Ruprecht-Karls-Universität

Improving the census of open clusters in the Milky Way with data from *Gaia*

Emily Lauren Hunt

Dissertation

submitted to the
Combined Faculty of Mathematics, Engineering and Natural Sciences
of Heidelberg University, Germany
for the degree of

Doctor of Natural Sciences

Put forward by

Emily Lauren Hunt
born in: Coventry, United Kingdom

Oral examination: July 12th, 2023

Improving the census of open clusters in the Milky Way with data from *Gaia*

Referees: PD Dr. Sabine Reffert
Prof. Dr. Hans-Walter Rix

Emily Lauren Hunt
Improving the census of open clusters in the Milky Way with data from Gaia
Ruprecht-Karls-Universität, May 2nd, 2023
Reviewers: PD Dr. Sabine Reffert and Prof. Dr. Hans-Walter Rix
Supervisor: PD Dr. Sabine Reffert

Ruprecht-Karls-Universität
Extrasolar Planet Research Group
Landessternwarte Königstuhl
Zentrum für Astronomie
Königstuhl 12
69117 Heidelberg

Abstract

TODO: abstract

Zusammenfassung

TODO: german abstract

Acknowledgement

TODO: write acknowledgements

Contents

1 Introduction

> Δέδυκε μὲν ἀ σελάννα
> *The moon and the Pleiades*
>
> καὶ Πληΐαδες, μέσαι δέ
> *have set, it is*
>
> νύκτες, πάρα δ΄ ἔρχετ΄ ὤρα,
> *midnight, time is passing,*
>
> ἔγω δὲ μόνα κατεύδω.
> *but I sleep alone.*
>
> — **Sappho, 'The Midnight Poem'**
> (c. 600 BC)

1.1 From seven sisters to a powerhouse of astronomy

In all of astronomy, few objects have retained relevance throughout the centuries as much as open clusters (OCs). Easily visible to the naked eye, the Pleiades has been observed since at least the dawn of civilisation (**rappengluck_palaeolithic_timekeepers_2001**; **mozel_sky_disk_2003**), along with a handful of other OCs visible without a telescope. In the present day, the now thousands of known OCs are a key tool in modern astronomy for understanding stellar and galactic evolution.

Star clusters are formed when clouds of cold molecular gas collapse due to gravity, forming stars. Sometimes, when star formation occurs densely enough, these stars fall further into gravitationally bound clusters that can survive in the galactic disk for as long as $\sim 10^9$ years (**lada_embedded_2003**; **portegies_zwart_young_2010**). It is this property of the formation of OCs that makes them so useful: having formed at the same time and from the same molecular cloud, all stars in an OC will have the same age and initial chemical composition, and will remain co-located in space in a dense cluster. Even hundreds of millions of years after an OC forms, the parameters of the cluster's member stars can be measured significantly more precisely than when studying stars in isolation.

For instance, when a parameter such as the distance of member stars can simply be averaged over all member stars, then the precision of the mean distance of an OC (and hence the distance to all of its member stars) will be a factor $\sqrt{n}$ more precise than the distance to any individual star. Alternatively, when a property such as chemical composition is highly time consuming to derive, it can be derived for a fraction of stars in an OC and be applied relatively safely to all stars in a cluster.

The ease of studying stellar astrophysics with OCs results in OCs having an extremely wide range of scientific use cases. For instance, OCs are used as testing grounds for stellar evolution models (**bressan_parsec_2012**), as tracers of galactic structure (**cantat-gaudin_painting_2020**; **castro-ginard_milky_2021**), or even as calibrators of Cepheid variable stars (**medina_revisited_2021**), which are an essential first rung on the cosmic distance ladder and are vital in the derivation of the cosmological parameters of the universe. It is somewhat of a cliché to describe OCs as 'the laboratories of stellar evolution', but it really is true: OCs are a fantastic way to observe stars of a given age and composition across a broad range of masses, and to do so with orders of magnitude more precision than when studying isolated field stars.

The best part of the modern story of the OC's contribution to astrophysics comes with the *Gaia* satellite, however. In just five years since its first full data release (**brown_gaia_2018**), *Gaia* has revolutionised the study of our galaxy, including the study of OCs; with dozens of papers reporting thousands of new objects (**liu_catalog_2019**; **castro-ginard_hunting_2019**; **castro-ginard_hunting_2020**; **castro-ginard_hunting_2022**), and a number of works deriving dramatically improved parameters and members for OCs in the Milky Way (**cantat-gaudin_gaia_2018**; **tarricq_3d_2020**). Arguably, there has never been a better time to do science with OCs, owing to the incredible quantity and quality of data that *Gaia* has provided.

There is, however, a catch. Even though the Milky Way is estimated to contain as many as 10^5 OCs (**dias_new_2002**), there are still only a few thousand currently known in the literature – representing a small fraction of the total number of OCs in our galaxy. It has been shown that the census of OCs is incomplete within even 1 kpc from the Sun (**castro-ginard_new_2018**), and the extent of the remaining incompleteness is unknown. Worse still, it has been shown that many of the OCs catalogued previously in the literature may not exist (**cantat-gaudin_clusters_2020**; **piatti_catching_2023**), with it being largely unknown which OCs are or are not real. The many fantastic uses of OCs in other areas of astronomy are contingent on a reliable, accurate, and complete census of OCs; and the many current caveats

with the census of OCs limit the science potential of these fantastic objects, in a time when we have more available data with which to study them than ever before.

In this thesis, I will present solutions to a number of the current issues with the OC census in the era of *Gaia*, using a range of data analysis and parameter inference techniques. I will then use these techniques to create the largest census of OCs to date and derive a range of parameters for these OCs. With this thesis, I also hope to present methods that could continue to be used to maximise the quality of the OC census for the coming decade of *Gaia* data releases – as well as for whatever instruments supercede *Gaia* in the future.

Before launching into the chapters detailing my work over the past three and a half years, it is worth first conducting an overview of the science behind OCs in the introduction to this thesis. In Sect. **??**, I will discuss the history of OC observations up to before the release of *Gaia* DR2 in 2018. Section **??** will then discuss the stunning data of *Gaia* and how it has already thoroughly revolutionised our understanding of OCs in just a handful of years. Section **??** reviews the current issues with the OC census and discusses the broad aims of this thesis. Finally, Sect. **??** will briefly discuss some key concepts from both observational and theoretical studies of OCs, providing background that will assist with the reading of this thesis.

The nomenclature and definition of star clusters varies throughout the literature. Hence, in the next section, I will discuss a working definition of OCs that I will adopt throughout the rest of this work.

1.2 The definition of an open cluster

There are many different types of star cluster in the universe. Avoiding confusion when talking about star clusters is important, particularly since observers and theorists often use very different nomenclature. Definitions of star clusters can differ significantly even in observational communities when comparing between galactic and extra-galactic astronomy. Hence, before going any further, it is important to define exactly what I will be discussing in this thesis; I will use the following definitions consistently throughout this thesis for clarity.

This thesis will almost exclusively discuss clusters observed in the Milky Way, which are traditionally divided into three broad categories. I will primarily discuss open

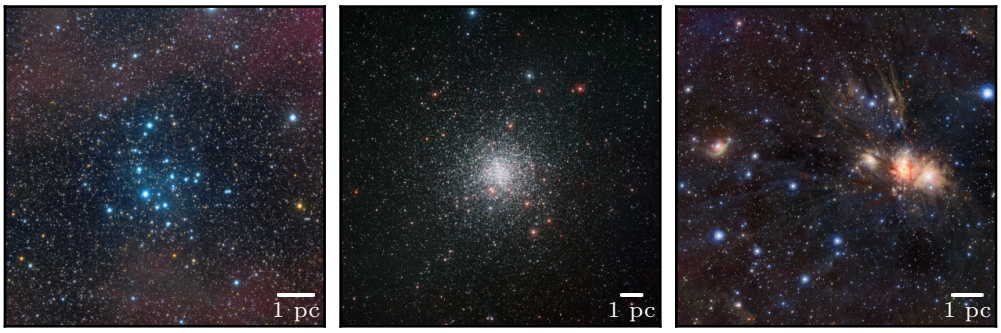

Fig. 1.1: A visual comparison between the three main types of star cluster found in the Milky Way. *Left:* the open cluster NGC 2547. *Middle:* the globular cluster M4. *Right:* the moving group/OB association Monoceros R2. All images contain a scale in the bottom right showing a length of 1 pc at the distance of each cluster. *Credit, left to right:* ESO / J. Pérez; ESO; ESO / J. Emerson / VISTA.

clusters, although I will also touch on globular clusters and moving groups. I differentiate between these three types of cluster as follows, matching the observational definitions in **portegies_zwart_young_2010**.

Open clusters (OCs) are gravitationally bound clusters with a typical age of around 100 Myr, although some are older than 1 Gyr and some are as young as 0.1 Myr. OCs have masses of typically no greater than 10^4 $M_\odot$ and may be made up of a few dozen to a few thousand stars, with a typical minimum being ten stars. OCs are remants of recent star formation, and are hence predominantly located in the galactic disk where the star formation rate is highest. Most OCs have a size of around 3 to 10 pc. Other than some rare, potential exceptions, OCs contain a single population of stars.

Globular clusters (GCs) are much older and more massive gravitationally bound clusters, with ages typically greater than 10 Gyr and masses typically greater than 10^5 $M_\odot$. The largest GCs can contain a million stars or more. GCs have a typical size of around 10 to 20 pc. GCs tend to reside in the galactic bulge or in the galactic halo. Many GCs contain multiple populations of stars. Almost all OCs have masses significantly lower than the typical present day mass of GCs, although observations of a handful of young massive clusters in the Milky Way such as Westerlund 1 (sometimes also referred to as 'super star clusters') as well as observations of galaxies with more active star formation suggest that the highest mass star clusters will be long-lived, and may evolve into GCs. However, this is not the case for almost all OCs that I will study in this thesis, as the only young massive clusters in the Milky Way are generally distant, heavily reddened, and outside of the reach of the visual-band observations of the *Gaia* telescope.

Tab. 1.1: Approximate definitions for the three types of star cluster that will be discussed in this thesis.

Type	Bound?	Age	Mass	Location
Open cluster (OC)	Weakly	$\lesssim 1$ Gyr	$\lesssim 10^4$ $M_\odot$	Disk
Globular cluster (GC)	Strongly	$\gtrsim 10$ Gyr	$\gtrsim 10^5$ $M_\odot$	Halo/Bulge
Moving group (MG)	No	$\lesssim 50$ Myr	$\lesssim 10^3$ $M_\odot$	Disk

Moving groups (MGs) are of a similar mass and number count to OCs, except they are not gravitationally bound. Due to this, they disperse much more quickly, and hence often have much younger ages. MGs have the widest definition, and encompass any group of stars that are comoving and coeval, but are specifically *not* gravitationally bound. Many MGs are referred to as 'OB associations' in the literature, due to them often containing a number of young, high mass O and B stars.

These definitions are summarised in Table **??** and compared visually in Fig. **??**. The figure shows three clusters; NGC 2547, M4, and Monoceros R2. NGC 2547 is a sparser OC that has a clear core of young blue stars at its center, about ~ 1 pc across. On the other hand, despite being only slightly larger, the GC M4 clearly contains significantly more stars. The stars in M4 are older, with the cluster having a whiter appearance along with more evolved red giant stars. Finally, the MG Monoceros R2 is simply a group of young blue stars, with no discernible core.

1.3 The pre-*Gaia* history of open cluster observations

While the results of this thesis are entirely derived using data from *Gaia*, to truly understand just how groundbreaking the current data of the *Gaia* satellite is, it is worth first briefly reviewing the history of OC observations.

1.3.1 Open clusters up to the 20th century

Our ability to observe OCs has progressed incredibly far throughout the history of astronomy (Fig. **??**). The invention of the refracting telescope allowed for early astronomers such as Galileo to observe that OCs and GCs are in fact clusters of many stars, as opposed to being dispered single sources as previously believed from unaided observations. It was, however, the invention and widespread adoption

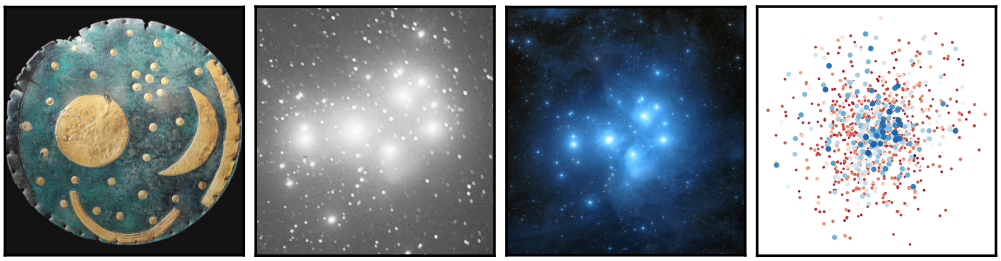

Fig. 1.2: The Pleiades, as depicted throughout history and showing the clear improvements in astronomical data gathering over time. *Left:* the Nebra Sky Disc, depicting the Pleiades with its seven naked-eye visible stars in the upper center. The disc was discovered in 1999 in northern Germany and is dated to between 1800-1600 BC. *Middle left:* the Pleiades, as imaged in 1909 with Wolf's Doppelastrograph at the Landessternwarte Heidelberg-Königstuhl. *Middle right:* the Pleiades, as imaged by Hubble. *Right:* the ∼1000 member stars for the Pleiades extracted from *Gaia* DR2 data and isolated from field stars by **cantat-gaudin_characterising_2018**. Each star is represented by a point scaled by its magnitude and coloured according to its $BP - RP$ colour. *Credits:* Frank Vincentz; Heidelberg Digitized Astronomical Plates; Davide De Martin & NASA/ESA Hubble.

of the reflecting telescope in the 17th and 18th centuries that led to catalogues of clusters like we use today.

The power of reflecting telescopes allowed astronomers to scan the sky to signficantly greater depth, searching for clusters of stars and discovering many new objects in the process (**herschel_catalogue_one_1786**), with the number of known OCs jumping from a few dozen to around 700 in a little over a century. Figure **??** shows the evolution in size of OC catalogues over time, showing the rise to around 700 clusters by the turn of the 20th century. Many of the OCs known and catalogued by astronomers at this point were some of the largest and most scientifically useful, with many of these OCs (especially those in the NGC catalogue) being some of the most frequently studied objects even today.

The 20th century saw improvements to data gathering and techniques, with early photometric and spectroscopic methods allowing authors such as **rosenberg_uber_zusammenhang_1910** and **hertzsprung_ueber_verwendung_1911** to plot the brightness of the stars in the Pleiades and the Hyades against their spectral features, noticing for the first time that the brightness of stars is related to their colour and spectral features. **russell_relations_spectra_1914** derived the absolute magnitude of stars in the Hyades and ploted this against an early spectral analogue of the temperature of its member stars, plotting the luminosity of stars against their temperature for the first time and inventing 'Hertzsprung-Russell' or 'colour-magnitude' diagrams (CMDs), a type of plot still used extensively in the present day as an essential tool to understand

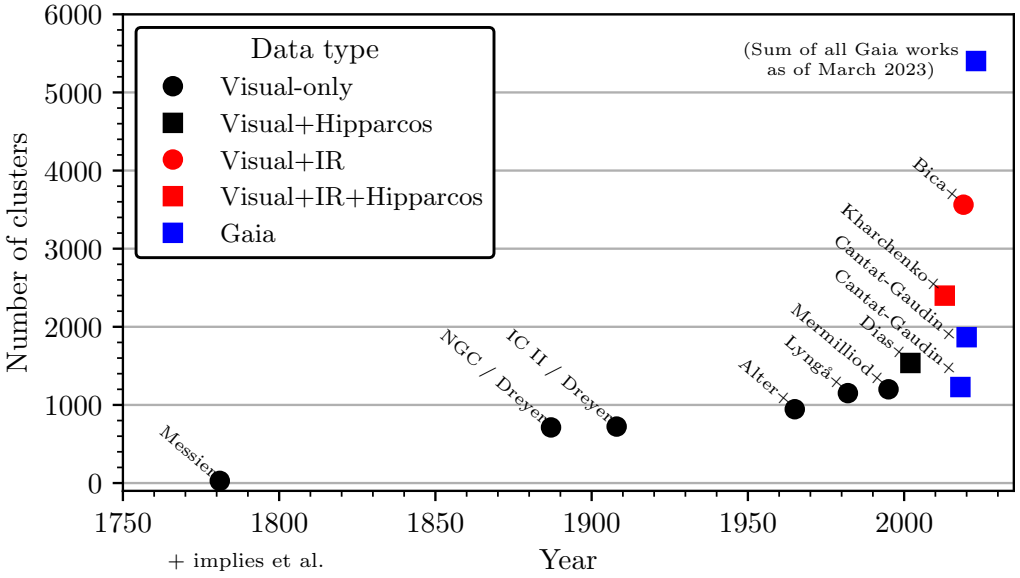

Fig. 1.3: The size of OC catalogues over time. After the initial rise in the size of catalogues due to the advent of reflecting telescopes in the 18th and 19th centuries, it was not until the past 25 years and the advent of large-scale astrometric and IR datasets that the OC census significantly increased in size.
N.B.: this is not an exhaustive plot of all catalogues, and a number of old catalogues such as **herschel_catalogue_one_1786** and **herschel_general_catalogue_1864** without digitised versions are not included.

stellar evolution. Later, the differences in CMDs between different clusters were noticed. This was interpreted as being a difference in age between the clusters, allowing for the ages of stars within star clusters to be estimated, and beginning the foundation of our knowledge of stellar evolution (Fig. **??**).

While the 20th century saw huge strides in our understanding of stars and star clusters, the size of OC catalogues went relatively unchanged (Fig. **??**). It was not until the 1990s and the arrival of new methodologies that the OC census itself has begun its largest upheaval since the widespread adoption of reflecting telescopes more than 200 years prior.

1.3.2 The advent of modern astrometry and infra-red datasets

The launch of the *Hipparcos* satellite and subsequent data releases (**perryman_hipparcos_1997**) produced a catalogue of around 10^5 sources with five-parameter milliarcsecond-precision astrometry. OCs stand out as overdensities in *Hipparcos* data, in particular

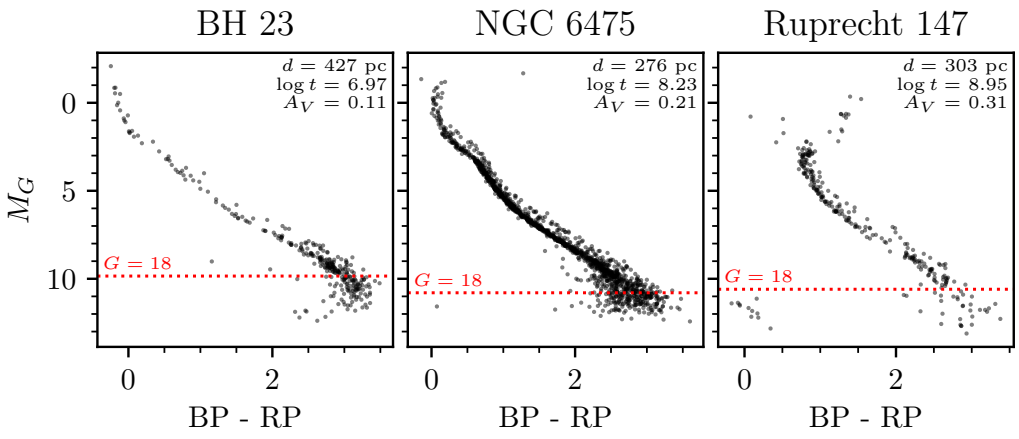

Fig. 1.4: A comparison of the CMDs of a number of nearby OCs, using membership lists from later in this thesis in Sect. **??** and plotted with their absolute magnitude M_G against colour $BP - RP$. The OCs are plotted from left to right in order of increasing age, with their distance d, logarithmic age $\log t$ and extinction A_V shown in the top right. The dashed red line indicates the approximate 100% completeness limit of these OC membership lists, with sources fainter than an apparent magnitude of $G = 18$ frequently being missed and often having underestimated $BP - RP$ colours. BH 23 is less than 10 Myr old and has almost no main sequence turn off; NGC 6475 is over 100 Myr old and has a clear turn off; Ruprecht 147 is around 1 Gyr old and even has a clear population of white dwarf stars.

in proper motions, as OCs are comoving groups of stars that often have different velocities to background field sources. This new data allowed works such as **platais_search_1998** to discover a number of new OCs, with many being small objects near to the Sun that evaded detection with only two-dimensional visual observations.

The catalogue of **dias_new_2002** included over 300 more objects than the roughly ten years prior catalogue of **mermilliod_database_1995** (Fig. **??**), representing the largest major jump in the size of the OC census in over a century, in addition to the much more accurate mean cluster proper motions and parallaxes provided by *Hipparcos*. However, this was just the beginning, and more new science was to come.

Data releases from the Two Micron All Sky Survey (**skrutskie_two_2006**) in the 2000s provided the next major jump in data availability for furthering OC science. The infrared (IR) data of 2MASS and its associated catalogue of 471 million point sources allowed works such as **dutra_new_2001**, **dutra_new_infrared_2003**, **bica_new_infrared_2003**, and **froebrich_systematic_2007** to uncover over a thousand new OC candidates in the galactic disk, using IR data to peer through inster-

stellar dust and unveil many previously-obscured objects for the first time. In addition, works around this time began to make increasing use of advances in computing power, with works such as **froebrich_systematic_2007** using automated retrieval to extract cluster candidates instead of simply scanning datasets by eye for overdensities.

Work predominantly with IR data culminated in the catalogue of **kharchenko_global_2013**, who derived homoegeneous membership lists, ages, extinctions, distances, proper motions, radii, and many other parameters for a total of 3006 clusters, 2399 of which are OCs or probable OCs, using a combination of 2MASS data and astrometric data from the PPMXL catalogue of proper motions (**roeser_ppmxl_catalog_2010**).

In around 20 years, the OC census more than doubled in size between the work of **mermilliod_database_1995** to the work of **kharchenko_global_2013**. This unprecedented shift represented the first time that the OC census had been significantly expanded in over a century, with improved datasets offering significantly better measurements of more clusters than ever before.

Yet the seismic shift in cluster catalogues brought about by IR datasets and *Hipparcos* was scarcely the beginning of the modern revolution in studies of OCs. *Gaia*'s first full data release in 2018, DR2 (**brown_gaia_2018**), sparked the next revolution in the census of OCs.

1.4 The *Gaia* revolution

For almost all of the history of astronomy, our view of the Milky Way has been strictly two-dimensional. Observing a three-dimensional galaxy in two dimensions is inherently limiting; it took until the 20$^{\text{th}}$ century to even discover that galaxies are separate from the Milky Way (**curtis_novae_spiral_1917**). Although astrometric parameters like parallaxes have been measured for stars for over a century, and can be used to view the stars of the galaxy in three dimensions, these datasets have always been limited to a few hundred or thousand stars until very recently.

1.4.1 Background on the *Gaia* satellite

Gaia is a space-based telescope launched in 2013 that aims to measure a wealth of parameters to an unprecedented level of precision for around 10^9 stars. *Gaia* is measuring precise positions, proper motions, parallaxes, and photometry for its full

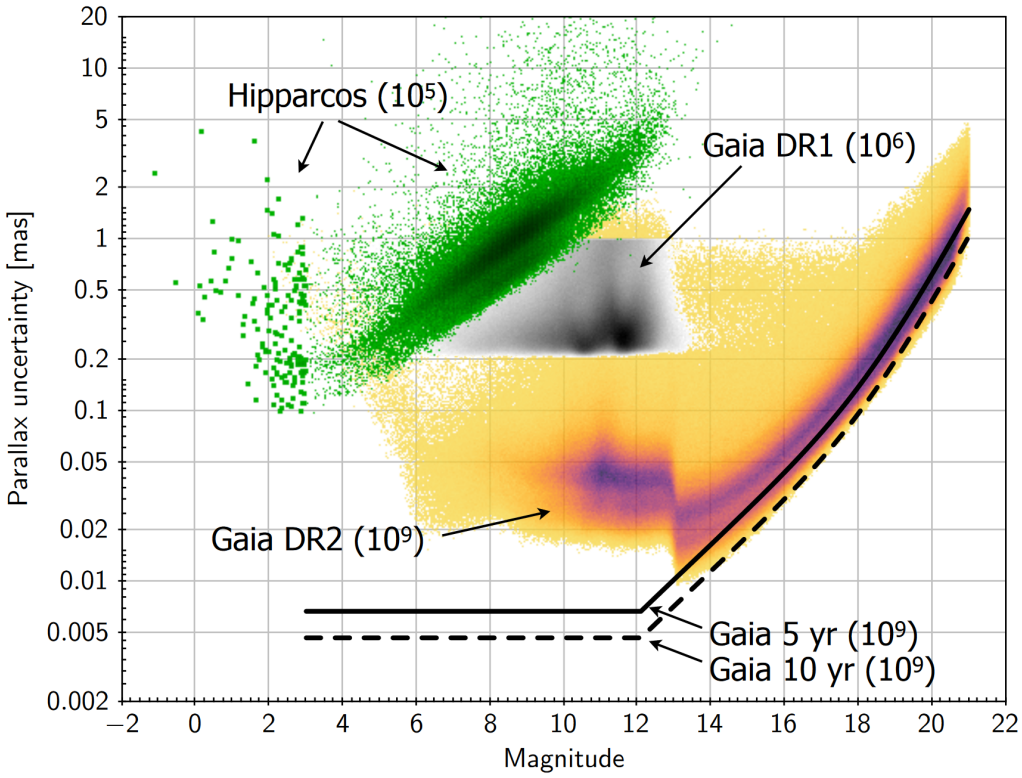

Fig. 1.5: Comparison between the astrometric accuracy for all sources in the final data release of Hipparcos, *Gaia* DR1, and *Gaia* DR2. The predicted accuracy of future data releases using 5 and 10 years of data is shown by the solid and dashed lines respectively. *Credit*: *Gaia* DPAC.

sample of stars, and also measures radial velocities and low-resolution spectra for brighter subsamples of sources (**gaia_collaboration_gaia_2016**). It is the incredible scale and precision of *Gaia* data that sets it apart from any previous datasets.

Figure **??** shows a comparison of the parallax uncertainty of *Gaia* data against data from the *Hipparcos* satellite. The difference in accuracy and quantity of data is clear: *Gaia* can measure parallaxes for 10^4 times as many stars at a projected eventual accuracy as much as 10^3 times better than *Hipparcos*. Inevitably, such a large increase in the amount (and quality) of data has huge implications for the study of all objects in the Milky Way, of course including OCs.

To truly understand the wonder of the *Gaia* satellite, it is first worth discussing how exactly it works. Although our galaxy is a dynamic system, with stars continually orbiting around the centre of the Milky Way (**binney_galactic_1987**), it is exceptionally difficult to capture the movement of our galaxy in real time. To the human eye, the night sky is static; even the closest stars with the highest proper motions and parallaxes have movements across the sky measured in arcseconds, with one arcsec-

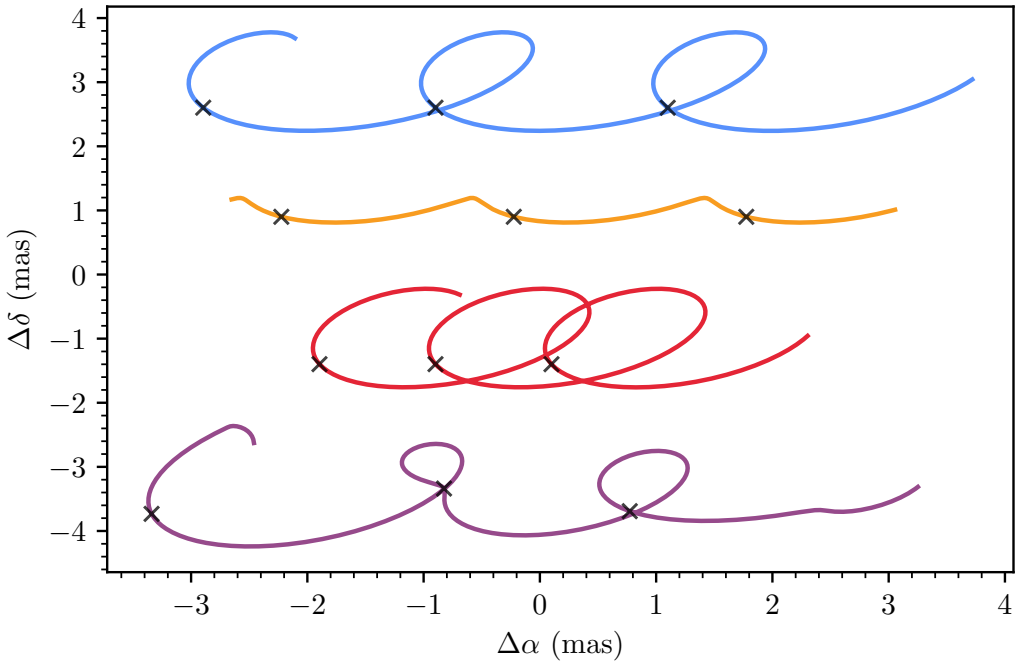

Fig. 1.6: The predicted on-sky astrometric tracks of stars with different parameters, generated using astromet (**penoyre_astrometric_2022**). All sources are at coordinates $\alpha, \beta = (0°, 45°)$, but are offset in the y direction for clarity of plotting. The first source has $\mu_{\alpha^*} = 2$ masyr^{-1}, $\mu_\delta = 0$, and is at a distance of 1 kpc. In the second example, the distance is quadrupled relative to the first. In the third example, the proper motion is halved relative to the first. In the final example, a binary with a period close to 1 yr, high eccentricity, and a low light ratio is added to the first example, producing a highly irregular track. The crosses denote the position of each source in one-year intervals.

ond being equivalent to just $1/3600$ of a degree. For stars at a distance of, say, 1 kpc, their parallax will amount to just 1 mas. With the Milky Way having an estimated radius of between roughly 15 to 25 kpc (**lopez-corredoira_disk_stars_2018**), it is clear that measuring precise astrometry for even a small fraction of the stars in the galaxy requires an incredible level of precision.

Using techniques originally pioneered with the *Hipparcos* satellite, *Gaia* operates quite unconventionally relative to traditional 'point and take a picture' telescopes. Instead, *Gaia* gathers data by rotating at a rate of exactly 1° per minute, spreading point sources into lines on its detector which are then processed into sources at a given location. Coupled with the field of view of the telescope, this scanning pattern means it visits every location on the celestial sphere around 14 times a year, allowing the complicated track of sources across the sky to be reconstructed to an exceptionally high level of precision for around 1 billion sources (see Fig. **??**). *Gaia*'s controlled rate of rotation, its view of the cosmos undisturbed by atmospheric distortion, and

its precise, modern detectors allow for *Gaia*'s revolutionary measurements to be possible (**gaia_collaboration_gaia_2016**).

1.4.2 The *Gaia* impact on the census of OCs

With so much data at an incredible level of quality, it is perhaps unsurprising that the OC census has been completely overhauled in just five years since the first full release of *Gaia* data (*Gaia* DR2). In many ways, *Gaia* is the perfect instrument for the study of OCs. Most OCs (such as NGC 2547 in Fig. **??**) have relatively low star counts and are situated on the galactic disk, where high numbers of field stars are present – making them challenging to isolate from background sources (**kharchenko_global_2012**). However, as OCs are comoving groups at a similar distance to one another and denser than the surrounding field, *Gaia*'s proper motions and parallaxes provide an excellent way to isolate clusters from the field (**gaiacollaboration_gaia_data_2017**; **cantat-gaudin_characterising_2018**).

