

# Improving the open cluster census.

## II. An all-sky cluster catalogue with *Gaia* DR3<sup>★</sup>

Emily L. Hunt<sup>1,★★</sup> and Sabine Reffert<sup>1</sup>

Landessternwarte, Zentrum für Astronomie der Universität Heidelberg, Königstuhl 12, 69117 Heidelberg, Germany  
e-mail: ehunt@lsw.uni-heidelberg.de

Received 1<sup>st</sup> March 2023; accepted 21<sup>st</sup> March 2023

### ABSTRACT

**Context.** Data from the *Gaia* satellite are revolutionising our understanding of the Milky Way. With every new data release, there is a need to update the census of open clusters.

**Aims.** We aim to conduct a blind, all-sky search for open clusters using 729 million sources from *Gaia* DR3 down to magnitude  $G \sim 20$ , creating a homogeneous catalogue of clusters including many new objects.

**Methods.** We used the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) algorithm to recover clusters. We validated our clusters using a statistical density test and a Bayesian convolutional neural network for colour-magnitude diagram classification. We inferred basic astrometric parameters, ages, extinctions, and distances for the clusters in the catalogue.

**Results.** We recovered 7200 clusters, 2420 of which are candidate new objects and 4780 of which crossmatch to objects in the literature, including 134 globular clusters. A more stringent cut of our catalogue contains 4114 highly reliable clusters, 749 of which are new. Owing to the scope of our methodology, we are able to tentatively suggest that many of the clusters we are unable to detect may not be real, including 1152 clusters from the Milky Way Star Cluster (MWSC) catalogue that should have been detectable in *Gaia* data. Our cluster membership lists include many new members and often include tidal tails. Our catalogue's distribution traces the galactic warp, the spiral arm structure, and the dust distribution of the Milky Way. While much of the content of our catalogue contains bound open and globular clusters, as many as a few thousand of our clusters are more compatible with unbound moving groups, which we will classify in an upcoming work.

**Conclusions.** We have conducted the largest search for open clusters to date, producing a single homogeneous star cluster catalogue which we make available with this paper.

**Key words.** open clusters and associations: general – Methods: data analysis – Catalogs – Astrometry

### 1. Introduction

The Milky Way galaxy is an intricate ecosystem of ongoing star formation, evolution, and destruction. Open clusters (OCs) are one such part of this system, which form when molecular clouds condense into stars and may further condense into gravitationally bound groups of a few dozen to a few thousand stars. Hence, OCs offer an important way to study the immediate aftermath of star formation, as well as the ongoing evolution of stars up to an age of around  $\sim 1$  Gyr, after which most OCs will have been broken up, with their member stars dissolving back into the galactic disk (Portegies Zwart et al. 2010; Krumholz et al. 2019; Krause et al. 2020).

Our view of OCs has always been complicated by their sparsity and their typical location in the galactic disk, making them challenging to isolate from field stars along the line of sight (Cantat-Gaudin 2022). However, dramatically improved astrometric and photometric data from the *Gaia* satellite (Gaia Collaboration et al. 2016) are revolutionising our understanding of OCs and the overall Milky Way. Compared with the *Hipparcos*

mission (Perryman et al. 1997), *Gaia* provides order of magnitude improvements in proper motion and parallax accuracy for around  $10^4$  times as many stars, with over 1 billion sources in total.

Because of these improvements, *Gaia* has enabled many new insights into all properties of OCs. Works such as Meingast et al. (2021) and Tarricq et al. (2022) have shown that many nearby OCs have tidal tails or comas of ejected member stars indicative of their ongoing tidal disruption by the Milky Way. Other works such as Bossini et al. (2019) and Cantat-Gaudin et al. (2020) have used *Gaia* photometry to infer cluster ages, extinctions, and distances, which can then be used to make wider inferences about the Milky Way, such as in Castro-Ginard et al. (2021) who used OCs to trace the spiral arms of the galaxy. Cleaned *Gaia* cluster membership lists also improve spectroscopic studies such as Baratella et al. (2020), who combined *Gaia* data with ground-based spectroscopic measurements to study the chemistry of OCs.

At the heart of all science with OCs, however, is the census of OCs itself. Particularly in the four years since *Gaia* Data Release 2 (DR2, Brown et al. 2018), many works have contributed major new insights into the census of OCs. Works such as Cantat-Gaudin et al. (2018), Cantat-Gaudin & Anders (2020), and Jaehnig et al. (2021) provide new membership lists for OCs with a significantly higher number of stars and reduced outliers from the field when compared to pre-*Gaia* works. Thousands of

<sup>★</sup> Tables 4, B.1, and the cluster members are only available in electronic form at the CDS via anonymous ftp to cdsarc.cds.unistra.fr (130.79.128.5) or via <https://cdsarc.cds.unistra.fr/cgi-bin/qcat?J/A+A/>

<sup>★★</sup> Fellow of the International Max Planck Research School for Astronomy and Cosmic Physics at the University of Heidelberg (IMPRS-HD).

new OCs have been reported using a range of unsupervised machine learning techniques, such as in Castro-Ginard et al. (2018, 2019, 2020, 2022), Cantat-Gaudin et al. (2019), or Liu & Pang (2019). The reliability of the census has also been improved, with works such as Cantat-Gaudin & Anders (2020) finding that a number of OCs discovered before *Gaia* are likely to be asterisms.

One might wonder how much further *Gaia* can improve the census of OCs, and what these improvements could reveal. In Hunt & Reffert (2021) (hereafter Paper 1), we compare three different approaches for recovering OCs in *Gaia* DR2 data, and find that the HDBSCAN clustering algorithm (Hierarchical Density-Based Spatial Clustering of Applications with Noise, Campello et al. 2013) is the most sensitive approach, although it is essential to reduce false positives with additional post-processing. In this work, we conduct the largest blind search for star clusters to date in *Gaia* data, using *Gaia* DR3 (Gaia Collaboration et al. 2021), methods developed in Paper 1, and additional validation criteria based on the photometry of every detected cluster.

In Sect. 2, we describe the *Gaia* DR3 data used in this work and the quality cuts we adopted to filter out unreliable sources. In Sect. 3, we briefly recap our clustering method from Paper 1 and tweaks made to improve cluster recovery within 1 kpc. We then outline a method to validate cluster candidates using their photometry in Sect. 4, which we generalise to additionally infer ages, extinctions, and photometric distances to our clusters in Sect. 5. In Sect. 6, we crossmatch our catalogue against literature works. Section 7 presents an overview of our catalogue. We discuss the non-detections of some literature clusters in Sect. 8, and discuss required steps a future work will take to improve the reliability of our new cluster candidates in Sect. 9. Section 10 summarises this work.

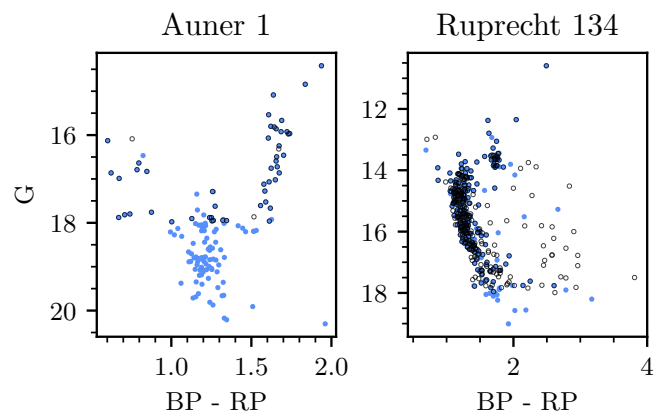
During the preparation of this work, we found that many of the star clusters we detect appear much more compatible with unbound moving groups than bound OCs, regardless of the quality of their photometry or how strong of an overdensity they are. In an upcoming third paper, we will classify the clusters resulting from this work into bound and unbound clusters, which will result in our final catalogue. This work will follow shortly (Hunt & Reffert, *in prep.*).

## 2. Data

In this section, we present a brief overview of *Gaia* DR3 data and the preprocessing steps applied to prepare it for clustering analysis.

### 2.1. Gaia DR3

The latest release of *Gaia* (Gaia Collaboration et al. 2016) astrometry and photometry, *Gaia* DR3, presents an update to *Gaia* DR2, based on an extra 12 months of data and various improvements to data processing. Astrometric and photometric data were released early in *Gaia* EDR3 (Gaia Collaboration et al. 2021), with the full DR3 release containing other data products such as low-resolution spectra and updated radial velocities that we also make limited use of in this work Gaia Collaboration et al. (2022). In total, DR3 contains 1.47 billion sources with 5- or 6-parameter astrometry, with a 30% improvement in parallax precisions and a roughly doubled accuracy in proper motions. These improvements have a large impact on the detectability of OCs in *Gaia* – particularly for proper motions, where distant OCs have a signal-to-noise ratio (S/N) increased by a factor of  $\sim 4$  in *Gaia*



**Fig. 1.** Comparison of cluster membership lists detected using *Gaia* DR3 data cut at  $G < 18$  (black empty circles) and a Rybizki et al. (2022) v1 criterion greater than 0.5 (blue filled circles) using separate runs of HDBSCAN and our pipeline for each cut, shown for Auner 1 (left) and Ruprecht 134 (right).

DR3 proper motion diagrams, owing to the halving in size of the Gaussian distribution of stars in both axes for distant clusters with proper motion dispersions smaller than *Gaia* errors.

In addition, many improvements have been made to the processing and understanding of *Gaia* data and systematics for *Gaia* DR3. Most notably for OCs, Lindegren et al. (2021b) provide a recipe for greatly reducing remaining parallax systematics for most sources in *Gaia* DR3 down to a few  $\mu\text{as}$  in the best cases, which should significantly improve the accuracy of distances to the most distant clusters. Cantat-Gaudin & Brandt (2021) provide a recipe for correcting the proper motions of certain bright stars around  $G \sim 13$ . While both of these corrections are too small to make a difference in unsupervised cluster searches, they are included in later cluster parameter determinations to improve the accuracy of final catalogue values.

### 2.2. Outlier removal

Despite improvements between *Gaia* DR2 and DR3, many sources in the catalogue are still unreliable due to a number of reasons. For instance, blending in crowded fields can cause both astrometric and photometric errors, with sources being erroneously combined or split for any or all *Gaia* measurements of the source. This is a particular issue in regions of the galactic disk with high numbers of sources. In addition, resolved and unresolved binary stars in DR3 may contribute significant errors to derived astrometric measurements for these sources, especially when their period is close to the one year baseline used to measure parallaxes (Penoyre et al. 2022; Lindegren et al. 2021a), as well as causing issues with photometric measurements due to blending (Riello et al. 2021; Golovin et al. 2023).

To remove unreliable sources, a number of different quality cuts were investigated, both in isolation and combined: firstly, simple magnitude cuts, including  $G < 18$  as adopted in works such as Paper 1 and Cantat-Gaudin et al. (2018),  $G < 19$ , and  $G < 20$ ; secondly, a cut on renormalised unit weight error (RUWE) values in the main *Gaia* source table; and finally, a cut presented in Rybizki et al. (2022), which uses a neural network and 17 diagnostic columns in the *Gaia* EDR3 data release to classify astrometric solutions as reliable and unreliable, where we required a quality value of at least 0.5.

To evaluate the performance of these cuts, the reliability of cluster recovery with HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise, Campello et al. 2013; McInnes et al. 2017) was inspected manually for 15 challenging to detect clusters given different combinations of these cuts. Notable clusters in this process include Ruprecht 134, a difficult to recover cluster located in the most crowded region of the galactic disk at  $l, b = (0.28^\circ, -1.63^\circ)$  and at a distance of  $\sim 3$  kpc, in addition to a number of clusters reported in Cantat-Gaudin & Anders (2020) but not detected in Paper 1 in *Gaia* DR2, such as Berkeley 91 and Auner 1.

A single, magnitude-independent cut based only on the quality flag of Rybizki et al. (2022) was found to outperform all other cuts trialed for cluster recovery. On average, for the trial set of 15 clusters, clusters recovered using this cut had the highest S/N of any recovered by any of the trialed cuts, with S/Ns being an average of 65% higher than clusters recovered using the  $G < 18$  cut common in the literature (see e.g. Cantat-Gaudin et al. 2018; Castro-Ginard et al. 2022). Clusters almost always had more member stars than a simple  $G < 18$  cut, with up to around twice as many member stars for distant, faint clusters where only giant stars can be resolved for magnitudes  $G < 18$ , such as for the distant cluster Auner 1 at a distance of 6.8 kpc. Inevitably, this cut should result in more complete membership lists and a more complete overall catalogue of clusters.

As a visual example, the CMDs of Auner 1 and Ruprecht 134 from clustering analyses using this cut and a  $G < 18$  cut are compared in Fig. 1. Auner 1 is a distant and difficult to detect cluster, for which only 51 stars are detected in the  $G < 18$  trial for a cluster S/N of  $10.8\sigma$ . However, the Rybizki cut cluster includes many additional faint sources, for a total of 139 member stars and an improved S/N of  $17.9\sigma$ . In the case of Ruprecht 134, a massive cluster in a crowded region near the galactic centre, the Rybizki cut cluster has fewer sources than the  $G < 18$  cut (277 to 355) but a higher S/N ( $24.7\sigma$  to  $16.6\sigma$ ), with the Rybizki cut removing a number of spurious sources from the cluster membership and the field – improving the cluster membership list and the cluster’s contrast against field stars.

Compared to having no cut at all, adoption of this cut typically has a minimal impact on the number of member stars for all clusters – it appears that sources with unreliable astrometry are already so unreliable that their position in 5D *Gaia* astrometry is too far from the bulk cluster position to be tagged as members, and few outliers are removed from cluster CMDs by this (or any) cut. Instead, in the crowded region at the galactic centre around Ruprecht 134, 85% of the sources in this field were removed by the cut, yet all reliable clusters in this field (including the nearby UFMG 88 reported by Ferreira et al. 2021) remained with a similar membership list to with no cut at all. In addition, the lack of a magnitude cut means that in sparse fields where faint sources have reliable astrometry, clusters such as the high galactic latitude Blanco 1 have membership lists down to fainter than  $G \sim 20$ , two magnitudes fainter than the membership list of Cantat-Gaudin & Anders (2020) for this cluster.

Only the v1 version of the Rybizki et al. (2022) quality flag was available during preparation of cluster membership lists in this work, for which a minimum value of 0.5 was adopted. Later versions of the initial Rybizki et al. (2022) pre-print and eventual published paper have a slightly improved version of the quality flag, although in practice it was found to make a negligible difference to the final results of this work and so clustering analysis was not revised to include it.

In total, 729.7 million sources in *Gaia* DR3 have a Rybizki et al. (2022) v1 quality flag of at least 0.5 and were selected for

further clustering analysis in this work. This represents significantly more sources than the 301.7 million sources with  $G < 18$ , a cut adopted in works such as Castro-Ginard et al. (2022) or Cantat-Gaudin & Anders (2020), and should result in a greater total number of both detected clusters and member stars.

### 2.3. Data partitioning

Finally, due to computational reasons, we partition the *Gaia* dataset into three separate collections for further analysis, as it is not possible to efficiently perform clustering analysis with 729.7 million sources at once. We aim to divide the *Gaia* dataset in such a way so that no more than 20 million sources are in any one field and so that a cluster of around 20 pc tidal radius can always be reliably detected regardless of its distance or location within adopted fields, which should be a reasonable upper size limit for almost all OCs based on Kharchenko et al. (2013) and Cantat-Gaudin & Anders (2020).

As in Paper 1, the HEALPix (Hierarchical Equal Area iso-Latitude Pixelation) tessellation scheme was used to segment the entire *Gaia* dataset (Górski et al. 2005), with calculations performed by the Python package Healpy (Zonca et al. 2019). This has advantages over other methods to subdivide spheres into a finite number of regions, in that all regions at a given tessellation level have the same area, and spherical distortions are minimised. However, unlike in Paper 1, the origin of the HEALPix grid was set at the origin of galactic coordinates ( $l, b = (0^\circ, 0^\circ)$ ), instead of the default ICRS origin at right ascension and declination values of  $\alpha, \delta = (0^\circ, 0^\circ)$  used in *Gaia* data releases, as this places most remaining spherical distortions at high galactic latitudes where we expect to find few clusters, meaning that all fields on the most important regions of the galactic disk are simple quadrilaterals.

We adopted three different partitioning schemes to detect clusters in three different distance ranges: those more distant than 750 pc, those closer than 750 pc, and those closer than 150 pc. Each scheme used large enough fields to detect clusters at each different distance range, but while minimising the number of stars in each field to keep the fields feasible to perform clustering analysis on. Firstly, for the most distant clusters, we adopted the same methodology as in Paper 1, dividing the entire *Gaia* dataset into 12288 HEALPix level five pixels. To avoid losing clusters on the edge of each pixel, each pixel is grouped into fields containing the pixel itself and its eight nearest neighbours, effectively overlapping each  $\approx 5.5^\circ \times 5.5^\circ$  field by  $1.8^\circ$  with all surrounding neighbours, with every pixel appearing in nine separate fields and in the centre of one. Next, to detect clusters closer than 750 pc, a HEALPix level two scheme with 192 pixels was adopted, containing only sources with  $\varpi > 1$  mas, using the same nine pixels per field system and resulting in overlapping fields of size  $\approx 44^\circ \times 44^\circ$ . Finally, for clusters closer than 150 pc, which can have large extents on the sky, a single field containing all stars closer than 250 pc was used, based on photo-geometric distances to sources in Bailer-Jones et al. (2021).

Between these three systems, all bound members of all open clusters of size 20 pc or smaller should be contained within these fields – although in reality, this is only a worst-case constraint at the 750 pc and 150 pc crossover points and for a cluster in the worst possible location in a field, and many significantly larger clusters (including tidal tails many times their size) would be detectable in other regions.

### 3. Cluster recovery

Next, we discuss the methodology we adopted to recover clusters in *Gaia* data, assign basic parameters, and crossmatch to existing cluster catalogues in the literature.

#### 3.1. HDBSCAN

Many different algorithms have been used to date to recover clusters in *Gaia* data. We present a review and full explanation of these algorithms in Paper 1, in which we found that the HDBSCAN algorithm (Campello et al. 2013; McInnes et al. 2017) is the most sensitive for recovering OCs in *Gaia* data.

Briefly, HDBSCAN is an updated version of the DBSCAN algorithm (Ester et al. 1996), for which only a minimum cluster size  $m_{clSize}$  and minimum number of points in the neighbourhood of a cluster core point  $m_{pts}$  must be specified, unlike DBSCAN which instead uses  $m_{pts}$  and a minimum, global distance between points in a cluster  $\epsilon$ . DBSCAN has seen much use in the literature so far for OC recovery, such as in Castro-Ginard et al. (2018, 2019, 2020, 2022) or He et al. (2021, 2022a). The main challenge of DBSCAN is that  $\epsilon$  must be set globally for an entire dataset, which can limit the sensitivity of the algorithm for datasets of varying density – such as the *Gaia* dataset, which has different densities at different distances and locations within the galaxy.

Instead, HDBSCAN copes with varying density datasets by effectively considering all possible DBSCAN  $\epsilon$  solutions for all regions of a dataset, selecting the best clusters based on the lower limit of cluster size  $m_{clSize}$ . HDBSCAN has so far been used to detect moving groups in *Gaia* data by Kounkel & Covey (2019) and Kounkel et al. (2020), as well as being used to find 41 new OCs in Paper 1, and being used by Tarricq et al. (2022) to reveal new tidal tails and comas of numerous OCs within 1.5 kpc. HDBSCAN has not yet been used to conduct a search through all *Gaia* data for OCs.

A major flaw of HDBSCAN, however, is its high false positive rate. In Paper 1, we show that this is due to the algorithm being overconfident, reporting dense random fluctuations of a given dataset as clusters. To mitigate this, we adopt the cluster significance test (CST) from Paper 1, which searches for field stars surrounding a cluster and compares the nearest neighbour distribution of cluster stars with that of field stars. This then produces a signal-to-noise ratio (S/N), with CST scores greater than  $5\sigma$  corresponding to highly likely clusters.

The issue of how to convert the five dimensions of *Gaia* astrometry into a form best usable by a clustering algorithm is an open problem. Converting proper motions and parallaxes to velocities and distances respectively is one such approach (e.g. as in Kounkel et al. 2020; He et al. 2022a), although a major issue is that converting *Gaia* parallaxes to distances is non-trivial and results in asymmetric errors and non-Gaussian parameter distributions (Bailer-Jones et al. 2021). Instead, we use the approach adopted in Paper 1, similar to that of works such as Castro-Ginard et al. (2018) and Liu & Pang (2019). We use *Gaia* positions, proper motions, and parallaxes directly, but with two preprocessing steps: firstly, recentring them into a coordinate frame with an origin at the centre of each respective field, which removes spherical distortions present at high declinations; secondly, rescaling all five axes of the dataset to have the same median and interquartile range, effectively removing the units of each axis of the data. Particularly for HDBSCAN, which can cope with varying density datasets, the choice to use these five simple recentred and rescaled features was found to have no im-

pact on the detectability and membership lists of nearby clusters, while having great benefits for clusters more distant than  $\sim 2$  kpc, for which a distance-based approach causes many clusters to have sparser, non-Gaussian, and more challenging to detect distributions.

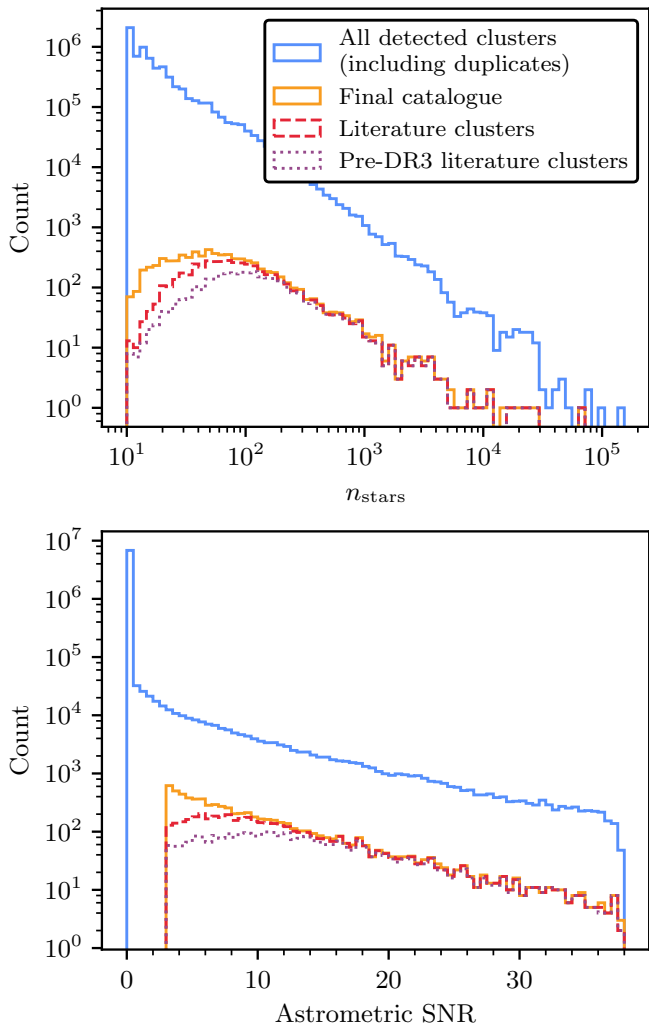
The one exception to this in this work is for the single field of all stars within 250 pc, which was adopted to help improve the accuracy of cluster membership lists for very nearby clusters with large angular extents on the sky such as the Hyades. Given that this field covers the entire sky, it is not possible to avoid high latitude spherical distortions with a simple recentring; instead, photo-geometric distances from Bailer-Jones et al. (2021) were used to convert positions and parallaxes to a Cartesian coordinate frame, with proper motions converted to tangential velocities. At such small distances, the uncertainties in Bailer-Jones et al. (2021) are small and not prior-dominated, and so reliance on *Gaia*-derived distances for the single nearby field should not cause any issues.

