Cold atomic gas identified by H I self-absorption

Cold atomic clouds toward giant molecular filaments

J. Syed[1], H. Beuther[1], P. F. Goldsmith[2], Th. Henning[1], M. Heyer[3], R. S. Klessen[4, 5], J. M. Stil[6], J. D. Soler[7], L. D. Anderson[8, 9, 10], J. S. Urquhart[11], M. R. Rugel[12, 13], K. G. Johnston[14], and A. Brunthaler[15]

[1] Max-Planck-Institut für Astronomie, Königstuhl 17, 69117 Heidelberg, Germany
e-mail: syed@mpia.de
[2] Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA 91109, USA
[3] Astronomy Department, University of Massachusetts, Amherst, MA, 01003 USA
[4] Universität Heidelberg, Zentrum für Astronomie, Institut für Theoretische Astrophysik, Albert-Ueberle-Str. 2, 69120 Heidelberg, Germany
[5] Universität Heidelberg, Interdisziplinäres Zentrum für Wissenschaftliches Rechnen, INF 205, 69120 Heidelberg, Germany
[6] Department of Physics and Astronomy, The University of Calgary, 2500 University Drive NW, Calgary AB T2N 1N4, Canada
[7] Istituto di Astrofisica e Planetologia Spaziali (IAPS). INAF. Via Fosso del Cavaliere 100, 00133 Roma, Italy
[8] Department of Physics and Astronomy, West Virginia University, Morgantown, WV 26506, USA
[9] Adjunct Astronomer at the Green Bank Observatory, P.O. Box 2, Green Bank, WV 24944, USA
[10] Center for Gravitational Waves and Cosmology, West Virginia University, Chestnut Ridge Research Building, Morgantown, WV 26505, USA
[11] Centre for Astrophysics and Planetary Science, University of Kent, Canterbury CT2 7NH, UK
[12] Harvard Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, MA, 02138, USA
[13] National Radio Astronomy Observatory, 1003 Lopezville Rd, Socorro, NM 87801, USA
[14] School of Physics & Astronomy, Sir William Henry Bragg Building, The University of Leeds, Leeds, LS2 9JT, UK
[15] Max-Planck-Institut für Radioastronomie, Auf dem Hügel 69, 53121 Bonn, Germany

Received XX March XXXX; accepted XX October XXXX

ABSTRACT

Context. Stars form in the dense interiors of molecular clouds. The dynamics and physical properties of the atomic interstellar medium (ISM) set the conditions under which molecular clouds and eventually stars form. It is, therefore, critical to investigate the relationship between the atomic and molecular gas phase to understand the global star formation process.
Aims. Using the high angular resolution data from The H I/OH/Recombination (THOR) line survey of the Milky Way, we aim to constrain the kinematic and physical properties of the cold atomic hydrogen gas phase toward the inner Galactic plane.
Methods. H I self-absorption (HISA) has proven to be a viable method to detect cold atomic hydrogen clouds in the Galactic plane. With the help of a newly developed self-absorption extraction routine (astroSaber), we built upon previous case studies to identify H I self-absorption toward a sample of giant molecular filaments (GMFs).
Results. We find the cold atomic gas to be spatially correlated with the molecular gas on a global scale. The column densities of the cold atomic gas traced by HISA are usually on the order of $10^{20}\,\mathrm{cm}^{-2}$ whereas those of molecular hydrogen traced by ^{13}CO are at least an order of magnitude higher. The HISA column densities are attributed to a cold gas component that accounts for a fraction of ∼5% of the total atomic gas budget within the clouds. The HISA column density distributions show pronounced log-normal shapes that are broader than those traced by H I emission. The cold atomic gas is found to be moderately supersonic with Mach numbers of approximately a few. In contrast, highly supersonic dynamics drive the molecular gas within most filaments.
Conclusions. While H I self-absorption is likely to trace just a small fraction of the total cold neutral medium within a cloud, probing the cold atomic ISM by the means of self-absorption significantly improves our understanding of the dynamical and physical interaction between the atomic and molecular gas phase during cloud formation.

Key words. ISM: clouds – ISM: atoms – ISM: molecules – radio lines: ISM – stars: formation

1. Introduction

Atomic hydrogen provides the raw material to form molecular clouds, the sites of star formation. The dynamical and physical conditions under which molecular clouds form are therefore critical to understand the global star formation process. On a large scale ($\gtrsim 10^2$ pc), molecular clouds form out of the diffuse atomic interstellar medium (ISM; for a review see Ferrière 2001; Draine 2011; Klessen & Glover 2016) and are shaped by galactic dynamics and turbulence, stellar feedback, and magnetic fields. One major constituent of the ISM is the neutral atomic gas that provides the raw material to molecular clouds out of which stars eventually form and this gas phase takes in most of the energy and momentum feedback from its environment. The kinematic and physical relationship between the atomic and molecular gas phase of the ISM is then of central interest in the understanding of cloud formation processes.

According to the classical photodissociation region (PDR) model, layers of cold atomic hydrogen can effectively shield the cloud from photo-dissociating UV radiation at sufficiently high densities, allowing a transition of atomic hydrogen to its molecular form. In this idealized picture, pockets of high-density molecular hydrogen are embedded in an envelope of cold atomic hydrogen.

Using high angular resolution data of H I emission, the structure of the atomic ISM can be studied in great detail. However, probing the physical properties of the atomic gas from H I emission studies alone is not straightforward. In thermal pressure equilibrium, theoretical considerations based on ISM heating and cooling processes predict two stable phases of atomic hydrogen at the observed pressures in the ISM, namely the cold neutral medium (CNM) and warm neutral medium (WNM; Field et al. 1969; McKee & Ostriker 1977; Wolfire et al. 2003; Bialy & Sternberg 2019). Observations of H I emission are thus generally attributed to both CNM and WNM, which have significantly different physical properties (see below). In an attempt to observationally isolate the CNM from the bistable emission, H I self-absorption (HISA; Gibson et al. 2000; Li & Goldsmith 2003; Wang et al. 2020b; Syed et al. 2020) is a viable method to identify cold atomic gas and study H I clouds in the inner Milky Way but it heavily depends on the presence of sufficient background emission. In the following, we refer to cold atomic gas traced by HISA as "CoAt" gas, to make a distinction between the CNM as a whole and HISA-traced cold gas, which is a subset of the CNM. Due to the Galactic rotation, any positive (negative) line-of-sight velocity in the first (fourth) Galactic quadrant generally corresponds to a near kinematic and far kinematic distance within the solar circle (see e.g., Burton 1988). HISA has been used to resolve this kinematic distance ambiguity for molecular clouds or Galactic H II regions. Sources of interest at the far distance are less likely to show HISA as there is less background to absorb. Any detection of corresponding HISA would then place molecular clouds or H II regions at the near kinematic distance (e.g., Jackson et al. 2002; Anderson & Bania 2009; Duarte-Cabral et al. 2021).

Since the warm component of atomic hydrogen is more diffuse and has a lower density, it fills up a larger volume than the cold component (McKee & Ostriker 1977; Stahler & Palla 2005; Kalberla & Kerp 2009). H I self-absorption occurs when a cold H I cloud is located in front of a warmer H I emitting cloud. Self-absorption can occur within the same cloud but can also be induced by an emitting cloud in the far background that has the same velocity as the absorbing medium with respect to the local standard of rest $\upsilon_{\rm LSR}$. Therefore, the clouds do not have to be physically associated for HISA to be observable.

The CNM is observed to have temperatures $\lesssim$300 K and number densities $\gtrsim n_{\rm min,CNM}$=10 cm^{-3}, while the thermally stable WNM exceeds temperatures of ~5000 K with number densities $\lesssim n_{\rm max,WNM}$=0.1 cm^{-3} (Heiles & Troland 2003; Kalberla & Kerp 2009). In contrast to the properties of the CNM, the atomic gas traced by HISA has typical spin temperatures below 100 K, in most cases even below 50 K (e.g., Gibson et al. 2000; Li & Goldsmith 2003; Krčo et al. 2008; Wang et al. 2020b), thus highlighting the limited sensitivity of the HISA method for higher-temperature gas. For densities between $n_{\rm min,CNM}$ and $n_{\rm max,WNM}$, the gas is thermally unstable – denoted by unstable neutral medium (UNM) – and it moves toward a stable CNM or WNM branch under isobaric density perturbations (Field 1965).

H I self-absorption is found throughout the Milky Way in various environments. Many studies have focused on the detection of HISA, first measured in 1954 (Heeschen 1954, 1955), in known sources, but statistical treatments of the kinematic properties and densities of the HISA-traced cold gas in large-scale high-resolution maps are still scarce.

Extensive investigations of the HISA properties are limited to individual observational case studies (e.g., Gibson et al. 2000; Li & Goldsmith 2003; Kavars et al. 2003; Krčo et al. 2008; Wang et al. 2020b; Syed et al. 2020) and few simulations (e.g., Seifried et al. 2022). In this paper, we aim to build upon these case studies and investigate the HISA properties toward a sample of giant molecular filaments (GMFs; Ragan et al. 2014) as they are likely to be at an early evolutionary stage of giant molecular clouds forming out of the atomic phase of the interstellar medium (Zucker et al. 2018). These giant filaments potentially trace the Galactic structure, such as spiral arms and spurs, and large concentrations of molecular gas.

Seifried et al. (2022) present synthetic H I observations including H I self-absorption toward molecular clouds and investigate the observational effects and limitations of HISA. Their synthetic observations are based on 3D-magnetohydrodynamic simulations within the scope of the SILCC-Zoom project that include out-of-equilibrium H I-to-H_2 chemistry, detailed radiative transfer calculations, as well as observational effects like noise and limited spatial and spectral resolution that are similar to those of the THOR survey (see Sect. 2.1). Using commonly employed methods to derive the HISA properties, their results show that the H I column densities inferred from self-absorption tend to underestimate the real column densities of cold H I in a systematic way.

Traditionally, HISA features are obtained through various methods. To quantify the absorption depth, the strength of the warm emission background inducing H I self-absorption has to be determined first. Since the warm atomic gas component is often assumed to have less spatial variation in intensity, the emission background is commonly estimated by taking an "off" position spectrum at a different line of sight close to the location where HISA is expected to occur (e.g., Gibson et al. 2000; Kavars et al. 2003). However, it is challenging to select a position that is close enough, such that the off position can serve as a good proxy for the true HISA background, but far enough not to be interfered with by (partially) self-absorbing medium. Therefore, the location and spatial distribution of self-absorbing gas has to be known prior to estimating the background, which may work for single isolated cases. But particularly toward the Galactic plane, where multiple emission components and the Galactic rotation can add to the confusion along the line of sight, finding a clean off position has proven to be difficult to accomplish since the off spectrum can vary significantly over the angular size of the background cloud (see Wang et al. 2020b).

Another approach is to recover self-absorption in the spectral domain of H I observations. If the location of HISA in the H I emission spectrum is known, the spectral baseline can be determined by fitting the emission range around a HISA feature with a simple polynomial or Gaussian function (Li & Goldsmith 2003; Kavars et al. 2003; Wang et al. 2020b; Syed et al. 2020). However, the assumption of a velocity range where HISA is located introduces an additional source of bias, together with the specific fitting function that is used to derive the background. To address this issue, Krčo et al. (2008) and Dénes et al. (2018) have employed second and higher derivatives of the emission spectra to search for narrow HISA features (HINSA; Li & Goldsmith 2003; Goldsmith & Li 2005; Goldsmith et al. 2007) over the entire spectral range in a more unbiased way. Sharp kinks and dips in the spectra that are due to self-absorption are therefore expected to become readily apparent when investigating the derivatives. This technique allows HISA features to be filtered out without prior knowledge of their central velocities but it relies on high sensitivity, a well-sampled HISA line width, and

HISA features that are much narrower than the average emission component. However, the spectral baselines of these identified absorption features would then still need to be obtained using, for example, polynomial fits or making other physical assumptions of the HISA properties (e.g., Krčo et al. 2008).

In this paper, we address the lack of a versatile self-absorption reconstruction algorithm that can be applied to any dataset, at any spectral resolution, and self-absorption line width, and without the prior assumption that the cold H I gas is tightly correlated with molecular gas. We present the algorithm astroSaber (**S**elf-**A**bsorption **B**aseline **E**xtracto**R**) that operates by smoothing emission spectra in an asymmetric way, such that it not only identifies signal dips in the spectrum but directly provides a spectral baseline[1] of potential self-absorption features. It works in multiple iterations, such that both narrow and broad absorption components can be recovered. An optimization step has been implemented that is designed to tune the amount of smoothing that is required to recover self-absorption features, irrespective of spectral resolution and line width. To test the performance and applicability of the algorithm, we apply astroSaber to the known sample of GMFs (Ragan et al. 2014) since they serve as a good laboratory to investigate the presence of CoAt gas. The properties of H I self-absorption toward two of these molecular filaments have already been investigated in dedicated case studies employing previous HISA extraction methods (GMF20.0-17.9 in Syed et al. 2020 and GMF38.1-32.4 in Wang et al. 2020b).

The paper is organized as follows: In Sect. 2 we briefly introduce the data used in this analysis and outline the methods of our newly developed H I self-absorption extraction routine and Gaussian decomposition. In Sect. 3 we present the kinematic and column density properties derived from the HISA extraction and spectral decomposition. We discuss the kinematic and spatial relationship between the CoAt gas and molecular gas as well as the column density properties in Sect. 4. We furthermore elaborate on some of the limitations of our HISA extraction method before concluding with our summary in Section 5.

2. Methods and observations

2.1. H I, CO, and continuum observations

The following analysis of the HISA properties toward molecular clouds is based on the H I and 1.4 GHz continuum observations as part of The H I/OH Recombination line survey of the inner Milky Way (THOR; Beuther et al. 2016; Wang et al. 2020a). The final THOR-H I and 1.4 GHz continuum data include observations taken with the Karl G. Jansky Very Large Array (VLA) in C-configuration that were combined with the H I Very Large Array Galactic Plane Survey (VGPS; Stil et al. 2006), which consists of VLA D-configuration data. To account for missing flux on short *uv* spacings, the VGPS also includes single-dish observations of H I and 1.4 GHz continuum taken with the Green Bank and Effelsberg 100m telescope, respectively. The final H I emission data, from which the continuum has been subtracted during the data reduction, have an angular and spectral resolution of $\Delta\Theta = 40''$ and 1.5 km s^{-1}, respectively. The rms noise in emission-free channels is $\sigma_{\rm rms} \sim 4$ K.

We selected six GMF regions to investigate the presence of CoAt gas. Our selection is based on the findings of Ragan et al. (2014) and Zucker et al. (2018). Ragan et al. (2014) identified seven mid-infrared extinction features as giant filaments that exhibit corresponding ^{13}CO emission and velocity coherence over their full length. Of these seven GMFs, six of the fields are covered by the THOR survey. We present an overview of the six fields covering the filament regions in Table 1. The indices of the source names refer to the approximate range in Galactic longitude the giant filaments cover. The selected filament regions are in close proximity to the Galactic midplane and are located in the inner disk of the Milky Way, a site where HISA is more likely to occur. These GMFs serve as a good laboratory to investigate the relationship between the atomic and molecular gas as they are molecular concentrations of lengths >50 pc and likely to be at an early evolutionary stage having formed out of the large-scale diffuse ISM (Zucker et al. 2018). More details about each region can be found in Ragan et al. (2014). In Sect. 3.2.3, we correct for optical depth effects to compute the atomic hydrogen column densities from H I emission. The optical depths are taken from the measurements provided by Wang et al. (2020a) and have been obtained from VLA C-configuration data only that have an angular resolution of ∼15″ and effectively filter out large-scale emission, such that H I absorption against discrete continuum sources can yield a direct measurement of the optical depth of atomic hydrogen.

In order to provide a comprehensive description of the kinematic and spatial relationship between the atomic gas and the molecular gas, we investigate the molecular gas properties toward the GMF regions using two different datasets. The kinematic information is based on the ^{13}CO(1–0) data of the Galactic Ring Survey (GRS; Jackson et al. 2006), with an angular and spectral resolution of 46″ and 0.21 km s^{-1}, respectively. Riener et al. (2020) present an overview of a Gaussian decomposition of the entire GRS using the fully automated GaussPy+ algorithm (Riener et al. 2019). Since the decomposition results are publicly available, we use these data to investigate the kinematic properties of the clouds.

In Sect. D we compute the ^{13}CO column densities from the ^{12}CO(1–0) and ^{13}CO(1–0) emission line data taken from the Milky Way Imaging Scroll Painting (MWISP) survey (Su et al. 2019). The GRS does not include ^{12}CO observations, that are required to estimate the CO excitation temperatures and ultimately ^{13}CO column densities. We therefore use both the ^{13}CO and ^{12}CO from the MWISP data to derive the column density properties in a consistent way, and to reduce systematic errors arising from observational biases. The MWISP ^{12}CO and ^{13}CO data have an angular resolution of ∼55″ and an rms noise of 0.5 K and 0.3 K at a spectral resolution of 0.16 km s^{-1} and 0.17 km s^{-1}, respectively. The ^{12}CO data have been reprojected onto the same spectral grid as the ^{13}CO data to infer the excitation temperatures on a voxel-by-voxel basis. The rms noise of the ^{12}CO data is then reduced to 0.4 K.

2.2. Absorption baseline reconstruction

In this section we describe the astroSaber method that we used to obtain self-absorption baselines to recover HISA features. The basic workflow of astroSaber is the following: 1) Generating mock H I spectra to use as "training data"[2] (described in Sect. A.1), 2) finding optimal smoothing parameters using gra-

[1] In this paper, all instances of the word "baseline" refer to the absorption-free spectrum that is used as a spectral baseline to extract clean HISA.

[2] While the terms "test data" and training data are commonly used in the context of machine learning algorithms, we note that the accuracy of astroSaber is not tested on unseen data but the underlying concepts are the same, such that these concepts can be used to integrate them in a machine learning algorithm.

Table 1. Properties of studied filament regions.

(1)	(2)	(3)	(4)	(5)	(6)
Source name[(a)]	Glon [°.°]	Glat [°.°]	$\upsilon_{\rm LSR}$ [km s^{-1}]	$d_{\rm near}$[(b)] [kpc]	$D_{\rm GC}$[(b)] [kpc]
GMF20.0-17.9	17.80 – 20.60	−1.00 – +0.30	37 – 50	3.2	5.2
GMF26.7-25.4	25.10 – 26.90	+0.40 – +1.20	41 – 51	2.9	5.7
GMF38.1-32.4a	33.30 – 37.30	−1.00 – +0.60	50 – 60	3.2	5.9
GMF38.1-32.4b	33.30 – 37.30	−1.00 – +0.60	43 – 46	2.6	6.2
GMF41.0-41.3	40.80 – 41.50	−0.70 – +0.50	34 – 42	2.2	6.7
GMF54.0-52.0	52.30 – 54.20	−0.50 – +0.40	20 – 26	1.4	7.4

Notes. Columns (2) and (3) give the Galactic longitude range and latitude range of the filament regions, respectively. Column (4) gives the line-of-sight velocity range of each GMF as defined in Ragan et al. (2014). Columns (5) and (6) give the kinematic near distance from us and the Galactocentric distance, respectively.
[(a)] as in Ragan et al. (2014).
[(b)] The distances are not taken from Ragan et al. (2014) but have been recalculated using the more recent spiral arm model by Reid et al. (2019).

dient descent (described in Appendix A.1 and Appendix A.3), 3) applying baseline extraction with optimal smoothing parameters found in step 2) (see below).

The publicly available python-based astroSaber[3] algorithm is an automated baseline extraction routine that is designed to recover baselines of absorption features that are superposed with H I emission spectra. In the following, the astroSaber algorithm is described in detail. A description of astroSaber parameters used throughout the paper, including their keywords and default values, can be found in Appendix A.5. The astroSaber method utilizes asymmetric least squares smoothing first proposed by Eilers (2004) in the context of Raman spectroscopy. The algorithm progresses iteratively in two cycles to obtain a smoothed baseline, the major (outer) cycle and the minor (inner) cycle executed at each iteration of the major cycle. The basis of the minor cycle is to find a solution that minimizes the penalized least squares function

$$F(\mathbf{z}) = (\mathbf{y}-\mathbf{z})^{\top}\mathbf{W}(\mathbf{y}-\mathbf{z}) + \lambda\, \mathbf{z}^{\top}\mathbf{D}^{\top}\mathbf{D}\mathbf{z}\,, \tag{1}$$

where $\mathbf{y}$ is the input signal (e.g., the observed H I spectrum) and $\mathbf{z}$ is the asymmetrically smoothed baseline to be found. The first and second term on the right-hand side describe the fitness of the data and the smoothness of $\mathbf{z}$ defined by the second order differential matrix $\mathbf{D}$, respectively. The parameter λ, which is a two-dimensional vector by default (see below), adjusts the weight of the smoothing term. The regularized smoothing allows the detection of less significant absorption features that would otherwise be missed by finite-difference detection methods (see the discussion in Appendix B). In order to correct the baseline with respect to peaks and dips in the spectrum, the asymmetry weighting matrix $\mathbf{W} = \mathrm{diag}(\mathbf{w})$ is introduced. The asymmetry weights are initialized to be $w_i = 1$. After a first iteration of the minor cycle with equal weights, the weights for channels containing signal are then assigned as follows:

$$w_i = \begin{cases} p, & y_i > z_i \\ 1-p, & y_i \leq z_i \end{cases}\,. \tag{2}$$

The asymmetry parameter $p \in [0, 1]$ is set to favor either peaks or dips while smoothing the spectra. Given both the parameters λ and p, a smoothed baseline $\mathbf{z}$ is updated iteratively. Depending on p and the deviation of $\mathbf{z}$ from $\mathbf{y}$ after each iteration, peaks (dips) in the spectrum are retained while dips (peaks) are given less weight during the smoothing. Since we only aim to asymmetrically smooth real signals, spectral channels containing only noise are given equal weights of 0.5, hence the baseline will be within the noise in emission-free channels. The signal range estimation is described in Appendix A.4. As can be seen in Eq. (1), there is a degeneracy in the solution of the least squares function introduced by the weighting factors $\mathbf{W}(p)$ and λ. It is then sensible to keep one of these parameters fixed while finding the best-fit solution for the other parameter in order to optimize the smoothing (see Appendix A.1). In the case of self-absorption features, we therefore chose to fix the asymmetry parameter at $p = 0.9$.

After $n_{\rm minor}$ iterations, the minor cycle converges, such that the iteratively updated baseline $\mathbf{z}$ does not change anymore given the input spectrum $\mathbf{y}$. However, in order to effectively smooth out dips while still retaining real signal peaks in the spectra, the smoothed baseline $\mathbf{z}$ is then passed to the next iteration of the major cycle as input (i.e., now $\mathbf{y}$) for its minor cycle smoothing.

After evaluating the THOR-H I data, the minor cycle has shown to already converge after three iterations. Hence, the number of minor cycle iterations has been fixed at $n_{\rm minor} = 3$ in the algorithm. This parameter affects the output of astroSaber only mildly since the final smoothed baseline is mostly dependent on the number of iterations in the major cycle and on the λ parameter that tunes the smoothing (see Appendix A.1).

The algorithm stops as soon as a convergence criterion in the major cycle is met, or if the maximum number of iterations $n_{\rm major}$ is reached. The convergence criterion is met if the change in baseline from one major cycle iteration to the next is below a threshold set by $s_{\rm thresh} \cdot \sigma_{\rm rms}$ for at least some number of iterations $n_{\rm converge}$. The default values set by astroSaber are $n_{\rm major} = 20$, $s_{\rm thresh} = 1$, and $n_{\rm converge} = 3$. There is a slight degeneracy between the actual number of iterations needed to make the baseline converge and a fixed smoothing parameter λ used for every smoothing iteration. For λ sufficiently high, fewer iterations are needed to smooth out sharp kinks and dips in the spectrum. In the case of the THOR-H I data where the continuum has been subtracted during data reduction, the maximum number of iterations can be reached for emission spectra that are contaminated by imperfect continuum subtraction toward very strong continuum sources. Inspecting the number of iterations can therefore serve as an additional quality check of the spectra. For high-sensitivity data at a spectral resolution that is much smaller than the HISA

[3] https://github.com/astrojoni89/astrosaber

line width, the optimal smoothing parameter λ might be too large to make the algorithm converge since the change in baseline will be significant after every major cycle iteration. It can then be sensible to decrease the convergence threshold or to reduce the maximum number of iterations to force the algorithm to terminate and thus break down the aforementioned degeneracy.

In order to recover both narrow and broad features and to account for the possibility of an absorption baseline that exceeds the intensity of that in adjacent velocity channels, the astroSaber routine can be set to add a residual $\mathcal{R}_+$, which is the absolute difference between the first and last iteration of the major cycle. An example of this is an isolated emission feature with a Gaussian shape that has an absorption dip at the line center, or the "flat-top" spectrum observed in [C II] emission toward the H II region RCW120 (Kabanovic et al. 2022, see their Fig. 6). To add flexibility to the baseline reconstruction, the very first major cycle iteration can be set to operate with its own individual smoothing parameter λ_1 while all following iterations use a smoothing parameter λ_2. A λ_2 smoothing parameter close to zero is then effectively equal to a spectral smoothing without adding the residual. In Appendix A.1 we investigate how to optimize the smoothing parameters using mock-H I data.

Figure 1 shows a step-by-step baseline extraction of a mock spectrum to illustrate the major cycle workflow. The mock-H I contains three emission components where two absorption features of different line widths have been added. Given the observed spectrum (black spectrum in Fig. 1), astroSaber is run with optimal smoothing parameters (λ_1, λ_2) (see Appendix A.1). The left panel in Fig. 1 shows the baseline after the first major cycle iteration, that is after the minor cycle smoothing converged given the input spectrum (i.e., after Eq. (1) has been solved for $\mathbf{z}$). The middle panel then shows the converged baseline after the last major cycle iteration before adding the residual. The right panel presents the final baseline obtained by astroSaber after adding the residual. The baseline so obtained is able to recover the pure emission spectrum well within the uncertainties. We note that if the $\mathcal{R}_+$ setting is turned off, the smoothing parameters obtained during the optimization (Appendix A.1) will be adjusted to have larger values in order to recover the baseline. The differences in baseline between these settings are likely to be small at the velocities of the absorption signals. However, real signal is then also more likely to be smoothed out by the higher smoothing weight.

An example of the final output of the extraction step is shown in Fig. 2. The figure shows maps of an example region toward a (100 × 100) pixels subsection of GMF20.0-17.9 (see Ragan et al. 2014; Syed et al. 2020) that is also made publicly available with the astroSaber code. The maps present the H I emission data, the baselines obtained with optimized smoothing parameters, and the resulting H I self-absorption data, respectively.

2.3. Gaussian decomposition

After astroSaber has been applied to all six giant filament regions with optimized smoothing parameters, in each case the resulting output gives four data cubes containing the reconstructed baseline spectra, the self-absorption features (i.e., the H I emission spectra subtracted from the baselines), a map of the number of iterations that were required for the baselines to converge, and a map with flags for spectra that did not meet the convergence criteria, either due to missing signal in the spectra or having reached the maximum number of iterations set by the user. Spectra are flagged with "missing signal" if there is no significant emission (defined by $s_{\mathrm{signal}} \cdot \sigma_{\mathrm{rms}}$) in more than a specified number of consecutive spectral channels Δv_{LSR} (in units of km s^{-1}), and these spectra are removed from the final data cubes by default, as is done for the strong continuum source in Fig. 2. For the THOR-H I data, we applied the default settings $s_{\mathrm{signal}} = 6$ and $\Delta v_{\mathrm{LSR}} = 15$ km s^{-1} (see Table A.4).

