A&A 664, A175 (2022)
https://doi.org/10.1051/0004-6361/202243580

Astronomy & Astrophysics

Low-mass young stars in the Milky Way unveiled by DBSCAN and *Gaia* EDR3: Mapping the star forming regions within 1.5 kpc⋆

L. Prisinzano[1], F. Damiani[1], S. Sciortino[1], E. Flaccomio[1], M. G. Guarcello[1], G. Micela[1], E. Tognelli[2], R. D. Jeffries[3], and J. M. Alcalá[4]

[1] INAF – Osservatorio Astronomico di Palermo, Piazza del Parlamento 1, 90134, Palermo, Italy
e-mail: loredana.prisinzano@inaf.it
[2] CEICO, Institute of Physics of the Czech Academy of Sciences, Na Slovance 2, 182 21 Praha 8, Czechia
[3] Astrophysics Group, Keele University, Keele, Staffordshire ST5 5BG, UK
[4] INAF – Osservatorio Astronomico di Capodimonte, via Moiariello 16, 80131 Napoli, Italy

Received 18 March 2022 / Accepted 26 May 2022

ABSTRACT

Context. With an unprecedented astrometric and photometric data precision, *Gaia* EDR3 provides, for the first time, the opportunity to systematically detect and map, in the optical bands, the low-mass populations of the star forming regions (SFRs) in the Milky Way.
Aims. We aim to provide a catalogue of the *Gaia* EDR3 data (photometry, proper motions and parallaxes) of the young stellar objects (YSOs) identified in the Galactic plane ($|b| < 30°$) within about 1.5 kpc. The catalogue of the SFRs to which they belong is also provided to study the properties of the very young clusters and put them in the context of the Galaxy structure.
Methods. We applied the machine learning unsupervised clustering algorithm density-based spatial clustering of applications with noise (DBSCAN) to a sample of *Gaia* EDR3 data photometrically selected on the region where very young stars ($t \lesssim 10$ Myr) are expected to be found, with the aim of identifying co-moving and spatially consistent stellar clusters. A sub-sample of 52 clusters, selected among the 7 323 found with DBSCAN, has been used as template data set to identify very young clusters from the pattern of the observed colour-absolute magnitude diagrams through a pattern-match process.
Results. We find 124 440 candidate YSOs clustered in 354 SFRs and stellar clusters younger than 10 Myr and within $\lesssim$1.5 kpc. In addition, 65 863 low-mass members of 322 stellar clusters located within ∼500 pc and with ages 10 Myr $\lesssim t \lesssim$ 100 Myr were also found.
Conclusions. The selected YSOs are spatially correlated with the well-known SFRs. Most of them are associated with well-concentrated regions or complex structures of the Galaxy, and a substantial number of them have been recognised for the first time. The massive SFRs, such as, for example, Orion, Sco-Cen, and Vela, located within 600–700 pc trace a very complex three-dimensional pattern, while the farthest ones seem to follow a more regular pattern along the Galactic plane.

Key words. methods: data analysis – open clusters and associations: general – catalogs – surveys – stars: formation – stars: pre-main sequence

1. Introduction

It is now well known that stars originate from the collapse of cold molecular clouds and mainly form in over-dense structures and clusters usually designated as star forming regions (SFRs). During the very early phases, young stellar objects (YSOs) can be identified in the near-, mid-, and far-infrared (IR) and radio wavelengths because of the presence of the optically thick infalling envelope or circumstellar disc around the central star. In the subsequent pre-main-sequence phase, they also become visible in the optical bands. However, when the final dispersal of the disc material occurs and non-accreting transition discs form, YSOs can no longer be identified in IR or radio surveys (Ercolano et al. 2021) and a complete census is only possible in the optical bands.

While a clean identification of YSOs is very hard using only optical photometry, an efficient way to systematically single out SFRs is by the identification of kinematical stellar groups with a common space motion. With an unprecedented astrometric precision and sky coverage, *Gaia* data offer the possibility to recognise the SFRs as common proper motion groups, at least within the *Gaia* observational limits.

Data from the *Gaia* mission are revolutionising our ability to map the youngest stellar populations of the Milky Way in the optical bands, which is one of the core science goals for an overall understanding of the Galactic components. The youngest stellar component is crucial to better characterising the Galactic thin disc and its spiral arms and to understanding its origin.

The characterisation of individual SFRs and their dynamics are also fundamental to understanding the local formation, evolution, and dispersion of star clusters, as well as the star formation history and the initial mass function (IMF). Finally, statistical studies of YSOs during the early years of their formation, when the proto-planetary discs are evolving and planets form, are crucial to shedding light on planet formation theory.

With more than 1.3 billion stars with precise proper motions and astrometric (positions and parallaxes) and photometric measurements, *Gaia* DR2 data allowed several studies aimed at identifying clustered populations of the Milky Way. Some of these studies have been dedicated to SFRs, associations, and moving groups. Zari et al. (2018) presented an analysis of

⋆ Tables 3 and 4 are only available at the CDS via anonymous ftp to cdsarc.u-strasbg.fr (130.79.128.5) or via http://cdsarc.u-strasbg.fr/viz-bin/cat/J/A+A/664/A175

the clustered and diffuse young populations within 500 pc, using a combination of photometric and astrometric criteria. Analogously, Kerr et al. (2021) studied the solar neighbourhood by applying the hierarchical density-based spatial clustering of applications with noise (HDBSCAN) algorithm (McInnes et al. 2017). They found 27 young groups, associations, and significant sub-structures, associated with known clusters and SFRs, and released a catalogue including $\sim 3 \times 10^4$ *Gaia* DR2 YSOs within 333 pc.

Cantat-Gaudin et al. (2018) started from a list of known clusters to assign them unsupervised membership and parameters. Other studies have been dedicated to systematically finding open clusters in the Galaxy. Castro-Ginard et al. (2018) used the density-based spatial clustering of applications with noise (DBSCAN) algorithm (Ester et al. 1996) to select a list of candidate open clusters (OC), which they then refined to identify real OCs with a well-defined main sequence (MS). Other papers have recently been published detailing the discoveries of new open clusters and the deduction of their parameters (e.g. Cantat-Gaudin & Anders 2020; Cantat-Gaudin et al. 2020; Castro-Ginard et al. 2020; Liu & Pang 2019).

A recent attempt to find Galactic plane (GP) clustered populations, including SFRs, was made by Kounkel & Covey (2019) and Kounkel et al. (2020), again using *Gaia* DR2 data and the HDBSCAN unsupervised algorithm in 5D space (l, b, π, $\mu_\alpha*$, μ_δ). In these works, the first limited to 1 kpc and the second to 3 kpc, they found clustered populations, associations, moving groups, and string-like structures, parallel to the GP, spanning hundreds of parsec in length. Clusters aged between 10 Myr and 1 Gyr have been found with an onion-like approach using the entire catalogue with different cut-offs in parallax and progressively merging the different catalogues.

A different approach was adopted by Bica et al. (2019), who used infrared (IR) data from 2MASS, WISE, VVV, *Spitzer*, and *Herschel* surveys to compile a catalogue of 10 978 Galactic star clusters, and associations, including 4234 embedded clusters.

With the advent of *Gaia* Early Data Release 3 (EDR3), based on 34 months of observations[1], available photometric and astrometric measurements improved significantly. In particular, photometric improvements have been made in the calibration models, in the different photometric systems, and in the treatment of the BP and RP local background flux (Riello et al. 2021).

In this work, we used *Gaia* EDR3 data to systematically identify the low-mass component of SFRs in the Galaxy, with ages approximately <10 Myr and within a distance limit of ~1.5 kpc imposed by our data selection. We focused our analysis on very young clusters by exploiting the significant progress achieved with *Gaia* EDR3 data. A full exploitation of the Gaia data and the results presented here would require further data, such as spectroscopic determination of individual stellar parameters, such as effective temperatures, gravities, and stellar luminosities, as well as rotational and radial velocities, which are crucial to deriving masses, ages, and 3D space velocities. Even though the results presented here cannot be used at this stage to determine the IMF, star formation history, and 3D kinematics of the SFRs, they can be used to trace the very young Galactic stellar component within 1.5–2 kpc through a systematic method that homogeneously identifies the bulk population of the SFRs. Such results can be used both for statistical and individual detailed analyses. The paper is organised as follows. In Sect. 2, we describe the requirements adopted to select the *Gaia* EDR3 data, and in Sect. 3 the photometric selection applied to obtain the starting sample of the YSO candidates. In Sect. 4, we describe the method adopted to identify SFRs and stellar clusters, the criteria adopted to validate them, and the age classification. Our results and the discussion are presented in Sects. 5 and 6, respectively; finally, our summary and conclusions are presented in Sect 7. In Appendix A we show the effects of the reddening in the *Gaia* colour-absolute magnitude diagrams, in Appendix B we estimate the effect of multiplicity in the selection of the YSOs, while in Appendix C we describe the comparison of specific regions with the literature.

2. *Gaia* data

In this analysis, we used the *Gaia* EDR3 data (Gaia Collaboration 2016, 2021), which provide precise astrometry and kinematics (l, b, π, $\mu_\alpha*$, μ_δ) as well as excellent photometry in three broad bands (G, $G_{\rm BP}$, $G_{\rm RP}$). Since our analysis is focussed on the Galactic midplane, where most of the YSOs are expected to be found, we selected sources within $|b| < 30°$. We limited our selection to $7.5 < G \leq 20.5$. The limit $G = 7.5$ was chosen in order to discard objects with magnitudes derived from saturated charge-coupled device (CCD) images, while $G = 20.5$ is the limit to include most of the objects with magnitude G uncertainties lower than 0.2 mag. This range includes the young, low-mass populations ($0.1 \lesssim M/M_\odot \lesssim 1.5$) of the known SFRs within the distance set by the limiting magnitude. In addition, we only considered positive parallax values. This choice does not introduce any bias since we do not expect to investigate stars with very small parallaxes that could have negative values (Luri et al. 2018). Finally, we imposed a relative parallax error lower than 20% in order to discard stars with a poorly constrained distance, and, to take into account the *Gaia* EDR3 systematics, we also considered the renormalised unit weight error (RUWE; Lindegren et al. 2021b), which is expected to be <1.4 for sources where the single-star model provides a good fit to the astrometric observations.