#### 3.2. Clustering analysis and catalogue merging

Using HDBSCAN and the same range of parameter choices as in Paper 1 ( $m_{clSize} \in \{10, 20, 40, 80\}$ ,  $m_{pts} = 10$ ), clustering analysis on all HEALPix level two and five fields was completed in around eight days of runtime on a machine with a 48 core Intel(R) Xeon(R) E5-2650 CPU with 48 GB of RAM. This run was mostly RAM-limited due to the worst-case  $O(n^3)$  memory use of the HDBSCAN implementation used for the largest fields. Given that fields overlap and that different parameter choices can detect the same cluster, each cluster can be duplicated up to four times within a single field, up to nine times by appearing in all neighbouring fields and a further time by appearing in different distance ranges (if the cluster has a distance between 0.7 to 1 kpc, or less than 250 pc). Hence, in the worst case, a single cluster could be duplicated 72 times. It is essential and non-trivial to merge the results of all fields accurately and without losing or duplicating any one individual cluster.

In total, 7.1 million different clusters were detected (including duplicates), almost all of which are astrometric false positives due to the oversensitivity flaws of HDBSCAN discussed in Paper 1. These clusters can be removed by using their astrometric S/N, as derived by the CST. Figure 2 shows histograms of the S/Ns of detected clusters, showing a clear spike in count for  $S/N < 0.5$  and an increasing trend in S/N for  $S/N \lesssim 3$  that deviates from the relatively straight log-linear relation in S/N present for  $S/N > 3$ , suggesting that an additional component of false positives is contributing to the otherwise log-linear component of reliable astrometric clusters at low S/Ns. This figure, our results from Paper 1, and the poor quality of the low-S/N clusters we detect strongly support that most low-S/N clusters are false positives; however, exactly where to set an S/N threshold is a non-trivial decision that has a large effect on the rest of the catalogue. A catalogue can choose to prioritise completeness, having a low threshold and including as many true positives as possible, but while inevitably including many false positives and sacrificing precision; or, a catalogue can do the opposite, having a lower completeness but also minimal false positives and maximised reliability of all objects in the catalogue.

For the purposes of this work, we chose to prioritise the precision and reliability of the catalogue, adopting a higher threshold on the minimum S/N of clusters. This sacrifices some completeness so that all final catalogue entries are likely to be real astrometric overdensities and not mere statistical fluctuations. This approach also comes with a key advantage. Our field tiling strat-



**Fig. 2.** Statistics of all detected clusters compared against the final catalogue. *Top*: distribution of the number of member stars of detected clusters,  $n_{\text{stars}}$ , for all detected clusters in all fields before catalogue merging and duplicate removal (solid blue line), for the final catalogue (solid orange line), and amongst clusters in the final catalogue that crossmatch to clusters in the literature, for all literature clusters (solid red line) and for only those detected before the release of *Gaia* EDR3 (dotted purple line). *Bottom*: as above, but for the astrometric S/N (CST score) for all clusters in these sets. S/Ns have a maximum value of 38 due to numerical reasons.

egy aimed to prevent any real clusters from being ‘lost’, aiming to recover  $> 99\%$  of real, good-quality OCs in a single catalogue. However, merging the results of so many separate clustering runs is a difficult and non-trivial task, and early experiments showed that the inclusion of false positives in the catalogue had a severe effect on the reliability and accuracy of the catalogue merging process. It was common that false positives and clear real OCs would share members in different clustering runs, meaning that low S/N thresholds on the final catalogue would adversely affect the catalogue’s completeness at higher S/Ns. For the purposes of this work, we set a higher threshold on the minimum S/N, requiring  $\text{S/N} > 3\sigma$ . This cut was found to maximise the quality of later catalogue merging steps, while removing a high number of false positives and retaining reliable clusters. Many false positives share member stars with real OCs, which greatly complicated the merging process and made the choice of which cluster

to keep challenging. A single S/N cut means that our incompleteness is well characterised and easy to understand, whereas lower cuts were found to adversely affect catalogue completeness even at high S/Ns in a difficult to characterise way. In addition, while our adopted cut is at an S/N of  $3\sigma$ , clusters with an S/N lower than even  $5\sigma$  may have minimal scientific usefulness, as they cannot be asserted as being real astrometric overdensities beyond any reasonable doubt; as such, it is not worth including such clusters in the catalogue at the expense of the recovery of better, real objects.

Inevitably, some low-S/N real OCs are likely to be lost in this process. We discuss the number of literature objects that are lost due to this cut in Sect. 8.1.3, and we briefly discuss some of the improvements to clustering algorithms that could be used to simplify the merging process and entirely remove the need for an S/N cut to ensure the catalogue’s reliability in Sect. 10.

After dropping unreliable low S/N clusters, the results of each parameter run in every field were merged. For clusters where every  $m_{\text{clsize}}$  detected an identical object, duplicates were simply dropped. In some cases (such as for the largest OCs and GCs), smaller  $m_{\text{clsize}}$  runs may split the cluster into two sub-clusters. Generally, it was possible to remove duplicate small subclusters by only keeping the single largest cluster. This process was extensively checked by hand, keeping smaller clusters instead in the case of some binary and coincident clusters which are better selected as being split, which was aided by fitting Gaussian mixture models to every cluster and evaluating the Bayesian information criterion of one and two-component fits, flagging clusters where a two component fit was preferred for potential splitting.

Secondly, cluster duplicates between fields must be removed. Using maximum likelihood distances calculated with the method presented in Cantat-Gaudin et al. (2018), clusters likely to be affected by edge effects or likely to be better detected at a different HEALPix level were removed. Clusters from the 250 pc run were only kept if they were closer than 175 pc. Clusters from the HEALPix level 2 run were only kept with distances between 150 and 750 pc. Finally, clusters from the HEALPix level 5 run were only kept if they had distances greater than 700 pc. The small overlaps in these distance ranges allow the best cluster to be selected later for clusters on the boundaries.

Next, duplicate clusters due to the overlap between fields must be removed. As each field is composed of nine pixels, a cluster can appear in up to nine separate fields. Keeping only clusters in the central pixel of every field is sufficient to mostly remove duplicates, retaining only the best cluster detection in the central pixel where edge effects are minimised. However, cluster membership lists are often not identical between fields, and it is hence possible that a cluster’s mean position could be different enough between runs to appear in the central pixel of multiple fields or to never appear in the central pixel of any field. Particularly for small clusters of 20 stars or less, the inclusion or removal of even a single star can have a reasonable impact on the mean position of the cluster. This effect is worst for the nearest clusters with the largest angular extents on the sky relative to the field they are in. While this effect only impacts a small number of clusters (causing around  $\sim 1\%$  of clusters reported in Cantat-Gaudin & Anders (2020) to be lost), it is nevertheless important to address to ensure the final catalogue is as complete as possible.

To mitigate this effect, clusters near to the edge of a central pixel were also kept. After extensive testing, it was found that cluster positions generally vary by no more than  $\sim 1$  pc at the distance of the cluster between different fields. We adopt a more



tolerant cut corresponding to  $\sim 5$  pc for a cluster at a worst-case distance, such that clusters within  $1.91^\circ$  (HEALPix level 2) or  $0.41^\circ$  (HEALPix level 5) of the edge of a central pixel were also kept. This is small compared to the overall field sizes of  $\approx 44^\circ \times 44^\circ$  (HEALPix level 2) or  $\approx 5.5^\circ \times 5.5^\circ$  (HEALPix level 5), but was nevertheless found to be sufficient to avoid losing any genuine clusters.

These processes removed most duplicated clusters while minimising the number of clusters lost during the merging process, although some duplicates still remained within the allowed overlaps between fields. These clusters were removed by looking for clusters with similar membership lists, mean positions, mean proper motions, and mean parallaxes, and selecting the cluster in each case with only the highest distance from any field edge. This process was also verified extensively by hand. For 23 large clusters (typically with tidal tails larger than the field they are in), duplicate clusters were similar but with both having additional members. In these cases, the clusters were merged into single clusters.

Finally, the catalogue was checked for clear, known binary clusters that were not correctly split by HDBSCAN. Four probable cases were identified, including the close binary Collinder 394/NGC 6716 as well as UBC 76/UBC 77. Generally, these binary clusters had very similar proper motion and parallax distributions, making them difficult or impossible for the HDBSCAN algorithm to split – particularly since HDBSCAN cannot assign members to two clusters at once, although this is necessary for such close and difficult to separate objects. These clusters were split with Gaussian mixture models by selecting the number of components with the highest Bayesian information criterion. In all four cases, multiple components were preferred over a single component. It is likely that some other objects in the catalogue may also be better described as binary clusters, although this would need to be investigated carefully on a case-by-case basis (see e.g. Kovaleva et al. 2020; Anders et al. 2022) or with analysis using improved astrometry of a future *Gaia* data release. This resulted in a list of 7788 clusters for further analysis.

### 3.3. Additional parameters and membership determination

Cluster parameters were mostly determined following the same approach as in Paper 1. However, it was noticed that many clusters are detected with tidal tails or comas, despite this study not being initially designed to detect cluster tidal tails. This is particularly common for clusters within  $\sim 2$  kpc. In many cases, this can cause clusters to have strongly biased mean parameters, such as for the cluster Mamajek 4 at a distance of 444 pc. Mamajek 4 has a tidal tail that stretches for  $15^\circ$  or 100 pc from its core, although only one side of the tail is detected due to limitations of the size of the field it was detected in. Using a simple mean position and proper motion for such clusters is hence affected by this asymmetry and is strongly biased.

Instead, we aim to derive cluster parameters for the central part of clusters only. In practice, particularly for dissolving clusters with a majority of their mass in their tidal tails, it can be difficult to decide where stars should be called members of the cluster or members of the field. For instance, Tarricq et al. (2022) attempted to derive structural parameters for 467 OCs within 1.5 kpc, but their method (based on fitting King (1962) profiles) only succeeded on 389 clusters. To allow for accurate parameters to be inferred for all clusters homogeneously, we adopt a simple methodology comparing the density of cluster members with that of the field.

Firstly, cluster members with a HDBSCAN membership probability of less than 50% were discarded. HDBSCAN membership probabilities are not based on *Gaia* uncertainties, but rather only on the proximity of a given member to the bulk of the cluster. It was noticed that membership probabilities lower than this limit always correspond to low-quality cluster members or members of tidal tails, and are hence not worth including in the determination of reliable parameters of clusters.

Next, using these members, cluster centres are derived in a way insensitive to asymmetries. Kernel density estimation was used to select the modal point of the cluster stellar distribution, with a bandwidth set to 1 pc at the distance of the cluster.

Finally, using this cluster centre, the radius at which the overall cluster has the best contrast to field stars was selected. In practice, this is similar to the King (1962) definition of tidal radius as the radius at which a cluster's density begins to exceed that of the density of the field, but is model-independent and can be easily and efficiently computed for the entire catalogue by selecting the radius at which a cluster has the highest CST against field stars. For instance, for well-defined clusters such as the Pleiades and Blanco 1, this radius was found to exclude cluster tidal tails while corresponding well with literature tidal radius values in Kharchenko et al. (2013) (see Sect. 7 for a discussion of our cluster radii.)

Mean parameters such as mean proper motion and parallax were then calculated given the members within the cluster's estimated tidal radius, in addition to maximum likelihood cluster distances calculated using the method of Cantat-Gaudin et al. (2018). To calculate more accurate distances, the parallax bias of member stars was corrected using the method in Lindegren et al. (2021b), which improved the accuracy of cluster distances particularly for distant clusters. As the Lindegren et al. (2021b) parallax correction can only be applied for certain parameter ranges, for six clusters, too few sources (or no sources) had available corrections, and so we applied a simple global offset of  $\varpi_0 = -17 \mu\text{as}$  as derived in Lindegren et al. (2021b). These six clusters are flagged in the final catalogue as having less accurate distances. Overall, although the Cantat-Gaudin et al. (2018) distance method assumes that the size of clusters is negligible compared to their distance, which introduces a bias for nearby clusters, our astrometric cluster distances were nevertheless found to agree well with the literature. For instance, we derive a distance of  $47.19^{+0.004}_{-0.005}$  pc to the Hyades, which is comparable to the  $47.34 \pm 0.21$  pc distance in McArthur et al. (2011), who use Hubble Space Telescope parallaxes to a subset of Hyades member stars to derive its distance.

In addition, King (1962) core radii were estimated given our estimated tidal radius  $r_t$  and radius containing 50% of members of the core  $r_{50}$ , since there exists only one solution to the number density equation in King (1962) (Eqn. 18) given  $n(r_{50})$  and  $r_t$ . While approximate and less accurate than full Markov chain Monte-Carlo (MCMC) fits such as those performed in Tarricq et al. (2022), these core radii still provide a good approximation of a King (1962) model fit and compared well to literature values for well-defined clusters for which different works have similar membership lists. Having calculated basic astrometric parameters for our clusters, we next calculate photometric parameters for our clusters using convolutional neural networks.

## 4. Photometric validation

In this section, we use photometry to validate members of the cluster catalogue as being compatible with single-population OCs and infer basic parameters, entirely using neural networks

**Table 1.** Probability distributions used for simulated clusters for training of the CMD classifier.

Param.	Range	Distribution
$\log t$	[6.4, 10.0]	$\mathcal{U}(6.4, 10.0)$
[Fe/H]	[−0.5, 0.5]	$\mathcal{B}(4.0, 4.0) - 0.5$
$m - M$	[3.2, 15.73]	$\mathcal{U}(3.2, 15.73)$
$A_V$	[0.0, 8.0]	$\mathcal{B}(\sqrt{d/3}, \sqrt{d/5}) \cdot 8 \tanh(d/2)^a$
$n_{\text{stars}}$	[10, 10000]	$10^{3 \cdot \mathcal{B}(2, 3.5) + 1}$
$\sigma_{\Delta A_V}$	[0.0, 0.6]	$0.4 \cdot \mathcal{T}(1.25)$
$l$	[0°, 360°]	$\mathcal{U}(0, 360)$
$b$	[−90°, 90°]	$90 \cdot \mathcal{S} \cdot \mathcal{R}(\mathcal{B}(1, 35), \mathcal{B}(1, 12), 2/3)$

**Notes.** Distributions of parameters are quoted as uniform distributions  $\mathcal{U}(a, b)$  between  $a$  and  $b$ , beta distributions  $\mathcal{B}(a, b)$  with parameters  $a$  and  $b$ , truncated exponential distributions  $\mathcal{T}(a)$  truncated at  $a$ ,  $\mathcal{R}(a, b, x)$  which is a weighted choice with probability  $x$  of choosing value  $a$  and probability  $1 - x$  of choosing value  $b$ , and  $\mathcal{S}$  which is a random sign with value  $+1$  or  $-1$ . <sup>(a)</sup> Distances  $d$  in kpc.

and simulated data. While Castro-Ginard et al. (2018, 2019, 2020, 2022) successfully use neural networks to classify candidate clusters as real or false with their photometry, and while Cantat-Gaudin et al. (2020) and Kounkel et al. (2020) use neural networks to infer the ages, extinctions, and distances of their catalogued clusters, all of these works rely partially or entirely on existing examples of OCs detected in *Gaia*.

While such an approach mitigates issues with simulated training data, namely that stellar isochrones such as Bressan et al. (2012) are typically an imperfect fit to the observed CMDs of OCs (Cantat-Gaudin et al. 2020), it is difficult to guarantee that a small training dataset that relies mostly or entirely on examples of OCs from *Gaia* accurately covers a full range in parameters such as absolute extinction, differential extinction, distance, metallicity, and age. In particular, due to the different cuts on *Gaia* data used in this work, we often detect significantly more member stars for many clusters and up to two magnitudes fainter than the membership lists of Cantat-Gaudin & Anders (2020); hence, particularly for more distant OCs, our membership differences have a significant impact on inferred parameters, making existing literature catalogues inappropriate to use as training data. Simulated data, if it can be simulated accurately enough, would offer an attractive way to quickly generate new training data applicable to new methodologies and new *Gaia* datasets or even other instruments, entirely based on a ground truth or ‘best estimate’ of how OCs should appear based on prior knowledge from stellar evolution models. Additionally, training data based on real clusters are biased towards an unknown selection effect of how a human defines a real cluster – whereas for simulated data, we are able to exactly state the distributions we assume real OCs are drawn from, hence giving more knowledge of any selection biases this may cause.

A key issue found in early experiments is that typical machine learning approaches are deterministic, and hence do not quantify the underlying uncertainties on their predictions. To aid with the use of simulated data, we adopt an approximate Bayesian neural network (BNN) framework using variational inference. In practice, true Bayesian machine learning is impractical to achieve with current methods; however, variational inference-based approaches offer an approximate and fast way to estimate the uncertainty of a neural network model by approximating parameters with simple probability distributions (Goan & Fookes 2020; Jospin et al. 2022), of which networks can then be sampled multiple times to produce a probability distribution

for their output. The BNN approach we trialed had similar accuracy to a purely deterministic one except while also outputting uncertainties, allowing us to estimate the uncertainty of our classifier. We provide a broader overview of our adopted variational inference-based approach in Appendix C. Next, we discuss the creation of training data for our CMD classifier.

#### 4.1. Simulated real OCs

A number of steps were used to generate examples of real OCs to train our CMD classifier. Basic OC generation was conducted using SPISEA (Hosek Jr et al. 2020) to simulate single-population clusters from PARSEC evolution models (Marigo et al. 2017), with extinction calculated star-by-star using a Cardelli et al. (1989) extinction law with  $R_V = 3.1$ . Stars were sampled from these isochrones with SPISEA using a Kroupa (2001) IMF. In addition, SPISEA was used to supplement simulated OC CMDs with unresolved binary stars based on general relations derived in Lu et al. (2013) for zero-age star clusters. The values in this work were found to correspond relatively well to *Gaia* observations, with a mass-dependent multiplicity frequency peaking at 100% for clusters of masses above  $5 M_\odot$ . In practice, unresolved binary stars have negligible impact on the final cluster CMDs fed to the network, as typical binary sequences observed in *Gaia* photometry are smaller than the size of the pixels in input CMD images. SPISEA was also used to apply Gaussian-distributed differential reddening, with values up to a standard deviation of 0.6 in the highest cases, reflecting the most extreme examples of differentially reddened reliable clusters found in Cantat-Gaudin & Anders (2020).

Next, a random location on the galactic disk was selected for each cluster, which was used to simulate a realistic selection function and photometric errors. The magnitude-dependent selection function of *Gaia* DR3 at each given location was queried using the `selectionfunctions` package presented in Boubert & Everall (2020) and Boubert et al. (2020), which gives the basic probability that a source appears in *Gaia* as a function of position and G-band magnitude. We use the online version of their package updated for *Gaia* DR3. The `selectionfunctions` package is based on the `dustmaps` package from Green (2018). In addition, the selection function of every cluster was also corrected for the cuts on *Gaia* data applied in Sect. 2.2. During the preparation of this work, Cantat-Gaudin et al. (2023) released a new selection function for *Gaia* DR3 which suggested that the earlier work of Boubert & Everall (2020); Boubert et al. (2020) can be over-confident at the faint end; however, given that our cluster membership lists are overwhelmingly dominated by the selection function of our cuts on *Gaia* data at magnitudes  $G > 18$ , and not the pure selection function of *Gaia*, we found that it made too small of a difference to our simulated clusters to be worth updating our training data for, although we will adopt their work in future works. Realistic photometric uncertainties were added to sources based on the distribution of source uncertainties at the selected location, which are generally larger in crowded fields. We added systematic offsets in simulated BP and RP *Gaia* photometry for faint sources using relations in Riello et al. (2021).

Outliers were not added to simulated cluster CMDs, as most clusters are already detected with very few or no outliers; instead, we wish the CMD classifier to quantify the evidence for a cluster being real based on its photometry alone, which photometric outliers inherently reduce. In this way, CMDs of clusters with a high number of outliers are scored more negatively by the network as they have less photometric evidence supporting them being real. Blue stragglers were also not added to cluster CMDs

as they are indistinguishable from photometric outliers, although in practice, real OCs with blue straggler stars were not found to be scored significantly lower by the trained network.

10 000 examples of simulated real clusters were generated to use as one half of the simulated cluster dataset. Distributions of parameters such as age  $\log t$ , extinction  $A_V$ , differential extinction  $\Delta A_V$  and distance modulus  $m - M$  were carefully chosen after many iterations to minimise systematics deriving from the overall distribution of training data in the dataset, while ensuring that the CMD classifier was trained on a representative set of simulated real OCs. Fundamentally, the objective of the training data are not to match the real distribution of OCs, but rather to yield an unbiased and representative sample of OCs to train the BNN on, such that the BNN can provide an unbiased classification of any object. For instance, while a distribution of the number of visible stars  $n$  based on the distribution of stars in Cantat-Gaudin & Anders (2020) (corrected for our deeper magnitude limit) was found to work well to produce an unbiased classifier, in other cases, such as for  $\log t$  and  $m - M$ , the use of a uniform distribution (instead of one based on the expected distribution of clusters) was essential to avoid biasing the classifier towards certain ages or distances. These distributions are listed in Table 1.

#### 4.2. Simulated fake OCs

A number of methods to simulate fake OCs reminiscent of false positives sometimes reported by HDBSCAN were trialed. As a clustering algorithm, the member stars of each cluster reported by the algorithm are spatially correlated, with a similar position, proper motion, and parallax. Hence, it is important that false positives contain member stars with similar astrometric parameters. Simply randomly selecting stars from *Gaia* data to construct each false positive was found to result in clusters that were too pessimistic.

Instead, to generate false positives with spatially correlated member stars, a star was first selected randomly from the entire *Gaia* dataset as an origin point. This ensures inherently that false positives are more likely to occur in the densest regions of the *Gaia* dataset, which was a behaviour observed inherently for HDBSCAN in Paper 1. A total number of stars for the cluster was selected from the same distribution as used for simulated real OCs. Then, a 5D hypersphere in position, proper motion, and parallax was expanded randomly around this star until the hypersphere contained the required number of stars. In this way, false positives with spatially correlated member stars were generated. Actual OCs make up a small enough portion of the *Gaia* dataset – 610 000 in the final version of the catalogue, or fewer than 0.1% – that it was not found to be necessary to first remove them from data used to generate false positives. This is similar to the false positive generation method used in Castro-Ginard et al. (2022).

10 000 false positives were generated using this methodology to provide the other half of the training dataset. While most false positives have obviously poor quality CMDs, false positives generated from regions of field stars with roughly homogeneous ages and composition (such as from the galactic halo) often had more homogeneous CMDs, that could be compatible with highly differentially reddened OCs. However, this is a useful property of the training dataset, given the variational inference approach used in the network: this ‘overlap’ between highly differentially reddened true positives and chance alignments of somewhat-similar field stars reflects on the real distributions of field stars in the galactic disk. Real *Gaia* cluster candidates with

**Table 2.** Human classifier performance.

Dataset	Size	Percent classified as			
		TP	TP?	FP?	FP
Test data	2000	53.6	26.5	11.0	8.9
Simulated real OCs	250	72.0	20.0	6.0	2.0
Simulated fake OCs	250	14.0	26.8	28.0	31.2

**Notes.** Results of human classification when applied to a test dataset of 2000 clusters detected by HDBSCAN in this work as well as two datasets of simulated real and fake clusters.

worse-quality CMDs making them compatible with both a real OC or a chance clustering of field stars hence have broad or bimodal PDFs from the BNN CMD classifier, reflecting how photometry alone offers only poor evidence of whether or not these objects are real or fake star clusters.

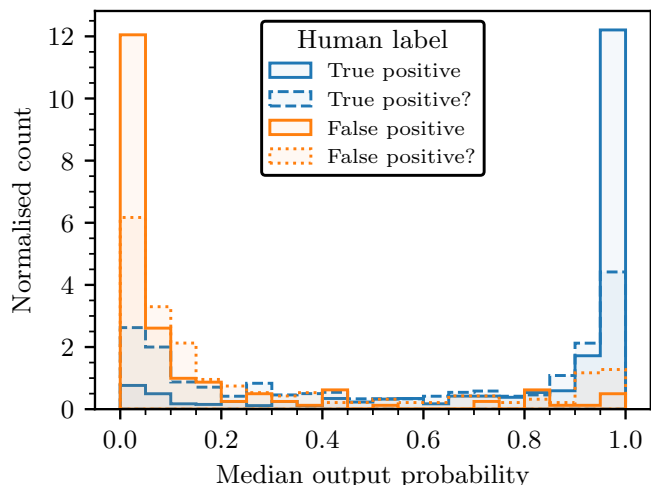
#### 4.3. Test dataset

In order to test the trained networks against real *Gaia* data and ensure that they can be generalised from their training on simulated data to use on real data, a test dataset of 2000 clusters randomly selected from the initial HDBSCAN clustering was selected and classified by hand, in addition to 250 simulated real clusters and 250 simulated fake ones to estimate the accuracy of human classification. These different datasets were classified in one classification run to avoid biasing the human classifier. Clusters were classified into ‘true positive’ (TP) and ‘false positive’ (FP) categories, in addition to two other categories for clusters that are most likely to be true or false clusters but are somewhat uncertain (abbreviated as ‘TP?’ or ‘FP?’), due to the presence of outliers, a small number of stars, or very high differential reddening that is compatible with both an association of field stars or a highly differentially reddened OC. The results of this classification are shown in Table 2.