The final self-absorption data cubes obtained from astroSaber contain what we refer to as HISA "candidates" since all signal dips have been extracted from the emission spectra. In the following steps, we decomposed the HISA candidate data cubes into their spectral components using the fully automated Gaussian decomposition algorithm GaussPy+[4] (Riener et al. 2019) and identified "real" H I self-absorption by cross-matching the centroid velocities with the molecular kinematics of the GMF regions given in Ragan et al. (2014).

GaussPy+ is a multicomponent Gaussian decomposition tool based on the earlier GaussPy algorithm (Lindner et al. 2015), and it provides additional preparatory steps and quality checks to improve the quality of the spectral decomposition. GaussPy has the fully automated means to decompose spectra using a supervised machine-learning technique. The algorithm automatically determines initial guesses for Gaussian fit components using derivative spectroscopy. To decompose the spectra, the spectra require smoothing to remove noise peaks while retaining real signal. The optimal smoothing parameters are found by employing a machine-learning algorithm that is trained on a few hundred well-fit spectra that are taken from a subsection of each dataset. GaussPy+ builds upon these results and introduces quality checks of the identified Gaussian components, such as FWHM values, signal-to-noise ratio, significance, and goodness of fit estimation (see Sect. 3.2 in Riener et al. 2019). These criteria are used to decide whether spectra are discarded or refit. Optional quality checks for broad or blended Gaussian components can also be imposed, depending on the specific dataset and expected physical cause of the spectral lines.

It is essential to reliably estimate the noise in the spectra to obtain good fit results. As we described above, GaussPy+ comes with an automated noise estimation routine as a preparatory step for the decomposition that also considers the median absolute deviation (MAD) of negative spectral channels to identify narrow spikes in the spectra that are masked before estimating the rms noise.[5]

For the full decomposition run of each dataset, we used the default parameters and standard quality control of GaussPy+ if not explicitly stated otherwise. A detailed description of all parameters and in-built quality checks is given in Riener et al. (2019). For each data cube, we ran the GaussPy+ training step with 300 randomly selected spectra from the HISA candidate data to find the optimal parameters for the fitting, as recommended in Riener et al. (2019). Owing to the absorption properties of the H I gas, we would naturally expect HISA to probe very cold gas so we opt to refit broad components in the GaussPy+ routine. GaussPy+ flags a component in a spectrum as broad if its line width is larger than the line width of the second broadest component by a user-defined factor (default: 2). We do not set a specific value as a line width limit. An absolute value that is used as a limiting line width might lead to unphysical fit solutions or artifacts, or can be difficult to determine since the range of expected values is not known.

[4] `https://github.com/mriener/gausspyplus`

[5] This step in the noise estimation of GaussPy+ is not included in astroSaber since we only want to identify signal ranges that are broad enough such that they can be used for generating mock self-absorption spectra.

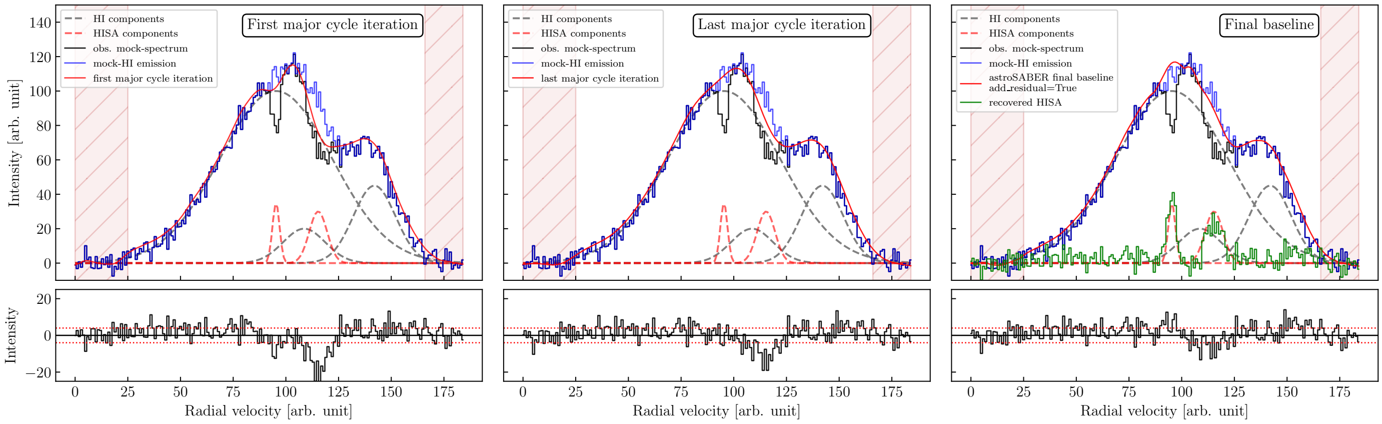

Fig. 1. Baseline extraction workflow of astroSaber. In each panel, the black mock spectrum represents the observed H i emission spectrum, which is the sum of the three gray dashed components, with self-absorption features (two red dashed components) superposed. The blue spectrum shows the "pure emission" spectrum that is to be recovered by the astroSaber algorithm. The algorithm is then applied to the observed spectrum using the optimal smoothing parameters (λ_1, λ_2). Hatched red areas indicate spectral channels that are masked out due to missing signal. *Left panel:* The astroSaber baseline (red) after the first major cycle iteration, that is, after the minor cycle smoothing converged given the input mock spectrum (i.e., after Eq. (1) has been solved for **z**). *Middle panel:* The astroSaber baseline (red) after the last major cycle iteration, that is, after the major cycle smoothing converged and before adding the residual, which is the absolute difference between the first and last major cycle iteration. *Right panel:* The final astroSaber baseline (red) after adding the residual. The baseline so obtained reproduces the pure emission spectrum (blue) well. The resulting HISA features expressed as equivalent emission features are shown in green, and show a good match with the real HISA absorption features. The smaller subpanels in each column show the residual, which is the difference between the red baseline and the blue emission spectrum, with the horizontal dotted red lines marking values of $\pm\sigma_{\rm rms}$.

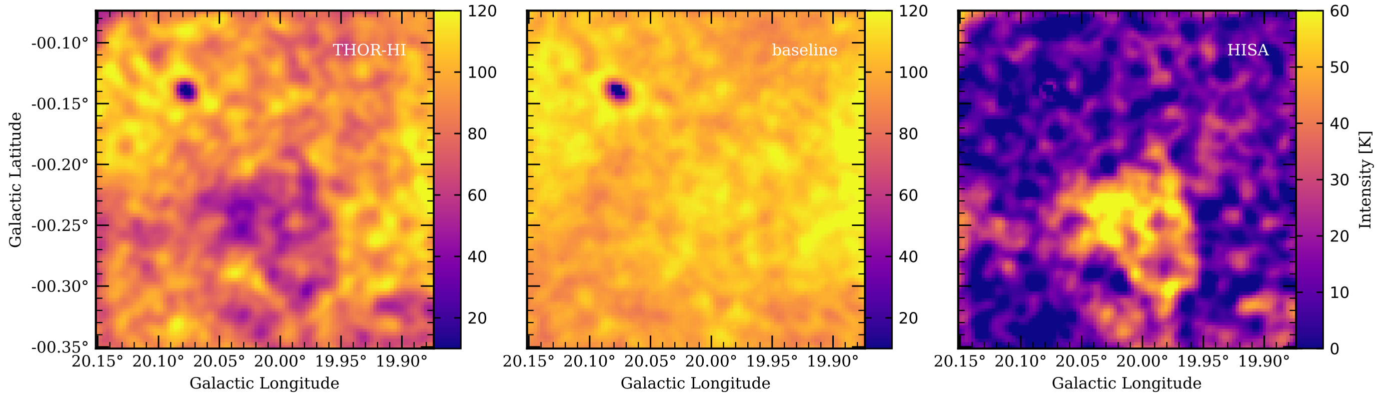

Fig. 2. Example H i self-absorption extraction. The *left panel* shows the observed THOR-H i emission channel map toward a (100 × 100) pixel subsection of the giant filament GMF20.0-17.9 at the velocity 44.5 km s^{-1}. The *middle panel* shows the map of the self-absorption baseline obtained with optimized smoothing parameters. The *right panel* gives the resulting HISA map, which is the difference between the baseline map and the H i emission map. The HISA feature in the bottom half of the map could be successfully recovered by astroSaber, while the strong continuum source in the top left was masked during a quality check of the spectra.

After the initial fitting, we apply the two-phase spatial coherence check implemented by GaussPy+ that can optimize the fit by refitting the components based on the fit results of neighboring pixels (see Sect. 3.3 in Riener et al. 2019). Mostly one velocity component was fit by GaussPy+ in the given velocity ranges of the filament regions. Only for some small isolated regions and single pixels more than one component was fit to the HISA spectra. As we show in a test environment in Appendix C, the centroid velocities recovered by the extraction and subsequent spectral decomposition are robust and have an uncertainty of ~0.35 km s^{-1}.

Spectra where the maximum number of iterations $n_{\rm major}$ is reached during the baseline extraction are flagged but not removed from the astroSaber routine. The affected spectra are usually toward positions where continuum emission contaminates the detection of self-absorption. We removed these spectra manually by masking pixels where there is strong continuum emission $T_{\rm cont} \geq 100$K. Due to the systematic uncertainty in the baselines and to ensure we only report reliable HISA features that are well detected, we additionally masked all pixels of the fit result maps where the corresponding fit amplitude is below $5\sigma_{\rm hisa} = \sqrt{2} \cdot 5\sigma_{\rm rms}$, with $\sigma_{\rm rms}$ being the rms noise of the THOR-H i emission data. The factor $\sqrt{2}$ accounts for the uncertainty in HISA amplitude that is due to the difference between the extracted HISA baseline and the H i emission.

3. Results

We show in Table 1 an overview of the filament regions analyzed in this paper, which are motivated by the results of Ragan et al.

(2014). We use their designated names (and shortened versions thereof) to refer to these regions throughout this paper.

We detect HISA toward all six filament regions. However, toward GMF26 and GMF41 only a small amount of CoAt gas could be recovered as HISA. The HISA-traced gas toward GMF26 does not appear to trace the distribution of the molecular gas well. Toward GMF20, GMF38a, and GMF38b we recovered a large cold atomic counterpart to the molecular gas within the filaments.

3.1. Kinematics

In this section, we discuss the kinematic properties of both HISA and their molecular counterpart as traced by ^{13}CO emission. As an example, we show the detected HISA and corresponding ^{13}CO emission map toward GMF20 in terms of their centroid velocities in Fig. 3. The kinematic maps of the remaining filament regions can be found in Appendix E. As we show in Appendix C, the centroid velocities and line widths are not heavily affected by our astroSABER routine and have an uncertainty of 0.4 km s^{-1} and 1.0 km s^{-1} (FWHM), respectively. Since the beam size of the THOR-H I data is similar to the one of the GRS survey (40″ and 46″, respectively), we chose to keep the original resolution (both spatially and spectrally) when comparing the kinematic maps. We tested smoothing the H I maps to the common beam size of 46″, which had a negligible effect.

For each of the kinematic histograms, we show every fit component along each line of sight within the velocity range of each filament region, thus taking into account multiple components if present. We furthermore only report fit components with an amplitude above the 5σ noise of the respective data cube ($\sigma \sim 5.7$ K for HISA, and $\sigma \sim 0.3$ K for ^{13}CO). The histograms of the centroid velocities of HISA and ^{13}CO show correlation for most of the filament regions (Fig. 4). The median peak velocity toward GMF20 is $v_{\rm LSR} = 44.7$ km s^{-1} for HISA and 44.0 km s^{-1} for ^{13}CO, which is in very good agreement with the results obtained in Syed et al. (2020). Particularly in the case of HISA, the histogram is mildly affected by components at higher velocities that might not be associated with the giant filament region or that might be troughs between two emission features erroneously picked up by the astroSABER routine. This effect is also evident in the histogram of GMF26. However, the median peak velocities also do agree toward GMF26, with $v_{\rm LSR} = 44.9$ km s^{-1} and 45.4 km s^{-1} for HISA and ^{13}CO, respectively. Toward GMF38a the histogram of peak velocities obtained with both astroSABER and the automated spectral line decomposition GAUSSPY+ reproduces the results presented in Wang et al. (2020b), with the median peak velocities agreeing to within 0.5 km s^{-1} ($v_{\rm LSR} = 54.3$ km s^{-1} and 54.8 km s^{-1} for HISA and ^{13}CO, respectively). The median HISA peak velocity of $v_{\rm LSR} = 44.6$ km s^{-1} toward GMF38b agrees with the ^{13}CO velocity of $v_{\rm LSR} = 44.4$ km s^{-1} within the uncertainty of our HISA extraction method. However, we caution that this agreement might be the result of a selection bias that only takes into account velocities in a rather narrow range. Since the GMF38b filament region is identified in the narrow velocity range between 43.0 km s^{-1} and 46.0 km s^{-1}, it is clear that the selection of velocities shows a smaller deviation between the two tracers. Toward GMF41 and GMF54 there is a more pronounced difference in median peak velocity. Within the GMF41 filament region, the median velocity traced by HISA is $v_{\rm LSR} = 38.0$ km s^{-1} while the median ^{13}CO velocity is $v_{\rm LSR} = 39.0$ km s^{-1}. The median peak velocities toward GMF54 are $v_{\rm LSR} = 24.2$ km s^{-1} and 23.1 km s^{-1} for HISA and ^{13}CO, respectively. We show the histograms of line width in Fig. 5. The peaks of the line width distributions are well resolved and above the spectral resolution limit, so the spectral resolution does not heavily affect the statistics of the kinematics. We find in general higher observed line widths in HISA than in ^{13}CO. The ^{13}CO line widths are 1.3–3.0 km s^{-1} while the HISA line widths are between 3.1 km s^{-1} and 5.2 km s^{-1}. The kinematic properties of the clouds are also summarized in Table 2. Assuming a kinetic temperature, we can estimate the expected thermal line width. In local thermodynamic equilibrium (LTE), the thermal line width (FWHM) is given by $\Delta v_{\rm th} = \sqrt{8\ln 2\, k_{\rm B}T_{\rm k}/(\mu m_{\rm H})}$, where $k_{\rm B}$, $T_{\rm k}$, and μ are the Boltzmann constant, kinetic temperature, and the mean molecular weight of H I ($\mu_{\rm H\,I} = 1.27$) and the CO molecule ($\mu_{\rm CO} = 2.34$; Allen 1973; Cox 2000) in terms of the mass of a hydrogen atom $m_{\rm H}$, respectively. The kinetic temperature can be well approximated by the estimated spin or excitation temperature of HISA and ^{13}CO, given the low temperatures and high densities of the cold gas (see Sects. 3.2.1 and D). If different line broadening effects are uncorrelated, the total observed line width will be

$$\Delta v_{\rm obs} = \sqrt{\Delta v_{\rm th}^2 + \Delta v_{\rm nth}^2 + \Delta v_{\rm res}^2}\,, \tag{3}$$

where $\Delta v_{\rm nth}$ is the line width due to nonthermal effects and $\Delta v_{\rm res}$ is the line width introduced by the spectral resolution of the data. The thermal line widths are on the order of ~0.5 km s^{-1} for ^{13}CO and ~1.0 km s^{-1} for HISA at the given temperatures. The observed line widths of both HISA and ^{13}CO therefore show that the line widths cannot be explained by thermal broadening alone. Nonthermal effects such as turbulent motions have a significant effect on the observed line widths and are most likely the dominant driver for the broadening of the lines. We investigate the turbulent Mach number of the gas in Sect. 4.2.

3.2. Column density and mass

In this section, we compute the column densities toward each filament region using HISA, ^{13}CO, and H I emission as a tracer for the CNM, molecular hydrogen, and bulk atomic hydrogen, respectively. For the derivation of the column density maps we integrated each filament region over the velocity range given in Table 1. The column density maps of each tracer can be found in Appendix F.

3.2.1. CNM column density traced by HISA

Following the derivation given in Gibson et al. (2000), we compute the optical depth of HISA as

$$\tau_{\rm HISA} = -\ln\left(1 - \frac{T_{\rm on} - T_{\rm off}}{T_{\rm HISA} - p_{\rm bg}T_{\rm off} - T_{\rm cont}}\right)\,, \tag{4}$$

with the dimensionless parameter $p_{\rm bg} \equiv T_{\rm bg}\left(1 - e^{-\tau_{\rm bg}}\right)/T_{\rm off}$ describing the fraction of background emission in the optically thin limit (Feldt 1993). Assuming a HISA spin temperature $T_{\rm s}$ ($= T_{\rm HISA}$), we can then calculate the H I column density of the cold H I gas using the general form (Wilson et al. 2013)

$$\frac{N_{\rm H}}{\rm cm^{-2}} = 1.8224\times10^{18}\,\frac{T_{\rm s}}{\rm K}\int \tau\left(T_{\rm s}, v\right)\left(\frac{{\rm d}v}{\rm km\,s^{-1}}\right)\,, \tag{5}$$

where $T_{\rm s}$ is the spin temperature of atomic hydrogen and $\tau\left(T_{\rm s}, v\right)$ describes the optical depth. We estimate the column density uncertainty by setting $T_{\rm on} = T_{\rm off} - \Delta T$ in Eq. (4) as the limit at which we can detect H I self-absorption, where ΔT is the rms noise in emission-free channels.

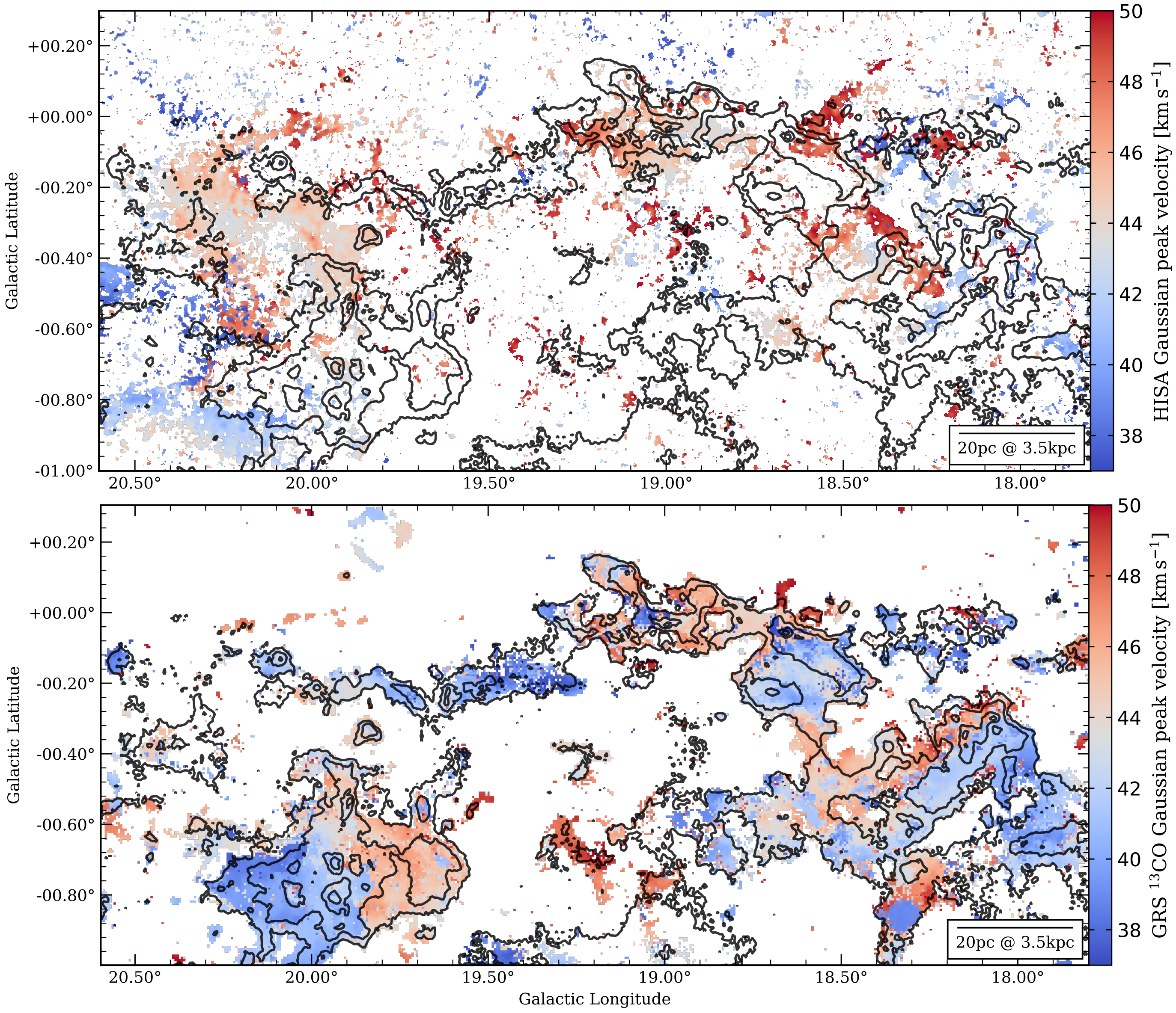

Fig. 3. Fit peak velocity toward GMF20. These maps show the peak velocities of fit components with amplitudes $\geq 5\sigma_{\rm rms}$ derived from the GaussPy+ decomposition of the spectra. If multiple components are present in a single pixel spectrum within the velocity range of the filament region, the component with the lowest peak velocity is shown. The black contours in both panels show the integrated GRS ^{13}CO emission at the levels 8.0, 16.0, 32.0, and 42.0 K km s^{-1}. *Top panel:* Fit HISA peak velocity. *Bottom panel:* Fit ^{13}CO peak velocity.

Table 2. Kinematic properties of the giant filament regions.

	HISA			^{13}CO		
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Source name	$\langle v \rangle$ [km s^{-1}]	$\langle \Delta v \rangle$ [km s^{-1}]	$\langle \mathcal{M} \rangle$	$\langle v \rangle$ [km s^{-1}]	$\langle \Delta v \rangle$ [km s^{-1}]	$\langle \mathcal{M} \rangle$
GMF20.0-17.9	44.7	3.9	4.8	44.0	3.0	9.5
GMF26.7-25.4	44.9	3.2	3.7	45.4	2.1	7.9
GMF38.1-32.4a	54.3	3.8	4.1	54.8	2.5	9.0
GMF38.1-32.4b	44.6	5.2	6.0	44.4	2.4	7.0
GMF41.0-41.3	38.0	3.6	3.7	39.0	2.5	8.2
GMF54.0-52.0	24.2	3.1	3.7	23.1	1.3	4.8

Notes. Columns (2) and (5) give the median peak velocity as traced by HISA and ^{13}CO for all six filament regions, respectively. Similarly, columns (3) and (6) present the median line width as traced by HISA and ^{13}CO, respectively. Columns (4) and (7) give the median sonic Mach number of HISA and ^{13}CO, respectively, which are computed in Sect. 4.2 using the sound speed at the temperatures estimated in Sects. 3.2.1 and D.

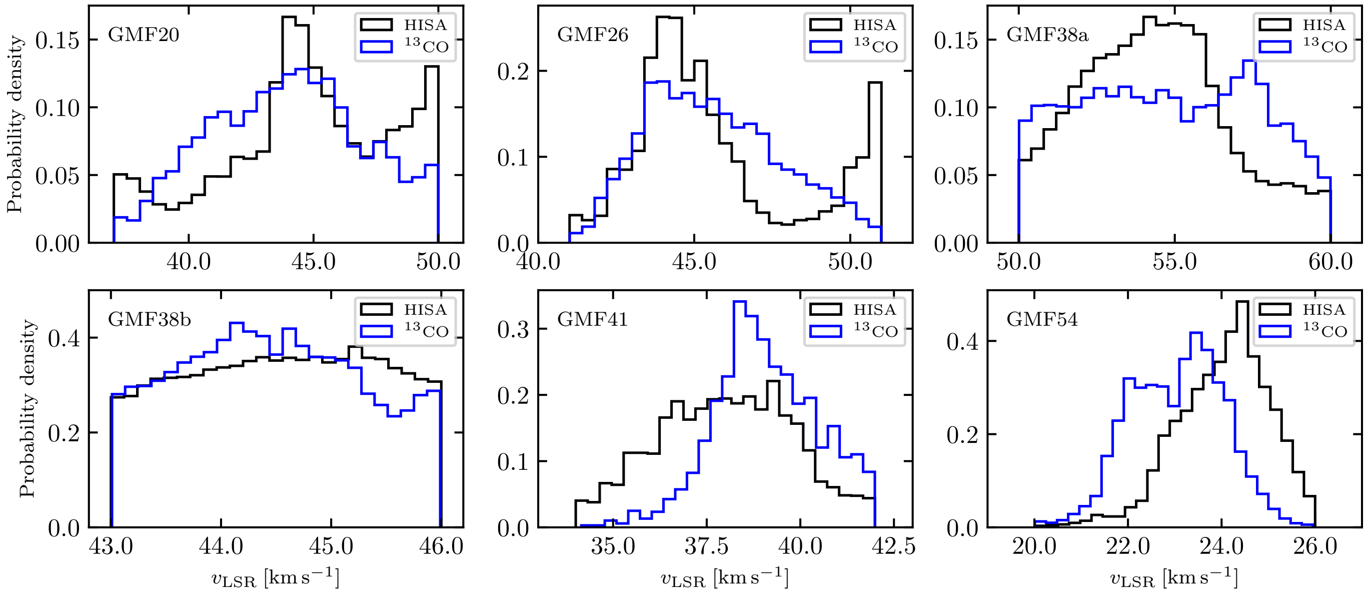

Fig. 4. Histograms of fit peak velocities. The panels show for each of the six giant filament regions the normalized histogram of peak velocities of HISA and ^{13}CO in black and blue, respectively.

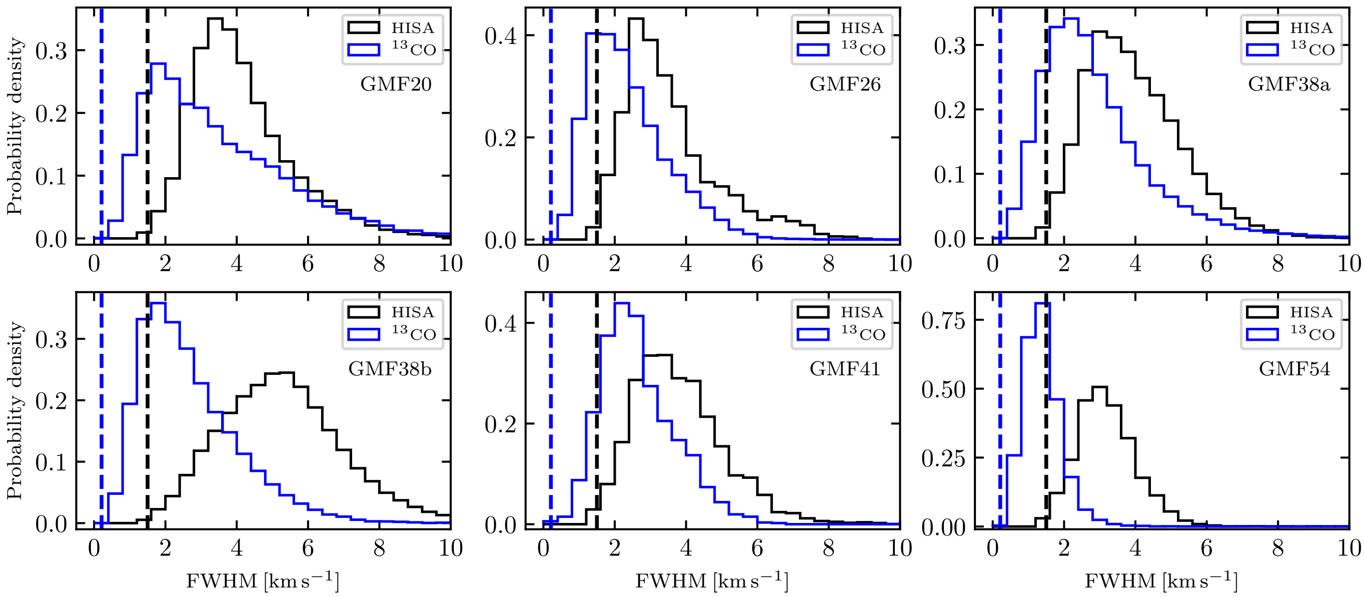

Fig. 5. Histograms of fit line widths. The panels show for each of the six giant filament regions the normalized histogram of FWHM line widths of HISA and ^{13}CO in black and blue, respectively. The black and blue vertical dashed lines mark the spectral resolution limit of the HISA and ^{13}CO data, respectively.

