



Universiteit  
Leiden  
The Netherlands

## Hunting for open clusters in Gaia EDR3: 628 new open clusters found with OCfinder

Castro Ginard, A.; Jordi, C.; Luri, X.; Cantat-Gaudin, T.; Carrasco, J.M.; Casamiquela, L.; ... ; Badia, R.M.

### Citation

Castro Ginard, A., Jordi, C., Luri, X., Cantat-Gaudin, T., Carrasco, J. M., Casamiquela, L., ... Badia, R. M. (2022). Hunting for open clusters in Gaia EDR3: 628 new open clusters found with OCfinder. *Astronomy & Astrophysics*, 661. doi:10.1051/0004-6361/202142568

Version: Accepted Manuscript

License: [Leiden University Non-exclusive license](#)

Downloaded from: <https://hdl.handle.net/1887/3515381>

**Note:** To cite this publication please use the final published version (if applicable).

# Hunting for open clusters in *Gaia* EDR3: 628 new open clusters found with OCfinder<sup>★</sup>

A. Castro-Ginard<sup>1,2</sup>, C. Jordi<sup>2</sup>, X. Luri<sup>2</sup>, T. Cantat-Gaudin<sup>2,3</sup>, J.M. Carrasco<sup>2</sup>, L. Casamiquela<sup>4</sup>, F. Anders<sup>2</sup>, L. Balaguer-Núñez<sup>2</sup>, and R.M. Badia<sup>5</sup>

<sup>1</sup> Leiden Observatory, Leiden University, Niels Bohrweg 2, 2333 CA Leiden, Netherlands  
e-mail: [acastro@strw.leidenuniv.nl](mailto:acastro@strw.leidenuniv.nl)

<sup>2</sup> Dept. Física Quàntica i Astrofísica, Institut de Ciències del Cosmos (ICCUB), Universitat de Barcelona (IEEC-UB), Martí i Franquès 1, E08028 Barcelona, Spain

<sup>3</sup> Max-Planck-Institut für Astronomie, Königstuhl 17, D-69117 Heidelberg, Germany

<sup>4</sup> Laboratoire d'Astrophysique de Bordeaux, Univ. Bordeaux, CNRS, B18N, allée Geoffroy Saint-Hilaire, 33615 Pessac, France

<sup>5</sup> Barcelona Supercomputing Center (BSC), Barcelona, Spain

Received date / Accepted date

## ABSTRACT

**Context.** The improvements in the precision of the published data in *Gaia* EDR3 with respect to *Gaia* DR2, particularly for parallaxes and proper motions, offer the opportunity to increase the number of known open clusters in the Milky Way by detecting farther and fainter objects that have thus far gone unnoticed.

**Aims.** Our aim is to continue to complete the open cluster census in the Milky Way with the detection of new stellar groups in the Galactic disc. We use *Gaia* EDR3 up to magnitude  $G = 18$  mag, increasing the magnitude limit and therefore the search volume explored in one unit in our previous studies.

**Methods.** We used the OCfinder method to search for new open clusters in *Gaia* EDR3 using a big data environment. As a first step, OCfinder identified stellar statistical overdensities in five-dimensional astrometric space (position, parallax, and proper motions) using the DBSCAN clustering algorithm. Then, these overdensities were classified into random statistical overdensities or real physical open clusters using a deep artificial neural network trained on well-characterised  $G$ ,  $G_{BP} - G_{RP}$  colour-magnitude diagrams.

**Results.** We report the discovery of 628 new open clusters within the Galactic disc, with most of them being located beyond 1 kpc from the Sun. From the estimation of ages, distances, and line-of-sight extinctions of these open clusters, we see that young clusters align following the Galactic spiral arms while older ones are dispersed in the Galactic disc. Furthermore, we find that most open clusters are located at low Galactic altitudes with the exception of a few groups older than 1 Gyr.

**Conclusions.** We show the success of the OCfinder method leading to the discovery of a total of 1 274 open clusters (joining the discoveries here with the previous ones based on *Gaia* DR2), which represents almost 50% of the known population. Our ability to perform big data searches on a large volume of the Galactic disc, together with the higher precision in *Gaia* EDR3, enable us to keep completing the census with the discovery of new open clusters.

**Key words.** Galaxy: disc — open clusters and associations: general — astrometry — Methods: data analysis

## 1. Introduction

Open clusters (OCs) have historically been used to study the structural, kinematical, and chemical properties of the disc of the Milky Way, and its evolution. For this reason, in recent years, there has been a growing interest in building an accurate and complete view of the OC population. In particular, after the *Gaia* Second Data Release (*Gaia* DR2, [Gaia Collaboration et al. 2018](#)), the study of this field was revolutionised by the re-definition of the OC population, with [Cantat-Gaudin et al. \(2018\)](#) refusing around 50% of the OCs reported in pre-*Gaia* catalogues as not real OCs ([Dias et al. 2002](#); [Kharchenko et al. 2013](#)). Furthermore, *Gaia* DR2 enabled the systematic detection of new OCs using machine-learning methods, which outperform traditional manual methods to search for these objects. [Castro-Ginard et al. \(2018\)](#), hereafter PaperI) were able to find 23 new OCs in *Gaia* DR1's TGAS subset ([Gaia Collaboration et al. 2016](#); [Michalik et al. 2015](#)), and then they used the same

methodology to detect more than 600 new OCs in the Galactic disc using *Gaia* DR2, which represents about one-third of the currently known OC population ([Castro-Ginard et al. 2019, 2020](#), hereafter PaperII and PaperIII, respectively). Since then, several publications have made use of machine-learning-based methods to detect new OCs ([Cantat-Gaudin et al. 2019](#); [Sim et al. 2019](#); [Liu & Pang 2019](#); [Ferreira et al. 2020](#); [Hunt & Reffert 2021](#)), computing membership lists ([Cantat-Gaudin & Anders 2020](#); [Jaehnig et al. 2021](#)) or characterising their astrophysical properties ([Bossini et al. 2019](#); [Cantat-Gaudin et al. 2020](#); [Dias et al. 2021](#)).

Several studies have combined *Gaia* astrometry with ground-based radial velocities with the purpose of studying the kinematics of the OC population (e.g. [Soubiran et al. 2018](#); [Carrera et al. 2022](#)). [Tarricq et al. \(2021\)](#) studied the 3D kinematics and age dependence of the OC population, also providing orbital parameters for 1 382 OCs. Other studies have used OCs' available astrometric and kinematic information to trace the spiral structure in the Milky Way ([Dias & Lépine 2005](#)). On this topic, [Monteiro](#)

<sup>★</sup> Full Table 1 and Table 2 are only available at the CDS.

et al. (2021) found that the behaviour of the spiral arms could be explained by classical density waves. However, Castro-Ginard et al. (2021) recently found a transient nature of the arms dis-favouring classic density waves as the main drivers of the spiral structure, which is also supported by studies using tracers other than OCs (Minniti et al. 2021; Colombo et al. 2022). These contradicting results show the need to keep improving the OC census.

The coupling of *Gaia* data with the detailed abundance results of large spectroscopic surveys has allowed for a more complete picture of the chemical composition of the OC population to be sketched. In all of these studies, an unbiased and complete census of the OCs of the Milky Way is needed to tackle the chemical evolution of our Galaxy. Detailed chemical abundance radial gradients in the Milky Way and their age dependence have been characterised (e.g. Carrera et al. 2019; Spina et al. 2021). Temporal dependencies of chemical abundance ratios are commonly calibrated using OCs due to their precise age determination (Casamiquela et al. 2021b). Finally, OCs are used as test cases to explore the feasibility of a diversity of techniques such as strong chemical tagging, that is to say the possibility of finding stars that were born in the same star-forming event (Casamiquela et al. 2021a).

The latest release of *Gaia* data (EDR3, Gaia Collaboration et al. 2021), providing astrometric measurements for about 1.8 billion stars with improved precision with respect to *Gaia* DR2, offers the opportunity to re-visit the OC census and keep improving it, both in terms of a better characterisation and new discoveries. The improved precisions in parallax, and particularly in proper motions with respect to *Gaia* DR2, allow for the application of machine-learning methods to search for new structures that would go unnoticed with traditional methods, which mostly relied on visual inspection. In this context, we adapted our methodology (developed and applied in Paper I, Paper II, and Paper III), which we dub OCfinder, to search for unknown OCs in *Gaia* EDR3 using its astrometric and photometric data.

This paper is organised as follows. In Sect. 2, we describe the data used to search for OCs. The OCfinder method used for that purpose is described in Sect. 3. Section 4 describes the OCs found, both re-detected and new findings. Finally, our conclusions are presented in Sect. 5.

## 2. Data

The data were used in this paper to search for unknown OCs in *Gaia* EDR3 (Gaia Collaboration et al. 2021). *Gaia* EDR3 is the first delivery of the third data release, and among other products it provides about 1.8 billion sources with astrometric and photometric observations, that is  $(l, b, \varpi, \mu_{\alpha^*}, \mu_{\delta}, G, G_{BP}, G_{RP})$ , with an improved precision with respect to *Gaia* DR2 due to an increase in the observational time baseline, now spanning a period of 34 months. Thanks to this improvement, we are able to search for new OCs with a deeper magnitude cut, which allows us to reach farther and less populated groupings. The adopted magnitude limit is  $G = 18$  mag, unlike our previous studies where the limit was  $G = 17$ . Additionally, and since OCs are usually found in the Galactic disc, we limited our search to Galactic latitudes within  $|b| \leq 20^\circ$ , where most of the OCs are expected to be found. Similarly to our previous searches, we rejected stars with parallaxes larger than 7 mas to avoid very close OCs, which will suffer from strong projection effects and will not be detectable by our method, and sources with negative parallaxes. We also filtered out stars with proper motions higher than  $|\mu_{\alpha^*}|$  and  $|\mu_{\delta}| \geq 30$

mas  $\text{yr}^{-1}$  to remove stars incompatible with disc rotation, which OCs are expected to follow.

The median parallax uncertainty in *Gaia* EDR3 at  $G = 18$  mag is 0.12 mas, while for proper motions the median uncertainties in  $\mu_{\alpha^*}$  and  $\mu_{\delta}$  at  $G = 18$  mag are of 0.123 and 0.111 mas  $\text{yr}^{-1}$ , respectively (Lindegren et al. 2021). These are similar uncertainty levels in comparison to that of *Gaia* DR2 at magnitude  $G = 17$  mag (Lindegren et al. 2018). For the photometry, the uncertainties in *Gaia* EDR3 at  $G = 18$  mag are at the level of a thousandth for  $G$ , and a hundredth of a magnitude for  $G_{BP}$  and  $G_{RP}$  (Riello et al. 2021). These magnitude uncertainty levels are also comparable to *Gaia* DR2 at  $G = 17$  mag; therefore we consider them to be a reasonable compromise to successfully achieve our goals. Altogether, and taking the aforementioned filters into account, the sample to be analysed contains 232 463 114 sources, and its analysis is enabled thanks to the use of a big data environment in our data analysis pipeline (Castro-Ginard et al. 2020).