Figure **??** shows the region around the high galactic latitude OC Blanco 1 in data from the Hipparcos-2 catalogue (**vanleeuwen_hipparcos_new_2007**), compared against the same region in data from *Gaia* DR3 (**gaia_collaboration_gaia_2022**). The difference in precision between the two datasets is dramatic. In *Hipparcos*, while the cluster is visible as an overdensity in proper motion space, in *Gaia*, the cluster becomes a small, compact group of stars that is trivially easy to separate from field stars. In addition, while Hipparcos parallaxes have accuracies on the order of 1 mas, *Gaia*'s $\sim 100\times$ better parallaxes make the cluster stand out as a clearly visible horizontal line as a function of right ascension. Combined together, proper motions and parallaxes make an exceptionally powerful tool to isolate OCs from field stars and derive clean, minimally contaminated membership lists. In addition, the difference in dataset size between the two telescopes is abundantly clear: *Gaia* DR3 can be used to probe the cluster eight to ten magnitudes fainter than *Hipparcos*, resulting in a membership list around $\sim 50\times$ larger than the cluster in *Hipparcos* data. This incredible level of astrometric precision is repeated across the entire galactic disk, and has powered the last five years of revolution in the OC census.

Not long after the release of *Gaia* DR2, **cantat-gaudin_gaia_2018** produced an updated catalogue of OCs and OC membership lists, using pre-*Gaia* works such as **kharchenko_global_2013** as input and trying to redetect their catalogued clusters in *Gaia* data. **cantat-gaudin_gaia_2018** were able to derive updated cluster membership lists with around twice as many members on average as in **kharchenko_global_2013**,

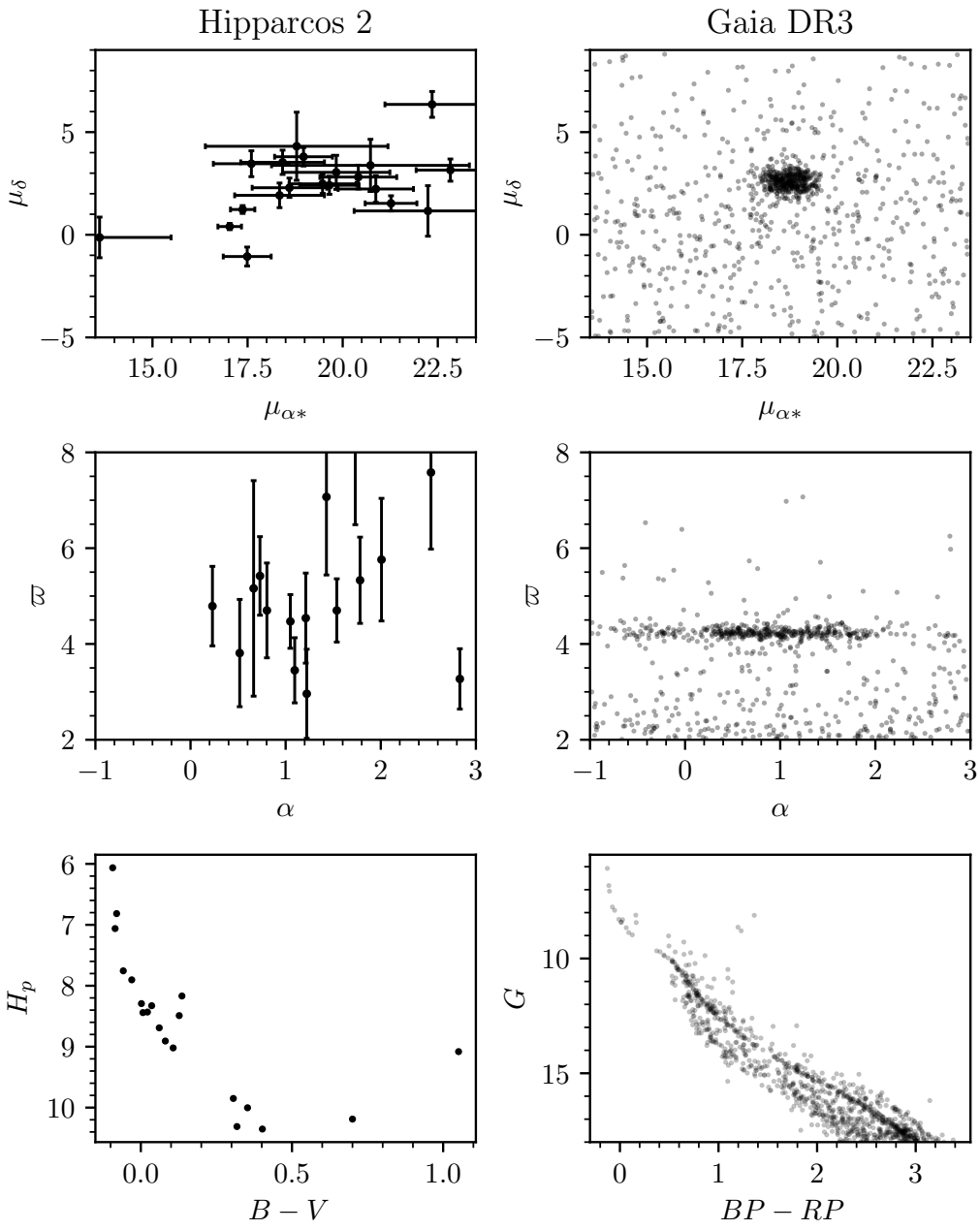

Fig. 1.7: Comparison between the regions around the star cluster Blanco 1 in data from *Hipparcos* and *Gaia*. *Hipparcos*-2 data (**vanleeuwen_hipparcos_new_2007**) is shown on the left and *Gaia* DR3 data (**gaia_collaboration_gaia_2022**) is shown on the right. The top row shows proper motions, the middle row shows parallax as a function of right ascension, and the bottom row shows the CMD of the stars in each region. While *Hipparcos* only sees a few dozen bright stars for the cluster, *Gaia* can detect up to 1000, and to a significantly higher degree of astrometric accuracy.

as well as deriving cluster proper motions to around two orders of magnitude greater precision than in **kharchenko_global_2013** and precise distances to clusters.

One of the largest results of the work on the OC census so far in the era of *Gaia* has been that many clusters catalogued before *Gaia* cannot be detected in *Gaia* data. This is clear in Fig. **??**, with the catalogue of **cantat-gaudin_gaia_2018** containing around half as many clusters as **kharchenko_global_2013**. The reasons for the non-detection of such a large number of OCs remain mostly unclear.

cantat-gaudin_clusters_2020 provided some answers to this question, searching again in *Gaia* DR2 data for some of the clusters they were unable to detect. They found that many clusters reported earlier in IR datasets continued to be undetectable in *Gaia*, being able to strongly rule out 38 objects as definite asterisms. The asterisms they found are generally older and at high galactic latitudes, and were typically reported in IR datasets. They comment that although *Gaia*'s visual observations should mean some clusters are too heavily reddened to be visible to *Gaia*, there are nevertheless many objects that *Gaia* should still be able to detect, owing to its deep visual photometry and ease of separating OCs from the field. However, the status of at least another ∼1000 clusters remains unknown, with the exact reasons for their non-detection in *Gaia* being only speculation at this time.

1.4.3 New open clusters found with *Gaia*

At the same time that *Gaia* has been an invaluable tool for better cataloguing already-known OCs, *Gaia* has also allowed for a large number of new OC discoveries; particularly for smaller, sparser objects that are otherwise impossible to find in 2D datasets (**cantat-gaudin_milky_2022**). Figure **??** shows the approximate number of papers reporting new OCs in the 21st century. Papers were found by searching the ADS[1] in Februrary 2023 for papers whose title or abstract contained the string '`new open cluster`'. The release of *Gaia* DR2 in 2018 (**brown_gaia_2018**) clearly corresponds with the number of papers reporting new OCs each year roughly tripling.

The central challenge of finding new OCs in data from *Gaia* is the sheer size of the *Gaia* dataset, with hundreds of millions of stars to search through in a five-dimensional dataset of positions, proper motions and parallaxes for each star. While traditional approaches in the 19th and 20th centuries searched for clusters by hand, and works such as **froebrich_systematic_2007** refined this approach by using kernel

[1] `https://ui.adsabs.harvard.edu/`

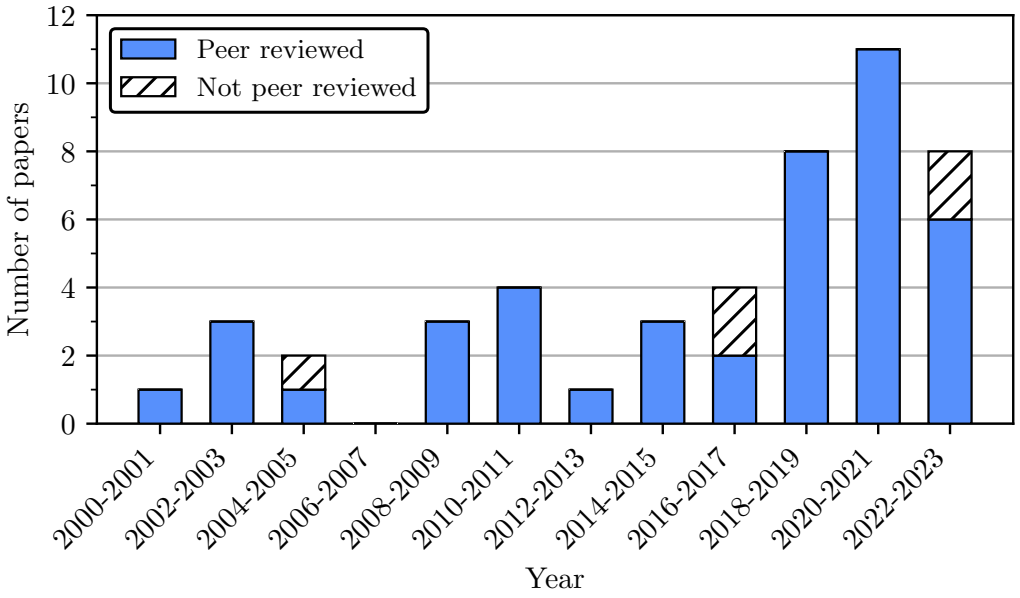

Fig. 1.8: The approximate number of papers reporting new open clusters in the 21st century, shown as a stacked bar chart of peer reviewed and non-peer reviewed works. Data for 2022-2023 are incomplete.

density estimation to identify overdense regions in the two-dimensional 2MASS dataset, the release of *Gaia* has also seen many new approaches for OC recovery.

Machine learning (ML) has exploded into observational astronomy over the last decade, developing from a niche method into a mainstay of astronomical data analysis methods (**ivezic_statistics_data_2020**). ML has two primary appeals. Firstly, it mostly automates the solving of complicated problems. ML can learn the relationship between input data and a desired output largely autonomously, with the user only being responsible for checking its work. Especially for arduous tasks like classification of large datasets (**killestein_transient-optimised_2021**), ML-based approaches can be orders of magnitude more straightforward to implement than creating a brand new algorithm or approach to solve every problem every time, or by simply solving a problem by hand as would be done traditionally. In this way, ML methods can be considered a 'Swiss army knife' of model fitting, with every method being applicable to a very wide range of potential problems. Secondly, ML-based approaches are generally much quicker than previous methodologies (**hunt_improving_2021**), leveraging the latest computing hardware such as graphics processing units (GPUs) significantly more efficiently than previous approaches.

While ML is not without its caveats (which will be discussed later in this thesis), ML has still been essential to the dramatic increase in newly reported OCs in the *Gaia* era.

castro-ginard_new_2018 were the first authors to adopt an ML-based approach for OC recovery, using two kinds of ML to automate tasks in cluster searches. Firstly, they used a clustering algorithm called DBSCAN (a form of unsupervised ML) to recover 31 new OCs in *Gaia* DR2 data, automating the process of cluster retrieval. Then, they used a neural network (a form of supervised ML) to classify OCs based on their CMD, also automating the process of assuring that OC CMDs have single stellar populations. Aside from **sim_207_2019** and a handful of works where small numbers of new OCs were noticed by mistake (**zari_3d_2018**; **bastian_gaia_2019**; **anders_ngc_2022-1**), all of the other roughly two dozen papers over the past few years that have found new OCs have used ML techniques to search for clusters.

Since then, many other works have used DBSCAN or variations on it to detect new clusters, with it proving to be an extremely popular method in the literature for OC retrieval (**castro-ginard_hunting_2019**; **castro-ginard_hunting_2020**; **castro-ginard_hunting_2022**; **liu_catalog_2019**; **he_catalogue_2021**; **he_new_2022**; **he_unveiling_hidden_2022**; **he_blind_allsky_2022**; **qin_discovery_2021**; **hao_sixteen_2020**; **hao_newly_2022**; **qin_hunting_2023**). In total, these works have reported nearly 4000 new OC candidates, which – if all of these objects are real – presents a major expansion in the size of the OC census. A handful of other works have used different methods, including **cantat-gaudin_gaia_2019** who used Gaussian mixture models (GMMs) and **jaehnig_membership_2021** who used extreme deconvolution (a probabilistic extension of GMMs).

The discovery of so many new OCs has brought a number of exciting new results. In particular, before *Gaia*, works such as **kharchenko_global_2013** believed that the OC census of 955 objects within 1.8 kpc was largely complete; however, around $\sim$400 new OCs have been reported in this range by *Gaia*-based OC searches since the release of *Gaia* DR1, firmly challenging the idea that the OC census is complete at close distances and providing many new objects for study.

The most recent analysis of the completeness of the OC census in the *Gaia* era (**anders_milky_2020**) found that the OC census within 2 kpc remains incomplete, although the full extent of this incompleteness is still an open question. It is unknown how many new OCs are remaining to be discovered and if existing methodologies could be improved upon.

1.4.4 *Gaia*'s brand new insights into open clusters

Although this thesis will mostly focus on methods to further improve the census of OCs, fundamentally, the reason why OCs are thoroughly important to modern

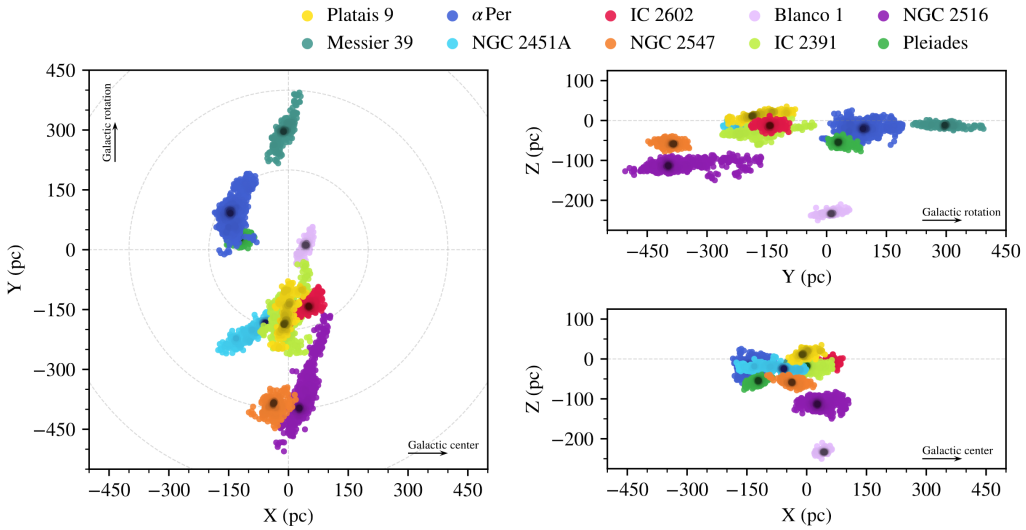

Fig. 1.9: The detected tidal tails and comas of ten OCs near to the Sun. Clusters are shown as coloured density plots and plotted in heliocentric coordinates with the galactic centre to the right. *Credit:* **meingast_extended_2021**

observational astronomy is the science that can be performed with them. Hence, I will also quickly discuss some of the main new results into OCs that *Gaia* data has enabled, giving an overview of the power and importance of these objects.

One of the most exciting results of the *Gaia* era is that the dissolution of OCs can now be observed. OCs have a typical age of around 100 Myr, which is significantly younger than the ≈ 13 Gyr age of the Milky Way, a difference that has long been argued as evidence that OCs are broken up by two-body interactions between stars ejecting some cluster members and the tidal forces of the Milky Way. Numerical simulations have shown that almost all OCs should have 'tidal tails' of stars stretching in front and behind the cluster's orbit due to such interactions with the Milky Way's potential (**portegies_zwart_young_2010**; **cantat-gaudin_milky_2022**), although such tidal tails had only been observed for GCs until *Gaia*. Now, thanks to *Gaia*, the detection and study of OC tidal tails and dissolution processes is possible in exquisite detail for dozens of clusters.

As the nearest OC to the Sun, the Hyades has been extensively studied, with its spatial elongation first being probed by **reino_gaia_study_2018** using *Gaia* DR1, and studied further by **lodieu_3d_view_2019**, **r\IeC {\"o}ser_hyades_tidal_2019**, and **meingast_extended_stellar_2019** with *Gaia* DR2. Similar analyses have been performed on many more clusters, with **meingast_extended_2021** analysing ten OCs in the solar neighbourhood and finding that not only do they all exhibit tidal tails, but most are also surrounded by 'comas' of stars ejected in all directions

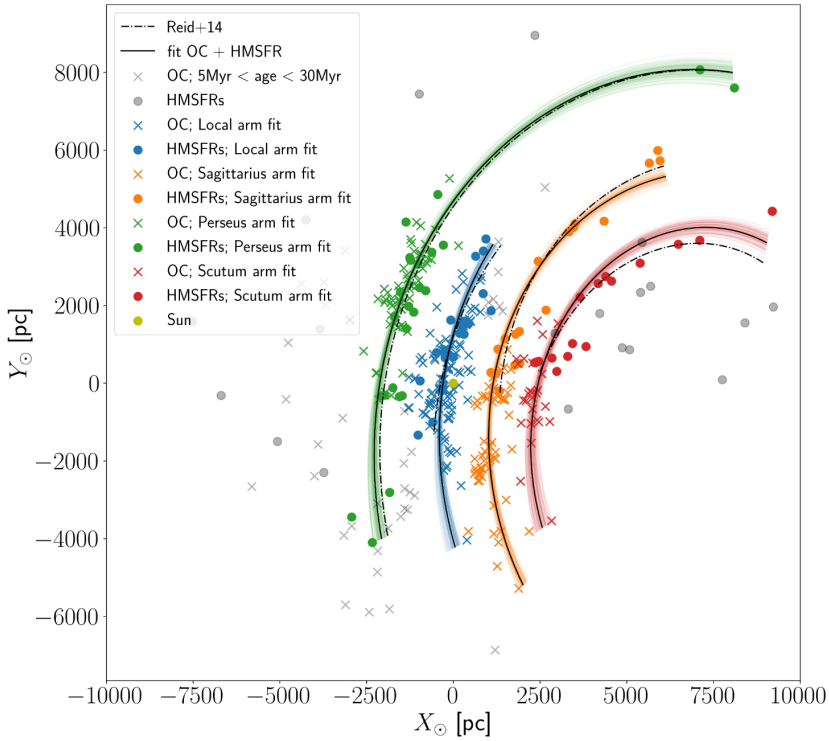

Fig. 1.10: A model of the Milky Way's spiral arm structure as traced by OCs and high-mass star forming regions. *Credit:* **castro-ginard_milky_2021**

from each cluster (Fig. **??**). **tarricq_structural_2022** studied 369 clusters within 1.5 kpc and detected tidal tails for 71 of them. Such clear visibility of the ongoing dynamical destruction of Milky Way OCs has been used by works such as **yeh_ruprecht_2019**, **oh_kinematic_modelling_2020**, and **pang_3d_2021** to study the dynamics of nearby OCs and make predictions on their future lifespan. It should be possible to expand these methods to more OCs and derive dynamical parameters for a wide range of star clusters, making wide-ranging inferences about the life of star clusters after their formation.

OCs have also been extensively used to probe the wider structure of the Milky Way. *Gaia*'s improved parallax accuracy allows for more accurate distances to OCs to be derived, and the improved OC membership lists possible with *Gaia* allow for better determination of photometric parameters. **cantat-gaudin_painting_2020** derived ages, extinctions, and distances for around 2000 OCs, showing that young clusters are generally correlated towards low galactic altitudes and appear to loosely trace spiral arm models derived from masers in works such as **reid_trigonometric_parallaxes_2014**, while older clusters are more uniformly dispersed and can be found at higher altitudes above or below the galactic plane, suggesting that their orbits have evolved while they aged. **castro-ginard_milky_2021** used these results to perform fits of a

spiral arm model to a combination of the distribution of young OCs and star forming regions (Fig. **??**), finding that the addition of young OCs slightly changes the most likely spiral arm model relative to the fit of **reid_trigonometric_parallaxes_2014**.

Finally, new OC results in the *Gaia* era have allowed for a number of new studies of stellar evolution. In particular, many more exotic phases of stellar evolution can now be studied more easily thanks to *Gaia* OC membership lists, which allow for significantly easier separation of OC member stars from field contamination.

A primary hot topic within the literature is blue straggler stars (BSSs), which are stars near to the main sequence turn-off of a cluster that are bluer and brighter than would otherwise be expected (e.g. four stars to the upper left of the turnoff point of Ruprecht 147 in Fig. **??**). These stars are interesting cases of non-ideal stellar evolution, with leading theories stating that BSSs may be caused by mass transfer, dynamical mergers, or a combination of multiple processes (**boffin_ecology_2015**). While BSSs have been extensively investigated in GCs, *Gaia* has allowed for many new investigations of BSSs in OCs (**cantat-gaudin_milky_2022**), such as in **rain_blue_2020** who investigated BSSs in Trumpler 5, Trumpler 20, and NGC 2477, or **vaidya_blue_2020** who studied BSSs in a further seven OCs and found that BSSs are not mass-segregated in just two of of the seven clusters they studied. **leiner_census_blue_2021** investigated BSSs in 16 OCs and found that standard population synthesis techniques do not produce enough BSSs when compared to *Gaia* observations. They found that changes to assumptions about binary mass transfer somewhat rectify differences between observations and theoretical predictions, although they found that it still remains difficult to create the observed number of BSSs from current theories, suggesting that theories of BSS formation may still require additional physics.

Another hot topic within stellar evolution that is more easily investigated within star clusters is extended main-sequence turnoffs (eMSTOs). Initially observed only in Magellanic cloud clusters (**bastian_effect_stellar_2009**), *Gaia*'s improved contrast between cluster and field stars has allowed for eMSTOs to be observed in a number of Milky Way OCs (**marino_discovery_2018**). eMSTOs challenge traditional theories of star formation for smaller clusters such as OCs, as they could be explained by multiple stellar populations of a range of ages. On the other hand, simpler theories such as different rates of stellar rotation or even circumstellar dust are also competing theories to explain the existence of eMSTOs (**milone_multiple_2022**; **dantona_role_dust_2023**).

Finally, OCs have also been used to study and calibrate variable stars. In particular, Cepheid variable stars are a critical first rung on the cosmic distance ladder,

useful for finding accurate distances galaxies within a few Mpc of the Milky Way. Currently, tension in the Hubble parameter H_0 could be explained in number of ways, ranging from the dominant ΛCDM cosmological model being wrong to simply being a miscalibration of one or more rungs on the cosmic distance ladder. Hence, in this context, accurate calibration and study of Cepheid variables is essential to ruling out or confirming issues with Cepheids as the source of any H_0 tension, a task that multiple authors have used OCs to aid in. **breuval_milky_way_2020** used OCs hosting Cepheid variables to derive a new Cepheid period-luminosity relation (Leavitt law) and derive an updated value for H_0, finding that the Hubble constant could be revised to a lower value still in some tension with Planck CMB results (**planckcollaboration_planck_2018_2020**) when using *Gaia* astrometry and Cepheid OC members. Works including **medina_revisited_2021**, **zhou_galactic_2021**, and **hao_open_2022** have searched for more Cepheid variable stars within OCs to assist in the further study of Cepheids.

It goes without saying that all scientific use cases of OCs rely on the OC census being accurate, and are greatly improved by it being as complete as possible. Even though this brief review may have presented a 'rosy-eyed' view of the status of OC science in the era of *Gaia*, there remain many issues with the current status of the OC census, with many unanswered questions and barriers to easier usability of OCs for science. In the next section, I will discuss some of these problems at length, and briefly introduce how I will try to solve some of them in the rest of this thesis.

1.5 Issues and solutions for the open cluster census

As detailed in Sects. **??** and **??**, there has been a huge amount of recent scientific progress in the OC census and in the study of OCs as a whole. However, the *Gaia* era of OC science is still relatively new, with many more years of data releases being anticipated. Inevitably, this will allow for a huge range of new scientific studies into OCs (**gaia_collaboration_gaia_2022**).

To maximise the scientific potential of OCs, it makes sense to improve the census of OCs as much as possible, as well as developing 'future-proof' methodologies that can be applied to future *Gaia* data releases as well as the current ones. The issues with the census of OCs in the Milky Way can be divided into five broad topics that I will discuss next.

1.5.1 The issues with the open cluster census

Problem 1. The methods used to detect open clusters (and their biases)

As mentioned in Sect. **??**, the *Gaia* mission has provided the OC community with a tremendous quantity of data. However, until now, many different works have tried many different approaches for OC recovery (both for recovery of existing clusters and for blind searches), with no direct comparison having been done between different approaches. Additionally, modern computer science is fast-paced, particularly in the field of machine learning. Many different approaches exist for clustering data, only a handful of which have been trialed for OC recovery (**xu_comprehensive_2015**), despite the fact that publically available open-source implementations of these algorithms are often available and ready to use (**scikit-learn**).

This causes a number of problems. Primarily, it is unclear whether or not existing approaches are subject to biases. Particularly since almost all blind searches for OCs have used DBSCAN (Sect. **??**), it could be that a bias with the algorithm could prevent certain clusters from being detected depending on their age, distance, or other parameters, which may mean that a whole type of new OC has been as-yet undiscovered within *Gaia* data. There is no certainty that all OCs that *can* be detected *have* been detected with *Gaia*.

It is also unclear how many false positives current approaches produce. Most works do not include an estimate of how many of their reported clusters are real (**castro-ginard_new_2018**; **liu_catalog_2019**; **he_catalogue_2021**). It is not known whether or not it is safe to assume that the results of a clustering algorithm can always be trusted, and it is not known whether certain algorithms are more or less trustworthy.

Additionally, there are many quirks with the usability of current approaches. For instance, the comprehensive DBSCAN-based works of **castro-ginard_new_2018**; **castro-ginard_hunting_2019**; **castro-ginard_hunting_2020**; **castro-ginard_hunting_2022** adopted a sky tiling scheme that requires a large number of algorithm re-runs, resulting in a method that must be applied on a supercomputer (**castro-ginard_hunting_2022**). It is not known whether a more efficient approach that requires fewer computational resources and is easier to repeat on future data releases is possible. This is a particular issue as future *Gaia* data releases are likely to contain higher numbers of reliable

sources (**gaia_collaboration_gaia_2022**), meaning that current approaches will need to be ran on four to eight times as many sources[2].

Problem 2. The status of clusters discovered before *Gaia*

Of the many clusters discovered before *Gaia*, fewer than 50% have so far been re-detected in *Gaia* data (**cantat-gaudin_gaia_2018**; **cantat-gaudin_clusters_2020**). The fact that so many objects are missing from *Gaia*-based OC studies could represent a total paradigm shift in the census of OCs in the Milky Way, or it could be indicative of the limitations of *Gaia*. For every cluster, there are two possibilities.

In the case that an object is real but cannot be detected in *Gaia* data, such as for heavily reddened clusters discovered using IR datasets that are obscured by dust in *Gaia* data (**cantat-gaudin_clusters_2020**), such an object would be a sign of the incompleteness of the *Gaia* OC census. If a significant number of IR clusters are in fact real, then to study all known OCs, it would be necessary to use both *Gaia* and IR datasets simultaneously.

On the other hand, it is also possible that such objects are not real. *Gaia* has significantly higher astrometric accuracy than all previous astrometric catalogues, and *Gaia* should be sensitive to a large number of real OCs, even for those with intermediate levels of reddening (**cantat-gaudin_clusters_2020**).

While some studies have performed small investigations into clusters missing from *Gaia* on a case-by-case basis (**cantat-gaudin_clusters_2020**; **piatti_catching_2023**), the status of most objects is still unknown. It should be possible to rule out many OCs reported previously in the literature given a large enough study. Alternatively, if *Gaia* is in fact a major limitation in recovering many OCs discovered before *Gaia* using IR datasets, then different datasets would need to be used to study such objects. This would also be a further strong science case for *Gaia* follow-up missions such as the proposed *GaiaNIR* mission for near-infrared astrometry (**hobbs_gaianir_combining_2016**).

Problem 3. The status of clusters discovered with *Gaia*

At the same time as the aforementioned 're-detection crisis' of clusters reported before *Gaia*, thousands of new OC candidates have been reported in the literature. Most

[2]Calculated for *Gaia* DR3, which contains $\sim$250 million sources with $G \leq 18$, which is a commonly adopted cut; however, the final *Gaia* data release is projected to contain at least 1 billion sources (**gaia_collaboration_gaia_2016**).

of these objects have not been independently verified (**cantat-gaudin_milky_2022**), meaning that a large number of objects exist in the literature and may be being used for studies of OCs and galactic structure but without knowing which objects are or are not real (**anders_milky_2020**; **castro-ginard_milky_2021**). Given that so many clusters cannot be detected from recent works reporting new OCs before *Gaia*, with some works such as (**scholz_global_2015**) having as many as 100% of their clusters being impossible to redetect (**cantat-gaudin_gaia_2018**), it is not far-fetched to suggest that there can be reproducibility issues between different studies when reporting new objects. Hence, there is a need to independently verify new OC candidates reported recently using *Gaia* data, preferably also with an alternative methodology and a thorough analysis of which objects are and are not real.

Additionally, it is also possible that some objects reported recently are duplicates. The large number of papers reporting new OCs since the release of *Gaia* DR2 (Fig. **??**) can make the literature difficult to keep up with (**cantat-gaudin_milky_2022**). There is likely a need to verify that new cluster candidates are unique and have not been previously reported in the literature. For instance, during the writing of this thesis, **chi_blind_search_2023** (accepted in ApJS) reported 1179 new OCs, which would represent a large increase in the number of newly discovered OCs in the *Gaia* era. However, many plots of their 'new' clusters are clearly compatible with OCs previously reported in the literature (e.g. candidate 14677, which is Blanco 1). The existence of works containing duplicates 'muddies the water' when attempting to use existing catalogues of OCs in combination with papers reporting new objects. There is a clear need to verify that newly reported OCs are real, unique clusters.

Problem 4. The completeness of the *Gaia* open cluster census

Despite the publication of many works that have used *Gaia* data to report thousands of new OCs (Sect. **??**), it is still unclear how complete the *Gaia* census of OCs is. It is not clear how many objects are missing or if any further biases contribute to certain objects being missed. Beyond the widespread disproving of the result in **kharchenko_global_2013** that the OC census is complete within 1.8 kpc, there has been little study in the *Gaia* era on the completeness of the OC census.

Nevertheless, the completeness of any catalogue, not least the OC census, is an interesting thing to know that would enable a large number of scientific studies. The study of star clusters in the Milky Way is unique in that we are able to study bound star clusters of significantly lower masses and luminosities than is possible in

extragalactic studies (**portegies_zwart_young_2010**). While extragalactic astronomy is able to probe the occurence rates of massive, highly luminous clusters in a large number of galaxies, it is only in the Milky Way that study of low-mass objects is possible, due to their low luminosity. Milky Way OCs are hence an important calibration point for understanding star formation at lower mass ranges.

Given that the Milky Way's cluster age and mass functions are uniquely important in the general study of star clusters, it is vital that the completeness of the OC census can be well known. Although this has been attempted in *Gaia* data by **anders_milky_2020**, who also derive a completeness estimate of the OC census, their work has two main limitations. Firstly, they used the blind searches of **castro-ginard_new_2018**; **castro-ginard_hunting_2019**; **castro-ginard_hunting_2020** to calibrate their completeness function. However, it is not known if the DBSCAN algorithm used in these works has any biases that their completeness estimate would inherit (see Problem 1/Sect. **??**). Secondly, they only create a selection function in terms of cluster age and distance. They expect that other parameters, such as cluster mass or size, could be major factors in the OC selection function. They were unable to include mass in their selection function due to the lack of cluster mass measurements in the *Gaia* era.

