



Full length paper

The evolutionary path of star-forming clumps in Hi-GAL

Y. Maruccia^{a,b,*}, S. Cavuoti^{a,b}, M. Brescia^{c,a}, G. Riccio^a, S. Molinari^d, D. Elia^d, E. Schisano^d^a INAF - Astronomical Observatory of Capodimonte, Via Moirariello 16, Napoli, I-80131, Italy^b INFN section of Naples, via Cinthia 6, Napoli, I-80126, Italy^c Department of Physics "E. Pancini", University Federico II of Napoli, Via Cinthia 21, Napoli, I-80126, Italy^d INAF - Istituto di Astrofisica e Planetologia Spaziale, Via Fosso del Cavaliere 100, Roma, I-00133, Italy

ARTICLE INFO

Keywords:

Star formation
Machine learning
Feature engineering
Feature selection
Classification

ABSTRACT

Star formation (SF) studies are benefiting from the huge amount of data made available by recent large-area Galactic plane surveys conducted between 2 μ m and 3 mm. Fully characterizing SF demands integrating far-infrared/sub-millimetre (FIR/sub-mm) data, tracing the earliest phases, with near-/mid-infrared (NIR/MIR) observations, revealing later stages characterized by Young Stellar Objects (YSOs) just before main sequence star appearance. However, the resulting dataset is often a complex mix of heterogeneous and intricate features, limiting the effectiveness of traditional analysis in uncovering hidden patterns and relationships. In this framework, machine learning emerges as a powerful tool to handle the complexity of feature-rich datasets and investigate potential physical connections between the cold dust component traced by FIR/sub-mm emission and the presence of YSOs. We present a study on the evolutionary path of star forming clumps in the Hi-GAL survey through a multi-step approach, with the final aims of (a) obtaining a robust and accurate set of features able to well classify the star forming clumps in Hi-GAL based on their evolutionary properties, (b) establishing whether a connection exists between the cold material reservoir in clumps, traced by FIR/sub-mm emission, and the already formed YSOs, precursors of stars. For these purposes, our designed experiments aim at testing whether the FIR/sub-mm properties related to clumps are sufficient to predict the clump evolutionary stage, without considering the direct information about the embedded YSOs at NIR/MIR. Our machine learning-based method involves a four-step approach, based on feature engineering, data handling, feature selection and classification. This workflow ensures the identification of the most relevant features driving the SF process, and rigorously evaluates the results through a classification analysis. Our findings suggest that FIR/sub-mm and NIR/MIR emissions trace different evolutionary phases of star forming clumps, highlighting the complex and asynchronous nature of the SF process.

1. Introduction

Recent large-scale Galactic plane surveys conducted in the wavelength range from 2 μ m and 3 mm have provided an invaluable wealth of data for the study of star formation processes. These surveys offer unprecedented insights into the birth and evolution of stars, thanks to the vast amount of information they have made available. By observing the Galactic plane across this broad spectrum, researchers are uncovering crucial details about the conditions, mechanisms, and environments that drive the formation of stars, enhancing our understanding of this phenomenon. Star formation occurs within vast and complex structures called giant molecular clouds (GMCs). These regions contain high concentrations of gas and dust and exhibit a network of interconnected clumps and filaments, where stars are born (Molinari et al., 2010a; Ballesteros-Paredes et al., 2020; Chevance et al.,

2022). Observations have revealed that clumps within these clouds often serve as the breeding grounds for star clusters (e.g. Elia et al., 2021). Within these clumps, smaller structures known as cores emerge through fragmentation processes (e.g. Motte et al., 2022) and undergo gravitational collapse. Cores are where the formation of individual stars or bound systems of protostars takes place. Notably, young stellar objects (YSOs), which are still in their early stages of development, reside within these dusty cores (McKee and Tan, 2003; Beuther et al., 2002; Wang et al., 2011, 2012, 2014; Zhang et al., 2015; Pillai et al., 2019). They are typically faint or even undetectable in the optical band, but their emissions, stemming from the contraction of protostellar cores and the surrounding dusty circumstellar material, become visible in the near-infrared (NIR), mid-infrared (MIR), and far-infrared (FIR) wavelengths (e.g. Gutermuth et al., 2004; Whitney et al., 2004).

* Corresponding author.

E-mail address: ylenia.maruccia@inaf.it (Y. Maruccia).

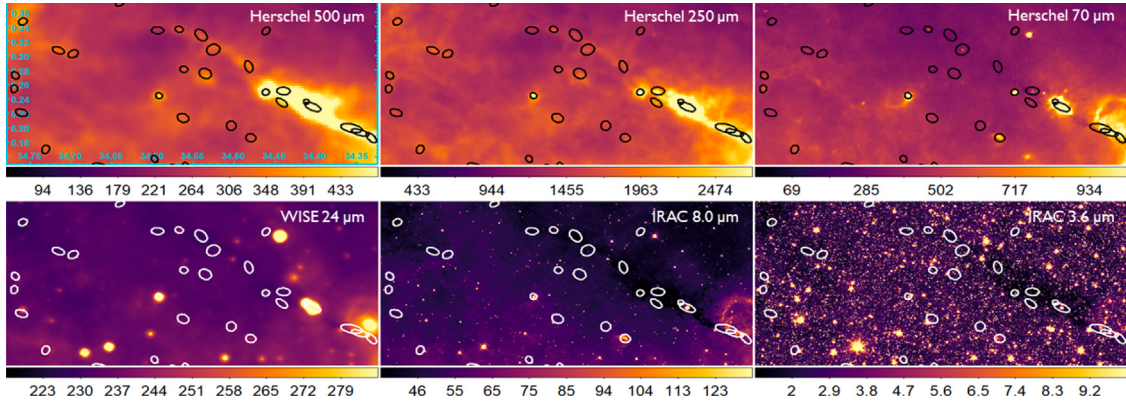


Fig. 1. Portion of the same Galactic plane at different wavelengths. Ellipses are clumps from the Hi-GAL catalog at 350 μm , and they are superimposed on different images.

During the early Class 0 stage (Andre et al., 1993), the emission of a YSO is predominantly governed by a dense circumstellar envelope. This envelope efficiently reprocesses all the radiation from the deeply embedded forming protostar into the FIR regime. At this stage, the YSO is detectable only at wavelengths longer than 60–100 μm . As the system evolves, continued accretion from the envelope progressively reduces its density. This allows shorter-wavelength radiation to escape, making the YSO observable in a broader range of wavelengths. Simultaneously, the initially toroidal structure of the envelope changes into a well-defined circumstellar disk. Accretion-driven outflows further sculpt the envelope by creating cavities, providing less obscured sightlines to the central regions. This marks the onset of the Class I stage, where the YSO becomes detectable at MIR wavelengths, extending well below 100 μm . In the subsequent evolutionary phases, namely Class II and III, the envelope continues to dissipate, resulting in a progressively less obscured view of the central protostar and its surrounding disk. During this period, the emission becomes increasingly dominated by radiation from the MIR and NIR bands, originating from the inner regions of the disk and the protostar itself, as the system transits toward the classical pre-Main Sequence phase (Greene et al., 1994).

Combining NIR to sub-millimeter (sub-mm) data provides a comprehensive view of the star formation process. The Hi-GAL (Herschel InfraRed Galactic Plane Survey, Molinari et al., 2010b) observations aptly trace the very early stages of the star formation process at clump scales, shedding light on the initial steps of the birth of YSOs. Elia et al. (2017) compiled a catalog of over 10^5 compact clumps located within the inner Galaxy, categorizing them based on their evolutionary stages. These clumps, with estimated sizes ranging from 0.1 to 1 pc, are substantially larger than the dense cores, typically measured at scales between 0.01 and 0.1 pc (Bergin and Tafalla, 2007). This disparity suggests that these clumps likely harbor clusters of YSOs. However, the limited angular resolution of single-dish telescopes operating in the FIR and sub-mm regimes makes it challenging, if not impossible, to resolve individual YSOs within these clumps. To overcome this limitation, FIR/sub-mm imaging must be combined with higher-resolution NIR/MIR observations, which provide a clearer view of individual YSOs. Surveys such as *Spitzer* and *WISE* have been instrumental in this regard.

Over the past few decades, investigations conducted in NIR and MIR spectra further have enhanced our understanding of the interstellar medium (ISM), identifying deeply embedded objects, and distinguishing YSO populations from background field stars. Observatories like the *Spitzer Space Telescope* (Werner et al., 2004) and the *Wide-field Infrared Survey Explorer* (WISE, Wright et al., 2010) have been particularly effective in detecting deeply embedded YSOs, providing insights into the later phases of star formation.

To illustrate the advantage of multi-wavelength observations, Fig. 1 presents a portion of the Galactic plane observed across different wavelengths, with Hi-GAL clumps overlaid as ellipses. This visualization

helps identify potential YSO counterparts (Maruccia et al., 2025; Koenig and Leisawitz, 2014; Evans et al., 2009) and their spatial distribution within the star forming regions (Retter et al., 2021; Megeath et al., 2015; Yadav et al., 2016). Furthermore, this approach facilitates the modeling of the spectral energy distribution (SED) of star-forming clumps across a wide wavelength range, providing critical statistical information to characterize their evolutionary stages (e.g. Persi et al., 2016; Tapia et al., 2018).

However, the final dataset is often made up of a rich set of heterogeneous and intricate features, and traditional analytical methods could not be able to uncover hidden patterns and relationships within this complex data landscape.

In this perspective, this work aims at characterizing the evolutionary path of star forming clumps in Hi-GAL by employing a systematic four-step approach. First, we perform feature engineering to construct and refine a comprehensive dataset of multi-wavelength parameters, characterizing the physical and observational properties of clumps. Second, we apply feature selection techniques to identify the most relevant and discriminative parameters for describing the star formation processes. Finally, we conduct a classification analysis using the full original parameter space and compare it with the reduced space defined by the selected features. Our goal is to identify key features, eliminate redundant parameters, and reveal hidden relationships within the dataset. This will enable a more precise and insightful characterization of star formation processes, with the final aims of: (a) obtaining a robust and informative set of features capable of accurately classifying the clumps according to their evolutionary properties, and (b) investigating whether a connection exists between the cold material reservoir within the cold dust component traced by FIR/sub-mm emission and the already formed YSOs, which are the precursors of stars. To this purpose, we design our experiments to assess whether the FIR/sub-mm properties alone are sufficient to predict the evolutionary stage of a clump, without relying on direct information from NIR/MIR tracers of the embedded YSOs. To the best of our knowledge, this data-driven approach is the first attempt aimed at replacing traditional hand-crafted analyses in this context. This motivated our focus on the exploration of the parameter space to identify the physical features that most effectively discriminate between different evolutionary stages of clumps. In this sense, the pioneering use of data-driven techniques for parameter space exploration on this astrophysical topic has the main merit of inferring possible new and deeper knowledge on the physics of the problem.

Outline: The paper is structured as follows. Section 2 presents the dataset, followed by Section 3, which details the feature engineering process. In Section 4, we outline the feature selection and classification methods, while Section 5 describes our experiments. Finally, we discuss the results in Section 6 and conclude in Section 7.

