the central 2.5 MHz of the band. The resulting signal is sampled at 2.5 Msps, with each sample being a 2-bit complex number, one bit real and one bit imaginary. The resulting data encode the entire 2.5 MHz frequency band centered at 1.42 GHz.

The sky position (RA/Dec) of each ALFA beam is sampled once per second. These *pointing records* are included in the data stream.

The digitized data were transported from Arecibo to Berkeley on disk drives. It was divided using a polyphase filter bank into 256 frequency *subbands* of 9765.625 Hz each. Each subband was sampled at that rate (with complex samples). These streams of samples were then divided into segments of length 2^{20} samples = 1 Mi samples (107.37 s in duration). We call these segments *workunits*. Workunits in a given subband overlap in time by approximately 20 seconds, so that the longest features of interest – 13 seconds or so – are always contained entirely in at least one workunit.

3.3. *Detections*

The workunits were analyzed on home computers using a program that finds detections. Details of this analysis are given in Korpela et al. (2025); we briefly summarize it here.

The data were converted to the frequency domain using the discrete Fourier transform (DFT). We used 15 DFT lengths, ranging by powers of two from 8 samples (1221 Hz resolution, 8.1×10^{-3} s) up to $128Ki$ samples (1221 Hz resolution, 13.4 s). Long DFTs are sensitive to continuous narrowband signals; short DFTs are sensitive to short or rapidly pulsing signals. (See Korpela et al. 2025 Sec. 5.3 for a discussion of sensitivity.) For each DFT length, we computed a sequence of DFTs on the data, producing a 2-D array of power as a function of time and frequency.

As described earlier, received signals may drift in frequency due to accelerations of transmitter and receiver. Rather than looking for features that drift in the power arrays, such signals can be detected with greater sensitivity using coherent integration, in which the data were *de-drifted* at a specific

Detection type	Number of detections
Spike	5,031,283,756
Gaussian	317,563,087
Pulse	3,078,135,522
Triplet	2,561,366,366
Autocorrelation	1,118,691,234

Table 1. Detection counts

Doppler drift rate before computing DFTs. This can put drifting features into single frequency bins where they can be more easily detected.

The program first performed a baseline smoothing operation on the data, removing any features wider than about 2 KHz. The data were then de-drifted at a range of Doppler drift rates corresponding to the range of planetary motions under consideration (typically 123 000 rates, ranging from $-100\,\mathrm{Hz\ s^{-1}}$ to $100\,\mathrm{Hz\ s^{-1}}$; see Korpela et al. (2025)). For each drift rate, the de-drifted data were analyzed at each of the 15 DFT lengths: we computed the 2-D power array, then looked for features in the array. We looked for several types of features, or *detections*:

Spike: A short continuous narrowband signal: a single DFT bin whose power is at least 24 times the mean noise power. This threshold was chosen so that workunits containing Gaussian noise yield an average of 1 spike; in real data the average is about 7 spikes.

Gaussian: A longer continuous narrowband signal: a sequence of DFT bins at the same frequency whose powers approximate the Gaussian-shaped envelope that would result from the beam moving over a fixed source, given the beam's angular velocity during the workunit.

Pulse: A pulsed narrowband signal: the time sequence of DFT powers at a given frequency approximates a pulsed signal. These are found using a folding algorithm (Staelin 1969; Korpela et al. 2025).

Triplet: A simple type of pulsed narrowband signal consisting of a group of three DFT bins at the same frequency, above a threshold, and equally spaced in time.

Autocorrelation: A signal having a repeating structure with periods up to ±6.7s. These are found by computing the autocorrelation of the data with a range of delays, looking for delays where the correlation is above a threshold. This detects any periodic structure in the data. See Korpela et al. (2025) for more details.

The number of detections of each type and their distribution among DFT lengths is shown in Tables 1 and 2.

We denote the properties of a detection D as follows:

$t(D)$**:** The time midpoint of the DFT bin, or range of bins, where D occurred.

$b(D)$**:** The ALFA beam (0 to 6) in which D was found.

DFT length samples	DFT duration s	Bandwidth Hz	Spike %	Gaussian %	Pulse %	Triplet %	Autocorrelation %
8	0.0008	1220.7			6.15	7.25	
16	0.0016	610.3			1.13	10.4	
32	0.0032	305.1	0.001		3.80	18.6	
64	0.0065	152.6	0.013		7.33	21.8	
128	0.0131	76.2	0.047		12.2	18.0	
256	0.0262	38.1	0.091		15.7	11.6	
512	0.0524	19.1	0.223	0.0	16.1	6.45	
1024	0.104	9.53	0.391	0.003	14.9	3.07	
2048	0.209	4.76	0.717	0.039	13.3	1.69	
4096	0.419	2.38	1.35	0.301	7.66	0.694	
8192	0.838	1.19	2.82	0.796	1.46	0.092	
16384	1.67	0.596	6.59	98.862		0.0	
32758	3.35	0.298	11.5				
65536	6.71	0.149	15.5				
131072	13.42	0.074	60.6				100.0

Table 2. Percentage of detections as a function of DFT length

$pos(D)$**:** The sky position of the beam's center at time $t(D)$. This is computed by linearly interpolating between the two beam pointings before and after $t(D)$.

$\tau(D)$**:** The duration of the detection. For spikes, this is the DFT bin duration. For Gaussians, it is twice the standard deviation of the Gaussian curve. For pulses and triplets, it is the beam-crossing time based on the workunit's average beam velocity (and at most the workunit's duration, 107.37 s). For autocorrelations it is the duration of the longest DFT length.

$S(D)$**:** An estimate of the probability of D occurring in noise. For spikes and triplets this is based on peak power. For Gaussians, it's based on power and goodness of fit. For pulses, its statistics based on the number of samples added into each bin of the folded array. These values are negated, so high scores are better. See Korpela et al. (2025) for details.

Detection types other than autocorrelation also have the following attributes:

$P(D)$**:** The detection power as a multiple of the mean noise power.

$\ell(D)$**:** The DFT length at which D was found.

$\nu_{\text{topo}}(D)$**:** The frequency in the reference frame of the observatory; that is, the frequency of the DFT bin where D occurred, correcting for de-drifting. This is used in RFI detection, since terrestrial RFI isn't Doppler-shifted by receiver movement.

$\nu_{\text{bary}}(D)$**:** The frequency of D adjusted for Doppler shift due to the receiver's velocity in the direction of the beam center at $t(D)$. This is used for all purposes other than RFI detection.

$\frac{\Delta\nu_{\text{topo}}}{\Delta t}(D)$**:** The dechirp value (see above) at which D was found.

Pulses and triplets have an additional attribute:

$p(D)$**:** The time between consecutive crests.

Autocorrelations have an additional attribute:

$\delta\tau(D)$**:** The delay at which the correlation occurred.

3.4. *Position and frequency uncertainty*

Detection positions are uncertain. Suppose, for example, that an ET signal emanates from a particular sky position P. The signal could result in a detection D from a particular beam. The position of this detection, $pos(D)$, is the center of the beam at time $t(D)$. Since the beam's sensitivity profile has nonzero width, P could differ from $pos(D)$. In fact, the beam has nonzero sensitivity over the entire sky, so a sufficiently powerful signal could be detected no matter where the telescope is pointing. However, ET signals are presumed to have low received power, so we assume that P is within the beam's half-power disk.

Thus, we define a *position uncertainty*, $\sigma_{pos} = \theta_{\text{beam}}$. If detections lie in a disk of radius σ_{pos} centered at P, it is plausible that they result from an ET signal emanating from P. Detections outside this disk are unlikely to result from such a signal.

Equivalently, if S is a set of detections and the maximum angle between their positions is at most $2\sigma_{pos}$, it is plausible that the detections in S originate from a single ET signal.

Similarly, the frequency of detections is uncertain. Suppose an ET signal is barycentric: that is, its transmission frequency is corrected for the Doppler drift due to the transmitter accelerations, or the transmitter is unaccelerated and no correction is needed. This signal could result in detections D with barycentric frequencies $\nu_{\text{bary}}(D)$ that are within an interval of nonzero width. The deviation can have several sources.

First, the correction for receiver Doppler drift is based on the direction of the beam center, but the sky position of the signal may differ from this. At an angular distance σ_{pos}, this results in a maximum error of about 40 Hz.

Second, assuming that the signal is being transmitted in a beam of nonzero width, there could be an analogous error in its Doppler drift correction. It is difficult to estimate the magnitude of such an error without knowing the distance to and design of the transmitter. It is likely, however, that deliberate transmission towards our Solar System would be directed with much better precision than an Arecibo beam width, so we estimate this uncertainty to be no more than a few Hz.

Third, recall that the SETI@home front end does coherent integration with a discrete set of drift rates. The step size is chosen so that the frequency between successive steps changes by one-half the bin width over the duration of a bin. If the actual drift of the signal due to receiver acceleration lies halfway between steps, this can result in a frequency difference. The maximum difference depends on the DFT length. For DFT lengths of 128 or more (which include 99.99% of spikes), it is 76 Hz.

Summing these uncertainties in quadrature gives an approximate maximum frequency uncertainty, $\sigma_\nu = 125\,\mathrm{Hz}$.

If a set of detections has barycentric frequencies lying in an interval $[\nu - \sigma_\nu, \nu + \sigma_\nu]$, it is plausible that the detections result from an ET signal with barycentric frequency ν. Detections with $\nu_{\mathrm{bary}}(D)$ outside this interval are unlikely to result from such a signal.

The quantities σ_{pos} and σ_ν play a key role in the identification of multiplets. We only consider sets of detections whose positions and frequencies vary by at most twice the corresponding uncertainty factor (see §7.10).

4. OBSERVATIONS AND SKY COVERAGE

4.1. *Observation modes*

SETI@home observed *commensally*; that is, it did not control the pointing of the telescope during its observations. Pointing was determined by other uses of the ALFA receiver: searching for pulsars near the plane of the Galaxy, mapping the distribution of hydrogen in all parts of the Galaxy visible from Arecibo, and searching for extragalactic hydrogen gas in isolated clouds and in nearby galaxies.

The pointing of the telescope may be roughly divided into four modes.

Tracking: The telescope tracks a fixed point in the sky. This mode is typically used for pulsar surveys, with dwell times ranging from 30 seconds to tens of minutes (Cordes et al. 2006).

Drift scan: The telescope is not moving and the beams move across the sky at the sidereal rate, $\sim 15'' \mathrm{s}^{-1}$. Objects in the sky pass through the beam of one of the ALFA receivers (0.°05) in about 12 seconds. This mode is typically used for extragalactic hydrogen surveys.

Basket-weave scan: The telescope does a zigzag scan of the sky, in a path designed to cover a large area of sky. The crossing points of this path can be used to determine the relative gain of the receivers, since each is measuring the same astronomical signal at that point. The angular velocity is typically near $90'' \mathrm{s}^{-1}$.

Slewing: The telescope moves rapidly between distant points in the sky. The angular velocity can be as low as $0.04^\circ\ \mathrm{s}^{-1}$ for pure zenith angle motions and as high as $0.4^\circ\ \mathrm{s}^{-1}$ for motions involving the azimuth axis.

See Peek et al. (2011) for details of the drift scan and basket-weave scan techniques.

SETI@home's observations alternated unpredictably between these modes. Our algorithms that involve pointing (such as RFI removal and candidate scoring) were required to work for the full range of pointing trajectories. The maximum continuous observation at Arecibo was 0.1 day. Any observations more widely separated than this value were at least 0.9 days apart. This 0.1 day value was used as a limiting case for some algorithms.

4.2. *Observation intervals*

During the 14 yearsof observing, SETI@home recorded about 400 days of data, or about 9% of the total time. The ALFA receiver was in use only part of the time, and the SETI@home data recorder was sometimes offline.

Scoring multiplets (see §7.9) requires knowing the periods during which we observed each sky position. For example, suppose that a multiplet consists of detections from a one-minute span. If we observed that sky position for only that minute, the multiplet should score higher than if we observed it for an hour and found no similar detections in the other 59 minutes. Calculating these *observation intervals* requires some approximations.

First, we quantize sky position using the HEALPix system (Górski et al. 2005), an equal-area pixelization of the sky. We use the resolution parameter $N_{SIDE} = 2^{11}$ for which there are $12 \times 2^{22} \sim 50$M pixels, resulting in an average pixel scale of 0.°0286, about half of the ALFA half-power beamwidth of 0.°05. Approximately 15M of the 50M pixels are visible using the ALFA receiver.

Second, we say that a beam observed a pixel at a particular time if the beam's center is within θ_{beam} of the pixel's center at that time.

Third, the rate at which we sample sky position is only 1 Hz (see §3.2). During rapid movement, the sampled pointings can skip over pixels. To fill in these gaps, we assume that the beam position is linear in RA and Dec between samples.

For a pixel P we let $I(P)$ denote the set of time intervals during which at least one beam observed P; these are the times during which a signal emanating from P could plausibly be detected in our data. An algorithm for calculating $I(P)$, given the above approximations, is given in Appendix A.

4.3. *Sky coverage as a function of frequency resolution*

The narrower the bandwidth of a signal, the longer the continuous observation needed to detect it with maximum sensitivity. Each DFT length ℓ has an associated duration $\Delta t(\ell)$ and frequency resolution $\Delta\nu(\ell)$. For some pixels P and DFT lengths ℓ, our longest observation of P is shorter than $\Delta t(\ell)$. Any DFT bin of length ℓ hence covers a band of sky that goes outside of P, limiting our sensitivity to signals narrower than $\Delta\nu(\ell)$. Thus, the effective sky coverage of SETI@home varies with signal bandwidth.

It is possible that a spike with DFT length ℓ is found during an observation shorter than $\Delta t(\ell)$. In that case, the spike's sky position is a stripe that is longer than a beam width. Such spikes are typically RFI.

DFT bins for a DFT length ℓ begin at regularly spaced times separated by $\Delta t(\ell)$. Multiplets comprise at least two time-disjoint detections (see §7.10). Thus, a pixel P can contain a multiplet composed of spikes of DFT length ℓ only if our observations of P include at least two intervals of duration $\Delta t(\ell)$ or more.

An observation of duration at least $3\Delta t(\ell)$ always contains at least two DFT intervals in their entirety; an observation of duration at least $2\Delta t(\ell)$ always contains at least one DFT interval and may contain two; an observation of duration at least $\Delta t(\ell)$ might contain only one.

DFT length	% of pixels strongly observed	% of pixels weakly observed
8	100	100
16	100	100
32	100	100
64	100	100
128	100	100
256	100	100
512	99.73	99.75
1024	97.42	98.98
2048	93.51	96.89
4096	78.29	91.49
8192	50.67	71.00
16384	42.56	48.44
32768	21.43	37.98
65536	2.41	11.61
131072	2.24	2.37

Table 3. Sky coverage as a function of DFT length

We say that a pixel P has been "strongly observed at DFT length ℓ" if its observation intervals always contain at least two DFT intervals. This is the case if P has an observation with duration at least $3\Delta t(\ell)$, or two observations with duration at least $2\Delta t(\ell)$.

We say that P has been "weakly observed at DFT length ℓ" if its observations may contain two DFT intervals. This is the case if P has an observation with duration at least $2\Delta t(\ell)$, or two observations with duration at least $\Delta t(\ell)$.

The fractions of pixels strongly and weakly observed at the various DFT lengths are shown in Table 3. For long DFT lengths (where SETI@home is the most sensitive; see §9.4), we have sufficient observations of only a small part of the sky.

5. CANDIDATE BIRDIES

Signal analysis systems can be tested by injecting artificial signals and verifying that these signals are reflected correctly in the output of the system. This technique is often called "signal injection and recovery." For example, we tested the SETI@home front end using hardware-based signal injection; this and other validations are described in Korpela et al. (2025).

Although hardware signal injection is useful for basic functional testing, it is difficult and expensive for hardware to generate complex and varied signals. The signal generation hardware used to test the front end can only inject simple signals, such as sinusoids and noise.

To test the SETI@home back end, we developed a sophisticated software signal injector. The artificial signals (*candidate birdies*) simulate persistent cosmic transmissions with fixed sky positions, and with a range of parameters: power, bandwidth, planetary motion, and possible correction for transmitter Doppler shift. For each birdie, we generate a set of *birdie detections*, mimicking as closely as possible

what the SETI@home front end would produce. We add these to the detection database at the start of the pipeline, before RFI removal.

The birdie mechanism serves several purposes, both in the development of the SETI@home back end and in its results.

RFI filtering: RFI filters can potentially remove ET signals; birdies let us estimate the extent of this. Birdie detections do not contain RFI, but some of them are inevitably removed by RFI filtering, for example, detections that lie in RFI frequency zones. The fraction of birdie detections that are removed indicates the likelihood that our RFI filters would remove a target signal. As we develop RFI filters, we monitor this fraction and keep it below about 10%.

Multiplet finding: Birdies help us develop effective algorithms for finding multiplets. If these algorithms fail to find multiplets for a birdie, or omit some of its detections, we can study these cases and improve the algorithms.

Score functions: We used birdies to develop our multiplet scoring functions by trying to find functions that rank birdie multiplets higher than other multiplets.

System sensitivity: By generating birdies with a range of parameters and seeing which of them are uncovered (i.e. produce highly ranked multiplets), we can estimate the sensitivity of the back-end system as a whole.

In this work, we created birdies that model continuous narrowband signals and generated only spikes for these birdies. The approach could be extended to model pulsed signals and generate detections of other types.

5.1. *Birdie parameters*

Each birdie B represents a signal with several parameters:

$(\alpha_{\mathrm{B}}, \delta_{\mathrm{B}})$**:** The signal's sky position (RA, dec).

$\nu(B)$**:** The center frequency of the signal.

$\Delta\nu(B)$**:** The bandwidth of the signal.

B_{is_bary}**:** whether the signal frequency is adjusted (by the sender) to cancel Doppler shift due to acceleration of the transmitter's reference frame.

$P(B)$**:** The power of the signal as it arrives at the receiver, in units of signal-to-noise ratio.

$F(B)$**:** The power of the signal as it arrives at the receiver, in W m^{-2}.