To summarise, data of our interest were selected from the Astronomical Data Query Language (ADQL) interface of the ESA *Gaia* Archive[2] using the following restrictions:

$$\begin{cases} |b| < 30° \\ 7.5 < G \leq 20.5 \\ \pi > 0\,\text{mas} \\ \sigma(\pi)/\pi < 0.2 \\ \text{RUWE} < 1.4 \end{cases} . \tag{1}$$

We also included a photometric condition in the query aimed to include the pre-main-sequence (PMS) region of the M_G versus $G - G_{\rm RP}$ colour-absolute magnitude diagram (CAMD) where all very young stars ($t \lesssim 10$ Myr) are expected to be found. We split our selection in two samples, namely bright and faint, according to the following criteria:

$$\text{Bright sample} = \begin{cases} M_G < 7.64\,(G - G_{\rm RP}) + 0.22 \\ 5 < M_G \leq 9 \\ (G - G_{\rm RP}) > 0.58 \end{cases} , \tag{2}$$

$$\text{Faint sample} = \begin{cases} M_G < 15.00\,(G - G_{\rm RP}) - 8.25 \\ M_G > 9 \\ (G - G_{\rm RP}) > 0.58. \end{cases} . \tag{3}$$

[1] *Gaia* DR2 data were based on 22 months of observations.

[2] https://gea.esac.esa.int/archive/

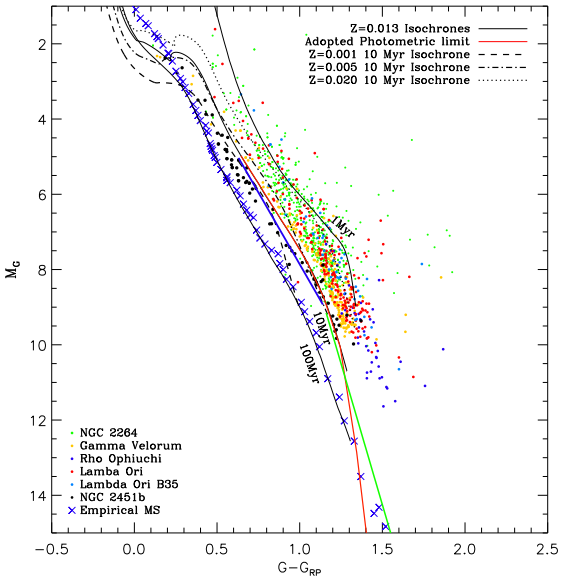

Fig. 1. CAMD of YSOs of some representative young clusters with membership probabilities >0.90 assigned by combining spectroscopic and *Gaia* EDR3 criteria (Jackson et al. 2022). Blue x symbols trace the empirical sequence by Pecaut & Mamajek (2013). Members of the clusters Gamma Velorum (18 Myr old) and NGC2451b (50 Myr old) are also shown. Black solid lines are the theoretical solar metallicity Pisa isochrones while the red solid line is the complete photometric limit adopted in this work including the low mass extrapolation. Dashed, dashed-dotted and dotted lines are the 10 Myr isochrones at different metallicities. Blue and green solid lines represent the limits described by the equations 2 and 3.

These limits are drawn as solid blue and green lines in Fig. 1. We note that in this work, for the reddening uncorrected absolute magnitudes, we adopted the definition $M_G = G + 5\log(\pi) - 10$, based on the inverted *Gaia* EDR3 parallaxes, since, as shown in Piecka & Paunzen (2021), within <2 kpc, the inverse-parallax method gives results comparable to distances derived by the Bayesian approach (Bailer-Jones et al. 2021).

The minimum value $M_G = 5$ was set to avoid the upper region of the colour-absolute magnitude diagram, where the overlap of the upper MS or PMS stars of the SFRs with giants, MS, or turn-off stars is expected to be very high, especially if the reddening is not corrected. This implies a cut of the massive population of the SFRs, but it does not represent an issue for our investigation since we are mainly interested in the rich low-mass component of these populations. In order to further reduce the fraction of contaminants, also we used the condition $G - G_{RP} > 0.58$, which is the minimum expected unreddened colour for low-mass ($M \lesssim 1.2\,M_\odot$) PMS (age ≤10 Myr) stars.

Our photometric selection and the subsequent analysis are based on the $G - G_{RP}$ colours. This choice allows us to avoid the use of the G_{BP} magnitudes that for $G \gtrapprox 20$ are strongly affected by the application of the minimum flux threshold, which overestimates the mean BP flux. This issue also affects the RP flux, but with a considerably lower effect in G_{RP} than in G_{BP} (Riello et al. 2021). Once the data had been retrieved by the ESA Gaia Archive, parallax values were corrected by the zero point bias reported in Lindegren et al. (2021a) using the Python code available to the community[3], which is a function of source magnitude, colour, and celestial position.

[3] https://gitlab.com/icc-ub/public/gaiadr3_zeropoint

In addition, we performed further data filtering by only considering objects with errors smaller than 0.14 mag in $G - G_{RP}$. Standard errors in the magnitudes were computed using the propagations of the flux errors with the following formulas:

$$\sigma(G) = \sqrt{(-2.5/ln(10)\sigma(FG)/FG)^2 + \sigma(G_0)^2}, \quad (4)$$

$$\sigma(G_{BP}) = \sqrt{(-2.5/ln(10)\sigma(FG_{BP})/FG_{BP})^2 + \sigma(G_{BP_0})^2}, \quad (5)$$

$$\sigma(G_{RP}) = \sqrt{(-2.5/ln(10)\sigma(FG_{RP})/FG_{RP})^2 + \sigma(G_{RP_0})^2}, \quad (6)$$

where FG, FG_{BP}, and FG_{RP} are the mean fluxes in the G, BP, and RP bands, respectively, and $\sigma(G_0) = 0.0027553202$, $\sigma(G_{BP_0}) = 0.0027901700$, and $\sigma(G_{RP_0}) = 0.0037793818$ are the *Gaia* EDR3 zero-point uncertainties[4].

3. Photometric selection of the input sample

In this section, we describe and discuss how we performed the final photometric selection of the sample used as input for the subsequent clustering analysis, which is based on the astrometric and kinematic *Gaia* EDR3 parameters, as described in Sect. 4. By considering the typical complexity of the environment of young stars and the dependence of the reddening law from the stellar effective temperature due to the large spectral range covered by the *Gaia* bands (Anders et al. 2019), we did not attempt to correct colours and magnitudes for reddening and absorption, but we used their observed values. This is certainly one of the main sources of contamination by older stars to be overcome, as we discuss later in the paper.

Our goal is to start from a complete sample, including all potential YSOs with ages <10 Myr, at least in the photometric range set described in Sect. 2. In particular, we selected the objects with M_G falling on the red side of the solar-metallicity 10 Myr isochrone computed using the PISA models (Dell'Omodarme et al. 2012; Randich et al. 2018; Tognelli et al. 2018, 2020) in the M_G versus $G - G_{RP}$ diagram shown in Fig. 1. To check if the selected photometric limit is compliant with our requirements, we compared it with the reddening uncorrected CAMD of some SFRs and young clusters for which membership was recently derived by Jackson et al. (2022) based on the 3D kinematics of the spectroscopic targets. We find that the adopted 10 Myr isocrone delimits the PMS region of clusters, such as NGC 2264, Lambda Ori, Lambda Ori B35, and Rho Ophiuchi, which are in our main age range ($t < 10$ Myr) of interest. However, members of ∼20 Myr old clusters, such as Gamma Velorum, also fall completely in the selected photometric region, while members of ∼50-Myr-old clusters, such as NGC 2451b, fall partially in the selected photometric region at $M_G \gtrsim 9$. Going to clusters with ages of $t > 50$ Myr the overlapping region occurs at fainter magnitudes.

Since the adopted isochrone is limited to 0.1 $M_\odot$, corresponding to $M_G = 10.7$, the photometric limit at fainter magnitudes was extrapolated using a linear extrapolation. To check the position of such extrapolation, we compared it with the empirical sequence by Pecaut & Mamajek (2013), for which mean stellar colours and effective temperatures are given down to M and L spectral types and that can be used as an upper limit to the region we are interested in. Our photometric limit approaches

[4] See https://www.cosmos.esa.int/web/gaia/edr3-passbands for further details

such a sequence and crosses it at $M_G \sim 13$. This ensures we set an inclusive photometric selection close to the MS at the lowest mass tail. In fact, even though this implies the inclusion of stars older than 10 Myr, it avoids a bias against the selection of very young stars.

We note that for the photometric selection, the minimum and maximum M_G associated with each observed star have been computed by considering the $1\,\sigma$ parallax uncertainties, which are dominant with respect to the magnitude uncertainties. The photometric selection with respect to the reference isochrone was performed by considering the compatibility of M_G magnitudes with respect to their minimum and maximum values; that is, they were selected if either their minimum or maximum value lay inside the selection region. At the end of this selection, we were left with a catalogue of 18 057 300 *Gaia* EDR3 entries.

Performing a photometric selection as inclusive as possible, as we have done, implies the introduction of a significant contamination by old field or open cluster stars, mainly due to the uncorrected reddening, binarity, or overlapping photometric region in the low-mass range, where the sensitivity of the $G - G_{RP}$ colours in distinguishing PMS or MS stars becomes very low. However, the contamination by field stars does not represent a significant issue for our clustering analysis, since they are not expected to share similar astrometric and kinematic properties. In addition, since we aim to investigate the low-mass component of the SFRs, which is also the most dominant ($\gtrsim$80%; Lada 2006), the statistical contrast with respect to field contaminants is expected to be favourable to detecting them.

A more complex effect of our inclusive photometric selection is that clusters older than 10 Myr can also partially fall in the selected region and be recognised as candidate clusters in the subsequent analysis. As shown in Fig. 1, at faint absolute magnitudes ($M_G > 9$), the model-computed isochrones are not very sensitive to stellar ages and tend to overlap, especially in the M_G versus $G - G_{RP}$ diagram. In addition, spectral synthesis of M dwarf stars suffers from the accuracy of the adopted atmosphere models and/or from incomplete molecular data. The model-predicted colours of very-low-mass stars are therefore uncertain. A further complication is the observed discrepancy between radii and colours of low-mass stars, likely due to the distorting effects of magnetic activity and star spots on the structure of active stars (Somers et al. 2020; Franciosini et al. 2021). All these effects cause a spread of the low-mass MS and can bring magnitudes and colours of ~100-Myr-old stars to the region selected by us as compatible with stars with $t < 10$ Myr. For all these reasons, as discussed, for example, in Jeffries et al. (2017), the ages judged from 'standard' isochrones are almost certainly underestimated due to a systematic bias.