We can estimate the amount of background emission from the radial H I volume density distribution in the Galactic plane. For Galactocentric radii $7 \lesssim D_{\rm GC} \lesssim 35$ kpc Kalberla & Dedes (2008) report an average mid-plane volume density distribution of $n(R) \sim n_0\, e^{-(D_{\rm GC}-D_{\odot})/D_n}$ with $n_0 = 0.9\,{\rm cm}^{-3}$, $D_{\odot} = 8.5$ kpc, and with the radial scale length $D_n = 3.15$ kpc (IAU recommendations). We assume a constant volume density distribution in the inner Galaxy of $n(D_{\rm GC} < 7\,{\rm kpc}) = n(D_{\rm GC} = 7\,{\rm kpc})$ as the volume density distribution is flattening off at lower Galactocentric distances (see Fig. 5 in Kalberla & Dedes 2008). This relation gives the averaged distribution of the northern and southern Galactic plane and could hold systematic differences in some regions (Kalberla & Kerp 2009).

In principle, only atomic gas located at a distance that corresponds to the radial velocity of HISA is relevant to the estimation of the background. Since we usually place any observed HISA at the near kinematic distance, most of the emission must stem from the background and thus move $p_{\rm bg}$ close to 1 at any given finite spectral resolution element. However, due to the velocity dispersion of H I, in particular the WNM, atomic hydrogen emission that has radial velocities slightly offset from the HISA velocity can blend into the velocity channels under consideration. Atomic hydrogen emission in the foreground or background that corresponds to radial velocities around HISA can therefore contribute to the observed feature and affect the optical depth computation. We estimate the background fraction using the volume densities in the kinematic near and far distance intervals Δd corresponding to radial velocity intervals around the mean velocities of the clouds. The length of the distance interval is estimated from an average velocity dispersion of $\sim 10\,\mathrm{km\,s^{-1}}$ that falls between typical CNM and WNM velocity dispersions found in the Milky Way (Heiles & Troland 2003; Haud & Kalberla 2007). Equal steps in radial velocity Δv_{LSR} are mapped into unequal steps in distance ΔD that are proportional to the inverse of the velocity gradient along the line of sight

$$\Delta D = \left|\frac{\mathrm{d}D}{\mathrm{d}v_{\mathrm{LSR}}}\right| \Delta v_{\mathrm{LSR}}\ . \tag{6}$$

Using the rotation curve by McClure-Griffiths & Dickey (2007) gives distance intervals between ∼0.6 kpc and ∼0.9 kpc for the considered clouds. The gas density is then integrated from $d_{\mathrm{near}} - \Delta d$ to d_{near} to obtain a foreground fraction of the emission at the velocity of HISA. The background gas fraction is inferred by adding the integrated gas density in the interval $[d_{\mathrm{near}}, d_{\mathrm{near}} + \Delta d]$ to the gas density integrated on the kinematic far side interval $[d_{\mathrm{far}} - \Delta d, d_{\mathrm{far}} + \Delta d]$. The derived background fractions are between 0.75 and 0.77. If we assume a continued exponential rise in volume density toward the inner Galaxy instead of a constant density distribution, the background fraction increases by up to three percentage points. This effect is strongest for sources at lower Galactic longitude. We do note that while considering many factors in the treatment of the background fraction, the uncertainties are substantial due to noncircular and streaming motions superposed with the Galactic rotation, or systematic differences in the density distribution of H I. This adds a considerable source of uncertainty to the column density derivation. The column density decreases by a factor of ∼2 if the background fraction increases from $p_{\mathrm{bg}} = 0.7$ to 0.9 (see a detailed discussion of these uncertainties in Wang et al. 2020b; Syed et al. 2020). These uncertainties are revisited later in this section. Since we expect most of the emission background to originate in more diffuse H I gas, we assume a constant p_{bg} for each filament region (see Table 3).

Depending on the assumed spin temperature T_{s} and background fraction p_{bg}, there might be no solution to Eq. (4) in some velocity channels of the spectra if the spin temperature is too high. Disregarding these channels in the line-of-sight integration, the column density computed in Eq. (5) underestimates the true column density toward some regions. To resolve this, we derived the maximum spin temperature limit by $T_{\mathrm{max}} = T_{\mathrm{on}} + T_{\mathrm{cont}} - (1 - p_{\mathrm{bg}})T_{\mathrm{off}}$ for $\tau \rightarrow \infty$ (see Eq. 4), and set the 0.1-percentile of the maximum spin temperature to be the spin temperature of the whole cloud, such that outliers in the extracted baseline data or strong continuum emission do not affect the temperature estimate while Eq. (4) gives a solution for 99.9% of pixels in the integrated column density map. The assumed spin temperatures T_{s} and estimated background fractions p_{bg} are shown in Table 3 for each GMF source. Since a lower assumed spin temperature (at constant background fraction) producing the same observed HISA feature results in a lower column density (see Eqs. 4 and 5), the total column density is yet again underestimated. This shortcoming of the HISA column density computation is addressed in the HISA simulations conducted by Seifried et al. (2022). However, assuming a constant spin temperature for the entire cloud appears to be the best approach to qualitatively recover the true column density structure of the cloud (Seifried et al. 2022).

If the spin temperature is varied by 10 K, the column density, and consequently mass, traced by HISA changes by a factor of ∼2. Hence, the largest uncertainty arises from the assumption of a spin temperature and the background fraction that is coupled to the optical depth computation. Even for an assumed spin temperature that comes close to the limit at which the optical depth computation gives an analytic solution (see Eq. 4), the column density is still underestimated due to line-of-sight variations in spin temperature and observational noise. By assuming an optically thick HISA cloud with $\tau \rightarrow \infty$, we are able to determine the spin temperature limit above which the line-of-sight geometry does not allow the computation of the column density. The uncertainty in column density and mass is further amplified by the background fraction p_{bg} in Eq. (4). If the background fraction is lowered, the column density will increase as the cold H I cloud would be more efficient in producing the same observed HISA feature given the weaker background. A moderate variation in the background fraction of 10% at fixed spin temperature results in a ∼30% change in column density and therefore mass. We derive a HISA mass uncertainty by varying the background fraction by 10% and adjusting the spin temperature accordingly, such that we have again a solution for most pixels in the map. We report these uncertainties in Table 5 as well. For a more detailed discussion of these uncertainties we refer to Wang et al. (2020b) and Syed et al. (2020).

The aforementioned statistical uncertainties add to the intrinsic systematic effects of the HISA method. We are generally limited by the background emission that enables the observation of HISA. Furthermore, HISA is only sensitive to gas that is colder than the gas that contributes to the emission background. The CNM is reported to have spin temperatures up to ∼300 K (e.g., Heiles & Troland 2003; Kalberla & Kerp 2009), rendering the HISA detection of the CNM in many cases impossible given the observed brightness temperatures. According to the simulations conducted by Seifried et al. (2022), the HISA-traced mass underestimates the mass of the CNM that could in principle be observed through HISA by a factor of 3–10. This underestimation is generally attributed to two effects. The proper estimation of the spin temperature that is required to compute the HISA properties is a challenging task because of its variation within a cold H I cloud. Due to the line-of-sight geometry, an assumed H I spin temperature that is too low results in an underestimate of the optical depth and the true column density (see Eqs. 4 and 5). An H I spin temperature that is too high causes the HISA-traced cloud to have no solution to the optical depth at least for some part of the spectrum (Eq. 4). This again underestimates the integrated column density as individual spectral channels are omitted. Varying the spin temperature along the line-of-sight or spatially can lead to an even larger deviation and might recover a column density structure that does not reflect the true distribution qualitatively. The challenges of unknown spin temperature consequently give rise to a large systematic uncertainty in the determination of the column density and mass (Seifried et al. 2022).

3.2.2. H_2 column density

We computed the ^{13}CO column densities following the standard procedure given in Wilson et al. (2013). Details of the derivation are given in Appendix D. In order to convert the ^{13}CO column densities to H_2 column densities, we used Galactocentric distance-dependent abundance relations to estimate an $[\mathrm{H_2}]/[^{13}\mathrm{CO}]$ conversion factor for each source. Giannetti et al. (2014) give a ^{12}CO-to-^{13}CO abundance relation of $[^{12}\mathrm{CO}]/[^{13}\mathrm{CO}] = 6.2D_{\mathrm{GC}} + 9.0$, and the H_2-to-^{12}CO abundance given in Fontani et al. (2012) is $[\mathrm{H_2}]/[^{12}\mathrm{CO}] = [8.5\times 10^{-5}\exp(1.105 - 0.13D_{\mathrm{GC}})]^{-1}$, where D_{GC} is the Galactocentric distance in units of kpc. We estimate the uncertainty in H_2

Table 3. Assumed spin temperatures and background fractions.

Source	T_s [K]	background fraction p_{bg}
GMF20	26	0.75
GMF26	27	0.75
GMF38a	32	0.75
GMF38b	33	0.75
GMF41	37	0.76
GMF54	24	0.77

Notes. The second column gives the spin temperature T_s assumed toward each GMF region. The third column gives the background fraction p_{bg} that is estimated from the ratio of foreground and background column density along the line of sight (see Sect. 3.2.1).

column density to be at least 50% due to the large uncertainties in these relations. Furthermore, CO might not always be a good tracer of H_2 as "CO-dark H_2" could account for a significant fraction of the total H_2 (Pineda et al. 2008; Goodman et al. 2009; Pineda et al. 2013; Smith et al. 2014; Tang et al. 2016), particularly at low column densities and early evolutionary stages as molecular clouds might not have become CO-bright yet (Goldsmith et al. 2008; Planck Collaboration et al. 2011). The $[H_2]/[^{13}CO]$ conversion factor for each source is given in Table 4.

Table 4. Limits of CO excitation temperatures and optical depths.

Source	$T_{ex,low}$ [K]	$T_{ex,up}$ [K]	τ_{low}	$X([H_2]/[^{13}CO])$
GMF20	5	29	0.05	3.1×10^5
GMF26	5	16	0.11	3.6×10^5
GMF38a	5	21	0.08	3.9×10^5
GMF38b	5	18	0.10	4.1×10^5
GMF41	5	12	0.26	4.7×10^5
GMF54	5	36	0.04	5.6×10^5

Notes. The second and third column gives the lower limit and upper limit of the CO excitation temperature, respectively. The fourth column shows the lower limit of the optical depth estimated from the 5σ ^{13}CO noise and the highest excitation temperature found toward each source (see also Appendix D for details). The last column gives the ^{13}CO-to-H_2 conversion factor that we have used for each source.

3.2.3. Atomic gas column density seen in H I emission

In addition to the cold atomic gas traced by HISA, we investigated the properties of the total atomic hydrogen gas budget (WNM+CNM) by measuring the column density from H I emission and correcting for optical depth effects and diffuse continuum. As the optically thin assumption might not hold for some regions, we can utilize strong continuum emission sources to directly measure the optical depth. H I continuum absorption (HICA) is a classical method to derive the properties of the CNM (e.g., Strasser & Taylor 2004; Heiles & Troland 2003). This method uses strong continuum sources, such as Galactic H II regions or active galactic nuclei (AGNs), to measure the optical depth of H I. As these sources have brightness temperatures that are larger than typical spin temperatures of cold H I clouds, we observe the H I cloud in absorption. The absorption feature is furthermore dominated by the CNM since the absorption is proportional to T_s^{-1} (e.g., Wilson et al. 2013).

The advantage of this method is the direct measurement of the optical depth. However, the HICA method requires strong continuum emission sources. As most strong continuum sources are discrete point sources, this method results in an incomplete census of optical depth measurements. However, Wang et al. (2020a) derived a velocity-resolved optical depth map computed from 228 continuum sources within the THOR survey that are above a 6σ noise threshold and interpolated the measurements using a nearest-neighbor method. For more details about the optical depth measurement we refer to Wang et al. (2020a). To the first approximation, we can use this optical depth map to correct the H I column density as confirmed in Syed et al. (2020), in spite of potential kinematic distance ambiguities and the location of a continuum source along each line of sight that might add or miss optical depth for each line-of-sight velocity, respectively. For each velocity channel, we take the spatial average of the optical depth map measured toward each filament region in order to avoid artifacts introduced by the interpolation.

In addition to strong continuum sources, we observe weak continuum emission throughout the Galactic plane. This component has brightness temperatures between 10 and 50 K. The continuum emission has been subtracted during data reduction as described in Sect. 2.1. As even weak continuum emission might suppress H I emission and therefore lead to an underestimate of the column density, we account for the weak emission component when computing the H I column density (see Bihr et al. 2015, Eq. 9). We estimate the column density and mass uncertainty by varying the optical depth by 10%, which roughly corresponds to the 1σ brightness variation of our weakest continuum sources.

3.3. Masses

Based on the column density estimates in the previous sections, we can directly estimate the (cold) atomic and molecular mass toward the filament regions (see Table 5). We compute the masses by summing up the mass pixels above a column density threshold corresponding to significant emission or H I self-absorption. These thresholds are then also used to derive column density PDFs (see Sect. 4.1). The change in mass that comes with varying thresholds is relatively small compared to the uncertainties of the column density derivation itself.

The CNM mass traced by HISA corresponds to 3–9% of the total atomic gas mass, depending on the region and assumed spin temperature. The HISA mass fraction toward GMF38b is 0.09 and exceeds that found in all other filament regions. We recovered column density regions off the main cloud that is defined as GMF38.1-32.4b (see Fig. F.4). These regions might not be tightly associated with the molecular gas that defines the GMF. The HISA mass fraction reduces to 3–4% if we only take into account the gas in close proximity to the main molecular feature of the cloud (gas beyond the lowest contour in Fig. F.4 to within 0.2° offset), thus being comparable to other filament regions. However, this example also illustrates that we recover cold atomic gas structures that do not have a molecular counterpart. The cold phase of the atomic ISM appears to be much more widespread than the molecular gas in Fig. F.4. The masses of both GMF20 and GMF38a are similar to the masses found by Syed et al. (2020) and Wang et al. (2020b). Given that we assume a spin temperature of 26 K for GMF20 (instead of 20 K and 40 K in Syed et al. 2020), the derived mass falls within the mass range 4.6×10^3–1.3×10^4 $M_\odot$ obtained in Syed et al. (2020).

The molecular hydrogen mass is on the order of 10^4–10^5 $M_\odot$. The total atomic gas fraction shows large differences among the filament regions. The atomic gas mass is generally comparable to the molecular gas mass. However, for GMF54 the atomic gas seen in H I emission and HISA accounts to a total that is just one quarter of the total hydrogen mass. With respect to the molecular gas phase, the total atomic gas fraction is found to increase with Galactocentric distance on average (e.g., Nakanishi & Sofue 2016; Miville-Deschênes et al. 2017). In spite of having the largest Galactocentric distance in our sample, GMF54 appears to have used up much of the atomic gas in which it was embedded to transition into a more complete molecular gas phase.

4. Discussion

4.1. Column density PDF

We employ the probability density function (PDF) of the column density to investigate the physical processes acting within the filament regions. The shape of column or volume density PDFs are commonly used as a means to describe the underlying physical mechanisms of a cloud (e.g., Federrath & Klessen 2013; Padoan et al. 2014; Kainulainen et al. 2014; Schneider et al. 2015, 2022). Turbulence is considered to be the dominant driver of a cloud's structure if its PDF shows a log-normal shape. Furthermore, the width of a log-normal PDF is linked to the Mach number as it changes with the magnitude of the turbulence driving the cloud's structure (e.g., Padoan et al. 1997; Passot & Vázquez-Semadeni 1998; Padoan & Nordlund 2002; Kritsuk et al. 2007; Federrath et al. 2008; Konstandin et al. 2012; Molina et al. 2012; Kainulainen et al. 2014; Beattie et al. 2022), while noting that the turbulence driving scale and CNM-WNM mass ratio also affect the width of the PDF (Bialy et al. 2017).

Molecular clouds that are subject to the increasing effect of self-gravity develop high-density regions, producing a power-law tail in their PDF (e.g., Klessen 2000; Girichidis et al. 2014; Burkhart et al. 2017). Many star-forming molecular clouds have been confirmed to show such power-law tails (Kainulainen et al. 2009; Schneider et al. 2013, 2016, 2022). Even before the effects of gravity become dominant, gravitationally unbound clumps can exhibit power-law tails due to pressure confinement from the surrounding medium (Kainulainen et al. 2011).

We show in Fig. 6 the column density PDFs (N-PDFs) of all filament regions as traced by H I emission, HISA, and ^{13}CO. We take into account only column densities above the noise threshold of each tracer and find that the widths of each N-PDF do not change significantly when considering higher thresholds. The column density thresholds are $\sim 2\times10^{21}$ cm^{-2} for H I emission, $\sim 8\times10^{19}$ cm^{-2} for HISA, and $\sim 1\times10^{21}$ cm^{-2} for molecular hydrogen. We fit all column density PDFs with a log-normal function and report their widths in Fig. 6. Since we use a consistent way in deriving the PDFs, systematic differences between the distributions should be small, such that they can be well compared in relative terms. All N-PDFs are well described by a log-normal function.

Toward all filaments, the HISA-traced cold atomic gas shows a column density distribution that is broader than the narrow distribution of the diffuse atomic gas (left panels of Fig. 6). The mean column densities of molecular hydrogen are at least an order of magnitude higher than the column densities traced by HISA. We note the narrow distributions in the molecular gas phase toward GMF26 and GMF41 that are comparable to the HISA distributions. Toward the other filament regions, the molecular gas N-PDF has a larger width than the HISA N-PDF, highlighting the spatially more concentrated distribution of the molecular gas. The relatively narrow distributions of GMF26 and GMF41 might be related to the low excitation temperatures we find toward these clouds. This might be an indication of an early evolutionary stage where gravity has not yet become dynamically important. This is further supported by the low number of YSOs identified toward GMF41 (see Zhang et al. 2019).

The narrow log-normal shaped N-PDFs are commonly observed in the diffuse H I emission toward well-known molecular clouds (Burkhart et al. 2015; Imara & Burkhart 2016; Rebolledo et al. 2017; Schneider et al. 2022). The HISA N-PDFs that trace the CNM show broader distributions, indicative of the clumpy structure and higher degree of turbulence. Considering the column density PDFs, HISA appears to trace the cold atomic gas phase that connects the diffuse state of the atomic ISM with the transition of a cloud becoming molecular.

4.2. Mach number distribution

In the following, we derive the turbulent Mach number distributions using the constant HISA spin temperatures given in Table 3 and the excitation temperatures of ^{13}CO derived in Sect. D. Given the low temperature regime, we approximate the kinetic gas temperatures of the CNM and the molecular gas with the spin temperature and mean excitation temperature, respectively. We then estimate the three-dimensional scale-dependent Mach number of the filaments assuming isotropic turbulence with $\mathcal{M} = \sqrt{3}\,\sigma_{\mathrm{turb}}/c_s$, where σ_{turb} and c_s are the turbulent one-dimensional velocity dispersion and sound speed, respectively. The turbulent line width is calculated by subtracting the thermal line width contribution from the observed line width as

$$\sigma_{\mathrm{turb}} = \sqrt{\sigma_{\mathrm{obs}}^2 - \sigma_{\mathrm{th}}^2 - \sigma_{\mathrm{res}}^2}\,, \tag{7}$$

where σ_{obs}, and σ_{th} are the observed, and thermal velocity dispersion, respectively. For completeness, we also account for the broadening introduced by the spectral resolution σ_{res}. Since the thermal line width and sound speed scale as $T_{\mathrm{k}}^{1/2}$, the variation with spin temperature is moderate and does not change the Mach number significantly. Seifried et al. (2022) showed that the Mach number estimate inferred through HISA is robust and can be determined with an accuracy within a factor of ∼2.

Five of the six filament regions show very similar Mach number distributions (Fig. 7). The Mach number distributions traced by HISA are generally much narrower than those traced by ^{13}CO emission, and peak round $\mathcal{M} \sim 3-6$, with few values as high as ∼10. Our findings are in very good agreement with recent HISA observations (Burkhart et al. 2015; Nguyen et al. 2019; Wang et al. 2020b; Syed et al. 2020) and the simulations conducted by Seifried et al. (2022).

With the exception of GMF54, the molecular gas is highly supersonic, and has median Mach numbers between $\mathcal{M} \sim 7-10$. The molecular gas toward GMF54 is moderately supersonic and has a median Mach number around ∼5. The total observed line widths are generally small with a few ∼km s^{-1} (see Fig. E.11) and we do find the highest excitation temperatures up to ∼35 K in GMF54. As the HISA Mach numbers are also smallest toward GMF54, we consider this an imprint of a different physical mechanism dominating the dynamics of the cloud. In combination with the high excitation temperatures, low atomic mass fraction, and the most pronounced power-law tail in its column density distribution that we find in our sample, GMF54 appears to be at a much more advanced stage in its evolution, at which gravity seems to be the dominant driver of the cloud's dynamics.

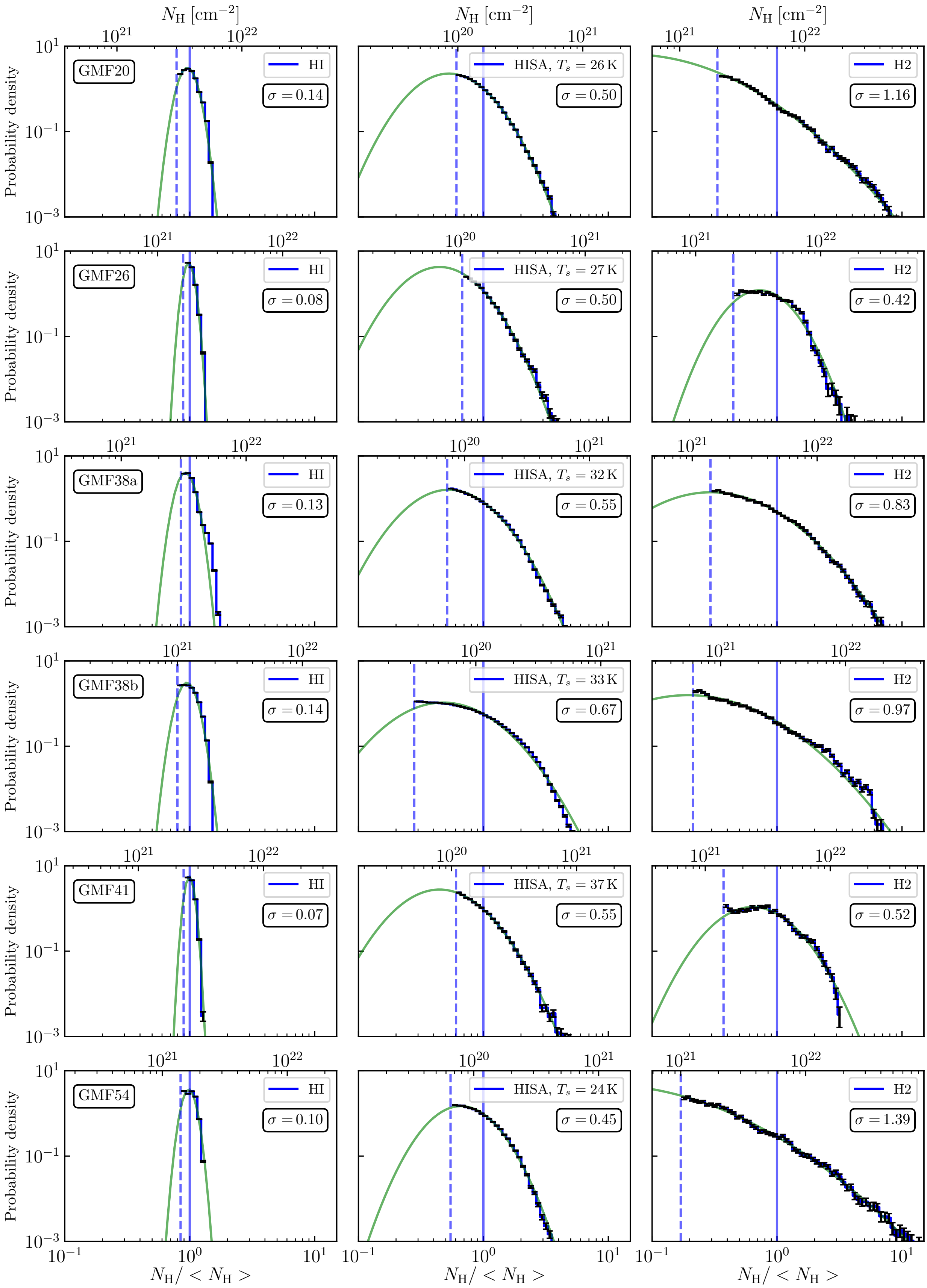

Fig. 6. Normalized column density PDFs of H I, HISA, and H_2 toward the giant filament regions. Each row shows the H I, HISA, and H_2 N-PDF toward one GMF region. *Left panels:* N-PDFs traced by H I emission. The distributions are derived from the H I column densities that have been corrected for optical depth and continuum emission. *Middle panels:* The N-PDFs of the gas traced by HISA. *Right panels:* H_2 N-PDFs traced by ^{13}CO in units of hydrogen atoms per cm^{-2}. The green curves indicate a log-normal fit to the distributions. The blue vertical dashed and solid lines mark the column density threshold and mean column density, respectively.

Table 5. Derived masses of the filament regions.

(1)	(2)	(3)	(4)	(5)	(6)
	M(HISA) [$M_\odot$]	M(H I) [$M_\odot$]	$M(H_2)$ [$M_\odot$]	f_{HISA}	f_{atomic}
GMF20.0-17.9	$7.5^{+1.3}_{-1.9} \times 10^3$	$2.7^{+0.2}_{-0.3} \times 10^5$	$2.3 \pm 1.2 \times 10^5$	0.03	0.55
GMF26.7-25.4	$1.1^{+0.3}_{-0.2} \times 10^3$	$3.0^{+0.3}_{-0.3} \times 10^4$	$5.5 \pm 2.8 \times 10^4$	0.04	0.36
GMF38.1-32.4a	$1.1^{+0.2}_{-0.3} \times 10^4$	$3.6^{+0.5}_{-0.7} \times 10^5$	$3.0 \pm 1.5 \times 10^5$	0.03	0.55
GMF38.1-32.4b	$6.2^{+1.3}_{-1.6} \times 10^3$	$6.2^{+0.7}_{-0.7} \times 10^4$	$5.5 \pm 2.8 \times 10^4$	0.09	0.55
GMF41.0-41.3	$5.3^{+0.7}_{-1.1} \times 10^2$	$1.4^{+0.2}_{-0.3} \times 10^4$	$1.7 \pm 0.9 \times 10^4$	0.04	0.46
GMF54.0-52.0	$3.3^{+0.7}_{-1.0} \times 10^2$	$5.2^{+1.1}_{-0.9} \times 10^3$	$1.6 \pm 0.8 \times 10^4$	0.06	0.26

Notes. The masses were calculated from the column density maps shown in Appendix F. Column (2) presents the mass of the cold atomic hydrogen traced by HISA. The uncertainties are statistical errors arising from the uncertainties in background fraction and spin temperature and do not include the systematic uncertainties due to the detection method. Column (3) shows the atomic hydrogen mass inferred from the optical depth and continuum corrected H I emission. The uncertainties are estimated from variations in the optical depth measurement. Column (4) gives the molecular hydrogen mass as traced by ^{13}CO emission along with a conservative 50% uncertainty owing to the large uncertainties in the CO-H_2 conversion. Column (5) gives the mass traced by HISA as a fraction of the total atomic gas mass M(HISA) + M(H I). Column (6) is the fraction of the total atomic gas mass traced by HISA and H I emission with respect to the total gas mass M(HISA) + M(H I) + $M(H_2)$.

