ARTICLES
https://doi.org/10.1038/s41550-020-01261-4

nature astronomy

Check for updates

An investigation of spectral line stacking techniques and application to the detection of $HC_{11}N$

Ryan A. Loomis[1]✉, Andrew M. Burkhardt[2], Christopher N. Shingledecker[3,4,5], Steven B. Charnley[6], Martin A. Cordiner[6,7], Eric Herbst[8,9], Sergei Kalenskii[10], Kin Long Kelvin Lee[2], Eric R. Willis[8], Ci Xue[8], Anthony J. Remijan[2], Michael C. McCarthy[2] and Brett A. McGuire[1,2,11]✉

As the inventory of interstellar molecules continues to grow, the gulf between small species, whose individual rotational lines can be observed with radio telescopes, and large ones, such as polycyclic aromatic hydrocarbons best studied in bulk via infrared and optical observations, is slowly being bridged. Understanding the connection between these two molecular reservoirs is critical to understanding the interstellar carbon cycle, but will require pushing the boundaries of how far we can probe molecular complexity while still retaining observational specificity. Towards this end, we present a method for detecting and characterizing new molecular species in single-dish observations towards sources with sparse line spectra. We have applied this method to data from the ongoing GOTHAM (GBT Observations of TMC-1: Hunting Aromatic Molecules) Green Bank Telescope large programme, discovering six new interstellar species. Here we highlight the detection of $HC_{11}N$, the largest cyanopolyyne in the interstellar medium.

As molecules increase in size, detection by rotational spectroscopy generally becomes more challenging. In large molecules, there are a substantially larger number of rotational energy levels over which the population is distributed, reducing the emission between any two that give rise to an observable transition. The rotational partition function for such species can be high even at low tempeartures, with a large number of thermally populated rotational levels, diluting the intensity of any given transition. Moreover, larger species are generally less abundant than smaller species[1]. Taken together, it is often far more difficult to detect individual rotational lines of a heavy species relative to those of a light species, even if both have identical dipole moments, rotational temperatures and column densities. Even for small polycyclic aromatic hydrocarbons (PAHs), for example, the total line intensity is diluted over potentially hundreds if not thousands of transitions, making it exceedingly difficult to detect any individual line in a reasonable amount of integration time.

Here, we describe a new method that combines the techniques of Markov chain Monte Carlo (MCMC) inference with spectral line stacking and matched filtering to counteract the effects of rotational dilution, improving detection efficiency and the characterization of weak emission from large molecules.

Molecular detection technique

MCMC inference has grown in popularity in recent years in the astrochemical community as a tool for analysing the properties of spectroscopic lines[2–4], allowing for straightforward characterization of parameter uncertainties and covariances. Similarly, line stacking and matched filtering techniques have regularly been applied to improve the signal-to-noise ratio (SNR) and detection efficiency of weak lines[5–7]. Here, we present a hybrid combination of these techniques to robustly infer the presence of large astronomical molecules of interest in single-dish spectra, as well as their emission parameters and associated uncertainties. In particular, this technique is ideal for identifying and characterizing species when no individual line is intense enough to be observed in a spectral line survey, but where many lines are present in the data itself, hidden under the noise. A flowchart providing an overview of our analysis method is shown in Fig. 1, and we explain each step of the process in the following subsections.

The GOTHAM dataset. Our method is best suited to a line-sparse single-dimensional spectral dataset, and here we investigate its application to data from the ongoing Green Bank Telescope (GBT) large programme GOTHAM. The details of these observations are presented in ref. [8]. In short, at the time of this analysis, the observations were ~30% complete, covering 13.1 GHz of the total bandwidth between 7.8 and 29.9 GHz. With a frequency resolution of 1.4 kHz (0.014–0.054 km s^{-1}), the dataset encompasses 9.3 million channels.

Despite the wide spectral range, the observations are relatively line-sparse. A total of 632 lines are detected above 5σ, yielding an effective average line density of 0.05 lines per MHz (one line every 20 MHz). The lines are also relatively narrow: ~0.3 km s^{-1} in aggregate,

[1]National Radio Astronomy Observatory, Charlottesville, VA, USA. [2]Center for Astrophysics, Harvard & Smithsonian, Cambridge, MA, USA. [3]Department of Physics and Astronomy, Benedictine College, Atchison, KS, USA. [4]Center for Astrochemical Studies, Max Planck Institute for Extraterrestrial Physics, Garching, Germany. [5]Institute for Theoretical Chemistry, University of Stuttgart, Stuttgart, Germany. [6]Astrochemistry Laboratory and the Goddard Center for Astrobiology, NASA Goddard Space Flight Center, Greenbelt, MD, USA. [7]Institute for Astrophysics and Computational Sciences, The Catholic University of America, Washington, DC, USA. [8]Department of Chemistry, University of Virginia, Charlottesville, VA, USA. [9]Department of Astronomy, University of Virginia, Charlottesville, VA, USA. [10]Astro Space Center, Lebedev Physical Institute, Russian Academy of Sciences, Moscow, Russia. [11]Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA, USA. ✉e-mail: rloomis@nrao.edu; brettmc@mit.edu

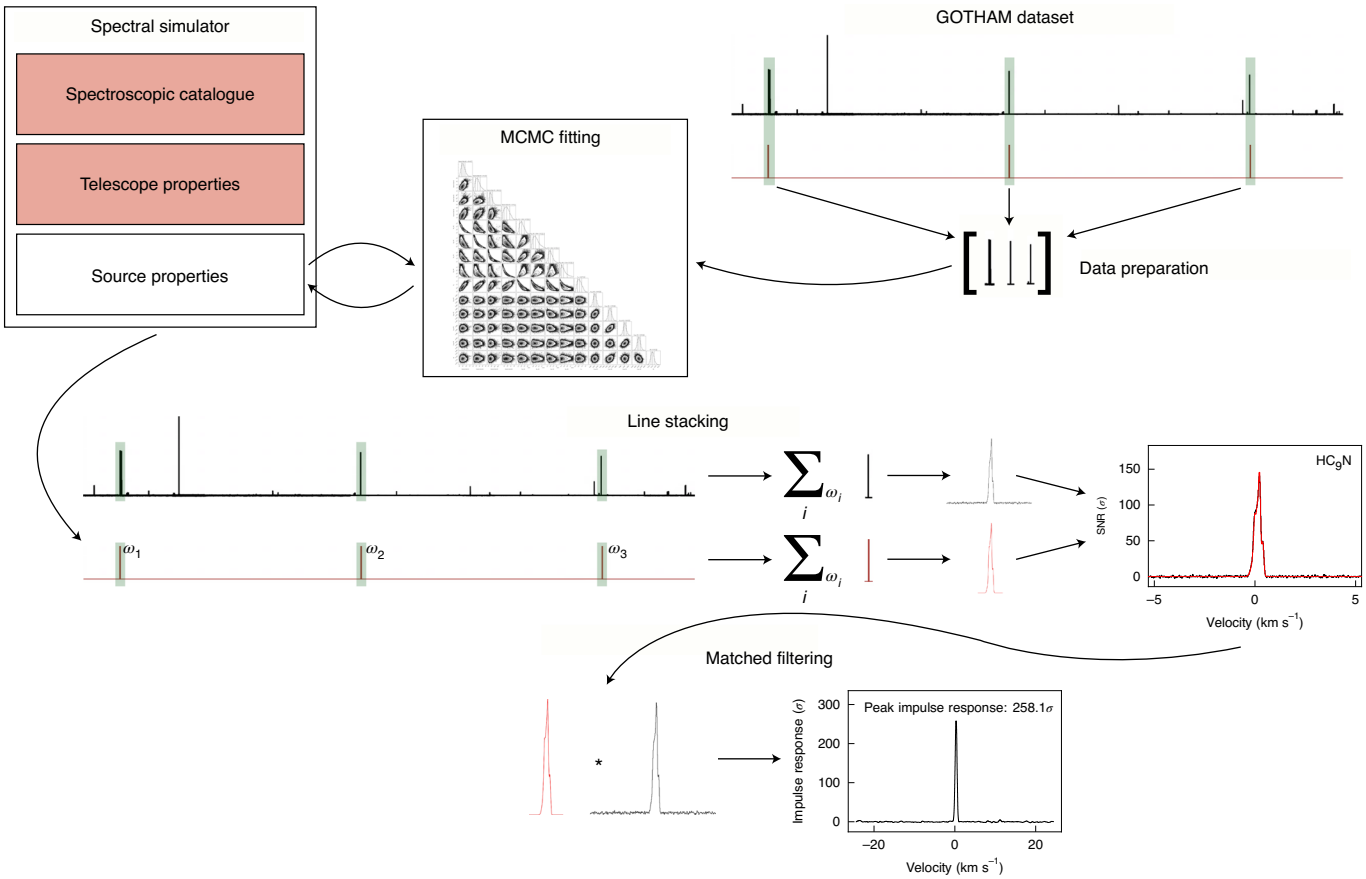


Fig. 1 | Schematic diagram of our method for molecular detection and characterization. In short, the GOTHAM dataset and an initial spectral simulation are used to select a relevant subsection of data (green shaded regions). A model is then fitted to the data and the source properties varied while the telescope properties and spectral catalogue are held fixed (shaded red). The best-fit model is used to weight the data for stacking (for example, ω_1, ω_2, ω_3, ... in the figure). To visualize the statistical significance of this detection, the stacked model is used as a matched filter and cross-correlated with the stacked data.

although we fit the contributions of several (2–4) ~0.11 km s^{-1} components to these features. The result is a spectrum that is sparse in 'bright' channels: only one channel in every ~1,400 is >5σ above the local noise level, which is equivalent to a filling factor of <0.1%. We discuss the importance of the line sparsity in more detail later.