At faint magnitudes, the fraction of old cluster members falling in the adopted photometric region decreases with cluster ages. Hence, clusters of about 20–30 Myr will be almost completely included in our selected sample, while at the age of 100–500 Myr only the low-mass tail will be included. However, because of the adopted photometric limit, the low-mass tails will be included only for relatively close clusters ($d < 500$ pc).

As already mentioned before, a partial contamination by old cluster members in our photometric sample can occur also for bright stars ($M_G < 9$–10) if their reddening or a binary status gives them observed magnitudes and colours compatible with the selected photometric region. As shown in Appendix A, the effects of using colours and magnitudes uncorrected for reddening are expected to be more severe for reddened stars with spectral types earlier than G, in comparison with later spectral types, in the sense that the selected sample is expected to be contaminated mainly by these objects, which fall in the brightest part of the photometric region adopted in this work. The implications of this contingency are discussed in the following sections.

Finally, we also considered the possible effects due to the metallicity on the selection by considering 10 Myr isochrones for a metallicity lower or higher than solar. The comparison shows that while YSOs with over-solar ($Z = 0.020$, [Fe/H] = 0.2) or sub-solar ($Z = 0.005$, [Fe/H] = −0.45) metallicities would fall in the selected photometric region, very-metal-poor YSOs ($Z = 0.001$, [Fe/H] = −1.10) would remain outside. However, as recently found by Spina et al. (2017) at galactocentric radii from ~6.5 kpc to 8.70 kpc, young open clusters and SFRs have close-to-solar or slightly sub-solar metallicities, and therefore we conclude that no SFRs are expected to be missed for metallicity effects with our photometric assumptions.

Based on the adopted photometric selection, our data set encompasses all YSOs of ages $t \lesssim 10$ Myr and observed $M_G > 5$, including the most reddened ($A_V < 3$–4) that can be detected with *Gaia*. Even YSOs with accretion (e.g. Gullbring et al. 1998) or that are seen in scattered light (Bonito et al. 2013) or flares in M-type stars (e.g. Mitra-Kraev et al. 2005) are expected to be included in our sample. In fact, these phenomena affect the $G_{BP} - G$ or the $G_{BP} - G_{RP}$ colours, causing the stellar colours to become bluer than their photospheric colours, while, on the contrary, their effect on the $G - G_{RP}$ colours goes in the same direction as the reddening, causing these latter colours to become redder.

We stress, however, that the constraint $M_G < 5$, adopted to strongly reduce the contamination due to reddened turn-off or MS stars, makes the selected photometric sample incomplete for the massive stellar component of the SFRs. A further expected missing stellar component is that of binary systems of the clusters, due to the restriction of the *Gaia* data to RUWE < 1.4 (see Appendix B). In addition, since available data do not allow us to obtain reliable corrections for the reddening affecting colours and magnitudes of the selected YSOs, accurate stellar parameters such as individual stellar ages and masses will not be derived in the subsequent analysis. However, even though the results we aim to achieve are not suitable for investigations based on complete young populations or accurate stellar parameters, they are expected to trace the dominant component of the SFRs, that is, their low-mass population, and will be crucial to an overall systematic view of the Galactic SFRs located within 1–2 kpc of the Sun, as well as for detailed individual or statistical investigations of these YSOs.

4. Method

4.1. Clustering with DBSCAN

This section describes the methodology used to search for candidate clusters with an unsupervised algorithm, such as overdensities in the 5D *Gaia* EDR3 astrometric and kinematics parameters (l, b, π, $\mu_{\alpha*}$, μ_δ). Starting from the data set selected as described in the Sect. 3, we performed a clustering analysis using the DBSCAN code (Ester et al. 1996), within the scikit-learn machine-learning package in Python. First of all, we prepared a grid of $5^\circ \times 5^\circ$ boxes, covering the entire range of the Galactic longitudes l and for $|b| < 30^\circ$. In this step, we took into account the discontinuity at $l = 0^\circ$. To homogenise the variables with different dimensions to comparable values, the five parameters (l, b, π, $\mu_{\alpha*}$, μ_δ) within each box were first re-scaled using

the `RobustScaler` Python code based on a statistics robust to outliers, according to the interquartile range.

The DBSCAN algorithm requires only two input parameters (ϵ, *minPts*). It identifies candidate clusters as overdensities in a multi-dimensional space (5D in our case) in which the number of sources exceeds the required minimum number of points *minPts*, within a neighbourhood of a particular linking length, ϵ, for all five parameters, using a statistical distance that is assumed to be Euclidean. DBSCAN does not require us to know an a priori number of clusters, and it is able to detect arbitrarily shaped clusters. This is crucial for our analysis aimed at finding SFRs that can be characterised by circular or elongated or asymmetric shapes, reminiscent of the native molecular clouds. In order to determine the best input parameters (ϵ, *minPts*) to give as input to DBSCAN, we experimented with several values in the direction of well-known SFRs, and we noted that in the same direction more than a combination of the two parameters is needed to reveal different real clusters located at different distances. This is due to the fact that close candidate clusters, such as associations and co-moving groups, can appear spatially (in l and b) sparse, while they are definitively clustered in distance and proper motions; yet, in the same direction it is possible to identify distant but spatially concentrated candidate clusters. In the two cases, the choice of two different ϵ values rather than a single ϵ is required to detect these kinds of clusters.

Based on this preliminary empirical analysis, we decided to run the DBSCAN codes in the entire GP, by adopting a total of 900 combinations of (ϵ, *minPts*) values with ϵ ranging from 0.1 to 9 in steps of 0.1 and *minPts* ranging from 5 to 50 in steps of 5. In addition, to account for candidate clusters falling in the borders of the defined boxes, we defined another four sets of grids by shifting the original boxes by $\delta l = \delta b = [1°, 2°, 3°, 4°]$ with respect to the original boxes. In the following, we refer to the five sets of grids as spatial configurations. At the end, we run DBSCAN within a total of $360/5 \times 60/5 \times 5 = 4320$ different boxes with 900 combinations of parameter sets (ϵ, *minPts*).

4.2. Candidate cluster validation

One of the most challenging phases of this analysis has been the validation of the recognised candidate clusters. In fact, DBSCAN is an unsupervised density-based algorithm, and, as a consequence, it picks up not only overdensities that correspond to real OCs, but also overdensities in purely statistical terms. For this reason, our a posteriori validation approach is based on the exploitation of two astrophysical constraints, based on the typical properties of the SFRs, by avoiding the introduction of strong biases.

Star forming regions are not characterised by well-defined age sequences, and they are typically observed in the Hertzsprung-Russell (HR) diagrams as ensembles showing an apparent luminosity spread, often associated with an age spread (e.g. Palla & Stahler 1999; Palla et al. 2005). On the other hand, such spreads have also been ascribed to complex phenomena affecting their photometry, such as variability, accretion and outflows, extinction, binarity, and our inability to quantify their contribution (Soderblom et al. 2014). Neverthless, SFRs are usually observed with a typical mass distribution that can be shaped by a standard (or closely resembling standard) IMF, characterised by an increasing fraction of members going towards decreasing masses, at least until masses of $\sim 0.3\,M_\odot$ (e.g. Salpeter 1955; Scalo 1998; Chabrier 2003).

Since we exploited the excellent *Gaia* EDR3 results down to $G = 20.5$, within reasonable reddening values ($A_V \lesssim 1$), with our data set we expect to detect YSOs with spectral types down to M-type and at distances $\lesssim$1.5 kpc. This is the case, for example, of the cluster NGC 6530, located at around 1.3 kpc, for which the low-mass population down to $0.4\,M_\odot$ has been detected at $V \sim 20$ (Prisinzano et al. 2005), roughly corresponding to our G magnitude limit.

Based on these considerations, a physically recognisable candidate cluster should include its tail of low-mass members. Hence, we imposed a minimum threshold of ten objects with $M_G > 7.7$, which means requiring candidate clusters to have at least ten stars with $M \lesssim 0.5\,M_\odot$, assuming the isochrone of 10 Myr from the Pisa models.

A further parameter that we considered as an indicator of reliability for the candidate cluster validation is the dispersion of the distances of each cluster. The observed total distance dispersion is a combination of the intrinsic dispersion plus the contribution due to the measurement errors. While the intrinsic dispersion does not depend on the distance, the contribution due to the measurement errors becomes dominant at large distances since *Gaia* EDR3 parallaxes become much more uncertain. Thus, among the parameters used to find overdensities by DBSCAN, the observed standard deviation of the distances is the most critical parameter to be constrained for the identification of real clusters. To this aim, for the cluster validation, we constrained the maximum allowed observed dispersion. For distances <1 kpc, the constraint is set on the ratio between the standard deviation of the distances of the putative members and the derived mean distance for the given candidate cluster. For a valid candidate cluster, the above ratio has to be <0.2. For more distant candidate clusters, we adopted the more stringent constraint that the standard deviation should be smaller than 200 pc. This limit was chosen considering that, for NGC 2244, located at ∼1.6 kpc and one of the most distant clusters that we detect, the distance dispersion is about 175 pc, and therefore we do not expect to find real physical clusters with a distance dispersion larger than this threshold. These choices may limit our ability to detect clusters at distance $\gtrsim$1.5 kpc, for which we could, in principle, detect, at the magnitude limit of our data set, the massive component of the clusters down to $\sim 1\,M_\odot$ regime. However, since the accuracy of *Gaia* EDR3 parallaxes and kinematic data beyond this limit becomes very low, we prefer to maintain our constraints at the cost of limiting our analysis to smaller distances.

The adopted constraints on the distance dispersion of cluster members have shown to be very effective in rejecting a large number of (unexpected) candidate massive clusters recognised by DBSCAN, typically with more than 1000 members located at distances $\gtrsim$1 kpc, which do not include M-type stars but only earlier stars and are characterised by very large dispersions in distance. These structures are likely those identified as strings in Kounkel & Covey (2019); Kounkel et al. (2020). However, since we do not recognise these structures as standard clusters, any further investigation of them is beyond the scope of this work.

The final cluster member selection was only performed for candidate clusters that satisfy the previous constraints. As a result of our choice of the DBSCAN input parameters (see Sect. 4.1) and of the adopted spatial configurations, a given candidate cluster can be identified by adopting similar input parameters, with possible small differences in the cluster membership. In addition, for a given pair of input parameters in two or more overlapping boxes, a given candidate cluster can be identified in more than one box (with the same membership result) if the candidate cluster is spatially small enough to be completely identified. Alternatively, it can be completely detected within

one box and only partially detected in a box where the candidate cluster falls at the borders. In order to assign the most likely membership for a given cluster, we proceeded by adopting the following strategy.

