Physical properties and real nature of massive clumps in the galaxy

Lu, Zu-Jia

2022-02

Lu , Z-J , Pelkonen , V-M , Juvela , M , Padoan , P , Haugbolle , T & Nordlund , A 2022 , ' Physical properties and real nature of massive clumps in the galaxy ' , Monthly Notices of the Royal Astronomical Society , vol. 510 , no. 2 , pp. 1697-1715 . https://doi.org/10.1093/mnras/stab3517

http://hdl.handle.net/10138/352555
https://doi.org/10.1093/mnras/stab3517

unspecified
publishedVersion

Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **510,** 1697–1715 (2022) https://doi.org/10.1093/mnras/stab3517
Advance Access publication 2021 December 3

Physical properties and real nature of massive clumps in the galaxy

Zu-Jia Lu [1]★ Veli-Matti Pelkonen [1] Mika Juvela [2] Paolo Padoan [1,3] Troels Haugbølle [4] and Åke Nordlund [4]

[1]*Institut de Ciències del Cosmos, Universitat de Barcelona, IEEC-UB, Martí i Franquès 1, E-08028 Barcelona, Spain*
[2]*Department of Physics, PO Box 64, University of Helsinki, 00014 Helsinki, Finland*
[3]*ICREA, Pg. Lluís Companys 23, E-08010 Barcelona, Spain*
[4]*Niels Bohr Institute, University of Copenhagen, Øster Voldgade 5-7, DK-1350 Copenhagen K, Denmark*

Accepted 2021 November 29. Received 2021 November 5; in original form 2021 July 6

ABSTRACT

Systematic surveys of massive clumps have been carried out to study the conditions leading to the formation of massive stars. These clumps are typically at large distances and unresolved, so their physical properties cannot be reliably derived from the observations alone. Numerical simulations are needed to interpret the observations. To this end, we generate synthetic Herschel observations using our large-scale star-formation simulation, where massive stars explode as supernovae driving the interstellar-medium turbulence. From the synthetic observations, we compile a catalogue of compact sources following the exact same procedure as for the Hi-GAL compact source catalogue. We show that the sources from the simulation have observational properties with statistical distributions consistent with the observations. By relating the compact sources from the synthetic observations to their 3D counterparts in the simulation, we find that the synthetic observations overestimate the clump masses by about an order of magnitude on average due to line-of-sight projection, and projection effects are likely to be even worse for Hi-GAL Inner Galaxy sources. We also find that a large fraction of sources classified as protostellar are likely to be starless, and propose a new method to partially discriminate between true and false protostellar sources.

Key words: MHD – radiative transfer – methods: numerical – catalogues – stars: formation.

1 INTRODUCTION

Massive stars are essential constituents of the ecosystem of galaxies, driving the thermodynamical and chemical evolution of their interstellar medium (ISM). Understanding their formation process is a prerequisite for modelling the evolution of galaxies and investigating the high-redshift universe. A complete theory of star formation is not available yet, massive-star formation being perhaps the main hurdle to overcome. An important limitation in the study of massive stars is the difficulty to test our theoretical models against observational data, because massive stars are more rare and shorter lived than low-mass stars. Regions of massive star formation tend to be at relatively large distances, obscured by high extinction levels, and confused by complex gas and stellar dynamics, as massive stars form in stellar clusters. Because of the complexity of such regions, their study can greatly benefit from the use of synthetic observations of realistic theoretical models.

Because of the complex nature of the dynamics of the ISM, synthetic and real observations of star-forming regions are better compared with statistical tools, which requires large samples. The 'Herschel Infrared Galactic Plane Survey' (Hi-GAL) has produced the largest catalogue to date of massive clumps (Molinari et al. 2016; Elia et al. 2017, 2021), usually viewed as potential progenitors of massive stars. Follow-up studies have characterized the dynamics of some of those clumps, including their infall rates (e.g. Traficante et al. 2017, 2018), providing important clues to the origin of massive stars. In this work, we compile the first synthetic catalogue of massive clumps that can be compared statistically to the Hi-GAL compact source catalogue, based on a star-formation simulation of a 250 pc region of the ISM driven by supernova (SN) explosions.

Using earlier evolutionary stages of the same simulation, we have previously shown that SNe alone can drive the observed turbulence in MCs (Padoan et al. 2016a; Pan et al. 2016) and can explain both the formation and dispersion of MCs (Lu et al. 2020). The star formation rate per free–fall time in the clouds was also found to be consistent with the observations (Padoan et al. 2017). The simulation was then used to study the formation of massive stars in Padoan et al. (2020), where it was also shown that observations could grossly overestimate the mass of protostellar cores, depending on distance and angular resolution. That study adopted a theoretician's perspective, by considering only the progenitor cores of massive stars, and only at the special moment when they have just started to collapse (the end of their pre-stellar phase).

In this work, we adopt an observer's perspective, by selecting compact sources from synthetic observations of individual simulation snapshots, following the same procedure as in Herschel's Hi-GAL compact source catalogue (Elia et al. 2017, 2021). This approach results in a very large catalogue of 51 831 synthetic sources (observing three simulation snapshots from three different directions and four distances), including both pre-stellar and protostellar ones (see Table 1), to be compared with the 22 932 sources from the Hi-

★ E-mail: luzujia@icc.ub.edu

Table 1. Number of sources, divided into starless and protostellar categories, and median values of diameter, mass, and temperature at different distances. (The total number of sources is 51 831, but becomes 107 453 when we normalize the numbers for distances >2 kpc to match the distance distribution in the Hi-GAL Inner Galaxy catalogue as described in Section 5).

	N	Diameter (pc)	Mass ($M_\odot$)	Temperature (K)
2 kpc	36 125	0.23	11.81	11.87
Starless	31 905	0.23	12.40	11.51
Protostellar	4220	0.23	7.24	16.84
4 kpc	11 330	0.45	41.19	12.68
Starless	9218	0.46	44.99	12.16
Protostellar	2112	0.43	25.75	17.17
8 kpc	3010	0.89	152.06	13.6
Starless	2100	0.90	179.57	12.59
Protostellar	910	0.84	104.59	17.41
12 kpc	1366	1.32	337.78	14.13
Starless	867	1.38	404.91	12.83
Protostellar	499	1.24	240.23	17.58

GAL catalogue of the Outer Galaxy (our main validation sample), and 78 325 sources of the Outer Galaxy catalogue (in the distance and temperature intervals considered for our comparison, $1.5 < d < 13.5$ kpc and $T < 40$ K). Our goal is twofold: to validate our synthetic catalogue through the comparison with real observations and, once validated, to use it for the interpretation of the observations. We show that the sources from the simulation have observational properties with statistical distributions consistent with the Outer Galaxy observations, while being systematically lower in surface density with respect to the Inner Galaxy sources, as expected. We then compare the compact sources from the synthetic observations to their 3D counterparts in the simulation. We find that the clump masses from the observations are generally overestimated due to line-of-sight projection and that a significant fraction of clumps classified as protostellar are likely to be starless.

The structure of the paper is as follows. In Section 2, we briefly summarize the numerical simulation. Radiative transfer and synthetic observations are presented in Section 3, while the procedure to compile the clump catalogue from the synthetic observations is described in Section 4. The observational properties of the synthetic clumps are presented in Section 5, where they are also compared with the corresponding properties of the clumps in the Hi-GAL catalogue. The synthetic sources are then compared with their 3D counterpart from the simulation in Section 6. Various implications of our results are discussed in Section 7, and the main conclusions are summarized in Section 8.

2 SIMULATION

This work is based on the same large-scale magnetohydrodynamic (MHD) simulation of star formation used in Padoan et al. (2017) to study the star-formation rate in molecular clouds, in Padoan et al. (2020) to study the formation of massive stars, and in Lu et al. (2020) to study the effect of SNe on the dispersion of molecular clouds (MCs). The simulation has been continuously run, during the past 2 yr, under a multiyear PRACE project, and will be run for another year, until it reaches ∼100 Myr of evolution. It describes an ISM region of size $L_{\rm box} = 250$ pc and total mass $M_{\rm box} = 1.9 \times 10^6\,{\rm M}_\odot$, where the turbulence is driven by SNe alone. The 3D MHD equations are solved with the AMR code RAMSES (Teyssier 2002; Fromang, Hennebelle & Teyssier 2006; Teyssier 2007), using periodic boundary conditions. We refer the reader to the papers cited above for details about the numerical setup. In the following, we briefly summarize only the main features relevant to this work.

The energy equation includes the pressure-volume work, the thermal energy introduced to model SN explosions, a uniform photoelectric heating as in Wolfire et al. (1995), with efficiency $\epsilon = 0.05$ and the FUV radiation field of Habing (1968) with coefficient $G_0 = 0.6$ (the UV shielding in MCs is approximated by tapering off the photoelectric heating exponentially above a number density of 200 cm^{-3}), and a tabulated optically thin cooling function constructed from the compilation by Gnedin & Hollon (2012) that includes all relevant atomic transitions. Molecular cooling is not included, due to the computational cost of solving the radiative transfer. The thermal balance between molecular cooling and cosmic ray heating in dense gas is emulated by setting a limit of 10 K as the lowest temperature of dense gas. However, to generate synthetic observations of the dust emission, the radiative transfer is computed post-processing individual snapshots, including all stars with mass $> 2\,{\rm M}_\odot$ as point sources (see Section 3).

The initial conditions of the simulation are zero velocity, uniform density, $n_{\rm H,0} = 5$ cm^{-3}, uniform temperature, $T_0 = 10^4$ K, and uniform magnetic field, $B_0 = 4.6\,\mu$G. During the first 45 Myr, self-gravity was not included and SN explosions were randomly distributed in space and time, at a rate of 6.25 SNe Myr^{-1}. The resolution was $dx = 0.24$ pc, achieved with a 128^3 root grid and three AMR levels. The minimum cell size was then decreased to $dx = 0.03$ pc, using a root-grid of 512^3 cells and four AMR levels, for an additional period of 10.5 Myr, still without self-gravity. At $t = 55.5$ Myr, gravity is introduced and the minimum cell size is further reduced to $dx = 0.0076$ pc by adding two more AMR levels. This resolution allows us to resolve the formation of individual massive stars, so the time and location of SNe are computed self-consistently from the evolution of the massive stars.

Individual stars are modelled with accreting sink particles, created when the gas density is larger than 10^6 cm^{-3} and other conditions are satisfied (see Haugbølle, Padoan & Nordlund 2018, for details of the sink particle model). A SN is created when a sink particle of mass larger than 7.5 $M_\odot$ has an age equal to the corresponding stellar lifetime for that mass (Schaller et al. 1992). The sink particle is removed and the stellar mass, momentum, and 10^{51} erg of thermal energy are added to the grid with a Gaussian profile (see Padoan et al. 2016a, for further details). By the last simulation snapshot used in this work, corresponding to a time of 34.2 Myr from the inclusion of self-gravity and star formation, 4400 stars with *final* mass $> 2\,{\rm M}_\odot$ have been generated. By *final* mass, we mean the mass they have achieved at the end of the simulation, 4 Myr later (at 38.2 Myr from the beginning of star formation), when the great majority of the stars have accreted most of their mass, so the stellar IMF is not expected to further evolve significantly (Haugbølle et al. 2018; Padoan et al. 2020).

Fig. 1 shows the IMF of those stars (including ∼538 that have already exploded as SNe), based on their final masses. The IMF is approximately a power law for massive stars, with a slope of $\Gamma = -1.8$ for masses between 7.5 and 50 $M_\odot$, but incomplete at lower masses (the power law should extend down to ∼1 $M_\odot$). The spatial resolution of the simulation was chosen to resolve the formation of massive stars, in order to model realistically the SN feedback. The apparent IMF cutoff above ∼50 $M_\odot$ is consistent with the predicted maximum stellar mass for the simulation. The reader is referred to Padoan et al. (2020, §3.1, 8.5, 8.6, 8.7) for a detailed discussion of

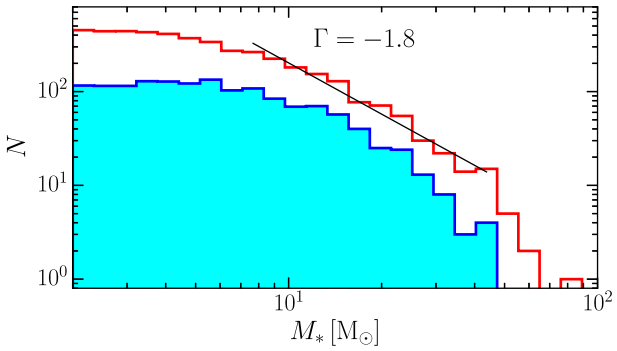

Figure 1. Mass distribution of the stars present in the last of the three snapshots analysed in this work, based on their *final* mass (red histogram). The power-law fit (black line), giving a slope of $\Gamma = -1.8$, is computed between 7.5 and 50 M$_\odot$. The cyan-shaded histogram includes only the stars that were still embedded in dense gas ($n > 10^3$ cm^{-3}) in the three snapshots used in this work.

the IMF in the simulation. In that work, we found that the IMF slope was $\Gamma = -1.5$, where the IMF was computed at ∼25 Myr, slightly steeper than Salpeter's value (Salpeter 1955), but consistent with the result of a study of many stellar clusters in M31, covering a similar range of stellar masses as in our simulation (Weisz et al. 2015). The apparent variation of the IMF slope over time will be addressed in a separate work. To estimate the number of stellar progenitors present in the three snapshots used in this work, we consider all the stars in the IMF of Fig. 1, and select those that were still embedded in dense gas ($n > 10^3$ cm^{-3}, as later explained in Section 4.4) in those snapshots. The mass distribution of these stars is shown by the cyan-shaded histogram in Fig. 1. Of a total of 1142 stars more massive than 7.5 M$_\odot$, 468 were still embedded in the three snapshots analysed in this work, and thus represent the sample of massive-star progenitors relevant for our synthetic catalogue.

3 SYNTHETIC OBSERVATIONS

To compute synthetic dust continuum maps, we select three snapshots of our simulation at times 15.4, 23.3, and 34.2 Myr from the beginning of self-gravity and star formation. These three snapshots were chosen to sample different conditions in the star-formation history of the simulation. The first and the third times correspond to periods of relatively low star-formation rate, while the middle time corresponds to a peak in the star-formation rate. The first snapshot has a relatively low star-formation efficiency, as less of the initial gas mass has gone to forming stars by that early time, whereas the star-formation efficiency has increased significantly by the last snapshot, as also evidenced by the number of stars in each snapshot mentioned below. We then calculate synthetic observations in Herschel's bands (70, 160, 250, 350, and 500 μm) to compare our results with Herschel's Hi-GAL compact-source catalogue (Elia et al. 2017, 2021).

The surface brightness maps were computed with the continuum radiative transfer program SOC (Juvela 2019). The spatial discretization uses the full density information from the MHD run, with a root grid of 512^3 cells and six levels of refinement in the octree hierarchy. For the dust properties, we tested both the diffuse-medium dust model of Compiègne et al. (2011) and the dust model of Ossenkopf & Henning (1994) that is more appropriate for dense medium. The final calculations were all carried out using the latter, which corresponds to dust grains with thin ice mantles, after 10^5 yr of coagulation at a density of 10^6 cm^{-3}. This model yields a value of $\beta \approx 1.8$ for the exponent of the power-law dust emissivity. For the radiation that enters the model from the outside, we used the values for the normal local interstellar radiation field (Mathis, Mezger & Panagia 1983).