## 3. The OCfinder method

The methodology developed to search for new OCs in *Gaia* data, OCfinder, is described in detail in Paper I. It was successfully applied to detect 23 new nearby OCs (Castro-Ginard et al. 2018) in the TGAS data set of *Gaia* DR1. It was also applied to *Gaia* DR2 where 53 new OCs were detected in a direction near the Galactic anticentre (Castro-Ginard et al. 2019) and hundreds of new OCs in a big data search on the whole Galactic disc (Castro-Ginard et al. 2020).

The OCfinder method consists of two main steps. The first step is a blind search for overdensities in the five-dimensional astrometric space of *Gaia*, that is  $(l, b, \varpi, \mu_{\alpha^*}, \mu_{\delta})$ , so as to find sets of stars which are more clustered than the average field stars for that region (Sect. 3.2). The second step makes use of the *Gaia* photometry to confirm whether OC member stars follow an isochrone pattern in a colour-magnitude diagram (CMD) using an artificial neural network trained on well-characterised CMDs (Sect. 3.3).

### 3.1. Data preparation

We divided the sky into small areas of size  $L \times L \text{ deg}^2$ , where  $L$  varies according to the local structure density. This was done in order to define local average densities, accounting for the varying densities of the Galactic disc, when searching for representative overdensities which may correspond to physical OCs. In our methodology, the sizes of these regions are not defined by following computational, but physical arguments. We used the *Gaia* Universe Model Snapshot<sup>1</sup> (GUMS, Robin et al. 2012) to represent the field star population, together with realistic OCs simulated using the *Gaia* Object Generator (GOG, Luri, X. et al. 2014) both including errors at the time of *Gaia* EDR3<sup>2</sup>, to find the size  $L$  of the regions that detect most of the simulated OCs (see Sect. 3 in Paper I for details). In this case, the sizes of the regions ranges from  $L = 10^\circ$  to  $L = 15^\circ$ , which corresponds to a maximum of about  $10^7$  stars per box to be simultaneously analysed with the clustering algorithm. The simultaneous analysis of such a large number of stars is not a problem for our method

<sup>1</sup> GUMS (true values of intrinsic simulated sources) and GOG (observed attributes with simulated observational uncertainties) can be found in the *Gaia* archive: <https://gea.esac.esa.int/archive/>

<sup>2</sup> Computed with the prescription given in <https://github.com/agabrown/PyGaia>

due to the inclusion of a big data environment (see details in Sect 3.2).

Once the stars are divided into the  $L \times L$  deg<sup>2</sup> regions, the five astrometric dimensions ( $l, b, \varpi, \mu_{\alpha^*}, \mu_{\delta}$ ) are standardised in order to balance their importance in the clustering algorithm. In our case, we used the StandardScaler method implemented in the scikit-learn Python library (Pedregosa et al. 2011), which transforms each dimension to have zero mean and unit variance.

### 3.2. Astrometric clustering with DBSCAN

In each of the  $L \times L$  deg<sup>2</sup> regions, we ran the density-based clustering algorithm DBSCAN (Ester et al. 1996) to find statistical overdensities that may belong to real OCs. DBSCAN relies on two input parameters, which are  $\epsilon$  and  $minPts$ , to define a density threshold and it searches for overdensities above the threshold. As a brief description, DBSCAN visits each source in the dataset, builds an  $N$ -dimensional  $\epsilon$  neighbourhood around the source, and counts how many sources are within the  $\epsilon$  neighbourhood. If at least  $minPts$  sources are found, they are considered to be a cluster (we refer readers to Sect. 2 in Paper I for a DBSCAN description relevant for this application).

Similarly to our previous applications of OCfinder, we selected several optimal values of  $minPts$  which were found, together with  $L$ , using simulated data. In this case, the values of  $minPts$  range from eight to 16 stars. The computation of the  $\epsilon$  parameter was done automatically in each  $L \times L$  deg<sup>2</sup> region taking advantage of the fact that OC member stars are closer than random field stars in multidimensional space including positions, parallax, and proper motions. In brief, we computed the distribution of distances for each star to its  $k_{th}$  nearest neighbour (defined as  $k = minPts - 1$ ), and we compared it to the distribution of  $k_{th}$  nearest neighbour distances between field stars (with no substructure present). Then,  $\epsilon$  is defined as the minimum  $k_{th}$  distance between field stars, below where the distribution starts to differ from the real distribution (due to the presence of clusters in the latter). Again, we refer the reader to Sect. 2.2 in Paper I for exact details on the computation of  $\epsilon$ .

After DBSCAN was applied to a given  $L \times L$  deg<sup>2</sup> region, we shifted these regions by  $L/3$  and  $2L/3$  and applied DBSCAN again in the new region in order to account for clusters in the borders of the grid. Then, we merge our duplicated or overlapping groupings that are in fact a single cluster. The whole process was applied using several values for the pairs  $(L, minPts)$ . This way, a Monte Carlo-like analysis of the results was enabled, and clusters with more findings within the different pairs of  $(L, minPts)$  are the most reliable.

The choice of DBSCAN is due to i) its ability to work with  $N$ -dimensional data, ii) the fact that it can handle noise (sources not assigned to any cluster), iii) being density-based, it can account for projections effects and clusters not having a predetermined shape, and iv) the fact that it does not require an a priori number of clusters to be found. The caveat of DBSCAN is that it is limited to a single density threshold (defined by  $\epsilon$  and  $minPts$ ), and it only finds overdensities above that limit. This has been improved with HDBSCAN (Hierarchical-DBSCAN, Campello et al. 2013), with which a whole range of  $\epsilon$  values is explored, therefore allowing there to be clusters with different density thresholds. In this case, however, the number of clusters found drastically increases, thus increasing the number of false positives and the complexity in the interpretability of the results. We consider that in our approach, using DBSCAN with several pairs of  $(L, minPts)$  which are found to be efficient in detect-

ing a large number of clusters using realistic simulated data, we cover the range of densities which may define OCs. A comparison between different clustering algorithms, including HDBSCAN, DBSCAN, and our specific approach in OCfinder, was carried out by Hunt & Reffert (2021), who found OCfinder to be the best among the explored options in terms of balance in sensitivity, specificity, and precision.

The whole clustering process is deployed at the MareNostrum supercomputer, located at the Barcelona Supercomputing Center<sup>3</sup>. Each DBSCAN run for each  $(L, minPts)$  pair was launched distributed in three nodes of MareNostrum (with a total of 144 cores and 48 cores for each node). To distribute the process, we used PyCOMPSSs (Tejedor et al. 2017), a Python-based application that distributes and schedules the execution of a job transparently to the user. Execution times for each DBSCAN application on the whole Galactic disc range from 12 to 27 hours depending on the  $(L, minPts)$  pairs, with higher values for both  $L$  and  $minPts$  being more computationally expensive due to the larger amount of sources to analyse. The advantage of using PyCOMPSSs, as well as the big data environment of MareNostrum, can be seen in Álvarez Cid-Fuentes et al. (2019), where the authors compare the performance of the clustering process of OCfinder in different environments.

### 3.3. Photometric confirmation with deep learning

The second step of OCfinder is the recognition of physical OCs among the statistical clusters found by DBSCAN. We built CMDs from the members of each statistical cluster using *Gaia*'s  $G$ ,  $G_{BP}$ , and  $G_{RP}$  photometry, and we used a deep artificial neural network (ANN, Hinton 1989) trained on well-characterised CMDs to distinguish real OCs by detecting their characteristic isochrone patterns. In our first applications of OCfinder, in Paper I and Paper II, we used a multi-layer perceptron with a single hidden layer to perform the classification. The big data search in the whole Galactic disc in Paper III resulted in a larger amount of statistical clusters found (with respect to Paper I and Paper II). There, we used a more robust classification through a deep ANN architecture, which outperformed the simpler ANN.

The deep ANN consists on an initial set of convolutional layers that extract the meaningful characteristics of the CMDs, followed by a set of fully connected layers to perform the classification. We used the PyTorch<sup>4</sup> package (Paszke et al. 2017), which provides powerful software particularly well suited for deep learning, to implement our deep ANN. We used the CUDA environment (NVIDIA et al. 2020) to implement and train the deep ANN on a NVidia RTX 2080Ti GPU, which provides fast and agile computations that allowed us to test different ANN architectures until reaching the optimal configuration.

The key ingredient for a good classification result is to ensure a high-quality training set. We used the largest homogeneous sample of known OCs present in Cantat-Gaudin et al. (2020) to represent positive isochrone identifications. From this list, we removed clusters with very few stars (at least  $minPts$ ) up to magnitude  $G = 18$  mag, clusters with diffuse isochrones, and highly contaminated cases. We used data augmentation techniques to increase the number of training examples, meaning that for each OC we built new CMDs from a subpopulation of the original OC members. We supplemented this training set with simulated clusters which were generated using synthetic isochrones assum-

<sup>3</sup> <https://www.bsc.es/marenostrum>

<sup>4</sup> <https://pytorch.org/>

ing solar metallicity ( $Z \approx 0.0152$  dex) from the PARSEC code<sup>5</sup> (Bressan et al. 2012), with ages ranging from 4 Myr to 13 Gyr, approximately. For each synthetic population, we built different sub-samples using the same data augmentation techniques as in the case of the real OCs, which were placed at different distances ranging from 300 pc to 4 kpc, as well as different values for extinction  $A_V$ , ranging from 0 to 2 mag, in order to represent all the possible configurations in the CMD. In order to mimic *Gaia* EDR3 photometric observations, we added photometric errors in each band (described in Appendix A).

On the negative identification side (identification of no clusters), we used random field stars queried from the *Gaia* EDR3 archive at locations that avoid known open clusters. We also applied our DBSCAN approach to the GUMS and used the resulting statistical clusters, which do not represent real objects since stellar groupings of such OCs are not present in GUMS, to increase the negative identification training set. Similarly to our previous studies, in order to feed the network and perform the classification, we converted each CMD to a 2D histogram to extract their pixels, which were then normalised. We used a logarithmic normalisation scheme in order to enhance the lower density regions which are key for the characterisation of contaminants in the CMD.

## 4. Results

### 4.1. Crossmatch to known cluster catalogues

The OCs found with our OCfinder methodology were cross-matched to known OC catalogues to identify which OCs are already known and which are new findings. The source of most known identifications is the catalogue provided by Cantat-Gaudin et al. (2020), representing the largest homogeneous catalogue including 2017 OCs with information about their mean astrometric parameters, as well as estimated values for age, distance, and line-of-sight extinction. We consider our OCs to match with a cluster in Cantat-Gaudin et al. (2020) if their centres are within a circle of radius  $0.5^\circ$  in  $l$  and  $b$  coordinates, and if their mean parallaxes and proper motions are compatible within  $2\sigma$  (where  $\sigma$  is the quadratic sum of the uncertainties quoted in both catalogues for each quantity). In the first step of OCfinder, the application of DBSCAN was able to find 1 559 clusters from Cantat-Gaudin et al. (2020) which represents nearly 80% of the catalogue. In the second OCfinder step, the photometric confirmation, the ANN validated 1 515 of the crossmatched OCs previously found with DBSCAN, which shows the high efficiency of the ANN in identifying OC CMDs against random statistical overdensities. This re-detection efficiency is similar to our previous work in Paper III using *Gaia* DR2, and it is mostly due to the selection of hyper-parameters for DBSCAN that are optimised for an all-sky search (see Sect. 3.2).