In addition, it is not clear how many more new OCs could be detected with future *Gaia* data releases. In theory, it should also be possible to extend such a prediction to other proposed surveys and instruments such as *GaiaNIR* (**hobbs_gaianir_combining_2016**). Given that the proposals for *GaiaNIR* describe science with OCs as a key scientific justification for the mission, a way to predict how many OCs would be discovered by a near-infrared astrometric mission such as *GaiaNIR* would be an interesting way to strengthen the science case for future astrometric missions and surveys.

Problem 5. The observational definition of open clusters

Finally, and somewhat amusingly, possibly the greatest issue with the OC census in the *Gaia* era is that no work can agree on what OCs actually are (at least observationally). While the theoretical definitions of star clusters in the Milky Way can now be reasonably clearly defined (**portegies_zwart_young_2010**), it is challenging to convert these theoretical definitions into a firm observational definition for OCs. Critically, this presents a number of issues when comparing between different works or when trying to combine the results of separate OC studies.

Most works reporting new OCs use different quality criteria to decide which objects are or are not included in their work. Almost all use some sort of criteria on colour-magnitude diagrams, requiring that the cluster CMD is narrow and compatible with a single stellar population. However, this is implemented in many different ways; including by using statistical criteria (**liu_catalog_2019**), a neural network classifier (**castro-ginard_new_2018**), or simple manual classification (**he_catalogue_2021**). Some works require that clusters are clear statistical overdensities in *Gaia*, deriving something analogous to a signal to noise ratio (S/N) for their cluster candidates (**cantat-gaudin_gaia_2019**). Some works also limit clusters based on their physical parameters, requiring that they are compact groups and hence more likely to be gravitationally bound (**liu_catalog_2019**). Finally, all works adopt a different minimum size for an OC, ranging from as low as 8 stars (**castro-ginard_new_2018**) to as high as 50 (**liu_catalog_2019**). With so many differing definitions of what constitutes a good enough OC candidate, it can be difficult to compare the results of multiple works or to combine them into singular catalogues without introducing biases. In addition, most works use simple binary 'yes/no' cuts on whether or not an OC passes a given constraint, which may not capture all of the uncertainty inherent in deciding whether an edge-case object is or is not a real OC.

cantat-gaudin_clusters_2020 outlined a set of empirical criteria to follow that all new OC candidates should meet, requiring that OCs are a clear overdensity in astrometric data, that they have a CMD with a clear homoegeneous population of stars, and that the cluster meets two cuts on its parameters intended to be a comparable test for being bound: that the radius containing 50% of members is smaller than 20 pc, and that the cluster's proper motion dispersion corresponds to an internal velocity dispersion of less than 5 kms^{-1}. While these criteria are an empirical minimum for an OC, as a thought experiment, it is still relatively straightforward in a dense region of the galactic disk to find $\sim$10 or more stars within 40 pc of each other and with a velocity dispersion below 5 kms^{-1}, and so these criteria are not infallible, and could allow unbound moving groups to be misclassified as OCs.

A 'gold standard' observational definition of an OC might be more directly derived from the theoretical definition presented in Sect. **??** – requiring that an OC is an overdensity, a single stellar population, and is unambiguously gravitationally bound. However, no such way to measure such parameters for a large number of clusters exists, principally due to the difficulty in measuring the dynamics and boundness of a large catalogue of star clusters.

1.5.2 The aims of this thesis

Even relative to the major improvements to OC science in the 1990s and 2000s, *Gaia* has still been utterly groundbreaking in the quality and quantity of data it provides on our galaxy. Never before has so much precise data been available for so many stars; *Gaia* will rewrite textbooks on the composition and characteristics of the Milky Way. Within the field of OCs, this is clear from the many incredible new *Gaia* results highlighted in Sect. **??**. Yet as the many problems discussed in the previous section show, many issues remain with the census of OCs. In this thesis, I hope to showcase timely research that can present solutions or partial solutions to the above problems, developing the methods used to analyse OCs in the *Gaia* era to maximise the scientific potential of these objects.

First and foremost, it is impossible to solve many of the other issues in the OC census without an understanding of the limitations and biases inherent to different methods for OC retrieval (Problem 1). In addition, numerous unexplored methods for cluster retrieval present in the computer science literature could provide better options for the recovery of OCs (**xu_comprehensive_2015**). To date, there has been no comparative study into the advantages and disadvantages of different approaches for cluster retrieval, despite the clear importance of understanding the limitations of different methods; hence, the first part of this thesis focuses on trialing different algorithms for OC recovery in *Gaia* DR2 data, performing a comparative study into their effectiveness. In this study, I will also aim to find optimal ways to divide *Gaia* data for OC retrieval, aiming to present a method that can be ran efficiently even on larger datasets, allowing for more sources to be incorporated in the future as *Gaia* data releases improve. In the best case scenario, a method can be found that can redetect the clusters reported by all other works with minimal bias.

With a best method found, other problems in the census will be more straightforward to solve. Finding the best methodology for OC retrieval and knowing its biases will allow for the application of the method to solve Problems 2 and 3, and to a lesser extent Problem 4. Specifically, in the second study of this thesis, I conduct a large-scale unbiased blind search for OCs using the best method found and data from *Gaia* DR3. Depending on which *Gaia*-discovered clusters can be found in this search and depending on the effectiveness of the method found in the first study, it will be relatively simple to solve Problem 3, as some clusters will (or will not) be possible to re-detect. This study will conduct the largest validation of OCs discovered using *Gaia* to date.

An unbiased all-sky search will also allow for a solution to Problem 2. Previously, studies have generally focused on small, case-by-case attempts to retrieve OCs that were originally catalogued in pre-*Gaia* works (**cantat-gaudin_clusters_2020**). However, an all-sky search ought to recover all OCs visible in *Gaia* within the limitations of the adopted methodology; given the understanding of this methodology gained from the first study, it should be possible to say with reasonable certainty whether or not *Gaia* should be able to detect many of the as-yet undetected OCs from before *Gaia*. This will greatly aid in bridging the OC census from pre-*Gaia* works to the *Gaia* era, tracing down any remaining missing clusters while suggesting that some are not real.

Inferring the completeness of the OC census is a major task, which this thesis will contribute towards but may not completely solve. An unbiased all-sky blind search for OCs is a good tool to find as many OCs as possible and reduce the incompleteness of the OC census. Despite the many works that have already searched for new OCs (Sect. **??**), there may still be many new objects left to discover. This search can be used as a drop-in replacement for the methodology of e.g. **anders_milky_2020**, serving as a better 'experiment' to detect a large sample of OCs. However, given that cluster masses are expected to be a major contributor to the selection function of OCs in *Gaia*, with less massive clusters being more difficult to detect, it will also be important to calculate accurate cluster masses for the entire sample of objects from the blind search. The final study of this thesis partly focuses on calculating cluster masses, which will help to solve Problem 4 while also deriving a generally useful parameter for OCs that has never been derived for such a large catalogue before. Cluster masses also require cluster ages, which I aim to derive estimates of in the second study to accompany the overall OC catalogue.

Finally, Problem 5 is likely to continue to plague the OC community for years to come, owing to the complexity of precisely defining OCs observationally. However, throughout this thesis, I will present new methods to try and convert a theoretical, first-principles oriented definition of an open cluster into a practical observational method to classify objects as OCs, MGs, GCs, false positives, or somewhere inbetween one of those categories. I aim to do so statistically, never presenting simple binary probabilities of an object being a false positive or one of the classes of real star cluster, but rather using a statistical treatment to aid in the definition of edge-case objects that could be between different classes. In the first study of this thesis, I augment clustering algorithms by trialing a number of different tests for the density of a cluster compared to its field, deriving a simple and efficient test of a cluster's astrometric signal to noise in *Gaia* data. In the second study of this thesis, I use an approximately Bayesian neural network to classify the likelihood of an OC being

compatible with a single stellar population given the predictions of stellar evolution models. In the final study of this thesis, I present a preliminary method to test the boundness of OC candidates and ascertain if they are a real bound object or simply an MG.

In total, with this thesis, I hope to contribute to the difficult task of cataloguing and characterising the OCs of the Milky Way in the era of *Gaia*. Implicitly, all of these methods will be 'future-proof' and applicable to future *Gaia* data releases, or future surveys that could replace the *Gaia* telescope. Inevitably, no method is perfect, and I will conclude by speculating on future avenues of research that could further develop the methods used to analyse OCs.

Before launching into the scientific content of this thesis, it is important to also present some theoretical background into OCs.

1.6 Further background into star clusters and associated common methods

To improve the reach and readability of this thesis, I feel it is important to review some common techniques and pieces of theory from the literature. For the seasoned open cluster astronomer, this section could be browsed quickly; for the non-specialist, I hope that this section provides more insight into pieces of theoretical knowledge that I will assume for the scientific parts of this thesis.

I begin by going into more depth on some of the most important methods to OC observers.

1.6.1 Analysis of CMDs

As discussed previously, CMDs are essential tools to derive many key parameters of a star cluster (Fig. **??**). The most common method to determine the age, extinction, and to a lesser extent the distance of a cluster is by fitting isochrones to cluster CMDs. An isochrone gives the predicted colour and luminosity of a population of stars with a range of masses given that the stars have the same age, extinction, composition, and distance. Stellar isochrones are derived from stellar evolution models such as PARSEC (**bressan_parsec_2012**) and are widely used in many areas of observational astronomy.

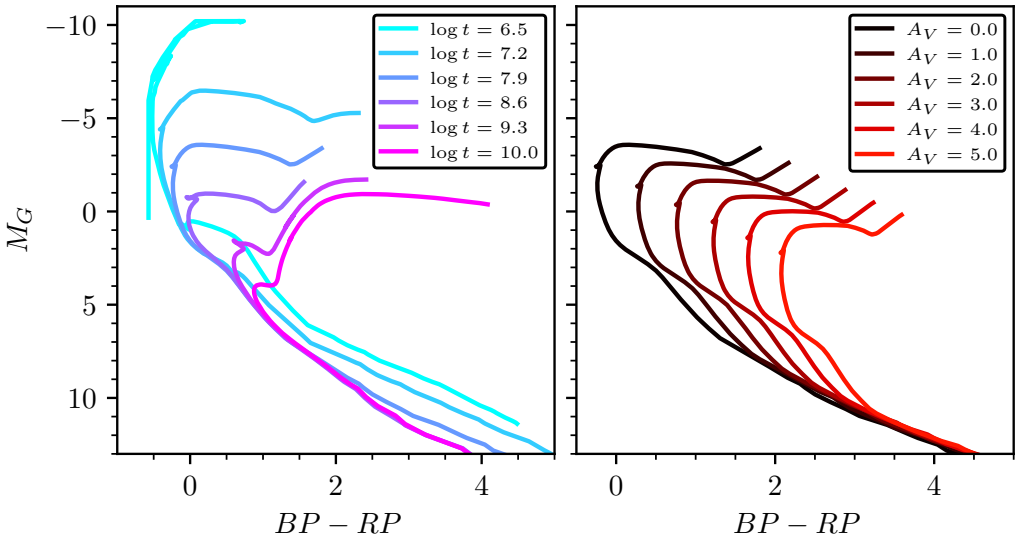

Fig. 1.11: A comparison between stellar isochrones of various different parameters, derived from PARSEC stellar evolution models (**bressan_parsec_2012**) and shown in *Gaia* photometric bands. *Left:* isochrones of solar metallicity and zero extinction shown for six different ages. Most noticeably, as cluster age increases, the magnitude of the turn-off point decreases, with ever-more stars evolving into red giants and eventually reaching the end of their lives. The rest of the stars in the cluster also move down slightly, relaxing onto the main sequence as they age. *Right:* the $\log t = 7.9$ isochrone from the left plotted at a range of different extinction values. Extinction reddens cluster stars as well as reducing their overall brightness. Extinction in *Gaia* photometry has a strong affect on the location of the turn-off point.

In practice, isochrones are difficult to fit, with age, extinction, distance, and metallicity all being somewhat degenerate with one another. Figure **??** shows the effect of varying age and extinction on stellar isochrones, with both age and extinction moving the location of the cluster turn-off point. Cluster distance merely shifts the isochrone up or down based on the cluster's distance modulus, although this is still slightly degenerate with age and extinction. Finally, the chemical composition of a cluster (most often parameterised with its metallicity [Fe/H]) has the smallest impact on cluster isochrones and is not shown, but will nevertheless slightly impact age and extinction determination.

Isochrone fitting is further complicated by the presence of other cluster features, such as blue stragglers, eMSTOs, or the presence of a binary sequence due to unresolved binaries (see binary sequences in Fig **??**, showing a clear second line of stars sat slightly above the main cluster population).

Probably unsurprisingly, there are hence many methods used in the literature to fit isochrones to data. Particularly as computational power is a major hindrance to

performing three or four-parameter fits with stellar isochrones, historically, it was common to simply fit isochrones by hand, which includes no robust uncertainty estimate and can open the door to human biases. For instance, the isochrones in **kharchenko_global_2013** were fit by eye, minimising χ^2 goodness-of-fit criterions manually. **yen_reanalysis_2018** developed this methodology further to perform χ^2 fitting of all cluster parameters autonomously. **hippel_inverting_2006** created a full Bayesian methodology to fit isochrones to cluster CMDs, which has been used by works in the *Gaia* era such as **bossini_age_2019**.

By far the main flaw of cluster isochrone fitting is speed. Three or four-parameter fits using complicated stellar isochrones simply cannot be performed quickly, requiring significant amounts of computation time to complete in e.g. **yen_reanalysis_2018**, making these key cluster parameters relatively time-intensive to derive using traditional isochrone fitting techniques. Alternatively, a recently developed approach used in **cantat-gaudin_painting_2020** and **kounkel_untangling_2020** uses neural networks trained on the results of a small subsample of precise isochrone fitting results for OCs to derive ages, extinctions, and distances to clusters. This uses significantly less computational time, although with the disadvantage that a subset of isochrone fits must be first created to then use as a training dataset.

1.6.2 Radial profiles

The physical size of OCs is another important property that can be measured. The size of observed clusters can be compared against theoretical predictions, and interesting relationships between parameters such as the size of clusters as a function of their age can be determined (**tarricq_structural_2022**).

The simplest commonly used measure of the size of an open cluster is the radius containing 50% of members, r_{50}, which is the median radius of all detected member stars from the cluster centre. This radius has been commonly measured in the literature for OCs (**cantat-gaudin_gaia_2018**; **cantat-gaudin_clusters_2020**). For a cluster where the mass of member stars is not correlated with their position in the cluster, such that high and low mass stars are equally distributed throughout the cluster (i.e., the cluster is not mass segregated), r_{50} is equivalent to a common theoretical definition – the half-mass radius r_{hm}, a radius commonly measured in theoretical works due to its use in various dynamical equations (**portegies_zwart_young_2010**).

However, simple measures of cluster radius are not informative about the shape of a cluster, as clusters have long been known to have different shapes, with some

clusters being more centrally concentrated in their 'core' and others being sparser. It is helpful to apply models to OC radial profiles, allowing for the shape of clusters to be compared given models of a small number of parameters.

king_structure_star_1962 models are the most common models applied to star clusters. While originally derived for GCs, these models have also been shown to be a good fit to many OCs (**piskunov_towards_2007**), with a radial distribution function f given by:

$$f = k\left\{\frac{1}{\sqrt{1+(r/r_c)^2}} - \frac{1}{\sqrt{1+(r_t/r_c)^2}}\right\}^2 \tag{1.1}$$

where r is the distance from the cluster centre, r_c is the radius of the core of the cluster (the radius at which the surface density drops to half that of the centre), and r_t is the tidal radius of the cluster beyond which the Milky Way's potential is dominant. This can also be convenient to express in terms of the total number of stars within a distance r from the center of a cluster $n(x)$, which is given by:

$$n(x) = \pi r_c^2 k\left[\ln(1+x) - 4\frac{\sqrt{1+x}-1}{\sqrt{1+x_t}} + \frac{x}{1+x_t}\right] \tag{1.2}$$

where $x = (r/r_c)^2$ and $x_t = (r_t/r_c)^2$. Within the tidal field of a galaxy (such as the Milky Way), it is also helpful to compare the King tidal radius with the theoretically predicted Jacobi radius of a spherically symmetric cluster r_J:

$$r_J = \left(\frac{GM}{4\Omega^2 - k^2}\right) \tag{1.3}$$

which relates the limiting radius of a cluster in a galactic potential field r_J to the cluster's mass M, given the circular frequency Ω and the epicyclic frequency k of the cluster's orbit. For a spherically symmetric cluster on a circular orbit $r_J \approx r_t$, relating a cluster's mass to the product of a King model fit (or vice versa).

As an empirical model, the **king_structure_star_1962** model is mostly useful for simple observational comparisons between clusters, such as comparisons between the core and tidal radii between clusters of different ages (**kharchenko_global_2013**; **tarricq_structural_2022**). **king_structure_1966** re-derives a similar model from theoretical principles, including by assuming that the velocity distribution of stars in the centre of a star cluster is isothermal and that the cluster is in virial equilibrium. The shape of these models is parameterised by a dimensionless variable W_0 which

parameterises the concentration of a cluster, with higher values corresponding to a more centrally concentrated cluster. For $W_0 \lesssim 7$, **king_structure_star_1962** and **king_structure_1966** models are very similar. In practice, almost all OCs have $W_0 < 7$, and so these models can be used somewhat interchangeably (**portegies_zwart_young_2010**). Due to the significantly simpler functional form of **king_structure_star_1962** models, they are used almost exclusively in the OC literature relative to **king_structure_1966** models (**portegies_zwart_young_2010**; **cantat-gaudin_milky_2022**).

Finally, it is worth mentioning the model of **plummer_problem_1911**. Once again originally designed for GCs, this model parameterises how centrally concentrated a star cluster is based on a single scale factor a. Unlike **king_structure_star_1962** and **king_structure_1966** models, the **plummer_problem_1911** model assumes star clusters do not have a physical limiting radius and extend to infinity, which is of course unrealistic. Nevertheless, the **plummer_problem_1911** model is still a satisfactory approximation of star cluster distribution functions, and it is still used in the literature due to its simple functional form for which many parameters can be solved analytically (**dejonghe_completely_analytical_1987**). **plummer_problem_1911** models are particularly popular in theoretical studies of star clusters due to this reason (**portegies_zwart_young_2010**).

1.6.3 Dynamics

Later in this thesis, I use measures of OC dynamics to test if OCs are bound. Some works such as **bravi_gaia-eso_2018** and **pang_3d_2021** have used similar methods on small scales to test if OCs are bound. The following useful definitions are all from **portegies_zwart_young_2010**.

Firstly, for a cluster with a one-dimensional velocity dispersion $\sigma_{\rm 1D}$, its total kinetic energy T is approximately

$$T = \frac{3}{2} M \sigma_{\rm 1D}^2 \tag{1.4}$$

where M is the total cluster mass. In addition, one can define the total potential energy of a cluster U as

$$U = -\frac{GM^2}{2r_{\rm vir}} \tag{1.5}$$

where G is the gravitational constant and $r_{\rm vir}$ is the theoretically definied virial radius of the cluster, a parameter that is difficult to calculate observationally as it requires three-dimensional positions. The three-dimensional virial radius can be converted to the two-dimensional deprojected median radius r_{50} with

$$r_{\rm vir} = \frac{\eta}{6} r_{50} \tag{1.6}$$

where η is a constant that is model-dependent. For an ideal **plummer_problem_1911** model, η is equal to 9.75, although in practice, this value can be out by a factor of two to four in extreme cases of star clusters with distributions that are poorly described by a **plummer_problem_1911** model (**portegies_zwart_young_2010**).

Finally, putting these together, one can define the virial ratio Q of a cluster, which is the ratio of kinetic to potential energy for a given bound system. Since the virial theorem predicts that $2T + U = 0$, Q is hence given by

$$Q = \frac{T}{|U|} = \frac{\eta r_{50} \sigma_{\rm 1D}^2}{2GM} \approx \frac{1}{2} \quad \text{for a bound cluster.} \tag{1.7}$$

Equation **??** is also commonly expressed in terms of the predicted one-dimensional velocity dispersion of a virialised cluster $\sigma_{\rm vir}$ for a cluster of a given mass and radius (**bravi_gaia-eso_2018**), as

$$\sigma_{\rm vir} = \sqrt{\frac{GM}{\eta r_{50}}}. \tag{1.8}$$

With these equations and measures of cluster mass, radius, and velocity dispersion, it is possible to probe the overall dynamical state of a cluster observationally or in the result of simulations (**banerjee_how_2017**; **bravi_gaia-eso_2018**; **pang_3d_2021**).

1.6.4 Formation, evolution and destruction

Finally, having considered various individual pieces of important theory, it is also worth discussing the current overall theory of star cluster formation and evolution, which should give some theoretical context to the clusters observed at different ages in this work.

Formation (up to ~1 Myr)

As discussed previously, stars form when clouds of cold molecular gas (giant molecular clouds, GMCs) within galaxies collapse due to gravity. This process is not believed or observed to be continuous: stellar winds from young stars rapidly heat and blow away any remaining gas within the cluster, preventing further star formation from occuring. Effectively, this process 'freezes' star formation, ensuring that the resulting group of stars is roughly homogeneous in age and chemical composition (**lada_embedded_2003**; **krumholz_how_2020**). If the parent GMC is dense enough, then the stars will eventually collapse into a bound star cluster (**portegies_zwart_young_2010**; **krumholz_star_2019**; **krumholz_how_2020**). GMCs in the local universe generally have masses in the range $\sim 10^3$ to $\sim 10^7$ $M_\odot$ (**krause_physics_2020**). For exceptionally large GMCs with masses in excess of $\sim 10^8$ $M_\odot$, it is believed that they are large enough to form clusters as massive as GCs. Such conditions are rare in the current universe, being much more common at redshifts $z \gtrsim 2$ (**krumholz_star_2019**); such high-mass GMCs and the clusters they form are hence outside of the range of open clusters in this study, which generally have ages of no more than $\sim$1 Gyr.

Historically, it was believed that all stars formed in bound star clusters. However, recent observations with *Gaia* have suggested that it may only be a minority of stars that form in bound clusters (**ward_not_2019**; **wright_ob_associations_2020**). Nevertheless, for GMCs that are dense enough to at least form an OC, a bound, virialised young cluster will emerge.

Supernova feedback and expansion (~1 to ~30 Myr)

Once stellar winds expel the initial gas a bound cluster formed from, star clusters in the disk of the Milky Way continue to have a somewhat tumultuous life. Young clusters were typically observed to have sizes larger than those predicted by N-body simulations; it is now believed that other processes must inject energy into early young star clusters in order for them to reach their present-day larger sizes (**banerjee_how_2017**).

After initial gas expulsion within the cluster stops additional star formation, star clusters are believed to continue interacting with surrounding gas in their parent GMC for multiple Myr. Stellar winds and feedback from supernovae will continue to disperse the GMC well beyond the radius of the cluster, forming a HII region. This causes a 'gravitational feedback' effect, where the massive GMC is blown away and

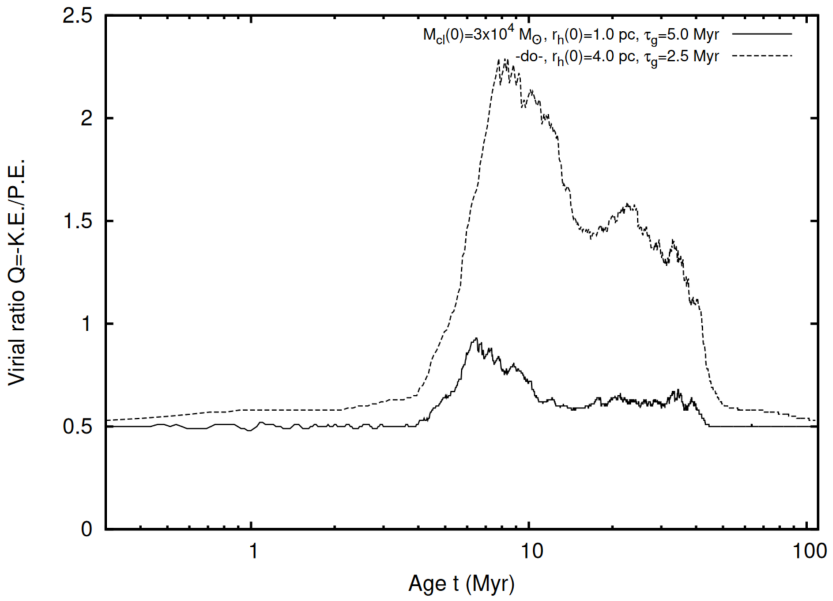

Fig. 1.12: The evolution of the virial ratio of simulated star clusters with a 'placid' model of explosive feedback. Simulations were conducted given a cluster of mass 3×10^4 $M_\odot$, with the dashed line showing a cluster with an initial half-mass radius of 1 pc and the solid line showing a cluster with an initial half-mass radius of 4 pc. *Credit*: **banerjee_how_2017**.

net forces on stars in the cluster also add energy to the system, causing the cluster to be supervirial and undergo a phase of expansion (**krause_physics_2020**).

The exact physics of this feedback are still under study, although Fig. **??** taken from **banerjee_how_2017** shows the virial ratio of simulated star clusters with respect to time, for an intermediate or 'placid' model of explosive stellar feedback. Within the first 10 Myr of the cluster's lives, feedback causes them to become supervirial. Once the parent GMC has been dispersed, the clusters continue to expand, before eventually reaching dynamical equilibrium and returning to a state with $Q \approx 0.5$ after a few tens of Myr. These simulations echo the results of **kuhn_kinematics_2019**, who studied 28 young stellar groups with *Gaia* DR2 and found that 75% were undergoing expansion (i.e. are supervirial).

All star clusters are expected to lose a significant portion of their initial mass during this phase of expansion. It has been theorised that bound clusters that form with low initial masses or high initial radii may even be completely destroyed in the initial phase of feedback and expansion, which ought to be visible as unbound cluster remnants (**krause_physics_2020**).

Evaporation and destruction (upto ~1 Gyr)

Since OCs are rarely observed at ages greater than ~ 1 Gyr, it is clear that some processes eventually destroy them over time. This is believed to happen in three ways.

Firstly, stellar evolution will gradually cause mass loss in a cluster, with more massive stars undergoing supernovae and evolving into compact remnants, losing a significant proportion of their mass in the process. However, since most stars are more compact M, K, or G stars with lifetimes significantly longer than the typical maximum OC lifetime of 1 Gyr, this effect is relatively insignificant during the lifespan of an OC (**krause_physics_2020**).

Secondly, clusters gradually lose stars over time in a process known as evaporation. The stars in a cluster will not all have the same velocity; a cluster in dynamical equilibrium will have an isothermal velocity dispersion that is approximately Maxwellian. Some stars with velocities at the tail end of this distribution will have velocities higher than the escape velocity of the cluster $v_{\rm esc}$, and will be ejected from the cluster. Over time, two and three-body interactions will accelerate some stars preferentially and ensure that a small proportion of cluster members always have velocities greater than the cluster's $v_{\rm esc}$ (**portegies_zwart_young_2010**; **krause_physics_2020**). Stars are preferentially ejected via the cluster's L_1 and L_2 Lagrange points relative to the tidal field of the Milky Way, which produces the observed tidal tails of many clusters in *Gaia* (**portegies_zwart_young_2010**; **tarricq_structural_2022**). Less commonly, stars will still be ejected in a random direction (opposed to via a Lagrange point), producing an additional spherical 'coma' or 'corona' of recently ejected stars around a cluster, an effect that has also been observed in *Gaia* data for a number of nearby OCs (**meingast_extended_2021**; **tarricq_structural_2022**).

Finally, star clusters are theorised to be heavily disrupted by tidal 'shocks' (perturbations). Reasonably often within every 1 Gyr, star clusters in the Milky Way's disk are expected to come close to or even collide with various pieces of massive galactic structure, such as GMCs or transient spiral arms. The tidal perturbations from these interactions increase the energy of stars in a cluster and should cause considerable mass loss. Due to the rarity of these events, they are yet to be directly observed for OCs in the Milky Way; nevertheless, simple theoretical arguments can show that these events will occur reasonably often for any star cluster in the disk. Especially within the first 1 Gyr of an OC's life, tidal shocks are expected to be a major (and potentially even dominant) method of star cluster destruction (**krause_physics_2020**).

1.7 The structure of this thesis

The remaining structure of this thesis will be as follows:

Chapter ??: ??

In the first study of this thesis, I compare clustering algorithms for OC retrieval in *Gaia* DR2 data. I review a number of different algorithms, before settling on three for use in the study. I develop methods to use these algorithms in a blind search, including a statistical density test to remove false positive clusters. I then use them to look for 100 OCs in *Gaia* data from the catalogue of **kharchenko_global_2013**, as well as a wider sample of 1385 OCs that were less well studied. A further development of the DBSCAN algorithm, HDBSCAN, was found to be the most effective algorithm for OC retrieval. Additionally, 41 new OCs were detected in the study.

Chapter ??: ??

Having refined an initial clustering methodology in Chapter **??**, I conduct the largest ever blind search for OCs in *Gaia* data, using HDBSCAN and data from *Gaia* DR3 down to magnitude $G \sim 20$. To further refine the results, I use a Bayesian neural network to derive the probablility of clusters being a single population of stars. In addition, I use a similar Bayesian neural network to derive ages, extinctions, and distances to the clusters in the work. A catalogue of 7167 clusters is produced, 2387 of which are candidate new objects and 4782 of which crossmatch to objects in the literature. 4105 clusters are in a high-quality sample of objects, including 739 of which are new. OC membership lists resulting from this method generally contain more member stars than in previous literature works, and often have tidal tails. It is possible to rule out over 1000 clusters from the pre-*Gaia* catalogue of **kharchenko_global_2013** that I am unable to detect. However, some of the clusters recovered in this chapter appear more compatible with unbound moving groups, and will require further classification with a dynamical methodology.

Chapter ??: ??

In the third scientific chapter of this thesis, I develop methodologies to accurate determine whether or not OC candidates in *Gaia* data are gravitationally bound, aiming to determine accurate masses, velocity dispersions, and radii for the clusters in the catalogue from Chapter **??**. **TODO: add more here once dynamics section is finalised.**

Chapter ??: ??

Finally, I present an overview of all the work contained within this thesis. I discuss the contribution of this thesis to the scientific literature. I also discuss the future of OC science in the *Gaia* era and beyond, and suggest a number of future avenues for research leading on from this thesis. **TODO: make sure this makes sense once conclusion is written**

2 Comparison of clustering algorithms applied to *Gaia* DR2 data

> *The reward of the young scientist is the emotional thrill of being the first person in the history of the world to see something or understand something. Nothing can compare with that experience.*
>
> — **Cecilia Payne-Gaposchkin**
> (1977)

Details of authorship. *The content of this chapter is almost entirely based on work published in* **hunt_improving_2021**. *I conducted all scientific work and wrote all of the text. Suggestions and corrections from my supervisor and the reviewer of the paper are included in the text. The formatting of figures and tables has been adjusted to better fit the formatting of this thesis.*

2.1 Introduction

Open clusters (OCs) are commonly known as the laboratories of stellar evolution, which form when large gas clouds collapse into dense, gravitationally bound regions of stars. The stars in OCs have roughly the same age and chemical composition, meaning that every OC is a unique 'experiment' showing the results of stellar evolution with stars across a range of masses given a certain set of initial conditions. In particular, OCs in our own galaxy are the most enlightening to study, since their proximity means that individual stars can be resolved and parameters can be determined to higher levels of precision.

The number of known open clusters has not changed significantly until recently. The New General Catalogue (NGC) listed ≈700 objects that we now know to be OCs (**dreyer_new_general_1888a**), the most comprehensive catalogue of its time

– yet over a century later, the catalogue of **mermilliod_database_1995** had only increased to a size of 1200 OCs, not even doubling the OC census despite the large strides in astronomical instrumentation and data analysis taken in the 20th century. In part, this is because numerous clusters in the literature were ruled out as associations by modern data, reducing the size of the census – yet it still persists that a century of work did not significantly increase the size of the OC census.