As internal sources for the radiative transfer calculations, all stars with masses $> 2\,M_\odot$ at the time of the snapshot that have not exploded as SNe yet, were included as point sources. Their luminosities were derived from the Zero Age Main Sequence mass–luminosity relations (Duric 2004; Salaris & Cassisi 2005), and their blackbody spectra from calculating the effective temperature using the above luminosity and the mass–radius relations in Kippenhahn & Weigert (1994). The number of stellar sources were 909, 2431, and 3868 for the three snapshots at times 15.4, 23.3, and 34.2 Myr respectively.

SOC was used to calculate the equilibrium dust temperature for each model cell and, based on that information, the surface brightness maps at the Herschel frequencies. The surface brightness maps are resampled to the same pixel sizes as the Hi-GAL maps: 3.2, 4.5, 6.0, 8.0, and 11.5 arcsec, for the five bands in the order of increasing wavelength. At each wavelength, the full width at half maximum (FWHM) values of the adopted Gaussian telescope beams are three pixels, giving 9.6, 13.5, 18.0, 24.0, and 34.5 arcsec. We added observational noise to the maps so that, after beam convolution appropriate for the assumed distances (2, 4, 8, and 12 kpc), the noise was consistent with actual observations. For the surface brightness relative noise values, we assumed 4 per cent in the PACS bands (70 and 160 μm) and 2 per cent in the SPIRE bands (250, 350, and 500 μm). We also added additional noise with absolute levels of 7.8, 6.0, 0.81, 0.42, and 0.28 MJy sr^{-1}, to the five bands in the growing wavelength order. This noise was estimated by extracting small 100 arcsec by 100 arcsec submaps along overlap regions of two individual Hi-GAL tiles at Galactic longitudes $l = 49°$, $89°$, and $144°$. The paired submaps of the same region were used to calculate the rms of the surface brightness difference of each pixel. The rms noise of the low surface brightness submaps was consistent regardless of the region and was adopted as the absolute noise level in the synthetic observations. SOC is based on the Monte Carlo method, which also contributes to the noise in the maps. However, the number of simulated photon packages was chosen to be large enough so that the Monte Carlo noise is a few times below the observational noise.

For each of the three snapshots, the surface brightness maps were computed along three different and orthogonal directions, and assuming four different distances of 2, 4, 8, and 12 kpc. This resulted in 36 maps at each wavelength: three snapshots seen from three orthogonal directions and at four assumed distances each. One of these 36 maps is shown in Fig. 2 as a three-colour image of the whole 250 pc volume. The image is made by using the 70, 160, and 250 μm maps for blue, green, and red colours, respectively. The lower panels of Fig. 2 show three-colour images of two dense regions of 25 and 50 pc size, hosting the formation of massive stars. All the maps and the three-colour images can be found at https://www.erda.dk/vgrid/massive-clumps/.

4 SYNTHETIC COMPACT SOURCE CATALOGUE

Using our synthetic Herschel observations, we compile a synthetic catalogue of compact sources extracted with the CUrvature Thresholding EXtractor (CuTEx) code (Molinari et al. 2011), with the same

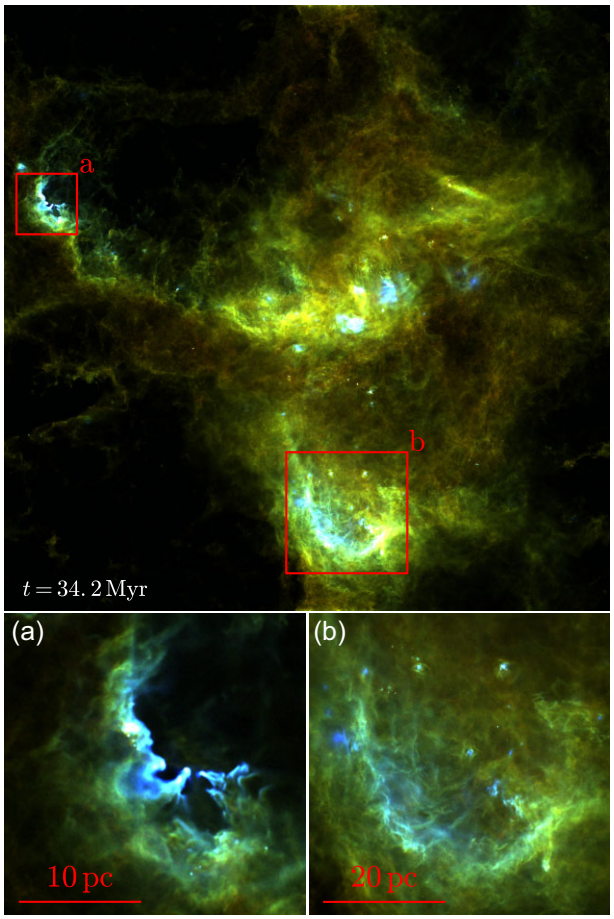

Figure 2. Upper panel: simulated Herschel's three-colour image of the whole 250 pc simulation at 34.2 Myr after the start of star formation, assuming a distance of 2 kpc, generated from single-band images at 70, 160, and 250 μm for blue, green, and red, respectively. The range of the linear colour scale is from 0 to 600 MJy sr^{-1}. The corresponding images for all directions and all three simulation snapshots can be found at https://www.erda.dk/vgrid/massive-clumps/. The two red boxes mark the positions of the two zoom-in regions shown in the lower panels. Lower panels: the 25 pc (a) and 50 pc (b) zoom-in regions marked in the upper panel. The linear colour scale of these zoom-in regions is from 0 to 1200 MJy sr^{-1}.

exact method used to generate the Hi-GAL compact source catalogue (Elia et al. 2017, 2021). All the steps of the source extraction are described in this section. The resulting catalogue of synthetic sources is described in Appendix A, and the number and median properties of the sources are summarized in Table 1.

4.1 CuTEx

CuTEx is a detection and extraction code originally written in IDL and now also available in GDL. It is presented in detail in Molinari et al. (2011), and summarized briefly here. CuTEx finds compact objects by calculating second-order derivative maps, ∂^2, in four directions: along each axis and along the diagonals. The pixels that exceed a 'curvature' threshold value, ζ_{th}, in all four maps are masked as possible sources, which are then clumped together into contiguous clusters. Molinari et al. (2011) report that using $\zeta_{\rm th} \geq 0.5\sigma_{\partial^2}$, where σ_{∂^2} is the rms in each particular ∂^2 map, the minimum number of contiguous pixels for reliable source detection is 3. The tentative location of the source is determined by finding a local maximum pixel that is $1\sigma_{\partial^2}$ above the neighbouring pixels in a map that is averaged from all four ∂^2 maps. If more local maxima pixels are found within the cluster, these become tentative other sources. If no statistically significant local maxima are found, the location is calculated as a mean of all the pixel coordinates in the cluster.

Once the source is located, the shape of the source is assumed to be a Gaussian. The size and the orientation is estimated from the local minima of the curvature around the source, using each of the four ∂^2 maps and measured along the gradient direction to find the minima before and after the source location, resulting in eight measurements of the minima. If the before and after minima in each direction agree within 20 per cent, both are kept; otherwise only the minimum nearest to the source location is kept. An ellipse is fitted to the minima, and semimajor axis, semiminor axis, and position angle are recorded. Finally, if the semimajor axis is larger than three times the point spread function (PSF) of the observations, the size estimate is flagged as uncertain and set back to one PSF. This limits the size range of the initial guess from one PSF to three PSFs.

The source photometry is estimated by fitting a 2D Gaussian shape of variable intensity, size and orientation, based on the initial guess on size and orientation. The background is estimated with a planar surface of a variable inclination and inclination direction, and is fit simultaneously with the source Gaussian, in a fitting area that is typically four times the PSF of the observations. If the source is in a crowded field, it and its nearest neighbours (typically within twice the PSF) are fitted simultaneously, although only the central source parameters are saved. The other neighbouring sources are fitted in their turn, based on their local background and neighbours. In this case of a crowded field, the fitting area is the minimum area that covers all the sources with an excess of one PSF around them. The Gaussian fitting routine is MPFIT (Markwardt 2009), which allows for the simultaneous adjusting of all source positions in a group, as well as varying the initial guesses on the source sizes by up to 30 per cent. If the initial guess was uncertain, the size is constrained only by the limits imposed by the photometric routine, which allows it to vary from a minimum of 0.95 PSF to the maximum of 3.9 PSFs, which can be achieved if the initial guess was already three PSFs. The photometric ASCII output file includes the source position, size, orientation, the integrated, peak, and background fluxes, as well as uncertainties for all the photometric parameters. In addition, CuTEx creates an SAO Image region file of the detected source ellipses, which can be easily overplotted on an image, with examples shown in Fig. 3.

In this study, we use the same parameters for CuTEx as Molinari et al. (2016), with an extraction threshold $\zeta_{\rm th} = 2\sigma_{\partial^2}$. Our synthetic maps are already resampled to the pixel resolution of three pixels per PSF, varying with the wavelength, similarly to the actual Herschel maps used by Molinari et al. (2016). We then perform a CuTEx detection at each of the five wavelength for the three selected snapshots seen from three orthogonal directions, and at 2, 4, 8, and 12 kpc distances. This results in 180 detection catalogues, which are used as inputs for the CuTEx photometric extraction routine to derive the photometry for the detected sources. In the end of the CuTEx detection and photometry, we have 180 single-band source catalogues for 36 different combinations of snapshots, viewing directions and distances.

4.2 Band-merging and final source selection

As in Elia et al. (2017), we band-merge the single-band catalogues, obtaining 36 multiwavelength catalogues, one for each unique

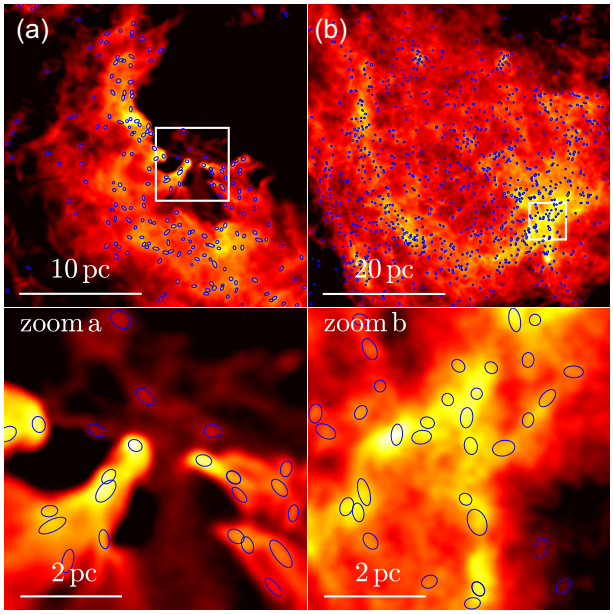

Figure 3. Upper panels: 250 μm surface brightness of the two zoom-in regions from Fig. 2 (a: 25 pc and b: 50 pc). The colour scale is logarithmic with minimal values set equal to 1/32 (a) and 1/8 (b) of the maximum surface brightness in each region. The blue ellipses are the clumps detected by CuTEx at 250 μm, before band-merging. Lower panels: additional 6 pc zoom-in regions corresponding to the white boxes in the upper panels. The logarithmic colour scale has the minimum values set at 1/32 (zoom a) and 1/3 (zoom b) of the maximum surface brightness in each zoom-in.

combination of snapshot, direction, and distance, using the following procedure:

(1) Starting from the 500 μm sources, we seek detections at 350 μm that are a positional match by taking the centre of the 350 μm sources, and checking which ones fall inside the circularized size of the 500 μm sources.

(2) We repeat the search for each lower wavelength by searching which 250 μm sources fall inside the circularized size of the 350 μm sources, which 160 μm sources fall inside the 250 μm sources, and finally which 70 μm sources are found inside the 160 μm sources.

(3) We select and retain only the band-merged sources that are detected in at least three consecutive Herschel bands (except for 70 μm band), meaning at 160-250-350 μm, at 250-350-500 μm, or at 160-250-350-500 μm, and without a dip in the spectral energy distribution (SED) between adjacent wavelengths or a peak at 500 μm (Giannini et al. 2012).

The circularized diameter of a source at each waveband is calculated as $\mathrm{FWHM}_{\mathrm{circ},\lambda} = \sqrt{\mathrm{FWHM}_{\mathrm{maj},\lambda} \times \mathrm{FWHM}_{\mathrm{min},\lambda}}$, where $\mathrm{FWHM}_{\mathrm{maj},\lambda}$ and $\mathrm{FWHM}_{\mathrm{min},\lambda}$ are the semimajor and the semiminor axis, respectively, of the ellipse of the source estimated by CuTEx. The beam-deconvolved diameter at each wavelength is estimated as $\theta_\lambda = \sqrt{\mathrm{FWHM}^2_{\mathrm{circ},\lambda} - \mathrm{HPBM}^2_\lambda}$, where HPBM_λ is the beam size at the given wavelength. However, if $\mathrm{FWHM}_{\mathrm{circ},\lambda} \leq \mathrm{HPBM}_\lambda$, we do not deconvolve. The fluxes at the wavelength $\lambda = 350$ and 500 μm are then scaled by the ratio between the deconvolved sizes at those wavelengths and at 250 μm,

$$\overline{F}_\lambda = F_\lambda \times \frac{\theta_{250}}{\theta_\lambda}, \quad \lambda \geq 350\ \mu\mathrm{m}, \tag{1}$$

if $\theta_\lambda > \theta_{250}$, $\mathrm{FWHM}_{\mathrm{circ},\lambda} > \mathrm{HPBM}_\lambda$, and $\mathrm{FWHM}_{\mathrm{circ},250} > \mathrm{HPBM}_{250}$ (Elia et al. 2013, 2017). If these conditions are not fulfilled, then we do not perform any flux scaling at that wavelength. The 250 μm wavelength is taken as the reference wavelength because it is the shortest one that is always present when it is imposed that the sources have detections for at least three consecutive wavelengths. Using that reference wavelength, the source diameter, D, in the catalogue is either the beam-deconvolved diameter, $D = \theta_{250}$, if $\mathrm{FWHM}_{\mathrm{circ},250} > \mathrm{HPBM}_{250}$, or the circularized diameter, $D = \mathrm{FWHM}_{\mathrm{circ},250}$, if $\mathrm{FWHM}_{\mathrm{circ},250} < \mathrm{HPBM}_{250}$. The error of the fluxes are estimated by the product of the area of the circularized clump size and the the RMS of the residual pixel fluxes after the subtraction of the best fit to the initial data by CuTEx.

