Astronomy & Astrophysics

Catalogue of new Herbig Ae/Be and classical Be stars[*]

A machine learning approach to *Gaia* DR2

M. Vioque[1,2], R. D. Oudmaijer[1], M. Schreiner[3], I. Mendigutía[4], D. Baines[5], N. Mowlavi[6], and R. Pérez-Martínez[2]

[1] School of Physics and Astronomy, University of Leeds, Leeds LS2 9JT, UK
   e-mail: pymvdl@leeds.ac.uk
[2] Ingeniería de Sistemas para la Defensa de España (Isdefe), XMM/Newton Science Operations Centre, ESA-ESAC Campus, PO Box 78, 28691 Villanueva de la Cañada, Madrid, Spain
[3] Desupervised, Desupervised ApS, Njalsgade 76, 2300 Copenhagen, Denmark
[4] Centro de Astrobiología (CSIC-INTA), Departamento de Astrofísica, ESA-ESAC Campus, PO Box 78, 28691 Villanueva de la Cañada, Madrid, Spain
[5] Quasar Science Resources for ESA-ESAC, ESAC Science Data Centre, PO Box 78, 28691 Villanueva de la Cañada, Madrid, Spain
[6] Department of Astronomy, University of Geneva, Ch. des Maillettes 51, 1290 Versoix, Switzerland

ABSTRACT

Context. The intermediate-mass pre-main sequence Herbig Ae/Be stars are key to understanding the differences in formation mechanisms between low- and high-mass stars. The study of the general properties of these objects is hampered by the lack of a well-defined, homogeneous sample, and because few and mostly serendipitously discovered sources are known.
Aims. Our goal is to identify new Herbig Ae/Be candidates to create a homogeneous and well defined catalogue of these objects.
Methods. We have applied machine learning techniques to 4 150 983 sources with data from *Gaia* DR2, 2MASS, WISE, and IPHAS or VPHAS+. Several observables were chosen to identify new Herbig Ae/Be candidates based on our current knowledge of this class, which is characterised by infrared excesses, photometric variabilities, and Hα emission lines. Classical techniques are not efficient for identifying new Herbig Ae/Be stars mainly because of their similarity with classical Be stars, with which they share many characteristics. By focusing on disentangling these two types of objects, our algorithm has also identified new classical Be stars.
Results. We have obtained a large catalogue of 8470 new pre-main sequence candidates and another catalogue of 693 new classical Be candidates with a completeness of $78.8 \pm 1.4\%$ and $85.5 \pm 1.2\%$, respectively. Of the catalogue of pre-main sequence candidates, at least 1361 sources are potentially new Herbig Ae/Be candidates according to their position in the Hertzsprung-Russell diagram. In this study we present the methodology used, evaluate the quality of the catalogues, and perform an analysis of their flaws and biases. For this assessment, we make use of observables that have not been accounted for by the algorithm and hence are selection-independent, such as coordinates and parallax based distances. The catalogue of new Herbig Ae/Be stars that we present here increases the number of known objects of the class by an order of magnitude.

Key words. catalogs – Hertzsprung-Russell and C-M diagrams – stars: emission-line, Be – stars: formation – stars: pre-main sequence – stars: variables: T Tauri, Herbig Ae/Be

1. Introduction

Herbig Ae/Be stars (HAeBes) are pre-main sequence (PMS) sources of intermediate-mass (canonically defined as $2 M_\odot \lesssim M \lesssim 10 M_\odot$, spectral type B, A, and F) that cover the gap between the lower-mass T-Tauri stars and the deeply embedded infrared-bright Massive Young Stellar Objects. HAeBes are thus key for understanding the properties of high-mass star formation. However, a large caveat in all of the studies dedicated to HAeBes is that 273 of them are known (108 in the master list of The et al. 1994, see Vioque et al. 2018). This is a very heterogeneous and biased set. In particular, few objects are known at the high-mass end (Herbig Be stars), with many of them having a doubtful nature as they are easily confused with classical Be stars (CBes,

rapidly rotating main sequence B stars with Keplerian gas discs, Rivinius et al. 2013). This situation contrasts with the thousands of T-Tauri stars known in the literature. As a consequence, many open problems involving high-mass star formation suffer from these biases and the lack of completeness.

For example, it is commonly accepted that T-Tauri stars accrete through magnetically-driven flows arising from the protoplanetary disc, which is truncated at a distance of a few stellar-radii (see Bouvier et al. 2007; Hartmann et al. 2016). However, higher-mass PMS objects have radiative envelopes and hence normally present negligible magnetic fields (Alecian et al. 2013; Villebrun et al. 2019). Therefore, the magnetospheric accretion model probably cannot apply to them. The transition from magnetospheric accretion to the still unknown accretion mechanism for higher-mass PMS objects takes place within the mass range of the Herbig Ae/Be stars. Indeed, near-IR interferometric (e.g. Monnier et al. 2005), optical- and near-UV spectro-polarimetric (e.g. Ababakr et al. 2017), and spectro-photometric observations (e.g. Mendigutía et al. 2011; Fairlamb et al. 2015;

Wichittanakom et al. 2020) have shown that the lower mass Herbig Ae stars show accretion signatures consistent with T-Tauri stars, whereas Herbig Be stars appear to be inconsistent with magnetospheric accretion. A large caveat in these studies is that they do not include the less evolved sources in high-mass PMS tracks (most Herbig Be stars observed to date are very close to the main sequence, Vioque et al. 2018), which are obviously of paramount importance for understanding high-mass accretion. In addition, there is observational evidence that points towards differences between the discs of low- and high-mass PMS sources. This can be seen in the amount of infrared excess, which is much lower for high-mass sources (Ribas et al. 2015; Vioque et al. 2018; Arun et al. 2019) or in morphology; for instance, spirals have only been found in early spectral type stars (Garufi et al. 2018). Similarly, there is a clear observational bias in these results, as so far mostly long-lived, massive discs around low-mass stars have been observed.

Independently, it is known that high-mass stars tend to form in clusters (Hillenbrand et al. 1995; Testi et al. 1999). Studies of massive field runaway stars have shown that at least a small fraction (∼4%, de Wit et al. 2005) of O-type stars are formed without a cluster environment. Nonetheless, recent publications question the existence of isolated high-mass star formation (e.g. Stephens et al. 2017). Again, the scarcity of known high-mass PMS sources makes the statistics non-robust.

It is thus useful to obtain a large homogeneous and low biased catalogue of new Herbig Ae/Be stars. *Gaia* Data Release 2 (DR2, Gaia Collaboration et al. 2016, 2018) provides a five dimensional astrometric solution for up to $G \lesssim 21$ mag (white G band, described in Evans et al. 2018) to over 1.3 billion objects (Lindegren et al. 2018). This large dataset allows for exploitation with statistical learning techniques (as done in e.g. Marton et al. 2019 or Cánovas et al. 2019; see Baron 2019 for a general description of these techniques into astronomy). In this paper we use an algorithm based on an artificial neural network (ANN) to identify new Herbig Ae/Be stars within *Gaia* DR2. ANNs are supervised learning classifiers, this means that they need to be trained with a list of known sources (training set) that have a set of characteristics (features) and a label (ground truth) that assign them to a certain category (e.g. a stellar class). Once trained, ANNs assign probabilities of belonging to every one of the chosen categories to each input source. The known HAeBes constitute a small, biased, and contaminated set. In order to achieve a good training performance the strategy adopted was to include T-Tauri stars in the training and use an algorithm focusing on the high-mass end. In the resulting catalogue of new PMS candidates, the most massive ones can be further selected by means of the Hertzsprung-Russell (HR) diagram.

The features that feed the ANN need to be relevant for identifying PMS sources. Hence, we want the features to trace the main observational characteristics of PMS sources, which are: infrared (IR) excesses, because of the radiation of the heated up protoplanetary disc, emission lines, that trace the surrounding material close to the forming star, and photometric variability. This PMS variability is caused by the presence of the disc in the line of sight (e.g. dippers, Bouvier et al. 1999, or UX Ori type sources, Vioque et al. 2018), because of episodic accretion events (EX Lup or FU Ori type sources, Cody et al. 2017), or pulsations due to internal instability (Zwintz et al. 2014). To feed the algorithm with these characteristics, we use observables belonging to five different surveys: *Gaia* DR2 for variability, 2MASS (Skrutskie et al. 2006) and WISE (Wright et al. 2010) for near- and mid-IR excess respectively, and IPHAS (Drew et al. 2005; Barentsen et al. 2014) and

VPHAS+ (Drew et al. 2014) for Hα emission. If HAeBes were unique in these properties, a simple linear separation in the parameter space would suffice for identifying more objects of the class (e.g. in a colour-colour plot). However, HAeBes share these characteristics with other types of objects, of which classical Be stars stand out, as their outwardly diffusing gaseous discs generate very similar observables (Grundstrom & Gies 2006; Rivinius et al. 2013; Klement et al. 2017). Therefore, our ANN-based algorithm spotlights on disentangling these two types of objects, and as a consequence we also find new classical Be candidates.

The paper is organised as follows: in Sect. 2, we describe the observables, features, and the metrics used for evaluating the performance of the algorithm as well as the sources that the algorithm classifies once it is trained. In Sect. 3 we present the labelled sources used for training the ANN, in Sect. 4 we describe and evaluate the output of the algorithm which we analyse in Sect. 5, describing its flaws and biases. Section 6 summarises the main conclusions. The algorithm itself is detailed in Appendix B.

2. Observables, features, and data

The features are the individual properties or characteristics that are used by the ANN to learn how to classify new sources. Feature selection is important, as the use of useless features or the lack of very relevant ones for differentiating the categories can heavily affect the performance of the algorithm.

2.1. Observables

As described in Sect. 1, we need observables contained within the catalogues *Gaia* DR2, 2MASS, WISE, and IPHAS and VPHAS+. These catalogues have information in several passbands ranging from the optical to the mid-infrared. We used the following passbands: from *Gaia* DR2, the broad white G band (0.59 μm), and the blue (G_{BP}) and red (G_{RP}) bands (0.50 μm and 0.77 μm respectively). A description of the *Gaia* filters can be found in Evans et al. (2018). From IPHAS and VPHAS+, we used the SLOAN passband r (0.62 μm) together with the $H\alpha$ narrow filter (0.66 μm). A description of the IPHAS passbands and associated footprints can be found in Drew et al. (2005) and Barentsen et al. (2014) (for the second data release that we are using) and in Drew et al. (2014) for VPHAS+. Finally, from 2MASS we used the three passbands J, H, and K_s (1.24 μm, 1.66 μm and 2.16 μm respectively) and from WISE the four passbands $W1$, $W2$, $W3$, and $W4$ (3.4 μm, 4.6 μm, 12 μm, and 22 μm respectively). These passbands of 2MASS and WISE were obtained from the AllWISE catalogue, which is described in Cutri et al. (2013).

It is important when setting up the features to be cautious about introducing unwanted bias regarding the selection we want to perform. An example of an unwanted bias is, for example, to introduce distance as a feature. Most of the known PMS objects are close-by because it is easier to study bright objects. If we introduce a distance dependent feature the algorithm would work with the idea that being close is an intrinsic property of PMS objects, and it would be biased to find PMS objects that are nearby. In addition, if we introduce position dependent features any posterior analysis about the clustering properties of Herbig Ae/Be stars would be biased towards the selected preferred positions of the training data. Therefore, we set the features to be distance and position independent, which implies that most of the observables used are colours. Of course, there

are unwanted biases in the resulting catalogues because of the selected features, and they are addressed in Sect. 5. For example, interstellar extinction results in colours that are not strictly distance-independent, and by demanding to have detections in all the WISE bands we are biasing ourselves to the most extreme IR-bright sources.

In total, we chose 48 observables from *Gaia* DR2, 2MASS, WISE, and IPHAS and VPHAS+ data. The colours used are $r - H\alpha$ plus all combinations of the passbands of *Gaia* DR2, 2MASS, and WISE (i.e. $G_{BP} - G$, $G_{BP} - G_{RP}$, $G_{BP} - J$, $G_{BP} - H$, $G_{BP} - K_s$, $G_{BP} - W1$, $G_{BP} - W2$, $G_{BP} - W3$, $G_{BP} - W4$, $G - G_{RP}$, $G - J$, $G - H$, $G - K_s$, $G - W1$, $G - W2$, $G - W3$, $G - W4$, $G_{RP} - J$, $G_{RP} - H$, $G_{RP} - K_s$, $G_{RP} - W1$, $G_{RP} - W2$, $G_{RP} - W3$, $G_{RP} - W4$, $J - H$, $J - K_s$, $J - W1$, $J - W2$, $J - W3$, $J - W4$, $H - K_s$, $H - W1$, $H - W2$, $H - W3$, $H - W4$, $K_s - W1$, $K_s - W2$, $K_s - W3$, $K_s - W4$, $W1 - W2$, $W1 - W3$, $W1 - W4$, $W2 - W3$, $W2 - W4$, $W3 - W4$). The idea behind using all these combinations is that we do not know which colours are ideal for selecting PMS objects, so we let principal component analysis (PCA) facilitate this (see Sect. 2.2). The reason why neither r nor $H\alpha$ passbands are combined with the other passbands is explained in Sect. 3.1.1.

In addition, we constructed two observables, G_{var} and V_{htg}, that trace optical photometric variability and are based on the *Gaia* passbands. We define G_{var} as:

$$G_{var} = \frac{F'_G \; e(F_G) \; \sqrt{N_{obs,G}}}{F_G \; e'(F_G) \; \sqrt{N'_{obs,G}}},$$ (1)

where F_G and $e(F_G)$ are the *Gaia* G band flux and its associated uncertainty for a certain source and $N_{obs,G}$ the number of times that source was observed in the G band. The idea is that variable sources have larger uncertainties (weighted with the square root of the number of observations) than non-variable ones. $F'_G/e(F'_G) \sqrt{N'_{obs,G}}$ refer to the median value of *Gaia* DR2 sources of the same brightness. This denominator is necessary as non-variable objects of different brightness show different median uncertainties. A similar indicator was used in Vioque et al. (2018) to study the variability of known Herbig Ae/Be stars. It was evidenced that this variability proxy mostly traces irregular (i.e. non-periodic) variabilities caused by material on the line of sight, so we expect it to be efficient in separating CBes from HAeBes. We define the heterogeneous variability (V_{htg}) as:

$$V_{htg} = \frac{F'_{B_p} \; e(F_{B_p}) \; \sqrt{N_{obs,B_p}}}{F_{B_p} \; e'(F_{B_p}) \; \sqrt{N'_{obs,B_p}}} - \frac{F'_{R_p} \; e(F_{R_p}) \; \sqrt{N_{obs,R_p}}}{F_{R_p} \; e'(F_{R_p}) \; \sqrt{N'_{obs,R_p}}}.$$ (2)

This V_{htg} observable is based on the same idea as G_{var} but it evaluates the heterogeneous variability that may be present among the blue (G_{BP}) and red (G_{RP}) filters. This may arise, for example, by circumstellar material causing irregular extinction episodes (as is the case in the reddening and blueing associated during the variations of UX Ori type stars, Grinin & Grinin 2000) or by variable accretion.

2.2. Features

We use PCA to select the optimal set of features for our problem. When applying PCA to our complete set of 48 observables we obtain 48 principal components. However, in our pipeline only 12 of those principal components carry 99.99% of the variance (see Appendix B.3). These principal components that carry

almost all of the variance of the space of observables constitute our set of features. In other words, these principal components are the features used by the ANN. PCA also removes any linear dependency between the observables.

2.3. Evaluation metrics

We use two correlated metrics, precision (P) and recall (R). They are defined as follows:

$$P = \frac{TP}{TP + FP},$$ (3)

where TP is the number of true positives, that is, the number of sources of a certain category correctly catalogued, and FP is the number of false positives, this is the number of sources of the same category wrongly classified[1]. In other words, of all objects for which we have predicted a certain category, P describes what fraction was correctly classified. Separately:

$$R = \frac{TP}{TP + FN},$$ (4)

where FN is the number of false negatives, that is, the number of sources that belong to a certain category but were not classified as such. In other words, of all objects that are actually of a certain class, R describes what fraction have we detected as belonging to that class, introducing a notion of completeness. These metrics are defined independently for each category.

2.4. Data

Before describing the training data, it is necessary to assess how many sources exist with all the observables we are using (Sample of Study, SoSt hereafter). The first step for generating this SoSt was to cross-match the catalogues that contain the required observables (*Gaia* DR2, AllWISE, IPHAS, and VPHAS+). Examples of works where this was done to a high level of accuracy are Scaringi et al. (2018) for *Gaia* DR2 with IPHAS and Marrese et al. (2019) and Wilson & Naylor (2018) for *Gaia* DR2 with AllWISE, among others. However, these cross-matches arrived at a high level of accuracy by sacrificing completeness. PMS sources in particular, because of their variability and preferred location in extincted and crowded regions, tend to be excluded in those general cross-matches (e.g. only ~52% of the known HAeBes are present in the AllWISE "Best-Neighbour table" of Marrese et al. 2019). Instead, we perform a more generous cross-match accepting that we may generate some incorrect associations.

We first cross-matched *Gaia* DR2 (using epoch 2000 adapted coordinates) with IPHAS and VPHAS+ independently with a 1 arcsec aperture because that is approximately the angular resolution of VPHAS+, IPHAS being slightly worse. We found that 95% of the sources are within 0.25 arcsec. These two catalogues present a further complication. They present different observations of the same source as different entries and hence produce duplications in the cross-match. Therefore, in those cases we chose the observation with data in all the passbands, if any. If none or more than one of the observations have information in all the passbands we chose the one with a higher quality flag and, in the case of having the same flags, we chose the object with the smaller angular distance to the *Gaia* DR2

---

[1] Other metrics, like the Area Under the Curve or the F_1 score, were discarded because FP is over-measured in this classification problem due to contamination in the training data, see Appendix B.1.

source. Similarly, whenever a *Gaia* DR2 source was present in both IPHAS and VPHAS+ we gave priority to the observation with all the passbands, followed by the one with a higher quality flag and, in the case of having the same flags, to the object with the smaller angular distance to the *Gaia* DR2 source. Then, we performed another cross-match using *Gaia* DR2 coordinates with AllWISE, using a cross-match aperture of 2 arcsec. This cross-match aperture, though large, was chosen after the experience in Vioque et al. (2018) where even a 3 arcsec aperture was still not sufficient for some HAeBes. We found that 95% of the sources are within 1.12 arcsec. This last cross-match provides us with a set of 51 548 230 sources. However, missing values are not allowed in ANNs and only 4 151 538 sources (8% of the original set) have all the 48 observables (see Sect. 2.1). This constitutes our SoSt, the master sample of all the objects with the data necessary to enter the ANN. This set has a mean of $G = 16.7 \pm 2.0$ mag (error is 1σ of the mean) so 98% of the sources are in the range $12.3 < G < 20.3$ mag. The mean parallax is $\varpi = 0.36 \pm 0.75$ mas. We note that the *Gaia* parallax is not available for all the sources. The sky footprint of the SoSt is not homogeneous as it is limited by the combined footprint of the surveys used. Primarily, IPHAS and VPHAS+ are limited to the galactic plane ($5.5° > b > −5.5°$) and VPHAS+ footprint ($29° > l > −145°$) is largely incomplete at the time of writing. In addition, spurious WISE photometric detections in the galactic plane are a known issue (Marton et al. 2019 and references therein). Furthermore, due to the *Gaia* scanning law, *Gaia* DR2 itself presents a heterogeneous footprint completeness. Finally, demanding proper detections up to W4 ($22\,\mu$m) and in the $H\alpha$ passband excludes many objects and it may be expected to have overdensities of SoSt sources around star forming regions. The impact of this footprint in the final catalogues is addressed in Sect. 5.2.