Spectral simulator. To infer the desired astrophysical properties (for example, excitation temperature and column density) of a given molecular species, we employ a forward modelling framework where spectra are iteratively simulated in a fashion similar to the observations themselves and then compared to the data. Our spectral simulator is based on the basic equations of molecular excitation and radiative transfer[9–11]. The simulator has three main inputs: a spectroscopic catalogue in SPCAT format from the CALPGM suite of programmes[12,13], a collection of telescope properties and a collection of source properties.

The most critical telescope property is the 100 m dish size of the GBT, required for calculating an effective beam filling factor to account for beam dilution effects. Source properties for each source component are left as free parameters, and include the effective source size (used for calculating filling factors and assumed to be a symmetric Gaussian), column density ($N_{\rm col}$), excitation temperature ($T_{\rm ex}$), source velocity ($v_{\rm LSRK}$) and linewidth (dv).

In our modelling of TMC-1, we have found four distinct velocity components at similar velocities to those previously identified[14,15] from which the majority of species emit from. Source size, column density and source velocity were allowed to vary freely for each component, and the excitation temperature and linewidth are fit jointly across components. It is likely that the excitation temperature and linewidth do vary slightly across the different cloud components, but our data are not sufficient to constrain these differences, which we discuss in more detail later. Several species in our analysis were best fitted by utilizing only a subset of three of these four cloud components, and their results are presented with a corresponding number of free parameters.

The spatial orientation of the four cloud components on the sky (Fig. 2) has a pronounced impact on how their emission is measured by the telescope. First, since our dataset is only from a single pointing position and we do not have spatial information about these cloud components, we make the simplifying assumption that each component is centred in the beam. The Gaussian full-width at half-maximum (FWHM) of the beam for the GBT (in arcseconds) is calculated for a given wavelength λ and dish size D as:

$$\theta_{\rm bm} = \frac{206{,}265 \times 1.22\lambda}{D} \quad (1)$$

as documented in the GBT Proposer's Guide[16] and the corresponding beam dilution factor for a Gaussian source centred in the beam with FWHM $\theta_{\rm source}$ is:

$$\frac{\theta^2_{\rm source}}{\theta^2_{\rm bm} + \theta^2_{\rm source}}. \quad (2)$$

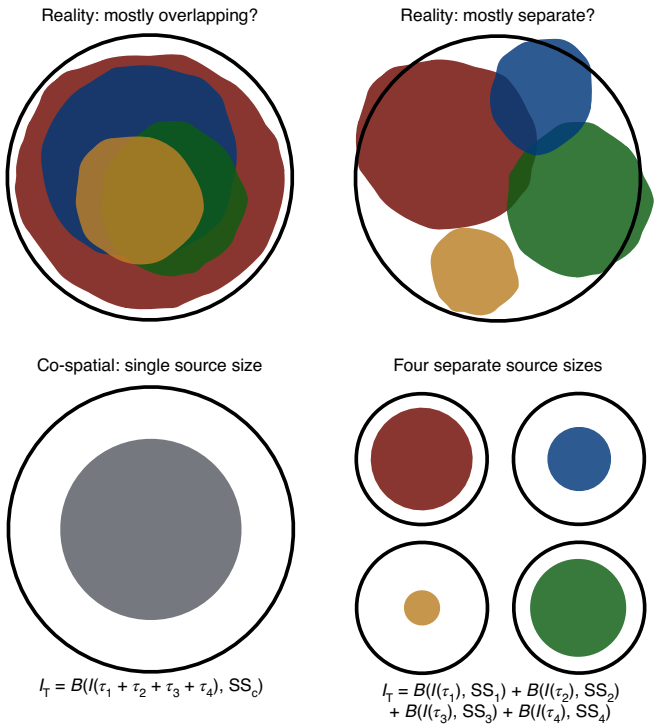


Fig. 2 | Schematic showing two spatial distribution regimes into which the emission from TMC-1 may fall and the approximations we use in our analysis. The FWHM primary beam of the GBT is denoted by the black circle. Emission may be mainly co-spatial, with substantial overlap between velocity components (top left) or originate from spatially distinct velocity components, which are all still mostly within the primary beam (top right). Optical depths (τ) in the co-spatial approximation (bottom left) are added linearly before converting to intensity (I), and a common source size (SS) is fit and applied to account for beam dilution ($B(I)$). In the separate components approximation (bottom right), each component is separately beam diluted and then these intensities are added linearly to calculate the total intensity I_T.

This beam dilution factor is applied at each frequency in the spectrum to all cloud components on the basis of their given source sizes. In reality, the sources may be unequally distributed throughout the beam, leading to varying beam dilution effects at different frequencies, as the source begins to exit the beam. We discuss this point in more detail later.

In the optically thin limit, the spatial distribution of components does not strongly impact how their emission co-adds. Thus, for species firmly within the optically thin limit, a beam diluted spectrum can be generated for each component and then summed. For species that may have lines that are more optically thick, however, the spatial distribution may have a more pronounced effect on how the emission co-adds.

If two optically thick lines lie at different velocities, and the linewidths are smaller than the separation between the central velocities of the components, then the components are radiative decoupled and can be added as in the optically thin case. This is the main assumption of the large velocity gradient approximation. Additionally, if two optically thick components are spatially distinct, they will add linearly in measured intensity. If there are two co-spatial optically thick lines that overlap in velocity, however, they need to be added in τ space before converting to intensity. We refer to these limiting cases as 'separate components' and 'co-spatial'. As we lack the spatial information to disentangle the more complicated (and more likely) scenario of a situation between these two limiting cases, we instead present results from the two limits and discuss both when relevant. For the co-spatial case, it makes more sense to fit a common source size across components (Fig. 2, bottom left), so the total number of model parameters is shrunk by three. As shown in the Supplementary Information and discussed in more detail later, a co-spatial model does a much better job of describing the smaller and more optically thick cyanopolyynes.

Initial data preparation. To begin the fitting process, it is first necessary to reduce the size of the dataset that will be simulated, as generating 9.3 million channels in every step of the MCMC process would not be computationally tractable. A dataset of much more manageable size would consist of only the small number of channels that are near lines of interest for a given species.

Using our spectral simulator and a nominal set of telescope and source properties, we generate an initial simulation for the target species across the full bandwidth of the GOTHAM observations. A dish size of 100 m, source size of 100″, excitation temperature of 8 K, column density of 10^{12} cm^{-2} and linewidth of 0.37 km s^{-1} are assumed. As this initial simulation is used only to select the regions of the spectrum to perform the fit on, relative line strengths need only be approximate, and knowing the exact source size, excitation temperature, column density or linewidth is not necessary. The linewidth and excitation temperature are estimated on the basis of previous observations of TMC-1[14,15,17], with the linewidth being large enough to encompass all of the known multiple velocity components.

Nominally, the method will work when including all lines in a catalogue file that fall within the range of the observations. For this work all lines were used with simple linear species, but for the analysis of species such as 2-cyanonaphthalene where there are thousands of extremely weak lines, applying a threshold substantially improves the computational efficiency. In these cases a threshold of 5% of the peak intensity in the initial simulation was used, discarding all lines below this threshold as they will not contribute substantially to the final fit or stacked detection. For each remaining line, a window was generated at 5.8 ± 0.5 km s^{-1} and applied to the GOTHAM spectrum, yielding a final sparse spectrum with a much smaller datasize, as shown in Fig. 1. Within each window, a local estimate of the noise was taken by calculating the standard deviation of all points less than 3.5σ (where σ is an initial standard deviation taken considering all points). This method reduces the impact of any strong lines on the estimate of the local noise. For the analysis of weaker species, a 6σ threshold was then applied to block any interloping lines from other species, preventing them from contaminating the model fit or final stack. Interloping lines were removed from the windowed dataset.

The final output of this procedure is a small, sparse spectrum for each species being considered, as well as a noise spectrum of identical dimensionality.

MCMC fitting. With a reasonably sized dataset now available for a given species, we then utilize an MCMC fitting method to derive posterior probability distributions and covariances for each free parameter in the spectral simulator model. This process is very similar to that described in ref. [4].

The degrees of freedom for each model are set by the considerations described earlier, with a maximum of 14 free parameters. The affine-invariant MCMC implementation emcee[18] was used with 100 walkers run for up to 10,000 steps. Convergence was assessed using a Gelman–Rubin convergence diagnostic[19].