Zhang et al. (2019) also find a star formation rate surface density that is among the highest in their sample of giant molecular filaments.

4.3. Spatial correlation between atomic and molecular gas

According to the classical idealized photodissociation region (PDR) picture, we would expect cold atomic gas to be spatially associated with its molecular counterpart (e.g., van Dishoeck & Black 1988; Andersson et al. 1991). We therefore employed the histogram of oriented gradients (HOG) tool[6] (Soler et al. 2019) to investigate the spatial correlation between HISA and ^{13}CO emission. The HOG method is based on machine vision to examine the spatial correlation between two spectral line tracers across their spectral domain. A detailed description of the HOG is given in Soler et al. (2019).

The underlying principle is the computation of intensity gradients in each velocity channel map of the respective line tracer. The relative angles between the intensity gradients of the line tracers (here HISA and ^{13}CO) are then computed for each pair of velocity channel maps. To statistically evaluate the significance of spatial correlation in terms of relative orientation between intensity gradients, the HOG uses the projected Rayleigh statistic V as a metric, which is a test to determine if the distribution is nonuniform and centered around 0°. It is tuned such that the sign of V is indicative of the angle distribution having a peak around $\theta = 0°$ ($V > 0$) or $\theta = 90°$ ($V < 0$) (Jow et al. 2018). The absolute value of V indicates the significance of that preferred orientation in the angle distribution. The projected Rayleigh statistic is therefore

$$V = \frac{\sum_{ij}^{m,n} w_{ij} \cos(2\theta_{ij})}{\sqrt{\sum_{ij}^{m,n} w_{ij}/2}} , \tag{8}$$

where the indices i and j run over the pixel locations in the two spatial dimensions for a given velocity channel and w_{ij} is the statistical weight of each angle θ_{ij}. We account for the spatial correlation between pixels introduced by the telescope beam and set the statistical weights to $w_{ij} = (\delta x/\Delta)^2$, where δx is the pixel size and Δ is the diameter of the derivative kernel that we used to calculate the gradients. We set the derivative kernel to $\Delta = 92''$, which is twice the beam size of the GRS.

We smoothed the extracted HISA cubes to a common beam size of 46″ and reprojected them onto the same spatial grid as the ^{13}CO data to run the HOG. Furthermore, we restricted the radial velocity range to $v_{LSR,low} - 25\,\mathrm{km\,s^{-1}}$ and $v_{LSR,up} + 25\,\mathrm{km\,s^{-1}}$ to save computational cost, where $v_{LSR,low}$ and $v_{LSR,up}$ are the lower and upper velocity limits given in Table 1, respectively. The extension of the velocity range $\pm 25\,\mathrm{km\,s^{-1}}$ provides a baseline measure of V (assuming there are signal-free channels over this velocity range). The projected Rayleigh statistic V should be ∼0 for these channels.

We use Monte Carlo sampling to propagate the errors introduced by the uncertainties in the flux measurement in each velocity channel (see e.g., Soler et al. 2020). For each velocity channel map, we generated ten random realizations per tracer with the same mean intensity and observational noise. Using this sampling, the uncertainty of the correlation can be determined by the variance of the correlation of different Monte Carlo realizations. Since we expect a contribution from non-Gaussian noise introduced by the observation, we report only $\geq 5\sigma$ confidence levels. We show in Fig. 8 the computed spatial correlation in terms of the projected Rayleigh statistic V for each filament region. We observe a strong spatial correlation between HISA and ^{13}CO toward GMF20, GMF38a, and GMF38b across multiple velocity channels. Even toward GMF26, GMF41, and GMF54 we detect significant spatial correlation in few velocity channels, despite little HISA detection. Although the spatial correlation between HISA and ^{13}CO appears to be poor in Fig. 3, we note that the intensity gradients of both tracers are in fact aligned where significant signal is overlapping, even if the intensity peaks do not exactly match. Small deviations in velocity are reflected by the width of the 2D distribution across the 1-to-1 correlation. We were able to reproduce the result obtained by Syed et al. (2020) toward GMF20, showing a significant spatial correlation within the entire filament region. Despite detecting significant spatial correlation between HISA and ^{13}CO toward all filament regions in our sample, we note that there might be little to no correlation when investigating subsections of filament regions, as observed in Syed et al. (2020) where the western part of the filament showed a strong agreement between the spatial

[6] https://github.com/solerjuan/astroHOG

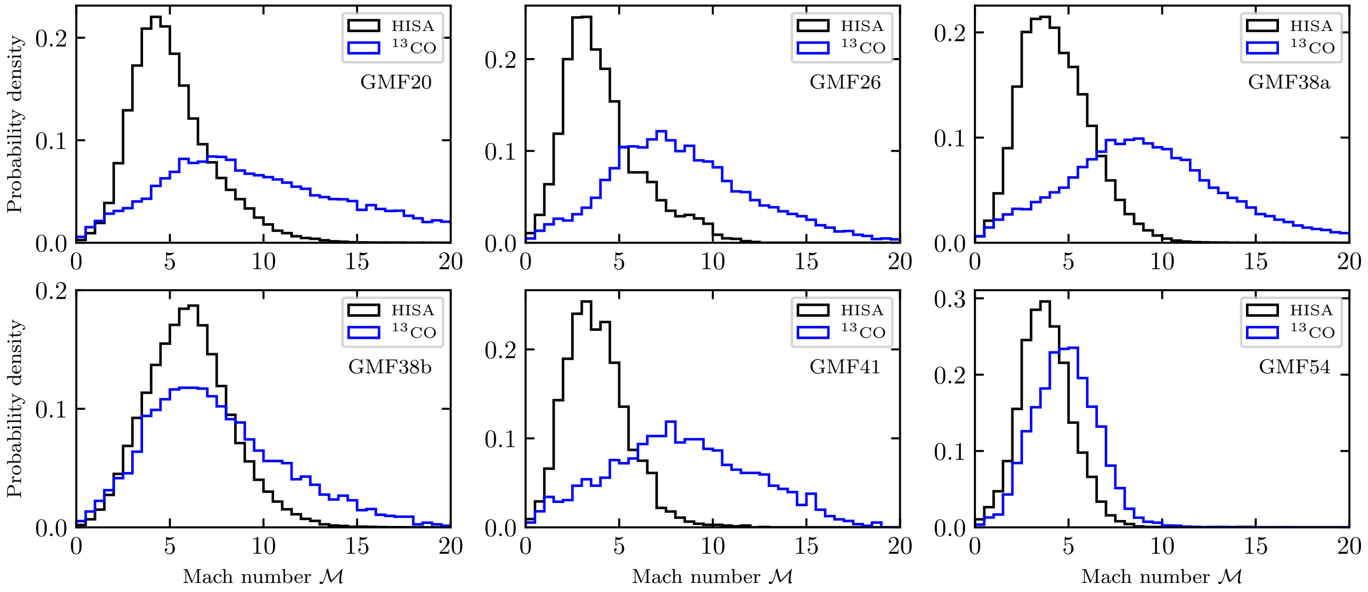

Fig. 7. Histograms of Mach numbers. The panels show for each of the six filament regions the normalized histograms of the Mach numbers of HISA and ^{13}CO in black and blue, respectively.

distribution of HISA and ^{13}CO while the eastern part entirely lacks correlation.

We conclude that the CNM traced by HISA generally appears to be associated with molecular gas in the giant filament regions on a large spatial scale. However, toward specific subregions within each filament systematic differences in spatial correlation can be evident that could be indicative of local events of star forming activity (see e.g., Soler et al. 2020, 2021).

5. Conclusions

We have investigated the properties of the cold atomic gas and molecular gas toward a sample of six giant molecular filament regions. We traced the cold atomic gas phase by H i self-absorption and obtained these features using the newly developed baseline extraction algorithm astroSaber. The kinematic properties of both the cold atomic gas and molecular gas were obtained using the spectral decomposition tool GaussPy+ (Riener et al. 2019). The main results are summarized as follows:

1. We detect HISA toward all giant filament regions. The mass traced by HISA accounts to a few percent of the total atomic hydrogen mass traced by H i self-absorption and emission. The total atomic mass is in most cases comparable to the molecular mass. Deviations from these mass fractions can be linked to different evolutionary stages of the clouds.
2. On a global average, the median centroid velocities of identified HISA and ^{13}CO appear to be similar, even though the agreement might be partially imposed by the restricted velocity range under consideration. A future large-scale HISA survey will facilitate an unbiased comparison of the global kinematics of HISA with molecular gas tracers. The well-resolved observed line widths of HISA are systematically larger than those of ^{13}CO. The CoAt gas traced by HISA is found to be moderately supersonic with Mach numbers of approximately a few, while the molecular gas within the majority of the filaments is driven by highly supersonic dynamics.
3. The derived column densities of the CoAt gas traced by HISA are on the order of $\sim10^{20}$ cm^{-2} and the column density distributions of the CoAt gas can be well described by a log-normal. The HISA-traced column density distributions are broader than the N-PDFs of the diffuse atomic gas traced by H i emission, indicating a spatially more concentrated cold gas distribution. The molecular gas has comparable or larger N-PDF widths than its cold atomic counterpart.
4. The recovered HISA features show a spatial correlation with the molecular gas toward many regions within the filaments. The Histogram of Oriented Gradients analysis (Soler et al. 2019) confirms a significant spatial correlation between HISA and ^{13}CO toward all filament regions at similar velocities.

Probing the cold atomic gas by means of H i self-absorption toward molecular clouds is a powerful tool to investigate the dynamical and physical interplay between the atomic and molecular gas during cloud formation. While molecular clouds are ideal targets to investigate the properties of HISA, we are looking to extend our findings and identify HISA without the bias of corresponding molecular line emission. We will investigate the global distribution of HISA in the inner Galactic plane in an upcoming paper.

Acknowledgements. J.S., H.B., and R.S.K. acknowledge support from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 138713538 – SFB 881 ("The Milky Way System", subproject A01, B01, B02, B08). This research was carried out in part at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration (80NM0018D0004). R.S.K. and J.D.S. thank for support from the European Research Council via the ERC Synergy Grant "ECOGAL" (project ID 855130). R.S.K. thanks support from the Heidelberg Cluster of Excellence (EXC 2181 - 390900948) "STRUCTURES", funded by the German Excellence Strategy, and from the German Ministry for Economic Affairs and Climate Action in project "MAINN" (funding ID 50OO2206). R.S.K. also is grateful for computing resources provided by the Ministry of Science, Research and the Arts (MWK) of the State of Baden-Württemberg through bwHPC and DFG through grant INST 35/1134-1 FUGG and for data storage at SDS@hd through grant INST 35/1314-1 FUGG. M.R.R. is a Jansky Fellow of the National Radio Astronomy Observatory. This research made use of the data from the Milky Way Imaging Scroll Painting (MWISP) project, which is a multiline survey in 12CO/13CO/C18O along the northern galactic plane with PMO-

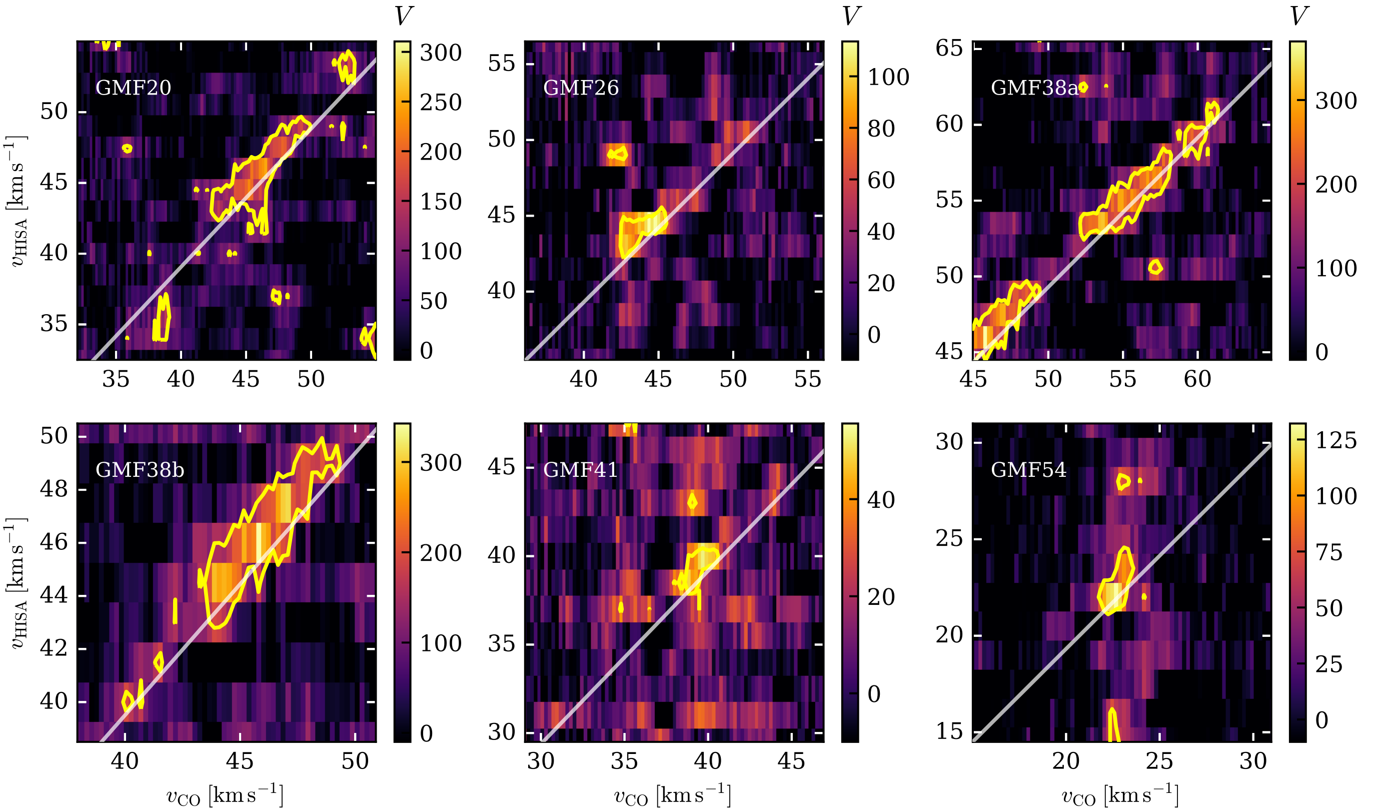

Fig. 8. Correlation in the distribution of HISA and ^{13}CO emission toward the six GMF regions as quantified by the projected Rayleigh statistic (V) in the HOG method (Soler et al. 2019). The panels present the computed spatial correlation between HISA and ^{13}CO across velocities in terms of the projected Rayleigh statistic V for each filament region. The values of V are indicated by the color bar to the right of each panel. The white line in each panel shows the 1-to-1 correlation. The yellow contours show the 5σ threshold estimated from the Monte Carlo sampling. Large values of V indicate a high spatial correlation. Values of V close to zero indicate a random orientation of the HISA structures with respect to ^{13}CO emission.

13.7m telescope. We are grateful to all the members of the MWISP working group, particularly the staff members at PMO-13.7m telescope, for their long-term support. MWISP was sponsored by National Key R&D Program of China with grant 2017YFA0402701 and CAS Key Research Program of Frontier Sciences with grant QYZDJ-SSW-SLH047.
We thank the anonymous referee for the detailed comments and appreciate the effort that has clearly gone into reviewing our work. We also thank M.B. Scheuck for his valuable contributions to the code development.

References

Allen, C. W. 1973, Astrophysical quantities
Anderson, L. D. & Bania, T. M. 2009, ApJ, 690, 706
Andersson, B. G., Wannier, P. G., & Morris, M. 1991, ApJ, 366, 464
Baek, S.-J., Park, A., Ahn, Y.-J., & Choo, J. 2015, Analyst, 140, 140
Beattie, J. R., Mocz, P., Federrath, C., & Klessen, R. S. 2022, MNRAS, 517, 5003
Beuther, H., Bihr, S., Rugel, M., et al. 2016, A&A, 595, A32
Bialy, S., Burkhart, B., & Sternberg, A. 2017, ApJ, 843, 92
Bialy, S. & Sternberg, A. 2019, ApJ, 881, 160
Bihr, S., Beuther, H., Ott, J., et al. 2015, A&A, 580, A112
Burkhart, B., Lee, M.-Y., Murray, C. E., & Stanimirović, S. 2015, ApJ, 811, L28
Burkhart, B., Stalpes, K., & Collins, D. C. 2017, ApJ, 834, L1
Burton, W. B. 1988, in Galactic and Extragalactic Radio Astronomy, ed. K. I. Kellermann & G. L. Verschuur, 295–358
Cox, A. N. 2000, Allen's astrophysical quantities
Dénes, H., McClure-Griffiths, N. M., Dickey, J. M., Dawson, J. R., & Murray, C. E. 2018, MNRAS, 479, 1465
Draine, B. T. 2011, Physics of the Interstellar and Intergalactic Medium - (Kassel: Princeton University Press)
Duarte-Cabral, A., Colombo, D., Urquhart, J. S., et al. 2021, The SEDIGISM survey: molecular clouds in the inner Galaxy, Monthly Notices of the Royal Astronomical Society, Volume 500, Issue 3, pp.3027-3049
Eilers, P. H. C. 2004, Analytical Chemistry, 76, 76
Federrath, C., Glover, S. C. O., Klessen, R. S., & Schmidt, W. 2008, Physica Scripta Volume T, 132, 014025
Federrath, C. & Klessen, R. S. 2013, ApJ, 763, 51
Feldt, C. 1993, A&A, 276, 531
Ferrière, K. M. 2001, Reviews of Modern Physics, 73, 1031
Field, G. B. 1965, ApJ, 142, 531
Field, G. B., Goldsmith, D. W., & Habing, H. J. 1969, ApJ, 155, L149
Fontani, F., Giannetti, A., Beltrán, M. T., et al. 2012, MNRAS, 423, 2342
Frerking, M. A., Langer, W. D., & Wilson, R. W. 1982, ApJ, 262, 590
Giannetti, A., Wyrowski, F., Brand, J., et al. 2014, A&A, 570, A65
Gibson, S. J., Taylor, A. R., Higgs, L. A., & Dewdney, P. E. 2000, ApJ, 540, 851
Girichidis, P., Konstandin, L., Whitworth, A. P., & Klessen, R. S. 2014, ApJ, 781, 91
Goldsmith, P. F., Heyer, M., Narayanan, G., et al. 2008, ApJ, 680, 428
Goldsmith, P. F. & Langer, W. D. 1999, ApJ, 517, 209
Goldsmith, P. F. & Li, D. 2005, ApJ, 622, 938
Goldsmith, P. F., Li, D., & Krčo, M. 2007, ApJ, 654, 273
Goodman, A. A., Pineda, J. E., & Schnee, S. L. 2009, ApJ, 692, 91
Haud, U. & Kalberla, P. M. W. 2007, A&A, 466, 555
Heeschen, D. S. 1954, AJ, 59, 324
Heeschen, D. S. 1955, ApJ, 121, 569
Heiles, C. & Troland, T. H. 2003, ApJ, 586, 1067
Imara, N. & Burkhart, B. 2016, ApJ, 829, 102
Jackson, J. M., Bania, T. M., Simon, R., et al. 2002, ApJ, 566, L81
Jackson, J. M., Rathborne, J. M., Shah, R. Y., et al. 2006, ApJS, 163, 145
Jow, D. L., Hill, R., Scott, D., et al. 2018, MNRAS, 474, 1018
Kabanovic, S., Schneider, N., Ossenkopf-Okada, V., et al. 2022, A&A, 659, A36
Kainulainen, J., Beuther, H., Banerjee, R., Federrath, C., & Henning, T. 2011, A&A, 530, A64
Kainulainen, J., Beuther, H., Henning, T., & Plume, R. 2009, A&A, 508, L35
Kainulainen, J., Federrath, C., & Henning, T. 2014, Science, 344, 183
Kalberla, P. M. W. & Dedes, L. 2008, A&A, 487, 951
Kalberla, P. M. W. & Kerp, J. 2009, ARA&A, 47, 27

Kavars, D. W., Dickey, J. M., McClure-Griffiths, N. M., Gaensler, B. M., & Green, A. J. 2003, ApJ, 598, 1048
Klessen, R. S. 2000, ApJ, 535, 869
Klessen, R. S. & Glover, S. C. O. 2016, Saas-Fee Advanced Course, 43, 85
Konstandin, L., Girichidis, P., Federrath, C., & Klessen, R. S. 2012, ApJ, 761, 149
Kritsuk, A. G., Norman, M. L., Padoan, P., & Wagner, R. 2007, ApJ, 665, 416
Krčo, M., Goldsmith, P. F., Brown, R. L., & Li, D. 2008, ApJ, 689, 276
Li, D. & Goldsmith, P. F. 2003, ApJ, 585, 823
Lindner, R. R., Vera-Ciro, C., Murray, C. E., et al. 2015, AJ, 149, 138
Liu, B., Wang, L., Wang, J., Peng, B., & Wang, H. 2022, PASA, 39, e050
McClure-Griffiths, N. M. & Dickey, J. M. 2007, ApJ, 671, 427
McKee, C. F. & Ostriker, J. P. 1977, ApJ, 218, 148
Miville-Deschênes, M.-A., Murray, N., & Lee, E. J. 2017, ApJ, 834, 57
Molina, F. Z., Glover, S. C. O., Federrath, C., & Klessen, R. S. 2012, MNRAS, 423, 2680
Nakanishi, H. & Sofue, Y. 2016, PASJ, 68, 5
Nguyen, H., Dawson, J. R., Lee, M.-Y., et al. 2019, ApJ, 880, 141
Padoan, P., Federrath, C., Chabrier, G., et al. 2014, in Protostars and Planets VI, ed. H. Beuther, R. S. Klessen, C. P. Dullemond, & T. Henning, 77
Padoan, P. & Nordlund, Å. 2002, ApJ, 576, 870
Padoan, P., Nordlund, A., & Jones, B. J. T. 1997, MNRAS, 288, 145
Passot, T. & Vázquez-Semadeni, E. 1998, Phys. Rev. E, 58, 4501
Pineda, J. E., Caselli, P., & Goodman, A. A. 2008, ApJ, 679, 481
Pineda, J. L., Langer, W. D., Velusamy, T., & Goldsmith, P. F. 2013, A&A, 554, A103
Planck Collaboration, Ade, P. A. R., Aghanim, N., et al. 2011, A&A, 536, A19
Ragan, S. E., Henning, T., Tackenberg, J., et al. 2014, A&A, 568, A73
Rebolledo, D., Green, A. J., Burton, M., et al. 2017, MNRAS, 472, 1685
Reid, M. J., Menten, K. M., Brunthaler, A., et al. 2019, ApJ, 885, 131
Riener, M., Kainulainen, J., Beuther, H., et al. 2020, A&A, 633, A14
Riener, M., Kainulainen, J., Henshaw, J. D., et al. 2019, A&A, 628, A78
Ruder, S. 2016, arXiv e-prints, arXiv:1609.04747
Schneider, N., André, P., Könyves, V., et al. 2013, ApJ, 766, L17
Schneider, N., Bontemps, S., Motte, F., et al. 2016, A&A, 587, A74
Schneider, N., Ossenkopf, V., Csengeri, T., et al. 2015, A&A, 575, A79
Schneider, N., Ossenkopf-Okada, V., Clarke, S., et al. 2022, A&A, 666, A165
Seifried, D., Beuther, H., Walch, S., et al. 2022, MNRAS, 512, 4765
Smith, R. J., Glover, S. C. O., Clark, P. C., Klessen, R. S., & Springel, V. 2014, MNRAS, 441, 1628
Soler, J. D., Beuther, H., Rugel, M., et al. 2019, A&A, 622, A166
Soler, J. D., Beuther, H., Syed, J., et al. 2020, A&A, 642, A163
Soler, J. D., Beuther, H., Syed, J., et al. 2021, A&A, 651, L4
Stahler, S. W. & Palla, F. 2005, The Formation of Stars (New York: John Wiley & Sons)
Stil, J. M., Taylor, A. R., Dickey, J. M., et al. 2006, AJ, 132, 1158
Strasser, S. & Taylor, A. R. 2004, ApJ, 603, 560
Su, Y., Yang, J., Zhang, S., et al. 2019, ApJS, 240, 9
Syed, J., Wang, Y., Beuther, H., et al. 2020, A&A, 642, A68
Tang, N., Li, D., Heiles, C., et al. 2016, A&A, 593, A42
Tikhonov, A. N. 1963, Soviet Math. Dokl., 4, 4
van Dishoeck, E. F. & Black, J. H. 1988, ApJ, 334, 771
Wang, Y., Beuther, H., Rugel, M. R., et al. 2020a, A&A, 634, A83
Wang, Y., Bihr, S., Beuther, H., et al. 2020b, A&A, 634, A139
Wilson, T. L., Rohlfs, K., & Hüttemeister, S. 2013, Tools of Radio Astronomy (Berlin Heidelberg: Springer Science & Business Media)
Wolfire, M. G., McKee, C. F., Hollenbach, D., & Tielens, A. G. G. M. 2003, ApJ, 587, 278
Zhang, F., Tang, X., Tong, A., et al. 2020, Spectroscopy Letters, 53, 53
Zhang, M., Kainulainen, J., Mattern, M., Fang, M., & Henning, T. 2019, A&A, 622, A52
Zucker, C., Battersby, C., & Goodman, A. 2018, ApJ, 864, 153

Appendix A: astroSaber optimization and parameters

Appendix A.1: Test data and smoothing optimization

We implemented a gradient descent method (see e.g., Ruder 2016) that uses generated mock data to find the optimal smoothing parameters λ_1 and λ_2. The $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$ parameter generally depends on the spectral resolution, noise level, line width of the absorption features, and on variations in the emission background. In a preparatory step astroSaber generates "pure emission" and "observed" mock data that are based on the actual observational data. The pure emission data represent emission line spectra that do not contain any absorption features and those are used as the test data that are to be recovered by the astroSaber algorithm. The observed data contain spectra where randomly generated (but known) absorption features were added to the pure emission data, and those are used as the input data for astroSaber. The mock data are generated by randomly selecting a defined number of spectra $N_{\rm train}$ taken from the real observational data on which the baseline extraction is to be applied later. The algorithm then uses asymmetric least squares smoothing with predefined and fixed parameters $(\lambda_1, \lambda_2) = (2.0, 2.0)$ (equivalent to one-phase smoothing with $\lambda_1 = 2.0$), without adding a residual, to smooth the spectra in the training set. The algorithm then adds a user-defined noise level to the spectra, thus creating spectra that will be used as pure emission data to be recovered. The reason for the smoothing in this preparatory step is to remove any dips that are present in the real data, such that the test data are free of absorption and that generated absorption features can be added anywhere in the spectrum. A moderate setting of the parameters p and $\boldsymbol{\lambda}$ in the preparation step does not heavily affect the optimization as these parameters are only used to generate data that show similarity to the overall structure of the real spectra as we show in Appendix A.2. We have taken samples of 200 spectra for each filament region to use as test data to find the optimal smoothing parameters. The randomly generated absorption spectra are created using Gaussian functions whose 1) amplitude, 2) mean, and 3) standard deviation parameters are drawn from normal distributions with the following mean and standard deviation (μ and σ) by default (see Table A.4): 1) the amplitude values follow a normal distribution with $\mu_{\rm amp} = 7\sigma_{\rm rms}$, and $\sigma_{\rm amp} = 1\sigma_{\rm rms}$, where $\sigma_{\rm rms}$ is the noise of the observational data, 2) the mean velocity values follow a normal distribution where the mean $\mu_{\rm mean}$ is set by the central velocity at which there is signal in each spectrum, its standard deviation is set accordingly, such that $3\sigma_{\rm mean}$ is at the edge of the signal range of the spectrum, 3) the magnitude of the $\boldsymbol{\lambda}$ parameters that is required for smoothing is crucially dependent on the width of the absorption features, so the standard deviation values of the absorption features drawn from a normal distribution have to be defined by the user. In the case of the THOR-H I data, Wang et al. (2020b) and Syed et al. (2020) report HISA FWHM values of ∼4 km s^{-1}. We have therefore set the mean and standard deviation of the line width distribution to $\mu_{\rm lw} = 4\,{\rm km\,s^{-1}}$ and $\sigma_{\rm lw} = 1\,{\rm km\,s^{-1}}$, respectively. The number of self-absorption components that are added to each spectrum are drawn from a normal distribution with $\mu_{\rm n} = 2.0$ and $\sigma_{\rm n} = 0.5$ by default, where all samples below 0.5 are set to 1.0 to add at least one self-absorption feature to each spectrum. In Appendix A.2 we discuss our fiducial parameter set.

