MNRAS **000**, 1–36 (2019) Preprint 11 June 2020 Compiled using MNRAS LATEX style file v3.0

The SEDIGISM survey: Molecular clouds in the inner Galaxy

A. Duarte-Cabral[1]★, D. Colombo[2], J. S. Urquhart[3], A. Ginsburg[4], D. Russeil[5], F. Schuller[2], L. D. Anderson[6], P. J. Barnes[7,8], M. T. Beltrán[9], H. Beuther[10], S. Bontemps[11], L. Bronfman[12], T. Csengeri[11], C. L. Dobbs[13], D. Eden[14], A. Giannetti[2], J. Kauffmann[15], M. Mattern[2], S.-N. X. Medina[2], K. M. Menten[2], M.-Y. Lee[2,16], A. R. Pettitt[17], M. Riener[10], A. J. Rigby[1], A. Traficante[18], V. S. Veena[19], M. Wienen[2], F. Wyrowski[2], C. Agurto[20], F. Azagra[20], R. Cesaroni[9], R. Finger[12], E. Gonzalez[20], T. Henning[10], A. K. Hernandez[21], J. Kainulainen[2,22], S. Leurini[2,23], S. Lopez[4], F. Mac-Auliffe[20], P. Mazumdar[2], S. Molinari[18], F. Motte[24], E. Muller[25], Q. Nguyen-Luong[15], R. Parra[20], J.-P. Perez-Beaupuits[20], F. M. Montenegro-Montes[20], T. J. T. Moore[14], S. E. Ragan[1], A. Sánchez-Monge[19], A. Sanna[2], P. Schilke[19], E. Schisano[18], N. Schneider[19], S. Suri[19], L. Testi[19], K. Torstensson[20], P. Venegas[20], K. Wang[26], and A. Zavagno[5]

Affiliations can be found after the references.

Accepted 2020 May 22. Received 2020 May 22; in original form 2019 October 10.

ABSTRACT

We use the ^{13}CO (2-1) emission from the SEDIGISM (Structure, Excitation, and Dynamics of the Inner Galactic InterStellar Medium) high-resolution spectral-line survey of the inner Galaxy, to extract the molecular cloud population with a large dynamic range in spatial scales, using the Spectral Clustering for Interstellar Molecular Emission Segmentation (SCIMES) algorithm. This work compiles a cloud catalogue with a total of 10663 molecular clouds, 10300 of which we were able to assign distances and compute physical properties. We study some of the global properties of clouds using a science sample, consisting of 6664 well resolved sources and for which the distance estimates are reliable. In particular, we compare the scaling relations retrieved from SEDIGISM to those of other surveys, and we explore the properties of clouds with and without high-mass star formation. Our results suggest that there is no single global property of a cloud that determines its ability to form massive stars, although we find combined trends of increasing mass, size, surface density and velocity dispersion for the sub-sample of clouds with ongoing high-mass star formation. We then isolate the most extreme clouds in the SEDIGISM sample (i.e. clouds in the tails of the distributions) to look at their overall Galactic distribution, in search for hints of environmental effects. We find that, for most properties, the Galactic distribution of the most extreme clouds is only marginally different to that of the global cloud population. The Galactic distribution of the largest clouds, the turbulent clouds and the high-mass star-forming clouds are those that deviate most significantly from the global cloud population. We also find that the least dynamically active clouds (with low velocity dispersion or low virial parameter) are situated further afield, mostly in the least populated areas. However, we suspect that part of these trends may be affected by some observational biases (such as completeness and survey limitations), and thus require further follow up work in order to be confirmed.

Key words: ISM: clouds – galaxies: ISM, star formation

★ E-mail: adc@astro.cf.ac.uk

1 INTRODUCTION

The evolution of the gas that makes up the interstellar medium (ISM), and the ultimate means by which that gas gives way to star formation, involve the tight interplay of a wealth of physical processes. Our understanding of those processes has relied upon the statistical characterisation of the molecular gas that is taking part in the star formation process. In particular, the star formation field has relied on a discretisation of the molecular component of the ISM into molecular clouds, across the Galactic disc, either as observed in 2D with dust continuum emission (e.g. the ATLASGAL survey, Schuller et al. 2009; the Hi-GAL survey, Molinari et al. 2010; or the Bolocam Galactic Plane Survey, Rosolowsky et al. 2010, Ginsburg et al. 2013), or with the 3D view of the Galactic plane from spectral-line observations, most commonly using the second-most abundant molecular species in the ISM, the CO molecule (and its isotopologues). Large survey observations of the Galactic plane in CO emission have allowed for a number of statistical studies of molecular clouds across the Galaxy (e.g. Scoville & Solomon 1975; Larson 1981; Solomon et al. 1987; Heyer et al. 2009; Roman-Duval et al. 2010; Rice et al. 2016; Miville-Deschênes et al. 2017), and have provided a large-scale view of the distribution of gas in the Milky Way, crucial for our understanding of its spiral structure (e.g. Dame et al. 2001; Vallée 2014; Pettitt et al. 2014, 2015).

These Galactic plane surveys, alongside some resolved studies of molecular clouds in nearby spiral galaxies, have also suggested a number of scaling relations (namely between the sizes of clouds, their line-widths, and their mass surface densities, e.g. Larson 1981; Solomon et al. 1987; Heyer et al. 2009; Sun et al. 2018), as well as some differences in the mass spectra of clouds towards different environments (e.g. Colombo et al. 2014; Rice et al. 2016; Miville-Deschênes et al. 2017). All of these findings have implications in our interpretation of the global properties of molecular clouds, and how they might evolve. Most of these surveys, however, were finding and describing molecular clouds that had typical sizes close to their resolution element - which can bias the interpretation of the results - and given their lower resolution they could also potentially suffer from severe blending of the emission along the same line of sight, especially with our edge-on perspective of the Milky Way (e.g. Duarte-Cabral et al. 2015; Duarte-Cabral & Dobbs 2016).

With the advent of new high-resolution and large-scale spectroscopic surveys of the Galactic plane (such as the Structure, Excitation, and Dynamics of the Inner Galactic InterStellar Medium survey - SEDIGISM, Schuller et al. 2017; the CO High Resolution Survey - COHRS, Dempsey et al. 2013; the ^{13}CO/C^{18}O (J=3-2) Heterodyne Inner Milky Way Plane Survey - CHIMPS, Rigby et al. 2016; the Three-mm Ultimate Mopra Milky Way Survey - ThrUMMS, Barnes et al. 2015; or the Galactic Census of High and Medium-mass Protostars - CHaMP Barnes et al. 2011), not only are these shortcomings now greatly minimised, but we can start to explore the details of the sub-structure within molecular clouds where star formation is actively taking place, and the clouds' link to the large-scale Galactic environment. This opens a new and exciting era in the study of star formation in a Galactic context. Given that molecular clouds are highly hierarchical systems, it is essential to be able to define molecular clouds with a large dynamic range in spatial scales (e.g. as in Colombo et al. 2019), and this is at the heart of this present work. In this paper, we explore the global properties of molecular clouds from the high-resolution ^{13}CO (2-1) emission from the SEDIGISM survey, covering the inner Galactic plane (from $+300^\circ \leqslant \ell \leqslant +18^\circ$, Schuller et al. 2017), which is described in Sect. 2. Section 3 contains the details of the method used for the extraction of molecular clouds from this dataset, along with a description of all the derived properties and data-products released with the molecular cloud catalogue. In Sect. 4 we describe the methods used to determine the distances and distinguish between derived near/far kinematic distances to all the clouds in the catalogue, essential to derive the physical properties. In Sect. 5 we explore the distributions of the global properties of the SEDIGISM clouds, and also compare these with other samples in the literature. In Sect. 6, we explore possible indications of environmental dependency of cloud properties, by isolating the most extreme clouds (i.e. clouds in the tails of the distributions), and comparing their Galactic distribution with that of the entire cloud population. Finally, our findings are summarised in Sect. 7.

2 DATA

In this paper, we use data from the SEDIGISM survey conducted with the Atacama Pathfinder Experiment 12 m submillimetre telescope (APEX, Güsten et al. 2006). In particular, we use the ^{13}CO (2-1) to extract and characterise the molecular clouds towards the inner Galaxy. The complete details on the observations, data reduction and data-quality checks can be found in the survey overview papers (Schuller et al. 2017, 2019).

In summary, the SEDIGISM survey observed a total of 84 deg^2, covering from $-60^\circ \leqslant \ell \leqslant +18^\circ$, and $|b| \leqslant 0.5^\circ$, plus a few extensions in b towards some regions, as well as an additional field towards the W43 region ($+29^\circ \leqslant \ell \leqslant +31^\circ$). The ^{13}CO (2-1) data that we use here is the DR1 dataset (fully described in Schuller et al. 2019), which has a typical 1σ sensitivity of 0.8–1.0 K (in $T_{\rm mb}$) per 0.25 km s^{-1} channel, and a FWHM beam size, $\theta_{\rm MB}$, of 28″.

In this paper we will use the complete contiguous dataset (i.e. the entire survey data except for the W43 field). This consists of 77 datacubes of roughly $2^\circ \times 1^\circ$ (note that the latitude range is sometimes larger than 1°), centred at all integer longitudes between $\ell = 301^\circ$ and $\ell = 17^\circ$ (i.e. spaced by 1° in longitude). This provides a 1° overlap in longitude between consecutive tiles, which ensures all sight lines (except for the first and last fields) are contained in two tiles. The velocity ranges from -200 to $+200$ km s^{-1} in all datacubes, and the pixel size is of 9.5″. Figure 1 shows the full ℓv map of the contiguous dataset from the SEDIGISM survey, that we use here.

3 MOLECULAR CLOUD EXTRACTION

3.1 The method: SCIMES

In order to decompose the ^{13}CO emission from the SEDIGISM survey into discrete clouds, we use the SCIMES algorithm (v.0.3.2)[1]. The original algorithm is fully described in Colombo et al. (2015), and the improvements included in the version we use here are detailed in Colombo et al. (2019). In brief, SCIMES brings a significant advancement with respect to other more commonly used cloud-extraction algorithms (e.g. Clumpfind by Williams et al. 1994, Gaussclumps by Stutzki & Guesten 1990, or Fellwalker by Berry 2015), as it is a fully automated method that uses spectral clustering and graph theory to analyse the dendrogram of the emission, and decompose the hierarchical structure of the ISM into "clusters" of molecular gas emission (i.e. molecular clouds, considering the resolution of

[1] https://github.com/Astroua/SCIMES

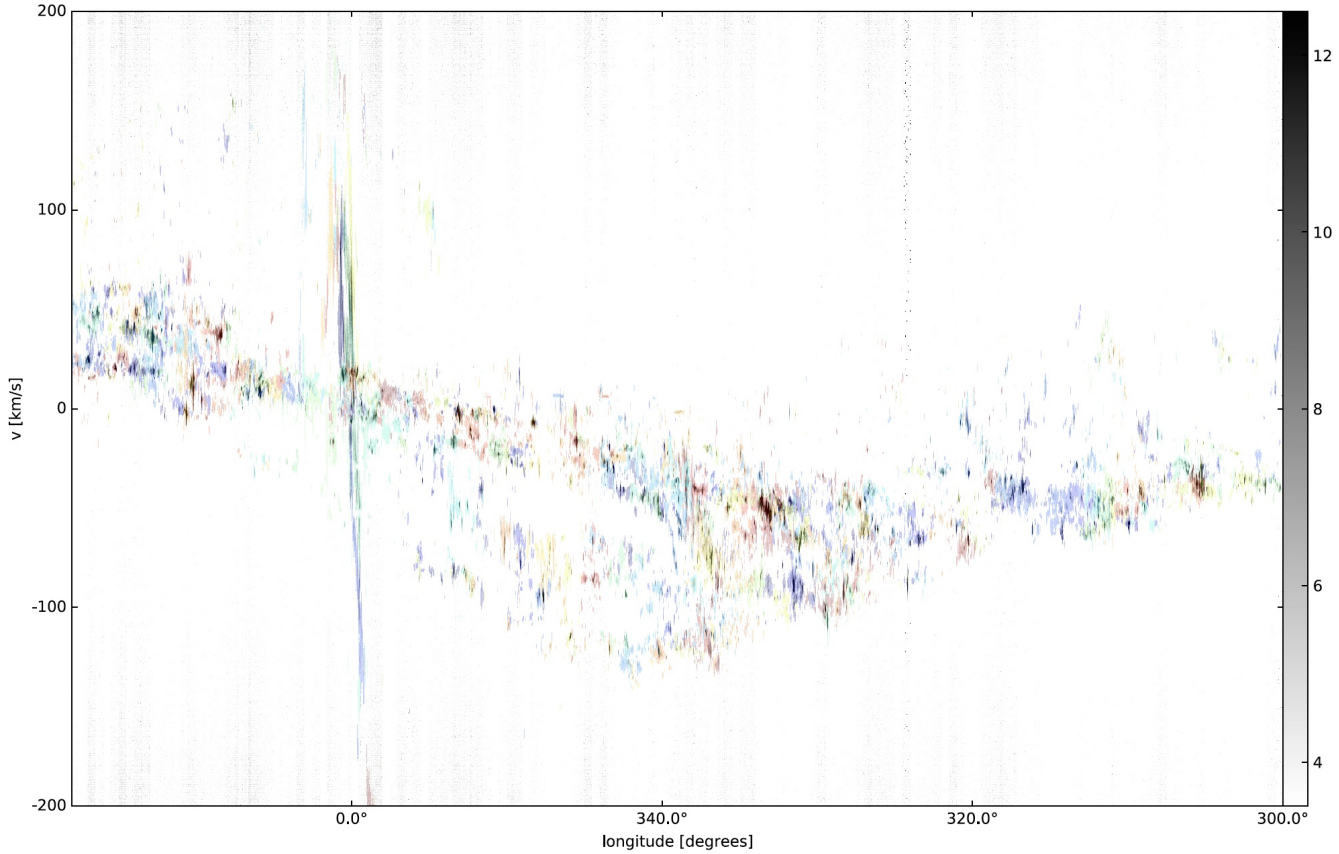

Figure 1. Longitude-velocity (ℓv) map of the ^{13}CO peak intensity (in greyscale) for the SEDIGISM coverage analysed in this paper. The peak intensity map was built after masking out voxels of the ^{13}CO datacube with intensities $< 2.5\sigma_{\rm rms}$ (estimated locally for each line of sight). The clouds extracted with SCIMES are overlaid as colours, where each cloud has a different (random) colour.

SEDIGISM). Unlike other cloud-extraction algorithms, SCIMES relies on the natural transitions in the emission to define discrete structures, and it is robust against changes in the input parameters (as demonstrated in Colombo et al. 2015).

The cloud extraction with SCIMES was performed on each of the 77 tiles of $2^\circ \times 1^\circ$. We ran SCIMES on these relatively small cubes because it would be extremely computationally expensive (and memory intensive) to generate a single dendrogram from the full SEDIGISM dataset, and perform SCIMES's affinity matrix analysis, where each cluster is equivalent to an additional dimension in the clustering space.

3.1.1 Input parameters and files

In order to optimise the performance of the SCIMES clustering algorithm, we have performed a few preparation steps on the original DR1 data. Firstly, we enhanced the signal-to-noise ratio of the data set prior to running SCIMES by smoothing the data in velocity. This was done by binning the data into 0.5 km s^{-1} channels. We then re-sampled these binned datacubes back into 0.25 km s^{-1} channels (using linear interpolation), simply so that the SCIMES assignment masks (Sect. 3.2) kept the same format as the original emission datacubes from DR1 (essential to have straight forward voxel-by-voxel match between the DR1 emission maps and the clouds' assignment masks). We have performed some tests on the science demonstration field (Schuller et al. 2017), with binned and non-binned data, and this step allows us to remove high-frequency noise spikes, speeding up the dendrogram construction and the SCIMES clustering, with minimal loss in the information retrieved.

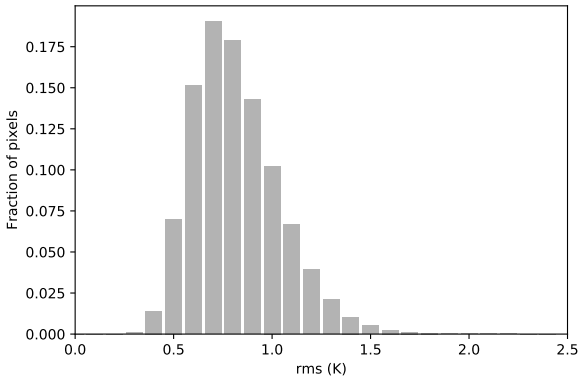

Figure 2. Histogram of the rms noise level of the entire survey, from the velocity-smoothed datacubes that we use for the SCIMES extraction, showing that it peaks at ~0.7 K, with a median value of 0.78 K.

Secondly, given that the noise in the survey is not perfectly uniform (due to different observing weather conditions), it is also essential to mask the datacubes using the local noise level, in order to prevent high-noise regions from being used in the dendrogram tree, and incorrectly identified as clouds. For this purpose, we estimated the local noise level at each pixel (i.e. each line of sight) in the velocity-smoothed datacubes, by taking the first 50 channels

(which are line-free, and on the high-frequency end, i.e. at negative systemic velocities), and computing the 1σ standard deviation. Figure 2 shows the distribution of this local 1σ-rms noise level for all the pixels in our velocity-smoothed dataset, showing that it peaks at ~0.7 K. We then create a mask of each datacube, by setting any 3D pixels (voxels) whose emission is lower that 2σ of the local noise to zero. Note that since we already go down to 2σ of the local noise, we do not perform any dilation of the masks after this step (which is a technique sometimes used to remove potential breaks in clouds in low signal-to-noise areas).

Using these masked datacubes, we computed the dendrogram tree of the 3D structures in the data (using the ASTRODENDRO[2] implementation, which is based on the original IDL procedures from Rosolowsky et al. 2008). The dendrogram is composed of three types of structures: *leaves*, which are at the top of the hierarchy and contain no substructure, i.e. they are associated with local peaks of emission; *branches*, which split into multiple substructures; and the *trunk*, which is at the bottom of the hierarchy (i.e. it has no parent structure), and comprises all *branches* and *leaves*. We built our dendrograms using the same input parameters as in the science demonstration field (Schuller et al. 2017): we considered a noise level ($\sigma_{\rm rms}$) of 0.7 K for all tiles (corresponding to the peak of the noise distribution in Fig. 2), a $4\sigma_{\rm rms}$ value as the minimum difference between two peaks for them to be considered as separate structures, and a lower threshold for detection of $2\sigma_{\rm rms}$, to maximise the connections between different structures at contiguous lower intensity levels[3]. Note that we specifically chose to use a single fixed value of $\sigma_{\rm rms}$ to build the dendrograms across the entire survey (rather than using a local signal-to-noise ratio approach) so that we could define our structures using a uniform criterion throughout. This not only makes it easier to replicate our results using other datasets, but it also ensures that the type of structures we extract are equivalent throughout the entire survey, and not dependent on the local noise conditions[4]. The choice of a $4\sigma_{\rm rms}$ for the significance of individual peaks, coupled with the fact that the dendrogram was built from datacubes that had been masked based on the pixel-based noise level, was so as to maximise the retained detailed information encoded in the survey, whilst minimising the inclusion of noise spikes. In addition, we set a minimum number of voxels for a structure to be considered as real to be 6 times the number of pixels per beam ($N_{\rm ppbeam} = 9$), so that structures are both resolved spatially (i.e. at least 3 beams), and in velocity (spanning at least 2 channels, which corresponds to our effective velocity resolution in the smoothed datacubes). Note, however, that the ASTRODENDRO implementation that we use to build the dendrogram does not separate the spatial axes from the spectral axis. In practice, this means that this criterion will still allow some clouds to be retained whilst being unresolved in one of the axes. Those sources are dealt with in a post-processing step (see Sect. 3.1.3).

Once the dendrograms were constructed for each tile, we ran SCIMES using both the "volume" and the "flux" (which in our case, refers simply to the surface brightness) as the clustering properties (cf. Colombo et al. 2015). This extraction recovered a total of 20387 gas clusters from the 77 tiles, but most of these are duplicated due to the overlap between consecutive fields. In order to build the final catalogue (and respective assignment masks), we performed a cleaning up procedure to handle clouds in overlapping areas. This is described in the following section.

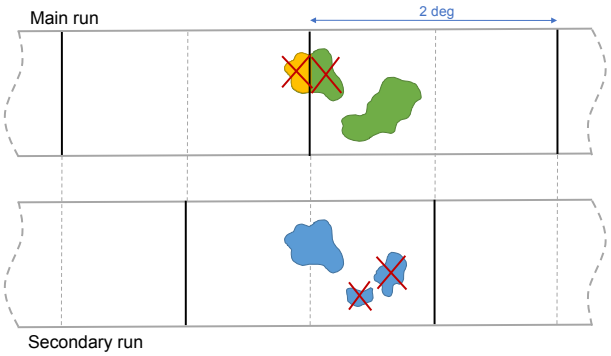

Figure 3. Schematic sketch of the procedure used to decide which clouds to retain from the two overlapping runs, namely the removal of clouds that touch the edge of the tiles (which are then recovered in full in the complementary run), and the selection of the larger clouds for overlapping cases between the two runs.

3.1.2 *Handling clouds in overlapping regions*

In order to handle the clouds that appear in overlapping areas, we have followed a procedure similar to that used by Colombo et al. (2019). This procedure is schematically described in Fig. 3. In essence, we have split our dataset into a *main* run (which is composed of all tiles centred at odd longitudes), and a *secondary* run (which is composed of all tiles centred at even longitudes). We then exclude all objects that touch a tile edge on the longitude axis, since their contours are not closed, and they should be fully recovered in the complementary run. We only made an exception for objects that touch the first and last longitude edges of the contiguous coverage (i.e. at $\ell = 18°$ and $\ell = 300°$), which are retained in the final catalogue with a tag that indicates that they are edge clouds. Similarly, we also retain clouds that touch the survey's upper and lower latitude edges, and tag them as being edge clouds. Finally, we proceed to checking the matches between the *main* and *secondary* runs. We start by including all objects that do not overlap between the two runs, and whenever two (or more) clouds overlap, we simply retain the larger object between the two runs. After this procedure, we have compiled a total of 11638 unique molecular clouds.

3.1.3 *Removal of spurious sources*

As mentioned in Sect. 3.1.1, despite our best efforts to avoid having any noisy spikes in the dendrogram (by imposing a noise level threshold) or unresolved sources (by imposing a minimum number of voxels), some spurious sources still persist to the dendrogram construction and into our final catalogue. One of the reasons for this is the fact that we have applied an average noise $\sigma_{\rm rms}$ for

[2] http://www.dendrograms.org

[3] These values are solely defined by the data quality, but tests using slight variations for the different parameters for the dendrogram construction were performed as part of our work on the science demonstration field (Schuller et al. 2017). Those tests have shown that the SCIMES clustering algorithm is robust against small differences on the parameters used to construct the dendrogram.

[4] Although the choice of a unique value for $\sigma_{\rm rms}$ does require an extra post-processing step to ensure that detected structures also have a high local signal-to-noise ratio (see Sec. 3.1.3), doing the cloud extraction on a signal-to-noise map with the SCIMES algorithm (which does not segment clouds at a fixed brightness threshold) means that we could end up with structures identified across the survey using different criteria, which is non-ideal.

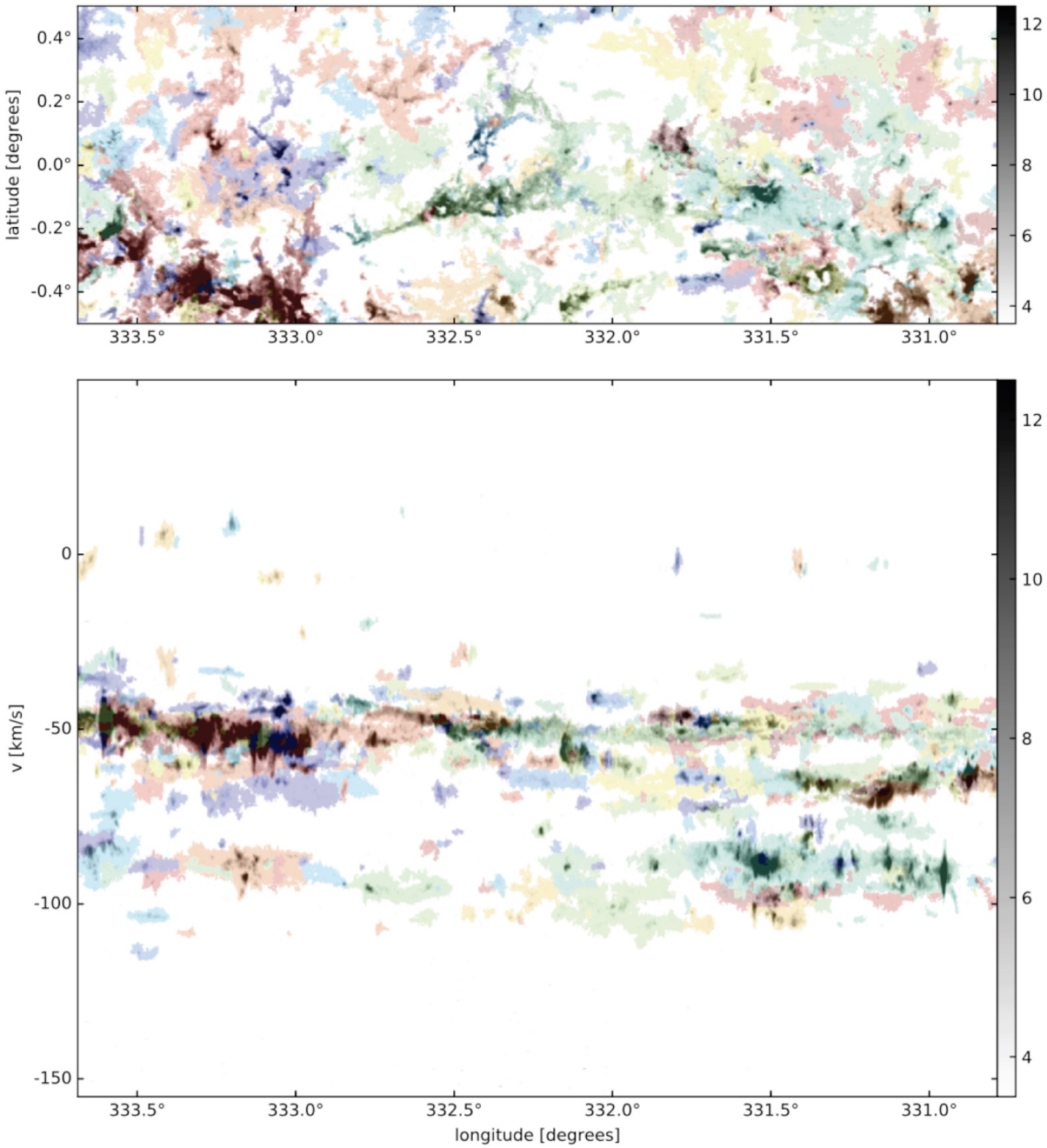

Figure 4. Example of the SCIMES cloud extraction results, on a small section of the SEDIGISM survey. The top panel shows the ℓb map with the ^{13}CO peak intensity in greyscale, and the SEDIGISM cloud masks overlaid as colours, where each cloud has a different (random) colour. The bottom panel shows the ℓv map of the same field, with the same colour-scheme as the top panel.

the entire survey (so that the dendrogram for all fields was built upon a fixed physical value of emission intensity). This means that in areas where the local noise level is higher than this average $\sigma_{\rm rms}$, some noisy peaks would have been considered as robust emission peaks. Most of these sources are located near the noisier edges of the observed fields, and are relatively small (close to the beam size). We therefore applied the following selection criteria to remove spurious sources from the final catalogue: 1) any source touching an edge that has a projected (footprint) size of less than 5 beams[5] (where the angular size of the beam is taken to be $\Omega_{\rm mb} = \theta^2_{\rm mb}\,\pi/(4ln(2)) \approx 888\,{\rm arcsec}^2$, e.g. Kauffmann et al. 2008);

[5] This size was determined by inspecting the datacubes. Unlike in the middle of the map where the noisy spikes are of the order of a beam size, the noisy spikes in the edges are typically much larger than a beam size due to gridding/convolution of the data whilst doing the data reduction. We also consider that even if some real sources were to be included in this criterion, those clouds would be both small and incomplete (since they touch an edge), and therefore their properties would be highly unusable.