2. Hi-GAL clumps and NIR/MIR: building a multi-faceted dataset

In the attempt to better investigate the star formation process, we adopt a collection of catalogs at different wavelengths. In particular, we employ the following surveys, as they provide near-complete coverage from the NIR/MIR to the sub-mm wavelength range, thus sampling the full SED of the clumps and their embedded YSOs:

- the Hi-GAL photometric catalog.¹ Introduced by [Molinari et al. \(2016\)](#), this comprehensive catalog of compact sources within the inner regions of the Milky Way, contained in a 2° latitude strip, comprises data gathered from the five Herschel bands at 70, 160, 250, 350, and 500 μm , each offering distinct perspectives with nominal angular resolutions of $5''$, $12''$, $18''$, $24''$, and $36''$, respectively. However, due to the limited available transmission bandwidth, the PACS 70 μm and 160 μm channels were co-added on-board Herschel. The resulting instrumental beams are $5.8'' \times 12.1''$ and $11.4'' \times 13.4''$, respectively. Both beams are elongated along the scan direction ([Molinari et al., 2016](#)). Compact sources were extracted using the CuTex algorithm ([Molinari et al., 2011](#)), specifically designed for detecting sources in a highly complex and non-uniform background. For our analysis, we use both the Hi-GAL fluxes (from 70 μm to 500 μm) and the physical parameters of the clumps. For details on how these parameters were derived, we refer the reader to [Elia et al. \(2021\)](#);
- the GLIMPSE I (Galactic Legacy Infrared Mid-Plane Survey Extraordinaire) catalog² ([Benjamin et al., 2003](#); [Churchwell et al., 2009](#)). This survey, conducted in the mid-infrared range, spans a critical longitude range between $10^\circ < \ell < 65^\circ$, $|b| \leq 1^\circ$, focusing on the inner regions of the Galactic plane. GLIMPSE I captures data across a range of wavelengths, in the 3.6, 4.5, 5.8, and 8.0 μm bands, by using the IRAC instrument ([Fazio et al., 2004](#));
- the AllWISE Source Catalog.³ This survey encompasses a vast array of data obtained from the Wide-field Infrared Survey Explorer (WISE) mission, which mapped the entire sky in four infrared bands: 3.4, 4.6, 12, and 22 μm (hereinafter, w1, w2, w3, and w4, respectively), boasting angular resolutions of 6.1, 6.4, 6.5, and 12 arcseconds, respectively ([Wright et al., 2010](#));
- the UKIDSS Galactic Plane Survey (GPS).⁴ Conducted by the United Kingdom Infrared Deep Sky Survey (UKIDSS) project, the GPS utilizes the United Kingdom Infrared Telescope (UKIRT) to observe the Galactic Plane at 1.25, 1.65, and 2.20 μm (namely, J, H, and K), covering $\sim 1868 \text{ deg}^2$ of the north plane, in the coordinate ranges $15^\circ < \ell < 107^\circ$ and $141^\circ < \ell < 230^\circ$, with $|b| < 5^\circ$ ([Lucas et al., 2008](#)).

3. Feature engineering

We follow the feature engineering workflow described in detail by [Maruccia et al. \(2025, hereinafter M25\)](#) and illustrated in [Fig. 2](#). Our goal is to create a multi-wavelength catalog that statistically and physically describes all clumps within the Galactic plane region $10^\circ \leq \ell \leq 60^\circ$, $|b| \leq 1^\circ$. We briefly summarize the whole procedure in the following.

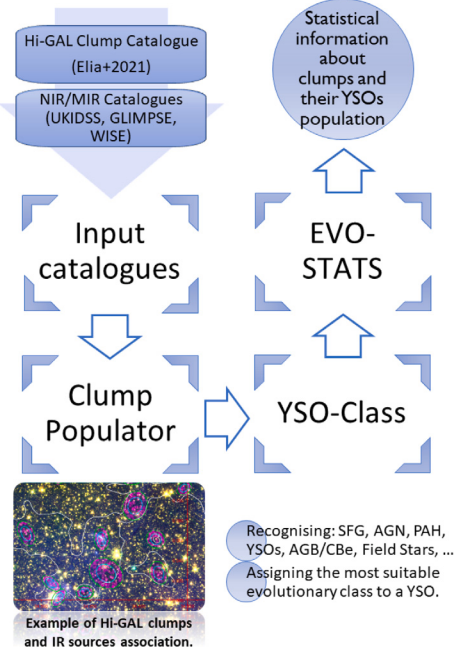


Fig. 2. The feature engineering process adopted for obtaining a multi-wavelength catalog of both statistical and physical features of the Hi-GAL clumps within the Galactic plane region $10^\circ \leq \ell \leq 60^\circ$, $|b| \leq 1^\circ$. For further details, see [M25](#).

3.1. ClumpPopulator

The first module is called *ClumpPopulator*, a specifically tailored version of the C^3 tool ([Riccio et al., 2017](#)). The *ClumpPopulator* module identifies all IR sources located within Hi-GAL clumps. Each clump is represented as an elliptical region centered on its Galactic coordinates, with axes defined by the Full Width at Half Maximum (FWHM) of a 2D Gaussian fit to its 350 μm source profile. Given the Hi-GAL clumps, the *ClumpPopulator* module has been run for three times, thus obtaining three lists of associations clumps-IR sources: one for GLIMPSE, one for WISE, and one for UKIDSS, respectively.

3.2. YSO-Class

The three lists in output from the previous step, have been then positionally cross-matched for obtaining a multi-wavelength sources catalog, which has been given in input of the second module of our evolutionary workflow, called YSO-Class. This component is crucial, as it assigns the most suitable evolutionary class to a source, investigating its color-color or magnitude-color diagrams through a set of specific prescriptions. These prescriptions finally classify each IR sources, distinguishing whether it belongs to one of the following categories: Star Forming Galaxies (SFGs), Active Galactic Nuclei (AGNs), unresolved knots of shock emission, sources where polycyclic aromatic hydrocarbons (PAH) emission contaminates the photometric apertures, YSOs (Class I/II), AGB/CBe Stars, Field Stars ([Gutermuth et al., 2008, 2009](#); [Koenig and Leisawitz, 2014](#)).

3.3. Evo-stats

A further step after classification, called *Evo-Stats*, provides the statistical information about the YSOs population associated to each Hi-GAL clump. Indeed, in this step the data are aggregated to extract key information about each clump. This includes determining the number and evolutionary stage distribution of YSOs, identifying the detected field stars, quantifying and categorizing contaminants, and assessing the spatial distribution and density of the YSOs within clumps.

¹ This dataset is available in VIALACTEA project, at http://vialactea.iaps.inaf.it/vialactea/public/HiGAL_360_clump_catalogue.tar.gz, and published in [Elia et al. \(2017\)](#).

² This catalog is available at <https://irsa.ipac.caltech.edu/data/SPITZER/GLIMPSE/overview.html>.

³ Data can be downloaded at <https://irsa.ipac.caltech.edu/Missions/wise.html>.

⁴ The UKIDSS Eighth Data Release (DR8), including catalogues and images, may be accessed from the WFCAM Science Archive (WSA) at <http://wsa.roe.ac.uk/>

Furthermore, for each clump a *clump class* is assigned, calculated as the average class of all the YSOs within the clump. This classification provides an intermediate value between the canonical Classes I and II, serving as an indicator of the clump's overall evolutionary stage.

3.4. Resulting features

From the evolutionary workflow outlined in Sections 3.1–3.3, we obtain a final catalog of the Hi-GAL clumps within the Galactic region $10^\circ \leq \ell \leq 60^\circ$, composed by 6940 clumps populated by a total number of 10,600 YSOs. In this catalog, each clump is described by a large set of features, which can be divided into photometric and physical features.

The photometric features refer to the fluxes and their combinations useful to derive the colors characterizing the clump emission. Firstly, it is worth noting that, since each clump may host multiple YSOs emitting in the NIR/MIR range, we need to characterize these clumps by aggregating the properties of the individual YSOs within them. To this aim, fluxes and magnitudes of the individual YSOs are extracted from the respective catalogs presented in Section 2. Unlike previous studies, if a YSO is not detected at some bands, we manage these missing values adopting the sensitivity limits provided by the respective surveys. Adopting these limits in cases of non-detection enables us to retain and maximize the use of available information, rather than discarding potentially valuable data. The adopted sensitivity limits in magnitude (corresponding to upper limits in flux) are: 15.5, 15.0, 13.0, and 13.0 for the *GLIMPSE* bands at 3.6, 4.5, 5.8, and 8.0 μm (Churchwell et al., 2009); 16.9, 16.0, 11.5, and 8.0 for the *WISE* bands w1 through w4 (Cutri et al., 2013); and 19.8, 19.0, and 18.1 for the *UKIDSS* J, H, and K bands, respectively (Hambly et al., 2008). Moreover, the Galactic plane in the range $10^\circ \leq \ell \leq 60^\circ$ is heavily affected by interstellar extinction. Therefore, to reliably apply color–magnitude diagnostics to stellar sources in this region, it is necessary to correct their observed fluxes for dust extinction along the line of sight. We adopt the extinction laws from Rieke and Lebofsky (1985) for the *JHK* bands and from Xue et al. (2016) for the MIR bands (*GLIMPSE* and *WISE*). Additionally, we assume an average extinction of 1 A_V per kiloparsec of heliocentric distance (Juvela and Montillaud, 2016). Since these YSOs belong to clumps, the clump distances, available in the Hi-GAL catalog, are used to estimate the extinction correction. More details about these steps can be found in M25. Starting from these data, we compute the following photometric features: the total flux for each band, calculated as the sum (hereinafter referred to as “sum”) of the fluxes of all YSOs within the same clump; the median (referred to as “med”) of the YSO fluxes within the clump; the absolute value of the difference between the average and, respectively, the minimum and the maximum of the YSOs fluxes within the clump (referred to as “D1” and “D2”, respectively). As regards the colors, we combine the magnitudes of each individual YSO, obtaining six colors for *GLIMPSE* and *WISE* and three for *UKIDSS*. Subsequently, we aggregate these values for each clump, taking into account the minimum value (“min”), the maximum (“max”), the median (“med”), and, as done for the fluxes, the absolute value of the difference between the average and, respectively, the minimum and the maximum of the given color (“D1” and “D2”). To provide a more comprehensive description of the nature of clumps in the MIR/FIR range, we utilize the clump fluxes at the wavelengths investigated by Herschel (as described in Section 2), along with the following colors: $[w4-70]$, which is a combination of the *WISE* flux at 22 μm and the Herschel 70 μm , $[70-250]$, and $[250-350]$. Finally, all fluxes have been rescaled to a reference distance of 1 kpc using the known distances of clumps.

The physical features mainly refer to the Hi-GAL clumps properties presented in the Elia et al. (2021)’s catalog. In particular, they include the bolometric luminosity (L_{bol}), the mass (M), the dust temperature (T_{dust}), the bolometric temperature (T_{bol}), the surface density (Σ), the L/M ratio, the ratio between the bolometric luminosity and its fraction computed over the range $\lambda \geq 350 \mu\text{m}$ (L_{ratio}). We refer to the

original paper for further details on how these physical properties were calculated.

Finally, we consider the YSO density (δ_{YSO}), a feature generated by the YSO–Class procedure, and calculated as the fraction of the YSOs among the total number of sources falling within the same clump. Table 1 summarizes all the features that we want to investigate in our work through experiments of feature selection and classification.

4. Workflow

The workflow adopted in this paper involves a structured, multi-step approach to ensure the identification of the most relevant features involved in the star formation process and evaluate the results with a classification step. This process, depicted in Fig. 3, is made up of the following steps:

- Step 1: Feature Engineering. It has already been discussed in detail in Section 3 and represented in Fig. 2. This step involves the construction of a parameter space (hereinafter, PS) that encapsulates the key aspects of the star formation process. Utilizing the input catalogs as a foundation, we systematically derived a comprehensive set of features essential for describing and analyzing the phenomenon under investigation.
- Step 2: Data handling. It consists in preparing the dataset for the next stages of analysis, in particular cleaning the missing data, normalizing the features to ensure equal contribution of each feature and prevent scale-related biases, and injecting random noise to improve model robustness, prevent overfitting, and increase diversity in the parameter space. More details are provided in Section 4.1.
- Step 3: Feature Selection. In this stage different algorithms have been used for identifying the most relevant features in the PS, as well described in Section 4.2. Results are shown in Section 5.1.
- Step 4: Classification. The final step involves classification using the Light Gradient Boosting Machine (hereinafter, LightGBM) to compare results across different PS. LightGBM’s leaf-wise growth and histogram-based algorithm improve accuracy and efficiency over traditional GBDT methods. More details are provided in Section 4.3. Results are discussed in Section 5.2.