Of clusters reported by HDBSCAN, 53.6% were hand-classified as being highly likely to be real, with a further 26.5% being potentially real, suggesting that most clusters we detect have a reliable CMD. Only 8.9% were highly unlikely to be real with a further 11.0% classified as probably not real, suggesting that around 80% of clusters reported by HDBSCAN are likely to have single stellar populations based on human classifications.

In testing the human classifier, 92.0% of simulated real clusters were correctly classified as real or potentially real, although only 59.2% of simulated fake clusters were classified as false or potentially false. 14.0% of simulated fake clusters were in fact classified as highly likely to be real. This shows the inherent limitations of using photometry to validate OCs, as spatially correlated groups of field stars can often have somewhat-homogeneous CMDs when all field stars in a given region have a similar age and chemistry (see Sect. 4.2), which can even fool a human classifier. This is particularly common in the halo and thick disk where most stars have a similar, old age. This is an important limitation of the human-classified test data to bear in mind, as a small fraction of clusters classified by hand as true positives will always in fact be false positives. Nevertheless, CMD classification is still a necessary validation tool to help ensure that detected cluster candidates are reliable, as many of the worst quality clusters can still be removed with this method.





**Fig. 3.** Performance of the CMD classifier on the independent test dataset of 2000 clusters detected by HDBSCAN in *Gaia* data and labelled by hand. Clusters are labelled as true positives or false positives, with clusters where the human classifier was less certain being additionally flagged.

#### 4.4. Network training and validation

The 20 000 simulated real and fake OCs were split randomly into a training set of 16 000 clusters and a validation dataset of 4 000 clusters to assess network overfitting. As the simulated fake OCs have a different distribution of distance moduli to the simulated real OCs, fake OCs at undersampled and oversampled distances were weighted to be emphasised more or less strongly during training, preventing systematics due to differences in distance distributions.

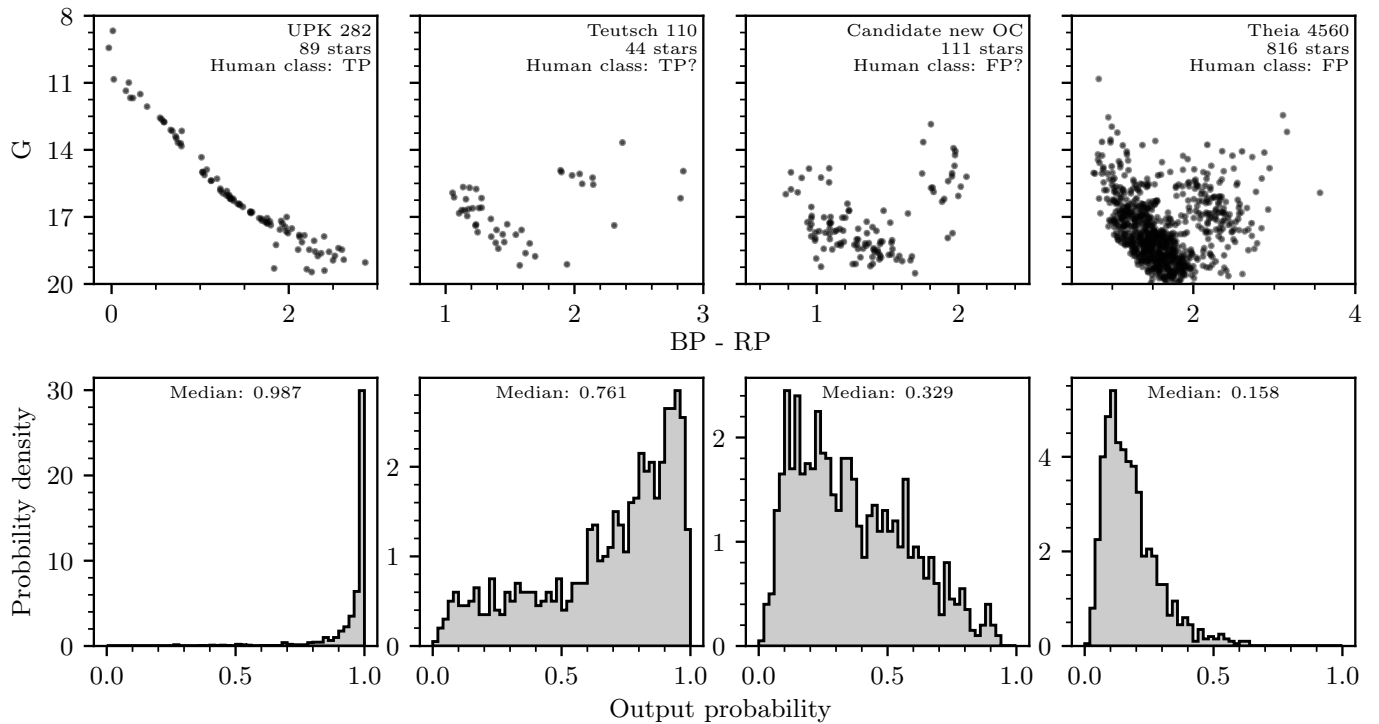
We used the implementations of neural networks and probabilistic layers in TensorFlow (Abadi et al. 2015, 2016) and TensorFlow Probability (Dillon et al. 2017) for all networks used in this work. Networks were trained with the Adam optimisation algorithm (Kingma & Ba 2017). A number of different neural network structures were trialed. Convolutional neural networks (CNNs), which convolve two-dimensional input with learnt filters, were found to perform ideally for the problem at hand, and have seen extensive use in the astronomical literature (e.g. Castro-Ginard et al. 2022; Becker et al. 2021; Killestein et al. 2021).

As input, the optimal network trialed used cluster CMDs converted to absolute magnitudes, with stars of absolute  $G$  magnitudes greater than 10 or lower than  $-2$  cut away. Generally, this cuts certain very low mass  $M$  stars and bright  $O$  stars from cluster CMDs, which were found to be poorly simulated by PARSEC isochrones with their inclusion only worsening network performance on real data. In practice, very few stars are cut due to this limitation, with  $O$  stars making up only a very small proportion of sources in young clusters and  $M$  dwarfs fainter than  $M_G = 10$  only being brighter than  $G = 20$  for clusters within 1 kpc, at which point the rest of the cluster CMD can be resolved well. In addition,  $BP - RP$  colours were cut between  $-0.4$  to  $4$ , which in practice is a wide enough colour range to include almost all sources but while providing a good range to discretise cluster CMDs between. Sources with very low  $BP$  and  $RP$  fluxes that have overestimated  $BP$  or  $RP$  magnitudes were removed using cuts from Riello et al. (2021), as these also only confused the network, despite these systematics being simulated in the training data. Finally, in terms of structure, the optimal net-

work trialed was trained on CMDs discretised into  $32 \times 32$  pixel images, corresponding to pixels of size  $0.38 \times 0.11$  mag. These images were first processed by three convolutional layers with  $5 \times 5$  pixel kernels of 6, 16, and 120 filters respectively. Max pooling layers were placed between these convolutional layers to speed up training and inference. Convolution layer output was connected to a single densely connected layer of 128 nodes, with a final single node for output. The distance modulus of the cluster based on the parallax-derived cluster distances was also fed to the network as an auxiliary input into the 128 node dense layer, in a similar way to the network of Cantat-Gaudin et al. (2020) which also uses both photometric and astrometric input simultaneously. All layers used Rectified Linear Unit (ReLU) activation other than a sigmoid activation function applied to the final output to constrain network output in the range  $[0, 1]$  as a probability distribution.

The final network had binary accuracies (the percentage of clusters given the correct true or false label) of 95% for both training and validation data, indicating that the network did not overfit to training samples when compared with other simulated data. Fig. 3 shows the performance of the network compared to the human-labelled test dataset of real clusters detected by HDBSCAN in *Gaia* after sampling the network 1000 times to generate PDFs for every object, with 85.5% of clusters labelled highly likely to be real and 91.3% of clusters labelled highly unlikely to be real having a median predicted probability greater or less than 0.5 respectively. Clusters where the human classifier was less certain have a much broader distribution, although this also reflects inherent uncertainties in the test dataset discussed in Sect. 4.3. Finally, only 4.3% and 2.5% of highly likely real and highly likely false clusters had predicted labels that disagree with human labels at more than the  $2\sigma$  level – namely, that 97.5% of their PDF is below or above 0.5 respectively. It is important to recall that these quantities merely validate the general agreement between two independent classifiers (the human classifier and the automated CMD classifier) on the same dataset, and do not exactly measure the ground truth sensitivity or accuracy of the CMD classifier, as the human class labels themselves are uncertain Sect. 4.3. Instead, these data show that the CMD classifier can perform comparably well to human classification, except with the added bonuses of speed and reproducibility.

Fig. 4 shows CMD classifier PDFs for four clusters from all human classes, including the names of any clusters that cross-matched to real objects. In general, CMD classifier predictions generally agreed well with the human-assigned labels, also generally with higher uncertainty and a broader PDF in cases where the human classifier was less certain. For clusters with clear, high-quality CMDs such as UPK 282, the CMD classifier outputs PDFs that strongly suggest they are real. Teutsch 110 is a less well-defined cluster that, if real, must have differential reddening and a few outliers, and is hence not classified as strongly. The candidate new cluster shown is a similar case albeit with a worse CMD, making it relatively unlikely to be real given this HDBSCAN detection. Finally, Theia 4560 is visible as a large and statistically significant overdensity in *Gaia* data as detected by Kounkel et al. (2020), although the overdensity as detected in this work does not appear to contain a homogeneous population of stars and is hence classified weakly. CMD classifier median probabilities and confidence intervals for all clusters are listed in Table 4, based on 1000 samples of the network for each cluster.



**Fig. 4.** Four examples of classified cluster CMDs from the test dataset, with cluster CMDs on the top row and their PDFs of predicted probabilities on the bottom row. Cluster names and human-assigned labels are indicated on the figures. PDFs are generated by sampling the CMD classifier 1000 times for every cluster.

## 5. Age, extinction, and distance inference

### 5.1. CMD classifier modifications

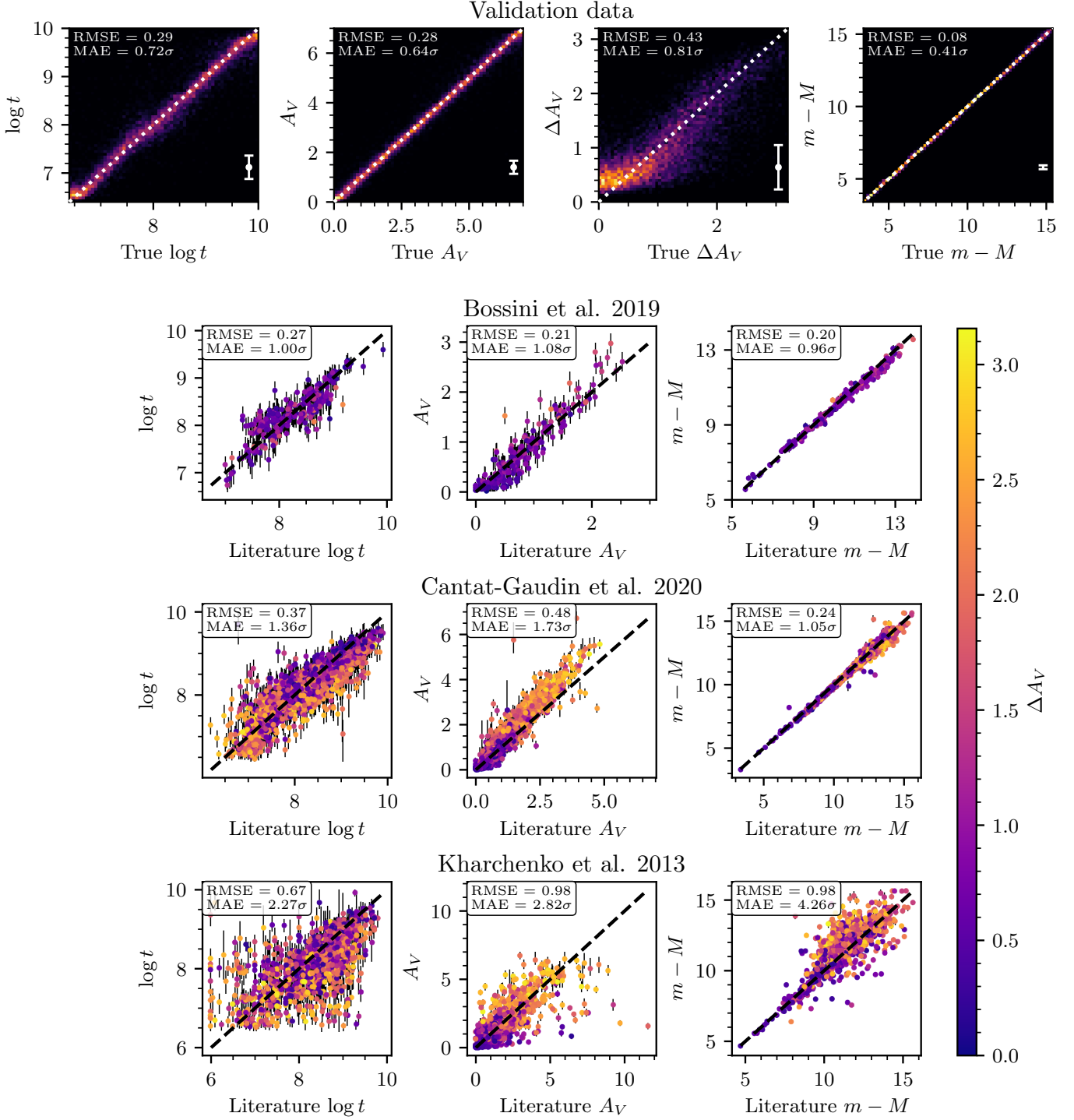
While not a main focus of this work, we also show that the approach based on simulated data and an approximate BNN using variational inference is also applicable for age  $\log t$ , extinction  $A_V$ , differential extinction  $\Delta A_V$  and distance modulus  $m - M$  inference. Recently, Cantat-Gaudin et al. (2020) use a neural network to infer  $\log t$ ,  $A_V$  and  $m - M$  for around 2000 OCs. In their work, a training dataset based on simulated OCs alone is not found to be sufficiently accurate to train a neural network. While simulated data were found to be accurate enough for the CMD classifier in Sect. 4, parameter inference is more challenging, as a network must learn to infer multiple parameters from a CMD alone and generalise this accurately to real data. However, our approach has a number of differences to theirs: firstly, we use a convolutional neural network, which may be better able to capture structure in CMDs due to its 2D approach, which may also reduce training data overfitting; secondly, our network is approximately Bayesian, and includes uncertainty estimates that quantify when it may have failed; finally, although Cantat-Gaudin et al. (2020) do not elaborate on how they simulate clusters in their work, our methodology is different and may produce different results. Hence, despite recent literature suggesting that using purely simulated data is not possible for parameter inference with CMDs, it is still worth attempting, as training on simulated data is attractive for reasons discussed in Sect. 4.

To create a parameter inference network, we used a similar network structure to that of Sect. 4.4, except with some tweaks to the network output to infer parameters. To better predict the aleatoric uncertainty of network output for this multiple-parameter network, network output was changed to a beta distribution for each parameter. These distributions can take any shape

from a uniform (completely uncertain) distribution to a single point-like estimate. The output was then scaled to be within the minimum and maximum ranges of the training data. To train the network, 50 000 simulated clusters were created using the same methodology as in Sect. 4.1, changing the distribution of cluster extinctions  $A_V$  (as defined in Table 1) to simply be uniform between 0 and 7.

In initial comparisons with literature results, differential reddening was found to strongly correlate with disagreements in extinction (and to a lesser extent, age) between this work and others. A primary cause of this is that while many works (e.g. Cantat-Gaudin et al. 2020; Bossini et al. 2019) use the so-called ‘blue edge’ of a CMD for isochrone fitting, meaning that  $\Delta A_V$  is only positive. This contrasts to SPISEA’s default  $\Delta A_V$  model, which is Gaussian – with cluster stars having both positive and negative  $\Delta A_V$  values.

However, changing SPISEA’s  $\Delta A_V$  model to also only be positive (and hence defining  $\Delta A_V$  in terms of the blue edge of cluster CMDs) was not found to be helpful. Owing to HDB-SCAN’s high sensitivity, we detect a higher number of stars outside of the core of clusters than in the membership lists of Cantat-Gaudin & Anders (2020), which are constructed with the UPMASK algorithm (Krone-Martins & Moitinho 2014) and for many clusters only select stars in the core. This means that our CMDs are constructed from clusters with significantly larger angular extents on the sky and are hence often more strongly differentially reddened than in Cantat-Gaudin & Anders (2020), with many clusters having a blue edge at an extinction value up to 1 magnitude lower than in Cantat-Gaudin & Anders (2020). For instance, NGC 884 is an example of this, with our membership list being larger and more strongly differentially reddened. A blue-edge based definition of  $A_V$  means that different works

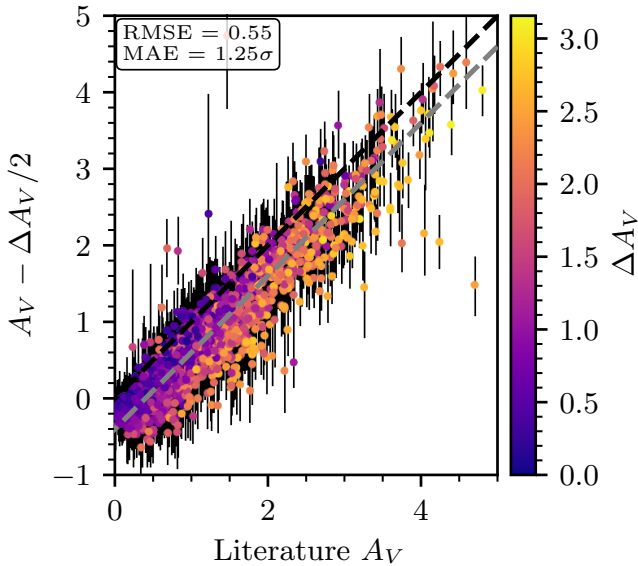


**Fig. 5.** Photometric parameters derived in this work compared against test datasets. *Top row:* 2D histograms showing the performance of the trained photometric parameter inference network on all 10 000 clusters from the validation dataset. The mean output uncertainty is shown with white error bars. As indicated by the dashed lines, predicted values on the y axis should be equal to true values on the x axis. The root mean square error (RMSE) and mean absolute error in terms of output network uncertainty (MAE) are given in the top left. All plots and the RMSE are in units of magnitude other than on age plots which are logarithms of cluster age in years. *Other rows:* comparison between network predicted parameters and ages, extinctions, and distance moduli for 247, 1753, and 1206 clusters in common with the catalogues of Bossini et al. (2019), Cantat-Gaudin et al. (2020), and Kharchenko et al. (2013) respectively. Points are shaded based on the differential extinction we infer for each cluster.

produce different values of  $A_V$  depending on how sensitive their membership recovery process is.

Instead, we continue using the default SPISEA  $\Delta A_V$  definition centred on the mean cluster  $A_V$ , but while also using the network to infer  $\Delta A_V$  for every cluster, which can then be used

as a correction to convert between extinctions in this work and others that use a blue-edge definition. In practice,  $\Delta A_V$  is very difficult to measure, as it is degenerate with other effects that broaden cluster CMDs, including unresolved binary stars and outliers. Against validation and test data, our median  $\Delta A_V$  values



**Fig. 6.** Extinction values from Cantat-Gaudin et al. (2020) compared against this work when corrected for differential extinction with an estimate of cluster differential extinction, plotted in the same style as Fig. 5. The dashed black line shows where  $y$  values equal  $x$  ones; the dashed grey line shows the same but offset by  $-0.4$ .

are found to be offset by around  $0.4$  due to unresolved binaries. Nevertheless, this parameter is helpful to aid comparisons with literature works.

Finally, we also updated our  $\Delta A_V$  model from the Gaussian default model in SPISEA to instead use the differential reddening as would be expected from stars sampled from a King profile (King 1962), assuming a first order (linear) gradient in differential extinction across a cluster. This model is narrower than the Gaussian model while retaining highly differentially reddened stars (which would be at the outskirts of a cluster), and was found to slightly improve  $\Delta A_V$  inference. This model depends on two parameters: the total differential extinction across a cluster, which was matched to have the same range as the previous Gaussian model at a  $3\sigma$  level; and the ratio between core and tidal radius, which was set to the median value for open clusters from Kharchenko et al. (2013).

Against our validation dataset of 10 000 simulated clusters, the network performs well with no clear systematics in  $\log t$ ,  $A_V$  or  $m - M$ . However, owing to the degeneracy between  $\Delta A_V$  and other effects such as unresolved binary stars, outliers, and photometric uncertainties, values of  $\Delta A_V$  smaller than  $0.4$  are not typically correctly predicted, although the true value is typically still within  $1\sigma$  uncertainty of the predicted value. These results are plotted on the top row of Fig. 5.

Using the best trained network after a number of experiments, all clusters in our catalogue closer than a maximum distance of 15 kpc have ages, extinctions, differential extinctions, and distance moduli listed in Table 4. These parameters are based on 1000 samples of the network for each cluster.

## 5.2. Comparison with other works

We briefly compare our photometric parameters to other works in the literature. Firstly, Fig. 7 shows example predicted isochrones for four OCs in this work. In the first case, NGC 2910 is a cluster with a well-behaved isochrone where all works agree

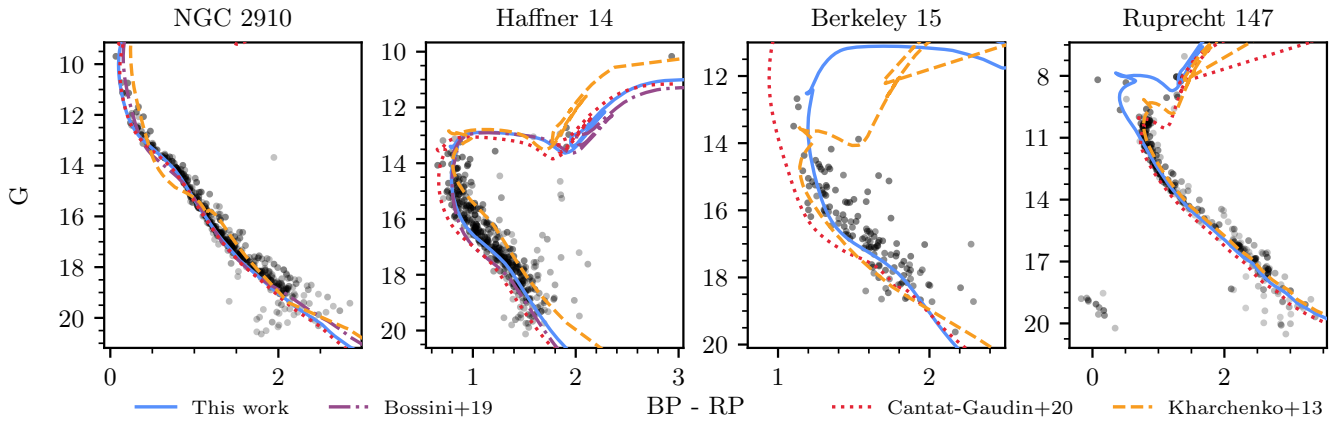
relatively well. On the other hand, Haffner 14 shows relatively strong differential reddening, and different definitions of differential reddening between different works cause isochrone fits to disagree. Berkeley 15 is a sparse cluster where both differential reddening and field star outliers affect different works in different ways, with our updated *Gaia* DR3 membership list having fewer outliers than that of Cantat-Gaudin et al. (2018). Ruprecht 147 is a nearby and particularly old cluster ( $\sim 1$  Gyr), where blue straggler stars systematically affected our network and caused an incorrect younger age value to be predicted for this cluster. It is clear from these plots that for all but the most well-behaved OCs, different works can have different photometric parameters.