2) any source whose projected footprint size is less than two beam sizes; and 3) any source whose signal-to-noise ratio was less than 3.5 (estimated by taking the peak of emission and comparing it to the local noise level). Most spurious sources were successfully removed with this set of criteria, but some remained, in particular towards the noisier high-velocity end of the spectrum (at positive velocities, which can be clearly seen on the ℓv plots of Fig. 1, and in Figs. A1 to A5). Therefore, we applied another criterion to our removal procedure: 4) any source outside the Galactic centre region (i.e. outside a $+355 < \ell < 10^{\circ}$ range), and with a centroid velocity $v_{\rm lsr} > 160\,{\rm km\,s^{-1}}$. The resulting catalogue contains 10663 molecular clouds (whose masks are shown as colours in Fig. 1).

By comparing the integrated intensities inside the cloud masks with the total integrated intensities along each sight line, we estimate that the extracted clouds contain $\sim$ 70% of the total integrated flux above $3\sigma_{\rm w}$ (similar to Barnes et al. 2016), and $\sim$ 50% of the flux above $2\sigma_{\rm w}$, where $\sigma_{\rm w}$ is the standard deviation of the total integrated intensity map, defined as $\sigma_{\rm w} = \sqrt{N_c}\sigma_{\rm rms}\Delta v$, with N_c being the total number of channels used for the integration, $\sigma_{\rm rms}$ the average noise level per channel (i.e. 0.7 K), and Δv the channel width (i.e. $0.25\,{\rm km\,s^{-1}}$). This suggests there is a non-negligible amount of molecular gas in a relatively diffuse inter-cloud medium. In addition, from the datacubes with the cloud masks, we find that of all ℓb pixels with clouds, we have ~82% of sight lines with a single cloud assignment, meaning that only ~18% of the lines of sight have multiple clouds (~16% with two clouds, ~2% with three clouds, and <1% with more than three clouds).

3.2 Data products: Cloud masks and catalogues

From our SCIMES extraction, we have produced two main data products: a catalogue with the properties of all the molecular clouds; and the respective assignment datacubes in the same format as the input 3D datacubes of emission. These data products are made publicly available alongside the data release of the survey[6].

In the assignment datacubes, each voxel holds the unique ID number of the cloud it has been assigned to by SCIMES, and the voxels with no assigned cloud take the value −1. These assignment datacubes are particularly useful for performing further studies on specific clouds, as they can be used to assign voxels to clouds, and therefore pull out the entire 3D structure of clouds from the original emission datacubes. Figure 4 shows an example of the results from the cloud extraction towards a small portion of the survey, with the ^{13}CO peak intensity map in greyscale, and the cloud masks overlaid as colours. In Appendix A, we show the same images for the entire survey coverage (from Fig. A1 to Fig. A5).

All the properties held in the catalogue of molecular clouds produced whilst running SCIMES are listed in Table A1 (in App. A). In essence, the catalogue contains two sets of properties: the directly measured quantities, and the physical properties derived from these after a distance has been assigned (see Sect. 4). Note that all the quantities we present in the catalogue were estimated using the default "bijection" paradigm, which is the most appropriate for characterising substructures within the nested dendrogram tree (Rosolowsky et al. 2008). Amongst the directly measured properties are the ID number, the cloud name, the clouds' centroid longitude (ℓ), latitude (b), and velocity (v), the velocity dispersion (σ_v), the projected footprint area ($Area$) and the respective equivalent radius

[6] **[website placeholder]**

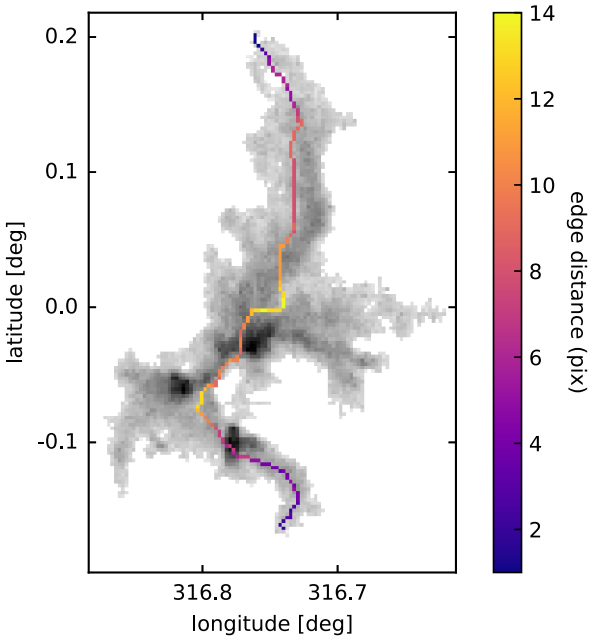

Figure 5. Example of the medial axis for a molecular cloud in our sample (SDG316.766-0.020), which corresponds to an IRDC (SDC316.786-0.044 from Peretto & Fuller 2009), and the larger cloud often shortened to G316.75 (e.g. studied in Watkins et al. 2019). The grey scale shows the ^{13}CO (2-1) integrated intensity, estimated using the voxels within the cloud's mask as defined by SCIMES, and the coloured pixels show the geometrical medial axis, colour-coded with the distance to the external cloud edge.

(R), the average integrated intensity ($< I_{^{13}\rm CO} >$), and the peak intensity ($T^{\rm peak}_{^{13}\rm CO}$). We also include a tag ($edge$) to indicate whether a cloud touches an edge of the survey coverage, in which case it is an incomplete object.

Given that some clouds will be close to the resolution element of our survey, a beam deconvolution on the sizes is needed. This will only affect the smaller objects, and has only very marginal effects on the statistical properties that we derive. Nevertheless, in the catalogue we also provide the equivalent radius deconvolved from the beam (R^d).

In addition to the properties already described, we also estimated some basic parameters to characterise the clouds' morphology. First, we estimated the projected semi-major and semi-minor axes from the second moment of the emission in 2D, weighted by the intensity ($major$ and $minor$), along with the respective position angle (PA), and the aspect ratio ($AR_{\rm mom} = major/minor$). However, this moment method is relatively limited in providing a good approximation of a cloud's morphology, and can easily underestimate the true aspect ratio. Therefore, we also determined the projected geometrical medial axis of the clouds, which is the longest running spine along the 2D-projected cloud's mask, which is farthest away from the external edges (any internal holes in the cloud's masks are filled before determining the medial axis). From that, we include in the catalogue also the medial axis length ($length_{\rm MA}$), as well as the medial axis width as being twice the average distance to the cloud edge ($width_{\rm MA}$), and the corresponding aspect ratio ($AR_{\rm MA} = length_{\rm MA}/width_{\rm MA}$). Figure 5 shows an example of this medial axis for a cloud in our sample. Note that this is a purely geometrical medial axis (i.e. it is built on the assignment masks, with no information on the actual structure of the emission), and

thus it is only a first approximation of the possible filamentary nature of clouds. A more accurate description of filamentary structures detected with ATLASGAL using the SEDIGISM survey data has been performed by Mattern et al. (2018), and shall be expanded to the entire SEDIGISM survey in future work.

The determination of the physical properties of the clouds requires a distance to be assigned. In Section 4 we detail the procedures that we followed to determine distances to the SEDIGISM clouds. Once the distances have been assigned, we can compute the physical properties of clouds. In the catalogue, besides the measured sizes in angular scales, we also present the sizes in physical scales, i.e. already converted using the assigned distance.

We then estimated a few other physical properties, which required using an $X_{^{13}\mathrm{CO}(2-1)}$ conversion factor between the integrated intensities of ^{13}CO (2-1) and the H_2 column densities. We adopted $X_{^{13}\mathrm{CO}(2-1)} = 1^{+1}_{-0.5} \times 10^{21}$ cm^{-2} (K km s^{-1})$^{-1}$, as estimated in the SEDIGISM science demonstration field (Schuller et al. 2017), by comparing the SEDIGISM ^{13}CO emission to the H_2 column densities as derived from the Hi-GAL survey data (Molinari et al. 2010)[7]. With this $X_{^{13}\mathrm{CO}(2-1)}$, and assuming a mean molecular weight μ_{H_2} of 2.8 (Kauffmann et al. 2008), we derived the clouds' masses (M), average gas surface densities (Σ), and virial parameter (α_{vir}), defined as $\alpha_{\mathrm{vir}} = 5\sigma_v^2 R/GM$ (Bertoldi & McKee 1992), where G is the gravitational constant, σ_v the velocity dispersion, and R is the equivalent radius. This formulation of α_{vir} assumes a spherical geometry and a uniform density, and it only takes into account the balance between kinetic and gravitational energies. Thus, α_{vir} is a very simplistic tool, and it should not be taken as a strict measurement of the gravitationally bound state of a cloud (e.g., Bertoldi & McKee 1992; Kauffmann et al. 2013; Traficante et al. 2018a,b). However, given its wide usage in the literature, we estimate it here to allow a direct comparison of our results with those of other surveys.

Finally, in the catalogue we provide the surface density and the virial parameters using both the measured R (noted as Σ and α_{vir}), and the deconvolved R^d (noted as Σ^d and α^d_{vir}). For the analysis presented in this paper we will use the deconvolved properties, although this choice has only a very marginal effect on the respective distributions, keeping the global trends virtually unchanged. Given the uncertainties on the distance estimates (which are of the order of $\sim 30\%$) and on the X_{CO} factor (of a factor two), all these quantities have an uncertainty of at least a factor two.

In the catalogue, we also provide the Heliocentric and Galactocentric coordinates of each cloud, determined as explained in App. B.

4 DISTANCE DETERMINATION

In order to compute the physical properties of clouds, we require knowledge of the distances. However, for a large survey such as SEDIGISM, there are very few existing direct measurements of the distances towards molecular clouds, and we mostly need to rely on estimates based on the kinematic distances (i.e. by assuming a Galactic rotation model, see Sect. 4.1), which rarely give a unique answer. Therefore, it is often required to search for ancillary indications to narrow down the distance assignment. In the following sections, we describe the computation of the kinematic distances, and how the problem of the kinematic distance ambiguities (KDA) were solved. The kinematic distance solutions, along with their uncertainties, and our final decisions are listed in the catalogue. For each cloud we include two distance tags: d_{sol} specifies the type of distance solution, and d_{flag} specifies the method used to reach the final distance assignment. The numbering of d_{flag} reflects the order by which we check the different methods. Once a cloud gets a distance as per a given tag, we stop testing further methods. The flowchart depicting this decision process is shown in Fig. 6. These methods are all described in detail in Sect. 4.2, and summarised in Table 1.

Table 1. Summary of the methods used to determine the distances of clouds, along with the number of clouds that had their distances assigned with each method.

d_{flag}	Description	Nb. clouds
−1	No distance information	363
0	Exact maser parallax distance	11
1	No distance ambiguity	551
2	Tangent distance	1080
3	Dark Cloud (near distance)	77
4	IRDC (near distance)	751
5	Literature HI SA (near distance)	91
6	Direct HI SA measurement (near distance)	828
7	ATLASGAL source at near distance	252
8	Solomon distance to GP (near distance)	34
9	Size-linewidth scatter (near or far distance)	2263
10	ATLASGAL source at far distance	142
11	Extinction (near or far distance)	3178
12	Ambiguity not solved (defaulted to far)	1042

4.1 Kinematic distances

To derive the kinematic distances of the clouds in our catalogue, we have used the Galactic rotation model of Reid et al. (2016), which has been constructed using maser parallax distance measurements. This model uses the revised values for R_0 and V_0 of 8.34 kpc and 240 km s^{-1}, respectively. Besides these rotation curve parameters, this model also uses a Bayesian approach that can consider the source's proximity to spiral arms, displacement from the Galactic mid-plane and proximity to parallax sources to estimate the most likely distance. Since molecular clouds are not always confined to the spiral arms or associated with star formation we have relaxed those constraints.

Some clouds, however, have velocities that lie outside those allowed by the rotation model, and thus we are unable to assign them a distance. This is the case for 363 clouds, and they can be identified in the catalogue with the distance solution tag d_{sol} = NULL (and $d_{\mathrm{flag}} = -1$)[8]. For the remaining clouds, if they lie outside the Solar circle, there is a unique kinematic distance solution. This is the case for 551 clouds, and these can be identified in the catalogue with the tag d_{sol} = NA, standing for *No Ambiguity* (and $d_{\mathrm{flag}} = 1$). When sources are located within the solar circle, there are two possible distance solutions, a *near* and a *far* one, which are equally spaced on

[7] The Hi-GAL column density maps for this calibration were built by fitting a pixel-by-pixel grey body curve to the spectral energy distribution from 160 to 500 μm (Elia et al. 2013), assuming a dust to gas ratio of 1:100, and an opacity law with a fixed spectral index $\beta = 2$, and $\kappa_0 = 0.1$ cm^2 g^{-1} at $\nu_0 =$ 1200 GHz (Hildebrand 1983).

[8] Note that in the catalogue we assign these clouds a distance of −1, which effectively means that we have estimated their physical properties as if they were at 1 kpc distance, and that properties that have a linear dependency with distance will appear with negative sign.

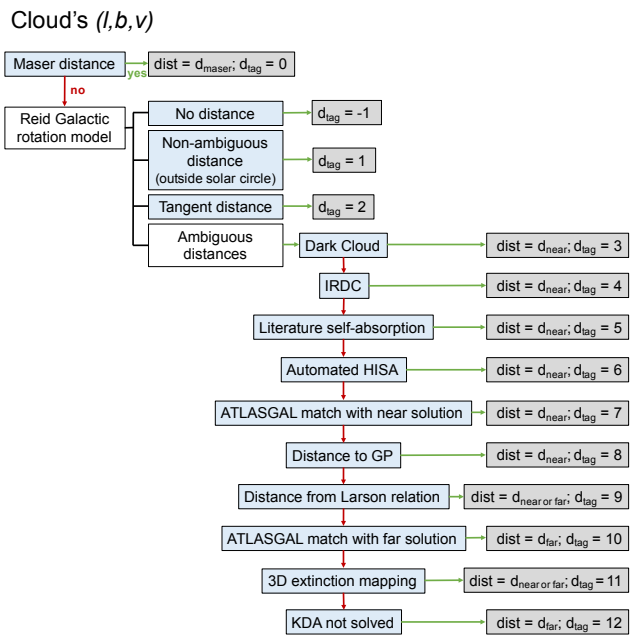

Figure 6. Flowchart showing the distance assignment procedure adopted for the SEDIGISM clouds. The blue boxes highlight the methods used, and the grey boxes show the corresponding assigned distance and tag. The green and red arrows show the directions taken if a specific method succeeds or fails in providing a distance solution, respectively.

either side of the tangent distance. Clouds that lie close to the tangent velocities (i.e. within 5 km s^{-1}, to accommodate for uncertainties due to streaming motions, e.g. Brand & Blitz 1993; Wienen et al. 2015) were assigned the tangent distance, and given the tag $d_{\rm sol}$ = T (and $d_{\rm flag}$ = 2). This is the case for 1080 clouds.

For sources with two possible distances, we performed an extensive cross-match with literature information, checked directly for HI self-absorption (HISA) in each cloud, and checked whether the cloud properties would make them statistically more likely to be at a specific distance solution, in order to solve the distance ambiguity. Upon completion of this procedure, clouds that were assigned a near distance were tagged in the catalogue with $d_{\rm sol}$ = N (corresponding to a total of 3679 clouds), while far distance clouds have $d_{\rm sol}$ = F (which amount to 4979 clouds). The full details on the procedure leading to our final distance decision are described in the following section.

Note that, despite our extensive effort in assigning distances to clouds, there are regions within our Galaxy for which we know that our kinematic distances are not reliable. We have therefore included a flag in the catalogue, $d_{\rm reliable}$, which identifies clouds for which the distances are unreliable or nonexistent ($d_{\rm reliable}$ = 0) and those that have a reliable distance estimate ($d_{\rm reliable}$ = 1). In particular, we have given a $d_{\rm reliable}$ = 0 for clouds with a $|v_{\rm lsr}| < 10$ km s^{-1} with a near distance assignment. For those, the kinematic distance is too uncertain, since the $v_{\rm lsr}$ of the clouds is dominated by local motions, and therefore the distance assigned from a global rotation model has a distance uncertainty on the order of the distance value itself. We also assigned a $d_{\rm reliable}$ = 0 to clouds for which we were not able to solve the distance ambiguity (i.e., clouds with a $d_{\rm flag}$ = 12, see Sect. 4.2.4). In addition, clouds towards the Galactic centre (and including most of the Galactic bar), i.e. within $+353° < \ell < 7°$, also have a very uncertain distance estimate (and are given a $d_{\rm reliable}$ = 0), as the Galactic rotation model used for our kinematic distance assignment is not tailored to reproduce the complex dynamics of the gas in the centre of the Galaxy. The only exception being the clouds for which we have a maser parallax distance (as that is an exact measurement, independent of kinematic considerations), which are retained with a $d_{\rm reliable}$ = 1. The Galactic Centre will be studied in more detail in future work, and we will then revise the catalogued distances for those clouds accordingly.

4.2 Solving the distance ambiguities

4.2.1 Maser parallaxes, Dark Clouds, IRDCs, and HISA from literature

We performed a cross-match of our entire catalogue with literature information for any known robust indication of the distance of our clouds. We started by cross-matching our clouds with a compilation of known maser parallax measurements (Reid et al. 2009, 2014; Wu et al. 2014; Honma et al. 2012; Bobylev & Bajkova 2013). The matches were performed by checking if the position of the masers (in 3D) fell inside the mask of one of our SEDIGISM clouds. Sources with a known maser parallax measurement were assigned their maser parallax distance (instead of the kinematic distance). If there were more than one maser parallax measurement for a given cloud, then we take the average parallax distance. Clouds with a maser distance were given a $d_{\rm sol}$ = M and $d_{\rm flag}$ = 0, and this was the case for 11 clouds. The small number of SEDIGISM clouds with a maser parallax is due to the fact that most of the maser parallax catalogues cover Quadrants 1, 2 or 3, hence only very few maser parallax distances have been measured for sources in our longitude range, and of those, about half lie outside our latitude range.

We then did a cross match with other literature catalogues (including dark clouds, infrared dark clouds (IRDCs), and HISA), using the clouds' centroid Galactic coordinates and velocity. For catalogues in which the major axes, minor axes and position angles are given, the match was done by checking if the centroid position of the SEDIGISM cloud falls in the elliptical footprint of the catalogued source. For catalogues that give no position angle, or provide only the beam size or a radius, we use the effective radius and the match is done by checking if the centroid of the SEDIGISM cloud falls in the defined circular footprint. For catalogues which have velocity information, besides the spatial match, we require that the velocity difference between the SEDIGISM cloud and the catalogued sources must be less than 6 km s^{-1} (assumed to be the typical cloud-cloud velocity dispersion, e.g. Stark & Lee 2006; Wilson et al. 2011).

Using these criteria, we cross-matched our clouds with catalogued Dark Clouds with velocity information (Otrupcek et al. 2000), as well as with IRDCs, some with and some without velocity information (Simon et al. 2006, Jackson et al. 2008, Du & Yang 2008, Peretto & Fuller 2009, Chira et al. 2013, Liu et al. 2013). Their extinction makes Dark Clouds and IRDCs appear in silhouette against a bright background (in the visible and in the IR, respectively). Dark Clouds typically reach high optical depths very quickly, and thus are typically tracing nearby clouds that absorb the stellar light from the Galactic disc. The IRDCs probe a higher column density regime, which means we observe deeper into the molecular clouds. Nevertheless, the concept is the same, in that we are more likely to see a cloud in extinction, if there is enough IR background to absorb against, thus placing such clouds preferably at their near distance solution (although this might not always be the case, e.g. Giannetti et al. 2015, found ~ 10% of IRDCs to be located at the far distance). For our purpose, we have assumed any

SEDIGISM sources which have a Dark Cloud, or an IRDC match to be at the near distance, and given a $d_{\rm flag}$ = 3 or 4, respectively. Note that, in cases where the cross-match with IRDCs was only spatial (i.e. in the absence of available velocity information), we only consider the match to be reliable if there is a single SEDIGISM cloud associated with each IRDC: if an IRDC is in the same line of sight as multiple SEDIGISM clouds, more information - such as velocity information or a more detailed morphological match - would be needed in order to produce a robust association.

We also cross-matched our catalogue with known HISA (or H_2CO absorption) features from the literature within the SEDIGISM coverage (Anderson & Bania 2009; Anderson et al. 2015; Wienen et al. 2015; Sewilo et al. 2004; Pandian et al. 2008; Busfield et al. 2006; Urquhart et al. 2012). HISA occurs when cold HI gas in the foreground absorbs the warmer HI emission from background gas at the same velocity (e.g., Gibson et al. 2000). Therefore, the existence of HISA at a given velocity is often used as an indication that the cold gas that is absorbing is at a near distance[9] - as this makes it more likely to have background emission to absorb against, and that emission is less likely to be filled by other warmer HI emission along the line of sight between the observer and the cold cloud (e.g., Roman-Duval et al. 2009). SEDIGISM sources with a known HISA feature from the literature were assumed to be at the near distances, and given a $d_{\rm flag}$ = 5.

4.2.2 *Direct and automated HISA determination*

Given that many of the clouds in our catalogue do not have a counterpart with literature sources (given the improved sensitivity and resolution of the SEDIGISM survey), we have also checked for the presence of HISA directly for each individual cloud. We have done so in an automated way, making use of HI 21 cm ATCA and Parkes data from both the Southern Galactic Plane Survey (SGPS; McClure-Griffiths et al. 2005), and the ATCA HI Galactic Centre Survey (McClure-Griffiths et al. 2012). Both these datasets have a spatial resolution of 2′, a spectral resolution of 1 km s^{-1} and an average noise level, $\sigma^{\rm HI}_{\rm rms}$, of ∼ 1 K. We combined the data from these surveys into a single datacube, using the CONVERT[10] and KAPPA[11] packages from the Starlink software (Currie et al. 2014), namely the WCSALIGN and WCSMOSAIC procedures. We then extracted subcubes covering the same spatial and velocity range of each of the SEDIGISM tiles, which we reprojected and resampled to the same pixel and channel sizes as the SEDIGISM data. Even though this procedure heavily oversamples the HI data, it facilitates the automated check of the HI emission in each of the SEDIGISM clouds, by directly using the assignment masks produced by SCIMES.

Our automated HISA procedure works in the following way: First, it selects all the voxels that belong to each cloud, and creates a projected 2D image of the cloud in the plane of the sky. This allows us to identify all lines of sight which belong to that cloud. Then, for each sight line, it determines the "background" HI emission, by taking the HI emission one channel before, and one channel after the cloud's velocity range in that specific sight line, and fitting it with a linear function (see illustration on the first column of Fig. 7). We then subtract the HI emission inside the cloud from the fit of the background HI emission, on a channel by channel basis. This "subtracted" datacube should have negative values whenever the HI content of a cloud is self-absorbing against the background HI emission (see the second column of Fig. 7). Therefore we use these background-removed HI datacubes as the basis for our decision on whether a given sight line in a cloud has a significant HISA or not (see last two columns of Fig. 7).

To determine if a specific line of sight has HISA, we impose two conditions:

(i) The minimum intensity of the background-removed HI emission signal is lower than $-3\sigma^{\rm HI}_{\rm rms}$. This ensures that the self-absorption is significant, given the noise in the HI data.

(ii) The sum of the background-removed HI signal is negative, and has an absolute value larger than 3 times the cumulative noise, given by $3\sqrt{N}\sigma^{\rm HI}_{\rm rms}$, where N is the number of velocity channels across which the signal was summed up.

Step (ii) ensures that false positives are rejected. A false positive typically occurs when our simple background fit does not capture properly the variations of the HI background emission (e.g. by under- or over-estimating the slope of the HI background emission), producing a signature similar to a p-Cygni profile, whose dip may be deeper than the HI noise – thus passing our criteria (i) (see Case 3 of Fig. 7). However, while a true self-absorbed profile would have negative emission throughout the entire cloud velocity range, resulting in the sum of the background-removed HI emission to be also negative (and significant), a false positive would have a sum that is within the noise of the HI data. We therefore use this criterion to remove potential false positives.

We then consider a cloud to have strong HISA only if the number of sight-lines (i.e. 2D pixels) that satisfy condition (i) amount to at least one beam size in the HI data, and that satisfy condition (ii) amount to at least one SEDIGISM beam size. The results from this automated HISA determination are compiled in the catalogue under the *tag_hisa* property, which is assigned a value of 1 for strong HISA, 0 if it is ambiguous (i.e. meeting only some of the criteria above), and −1 if there is no HISA. Clouds with a strong HISA from this method are taken to be at a near distance, and given a $d_{\rm flag}$ = 6.

4.2.3 *ATLASGAL distances*

The ATLASGAL survey (Schuller et al. 2009; Beuther et al. 2011) observed the dust continuum emission towards the inner Galactic plane at 870 μm, and produced a catalogue of 10163 compact sources[12] (CSC catalogue; Contreras et al. 2013; Urquhart et al. 2014c). In order to determine the distances to these clumps, there was a significant effort in assigning velocities to the continuum emission through a combination of extensive cross-match with molecular line data reported in the literature and dedicated follow-up observations (Wienen et al. 2012, Csengeri et al. 2016, Wienen et al. 2018, Urquhart et al. 2019). This was then combined with the

[9] Note that not all near-distance clouds are expected to show a strong HISA feature, given the simultaneous requirement of: 1) the existence of significant cold HI gas at the same velocities as the molecular cloud traced by ^{13}CO; 2) the existence of warm HI gas in the background, at the same velocity as the cloud; and 3) the non-existence of intervening warm HI gas between us and the cloud, that could fill in the cloud's intrinsic HISA. In addition, HII regions can also produce a direct HI absorption feature, even at the far distances, which can be mistakenly interpreted as a HISA feature. The HISA method for solving the KDA is therefore estimated to be ∼80% reliable (e.g. Anderson & Bania 2009)

[10] http://starlink.eao.hawaii.edu/docs/sun55.htx/sun55.html

[11] http://starlink.eao.hawaii.edu/docs/sun95.htx/sun95.html

[12] Note on nomenclature: we will refer to the ATLASGAL compact sources as "clumps", as opposed to the larger scale SEDIGISM structures that we refer to as "clouds".

