A&A, 706, A242 (2026)
https://doi.org/10.1051/0004-6361/202452293

Astronomy & Astrophysics

Catalog of ^{13}CO clumps from FUGIN in the Milky Way at l = 10°–50°

Sheng Zheng[1], Xuejiao Pan[1], James Stuart Urquhart[2], Xiaoyu Luo[1,4,⋆], Yao Huang[1,⋆], Zhibo Jiang[3], Zhiwei Chen[1], Shuguang Zeng[1], Xiangyun Zeng[1], and Junjie Zhang[1]

[1] Center for Astronomy and Space Sciences, China Three Gorges University, Yichang 443000, China
[2] Centre for Astrophysics and Planetary Science, University of Knet, Canterbury CT2 7NH, UK
[3] Purple Mountain Observatory, Chinese Academy of Sciences, Nanjing 210023, China
[4] College of Electrical Engineering and New Energy, China Three Gorges University, Yichang 443000, China

Received 18 September 2024 / Accepted 10 December 2025

ABSTRACT

Context. Since stars and star clusters emerge from the gravitational collapse of clumps and cores, studying molecular clumps is fundamental to understanding the processes of star formation. The FOREST Unbiased Galactic Plane Imaging (FUGIN) survey offers insights into the distribution of clumps and physical properties across different environments, aiding in studies of environmental effects, such as the location within the galaxy on star formation.
Aims. This study aims to produce a catalog of clumps from the FUGIN survey to understand the complete mechanism of high-mass star formation in giant molecular clouds (GMCs). We use the catalog to analyze the physical properties of clumps in high-mass star-forming regions, enhancing our understanding of how different environments impact the star-formation process.
Methods. Our process for the detection and verification of ^{13}CO clumps in the FUGIN survey comprised two steps. First, the source extraction code FacetClumps was used to detect as many molecular clump candidates as possible from the FUGIN ^{13}CO data. Second, a trained and validated semi-supervised deep learning model, SS-3D-Clump, was applied to verify these candidates, providing confidence levels for the clumps and filtering out false candidates to enhance the accuracy of the detection results.
Results. The resulting catalog containing 23 150 clumps extracted from the ^{13}CO (J=1–0) data covers the first quadrant ($10° \leq l \leq 50°$, $|b| \leq 1°$). By matching with CHIMPS and inheriting the distances of the matched CHIMPS clumps, we found that the sizes of the FUGIN clumps range from 0.1 to 3 pc, demonstrating that the dense structures belong to the clump scale. The catalog achieves an 80% completeness level above 466 K km s^{-1}.
Conclusions. The proposed two-step approach effectively integrates clump detection algorithms with semi-supervised deep learning, achieving an accuracy comparable to manual verification and thereby improving the extraction of clumps from large-scale survey data. The resulting clump catalog enables the analysis of the physical properties of clumps in high-mass star-forming regions, contributing to a better understanding of environmental influences on clump formation and the star formation process.

Key words. molecular data – techniques: image processing – ISM: molecules

1. Introduction

The study of molecular clumps is crucial for advancing theories of star formation since stars and star clusters are formed through the gravitational collapse of these clumps or cores (Krumholz & McKee 2005; Zinnecker & Yorke 2007; Rathborne et al. 2009; Alves de Oliveira et al. 2014; Takekoshi et al. 2019; Rigby et al. 2019; Yoo et al. 2023). Star-forming regions are messy and chaotic environments, with structures existing on many scales (Wurster & Rowan 2023). The entire region is typically referred to as a cloud; dense regions embedded within the cloud are clumps and the very dense regions in the clumps are cores (Alves et al. 2007). The clumps are defined as compact (~1 pc) and dense (~10^4 H_2 cm^{-3}) structures (e.g., Williams et al. 2000; Zhang et al. 2009; Heyer & Dame 2015; Ohashi et al. 2016; Motte et al. 2018; Takekoshi et al. 2019). The mass distribution of clumps provides essential information not only on the mechanisms that influence clump formation, evolution, and destruction (e.g., Rosolowsky 2005; Colombo et al. 2014; Faesi et al. 2016), but also on the crucial factor in constraining theories of star formation (e.g., Alves et al. 2007; Clark et al. 2007; Liu et al. 2022). To determine the role played by environmental effects in the formation and evolution of molecular clumps and how these, in turn, affect star formation, we need to study the properties of clumps in different environments.

Many systematic CO survey projects have focused on the inner and outer Galaxy, such as the Galactic Ring Survey (GRS; Jackson et al. 2006), the CO Heterodyne Inner Milky Way Plane Survey (CHIMPS; Rigby et al. 2016), the Structure, Excitation, and Dynamics of the Inner Galactic Inter-Stellar Medium Survey (SEDIGISM; Schuller et al. 2017), the Milky Way Imaging Scroll Painting Survey (MWISP; Su et al. 2016, 2019), the FOREST Unbiased Galactic Plane Imaging Survey with Nobeyama 45 m telescope (FUGIN; Umemoto et al. 2017), and the Outer Galaxy High-Resolution Survey (OGHReS; Urquhart et al. 2024). These surveys, with their different levels of sensitivity and spatial resolution, can help us to detect the distribution of the molecular component at different scales (Benedettini et al. 2021) from dense clouds and filamentary structures (André et al. 2014) to pre-stellar clumps and young stellar objects.

⋆ Corresponding authors: vastlxy@163.com; huangyao@ctgu.edu.cn

In the northern inner Galactic plane, GRS (Jackson et al. 2006) covers the region between $18° \leq l \leq 55.7°$ and $|b| \leq 1°$ in ^{13}CO (J=1–0) at 46″ resolution, providing a benchmark in high-resolution, unbiased spectral imaging. Rathborne et al. (2009) employed the ClumpFind algorithm (Williams et al. 1994) to identify 829 molecular clouds and 6124 clumps, finding that clouds within the 5 kpc ring typically have warmer temperatures, higher column densities, larger areas, and a larger number of clumps. CHIMPS (Rigby et al. 2016) carried out a ^{13}CO and C^{18}O (3–2) survey with the 15-m *James Clerk Maxwell* Telescope (JCMT), covering a large portion of the GRS survey region ($27.8° \leq l \leq 46.2°$ and $|b| \leq 0.5°$) with a higher angular resolution of 15″. Based on the CHIMPS ^{13}CO data, Rigby et al. (2019) obtained the first source catalog identified by FellWalker (Berry 2015) and found no significant systematic variations in the physical properties of the sources (e.g., mass, column density, or virial parameter) across the probed range of Galactocentric distances. In the outer Galaxy, OGHReS (Urquhart et al. 2024) is a systematic, high-resolution survey (i.e., $\theta_{\rm FWHM} \approx 30''$) in ^{12}CO ($J$=2–1) and ^{13}CO ($J$=2–1), covering the region between $180° \leq l \leq 280°$ and approximately 1° in b. Urquhart et al. (2024) used a subset of the data ($250° \leq l \leq 280°$, $-2° \leq b \leq 1°$) to verify the velocities and distances assigned to the Hi-GAL clumps in this part of the Galaxy (Mège et al. 2021) and to refine their physical properties (Elia et al. 2021), which is crucial for understanding how different environmental conditions affect star formation.

The FOREST Unbiased Galactic Plane Imaging (FUGIN, Umemoto et al. 2017) Survey utilized the Nobeyama 45-meter radio telescope to simultaneously observe the emission spectral lines of three CO isotopologues: ^{12}CO, ^{13}CO, and C^{18}O (J=1–0) in the first quadrant region ($10° \leq l \leq 50°$, $|b| \leq 1°$, hereafter QI, shown as the red area in Fig. 1) and the third quadrant region ($198° \leq l \leq 236°$, $|b| \leq 1°$, hereafter QIII, shown as the green area in Fig. 1) of the Galaxy. The project achieved 20″ angular resolution allows to resolve down to 0.19 pc at 2 kpc distances, which can provide detailed knowledge of the structure and physical properties of the nearest GMCs (Heyer & Dame 2015), ranging from the core scales (~0.1 pc) to cloud scales (10–100 pc). The survey maps areas include the spiral arms (Perseus, Sagittarius, Scutum, and Norma) and star-forming regions (e.g., W31, W33, W39, M16, M17; Beuther et al. 2011; Khan et al. 2022; Kerton et al. 2013; Tremblin et al. 2014; Chen et al. 2021), the bar structure, and the molecular gas ring. It is essential to understand the full mechanism of high-mass star formation in giant molecular clouds (GMCs), particularly by examining the physical properties of cores and clumps in high-mass star-forming regions through large-field surveys (Takekoshi et al. 2019). Comparing these regions, along with the outer Galaxy where the metallicity is much lower (Smartt & Rolleston 1997) and their bar-swept radii, will expand our understanding of the impact of the environment on the star formation process (Eden et al. 2020).

With the progress of Galactic survey projects, various molecular clumps detection algorithms have emerged (Rosolowsky et al. 2008; Berry 2015; Luo et al. 2022; Jiang et al. 2023). The Dendrograms algorithm (Rosolowsky et al. 2008) is well-suited to representing the hierarchical structure of isosurfaces in molecular line data cubes, illustrating variations in topology as contour levels change (Rani et al. 2023). This method has been widely used in conjunction with continuum, atomic hydrogen (HI), and molecular line data (Takekoshi et al. 2019; Nakanishi et al. 2020; Zhang et al. 2021). The FacetClumps algorithm (Jiang et al. 2023) utilizes morphological methods to extract signal regions from raw data and applies the Gaussian facet model

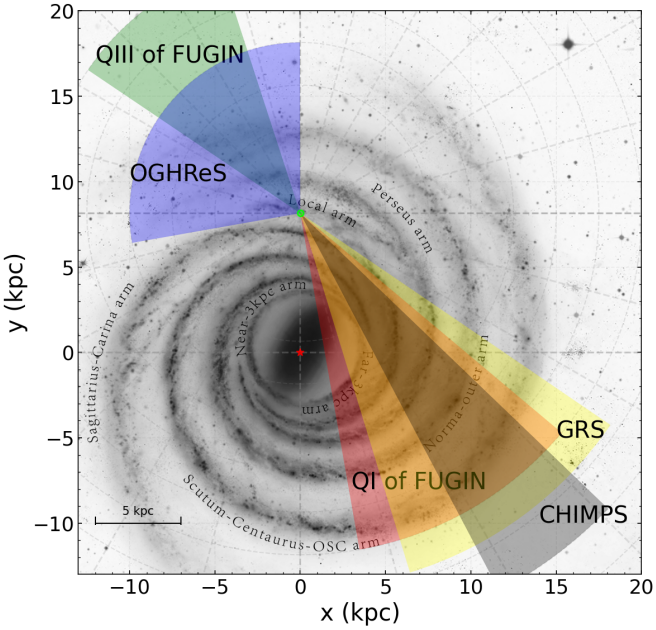

Fig. 1. Area of the Galaxy covered by the FUGIN survey. Face-on view of an imaginary Milky Way (credit: R. Hurt, NASA/JPL-Caltech/SSC). The Galactic center (red asterisk) is at (0, 0) and the Sun (green filled circle) is at (0, 8.5).

(Ji & Haralick 2002) and extremum theory of multivariate functions to locate the centers of clumps. This algorithm enhances the accuracy of region segmentation by dividing signal regions into local areas and subsequently clustering these local regions to the centers of clumps based on connectivity and minimum distance.

A manual confirmation is crucial for the clump candidates obtained using the FacetClumps algorithm noted above to eliminate false positives and ensure the reliability of clump targets in scientific analysis (Rigby et al. 2019). However, large-scale survey projects often yield many clump candidates, making extensive manual verification impractical. Therefore, there is an urgent need for an automated clump verification algorithm to substitute for manual inspection. Luo et al. (2024a) attempted to integrate deep learning into the clump verification process as deep learning has shown outstanding performance in classifying galaxies (Cheng et al. 2020; He et al. 2021), developing the semi-supervised deep clustering algorithm called SS-3D-Clump. The model utilizes a 3D convolutional neural network (3D CNN) to extract features of clumps and classify them using a constrained K-means approach to obtain a pseudo-label. Subsequently, it uses these pseudo labels as supervision to update the weights of SS-3D-Clump. Therefore, SS-3D-Clump can leverage unlabeled samples via a semi-supervised learning to overcome limitations in supervised learning for limited labeled samples and enhance the generalization ability.

This study builds upon the Facet-SS-3D-Clump framework introduced in Luo et al. (2024b), which established a workflow for detecting and verifying ^{13}CO clumps in the MWISP survey. The workflow first uses FacetClumps to identify clump candidates, and then applies SS-3D-Clump for verification. In the present work, we extend the framework by applying it to a broader Galactic region. The paper is organized as follows. Section 2 introduces the observations of FUGIN. Section 3 describes the molecular clump extraction algorithm and the use of SS-3D-Clump to verify clump candidates. In Section 4, we present

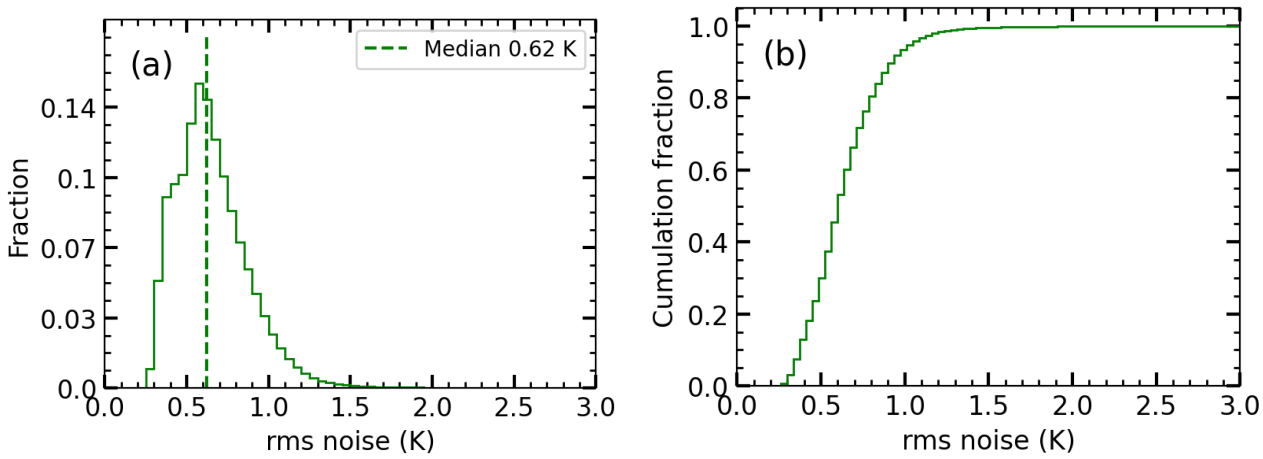

Fig. 2. Noise level of the ^{13}CO data. Panel a: histogram of the rms noise levels. Panel b: cumulative distribution of the rms noise levels.