Fig. 5 compares all network predictions with values from four test datasets. An advantage of our simulated training approach is that network predictions can now be compared to other literature works, which act as independent test datasets which can verify the accuracy of our network. It is important to note that our results never agree perfectly, however, particularly since all works we compare to are based on *Gaia* DR2 or pre-*Gaia* OC membership lists that may be significantly less clean or have significantly fewer stars than our *Gaia* DR3 membership lists.

Bossini et al. (2019) provide a catalogue of precise OC parameters from Bayesian isochrone fitting using the BASE-9 algorithm (von Hippel et al. 2006). A key difference is that their work uses metallicity estimates from the literature where available, whereas our approach is based entirely on *Gaia* DR3 parameters and assumes a given cluster can have any metallicity as drawn from a broad probability distribution based on literature values (Table 1). Nevertheless, our results still agree well with theirs in  $\log t$ ,  $A_V$  and  $m - M$ . In cases where our  $\log t$  estimates disagree most strongly, this is typically due to differences in OC membership list. There is however a possible minor systematic between our two works for OCs with extinctions below  $0.6$ , many of which we infer smaller extinctions for than them; this may be as a result of  $A_V$  vs. metallicity degeneracies. However, their values are typically only  $1$  to  $2\sigma$  from ours.

Our parameters agree less strongly with the results of Cantat-Gaudin et al. (2020), which are derived from a neural network trained on isochrone fits from a variety of works (including Bossini et al. 2019). This is to be expected to some extent, as while Bossini et al. (2019) only fit isochrones to a subset of OCs with clean membership lists and the least differential reddening, Cantat-Gaudin et al. (2020) fit isochrones to all known OCs at the time, including many sparse objects which may now have significantly different membership lists in our current *Gaia* DR3 work. However, some differences persist. A clear systematic in our and their  $A_V$  values is clear, although this is likely due to their different blue edge definition of extinction (whereas our network fits to the mean extinction in a cluster.) Figure 6 shows a crude conversion between our  $A_V$  values and their blue-edge  $A_V$  values. While this removes the systematic difference in gradient, our converted  $A_V$  values are still generally smaller than theirs by around  $0.4$  to  $0.5$  on average. This is likely due to two effects; firstly, as shown by the results on validation data,  $\Delta A_V$  is generally overestimated for our validation data by around  $\sim 0.4$  due to degeneracies with unresolved binary stars, outlier non-member stars, and photometric uncertainties, which may explain some of this discrepancy, particularly for clusters with lower  $\Delta A_V$  values. Secondly, our membership lists generally cover a wider extent on the sky than those used in Cantat-Gaudin et al. (2020), meaning that our clusters are often larger and hence are more extremely differentially reddened between separate sides of the cluster; hence, a conversion between the works based on our  $\Delta A_V$  values is likely to frequently over-correct for the difference





**Fig. 7.** Predicted cluster isochrones from this work (solid blue line) compared with those from other works. Cluster members are plotted in black and shaded according to their membership probability.

in  $A_V$  definition. Finally, some of our ages for the oldest clusters ( $\log t > 9$ ) appear systematically younger, on average by around  $2\sigma$ ; in some cases, this may be due to our fits being disrupted by blue straggler stars (Fig. 7, see Ruprecht 147.) The training data we use for our photometric parameter inference are adapted from our CMD classifier in Sect. 4, for which blue straggler stars were not found to have a negative impact on the accuracy of our network and were hence not included. Future works using purely simulated data to train a photometric parameter inference neural network would benefit from inclusion of blue straggler stars in their training data, although in practice the origin of blue stragglers is still disputed, and these stars may hence be challenging to simulate accurate photometry for (Boffin et al. 2015; Cantat-Gaudin 2022).

Finally, our results have limited agreement with those of Kharchenko et al. (2013). While some clusters have similar values between their work and ours, particularly for  $A_V$  and particularly for the largest and most clearly defined clusters (Fig. 7), many sparse clusters that were difficult to detect before *Gaia* have very different photometric parameters. This typically appears to be caused by extremely different cluster membership lists. Before *Gaia*, OCs were often challenging to separate from field stars (Cantat-Gaudin 2022), requiring that suspected outliers be removed iteratively to improve CMD quality (Kharchenko et al. 2012). However, this process can also remove true cluster members, which can cause resulting cluster membership lists to be incorrect (Cantat-Gaudin & Anders 2020). This discrepancy with the results of Kharchenko et al. (2013) is also reported by Cantat-Gaudin et al. (2020), who also find that many photometric parameters derived before *Gaia* are strongly discrepant with current results. In addition, while the number of member stars reported in Kharchenko et al. (2013) is generally a poor predictor for whether or not a given cluster in their work has very different parameters to ours, there are some cases (such as clusters in their work with  $A_V > 5$  that we derive much smaller values for) where the most discrepant clusters were also the smallest, with fewer than 20 member stars in reported in Kharchenko et al. (2013).

Although approximate, these results still agree well within the sample-limited but accurate Bayesian isochrone fits of Bossini et al. (2019) and agree relatively well (albeit with some caveats) with the machine learning derived parameters of Cantat-Gaudin & Anders (2020). This work offers a large and homogeneously derived catalogue of photometric parameters with suffi-

cient accuracy for basic analysis. In the next section, we use the ages and extinctions we derived here to aid with discussion of our cluster sample.

## 6. Crossmatch to existing catalogues

### 6.1. Crossmatch strategy

Before conducting further analysis on the cluster catalogue, such as restricting it to only clusters with reliable colour-magnitude diagrams or removing moving groups, it is helpful to crossmatch our results to literature catalogues to allow for easier comparisons between derived parameters and other works. In particular, this makes it possible to compare whether clusters reported in other works are compatible with real open clusters given further parameters derived in Sect. 4 and the third paper in this series, Hunt & Reffert, *in prep.*, where we will derive dynamical parameters for our census of star clusters.

In Paper 1, we crossmatched by assigning matches to clusters when their mean positions were compatible to within their tidal radii and when their mean proper motions and parallaxes were compatible within five standard errors. In initial testing, the crossmatch strategy of Paper 1 was found to be insufficient for two reasons when comparing between *Gaia* DR3 astrometry and *Gaia* DR2 astrometry, in addition to a further issue with the positional strategy used.

Firstly, the standard errors on mean proper motions and parallaxes in *Gaia* DR2 can be as small as 5 to 10  $\mu\text{as}$  for the largest clusters in catalogues such as Cantat-Gaudin & Anders (2020), although this is smaller than estimated upper limits on systematics in *Gaia* DR2 of 50  $\mu\text{as}$  (Lindgren et al. 2018). Many reliable clusters are hence missed when treating DR2 positions exactly, as they have systematics significantly larger than their standard errors, with positions in DR3 that can deviate systematically from their DR2 positions by 50  $\mu\text{as}$  or more.

Secondly, membership lists can differ between works and can be significantly different for the same cluster – for instance, works such as Castro-Ginard et al. (2020) only used stars down to  $G = 17$ , whereas this work often has membership lists down to  $G \sim 20$ . Many clusters hence have significantly different membership lists that can result in different mean parameters, particularly for asymmetric clusters.

Our positional crossmatch strategy was also revised and improved. Paper 1 used a conservative strategy for matching on position, which assumed that a cluster is a positional match if the

centre of the literature cluster is closer than either the Paper 1 or literature radius for a given cluster. However, in practice, this strategy appears almost always too conservative, as many distant, compact clusters reported in catalogues such as Froebrich et al. (2007) would match to large, nearby clusters that happen to contain the distant object within one radius, despite the cluster centres being strongly incompatible given the smaller (literature) radius.

To improve positional crossmatching, we instead define a positional match to require that the centre of the literature cluster is closer than both the current and literature radius, which in almost all cases still recovers reliable matches but while not erroneously matching to compact, distant objects with significantly different sizes and cluster centres. Then, for catalogues with *Gaia* astrometry available, we also match on proper motions and parallaxes, requiring that the new mean proper motion and parallax are within two standard deviations of the literature value (with both current and literature standard deviations summed in quadrature.) This approach with standard deviations matches clusters if a new cluster is within allowed ranges of the dispersion of the current and literature entries, with the principles that exact statistical matching based on standard errors is not possible as unknown systematic errors dominate, and that a cluster within the dispersion of a literature entry is likely to be the same object. Using a higher maximum value of the dispersion was not found to significantly increase the number of literature clusters recovered by more than 1%, but while adding many false crossmatches to other nearby objects that greatly worsen the reliability of the overall crossmatching process.

Some special cases are also worth mentioning: the catalogue of Kharchenko et al. (2013) is based on PPMXL proper motions and distances from isochrone fitting by hand, which are generally significantly less accurate than *Gaia* astrometry. Hence, we crossmatch to Kharchenko et al. (2013) with both a position-only and a second positions, proper motions, and distances crossmatch which can more strongly confirm the most reliable matches. Some catalogues list only a radius containing 50% of members for entries (e.g. Cantat-Gaudin & Anders 2020); for these catalogues, we use twice this radius to approximate the total size of the cluster. Other works (e.g. Castro-Ginard et al. 2020; He et al. 2022a) list only standard deviations of the mean position; for these catalogues, we use twice the geometric mean of this standard deviation on position to approximate the total size of the cluster. Finally, Kounkel et al. (2020) does not list uncertainties or dispersions on mean parameters, and so these were manually recalculated with our own pipeline using their lists of members.

After an extensive search of the literature for recent catalogues, excluding works already listed entirely in other catalogues (such as Froebrich et al. (2007), which appears in its complete form within Bica et al. (2018)), we crossmatch against 26 different works listed in Table 3. In addition, as our catalogue contains many moving groups, globular clusters, and a handful of clusters associated with the Magellanic clouds, we also crossmatch against the Kounkel et al. (2020) catalogue of predominantly moving groups, the Vasiliev & Baumgardt (2021) *Gaia* DR3 catalogue of globular clusters and the Bica et al. (2008) catalogue of star clusters in the Magellanic clouds. Names between catalogues were standardised as much as possible to facilitate easier comparison and remove duplicated clusters. One such example are ESO clusters, which are numbered based on their position in the form ‘ESO XXX-XX’ in the original work and Kharchenko et al. (2013), but with numbers that are separated by a space instead of a dash in Cantat-Gaudin & Anders

**Table 3.** Results of crossmatching against literature catalogues sorted by  $n_{\text{clusters}}$ .

Work	$n_{\text{clusters}}$	$n_{\text{detected}}$	%
Bica et al. (2018)	4391	1251	28.5
Kharchenko et al. (2013)	2935	1513	51.6
Dias et al. (2002)	2161	1160	53.7
He et al. (2022b)	1656	737	44.5
Cantat-Gaudin & Anders (2020)	1481	1431	96.6
Hao et al. (2022)	704	501	71.2
Castro-Ginard et al. (2022)	628	558	88.9
Castro-Ginard et al. (2020)	582	519	89.2
He et al. (2022a)	541	440	81.3
He et al. (2022c)	270	122	45.2
Sim et al. (2019)	208	180	86.5
Qin et al. (2023)	101	74	73.3
Chi et al. (2023)	82	18	22.0
Liu & Pang (2019) <sup>a</sup>	76	57	75.0
He et al. (2021) <sup>b</sup>	74	69	93.2
Li et al. (2022)	64	44	72.1
Chi et al. (2022) <sup>b</sup>	46	11	23.9
Hunt & Reffert (2021)	41	41	100.0
Li & Mao (2023)	35	0	0.0
Ferreira et al. (2021)	34	32	94.1
Ferreira et al. (2020)	25	25	100.0
Casado (2021)	20	15	75.0
Hao et al. (2020) <sup>b</sup>	16	5	31.3
Jaehnig et al. (2021)	11	7	63.6
Santos-Silva et al. (2021)	5	4	80.0
Qin et al. (2021) <sup>b</sup>	4	4	100.0
Ferreira et al. (2019)	3	0	0.0
Casado & Hendy (2023)	2	2	100.0
Anders et al. (2022)	1	1	100.0
Bastian (2019)	1	1	100.0
Tian (2020)	1	1	100.0
Zari et al. (2018) <sup>b</sup>	1	1	100.0
Kounkel et al. (2020) <sup>c</sup>	8281	1498	18.1%
Bica et al. (2008) <sup>d</sup>	3740	22	0.6%
Vasiliev & Baumgardt (2021) <sup>e</sup>	170	134	78.8%

**Notes.** 32 catalogues of OCs are listed in the first section of the table, in addition to three catalogues at the bottom of other star clusters. <sup>(a)</sup> Original work and this work uses the acronym ‘FoF’ to name clusters, although others list with acronym ‘LP’. <sup>(b)</sup> Cluster(s) in these works were unnamed, and so cluster acronyms were adopted based on first letters of surnames of authors. <sup>(c)</sup> Catalogue of predominantly moving groups, although many are also open clusters. <sup>(d)</sup> Position-only catalogue of objects in the Magellanic clouds. <sup>(e)</sup> Catalogue of globular clusters.

(2020) and Dias et al. (2002), or often miss leading zeroes in Bica et al. (2018).

## 6.2. Recovery of clusters from prior works

Table 3 shows that this work has a high recovery rate of OCs from other works. As shown in Table 3, we recover 96.6% of clusters from Cantat-Gaudin & Anders (2020), higher than the 86.4% of clusters recovered in Paper 1. Generally, clusters not recovered in Paper 1 were sparse, barely-visible overdensities in *Gaia* DR2 which often now stand out strongly in *Gaia* DR3, including clusters such as Berkeley 91 and Auner 1, which we now detect reliably at S/Ns of  $9.7\sigma$  and  $12.5\sigma$  respectively. The fact that only Cantat-Gaudin & Anders (2020) was able to detect these clusters in DR2 is likely due to a difference in methodology



– by starting with prior cluster positions, their search regions for these clusters are smaller and may help the clusters to stand out. However, the disadvantage of such an approach is that it may also introduce a handful of false positives, due to poor statistics inherent in such small search regions – in Paper 1, we comment that a handful of clusters in Cantat-Gaudin et al. (2020) may not exist, which may be the case for some of the 3.4% of clusters we are still not able to recover in *Gaia* DR3 despite the greatly improved astrometry and clear benefits to the S/N of other previously undetected clusters.

We recover most of the new clusters reported in Castro-Ginard et al. (2020) (a work based on *Gaia* DR2) and Castro-Ginard et al. (2022) (a work based on *Gaia* EDR3), recovering almost exactly 89% of both catalogues, showing that a majority of these objects can be confirmed independently. The reason for the non-recovery of around 11% of clusters in both cases is not clear, although the fact that this amount is similar between both clusters detected with *Gaia* DR2 and EDR3 suggests that it is a fundamental methodological difference (their works use the DBSCAN algorithm, see Paper 1 for a review) rather than a data one.

However, we recover fewer of the new clusters reported by other DBSCAN-based works such as Hao et al. (2020, 2022) and He et al. (2021, 2022c,a,b), recovering fewer than 50% of the clusters reported in He et al. (2022c,b) using *Gaia* EDR3 data.

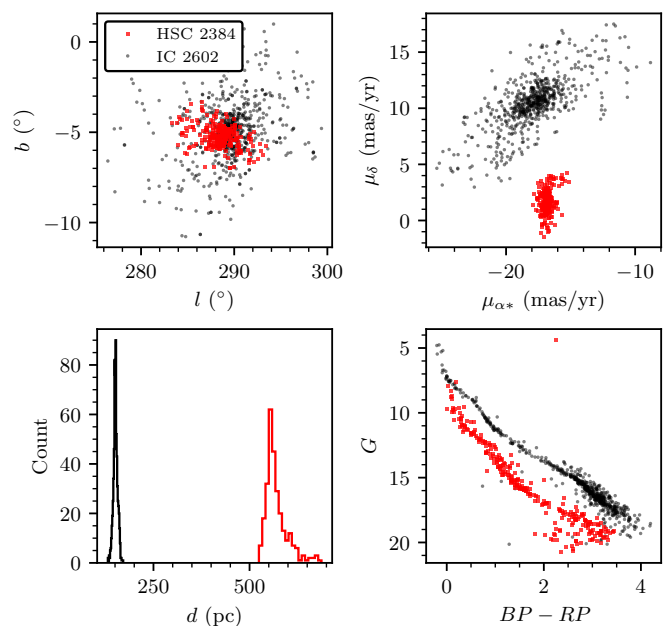
Additionally, while a large fraction of clusters reported before *Gaia* and catalogued in works such as Dias et al. (2002), Kharchenko et al. (2013), and Bica et al. (2018) still do not appear in *Gaia* DR3, we are able to reliably detect an additional 277 clusters from Dias et al. (2002), 292 clusters from Kharchenko et al. (2013), and 127 clusters from Bica et al. (2018) that do not appear in the *Gaia* DR2 catalogue of Cantat-Gaudin & Anders (2020) (excluding GCs in all cases, as the catalogue of Cantat-Gaudin & Anders (2020) does not contain them.)

Notably, we are unable to detect any of the high galactic latitude OCs that have been reported recently in Li & Mao (2023), despite the fact that OCs at such high latitudes should stand out clearly against the low number of field stars in the galactic halo. This echoes the results of Cantat-Gaudin et al. (2018) and Cantat-Gaudin & Anders (2020), who also find that high latitude OCs that have been reported in works such as Schmeja et al. (2014) are undetectable in *Gaia* data.

We discuss possible reasons for the non-detection of many literature OCs further in Sect. 8.

Finally, it is worth commenting on our detections of moving groups, globular clusters, and Magellanic cloud objects. We are only able to detect 18.1% of moving groups and clusters from the catalogue of Kounkel et al. (2020), despite this work using the same algorithm (HDBSCAN). Many of the groups reported in Kounkel et al. (2020) have large on-sky extents that are larger than the fields used in this work. However, although 2276 of their 8281 clusters are compact enough to be easily detectable in our fields, we only recover 622 (27.3%) of these compact groups, many of which correspond anyway to known nearby OCs. In Paper 1, we found that while HDBSCAN is the most sensitive clustering algorithm for application to *Gaia* data, it also reports a large number of false positives without additional postprocessing to remove clusters based on their statistical significance. It may be that these clusters are false positives, although this should be investigated further in detail (see e.g. Zucker et al. 2022).

The recovery of a large fraction of GCs in Vasiliev & Baumgardt (2021) shows that HDBSCAN can be used to effectively recover GCs. The non-recovered objects are mostly distant and



**Fig. 8.** Member stars for the candidate new cluster HSC 2384 (red squares) compared against the nearby cluster IC 2602 (black circles). Four plots of are shown, comparing positions (top left), proper motions (top right) and photometry (bottom right). The bottom left plot shows a histogram of all distances to individual member stars.

heavily reddened GCs whose member stars can only be recovered with a prior position and distance to narrow the search region. Finally, while not a focus of this work, the recovery of 22 Magellanic cloud star clusters from Bica et al. (2008) shows that *Gaia* data could be used to make limited inferences on existing Magellanic cloud clusters in a future work, although we do not appear to detect any new clusters in the Magellanic clouds as their distance is too high.

### 6.3. Assignment of names

As many of the objects we detect crossmatch to multiple entries in the literature (or vice-versa), assigning detected clusters to literature names can be non-trivial. A total of 7022 literature clusters crossmatch to 4944 of the entries in our catalogue, of which only 2749 matches are direct one-to-one matches where a single detected cluster can be easily assigned a single name.

1396 detected clusters each match to multiple literature entries. In these cases, the main cluster name was assigned based on the date of submission to a journal, with other names recorded in a separate column of alternative names for this object.

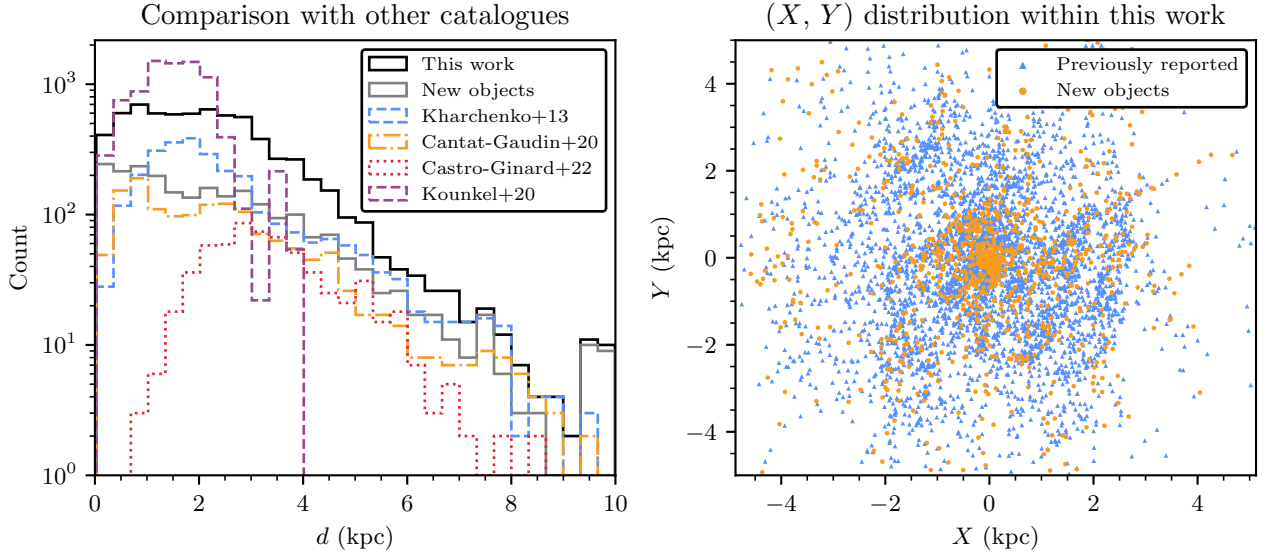
In 64 cases, multiple detected clusters crossmatched to the same literature object. The best match was selected based on position (or proper motions and distances, if available), with other objects instead recorded as new clusters.

Finally, there were 265 groups of crossmatches where multiple detected clusters crossmatched to multiple literature clusters, where assigning one match affects other matches. This is common in regions where many clusters are in a small area, such as in star formation regions like the Carina nebula. For simplicity, and since many of these groups contain literature entries with only positions available, we assign the best match on cluster positions only, iterating over all matches within a group accepting the match with the smallest positional separation and then removing all other literature entries with the same name within

**Table 4.** Mean parameters for the clusters detected in this study.

Name	ID <sup>a</sup>	S/N	$n_{\text{stars}}$	$\alpha$ (°)	$\delta$ (°)	$r_{50}$ (°)	$\mu_{\alpha^*}$ (mas yr <sup>-1</sup> )	$\mu_{\delta}$ (mas yr <sup>-1</sup> )	$\varpi$ (mas)	$\log t$
HSC 1	1805	8.21	64	289.61	-38.03	3.32	-1.029 (0.054)	-8.941 (0.085)	2.097 (0.006)	7.87 <sup>+0.24</sup> <sub>-0.27</sub>
HSC 2	1806	3.79	16	268.63	-29.53	0.13	1.680 (0.031)	-1.182 (0.032)	0.634 (0.003)	7.92 <sup>+0.24</sup> <sub>-0.22</sub>
HSC 3	1807	3.89	24	273.73	-31.87	0.12	0.371 (0.019)	0.210 (0.025)	0.647 (0.005)	8.75 <sup>+0.18</sup> <sub>-0.20</sub>
HSC 4	1808	3.32	17	269.07	-29.64	0.02	2.125 (0.067)	-11.895 (0.060)	0.112 (0.015)	7.54 <sup>+0.45</sup> <sub>-0.50</sub>
HSC 5	1809	4.38	18	276.78	-33.09	0.12	0.150 (0.047)	-6.676 (0.049)	0.657 (0.004)	9.70 <sup>+0.17</sup> <sub>-0.17</sub>
HSC 6	1810	4.57	21	267.71	-28.82	0.05	-0.292 (0.017)	-1.516 (0.023)	0.252 (0.004)	7.84 <sup>+0.27</sup> <sub>-0.32</sub>
HSC 7	1811	3.12	18	261.40	-25.13	0.09	-5.033 (0.061)	-0.983 (0.060)	0.464 (0.005)	9.68 <sup>+0.15</sup> <sub>-0.23</sub>
HSC 8	1812	3.33	28	267.67	-28.63	0.06	0.207 (0.014)	-0.211 (0.026)	0.340 (0.004)	7.86 <sup>+0.23</sup> <sub>-0.22</sub>
HSC 9	1813	5.88	25	269.05	-29.33	0.16	2.120 (0.020)	-0.289 (0.021)	0.549 (0.005)	7.61 <sup>+0.22</sup> <sub>-0.19</sub>
HSC 10	1814	4.56	12	268.23	-28.80	0.06	-0.200 (0.011)	-1.753 (0.013)	0.351 (0.003)	8.20 <sup>+0.30</sup> <sub>-0.31</sub>

**Notes.** Standard errors for mean proper motions and parallaxes are shown in the brackets. The full version of this table with 7200 rows and many extra columns is available at the CDS only, with a complete description of the included additional data in Appendix A. <sup>(a)</sup> Internal designation used to link final catalogue entries to their crossmatching results in Table B.1.