4.3 SED Fitting

As in Elia et al. (2017), we use a single greybody function to fit the $\lambda \geq 160$ μm SEDs from the CuTEx photometric output to estimate the temperature of the sources, T, and thus derive their masses, M_{SED},

$$F_\nu = (1 - \mathrm{e}^{-\tau_\nu})\, B_\nu(T)\, \Omega\,, \tag{2}$$

where F_ν is the observed flux density at the frequency ν, τ_ν is optical depth of the medium, $B_\nu(T)$ is the Planck's function at the dust temperature T, and Ω is the source solid angle in the sky. The optical depth is parametrized as:

$$\tau_\nu = \left(\frac{\nu}{\nu_0}\right)^\beta, \tag{3}$$

where the parameter $\nu_0 = c/\lambda_0$ is such that $\tau_{\nu_0} = 1$ and β is the exponent of the power-law dust emissivity at large wavelengths. While Elia et al. (2017) adopt the value $\beta = 2$, we use $\beta = 1.8$, which is more appropriate for the dust model of Ossenkopf & Henning (1994) that we have used in the radiative transfer. By keeping self-consistency between the dust model and the SED fitting, we avoid the introduction of artificial discrepancies between the mass in the model and that deduced from the SED fitting. Ω is taken to be equal to the surface area measured by CuTEx at 250 μm, and we find the values of T and λ_0 from a least-square-fit of the SED using equation (2), within the ranges 5 K$\leq T \leq$ 40 K and 5 μm$\leq\lambda_0 \leq$ 350 μm, as in Elia et al. (2017). For a small number of sources the best-fitting temperature would be larger than 40 K, but it is forced to be equal to 40 K as that is the maximum value of the temperature range adopted for the fit. To avoid uncertainties related to these sources, we drop all sources with $T = 40$ K from both the Hi-GAL catalogue (396 sources) and the synthetic catalogue (311 sources).

The mass is then derived as

$$M_{\mathrm{SED}} = (d^2\Omega/\kappa_{\mathrm{ref}})\tau_{\mathrm{ref}}\,, \tag{4}$$

where κ_{ref} is the opacity estimated at a reference frequency ν_{ref}. As in Elia et al. (2017), we adopt $\lambda_{\mathrm{ref}} = 300$ μm and $\kappa_{\mathrm{ref}} = 0.1$ cm^2 g^{-1}. This value of the opacity is only ≈30 per cent lower than in the dust model of Ossenkopf & Henning (1994).

If $\lambda_0 \leq 44.5$ μm, then $\tau \leq 0.1$ at 160 μm, and the SEDs are instead fitted with the optically thin expression for the flux density,

$$F_\nu = \frac{M_{\mathrm{SED}}\,\kappa_{\mathrm{ref}}}{d^2}\left(\frac{\nu}{\nu_{\mathrm{ref}}}\right)^\beta B_\nu(T)\,, \tag{5}$$

which gives directly the values of both T and M_{SED}.

4.4 Source classification

The 70 μm flux is not used in the SED fitting procedure, but only to classify the sources. Following Elia et al. (2017), we classify the synthetic compact sources in our catalogue as protostellar if they

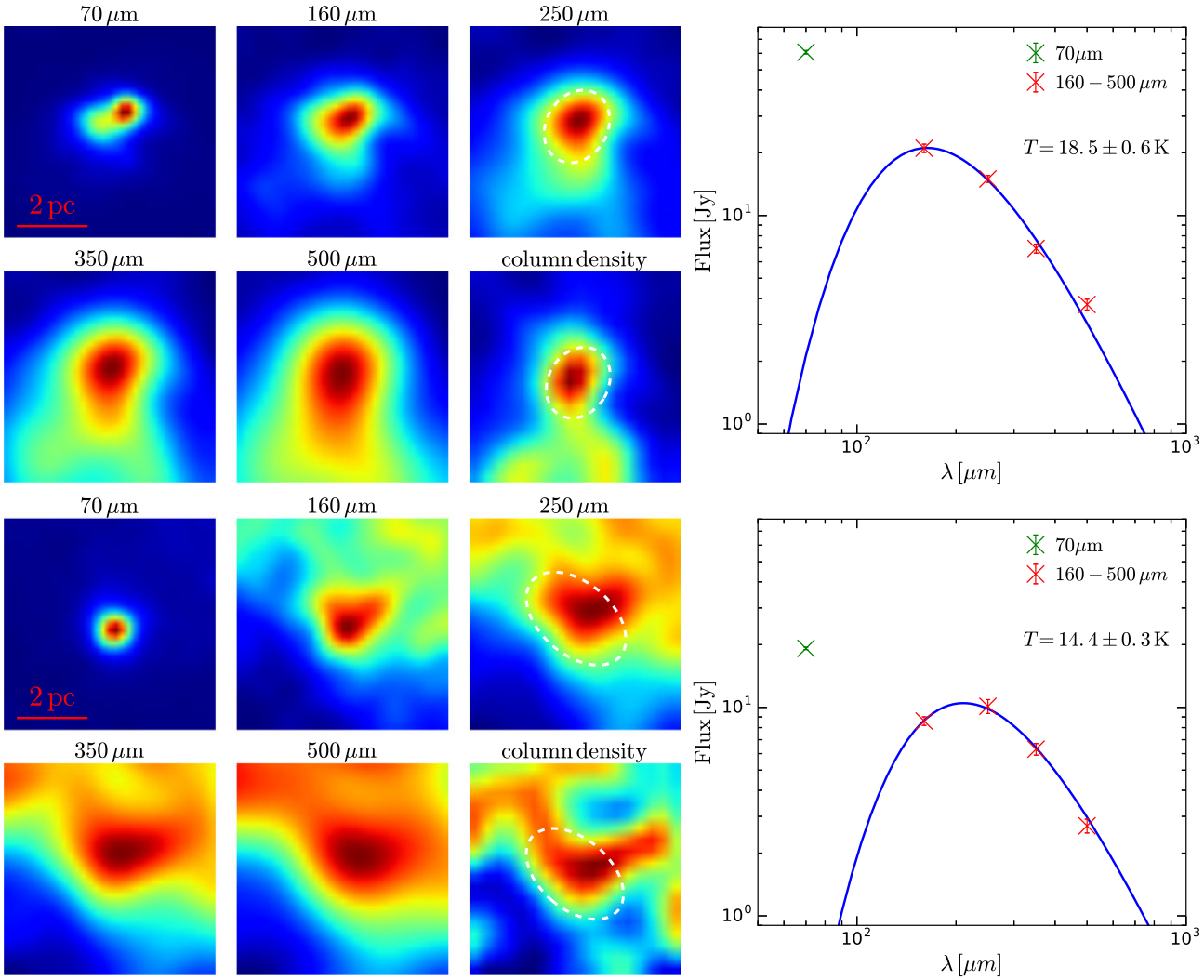

Figure 4. Surface brightness at all five Herschel's bands and column density, in maps of 6 pc size, of two examples of protostellar sources assumed to be at a distance of 12 kpc. The dashed white ellipse is the CuTEx source detected at 250 μm. The corresponding SEDs are shown in the right panels, where the blue lines are the best fits. The 70 μm flux (green cross symbol), detected in both sources, is not included in the SED fitting.

have a CuTEx counterpart at 70 μm, or starless if they do not. The number of the classified synthetic sources, and the median values of their diameters, masses, and temperatures at the different distances are summarized in Table 1.

In addition, we check if our protostellar sources have actual protostars within them. This requires us to discriminate between embedded stars and other stars that may be in the line of sight by pure chance, rather than by birth. For that purpose, we adopt a simple density threshold, such that a star is considered to be embedded if the mean density in a $(0.12\,\mathrm{pc})^3$ cell around it is larger than $10^3\,\mathrm{cm}^{-3}$ (see Appendix B for details). Among all the protostellar sources in our catalogue, we then define as true protostellar sources those with at least one embedded star in their line of sight, and as false protostellar sources the ones without embedded stars, irrespective of their 70 μm flux.

Single waveband images, column density maps (at the resolution of the 250 μm observations) and the corresponding SEDs are shown in Fig. 4 for two example sources at 12 kpc distance. As shown by the right-hand panels of Fig. 4, both sources have a strong 70 μm excess, so they are classified as protostellar. As discussed later in Sections 6.1 and 6.2, the source in the upper panels is an example of a true protostellar clump, while the one in the lower panels is primarily a projection effect and a false protostellar source.

4.5 Source duplication

One of the goals of this work is to relate the compact sources from the synthetic observations to their 3D counterparts in the simulation. Because each of the three snapshots of the simulation is observed from three different directions and at four different distances, it is possible that some of our synthetic sources correspond to the same 3D clumps. Thus, although the catalogue contains a total of 51 831 synthetic sources, only a fraction of those are truly independent, meaning that they do not correspond to the same 3D clumps.

In Section 6.1, we will define the real 3D clump that corresponds to a 2D source as the region around the maximum density peak, along the line of sight of that source. Based on this definition, we compute the number of sources whose corresponding 3D clumps do not overlap significantly. Fig. 5 shows this number of independent sources as a function of the maximum fractional volume overlap, $f_{\mathrm{vol,\,max}}$, defined in the following. For each pair of sources, we consider

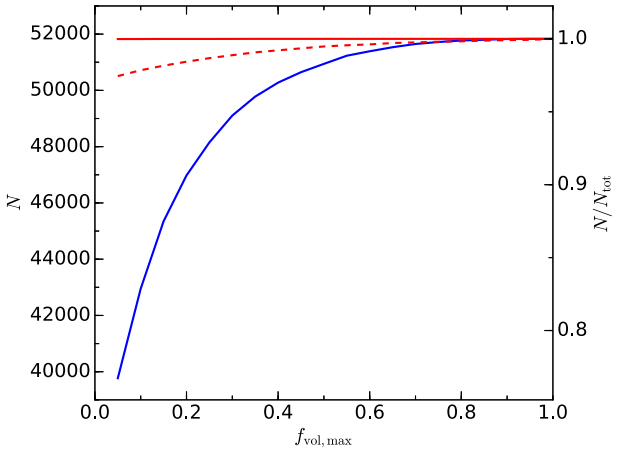

Figure 5. Number (left y-axis) and fraction (right y-axis) of independent synthetic sources as a function of the fractional volume overlap between the corresponding main 3D clumps (solid lines) or between a clump and a line-of-sight column (dashed line), as explained in Section 4.5. The blue line is computed by searching for duplicate sources using pairs from different directions and distances, while the red curves do not exclude pairs if the two sources are at different distances.

the pair of their corresponding 3D clumps, and we measure the volume of the two clumps, V_1 and V_2, and the volume of their intersection, $V_{\rm in}$. We then compute the two volume ratios, $V_{\rm in}/V_1$ and $V_{\rm in}/V_2$, and define the fractional volume overlap as the smaller of the two ratios, $f_{\rm vol} \equiv \min(V_{\rm in}/V_1, V_{\rm in}/V_2)$. Two synthetic sources are considered as independent if their fractional volume overlap is smaller than a chosen value, $f_{\rm vol,\,max}$, which is the quantity in the abscissa of Fig. 5.

The blue line shows the number (left y-axis), or fraction (right y-axis), of independent sources when comparing sources detected both from different directions and from different distances, and the red line shows the fraction of independent sources considering duplication only between different directions, that is assuming that sources at different distances are always independent. The red line is essentially a constant equal to nearly the total number of sources in the catalogue, even for very small values of $f_{\rm vol,\,max}$, showing that a main 3D clump is almost never found as the main clump from more than one direction. Thus, source duplication of the same 3D clump occurs only between synthetic maps from the same data cube at different distances, but in the same direction. However, even in that case, the fraction of independent sources remains very large. For example, even considering a rather small volume overlap of 20 per cent, $f_{\rm vol,\,max} = 0.2$, the fraction of independent sources is still larger than 90 per cent.

The fact that source duplication from different directions is rare in the synthetic catalogue illustrates the difficulty of identifying real 3D clumps from the 2D projections of the observations. To further investigate if a main 3D clump detected in one direction contributes at least partially to sources selected from the two orthogonal directions, we also compute the fractional volume of the intersection of the clump with any of the columns whose projections correspond to the synthetic sources. In other words, we test if the clump is found in the line of sight of a deconvolved source, even if it is not the main clump for that source. If that is the case, the source corresponding to that clump is not considered independent. The result is shown by the red dashed line in Fig. 5. Although the number of independent sources decreases slightly with decreasing $f_{\rm vol,\,max}$, in the great majority of cases the main 3D clumps do not contribute to other sources detected from different directions.

5 OBSERVATIONAL PROPERTIES OF SYNTHETIC SOURCES

In this section, we compute the observational properties of the compact sources from our synthetic observations and compare them with those of the sources in Hershel's Hi-GAL catalogue (Elia et al. 2021) to validate our synthetic compact source catalogue. We first remove Hi-GAL sources that have temperatures of 40 K, the maximum value allowed in the Hi-GAL SED fitting, as we have done with our sources (see Section 4.3). Furthermore, we split the Hi-GAL catalogue into the Outer Galaxy, defined by the Galactic longitude range 67°–289°,[1] and the Inner Galaxy, defined by 0°–67° and 289°–360° longitude. The temperature cut and longitude division result in 32 691 sources in the Outer Galaxy and 86 896 sources in the Inner Galaxy. Finally, since our synthetic catalogue has only distances from 2 to 12 kpc, we apply a distance cut to both samples, taking only sources with valid distances between 1.5 and 13.5 kpc, resulting in a final sample of 22 932 and 78 325 sources for the Outer and Inner Galaxy, respectively, comparable to the sample size of our synthetic catalogue, containing 51 831 sources.

Although the star-formation rate in the simulation is consistent with both the Kennicutt–Schmidt relation, globally, and with the observed star-formation rate in molecular clouds, at smaller scales (Padoan, Haugbølle & Nordlund 2012), its size, 250 pc, and mean column density, 30 $\rm M_\odot pc^{-2}$, makes the synthetic observations more suitable for comparison with a single spiral arm than with the entire Galactic plane (e.g. the column density of the Perseus arm is estimated to be $\sim$23 $\rm M_\odot\,pc^{-2}$ (Heyer & Terebey 1998)). In the case of the Outer Galaxy, a large fraction of sources are located in the outer Perseus arm. Therefore, projection effects should not be much stronger than in our simulation, as the number of sources drops almost exponentially with increasing distance, similarly to the source counts in the synthetic catalogue (see Table 1). In Padoan et al. (2016b) it was already found, through synthetic CO observations, that molecular clouds extracted from the simulation are consistent with those from the FCRAO CO Survey of the Outer Galaxy (Heyer et al. 1998). Thus, we validate our synthetic catalogue against the Outer Galaxy sample of the Hi-GAL catalogue.

Besides the comparison with the Outer Galaxy, we also carry out a parallel comparison with the Hi-GAL catalogue of the Inner Galaxy, in order to highlight potential differences arising from the much larger depth of the observations. The number of sources in the Inner Galaxy is nearly constant between $\sim$1.5 and $\sim$13.5 kpc, so projection effects are expected to be significantly enhanced, in comparison to the Outer Galaxy. Thus, in the case of the comparison with the Inner Galaxy, the number of the synthetic sources in the histograms and scatter plots is multiplied by a factor $(d/2\,{\rm kpc})^2/2$, for distances $d >$ 2 kpc (and left unchanged for sources at 2 kpc), to approximate the distance distribution of sources in the Inner Galaxy sample. This is necessary to be able to compare the probability distributions of source properties, as sizes and masses increase on average with distance, as shown below.

[1]Elia et al. (2021) adopted a galactocentric definition of the Inner/Outer Galaxy, separated by the Solar circle. However, due to our desire to avoid strong projection effects in the Outer Galaxy sample, we use the longitudinal definition of Inner/Outer Galaxy as in Elia et al. (2017).