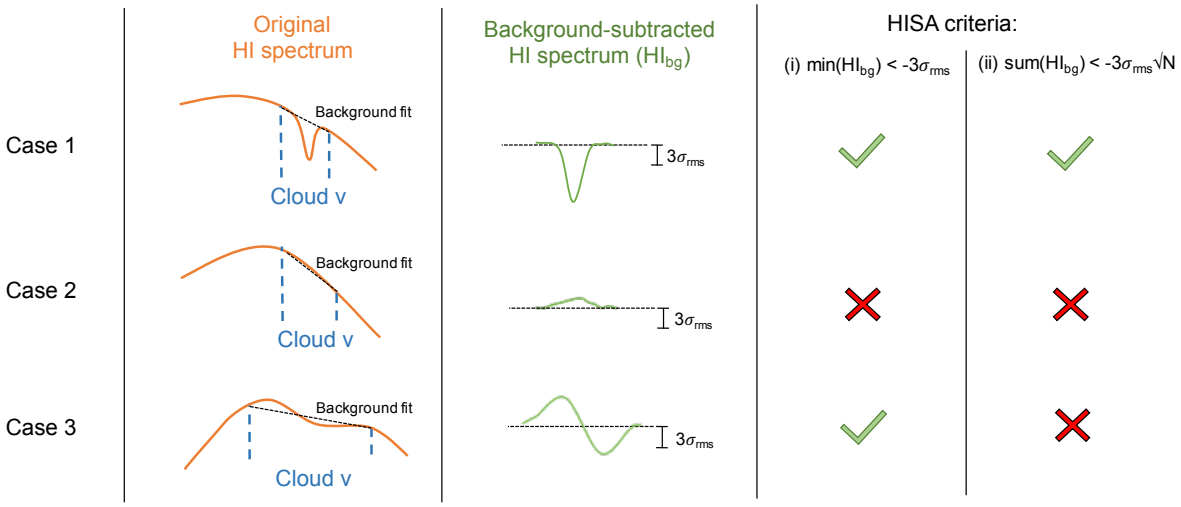

Figure 7. Three sketch examples of the automated HISA method, showing the original HI spectrum on the first column (in orange), with the cloud's velocity ranges denoted in blue, and the linear background fit done to the HI spectrum in black dotted line. The second column shows the cloud's background-subtracted HI spectra (in green), with the dotted dashed line representing the 0-emission level, and the vertical bar representing $3\sigma_{\rm rms}$ of the HI emission. The two last columns show the criteria that we use to infer whether there is HISA in that particular sight line. Case 1 represents a line of sight with strong HISA, but the other two cases are not considered to have HISA. Case 3 shows an example of a false positive arising from criterion (i) alone, but which is mitigated by introducing criterion (ii).

Reid et al. (2016) Galactic rotation curve to calculate kinematic distances, and the distance ambiguities were resolved using the HISA method and using a friends-of-friends clustering algorithm to identify complexes. This successfully determined the distances to ~8000 ATLASGAL clumps (see Urquhart et al. 2018 for details).

Since all of the SEDIGISM survey is covered by ATLASGAL, we performed a cross match between all clouds in our sample, to the ATLASGAL clumps with known $v_{\rm lsr}$ from Urquhart et al. (2018). This cross match was done by considering the centroid positions and velocities of the ATLASGAL clumps, and placing them in the respective voxel in our 3D datacubes. We then checked whether that voxel falls within the mask of a SEDIGISM cloud (i.e. a *perfect match*), and otherwise estimate the distance to the nearest SEDIGISM cloud (in all 3 dimensions). We then consider ATLASGAL clumps that lie within one beam size of the edge of the nearest cloud, or within one σ_v of the cloud, to be a *partial match*. Out of the 5067 ATLASGAL sources within the SEDIGISM coverage, 4376 were matched as a *perfect match* to a SEDIGISM cloud, and 448 as being a *partial match*, leaving only 243 ATLASGAL clumps without a SEDIGISM counterpart. Most of these unmatched ATLASGAL clumps are either small clumps whose corresponding SEDIGISM emission did not satisfy our minimum size requirement, or they are in regions that form part of a smoother background that does not get assigned to a specific cloud (i.e. where the ^{13}CO emission does not have a local peak rising above the 4-$\sigma_{\rm rms}$ requirement to be considered as independent peaks/leaves within the dendrogram). In total, these 4824 ATLASGAL clumps are contained within 1709 SEDIGISM clouds (i.e. ~ 16% of SEDIGISM clouds).

Given that the distances to the ATLASGAL sample were estimated with the individual $v_{\rm lsr}$ of clumps (rather than that of the parent cloud), we do not use their distances directly. Instead, we are only interested in the type of distance solution determined for each ATLASGAL clump (*near* or *far*), in order to incorporate it in our distance assignment. In most cases, all ATLASGAL clumps within a given SEDIGISM cloud have a distance solution that agrees amongst them. However, there are a few cases where, within a SEDIGISM cloud, there are ATLASGAL clumps with both a *near* and *far* solution. In those cases, we define the "global" ATLASGAL solution as being *near*, under the assumption that an indication for a near distance solution is more reliable than the absence of one (which is the most common reason for a far distance assignment). Note that, even though we had to do this step to provide a complete list of "ATLASGAL distance solutions" for our SEDIGISM sample, none of the clouds for which the ATLASGAL distance solutions disagreed, actually took their final solution from ATLASGAL (instead they had their KDA lifted by other methods).

For SEDIGISM clouds with an ATLASGAL match, and for which the criteria in Sect. 4.2.1 and 4.2.2 did not have an indication for a near distance, we check the distance solution from ATLASGAL. If that solution is *near*, then we adopt the near distance, and assign a $d_{\rm flag} = 7$. If the ATLASGAL solution is *far*, and there are no other indications of a *near* distance solution (from methods 8 and 9, see Sect. 4.2.4), then we assign the far distance, and a $d_{\rm flag} = 10$.

4.2.4 *Other distance indicators*

In addition to the above methods, we also checked two often-used techniques that take the statistical distribution of the properties of molecular clouds into account. The first one is the method used by Solomon et al. (1987), which considers the physical distance of a cloud to the Galactic plane, should the cloud be assigned the far distance. If by taking the far distance the cloud is too far off the Galactic plane (i.e. > 140 pc, which is the scale height of the gaseous Galactic disc, e.g. Solomon et al. 1987; Tavakoli 2012), then the near distance is favoured, and the cloud is given a $d_{\rm flag} = 8$. Note that towards the far side of the Sagittarius and Scutum-Centaurus arm (around $\ell \sim 290°$), the Galactic mid-plane is known to be warped towards negative latitudes (e.g. Chen et al. 2019; Romero-Gómez et al. 2019). This implies that on the far-distance side, in the latitude range of $300° < \ell < 318°$, the Galactic plane descends below a latitude of $-0.5°$ (e.g. Reid et al. 2016), and therefore this

area of the Galaxy is not well covered by our survey (since we cover a relatively narrow b range). Nevertheless, since the Galactic warp only becomes significant at Galactocentric distances of 8 kpc and beyond, any clouds in the longitude range of $300° < \ell < 318°$ possibly following the warp are beyond the Solar circle, and should have unambiguous distances. Therefore, our criterion checking for the height above the Galactic plane is not affected by the existence of the Galactic warp.

The second method places each cloud on the size-linewidth relation ($\sigma_v - R$) (Larson 1981; Solomon et al. 1987, e.g.), for both near and far distance solutions, and checks which solution provides the smaller distance to the empirical relation. We use this method to favour a given distance solution *only* if one solution is significantly closer to the empirical relation than the other solution (i.e. at least a factor 3 difference in log-space). More details on this method can be found in App.D, and clouds that used this criteria were given a $d_{\rm flag} = 9$.

Finally, we also used a method based on datacubes of the visual extinction in K-band, as a function of distance (Marshall et al. 2006, Marshall et al. in prep, Elia et al. in prep). From those cubes, the structures of significant extinction can be identified along each line of sight, by taking the distances at which the extinction has a significant jump. We then compare the extinction distances with the near and far kinematic distance estimates, by taking into account a 30% uncertainty on the extinction distance as well as the kinematic distance uncertainties. We solve the KDA by taking the kinematic distance which has an extinction counterpart, if one exists. Clouds that used this criteria were given a $d_{\rm flag} = 11$[13]. In the future, this could potentially be expanded to also include Gaia-based 3D dust extinction maps (e.g. Lallement et al. 2019), although at the moment these only probe distances up to 3 kpc.

4.3 Revisiting previous distance estimates

In order to gauge how our distance estimates compare to the results from other surveys that covered the same area of the Galactic plane, we have compared the results from our distance solutions to those of the ATLASGAL survey, as a reference (given that ATLASGAL had already performed a detailed comparison with other surveys, e.g. Urquhart et al. 2014b, 2018). Note however, that as per Section 4.2.3, only 1709 SEDIGISM clouds have an ATLASGAL counterpart (i.e. only $\sim$ 16% of our sample), although this includes 95% of all ATLASGAL clumps in our coverage (i.e. 4814 clumps). Of those, 1253 ATLASGAL clumps did not have an assigned distance, which we have now assigned[14]. For the ATLASGAL sources with a distance, the KDA solution between the two surveys agrees for 3080 ATLASGAL clumps. This leaves a total of 481 clumps (i.e. 13.5% of the ATLASGAL clumps with distances) with a distance solution that was revised by us, in most cases from a far distance to a near distance, by one of the other methods listed in Section 4.1. Most of these revisions were done using our HISA method (321 clumps, $d_{\rm flag} = 6$), followed by 60 clumps revised using the IRDC matches ($d_{\rm flag} = 4$), and 53 clumps with literature HISA ($d_{\rm flag} = 5$). A further 23 clumps were revised using the maser parallax measurements ($d_{\rm flag} = 0$), 7 clumps using the distance around the Larson relation ($d_{\rm flag} = 9$), and 2 clumps using the Dark Cloud association ($d_{\rm flag} = 3$). Finally, 2 clumps were re-assigned a near distance for being in the same complex as other ATLASGAL sources with a near distance ($d_{\rm flag} = 7$), 1 clump was revised as having a non-ambiguous solution ($d_{\rm flag} = 1$), and 12 clumps had their distances revised to a tangent distance ($d_{\rm flag} = 2$), although for these cases the change from near or far solutions into the assumed tangent distance is within the uncertainties.

With the large survey coverage, and improved resolution and sensitivity of the SEDIGISM survey compared to other spectroscopic surveys covering the same Galactic longitudes (e.g. the MopraCO survey Burton et al. (2013), the ThrUMMS Barnes et al. (2015), and the Dame et al. (2001) survey), here we present the most extensive sample of molecular clouds towards the inner Galaxy yet, with 10663 clouds in total. With our comprehensive effort to combine different independent methods to determine the distance solutions for each SEDIGISM cloud, we have been able to assign distances to 10300 clouds, 7993 of which have well-characterised (reliable) distance assignments.

5 GLOBAL PROPERTIES OF THE SEDIGISM SAMPLE

For our analysis of the statistical properties of the SEDIGISM molecular clouds, we have excluded any clouds whose projected footprint size is smaller than 3 beams (i.e. any clouds that are barely resolved). We also excluded clouds with an unreliable distance ($d_{\rm reliable} = 0$), and those that are incomplete because they touch a survey coverage edge ($edge = 1$). With these criteria, we select a total of 6664 clouds for our analysis, which we will refer to as our "science sample". In addition, we will refer to the science sample above the completeness limits (as per App. C) as our "complete science sample" (which also exclude clouds at a tangent distance - see Sect. 6 for more details). Table 2 summarises the specific details of the several samples that we use in the paper.

5.1 Distribution of individual properties

Figure 8 shows the distributions of a number of different properties, namely the total mass (M), the velocity dispersion (σ_v), the medial axis length ($length_{\rm MA}$), the average surface density (Σ), the virial parameter ($\alpha_{\rm vir}$), and the aspect ratio from the medial axis ($AR_{\rm MA}$). The histograms correspond to the full science sample (in light grey), from which we highlight the subset of clouds with an ATLASGAL counterpart (in dark grey), and from those, also clouds with a signpost of high-mass star formation (HMSF, in red), as per Urquhart et al. (2014b). These signposts of HMSF include the existence of methanol masers (Urquhart et al. 2013a, 2015, which used the masers from the Methanol Multibeam Survey, Caswell et al. 2010; Green et al. 2012); HII regions (Urquhart et al. 2013b, which combined information from the CORNISH survey, Hoare et al. 2012; Purcell et al. 2013, and the GLIMPSE survey, Benjamin et al. 2003); or massive young stellar objects, YSOs (Urquhart et al. 2014b, which matched ATLASGAL sources with YSOs and HII regions identified by the Red MSX Source (RMS) survey, Lumsden et al. 2013; Urquhart et al. 2014a). In total, we have 435 SEDIGISM clouds within the full sample (330 in the science sample, i.e. $\sim$ 4%

[13] This method does have a few limitations, one being that it becomes less reliable for far distances, mainly as the extinction cubes have a pixel of 5′, and therefore roughly 10 times larger than the SEDIGISM beam size. Small clouds assigned a distance using this method should therefore be used with caution.

[14] Note that most of these are towards the central Galaxy, for which the kinematic distances are less reliable. If we consider only the sample we use for science (as per Sect. 5), the number of ATLASGAL clumps that so far did not have a distance assigned and for which we are able to assign a reliable distance is 308.

Table 2. Summary of different samples

Sample Name	Description/conditions for selection	nb of sources
Full sample	Entire catalogue, with distances ($d_{\rm flag} \neq -1$)	10300
Science sample	$d_{\rm reliable} = 1$, $Area > 3\Omega_{\rm beam}$, edge = 0	6664
Distance limited sample	$d_{\rm reliable} = 1$, $Area > 3\Omega_{\rm beam}$, edge = 0, 2.5 kpc $< d <$ 5 kpc	1743
Complete science sample	$d_{\rm reliable} = 1$, $Area > 3\Omega_{\rm beam}$, edge = 0, $M > 2.6 \times 10^3$ M$_\odot$, $R >$ 2.9 pc, $d <$ 14.5 kpc, $d_{\rm flag} \neq 2$	1680

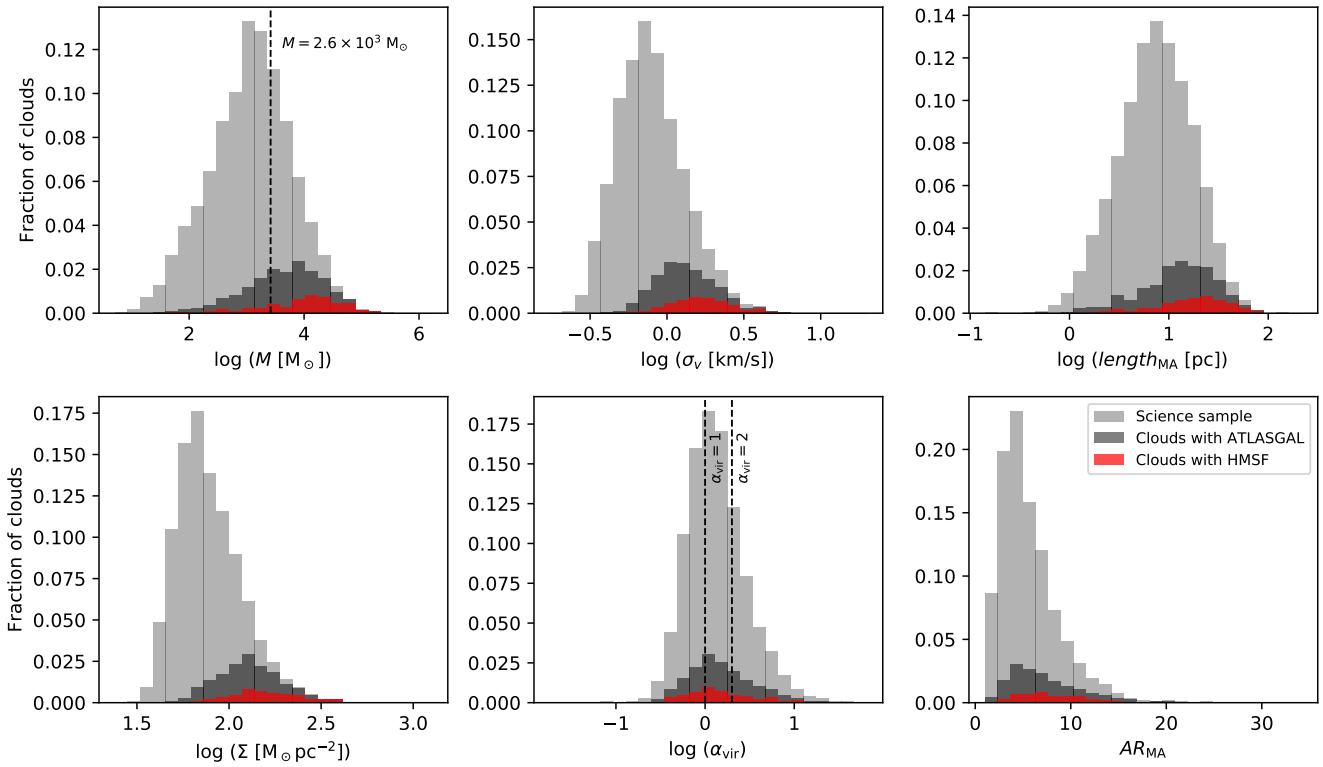

Figure 8. Histograms of global properties: Mass (top-left), velocity dispersion (top-centre), medial axis length (top-right), average surface density (bottom-left), virial parameter (bottom-centre), and aspect ratio from the medial axis (bottom-right). The histograms are for the science sample (light grey), clouds that have an ATLASGAL counterpart (dark grey), and clouds that have a HMSF signpost (red). The normalisation of all histograms was made with respect to the total number of clouds in the science sample. The vertical dashed line on the mass histogram shows our mass completeness limit (see App. C), and the dashed lines on the virial parameter histogram represent an $\alpha_{\rm vir} = 1$ and 2.

of clouds) that have signposts of active HMSF (similar to the fraction of high-mass star forming clouds found by Barnes et al. 2011). We note, however, that for this work, we did not cross-match our SEDIGISM clouds with HMSF tracers directly: our sample of HMSF clouds is purely a subsample of the ATLASGAL sources, and so any HMSF signposts outside that are not accounted for. This will be explored in future work. We also computed the main statistics (i.e. the median, lower and upper quartiles, skewness and kurtosis) of these distributions, plus that of the equivalent radius (R), which we compile in Table 3. These distributions, however, could potentially be affected by our different completeness at different distances within our science sample. In order to check how this might affect the global results, we have also computed the histograms using a distance limited sample (with 2.5 kpc $< d <$ 5.0 kpc), shown in App. E (Fig. E1). The statistics for the distance limited sample are also compiled in Table 3, showing that they follow broadly the same trends as the science sample.

Noticeably, the median values in Table 3 and the histograms from Fig. 8 show that clouds with an ATLASGAL counterpart tend to be at the higher end of the distributions of mass, velocity dispersion, size, aspect ratio, and surface density, as compared to the science sample. This is even more so for clouds with a HMSF signpost (whose median values are again higher than those of the ATLASGAL sub-sample). The increase in the median values of those properties as we go from the science sample to the HMSF sub-sample range from a modest increase of a factor 2 (e.g. for the aspect ratio and velocity dispersion) up to an order of magnitude increase (for the mass). The only exception to this trend is the virial parameter, for which the median values (and the quartiles) are similar between all three subsets.

Interestingly, while the science sample typically has a distribution with a significant tail (i.e. with high kurtosis values), as we move from the full sample to the ATLASGAL sub-sample and then to clouds with a HMSF signpost, the shape of the distribution of

Table 3. Statistics of some of the physical properties of the SEDIGISM clouds, namely the mass (M), velocity dispersion (σ_v), equivalent radius ($R_{\rm eq}$), medial axis length ($length_{\rm MA}$), medial axis aspect ratio ($AR_{\rm MA}$), surface density (Σ), and virial parameter ($\alpha_{\rm vir}$), for the entire science sample, and for a distance-limited sample (to minimise distance-biased results). Within these samples we also list the statistics for the subsets of clouds with an ATLASGAL counterpart or with a HMSF signpost. Q25 and Q75 represent the lower (25%) and upper (75%) quartiles of the distributions.

	Science sample					Distance limited sample (2.5 kpc $< d <$ 5.0 kpc)				
Sub-set	Median	Q25	Q75	Skewness	Kurtosis	Median	Q25	Q75	Skewness	Kurtosis
M [$\times 10^3$ M$_\odot$]										
Science	1.32	0.42	3.69	53.6	3638.0	0.44	0.14	2.05	7.7	95.0
With ATLASGAL	5.19	1.74	13.86	24.2	698.0	3.75	1.21	10.52	4.3	32.4
With HMSF	11.94	3.48	27.24	13.8	220.9	10.21	3.14	23.00	3.1	17.1
σ_v [km/s]										
Science	0.76	0.55	1.08	6.8	160.5	0.73	0.51	1.18	2.8	16.8
With ATLASGAL	1.29	0.97	1.80	7.5	127.5	1.35	0.99	1.93	2.5	14.0
With HMSF	1.66	1.25	2.20	7.7	95.4	1.68	1.28	2.29	2.4	11.5
$R_{\rm eq}$ [pc]										
Science	2.36	1.42	3.68	3.3	36.7	1.34	0.82	2.62	1.9	7.2
With ATLASGAL	3.55	2.19	5.66	1.5	6.9	3.07	1.79	4.54	1.0	3.8
With HMSF	4.79	2.76	6.92	1.4	6.0	4.09	2.66	5.80	0.7	3.1
$length_{\rm MA}$ [pc]										
Science	7.70	4.33	13.67	3.8	41.8	4.90	2.63	10.81	2.1	8.8
With ATLASGAL	13.51	7.80	23.08	1.6	6.4	12.62	6.32	21.10	1.2	4.5
With HMSF	18.78	10.72	29.74	1.2	4.5	16.82	10.73	26.27	0.8	3.4
$AR_{\rm MA}$										
Science	4.9	3.4	7.0	1.6	7.7	5.6	3.9	8.3	1.6	7.0
With ATLASGAL	6.5	4.5	9.5	1.4	6.3	7.6	5.1	10.9	1.4	6.0
With HMSF	7.6	5.3	10.8	1.6	7.5	9.0	6.2	11.7	1.6	7.7
Σ [M$_\odot$pc^{-2}]										
Science	73.2	58.1	99.6	5.1	71.5	73.7	57.7	110.0	3.1	19.8
With ATLASGAL	128.1	98.3	170.2	4.2	42.5	140.0	103.7	190.8	2.2	12.1
With HMSF	158.1	120.4	221.1	3.7	27.6	183.9	137.0	252.6	1.9	8.5
$\alpha_{\rm vir}$										
Science	1.30	0.81	2.11	9.9	187.8	1.84	1.26	2.77	4.2	32.0
With ATLASGAL	1.36	0.81	2.55	7.8	91.3	1.78	1.05	2.97	3.0	16.6
With HMSF	1.28	0.76	2.62	7.0	61.3	1.49	0.93	2.77	2.5	10.3

all properties (except for the aspect ratio) becomes progressively flatter (smaller values of kurtosis) and symmetric (smaller values of skewness) – with HMSF clouds occupying nearly the same parameter space as clouds with an ATLASGAL counterpart but without HMSF signpost. This is rather interesting as it suggests that there is no single "global" property of clouds that is sufficient to determine, on its own and unambiguously, their potential to host high-mass star formation, and perhaps a complex combination of several conditions is needed. It is worth noting that some global properties like magnetic fields are, of course, not considered here. In Section 6.5 we will investigate if the ability to form high-mass stars might instead be influenced by the Galactic environment.

5.2 Scaling relations

Figure 9 shows two of the most common scaling relations in the literature: the size-linewidth relation in the top panels, where the dashed-line represents the Larson relation, $\sigma_v^2 \propto R$ (Larson 1981; Heyer et al. 1998); and the Heyer relation, $\sigma_v^2/R \propto \Sigma$ (Heyer et al. 2009), in the lower panels, where the solid black line shows $\alpha_{\rm vir} = 1$ as defined in Sect. 3.2, and the dashed lines correspond to a $\alpha_{\rm vir} = 1$ when including the contribution of external pressure ($P_{\rm ext}$ = 1, 10 and 100 M$_\odot$ pc^{-3} km^2 s^{-2}). On the left panels, we show our SEDIGISM science sample in grey scale, and the subset of clouds with an ATLASGAL counterpart in blue, and those with a signpost of HMSF in red. From these, we can see that although our SEDIGISM clouds do show some correlation on both plots, neither of these follow the scaling relations proposed by previous works.