$F(B)$ is related to $P(B)$ by

$$F(B) = P(B)\Delta\nu(B)\frac{(T_{\mathrm{sys}} + T_{\mathrm{sky}}(\alpha_{\mathrm{B}}, \delta_{\mathrm{B}}))}{(\epsilon\Gamma_{\mathrm{B}}\langle M_{\mathrm{bin}}\rangle\langle M_{\mathrm{beam}}\rangle)} \tag{1}$$

where the terms include

T_{sys}: the Arecibo/ALFA system temperature (27 K).

$T_{sky}(\alpha_B, \delta_B)$: the galactic sky brightness at position (α_B, δ_B) expressed as a brightness temperature.

Γ_B: the geometric gain of Arecibo / ALFA (8 K Jy^{-1}).

ϵ: a non-geometric system loss term dominated by quantization losses (0.56, Korpela et al. 2025).

$\langle M_{bin} \rangle$: the median DFT bin response (0.81).

$\langle M_{beam} \rangle$: the median beam response (0.844).

A nonbarycentric birdie B has additional parameters:

$B_{orbital}$: The frequency, amplitude and phase of sinusoidal frequency variation due to the transmitter's orbital motion.

$B_{rotational}$: The parameters of the sinusoidal frequency variation due to the rotational motion of the transmitter.

During the development of the SETI@home back end, we generated sets of birdies of various sizes and parameter distributions. The set of birdies that we used to estimate sensitivity is described in §9.4.

5.2. *Generating birdie detections*

Given a birdie – a virtual ET signal with parameters as above – we must approximate the set of detections that would be produced by the SETI@home front end, were the signal to exist. This is a complex task, as we must simulate the changing Doppler shift of the signal, the movement and sensitivity of the telescope beams, the algorithms and parameters of the front-end signal analysis, the effects of noise, and so on. We do not model the effects of scintillation.

The algorithm for generating birdie detections is given in Appendix B

The number of detections generated for a birdie depends on its power and how closely beam trajectories approach its position. For some birdies, no detections are generated.

6. RFI REMOVAL

The Arecibo telescope detects anthropogenic radio frequency interference (RFI). Sources of RFI include side bands of TV and radio stations, aviation radar, cell phones, and other electronic devices. RFI can reach a receiver by various paths: reflection from ionized atmospheric layers, refraction around the structural components of the telescope, and direct detection by the receiver and associated electrical systems.

RFI often resembles target signals and is typically more powerful than cosmic sources. If we did not remove RFI, most of the top-scoring multiplets would be RFI; it would not be feasible to manually scan these looking for non-RFI multiplets. Thus, we need to remove as much RFI as possible before finding multiplets. At the same time, we need to minimize the possibility of removing target signals.

Many types of RFI are present at Arecibo, each with specific characteristics. There are a number of radars that produce strong pulsed signals. The SETI@home front end detects and removes these from the data before it is analyzed (see Korpela et al. 2025).

To remove other types of RFI, we developed several software filters, each designed to detect a particular type of RFI. Most of the filters are based on the fact that the telescope's sensitivity to RFI does not depend much on pointing: if RFI is present in a beam at a particular sky position P, it is likely to be present at positions several beamwidths from P, whereas this is not true for cosmic sources. Thus, if we find groups of detections that are similar but are separated in position, they may be RFI. We applied this principle at several times scales:

Long-term: the entire 14 years of observation. This gave us sensitivity to frequent RFI sources, even if they are faint.

Medium-term: roughly 10 minutes. This worked well for intermittent or unstable sources.

Short-term: the timescale of signals detected by the SETI@home client (1 ms to 13 s).

We developed filters for each of these timescales and for different types of detection. We also used single-detection filters that flag detections on the basis of their attributes, with no reference to other detections.

We ran the filters in parallel; that is, each was run against the entire set of detections, and the detections flagged by any of the filters were removed.

6.1. *Long-term RFI filters*

We developed separate filters for continuous detections (spikes and Gaussians), pulsed detections (pulses and triplets), and autocorrelations.

6.1.1. *Frequency-zone RFI*

This type of RFI involves terrestrial signals that occupy a limited, stable frequency band and are present during a significant fraction of the observation period. Typical sources are TV and radio transmitters and leakage from oscillators at the observatory. This RFI appears as spikes and Gaussians.

We developed an algorithm to identify this type of RFI. The algorithm divides the overall frequency band into small zones. For each zone change$Z$, we computed the fraction of time $E(Z)$ during which there is a statistical excess of detections in that zone. We then removed a fraction of the overall frequency band, consisting of the zones for which $E(Z)$ is greatest. The algorithm is given in Appendix C.

6.1.2. *Period zone RFI*

Periodic detections (pulses and triplets) have a type of RFI that is analogous to zone RFI, but the zones are in period rather than frequency. For example, an aviation radar near Arecibo produces RFI

with a ∼12 second period; the injected noise sources used for calibration typically have on/off cycles of .5, .2, and .1 sec, and 18 Hz is a common computer interrupt frequency. So there is a statistical excess of detections at these periods and their multiples by powers of 2.

The distribution of pulse and triplet periods is not uniform. They are concentrated at multiples or integer fractions of DFT durations. The frequency-zone algorithm assumes a uniform distribution across zones, so it must be modified to work for period zones. To do this, we compute the histogram of detection counts as a function of log period, smooth it with a 30-point median filter, and remove period zones whose counts exceed a probability threshold.

We use a similar algorithm for autocorrelations, for which the time parameter is the correlation delay.

6.2. *Medium-term filters*

6.2.1. *Drifting RFI*

We observed in spikes and Gaussians, narrowband RFI with frequencies that vary over time. An example of this type of RFI is shown in Figure 3. This RFI is not detected by the frequency-zone algorithm due to this frequency variation. We speculate that most of this *drifting RFI* is produced by consumer electronic devices whose oscillator frequencies vary with temperature.

To detect drifting RFI we adapted an algorithm that had been used in earlier SETI projects (Cobb et al. 2000). The basic idea is to construct for each detection D, two *fans* of triangles in time/frequency space originating at D, and extending forward and backward in time; see Figure 2. The slope of the triangles corresponds to the drift rate of the RFI. Within each triangle, we count detections whose sky positions are sufficiently far from D that they are not likely to have the same source. Then we look for triangles and pairs of opposed triangles, with a statistical excess of detections.

Detections often occur in clusters with adjacent or overlapping DFT bins; these may, for example, be one signal detected at different DFT lengths. Such clusters can inflate the count of detections in a triangle and erroneously trigger the algorithm. So we identify such clusters, select a *master detection* from each cluster, and use only these detections in the rest of the algorithm.

The algorithm is given in Appendix D

Gaussians near spikes flagged as drifting RFI typically are RFI as well, but often are not flagged as RFI because there are far fewer Gaussians than spikes. So we first compute drifting RFI for spikes, save the list of the time/frequency triangles with probabilities below threshold, and flag as RFI Gaussians lying in any of these triangles.

6.2.2. *Medium-term pulse and triplet RFI*

The pulsed detection types (pulses and triplets) also have RFI that occurs on timescales of the order of 10 minutes: groups of detections, similar in period, spread out in sky position.

We use the following filter to identify this type of RFI. It operates on windows of at most 10 minutes duration, identifies the detection D at the midpoint of this interval, and counts the detections that are far from D in sky position in both the positive and negative time directions. For typical observation

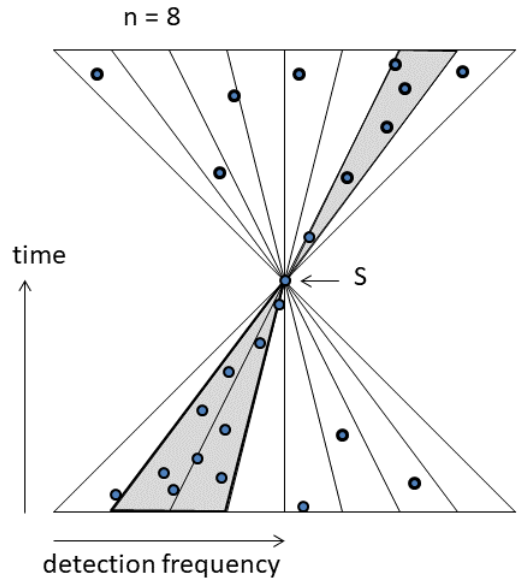

Figure 2. The drifting RFI algorithm constructs fans of triangles in time/frequency space, and looks for triangles containing an excess of detections.

modes, these two sets are not only far from D, but are also far from each other. If, in a given frequency bin, both sets have a statistical excess of detections, then they are flagged as RFI.

The algorithm is given in Appendix E.

6.3. *Short-term RFI filter*

If two detections are close in time and have similar properties (frequency or period, depending on type) but significantly different positions, it is likely that they are both RFI. This motivates the following filter. Typically, the two detections are from different beams, so we call it the *multi-beam filter*; however, the detections may be from the same beam if the telescope is slewing quickly.

The algorithm is given in Appendix F

6.4. *Signal property filters*

Two additional RFI filters consider the properties of each signal rather than the properties of a group of signals:

6.4.1. *Doppler drift rate consistency*

Because RFI is terrestrial in origin, it is detected at drift rates close to zero. Hence we flag as RFI all spikes and Gaussians D for which

$$\left|\frac{\Delta\nu_{\text{topo}}}{\Delta t}(D)\right| < 0.086 \text{ Hz s}^{-1} \tag{2}$$

This is done only for detections with $\ell(D)>32$ki. At DFT lengths of 32ki and less, the frequency bin width is such that the terrestrial drift rate is indistinguishable from zero.

6.4.2. *Duration consistency*

Each detection D has an associated duration $\tau(D)$, as described in §6.3. In the case of spikes, this is the duration of its DFT bin. During this time, the beam $b(D)$ is moving at an angular velocity V. Let $C(V)$ denote the corresponding beam crossing time, that is, the time it takes for the half-power beam to cross a sky point.

If $T < C(V)$, then in terms of sky position D is spread beyond a single half-power beam; in a sense it is too long for the observation in which it occurs.

For example, suppose that a 13.4 s spike D (the longest spike, hence the narrowest bandwidth) occurs at a moment when the beam-crossing time is 1 second. Then it is unlikely that the source of D is a cosmic signal; it is more likely RFI or noise. And if D is from a cosmic signal, that signal will probably be detected with greater power at a shorter DFT length.

We identify such spikes using an approximate algorithm based on pixel observation intervals. Given a spike D, let P be the pixel containing $pos(D)$, and let $b = b(D)$. Recall (§4.2) that $I(P, b)$ is the set of time intervals during which beam b observed pixel P. Let I_D denote the time interval of duration $\tau(D)$ centered at $t(D)$. If

$$|I(P, b) \cap I_D| < |I_D|/2 \tag{3}$$

(i.e. if at least half of I_D lies outside the observation intervals of its pixel) then flag D as RFI.

About 56% of spikes were flagged by this filter. The search for other detection types is already limited by observation duration, so no filter is needed.

6.5. *Developing RFI filters*

We do not know the characteristics of all types of RFI in advance, so the development of RFI filters, and the selection of their parameters, has been iterative and heuristic. The development cycle is as follows:

1. Run the current set of filters.
2. From the detections that remain, find the top-scoring multiplets. (see §7.10)
3. Examine these multiplets and the waterfall plots (see below) of their component detections, looking for RFI.
4. For RFI that should have been removed by an existing filter, modify the filter's algorithm or its parameters appropriately.
5. For new types of RFI, study their characteristics and add a new filter.
6. Examine birdie spikes that were flagged as RFI; in cases where they do not resemble RFI, modify the filters to not flag them.

To support this process, we developed a set of web-based tools. Using these tools, one can view lists of top-scoring multiplets of different types. For each multiplet, one can view a list of the component

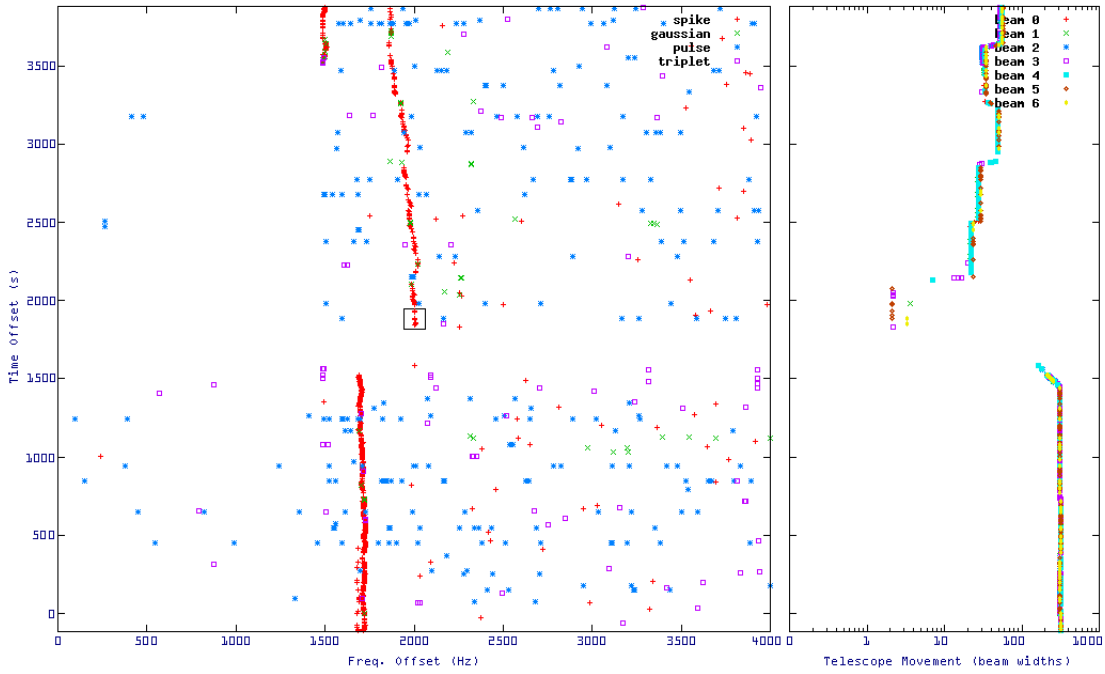

Figure 3. A waterfall plot showing drifting RFI. The left panel shows detections in time/frequency space, and the right panel shows the distance in sky position from the center detection to each of the seven beams. The black box in the left panel is centered on the center detection.

detections. For each detection, one can view a graphical *waterfall plot* showing the nearby detections in time/frequency space. RFI is generally visually apparent in these plots.

An example of a waterfall plot is shown in Figure 3. The left panel shows the detections, of all types. The right panel shows the angular distance (in beamwidths, on a log scale) of each of the 7 beams from the sky position of the detection, as a function of time. This is useful for identifying RFI, since groups of detections close in frequency and time, but at different sky positions, are generally RFI.

The waterfall plot web interface lets users move up or down, or zoom in or out, in either dimension. It lets them view the detections with or without RFI removal. It also lets them specify parameter values for the different RFI filters, and re-run the filters.

The interface also lets users bookmark detections of interest so that they can be reexamined later. This is useful for checking that algorithmic or parameter changes made for a particular case do not fail in other cases.

6.6. *Evaluating RFI removal*

The fractions of detections of each type flagged by each RFI filter, and in total, are shown in Table 4.

In evaluating the RFI removal system, there are two main criteria. First, it must remove enough RFI so that a significant fraction of top-ranking multiplets are not obvious RFI; otherwise, finding the non-RFI multiplets manually would take too much time. Our system has, in fact, done this; see §9.

Table 4. Fraction of signals flagged as RFI.

Signal type	Zone	Multibeam	Low drift	Drifting	In spike drifting	All
all	7.19%	4.46%	4.73%	5.47%	0.06%	10.83%
spike	13.31%	6.12%	8.44%	11.82%	0.00%	17.49%
Gaussian	1.69%	0.79%	0.00%	0.24%	2.38%	4.15%
pulse	2.33%	6.61%	0.00%	2.15%	0.00%	7.36%
triplet	0.17%	1.01%	0.00%	0.01%	0.00%	1.14%
autocorrelation	10.67%	0.00%	13.17%	0.00%	0.00%	14.49%
birdie spikes	8.78%	0.10%	0.56%	2.90%	0.00%	11.46%

We can also confirm this by looking at statistical measures. For example, the distribution of spike frequencies outside of a narrow band around the hydrogen line would be uniform in the absence of RFI. The zone algorithm (§6.1.1) is designed to enforce this and works as intended. Similarly, the distribution of spike power in noise should be a negative exponential, while RFI skews this distribution towards higher power. Indeed, the output of our RFI removal system has about the right distribution of power.

The second criterion is that the RFI removal system should not remove target signals. We used birdies (§5) – surrogates for target signals – to study this. RFI removal inevitably flags some birdie detections. For example, because birdie frequencies are random, some of their detections are in RFI frequency zones. However, as shown in Table 4, the fraction of birdie spikes flagged as RFI (11.36%) is significantly less than the fraction of flagged non-birdie spikes (17.49%), and as shown in §9.4, sufficient birdie spikes are not flagged that multiplets are found for most birdies.

7. TARGET SIGNAL CANDIDATES

After RFI removal, the remaining *clean* detections consist primarily of noise from the receiver and astrophysical sources, birdie spikes, and possibly target signals (see §2). The next analysis step is to look for sets of these detections that could plausibly be artifacts of a target signal. This identifies sets for which a) the detections' frequencies and drift rates are compatible with one of the target signal types, and b) the detections' sky positions are close enough for them to plausibly have the same source.

We call these sets *multiplets*. The SETI@home ALFA observations span 14 years, and most sky locations have been observed multiple times. The detections in a multiplet may come from many observations over a long period of time.

Multiplets are assigned scores designed to reflect the likelihood that they are the result of a target signal transmission rather than noise (see 7.9). The multiplet-finding algorithm searches for multiplets with high scores.