4.1. Data handling

The second step of the workflow consists in managing our dataset, to ensure consistency and mitigate biases. This phase encompasses three processes aimed at preparing data for subsequent analysis: cleaning missing data, feature normalization, and the injection of random noise. The first process involves cleaning missing data, replacing missing values with statistical measures or removing incomplete entries, generating new classes, and removing irrelevant features. These preliminary actions are essential for ensuring that the dataset is both accurate and relevant. Normalizing the data is an important step (Huang et al., 2023), particularly when the features have different scales. Techniques like Min–Max normalization or Z-score standardization are the most popular choice for normalizing data (Singh and Singh, 2022) and ensure that each feature contributes equally to the model by adjusting the range of the data. This helps avoid problems like certain features dominating the model due to their larger scale or dynamics. Furthermore, random noise can be injected into the data (Grandvalet et al., 1997; Hua et al., 2006), introducing artificial features that help improve the robustness of the model by preventing it from overfitting to the noise present in real-world data. This technique can also help increase the diversity of the PS, making the model more generalizable. Once the data has been cleaned and normalized, the dataset is ready to be used in the next steps of our analysis.

Table 1

Overview of all the features adopted for and obtained from feature selection in our experiments. The first half of the table (fluxes and colors) is related to the photometric features, while the second half corresponds to the physical features. See the text for further details.

	Source	Features	Base PS	Enriched PS
Fluxes	<i>GLIMPSE</i>		—	D1_i4, sum_i4
	<i>WISE</i>	sum, med D1, D2 of individual fluxes	—	D2_w1, sum_w1, med_w1, sum_w2, med_w2, D1_w3, D1_w4, sum_w4, med_w4
	<i>UKIDSS</i>		—	D2_J, med_J, D2_H
	<i>Hi-GAL</i>	$\bar{F}_{70}, \bar{F}_{160}, \bar{F}_{250},$ $\bar{F}_{350}, \bar{F}_{500}$	$\bar{F}_{160}, \bar{F}_{500}$	—
Colors	<i>GLIMPSE</i>		—	D2_i1-i2, min_i1-i2, med_i1-i2, max_i1-i2, med_i1-i3, max_i1-i3, med_i1-i4, min_i2-i3, med_i2-i3, max_i2-i3, min_i2-i4, med_i2-i4, max_i2-i4, min_i3-i4, med_i3-i4, max_i3-i4
	<i>WISE</i>	min, max, med D1, D2 of individual colors	—	min_w1-w2, med_w1-w2, max_w1-w2, min_w1-w3, med_w1-w3, max_w1-w3, min_w2-w3, med_w2-w3, max_w2-w3
	<i>UKIDSS</i>		—	min_H-K, med_H-K, max_H-K
	<i>Hi-GAL</i>	$[70-250],$ $[250-350]$	$[250-350]$	—
	<i>WISE/Hi-GAL</i>	$[w4-70]$	$[w4-70]$	—
	<i>Hi-GAL</i>	$M, L_{\text{ratio}},$ $L/M, L_{\text{bol}},$ $T_{\text{dust}}, T_{\text{bol}},$ Σ	$L_{\text{ratio}},$ $T_{\text{bol}},$ Σ	—
Physical	<i>YSO-Class</i>	$\bar{\delta}_{YSO}$	$\bar{\delta}_{YSO}$	$\bar{\delta}_{YSO}$

4.2. Feature selection

Finally, in the third step of our workflow we focus on feature selection (hereinafter, FS), a critical phase for identifying the most relevant features in our PS. In this phase, we adopt two different algorithms for the FS. In particular:

- the Recursive Feature Elimination (RFE), which evaluates recursively the importance of a feature by removing one feature at a time from PS and calculating its impact on classification prediction. Then, it selects those features (columns) in a training dataset that are more or the most relevant in predicting the target variable.
- the Boruta algorithm (Kursa and Rudnicki, 2010), which identifies the most relevant features in a dataset by comparing their importance to randomly generated “shadow features”. It begins by creating a copy of the original dataset’s features, where the values in each column are shuffled to eliminate any inherent relationship with the target variable. These shuffled features, known as shadow features, are then merged with the original dataset to form a new feature space whose dimension is twice the original dataset. At this point, the Random Forest (Breiman, 2001) works on these extended features space and determines their importance (Kursa and Rudnicki, 2011) using a statistical test, often the Z-Score, which quantifies the contribution of a feature to the model’s performance. The algorithm checks if each original feature has a higher importance score than the maximum importance of the shadow features. If an original feature has a higher importance score than all shadow features, the feature is considered significant and kept in the dataset. Otherwise, features having a lower score than the shadow features are considered insignificant and dropped from the dataset. Features whose importance cannot be definitively determined are labeled as tentative and re-evaluated in subsequent iterations. In each

iteration, the Boruta algorithm uses only the retained and tentative features to create a new set of shadow features, repeating the process of evaluating their importance relative to noise. This iterative approach ensures that features are either confirmed as significant or discarded as irrelevant. This is repeated until a specified number of iterations has been achieved or when all features have been confirmed or dropped. It is worth noting that, unlike traditional FS techniques, the Boruta algorithm is tasked with finding a minimal optimal feature set rather than finding all the features relevant to the target variable. By using shadow features as a baseline, it offers a stringent and interpretable mechanism to filter out noise and irrelevant variables, making it especially effective in high-dimensional datasets where irrelevant features could obscure meaningful patterns.

Furthermore, each FS algorithm is paired with two methods for calculating the feature importance:

- the standard tree-based importance, which evaluates the contribution of each feature considering the standard tree-based feature importance calculation. This evaluation is typically based on metrics like the Gini importance or mean decrease in impurity (Breiman, 2017), which measure how much a feature reduces uncertainty or improves model performance at decision splits across the trees in the ensemble. Features are ranked according to the magnitude of their importance scores, and the less important features (i.e. those with lower scores) are considered redundant or noisy as they contribute minimally to reducing prediction error, thus they are excluded from the final selection. This approach not only simplifies the model by reducing the number of input features but also helps prevent overfitting, as irrelevant features can introduce noise and decrease the model’s generalization capability.
- the SHAP (SHapley Additive exPlanations, Lundberg, 2017) importance, which leverages the Shapley’s values, a concept rooted

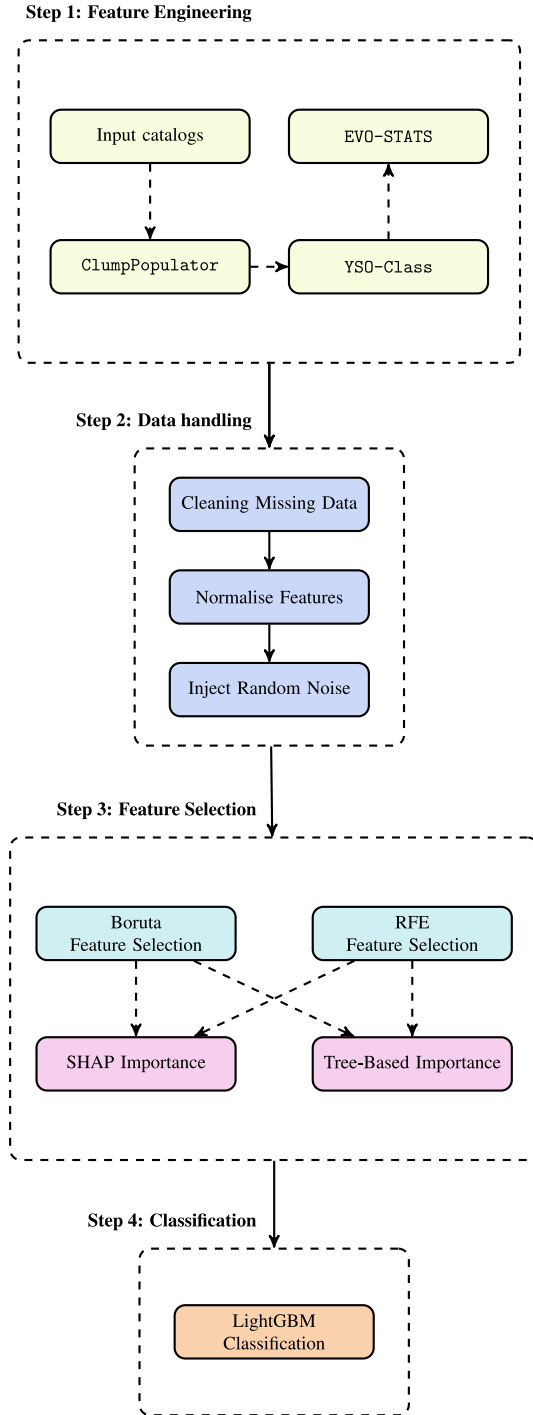


Fig. 3. The four-step workflow adopted in this work. Firstly, the feature engineering step builds the parameter space from several input catalogs (**Step 1: Feature Engineering**). More details about this step can be found in Section 3, while the feature engineering process is depicted in Fig. 2. After cleaning and normalizing the dataset (**Step 2: Data Handling**), the most relevant features are identified (**Step 3: Feature Selection**) and used for classification tests (**Step 4: Classification**).

in game theory, to calculate the contribution of each feature to a model's predictions. This approach determines a feature's importance by comparing the model's predictions with and without the inclusion of that feature, effectively measuring how much the feature influences the output. Inspired by game theory, which evaluates the role of individual players in a collaborative team, SHAP adapts this framework to assess features within a machine

learning model. In game theory, a player can either participate in or abstain from a game, and SHAP translates this idea by defining a feature as “joining the model” when its value is known and “not joining the model” when its value is missing or replaced. The calculation involves considering all possible combinations of features, assessing how the model's output changes when the feature of interest is added or removed in each scenario. This ensures that SHAP provides a fair and consistent estimate of feature importance, accounting for the interactions between features. The result is a detailed, instance-specific measure of importance that not only identifies which features are influential but also quantifies their impact in a way that is interpretable and intuitive, making SHAP a powerful tool for explaining complex models.

This results in four main combinations (RFE + standard tree-based importance, RFE + SHAP importance, Boruta + standard tree-based importance, Boruta + SHAP importance), and we evaluate the each outcomes to determine which combination performs best for our PS. The selected features from this step will then be used in the subsequent classification stage.

4.3. LightGBM classification

We now focus on the classification, with the primary goal of comparing results across the different PS identified in the FS phase. For this purpose, we employed LightGBM, an advanced implementation of the Gradient Boosting Decision Tree (GBDT) algorithm (Ke et al., 2017). LightGBM avoids the typical hierarchical growth approach used by most GBDT algorithms and instead employs a leaf growth algorithm with depth constraints. In LightGBM, the tree grows leaf-by-leaf, where the leaf with the highest split gain among all current leaves is selected and split, and this process continues iteratively. This leaf-by-leaf growth method tends to reduce errors and achieve better accuracy with the same number of splits compared to the traditional layer-by-layer approach. The histogram for the next layer node is calculated by finding the difference between its parent and sibling nodes. The histogram is computed by traversing the leaf nodes according to the growth strategy. LightGBM uses a histogram-based algorithm, which has the advantage of efficiently calculating the histogram of a leaf node while simultaneously obtaining the histogram of its sibling node at half the computational cost, making the process twice as fast (Ke et al., 2017).

5. Experiments

As described in the previous sections, all the considered features are traditionally essential for comprehensively describing the star formation process. They encompass various evolutionary phases observable across a broad spectrum of wavelengths. FS serves as a crucial step in streamlining our analysis. By identifying and eliminating redundant features, we can focus solely on those most pertinent to characterize the star formation process effectively.