Parameter initialization and priors were determined using two well-characterized 'template' species. While initially investigating the properties of species within the GOTHAM data, we found that, as one might expect from chemical intuition, linear species seemed

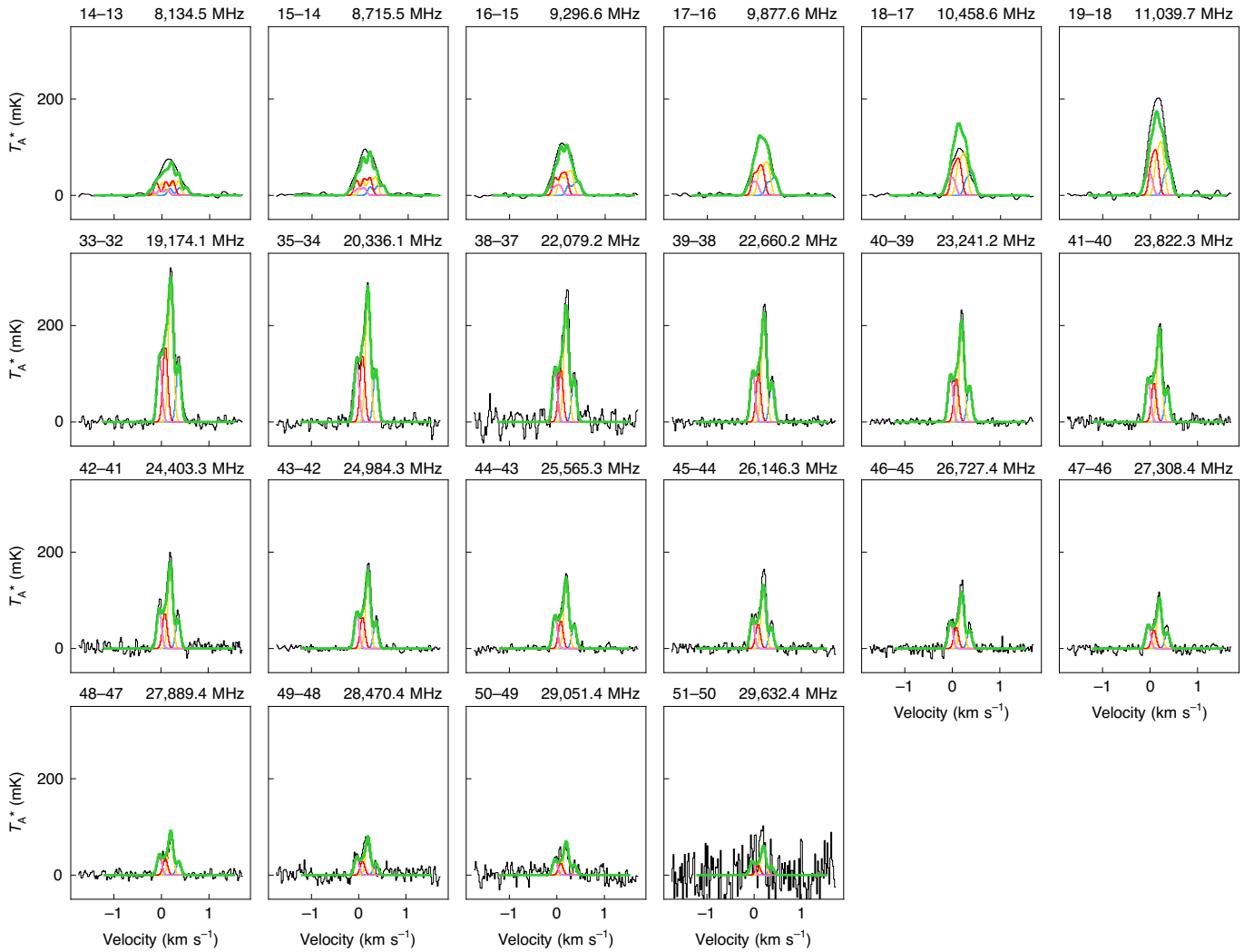


Fig. 3 | Individual line detections of HC_9N in the GOTHAM data. The spectra (black) are displayed in antenna temperature (T_A^*) vs velocity relative to 5.8 km s^{-1}, using the rest frequencies given in the top right of each panel. Quantum numbers are given in the top left of each panel, neglecting hyperfine splitting. The best-fit model to the data, including all velocity components, is overlaid in green. Simulated spectra of the individual velocity components are shown in blue (5.63 km s^{-1}), yellow (5.79 km s^{-1}), red (5.91 km s^{-1}) and violet (6.03 km s^{-1}). See Extended Data Fig. 2.

to share source properties with HC_9N, whereas cyclic species seemed to share source properties more similar to benzonitrile. These two species, which both have easily identified bright individual lines, were therefore fitted first with very simple priors—source velocities were forced to be in a sequential order and all other values had physical bounds set on them (for example, positivity constraints). An example corner plot of the HC_9N fit is shown in Extended Data Fig. 1.

The quality of these fits was then assessed visually, ensuring the suitability of the model for the data. As seen in Fig. 3, these nominal fit parameters reproduced all observed lines within uncertainties. The HC_9N and benzonitrile posteriors were then used as priors for their respective template families for all values other than column density, and 50th percentile values were used to initialize walkers in a tight ball. Column densities were initialized via quick maximum likelihood fits, holding the other initialized values fixed.

From these fits for each species, we report parameters and their uncertainties using 16th, 50th and 84th percentile intervals (for example, Extended Data Fig. 2 for HC_9N). These intervals are also denoted in the corner plots (for example, Extended Data Fig. 1). The 50th percentile values are used for all stacking analyses.

Line stacking. The posterior probability distributions from the MCMC fitting describe the range of parameter values consistent with the data, but are predicated on the assumption that our model does a good job of describing the underlying data. This is easily justified when individual lines can be detected and compared to the model predictions (Fig. 3), but is less easy to visualize when individual lines are not seen above the noise level. Calculating a detection significance is therefore crucial to interpreting the MCMC constraints. To provide a visually intuitive interpretation of detection significance, we break this process down into two steps. First, we stack all of the windowed lines that have no interlopers, and second, we apply the stacked best-fit simulation as a matched filter to the data stack.

The application of line stacking techniques to increase the SNR in spectroscopic data is a well-known technique, particularly in an astrochemical context for the detection of new species[4,5,20]. Here we follow the normal prescription of SNR weighted stacking of each line (Fig. 1), but with a minor modification for some species. When a species has a more complex spectrum where transitions are not always well separated (for example, closely spaced hyperfine components), a naive stack of every transition will overcount the

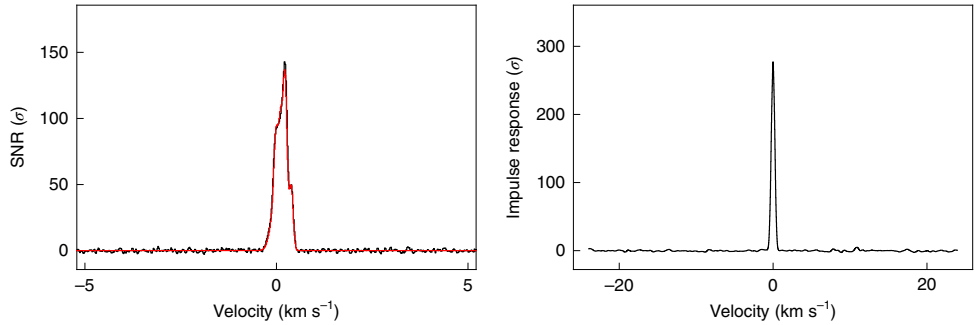


Fig. 4 | HC_9N spectra in TMC-1. Left: velocity-stacked spectra of HC_9N in black, with the corresponding stack of the simulation using the best-fit parameters to the individual lines in red. The data have been uniformly sampled to a resolution of 0.02 km s^{-1}. The intensity scale is the SNR of the spectrum at any given velocity. Right: impulse response function of the stacked spectrum using the simulated line profile as a matched filter. The intensity scale is the SNR of the response function when centred at a given velocity. The peak of the impulse response function provides a minimum significance for the detection of 277.3σ.

contributions from other nearby transitions, and may also contaminate the signal-free noise regions of the stack with signal from these nearby lines. To avoid these issues, we treat groups of transitions that are blended or closely spaced (typically <3 FWHM) as a single spectral line feature. This has the effect of slightly blurring the contribution to the total line stack, but avoids any overcounting. As the stacking procedure is performed identically for both the data spectrum and the predicted spectrum, the full signal is recovered during the matched filtering stage.

An example of this line stacking for the HC_9N lines in Fig. 3 is shown in Fig. 4. Even though each of the individual lines were strongly detected, the overall significance of the detection is greatly enhanced, now with a peak value of ~140σ. A similar stack of our best-fit model is overlaid in red, illustrating the quality of the fit. Demonstrations of the robustness of this line stacking method for our dataset are shown in the Supplementary Information. We discuss its limitations later, particularly with respect to source line density.

Matched filtering. As described in Loomis et al.[6], the technique of matched filtering first presented by Woodward[21] and North[22] can be used on astronomical spectroscopic data to optimally extract a detection significance when the shape of the signal is known. In our case, the stacked line signal still retains velocity structure, as seen in Fig. 4, and is thus not yet the maximum SNR attainable.

As shown in Fig. 1, we select a narrow region around the stacked predicted spectrum to use as the template filter, and then cross-correlate this filter with the stacked data spectrum, yielding an impulse response spectrum. The spectrum is then normalized by calculating the standard deviation of the spectrum (excluding the central region where we expect to see a signal) and dividing by this standard deviation[6]. The units of the impulse response are now σ, rather than a flux unit, and describe the SNR of the response. The peak response can therefore be thought of as a minimum detection significance for the species. An example of this impulse response spectrum for HC_9N is shown in Fig. 4, where the peak detection significance is now almost doubled, at 258.1σ.

With a better model and hence a better matching filter, the significance of the detection could be improved, but it cannot be lower than the current peak response. We discuss this point in more detail later, along with an exploration of the effects of spectroscopic catalogue accuracy on the recovered detection significance.

Upper limits. In cases where our matched filtering analysis yields an impulse response with a significance not large enough to claim a detection (for example <4σ), we refit the data using a modified MCMC process to yield more useful posteriors on the column densities. Instead of letting all of the parameters described above vary freely, we instead fix the source sizes, velocities and excitation temperatures to the values reported for a similar molecule, as was done for the priors described earlier (for example, HC_9N for linear species and benzonitrile for cyclic species). From the resultant posterior distributions, 95th percentile confidence interval values are reported as 2σ upper limit column densities. An example upper limit posterior is shown in Extended Data Fig. 3 for $HC_{13}N$, which we do not currently detect above a 4σ significance in the GOTHAM data.

