Gaia Data Release 3

Validating the classification of variable young stellar object candidates[*]

Gábor Marton[1,2] [ORCID], Péter Ábrahám[1,2,3] [ORCID], Lorenzo Rimoldini[5], Marc Audard[5,6] [ORCID], Mária Kun[1,2] [ORCID], Zsófia Nagy[1,2] [ORCID], Ágnes Kóspál[1,3,4,2] [ORCID], László Szabados[1,2] [ORCID], Berry Holl[5,6], Panagiotis Gavras[7], Nami Mowlavi[5,6] [ORCID], Krzysztof Nienartowicz[8], Grégory Jevardat de Fombelle[5], Isabelle Lecoeur-Taïbi[5], Lea Karbevska[5,9], Pedro Garcia Lario[10], and Laurent Eyer[6]

1   Konkoly Observatory, Research Centre for Astronomy and Earth Sciences, Eötvös Loránd Research Network (ELKH), Konkoly Thege Miklós Út 15-17, 1121 Budapest, Hungary
2   CSFK, MTA Centre of Excellence, Budapest Konkoly Thege Miklós Út 15-17, 1121, Hungary
    e-mail: marton.gabor@csfk.org
3   ELTE Eötvös Loránd University, Institute of Physics, Pázmány Péter sétány 1/A, 1117 Budapest, Hungary
4   Max Planck Institute for Astronomy, Königstuhl 17, 69117 Heidelberg, Germany
5   Department of Astronomy, University of Geneva, Chemin d'Ecogia 16, 1290 Versoix, Switzerland
6   Department of Astronomy, University of Geneva, Chemin Pegasi 51, 1290 Versoix, Switzerland
7   RHEA for European Space Agency (ESA), Camino bajo del Castillo, s/n, Urbanizacion Villafranca del Castillo, Villanueva de la Cañada, 28692 Madrid, Spain
8   Sednai Sàrl, Geneva, Switzerland
9   Université de Caen Normandie, Côte de Nacre Boulevard Maréchal Juin, 14032 Caen, France
10 European Space Agency (ESA), European Space Astronomy Centre (ESAC), Camino bajo del Castillo, s/n, Urbanizacion Villafranca del Castillo, Villanueva de la Cañada, 28692 Madrid, Spain

ABSTRACT

Context. The *Gaia* third Data Release (DR3) presents the first catalogue of full-sky variable young stellar object (YSO) candidates observed by the *Gaia* space telescope during the initial 34 months of science operations.
Aims. Numerous types of variable stars were classified using photometric data collected by *Gaia*. One of the new classes presented in the *Gaia* DR3 is the class of YSOs showing brightness variability. We analysed 79 375 sources classified as YSO candidates in order to validate their young nature and investigate the completeness and purity of the sample.
Methods. We cross-matched the *Gaia* DR3 YSO sample with numerous catalogues from the literature, including YSO catalogues based on optical and infrared data, as well as catalogues of extragalactic sources and Galactic variable stars. YSO catalogues were used to quantify the completeness of the *Gaia* DR3 YSO sample, while others were inspected to calculate the contamination.
Results. Among the 79 375 potential YSO candidates published in the *Gaia* DR3 variable star catalogue, the majority of these objects are distributed along the line of sight of well-known star forming regions and the Galactic midplane. We find that the upper limit of contamination is 26.7%, depending on the external catalogue used for the estimation, but find an average of ~10% in general, while the completeness is at the percent level, taking into account that the *Gaia* DR3 YSO sample is based on sources that showed significant variability during the data-collection period. The number of sources in our sample that had not previously been catalogued as YSO candidates is ~40 000 objects.

Key words. catalogs – stars: early-type – Galaxy: stellar content – stars: formation – stars: pre-main sequence – stars: statistics

1. Introduction

Gaia (Gaia Collaboration 2016) is a cornerstone mission of the European Space Agency. The mission is not only the most ambitious stellar (and extragalactic) astrometric project ever, but also one of the best transient discovery machines today. While other surveys, such as YSOVAR (Morales-Calderón et al. 2011), ASAS-SN (Shappee et al. 2014; Kochanek et al. 2017), and ZTF (Masci et al. 2019), provide photometric data and light curves for millions of sources, including young stellar objects

(YSOs), *Gaia* has the advantage of observing the whole sky. It collects photometric observations of some 1.8 billion stars down to a faint limit of 20.7 mag in the *G* band and obtains low-resolution spectra down to ~19 mag for an average of ~80 epochs during the nominal five-year mission (which was completed and then extended until the end of 2022; hopefully it will be extended to the end of 2025 but this remains to be confirmed), although the cadence is highly dependent on the scanning law (Gaia Collaboration 2016). *Gaia* started monitoring the whole sky on 25 July 2014 and collects multi-epoch multi-band spectrophotometry and astrometric data for sources crossing its two fields of view (FoVs). A description of the *Gaia* mission (spacecraft, instruments, survey, and measurement principles) as well as the structure and activities of the *Gaia*

---

[*] The KYSO table is only available at the CDS via anonymous ftp to cdsarc.cds.unistra.fr (130.79.128.5) or via https://cdsarc.cds.unistra.fr/viz-bin/cat/J/A+A/674/A21

Data Processing and Analysis Consortium (DPAC) can be found in Gaia Collaboration (2016). The *Gaia* Data Release 3 (DR3) became public on 13 June 2022, and among many different products that have the potential to increase our fundamental understanding of the Galaxy it provides photometry in three passbands (*Gaia G*, G_{BP}, and G_{RP}), five-parameter astrometry, and radial velocities (RVs) collected over the initial 34 months of observations; these are the most relevant quantities for our case. A summary of the *Gaia* DR3 contents and survey properties is provided in Gaia Collaboration (2023).

One important research area, where collecting large amounts of data brings fundamental new results, is star formation. How the Sun – and stars in general – was born is identified as one of the most important questions of modern astronomy. Throughout their early evolution, stars show different features in their spectral energy distribution (SED). Initially, protostars are deeply embedded in their parental clouds, surrounded by dense dust and gas envelopes, which allow their detection only at sub-millimetre (mm) and far-infrared (FIR) wavelengths. At later stages when the envelope dissolves, young stars become apparent at optical wavelengths as well, which allows us to detect them with the *Gaia* space telescope. However, at these evolutionary stages they still have circumstellar matter, their protoplanetary discs are still evolving, the stellar magnetosphere drives accretion columns, and accretion shocks are seen, all of which lead to variability in their emitted light, which can be studied to infer the underlying physics of star and planet formation.

Gaia scans the sky repeatedly with successive observations following an irregular sequence in time – based on the scanning law – on a timescale from hours to months, and with a varying number of FoV transits depending on the celestial location (e.g. see Appendix A of Eyer et al. 2017). This allows the construction of light curves, which can be analysed for variability. More details about the general variability processing in *Gaia* DR3 are published in Eyer et al. (2023). Variability in YSOs occurs on timescales that span a wide range and depend on the physical processes (see Hillenbrand & Findeisen 2015; Fischer et al. 2022). Circumbinary disc occultations last for ~100 days and can cause dimmings of up to 4 mag. The EX Lup-type outbursts occur on a timescale of couple of hundred days and can also cause brightenings of 2–4 mag, while FU Ori-type outbursts last even longer and cause brightenings of even greater magnitude (see e.g., Audard et al. 2014). These latter are therefore also a potential target of *Gaia* observations.

In this study, we focus on a new class introduced in the *Gaia* DR3 variability classification, the class of YSOs, for which the classification process resulted in a list of 79 375 candidates. This output of the classification process was used in the current study. Our goal was threefold: (1) to validate their young nature, (2) to put constraints on the completeness of the catalogue, and (3) to check the purity and estimate the level of contamination. Having a large and reliable catalogue of variable YSOs can help further studies by providing a list of potential targets for more detailed analysis. Such analyses can be used to infer the underlying physics of the star and disc evolution, and their interaction.

2. Data

2.1. Gaia data

Gaia DR3 represents a significant improvement over *Gaia* DR2. The parallax precisions increased by 30 percent and proper motion precision improved by a factor of 2. Also, the systematic errors on the astrometry are lower by 30%–40% for the parallaxes, while those on the proper motions are lower by a factor ~2.5. The longer temporal baseline of the observations also resulted in increased precision of the photometry and much better homogeneity across colour, magnitude, and celestial position. While *Gaia* measures the position and brightness of the objects in its FoV, many parameters are provided that describe the quality of the data. Through the variability processing pipeline (Eyer et al. 2023), many other parameters are derived, which are helpful during the identification of variable sources and their verification.

One of the *Gaia* DR3 products is the supervised classification of variable sources (Rimoldini et al. 2023). Supervised classification was applied to classify several variability types using per-FoV epoch photometry in the three *Gaia* bands. Among the 60 000 training sources, 5148 were YSOs and included subtypes such as T Tauri stars (classical, weak-lined, intermediate mass of types F to early G, and late G to early K types), FU Orionis, Herbig Ae/Be, and UX Orionis type stars. The time series were characterised by about two dozen features (described in Sect. 10.3.3 of Rimoldini et al. 2022), which were used to train classifiers employing Random Forest (Breiman 2001) and XGBoost (Chen & Guestrin 2016) algorithms.

The results of this supervised classification were verified in a post-processing phase, where the parameter space was analysed in order to maximise the purity of the sample and minimise its contamination. Several cuts were applied to variability parameters and astrometric values, such as parallax errors and proper motion, to exclude sources that have distances that are too small or too large to be taken into account as potential young stars (see Rimoldini et al. 2023, for details).

2.2. KYSO catalogue

To help the identification of YSOs among the *Gaia* DR3 variable stars, we created a catalogue, the Konkoly Optical YSO (hereafter KYSO) catalogue (available at the CDS), part of which served as a training sample for the supervised classification of *Gaia* DR3 variable stars (Rimoldini et al. 2023) and was also used in the post-processing verification phase. The KYSO contains nearly 12 000 objects, which are carefully selected young stars identified in the optical domain, and their young nature is mostly confirmed using spectroscopic data. More details about how the KYSO catalogue was compiled are described in Appendix A. The KYSO catalogue is published as a Vizier table so that future classification studies can benefit from it. The KYSO catalogue was also used as a first step in our validation process and a detailed analysis was performed. We note that the KYSO catalogue has evolved since it was provided for training purposes for the supervised classification, as new sources have been added, and less reliable YSOs removed in order to increase the reliability of the catalogue.

2.3. Other catalogues

We used several existing YSO catalogues from the literature based on optical and infrared surveys and observations, including data from *Gaia* DR2, infrared data from 2MASS, and the *Spitzer* and WISE surveys. These catalogues are the Marton et al. (2019) probabilistic YSO catalogue, based on *Gaia* DR2 optical data, 2MASS near-infrared (NIR) photometry, and WISE mid-infrared (MIR) observations from the AllWISE catalogue, the SPICY (*Spitzer*/IRAC Candidate YSO) catalogue of Kuhn et al. (2021) based on MIR observations of *Spitzer,* and the Großschedl et al. (2018) study of Orion A based on the ESO–

VISTA NIR survey. While the *Gaia* DR3 YSO candidate sample is based on optical variability, the listed catalogues from the literature used the IR excess in the SED as a signature of the young nature. We also cross-matched the *Gaia* DR3 YSO candidates with the SIMBAD database (Wenger et al. 2000) and other large catalogues listing extragalactic sources and other types of variable stars identified in the optical domain, as these sources can show colours and light-curve features that are similar to those of YSOs. Based on the results of the cross-matches we calculated an estimated completeness and purity for the *Gaia* DR3 YSO candidates.

It is important to note that the *Gaia* observations were of different angular resolution and sensitivity when compared to other observations in the literature. The 2MASS survey (Skrutskie et al. 2006) collected NIR data in the J (1.2 μm), H (1.65 μm), and K_s (2.16 μm) bands; the system PSF was ∼2″.5, and the average limiting magnitude was ∼14. The WISE mission (Wright et al. 2010) observed the whole sky with an angular resolution 6″.1, 6″.4, 6″.5, and 12″.0 at 3.4, 4.6, 12, and 22 μm, respectively. For the WISE data, we used the AllWISE catalogue, which is >95% complete for sources with $W1 < 17.1$ and $W2 < 15.7$ mag. These differences and the apparent motions of several objects make it difficult to simply cross-match such large catalogues and make it challenging to avoid source confusion. In the case of the AllWISE catalogue, 2MASS data are already included, where the position reconstruction was done using bright 2MASS point sources as the astrometric reference. For AllWISE, the proper motion of the reference stars in the 11 yr separating the WISE and 2MASS surveys has been integrated into the solutions to improve the absolute astrometric accuracy[1]. The cross-match of the *Gaia* DR2 and the AllWISE catalogues was done by Marrese et al. (2019). In cases where we used 2MASS and AllWISE data for the validation, we relied on these catalogues, matched the *Gaia* DR3 YSO sources to the DR2 positions, and inferred the data from the DR2xAllWISE (including 2MASS) table. In other cases, when the cross-match with such accuracy was not already available, we used the TOPCAT software (Taylor 2005) with a search radius of 1″.

3. Validation

In this section, we consider three different aspects of the validation. In Sect. 3.1, we compare the parameters of the *Gaia* DR3 YSO sample to those of objects listed in different YSO catalogues from the literature in order to validate the young nature of the *Gaia* DR3 objects. These catalogues are based on different domains of the electromagnetic spectrum. The KYSO is based on optical data, the Großschedl et al. (2018) uses NIR data, the Kuhn et al. (2021) is based on MIR data, and finally the Marton et al. (2019) used optical, NIR, and MIR photometry. We also investigate the distance distributions of the *Gaia* DR3 YSOs on the all-sky and in individual star forming regions and compare them to literature values.

In Sect. 3.2, we estimate the completeness of the *Gaia* DR3 YSO candidates by comparing the total number of the sources listed in a given YSO catalogue to the number of *Gaia* DR3 counterparts (meaning *Gaia* actually observed these sources), to the number of counterparts that were part of the classification process (for sufficiently sampled signals that were considered as variable in *Gaia* DR3), and to the number of counterparts existing in the final *Gaia* DR3 YSO sample.

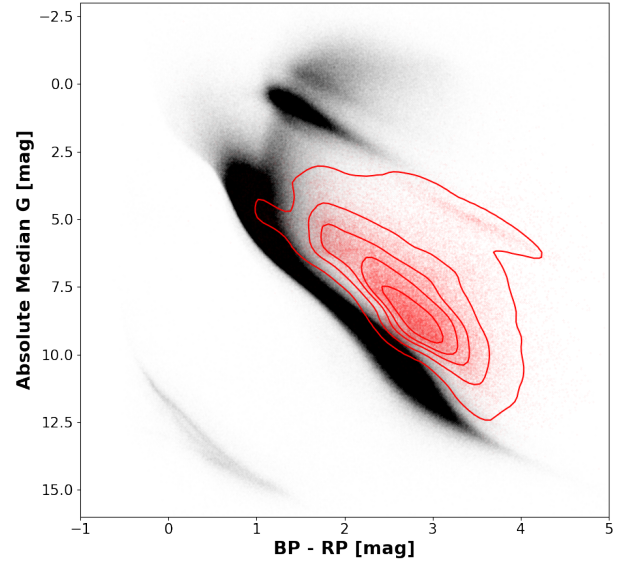[1] https://wise2.ipac.caltech.edu/docs/release/allwise/expsup/sec2_1.html

Fig. 1. Observational HRD of *Gaia* DR3 YSOs (red dots) and reference sources based on 4.2 million *Gaia* objects (black dots) selected based on their highly reliable parallax, sufficient S/N in both BP and RP bands, and the sufficiently high number of data points in their light curves. The *Gaia* DR3 YSOs occupy a specific region above the main sequence and below the giant branch. The contour levels are at 5%, 25%, 45%, 65%, and 85% of the maximum density value. In the comparison with other catalogues, we use only the contours of the DR3 distribution for better visibility of the underlying data points.