5.1 Protostellar fraction

In the comparison with the Hi-GAL catalogue samples, we distinguish between protostellar and starless sources. The fraction of protostellar sources, as a function of distance, is shown in the top panels of Fig. 6 for both the synthetic catalogue (solid line) and the Hi-GAL catalogues (dashed lines). In our catalogue, the fraction increases slightly with increasing distance at all distances, while in the Hi-GAL catalogues it remains approximately constant, except for a slight increase in the Inner Galaxy catalogue (top right panel) for distances larger than approximately 7 kpc.

Two competing effects contribute to the distance dependence of the protostellar fraction. On the one hand, at larger distances neighbouring sources may blend together. If a blended source consists of a starless and a protostellar source, it is marked as protostellar due to the detected 70 μm emission, with the effect of increasing the protostellar fraction with increasing distance. On the other hand, protostars of increasingly smaller mass and luminosity can be detected with decreasing distance, which tends to increase the protostellar fraction toward smaller distances and to cancel the effect of blending. While both effects are present in the observations, explaining the approximately constant protostellar fraction, the second one is not fully captured in the synthetic observations. Although we include as radiation sources for all the stars down to 2 $M_\odot$, the stellar IMF in our simulation is incomplete below $\sim$ 8 $M_\odot$. Thus, we underestimate the relative number of low-mass stars that may be potentially detected as low luminosity protostars at small distances, explaining the distance dependence of the protostellar fraction in the synthetic catalogue.

The protostellar fraction in the synthetic catalogue is very close to that of the Outer Galaxy at intermediate distances ($\sim$6–9 kpc), with a value of $\sim$0.3. It is also the same as that of the Inner Galaxy at 6 kpc.

5.2 Size, mass, and temperature distributions

The three bottom rows of panels of Fig. 6 show the distance dependence of the source diameters, masses, and temperatures as error-bar plots for the synthetic sources at distances 2, 4, 8, and 12 kpc, and as shaded areas for the Hi-GAL sources in bins of 1 kpc between 0.5 and 13.5 kpc. The red and blue error bars give the median and the dispersion of the diameters, masses, or temperatures of our starless and protostellar sources, respectively. The dispersion is estimated as the interval between the 2.5 and 97.5 percentiles. The red- and blue-dashed lines and shaded areas give the median and the dispersion for the starless and protostellar sources in the Hi-GAL catalogues.

Fig. 6 shows that our synthetic observations reproduce well the range of clump sizes found in Herschel's sources and their distance dependence, as well as the lower median values of the diameters of the protostellar clumps relative to that of the starless ones. Although the range of source sizes at a fixed distance is a direct result of the CuTEx detection procedure, not a good measure of the match between the simulation and the observations, the median value of the sizes is sensitive to the size distribution within that fixed range, which is not set by CuTEx.

In the middle panels of Fig. 6, we show the masses of the clumps from the synthetic and the Hi-GAL catalogues. In general, the synthetic catalogue reproduces well the median values of the observed masses as a function of distance, and their dispersion, in the case of the Outer Galaxy. The masses of protostellar sources at 2 and 4 kpc are underestimated by a factor of 2–3, which is proposed to be related to the temperature discrepancy due to the incompleteness

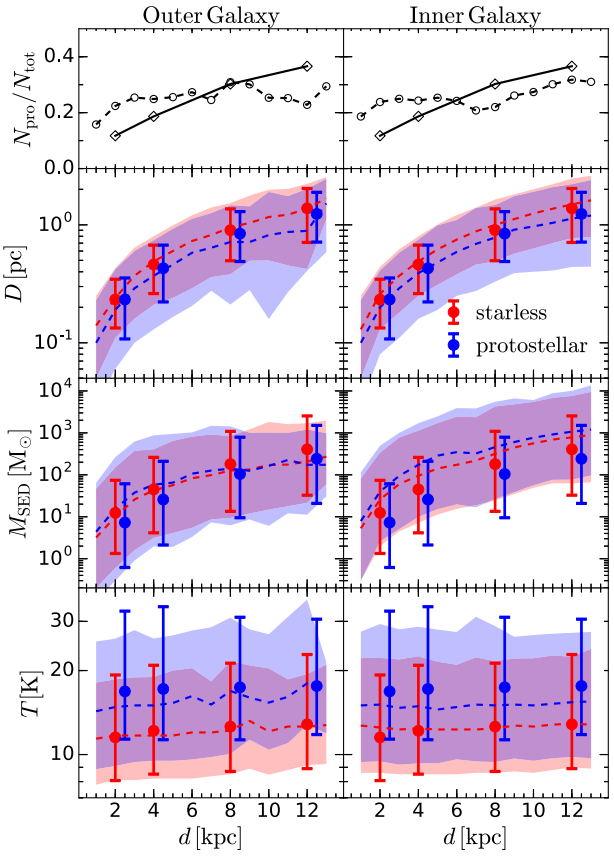

Figure 6. Source classification and derived source parameters as a function of distance. Top panels: fraction of protostellar sources in the synthetic catalogue (solid line) and in the Hi-GAL catalogues (dashed lines) for the Outer Galaxy (left) and the Inner Galaxy (right). Second row: median values of the synthetic source diameters, separated into starless (red) and protostellar (blue) sources. Error bars are between the 2.5 and 97.5 percentiles. The median values for the Hi-GAL sources are shown by the dashed lines, while the shaded regions correspond to the same percentile range as for the error bars. Third row: the same as the second row of panels, but for the source mass derived from the SED fitting. Fourth row: the same as the previous row, but for the source temperature.

of the IMF discussed below. Compared with the Inner Galaxy, our source masses, particularly the protostellar ones, are significantly lower than in the observations at all distances. We attribute this discrepancy primarily to the much longer lines of sight through the Inner Galaxy (several kpc) than through our simulation (250 pc), or through the Outer Galaxy (dominated by the Perseus arm), as discussed later in Section 7.2. We will argue that the difference between intrinsic and projected clump masses that will be discussed in Section 6.1 must be even larger in the Inner Galaxy survey than in the synthetic observations.

Finally, the bottom panels of Fig. 6 show that the observed clump temperatures are well reproduced by the synthetic catalogue, except for slightly larger values in the case of protostellar sources at 2 and 4 kpc in the Outer Galaxy, and at all distances in the Inner Galaxy. This temperature difference in Outer Galaxy may be due to the incompleteness of the IMF in the simulation, as explained below in the context of Fig. 7. In the Inner Galaxy, the discrepancy is larger, and is primarily due to the same projection effect that causes the mass discrepancy discussed above: the inclusion of colder gas in the line of sight brings the protostellar sources' brightness temperature

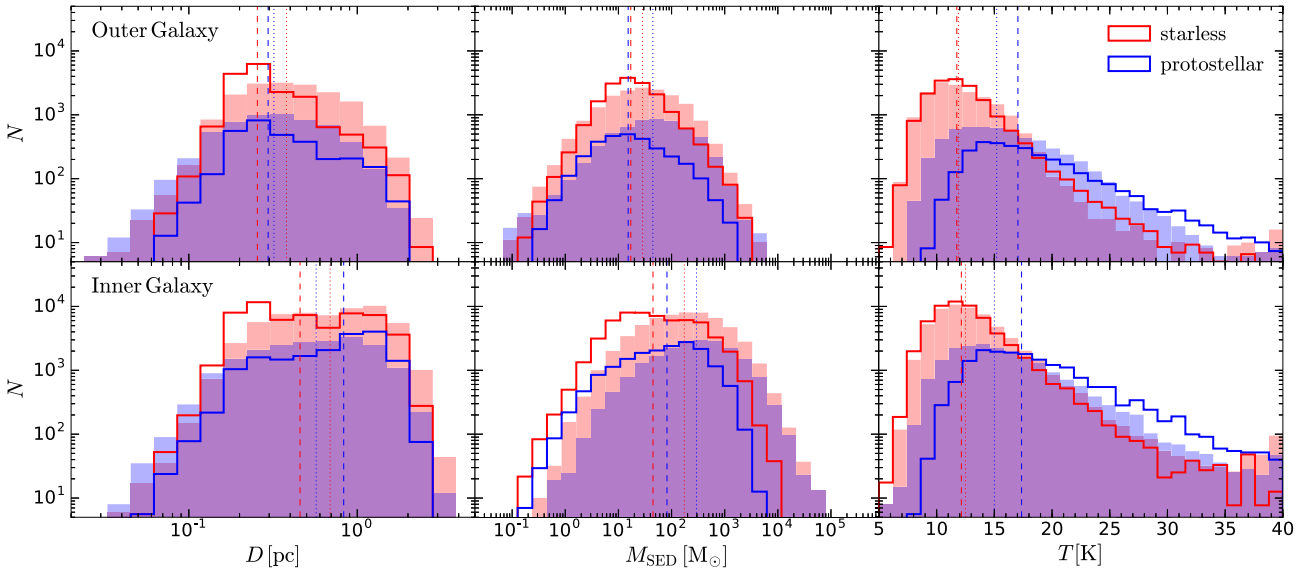

Figure 7. Upper panels: distributions of the source diameters, masses, and temperatures for the starless (red solid line) and protostellar (blue solid line) sources from the synthetic catalogue, compared to the sources from the Hi-GAL Outer Galaxy catalogue (red and blue bars) in the distance interval between 1.5 and 13.5 kpc (for the synthetic catalogue, the spread in the size distributions is dominated by the imposed distance distribution). The number of synthetic sources (both starless and protostellar) is normalized by the ratio of Hi-GAL and synthetic starless sources. The vertical dashed and dotted lines are the median values for the synthetic and Hi-GAL catalogues, respectively. Lower panels: the same as the upper panels, but comparing with the Hi-GAL Inner Galaxy catalogue. To approximate the distance distribution of the Hi-GAL sources in the Inner Galaxy, the number of synthetic sources at distances $d > 2$ kpc is scaled by a factor $(d/2\,\mathrm{kpc})^2/2$ (see Section 5).

down from their intrinsic one. This is consistent with the fact that the difference in median temperatures between starless (dashed red line) and protostellar (dashed blue line) sources is smaller in the Inner Galaxy than in the Outer Galaxy. By contrast, the temperature of starless sources is perfectly reproduced in both median values and dispersion at all distances for both the Outer and the Inner Galaxy.

Fig. 7 shows the probability distribution of the clump diameters (left-hand panels), clump masses (middle panels), and clump temperatures (right-hand panels) for starless and protostellar synthetic sources (solid-line histograms), compared with the same distributions for the Hi-GAL sources (shaded-area histograms) in the Outer Galaxy (upper panels) and in the Inner Galaxy (lower panels). All Hi-GAL sources with distances between 1.5 and 13.5 kpc are included in the shaded-area histograms. The histograms from the synthetic catalogue are normalized with the total number of synthetic starless sources to the same total number of observational starless sources, and using the same normalization factor for the synthetic protostellar clumps, too. Thus, the area below the histograms of synthetic and observed starless sources are the same, while the areas below the histograms of protostellar sources reflect the actual ratios of protostellar to starless sources in both the synthetic and Hi-GAL catalogues. The vertical dashed and dotted lines are the median values of the synthetic and Hi-GAL catalogues, respectively.

As shown by the left-hand panels of Fig. 7, there is very good agreement between the synthetic and Hi-GAL sources with respect to the source diameters in both the median values and the shapes of the probability distributions. The difference in the shape of the histograms between the Inner and Outer Galaxy is due to the different distance distributions of the sources. The more uniform distance distribution of the Inner Galaxy sources relative to the Outer Galaxy ones causes their flatter diameter distribution in the approximate diameter range between 0.2 and 2 pc, while the Outer Galaxy sources are picked around 0.3 pc. Because these size distributions are mainly due to the distance distributions of the sources, which was matched to the observational ones *a posteriori*, they should not be considered as further evidence of the agreement between the simulation and the observations.

The central upper panel of Fig. 7 shows that the mass distribution of the synthetic starless sources fits well the corresponding distribution from the Outer Galaxy, while that of the protostellar sources slightly underestimates the number of massive sources. In the case of the Inner-Galaxy comparison (central lower panel of Fig. 7), both the starless and the protostellar mass distributions of the synthetic sources are clearly shifted to smaller masses relative to the observations (see discussion in Section 7.2).

Finally, the right-hand panels of Fig. 7 show that the temperature distributions of starless sources are nearly perfectly matched by the synthetic sources, both for the Outer and the Inner Galaxy. On the other hand, the temperatures of the synthetic protostellar sources have median values $\sim$2 K larger than in the observations, as well as high-temperature tails that are a bit shallower than those of the Hi-GAL sources. These differences in the temperature distributions of protostellar sources are consistent with the expected consequence of the incompleteness of our stellar mass distribution. As mentioned in Section 2, the stellar mass distribution in the simulation is consistent with Salpeter's IMF (Salpeter 1955) above $\sim$8 $\mathrm{M}_\odot$, but is incomplete at lower masses (and no stars below $\sim$2 $\mathrm{M}_\odot$ are included as stellar sources in the radiative transfer calculation). At the lower distances (2 and 4 kpc), the 70 μm flux of most stars in the range $\sim 2 - 8\,\mathrm{M}_\odot$ would be detected in the observations, but many of these 70 μm sources would be missing in our simulation. The fraction of missing stars increases with decreasing stellar masses, thus we are deficient in lower luminosity stars that are less able to heat their surroundings. Therefore, we miss preferentially colder protostellar sources, which matches the discrepancy in the temperature distributions of

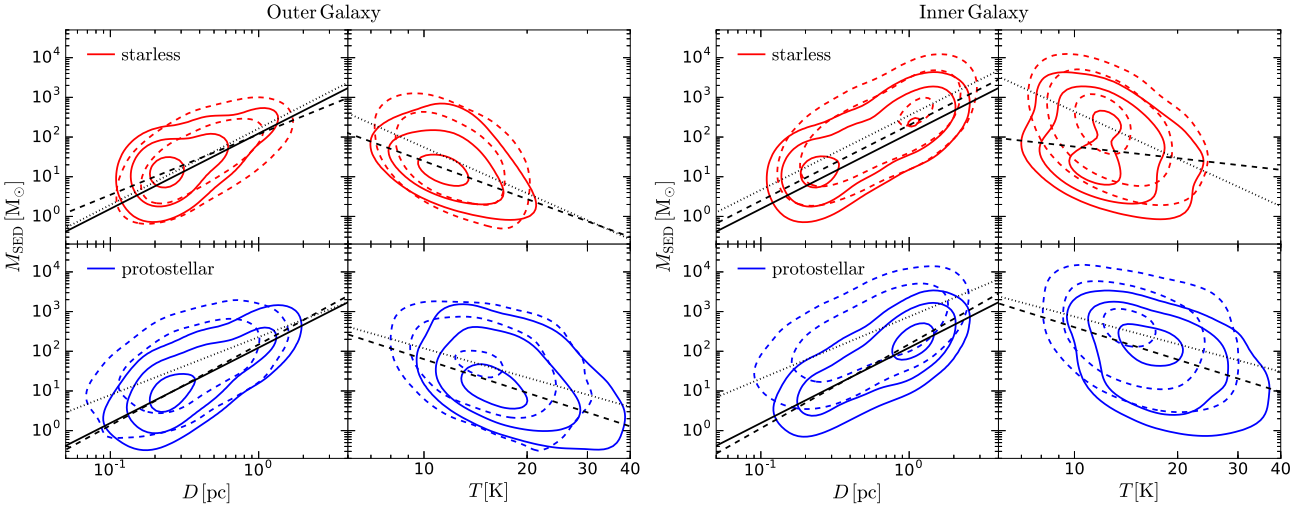

Figure 8. Left four panels: mass versus diameter and mass versus temperature for starless (upper panels) and protostellar (lower panels) synthetic sources (solid contour lines). The black line is Larson's mass–size relation, $M(\mathrm{r}) = 460\ \mathrm{M}_\odot(\mathrm{r/pc})^{1.9}$ (Larson 1981). Source number density isocontours representing the starless and protostellar sources from the Hi-GAL Outer Galaxy catalogue are also displayed with dashed lines. The number densities have been computed subdividing the area of the plot with a grid of 70 × 70 cells in logarithmic intervals. Once the global maximum of both distributions in the same panel (starless or protostellar) has been found, the plotted contours are chosen to show the number density of 5 per cent, 20 per cent, and 70 per cent of the maximum, as in Fig. 7 of Elia et al. (2017). Right four panels: the same as in left four panels, but comparing with the sources from the Hi-GAL Inner Galaxy survey. In this case the number of synthetic sources at distances $d > 2$ kpc has been rescaled by a factor $(d/2\,\mathrm{kpc})^2/2$, to mimic the distance distribution of the sources in the Inner Galaxy (see Section 5). In all panels, the dashed lines show the least-squares fits to the underlying scatter plots from the synthetic catalogue, while the dotted lines are the fits to the observations.

protostellar sources in the lower left panel of Fig. 6 and in the right-hand panels of Fig. 7.