The right-hand side panels show a compilation of literature catalogues of molecular clouds in green colour scale, including both Galactic studies (Oka et al. 2001; Heyer et al. 2009; Roman-Duval et al. 2010; Rice et al. 2016; Barnes et al. 2016; Miville-Deschênes et al. 2017; Colombo et al. 2019; Rigby et al. 2019) and extragalactic studies (Rosolowsky & Blitz 2005; Bolatto et al. 2008; Wong et al. 2011; Gratier et al. 2012; Wei et al. 2012; Donovan Meyer et al. 2013; Colombo et al. 2014; Leroy et al. 2015; Utomo et al. 2015; Faesi et al. 2016; Tosaki et al. 2017; Freeman et al. 2017; Pan & Kuno 2017; Schruba et al. 2017). On those, we overplot the loci of the distribution of our science sample as the black ellipse, produced from a principal component analysis[15] (PCA, Pearson 1901), similar

[15] The PCA analysis (Pearson 1901) can be useful to identify the directions of maximal and minimal variance of data with large intrinsic scatter, thus equivalent to finding the direction and scatter of the underlying scaling relation (which are typically estimated using a linear regression fit). As we are simply interested in using the PCA as a representation of the loci of the

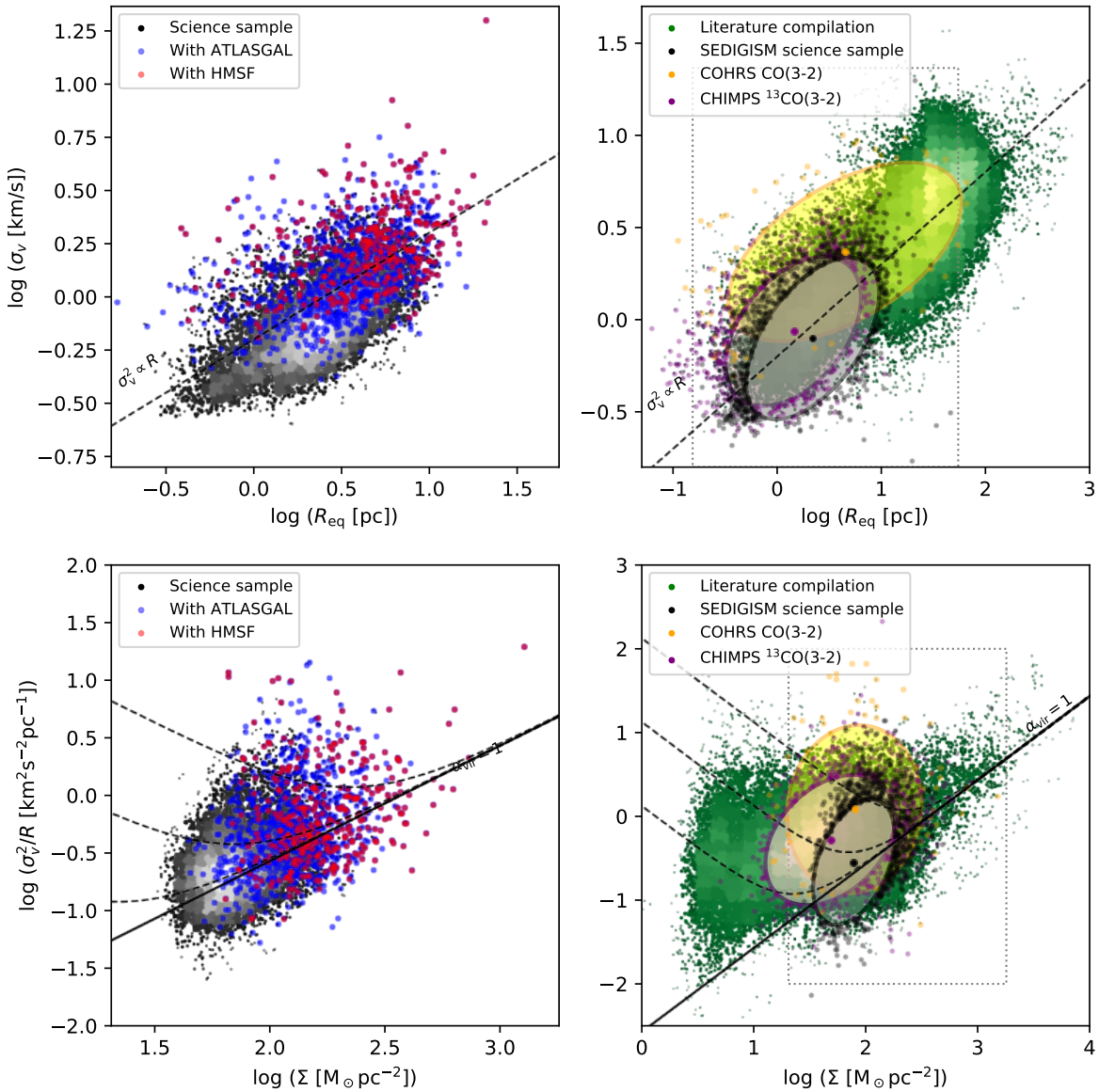

Figure 9. Top row: size-linewidth relation (σ_v versus R_{eq}), where the dashed-line represents the Larson relation. Bottom row: scaling relation between σ_v^2/R and gas surface density Σ, where the lines correspond to $\alpha_{vir} = 1$: the solid line is without external pressure, and the dashed lines are when including external pressure (from top down, at a constant P_{ext} = 100, 10 and 1 M$_\odot$ pc^{-3} km^2 s^{-2}). The left panels show these relations for the SEDIGISM sample alone, where the grey scale represents the density of points for the entire science sample, the blue circles show the clouds with an ATLASGAL counterpart, and the red circles show the clouds that have a HMSF signpost. The panels on the right show, in green, the density of points from a compilation of literature catalogues which include both Galactic and extragalactic studies (see text for full list of references). Our SEDIGISM sample is represented by the black ellipse (from a PCA analysis, and where the ellipse contour contains 95% of the data) and black points (which show the remaining 5% of clouds). Similarly, we also show the PCA ellipses for the fiducial sample of the COHRS survey in orange (Colombo et al. 2019), and the CHIMPS survey in purple (Rigby et al. 2019), both of which are high-resolution surveys towards the 1st Galactic quadrant - complementary to SEDIGISM. For reference, the dashed grey boxes on the right panels show the plotting range of the corresponding left panel.

to Colombo et al. (2019). The ellipse contours in the right panels of Fig. 9 correspond to a 2-sigma level, i.e. it contains ∼ 95% of the data points, while the central point corresponds to the mean. The remaining 5% of data points are overplotted as circles.

We have also performed this PCA analysis for the cloud catalogues from the fiducial sample of the COHRS survey (in ^{12}CO (3-2), Colombo et al. 2019), and from the CHIMPS survey (in ^{13}CO (3-2), Rigby et al. 2019), which we plot in Fig. 9 as yellow and purple ellipses, respectively. Although both of these surveys have a slightly higher spatial resolution than SEDIGISM (17″ versus 28″), they both cover the 1st quadrant, making them highly complementary to the SEDIGISM survey. In fact, the native resolution of CHIMPS was smoothed to 27″ for their source extraction and derivation of cloud properties that we use here, thus making it very similar to

distributions, we did not take into account the uncertainties in the measured quantities for this analysis.

Table 4. Slopes (α and b) recovered from a PCA analysis on the scaling relations, where $\sigma_v \propto R^{\alpha}$, and $(\sigma_v^2/R) \propto \Sigma^b$. The mean values of each pair or quantities (i.e. the centres of the ellipses in Fig. 9), are noted with the upper-script m.

Sample	α	$[\sigma_v^m, R^m]$	b	$[(\sigma_v^2/R)^m, \Sigma^m]$
SEDIGISM	0.52	[2.19, 0.79]	3.91	[78.3, 0.29]
CHIMPS	0.39	[1.43, 0.86]	2.15	[48.3, 0.52]
COHRS	0.27.	[4.48, 2.33]	14.79	[79.1, 1.22]
Expected	0.5[a]		1.0[b]	

[a] Larson (1981)
[b] Heyer et al. (2009)

that of the SEDIGISM survey. For completeness, we summarise the directions of major variance from the PCA analysis for these three surveys in Table 4, which can be compared to the expected slopes from the literature. Note, however, that even though the slopes from the PCA analysis can be suggestive of a correlation, in all the cases we performed the PCA here, the major and minor axis are similar (within a maximum of a factor 3 difference), which indicates that these are not tight correlations.

The clouds from the COHRS survey were extracted using the same method as us (SCIMES) but, because it uses ^{12}CO (3-2), it typically traces larger clouds, with larger velocity dispersions (partly due to the fact that ^{12}CO traces more diffuse gas than ^{13}CO, but also due to line broadening from optical depth effects, and from a coarser spectral resolution of 1 km s^{-1})[16]. The CHIMPS survey coverage overlaps with COHRS, but it uses the optically thinner ^{13}CO (3-2). Even though the clouds from CHIMPS were extracted using Fellwalker (Berry 2015), which segments the emission into their individual peaks (hence not allowing for the grouping of several peaks into complexes), and their line tracer is not the same as ours (using a higher-energy transition of ^{13}CO), the properties of the CHIMPS clouds agree remarkably well with those of our SEDIGISM sample. There is only a small shift in the sizes of the SEDIGISM clouds towards larger values (as we can see in the top-right panel of Fig. 9) and, although the CHIMPS sample spans to lower average surface densities than the SEDIGISM sample (as we can see on the lower-right panel), both samples have clouds reaching similar values towards the high surface-density end. These differences can be easily understood as a consequence of: 1) the cloud segmentation used by the CHIMPS survey, breaks up the emission more, thus extracting smaller (and less dense) clouds, whilst the grouping of individual clumps into larger cloud complexes achieved by our usage of SCIMES for the SEDIGISM segmentation will tend to incorporate such small diffuse clumps into larger complexes; and 2) the ^{13}CO (3-2) transition used in CHIMPS has a higher critical density and will typically trace warmer gas than the brighter ^{13}CO (2-1) transition of SEDIGISM, which will mean that the CHIMPS clouds will typically be able to trace less mass for a given brightness temperature.

Most interestingly, these plots show that the choice of tracer and the specific limitations of the surveys change our global view of the properties of molecular clouds. Looking at the ^{12}CO emission from the COHRS survey, we could argue that these clouds are in a pressure-confined regime (i.e. lying above the $\alpha_{\rm vir} = 1$ line when external pressure is not included, but could be consistent with being virialised if a moderate external pressure is at play). However, looking at the same clouds with an optically thinner tracer (i.e. with CHIMPS) changes our perception of their energy balance, with clouds moving closer to a more gravitationally bound regime, or else requiring only a very weak external pressure to be virialised. This points out a rather important issue: although molecular clouds are highly hierarchical, they are part of a continuous medium that smoothly blends into the diffuse warm neutral medium, with no hard boundary. We know that the ISM is not composed of a discrete set of entities, and yet this discretisation is (and has been) a crucial step in our understanding of the cold molecular medium. What we use to define them thus changes what we actually trace. Simple measures of the energy balance of clouds at any one single level are incapable of providing a complete picture of the true physics that describe and regulate the evolution of clouds. Instead, we need to move into trying to put a sequence together for the general trend of the change in molecular cloud properties with tracer density (which could even perhaps be used as a proxy for time). Studies looking into the evolution of these global properties, within molecular clouds - i.e. as we move inside the internal hierarchy of clouds - are necessary for taking our understanding of the physics inside molecular clouds to the next level. This is one of the key advantages of using a dendrogram-based segmentation of the ISM, that we shall explore in future work.

6 GALACTIC DISTRIBUTION OF THE MOST EXTREME CLOUDS

Using the longitude (ℓ) and distance (d) of the clouds in our catalogue, we can estimate their Galactocentric coordinates, which we use to plot our clouds on a "top-down" view of the Galaxy. These are shown in Fig. 10, overlaid on an artist's impression of the Milky Way (by NASA/JPL-Caltech/R. Hurt (SSC/Caltech)). The main known gaseous spiral arms are labeled in the bottom-left panel. The top-left panel of Fig. 10 shows our full SEDIGISM catalogue with distances, the top-right panel shows the distribution of our science sample, and the bottom-right panel shows the science sample colour-coded depending on whether the clouds have an ATLASGAL counterpart (blue), or a HMSF signpost (red). Using this top-down Galactic distribution of clouds in the science sample, we estimate a typical mass surface density of gas associated with clouds to be of the order of 1×10^5 M$_\odot$kpc^{-2} (and ranging from $\sim 4.4 \times 10^2$ to 1.3×10^6 M$_\odot$kpc^{-2}). Note that the values for the average and minimum mass surface densities are only lower limits, as they are likely affected by our completeness limits. On the bottom-left panel of Fig. 10 we show our complete science sample, i.e. clouds within the science sample that lie above our mass and radius completeness limit (as detailed in App. C), and are located within a heliocentric distance of 14.5 kpc (the distance used to determine our completeness limit). The complete science sample also excludes clouds with a tangent distance. For those clouds, although the physical properties are reliable (since the near and far distances are relatively close together), their Galactic position falls into a single line at the tangent distance, which introduces some biases for the statistical tests we will be performing with this sample (see App. F for more details). Our complete science sample consists of 1680 clouds.

We caution that showing clouds with this top-down perspective, although suggestive, can be misleading - indeed we know that

[16] A comprehensive comparison of the COHRS cloud population with other surveys can be found in Colombo et al. (2019), namely their Fig. 13, which can be used to compare with the relative position of the SEDIGISM cloud catalogue.

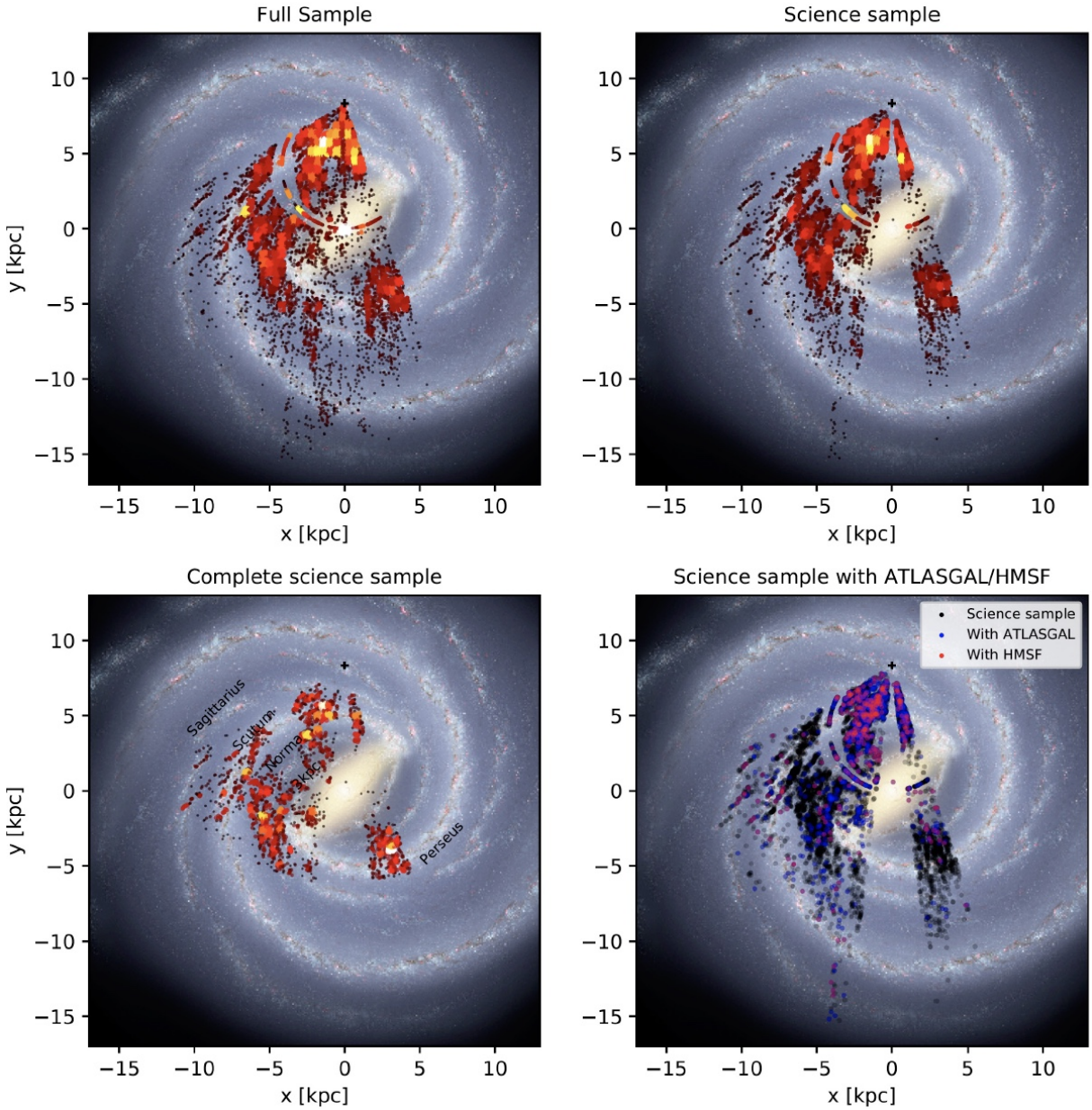

Figure 10. Top down view of the Galaxy, with the deprojected position of SEDIGISM clouds overplotted on an artistic impression of the Milky Way (NASA/JPL-Caltech/R. Hurt (SSC/Caltech)). The position of the Sun is marked with a '+' in all panels. The top-left panel shows the density plot of the entire catalogue, and the top-right panel shows the science sample. The bottom-left panel shows the Galactic distribution of the clouds in the complete science sample (i.e. above our completeness limit, and excluding clouds with a tangent distance assignment). For these three panels, the colour scale and the size of the symbols is related to the local density of clouds (more crowded areas are shown in white, and with larger symbols). The bottom-right panel shows all the sources in the science sample colour-coded depending on whether they have an ATLASGAL counterpart (in blue), a HMSF signpost (in red), or neither (in black).

the uncertainties on the distances can amount to $\sim$ 1 kpc, particularly when streaming motions around spiral arms can be important, and this can easily displace clouds across entire spiral arms. In addition, the exact position and strength of these arms is still quite uncertain (e.g. Taylor & Cordes 1993; Reid et al. 2014; Vallée 2017). In fact, the very existence of four strong spiral arms is still subject of debate, especially as studies in the Optical/near-IR (e.g. Drimmel 2000; Siebert et al. 2011, 2012; Gaia Collaboration et al. 2018), suggest that we only have two main stellar spiral arms - which could indicate that the four spiral arms that we see in the gas, are not as well defined as this figure depicts, and are perhaps more flocculent in nature. This idea is also supported by our relatively low values of molecular gas mass surface densities, which place the Milky Way at the bottom of the distribution of the values retrieved for a sample of 15 nearby spiral galaxies Sun et al. (2018), whose typical molecular gas mass surface densities are of the order of $10^6 - 10^8$ M$_\odot$kpc^{-2}. Hence these top-down perspective plots are used here merely as a first look at the Galactic distribution of clouds. A more detailed study of arm/inter-arm dependency requires using a model of the spiral pattern, and is most accurately done in the ℓbv space, which is beyond the scope of this paper.

In order to look for effects that could depend on the Galactic

environment, without the need to assume any specific spiral arm model, we have examined the spatial distribution of clouds with extreme properties (i.e. clouds that form the tails of a distribution), and compared those to the global Galactic distribution of clouds. The idea behind this exercise is a purely statistical one, which will test whether the most extreme clouds follow the same spatial distribution as the global population of clouds, or whether they show significant deviations from it. As an attempt to take this analysis a step further, we can make the loose assumption that the spiral arms should preferentially be represented by the crowded regions of the global population, while the inter-arm regions would be preferentially associated with the least crowded places. This assumption is purely qualitative (due to the uncertainties in the distances), and we make no attempt to effectively associate clouds with spiral arms or inter-arm regions. For our purpose, we use the complete science sample as our global cloud population (bottom-left panel of Fig. 10), from which we selected a number of sub-samples that comprise the most extreme clouds. This selection was made by taking the most extreme 100 clouds of each distribution (corresponding to the top or bottom 6%), and the specific selection criterion is indicated at the top of each panel in Figs. 11 and 12.

The comparison between the sub-samples and the global cloud population was done by performing the Pearson's χ^2 statistical test, which tests whether the frequency distribution of certain events observed in a sample is consistent with a particular theoretical distribution. The full details of the χ^2 statistical test that we performed are explained in App. F. In brief, for our purpose, we used the 2D Galactic distribution of clouds in the complete science sample as our theoretical distribution. In practice, we built a normalised 2D histogram with the spatial distribution of clouds in the complete science sample (using their Galactocentric coordinates), using a spatial bin of 0.3×0.3 kpc – this map represents the probability of an observation falling in a specific spatial bin (see left panels of Figs. F1 to F3). We then compute the χ^2 statistics using the observed 2D distribution of each sub-sample (shown in the central panels of Figs. F1 to F3), and the observed χ^2-values are compared to the values obtained from a pure random draw of clouds from the theoretical distribution (i.e. effectively obtaining a p-value, which we call $p_{\rm rnd}$, see Fig. F4). Given the statistical fluctuations, as well as the uncertainties in the distributions, and binning effects (neither of which are taken into account for this exercise), the exact χ^2 values and $p_{\rm rnd}$ that we derive should not be taken at face value. Instead, they are more useful for a relative comparison of the sub-samples, as an indication for which sub-samples are most different to the global cloud population. The results from our χ^2 statistical test are summarised in Table F1. We describe all of the studied tails of distributions in the following Sects. 6.1 to 6.5.

6.1 The most massive molecular cloud complexes

Some observations of nearby spiral galaxies (e.g. Koda et al. 2009, Colombo et al. 2014), as well as some galaxy-scale numerical models (e.g. Dobbs et al. 2008, Fujimoto et al. 2014, Duarte-Cabral & Dobbs 2016, Pettitt et al. 2018) - both of which benefit from a more straightforward association of clouds to spiral arms - have suggested that the most massive clouds are preferentially located along spiral arms. This is widely accepted and understood in the context of spiral arms being able to concentrate more material, and thus able to form larger and more massive giant molecular clouds (GMCs). Similarly to the argument for encountering the most massive clouds in the arms, with a higher concentration of material in the spiral arms we ought to expect the highest surface density clouds to lie in the spiral arms as well. In this spirit, we have plotted the Galactic distribution of the 100 most massive clouds in our sample, and the 100 clouds with the highest surface density, in the top-left and top-right panels of Fig. 11, respectively.

Our χ^2 tests comparing these two distributions to the global cloud population, give us χ^2 values of 670 and 638 (which corresponds to a $p_{\rm rnd}$ of 0.05 and 0.16), for the extreme mass and surface density clouds respectively. This suggests that the distribution of high-surface density clouds still follows the original distribution of clouds, implying that such clouds might be found in crowded areas (or spiral arms), simply from statistics. The distribution of the most massive clouds, however, is less consistent with a pure random draw of clouds from the parent distribution. If the disparities between the two distributions were caused by having more high-mass clouds in the spiral arms than what is statistically expected, then we should see an excess of high-mass clouds in the most crowded areas of the global distribution. However, considering the spatial distribution of these clouds on Fig. 11 (top panels), and the relative difference between the predicted and measured counts shown in Fig. F1 (top and middle rows), it is not obvious that this is the case, with clouds having both an excess and lack of counts in different crowded areas. The specific regions where the most high-mass clouds are found to be in excess or lacking, are not particularly striking in terms of their environment, leaving our interpretation inconclusive.

6.2 The most elongated clouds

A subject of increasing interest in the SF community is the origin and properties of the most elongated clouds. While some numerical and observational studies suggest that extremely long filamentary clouds would be formed as the result of the Galactic shear in the inter-arm regions (e.g. Kim & Ostriker 2002; Shetty & Ostriker 2006; Ragan et al. 2014; Duarte-Cabral & Dobbs 2016, 2017), other studies suggest that at least some of these might trace the "spines" of the spiral arms (e.g. Goodman et al. 2014; Wang et al. 2015; Zucker et al. 2015).

We have thus looked at the Galactic distribution of the 100 longest clouds in the SEDIGISM sample, as well as the 100 clouds with the largest aspect ratio. These are shown in the bottom panels of Fig. 11 (left and right respectively). Our χ^2 tests for these two distributions, give us a χ^2 value of 715 and 671 (corresponding to a $p_{\rm rnd}$ of 0.005 and 0.04) for the extreme length and aspect ratio clouds respectively. This suggests that the Galactic distribution of both these sub-samples are different from the global cloud population (although this is most evident for the sample of largest clouds). However, neither of them seem to show any clear preference for crowded or non-crowded areas (see also Fig. F1 bottom panel and Fig. F2 top panel).

This analysis has a few caveats, though. The first one is that there are more of these elongated clouds located at the near distance, than there are at the far distance. This could be linked to resolution limitations which will result in more distant filaments appearing less elongated. The second caveat is the fact that both of these quantities are purely the projected ones (the length and aspect ratio on the plane of the sky). If long filamentary clouds are indeed shaped by the shear from the Galactic differential rotation, we do not expect them to be randomly orientated. Therefore, this projection is likely to affect our ability to select the truly elongated structures, in specific parts of the Galaxy, being particularly critical in lines of sight where we expect the clouds' elongations to be roughly along our line of sight. The third caveat is the fact that even the longest molecular filaments in our Galaxy (such as the ~100 pc long Nessie filament,

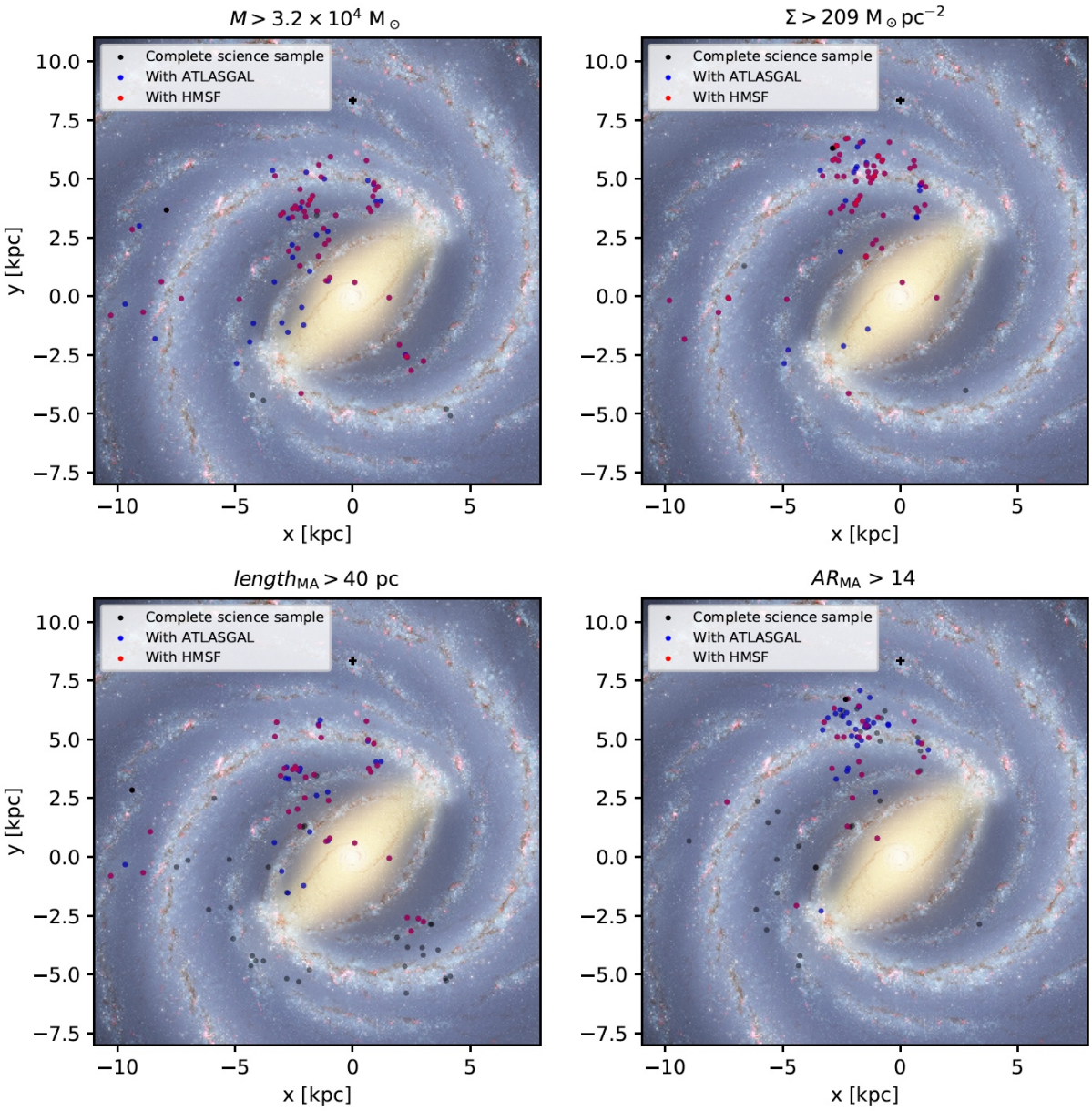

Figure 11. Top down view of the Galaxy as in Fig. 10, showing the SEDIGISM clouds of the complete science sample that are part of the top 10% of clouds in terms of Mass (top-left), surface density (top-right), medial axis length (bottom-left), and aspect ratio from the medial axis (bottom-right). The specific condition that corresponds to this cut-off is indicated at the top of each panel. Clouds are colour-coded depending on whether they have an ATLASGAL counterpart (in blue), a HMSF signpost (in red), or neither (in black).

Jackson et al. 2010), do not appear in our segmentation as a single entity - they are instead composed of several (smaller) filamentary sections. Finally, the relative lack of large (and massive) clouds nearby (with $d < 2.5$ kpc), can also point at a possible bias from the cloud segmentation, in which we might still be more likely to break the most nearby clouds into smaller sub-structures (even though we use a clustering algorithm designed to minimise this effect). All of these effects could result in the underestimation of both the length and aspect ratio of clouds, and thus the 100 most extreme clouds we take for this analysis might not correspond to the most extreme cases in physical space.