As the beams of IPHAS, VPHAS+, and AllWISE are larger than *Gaia*'s, different *Gaia* sources could have been assigned to the same IPHAS, VPHAS+, or AllWISE source. This can be the case if various *Gaia* sources are present within the same beam or if a wrong assignation was done in our generous cross-match. Indeed, 4.9% of the AllWISE sources are repeated and 0.31% of the IPHAS or VPHAS+ objects. These do not affect the classification, as the values are too small to have a significant impact on the training or the final catalogues (see Sects. 3.4 and 5.2 respectively). However, this implies that on average 1/42 (regarding AllWISE) and 1/625 (regarding IPHAS or VPHAS+) sources of the SoSt are fake, in the sense that its associated photometry does not belong to them, or it is a mixture of all the *Gaia* sources within the same beam. Another way of estimating the number of purely incorrect cross-matches is by comparing the *Gaia* passbands and colours with the AllWISE and IPHAS or VPHAS+ ones. In the case of AllWISE we compared $G_{RP} − J$ vs. $J − H$, which are strongly linearly correlated, and labelled as potential incorrect cross-matches those sources that were beyond 0.5 mag of the best linear fit. This results in about 2.3% bad matches for AllWISE. In the case of $H\alpha$ we compared G_{BP} vs. r (there is no linear relation between colours) and labelled as potential incorrect matches those sources that were beyond 1 mag (to account for variability) of the best linear fit. This results in a contamination of 1.3% for IPHAS or VPHAS+. Therefore, we conclude that the cross-matches are good to the ~98% level.

We did not take into account the quality flags of the catalogues. This decision was made for two reasons. First, IPHAS and VPHAS+ have very stringent quality indicators, and by limiting ourselves to sources with a good flag in these catalogues we reduce the size of our training set significantly (e.g. the SoSt

would be reduced to 47%). Similarly, the mid-IR colours W3 and W4 tend to have very poor quality flags. Only ~10% of the sources within *Gaia* and AllWISE with information in all passbands have reliable mid-IR measurements (Marton et al. 2019). However, in this work the mid-IR is of paramount importance and cannot be excluded, as it is where the discs around HAeBes start to differ from the dust-free discs around classical Be stars (Waters et al. 1988; Rivinius et al. 2013). Second, because introducing cuts in the training data based upon quality criteria can introduce uncontrolled biases in the subsequent selection. This is because these quality flags are a result of a combination of very different factors. It is preferable to let the ANN deal with bad quality photometry as well as contaminants. Nonetheless, these quality flags are added to the final catalogues of new PMS and CBe candidates (Sect. 4 and Tables D.1 and D.2). The consequences of using low-quality data are discussed in Sect. 5.2.

3. Labelled sources

As described in Sect. 1, we need to select which categories we want the ANN to learn how to classify. Then we need to label a set of sources as belonging to these categories and use them to train the ANN. These labels are considered as ground truth and any bias, trend or contamination of this sample inevitably results in a bias in the final classification. In this section we describe the construction of this set of Labelled Sources, which is a subset of the Sample of Study. The complementary subset of the SoSt that is not labelled (Input Set) is the one classified by the trained ANN (see Appendix B and Fig. B.1 for further details).

We use one category of PMS sources and another category of classical Be stars, as telling the difference between these two groups is the main goal of the algorithm. In addition of learning from the characteristics of PMS and CBe objects we need the algorithm to learn from the existence of other similar or distinct sources that do not belong to these categories. This includes the erroneous or spurious data present in every catalogue. In other words, we need to construct a representation of what the algorithm finds when being applied to the Input Set. Hence, we use a third category of other objects, which comprises all types of sources present within the catalogues used that are neither a PMS source nor a CBe star. Therefore, the set of Labelled Sources contains already known PMS sources (Sect. 3.1), already known CBes (Sect. 3.2), and other objects (Sect. 3.4). In the following sections the construction of these three categories is described.

All known PMS and CBe sources considered with a good astrometric solution appear on the *Gaia* HR diagram (colour vs. absolute magnitude diagram) in Fig. 1. We define as sources with a "good astrometric solution" those with a re-normalised unit weight error (RUWE parameter of *Gaia* DR2) of smaller than 1.4 and $\varpi/\sigma(\varpi) \geq 10$. Only these astrometrically well behaved sources have trustworthy positions in the HR diagram, as astrometry carries most of the uncertainty (see e.g. Vioque et al. 2018). However, those with a bad astrometric solution are still used by the algorithm as the observables are astrometry-independent (see Sect. 2.1). In this work we use the parallax to distance conversion of Bailer-Jones et al. (2018). In order to achieve the most accurate HR diagram positions we also needed to correct for extinction. Unfortunately, it cannot be totally taken into account as in general the intrinsic extinctions are unknown. However, we corrected for interstellar extinction by using the dust map of Lallement et al. (2019) and the extinction coefficients of *Gaia* of Casagrande & VandenBerg (2018). This interstellar extinction shall only be considered as a lower limit to the total extinction. This procedure for generating HR diagrams is standard
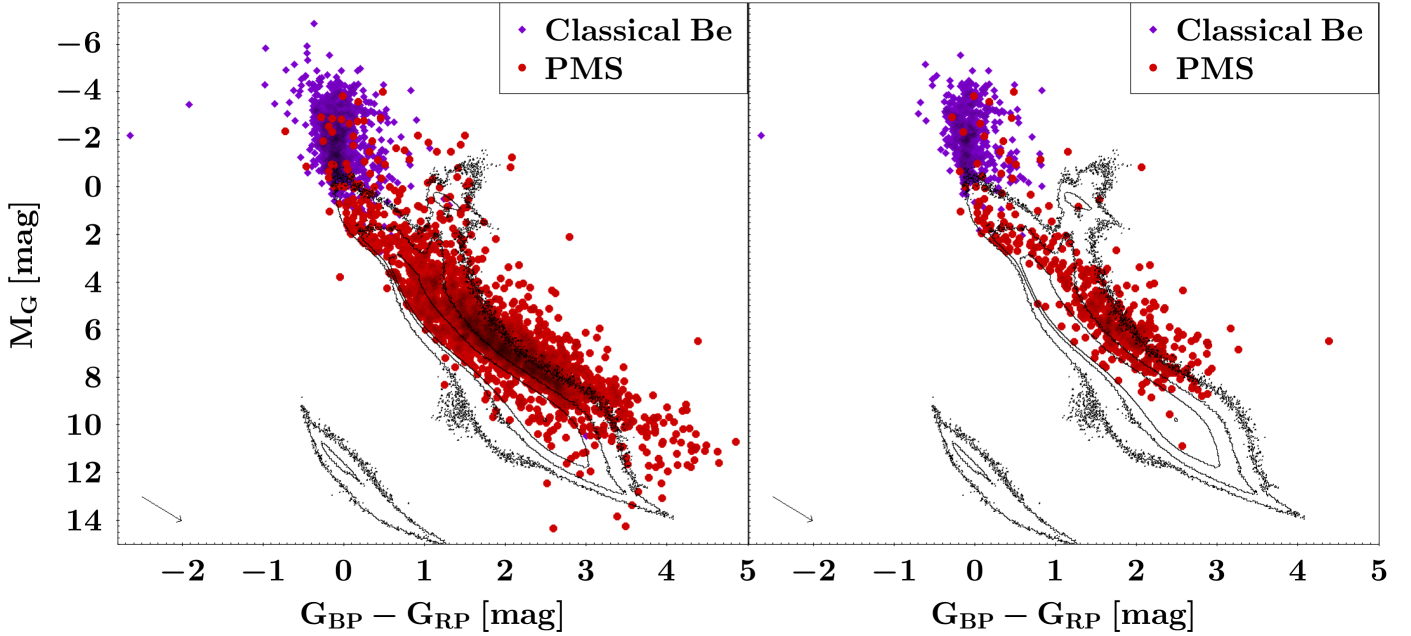
Fig. 1. *Gaia* colour vs. absolute magnitude diagram. Known PMS (in red circles) and classical Be stars (in violet diamonds) with good astrometric solution and corrected from interstellar extinction are plotted. An extinction vector corresponding to $A_G = 1$ is shown at the bottom left of each plot. Black contours trace *Gaia* sources within 500 pc with good astrometric solution. *Left*: all known sources. *Right*: subset of sources with all the observables that are used for training. The very blue classical Be star is ω CMa and it probably has a spurious G_{RP} magnitude because of being brighter than the bright limit of *Gaia* DR2.

throughout the paper, so all the HR diagrams presented can be directly compared.

3.1. PMS object category

Although for the algorithm there is just a single class of PMS objects, we create that class by combining intermediate-mass Herbig Ae/Be stars and lower mass T-Tauris, so we cover the whole mass range.

3.1.1. Herbig Ae/Be stars

Regarding the Herbig Ae/Be stars, we start with the compilation of Vioque et al. (2018) where most known HAeBes could be matched with *Gaia* DR2 data. The main issue with Herbig Ae/Be stars is that almost all of them are brighter than the bright limit of IPHAS and VPHAS+ ($12-13$ mag). Using Hα equivalent widths (EWs) we derived the IPHAS and VPHAS+ like colour $r - H\alpha$ using the synthetic tracks of Drew et al. (2005, see their Fig. 6, extinctions and effective temperatures are present in Vioque et al. 2018 and references therein). Combining the Hα EWs of Vioque et al. (2018) and Wichittanakom et al. (2020) with the few sources present in IPHAS or VPHAS+ gave us $r - H\alpha$ colour for 215 HAeBes. This is why neither r nor $H\alpha$ passbands are combined with the rest in Sect. 2.1, as we do not have them for many sources. There is a bias in this conversion from Hα EWs to $r - H\alpha$ colour because it can only be applied to those objects with observed Hα in emission above the continuum. Hence, it could not be applied to the many HAeBes with intrinsic emission filling in the underlying absorption but below the continuum level. This bias also appears later for T-Tauri stars and CBes in Sects. 3.1.2 and 3.2 and its impact is addressed in Sect. 5.2.

The cross-match with AllWISE to obtain 2MASS and WISE passbands was already performed in Vioque et al. (2018). The

final number of Herbig Ae/Be stars considered is 255, of which 163 have all observables. We did not include Massive Young Stellar Objects (Lumsden et al. 2013) in this sample as in general they are not optically visible so they are not present in *Gaia* DR2 (except those that are already in Vioque et al. 2018 list which were included in this study).

3.1.2. T-Tauri stars

To the set of intermediate-mass Herbig Ae/Be stars we add a set of T-Tauri stars to complete the low-mass regime. If we use those objects catalogued as T-Tauris in the SIMBAD database (around 3500 objects at the time of writing) we end up, after the cross-matches, with most of the objects having being catalogued by a few papers dedicated to very specific regions (e.g. Venuti et al. 2014 on NGC 2264 open cluster or Sicilia-Aguilar et al. 2013 on Tr 37). In order to minimise the possible implications due to this we add the sources of the Herbig-Bell (HB) Catalogue (Herbig & Bell 1988) which, although focused in the Orion region, has sources distributed all over the sky. We cross-matched the set of T-Tauris with *Gaia* DR2 with a 0.5 arcsec aperture (close to the 0.4 arcsec angular resolution of *Gaia* DR2). We double checked that the cross-matched sources have a similar V and G band (within ± 2 mag, the range is rather generous to avoid biasing to exclude very variable sources) when possible to discard bad cross-matches. Then, we cross-matched the *Gaia* source identifications with those of the SoSt (see Sect. 2.4) to obtain the T-Tauri stars with all the observables.

In addition, the HB catalogue provides us with Hα EWs and spectral types that allow us to derive $r - H\alpha$ colour for 297 more T-Tauris. To this end, we used the HB B-V colour, which come from simultaneous passbands at maximum brightness, and the spectral types provided by the HB catalogue to derive extinctions for these T-Tauris. Whenever B-V colours were not available we used those of the APASS survey (Henden et al. 2018) with

a 3 arcsec cross-match. A small error is introduced for objects colder than roughly a G2 V star which are typically given slightly smaller $r - H\alpha$ magnitudes than those that correspond to them (see Drew et al. 2005 for further details). The overall result is a sample of 3171 T-Tauri stars, of which 685 have information in all the observables.

3.2. Classical Be stars

For the classical Be stars, we use the Be Star Spectra Database (BeSS Database, Neiner et al. 2011) which comprises 2264 CBes. This includes the candidates of Raddi et al. (2013, 2015). To these we add 35 more CBes from Shokry et al. (2018, those they claim as secure detections). We cross-matched that catalogue with *Gaia* DR2 using a 0.5 arcsec aperture. Again, we double checked that the cross-matched sources have a similar V and G band photometry (within ± 2 mag) when possible, in order to discard bad cross-matches. Then, we cross-matched the *Gaia* source identifications with those in the SoSt (Sect. 2.4) to assess how many CBe stars are there with all the observables.

In order to increase the number of stars in this category, we complemented it with Hα EWs from the spectra available in the BeSS database. We estimated an uncertainty measuring EWs of 15%, which is probably within the intrinsic EW variations of these objects. Then, we used again the synthetic tracks of Drew et al. (2005) to transform Hα EWs to IPHAS and VPHAS+ $r - H\alpha$ colour for 442 sources. In order to do this, we used the spectral types of the BeSS database to estimate effective temperatures and, if undetermined, we estimated them from the positions in the HR diagram (Fig. 1). We assumed no extinction, which is roughly safe for this kind of object (only the faint ones from Raddi et al. 2015 suffer significantly from interstellar extinction). To assess whether this is a valid assumption we studied the extinction in the G band provided by *Gaia* DR2 for all the CBe stars for which it is available. If we take the central values we found that 94% of the sources have an A_G lower than 1.55, which is roughly the value beyond where the extinction becomes significant for the colour conversion of Drew et al. (2005). The final number of classical Be stars considered is 1992 of which 775 have information in all the observables.

3.3. Disentangling Herbig Ae/Be, CBe stars, and B[e] stars

There is some inevitable contamination between categories. For example, the set of known PMS objects is contaminated in its massive end by classical Be stars and vice-versa. Indeed, there were 15 sources that appeared both as PMS and CBe star in the previous selections. Therefore, we needed to take decisions on how to catalogue them, even though in many cases there is no clear answer in the literature. We did not exclude these objects as they are the most interesting ones for the algorithm to learn from. The particular cases and the decision made upon them are detailed in Appendix A.

In addition, within our sets of known sources there were many "unclassified B[e]" stars (also known as FS CMa stars, Miroshnichenko 2017; Arias et al. 2018). FS CMa objects are an inhomogeneous group of B stars with forbidden lines and a very unclear nature. These forbidden lines and the dust-type infrared excess exclude them from being PMS or CBe sources (Rivinius et al. 2013) and we removed them from the sets of known objects in order to not bias the results. The 17 excluded FS CMa sources are also detailed in Appendix A. As a word of caution, independently, around half of the HAeBes display the B[e] phenomenon (see Oudmaijer 2017).
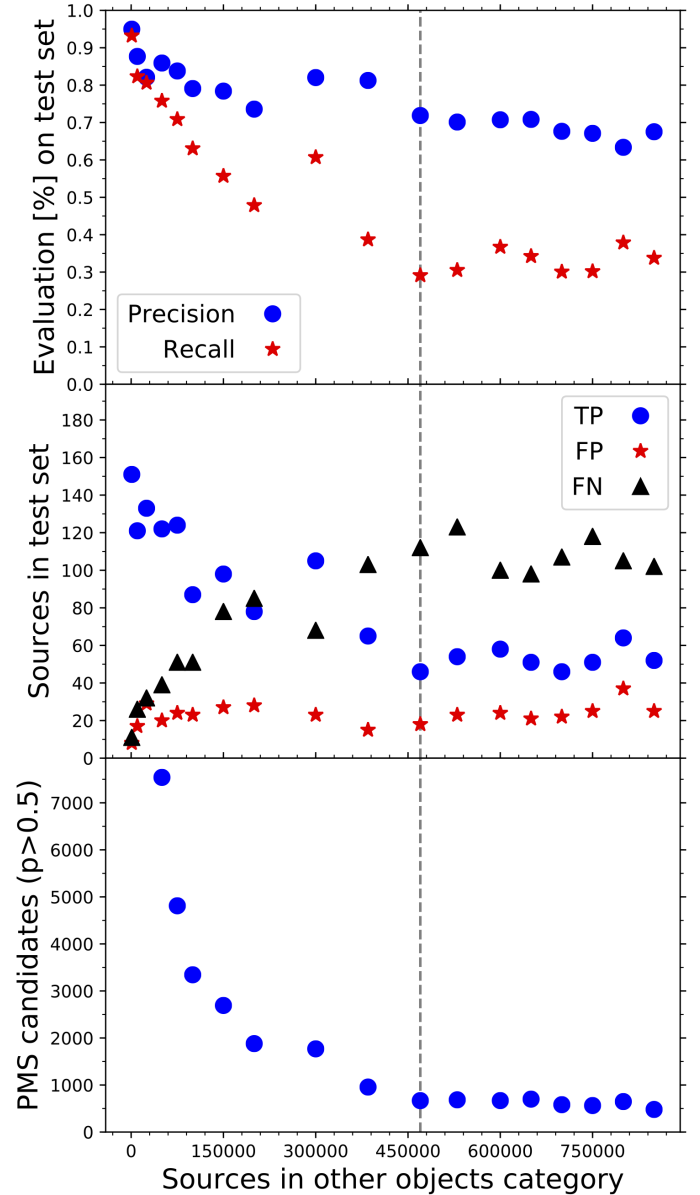
Fig. 2. Different metrics of the ANN on the PMS category vs. different sizes of the other objects category. Stabilisation point is marked with a vertical grey line. *Top*: precision and recall. As the size of the category gets larger the recall drops drastically up to the stabilisation point whereas the precision is roughly stable at all sizes. *Middle*: TP, FP, and FN. Similarly, TP and FN have equal stabilisation point whereas FP is stable for all sizes. *Bottom*: number of PMS candidates obtained when generalizing the trained ANN, we note the same stabilisation point.

3.4. Other objects

We construct the category of other objects by randomly sampling sources from the Sample of Study. We would like to have a representative set of whatever else might be present in the SoSt that is not a PMS object or a classical Be star. The question is how large this category should be in order for the algorithm to generalise properly. In other words, we want to know how many random sources from the SoSt are necessary so all populations present in the cross-matched catalogues have been represented in this category of other objects.