We first considered the candidate clusters detected within the same spatial configuration but with different set of parameters (ϵ, $minPts$). For each of the selected candidate clusters, we computed the median values of the five parameters (l, b, π, $\mu_{\alpha}*$, μ_{δ}) and then selected all the candidate clusters that were simultaneously compatible in these five parameters; that is, if the two compared distributions of each parameter overlap around the median, within half of the total width. Among the compatible candidate clusters, we selected the most populated and discarded the others. This strategy allowed us to identify the most persistent candidate clusters on different scales.

In the subsequent step, we compared the candidate clusters identified in each of the five spatial configurations to select the best configuration, or, likewise, the best box in which the spatial coverage of the candidate cluster is maximised. Since we can have more than one detection of the same cluster, for each member we only selected the configuration for which it is associated with the most populated candidate cluster, and that member was removed from the less populated clusters as identified by DBSCAN. The peripheral members of candidate clusters covering a spatial region larger than the area of the box ($5^\circ \times 5^\circ$), left out from the richest centred candidate cluster, were only considered as additional candidate clusters if they included at least ten elements[5]; the same limit was also assumed in other similar works (e.g. Castro-Ginard et al. 2018; Kerr et al. 2021). This selection strategy allowed us to also include likely members at the candidate cluster's periphery, providing data for further investigations on the dynamics of these stellar clusters. At the end of this process, we are left with a total of 449 849 detected stars within 14 178 single candidate clusters.

Many SFRs are associated with giant molecular clouds, and thus they can have a spatial extension larger than the box of $5^\circ \times 5^\circ$ used for our analysis. In order to merge candidate clusters belonging to the same complex, we proceeded as follows: we computed the median and the 16th and 84th percentiles of the distance and proper motion distributions. Then, we merged all neighbouring clusters for which distances and proper motions were compatible within 1 σ. The total number of merged clusters is 7 323.

4.3. Cluster age classification

From a visual inspection of the photometric properties of the clusters found with this analysis, we note that, while for most of the recognised clusters their selected members of any mass stay in the PMS region of the CAMD as expected, there is a fraction of recognised clusters for which only the low-mass members stay in that PMS region. This is, for example, the case of clusters with low or moderate extinction ($A_V \lesssim 1$) and ages of 10 Myr $\lesssim t \lesssim$ 50 Myr, such as IC 2602, Melotte 20, NGC 2451 A, and NGC 2451 B, where part of the MS or PMS low-mass tail ($M_G \gtrsim 9$) overlaps the photometric region considered here. For clusters with ages of $t \sim$ 100–200 Myr, such as Melotte 22 (Pleiades), NGC 2422, and NGC 2516, a smaller fraction of the MS low-mass tail, likely composed of reddened members, cluster binaries or PMS members, is selected.

Further reddening effects or poorly constrained magnitudes or parallaxes can bring colours or magnitudes of members of even older clusters within the PMS photometric region considered in this work. For clusters with extinctions of $A_V \gtrsim 1$, the MS of $t \gtrsim$ 100 Myr old clusters in the $5 < M_G \lesssim 8$ range fall to the right of the unreddened 10 Myr isochrone. Thus, depending on the cluster age, binaries or reddened members of clusters with ages of $t >$ 10 Myr can also fall in the selected photometric region. Since these objects share the same proper motions and are at the same distance, they are recognised as belonging to a cluster and are therefore included in our catalogue.

To distinguish SFRs from old clusters, we adopted a pattern match procedure based on the extraction of the different patterns that characterise the observed CAMD of clusters of different ages. Among the clusters identified as described in the previous sections, we selected those listed in Table 1 (52 in total) and we used them as a template data set.

In the template data set, we identified 28 clusters, shown in Fig. 2, that we used as a proxy for clusters with ages of $t \lesssim$ 10 Myr. Such clusters were selected since most of them show a consistent luminosity spread, typical of the SFRs, starting from our brightest limit, $M_G = 5$. However, their general shape is also set by the reddening and the distance, with the observed M_G maximum limit that increases as distance decreases. All these cases have been included in the template data set to retrieve all the possible patterns observed in the CAMD due to different ages, distances, reddening, and cluster richness. For each of these clusters, we assigned an increasing flag from 1 to 28, aimed at representing the different shapes of the observed CAMD shown in Fig. 2.

We also identified eight clusters as representative of the ages $10 \lesssim t/\mathrm{Myr} \lesssim 100$, flagged from 29 to 36, according to the ages given in Cantat-Gaudin & Anders (2020). The observed CAMDs of these clusters are shown in Fig. 3. These clusters show an evident PMS region that is mainly populated in the range of $M_G \gtrsim 8$ (e.g. NGC 2451B, NGC 2232), as per our photometric selection. Such a region becomes thinner and thinner for older clusters such as Melotte 20 and Melotte 22. Finally, we selected 16 clusters, flagged from 37 to 52 as a proxy for clusters with ages of $t \gtrsim$ 100 Myr, in agreement with Cantat-Gaudin & Anders (2020). Most of these clusters have been included in the template sample to take into account the non-uniform distribution of the absolute magnitudes of their members in the observed CAMD. In fact, while for very young clusters it is uniformly populated, accordingly to their age and the IMF, the population is not entirely identified for these reddened and old clusters. For example, the clusters with flags from 43 to 52 are characterised in the CAMD by an overdensity of members with $M_G \lesssim 9$. Most of them are quite distant clusters ($d \gtrsim$ 500 pc) and thus very likely affected by reddening. As shown in Appendix A, the effect of the reddening for the *Gaia* bands depends on the stellar effective temperature (Anders et al. 2019), and for high mass stars such an effect is greater than for low-mass stars. This would explain the presence of the peak at higher masses in the observed magnitudes of the CAMD for most of these clusters. Depending on the cluster distance, part of the low-mass tail is also detected, but the overall non-uniform pattern of their CAMD is different from that expected for young clusters. Since most of the clusters show asymmetric structures, to evaluate their extension we estimated the radius in which half of the identified members are concentrated as $r_{50} = 0.5 \times \sqrt{(width^2 + height^2)}$, as was done in Cantat-Gaudin & Anders (2020).

In our final catalogue, we also noted the presence of other photometrically unphysical aggregates including mostly only faint stars (with $G > 18.5$) with very red $G - G_{RP}$ colours and a horizontal distribution in the CAMD likely compatible with

[5] For this reason, our catalogue includes cases in which a single physical cluster is identified by more than one DBSCAN cluster.

Table 1. Clusters used as template data set to select SFRs and other stellar clusters.

Literature name	Flag	Reference	l (deg)	b (deg)	r_{50} (deg)	d (pc)	log t (yr)	N
[LK2002] Cl10	1	Le Duigou & Knödlseder (2002)	79.867	−0.908	0.886	1557		167
65.78–2.61	2	Avedisova (2002)	66.153	−3.123	1.194	1324		134
Rosette	3	Zucker et al. (2020)	206.438	−1.903	2.025	1571	7.1	810
NGC 6530	4	Dias et al. (2002)	6.060	−1.287	1.020	1364		635
NGC 6531	5	Dias et al. (2002)	7.585	−0.338	1.634	1350	8.6	804
UBC 386	6	Cantat-Gaudin & Anders (2020)	100.562	8.694	1.147	1280	6.8	193
Ass Cyg OB 9	7	Sitnik (2003)	78.753	1.778	2.293	1339	8.1	616
Serpens South molecular cloud	8	Fernández-López et al. (2014)	29.364	2.870	0.976	920		123
CygOB7 CO Complex	9	Dutra & Bica (2002)	92.653	2.529	0.950	1123		46
BRC 27	10	Rebull et al. (2013)	224.621	−2.244	3.027	1233	6.9	1709
[DB2002b] G352.16+3.07	11	Otrupcek et al. (2000)	−7.866	3.002	4.764	1169	7.0	2357
IC 1396	12	Zucker et al. (2020)	99.236	4.733	7.407	945	7.4	3140
[MML2017] 2399	13	Miville-Deschênes et al. (2017)	33.890	0.643	2.543	609		130
Chamaeleon II	14	Zucker et al. (2020)	−56.363	−14.720	2.452	200		41
Cepheus	15	Zucker et al. (2020)	108.911	4.359	9.748	923	8.2	11445
NGC 7039	16	Cantat-Gaudin & Anders (2020)	88.350	−1.717	5.322	767	7.3	1048
[YDM97] CO 14	17	Yonekura et al. (1997)	104.508	13.950	3.039	350		124
Serpens	18	Zucker et al. (2020)	28.783	3.082	10.166	455	7.2	2388
IC 348	19	Cantat-Gaudin & Anders (2020)	160.790	−15.812	11.430	334	7.4	2661
Chamaeleon I	20	Zucker et al. (2020)	−62.781	−15.444	3.099	192		156
Taurus	21	Zucker et al. (2020)	172.114	−15.302	4.551	131		112
Ophiuchus	22	Zucker et al. (2020)	−8.024	18.781	12.655	144		2398
Corona Australis	23	Zucker et al. (2020)	−0.132	−17.592	3.291	155		107
[DB2002b] G302.72+4.67	24	Dutra & Bica (2002)	−57.143	4.739	5.854	112		235
Pozzo 1	25	Cantat-Gaudin & Anders (2020)	261.858	−8.321	13.343	398	8.3	6001
ASCC 32	26	Cantat-Gaudin & Anders (2020)	237.327	−9.186	9.878	818	8.4	4416
Lac OB1	27	Chen & Lee (2008)	96.762	−15.032	11.268	548	7.4	2367
RSG 8	28	Cantat-Gaudin & Anders (2020)	109.331	−1.212	12.055	468	7.4	2900
NGC 2451B	29	Cantat-Gaudin & Anders (2020)	253.198	−7.499	9.513	401	7.6	2826
NGC 2232	30	Cantat-Gaudin & Anders (2020)	215.533	−7.983	13.427	372	7.2	1703
Sco OB2 UCL	31	de Zeeuw et al. (1999)	−29.000	16.813	15.052	145		1189
IC 2602	32	Cantat-Gaudin & Anders (2020)	−70.259	−5.011	6.825	151	7.6	315
NGC 2516	33	Cantat-Gaudin & Anders (2020)	−86.236	−15.931	6.881	427	7.6	1156
Melotte 20	34	Cantat-Gaudin & Anders (2020)	147.504	−6.461	8.867	174	7.7	414
Melotte 22	35	Cantat-Gaudin & Anders (2020)	166.573	−23.406	5.882	137	7.9	296
NGC 2422	36	Cantat-Gaudin & Anders (2020)	230.995	3.061	6.238	500	8.0	347
Alessi 12	37	Cantat-Gaudin & Anders (2020)	67.678	−11.723	3.977	546	8.1	127
NGC 3532	38	Cantat-Gaudin & Anders (2020)	−72.815	2.279	4.851	561	8.6	88
IC 6451	39	Cantat-Gaudin & Anders (2020)	−19.939	−7.821	1.257	1068	9.2	86
NGC 6087	40	Cantat-Gaudin & Anders (2020)	−32.077	−5.426	2.532	1007	8.0	77
Alessi 62	41	Cantat-Gaudin & Anders (2020)	53.676	8.773	3.561	622	8.4	87
UPK 33	42	Cantat-Gaudin & Anders (2020)	27.965	0.108	3.931	518	8.4	111
NGC 1647	43	Cantat-Gaudin & Anders (2020)	180.355	−16.861	2.141	606	8.6	272
NGC 6124	44	Cantat-Gaudin & Anders (2020)	−19.205	6.078	5.404	648	8.3	1102
NGC 6494	45	Cantat-Gaudin & Anders (2020)	9.714	2.980	5.537	755	8.6	680
IC 4725	46	Cantat-Gaudin & Anders (2020)	14.022	−4.595	4.807	669	8.1	788
Alessi 44	47	Cantat-Gaudin & Anders (2020)	37.075	−11.510	7.285	587	8.2	637
Stock 2	48	Cantat-Gaudin & Anders (2020)	133.371	−1.160	8.292	384	8.6	727
NGC 2168	49	Cantat-Gaudin & Anders (2020)	186.647	2.327	2.616	928	8.2	118
DSH J2320.1+5821A	50	Kronberger et al. (2006)	111.248	−2.785	2.394	1131		243
UPK 143	51	Cantat-Gaudin & Anders (2020)	91.810	0.514	1.752	934	8.4	262
Collinder 421	52	Cantat-Gaudin & Anders (2020)	79.429	2.527	1.061	1265	8.4	154