6.3 The most dynamically active clouds

Given the complex global Galactic dynamics, we would expect to see, at least at first order, some link between the most dynamic places in the Galaxy, with the kinetic properties of clouds. In this sense, we have isolated the 100 clouds with the highest virial parameter, and highest velocity dispersion. Their Galactic distribution is shown in Fig. 12, top panels. Our χ^2 test for clouds with a large virial parameter gives us a χ^2 value of 699 (corresponding to a $p_{\rm rnd}$ of 0.01), indicating that they differ from a random statistical subset of the global cloud population, in terms of their Galactic placement (see also Fig. F2 middle panel). On the other hand, clouds with a large velocity dispersion have a higher χ^2 value of 747 (which corresponds to a much smaller $p_{\rm rnd}$ of 0.001), making this distribution

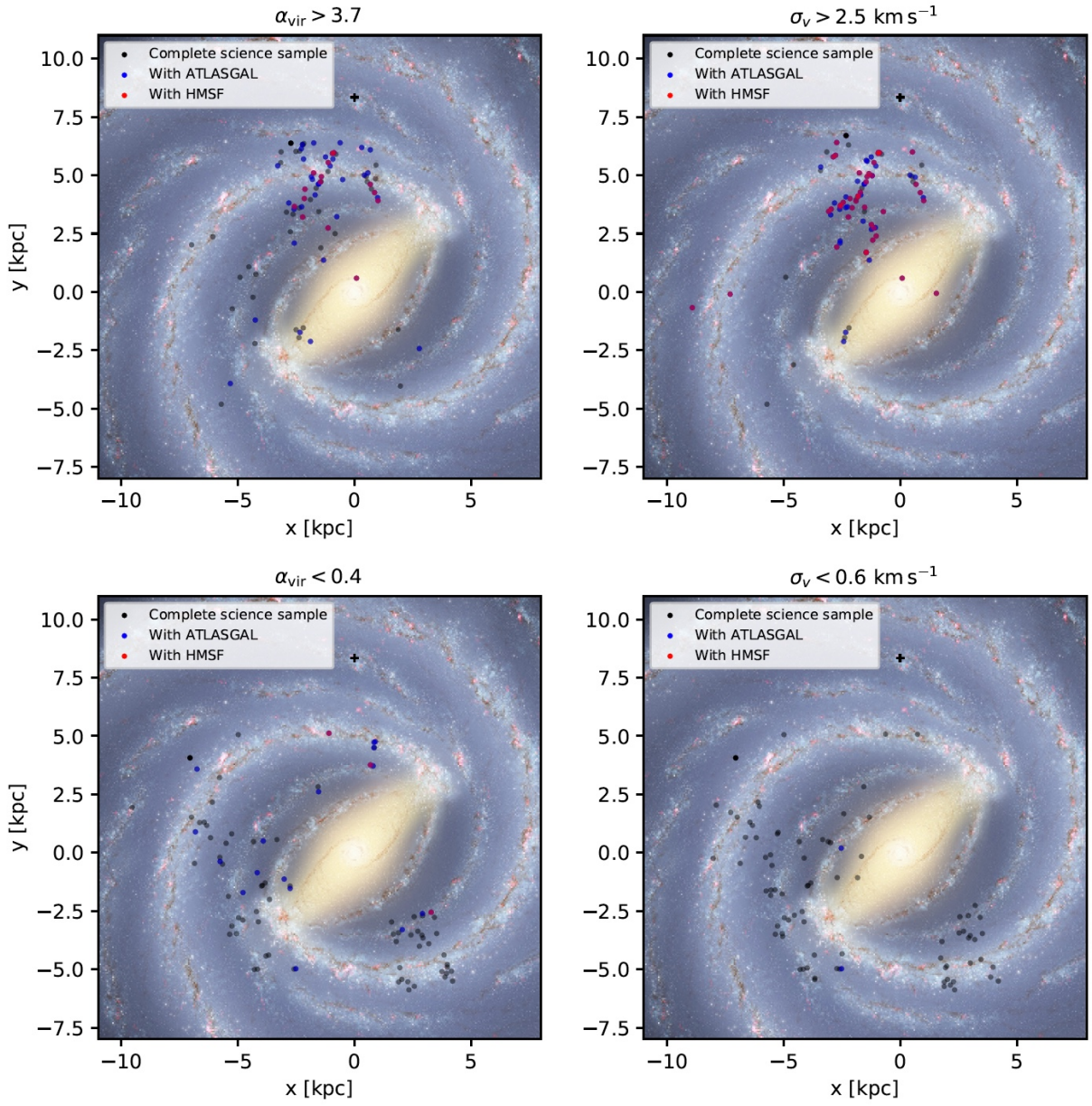

Figure 12. Same as Fig. 11, showing the SEDIGISM clouds of the complete science sample that are part of the top 10% of clouds with a high virial parameter (top-left), and high velocity dispersion (top-right). The lower panels show the bottom 10% of clouds in the same properties: with a low virial parameter (bottom-left) and a low velocity dispersion (bottom-right). The specific condition that corresponds to this cut-off is indicated at the top of each panel. Clouds are colour-coded depending on whether they have an ATLASGAL counterpart (in blue), a HMSF signpost (in red) or neither (black).

less like the global cloud population. Most of the differences in the statistics of this sub-sample comes from an excess of high-velocity dispersion clouds relatively nearby (see top-right panel of Fig. 12, and bottom panel of Fig. F2), which then also propagates (although less severely) into clouds with high-virial parameters also being mostly nearby. We believe that these trends could be partly due to observational biases (see Fig. C2, and the discussion in App. C).

Interestingly, these dynamically active clouds typically make up two types of populations. The first is most closely associated with crowded regions (potentially associated with the near Sagittarius, Scutum and Norma spiral arms), which is where we expect more frequent cloud-cloud interactions, in line with the results from numerical simulations of spiral galaxies (e.g. Duarte-Cabral & Dobbs 2017; Pettitt et al. 2018). This population of clouds is also actively forming high-mass stars. The larger values of velocity dispersion and virial parameters could thus be also an indication of larger internal motions of clouds, perhaps partly driven by their active gravitational contraction, or by internal feedback from the forming stars, or both.

The second population of clouds are devoid of HMSF signposts, and some even lacking an ATLASGAL counterpart (i.e. less dense). Most of these are also at large distances, which could suffer from a completeness effect in the ATLASGAL and HMSF tracers. Alternatively, this second population could represent clouds relatively close to the Galactic bar, and/or in the streams of gas feeding the Galactic centre region – all regions prone to experiencing a significant shear driven by the global Galactic dynamics. This dichotomy (of clouds in the two extremes of their SF history sharing

the same integrated dynamical properties) highlights the caveats of performing a standard virial analysis and deriving any conclusions therefrom alone.

6.4 The most dynamically quiescent clouds

On the opposite extreme of the dynamical status of molecular clouds, we have also explored the location of the clouds that are relatively quiet (which we refer to as the most "dynamically quiescent" clouds), which include clouds with a low virial parameter, or a low velocity dispersion. These types of clouds are often not subject of much attention (mostly as they typically lie close to survey limitations in terms of spectral resolution). Nevertheless, some recent numerical work by Pettitt et al. (2018) has suggested that, in grand-design spiral galaxies, while clouds with high virial parameter are most often associated with spiral arms, clouds with low virial parameters have a weaker correspondence with the spiral arms, with many inter-arm clouds being remnants of large arm complexes or simply formed in-situ from small over-densities in filaments and arm spurs.

To investigate these dynamically quiescent clouds in SEDIGISM, we have selected the 100 clouds in the complete science sample with the lowest virial parameter, and the lowest velocity dispersion. Their Galactic distribution is shown in the bottom panels of Fig. 12. Our χ^2 tests give us χ^2 values of 675 and 669 (corresponding to a $p_{\rm rnd}$ of 0.04 and 0.05) for the low virial parameter and low velocity dispersion respectively. This suggests that the Galactic distribution of the most dynamically quiescent clouds is only mildly different to that of the global cloud population. Their distribution in Fig. 12 (see also Fig. F3 top and middle row) suggests that they are not found in very crowded areas (possibly favouring inter-arm locations).

Clouds with a low virial parameter are often interpreted to be gravitationally bound (i.e. where gravity dominates over turbulence). However, these clouds are not necessarily collapsing - indeed if they were, the collapse itself would increase the virial parameter again (e.g. Kauffmann et al. 2013). Our results show that these dynamically quiescent clouds are mostly devoid of HMSF or even high-column densities (which would result in an ATLASGAL counterpart), perhaps indicating that their evolution is not regulated by their own gravity but by interaction with the Galactic potential, the large scale shear motions and perhaps also by large scale magnetic fields.

We caution however, that even though a handful of dynamically quiescent clouds are relatively nearby, most of them are at $d > 8.0$ kpc. In terms of absolute numbers, the science sample does contain nearby low-velocity-dispersion clouds, but most of those are below the size and/or mass threshold used to build the complete science sample. The usage of a completeness limit for the whole SEDIGISM sample (and especially one largely above the resolution element) was an attempt to remove any bias from the resolution and distance. However, our intrinsic observational limitations may still be responsible for at least part of this signature, as we can see that the average measured velocity dispersion of the complete science sample has a correlation with distance (see Fig. C2, and the respective discussion in App. C). Furthermore, at the far distances our sample may also not be complete in terms of the detection of an ATLASGAL counterpart or HMSF signposts, potentially biasing the interpretation above.

Nevertheless, these type of clouds could potentially be interesting to follow up with the goal to investigate whether this tentative trend does hold up, with a more in-depth analysis, considering the survey limitations and a detailed modelling of the spiral pattern of the Galaxy.

6.5 The high-mass star-forming clouds

One of the questions we wanted to address here is whether the Galactic distribution of clouds that host ongoing high-mass star formation is uniform, or whether they are preferably located in spiral arms as our preliminary study of the SEDIGISM science verification field suggested (Schuller et al. 2017). In particular, if high-mass star-forming clouds are tracing the arms, we are also interested in exploring whether that is purely due to a statistical sampling (as suggested by e.g. Elmegreen & Elmegreen 1986; Moore et al. 2012; Eden et al. 2013); or whether there is an excess of high-mass star-forming regions in the crowded spiral arms, suggestive of SF triggering from the passage of a spiral wave (e.g. Lin & Shu 1964, Roberts 1969, Toomre 1977, Martínez-García et al. 2009).

Figure 10 (bottom-right panel), shows the distribution of all clouds with a HMSF signpost in our science sample (in red). The χ^2 statistical test, performed using only the clouds in the complete science sample (from which only 211 clouds have a HMSF signpost) gives a χ^2 value of 735, which translates into a $p_{\rm rnd}$ of 0.001. This indicates that the distribution of clouds with a HMSF signpost does not mimic the global distribution of clouds. Upon closer inspection of Fig. 10 and F3, it becomes clear, however, that most of the deviations from the global distribution of clouds do not arise from crowded or non-crowded areas, but rather shows a distance effect. Indeed, most of the clouds with signs of on-going high-mass star formation are located relatively close to us. The extremely high density of points there (compared to elsewhere in the Galaxy), is likely to be a simple consequence of completeness in the HMSF signposts (namely HII regions and massive YSOs).

Interestingly, if we look at the higher-mass clouds or the higher-surface density clouds (Fig. 11 top panels), not all of these host high-mass star formation. This is true even if we just consider the most nearby clouds, where we should be less affected by completeness issues in terms of HMSF signposts. As we have seen in Sect. 5, there does not seem to be a unique global property of a molecular cloud that defines the ability of a cloud to form high-mass stars - and the same applies for the Galactic environment. Perhaps to isolate clouds with a potential to form massive stars, we need to use a combination of conditions that need to be satisfied, or even just the most extreme conditions within a cloud (rather than the integrated properties). Applying a single global threshold law (such as a gas surface density threshold or mass-radius threshold, e.g. Krumholz & McKee 2008; Kauffmann & Pillai 2010; Baldeschi et al. 2017) to define the potential to form massive stars, is probably not a single unique descriptor. Figure 13 highlights this issue, where we can see clouds with and without high mass star formation that have the same mass and radius. In this figure, we also show as a dashed line, the empirical relation for high-mass star formation inferred by Kauffmann & Pillai (2010), and confirmed by other works (e.g., Kauffmann et al. 2010a,b; Urquhart et al. 2018). Note that the plotted line is the Kauffmann & Pillai (2010) original threshold scaled up so as to be consistent with our adopted opacity law (see App. G for more details).

Although this empirical relation was determined for clumps, rather than for clouds (as we use here), the bulk of the parameter space that we probe is similar to that in Kauffmann & Pillai (2010): their sizes range from <0.1 pc to 10 pc (compared to our range of 0.3 pc to ~30 pc), and their masses range from 1 $M_\odot$ to > 10^4 $M_\odot$ (compared to our range of 10 $M_\odot$ to > 10^5 $M_\odot$).

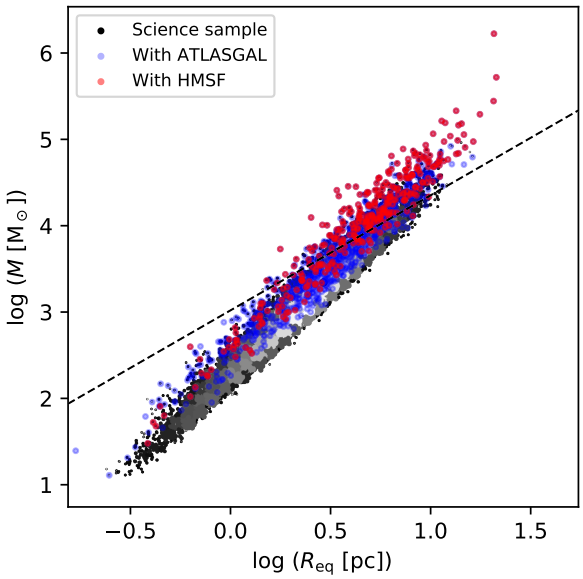

Figure 13. Mass-radius relation for the SEDIGISM clouds, where the grey scale represents the density of points for the entire science sample, the blue circles show the clouds with an ATLASGAL counterpart, and the red circles show the clouds that have a HMSF signpost. The dashed line shows the empirical relation from Kauffmann & Pillai (2010), where clouds above this line are expected to form high-mass stars. The plotted threshold is at $M\,[\mathrm{M}\odot] = 1053\ (R\,[\mathrm{pc}])^{1.33}$, which is scaled up from the original HMSF threshold from Kauffmann & Pillai (2010), to account for the different opacity law used (see App. G).

If we use that relation directly with our clouds, we would miss some true positives (107 out of 330 clouds with a HMSF signpost lie below the empirical threshold, i.e. missing ∼33% of all clouds that we know are actively forming massive stars), as well as potentially provide a significant number of false positives (455 out of a total of 678 clouds above the HMSF threshold do not have a detected HMSF signpost, i.e. ∼ 70% of clouds above the empirical line for HMSF). Since completeness limits could play a role in the non-detection of the signposts for HMSF, we estimate that the number of false negatives (i.e. missed true positives) is a lower limit, while the number of false positives is an upper limit.

The detection of potential false positives was not ruled out by Kauffmann & Pillai (2010): indeed they note that their threshold appears to capture a necessary condition for HMSF, but not a sufficient one. Alternatively, it could also be that part of the clouds above the HMSF threshold line but for which we have no detected HMSF (i.e. the false positives), are in fact clouds that simply have not done so yet, because of the potential large latency periods prior to star formation. In that sense, the trends in properties going from the science sample to clouds with an ATLASGAL counterpart and then clouds with a HMSF signpost (from Sect. 5) could be an indication of the cloud evolution towards HMSF during this latency period (with clouds progressively building up their mass, becoming larger, denser, and more dynamically active - with larger velocity dispersions), even if this remains a stochastic process for each individual cloud (e.g. Barnes et al. 2018).

More intriguing, however, are the missed true positives. These clouds lie below the empirical line supposedly representing the threshold below which HMSF would not occur, and yet they have tracers of ongoing HMSF. Nevertheless, it is possible that the material probed by Kauffmann & Pillai (2010) is intrinsically tracing higher density material than what we do, which could shift the exact position of the cloud sample with respect to the empirical line for HMSF, thus potentially making this relation inappropriate for usage with our sample. An indication that this might indeed be the case, is the fact that the subsample of SEDIGISM clouds with a HMSF signpost that we present here, is purely a subsample of the ATLASGAL clumps, which seem to confirm the Kauffmann & Pillai (2010) relation on clump scales (e.g. Urquhart et al. 2018). This highlights a potential caveat of using such relations blindly, as perhaps they are not applicable on cloud scales, when the density profiles become shallower, and the more diffuse material contributes to increasing the sizes of the clouds, whilst providing only moderate increase to the enclosed mass. A hierarchical study of this transition within clouds would be required to understand where this relation might break.

7 SUMMARY AND CONCLUSIONS

The SEDIGISM survey has covered ∼ 84 square degrees of the inner Galaxy with ^{13}CO (2-1). From the contiguous portion of the survey (i.e. excluding the W43 field), we extracted the entire molecular cloud population with a large dynamic range in spatial scales, using the Spectral Clustering for Interstellar Molecular Emission Segmentation (SCIMES) algorithm. We determined the distances to the clouds, using the kinematic distances, and a number of methods to solve the distance ambiguities (including masers, IRDC, Dark Clouds, HiSA, distance to the Larson's size-linewidth relation, distance to the Galactic plane, and extinction distances). The full catalogue that we release contains 10663 molecular clouds, 10300 of which with measurements of physical properties.

In this paper, we have explored some of the global properties of clouds using a sub-sample of the full catalogue (i.e. our "science sample"), consisting of 6664 well resolved sources and for which the distance estimates are reliable. In particular, we compare the scaling relations retrieved from SEDIGISM to those of other surveys, including Galactic and extragalactic work. We find that the locus of the SEDIGISM clouds is similar to that of other surveys, but that the specific scaling relations vary widely between surveys - even between those that cover the same area in the Galaxy, just with different tracers. The intrinsic scatter in these relations is very large, making all the correlations rather unconstrained.

We also explored the properties of clouds with and without tracers of high-mass star formation, and we find that for most distributions (mass, size, surface density, velocity dispersion), the median values of the distributions is higher for clouds with a HMSF signpost, potentially indicative of an evolutionary sequence. However, the distributions become progressively flatter, with the clouds with HMSF spanning a wide range of values for all properties we looked at. These results suggest that there is no single global property of a cloud that is able to define their ability to form massive stars, and the usage of a simple threshold to isolate clouds forming high-mass stars is not complete (providing both false negatives and false positives).

Finally, we have looked into potential links between the Galactic environment of clouds and their properties, by looking at the Galactic distribution of the most extreme clouds. For that purpose, we have isolated the most extreme 100 clouds in each distribution (i.e. clouds that make up the tails of the distributions), and compared their Galactic distribution to that of the cloud population above our completeness limits (i.e. our complete science sample), using a χ^2 statistical test. This provides a means to determine whether extreme

clouds follow a Galactic distribution that differs significantly from the global cloud population. We find that, for most properties, the Galactic distribution of the most extreme molecular clouds is is only marginally different to that of the global cloud population. The Galactic distribution of the largest clouds, the most turbulent clouds and the high-mass star-forming clouds are those that deviate most significantly from the global cloud population. We also find that the least dynamically active clouds (with low velocity dispersion or low virial parameter) are situated further afield, mostly in the least populated areas, and therefore could hint at those being mostly in inter-arm regions. However, we find that part of these trends might be due to completeness limits (e.g. in case of the HMSF tracers), and intrinsic survey limitations, which result in a trend of decreasing velocity dispersion with distance, hampering our ability to make any firm conclusions from this data alone.

In future work, we shall follow up some of these tentative trends using distance-limited samples, with the incorporation of detailed models of the spiral arms, and with more complete cross-match with signposts of HMSF (e.g. by comparing with the Hi-GAL samples, and their L/M ratio as an indicator for more embedded HMSF and their respective evolutionary stage) to mitigate some of the observational biases that are potentially at play in the work presented here.

ACKNOWLEDGEMENTS

ADC acknowledges the support from the Royal Society University Research Fellowship (URF/R1/191609). ADC and AJR acknowledge the support from the UK STFC consolidated grant ST/N000706/1. DC acknowledges support by the Deutsche Forschungsgemeinschaft, DFG, through project number SFB956C. LB and RF acknowledge support from CONICYT grant Basal AFB-170002. HB acknowledges support from the European Research Council under the Horizon 2020 Framework Program via the ERC Consolidator Grant CSF-648505. HB furthermore thanks for financial help from the DFG via the SFB881 "The Milky Way System" (subproject B1). CLD acknowledges funding from the European Research Council for the FP7 ERC consolidator grant project ICYBOB, grant number 818940. S.B. and N.S. acknowledge support from the Agence National de Recherche (ANR/France) and the Deutsche Forschungsgemeinschaft (DFG/Germany) through the project GENESIS (ANR-16-CE92-0035-01/DFG1591/2-1).The Starlink software (Currie et al. 2014) is currently supported by the East Asian Observatory. This publication is based on data acquired with the Atacama Pathfinder Experiment (APEX) under programmes 092.F-9315 and 193.C-0584. APEX is a collaboration among the Max-Planck-Institut fur Radioastronomie, the European Southern Observatory, and the Onsala Space Observatory.

REFERENCES

Anderson L. D., Bania T. M., 2009, ApJ, 690, 706
Anderson L. D., Armentrout W. P., Johnstone B. M., Bania T. M., Balser D. S., Wenger T. V., Cunningham V., 2015, ApJS, 221, 26
Baldeschi A., et al., 2017, MNRAS, 466, 3682
Barnes P. J., et al., 2011, ApJS, 196, 12
Barnes P. J., Muller E., Indermuehle B., O'Dougherty S. N., Lowe V., Cunningham M., Hernandez A. K., Fuller G. A., 2015, ApJ, 812, 6
Barnes P. J., Hernandez A. K., O'Dougherty S. N., Schap William J. I., Muller E., 2016, ApJ, 831, 67
Barnes P. J., Hernandez A. K., Muller E., Pitts R. L., 2018, ApJ, 866, 19
Battersby C., et al., 2011, A&A, 535, A128
Benjamin R. A., et al., 2003, PASP, 115, 953
Berry D. S., 2015, Astronomy and Computing, 10, 22
Bertoldi F., McKee C. F., 1992, ApJ, 395, 140
Beuther H., Kainulainen J., Henning T., Plume R., Heitsch F., 2011, A&A, 533, A17
Bobylev V. V., Bajkova A. T., 2013, Astronomy Letters, 39, 809
Bolatto A. D., Leroy A. K., Rosolowsky E., Walter F., Blitz L., 2008, ApJ, 686, 948
Brand J., Blitz L., 1993, A&A, 275, 67
Burton M. G., et al., 2013, Publ. Astron. Soc. Australia, 30, e044
Busfield A. L., Purcell C. R., Hoare M. G., Lumsden S. L., Moore T. J. T., Oudmaijer R. D., 2006, MNRAS, 366, 1096
Caswell J. L., et al., 2010, MNRAS, 404, 1029
Chen X., Wang S., Deng L., de Grijs R., Liu C., Tian H., 2019, Nature Astronomy, 3, 320
Chira R.-A., Beuther H., Linz H., Schuller F., Walmsley C. M., Menten K. M., Bronfman L., 2013, A&A, 552, A40
Colombo D., et al., 2014, ApJ, 784, 3
Colombo D., Rosolowsky E., Ginsburg A., Duarte-Cabral A., Hughes A., 2015, MNRAS, 454, 2067
Colombo D., et al., 2019, MNRAS, 483, 4291
Contreras Y., et al., 2013, A&A, 549, A45
Csengeri T., et al., 2016, A&A, 586, A149
Currie M. J., Berry D. S., Jenness T., Gibb A. G., Bell G. S., Draper P. W., 2014, in Manset N., Forshay P., eds, Astronomical Society of the Pacific Conference Series Vol. 485, Astronomical Data Analysis Software and Systems XXIII. p. 391
Dame T. M., Hartmann D., Thaddeus P., 2001, ApJ, 547, 792
Dempsey J. T., Thomas H. S., Currie M. J., 2013, ApJS, 209, 8
Dobbs C. L., Glover S. C. O., Clark P. C., Klessen R. S., 2008, MNRAS, 389, 1097
Donovan Meyer J., et al., 2013, ApJ, 772, 107
Drimmel R., 2000, A&A, 358, L13
Du F., Yang J., 2008, ApJ, 686, 384
Duarte-Cabral A., Dobbs C. L., 2016, MNRAS, 458, 3667
Duarte-Cabral A., Dobbs C. L., 2017, MNRAS, 470, 4261
Duarte-Cabral A., Acreman D. M., Dobbs C. L., Mottram J. C., Gibson S. J., Brunt C. M., Douglas K. A., 2015, MNRAS, 447, 2144
Eden D. J., Moore T. J. T., Morgan L. K., Thompson M. A., Urquhart J. S., 2013, MNRAS, 431, 1587
Elia D., et al., 2013, ApJ, 772, 45
Ellsworth-Bowers T. P., et al., 2013, ApJ, 770, 39
Elmegreen B. G., Elmegreen D. M., 1986, ApJ, 311, 554
Faesi C. M., Lada C. J., Forbrich J., 2016, ApJ, 821, 125
Freeman P., Rosolowsky E., Kruijssen J. M. D., Bastian N., Adamo A., 2017, MNRAS, 468, 1769
Fujimoto Y., Tasker E. J., Wakayama M., Habe A., 2014, MNRAS, 439, 936
Gaia Collaboration et al., 2018, A&A, 616, A11
Giannetti A., Wyrowski F., Leurini S., Urquhart J., Csengeri T., Menten K. M., Bronfman L., van der Tak F. F. S., 2015, A&A, 580, L7
Gibson S. J., Taylor A. R., Higgs L. A., Dewdney P. E., 2000, ApJ, 540, 851
Ginsburg A., et al., 2013, ApJS, 208, 14
Goodman A. A., et al., 2014, ApJ, 797, 53
Gratier P., et al., 2012, A&A, 542, A108
Green J. A., et al., 2012, MNRAS, 420, 3108
Güsten R., Nyman L. Å., Schilke P., Menten K., Cesarsky C., Booth R., 2006, A&A, 454, L13
Heyer M. H., Brunt C., Snell R. L., Howe J. E., Schloerb F. P., Carpenter J. M., 1998, ApJS, 115, 241
Heyer M. H., Carpenter J. M., Snell R. L., 2001, ApJ, 551, 852
Heyer M., Krawczyk C., Duval J., Jackson J. M., 2009, ApJ, 699, 1092
Hildebrand R. H., 1983, QJRAS, 24, 267
Hoare M. G., et al., 2012, PASP, 124, 939
Honma M., et al., 2012, PASJ, 64, 136
Jackson J. M., Finn S. C., Rathborne J. M., Chambers E. T., Simon R., 2008, ApJ, 680, 349