We designed two experiments based on different PS. In the first one, referred to as the *base PS*, we include fluxes and colors of the Hi-GAL clumps, along with their physical properties. The second experiment builds upon this feature set by incorporating the additional photometric features described in Section 3. Hereinafter, we will refer to this PS as *enriched PS*. For studying the clump evolution, we take into account the *clump class* as target, defined as the average of the classes of all the YSOs associated with a given clump, as described in Section 3.3. This setup serves a dual purpose. First, it allows us to obtain a robust and accurate classification of the Hi-GAL clumps in terms of their evolutionary properties. Second, it enables us to assess whether there is a link between the cold material traced in the FIR/sub-mm (associated with the clumps themselves) and the stellar content already formed within them, observable in the NIR/MIR. Moreover, since the clump class is

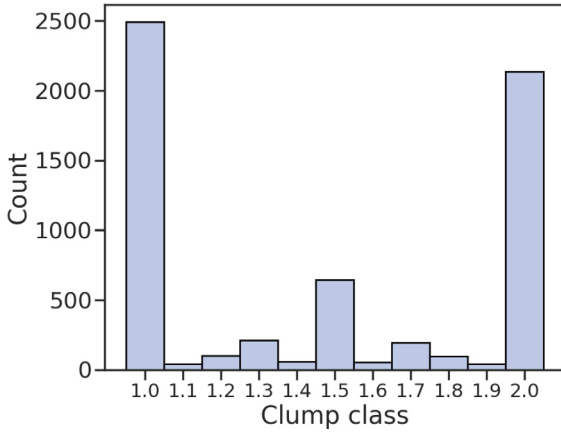


Fig. 4. Clump class distribution.

derived through a process that involves the additional features present only in the *enriched PS*, our experimental design is also aimed at testing whether the *base PS* alone – i.e., without direct information about the embedded YSOs – is sufficient to predict the clump evolutionary stage. This provides insights into the predictive power of the clump’s intrinsic properties and their observed fluxes/colors.

The starting dataset is made up of 6079 objects. During the pre-processing step, four clumps have been discarded as they presented *NaN* values in correspondence of some physical features, thus the final clump dataset is made up of 6075 objects. Before going on with our analysis, the dataset has been normalized using the Z score. In Fig. 4 it is shown the distribution of the clump classes for our dataset. Clumps with Class 1.0 and 2.0 are well distributed in the dataset, and are more numerous than clumps with intermediate classes. For our experiments, we decide to label as Class 0 all clumps with class ≤ 1.5 , as Class 1 all clumps with class > 1.5 , resulting, respectively, 3552 (58%) and 2523 (42%) clumps. Finally, in all the following experiments the train+validation/test split was always based on, respectively, 70% training and validation (the latter amounting to 10%) and 30% test of the dataset.

5.1. Feature selection

FS is an important aspect of our workflow. It aims at simplifying complex datasets into a more manageable form without losing essential information, and selecting only those features that are more involved in the description of the star formation process. Since we are looking to all the relevant features this should bring also useful insight on the physics driving the selection. We perform the FS on the two different PS, the *base PS* made up of 16 features and the *enriched PS*, made up of 135 features. Moreover, as described in Section 4.2, we combined both the RFE and the Boruta algorithm with two measures of importance, the standard tree-based and the SHAP importances. In Table 2 results of these four experiments are summarized. In the experiments with the standard tree-based importance, some of the shadow features have been selected as good features, thus invalidating the final results. The best solution is that one in which we adopted the combination of Boruta and the SHAP importance, both for the *base PS* and the *enriched PS*.

As it can be seen, when considering the *base PS*, the algorithm selected 8 features, while for the *enriched PS* the algorithm selected 43 features. In the third and fourth columns of Table 1, the selected features are indicated. In the experiment with the *enriched PS* all the selected features are related to the NIR/MIR properties, which are effectively used in the workflow that computes the *clump class* starting from the color–color diagrams. Only one selected feature, the δ_{YSO} , is not related to the MIR properties. On the contrary, in the experiment with

Table 2

Results of the FS on both the two PS, using the combination of RFE + standard tree-based importance, RFE + SHAP importance, Boruta + standard tree-based importance, Boruta + SHAP importance. The numbers indicate the total number of features selected for each experiment, while the row *All features* indicates the initial number of the features in each PS. The colored row indicates the best solution obtained.

Experiment	<i>base PS</i>	<i>enriched PS</i>
All features	16	135
RFE + Std tree-based	15	102
RFE + SHAP	14	98
Boruta + Std tree-based	4	28
Boruta + SHAP	8	43

Table 3

LightGBM and experiment setup.

LightGBM and experiment setup	Base PS All features	Base PS Sel. Features	Enriched PS All features	Enriched PS Sel. Features
PS dimension	16	8	135	43
FS method	–	Boruta+SHAP	–	Boruta+SHAP
Learning rate	0.2	0.2	0.01	0.2
$N_{estimators}^a$	800	800	800	800
Num_leaves^b	25	40	35	35
Validation set score ^c	0.605	0.606	0.122	0.115
Blind test set score	0.655	0.656	0.953	0.953

^a Number of boosting rounds.

^b Maximum number of leaves in one tree.

^c Validation score computed on a held-out validation set.

the *base PS* (without the MIR properties), all the selected features are related to the clump properties. Also in this case the δ_{YSO} is selected. Finally, only the color [w4–70] contains the information from the MIR. Further details about the role of these indicators in understanding the evolution stage of clumps are discussed in Section 6.

5.2. Classification

As explained in Section 4.3, the results obtained previously by the FS phase are finally used in the classification step. In particular, we run the LightGBM for obtaining a classification in four cases:

- *base PS* and all the 16 features;
- *base PS* and the selected 8 features;
- *enriched PS* and all the 135 features;
- *enriched PS* and the selected 43 features.

The hyperparameters used for these experiments are reported in Table 3. They were selected using a standard grid search optimization strategy. All remaining LGBM hyperparameters were kept at their default values⁵ as provided by the official implementation.

In Fig. 5 we present the results of the four classification experiments, reporting for each case the corresponding confusion matrix. In particular, the top panels illustrate the outcomes obtained when adopting the *base PS* (Fig. 5(a)), made up of 16 features, alongside the results obtained using the corresponding 8 features (Fig. 5(b)) selected during the FS step. As expected, when the classification is performed on the PS with only the clump properties, the experiment suffers the missing information about the NIR/MIR photometric properties. In fact, there is high contamination, especially for the Class 0. This limitation suggests that the information contained in the clump-only PS is insufficient to

⁵ <https://lightgbm.readthedocs.io/en/latest/pythonapi/lightgbm.LGBMClassifier.html>

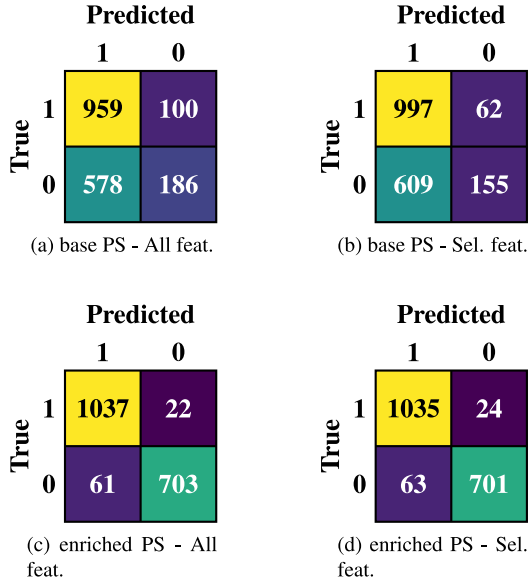


Fig. 5. Confusion matrix for the four experiment of classification adopting the LightGBM. Top panel: *base PS* with all the features (on the left) and the selected features (on the right). Bottom panel: *enriched PS* with all the features (on the left) and the selected features (on the right).

effectively distinguish between the classes. It is worth noting that, as well explained in Section 3, the clump class is a parameter derived through a process that involves the additional features present only in the *enriched PS*, not in the *base PS*. Analyzing the experimental results from this perspective, we conclude that the clumps intrinsic properties and their fluxes and colors do not have a good predictive power of their evolutionary stages derived from the information about the embedded YSOs. Note that, when the classification is performed using only the 8 selected features, the results do not get worse, as is evident from the upper panels in Fig. 5 and confirmed by the metrics reported in Table 4 (panels a and b). Thus, the *base PS* appears insufficient to give proper information about the clump class. Nevertheless, the FS step successfully identified the most informative features within the *base PS*, allowing the model to maintain its performance despite the reduced dimensionality.

When we consider the *enriched PS*, the results improve substantially compared to the previous experiment. This is evident in the confusion matrices shown in Fig. 5(c) (with all the features) and Fig. 5(d) (with the 43 selected features). In both cases, there is a very low contamination between the two classes, underscoring the superior predictive power of the full parameter set. Furthermore, the metrics reported in the panel c and d of Table 4 confirm that the classification performance remains consistent when transitioning from the *enriched PS* to the subset of selected features. These results demonstrate that the *enriched PS* is more effective in capturing the characteristics required to accurately classify clumps, probably due to the fact that the clump class depends on the information about the embedded YSOs. Moreover, the selected features still appear a good choice for modeling the problem, ensuring good results in predicting the right *clump class*. This highlights the robustness of the FS process, which ensures that the selected features retain the critical information necessary for modeling the problem effectively while reducing computational complexity.

6. Discussion

We performed ten different runs for each kind of experiment, both for classification and feature selection use cases, by randomly shuffling the training/test data selection, in order to keep under control the performance of models with respect to the potential variability of the

Table 4

Performance evaluation of the classification step across the four experiments, presenting: *precision* (the accuracy of positive predictions), *recall* (the model's ability to find all positive instances), *F1-score* (a balanced measure of precision and recall), *support* (the number of samples in each class), *accuracy* (the percentage of correctly classified samples), *macro avg* (the average of precision, recall, and F1-score across classes without considering class imbalance), and *weighted avg* (the average of precision, recall, and F1-score across classes, weighted by the support of each class).

	precision	recall	f1-score	support
0	0.62	0.91	0.74	1059
1	0.65	0.24	0.35	764
accuracy			0.63	1823
macro avg	0.64	0.57	0.55	1823
weighted avg	0.64	0.63	0.58	1823
(a) <i>base PS</i> — All feat.				
	precision	recall	f1-score	support
0	0.62	0.94	0.75	1059
1	0.71	0.20	0.32	764
accuracy			0.63	1823
macro avg	0.67	0.57	0.53	1823
weighted avg	0.66	0.63	0.57	1823
(b) <i>base PS</i> — Sel. feat.				
	precision	recall	f1-score	support
0	0.94	0.98	0.96	1059
1	0.97	0.92	0.94	764
accuracy			0.95	1823
macro avg	0.96	0.95	0.95	1823
weighted avg	0.96	0.95	0.95	1823
(c) <i>enriched PS</i> — All feat.				
	precision	recall	f1-score	support
0	0.94	0.98	0.96	1059
1	0.97	0.92	0.94	764
accuracy			0.95	1823
macro avg	0.95	0.95	0.95	1823
weighted avg	0.95	0.95	0.95	1823
(d) <i>enriched PS</i> — Sel. feat.				

parameter space coverage induced by the different choice of training and blind test samples, as well as to prevent possible selection criteria due to a particular model setup.

In the specific case of classification, this mechanism is always coupled with the model hyperparameter optimization, based on grid search, ensuring the optimal performance at each repeated experiment run. As the final choice of the best result, we always adopt the “worst case” strategy, by considering as final result the worst statistical performance in terms of confusion matrix, in order to remain as much as possible conservative regarding the final classification trade-off between the class purity and completeness.

In the case of feature selection, the strategy of repeating ten times the experiment was particularly useful for Boruta method, in order to keep under control the potential fluctuation of the measured relevance of features, induced by the random mechanism of shadow feature construction. As expected, the stability of the model, in principle declared by Kursa and Rudnicki (2010), was confirmed in our experiment, in which always the same features were labeled as true (rank 1, i.e. selected) and false (rank 4, i.e. rejected), without any presence of “tentative” features, which can be interpreted as a high reliability of the selected features with respect to the random noise injected into the parameter space and a high robustness with respect to the feature distributions within the parameter space.