Recently, Dias et al. (2021) have provided fundamental parameters for 1 743 OCs in our Galaxy based on *Gaia* DR2. The vast majority of these clusters are also included in Cantat-Gaudin et al. (2020), and therefore they have already been crossmatched. However, there are clusters which could not be characterised by Cantat-Gaudin et al. (2020) or new clusters which were detected afterwards (e.g. Monteiro et al. 2020; Ferreira et al. 2020). From them, we were able to re-detect 110 in our clustering step of which 104 were confirmed in the photometric validation using the CMD. Dias et al. (2021) also provide a list of dubious and likely not real OCs. We could not detect any of the OCs

listed as they are not likely real, thus confirming the results by Dias et al. (2021) for these cases. On the other hand, we were able to re-detect eight out of 11 UBC clusters detected in Paper III which are listed as dubious; these include UBC 359, UBC 416, UBC 505, UBC 573, UBC 575, UBC 577, UBC 579, and UBC 593. We were not able to re-detect any of the Liu & Pang (2019) candidates listed as dubious.

In the aforementioned catalogues, there is a large contribution of UBC clusters detected in Paper I, Paper II, and Paper III. From the  $\sim 650$  UBC clusters, we were able to re-detect 514 ( $\approx 80\%$ ) of them with the *Gaia* EDR3 data, which is a similar re-detection efficiency as in the general case. For the different releases of UBC clusters, our re-detection efficiency is  $\approx 30\%$  for Paper I,  $\approx 60\%$  for Paper II, and  $\geq 80\%$  for Paper III. The differences in the detection efficiency for these UBC clusters are due to the different star densities, and nature, of the datasets analysed, captured in the hyper-parameters used for the search. In Paper I and Paper II, the OC search was performed on the TGAS subset of *Gaia* DR1 (where the limiting magnitude is  $G = 12$  mag) and a low density region localised near the Galactic anti-centre in *Gaia* DR2, respectively. In those searches, the hyper-parameters for the DBSCAN were adapted to the corresponding regions, and they are different from the all-sky search performed in this work. For OCs in Paper III, the detection efficiency is slightly higher because of the similarity between both datasets (see Sect. 2), and thus this is also the case in the method hyper-parameters.

Hunt & Reffert (2021) recently reported the discovery of 41 new OCs using a HDBSCAN clustering algorithm on *Gaia* DR2. Out of these, we were able to re-detect 20 of them in *Gaia* EDR3 with our OCfinder method. From the 21 remaining clusters, nine were detected in the first DBSCAN step, but they were not validated in the second ANN step. Therefore, they need to be further investigated. The reason for the non-detection of the other 12 clusters can be related to the choice of the algorithm, among other causes. In our OCfinder method, we used DBSCAN to detect astrometric overdensities which are limited to a single density threshold, usually limited to the densest cluster in the region (a limitation we minimised by performing a Monte-Carlo-like analysis, see Sect. 3.2). HDBSCAN runs the clustering in a hierarchy of density thresholds, thus it is able to detect varying density clusters on the same region, at the cost of increasing the number of false positives. This is seen in the compactness of the OCs found by Hunt & Reffert (2021), from which the ones we were able to detect have mean dispersions of  $\overline{\sigma_{\overline{\omega}}} = 0.03$  mas and  $\overline{\sigma_{\mu_{\alpha^*}}, \sigma_{\mu_{\delta}}} = 0.11$  mas yr<sup>-1</sup>, and the cluster that were not re-detected have  $\overline{\sigma_{\overline{\omega}}} = 0.05$  mas and  $\overline{\sigma_{\mu_{\alpha^*}}, \sigma_{\mu_{\delta}}} = 0.18$  mas yr<sup>-1</sup>, showing that the clusters we are able to recover are more compact.

Comparisons to pre-*Gaia* cluster catalogues are more difficult because they do not allow for sufficiently good comparisons in proper motion space. These catalogues contained around 3 000 catalogued objects gathered from different data sources (Dias et al. 2002; Kharchenko et al. 2013). Most of their reported clusters identified with our findings were taken into account when crossmatching with the catalogue by Cantat-Gaudin et al. (2018) since they are based on the same data. For the clusters not found within Cantat-Gaudin et al. (2018), we performed a 10 arcmin positional crossmatch based on sky coordinates only. We flagged the coinciding candidates in our main Table 1 (see Sect. 4.2); however, we did not explore further coincidences in any other dimension.

Similarly to the OC catalogues, we crossmatched our findings to globular cluster (GC) catalogues. Recently, Vasiliev &

<sup>5</sup> <http://stev.oapd.inaf.it/cgi-bin/cmd>

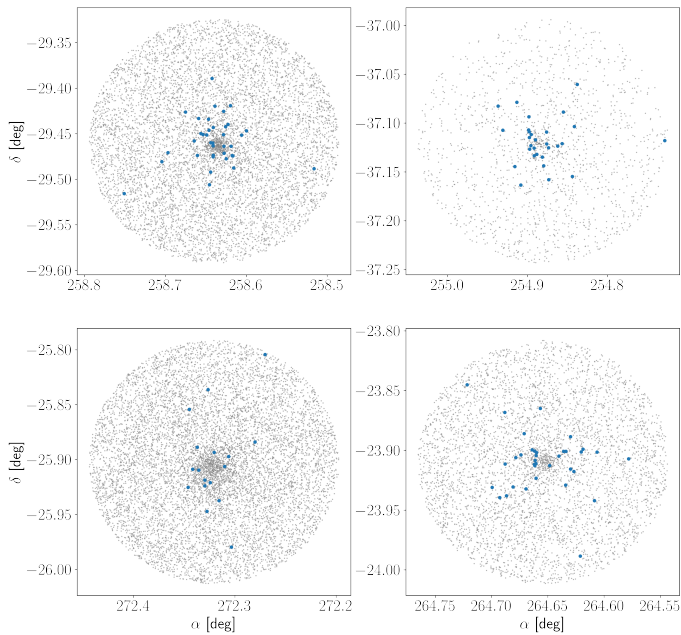


Fig. 1: Distribution in  $\alpha$  and  $\delta$  for GCs NGC 6304 (top left), NGC 6256 (top right), NGC 6553 (bottom left), and NGC 6401 (bottom right). The blue dots represent the stars found as overdensities with OCfinder, while grey dots are all stars in a cone search around the GC centre.

Baumgardt (2021) reported a catalogue of known GCs with mean astrometric parameters computed from *Gaia* EDR3. There are 113 GCs that can be found by OCfinder within our filters. We were able to find 94 of them in the clustering step, which represents 83% of the target catalogue, showing a re-detection efficiency similar to the OC case. Out of these, 84 GCs were validated with the ANN in our photometric step. This shows that the addition of the simulated isochrones in the ANN training (see Sect 3.3), which are the only contribution for clusters with these old ages, improves the validation of cluster CMDs. There are four cases that escaped the crossmatch with a known GC due to differences in parallax and proper motions higher than our threshold. These cases are shown in Fig. 1, where we show all stars (in grey) around the centres of NGC 6304, NGC 6256, NGC 6553, and NGC 6401 as well as the cluster stars we were able to find around them (in blue). In these cases, we were able to find both the main host GC (already crossmatched) and a structure which is more dispersed than the catalogued GC, with differences of  $3\sigma$  either in  $\varpi$ ,  $\mu_{\alpha^*}$ , or  $\mu_{\delta}$ , and with a CMD compatible with being a very old object. Therefore, we consider these objects to be part of the same structure as the host GC, revealing the presence of extensive halos around these distant objects.

#### 4.2. New UBC clusters

After crossmatching our findings with known cluster catalogues, we were able to report 628 new OC candidates, which are numbered from UBC 1001 in order to differentiate from the UBC clusters found in *Gaia* DR2. These candidates were further divided into class A, class B, and class C based on a visual inspection of their distributions in  $(\alpha, \delta, \varpi, \mu_{\alpha^*}, \mu_{\delta})$  and their CMDs, aided with the distribution of radial velocities when available. We classify 566 OC candidates as class A (90% of the total), 26 (4%) as class B, and 36 (6%) as class C. Candidates in class A

usually show a clustered distribution in the five astrometric dimensions and a clear sequence in the CMD. On the other hand, we generally classified candidates into class B if the main sequence formed by the candidate member stars in the CMD is truncated before  $G = 18$  mag, and into class C if they also contain less than 15 members (where we consider the validation to be less reliable due to small numbers). In Fig. 2 we show examples of class A (two first rows), class B (third row), and class C (fourth row) OC candidates, which give a visual idea of the features of OC candidates in each of the classes.

A sample of the list of the new OC candidates, divided in their classes, can be found in Table 1. It contains the mean astrometric parameters ( $\alpha, \delta, l, b, \varpi, \mu_{\alpha^*}, \mu_{\delta}$ ) and their dispersions for each OC candidate together with radial velocities when available. It also contains the apparent angular radius ( $\theta$ ) of the OC candidate, computed as the quadratic sum of  $\sigma_l$  and  $\sigma_b$ . A distance estimation computed from the CMD and mean parallax (see Sect. 4.2.2) is also provided, together with an estimation of age and line-of-sight extinction ( $A_V$ ). Finally, the number of stars considered as members, and member stars with available radial velocities, are also reported. We flagged the candidates that are positionally crossmatched to Kharchenko et al. (2013) (see Sect. 4.1). The full version of Table 1 can be found online at the CDS, together with Table 2 reporting the membership lists that resulted from our OCfinder method.

##### 4.2.1. Characteristics of the new OC candidates

A small subset of the *Gaia* EDR3 stars have radial velocity measurements from *Gaia* DR2. We did not use these radial velocities in the clustering process; however, when available, they are useful to assess the reliability of the classification of the OC candidate. For the OC candidates in class A, 178 of them have radial velocity measurements, of which 72 are based on more than one star, and 25 are based on more than two stars. From these 25 OC candidates with radial velocities averaged over more than two stars, the median value of the radial velocity dispersions is  $2.31 \text{ km}\cdot\text{s}^{-1}$  with a median absolute dispersion of  $2.03 \text{ km}\cdot\text{s}^{-1}$ , and 17 of them have radial velocity dispersions of  $3 \text{ km}\cdot\text{s}^{-1}$  at most. For class B and class C candidates, only four and six clusters have a mean radial velocity available, respectively, with 1 OC with more than two stars with radial velocity measurements in both cases. In these OCs, the radial velocity dispersions are  $13.34$  and  $15.03 \text{ km}\cdot\text{s}^{-1}$ , respectively.

