Characterizing the Physical Properties of the HI Line at Cygnus X

Avery Books
University of California, Berkeley, Dept. of Astronomy
March 6, 2026

ABSTRACT

The hyperfine transition of neutral hydrogen (HI line) is a primary tool for looking through interstellar dust and characterizing galactic structures. It is critical to determine information such as the thermal and kinematic properties of the Interstellar Medium in the Milky Way Galaxy and the structures within. The main objective of this paper is to characterize thermal and Doppler velocity properties of Hydrogen in the Cygnus X and the line of sight local arm while looking at Cygnus X. We utilized a radio telescope capable of detecting the 1420.4 MHz emission line to obtain power spectra data. With this data we used gaussian fitting to decompose the multi-peak spectra and statistical analysis to determine Doppler and temperature quantities. We were able to determine the LSR corrected velocity of the main parcel of gas to be 12.75 km/s and to determine the readings are made of 4 distinct parcels of gas each experiencing different parameters. Our best constrained upper limit temperature measurement is $T_{upper} = 3984 \pm 706$ K, which still has an error of $\approx 17\%$ which is poorly constrained.

Contents

1. Introduction

Hydrogen is the most abundant element in the universe, and neutral hydrogen is the vast majority of baryonic mass in the galactic disk. The space between stars is not empty, but filled with neutral Hydrogen gas and that serves as potential nurseries for new stars through cooling and collapse into molecular clouds. These stars then explode and send the neutral hydrogen back into the interstellar medium in a galactic recycling process. Neutral hydrogen is cold, and unlike stars it does not emit visible light, so we can detect it when the electron's spin flips from parallel to anti-parallel relative to the proton. This spin flip releases a photon with a wavelength $\approx$ 21.1 cm. A major motivation for using the neutral hydrogen hyperfine transition line (HI line), is that radio waves at this frequency are unaffected by interstellar extinction (dust particles). So, while optical telescopes have difficulty looking through the optically thick galactic plane, HI line radio telescopes allow us to see through the Galaxy.

Using a radio telescope located on the roof of New Campbell Hall at UC Berkeley, we will target the star system of Cygnus X, as it is a massive nearby star system and should demonstrate different physics through the local galactic arm that is along the telescopes line of sight. The primary objectives of the data analysis in this paper is to determine the Doppler velocities with consideration of the Local Standard of Rest (LSR), to use this data to characterize the thermal parameters of the gas detected, and to demonstrate the characterization and error propagation of calibrations through the data and interpretation.

In §2 we introduce the data analysis techniques and theory relevant to this paper. In §3 we discuss the characterization of our radio telescope and the specific utilization of data analysis. In §4 we provide and outline the figures and data relevant to our analytical goals. Finally, in §5 we discuss the interpretation of the data, the physical quantities we can derive from it, and it's statistical analysis.

2. Background

The HI line is a difficult to detect signal, and is often masked within cosmic noise data. This requires data manipulation to extract the HI line from the raw data taken from the telescope.

2.1. Isolating the HI Line

First we utilize frequency switching to crease "on-line" and "off-line" spectra. The on-line spectrum is a long integration that has an LO (local oscillation frequency) centered at 1420 Mhz, adjacent to the rest frequency of the HI line. The off-line spectrum has the LO shifted so the line is either moved or removed from the band. When the HI line is moved in the band you can perform "in-band" frequency switching, and standard frequency switching when the off line spectrum does not contain the HI line. The purpose of this is to utilize a ratio $r_{line} = s_{on}/s_{off}$ to remove the instrumental band pass shape. However this only returns the spectra shape and removes power information from the data, as it divides out with the ratio.

Using this ratio, we can multiply by a

gain function

$$G = \frac{T_{\text{sys,cal}} - T_{\text{sys,cold}}}{\sum(s_{\text{cal}} - s_{\text{cold}})} \sum s_{\text{cold}} \quad (1)$$

to recover a power-calibrated spectrum, $T_{line} = r_{line} \times G$, in units of Brightness Temperature (Kelvin).