7.1. *Barycentric and nonbarycentric multiplets*

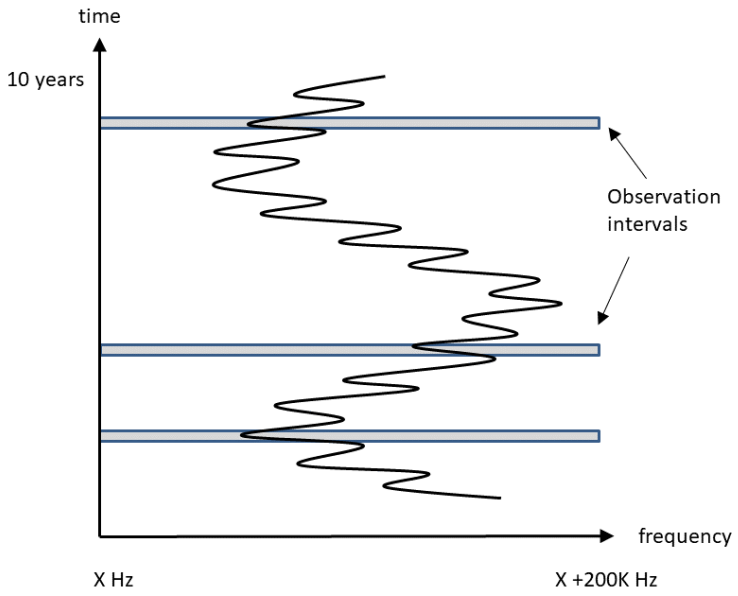

Figure 4. Nonbarycentric multiplets may include detections over a frequency range of up to 308 kHz.

Recall from §2 that target signals may or may not be corrected for transmitter Doppler shift; if a transmission is corrected in this way, we call it barycentric.

A barycentric signal will arrive at a nearly constant frequency. However, due to factors described in §3.4, the frequencies $\nu_{\text{bary}}(D)$ of the detections resulting from the signal can differ from this frequency by up to σ_ν. So, multiplets arising from barycentric signals will have detections for which $\nu_{\text{bary}}(D)$ varies by at most $M_{\Delta\nu}(bary) = 2\sigma_\nu = 250$ Hz. We call these *barycentric multiplets*.

The frequency of nonbarycentric signals is Doppler shifted due to the sender's orbital and rotational planetary motion. For the range of planetary parameters that we consider, the maximum range of this shift is about 308 kHz. The shift resulting from orbital motion varies slowly but covers a wide range; the shift resulting from rotational motion varies faster over a smaller range. See Figure 4.

Thus, the maximum frequency range of nonbarycentric multiplets is $M_{\Delta\nu}(nonbary) = 308$ kHz.

For the planetary parameters we consider, variations of this magnitude occur over long timescales (months or years). On short timescales (minutes), sender shifts are small; detection frequencies can vary within the uncertainty range of $2\sigma_\nu$, or 250 Hz.

The techniques for finding and scoring multiplets of these two classes are somewhat different. Thus, our algorithm for finding signal candidates has two variants: one that looks for barycentric multiplets in frequency windows of $M_{\Delta\nu}(bary)$, and one that looks for nonbarycentric multiplets in frequency windows of $M_{\Delta\nu}(nonbary)$.

7.2. *Multiplet categories*

A continuous narrowband signal could produce Gaussians when the telescope moves slowly across its source position, and spikes at other times. To maximize sensitivity, we form multiplets from the

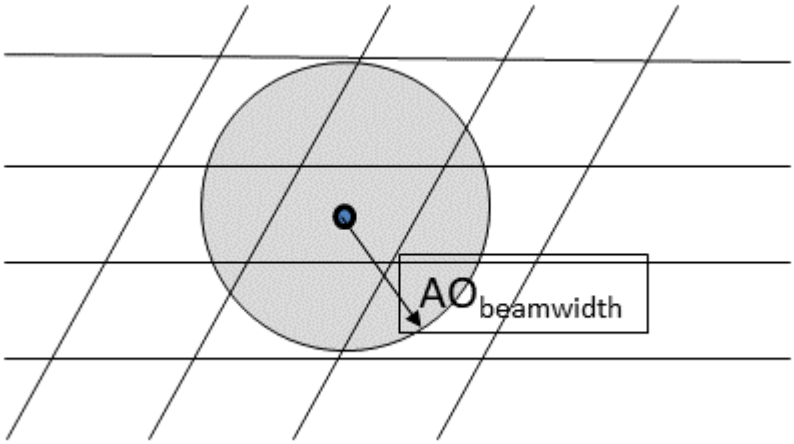

Figure 5. A pixel disk is centered at a pixel and has radius θ_{beam}.

combined set of spikes and Gaussians. Similarly, a pulsed transmission could produce both pulses and triplets, so we form multiplets from the combined set of pulses and triplets. We form autocorrelation multiplets separately.

Thus, multiplets can be from any of three detection sets (spike+Gaussian, pulse+triplet, and autocorrelation) and can be barycentric or nonbarycentric. Each of the six combinations is called a *multiplet category*. The categories differ in many respects; for example, their score distributions differ. We generate a separate list of top-scoring multiplets for each category.

7.3. *Sky position of multiplets*

A target signal is assumed to have a fixed sky position. However, the detections resulting from the signal will generally have different sky positions due to position uncertainty (see §3.4).

Our multiplet-finding algorithm makes two approximations involving sky position. First, when we form multiplets, we require that the sky positions of the detections lie in a disk of radius θ_{beam}. This ensures that all detections in the disk are within the uncertainty radius σ_{pos} of its center and therefore could be from the same source.

Second, we consider only disks centered at the center of HEALPix pixels. For a pixel P, disk(P) denotes the set of sky positions centered on P and with radius θ_{beam}. These disks overlap (see Figure 5). Note that the set of detections resulting from a target signal might not be contained in a single pixel disk, in which case we will not necessarily find the best multiplet for that signal.

7.4. *Constraints on multiplets*

To ensure that multiplets plausibly result from a target signal, we require that they satisfy a number of constraints.

7.4.1. *Overall frequency range*

As explained in §7.1, the detections D forming a multiplet must have values of $\nu_{\text{bary}}(D)$ that lie in a range of $M_{\Delta\nu}(bary)$ (250 Hz) for barycentric multiplets and $M_{\Delta\nu}(nonbary)$ (308 kHz) for non-barycentric multiplets.

7.4.2. *Drift rate constraints*

The terrestrial drift rate in the SETI@home band (1418.75–1421.25 MHz) at AO ranges from -0.12 to $-0.16\,\text{Hz s}^{-1}$. Most of the variation is due to the difference between the observation direction and the rotational acceleration vector. For barycentric multiplets, we constrained the reported detection drift rate to those consistent with this drift range, given the drift rate step used at each DFT length.

When determining whether a detection is consistent with having a stable barycentric frequency, we compare the reported drift rate with the range of allowed drift rates for a signal detected at that DFT bandwidth. For the 128ki DFT length, the allowed range is -0.19 Hz s^{-1} to -0.09 Hz s^{-1}. For DFT lengths shorter than 16ki, any nonzero drift rate disqualifies the detection.

7.4.3. *Period and delay consistency*

For pulse/triplet multiplets, we require that the periods of the detections be about the same, and similarly for the delays in autocorrelation multiplets. For pulse/triplet multiplets, two detections are considered *consistent* if their periods differ by at most twice the maximum of the two DFT durations.

For barycentric autocorrelation multiplets, two detections are considered consistent if their delays differ by at most $M_{\Delta p}(bary) = 0.01$ s.

For nonbarycentric multiplets (for which the delays may be Doppler-shifted), the delays must differ by at most $M_{\Delta p}(nonbary) = 0.1$ s.

7.4.4. *Local drift rate/frequency consistency for nonbarycentric multiplets*

Given our assumptions about target signals (§2), planetary motion on timescales of hours or less causes only small changes in drift rate. Therefore, if two detections are close in time and have very different drift rates, they are not from the same source. In addition, if a detection D has a drift rate of $1\,\text{Hz s}^{-1}$, we would expect a detection D' 10 seconds later to have a frequency approximately 10 Hz higher. The more it differs from this, the less likely that D and D' have the same source.

To express this, we impose a constraint with the following parameters:

$M_{\Delta t}(local)$**:** The time period over which we enforce consistency: 0.1 day. Over this duration, drift rate is fairly constant for our range of target signals.

$M_{\Delta c}(local)$**:** The maximum allowed variation in drift rate over periods of duration $M_{\Delta t}(local)$: $10\,\text{Hz s}^{-1}$.

$M_{\Delta\nu}(local)$**:** The maximum variation in drift-adjusted frequency during that period: $2\sigma_\nu$, or 250 Hz.

A drift rate change of $10\,\text{Hz s}^{-1}$ corresponds to a change in the line-of-sight acceleration of 2.2ms^{-2}. While this could potentially occur in a long observation of a short period satellite, because the vast majority of our observations are significantly less than one minute it is unlikely to occur in practice.

7.4.5. *Global drift rate/frequency consistency for nonbarycentric multiplets*

We require that the change in Doppler drift rate and frequency over longer time periods be consistent with planetary motion of the range we are considering. We approach this problem by asking: If (given our assumptions on planetary motion) a signal has drift-rate and frequency (C_1, F_1) and time t_1, is it possible that it has drift-rate/frequency (C_2, F_2) at a later time t_2? Based on an analysis of a large number of nonbarycentric birdies, we empirically found that the inequality

$$|C_1 - C_2|\ |F_1 - F_2| < 0.75{\times}10^7\sqrt{t_2 - t_1} \tag{4}$$

provides a fairly tight bound for our range of planetary and stellar parameters.

We use this as our global consistency constraint. This is necessary but not sufficient for the multiplet to match an actual orbital/rotational system. It does not ensure that the detection frequency and drift rate fit plausible sum-of-sinusoids functions that would be expected of target signals.

7.4.6. *Frequency-variation consistency for barycentric multiplets*

The detections comprising a barycentric multiplet should vary little in barycentric frequency, but their topocentric frequencies can (and should) vary because of receiver acceleration.

At one point, we noticed a number of high-scoring barycentric multiplets for which the opposite was true: There was more variation in barycentric frequency than in topocentric frequency. This is evidence that these groups were due to a frequency-stable terrestrial source. Often these multiplets resembled weak drifting RFI but did not have sufficient detections to trigger the drifting RFI algorithm.

To disallow such multiplets, we added the following constraint for spike/Gaussian barycentric multiplets. Such a multiplet may not contain a group of detections that are within a 0.1 day time interval (about the maximum amount of time that a celestial source could be observed continuously from Arecibo), and for which

$$B > 2\max(R, T) \tag{5}$$

where B is the RMS variation in barycentric frequency, T is the RMS variation in topocentric frequency, and R is the coarsest frequency resolution of the detections in the group.

7.4.7. *Time disjointness of detections*

Recall (§3.3) that each detection occupies a particular time interval. A cosmic signal may produce detections for which these intervals overlap: perhaps spikes in frequency-adjacent DFT bins or of different DFT lengths. These detections are redundant, and including all of them in a multiplet would inflate its score. So we require that a multiplet's detections not overlap in time; in other words, that the intervals determined by $t(D)$ and $\tau(D)$ not overlap.

7.4.8. *Frequency disjointness of multiplets*

Our initial implementation of multiplet finding produced a large number of multiplets differing by only one or two detections. This is undesirable; it fills the high-score lists with copies of essentially

the same multiplet. Ideally, there should be one multiplet for a given target signal candidate. So, we require that multiplets of a given category be disjoint. We enforce this with a stronger constraint: If M_1 and M_2 are multiplets of the same category, the ranges of the frequencies of their detections (that is, $\nu_{\text{bary}}(D)$) must be disjoint.

7.4.9. *Multiple detections*

Finally, all multiplets must include at least two detections. We originally allowed one-detection *singlets*, but these seemed to be entirely RFI. Only a small fraction of RFI detections leak through the RFI rejection algorithms, but these detections are highly selected when searching for the best singlets.

Because 97% of the Arecibo sky is covered two or more times in our commensal sky survey, and the median number of observations of a pixel is 13, our requirement that a signal be detected two or more times does not significantly lower the probability of detection of a high duty cycle signal.

7.5. *Finding nonbarycentric multiplets*

We now describe how we find multiplets: That is, given a pixel disk of detections of a given category (spike/Gaussian, pulse/triplet, or autocorrelation) how we find subsets of these detections that satisfy the above constraints. We do this in a way that tries to maximize the scores of the resulting multiplets. In practice, this means trying to find multiplets that have as many detections as possible and whose detections have the highest scores as possible (although these goals conflict in some cases).

We start with the following subproblem: Given the set S of detections in a given sky disk and in frequency band of $M_{\Delta\nu}(nonbary)$ (the maximum width of a nonbarycentric multiplet), what is the subset $T \subseteq S$ that

1. satisfies the consistency constraints described in §7.4
2. has the highest multiplet score subject to 1).

Finding the highest scoring subset is probably not feasible. Brute force search – examining all subsets of S – uses computing time exponential in the size of S, which can exceed one million. So we use a heuristic algorithm: starting with S, we prune detections in several stages, in a way that satisfies the constraints while maximizing the sum of detection scores. This increases the multiplet's *power factor*, one of three independent multiplet scores (see §7.9). A more sophisticated algorithm might try to increase the other factors as well.

7.5.1. *Local drift-rate/frequency pruning*

To enforce the local drift-rate/frequency constraint (see §7.4.4), we use an algorithm that, given a set of detections in a time window $M_{\Delta t}(local)$, returns a subset that satisfies the constraint. It tries to find a subset that will result in a high multiplet score. This algorithm is given in Appendix G.

The above algorithm produces a locally consistent group of detections in a short time window. We then need to assemble multiple such groups across the full time range in a way that is globally consistent (§7.4.5). An algorithm for this is given in Appendix H.

7.5.2. *Period and delay pruning*

For pulse/triplet and autocorrelation multiplets, we enforce the period/delay consistency constraint (see §7.4.3) by removing detections. As with drift-rate/frequency pruning, we want to keep as many detections as possible, especially high-scoring ones. We use a similar algorithm: we sort the detections by period or delay, then scan this list and find the interval over which a) the detections have consistent periods or delays and b) the sum of detection scores is greatest.

7.5.3. *Time overlap pruning*

To enforce the time-disjointness constraint (§7.4.7), we must remove overlapping detections. Specifically, given a set S of detections, some of which may overlap in time, we want to find $S' \subseteq S$ for which no detections overlap in time and which maximizes the sum of detection scores.

This is equivalent to the scheduling problem known as the Weighted Activity Selection Problem: given a set of *activities*, each with a value and start and end times, find the non-overlapping subset with the greatest total value. There is an efficient ($O(Nlog(N))$ algorithm to solve this problem, using dynamic programming (Cormen et al. 2022).

7.6. *Finding barycentric multiplets*

We have now described the algorithm for finding nonbarycentric multiplets in a given frequency band, from the detections in a pixel's disk. The analogous algorithm for barycentric multiplets is somewhat simpler and is given in Appendix I.

7.7. *Finding multiplets in a detection disk*

We have described the algorithms for finding barycentric and nonbarycentric multiplets in a particular frequency band; Appendix J describes the algorithm for finding multiplets across all frequencies.

The algorithm for finding nonbarycentric multiplets in a detection disk is identical, except that the frequency band is $M_{\Delta\nu}(nonbary)$, the drift rate constraint for nonbarycentric multiplets is used (§7.4.2), and the algorithm described in §7.5 is used.

7.8. *Pruning overlapping multiplets across pixels*

For a given pixel, the algorithm in the previous section generates multiplets that are disjoint in frequency range and therefore disjoint in terms of detections (see §7.4.8). However, it is possible that multiplets generated for nearby pixels are not disjoint. This can fill the top-ranked lists with several copies of essentially the same multiplet.

A final stage in multiplet-finding identifies pairs of multiplets in nearby pixels that are of the same category and have one or more detections in common. It removes the lower-scoring multiplet from each such pair.

7.9. *Scoring multiplets*

The algorithms described in the previous subsections produce tens of millions of multiplets. We extracted about one thousand of these for manual examination and possible reobservation. For this purpose, we developed three functions, or score factors, that take a multiplet M and return a score – a measure of a property that we would expect to find in ET signals and not in noise. We used these scores to decide which multiplets to examine.

Each factor estimates a probability that M would be found in noise, assuming Poisson statistics. These probabilities can be very small, so we do the computation in log space to avoid loss of numerical precision. We then negate the log values so that larger values are better.

7.9.1. *Power factor*

Recall from §3.3 that a detection has a probability score $S(D)$. The first multiplet score factor, $S_{\text{prob}}(M)$, rewards (i.e. gives high scores to) multiplets that have detections with high scores, for example because they have high power.

$S_{\text{prob}}(M)$ is essentially the median score of the detections in M. However, the distribution of $S(D)$ varies significantly between the detection types, because it is defined differently and the number of detections varies between types. Millions of spikes have scores higher than the highest Gaussian score.

To ensure that $S_{\text{prob}}(M)$ selects multiplets containing Gaussians, we normalize scores across detection types. For each detection type, we find the score of the 30 millionth highest detection and subtract this from the scores of detections of that type. $S_{\text{prob}}(M)$ is then defined as the median over the detections in M of these normalized scores. This has the desired effect: in the high-ranking multiplets of the mixed types (spike/Gaussian and pulse/triplet) both types are well represented.

7.9.2. *Density factor*

Suppose that for a pixel P, disk(P) contains N detections of a particular type. If A is a section of sky contained within disk(P), with area Xarea(P) for some $X < 1$, then the expected number of detections positioned in A is NX. Similarly, if a frequency band F is contained in SETI@home's 2.5 MHz band, and its width is $Y \times 2.5$ MHz for some $Y < 1$, the expected number of detections in $disk(P)$ with frequencies in F is NY, and the expected number of detections with positions in A and frequencies in F is NXY.

We would expect an ET signal to produce a multiplet M with many detections closely spaced in both position and frequency. Suppose a multiplet M is in a pixel disk, disk(P), with N detections, and the positions of M's detections cover (in the sense defined below) a fraction $X(M)$ of the area of disk(P), and the frequencies of M's detections cover a fraction $Y(M)$ of SETI@home's 2.5 MHz range.

The expected number of detections in these ranges of positions and time is then

$$E = NX(M)Y(M) \tag{6}$$

The probability of finding at least $|M|$ detections in these ranges is then

$$P(M) = \Gamma(|M|, E) \tag{7}$$

where Γ is the incomplete Gamma function.