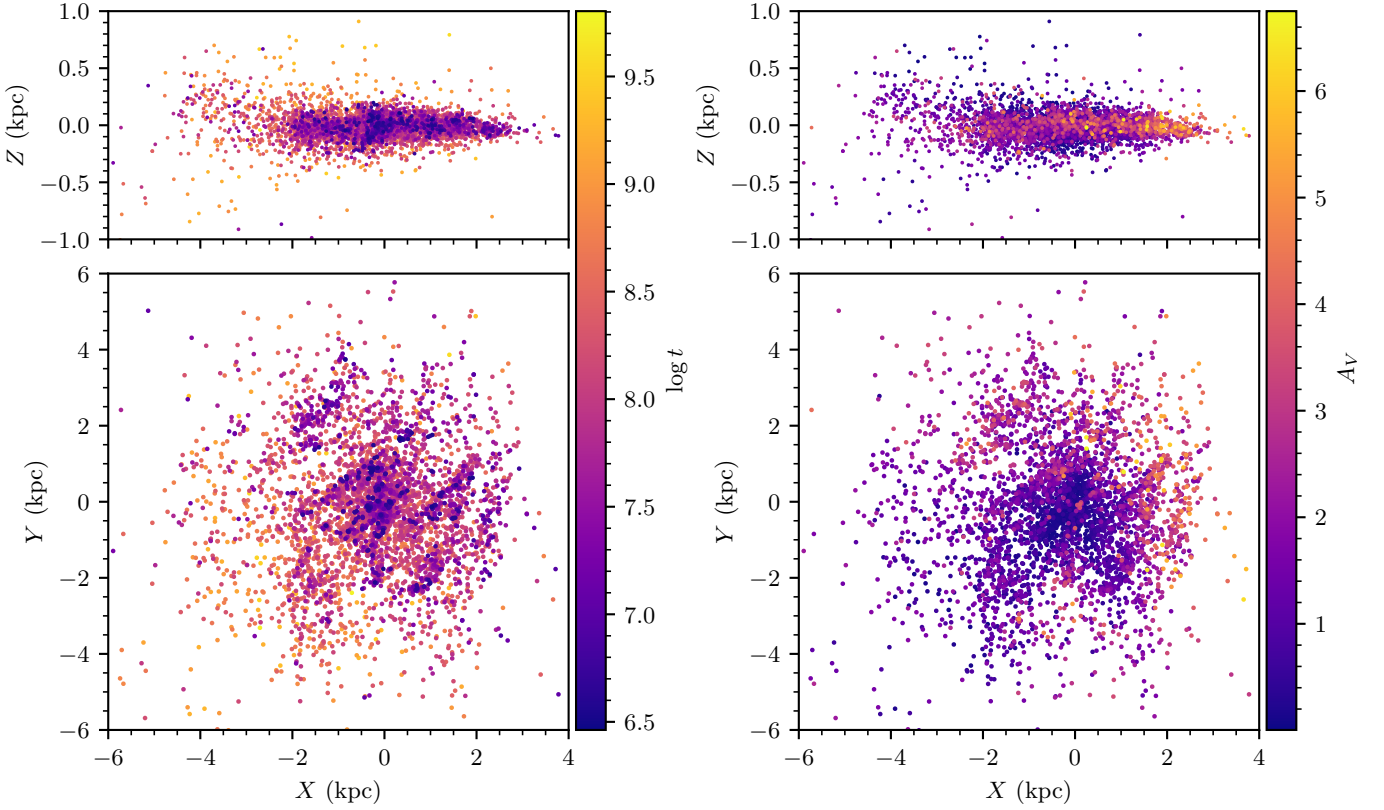
**Fig. 9.** Distance and spatial distributions of clusters in this work. *Left:* the distance distribution of all clusters in this work that do not crossmatch to known GCs compared to other catalogues. *Right:* The distribution of clusters in this work in Cartesian coordinates centred on the Sun, cut to only those within 5 kpc in the X or Y directions. All previously reported clusters that we redetect are shown as blue triangles, and all objects new in this work shown as orange circles.

this group. All valid matches for every cluster are recorded in a separate column, and as these crossmatches represent the most difficult to assign reliably, clusters where their name has been assigned in this way are flagged in the catalogue as crossmatches that were particularly difficult to assign.

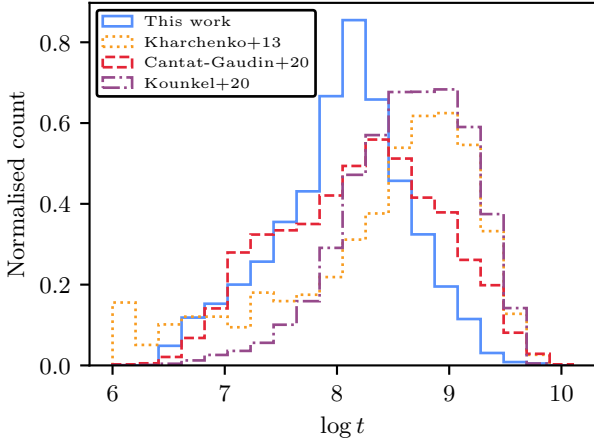
After assigning names to clusters, removing 22 objects associated with the Magellanic clouds, and removing 566 objects clearly associated with stellar streams in the galactic halo, our catalogue contains 7200 clusters, and is listed in Table 4 and online at the CDS, with tables of member stars and the rejected Magellanic cloud objects and stellar streams available online only. 2420 of these clusters are unreported in the literature and are candidate new objects, which we label with the acronym ‘HSC’ (standing for HDBSCAN Star Cluster.) Most of these objects have good-quality CMDs, and some are likely to be new OCs. For instance, HSC 2384 is a nearby new OC candidate at a distance of only 551 pc with 273 member stars and a high astro-

metric S/N of  $23.6\sigma$ , which likely avoided prior detection due to being obscured by IC 2602 and mis-crossmatched to it (shown in Fig. 8.) However, many appear to be more consistent with unbound moving groups, and will require further classification based on their structure and dynamics. In addition, we provide a table of all crossmatches and non-crossmatches against the clusters in this work in Table B.1.

In the next sections, we discuss multiple aspects of the overall catalogue. Firstly, we discuss the overall catalogue of existing clusters in Sect. 7, including its distribution and the quality of its membership lists. Section 8 discusses why some literature clusters are undetected. Finally, Sect. 9 discusses why existing approaches to differentiate between moving groups and OCs are inadequate to classify the new clusters detected in this work, a topic that will be explored further in a future work (Hunt & Reffert, *in prep.*).



**Fig. 10.** Spatial distributions of clusters detected in this work shaded on our derived  $\log t$  and  $A_V$  values. *Left:* side-on and top-down distribution of clusters in heliocentric coordinates that do not crossmatch to known GCs. The galactic centre is to the right, with the Sun at (0,0). Only clusters passing two quality cuts are plotted: firstly, those with a CST score above  $5\sigma$ , meaning they are highly probable astrometric overdensities; and secondly, a median CMD class above 0.5, which are those compatible with single population star clusters. Clusters are plotted in descending age order, meaning points representing young clusters are most visible in crowded regions. *Right:* as left, except clusters are colour-coded by extinction  $A_V$ . Clusters are plotted in ascending order of extinction.



**Fig. 11.** Histogram of ages of all clusters in this work with median CMD classes greater than 0.5 – specifically, all clusters with photometry that is compatible with a single population of stars. These are compared to the ages of all clusters in the catalogues of Kharchenko et al. (2013), Kounkel et al. (2020), and Cantat-Gaudin et al. (2020). Known GCs are excluded from the results of this work and the results of previous works for this plot.

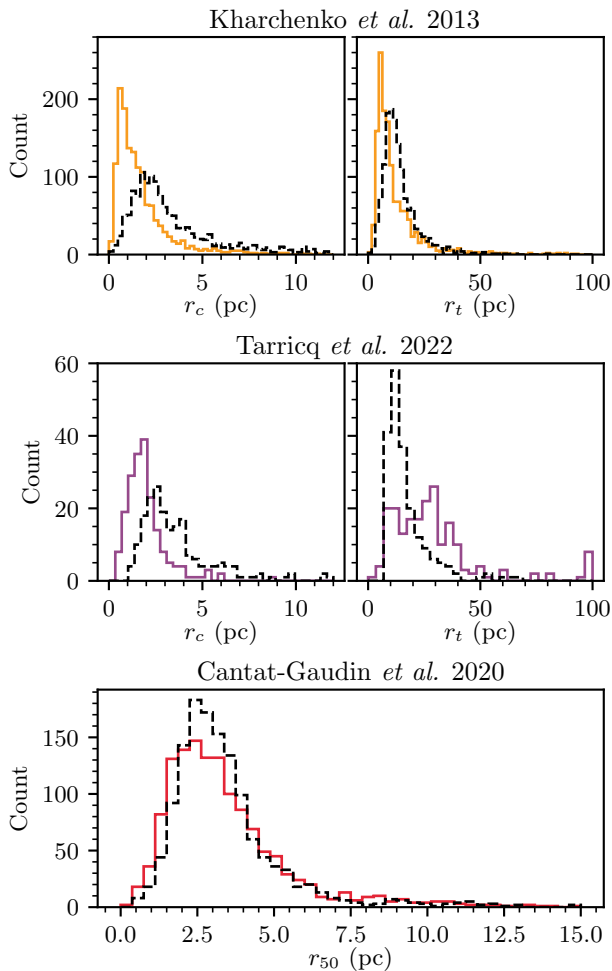
## 7. Overall results

In this section, we briefly discuss the structure and characteristics of the overall catalogue of 7200 clusters.

### 7.1. Suggested cuts on the catalogue for a high-quality cluster sample

Our catalogue also includes objects that we detect with CST scores as low as  $3\sigma$ , and objects with low-quality CMDs given the results of our classifier in Sect. 4. Such clusters are included in our catalogue for completeness, as a low-quality CMD may be caused by a poor detection of a real OC by our cluster recovery method, and a cluster with a low CST that is not a guaranteed astrometric overdensity may still be a real cluster that could be validated by a future *Gaia* data release. However, these clusters are not particularly scientifically useful for studies of star clusters, as they cannot be validated as real within this work, or even with any currently available data.

Hence, in discussions of the overall structure of our results, we predominantly discuss the most reliable sample of 4114 clusters within the catalogue: those with a median CMD class greater than 0.5, meaning that they are likely to be a largely homogeneous single population of stars as in OCs and moving groups, allowing some tolerance for blue stragglers and extended main-sequence turnoffs; and a CST of greater than  $5\sigma$ , corresponding to clusters with a high likelihood of being real overdensities within *Gaia* data and not simply a statistical fluctuation. The



**Fig. 12.** Cluster radii derived in this work (dashed black line) compared against the distributions of cluster radii in various literature works. *Top row:*  $r_c$  (top left) and  $r_t$  (top right) of 1446 clusters from Kharchenko et al. (2013) that we re-detect in this work (solid orange curve) compared against our approximately estimated King (1962) radii for these 1446 clusters. *Middle row:* same as top, except for radii of 202 clusters from Tarricq et al. (2022) that have derived King radii (solid purple curve). *Bottom:*  $r_{50}$  measurements from Cantat-Gaudin & Anders (2020) compared against our  $r_{50}$  measurements for the 1343 clusters from their work that we re-detect.

more tenuous 3080 objects excluded by this cut may still be used in some analyses, although with the caveat that these objects are less likely to be real star clusters.

## 7.2. General distribution

The distribution of clusters in our catalogue is generally similar to that of other *Gaia*-based works such as Cantat-Gaudin & Anders (2020), albeit with more stark differences when compared to those compiled before *Gaia*, such as Kharchenko et al. (2013). Comparisons are also useful to the catalogue of structures, moving groups, and star clusters of Kounkel et al. (2020) and papers based on *Gaia* DR3 data that report new clusters, such as Castro-Ginard et al. (2022).

Figure 9 shows the distance distribution of clusters in this work, as well as the  $X$ ,  $Y$  distribution of clusters we re-detect and objects new to this work. Owing to the improved astrometry of *Gaia* DR3 and the clustering method we use (see Paper 1), our

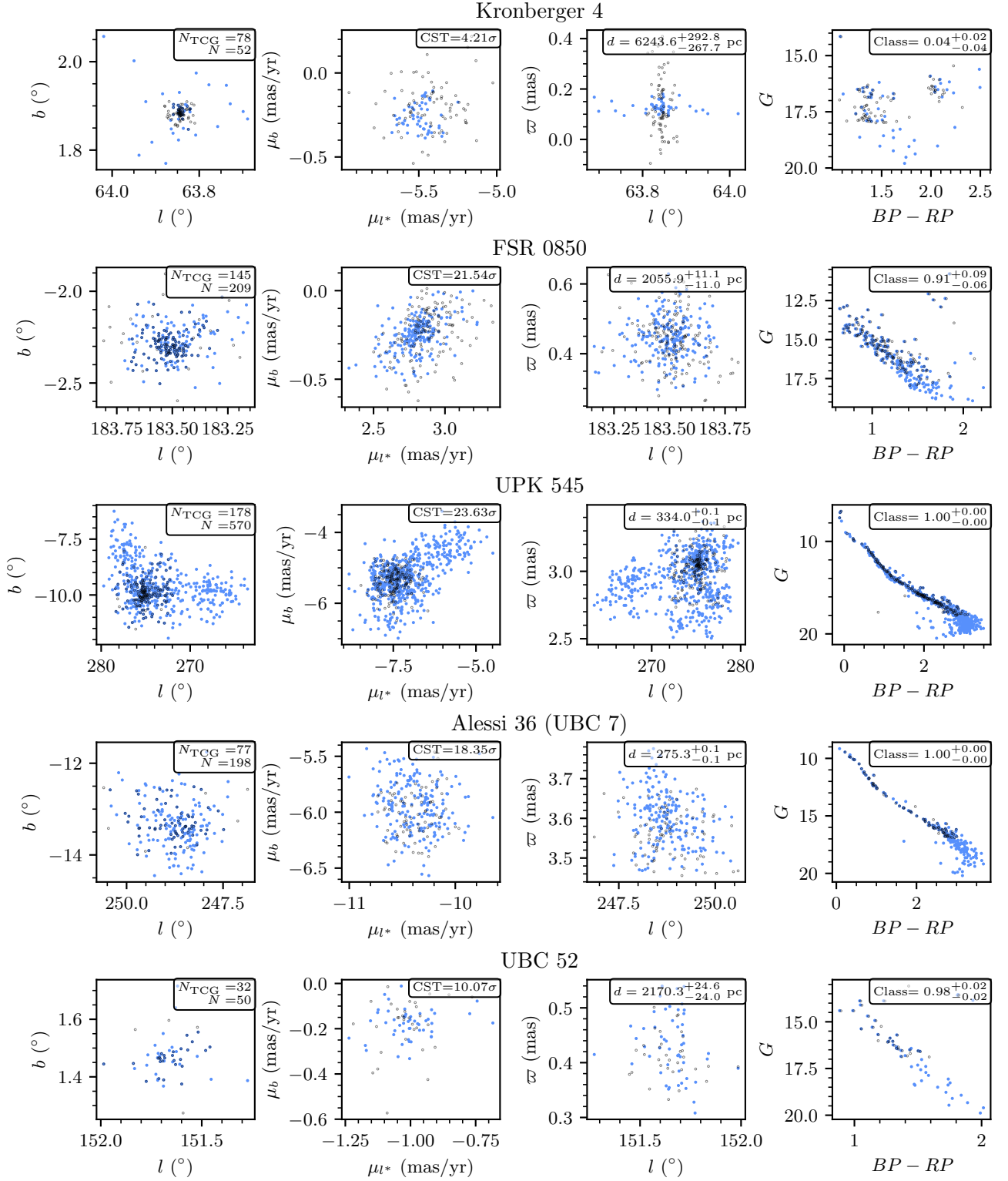
catalogue has a high total number of clusters in most distance bins relative to other catalogues. As expected from the results in Paper 1, HDBSCAN is a cluster recovery technique sensitive across all distance ranges. However, HDBSCAN is sensitive to all clusters within *Gaia* data, as it is unbiased on the shape of clusters it reports; hence, the catalogue contains a large number of moving groups, which are generally detected near to the Sun. The catalogue contains around 8x as many objects as the open cluster catalogue of Cantat-Gaudin & Anders (2020) within 500 pc, clearly visible as an overdensity of new objects and in the distance distribution of Fig. 9. These objects are often difficult to classify as being OCs or moving groups (see Sect. 9).

The age and extinction distribution of Fig. 10 is similar to that of Cantat-Gaudin et al. (2020). A number of structures stand out, including: the imprint of the galactic warp in  $X$ ,  $Z$  plots for  $X < -2$  kpc; the presence of spiral arm structure amongst young clusters very similar to that reported in works such as Castro-Ginard et al. (2021); and the general flatness of the distribution of compact star clusters in the Milky Way other than GCs, with few existing at heights of  $|Z| > 250$  pc. Additionally, clusters towards the galactic centre generally have high  $A_V$  values of 5 or greater, suggesting that extinction may be a limiting factor in the detection of clusters in this direction.

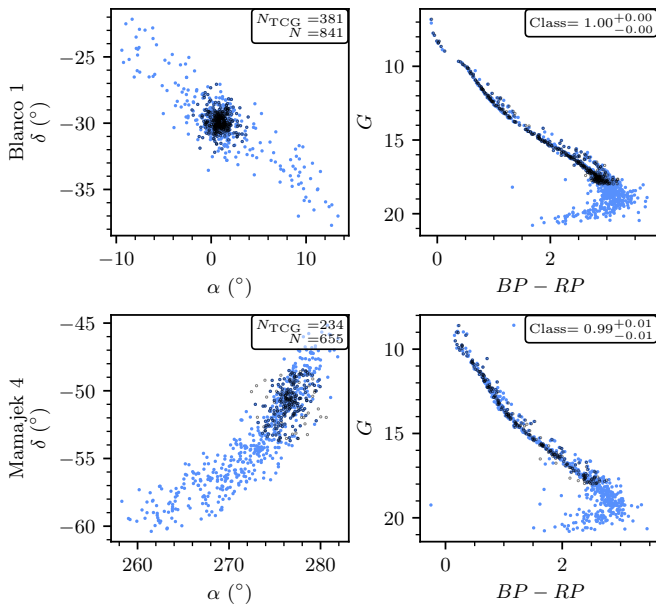
Differences to pre-*Gaia* works are most apparent in the age histogram of Fig. 11, however. Our combined age distribution is relatively similar to that of Cantat-Gaudin et al. (2020), albeit with a slightly lower median age around  $\log t \approx 8$  and no additional bump between  $7 < \log t < 8$ . However, the star cluster catalogue of Kharchenko et al. (2013) skews significantly older, with the most common (modal) age for clusters being around  $\log t \approx 9$ , an age range where we detect few clusters. A similar pattern is also visible for the catalogue of Kounkel et al. (2020), whose moving group and star cluster catalogue contains many unbound, old structures. Many of these objects have similar ages to the typical ages of unclustered stars in the Milky Way disk. In Sect. 8, we elaborate on how some of these age differences may be caused by these catalogues containing a number of old false positive clusters.

Finally, Fig. 12 shows the distribution of cluster radii compared between this work and the works of Kharchenko et al. (2013), Tarricq et al. (2022), and Cantat-Gaudin & Anders (2020). Our cluster radii agree most strongly with those in Cantat-Gaudin & Anders (2020), with a similar distribution of cluster radii containing 50% of members  $r_{50}$ . The King (1962) core radii  $r_c$  that we derive, when compared against those in Kharchenko et al. (2013) and Tarricq et al. (2022), are generally larger. This may be due to our more populated membership lists, particularly for faint stars, due to our lack of a magnitude cut in our clustering analysis. Particularly for clusters with a high degree of mass segregation, this difference in memberships would cause our clusters to have larger observed cores. Our tidal radii  $r_t$  are slightly larger than those in Kharchenko et al. (2013), but much smaller than those in Tarricq et al. (2022). In the first case, the difference may be due to the improved precision of *Gaia* data compared to pre-*Gaia* works, causing us to detect more member stars at the outskirts of clusters and hence derive larger cluster tidal radii, with this effect again being stronger for mass segregated clusters. In the second case, since Tarricq et al. (2022) also explicitly searched for cluster tidal tails and comas in their work, it may be that their extended cluster membership lists mean that they report higher cluster tidal radii.





**Fig. 13.** Membership list comparisons between this work and the catalogue of Cantat-Gaudin & Anders (2020), using three clusters selected at random (upper three) and two clusters selected at random that were detected in Castro-Ginard et al. (2018) using *Gaia* DR1 data. Stars assigned as members by this work are plotted with filled blue circles, while members reported by Cantat-Gaudin & Anders (2020) are plotted with empty black circles. The first three columns compare the astrometry of cluster members in galactic coordinates, proper motions, and parallax as a function of  $l$ . The final column compares colour-magnitude diagrams of each resulting membership list. For every cluster, various parameters are labelled on the plots: number of member stars in Cantat-Gaudin & Anders (2020)  $N_{\text{TCG}}$ , number of member stars in this work  $N$ , astrometric S/N as estimated by the CST, distance  $d$ , and probability of being a single stellar population given the neural network in Sect. 4.



**Fig. 14.** Two examples of clusters in the catalogue that have detected tidal structures. The spatial distribution of the clusters Blanco 1 (top row) and Mamajek 4 (bottom row) are plotted on the left, with member stars reported in this work shown as filled blue circles and compared against member stars from Cantat-Gaudin & Anders (2020) which are plotted as empty black circles. CMDs are shown in the two plots on the right for both clusters.

### 7.3. Membership lists for individual clusters

Owing to the improved quality of *Gaia* DR3 data and the expanded selection of 729 million stars from *Gaia* data used as input into our cluster recovery pipeline, clusters in this work generally have more populated membership lists than in previous catalogues. Fig. 13 compares our membership lists with those from Cantat-Gaudin & Anders (2020) for five clusters randomly selected from our catalogue. Our membership lists typically have a higher total number of stars, with virtually all new member stars being compatible with the existing cluster CMD. This is particularly the case for clusters in regions with minimal crowding, where *Gaia* has a high completeness of stars with 5-parameter astrometry down to  $G \sim 20$ , with our membership lists containing stars down to approximately this limit. For more distant clusters such as Kronberger 4, membership lists are comparable in quality to those of Cantat-Gaudin & Anders (2020), as *Gaia* DR3 data does not present a large improvement in the astrometric quality of these distant sources compared to DR2. On average, our work contains 2.1 times as many member stars as the clusters we have in common with Cantat-Gaudin & Anders (2020), and 4.1 times as many member stars as the clusters we have in common with Kharchenko et al. (2013).

A second major advantage of our pipeline is that clusters are not forced to take a spherical shape, as with other methods such as Gaussian mixture models (Paper 1). Hence, we are able to detect tidal tails for many of the clusters in the catalogue, especially for those that are nearby and within 1 – 2 kpc. Tarricq et al. (2022) use HDBSCAN to detect tidal tails for 71 nearby OCs, many of which we are also able to detect. Figure 14 shows two examples of nearby clusters with well-resolved tidal tails using our methodology, Blanco 1 and Mamajek 4, both of which have reported tidal tails stretching around 50 pc from the centre of the cluster. Virtually all stars within the tidal structures appear

compatible with the isochrone of the cluster core, suggesting that they are stars with the same age, composition, and origin as the stars in the cluster cores. Particularly for clusters within 1 kpc, many of the clusters in our catalogue have tidal tails or comas.

However, as no current methodology for star cluster recovery from *Gaia* data is perfect (Paper 1), our membership lists are not without caveats – both of which are consistent with our results from Paper 1, but that are still worth mentioning in the main work of this catalogue.