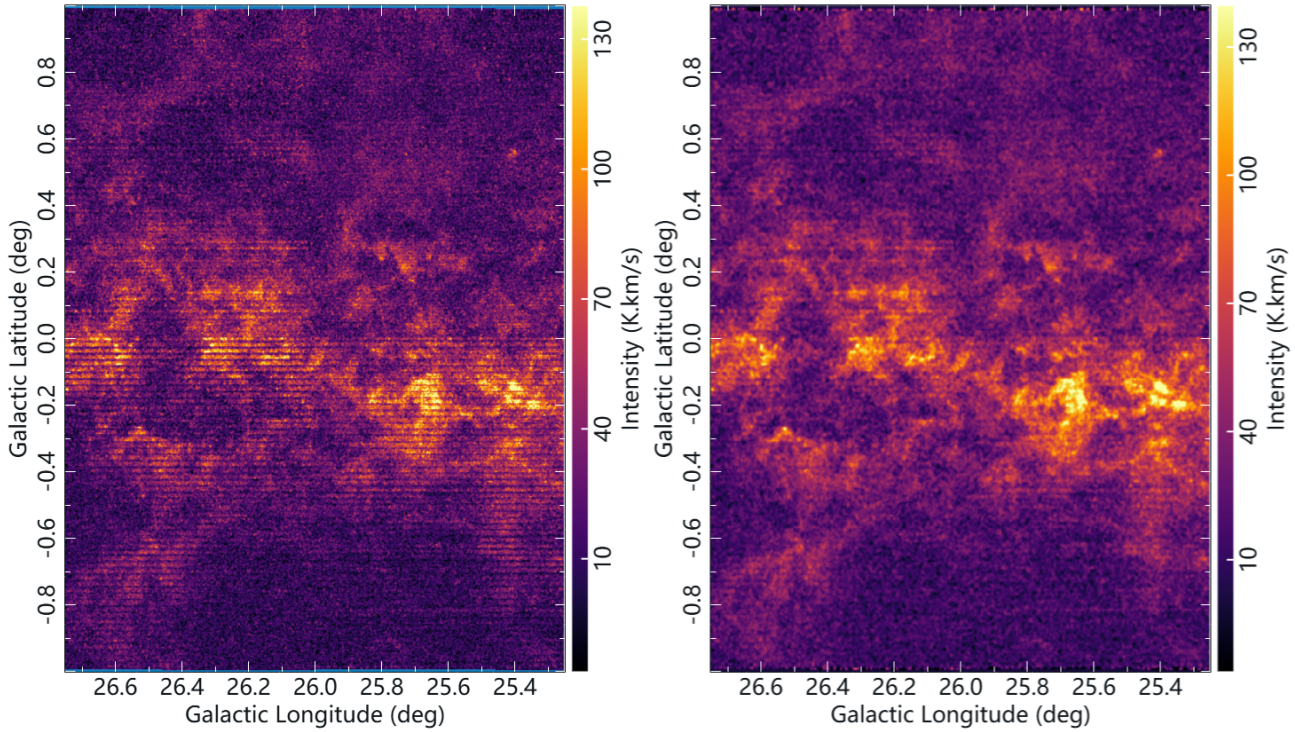

Fig. 3. Selected region of the Milky Way ($25.25° \leq b \leq 26.75°$, $-1° \leq l \leq 1°$, and $0 \leq v \leq 130$ km/s). Left panel displays the integrated map of the S/N data, revealing noticeable horizontal stripes. Right panel shows the result after interpolation, effectively removing the stripes.

the catalog containing 23 150 ^{13}CO clumps extracted from QI and provide a brief statistical analysis of the spatial distribution of clumps and comparison with CHIMPS clumps. Section 5 presents the conclusion.

2. Data

2.1. Data introduction

The FUGIN project utilizes the new multi-beam FOREST system, installed on the Nobeyama 45 m telescope to support legacy projects. This four-beam receiver system on the 45 m Telescope is an integrated, dual-polarization, sideband-separating SIS receiver. The four beams are arranged in a 2×2 configuration with approximately a 50″ grid, and each beam has angular resolution of around 14″ at 115 GHz. The main beam efficiencies at 86, 110, and 115 GHz are 0.56 ± 0.03, 0.45 ± 0.02, and 0.43 ± 0.02, respectively. FUGIN conducts Galactic plane CO observations[1] with a 20″ angular resolution for the ^{12}CO and 21″ angular resolution for the ^{13}CO and C^{18}O (J=1–0). The effective velocity resolution is 1.3 km s^{-1} at 115 GHz. The root mean square (rms) distribution is shown in Fig. 2, and the median noise level for the ^{13}CO spectra is 0.62 K.

2.2. Preprocessing of ^{13}CO data

We first selected a local region in the first quadrant of the Milky Way: $25.25° \leq l \leq 26.75°$, $-1° \leq b \leq 1°$, and 0 km s^{-1} $\leq v \leq$ 130 km s^{-1}, as shown in Fig. 3. The left panel of Fig. 3 shows the

[1] https://nro-fugin.github.io/release/

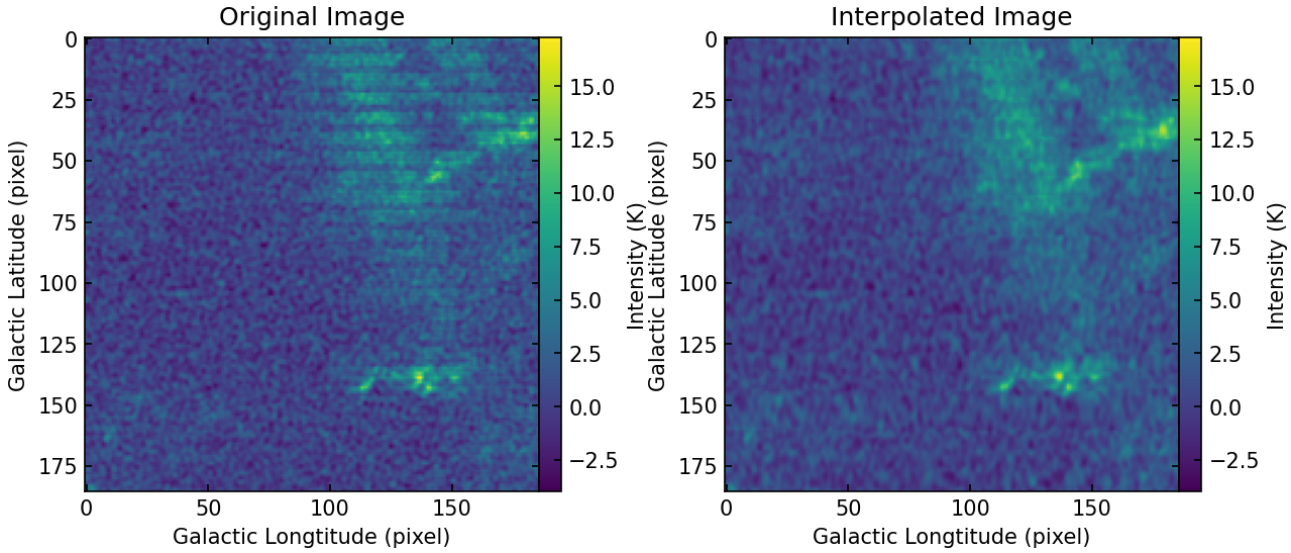

Fig. 4. Example illustrating the interpolation of stripes in the data. Left panel shows the original data with stripes. Right panel shows the data after the interpolation processing.

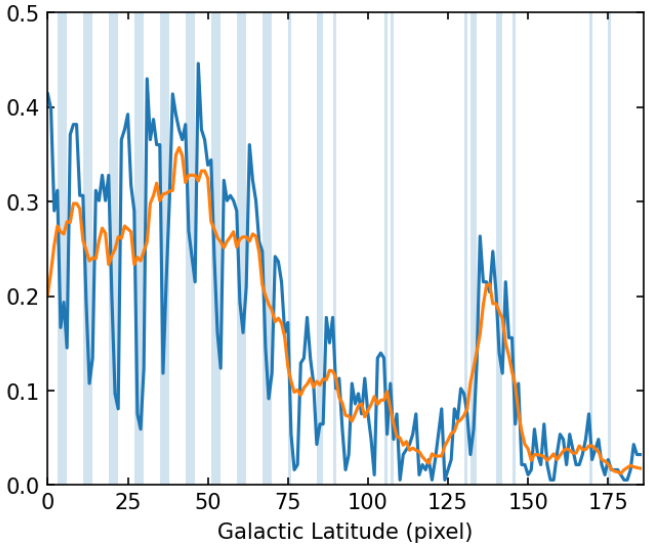

Fig. 5. Horizontal projection curve of the original image after binarization is shown, where the blue curve represents the original projection curve, and the orange is the result of applying a moving average to the original curve. The blue shaded areas indicate the regions of stripes in the image. Note: the vertical axis in the figure has no actual physical significance; it only represents the ratio of pixels exceeding 3×rms to the total number of pixels in the horizontal direction.

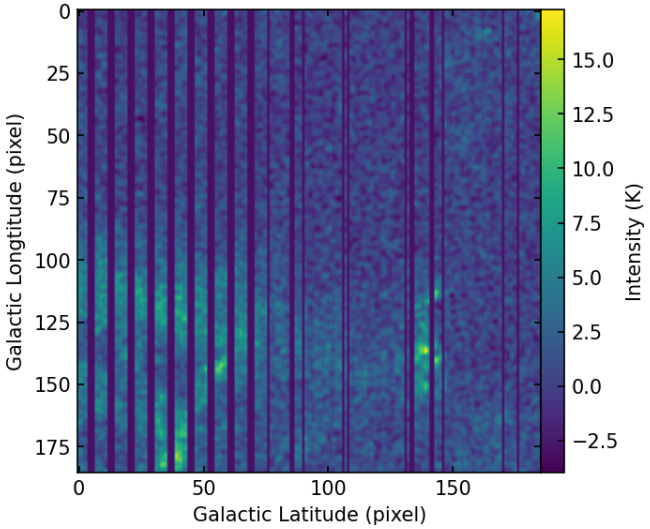

Fig. 6. Stripe regions are overlaid onto the original data and for easier observation, the X and Y axes of the image have been swapped.

integrated map of the region's signal-to-noise ratio (S/N) data in velocity. The map reveals noticeable horizontal stripes in the data. The right panel of Fig. 3 displays the result after interpolating the striped data. It can be seen from the map that the stripes have been reduced, although some faint residual patterns remain. As an example, we use a section of data from one velocity channel (see the left panel of Fig. 4), which clearly shows horizontal stripes. The specific processing steps are as follows:

The first step is to apply a threshold of 3×rms to the image to generate a binary map. A horizontal projection (i.e., integration along the Galactic longitude direction) is then performed on this binary map, resulting in the blue curve (shown in Fig. 5).

In the second step, a moving average filter with a fixed window size (9 pixels) is applied to the blue curve to obtain the orange curve. This filter effectively suppresses high-frequency noise while preserving the overall trend of the data by computing the arithmetic mean within the window. Stripe regions are identified based on the peak-valley structures of the two curves; specifically, areas where the orange curve significantly exceeds the blue curve are considered candidate stripe regions, as these correspond to weakened intensity features. These regions are highlighted as blue-shaded areas in Fig. 5 and the final detected stripe regions are mapped back onto the original data, as shown in Fig. 6.

The third step involves using the data surrounding the detected stripe regions to interpolate and fill in the stripes using the cubic spline interpolation algorithm (Ooyama 2002; Karpfinger 2022). The cubic spline method fits a set of piecewise cubic polynomials between data points, ensuring the continuity of the first and second derivatives across segment boundaries,

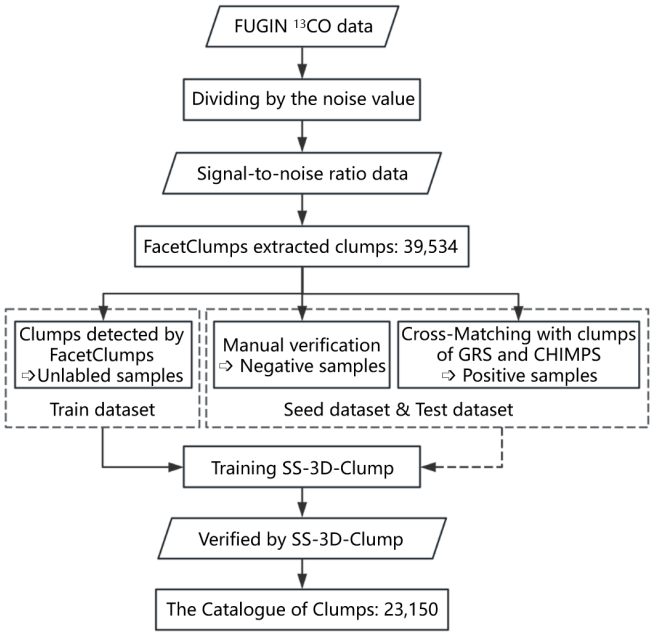

Fig. 7. Workflow for obtaining a clumps catalog.