We define the density factor as

$$S_{density}(M) = -\log(P(M)) \tag{8}$$

The density factor rewards multiplets with more detections than would be expected given their range of sky position and frequency. Thus, it gives high scores to multiplets that have a large number of detections and are compact in both position and frequency.

We now define the fractional coverage factors $X(M)$ and $Y(M)$. Let A be the mean sky position of the detections in M, let B be the standard deviation of the angle between A and the positions of detections in M. X(M) is then the area of a disk of radius B divided by the area of the sky disk disk(P).

To define $Y(M)$, we first define the *frequency deviation*, D_{dev}, of a detection D in a multiplet as the difference between $\nu_{\text{bary}}(D)$ and the *center frequency* of the multiplet at time $t(D)$. For barycentric multiplets the center frequency is constant; it is defined as the mean of the $\nu_{\text{bary}}(D)$ over all the detections in the multiplet. For nonbarycentric multiplets, the center frequency varies linearly over time within an observation interval; it is the center of the time/frequency bands computed during drift-rate/frequency pruning. (See §7.5.1) Note that in all cases the frequency deviation is at most $2\sigma_\nu$. $Y(M)$ is then the standard deviation of D_{dev} divided by 2.5 MHz.

7.9.3. *Time factor*

A multiplet M consists of detections within the disk centered at a pixel P. We know $I(P)$, the set of time intervals during which we observed P (see §4.2). If M results from a beacon that is always on, it could contain detections from throughout these intervals. The *time factor* $S_{time}(M)$ attempts to quantify the extent to which it does. We do this by estimating the probability that a random set of detections, with constraints similar to those of multiplets, would cover at least as much time as the detections in M.

We introduce some terms to formalize this We call a set of non-overlapping time intervals an *interval set*. If I is an interval set, $D(I)$ denotes its duration, that is, the sum of the lengths of its component intervals.

Recall that detections have associated time intervals, centered at $t(D)$ and of duration $\tau(D)$. The detections in a multiplet M have non-overlapping intervals. Let $I(M)$ denote the corresponding interval set and $D(M)$ its duration.

If P is a pixel, let $I(P)$ denote the interval set consisting of its observations and $D(P)$ the corresponding duration. Typically, if M is a multiplet from pixel P, then $I(M) \subseteq I(P)$ and therefore $D(M) < D(P)$.

The detections in a multiplet lie in a frequency band of width $2\sigma_\nu$ (250 Hz) (for barycentric multiplets, this band is fixed; for nonbarycentric multiplets, it changes over time). For a multiplet of type T (spike/Gaussian, pulse/triplet, or autocorrelation), let $F(T)$ denote the average fraction of time that a 250 Hz band contains detections of type T. We can estimate this as follows: enumerate the detections of type T in time order. We group them into contiguous sets in which the gap between

detections does not exceed 30 seconds. For each such time interval, divide the detections into 250-Hz bins. Within each bin compute the amount of the interval that is covered by the detections. $F(T)$ is then the average coverage over all bins and time intervals.

We assume that for a given interval set I and 250-Hz band B, the coverage of I by detections in B has a roughly Poisson distribution with mean $F(T)D(I)$.

We can now define the time factor for a multiplet M:

$$X = \Gamma(D(M), F(T)D(P)) \tag{9}$$

$$S_{time}(M) = -log(X) \tag{10}$$

where P is the pixel that contains M. X is the probability that the coverage of $I(P)$ by a random set of detections of type T, within a 250 Hz window, would exceed the coverage of M.

7.9.4. *Evaluating and combining score factors*

We have defined three *score factors* designed to measure the multiplet properties that we expect ET signals to have. Two questions arise:

- Do the factors, as we intend, rank ET signals higher than noise?
- How do we combine the factors into an aggregate score?

For the spike/Gaussian categories, the birdie mechanism gives us a way to answer the first question: we can see whether birdie multiplets score higher than non-birdie multiplets. In fact, this idea was used to develop and refine the score factors. Given two alternative functions, we chose the one that more favors birdie multiplets.

We originally combined the three scoring factors into a single score. Since each factor is conceptually a probability, it was natural to multiply them to form a single score. However, some factors have a much wider range than others and they dominated the product. So we experimented with using different weights for the factors in an effort to boost the scores of birdie multiplets relative to non-birdie multiplets. We tried using both neural networks and optimization. Neither approach was successful; when we trained with a set of birdies, the resulting weights worked well for those birdies but poorly for others.

Currently, we use a simpler approach. First, we linearly scale each factor so that the 25th percentile and 75th percentile are -1 and 1 respectively. The score factors have different distributions for different categories, so we do this scaling separately for each category.

Second, we consider sums of just one or two of the score factors, as well as the sum of all three. There are seven such combinations; we call them *score variants*. An ET signal could score high in one or two factors but not in the others. For example, a weak but persistent signal could score high in time and density factors but low in power factor. A strong but brief signal would do the opposite.

To select multiplets to reobserve, we manually examined the top multiplets from each score variant.

table	each row represents	number of rows
tape	Several hours of data from a beam	117,466
workunit group	107 seconds of data from a tape	7,691,384
workunit	One of 256 frequency subbands of a workunit group	1,968,994,304
detection	A detection (spike, Gaussian, etc.)	12,107,039,965

Table 5. Database tables and row count.

7.10. *Evaluating the multiplet-finding and scoring algorithms*

The algorithms for finding and ranking multiplets described in this section went through many stages of development and refinement. For continuous narrowband signals, this process was guided by birdies. We evolved the multiplet-finding algorithms to find most of the non-overlapping detections in a birdie, and few extraneous detections; we evolved the scoring functions to rank birdie multiplets high compared to real multiplets. This process resulted in algorithms that perform these functions well. As will be shown in §9.4, they successfully uncover most birdies.

For other detection types, it is hard to quantify the performance of the algorithms because we have no birdies of those types. In general, we have proceeded empirically. We manually examined the top-ranking multiplets. For each multiplet, we examined the detections in its range of frequency and/or period, and made sure that we were including those we should. We examined its score factors and determined whether its high scores were merited. By these subjective criteria, our algorithms work well.

8. IMPLEMENTATION AND PERFORMANCE

The algorithms comprising the SETI@home back end have been developed empirically in many iterations. Each iteration required running some or all of the pipeline. An early version of the back-end pipeline, called NTPCkr, (Korpela et al. 2011; Korpela et al. 2011) took years to run, making algorithm development infeasible. Starting in 2016 we redesigned the back end, with the goal of reducing the time for a complete run to about a day.

8.1. *Data storage and I/O*

Much of the back-end processing is data intensive. The SETI@home front end stores its output in a relational database. Information is stored in a hierarchy of tables (see Table 5). A row in each table links to a parent row in the next higher table. Retrieving data from this database is slow. Traversing links up the table hierarchy compounds the problem.

We speed up data access as follows. First, we dump the relational database to comma-separated value (CSV) files; the database is not used further. We then flatten the table hierarchy. For example, given a detection, we need the angle range of its workunit group. For this purpose, we create an array, indexed by workunit group ID, of the angle range of that workunit group. We store this array in a file, accessed using memory mapping. Given a detection, we can find the angle range by using the detection's workunit group ID as an index into this array. This involves a single memory reference rather than a chain of database queries.

Multiplet category	#multiplets	avg. #detections	avg time span (sec)
spike/gaussian, barycentric	486092	3.51 spikes + 0.07 Gaussians	110.44
spike/gaussian, nonbarycentric	2260066	2.56 spikes + 0.20 Gaussians	1,152.81
pulse/triplet, barycentric	16668902	0.11 pulses + 2.00 triplets	666.19
pulse/triplet, nonbarycentric	134247	0.00 pulses + 2.09 triplets	890.29
autocorr, barycentric	287	2.13 autocorrs	53.83
autocorr, nonbarycentric	880346	2.08 autocorrs	1,103.20

Table 6. Multiplet statistics.

We need to access detections in time order during RFI removal and by pixel number during multiplet finding. We do this in each case by sorting the detection files on that parameter (using the Unix "sort" utility), then building an index that allows random access based on the parameter. The RFI detection program makes a single pass through the time-ordered list of detections of a given type, minimizing disk I/O. Multiplet finding makes a single pass through frequency-ordered detections.

8.2. *Efficient data structures and algorithms*

Many of the algorithms involve processing N items where N is on the order of a million. To avoid $O(N^2)$ runtime, the back-end programs use a data structure called R-trees (Guttman 1984). An R-tree stores a set of geometries such as polygons or points, and allows the set to be queried (e.g. to see if a given point is in any of a set of N rectangles) in $O(log(N))$ time.

For example, while generating birdie detections (§5.2), we store the set of birdie sky positions in an R-tree. Given a telescope pointing, we can then efficiently identify the birdies that are close to it. In RFI removal (§6) we use R-trees to store drift triangles and detection uncertainty rectangles.

8.3. *Parallelism*

RFI removal is done by a multithreaded program; on a machine with N CPUs, each CPU processes a time range containing about $1/N$ of the detections. With 10 billion detections and 56 CPUs, RFI removal takes about 15 hours.

Multiplet finding and scoring are performed by a computing cluster with several thousand nodes, using HTCondor (Thain et al. 2005). The task is divided into jobs of 64 pixels each; there are about 250,000 jobs. Each job takes an average of 1120 seconds, so with 2000 CPUs the task takes about 1.6 days.

9. RESULTS

Once the algorithms were finalized, we performed two complete runs of the SETI@home back end. For the first run, the input was the detections generated by the SETI@home front end. The result of this run is a set M_R of *real multiplets*. Table 6 shows some statistics of these multiplets.

We manually evaluated the top-ranking multiplets, skipping those consisting of obvious RFI; see §9.5. The result of this is a set of sky positions and frequency ranges to be reobserved.

Next, we generated a number of birdies that span our range of continuous target signals and generated detections as described in §5.2. We combined these detections with the real detections and processed this with the back-end pipeline. We used the results to estimate the sensitivity of SETI@home to signals with various characteristics. To do this, we checked which birdies were uncovered: i.e., which one produced a multiplet that ranked in the top 100 non-birdie multiplets for some score variant. For each type of birdie, there was a power level above which most birdies were uncovered. This minimum power level, or *candidate sensitivity*, depends on the intrinsic bandwidth of the birdie and whether it is barycentric.

9.1. *Birdie generation*

For the second run, we created 3000 birdies with parameters chosen as follows:

- Half of the birdies (1500) are barycentric, half are nonbarycentric.
- $\Delta\nu(B)$ is chosen randomly in a way that distributes the values evenly (in log space) over SETI@home's range of frequency resolution. For each of the 15 DFT lengths ℓ, there are 100 barycentric and 100 nonbarycentric birdies with bandwidths in the range of resolutions closest to $\Delta\nu(\ell)$.
- For a birdie with $\Delta\nu(B)$ closest to DFT length ℓ, the position $(\alpha_\mathrm{B}, \delta_\mathrm{B})$ was chosen randomly from pixels that were strongly observed at that DFT length (see §4.2). In other words, we placed the birdies only in areas of the sky where we had long enough observations to detect them with maximum sensitivity.
- For nonbarycentric birdies, the planetary motion parameters were chosen randomly from ranges consistent with habitable-zone planets orbiting type F and G stars. We assume that the signal is always on, so that if the transmitter is on the surface of a planet, there is a second transmitter at the antipode.
- $P(B)$ is between 18 and 33 for barycentric birdies, and 18 and 50 for nonbarycentric birdies. Within the 100 birdies in each bandwidth range, the powers are regularly spaced within these limits.
- $\nu(B)$ is chosen randomly from the 2.5 MHz band with a band of $M_{\Delta\nu}(bary)$ Hz (barycentric) or $M_{\Delta\nu}(nonbary)$ Hz (nonbarycentric) removed from both ends; this ensures that all detections generated for the birdie lie in the 2.5 MHz band.

We generated spikes for the birdies as described in §5. This work predates the publication of work by Li et al. (2022), but is generally consistent. Because Li et al. concentrated on exoplanets around less massive stars, the orbital periods of the planets were shorter and the rotational periods typically longer due to tidal resonance. The equivalent drift rates for non-barycentric signals from Li et al. (2022) are well within the SETI@home range, and would not be excluded from being detected. However, the high orbital velocities expand the frequency range of the signals, making detection of multiplets more difficult within a limited number of observations. In essence, the number of non-birdie detections included in a multiplet is proportional to the orbital velocity, whereas the number of

Birdie BW	All	power+time	density+time	time	density+power	power	density
			Barycentric birdies				
All	11.27%	4.29%	13.46%	6.34%	22.54%	0.05%	42.06%
< 0.11 Hz	19.63%	15.89%	22.90%	19.16%	7.48%	0.00%	14.95%
> 863.17 Hz	7.59%	0.00%	7.59%	0.00%	34.18%	0.00%	50.63%
			nonbarycentric birdies				
All	6.02%	3.04%	8.76%	4.89%	7.81%	0.77%	68.71%
< 0.11 Hz	15.53%	12.42%	22.36%	15.53%	6.83%	4.35%	22.98%
> 863.17 Hz	0.00%	0.00%	0.00%	0.00%	5.10%	0.00%	94.90%

Table 7. Percent of birdies ranked highest by each scoring variant.

real signals would be constant. Hence we limited our birdie generation to orbits around more sunlike stars and acknowledge that uncorrected transmissions from rapidly accelerating systems is difficult. These spikes were combined with real spikes prior to RFI removal. We ran the multiplet-finding algorithm for pixels containing at least one birdie spike and retained only multiplets having at least one birdie spike. For each birdie B, $M(B)$ denotes the set of multiplets in M_B that contain at least one spike from B; these are collectively called *birdie multiplets*.

Each multiplet is scored using the three score factors described in §7.9, and these are combined into seven score variants. For each multiplet M and score variant V, $S(M, V)$ denotes the score of M in variant V.

For each birdie multiplet M and each score variant V, $rank(M, V)$ denotes the rank of $S(M, V)$ relative to the scores of the real (non-birdie) multiplets. For example, $rank(M, V)$ is zero if its V score is greater than that of all real multiplets.

9.2. *Effectiveness of multiplet score factors and variants*

In §7.9, we defined three multiplet score factors: S_{power}, $S_{density}$, and S_{time}. These were designed to measure properties of multiplets that are present in target signals more than in noise. How well do they do this, and which factors or combination of factors work best?

We studied this for continuous signals by examining the ranks of birdie multiplets and seeing which score variant produced the best ranks. Specifically: for a birdie B and score variant V, let $rank(B, V)$ denote the minimum of $rank(M, V)$ over multiplets M containing detections from B. For a score variant V, let $F(V)$ be the fraction of birdies for which $rank(B, V)$ is minimal over score variants; i.e. the birdies for which V is the best selector.

Table 7 shows values of $F(V)$, broken down by birdie type (barycentric or not) and by the intrinsic bandwidth of the birdie (only the smallest and largest ranges are shown). From this data we can conclude that a) the best score variant depends on birdie type and bandwidth; b) each of the score variants is best for a significant fraction of birdies. Hence, in selecting multiplets for manual evaluation, we examined the top-scoring multiplets in all seven score variants.

9.3. *Defining event and candidate sensitivity*

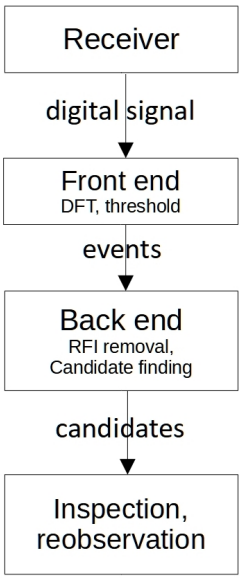

Figure 6. The general structure of radio SETI projects.

Radio SETI projects typically consist of a front end that finds *detections*, and a back end that finds a relatively small set of *candidates*, which are inspected manually and possibly reobserved. In the case of SETI@home, candidates are high-scoring multiplets.

The sensitivity of the system is, generally speaking, the minimum signal strength that it can detect. But this must be defined carefully, and it is important to distinguish between events and candidates.

Suppose a signal, with a flux F and other parameters such as bandwidth $d\nu$, arrives at the receiver at a particular time. The signal may be detected, producing an *event*. Whether it is detected is probabilistic because it depends on:

- noise from space (which varies with sky position);
- noise in the receiver electronics;
- RFI (which varies with time and frequency);
- where the telescope was pointing relative to the signal source;
- whether the telescope pointing was changing, and how fast;
- sender and receiver acceleration, and corresponding Doppler drift.

We first define the system's *event sensitivity*, S_{event}. To do this, we pick a probability threshold P (say, 0.5). S_{event} is then defined as the least value of F for which the probability of detecting the signal is at least P. (We assume that the detection probability is a monotonic function of F: the stronger the signal, the more likely it is to be detected).

S_{event} depends on the signal bandwidth $d\nu$, so it should be notated as $S_{event}(d\nu)$. It also depends on the various factors listed above; we can average these factors over the project's observations to reduce it to a single number.

It is difficult to estimate a project's event sensitivity. The sensitivity reported by most projects is a best-case value that ignores most of the above factors.

But in any case, what matters for the purpose of detecting ET signals is not event sensitivity, but rather what we call *candidate sensitivity*. Again, we pick a probability threshold P. The candidate sensitivity $S_{candidate}$ is the least flux F for which a signal with flux F results in a candidate with probability at least P.

The $S_{candidate}$ and S_{event} are related, but they generally differ. $S_{candidate}$ could be larger than S_{event} for various reasons:

- In a noisy RFI environment faint signals might be detected by the front end, but almost all of them would be discarded as RFI and hence would not produce candidates.
- a faint signal might produce events, but not enough to satisfy the requirements of the candidate detection algorithm.
- RFI removal algorithms might be too aggressive or the candidate detection algorithm might fail to find possible candidates.

Conversely, $S_{candidate}$ could be less than S_{event}. Assume that the signals are persistent (always on) and that the project observes sky positions repeatedly or for long periods. If a persistent signal is weaker than S_{event}, it might be detected with below-threshold probability, say 0.4. Hence, the signal could result in multiple events which could be identified as a candidate by appropriate algorithms.[1]

Using the birdie mechanism, we can estimate the candidate sensitivity of SETI@home.