This can be estimated by training an ANN with different sizes of this third category, and studying how well it generalises

in each case. The size of the previously described categories is kept constant. Using an ANN (3 fully connected hidden layers of 300 neurons each) we evaluate how the precision and recall of the network on the PMS group behave on a test set (sized 20% of the training set) for different sizes of this other objects category (Fig. 2). The architecture of this ANN is a bit more sophisticated than the complexity demanded by our problem (as can be seen by the chosen architecture in Appendix B.2), but we wanted to be sure to not underfit in any case so the ANN is always sensitive to new data. If the category of other objects is very small the algorithm is very precise and has a high recall (Fig. 2 on top); few other objects appear in the regions of the feature space where PMS and CBe stars tend to be placed, so they have little impact in the classification (it also indicates that the ANN is good in telling the difference between PMS and CBe stars, although we note that a large fraction of the PMS category are low-mass T-Tauri stars). The more we populate the feature space with other objects the algorithm is less able to recognise PMS stars (the false negatives rise, Fig. 2 at middle) as the regions of the feature space with the more common PMS sources start to be highly populated by objects similar to PMS stars and undiscovered PMS objects. The number of false positives stays the same as the algorithm is still being efficient in the less populated regions. In other words, the PMS candidate region in the feature space gets smaller and localised around the less common PMS sources. The number of true positives drops as a consequence of the increase of false negatives. This causes the precision to barely change (Eq. (3)) but the recall to drop (Eq. (4), Fig. 2 on top) up to a stabilisation point (grey line in Fig. 2) where most of the different types of objects that populate the feature space differently have appeared, and hence adding more sources does not further constrain significantly the locus of PMS candidates in the feature space.

This stabilisation point can also be found if we study the number of PMS candidates retrieved after generalizing the trained ANN to the unlabelled sources of the SoSt (Fig. 2 at bottom, selecting as candidates those with a probability $p \geq 50\%$ of belonging to the PMS category). This number drops quickly from a very high value, where the algorithm does not know about the existence of anything but PMS stars and CBes, up to the stabilisation point where it becomes roughly stable. Of course, adding more other objects always diminish the region of PMS and CBe candidates in the feature space, but this stabilisation point constitutes the optimal size for the category of other objects, as larger sizes do not compensate the amount of extra information for the contamination they introduce.

Therefore, for constructing the category of other objects, we randomly sample sources from the SoSt (excluding the sources in the PMS and CBe categories) so that they are in a proportion of 99.82% with the number of sources in the category of PMS objects (848), this being the observed stabilisation point. This scales to 470 263 objects. Some of these sources might have been classified previously by different catalogues, although most remain unclassified.

We can approximate what is the proportion of other objects to PMS objects in the SoSt from a simulation. Robin et al. (2012), using the *Gaia* Universe Model Snapshot (GUMS) simulation estimate that the percentage of PMS objects within $G < 20$ mag in *Gaia* is 0.18%. The real proportion of PMS sources in the SoSt is somewhat larger as we are demanding detections up to $22\,\mu$m and in $H\alpha$. This implies, theoretically, that roughly there are as many undiscovered PMS stars in the other objects category as known PMS stars in the PMS category.
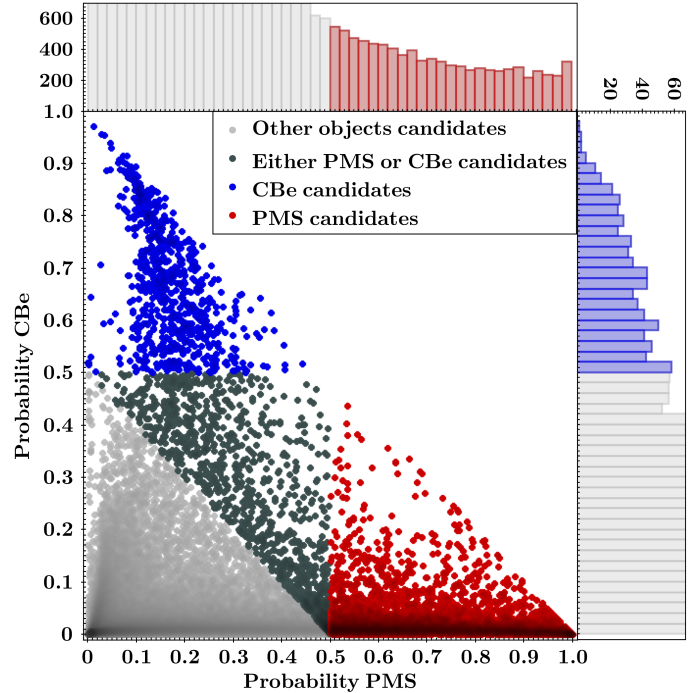

Fig. 3. Output probability map of the Sample of Study. A probability threshold of $p \geq 50\%$ is used to select the PMS (in red) and classical Be candidates (in blue). On top and right number histograms of the candidates for different probabilities. In dark grey the sources which belong to either category ($p(\text{PMS}) + p(\text{CBe}) \geq 50\%$ but $p(\text{PMS}) < 50\%$, $p(\text{CBe}) < 50\%$). Uncertainties are not indicated for clarity. Numbers are: PMS candidates (8470), classical Be candidates (693), either (1309), other objects (4 140 511).

4. Results

The set of Labelled Sources described in the previous section and the Sample of Study of Sect. 2.4 are introduced in a pipeline of algorithms which core is an ANN. As an output, the pipeline gives probabilities and associated uncertainties for all the sources in the SoSt of belonging to either the PMS, classical Be, or other objects category (excluding the known PMS and CBe sources included in the categories of the Labelled Sources set). These probabilities, that sum up to one in each source, are presented in Fig. 3. The pipeline, algorithms, and methodology are described in detail in Appendix B.

The SoSt with probabilities is available at the CDS in its entirety (4 150 983 sources). This data is made available so the user can choose their own probability threshold (p) to select PMS and classical Be candidates. Choosing p implies fixing a precision (P) and a recall (R). The pipeline also gives a solid estimate of the precisions and recalls for different p thresholds. However, due to the nature of the pipeline the values for the precision are only lower limits (see Appendix B.1). Ideally these two metrics should be as high as possible but there is a trade-off between them. This is shown in Fig. B.2, where the precision and recall for both the PMS and CBe category are plotted for different p probability thresholds. Rising the threshold to $p \geq 99\%$ maximises the precision to almost 1, but as a consequence the recall lowers to almost 0. The opposite also applies and neither of both extremes is close to be representative of a good selection; as it would be either largely incomplete or largely imprecise. The general shape of the curves is determined by the architecture of the algorithm and the peculiarities of the classification problem (see Appendix B.2).

In practice, using probability thresholds below 50% is possible, but entering the regime where the algorithm assigns larger probabilities to other categories is not advisable as p does not correlate linearly with the precision and recall (see Fig. B.2). At $p \geq 50\%$ the resulting catalogues are: new PMS candidates (8,470 sources, $P = 40.7 \pm 1.5\%$, $R = 78.8 \pm 1.4\%$), new classical Be candidates (693 sources, $P = 88.6 \pm 1.1\%$, $R = 85.5 \pm 1.2\%$). We note that the precisions are lower limits. These catalogues of new candidates are presented in Tables D.1 and D.2 respectively and highlighted in Fig. 3 (full tables available at the CDS)[2]. In those tables, together with the probabilities we present the observables used for the training (Sect. 2.1) and *Gaia* astrometric information. In addition, we included the derived interstellar extinction (A_G') and A_G' corrected M_G and $G_{BP} - G_{RP}$ for those sources with RUWE<1.4 and $\varpi/\sigma(\varpi) \geq 5$. These allow for a better positioning in the HR diagram (see Sect. 3 and Figs. 1, 4, and 5).

In the CBe case the precision does not drop drastically (see Fig. B.2). This implies that for the algorithm it is easier to find CBe stars than PMS stars as their locus in the feature space is less prone to contaminants but mostly because there are fewer unclassified CBes in the SoSt (see Appendix B.1). A consequence of this is that we retrieve an order of magnitude less CBe candidates than PMS candidates.

Following the discussion of Sect. 3.4, the size of the other objects category roughly coincides with the point where there is approximately one undiscovered PMS source per known PMS source in the training and test sets. Taking this into account, the lower limit on the precision of $P \sim 40\%$ for the PMS group obtained with $p \geq 50\%$ is an adequate enough result (i.e. the real precision is roughly double). However, as the precision is a lower limit, it is hard to assess whether a higher probability threshold is better to retrieve a stronger catalogue of PMS candidates. In order to decide this, we need to use parameters and observables that have not been used in the training, and are hence independent of the selection. As explained in Sect. 2, the set of features (and hence the classification) is distance and position independent, at least at first order. This means that we can use the HR diagram and the sky locations to assess this issue.

Before analysing these catalogues, we first remove the sources brighter than the typical bright limit of IPHAS and VPHAS+ that show significant differences between their IPHAS or VPHAS+ magnitudes and their *Gaia* magnitudes (marked in Tables D.1 and D.2 with a "X-mtch" flag). These objects did not affect the training as they barely account for 0.5% of the other objects category. There are 18 PMS and 57 CBe candidates with this flag. These sources are likely to be incorrect cross-matches and they are left out in the following analyses.

4.1. Evaluation using the HR diagram

The HR diagram is not entirely selection independent, as we used different colours in the classification and we do not correct for the unknown intrinsic extinctions. However, the location in the HR diagram, which carries information about evolutionary status, is almost independent of the classification.

The *Gaia* HR diagram of the PMS and classical Be candidates (those with $p > 50\%$) can be seen in Fig. 4 at left panels. In Fig. 4 at right panels we also distinguish those with a good astrometric solution (RUWE<1.4 and $\varpi/\sigma(\varpi) \geq 10$) to which interstellar extinction corrections have been applied. These well behaved sources have trustworthy positions in the HR diagram. The candidates have been colour-coded according to their membership probability to the corresponding category. In this HR diagram we can evaluate the quality of the retrieved catalogues. Regarding the PMS candidates, the majority are placed to the right of the main sequence, as expected for PMS sources. Moreover, if we move up to higher probability thresholds or to those sources with a good astrometric solution we constrain the selection of PMS candidates to those sources that are located in the more likely PMS positions. Something similar happens with the CBe candidates, as they are placed where CBe stars are supposed to be. This can be better appreciated when comparing these candidates with the locus of the known PMS and classical Be stars in Fig. 1. We note that ~6% of PMS candidates and ~1% of CBe candidates lack parallax information.

This, in addition of evidencing the quality of the selection, allows to select a higher probability threshold by looking at the retrieved candidates in the HR diagram of astrometrically well behaved sources corrected from interstellar extinction (Fig. 4 at top right). This threshold can be adapted to the requirements of different studies or situations. Here, we stick to the probability threshold of $p \geq 50\%$ for constructing the catalogues of candidates. This is because the candidates with the higher probabilities are the easier ones to find. Hence, as can be seen in Fig. 4, most of the high-mass PMS candidates do not have high associated probabilities as the algorithm struggles more to differentiate them from classical Be stars and vice-versa. In addition, Fig. 3 shows that there are very few CBe candidates with a negligible PMS probability. Therefore, a more conservative selection of the probability threshold would exclude many of the high-mass objects (see histograms of Fig. 3).

Finally, although in the rest of the paper we discuss the catalogues as of $p \geq 50\%$, the user can construct their own catalogues by means of Tables D.1, D.2, and the SoSt with probabilities. As any new catalogue is likely to have a higher probability threshold, the discussion and analysis that follows holds true. Hence, from here onward we refer to PMS and CBe candidates as those with a $p > 50\%$ in their respective categories.

4.2. Evaluation using sky locations

The selection has been totally independent of coordinates. This, though true, is limited by the combined footprint of the surveys used; for example, it is limited to the Galactic plane, see Sect. 2.4. Now, in Fig. 6 at top, we plot the contours of the catalogue of new PMS candidates on the Sample of Study footprint. The PMS candidates trace some of the overdensities of the SoSt. This is because any random selection of sources traces the footprint of the SoSt but might also be because the new PMS candidates are mostly appearing in star forming regions, which would be a strong assessment of the selection.

In this respect, it is noticeable that some overdensities are not particularly populated by PMS candidates. Moreover, if we look at the small scale (examples in Fig. 6 at middle and bottom panels) we see that the PMS candidates are not strictly following the SoSt overdensities but are more likely associated to nebulosities. In addition, in Fig. 7 we plot all the PMS candidates in the sky. They appear distributed all over the Galactic plane but there are associations of candidates, regions of around ~0.5 to 1 squared

---

[2] As a word of caution, these recalls do not imply that the presented catalogues contain ~80% of the existing PMS and CBe stars within any region. They imply that ~80% of the known PMS and CBe stars in the test set are recovered by the algorithm. This is for example affected by what the different surveys used are probing and the distribution of the SoSt (see Sect. 2.4). As explained in Sect. 5, probably some of the less extreme objects in the observables used have not been classified. Similar reasoning can be applied to the precision values.
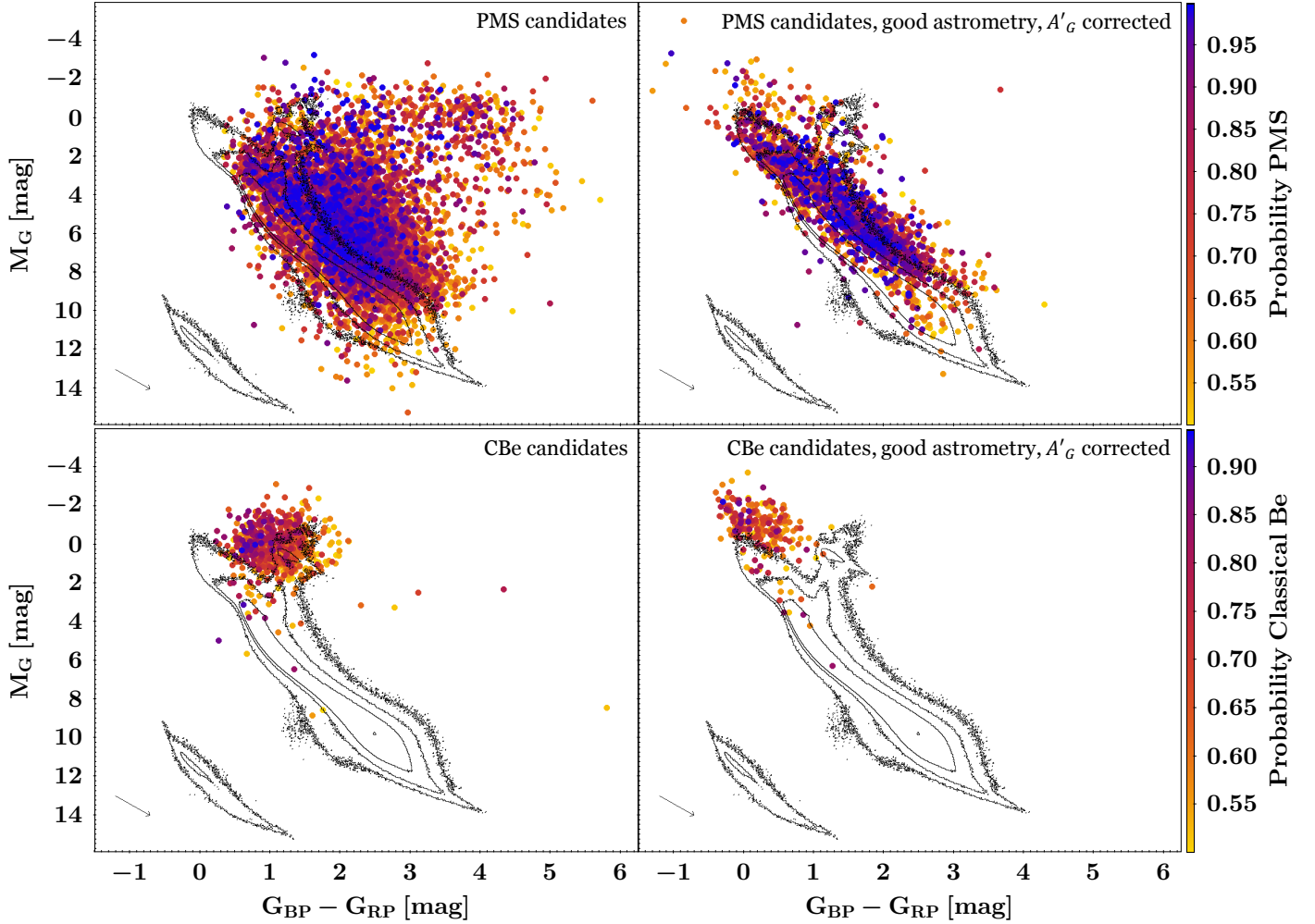
Fig. 4. *Gaia* colour vs. absolute magnitude HR diagram. An extinction vector corresponding to $A_G = 1$ is shown on the bottom left of each plot. Black contours trace *Gaia* sources within 500 pc with good astrometric solution. *Top left*: HR diagram of PMS candidates ($p \geq 50\%$) colour-coded by their associated membership probability. *Top right*: HR diagram of PMS candidates ($p \geq 50\%$) with good astrometric solution colour-coded by their associated membership probability and corrected from interstellar extinction. *Bottom left*: HR diagram of classical Be candidates ($p \geq 50\%$) colour-coded by their associated membership probability. *Bottom right*: HR diagram of classical Be candidates ($p \geq 50\%$) with good astrometric solution colour-coded by their associated membership probability and corrected from interstellar extinction.

degrees where there are ten to a hundred more PMS candidates than normally distributed. These associations also appear if we include the distances (Fig. 8).

This means that the *Gaia* coordinates are assessing the efficiency of the algorithm, as the retrieved PMS candidates are prone to appear around nebulosities and star forming regions, even though these regions are not over-represented in the input data. Further evidence in this respect can be seen in Fig. 7, where the new classical Be candidates are also plotted in the sky. These candidates are distributed all over the Galactic plane but they are not tracing the associations of PMS candidates or nebulosities (see e.g. Fig. 6 at bottom panels), which implies that they are indeed of an independent nature for the algorithm. Moreover, if we include the distances (Fig. 8), CBe candidates also appear decoupled from the PMS candidates. CBe candidates are typically further away, something expected from bright B-type stars.

Although the clustering properties of this new set of PMS sources is beyond the scope of this study, we can make some remarks. Firstly, on a global scale, PMS candidates trace some of the regions with more data available. This is likely because these zones contain star forming regions, as not all the regions

with more data available are overpopulated with PMS candidates (Fig. 6 at top panel). However, on a local scale, PMS candidates do not trace overdensities in the available data, and the associations of candidates appearing are normally not related to those overdensities (e.g. Fig. 6 at middle panels). Secondly, the PMS candidates seem to trace nebulosities, and the large sky associations obtained are mostly related to them (e.g. Fig. 6 at bottom panels). There are also a few smaller associations of PMS candidates unrelated to footprint overdensities that seem to trace dark nebula and are placed on their edges. Lastly, among the PMS candidates there is no significant correlation between PMS probability ($0.5 \geq p \geq 1$, see Fig. 3) and coordinates.