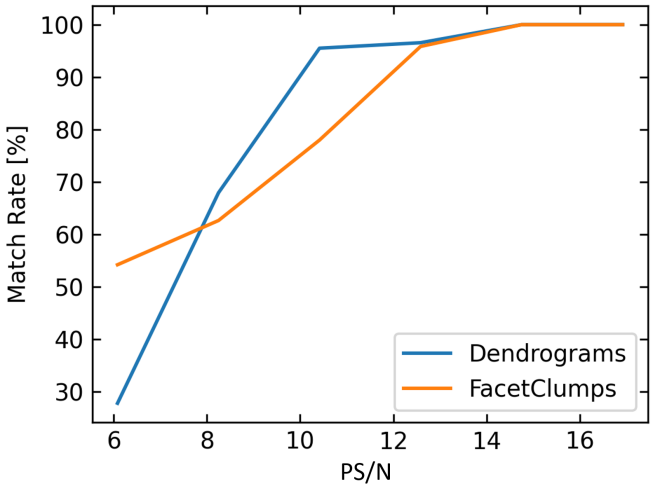

Fig. 8. Matching ratios of FacetClumps and Dendrograms in the three regions data as a function of the PS/N of the clumps. The orange curve represents the fraction of FacetClumps that have a corresponding Dendrogram clump, while the blue curve represents the fraction of Dendrogram clumps that have a corresponding FacetClump clump.

which results in a smooth and natural-looking reconstruction. The final processed result is shown in the right panel of Fig. 4. After processing the striped data in FUGIN, the map of ^{13}CO integrated with velocity from -9 to 128 km s^{-1} for the FUGIN (Umemoto et al. 2017) in QI is shown in Appendix D (Fig. D.1). It was drawn using the Cube Analysis and Rendering Tool[2] for Astronomy (Carta, Angus Comrie et al. 2021).

3. Generation of a ^{13}CO clump catalog from FUGIN

Stars are formed within interstellar molecular clouds, primarily composed of molecular gases. The hierarchical structure of the molecular phase of the ISM can be broadly categorized into three structures: clouds, clumps, and cores (Alves et al. 2007; Wurster & Rowan 2023). Clumps refer to regions of increased density within more giant clouds. They can also be identified as continuous areas in the l–b–v space (where l, b, and v represent Galactic longitude, latitude, and velocity, respectively). Cores are highly dense regions within clumps. They result from the fragmentation of larger clumps and may eventually collapse to form individual stars or star clusters (Blitz & Williams 1999; Bergin & Tafalla 2007). This study defines the compact structure traced by ^{13}CO as a "clump".

The ^{13}CO clumps detection and verification process of FUGIN is illustrated in Fig. 7, which comprises two major steps. First, FacetClumps (Jiang et al. 2023) was used to detect and obtain molecular clump candidates from FUGIN ^{13}CO data. As mentioned by Medina et al. (2019), it is doubtful that a false detection would randomly appear in the same position in other independent surveys. Therefore, we assumed that any candidate with a counterpart in a published catalog was real. We labeled the candidates matched with clumps in GRS (Rathborne et al. 2009) and CHIMPS (Rigby et al. 2019) as real sources, which then served as the seed set in the training process of SS-3D-Clump (Luo et al. 2024a). Negative samples were obtained through manual verification (see the right panel of Fig. C.1 in Appendix C). Second, the trained and stabilized SS-3D-Clump is applied to verify the candidates, obtaining confidence levels (i.e., predicted probabilities between 0 and 1) for the molecular clumps. After the process above, we obtained a catalog of molecular clumps, including their positions, peak, integrated intensity, and confidence levels associated with the clumps.

3.1. Clumps extraction

We observed a noticeable and uneven noise level in FUGIN data due to the non-uniform sensitivity of the survey, which could lead to overlooking clumps in regions with noise below the average level and, conversely, misidentifying false clumps as real ones in high-noise areas. Following the approach of Rigby et al. (2016), who conducted source extraction on the CHIMPS survey on S/N data cubes, these were obtained by dividing the observational data by the corresponding noise. We also applied FacetClumps to the S/N data to detect molecular clumps and generate masks. We selected three local regions along the Galactic longitude in QI, spanning different density ranges. Using these three representative regions, we compared and analyzed the detection results of the FacetClumps and Dendrograms algorithms as a benchmark against existing mature algorithms.

Figure 8 shows the matching ratios between the FacetClumps and Dendrogram extractions in the three regions as a function of the peak S/N (PS/N) of the clumps. The matching ratio is defined as the fraction of clumps in one catalog that have at least one spatially overlapping counterpart in the other. Specifically, the orange curve represents the fraction of FacetClumps that have corresponding Dendrogram clumps, while the blue curve represents the fraction of Dendrogram clumps that have corresponding FacetClump clumps. As shown in Fig. 8, when the PS/N of the clumps rises above 12, the matching rate exceeds 90% across the data of three regions with different densities, and the detection results of the two algorithms become consistent. Detailed experiments and results are presented in Appendix A.

FacetClumps[3] (Jiang et al. 2023) employs the Gaussian facet model to fit the local surfaces and determine clump centers through the extremum determination theorem of multivariate

[2] https://carta.readthedocs.io/en/4.1/

[3] https://github.com/JiangYuTS/FacetClumps

functions. Based on the identified clump centers, FacetClumps clusters the regions near the centers by considering connectivity and minimum distance, thereby obtaining molecular clumps. Since the detection data is based on S/N data, the background rms equals unity. The parameter *Threshold* is the minimum intensity used to truncate the signal, defined by its relationship to the background rms and set to 4. Another parameter, *SRecursionLBV*, is used to determine the minimum area of a region in the spatial direction (*SRecursionLB*) and the minimum length of a region in the velocity channels (*SRecursionV*) when a recursion terminates. The parameter breaks down the volume parameter used in other algorithms, such as Dendrograms (Rosolowsky et al. 2008) and FellWalker (Berry 2015), into the minimum area in the spatial direction and the minimum extent in the velocity direction. It helps further eliminate false detections caused by elongated noise, as such noise can still meet the minimum volume parameter criteria. *SRecursionLB* is set to 25 square pixels, corresponding to an angular size of approximately 0.5 square arcminutes, and *SRecursionV* is set to 5 pixels, corresponding to a velocity of 3.25 km s^{-1}. The other parameters of FacetClumps are set to the default values of the algorithm.

3.2. The seed dataset and test dataset

There are roughly three methods to obtain high-confidence molecular clumps: (1) using different algorithms to detect the same dataset. Thus, if multiple algorithms detect a target, the probability that this target is real will be high; (2) using different data for validation. Thus, if a molecular clump detected in the ^{13}CO data is also detected in the C^{18}O data, the probability that the corresponding ^{13}CO clump is real will also be high; and (3) by matching with other published catalogs. Thus, if a molecular clump can match with an existing source catalog, the probability that it is real will also be high, as reported in Medina et al. (2019).

This paper adopts the same strategy as in Medina et al. (2019), based on establishing matches with published catalogs to obtain a reliable sample of molecular clumps, which would then be considered a robust sample. We used the topcat[4] tool from the Starlink package to match the clump candidates with the clumps in GRS (Rathborne et al. 2009) and CHIMPS (Rigby et al. 2019). The matching was performed based on the centroid positions of the clumps, requiring deviations of less than 1′ in Galactic longitude and latitude and less than 2 km s^{-1} in velocity. In the overlapping region between FUGIN and GRS, which contains 16 292 FUGIN clumps and 5868 GRS clumps, we have 3242 GRS clumps that were successfully matched with at least one FUGIN counterpart. Between FUGIN and CHIMPS, which have 7171 and 4473 clumps in the overlapping region, respectively, there were 2476 CHIMPS clumps found to have FUGIN counterparts. The strict matching criteria were adopted to ensure a high-confidence sample of molecular clumps for subsequent analysis. Figure 9 shows the results of two matched molecular clumps in CHIMPS and FUGIN. The first column shows two clumps from CHIMPS, while the second column shows the corresponding matched clumps from FUGIN.

To obtain sample of false detections (i.e., incorrectly identified molecular clumps), we first need to identify the clumps that can be confidently regarded as real based on the following criteria: 1. the molecular clumps must first exhibit enhanced intensity in the local region, showing significant intensity peaks on the integrated maps in three directions (Galactic longitude, Galactic latitude, and velocity); and 2. in the spectral line information of the velocity direction, the clump's average spectrum and peak spectrum must display a Gaussian profile.

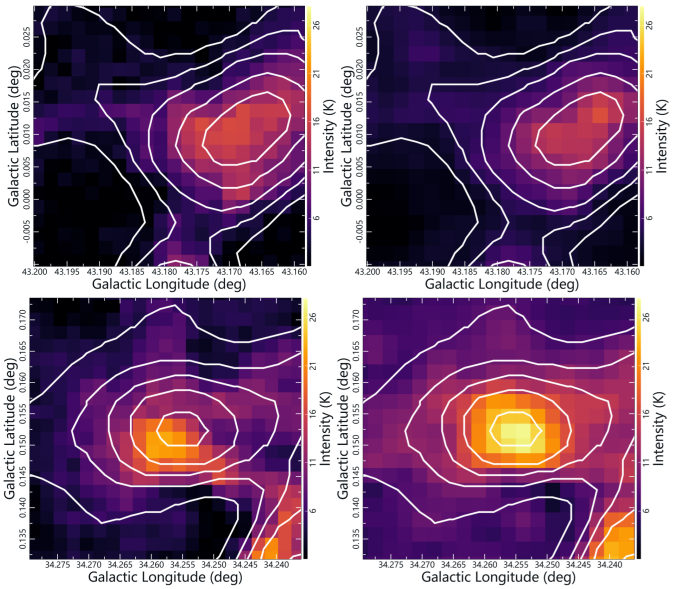

Fig. 9. Examples of two matched clumps in CHIMPS and FUGIN are shown. The first column shows two clumps from CHIMPS, while the second column shows the corresponding matched clumps from FUGIN. The contours in the figure are drawn on the intensity map of the FUGIN clumps and overlaid onto the map of the CHIMPS clumps.

Clumps that satisfy these two conditions are considered real by manual verification and the remaining ones are regarded as false detections. For example, Fig. C.1 in Appendix C shows two examples of clumps verified as "true" and as "false."

We compiled a test dataset containing 2000 true detections and 1000 false detections. The true detections are the ones that were matched with clumps in the GRS and CHIMPS. The false detections were identified through visual inspection and confirmed not to correspond to any real clump structures. Additionally, we separated some samples from the test dataset to form a seed dataset, which includes 300 real and 300 false detections. We note that the test dataset does not participate in the model training. The seed sample set was used to ensure proper model convergence. The primary data used for model training consists of many unlabeled sample data. The model principles and the role of the data will be explained in detail in the next section.

3.3. Verification by SS-3D-Clump

SS-3D-Clump[5] is a semi-supervised learning method designed to verify molecular clumps by Luo et al. (2024a). It consists of three main modules: a feature extraction module utilizing a 3D CNN, a pseudo-label acquisition module employing an unsupervised clustering algorithm and a fully connected network (FCN) for label prediction. The input to SS-3D-Clump is a 3D cube with the size set to 30 px × 30 px × 30 px, centered on the clump candidate. Each cube is extracted from the ^{13}CO spectral line data. During training, the model is fed shuffled mini-batches sampled from the full training dataset in each iteration, and the entire dataset is reshuffled at the start of every epoch to enhance generalization. The SS-3D-Clump outputs a confidence value between

[4] https://www.star.bristol.ac.uk/mbt/topcat/

[5] https://github.com/Luoxiaoyu828/SS-3D-Clump

0 and 1, indicating the likelihood that a given candidate is a real clump.

From an intuitive perspective, if two clumps share similar physical characteristics, they are expected to remain close in the learned feature space after being processed by a feature extraction model. This means their separation (i.e., feature distance) will be small and their similarity high. The SS-3D-Clump framework is built on this principle. Initially, the trainable parameters of the feature extraction and classification networks in the SS-3D-Clump are randomly initialized and tasked with learning a feature mapping that transforms clump candidates into a high-dimensional space where structurally similar clumps are grouped together. To generate training labels without requiring manual annotation, we adopted an unsupervised strategy using the K-means clustering algorithm (Han et al. 2012; Ay et al. 2023), whose initial cluster centers were determined on the basis of the high-confidence seed samples. In each training iteration, K-means is applied to the extracted features to assign each sample to one of two clusters, interpreted as pseudo-labels. It should be noted that the two clusters obtained by the K-means algorithm are not intended to represent the physical subclasses of molecular clumps. Instead, they provide a fundamental distinction between "normal" and "potentially anomalous" samples, serving as pseudo-labels for the subsequent semi-supervised classification. The number of clusters was empirically determined based on the stability and convergence behavior of the SS-3D-Clump, as increasing the cluster number often resulted in fragmented or unstable groupings without improving the detection accuracy, while simply using two clusters ensured a more stable convergence and consistent classification results.

These pseudo-labels are then used to construct a classification loss function, enabling supervised-style training via backpropagation. During the early stages of training, the model's feature extraction capability is still under development and the resulting pseudo-labels are often unstable. As training progresses, however, the model gradually learns to represent similar clumps with increasingly similar features, leading to more consistent and accurate cluster assignments. To further guide the convergence process, we introduced a small number of high-confidence "seed" samples with fixed labels to constrain the K-means algorithm (Basu et al. 2002) and accelerate model learning.

Some statistical metrics (Cunha et al. 2024), such as precision (P), recall (R), F1 score (F_1), and accuracy (Acc), were used to quantify the performance of SS-3D-Clump in the test dataset. The calculation formulas of R, P, F_1, and Acc are as follows:

$$R = \frac{TP}{TP + FN}, \tag{1}$$

$$P = \frac{TP}{TP + FP}, \tag{2}$$

$$F_1 = 2 \times \frac{P \times R}{P + R}, \tag{3}$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}, \tag{4}$$

where TP (true-positive) is the number of sources predicted to be true and actual is true, FP (false-positive) is the number of sources predicted to be true and actual is false, TN (true-negative) is the number of sources predicted to be false and actual is false, and FN (false-negative) is the number of sources predicted to be false and actual is true.