Firstly, for distant OCs, our method may return fewer members than some other approaches. At high distances ( $d \gtrsim 5$  kpc), the errors on *Gaia* parallaxes and proper motions generally become much higher than the intrinsic dispersion of OCs, meaning that many members have low membership probabilities and can only be reliably assigned as members by incorporating error information. Our methodology does not use error information in the clustering analysis for reasons of speed and the fact that HDBSCAN does not directly include a way to consider errors on data in clustering analysis, although other methods such as UPMASK (Krone-Martins & Moitinho 2014) which do consider error information could return better membership lists for these distant clusters. This is visible for Kronberger 4 in Fig. 13, where the membership list of Cantat-Gaudin & Anders (2020) (which was compiled using UPMASK) has a slightly higher number of sources than our membership list, even though our list was compiled from a greater number of input sources due to our lack of a  $G$ -magnitude cut.

Secondly, HDBSCAN may sometimes return too many members, selecting regions larger than just an OC’s core and tidal tails. This is particularly common for young clusters, which are often embedded in regions of high stellar density where recent hierarchical star formation has occurred (Portegies Zwart et al. 2010). These clusters can be difficult for HDBSCAN to isolate from other surrounding stars and sub-clusters. One particular example can be seen for UPK 545 in Fig. 13. Although the tail emerging from the cluster core in the upper-left of the  $(l, b)$  plot appears compatible with a tidal tail, the connected structure to the right of the cluster is not. It appears to have the same age and composition as the cluster core, with all members of the tail being photometrically consistent with it. However, this ‘offshoot’ from the cluster may be better described as a separate cluster, which may also be bound to the core of UPK 545 in a binary pair of clusters, due to their proximity. Edge cases such as these are impossible to deal with autonomously with our current methodology and HDBSCAN alone, and require manual selection and separation of certain clusters in the catalogue into multiple separate components.

On a whole, the primary advantage of our catalogue is its completeness, generally reporting more member stars than previous works in the literature and doing so with a homogeneous methodology for a high number of total clusters. However, this is also the primary disadvantage of our catalogue: there are too many clusters and too many edge cases for all membership lists to be perfect, given only one clustering methodology. Hence, users of the catalogue who work with a small enough number of clusters are encouraged to manually check cluster membership lists and refine them depending on their application. To give one example, a user who wishes to only study cluster cores could refine our cluster membership lists by selecting a subset of them with Gaussian mixture models. With careful manual tweaking of the parameters of the mixture models, such a method could be used to remove tidal tails or possible other cluster components from our membership lists where necessary. Having discussed the general results of clusters in our catalogue, we next discuss



the reasons why many clusters reported in the literature may not appear in our catalogue.

## 8. Reasons for the non-detection of some literature objects

Thousands of new OCs and moving groups have been reported since the release of *Gaia* DR2 (Brown et al. 2018), with over 2000 reported in the last two years using *Gaia* DR3 data alone (Gaia Collaboration et al. 2021). While multiple works have commented on the reliability of individual clusters in the literature at-length (e.g. Cantat-Gaudin & Anders 2020; Piatti et al. 2023), as an unbiased search for all clusters within all of *Gaia* DR3, the results of this work offer a unique way to review the reliability of recently detected OCs on a large scale. In addition, with hundreds of literature OCs newly redetected in this work, this work also offers a chance to update the status of many older clusters reported in the pre-*Gaia* era.

The non-detection of a cluster by this work can be a result of multiple different factors. It is important to first rule out any possible methodological reasons before claiming that a given cluster does not exist. In Paper 1, we showed that our methodology has a high sensitivity, and hence a literature cluster being non-detected in this work can nevertheless raise strong doubts about whether or not it is real. With thousands of non-detected clusters, there are far too many to review all clusters individually, and hence we do not aim to decisively prove that some literature clusters are not real. We discuss the six main methodological and data-related reasons why a cluster may not appear in this work, concluding with questioning the existence of many objects reported in existing literature works.

### 8.1. Methodological reasons for the non-detection of a cluster

#### 8.1.1. Limitations of the clustering algorithm used

An obvious reason why we may not detect a given literature OC is due to limitations of the HDBSCAN algorithm that we use in this work. While we found in Paper 1 that HDBSCAN is the most sensitive clustering algorithm overall, DBSCAN was slightly more sensitive for clusters at distances greater than 5 kpc when applied to *Gaia* DR2 data. On the other hand, with respect to cluster size, HDBSCAN was the most sensitive algorithm for all sizes of cluster, although HDBSCAN and DBSCAN had similar or identical sensitivity for clusters with a number of members stars of  $n_{\text{stars}} = 10$ . Age and extinction were not found to have any significant differential impact on the sensitivity of the algorithms trialed, with all algorithms being more or less equally affected by older and/or heavily reddened clusters having fewer visible member stars, and hence being harder to detect.

The main limitation of HDBSCAN should be for clusters at distances greater than 5 kpc. However, only 6% and 21% of clusters from the DBSCAN-based works of Castro-Ginard et al. (2020) and Castro-Ginard et al. (2022) respectively that we are unable to detect have reported parallaxes of less than 0.2 mas, suggesting that distance-related detection issues alone are not enough to explain why certain clusters from these works are not detected. Additionally, we note that Castro-Ginard et al. (2022) using *Gaia* EDR3 were only able to recover  $\geq 80\%$  of clusters they found in DR2 in Castro-Ginard et al. (2020), and so DBSCAN itself between *Gaia* data releases is not able to reliably reconfirm all clusters it detected previously.

Nevertheless, Fig. 15 shows that our chance of recovering clusters at high distances can be lower for certain works. In particular, although we are unable to recover only 3.4% of clusters reported in Cantat-Gaudin & Anders (2020), most of the clusters from their work that we are unable to recover are small clusters at distances above 5 kpc, suggesting that an algorithmic limitation may contribute to why we are unable to recover remaining objects from Cantat-Gaudin & Anders (2020). A key difference between our work and Cantat-Gaudin & Anders (2020) is that their work used locations of clusters reported in the literature to narrow their search regions, which may in some cases be enough to make very distant clusters at the absolute limit of detectability in *Gaia* stand out. Future *Gaia* data releases with better data should provide additional clarity on whether or not such objects are real.

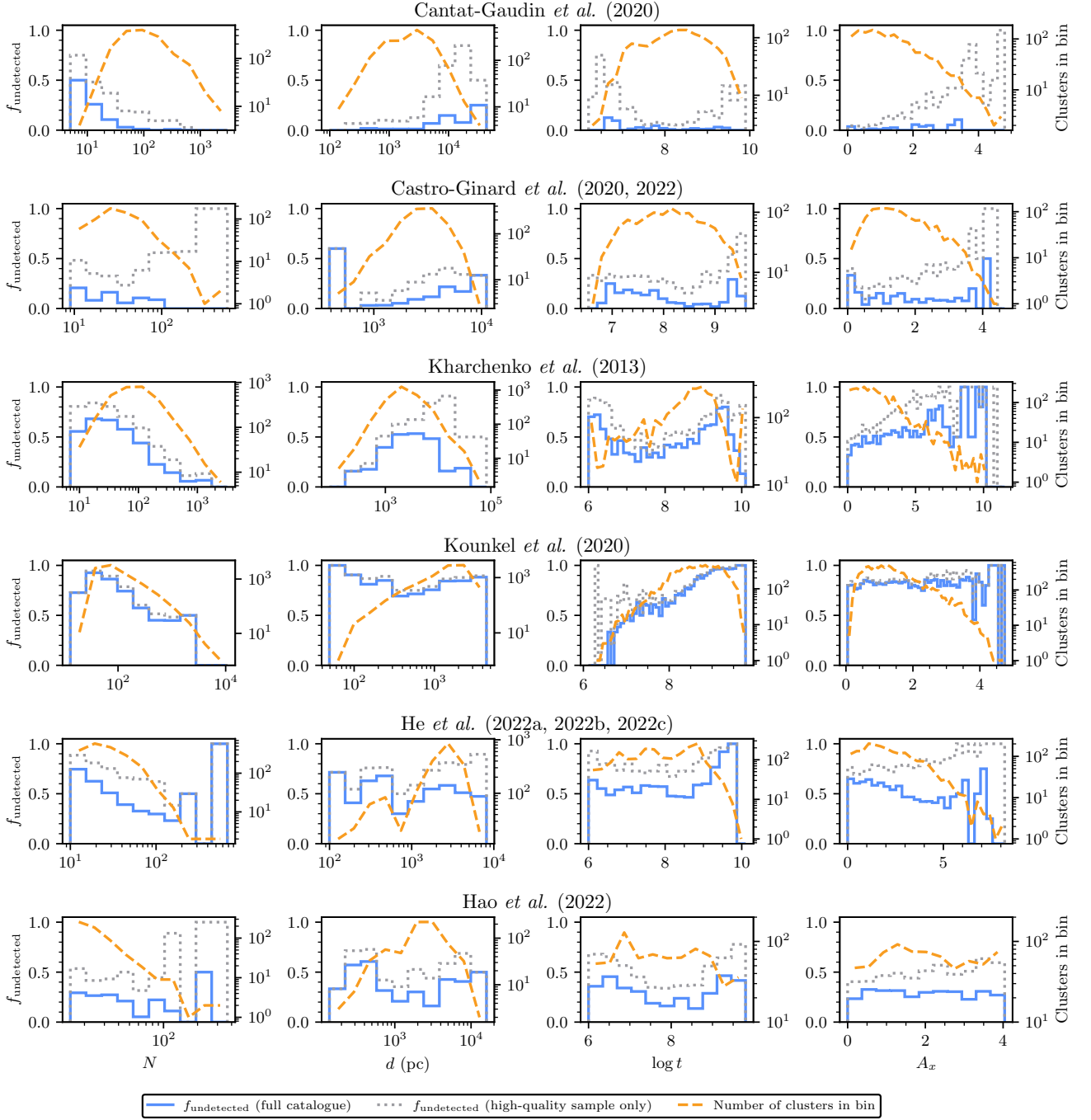
#### 8.1.2. Differences in the definition of an OC

There is no single agreed upon definition of an OC in the literature, and the slight differences in definition between works could cause some clusters to be detected or missed.

Principle amongst these definitions is the minimum number of observed member stars for a valid cluster,  $n_{\text{stars}, \text{min}}$ , which is important to distinguish star clusters from multiple star systems, also being used by some works as a proxy for the significance of a cluster relative to the field. In the literature, values of  $n_{\text{stars}, \text{min}}$  range from 8 in Castro-Ginard et al. (2022) to as high as 50 in Liu & Pang (2019), with most works coalescing around a value of between 10 and 12 (Krumholz et al. 2019). For the purposes of this work, we adopt a value of 10, and we should hence miss very few literature clusters due to this constraint alone.

Secondly, OCs generally have a population of stars with the same age and chemical composition, due to forming at the same time from the same molecular cloud (Cantat-Gaudin 2022). In practice, this is a difficult definition to constrain observationally, with the CMDs of OCs being broadened by effects such as differential extinction or outliers which are not true member stars, with these effects being worse with increasing distance and field star density. In addition, many OCs are not perfect single populations, with some hosting blue stragglers or having a clear second population in the form of an extended main-sequence turnoff (Cantat-Gaudin 2022). For the purposes of this work, we classify our clusters with our CMD classifier (see Sect. 4) and include all clusters in the final catalogue, instead leaving the task of removing clusters with poor photometry to the end user (recommending a minimum class value of 0.5). This means that no clusters are missing from the catalogue due to photometric reasons.

Finally, OCs must be distinguished from other types of single-population stellar overdensities. Star clusters can be divided into bound clusters (such as OCs and GCs) and unbound clusters (typically referred to as moving groups). Some works, such as Cantat-Gaudin & Anders (2020), use basic cuts on mean parameters to remove clear moving groups from their catalogue; we leave the classification of moving groups in our catalogue to a future work (Hunt & Reffert, *in prep.*) for reasons discussed in Sect. 9, and hence, no OCs are missing from this work due to being catalogued as moving groups. We do, however, flag known GCs in our catalogue by crossmatching against the catalogue of GCs of Vasiliev & Baumgardt (2021), with GCs in the Milky Way being distinguished from OCs by their age, which is typically greater than  $\sim 6\text{Gyr}$ , and their mass, which is typically greater than  $\sim 10^4 M_{\odot}$ , whereas most OCs have masses no higher than  $\sim 5000 M_{\odot}$  (Kharchenko et al. 2013). In total, differences in the fundamental definition of an OC between works



**Fig. 15.** Plots showing the fraction of clusters undetected by this work when compared to various literature works or series of literature works, shown as a histogram of various parameters as a solid blue line for all clusters in the catalogue, and a dashed grey line for clusters in the high quality sample defined in Sect. 7.1. The dashed orange lines show the number of clusters in each bin. Optimum histogram bin widths were selected automatically using `numpy` (Harris et al. 2020). From left to right, each column shows the number of stars  $N$ , distance  $d$ , age  $\log t$  and extinction  $A_x$  reported in each catalogue. For the top four groups of catalogues, extinctions were given in the  $V$  band. For the lower two, extinctions were given in *Gaia*'s  $G$  band, which are generally slightly lower.

should have a small impact on the inclusion of OCs in this work when compared to others.

### 8.1.3. Different quality cuts between different works

Different works in the literature often place different quality cuts on their catalogues, meaning that another possible reason why a given literature cluster does not appear in this catalogue would

be if it has been cut for quality reasons. Our catalogue adopts a philosophy of allowing users to decide their own quality cuts as much as possible, and hence includes all objects with bad photometry as well as moving groups that are unlikely to be bound OCs. The approach of allowing end users of the catalogue to define their own quality cuts is a similar philosophy to how *Gaia* data releases include many poor-quality sources, instead allowing users decide how strongly they wish to cut the *Gaia* catalogue (Gaia Collaboration et al. 2021). Poor photometry and the

bound or unbound status hence do not impact our recoverability of clusters in Fig. 15.

However, the sole quality cut applied to the catalogue that would affect its sensitivity is a cut on the astrometric S/N of detected clusters (derived using the CST) at  $3\sigma$ . This was performed because clusters with an S/N below this threshold are likely to be false positives, and because the high number of clusters below this threshold greatly complicated the process of merging results between different runs (see Sect. 3). Including such a quality cut dramatically improved the run merging process and hence our membership lists and completeness for reliable clusters, which is a more important scientific product than a list of low quality clusters that we cannot deem likely to be real clusters based on their S/N alone.

While we believe this is a fair trade-off to produce a catalogue that is as reliable as possible overall, it is likely that some real clusters are missed due to this cut on S/N. For instance, in Paper 1 using *Gaia* DR2 data, we tentatively detected Teutsch 156 with an S/N of  $0.68\sigma$ , which counted as a non-detection; however, using *Gaia* DR3, we clearly detect Teutsch 156 with an S/N of  $16.3\sigma$ . It is difficult to know exactly how many real literature clusters are missed due to this cut, particularly since some clusters in the literature with an S/N below  $3\sigma$  are likely to be statistical fluctuations and not real clusters, especially for S/Ns below  $1\sigma$ . This can be approximately estimated using the histogram of detected cluster S/Ns in Fig. 2. Since the distribution of literature cluster S/Ns is roughly flat for S/Ns below  $10\sigma$ , assuming that this trend continues for S/Ns below  $3\sigma$ , we may have missed approximately  $\sim 300$  crossmatches to clusters reported before *Gaia* DR3 and an additional  $\sim 400$  reported using *Gaia* DR3 data – although, owing to the low S/Ns that such objects would inevitably have, it is also likely that a number of these crossmatches would be false positives.

Inevitably, a repeat of this work with better data (such as *Gaia* DR4) would likely detect more of the objects that we do not recover with a sufficient statistical significance using *Gaia* DR3 data. In the future, further development of clustering algorithms that produce fewer false positives and can be ran on more data at once (both of which would tremendously simplify the run-merging process) would allow the minimum S/N threshold to be lowered.

#### 8.1.4. When two clusters are catalogued as one cluster

Certain other non-detections can be explained by further methodological differences. Sometimes, clusters reported as multiples in the literature are reported as a single object by HDBSCAN, even across all of its  $m_{clSize}$  runs. A notable example is UPK 533 from Sim et al. (2019), which was re-detected by Cantat-Gaudin & Anders (2020), but which HDBSCAN assigns as simply being a member of a tidal tail of a different and significantly larger nearby cluster, UPK 545, with no HDBSCAN  $m_{clSize}$  run separating the two objects. UPK 545 is shown in Fig. 13 on the third row. In this and other edge cases, our catalogue merges the two objects. An improved clustering algorithm that can separate edge-case binary clusters such as these autonomously would be helpful. However, only a small fraction of clusters (fewer than 1%) are affected by this issue.

#### 8.1.5. When a literature catalogue's parameters deviate too strongly from a detected cluster

While our crossmatching procedure as outlined in Sect. 6 aims to be as fair as possible, generally giving the benefit of the doubt to potential crossmatches, there are nevertheless cases where clusters reported in the literature still remain outside of our bounds for an accepted match. Generally, in all cases where this occurs, our detected cluster is significantly different to the literature object in at least one of the parameters considered for crossmatching, with these clusters representing ambiguous cases where it is not clear that the reported literature cluster is truly the same object.

CWNU 528 as reported in He et al. (2022a) is one example of a cluster reported in the literature that we are unable to detect within our crossmatching criteria. CWNU 528 is reported in He et al. (2022a) with 24 member stars, but appears to be a small offshoot of the recently reported new cluster OCSN 82 from Qin et al. (2023), which has an overall position different by around  $3^\circ$  and a total of 157 member stars. CWNU 528 is so much smaller than OCSN 82 and at such a different location that it does not crossmatch to it given our adopted crossmatching scheme, even though a few of the member stars in our detection of OCSN 82 are in common with CWNU 528 and they have similar proper motions and parallaxes.

This case is likely to have been repeated a few times, and appears particularly common with clusters detected in *Gaia* data using the DBSCAN algorithm (as in He et al. 2022a). In Paper 1, we commented that while DBSCAN has an excellent sensitivity and low false positive rate (depending strongly how the  $\epsilon$  parameter is chosen), it often had the sparsest and most incomplete membership lists of all algorithms we studied. Hence, detections of clusters may be so different or poor compared to what another algorithm recovers that crossmatch criteria may not be fulfilled, even when using a very permissive crossmatching scheme. In these cases, it is debatable whether the literature cluster is even the same object as the newly detected one.

#### 8.1.6. Limitations of *Gaia* data

Finally, it is worth considering the limitations of *Gaia* data itself, particularly when comparing our catalogue to works created from different data sources. Notably, the catalogue of Kharchenko et al. (2013) was compiled before *Gaia* and used infrared data from 2MASS (Skrutskie et al. 2006). Cantat-Gaudin & Anders (2020) are unable to recover a majority of the clusters from Kharchenko et al. (2013) using *Gaia* DR2 data, and we are unable to recover 48.4% of the clusters reported in their catalogue in *Gaia* DR3 data. Given that infrared light is significantly less affected by extinction than the visual light used to compile *Gaia* data, it begs the question of whether many clusters from Kharchenko et al. (2013) may still be missing from *Gaia*-based catalogues due to extinction limits.

However, Fig. 15 shows that extinction does not appear to play a major role in the non-detection of many clusters from Kharchenko et al. (2013). If extinction was a major contributor to why we are unable to detect so many of the clusters in their catalogue, then one would expect to see a linear trend in  $f_{undetected}$ ; all of their low-extinction clusters would be easily detected in *Gaia*, until some cut-off value beyond which *Gaia* detects no further clusters. On the contrary, most of their clusters have  $A_V < 5$ , and we are unable to detect around 50% of all clusters in this range with an approximately flat and uncorrelated distribution in the fraction of clusters recovered.

A few dozen of their reported clusters may be genuinely challenging to detect in *Gaia* data, since some of their clusters have  $A_V > 5$  and are at high distances of greater than 10 kpc. However, the majority of their clusters are within 10 kpc and have  $A_V < 5$ . Given that *Gaia* data have  $\sim 10^3$  times greater astrometric precision than *Hipparcos* data for  $\sim 10^5$  times as many stars (Gaia Collaboration et al. 2021), and given that our chance of detecting a cluster reported in Kharchenko et al. (2013) is uncorrelated with extinction for  $A_V < 5$ , limitations of *Gaia* data do not appear to be responsible for the bulk of non-detections of clusters from pre-*Gaia* works, despite assertions in recent works that *Gaia* data may be extinction-limited and unable to recover many highly reddened OCs from infrared datasets. Nevertheless, a handful of high-extinction clusters with  $A_V > 5$  reported in the literature may still be challenging to recover in *Gaia* data.

## 8.2. The cluster does not exist

Having exhausted all other major possibilities for why a cluster may not appear in our catalogue, the final potential reason would be that the cluster simply does not exist. As stated in the introduction to this section, far too many clusters are non-detected in this work for us to individually review them all and decisively prove that they are not real; however, we can give a broad overview of the typical characteristics of non-detected clusters, and contrast the similarities and differences between non-detected clusters in this work.

Figure 15 shows that the parameter most strongly correlated with  $f_{\text{undetected}}$  is the number of member stars  $N$ , with the smallest clusters from all papers being the least likely to be redetected. Few works report the statistical likelihood of a cluster being real in a way similar to the CST used in this work; however,  $N$  can be thought of as a good proxy for the statistical significance of a cluster, as it stands that a cluster with fewer member stars is probably less likely to be real. Clusters with fewer than 20 reported sources are often the most difficult to redetect.

In general, since most works in Fig. 15 use *Gaia* DR2 data or stronger cuts on *Gaia* data than our methodology, there are many cases where we should be able to detect their reported clusters easily and with a higher number of member stars and statistical significance. The fact that we cannot suggests that some of these clusters may have been statistically insignificant associations of a small number of member stars.

The distance of undetected reported literature clusters is similarly revealing. In Sect. 8.1.1, we suggest that some clusters may be undetected in this work at high distances due to limitations of the HDBSCAN algorithm. However, given that HDBSCAN should be the most sensitive algorithm for recovery of nearby clusters (Paper 1), it makes little sense that we are unable to recover a number of nearby clusters within 1 kpc for most of the works in Fig. 15. Many of these nearby and undetected objects may not be real, as there is no reason why we should not be able to detect them using the improved data of *Gaia* DR3 and the most sensitive algorithm for recovery of nearby OCs.

The age of undetected clusters paints a complicated picture. In principle, detecting an old cluster has two challenges. Firstly, as the cluster ages, the brightest stars in the cluster evolve into faint remnants, which reduces the number of stars visible in the cluster. This is a particular issue for distant old clusters, as the remaining fainter and longer-lived stars in a cluster may be below a survey's magnitude limit. In the case of *Gaia*, stars near its magnitude limit have the lowest accuracy astrometry, reducing the signal-to-noise ratio of a given old, distant cluster in proper motion and parallax space – further complicating its detection.

Secondly, as clusters age, they are theorised to take a sparser and less centrally concentrated distribution (Portegies Zwart et al. 2010), reducing their signal-to-noise ratio relative to background field stars in positional data.