Table A.1. Optimal smoothing parameters.

Source	λ_1	λ_2
GMF20.0-17.9	3.10	0.56
GMF26.7-25.4	3.40	0.46
GMF38.1-32.4a	2.76	0.50
GMF38.1-32.4b	2.76$^{(a)}$	0.50$^{(a)}$
GMF41.0-41.3	3.45	0.43
GMF54.0-52.0	3.70	0.39

Notes. The second column and third column give the best-fit λ_1 and λ_2 smoothing parameters obtained during the optimization step of astroSaber, respectively. Minor differences in these optimal parameters are expected due to different noise and fluctuations in the emission spectra.
$^{(a)}$ Since the optimal smoothing parameters are obtained from the same data cube as GMF38.1-32.4a, the λ values are the same.

Once the mock spectra have been generated, a gradient descent algorithm is run to find the optimal smoothing parameters λ_1 and λ_2. The gradient descent is designed to minimize the residual between the actual test data (pure emission mock spectra) and the baselines obtained by the astroSaber smoothing routine. Since we do not expect large variations in the emission spectra and absorption baselines within single HISA regions, we aim to find single λ_1 and λ_2 values that we then apply to the whole filament region in each case. We make use of the statistics of the training data and select the median of the reduced chi square values as the cost function for the optimization. The median value is robust against individual outliers in the training data and represents on average the best solution for the entire training dataset. The reduced chi square is only evaluated in channels where artificial absorption features have been added. More details about the gradient descent method applied in this paper are given in Appendix A.3.

With increasingly complex emission spectra containing multiple broad and narrow emission peaks as well as absorption features, adding a residual to a moderately smoothed spectrum has shown to give the best results for all the mock data that we have tested. In the subsequent analysis, we have therefore used the two-phase smoothing with two $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$ parameters and added a residual that is the difference between the very first major cycle iteration (using λ_1) and the last major cycle iteration (using λ_2). We note, however, that a simpler one-phase smoothing without adding a residual might give generally good results, depending on the signal-to-noise and complexity in the spectrum. The final smoothing parameters obtained in the optimization step of astroSaber that were used for the final baseline reconstruction are listed in Table A.1. The inferred smoothing parameters are similar toward all filament regions and compare well to each other. We expect small differences between the samples because of the training datasets containing different emission spectra and noise. These minor differences in the smoothing parameters only have a limited impact on the extraction results (see also Fig. A.1). In particular, the accuracy of the fit results does not heavily depend on λ_1. Figure A.1 shows similar accuracy in the fits for a range between 3–6.

Appendix A.2: Mock data parameter testing

We tested the final output of the optimization step using different parameter settings for generating the test and training datasets. As mentioned above, we apply a prior smoothing to test spectra in order to remove any pronounced absorption dips, such that absorption features can be added anywhere in the spectrum while generating the training data. The ideal test data should only contain pure emission features free of self-absorption. If an absorption feature were present in the test data, as might be the case if the real observations were used as test data, and an additional absorption feature were added at the same location, the optimization would falsely result in a smoothing parameter $\boldsymbol{\lambda}$ that is too small to recover the absorption. This might only affect a fraction of test data spectra as we randomly sample self-absorption features but these spectra would not yield reliable results during optimization. In a first test, we have therefore applied varying asymmetric least squares smoothing weights $\boldsymbol{\lambda}$ to the observational spectra and added noise to generate different test datasets. We used a one-phase smoothing ($\boldsymbol{\lambda} = \lambda_1$), without adding a residual, and smoothing values $\lambda_1 = 1.0, 2.0, 4.0$. In addition, we used a sample of real observations as test data to compare the optimization with the smoothed test datasets. In each case, we selected 100 spectra (same spectra for all tests) for the optimization. We then added self-absorption features and ran the optimization for all training datasets and report the optimal smoothing parameters in Table A.2. The optimal parameters of the various test and training datasets, that were generated using varying degrees of prior smoothing, show only marginal variations. Only the time for reaching a stable convergence might increase with less prior smoothing due to contamination of the pure emission test data, which can lead to fluctuations in the cost function (see Appendix A.3). The final baseline emission shows small differences, that are well within the noise of the observations. We conclude that a moderate prior smoothing of the data does not affect the final result of the optimization but can help achieve a stable convergence toward the optimal parameters more rapidly. We therefore chose to apply a prior smoothing weight of $(\lambda_1, \lambda_2) = (2.0, 2.0)$ (equivalent to one-phase smoothing) to generate the test and training data. However, astroSaber users can adjust these predefined smoothing parameters or opt to use real observations as test data instead.

Table A.2. Optimal smoothing parameters for test data with varying prior smoothing.

Prior smoothing weight	λ_1	λ_2
None	3.47	0.11
1.0	3.37	0.41
2.0	3.84	0.45
4.0	3.17	0.58

Notes. The first column gives the prior smoothing weight that was applied to the observations to generate pure emission test data. In the first row, no smoothing has been applied and real observations have been used as test data. The second and third column give the optimal smoothing parameters λ_1 and λ_2 that best recover the test data. Minor differences are expected due to noise fluctuations.

We have also varied the mean values of the Gaussian distributions, from which the parameters of self-absorption features are sampled, to investigate how this affects the final output of the astroSaber optimization. In each test, we vary one Gaussian distribution from which self-absorption parameters are drawn (i.e., either amplitude, line width, or number of self-absorption features), while fixing the remaining distributions to our fiducial values ($\mu_{\rm amp} = 7\sigma_{\rm rms}$, and $\sigma_{\rm amp} = 1\sigma_{\rm rms}$; $\mu_{\rm lw} = 4\,{\rm km\,s^{-1}}$ and $\sigma_{\rm lw} = 1\,{\rm km\,s^{-1}}$; $\mu_{\rm n} = 2.0$ and $\sigma_{\rm n} = 0.5$; see Appendix A.1). Table A.3 shows the final output of the optimization with varying self-absorption input parameters. The optimal smoothing weights (λ_1, λ_2) slightly increase with increasing absorption depth. When using a finite and fixed asymmetry weight p, larger smoothing weights are required to effectively smooth out strong absorption features and recover their baselines. We chose to set a mean amplitude of self-absorption features to $7\sigma_{\rm rms}$ in order to ensure a good balance between recovering significant (i.e., $\gtrsim 5\sigma_{\rm rms}$) self-absorption and retaining real emission signals. Stronger self-absorption features are then still identified and extracted with astroSaber, with their amplitudes being slightly underestimated.

As expected, the optimal smoothing weights increase with increasing line widths of self-absorption features. The output of astroSaber is most sensitive to the input line width of the absorption components, and therefore has to be provided by the user. As described in Appendix A.1, we set our fiducial value to $4\,{\rm km\,s^{-1}}$ that is in agreement with the reported line widths in Wang et al. (2020b) and Syed et al. (2020).

The output of astroSaber does not significantly change with the number of self-absorption components that are added to each test spectrum. The number of components effectively changes the number of samples that are tested and optimized in the training data. A larger number of samples allows a statistically more robust conclusion of optimal parameters, but comes at the cost of increased time to reach a stable convergence. We have therefore chosen to add an average of two components to each of the 200 test spectra, such that approximately 400 self-absorption features per training dataset are evaluated for the optimization. The self-absorption parameter distributions can also be modified by the user (see *optimization parameters* listed in Table A.4).

Appendix A.3: Momentum-driven gradient descent

The smoothing parameter $\boldsymbol{\lambda}$ (which is in our case a two-component vector by default) is tuned to maximize the fitness of the self-absorption baselines using a batch gradient descent with momentum (Ruder 2016). We define the median reduced chi square $\langle\chi^2_{\rm red}\rangle$ as the cost function C that we wish to minimize in order to achieve the highest goodness of fit result:

$$C(\boldsymbol{\lambda}) = \langle\chi^2_{\rm red}\rangle = \left\langle \frac{\sum_{i=1}^{N} \frac{(y_i - z_i(\boldsymbol{\lambda}))^2}{\sigma^2_{\rm rms}}}{N-k} \right\rangle , \qquad \text{(A.1)}$$

with y_i and $z_i(\boldsymbol{\lambda})$ denoting the data and baseline value at channel position i, respectively, N is the sample size (in this case the number of spectral channels containing self-absorption features), k denotes the degrees of freedom, which is in our case $k = 1$ for one-phase smoothing or $k = 2$ for two-phase smoothing, and $\sigma_{\rm rms}$ is the rms noise of the data.

In a classical gradient descent, updates to the smoothing weight $\boldsymbol{\lambda}$ are made by moving in the direction of greatest decrease in the cost function, that is $\Delta\boldsymbol{\lambda} = -\ell\,\nabla C(\boldsymbol{\lambda})$, where the learning rate ℓ controls the step size. Since the cost function is usually highly nonconvex, we implemented a gradient descent with added momentum to overcome local minima that might be due to noise or fluctuations in the spectra. Therefore, at the n-th iteration, the change in $\boldsymbol{\lambda}$ is given by

$$\Delta\boldsymbol{\lambda}^{(n)} = -\ell\,\nabla C(\boldsymbol{\lambda}) + \phi\,\Delta\boldsymbol{\lambda}^{(n-1)} , \qquad \text{(A.2)}$$

where the momentum ϕ controls the degree to which the previous step influences the current one. The gradient $\nabla C(\boldsymbol{\lambda})$ in Eq. (A.2) is defined as

$$\nabla C(\boldsymbol{\lambda}) = \begin{pmatrix} \frac{C(\lambda_1+\epsilon,\lambda_2)-C(\lambda_1-\epsilon,\lambda_2)}{2\epsilon} \\ \frac{C(\lambda_1,\lambda_2+\epsilon)-C(\lambda_1,\lambda_2-\epsilon)}{2\epsilon} \end{pmatrix} , \qquad \text{(A.3)}$$

Table A.3. Optimal smoothing parameters for test data with varying self-absorption parameters.

	(1)			(2)			(3)		
	$\mu_{\rm amp}$ [$\sigma_{\rm rms}$]			$\mu_{\rm lw}$ [km s^{-1}]			$\mu_{\rm n}$		
	5.0	7.0$^{(*)}$	9.0	2.0	4.0$^{(*)}$	6.0	2.0$^{(*)}$	3.0	4.0
λ_1	3.08	3.84	4.09	2.96	3.84	4.35	3.84	3.49	3.47
λ_2	0.29	0.45	0.60	0.20	0.45	1.26	0.45	0.80	1.15

Notes. The first test (1) describes training datasets with varying mean amplitude $\mu_{\rm amp}$ of self-absorption features. The second test (2) shows the optimal smoothing parameters for test data with varying mean line width $\mu_{\rm lw}$. The third test (3) gives the optimal smoothing with varying numbers of self-absorption components $\mu_{\rm n}$ that are added to each test spectrum. The remaining parameter distributions of each test are always set to the fiducial values $\mu_{\rm amp} = 7.0\,\sigma_{\rm rms}$, $\mu_{\rm lw} = 4.0\,{\rm km\,s^{-1}}$, and $\mu_{\rm n} = 2.0$.
$^{(*)}$ Fiducial values.

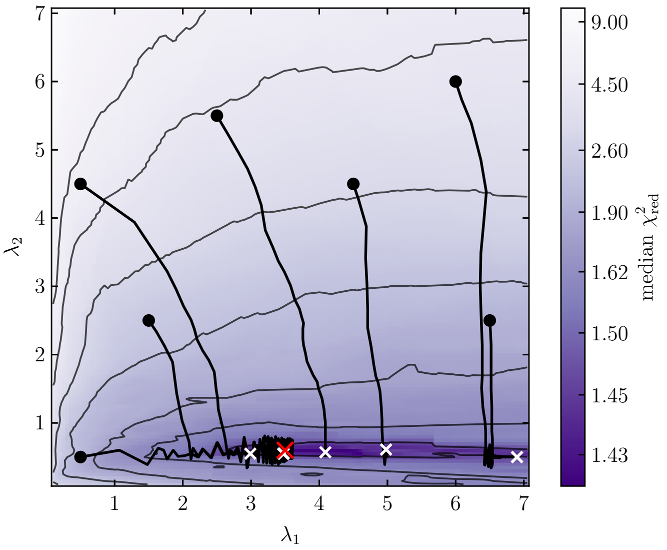

Fig. A.1. Smoothing parameter optimization using gradient descent. The map shows a sampled representation of the underlying $\boldsymbol{\lambda}$ parameter space in terms of the median value of the reduced chi square results. Initial values, tracks, and convergence locations of the (λ_1, λ_2) parameters during the optimization are represented by black circles, black lines, and white crosses, respectively. The red cross marks the global minimum in the sampled parameter space. Initial locations that start off too far from the global best solution $(\lambda_1 = 3.5, \lambda_2 = 0.6)$ might converge to local minima with less accurate fit results.

where we set the finite-difference step $\epsilon = 0.1$. Figure A.1 shows example tracks of $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$ when using the gradient descent with different initial values for λ_1 and λ_2 during the two-phase optimization on THOR-H I data. We find that small-scale local optima are ignored effectively during the search for large-scale optima.

Appendix A.4: Noise and signal range estimation

The signal ranges of the spectra are determined by borrowing parts of the noise estimation routine included in the GaussPy+ tool described in Sect. 2.3. For a detailed description, we refer the reader to Sect. 3.1.1 in Riener et al. (2019). The underlying assumptions to determine signal or noise ranges in the spectra are as follows: 1) the noise distribution is Gaussian, 2) the spectral channels are uncorrelated, and 3) the noise has a mean around zero. The assumption is that a spectrum containing white noise has on average an equal number of negative and positive channels. We can then estimate the probability of a given positive or negative feature observed in consecutive spectral channels to be caused by white noise (as opposed to be due to real signal) using a Markov chain. The routine to determine signal ranges then selects all features in spectral channels that have a probability to be caused by white noise of less than a user-defined threshold. We set this probability limit value to $P_{\rm limit} = 1\%$. In the case of the THOR-H I data with 185 channels per spectrum, we find that all features with more than 15 consecutive positive channels have a probability to be caused by noise of less than $P_{\rm limit} = 1\%$. To set the mean velocity in the velocity distribution of the absorption features, we additionally clipped the determined signal ranges by five channels on either side to ensure that the signal has sufficient intensity from which absorption features can be subtracted.

Appendix A.5: Symbols, astroSaber keywords, and default values

Depending on the scientific application and dataset, different parameters might be necessary to achieve satisfactory results from the astroSaber extraction. We have designed astroSaber such that most parameters can be easily modified by the user in order to allow a broad applicability of the algorithm. Table A.4 gives an overview of the parameter settings of astroSaber, listing their corresponding default values and symbols used throughout the text. In order to get first extraction results, only a small number of parameters (listed as *essential parameters*) need to be provided by the user. If the extraction results are deemed not satisfactory after adjusting these parameters, more *advanced settings* might be modified. While the optimization step should yield good results in most cases, the *optimization parameters* listed in Table A.4 also allow to customize the parameters used to generate the training data.

Appendix A.6: The astroSaber method and physical implications

When dealing with finite spectral resolution, one of the shortcomings of classical approaches using finite-difference derivatives is the strong dependence on sensitivity and line width. Noise fluctuations are greatly amplified in second (or higher order) derivatives of a spectrum. Only HINSA with line widths $\lesssim$1 km s^{-1} might be identified using this approach. It is then often assumed that there is a tight physical correlation in temperature between the cold H I gas traced by self-absorption and the molecular gas within a cloud. This correlation is then used to

Table A.4. astroSaber keywords mentioned throughout the text.

Symbol	Description	astroSaber keyword	Default
	Essential parameters		
Φ	Smoothing mode of the extraction (Sect. 2.2)	`phase`	`'two'`
λ_1	If `phase='two'` (=`'one'`), smoothing parameter of the first major cycle iteration (all major cycle iterations) (Sect. 2.2)	`lam1`	`None`
p_1	If `phase='two'` (=`'one'`), asymmetry weight of the first major cycle iteration (all major cycle iterations) (Sect. 2.2)	`p1`	0.9
λ_2	Smoothing parameter of the remaining major cycle iterations (Sect. 2.2)	`lam2`	`None`
p_2	Asymmetry weight of the remaining major cycle iterations (Sect. 2.2)	`p2`	0.9
σ_{rms}	Observational noise of the data (Sect. 2.2)	`noise`	`None`
	Advanced settings		
$\mathcal{R}_+$	Option to add residual (= absolute difference between first and last major cycle iteration; Sect. 2.2)	`add_residual`	`True`
n_{major}	Maximum number of major cycle iterations (Sect. 2.2)	`niters`	20
s_{thresh}	Multiplying factor of the noise setting the convergence threshold (Sect. 2.2)	`sig`	1.0
n_{converge}	Number of iterations of the major cycle to determine convergence (Sect. 2.2)	`iterations_for_convergence`	3
s_{signal}	Significance of emission to be considered in the extraction (Sect. A.4)	`check_signal_sigma`	6.0
Δv_{LSR}	Velocity range of the spectrum containing significant emission (set by σ_{signal}) to be considered in the baseline extraction (in units of km s^{-1}) (Sect. A.4)	`velo_range`	15.0
P_{limit}	The probability threshold of the Markov chain to estimate signal ranges in the spectra (Sect. A.4)	`p_limit`	0.01
	Optimization parameters		
$\mathcal{S}_{\mathrm{prior}}$	Option to apply prior asymmetric least squares smoothing to observations to generate test data (Sect. A.2)	`smooth_testdata`	`True`
N_{train}	Number of training spectra used for the optimization (Sect. A.1)	`training_set_size`	100
μ_{amp}	Mean of normal distribution to draw amplitude values from (in units of σ_{rms}; Sect. A.1)	`mean_amp_snr`	7.0
σ_{amp}	Standard deviation of normal distribution to draw amplitude values from (in units of σ_{rms}; Sect. A.1)	`std_amp_snr`	1.0
μ_{mean}	Mean of normal distribution to draw mean velocities from (Sect. A.1)	–	set by signal range
σ_{mean}	Standard deviation of normal distribution to draw mean velocities from (Sect. A.1)	–	set by signal range
μ_{lw}	Mean of normal distribution to draw line width (FWHM) values from (in units of km s^{-1}; Sect. A.1)	`mean_linewidth`	`None`
σ_{lw}	Standard deviation of normal distribution to draw line width (FWHM) values from (in units of km s^{-1}; Sect. A.1)	`std_linewidth`	`None`
μ_{n}	Mean of normal distribution to draw number of components from (Sect. A.1)	`mean_ncomponent`	2.0
σ_{n}	Standard deviation of normal distribution to draw number of components from (Sect. A.1)	`std_ncomponent`	0.5

Notes. A full documentation of all parameters is given at `https://github.com/astrojoni89/astrosaber`.

constrain the baselines of the self-absorption features (see Krčo et al. 2008), which is a reasonable approximation given the projected spatial correlation and small line widths often observed toward the central regions of molecular clouds. However, the tight correlation observed through HINSA is likely to trace only the cold (∼10 K) H I gas that is well mixed with the molecular gas in well-shielded regions (Li & Goldsmith 2003; Goldsmith & Li 2005), where the UV photo-dissociation rate of H_2 might become comparable to the cosmic ray dissociation in the central region of a cloud. By construction of the detection method, the CNM traced by HINSA likely results in atomic gas not being detected far beyond the inner regions of a molecular cloud (Goldsmith et al. 2007). However, once it is shown the HINSA-traced gas is coincident with ^{13}CO emission, the uncertainty in kinetic temperature should be considerably less than with our method.

With the newly developed algorithm astroSaber we identify H I self-absorption in an unbiased way, independent of the occurrence of molecular gas. The astroSaber algorithm can therefore complement the detection of CNM in the outer layers of molecular clouds or even the detection of CoAt gas that has no CO-bright molecular counterpart, which is likely to have larger line widths and would otherwise be missed by a second derivative approach, as we show in Appendix B.

In the following, we discuss some of the limitations and ways that might boost the performance of the astroSaber routine. Since the observed spectra also contain noise where there is signal, the baseline smoothing slightly overestimates the baselines within signal ranges as it also weights the noise asymmetrically that is superposed with the emission. One way to take the noise within signal ranges into account is to adjust the weightings in Eq. (2) according to the mean and standard deviation of the positive and negative difference values between the spectrum and the baseline after each iteration (see e.g., Baek et al. 2015; Liu et al. 2022). However, we are only interested in ranges where we expect self-absorption to be present. As we tune the smoothing parameter such that significant dips in the spectra are smoothed out, any variation of the obtained baselines within emission ranges without absorption should be limited to the noise. Any features in those ranges are therefore not identified as significant absorption anyway.

As we show in Appendix C, the centroid velocities recovered by astroSaber and Gaussian fitting show little deviation from the input velocities within our test environment. The distribution of centroid velocity differences has a mean and standard deviation of $-0.01\,\mathrm{km\,s^{-1}}$ and $0.35\,\mathrm{km\,s^{-1}}$, respectively. Based on the findings by Wang et al. (2020b) and Syed et al. (2020), the input line widths of $\sim 4\,\mathrm{km\,s^{-1}}$ (FWHM) could be recovered with a standard deviation of $\sim 1\,\mathrm{km\,s^{-1}}$. The larger scatter in line widths is likely due to employing a constant smoothing parameter for both narrow and broad absorption components. The difference in amplitude shows the largest scatter around the mean as a single smoothing parameter is used for the entire region.

Since we set a constant λ value for all spectra in each field, we account for significant broad absorption features by performing multiple iterations to obtain their baselines. However, depending on the number of iterations, the final baseline might not reflect the original spectrum within emission ranges accurately as in each iteration an updated baseline is used as input for the next major cycle iteration. One way to address this is to not have multiple major cycle iterations but instead adjust the smoothing parameter channel by channel as broader absorption features would require more iterations than narrow ones at constant λ.

With a single iteration, broader absorption features require a larger smoothing parameter λ if the asymmetry weighting for negative differences (i.e., absorption dips) is constant but nonzero. This baseline "drag" because of nonzero weighting could be corrected for if we introduced another coefficient vector $\boldsymbol{\alpha}$ that adjusts the smoothing parameter in Eq. (1) for each channel in the spectrum, with its components being defined as

$$\alpha_i = \frac{\mathrm{abs}(y_i - z_i)}{\max(\mathrm{abs}(\mathbf{y} - \mathbf{z}))}\,, \tag{A.4}$$

where the numerator is the absolute difference of the spectrum and baseline at channel i, and the denominator is the maximum of the absolute differences in the spectrum (see e.g., Zhang et al. 2020). Equation (1) would then change to

$$F(\mathbf{z}) = (\mathbf{y} - \mathbf{z})^{\top}\mathbf{W}(\mathbf{y} - \mathbf{z}) + (\lambda\,\alpha)\,\mathbf{z}^{\top}\mathbf{D}^{\top}\mathbf{D}\mathbf{z}\,. \tag{A.5}$$

This could be a way to tune the smoothing parameter to an optimum without the need of having to perform multiple iterations. Since the weight curve and smoothing coefficients would also be fixed in that case, the smoothing parameter λ would still be the only parameter to be optimized.

In summary, the results could be improved by utilizing parameterized smoothing and asymmetry weights. Ultimately, these training and test data could then be used to feed a machine learning algorithm that sets an optimized smoothing parameter for each spectrum. However, as we have achieved good results with astroSaber in its current state, that already outperforms our traditional approach of using polynomial fits to specific ranges of the emission spectra, we leave the optimization of performance and efficiency to future investigations.

Appendix B: Classical second derivative approach

Another way to identify HISA features uses the second derivative of the observed H I spectrum as described in Krčo et al. (2008). Pronounced self-absorption features would therefore become readily apparent as signatures in the second derivative representation of the spectrum. In the following, we discuss the limitations of this method and how astroSaber overcomes the issues imposed by finite spectral resolution and noise.

Calculating the second (or higher) derivatives using finite-difference techniques might not always give reliable results as noise in the observational data is greatly amplified. This is illustrated in Fig. B.1. While the top panel shows a mock-H I spectrum including two self-absorption components that does not contain noise, the bottom panel presents the same spectrum with added noise that is comparable to the THOR data (same spectrum as in Fig. 1). The green spectrum in each panel shows the finite-difference second derivative of the spectrum. For the noise-less data, the narrow HISA component emerges as a signature in the second derivative. Although less pronounced, even the broader absorption feature can be identified through an enhancement in its second derivative. On the other hand, given the observed spectrum that contains noise (lower panel in Fig. B.1) the HISA components do not become visible as noise fluctuations dominate the second derivative of the spectrum. To overcome this, regularized differentiation can be used to mitigate the effect of noise fluctuations. It is a method of regularization of ill-posed problems that commonly occur in models with large numbers of parameters or inverse-solving during optimization. For example, this so-called Tikhonov regularization (Tikhonov 1963) may be used to enforce smoothness of a given vector, giving preference to solutions that minimize the second derivative.

In a similar way, astroSaber uses this type of regularization when it introduces a penalty term to the (asymmetric) least squares function that minimizes the second derivative (see Eq. 1). This is demonstrated in a simplified way in the lower panel of Fig. B.1. The dashed blue spectrum shows a (in this case symmetric) least squares solution to the mock-H I spectrum with a regularization term as in Eq. (1). The second derivative of the smooth representation of the spectrum now responds to the narrow absorption feature (blue spectrum) and shows a peak. However, the broader feature cannot be identified in the second derivative.

For conceptual purposes, if we assume that a self-absorption feature is Gaussian $g(v_{\rm LSR})$, the second derivative of the feature will be

$$\frac{{\rm d}^2 g(v_{\rm LSR})}{{\rm d}v_{\rm LSR}^2} = \left(-\frac{1}{\sigma^2} + \frac{v_{\rm LSR}^2}{\sigma^4}\right) g(v_{\rm LSR})\,, \tag{B.1}$$

where σ is the standard deviation of the Gaussian, thus showing a strong dependence on line width. Narrow self-absorption features can then be identified through their second derivative more easily. In conclusion, the second derivative alone only works reliably well for high sensitivity, sufficient spectral resolution, and HISA line widths that are much smaller than the average emission component. Furthermore, even if the spectral ranges of HINSA (Li & Goldsmith 2003; Goldsmith & Li 2005; Goldsmith et al. 2007) were determined with second derivatives, the baselines would still need to be inferred using, for example, polynomial fits or making physical assumptions of the HINSA properties (see Krčo et al. 2008). By introducing an asymmetry weighting and an optimized regularization term, that simultaneously mitigates the undesirable effect of noise fluctuations, astroSaber is able to recover baselines while identifying absorption dips without the necessity of assuming a fitting function or a tight physical correlation between the cold H I gas and the molecular gas.