9.4. *Candidate sensitivity for continuous signals*

For some birdies – those with low power or whose positions were not sufficiently close to any beam trajectory – no spikes were generated. Other birdies had spikes but no multiplets were produced containing any of these spikes. Of the birdies for which multiplets were produced, we determined whether they were uncovered in the above sense.

We graphed, as a function of birdie power, the fraction of birdies falling into these three classes: those with spikes, those with multiplets, and those with multiplets that were uncovered. We did this separately for each of the 15 signal frequency ranges, and for both barycentric and nonbarycentric signals.

In all cases, the fraction of birdies uncovered increased with birdie power, and above some power at least 80% of birdies were uncovered. This power is an estimate of our candidate sensitivity to signals of that type. Examples of this for barycentric birdies are shown in Figures 7 to 9. We show the graphs for the 1st, 8th, and 15th bandwidth ranges; the others are similar. Analogous graphs for nonbarycentric birdies are shown in Figures 10 to 12.

[1] It is important for SETI researchers to acknowledge this distinction. A project claiming a detection sensitivity of, for example, 10^{-23} W m^{-2} but only publishes and reobserves candidates with power $>10^{-20}$ W m^{-2} has a true sensitivity of $\sim 10^{-20}$ W m^{-2}.

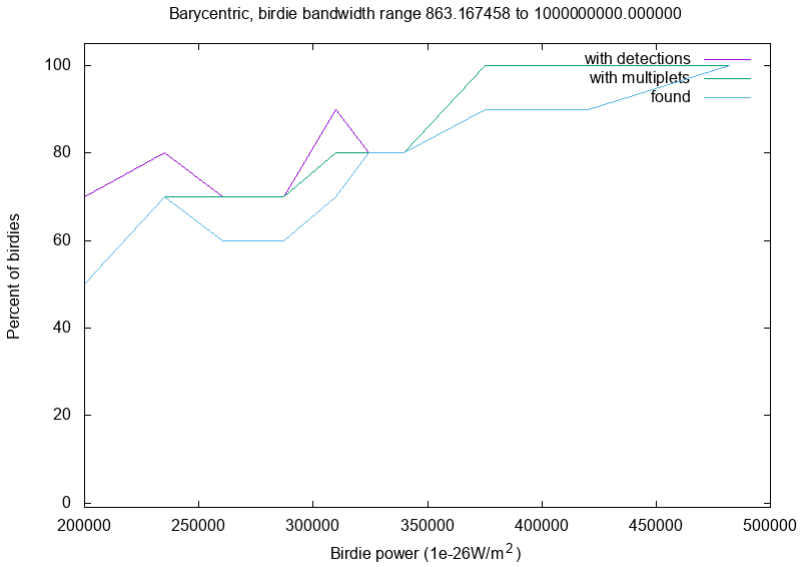

Figure 7. Candidate sensitivity to barycentric signals with bandwidth $\geq$ 863 Hz

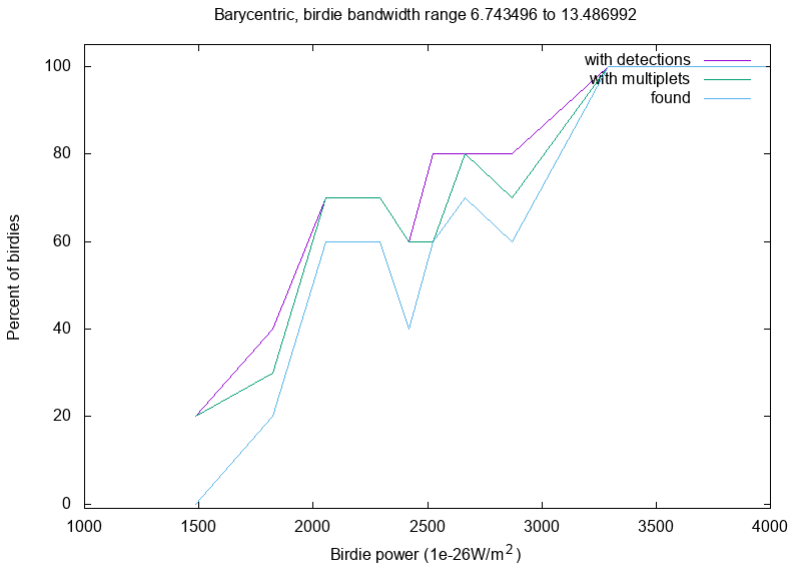

Figure 8. Candidate sensitivity to barycentric signals with medium bandwidth (6 Hz to 13 Hz)

Note that although birdies have the same range of SNR power, the range of powers in terms of flux varies widely with signal bandwidth. This reflects the fact that our DFT frequency resolutions are as small as 0.07 Hz, and we are much more sensitive to signals with bandwidths in this range.

These results are summarized in Table 8. For each range of signal bandwidth we show SETI@home's candidate sensitivity to barycentric and nonbarycentric signals in that range and the fraction of the sky in which we achieve this sensitivity. As can be seen, SETI@home is most sensitive to signals of bandwidths less than 0.1 Hz, but achieves this sensitivity in only a small fraction of the sky ($\sim$2%).

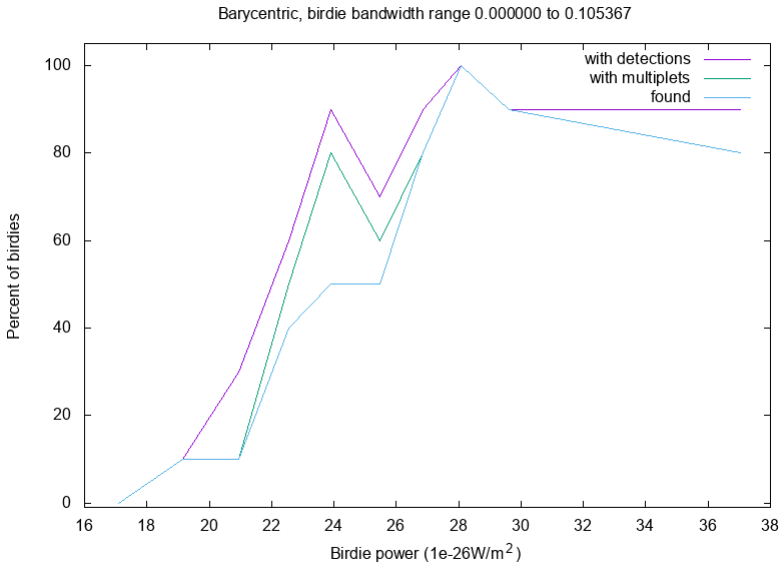

Figure 9. Candidate sensitivity to barycentric signals with bandwidth ≤ 0.1 Hz

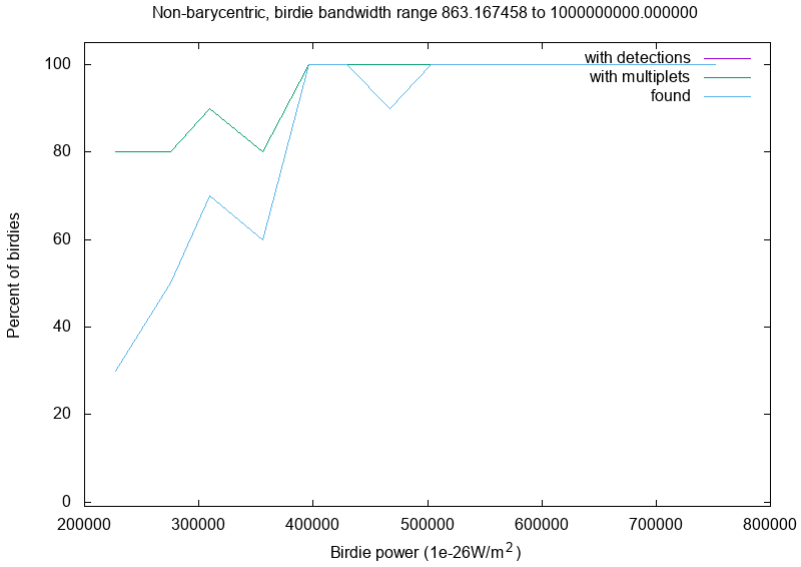

Figure 10. Candidate sensitivity to nonbarycentric signals with bandwidth ≥ 863 Hz

This is because the telescope was moving too quickly to detect such signals during most of our observations.

We conjectured that the total observing time in a birdie's pixel might be correlated with the probability of finding the birdie, but this turned out not to be the case. The fraction of birdies that were uncovered did not consistently exceed 80% as this value increased.

9.5. *Selecting signal candidates for reobservation*

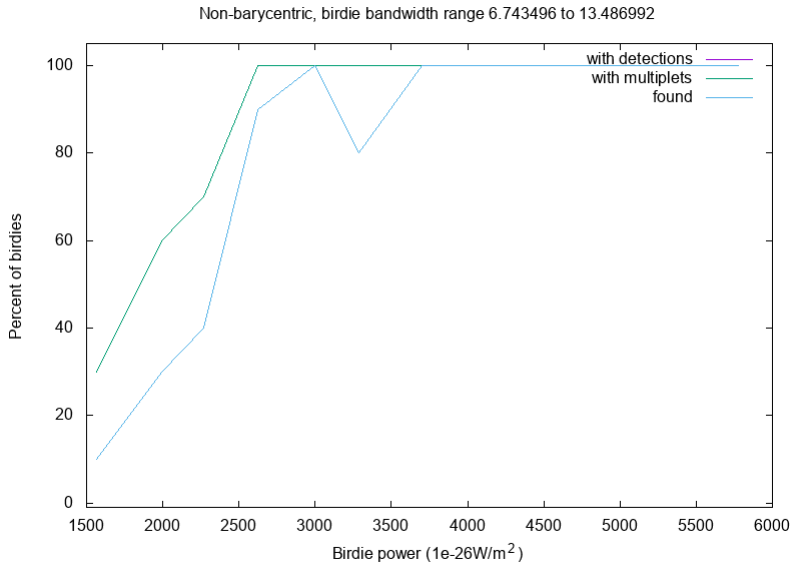

Figure 11. Candidate sensitivity to nonbarycentric signals with bandwidth from 6 Hz to 13 Hz

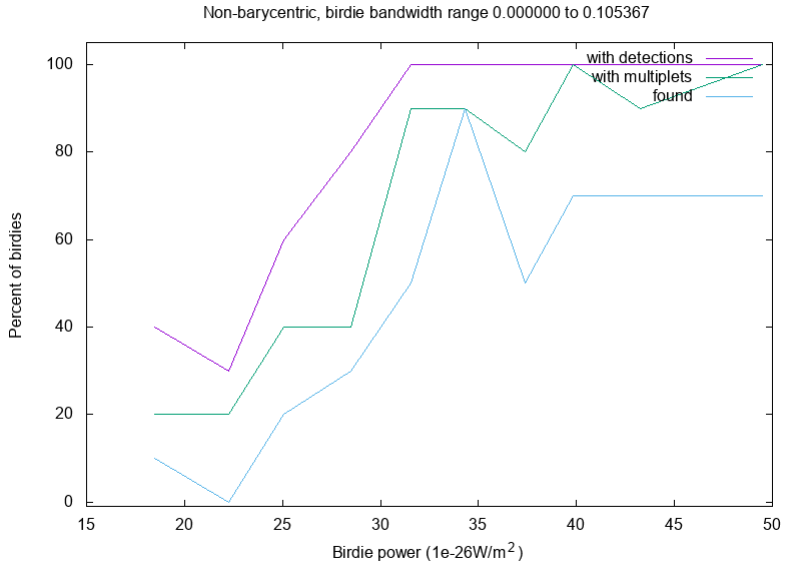

Figure 12. Candidate sensitivity to nonbarycentric signals with bandwidth < 0.1 Hz

The 300 meter Arecibo telescope collapsed in December 2020. Only the Five-hundred-meter Aperture Spherical Telescope (FAST) in China has sufficient sensitivity to check SETI@home candidates. Together with our Chinese colleagues, we were granted 23 hours of dedicated observing time on FAST to reobserve SETI@home candidates. Given the telescope time per reobservation (see §11.1) this allows us to reobserve 92 candidates.

We selected candidates for reobservation from among the top-ranked multiplets in the various categories and score variants. SETI@home is optimized for finding continuous narrowband signals. In deciding how to budget our 92 reobservations among multiplet categories, we allocated more to the

Signal Bandwidth (Hz)	Candidate Sensitivity (bary) (10^{-26} W m^{-2})	Candidate Sensitivity (nonbary) (10^{-26} W m^{-2})	Sky Coverage (%)
863–1726	320000	400000	100
431–863	150000	150000	100
215–431	90000	90000	100
107–215	50000	45000	100
53.9–107	21000	25000	100
26.9–53.9	12000	12000	100
13.4–26.9	5100	5100	99.73
6.74–13.4	3300	2600	97.42
3.37–6.74	1800	1500	93.51
1.68–3.37	800	900	78.29
0.842–1.68	350	650	50.67
0.421–0.842	210	260	42.56
0.210–0.421	120	160	21.43
0.105–0.210	52	60	2.41
0.052–0.105	28	40	2.24

Table 8. Summary of estimated candidate sensitivity and sky coverage

spike/Gaussian categories. In total, we selected 70 spike/Gaussian candidates (50 barycentric and 20 nonbarycentric), 12 pulse/triplet candidates, and 10 autocorrelation candidates.

Our algorithms for finding and ranking multiplets are not perfect; manual evaluation is still needed. For this purpose, we use the visualization tools described in §6.5. These let us examine waterfall plots of the multiplet's detections and see how detection times are distributed over the observation intervals for the pixel. A significant fraction of top-ranking multiplets were identified as consisting of RFI and/or noise.

For each multiplet category and scoring variant, the back-end software creates a list of multiplets ordered by descending score. Each multiplet has a summary page showing the factors that went into the multiplet score, information on each detection in the multiplet, and rankings and comments by experts inspecting this multiplet and its detections.

Figure 13 shows an example of a multiplet summary page. This particular multiplet is a nonbarycentric pulse/triplet candidate that is highly ranked manually and by algorithms because: it consists of detections with high SNR from seven different observations spanning 2.3 years at the same position on the sky (within 1 beamwidth). These detection periods are similar (0.837 Hz), the detections have a low negative drift rate (0.98 Hz/sec), and their frequencies are within the span expected from a nonbarycentic candidate (50 KHz RMS). This multiplet was selected for reobservation at the FAST telescope.

Figure 14 is a time vs. frequency waterfall plot from a multiplet that was rejected manually because its detections are likely RFI not detected by our RFI algorithms. The left panel of the figure shows

SETI HOME Project Science Computing Community Site Dan Log out

SETI@home multiplet: pixel 4922510, pulse/triplet, nonbarycentric, #3

Position	mean (8.33039 RA, 22.2988 dec), angular stddev 0.0319363 degrees
Entry	8.33039 22.2988 pixel_4922510
Local freq	stddev 0.807471 Hz
Chirp	mean -0.980773, stddev 0.475106
Period	mean 0.837456, stddev 0.0785489
SNR	mean 14.1702, stddev 0
Pixel observations	3764 observations; 30570 sec
Score	All 20.45 signal+time -1.90 ND+time 20.85 time -1.49 ND+signal 21.94 signal -0.41 ND 22.35
Score factors	nd_factor 22.347332 p_factor 0.657971; f_factor 0.000020; nsigs_mp 38 nsigs_tot 8833 signal_factor -0.406544 time factor -1.492676 entropy 0.000000 tg 0.000000
Detection freq stddev	50090.839569237
Bary freq stddev	69443.129032054
Signal time range	2.31 years

Ratings

Who	When	Rating	Comment
Dan	1 May 2023, 20:51:36 UTC	-1	EK - example of interesting barycentric candidate - seen on seven different days
Dan	9 Dec 2022, 21:27:06 UTC	2	7 signals, 7 different days, low chirp

Edit rating

Freq/time plot Show all detections in pixel disc? Show topo freq? min score: limit chirp rate?