The largest increases to the size of the census came with the advent of new techniques. The space-based astrometric survey of the *Hipparcos* satellite (**hog_tycho-2_2000**) revealed a number of new, often relatively sparse OCs in studies such as **platais_search_1998** and **chereul_distribution_1999**, while wide-field infrared surveys looked through interstellar extinction to find new OCs in studies such as **dutra_new_2001** and **froebrich_systematic_2007**. The catalogue of **kharchenko_global_2013** (hereafter MWSC) lists 2267 probable OCs and a further 132 that showed nebulosity, a major increase from the figure of **mermilliod_database_1995** just two decades prior.

The next major increase to the size of the OC catalogue is currently in progress thanks to the *Gaia* satellite (**brown_gaia_2018**). *Gaia* maps the stars of the Milky Way in five dimensions (positions, proper motions, and parallax), while also providing visual photometry and colours in its own G, G_{BP}, and G_{RP} photometric bands, and spectroscopic radial velocities for a small sample of bright stars. Compared with *Hipparcos*, *Gaia* has roughly an order of magnitude more precision in astrometric parameters for 10^4 times as many stars, resulting in a groundbreaking dataset that has full astrometric solutions and photometry for 1.3 billion stars as of *Gaia* DR2, 7 million of which also have radial velocities.

It is perhaps unsurprising that such a large improvement in our ability to map the galaxy is also greatly improving the OC census. In terms of quantity, works such as **castro-ginard_new_2018**; **castro-ginard_hunting_2019**; **castro-ginard_hunting_2020**, **liu_catalog_2019**, **sim_207_2019**, and **cantat-gaudin_gaia_2019** have recently reported hundreds of candidate OCs using *Gaia* data. Typically, this is done using automated blind searches of the *Gaia* dataset with clustering algorithms – a type of unsupervised machine learning that can find the most natural groupings or clusters within a dataset, requiring only basic parameters and minimal prior knowledge about the structure of the data.

The precision of *Gaia* is also improving the quality of the OC census. While traditionally, distances to OCs would be derived using photometry alone and fitting a model-dependent stellar isochrone to the OC's colour-magnitude diagram, *Gaia* parallaxes provide an unbiased and model-independent distance es-

timator – allowing parameters for OCs to be derived to greater levels of precision (**cantat-gaudin_painting_2020**) . **cantat-gaudin_gaia_2018** derive membership lists and parameters for 1229 OCs using *Gaia* DR2 data alone, which has been expanded with some re-analysis and by including recently detected clusters in **cantat-gaudin_clusters_2020**. It is expected that some OCs listed in MWSC will not be detectable in *Gaia* data. *Gaia*'s visual band observations are unable to see into areas of high dust extinction unlike the infrared photometry used by MWSC – obscuring small, distant OCs from view in regions with high extinction, such as towards the galactic centre at distances greater than ~ 3 kpc. In addition, parallaxes and proper motions have fractional uncertainties that increase with distance, which has a significant negative effect on the signal to noise ratio of OCs in *Gaia* data at distances larger than $\sim 1 - 3$ kpc.

However, despite astrometric uncertainties or dust obscuring *Gaia*'s view of some clusters, it is also possible to rule out a number of OCs that should still be detectable in *Gaia* data based on their existing parameters. **cantat-gaudin_clusters_2020** have ruled out 38 OCs in the literature as asterisms, all of which should be bright enough to detect in the *Gaia* dataset based on their reported parameters but do not appear to exist. Future studies will be able to rule out yet more putative OCs based on *Gaia* data alone, particularly as *Gaia* data improves in the coming years with future releases.

In its current state, the OC census is difficult for astronomers to use. Despite MWSC deriving that the OC census was complete to within 1.8 kpc, the recent myriad of studies using *Gaia* data have shown that many more OCs are yet to be discovered within the immediate solar neighbourhood. Until the OC census is shown to be complete to within a certain radius, it is impossible to calculate accurate population statistics about OCs in the Milky Way. In addition, the many asterisms that are not yet concretely ruled out in the literature make the OC census more difficult to use, as not all reported OCs in the literature are really there and many do not make good targets for precious telescope time.

Within the next decade, future data releases of the *Gaia* satellite and large-scale spectroscopic surveys such as 4MOST (**de_jong_4most_2012**) will provide astronomers with a wealth of data on on our galaxy. OCs are an important piece of the jigsaw puzzle of the Milky Way's current and past star formation. An OC census with greatly improved quality and quantity will allow astronomers to use the census reliably for a range of scientific purposes, including mapping the age distribution of OCs across the galaxy (**cantat-gaudin_painting_2020**; **yen_reanalysis_2018**), studies of the chemical composition of OCs (**baratella_gaia-eso_2020**; **donor_open_2020**), to

even studying the conditions of planet formation in OCs and the implications that may have for the distribution of the wider exoplanet census (**fujii_survival_2019**).

To date, a number of different methods have been used to search for new or existing OCs in *Gaia* data. While UPMASK (**krone-martins_upmask:_2014**) as used by **cantat-gaudin_gaia_2018** and **cantat-gaudin_clusters_2020** is a highly successful tool for producing membership lists of existing OCs, it is too slow to conduct a large-scale blind search across the billion star dataset of *Gaia*. In turn, while approaches such as the one applied in **castro-ginard_hunting_2020** has detected hundreds of new OCs in the *Gaia* dataset, their method is unable to detect a large fraction of literature OCs, suggesting that their approach may also be unable to detect a large fraction of as yet undiscovered OCs with similar properties. Different approaches have advantages and disadvantages that have never before been compared side-by-side on *Gaia* data, and no single approach has yet been developed that can simultaneously detect new OCs in a large-scale blind search while also detecting a majority of already-reported objects.

In this series of papers, we will work to improve the OC census: primarily by attempting to detect new OCs, but also by re-detecting a large fraction of literature OCs with a different methodology and complementing cataloguing efforts such as **cantat-gaudin_clusters_2020**. In this study, we create an unbiased preprocessing pipeline to prepare *Gaia* data for analysis by clustering algorithms, and test the ability of three clustering algorithms to detect OCs in *Gaia* data. In Sect. **??**, we describe the *Gaia* data used and the applied pre-processing steps. Section **??** outlines the requirements for any clustering algorithm to be applied to *Gaia* data and describes three chosen algorithms that meet these criteria. Our analysis process for the algorithms applied to our data and the results of this are presented in Sect. **??**. In Sect. **??**, we discuss the strengths and weaknesses of each approach and the implications for future studies. We report on 41 new OC candidates discovered during the preparation of this paper in Sect. **??**. Finally, Sect. **??** summarises our results.

2.2 Data

2.2.1 The *Gaia* DR2 dataset and the HEALPix system

The results of any unsupervised search for OCs are always highly dependent on the input data and how it is preprocessed: assumptions must be made for reasons of

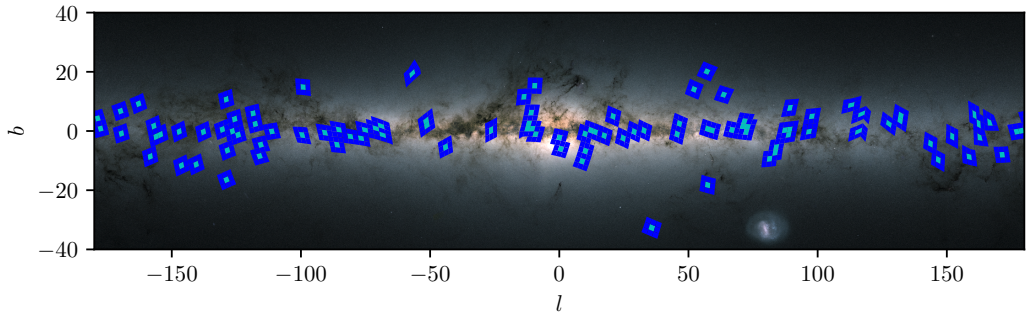

Fig. 2.1: Target fields for this study plotted above a *Gaia* map of stellar density in an equirectangular projection. Cyan regions show the 100 main HEALPix level five pixels. Each main pixel was merged with its eight nearest neighbours, which are shown in blue. Some nearest neighbour pixels overlap between different fields.

computational efficiency (for instance, splitting the dataset into separate chunks to improve runtime), and dimensions of the data with different units and coordinate systems must be intelligently preprocessed to allow an unsupervised algorithm to take full advantage of *Gaia* data. We briefly introduce the *Gaia* satellite and explain the preprocessing pipeline we developed to prepare its data for use with unsupervised clustering algorithms.

The *Gaia* satellite is producing a previously unprecedented quantity and quality of astrometric and photometric data for stars in the Milky Way. 1.7 billion sources brighter than $G = 21$ are included in *Gaia* DR2, where 1.3 billion have full five-parameter astrometric solutions. Uncertainties on derived parameters for each source depend strongly on the brightness of the source. While as many detected sources as possible are included in *Gaia* DR2 for completeness, the majority of faint sources are not useful for studies of galactic structure as the uncertainties on their parameters are too large.

For example, a star with brightness $G = 17$ would have corresponding uncertainties of 0.1 mas in parallax and 0.2 mas yr^{-1} in proper motion (**brown_gaia_2018**). If this star is 1 kpc away and hence has a true parallax of 1 mas, a measured parallax for this star would be informative to within roughly 10% of the true distance. The uncertainty on proper motion for this star would easily allow it to be distinguished as a member of an open cluster, as open clusters at this distance typically have an inherent proper motion dispersion of $\sim$1 mas yr^{-1} which is larger than the star's proper motion uncertainty. However, a faint star with $G = 20$ at a true distance of 1 kpc will have corresponding uncertainties of 0.7 mas in parallax and 1.2 mas yr^{-1} in proper motion. Any parallax measurement for this star will be much less informative about the star's true distance, with a near 100% fractional uncertainty. Its proper

motion uncertainty is larger than the typical dispersion of proper motion in OCs at 1 kpc, meaning that this faint star could never be reliably assigned as a member of an OC with *Gaia* DR2.

As such, most studies adopt a cut on the dataset to ignore uninformative stars and improve the signal to noise ratio of open clusters in the *Gaia* data. For the purposes of this study, we cut all stars fainter than $G = 18$, corresponding to typical maximum uncertainties of 0.15 mas in parallax and 0.3 mas yr^{-1} in proper motion, which is the same magnitude cut as used by **cantat-gaudin_gaia_2018** and **liu_catalog_2019**, although **castro-ginard_hunting_2020** adopt a stronger cut at $G = 17$.

Some studies (**castro-ginard_hunting_2020**; **liu_catalog_2019**) also remove outlier stars based on the magnitude of their proper motions or parallaxes. We choose not to remove stars with negative parallaxes, as this would make our study less sensitive to the most distant clusters for which member stars may have zero or negative parallax values. We also do not remove stars with high proper motions – while very few open clusters have proper motions $\mu_{\alpha*}$ or μ_δ of greater than 30 mas yr^{-1}, we still wish to have as few biases as possible in this study.

To select regions of the sky for study, we use the HEALPix[1] (Hierarchical Equal Area isoLatitude Pixelization) scheme (**gorski_healpix:_2005**) to select equal-area approximately quadrilateral regions of the *Gaia* dataset. HEALPix has advantages over rectangular tesselation schemes (**castro-ginard_hunting_2020**) since spherical distortions from projecting quadrilaterals onto the sky are spread out, allowing algorithms to be ran on equal-area pixels. In addition, *Gaia* sources are numbered based on a HEALPix system, making the HEALPix system convenient for a study of *Gaia* data to implement.

We aim to tile the sky into manageable chunks: large enough to contain OCs beyond a certain distance, but not so large that the amount of data in each chunk becomes prohibitively computationally expensive to run on. These chunks should also be easy to overlap, so that our future blind searches would not have edge effects. To do this, we select a region of study and its HEALPix level five pixel. The eight nearest pixels to each central pixel are added to each region of data to analyse, creating chunks each of area ~ 31 deg^2 and side length approximately 5°. 0.2% of HEALPix pixels only have seven neighbours and will hence have slightly smaller areas. It may be difficult to detect OCs closer than ∼350 pc with this tiling scheme since a typical OC at this distance may have a larger tidal radius that the data field. Any future blind search would be supplemented with a clustering analysis in Cartesian co-ordinates of all stars within 500 pc, solving this issue and also allowing nearby

[1] http://healpix.sourceforge.net

OCs to be properly detected without issues stemming from spherical distortions at angular separations greater than roughly $\sim 10°$.

HEALPix is straightforward to use with *Gaia* data, since all stars are numbered based on the HEALPix pixels they are present in. For a given `source_id`, its HEALPix pixel at level n is given by:

$$\text{HEALPix pixel} = \text{FLOOR}\left(\frac{\texttt{source\_id}}{2^{35} \cdot 4^{12-n}}\right). \tag{2.1}$$

For efficiency when querying the *Gaia* database with ADQL, this formula is inverted and used to select all stars with a `source_id` in the correct range of values. When downloading the stars in a given pixel, we require that all sources have a full five-parameter astrometric solution, valid G_{BP} and G_{RP} photometry, and a G-band magnitude less than 18. An exact copy of the ADQL query used for this study is included in Appendix **??**.

2.2.2 Selection of target fields

To study the effectiveness of clustering algorithms across a representative sample of stellar densities, we randomly selected 100 objects from MWSC that were each in unique HEALPix level five pixels. This list of 100 objects formed a list of 100 'main objects' to study. The eight nearest pixels to each central pixel were added to each of the 100 selected pixels to analyse. This resulted in 100 separate chunks, with 733 unique HEALPix level five pixels out of 900 in total since the neighbour pixels of different chunks were allowed to overlap. The fields are shown in Fig. **??** and listed in Appendix **??**. The fields contained between 100 000 to 4.2 million stars, with a mean of 734 000 stars. In total, all fields contain 56.8 million unique stars, representing ≈20% of the 260 million stars in *Gaia* DR2 brighter than $G = 18$.

All but one of these fields are in the galactic disk with $|b| < 25°$, and many of them are situated in areas of dense star formation where many OCs are present. We accidentally selected two globular clusters in our main list that did not contain any OCs centrally located in their field, which we replaced in our main list of 100 OCs with OCs from fields 14 and 57. To expand the target list from the initial 100 OCs to include other OCs contained within these fields, we searched the catalogues of MWSC, **cantat-gaudin_clusters_2020**, **castro-ginard_hunting_2020** and **liu_catalog_2019**, which contain a total sum of 4002 reported OCs. We required that reported OCs in the literature be entirely contained by a field given

their reported radius and distance to mitigate edge effects which could cause non-detections. For **cantat-gaudin_clusters_2020** OCs, $2 \cdot r_{50}$ (the radius containing half of the members of the OC) was used as a proxy for tidal radius. For the catalogue of **castro-ginard_hunting_2020**, which lists Gaussian angular dispersions θ containing $\sim 68\%$ of members, $2 \cdot \theta$ was used as a proxy for tidal radius. In total, the literature reports 1385 unique OCs contained within the 100 fields, all of which should be entirely visible and not partially clipped by the fields' edges. This represents roughly a third of the total number of clusters that the above four works report.

The objects in MWSC that remain undetected in *Gaia* data present a particular challenge for the algorithms in this study. Since **cantat-gaudin_clusters_2020** have found that a number of clusters listed in MWSC are not real, we do not expect any algorithm to detect OC candidates corresponding to all listed targets, meaning that most MWSC targets are false positives that should be discarded by the algorithms. However, a small number of MWSC objects may be real but are simply as yet undetected in *Gaia* data due to limitations of the methodologies used. Hence, the inclusion of MWSC objects allows us to test the ability of the algorithms to rule out putative objects, corresponding to their true and false negative rates, while also testing the algorithms against the sensitivity of existing approaches and seeing if any additional MWSC targets can be recovered in *Gaia* data with new methodologies. We chose to use real *Gaia* data for our study instead of simulated data so that we can develop our full pipeline from start to finish to work with real data, which includes a number of challenging aspects **lindegren_gaia_2018** that would not be adequately tested by using simulated data only.

2.2.3 Preprocessing steps

In the limit of small errors, clustering analysis could be performed with three-dimensional spatial data in a Cartesian frame. However, parallaxes are inherently difficult to measure and have large fractional uncertainties in *Gaia*, and transforming the spherical co-ordinate system of *Gaia* data to Cartesian co-ordinates is non-trivial and would introduce large errors to other axes of the data. As such, it is easier to remain in a spherical co-ordinate system to avoid contaminating positional data with the large errors of parallax measurements. Searches for OCs are helped immensely by proper motions, as OCs are gravitationally bound groups of stars that appear tightly clumped in proper motion space. These could be changed to Cartesian velocities with parallaxes, but this is avoided for the same reasons as with positions.

Attempts were made to use distances instead of parallaxes with the distance catalogue of **bailer-jones_estimating_2018**. However, while their method is appropriate for macroscopic studies of galactic structure, it places stars with uncertain parallaxes at a prior-defined distance, which moves low magnitude member stars further from their parent OCs in the data and was found to reduce the signal to noise ratio of OCs in the data. Alternative distance estimators that are better at preserving small scale galactic structure could be investigated in the future, such as StarHorse (**anders_photo-astrometric_2019**) which uses magnitude information and stellar models to increase the accuracy of *Gaia*-derived distances.

Some pre-processing can be done to reduce the effect of remaining in a spherical co-ordinate system and using parallaxes instead of distances, but without contaminating other dimensions of the data with the large uncertainties of parallax measurements. To remove spherical distortions that occur at high latitudes in position and proper motions, every field is rotated to an arbitrary co-ordinate frame $(\lambda,\, \phi)$ centred at (0,0) and rotated to have edges parallel with the co-ordinate axes for neater plotting of individual fields, with proper motions $\mu_{\alpha*}$, μ_δ also transformed to the new frame as $\mu_{\lambda*}$, μ_ϕ.

Machine learning algorithms benefit from having scaled inputs, so the five dimensions of data for each field $(\lambda,\, \phi,\, \mu_{\lambda*},\, \mu_\phi,\, \varpi)$ are re-scaled to have a median of zero and a unit inter-quartile range using a `RobustScaler` object from `scikit-learn` (**pedregosa_scikit-learn_2011**). This process is resilient to outliers, unlike scaling to have zero mean and unit variance as is sometimes used in the literature. This re-scaling process also ensures that each co-ordinate axis has an equal weight when passed to clustering algorithms. We choose not to experiment with re-weighting dimensions of the dataset as was performed tentatively by **liu_catalog_2019**, although this could be explored in future works.

2.3 Selection and implementation of clustering algorithms

2.3.1 Criteria

While many clustering algorithms exist in the literature, the complexities of *Gaia* data make only a few appropriate for a large-scale unsupervised OC search. In future works, we will run on the ≈ 200 million stars in *Gaia* data brighter than G $= 18$,

Tab. 2.1: Algorithms considered for inclusion by this study.

Algorithm	Runtime scaling[a]	Deals with noise	Open-source
KMeans	n	No	`sklearn`[b]
Affinity propagation	n^2	No	`sklearn`[b]
Mean-shift	n^2	No	`sklearn`[b]
Spectral	n^3	No	`sklearn`[b]
Ward	n^3	No	`sklearn`[b]
Agglomerative	n^3	No	`sklearn`[b]
DBSCAN	$n \log n$	Yes	`sklearn`[b]
OPTICS	n^2	Yes	`sklearn`[b]
Gaussian mixtures	n	No	`sklearn`[b]
Birch	n	No	`sklearn`[b]
Friend of Friends	$n \log n$	No	`pyfof`[c]
HDBSCAN	$n \log n$	Yes	`HDBSCAN`[d]

Notes. [(a)] Runtime scalings are best case estimates and are only given with respect to number of data points n. [(b)] https://scikit-learn.org/ [(c)] https://pypi.org/project/pyfof/ [(d)] https://pypi.org/project/hdbscan/

only a small fraction of which reside in OCs. Individual fields can contain up to approximately five million stars. Hence, any clustering algorithm would need to be extremely efficient at searching through a large quantity of data to find rare objects that require a high degree of sensitivity to detect.

In later parts of this work, we compare the performance of the clustering algorithms we selected. However, to be selected for further study, the algorithms must be even remotely practical for use with *Gaia* data. We set the following basic requirements on clustering algorithms for inclusion in this study. Firstly, it must be fast enough to run on the entire *Gaia* dataset with a few weeks of wall time on a relatively powerful computer. Secondly, it must be able to deal with unclustered field stars (noise), as only a small fraction of stars in the Milky Way reside in OCs and the rest must be discarded. Finally, an open-source implementation must be readily available in the literature for the algorithm.

The performance of all clustering algorithms against these criteria listed in the `scikit-learn` (**pedregosa_scikit-learn_2011**) Python library, in addition to two other common algorithms considered here, is listed in Table **??**. The galaxy cluster detection algorithm AMICO (**bellagamba_amico:_2018**) was also investigated for this work, but necessary modifications to the algorithm were not made in time to adapt it for use with *Gaia* data. AMICO was a top performing algorithm on

mock *Euclid* data (**euclid_collaboration_euclid_2019**), and so its application to OC detection would still be worth investigating in the future.

The first criterion disqualifies the vast majority of clustering algorithms in the literature. Practically, algorithms with runtime complexities of $\mathcal{O}(n^2)$ or worse (where n is the number of stars) are too slow to run on large segments of the *Gaia* dataset. For instance, while OPTICS (**ankerst_optics_1999**) has seen some use in astronomy analysing smaller portions of the *Gaia* dataset – such as by **ward_not_2019**, who used OPTICS to detect OB associations – its $\mathcal{O}(n^2)$ runtime complexity was found to be prohibitively slow for inclusion in this work.

The second criterion favours density-based clustering algorithms such as DBSCAN (**ester_density-based_1996**) and HDBSCAN (**hutchison_hdbscan_2013**), which are the only class of clustering algorithm that can discard points that are not in locally dense regions. These algorithms use nearest-neighbour distances to infer the local density around points, with points in low density regions discarded as field stars. However, some success has also been had in the literature with using fast algorithms to partition all data and only keep partitions that look like OCs, such as with Gaussian mixture models (**dempster_maximum_1977**) by **cantat-gaudin_gaia_2019**. A simple cut on proper motion dispersion can be enough to discard most non-OC partitions. The third criterion is unrestrictive, as open-source implementations exist for all algorithms that will be considered in this study.

Three algorithms were selected for further study. Firstly, DBSCAN, as mentioned previously, is a fast density-based clustering algorithm with excellent scalability, that has already proven itself in the literature in the blind searches of **castro-ginard_new_2018**; **castro-ginard_hunting_2019**; **castro-ginard_hunting_2020**, recently finding hundreds of new OCs in *Gaia* data.

The second algorithm, HDBSCAN, is also density-based but improves upon DBSCAN by clustering the data hierarchically, allowing it to deal with areas of different densities better and theoretically giving it greater sensitivity. Its parameters are different to DBSCAN, and may or may not be easier to tune. It has been used by **kounkel_untangling_2019** and **kounkel_untangling_2020** to probe the *Gaia* dataset for spatially correlated moving groups within 3 kpc, but has never been used purely to search for OCs and across all distance scales in the *Gaia* dataset.

Finally, GMMs were selected for trial, an algorithm unlike DBSCAN or HDBSCAN in that it must partition all data, and partitions not containing OCs must be discarded. In principle, this could be a fast method, as GMMs have $\mathcal{O}(n)$ runtime. A method similar to that of **cantat-gaudin_gaia_2019** should be used to discard unclustered

field stars with this algorithm. In addition, since it fits a model directly to the data instead of using nearest-neighbour distances, it should be less sensitive to the preprocessing or underlying shape of the data.

Some algorithms were not included in this study as their performance is clearly superseded by one of the three above. K-Means (**macqueen_methods_1967**) is a partitioning algorithm similar to GMMs that fits a user-specified number of centroids to a dataset. Points are assigned to their nearest centroid. However, this algorithm was found to perform poorly on the *Gaia* dataset, as the five dimensions of the data have intrinsically different scales. A distant cluster will have a near-negligible size in positional space, but will still form a Gaussian clump in proper motion and parallax spaces due to the dominance of *Gaia* errors at these distances. Alternatively, a nearby cluster will have large sizes in position and proper motion spaces, but still a relatively small parallax dispersion as all stars are at roughly the same distance. K-Means will routinely over or under-select cluster stars without extremely careful pre-processing, as it cannot re-scale its model independently for each axis of the data. However, GMMs can, since they fit a multivariate Gaussian. As such, including K-Means in this study was unnecessary, as GMMs are effectively a generalisation of the K-Means algorithm that allows each cluster to have a covariance matrix (i.e. a different scale for each axis.) In a similar vein, while Birch (**zhang_birch_1996**) is similar to K-Means clustering but makes a number of improvements, it also struggles to deal with clusters that have different scales in each dimension for the same reasons.

The Friend of Friends (FoF) algorithm has also been used in the literature (**liu_catalog_2019**) and has a history of use in astronomy, especially in searches for dwarf galaxies (**duarte_how_2014**) or dark matter haloes. However, it was not included in this study as it is the same as running DBSCAN with the minimum number of points (m_{Pts}) parameter set to 1, since both algorithms use a global density parameter and a notion of core or border points – or 'friends' and 'friends of friends' in the FoF algorithm. However, the addition of m_{Pts} to DBSCAN allows it to deal with unclustered points by discarding small clusters, which is a clear advantage for *Gaia* data and in line with our second criterion – whereas users of the FoF algorithm must manually discard small clusters.

In the following sub-sections, each selected algorithm will be explained in brief detail, along with any steps necessary to determine their parameters.

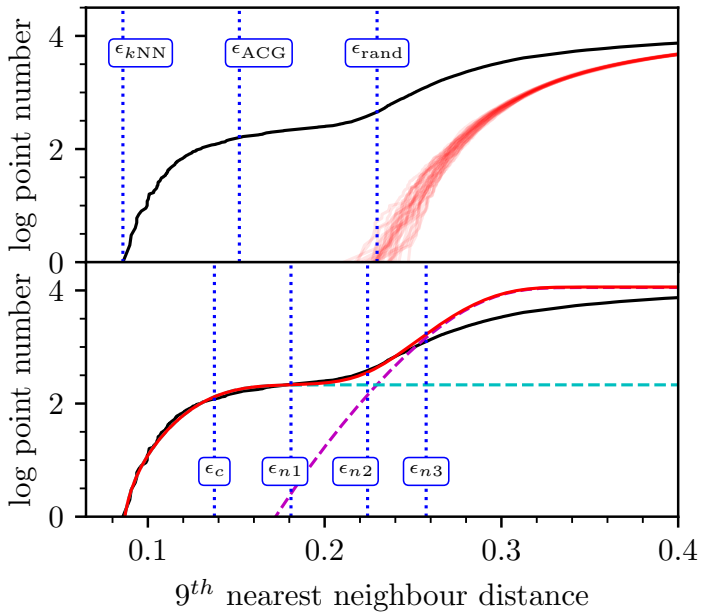

Fig. 2.2: Nearest neighbour graphs for both methods of determining the optimum ϵ for DBSCAN. To produce this plot, which is effectively an unnormalised log cumulative density function (CDF) of nearest neighbour distances, stars are sorted based on their k^{th}NN distances and numbered from one to n. These labels as a function of k^{th}NN distance are then plotted to form a continuous curve. The black line on both plots is the 9^{th}NN distances of a 2.5° field around the nearby OC Blanco 1. On the upper plot, ϵ estimates are determined by re-sampling the field 30 times (shown in red) to smooth out the signature of clustered stars. On the lower plot, a model of the signature of the cluster (cyan, dashed) and the field (magenta, dashed) is summed (red, solid) to approximate the curve and produce four ϵ estimates.

2.3.2 DBSCAN

Description of algorithm

DBSCAN (**ester_density-based_1996**) is one of the oldest and most widely used density-based clustering algorithms in the literature. It works by using the distances between points as a proxy for the local density of an area in a dataset, with the densest areas labelled as clusters and sparse regions labelled as unclustered background noise. Clusters are selected using two parameters. Firstly, points in a dataset are labelled as ϵ-reachable if the distance between them is lower than some threshold ϵ. Secondly, points are labelled as core points if they are ϵ-reachable to at least m_{Pts} other points, or border points if they are not core points but are ϵ-reachable to a core point, where m_{Pts} also includes the considered point itself. Finally, clusters are selected as density-connected groups of points that are ϵ-reachable via a core point, with all other points labelled as noise.

In this way, it follows that setting ϵ to a very large value would cause all points to be labelled as one cluster, and setting ϵ to a very small value would cause no points to be labelled as cluster members. The key is to set ϵ to an appropriate value, such that separate clusters are not accidentally merged by the algorithm, and such that the algorithm is still sensitive to sparse clusters that are only marginally denser than surrounding noise points. However, this is difficult for datasets of variable density, since ϵ is a global parameter. This is a particular issue for *Gaia* data, since the density of the dataset is highly variable: due to the spherical projection of *Gaia* data, the density of the dataset changes with distance, since higher distances sample a larger angular volume. In addition, fields that include opaque clouds have variable densities on scales of less than 1°, since the high levels of extinction in the clouds reduces the completeness of the *Gaia* instrument. Hence, a key challenge with using DBSCAN on *Gaia* data is choosing values of ϵ that are a good enough fit to the entirety of every field under study.

The m_{Pts} parameter must be set high enough to restrict the core point label to only the most densely connected points, but not so high that even real clusters do not contain enough points to generate core points. In practice, m_{Pts} and ϵ do not act independently, with a different choice of ϵ able to largely reproduce the same result for most values of m_{Pts}. As such, m_{Pts} can be set to the most efficient choice. **ester_density-based_1996** suggest setting m_{Pts} to twice the number of dimensions of the dataset, as higher values are more computationally intensive but do not appear to include more information. For the 5D *Gaia* dataset, this would imply setting $m_{Pts} = 10$, which also sets a threshold on the minimum size of an OC candidate at ten stars.

By far the most computationally expensive part of the algorithm is the computation of nearest neighbour distances. This is greatly sped up by using a k-d tree to calculate nearest neighbour distances efficiently, which is used by the `scikit-learn` (**pedregosa_scikit-learn_2011**) implementation of DBSCAN which is used in this work.

In the following subsections, two methods for determining ϵ for each field are presented, which are both be compared by this study.

Parameter determination with the Castro-Ginard et al. (ACG) method

castro-ginard_new_2018 have developed a method for determining ϵ for *Gaia* data that exploits the random, unclustered nature of field stars to produce consistent ϵ

estimates (hereafter abbreviated as the ACG method.) A brief description of how it works follows.

Firstly, a k^{th} nearest neighbour graph is computed for the dataset, where $k = m_{Pts} - 1$ (since m_{Pts} includes each point itself whereas k is the distance to the nearest neighbouring point.) The smallest k^{th} nearest neighbour distance $\epsilon_{k\text{NN}}$ is recorded.

Secondly, the data are randomly re-sampled according to the overall distribution of astrometric parameters in a given field. Assuming that the contribution of a cluster to this distribution is small as very few stars reside in OCs, the signature of the cluster is removed in the randomly redrawn nearest neighbour graph, allowing it to approximate the distribution of field stars in the dataset. Its minimum k^{th} nearest neighbour distance ϵ_{rand} is recorded. This step can be repeated multiple times to take a more accurate mean value of ϵ_{rand}. **castro-ginard_new_2018** repeat this step 30 times.

Finally, the average of these two values $\epsilon_{\text{ACG}} = (\epsilon_{k\text{NN}} + \epsilon_{\text{rand}}) / 2$ is used as ϵ by DBSCAN. When a cluster is present in a field, ϵ_{ACG} roughly approximates the modal k^{th} nearest neighbour value for the cluster. When no cluster is present in a field, $\epsilon_{\text{ACG}} \approx \epsilon_{k\text{NN}} \approx \epsilon_{\text{rand}}$, and no clusters will be erroneously detected by DBSCAN.

The ACG method is explained in more depth in **castro-ginard_new_2018**. The top panel of Fig. **??** shows how random re-sampling allows ϵ_{ACG} to be calculated. The implementation of this method differs slightly from the original used by **castro-ginard_new_2018**, as random re-draws in the second step are performed by randomly re-using existing parameter values for stars instead of first averaging them with kernel density estimation, as this was found to produce equivalent results while being somewhat faster.