Broader applicability and limitations of method. Both the MCMC fitting and stacking analysis presented here are predicated on the assumption that signal (that is, coherent information content) within the windowed data being fitted or stacked is dominated by species of interest, rather than some red noise or contribution from competing species. In the context of well-calibrated single-dish spectra, this can be more simply stated as a requirement of line sparsity. Analysis of interferometric data with this technique is possible, but beyond the scope of this paper. The degree of line sparsity necessary for a given analysis will be different for each species of interest. As discussed earlier, thresholding data is able to prevent the most egregious interloping lines from contaminating an analysis, but low-level line confusion would prevent successful stacking of the thousands of lines necessary to detect a species such as 2-cyanonaphthalene. In contrast, the several dozen lines of $HC_{11}N$ would be more tolerant to a low level of line confusion (as each individual line of interest would be brighter in comparison with the confusing lines).

Of the handful of astronomical sources that have yielded the vast majority of new interstellar molecular detections, TMC-1 has by far the most sparse spectra. Application of our technique to other sources, such as Sgr B2(N), IRAS 16293-2422 or Orion KL, is probably not as straightforward due to their higher line density. A more fruitful approach may be to take inspiration from other solutions to the analogous problems of detrending and source separation, where advancements in Bayesian methods such as probabilistic cataloguing[23] hold promise for the bulk analysis of large datasets[24].

Finally, thus far we have made the assumption that the spectroscopic catalogues used in our spectral simulator are a fixed input, with no error. In reality, few large species of astronomical interest have precise laboratory constraints on their spectra, and several of our newly detected species in GOTHAM required substantial refinement via new laboratory spectroscopic investigation[8,25]. To better understand the sensitivity of our stacking method to spectroscopic errors, we systematically introduced increasing amounts

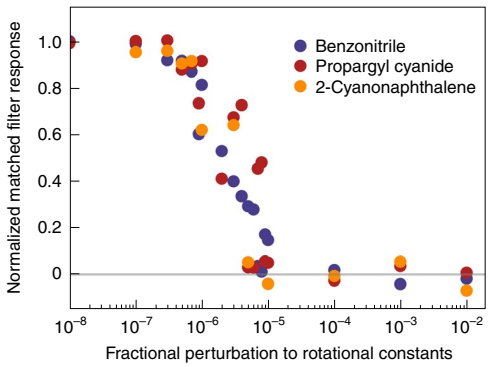


Fig. 5 | Fractional modification to rotational constants plotted versus normalized matched filter response. Three species are shown: benzonitrile, propargyl cyanide and 2-cyanonaphthalene. A relative precision of ~100 ppb allows recovery of most of the signal.

of Gaussian random noise to the rotational constants used to generate the catalogues for benzonitrile, propargyl cyanide[8], and 2-cyanonaphthalene. A plot of the fractional level of modification to the rotational constants versus the fractional peak filter response (normalized to the peak filter response for the nominal catalogue) is shown in Fig. 5. We find that for all three species, a modification of ten parts per million (ppm) is sufficient to effectively nullify the molecular detection. A relative precision of ~100 parts per billion (ppb) is sufficient to recover most of the signal. This is roughly equivalent to the accuracy of a state-of-the-art high-resolution microwave spectrometer[26], highlighting the necessity of modern laboratory constraints for the identification of large molecules in the interstellar medium.

This analysis also doubles as evidence that our stacking method is not likely to yield false positives given the line sparsity of TMC-1: a small change of a few parts per million to rotational constants is sufficient to reduce the signal in stacked spectrum to nothing, making it unlikely that our stacking analysis would recover spurious signal. This point is discussed further in the Supplementary Information, where we demonstrate the robustness of the method via jack-knifing the data.

Detection of $HC_{11}N$

$HC_{11}N$ has a long and colourful history in radio astronomy. Three radio lines were first reported toward IRC+10216 on the basis of a rotational constant derived by extrapolation from those measured experimentally for shorter members in this homologous series[27,28]. Any lingering doubt of the astronomical identification seemed to be put to rest with the observation of a fourth transition towards TMC-1 in 1985[29]. The subsequent laboratory detection of $HC_{11}N$ (ref. [30]), however, established that its rotational lines actually lie 0.13% lower in frequency (a shift equivalent to 13 linewidths in IRC+10216 and nearly 800 linewidths in TMC-1) relative to those originally reported[27,29]. The observed lines thus could not arise from $HC_{11}N$. Subsequently, two new astronomical lines were detected in TMC-1 with the NRAO 43 m radio telescope[31], both in apparent agreement with the laboratory rest frequencies. Albeit based on slender astronomical data, the detection of $HC_{11}N$ in space now seemed secure. In 2016, an attempt was made to verify the detection of $HC_{11}N$ by analysing archival observations towards TMC-1 with the 100 m GBT[4]. Even with substantially deeper integrations, no evidence was found for six consecutive transitions between 12.9 and 14.6 GHz. The non-detection of $HC_{11}N$ towards TMC-1 was further supported by observations that were unable to detect two higher-frequency transitions in a sensitive observation in the K band with the GBT[32].

The apparent absence of $HC_{11}N$ in TMC-1 and corresponding column density upper limit combined with a nonlinear relationship between column density and chain length for shorter cyanopolyynes (HC_xN)[31,33,34] led Loomis et al.[4] to hypothesize that cyclization reactions may become important once a carbon chain reaches a critical size. If correct, the formation of ring isomers could then directly compete with linear isomers via 'bottom-up' pathways. The detections of benzonitrile (cyclo-C_6H_5CN), the simplest aromatic nitrile[35], and now individual PAHs (B.A.M. et al., manuscript in preparation) in TMC-1 suggest that cyclic chemistry is far more widespread at these earliest stages of star formation than previously thought.

With confidence from the aforementioned tests that our method is able to rigorously detect not only species that show individual lines, but also those that sit below the visible noise, we turn back to the previous mysterious non-detection of $HC_{11}N$ (ref. [4]). A similar stacking and MCMC analysis was undertaken[4], but with substantially fewer data than are now available in the GOTHAM observations.

Unsurprisingly, we find that none of the brightest $HC_{11}N$ lines are individually detected in our observations (Fig. 6). By fitting for $HC_{11}N$ using priors from our HC_9N fit, however, we find column density posteriors that are consistent with a detection of $HC_{11}N$ (Extended Data Figs. 4 and 5). We visualize the significance of these posteriors through the same line stacking and matched filter analysis. The line stack shown in Fig. 7 displays a tentative but encouraging 3.8σ signal, and with a matched filter applied, the signal increases to a 5.0σ detection (Fig. 7).

The column density constraints from this analysis of $HC_{11}N$ yield a total column density of $7.8^{+21.27}_{-5.08} \times 10^{11}$ cm^{-2}. Three of the velocity components show well-constrained column densities, whereas the fourth component column density is best viewed as an upper limit. The total column density value is not directly comparable, however, with the 2σ upper limit of 9.4×10^{1} cm^{-2} from Loomis et al.[4], as that analysis did not constrain the $HC_{11}N$ source size, instead assuming a much larger fixed source size of $6.0' \times 1.3'$, which would fill the GBT beam (based on previous mapping observations of HC_3N). As seen in Extended Data Fig. 4, column density is highly covariant with our derived source size, and the largest contribution to the total $HC_{11}N$ column density comes from the fourth velocity component with a source size of ~9″. With the brightest $HC_{11}N$ lines originating in the X band, where the GBT beam size is ~1.2′, this source size would correspond to a beam dilution factor using equation (2) of ~0.015. Thus, under the same assumptions as Loomis et al.[4], our newly measured total $HC_{11}N$ column density would be roughly $1.2^{+3.2}_{-0.8} \times 10^{10}$ cm^{-2}, entirely consistent with their upper limit of 9.4×10^{10} cm^{-2}.

Discussion

Now with a detection of $HC_{11}N$, it is useful to reconsider the overall chemistry of cyanopolyynes in TMC-1, particularly focusing on both their relative column densities and distributions.

Cyanopolyyne column densities. The previous analysis of relative cyanopolyyne column densities synthesized both GBT observations reported in that paper[4] as well as previous literature values. In all cases, an assumption was made that emission filled the beam, and the individual velocity components were not considered.