Figure 3 shows the distribution of the new OC candidates in Galactic  $l$  and  $b$  coordinates. We see that the location of the new candidates matches that of the known OCs. Even if the search is performed up to  $|b| \leq 20^\circ$ , new OCs are preferentially located within the Galactic disc at low latitudes. The vast majority of the new OC candidates ( $\sim 99\%$ ) are located at  $|b| \leq 10^\circ$  (with only UBC 1186 and UBC 1530 with  $|b| > 10^\circ$ , both belonging to class A), and  $\sim 93\%$  of them within  $|b| \leq 5^\circ$ . We are also able to confirm some structures seen in previous studies, such as the lack of OCs in a region near  $l \sim 140^\circ$  previously dubbed the Gulf of Camelopardalis (Cantat-Gaudin et al. 2019; Castro-Ginard et al. 2019).

The fact that our new detections are generally at larger distances can be seen in Fig. 4, where we show a histogram of the OC mean parallaxes for both the known population (Cantat-Gaudin et al. 2020) and new OC candidates. We see that the parallax distribution of the new OCs is positively skewed with respect to the distribution of known OC, meaning that the mode is slightly towards smaller mean parallaxes. Also, the drop in the distribution towards larger parallaxes is steeper in the new

Table 1: Some examples of the OCs found in this paper.

Name	$\alpha$ [deg]	$\delta$ [deg]	$l$ [deg]	$b$ [deg]	$\theta$ [deg]	$\varpi$ [mas]	$\mu_{\alpha^*}$ [mas yr $^{-1}$ ]	$\mu_{\delta}$ [mas yr $^{-1}$ ]	$V_{\text{rad}}$ [km s $^{-1}$ ]	$N$ ( $N_{V_{\text{rad}}}$ )	log age [dex]	$d$ [pc]	$A_v$ [mag]
Class A													
UBC 1001 <sup>a</sup>	267.36(0.10)	-29.23(0.04)	0.18(0.05)	-0.86(0.08)	0.10	0.37(0.01)	0.42(0.08)	-1.94(0.06)	-	29(0)	7.46	2786	1.91
UBC 1002	268.11(0.05)	-28.20(0.07)	1.40(0.06)	-0.90(0.05)	0.08	0.28(0.03)	0.01(0.06)	-1.96(0.06)	-14.88(-)	71(1)	8.34	3233	2.31
UBC 1003	268.60(0.05)	-24.89(0.05)	4.48(0.05)	0.40(0.05)	0.07	0.39(0.02)	-0.58(0.09)	-3.11(0.06)	-	44(0)	7.78	2809	2.66
UBC 1004	270.56(0.07)	-25.07(0.06)	5.21(0.07)	-1.23(0.06)	0.09	0.38(0.04)	-0.90(0.11)	-2.40(0.10)	-5.24(2.12)	418(2)	8.50	2436	2.32
UBC 1005 <sup>a</sup>	272.18(0.06)	-21.42(0.08)	9.12(0.08)	-0.75(0.06)	0.10	0.66(0.01)	-0.71(0.07)	-1.35(0.06)	-	31(0)	7.37	1542	1.23
UBC 1006	272.45(0.07)	-21.48(0.08)	9.19(0.07)	-0.99(0.07)	0.10	0.37(0.02)	0.07(0.08)	-1.06(0.07)	10.21(0.59)	77(2)	8.02	2987	2.28
UBC 1007	270.80(0.05)	-19.35(0.05)	10.30(0.06)	1.40(0.05)	0.08	0.65(0.02)	0.05(0.07)	-1.08(0.09)	-	32(0)	7.89	1602	1.93
UBC 1008	273.34(0.06)	-19.03(0.08)	11.74(0.08)	-0.54(0.07)	0.10	0.34(0.02)	-0.84(0.06)	-1.82(0.06)	-	29(0)	7.84	3182	1.28
UBC 1009	273.50(0.07)	-19.05(0.07)	11.79(0.08)	-0.69(0.05)	0.09	0.33(0.02)	-0.36(0.09)	-1.19(0.09)	-	46(0)	7.74	2753	1.27
UBC 1010 <sup>a</sup>	274.50(0.06)	-18.54(0.07)	12.69(0.07)	-1.28(0.05)	0.09	0.37(0.02)	-0.01(0.05)	-0.50(0.08)	-	46(0)	8.34	2634	1.08
UBC 1011	274.10(0.06)	-18.11(0.07)	12.89(0.07)	-0.74(0.06)	0.09	0.34(0.04)	-1.11(0.10)	-2.64(0.09)	9.39(8.93)	235(2)	7.73	2804	2.74
UBC 1012 <sup>a</sup>	273.03(0.05)	-15.98(0.04)	14.27(0.05)	1.18(0.04)	0.06	0.39(0.03)	0.36(0.09)	-0.69(0.09)	-	44(0)	8.89	2871	2.39
UBC 1013	273.13(0.11)	-15.33(0.11)	14.89(0.11)	1.41(0.10)	0.15	0.58(0.02)	-0.24(0.07)	-1.76(0.09)	-	50(0)	7.04	1839	1.20
UBC 1014	276.55(0.09)	-15.15(0.08)	16.61(0.09)	-1.42(0.07)	0.12	0.48(0.01)	-0.72(0.06)	-2.73(0.06)	-	28(0)	6.99	2126	1.90
UBC 1015	276.74(0.08)	-14.11(0.08)	17.62(0.08)	-1.09(0.08)	0.11	0.30(0.05)	-0.77(0.10)	-2.12(0.11)	3.74(17.66)	565(4)	8.40	3323	2.61
Class B													
UBC 1567	269.01(0.06)	-29.51(0.04)	0.67(0.03)	-2.25(0.06)	0.06	0.38(0.02)	-0.10(0.09)	-0.47(0.07)	-	44(0)	7.02	2929	1.94
UBC 1568	283.34(0.03)	-9.31(0.03)	24.84(0.03)	-4.66(0.02)	0.04	0.43(0.01)	0.86(0.04)	-0.83(0.05)	-	10(0)	7.54	2798	1.46
UBC 1569	282.29(0.06)	-7.15(0.04)	26.31(0.05)	-2.75(0.06)	0.08	0.37(0.01)	-0.31(0.03)	-1.04(0.03)	-	11(0)	8.24	2554	1.02
UBC 1570	298.12(0.06)	32.54(0.05)	68.42(0.06)	2.74(0.04)	0.07	0.40(0.01)	-2.41(0.04)	-5.39(0.05)	-2.46(-)	25(1)	8.46	2761	1.30
UBC 1571	314.26(0.15)	46.45(0.11)	86.91(0.10)	0.63(0.12)	0.15	0.18(0.04)	-2.95(0.10)	-3.62(0.13)	-45.89(13.34)	233(6)	8.91	5053	2.51
UBC 1572	329.04(0.03)	53.98(0.03)	98.90(0.02)	-0.50(0.02)	0.03	0.45(0.01)	-2.24(0.03)	-2.40(0.05)	-	10(0)	8.03	2320	1.55
UBC 1573 <sup>a</sup>	343.45(0.08)	62.61(0.03)	109.86(0.04)	2.76(0.02)	0.05	1.18(0.02)	-1.32(0.11)	-2.70(0.10)	-	17(0)	7.29	863	4.09
Class C													
UBC 1593	267.03(0.05)	-29.01(0.04)	0.22(0.02)	-0.50(0.05)	0.06	0.58(0.01)	0.57(0.07)	-1.43(0.07)	-	15(0)	8.04	1829	1.65
UBC 1594	276.49(0.02)	-13.74(0.03)	17.83(0.03)	-0.71(0.02)	0.03	0.33(0.01)	-0.66(0.04)	-3.65(0.06)	-	13(0)	8.21	3270	2.80
UBC 1595	289.95(0.07)	-1.87(0.04)	34.50(0.03)	-7.16(0.08)	0.08	1.07(0.02)	-0.78(0.07)	-2.12(0.04)	-	11(0)	8.10	1028	2.13
UBC 1596	285.62(0.04)	11.89(0.04)	44.77(0.05)	2.98(0.03)	0.06	0.25(0.01)	-1.81(0.07)	-5.09(0.05)	18.88(15.03)	13(2)	8.55	4133	2.39
UBC 1597	293.17(0.07)	16.81(0.07)	52.54(0.07)	-1.19(0.06)	0.09	0.34(0.01)	-2.28(0.04)	-5.59(0.05)	-	13(0)	7.12	3225	3.04
UBC 1598	313.90(0.03)	49.11(0.04)	88.78(0.04)	2.54(0.03)	0.04	0.36(0.02)	-2.71(0.04)	-4.54(0.03)	-36.83(-)	10(1)	9.04	3585	1.83

**Notes.** The different panels are for OCs classified in class A, class B, or class C, respectively. The parameters for each OC are the mean astrometric parameters and their standard deviations, computed from the ( $N$ ) member stars found in this work ( $N_{V_{\text{rad}}}$  with radial velocity measurements). The apparent angular radius ( $\theta$ ) and an estimation for age, astro-photometric distance, and extinction are also included. The numeration starts from UBC 1001 to differentiate from UBC clusters found in *Gaia* DR2. The full list can be found online at the CDS. <sup>(a)</sup> For information regarding the positional coincidence with [Kharchenko et al. \(2013\)](#), see Sect. 4.1