RA/dec plot

Detections

	ID Click for waterfall plot	Score	Time Show as Julian date	Detection freq	Barycentric freq	RA/dec Click for sky map	Pixel	Freq res Hz	Chirp rate	Beam	Period	Type
1	3548246067	43.293148	14 Nov 2008, 8:42:50	1,420,783,653	1,420,650,313	(8.330288, 22.267595)	4922505	19.07349	-0.982489	6	0.865075	triplet
2	3668359976	43.602433	22 Dec 2008, 8:05:08	1,420,785,900	1,420,712,416	(8.330923, 22.286181)	4922508	38.14697	-0.73756	5	0.943718	triplet
3	3351105050	44.32298	22 Mar 2009, 23:46:46	1,420,723,074	1,420,844,324	(8.33189, 22.327034)	4922511	38.14697	-0.244929	1	0.904397	triplet
4	3045682265	45.780214	27 Nov 2010, 7:17:52	1,420,814,455	1,420,694,591	(8.328176, 22.304052)	4922529	152.58789	-0.981565	5.5	0.799539	triplet
5	2978012454	44.566656	14 Dec 2010, 7:03:41	1,420,883,820	1,420,792,376	(8.329615, 22.342298)	4922533	9.53674	-1.103567	1	0.734003	triplet
6	2791821377	47.063011	14 Jan 2011, 4:13:49	1,420,819,982	1,420,797,750	(8.331305, 22.284298)	4922508	152.58789	-0.975095	0.5	0.750387	triplet
7	2825820595	43.714037	7 Mar 2011, 2:29:00	1,420,737,315	1,420,834,867	(8.330508, 22.280205)	4922505	19.07349	-1.840203	6	0.865075	triplet

Figure 13. Example of a multiplet summary page.

detections from different receiver beams. The detection under examination is the blue X in the center, inside the small box. The right panel shows the relative telescope angle (in units of beam width) versus time. The relative angle plotted is the difference between the telescope coordinates (which change with time) and the coordinates of the detection being examined. The detection in the center of the left plot is probably caused by RFI because it was seen four other times at roughly the same

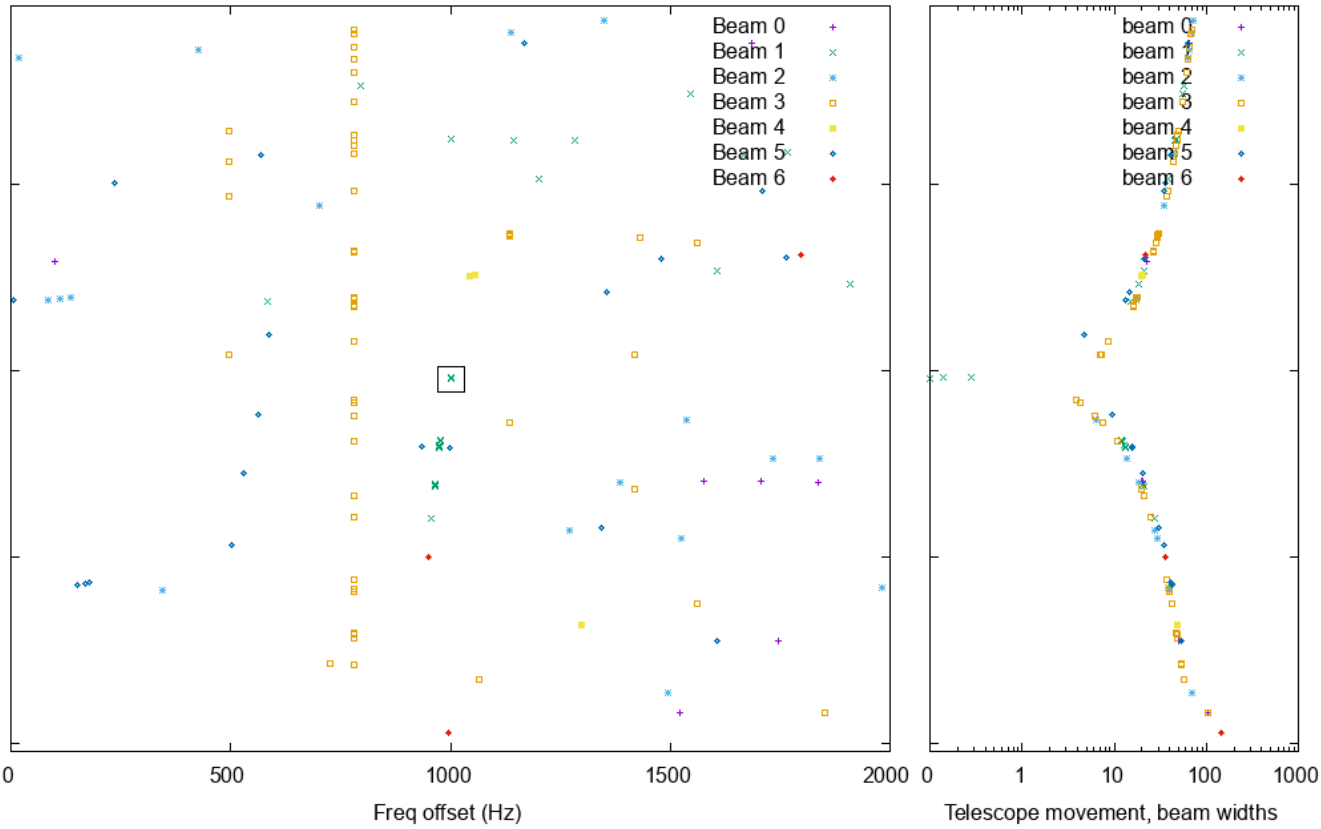

Figure 14. Waterfall plot showing detections from a high scoring multiplet that is likely due to RFI. The detection at the center of the plot is indicated with a square.

frequency while the telescope was pointing several beams away from the center detection. (the four detections with blue X symbols sloping down and left from the center detection).

In contrast, Figure 15 shows an example of a detection from a barycentric multiplet that was given a high manual rating as well as a high machine rating. The axes and symbols are similar to those in Figure 14. This multiplet was given high ratings for several reasons: it was detected several different times during several different observations; all these detections are clustered closely on the sky (within one beamwidth); all detections are clustered tightly in barycentric frequency (50 Hz RMS); none of the detections in the multiplet looked like RFI; each time that sky position was observed, there were several detections made while the telescope was tracking that sky position, and when the telescope moved away from that position, detections immediately ceased (this is consistent with a point source; it is unlikely RFI would do this, and extremely unlikely that RFI would do this multiple times). This multiplet was selected for reobservation at the FAST telescope.

For spike/Gaussian multiplets we concentrated on the density and time score variants, since that was most effective for finding both barycentric and nonbarycentric birdies (§9.2). Many of the top-ranked multiplets consisted of two detections within one minute. We decided to rule out these and consider only multiplets with a time span of at least 0.15 days. We only selected spike/Gaussian barycentric multiplets whose detections had drift rates consistent with the barycentric reference frame (drift rates < 0.7 Hz s^{-1}). We first inspected highly ranked barycentric multiplets containing three or more

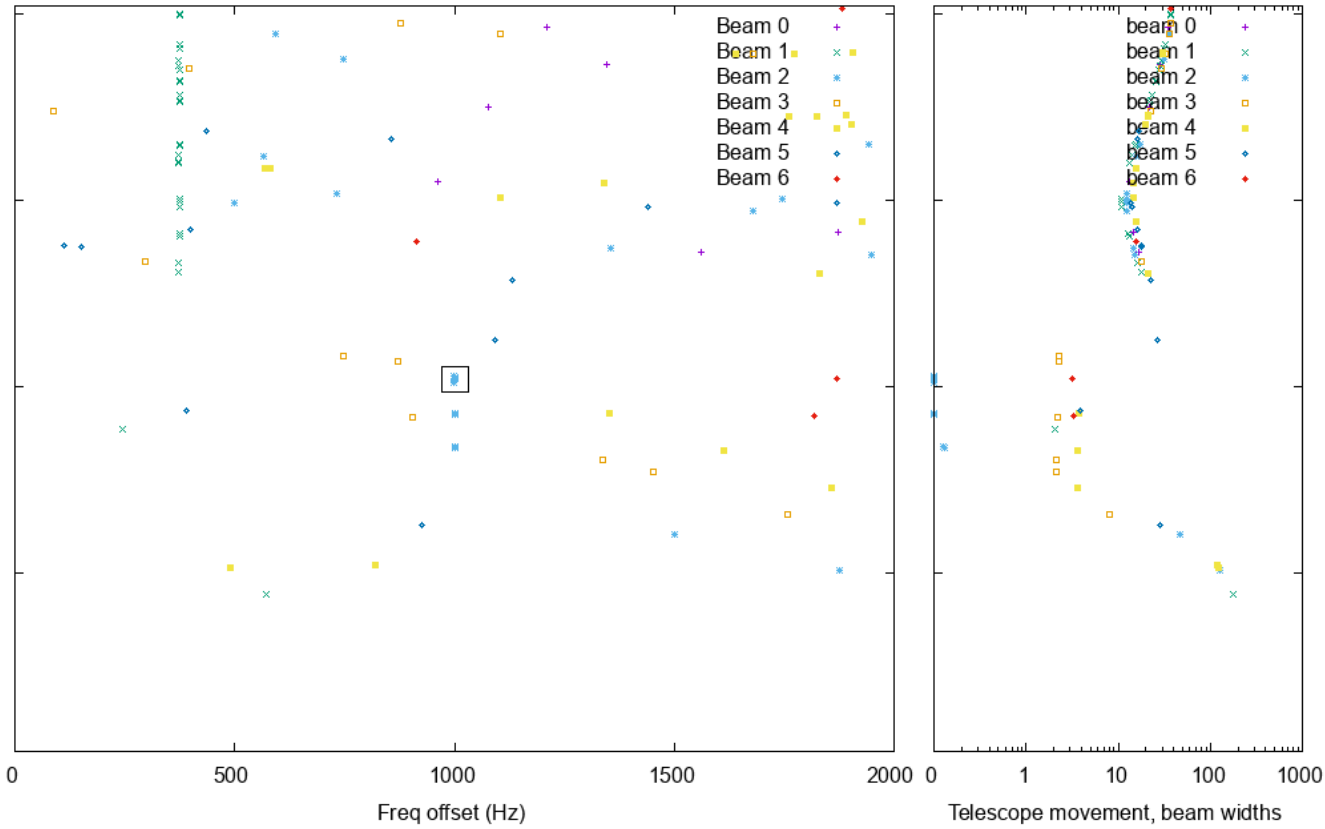

Figure 15. Detections from a multiplet that was scored highly by both algorithms and manually.

spikes or Gaussian signals; we then considered highly ranked multiplets composed of 1 Gaussian and 1 spike, and finally we selected a few multiplets with 2 spikes.

10. RELATED WORK AND CONTRIBUTIONS OF SETI@HOME

As discussed in §1, there have been many radio SETI projects using both sky survey and targeted search methods. Like SETI@home, each of these projects has a front end and a back end. Differences among front ends are discussed in (Korpela et al. 2025). The differences among the back ends include the following.

Candidate birdies: Some radio SETI projects (including SETI@home) tested their event detection by injecting constant-frequency or drifting sinusoids (Margot et al. 2023). SETI@home is the first project to use high-level candidate birdies that simulate persistent ET signals with a range of powers, frequencies, bandwidths, and transmitter orbital parameters. This lets us estimate the probability of detecting ET signals with different combinations of parameters and provides a basis for evaluating and refining our RFI rejection and candidate detection algorithms.

RFI detection algorithms: SETI@home's RFI detection algorithms are novel to varying extents and may be useful in future radio astronomy projects. In particular, the algorithms for detecting zone and drifting narrowband RFI are new. Several projects have used some form of cross-beam

RFI rejection (Korpela et al. 2009; Parsons et al. 2004; Harp et al. 2016; Chennamangalam et al. 2017).

Repeated observations separated in time: SETI@home recorded a decade-long archive of detections. During twelve years of observation, SETI@home has observed most of the Arecibo sky several times. The observations of a given sky position are typically widely separated in time. Compared to projects in which a given sky position is observed only once, this increases the likelihood of detecting these types of signals:

a) Sporadic signals, such as signals from low duty cycle transmitters (for example, a transmitter on the surface of a planet may be invisible 50% of the time as the planet rotates).

b) Scintillated signals: scintillation can cause a narrowband signal to be amplified or attenuated when propagating through the interstellar medium, on timescales from hours to months, possibly causing the signal to go above or below detectable thresholds (Cordes et al. 1997).

Search for long-duration nonbarycentric signals: SETI@home searches for nonbarycentric signals which, due to the transmitter's planetary motion, vary widely (± 125 kHz) over time in received frequency. The companion SERENDIP IV project at AO also searched for such signals (Cobb et al. 2000).

SETI@home's algorithms and techniques may be useful in future radio SETI projects. Its software is open source, so it can be used and adapted by such projects. The source code is available at https://sourceforge.net/projects/seti-science/.

The web interfaces described here are visible online at https://setiathome.berkeley.edu/nebula/. We maintained a public blog describing the evolution of the back end:

https://setiathome.berkeley.edu/forum_forum.php?id=1511.

11. FUTURE WORK

11.1. *Reobservation of signal candidates*

We are currently reobserving the selected signal candidates (see §9.5) at the Five-hundred-meter Aperture Spherical Telescope (FAST) in China. For each candidate, we do a single scan across its sky position at 0.7 times the sidereal rate, recording data with all 19 beams of the FAST telescope. This scan rate allows the SETI@home client to find all detection types, including spikes and Gaussians at their longest DFT lengths. The scan covers 0.47 degrees and lasts 172 seconds. Including slewing time, it takes about 15 minutes to reobserve each candidate.

We analyze the resulting data in three ways. First, we use the SERENDIP VI system, described in Zhang et al. (2020) which operates commensally at FAST, to search for strong narrowband signals near the candidate's frequency.

Second, we use the SERENDIP VI system to record baseband data. A polyphase filter bank in SERENDIP VI generates 32768 15.26-kHz wide "coarse channels" from the 500 MHz (1.0-1.5 GHz) bandwidth of the FAST receivers. The coarse channels that overlap the 2.5 MHz SETI@home band

are recorded as 16 bit complex samples in baseband format and analyzed in this format to avoid quantization losses. We generate SETI@home workunits from these, analyze them using the SETI@home client program, and manually check the output for detections that match the candidate parameters.

We can use Eqn. 13 of Korpela et al. (2025) to evaluate the sensitivity of this method. Our targets can be up to 27° from the FAST zenith. The beam performance from Jiang et al. (2020) gives a worst case system temperature of 19 K at the zenith and 24 K at 27° from the zenith, versus 29 K for ALFA, and a gain of 13.2 K Jy^{-1} relative to ALFA's 8.6K Jy^{-1}. Minimal quantization losses result in a minimum effective area of 28 200 m^2, or 2.56× that of our Arecibo observations. However, because of the increased workunit bandwidth, our channel bandwidth has increased by 1.56×. For spikes, triplets and autocorrelation, the integration time has decreased by the same factor. Therefore overall sensitivity to these signal types is improved about 2.0× over our Arecibo observations. Because we conduct our observations at 0.7× the sidereal rate in order to compensate for the narrower beamwidths at FAST, integration times for Gaussians and pulses are essentially unchanged, resulting in a sensitivity improvement of 2.5× over our Arecibo data.

Third, we manually investigate the expected frequency bands using coherent Doppler correction at the barycentric drift rate. We generate logarithmically scaled waterfall images at multiple time and frequency resolutions, and look for features missed by the SETI@home client.

We began to reobserve candidates at FAST on September 24, 2022. So far we have reobserved 80 candidates.

11.2. *Possible refinements of the front end*

As described by Korpela et al. (2025), there are several ways in which the SETI@home front end could be improved; we could have workunits include data from all 14 beams and polarizations instead of just one. We could merge the spike and Gaussian detection types. These changes would improve event sensitivity. This would also allow us to do cross-beam RFI rejection in the client, which would effectively improve event sensitivity by letting us lower detection thresholds.

We have archived the ALFA data that we recorded, so we could re-analyze this data with an improved front end. However, it would be preferable to do a new sky survey using a telescope such as FAST or SKA, for two reasons:

- Since the start of SETI@home network bandwidth to the home and PC storage capacity have increased greatly, making it feasible to have much larger workunits. Thus, it would be feasible to cover a much larger frequency range: perhaps a factor of 10 or 100.
- The archived ALFA data was recorded via commensal observation, and its pointing trajectory was not well suited to detecting continuous signals. Future sky surveys should use, at least in part, a pointing strategy optimized for continuous signal detection (see §11.4).

11.3. *Possible refinements of the back end*

As described in §5, our use of birdies has served several purposes. We currently use only continuous birdies, and we generate only spike detections for them. Generating Gaussians as well could improve

our sensitivity estimates. We could also generate pulsed birdies, and generate pulses and triplets for them. This would help us improve our RFI algorithms for pulsed signals and would allow us to estimate our sensitivity to them.

Our multiplet-finding algorithms (Sections 7.5 and 7.6) consist of a sequence of stages, each of which removes detections to satisfy a constraint. This approach may not lead to the highest-scoring multiplets; for example, one stage may select detections that must be removed by a later stage. It might be better to use an algorithm, perhaps based on clustering or other AI methods, that combines constraints with score maximization.

The global drift-rate/frequency consistency constraint for nonbarycentric multiplets (§7.4.5) ensures that detections do not change too fast for our planetary motion limits, but it does not ensure that they match a frequency trajectory resulting from a set of orbital and rotational parameters within the limits. A more sophisticated algorithm could do this.

It would also be possible to search the SETI@home detection archive for signals co-located (and co-accelerating) with Solar System objects such as planets, moons, asteroids, and Kuiper Belt Objects (KBO). Such an analysis would use object ephemerides for position clustering, with frequency correction based on the changing object frames.

11.4. *Pointing strategies for radio SETI sky surveys*

Perhaps the most important lesson from SETI@home is that, in a sky survey with high frequency resolution, pointing matters. Sensitivity to continuous narrowband signals requires long observations during which the telescope is drifting slowly or not at all. Commensal observation may not provide such data.

We can estimate how much observing time is needed to survey the sky with high frequency resolution. As an example: in SETI@home there are 15M observable pixels. Observing each for 13.4 s (corresponding to a frequency resolution of 0.075 Hz) would take a total of 6.37 years. A multibeam receiver can observe multiple pixels simultaneously; so, for example, with the FAST 19-beam receiver the observing time could in principle be reduced to 0.33 years. (Ideally, for the reasons given in §10, this survey would be done at least twice).

What are the optimal pointing strategies? This question merits study; there is a trade-off between sensitivity and telescope time. If we do a continuous sweep, the beam crossing time should be at least the bin duration of the longest DFT length. And if it is longer – say 2× or 3× – this would provide the RFI rejection that is built into the Gaussian fitting. Looking at it the other way: if a constant slew rate is given, we can avoid needlessly long DFT lengths.

For RFI detection purposes, it is useful to observe each point after a short delay (a minute or so). This can be done with a back-and-forth pointing pattern. With multibeam receivers, a similar effect can be achieved by rotating the receiver so that 2 or more beams lie in the direction of motion. However, in both cases, the benefit comes at the cost of increased observing time.

If a sky survey's telescope movement were regular – for example, if beams always moved at the same rate – the algorithms for some types of RFI detections, for signal detection, and for candidate evaluation could be more simple and probably more effective than the flexible but complex algorithms

we developed for SETI@home. However, in light of the expense of large radio telescopes and the high demand for telescope time from other areas of radio astronomy, sensitive sky surveys may be limited to commensal observing for the foreseeable future.

12. ACKNOWLEDGEMENTS

Millions of SETI@home volunteers supplied computing power for front-end data processing. Volunteers contributed in many other ways, such as providing technical support to other volunteers, moderating message boards, translating Web site text, and porting the SETI@home application to GPUs; see Korpela et al. (2025). Hundreds of people have helped develop SETI@home's software and hardware systems; the contributions of Matt Lebofsky and Charlie Fenton were especially valuable. David Gedye had the idea for SETI@home and assembled its initial team.

SETI@home has been supported by grants from Starwave, The Planetary Society, the state of California, National Science Foundation grant 1407804, the Marilyn and Watson Alberts SETI Chair fund, and by donations from individuals. We received equipment donations from Sun Microsystems, Intel, NetApp, NVIDIA, AMD/Xilinx, Hewlett Packard, Fujitsu, Quantum, Seagate, Western Digital, and Packet Clearing House.