These assumptions are reasonable for the smaller cyanopolyynes—we find that co-spatial fits to HC_3N and HC_5N substantially better replicate the observed line profiles. Although the more optically thin larger cyanopolyyne species such as HC_9N and $HC_{11}N$ are well fitted by a separate components fit, the varying source sizes in these fits make it very difficult to compare column densities across the two different fitting methods. For the purposes of this comparison, we have therefore additionally fitted all cyanopolyyne

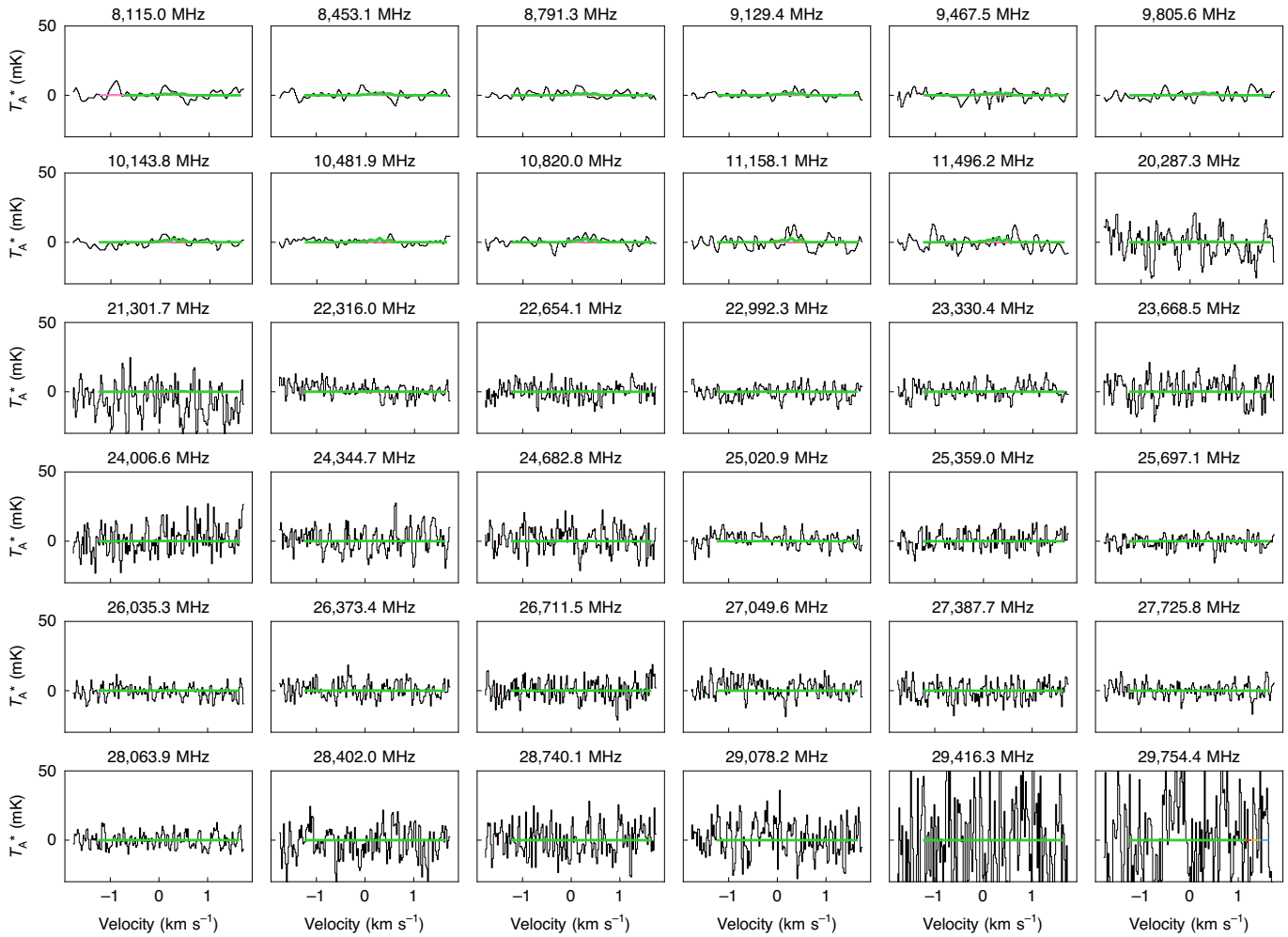


Fig. 6 | Individual line observations of $HC_{11}N$ in the GOTHAM data. The spectra (black) are displayed in velocity space relative to 5.8 km s^{-1}, and using the rest frequencies given in the top right of each panel. The best-fit model to the data, including all velocity components, is overlaid in green. Simulated spectra of the individual velocity components are shown in blue (5.63 km s^{-1}), yellow (5.79 km s^{-1}), red (5.91 km s^{-1}) and violet (6.03 km s^{-1}).

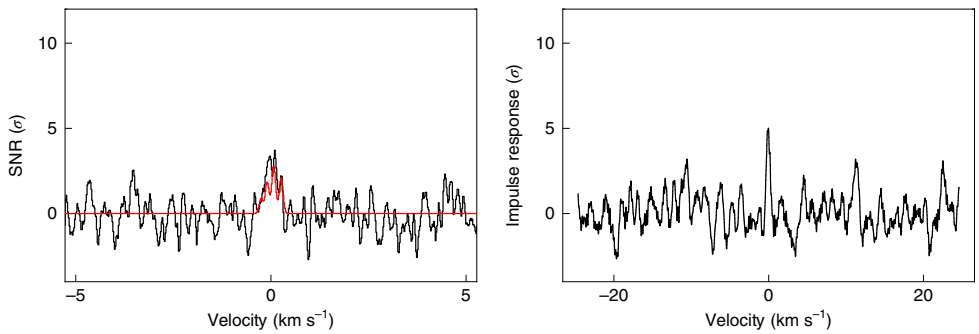


Fig. 7 | $HC_{11}N$ spectra in TMC-1. Left: velocity-stacked spectra of $HC_{11}N$ in black, with the corresponding stack of the simulation using the best-fit parameters to the individual lines in red. The data have been uniformly sampled to a resolution of 0.02 km s^{-1}. The intensity scale is the SNR of the spectrum at any given velocity. Right: impulse response function of the stacked $HC_{11}N$ spectrum using the simulated line profile as a matched filter. The intensity scale is the SNR of the response function when centred at a given velocity. The peak of the impulse response function provides a minimum significance for the detection of 5.0σ.

species with a co-spatial method, with results presented in detail in the Supplementary Information. The separate components and co-spatial results for the larger species are very similar, as would be expected for optically thin species. Further discussion of the relative source sizes and distributions of the cyanopolyynes and the effect on their fits is presented below.

Using the column densities derived from the fits presented in the Supplementary Information, an updated version of Fig. 5 from

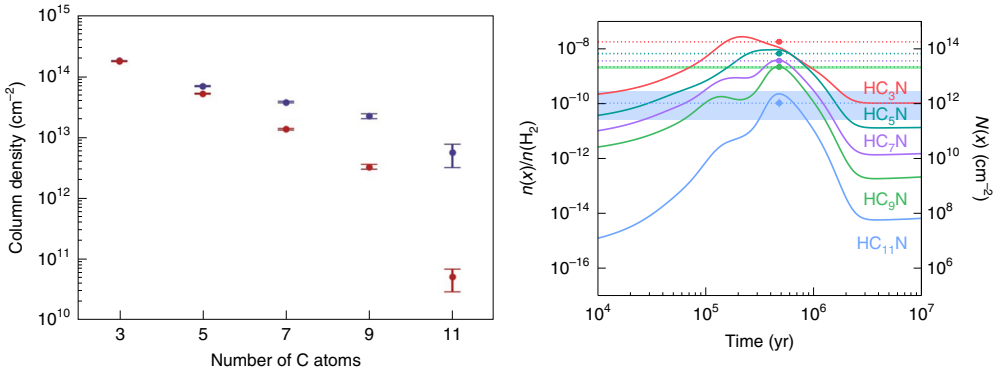


Fig. 8 | A comparison of cyanopolyyne column densities in TMC-1 from observations and a chemical model. Left: cyanopolyyne total column densities from fits using the co-spatial approximation (Supplementary Information) are plotted against carbon-chain size, as in ref. [4]. Blue data points are taken directly from our fits, where the source size of each component was taken into account. Red data points have been adjusted back to the value that one would calculate under the assumption that the source fills the beam. These red data points are directly comparable to those shown in fig. 5 in Loomis et al.[4]. Error bars reflect 1σ uncertainties in the fit column density. Right: calculated abundances (solid lines), abundances from the co-spatial MCMC analysis (dotted lines) and best-fit times (dots) for the cyanopolyynes HC$_n$N, $n \in [3, 5, 7, 9, 11]$. Abundance ranges from the separate components MCMC analysis for HC$_9$N and HC$_{11}$N are shown by the green and blue bars, respectively. Equivalent column densities assuming $N(H_2) = 10^{22}$ cm^{-2} (refs. [3,43]) are shown on the right-hand y axis. Error bars for HC$_3$N–HC$_9$N are not visible at the scale used, but can be found in Supplementary Tables 2–6.

Loomis et al.[4] is shown in Fig. 8, along with a comparison to predictions from a chemical model, discussed in more detail in the Supplementary Information. The general qualitative trend noted in that work is maintained, with a log-linear trend at smaller sizes, and a sharp decline at HC$_{11}$N.

Spatial variations in cyanopolyyne chemistry. Previous spatially resolved observations of HC$_3$N, HC$_5$N and HC$_7$N towards TMC-1 have shown them to be spatially extended on scales large enough to fill the GBT beam at the frequencies probed by GOTHAM[36–38]. These observations were all taken at relatively coarse spatial resolution, however, and the detailed distribution of these species is unknown, as is the distribution of larger cyanopolyynes such as HC$_9$N. In particular, observations of cyanopolyynes at both high spectral and spatial resolution do not exist to date, making it difficult to spatially disentangle the four known velocity components in TMC-1.

Several pieces of evidence suggest that our two limiting sets of assumptions in this analysis of cyanopolyynes are insufficient, but also provide some hints at the true cyanopolyyne distribution. First, we note that separate component fits for HC$_3$N and HC$_5$N yield line profiles that poorly represent the data, whereas the co-spatial fits shown in the Supplementary Information provide reasonable fits to the observational line profiles. This suggests that the velocity components are sufficiently co-spatial that when source sizes are large, they overlap substantially along the line of sight. Second, we find that for both the co-spatial and separate component fits, the source size(s) decrease with cyanopolyyne size as previously noted[39], possibly suggesting spatially segregated chemical evolution within the source. Finally, for more optically thin species such as HC$_9$N, separate component fits yield widely varying source sizes for the components. This suggests that the source components are not purely co-spatial, and probably have some scatter within the beam.