The impact of the repeated experiments on the feature importance was also useful to evaluate their potential fluctuation in the case of standard Z score used by the RFE method. Also in this case no any significant variation was detected in terms of selected features, partially

due to the LightGBM, which is known to be less sensible to randomness than the classical Random Forest when a hyperparameter optimization, based on the grid search, is performed (which also helps to prevent the occurrence of overfitting, which boosting models may typically suffer with respect to bagging models). Finally, in the case of SHAP values, the intrinsic strategy of the method itself, based on the exploration of the full set of combinations among data patterns and related parameter space features, was insensitive to their potential variability.

6.1. Evolutionary parameters

The SEDs spanning the NIR to sub-mm range are widely recognized as valuable tools for assessing the evolutionary stage of a given object (Lada and Wilking, 1984; Allen et al., 2004). An important step for clump classification is the synergic exploitation of NIR/MIR and FIR/sub-mm data. This multi-wavelength approach is critical as it probes distinct physical components and processes inherent to star formation. While NIR/MIR observations allow us to penetrate some of the obscuring dust to identify embedded YSOs and trace warmer dust emission associated with protostars, FIR/sub-mm wavelengths are essential for characterizing the bulk properties of the cold, dense molecular gas and dust reservoirs from which stars form. In this perspective, the combination of these spectral windows provides a more complete understanding of clump evolutionary stages, mitigating biases inherent in single-wavelength studies and enabling a more robust assessment of their star-forming embedded stellar populations.

Typically, the clump SED peak shifts progressively from wavelengths longer than 100 μm in the earlier-stage objects (where the infrared excess comes from radiation from the disk and envelope for the youngest objects, and average values of dust temperatures do not go beyond ~ 15 K, see for example Elia et al. 2017) to shorter wavelengths as the object evolves. Among our selected features, the Herschel fluxes at 160 μm and 500 μm stand out. As reported in Elia et al. (2017), in the absence of a complete grey-body fit, the combination of these two fluxes are an excellent diagnostic for estimating the mean dust temperatures of clumps.

As regards the colors, among the selected features we find the Herschel [250–350], and the [w4–70]. The latter has been widely used in literature in combination with other Herschel colors for diagnostic. For example, Paladini et al. (2012) highlighted that it is useful for the segregation of the HII regions with respect to the YSOs population with the color [160–250], and also (Molinari et al., 2008) emphasized the role of [24–70] to disentangle different kind of objects.

As far as physical properties are concerned, in the last years different indicators has been useful for studying the evolutionary stages of clumps. Among the most known evolutionary indicators in the literature, there are the bolometric luminosity (L_{bol}) and its ratio to the mass (M), namely the L/M . The L/M is a distance-independent quantity which is expected to rapidly increase during the accretion phase of star formation (Molinari et al., 2008; Elia et al., 2013). In our experiment, these two properties are not among the most important features. However, both L_{ratio} and the surface density appear as important features. The L_{ratio} is related to the bolometric luminosity, this being defined as the ratio between the bolometric luminosity and the luminosity at wavelengths longward of 350 μm . It is a distance-independent evolutionary indicator usually adopted, which is expected to be larger at more evolved star formation stages (Andre et al., 1993). For example, it has been investigated in studies about the low-mass star formation (Andre et al., 1999; Maury et al., 2011). In general, a significant emission observable at sub-mm wavelengths indicates an object in a very early stage of star formation (Elia et al., 2017). Beside the L_{ratio} , the surface density (Σ) is the parameter related to the mass of the clump. This is in fact obtained by dividing the mass of the clump by the area of a circle whose diameter is the linear diameter of the clump (Elia et al., 2021), thus it derives from both photometric measurements and the source linear sizes. It has been shown in Elia

et al. (2017) that more evolved sources are on average slightly smaller and denser than in the early stages of star formation, with a drop at increasing ages (Elia et al., 2017; Giannetti et al., 2013).

Another diagnostic is the bolometric temperature, T_{bol} , used in different works, such as Strafella et al. (2015), for understanding the evolutionary stages of sources in Vela-D using Hi-GAL data. The T_{bol} is often used with the L_{bol} for building the corresponding L_{bol} versus T_{bol} diagram as a diagnostic tool for the characterization of YSOs from an evolutionary point of view (for example, Strafella et al., 2015).

Lastly, the YSO density, δ_{YSO} , is a feature selected in both the experiments. Coming from the YSO-Class step, it is the percentage of YSOs within clumps, computed as the ratio of the number of YSOs to the total number of sources detected within the clump's spatial boundaries (encompassing all source types and including potential contaminants). Interestingly, our analysis reveals that δ_{YSO} shows no significant correlation with commonly used physical or photometric parameters. This suggests that data-driven approaches can uncover observables that, while not grounded in physical modeling, encapsulate latent evolutionary information, offering a different perspective on clump evolution.

As previously discussed, the feature selection process primarily favored photometric features, likely because they trace dust emission and spectral signatures typical of embedded YSOs within clumps. This outcome is likely influenced by the way the *clump class* itself is defined. As explained in Sections 3.1–3.2, the class assigned to each clump is derived from the classification of the YSOs located within their elliptical area. The classification of individual YSOs is based on their fluxes and corresponding colors in the NIR/MIR range. Consequently, the *clump class* is the average over the YSO classes in the clump and intrinsically tied to their NIR/MIR emission. From this analysis, it appears that the NIR/MIR properties, by being inherently linked to the definition of the target variable, may overshadow the contribution of the physical properties in the feature selection process. This does not necessarily imply that key physical properties, such as L_{bol} and the L/M ratio, are irrelevant for tracing evolution. Rather, their exclusion may stem from other factors, such as strong correlations with selected photometric features, larger uncertainties in their estimates, or a reduced completeness across the sample. Additionally, their dynamic range or distribution may not be well captured by the selection algorithm, especially if it assumes linear relationships or normalized inputs. On the contrary, in the absence of features directly related to the definition of the *clump class*, the algorithm relies entirely on structural and physical descriptors to discriminate among classes. While this shift in feature relevance is partly driven by the available input parameter space, it is particularly encouraging from a physical point of view: the selected features may describe the physical nature of the clumps, rather than simply reflecting how the clump class has been defined. In this sense, identifying this subset of relevant physical features could provide valuable hints to better understand the evolutionary processes at play in the formation and development of clumps. Future work will aim to disentangle these effects and explicitly investigate how the selected photometric features are related to the underlying physical parameters, in order to build a more comprehensive picture of the evolutionary sequence.

6.2. Statistical test analysis

In Fig. 6, a grid of plots of the selected features is shown for providing a comprehensive overview of the data and enabling the identification of possible trends.

First, it can be observed that while some pairs of features exhibit a relationship, albeit weak, others show no apparent correlation, with distributions suggesting independence between the variables. Second, the behavior of some features as a function of the *clump class* shows significant overlap for certain features, whereas for others, the distributions appear to be quite distinguishable. For example, the distribution of [w4–70] is steeper for Class 0, assuming lower values than Class 1.

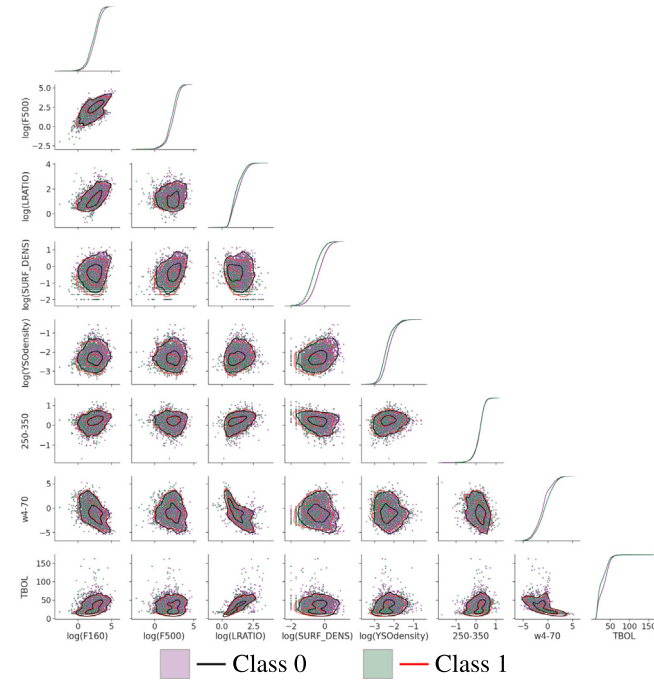


Fig. 6. Grid of plots of the selected features for the experiment with the *base PS*. In this grid, each plot corresponds to a pair of variables. Class 0 are plotted in violet, Class 1 in green. The contours represent the 20, 50 and 80 per cent levels of the density distribution of Class 0 (black) and Class 1 (red), respectively. Along the diagonal, the empirical cumulative distribution functions (ECDFs) of the given features are shown.

In particular, we observe that high values of this color correspond to low values of L_{ratio} . This is expected, since in the early stages of star formation the sub-mm luminosity is higher, and it is difficult to have a strong emission in the MIR. However, the distinction between Class 0 and Class 1 is not as clear-cut as, for example, in the $\log(\Sigma)$ vs $[w4-70]$ plot. Generally, it can be observed that Class 0 exhibits higher values of Σ compared to Class 1, as Σ tends to be higher in the early stages of the star formation process and decreases as the process goes on (e.g. [Elia et al., 2021](#)). Similarly, the $\delta_{Y_{SO}}$ exhibits the same behavior. It is important to note that these two parameters have been obtained independently. In fact, while the first is directly related to the mass of the clump ([Elia et al., 2017](#)), the second is a statistical parameter derived from YSO-CLASS and defined as the ratio between the number of YSOs in the clump and the total number of sources within the clump boundaries (M25).

To determine whether the distributions of the selected features differ significantly between the two classes defined by the *clump class*, we employed three statistical tests: the Kolmogorov–Smirnov (KS) test ([Massey, 1951](#)), the Mann–Whitney U (MWU) test ([Gibson, 1992](#)), and Linear Discriminant Analysis (LDA) ([McLachlan, 2005](#)). These tests provide complementary insights into the separability of the distributions. The KS test is a non-parametric test that evaluates whether two samples originate from the same continuous distribution by comparing their empirical cumulative distribution functions (ECDFs, represented also in [Fig. 6](#)), making it particularly effective for detecting both location (along the x-axis) and shape differences between distributions. The MWU test, on the other hand, is a non-parametric test that assesses whether one distribution tends to have systematically higher or lower values compared to the other, focusing primarily on differences in central tendency without assuming a specific distributional shape. Finally, LDA is not a statistical test but a supervised classification method, used to separate two or more classes in a space defined by multiple variables. Unlike the KS and MWU tests, LDA is specifically designed to address problems of separation and classification in multivariate contexts. In particular, it provides a quantitative

Table 5

Statistical test results for feature separation adopting the *clump class*. The first column reports the selected features for the experiment with the *base PS*. The second and third columns show the distance between the two distributions and the probability of obtaining that distance by chance according to the KS test, while the fourth column those related to the MWU test.

Feature	KS Statistic	KS p -value	MWU p -value
F_{160}	0.1019	8.27×10^{-14}	$< 10^{-8}$
F_{500}	0.1217	1.64×10^{-19}	$< 10^{-8}$
L_{ratio}	0.1208	3.20×10^{-19}	$< 10^{-8}$
Σ	0.2477	5.49×10^{-80}	$< 10^{-8}$
$\delta_{Y_{SO}}$	0.1724	8.77×10^{-39}	$< 10^{-8}$
[250–350]	0.0356	4.63×10^{-2}	5.68×10^{-2}
[w4–70]	0.1460	6.83×10^{-28}	$< 10^{-8}$
T_{bol}	0.1698	1.22×10^{-37}	$< 10^{-8}$

measure of separability by constructing linear combinations of the features to maximize class distinction, offering a direct assessment of the discriminatory power of the variables. Together, these tests form a robust framework for evaluating the degree of separation between the two groups, each contributing unique strengths to the analysis: the KS test for overall distributional differences, the Mann–Whitney U test for central tendency shifts, and LDA for assessing classification performance. Results are summarized in [Table 5](#).