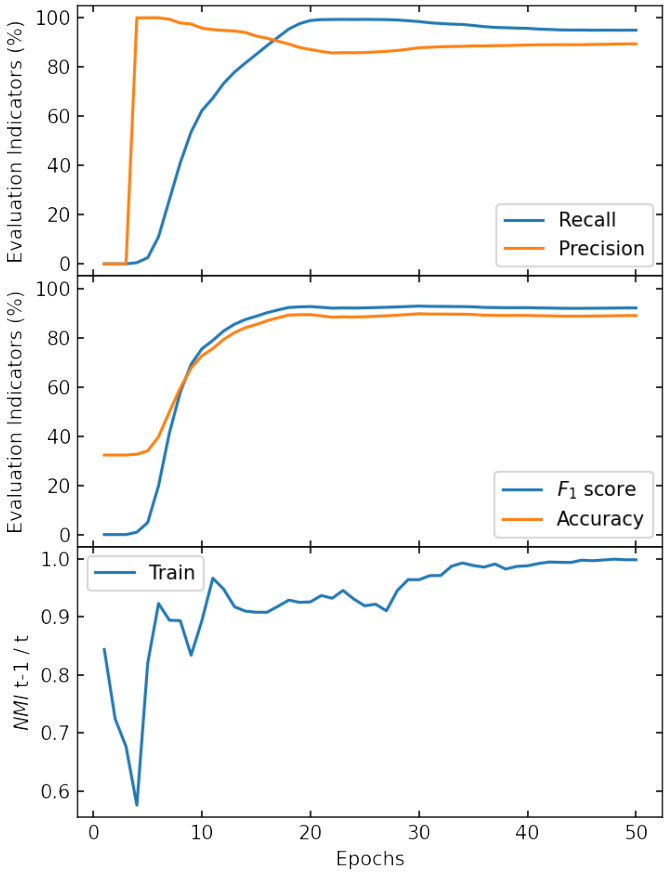

Fig. 10. Accuracy and recall metrics of SS-3D-Clump on the test dataset as the epoch of the training process changes. Evolution of cluster reassignments (NMI) in each clustering iteration of SS-3D-Clump.

Acc measures the proportion of correct predictions across all classes relative to the total number of predictions. P indicates the proportion of correct predictions for a specific class compared to the total number of predictions for that class, focusing on the quality of positive predictions. R represents the proportion of correct predictions for a specific class relative to that class's total number of positive cases. Combining P and R yields informative metrics like F_1, the harmonic mean of P and R.

Figure 10 shows those evaluation indications of SS-3D-Clump in verifying the samples in the test dataset after each iteration during the training process. As shown from the top two subplots in Fig. 10, the model exhibits high accuracy, but low recall during the first ten training epochs, as the insufficiently trained model tends to classify most samples as negative. As training iterations increase, the recall rate rises sharply while the accuracy decreases slightly. After 30 training iterations, all metrics stabilize, with precision and recall approaching 90% and 95%. It is noteworthy that this level of accuracy is achieved using only seed samples of 300 real and 300 false-positive clumps, indicating that the SS-3D-Clump can find similarities and differences among data through training, enabling it to recognize molecular clumps effectively, even in large-scale training projects with massive sample sizes.

We measured the information shared between two different cluster assignments A and B, each corresponding to a different training epoch, using normalized mutual rnformation (NMI, Caron et al. 2018). In this context, an epoch refers to a complete pass through the training dataset during model optimization. As the SS-3D-Clump model trains, it generates different cluster

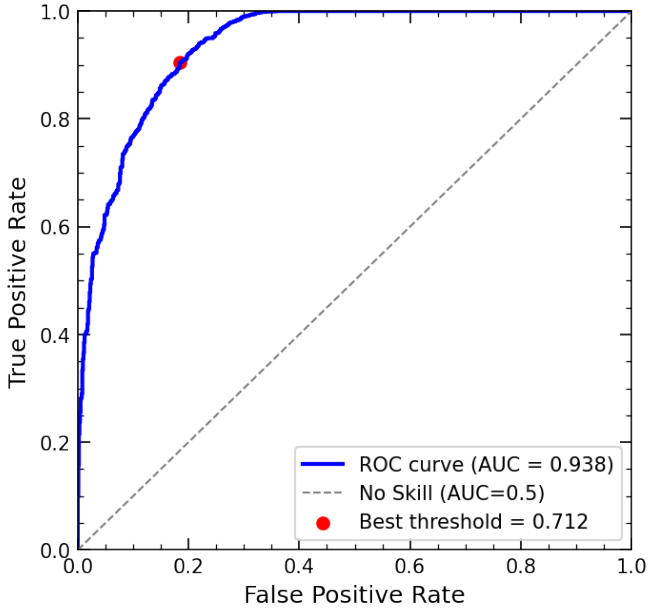

Fig. 11. Receiver operating characteristic (ROC) curve of the SS-3D-Clump model. The optimal threshold of 0.712 yields an area under the curve (AUC) of 0.938.

assignments at each epoch. By computing $NMI(A; B)$ between successive epochs (e.g., A at epoch $t-1$ and B at epoch t), we can monitor the evolution of the predicted assignments and assess the convergence and stability of the clustering process. The NMI is calculated as

$$NMI(A; B) = \frac{MI(A; B)}{\sqrt{H(A)H(B)}}, \tag{5}$$

where MI denotes the mutual information and H the entropy. Overall, NMI can be applied to any assignment from the clusters or the true labels. The value of NMI typically falls between 0 and 1, where 1 indicates complete similarity between two clustering results, and 0 signifies no resemblance. Therefore, the NMI curves can be utilized to observe how the performance of SS-3D-Clump changes with increasing number of training epochs, facilitating the optimization of the model training process. The bottom panel in Fig. 10 illustrates the NMI curve changing with the epochs during the training process of SS-3D-Clump, it can be observed that the NMI increases to ∼0.98 at the 30th epoch and remains stable thereafter. This indicates that SS-3D-Clump is experiencing fewer reassignments and the clusters are stabilizing.

Utilizing the stabilized SS-3D-Clump model, we verified the candidates detected by FacetClumps by computing their confidence values, which represent the predicted probability that a given candidate is a real molecular clump. As shown in Fig. 11, the receiver operating characteristic (ROC) curve (Rojas et al. 2022; Demianenko et al. 2023) yields an optimal classification threshold of 0.712, with an area under the curve (AUC) of 0.938, indicating that the model effectively distinguishes between clump and non-clump samples. To simplify interpretation while preserving high-confidence selections, we adopted a slightly rounded threshold of 0.7. Candidates with confidence values below this threshold were rejected.

As a result, SS-3D-Clump excluded 16 384 of the 39 534 clump candidates, leaving a final catalog of 23 150 molecular clumps. Figure 12 shows the histograms of clump PS/N values.

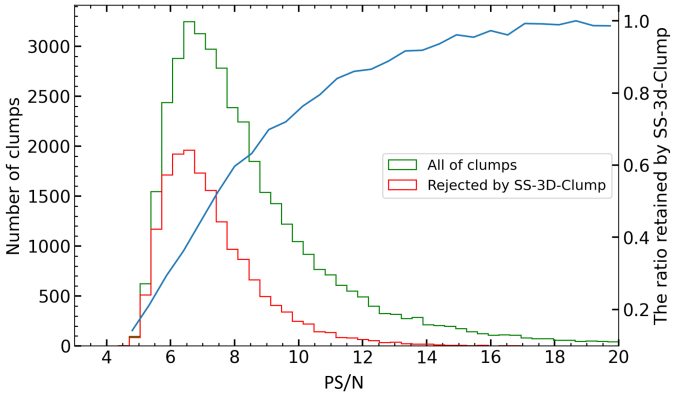

Fig. 12. Histogram illustrating the PS/N distribution of clumps molecular clump candidates obtained by Facet-SS-3D-Clump. The green bars represent the distribution of clumps obtained by FacetClumps, while the red bars represent these rejected by SS-3D-Clump. The blue curve indicates the proportion of clumps retained by SS-3D-Clump as a function of the PS/N.

The green histogram represents the PS/N distribution of clumps obtained by FacetClumps, while the red histogram shows the distribution of PS/N values for clumps excluded by SS-3D-Clump. The blue curve indicates the proportion of clumps retained by SS-3D-Clump as a function of the PS/N. Figure 12 shows that the PS/N of the clumps verified as false by SS-3D-Clump is primarily below 10. Combining this with the results from the algorithm benchmark tests, where detection results of various algorithms converge and show high accuracy when the PS/N exceeds 12 (see Fig. 8), it is evident that SS-3D-Clump is able to effectively exclude false-positive detections. This experiment demonstrates that combining detection algorithms with deep learning-based verification techniques can allow us to achieve results that match the reliability of combining the results from detections made with multiple algorithms. In practical scenarios, this approach enables the detection algorithms to identify as many candidates as possible for high recall rates, while SS-3D-Clump enhances the accuracy of the resulting clump catalog.

4. Result

Using the two-step method for detecting and verifying clumps, there were 23 150 ^{13}CO clumps extracted from QI of FUGIN, covering the first quadrant region ($10° \leq l \leq 50°$, $|b| \leq 1°$). The names, units, symbols, and detailed descriptions of the columns in the molecular clump catalog are provided in Table 1. Following the definition of Berry (2015), we define the intensity-weighted second-moment sizes (or RMS sizes) along each axis as

$$S_l = \sqrt{\frac{\sum_i I_i(l_i - \bar{l})^2}{\sum_i I_i}}, \quad S_b = \sqrt{\frac{\sum_i I_i(b_i - \bar{b})^2}{\sum_i I_i}}, \tag{6}$$

where I_i is the intensity of the voxel, i, while $\bar{l}$ and $\bar{b}$ are the intensity-weighted centroids defined as

$$\bar{l} = \frac{\sum_i I_i l_i}{\sum_i I_i}, \quad \bar{b} = \frac{\sum_i I_i b_i}{\sum_i I_i}. \tag{7}$$

For sources with purely Gaussian intensity profiles, this RMS size corresponds to the standard deviation of the Gaussian

Table 1. FUGIN ^{13}CO (1–0) clump catalog.

ID	l_c	b_c	v_c	S_l	S_b	Δv	θ	T_{peak}	W	PS/N	α
–	(deg)	(deg)	(km s^{-1})	($''$)	($''$)	(km s^{-1})	(deg)	(K)	(10^3 K km s^{-1})	–	–
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
FUGIN034.231+00.132+057.36	34.231	0.132	57.36	36.4	29.9	1.72	9	25.4	14.4	28.9	1.00
FUGIN034.255+00.152+056.95	34.255	0.152	56.95	53.4	41.7	2.62	22	28.0	26.2	37.5	1.00
FUGIN043.173-00.022+011.98	43.173	–0.022	11.98	45.4	36.7	2.90	–25	12.6	28.6	28.5	1.00
FUGIN029.906-00.045+099.40	29.906	–0.045	99.40	48.5	51.2	2.69	–57	24.3	36.9	26.9	1.00
FUGIN030.825-00.062+096.35	30.825	–0.062	96.35	32.9	43.3	1.99	–59	14.4	10.4	18.4	1.00
FUGIN030.719-00.098+095.33	30.719	–0.098	95.33	51.1	44.2	2.27	54	11.8	20.7	16.1	1.00
FUGIN029.970-00.016+096.85	29.970	–0.016	96.85	38.1	51.5	1.71	–74	20.7	23.6	25.6	1.00
FUGIN030.447-00.231+103.45	30.447	–0.231	103.45	36.4	51.1	1.97	72	15.2	12.5	18.8	1.00
FUGIN029.854-00.056+099.85	29.854	–0.056	99.85	35.9	44.1	2.11	47	30.6	15.0	30.3	1.00
FUGIN012.297-00.441+047.68	12.297	–0.441	47.68	26.7	22.1	0.81	–27	10.0	1.9	18.2	1.00

Notes. Columns are as follows: (1) source designation; (2–4) centroid coordinates; (5–7) intensity-weighted RMS sizes in each dimension; S_l and S_b the represent intensity-weighted RMS deviation of voxels from the centroid in the orthogonal l, b axes, the same as Δv but in the v axis. For sources with purely Gaussian profiles, these RMS sizes would return the standard deviation of the profile in the corresponding axis, which may be converted into FWHM by multiplying by a factor of $\sqrt{8ln2}$. (8) position angle of the Clump; (9) the brightest intensity of the clump; (10) total integrated intensity over all voxels in the clump; (11) PS/N; (12) confidence associated with the clump verified by SS-3D-Clump. Only a portion of the full table is shown here to illustrate its form and content.

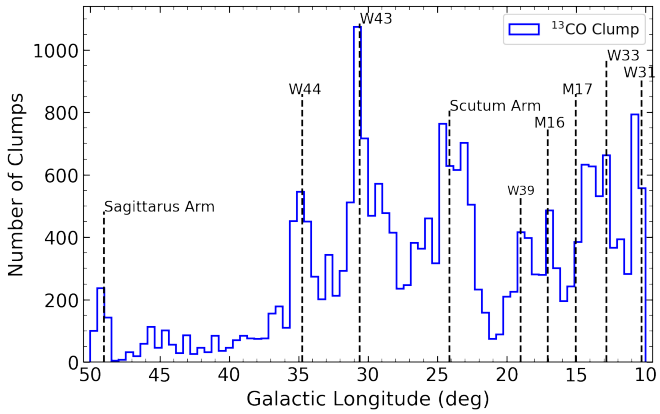

Fig. 13. Distribution of ^{13}CO clumps along the Galactic longitude. The dashed line represents the star-forming region or spiral arm in the QI.