Our beam dilution and source-size fitting analysis is limited by both the sensitivity of our observations and the assumption that each source is centrally located within the beam. It is possible that the larger species have a broader distribution that is not well probed by our observations due to sensitivity limitations. If the spatiokinematic structure of the cyanopolyynes is shared by other species, it may be possible to use a single set of interferometric observations as a template to unlock the GOTHAM observations, enabling more complicated fitting and thus better characterization of the true spatial distribution of the column density.

In conclusion, we have presented a new method for robustly characterizing and visualizing detections of new interstellar species in line-sparse sources, even when individual lines of the species are not detected. These results of applying this method to the GOTHAM dataset have resulted in a total of six new interstellar species have been detected in TMC-1 (see refs. [8,25,40,41] and B.A.M. et al. (manuscript in preparation)) . In particular, we have detected HC$_{11}$N in TMC-1 and derived a column density consistent with the previous upper limit presented in Loomis et al.[4].

Data availability

The datasets analysed during the current study are available in the Green Bank Telescope archive (https://archive.nrao.edu/archive/advquery.jsp; PI: B.A.M.). A user manual for their reduction and analysis is also available (https://greenbankobservatory.org/science/gbt-observers/visitor-facilities-policies/data-reduction-gbt-using-idl/). The complete, reduced survey data in the X band are available as supplementary information in ref. [8]. The individual portions of the reduced spectra used in the analysis of the individual species presented here are available in the Harvard Dataverse Archive[42].

Code availability

All the codes used in the MCMC fitting and stacking analysis presented in this paper are open source and publicly available at https://github.com/ryanaloomis/TMC1_mcmc_fitting. The open source code for our spectral simulator can be found at https://github.com/ryanaloomis/spectral_simulator.

Received: 16 March 2020; Accepted: 22 October 2020;
Published online: 11 January 2021

References

1. McGuire, B. A. 2018 census of interstellar, circumstellar, extragalactic, protoplanetary disk, and exoplanetary molecules. *Astrophys. J. Suppl. Ser.* **239**, 17 (2018).
2. Czekala, I. DiskJockey: protoplanetary disk modeling for dynamical mass derivationAstrophys. Source Code Libr. (2016). ascl:1603.011.
3. Gratier, P. et al. A new reference chemical composition for TMC-1. *Astrophys. J. Suppl. Ser.* **225**, 25 (2016).

4. Loomis, R. A. et al. Non-detection of $HC_{11}N$ towards TMC-1: constraining the chemistry of large carbon-chain molecules. *Mon. Not. R. Astron. Soc.* **463**, 4175–4183 (2016).
5. Walsh, C. et al. First detection of gas-phase methanol in a protoplanetary disk. *Astrophys. J. Lett.* **823**, L10 (2016).
6. Loomis, R. A. et al. Detecting weak spectral lines in interferometric data through matched filtering. *Astron. J.* **155**, 182 (2018).
7. Loomis, R. A. et al. An unbiased ALMA spectral survey of the LkCa 15 and MWC 480 protoplanetary disks. *Astrophys. J.* **893**, 101 (2020).
8. McGuire, B. A. et al. Early science from GOTHAM: Project overview, methods, and the detection of interstellar propargyl cyanide ($HCCCH_2CN$) in TMC-1. *Astrophys. J. Lett.* **900**, L10 (2020).
9. Liu, S.-Y., Mehringer, D. M. & Snyder, L. E. Observations of formic acid in hot molecular cores. *Astrophys. J.* **552**, 654–663 (2001).
10. Remijan, A. J., Hollis, J. M., Lovas, F. J., Plusquellic, D. F. & Jewell, P. R. Interstellar isomers: the importance of bonding energy differences. *Astrophys. J.* **632**, 333–339 (2005).
11. Mangum, J. G. & Shirley, Y. L. How to calculate molecular column density. *Publ. Astron. Soc. Pac.* **127**, 266 (2015).
12. Pickett, H. M. The fitting and prediction of vibration-rotation spectra with spin interactions. *J. Mol. Spectrosc.* **148**, 371–377 (1991).
13. Drouin, B. J. Practical uses of SPFIT. *J. Mol. Spectrosc.* **340**, 1–15 (2017).
14. Dobashi, K. et al. Spectral tomography for the line-of-sight structures of the Taurus Molecular Cloud 1. *Astrophys. J.* **864**, 82 (2018).
15. Dobashi, K. et al. Discovery of CCS velocity-coherent substructures in the Taurus Molecular Cloud 1. *Astrophys. J.* **879**, 88 (2019).
16. GBT Support Staff. *Proposer's Guide for the Green Bank Telescope* (Green Bank Observatory, 2020); https://www.gb.nrao.edu/scienceDocs/GBTpg.pdf
17. Remijan, A. J., Hollis, J. M., Snyder, L. E., Jewell, P. R. & Lovas, F. J. Methyltriacetylene (CH_3C_6H) toward TMC-1: the largest detected symmetric top. *Astrophys. J. Lett.* **643**, L37–L40 (2006).
18. Foreman-Mackey, D., Hogg, D. W., Lang, D. & Goodman, J. emcee: the MCMC hammer. *Publ. Astron. Soc. Pac.* **125**, 306–312 (2013).
19. Gelman, A. & Rubin, D. B. Inference from iterative simulation using multiple sequences. *Stat. Sci.* **7**, 457–472 (1992).
20. Langston, G. & Turner, B. Detection of ^{13}C isotopomers of the molecule HC_7N. *Astrophys. J.* **658**, 455–461 (2007).
21. Woodward, P. *Probability and Information Theory: With Applications to Radar* Vol. 3 (Elsevier Science & Technology, 1953).
22. North, D. O. An analysis of the factors which determine signal/noise discrimination in pulsed-carrier systems. *Proc. IEEE* **51**, 1016–1027 (1963).
23. Portillo, S. K. N., Lee, B. C. G., Daylan, T. & Finkbeiner, D. P. Improved point-source detection in crowded fields using probabilistic cataloging. *Astron. J.* **154**, 132 (2017).
24. Siemiginowska, A. et al. The next decade of astroinformatics and astrostatistics. *Bull. Am. Astron. Soc.* **51**, 355 (2019).
25. McCarthy, M. C. et al. Detection of the highly polar five-membered ring cyanocyclopentadiene. *Nat. Astron.* https://doi.org/10.1038/s41550-020-01213-y (2020).
26. Crabtree, K. N. et al. Microwave spectral taxonomy: a semi-automated combination of chirped-pulse and cavity Fourier-transform microwave spectroscopy. *J. Chem. Phys.* **144**, 124201 (2016).
27. Bell, M. B., Feldman, P. A., Kwok, S. & Matthews, H. E. Detection of $HC_{11}N$ in IRC+10°216. *Nature* **295**, 389–391 (1982).
28. Oka, T. The prediction of the rotational constants of polyacetylene compounds H-(H ≡ H)*n*-H ≡ N. *J. Mol. Spectrosc.* **72**, 172–174 (1978).
29. Bell, M. B. & Matthews, H. E. Detection of $HC_{11}N$ in the cold dust cloud TMC-1. *Astrophys. J. Lett.* **291**, L63–L65 (1985).
30. Travers, M. J., McCarthy, M. C., Kalmus, P., Gottlieb, C. A. & Thaddeus, P. Laboratory detection of the linear cyanopolyyne $HC_{11}N$. *Astrophys. J. Lett.* **469**, L65–L68 (1996).
31. Bell, M. B. et al. Detection of $HC_{11}N$ in the cold dust cloud TMC-1. *Astrophys. J. Lett.* **483**, L61–L64 (1997).
32. Cordiner, M. A., Charnley, S. B., Kisiel, Z., McGuire, B. A. & Kuan, Y.-J. Deep K-band observations of TMC-1 with the Green Bank Telescope: detection of HC_7O, nondetection of $HC_{11}N$, and a search for new organic molecules. *Astrophys. J.* **850**, 187 (2017).
33. Bujarrabal, V., Guelin, M., Morris, M. & Thaddeus, P. The abundance and excitation of the carbon chains in interstellar molecular clouds. *Astron. Astrophys.* **99**, 239–247 (1981).
34. Ohishi, M. & Kaifu, N. Chemical and physical evolution of dark clouds: molecular spectral line survey toward TMC-1. *Faraday Discuss.* **109**, 205–216 (1998).
35. McGuire, B. A. et al. Detection of the aromatic molecule benzonitrile (*c*-C_6H_5CN) in the interstellar medium. *Science* **359**, 202–205 (2018).
36. Toelle, F., Ungerechts, H., Walmsley, C. M., Winnewisser, G. & Churchwell, E. A molecular line study of the elongated daark dust cloud TMC 1. *Astron. Astrophys.* **95**, 143–155 (1981).
37. Churchwell, E., Winnewisser, G. & Walmsley, C. M. Molecular observations of a possible proto-solar nebula in a dark cloud in Taurus. *Astron. Astrophys.* **67**, 139–147 (1978).
38. Olano, C. A., Walmsley, C. M. & Wilson, T. L. The relative distribution of NH_3, HC_7N and C_4H in the Taurus Molecular Cloud 1 (TMC 1). *Astron. Astrophys.* **196**, 194–200 (1988).
39. Bell, M. B., Watson, J. K. G., Feldman, P. A. & Travers, M. J. The excitation temperatures of HC_9N and other long cyanopolyynes in TMC-1. *Astrophys. J.* **508**, 286–290 (1998).
40. Burkhardt, A. M. et al. Ubiquitous aromatic carbon chemistry at the earliest stages of star formation. *Nat. Astron.* https://doi.org/10.1038/s41550-020-01253-4 (2020).
41. Xue, C. et al. Detection of interstellar HC_4NC and an investigation of isocyanopolyyne chemistry in TMC-1 conditions. *Astrophys. J. Lett.* **900**, L9 (2020).
42. GOTHAM Collaboration. Spectral stacking data for Phase 1 science release of GOTHAM. *Harvard Dataverse* https://doi.org/10.7910/DVN/PG7BHO (2020).
43. Fuente, A. et al. Gas phase elemental abundances in molecular cloudS (GEMS). I. The prototypical dark cloud TMC 1. *Astron. Astrophys.* **624**, A105 (2019).