In Sect. 3.3, we estimate the contamination level of the *Gaia* DR3 YSO sources by matching them to catalogues listing a specific type of objects, mostly identified in the optical domain to make the comparison reasonable.

3.1. The young nature of the Gaia DR3 YSO candidates

3.1.1. Cross-match with the KYSO catalogue

As a first step of the validation of the *Gaia* DR3 YSO candidate sample, we cross-matched it with the KYSO catalogue. The KYSO catalogue provides a strong bias as the sources included are from various catalogues and from well-known star forming regions, while *Gaia* provides a relatively homogeneous all-sky survey.

In the case of a good classification, one would expect to see that YSOs occupy the same regions on the observational HRD and also on the various colour–colour diagrams. The position of the DR3 YSO candidates compared to a set of reference sources on the observational HRD is shown in Fig. 1. The 2MASS colour-colour diagram for the DR3 YSO candidates is shown in Fig. 2. YSOs are seen in specific regions of the sky, and therefore we expect overdensities in the YSO candidate surface density distribution in the line of sight of the (1) regions where the training sample showed overdensities and (2) the known star forming regions and the Galactic midplane. Figure 3 shows their distribution on the sky in Galactic coordinates, which clearly reflects the expected distribution.

As a next step, we checked the location of the *Gaia* DR3 YSO sources and that of the KYSO objects on the observational Hertzsprung–Russell Diagram (HRD, often referred to as simply the colour–magnitude diagram). We note here that HRDs in this study are not corrected by Galactic extinction or reddening. Figure 4 shows how the KYSOs are distributed in the median $G_{BP} − G_{RP}$ colour versus G band absolute magnitude diagram.
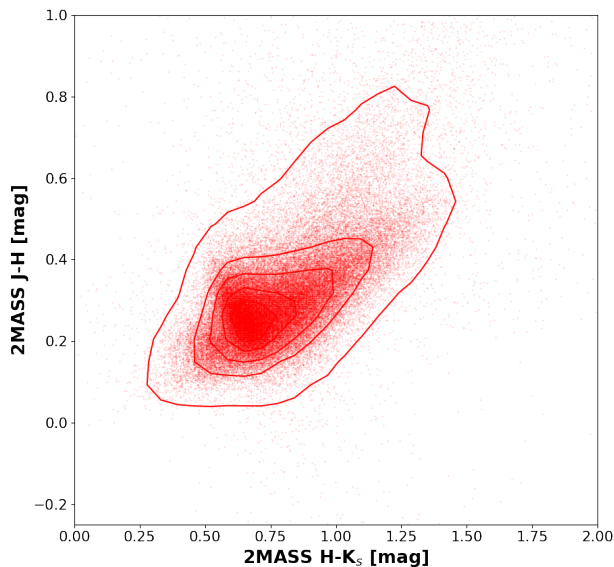
Fig. 2. *Gaia* DR3 YSOs on the 2MASS colour–colour diagram. The contour levels are at 5%, 25%, 45%, 65%, and 85% of the maximum density value. In the comparison with other catalogues, we use only the red contours of the DR3 distribution for better visibility of the underlying data points. The median, mean, standard deviation, and 5% and 95% quantiles of both colours are listed in Table 1.

As explained in Bailer-Jones et al. (2021), a simple inversion of the parallax values does not give a precise distance, because the inversion can lead to bias. Therefore, the G band absolute magnitudes were calculated using their distance values obtained from the Bailer-Jones et al. (2021) distance catalogue instead of the parallax listed in the *Gaia* DR3. In general, YSOs are mostly located above the main sequence, but depending on their mass and evolutionary stage, they can appear in other parts of the HRD as well (see Sect. 3.2.1). The *Gaia* DR3 YSOs are located in a smaller region on the HRD, but highly overlapping with the KYSO catalogue, which shows that the classification process successfully identified those candidates that are located in the same place as the majority of the KYSOs.

There are only a few sources located at the blue side of the HRD. The $G_{BP} - G_{RP}$ colour distribution is shown in Fig. B.1. While the histograms show similarities, it is seen that there are more KYSOs at the blue end of the diagram, where mostly high-mass YSOs and Ae/Be stars, are located. This means that the YSO classification was more sensitive to the redder, fainter, and lower-mass objects, because very few training objects were included below the main sequence, and also because of the strong competition with other classes at the blue end.

We also analysed the *Gaia* G band absolute magnitude distribution of the KYSO sources and that of the sources classified as *Gaia* DR3 YSOs; the results are shown in Fig. B.5. The distribution of KYSO sources is wider than that of the *Gaia* DR3 YSOs, but the two samples show a significant overlap.

To see the IR excess distribution of the KYSOs and the *Gaia* DR3 YSOs, we cross-matched them with the 2MASS catalogue using a 1″ matching radius. Of the 79 375 *Gaia* DR3 YSOs, 76 879 (97%) had a NIR counterpart, while 10 613 (91%) of the 11 665 KYSOs had a match in the 2MASS database.

Figure 5 shows the distributions of the *Gaia* DR3 YSO candidates and the KYSO sources on the 2MASS $J-H$ versus $H-K_s$ colour–colour diagram. Both samples occupy the same region on the diagram, and are mostly located between $0 < J - H < 1.5$ and $0 < H - K_s < 1$.

Further details of the above-mentioned distributions are listed in Table 1, where we list the median and mean values, the standard deviation, and the threshold values below and above which 5% of the samples are located. Figures showing the distributions are found in Appendix B.

We also investigated the distance distribution of the KYSOs and *Gaia* DR3 YSOs based on the values of the median photogeometric distance posterior listed for *Gaia* EDR3 by Bailer-Jones et al. (2021). While the overall distributions of the distance values show strong similarities, the KYSO catalogue shows a small excess of very nearby stars (<100 pc); see Fig. 6.

3.1.2. Cross-match with the Marton et al. (2019) YSO catalogue

In Marton et al. (2019; hereafter M19), 101 million objects were classified from the DR2xAllWISE cross-match table into four main categories (evolved stars (E), main sequence stars (MS), extragalactic sources (EG), and YSOs (Y)) with the help of a Random Forest classifier, and class membership probabilities were assigned to each source. Koenig & Leisawitz (2014) investigated the spurious detections in the WISE $W3$ and $W4$ bands. Because M19 was heavily reliant on the WISE data, the reliability of the $W3$ and $W4$ photometry values was also investigated. Another Random Forest classifier was built to decide whether the $W3$ and $W4$ band detections are spurious or reliable, and a probability value R was given for each detection, as detailed in Sect. 2.7 in Marton et al. (2019). If $R \geq 0.5$, one can assume that the $W3$ and $W4$ band photometry can be used and the LY probability values are used from the M19 sample, which gives the probability that a source is a YSO; in the LY acronym, the letter Y refers to YSO, while the letter L refers to the inclusion of longer wavelength $W3$ and $W4$ bands in the classification. In cases where $R < 0.5$, the SY (where the letter S refers to the shorter wavelength $W1$ and $W2$ bands; in these cases the WISE $W3$ and $W4$ bands were not part of the classification process) probability was taken into account.

Based on the probability values, one can define a threshold above which one feels confident to accept the source as a reliably classified object. For analysing the properties of the *Gaia* DR3 sources and comparing it to the M19, we used YSOs from the M19 catalogue, where $R \geq 0.5$ and $LY \geq 0.95$, or $R < 0.5$ and $SY \geq 0.95$, meaning 259 363 sources in total.

The observational HRD of the M19 and *Gaia* DR3 YSO candidates are shown in Fig. 7. Similarly to the comparison with the KYSOs, the *Gaia* DR3 YSO candidates also show significant overlap with the distribution of the M19 sources, but the M19 distribution shows additional peaks at the bluer and brighter part of the HRD as well as close to the giant branch. The bright-blue tail of the distribution can be explained by the less accurate training sample used in the Marton et al. (2019) study, which allowed more evolved sources to be classified as YSOs with relatively high probability without the long-wavelength $W3$ and $W4$ measurements of the WISE telescope. The peak close to the giant branch is also due to the uncertainty in the distance estimation of the source, and therefore these sources may appear more distant than they actually are. However, as listed in Table 1, 5% of the *Gaia* DR3 YSO candidates are fainter than ~10.3 mag, while in the M19 catalogue this quantile value is ~10.7 mag. The 5% quantile value at the bright end for the *Gaia* DR3 YSO candidates is ~4.4 mag, while for the M19 YSOs, the same threshold is at ~3.6 mag. Therefore, these differences cannot be considered significant.

On the 2MASS $J - H$ versus $H - K_s$ colour–colour diagram (Fig. 8), the *Gaia* DR3 YSOs are located in the same
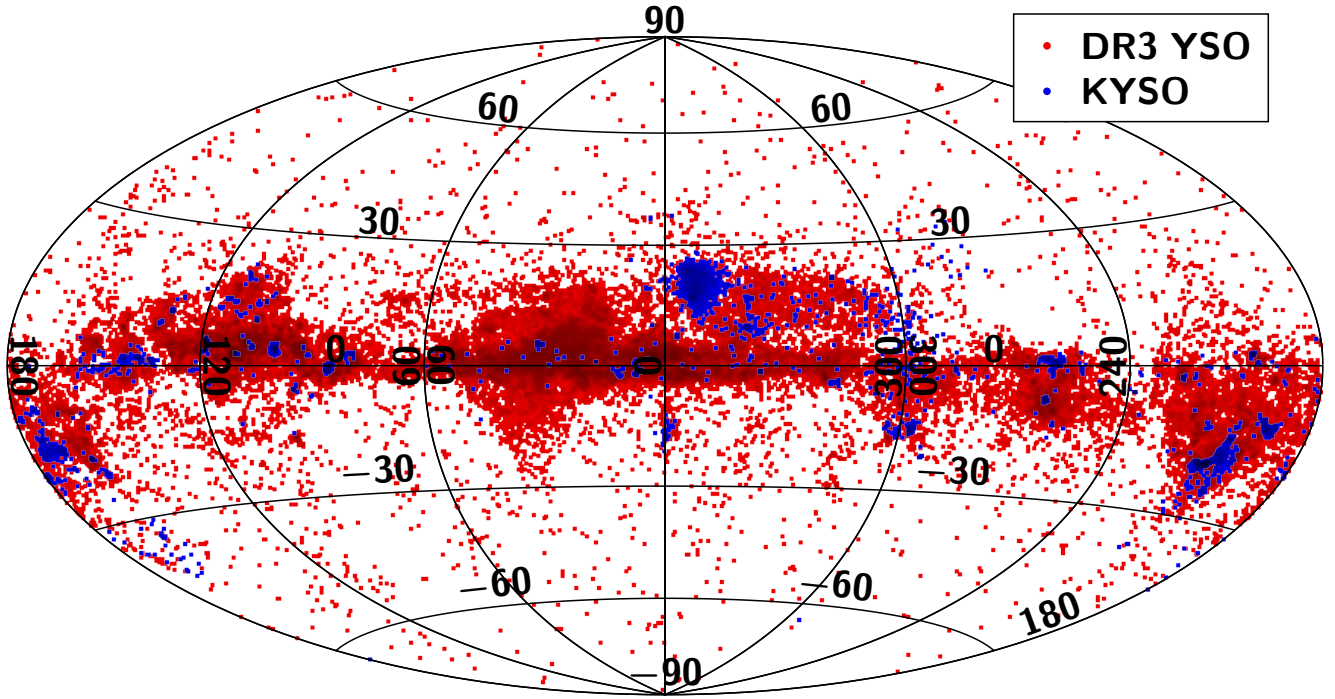
Fig. 3. DR3 YSO sample (red dots) and the KYSO objects (blue dots) presented in Hammer-Aitoff projection in Galactic coordinates. The KYSO sample is heavily biased as it lists sources from well studied star forming regions, while *Gaia* is an unbiased all-sky survey, and therefore *Gaia* DR3 YSO candidates can be seen in all directions. YSO candidates are seen at high Galactic latitudes ($|b| \geq 30°$), their number is 940, which is only 1.2% of all the candidates. The median distance of these high-latitude sources is 363.6 pc, and 95% of them are within a distance of 869.4 pc.
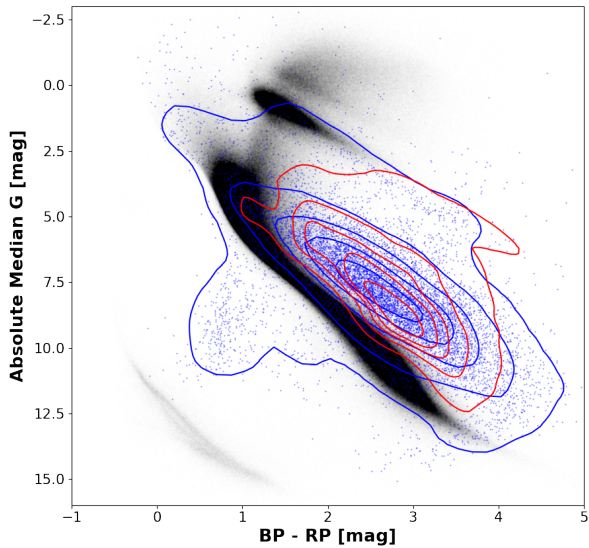

Fig. 4. Observational HRD of *Gaia* DR3 YSOs (red contours), KYSOs (blue dots and blue contours), and reference sources (grey dots) as described in the caption of Fig. 1. Most of the KYSOs are located above the main sequence, but other parts of the HRD are also covered. The probability distributions of the $G_{BP} - G_{RP}$ colour and the G_{Abs} are shown in Figs. B.1 and B.5. The main parameters of the distributions are listed in Table 1.
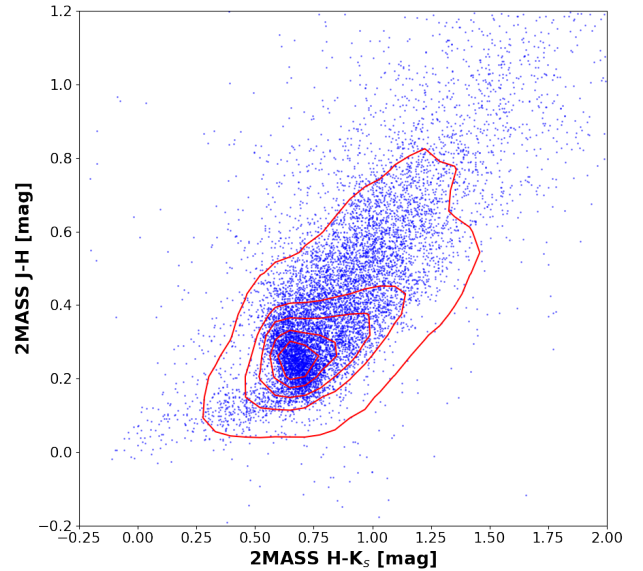

Fig. 5. KYSO sources (blue dots) and *Gaia* DR3 YSOs (red contours) on the 2MASS colour–colour diagram. The median, mean, standard deviation, and 5% and 95% quantiles of both colours are listed in Table 1. The colour probability distributions are shown in Figs. B.9 and B.13. The main parameters of the distributions are listed in Table 1.

region as the M19 objects, but only show one maximum in the distribution, unlike the M19 sources, which show bimodality. As explained earlier, the method used for the M19 classification allowed more evolved objects to be classified as YSOs with relatively high probability when the long-wavelength WISE measurements were not taken into account.