Notes. Flag is the value assigned to each cluster to characterise a given observed CAMD shape. r_{50} is the radius in which half of the identified members are concentrated, d is the distance obtained by inverting the median value of the member parallaxes and N is the number of members. Flag = [1, 28] are assigned to clusters with ages $t \lesssim 10$ Myr, Flag = [29, 36] are assigned to clusters with ages $10 \lesssim t/\,\mathrm{Myr} \lesssim 100$, Flag = [37,52] are assigned to clusters with ages $t \gtrsim 100$ Myr.

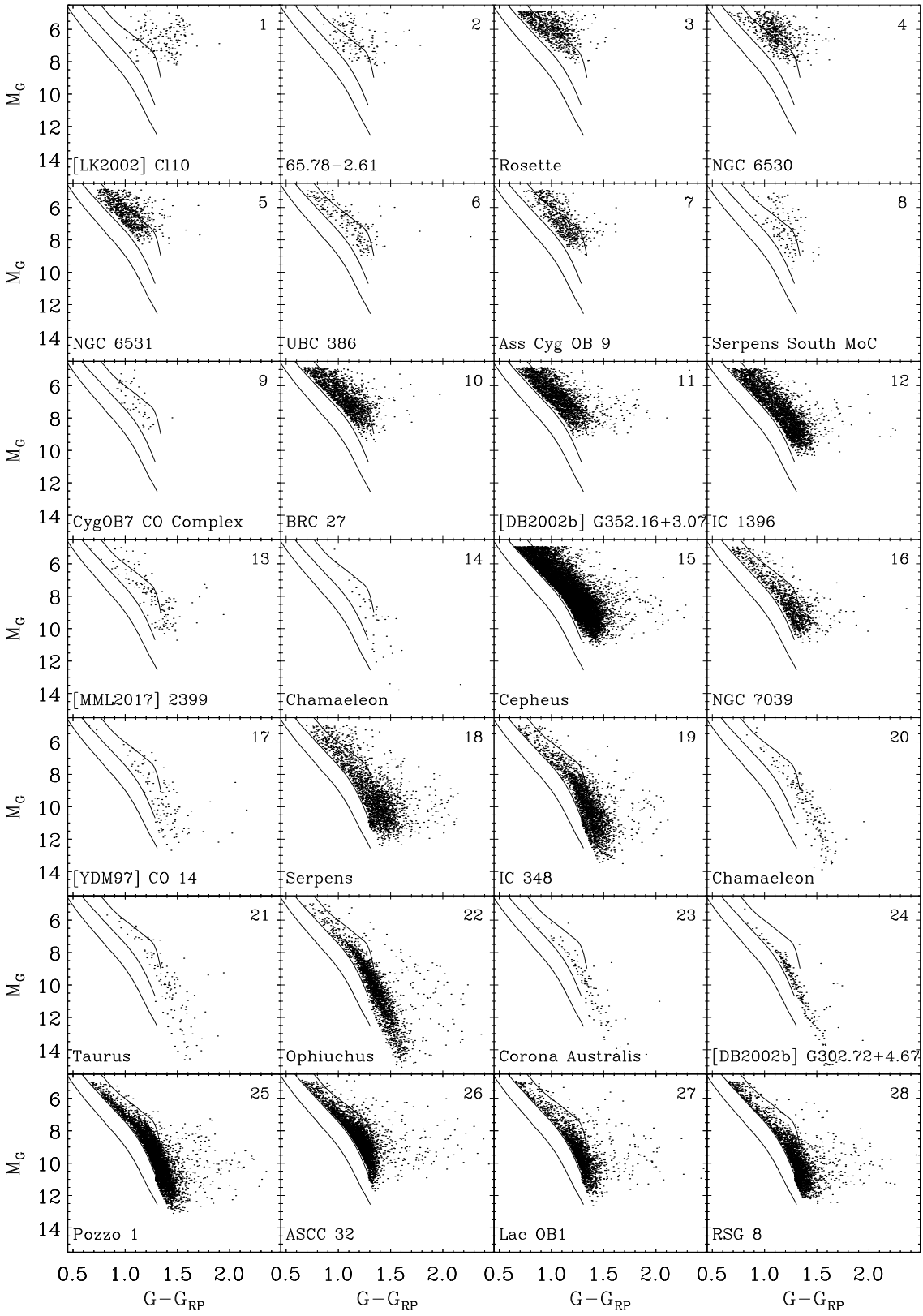

Fig. 2. CAMD of YSOs identified in clusters with ages $t \lesssim 10$ Myr included in the template data set. Black solid lines are the theoretical solar metallicity Pisa isochrones of 1, 10, and 100 Myr isochrones (*from right to left*). The number on the top right edge of each panel is the flag assigned to each cluster.

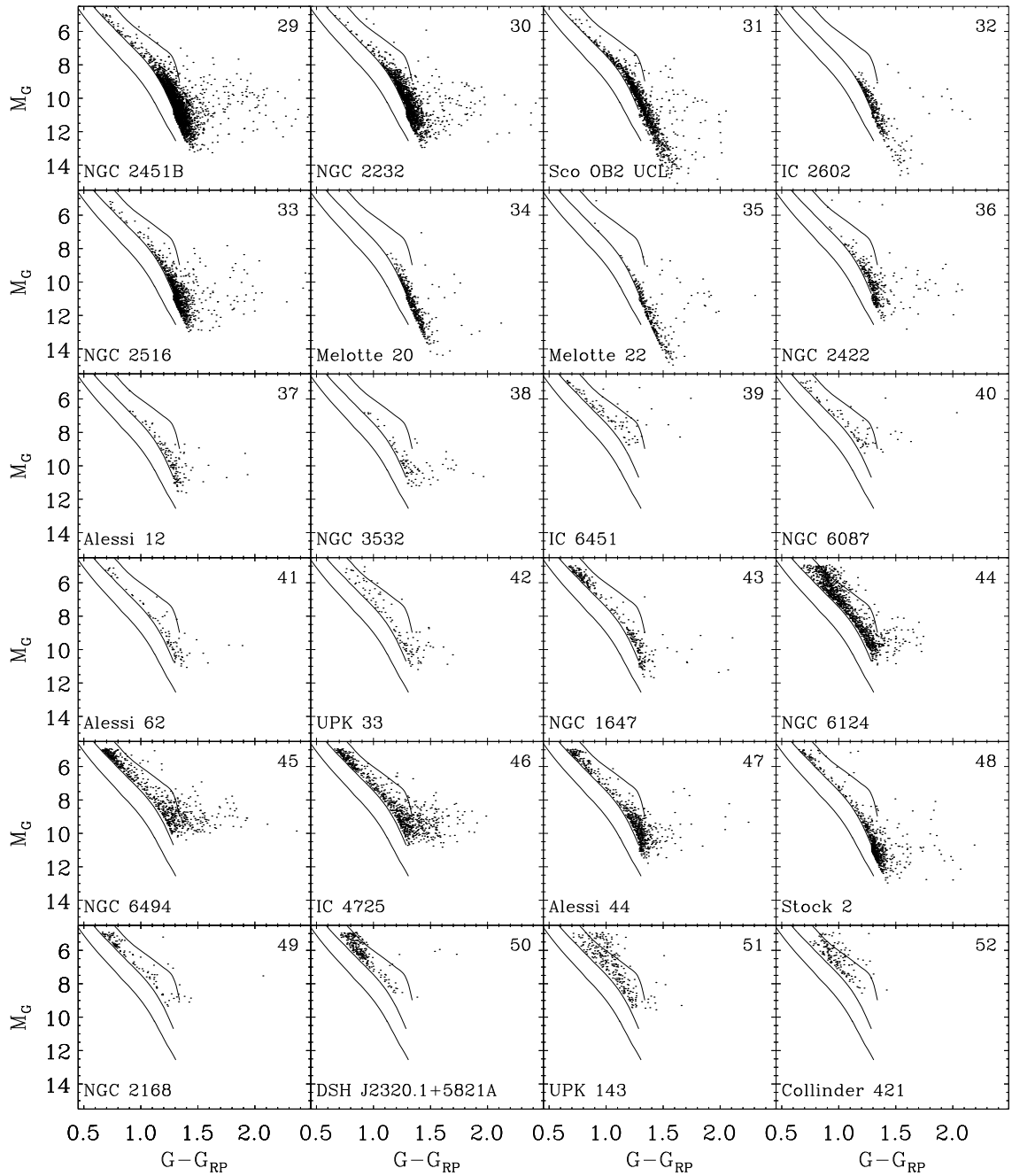

Fig. 3. CAMD of clusters with ages 10 Myr $\lesssim t \lesssim$ 100 Myr, flagged from 29 to 36, and with ages $t \gtrsim$ 100 Myr, flagged from 37 to 52, included in the template data set. Black solid lines are as in Fig. 2. The number on the top right edge of each panel is the flag assigned to the clusters.

those of giant stars and where M_G is nearly constant. Since most of these peculiar clusters are in the direction of the Galactic centre, we infer that they correspond to very distant giants for which *Gaia* EDR3 parallaxes are systematically wrong due to the strong effects of crowding and high extinction in the direction of the Galactic centre. To separate these aggregates from SFRs or stellar clusters, we included a further 27 cases of these peculiar aggregates (flagged from −27 to −1, with a median M_G from 7.6 to 15.8), covering their observed magnitude values.