4.3. Herbig Ae/Be candidates

We have constructed a new catalogue of PMS candidates, of which some can be plotted accurately in the HR diagram (Fig. 4 at top right), a selection-independent plot. Therefore, we can further select the Herbig Ae/Be stars. In order to do this we study where the known HAeBes (Sect. 3.1.1) are placed in the HR diagram using the same quality constrains and extinction
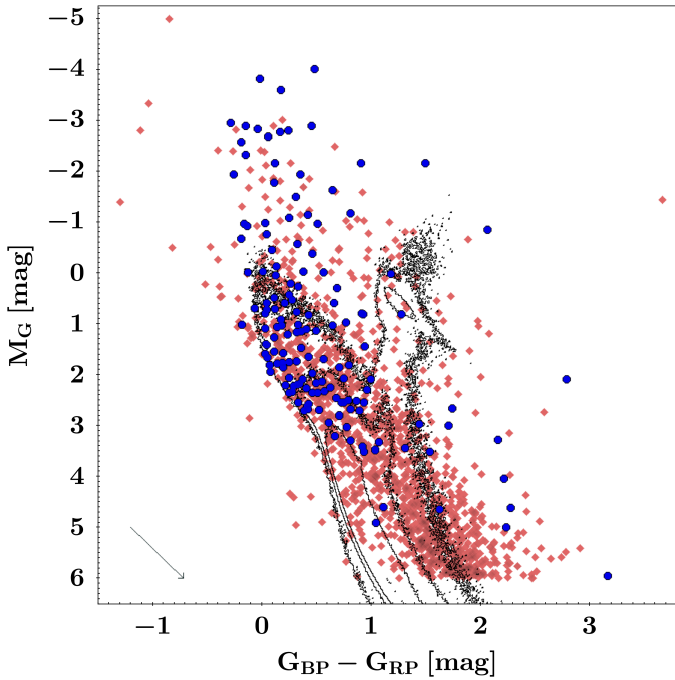
Fig. 5. *Gaia* colour vs. absolute magnitude HR diagram. Blue dots are previously known Herbig Ae/Be stars with good astrometric solution corrected from interstellar extinction. Red diamonds are new Herbig Ae/Be candidates corrected from interstellar extinction. An extinction vector corresponding to $A_G = 1$ is shown on the bottom left. Black contours trace *Gaia* sources within 500 pc with good astrometric solution.

corrections. This is done in Fig. 5. There, we estimate that PMS candidates with absolute magnitude $M_G < 6$ are possible HAeBe candidates. This is taking into account that the intrinsic extinction, typically large for these objects, has not been considered (most of these sources do not have measured spectral types). This way, we retrieve 1361 new Herbig Ae/Be candidates which are marked in Table D.1 (end of the pipeline, Fig. B.1). This constitutes an improvement of one order of magnitude with respect to the ~273 previously known HAeBes (Vioque et al. 2018). The new Herbig Ae/Be candidates are shown in Fig. 5.

By construction, these HAeBe candidates are astrometrically well behaved sources (RUWE<1.4 and $\varpi/\sigma(\varpi) \geq 10$). Hence, we expect many more HAeBes among the PMS candidates, as most of them do not satisfy these conditions (~60%, see Fig. 4 at top left) or do not even have parallax information (~6%). For example, using a more relaxed $\varpi/\sigma(\varpi) \geq 5$ parallax constraint gives 2226 HAeBe candidates, but their distance and interstellar extinction uncertainties do not allow us to separate them as nicely from the low-mass candidates. In contrast, using less precise parallaxes allow to retrieve candidates at farther distances and typically more massive. As the list of all PMS candidates is available in Table D.1, future studies may want to use less conservative thresholds to the astrometric quality and select their own set of Herbig Ae/Be candidates.

4.4. Variable candidates

UX Ori type objects (UXORs) are sources with irregular brightness variations from 2–3 mag in the optical. Observed light gets bluer in the deep minima, and the fraction of polarised light increases. Many of them are catalogued as HAeBes and their extreme variability is explained by eclipsing dust clouds in

nearly edge-on sources and scattering radiation in the circumstellar environment (Natta et al. 1997; Grinin & Grinin 2000; Natta & Whitney 2000; Oudmaijer et al. 2001). In Vioque et al. (2018) we provided strong support to the edge-on disc explanation using Hα line profiles of known HAeBes. In addition, using a similar variability indicator to G_{var} (Eq. (1)), we found that all catalogued UXORs with V band detected variabilities above 0.5 mag are strongly variable (17 objects). This implies that these indicators effectively trace irregular photometric variability.

By using a variability threshold, in Vioque et al. (2018) we proposed 31 new UX Ori candidates among the previously known HAeBes. The equivalent G_{var} threshold is $G_{var} \geq 10$. PMS candidates with $G_{var} \geq 10$ are marked in Table D.1 with the flag of "Var" (3436 sources) and for the HAeBe candidates the UXOR phenomenon is the most likely explanation. As we are tracing variability by using the *Gaia* DR2 uncertainties, sources without intrinsically irregular photometry like binaries or extended objects can pop out as strongly variable.

This means that ~41% of the new PMS candidates are of this variability type. This proportion increases to ~49% when it comes to the HAeBe candidates. This number is consistent with the UXOR behaviour caused by an inclination effect (50% predicted by Natta & Whitney 2000). Very probable PMS candidates are in general very variable (see Sects. 5.1 and 5.2), so most of the best PMS candidates appear as "Var" in Table D.1.

Another assessment of our variability proxies can be achieved by cross-matching the PMS candidates with variability surveys. A 5 arcsec cross-match with ASAS-SN (Jayasinghe et al. 2019) gives 949 sources, of which 830 (87%) have $G_{var} \geq 10$. In addition, 557/949 (59%) appear as of young-stellar object (YSO) variability type. A 5 arcsec cross-match with the Zwicky Transient Facility (ZTF, Masci et al. 2019) gives 6438 sources, including 95% of those with $G_{var} \geq 10$ in the sky region covered by the survey. A 5 arcsec cross-match with ATLAS-VAR (Heinze et al. 2018) gives 2216 objects. Of these, 1960 (88%) have variabilities which are hard to classify by the ATLAS-VAR machine classifier, suggesting that they are likely of an irregular type, similar to those of PMS sources. Finally, if we cross-match our results with the catalogue of long-period variable candidates of Mowlavi et al. (2018), which also contains a small set of YSO candidates, we obtain 491 matches with the set of PMS candidates, of which 444 (90%) have $G_{var} \geq 10$. According to Mowlavi et al. (2018) classification, 297/491 (60%) are YSO candidates but 190/491 (39%) are long-period variable candidates (4/491 are undetermined because they lack parallaxes). These possible contaminants are addressed in Sect. 5.2.

4.5. Comparison with *Marton et al.* (2019) and other catalogues and surveys

Marton et al. (2019) did a similar study to the one presented here but looking for YSOs in general and only using *Gaia* DR2, 2MASS, WISE colours and passbands, and the optical depth from the *Planck* dust opacity map. Therefore, they did not use Hα or variability information and did use distance-dependent features. In addition, they restricted the search to high dust opacity regions. They found 1 768 628 potential new YSO candidates (with the recommended $p \geq 0.9$) using a Random Forest algorithm. Giving the differences between the two approaches in terms of considered sources, training data, and features; it should not be surprising if there are not many objects in common between the two studies and yet both are highly accurate. However, we find that 48% of our PMS candidates are within the Marton et al. (2019) catalogue. Moreover, this percentage
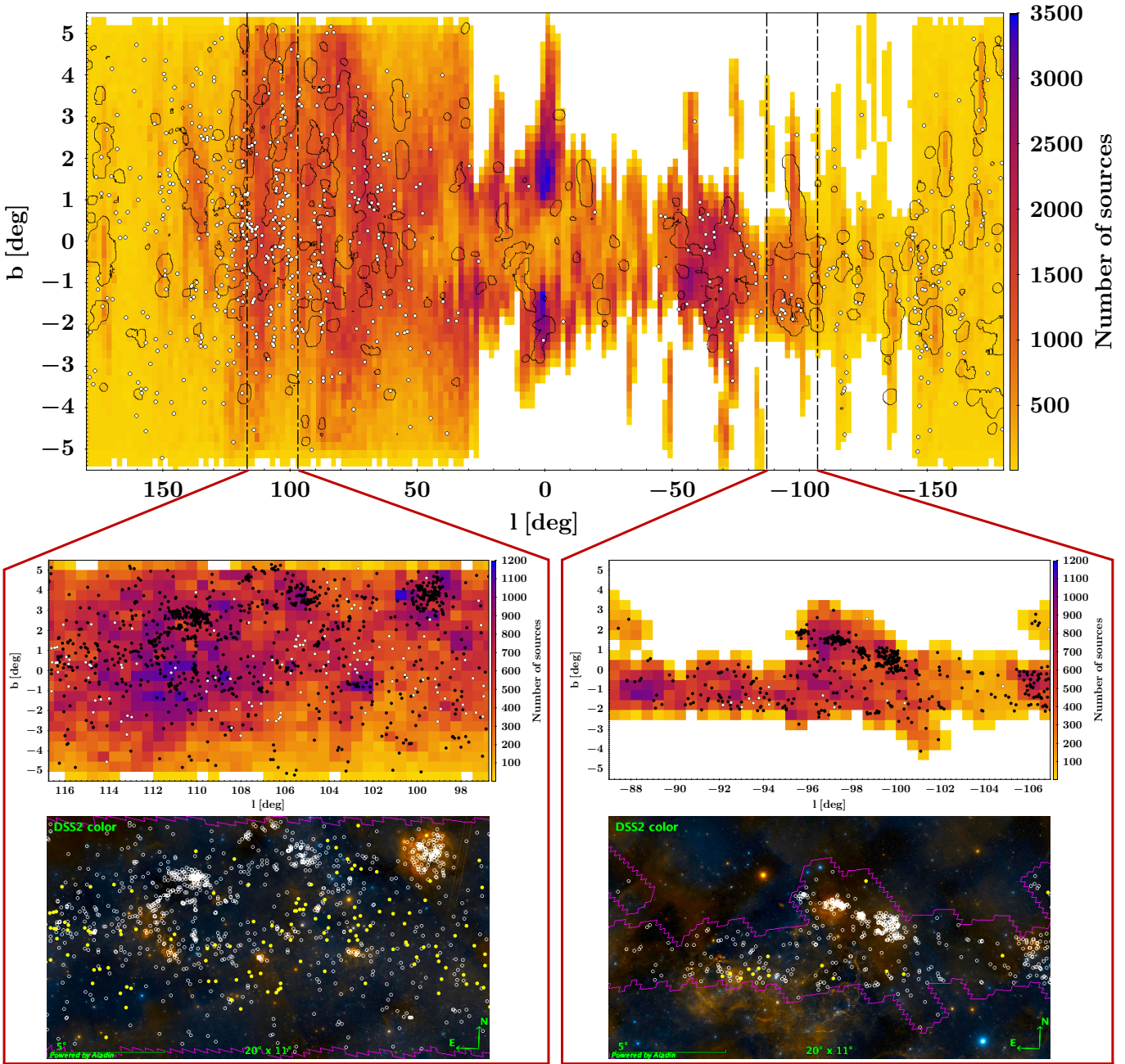
Fig. 6. *Top*: sky footprint of the Sample of Study in galactic coordinates, colour-coded by number density. We note the heterogeneity of the footprint. The scarcity of sources between $29° > l > -145°$ is due to the incompleteness of VPHAS+ at the time of writing. Each pixel is $2° \times 0.2°$. PMS candidates overdensities appear as black contours. There are ten time more candidates inside than outside the contours. CBe candidates appear as white dots. Expanded regions at bottom panels appear between dashed lines. *Middle*: expanded regions. Each region is $20° \times 11°$. PMS contours are replaced by PMS candidates (black dots). Each pixel is $0.5° \times 0.5°$. *Bottom*: same expanded regions in DSS2 colour with the PMS candidates as white circles and the CBe candidates as yellow dots. Contours trace the footprint of the Sample of Study.

slightly increases at higher probability thresholds of our catalogue (56% at $p \geq 0.95$). Regarding the Herbig Ae/Be candidates (see Sect. 4.3), 56% of them are present among the YSOs of Marton et al. (2019). In contrast, only 11% of our catalogue of classical Be stars appear as YSO in Marton et al. (2019). When moving to $p \geq 0.85$ this number goes down to 0%.

This is a good assessment of the quality of our categorisation as an independent study, using a different algorithm and training data, has achieved relatively similar results regarding PMS sources (taking into account the differences between methodologies) but has not found almost any of our CBe candidates (and

none of the best CBe candidates). This, in addition to support our selection, proves that our HAeBe candidates are nicely separated from the population of CBe stars. The differences between the two studies probably lie in that we are using Hα and variability information and that Marton et al. (2019) searched only in dusty environments, being this way position-dependent. In addition, we demand detections up to W4 ($22 \, \mu$m), whereas these authors only demand detections up to W2 ($4.6 \, \mu$m). Further assessment is that, as in Marton et al. (2019), we find that 62 of our PMS candidates are within the *Gaia* Photometric Science Alerts published at the time of writing (a project that looks for transient
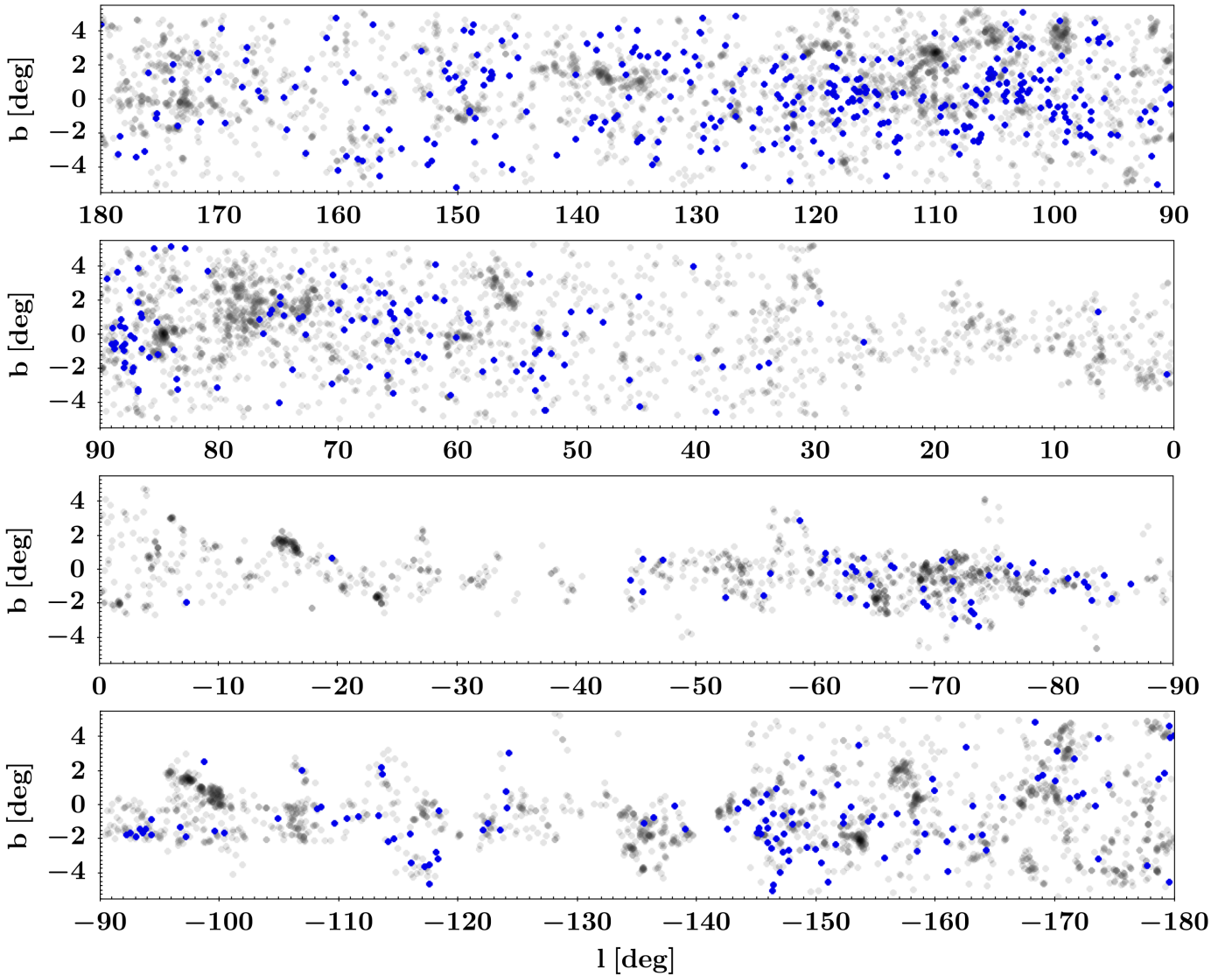
Fig. 7. Classical Be candidates in blue distributed in the sky in galactic coordinates plotted on top of the PMS candidates in light grey. The densest regions of PMS candidates appear darker. The scarcity of sources between $29° > l > -145°$ is due to the incompleteness of VPHAS+ at the time of writing.

events in the *Gaia* data, Delgado et al. 2019); 13 of them appear as YSO, 47 as unknown, and only two appear as non-PMS. Conversely, of the 87 YSOs in the Alerts, 18 are in the SoSt, which means that we only missed five that were classified as "other source".

Similarly, of the PMS candidates in SIMBAD (2607 within 1 arcsec cross-match at the time of writing) 974 (~37%) appear catalogued as PMS or PMS candidate. There are 18 objects appearing as CBe, but these were mostly catalogued by Mathew et al. (2008) and Gkouvelis et al. (2016). These papers selected CBes using simple cuts in IPHAS+, 2MASS, or WISE observables and hence we understand that our analysis supersedes theirs. In addition to this, 663 sources (~25%) appear as with emission lines, infrared bright, or variable. Only 356 sources (~14%) appear as clearly non-PMS. This includes 101 AGB candidates and 16 carbon star candidates that are addressed in Sect. 5.2. As explained in that section, we expect this number of 356 PMS candidates classified as non-PMS by other studies to be considerably lower, so this cross-match with SIMBAD is consistent with the estimated precision in Sect. 4 of $P \gtrsim 80\%$ for

the catalogue of PMS candidates. The other 596/2607 sources do not have a defined category in SIMBAD. VES 263, the new Herbig Ae/Be star discovered by Munari et al. (2019) is not within the SoSt.

Of the classical Be candidates in SIMBAD (280 within 1 arcsec cross-match at the time of writing) 17 appear as CBe (again, most from Mathew et al. 2008 and Gkouvelis et al. 2016) and 197 as with emission lines. Only nine are clearly not CBes, of which four are of PMS nature and three appear as variable. This reinforces the idea that the algorithm is efficiently separating PMS sources from classical Be stars. The other 57/280 sources do not have a defined category in SIMBAD.