Acknowledgements

A.M.B. acknowledges support from the Smithsonian Institution as a Submillimeter Array (SMA) Fellow. M.C.M. and K.L.K.L. acknowledge support from NSF grant number AST-1615847 and NASA grant number 80NSSC18K0396. Support for B.A.M. during the initial portions of this work was provided by NASA through Hubble Fellowship grant number HST-HF2-51396 awarded by the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy, Inc., for NASA, under contract number NAS5-26555. C.N.S. thanks the Alexander von Humboldt Stiftung/Foundation for their support, as well as V. Wakelam for use of the NAUTILUS v1.1 code. C.X. is a Grote Reber Fellow, and support for this work was provided by the NSF through the Grote Reber Fellowship Program administered by Associated Universities, Inc./National Radio Astronomy Observatory and the Virginia Space Grant Consortium. E.H. thanks the National Science Foundation for support through grant number AST 1906489. S.B.C. and M.A.C. were supported by the NASA Astrobiology Institute through the Goddard Center for Astrobiology. The National Radio Astronomy Observatory is a facility of the National Science Foundation operated under cooperative agreement by Associated Universities, Inc. The Green Bank Observatory is a facility of the National Science Foundation operated under cooperative agreement by Associated Universities, Inc.

Author contributions

R.A.L. wrote the manuscript and developed the MCMC and spectral stacking analysis code described here. M.C.M. and K.L.K.L. performed the laboratory experiments and theoretical calculations for several of the catalogues used in this analysis, and helped revise the manuscript. A.M.B and B.A.M. performed the astronomical observations and subsequent data reduction. E.H. determined and/or estimated rate coefficients and designed many of the original chemical simulations. A.M.B. and C.N.S. contributed or undertook the astronomical modelling and simulations. E.R.W., M.A.C., S.B.C., S.K. and B.A.M. contributed to the design of the GOTHAM survey, and helped revise the manuscript. C.X. modified and contributed the chemical networks of the related species and helped revise the manuscript. C.X. and A.J.R. performed the astronomical observations.

Competing interests

The authors declare no competing interests.

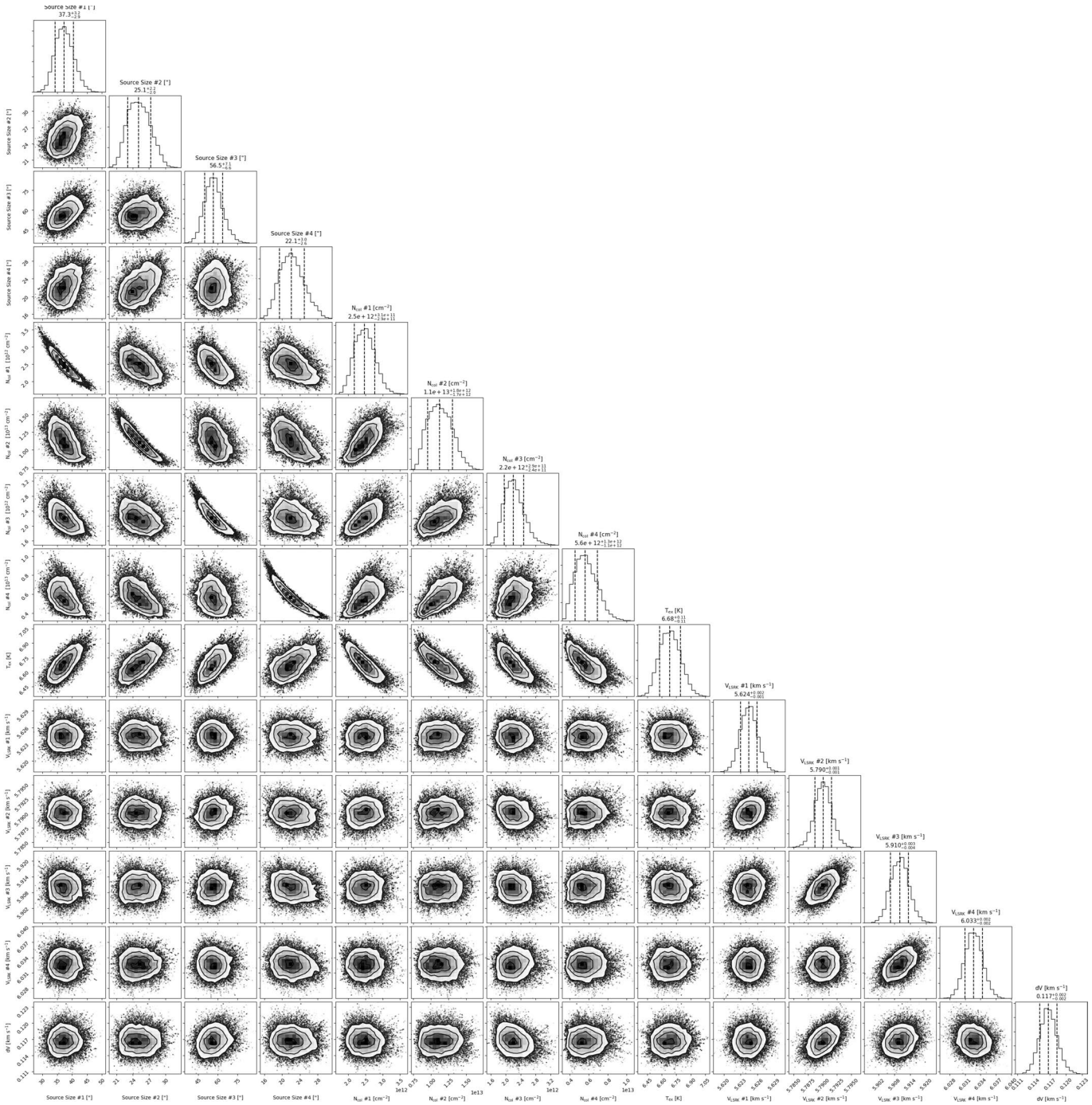
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41550-020-01261-4.

Supplementary information is available for this paper at https://doi.org/10.1038/s41550-020-01261-4.

Correspondence and requests for materials should be addressed to R.A.L. or B.A.M.

Peer review information *Nature Astronomy* thanks the anonymous reviewers for their contribution to the peer review of this work.



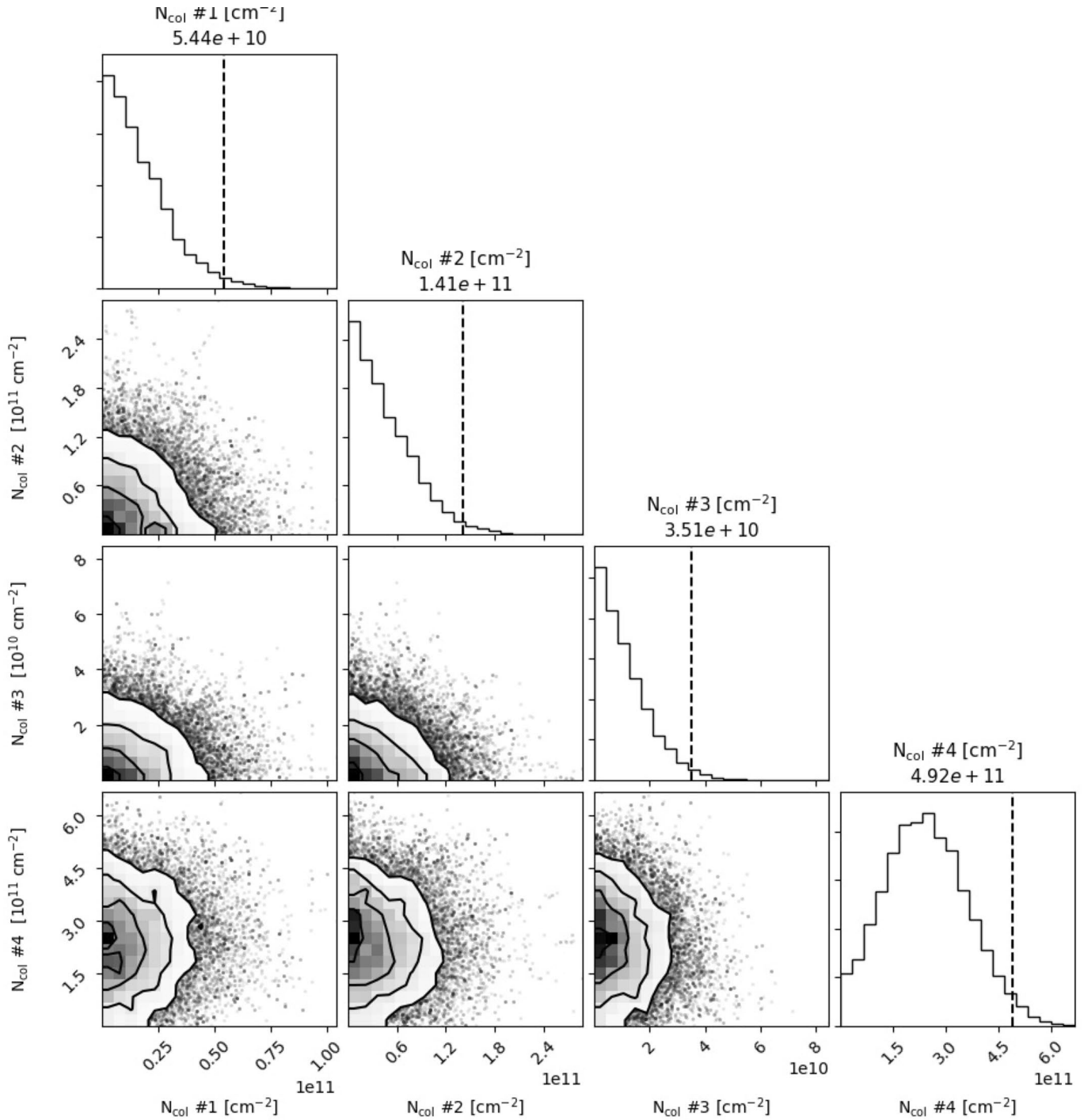
Extended Data Fig. 1 | Parameter covariances and marginalized posterior distributions for the HC_9N MCMC fit. 16th, 50th, and 84th confidence intervals (corresponding to $\pm$ 1 sigma for a Gaussian posterior distribution) are shown as vertical lines.