3.1.3. Cross-match with the Kuhn et al. (2021) YSO catalogue

Based on *Spitzer* space telescope surveys of the Galactic mid-plane between $l \sim 255°$ and $110°$, including the GLIMPSE I, II, and 3D, Vela–Carina, Cygnus X, and SMOG surveys

Table 1. Median, mean, standard deviation, and 5% and 95% quantiles of the G_{Abs}, $G_{BP} - G_{RP}$, $J - H$ and $H - K_s$ distributions of the *Gaia* DR3 YSO sample and other public catalogues used for the validation process.

Parameter	Median	Mean	Stdev	5%	95%
	DR3 YSO candidate				
G_{Abs}	7.623	7.506	1.854	4.384	10.316
$G_{BP} - G_{RP}$	2.770	2.748	0.601	1.749	3.697
$J - H$	0.746	0.791	0.235	0.513	1.195
$H - K_s$	0.287	0.318	0.180	0.152	0.589
	KYSO				
G_{Abs}	7.764	7.609	2.431	3.321	11.281
$G_{BP} - G_{RP}$	2.569	2.534	0.879	0.935	3.889
$J - H$	0.796	0.881	0.368	0.495	1.549
$H - K_s$	0.377	0.455	0.306	0.143	1.025
	G19 Großschedl et al. (2019)				
G_{Abs}	9.838	8.594	1.990	5.133	11.351
$G_{BP} - G_{RP}$	2.562	2.465	0.884	0.778	3.709
$J - H$	0.951	1.203	0.745	0.568	2.778
$H - K_s$	0.549	0.719	0.521	0.246	1.783
	SPICY Kuhn et al. (2021)				
G_{Abs}	6.228	6.024	1.657	2.947	8.386
$G_{BP} - G_{RP}$	2.754	2.777	0.752	1.580	4.066
$J - H$	1.372	1.459	0.657	0.552	2.673
$H - K_s$	0.899	1.008	0.606	0.287	2.166
	M19 Marton et al. (2019)				
G_{Abs}	6.182	6.546	2.191	3.608	10.705
$G_{BP} - G_{RP}$	2.709	2.466	0.811	1.173	3.491
$J - H$	0.873	0.845	0.340	0.353	1.394
$H - K_s$	0.384	0.383	0.217	0.100	0.758

Notes. Each quantity is presented in magnitudes. Histograms for each sample and each quantity are presented in Appendix B.

(613 square degrees), Kuhn et al. (2021) published a probabilistic YSO catalogue, the SPICY catalog. These authors presented 117 446 *Spitzer*/IRAC candidate YSOs.

Because *Spitzer* observes the IR domain, it is an ideal tool for YSO discovery. Therefore, it is expected that more of the YSOs in the SPICY catalogue are in the earlier stages of evolution, showing more IR excess. Also, because the SPICY contains sources seen towards the Galactic midplane, where the amount of interstellar dust is very high and obscures the visible light, one can expect that only the brighter, higher-mass YSOs are seen with *Gaia*. Figure 9 shows the HRD for the *Gaia* DR3 YSO candidates and SPICY YSOs. The overlap is significant, but we see more *Gaia* DR3 YSO candidates in the $3 < G_{BP} - G_{RP} < 4.5$ and $5 < G_{Abs} < 10$ region. These are objects in nearby star forming regions and only a few of them are located in the region covered by the SPICY catalogue. While the median G_{Abs} brightness of the *Gaia* DR3 YSOs is 7.6 mag, the SPICY objects are brighter by 1 mag, and the median value is 6.2 magnitude. The median value of the $G_{BP} - G_{RP}$ is very similar, 2.7 for both samples.

The 2MASS colour–colour distribution is shown in Fig. 10. As expected, the SPICY objects cover a larger area on the diagram. While the median $J - H$ colour of the *Gaia* DR3 YSOs is 0.7 mag, this value is 1.4 mag for the SPICY objects. This difference is also seen in the median value of the $J - K_s$ colour distribution, as the median value for the *Gaia* DR3 YSO candidates is 0.3 magnitude while it is 0.9 for the SPICY objects.
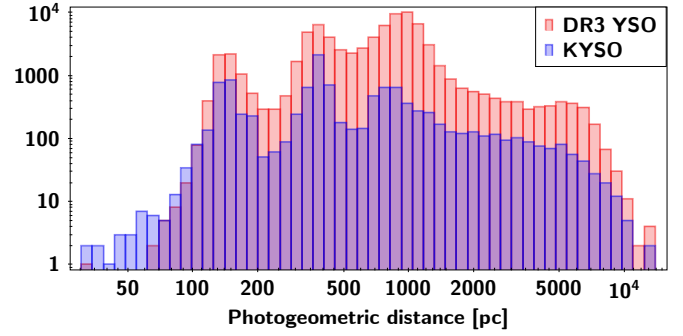
Fig. 6. Distance distribution of KYSO sources (blue bars) and of sources classified as *Gaia* DR3 YSOs (red bars). The values show strong similarities, except that there are more KYSOs within 100 pc of the Sun. These nearby KYSO sources (87 have distances of smaller than 100 pc) are mostly members of the TW Hydrae association. There are also three strong peaks in the distance distribution. The first one, between ∼125 and ∼175 pc, corresponds to the Ophiuchus, Sco OB2 association, and Taurus regions. The second strong peak is seen at distances of between ∼300 and ∼500 pc. The vast majority of these sources are located in Orion. The third peak is between ∼300 and ∼1000 pc. These sources are also seen towards Orion and towards the Galactic midplane as well.
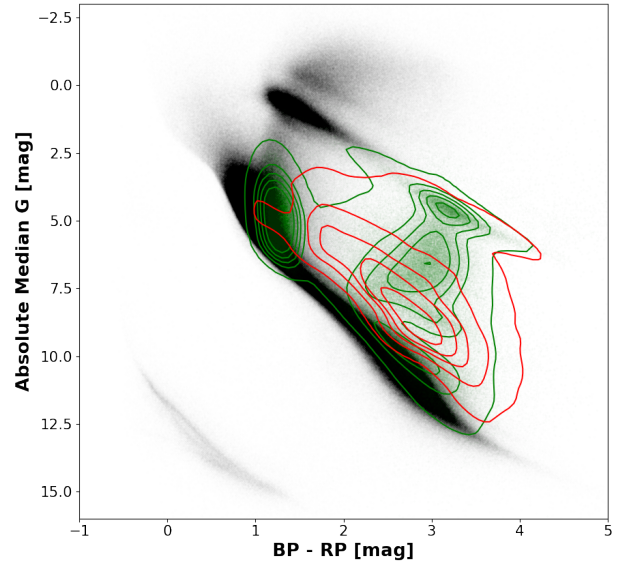

Fig. 7. Same as Fig. 4 but for the *Gaia* DR3 (red contours) and M19 YSO candidates with $R \geq 0.5$ and $LY \geq 0.95$, or $R < 0.5$ and $SY \geq 0.95$ (green dots and contours). The probability distributions are shown in Figs. B.4 and B.8. The main parameters of the distributions are listed in Table 1. Because the *Gaia* DR3 YSO distribution is narrower, for better visibility we plotted them on the top of the M19 distribution.

3.1.4. Cross-match with the Großschedl et al. (2019) paper

Großschedl et al. (2018) used the *Gaia* DR2 distances of MIR-selected YSOs in the benchmark giant molecular cloud Orion A to infer its 3D shape and orientation based on the ESO–VISTA NIR survey. At a later stage, an updated source list was published in Großschedl et al. (2019). We used this source list (hereafter G19) to match our sources to YSOs in the Orion A star forming region.

Again, as a first step, we checked the observational HRD and analysed the location of the objects shown in Fig. 11. The G19 YSOs are seen not only above the main sequence, but also in the region between the main sequence and the location of the
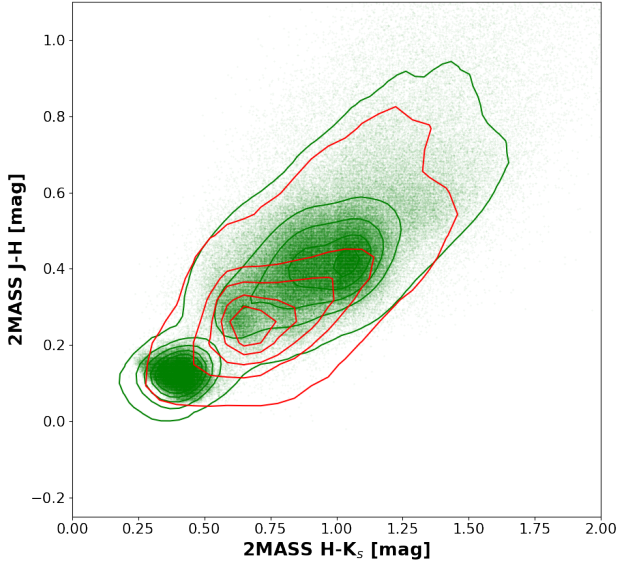
Fig. 8. Same as Fig. 5, but for the M19 sources (green dots and contours) and *Gaia* DR3 YSO candidates (red contours). The colour probability distributions are shown in Figs. B.12 and B.16, and the main parameters of the distributions are listed in Table 1.
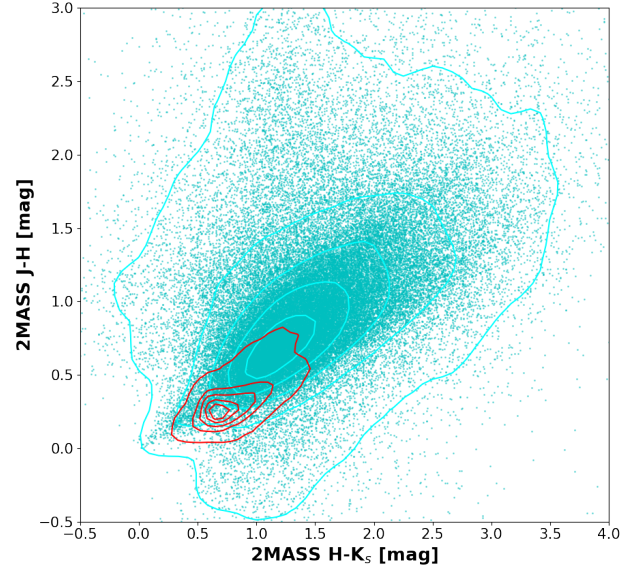

Fig. 10. SPICY sources (cyan dots and contours) and *Gaia* DR3 YSOs (red contours) on the 2MASS colour–colour diagram. The $G_{BP} - G_{RP}$ and G_{Abs} distributions are shown in Figs. B.11 and B.15, and the main parameters of the distributions are listed in Table 1.
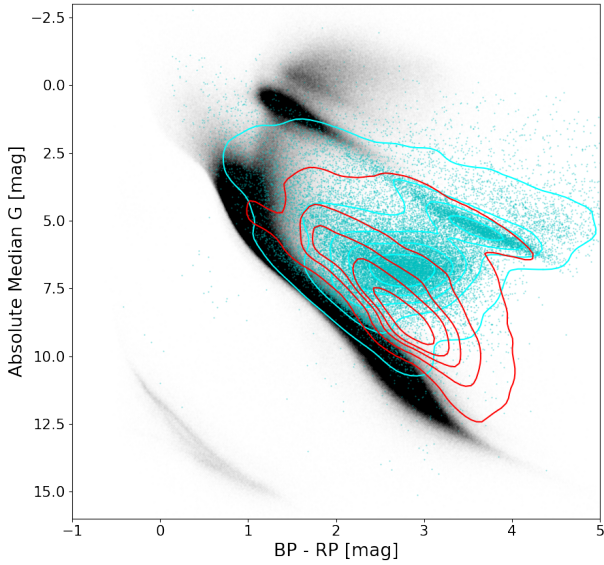

Fig. 9. DR3 YSO candidates (red contours) and SPICY YSOs (cyan dots and contours) on the HRD. The $G_{BP} - G_{RP}$ and G_{Abs} distributions are shown in Figs. B.3 and B.7, and the main parameters of the distributions are listed in Table 1.
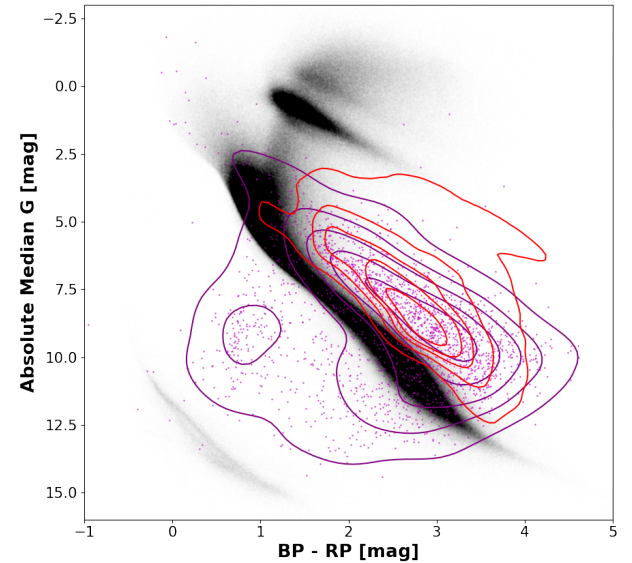

Fig. 11. Same as Fig. 4 but for *Gaia* DR3 (red contours) and G19 (purple dots and contours) YSO candidates. Histograms of the two parameter distributions are shown in Figs. B.6 and B.2. The main parameters of the distributions are listed in Table 1. The distribution shows strong similarities to that seen in Fig. 4.

white dwarfs. This feature of their distribution is discussed in Sect. 3.2.1. It is also clear from Fig. 11 that the G19 objects tend to have lower G_{Abs} values. As listed in Table 1, the median G_{Abs} of the G19 objects is 9.8 mag, while this value is 7.6 for the *Gaia* DR3 YSO candidates. Because the Orion A is a nearby star forming region with an average distance of ~420 pc, it is expected that more of the fainter YSOs are seen with *Gaia*.

The distributions of G19 sources and *Gaia* DR3 YSOs on the 2MASS colour–colour diagram are presented in Fig. 12. The two samples show significant overlap, except that fewer G19 objects have $J-H < 0.6$ colour, but sources with larger IR excess are also detected. This can be also explained with the fact that the Orion A is a nearby system, and therefore those sources that do not

emit the majority of their energy in the optical domain are still observable with *Gaia*.