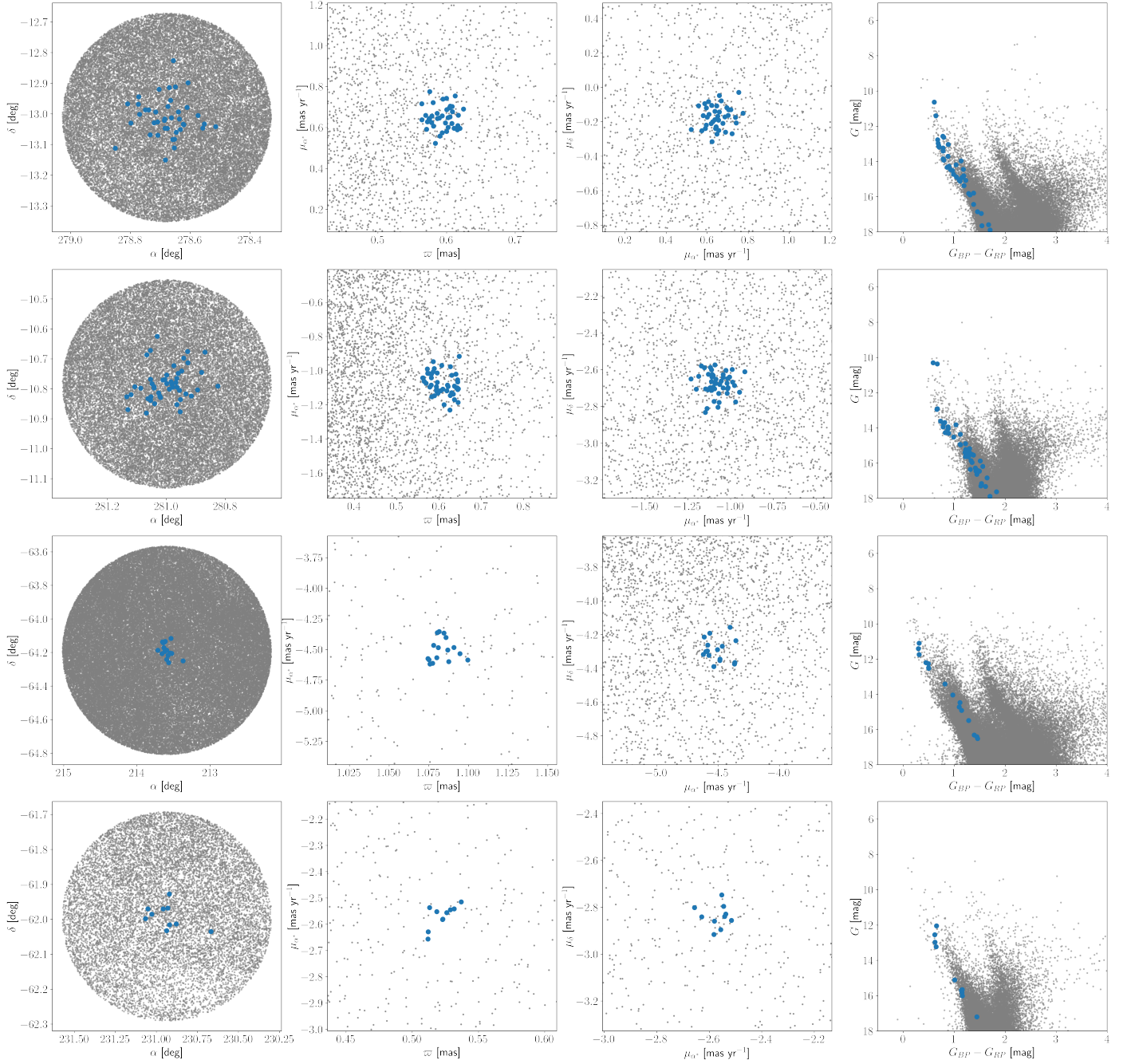


Fig. 2: Examples of the detected OCs for the different classes. The blue dots represent the detected member stars for each OC, while the grey dots are field stars queried in the *Gaia* archive using a cone search within 10 pc radius at the distance of the OC. From left to right, different panels represent : i) positional distribution in  $\alpha$ ,  $\delta$ , ii)  $\varpi$  versus  $\mu_{\alpha^*}$  distribution, iii) proper motion diagram, and iv) the CMD. From top to bottom, different rows are for different OCs: i) UBC 1023 and UBC 1029 from class A, ii) UBC 1592 from class B, and iii) UBC 1628 from class C.

OCs' distribution, showing the increasing difficulty in finding new nearby ones. In fact, only four OC candidates (0.60%) are closer than 1 kpc (see Sect. 4.2.4), and 75 (11.3%) are located within 1 and 2 kpc, probably due to a better completeness of previous surveys in these regions among other methodological effects. The relative parallax errors ( $\sigma_{\varpi}/\varpi$ ) for our new OCs range from 0.003 to 0.05 in the case of class A OCs and from 0.002 to 0.02 for class B and class C.

The heliocentric distances for the new OC candidates range from 860 pc to 9.6 kpc, computed from the distance modulus

(see Sect. 4.2.2). In Fig. 5 we show the distribution in the  $X_{\odot}$  and  $Y_{\odot}$  coordinates, where we see that very few new OC candidates are detected within 1.5 kpc. This may be due to the combination of the following: i) the approach adopted in OCfinder, where we are limited to the most compact object in the search region (see Sect. 3.2), and those are more likely to be already known at these close distances; and ii) we expanded the search to  $G = 18$  mag, which naturally pushes the search to farther distances (see Anders et al. 2021, to see the performance of OCfinder in terms of



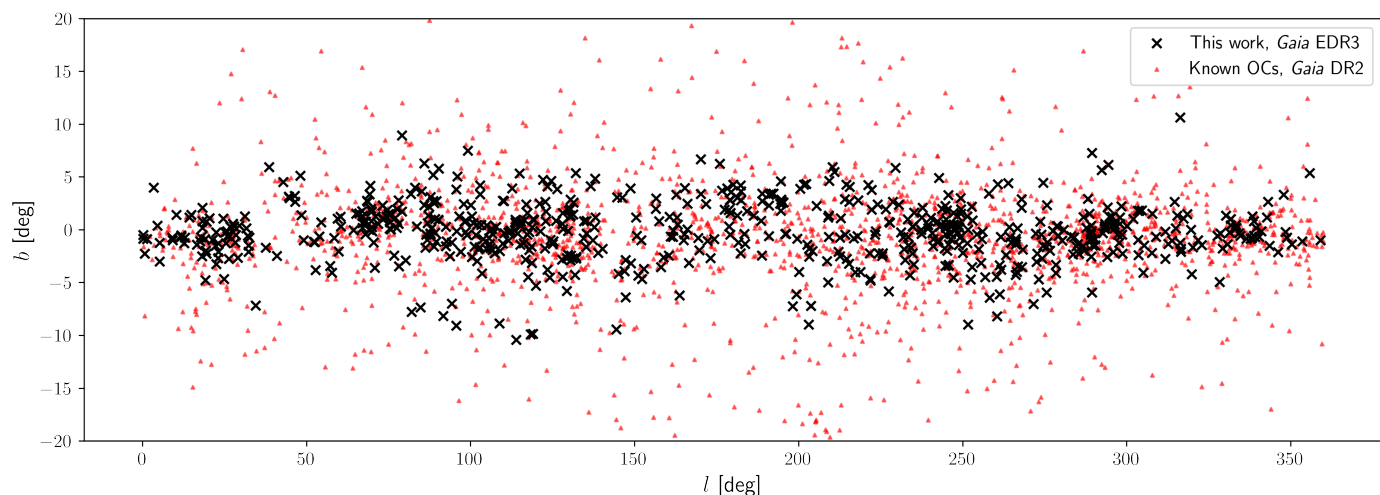


Fig. 3: Distribution of the OC population in  $l, b$  Galactic coordinates. Red triangles represent the OCs known prior to this study, reported in Cantat-Gaudin et al. (2020). Black crosses represent the new OCs found in this work using *Gaia* EDR3.

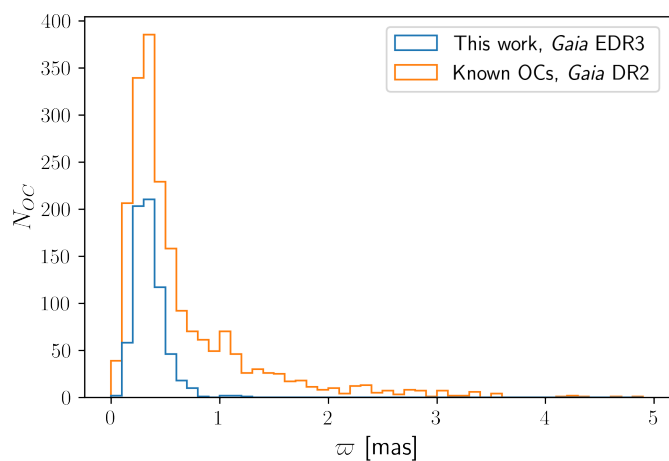


Fig. 4: Histogram of parallaxes for the OC population. The orange line shows the known population in Cantat-Gaudin et al. (2020), while the blue line shows the new findings in this study.

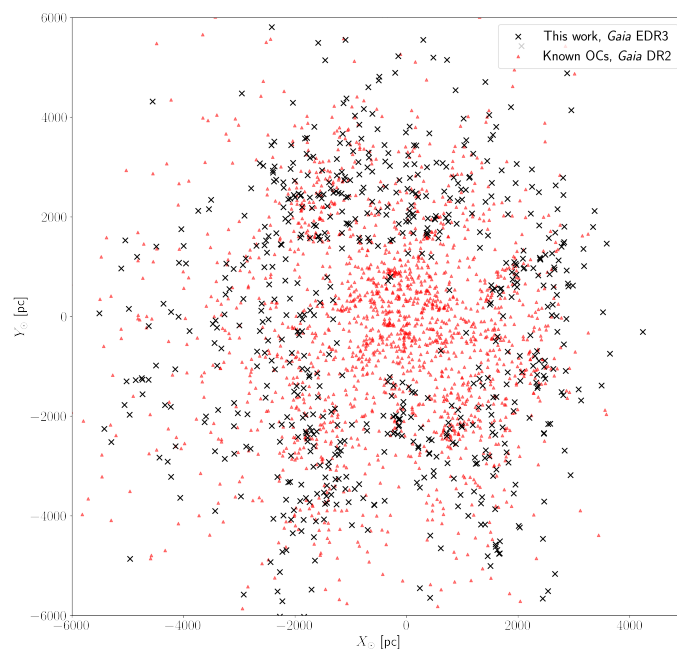


Fig. 5: Distribution in  $X_{\odot}$  and  $Y_{\odot}$  heliocentric coordinates. Symbols are the same as in Fig. 3.

completeness for nearby objects), together with the above consideration of a better completeness of the nearby population.

The effect of the improvements in the *Gaia* EDR3 data is also seen in Fig. 6. There, we show a contour plot of the total proper motion dispersion as a function of the mean parallax, mimicking Fig. 1 in Cantat-Gaudin & Anders (2020). The densest part of the distribution, where most of the clusters are, is moved towards smaller  $\sigma_{\mu}$  showing the huge improvement in the proper motion determinations (the effect is smaller in the parallax). This results in OCs being more compact (in proper motion and parallax), and thus it is easier to detect them as overdensities with respect to *Gaia* DR2.