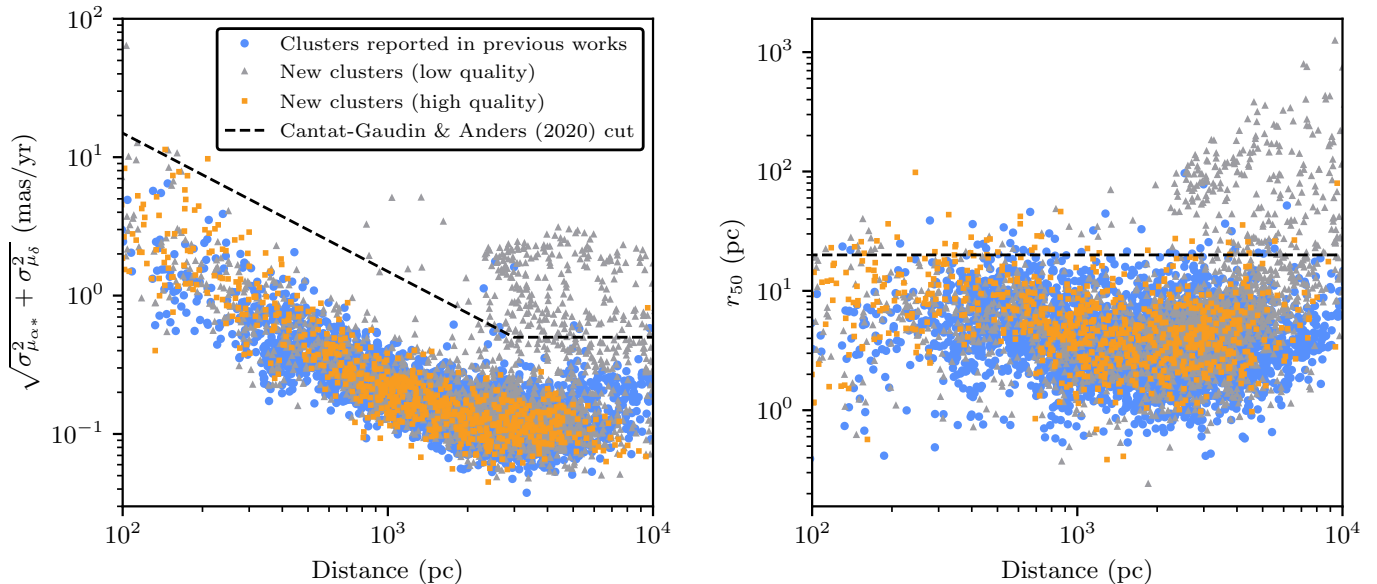
Although old clusters are likely to be harder to detect, in Paper 1, we found that the age of a reported cluster generally has the same effect on all algorithms: their lower number counts and sparsity affect all algorithms more or less equally in making them harder to detect. However, there are correlations between  $f_{\text{undetected}}$  and  $\log t$  for almost all papers in Fig. 15, despite all of them other than Kharchenko et al. (2013) being based on *Gaia* data and using methods found in Paper 1 to be equally affected by cluster age. Hence, these correlations may be more informative about the types of cluster in other catalogues that are false positives than on whether or not a given catalogue used a better method.

For all works other than Cantat-Gaudin & Anders (2020), clusters older than an age of around 1 Gyr ( $\log t > 9$ ) are much less likely to be redetected. Zucker et al. (2022) have recently investigated the nature of the groups reported in Kounkel et al. (2020), and find that many of them have ages  $\sim 120$  times larger than their dispersal times while being unbound and chemically homogeneous with their surrounding field stars – strongly suggesting that they are merely associations of field stars and not physical groupings. The fact that we are unable to redetect almost any of the groups older than 1 Gyr reported in Kounkel et al. (2020) supports this conclusion, with it being plausible that many of their oldest groups are instead associations of field stars, consistent with the mean ages of field stars in the galactic thin and thick disks of a few Gyr. The similar correlations with old clusters being undetected for other works may also suggest that a number of other old clusters reported in the literature are also associations of field stars with mean ages similar to that of the typical ages of unclustered field stars in the galactic disk.

The reasons for the non-detection of some young clusters are less clear, and are more surprising given that young clusters should be easier to detect. In the case of Cantat-Gaudin & Anders (2020), the handful of young clusters that we are unable to detect are also at high distances, which may mean that their non-detection is entirely a result of our own methodological limitations (see Sect. 8.1.1.) On the other hand, these distant, young clusters may have originally been detected by hand-searching for OB stars in pre-*Gaia* works and cataloguing them as OCs, but without a test of their physical nature, which could mean that they are associations. Similar reasoning could also be applied to the non-detected young clusters from Kharchenko et al. (2013). Both possibilities are plausible, and this should be investigated further in another work.

Finally, the reasons for the spikes in non-detected clusters between  $7 < \log t < 8$  for Castro-Ginard et al. (2020, 2022) and between  $6 < \log t < 7$  in Hao et al. (2022) remain unclear. These works are entirely compiled from *Gaia* DR2 and EDR3 data using the DBSCAN algorithm. Given that our results in Paper 1 suggest that clustering algorithms applied to *Gaia* data have no differences between themselves in their ability to detect clusters based on their age, there is no clear reason why these clusters would be undetectable. The non-detection of these clusters should be investigated further.

For most works, extinction  $A_V$  does not predict the chance of redetecting a given cluster. In Sect. 8.1.6, we discuss that  $A_V$  values of greater than  $\sim 5$  appear to reduce the chance of a cluster being recovered in *Gaia* data. The increasing trend in  $f_{\text{undetected}}$  for Cantat-Gaudin & Anders (2020) as a function of  $A_V$  appears to entirely be due to our lower chance of detecting clusters with



**Fig. 16.** Geometric mean of the proper motion dispersion (left) and radius containing 50% of members (right) for the clusters reported in this work, as a function of distance. Clusters are split between those detected in previous works (blue circles) and those newly reported in this work, divided between the high quality (orange squares) and low quality (grey triangles) samples defined in Sect. 7.1. The cuts on cluster parameters to distinguish between bound OCs and unbound moving groups or associations proposed in Cantat-Gaudin & Anders (2020) are shown as a dashed black line.

$d > 10$  kpc, since distant clusters also often have a high  $A_V$ . No other clear correlations exist for other works in Fig. 15 with respect to extinction, other than for a few dozen pre-*Gaia* clusters from the infra-red catalogue of Kharchenko et al. (2013) with  $A_V \gtrsim 5$  that we are unable to redetect with *Gaia* data.

In summary, we find that there are many potential reasons for the non-detection of given clusters from the literature, all of which should be investigated in more depth in future works. Verifying that new clusters reported in the literature are real is arguably as important as reporting them. While we cannot provide conclusive reasons for the non-detection of given clusters, given the scope of this survey, the overall trends we have identified should still be helpful and suggestive in whether or not given objects are real. We provide a table of all clusters non-detected by this work in Table B.1 and at the CDS.

## 9. The difficulties of distinguishing between open clusters and moving groups

Having discussed the catalogue’s overall quality for the verification and study of clusters reported previously in the literature, it is worth discussing the 2420 new objects reported in this work – 749 of which have a median CMD class above 0.5 and a CST of greater than  $5\sigma$ , and are hence the most reliable new objects that we report.

### 9.1. The case against many of our new clusters being OCs

On first inspection, despite having reliable CMDs and being statistically significant astrometric overdensities, many of our most reliable new objects have sparse density and proper motion distributions that appear more compatible with moving groups than spherically symmetric OCs with King (King 1962) or Plummer-like (Plummer 1911) profiles. Figure 17 shows three clusters randomly selected from the 749 most reliable objects. HSC 1131 is a sparse, elongated grouping of stars in the thin disk, with a

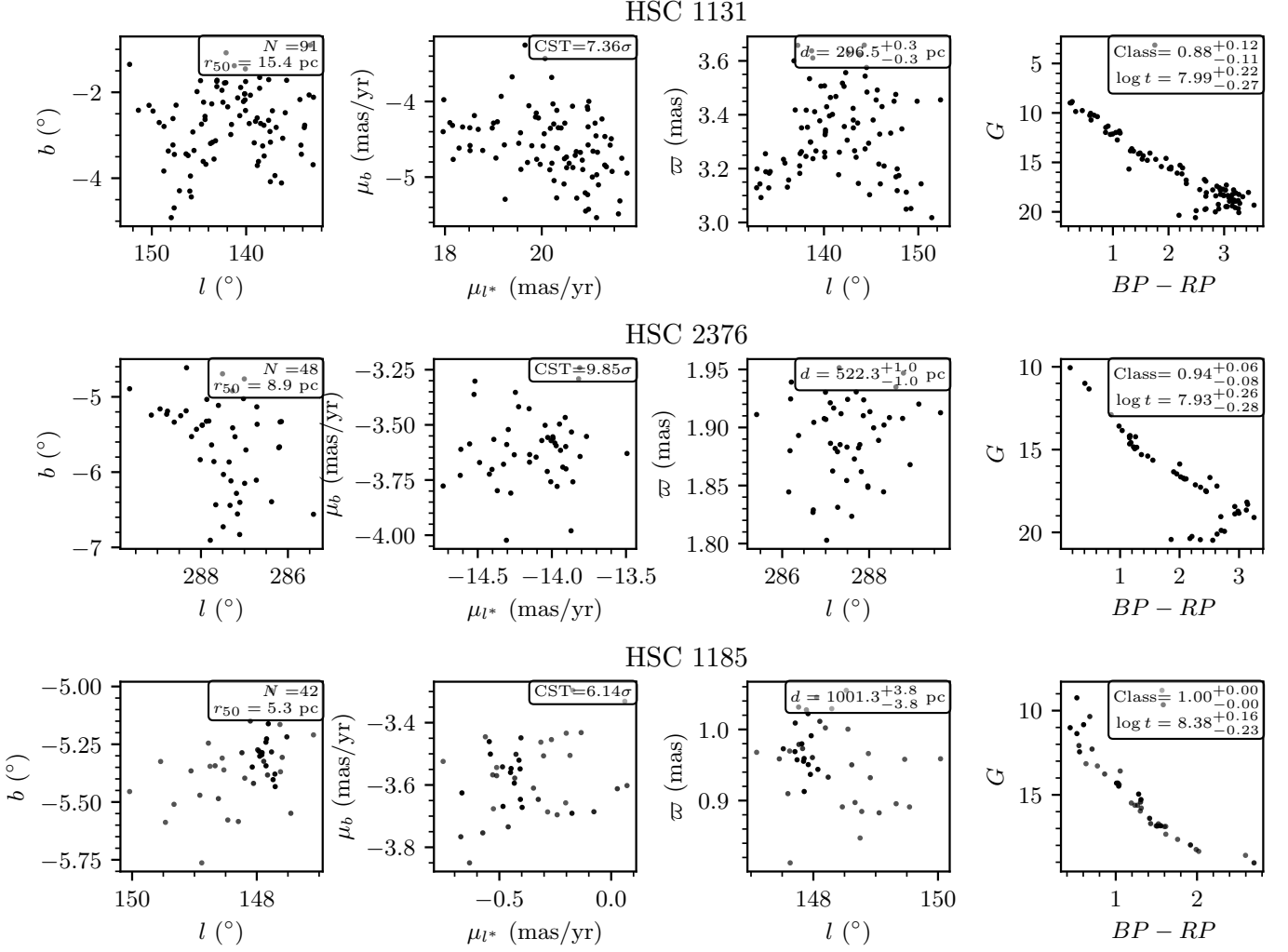
stringy nature much more compatible with a moving group than an OC. HSC 2376 is less clear, showing a more Gaussian clumping reminiscent of an OC within proper motion space but while still being relatively sparse, with  $r_{50} = 8.9$  pc. HSC 1185 appears visually to be the most OC-like cluster, with its distribution of member stars forming compacter Gaussian-like overdensities in spatial and proper motion plots.

While we have used tests on statistical significance and cluster CMDs to determine the reliability of clusters in the catalogue, it is clear that a further test on the astrometric parameters of clusters (such as sparsity and proper motion dispersion) is necessary. Cantat-Gaudin & Anders (2020) propose two tolerant cuts on cluster parameters, finding that requiring the geometric mean of proper motion dispersion to be less than a criterion (corresponding to  $\sim 5 \text{ km s}^{-1}$ ) and  $r_{50} < 20 \text{ pc}$  removed objects highly unlikely to be OCs from their sample.

However, Fig. 16 shows that with the exception of some clusters that are clearly associated with stellar streams (based on their location, CMD, and sparsity at distances greater than  $\sim 3$  kpc), most new clusters detected in this work are compatible with OCs given the tolerant cuts in Cantat-Gaudin & Anders (2020).

If almost all of the new clusters that we detect within 1 kpc of the Sun are in fact OCs, then this would represent a total paradigm shift in the census of OCs – with a large number of previously unseen low number count, low mass, and sparse clusters being detectable nearby with *Gaia* data. In reality, there are good reasons for this not being the case, and a more stringent cut on the astrometric parameters of candidate OCs is necessary.

In the preparation of this work, much effort was put in to attempting to find a more stringent cut on basic astrometric parameters (or some combination of them) to distinguish OCs from moving groups. We found that whether or not a cluster is a bound OC cannot be decided accurately based on individual cuts on  $r_{50}$  or proper motion dispersions alone, and instead requires at least some modelling of the cluster’s spatial profile, its velocity pro-



**Fig. 17.** Three newly reported clusters randomly selected from the cluster catalogue and ordered by increasing distance, with member stars plotted as a function of their astrometric and photometric data as in Fig. 13. All clusters pass the cuts proposed in Cantat-Gaudin & Anders (2020), have good-quality CMDs passing the cuts from Sect. 4, and have astrometric significances of greater than  $5\sigma$ , meaning they are almost certainly real overdensities in *Gaia* data.

file, and its mass. In the next section, we discuss the difficulties of such a method, which will be applied in the next paper in this series.

## 9.2. A test for if our OC candidates are bound

A given system is said to be in virial equilibrium if the absolute value of its potential energy  $|V|$  is equal to twice its kinetic energy  $T$ . A number of works have recently used a relationship derived from the virial theorem, which predicts a velocity dispersion that a cluster should have if it is bound,  $\sigma_{\text{vir}}$ , based on its mass and radius. This can be compared to the cluster's measured 1D velocity dispersion  $\sigma_{1D}$ , which should equal  $\sigma_{\text{vir}}$  if the cluster is bound:

$$\sigma_{\text{vir}} = \sqrt{\frac{GM}{\eta r_{\text{hm}}}} \approx \sigma_{1D} \text{ for a bound cluster,} \quad (1)$$

where  $r_{\text{hm}}$  is the cluster's half-mass radius,  $M$  is the cluster's mass,  $G$  is the gravitational constant and  $\eta$  is a constant depending on the cluster's density profile that is usually set to 10 (Porte-

gies Zwart et al. 2010). In the case when  $\sigma_{1D} \gg \sigma_{\text{vir}}$ , the cluster is likely to be unbound. This relationship has been used by works such as Bravi et al. (2018), Kuhn et al. (2019), and Pang et al. (2021) to test the virial nature of OCs using *Gaia* data, albeit in limited studies of no more than 28 clusters in one work.

While this relation is a promising way to distinguish between bound OCs and unbound moving groups, scaling this methodology to apply across our entire catalogue is extremely challenging. There are many systematics that can enter velocity dispersion, mass, and radius measurements, all of which must be reduced as much as possible to produce meaningful classifications. The clusters in our catalogue range across two orders of magnitude in distance, many orders of magnitude in mass, and two orders of magnitude in radius, with clusters of different parameters having fundamentally different challenges. For instance, nearby clusters may have tidal tails that must be removed from membership lists and may suffer from projection effects due to their radial velocity that would skew the measurement of their velocity dispersion with proper motions. On the other hand, distant clusters will push the limits of *Gaia*'s astrometric measurements, with velocity dispersions being difficult to measure precisely.



Given the scope of such a method, we leave its implementation to a future work. To restrict our catalogue to a reliable sample of OCs, users of our catalogue may for now use our CST scores, CMD classifications, and the criteria from Cantat-Gaudin & Anders (2020) to remove objects highly unlikely to be OCs. The next work to follow this one will provide a more accurate way to separate OCs from moving groups, and is anticipated to be submitted soon (Hunt & Reffert, *in prep.*).

## 10. Conclusions and future prospects

In this work, we conducted a blind all-sky search for Milky Way star clusters using *Gaia* DR3 data. We show that a single blind search can be used to produce a homogeneous star cluster catalogue in the *Gaia* era. We used the HDBSCAN algorithm, a density-based test of cluster significance, and a data partitioning scheme to detect as many reliable clusters as possible, producing a catalogue that is as complete and reliable as possible given current data. In total, the catalogue contains 7200 clusters, of which 4114 clusters form the most reliable sub-sample of objects with median CMD classifications greater than 0.5 and S/Ns greater than  $5\sigma$ .

We provide a wide range of parameters for clusters in the catalogue, including: basic astrometric parameters, S/Ns that correspond to their statistical significance given *Gaia* astrometry, CMD quality classifications, ages, extinctions, distances, and *Gaia* DR3 radial velocities. We recover large, expansive membership lists for many OCs, often including tidal tails for clusters within  $\sim 1$  kpc. Membership lists for all of our clusters are also available as a part of the catalogue (see Appendix A and the CDS).

Extensive care was taken to crossmatch our catalogue against 35 other works. To the best of the authors' knowledge, these works catalogue all OCs reported in the literature, including many thousands of OCs recently reported in the literature using *Gaia* data that are yet to be verified independently. 7022 clusters reported in the literature crossmatch against 4944 of the entries in our catalogue, including around 2000 of which we are able to independently verify for the first time. The spatial and age distribution of our catalogue traces the spiral arms in a similar way to many other recent works (e.g. Cantat-Gaudin et al. 2020; Castro-Ginard et al. 2021).

However, we are unable to recover many of the clusters reported in the literature, despite our methodology having the highest sensitivity for OC recovery of all methods we trialed in Paper 1. We discuss reasons why we may be unable to detect an OC and are able to tentatively suggest that many thousands of clusters reported in the literature may not be real, including calling into question the common assertion that *Gaia* is unable to recover a large fraction of OCs reported before *Gaia* due to being extinction-limited. Further investigations into whether or not many of the OCs we are unable to detect are real would be helpful to improve the accuracy of the OC census.

Our catalogue contains 2420 new objects as yet unreported in the literature, 749 of which are a part of our most reliable sample of clusters with median CMD classifications of greater than 0.5 and an S/N of greater than  $5\sigma$ . While some of these objects are likely to be new OCs, we find that many are more compatible with unbound moving groups, as our methodology is sensitive to all kinds of stellar overdensity in *Gaia* data. We find there is often no simple way to distinguish between the sparse, compact moving groups we detect and OCs, with the cuts on basic parameters proposed in Cantat-Gaudin & Anders (2020) being too lenient. In an upcoming work, we will use the virial

theorem to distinguish between bound and unbound clusters with a probabilistic methodology (Hunt & Reffert, *in prep.*).

The coming decade of Milky Way star cluster research is likely to continue to be exciting and fast-paced. Firstly, the quality of available data will increase ever-higher. *Gaia* DR4 will be produced from  $\sim 66$  months of data, almost double that of *Gaia* DR3, which will result in a large jump in the accuracy of available astrometric and photometric data. DR4 is currently slated for release no sooner than the end of 2025. The current planned final *Gaia* data release, DR5, may be based on around ten years of data, again roughly doubling the amount of input data used (Gaia Collaboration et al. 2021). Such large improvements in the accuracy of available astrometric data will inevitably result in more new clusters and improvements in the S/N and membership lists of existing clusters, further increasing the completeness and purity of the OC census.

Secondly, methodological improvements will continue to ease the process of star cluster recovery and characterisation. In the preparation of this work, it was still necessary to extensively verify many results by hand and develop postprocessing techniques to clean false positives from our catalogue. Improvements in clustering algorithms and techniques over the coming decade could make the process of cluster recovery more straightforward, accurate, and sensitive, with new methodologies such as Significance Mode Analysis (SigMA) methodology (Ratzenböck et al. 2022) showing promise in this area. As we discussed in Paper 1, there is currently no known perfect way to recover OCs from *Gaia* data; much work remains to be done to try and find one.

**Acknowledgements.** We thank the anonymous referee for their helpful comments that improved the clarity of this work. E.L.H. and S.R. gratefully acknowledge funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 138713538 – SFB 881 (“The Milky Way System”, sub-project B5). This work has made use of data from the European Space Agency (ESA) mission *Gaia* (<https://www.cosmos.esa.int/gaia>), processed by the *Gaia* Data Processing and Analysis Consortium (DPAC, <https://www.cosmos.esa.int/web/gaia/dpac/consortium>). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the *Gaia* Multilateral Agreement. This research has made use of NASA’s Astrophysics Data System Bibliographic Services. This research also made use of the SIMBAD database, operated at CDS, Strasbourg, France (Wenger, M. et al. 2000). In addition to those cited in the main body of the text, this work made use of the open source Python packages NumPy (Harris et al. 2020), SciPy (Virtanen et al. 2020), IPython (Pérez & Granger 2007), Jupyter (Kluyver et al. 2016), Matplotlib (Hunter 2007), pandas (McKinney 2010; Reback et al. 2020), Astropy (Robitaille et al. 2013; Astropy Collaboration et al. 2018), healpy (Zonca et al. 2019), and scikit-learn Pedregosa et al. (2011).

## References

- Abadi, M., Agarwal, A., Barham, P., et al. 2015, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems
- Abadi, M., Barham, P., Chen, J., et al. 2016, Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation, 21
- Anders, F., Castro-Ginard, A., Casado, J., Jordi, C., & Balaguer-Núñez, L. 2022, Research Notes of the American Astronomical Society, 6, 58
- Astropy Collaboration, Price-Whelan, A. M., Sipőcz, B. M., et al. 2018, The Astronomical Journal, 156, 123
- Bailer-Jones, C. A. L., Rybizki, J., Fouesneau, M., Demleitner, M., & Andrae, R. 2021, The Astronomical Journal, 161, 147
- Baratella, M., D’Orazi, V., Carraro, G., et al. 2020, Astronomy and Astrophysics, 634, A34
- Bastian, U. 2019, Astronomy & Astrophysics, 630, L8
- Becker, B., Vaccari, M., Prescott, M., & Grobler, T. L. 2021, Monthly Notices of the Royal Astronomical Society, 503, 1828
- Bica, E., Bonatto, C., Dutra, C. M., & Santos, J. F. C. 2008, Monthly Notices of the Royal Astronomical Society, 389, 678
- Bica, E., Pavani, D. B., Bonatto, C. J., & Lima, E. F. 2018, AJ, 157, 12
- Blundell, C., Corneise, J., Kavukcuoglu, K., & Wierstra, D. 2015, arXiv e-prints, arXiv:1505.05424