Jackson J. M., Finn S. C., Chambers E. T., Rathborne J. M., Simon R., 2010, ApJ, 719, L185
Kauffmann J., Pillai T., 2010, ApJ, 723, L7
Kauffmann J., Bertoldi F., Bourke T. L., Evans N. J. I., Lee C. W., 2008, A&A, 487, 993
Kauffmann J., Pillai T., Shetty R., Myers P. C., Goodman A. A., 2010a, ApJ, 712, 1137
Kauffmann J., Pillai T., Shetty R., Myers P. C., Goodman A. A., 2010b, ApJ, 716, 433
Kauffmann J., Pillai T., Goldsmith P. F., 2013, ApJ, 779, 185
Kim W.-T., Ostriker E. C., 2002, ApJ, 570, 132
Koda J., et al., 2009, ApJ, 700, L132
Krumholz M. R., McKee C. F., 2008, Nature, 451, 1082
Lallement R., Babusiaux C., Vergely J. L., Katz D., Arenou F., Valette B., Hottier C., Capitanio L., 2019, A&A, 625, A135
Larson R. B., 1981, MNRAS, 194, 809
Leroy A. K., et al., 2015, ApJ, 801, 25
Lin C. C., Shu F. H., 1964, ApJ, 140, 646
Liu X.-L., Wang J.-J., Xu J.-L., 2013, MNRAS, 431, 27
Lumsden S. L., Hoare M. G., Urquhart J. S., Oudmaijer R. D., Davies B., Mottram J. C., Cooper H. D. B., Moore T. J. T., 2013, ApJS, 208, 11
Marshall D. J., Robin A. C., Reylé C., Schultheis M., Picaud S., 2006, A&A, 453, 635
Martínez-García E. E., González-López lira R. A., Bruzual-A G., 2009, ApJ, 694, 512
Mattern M., et al., 2018, A&A, 619, A166
McClure-Griffiths N. M., Dickey J. M., Gaensler B. M., Green A. J., Haverkorn M., Strasser S., 2005, ApJS, 158, 178
McClure-Griffiths N. M., Dickey J. M., Gaensler B. M., Green A. J., Green J. A., Haverkorn M., 2012, ApJS, 199, 12
Miville-Deschênes M.-A., Murray N., Lee E. J., 2017, ApJ, 834, 57
Molinari S., et al., 2010, A&A, 518, L100
Moore T. J. T., Urquhart J. S., Morgan L. K., Thompson M. A., 2012, MNRAS, 426, 701
Oka T., Hasegawa T., Sato F., Tsuboi M., Miyazaki A., Sugimoto M., 2001, ApJ, 562, 348
Ossenkopf V., Henning T., 1994, A&A, 291, 943
Otrupcek R. E., Hartley M., Wang J.-S., 2000, Publ. Astron. Soc. Australia, 17, 92
Pan H.-A., Kuno N., 2017, ApJ, 839, 133
Pandian J. D., Momjian E., Goldsmith P. F., 2008, A&A, 486, 191
Pearson K., 1901, Philosophical Magazine, 2, 559
Peretto N., Fuller G. A., 2009, A&A, 505, 405
Pettitt A. R., Dobbs C. L., Acreman D. M., Price D. J., 2014, MNRAS, 444, 919
Pettitt A. R., Dobbs C. L., Acreman D. M., Bate M. R., 2015, MNRAS, 449, 3911
Pettitt A. R., Egusa F., Dobbs C. L., Tasker E. J., Fujimoto Y., Habe A., 2018, MNRAS, 480, 3356
Purcell C. R., et al., 2013, ApJS, 205, 1
Ragan S. E., Henning T., Tackenberg J., Beuther H., Johnston K. G., Kainulainen J., Linz H., 2014, A&A, 568, A73
Reid M. J., et al., 2009, ApJ, 700, 137
Reid M. J., et al., 2014, ApJ, 783, 130
Reid M. J., Dame T. M., Menten K. M., Brunthaler A., 2016, ApJ, 823, 77
Rice T. S., Goodman A. A., Bergin E. A., Beaumont C., Dame T. M., 2016, ApJ, 822, 52
Rigby A. J., et al., 2016, MNRAS, 456, 2885
Rigby A. J., et al., 2019, arXiv e-prints, p. arXiv:1909.04714
Roberts W. W., 1969, ApJ, 158, 123
Roman-Duval J., Jackson J. M., Heyer M., Johnson A., Rathborne J., Shah R., Simon R., 2009, ApJ, 699, 1153
Roman-Duval J., Jackson J. M., Heyer M., Rathborne J., Simon R., 2010, ApJ, 723, 492
Romero-Gómez M., Mateu C., Aguilar L., Figueras F., Castro-Ginard A., 2019, A&A, 627, A150
Rosolowsky E., Blitz L., 2005, ApJ, 623, 826
Rosolowsky E. W., Pineda J. E., Kauffmann J., Goodman A. A., 2008, ApJ, 679, 1338
Rosolowsky E., et al., 2010, ApJS, 188, 123
Schruba A., et al., 2017, ApJ, 835, 278
Schuller F., Menten K. M., Contreras Y., Wyrowski F., Schilke e. a., 2009, A&A, 504, 415
Schuller F., et al., 2017, A&A, 601, A124
Schuller F., Colombo D., Csengeri T., Duarte-Cabral A., Ginsburg A., Urquhart J. S., 2019, in prep.
Scoville N. Z., Solomon P. M., 1975, ApJ, 199, L105
Sewilo M., Watson C., Araya E., Churchwell E., Hofner P., Kurtz S., 2004, ApJS, 154, 553
Shetty R., Ostriker E. C., 2006, ApJ, 647, 997
Siebert A., et al., 2011, MNRAS, 412, 2026
Siebert A., et al., 2012, MNRAS, 425, 2335
Simon R., Rathborne J. M., Shah R. Y., Jackson J. M., Chambers E. T., 2006, ApJ, 653, 1325
Solomon P. M., Rivolo A. R., Barrett J., Yahil A., 1987, ApJ, 319, 730
Stark A. A., Lee Y., 2006, ApJ, 641, L113
Stutzki J., Guesten R., 1990, ApJ, 356, 513
Sun J., et al., 2018, The Astrophysical Journal, 860, 172
Tavakoli M., 2012, arXiv e-prints, p. arXiv:1207.6150
Taylor J. H., Cordes J. M., 1993, ApJ, 411, 674
Toomre A., 1977, ARA&A, 15, 437
Tosaki T., et al., 2017, PASJ, 69, 18
Traficante A., Fuller G. A., Smith R. J., Billot N., Duarte-Cabral A., Peretto N., Molinari S., Pineda J. E., 2018a, MNRAS, 473, 4975
Traficante A., Lee Y. N., Hennebelle P., Molinari S., Kauffmann J., Pillai T., 2018b, A&A, 619, L7
Urquhart J. S., et al., 2012, MNRAS, 420, 1656
Urquhart J. S., et al., 2013a, MNRAS, 431, 1752
Urquhart J. S., et al., 2013b, MNRAS, 435, 400
Urquhart J. S., Figura C. C., Moore T. J. T., Hoare M. G., Lumsden S. L., Mottram J. C., Thompson M. A., Oudmaijer R. D., 2014a, MNRAS, 437, 1791
Urquhart J. S., et al., 2014b, MNRAS, 443, 1555
Urquhart J. S., et al., 2014c, A&A, 568, A41
Urquhart J. S., et al., 2015, MNRAS, 446, 3461
Urquhart J. S., et al., 2018, MNRAS, 473, 1059
Urquhart J. S., et al., 2019, MNRAS, 484, 4444
Utomo D., Blitz L., Davis T., Rosolowsky E., Bureau M., Cappellari M., Sarzi M., 2015, ApJ, 803, 16
Vallée J. P., 2014, ApJS, 215, 1
Vallée J. P., 2017, The Astronomical Review, 13, 113
Wang K., Testi L., Ginsburg A., Walmsley C. M., Molinari S., Schisano E., 2015, MNRAS, 450, 4043
Watkins E. J., Peretto N., Marsh K., Fuller G. A., 2019, arXiv e-prints, p. arXiv:1906.09275
Wei L. H., Keto E., Ho L. C., 2012, ApJ, 750, 136
Wienen M., Wyrowski F., Schuller F., Menten K. M., Walmsley C. M., Bronfman L., Motte F., 2012, A&A, 544, A146
Wienen M., et al., 2015, A&A, 579, A91
Wienen M., Wyrowski F., Menten K. M., Urquhart J. S., Walmsley C. M., Csengeri T., Koribalski B. S., Schuller F., 2018, A&A, 609, A125
Williams J. P., de Geus E. J., Blitz L., 1994, ApJ, 428, 693
Wilson C. D., et al., 2011, MNRAS, 410, 1409
Wong T., et al., 2011, ApJS, 197, 16
Wu Y. W., et al., 2014, A&A, 566, A17
Zucker C., Battersby C., Goodman A., 2015, ApJ, 815, 23

[1] *School of Physics & Astronomy, Cardiff University, Queen's building, The parade, Cardiff, CF24 3AA, U.K.*
[2] *Max-Planck-Institut für Radioastronomie (MPIfR), Auf dem Hügel 69, 53121 Bonn, Germany.*
[3] *School of Physical Sciences, University of Kent, Ingram Building, Canterbury, Kent CT2 7NH, U.K.*
[4] *Department of Astronomy, University of Florida, 211 Bryant Space*

Sciences Center, Gainesville, FL, USA
[5] *Laboratoire d'Astrophysique de Marseille, Aix Marseille Université, CNRS, UMR 7326, F-13388 Marseille, France*
[6] *West Virginia University, Department of Physics & Astronomy, P. O. Box 6315, Morgantown, WV 26506, USA*
[7] *Space Science Institute, 4765 Walnut St. Suite B, Boulder, CO 80301, USA*
[8] *School of Science and Technology, University of New England, NSW 2351, Australia*
[9] *Observatorio Astrofisico di Arcetri, Largo Enrico Fermi 5, I-50125 Firenze, Italy*
[10] *Max-Planck-Institut für Astronomie, Königstuhl 17, D-69117 Heidelberg, Germany*
[11] *Laboratoire d'astrophysique de Bordeaux, Univ. Bordeaux, CNRS, B18N, allée Geoffroy Saint-Hilaire, 33615 Pessac, France.*
[12] *Departamento de Astronomía, Universidad de Chile, Casilla 36-D, Santiago, Chile*
[13] *Department of Physics & Astronomy, University of Exeter, Stocker Road, Exeter, EX4 4QL, United Kingdom*
[14] *Astrophysics Research Institute, Liverpool John Moores University, 146 Brownlow Hill, Liverpool, L3 5RF, United Kingdom*
[15] *Haystack Observatory, Massachusetts Institute of Technology, 99 Millstone Road, Westford, MA 01886, USA*
[16] *Korea Astronomy & Space Science Institute, 776 Daedeokdae-ro, 34055 Daejeon, Republic of Korea*
[17] *Department of Physics, Faculty of Science, Hokkaido University, Sapporo 060-0810, Japan*
[18] *Istituto di Astrofisica e Planetologia Spaziali, INAF, via Fosso del Cavaliere 100, I-00133 Roma, Italy*
[19] *I. Physikalisches Institut, Universität zu Köln, Zülpicher Str. 77, D-50937 Köln, Germany*
[20] *European Southern Observatory, Alonso de Cordova 3107, Casilla 19001, Santiago 19, Chile*
[21] *Astronomy Department, University of Wisconsin, 475 North Charter St, Madison, WI 53706, USA*
[22] *Dept. of Space, Earth and Environment, Chalmers University of Technology Onsala Space Observatory, 439 92 Onsala, Sweden*
[23] *INAF - Osservatorio Astronomico di Cagliari, Via della Scienza 5, 09047 Selargius (CA), Italy*
[24] *Univ. Grenoble Alpes, CNRS, IPAG, 38000 Grenoble, France*
[25] *School of Engineering, Macquarie University, NSW 2109, Australia*
[26] *European Southern Observatory, Karl-Schwarzschild-Str. 2, D-85748 Garching bei München, Germany*

APPENDIX A: DATA-PRODUCTS AND CATALOGUES

With this paper, we release the complete catalogue of SEDIGISM molecular clouds as extracted using the SCIMES code, alongside the masks of each molecular cloud in the catalogue as fits files, in **[website placeholder]**. In Fig. A1 to A5 we show the sequence of ℓb and ℓv plots of the survey, with the ^{13}CO peak intensity as the background greyscale, and the SEDIGISM cloud masks overlaid as colours. The full details of the extraction are given in Sect. 3.2, and Table A1 has a description of all the properties recorded in the released catalogue.

Besides these two main data-products, we also provide a few other ancillary materials online, which include dictionaries with the medial axis, and extra tables with more detailed information on the SEDIGISM-ATLASGAL matches, as well as the SEDIGISM matches with the other literature catalogues used for our distance assignment procedure. The full details on the format and content of this extra material are provided alongside those, as README files.

APPENDIX B: HELIOCENTRIC AND GALACTOCENTRIC COORDINATES

In order to calculate the de-projected position of each cloud in the Galaxy, we have converted the (ℓ, b, d) triad into a Heliocentric coordinate system ($x_\odot$,$y_\odot$,$z_\odot$), where the x-axis is defined along the line that connects the sun to the Galactic centre (GC), pointing towards the GC, and the z-axis points north out of the plane (similar to Ellsworth-Bowers et al. 2013). Since the latitude (b) across the SEDIGISM coverage is always below 1°, the contribution of the latitude is negligible in the determination of the x and y coordinates, and we have thus simplified the equations from Ellsworth-Bowers et al. (2013) as:

$$\begin{aligned} x_\odot &= d\,cos(l) \\ y_\odot &= d\,sin(l) \\ z_\odot &= d\,sin(b) \end{aligned} \tag{B1}$$

We then also estimate the coordinates in a Galactocentric reference frame ($x_{\rm gal}$,$y_{\rm gal}$,$z_{\rm gal}$), centred in the GC, and in which the y-axis is now the line connecting the GC to the Sun, pointing outwards (note that this is rotated by 90° with respect to the reference frame used in Ellsworth-Bowers et al. 2013). For $z_{\rm gal}$, we need to include the correction for the fact that the Sun does not lie exactly in the Galactic plane but is slight above (e.g. Ellsworth-Bowers et al. 2013), by introducing a rotation angle $\theta = sin^{-1}(z_0/R_0)$, where $z_0 = 0.025$ kpc is the vertical displacement of the Sun above the Galactic midplane, and $R_0 = 8.34$ kpc is the distance of the Sun to the Galactic centre (Reid et al. 2016). As for the Heliocentric coordinates, we ignore the negligible contributions from the small latitude b across the SEDIGISM coverage, as well as from z_0, on the calculations of $x_{\rm gal}$ and $y_{\rm gal}$. As such, our simplified equations for the determination of the Galactocentric coordinates are:

$$\begin{aligned} x_{\rm gal} &= d\,sin(l) \\ y_{\rm gal} &= R_0 - d\,cos(l) \\ z_{\rm gal} &= R_0\,sin(\theta) - d\,cos(l)\,sin(\theta) \end{aligned} \tag{B2}$$

The Galactocentric distance ($R_{\rm gal}$), can then be determined as $R_{\rm gal} = \sqrt{x_{\rm gal}^2 + y_{\rm gal}^2}$.

APPENDIX C: SENSITIVITY/COMPLETENESS LIMIT OF THE SEDIGISM CLOUD CATALOGUE

We have estimated a proxy for the completeness limit of the SEDIGISM dataset, based on the sensitivity and resolution of the data (i.e. estimating a robust detection limit), following the approach used by Heyer et al. (2001) and Colombo et al. (2019). To that purpose, we start by estimating a "luminosity" completeness limit, L^c, at a 5σ confidence level, defined as:

$$L^c = L^{\rm min} + 5\sigma_L \tag{C1}$$

where $L^{\rm min}$ is the minimum luminosity we detect, defined as

$$L^{\rm min}\,[{\rm K\,km\,s^{-1}\,pc^2}] = N_{vox} T_{th} \Delta v \Omega_p d^2, \tag{C2}$$

and

$$\sigma_L\,[{\rm K\,km\,s^{-1}\,pc^2}] = \sigma_{rms} \sqrt{N_{vox}} \Delta v \Omega_p d^2, \tag{C3}$$

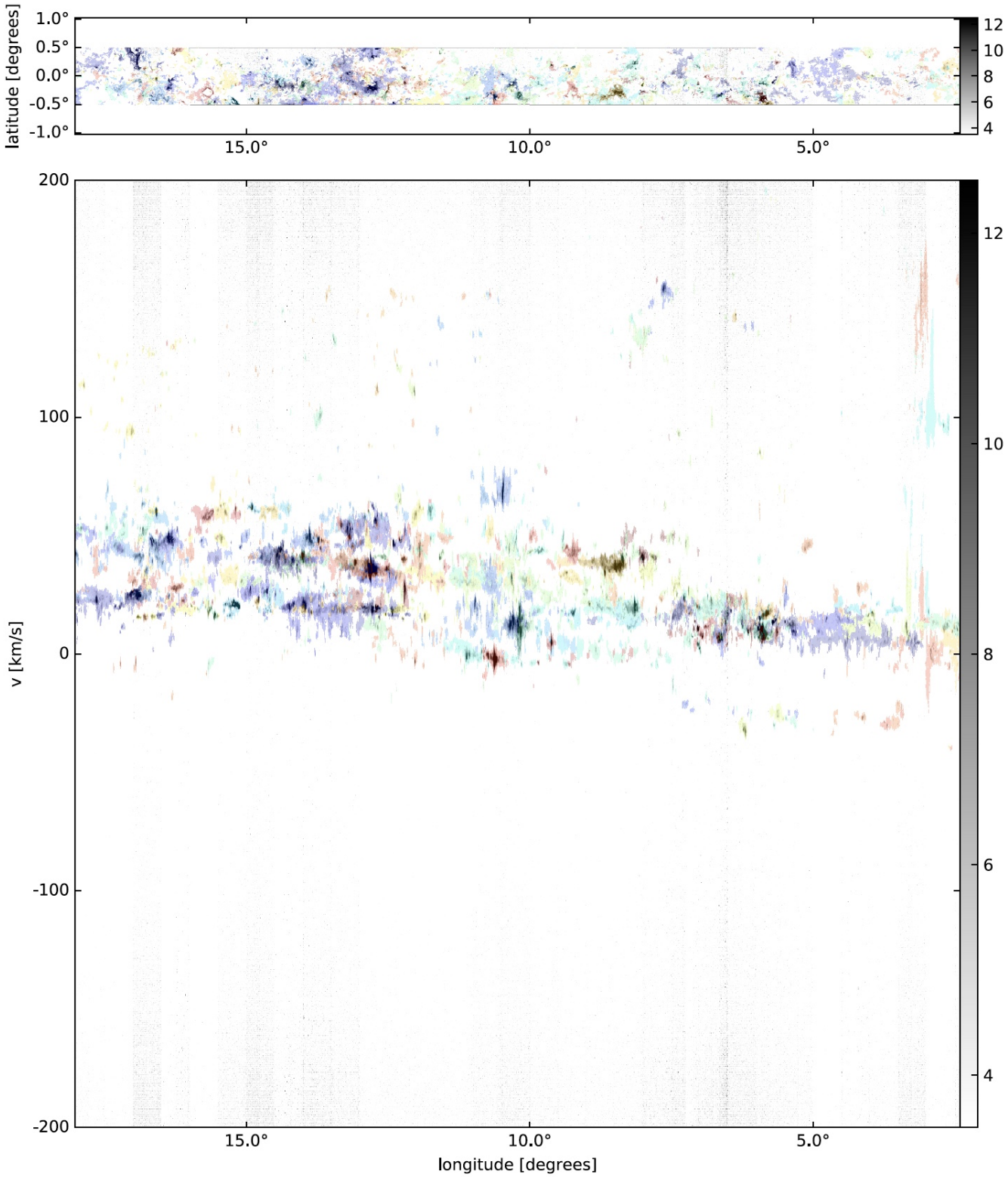

Figure A1. ℓb and ℓv plots of the SEDIGISM survey, with the ^{13}CO peak intensity as the background greyscale, and the SEDIGISM clouds overlaid as colours (each cloud has a different colour, and the colouring scheme is random, but consistent between ℓb and ℓv plots).

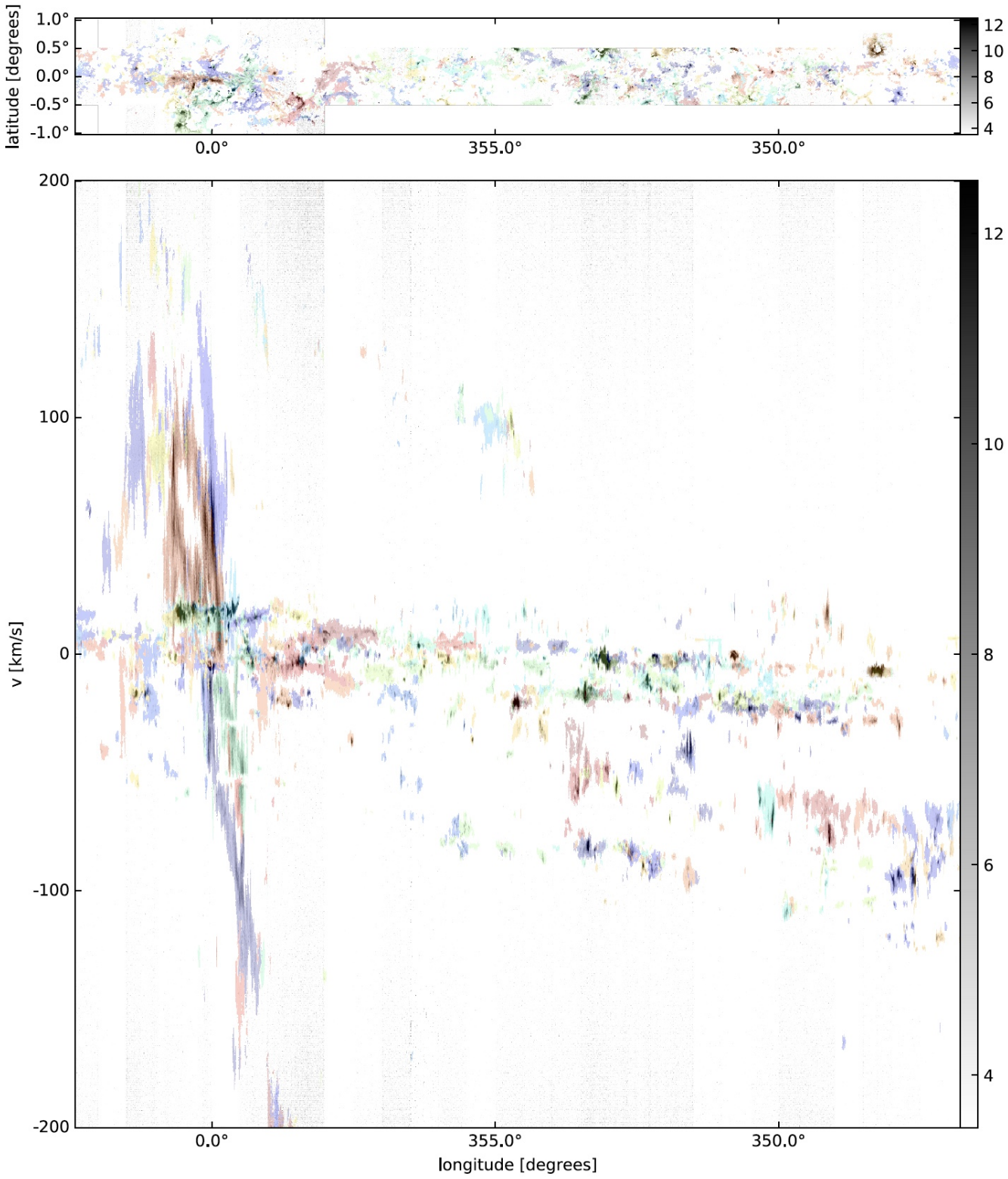

Figure A2. Fig.A1 continued.

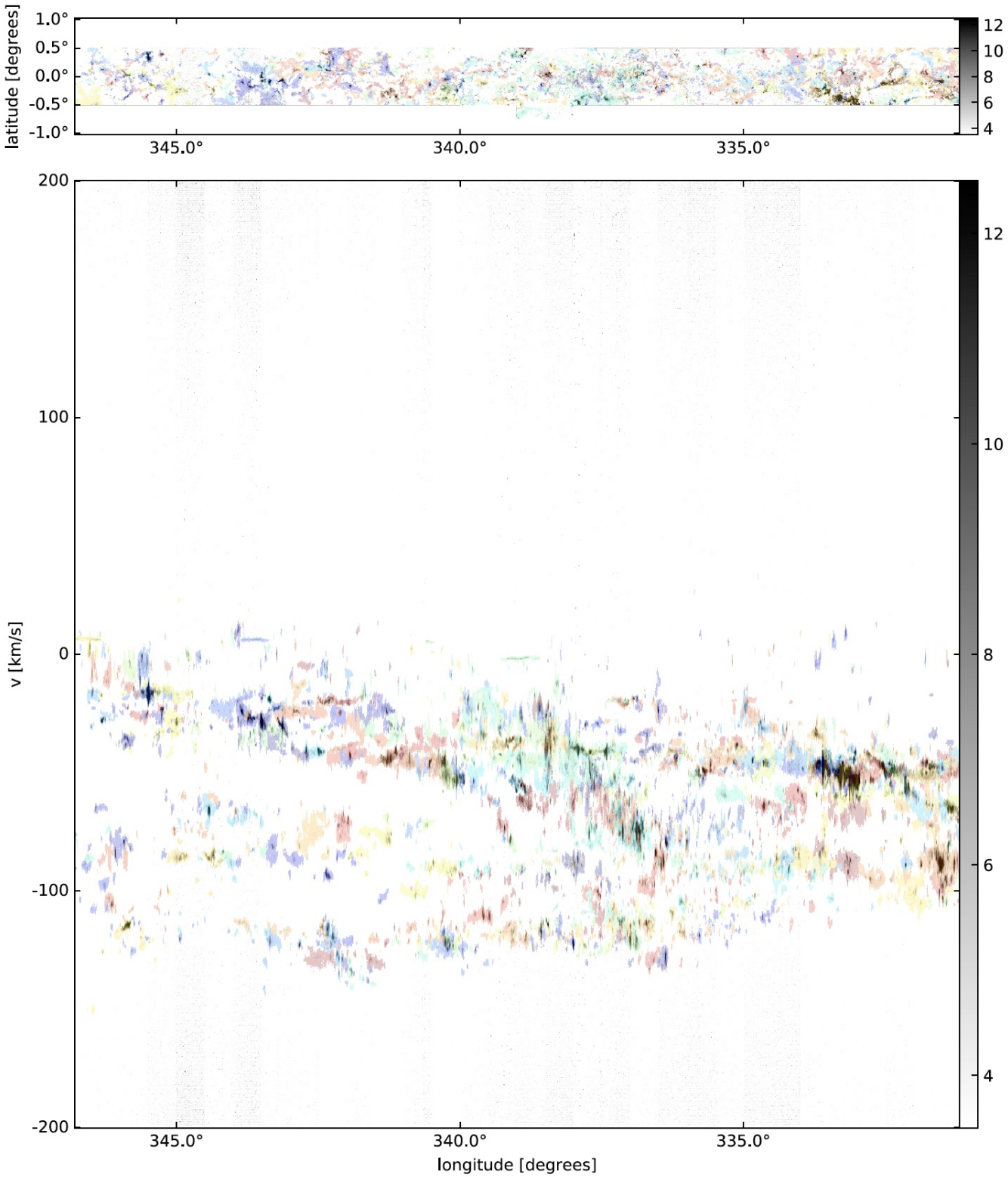

Figure A3. Fig.A1 continued.