In **castro-ginard_new_2018**; **castro-ginard_hunting_2020**, the size of the field under study and the parameter m_{Pts} are also varied across a number of different values, helping to reduce the effect of DBSCAN's global density parameter and detect OCs of different densities. Instead, we trialed varying only ϵ with the following method, as we expect this will produce similar results while being more computationally efficient: changing the size of the field requires re-calculating the array of nearest neighbour distances, whereas only varying ϵ means that the array can be cached and efficiently re-used for new parameter values. In practice, one could also vary ϵ with the ACG method by using different multiples of ϵ_{ACG} (e.g. $1.5 \cdot \epsilon_{\text{ACG}}$ or $2 \cdot \epsilon_{\text{ACG}}$).

Parameter determination with a model-fitting method

While consistent, the ACG method is slow. Since k^{th} nearest neighbour determination is the most computationally expensive part of DBSCAN, repeating it 30 times to randomly estimate ϵ_{rand} increases the runtime of DBSCAN by a factor of about 30. Instead, a model-fitting method was devised in this study to perform fast approximate analyses of the k^{th} nearest neighbour graph of a field.

Instead of numerically differentiating this graph to find turning points and hence an optimum value for ϵ, fitting a simple, approximate model is significantly more consistent and numerically stable. A cluster can be made up of just a few dozen stars projected against tens of thousands of background stars, complicating numerical differentiation since the signal of a cluster in such a graph is small and noisy.

chandrasekhar_stochastic_1943 derived a law for the nearest neighbour distribution of a uniformly distributed set of points in 3D, which can be converted to an arbitrary dimensionality d as:

$$P\left(x,\, A,\, a,\, d,\, k\right) = A\frac{x^{d+k-1}}{a^d}\exp\left[-\left(\frac{x}{a}\right)^d\right], \tag{2.2}$$

where x is the k^{th} nearest neighbour distance, A is a normalisation constant calculated numerically, and a is a constant that can be expressed in terms of the modal k^{th} nearest neighbour distance x_{max} as

$$a = x_{max}\left(\frac{k-1}{d}+1\right)^{\frac{-1}{d}}. \tag{2.3}$$

Many different density scales exist across the 5D *Gaia* dataset. OCs or globular clusters are the densest regions, with spatially correlated moving groups (**kounkel_untangling_2019**; **kounkel_untangling_2020**) also forming dense groups. Unclustered field stars exist across a range of different densities: fewer stars further from the *Gaia* instrument are bright enough to be detected, so distant or dust-obscured regions have lower densities; while regions in the galactic thin disk (especially towards the galactic centre) have high densities.

Ideally, this complicated structure would be captured by fitting many instances of Eqn. **??** simultaneously, effectively integrating across all density levels to perfectly fit a k^{th} nearest neighbour model to a field. However, this would be time intensive, and was found to be unnecessary, since a simple two-instance fit in log-log space could

achieve good results in less than a second of runtime. A single function P_c with parameters $\theta_c = \{a_c,\, d_c,\, k\}$ was combined with a single function P_f with parameters $\theta_f = \{a_f,\, d_f,\, k\}$, where the former and the latter represent the signal of the cluster and the field respectively:

$$P_{total}(x,\, A_t,\, C,\, \theta_c,\, \theta_f) = A_t\,[C \cdot P_c(x,\, \theta_c) + (1 - C) \cdot P_f(x,\, \theta_f)] \tag{2.4}$$

where a single normalisation constant A_t was used. C, a number between 0 and 1, represents the cluster fraction, corresponding to the strength of the signal of the cluster in the k^{th} nearest neighbour graph relative to unclustered field stars. k was set to 9, and the fit was constrained using Eqn. **??** such that $x_{max,c} < x_{max,f}$.

The fit was further stabilised by finding $x_{max,f}$ numerically in a histogram of k^{th} nearest neighbour distances, which was then used in conjunction with Eqn. **??** to fix a_f, leaving just four free parameters: a_c, d_c, d_f and C, with A_t determined numerically at every fitting iteration. The dimensionalities d_c and d_f were allowed to be non-integer to give the fit access to a greater range of shapes.

An example fit is shown in Fig. **??**. Once a field has been fit, points of interest in the curve can be used to estimate ϵ. The first, ϵ_c, is the modal k^{th}NN distance of the cluster component of the model, and physically corresponds to the most optimum ϵ value for the most prominent cluster in a given field. ϵ_c was often similar to ϵ_{ACG}.

When multiple OCs are in a single field, sparser objects with less contrast against field stars were not detected at the ϵ_c level, and so we also took three additional values from the curve: ϵ_{n1} is the first inflection point in the second derivative of the overall model, ϵ_{n2} is the point in the model with the highest second derivative (i.e. the highest rate of change of curvature) and ϵ_{n3} is the third inflection point in the second derivative of the overall model. These additional points correspond to where unclustered stars become increasingly dominant in the nearest neighbour distribution of the entire field, and allow low contrast objects in a field to be detected even if the shape of the fit and the value of ϵ_c has been primarily influenced by a denser object in the field. However, there is a trade-off: these higher values of ϵ are also likely to produce more false positives. Values higher than ϵ_{n3} were briefly investigated but were found to have false positive rates that were too high to be useful.

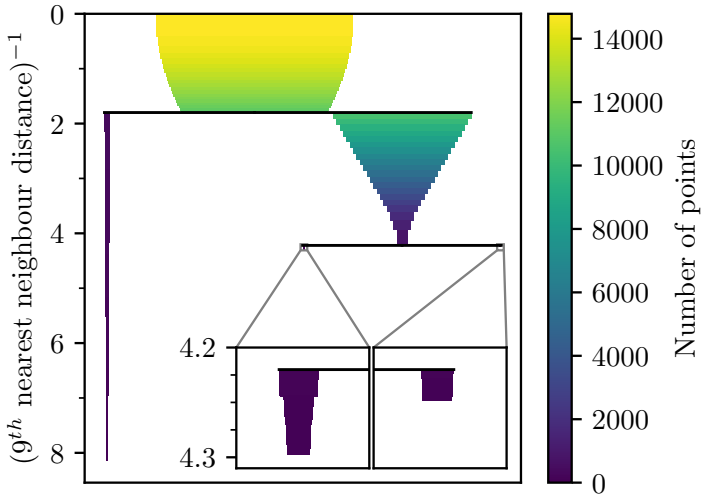

Fig. 2.3: Condensed tree graph for HDBSCAN with $m_{clSize} = 80$ applied to a 2.5° field around Blanco 1, a nearby cluster without any other known OCs in the field. The colour and width of each icicle denotes the number of stars remaining in the cluster. Horizontal splits occur when clusters are no longer connected. The long icicle on the left is Blanco 1, which is an extremely clear, nearby cluster and hence splits early from field stars. On the right, the algorithm continues discarding field stars, splitting into two very short icicles at the end which are false positive clusters. The two small sub-plots in the lower right are zoomed in on the two small icicles.

2.3.3 HDBSCAN

Description of algorithm

HDBSCAN (**hutchison_hdbscan_2013**) is a more recently developed clustering algorithm that attempts to improve the performance and usability of previous approaches. HDBSCAN combines the density-based approach of DBSCAN with hierarchical clustering, allowing it to deal with datasets of varying densities. Despite the extra computations, HDBSCAN does not have a significant increase in runtime compared to DBSCAN.

To evaluate possible clustering, nearest neighbour distances are calculated as with DBSCAN. However, HDBSCAN then effectively considers all possible DBSCAN solutions for all possible values of ϵ, constructing a hierarchical tree representation of the possible clusterings of the dataset. As with DBSCAN, clusters are defined using an m_{Pts} parameter to define core and border points. HDBSCAN 'replaces' the ϵ parameter of DBSCAN with a minimum cluster size m_{clSize}, which is used to define the minimum possible size of a cluster before all points within it are instead classified as noise. Smaller values of m_{clSize} cause the hierarchical graph to be split

more, as deeper, more nested solutions become valid. Larger m_{clSize} values will merge small groups, negating the algorithm's sensitivity to clusters smaller than m_{clSize} but while reducing the number of false positive associations of points in the dataset that are reported as clusters.

Figure **??** shows a representation of the HDBSCAN hierarchical graph for clustering analysis performed on a 2.5° field centred on Blanco 1, with parameters $m_{clSize} = 80$ and $m_{Pts} = 10$. Having produced a hierarchical graph representation of the dataset, clusters can be selected from it in one of two ways. In the Excess of Mass (EoM) method, the clusters with the largest area in this plot are selected. Alternatively, in the leaf method, more fine-grained structure is revealed, as clusters at the bottom of the tree are always selected.

HDBSCAN solves a number of issues encountered by previous approaches in the literature. For instance: whereas DBSCAN requires setting the ϵ parameter homogeneously across an entire dataset, giving it poor performance when detecting clusters of different densities, HDBSCAN's consideration of all DBSCAN solutions simultaneously gives it equal sensitivity across all density ranges of a dataset. In addition, m_{clSize} is a much more intuitive parameter to set than ϵ for detecting OCs, since the minimum allowable size of an OC can be decided beforehand and does not require an additional method to try and estimate it for a given dataset as with ϵ.

However, use of HDBSCAN comes with some challenges when running on largely unclustered data - such as the *Gaia* dataset, where very few stars reside in OCs. HDBSCAN is sensitive to all regions of a dataset where points appear statistically more clustered than the local background, particularly when setting the parameter m_{clSize} low to ensure that HDBSCAN is sensitive to the smallest galactic OCs. In the *Gaia* regime, it is unsurprising that a field of one million stars will contain many low signal to noise ratio false positive associations, where groups of $10-20$ stars will appear more clustered than the background by statistical chance. These false positives will be reported by HDBSCAN and must be later removed to use HDBSCAN successfully at high sensitivities.

The Python implementation of HDBSCAN by **mcinnes_hdbscan_2017** was used for this work, which differs from the original publication in a few small ways (such as using a k-d tree for nearest neighbour computation) that allow the algorithm to run faster.

Parameter tuning

HDBSCAN parameters were straightforward to set in a number of small experiments conducted on well-characterised OCs. While the original HDBSCAN paper recommends setting m_{Pts} and m_{clSize} to the same value, setting $m_{Pts} = 10$ was found to offer the best sensitivity and speed when running the algorithm, a decision also supported by the arguments for setting $m_{Pts} = 10$ for DBSCAN.

To select candidate clusters from the hierarchical tree, the leaf selection method was almost always superior to the EoM method. OCs contain a very small number of stars ($\sim$100) compared to the fields they occupy ($\sim$100 000+), and the leaf selection method was significantly better at recovering the smallest objects (OCs) in a given field.

However, there is no perfect setting for m_{clSize} having investigated the effect of the parameter in a number of small experiments, for which we tested different parameter values against a representative set of OCs. An exact m_{clSize} setting must be found empirically based on the properties of a dataset. While theory suggests that m_{clSize} should not be smaller than the smallest size of an OC (which we define as ten stars in this work), such low values were found to produce a large number of false positives, also sometimes erroneously splitting the largest OCs into two or more sub-clusters that miss many valid members of the cluster. Alternatively, setting it high (e.g. $m_{clSize} = 80$) makes the algorithm's output significantly less noisy at the cost of missing the smallest objects. Values larger than 80 had no added advantages despite further decreasing the algorithm's sensitivity to smaller OCs. High values also sometimes select dense regions of field stars that must be removed later as they are not OCs. They may correspond to moving groups such as those reported by **kounkel_untangling_2019** and **kounkel_untangling_2020**. A range of settings (10, 20, 40 and 80) will be compared in this paper.

2.3.4 Gaussian mixture models

Description of algorithm

GMMs differ from the other methods considered in this study in a number of ways, offering an interesting alternative viewpoint on how an entirely different and much older method performs when trying to detect OCs in a large, modern dataset. The data are assumed to be drawn from a number of Gaussian distributions, to which the algorithm fits a mixture of m Gaussian components across a series of iterations.

The likelihood of consecutive iterations is maximised until convergence is achieved. Covariances between dimensions allow the fitted Gaussians to have an elliptical or diagonal shape, which is important for OCs as many are elongated due to tidal effects.

Since all points must be assigned as a member of a Gaussian, GMMs do not have a natural way to deal with unclustered field stars. Instead, mixture components must be ruled out if their properties are incompatible with OCs. The means and standard deviations of mixture components in the different scaled dimensions (λ, ϕ, $\mu_{\lambda*}$, μ_ϕ, ϖ) can be quickly used as proxies for the properties of a candidate OC, with any targets wholly incompatible with an OC ruled out as groupings of field stars. We adopt a similar approach to **cantat-gaudin_gaia_2019** and require that the following constraints are met on the proper motion dispersion ($\sigma_{\mu_{\lambda*}}$, σ_{μ_ϕ}) and the dispersion in positional space (σ_λ, , σ_ϕ) respectively:

$$\sqrt{\sigma^2_{\mu_{\lambda*}} + \sigma^2_{\mu_\phi}} \leq \begin{cases} 1 \text{ mas yr}^{-1} & \varpi \leq 0.67 \text{ mas} \\ 1.49 \cdot \varpi \text{ mas yr}^{-1} & \varpi > 0.67 \text{ mas} \end{cases} \tag{2.5}$$

$$\sqrt{\sigma^2_{\lambda} + \sigma^2_{\phi}} \leq \begin{cases} 0.1^\circ & \varpi \leq 0.17 \text{ mas} \\ \arctan(\varpi/100)^\circ & \varpi > 0.17 \text{ mas} \end{cases} \tag{2.6}$$

which differs from the constraints of **cantat-gaudin_gaia_2019**, who only include a proper motion constraint. The addition of a latter radius constraint helps to remove clear false positives that are significantly larger than the typical size of OCs.

The implementation of GMMs freely available in `scikit-learn` (**pedregosa_scikit-learn_2011**) was used in this study. In addition, **cantat-gaudin_gaia_2019** have used UPMASK (**krone-martins_upmask:_2014**) to verify OC candidates. However, this was deemed unnecessary for this study, as the sample of OC candidates after the application of the constraints was already relatively clean with few false positives.

Parameter tuning & dataset sub-partitioning

Issues were encountered when attempting to tune the number of mixtures m. Firstly, larger fields required linearly more mixtures to ensure that enough were available for fitting to field stars, such that $m \propto n$. Instead, it is easier to set the parameter m_s, the number of stars per mixture – where $m = n/m_s$. This causes the method to be n times slower, since the GMM runtime complexity also scales linearly with the number

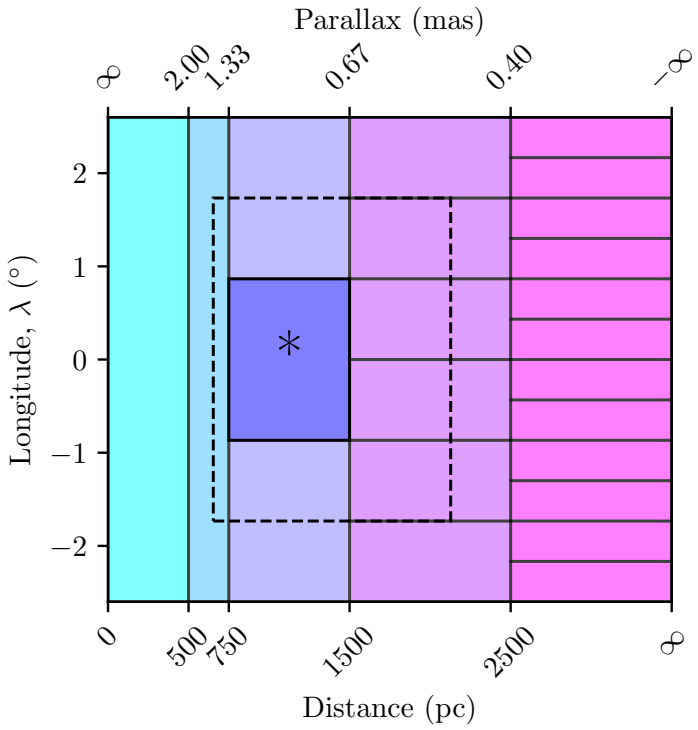

Fig. 2.4: Schematic, top-down representation of the GMM partitioning system. Each box represents a column of sub-partitions viewed from the top. For the highlighted sub-partition also marked with an asterisk (*), the dashed width of the box shows the region in which extra stars with a parallax uncertainty of greater than 1 mas would be included. The height of the dashed box shows the extra overlap between this sub-partition and nearby other sub-partitions. Any cluster with a centroid within the dashed region but not within the main highlighted region was automatically discarded, as it will be better characterised by the neighbouring sub-partition its centroid is in.

of mixtures $\mathcal{O}(nm)$, which for this choice of parameters means it is equivalent to $\mathcal{O}(n^2)$ since the number of mixtures linearly increases with the number of stars. This causes the method to fail the speed criterion (criterion one) from Sect. **??**, even though it was initially believed to be the fastest method under consideration.

To rectify this and ensure that GMMs can still be included in this study, a method for sub-partitioning *Gaia* data chunks was devised. While **cantat-gaudin_gaia_2019** used a k-d tree to partition fields into groups of 8000 stars, this method had no overlap between partitions, and hence may miss OCs that are split between partitions. k-d trees work by splitting random dimensions of a dataset along their median until each branch of the tree is small enough, a process that has no guarantee against splitting a possible OC into many different branches.

Tab. 2.2: Specifications of the GMM sub-partitioning scheme.

distance range (pc)	Max. HEALPix level (overlap)	Max. sub partitions[a]	Optimum m_s
0 - 500	None (None)	1	1000
500 - 750	None (None)	1	1000
750 - 1500	5 (6)	9	800
1500 - 2500	6 (7)	36	600
2500 - ∞	7 (8)	144	250

Notes. [(a)] When fewer than $10m_s$ stars were in a sub-partition, the main HEALPix level was decreased by one to make the sub-partitions a factor of four larger.

Instead, stars in a given field were divided into five segments based on parallax, where stars may be a member of any segment that they have a better than $2\sigma_{\varpi}$ agreement with. Each parallax segment was sub-divided into smaller HEALPix pixels at a specific level. Neighbouring HEALPix pixels were also selected to overlap sub-partitions between each other. The levels of the primary and overlap pixels were carefully selected to ensure that the nearest edge of every sub-partition could always fully contain an OC of 10 pc radius. In the case of the most diffuse OCs, this method could miss some stars that are far from the OC's centre, but should always be able to detect the core of all OCs. When any sub-partition in a parallax range contained fewer than $10m_s$ stars, the main HEALPix level for the parallax segment was decreased by one to increase the number of stars in the sub-partitions. This ensured that no sub-partition was impractically small for later GMM fitting. A schematic representation of this is shown in Fig. **??**, and the values for the sub-partitions are listed in Table **??**.

Any OC candidate with a centroid (λ, ϕ) in an overlap pixel is automatically discarded, as it is assumed to be better characterised in the neighbouring sub-partition for which it would be more fully selected. The sub-partitioning scheme improved the runtime of the method by a factor of about five and the memory use by a factor of about 80. This could be improved further by reducing the pixel sizes or overlap levels, albeit at the cost of sensitivity to OCs on the boundaries between sub-partitions.

Two scenarios were tested in this study for a value of m_s. Firstly, m_s was fixed to 800 stars per mixture component, which was found to be a good general value across the entire dataset. Secondly, m_s was varied depending on the parallax range, as in Table **??**. This was found to greatly improve the sensitivity of the method at high distances where OCs have fewer visible stars in *Gaia* data and are much smaller.

Since GMMs are a method that relies on convergence, the randomly selected starting parameters of the Gaussian mixtures can affect the final result found by the method. Selecting the best result after multiple initialisations was not found to significantly improve results, so the `n_init` parameter of the `scikit-learn` implementation was left at 1. However, the maximum number of iterations of the method, `max_iter`, was set to 1000, to ensure that the method was always able to converge.

2.4 Analysis

2.4.1 Evaluation criteria for clustering algorithms

So far, we prepared *Gaia* data for clustering analysis, selected three algorithms for further study, and developed techniques to optimise them for use on *Gaia* data. In this section, we explain how we quantify the performance of the algorithms against each other by crossmatching to existing objects in the literature, and we present those results.

We quantify the performance of the algorithms using a number of standardised statistics. For our existing literature OCs, we expect that a number of them are real, or true positives (TP). However, literature catalogues such as MWSC have been shown to have a number of erroneous entries (**cantat-gaudin_clusters_2020**), which are in reality true negatives (TN). While a perfect algorithm would report all true positive OCs, missed objects are defined as false negatives (FN). Similarly, when a putative object is erroneously reported as real, it is defined as a false positive (FP).

It is convenient to use these quantities to derive performance statistics normalised to be between 0 and 1, and so we also derive the sensitivity, specificity and precision of the algorithms, which are defined as:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{2.7}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}, \tag{2.8}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{2.9}$$

Effectively, the sensitivity is a measure of an algorithm's ability to detect real objects, the specificity is its ability to reject putative objects, and the precision is the fraction of reported objects that the user could expect are actually real. It follows that a perfect algorithm would have all three quantities at 1, since FN and FP would be zero. However, this is of course unrealistic as no algorithm is likely to be perfect, and different studies may wish to prioritise different statistics over one another. For instance, a search for new OCs would wish to use an algorithm with a maximised sensitivity, such that as many new OCs as possible could be discovered – although the precision of such a study is also of concern, so that as few false positive OCs as possible are reported. A search for existing OCs that attempts to improve the general quality of the OC census would need to maximise all three quantities, and may be especially concerned with maximising the specificity of the method used, such that as many putative literature OCs as possible can be ruled out.

We look in detail at the 100 main OCs of this study and derive sensitivity, specificity and precision statistics for all algorithms in these cases, giving the usefulness of each algorithm and parameter combination when searching for a given literature OC. Then, in the second part of our results, we derive true positive rates for all algorithms across all OCs in the fields in this study, giving supplementary information on the sensitivity of each algorithm as a function of the reported literature distance and size of the OCs.

2.4.2 False positive identification

It is likely impossible to maximise both the specificity and sensitivity of any algorithm simultaneously: for all algorithms studied, increasing their sensitivity would always decrease their specificity. The detection of more true positive OCs always also resulted in more false positive OCs. We explore two techniques to reduce the number of false positives of the algorithms: firstly, by dropping all OC candidates with parameters unrealistic for an OC, and secondly, by using a density-based criterion to discard OC candidates that have a density compatible with being drawn from unclustered local field stars.

To reduce the number of false positive crossmatches – particularly since algorithms such as HDBSCAN and DBSCAN when ran at maximum sensitivity reported over 40 000 OC candidates, the majority of which are false positives – OC candidates with mean parameters extremely incompatible with a real OC were first removed, using criteria presented in **cantat-gaudin_clusters_2020**. The proper motion dispersion of OC candidates was required to satisfy

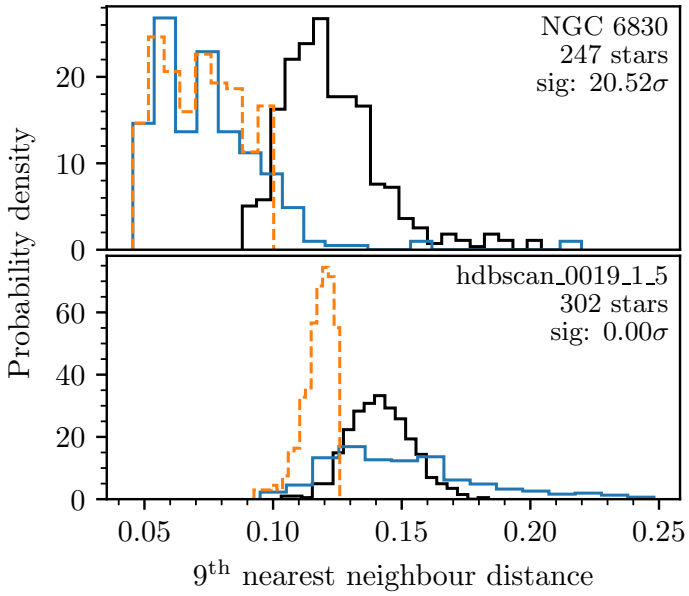

Fig. 2.5: Two examples of NNDs used to test the significance of OC candidates. The solid black line shows the NND of nearby field stars. The blue line shows the NND of distances between cluster members. For a cluster to be significant and not simply a selection of unclustered field stars, the cluster NND must be incompatible with being drawn from the field distribution. For later illustrative purposes, the NND of a cluster member to the nearest field star is shown by the dashed orange line, although this is not used for the CST. In the upper plot, an OC candidate detected by HDBSCAN and crossmatched to the well-characterised OC NGC 6830 is shown, which has a clearly different NND to field stars with a significance of over 20σ. In the lower plot, a false positive OC detected by HDBSCAN in field 19 is shown that has a significance of 0σ.

$$\sqrt{\sigma^2_{\mu_{\alpha*}} + \sigma^2_{\mu_\delta}} \leq \begin{cases} 1 \text{ mas yr}^{-1} & \varpi \leq 0.67 \text{ mas} \\ 1.49 \cdot \varpi \text{ mas yr}^{-1} & \varpi > 0.67 \text{ mas.} \end{cases} \tag{2.10}$$

The radius containing half of the members of the OC candidate was also required to satisfy $r_{50} < 20$ pc. These constraints were relatively weak, only removing a small number of clearly anomalous OC candidates that had velocity dispersions or radii that were clearly incompatible with real OCs.

Secondly, a method was implemented to compare the density of OC candidates with the density of local field stars and evaluate the significance of the OC candidate, hence referred to as the cluster significance test (CST). Ninth nearest neighbour distances between stars were used as a proxy for density, as this corresponds exactly to how two of the three methods in this study performed clustering analysis (since they used $m_{Pts} = 10$, i.e. $k = 9$) and since this value is free of contamination from

binary or multiple star systems, since they will have significantly smaller first or second nearest neighbour distances.

To calculate the density distribution of a cluster, the nearest neighbour distribution (NND) of intra-cluster distances between stars within an OC candidate was calculated. Then, in an iterative approach, a minimum of 100 and a maximum of 500 local field stars were found around the OC candidate by traversing the graph of nearest neighbours and looking for field stars with NNDs uncontaminated by proximity to the cluster, meaning that none of their 1^{st} to 9^{th} nearest neighbours were labelled as cluster members. This approach was found to generate reliable and quick approximations of the NND of local field stars.

Since a good OC candidate is a clear overdensity in the parameter space, its NND should be incompatible with being drawn from the distribution of field stars. A number of statistical tests were investigated to test this, with a Mann-Whitney U test (**mann_test_1947**) found to be the most reliable, since it makes no assumptions about the shape of the distribution and does not require the distribution to be continuous. Significance values for each OC candidate are then derived from a one-tailed test where the alternate hypothesis is that the OC candidate has an NND incompatible with and with a lesser median than the field NND.

Requiring a CST value of at least 3σ was found to keep the vast majority of good OCs while identifying and removing a large number of false positives for all algorithms. For instance, for DBSCAN when running with ϵ_{n3} (the algorithm and parameter combination that produced the highest number of OC candidates), the CST constraint reduced the number of reported OC candidates from 51920 to just 1111 objects.

2.4.3 Crossmatches with existing catalogues

Having greatly reduced the number of false positives identified by all algorithms, we crossmatched OC candidates against literature clusters to estimate the number of true positives detected by each algorithm. However, this process is non-trivial, with each catalogue reporting OCs in different ways.

A number of approaches were trialed to crossmatch OC candidates. The best approach found to crossmatch OC candidates' positions was that of **liu_catalog_2019**. The tidal radius of OC candidates is estimated as the maximum distance of a member star from its mean α and δ. The OC candidate must be within one tidal radius of the reported position in the literature, where whichever tidal radius is larger (that of the

candidate or that of the literature cluster) is used. This would typically correspond to searching in a radius of no more than $0.5°$.

$\mu_{\alpha*}$, μ_δ and ϖ for OC candidates were required to be within 5σ (5 standard errors) of literature values. It has been shown that *Gaia* DR2 has a number of small unaccounted for systematic effects, including a parallax zero-point offset ϖ_0 that may be magnitude-dependent (**lindegren_gaia_2018**). As such, even when crossmatching to other OCs detected in *Gaia* data, magnitude-dependent systematic errors could cause crossmatches to fail. For instance, **castro-ginard_hunting_2020** have only studied *Gaia* data to $G = 17$. Extra stars introduced by this study using a magnitude cut of $G = 18$ will have a different mean systematic effect on derived astrometric parameters. Additionally, every clustering algorithm will report slightly different membership lists for each OC, and the differences in parameters of included or ignored members could introduce different systematic errors. In (α, δ), these effects are small, since tidal radii (often no smaller than $\sim 0.1°$) are much larger than the small systematic errors in position of the *Gaia* reference frame. However, large OCs especially may have standard errors on their mean parallax or proper motion as small as 10 μas or 10 μas yr^{-1}, smaller than the reported *Gaia* systematic errors.

To rectify missed crossmatches, small tolerances to uniform systematic errors of 50 μas yr^{-1} and 50 μas were accounted for in crossmatching of proper motions and parallaxes respectively. These values were selected to roughly account for the scatter in parallax and proper motion offsets as a function of magnitude as reported by **lindegren_gaia_2018**. This allowed a number of larger OC candidates with very small uncertainties (**cantat-gaudin_clusters_2020**) to be successfully crossmatched. Many of these large OC candidates were visible by eye in the *Gaia* data and in the reported results of the clustering algorithms, and were being missed in the crossmatch procedure by a lack of tolerance to systematic error and due to their small uncertainties on parameters owing to their large size.

As the only non- *Gaia* catalogue, MWSC was more complicated to crossmatch against. Reported distances to OCs were converted to parallaxes. While distance measurements in MWSC do not include uncertainties, the estimated 11% systematic uncertainty on distance measurements reported by **kharchenko_global_2013** was accounted for. A parallax offset of $\varpi_0 = -0.029$ mas was applied to MWSC parallaxes, ensuring that they have the same mean systematic offset as parallaxes in the *Gaia* DR2 dataset as reported by **lindegren_gaia_2018**. The additional ± 0.8 mas yr^{-1} external error in MWSC proper motions was also accounted for, which resulted in a handful of extra crossmatches to objects clearly crossmatched in other dimensions that had large offsets in their proper motions relative to *Gaia* DR2.

2.4.4 Results

Finally, we present analysed results of the algorithms for discussion in three parts.

Firstly, we inspected *Gaia* data manually to assign the 100 main OCs as either true positives or true negatives. An interactive data viewer was used to explore the region around the reported locations of the OCs, searching for significant overdensities within the possible crossmatch region. We also required that the detected overdensity had a colour magnitude diagram (CMD) compatible with an OC, for which we define the following criteria.

A class one OC has a clear, difficult to dispute CMD, with a realistic shape. The CMD may be somewhat broadened by differential extinction or inhomogeneities, but there should be enough stars present to make the probability of a false alarm very small. However, a class two OC is a possible OC that may be too small or too inhomogeneous for its true existence to be clear. It may be that only the brightest stars (near the turnoff point) are detected, making its shape difficult to discern as a true isochrone. There may be a small number of outlier stars incompatible with an isochrone, owing to a poor detection by the algorithm. This class signifies that more work would be needed to confirm this object as an OC. Finally, class three OCs are very unlikely to be an OC and much more compatible with random noise. Even if some stars follow an isochrone, a significant number are outliers, owing to this being a selection of unclustered, inhomogeneous stars.

After an overdensity was isolated in position, proper motion and parallax, it was required to have a class one or two CMD to confirm it as a true positive OC. We assigned 40 OCs as true positives and the remaining 60 as true negatives.

31 of the 33 OCs in the list of main OCs from MWSC that are also in the catalogue of **cantat-gaudin_clusters_2020** were entered as true positives. Most of these objects were good OCs that were clearly visible at their reported location. We did not detect significant overdensities with class one or two CMDs corresponding to Patchick 75 or Auner 1, both of which are distant OCs with distances in **cantat-gaudin_clusters_2020** of $\sim$7 kpc and $\sim$8 kpc respectively. These OCs are heavily polluted in the literature membership lists. If real, they are scarcely detectable in *Gaia* data. Alternatively, they may simply not be real objects.