#### 4.2.2. Ages, distances, and line-of-sight extinctions

Cantat-Gaudin et al. (2020) trained an ANN on a set of well-characterised OCs with reliable estimations for ages, distances, and line-of-sight extinctions to estimate these parameters for almost the whole OC population characterised with *Gaia* DR2 data. We fine-tuned this ANN to estimate these astrophysical pa-

rameters for the newly discovered OCs with *Gaia* EDR3 data and we include them in our Table 1. The ANN takes the CMD of the OC member stars into account, together with the mean parallax plus two other quantities derived from the CMD to aid in the estimation (see Sect. 3.1 from Cantat-Gaudin et al. 2020, for details). For each OC, the ANN estimates its age, absorption ( $A_V$ ), and the distance modulus, and from this we were able to estimate the distance. The authors compared the values from a set of reference clusters with their estimated values to account for their uncertainties. They report that the uncertainties on the determination of the  $\log age$  range from 0.15 to 0.25 dex for young OCs ( $\leq 8.5$  dex), and from 0.1 to 0.2 dex for old OCs. In the case of extinction and distance modulus, the reported typical uncertainties range from 0.1 to 0.2 mag for  $A_V$ , and from 0.1 to 0.2 mag in the distance modulus which corresponds to a 5% to

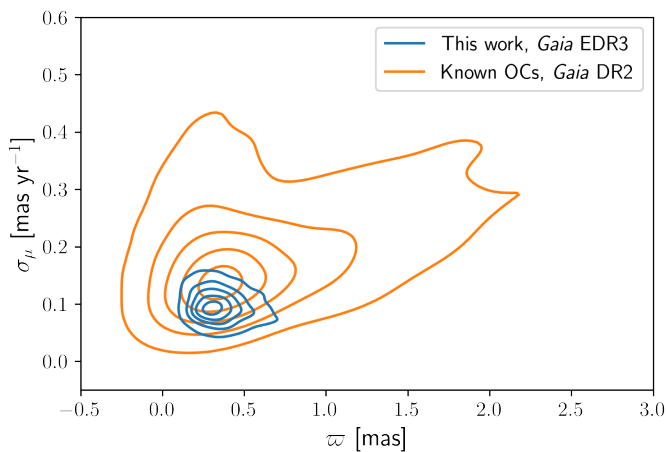


Fig. 6: Total proper motion dispersion ( $\sqrt{\sigma_{\mu_{\alpha^*}}^2 + \sigma_{\mu_{\delta}}^2}$ ) as a function of parallax. The orange lines show the OC density contours for the OC characterised with *Gaia* DR2 (Cantat-Gaudin et al. 2020). The blue lines show the same OC density contours for the OCs detected in this study using *Gaia* EDR3. The density contours are at the 10, 30, 50, 70, and 90% levels.

10% distance uncertainty. For further details, readers can refer to Sect. 3.4 from Cantat-Gaudin et al. (2020).

In Fig. 7 we plotted the distribution of the new OCs (crosses), together with the previously known population (triangles), in the Galactic disc for different age intervals: i) younger than 100 Myr counting with 276 and 703 new and previously known OCs, respectively; ii) from 100 to 500 Myr, with 248 new OCs and 675 previously known OCs; iii) 500 Myr to 1 Gyr, 58 new OCs and 229 previously known OCs; and iv) older than 1 Gyr, with 46 new OCs and 260 previously known OCs. These OCs are also colour-coded by their age. In the younger age bin ( $\leq 100$  Myr), we find clear overdensities of OCs, which are following the different spiral arms (as fitted by Castro-Ginard et al. 2021). In the following age intervals, the distribution of the OCs is more dispersed, not showing significant overdensities, as expected.

The consistency of the age estimations with the previously known OC population is also shown in Fig. 8, where we show the distribution of the new OCs in the galactocentric radius ( $R_{GC}$ ) and altitude above the Galactic plane ( $Z$ ) coordinates. We did not find young open clusters at high  $|Z|$  in the inner disc, but old OCs. The black circles highlight the OCs identified as old and with high  $|Z|$ , and specific plots for those are shown in show Fig. 9. These are interesting OCs, all of them are older than 1 Gyr and up to 4 Gyr (which is the oldest OC in our findings), and follow-up studies will be needed to explore their nature. At large  $R_{GC}$ , we are also able to see the flare of the Galactic disc already seen with the OC population after *Gaia* DR2 Cantat-Gaudin et al. (2020). The extinction values range from 0.14 to 4.65 mag, with a distribution that is shown in Fig. 10.

#### 4.2.3. Comments on UBC 1061

In Paper III, we reported the discovery of UBC 274, an old OC with an extended profile due to disruption by tidal forces. Here, we were also able to detect stellar groups that are probably undergoing disruption processes. This is the case for UBC 1061, which is a new OC found at  $l = 52.75^\circ$  and  $b = -3.82^\circ$ . It is located at a galactocentric distance of  $R_{GC} = 6.77$  kpc, associ-

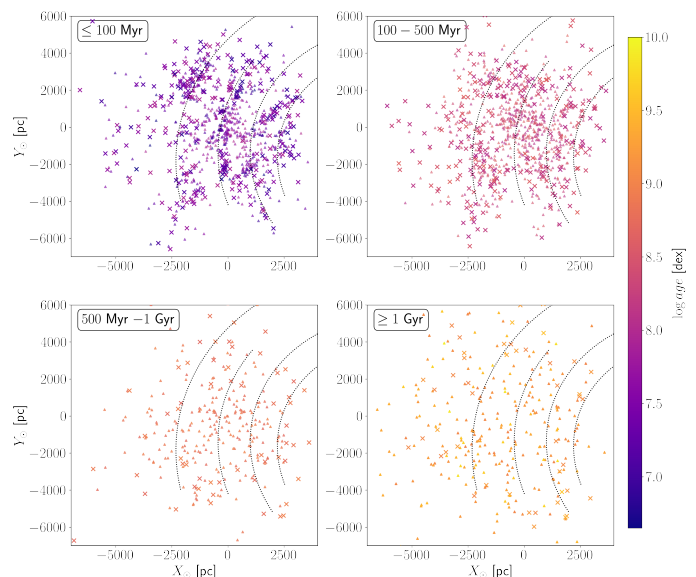


Fig. 7: Distribution of the new (crosses) and known (triangles, Cantat-Gaudin et al. 2020) OCs in the  $X_{\odot}$  versus  $Y_{\odot}$  coordinates for different age bins: 0 – 100 Myr (top left), 100 – 500 Myr (top right), 500 Myr – 1 Gyr (bottom left), and more than 1 Gyr (bottom right). The dotted lines show the spiral arms as described by Castro-Ginard et al. (2021).

ated with the Sagittarius arm (Reid et al. 2014; Castro-Ginard et al. 2021), and at  $Z = -247.95$  pc. We estimate the age of UBC 1061 to be 1.3 Gyr, with an extinction value of  $A_V = 0.74$  mag. Figure 11 shows the distribution of its member stars in five astrometrical dimensions, in addition to its CMD where two blue straggler star candidates can be seen. While in parallax and proper motion UBC 1061 shows a clustered structure, it presents an elongation in the sky position diagram along the  $l$  coordinate, probably because it is undergoing disruption processes by the tidal forces (see Röser et al. 2019; Piatti 2020, for more examples). Such an old OC is rare in current OC catalogues, fewer than 20% of the reported OCs are older than 1 Gyr, and this is particularly the case for an inner disc cluster where young OCs are more frequent (Cantat-Gaudin et al. 2020). None of the 79 identified member stars of UBC 1061 have radial velocities in *Gaia* EDR3, and its red clump stars (at  $G \sim 14$  mag) are at the faint limit for the  $G_{RVs}$  sample. Therefore, UBC 1061 is an interesting case to follow up with on-ground spectroscopic surveys.

#### 4.2.4. Candidates within 1 kpc

We found four OC candidates within 1 kpc from the Sun, where the OC census was assumed to be complete before *Gaia*. These OC candidates are UBC 1187, UBC 1573, and UBC 1592, which are located at 861, 862, and 977 pc, respectively, and all of them have angular sizes  $\leq 0.10^\circ$ . They are young objects, with none of them showing red clump stars in their CMDs, and they are poorly populated, with  $\sim 15$  stars per cluster. This confirms the assumption that our methodology OCfinder is better suited to find objects with small angular sizes, and thus farther objects. However, since nearby objects are still detected, this opens the possibility of performing dedicated searches in the Solar neighbourhood which will be enhanced with the future *Gaia* DR3 thanks to the  $\sim 33 \cdot 10^6$  stars with radial velocity measurements<sup>6</sup>.

<sup>6</sup> <https://www.cosmos.esa.int/web/gaia/dr3>

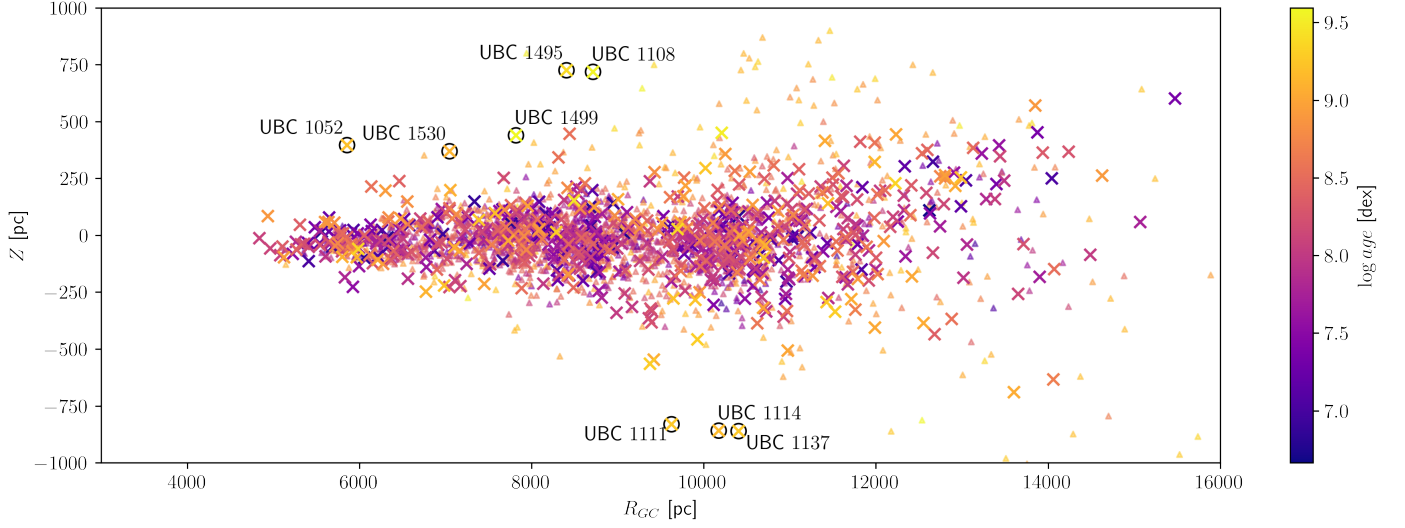


Fig. 8: Distribution of the newly found (crosses) and known (triangles) OCs in the  $R_{GC}$  and  $Z$  coordinates, colour-coded by age. The black circles represent the selected OC shown in Fig. 9.