The statistical analysis performed on the selected variables highlights significant differences in the distributions when comparing the two classes identified by *clump class*. The KS test reveals notable separability for nearly all variables, and this is evidenced by very low p -values, which indicate that the null hypothesis of identical distributions can be confidently rejected. Indeed among these variables, Σ stands out with the highest KS statistic of 0.2477 and a p -value on the order of 10^{-80} , suggesting a particularly strong separation between the two classes. Similarly, $\delta_{Y_{SO}}$, T_{bol} , and [w4–70] demonstrate substantial separability, with KS statistics ranging from 0.1460 to 0.1724 and corresponding p -values that remain far below any conventional significance threshold. The variable [250–350] presents a weaker signal, as indicated by a lower KS statistic of 0.0356 and a p -value of 0.046, which, while still statistically significant, suggests that this feature contributes less to class differentiation compared to the others.

The MWU test further corroborates these findings, as it consistently returns p -values close to zero for the majority of the variables, reinforcing the evidence that the distributions across the *clump class* are significantly different. Only the variable [250–350] approaches the threshold of significance with a p -value of approximately 0.057, aligning with the weaker signal observed in the KS test. These results collectively suggest that Σ , $\delta_{Y_{SO}}$, and T_{bol} are among the most discriminative features in separating the two classes, followed closely by F_{160} , F_{500} , and [w4–70].

The effectiveness of these features in capturing class separability is further supported by the LDA, which achieves an accuracy score of 0.6532. While this result does not indicate perfect separation, it demonstrates that the selected features contain meaningful discriminatory power when distinguishing between the two classes. The fact that LDA achieves this level of accuracy using linear combinations of the features suggests that the observed separability is not an artifact of the tests but reflects genuine differences between the two groups.

7. Conclusions

In this work we studied the evolutionary path of star forming clumps in Hi-GAL through the adoption of a multi-step approach. First, we conducted feature engineering on data derived from several catalogs, from NIR to sub-mm, in order to build a broad parameter space able to trace the entire process of star formation, from the very early stage of the birth of YSOs to the classical pre-Main Sequence phase. Once the dataset was complete and processed to address any missing data

and ensure normalization, we identified two distinct parameter spaces: the *base PS*, made up of 16 features, and the *enriched PS*, composed by 135 features. Following this, we applied a feature selection step, leveraging a combination of different algorithms (RFE + standard tree-based importance, RFE + SHAP importance, Boruta + standard tree-based importance, Boruta + SHAP importance) for improving the performance of our workflow, using the *clump class* provided by M25 as target variable. For each experiment, we evaluated the results and determined the optimal combination for our parameter spaces. This process allowed us to define a reduced parameter space for each, consisting solely of the selected features: for the *base PS*, 8 features have been selected as relevant, 43 for the *enriched PS*. By adopting feature selection methods, we refined our models to focus solely on the most meaningful and relevant features. This strategy was essential for faithfully capturing the processes underlying star formation, ensuring a robust and insightful representation of this complex phenomenon. Finally, these four parameter spaces (the enriched, the base and the corresponding with only the selected features) have been used in the classification step of our workflow. This step validated the outcomes of the feature selection process. Main results are summarized below:

- (i) The FS experiment with the *enriched PS* revealed that the most significant features predominantly belong to the photometric group. It is worth noting that our method selects the most relevant features, which may result in similar features being included in the selection.
- (ii) The FS experiment with the *base PS* suffers the scarcity of information related to the photometric features. This result was quite expected, due to the fact that the *clump class* is related to the presence of YSOs within them, which in turn have been classified using their magnitudes at different wavelengths in the NIR/MIR.
- (iii) Among the selected features, there are well-known indicators of the evolutionary stages of clumps in literature: the Herschel fluxes F_{160} and F_{500} , the colors [250–350] and [w4–70], the ratio L_{ratio} of the bolometric luminosity to the luminosity at wavelengths longward of 350 μm , bolometric temperature T_{bol} , the surface density Σ , and the δ_{YSO} .
- (iv) The classification experiment reveals interesting insights into the role of the selected features. While the *base PS* lacked the necessary information to clearly distinguish between the classes (leading to notable contamination, particularly in Class 0), the feature selection process proved remarkably effective. By isolating the 8 most statistically significant variables, the model was able to retain its predictive performance even after dimensionality reduction. This outcome highlights that while the *base PS* alone may lack the full range of information needed to clearly distinguish between the classes, the FS step effectively preserves key discriminative features (Table 4, panels a and b).
- (v) The combination of KS tests, MWU tests and LDA accuracy highlights a robust pattern of separation across the *clump class* groups. Variables such as Σ , δ_{YSO} , and T_{bol} emerge as the strongest candidates for characterizing the underlying differences between the two classes, while [250–350] shows more limited discriminatory capacity.

In summary, what emerges from our work is that some features lack a clear and consistent behavior in distinguishing the different evolutionary phases of the clumps. In particular, the NIR/MIR features do not appear to be strongly reflected in the evolutionary indicators of the clumps, which are primarily based on FIR features derived from their global emission. This result likely indicates that the global FIR/sub-mm emission of clumps and the radiation from the NIR/MIR resolved YSOs within them are not necessarily coeval but rather correspond to different stages of star formation (M25). These findings highlight the complex nature of star formation processes, where the physical conditions probed by different observational indicators can evolve asynchronously within the same clump environment.

This complex behavior obviously makes the problem of object classification complex and challenging. Traditional approaches to the classification of YSOs, in particular, have typically relied only on limited information, often related to color–color and color–magnitude diagrams. These include combination of *Spitzer* and MIPS data complemented by *J*, *H* and *K* from 2MASS (Gutermuth et al., 2008, 2009), as well as *WISE* -based diagnostics combined with 2MASS (Koenig et al., 2012; Koenig and Leisawitz, 2014) or AKARI (Pollo et al., 2010), and so on. Although such approaches introduced a more formal statistical treatment, they remained constrained to low-dimensional parameter spaces and a limited set of features. A step toward more automated classification was introduced by Tóth et al. (2014), who applied Quadratic Discriminant Analysis (QDA, Hastie et al., 2009) to a combination of *WISE* and AKARI photometry, successfully recovering a high fraction of known YSO candidates with limited contamination. Similarly, Marton et al. (2016) applied Support Vector Machine (SVM, Cristianini and Shawe-Taylor, 2000) trained on a dataset obtained combining Reliable 2MASS and *WISE* photometric data with *Planck* dust opacity values, successfully identifying more known YSOs of the training sample than other methods based on color–color and magnitude–color selection. In contrast, our study leverages a richer and more diverse dataset, combining all available photometric information for identifying YSOs in the Galactic plane region $10^\circ \leq \ell \leq 60^\circ$, $|b| \leq 1^\circ$, including the *Herschel* bands and the physical properties associated to the Hi-GAL clumps containing the YSOs. Our machine learning approach enables the integration of multiple photometric and physical features, capturing complex, possibly nonlinear relationships between observables and evolutionary classes. This facilitates a more objective, reproducible, and scalable classification process, which not only supports current models of clump evolution but also reveals the relative importance of diagnostic features that may be overlooked by traditional methods.

Although our analysis is based on Hi-GAL clumps, which are primarily associated with intermediate-to-high-mass star forming regions in the Galactic plane, the proposed methodology is general and can be applied to other star forming environments. The feature selection and classification framework does not rely on specific assumptions tied to Hi-GAL data and can, in principle, be adapted to different surveys. The inclusion of additional datasets in an experiment of this kind could, on the one hand, enrich the physical interpretation of the clump classification scheme, particularly when the new tracers differ significantly from the infrared continuum examined so far (for instance, data from line surveys or radio continuum observations). On the other hand, such an expansion toward other surveys would effectively constitute a new experiment, which might reveal that the newly introduced parameters are entirely independent of those identified in the previous analysis, or alternatively, that they add only limited relevance or novelty.

The advantage of performing a FS, besides its intrinsic capability to infer knowledge on the physics of the problem, is evident by comparing the statistical performance of the classification between the full and reduced parameter spaces. In fact, the comparable results in terms of purity and completeness demonstrates the presence of redundant information in the full case, useless from the classification point of view and highly impacting the experiment computation complexity.

Finally, the interpretability of the resulting selected PS is intrinsically ensured by the use of SHAP method, developed in the framework of the Explainable Artificial Intelligence (XAI) methodology and proven to enhance the interpretability of its results in the context of data-driven approaches. In particular, SHAP provides a unified and theoretically grounded approach to interpret model predictions by computing the marginal contribution of each feature based on cooperative game theory. This allows for a detailed attribution of the model's output to individual input parameters, ensuring transparency and enabling a rigorous assessment of the physical relevance of the selected features. In this context, SHAP serves not only as a post-hoc explanation tool but also as a guide for domain-driven interpretation, strengthening the methodological robustness and scientific interpretability of the proposed approach.

CRedit authorship contribution statement

Y. Maruccia: Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Data curation. **S. Cuvuoti:** Writing – review & editing, Visualization, Validation, Methodology, Formal analysis. **M. Brescia:** Writing – review & editing, Visualization, Validation, Methodology, Formal analysis. **G. Riccio:** Writing – review & editing, Visualization, Validation, Methodology, Formal analysis. **S. Molinari:** Validation, Supervision, Methodology. **D. Elia:** Writing – review & editing, Validation, Methodology. **E. Schisano:** Writing – review & editing, Validation, Methodology.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors thank the anonymous referee for the valuable comments and suggestions that have improved the quality of this manuscript. The authors acknowledge funding by the European Research Council via the ERC Synergy Grant “ECOGAL – Understanding our Galactic ecosystem: From the disk of the Milky Way to the formation sites of stars and planets” (project ID 855130). YM, SC, GR acknowledge the Spoke 3 “Astrophysics and Cosmos Observations” of the ICSC–Centro Nazionale di Ricerca in High Performance Computing, Big Data and Quantum Computing and hosting entity, funded by European Union–Next GenerationEU. YM expresses gratitude to Professor Francesco Strafella for his helpful comments and feedback, which greatly strengthened the overall manuscript. SC and GR acknowledge support from PRIN MUR 2022, Italy (20224MNC5A), “Life, death and after-death of massive stars”, funded by European Union – Next Generation EU. MB, SC and GR acknowledge the ASI-INAF TI agreement, Italy, 2018-23-HH.0 “Attività scientifica per la missione Euclid - fase D”.

Data availability

The data that has been used is confidential.