Most of the additional 67 OCs listed in MWSC do not appear detectable in *Gaia* data, and may simply not be real objects. However, we did find sparse overdensities corresponding to nine objects from MWSC: ASCC 28, ASCC 100, ASCC 130, BDSB 124, Berkeley 64, DBSB 164, IRAS 06046-0603, SAI 90 and Teutsch 146.

After reducing the number of false positives in the results of the algorithms using the techniques from Sect. **??**, we crossmatched their results to the main list of 100 OCs and derived performance statistics. To quantify uncertainty on derived statistics, we used the method for computing Bayesian binomial confidence intervals described in **cameron_estimation_2011**, where a Beta distribution with an uninformative prior is used to estimate a confidence interval containing the true success fraction given the measured success fraction. The performance of the algorithms is listed in Table **??**.

Five OCs from the 40 true positives are never detected by any algorithm within our constraints, which we discuss here for completeness: Berkeley 91 and Teutsch 156 (**cantat-gaudin_clusters_2020**) as well as ASCC 28, BDSB 124 and Teutsch 146 from MWSC. Berkeley 91 is relatively distant (~ 4 kpc) OC with a polluted CMD in the catalogue of **cantat-gaudin_clusters_2020**. If real, it is barely detectable in *Gaia* data. Teutsch 156 appears to be detected by HDBSCAN, but only tentatively with a CST of 0.68σ. ASCC 28 should be detected, as the detected overdensity was nearby with a parallax of 0.85 mas. It may be too sparse for an algorithm to detect or may be an association mis-classified by the expert classifier. The BDSB 124 and Teutsch 146 overdensities are distant ($\varpi \approx 0.3$ mas and 0.25 mas respectively) with polluted CMDs. These objects may be difficult for algorithms to detect in *Gaia* data or may simply be associations.

Six OCs from MWSC listed as true negatives are reported at some point by any algorithm (five by HDBSCAN, two by DBSCAN), although only OC candidates crossmatched to FSR 0316 are detected by two different algorithms (HDBSCAN and DBSCAN). In all of these cases, the objects are sparse and relatively separated from the reported literature locations on the sky, and may be new OCs that marginally coincide with the existing locations.

Secondly, we performed crossmatches to all 1385 targets listed in the literature in the fields of this study. While there are too many OCs to conduct a precise by-hand treatment of the results, these results give a better indication of the dependence of the algorithms' sensitivities on OC features like distance, age and size. Clear dependencies on distance and size are found, which are shown in Fig. **??**. No significant dependence on ages listed in MWSC is found for any of the algorithms, although the more recent and accurate age catalogue of **cantat-gaudin_painting_2020** did reveal a slight dependence on age that appears to result from the smaller size of older OCs. HDBSCAN is the most sensitive algorithm across all ages. When combining all ϵ runs, DBSCAN is as sensitive for well populated young OCs, but is less sensitive to older, typically smaller OCs. GMMs are the worst algorithm across all ages.

Finally, to compare the general usability of the algorithms, we list the runtimes and the total number of OC candidates with valid proper motion dispersions and radii reported by each algorithm in Table **??**. Ideally, an algorithm would report a realistic number of OC candidates in as little time as possible.

We also list additional comparisons of our results with other catalogues in Appendix **??**. Full tables of detected and non-detected objects (including OC membership lists) are available in the online material only, with descriptions of their content in Appendix **??**.

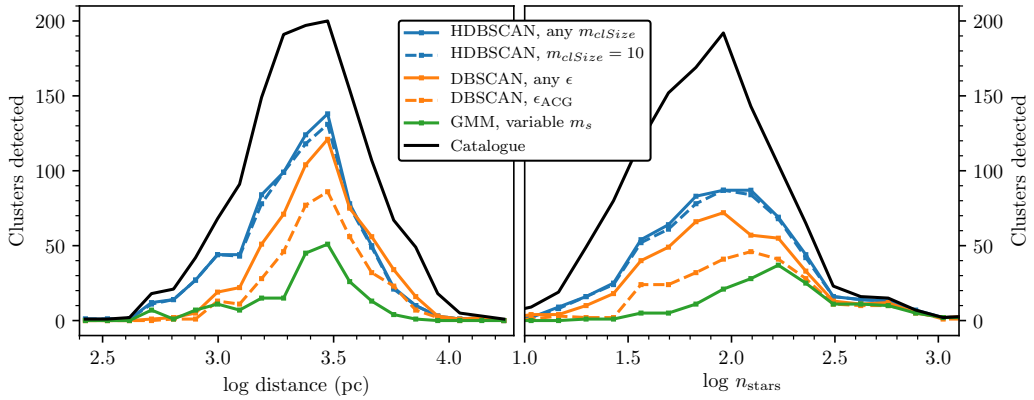

Fig. 2.6: Distance and size dependence of detections by different algorithm and parameter combinations for all 1385 OCs in all studied fields, plotted against the reported size and distance of the OCs in the literature. OC candidates not passing the criterion in Sect. **??** and with a CST of less than 3σ were discarded. HDBSCAN detects the most OCs, especially at nearby distances. GMMs only perform well at detecting well populated OCs. While individual DBSCAN results at different ϵ values do not detect especially many OCs, combining them all together nearly matches the performance of HDBSCAN – even exceeding it slightly at large distances.

2.5 Comparison of algorithms

Finally, having selected three algorithms for further study and having ran them on 100 representative fields across the galactic disk, we address the central topic of this work as to which clustering algorithm is best at detecting OCs in *Gaia* data. We discuss the pros and cons of each algorithm in subsections before presenting an opinion.

Tab. 2.3: Performance of different algorithm and parameter combinations on the 100 main OCs.

Algorithm	Parameters	TP	FP	TN	FN	Sensitivity	Specificity	Precision
DBSCAN	ϵ_{ACG}, 1 repeat	$22^{25.0}_{18.8}$	$0^{1.8}_{0.0}$	$60^{60.0}_{58.2}$	$18^{21.2}_{15.0}$	$0.55^{0.62}_{0.47}$	$1.00^{1.00}_{0.97}$	$1.00^{1.00}_{0.91}$
-	ϵ_{ACG}, 30 repeats	$20^{23.1}_{16.9}$	$0^{1.8}_{0.0}$	$60^{60.0}_{58.2}$	$20^{23.1}_{16.9}$	$0.50^{0.58}_{0.42}$	$1.00^{1.00}_{0.97}$	$1.00^{1.00}_{0.90}$
-	ϵ_c	$20^{23.1}_{16.9}$	$1^{3.2}_{0.7}$	$59^{59.3}_{56.8}$	$20^{23.1}_{16.9}$	$0.50^{0.58}_{0.42}$	$0.98^{0.99}_{0.95}$	$0.95^{0.97}_{0.84}$
-	ϵ_{n1}	$20^{23.1}_{16.9}$	$0^{1.8}_{0.0}$	$60^{60.0}_{58.2}$	$20^{23.1}_{16.9}$	$0.50^{0.58}_{0.42}$	$1.00^{1.00}_{0.97}$	$1.00^{1.00}_{0.90}$
-	ϵ_{n2}	$7^{10.0}_{5.2}$	$2^{4.5}_{1.4}$	$58^{58.6}_{55.5}$	$33^{34.8}_{30.0}$	$0.17^{0.25}_{0.13}$	$0.97^{0.98}_{0.93}$	$0.78^{0.88}_{0.54}$
-	ϵ_{n3}	$2^{4.4}_{1.3}$	$0^{1.8}_{0.0}$	$60^{60.0}_{58.2}$	$38^{38.7}_{35.6}$	$0.05^{0.11}_{0.03}$	$1.00^{1.00}_{0.97}$	$1.00^{1.00}_{0.43}$
-	$\{\epsilon_c, \epsilon_{n1}, \epsilon_{n2}, \epsilon_{n3}\}$	$25^{27.8}_{21.8}$	$2^{4.5}_{1.4}$	$58^{58.6}_{55.5}$	$15^{18.2}_{12.2}$	$0.62^{0.69}_{0.54}$	$0.97^{0.98}_{0.93}$	$0.93^{0.95}_{0.83}$
HDBSCAN	$m_{clSize} = 80$	$17^{20.2}_{14.1}$	$3^{5.7}_{2.1}$	$57^{57.9}_{54.3}$	$23^{25.9}_{19.8}$	$0.42^{0.50}_{0.35}$	$0.95^{0.97}_{0.91}$	$0.85^{0.91}_{0.71}$
-	$m_{clSize} = 40$	$24^{26.8}_{20.8}$	$6^{9.2}_{4.4}$	$54^{55.6}_{50.8}$	$16^{19.2}_{13.2}$	$0.60^{0.67}_{0.52}$	$0.90^{0.93}_{0.85}$	$0.80^{0.86}_{0.69}$
-	$m_{clSize} = 20$	$31^{33.1}_{27.9}$	$7^{10.3}_{5.2}$	$53^{54.8}_{49.7}$	$9^{12.1}_{6.9}$	$0.78^{0.83}_{0.70}$	$0.88^{0.91}_{0.83}$	$0.82^{0.86}_{0.73}$
-	$m_{clSize} = 10$	$33^{34.8}_{30.0}$	$7^{10.3}_{5.2}$	$53^{54.8}_{49.7}$	$7^{10.0}_{5.2}$	$0.82^{0.87}_{0.75}$	$0.88^{0.91}_{0.83}$	$0.82^{0.87}_{0.74}$
GMM	$m_s = 800$	$7^{10.0}_{5.2}$	$0^{1.8}_{0.0}$	$60^{60.0}_{58.2}$	$33^{34.8}_{30.0}$	$0.17^{0.25}_{0.13}$	$1.00^{1.00}_{0.97}$	$1.00^{1.00}_{0.75}$
-	$m_s =$ variable	$13^{16.2}_{10.4}$	$0^{1.8}_{0.0}$	$60^{60.0}_{58.2}$	$27^{29.6}_{23.8}$	$0.33^{0.41}_{0.26}$	$1.00^{1.00}_{0.97}$	$1.00^{1.00}_{0.85}$

Notes. True positive (TP), false positive (FP), true negative (TN) and false negative (FN) counts of detected clusters are given along with the sensitivity, specificity and precision. 68.3% confidence intervals are shown for all numbers. Confidence intervals for a handful of values (e.g. measured precisions of exactly 0.0 or 1.0) were adjusted to include the measured values. This corrects for approximations in the calculation of binomial confidence intervals where the measured success probability is exactly 0 or 1. All objects that did not pass the criterion in Sect. **??** with a CST greater than 3σ were discarded before crossmatching and producing this table.

Tab. 2.4: Extra information on the algorithms' performance.

Algorithm	Reported OC candidates[a]	Fraction with CST $> 3\sigma$	Total crossmatches[b]	Mean runtime (mins)[c]
DBSCAN (ACG)	1518 to 1538	58.9% to 59.6%	382	1.19 (1 repeat) to 10.3 (30 repeats)
DBSCAN (model)	5212 to 51920	22.4% to 2.1%	593	0.885
HDSBCAN	1196 to 49693	82.0% to 5.2%	756	2.36
GMM	314 to 2465	60.5% to 20.5%	213	21.9 ($m_s = 800$) 47.0 (variable m_s)

Notes. [(a)] Total number of OC candidates for all fields that passed the proper motion dispersion and radius constraints from Sect. **??**. The range is between the minimum and maximum reported number for the least and most sensitive parameters. [(b)] Total crossmatches is given as the union of all results from all parameter sets for a given method. A total of 1385 literature OCs were crossmatched against. [(c)] Mean runtime for a single field out of the 100 in this study. All runs were conducted on the same workbench computer with a 3.1 GHz 4-core CPU.

2.5.1 DBSCAN is effective at searching for OCs

DBSCAN is a well proven algorithm on *Gaia* data, having recently detected over 500 new OCs in **castro-ginard_hunting_2020**. It performed relatively well on the 100 main OCs in this study. ϵ_{ACG} has particularly high specificity and precision values of ≈ 1.00 (Table **??**), suggesting that DBSCAN can produce consistent and reliable results when not ran sensitively. However, even when greatly increasing its theoretical sensitivity (e.g. ϵ_{n3}, the highest ϵ value used in this study), DBSCAN still is not able to detect all OCs present in a field.

At individual values of ϵ, Fig. **??** shows that DBSCAN is most sensitive to OCs at certain distances, with the ϵ_{ACG} sensitivity peaking at 3.1 kpc. This is likely due to how distant OCs are very compact in all dimensions, while nearby OCs in the sample may have radii of up to 0.5° or more, and are hence much sparser in the two positional dimensions and require the global density threshold ϵ to be higher for them to be completely detected. This is a key disadvantage of DBSCAN: single, global ϵ parameters rarely seem to be perfect for individual OCs, especially when the global parameter is influenced by density contributions from many different OCs in a single field.

Manual comparison between algorithm results shows that DBSCAN often under or over-selects OCs and produces less reliable membership lists than HDBSCAN or GMMs. Over-selection is a particular issue as CMDs become polluted and the performance of OC candidates in the CST is reduced, as the nearest neighbour distribution becomes dominated by contaminating field stars. Many OC candidates detected by DBSCAN at CST values of less than 3σ appear to correspond to real OCs, but are too polluted or too sparse to pass criterion to verify the candidate objects as real. In addition, these membership lists containing too few or too many members are of less use to other scientific applications, and would need to be followed up with another algorithm to improve their quality.

This can be partially mitigated by combining all DBSCAN results across all ϵ values, which approaches a similar degree of sensitivity to HDBSCAN, albeit still with a deficit of detections for small distances at less than 1 kpc. This result is in good agreement with what theory presented in Sect. **??** suggests: that a single run of DBSCAN's global density parameter will only be sensitive to a certain size of OC at a given distance, and that running HDBSCAN is equivalent to running DBSCAN across all possible values of ϵ. However, HDBSCAN is better still – able to detect the majority of OCs in the sample in a single run at $m_{clSize} = 10$.

Combining multiple DBSCAN results is similar to the effective approach that **castro-ginard_hunting_20** use to detect over 500 new OCs, since they vary the m_{Pts} parameter and the size of the field analysed by the algorithm. ϵ_{ACG} results presented here should be less sensitive than the results of **castro-ginard_hunting_2020** as they are based on a single DBSCAN run at a single value of $m_{Pts} = 10$ and a single size of field. However, m_{Pts}, ϵ and the size of the field under consideration are not entirely independent parameters, since changing ϵ has a similar effect to how changing either of the others improves DBSCAN sensitivity in **castro-ginard_hunting_2020**. It is not possible to quantitatively compare the sensitivity of DBSCAN methods in this study to that of **castro-ginard_hunting_2020**, since they use a different cut on the *Gaia* dataset ($G = 17$) and autonomous CMD classification to remove false positives, which will have reduced their sensitivity to faint OCs or distant objects with high CMD contamination. They detect 688 (55.9%) of the total number of OCs reported in **cantat-gaudin_gaia_2018**, although the 688 correspond to 81% of objects in **cantat-gaudin_gaia_2018** with a significant number of members brighter than $G = 17$ and a well defined isochrone, which are the objects that their study would be theoretically sensitive to. The combination of all DBSCAN runs in this study was able to detect 343 out of 537 (63.9%) of the OCs in this study from **cantat-gaudin_clusters_2020** at a CST of greater than 3σ.

Future use of DBSCAN could benefit from repeated runs while only changing ϵ opposed to the size of the field, which we expect to be both as sensitive but more computationally efficient. The most computationally expensive step (calculation of nearest neighbour distances for a given field) would only need to be performed once, with DBSCAN then evaluated quickly on the same matrix of nearest neighbour distances but simply with a wide range of different ϵ values.

For ϵ determination for DBSCAN, both methods appear viable, although the ACG method is more numerically stable. ϵ_c results were largely analogous to results produced by ϵ_{ACG} – although occasionally, on more difficult fields, the model fit would be less stable and would over-estimate the optimum value of ϵ. This is clear in Table **??**, where the ACG method has a very good precision of 1.00 compared with 0.95 for ϵ_c results.

When running on large fields such as those in this study, the ACG method's random field resampling only needs to be repeated once, since no improvement is visible in crossmatch statistics between resampling 30 times versus only performing it once. The measured sensitivity of the ACG method with a single repeat is slightly better than using 30 repeats (0.55 vs. 0.50), although this difference is not statistically significant. Using only a single repeat is also an order of magnitude faster than

doing 30 repeats, although the field modelling approach is faster still, being 25% faster than the ACG method with a single repeat.

While naturally, the ACG method only produces a single ϵ estimate, it could easily produce more by using multiples of ϵ_{ACG} in the range $[1,\, 2.5]$ to approximate results between ϵ_c and ϵ_{n3} and to combine multiple different-ϵ runs to improve sensitivity.

Overall, DBSCAN is an effective and well-proven methodology. In particular, its high precision at low sensitivities when used with the ACG method makes it an excellent choice for a limited blind search for good OC candidates. However, it was unable to detect all OCs present in a given field, and is not reliable for producing complete, minimally polluted membership lists for OCs, since ϵ is a global parameter and will inherently not be optimised for individual OCs in a given field when a field contains multiple different objects.

2.5.2 HDBSCAN solves many issues encountered by DBSCAN, but is not without flaws

Many of the issues with DBSCAN are solved with HDBSCAN. Parameter determination and setup of the method for HDBSCAN is significantly easier, since the minimum size of an OC m_{clSize} is a much more intuitive choice for a parameter than one based on nearest neighbour distances for a given dataset, ϵ. In terms of sensitivity, individual runs (such as $m_{clSize} = 10$) are able to outperform all DBSCAN results combined, detecting the highest number of true positive OCs in the study. However, this increased sensitivity comes at a cost: HDBSCAN results generally have the worst specificity and precision scores of any algorithm in this study, with a large number of false positives and poor characterisation of true negatives (especially for $m_{clSize} = 10$.) This would be even worse when not using the CST to reduce false positives: $m_{clSize} = 10$ results without a CST restriction had a precision of just 0.47 and a specificity of 0.28, owing to a huge number of reported false positives. Clearly, to be used effectively, HDBSCAN must also be used with criteria to select valid clusters from its results.

This appears to happen because of how HDBSCAN autonomously decides local thresholds for if objects are or are not a cluster. Often, HDBSCAN reports OC candidates in the densest regions of the dataset. These objects are clearly not OCs, but simply features of the underlying shape of the data, since the *Gaia* satellite samples a magnitude-limited spherical volume with different observed densities at different distances. This is demonstrated well by some of the existing analysis in

this work. In Fig. **??**, two false positive clusters were reported alongside Blanco 1 due to how HDBSCAN effectively considers all possible DBSCAN solutions, which includes erroneously reporting two small and impersistent clusterings of field stars as OC candidates. Uniform noise follows a nearest neighbour distribution given by Eqn. **??**, which implies that field stars will have a smooth range of different nearest neighbour distances. However, when more than m_{clSize} stars exist on the dense end of this curve, HDBSCAN erroneously assigns them into a cluster, even though they are simply a feature of the random nature of the unclustered stars.

This is also demonstrated by the orange curve in the lower panel of Fig. **??**, where the nearest neighbour distributions of a false positive OC are plotted. The 302 stars in this false positive OC have an external (i.e. to the nearest star) density distribution that is simply a dense slice of the field star nearest neighbour distance distribution. This orange curve is analogous to what HDBSCAN and other density-based clustering algorithms use to assign stars as cluster members. However, when looking at the internal nearest neighbour distance distribution (i.e. distance to the nearest cluster member), it becomes clear that these stars are not self-consistent with being a separate, dense object, and still appear to be drawn from a nearest neighbour density distribution that is the same as that of the local field stars.

An additional issue is that increasing the sensitivity of HDBSCAN sometimes causes it to miss certain OCs. These are typically large, clearly real objects that are mistakenly split apart into multiple substructures for low m_{clSize} values. To detect all OCs in a future all-sky survey, it would be necessary to run with multiple parameters and combine runs as with DBSCAN. While Fig. **??** shows that the effect is not as significant as with DBSCAN, combining multiple runs still provides a small increase in the total number of OCs detected.

HDBSCAN detects slightly fewer OCs than the combination of all DBSCAN results for distances of greater than 4 kpc, as shown in Fig. **??**. This appears strange at first glance, as HDBSCAN should consider all DBSCAN solutions and should in theory be able to detect all objects DBSCAN can detect. On closer inspection, it appears that this is due to HDBSCAN's approach to membership lists, since HDBSCAN includes all objects that could be cluster members but will assign them correspondingly low membership probabilities. Distant OCs are difficult to separate from field stars, as proper motions and parallaxes become decreasingly informative at large distances – and for these objects, HDBSCAN often includes many low probability members that reduce the quality of the detection and of the CMD of the objects.

At relatively large distances, these low probability members cause HDBSCAN to perform worse in our study. The CST does not currently consider membership

probabilities, meaning that low probability members that are more likely to be members of the field would reduce the measured significance of some distant OC candidates. In the future, the CST should be modified to also include membership probabilities.

Despite some shortcomings, it appears that properly handling these (e.g. with the CST or another test to remove false positives based on their density) allows HDBSCAN to be used as a powerful method for OC detection. Its runtime is not significantly longer than DBSCAN (see Table **??**), yet it is able to detect more OCs across a wider range of distances. In addition, HDBSCAN's membership lists for validated OC candidates were typically very clean, often even detecting tidal structures for OCs due to its excellent recovery of clusters across all density levels. There is room for more improvement of HDBSCAN results at distances greater than 4 kpc by optimising validation criteria to also make use of its membership probabilities.

2.5.3 GMMs are inappropriate for large-scale OC blind searches in the *Gaia* era

While HDBSCAN is somewhat slower than DBSCAN, both are significantly faster than GMMs. Despite receiving the most time investment from the authors into optimising the algorithm for use on OCs, it still under-performed relative to HDBSCAN and DBSCAN by an order of magnitude in runtime. As an $\mathcal{O}(n^2)$ algorithm when used with the optimum parameters, it scales poorly to the large *Gaia* dataset of many millions of stars per field in the densest regions. This is especially noticeable in the maximum single-field runtimes of GMMs. The single densest field took 20.1 hours to run for $m_s =$ variable, a factor of around 40 times slower than HDBSCAN on the same field. GMMs are simply too slow for practical use with unsupervised searches through *Gaia* data, and it would take many months to run on the entirety of *Gaia* DR2 in its current implementation in this study without using a supercomputer.

In the test on the 100 main OCs, GMMs had the lowest maximum sensitivity of any algorithm, detecting just 33% of the true positive OCs. However, it did perform well in the specificity and precision metrics, even without the CST. The built-in validity constraints of the GMM method on the proper motion and radial dispersion of OC candidates ensure that all reported candidates are already of high quality. This is also evident in Table **??**, where GMMs reported a relatively small maximum number of OC candidates (2465 for the $m_s =$ variable run), although this is still not as good as the DBSCAN ACG method, for which 1538 OC candidates were reported at most

– yet the sensitivity in the study of 100 OCs was around 60% greater, suggesting that the DBSCAN ACG method is still more efficient at producing crossmatches to existing OCs.

A further disadvantage of GMMs is their sensitivity to the number of stars per component, m_s: reducing this number allows the method to detect smaller OCs, but greatly increases the runtime of the algorithm, since the runtime complexity linearly scales with the number of components of the GMM. As shown in Fig. **??**, the variable m_s values used in this study are still too large to detect many smaller clusters, with the algorithm performing poorly for any objects with fewer than around 160 reported members. In addition, low values of m_s begin to cause larger OCs to be erroneously split into separate objects that would require either multiple runs of the GMM algorithm at different parameters or a scheme to merge nearby clusters after a run.

More OCs may be detected by GMMs by removing outlier stars from the dataset. While this approach is not favoured by this study as it introduces biases into the running of the algorithms, cutting stars with high proper motions (**cantat-gaudin_gaia_2019**) simplifies the likelihood maximisation process for the algorithm. Sometimes, the algorithm would place individual stars with extremely high proper motions into single-star clusters, since this maximises the likelihood of the overall model fit. However, this is counterproductive, as GMM components are wasted on individual stars at high proper motions and are no longer available to fit to OCs.

While this study shows that GMMs are not scalable to a large-scale blind search, it is still a useful method for deriving membership lists for the cores of OCs. GMM OC membership lists are typically very clean. When the location of an OC is known to high accuracy beforehand, GMMs can be applied quickly to a heavily cut dataset to derive a membership list. This mirrors the success of works using UPMASK to derive membership lists for existing OCs (**cantat-gaudin_gaia_2018**; **cantat-gaudin_clusters_2020**), since UPMASK uses K-Means clustering (an algorithm closely related to GMMs) to derive OC membership lists. This approach assumes that the reported location of an OC is accurate enough to allow a dataset to be effectively cut such that an algorithm with an effective runtime complexity of $\mathcal{O}(n^2)$ can be applied in a reasonable amount of time.

Tab. 2.5: Mean parameters for a selection of the new OCs detected in this study.

Name	α (°)	δ (°)	l (°)	b (°)	$\mu_{\alpha*}$ (mas yr^{-1})	μ_δ (mas yr^{-1})	ϖ (mas)	r_{50} (°)	n	$\sigma_{\rm CST}$
PHOC 1	126.99	-42.77	260.83	-2.44	-5.74 (0.03)	4.79 (0.02)	0.67 (0.00)	0.11	32	8.64
PHOC 2	280.11	-3.75	28.34	0.73	0.46 (0.02)	-1.59 (0.02)	0.36 (0.00)	0.09	47	5.94
PHOC 3	115.83	-30.48	245.65	-3.39	-2.03 (0.01)	2.35 (0.02)	0.40 (0.01)	0.09	30	5.72
PHOC 4	106.79	-7.69	221.57	-0.03	-3.80 (0.04)	1.10 (0.02)	0.91 (0.01)	0.22	71	9.96
PHOC 5	105.88	-7.78	221.23	-0.88	-0.66 (0.01)	-1.07 (0.02)	0.79 (0.01)	0.13	39	6.91
PHOC 6	280.59	-7.23	25.46	-1.29	0.88 (0.02)	-2.85 (0.02)	0.38 (0.01)	0.06	38	7.46
PHOC 7	285.73	14.58	47.21	4.11	-0.41 (0.02)	-3.15 (0.01)	0.49 (0.01)	0.14	28	5.92
PHOC 8	288.83	14.43	48.47	1.37	-1.69 (0.02)	-2.50 (0.02)	0.34 (0.01)	0.08	39	9.26
PHOC 9	79.59	41.99	166.04	2.51	0.13 (0.02)	-0.45 (0.01)	0.20 (0.00)	0.07	43	6.56
					⋮					
PHOC 39	277.78	-3.81	27.22	2.77	1.89 (0.05)	-8.75 (0.05)	2.49 (0.02)	0.48	139	15.10
PHOC 40	287.77	14.27	47.85	2.21	-1.57 (0.06)	-9.41 (0.09)	2.99 (0.02)	0.49	36	7.65
PHOC 41	282.55	33.41	63.24	14.78	1.85 (0.08)	-3.84 (0.07)	3.42 (0.02)	0.37	63	9.42

Notes. Standard errors for mean proper motions and parallaxes are shown in the brackets. The full version of this table (including extra columns) is available in the online material only, following the format of Table **??** except with column 26 omitted.

2.6 New OC candidates in the galactic disk

2.6.1 Methodology

During the preparation of this work, we discovered that many of the algorithms' reported OC candidates did not crossmatch to literature targets and appeared to be distinct, new objects. We investigated this further to see if any of the objects are genuine new OC candidates.

Firstly, we made conservative cuts on our reported OC candidates to select only high-quality objects. All objects failing the criteria from Sect. **??** or with a CST of less than 5σ were discarded, meaning that our sample of candidates only represents definitive astrometric overdensities. In addition, any objects with a centre closer than 1.5 estimated tidal radii to the edge of the field they were detected in were discarded, removing any objects that could have a remote possibility of issues due to edge effects.

Secondly, we performed extra crossmatching to the catalogues of **dias_new_2002**, **bica_multi-band_2018**, **sim_207_2019**, **ferreira_discovery_2020** and **qin_discovery_2021**. To the best of our knowledge, they in addition to the four catalogues from earlier in this work include all reported literature OCs from at least the past two decades.

After the crossmatching and the cuts, all algorithms still appeared to detect new OCs, but the most were found by HDBSCAN. At the high CST threshold of 5σ, any objects found by DBSCAN or GMMs were almost always also found by HDBSCAN, and so for simplicity we only looked at the results of HDBSCAN with $m_{clSize} = 20$, since merging the results of different algorithm and parameter combinations would be non-trivial and is beyond the scope of this work.

This produced a list of 102 tentative objects based on astrometry and crossmatching alone. A small fraction of these had CMDs that were clearly random selections of unassociated stars that followed no clear isochrone, although many others were borderline objects with poor quality CMDs. We manually selected only objects with good or relatively good quality CMDs, leaving a list of 38 new OCs. While this study was not optimised to find nearby objects, we noticed that some of the 76 objects closer than 1 kpc that were discarded because of edge effects could be real, new OCs. We investigated the 12 most promising objects by downloading new regions of *Gaia* data around them and re-running HDBSCAN, finding that three of these objects are of a good quality and bringing our total to 41 new objects.

We name the objects with the acronym PHOC (Preliminary HDBSCAN Open Cluster) as we expect to characterise these objects further in future works. Mean parameters for a selection of these objects are shown in Table **??**, with a full list of mean parameters and members for these new objects included in the online material. Extra descriptions of the contents of these tables are included in Appendix **??**. In addition, plots of all new objects are included in Appendix **??**.

2.6.2 Comments on the new OC candidates

We present brief remarks on some of the newly reported OC candidates.

Comparing our list of new OC candidates with the DBSCAN blind search of **castro-ginard_hunting_2020** reveals patterns similar to those shown earlier in this work in Fig. **??**. Whereas the 209 OC candidates from their work have a median distance of 2650 pc with a highest individual distance of 8400 pc, our 41 objects detected by HDBSCAN have a closer median distance of 1940 pc with a maximum distance of 4400 pc. These results are in agreement with our earlier finding that HDBSCAN is more sensitive to nearby OCs whereas DBSCAN is more sensitive to more distant OCs. However, as discussed in Sect. **??**, this may simply be an artefact of our study as our CST as-implemented gives higher scores to more nearby clusters. Our future works will report more tentative candidates with lower CST scores and may be able to achieve similar sensitivities as DBSCAN with HDBSCAN.

Our three very nearby candidates within 500 pc (PHOC 39, PHOC 40, and PHOC 41) are some of the most scientifically interesting. If real, these objects demonstrate that new clusters are yet to be found even at close distances. We estimated approximate distances to these objects as the inverse parallax after correcting for the *Gaia* zero-point offset (**lindegren_gaia_2018**), although our future works will use a more sophisticated inference-based approach. All three objects are within the galactic disk.

PHOC 39 has an estimated distance of 396 pc, 139 member stars and a CST score of 15.1σ. While it has a broad CMD as reported, plotting only member stars with a membership probability of at least 80% gives a much cleaner and less broadened CMD. PHOC 40 and PHOC 41 are more compact, composed of 36 and 63 stars respectively with CSTs of 7.7σ and 9.4σ and at estimated distances of just 331 pc and 290 pc. Both objects have good quality CMDs. All three OC candidates would be excellent candidates for further study (especially with spectroscopy) thanks to their proximity.

The other 38 candidates have a mean size of 49 stars, with the largest having 118 stars and the smallest 29. Additionally, the objects had a mean CST of 7.9σ – many of the new OC candidates are well above our 5σ CST threshold and represent clear astrometric overdensities.

2.7 Conclusions and future prospects

In this work, we created a preprocessing pipeline for future searches of the *Gaia* dataset for OCs. We selected three viable clustering algorithms from the literature and developed new methodologies to apply them effectively to the large-scale *Gaia* dataset. We compare the three algorithms side-by-side on *Gaia* data for the first time. We find that GMMs are an inefficient algorithm inappropriate for large-scale blind searches of the *Gaia* dataset, although they are relatively effective at producing accurate membership lists of known OCs. DBSCAN is found to be feasible and successful for finding OCs, but still struggles to detect certain objects since it operates with a single global density parameter that is rarely optimal across the variable densities of *Gaia* data. In particular, when DBSCAN is used with the method of **castro-ginard_new_2018** for ϵ determination, we find that it has very good precision and specificity, producing only very small numbers of false positives – although the ACG method is only sensitive to ≈50% of the 40 true positive OCs in our main sample of 100. HDBSCAN is found to solve many of the issues encountered by DBSCAN and was the most sensitive algorithm of the three, although it also produces many false positives that need to be mitigated with additional post-processing. We will use HDBSCAN in future work to conduct a large-scale blind search for OCs. We expect that HDBSCAN's improved sensitivity over other methods trialed to date will reveal many more new OCs.