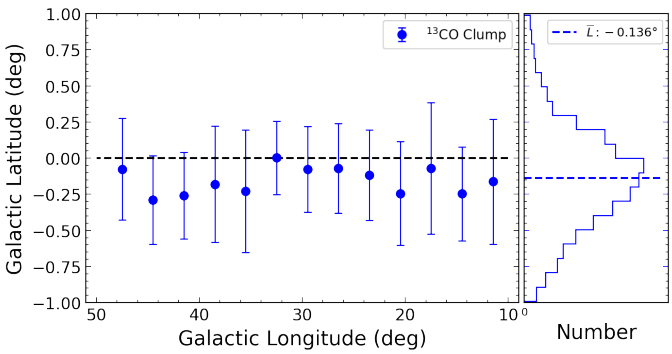

Fig. 14. Latitude distribution of clumps in the QI. The left part illustrates the mean and standard deviation of clump distribution along latitude for various longitude intervals. Blue circular points denote the mean latitude within each interval, with error bars representing the standard deviation. The right part shows a latitude distribution of QI clumps, with the blue dashed line indicating the mean latitude.

along the corresponding axis, namely,

$$I(l) \propto \exp\left[-\frac{(l-\bar{l})^2}{2\sigma_l^2}\right], \quad I(b) \propto \exp\left[-\frac{(b-\bar{b})^2}{2\sigma_b^2}\right], \tag{8}$$

the intensity-weighted RMS sizes reduce exactly to the Gaussian standard deviations. Thus, for purely Gaussian clumps, the RMS sizes directly correspond to the standard deviations of the profiles along each axis.

4.1. Galactic distribution

Due to the sparse distribution of molecular gas in the QIII, only a few molecular clumps (167) are detected, we consider exploring the analysis of the outer Galaxy as a separate case study. This would allow for a more detailed investigation by matching the outer Galaxy data with OGHReS (Urquhart et al. 2024) observations to examine star formation metrics. In this study, the spatial analysis is limited to the QI. The longitude distribution of ^{13}CO clumps is shown in Fig. 13, where the dashed lines mark the locations of known star-forming regions or spiral arms within QI. Similarly to Beuther et al. (2012); Molinari et al. (2016), we identify in Fig. 13 prominent features that can be associated with major star-forming complexes or with the tangents of spiral arms. Specifically, the overdensities near $l \sim 23°$ and $l \sim 49°$ correspond to the Scutum and Sagittarius arms, respectively, where the line of sight intersects dense molecular gas and active star-forming regions. This figure reveals clustering of ^{13}CO clumps near the longitude of the star-forming region or spiral arm.

The latitude distribution of clumps in the QI is shown in Fig. 14. In the left panel of this figure, we show the mean and standard deviation of the distribution of clumps along latitude for different longitude intervals. The filled blue circles represent the mean latitude of the clumps within each longitude interval, with error bars indicating the size of their standard deviation. The blue histogram shown in the right panel of Fig. 14 shows the distribution of molecular clumps obtained by Facet-SS-3D-Clump and the blue dashed line indicating the mean value of latitude. The histogram peaks at slightly negative values with a mean value of $b = -0.136°$. As Molinari et al. (2016) pointed out, this may

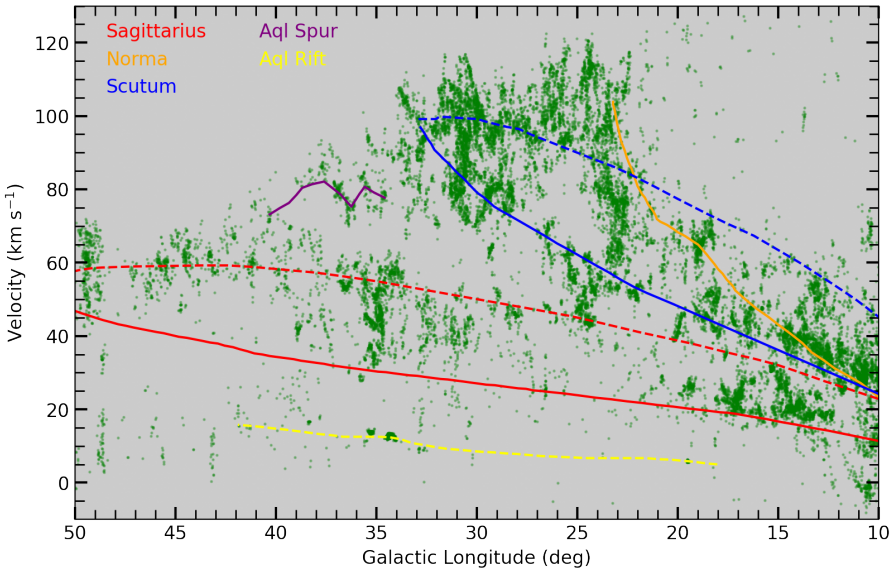

Fig. 15. Centers of clumps in the l–v plane are illustrated. The lines indicate the spiral arm loci of Norma, Scutum, and Sagittarius, as well as the smaller Aquila spur and Aquila Rift as estimated by Reid et al. (2014). The near and far sides of the arms are dotted and solid lines, respectively.

be attributed to an incorrect assumption about the vertical position of the Sun in the Milky Way. The latitude distribution of molecular clumps exhibits an asymmetrical pattern.

In Fig. 15, we show the centers of clumps in the longitude-velocity ($l - v$) plane and the relevant sprial arm features (Reid et al. 2014). Dashed lines indicate the near side of the arms and solid lines represent the respective far side. In addition to the correlation between the spiral arms, we also find significant numbers of sources coincident with the Aquila Rift (34°, 11 km s^{-1}) and Aquila Spur (37°, 80 km s^{-1}). The good correlation with these spiral features provides confidence that the detected sources are genuine and that the number of false detections is low.

4.2. Comparison with CHIMPS

CHIMPS is a high resolution ^{13}CO and C^{18}O (J=3–2) survey of ∼20 square degrees of the first quadrant (θ =15″ and 27.8° ≤ l ≤ 46.2° and $|b|$ ≤ 0.5°; Rigby et al. 2016). Rigby et al. (2019) obtained the first source catalog consisting of 4999 clumps identified by FellWalker Berry (2015) in the ^{13}CO emission maps. Comparing the CHIMPS sources with the sources extracted from FUGINS provides a benchmark for our method. The selection of CHIMPS over GRS was motivated by two principal considerations. First, the angular resolution of CHIMPS (15″) closely matches that of FUGIN (20″), while GRS exhibits a significantly coarser resolution (46″). This resolution compatibility ensures more reliable one-to-one (or one-to-many) matching between the catalogs, effectively minimizing systematic biases arising from spatial scale mismatches. Second, while both surveys observe the ^{13}CO line, they probe different rotational transitions (J=3–2 for CHIMPS versus J=1–0 for FUGIN). This configuration not only facilitates future investigations of excitation conditions across transitions but also provides unique opportunities to study the physical properties and evolutionary stages of star-forming structures through multi-transition analysis. Consequently, CHIMPS offers both methodological robustness and scientific added value for verifying and expanding upon our results.

Using the matching criteria described in Section 3.3, we obtained 2476 CHIMPS sources with a FUGIN counterpart. The FUGIN catalog for the overlapping CHIMPS region contains 7171 clumps and 5934 FUGIN clumps have corresponding CHIMPS clumps. The observed asymmetry in matched clumps (5934 FUGIN clumps associated with 2476 CHIMPS clumps) reflects a complex interplay between detection thresholds and molecular line properties. The higher S/N threshold used in FUGIN (≥4) preferentially isolates prominent emission peaks, especially within the more diffuse ^{13}CO (J=1–0) line, which is sensitive to cold and extended molecular gas. This tends to fragment extended structures into multiple smaller clumps. In contrast, CHIMPS employs a lower S/N threshold (≥2), allowing for more extended J=3–2 emission to be grouped into single clumps, often merging adjacent substructures into larger ones. Moreover, the J=3–2 line traces warmer and denser regions, typically found in the compact centers of star-forming molecular clouds, which may appear as continuous, larger-scale structures. As a result, many small FUGIN clumps can fall within the extent of a single CHIMPS clump, naturally leading to a one-to-many matching relationship.

It should be noted that the process of cross-matching clumps identified in different transitions is subject to intrinsic physical limitations. The ^{13}CO (J=1–0) transition used in FUGIN primarily traces colder and more diffuse molecular gas, while the ^{13}CO (J=3–2) line observed in CHIMPS is excited under warmer and denser conditions. As a result, some clumps that are clearly detected in the J=1–0 emission might fall below the excitation threshold of the J=3–2 line and vice versa. These differences in excitation temperature, optical depth, and critical density naturally lead to incomplete one-to-one correspondence between the catalogs.

Figure 16 shows the relative deviations in Galactic longitude and latitude between FUGIN clumps that have matched counterparts in CHIMPS. The relative deviation (ε) is defined as

$$\varepsilon = \frac{d_{cen}}{d_{size}}, \tag{9}$$

where d_{cen} is the distance between the centroids of the matched clumps and d_{size} is the sum of the size of the matched clumps in Galactic longitude or latitude. The blue and orange bars in Fig. 16 represent the relative deviations of the matched clumps in Galactic longitude and latitude, respectively. The figure shows that most matched clumps (∼72%) have a relative deviation of less than 25%.

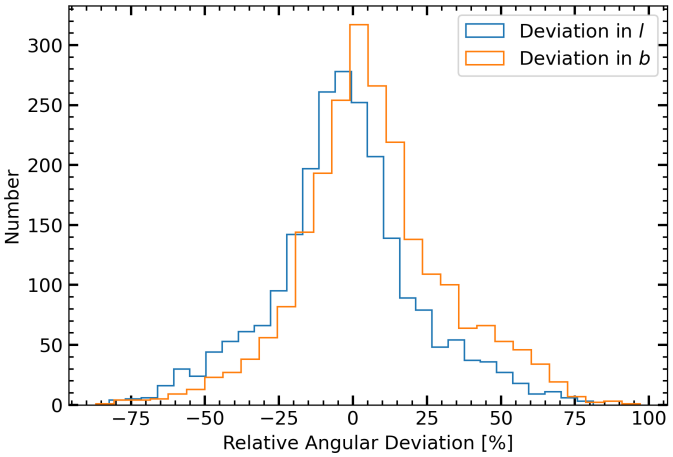

Fig. 16. Relative angular deviations of matched clumps in Galactic longitude and latitude. Blue bars represent the deviations in Galactic longitude, while orange bars represent the deviations in Galactic latitude.

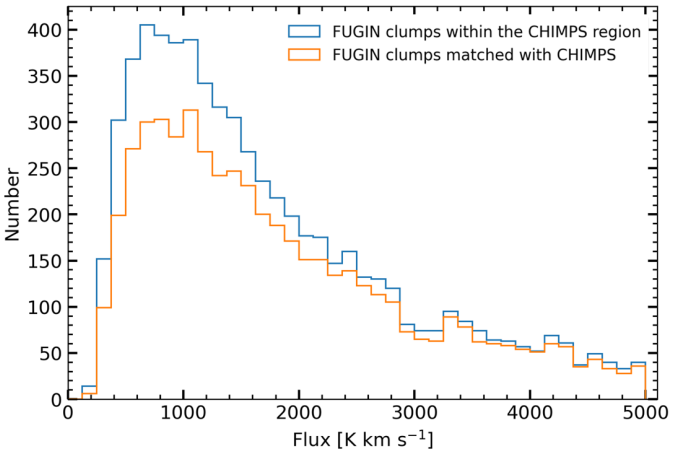

Fig. 17. Flux distribution of all FUGIN clumps within the CHIMPS region (blue bars) and the flux distribution of FUGIN clumps matched with CHIMPS (orange bars).

Figure 17 shows the flux distribution of FUGIN clumps within the CHIMPS region and the flux distribution of clumps matched with CHIMPS. The blue bars represent the flux distribution of FUGIN clumps in the CHIMPS region, while the orange bars represent the flux distribution of clumps with CHIMPS counterparts. As can be seen from the figure, the flux of clumps without a CHIMPS counterpart is primarily concentrated on the left side of the distribution, indicating that these clumps tend to be smaller or less intense.

Based on the matching results, we adopted the distances (d) from the clumps in CHIMPS and calculated the equivalent radii for the clumps in FUGIN. The equivalent radius was defined consistently with the value given in CHIMPS (Rigby et al. 2019), calculated as $R_\sigma = d\sqrt{S_l S_b}$, where d is the distance in parsecs, and S_l and S_b are angular sizes in radians. Here, S_l and S_b represent the intensity-weighted root-mean-square (rms) deviations of voxels from the clump centroid along the Galactic longitude and latitude axes, respectively. Figure 18 shows the equivalent radii of the matched clumps in FUGIN and CHIMPS. The horizontal axis represents the equivalent radii of the clumps in FUGIN, while the vertical axis represents the equivalent radii of the corresponding matched clumps in CHIMPS. As seen in Fig. 18,

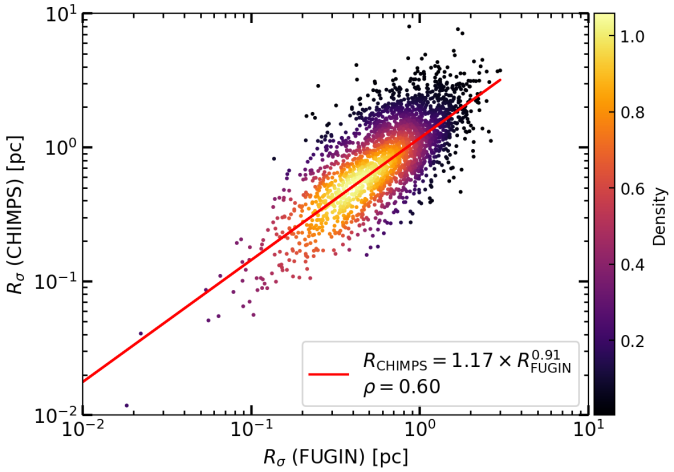

Fig. 18. Comparison of the radii of FUGIN clumps matched to CHIMPS clumps. The distances of the FUGIN clumps are set to match these of their corresponding CHIMPS clumps. The red line represents the fitted relationship, with a correlation coefficient of 0.6. The color bar on the right indicates the local density of data points, normalized to unity, meaning that the region with the highest concentration of points has a normalized density of 1.