Appendix C: Robustness of kinematics

To test how well the kinematics of the recovered absorption features match the input data, we ran astroSaber on an example data cube taken from a subsection of GMF20.0-17.9 (see Ragan et al. 2014; Syed et al. 2020). This example cube is also made available along with astroSaber source code. We created mock data as described in Appendix A.1 containing 100 test spectra where known self-absorption have been added. We then ran astroSaber to extract the self-absorption baselines and spectra after finding the optimal smoothing parameters. To obtain the kinematic properties of the extracted self-absorption features, we fit several Gaussian components to the self-absorption spectra, depending on the number of components that were added. In total, 207 self-absorption components have been added while generating the mock spectra.

In Fig. C.1 we present histograms showing the residuals between the true amplitudes, centroid velocities, line widths (FWHM) and their respective fit results. All distributions show a mean around zero. The line widths show a weak systematic trend as the mean of the residuals is 0.25 km s^{-1}, implying that the line width fits slightly underestimate the true line width. As expected, the amplitude distribution shows the largest dispersion as we use only one set of smoothing parameters for the entire region. The recovered centroid velocities are very robust as the histogram shows a mean around zero and a standard deviation of 0.35 km s^{-1}.

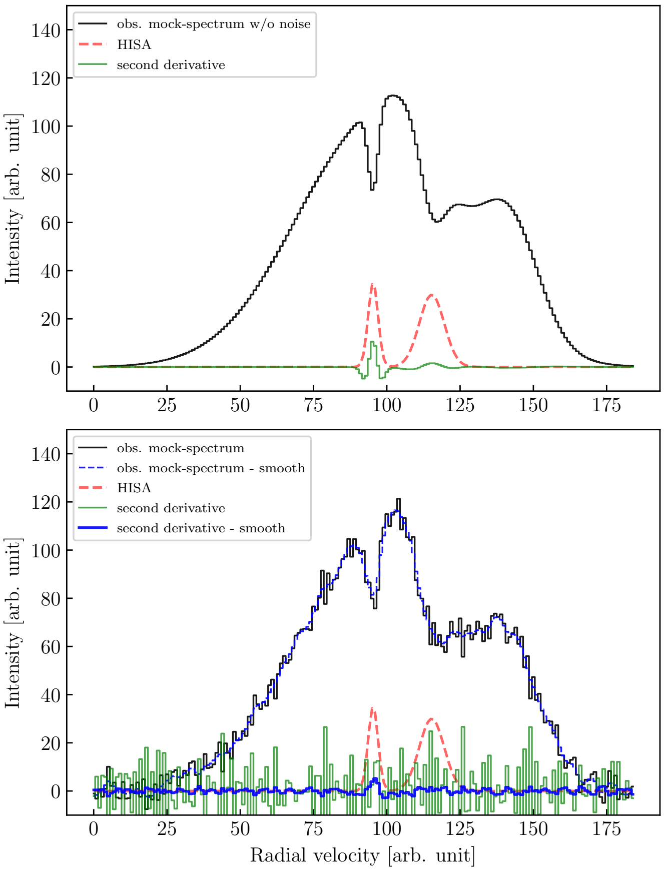

Fig. B.1. Second derivative representation as a means to identify self-absorption. *Top panel:* The black mock spectrum represents the H I emission spectrum, with two self-absorption features superposed (red dashed components) and without any observational noise. The green spectrum shows the second derivative of the black mock spectrum, obtained from the finite differences between spectral channels. *Bottom panel:* The black mock spectrum represents the H I emission spectrum, with two self-absorption features superposed (red dashed components) and with added noise that is comparable to the noise of the THOR-H I observations (same spectrum as in Fig. 1). The green spectrum shows the second derivative of the black mock spectrum, obtained from the finite differences between spectral channels. The dashed blue spectrum represents a regularized least squares solution to the H I spectrum, which minimizes the second derivative. The corresponding second derivative is shown in blue, which is now less affected by noise fluctuations.

Appendix D: H$_2$ column density traced by ^{13}CO emission

Assuming that ^{12}CO is becoming optically thick toward these GMFs, we estimated the column density of molecular hydrogen from ^{13}CO emission. In the optically thin limit, the ^{13}CO column density is computed by (Wilson et al. 2013)

$$N(^{13}{\rm CO}) = 3.0 \times 10^{14}\,\frac{\int T_{\rm B}(v)\,{\rm d}v}{1-\exp(-5.3/T_{\rm ex})}\,, \tag{D.1}$$

where $N(^{13}{\rm CO})$ is the column density of ^{13}CO molecules in cm^{-2}, dv is in units of km s^{-1}, $T_{\rm B}$ and $T_{\rm ex}$ are the brightness temperature and excitation temperature of the ^{13}CO line in units of Kelvin, respectively. Under the assumption that the excitation temperature $T_{\rm ex}$ of ^{12}CO and ^{13}CO are the same in LTE, we computed the ^{13}CO excitation temperature from ^{12}CO line emission. Both the ^{12}CO and ^{13}CO data are taken from the high-resolution

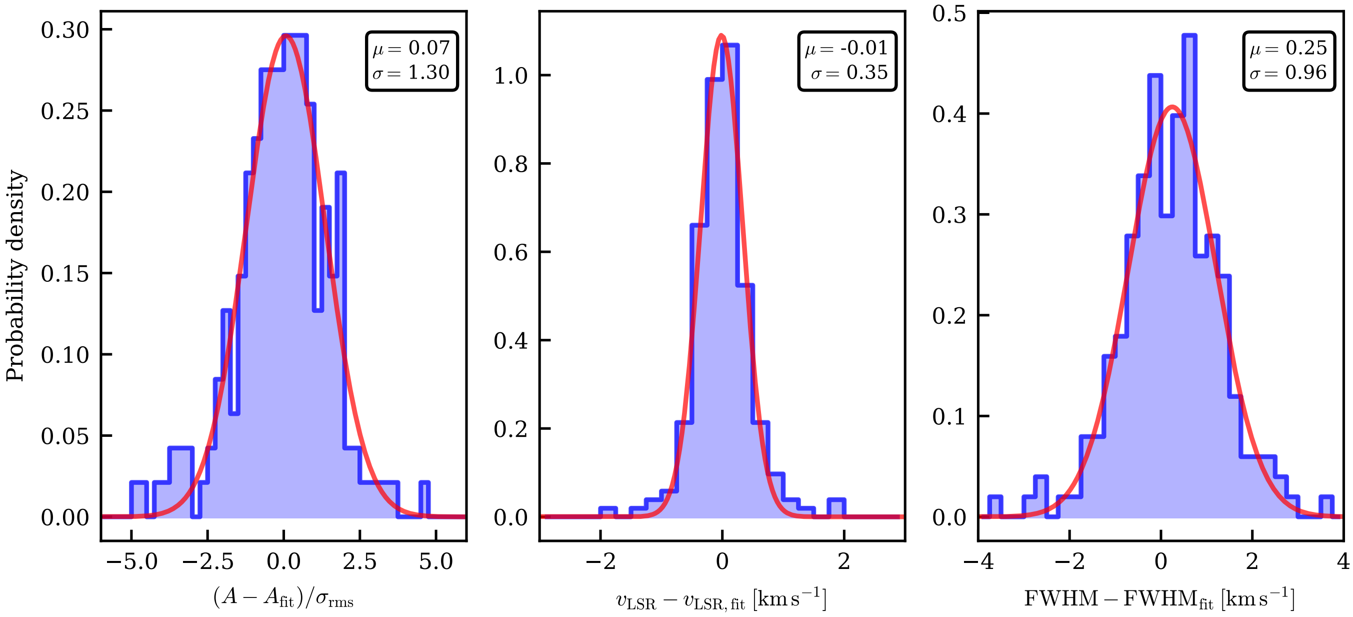

Fig. C.1. Histograms of residuals between input features and their respective fit results. *Left panel:* The distribution shows the residuals between the amplitudes that were used to generate self-absorption features and the fit results (in units of the observational noise) after running astroSaber. *Middle panel:* The distribution shows the residuals between the input velocities of self-absorption features and the recovered fit velocities. *Right panel:* Similarly, the distribution in the right panel shows the residuals of the line widths. The red curve in each panel shows a Gaussian fit to the distribution.

survey MWISP (Su et al. 2019) to compute the excitation temperatures and column densities, as described in Sect. 2.1. The excitation temperature is computed as (Wilson et al. 2013)

$$T_{\rm ex} = 5.5 \cdot \left[\ln\left(1 + \frac{5.5}{T_{\rm B}^{12} + 0.82}\right)\right]^{-1} , \tag{D.2}$$

where $T_{\rm B}^{12}$ is the brightness temperature of the ^{12}CO line in units of Kelvin. To calculate the excitation temperature for each voxel, we reprojected the ^{12}CO data cubes onto the same spectral grid as the ^{13}CO data.

We set a lower limit to the excitation temperatures for regions where the ^{12}CO brightness temperatures reach the 5σ noise level. We can then derive the optical depth of the ^{13}CO line from the excitation and brightness temperature, using (see e.g., Wilson et al. 2013; Schneider et al. 2016)

$$\tau = -\ln\left[1 - \frac{T_{\rm B}}{5.3} \cdot \left(\left[\exp\left(\frac{5.3}{T_{\rm ex}}\right) - 1\right]^{-1} - 0.16\right)^{-1}\right] . \tag{D.3}$$

We then estimated a lower limit of the optical depth for ^{13}CO brightness temperatures at the 5σ noise level and the highest excitation temperatures we find toward each GMF region. The lower and upper limits to the excitation temperatures as well as the lower limits of the optical depth are listed in Table 4 for each source. To account for high optical depth effects, we employ a correction factor by replacing the integral in Eq. (D.1) with (Frerking et al. 1982; Goldsmith & Langer 1999)

$$\int T_{\rm B}(v)\,{\rm d}v \rightarrow \frac{\tau}{1 - e^{-\tau}} \int T_{\rm B}(v)\,{\rm d}v\,. \tag{D.4}$$

This correction factor is accurate to 15% for $\tau < 2$.

Appendix E: Kinematics maps

The kinematic properties are presented in this section. The following maps show the fit peak velocities and line widths (FWHM) obtained with GaussPy+ for both HISA and ^{13}CO emission toward all remaining filament regions. If multiple components are identified within the velocity range of the filament, we only show the component with the lowest peak velocity.

Appendix F: Column density maps

The column density maps are presented in this section. The following maps show the column density maps for both HISA and H_2 as traced by ^{13}CO emission integrated over the velocity range of the respective filament region. Details about the column density derivation of each tracer can be found in Sect. 3.2.

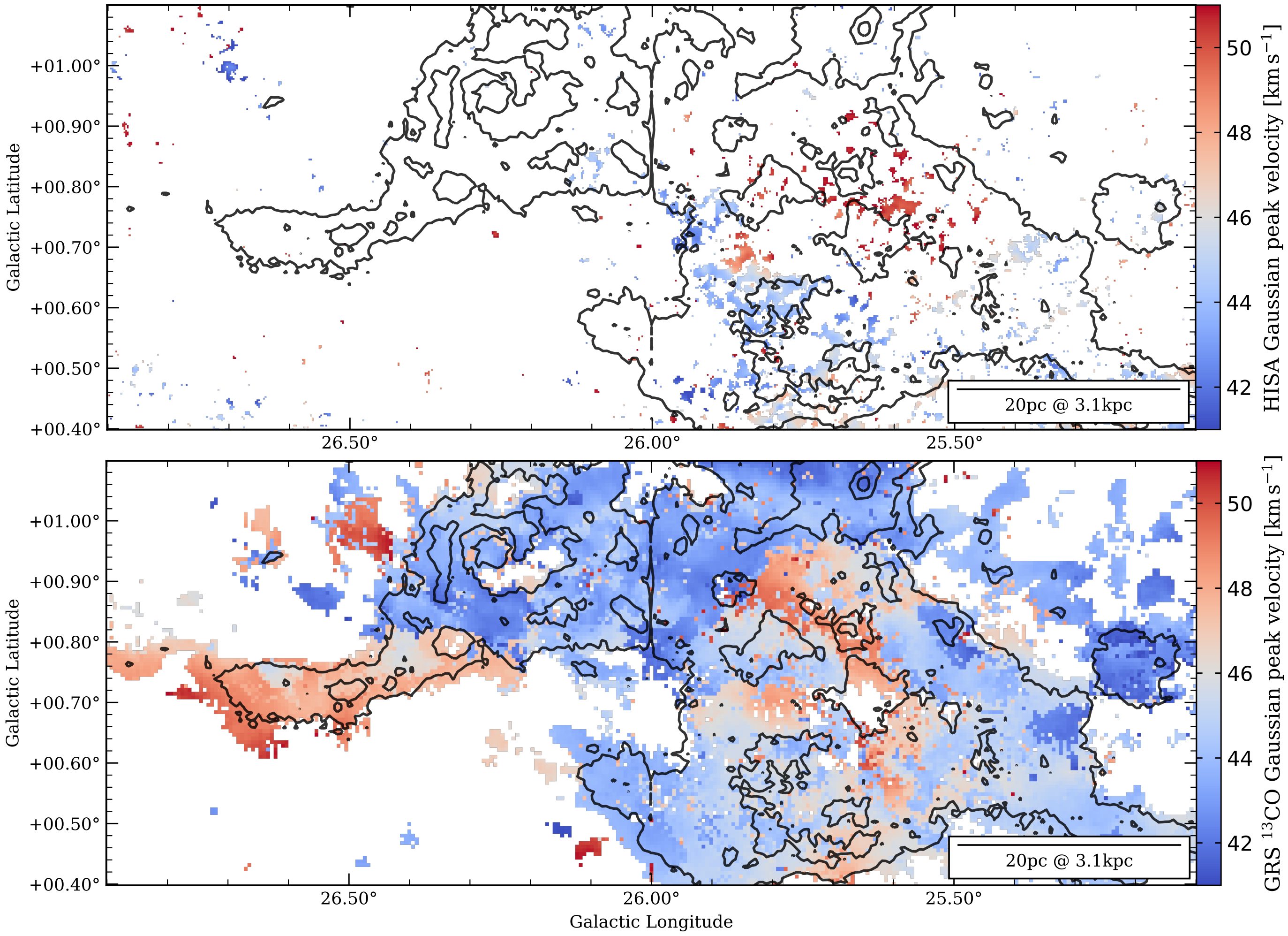

Fig. E.1. Fit peak velocity toward GMF26. These maps show the peak velocities of fit components derived from the GaussPy+ decomposition of the spectra. If multiple components are present in a single pixel spectrum within the velocity range of the filament region, the component with the lowest peak velocity is shown. The black contours in both panels show the integrated GRS ^{13}CO emission at the levels 6.0, 12.0, 24.0, and 34.0 K km s^{-1}. The contour feature at longitude $\ell = 26°$ is an artifact in the observational data. *Top panel:* Fit HISA peak velocity. *Bottom panel:* Fit ^{13}CO peak velocity.

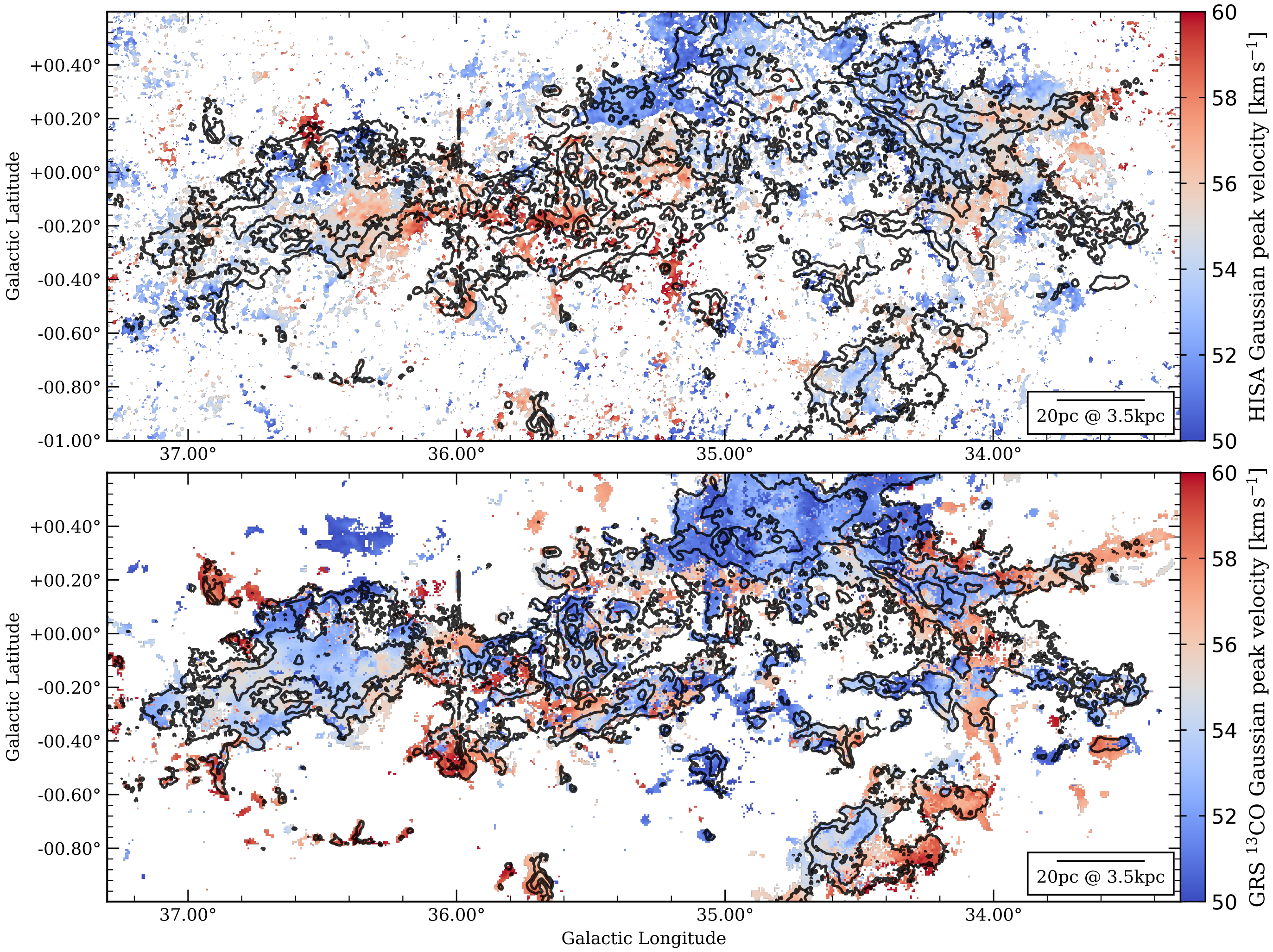

Fig. E.2. Fit peak velocity toward GMF38a. These maps show the peak velocities of fit components derived from the GaussPy+ decomposition of the spectra. If multiple components are present in a single pixel spectrum within the velocity range of the filament region, the component with the lowest peak velocity is shown. The black contours in both panels show the integrated GRS ^{13}CO emission at the levels 5.0, 10.0, 20.0, and 30.0 K km s^{-1}. The contour feature at longitude $\ell = 36°$ is an artifact in the observational data. *Top panel:* Fit HISA peak velocity. *Bottom panel:* Fit ^{13}CO peak velocity.

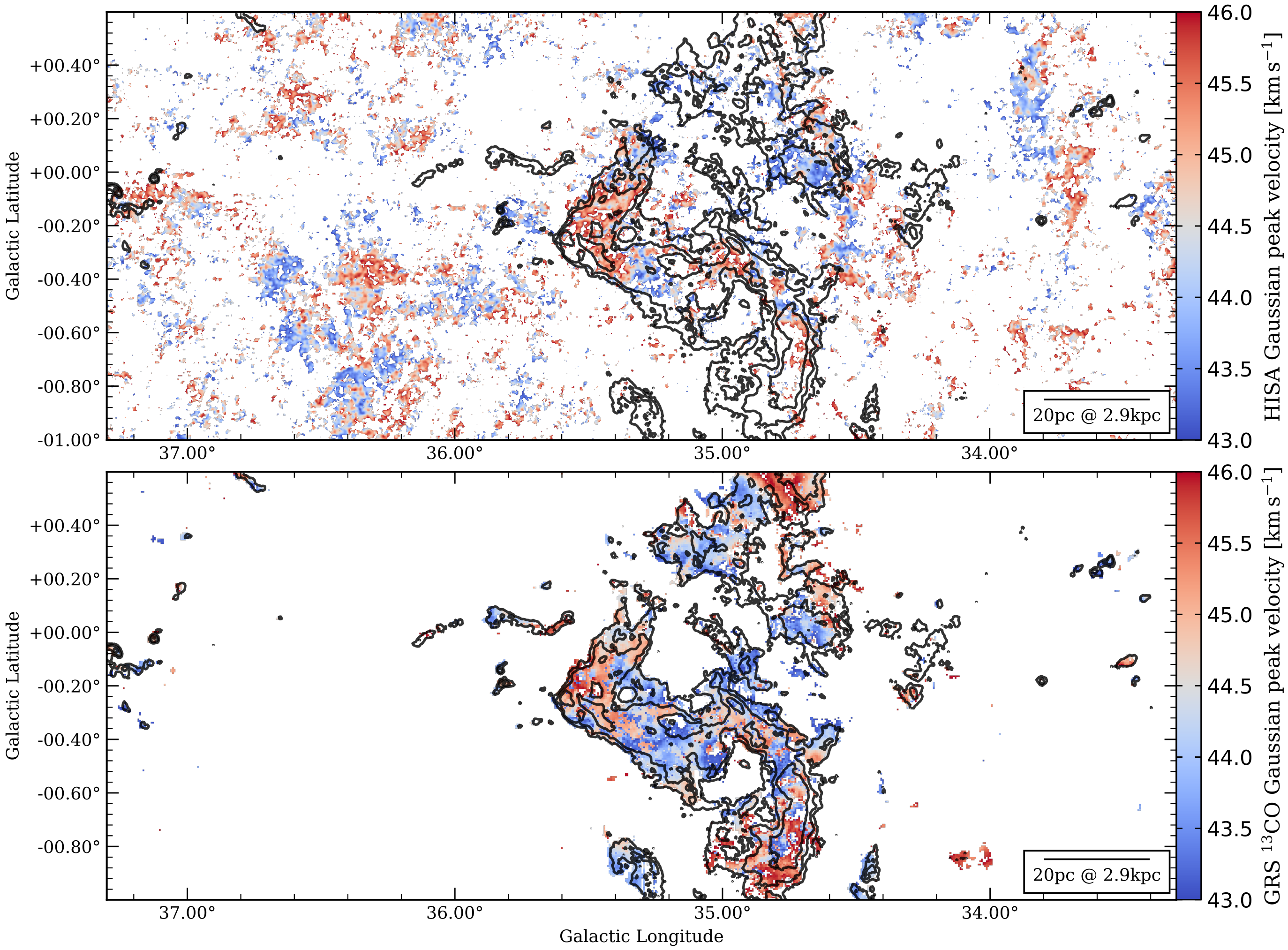

Fig. E.3. Fit peak velocity toward GMF38b. These maps show the peak velocities of fit components derived from the GaussPy+ decomposition of the spectra. If multiple components are present in a single pixel spectrum within the velocity range of the filament region, the component with the lowest peak velocity is shown. The black contours in both panels show the integrated GRS ^{13}CO emission at the levels 2.5, 5.0, 10.0, and 20.0 K km s^{-1}. *Top panel:* Fit HISA peak velocity. *Bottom panel:* Fit ^{13}CO peak velocity.

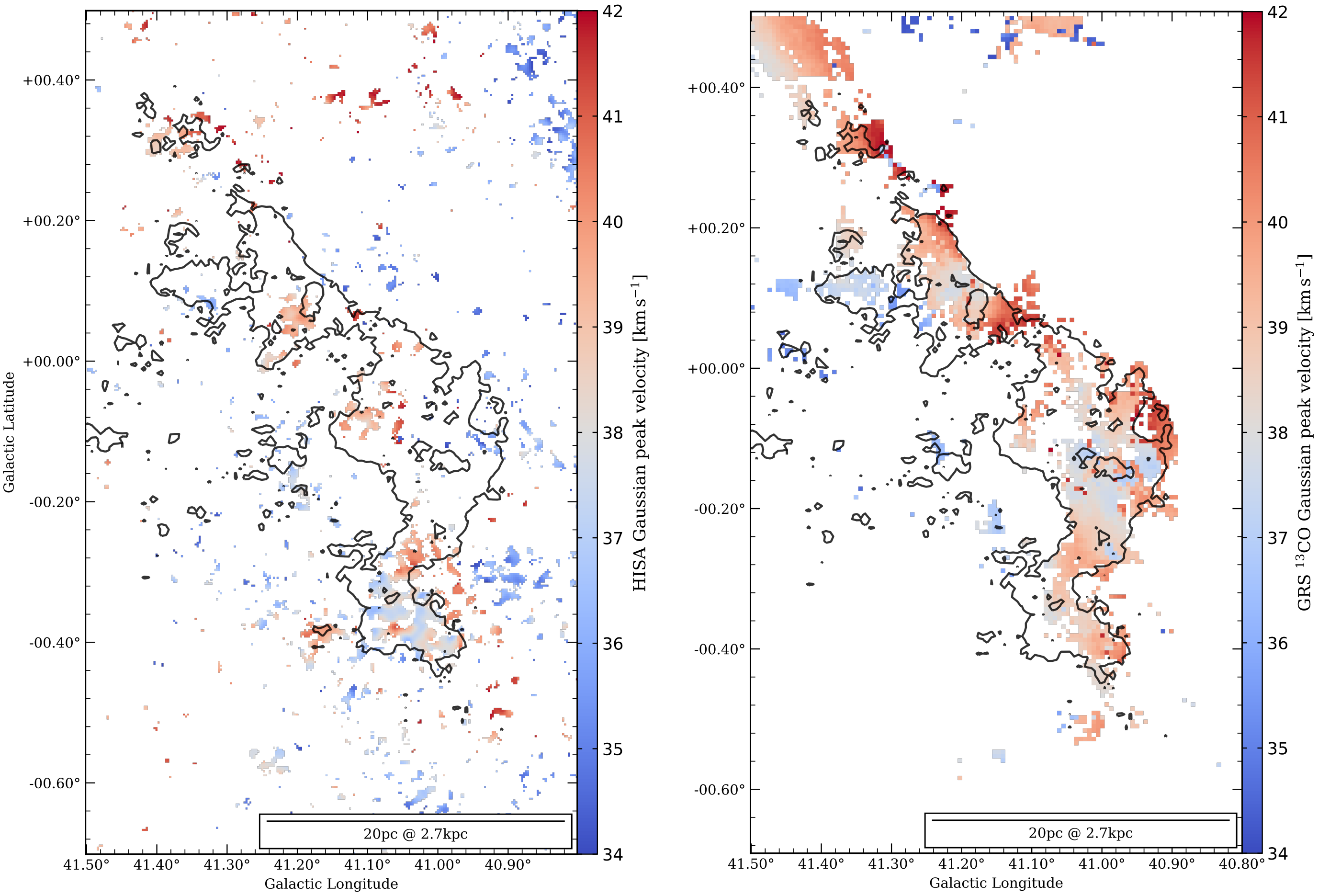

Fig. E.4. Fit peak velocity toward GMF41. These maps show the peak velocities of fit components derived from the GaussPy+ decomposition of the spectra. If multiple components are present in a single pixel spectrum within the velocity range of the filament region, the component with the lowest peak velocity is shown. The black contours in both panels show the integrated GRS ^{13}CO emission at the levels 6.0, 12.0, 24.0, and 34.0 K km s^{-1}. *Top panel:* Fit HISA peak velocity. *Bottom panel:* Fit ^{13}CO peak velocity.