We also tested the *Gaia* DR3 YSO sample against the 3D shape of the Orion A cloud that was analysed in Großschedl et al. (2018). Figures 13 and 14 clearly show that the median distance values for the *Gaia* DR3 YSO sample as a function of Galactic longitude are in very good agreement with the findings of these latter authors, and we can confirm that the Orion A cloud is an elongated structure with its head part closer to us and its tail at a greater distance towards higher longitude values.
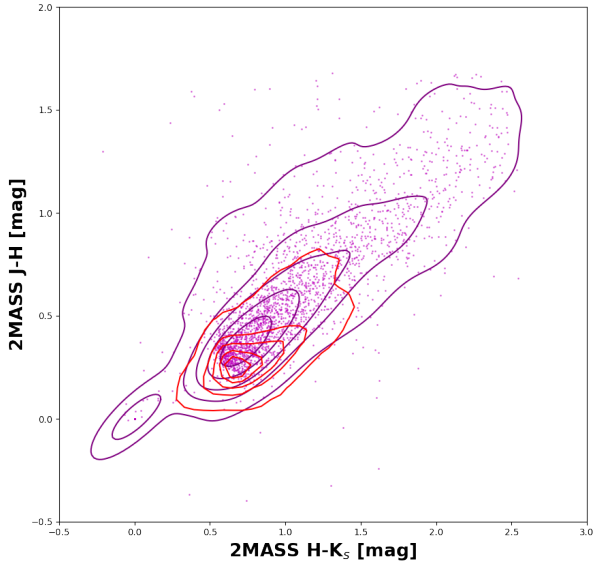
Fig. 12. G19 sources (purple dots and contours) and *Gaia* DR3 YSOs (red contours) on the 2MASS colour–colour diagram. Probability distributions are shown in Figs. B.10 and B.14. The main parameters of the distributions are listed in Table 1.
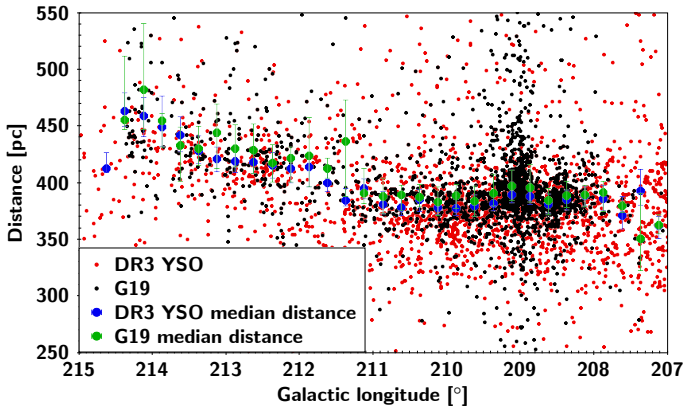

Fig. 13. Distance of the *Gaia* DR3 YSO candidates (red dots) and the Großschedl et al. (2018) YSOs (black dots) as a function of Galactic longitude. Large dots represent the *med(D)*, which is median of the individual distance values in 0.25° bins, while error bars represent the *med(D − D_L)* and *med(D − D_U)*, where D_L and D_U are the lower and upper limits of the individual distance values of the *Gaia* DR3 YSO candidates (blue dots) and of the Großschedl et al. (2018) YSOs (green dots).

3.1.5. Distances of star forming regions

Zucker et al. (2019) presented a uniform catalogue of accurate distances to local molecular clouds based on *Gaia* DR2. These authors used a sophisticated method to derive distances to 27 nearby star forming clouds, and reported upper and lower corner coordinates of regions in which they calculated average distances based on *Gaia* astrometry and optical–NIR photometry. In this comparison, we simply calculated a median distance of the *Gaia* DR3 YSOs based on the distance values of Bailer-Jones et al. (2021). The values are listed in Table 2. As shown in Fig. 15, with the exception of two regions, the *Gaia* DR3 YSO results are in good agreement with the findings of Zucker et al. (2019).

The two exceptions are the Gem OB1 and the Maddalena regions. In the case of the Gem OB1, the reported distance was
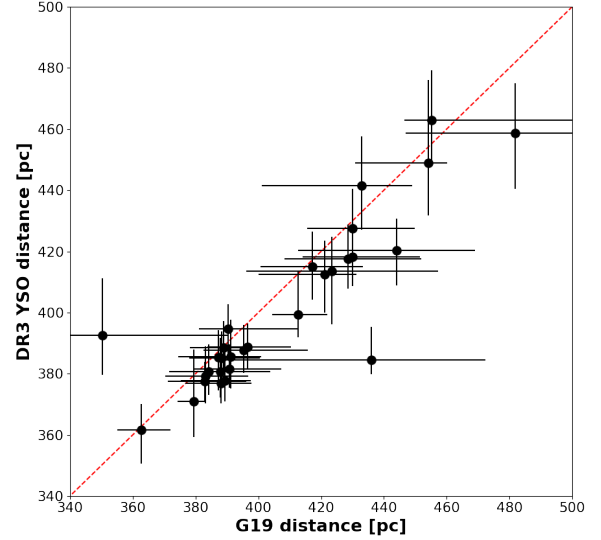
Fig. 14. Distances of the DR3 YSO candidates versus those of the G19 YSOs in the star forming region Orion A. The error bars represent the same values as in Fig. 13.

1786±89 pc, while the *Gaia* DR3 YSO sources showed a strong peak between 320 and 450 pc, and the farthest object was found to be at a distance of 1067.5 pc. In the direction of the Maddalena star forming region, we also see a peak in the distance distribution between 320 and 420 pc and the farthest object being at a distance of 1079.5 pc. For the case of Maddalena, it is known that despite its large mass, only low level star formation is happening in the cloud (e.g. see Schneider et al. 2015).

### 3.1.6. Cross-match with the *Gaia* Photometric Science Alerts

The KYSO catalogue was also used to improve the classification of the *Gaia* photometric science alerts (GSA; Hodgkin et al. 2021) with other YSO catalogues from the literature. Still, numerous objects remained unknown among the alerting sources, including possible YSOs. Therefore, we also checked which of the *Gaia* DR3 YSOs were also among the alerts. At the beginning of February 2022, 478 alerts appeared on the GSA Index website[2] classified as YSOs or mentioned in the comment section as possible YSOs from the total of 18 976 alerts. A cross-match revealed that 8905 (46.9%) of them are in the *Gaia* DR3 catalogue. The number of *Gaia* DR3 YSO candidates present in the alert list was found to be 159, which is 33.3% of the possible YSOs alerts. Also, 41 sources were not identified as confirmed or possible YSOs. In the alert system, these are listed as mostly unknown, red stars in the direction of the Galactic midplane. Below, we present a list of YSO alerts that were or are being investigated with follow-up observations in more detail to infer the underlying physics of their brightness changes and in uncertain cases to confirm their YSO nature. The distance estimates for each source are based on the *Gaia* EDR3 distance catalogue of Bailer-Jones et al. (2021), while the probability value of being a YSO is from Marton et al. (2019).

– Gaia22afv ($\alpha_{J2000} = 04^h 03^m 38^s.86$, $\delta_{J2000} = 32° 15' 49''.93$) is a candidate YSO, which triggered the *Gaia* Alerts system on 2022 January 18 because of its brightening by ~0.7 mag. Its distance is $263.7^{+6.5}_{-4.8}$ pc and the probability of being a YSO is 67.22%

---

[2] http://gsaweb.ast.cam.ac.uk/alerts/alertsindex

Table 2. Distances of star-forming clouds based on the *Gaia* DR3 YSO sample in comparison with those reported by Zucker et al. (2019).

Cloud name	DR3 median parallax [mas]	DR3 parallax MAD [mas]	B.-J.(2021) med. dist. [pc]	B.-J.(2021) dist. MAD [pc]	Zucker et al. dist. [pc]	Zucker et al. dist. err. [pc]	DR3 YSO member count
AquilaS	2.473	1.581	385.138	144.398	133	7	62
California	2.771	0.633	352.502	147.169	470	24	770
Chamaeleon	2.968	1.733	337.852	253.633	183	9	394
CMaOB1	0.892	0.188	1098.853	158.152	1209	60	408
CoronaAustralis	6.351	2.106	156.923	275.613	151	8	28
Crossbones	1.051	0.304	918.044	140.597	886	44	176
GemOB1	2.348	0.931	414.320	209.351	1786	89	64
Lupus	6.141	2.233	161.129	267.720	189	9	1059
Maddalena	2.351	0.835	398.265	207.223	2072	104	17
MonOB1	1.373	0.175	709.091	121.870	745	37	577
MonR2	1.146	0.264	844.402	132.447	778	39	473
Ophiuchus	6.978	1.435	141.990	150.515	144	7	791
OrionA	2.531	0.267	388.360	65.817	432	22	2072
OrionB	2.439	0.457	402.371	136.768	423	21	854
OrionLam	2.493	0.374	393.914	87.140	402	20	1075
Perseus	3.164	0.496	310.011	117.780	294	15	435
Polaris	1.420	0.439	682.982	174.088	352	18	15
Rosette	0.950	0.581	1092.109	233.491	1304	65	78
SerpensAqr	1.576	0.470	650.556	273.240	484	24	1136
Taurus	4.355	2.535	232.237	260.722	141	7	432
UrsaMajor	2.154	1.596	453.741	258.613	371	19	12
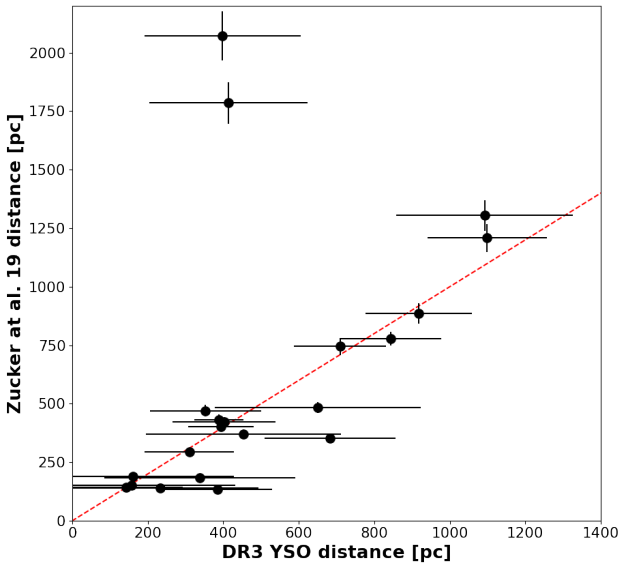
Fig. 15. Median distance of *Gaia* DR3 YSOs in regions defined by Zucker et al. (2019) versus the distances reported by these latter authors based on *Gaia* DR2 data. Horizontal error bars represent the standard deviation of distances of *Gaia* DR3 YSOs in the given region. Vertical error bars are the systematic errors given in Table 1 of Zucker et al. (2019).

– Gaia18dlf (α_{J2000} = 20^h 57^m 03^{s}36, δ_{J2000} = 43° 41′ 44″45) is a known YSO, which had a *Gaia* alert on 2018 November 19 because of its long-term (on a timescale of a year) brightening by more than 1 mag. Its distance is 782^{+31}_{-34} pc. Rebull et al. (2011) classified it as a flat-spectrum source based on the near- to mid-IR SED slope.

– Gaia21egm (or V733 Cep, α_{J2000} = 22^h 53^m 33^{s}25, δ_{J2000} = 62° 32′ 23″60) is a known FUor (Reipurth et al. 2007), which triggered a *Gaia* alert on 2021 September 25 because of a drop in its brightness following a long-term fading. Its distance is $724.3^{+31.3}_{-26.8}$ pc. It exhibited a slow rise in its brightness by

~4.5 mag between 1971 and 1993, after which it stayed at maximum brightness for several decades (Peneva et al. 2010).

– Gaia21arq (α_{J2000} = 05^h 34^m 28^{s}94, δ_{J2000} = −05° 08′ 38″40) is a known YSO, which triggered a *Gaia* alert on 2021 February 9, due to its brightening by 0.8 mag. Its distance is $354.6^{+17.7}_{-14.9}$ pc. It is a classical T Tauri star, which was part of the Hα survey of the Orion Nebula Cluster by Szegedi-Elek et al. (2013).

– Gaia18eap (α_{J2000} = 02^h 34^m 34^{s}63, δ_{J2000} = 61° 21′ 53″64) is a YSO candidate, which had a *Gaia* alert on 2018 December 29 because of its more than 1 mag brightening. The duration of the brightening event was about a year: the source returned to its original brightness by early 2020. Its distance is 980^{+248}_{-189} pc. It is classified as a flat-spectrum source (Sung et al. 2017).

– Gaia21eox (α_{J2000} = 00^h 04^m 13^{s}63, δ_{J2000} = 67° 24′ 45″50) is a YSO candidate at a distance of $977.02^{+56.72}_{-53.84}$ pc. It had a *Gaia* alert on 2021 October 10 because of a 1 mag dimming episode, which ended by January 2022. Its probability of being a YSO is 98%.

As shown by the list above, the DR3 YSO candidate sample has the potential to contain eruptive YSOs, which are excellent targets for detailed observations aiming to better understand the fundamentals of the disc- and planet formation and evolution.

3.2. Completeness

3.2.1. Completeness based on the KYSOs

The KYSOs also occupy a region on the observational HRD close to or above the giant branch and below the main sequence, while the *Gaia* DR3 YSOs have a narrower distribution as seen in Fig. 4. After visualising their distances, as shown in Fig. 16, we realised that many of them are distant objects (plotted with dark blue or black colours) and also their ϖ/σ_ϖ value is below 3, while for the *Gaia* DR3 YSOs we set a requirement for $\varpi/\sigma_\varpi \geq 3$. This also means that the distance estimates of the KYSOs are less reliable in some cases and this can cause a less precise position on the colour–magnitude diagrams. Another feature in the distribution of the KYSOs on the observational
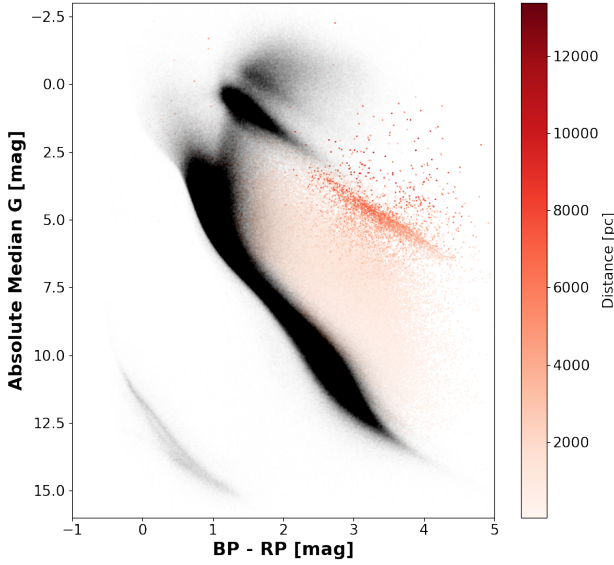
Fig. 16. KYSOs on the observational HRD. Colour coding corresponds to the distance of the individual objects. Grey dots represent the reference objects on the HRD. Many of the YSOs close to the giant branch or above it seem to be distant objects. Also, many of them have parallax errors that make the proper distance estimation problematic, and therefore their position on the diagram is more uncertain.

Fig. 17. Distribution of parallax over its uncertainty ϖ/σ_ϖ for KYSO sources classified as *Gaia* DR3 YSOs (green bars) and those KYSOs that were excluded from the final *Gaia* DR3 YSO candidate list (grey bars). The vertical red line at $\varpi/\sigma_\varpi = 3$ indicates the threshold below which we excluded all sources.

HRD is an overdensity of sources with $G_{BP}-G_{RP}$ colour between 0 and 1 mag and with an absolute median G band brightness of between 11 and 5 mag, located between the main and white dwarf sequences. The vast majority (>95%) of these sources are located in the Orion star forming region and have M spectral type, and all are reported in the study of Hillenbrand et al. (2013). They are most likely YSOs for which we do not see their photosphere but scattered light from their protoplanetary discs. At high inclination, when the disc is seen edge-on, the sources appear to be fainter and bluer than they actually are (Guarcello et al. 2010). We note that this feature on the colour–magnitude diagram also appears in Fig. 11, where the literature YSOs are also from the Orion.

As a further step in our investigation, we checked the distribution of the parallax over its uncertainty (ϖ/σ_ϖ) for those sources that were included in the *Gaia* DR3 YSO sample from the KYSO catalogue and the excluded ones. The distribution of the values is shown in Fig. 17. The excluded KYSO objects clearly tend to have lower values, meaning that their parallax is less reliable.

In total, 4656 KYSO sources were included in the final *Gaia* DR3 YSO sample, which means a ~40% completeness in this comparison. Of the KYSOs, 5286 were included in the classification process with the supervised classifier (Rimoldini et al. 2023), and 361 of them ended up in a class other than YSO. A quarter of the 11 671 sources listed in the KYSO table were rejected based on data quality cuts applied in the classification verification phase to ensure that the *Gaia* DR3 YSO sample, while not complete, provides a reliable list of confirmed variable YSOs and potential YSO candidates.