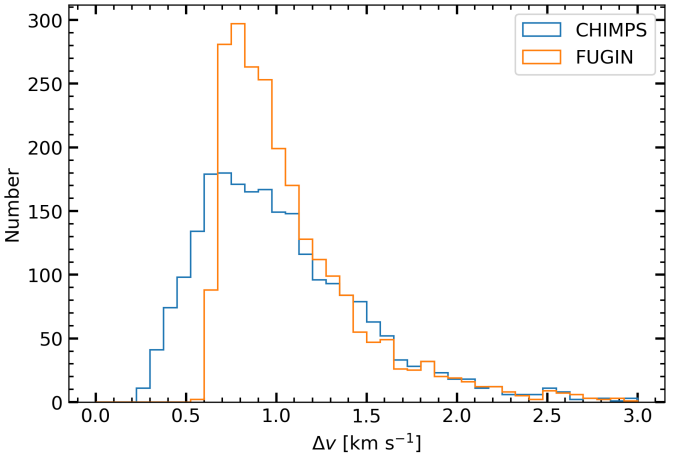

Fig. 19. Comparison of Δv of the matched FUGIN-CHIMPS clumps.

most clumps in FUGIN have equivalent radii within the range of 0.1–1 pc, with a few exceeding 1 pc but all within 3 pc. It indicates that the dense structures detected in the FUGIN ^{13}CO data primarily correspond to clump-scale structures.

A comparison between the equivalent radii of the matched FUGIN and CHIMPS clumps reveals a power-law relation expressed as $R_{\rm CHIMPS} = 1.17 \times R^{0.91}_{\rm FUGIN}$, as shown in Fig. 18. The Pearson correlation coefficient between the two sets of radii is 0.60, indicating a moderate positive correlation. This suggests that the sizes of CHIMPS clumps are generally comparable to, but slightly larger than, those in FUGIN, and they tend to scale consistently on average. It should be noted that the two extractions were performed on different ^{13}CO transitions J=3–2 for CHIMPS and J=1–0 for FUGIN. A likely explanation lies in the different detection thresholds: the FellWalker algorithm used for CHIMPS allows voxels with S/N $\geq$ 2 to be included in a clump, while in FUGIN, a stricter threshold of S/N $\geq$ 4 is applied. As a result, CHIMPS clumps tend to include more low-intensity surrounding voxels, leading to slightly larger segmented regions.

Figure 19 shows the distribution of velocity dispersion for the matched FUGIN-CHIMPS clumps. The velocity dispersion

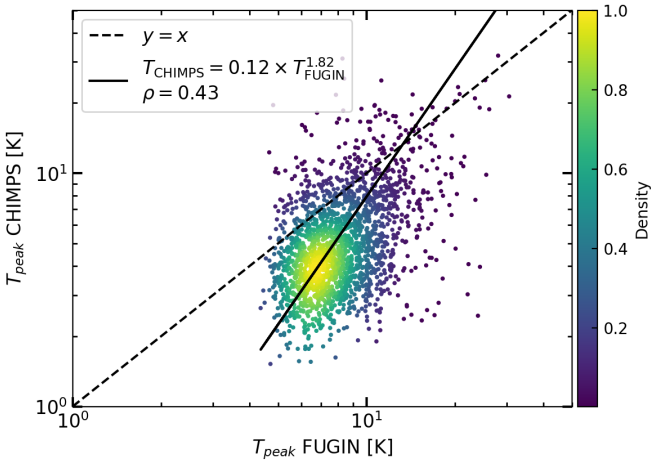

Fig. 20. Comparison of T_{peak} of the matched FUGIN-CHIMPS clumps.

of FUGIN clumps is mainly concentrated between 0.6 and 1.5 km s^{-1}, with the distribution peaking at 0.8 km s^{-1}. Notably, the FUGIN distribution lacks clumps with velocity dispersions below 0.6 km s^{-1}, which are present in the CHIMPS sample. This discrepancy arises from the extraction criteria applied in the FUGIN clump identification process, which require a minimum velocity extent of five consecutive channels (corresponding to 3.25 km s^{-1}) for a clump to be considered. Since the velocity dispersion Δv is defined as the intensity-weighted rms deviation from the centroid in velocity axis, this threshold effectively excludes clumps with narrow linewidths. Assuming a Gaussian line profile with a full width of five channels (approximately 6σ), the minimum Δv would be around 0.504 km s^{-1}, which is consistent with the lower bound observed in the FUGIN distribution. As a result, some low-Δv clumps present in CHIMPS were not recovered in FUGIN due to this velocity constraint.

We also compared the peak intensities of molecular clumps matched between FUGIN and CHIMPS. As shown in Fig. 20, the peak intensities of most FUGIN clumps are stronger than those of CHIMPS clumps, which may be because the excitation conditions for the $J = 3 \rightarrow 2$ transition are higher than those for the $J = 1 \rightarrow 0$ transition, resulting in lower radiation temperatures. Another noteworthy point is that the scatter plot of the matched FUGIN and CHIMPS clumps exhibits a power-law distribution. The fitted power-law index is 1.82, indicating that as the intensity of the molecular clumps increases, the intensity of the $J = 3 \rightarrow 2$ clumps increases more rapidly. The Pearson correlation coefficient between the logarithmic peak intensities of the two datasets is 0.43, indicating a weak to moderate linear correlation in log-log space. The considerable scatter visible in Fig. 20 suggests that this relationship is not tightly constrained, likely reflecting intrinsic physical differences between the tracers: the $J = 3 \rightarrow 2$ transition traces warmer and denser gas with higher critical densities, whereas the $J = 1 \rightarrow 0$ line is more sensitive to extended, lower-density molecular envelopes.

The FUGIN ^{13}CO (J=1–0) clump catalog presented in this study complements existing large-scale molecular clump catalogs such as GRS (^{13}CO J=1–0) and CHIMPS (^{13}CO J=3–2). With its high-resolution coverage and consistent clump identification criteria applied across the first Galactic quadrant, this catalog provides a valuable resource for investigating environmental variations in molecular clump structure and star formation efficiency in the inner Milky Way. Moreover, the dataset can be directly compared with GRS and CHIMPS to probe excitation conditions, evolutionary stages, and line-ratio diagnostics, thereby facilitating comprehensive multi-transition analyses of Galactic molecular gas. A detailed excitation analysis for the FUGIN and CHIMPS-matched molecular clumps is presented in a separate study.

5. Conclusion

We utilized Facet-SS-3D-Clump for the detection and verification of molecular clumps in the ^{13}CO data within FUGIN, covering the longitude range $10° \leq l \leq 50°$ and a latitude strip of $|b| \leq 1°$. We obtained a catalog containing 23 150 ^{13}CO clumps. We summarize our conclusions below.

First, the Facet-SS-3D-Clump workflow leverages a clump detection algorithm combined with semi-supervised deep learning to effectively search for and verify dense structures in large-scale survey projects, achieving accuracy comparable to manual verification and making comprehensive molecular clump surveys feasible. Second, these molecular clumps align well with the star-forming regions and spiral arms, exhibiting a mean Galactic latitude significantly below the midplane, with $b = -0.136°$. Third, this catalog provides data support for analyzing the physical properties of clumps in high-mass star-forming regions, enhancing our understanding of the role of environmental effects in the formation and evolution of clumps and how different environments impact the star formation process. Finally, the flux completeness limits were presented: the catalog is 80% complete above 466 K km s^{-1}.

Data availability

Full table 1 is available at the CDS via https://cdsarc.cds.unistra.fr/viz-bin/cat/J/A+A/706/A242.

Acknowledgements. This publication makes use of data from FUGIN, the FOREST Unbiased Galactic plane Imaging survey with the Nobeyama 45-m telescope, a legacy project of the Nobeyama 45-m radio telescope. We are grateful to all the members of the FUGIN working group. This work was supported by the National Natural Science Foundation of China (U2031202, 11903083, 11873093, and 12203029). Software: CARTA (Angus Comrie et al. 2021), Astropy (Astropy Collaboration 2013, 2018; Astropy Collaboration 2022), TensorFlow (Abadi et al. 2015), Scikit-Learn (Pedregosa et al. 2011).

References

Abadi, M., Agarwal, A., Barham, P., et al. 2015, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, software available from https://tensorflow.org
Alves, J., Lombardi, M., & Lada, C. J. 2007, A&A, 462, L17
Alves de Oliveira, C., Schneider, N., Merín, B., et al. 2014, A&A, 568, A98
André, P., Di Francesco, J., Ward-Thompson, D., et al. 2014, in Protostars and Planets VI, eds. H. Beuther, R. S. Klessen, C. P. Dullemond, & T. Henning, 27
Angus Comrie, Kuo-Song Wang, Shou-Chieh Hsu, et al. 2021, CARTA: The Cube Analysis and Rendering Tool for Astronomy
Astropy Collaboration (Robitaille, T. P., et al.) 2013, A&A, 558, A33
Astropy Collaboration (Price-Whelan, A. M., et al.) 2018, AJ, 156, 123
Astropy Collaboration (Price-Whelan, A. M., et al.) 2022, ApJ, 935, 167
Ay, M., Özbakır, L., Kulluk, S., et al. 2023, Expert Syst. Appl., 211, 118656
Basu, S., Banerjee, A., & Mooney, R. J. 2002, in Machine Learning, Proceedings of the Nineteenth International Conference (ICML 2002), University of New South Wales, Sydney, Australia, July 8–12, 2002
Benedettini, M., Traficante, A., Olmi, L., et al. 2021, A&A, 654, A144
Bergin, E. A., & Tafalla, M. 2007, ARA&A, 45, 339
Berry, D. S. 2015, Astron. Comput., 10, 22
Beuther, H., Linz, H., Henning, T., et al. 2011, A&A, 531, A26
Beuther, H., Tackenberg, J., Linz, H., et al. 2012, ApJ, 747, 43
Blitz, L., & Williams, J. P. 1999, The Origin of Stars and Planetary Systems, eds. C. J. Lada & N. D. Kylafis, 540, 3

Caron, M., Bojanowski, P., Joulin, A., & Douze, M. 2018, in Proceedings of the European Conference on Computer Vision (ECCV)
Chen, Z., Sun, W., Chini, R., et al. 2021, ApJ, 922, 90
Cheng, T.-Y., Conselice, C. J., Aragón-Salamanca, A., et al. 2020, MNRAS, 493, 4209
Clark, P. C., Klessen, R. S., & Bonnell, I. A. 2007, MNRAS, 379, 57
Colombo, D., Hughes, A., Schinnerer, E., et al. 2014, ApJ, 784, 3
Cunha, P. A. C., Humphrey, A., Brinchmann, J., et al. 2024, A&A, 687, A269
Demianenko, M., Malanchev, K., Samorodova, E., et al. 2023, A&A, 677, A16
Eden, D. J., Moore, T. J. T., Currie, M. J., et al. 2020, MNRAS, 498, 5936
Elia, D., Merello, M., Molinari, S., et al. 2021, MNRAS, 504, 2742
Faesi, C. M., Lada, C. J., & Forbrich, J. 2016, ApJ, 821, 125
Han, J., Kamber, M., & Pei, J. 2012, in Data Mining, third edn., eds. J. Han, M. Kamber, & J. Pei, The Morgan Kaufmann Series in Data Management Systems (Boston: Morgan Kaufmann), 497
He, Z., Qiu, B., Luo, A. L., et al. 2021, MNRAS, 508, 2039
Heyer, M., & Dame, T. 2015, ARA&A, 53, 583
Jackson, J. M., Rathborne, J. M., Shah, R. Y., et al. 2006, ApJS, 163, 145
Ji, Q., & Haralick, R. M. 2002, Pattern Recognit., 35, 689
Jiang, Y., Chen, Z., Zheng, S., et al. 2023, ApJS, 267, 32
Karpfinger, C. 2022, Polynomial and Spline Interpolation (Berlin, Heidelberg: Springer Berlin Heidelberg), 311
Kerton, C. R., Arvidsson, K., & Alexander, M. J. 2013, AJ, 145, 78
Khan, S., Pandian, J. D., Lal, D. V., et al. 2022, A&A, 664, A140
Krumholz, M. R., & McKee, C. F. 2005, ApJ, 630, 250
Liu, L., Bureau, M., Li, G.-X., et al. 2022, MNRAS, 517, 632
Luo, X., Zheng, S., Huang, Y., et al. 2022, Res. Astron. Astrophys., 22, 015003
Luo, X., Zheng, S., Jiang, Z., et al. 2024a, A&A, 683, A104
Luo, X., Zheng, S., Jiang, Z., et al. 2024b, Res. Astron. Astrophys., 24, 055018
Medina, S. N. X., Urquhart, J. S., Dzib, S. A., et al. 2019, A&A, 627, A175
Mège, P., Russeil, D., Zavagno, A., et al. 2021, A&A, 646, A74
Molinari, S., Schisano, E., Elia, D., et al. 2016, A&A, 591, A149
Motte, F., Bontemps, S., & Louvet, F. 2018, ARA&A, 56, 41
Nakanishi, H., Fujita, S., Tachihara, K., et al. 2020, PASJ, 72, 43
Ohashi, S., Sanhueza, P., Chen, H.-R. V., et al. 2016, ApJ, 833, 209
Ooyama, K. V. 2002, Monthly Weather Rev., 130, 2392
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, J. Mach. Learn. Res., 12, 2825
Rani, R., Moore, T. J. T., Eden, D. J., et al. 2023, MNRAS, 523, 1832
Rathborne, J. M., Johnson, A. M., Jackson, J. M., Shah, R. Y., & Simon, R. 2009, ApJS, 182, 131
Reid, M. J., Menten, K. M., Brunthaler, A., et al. 2014, ApJ, 783, 130
Rigby, A. J., Moore, T. J. T., Plume, R., et al. 2016, MNRAS, 456, 2885
Rigby, A. J., Moore, T. J. T., Eden, D. J., et al. 2019, A&A, 632, A58
Rojas, K., Savary, E., Clément, B., et al. 2022, A&A, 668, A73
Rosolowsky, E. 2005, PASP, 117, 1403
Rosolowsky, E. W., Pineda, J. E., Kauffmann, J., & Goodman, A. A. 2008, ApJ, 679, 1338
Schuller, F., Csengeri, T., Urquhart, J. S., et al. 2017, A&A, 601, A124
Smartt, S. J., & Rolleston, W. R. J. 1997, ApJ, 481, L47
Su, Y., Sun, Y., Li, C., et al. 2016, ApJ, 828, 59
Su, Y., Yang, J., Zhang, S., et al. 2019, ApJS, 240, 9
Takekoshi, T., Fujita, S., Nishimura, A., et al. 2019, ApJ, 883, 156
Tremblin, P., Schneider, N., Minier, V., et al. 2014, A&A, 564, A106
Umemoto, T., Minamidani, T., Kuno, N., et al. 2017, PASJ, 69, 78
Urquhart, J. S., König, C., Colombo, D., et al. 2024, MNRAS, 528, 4746
Williams, J. P., de Geus, E. J., & Blitz, L. 1994, ApJ, 428, 693
Williams, J. P., Blitz, L., & McKee, C. F. 2000, in Protostars and Planets IV, eds. V. Mannings, A. P. Boss, & S. S. Russell, 97
Wurster, J., & Rowan, C. 2023, MNRAS, 523, 3025
Yoo, H., Lee, C. W., Chung, E. J., et al. 2023, ApJ, 957, 94
Zhang, Q., Wang, Y., Pillai, T., & Rathborne, J. 2009, ApJ, 696, 268
Zhang, S., Zavagno, A., López-Sepulcre, A., et al. 2021, A&A, 646, A25
Zinnecker, H., & Yorke, H. W. 2007, ARA&A, 45, 481