5.3 Size, mass, and temperature correlations

The left group of four panels of Fig. 8 shows the relation between the mass and diameter (left), and the mass and temperature (right) of the synthetic sources (solid contour lines) and of the Hi-GAL Outer Galaxy sources (dashed contour lines) for starless (upper panel) and protostellar (lower panel) sources separately. The right group of four panels shows the same, but for the Hi-GAL Inner Galaxy survey. The contours are lines of equal number density of sources. The number density is computed separately for the synthetic and Hi-GAL sources, by dividing the plane into 70 × 70 logarithmic intervals of diameter and mass. The contours correspond to 5 per cent, 20 per cent, and 70 per cent of the maximum number density among all the cells from both samples. The comparison shows that our synthetic sources follow very closely the mass–diameter and the mass–temperature relations of the Outer Galaxy sources, with respect to both the contour lines and the median least-square-fit relations (dashed-dotted and dotted lines).

In the case of the Inner Galaxy, as shown by previous figures, the observed clump masses are on average a few times higher than those of the synthetic observations, the discrepancy being stronger in the case of protostellar sources. In both Hi-GAL catalogues, the median mass–size relations have nearly the same slope as Larson's mass–size relation (solid line), corresponding to roughly constant surface density (Larson 1981). This slope is well reproduced in the synthetic catalogue as well.

All the mass-temperature contour plots and median relations from our catalogue and from the observations show the same trend of decreasing mass with increasing temperature. However, this trend is not as significant as the correlation between mass and diameter, as shown by the values of Pearson's correlation coefficients, between approximately −0.4 and −0.5 (compared to ∼0.7 − 0.8 for the mass–size relations). It is possible that most of the anticorrelation between mass and temperature originates from the uncertainty in the temperature estimate from the SED fitting, rather than from a real physical anticorrelation between mass and temperature.

6 OBSERVATIONAL VERSUS INTRINSIC CLUMP PROPERTIES

Projection effects along the line of sight can strongly affect the derivation of observational clump properties, particularly when no kinematic information is available. To establish the importance of projection effects, we search for the 3D counterparts of the synthetic sources in the simulation data cubes used to generate the synthetic observations. By relating the synthetic sources to their 3D counterparts, we can then compare the observational properties with the intrinsic clump properties. In this section, 'Outer Galaxy' refers to our 51 831 synthetic sources, and 'Inner Galaxy' to the 107 453 synthetic sources obtained from the synthetic catalogue by duplicating more distant sources in order to match the distance distribution of the Inner Galaxy sample of the Hi-GAL catalogue (see Section 5).

6.1 Observational versus intrinsic clump mass

Fig. 9 shows the density profile along the lines of sight of two sources of similar mass (618 and 1037 $\mathrm{M}_\odot$) at 12 kpc, both classified as protostellar in the synthetic catalogue, chosen to be an example of a line of sight dominated by a real 3D clump (left-hand panel) and of a line of sight without a dominant 3D clump (right-hand panel). The gas density is computed within a column of size equal to the source diameter. The open cyan diamonds mark the positions of stars located inside the main 3D clump, while the open red circles are for the stars located in the column of the source footprint, but outside

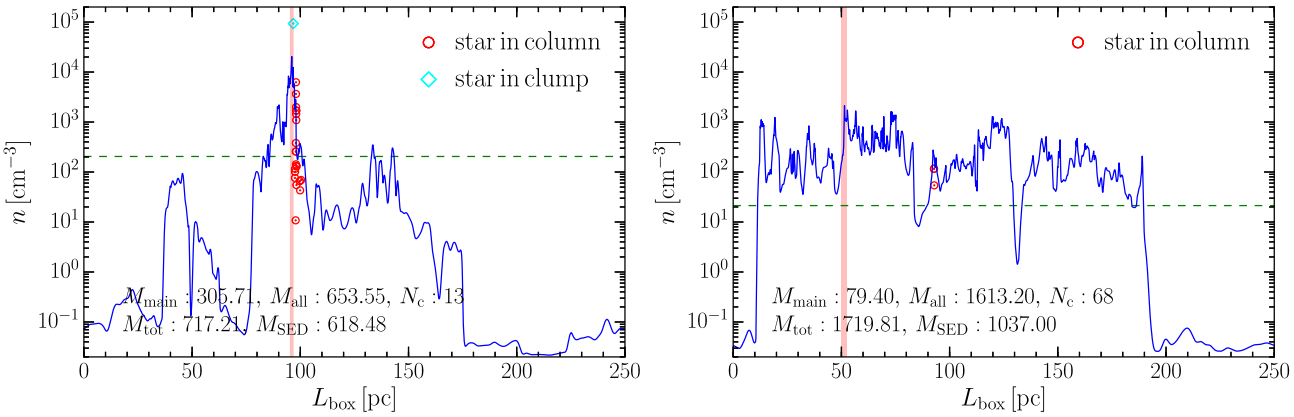

Figure 9. Density profile along the line of sight for two example sources (the same two as in Fig. 4): one source has a dominant clump (left-hand panel) and the other has many clumps of approximately the same peak density (right-hand panel). The red vertical line shows the position of the main clump (the highest density peak), with the thickness of the line corresponding to the source diameter (we assume that the 3D clump has the same diameter as the corresponding source). The open cyan diamonds are the stars located inside the main clump. The open red circles are the stars located in the column of the source footprint, but outside of the dominant clump. The green dashed horizontal line shows the 1 per cent of the maximum density, which is a minimum density we have chosen to define secondary clumps.

of the main 3D clump. The y-axis coordinate of the diamonds and circles corresponds to the mean density around the star in a $(0.12\,{\rm pc})^3$ volume. While in the example dominated by the main 3D clump (left-hand panel), one star is found inside the main clump, and several more in its vicinity, only two stars are found in the line of sight of the other example (right-hand panel) and at relatively low densities, and none inside the main 3D clump.

Density profiles like those shown in Fig. 9 are computed for all the sources in our synthetic catalogue, and are then used to identify the main 3D clump corresponding to each synthetic source. The main 3D clump in each line of sight is defined as the fraction of the corresponding column centred at the highest density maximum and with a size along the line of sight equal to the synthetic source diameter (the size on the plane of the sky). The position of the maximum density is shown by the vertical transparent line in Fig. 9, whose thickness corresponds to the diameter of the synthetic source and thus of the main 3D clump. We also define as independent 3D clumps all other density maxima with value larger than 1 per cent of the absolute maximum that defines the main 3D clump (see the horizontal dashed line). All 3D clumps are assumed to have a size in the direction of the line of sight equal to the source diameter as well.

With these definitions of main and secondary 3D clumps, we can then associate three different mass values to each line of sight, the total mass of the column, $M_{\rm tot}$, the sum of the masses of all clumps, $M_{\rm all}$, and the mass of the main clump, $M_{\rm main}$, and compare them with the mass of the corresponding synthetic source, $M_{\rm SED}$. In the two examples of Fig. 9, the main 3D clump in the left-hand panel has a mass $M_{\rm main} = 305.7\,{\rm M}_\odot$, approximately half the value of the estimated source mass, $M_{\rm SED}$, while the main 3D clump in the right-hand panel has a mass $M_{\rm main} = 79.4\,{\rm M}_\odot$, approximately 13 times smaller than $M_{\rm SED}$.

The comparison of $M_{\rm all}$ and $M_{\rm main}$ with $M_{\rm SED}$ for all the sources in our catalogue is shown in Fig. 10, where the black solid line marks the one-to-one relation. The least-squares fits to the median value of each of the two masses as a function of $M_{\rm SED}$ are shown in Fig. 10 as dashed and dotted lines. The most striking result of this comparison is that $M_{\rm SED}$ overestimates the true mass of the main 3D clump, $M_{\rm main}$, by a factor greater than 10, on average (dashed contour lines and dotted least-squares fit). Because the median relation given by the fit is rather shallow, $M_{\rm main} = (0.230 \pm 0.014) M_{\rm SED}^{(0.587\pm0.020)}$ for the

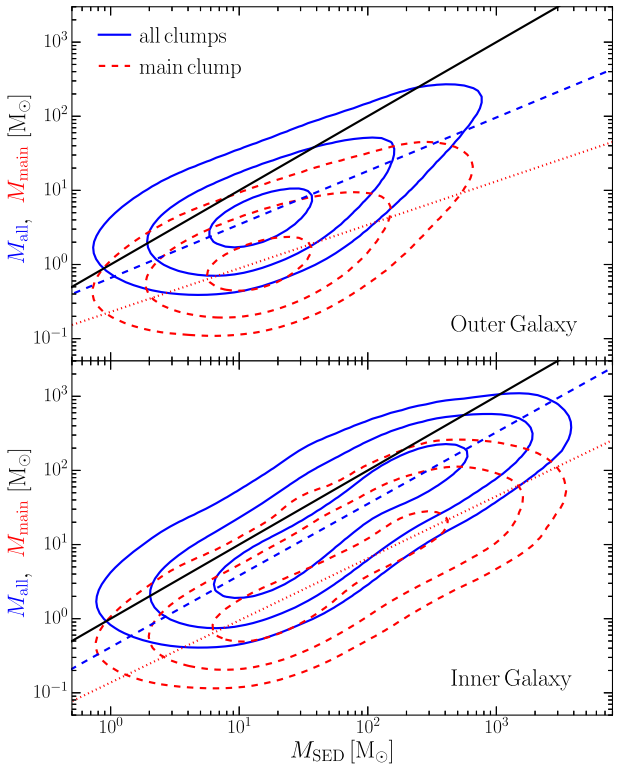

Figure 10. The total mass of the dense structures along the line of sight, $M_{\rm all}$ (blue solid contour lines), and main 3D clump mass, $M_{\rm main}$ (red dashed contour lines), versus the source mass derived from the SED fitting, $M_{\rm SED}$, for all the sources in the synthetic catalogue, normalized with the Outer Galaxy distance distribution (upper panel) and the Inner Galaxy distance distribution (lower panel). The solid black line is the one-to-one relation. The blue dashed straight line is the least-squares fit to the $M_{\rm all} - M_{\rm SED}$ relation, $M_{\rm all} = (0.658 \pm 0.046) M_{\rm SED}^{(0.722\pm0.023)}$ for the Outer Galaxy and $M_{\rm all} = (0.412 \pm 0.027) M_{\rm SED}^{(0.967\pm0.015)}$ for the Inner Galaxy case, with the masses in units of M$_\odot$. The dotted line is the least-squares fit to the $M_{\rm main} - M_{\rm SED}$ relation, $M_{\rm main} = (0.230 \pm 0.014) M_{\rm SED}^{(0.587\pm0.020)}$ for the Outer Galaxy, and $M_{\rm main} = (0.137 \pm 0.009) M_{\rm SED}^{(0.838\pm0.015)}$ for the Inner Galaxy case.

Outer Galaxy and $M_{\rm main} = (0.137 \pm 0.009) M_{\rm SED}^{(0.838\pm0.015)}$ (the masses are in $M_\odot$), the ratio $M_{\rm SED}/M_{\rm main}$ increases with increasing $M_{\rm SED}$: the larger the value of $M_{\rm SED}$, the larger the contribution to that mass from structures along the line of sight unrelated to the main 3D clump. In other words, *most of the mass of a CuTEx source is due to the overlap of different structures, and increasingly so towards larger masses.* As a confirmation of this, we find that $M_{\rm all} \sim M_{\rm tot}$, which means that most of the mass in the line of sight is in relatively dense structures (secondary clumps) rather than in a diffuse background. In the case of the Hi-GAL catalogue, this effect is likely to be even stronger than in our simulation, as discussed in Section 7.2.

The relation between $M_{\rm all}$ and $M_{\rm SED}$, $M_{\rm all} = (0.658 \pm 0.046) M_{\rm SED}^{(0.722\pm0.023)}$ (Outer Galaxy) and $M_{\rm all} = (0.412 \pm 0.027) M_{\rm SED}^{(0.967\pm0.015)}$ (Inner Galaxy), is shallower than linear. $M_{\rm all}$ is, on average, only slightly smaller than $M_{\rm SED}$, though with a very large scatter, which demonstrates that the SED fitting approximately recovers the total mass of the dense clumps in the line of sight, rather than that of a single, isolated clump. In the case of aperture photometry, one would expect $M_{\rm SED} \approx M_{\rm all}$ (within the errors). However, in some cases, we derive $M_{\rm SED} > M_{\rm all}$, which shows the impact of band-merging occasionally leading to a lower temperature (compared to the mass-averaged temperature in the line of sight) from the SED fit and hence an overestimate of $M_{\rm SED}$.

6.2 Starless versus protostellar clumps

Having associated our synthetic sources with 3D clumps, we can now ask whether the sources classified as protostellar are associated with actual 3D protostellar clumps or not. For this possible association, we consider both the main 3D clump and other secondary clumps along the line of sight of each source, using the existence of embedded stars along the line of sight discussed in Section 4.4 to divide protostellar sources into true and false protostellar sources. In the two examples of Fig. 9, the source from the left-hand panel is a true protostellar clump (it has seven stars at density larger than 10^3 cm^{-3}), while the source from the right-hand panel is a false protostellar clump (it has no stars above the density threshold).

We have found that true protostellar sources can be partially differentiated from false protostellar sources based on purely observable quantities, using the dependence of their 70 μm excess on their temperature. Before demonstrating this method on the synthetic sources, we want to verify that the relation between 70 μm excess and temperature of the synthetic protostellar sources reproduces well that of the observations. The 70 μm excess is defined as the ratio of the 70 μm flux of the source, $F_{70,\rm ob}$, and the greybody 70 μm flux, $F_{70,\rm gb}$, that is the 70 μm flux extrapolated from the SED fit of the fluxes at the other wavelengths. The upper panel of Fig. 11 shows that the temperature dependence of $F_{70,\rm ob}/F_{70,\rm gb}$ for the protostellar sources in our synthetic catalogue (blue shaded area) has approximately the same median values and dispersion as for the Hi-GAL sources in the Outer Galaxy (red shaded area). Compared with the Inner Galaxy, the 70 μm excess of the synthetic sources is slightly larger than for the Hi-GAL sources (lower panel).