Finally, using a cross-match aperture of 20 arcsec we find 26 matches between the set of PMS candidates and the *Gaia*-ESO Public Spectroscopic Survey (Gilmore et al. 2012). A fraction of 24/26 sources have hydrogen lines in emission: 14/26 show double-peaked emission (although two might be considered P-Cygni), 6/26 single-peaked emission, 3/26 are either single-peaked or double-peaked, and one shows a clear inverse P-Cygni profile. Only 2/26 spectra have Hα line in absorption. The line
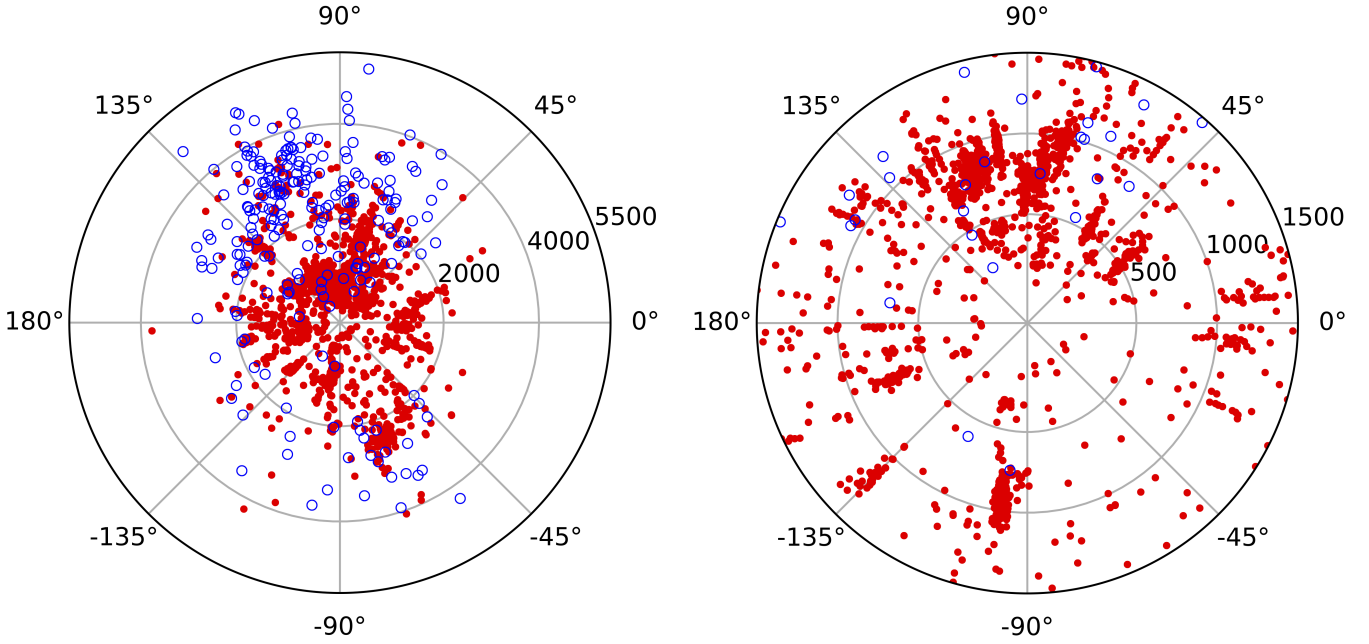
Fig. 8. Galactic longitude vs. distance (in parsec [pc], from Bailer-Jones et al. 2018) of PMS (red dots) and classical Be (blue circles) candidates with good astrometric solution. *Left*: all candidates, *right*: candidates up to 1500 pc.

profile fractions agree with those studied in Vioque et al. (2018) for known HAeBes (31% single-peaked, 52% double-peaked, and 17% P-Cygni). This gives independent spectroscopic evidence for the PMS nature of the new PMS candidates.

5. Quality assessment

Table 1 summarises the final number of sources in the resulting catalogues of PMS and CBe candidates. Table 1 also indicates the number of known sources considered in Sect. 3, of which those having all observables were used for the set of Labelled Sources. In this section we evaluate the classification from different perspectives and give insights on the relative importance of the different observables used for the selection. In addition, we discuss any detected bias or flaw in the final catalogues of PMS and classical Be candidates. In general, these mostly affect sources with a bad astrometric solution in *Gaia* DR2 so they do not implicate the catalogue of new Herbig Ae/Be candidates (Sect. 4.3).

5.1. Classification on the test sets

One way to analyse the classification is to study the evaluation on the test sets. As described in Appendix B.4, we evaluated the performance of the ANN in 30 different test sets. As the selection of the test set is random in every iteration, almost all of the known PMS and classical Be sources were in the test set at some point. If we average these 30 evaluations we end up with 793 PMS and 733 CBe known sources that have been independently assessed by the algorithm.

Regarding the classification of known PMS sources, the most noteworthy trend is that very variable PMS stars in either indicator (G_{var} and V_{htg}) are identified. Although those known PMS stars with $r-H\alpha > 1.3$ are identified, objects with $0 < r-H\alpha < 1.3$ are spread over the whole range of probabilities. Thus, $r-H\alpha$ does not seem to play an important role in detecting PMS sources (see Sect. 5.4). This also happens with $G_{BP} - G_{RP}$. However, in

these two cases, known PMS sources with low $r-H\alpha$ or bluer are those who tend to be given high CBe probabilities. The known PMS sources that were not identified were mostly stars with low near-IR excess ($J - K_s$), which are also the ones that are given high CBe probabilities. This is probably because these are more similar to CBe stars. Surprisingly, we miss many known PMS sources with high mid-IR excess ($W1-W4$) and those that had very low $W1-W4$ values were mostly not identified, which again are the ones with higher CBe probabilities. In general, very few known PMS sources are assigned to the CBe category, although many known PMS sources are not classified as such (algorithm recall on the PMS group is $R = 78.8 \pm 1.4\%$, Sect. 4.1).

Regarding the known classical Be sources, the algorithm also identifies the very variable ones as CBe. This implies that it uses variability to differentiate PMS and CBe sources from other objects. CBe sources with high $r - H\alpha$ are normally given high PMS probabilities but in general they are not misclassified. There is no trend between $r-H\alpha$ values and CBe assigned probabilities. In contrast, there is a trend with $G_{BP} - G_{RP}$ and redder objects are less likely to be classified as CBe and are given higher PMS probabilities, but are rarely misclassified as PMS. In addition, CBe sources show no CBe probability trend with $J - K_s$ or $W1 - K_s$ although sources with $W1 - W4 \gtrsim 3$ are normally not classified as CBe. Similarly, CBe sources with higher near- and mid-IR excesses are given higher PMS probabilities but are infrequently assigned to the PMS category.

Evaluation on the test sets indicates that the algorithm effectively identifies sources of different categories and uses the various observables to trace the main characteristics of PMS and classical Be stars.

5.2. Final catalogues assessment

In the following points we discuss a few biases and flaws detected in the final catalogues of PMS and CBe candidates:

1. We demand to have detections up to W4 (22μm) and in the Hα passbands. Although we are training with sources that span the whole range of values in these observables, this induces

Table 1. Summary of the number of known sources of each type considered together with those included in the set of labelled sources because of having all observables.

	Considered sources	Labelled sources set	Classified $p \geq 0.5$
Herbig Ae/Be	255	163	−
T-Tauri	3171	685	−
PMS	3426	848	8470
Classical Be	1992	775	693

Notes. The last column indicates the number of new candidates of each type classified by the algorithm. To be considered, we demanded the sources to be present in *Gaia* DR2. For the algorithm there is just a single class of PMS objects, which is constructed by combining known Herbig Ae/Be and T-Tauri stars. Herbig Ae/Be and T-Tauri candidates can be selected from the set of PMS candidates using the HR diagram. In Sect. 4.3 1361 Herbig Ae/Be candidates were obtained this way.

some biases as we are excluding in the training many of the less extreme sources and hence biasing the posterior selection. This is aggravated given that the Hα EW to $r − H\alpha$ colour conversion of Sect. 3 can only be applied to sources with observed emission above the continuum. This effect can be quantified if we compare the output catalogues with all the known sources gathered in Sect. 3 (see Table 1). This way, the mean value of $r − H\alpha$ for known PMS (classical Be) sources (using one standard deviation as error) is $r − H\alpha = 0.74 \pm 0.36$ mag (0.38 ± 0.18 mag) and for the candidates is $r − H\alpha = 0.87 \pm 0.46$ mag (0.63 ± 0.20 mag). The mean value of $W1−W4$ for known PMS (classical Be) sources is $W1−W4 = 4.0 \pm 2.2$ mag (1.7 ± 1.3 mag) and for the candidates is $W1−W4 = 5.2 \pm 1.4$ mag (2.24 ± 0.71 mag). In short, the retrieved candidates are the more extreme of their kind in terms of Hα emission and IR excess (specially mid-IR excess). This particularly affects the catalogue of CBe candidates, as these have typically less extreme values. In Fig. 9 we present the frequency density distribution of the final catalogues of PMS and CBe candidates for a subsection of key observables ($G_{BP} − G_{RP}$, $J − K_s$, $W1−W4$, $r − H\alpha$, G_{var}, and V_{htg}) together with the distribution of all known sources.

2. As mentioned in Sect. 2.4, WISE presents many spurious photometric detections in the Galactic plane. To investigate this, Koenig & Leisawitz (2014) used a set of AllWISE quality parameters and additional selection criteria to determine that only ~28% of the sources in their study have reliable W3 and W4 detections. Marton et al. (2019), using a different approach, concluded that only 10% of their set have reliable W3 and W4 photometry. These authors used very stringent criteria for the sake of purity and these percentages may be slightly pessimistic.

We decided to use these passbands because of their expected importance in separating CBes from PMS sources (see Sect. 5.4). A more relaxed constraint, using the extended source flag of AllWISE distinct to 0 gives 44% and 27% of badly behaved PMS and CBe candidates respectively (marked in Tables D.1 and D.2 with a "W3W4" flag). We note that, in contrast to Marton et al. (2019), we are using many observables in addition to W3 and W4 so the algorithm can deal better with these being spurious.

3. As described in Sect. 2.4, because of the cross-match, we estimated that 1/42 (1/625) sources of the SoSt on average are fake, in the sense that their associated IR (Hα) photometry do not belong to them. However, only 17 (6) PMS and no CBe candidates appear with duplicated IR (Hα) information. The sources

that have the AllWISE, or IPHAS or VPHAS+ name repeated in the SoSt are marked in Tables D.1 and D.2 with the "ID AllW" or "ID IPH/VPH" flag respectively.

4. There are 104 SIMBAD AGB stars in the catalogue of PMS candidates (only three appear as of confirmed AGB nature and the rest appear as candidates). This is because they were all classified in one single paper (Robitaille et al. 2008), where they attempted to separate YSOs from AGB stars using a simple colour and magnitude selection criteria in the near- and mid-IR. We understand that these are contaminants in that work as our analysis supersedes theirs.

5. We have detected a high incidence of planetary nebulae (PN) detected as PMS candidates. Observational similarities between YSOs, B[e] stars, and PN have been reported before (e.g. Frew & Parker 2010; Boissay et al. 2012; Miroshnichenko & Zharikov 2015 and references therein). This is mainly caused because PN show high $r − H\alpha$ colours. In addition, as they are extended they present large *Gaia* uncertainties, so they can appear as highly photometrically variable in our indicators (Eqs. (1) and (2)). Of the PMS candidates in SIMBAD (Sect. 4.5), there are 57 (~3.5%) catalogued as PN and 34 as possible PN. By studying their location in the observable space we concluded that any candidate with a $r − H\alpha \gtrsim 1.3$ should be treated with caution (16% of the sample of PMS candidates). Below that number we estimate the possible contamination by PN to be below 5%. Candidates with $r − H\alpha \geq 1.3$ are marked in Tables D.1 and D.2 with a "PN" warning flag. Indeed, there are eight PN (within 5 arcsec) from Sabin et al. (2014), 40 from Kerber et al. (2003), and three from Stanghellini et al. (2008) among our PMS candidates. A fraction of 46/51 (90%) have $r − H\alpha \geq 1.3$. We expect also some contamination in these works. PMS candidates with $r − H\alpha \geq 1.3$ have mostly absolute magnitudes $M_G > 6$, so they barely contaminate the sample of Herbig Ae/Be candidates (Sect. 4.3). In contrast, the few candidates with a "PN" warning flag and $M_G < 6$ have a significant probability of being unclassified B[e] (FS CMa) stars.

6. Similarly, we detect a high number of carbon stars among our PMS candidates (71 confirmed and 16 candidates, according to SIMBAD). They stand out in variability and near-IR excess, but not in mid-IR excess, where they have a smaller excess than the rest of the candidates. Only two were classified as Herbig Ae/Be (*Gaia* DR2 1828276425855506304 and *Gaia* DR2 5336019093122634624 with PMS probability of 0.52 and 0.53 respectively) as the other 85 do not have reliable astrometry. Not surprisingly, 80/87 (92%) were identified as contaminants in Appendix C and marked in Table D.1 with a "G_{UMAP}" warning flag (see Appendix C for further details). Therefore, we do not expect them to have a high impact on the final catalogue of PMS candidates. In addition, 29 PMS candidates appear as variable stars of Mira Cet type in the cross-match with SIMBAD. A fraction of 17/29 (59%) are flagged as "G_{UMAP}" in Table D.1.

7. 51 PMS candidates are in the catalogue of OH/IR stars of Engels & Bunzel (2015, within 5 arcsec). These 51 sources have different categories in SIMBAD and may have been assessed in previous points and sections as of other nature (even YSO). From the catalogue of AGBs of Suh & Hong (2017), 26 (within 5 arcsec) are among our PMS candidates (14, 54%, are flagged as "G_{UMAP}" in Appendix C). Similarly, a very small fraction of the CBe catalogue is potentially contaminated by sources with very strong mid-IR excess or that seem evolved. There are 0 post-AGB stars of Szczerba et al. (2007) in either catalogue.

Most of the contaminants discussed in the previous points can be avoided by constraining the position in the HR diagram and moving away from the giant and supergiant region. One way
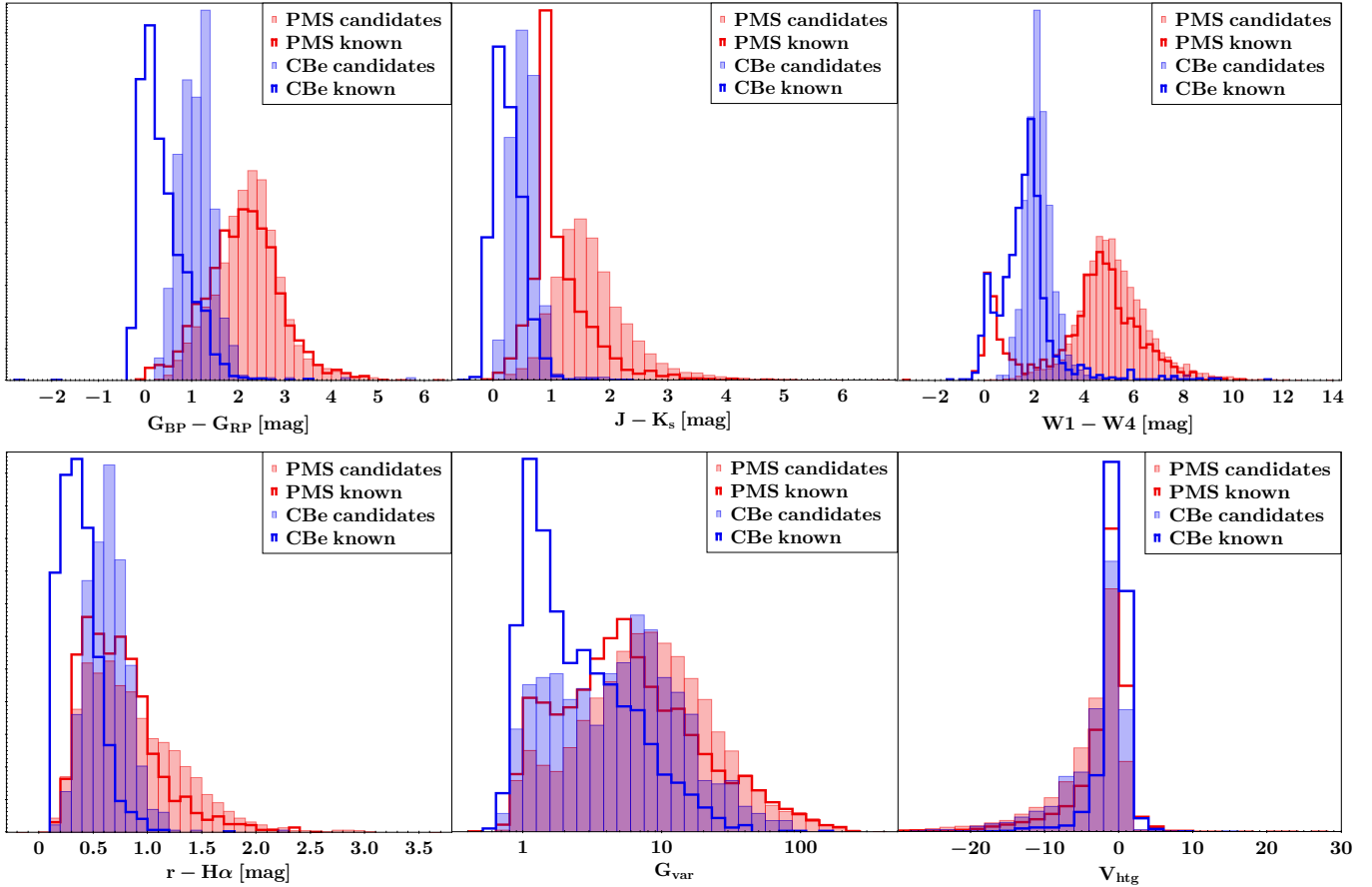
Fig. 9. Frequency density distribution of PMS candidates in shaded red and classical Be candidates in shaded blue for different selected observables. The red and blue lines respectively trace all considered known PMS and CBe objects, including those without all the observables. Area of each histogram has been normalised to one. For clarity, some individual extreme sources are out of bounds in the $r - H\alpha$, G_{var}, and V_{htg} plots.

to do so is with the constraint applied in Fig. 4 at right. This also implies that the sample of Herbig Ae/Be candidates (see Sect. 4.3) is barely affected by these contaminants. Conversely, the HR diagram is a powerful tool to discard contaminants in other catalogues which were correctly classified in this work.

5.3. Probability of being either PMS or classical Be

There are 1309 sources whose probabilities of being PMS and CBe sum up to a probability ≥50%, but that are below 50% in either category (i.e. $p(PMS) + p(CBe) \geq 50\%$ but $p(PMS) < 50\%$, $p(CBe) < 50\%$, see Fig. 3). This means that the algorithm thinks that they belong to one of those two categories, but it is unable to say which. A closer look at these objects reveals that they behave very similarly to known CBe stars in terms of G_{var}, V_{htg}, and $r - H\alpha$ but their mean $G_{BP} - G_{RP}$ colour is redder and they have slightly larger near-IR excess ($J - K_s$). The $W1-W4$ colour peaks where CBe stars do but it is quite homogeneously distributed. This is not surprising as in Fig. C.1 at left they appear mostly mixed with the CBe candidates, in the region where the PMS candidates that are more similar to known CBes are placed.

These borderline objects are interesting in their own right and contain probably most of the less active PMS sources and in particular, most of the less active and probably more evolved Herbig Ae/Be stars. These sources are listed in a table equivalent to Tables D.1 and D.2 which is only available at the CDS.

5.4. Important observables

In this section, we try to assess how important the different observables are for identifying PMS and classical Be stars. This is not trivial because of the intrinsic nature of the ANN-based algorithm, as the selection itself is not an obvious process.

What we did was to repeat the pipeline explained in Appendix B excluding some observables. We did not include the sources that were removed by demanding detections in those observables (see Sect. 3), as this would make it impossible to know whether the new results are caused by the new sources or the lack of those observables. Similarly, the ANN architecture was optimized for the whole set of observables (Appendix B.2), so using fewer observables might alter the performance in an uncontrolled manner that can affect our conclusions. In order to minimise the impact of this we removed only a few observables at a time. Another problem is that by using fewer observables without changing the complexity of the algorithm we may start overfitting the selection. To assess this we checked that the retrieved precisions and recalls are within reasonable limits in each case.

In Table 2 we present the results (precision, recall, PMS and CBe candidates with $p \geq 0.5$) obtained when applying the same algorithm of Appendix B to the same sources of Sects. 2.4 and 3 but excluding certain observables (in the case of passbands this implies excluding the colours they appear on, see Sect. 2.1). In the last two columns we express the percentage of PMS and CBe candidates of those selections that were also retrieved when

Table 2. Evaluation of the impact of the different observables in the final selection.