3.2.2. Completeness based on Marton et al. (2019)

A 1″ radius was used to match the *Gaia* DR3 YSO source positions to the M19 catalogue, resulting in 40 320 objects being in both the M19 and *Gaia* DR3 YSO samples, which means 51%
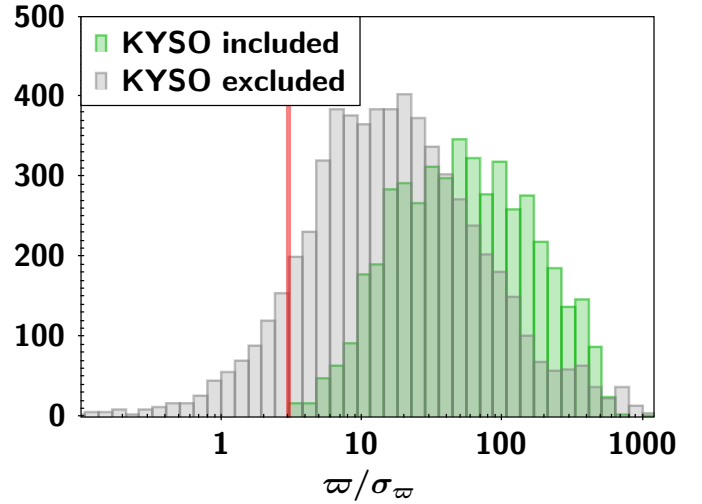
of the *Gaia* DR3 YSO candidates. While *Gaia* DR3 provides information on sources from all over the sky, the M19 catalogue was restricted to an area where the dust opacity value based on the *Planck* foreground maps (Planck Collaboration X 2016) was higher than 1.3×10^{-5}. This means that YSOs were searched for on 25.4% of the sky. Because of this cut, as many as 6037 *Gaia* DR3 YSO candidates cannot have a counterpart in the M19 sample. In the common area, 55% of the sources have a match.

In this completeness study, we used sources from the M19 catalogue that had YSO as their best label, meaning that the probability of being a YSO was higher than any of the other probabilities, which means that in cases where $R \geq 0.5$, then $LY > LMS$, $LY > LEG$, and $LY > LE$, while if $R < 0.5$, then $SY > SMS$, $SY > SEG$, and $SY > SE$, where MS, EG, and E refer to the probability of being a main sequence source, an extragalactic object, and an evolved star, respectively. The number of sources in the M19 catalogue fulfilling these criteria is 15 052 388.

In this case, 6 488 *Gaia* DR3 YSO candidates were found to have $R \geq 0.5$ and the highest probability label was LY from the M19 catalogue, while 21 897 candidates had $R < 0.5$ and best label SY. This means that, in total, 28 385 *Gaia* DR3 YSO candidates were classified as possible YSOs according to the M19 catalogue. Compared to the total number of such sources in the M19 catalogue, the estimated completeness level is 0.19%.

If we compare to those M19 sources with $R \geq 0.5$ and $LY \geq 0.95$ or $R < 0.5$ and $SY \geq 0.95$ (259 363 in total), the overlap with the *Gaia* DR3 YSO candidates is 6087 sources, and the completeness level is 2.35%. The best labels (the object type with the highest probability for a given object) of possible evolved star, possible extragalactic source, and possible main sequence star were assigned to 9482 (23.5%), 2143 (5.3%), and 307 (0.8%) sources, respectively. We cross-matched these sources with SIMBAD in order to investigate the completeness and contamination in more detail. Our findings are summarised in Table 3. Based on the numbers, we find that 879 (65.5%) out of the 1 343 sources classified as evolved stars in the M19 are listed as some kind of YSO in SIMBAD, 142 (67%) of the 212 M19 extragalactic sources that are in SIMBAD are also potential

Table 3. SIMBAD classification of *Gaia* DR3 YSO candidates classified as possible contaminants in the M19 catalogue.

	M19 E	M19 EG	M19 MS
Total	9482	2143	307
in SIMBAD	1343	212	264
YSO	278	24	41
Candidate_YSO	392	112	54
Orion V*	54	2	12
TTau*	155	4	39
Probably not YSO	464	70	118

Notes. The first row lists the total number of *Gaia* DR3 YSO candidates classified as either evolved star (E), extragalactic source (EG), or main sequence star (MS) in the M19. The second row shows the number of objects also found in SIMBAD. Rows (3)–(6) detail how many of these are classified in SIMBAD as some sort of YSO. The last row shows the number of sources that are not YSO related.

YSOs, and finally 146 (55.5%) of the 264 SIMBAD counterparts of the M19 main sequence stars are also possibly YSOs.

Combining the M19 classifications with the SIMBAD classifications results in 28 385+1167 = 29 552 YSOs. This means a small increase in completeness to 0.20%.

3.2.3. Completeness based on Kuhn et al. (2021)

As the SPICY catalogue did not require data from the optical domain and is heavily concentrated on the Galactic midplane, one may expect a low fraction of matching sources. Using a 1″ radius, we find only 753 objects present in both the *Gaia* DR3 YSO sample and the SPICY catalogue. These can all be considered as possible YSOs, because the SPICY catalogue includes only sources that were classified as YSO with a probability of greater than 0.5 in at least one of the three classifiers used in their study.

SPICY also provided an estimated evolutionary class based on the IR slope of the SED. The age of a YSO is increasing in the following order: Class I, flat spectrum, Class II, and finally Class III. We found 8 Class I, 87 flat spectra, 598 Class II, 57 Class III objects, and 3 with uncertain class. In the SPICY catalogue, the numbers of objects in the respective classes are 15 943, 23 810, 59 949, and 5352, meaning that 0.05%, 0.4%, 1%, and 1.1% of each class was found, which reflects that more evolved sources – which we expect to be more visible in *Gaia* – are found more reliably. The detailed numbers are listed in Table 4, which also shows that among the sources that were present in the variability analysis, the completeness is much higher as it was found to be 20.3% considering all evolutionary classes, being the highest among Class II YSOs (23.36%).

3.2.4. Completeness based on Großschedl et al. (2019)

Similarly to the SPICY catalogue, this study also reported evolutionary stages for the YSOs, and therefore we were able to estimate completeness as a function of age. The study lists 3117 sources and they are classified into several types as listed in the first column of Table 5. The number of sources classified into each type is listed in Col. 2. In order to obtain an estimate of the completeness of the *Gaia* DR3 YSO sample, we calculated three quantities. First, we checked how many of the different types of sources were visible to *Gaia* based on *Gaia* DR3 (Col. 3), how many of them participated in the variability classification

(Col. 4), and how many are among the *Gaia* DR3 YSOs (Col. 5). The ratio of Cols. 5 and 2 is listed in Col. 6. Column 7 is the ratio of Cols. 5 and 3, and Col. 8 is the ratio of Cols. 5 and 4.

The total completeness was found to be 19.9%. As expected, the completeness is very low for Class 0 and I sources, as they emit most of their radiation at IR wavelengths, but completeness grows with the evolutionary stage and for transition discs it reaches 38.3%.

3.2.5. Cross-match with the SIMBAD database

The Object Type in SIMBAD is defined as a hierarchical classification, which emphasises the physical nature of the object rather than a peculiar emission in some region of the electromagnetic spectrum or the location in a peculiar cluster or external galaxy. Therefore, objects are only classified as peculiar emitters (in radio, IR, red, blue, UV, X-ray, or gamma ray) if nothing more about the nature of the object is known; that is, if it cannot be decided whether the object is a star, a multiple system, a nebula, or a galaxy (Wenger et al. 2000).

The total number of objects we find in the SIMBAD database using a 1″ search radius is 20 531. The first column in Table 6 is the SIMBAD main_type, the second column shows how many objects with the given SIMBAD main_type were found in the *Gaia* DR3 YSO sample, while the last column is the percentage of the given main_type according to the total number of 20 531 associated SIMBAD objects. As one can see from Table 6, there are multiple object types that we can use to estimate the completeness, because they can be considered as potential YSOs. We calculated the sum of objects belonging to the following types: Candidate_YSO, YSO, TTau*, Candidate_TTau*, Orion_V*, Ae*, Candidate_Ae*, and Be*. The total number of these objects is 15 363, which means that 74.8% of the total SIMBAD associations can be considered as potential YSOs. However, the number of sources listed in SIMBAD with the same main_types is the following: 49 945 YSO, 99 097 Candidate_YSO, 5 079 TTau*, 317 Candidate_TTau*, 2 887 Orion_V*, 147 Ae*, 67 Candidate_Ae*, and 2 215 Be*; in total, 159 754 sources can be considered as possible YSOs. Based on these numbers, the estimated completeness using the SIMBAD catalogue is 9.6%. This number is mostly determined by the large number of Candidate_YSO type objects. If we do not take these into account, the estimated completeness is 11%.

3.2.6. Magnitude and extinction limitations

The completeness in astronomy is mainly determined by the apparent brightness of the objects in the sky. This is especially true for YSOs, as in their early evolutionary stages most of their energy is emitted at wavelengths invisible to *Gaia*. Moreover, they are located in obscure regions where the interstellar dust can modify their observed SEDs, resulting in lower apparent brightness at the visible wavelengths.

Therefore, we also analysed the completeness of the *Gaia* DR3 YSO candidate sample as a function of the *Planck* dust opacity value (τ, Planck Collaboration X 2016). As shown in Fig. 18, we had no objects in the *Gaia* DR3 YSO candidate sample above $\tau = 0.0064$. We find that 99% of the YSOs from the combined KYSO, SPICY, M19, and G19 catalogues are located above $\tau = 0.0018$, while 99% of the *Gaia* DR3 YSO candidates are in regions where $\tau \geq 0.0008$.

We also analysed the completeness as a function of the absolute *G* band magnitude and *Planck* τ. As shown in Fig. 19, very faint YSOs were not recovered by the classification process and

Table 4. Number of objects in the different samples from the SPICY catalogue.

Source type	SPICY	*Gaia* DR3	In var. class.	*Gaia* DR3 YSO	*Gaia* DR3 YSO % of SPICY	*Gaia* DR3 YSO % of *Gaia* DR3	*Gaia* DR3 YSO % of var. class.
Total	117 446	37 693	3693	751	0.64	1.99	20.30
Class I	15 943	2529	130	8	0.05	0.32	6.15
Flat spectrum	23 810	5480	514	87	0.37	1.59	16.93
Class II	59 949	24 486	2560	598	1.00	2.44	23.36
Class III	5352	3413	462	57	1.07	1.67	12.34

Notes. The meanings of the columns are as follows: (1) source type as reported in SPICY; (2) number of sources reported in SPICY; (3) number of sources from the SPICY catalogue also present in the *Gaia* DR3; (4) number of sources participating in the variability classification; (5) number of sources classified as YSO during variability classification; (6) fraction of SPICY sources classified as YSO, i.e. Col. 5 divided by Col. 2; (7) fraction of *Gaia* DR3 sources classified as YSO, i.e. Col. 5 divided by Col. 3; (8) fraction of those sources participating in the variability classification and classified as YSO, i.e. Col. 5 divided by Col. 4.

Table 5. Number of objects in the different samples from the G19 catalogue.

Source type	Großschedl et al.	*Gaia* DR3	in var. class.	*Gaia* DR3 YSO	*Gaia* DR3 YSO % of G19	*Gaia* DR3 YSO % of *Gaia* DR3	*Gaia* DR3 YSO % of var. class.
Total	3117	2339	672	621	19.9	26.5	92.4
Class 0	60	0	0	0	0.0	0.0	0.0
Class I	128	9	1	1	0.8	11.1	11.1
Flat spectrum	185	91	20	13	7.0	14.3	65.0
Class II/III PMS with disc	2012	1685	503	473	23.5	28.1	94.0
Anemic disc	394	332	63	57	14.5	17.2	90.5
Transition disc	201	195	83	77	38.3	39.5	92.8
Galaxy	39	7	1	0	0.0	0.0	0.0
Nebulosity, fuzzy contamination	45	2	0	0	0.0	0.0	0.0
MS star	4	2	0	0	0.0	0.0	0.0
Class III w/o IR excess	3	3	0	0	0.0	0.0	0.0
Image artefact	5	0	0	0	0.0	0.0	0.0
Galaxy candidate	20	4	0	0	0.0	0.0	0.0
Uncertain YSO candidate	21	9	0	0	0.0	0.0	0.0

Notes. The columns list: (1) the source type as reported in G19; (2) the number of sources reported by G19; (3) the number of sources from the G19 catalogue also present in *Gaia* DR3; (4) the number of sources participating in the variability classification; (5) the number of sources classified as YSOs during variability classification; (6) the fraction of G19 sources classified as YSO, i.e. Col. 5 divided by Col. 2; (7) the fraction of *Gaia* DR3 sources classified as YSO, i.e. Col. 5 divided by Col. 3; (8) the fraction of those sources participating in the variability classification and classified as YSO; i.e. Col. 5 divided by Col. 4.

also sources at high τ values are missing, even if they are bright. On the other hand, sources brighter than 19 mag were found with a completeness of higher than 30% in regions where τ is very low, but for all magnitude bins, the completeness decreases as τ increases.

3.3. Contamination

In the previous sections, we mainly focus on the confirmation of the young nature of the sources and show that the *Gaia* DR3 YSO sample shows strong similarities to known YSOs in the colour–magnitude and colour–colour space, and that their distribution in the sky shows overdensities in the directions of the known star forming regions and towards the Galactic midplane. We also give estimations about the completeness of our catalogue, not only in general, but also as a function of evolutionary stage. In this subsection the SIMBAD and M19 catalogues

are used to estimate the contamination rate of the *Gaia* DR3 YSO sample, complemented with various catalogues accessible through the VizieR service (Ochsenbein et al. 2000). Our goal is to find large catalogues listing sources classified into several object types that are potential contaminants of the *Gaia* DR3 YSO set. In all cases, we used a 1″ radius to find a counterpart to *Gaia* DR3 YSOs.

#### 3.3.1. Marton et al. (2019)

In the M19 catalogue, the number of sources that are likely not related to YSOs after cross-matching them with SIMBAD and also do not have the best label of *S Y* or *LY* (depending on the *R* value) but are either main sequence stars, extragalactic sources, or evolved stars was found to be 652 (35.8% of the 1819 sources listed in SIMBAD). If we assume that the contamination rate is 35.8% among that sample of sources for which the best label is

Table 6. Number of sources associated with different SIMBAD objects types.

Simbad main_type	Number	%
Candidate_YSO	8683	42.29
YSO	3975	19.36
Star	2950	14.37
TTau*	1898	9.24
Em*	821	4.00
Orion_V*	773	3.77
RotV*	389	1.89
low-mass*	359	1.75
V*	187	0.91
Ae*	48	0.23
Candidate_LP*	48	0.23
EB*	46	0.22
**	33	0.16
BYDra	25	0.12
RGB*	25	0.12
RSCVn	20	0.10
SB*	20	0.10
X	20	0.10
Pec*	19	0.09
Candidate_TTau*	17	0.08
NIR	15	0.07
RRLyr	14	0.07
EllipVar	13	0.06
PulsV*	13	0.06
Eruptive*	9	0.04
MIR	8	0.04
Candidate_EB*	7	0.03
LPV*	7	0.03
Be*	6	0.03
Candidate_AGB*	6	0.03
Candidate_brownD*	6	0.03
HB*	6	0.03
Irregular_V*	6	0.03
PM*	6	0.03
brownD*	5	0.02
deltaCep	4	0.02
Mira	4	0.02
Planet	4	0.02
PulsV*WVir	4	0.02
V*?	4	0.02
Candidate_Ae*	3	0.01
IR	3	0.01
AGN_Candidate	2	0.01
Candidate_Cepheid	2	0.01
Candidate_Mi*	2	0.01
Candidate_RRLyr	2	0.01
Cepheid	2	0.01
BLLac_Candidate	1	0.00
denseCore	1	0.00
FIR	1	0.00

Table 6. continued.