The relation between the 70 μm excess and the temperature of false protostellar synthetic sources is shown by the magenta shaded areas in Fig. 12. The cyan shaded areas corresponds to the true protostellar synthetic sources. In the upper panel, the number of sources at different distances is the same as in the synthetic catalogue, to mimic the distance distribution of sources in the Outer Galaxy, while in the lower panel the number of sources has been renormalized according to their distance to mimic the distance distribution in the Inner Galaxy, as in Fig. 7. The dashed red line and the cyan solid line show the median values of $F_{70,\rm ob}/F_{70,\rm gb}$ as a function of temperature, measured in logarithmic temperature intervals, for the false and true pre-stellar sources, respectively. The true protostellar sources are systematically warmer, at equal value of $F_{70,\rm ob}/F_{70,\rm gb}$, or, equivalently, they have systematically larger $F_{70,\rm ob}/F_{70,\rm gb}$ at equal value of temperature (by almost an order of magnitude in the case of the Outer Galaxy's distance distribution shown in the upper panel). However, there is also a considerable overlap between the two types of sources.

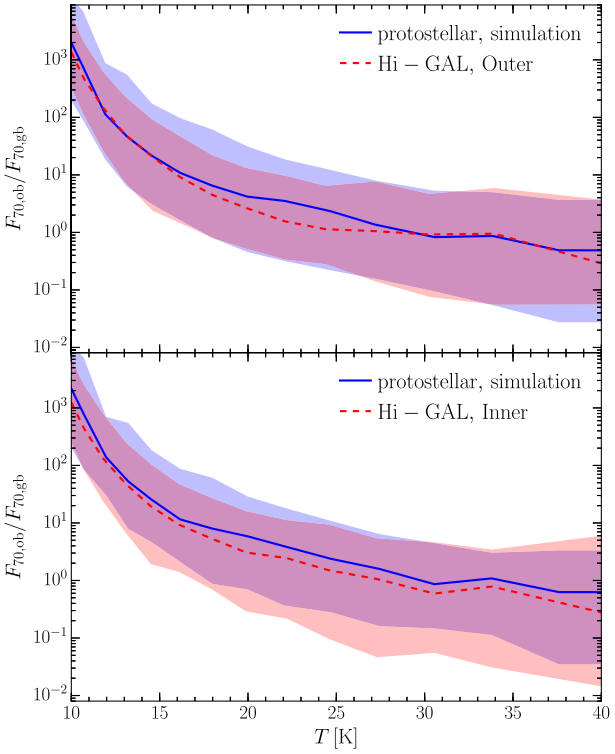

Figure 11. The ratio of the observed 70 μm flux, $F_{70,\rm ob}$, and 70 μm flux extrapolated from the greybody SED fitting, $F_{70,\rm gb}$, versus the temperature for the protostellar sources in the synthetic and the Hi-GAL Outer Galaxy catalogues (upper panel), and the Inner Galaxy catalogue (lower panel). The solid blue and dashed red lines are the median values for the synthetic protostellar sources and Hi-GAL protostellar sources, respectively. The shaded areas (blue and red, respectively) show the interval between the 2.5 and 97.5 percentiles of the flux ratio.

To partially separate true from false protostellar sources, we propose to separate the protostellar sources in two groups using the median relation for the true sources (cyan line). We first obtain an analytical fit to the cyan line for the Outer Galaxy distance distribution,

$$F_{70,\rm ob}/F_{70,\rm gb} = \frac{2877}{(T/10\,{\rm K})^{15}} + \frac{367}{(T/10\,{\rm K})^{5}} + 1.3\,, \quad (6)$$

and for the Inner Galaxy case,

$$F_{70,\rm ob}/F_{70,\rm gb} = \frac{2051}{(T/10\,{\rm K})^{15}} + \frac{365}{(T/10\,{\rm K})^{5}} + 0.5\,, \quad (7)$$

which is shown by the solid black lines in Fig. 11. The dashed black lines connecting the diamond symbols show the ratio between true protostellar sources and total protostellar sources found above the analytical fits. The values of this ratio are shown in the y-axis on the right-hand side of the plots. The lower black dotted lines,

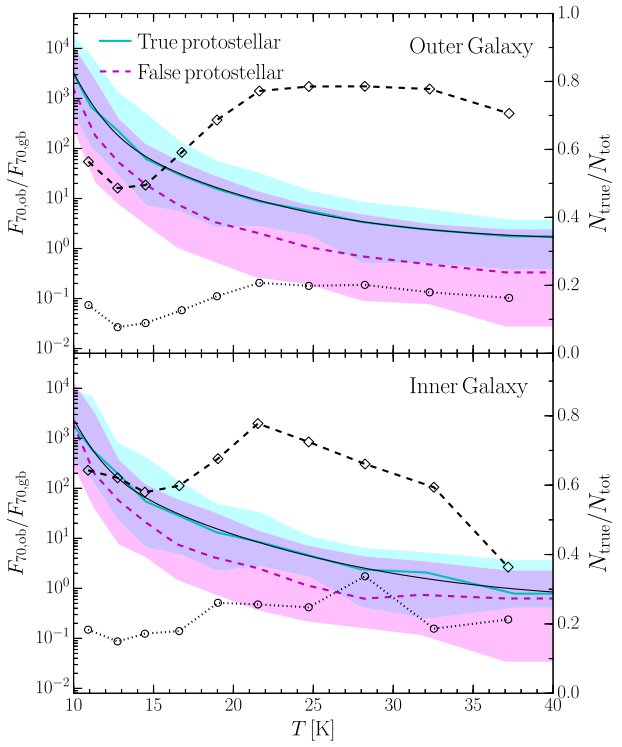

Figure 12. The ratio of the observed 70 μm flux, $F_{70,\,ob}$, and 70 μm flux extrapolated from the greybody SED fitting, $F_{70,\,gb}$, versus the temperature for the protostellar sources in the synthetic catalogue. The cyan line is the median values for the 'true' protostellar sources, which are those with at least one embedded star (with local density $\geq 1000\,\mathrm{cm}^{-3}$) in their line of sight. The magenta line shows the median values for the 'false' protostellar sources, which are those without embedded stars in their line of sight. The shaded areas show the 2.5 and 97.5 percentiles of the 70 μm flux ratio. The solid black line is an analytical fit to the cyan median line, and the dashed and dotted black lines are the fraction of 'true' protostellar sources of all sources above and below the solid black line, respectively, with the value of these fractions given by the scale on the right y-axis.

connecting the circle symbols, show the same ratio for sources below the analytical fit to the median line. One can see that, in the case of the Outer Galaxy, approximately 80 per cent of the protostellar sources are true ones in the approximate temperature range between 21 and 33 K, while below the lines the fraction of true protostellar sources is always lower than 20 per cent. Averaging over all temperatures, the fraction above the line is 0.64 and 0.66 for the Outer and Inner Galaxy, respectively, while the fraction below the line is 0.14 and 0.21, for the same cases. In summary, the method proposed here allows to select a class of sources where 60–70 per cent are true protostellar ones (they have real 3D clumps along the line of sight containing embedded protostars), by selecting half of the total number of candidate protostellar sources.

Because we can partially extract true protostellar sources, it is interesting to check the relation of their estimated mass, $M_{\rm SED}$, with the mass of the corresponding main 3D clumps, $M_{\rm main}$, for this specific subset of sources. Fig. 13 shows contour plots of the relation between $M_{\rm main}$ and $M_{\rm SED}$ for the synthetic sources classified as starless (left-hand panel), false protostellar (second panel from the left), true protostellar (third panel from the left), and for the sources above the line defined by equation (6; fourth panel from the left). While for false protostellar sources the mass discrepancy, $M_{\rm SED} \gg M_{\rm main}$, is comparable to that of starless sources, for true protostellar sources it is significantly reduced. The fifth panel of Fig. 13 shows the probability distributions of the ratio $M_{\rm SED}/M_{\rm main}$ for the four populations, with the vertical dashed lines corresponding to the median values. The median values decrease from nearly 20 for starless sources to ≈4 for true protostellar ones.

Even sources selected with the observational criterion suggested above, based on the relation in equation (6), have a median ratio $M_{\rm SED}/M_{\rm main} \approx 4$. This analysis shows that, a source selection that gives a majority of true protostellar sources also gives sources that are less affected by (though not free from) projection effects.

7 DISCUSSION

7.1 Implications for clump dynamics and evolution

We have found that the estimated mass of our synthetic sources, $M_{\rm SED}$, is typically of the order of the total mass in the line of sight of the source, and an order of magnitude larger than the densest real 3D clump identified in the simulation along the line of sight: $M_{\rm SED} \sim M_{\rm tot} \gg M_{\rm main}$. This implies that our synthetic sources, and by extension the Hi-GAL sources, should not be considered as individual clumps. Instead, they are generally a collection of separate high-density structures along the line of sight, because most of the mass is contained in secondary clumps: $M_{\rm all} \sim M_{\rm tot} \sim M_{\rm SED}$. If interpreted as individual clumps, one should keep in mind that the masses of such clumps are most likely overestimated by a very large factor, which complicates the analysis of their dynamical state (e.g. bound versus unbound (Elia et al. 2017)), their evolutionary state (e.g. the clump mass–luminosity relation (Molinari et al. 2008)), their statistical properties (e.g. the velocity–size relation (Traficante et al. 2018)), and their role in the formation of massive stars (e.g. the estimated infall rates (Traficante et al. 2018), or the estimated column density threshold (Tan et al. 2014; Urquhart et al. 2014)). Even the selection of a subset of sources with molecular emission-line spectra without multiple components (e.g. Traficante et al. 2018) may not be sufficient to prevent projection effects, as the existence of a large number of unrelated sources along the line of sight could presumably result in the appearance of an approximately Gaussian velocity profile.

In Section 5.3, we found that our synthetic sources, as well as the Hi-GAL sources, follow very closely Larson's mass–size relation, corresponding to their surface density being nearly independent of size, on average. Elia et al. (2017) implicitly assumed that this mass–size relation of Hi-GAL sources had to stem from the combination of Larson's velocity–size and velocity–mass relations (Larson 1981), so that sources above (more massive than) the average mass–size relation could be interpreted as gravitationally bound, and sources below (less massive than) the relation would be unbound. In view of our finding that $M_{\rm SED}$ is not a reliable estimate of a real clump mass, the interpretation of the mass–size relation of Hi-GAL sources is not straightforward. In fact, it was later found that Hi-GAL sources do not follow Larson's velocity–size relation at all (Traficante et al. 2018), which is in itself a result of difficult interpretation if the velocity dispersion has potentially multiple contributions along the line of sight (not necessarily spotted as multiple spectral components), but certainly invalidates the bound versus unbound classification of Hi-GAL sources based on the mass–size relation.

Observed infall rates of massive clumps (e.g. Fuller, Williams & Sridharan 2005; Beuther, Linz & Henning 2012; Peretto et al. 2013; Beuther, Linz & Henning 2013; Wyrowski et al. 2016; Traficante et al. 2017, 2018; Contreras et al. 2018; Yuan et al. 2018) are often used to constrain the formation time-scale of massive stars.

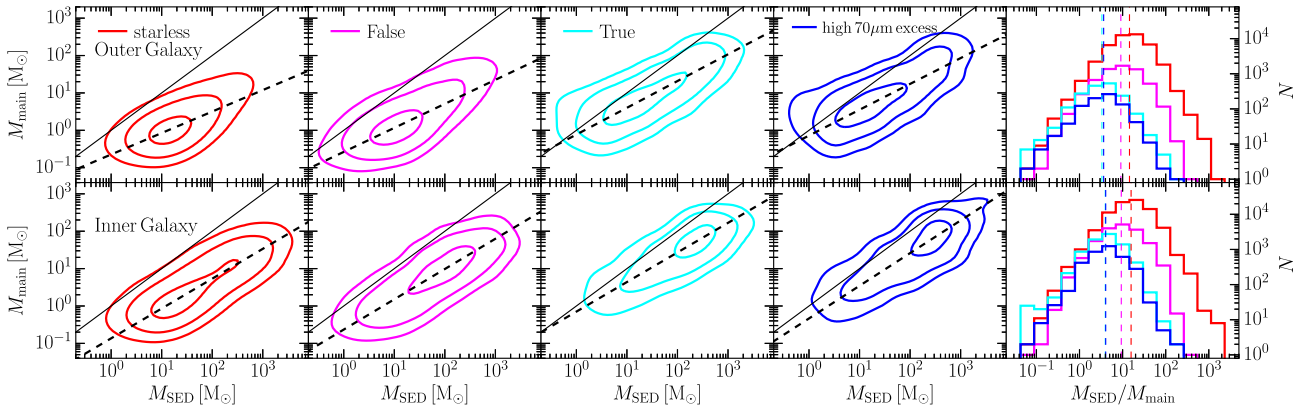

Figure 13. Main 3D clump mass versus SED mass, as in Fig. 10, for starless, 'false' protostellar, 'true' protostellar, and protostellar sources selected based on the 70 μm excess (see Fig. 12), from left to right. The thin solid black line is the one-to-one relationship. The fits shown by the dashed black lines are: $M_{\rm starless} = (0.226 \pm 0.012) M_{\rm SED}^{(0.574\pm0.018)}$, $M_{\rm false} = (0.264 \pm 0.018) M_{\rm SED}^{(0.643\pm0.023)}$, $M_{\rm true} = (0.769 \pm 0.057) M_{\rm SED}^{(0.716\pm0.021)}$, and $M_{\rm high} = (0.722 \pm 0.073) M_{\rm SED}^{(0.691\pm0.030)}$ for the Outer Galaxy. For the Inner Galaxy: $M_{\rm starless} = (0.136 \pm 0.008) M_{\rm SED}^{(0.798\pm0.013)}$, $M_{\rm false} = (0.234 \pm 0.018) M_{\rm SED}^{(0.810\pm0.017)}$, $M_{\rm true} = (0.707 \pm 0.095) M_{\rm SED}^{(0.791\pm0.024)}$, and $M_{\rm high} = (0.519 \pm 0.087) M_{\rm SED}^{(0.862\pm0.029)}$. All the masses in the previous relations are in units of M$_\odot$. The rightmost panels show the distributions of the $M_{\rm SED}/M_{\rm main}$ ratio.

Besides the difficulty of interpreting the kinematic information from emission line spectra in terms of infall, and the need to explain to low detection rate of infall signatures in massive clumps, it is important to remember that a massive clump may host the formation of multiple stars. Furthermore, in view of the results of this study, the clump mass may be grossly overestimated. The infall rate is usually estimated assuming that the infalling gas has a density equal to the clump's mean density. Thus, if the clump mass is overestimate by a factor of 10, the infall rate is also overestimated by the same factor.