Appendix A: Comparison with Dendrograms

As a benchmark against existing mature algorithms, three local regions along the Galactic longitude in QI, spanning different density ranges, were selected to compare and analyze the detection results of the FacetClumps and Dendrograms algorithms. Three selected regions are centered at Galactic longitudes 12°, 28°, and 44°, each covering an area of one square degree. The first region spans a Galactic longitude range of 11.5 to 12.5°, a Galactic latitude range of -0.87 to 0.13 °, and a velocity range of -10 to 80 km s^{-1} (hereafter called Cell_12). The second region: $27.65° \leq l \leq 28.65°$, $-0.6° \leq b \leq 0.4°$, and 12 km s^{-1} $\leq v \leq$ 112 km s^{-1} (hereafter called Cell_28). The third region: $43.65° \leq l \leq 44.65°$, $-0.46° \leq b \leq 0.54°$, and 0 km s^{-1} $\leq v \leq$ 100 km s^{-1} (hereafter called Cell_44). Figure A.1 shows the intensity information for Cell_12, Cell_28, and Cell_44 in the top, middle, and bottom panels, respectively. For each cell, the first, second, and third columns display the integrated intensity maps projected along the velocity, Galactic latitude and Galactic longitude axes, respectively.

Dendrograms (Rosolowsky et al. 2008) algorithm implemented in astrodendro[6] is an abstraction of the changing topology of the isosurfaces as a function of contour level. It uses a tree diagram to describe hierarchical structures over various scales in a 2D or 3D datacube (Zhang et al. 2021). The results return two types of structures: leaves, which have no substructure, and branches, which can split into multiple branches or leaves. The algorithm has two main parameters: T_{min} and ΔT. The parameter T_{min} represents the minimum data value to be considered, and is set to 4. The parameter ΔT defines the required contrast for a leaf to be identified as an independent clump and a value of $\Delta T = 1$ is used, implying that a clump must have a peak intensity of at least 5. The other parameter is the minimum volume parameter ($V_{min} = 125$), which ensures that the retained clumps have a minimum pixel volume. The above parameter settings are all configured for S/N data and are consistent with the parameters used in FacetClumps (see detail in Section 3.1). FacetClumps detected 463, 317, and 128 clumps in Cell_12, Cell_28, and Cell_44, respectively, while Dendrograms detected 318, 198, and 88 clumps.

We matched the detection results of FacetClumps and Dendrograms, taking into account the region and intensity information of the clumps. The specific matching criterion was considered a match when the sum of intensities in the overlapping region accounted for 25% of the sum of intensities within the clump regions detected by each algorithm. We did not use the intersection-over-union (IoU) as the matching standard because different algorithms may produce varying segmentations of the same data and detect different numbers of clumps. For example, in the first rows of Fig. A.3, Dendrograms detected one clump, while FacetClumps detected two. If IoU were used,

[6] https://dendrograms.readthedocs.io/en/stable/

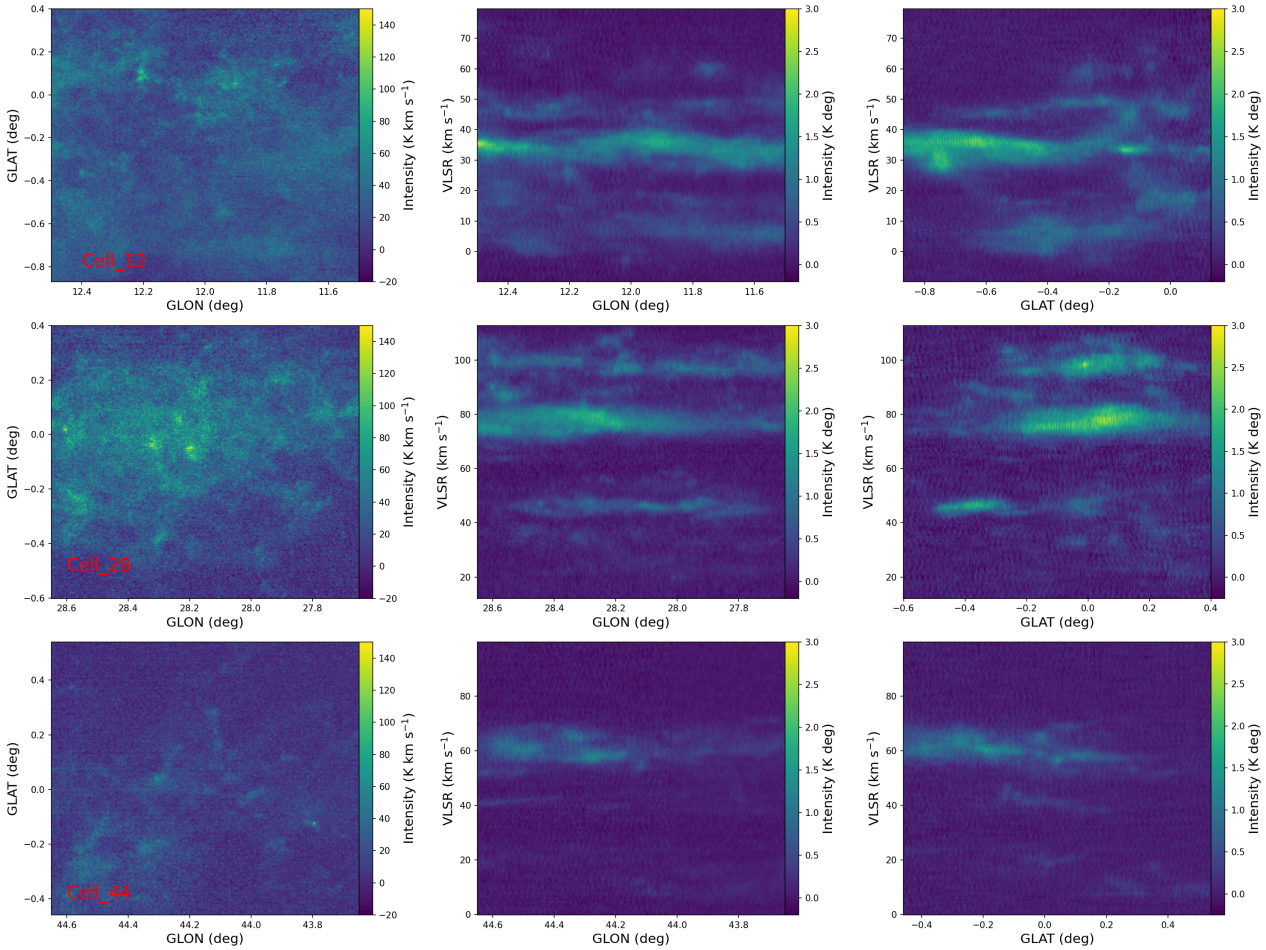

Fig. A.1. The information for Cell_12, Cell_28, and Cell_44 are shown in the top, middle, and bottom panels, respectively. The first, second, and third columns show the integrated maps for these regions in Galactic longitude, Galactic latitude, and velocity directions.

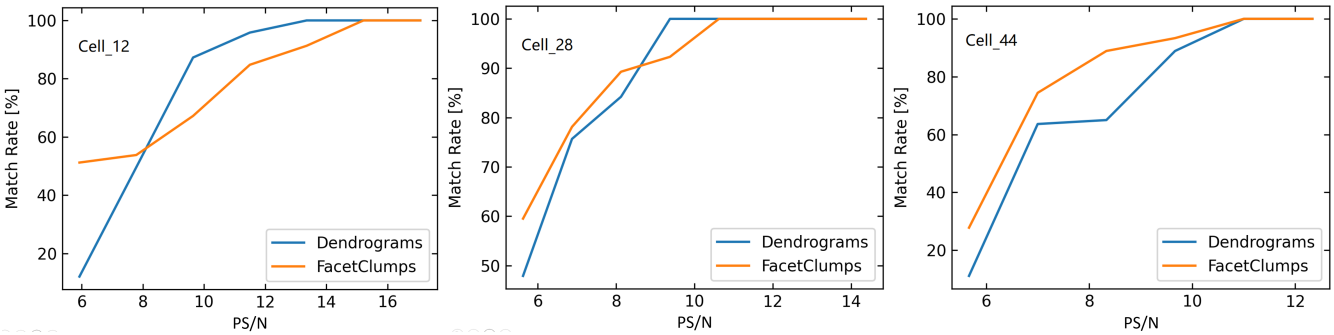

Fig. A.2. Matching ratios of FacetClumps and Dendrograms for Cell_12, Cell_28, and Cell_44 in the left, middle, and right subplots, respectively, as a function of the PS/N of the clumps.

Table A.1. The detailed information on the detection and matching results of the two algorithms across the three regions.

Data region	FacetClumps			Dendrograms		
	Detect number	Match number	Match ratio	Detect number	Match number	Match ratio
Cell_12	463	298	64.4%	318	168	52.8%
Cell_28	317	208	65.6%	198	126	63.6%
Cell_44	128	99	78.0%	88	60	68.2%

one of the clumps detected by FacetClumps (marked by a yellow line) would have a low IoU, even though both algorithms detected this clump. Table A.1 provides detailed information on the detection and matching results of the two algorithms across the three regions. We defined a matching ratio, the ratio of matches between the two algorithms to each algorithm's total number of detections. This ratio represents the matching ratio for the corresponding algorithm. By comparing the detection and matching results across different regions, it can be observed that as the data density decreases, the detection results of the two algorithms become more consistent.

Table A.1 presents the overall matching results of the two algorithms across the three regions. Clumps with higher intensities are more likely to be detected by both algorithms. To investigate this effect quantitatively, we grouped the clumps by their PS/N values and calculated the corresponding detection and matching statistics. As shown in Fig. A.2, the left, middle, and right subplots display the matching rates of FacetClumps and Dendrograms for Cell_12, Cell_28, and Cell_44 as a function of the PS/N of the clumps, respectively. As shown in Fig. A.2, when the PS/N of the clumps rises above 10, the matching rate exceeds 90% across the data of three regions with different densities, and the detection results of the two algorithms become consistent. When the matching rate of both algorithms reaches 100%, it indicates that the detection results of the two algorithms are fully consistent in terms of coverage, with every clump detected by one algorithm having at least one corresponding match in the other, and no clumps left unmatched.

As shown in Fig. A.2, in the Cell_12 data, when the PS/N exceeds 15, the matching rates of both algorithms reach 100%, indicating that all clumps above this threshold are matched between the two algorithms. In the Cell_28 and Cell_44 data, this occurs at approximately PS/N > 11. We then counted the clumps in these three regions that met these PS/N conditions. FacetClumps detected 65 clumps across these regions, while Dendrograms detected 45. The experimental results suggest that FacetClumps detects slightly more clumps than Dendrograms. One possible reason, as illustrated in the first row of Fig. A.3, is that the results detected by Dendrograms are more extended, potentially including multiple local peaks. At the same time, FacetClumps is more compact, enabling the detection of some of these local peaks.

Appendix B: Completeness Test

The completeness limit refers to the total flux or mass above which an algorithm can detect a clump at a certain level. To quantify the completeness of the extracted clumps, we conducted extensive synthesis data experiments by injecting simulated clumps into the ^{13}CO ($J = 1 - 0$) data in FUGIN. The simulated clumps had randomly assigned axis sizes and peak intensities within specified ranges: the peak intensity ranged from 1.5 to 15 K, the velocity-axis size ranged from 1 to 4 pixels, and the spatial size (in Galactic longitude and latitude) ranged from 2 to 6 pixels. These simulated clumps were randomly injected into the FUGIN data, with the constraint being to avoid overlap with another clump.

The recovery fraction as a function of the integrated flux density in different latitude intervals in QI is shown in Fig. B.1. From Fig. B.1, the completeness curves decrease sharply when the flux of clumps is lower than 500 K km s^{-1} in different latitude intervals. When the completeness reaches 80%, the total flux of clumps is approximately 466 K km s^{-1} in the QI (black line).

In Fig. B.2, the dashed line represents the completeness limit as a function of Galactic latitude. Meanwhile, the histogram shows the distribution of missed simulated clumps in each Galactic latitude bin. From Fig. B.2, it can be observed that within the Galactic longitude range of 10° to 36°, the flux corresponding to 80% completeness remains around 500 K km s^{-1}, significantly higher than the 200 K km s^{-1} in the 36° to 50°. It implies that in the 10° to 36°, only larger and brighter molecular clumps can be detected. It can be explained by the fact that numerous star-forming regions within the 10° to 36° lead to a higher data volume and density of molecular clumps, as depicted in Fig. 13.