Simbad main_type	Number	%
gammaDor	1	0.00
Nova	1	0.00
Planet?	1	0.00
PN	1	0.00
post-AGB*	1	0.00
PulsV*delSct	1	0.00
Radio	1	0.00
RotV*alf2CVn	1	0.00
S*	1	0.00
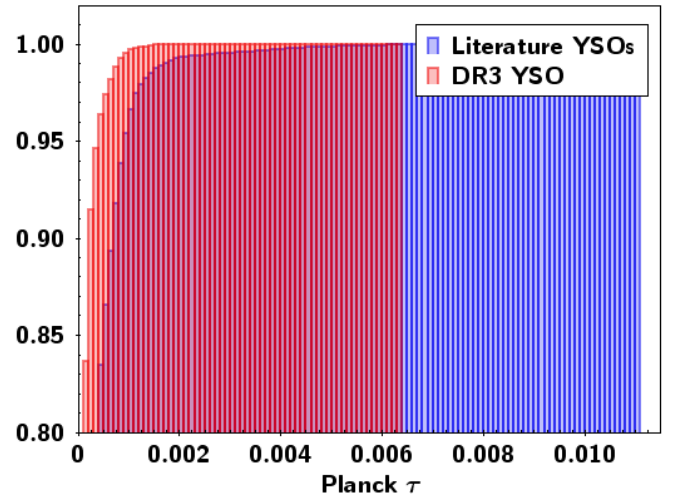
Fig. 18. Cumulative distribution of the *Planck* τ values in the directions of the YSOs from the KYSO, SPICY, M19, and G19 catalogues (blue bars) and that of the *Gaia* DR3 YSO candidates (red bars).
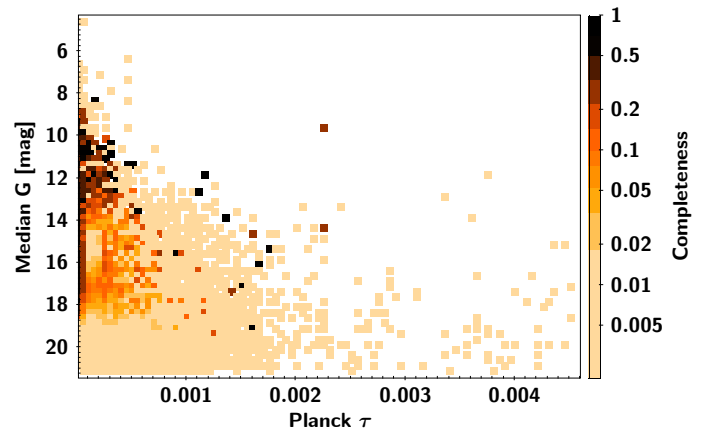
Fig. 19. Completeness calculated based on objects from the KYSO, SPICY, M19, and G19 catalogues as a function of *Planck* dust opacity (τ) and the absolute median *Gaia* G band magnitude. Colour coding presents the fraction of objects that are in any of the mentioned catalogues and also in the *Gaia* DR3 YSO sample in $5 \times 10^{-5} \times 0.25$ mag size bins on a logarithmic scale.

not YSO, then the expected number of such sources is 4272. This is 10.6% of the total number of sources present in both the M19 and *Gaia* DR3 YSO samples and gives a lower limit of contamination for the whole *Gaia* DR3 YSO candidate list. However, a more realistic approach could be to estimate the upper limit of the contamination. As described in Sect. 3.2.2, there are

40 320 sources common in the M19 and *Gaia* DR3 YSO candidate catalogues and 29 552 of them were identified as a possible YSO by either the M19 classification or the SIMBAD classification. Assuming that these classifications are realistic and that all

the other sources are contaminant, 26.7% of the *Gaia* DR3 YSO candidates are not young stars.

3.3.2. SIMBAD

In some cases, as a result of the cross-match with SIMBAD, the listed main type is too generic (like Star, Em*, low-mass*, etc.; see Table 6). Therefore, in order to estimate the contamination, we took into account only the well-defined object types that definitely cannot be YSOs. The number of objects included the following object types: BYDra, RGB*, RSCVn, SB*, RRLyr, Candidate_AGB*, Candidate_brownD*, HB*, brownD*, delta-Cep, Mira, Planet, PulsV*WVir, AGN_Candidate, Candidate_Cepheid, Candidate_Mi*, Candidate_RRLyr, Cepheid, BLLac_Candidate, gammaDor, Nova, Planet?, PN, post-AGB*, PulsV*delSct, Radio, RotV*alf2CVn, and S*. In total, these add up to 130 sources, 0.6% of the *Gaia* DR3 YSO associated with sources in Simbad. We consider this number as a lower limit for the contamination. As an upper limit, we consider all types that cannot be YSOs as contaminants, which means all types except those in Sect. 3.2.2 (YSO, Candidate_YSO, OrionV* and TTau*). Subtracting the number of these objects from the total number of 20 531 objects common with SIMBAD leaves 5 202 sources, meaning 25.3% contamination.

3.3.3. Cross-match with Gavras et al. (2023)

Rimoldini et al. (2023) compared the DR3 variable YSO candidates with the catalogues of variable objects collected by Gavras et al. (2023) as part of the evaluation process of the machine-learning-based classification. As a result Rimoldini et al. (2023) found a contamination rate of 79.8%. We analysed the 5236 matching sources, including 1 057 true positives (after excluding the trained YSO sources from the evaluation of completeness and contamination) and 4 179 false positives. The majority of the contamination comes from two types of stars: RS CVn binaries (1542) and BY Dra-type stars (1435).

About 99% of the contaminants that have the label BY and 98% of those with the label RS are from the Zwicky Transient Facility (ZTF) catalogue of Chen et al. (2020). Chahal et al. (2022) investigated this ZTF catalogue, paying special attention to the BY Dra-type objects, and collected additional photometric data to validate their nature. These authors found that most of the sources in their catalogue are rapid rotators, and are therefore very likely young stars for which a spin-down has not yet occurred.

To investigate the other main type of contamination, RS CVn-type stars, we did not find any paper that looks at ZTF data in a similar way to Chahal et al. (2022), but found that Martínez et al. (2022) analysed the activity cycles in RS CVn-type stars. We analysed the reported rotational periods and found that while 90% of the DR3 YSO candidates labelled RS have periods shorter than 9.41 days, 54% of the Martínez et al. (2022) RS CVn stars have periods longer than this threshold. The median period for those YSO candidates labelled RS is 3.67 days (MAD = 2.27 days), while the median period for the Martínez et al. (2022) objects is 11.26 days (MAD = 8.68 days). Therefore, we conclude that a significant portion of the objects in our sample labelled RS CVn were likely misclassified by Chen et al. (2020) and are more likely YSOs.

Among the 1202 remaining sources that might contaminate the YSO sample, 487 are rotational variables (ROTs), which are defined as a generic class of spotted stars and are scattered all over the colour–magnitude diagram (Gavras et al. 2023), and therefore cannot be considered as contamination with certainty.

3.3.4. Other catalogues

In Table 7, we summarise the results of the cross-matches with large tables from the VizieR listing a total of 4 828 588 objects of specific types. We find 11 sources that might be galaxies, 16 sources that are listed as eclipsing or ellipsoidal binaries, 32 sources that are also listed as Cepheid/RR Lyrae-type variables, 5 objects listed as Miras, and 1 star identified as a Mira-type variable. Also 36 sources are listed in the ASAS-SN catalogue of variable stars associated with objects that are certainly not YSO related. The sum of these numbers is 101 objects, which is 0.12% of the variable *Gaia* DR3 YSO candidates. We also checked how many of our sources are located in the same regions as the catalogues used for the cross-match. To do so we used the HEALPix (Górski et al. 2005) pixelisation of the sky with nside = 256 resolution. This corresponds to 1.5978967×10^{-5} sr pixel size. In some cases, the contamination in the areas covered by the different catalogues increases by an order of magnitude, but still remains at the percent level, except for the Stringer et al. (2019), where the number of cross-matches suggests a contamination level of ~15.6%. in the overlapping area.

4. Summary and conclusions

We validate the first catalogue of variable YSO candidates observed during the first 34 months of operation of the *Gaia* space telescope. As a result of the classification process, the catalogue lists 79 375 YSO candidates as part of the third *Gaia* Data Release (DR3). After analysing the different parameter distributions, such as colours, brightness, distance, and apparent positions on the sky, we conclude that the *Gaia* DR3 YSO sample contains sources that are very similar to confirmed YSOs, mainly including the lower-mass objects. By comparing the *Gaia* DR3 YSOs to catalogues listing evolutionary stages, we confirm that *Gaia* is more sensitive to the more evolved Class II/III YSOs, which are visible in the optical bands covered by *Gaia*. The estimated completeness at the very early stages of star formation, when most of the energy of YSOs is emitted at FIR and MIR wavelengths, is close to zero, but it can range from a few percent to ~40% for objects with transitional discs, depending on the distance of the star forming regions they are located in. The lower limit of contamination level is at least 0.6% (according to SIMBAD), but the upper limit is 26.7%. More realistically, the contamination is lower than the upper limit, but still around the 10% level. Estimates of the contamination based on the cross-match with several other catalogues containing specific object types are even lower (below 0.2%), but these objects are mostly located at higher galactic latitudes where only a small fraction of our sources are located. The estimated number of potentially new YSOs presented in the *Gaia* DR3 variable star catalogue is on the order of a few tens of thousands of objects, but below 44 651. These objects were not classified as YSOs in the YSO catalogues we used for the validation and were not classified as a different type of object by the other catalogues used in this study.

Table 7. Catalogues cross-matched with the *Gaia* DR3 YSO sample in order to quantify its contamination.

Paper title	Reference	Nr. of entries in the catalogue	Nr. of *Gaia* DR3 YSOs in the same region	Nr. of matches with *Gaia* DR3 YSOs
HYPERLEDA I. Catalogue of galaxies	Paturel et al. (2003)	983 261	7600	2
6dF galaxy survey final redshift release	Jones et al. (2009)	124 647	1867	0
The Half Million Quasars (HMQ) catalogue	Flesch (2015)	510 764	988	0
OGLE fundamental-mode RR Lyrae in Galactic Bulge RRab sample	Collinge et al. (2006)	1913	334	0
Galactic bulge eclipsing and ellipsoidal binaries	Soszyński et al. (2016)	450 598	5677	16
Identification of RR Lyrae stars from the DES	Stringer et al. (2019)	713 816	192	30
JHKs photometry of VVV RR Lyrae stars	Dékány et al. (2018)	1892	965	0
OGLE Galactic center Cepheids and RR Lyrae	Soszyński et al. (2017)	40 112	5319	2
SDSS quasar catalogue, fourteenth data release	Pâris et al. (2018)	526 356	345	0
General Catalogue of Galactic Carbon Stars	Alksnis et al. (2001)	6891	2994	0
LCs re-analysis of Mira variables in ASAS	Vogt et al. (2016)	2875	363	0
Mira-like variables from the KELT survey	Arnold et al. (2020)	4132	997	5
Mira Variables in the OGLE Bulge fields	Groenewegen & Blommaert (2005)	2691	345	0
Compiled catalogue of stellar data of Miras	Kharchenko et al. (2002)	1103	78	0
Extended catalogue of NSVS red AGB variable stars	Usatov & Nosulchik (2008)	794	374	1
LEDA galaxies with DENIS measurements	Paturel et al. (2005)	753 153	5917	2
PSCz catalogue	Saunders et al. (2000)	18 351	1272	7
ASAS-SN catalogue of variable stars (*)	Jayasinghe et al. (2018)	666 502	55 160	36
Groups of galaxies in 2MASS survey	Crook et al. (2007)	18 737	394	0

Notes. Column 1 lists the titles of the corresponding papers. Column 2 lists the references to the papers. Column 3 gives the number of sources in the given catalogue. Column 4 is the number of *Gaia* DR3 YSO candidates that are found in the same HEALpixels as the sources of the given catalogue are located. Column 5 is the number of cross-matching *Gaia* DR3 YSOs found in the catalogue. (*)The ASAS-SN catalogue provides a classification for each of their sources. Herein, we take into account sources that should not be considered as YSOs, such as CWB, EA, EB, EW, GCAS, and RCB: variables.

References

Alksnis, A., Balklavs, A., Dzervitis, U., et al. 2001, Balt. Astron., 10, 1
Arnold, R. A., McSwain, M. V., Pepper, J., et al. 2020, ApJS, 247, 44
Audard, M., Ábrahám, P., Dunham, M. M., et al. 2014, in Protostars and Planets VI, eds. H. Beuther, R. S. Klessen, C. P. Dullemond, & T. Henning, 387
Bailer-Jones, C. A. L., Rybizki, J., Fouesneau, M., Demleitner, M., & Andrae, R. 2021, AJ, 161, 147
Breiman, L. 2001, Mach. Learn., 45, 5
Carpenter, J. M., Hillenbrand, L. A., & Skrutskie, M. F. 2001, AJ, 121, 3160
Chahal, D., de Grijs, R., Kamath, D., & Chen, X. 2022, MNRAS, 514, 4932
Chen, T., & Guestrin, C. 2016, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16 (New York, NY, USA: ACM), 785
Chen, X., Wang, S., Deng, L., et al. 2020, ApJS, 249, 18
Collinge, M. J., Sumi, T., & Fabrycky, D. 2006, ApJ, 651, 197
Covey, K. R., Lada, C. J., Román-Zúñiga, C., et al. 2010, ApJ, 722, 971
Crook, A. C., Huchra, J. P., Martimbeau, N., et al. 2007, ApJ, 655, 790
Dékány, I., Hajdu, G., Grebel, E. K., et al. 2018, ApJ, 857, 54
Eyer, L., Mowlavi, N., Evans, D. W., et al. 2017, ArXiv e-prints [arXiv:1702.03295]
Eyer, L., Audard, M., Holl, B., et al. 2023, A&A, 674, A13 (*Gaia* DR3 SI)
Fischer, W. J., Hillenbrand, L. A., Herczeg, G. J., et al. 2022, ArXiv e-prints [arXiv:2203.11257]
Flesch, E. W. 2015, PASA, 32, e010
Gaia Collaboration (Prusti, T., et al.) 2016, A&A, 595, A1
Gaia Collaboration (Vallenari, A., et al.) 2023, A&A, 674, A1 (*Gaia* DR3 SI)
Gavras, P., Rimoldini, L., Nienartowicz, K., et al. 2023, A&A, 674, A22 (*Gaia* DR3 SI)
Getman, K. V., Feigelson, E. D., Luhman, K. L., et al. 2009, ApJ, 699, 1454
Górski, K. M., Hivon, E., Banday, A. J., et al. 2005, ApJ, 622, 759
Groenewegen, M. A. T., & Blommaert, J. A. D. L. 2005, A&A, 443, 143
Großschedl, J. E., Alves, J., Meingast, S., et al. 2018, A&A, 619, A106
Großschedl, J. E., Alves, J., Teixeira, P. S., et al. 2019, A&A, 622, A149
Guarcello, M. G., Damiani, F., Micela, G., et al. 2010, A&A, 521, A18
Herbig, G. H., & Bell, K. R. 1988, Third Catalog of Emission-Line Stars of the Orion Population : 3: 1988 (Santa Cruz: Lick Observatory)
Herbst, W., & Shevchenko, V. S. 1999, AJ, 118, 1043
Hillenbrand, L. A., & Findeisen, K. P. 2015, ApJ, 808, 68
Hillenbrand, L. A., Hoffer, A. S., & Herczeg, G. J. 2013, AJ, 146, 85
Hodgkin, S. T., Harrison, D. L., Breedt, E., et al. 2021, A&A, 652, A76
Jayasinghe, T., Kochanek, C. S., Stanek, K. Z., et al. 2018, MNRAS, 477, 3145
Jones, D. H., Read, M. A., Saunders, W., et al. 2009, MNRAS, 399, 683
Kharchenko, N., Kilpio, E., Malkov, O., & Schilbach, E. 2002, A&A, 384, 925
Kochanek, C. S., Shappee, B. J., Stanek, K. Z., et al. 2017, PASP, 129, 104502
Koenig, X. P., & Leisawitz, D. T. 2014, ApJ, 791, 131
Kuhn, M. A., de Souza, R. S., Krone-Martins, A., et al. 2021, ApJS, 254, 33
Mamajek, E. E., Meyer, M. R., & Liebert, J. 2002, AJ, 124, 1670
Marrese, P. M., Marinoni, S., Fabrizio, M., & Altavilla, G. 2019, A&A, 621, A144
Martínez, C. I., Mauas, P. J. D., & Buccino, A. P. 2022, MNRAS, 512, 4835
Marton, G., Ábrahám, P., Szegedi-Elek, E., et al. 2019, MNRAS, 487, 2522
Masci, F. J., Laher, R. R., Rusholme, B., et al. 2019, PASP, 131, 018003
Morales-Calderón, M., Stauffer, J. R., Hillenbrand, L. A., et al. 2011, ApJ, 733, 50