Now, to remove any residual instrumental noise, interference, or otherwise non-HI line information from the data, we can apply a low-order polynomial fit to the off line channels in the data and subtract the polynomial from our T_{line} spectrum. This residual noise can be in part attributed to frequency-dependent reflections in the cable connection interfaces. That is, when there is an interface between two components that do not share the exact same impedance property, there is some fraction of the signal that is transmitted and some fraction that is reflected. This reflection causes interference in the data that does not perfectly cancel out in the on and off line division leaving a baseline in the data. For example, if you plug in a perfect square wave from a function generator, and connect one end of a T junction to an oscilloscope, and the other end is not connected to anything, one will notice the square wave has interference manifesting as "steps" on the oscilloscope. This is due to the signal reflecting off the non-connected end (floating end) and going through the cable into the oscilloscope. These steps disappear if a correct impedance terminator is attached to the end of the floating cable. Another cause of this residual noise is that gain and noise characteristics of the SDR and amplifiers can drift between the time of on and off line measurements, so during division, they do not cancel perfectly and can be addressed in the same way.

Finally, we can subtract the system temperature baseline from the T_{line} spectrum to center the parts of the sky without Hydrogen at 0 Kelvin.

2.2. Thermal and Turbulent Broadening

Thermal broadening, a characteristic of the random motion of gas at some temperature T. Even without bulk movement (i.e. moving in the galactic plane), this random motion will occur. In data, this causes the ensemble of atoms at different velocities to produce a spread of observed frequencies rather than just a single sharp line at the bulk motion frequency. Since the Maxwell-Boltzmann distribution that governs this random motion is a Gaussian, the thermally broadened profile is also Gaussian and the change in velocity due to thermal broadening is

$$\Delta v_{thermal} = \sqrt{\frac{8 k_B T \ln 2}{m_H}} \quad (2)$$

and can be solved for T to get the temperature from a measured line width.

Turbulent broadening is when there is some internal bulk motions where different packages of gas within the same cloud move at different velocities relative to each other. These are not random thermal motions and can occur on a variety of scales within the cloud. Due to potential differences in velocity, the Doppler shift measured is a superposition of all the different Doppler shifts. We can add these two components as

$$\Delta v_{obs}^2 = \Delta v_{thermal}^2 + \Delta v_{turbulent}^2 \quad (3)$$

In this paper we do not resolve each component individually, so we can only report upper limits of temperature readings, as our Doppler measurements may include turbulent broadening, and we do not have a way to rule it out without decomposition.

2.3. LSR Calculation

In order for Doppler velocity data to have significance, and to allow for reproduction in future experiments, we must implement the Local Standard of Rest which accounts for the the reference frame of when the data was taken. The LSR corrects for the Earth's rotation, the Earth's orbit, and the Solar motion within the galaxy. The LSR itself is a theoretical point located at the Sun's current distance from the galactic center that follows and exactly circular orbit. We first determine a corrected velocity

$$V_{corr} = \vec{V}_{Earth} \cdot \hat{n}_{target}$$

by projecting the velocity measurement along the line of sight of the telescope. Then to get the final velocity:

$$V_{LSR} = V_{measured} + V_{corr}$$

To make the right corrections, we must first characterize the gain function for the entire system, utilizing cable length and understanding the power loss per unit length in the cable, and getting a baseline system temperature using a known temperature blackbody.

3. System Characterization and Data Collection Methods

3.1. System Characterization

To characterize the gain function we first need to determine the length of the cable system. To do this we can utilize the phase difference in the interference mentioned in §2.1. First, using the same type of cable in our observing system, we plugged a function generator into a T junction, and attached a cable to the other end of the junction and left it unattenuated. These are cables of known length, with error ± 0.5 inch such that we could use the oscilloscope to measure the phase difference duration in time to an error of ± 2 ns.Our measurements were $t_{cable} = 157.0 \pm 2$ ns and we used lengths of 26'8" and 27'2", and converted to meters to get a total system length of $L_{cable} = 16.41 \pm 0.02$ m. Then to solve for the fraction of the speed of light in the cable we use: $(2L_{cable}/t_{cable}) = (2.09 \pm 0.04)e8$ m/s. Then we swap the function generator cable for the cable directly to our radio telescope and rearrange the equation to yield $Lscope = (c_{cable}t_{scope})/2 = 65.22 \pm 1.25$ meters.