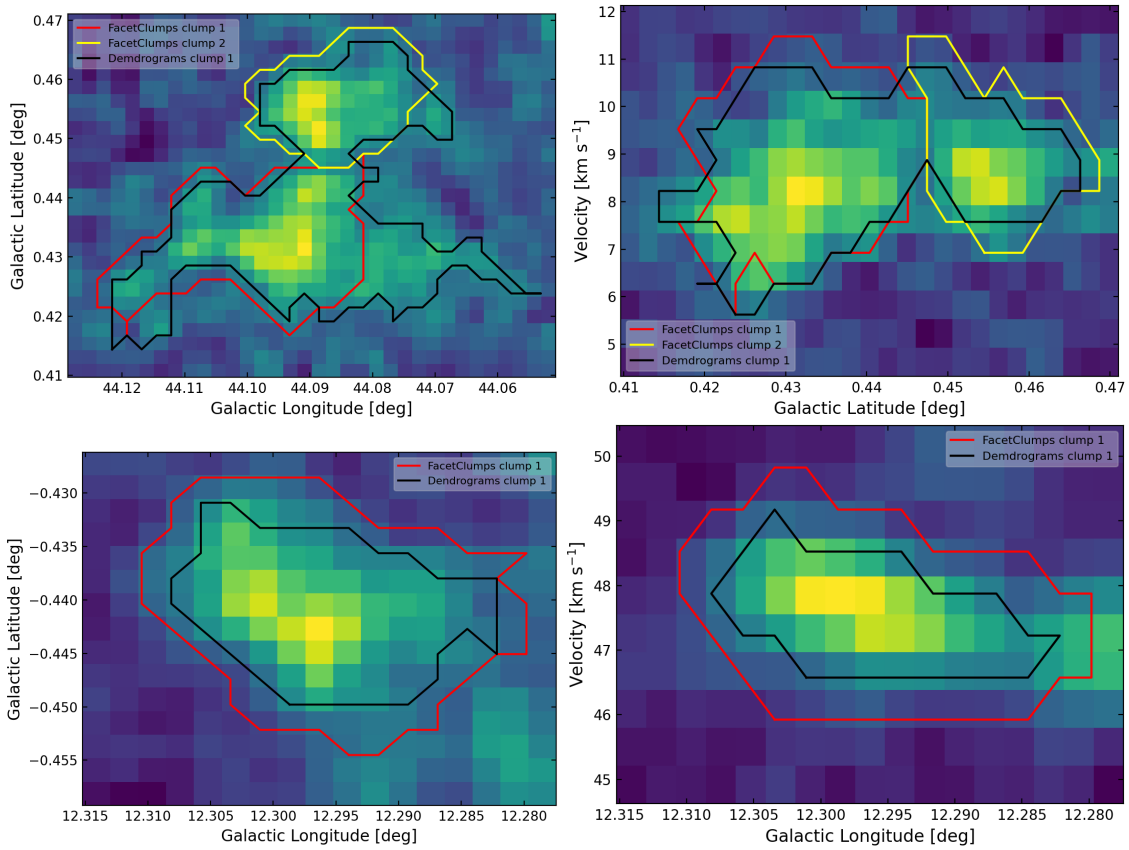

Fig. A.3. Examples of detection and matching results between FacetClumps and Dendrograms: The first row shows the matching result of two clumps detected by FacetClumps and one detected by Dendrograms. The left subplot displays the velocity-integrated map of the subregion containing the detected clumps, with the black line marking the boundary of the clump detected by Dendrograms and the red and yellow lines marking the boundaries of the two clumps detected by FacetClumps. The right subplot is similar, except the integration is along the Galactic longitude. The second row shows the matching result of one clump detected by each algorithm, with the rest of the details consistent with the first row.

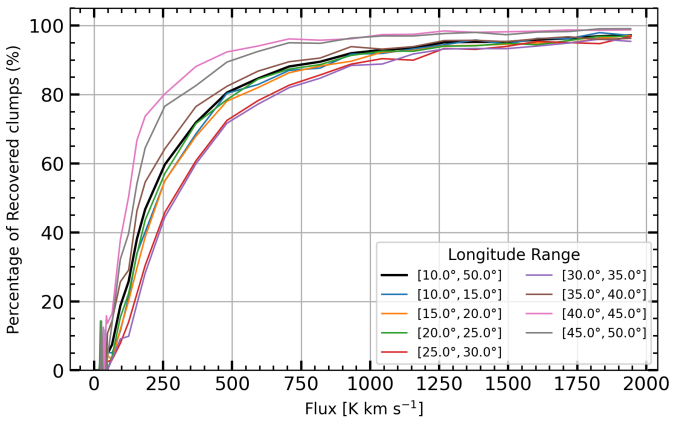

Fig. B.1. Completeness fractions as a function of flux density in different Galactic latitude intervals within QI, corresponding to molecular clumps in the extracted catalog with statistically the same sizes. The Galactic latitude range from 10° to 50° is divided into 5-degree bins. The bold black curve represents the completeness curve throughout the entire QI region.

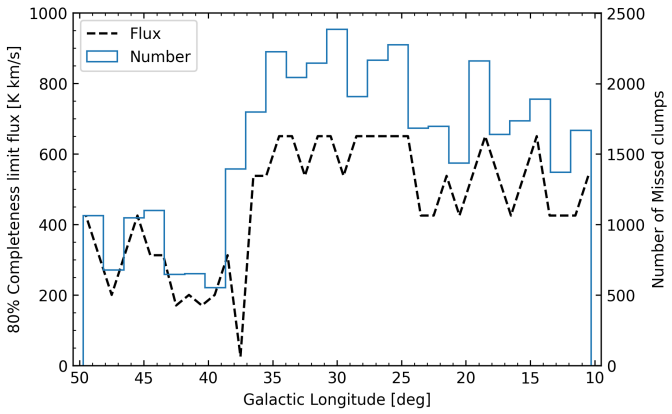

Fig. B.2. Eighty percent completeness limits in flux density for a population of clumps with the same distribution of sizes as the one extracted by FacetClumps as a function of Galactic longitude.

Appendix C: Manual Verification

A small portion of the clumps was manually verified according to the criteria described in Section 3.2 to obtain accurate labels for clump samples. To illustrate the differences between true and false clumps, Fig. C.1 presents two representative examples. The top set corresponds to a true molecular clump, while the bottom set shows a false detection. For each case, the first row displays integrated maps along the velocity, Galactic latitude and Galactic longitude axes. The second row shows the same projections, but only retaining emission within the clump boundary to highlight its morphology. The third row shows the clump's average spectrum, with the red vertical line indicating the centroid velocity and the shaded region representing the velocity range.

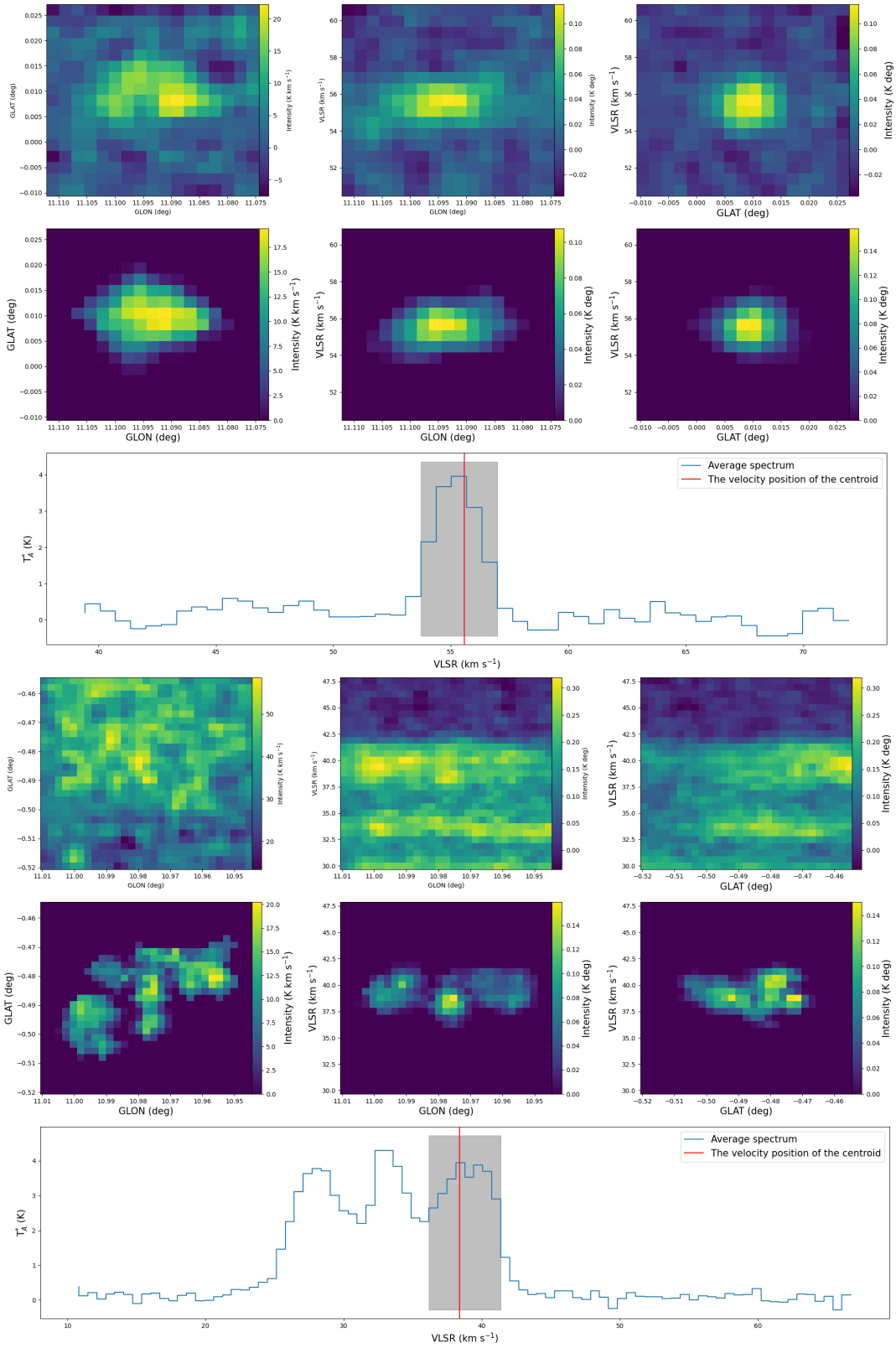

Fig. C.1. Example of a true (top) and false (bottom) molecular clump. Each example includes integrated maps (first row), masked maps retaining the clump region (second row), and the clump's average spectrum (third row), with red lines marking centroid velocity, and the shaded area indicating the velocity range.

Appendix D: The maps of the FUGIN ^{13}CO data

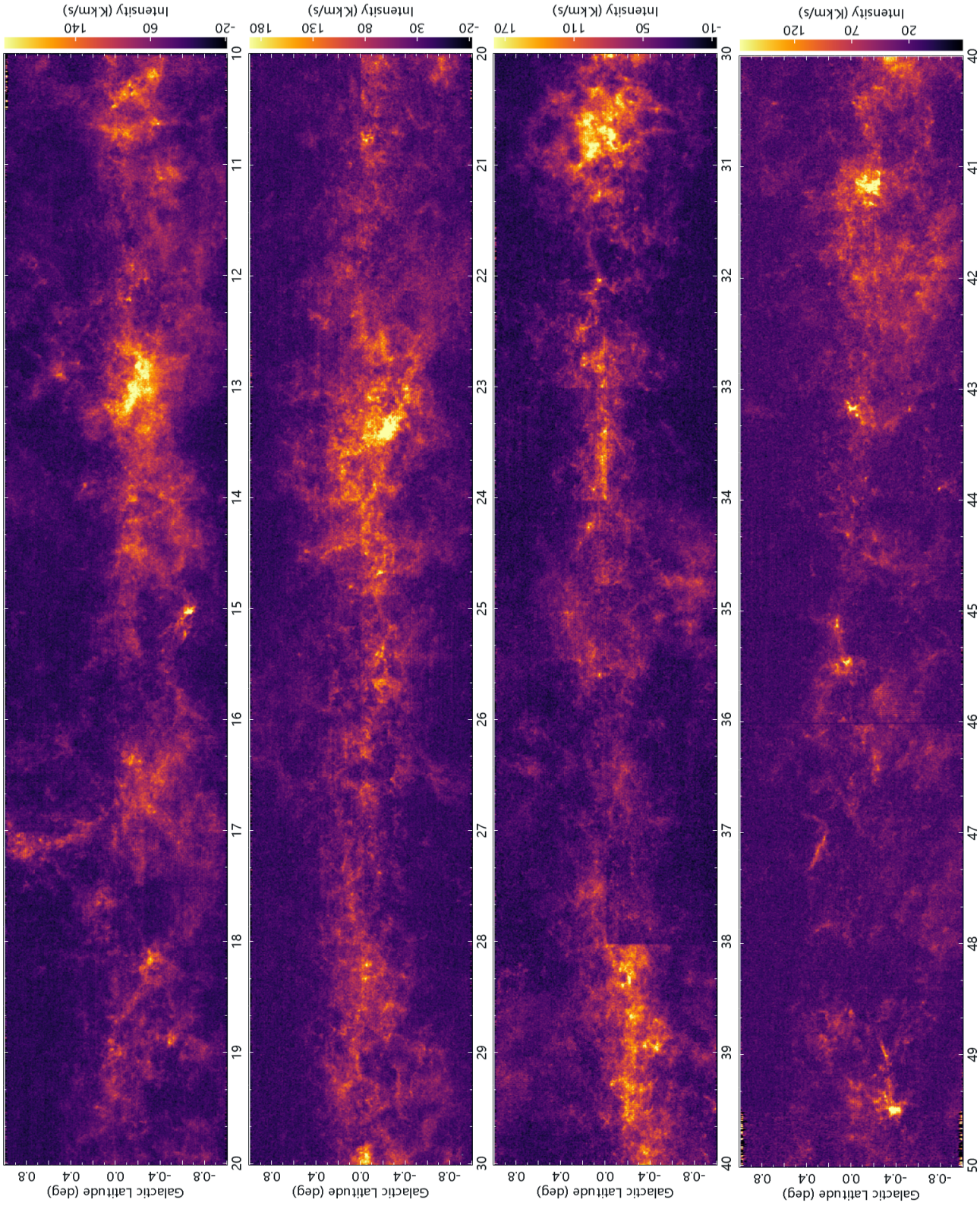

Fig. D.1. Integrated intensity maps of the ^{13}CO data over the velocity range -9 to 128 km s^{-1}. A S/N threshold of 2 was applied before integration, and pixels below this threshold were masked.