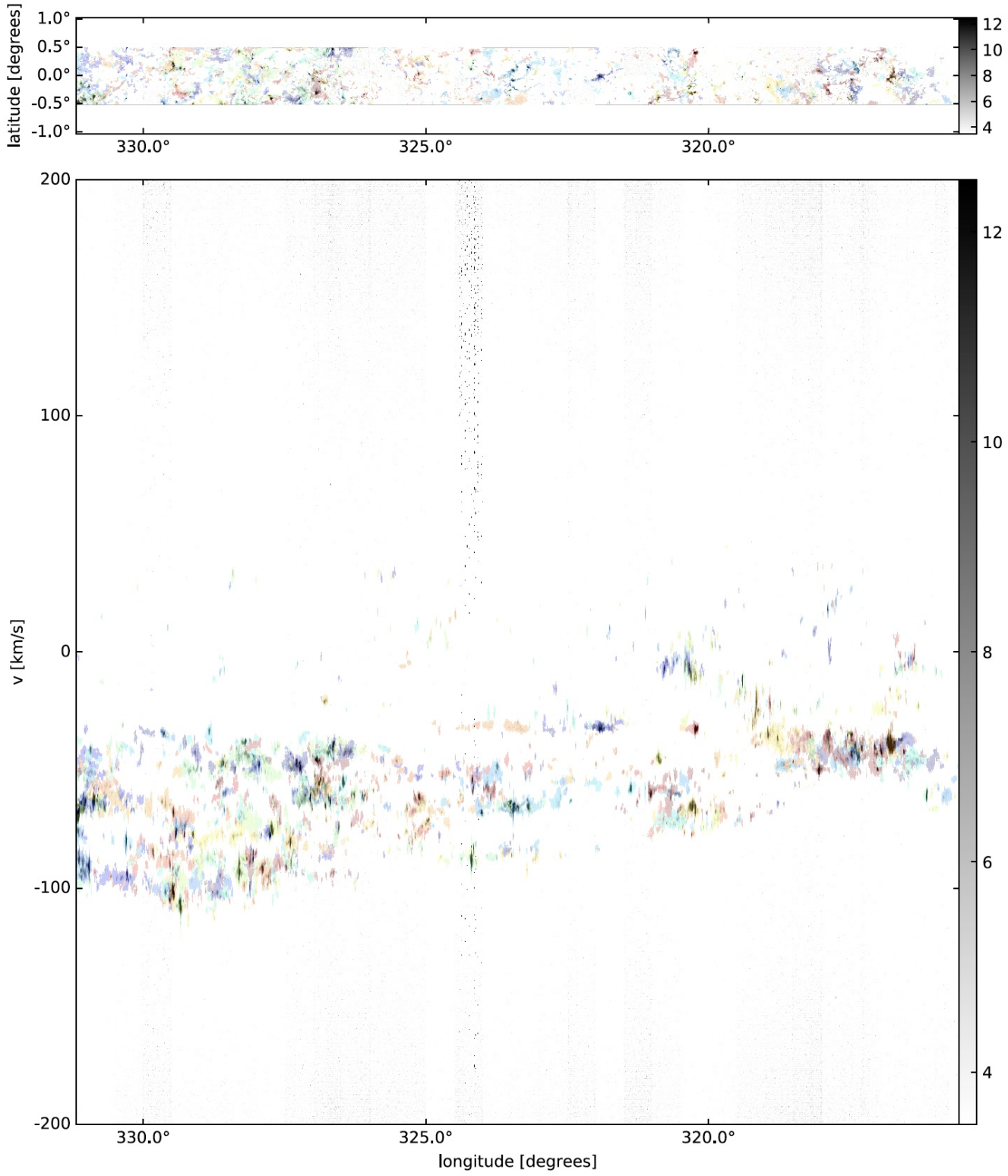

Figure A4. Fig.A1 continued.

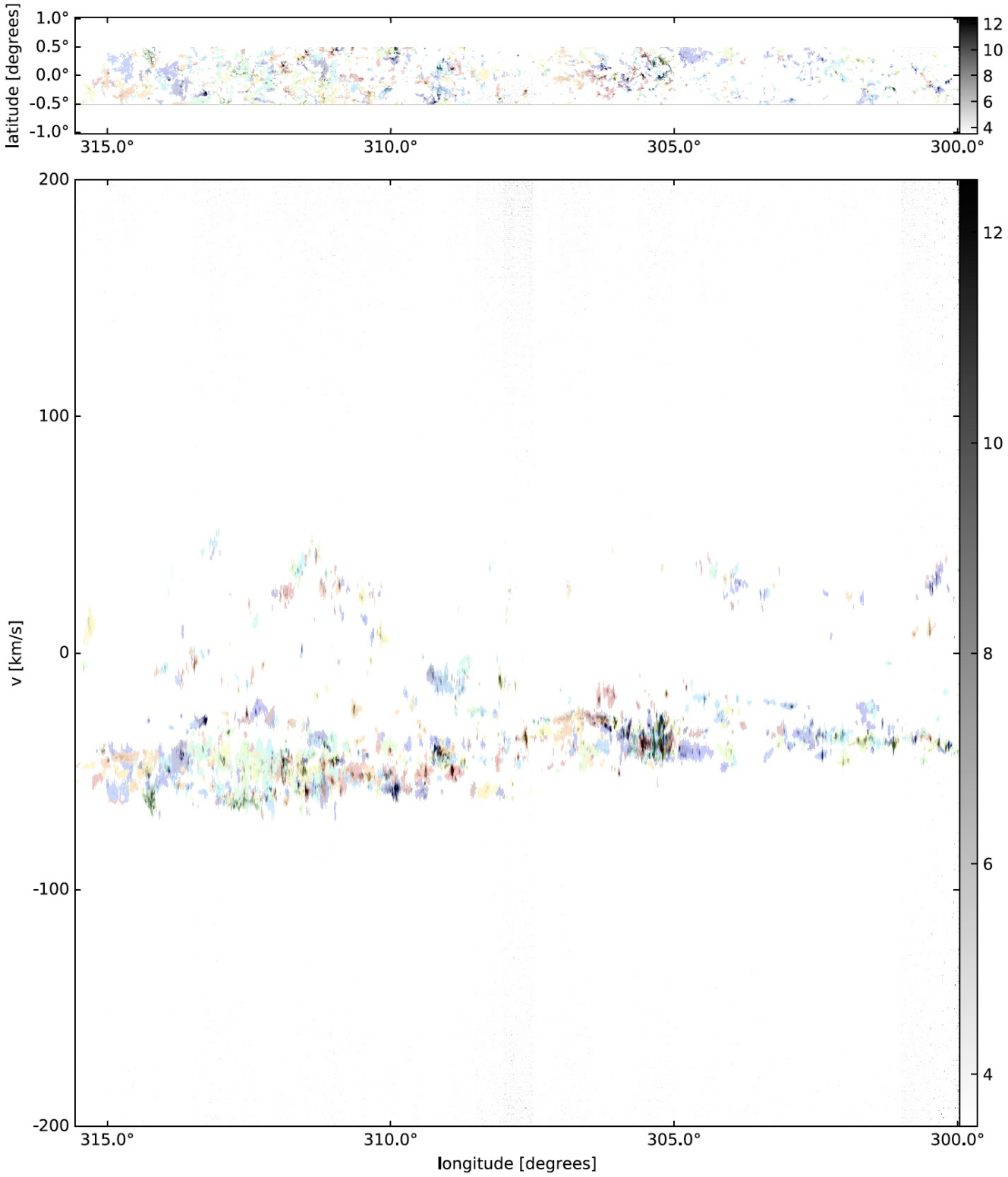

Figure A5. Fig.A1 continued.

Table A1. Description of the SEDIGISM catalogue contents

Catalogue column	Description
cloud_id	Unique cloud ID number
cloud_name	Cloud name as per the SEDIGISM naming scheme, i.e. SDG followed by the Galactic coordinates of the cloud
lon_deg	Galactic Longitude of the cloud's centroid, ℓ (deg)
lat_deg	Galactic Latitude of the cloud's centroid, b (deg)
vlsr_kms	Systemic velocity, $v_{\rm lsr}$ (km s^{-1})
sigv_kms	Velocity dispersion, σ_v (km s^{-1})
area_as	Exact footprint area (arcsec^2)
radius_eq_as	Equivalent radius, estimated using the footprint area, R (arcsec)
major_as	Semi-major axis, *major* (arcsec)
minor_as	Semi-minor axis, $minor$ (arcsec)
pa_deg	Position angle of the major axis, with 0° being along the x/ℓ axis, PA (degrees)
pca_axis_ratio	Aspect ratio from the moments, $AR_{\rm mom}$ (i.e. major_as/minor_as)
medaxis_length_as	Projected geometrical medial axis length, $length_{\rm MA}$ (as)
medaxis_width_as	Projected geometrical medial axis width, $width_{\rm MA}$ (as)
medaxis_ratio	Aspect ratio from the medial axis, $AR_{\rm MA}$ (i.e. medaxis_length_as /medaxis_width_as)
ave_wco_Kkms	Average ^{13}CO (2-1) integrated intensity, $< I_{^{13}\rm CO} >$ (K km s^{-1})
peak_Ico_K	Peak ^{13}CO (2-1) intensity, $T_{^{13}\rm CO}^{\rm peak}$ (K)
sn_ratio	signal-to-noise ratio (SNR = peak intensity / local noise level)
n_pixel	Number of 3D pixels (i.e. voxels) in the cloud
n_leaves	Number of individual dendrogram leaves comprised in the cloud
orig_file	Name of the field from which the cloud was originally extracted
edge	tag identifying whether a cloud touches an edge of the field (yes=1, no=0)
d_near	Near kinematic distance (kpc)
d_near_err	Uncertainty on near distance (kpc)
d_far	Far kinematic distance (kpc)
d_far_err	Uncertainty on far distance (kpc)
dist_kpc	Final adopted distance, d (kpc)
dist_err_kpc	Uncertainty on final distance (kpc)
d_flag	Flag describing the method by which the final distance was decided, $d_{\rm flag}$ (as per Table 1)
d_solution	Flag describing the type of distance solution, d_{sol} (NA = Not Ambiguous, T = tangent, N = Near, F = Far, M = Maser)
d_reliable	Flag to indicate sources with a reliable distance, $d_{reliable}$ (1 = reliable, 0 = non-reliable – as per Sect. 4.1)
tag_hisa	Flag with the result from our automated HI SA determination (1 = strong HI SA; 0 = ambiguous; -1 = no HI SA).
nb_AGAL_matches_total	Total number of ATLAGAL matches
nb_AGAL_matches_perfect	Number of ATLASGAL perfect matches
nb_AGAL_matches_partial	Number of ATLASGAL partial matches
nb_AGAL_nodistance	Number of ATLASGAL matches with no distance assigned
HMSF	Tag identifying whether a cloud has a HMSF tracer (1 = yes, 0 = no)
area_pc2	Exact footprint area (pc^2)
radius_eq_pc	Equivalent radius, estimated using the footprint area, R (pc)
major_pc	Semi-major axis, *major* (pc)
minor_pc	Semi-minor axis, $minor$ (pc)
medaxis_length_pc	Projected geometrical medial axis length, $length_{\rm MA}$ (pc)
medaxis_width_pc	Projected geometrical medial axis width, $width_{\rm MA}$ (pc)
Mass	Cloud mass, M ($M_\odot$)
Column_density_cm2	Cloud's average column density, N (cm^{-2})
Surf_density_Mpc2	Cloud's average gas surface density, Σ ($M_\odot$ pc^{-2})
alpha_vir	Virial parameter, α_{vir}
radius_dec_pc	Deconvolved equivalent radius, R^d (pc)
Surf_density_dec_Mpc2	Surface density, calculated using the deconvolved radius, Σ^d ($M_\odot$ pc^{-2})
alpha_vir_dec	Virial parameter, calculating using the deconvolved radius, α_{vir}^d
x_sun_kpc	x in Heliocentric coordinates, $x_\odot$ (kpc)
y_sun_kpc	y in Heliocentric coordinates, $y_\odot$ (kpc)
z_sun_kpc	z in Heliocentric coordinates, $z_\odot$ (kpc)
x_gal_kpc	x in Galactocentric coordinates, $x_{\rm gal}$ (kpc)
y_gal_kpc	x in Galactocentric coordinates, $y_{\rm gal}$ (kpc)
z_gal_kpc	x in Galactocentric coordinates, $z_{\rm gal}$ (kpc)
R_gal	Galactocentric distance, $R_{\rm gal}$ (kpc)

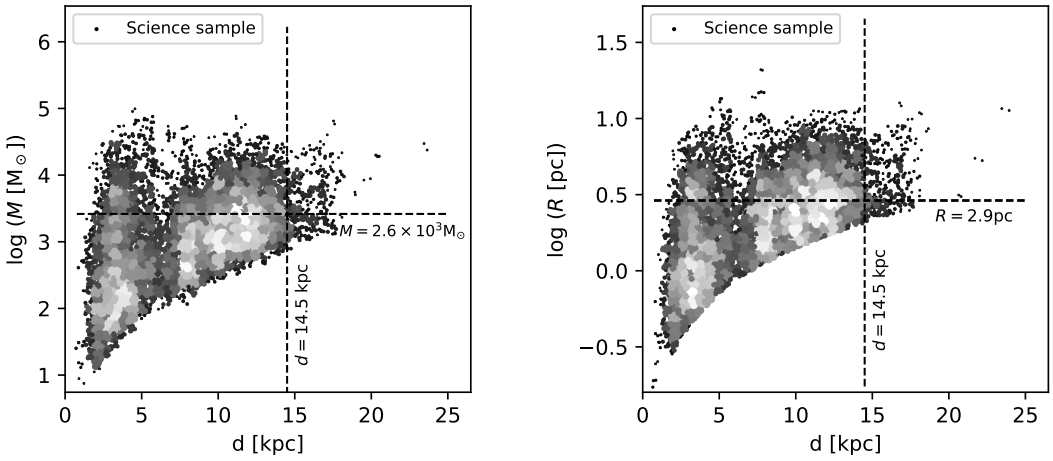

Figure C1. Scatter-density plots showing the distances to the clouds in the science sample (in grey scale), with the completeness limits adopted for the complete science sample (in terms of mass on the left, and equivalent radius on the right) plotted as horizontal dashed lines. The vertical line delineates the distance at which the completeness limits were estimated from.

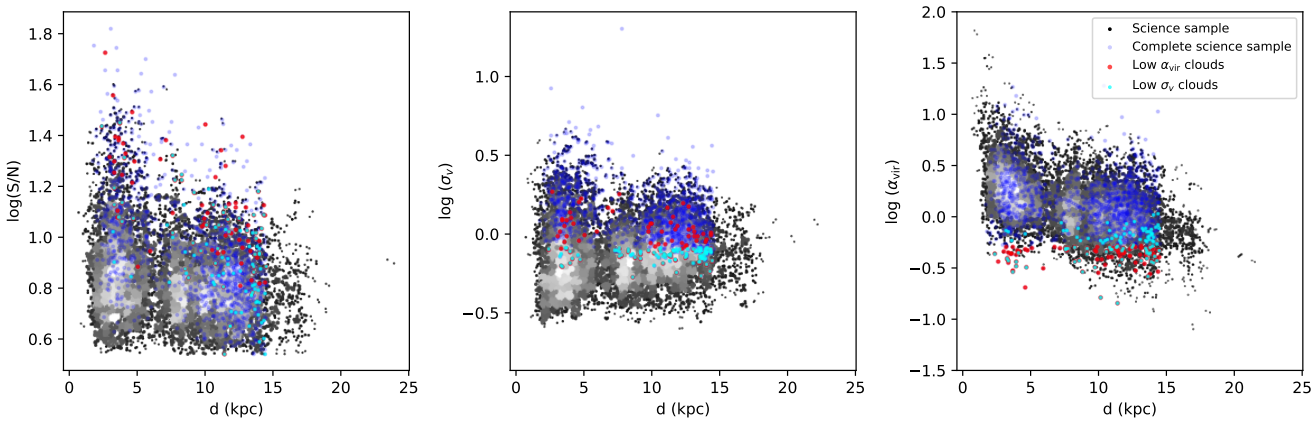

Figure C2. Scatter-density plots showing potential observational biases in some of the derived properties. In particular, we show the signal-to-noise ratio (S/N) as a function of distance on the left panel, the same for the velocity dispersion (σ_v) on the middle panel, and virial parameter (α_{vir}) on the right panel. The grey scale shows all the clouds in the science sample, and the dark-blue points are the clouds within the complete science sample. The red and turquoise points are the sub-samples of 100 clouds with low velocity dispersion, and with low virial parameter, respectively, which are the two distributions that could be potentially most affected by observational biases.

where N_{vox} is the minimum number of 3D pixels (voxels) in the cloud, defined as $N_{vox} = N_p N_c$, where N_p is the minimum number of (spatial) pixels that the cloud has to cover, and N_c the minimum number of (spectral) channels. For this purpose, we want to be as conservative as possible, so that we minimise the possible biases, particularly when considering clouds that are at lower signal-to-noise levels. Therefore, we take $N_p = 6 N_{\rm ppbeam}$, where $N_{\rm ppbeam} = 9$ is the number of pixels per beam (i.e. corresponding to clouds whose footprint size is twice larger than the smallest clouds allowed into the science sample), and $N_c = 4$ (with 2 channels being our spectral resolution element). In essence, this corresponds to clouds that are $\sim$ 4 times larger (in terms of 3D pixels) than those allowed to go through to the dendrogram construction. T_{th} is the sensitivity threshold as a main-beam antenna temperature, which we take as being $6\,\sigma_{rms}$, with $\sigma_{rms} = 0.7$ K being the average noise level used for the dendrogram. This T_{th} effectively corresponds to the first level leaves allowed into the dendrogram, i.e. starting at a level of $2\,\sigma_{rms}$, and with a leaf height of $4\,\sigma_{rms}$ above that. Δv is the channel width (i.e. 0.25 km s^{-1}), Ω_p is the pixel size (Ω_p = 9.5″ × 9.5″ $\approx$ 2.12 × 10^{-9} sr), and finally, d is the distance to the cloud. Given that the vast majority of the SEDIGISM clouds are within 14.5 kpc we use that as our maximum distance for these calculations, although we note that we are still sensitive to clouds beyond those distances, with clouds lying up to kinematic distances of 23 kpc.

From this luminosity completeness limit, we derive a mass completeness limit as $M^c = \alpha_{^{13}{\rm CO}(2-1)}\, L^c$, using a conversion factor $\alpha_{^{13}{\rm CO}(2-1)} = 22.43\,{\rm M}_\odot\,({\rm K\,km\,s^{-1}})^{-1}\,{\rm pc}^{-2}$, estimated from our $X_{^{13}{\rm CO}(2-1)}$ (Sect. 3.2), assuming a molecular weight per hydrogen molecule of 2.8 (Kauffmann et al. 2008). We thus retrieve a mass completeness limit of $M^c = 2.6 \times 10^3$ M$_\odot$ at 14.5 kpc.

We also derive a size completeness limit, which is directly linked to the minimum source size that we can robustly recover. As for the luminosity completeness limit, we take that to be 6 beam sizes, which corresponds to a completeness radius $R^c = 2.9$ pc, at 14.5 kpc distance.

We consider these completeness/detection limit estimates to be rather conservative, since the majority of the sources in our catalogue lie well below a distance of 14.5 kpc, but also because our full catalogue does contain sources that are as small as two beam sizes, making it possible to find clouds that lie well below our completeness limits at 14.5 kpc and beyond. This is clear from Fig. C1 which shows the mass (M) and radius (R) of the science sample, as a function of distance (d), and where we overplot the respective completeness limits at 14.5 kpc (as horizontal lines). For reference, if we simply take clouds up to a distance of 5 kpc (e.g. as in our distance limited sample), then our completeness limits decrease to $M^c = 3.1 \times 10^2\,M_\odot$ and $R^c = 1$ pc.

By using the SCIMES cloud extraction algorithm, which clusters the individual peaks of emission into clusters, we are less prone to have severe observational biases that often arise from the large range of distances probed and the extraction of objects close to the resolution elements of the survey (i.e. both spectrally and spatially). In addition, the rather conservative limits imposed to build the complete science sample should, in principle, also guarantee that the cloud properties that we extract are comparable across the entire survey. However, to test whether this was truly the case, we have investigated whether we could see evidence for any remaining observational biases in the complete science sample, that could potentially affect our results.

The left panel of Fig. C2 shows the peak signal-to-noise ratio (S/N) as a function of distance, for the science sample (in grey) and for the complete science sample (in blue). We can see that while across the entire science sample the average S/N is roughly constant, we have a larger amount of clouds with a high S/N nearby. When we consider only the complete science sample (in dark blue), we effectively discard most of the nearby clouds with lower S/N, which then produces a trend of decreasing S/N with increasing distance from the sun. This could become a problem, as the properties of clouds with lower S/N are less well constrained than those with high S/N. A particularly problematic property is the velocity dispersion (and the respective virial parameter derived from it), because for further away clouds - since we do not do any bootstrap to extrapolate the emission into the noise - we might artificially recover a lower FWHM than we would have, had we been able to detect the full extent of those clouds at higher S/N. The middle and left panels of Fig. C2, highlight this issue, where we can see a clear trend of decreasing σ_v and $\alpha_{\rm vir}$ as a function of distance. There is no obvious physical reason why we should expect these two properties to correlate with their distance to the sun, suggesting that there are still some remaining observational biases at play, despite our best efforts to neutralise them. This could therefore be responsible (at least in part) for the signatures seen in Sect. 6.3 and 6.4. Surprisingly, though, Fig. C2 also shows that the clouds with the low velocity dispersion (marked in turquoise) and those with low virial parameter (shown in red), are not the clouds with the lower S/N values of the sample (which we ought to expect, if these trends were purely driven by cloud segmentation biases and survey limitations). It thus remains to be seen whether some of the signal seen in Sect. 6.3 and 6.4 could be physically driven.

APPENDIX D: SOLVING KDA USING THE SIZE-LINEWIDTH RELATION

In an attempt to solve the kinematic distance ambiguity for clouds for which our methods 0-9 did not work, we explored the position of clouds in a size-linewidth plot, using both near and far distances.

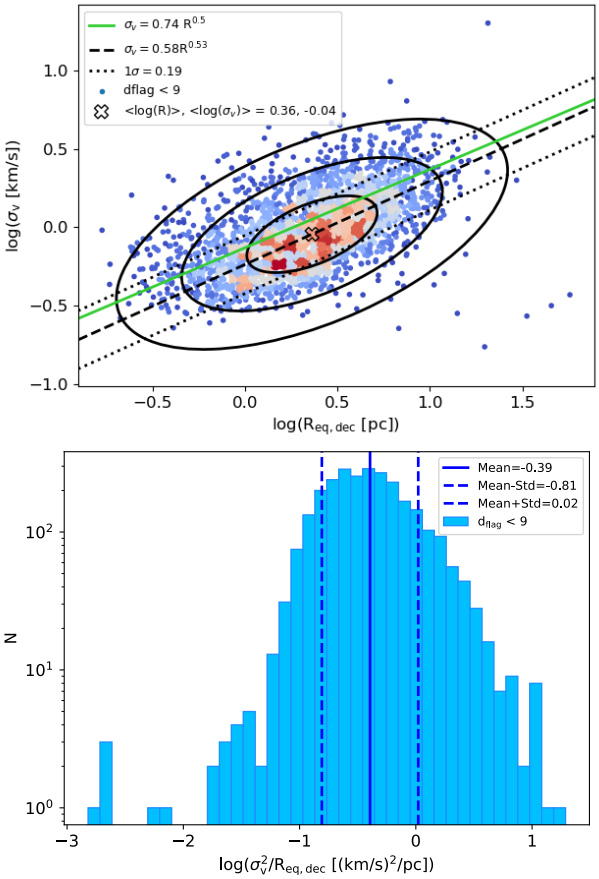

Figure D1. Top: Scatter-density plot of the size-linewidth relation for our full sample of clouds with a solved KDA using methods 0-9. The ellipses show the 1, 2 and 3σ from PCA analysis on our sample, and the dashed line shows the respective slope (i.e. $\sigma_v \propto R^{0.53}$). This slope is consistent with the relation from Solomon et al. (1987), shown as a green line, that lies well within 1σ of our relation. Bottom: Histogram of the size-linewidth relation for all clouds with a solved KDA using methods 0-9, assuming the original $\sigma_v \propto R^{0.5}$ relation. The blue solid and dashed lines show the mean and 1σ standard deviation respectively (in log-space), that we use to determine if a distance solution of a cloud is significantly more likely than the other.

We then determined whether one of those solutions was more likely, based on their position relative to the bulk distribution of clouds.

The original size-linewidth relation was first looked at by Larson (1981), but redefined by Solomon et al. (1987), taking the form of $\sigma_v \propto R^{0.5}$. However, the exact positioning of this relation (with Solomon et al. 1987 placing it at $\sigma_v^2/R = 0.55$) is sensitive to the specific way by which the radius and velocity dispersion are estimated (e.g. sensitive to the specific tracer, cloud extraction algorithm, etc.). Therefore, for our purpose, we calibrate the size-linewidth relation using our data, for clouds with a solved KDA using methods 0-9. The size-linewidth relation for our data is shown in Fig. D1 (top panel), and has an exponent (from a PCA analysis), that is consistent to the exponent $\sigma_v \propto R^{0.5}$ found by Solomon et al. (1987). The bottom panel of Fig. D1 shows the histograms of the values of σ_v^2/R for our data, showing that they follow a log-normal distribution. We use these mean and standard deviation of the σ_v^2/R values (in log-space) to compare to the σ_v^2/R values of the clouds using both the near and far distance solutions. We favour a given distance solution *only* if that solution is significantly closer to the empirical relation than the other solution (i.e. at least a factor 3

difference in log-space), and *only* if we do not have both solutions placing the clouds within 1σ of the underlying distribution (as in that case, both near and far solutions would be equally plausible).

APPENDIX E: GLOBAL PROPERTIES FOR DISTANCE-LIMITED SCIENCE SAMPLE

Figure E1 shows the distributions of a number of different properties, namely the mass (M), the velocity dispersion (σ_v), the medial axis length ($length_{\rm MA}$), the average surface density (Σ), the virial parameter ($\alpha_{\rm vir}$), and the aspect ratio from the medial axis ($AR_{\rm MA}$), for a distance-limited sample, with 2.5 kpc $< d <$ 5.0 kpc, in order to minimise the selection effects due to distances/resolution. Note that at these distances, we are mostly looking at a spiral arm (the Scutum arm), and thus we tend to focus on the sample that has more ATLASGAL matches, missing some of the more diffuse clouds. Overall, the histograms are consistent with the full sample, and the shapes of the distributions follow the same trends - as quantified through the statistics for both the full and the distance-limited samples reported in Table 3. This would suggest that the results inferred from the global sample are significant.

APPENDIX F: 2D STATISTICAL TEST

In order to determine if the spatial distribution of clouds in our sub-samples with extreme properties has any dependency on Galactic environment, we would require a proper modelling of the spiral arms and bar in PPV space, and even then, the uncertainties in the distances would always be a limitation to the interpretation of the results. Therefore, we have instead simply chosen to test whether the spatial distribution of clouds in our sub-samples (as per their de-projection onto the top-down view of the Galaxy) is statistically consistent with the global distribution of clouds. This makes no assumption on the Galactic structure, and is less affected by distance uncertainties - since the sub-sample is purely drawn out of the global population; both are affected in the exact same way. We have used the Pearson's χ^2 statistical test, which tests whether the frequency distribution of certain events observed in a sample is consistent with a particular theoretical distribution.