Now we can characterize the amplitude in the system by determining amplitude loss per unit length. To determine this, we measured our amplitude of signal before and after attaching our telescope cable, and converted to a loss parameter in dB using $l_{dB} = 20log(A_{init}/A_{cable}) \pm 0.001$. Then, dividing by the cable length we got a loss per meter of $0.23\,\mathrm{dB/m} \pm 0.01\,\mathrm{dB}$. Then applying this to our telescope we get a total amplitude loss of $15.31 \pm 0.73\,\mathrm{dB}$.

Finally to get a general gain function

we must determine the amplification for all system components. The system comprises, in order, of: the horn telescope, 3 pre-amplifiers (+26.8 dB) each, attenuator (−2.2 dB), the cable (−15.31(73) dB), a BPF, 2 amplifiers (+20.80(1) dB) each, and the SDR (+45 dB) by defined by our code. This gives us a total system gain of 149.66(73) dB.

Note, this system gain is $\mathcal{O}(100)$ lower than the gain calculated using Equation(1). Our cold sky zenith data and our attempted 300 K blackbody (human) calibration data is not significantly distinguishable in arbitrary power. This indicates that there may be an antenna blackbody coupling error and the data is not significant. Therefore the calibration using that technique provided an incorrect value $G \approx 17000$ whereas the calibration utilizing system characterization provides $G \approx 140$. So we will use the latter, as it provides more physically meaningful readings in §4.

3.2. Data Analysis

The data analysis section begins after we have applied the adequate noise removal, gain correction, and conversion to Kelvin to determine our T_{line} as described in §2.

To characterize the HI spectral data, a multiple component Gaussian model was fit to the isolate line profile in the ON-OFF difference spectrum. A Gaussian functional form was chosen as it is the expected line shape for optically thin HI emissions and should stay consistent for any thermal broadening. The model takes the analytical form

$$S(\nu) = \sum_{i=1}^{N} A_i \exp\left(-\frac{(\nu - \mu_i)^2}{2\sigma_i^2}\right) \quad (4)$$

where A_i is the peak amplitude of each component in units Kelvin, μ_i is the central frequency in MHz, and σ_i is the standard deviation in MHz. N was determined by inspection of the line profile and minimizing residuals.

The fit was performed using the `gaussfit` function from the `ugradio` package, which wraps a nonlinear least squares routine and accepts initial parameter guesses for the amplitude, center, and width of each Gaussian component. Multiple guesses for N needed to be individually graphed, and fit quality was assess by computing the residuals between the model and data, with a good fit having residuals consistent with the noise floor across the spectrum.

Each fitted component has three physically meaningful quantities. A_i is proportional to the column density of HI gas at that velocity. The central frequency μ_i can be converted to a light of sign velocity using the Doppler formula and subsequently used to check our other LSR corrected velocities. This tells us about the thermal broadening from §2. The following is a more detailed analytical approach.

The FWHM is extracted directly from the fitted Gaussian standard deviation σ_i returned by `gaussfit`, via:

$$\text{FWHM} = 2\sqrt{2\ln 2}\,\sigma_i \quad (5)$$

This value is in units of frequency [MHz], and is converted to a velocity width using the Doppler relation:

$$\Delta v_{\mathrm{FWHM}} = \frac{\mathrm{FWHM}}{\nu_0} \cdot c \qquad (6)$$

where $\nu_0 = 1420.406$ MHz is the HI rest frequency. The resulting Δv_{FWHM} in km/s can then be substituted directly into Equation (2) to estimate an upper limit on the kinetic temperature of the emitting gas:

$$T \leq \frac{m_H}{8 k_B \ln 2} \left(\Delta v_{\mathrm{FWHM}}\right)^2 \qquad (7)$$

where the inequality reflects the fact that turbulent broadening may contribute to the observed line width, meaning the true thermal temperature could be lower than this estimate.