Ochsenbein, F., Bauer, P., & Marcout, J. 2000, A&AS, 143, 23
Pâris, I., Petitjean, P., Aubourg, É., et al. 2018, A&A, 613, A51
Paturel, G., Petit, C., Prugniel, P., et al. 2003, A&A, 412, 45
Paturel, G., Vauglin, I., Petit, C., et al. 2005, A&A, 430, 751
Peneva, S. P., Semkov, E. H., Munari, U., & Birkle, K. 2010, A&A, 515, A24
Planck Collaboration X. 2016, A&A, 594, A10
R Core Team 2018, R: A Language and Environment for Statistical Computing (Vienna, Austria: R Foundation for Statistical Computing)
Rebull, L. M., Guieu, S., Stauffer, J. R., et al. 2011, ApJS, 193, 25
Reipurth, B. 2008a, Handbook of Star Forming Regions, Volume I: The Northern Sky (San Francisco, CA: ASP)
Reipurth, B. 2008b, Handbook of Star Forming Regions, Volume II: The Southern Sky (San Francisco, CA: ASP)
Reipurth, B., Aspin, C., Beck, T., et al. 2007, AJ, 133, 1000
Rimoldini, L., Eyer, L., Audard, M., et al. 2022, Gaia DR3 documentation Chapter 10: Variability, European Space Agency; Gaia Data Processing and Analysis Consortium, https://gea.esac.esa.int/archive/documentation/GDR3/Data_analysis/chap_cu7var/
Rimoldini, L., Holl, B., Gavras, P., et al. 2023, A&A, 674, A14 (Gaia DR3 SI)
Saunders, W., Sutherland, W. J., Maddox, S. J., et al. 2000, MNRAS, 317, 55
Schneider, N., Ossenkopf, V., Csengeri, T., et al. 2015, A&A, 575, A79

Shappee, B. J., Prieto, J. L., Grupe, D., et al. 2014, ApJ, 788, 48
Skrutskie, M. F., Cutri, R. M., Stiening, R., et al. 2006, AJ, 131, 1163
Soszyński, I., Pawlak, M., Pietrukowicz, P., et al. 2016, Acta Astron., 66, 405
Soszyński, I., Udalski, A., Szymański, M. K., et al. 2017, Acta Astron., 67, 297
Straižys, V., & Kazlauskas, A. 2010, Balt. Astron., 19, 1
Stringer, K. M., Long, J. P., Macri, L. M., et al. 2019, AJ, 158, 16
Sung, H., Bessell, M. S., Chun, M.-Y., et al. 2017, ApJS, 230, 3
Szegedi-Elek, E., Kun, M., Reipurth, B., et al. 2013, ApJS, 208, 28
Taylor, M. B. 2005, in Astronomical Data Analysis Software and Systems XIV, eds. P. Shopbell, M. Britton, & R. Ebert, ASP Conf. Ser., 347, 29
Thé, P. S., de Winter, D., & Pérez, M. R. 1994, A&AS, 104, 315
Usatov, M., & Nosulchik, A. 2008, Open Eur. J. Var. Stars, 0087, 1
Varga-Verebélyi, E., Kun, M., Szegedi-Elek, E., et al. 2020, in Origins: From the Protosun to the First Steps of Life, eds. B. G. Elmegreen, L. V. Tóth, & M. Güdel, 345, 378
Vieira, S. L. A., Corradi, W. J. B., Alencar, S. H. P., et al. 2003, AJ, 126, 2971
Vogt, N., Contreras-Quijada, A., Fuentes-Morales, I., et al. 2016, ApJS, 227, 6
Wenger, M., Ochsenbein, F., Egret, D., et al. 2000, A&AS, 143, 9
Wolk, S. J., Spitzbart, B. D., Bourke, T. L., et al. 2008, AJ, 135, 693
Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., et al. 2010, AJ, 140, 1868
Zucker, C., Speagle, J. S., Schlafly, E. F., et al. 2019, ApJ, 879, 125

Appendix A: KYSO - The Konkoly Optical YSO catalogue

We compiled a large catalogue of optically detected, bona fide YSOs.[3] The main source of our compilation was the Handbook of Star Forming Regions (Volumes I and II, Reipurth 2008a,b) which consists of 62 chapters, with each chapter describing a region of the sky (mostly one constellation or a part of it) and not a well-defined star forming region. All of the included regions are located within ~2 kpc of the Sun. Most chapters contain a list of YSOs located in the specific star forming regions, but sometimes only references to the catalogues are given. In the latter case, we took the list of the YSOs from the original (discovery) papers. We also performed an extensive literature search to include YSO catalogues published after the Handbook of Star Forming Regions. Furthermore, we included the stars of the comprehensive Herbig and Bell (1988) catalogue of pre-main sequence stars, and Herbig Ae/Be stars from Thé et al. (1994) and Vieira et al. (2003). The celestial distribution of the KYSOs is shown in Fig. A.1.

Most of the stars of the KYSO catalogue were classified as YSOs based on optical spectra, i.e. they exhibited strong emission lines and/or strong lithium absorption. Moreover, we included Hα emission stars located in star forming regions and detected by slitless spectroscopy. In addition to the spectroscopically identified YSOs, we included a few datasets containing optical counterparts of candidate YSOs selected by IR and X-ray observations. These are listed as follows: (i) IR variable stars detected by Carpenter et al. (2001) in the Orion A cloud; (ii) optically visible low-mass stars classified as YSOs based on combined X-ray and IR criteria by Getman et al. (2009) in Cepheus B, and by Wolk et al. (2008) in RCW 108; (iii) optical counterparts of IR sources in Barnard 59, classified as YSOs by IR spectroscopy (Covey et al. 2010); (iv) candidate YSOs in the Camelopardalis region, classified by optical–IR SED (Straižys & Kazlauskas 2010).

The individual YSO catalogues that we collected are very diverse regarding data structure, detection methods, limiting magnitude, and angular resolution. Our aim is to create a unified database from these various sources. As a first step, we extracted the celestial coordinates and names of the objects from the individual catalogues. In the case of older measurements, we converted the epoch from B1950 to J2000 and checked the positions in original finding charts. These data were then loaded into a database using a common scheme. Apart from the name, coordinates, and the detection method, no other information was used from the YSO catalogues.

A major issue with the data is that a source can be present in several catalogues under different names and with slightly different coordinates. For the detection of duplicate sources, we applied a semi-automatic approach. The automatic part of the process was the identification of clusters of objects within a search radius of 1″5. Then we manually examined the duplicate candidates. We checked the probability of being a duplicate based on the distribution of positions from the various catalogues, known binarity, and the naming of objects from the SIMBAD database. Wh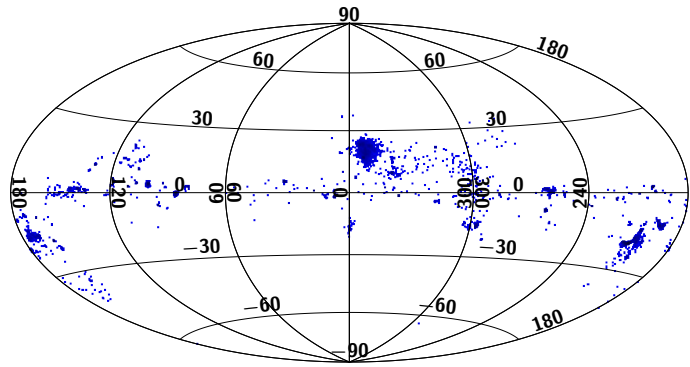en a duplicate was found, we removed the duplicate entries from our database and retained the data from the most recent catalogue. Binary stars can also appear as duplicate candidates. Three criteria were used in the identification of multiple stars: (1) When duplicate candidates are present in the same catalogue, this suggests that the sources are indeed distinct, probably binaries. (2) In other cases, the names of the objects indicate binarity. (3) The source is present in the Young Visual Binary Star Database[4].

Next, we consistently associated all of the KYSO objects with their parent star forming regions. Generally, we took the names of the regions from the individual papers describing the catalogues. In some cases, we found inconsistencies in the naming, when, e.g. the same region had different designations. To resolve these issues, we chose the naming convention present in SIMBAD. In the case of some smaller regions, where there is a known larger star forming region to which the given smaller region belongs, we used both names in the following arrangement: 'Large star forming region: Small region' (e.g. 'Rosette Complex: NGC 2244'). In total, the KYSO catalogue contains 124 star forming regions.

We supplemented the catalogue with the columns of YSO type and variability type of the sources. YSO types are as follows: CTT* – classical T Tauri star, K–M type stars with emission spectra and accretion disc, GTT* – similar to CTT* with G0–K0 spectral type (see Herbst & Shevchenko 1999), IMTT* – intermediate-mass TT* similar to CTT* with F5–F8 spectral type; WTT* – weak-line T Tauri stars identified by G–M spectral types and strong Li I absorption line at 6707 Å. PTT* is for post-T Tauri stars (Mamajek et al. 2002). HAeBe stars are B–F5 type pre-main sequence stars. FU Ori and EX Lupi type stars are regarded as distinct classes. TT* is assigned when no data were available on the accretion; Y*O is assigned when this was the only information. We have not included normal, main sequence OB stars, but included a few objects classified as high-mass YSOs (HMY*O). Variability data were obtained from an extensive search in VizieR. In addition to the published variability types, we inspected the light curves available in various VizieR catalogues. A previous version of the catalogue is described in Varga-Verebélyi et al. (2020).

Fig. A.1. Konkoly Optical YSO (KYSO) catalogue of optically selected young stars presented in Aitoff projection in Galactic coordinates.

---

[3] The KYSO table is only available in electronic form at the CDS via anonymous ftp to cdsarc.cds.unistra.fr (130.79.128.5) or via https://cdsarc.cds.unistra.fr/cgi-bin/qcat?J/A+A/.

[4] http://www2.lowell.edu/users/lprato/YBIN/Binary_Star_Database.html

Appendix B: Parameter distributions

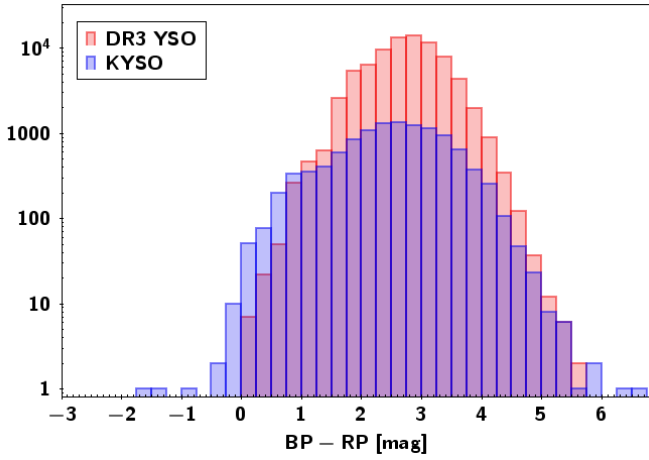## B.1. $G_{BP} - G_{RP}$ distributions

Fig. B.1. $G_{BP}-G_{RP}$ colour distribution of the *Gaia* DR3 YSOs (red bars) and KYSOs (blue bars) on a logarithmic scale. KYSOs show an excess towards bluer colours.
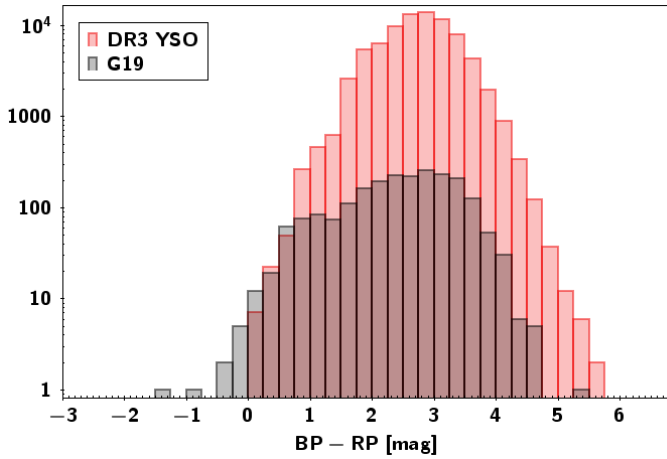

Fig. B.2. Same as Fig. B.1 but for the *Gaia* DR3 YSOs (red bars) and the Großschedl et al. (2019) G19 YSOs (grey bars).
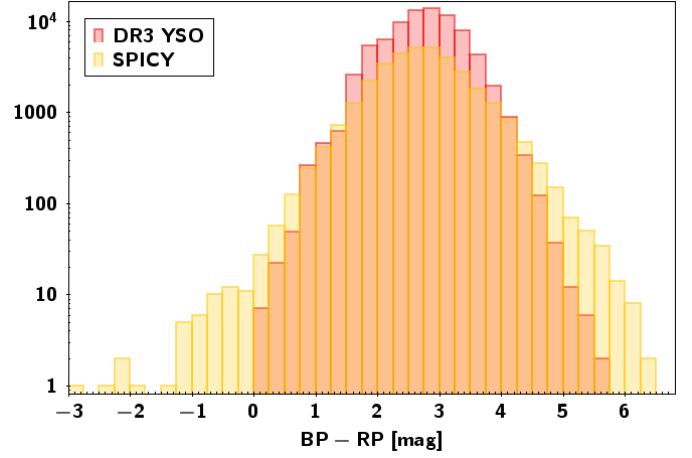

Fig. B.3. Same as Fig. B.1 but for the *Gaia* DR3 YSOs (red bars) and the Kuhn et al. (2021) SPICY YSOs (yellow bars).
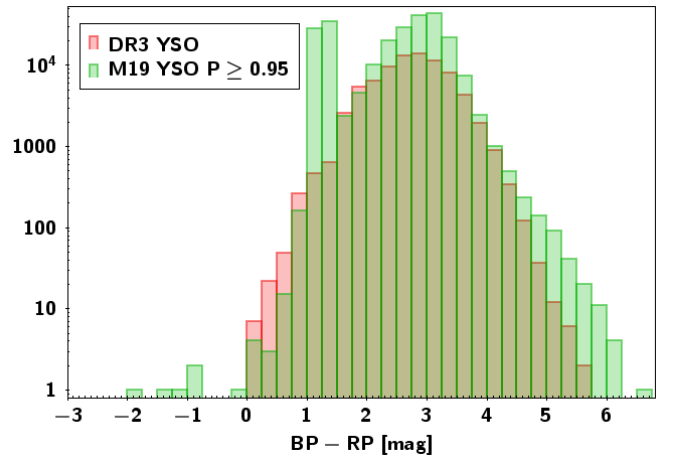
Fig. B.4. Same as Fig. B.1 but for the *Gaia* DR3 YSOs (red bars) and the Marton et al. (2019) M19 YSOs with $R \geq 0.5$ and $LY \geq 0.95$ or $R < 0.5$ and $SY \geq 0.95$ (green bars). The M19 YSOs show a significant excess at $G_{BP} - G_{RP} = 1$. These are all YSO candidates seen towards the Galactic midplane.