	Precision PMS (%)	Recall PMS (%)	Precision CBe (%)	Recall CBe (%)	PMS ($p \geq 0.5$)	CBe ($p \geq 0.5$)	PMS P/R (%)	CBe P/R (%)
All observables	**40.7 ± 1.5**	**78.8 ± 1.4**	**88.6 ± 1.1**	**85.5 ± 1.2**	**8470**	**693**	**100/100**	**100/100**
No r, $H\alpha$	35.5 ± 1.4	73.1 ± 1.5	87.1 ± 1.1	85.52 ± 0.97	9108	718	77/82	80/83
No G_{BP}, G, G_{RP}	32.11 ± 0.99	74.2 ± 1.2	87.5 ± 1.0	82.2 ± 1.3	10772	650	63/80	80/75
No J, H, K_s	32.4 ± 1.2	81.9 ± 1.4	87.42 ± 0.84	86.59 ± 0.64	11123	758	69/91	81/88
No $W1$, $W2$, $W3$, $W4$	28.33 ± 0.68	80.4 ± 1.3	70.2 ± 1.5	76.6 ± 1.5	13303	1890	47/74	25/69
No $W1$, $W2$	37.6 ± 1.1	75.1 ± 1.3	86.0 ± 1.2	85.8 ± 1.1	8398	742	76/75	75/80
No $W3$, $W4$	31.7 ± 1.1	81.0 ± 1.2	83.9 ± 1.2	75.4 ± 1.9	11889	820	68/95	58/69
No G_{var}, V_{htg}	23.49 ± 0.66	74.0 ± 1.3	91.98 ± 0.83	80.7 ± 1.1	18055	358	40/84	100/52
No G_{var}	27.78 ± 0.84	74.7 ± 1.4	92.62 ± 0.80	80.2 ± 1.8	13667	355	53/85	99/51
No V_{htg}	40.0 ± 1.1	73.1 ± 1.4	91.89 ± 0.97	83.8 ± 1.6	7543	468	97/87	100/67

Notes. Metrics and number of sources obtained for the different classifications excluding the indicated observables. Precisions are lower limits (see Appendix B.1). Last two columns indicate the "precision" and "recall" of these new catalogues with respect to the catalogues obtained when using all the observables. This is, the proportion of sources of the different catalogues that are in the catalogues obtained using all the observables and the proportion of the candidates obtained when using all the observables that are still retrieved when excluding the respective observables.

using all the observables and the percentage of sources classified when using all the observables present in these new catalogues. These two values, in some sense, are equivalent to precisions and recalls if we assume the catalogues obtained using all observables as reference. As the algorithm was optimized to maximise the precision on PMS sources (see Appendix B.3), this is maximum when using all observables. Similarly, when applying the algorithm to a smaller set of observables the results are going to be inevitably worse (as we do not include more sources). However, there is information in how much worse they get, although we can only talk in relative terms.

As outlined above, this table should be treated with caution but it gives information about the relative importance of the different observables in the selection of PMS and CBe candidates. We discuss the main outcomes of Table 2 in the following points:

1. Not using $r - H\alpha$ does not change the output tremendously. The number of PMS and CBe candidates retrieved is similar and there is only a small discrepancy (of ~20%) with the case of using this colour, in the sense that mostly the same sources are identified and not many sources that were not identified when using $r - H\alpha$ are included. This is because cool stars have the same $r - H\alpha$ colour than hot stars with emission, see Drew et al. (2005), and hence this observable is not efficient in separating PMS and CBe sources from other objects.

2. G_{BP}, G, and G_{RP} are more relevant for the selection. If we do not use them the discrepancy with the original set of PMS candidates obtained using all the observables is higher than in the $r - H\alpha$ case. Many more sources are obtained (which we can consider a sort of contaminants) without losing many of the catalogued ones using all observables. In the case of the CBes, the effect is the opposite, not many contaminants are added but a few of the sources identified with these colours are lost. This might be caused by these colours carrying the photospheric information less affected by disc emission and hence more representative of temperature. Therefore, for the PMS case including them helps to remove candidates with unfeasible temperatures (like white dwarfs) and in the CBe case it helps a bit the selection as they are mostly blue with low extinctions (see Figs. 1 and 4).

3. J, H, and K_s: Not using these 2MASS passbands makes us lose very few PMS and CBe candidates (less than in the previous cases) but we get many PMS contaminants, implying that the colours involving these passbands are relevant for differentiating PMS sources from other objects, although they are not critical for the selection. These observables do not seem to have a big impact for the classification of CBes.

4. As expected, $W1$, $W2$, $W3$, and $W4$ are very important. Not using these WISE passbands drastically increases the number of contaminants and significantly reduces the number of PMS and CBe candidates obtained when compared to the case of using this information. However, removing so many observables at a single time can cause the algorithm to start overfitting, so these results might be a bit exaggerated. If we choose to not use only $W1$ and $W2$ the selection is not much affected (only a bit more than in the $r - H\alpha$ case). Not surprisingly, it is if we choose not to use $W3$ or $W4$ when we obtain very poor results. We retrieve mostly the same PMS candidates but also many PMS contaminants. The larger impact is in the case of the CBes, as we miss a lot of them and misclassify almost half of the obtained catalogue. This is expected, as at this wavelength range is where the discs of PMS stars and CBes start differing. Therefore, probably many CBes are moved to the PMS catalogue lowering its precision. This is the reason we opted to keep these passbands even though they suffer from a high incidence of spurious detections (see Sects. 2.4 and 5.2).

5. G_{var} and V_{htg}: The observable G_{var} proves to be of the utmost importance for identifying PMS sources. When excluding both variability indicators, we get twice as many PMS candidates than in the catalogue using all observables, with an almost full recovery of the later ones. Curiously, half of the CBes are lost, but not a single contaminant is added. If we exclude them independently we find that V_{htg} is almost irrelevant. It only helped to classify several CBes. In contrast, not using G_{var} doubles the number of PMS candidates retrieved (so half the sources can be considered contaminants) and halves the number of CBes obtained. The number of CBe contaminants is close to zero in every case. All these imply that this indicator is very useful for separating PMS sources from other objects and, in some cases, to differentiate them

from CBes, but ineffective to identify CBes from the background sources. This is just as expected as we know from Vioque et al. (2018) that G_{var} mostly traces irregular photometric variations caused by edge-on dusty discs.

It is clear that if we had optimized the algorithm for each situation using all the sources available in each case for the training we would have obtained more candidates and, from Table 2, it is safe to say that these would have been more contaminated.

6. Conclusions

In this work we have used machine learning techniques (mainly artificial neural networks) to produce a catalogue of new PMS candidates and a catalogue of new classical Be stars from 4 150 983 sources resulting from the cross-match of *Gaia* DR2, AllWISE, IPHAS, and VPHAS+. To each of the 4 150 983 sources we assigned a PMS and a classical Be probability. The entire set of sources is available at the CDS so the users can choose the probability thresholds (p) that fit their needs. The categorisation is distance and position independent. We have given independent evidence that the categorisation is accurate and consistent, having a high efficiency at separating PMS sources from classical Be stars.

1. At $p \geq 50\%$ the catalogue of PMS candidates is: 8470 sources, recall (completeness) of $78.8 \pm 1.4\%$ and lower limit to precision of $40.7 \pm 1.5\%$. Independent analyses indicate that the real precision is around double this value. The PMS candidates are distributed all over the Galactic plane, tend to be associated with nebulosities and appear mostly in PMS locations in the HR diagram. This catalogue (Table D.1) is available at the CDS independently.

2. Out of the PMS candidates, 2052 have a good astrometric solution in *Gaia* DR2 (RUWE<1.4 and $\varpi/\sigma(\varpi) \geq 10$). Of those, 1361 have a location in the HR diagram compatible with that of known Herbig Ae/Be stars. Many more Herbig Ae/Be candidates can be obtained from the set of PMS candidates by relaxing the constraint to the parallax quality. This comes at a price, as the larger errors on the absolute magnitudes make it more difficult to distinguish low- and high-mass objects from each other.

3. At $p \geq 50\%$ the catalogue of classical Be candidates is: 693 sources, recall (completeness) of $85.5 \pm 1.2\%$ and lower limit to precision of $88.6 \pm 1.1\%$. The classical Be candidates are distributed all over the Galactic plane and appear mostly in classical Be locations in the HR diagram. This catalogue (Table D.2) is available at the CDS independently.

4. There are 1309 sources that have a combined probability of larger than 50% of belonging to either of these categories but each individual category has a probability of below 50%. In general these objects have characteristics of classical Be stars or the less extreme PMS sources in the observables used. These sources are listed in a table equivalent to Tables D.1 and D.2 which is only available at the CDS.

5. We have made a thorough analysis of the possible biases and contaminants present in the selection. The biases can be summarised in that we are retrieving the most extreme PMS and classical Be sources in the observables used. The contaminants are mostly giants, with the special case of planetary nebulae appearing as PMS. These contaminants are sparse and easy to avoid. Instructions are given to minimise their impact (in Sects. 4, 5.1, 5.2, and Appendix C). The new HAeBe candidates are very little affected by these contami-

nants, mainly as by construction they have a good astrometric solution.

6. 3436 PMS candidates (at $p \geq 50\%$) show strong irregular photometric variabilities. For the HAeBe candidates the UXOR phenomenon is the most likely explanation. The proportion of variable HAeBe candidates is consistent with the inclination explanation for the UX-Ori type variability.

7. An analysis of the relative importance of the different observables used shows that irregular photometric variability is extremely important for identifying PMS sources among other objects and W3 [12 μm] and W4 [22 μm] are very powerful to separate PMS sources from classical Be stars. On the other hand, $r - H\alpha$ is not very relevant for the selection of these two types of objects.

These new catalogues of PMS and classical Be candidates will be subjected to follow up studies of their properties using independent spectroscopic observations. The catalogue of new PMS candidates was accepted as target list for the WEAVE survey, a wide-field spectroscopic survey which will be carried out at the *William Herschel* Telescope in the forthcoming years (Dalton et al. 2018).

References

Ababakr, K. M., Oudmaijer, R. D., & Vink, J. S. 2017, MNRAS, 472, 854
Acke, B., & van den Ancker, M. E. 2006, A&A, 457, 171
Alecian, E., Wade, G. A., Catala, C., et al. 2013, MNRAS, 429, 1001
Arias, M. L., Cidale, L. S., Kraus, M., et al. 2018, PASP, 130, 114201
Arun, R., Mathew, B., Manoj, P., et al. 2019, AJ, 157, 159
Astropy Collaboration (Robitaille, T. P., et al.) 2013, A&A, 558, A33
Astropy Collaboration (Price-Whelan, A. M., et al.) 2018, AJ, 156, 123
Bailer-Jones, C. A. L., Rybizki, J., Fouesneau, M., Mantelet, G., & Andrae, R. 2018, AJ, 156, 58
Barentsen, G., Farnhill, H. J., Drew, J. E., et al. 2014, MNRAS, 444, 3230
Baron, D. 2019, ArXiv e-prints [arXiv:1904.07248]
Boissay, R., Parker, Q. A., Frew, D. J., & Bojicic, I. 2012, IAU Symp., 283, 316
Bouvier, J., Chelli, A., Allain, S., et al. 1999, A&A, 349, 619
Bouvier, J., Alencar, S. H. P., Harries, T. J., Johns- Krull, C. M., & Romanova, M. M. 2007, Protostars and Planets V, (eds.B. Reipurth, D. Jewitt, & K. Keil), 479
Cánovas, H., Cantero, C., Cieza, L., et al. 2019, A&A, 626, A80
Casagrande, L., & VandenBerg, D. A. 2018, MNRAS, 479, L102

Castro-Ginard, A., Jordi, C., Luri, X., et al. 2018, A&A, 618, A59
Cauley, P. W., & Johns-Krull, C. M. 2014, ApJ, 797, 112
Chollet, F., et al. 2015, Keras, https://keras.io
Cochetti, Y. R., Arcos, C., Kanaan, S., et al. 2019, A&A, 621, A123
Cody, A. M., Hillenbrand, L. A., David, T. J., et al. 2017, ApJ, 836, 41
Cutri, R. M., Wright, E. L., Conrow, T., et al. 2013, Explanatory Supplement to the AllWISE Data Release Products, Tech. Rep.
Dalton, G., Trager, S., Abrams, D. C., et al. 2018, SPIE Conf. Ser., 10702, 107021B
Delgado, A., Hodgkin, S., Evans, D. W., et al. 2019, Gaia Photometric Science Alerts Data Flow, eds. P. J. Teuben, M. W. Pound, B. A. Thomas, & E. M. Warner, ASP Conf. Ser., 523, 261
de Wit, W. J., Testi, L., Palla, F., & Zinnecker, H. 2005, A&A, 437, 247
Donehew, B., & Brittain, S. 2011, AJ, 141, 46
Drew, J. E., Greimel, R., Irwin, M. J., et al. 2005, MNRAS, 362, 753
Drew, J. E., Gonzalez-Solares, E., Greimel, R., et al. 2014, MNRAS, 440, 2036
Engels, D., & Bunzel, F. 2015, A&A, 582, A68
Evans, D. W., Riello, M., De Angeli, F., et al. 2018, A&A, 616, A4
Fairlamb, J. R., Oudmaijer, R. D., Mendigutía, I., Ilee, J. D., & van den Ancker, M. E. 2015, MNRAS, 453, 976
Frasca, A., Miroshnichenko, A. S., Rossi, C., et al. 2016, A&A, 585, A60
Frew, D. J., & Parker, Q. A. 2010, PASA, 27, 129
Gaia Collaboration (Prusti, T., et al.) 2016, A&A, 595, A1
Gaia Collaboration (Brown, A. G. A., et al.) 2018, A&A, 616, A1
Garufi, A., Benisty, M., Pinilla, P., et al. 2018, A&A, 620, A94
Gilmore, G., Randich, S., Asplund, M., et al. 2012, The Messenger, 147, 25
Gkouvelis, L., Fabregat, J., Zorec, J., et al. 2016, A&A, 591, A140
Grinin, V. P. 2000, Planetesimals, and Planets, eds. G. Garzón, C. Eiroa, D. de Winter, & T. J. Mahoney, 219, 216
Grundstrom, E. D., & Gies, D. R. 2006, ApJ, 651, L53
Hampton, E. J., Medling, A. M., Groves, B., et al. 2017, MNRAS, 470, 3395
Hartmann, L., Herczeg, G., & Calvet, N. 2016, ARA&A, 54, 135
Hedges, C., Hodgkin, S., & Kennedy, G. 2018, MNRAS, 476, 2968
Heinze, A. N., Tonry, J. L., Denneau, L., et al. 2018, AJ, 156, 241
Henden, A. A., Levine, S., & Terrell, D. 2018, Am. Astron. Soc. Meet. Abstracts, 232, 223.06
Herbig, G. H., & Bell, K. R. 1988, Third Catalog of Emission-Line Stars of the Orion Population, 3, 1988
Hernández, J., Calvet, N., Briceño, C., Hartmann, L., & Berlind, P. 2004, AJ, 127, 1682
Hillenbrand, L. A., Meyer, M. R., Strom, S. E., & Skrutskie, M. F. 1995, AJ, 109, 280
Jayasinghe, T., Stanek, K. Z., Kochanek, C. S., et al. 2019, MNRAS, 485, 961
Jiang, Y.-F., Cantiello, M., Bildsten, L., et al. 2018, Nature, 561, 498
Kerber, F., Mignani, R. P., Guglielmetti, F., & Wicenec, A. 2003, A&A, 408, 1029
Khokhlov, S. A., Miroshnichenko, A. S., Zharikov, S. V., et al. 2018, ApJ, 856, 158
Kingma, D. P., & Ba, J. 2014, ArXiv e-prints [arXiv:1412.6980]
Klement, R., Carciofi, A. C., Rivinius, T., et al. 2017, A&A, 601, A74
Koenig, X. P., & Leisawitz, D. T. 2014, ApJ, 791, 131
Ksoll, V. F., Gouliermis, D. A., Klessen, R. S., et al. 2018, MNRAS, 479, 2389
Labadie-Bartz, J., Pepper, J., McSwain, M. V., et al. 2017, AJ, 153, 252
Lallement, R., Babusiaux, C., Vergely, J. L., et al. 2019, A&A, 625, A135
Lamers, H. J. G. L. M., Zickgraf, F.-J., de Winter, D., Houziaux, L., & Zorec, J. 1998, A&A, 340, 117
Lindegren, L., Hernández, J., Bombrun, A., et al. 2018, A&A, 616, A2
Lumsden, S. L., Hoare, M. G., Urquhart, J. S., et al. 2013, ApJS, 208, 11
Małek, K., Solarz, A., Pollo, A., et al. 2013, A&A, 557, A16
Marrese, P. M., Marinoni, S., Fabrizio, M., & Altavilla, G. 2019, A&A, 621, A144
Martayan, C., Lobel, A., Baade, D., et al. 2016, A&A, 587, A115
Marton, G., Tóth, L. V., Paladini, R., et al. 2016, MNRAS, 458, 3479
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