The LSR correction was implemented using the `astropy.coordinates` package, and the barycentric correction was using the `SkyCoord.radial_velocity_correction` method. Julian date and observation location information was saved into the metadata via the `ugradio` package.

For Cygnus X data the galactic coordinates observed are $\approx (81, 0)$.

4. Results

The HI line from the Cygnus X region was detected through on-off beam switching with the averaged spectra for each being shown in Figure 4. A clear excess in the on-source spectrum is visible above the off source baseline starting at about 1420.3 MHz which is consistent with the HI line rest frequency of 1420.406 MHz. The on-off difference spectrum isolates the excess and reveals a broad asymmetric emission feature with a peak brightness of 10.60 ± 0.08 Kelvin, demonstrating a possible detection of neutral hydrogen with a signal to noise ratio of 172. Table 1 is a table of relevant parameters extracted from the data, including LSR corrected Doppler velocities and peak frequency.

Parameter	**Value**
HI Rest Frequency	1420.406 MHz
Peak Frequency	1420.45 MHz
Δf (Obs to Rest)	42.73 kHz
Doppler Velocity	-9.02 km/s
LSR Correction	21.77 km/s
LSR Velocity	12.75 km/s

Table 1: Doppler and LSR velocity parameters derived from the peak of the ON-OFF difference spectrum toward Cygnus X.

Parameter	**Value**
T_{sys}	35.57 K
$\delta\nu$	488.28 Hz
t_{on}	22.42 s
t_{off}	22.32 s
σ_{on}	0.3400 K
σ_{off}	0.3408 K
σ_{diff}	0.4813 K
χ^2	2554.96
DOF	1422
χ^2_ν	1.7967

Table 2: Radiometer noise and chi-squared fit parameters for Cygnus X HI observation.

To verify if the detection is the HI line, we must demonstrate a Gaussian fit. This will also assist in verification of the velocity structure. The fit and it's individual components are shown in Figure 4, with relevant parameters summarized in the Table 2.

The total fit reproduces the observed profile with a reduced chi-squared of $\chi^2_\nu =$

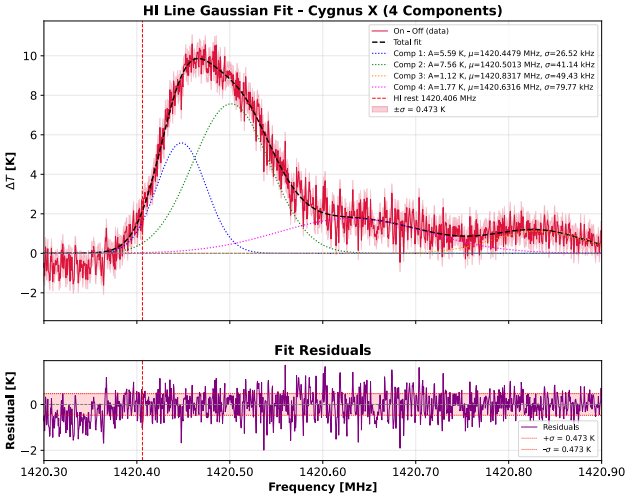

Fig. 1.— *Top:* The ON-OFF difference spectrum (red) with the total four-component Gaussian fit (black dashed) and individual components shown in blue, green, orange, and magenta. *Bottom:* Residuals after subtracting the total fit from the data, with the red $\pm\sigma$ noise envelope derived from the radiometer equation: ($\sigma_{\rm diff} = 0.481$ K). The residuals are broadly consistent with the thermal noise floor. Fit quality parameters are summarized in the accompanying table, with a reduced chi-squared of $\chi^2_\nu = 1.797$.