In addition, we detect a number of literature OCs that have previously gone undetected in *Gaia* data. We expect that many more literature objects from the MWSC catalogue remain to be detected in *Gaia* in future works and data releases, although the majority appear to be associations or simply undetectable in *Gaia*. We found that a handful of OCs from **cantat-gaudin_clusters_2020** may be associations – either due to being undetectable by any of the approaches we tried or due to having very poor CMDs. We hope that future work expanding our analysis to the entire *Gaia* dataset will contribute further to improving the quality and completeness of the OC catalogue of the Milky Way.

Finally, we searched our existing results for new objects and produced a list of 41 good quality new OC candidates, the nearest of which is at an estimated distance of just 290 pc. While many authors have performed searches for new OCs in *Gaia* data, our comparison of algorithms suggests that existing surveys have gaps in their sensitivity and that many new objects are yet to be detected. Our tentative new detections demonstrate this, suggesting that the OC census is still incomplete within 2 kpc to an unknown extent. Future searches with new and improved methodologies will be essential to increase the completeness of the local OC census.

We plan to develop improved processes and statistical quantifiers of the strength of all OC candidate detections, including developing supervised machine learning techniques to classify OC candidate CMDs, owing to their success in other works such as **castro-ginard_hunting_2020**. As methods for improved distance determination with parallaxes develop further (**anders_photo-astrometric_2019**), we hope to include these in our work to increase the signal to noise ratio of OCs in the *Gaia* dataset and provide cleaner membership lists.

Data from the *Gaia* satellite is overhauling our understanding of the Milky Way's structure. By continuously developing, comparing and improving our methodologies, astronomers can maximise the productivity of *Gaia* data and improve our understanding of the galaxy.

3 An all-sky cluster catalogue with *Gaia* DR3

> *These circumstances, but more especially the last-mentioned, render it extremely desirable to have presented in one work, without the necessity of turning over many volumes, a general catalogue of all the nebulae and clusters of stars actually known.*
>
> — **John Herschel**
> (1864)

Details of authorship. *The content of this chapter is almost entirely based on work published in* **hunt_improving_open_2023**. *I conducted all scientific work and wrote all of the text. Suggestions and corrections from my supervisor and the reviewer of the paper are included in the text. The formatting of figures and tables has been adjusted to better fit the formatting of this thesis.*

3.1 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_young_2010**; **krumholz_star_2019**; **krause_physics_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_milky_2022**). However, dramatically improved astrometric and photometric data from the *Gaia* satellite (**gaia_collaboration_gaia_2016**) are revolutionising our understanding of OCs and the overall Milky Way. Compared with the *Hipparcos* mission (**perryman_hipparcos_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_extended_2021** and **tarricq_structural_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_age_2019** and **cantat-gaudin_painting_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_milky_2021** who used OCs to trace the spiral arms of the galaxy. Cleaned *Gaia* cluster membership lists also improve spectroscopic studies such as **baratella_gaia-eso_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 (**brown_gaia_2018**), many works have contributed major new insights into the census of OCs. Works such as **cantat-gaudin_characterising_2018**, **cantat-gaudin_clusters_2020**, and **jaehnig_membership_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 new OCs have been reported using a range of unsupervised machine learning techniques, such as in **castro-ginard_new_2018**; **castro-ginard_hunting_2019**; **castro-ginard_hunting_2020**; **castro-ginard_hunting_2022**, **cantat-gaudin_gaia_2019**, or **liu_catalog_2019**. The reliability of the census has also been improved, with works such as **cantat-gaudin_clusters_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_improving_2021** (hereafter Paper 1), we compare three different approaches for recovering OCs in *Gaia* DR2 data, and find that the HDBSCAN clustering algorithm (**hutchison_hdbscan_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_gaia_2021**), methods

developed in Paper 1, and additional validation criteria based on the photometry of every detected cluster.

In Sect. **??**, we describe the *Gaia* DR3 data used in this work and the quality cuts we adopted to filter out unreliable sources. In Sect. **??**, 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. **??**, which we generalise to additionally infer ages, extinctions, and photometric distances to our clusters in Sect. **??**. In Sect. **??**, we crossmatch our catalogue against literature works. Section **??** presents an overview of our catalogue. We discuss the non-detections of some literature clusters in Sect. **??**, and discuss required steps a future work will take to improve the reliability of our new cluster candidates in Sect. **??**. Section **??** 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.*).

3.2 Data

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

3.2.1 *Gaia* DR3

The latest release of *Gaia* (**gaia_collaboration_gaia_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_gaia_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_gaia_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

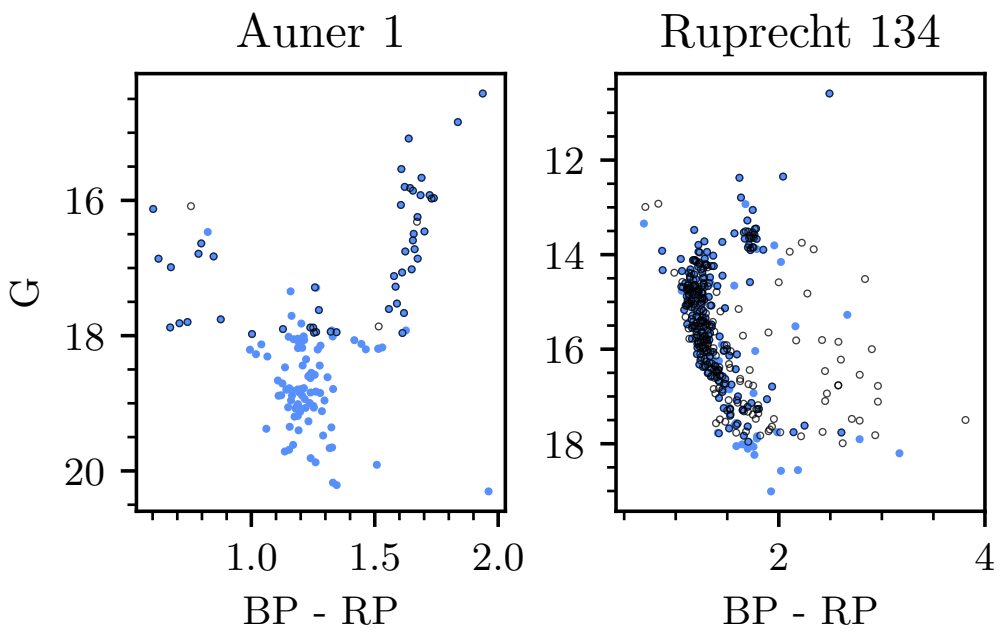

Fig. 3.1: Comparison of cluster membership lists detected using *Gaia* DR3 data cut at $G < 18$ (black empty circles) and a **rybizki_classifier_2022** v1 criterion greater than 0.5 (blue filled circles) using separate runs of HDBSCAN and our pipeline for each cut, shown for Auner 1 (left) and Ruprecht 134 (right).

distant OCs have a signal-to-noise ratio (S/N) increased by a factor of ~ 4 in *Gaia* 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_gaia_2021** provide a recipe for greatly reducing remaining parallax systematics for most sources in *Gaia* DR3 down to a few μas in the best cases, which should significantly improve the accuracy of distances to the most distant clusters. **cantat-gaudin_characterizing_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.

3.2.2 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_astrometric_2022**; **lindegren_gaia_2021-1**), as well as causing issues with photometric measurements due to blending (**riello_gaia_2021**; **golovin_fifth_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_characterising_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_classifier_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 (**hutchison_hdbscan_2013**; **mcinnes_hdbscan_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 $l, b = (0.28°, -1.63°)$ and at a distance of ~ 3 kpc, in addition to a number of clusters reported in **cantat-gaudin_clusters_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_classifier_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 (**cantat-gaudin_characterising_2018**; **castro-ginard_hunting_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. **??**. 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 10.8σ. However, the Rybizki cut cluster includes many additional faint sources, for a total of 139 member stars and an improved S/N of 17.9σ. 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 (24.7σ to 16.6σ), 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 (**ferreira_new_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_clusters_2020** for this cluster.

Only the v1 version of the **rybizki_classifier_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_classifier_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_classifier_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_hunting_2022** or **cantat-gaudin_clusters_2020**, and should result in a greater total number of both detected clusters and member stars.

3.2.3 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 $20pc$ 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_global_2013** and **cantat-gaudin_clusters_2020**.

As in Paper 1, the HEALPix (Hierarchical Equal Area isoLatitude Pixelation) tessellation scheme was used to segment the entire *Gaia* dataset (**gorski_healpix:_2005**), with calculations performed by the Python package `Healpy` (**zonca_healpy_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 ($l, b = (0°, 0°)$), instead of the default ICRS origin at right ascension and declination values of $\alpha, \delta = (0°, 0°)$ 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 $\approx 5.5° \times 5.5°$ field by $1.8°$ 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 $\approx 44° \times 44°$. 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_estimating_2021**.

Between these three systems, all bound members of all open clusters of size $20pc$ 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.

3.3 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.

3.3.1 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 (**hutchison_hdbscan_2013**; **mcinnes_hdbscan_2017**) is the most sensitive for recovering OCs in *Gaia* data.

Briefly, HDBSCAN is an updated version of the DBSCAN algorithm (**ester_density-based_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 ϵ. DBSCAN has seen much use in the literature so far for OC recovery, such as in **castro-ginard_new_2018**; **castro-ginard_hunting_2019**; **castro-ginard_hunting_2020**; **castro-ginard_hunting_2022** or **he_catalogue_2021**; **he_new_2022**. The main challenge of DBSCAN is that ϵ 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 ϵ 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_untangling_2019** and **kounkel_untangling_2020**, as well as being used to find 41 new OCs in Paper 1, and being used by **tarricq_structural_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 5σ 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 (**kounkel_untangling_2020**; **he_new_2022**), 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_estimating_2021**). Instead, we use the approach adopted in Paper 1, similar to that of works such as **castro-ginard_new_2018** and **liu_catalog_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 impact on the detectability and membership lists of nearby clusters, while having great benefits for clusters more distant than ~ 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_estimating_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_estimating_2021** are small and not prior-dominated, and so reliance on *Gaia*-derived distances for the single nearby field should not cause any issues.

3.3.2 Clustering analysis and catalogue merging

Using HDBSCAN and the same range of parameter choices as in Paper 1 ($m_{clSize} \epsilon \{10, 20, 40, 80\}$, $m_{Pts} = 10$), 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 $\mathcal{O}(n^3)$ 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 **??** 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 strategy 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

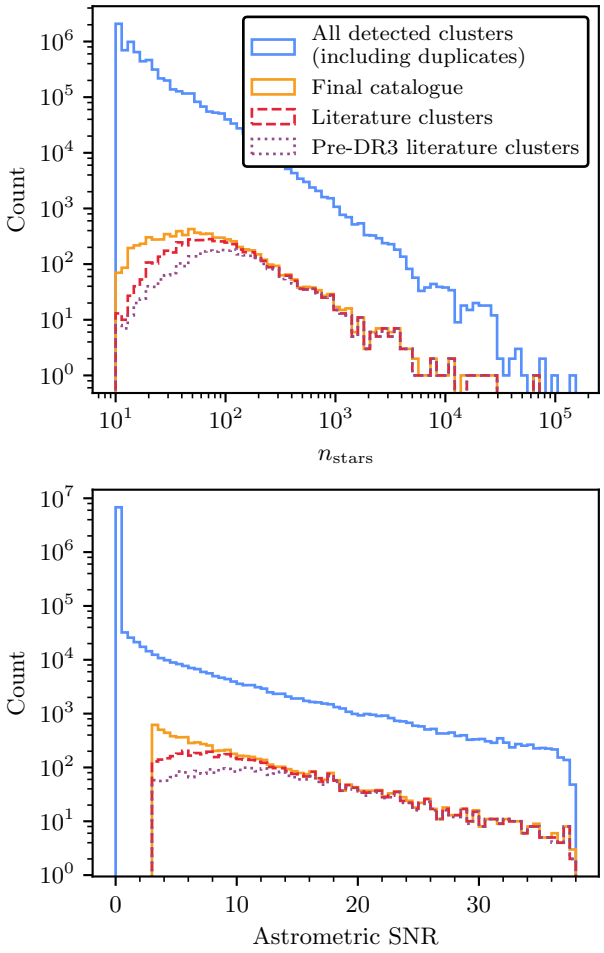

Fig. 3.2: Statistics of all detected clusters compared against the final catalogue. *Top*: distribution of the number of member stars of detected clusters, n_{stars}, for all detected clusters in all fields before catalogue merging and duplicate removal (solid blue line), for the final catalogue (solid 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.

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 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 3σ, clusters with an S/N lower than even 5σ 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. **??**, 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. **??**.

After dropping unreliable low S/N clusters, the results of each parameter run in every field were merged. For clusters where every m_{clSize} detected an identical object, duplicates were simply dropped. In some cases (such as for the largest OCs and GCs), smaller m_{clSize} runs may split the cluster into two subclusters. 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_characterising_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_clusters_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 ~ 1 pc at the distance of the cluster between different fields. We adopt a more tolerant cut corresponding to ~ 5 pc for a cluster at a worst-case distance, such that clusters within 1.91° (HEALPix level 2) or 0.41° (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 (**kovaleva_collinder_2020**; **anders_ngc_2022-1**) or with analysis using improved astrometry of a future *Gaia* data release. This resulted in a list of 7788 clusters for further analysis.

3.3.3 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 ~ 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° 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_structural_2022** attempted to derive structural parameters for 467 OCs within 1.5 kpc, but their method (based on fitting **king_structure_star_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_structure_star_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_global_2013** (see Sect. **??** 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_characterising_2018**. To calculate more accurate distances, the parallax bias of member stars was corrected using the method in **lindegren_gaia_2021**, which improved the accuracy of cluster distances particularly for distant clusters. As the **lindegren_gaia_2021** 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 $\varpi_0 = -17$ μas as derived in **lindegren_gaia_2021**. These six clusters are flagged in the final catalogue as having less accurate distances. Overall, although the **cantat-gaudin_characterising_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 ± 0.21 pc distance in **mcarthur_astrometry_2011**, who use Hubble Space Telescope parallaxes to a subset of Hyades member stars to derive its distance.

In addition, **king_structure_star_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_structure_star_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_structural_2022**,

Tab. 3.1: 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_{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. $^{(a)}$ Distances d in kpc.

these core radii still provide a good approximation of a **king_structure_star_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.

3.4 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 and simulated data. While **castro-ginard_new_2018**; **castro-ginard_hunting_2019**; **castro-ginard_hunting_2020**; **castro-ginard_hunting_2022** successfully use neural networks to classify candidate clusters as real or false with their photometry, and while **cantat-gaudin_painting_2020** and **kounkel_untangling_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_parsec_2012** are typically an imperfect fit to the observed CMDs of OCs (**cantat-gaudin_painting_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_clusters_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_bayesian_2020**; **jospin_hands-bayesian_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 **??**. Next, we discuss the creation of training data for our CMD classifier.

3.4.1 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_pypopstar_2020**) to simulate single-population clusters from PARSEC evolution models (**marigo_new_2017**), with extinction calculated star-by-star using a **cardelli_relationship_1989** extinction law with $R_V = 3.1$. Stars were sampled from these isochrones with SPISEA

using a **kroupa_variation_2001** IMF. In addition, SPISEA was used to supplement simulated OC CMDs with unresolved binary stars based on general relations derived in **lu_stellar_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_clusters_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_completeness_2020** and **boubert_completeness_2020-1**, 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_dustmaps_2018**. In addition, the selection function of every cluster was also corrected for the cuts to *Gaia* data applied in Sect. **??**. During the preparation of this work, **cantat-gaudin_empirical_model_2023** released a new selection function for *Gaia* DR3 which suggested that the earlier work of **boubert_completeness_2020**; **boubert_completeness_2020-1** 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_gaia_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 ΔA_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_clusters_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 **??**.

3.4.2 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

Tab. 3.2: Human classifier performance.

		Percent classified as			
Dataset	Size	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.

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_hunting_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 worse-quality CMDs making them compatible with both a real OC or a chance clustering of field stars hence have broad or bi-modal 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.

3.4.3 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 **??**.

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. **??**), 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.

3.4.4 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_tensorflow_2015**; **abadi_tensorflow_2016**) and TensorFlow Probability (**dillon_tensorflow_2017**) for all networks used in this work. Networks were trained with the Adam optimisation algorithm (**kingma_adam_2017**). A

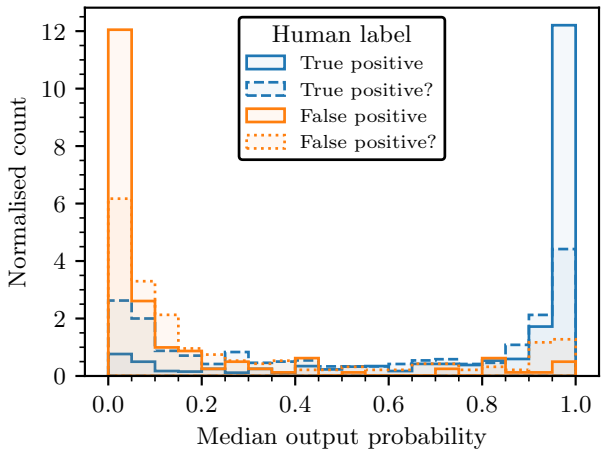

Fig. 3.3: Performance of the CMD classifier on the independent test dataset of 2000 clusters detected by HDBSCAN in *Gaia* data and labelled by hand. Clusters are labelled as true positives or false positives, with clusters where the human classifier was less certain being additionally flagged.

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 **castro-ginard_hunting_2022**; **becker_cnn_2021**; **killestein_transient-optimised_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_gaia_2021**, as these also only confused the network, despite these systematics being simulated in the training data. Finally, in terms of structure, the optimal network trialed was trained on CMDs discretised into 32×32 pixel images, corresponding to pixels of size 0.38×0.11 mag. These images were first processed by three convolutional layers with 5×5 pixel kernels of 6, 16, and

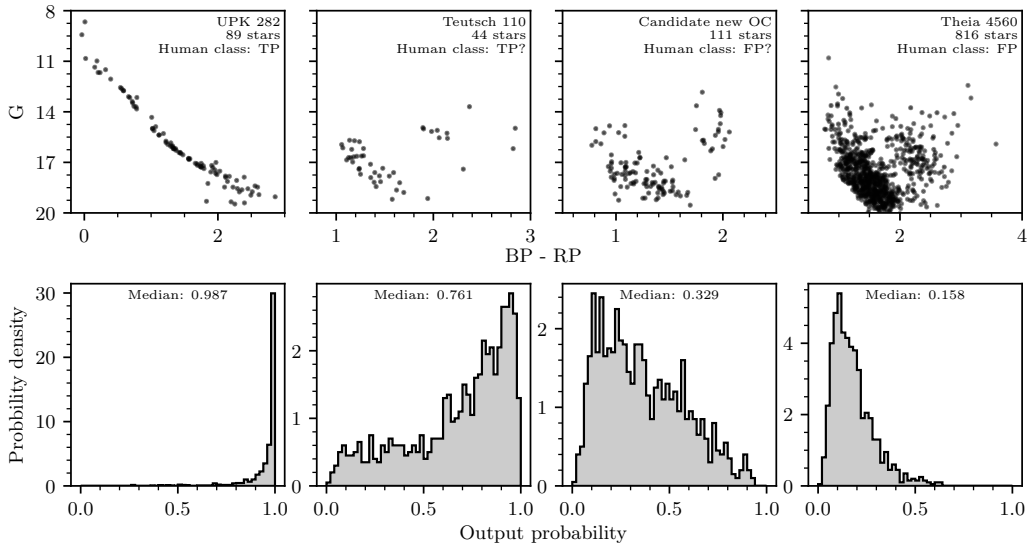

Fig. 3.4: Four examples of classified cluster CMDs from the test dataset, with cluster CMDs on the top row and their PDFs of predicted probabilities on the bottom row. Cluster names and human-assigned labels are indicated on the figures. PDFs are generated by sampling the CMD classifier 1000 times for every cluster.

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_painting_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. **??** 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. **??**. 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 2σ 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 themself are uncertain Sect. **??**. 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. **??** shows CMD classifier PDFs for four clusters from all human classes, including the names of any clusters that crossmatched 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_untangling_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 **??**, based on 1000 samples of the network for each cluster.

3.5 Age, extinction, and distance inference

3.5.1 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 ΔA_V and distance modulus $m-M$ inference. Recently, **cantat-gaudin_painting_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. **??**, parameter inference is more challenging, as a network must learn to infer multiple parameters from a CMD alone and generalise this accurately to real

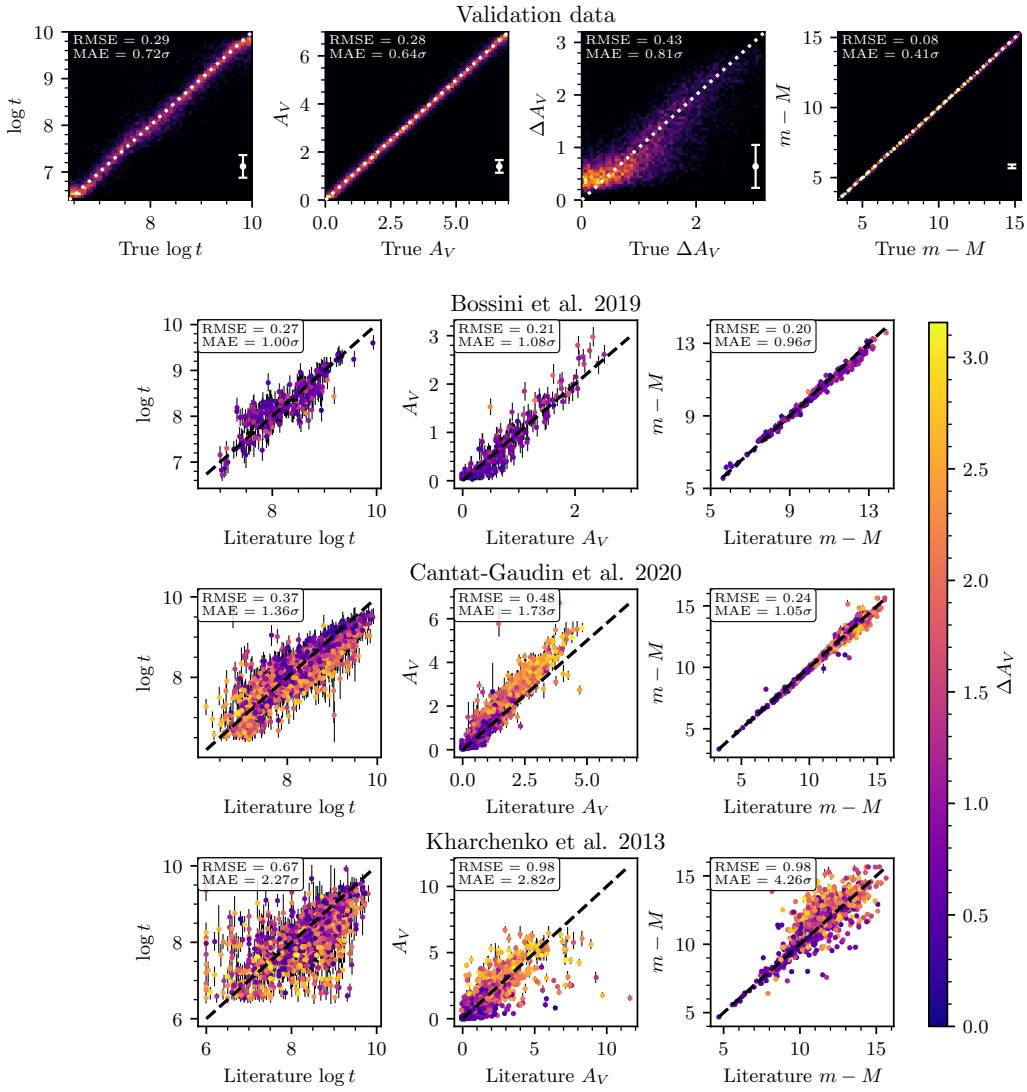

Fig. 3.5: Photometric parameters derived in this work compared against test datasets. *Top row*: 2D histograms showing the performance of the trained photometric parameter inference network on all 10 000 clusters from the validation dataset. The mean output uncertainty is shown with white error bars. 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_age_2019**, **cantat-gaudin_painting_2020**, and **kharchenko_global_2013** respectively. Points are shaded based on the differential extinction we infer for each cluster.

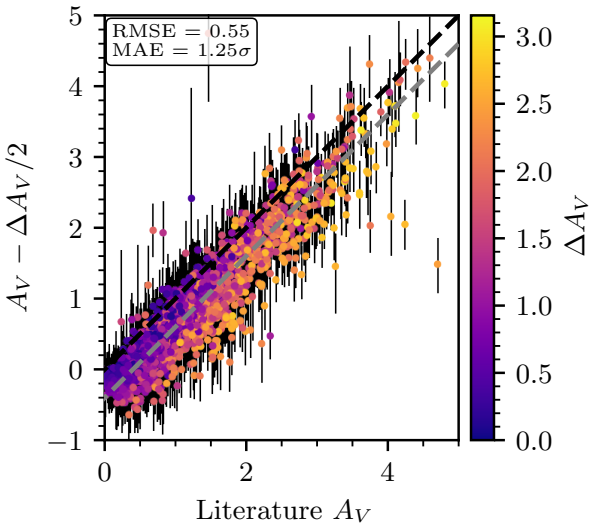

Fig. 3.6: Extinction values from **cantat-gaudin_painting_2020** compared against this work when corrected for differential extinction with an estimate of cluster differential extinction, plotted in the same style as Fig. **??**. The dashed black line shows where y values equal x ones; the dashed grey line shows the same but offset by -0.4.

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_painting_2020** do not elaborate on how they simulate clusters in their work, our methodology is be 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. **??**.

To create a parameter inference network, we used a similar network structure to that of Sect. **??**, 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. **??**, changing the distribution of cluster extinctions A_V (as defined in Table **??**) to simply be uniform between 0 and 7.

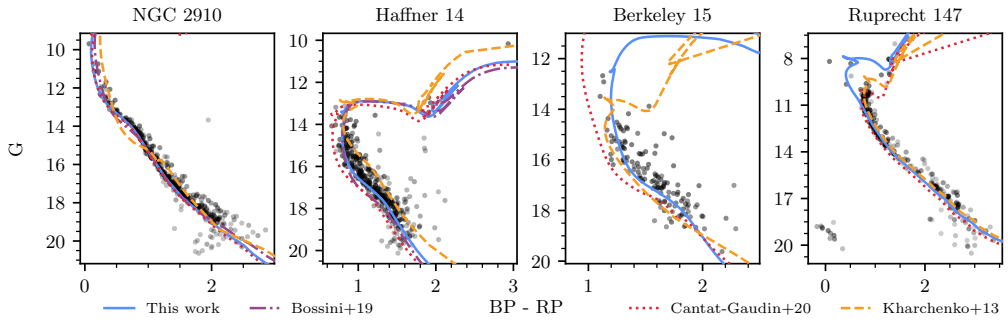

Fig. 3.7: Predicted cluster isochrones from this work (solid blue line) compared with those from other works. Cluster members are plotted in black and shaded according to their membership probability.

In 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 (**cantat-gaudin_painting_2020**; **bossini_age_2019**) use the so-called 'blue edge' of a CMD for isochrone fitting, meaning that ΔA_V is only positive. This contrasts to SPISEA's default ΔA_V model, which is Gaussian – with cluster stars having both positive and negative ΔA_V values.

However, changing SPISEA's ΔA_V model to also only be positive (and hence defining ΔA_V in terms of the blue edge of cluster CMDs) was not found to be helpful. Owing to HDBSCAN's high sensitivity, we detect a higher number of stars outside of the core of clusters than in the membership lists of **cantat-gaudin_clusters_2020**, which are constructed with the UPMASK algorithm (**krone-martins_upmask:_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_clusters_2020**, with many clusters having a blue edge at an extinction value up to 1 magnitude lower than in **cantat-gaudin_clusters_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 produce different values of A_V depending on how sensitive their membership recovery process is.

Instead, we continue using the default SPISEA ΔA_V definition centred on the mean cluster A_V, but while also using the network to infer ΔA_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, ΔA_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 ΔA_V values

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 ΔA_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_structure_star_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 ΔA_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 3σ level; and the ratio between core and tidal radius, which was set to the median value for open clusters from **kharchenko_global_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 ΔA_V and other effects such as unresolved binary stars, outliers, and photometric uncertainties, values of ΔA_V smaller than 0.4 are not typically correctly predicted, although the true value is typically still within 1σ uncertainty of the predicted value. These results are plotted on the top row of Fig. **??**.

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 **??**. These parameters are based on 1000 samples of the network for each cluster.

3.5.2 Comparison with other works

We briefly compare our photometric parameters to other works in the literature. Firstly, Fig. **??** 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_characterising_2018**. Ruprecht 147 is a nearby and particularly old cluster (~ 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. **??** 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_age_2019 provide a catalogue of precise OC parameters from Bayesian isochrone fitting using the BASE-9 algorithm (**hippel_inverting_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 **??**). 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 2σ from ours.

Our parameters agree less strongly with the results of **cantat-gaudin_painting_2020**, which are derived from a neural network trained on isochrone fits from a variety of works (**bossini_age_2019**). This is to be expected to some extent, as while **bossini_age_2019** only fit isochrones to a subset of OCs with clean membership lists and the least differential reddening, **cantat-gaudin_painting_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 **??** 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, ΔA_V is generally overestimated for our validation data by around ~ 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 ΔA_V values. Secondly, our membership lists generally cover a wider extent on the sky than those used in **cantat-gaudin_painting_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 ΔA_V values is likely to frequently over-correct for the difference in A_V definition. Finally, some of our ages for the oldest clusters ($\log t > 9$) appear systematically younger, on average by around 2σ; in some cases, this may be due to our fits being disrupted by blue straggler stars (Fig. **??**, see Ruprecht 147.) The training data we use for our photometric parameter inference are adapted from our CMD classifier in Sect. **??**, 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_ecology_2015**; **cantat-gaudin_milky_2022**).

Finally, our results have limited agreement with those of **kharchenko_global_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. **??**), 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_milky_2022**), requiring that suspected outliers be removed iteratively to improve CMD quality (**kharchenko_global_2012**). However, this process can also remove true cluster members, which can cause resulting cluster membership lists to be incorrect (**cantat-gaudin_clusters_2020**). This discrepancy with the results of **kharchenko_global_2013** is also reported by **cantat-gaudin_painting_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_global_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_global_2013**.

Although approximate, these results still agree well within the sample-limited but accurate Bayesian isochrone fits of **bossini_age_2019** and agree relatively well (albeit with some caveats) with the machine learning derived parameters of

cantat-gaudin_clusters_2020. This work offers a large and homogeneously derived catalogue of photometric parameters with sufficient 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.

3.6 Crossmatch to existing catalogues

3.6.1 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. **??** 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 μas for the largest clusters in catalogues such as **cantat-gaudin_clusters_2020**, although this is smaller than estimated upper limits on systematics in *Gaia* DR2 of 50 μas (**lindegren_gaia_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 μas 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_hunting_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_systematic_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_global_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_global_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 (**cantat-gaudin_clusters_2020**); for these catalogues, we use twice this radius to approximate the total size of the cluster. Other works (**castro-ginard_hunting_2020**; **he_new_2022**) 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_untangling_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_systematic_2007**, which appears in its complete form within **bica_multi-band_2018**), we crossmatch against 26 different works listed in Table **??**. 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_untangling_2020** catalogue of predominantly moving groups, the **vasiliev_gaia_2021** *Gaia* DR3 catalogue of globular clusters and the **bica_general_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_global_2013**, but with numbers that are separated by a space instead of a dash in **cantat-gaudin_clusters_2020** and **dias_new_2002**, or often miss leading zeroes in **bica_multi-band_2018**.