According to the known ages of the clusters of the template data set, we defined the three age bins, $t \lesssim 10$ Myr, $10 \lesssim t/\,\mathrm{Myr} \lesssim 100$, and $t \gtrsim 100$ Myr, including the clusters with flags in the [1, 28], [29, 36], and [37, 52] ranges, respectively. Then, we used a python implementation of the 2D version of the Kolmogorov–Smirnov (KS) test[6], developed by Peacock (1983) and generalised by Fasano & Franceschini (1987), to identify the most similar amongst the chosen template clusters in the CAMD for each of the 7 323 clusters; that is, the one for which the KS statistic is lowest.

The procedure is not intended to derive any best fitting parameter, but its aim is to only assign a flag to each cluster and

[6] Available at https://github.com/syrte/ndtest

Table 2. Results of the cluster age classification.

Classification	# Stars	# clusters	Flag
$t \lesssim 10$ Myr	124 440	354	[1, 28]
$10 \lesssim t/$ Myr $\lesssim 100$	65 863	322	[29, 36]
$t \gtrsim 100$ Myr	43 936	524	[37, 52]
Phot. unphysical aggregates	68 491	250	[−27, −1]
Unclassified	147 119	5 887	

then a "coarse" age range to which it belongs. At the end, we selected only the 1 450 clusters with more than 20 members (corresponding to 302 730 objects), for which the KS test statistic is <0.2.

In conclusion, we classified 124 440 candidate YSOs that belong to 354 structures with $t \lesssim 10$ Myr, distributed within $\lesssim$1.5 kpc. From now on, we indicate these structures as SFRs, meaning regions that can include at least one very young cluster and mostly consistent YSOs with $t \lesssim 10$ Myr. In addition, we classified 65 863 low-mass members of 322 stellar clusters, mainly located within ∼500 pc and with ages of 10 Myr $\lesssim t \lesssim$ 100 Myr, and, finally, 43 936 members of 524 clusters with $t \gtrsim 100$ Myr. The objects that belong to photometrically unphysical aggregates are 68 491. The results are summarised in Table 2. From our catalogue, we reject all clusters with ages of $t \gtrsim$ 100 Myr; the photometrically unphysical aggregates and those that remain unclassified are mainly poorly populated with a CAMD that does not allow us to properly classify them.

Star forming regions and stellar clusters with ages of $t \lesssim 100$ Myr are listed in Table 3, while cluster members are given in Table 4[7]. Most of the clusters listed in the table are very extended complex regions including several sub-clusters known in the literature, merged here within single structures. Since the aim here is to detect these Galactic young structures, the literature cluster names given in Table 3, mainly taken from Cantat-Gaudin & Anders (2020) or Zucker et al. (2020) or from Simbad, are only indicative of the region.

5. Results

5.1. Photometric completeness

Within the magnitude range explored in this work and assuming the restrictions on *Gaia* data defined in Sect. 2, the photometric cluster completeness for clusters with $t \lesssim 10$ Myr is expected to be near 100% for non-embedded YSOs. This is because, as shown in Fig. 1, all members detectable in this age range and in the optical bands are expected to lie in the selected photometric region.

Nevertheless, the adopted restriction, RUWE < 1.4, introduced a bias in the selection of multiple members of the SFRs. To estimate the fraction of missed binary members with the *Gaia*-based selection used in this paper, we used the Taurus-Auriga binary-star list by Kraus et al. (2012) as a reference. Details about the comparison of this list with our catalogue and *Gaia* EDR3 data are given in Appendix B. This comparison shows that, due to the RUWE restriction, in SFRs at distances similar to Taurus-Auriga, we have lost about 72% of their binary populations. Assuming a binary frequency of ∼50% (Mathieu 1994), a loss of ∼35% of PMS members can be expected. However, at large distances, the projected binary motions become smaller, and therefore we expect a less significant binary member loss for the farther-out SFRs.

For clusters with ages of $t \gtrsim 10$ Myr, the cluster completeness decreases with ages and strongly depends on the cluster distance. In fact, clusters with 10 Myr $\lesssim t \lesssim 100$ Myr (indexed from 29 to 36), are mainly in the solar neighbourhood ($d <$ 500 pc). For these clusters, even though we are not able to detect the entire cluster population, we are, however, able to detect part of the very-low-mass tail component. The fraction of the detected very-low-mass tail component decreases with age, and, for clusters with $t \gtrsim 100$ Myr (indexed from 37 to 52), mainly concentrated at $d \gtrsim 500$ pc, the completeness is very low. The latter were discarded from our final catalogue since they include only a small fraction of the cluster members and are not in the age range of interest for this work. Clusters with ages of 10 Myr $\lesssim t \lesssim$100 Myr are included in our catalogue since the age transition to the clusters with $t \lesssim$10 Myr is not sharply defined, and, in addition, there are structures such as Sco OB2 that include clusters in both age ranges that very likely belong to correlated star forming processes.

5.2. Spatial distribution

Figure 4 shows the maps of the 124 440 YSOs associated with the 354 SFRs with ages $t \lesssim 10$ Myr, while Fig. 5 shows the maps of the 65 863 stars associated with the stellar clusters with ages of 10 Myr $\lesssim t \lesssim 100$ Myr. Each map has been obtained as a 2D histogram smoothed with a Gaussian kernel at 3σ, adopting a pixel size of 3 pc × 3 pc.

Most of the overdensities in Fig. 4 are associated with known SFRs, some of which are labelled in the figure. With the exception of those within 200–300 pc, all clusters present a radial, elongated shape, tracing the increasing uncertainties in the distances.

The clusters with ages of 10 Myr $\lesssim t \lesssim 100$ Myr are mainly limited within ∼600 pc (see Fig. 5) and show a much more diffuse spatial distribution. Very rich clusters such as NGC 2232, NGC 2451B, Gamma Velorum, NGC 2547, NGC 2516, and Alessi 5 at distance of ∼400 pc, seem to belong to a common giant complex, mostly lying in the third Galactic quadrant.

5.3. Literature comparison

In this section, we present the comparison of our results with those previously obtained in the literature for two particular regions, Sco OB2 and NGC 2264. These comparisons were used to estimate our completeness and the contamination level, at least when the completeness of the comparison sample enabled us to do so. We note that we considered each of the merged clusters as a unique ensemble. A detailed sub-clustering analysis, with the identification of possible sub-structures with age-gradient or kinematic sub-clusters is deferred to a future paper. A detailed comparison with the literature for other SFRs is presented in Appendix C, where we also compare the whole catalogue with other all-sky catalogues, mainly derived with *Gaia* DR2 data.

5.3.1. The Sco-OB2 association

The Sco-Cen or Sco-OB2 association is a very extended SFR (∼120° × 60°) quite close to the Sun ($d \sim 150$ pc), which, in the last years, has been the subject of several studies focussed on the low-mass population. By exploiting available all-sky surveys, these studies finally allowed us to study the entire region and its

[7] Tables 3 and 4 are only available in electronic form.