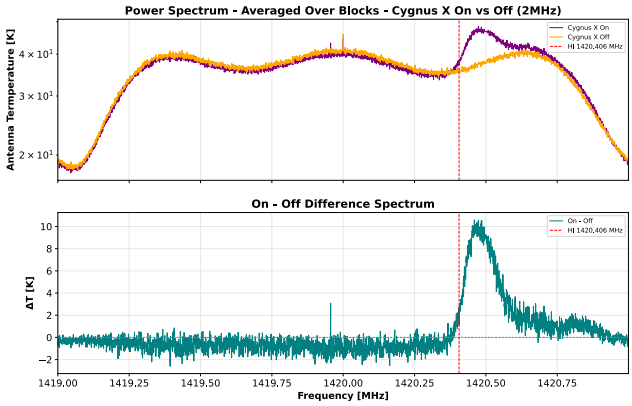

Fig. 2.— Averaged power spectra for the on-source (purple) and off-source (orange). The dashed red line indicates the HI rest frequency at 1420.406 MHz. *Bottom:* The ON-OFF difference spectrum $\Delta T = T_{\rm on} - T_{\rm off}$, isolating the excess emission from Cygnus X above the off-source background. A broad asymmetric emission feature is visible above the noise floor beginning at approximately 1420.30 MHz. All spectra are calibrated to units of antenna temperature [K] using a gain of $G = 149.66$ K/unit.

1.797, indicating a good but not perfect fit. The residuals are mostly consistent with the thermal noise floor of 0.481 K and does not demonstrate a systematic structure suggesting this four component model captures the HI line effectively. The four component span a frequency range of 1420.35 to 1420.90 MHZ, corresponding to a total velocity of -104.3 to 11.8 km/s, which reflects a superposition of multiple HI clouds and spiral arm features. The dominant component, centered at 1420.5 MHz accounts for a majority of the flux, while the remaining three components describe more broad asymmetries at higher LSR velocities.

In table §4, there are listed error values that propagate through the measurements and will be further discussed in §5. From the FWHM data from each gaussian fit, we are able to recover some upper limits on the thermal characteristics and determine which thermal state the gas is in, noted in §4. All four components are within the intermediate to warm neutral medium regime, and none are in the cold neutral medium regime.

Parameter	Uncertainty
G	149.66 ± 0.73 K/unit (0.49%)
$T_{\rm sys}$	35.57 ± 0.17 K
$\sigma_{\rm diff}$ (stat)	0.4728 K (radiometer equation)
$\sigma_{\rm diff}$ (sys)	0.0067 K (gain systematic)
$\sigma_{\rm diff}$ (total)	0.4728 K (combined in quadrature)
$T_{\rm peak}$	10.60 ± 0.47 K

Table 3: Summary error budget for the Cygnus X HI 21cm observation. The total uncertainty on the difference spectrum is dominated by thermal noise, with the gain systematic contributing less than 2% of the total error.

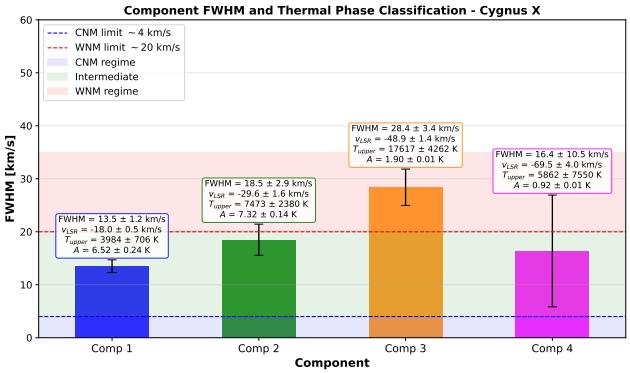

Fig. 3.— FWHM and thermal phase classification for each Gaussian component toward Cygnus X. Horizontal dashed lines mark the CNM (∼4 km/s) and WNM (∼20 km/s) velocity boundaries. All four components fall in the intermediate to WNM regime, with annotated values showing the fitted amplitude, LSR velocity, FWHM, and kinetic temperature upper limit with propagated uncertainties.