References

- Allen, L.E., Calvet, N., D'Alessio, P., Merin, B., Hartmann, L., Megeath, S.T., Gutermuth, R.A., Muzerolle, J., Pipher, J.L., Myers, P.C., et al., 2004. Infrared array camera (IRAC) colors of young stellar objects. *Astrophys. J. Suppl. Ser.* 154, 363.
- Andre, P., Ward-Thompson, D., Barsony, M., 1993. Submillimeter continuum observations of Rho Ophiuchi A-the candidate protostar VLA 1623 and prestellar clumps. *Astrophys. J.* (ISSN: 0004-637X) 406, 122–141.
- Andre, P., Ward-Thompson, D., Barsony, M., 1999. From pre-stellar cores to protostars: The initial conditions of star formation. *arXiv Preprint Astro-Ph/9903284*.
- Ballesteros-Paredes, J., André, P., Hennebelle, P., Klessen, R.S., Kruijssen, J.M.D., Chevance, M., Nakamura, F., Adamo, A., Vazquez-Semadeni, E., 2020. From diffuse gas to dense molecular cloud cores. *Space Sci. Rev.* 216, 1–58.
- Benjamin, R.A., Churchwell, E., Babler, B.L., Bania, T.M., Clemens, D.P., Cohen, M., Dickey, J.M., Indebetouw, R., Jackson, J.M., Kobulnicky, H.A., Lazarian, A., Marston, A.P., Mathis, J.S., Meade, M.R., Seager, S., Stolovy, S.R., Watson, C., Whitney, B.A., Wolff, M.J., Wolfire, M.G., 2003. GLIMPSE. I. An SIRTf legacy project to map the inner galaxy. *Publ. Astron. Soc. Pac.* 115, 953–964. doi:10.1086/376696, arXiv:astro-ph/0306274.
- Bergin, E.A., Tafalla, M., 2007. Cold dark clouds: The initial conditions for star formation. *Annu. Rev. Astron. Astrophys.* 45, 339–396.
- Beuther, H., Schilke, P., Sridharan, T.K., Menten, K.M., Walmsley, C.M., Wyrowski, F., 2002. Massive molecular outflows. *Astron. Astrophys.* 383, 892–904.
- Breiman, L., 2001. Random forests. *Mach. Learn.* 45, 5–32.
- Breiman, L., 2017. Classification and Regression Trees. Routledge.
- Chevance, M., Krumholz, M.R., McLeod, A.F., Ostriker, E.C., Rosolowsky, E.W., Sternberg, A., 2022. The life and times of giant molecular clouds. *arXiv preprint arXiv:2203.09570*.
- Churchwell, E., Babler, B.L., Meade, M.R., Whitney, B.A., Benjamin, R., Indebetouw, R., Cyganowski, C., Robitaille, T.P., Povich, M., Watson, C., Bracker, S., 2009. The spitzer/GLIMPSE surveys: A new view of the Milky Way. *Publ. Astron. Soc. Pac.* 121, 213. doi:10.1086/597811.
- Cristianini, N., Shawe-Taylor, J., 2000. An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge University Press.
- Cutri, R.M., Wright, E.L., Conrow, T., Fowler, J.W., Eisenhardt, P.R.M., Grillmair, C., Kirkpatrick, J.D., Masci, F., McCallon, H.L., Wheelock, S.L., et al., 2013. Explanatory supplement to the AllWISE data release products. *Explan. Suppl. AllWISE Data Release Prod.* 1.
- Elia, D., Merello, M., Molinari, S., Schisano, E., Zavagno, A., Russeil, D., Mège, P., Martin, P.G., Olmi, L., Pestalozzi, M., et al., 2021. The hi-GAL compact source catalogue–II. The 360° catalogue of clump physical properties. *Mon. Not. R. Astron. Soc.* 504, 2742–2766.
- Elia, D., Molinari, S., Fukui, Y., Schisano, E., Olmi, L., Veneziani, M., Hayakawa, T., Pestalozzi, M., Schneider, N., Benedettini, M., di Giorgio, A.M., Ikhenade, D., Mizuno, A., Onishi, T., Pezzuto, S., Piazzo, L., Polychroni, D., Rygl, K.L.J., Yamamoto, H., Maruccia, Y., 2013. The first Hi-GAL observations of the outer galaxy: A look at star formation in the third galactic quadrant in the longitude range $216.5^\circ < l < 225.5^\circ$. *Astrophys. J.* 772, 45. doi:10.1088/0004-637X/772/1/45, arXiv:1304.7358.
- Elia, D., Molinari, S., Schisano, E., Pestalozzi, M., Pezzuto, S., Merello, M., Noriega-Crespo, A., Moore, T.J.T., Russeil, D., Mottram, J.C., Paladini, R., Strafella, F., Benedettini, M., Bernard, J.P., Di Giorgio, A., Eden, D.J., Fukui, Y., Plume, R., Bally, J., Martin, P.G., Ragan, S.E., Jaffa, S.E., Motte, F., Olmi, L., Schneider, N., Testi, L., Wyrowski, F., Zavagno, A., Calzoletti, L., Faustini, F., Natoli, P., Palmeirim, P., Piacentini, F., Piazzo, L., Pilbratt, G.L., Polychroni, D., Baldeschi, A., Beltrán, M.T., Billot, N., Cambresy, L., Cesaroni, R., García-Lario, P., Hoare, M.G., Huang, M., Jonas, G., Liu, S.J., Maiolo, B.M.T., Marsh, K.A., Maruccia, Y., Mège, P., Peretto, N., Rygl, K.L.J., Schilke, P., Thompson, M.A., Traficante, A., Umam, G., Veneziani, M., Ward-Thompson, D., Whitworth, A.P., Arab, H., Bandieramonte, M., Becciani, U., Brescia, M., Buemi, C., Bufano, F., Butora, R., Cuvuoti, S., Costa, A., Fiorellino, E., Hajnal, A., Hayakawa, T., Kacsuk, P., Leto, P., Li Causi, G., Marchili, N., Martinavarro-Armengol, S., Mercurio, A., Molinaro, M., Riccio, G., Sano, H., Sciacca, E., Tachihara, K., Torii, K., Trigilio, C., Vitello, F., Yamamoto, H., 2017. The Hi-GAL compact source catalogue - I. The physical properties of the clumps in the inner galaxy ($-71.0^\circ < \ell < 67.0^\circ$). *Mon. Not. R. Astron. Soc.* 471, 100–143. doi:10.1093/mnras/stx1357, arXiv:1706.01046.
- Evans, N.J., Dunham, M.M., Jørgensen, J.K., Enoch, M.L., Merín, B., Van Dishoeck, E.F., Alcalá, J.M., Myers, P.C., Stapelfeldt, K.R., Huard, T.L., et al., 2009. The spitzer c2d legacy results: Star-formation rates and efficiencies; evolution and lifetimes. *Astrophys. J. Suppl. Ser.* 181, 321.
- Fazio, G.G., Hora, J.L., Allen, L.E., Ashby, M.L.N., Barmby, P., Deutsch, L.K., Huang, J.-S., Kleiner, S., Marengo, M., Megeath, S.T., Melnick, G.J., Pahre, M.A., Patten, B.M., Polizotti, J., Smith, H.A., Taylor, R.S., Wang, Z., Willner, S.P., Hoffmann, W.F., Pipher, J.L., Forrest, W.J., McMurt, C.W., McCreight, C.R., McKelvey, M.E., McMurray, R.E., Koch, D.G., Moseley, S.H., Arendt, R.G., Mentzell, J.E., Marx, C.T., Losch, P., Mayman, P., Eichhorn, W., Krebs, D., Jhabvala, M., Gezari, D.Y., Fixsen, D.J., Flores, J., Shakoortzadeh, K., Jungo, R., Hakun, C., Workman, L., Karpati, G., Kichak, R., Whitley, R., Mann, S., Tollestrup, E.V., Eisenhardt, P., Stern, D., Gorjian, V., Bhattacharya, B., Carey, S., Nelson, B.O., Glaccum, W.J., Lacy, M., Lowrance, P.J., Laine, S., Reach, W.T., Stauffer, J.A., Surace, J.A., Wilson, G., Wright, E.L., Hoffman, A., Domingo, G., Cohen, M., 2004. The infrared array camera (IRAC) for the spitzer space telescope. *Astrophys. J. Suppl.* 154, 10–17. doi:10.1086/422843, arXiv:astro-ph/0405616.
- Giannetti, A., Brand, J., Sánchez-Monge, Á., Fontani, F., Cesaroni, R., Beltrán, M.T., Molinari, S., Dodson, R., Rioja, M.J., 2013. Physical properties of high-mass clumps in different stages of evolution. *Astron. Astrophys.* 556, A16.
- Gibson, J.D., 1992. Nonparametric statistics: An introduction. Quantitative applications in social sciences.
- Grandvalet, Y., Canu, S., Boucheron, S., 1997. Noise injection: Theoretical prospects. *Neural Comput.* 9, 1093–1108.
- Greene, T.P., Wilking, B.A., Andre, P., Young, E.T., Lada, C.J., 1994. Further mid-infrared study of the rho ophiuchi cloud young stellar population: Luminosities and masses of pre-main-sequence stars. *Astrophys. J.* 434, 614. doi:10.1086/174763.
- Gutermuth, R.A., Megeath, S.T., Muzerolle, J., Allen, L.E., Pipher, J.L., Myers, P.C., Fazio, G.G., 2004. The NGC 7129 young stellar cluster: A combined spitzer, MMT, and two micron all sky survey census of disks, protostars, and outflows. *Astrophys. J. Suppl. Ser.* 154, 374.
- Gutermuth, R.A., Megeath, S.T., Myers, P.C., Allen, L.E., Pipher, J.L., Fazio, G.G., 2009. A spitzer survey of young stellar clusters within one kiloparsec of the sun: Cluster core extraction and basic structural analysis. *Astrophys. J. Suppl.* 184, 18–83. doi:10.1088/0067-0049/184/1/18, arXiv:0906.0201.
- Gutermuth, R.A., Myers, P.C., Megeath, S.T., Allen, L.E., Pipher, J.L., Muzerolle, J., Porras, A., Winston, E., Fazio, G., 2008. Spitzer observations of NGC 1333: A study of structure and evolution in a nearby embedded cluster. *Astrophys. J.* 674, 336–356. doi:10.1086/524722, arXiv:0710.1860.
- Hambly, N.C., Collins, R.S., Cross, N.J.G., Mann, R.G., Read, M.A., Sutorius, E.T.W., Bond, I., Bryant, J., Emerson, J.P., Lawrence, A., Rimoldini, L., Stewart, J.M., Williams, P.M., Adamson, A., Hirst, P., Dye, S., Warren, S.J., 2008. The WFCAM