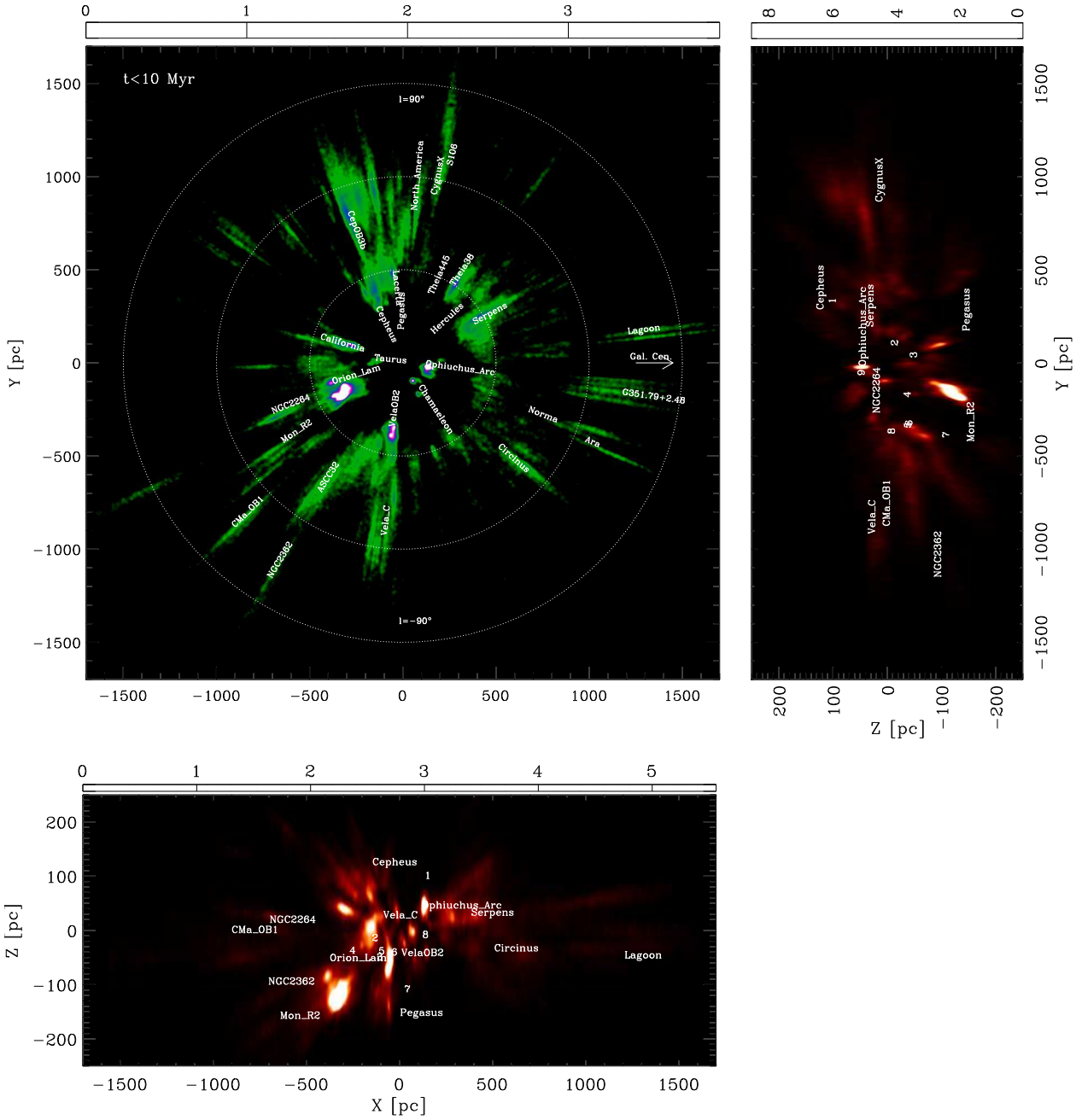

Fig. 4. Density map of three orientations of the GP of YSOs associated with SFRs with ages of $t \lesssim 10$ Myr. In the *upper left panel*, the Sun is at (0, 0), the x-axis is directed towards the Galactic centre, and the y-axis is towards the direction of the Galactic rotation. Dashed white circles are drawn at distance steps of 500 pc. In the *upper right* and *lower left panels*, the z-axis is perpendicular to the GP. Colour bars indicate the surface densities, that is, the number of stars per bin and per pc^2. Some known SFRs are indicated. The numbers from 1 to 9 in the *upper right* and *bottom panels* indicate the position of the clusters indicated in Fig. 5 (*upper left panel*).

complexity (e.g. Zari et al. 2018; Damiani et al. 2019; Kounkel & Covey 2019; Kerr et al. 2021; Luhman 2022).

We spatially selected the members of this region by considering all stars with $-102° < l < 10°$, $-30° < b < 40°$, and, as assumed in Damiani et al. (2019), a distance of $d < 200$ pc. This gave a total of 9 663 YSOs with ages $t \lesssim 100$ Myr, distributed as in Fig. 6. In the (l, b) plane, the pattern of the YSOs associated with Sco-OB2 is that already found in the literature (e.g. de Zeeuw et al. 1999; Damiani et al. 2019; Kerr et al. 2021). Among the selected objects, 4 232 YSOs have been classified in the $t < 10$ Myr range. 2 472 are concentrated in the upper Sco (US) region. They correspond to the youngest subpopulation of Rho Ophiuchi. Another prominent sub-population, classified in the 10 Myr $\lesssim t \lesssim$100 Myr (flag 31) range, includes 3 741 YSOs falling in the upper Centaurus-Lupus (UCL) and lower Centaurus-Crux (LCC) regions. This represents the first

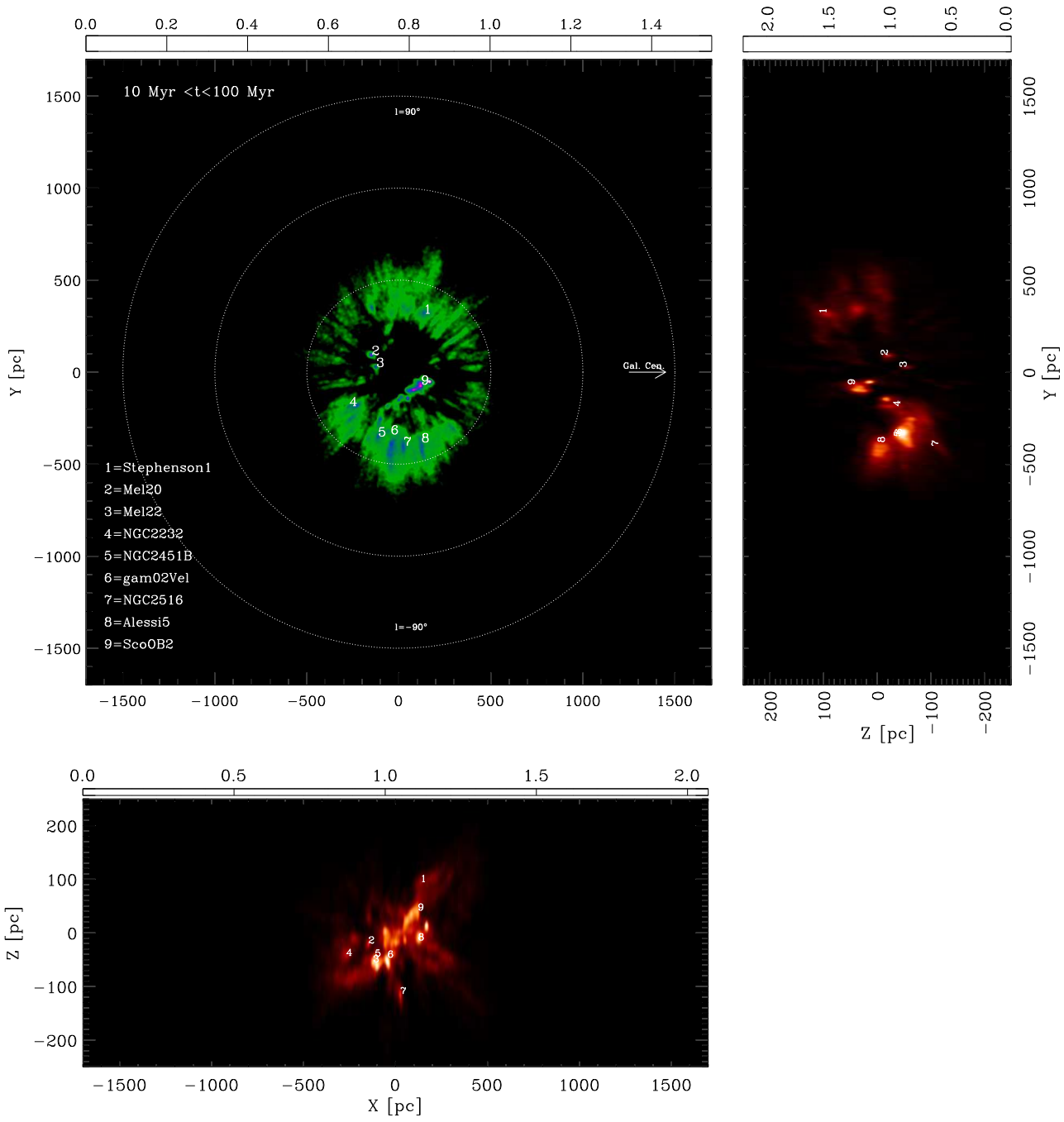

Fig. 5. Same as Fig. 4, but for stars associated with clusters of ages 10 Myr $\lesssim t \lesssim$ 100 Myr.

generation of stars of the Sco OB2 region, in agreement with recent results (e.g. Damiani et al. 2019; Luhman 2022).

Proper motions, parallaxes, and the CAMDs of the different sub-populations detected in the Sco OB2 association are shown in Fig. 7. The proper motions of the YSOs associated with Sco-OB2 show a quite complicated pattern, confirming the complex kinematic structure of this association. The values of parallaxes of YSOs in Sco-OB2 are mostly enclosed between ∼5 mas and ∼10 mas, corresponding to a mean distance of 152 pc and standard deviation σ = 28 pc. Finally, in the CAMD, we can recognise a usual distribution of YSOs in the PMS region. As already noted, the census of the first-generation stars of the Sco OB2 association is likely incomplete since it is expected to lie in the region of the CAMD that has not been considered in this work.

To estimate the completeness level of our census, we compared our list of Sco-OB2 YSOs with the ones recently published by Damiani et al. (2019) and Kerr et al. (2021), based on *Gaia* DR2 data, and by Luhman (2022), based on *Gaia* EDR3 data. To perform these comparisons, we used the *Gaia* identification number of YSOs in our catalogue, retrieved as described in Appendix C.4. We find that there are 6 492 YSOs

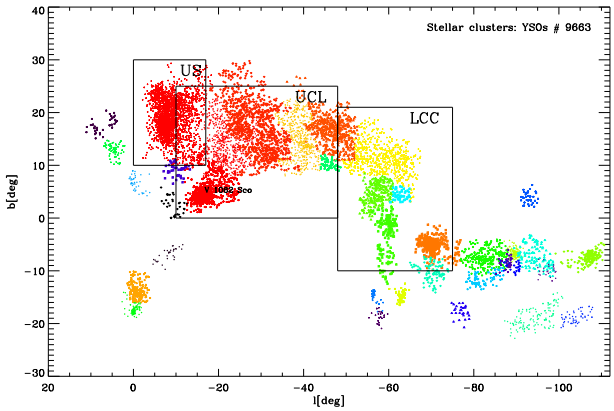

Fig. 6. Spatial distributions in Galactic coordinates of YSOs associated with the Sco-OB2 Association. The de Zeeuw et al. (1999) sub-regions of US, UCL, LCC are shown. The different colours indicate all the different substructures found in this region.

in common with the Damiani et al. (2019) catalogue, which includes a total 10 839 members. That is about 60% of the Damiani et al. (2019) list. Among the 9663 YSOs we selected in the Sco-OB2 association, 7553 fall in the spatial region and magnitude range of $G < 19.5$ covered by Damiani et al. (2019). Therefore, the objects in common are 86% of our sample in the same field. Many of the remaining 1 061 YSOs (14%) not selected by Damiani et al. (2019) show a spatial distribution consistent with that of the other members, and thus we discard the hypothesis that they are contaminants and suggest that they are likely YSOs missed by Damiani et al. (2019; i.e. those based on the less complete *Gaia* DR2 catalogue).

Adopting the same spatial constraints, we retrieved 9 083 objects in the Sco-OB2 region that were selected as candidate YSOs in the Kerr et al. (2021) catalogue, independently from their clustering type of classification. Among these, 5203 are in common with our catalogue, but those classified as YSOs are 5109; that is, ~56.2% of the Kerr et al. (2021) sample[8].

The Luhman (2022) catalogue includes a total of 10 509 YSOs; however, to be consistent with our selection, we selected those with $M_G > 5$, $7.5 < G < 20.5$, RUWE < 1.4 (7925 in total). Using the *Gaia* EDR3 ID and considering the 7408 counterparts falling in the region covered by Luhman (2022), we found that 6341 YSOs are in common with our catalogue, representing 80% of the Luhman (2022) catalogue and 85.6% of our list of YSOs in the Sco Cen.