SETI@home's back-end processing used the Atlas cluster at the Albert Einstein Institute in Hanover Germany. We thank Bruce Allen for making this available.

This research used resources from the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the US Department of Energy under Contract No. DE-AC02-05CH11231.

REFERENCES

Anderson, D. P. 2020, J. Grid Comput., 18, 99, doi: 10.1007/s10723-019-09497-9

Backus, P. R., & Project Phoenix Team. 2002, in ASP Conf. Ser., Vol. 278, Single-Dish Radio Astronomy: Techniques and Applications, ed. S. Stanimirovic, D. Altschuler, P. Goldsmith, & C. Salter, 525–527. https://ui.adsabs.harvard.edu/abs/2002ASPC..278..525B

Bowyer, S., Lampton, M., Korpela, E., et al. 2016, ArXiv e-prints. https://arxiv.org/abs/1607.00440

Chennamangalam, J., MacMahon, D., Cobb, J., et al. 2017, ApJS, 228, 21, doi: 10.3847/1538-4365/228/2/21

Cobb, J., Lebofsky, M., Werthimer, D., Bowyer, S., & Lampton, M. 2000, in ASP Conf. Ser., Vol. 213, Bioastronomy 99: A New Era In Bioastronomy, ed. G. Lemarchand & K. Meech, 485. https://ui.adsabs.harvard.edu/abs/2000ASPC..213..485C

Cocconi, G., & Morrison, P. 1959, Nature, 184, 844, doi: 10.1038/184844a0

Cohen, R. J., Downs, G., Emerson, R., et al. 1987, MNRAS, 225, 491, doi: 10.1093/mnras/225.3.491

Cordes, J. M., Lazio, J. W., & Sagan, C. 1997, ApJ, 487, 782, doi: 10.1086/304620

Cordes, J. M., Freire, P. C. C., Lorimer, D. R., et al. 2006, ApJ, 637, 446, doi: 10.1086/498335

Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. 2022, Introduction to Algorithms (MIT press)

Court, R. W., & Sephton, M. A. 2012, Planet. Space Sci., 73, 233, doi: 10.1016/j.pss.2012.08.026

Czech, D., Isaacson, H., Pearce, L., et al. 2021, PASP, 133, 064502, doi: 10.1088/1538-3873/abf329

Drake, F. D. 1960, S&T, 19, 140. https//archive.org/details/sim_sky-and-telescope_1960-01_19_3/page/140

—. 1974, in Interstellar Communication: Scientific Perspectives, ed. C. Ponnamperuma & A. G. W. Cameron, 118–139

Drake, F. D. 1986, in NRAO Workshop on the Search for Extraterrestrial Intelligence, ed. K. I. Kellermann & G. A. Seielstad, 17–26. https://ui.adsabs.harvard.edu/abs/1986seti.work...17D

Enriquez, J. E., Siemion, A., Foster, G., et al. 2017, ApJ, 849, 104, doi: 10.3847/1538-4357/aa8d1b

Górski, K. M., Hivon, E., Banday, A. J., et al. 2005, ApJ, 622, 759, doi: 10.1086/427976

Gray, R. H. 2021, JAHH, 24, 981. https://ui.adsabs.harvard.edu/abs/2021JAHH...24..981G

Guttman, A. 1984, SIGMOD Rec., 14, 47, doi: 10.1145/971697.602266

Haqq-Misra, J. 2024, in Oxford Research Encyclopedia of Planetary Science, 275, doi: 10.1093/acrefore/9780190647926.013.275

Harp, G. R., Richards, J., Tarter, J. C., et al. 2016, AJ, 152, 181, doi: 10.3847/0004-6256/152/6/181

Hort, E., Sheikh, S., Farah, W., & Tusay, N. 2024, in American Astronomical Society Meeting Abstracts, Vol. 243, American Astronomical Society Meeting Abstracts, 109.03

Jiang, P., Tang, N.-Y., Hou, L.-G., et al. 2020, Research in Astronomy and Astrophysics, 20, 064, doi: 10.1088/1674-4527/20/5/64

Korpela, E., Werthimer, D., Anderson, D., Cobb, J., & Leboisky, M. 2001, Computi. Sci. & Eng., 3, 78, doi: 10.1109/5992.895191

Korpela, E. J., Anderson, D. P., Cobb, J., Lebofsky, M., & Werthimer, D. 2025, Submitted to AJ

Korpela, E. J., Cobb, J., Lebofsky, M., et al. 2011, in Communication with Extraterrestrial Intelligence (CETI), ed. D. A. Vakoch (Albany, NY, USA: SUNY Press), 37–44. https://arxiv.org/abs/1109.1595

Korpela, E. J., Demorest, P., Heien, E., Heiles, C., & Werthimer, D. 2002, in ASP Conf. Ser., Vol. 276, Seeing Through the Dust: The Detection of HI and the Exploration of the ISM in Galaxies, ed. A. R. Taylor, T. L. Landecker, & A. G. Willis, 100–103. https://arxiv.org/abs/astro-ph/0112300

Korpela, E. J., Anderson, D. P., Bankay, R., et al. 2009, in ASP Conf. Ser., Vol. 420, Bioastronomy 2007: Molecules, Microbes and Extraterrestrial Life, ed. K. J. Meech, J. V. Keane, M. J. Mumma, J. L. Siefert, & D. J. Werthimer , 431–438. https://aspbooks.org/custom/publications/paper/420-0431.html

Korpela, E. J., Anderson, D. P., Bankay, R., et al. 2011, in Proc. SPIE, Vol. 8152, Instruments, Methods, and Missions for Astrobiology XIV, ed. R.B. Hoover, P.C.W. Davies, G.V. Levin, & A.Y. Rozanov, 815212 (8 pages), doi: 10.1117/12.894066

Kraus, J. D. 1977, Vistas in Astronomy, 20, 445, doi: 10.1016/0083-6656(77)90027-7

Li, J.-K., Zhao, H.-C., Tao, Z.-Z., Zhang, T.-J., & Xiao-Hui, S. 2022, ApJ, 938, 1, doi: 10.3847/1538-4357/ac90bd

Luan, X.-H., Tao, Z.-Z., Zhao, H.-C., et al. 2023, AJ, 165, 132, doi: 10.3847/1538-3881/acb706

Margot, J.-L., Li, M. G., Pinchuk, P., et al. 2023, AJ, 166, 206, doi: 10.3847/1538-3881/acfda4

Miller, S. L. 1953, Science, 117, 528, doi: 10.1126/science.117.3046.528

Miller, S. L., & Urey, H. C. 1959, Science, 130, 245, doi: 10.1126/science.130.3370.245

Parsons, A., Werthimer, D., Anderson, D., et al. 2004, in 55th International Astronautical Congress (International Astronautical Federation, 8-10 rue Mario-Nikis, Paris Cedex, 15, France). http://setiathome.berkeley.edu/~korpela/papers/2004-10-01_Searching_For_ET_SETI_overview.pdf

Pearce, B. K. D., & Pudritz, R. E. 2015, The Astrophysical Journal, 807, 85, doi: 10.1088/0004-637X/807/1/85

Peek, J. E. G., Begum, A., Douglas, K. A., et al. 2010, in ASP Conf. Ser., Vol. 438, The Dynamic Interstellar Medium, ed. R. Kothes, T. L. Landecker, and A. G. Willis, 393–401. https://aspbooks.org/custom/publications/paper/438-0393.html

Peek, J. E. G., Heiles, C., Douglas, K. A., et al. 2011, ApJS, 194, 20 (13 pages), doi: 10.1088/0067-0049/194/2/20

Price, D. C., Enriquez, J. E., Brzycki, B., et al. 2020, AJ, 159, 86, doi: 10.3847/1538-3881/ab65f1

Rivilla, V. M., Sanz-Novo, M., Jiménez-Serra, I., et al. 2023, ApJL, 953, L20, doi: 10.3847/2041-8213/ace977
Sagan, C., & Drake, F. 1975, SciAm, 232, 80. http://www.jstor.org/stable/24949801
Siemion, A. P. V., Bower, G. C., Foster, G., et al. 2012, ApJ, 744, 109, doi: 10.1088/0004-637X/744/2/109
Staelin, D. H. 1969, IEEE Proc., 57, 724, doi: 10.1109/PROC.1969.7051
Thain, D., Tannenbaum, T., & Livny, M. 2005, Concurrency Computat.: Pract. Exper. , 17, 323, doi: 10.1002/cpe.938
Tokadjian, A., Hu, R., & Damiano, M. 2024, AJ, 168, 292, doi: 10.3847/1538-3881/ad88eb
Tremblay, C. D., Varghese, S. S., Hickish, J., et al. 2024, AJ, 167, 35, doi: 10.3847/1538-3881/ad0fe0
Tusay, N., Sheikh, S. Z., Sneed, E. L., et al. 2024, AJ, 168, 283, doi: 10.3847/1538-3881/ad823c
Werthimer, D., Brady, R., Berezin, A., & Bowyer, S. 1988, Acta Astronautica, 17, 123, doi: 10.1016/0094-5765(88)90135-X
Zeng, S., Quénard, D., Jiménez-Serra, I., et al. 2019, MNRAS, 484, L43, doi: 10.1093/mnrasl/slz002
Zhang, Z.-S., Werthimer, D., Zhang, T.-J., et al. 2020, ApJ, 891, 174, doi: 10.3847/1538-4357/ab7376

APPENDIX

A. ALGORITHM: CALCULATE PER-PIXEL OBSERVATION INTERVALS

Input: The pointing history of each beam b, represented as a time-ordered list of (time, position) pairs.

Output: For each pixel P, a list $I(P)$ of non-adjacent time intervals during which some beam's sky position was within θ_{beam} of the center of P.

The algorithm loops over beams b, calculating for each pixel P a list $I(P, b)$ of non-adjacent time intervals during which the position of b was within θ_{beam} of the center of P.

Since the pointing sampling interval is 1 Hz, we assume that if consecutive pointings are separated by at least 10 seconds, there is a gap in the data.

We first add interpolated virtual pointings so that no pixels are skipped. To do this, we scan the list of b's pointings in time order. Let A_1 and A_2 be consecutive pointings. If the time difference between A_1 and A_2 is more than 10 seconds, we skip the pair. Otherwise, for a point A, let $D(A)$ be the set of pixels whose centers are within θ_{beam} of A. $D(A)$ approximates the sky area seen by b (in the above sense) at the time of the pointing. If $D(A_1)$ and $D(A_2)$ differ, we add a virtual pointing A_{mid} halfway between A_1 and A_2 in time, RA and dec. We then recursively do the same check for (A_1, A_{mid}) and (A_{mid}, A_2), possibly adding further pointings.

Having added the interpolated pointings, we compute, for each pixel P, a list $I(P, b)$ of observation intervals, initially empty. We scan the interpolated pointing list in time order, again skipping pairs separated by more than 10 seconds. For the other consecutive pairs (A_1, A_2), we add the time interval to $I(P, b)$ for each pixel P in the union of $D(A_1)$ and $D(A_2)$.

After completing this scan, we merge adjacent intervals. The resulting lists $I(P, b)$ are the output of the function.

Having calculated $I(P, b)$ for the seven beams b, we merge these to form $I(P)$.

B. ALGORITHM: GENERATE BIRDIE DETECTIONS

Input: a set S of birdies.

Output: a set of spike detections, approximating what the SETI@home front end would have produced had the birdie signals been present in its input data.

For each of the 7 receiver beams b, we scan through its pointing history: its sky positions over the SETI@home observing period, interpolated as described in §4.2. See Figure B.1

At each pointing (P_{time}, P_{pos}), we find the set $S_0 \subseteq S$ of birdies B with position close enough to P_{pos} (within two beam widths) that could potentially produce a detection. We consider the time interval I from P_{time} to the time of the next pointing. We loop over each DFT length ℓ as the birdie could potentially produce detections at any of these. For each DFT length, we loop over the DFT time intervals that overlap I.

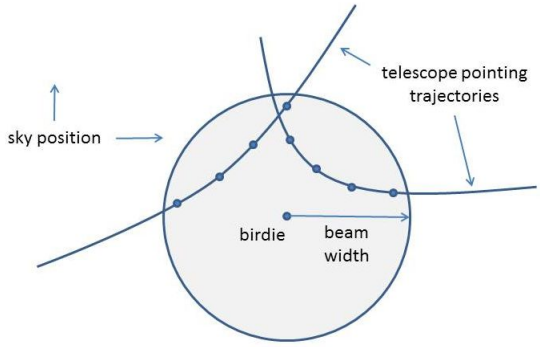

Figure B.1. Birdie detections are generated when a telescope beam passes close to the birdie sky position.

For each birdie $B \in S_0$ we compute the angle θ between P_{pos} and $(\alpha_{\rm B}, \delta_{\rm B})$. Since the sensitivity of the beam is a Gaussian function, the attenuation due to this angular separation is

$$A_{angle} = e^{-\frac{\theta}{\theta_{\rm beam}}} \tag{B1}$$

Recall that $\Delta\nu(\ell)$ is the frequency resolution that corresponds to ℓ. The attenuation due to the disparity between $\Delta\nu(\ell)$ and $\Delta\nu(B)$ is

$$A_{freq} = \sqrt{\frac{\Delta\nu(\ell)}{\Delta\nu(B)}} \tag{B2}$$

if $\Delta\nu(B) > \Delta\nu(\ell)$, and

$$A_{freq} = \sqrt{\frac{\Delta\nu(B)}{\Delta\nu(\ell)}} \tag{B3}$$

otherwise.

To model the effects of noise, we add a random number R from a normal distribution with mean 0 and stddev 1.

The resulting total power is

$$P = (B_{power} A_{angle} A_{freq}) + R \tag{B4}$$

If P is above the threshold for spikes, we create a spike D with $P(D) = P$, $pos(D) = P_{pos}$, and $t(D)$ equal to the midpoint of the DFT time interval.

$\nu_{\rm bary}(D)$ is $\nu(B)$ plus the Doppler shifts resulting from the movements of the sender and receiver. If B is barycentric, the former is zero; otherwise it is the Doppler shift due to the motion of B' at time t. The latter is $\Delta\nu_{\rm AO}(b,t)$.

$\nu_{\rm topo}(D)$ is $\nu_{\rm bary}(D) + \Delta\nu_{\rm AO}(b,t)$. We then round $\nu_{\rm topo}(D)$ down to a multiple of the frequency resolution of DFT length ℓ.

$\frac{\Delta\nu_{\rm topo}}{\Delta t}(D)$ is the sum of sender and receiver components. The former is the derivative of B's Doppler shift as determined by its planetary motion parameters, or zero for barycentric birdies. The latter is $\frac{d(\Delta\nu_{\rm AO})}{dt}(b,t)$.

Detections of a given DFT length ℓ are spaced in time by at least $\Delta t(\ell)$, the duration corresponding to ℓ. So we must ensure that this holds for birdie detections. For long DFT lengths this may require skipping over pointing intervals For short DFT lengths, we may create several detections within a single pointing interval.

C. ALGORITHM: DETECT FREQUENCY-ZONE RFI

The prevalence of zone RFI varies with detection type T and DFT length ℓ, so we process each combination separately. The band fraction removed, denoted $Z_{frac}(t, \ell)$, depends on the combination.

For each detection type and DFT length, we divide the set of detections into 0.1-day *time windows* and, for each window, we flag frequency bins that have a statistical excess of detections. Then, for each frequency bin, we count the number of windows during which the bin was flagged. We find the count threshold for which at least $Z_{frac}(t, \ell)$ of the bins are over threshold, and flag detections in these bins as RFI.

Input: A set S of detections of a given type T (spike or Gaussian) and DFT length ℓ.

Output: A set $S' \subseteq S$ of detections marked as RFI.

Algorithm parameters:

$Z_{\Delta\nu}$**:** The size of the frequency bin. This is twice the DFT frequency resolution, to account for frequency imprecision due to Doppler drift.

$Z_{\Delta t}$**:** The duration of a window: 0.1 days.

Z_{prob}**:** A probability threshold used in deciding whether a bin has an excess of detections: 10^{-7}.

$Z_{frac}(t, \ell)$**:** The fraction of bins to flag as RFI, determined separately for each combination of detection type and DFT length, as described above.

We divide the overall frequency range into bins of size $Z_{\Delta\nu}$ Let *nbins* be the number of bins.

We scan S in time order, accumulating a window W of detections. W is complete when

- adding the next detection to W would cause its time span to exceed $Z_{\Delta t}$, and
- the number of detections in W is at least half the average number of detections of the given type and DFT length during a period of duration $Z_{\Delta t}$, averaged over the entire SETI@home observation period.

As we process windows, we maintain an array $FC[i]$ containing, for each frequency bin i, the count of windows during which the bin had an excess of detections.

When a window is complete, we analyze it to identify bins with a statistical excess of detections. An RFI source may not be contained in a single bin; it may have a significant bandwidth, or its frequency may vary. So, after analyzing the bins singly, we combine them into groups of 2, 4 and so on. If one of these combined bins has an excess of detections, all of its component bins are flagged. This is described in the pseudocode in Fig. C.1.

```
stride = 1
nsigs = |W|
for i = 0; i < nbins; i += 1 do
    C[i] =number of detections in bin i
end for
while true do
    thresh = stride * Z_prob/nbins
    avg = stride * nsigs/nbins
    n = the smallest number for which Γ(n, avg) < thresh
    for i = 0; i < nbins; i += stride do
        if C[i] is greater then n then
            F[i]...F[i + stride − 1] = True
            C[i] = 0
        end if
    end for
    if stride * 2 > nbins/2 then
        break
    end if
    stride = stride * 2
    for i = 0; i < nbins; i += stride do
        C[i] = ∑(C[i]..C[i + stride − 1])
    end for
end while
for i = 0; i < nbins; i += stride do
    if F[i] then
        FC[i] += 1
    end if
end for
```

Figure C.1. Frequency-zone RFI Algorithm

Having processed all windows, the array $FC[i]$ contains the number of windows in which bin i was flagged as having an excess of detections. Let X be the $1 - Z_{frac}(t, \ell)$ quantile of FC; i.e. the value for which $Z_{frac}(t, \ell)$ of the bins are above X.