Mathew, B., Subramaniam, A., & Bhatt, B. C. R. 2008, MNRAS, 388, 1879
Mathew, B., Manoj, P., Narang, M., et al. 2018, ApJ, 857, 30
McInnes, L., Healy, J., & Melville, J. 2018, ArXiv e-prints [arXiv:1802.03426]
Mendigutía, I., Calvet, N., Montesinos, B., et al. 2011, A&A, 535, A99
Miroshnichenko, A. S. 2007, ApJ, 667, 497
Miroshnichenko, A. S. 2017, in The B[e] Phenomenon: Forty Years of Studies, eds. A. Miroshnichenko, S. Zharikov, D. Korčáková, & M. Wolf, ASP Conf. Ser., 508, 285
Miroshnichenko, A. S., & Zharikov, S. V. 2015, EAS Publ. Ser., 71–72, 181
Miroshnichenko, A. S., Manset, N., Kusakin, A. V., et al. 2007, ApJ, 671, 828
Miroshnichenko, A. S., Polcaro, V. F., & Rossi, C. 2017, ASP Conf. Ser., 508, 387
Monnier, J. D., Millan-Gabet, R., Billmeier, R., et al. 2005, ApJ, 624, 832
Mowlavi, N., Lecoeur-Taïbi, I., Lebzelter, T., et al. 2018, A&A, 618, A58
Munari, U., Joshi, V., Banerjee, D. P. K., et al. 2019, MNRAS, 488, 5536
Natta, A., & Whitney, B. A. 2000, A&A, 364, 633
Natta, A., Grinin, V. P., Mannings, V., & Ungerechts, H. 1997, ApJ, 491, 885
Neiner, C., de Batz, B., Cochard, F., et al. 2011, AJ, 142, 149
Oudmaijer, R. D. 2017, The B[e] Phenomenon: Forty Years of Studies, eds. A. Miroshnichenko, S. Zharikov, D. Korčáková, & M. Wolf, 508, 175
Oudmaijer, R. D., Palacios, J., Eiroa, C., et al. 2001, A&A, 379, 564
Pashchenko, I. N., Sokolovsky, K. V., & Gavras, P. 2018, MNRAS, 475, 2326
Patel, P., Sigut, T. A. A., & Landstreet, J. D. 2017, ApJ, 836, 214
Pérez-Ortiz, M. F., García-Varela, A., Quiroz, A. J., Sabogal, B. E., & Hernández, J. 2017, A&A, 605, A123
Polster, J., Korčáková, D., & Manset, N. 2018, A&A, 617, A79
Raddi, R., Drew, J. E., Fabregat, J., et al. 2013, MNRAS, 430, 2169
Raddi, R., Drew, J. E., Steeghs, D., et al. 2015, MNRAS, 446, 274
Reiter, M., Calvet, N., Thanathibodee, T., et al. 2018, ApJ, 852, 5
Ribas, Á., Bouy, H., & Merín, B. 2015, A&A, 576, A52
Rimoldini, L., Holl, B., Audard, M., et al. 2019, A&A, 625, A97
Rivinius, T., Carciofi, A. C., & Martayan, C. 2013, A&ARv, 21, 69
Robin, A. C., Luri, X., Reylé, C., et al. 2012, A&A, 543, A100
Robitaille, T. P., Meade, M. R., Babler, B. L., et al. 2008, AJ, 136, 2413
Sabin, L., Parker, Q. A., Corradi, R. L. M., et al. 2014, MNRAS, 443, 3388
Sartori, M.J., Gregorio-Hetem, J., Rodrigues, C.V., et al. 2010, AJ, 139, 27
Scaringi, S., Knigge, C., Drew, J. E., et al. 2018, MNRAS, 481, 3357
Shokry, A., Rivinius, T., Mehner, A., et al. 2018, A&A, 609, A108
Sicilia-Aguilar, A., Kim, J. S., Sobolev, A., et al. 2013, A&A, 559, A3
Skrutskie, M. F., Cutri, R. M., Stiening, R., et al. 2006, AJ, 131, 1163
Snider, S., Allende Prieto, C., von Hippel, T., et al. 2001, ApJ, 562, 528
Solarz, A., Bilicki, M., Gromadzki, M., et al. 2017, A&A, 606, A39
Stanghellini, L., Shaw, R. A., & Villaver, E. 2008, ApJ, 689, 194
Stephens, I. W., Gouliermis, D., Looney, L. W., et al. 2017, ApJ, 834, 94
Suh, K.-W., & Hong, J. 2017, J. Kor. Astron. Soc., 50, 131
Szczerba, R., Siódmiak, N., Stasińska, G., & Borkowski, J. 2007, A&A, 469, 799
Taylor, M. B. 2005, ASP Conf. Ser., 347, 29
Testi, L., Palla, F., & Natta, A. 1999, A&A, 342, 515
The, P. S., de Winter, D., & Perez, M. R. 1994, A&AS, 104, 315
Venuti, L., Bouvier, J., Flaccomio, E., et al. 2014, A&A, 570, A82
Villebrun, F., Alecian, E., Hussain, G., et al. 2019, A&A, 622, A72
Vioque, M., Oudmaijer, R. D., Baines, D., Mendigutía, I., & Pérez-Martínez, R. 2018, A&A, 620, A128
Waters, L. B. F. M., Cote, J., & Geballe, T. R. 1988, A&A, 203, 348
Wheelwright, H. E., Oudmaijer, R. D., & Goodwin, S. P. 2010, MNRAS, 401, 1199
Wichittanakom, C., Oudmaijer, R. D., Fairlamb, J. R., et al. 2020, MNRAS, 493, 234
Wilson, T. J., & Naylor, T. 2018, MNRAS, 481, 2148
Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., et al. 2010, AJ, 140, 1868
Zwintz, K., Fossati, L., Ryabchikova, T., et al. 2014, Science, 345, 550

Appendix A: Disentangling Herbig Ae/Be, CBe stars, and B[e] stars

In the following we specify which classification decision was made regarding the sources that appear both as Herbig Ae/Be and classical Be in Sect. 3. Not all of them have all the observables.

– BD+41 3731 – Classical Be – It appears as a Herbig Ae/Be star in Alecian et al. (2013) and Reiter et al. (2018). However, Labadie-Bartz et al. (2017) consider it a classical Be star and Cauley & Johns-Krull (2014) suggest not to treat it as a PMS object and so did we.
– GU CMA – Herbig Ae/Be – It is generally considered as a Herbig Ae/Be star (e.g. Ababakr et al. 2017; Fairlamb et al. 2015; Reiter et al. 2018; Mathew et al. 2018).
– HBC 7 – Herbig Ae/Be – It is a bit doubtful but Hernández et al. (2004) argue that it shows chariecteristics of PMS objects.
– HD 114981 – Classical Be – It appears as Herbig Ae/Be in many papers (Reiter et al. 2018; Fairlamb et al. 2015) but as CBe in Labadie-Bartz et al. (2017). Cauley & Johns-Krull (2014) found evidence for it to be a CBe star.
– HD 130437 – Classical Be – Although it appears in The et al. (1994) as a Herbig Ae/Be star, the situation is very unclear. We decided to follow the intuition of Acke & van den Ancker (2006).
– HD 158643 – Herbig Ae/Be – Shokry et al. (2018)
– HD 174571 – Herbig Ae/Be – It displays a doubtful nature in many papers (Reiter et al. 2018; Ababakr et al. 2017; Cauley & Johns-Krull 2014; Labadie-Bartz et al. 2017) but there is a general consensus that it is a Herbig Ae/Be star.
– HD 36408 – Herbig Ae/Be – Donehew & Brittain (2011) and Cauley & Johns-Krull (2014)
– HD 37490 – Classical Be – Cauley & Johns-Krull (2014) and Cochetti et al. (2019)
– HD 50083 – Herbig Ae/Be – Reiter et al. (2018), Alecian et al. (2013), Sartori et al. (2010), Wheelwright et al. (2010), and Cauley & Johns-Krull (2014)
– HD 76534 – Herbig Ae/Be – Patel et al. (2017)
– HD 94509 – Herbig Ae/Be – Fairlamb et al. (2015)
– LkHA 350 – Herbig Ae/Be – Hernández et al. (2004)
– MWC 655 – Herbig Ae/Be – Ababakr et al. (2017) and Wheelwright et al. (2010)
– V1493 Cyg – Herbig Ae/Be – It has been little studied in the recent years but appears as a Herbig Ae/Be star in The et al. (1994) and in a few papers since then (e.g. Mathew et al. 2018) although Hernández et al. (2004) was unable to classify it.

Regarding the unclassified B[e] or FS CMa stars, almost all the confirmed FS CMa objects are listed in Lamers et al. (1998), Miroshnichenko (2007), Miroshnichenko et al. (2007, 2017), and Khokhlov et al. (2018) and they add up to 53 objects (around 70 proposed in total, Miroshnichenko & Zharikov 2015). A total of 17 FS CMa stars from this list were discarded from the sets of known PMS and CBe stars:

– BD+23 3183
– CD-24 5721
– CD-49 3441
– AS 119
– HD 328990
– HD 45677
– HD 50138
– HD 85567
– Hen 3-847
– LkHA 348
– MWC 1055
– MWC 342
– MWC 657
– PDS 021
– PDS 211
– V2211 Cyg
– V669 Cep

Separately, but related, in Vioque et al. (2018) it was found that because of their positions on the HR diagram: MWC 314, MWC 623, and MWC 930 were not very likely to be PMS objects. Indeed, MWC 314 seems to be a supergiant B[e] star (Frasca et al. 2016), MWC 930 looks like a luminous blue variable (Martayan et al. 2016; Jiang et al. 2018) and MWC 623 seems clear to be a FS CMa star (Miroshnichenko 2007; Polster et al. 2018). Therefore, we also removed these three objects from our set of known HAeBes (Sect. 3.1).

Appendix B: Algorithm and methodology

The pipeline used can be seen in Fig. B.1. Most of the algorithm described in it is available on a GitHub repository[3] under the name of YODA (Young Object Discoverer Algorithm). In this appendix we describe this pipeline in detail.

B.1. Class weights

As explained in Sect. 3.4, the set of other objects is contaminated by undiscovered PMS and CBe stars. This causes the ANN used in Sect 3.4 to obtain a decent precision but a terrible recall (see Fig. 2). As discussed in Sect. 3.4, this is because we are only retrieving the less common sources. Any pre-classifications performed with the observables used by the algorithm would artificially bias the results. One option is to use simulations to generate this well defined category of other objects without PMS and CBe stars (as done for example by Castro-Ginard et al. 2018) but there is none that lists PMS and CBe objects and contains IR and $H\alpha$ information.

We can address this issue by changing the weights of the sources used in the training, in a way that they are balanced for the different category sizes. Hence, during training the ANN is heavily penalized when failing at categorising PMS or CBe sources, but lightly affected by mistakes on the other objects category, which is much larger. Therefore, the training is not dominated by contaminants or undiscovered sources, although they still are considered. This weighting technique produces a decent recall and a very low precision, but this precision is just a lower limit as the candidate regions of the feature space contain many undiscovered PMS and CBe stars. In other words, FP is over-measured (Eq. (3)). In addition, these class weights stress the algorithm to focus on the differences between PMS sources and CBe stars, which also bias the selection of PMS sources towards the high-mass end.

B.2. Architecture selection

Different machine learning algorithms were considered for this classification problem. A variety of them have been used so far for similar matters. For example: random forests (Hedges et al. 2018; Marton et al. 2019; Rimoldini et al. 2019), support vector machines (Małek et al. 2013; Marton et al. 2016;
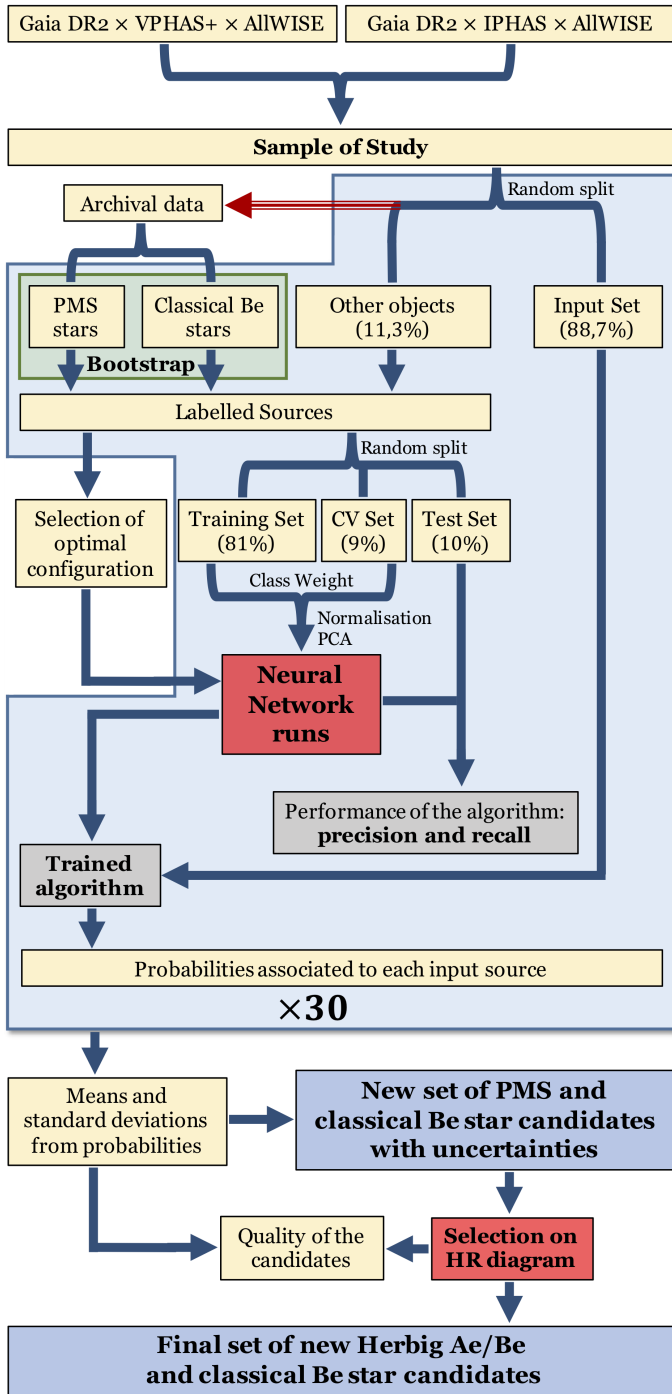
Fig. B.1. Pipeline of the whole algorithm, from the cross-match of *Gaia* DR2, AllWISE, IPHAS, and VPHAS+ to the set of new HAeBes and classical Be stars. The light blue area indicates the set of processes that are repeated in a loop 30 times, each time generating a different set of probabilities associated to each input source. The green area shows the bootstrapped sets. The red arrow indicates that the Archival data is partially contained within the Sample of Study, but was also constructed using external information like the Hα EWs.

Solarz et al. 2017; Ksoll et al. 2018), or artificial neural networks (Snider et al. 2001; Hampton et al. 2017). However, similar performances are achieved with most of the algorithms and it is evident that the output is mainly dominated by the quality of the training data (Pérez-Ortiz et al. 2017; Pashchenko et al. 2018, or Marton et al. 2019 in a similar problem of identifying

YSOs). Therefore, we decided to use a shallow artificial neural network as it has the advantage of flexibility and non-linearity, being able to describe very complex and subtle relations. In addition, its output can be a probability vector, which eases the catalogue construction. Cons are the number of hyper-parameters required, which are normally hard to interpret.

Therefore, we needed to find the architecture or optimal configuration of the ANN for our particular problem. This means choosing the hyper-parameters of the ANN (e.g. layers, neurons per layer, regularisation). Ideally, this architecture would be selected with cross-validation (CV) sets that are not used for the training. However, our sample of known PMS and CBe stars is too small to have the number of CV sets necessary to test a large enough grid of ANN configurations. Instead, we used the set of Labelled Sources to select the optimal hyper-parameters and then, independently, used those hyper-parameters and that same set of Labelled Sources for the training (see Fig. B.1). We ran an ANN over the Labelled Sources set 100 times (each time with a 10% test set random split); evaluating at each training iteration on a CV set (sized 10%) and early-stopping whenever the precision of the algorithm on the PMS category (selecting as candidates those with a probability $p \geq 50\%$ of belonging to such category) stopped increasing over 250 iterations. In addition, we imposed that the recall had to be at least 90%. In each run a different grid of hyper-parameters was used. After the 100 runs, the best architecture resulted in two fully connected hidden layers of 580 neurons, with a dropout rate of 50% and 0.01 L2 regularisation. Batch Normalisation was applied after every layer, though no batches were used and the whole training data was evaluated in each training iteration (as it is a very skewed training set, see Sect. 3).

The activation functions used were "ReLU" for the hidden layers and "softmax" for the output layer. This is because softmax output can be interpreted as a probability distribution. The loss function used was "categorical crossentropy" with the "AdaMax" optimizer (Kingma & Ba 2014). To construct the ANNs of this project we used Keras (Chollet et al. 2015), a high-level neural networks application programming interface.

B.3. Training, cross-validation, and test set

We shuffle the Labelled Sources set and randomly split it into two subsets (see Fig. B.1). One contains 90% of the sources and is used to train the algorithm (training set). The other, containing 10% of the sources is used to evaluate its performance (test set).

The first step is to perform feature scaling and mean normalisation to the observables, so they all have the same mean and standard deviation. Then we apply PCA to the scaled observables to get the set of features used by the ANN (12 principal components of the 48 carry 99.99% of the variance, see Sect. 2). Next, we train the ANN, which has the architecture chosen in Appendix B.2, with the training set and use a CV set (sized 10% of the training set) to evaluate the ANN performance after every training iteration. Early-stopping finishes the training whenever the precision on the PMS category (with $p \geq 50\%$) stops increasing over 50 iterations. We note that, as discussed in Appendix B.1, the precision retrieved is just a lower limit. Once the ANN is trained, we run it on the test set, that needs to be scaled and feature extracted as done for the training set. Evaluation on test set gives a value of precision and recall for each probability threshold for classification p (i.e. the performance of the algorithm, see Fig. B.2). Finally, we can apply the trained ANN to the Input Set, giving a probability for every source of belonging to each of the chosen categories.
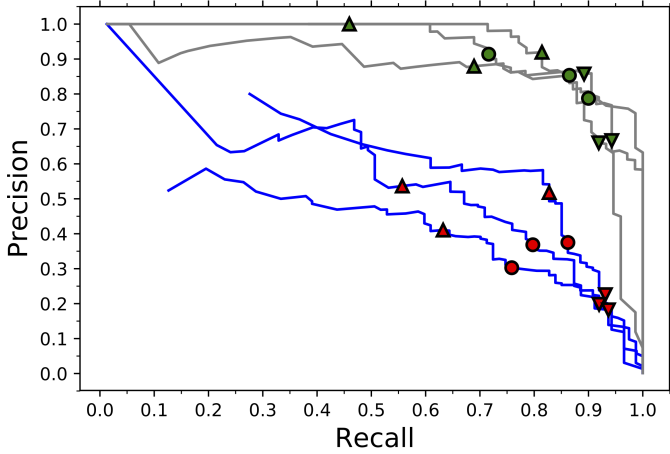
Fig. B.2. Precision vs. recall trade-off plot resulting after evaluation on test set for three different bootstrapped iterations. Blue lines correspond to PMS classification and grey lines to classical Be classification. Different probability thresholds (p) for selecting candidate objects correspond to different locations on the line. Upwards arrows ($p \geq 75\%$), circles ($p \geq 50\%$), and downwards arrows ($p \geq 25\%$) are examples of such probability thresholds. Some lines do not cover the whole metric space because evaluation stops whenever there are no longer true positives in the corresponding test set. The precision values are lower limits to the real precisions.

B.4. Bootstrap

A major issue is the small size of the PMS and classical Be categories in the training data (see Sect. 3). Small training sets imply that outliers and contaminants have a very strong influence and might dominate the posterior generalisation. In addition, the training might be biased to any hidden trend or pattern.

One way to minimise the impact of this is by means of the bootstrap. The key idea is to fake the construction of new training sets. It works by repeatedly sampling the original training data and randomly substituting sources with others of the same data set. If we run the same algorithm over two bootstrapped sets we obtain similar, but slightly different metrics as a result. If we repeat this bootstrapping process a large enough number of times we end up with a distribution of precisions and recalls characteristic of our method, which allows us to estimate the uncertainty of the metrics for each probability threshold. Bootstrapping has another advantage, which is to better represent the distribution of the categories on the feature space and minimise the impact of outliers.

Therefore, we run the processes described in Appendix B.3 (blue area of Fig. B.1) 30 times in a loop. In each iteration, we create a bootstrapped version of the combination of the categories of known PMS and classical Be stars (so the number of objects in each group is not conserved). In the case of the other objects category, we just withdraw another random set of sources from the Sample of Study. Once the algorithm is trained with a certain Labelled Sources bootstrapped set, we obtain by evaluating on the corresponding test set values for the precision and recall at different probability thresholds (see Fig. B.2). When we run the trained ANN over the Input Set, we retrieve probabilities associated to every source of belonging to each of the

three categories. Hence, after the bootstrapped iterations we end up with 30 values for precision and recall at different thresholds (in Fig. B.2 only three bootstrapped iterations are shown for clarity) and 30 sets of probabilities associated to each source of the Sample of Study. This is because as the category of other objects has been randomly sampled 30 times, the whole SoSt has been covered eventually. To obtain the final values presented in Tables D.1 and D.2 we average these 30 repetitions and take the standard deviation of the mean as the uncertainty of each measurement.

Appendix C: Visualisation

It is possible to visualise the feature space and the selection using a dimensionality reduction algorithm. We used the UMAP algorithm (Uniform Manifold Approximation and Projection for Dimension Reduction, McInnes et al. 2018) to project the 12-dimensional feature space (Sect. 2.2) into two dimensions. This is done in Fig. C.1. At left we project the candidates using an euclidean metric (15% of the number of sources as number of neighbours and a minimum distance of 0.4). In Fig. C.1 at right we project the known HAeBe, T-Tauri, and CBe stars used for the training (Sect. 3) onto this same plane, which is colour-coded following the PMS probability distribution of the sources at left. This dimensionality reduction helps to understand the catalogue construction and to find trends within the data. However, information is lost when moving to 2D. The category of other objects was not included here because of size limitations.