- Boffin, H. M. J., Carraro, G., & Beccari, G. 2015, *Astrophysics and Space Science Library*, Vol. 413, *Ecology of Blue Straggler Stars* (Springer Berlin, Heidelberg)
- Bossini, D., Vallenari, A., Bragaglia, A., et al. 2019, *A&A*, 623, A108
- Boubert, D. & Everall, A. 2020, *Monthly Notices of the Royal Astronomical Society*, 497, 4246
- Boubert, D., Everall, A., & Holl, B. 2020, *Monthly Notices of the Royal Astronomical Society*, 497, 1826
- Bravi, L., Zari, E., Sacco, G. G., et al. 2018, *Astronomy and Astrophysics*, 615, A37
- Bressan, A., Marigo, P., Girardi, Léo., et al. 2012, *Monthly Notices of the Royal Astronomical Society*, 427, 127
- Brown, A. G. A., Vallenari, A., Prusti, T., et al. 2018, *A&A*, 616, A1
- Campello, R. J. G. B., Moulavi, D., & Sander, J. 2013, *Advances in Knowledge Discovery and Data Mining*, 7819, 160
- Cantat-Gaudin, T. 2022, *Universe*, 8, 111
- Cantat-Gaudin, T. & Anders, F. 2020, *Astronomy and Astrophysics*, 633, A99
- Cantat-Gaudin, T., Anders, F., Castro-Ginard, A., et al. 2020, *A&A*, 640, A1
- Cantat-Gaudin, T. & Brandt, T. D. 2021, *Astronomy & Astrophysics*, 649, A124
- Cantat-Gaudin, T., Fouesneau, M., Rix, H.-W., et al. 2023, *Astronomy & Astrophysics*, 669, A55
- Cantat-Gaudin, T., Krone-Martins, A., Sedaghat, N., et al. 2019, *A&A*, 624, A126
- Cantat-Gaudin, T., Vallenari, A., Sordo, R., et al. 2018, *A&A*, 615, A49
- Cardelli, J. A., Clayton, G. C., & Mathis, J. S. 1989, *The Astrophysical Journal*, 345, 245
- Casado, J. 2021, *Research in Astronomy and Astrophysics*, 21, 117
- Casado, J. & Hendy, Y. 2023, *Monthly Notices of the Royal Astronomical Society*, stad071
- Castro-Ginard, A., Jordi, C., Luri, X., Cantat-Gaudin, T., & Balaguer-Núñez, L. 2019, *A&A*, 627
- Castro-Ginard, A., Jordi, C., Luri, X., et al. 2022, *Astronomy & Astrophysics*, 661, A118
- Castro-Ginard, A., Jordi, C., Luri, X., et al. 2020, *Astronomy & Astrophysics*, 635
- 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, *Astronomy & Astrophysics*, 652, A162
- Chi, H., Wang, F., & Li, Z. 2023, arXiv e-prints, arXiv:2302.08926
- Chi, H., Wei, S., Wang, F., & Li, Z. 2022, arXiv e-prints, arXiv:2212.11569
- Dias, W. S., Alessi, B. S., Moitinho, A., & Lépine, J. R. D. 2002, *Astronomy and Astrophysics*, 389, 871
- Dillon, J. V., Langmore, I., Tran, D., et al. 2017, arXiv e-prints, arXiv:1711.10604
- Ester, M., Krieger, H.-P., & Xu, X. 1996, in *KDD-96 Proceedings*, 6
- Ferreira, F. A., Corradi, W. J. B., Maia, F. F. S., Angelo, M. S., & Santos, Jr., J. F. C. 2021, *Monthly Notices of the Royal Astronomical Society*, 502, L90
- Ferreira, F. A., Santos, J. F. C., Corradi, W. J. B., Maia, F. F. S., & Angelo, M. S. 2019, *Monthly Notices of the Royal Astronomical Society*, 483, 5508
- Ferreira, F. A., Santos Jr., J. F. C., Corradi, W. J. B., Maia, F. F. S., & Angelo, M. S. 2020, *Monthly Notices of the Royal Astronomical Society*, 496, 2021
- Freibrich, D., Scholz, A., & Raftery, C. L. 2007, *Monthly Notices of the Royal Astronomical Society*, 374, 399
- Gaia Collaboration, Brown, A. G. A., Vallenari, A., et al. 2021, *Astronomy & Astrophysics*, 649, A1
- Gaia Collaboration, Prusti, T., de Bruijne, J. H. J., et al. 2016, *Astronomy & Astrophysics*, 595, A1
- Gaia Collaboration, Vallenari, A., Brown, A., Prusti, T., et al. 2022, *A&A*, arXiv:2208.00211
- Gal, Y. & Ghahramani, Z. 2015, arXiv e-prints, arXiv:1506.02142
- Goan, E. & Fookes, C. 2020, arXiv e-prints, arXiv:2006.12024 [cs, stat]
- Golovin, A., Reffert, S., Just, A., et al. 2023, *Astronomy & Astrophysics*, 670, A19
- Górski, K. M., Hivon, E., Banday, A. J., et al. 2005, *ApJ*, 622, 759
- Green, G. M. 2018, *Journal of Open Source Software*, 3, 695
- Hao, C., Xu, Y., Wu, Z., He, Z., & Bian, S. 2020, *Publications of the Astronomical Society of the Pacific*, 132, 034502
- Hao, C. J., Xu, Y., Wu, Z. Y., et al. 2022, *A&A*, 660, A4
- Harris, C. R., Millman, K. J., van der Walt, S. J., et al. 2020, *Nature*, 585, 357
- He, Z., Li, C., Zhong, J., et al. 2022a, *The Astrophysical Journal Supplement Series*, 260, 8
- He, Z., Liu, X., Luo, Y., Wang, K., & Jiang, Q. 2022b, *The Astrophysical Journal Supplement Series*, 264, 8
- He, Z., Wang, K., Luo, Y., et al. 2022c, *The Astrophysical Journal Supplement Series*, 262, 7
- He, Z.-H., Xu, Y., Hao, C.-J., Wu, Z.-Y., & Li, J.-J. 2021, *Research in Astronomy and Astrophysics*, 21, 093
- Hosek Jr, M. W., Lu, J. R., Lam, C. Y., et al. 2020, *The Astronomical Journal*, 160, 143
- Hron, J., Matthews, A. G. d. G., & Ghahramani, Z. 2017, arXiv e-prints, arXiv:1711.02989
- Huertas-Company, M., Rodriguez-Gomez, V., Nelson, D., et al. 2019, *Monthly Notices of the Royal Astronomical Society*, 489, 1859
- Hunt, E. L. & Reffert, S. 2021, *A&A*, 646, A104
- Hunter, J. D. 2007, *Computing in Science & Engineering*, 9, 90
- Jaehnig, K., Bird, J., & Holley-Bockelmann, K. 2021, *The Astrophysical Journal*, 923, 129
- Jospin, L. V., Buntine, W., Boussaid, F., Laga, H., & Benmamoun, M. 2022, arXiv e-prints, arXiv:2007.06823
- Kharchenko, N. V., Piskunov, A. E., Schilbach, E., Röser, S., & Scholz, R.-D. 2012, *A&A*, 543, A156
- Kharchenko, N. V., Piskunov, A. E., Schilbach, E., Röser, S., & Scholz, R.-D. 2013, *A&A*, 558, A53
- Killestein, T. L., Lyman, J., Steeghs, D., et al. 2021, *Monthly Notices of the Royal Astronomical Society*, 503, 4838
- King, I. 1962, *The Astronomical Journal*, 67, 471
- Kingma, D. P. & Ba, J. 2017, arXiv e-prints, arXiv:1412.6980
- Kluyver, T., Ragan-Kelley, B., Pérez, F., et al. 2016, in *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, ed. F. Loizides & B. Schmidt (IOS Press), 87–90
- Kounkel, M. & Covey, K. 2019, *AJ*, 158, 122
- Kounkel, M., Covey, K., & Stassun, K. G. 2020, *The Astronomical Journal*, 160, 279
- Kovaleva, D., Ishchenko, M., Postnikova, E., et al. 2020, *Astronomy & Astrophysics*, 642, L4
- Krause, M. G. H., Offner, S. S. R., Charbonnel, C., et al. 2020, *Space Science Reviews*, 216, 64
- Krone-Martins, A. & Moitinho, A. 2014, *Astronomy and Astrophysics*, 561, A57
- Kroupa, P. 2001, *Monthly Notices of the Royal Astronomical Society*, 322, 231
- Krumholz, M. R., McKee, C. F., & Bland-Hawthorn, J. 2019, *Annu. Rev. Astron. Astrophys.*, 57, 227
- Kuhn, M. A., Hillenbrand, L. A., Sills, A., Feigelson, E. D., & Getman, K. V. 2019, *The Astrophysical Journal*, 870, 32
- Leung, H. W. & Bovy, J. 2019, *Monthly Notices of the Royal Astronomical Society*, 483, 3255
- Li, Z., Deng, Y., Chi, H., et al. 2022, *ApJS*, 259, 19
- Li, Z. & Mao, C. 2023, *ApJS*, 265, 3
- Lin, Y.-C. & Wu, J.-H. P. 2021, *Physical Review D*, 103, 063034
- Lindgren, L., Bastian, U., Biermann, M., et al. 2021a, *Astronomy & Astrophysics*, 649, A4
- Lindgren, L., Hernández, J., Bombrun, A., et al. 2018, *Astronomy and Astrophysics*, 616, A2
- Lindgren, L., Klioner, S. A., Hernández, J., et al. 2021b, *Astronomy & Astrophysics*, 649, A2
- Liu, L. & Pang, X. 2019, *The Astrophysical Journal Supplement Series*, 245, 32
- Lu, J. R., Do, T., Ghez, A. M., et al. 2013, *ApJ*, 764, 155
- Marigo, P., Girardi, L., Bressan, A., et al. 2017, *ApJ*, 835, 77
- McArthur, B. E., Benedict, G. F., Harrison, T. E., & van Altena, W. 2011, *The Astronomical Journal*, 141, 172
- McInnes, L., Healy, J., & Astels, S. 2017, *Journal of Open Source Software*, 2, 205
- McKinney, W. 2010, in *Proceedings of the 9th Python in Science Conference*, Austin, Texas, 56–61
- Meingast, S., Alves, J., & Rottensteiner, A. 2021, *A&A*, 645, A84
- Pang, X., Li, Y., Yu, Z., et al. 2021, *The Astrophysical Journal*, 912, 162
- Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, *Journal of Machine Learning Research*, 12, 2825
- Penoyre, Z., Belokurov, V., & Evans, N. W. 2022, *Monthly Notices of the Royal Astronomical Society*, 513, 2437
- Pérez, F. & Granger, B. E. 2007, *Computing in Science and Engineering*, 9, 21
- Perryman, M. a. C., Lindgren, L., Kovalevsky, J., et al. 1997, *Astronomy and Astrophysics*, Vol. 323, p.L49-L52, 323, L49
- Piatti, A. E., Illesca, D. M. F., Massara, A. A., et al. 2023, *Monthly Notices of the Royal Astronomical Society*, 518, 6216
- Plummer, H. C. 1911, *Monthly Notices of the Royal Astronomical Society*, 71, 460
- Portegies Zwart, S. F., McMillan, S. L. W., & Gieles, M. 2010, *Annual Review of Astronomy and Astrophysics*, 48, 431
- Qin, S., Zhong, J., Tang, T., & Chen, L. 2023, *The Astrophysical Journal Supplement Series*, 265, 12
- Qin, S.-m., Li, J., Chen, L., & Zhong, J. 2021, *Research in Astronomy and Astrophysics*, 21, 045
- Ratzenböck, S., Großschedl, J. E., Möller, T., et al. 2022, arXiv e-prints, arXiv:2211.14225
- Reback, J., McKinney, W., jbrockmendel, et al. 2020, *Pandas-Dev/Pandas: Pandas 1.0.3*, Zenodo
- Riello, M., De Angeli, F., Evans, D. W., et al. 2021, *Astronomy & Astrophysics*, 649, A3
- Robitaille, T. P., Tollerud, E. J., Greenfield, P., et al. 2013, *A&A*, 558, A33

- Rybizki, J., Green, G., Rix, H.-W., et al. 2022, *Monthly Notices of the Royal Astronomical Society*, 510, 2597
- Santos-Silva, T., Perottoni, H. D., Almeida-Fernandes, F., et al. 2021, *Monthly Notices of the Royal Astronomical Society*, 508, 1033
- Schmeja, S., Kharchenko, N. V., Piskunov, A. E., et al. 2014, *A&A*, 568, A51
- Sim, G., Lee, S. H., Ann, H. B., & Kim, S. 2019, *Journal of the Korean Astronomical Society*, 52, 145
- Skrutskie, M. F., Cutri, R. M., Stiening, R., et al. 2006, *The Astronomical Journal*, 131, 1163
- Tarricq, Y., Soubiran, C., Casamiquela, L., et al. 2022, *Astronomy & Astrophysics*, 659, A59
- Tian, H.-J. 2020, *The Astrophysical Journal*, 904, 196
- Vasiliev, E. & Baumgardt, H. 2021, *Monthly Notices of the Royal Astronomical Society*, 505, 5978
- Virtanen, P., Gommers, R., Oliphant, T. E., et al. 2020, *Nature Methods*, 17, 261
- von Hippel, T., Jefferys, W. H., Scott, J., et al. 2006, *ApJ*, 645, 1436
- Wen, Y., Vicol, P., Ba, J., Tran, D., & Grosse, R. 2018, *arXiv e-prints*, arXiv:1803.04386
- Wenger, M., Ochsenbein, F., Egret, D., et al. 2000, *Astron. Astrophys. Suppl. Ser.*, 143, 9
- Zari, E., Hashemi, H., Brown, A. G. A., Jardine, K., & de Zeeuw, P. T. 2018, *Astronomy & Astrophysics*, 620, A172
- Zonca, A., Singer, L. P., Lenz, D., et al. 2019, *Journal of Open Source Software*, 4, 1298
- Zucker, C., Peek, J. E. G., & Loebman, S. 2022, *The Astrophysical Journal*, 936, 160

## Appendix A: Description of contents of online tables

We provide tables of clusters, rejected clusters, member stars, and members stars for rejected clusters at the CDS. Tables of clusters follow the table format in Table A. Tables of members follow the same columns and column naming scheme as in *Gaia* DR3 (Gaia Collaboration et al. 2022), except while also having columns referencing the cluster name and cluster ID we assign them to, the cluster membership probability, and a flag for if the star is a member within our estimated tidal radius  $r_t$ .

## Appendix B: Table of crossmatch results

Here we provide a table of all crossmatches to all literature clusters that meet our adopted crossmatch criteria from Sect. 6 in Table B.1. For every cluster in the literature that we detect in this work, the table lists the internal cluster ID corresponding to our table of clusters in Table 4 that corresponds to this object. For clusters that we do not redetect, only a blank row with the cluster name, source paper, and type of crossmatch is shown.

## Appendix C: Bayesian neural networks

Given that Bayesian neural networks (BNNs) are only just beginning to see use in the astronomical literature (e.g. Huertas-Company et al. 2019), here we provide a brief background overview of the advantages and caveats of the approximate BNN methodology we adopted in Sect. 4 and Sect. 5.

BNNs are a somewhat elusive area of open research in machine learning. Their appeal is clear: unlike a deterministic approach or an approach based on simply perturbing network inputs, a perfect BNN would be able to estimate both aleatoric uncertainties, which are uncertainties that result from random phenomena, such as uncertainty on photometric measurements; and epistemic uncertainties, which are uncertainties that result from a lack of knowledge about the underlying processes being modelled. For instance, any remaining gaps or issues in the simulated training data we use would cause a traditional deterministic neural network to always output an incorrect answer, whereas a probabilistic neural network should at least output a wide range of answers that demonstrate its uncertainty in such difficult cases (Goan & Fookes 2020; Jospin et al. 2022).

In practice, there is currently no perfect BNN architecture, with all approaches having some flaws (Goan & Fookes 2020; Jospin et al. 2022). While a Monte-Carlo Markov chain (MCMC)-based approach should in theory be superior, where every network weight has an arbitrary posterior distribution, MCMC-based BNNs are extremely difficult or impossible to train accurately, with current sampling techniques being inadequate (Goan & Fookes 2020). In addition, BNNs are often time consuming to train. Instead, ‘variational inference’ is widely used to approximate BNNs. In this technique, an ideal BNN is approximated by perturbing network features, approximating a BNN by ‘emphasising or de-emphasising’ certain parts of a trained model when the model is sampled. This can then be used to estimate the epistemic uncertainty of a model by sampling a variational network multiple times.

Many approaches for variational inference exist in the literature, with a common approach being dropout regularisation as an approximation of a BNN (Gal & Ghahramani 2015), having also been used within astronomy (e.g. Huertas-Company et al. 2019; Leung & Bovy 2019). However, this approximation is not inherently Bayesian (Hron et al. 2017), and may be improved

**Table A.1.** Description of the columns in the tables of detected clusters.

Col.	Label	Unit	Description
1	Name	–	Designation
2	Internal ID	–	Internal designation
3	All names	–	All literature names
4	Kind	–	Estimated object type <sup>c</sup>
5	$n_{\text{stars}}$	–	Num. of member stars
6	S/N	–	Astrometric S/N
7	$n_{\text{stars}} _{r_t}$	–	$n_{\text{stars}}$ within $r_t$
8	S/N $ _{r_t}$	–	S/N within $r_t$
9-10	$\alpha, \delta$	deg	ICRS position
11-12	$l, b$	deg	Galactic position
13-16	$r_{50, c, t, \text{tot}}$	deg	Angular radii
17-20	$R_{50, c, t, \text{tot}}$	pc	Physical radii
21-26 <sup>a</sup>	$\mu_{\alpha^*}, \mu_{\delta}$	mas yr <sup>-1</sup>	ICRS proper motions
27-29 <sup>a</sup>	$\varpi$	mas	Parallax
30-32 <sup>b</sup>	$d$	pc	Distance
33	$n_d$	pc	$n_{\text{stars}}$ for distance calc.
34	$\varpi_0$ type	–	Parallax offset type <sup>d</sup>
35-37	X, Y, Z	pc	Galactocentric coords.
38-40 <sup>a</sup>	RV	km s <sup>-1</sup>	Radial velocity <sup>e</sup>
41	$n_{\text{RV}}$	–	$n_{\text{stars}}$ with RVs
42-46 <sup>b</sup>	CMD class	–	CMD class quantiles <sup>f</sup>
47	Human class	–	(where available) <sup>f</sup>
48-50 <sup>b</sup>	$\log t$	log [yr]	Cluster age
51-53 <sup>b</sup>	$A_V$	mag	V-band extinction
54-56 <sup>b</sup>	$\Delta A_V$	mag	Differential $A_V$
57-59 <sup>b</sup>	$m - M$	mag	Photometric dist. mod.
60	$m_{\text{clsize}}$	–	HDBSCAN parameter
61	merged	–	Flag if merged <sup>g</sup>
62	is_gmm	–	Flag if GMM used <sup>h</sup>
63	$n_{\text{crossmatches}}$	–	Num. crossmatches
64	Xmatch type	–	Type of crossmatch <sup>i</sup>

**Notes.** The full version is available at the CDS. <sup>(a)</sup> Mean value, standard deviation  $\sigma$ , and standard error  $\sigma / \sqrt{n}$  are given. <sup>(b)</sup> Median value and various confidence intervals are given. <sup>(c)</sup> g for objects in the Vasiliev & Baumgardt (2021) GC catalogue, otherwise o (OC) or m (moving group) for clusters according to the empirical cuts in Cantat-Gaudin & Anders (2020). <sup>(d)</sup> Flag indicating six clusters for which parallax bias correction using the method of Lindegren et al. (2021b) was not possible, and a global offset was used instead (see Sect. 3.3). <sup>(e)</sup> Corrected using cluster distances to be relative to cluster centre. <sup>(f)</sup> Cluster CMD classes (the probability of a given cluster being a single coeval population of stars) derived using the neural network in Sect. 4. Some clusters also appeared in our human-labelled test dataset, for which human classes are also listed. <sup>(g)</sup> Indicates 25 clusters merged by hand where initial HDBSCAN clustering was unsatisfactory (see Sect. 3). <sup>(h)</sup> Indicates nine clusters with members from an additional Gaussian mixture model clustering step, typically applied to difficult to separate binary clusters. <sup>(i)</sup> Method used to assign name to cluster. In particular, ‘many to many’ crossmatches are the most difficult to assign due to multiple objects crossmatching to multiple literature entries co-dependently (see Sect. 6.3 for full discussion of final cluster name assignments.)

upon with recent developments in the literature. Another common approximation is to assume that all layer kernel and bias weights are drawn from simple distributions, such as independent Gaussian distributions. This allows for gradients during network training to be calculated straightforwardly using Bayes by backpropagation (Blundell et al. 2015). This approximation can hold relatively well for (simple) neural networks, which often have normally distributed weights, but may cause underfitting on

**Table B.1.** All cluster crossmatches, including literature clusters that have no match.

ID	Name	Source	Type	$\theta$ ( $^{\circ}$ )	$\theta_r^a$	$s_{\mu_{\alpha^*}}$ (mas yr $^{-1}$ )	$\sigma_{\mu_{\alpha^*}}$	$s_{\mu_{\delta}}$ (mas yr $^{-1}$ )	$\sigma_{\mu_{\delta}}$	$s_{\varpi}$ (mas)	$\sigma_{\varpi}$
...											
176	Basel 1	Cantat-Gaudin+20	gaia dr2	0.01	0.04	0.03	0.00	0.01	0.00	0.01	0.00
176	Basel 1	Dias+02	position	0.03	0.12	-	-	-	-	-	-
176	Basel 1	Kharchenko+13	hipparcos	0.01	0.04	0.40	0.09	1.04	0.24	0.06	0.17
179	Basel 10	Bica+18	position	0.01	0.04	-	-	-	-	-	-
179	Basel 10	Dias+02	position	0.01	0.04	-	-	-	-	-	-
179	Basel 10	Cantat-Gaudin+20	gaia dr2	0.01	0.07	0.03	0.00	0.05	0.01	0.01	0.00
179	Basel 10	Kharchenko+13	hipparcos	0.01	0.07	0.30	0.05	2.49	0.51	0.02	0.00
179	Basel 10	Kharchenko+13	position	0.01	0.07	-	-	-	-	-	-
183	Basel 11A	Cantat-Gaudin+20	gaia dr2	0.01	0.01	0.02	0.00	0.05	0.00	0.02	0.00
183	Basel 11A	Kharchenko+13	hipparcos	0.01	0.04	0.52	0.12	1.66	0.42	0.11	0.81
183	Basel 11A	Dias+02	position	0.02	0.06	-	-	-	-	-	-
183	Basel 11A	Bica+18	position	0.03	0.06	-	-	-	-	-	-
183	Basel 11A	Kharchenko+13	position	0.01	0.04	-	-	-	-	-	-
3003	Basel 11B	Kharchenko+13	position	0.11	0.25	-	-	-	-	-	-
184	Basel 11B	Kharchenko+13	hipparcos	0.02	0.06	1.28	0.37	0.24	0.06	0.17	1.40
184	Basel 11B	Kharchenko+13	position	0.02	0.06	-	-	-	-	-	-
184	Basel 11B	Dias+02	position	0.01	0.02	-	-	-	-	-	-
184	Basel 11B	Cantat-Gaudin+20	gaia dr2	0.01	0.03	0.02	0.00	0.01	0.00	0.03	0.00
184	Basel 11B	Bica+18	position	0.00	0.01	-	-	-	-	-	-
6363	Basel 11B	Kharchenko+13	hipparcos	0.11	0.39	2.15	0.64	1.99	0.59	0.22	1.98
6363	Basel 11B	Kharchenko+13	position	0.11	0.39	-	-	-	-	-	-
-	Basel 12	Dias+02	position	-	-	-	-	-	-	-	-
-	Basel 12	Kharchenko+13	hipparcos	-	-	-	-	-	-	-	-
-	Basel 12	Bica+18	position	-	-	-	-	-	-	-	-
180	Basel 13	Kharchenko+13	position	0.11	0.74	-	-	-	-	-	-
-	Basel 13A	Bica+18	position	-	-	-	-	-	-	-	-
-	Basel 14	Dias+02	position	-	-	-	-	-	-	-	-
-	Basel 14	Kharchenko+13	hipparcos	-	-	-	-	-	-	-	-
-	Basel 15	Bica+18	position	-	-	-	-	-	-	-	-
-	Basel 15	Kharchenko+13	hipparcos	-	-	-	-	-	-	-	-
-	Basel 15	Dias+02	position	-	-	-	-	-	-	-	-
181	Basel 17	Kharchenko+13	position	0.00	0.02	-	-	-	-	-	-
181	Basel 17	Cantat-Gaudin+20	gaia dr2	0.01	0.07	0.01	0.00	0.06	0.08	0.01	0.00
181	Basel 17	Kharchenko+13	hipparcos	0.00	0.02	0.62	0.10	2.37	0.43	0.11	0.93
181	Basel 17	Dias+02	position	0.00	0.02	-	-	-	-	-	-
...											

**Notes.** The full version is available at the CDS; the above only shows crossmatches against a selection of Basel clusters. Depending on the type of work crossmatched against, only separations in terms of position  $\theta$  may be listed. For works with astrometry, separations  $s$  with respect to  $\mu_{\alpha^*}$ ,  $\mu_{\delta}$ , and  $\varpi$  are shown, in addition to separations  $\sigma$  which are in terms of standard deviations about the mean of the astrometry of these clusters added together in quadrature, after accounting for worst-case systematics. Cluster entries in the literature that did not have a valid crossmatch against any cluster detected in this study are listed with only the name, source, and source type columns filled. Recalling Sect. 6, for a valid crossmatch, we require  $\theta_r < 1$ , and additionally, when crossmatching to a work with full five parameter astrometry, all  $\sigma$  values to be less than two. <sup>(a)</sup> The separation between cluster centres in terms of the largest cluster radius available,  $\theta_r = \theta / \max(r_t, r_{t,\text{lit}})$

more complicated problems (Goan & Fookes 2020). Due to the time-consuming nature of repeated samples of all kernel and bias posterior distributions, we also apply an approximation known as Flipout to more efficiently sample them with a lower runtime while preserving good training characteristics (Wen et al. 2018). Similar approaches using Bayes by backpropagation and Flipout have seen some use in the astronomy literature (e.g. Lin & Wu 2021). We use the implementations of DenseFlipout and Convolution2DFlipout layers in TensorFlow Probability (Dillon et al. 2017), minimising the evidence lower bound (ELBO) loss (Blundell et al. 2015).

In initial tests, these approximations produced network outputs with reliable uncertainty estimates that correspond well to the uncertainty inherent to classifying star cluster CMDs. It is

worth noting from the literature that variational-inference based approaches are still more overconfident than a true BNN when applied to unseen data (Goan & Fookes 2020), and that this approach is still an imperfect estimator of the true uncertainty of our model; nevertheless, our adopted method was found to be as accurate as a traditional deterministic network architecture of the same configuration when applied to our training data, but while providing an estimate of its uncertainty and without dramatically increasing runtime during training or sampling.