Let B be the set of bins i for which $FC[i] > X$. Let S' be the subset of S consisting of detections whose bin is in B; these detections are flagged as RFI.

To determine appropriate values for $Z_{frac}(t, \ell)$, we examined graphs of the fraction of signals removed as a function of the fraction of frequency zones removed, assuming that zones are removed in decreasing order of $E(Z)$. These graphs – one per detection type and DFT length – have a knee because a small fraction of zones contain disproportionate numbers of detections. An example is shown in Figure C.2. We choose $Z_{frac}(t, \ell)$ to be a value slightly above this knee.

Note that in deciding whether to remove a zone, we use window count instead of number of detections in the zone. If we used the number of detections, a zone could be removed due to a large excess of detections in a short time range. We do not want to remove that zone at all times because an ET

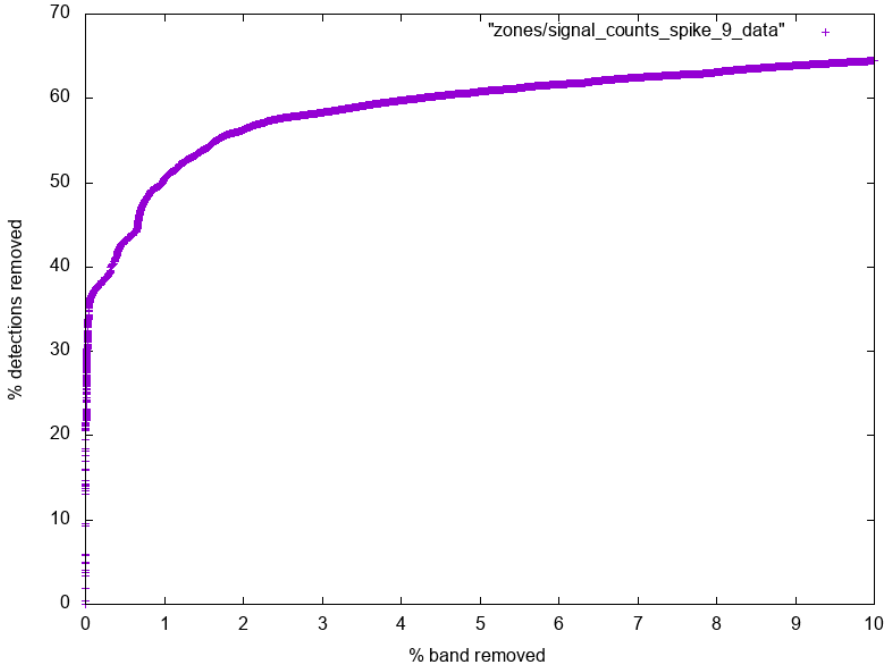

Figure C.2. Spikes removed as a function of frequency band removed for DFT length 4096. The knee of this curve (in this case about 2%) determines the fraction of the frequency band removed by the zone RFI algorithm.

signal could occur in that zone at a different time. This RFI would be removed by other RFI filters (such as the drifting filter; see below).

D. ALGORITHM: DETECT DRIFTING RFI

Input: A set S of detections.

Output: A set $S' \subseteq S$ of detections marked as drifting RFI.

Algorithm parameters:

$D_{\Delta t}$**:** The duration of a triangle: 600 s.

$D_{drift_{max}}$**:** The maximum drift rate: $5\,\mathrm{Hz\ s^{-1}}$.

$D_{\theta_{min}}$**:** Count only detections at least this far from D: $\frac{3}{2}\theta_{\mathrm{beam}}$.

D_n**:** The number of triangles in the fan: 21.

D_{prob_1}**:** Probability threshold for single triangles (see below): 10^{-8}.

D_{prob_2}**:** Probability threshold for opposed pairs of triangles: 10^{-4}.

First, divide S into clusters as described above. For each cluster, select one representative detection. Let C denote the set of these representatives.

For a detection $D \in C$, consider the rectangle in time/frequency space centered at $t(D)$, $\nu_{\text{topo}}(D)$ and of duration $D_{\Delta t}$ and frequency width $D_{\Delta t} D_{drift_{max}}$. Construct $2D_n$ equal-area triangles with one vertex at D and the other two on the top or bottom edge of the rectangle, as shown in Figure 2.

For each triangle T, compute the number $F(T) \subseteq C$ of detections that a) lie within T and b) whose angular distance from D is greater than $D_{\theta_{min}}$.

For each triangle T, compute an estimate $P(T)$ of the probability that T contains at least $F(T)$ detections. This is done as follows. For each of the two triangle fans (before and after D), let A and B be the median and mean, respectively, of $F(T)$ over the triangles T in the fan. If A is greater than zero, let

$$M = \min(A, B) \tag{D5}$$

otherwise let

$$M = \min(1, B) \tag{D6}$$

For each triangle T in the fan, let

$$P(T) = \Gamma(M, F(T)) \tag{D7}$$

where Γ is the incomplete Gamma function. This estimates the probability of a triangle containing at least $F(T)$ detections, given an average of M.

For each triangle T, consider the opposite triangle and the two adjacent triangles. If for T and an opposing triangle S, $P(T)$ and $P(S)$ are less than D_{prob_2}, then flag D as RFI. This tracks bands of drifting RFI whose drift rate changes over time. In addition, if for any triangle T, $P(T)$ is less than D_{prob_1}, then flag D as RFI. This flags the start and end of bands of drifting RFI.

If D is flagged as RFI by the above procedure, also flag the detections in the same cluster as D.

Repeat the above procedure for each $D \in C$.

E. ALGORITHM: DETECT MEDIUM-TERM PULSE AND TRIPLET RFI

Input: A set S of pulses or triplets.

Output: A subset $S' \subseteq S$ of detections flagged as RFI.

Algorithm parameters:

$P_{\Delta t}$. The period over which we look for RFI features: 10 minutes.

P_{prob}. The probability threshold for a statistical excess: 10^{-3}.

$P_{\Delta\theta}$. Detections are considered far if the angle between their positions exceeds $\frac{3}{2}\theta_{\text{beam}}$.

Scan S in time order, maintaining a window W of duration at most $P_{\Delta t}$. When adding a detection would exceed $P_{\Delta t}$, process W as follows:

Bin the detections in W by frequency, with bin size twice the median bandwidth for detections of that type (38 Hz for pulses, 305 Hz for triplets).

For each bin B_i, find the detection D_i whose time is closest to the midpoint of W. Let L_i be the set of detections in B_i with time before D_i that are far from D_i, that is, whose angular distance from D_i is at least $P_{\Delta\theta}$. Let U_i be the set of far detections after D_i.

Let N_{far} be the number of far signals in all bins.

Let M be the median size of the L_i and U_i.

Let X be the least integer such that

$$\Gamma(X, M) < P_{prob} \tag{E8}$$

X is the threshold for a statistical excess.

If, for a bin i, $|L_i| > X$ and $|U_i| > X$, then flag all the detections in that bin as RFI, including those that are not far.

Having processed W, advance the window by $P_{\Delta t}/2$ and continue scanning S.

F. ALGORITHM: DETECT MULTI-BEAM RFI

Input: A set S of detections.

Output: A subset $S' \subseteq S$ of detections flagged as RFI.

For each detection D, we define a rectangle $R(D)$ in time/frequency space. The center of the rectangle is $(t(D), \nu_{\text{topo}}(D))$. The size in each dimension represents the uncertainty in that attribute. If another detection D_2 is contained in this rectangle, then D_2 probably has the same source (cosmic or RFI) as D.

The size of $R(D)$ in the frequency dimension is $\Delta\nu(\ell)$, the frequency resolution corresponding to the detection's DFT length ℓ. The size in the time dimension is $\tau(D)$.

For each detection $D' \in S$, we find the set T of detections D of the same type for which $(t(D), \nu_{\text{topo}}(D))$ lies in $R(D')$. If D is a pulse or triplet, we include only detections with periods close to $p(D)$.

We then count the number N of detections $D' \in T$ for which the angle between $pos(D)$ and $pos(D')$ exceeds $1.75\theta_{\text{beam}}$. These detections are unlikely to have the same source as D. If N is at least half the size of T, we flag D as RFI.

G. ALGORITHM: LOCAL DRIFT-RATE/FREQUENCY PRUNING

Input: A set S of detections with a time span of at most $M_{\Delta t}(local)$.

Output: A subset $S' \subseteq S$ that satisfies the local drift-rate/frequency consistency constraints.

We enforce the constraints by pruning detections that violate them. First, we find a range of drift rates that contains a concentration of detections. To do this, we scan S in order of drift rate and find the range of drift rates $[C_0..C_1]$ of size $M_{\Delta c}(local)$ in which the sum of detection scores is greatest. We then discard detections with drift rates outside the interval. See Figure G.1.

Let C_{median} be the median drift rate of the remaining detections. For each detection D, we compute the barycentric frequency adjusted for the sender component of C_{median}:

$$\begin{aligned}\nu_{\text{adj}} =& \nu_{\text{bary}}(D) + C_{median}(t(D) - T_0)) \\ &- (\nu_{\text{bary}}(D) - \nu_{\text{topo}}(D))\end{aligned} \tag{G9}$$

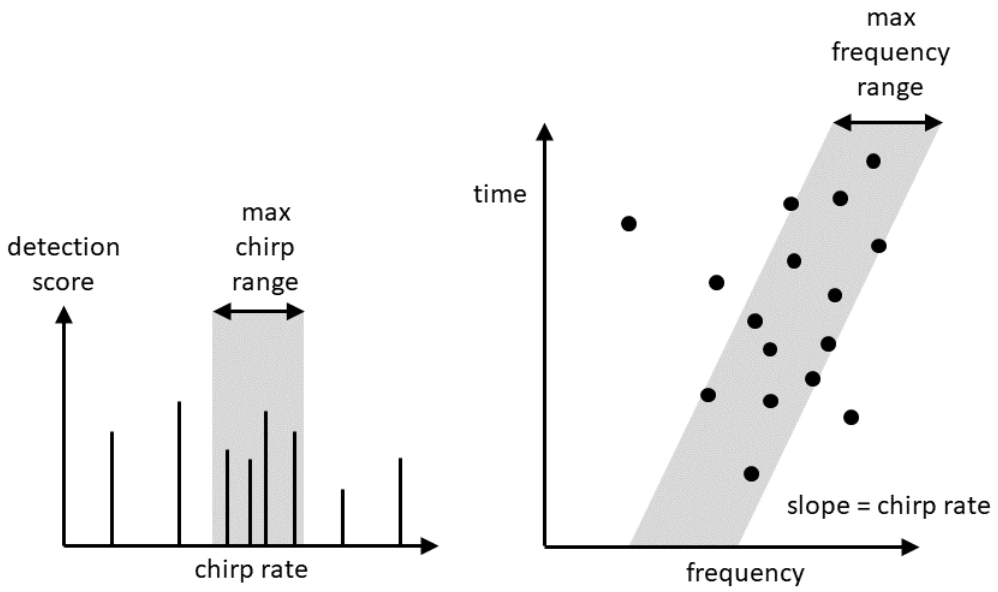

Figure G.1. Local drift-rate/frequency pruning. We a) find the drift rate interval containing the most detection power and b) find the frequency interval which when time-shifted by C_{median} contains the most detection power.

where T_0 is the time of the earliest detection in S.

We then scan the detections in order of increasing ν_{adj}, and find the interval $[f_0..f_1]$ of width $M_{\Delta\nu}(nonbary)$ in which the sum of the detection scores is greatest. S' is then the set of detections D for which ν_{adj} lies in that interval.

H. ALGORITHM: GLOBAL DRIFT-RATE/FREQUENCY PRUNING

Input: A set S of detections.

Output: A subset $S' \subseteq S$ satisfying both the local and global drift-rate/frequency constraints.

We scan M's detections in time order, using the above algorithm to process windows of duration at most $M_{\Delta t}(local)$. This produces, for each window, a locally consistent set of detections W_i.

We maintain a list $W_1...W_n$ of windows that we have already processed and which satisfy the global constraint. For each such window W, we store its set of detections, the median time, drift rate, and frequency of these detections, and the sum of their scores; the latter approximates the contribution of W to the overall multiplet score.

We say that two windows are consistent if their median times, drift rates, and frequencies satisfy the constraint in §7.4.5. Having processed a window W_{n+1}, we scan the list $W_1...W_n$ and find the set of windows that is inconsistent with W_{n+1}. If the sum of the values of these windows is less than the value of W_{n+1}, we discard the conflicting windows and add W_{n+1} to the window list. Otherwise, we remove the detections we selected from W_{n+1}, rerun the local consistency algorithm with the remaining detections in that time window, and check for consistency with the other windows. This is repeated until no detections remain, in which case we discard W_{n+1}. see Figure H.1.

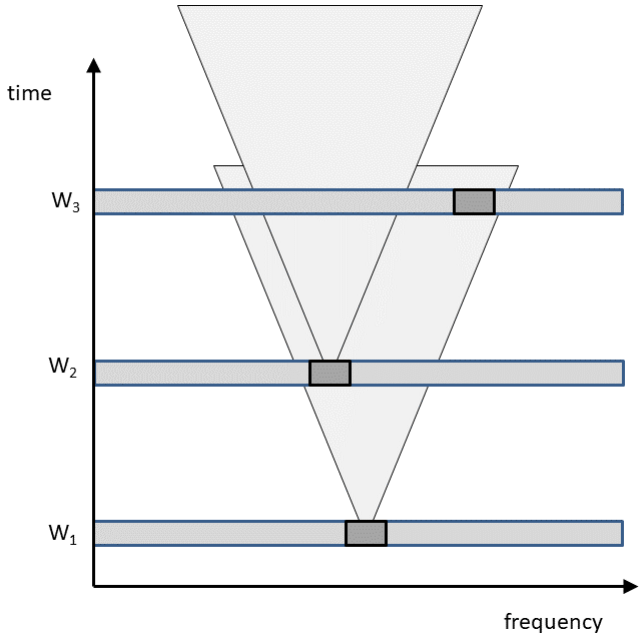

Figure H.1. Global drift-rate/frequency pruning. The detections selected in window W_3 are not compatible with those selected in W_2; we must discard the one with lower value.

I. ALGORITHM: FIND A BARYCENTRIC MULTIPLET IN A FREQUENCY BAND

Input: A set S of detections in a pixel disk and a frequency band of $M_{\Delta\nu}(bary)$ Hz.

Output: A multiplet M, consisting of detections in S, that satisfies the multiplet constraints, or null if no such M is found.

We first remove detections whose drift rate is outside the limits described in §7.4. We then do time-overlap pruning (see §7.5.3) on the remaining detections. Finally, for spike/Gaussian multiplets, we enforce frequency-variation consistency (see §7.4.6) by scanning the remaining detections in time order and removing groups of detections that violate the constraint.

If at least two detections remain, return M. Otherwise, return null.

J. ALGORITHM: FIND BARYCENTRIC MULTIPLETS IN A DETECTION DISK

Input: A set S of detections (the detection disk of a pixel).

Output: A set of barycentric multiplets made up of detections in S.

We scan through the detections in S in order of increasing $\nu_{\text{bary}}(D)$, maintaining a window W whose range of $\nu_{\text{bary}}(D)$ is at most $M_{\Delta\nu}(bary)$. We skip detections that do not satisfy the drift rate constraint for barycentric multiplets (see §7.4.2).

When adding a detection to W would exceed $M_{\Delta\nu}(bary)$, we look for a multiplet within W, using the algorithms described in Section 7.6. If this produces a multiplet M, we advance the window beyond the maximum frequency of M and continue. Otherwise (if no multiplet is found in W) we append

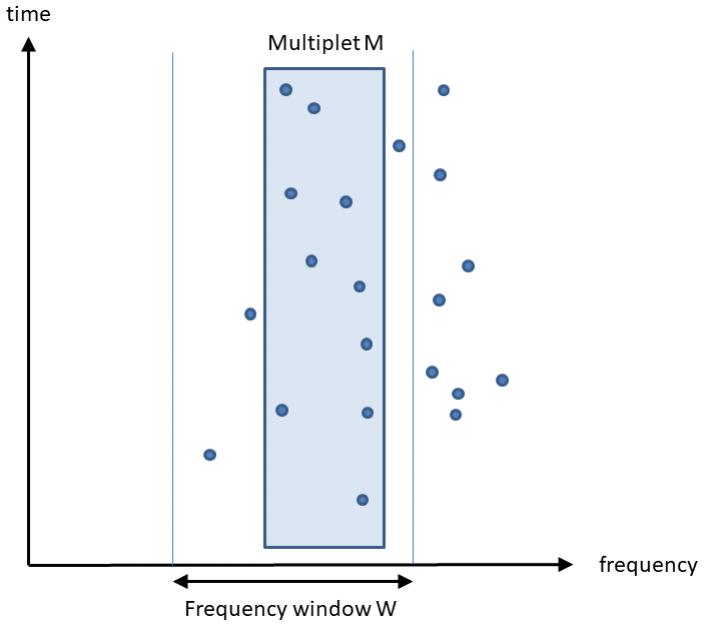

Figure J.1. M is the highest scoring multiplet in W, but there may be a higher scoring multiplet that overlaps M using higher-frequency detections.

the next detection to W and remove detections from the start of W as needed to limit its frequency range to $M_{\Delta\nu}(bary)$.

When we identify a multiplet M in a window W, we do not immediately output it because there may be a multiplet M_2, including detections with frequencies above W, that overlaps M in frequency and has a higher score; see Figure J.1.

In this case, we want to output M_2 rather than M. To do this, we maintain a *reserved multiplet* M_R that is possibly null. When we identify a new multiplet M and M_R is not null: if the frequency range of M is completely above that of M_R, we output M_R and reserve M. If the frequency ranges overlap and M_R has a higher score than M, we output M_R and advance the window beyond its maximum frequency; otherwise, we reserve M.

If we find a multiple M and there is no reserved multiplet, let A and B denote the minimum and maximum frequencies of M. If the next detection to be scanned has a frequency of at most $A + M_{\Delta\nu}(bary)$, we reserve M and move the window W to start at A (that is, remove detections from W with frequencies less than A). Otherwise, we output M, and move W to start at B.