Component	v_{lsr} (km s^{-1})	Size ($''$)	$N_T^{\dagger}$ (10^{12} cm^{-2})	T_{ex} (K)	ΔV (km s^{-1})
C1	$5.624^{+0.002}_{-0.001}$	37^{+3}_{-2}	$2.47^{+0.31}_{-0.29}$	$6.7^{+0.1}_{-0.1}$	$0.117^{+0.002}_{-0.002}$
C2	$5.790^{+0.001}_{-0.001}$	25^{+2}_{-2}	$11.19^{+1.83}_{-1.67}$		
C3	$5.910^{+0.003}_{-0.004}$	56^{+7}_{-6}	$2.20^{+0.29}_{-0.24}$		
C4	$6.033^{+0.002}_{-0.002}$	22^{+2}_{-2}	$5.64^{+1.30}_{-1.07}$		
N_T (Total)††	$2.15^{+0.23}_{-0.20} \times 10^{13}$ cm^{-2}				

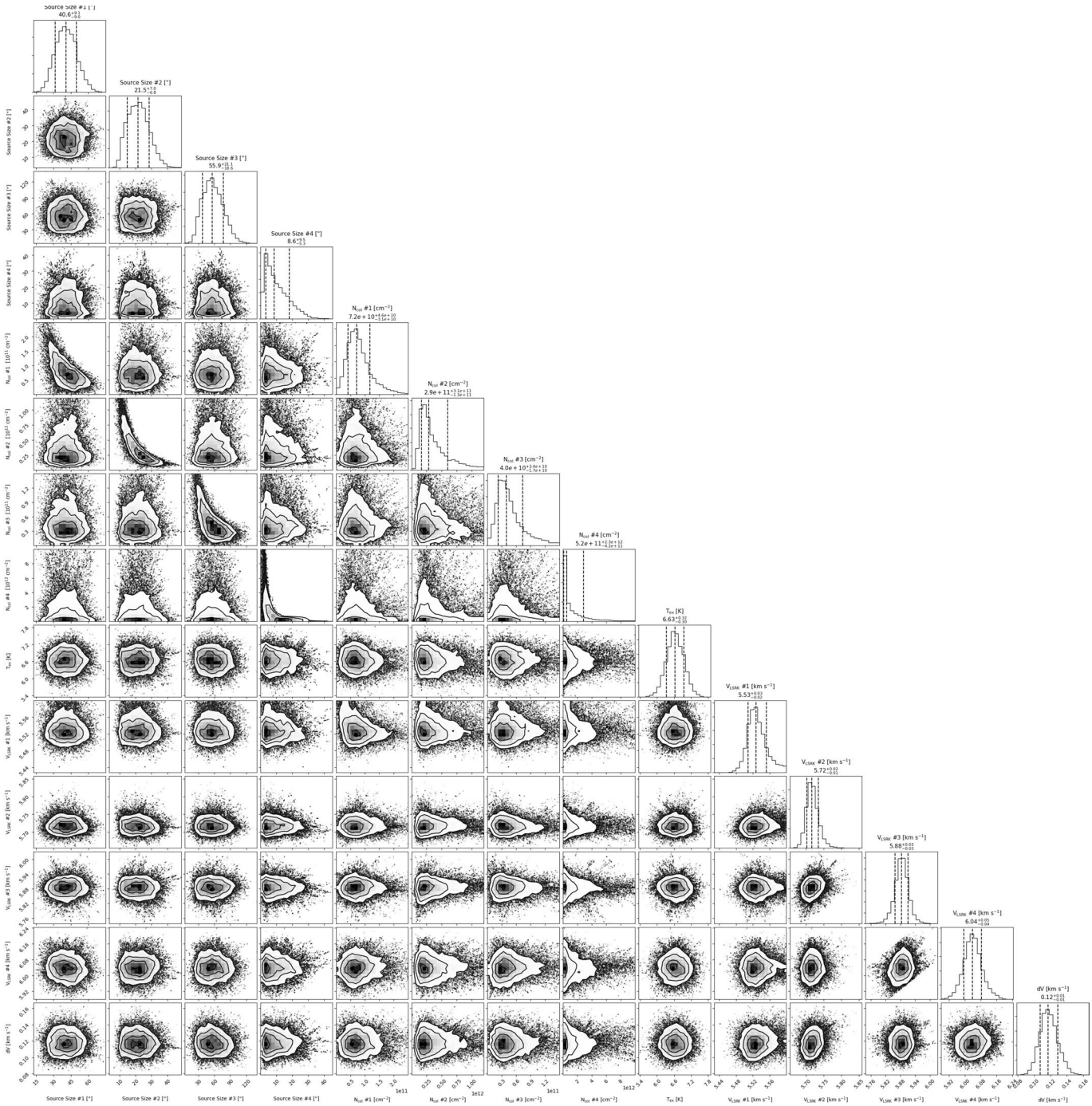
Note – The quoted uncertainties represent the 16^{th} and 84^{th} percentile (1σ for a Gaussian distribution) uncertainties. Values in the table are also available in the files provided at.[44]

†Column density values are highly covariant with the derived source sizes. The marginalized uncertainties on the column densities are therefore dominated by the largely unconstrained nature of the source sizes, and not by the signal-to-noise of the observations. ††Uncertainties derived by adding the uncertainties of the individual components in quadrature.

Extended Data Fig. 2 | HC_9N best-fit parameters from MCMC analysis. The quoted uncertainties represent the 16th and 84th percentile (1σ for a Gaussian distribution) uncertainties. †Column density values are highly covariant with the derived source sizes. The marginalized uncertainties on the column densities are therefore dominated by the largely unconstrained nature of the source sizes, and not by the signal-to-noise of the observations. ‡Uncertainties derived by adding the uncertainties of the individual components in quadrature.



Extended Data Fig. 3 | Parameter covariances and marginalized posterior distributions for the $HC_{13}N$ MCMC fit. The 97.8th confidence interval (corresponding to 2σ for a Gaussian posterior distribution) is shown as a vertical line.



Extended Data Fig. 4 | Parameter covariances and marginalized posterior distributions for the $HC_{11}N$ MCMC fit. 16th, 50th, and 84th confidence intervals (corresponding to $\pm$ 1σ for a Gaussian posterior distribution) are shown as vertical lines.

Component	v_{lsr} (km s^{-1})	Size ($''$)	$N_T^{\dagger}$ (10^{11} cm^{-2})	T_{ex} (K)	ΔV (km s^{-1})
C1	$5.532^{+0.113}_{-0.022}$	39^{+9}_{-8}	$0.73^{+0.54}_{-0.32}$	$6.6^{+0.3}_{-0.3}$	$0.117^{+0.012}_{-0.011}$
C2	$5.722^{+0.043}_{-0.017}$	21^{+7}_{-6}	$2.60^{+3.73}_{-1.31}$		
C3	$5.887^{+0.027}_{-0.023}$	56^{+18}_{-19}	$0.36^{+0.32}_{-0.17}$		
C4	$6.034^{+0.052}_{-0.041}$	9^{+10}_{-5}	$4.12^{+16.68}_{-3.28}$		
N_T (Total)††			$7.8^{+21.27}_{-5.08} \times 10^{11}$ cm^{-2}		

Note – The quoted uncertainties represent the 16^{th} and 84^{th} percentile (1σ for a Gaussian distribution) uncertainties. Values in the table are also available in the files provided at.[44]

†Column density values are highly covariant with the derived source sizes. The marginalized uncertainties on the column densities are therefore dominated by the largely unconstrained nature of the source sizes, and not by the signal-to-noise of the observations. ††Uncertainties derived by adding the uncertainties of the individual components in quadrature.

Extended Data Fig. 5 | HC_nN best-fit parameters from MCMC analysis. The quoted uncertainties represent the 16th and 84th percentile (1σ for a Gaussian distribution) uncertainties. Values in the table are also available in the files provided at ref. [42]. †Column density values are highly covariant with the derived source sizes. The marginalized uncertainties on the column densities are therefore dominated by the largely unconstrained nature of the source sizes, and not by the signal-to-noise of the observations. ‡Uncertainties derived by adding the uncertainties of the individual components in quadrature.