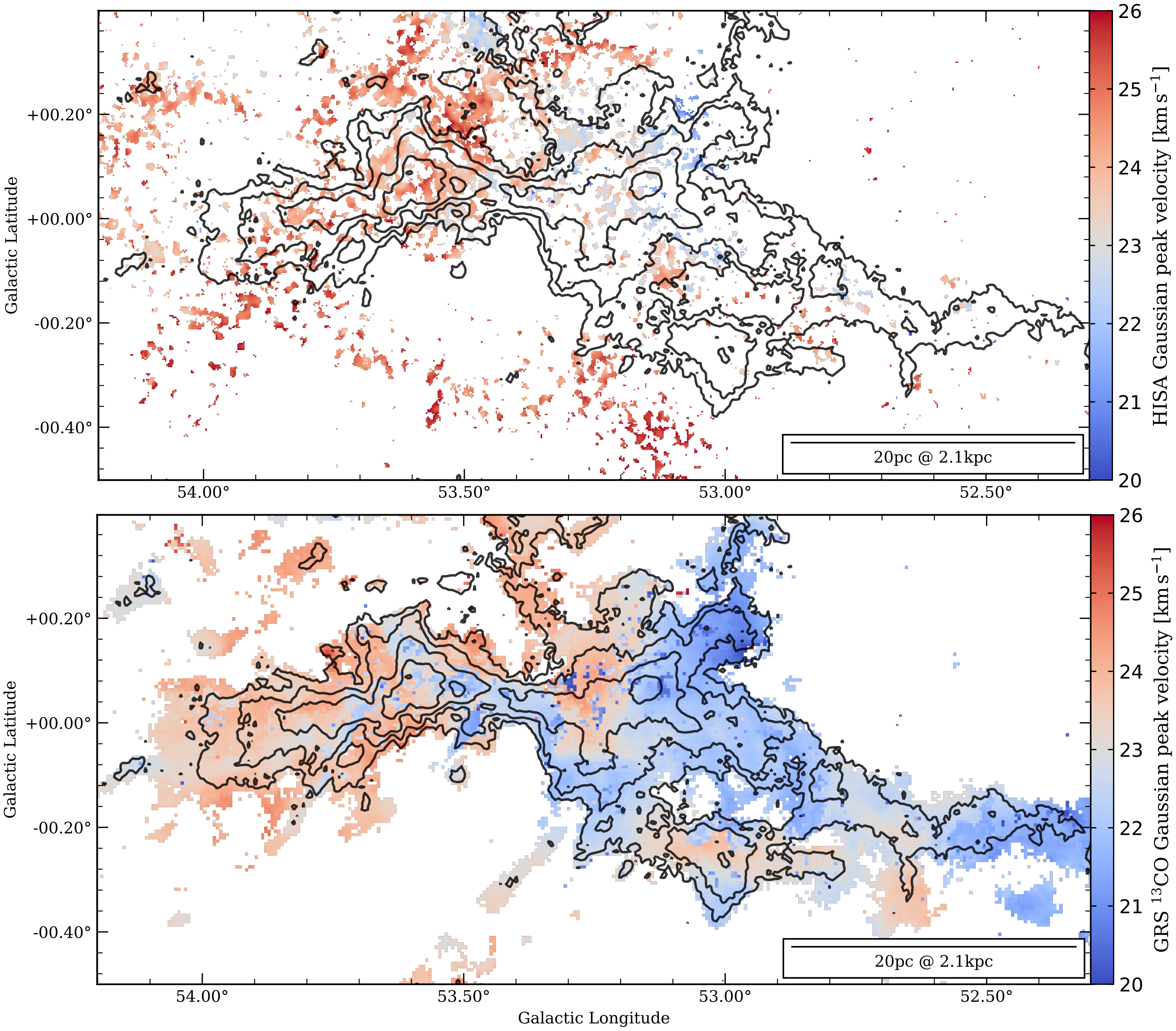

Fig. E.5. Fit peak velocity toward GMF54. These maps show the peak velocities of fit components derived from the GaussPy+ decomposition of the spectra. If multiple components are present in a single pixel spectrum within the velocity range of the filament region, the component with the lowest peak velocity is shown. The black contours in both panels show the integrated GRS ^{13}CO emission at the levels 2.5, 5.0, 10.0, and 20.0 K km s^{-1}. *Top panel:* Fit HISA peak velocity. *Bottom panel:* Fit ^{13}CO peak velocity.

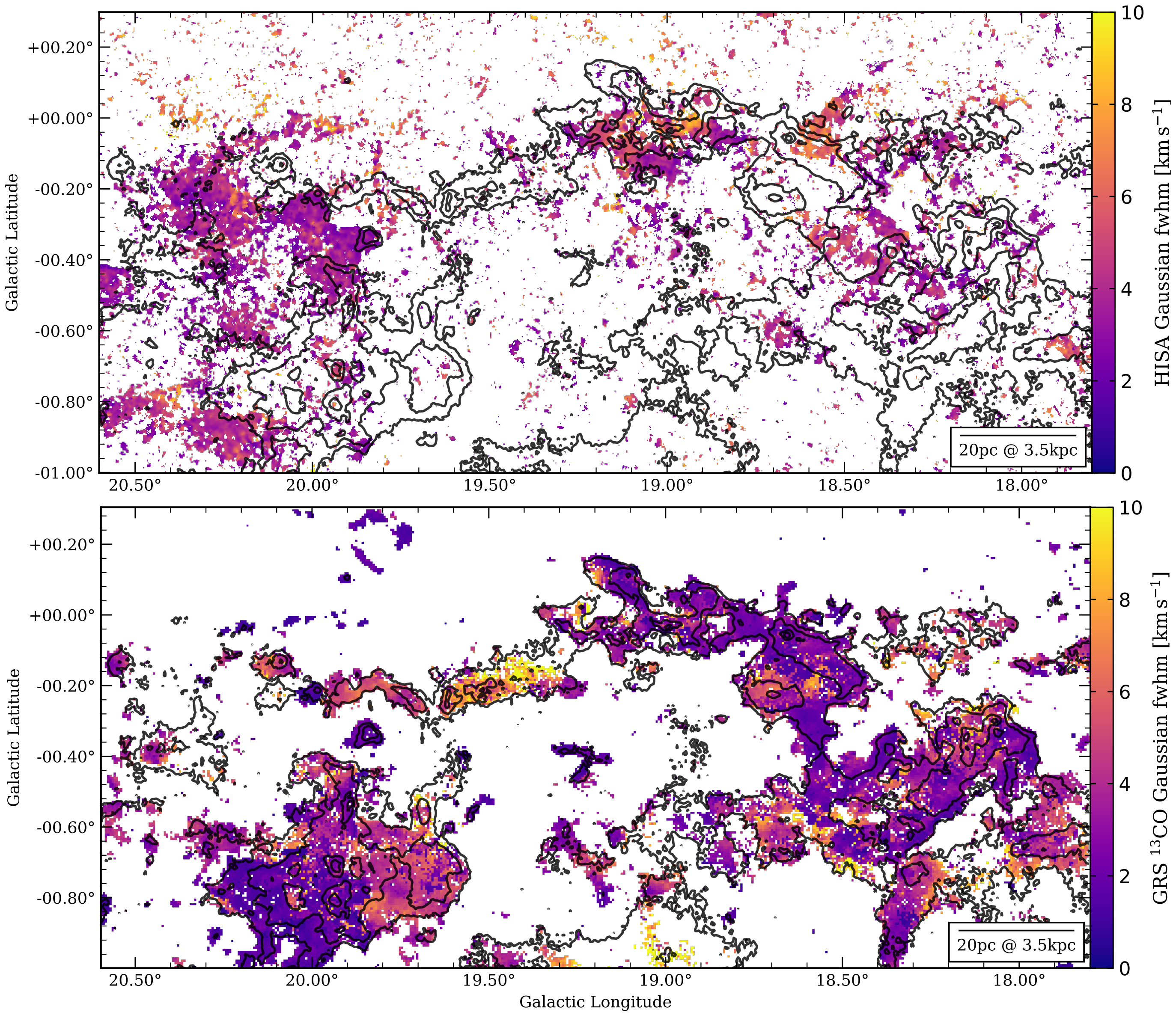

Fig. E.6. Fit line width (FWHM) toward GMF20. These maps show the line widths of fit components derived from the GaussPy+ decomposition of the spectra. If multiple components are present in a single pixel spectrum within the velocity range of the filament region, the component with the lowest peak velocity is shown. The black contours in both panels show the integrated GRS ^{13}CO emission at the levels 8.0, 16.0, 32.0, and 42.0 K km s^{-1}. *Top panel:* Fit HISA line width. *Bottom panel:* Fit ^{13}CO line width.

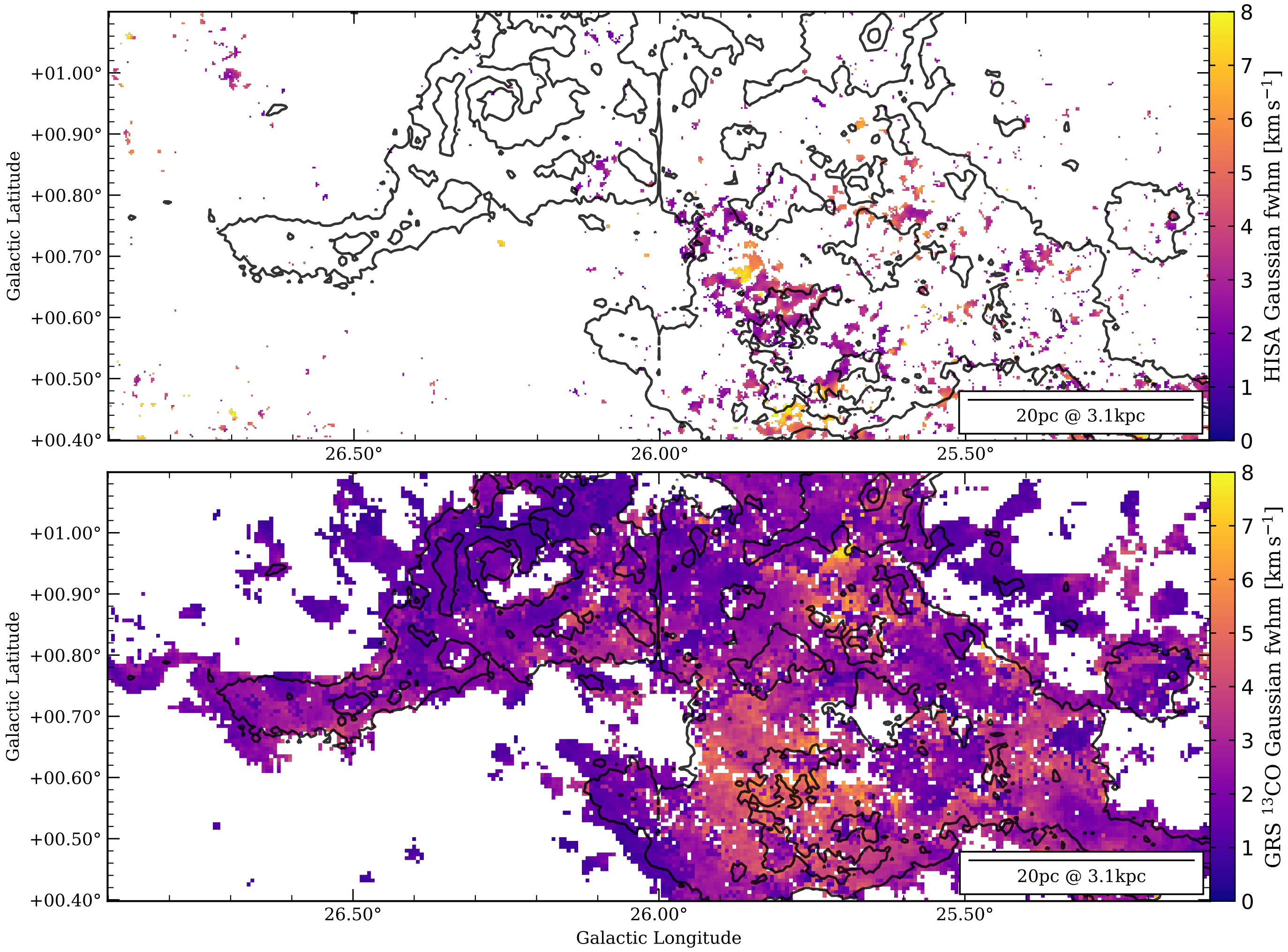

Fig. E.7. Fit line width (FWHM) toward GMF26. These maps show the line widths of fit components derived from the GaussPy+ decomposition of the spectra. If multiple components are present in a single pixel spectrum within the velocity range of the filament region, the component with the lowest peak velocity is shown. The black contours in both panels show the integrated GRS ^{13}CO emission at the levels 6.0, 12.0, 24.0, and 34.0 K km s^{-1}. The contour feature at longitude $\ell = 26°$ is an artifact in the observational data. *Top panel:* Fit HISA line width. *Bottom panel:* Fit ^{13}CO line width.

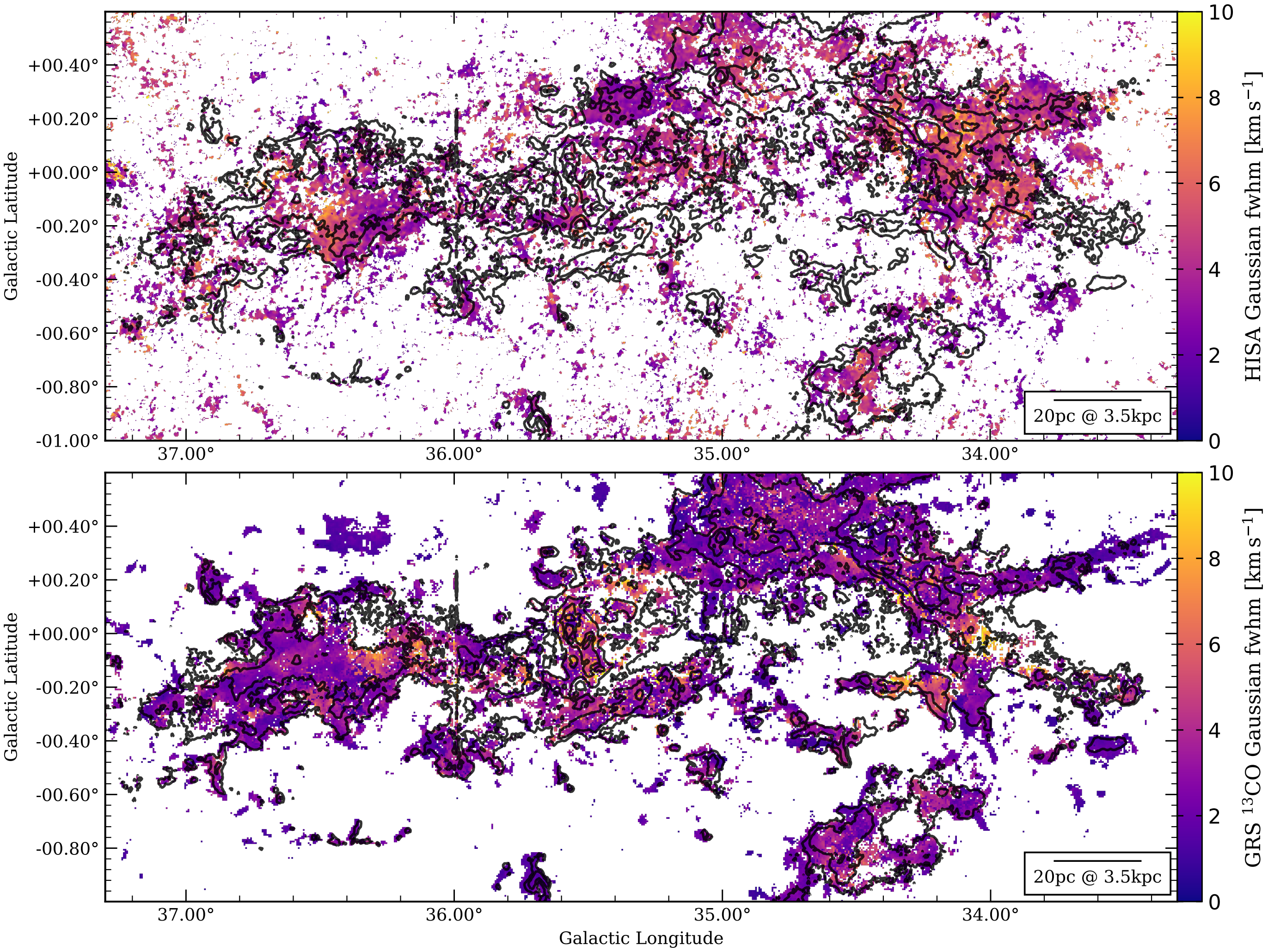

Fig. E.8. Fit line width (FWHM) toward GMF38a. These maps show the line widths of fit components derived from the GaussPy+ decomposition of the spectra. If multiple components are present in a single pixel spectrum within the velocity range of the filament region, the component with the lowest peak velocity is shown. The black contours in both panels show the integrated GRS ^{13}CO emission at the levels 5.0, 10.0, 20.0, and 30.0 K km s^{-1}. The contour feature at longitude $\ell = 36°$ is an artifact in the observational data. *Top panel:* Fit HISA line width. *Bottom panel:* Fit ^{13}CO line width.

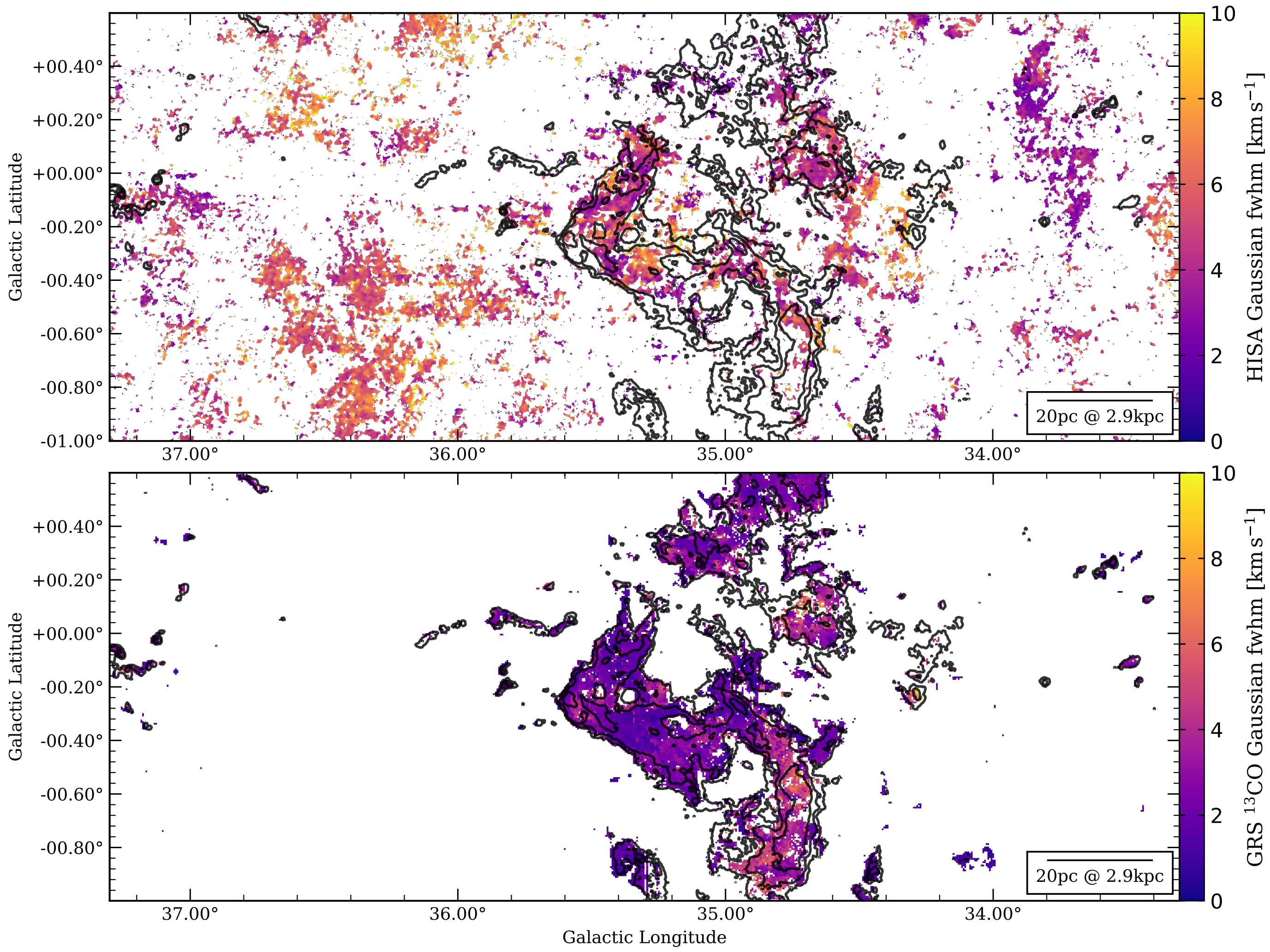

Fig. E.9. Fit line width (FWHM) toward GMF38b. These maps show the line widths of fit components derived from the GaussPy+ decomposition of the spectra. If multiple components are present in a single pixel spectrum within the velocity range of the filament region, the component with the lowest peak velocity is shown. The black contours in both panels show the integrated GRS ^{13}CO emission at the levels 2.5, 5.0, 10.0, and 20.0 K km s^{-1}. *Top panel:* Fit HISA line width. *Bottom panel:* Fit ^{13}CO line width.

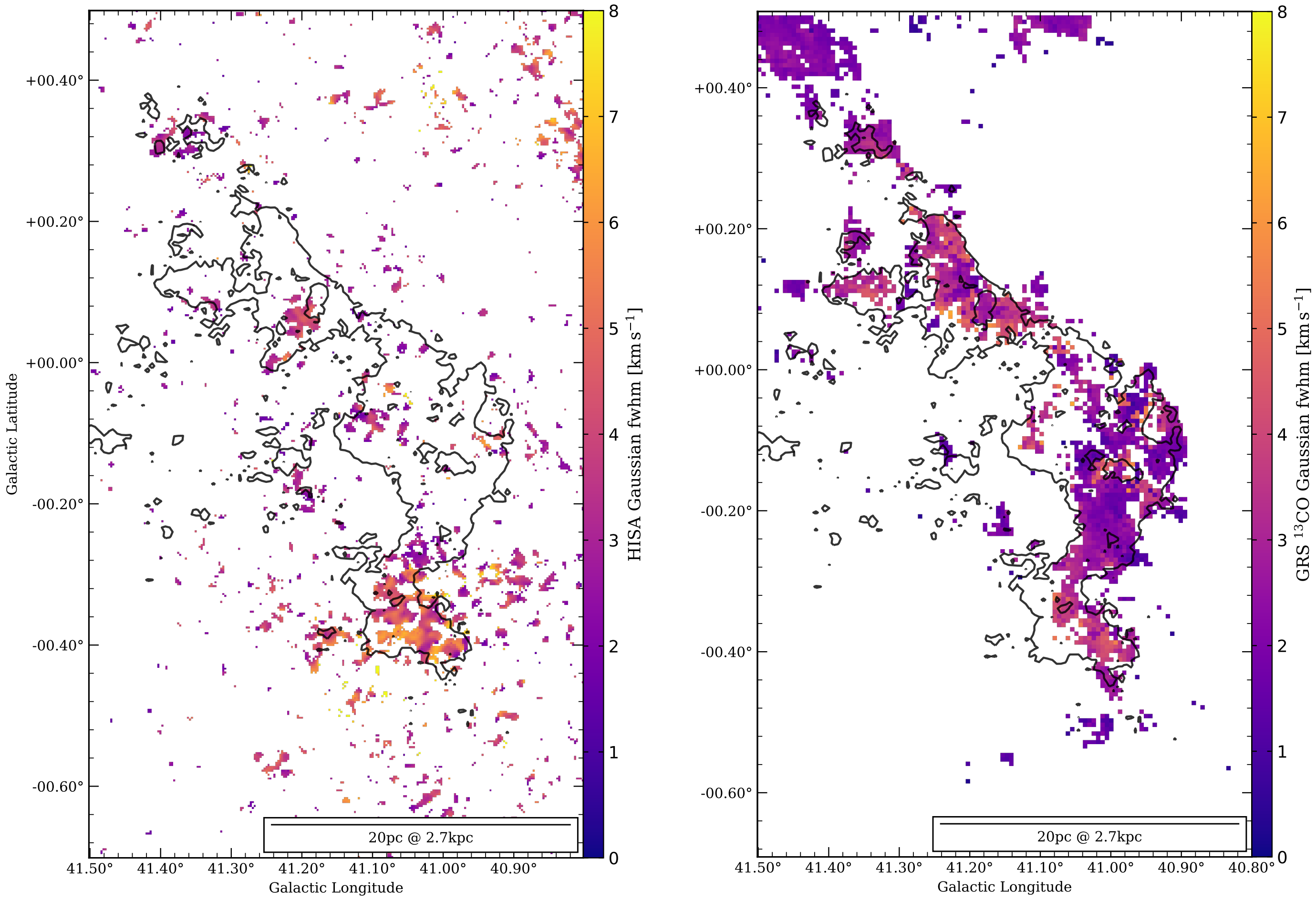

Fig. E.10. Fit line width (FWHM) toward GMF41. These maps show the line widths of fit components derived from the GaussPy+ decomposition of the spectra. If multiple components are present in a single pixel spectrum within the velocity range of the filament region, the component with the lowest peak velocity is shown. The black contours in both panels show the integrated GRS ^{13}CO emission at the levels 6.0, 12.0, 24.0, and 34.0 K km s^{-1}. *Top panel:* Fit HISA line width. *Bottom panel:* Fit ^{13}CO line width.