The specific calculation of the χ^2 statistics is given by:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} = N \sum_{i=1}^{n} \frac{(O_i/N - p_i)^2}{p_i} \quad \text{(F1)}$$

where n is the number of bins (or cells) considered, O_i is the number of observed counts into bin i, N is the total number of observations, $E_i = Np_i$ is the number of expected counts in bin i where p_i in the probability of an observation falling into bin i.

In our case, we assume that the global distribution of clouds in the complete science sample represents our probability function (i.e. our "theoretical" distribution), and we want to assert if the spatial distribution of the most extreme clouds simply follows (statistically) the same distribution of the entire sample, or whether it shows significant deviations. We thus constructed our p_i, by building a 2D probability density function (pdf), of the complete science sample (i.e. a normalised 2D histogram of the spatial distribution - in Galactocentric coordinates - of clouds within the complete science sample), using a spatial bin of 0.3×0.3 kpc[17] (see left panels of Figs. F1, F2 and F3). Note that because our data are sparse, p_i can be 0 in many bins, but we only compute the χ^2 statistics for the n bins that have $p_i > 0$. Similarly to p_i, we then construct O_i as the 2D pdf of the distribution of the N clouds in our sub-sample, using the exact same bins for p_i (see the panels in the central column of Figs. F1, F2 and F3). N in our case is always 100 clouds, i.e. 5% of the complete science sample, except for the HMSF clouds which amount to a total of 211 clouds. The results from this χ^2 statistics, for all of our tested sub-samples, are summarised in Table F1.

In order to quantify the statistical significance of these results for our specific purpose, we have performed a test to determine the likelihood of obtaining a given χ^2-value purely out of a random sampling of our theoretical distribution. To do so, we performed 100 000 draws of $N = 100$ clouds randomly selected from the original sample of clouds (i.e. from the complete science sample), without replacement. For each of those draws, we construct the O_i as the 2D pdf of the distribution of the N clouds, and perform the χ^2 test in the exact same way as for the extreme cloud samples. Figure F4 shows the distribution of χ^2 values obtained as a result of these 100 000 random draws (left panels). The right panel of Fig. F4 shows the cumulative fraction of runs with a χ^2-value above a certain value. From this, we derive our probability, $p_{\rm rnd}$, that the observed χ^2-value comes from a pure random sampling of the theoretical distribution. For instance, there is a 1% change of obtaining a χ^2 above 705 from a pure random sampling of the theoretical distribution. Similarly, there is a 2% chance that the χ^2-value lies above 690, 5% above 670, 10% above 650, 20% above 630, and 30% above 615. Table F1 compiles the $p_{\rm rnd}$ for each of the extreme cloud samples that we studied.

Although the exact χ^2 values and $p_{\rm rnd}$ should not be taken at face value (given the statistical fluctuations, as well as the uncertainties in the distributions, and binning effects, neither of which are taken into account), they can be useful for a relative comparison of the sub-samples. Indeed, the lower the χ^2 value, the higher the $p_{\rm rnd}$, and the closer the distribution of the sub-sample matches the global one. From these statistics we can start to quantify which distributions are less-like the original distribution of clouds, and identify which properties of clouds could be most affected by the Galactic environment.

To support this interpretation, besides the pure statistical test that compares our sub-samples to the global cloud population, we also checked where (spatially) the disparities between the theoretical and observed distributions were coming from. In order to do that, we produced "difference maps" between the observed and expected distributions, after rescaling the observed distribution to have the same mean and standard deviation as the theoretical one (so that they share a common reference frame). These difference maps are shown on the right panels of Figs. F1, F2 and F3. Note that because we performed a re-scaling of the observed distribution, the absolute values of the difference are not meaningful. Instead, these plots are only meant to illustrate, qualitatively, where the differences between observed and predicted distributions are, with regions that have a

[17] The need to produce these regular spatial bins is the reason why we have excluded clouds at the tangent distance for this exercise. If using those, we would be effectively including bins that have sources regrouped from a larger spatial range than the bin size. In order to include tangent clouds, we would have to introduce a different weight to the bins at the tangent distances, to effectively account for the larger areas covered. This, however, is not straight forward to produce, since the binning of sources onto their tangent distance was made based on their line of sight velocity, which effectively means a variable spatial range, and also directed solely along the line of sight (rather than along any of the Galactocentric cartesian coordinates).

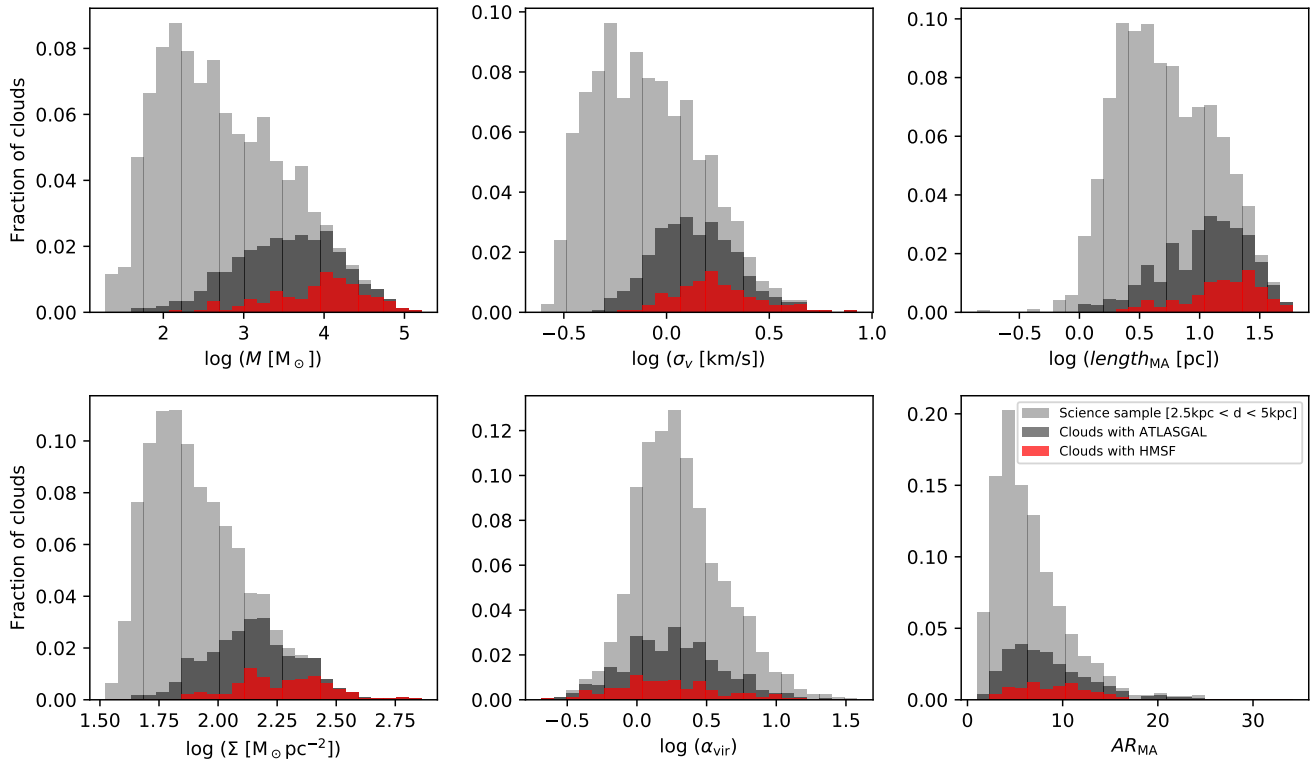

Figure E1. Same as Fig. 8, showing the histograms of global properties but for a distance-limited sample (2.5 kpc $< d <$ 5.0 kpc), in order to minimise distance biases. The panels represent the distributions of: Mass (top-left), velocity dispersion (top-centre), medial axis length (top-right), average surface density (bottom-left), virial parameters (bottom-centre), and aspect ratio from the medial axis (bottom-right). The histograms shown are for the distance-limited science sample (light grey), along with clouds that have an ATLASGAL counterpart (dark grey), and clouds that have a HMSF signpost (red). The normalisation of all histograms was made with respect to the total number of clouds in the distance-limited sample.

Table F1. Results from the 2D χ^2 statistical tests, for all the sub-samples of extreme clouds, compared to the global spatial distribution of clouds in the complete science sample. p_{rnd} specifies the likelihood of obtaining the respective χ^2-value from a pure random sampling of $N = 100$ clouds from the theoretical distribution.

Condition	χ^2	p_{rnd}
$M > 3.2 \times 10^4$ M$_\odot$	670	0.05
$\Sigma > 209$ M$_\odot$pc^{-2}	638	0.16
$length_{MA} > 40$ pc	715	0.005
$AR_{MA} > 14$	671	0.04
$\alpha_{vir} > 3.7$	699	0.01
$\sigma_v > 2.5$ km s^{-1}	747	0.001
$\alpha_{vir} < 0.4$	675	0.04
$\sigma_v < 0.6$ km s^{-1}	669	0.05
HMSF	735	0.001$^{(*)}$

$^{(*)}$ estimated for a random sampling of $N = 211$ clouds.

relative excess of counts shown in dark red, and regions with a relative lack of counts shown in dark blue.

APPENDIX G: OPACITY LAWS AND THEIR EFFECT ON THE HMSF THRESHOLD

The original empirical threshold for HMSF, of $M[\mathrm{M}\odot] = 870(R[\mathrm{pc}])^{1.33}$, from Kauffmann & Pillai (2010) was determined using a combination of dust extinction and dust emission measurements. When determining gas masses from dust emission, however, we are required to adopt an opacity law and specific dust opacities, both of which are still largely uncertain. Some works (such as Battersby et al. 2011), adopt the Ossenkopf & Henning (1994) specific opacity of $k_0 = 4\,\mathrm{cm^2 g^{-1}}$ at 505 GHz, and an opacity law as $k_\nu = k_0(\nu/505\mathrm{GHz})^{1.75}$. Kauffmann & Pillai (2010) also use the Ossenkopf & Henning (1994) opacities, but include an additional correction of a factor 1.5, so that the mass estimates from dust extinction were consistent to those from dust emission at 1 mm. In other words, the equivalent opacity law from Kauffmann & Pillai (2010) would be $k_\nu = 12.1\,\mathrm{cm^2 g^{-1}}(\nu/1200\mathrm{GHz})^{1.75}$ - and this is the opacity law for which the original threshold is applicable.

For works that do not include this correction for their mass estimates (such as the masses from Battersby et al. 2011), would therefore require to be compared to the equivalent threshold relation, without the 1.5 scaling, i.e. $M[\mathrm{M}\odot] = 580(R[\mathrm{pc}])^{1.33}$. The ATLASGAL works (such as Urquhart et al. 2018), adopt a slightly different opacity law: they take the same specific opacities from Ossenkopf & Henning (1994), but without the 1.5 factor correction, and using a spectral index of 2 (instead of 1.75). With this change in the power law index plus the non-adoption of the 1.5 correction factor in the opacities, makes the ATLASGAL dust masses lower

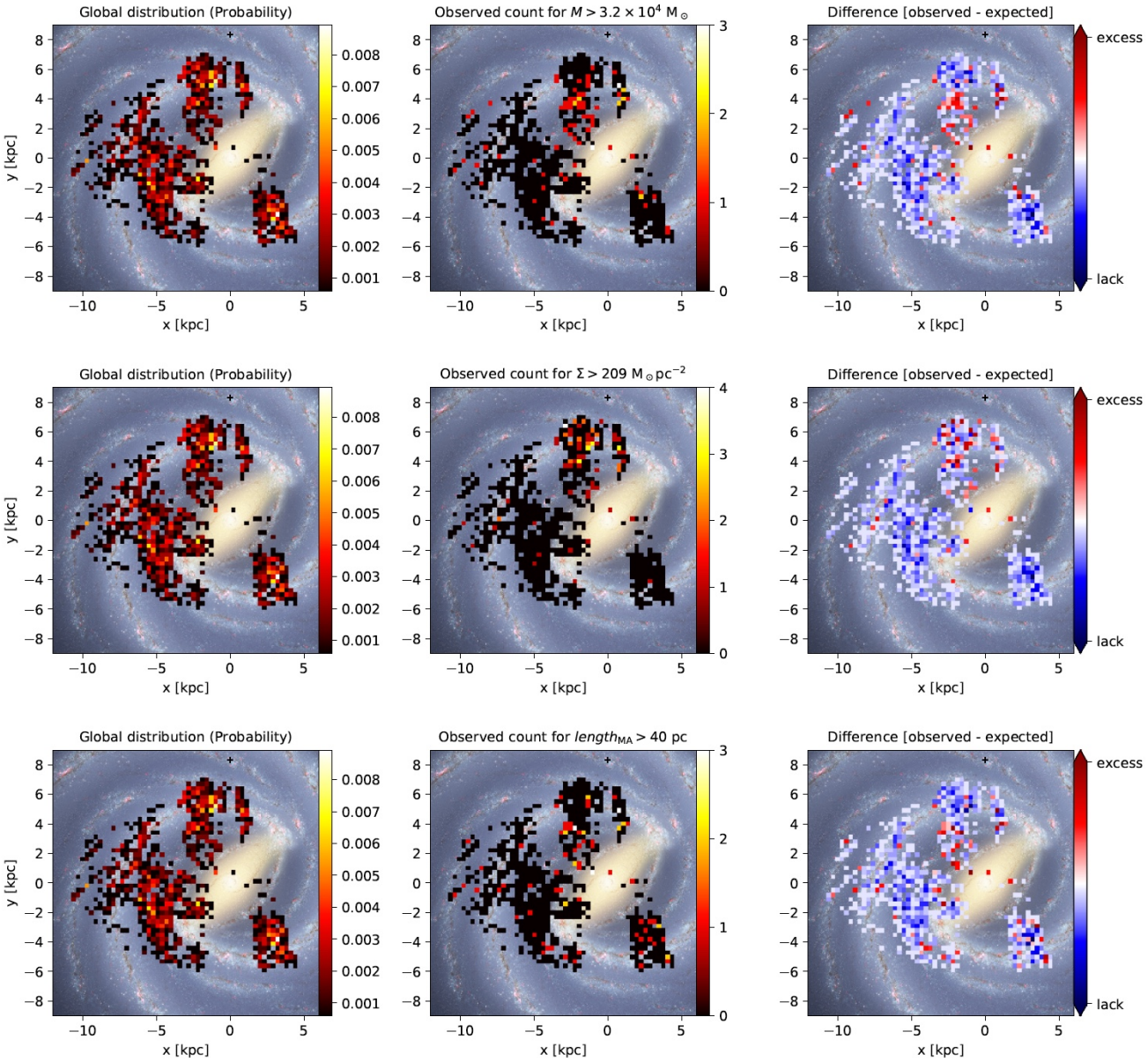

Figure F1. Left column: Normalised 2D histogram of the Galactic distribution of all clouds within the complete science sample, representing our "theoretical probability" for the χ^2 test. Each 2D bin corresponds to a cell of 0.3 kpc × 0.3 kpc. Middle: Observed number of clouds from a specific sub-sample in each of the 2D bins (defined as in the left panels). Each row shows a different tail of a distribution, using the 100 clouds with highest mass (top row), surface density (middle row), and length (bottom row). The specific condition to select these subsamples is specified on the top of each of these middle panels. Right column: Rescaled difference between the observed and expected distributions (i.e. between the middle and left panels), with red representing a relative excess, and blue a relative lack of counts with respect to the statistical prediction.

by roughly a factor 2 with respect to the masses from Kauffmann & Pillai (2010).

In our work, the gas masses were derived by calibrating the ^{13}CO (2 – 1) emission against the column density maps from the Herschel Hi-GAL survey (Molinari et al. 2010), which use a different opacity law from the above (Elia et al. 2013). It adopts the Hildebrand (1983) opacity at 250 μm and a spectral index of 2, resulting in $k_\nu = 10\,\mathrm{cm^2 g^{-1}} (\nu/1200\mathrm{GHz})^2$. We note that our conversion factor from ^{13}CO (2 – 1) integrated intensities into H_2 column densities as per this comparison with the Hi-GAL data was also remarkably consistent with the conversion factor derived with a multi-transition modelling of the line emission (combining SEDIGISM and THrUMMS data, Schuller et al. 2017).

At Herschel wavelengths, the difference between our opacity law and that used by Kauffmann & Pillai (2010) introduces only a small difference of ~ 20% on the masses (well within the overall uncertainties in our mass estimates). Nevertheless, for consistency, we scale the Kauffmann & Pillai (2010) HMSF threshold line to match our particular opacity law, bringing the threshold to $M[\mathrm{M}\odot] = 1053(R[\mathrm{pc}])^{1.33}$. We note that the fraction of SEDIGISM clouds above and below the threshold line fluctuates only by ~ 10% if we were to adopt the original relation, thus not changing the global trends and results that we find.

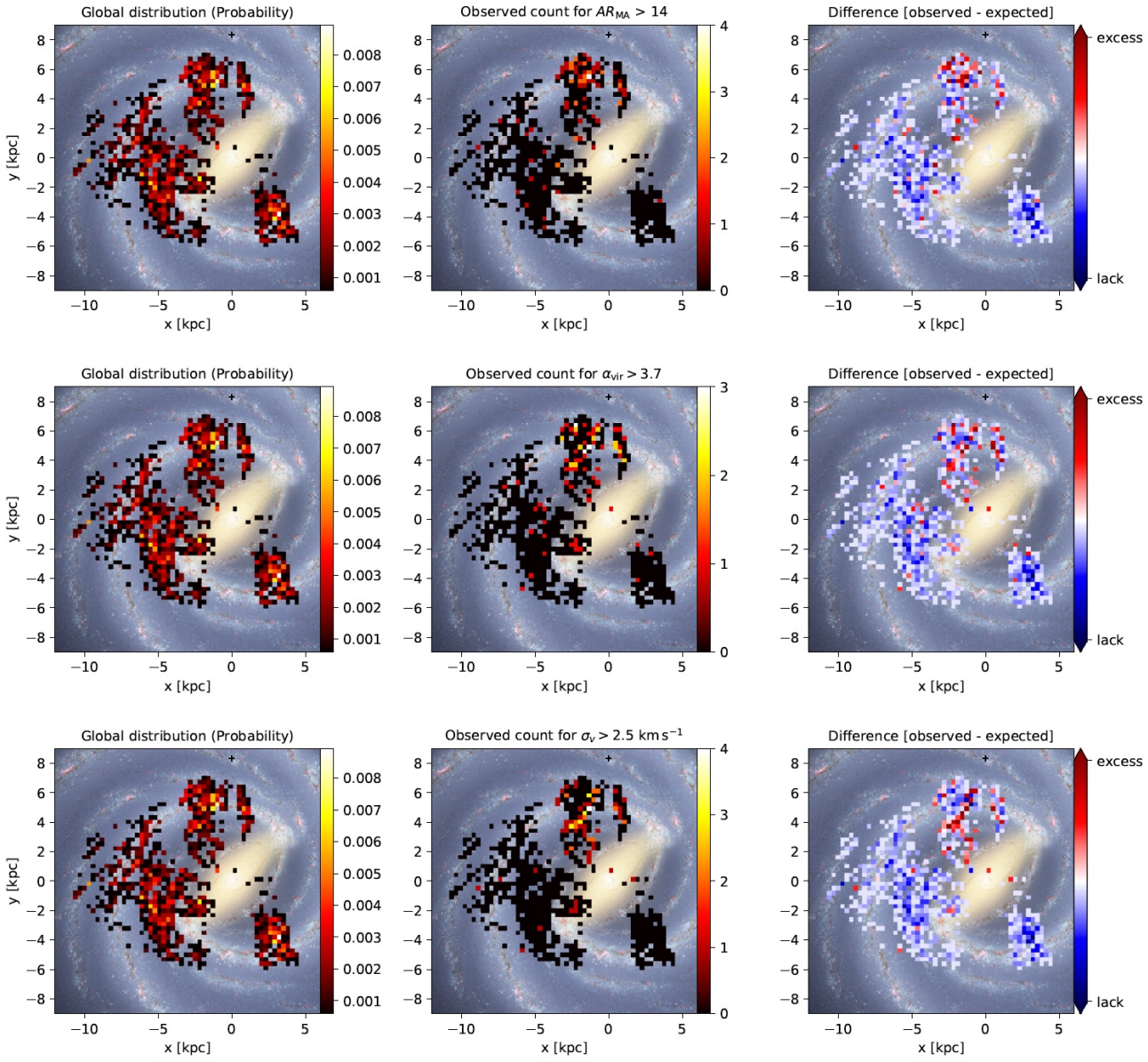

Figure F2. Same as Fig. F1, but for the 100 clouds with the largest aspect ratio (top), highest virial parameter (middle), and highest velocity dispersion (bottom).

This paper has been typeset from a TEX/LATEX file prepared by the author.

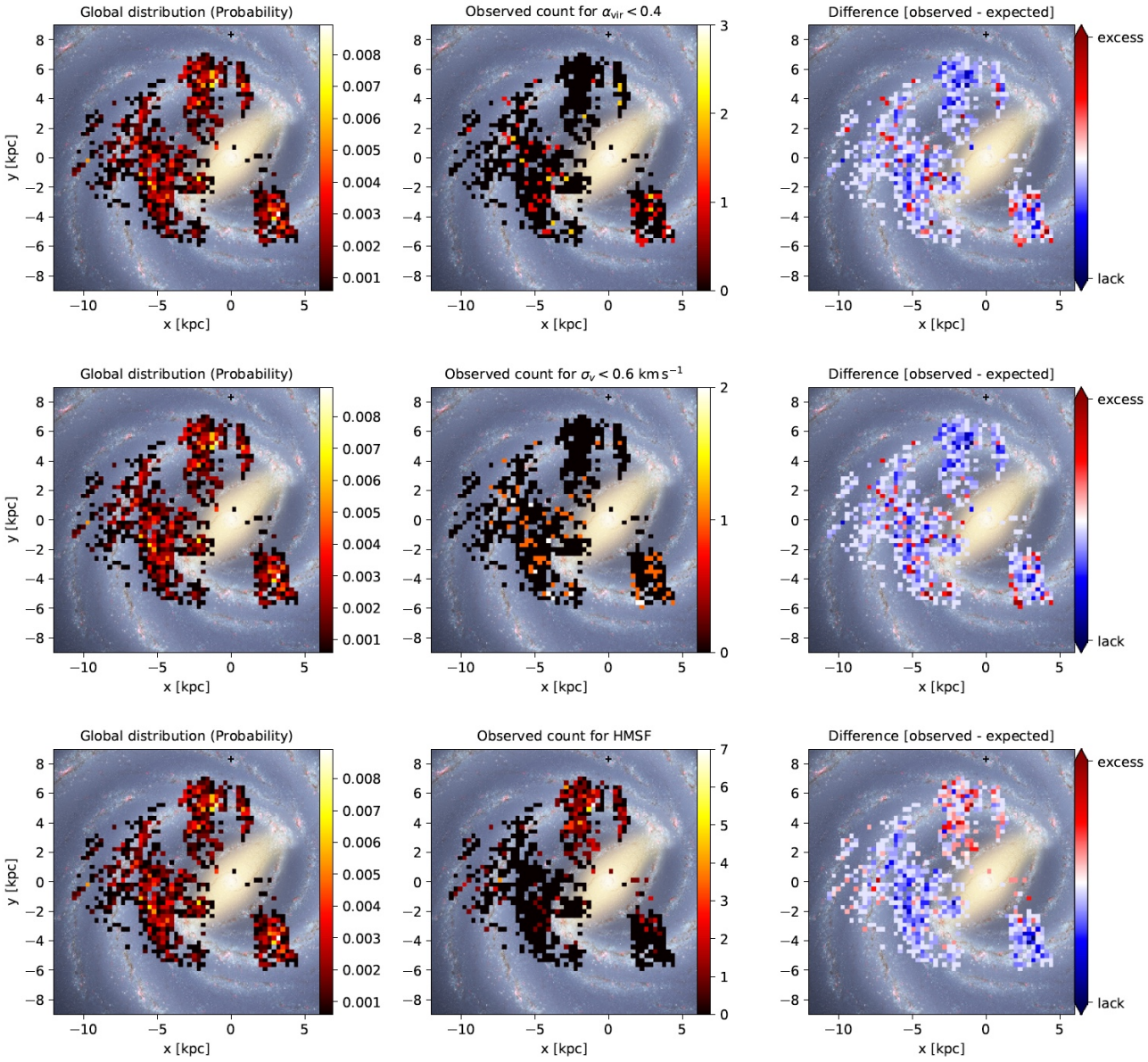

Figure F3. Same as Fig. F1 and F2, for the 100 clouds with lowest virial parameter (top) and velocity dispersion (middle), and also for the 211 clouds within the complete science sample that have a HMSF signpost (bottom).

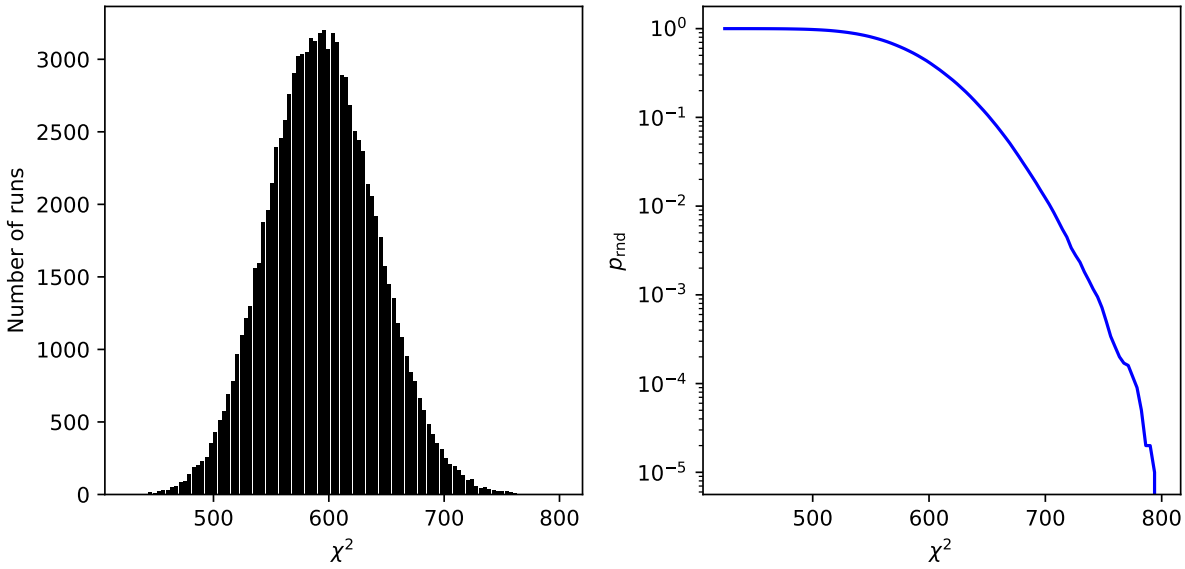

Figure F4. Left panel: Distribution of χ^2 values obtained as a result of performing our χ^2 statistical test on 100 000 random draws of 100 clouds from the complete science sample. Right panel: cumulative fraction of those runs (in log-scale) that had a χ^2 value above a certain value, i.e. the probability, $p_{\rm rnd}$, of observing χ^2-values above a specific value, purely from a random draw of clouds from our theoretical distribution.