- science archive. Mon. Not. R. Astron. Soc. 384, 637–662. doi:10.1111/j.1365-2966.2007.12700.x, arXiv:0711.3593.
- Hastie, T., Tibshirani, R., Friedman, J.H., Friedman, J.H., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, vol. 2. Springer.
- Hua, J., Lowey, J., Xiong, Z., Dougherty, E.R., 2006. Noise-injected neural networks show promise for use on small-sample expression data. BMC Bioinformatics 7, 1–14.
- Huang, L., Qin, J., Zhou, Y., Zhu, F., Liu, L., Shao, L., 2023. Normalization techniques in training dnns: Methodology, analysis and application. IEEE Trans. Pattern Anal. Mach. Intell. 45, 10173–10196.
- Juvela, M., Montillaud, J., 2016. Allsky NICER and NICEST extinction maps based on the 2MASS near-infrared survey. Astron. Astrophys. 585, A38.
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y., 2017. LightGBM: A highly efficient gradient boosting decision tree. Adv. Neural Inf. Process. Syst. 30.
- Koenig, X.P., Leisawitz, D.T., 2014. A classification scheme for Young stellar objects using the wide-field infrared survey explorer allwise catalog: Revealing low-density star formation in the outer galaxy. Astrophys. J. 791, 131. doi:10.1088/0004-637X/791/2/131, arXiv:1407.2262.
- Koenig, X.P., Leisawitz, D.T., Benford, D.J., Rebull, L.M., Padgett, D.L., Assef, R.J., 2012. Wide-field infrared survey explorer observations of the evolution of massive star-forming regions. Astrophys. J. 744, 130. doi:10.1088/0004-637X/744/2/130.
- Kursa, M.B., Rudnicki, W.R., 2010. Feature selection with the boruta package. J. Stat. Softw. 36, 1–13. doi:10.18637/jss.v036.i11, URL: <https://www.jstatsoft.org/index.php/jss/article/view/v036i11>.
- Kursa, M.B., Rudnicki, W.R., 2011. The all relevant feature selection using random forest. arXiv preprint arXiv:1106.5112.
- Lada, C.J., Wilking, B.A., 1984. The nature of the embedded population in the rho ophiuchi dark cloud-mid-infrared observations. Astrophys. J. (ISSN: 0004-637X) 287, 610–621.
- Lucas, P.W., Hoare, M.G., Longmore, A., Schröder, A.C., Davis, C.J., Adamson, A., Bandyopadhyay, R.M., de Grijs, R., Smith, M., Gosling, A., Mitchison, S., Gáspár, A., Coe, M., Tamura, M., Parker, Q., Irwin, M., Hambly, N., Bryant, J., Collins, R.S., Cross, N., Evans, D.W., Gonzalez-Solares, E., Hodgkin, S., Lewis, J., Read, M., Riehl, M., Sutorius, E.T.W., Lawrence, A., Drew, J.E., Dye, S., Thompson, M.A., 2008. The UKIDSS galactic plane survey. Mon. Not. R. Astron. Soc. 391, 136–163. doi:10.1111/j.1365-2966.2008.13924.x, arXiv:0712.0100.
- Lundberg, S., 2017. A unified approach to interpreting model predictions. arXiv preprint arXiv:1705.07874.
- Marton, G., Tóth, L.V., Paladini, R., Kun, M., Zahorec, S., McGehee, P., Kiss, C., 2016. An all-sky support vector machine selection of WISE YSO candidates. Mon. Not. R. Astron. Soc. 458, 3479–3488.
- Maruccia, Y., Elia, D., Riccio, G., Cavuoti, S., Schisano, E., Molinari, S., Brescia, M., Merello, M., Pezzuto, S., Traficante, A., Koll, V.F., Hennebelle, P., Klessen, R.S., Testi, L., 2025. Characterisation of the evolutionary stage of the Hi-GAL clumps in the galactic plane region $10^\circ \leq \ell \leq 60^\circ$. (submitted for publication).
- Massey, Jr., F.J., 1951. The Kolmogorov-Smirnov test for goodness of fit. J. Amer. Statist. Assoc. 46, 68–78.
- Maury, A.J., André, P., Men'shchikov, A., Könyves, V., Bontemps, S., 2011. The formation of active protoclusters in the aquila rift: A millimeter continuum view. Astron. Astrophys. 535, A77.
- McKee, C.F., Tan, J.C., 2003. The formation of massive stars from turbulent cores. Astrophys. J. 585, 850.
- McLachlan, G.J., 2005. Discriminant Analysis and Statistical Pattern Recognition. John Wiley & Sons.
- Megeath, S.T., Gutermuth, R., Muzerolle, J., Kryukova, E., Hora, J.L., Allen, L.E., Flaherty, K., Hartmann, L., Myers, P.C., Pipher, J.L., et al., 2015. The spitzer space telescope survey of the orion A and B molecular clouds. II. The spatial distribution and demographics of dusty young stellar objects. Astron. J. 151, 5.
- Molinari, S., Pezzuto, S., Cesaroni, R., Brand, J., Faustini, F., Testi, L., 2008. The evolution of the spectral energy distribution in massive young stellar objects. Astron. Astrophys. 481, 345–365.
- Molinari, S., Schisano, E., Elia, D., Pestalozzi, M., Traficante, A., Pezzuto, S., Swinyard, B.M., Noriega-Crespo, A., Bally, J., Moore, T.J.T., Plume, R., Zavagno, A., di Giorgio, A.M., Liu, S.J., Pilbratt, G.L., Mottram, J.C., Russell, D., Piazzi, L., Veneziani, M., Benedettini, M., Calzoletti, L., Faustini, F., Natoli, P., Piacentini, F., Merello, M., Palmese, A., Del Grande, R., Polychroni, D., Rygl, K.L.J., Polenta, G., Barlow, M.J., Bernard, J.P., Martin, P.G., Testi, L., Ali, B., André, P., Beltrán, M.T., Billot, N., Carey, S., Cesaroni, R., Compiègne, M., Eden, D., Fukui, Y., García-Lario, P., Hoare, M.G., Huang, M., Jomcas, G., Lim, T.L., Lord, S.D., Martínvarro-Armengol, S., Motte, F., Paladini, R., Paradis, D., Peretto, N., Robitaille, T., Schilke, P., Schneider, N., Schulz, B., Sibthorpe, B., Strafella, F., Thompson, M.A., Umana, G., Ward-Thompson, D., Wyrowski, F., 2016. Hi-GAL, the herchel infrared galactic plane survey: Photometric maps and compact source catalogues. First data release for the inner Milky Way: $+68^\circ \geq l \geq -70^\circ$. Astron. Astrophys. 591, A149. doi:10.1051/0004-6361/201526380, arXiv:1604.05911.
- Molinari, S., Schisano, E., Faustini, F., Pestalozzi, M., di Giorgio, A.M., Liu, S., 2011. Source extraction and photometry for the far-infrared and sub-millimeter continuum in the presence of complex backgrounds. Astron. Astrophys. 530, A133. doi:10.1051/0004-6361/201014752, arXiv:1011.3946.
- Molinari, S., Swinyard, B., Bally, J., Barlow, M., Bernard, J.-P., Martin, P., Moore, T., Noriega-Crespo, A., Plume, R., Testi, L., et al., 2010a. Clouds, filaments, and protostars: The herchel Hi-GAL Milky Way. Astron. Astrophys. 518, L100.
- Molinari, S., Swinyard, B., Bally, J., Barlow, M., Bernard, J.-P., Martin, P., Moore, T., Noriega-Crespo, A., Plume, R., Testi, L., et al., 2010b. Hi-GAL: The herchel infrared galactic plane survey. Publ. Astron. Soc. Pac. 122, 314.
- Motte, F., Bontemps, S., Csengeri, T., Pouteau, Y., Louvet, F., Stutz, A.M., Cunningham, N., López-Sepulcre, A., Brouillet, N., Galván-Madrid, R., et al., 2022. ALMA-IMF-I. Investigating the origin of stellar masses: Introduction to the large program and first results. Astron. Astrophys. 662, A8.
- Paladini, R., Umana, G., Veneziani, M., Noriega-Crespo, A., Anderson, L.D., Piacentini, F., Gonçalves, D.P., Paradis, D., Tibbs, C.T., Bernard, J.-P., et al., 2012. Spitzer and herchel multiwavelength characterization of the dust content of evolved H II regions. Astrophys. J. 760, 149.
- Persi, P., Tapia, M., Roth, M., Elia, D., López-Vázquez, J.A., 2016. The complex high-mass star-forming region IRAS 15507-5359. Mon. Not. R. Astron. Soc. 459, 1946–1958.
- Pillai, T., Kauffmann, J., Zhang, Q., Sanhueza, P., Leurini, S., Wang, K., Sridharan, T.K., König, C., 2019. Massive and low-mass protostars in massive “starless” cores. Astron. Astrophys. 622, A54.
- Pollo, A., Rybka, P., Takeuchi, T.T., 2010. Star-galaxy separation by far-infrared color-color diagrams for the AKARI FIS all-sky survey (bright source catalog version β -1). Astron. Astrophys. 514, A3. doi:10.1051/0004-6361/200913428, arXiv:1003.1949.
- Retter, B., Hatchell, J., Naylor, T., 2021. Spatial statistics in star-forming regions: Is star formation driven by column density alone? Mon. Not. R. Astron. Soc. 507, 1904–1922.
- Ricco, G., Brescia, M., Cavuoti, S., Mercurio, A., di Giorgio, A.M., Molinari, S., 2017. C³, a command-line catalog cross-match tool for large astrophysical catalogs. Publ. Astron. Soc. Pac. 129, 024005. doi:10.1088/1538-3873/129/972/024005, arXiv:1611.04431.
- Rieke, G.H., Lebofsky, M.J., 1985. The interstellar extinction law from 1 to 13 microns. Astrophys. J. (ISSN: 0004-637X) 288, 618–621, NSF-supported research.
- Singh, D., Singh, B., 2022. Feature wise normalization: An effective way of normalizing data. Pattern Recognit. 122, 108307.
- Strafella, F., Lorenzetti, D., Giannini, T., Elia, D., Maruccia, Y., Maiolo, B., Massi, F., Olmi, L., Molinari, S., Pezzuto, S., 2015. The young stellar object population in the Vela-D molecular cloud. Astrophys. J. 798, 104.
- Tapia, M., Persi, P., Román-Zúñiga, C., Elia, D., Giovannelli, F., Sabau-Graziati, L., 2018. The star-forming cores in the centre of the trifold nebula (M 20): From herchel to the near-infrared. Mon. Not. R. Astron. Soc. 475, 3029–3045.
- Tóth, L.V., Marton, G., Zahorec, S., Balázs, L.G., Ueno, M., Tamura, M., Kawamura, A., Kiss, Z.T., Kitamura, Y., 2014. The AKARI far-infrared surveyor young stellar object catalog. Publ. Astron. Soc. Japan 66, 17.
- Wang, K., Zhang, Q., Testi, L., Tak, F.v.d., Wu, Y., Zhang, H., Pillai, T., Wyrowski, F., Carey, S., Ragan, S.E., et al., 2014. Hierarchical fragmentation and differential star formation in the galactic “snake”: Infrared dark cloud G11. 11- 0.12. Mon. Not. R. Astron. Soc. 439, 3275–3293.
- Wang, K., Zhang, Q., Wu, Y., Li, H.-b., Zhang, H., 2012. Protostellar outflow heating in a growing massive protocluster. Astrophys. J. Lett. 745, L30.
- Wang, K., Zhang, Q., Wu, Y., Zhang, H., 2011. Hierarchical fragmentation and jet-like outflows in IRDC G28. 34+ 0.06: A growing massive protostar cluster. Astrophys. J. 735, 64.
- Werner, M.W., Roellig, T.L., Low, F.J., Rieke, G.H., Rieke, M., Hoffmann, W.F., Young, E., Houck, J.R., Brandl, B., Fazio, G.G., et al., 2004. The spitzer space telescope mission. Astrophys. J. Suppl. Ser. 154, 1.
- Whitney, B.A., Indebetouw, R., Babler, B.L., Meade, M.R., Watson, C., Wolff, M.J., Wolfire, M.G., Clemens, D.P., Bania, T.M., Benjamin, R.A., et al., 2004. A glimpse of star formation in the giant H II region RCW 49. Astrophys. J. Suppl. Ser. 154, 315.
- Wright, E.L., Eisenhardt, P.R.M., Mainzer, A.K., Ressler, M.E., Cutri, R.M., Jarrett, T., Kirkpatrick, J.D., Padgett, D., McMillan, R.S., Skrutskie, M., Stanford, S.A., Cohen, M., Walker, R.G., Mather, J.C., Leisawitz, D., Gautier, III, T.N., McLean, I., Benford, D., Lonsdale, C.J., Blain, A., Mendez, B., Irace, W.R., Duval, V., Liu, F., Royer, D., Heinrichsen, I., Howard, J., Shannon, M., Kendall, M., Walsh, A.L., Larsen, M., Cardon, J.G., Schick, S., Schwalm, M., Abid, M., Fabsky, B., Naes, L., Tsai, C.-W., 2010. The wide-field infrared survey explorer (WISE): Mission description and initial on-orbit performance. Astron. J. 140, 1868–1881. doi:10.1088/0004-6256/140/6/1868, arXiv:1008.0031.
- Xue, M., Jiang, B.W., Gao, J., Liu, J., Wang, S., Li, A., 2016. A precise determination of the mid-infrared interstellar extinction law based on the APOGEE spectroscopic survey. Astrophys. J. Suppl. Ser. 224, 23.
- Yadav, R.K., Pandey, A.K., Sharma, S., Ojha, D.K., Samal, M.R., Mallick, K.K., Jose, J., Ogura, K., Richichi, A., Irawati, P., et al., 2016. A multiwavelength investigation of the H II region S311: Young stellar population and star formation. Mon. Not. R. Astron. Soc. 461, 2502–2518.
- Zhang, Q., Wang, K., Lu, X., Jiménez-Serra, I., 2015. Fragmentation of molecular clumps and formation of a protocluster. Astrophys. J. 804, 141.