Fig. 14 shows the distributions of the mean volume density of our synthetic sources (blue histograms) and of the Hi-GAL sources (red dashed line histograms) for the Outer Galaxy (upper panel) and the Lower Galaxy (lower panel). The cyan histograms show the distribution of the mean density of the main 3D clump associated to each of our synthetic sources. This distribution is shifted to lower densities by a factor of 10 relative to that of the synthetic sources, as expected from the mass comparison in Section 6.1. In the case of the Hi-GAL sources, the factor may be even larger, as discussed below in Section 7.2. Based on the median values shown by the vertical lines in Fig. 14, infall rates based on estimated densities of Hi-GAL Inner Galaxy sources may be overestimated by more than a factor of 20, since the median density in the Hi-GAL Inner Galaxy catalogue is more than a factor of two larger than in our synthetic catalogue, which is more than a factor of 10 larger than the median for the corresponding 3D clumps. This uncertainty could be reduced, by approximately a factor of three, if true protostellar sources were selected according to the method suggested in Section 6.2, but even in that case over 30 per cent of sources would still be false protostellar sources suffering a larger projection effect.

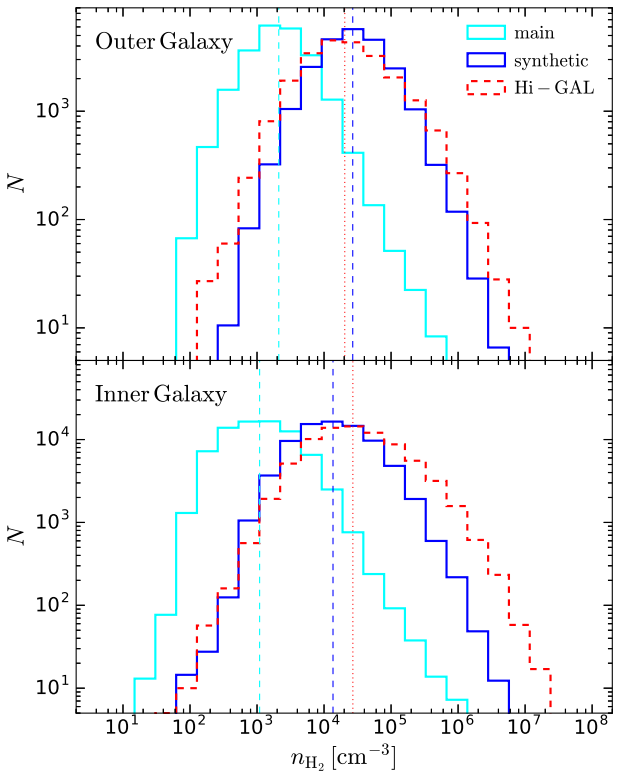

Figure 14. Distributions of the mean gas number density of the sources in the synthetic catalogue (solid blue lines) and in the Hi-GAL catalogue with distances between 1.5 and 13.5 kpc (red dashed lines), for the Outer Galaxy (upper panel) and the Inner Galaxy (lower panel). In the lower panel, the numbers of synthetic sources at distances >2 kpc have been rescaled as in previous figures. The cyan lines show the distributions of the mean density of the main 3D clumps in the lines of sight to the synthetic sources.

7.2 From the simulation volume to the galactic plane

The synthetic observations of our simulation reproduce the main observational properties, and their statistical distributions, of the Hi-GAL catalogue, with the caveat of a systematic shift of the mass distribution towards lower masses. The mass discrepancy is significant only in the high-mass tail of the protostellar mass distribution in the case of the Outer-Galaxy comparison, while the mass distributions of both starless and protostellar sources in the Inner Galaxy are clearly shifted towards larger masses in both high- and low-mass tails, by a factor of ≈4 in mass relative to our synthetic catalogue (see Fig. 7). This discrepancy is to be expected, because the 250 pc depth of the computational volume cannot match the full complexity and the highest column densities of the lines of

sight towards the Inner Galactic plane sampled by the Hi-GAL survey.

However, the larger column density available does not mean that the masses of Hi-GAL's sources are more real (closer to the mass of actual 3D clumps) than those in our synthetic catalogue. As shown in Section 6.1, the estimated source masses are in large part the result of a projection effect. Thus, we interpret the larger Hi-GAL masses as the result of stronger projection effects in the observations than in the simulation. This is supported by the fact that the mass discrepancy is significantly stronger in the comparison with the Inner Galaxy than with the Outer Galaxy.

Besides the lower values of maximum column densities, the synthetic observations also lack sources of confusion both in the line of sight and on the plane of the sky that afflict Galactic disc surveys. Our synthetic observations should be more sensitive to fainter sources, as they are not limited by the confusion due to the Galactic structures, including the cirrus clouds. These cirrus clouds are real objects, but from an observational standpoint, their effect is similar to noise when trying to detect compact sources, and dominates over the purely instrumental noise for wavelengths longer than 70 μm, as shown in fig. 3 of Molinari et al. (2016). Our simulation lacks the equivalent of this cirrus noise, as the AMR too has low resolution of intermediate and low-density structures, smoothing them away, and the line-of-sight depth would not be enough to accumulate enough surface brightness to cause comparable confusion.

In Section 7.1, we have argued that the subdivision of Hi-GAL's starless sources into bound pre-stellar and unbound ones based on Larson's mass–size relation is most likely incorrect, because of the uncertainty of the source masses demonstrated in this work (see Section 6.1) and because Hi-GAL's sources do not follow Larson's velocity–size relation (Traficante et al. 2018). However, the separation of the sources into these two classes, essentially a cut at nearly constant surface density, offers an additional comparison test between the synthetic catalogue and Hi-GAL. For simplicity, we refer to these sources as pre-stellar and starless unbound, as in Elia et al. (2017), even if these names do not reflect their real nature. The only significant discrepancy between our catalogue and the observations, with respect to this column density cut, is the temperature distribution, which we show in Fig. 15 for the Outer and Inner Galaxy catalogues in the upper and lower panels, respectively.

The temperature distributions of the synthetic sources reproduce reasonably well the observations, except for a significantly deficient tail of high-temperature pre-stellar sources relative to the Inner Galaxy. The Hi-GAL Inner Galaxy catalogue contains an excess of high column density sources (classified as bound pre-stellar sources) with relatively high temperature, between ≈15 K and ≈30 K, which cannot be explained based on our radiative transfer calculations. The most straightforward explanation is that most of the warmest sources at relatively high column densities are a projection of a number of (starless unbound) sources of low enough column density that they are heated by their local interstellar radiation field to the observed temperatures, whereas a single, dense clump with that high column density would be shielded from radiation and hence be colder. This projection effect happens also in the synthetic observations, but it is clearly much stronger for the Inner-Galaxy sources. This interpretation is consistent with the fact that the temperature discrepancy of pre-stellar sources is insignificant in the case of the Outer Galaxy catalogue (upper panel of Fig. 15), where projection effects are expected to be significantly reduced, due to the nearly exponential drop in the number of sources as a function of distance beyond ≈2 kpc.

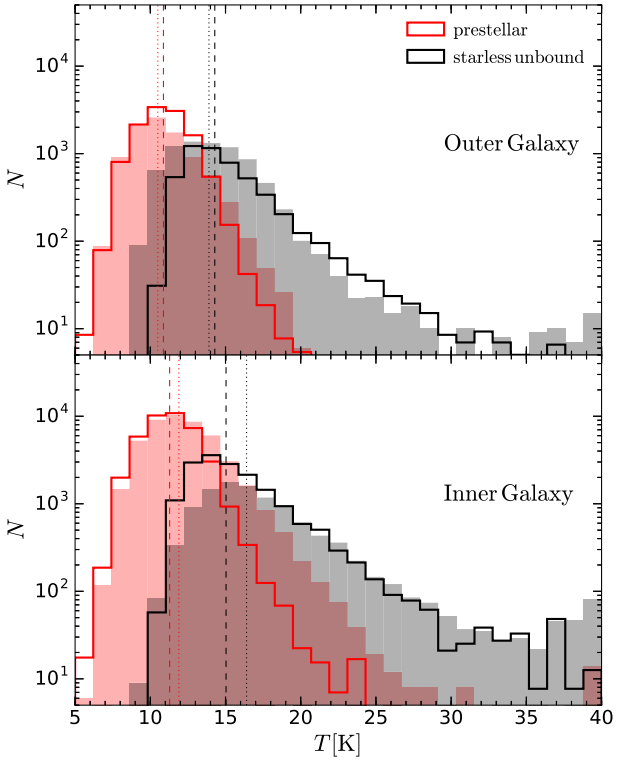

Figure 15. Greybody temperature distributions for the synthetic pre-stellar sources (red line) and the synthetic starless unbound sources (black line), compared with Hi-GAL sources in the distance interval between 1.5 and 13.5 kpc (shaded histograms), for the Outer Galaxy (upper panel) and the Inner Galaxy (lower panel). The dashed and dotted vertical lines are the median values of the synthetic and Hi-GAL sources, respectively.

7.3 Angular resolution and true nature of clumps

So far, we have focused on the uncertainty (mostly the overestimate) of the observed source masses due to blending of different density structures along the line of sight. We have shown that this uncertainty is large (typically a factor of 10) in our synthetic observations (Section 6.1) and thus in the Hi-GAL catalogue, as the synthetic observations were obtained by following closely the Hi-GAL data-analysis pipeline. We have further argued that projection effects are likely to be even worse in the Hi-GAL Inner Galaxy catalogue, due to the larger column density of lines of sight through the Galactic plane than through our 250 pc volume (Section 7.2). Besides these projection effects, the limited spatial resolution of the observations is another important factor that undermines the interpretation of compact sources as individual clumps. At the characteristic distance of Hi-GAL sources, primarily in the approximate range between 2 and 14 kpc, most of the compact sources at the angular resolution of Herschel's observations are expected to be highly fragmented. The effect of the angular resolution (or distance to the source) on the mass determination was already quantified in Padoan et al. (2020), where it was shown that the mass of true pre-stellar cores observed by Herschel at distances larger than 1 kpc would on average be overestimated by more than a factor of 10. The error grows with increasing distance, being a factor of ∼40 at a distance of ∼2 kpc (see fig. 28 in Padoan et al. 2020).

It is generally understood that compact sources from Herschel's observations do not represent individual pre-stellar cores, but are more likely to be the sites of formation of multiple stars. Their large internal velocity dispersions, revealed by follow-up studies,

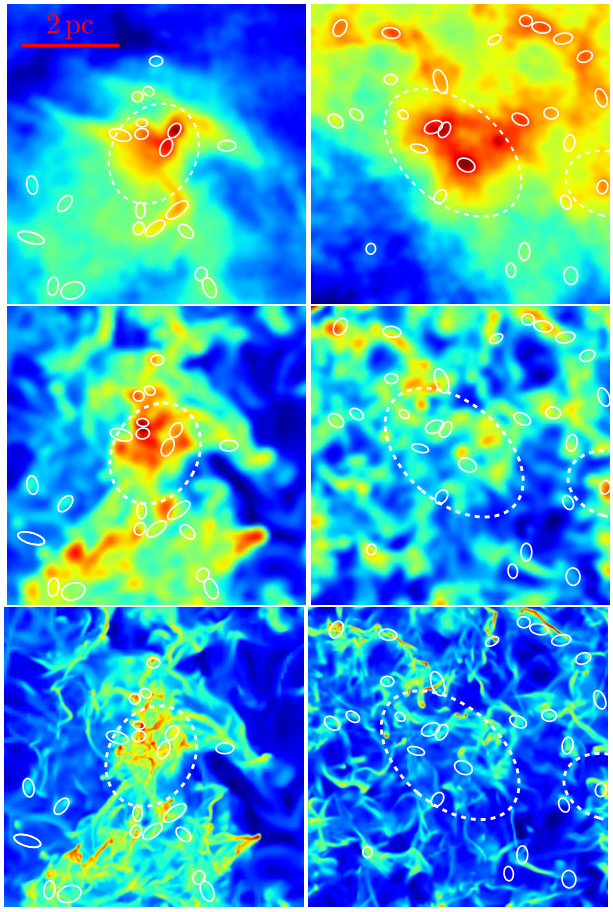

Figure 16. The two sources at 12 kpc from Fig. 4 are shown by the white dashed ellipses, in the left and right columns. The solid white ellipses are the sources found in the same 6 pc regions assumed to be at 2 kpc. The top panels show the 250 μm surface brightness maps at the distance of 2 kpc, while the middle panels show the column density at a comparable resolution (4096^2). The lower panels show again the column density in the same regions, but at the highest resolution (32, 768^2). All the maps have a logarithmic colour scale.

also implies supersonic turbulence, hence fragmentation, inside the clumps. Nevertheless, unresolved clumps are still often interpreted and modelled as well-defined individual objects within the boundaries of their estimated (and unresolved) size. Rather than attempting a quantitative analysis of specific uncertainties that arise from the limited angular resolution of the observations, we use our synthetic observations to describe the problem graphically through two examples. We consider the same two example sources of Figs 4 and 9, and show their maps at the same physical scale (∼5 pc) as in the panels of Fig. 4. While in that figure, the sources were assumed to be at a distance of 12 kpc, in the top panels of Fig. 16 they are assumed to be at a distance of 2 kpc. At this higher spatial resolution, the 250 μm sources now appear to be connected to filaments at larger scales, and the original sources have broken into five sources (left-hand panel) and six sources (right-hand panel). The middle panels show that the structure of the column density maps, at the same resolution of Herschel's 250 μm beam, is even more fragmented than that of the dust emission maps. Finally, the bottom panels show the column density of the same regions at the maximum resolution of the simulation, 0.008 pc. While the source associated with a dominant 3D clump (a true protostellar source) appears to be like a complex conglomerate of dense cores and filaments (left-hand panel), the source that did not have a dominant 3D clump (a false protostellar source) has turned into a looser association of filaments and cores that barely stands out of the background.

These examples show that the limited angular resolution may easily lead to masses of compact sources many times larger than the actual cores and filaments they contain, as previously quantified in Padoan et al. (2020). It is important to appreciate that this effect is conceptually distinct from the blending along the line of sight, and so it may further contribute to the mass uncertainty. However, as the spatial resolution of the observations is increased and more fragments are resolved on the plane of the sky, the chance of blending of these smaller structures along the line of sight is decreased, so the mass estimates become more accurate, as demonstrated with synthetic higher angular resolution ALMA observations in Padoan et al. (2020).

8 CONCLUSIONS

We have used synthetic Herschel observations of a star-formation simulation on a scale of 250 pc to generate a catalogue of compact sources, with a range of distances between 2 and 12 kpc. The sources have been selected from the synthetic observations with the CuTEx code, following the same procedure as in the compilation of the Hi-GAL compact source catalogue. Our synthetic catalogue contains 51 831 compact sources and is an invaluable tool to interpret the nature of the 150 223 Herschel's compact sources in the Hi-GAL catalogue. This work serves both as a validation of the synthetic observations and as a first interpretation of the nature of the Hi-GAL sources.

To validate the synthetic observations, we have compared statistical distributions and correlations of size, mass and temperature of the synthetic sources with those of Herschel's sources from the Outer Galaxy catalogue. The comparison with the Inner Galaxy catalogue has been carried out in parallel to stress the importance of projection effects. We have found a good agreement with the Outer Galaxy catalogue, except for details related to the incompleteness of the stellar IMF in the simulation, and discrepancies in the comparison with the Inner Galaxy that can be understood as due to stronger projection effects there. We have then investigated the nature of the selected sources by searching for their counterparts in the 3D data cubes of the simulations. Our main results are listed in the following.