5. Analysis and Discussion

5.1. Doppler and Data Shape Interpretation

We can confirm the detection of the HI line toward Cygnus X through the clear excess in brightness temperature of 10.60 ± 0.08 K and SNR of 172, that exhibits a multiple component Gaussian construction. The asymmetric profile characterized by the blue-shifted rise and a slight red-shifted tail is consistent with the superposition of multiple HI clouds at distinct LSR velocities along the line of sight through the region. The dominant component at $v_{LSR} = 12.75$ km/s is consistent with emission from local spiral arm gas, as emission within $v_{LSR}| < 20$ km/s along this line of sight is generally attributed to the local Orion-Cygnus arm (1). The smaller redshifted components are possibly associated with individual parcels of gas moving separately to the blueshifted parcels.

$$T \leq \frac{m_H}{8k_B \ln 2} \left(\Delta v_{\rm FWHM}\right)^2 \approx 3984 \pm 706K \tag{8}$$

Utilizing this equation we calculate the values from §4. With the most constrained temperature parcel measured at 3984±706 K

Comp.	$T_{\rm upper}$ [K]	Phase
1	3984 ± 706 K	Intermediate
2	7472 ± 2379 K	WNM
3	17616 ± 4262 K	WNM
4	5862 ± 7550 K	Intermediate
Most constrained: Comp. 1, $T_{upper} = 3984 \pm 706$ K		

Table 4: Kinetic temperature upper limits derived from the thermal broadening equation for each Gaussian component toward Cygnus X. Note that only components 1 and 3 are reasonably constrained. Phase classifications follow CNM ($T < 200$ K), intermediate ($200 < T < 5000$ K), and WNM ($T > 5000$ K) boundaries.

The values from the FWHM are 13.5 ± 1.2 km/s, 18.5±2.9 km/s, 28.4±3.4 km/s,

and 16.4 ± 10.5 km/s and exist primarily in the intermediate zone, with the third component being inside the warm neutral medium zone within error.

However, the error in these temperature measurements are too high to consider the values well constrained. The first component with the lowest uncertainty has 17.7% uncertainty, with the second lowest being component 3 with 31.8% uncertainty. This demonstrates that we cannot attribute these measurements to be statistically meaningful and likely have a systematic error propagating through the calculations. It is of note though, that these values are within a reasonable range of temperatures for the objective characteristics, so while they are unconstrained and not statistically significant, they are at the minimum physically reasonable.

5.2. Chi Squared and Error Propagation

The evaluation of $\chi^2_\nu = 1.797$ in the Cygnus X data is slightly above unity, indicating that while the model captures the dominant structure, there are some systematic residuals that are not characterized by the model. There could be unresolved velocity substructures in the individual components of the fit that are not distinguishable at the measured spectral resolution. The noise estimate in the chi-squared calculation was derived from the radiometer equation, which assumes purely thermal noise as $\sigma_{diff} = 0.481$ K. So, any correlated noise or baseline structure not captured by this model would inflate χ^2_ν above unity.

An F-test indicated marginal statistical justification for a fifth component ($p = 0.027$). However, the additional component, and subsequent adjustments to other components were poorly constrained with fractional uncertainties that exceeded 100% and therefore is not considered physically meaningful. The four-component model is adopted in this paper as it provides a more physically interpretable description of the data.

The total uncertainty in the measured antenna temperatures arises from two independent sources: statistical noise characterized by the radiometer equation, and systematic uncertainty from the gain calibration. These are propagated separately and combined in quadrature to yield the total error budget summarized in Table 4. The statistical noise on each averaged power spectrum is given by the radiometer equation:

$$\sigma = \frac{T_{\rm sys}}{\sqrt{\Delta\nu \cdot t}} \tag{9}$$

where $T_{sys} = 35.57$ K is the system temperature estimated from the mean of the off-source spectrum, $\Delta\nu = 488.28$ Hz is the frequency resolution per bin, and t is the total integration time. This yields $\sigma_{\rm on} = 0.340$ K and $\sigma_{\rm off} = 0.341$ K for the on and off source spectra respectively. Since each spectra are independent measurements, their uncertainties add in quadrature to give the noise on the difference spectrum:

$$\sigma_{\rm diff} = \sqrt{\sigma^2_{\rm on} + \sigma^2_{\rm off}} = 0.473 \text{ K} \tag{10}$$