These percentages cannot be used to accurately estimate our level of completeness or contamination, since the catalogues were obtained starting from different initial constraints, both for the photometric and the astrometric selection, which can inevitably introduce several biases. However, these comparisons are useful to confirm membership for ~85% of the selected members. The remaining 1067 objects not retrieved by Luhman (2022) but selected by us as YSOs show a spatial distribution consistent with that of the other members with two strong concentrations of them in the US region and around V 1062 Sco. We therefore conclude that they are genuine members rather than contaminants, which were likely missed by Luhman (2022) in the photometric selection based on the $G_{BP} - G_{RP}$ colours.

[8] Using the *Gaia* DR2 number, we cross-matched the Kerr et al. (2021) and Damiani et al. (2019) lists and found 6423 objects in common.

5.3.2. The Monoceros OB1/NGC 2264 complex and the Rosette nebulae

Another well-studied region that we used to test our results is the cluster NGC 2264 in the Monoceros OB1 complex. This relatively compact and close (~720 pc from the Sun) SFR, devoid of background and foreground emission, has been the subject of many detailed studies, including, for example, X-ray observations (Flaccomio et al. 2006), optical and near-IR analysis of its low-mass population (Venuti et al. 2019), and coordinated synoptic investigations with optical and IR light curves with CoRot and Spitzer (Cody et al. 2014). Flaccomio et al. (in prep.) compiles the most complete data set of NGC 2264, based both on all-sky surveys (*Gaia* EDR3, 2MASS, VPHAS) and dedicated observations falling in the $98.93° < RA < 101.47°$ and $8.45° < Dec < 10.95°$ regions. The young structure we identified in this field includes a total of 1 916 YSOs, but only 1 062 of them (~55%) fall in the region investigated by Flaccomio et al. (in prep.). The remaining YSOs are in part (404 YSOs) concentrated in the region corresponding to the cluster IC 446, while a further unknown group of 450 YSOs are sparsely distributed in the southern region of NGC 2264. As shown in Fig. 8, a sub-group of the latter form a visual bridge along a filamentary structure, which is clearly visible in the IR IRIS image, down to the location corresponding to the more distant Rosette nebula, which is located at ~1.5 kpc and hosts the SFR NGC 2244. Thus, our finding is that the known cluster NGC 2264 actually belongs to a structure larger than the ~2° × ~2° region, typically considered in the literature for this SFR. The mean distance of YSOs associated with the complex NGC2264-IC 446 is 731.86 ± 95.5 pc, even though the proper motion distributions of the three subgroups suggest they share similar but not equal values[9].

In the same region, we identified a further five sub-structures with distance >0.5 kpc[10], the most populated being the cluster in the CMa OB1 association, centred around RA = 106.3° Dec = –11.47°. It is found at a distance of 1250 ± 162 pc, is associated with the reflection nebula IC 2177, and includes 1709 YSOs. In addition, we identified the cluster NGC 2244, which includes 810 YSOs, is centred at RA = 98.3° Dec = 4.9°, and is at a distance of 1580 ± 199 pc. We also identified the cluster associated with Mon R2, which is at a distance of 897 ± 112 pc and includes 1272 YSOs. In addition, we detected the cluster indicated in Cantat-Gaudin & Anders (2020) as UPK 436 with 620 members and a minor sparse cluster in the region of CMa OB1 located at 807 pc. Figure 9 shows the PM, parallaxes, and CAMD of all these sub-structures, where it is clearly visible that they are spatially and kinematically uncorrelated, while in the PMS region of the CAMD they are indistinguishable since they consist of similarly aged stars.

The membership defined in Flaccomio et al. (in prep.) includes two confidence levels. One is based on the combination of several criteria derived by dedicated X-ray, spectroscopic, and IR observations, including 2263 confirmed members (sample C). Moreover, the fraction of false positives is negligible. Another list (sample C-Wide) is based exclusively on all-sky surveys and includes 1542 YSOs. The membership was deduced by a smaller number of criteria, and thus the number of false positives is expected to be higher. We find that 972 (960) YSOs of the sample C (sample C-wide) are in common with our list of YSOs in the NGC 2264 region, corresponding to a fraction of 43%

[9] A detailed kinematic analysis of these sub-regions is beyond the aims of this work.

[10] This limit was adopted to avoid the Orion sub-structures.

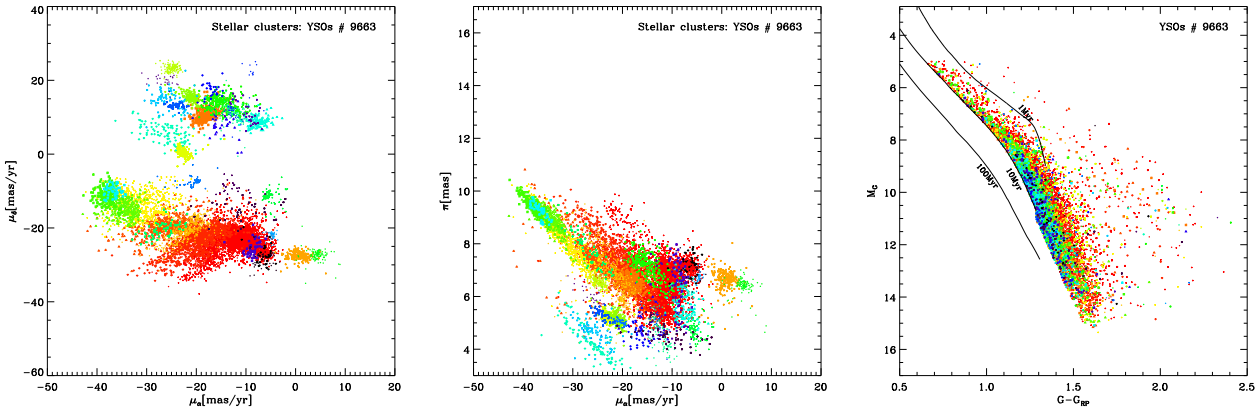

Fig. 7. Proper motions in RA and Dec, parallaxes, and CAMDs of the groups of YSOs associated with the Sco OB2 association. The symbol colours are as in Fig. 6. Three representative solar metallicity isochrones from the Pisa models are also shown as solid lines in the *right panels*.

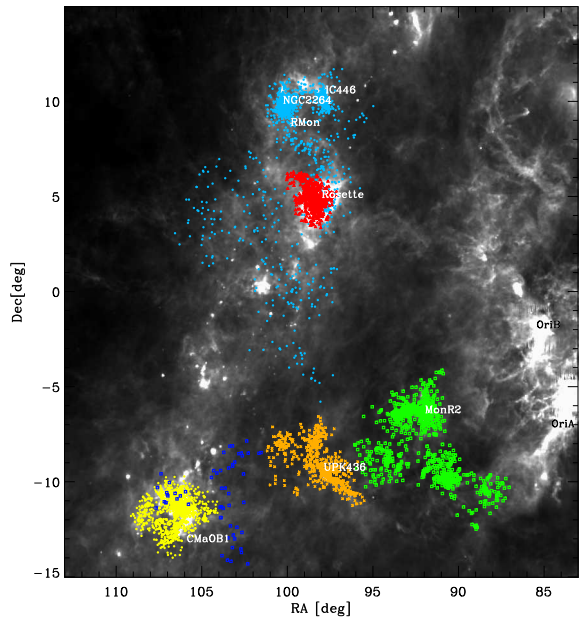

Fig. 8. Spatial distribution in equatorial coordinates of the YSOs associated with NGC 2264, NGC 2244, Mon R2, CMA OB1, and UPK 436. YSOs are overplotted on an IRIS 100 μm image (Miville-Deschênes & Lagache 2005). For clarity, Orion members have not been plotted.

(62%) with respect to the Flaccomio sample. These fractions are considered here as indicators of our level of completeness of the entire SFR population. However, these results are strongly conditioned by the starting photometric selection ($M_G > 5$) and the restrictions on the *Gaia* EDR3 data that we adopted in this work. In addition, the Flaccomio et al. (in prep.) sample C includes 497 of the 2263 confirmed members that do not have a Gaia counterpart.

To estimate the efficiency of our method in recovering YSOs, we considered the members selected by Flaccomio et al. (in prep.) with a Gaia counterpart, which fall in the photometric region considered in this work and are compliant with our initial data restrictions (i.e. RUWE < 1.4 and parallax relative error <0.2). Adopting this sample, the fraction of the YSOs we selected in common with the Flaccomio et al. membership is 95–96%, considering both the samples C and C-Wide. We note that this is the efficiency of our clustering method but is not the efficiency of the *Gaia* data. In fact, if for the two lists we consider the members falling in the same photometric region but we do not consider the restrictions in RUWE and in the parallax error, the fraction of YSOs in common is 72% for sample C and 77% for sample C-wide. This suggests that we missed 23–28% of genuine YSOs due to remaining issues in the *Gaia* data, at least in the current *Gaia* EDR3 release.

Finally, we find that among the 1052 YSOs we selected in the NGC 2264 region, a total of 1034 are included in the list of objects collected by Flaccomio et al., but 62 of them are not members in the more complete and less contaminated sample C. This means that about 92% (i.e. (1034–62)/1052) of the YSOs we selected are confirmed members. Hence, we conclude that the contamination level of the sample that we selected is ~8%.

For comparison, Kounkel & Covey (2019) found 637 YSOs belonging to the clusters named as Theia 41 and 189 in the same region, with 548 and 89 objects, respectively. Of them, 420 (about 66%) are in common with our list.

6. Discussion

In the previous sections, we describe how overdensities in the 5D parameter space (l, b, π, $\mu_\alpha*$, μ_δ) were identified, starting from a photometrically selected sample that covers the expected PMS region of YSOs with ages $t < 10$ Myr. Since no attempt has been made to correct for interstellar reddening, the starting sample was also contaminated by older reddened stars. Another possible reason for the contamination of older stars is derived from the adopted strategy to select the starting sample in the M_G versus $G - G_{RP}$ plane, where the sensitivity to stellar ages of the available isochrones is quite low for the low-mass population. In fact, for faint and very-low-mass stars, isochrones become closer and closer for ages over about 50–100 Myr, and, consequently, it is difficult to separate young populations from older ones. As a result, the DBSCAN clustering algorithm, adopted to resolve spatially concentrated and/or co-moving stellar populations located at the same distance, can also select clusters older than 10 Myr.

A pattern match procedure has been adopted to disentangle SFRs and young clusters from older and photometrically unphysical clusters. We found 354 SFRs with ages of $t \lesssim 10$ Myr and 322 young clusters with ages of approximately 10–100 Myr. We