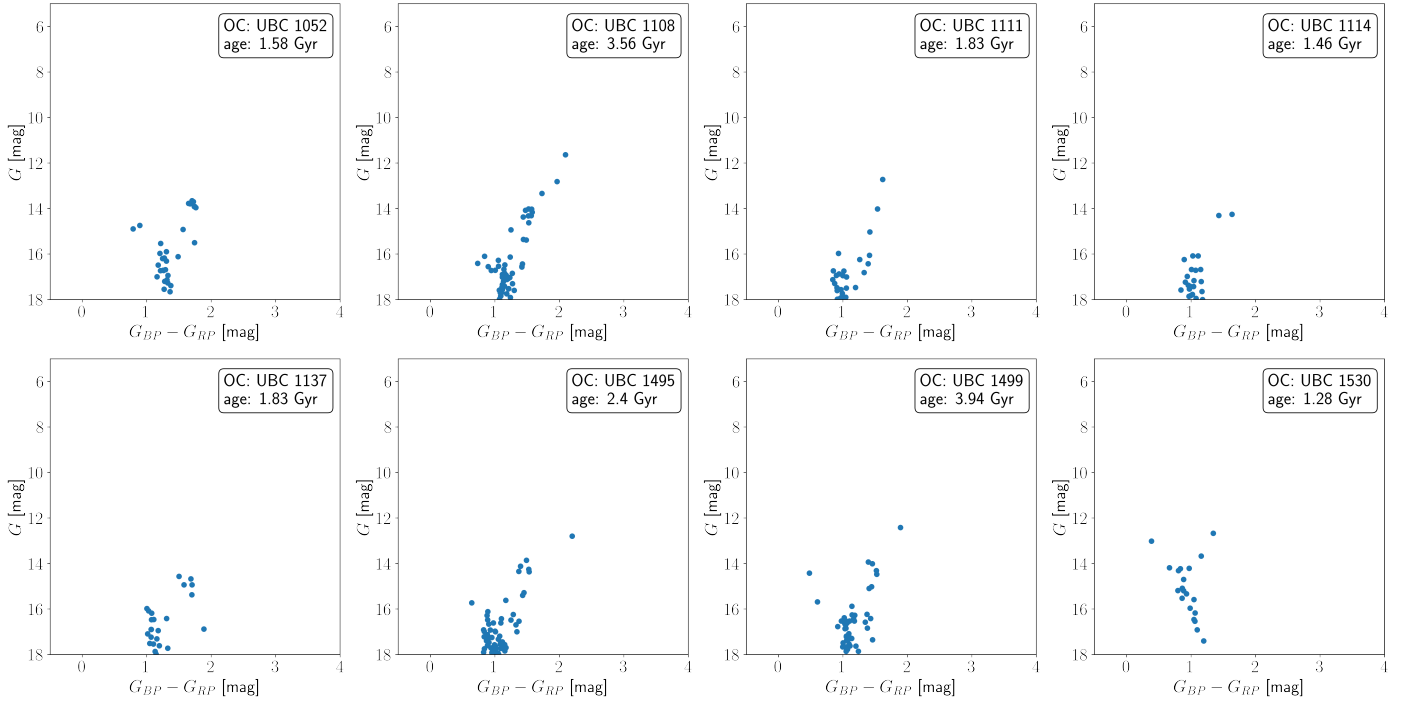


Fig. 9: CMDs of the high  $|Z|$  OCs, selected in Fig. 8.

## 5. Conclusions

We report the detection of 628 new OCs within the Galactic disc, as the result of the adaptation and application of the OCfinder method to *Gaia* EDR3. For all of the new OCs, we report mean astrometric values ( $\alpha$ ,  $\delta$ ,  $l$ ,  $b$ ,  $\varpi$ ,  $\mu_{\alpha^*}$ ,  $\mu_{\delta}$ ) and radial velocity when available. We also estimate their ages, distances, and line-of-sight extinctions using an ANN trained on well-characterised OCs that was successfully applied to the OCs known in *Gaia* DR2. We divided the new OCs into class A, class B, and class C depending on the reliability of the candidate by inspecting the distribution of their member stars in all the available dimensions.

We find that the new OCs are located, in general, at farther distances than the clusters known from *Gaia* DR2. In fact, we

were only able to detect three new objects closer than 1 kpc from the Sun. This is thanks to the improvements in parallax and proper motion precisions of *Gaia* EDR3, with respect to DR2 which allowed us to find more clustered objects (and thus we were able to search farther), and better knowledge of the OC population at close distances.

The estimation of astrophysical parameters also adds reliability to the OCs found with OCfinder. We see that young OCs follow the Galactic spiral arms, and they disperse on the Galactic disc as we explore older ages. Also, we find most OCs to be located at low  $|Z|$  in the inner disc, with the exception of some old ( $> 1$  Gyr) newly found OCs. In the outer disc, we are able to see the flaring of the Galactic disc with young OCs.

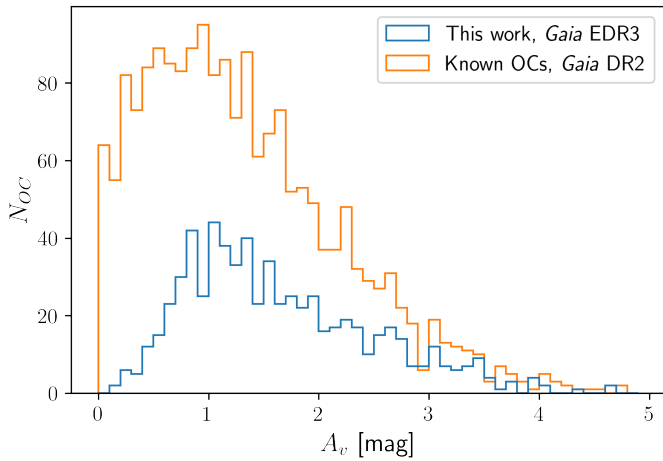


Fig. 10:  $A_V$  histogram for the known (orange) and the new OC (blue) population.

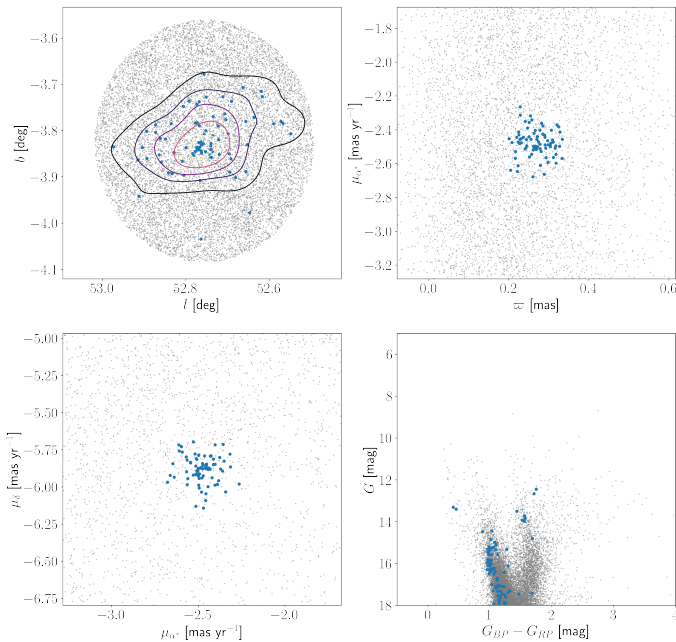


Fig. 11: Diagrams for UBC 1061. The blue dots are the stars selected as members with our OCfinder method, while the grey dots are field stars around 15 pc from the cluster centre. The diagrams correspond to the sky distribution, with density contour plots (top left),  $\varpi$  versus  $\mu_{\alpha^*}$  (top right), proper motion diagram (bottom left), and the CMD (bottom right).

The use of a big data environment in the OCfinder methodology is key for a successful search of OCs. So far, 1 274 OCs have been discovered using the OCfinder method, which represents almost 50% of the currently known OCs population. We have shown that improvements in the OCs census, both in terms of new detections or characterisation of known OCs, need to be aided by machine-learning methods to extract knowledge of the huge volume of high-quality data that is provided by *Gaia* EDR3. Future *Gaia* releases, as well as future photometric and spectroscopic Galactic surveys, will only increase this huge volume of high-quality data and thus the need for machine-learning methods.

**Acknowledgements.** This work has made use of results from the European Space Agency (ESA) space mission *Gaia*, the data from which were processed by the *Gaia* Data Processing and Analysis Consortium (DPAC). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the *Gaia* Multilateral Agreement. The *Gaia* mission website is <http://www.cosmos.esa.int/gaia>. The authors are current or past members of the ESA *Gaia* mission team and of the *Gaia* DPAC. This work was (partially) funded by the Spanish MICIN/AEI/10.13039/501100011033 and by "ERDF A way of making Europe" by the "European Union" through grant RTI2018-095076-B-C21, and the Institute of Cosmos Sciences University of Barcelona (ICCUB, Unidad de Excelencia "María de Maeztu") through grant CEX2019-000918-M. ACG acknowledges Spanish Ministry FPI fellowship n. BES-2016-078499. This work has been partially supported by the Spanish Government (PID2019-107255GB), by Generalitat de Catalunya (contract 2014-SGR-1051). This research has made use of the VizieR catalogue access tool, CDS, Strasbourg, France. The original description of the VizieR service was published in A&AS 143, 23. This research has made extensive use of the TOPCAT software (Taylor 2005).