The systematic uncertainty from the gain calibration propagates multiplicatively into all absolute temperature measurements. For a quantity $T = P/G$, the fractional gain uncertainty gives:

$$\frac{\sigma_T}{T} = \frac{\sigma_G}{G} = \frac{0.73}{149.66} = 0.49\% \qquad (11)$$

This contributes $\sigma_{sys} = 0.007$ K to the difference spectrum, which is negligible compared to the statistical noise. The total uncertainty on the difference spectrum is therefore:

$$\sigma_{\text{total}} = \sqrt{\sigma_{\text{diff}}^2 + \sigma_{\text{sys}}^2} = 0.473 \text{ K} \qquad (12)$$

confirming that the error budget is dominated entirely by thermal noise. It is important to note that the gain systematic is a correlated error that scales all temperature measurements by the same factor, and therefore does not affect frequency-derived quantities such as v_{LSR}, FWHM, or chi^2_ν.

The uncertainties on the individual Gaussian fit parameters were propagated from the covariance matrix returned by `gaussfit`, with the gain systematic added in quadrature to the amplitude uncertainties. Of the four fitted components, only component 1 is well constrained with fractional amplitude uncertainties of 53%, which is only marginally acceptable. Components 2, 3, and 4 have statistical uncertainties exceeding their fitted amplitudes and are not individually significant at the 1σ level, though they contribute to the overall fit quality as assessed by the reduced chi-squared $\chi^2_\nu = 1.797$.

6. Conclusion

This experiment successfully detected HI 21cm emission toward Cygnus X using a software-defined radio telescope, confirming the presence of neutral hydrogen along the line of sight with a peak brightness temperature of $T_{\text{peak}} = 10.6 \pm 0.08$ K. The asymmetric line profile was decomposed into four Gaussian components, adopted as the most parsimonious physically meaningful model following F-testing, which rejected a fifth component despite marginal statistical justification ($p = 0.027$) due to unconstrained fit parameters. The reduced chi-squared of $\chi^2_\nu = 1.797$ indicates good but imperfect fit quality, with residuals attributed to unresolved velocity substructure.

The dominant component at $v_{\text{LSR}} = 12.75$ km/s confirms emission from local spiral arm HI, with remaining components consistent with more distant Perseus and outer Cygnus arm gas. FWHM-derived temperature upper limits place the narrowest component (13.50 ± 1.22 km/s) in the intermediate regime ($T_{upper} = 3984 \pm 706$ K). However, error in the temperature readings were statistically insignificant but within a realm of physical reason. So further calibration and observation is required to confirm data and derivations along the Cygnus X line of sight.

Acknowledgements

I would like to acknowledge everyone from Team RTFM for help in data gathering and code troubleshooting. Specifically, I would like to acknowledge Evan Wille for help in the system characterization and cable parameter calculations such as attenuation per foot. Additional, acknowledgments to Ben Jacobson-Bell for debugging help and to Prof Aaron Parsons for theory instruction and lab manual instructions.

This paper utilized python packages for data analysis including `numpy`, `ugradio`,

`scipy`, `astropy`, `matplotlib`, `rafpy` (which is my own package).

Additionally, Claude AI model was used for code efficiency debugging. Particular data files had 10,000 blocks with 4096 samples each. So, after I wrote data analysis code, if the algorithm took upwards of 2 minutes, I utilized Claude to identify and rewrite parts of code that had significant runtimes.

REFERENCES

P. M. W. Kalberla, W. B. Burton, D. Hartmann, E. M. Arnal, E. Bajaja, R. Morras, and W. G. L. Pöppel. The Leiden/Argentine/Bonn (LAB) survey of galactic HI. final data release of the combined LDS and IAR surveys with improved stray-radiation corrections. *Astronomy & Astrophysics*, 440:775–782, 2005.

This 2-column preprint was prepared with the AAS LaTeX macros v5.2.