B.2. Absolute median G
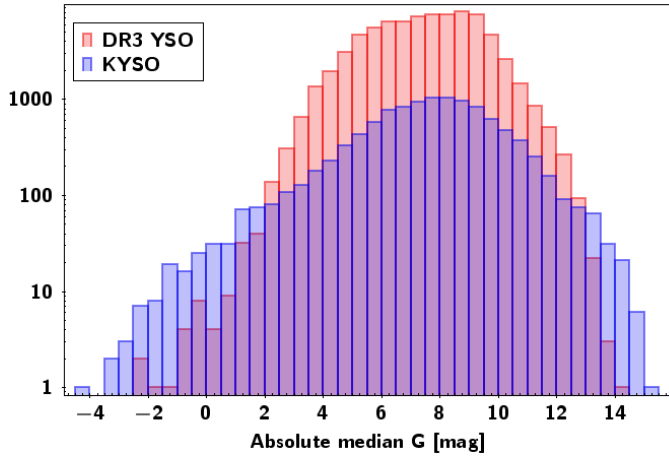
Fig. B.5. *Gaia* absolute *G*-band magnitude distribution of the *Gaia* DR3 YSOs (red bars) and KYSOs (blue bars) on a logarithmic scale. KYSOs show a slight excess towards both ends, showing that very faint and bright sources were not classified as YSOs.
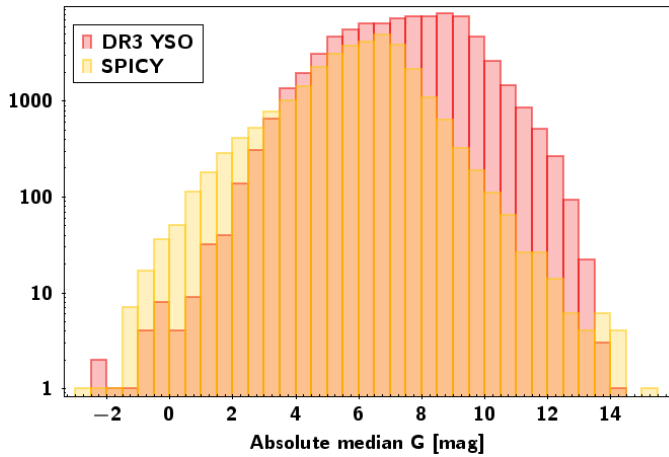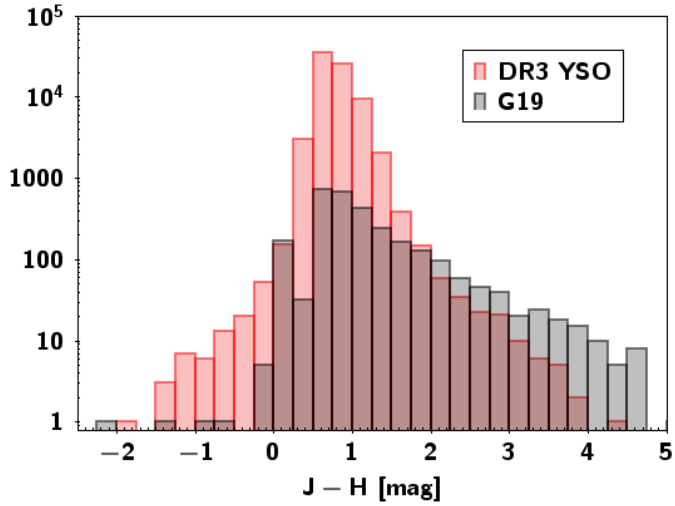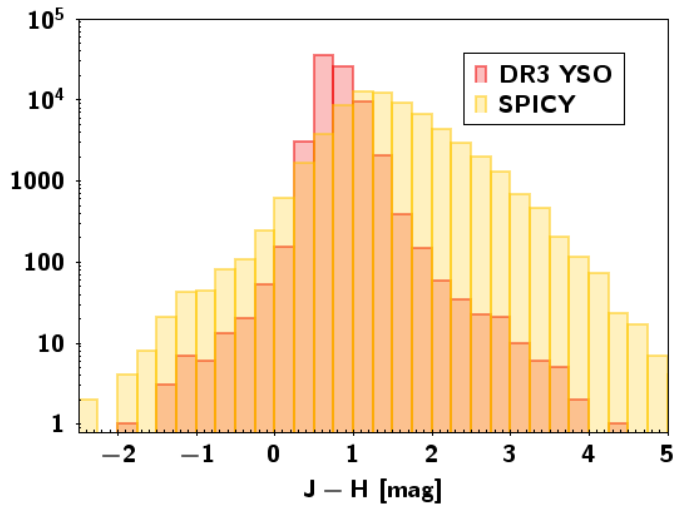

Fig. B.6. Same as Fig. B.5, but for the *Gaia* DR3 YSOs (red bars) and the G19 YSOs (grey bars).

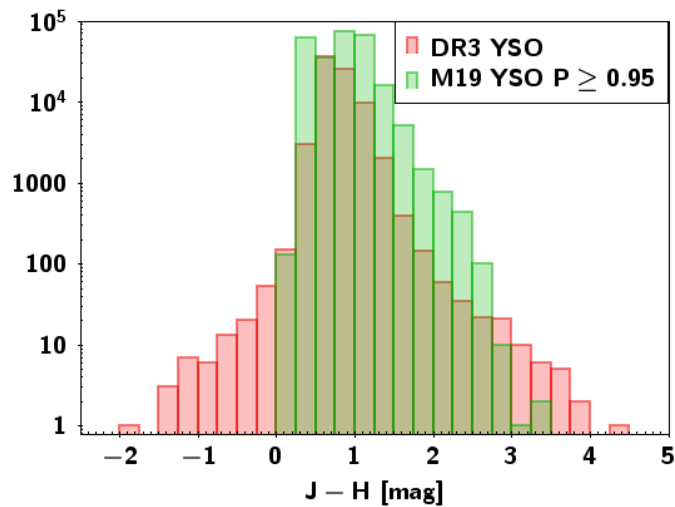**Fig. B.7.** Same as Fig. B.5, but for the *Gaia* DR3 YSOs (red bars) and the SPICY YSOs (yellow bars).
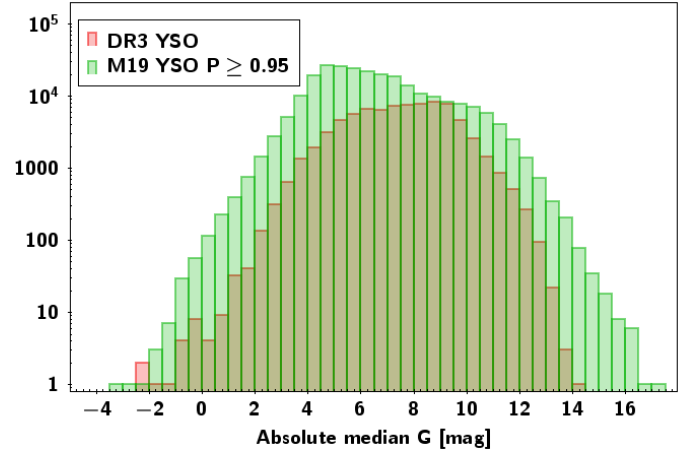
Fig. B.8. Same as Fig. B.5, but for the *Gaia* DR3 YSOs (red bars) and the M19 YSOs (green bars).
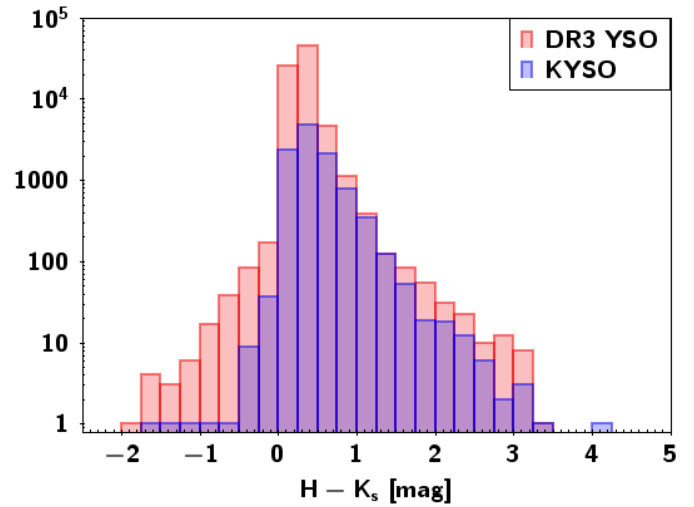
B.3. J-H 2MASS colour
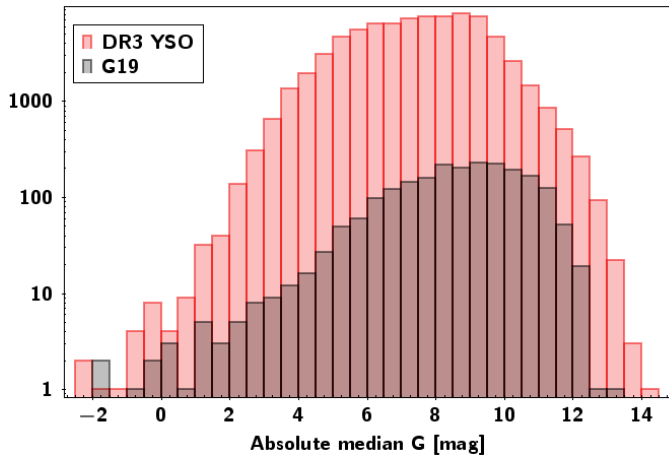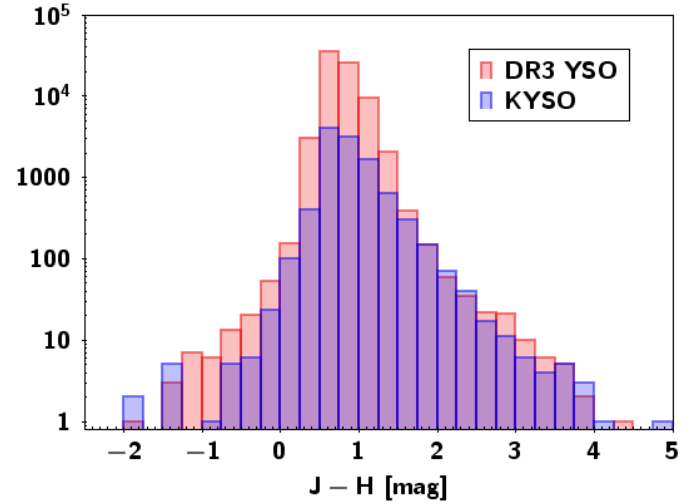
Fig. B.9. 2MASS J−H colour distribution of the *Gaia* DR3 YSOs (red bars) and KYSOs (blue bars) on a logarithmic scale.

Fig. B.10. Same as Fig. B.9, but for the *Gaia* DR3 YSOs (red bars) and the G19 YSOs (grey bars).

Fig. B.11. Same as Fig. B.9, but for the *Gaia* DR3 YSOs (red bars) and the SPICY YSOs (yellow bars).

Fig. B.12. Same as Fig. B.9, but for the *Gaia* DR3 YSOs (red bars) and the M19 YSOs (green bars).

B.4. H-K$_s$ 2MASS colour

Fig. B.13. 2MASS H−K$_s$ colour distribution of the *Gaia* DR3 YSOs (red bars) and KYSOs (blue bars) on a logarithmic scale.
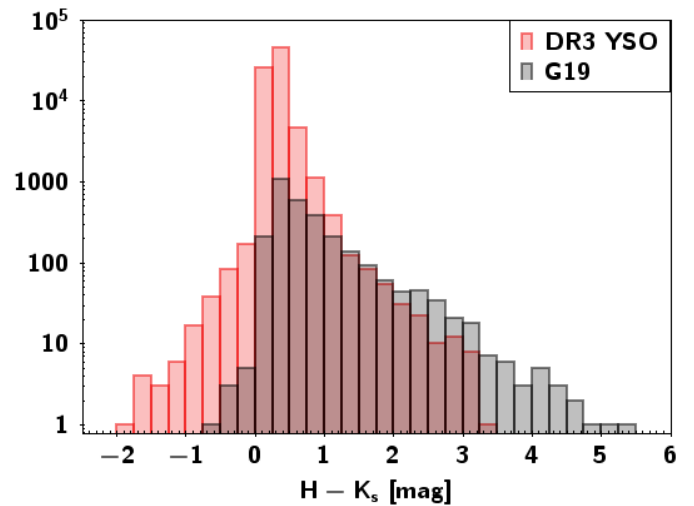

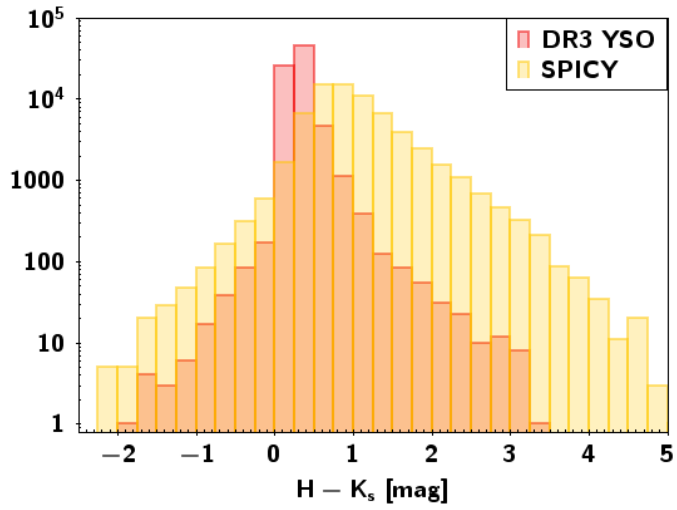**Fig. B.14.** Same as Fig. B.13, but for the *Gaia* DR3 YSOs (red bars) and the G19 YSOs (grey bars).

Fig. B.15. Same as Fig. B.13, but for the *Gaia* DR3 YSOs (red bars) and the SPICY YSOs (yellow bars).
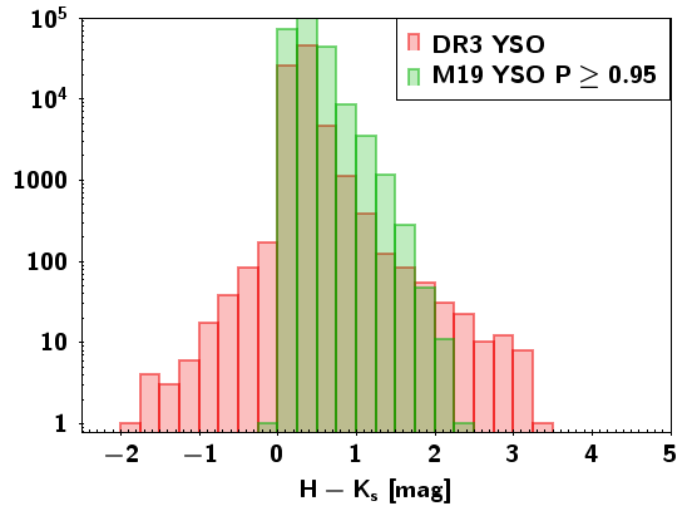
Fig. B.16. Same as Fig. B.13, but for the *Gaia* DR3 YSOs (red bars) and the M19 YSOs (green bars).