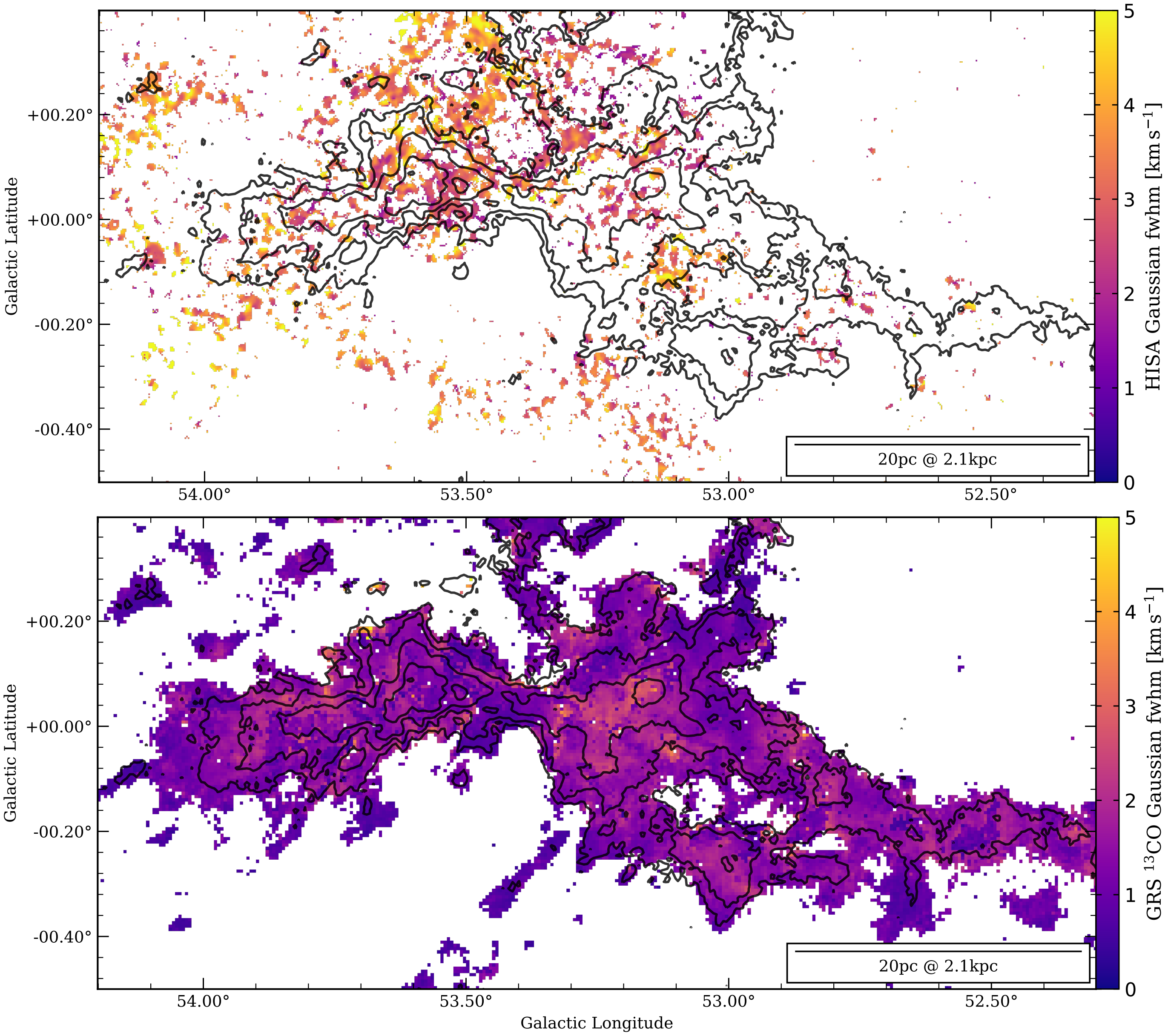

Fig. E.11. Fit line width (FWHM) toward GMF54. These maps show the line widths of fit components derived from the GaussPy+ decomposition of the spectra. If multiple components are present in a single pixel spectrum within the velocity range of the filament region, the component with the lowest peak velocity is shown. The black contours in both panels show the integrated GRS ^{13}CO emission at the levels 2.5, 5.0, 10.0, and 20.0 K km s^{-1}. *Top panel:* Fit HISA line width. *Bottom panel:* Fit ^{13}CO line width.

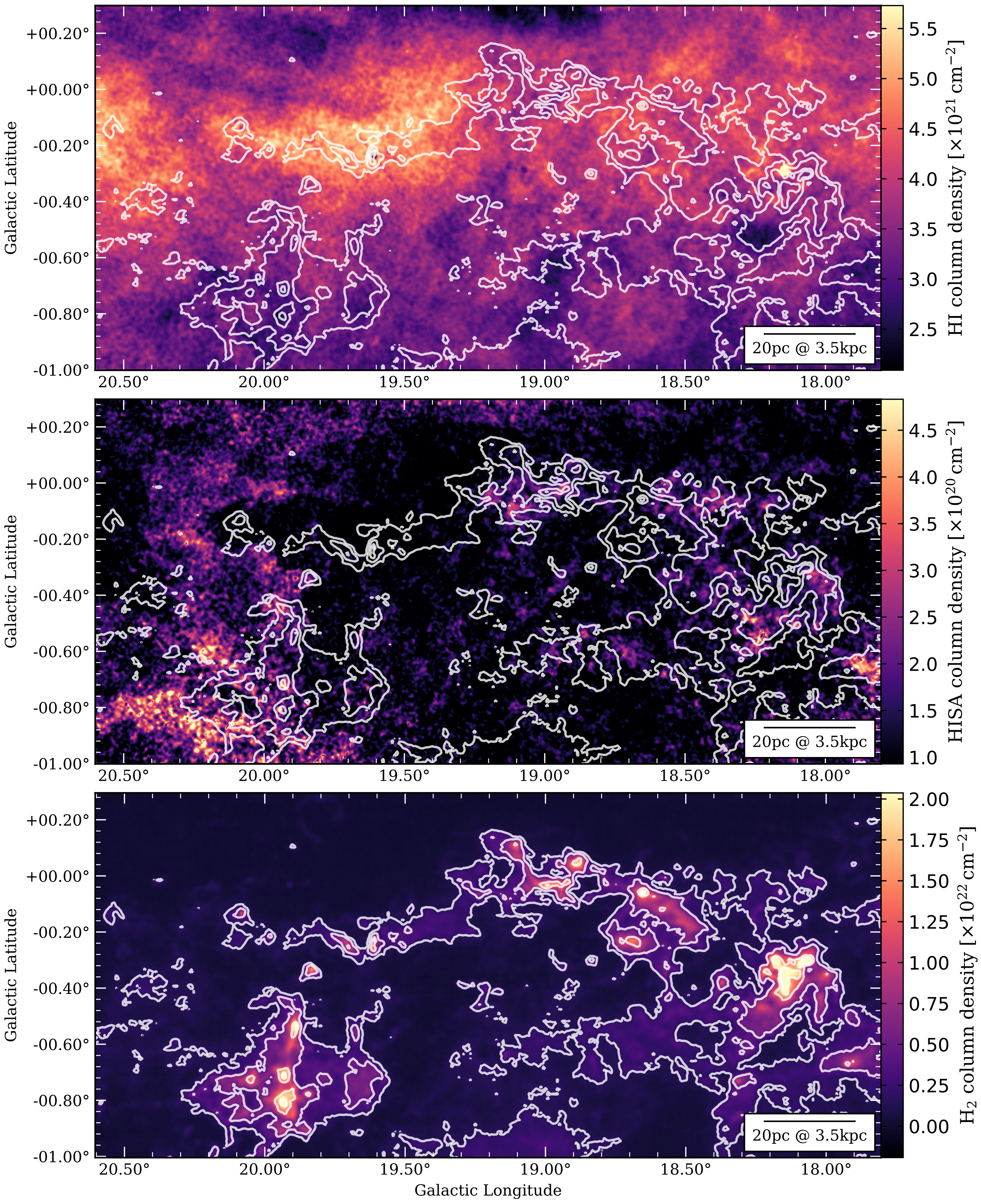

Fig. F.1. Column density toward GMF20. These maps show the column densities of atomic hydrogen traced by H I emission, the cold hydrogen gas traced by HISA, and molecular hydrogen traced by ^{13}CO emission, respectively. The column densities are integrated over the velocity range of the filament region given in Table 1. The white contours in both panels show the integrated MWISP ^{13}CO emission at the levels 8.0, 16.0, 32.0, and 42.0 K km s^{-1}. *Top panel:* H I column density traced by H I emission. *Middle panel:* HISA column density. *Bottom panel:* H_2 column density traced by ^{13}CO.

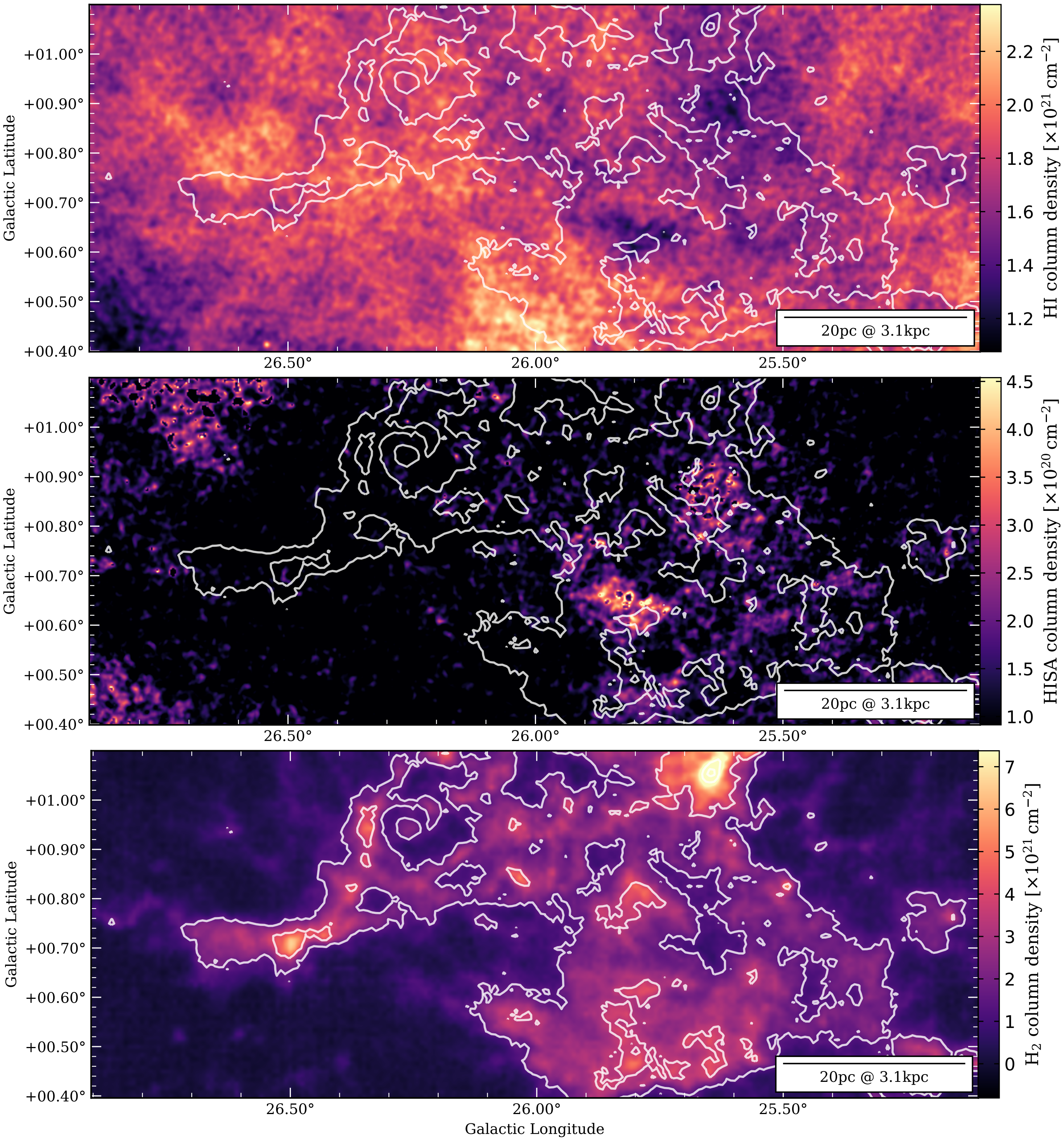

Fig. F.2. Column density toward GMF26. These maps show the column densities of atomic hydrogen traced by H I emission, the cold hydrogen gas traced by HISA, and molecular hydrogen traced by ^{13}CO emission, respectively. The column densities are integrated over the velocity range of the filament region given in Table 1. The white contours in both panels show the integrated MWISP ^{13}CO emission at the levels 6.0, 12.0, 24.0, and 34.0 K km s^{-1}. *Top panel:* H I column density traced by H I emission. *Middle panel:* HISA column density. *Bottom panel:* H_2 column density traced by ^{13}CO.

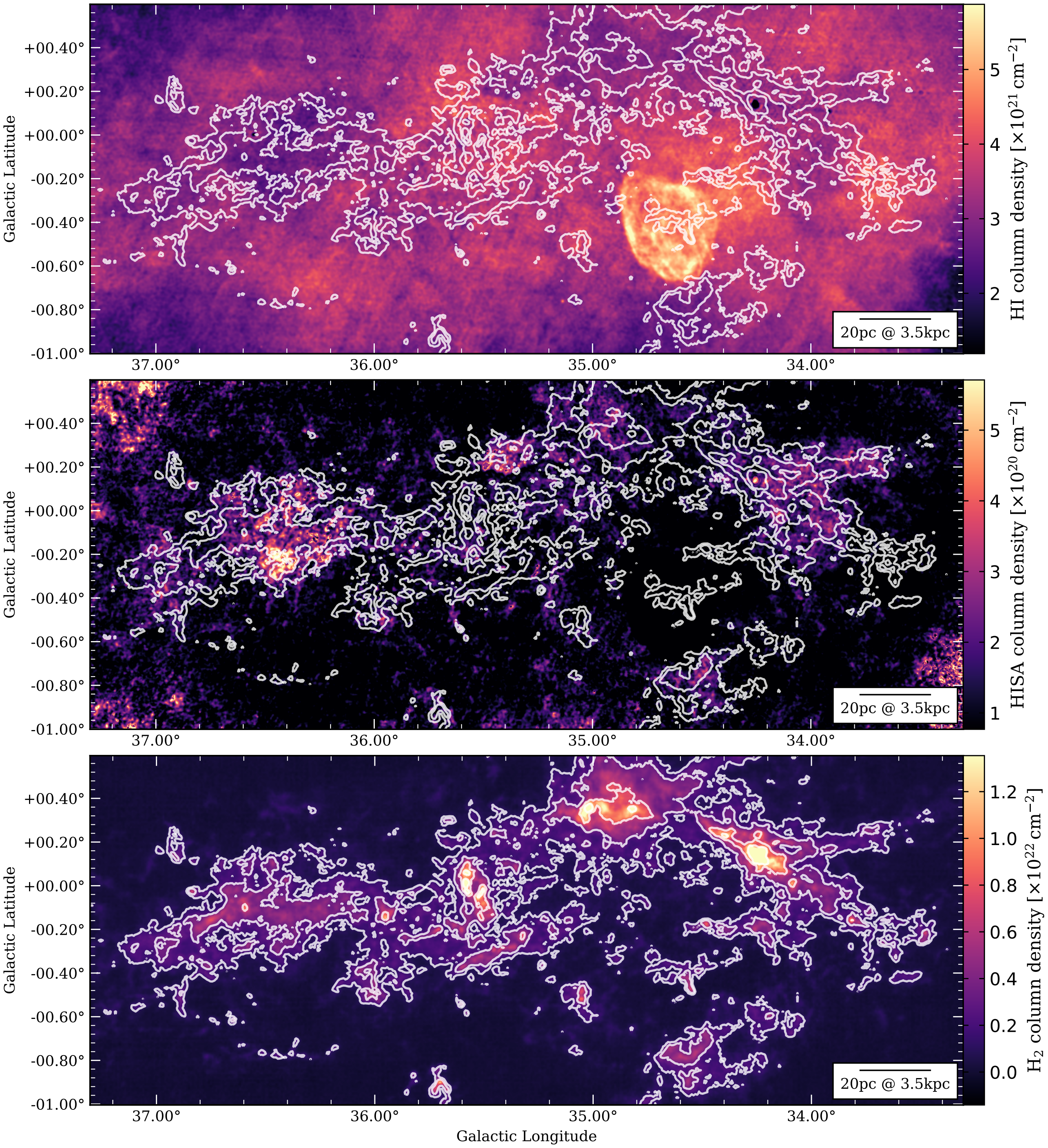

Fig. F.3. Column density toward GMF38a. These maps show the column densities of atomic hydrogen traced by H I emission, the cold hydrogen gas traced by HISA, and molecular hydrogen traced by ^{13}CO emission, respectively. The column densities are integrated over the velocity range of the filament region given in Table 1. The white contours in both panels show the integrated MWISP ^{13}CO emission at the levels 5.0, 10.0, 20.0, and 30.0 K km s^{-1}. *Top panel:* H I column density traced by H I emission. *Middle panel:* HISA column density. *Bottom panel:* H$_2$ column density traced by ^{13}CO.

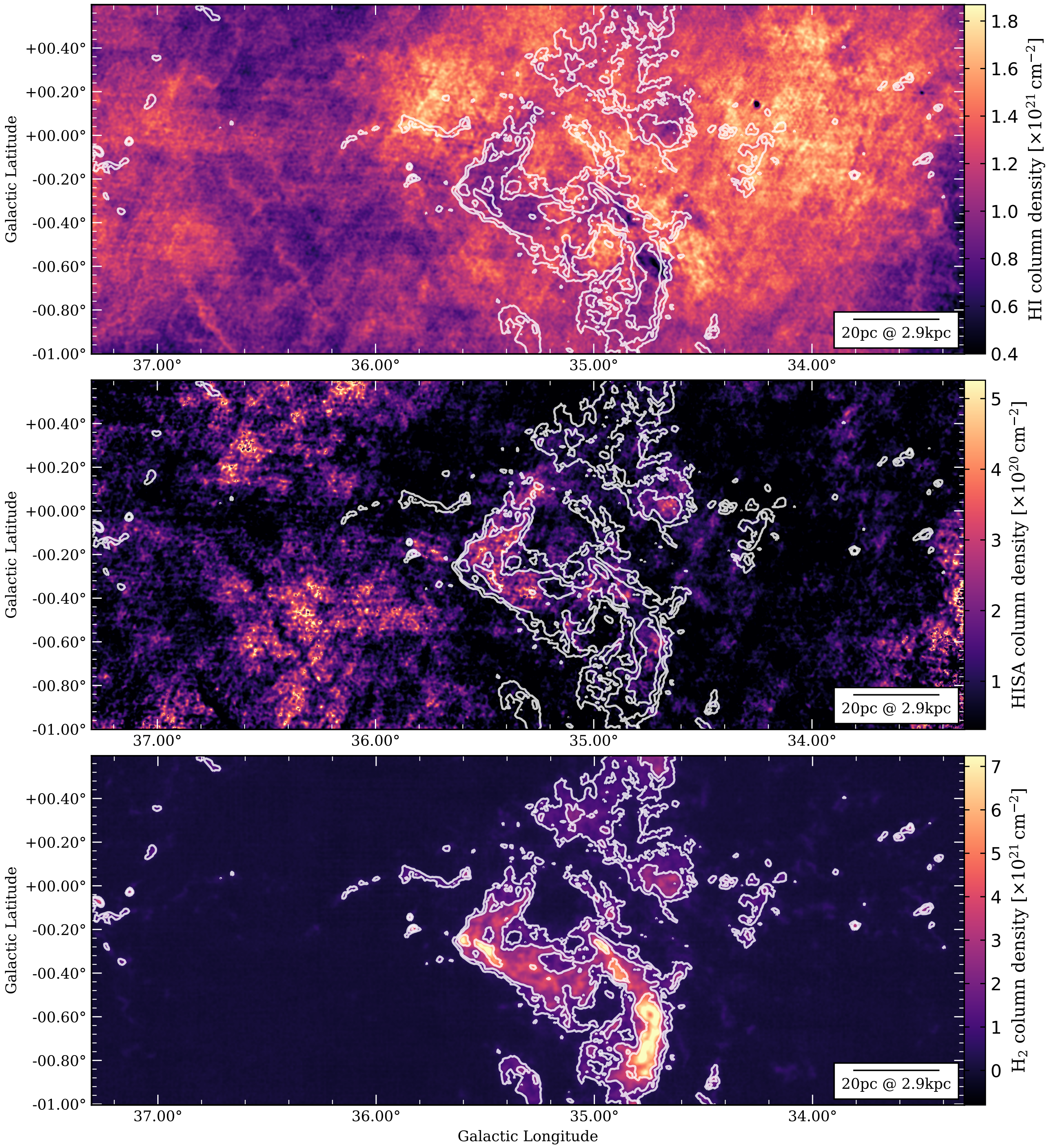

Fig. F.4. Column density toward GMF38b. These maps show the column densities of atomic hydrogen traced by H I emission, the cold hydrogen gas traced by HISA, and molecular hydrogen traced by ^{13}CO emission, respectively. The column densities are integrated over the velocity range of the filament region given in Table 1. The white contours in both panels show the integrated MWISP ^{13}CO emission at the levels 2.5, 5.0, 10.0, and 20.0 K km s^{-1}. *Top panel:* H I column density traced by H I emission. *Middle panel:* HISA column density. *Bottom panel:* H_2 column density traced by ^{13}CO.

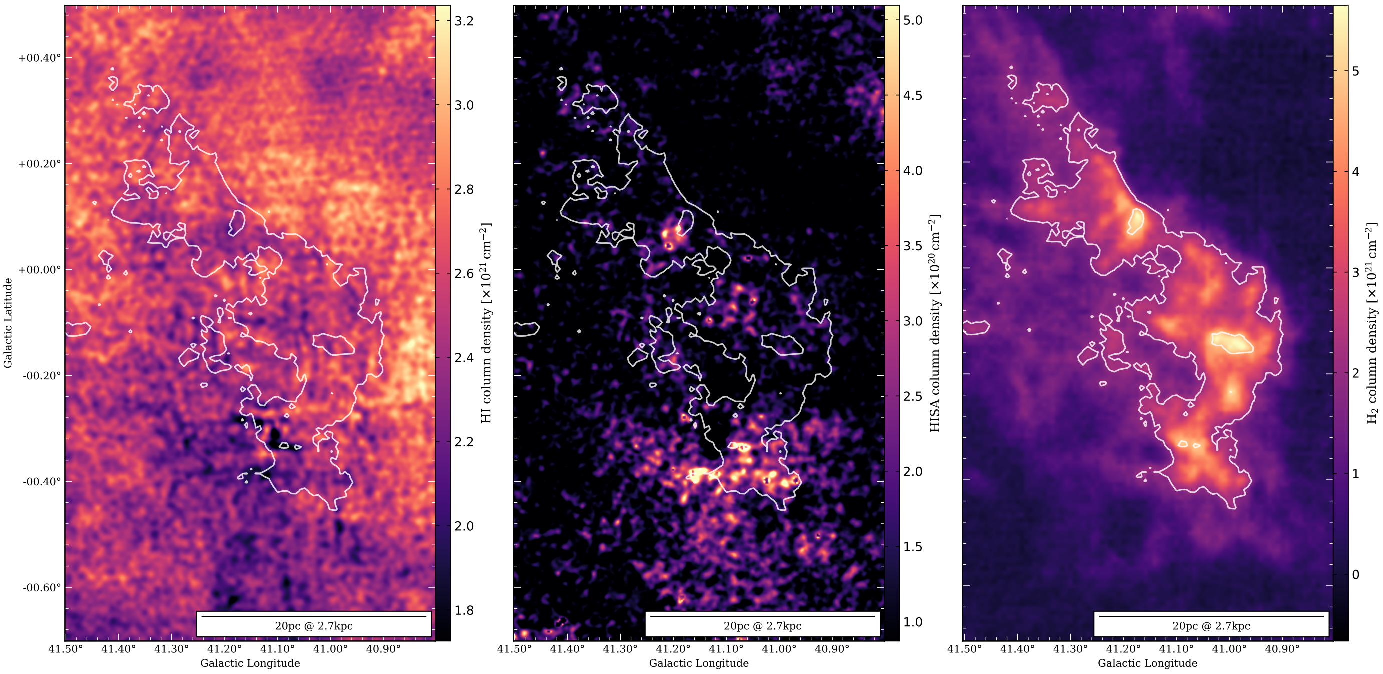

Fig. F.5. Column density toward GMF41. These maps show the column densities of atomic hydrogen traced by H I emission, the cold hydrogen gas traced by HISA, and molecular hydrogen traced by ^{13}CO emission, respectively. The column densities are integrated over the velocity range of the filament region given in Table 1. The white contours in both panels show the integrated MWISP ^{13}CO emission at the levels 6.0, 12.0, 24.0, and 34.0 K km s^{-1}. *Left panel:* H I column density traced by H I emission. *Middle panel:* HISA column density. *Right panel:* H$_2$ column density traced by ^{13}CO.

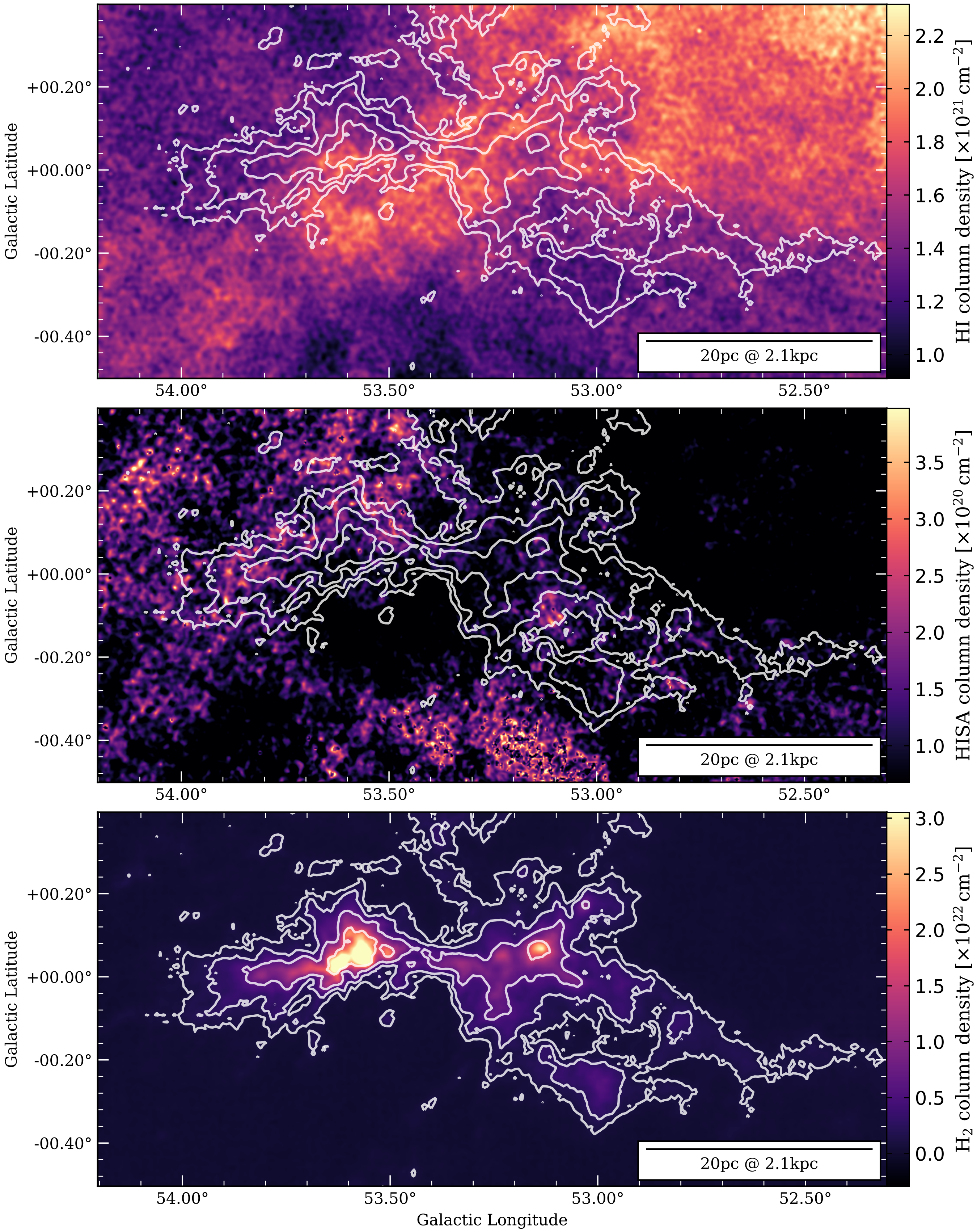

Fig. F.6. Column density toward GMF54. These maps show the column densities of atomic hydrogen traced by H I emission, the cold hydrogen gas traced by HISA, and molecular hydrogen traced by ^{13}CO emission, respectively. The column densities are integrated over the velocity range of the filament region given in Table 1. The white contours in both panels show the integrated MWISP ^{13}CO emission at the levels 2.5, 5.0, 10.0, and 20.0 K km s^{-1}. *Top panel:* H I column density traced by H I emission. *Middle panel:* HISA column density. *Bottom panel:* H$_2$ column density traced by ^{13}CO.