## References

- Álvarez Cid-Fuentes, J., Solà, S., Álvarez, P., Castro-Ginard, A., & Badia, R. 2019, in Proceedings of the 15th International Conference of eScience, 96–105
- Anders, F., Cantat-Gaudin, T., Quadrino-Lodoso, I., et al. 2021, A&A, 645, L2
- Bossini, D., Vallenari, A., Bragaglia, A., et al. 2019, A&A, 623, A108
- Bressan, A., Marigo, P., Girardi, L., et al. 2012, MNRAS, 427, 127
- Campello, R. J. G. B., Moulavi, D., & Sander, J. 2013, in Advances in Knowledge Discovery and Data Mining, ed. J. Pei, V. S. Tseng, L. Cao, H. Motoda, & G. Xu (Berlin, Heidelberg: Springer Berlin Heidelberg), 160–172
- Cantat-Gaudin, T. & Anders, F. 2020, A&A, 633, A99
- Cantat-Gaudin, T., Anders, F., Castro-Ginard, A., et al. 2020, A&A, 640, A1
- Cantat-Gaudin, T., Jordi, C., Vallenari, A., et al. 2018, A&A, 618, A93
- Cantat-Gaudin, T., Krone-Martins, A., Sedaghat, N., et al. 2019, A&A, 624, A126
- Carrera, R., Bragaglia, A., Cantat-Gaudin, T., et al. 2019, A&A, 623, A80
- Carrera, R., Casamiquela, L., Carballo-Hijarrubia, J., et al. 2022, A&A, 658, A14
- Casamiquela, L., Castro-Ginard, A., Anders, F., & Soubiran, C. 2021a, A&A, 654, A151
- Casamiquela, L., Soubiran, C., Jofré, P., et al. 2021b, A&A, 652, A25
- Castro-Ginard, A., Jordi, C., Luri, X., et al. 2020, A&A, 635, A45
- Castro-Ginard, A., Jordi, C., Luri, X., Cantat-Gaudin, T., & Balaguer-Núñez, L. 2019, A&A, 627, A35
- Castro-Ginard, A., Jordi, C., Luri, X., et al. 2018, A&A, 618, A59
- Castro-Ginard, A., McMillan, P. J., Luri, X., et al. 2021, A&A, 652, A162
- Colombo, D., Duarte-Cabral, A., Pettitt, A. R., et al. 2022, A&A, 658, A54
- Dias, W. S., Alessi, B. S., Moitinho, A., & Lépine, J. R. D. 2002, A&A, 389, 871
- Dias, W. S. & Lépine, J. R. D. 2005, ApJ, 629, 825
- Dias, W. S., Monteiro, H., Moitinho, A., et al. 2021, MNRAS, 504, 356
- Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. 1996, in Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD'96 (AAAI Press), 226–231
- Ferreira, F. A., Corradi, W. J. B., Maia, F. F. S., Angelo, M. S., & Santos, J. F. C., J. 2020, MNRAS, 496, 2021
- Gaia Collaboration, Brown, A. G. A., Vallenari, A., et al. 2018, A&A, 616, A1
- Gaia Collaboration, Brown, A. G. A., Vallenari, A., et al. 2021, A&A, 649, A1
- Gaia Collaboration, Brown, A. G. A., Vallenari, A., et al. 2016, A&A, 595, A2
- Hinton, G. 1989, Artificial Intelligence, 40, 185
- Hunt, E. L. & Reffert, S. 2021, A&A, 646, A104
- Jaehnig, K., Bird, J., & Holley-Bockelmann, K. 2021, ApJ, 923, 129
- Kharchenko, N. V., Piskunov, A. E., Schilbach, E., Röser, S., & Scholz, R.-D. 2013, A&A, 558, A53
- Lindgren, L., Hernández, J., Bombrun, A., et al. 2018, A&A, 616, A2
- Lindgren, L., Klioner, S. A., Hernández, J., et al. 2021, A&A, 649, A2
- Liu, L. & Pang, X. 2019, ApJS, 245, 32
- Luri, X., Palmer, M., Arenou, F., et al. 2014, A&A, 566, A119
- Michalik, D., Lindgren, L., & Hobbs, D. 2015, A&A, 574, A115
- Minniti, J. H., Zoccali, M., Rojas-Arriagada, A., et al. 2021, A&A, 654, A138
- Monteiro, H., Barros, D. A., Dias, W. S., & Lépine, J. R. D. 2021, Frontiers in Astronomy and Space Sciences, 8, 62
- Monteiro, H., Dias, W. S., Moitinho, A., et al. 2020, MNRAS, 499, 1874
- NVIDIA, Vingelmann, P., & Fitzek, F. H. 2020, CUDA, release: 10.2.89
- Paszke, A., Gross, S., Chintala, S., et al. 2017, in NIPS-W
- Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, Journal of Machine Learning Research, 12, 2825
- Piatti, A. E. 2020, A&A, 639, A55
- Reid, M. J., Menten, K. M., Brunthaler, A., et al. 2014, ApJ, 783, 130
- Riello, M., De Angeli, F., Evans, D. W., et al. 2021, A&A, 649, A3

- Robin, A. C., Luri, X., Reylé, C., et al. 2012, *A&A*, 543, A100
- Röser, S., Schilbach, E., & Goldman, B. 2019, *A&A*, 621, L2
- Sim, G., Lee, S. H., Ann, H. B., & Kim, S. 2019, *Journal of Korean Astronomical Society*, 52, 145
- Soubiran, C., Cantat-Gaudin, T., Romero-Gómez, M., et al. 2018, *A&A*, 619, A155
- Spina, L., Ting, Y. S., De Silva, G. M., et al. 2021, *MNRAS*, 503, 3279
- Tarricq, Y., Soubiran, C., Casamiquela, L., et al. 2021, *A&A*, 647, A19
- Taylor, M. B. 2005, in *Astronomical Society of the Pacific Conference Series*, Vol. 347, *Astronomical Data Analysis Software and Systems XIV*, ed. P. Shopbell, M. Britton, & R. Ebert, 29
- Tejedor, E., Becerra, Y., Alomar, G., et al. 2017, *The International Journal of High Performance Computing Applications*, 31, 66
- Vasiliev, E. & Baumgardt, H. 2021, *MNRAS*, 505, 5978

## Appendix A: Fitting magnitude uncertainties in *Gaia* EDR3

The *Gaia* EDR3 photometric error model used in this work is based on some relatively simple analytical functions derived fitting magnitude uncertainties for one million random sources in the *Gaia* EDR3 catalogue, covering all magnitude ranges. This error model was used to produce mock *Gaia* CMDs of the synthetic clusters used in the training of the ANN, described in Sect. 3.3.

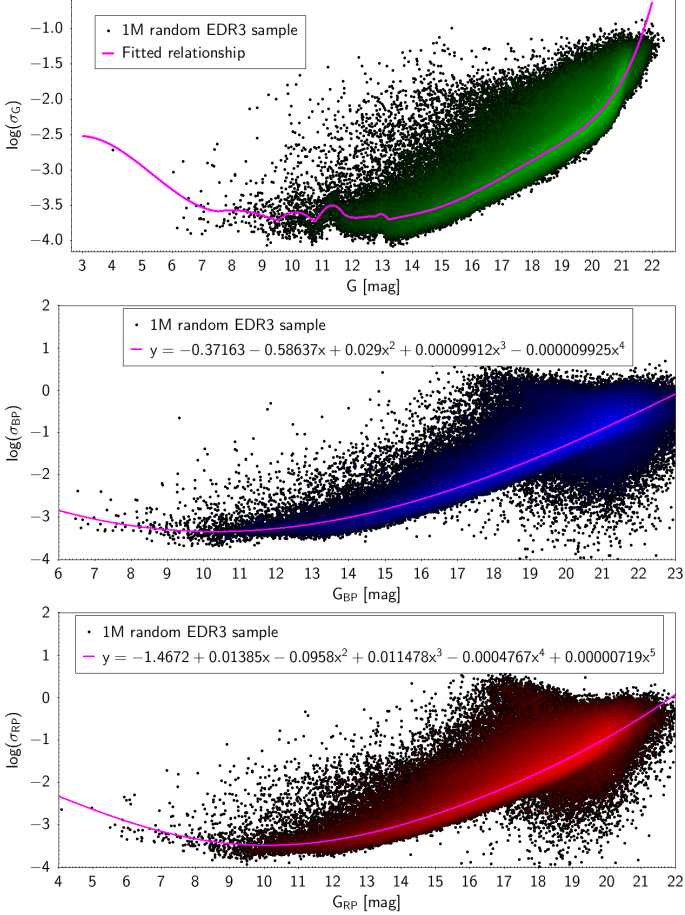


Fig. A.1: *Gaia* EDR3 magnitude uncertainties for  $G$  (top),  $G_{BP}$  (centre), and  $G_{RP}$  (bottom) and the fitted laws included in Tables A.1.

In order to fit the uncertainties, we assumed a polynomial relationship between the logarithm of the *Gaia* EDR3 magnitude uncertainties,  $\log(\sigma_{G_{XP}})$ , and their magnitudes, following the expression

$$\log(\sigma_{G_{XP}}) = \sum A_k \cdot (G_{XP})^k, \quad (\text{A.1})$$

with  $G_{XP}$  being either  $G$ ,  $G_{BP}$ , or  $G_{RP}$  in every case. The resulting coefficients derived are shown in Table A.1.

For the  $G$  case, in some magnitude ranges the uncertainties in *Gaia* EDR3 show some complicated features due to calibration issues (see Riello et al. 2021), which cannot be fitted with a simple polynomial. A Gaussian function (Eq. A.2) was used instead of the polynomial (Eq. A.1) in these ranges:

$$\log(\sigma_G) = a + \exp\left(\frac{-(G-b)^2}{2c^2}\right). \quad (\text{A.2})$$

The actual fit used the following equation obtained taking the natural logarithm of Eq. A.2:

$$\ln[\log(\sigma_G) + 4] = P \cdot G^2 + Q \cdot G + R, \quad (\text{A.3})$$

where we added 4 to  $\log(\sigma_G)$  in order to avoid negative values when deriving its neperian logarithm. In order to retrieve the coefficients in Eq. A.2 from the fitted coefficients (Table A.1), the following transformation can be done from the values of  $P$ ,  $Q$ , and  $R$ :

$$a = e^{(R+\frac{Q}{2})}, \quad (\text{A.4})$$

$$b = \frac{-Q}{2P}, \quad (\text{A.5})$$

$$c = \sqrt{\frac{-1}{2P}}. \quad (\text{A.6})$$

In order to improve the behaviour of the obtained predictions, for all fitted passbands, we restricted our fitting to the median of the observations as a function of the magnitude, and not considering the individual observations of all sources used to derive these medians. Table A.1 shows the  $G$  magnitude intervals and the corresponding coefficients for the different fitted laws. The resulting fitted laws in Table A.1 are plotted in Fig. A.1.

Table A.1: Coefficients of the polynomial (Eq. A.1, first rows in the table with parameters  $A_k$ ) and Gaussian (Eq. A.2, second part of the table with parameters  $P$ ,  $Q$ , and  $R$ ) equations for each magnitude range.

Fitted uncertainty	Magnitude range	$A_0$	$A_1$	$A_2$	$A_3$	$A_4$	$A_5$
$\log(\sigma_{BP})$	$6 < G_{BP} < 21$	-0.37163	-0.58637	0.029	0.00009912	-0.000009925	-
$\log(\sigma_{RP})$	$6 < G_{RP} < 21$	-1.4672	0.01385	-0.0958	0.011478	-0.0004767	0.00000719
Fitted uncertainty	Magnitude range	$A_0$	$A_1$	$A_2$	$A_3$	$A_4$	$A_5$
$\log(\sigma_G)$	$4 \leq G < 7.5313$	-4.31185	1.39006	-0.32577	0.020446	-	-
$\log(\sigma_G)$	$11.7230 \leq G < 12.7819$	467.32190	-114.01847	9.19605	-0.247111	-	-
$\log(\sigma_G)$	$13.2107 \leq G$	-363.52090	113.42543	-14.20340	0.88228	-0.027177	0.00033255
Fitted uncertainty	Magnitude range	$R$	$Q$	$P$			
$\ln[\log(\sigma_G) + 4]$	$7.5313 \leq G < 9.5060$	-12.16197	2.83670	-0.17781	-	-	-
$\ln[\log(\sigma_G) + 4]$	$9.5060 \leq G < 10.7513$	-98.88558	19.38812	-0.95925	-	-	-
$\ln[\log(\sigma_G) + 4]$	$10.7513 \leq G < 11.7230$	-241.49901	42.51819	-1.87684	-	-	-
$\ln[\log(\sigma_G) + 4]$	$12.7819 \leq G < 13.2107$	-571.61430	88.12748	-3.402567	-	-	-