First, we can see that there is indeed a separation between known PMS sources and classical Be stars (Fig. C.1 at right), which is used by the algorithm to learn how to separate these populations (Fig. C.1 at left). It is remarkable that most of the retrieved CBe candidates are not found where known CBe stars are, but even farther away from the PMS region, which might imply that we are retrieving very extreme CBe candidates (see Sect. 5.2). In addition, most of the PMS candidates that are located very close to the CBe region are those with a high CBe probability and vice-versa (see Fig. 3). These PMS candidates have $r - H\alpha$ and $G_{BP} - G_{RP}$ values typical of known CBes and low IR-excesses for PMS objects. However, these are still typically larger than the ones of known CBes. This, together with strong G_{var} variability explains their selection as PMS.

Second, in Fig. C.1 an arm structure at the very top left of the space of PMS candidates seems special. A closer look at the 204 PMS candidates located in that arm shows that they are all placed in the red giant region of the HR diagram (see Fig. 4 at top left). They differentiate from the rest of the PMS candidates in that they have more variability in our indicators and are typically brighter. In addition, they show on average larger near-IR excesses but lower mid-IR excesses. None of them have reliable astrometry so they do not contaminate our sample of Herbig Ae/Be candidates. Potentially, they are evolved contaminants and are flagged in Table D.1 as "G_{UMAP}" (see Sect. 5.2 for further details of their nature). We cannot exclude sources only because of their HR diagram position, as we might remove HAeBes with high extinctions and some candidates do not have parallax information. Hence, candidates in similar red giant HR diagram positions but not in this UMAP region are not flagged in the final catalogues of Tables D.1 and D.2.

Fig. C.1. UMAP dimensionality reduction from the 12D space of features to 2D. *Left*: dimensionality reduction of the PMS and classical Be candidates, together with those sources which belong to either category (i.e. $p(\text{PMS}) + p(\text{CBe}) \geq 50\%$ but $p(\text{PMS}) < 50\%$, $p(\text{CBe}) < 50\%$). *Right*: we project the known Herbig Ae/Be, T-Tauri, and classical Be stars used for the training onto the same plane, which is colour-coded following the PMS probability distribution of the sources at left.

Appendix D: Catalogue of new PMS and classical Be stars

Here we present a portion of the catalogue of new pre-main sequence (Table D.1) and classical Be (Table D.2) stars for guidance regarding its form and content. These tables are available at the CDS in their entirety with uncertainties of the magnitudes, quality flags and rest of *Gaia* parameters together with angular distances from AllWISE and IPHAS or VPHAS+ sources to *Gaia* sources. Below a description of the possible warning flags of Tables D.1 and D.2 by alphabetical order. See the main text for further discussion.

– G_{UMAP}: Possible evolved star contaminant. Identified through UMAP visualisation. Discussed in Appendix C.

– ID AllW: Source with an AllWISE name repeated in the Sample of Study. Discussed in Sect. 5.2, point 3.

– ID IPH/VPH: Source with an IPHAS or VPHAS+ name repeated in the Sample of Study. Discussed in Sect. 5.2, point 3.

– PN: Possible planetary nebula or "unclassified B[e]" contaminant. Defined as those candidates with $r - H\alpha \geq 1.3$. Discussed in Sect. 5.2, point 5.

– Var: Photometrically variable PMS candidate. Defined as those PMS candidates with $G_{var} \geq 10$. Discussed in Sect. 4.4.

– W3W4: Source which extended source flag of AllWISE catalogue is different of 0. Discussed in Sect. 5.2, point 2.

– X-mtch: Likely false candidate because of incorrect cross-match with IPHAS or VPHAS+. Discussed in Sect. 4.

Table D.1. PMS candidates ($p \geq 50\%$, 8470 sources) ordered by probability.

Gaia source id	RA h:m:s	Dec deg:m:s	Probability PMS	Probability CBe	Probability Other	V_{htg}	G_{var}
189190423171051136	**05:39:18.1**	**+36:17:16**	**1.00±0.00**	**$(7.8 \pm 2.0) \times 10^{-6}$**	**$(4.2 \pm 1.6) \times 10^{-6}$**	**−52.13**	**158.74**
513068993519575808	01:39:32.5	+64:53:02	1.00±0.00	$(4.18 \pm 0.91) \times 10^{-5}$	$(7.7 \pm 2.9) \times 10^{-6}$	−66.16	165.17
5546522453993928960	08:12:40.5	−34:14:12	1.00±0.00	$(4.4 \pm 2.0) \times 10^{-5}$	$(1.38 \pm 0.58) \times 10^{-5}$	−117.08	225.56
181175743617347328	05:23:04.3	+33:28:46	1.00±0.00	$(1.67 \pm 0.44) \times 10^{-4}$	$(4.7 \pm 2.0) \times 10^{-6}$	−47.67	131.31
3430718965791698176	**05:52:51.7**	**+26:47:30**	**1.00±0.00**	**$(3.04 \pm 0.96) \times 10^{-4}$**	**$(3.6 \pm 1.0) \times 10^{-5}$**	**−67.90**	**152.91**
3046391307734381184	07:09:22.3	−10:30:57	1.00±0.00	$(2.06 \pm 0.55) \times 10^{-4}$	$(1.52 \pm 0.57) \times 10^{-4}$	−44.33	77.08
3449189833426211840	05:32:13.9	+34:06:01	1.00±0.00	$(8.0 \pm 2.8) \times 10^{-5}$	$(2.8 \pm 1.8) \times 10^{-4}$	−107.39	183.40
189577661722195840	05:37:53.2	+37:24:56	1.00±0.00	$(2.61 \pm 0.60) \times 10^{-4}$	$(1.52 \pm 0.51) \times 10^{-4}$	−66.15	117.40
2006046771487587328	22:24:13.8	+56:11:33	1.00±0.00	$(2.3 \pm 1.1) \times 10^{-4}$	$(2.38 \pm 0.99) \times 10^{-4}$	−117.72	162.47
4279295067717293696	**18:47:32.6**	**+02:12:06**	**1.00±0.00**	**$(1.31 \pm 0.21) \times 10^{-4}$**	**$(3.82 \pm 0.94) \times 10^{-4}$**	**−33.04**	**96.61**
3114712967419900544	07:09:42.8	+01:53:11	1.00±0.00	$(5.5 \pm 2.4) \times 10^{-5}$	$(5.4 \pm 4.0) \times 10^{-4}$	−46.52	143.75
3102576519414606464	06:58:34.4	−03:56:46	1.00±0.00	$(1.75 \pm 0.58) \times 10^{-4}$	$(4.3 \pm 1.2) \times 10^{-4}$	−26.55	58.08
...	...	...	...	...	...	...	...

G_{BP} (mag)	G (mag)	G_{RP} (mag)	IPHAS or VPHAS+ name	r (mag)	i (mag)	Hα (mag)	AllWISE name	J (mag)	H (mag)	K_s (mag)
15.97	**15.04**	**14.04**	**J053918.09+361716.2**	**13.78**	**13.06**	**13.21**	**J053918.08+361716.1**	**12.98**	**11.23**	**9.79**
14.34	13.80	12.98	J013932.55+645302.3	12.91	12.43	12.35	J013932.53+645302.2	11.43	10.21	9.11
14.97	14.18	13.20	J081240.5−341411.7	12.86	12.19	11.99	J081240.50−341411.7	10.86	10.08	9.59
14.89	14.40	13.66	J052304.26+332846.5	13.80	13.24	12.90	J052304.26+332846.4	12.01	11.12	10.25
14.40	**13.94**	**13.05**	**J055251.75+264730.2**	**12.62**	**12.22**	**12.28**	**J055251.74+264730.0**	**12.16**	**11.04**	**10.06**
16.75	15.56	14.31	J070922.3−103057.0	16.04	14.94	14.78	J070922.28−103057.0	12.47	10.98	9.90
16.17	14.76	13.63	J053213.92+340601.5	13.70	12.39	13.25	J053213.92+340601.4	10.80	9.83	8.83
14.73	13.68	12.63	J053753.15+372456.1	14.30	13.07	13.80	J053753.15+372456.0	10.54	9.83	9.31
14.00	13.31	12.40	J222413.78+561133.2	13.63	12.92	13.29	J222413.75+561133.2	9.56	8.69	8.01
15.15	**14.11**	**13.05**	**J184732.62+021205.9**	**15.01**	**13.84**	**14.23**	**J184732.61+021205.7**	**11.34**	**9.97**	**8.79**
13.62	11.79	10.19	J070942.85+015311.0	13.86	11.84	13.15	J070942.75+015311.1	8.20	6.62	5.25
18.24	17.57	16.41	J065834.4−035645.6	21.41	20.84	20.13	J065834.39−035645.6	13.73	12.40	11.38
...	...	...	...	...	...	...	...	...	...	...

$W1$ (mag)	$W2$ (mag)	$W3$ (mag)	$W4$ (mag)	RUWE	Parallax (ϖ) [mas]	Distance [pc]	A'_G (mag)	M_G (mag)	$G_{BP} - G_{RP}$ (mag)	HAeBe	Flag
7.90	**7.05**	**4.84**	**3.23**	**1.36**	**0.636 ± 0.048**	**1510^{+120}_{-100}**	**1.06**	**3.09**	**1.41**	**Yes**	**Var**
7.95	7.26	5.41	3.90	1.86	1.116 ± 0.029	874^{+23}_{-22}	–	–	–	–	Var
8.63	7.81	5.37	3.32	1.33	2.877 ± 0.025	$344.2^{+3.0}_{-2.9}$	0.06	6.43	1.75	–	Var, W3W4
9.12	8.39	6.35	4.21	1.37	0.234 ± 0.040	3710^{+650}_{-490}	1.04	0.51	0.72	–	Var, W3W4
9.26	**8.76**	**7.23**	**5.56**	**1.09**	**1.950 ± 0.032**	**$505.6^{+8.3}_{-8.0}$**	**0.27**	**5.15**	**1.22**	**Yes**	**Var**
7.19	5.90	3.19	1.06	3.28	1.00 ± 0.20	1020^{+320}_{-200}	–	–	–	–	Var, W3W4
6.92	5.79	3.95	2.94	1.01	−0.068 ± 0.080	7200^{+2500}_{-1800}	–	–	–	–	G_{UMAP}, Var
8.34	7.37	5.01	3.66	1.32	0.025 ± 0.041	8200^{+2300}_{-1600}	–	–	–	–	Var, W3W4
7.06	5.99	3.79	2.43	1.72	0.326 ± 0.027	2810^{+230}_{-200}	–	–	–	–	Var, W3W4
7.61	**6.97**	**4.84**	**3.10**	**1.30**	**1.614 ± 0.046**	**609^{+18}_{-17}**	**1.99**	**3.19**	**1.13**	**Yes**	**Var**
3.99	1.79	1.93	1.44	1.01	0.103 ± 0.089	4500^{+1600}_{-1100}	–	–	–	–	G_{UMAP}, Var, W3W4
10.58	9.25	6.40	4.02	1.15	0.36 ± 0.11	2580^{+1120}_{-640}	–	–	–	–	Var, W3W4
...	...	...	...	...	...	...	...	...	...	...	...

Notes. In boldface sources that are strong Herbig Ae/Be candidates according to their position on the HR diagram ($M_G < 6$, 1361, see Fig. 5). These Herbig Ae/Be candidates are also indicated in the column "HAeBe". This table is available in its entirety at the CDS with uncertainties of the magnitudes, quality flags, and rest of Gaia parameters together with angular distances from AllWISE and IPHAS or VPHAS+ sources to Gaia sources. A portion is shown here for guidance regarding its form and content. The probabilities are expressed to the precision of their uncertainties. Distances from Bailer-Jones et al. (2018). A'_G only traces the interstellar extinction. It is used to correct M_G and $G_{BP} - G_{RP}$. A'_G, M_G, and $G_{BP} - G_{RP}$ are only presented for sources with RUWE<1.4 and $\varpi/\sigma(\varpi) \geq 5$. By construction Herbig Ae/Be candidates are astrometrically well behaved (RUWE<1.4 and $\varpi/\sigma(\varpi) \geq 10$). These constraints can be relaxed to obtain more Herbig Ae/Be candidates. The probabilities of the three categories sum up to 1. The different catalogue warning flags are discussed throughout the text and summarised in Appendix D.

Table D.2. Representative sample of the full table of classical Be candidates ($p \geq 50\%$, 693 sources) ordered by probability.

Gaia source id	RA h:m:s	Dec deg:m:s	Probability PMS	Probability CBe	Probability Other	$V_{\rm htg}$	$G_{\rm var}$
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
2012831922144678272	23:43:35.7	+61:27:44	0.119±0.010	0.831±0.015	0.0500±0.0061	−1.87	7.75
3106114885277650944	06:49:27.0	−02:33:30	0.105±0.014	0.831±0.020	0.064±0.010	−1.36	2.40
509370275062324608	01:34:15.6	+59:30:59	0.146±0.012	0.830±0.014	0.0239±0.0032	−7.87	13.95
188951970886326656	05:11:53.5	+40:13:11	0.1127±0.0099	0.830±0.015	0.0572±0.0081	0.14	1.62
2173867430250308608	21:50:31.8	+54:04:53	0.122±0.016	0.830±0.019	0.0483±0.0058	−10.14	15.16
4268987932187374848	19:07:10.6	+03:29:22	0.107±0.010	0.828±0.017	0.0644±0.0090	−0.42	6.07
518388499496394752	01:58:16.4	+65:49:52	0.107±0.010	0.828±0.015	0.0647±0.0071	−7.34	11.67
2201241387128842240	22:18:59.0	+59:47:10	0.137±0.018	0.827±0.022	0.0356±0.0083	0.00	0.67
2060197062012903296	20:00:28.0	+37:06:52	0.103±0.012	0.827±0.017	0.0700±0.0080	0.08	1.18
3325867711602501504	06:21:13.0	+09:07:21	0.101±0.011	0.827±0.019	0.072±0.016	−0.23	0.99
3106856785040836224	06:37:36.4	−01:54:03	0.107±0.010	0.824±0.016	0.0694±0.0079	−6.52	12.02
3356939124228643968	06:35:02.5	+15:17:52	0.125±0.014	0.823±0.017	0.0518±0.0061	−0.27	0.87
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .

$G_{\rm BP}$ (mag)	G (mag)	$G_{\rm RP}$ (mag)	IPHAS or VPHAS+ name	r (mag)	i (mag)	Hα (mag)	AllWISE name	J (mag)	H (mag)	$K_{\rm s}$ (mag)
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
13.14	12.80	12.26	J234335.71+612744.0	12.92	12.39	12.28	J234335.70+612743.9	11.59	11.36	11.15
12.93	12.67	12.18	J064926.96–023330.0	12.56	12.14	11.88	J064926.98–023330.5	11.41	11.24	10.87
12.27	11.99	11.51	J013415.63+593058.7	12.00	11.60	11.40	J013415.64+593058.6	10.98	10.69	10.48
12.66	12.43	11.99	J051153.45+401310.9	12.50	11.92	11.64	J051153.44+401310.9	11.55	11.38	11.21
12.87	12.32	11.58	J215031.85+540452.9	12.73	11.91	11.94	J215031.83+540452.7	10.71	10.41	10.24
12.30	12.04	11.59	J190710.55+032922.3	12.15	11.60	11.37	J190710.55+032922.4	10.95	10.74	10.54
12.81	12.32	11.67	J015816.40+654951.9	12.16	11.56	11.61	J015816.37+654951.9	10.89	10.65	10.46
12.73	12.33	11.75	J221859.05+594710.0	12.60	11.84	12.18	J221859.03+594709.9	11.02	10.90	10.77
12.43	12.25	11.94	J200028.00+370651.9	12.33	11.93	11.99	J200028.00+370653.0	11.54	11.45	11.37
11.99	11.90	11.66	J062112.96+090721.4	12.48	11.71	11.44	J062112.95+090721.4	11.37	11.28	11.13
13.13	12.68	12.06	J063736.41–015403.0	12.58	12.04	12.21	J063736.41–015402.9	11.26	11.02	10.84
12.56	12.37	12.01	J063502.46+151752.5	12.50	12.05	12.11	J063502.46+151752.5	11.57	11.52	11.37
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .

$W1$ (mag)	$W2$ (mag)	$W3$ (mag)	$W4$ (mag)	RUWE	Parallax [mas]	Distance [pc]	A'_G (mag)	M_G (mag)	$G_{\rm BP}-G_{\rm RP}$ (mag)	Flag
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
10.82	10.56	9.76	9.15	0.96	0.321 ± 0.023	2850^{+200}_{-180}	2.31	−1.78	−0.25	–
10.23	9.99	9.35	9.00	0.89	0.356 ± 0.035	2600^{+270}_{-230}	1.37	−0.78	0.08	W3W4
10.14	9.83	8.92	7.70	1.11	0.286 ± 0.033	3130^{+360}_{-300}	1.36	−1.85	0.09	–
10.63	10.33	9.40	8.21	1.30	0.108 ± 0.044	6000^{+1700}_{-1200}	–	–	–	–
9.17	8.91	8.33	7.50	1.03	0.256 ± 0.028	3470^{+390}_{-320}	2.11	−2.50	0.25	W3W4
10.23	10.03	9.32	8.71	0.82	0.388 ± 0.078	2400^{+590}_{-400}	–	–	–	W3W4
10.25	10.00	9.31	8.19	0.99	0.347 ± 0.034	2660^{+270}_{-230}	2.47	−2.27	−0.07	–
10.68	9.75	10.42	8.46	0.95	1.137 ± 0.032	858^{+24}_{-23}	1.52	1.14	0.24	W3W4
10.15	9.53	9.55	8.86	1.06	0.864 ± 0.027	1121^{+35}_{-33}	0.61	1.39	0.18	W3W4
11.06	10.84	9.91	8.63	0.93	0.340 ± 0.046	2710^{+410}_{-320}	0.81	−1.08	−0.06	X-mtch
10.47	10.22	9.40	8.70	0.90	0.217 ± 0.036	3990^{+670}_{-510}	1.48	−1.80	0.35	W3W4
11.20	10.78	10.53	8.66	0.97	0.446 ± 0.056	2110^{+290}_{-230}	0.91	−0.16	0.11	–
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .

Notes. This table is available in its entirety at the CDS with uncertainties of the magnitudes, quality flags, and rest of *Gaia* parameters together with angular distances from AllWISE and IPHAS or VPHAS+ sources to *Gaia* sources. A portion is shown here for guidance regarding its form and content. The probabilities are expressed to the precision of their uncertainties. Distances from Bailer-Jones et al. (2018). A'_G only traces the interstellar extinction. It is used to correct M_G and $G_{BP} - G_{RP}$. A'_G, M_G, and $G_{BP} - G_{RP}$ are only presented for sources with RUWE<1.4 and $\varpi/\sigma(\varpi) \geq 5$. The probabilities of the three categories sum up to 1. The different catalogue warning flags are discussed throughout the text and summarised in Appendix D. We note that some of the sources presented here have a "X-mtch" flag (57 out of 693), and hence they are likely false candidates because of an incorrect cross-match with IPHAS or VPHAS+.