(i) The source masses overestimate the clump masses by an order of magnitude on average, due to line-of-sight projection. The estimated mass roughly corresponds to the whole mass, along the line of sight, while the most massive clump in the line of sight usually contains only about one 10th of the total mass on average.

(ii) A large fraction of sources classified as protostellar are likely to be starless at all values of temperature and 70 μm excess, as the 70 μm excess may be caused by stellar sources outside dense clumps.

(iii) We have proposed a method to partially discriminate between false and true protostellar sources based on the dependence of the 70 μm excess on the temperature.

(iv) We have found evidence of significantly stronger projection effects in the Inner Galaxy catalogue than in the Outer Galaxy and synthetic catalogues, from the mass distribution of the sources and their temperature distributions above and below Larson's mass-size relation. This would suggest that the mass of Hi-GAL sources in the Inner Galaxy may be on average over 20 times larger than the main 3D clumps in their lines of sight.

(v) At higher angular resolution, most Hi-GAL clumps should reveal a strongly fragmented structure, so future ALMA observations should confirm the important role of the projection effects demonstrated in this work.

Our synthetic catalogue will be used to interpret the results of follow-up studies of Hi-GAL sources, including single-dish or higher-resolution ALMA observations of molecular emission lines. In a future work, we will focus on the interpretation of the observational estimates of infall rates of massive clumps, as these have direct consequences for our understanding of the origin of massive stars. We can already conclude from the results of this work that the observed infall rates of massive clumps may have been overestimated by more than one order of magnitude, as the derived masses of Hi-GAL sources cannot be interpreted as the masses of individual clumps. The implications of our results for the formation of massive stars should also be addressed in future works.

ACKNOWLEDGEMENTS

We thank Davide Elia, Alessio Traficante, and Sergio Molinari for many useful explanations about the Hi-GAL data analysis method. We are grateful to the anonymous referee for a detailed and useful referee report. ZJL acknowledges financial support from China Scholarship Council (CSC) under grant no. 201606660003. PP and VMP acknowledge support by the Spanish MINECO under project AYA2017-88754-P. Computing resources for this work were provided by the NASA High-End Computing (HEC) Program through the NASA Advanced Supercomputing (NAS) Division at Ames Research Center. We acknowledge PRACE for awarding us access to Curie at GENCI@CEA, France. Storage and computing resources at the University of Copenhagen HPC centre, funded in part by Villum Fonden (VKR023406), were used to carry out part of the data analysis.

DATA AVAILABILITY

The full synthetic source catalogue (the description of which is in Appendix A), the surface brightness maps at all wavelengths, three colour images at ∼2 kpc distance, and the column density maps at a resolution of ∼0.06 pc (comparable to the pixel size of the surface brightness maps at ∼250 μm at a distance of ∼2 kpc) and at a full resolution of ∼0.008 pc can be obtained from a dedicated public URL (https://www.erda.dk/vgrid/massive-clumps/).

REFERENCES

Beuther H., Linz H., Henning T., 2012, A&A, 543, A88
Beuther H., Linz H., Henning T., 2013, A&A, 558, A81
Compiègne M. et al., 2011, A&A, 525, A103
Contreras Y. et al., 2018, ApJ, 861, 14
Duric N., 2004, Advanced Astrophysics. Cambridge Univ. Press, Cambridge
Elia D. et al., 2013, ApJ, 772, 45
Elia D. et al., 2017, MNRAS, 471, 100
Elia D. et al., 2021, MNRAS, 504, 2742
Fromang S., Hennebelle P., Teyssier R., 2006, A&A, 457, 371
Fuller G. A., Williams S. J., Sridharan T. K., 2005, A&A, 442, 949
Giannini T. et al., 2012, A&A, 539, A156
Gnedin N. Y., Hollon N., 2012, ApJS, 202, 13
Habing H. J., 1968, Bull. Astron. Inst. Neth., 19, 421
Haugbølle T., Padoan P., Nordlund Å., 2018, ApJ, 854, 35
Heyer M. H., Terebey S., 1998, ApJ, 502, 265
Heyer M. H., Brunt C., Snell R. L., Howe J. E., Schloerb F. P., Carpenter J. M., 1998, ApJS, 115, 241
Juvela M., 2019, A&A, 622, A79
Kippenhahn R., Weigert A., 1994, Stellar Structure and Evolution. Springer-Verlag, Berlin Heidelberg New York
Larson R. B., 1981, MNRAS, 194, 809
Lu Z.-J., Pelkonen V.-M., Padoan P., Pan L., Haugbølle T., Nordlund Å., 2020, ApJ, 904, 58
Markwardt C. B., 2009, in Bohlender D. A., Durand D., Dowler P., eds, ASP Conf. Ser. Vol. 411, Astronomical Data Analysis Software and Systems XVIII. Astron. Soc. Pac., San Francisco, p. 251
Mathis J. S., Mezger P. G., Panagia N., 1983, A&A, 500, 259
Molinari S. et al., 2016, A&A, 591, A149
Molinari S., Pezzuto S., Cesaroni R., Brand J., Faustini F., Testi L., 2008, A&A, 481, 345
Molinari S., Schisano E., Faustini F., Pestalozzi M., di Giorgio A. M., Liu S., 2011, A&A, 530, A133
Ossenkopf V., Henning T., 1994, A&A, 291, 943
Padoan P., Haugbølle T., Nordlund Å., 2012, ApJ, 759, L27
Padoan P., Pan L., Haugbølle T., Nordlund Å., 2016a, ApJ, 822, 11
Padoan P., Juvela M., Pan L., Haugbølle T., Nordlund Å., 2016b, ApJ, 826, 140
Padoan P., Haugbølle T., Nordlund Å., Frimann S., 2017, ApJ, 840, 48
Padoan P., Pan L., Juvela M., Haugbølle T., Nordlund Å., 2020, ApJ, 900, 82
Pan L., Padoan P., Haugbølle T., Nordlund Å., 2016, ApJ, 825, 30
Peretto N. et al., 2013, A&A, 555, A112
Salaris M., Cassisi S., 2005, Evolution of Stars and Stellar Populations. Wiley-VCH, Weinheim
Salpeter E. E., 1955, ApJ, 121, 161
Schaller G., Schaerer D., Meynet G., Maeder A., 1992, A&AS, 96, 269
Tan J. C., Beltrán M. T., Caselli P., Fontani F., Fuente A., Krumholz M. R., McKee C. F., Stolte A., 2014, in Beuther H., Klessen R. S., Dullemond C. P., Henning T., eds, Protostars and Planets VI. University of Arizona Press, Tucson, p. 149
Teyssier R., 2002, A&A, 385, 337
Teyssier R., 2007, Geophys. Astrophys. Fluid Dyn., 101, 199
Traficante A. et al., 2018, MNRAS, 477, 2220
Traficante A., Fuller G. A., Billot N., Duarte-Cabral A., Merello M., Molinari S., Peretto N., Schisano E., 2017, MNRAS, 470, 3882
Urquhart J. S. et al., 2014, MNRAS, 443, 1555
Weisz D. R. et al., 2015, ApJ, 806, 198
Wolfire M. G., Hollenbach D., McKee C. F., Tielens A. G. G. M., Bakes E. L. O., 1995, ApJ, 443, 152
Wyrowski F. et al., 2016, A&A, 585, A149
Yuan J. et al., 2018, ApJ, 852, 12

APPENDIX A: DESCRIPTION OF THE SYNTHETIC SOURCE CATALOGUE

The catalogue of synthetic compact sources is a text table with 34 columns that can be downloaded from https://www.erda.dk/vgrid/massive-clumps/. In the following, we describe the content of each column.

(i) Column[1], *ID*: running number (starting from 1 to 51 831).
(ii) Column[2], *SNAPSHOT*: the number of snapshots (839, 1107, and 1479, respectively).
(iii) Column[3], *DIRECTION*: the direction of the maps (0, 1, and 2 for x, y, and z directions, respectively).
(iv) Column[4], *DISTANCE*: the distance of the sources, in pc.
(v) Column[5], $X250$, [6], $Y250$: the map coordinates of the sources detected at 250 μm, in box units (from 0 to 1). Column [7], $Z250$: the line-of-sight coordinate of the maximum density in the 3D density cube, in box units (from 0 to 1).
(vi) Column[8], $FWHM\_X$, and [9], $FWHM\_Y$: the FWHM of the fitted bi-dimensional Gaussian along the x-axis and y-axis, in box units (from 0 to 1).

(vii) Column[10], *ANGLE*: the orientation angle of the major axis of the Gaussian with respect to the *x*-axis, in degree.

(viii) Column[11], *R*: the deconvolved radius of the source estimated at 250 μm, in pc.

(ix) Column[12], *F*70, [13], *DF*70, [14], *F*160, [15], *DF*160, [16], *F*250, [17], *DF*250, [18], *F*350, [19], *DF*350, [20], *F*500, and [21], *DF*500: the integrated flux of the source, in Jy, and noise of the flux, in Jy, estimated from the RMS for 70, 160, 250, 350, and 500 μm. Null value is -1.

(x) Column[22], $LAMBDA\_0$: the wavelength λ_0, in μm, from SED fitting for optical thick cases, and 0 for the optical thin cases.

(xi) Column[23], *T*, and [24], *DT*: temperature and its error from SED fitting, in K.

(xii) Column[25], M_SED, and [26], DM_SED: the mass of the clump and its error calculated from SED fitting with a greybody function, in $M_\odot$.

(xiii) Column[27], $F70\_GB$: the flux at 70 μm calculated from SED fitting with greybody function, in Jy.

(xiv) Column[28], *TYPE*: the classification of clumps: starless unbound, pre-stellar, and protostellar are denoted as 0, 1, and 2, respectively.

(xv) Column[29], M_MAIN: the mass of the main 3D clump along the light of sight, in $M_\odot$.

(xvi) Column[30], M_ALL: the mass of all the 3D clumps with density peaks $\geq$ 1 per cent of the maximum density peak along the light of sight, in $M_\odot$.

(xvii) Column[31], M_TOT: the total mass of column along the light of sight, in $M_\odot$.

(xviii) Column[32], $LOCAL\_GAS\_DENSITY$: the maximum of local gas densities around stars along the line of sight, in cm^{-3}. If the star is in the main clump, the value is positive, and if it is in the column, the value is negative. Null value is 0.

(xix) Column[33], $STAR\_AGE$: the maximum age of the stars along the line of sight, in Myr. If the star is in the main clump, the value is positive, and if it is in the column, the value is negative. Null value is 0.

(xx) Column[34], $STAR\_MASS$: the maximum mass of the stars along the line of sight, in $M_\odot$. If the star is in the main clump, the value is positive, and if it is in the column, the value is negative. Null value is 0.

APPENDIX B: DENSITY CRITERIA OF PROTOSTELLAR SOURCES

The value of the density threshold is justified by a comparison of the gas density distribution over the whole box with the density distribution sampled by the positions of the stars. The solid black line histogram in the three panels of Fig. B1 shows the gas density distribution computed from the final simulation snapshot analysed in this work at a uniform resolution of 2048^3 ($\sim$0.12 pc). The solid red line histogram in the same panels is the gas density distribution sampled at the same resolution, but only at the positions of the stars at the time of the final snapshot (the density is correctly centred at the stellar positions by using a higher-resolution extraction of the density field and then averaging the density around each star within a cell of 0.12 pc). Fig. B1 shows that the uniformly-sampled density distribution is bimodal, with the lower density peak at $\sim 0.01\,cm^{-3}$ corresponding primarily to hot gas, and the higher density peak at $\sim 3\,cm^{-3}$ primarily to colder gas. The density distribution sampled by the stars, instead, has three peaks. The two lower density peaks are the same as for the global distribution, while the highest density one, at $\sim 3 \times 10^4\,cm^{-3}$, corresponds to a characteristic density of protostellar cores.

Because protostellar cores occupy a tiny fraction of the computational volume, their density is sampled by the global gas distribution (black line) only at extremely low probability, not shown in the plots of Fig. B1. The densest peak is instead visible in the density distribution sampled by the stars because the fraction of stars that are found within their parent protostellar cores is orders of magnitude in excess of the fraction of the total volume occupied by the cores. The fact that the highest density peak is associated with protostellar cores is confirmed by the comparison of the three panels in Fig. B1, where the density distribution sampled only by stars in a limited age range is shown in each panel by the blue-line histogram. Considering only stars younger than 1 Myr (left-hand panel), one can see that they sample only density around the peak, while the stars older then 5 Myr (right-hand panel) sample only densities outside of that peak. Furthermore, when stars are older (right-hand panel) they sample well the global bimodal density field, meaning that their positions not only do not correspond to dense protostellar cores any longer, but are also completely random with respect to the density field. In other words, their position no longer depends on the local density, and hence by measuring their local density, we are measuring the density distribution in the simulation box.

Having identified the highest density peak as due to protostellar cores, the position of the minimum between the two denser peaks in the density distribution sampled by the stars, at $\sim 10^3\,cm^{-3}$, can be adopted as the threshold between the overall density distribution and the density of protostellar cores. This justifies the density threshold used in our criterion to define embedded stars and thus to discriminate between true and false protostellar sources.

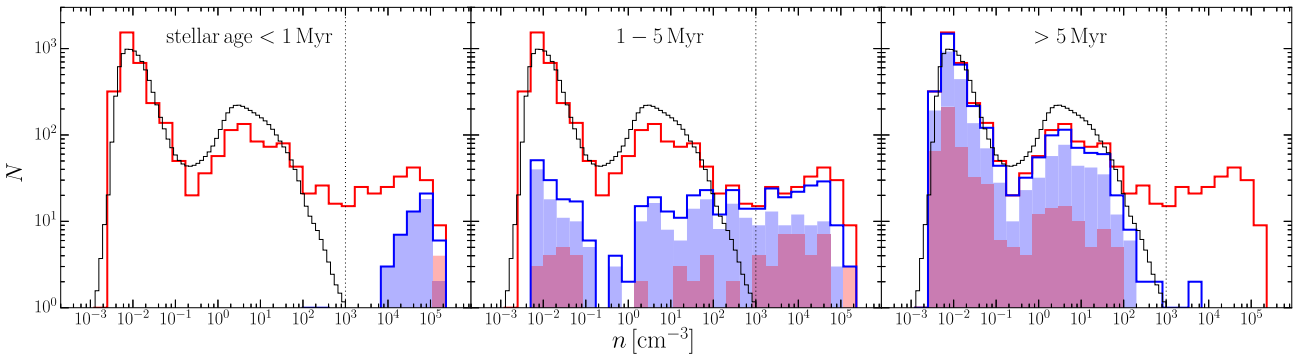

Figure B1. Gas density distributions sampled uniformly at a 2048^3 resolution in the last snapshot of the simulation analysed in this work, at $t = 34.2$ Myr (solid black line) and sampled at the positions of the stars (solid red line). The density at the position of a star is the local mean gas density calculated in a ∼0.12 pc box centred around the star. The density around the stars is also plotted as a blue step histogram after separating the stars into three categories based on their age, <1, 1–5, and > 5 Myr (left-hand, middle, and right-hand panels, respectively). From each of these age intervals, we also show two sub-samples corresponding to stars with mass $< 5\,M_\odot$ (blue filled histogram), and with mass $> 8\,M_\odot$ (red filled histogram). The overlap region of the two filled histogram is in purple.

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