3.6.2 Recovery of clusters from prior works

Table **??** shows that this work has a high recovery rate of OCs from other works. As shown in Table **??**, we recover 96.6% of clusters from **cantat-gaudin_clusters_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 9.7σ and 12.5σ respectively. The fact that only **cantat-gaudin_clusters_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_painting_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_hunting_2020** (a work based on *Gaia* DR2) and **castro-ginard_hunting_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

Tab. 3.3: Results of crossmatching against literature catalogues sorted by n_{clusters}.

Work	n_{clusters}	n_{detected}	%
bica_multi-band_2018	4391	1251	28.5
kharchenko_global_2013	2935	1513	51.6
dias_new_2002	2161	1160	53.7
he_unveiling_hidden_2022	1656	737	44.5
cantat-gaudin_clusters_2020	1481	1431	96.6
hao_newly_2022	704	501	71.2
castro-ginard_hunting_2022	628	558	88.9
castro-ginard_hunting_2020	582	519	89.2
he_new_2022	541	440	81.3
he_blind_allsky_2022	270	122	45.2
sim_207_2019	208	180	86.5
qin_hunting_2023	101	74	73.3
chi_lisc_2023	82	18	22.0
liu_catalog_2019[a]	76	57	75.0
he_catalogue_2021[b]	74	69	93.2
li_lisc_2022	64	44	72.1
chi_identify_2022[b]	46	11	23.9
hunt_improving_2021	41	41	100.0
li_lisc_2023	35	0	0.0
ferreira_new_2021	34	32	94.1
ferreira_discovery_2020	25	25	100.0
casado_new_2021	20	15	75.0
hao_sixteen_2020[b]	16	5	31.3
jaehnig_membership_2021	11	7	63.6
santos-silva_canis_2021	5	4	80.0
qin_discovery_2021[b]	4	4	100.0
ferreira_three_2019	3	0	0.0
casado_discovery_2023	2	2	100.0
anders_ngc_2022-1	1	1	100.0
bastian_gaia_2019	1	1	100.0
tian_discovery_2020	1	1	100.0
zari_3d_2018[b]	1	1	100.0
kounkel_untangling_2020[c]	8281	1498	18.1%
bica_general_2008[d]	3740	22	0.6%
vasiliev_gaia_2021[e]	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. [(a)] Original work and this work uses the acronym 'FoF' to name clusters, although others list with acronym 'LP'. [(b)] Cluster(s) in these works were unnamed, and so cluster acronyms were adopted based on first letters of surnames of authors. [(c)] Catalogue of predominantly moving groups, although many are also open clusters. [(d)] Position-only catalogue of objects in the Magellanic clouds. [(e)] Catalogue of globular clusters.

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_sixteen_2020**; **hao_newly_2022** and **he_catalogue_2021**; **he_blind_allsky_2022** **he_new_2022**; **he_unveiling_hidden_2022**, recovering fewer than 50% of the clusters reported in **he_blind_allsky_2022**; **he_unveiling_hidden_2022** using *Gaia* EDR3 data.

Additionally, while a large fraction of clusters reported before *Gaia* and catalogued in works such as **dias_new_2002**, **kharchenko_global_2013**, and **bica_multi-band_2018** still do not appear in *Gaia* DR3, we are able to reliably detect an additional 277 clusters from **dias_new_2002**, 292 clusters from **kharchenko_global_2013**, and 127 clusters from **bica_multi-band_2018** that do not appear in the *Gaia* DR2 catalogue of **cantat-gaudin_clusters_2020** (excluding GCs in all cases, as the catalogue of **cantat-gaudin_clusters_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_lisc_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_characterising_2018** and **cantat-gaudin_clusters_2020**, who also find that high latitude OCs that have been reported in works such as **schmeja_global_2014** are undetectable in *Gaia* data.

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

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_untangling_2020**, despite this work using the same algorithm (HDBSCAN). Many of the groups reported in **kounkel_untangling_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 (**zucker_disconnecting_dots_2022**).

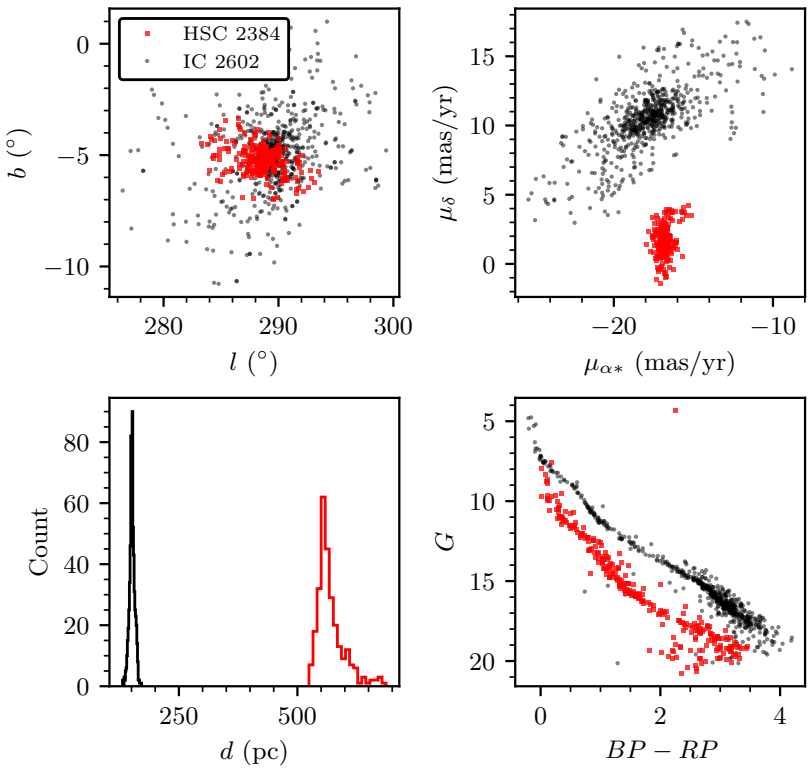

Fig. 3.8: Member stars for the candidate new cluster HSC 2384 (red squares) compared against the nearby cluster IC 2602 (black circles). Four plots of are shown, comparing positions (top left), proper motions (top right) and photometry (bottom right). The bottom left plot shows a histogram of all distances to individual member stars.

The recovery of a large fraction of GCs in **vasiliev_gaia_2021** shows that HDBSCAN can be used to effectively recover GCs. The non-recovered objects are mostly distant and 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_general_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.

3.6.3 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,

Tab. 3.4: Mean parameters for the clusters detected in this study.

Name	ID[a]	S/N	$n_{\rm stars}$	α (°)	δ (°)	r_{50} (°)	$\mu_{\alpha*}$ (mas yr^{-1})	μ_δ (mas yr^{-1})	ϖ (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)	$7.87^{+0.24}_{-0.27}$
HSC 2	1806	3.79	16	268.63	-29.53	0.13	1.680 (0.031)	-1.182 (0.032)	0.634 (0.003)	$7.92^{+0.24}_{-0.22}$
HSC 3	1807	3.89	24	273.73	-31.87	0.12	0.371 (0.019)	0.210 (0.025)	0.647 (0.005)	$8.75^{+0.18}_{-0.20}$
HSC 4	1808	3.32	17	269.07	-29.64	0.02	2.125 (0.067)	-11.895 (0.060)	0.112 (0.015)	$7.54^{+0.45}_{-0.50}$
HSC 5	1809	4.38	18	276.78	-33.09	0.12	0.150 (0.047)	-6.676 (0.049)	0.657 (0.004)	$9.70^{+0.30}_{-0.17}$
HSC 6	1810	4.57	21	267.71	-28.82	0.05	-0.292 (0.017)	-1.516 (0.023)	0.252 (0.004)	$7.84^{+0.29}_{-0.27}$
HSC 7	1811	3.12	18	261.40	-25.13	0.09	-5.033 (0.061)	-0.983 (0.060)	0.464 (0.005)	$9.68^{+0.32}_{-0.15}$
HSC 8	1812	3.33	28	267.67	-28.63	0.06	0.207 (0.014)	-0.211 (0.026)	0.340 (0.004)	$7.86^{+0.22}_{-0.23}$
HSC 9	1813	5.88	25	269.05	-29.33	0.16	2.120 (0.020)	-0.289 (0.021)	0.549 (0.005)	$7.61^{+0.22}_{-0.19}$
HSC 10	1814	4.56	12	268.23	-28.80	0.06	-0.200 (0.011)	-1.753 (0.013)	0.351 (0.003)	$8.20^{+0.30}_{-0.31}$
							...			

Notes. Standard errors for mean proper motions and parallaxes are shown in the brackets. The full version of this table with 7167 rows and many extra columns is available at the CDS only, with a complete description of the included additional data in Appendix **??**.
[(a)] Internal designation used to link final catalogue entries to their crossmatching results in Table **??**.

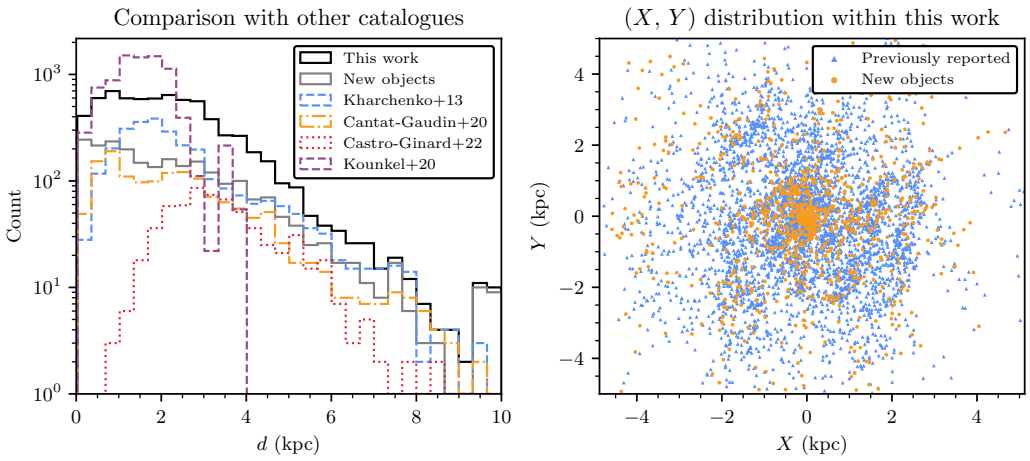

Fig. 3.9: Distance and spatial distributions of clusters in this work. *Left*: the distance distribution of all clusters in this work that do not crossmatch to known GCs compared to other catalogues. *Right*: The distribution of clusters in this work in Cartesian coordinates centred on the Sun, cut to 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.

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 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, 17 objects associated with galaxies or dwarf galaxies, and 582 objects clearly associated with stellar streams in the galactic halo, our catalogue contains 7167 clusters, and is listed in Table **??** and online at the CDS, with tables of member stars and the rejected Magellanic cloud objects, galaxies, and stellar streams available online only. 2387 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 astrometric S/N of 23.6σ, which likely avoided prior detection due to being obscured by IC 2602 and mis-crossmatched to it (shown in Fig. **??**.) 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 **??**.

In the next sections, we discuss multiple aspects of the overall catalogue. Firstly, we discuss the overall catalogue of existing clusters in Sect. **??**, including its distribution and the quality of its membership lists. Section **??** discusses why some literature clusters are undetected. Finally, Sect. **??** 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.*).

3.7 Overall results

In this section, we briefly discuss the structure and characteristics of the overall catalogue of 7167 clusters.

3.7.1 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 3σ, and objects with low-quality CMDs given the results of our classifier in Sect. **??**. 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

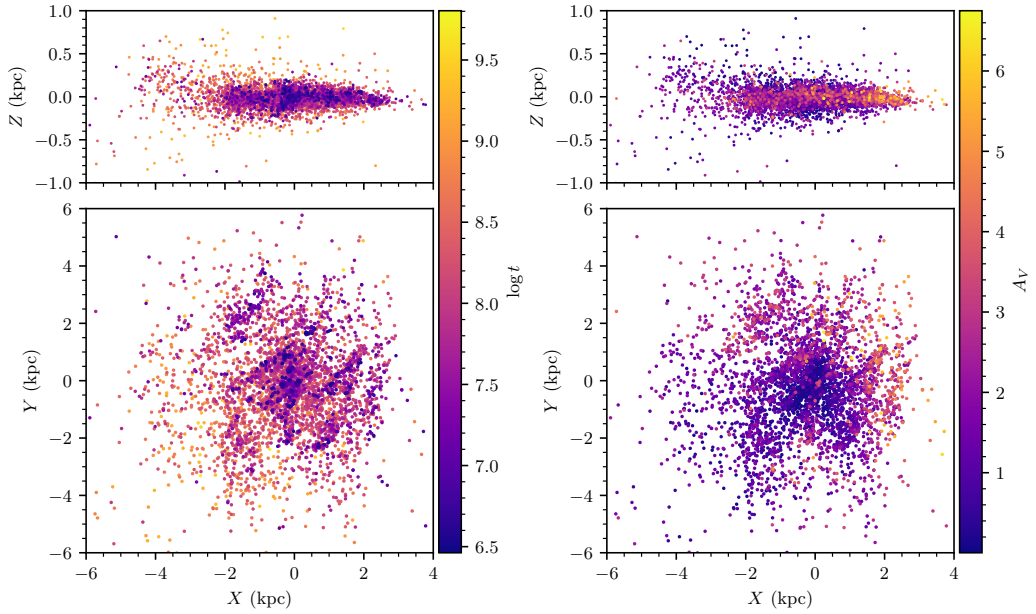

Fig. 3.10: Spatial distributions of clusters detected in this work shaded on our derived $\log t$ and A_V values. *Left:* side-on and top-down distribution of clusters in heliocentric coordinates that do not crossmatch to known GCs. The galactic centre is to the right, with the Sun at $(0, 0)$. Only clusters passing two quality cuts are plotted: firstly, those with a CST score above 5σ, 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.

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 4105 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 5σ, corresponding to clusters with a high likelihood of being real overdensities within *Gaia* data and not simply a statistical fluctuation. The more tenuous 3062 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.

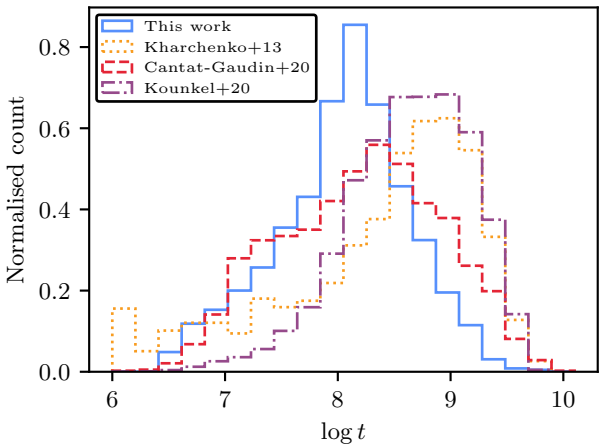

Fig. 3.11: Histogram of ages of all clusters in this work with median CMD classes greater than 0.5 – specifically, all clusters with photometry that is compatible with a single population of stars. These are compared to the ages of all clusters in the catalogues of **kharchenko_global_2013**, **kounkel_untangling_2020**, and **cantat-gaudin_painting_2020**. Known GCs are excluded from the results of this work and the results of previous works for this plot.

3.7.2 General distribution

The distribution of clusters in our catalogue is generally similar to that of other *Gaia*-based works such as **cantat-gaudin_clusters_2020**, albeit with more stark differences when compared to those compiled before *Gaia*, such as **kharchenko_global_2013**. Comparisons are also useful to the catalogue of structures, moving groups, and star clusters of **kounkel_untangling_2020** and papers based on *Gaia* DR3 data that report new clusters, such as **castro-ginard_hunting_2022**.

Figure **??** 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_clusters_2020** within 500 pc, clearly visible as an overdensity of new objects and in the distance distribution of Fig. **??**. These objects are often difficult to classify as being OCs or moving groups (see Sect. **??**).

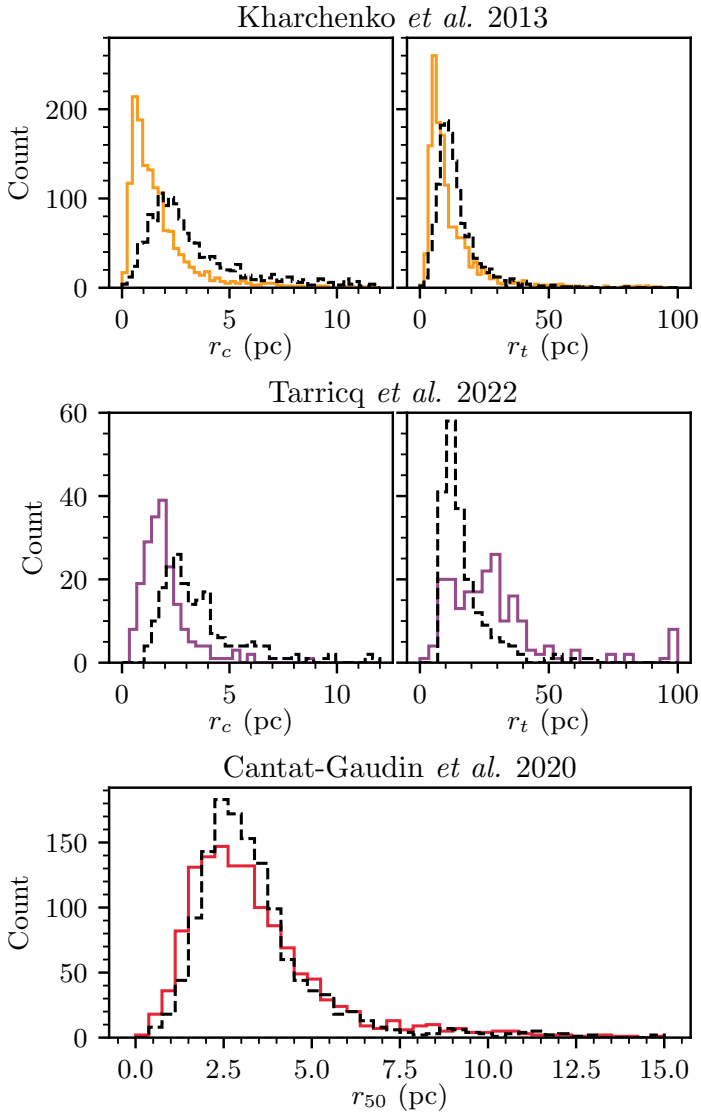

Fig. 3.12: Cluster radii derived in this work (dashed black line) compared against the distributions of cluster radii in various literature works. *Top row:* r_c (top left) and r_t (top right) of 1446 clusters from **kharchenko_global_2013** that we redetect in this work (solid orange curve) compared against our approximately estimated **king_structure_star_1962** radii for these 1446 clusters. *Middle row:* same as top, except for radii of 202 clusters from **tarricq_structural_2022** that have derived King radii (solid purple curve). *Bottom:* r_{50} measurements from **cantat-gaudin_clusters_2020** compared against our r_{50} measurements for the 1343 clusters from their work that we redetect.

The age and extinction distribution of Fig. **??** is similar to that of **cantat-gaudin_painting_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_milky_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. **??**, however. Our combined age distribution is relatively similar to that of **cantat-gaudin_painting_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_global_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_untangling_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. **??**, 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. **??** shows the distribution of cluster radii compared between this work and the works of **kharchenko_global_2013**, **tarricq_structural_2022**, and **cantat-gaudin_clusters_2020**. Our cluster radii agree most strongly with those in **cantat-gaudin_clusters_2020**, with a similar distribution of cluster radii containing 50% of members r_{50}. The **king_structure_star_1962** core radii r_c that we derive, when compared against those in **kharchenko_global_2013** and **tarricq_structural_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_global_2013**, but much smaller than those in **tarricq_structural_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_structural_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.

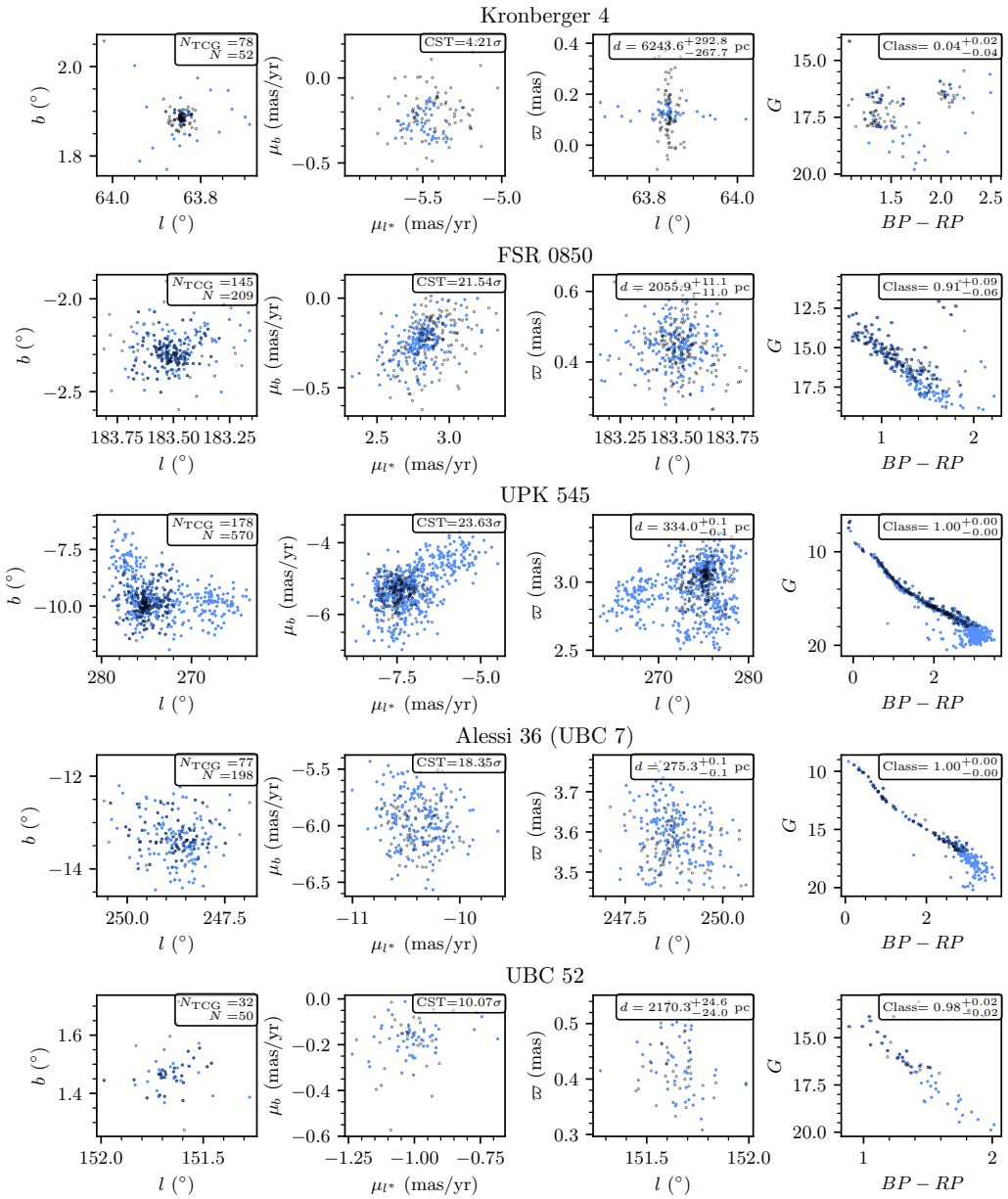

Fig. 3.13: Membership list comparisons between this work and the catalogue of **cantat-gaudin_clusters_2020**, using three clusters selected at random (upper three) and two clusters selected at random that were detected in **castro-ginard_new_2018** using *Gaia* DR1 data. Stars assigned as members by this work are plotted with filled blue circles, while members reported by **cantat-gaudin_clusters_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_clusters_2020** $N_{\rm TCG}$, 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. **??**.

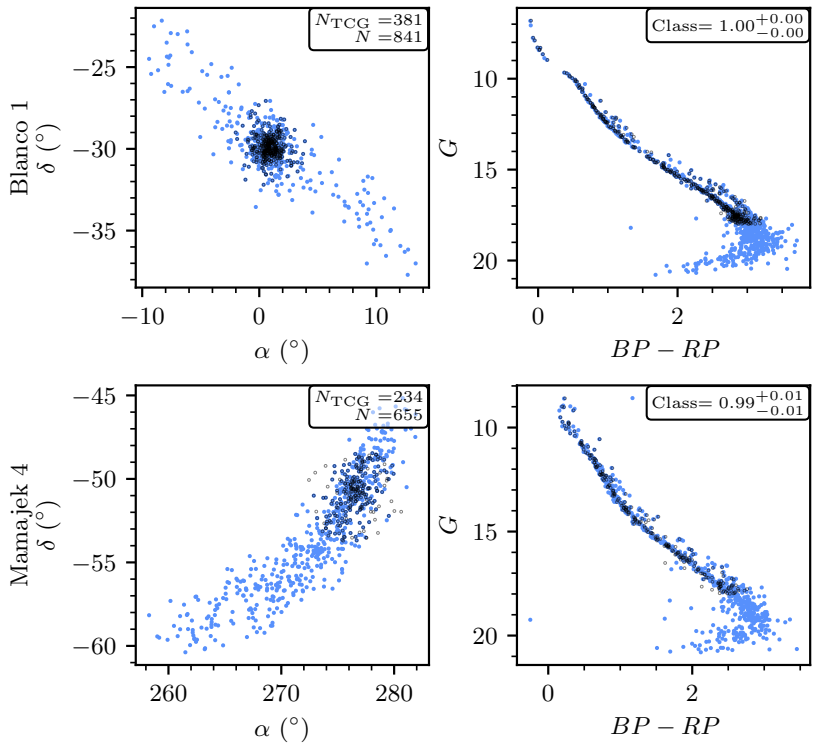

Fig. 3.14: Two examples of clusters in the catalogue that have detected tidal structures. The spatial distribution of the clusters Blanco 1 (top row) and Mamajek 4 (bottom row) are plotted on the left, with member stars reported in this work shown as filled blue circles and compared against member stars from **cantat-gaudin_clusters_2020** which are plotted as empty black circles. CMDs are shown in the two plots on the right for both clusters.

3.7.3 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. **??** compares our membership lists with those from **cantat-gaudin_clusters_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_clusters_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_clusters_2020**, and 4.1 times as many member stars as the clusters we have in common with **kharchenko_global_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_structural_2022** use HDBSCAN to detect tidal tails for 71 nearby OCs, many of which we are also able to detect. Figure **??** 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_upmask:_2014**) which do consider error information could return better membership lists for these distant clusters. This is visible for Kronberger 4 in Fig. **??**, where the membership list of **cantat-gaudin_clusters_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_young_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. **??**. 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.

3.8 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_gaia_2018**), with over 2000 reported in the last two years using *Gaia* DR3 data alone (**gaia_collaboration_gaia_2021**). While multiple works have commented on the reliability of individual clusters in the literature at-length (**cantat-gaudin_clusters_2020**; **piatti_assessing_2023**), as an unbiased search for all clusters within all of *Gaia* DR3, the results of this work offer a unique way to

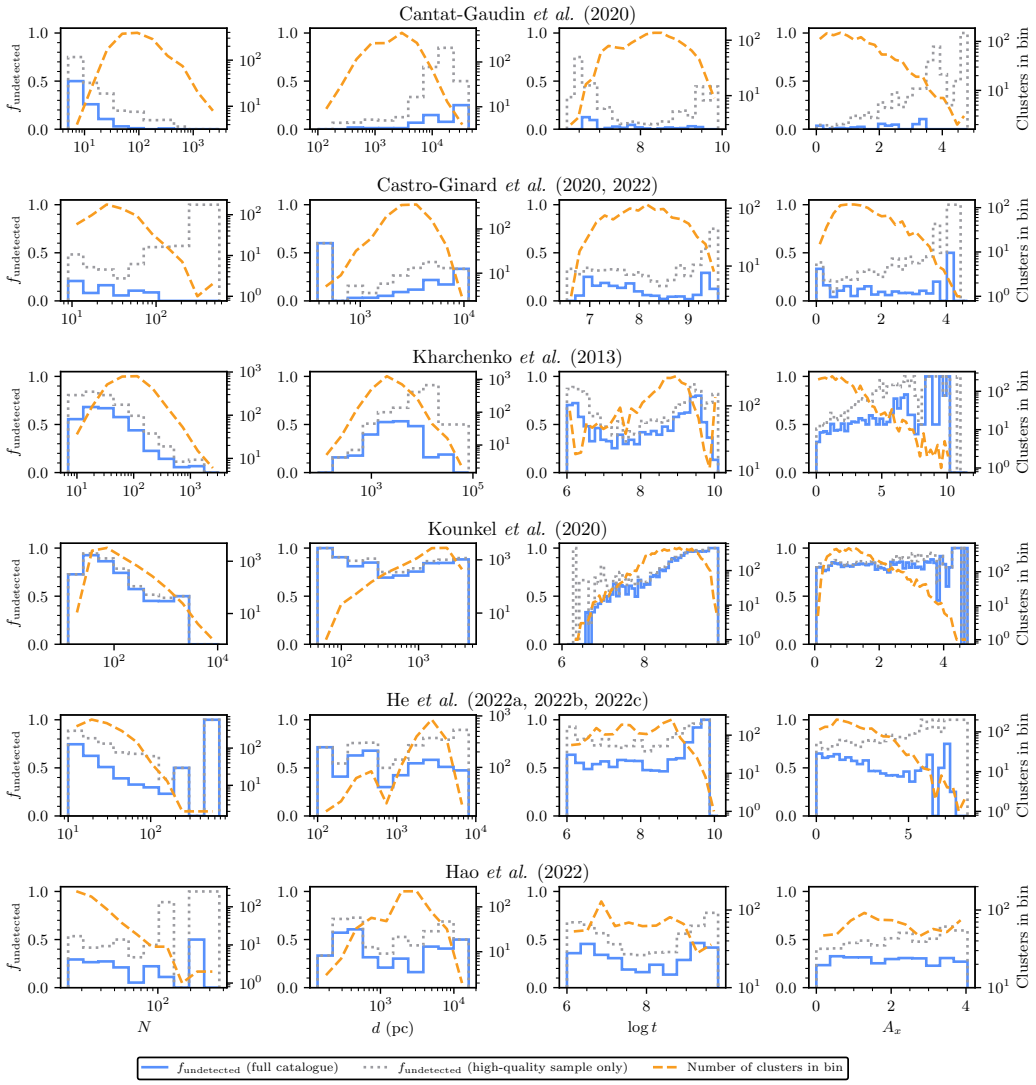

Fig. 3.15: Plots showing the fraction of clusters undetected by this work when compared to various literature works or series of literature works, shown as a histogram of various parameters as a solid blue line for all clusters in the catalogue, and a dashed grey line for clusters in the high quality sample defined in Sect. **??**. The dashed orange lines show the number of clusters in each bin. Optimum histogram bin widths were selected automatically using `numpy` (**harris_array_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.

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.

3.8.1 Methodological reasons for the non-detection of a cluster

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_hunting_2020** and **castro-ginard_hunting_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_hunting_2022** using *Gaia* EDR3 were only able to recover $\gtrsim$80% of clusters they found in DR2 in **castro-ginard_hunting_2020**, and so DBSCAN itself

between *Gaia* data releases is not able to reliably reconfirm all clusters it detected previously.

Nevertheless, Fig. **??** 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_clusters_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_clusters_2020**. A key difference between our work and **cantat-gaudin_clusters_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.

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, $n_{\text{stars, min}}$, 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 $n_{\text{stars, min}}$ range from 8 in **castro-ginard_hunting_2022** to as high as 50 in **liu_catalog_2019**, with most works coalescing around a value of between 10 and 12 (**krumholz_star_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_milky_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_milky_2022**). For the purposes of this work, we classify our clusters with our CMD classifier (see Sect. **??**) 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_clusters_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. **??**, and hence, no OCs are be 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_gaia_2021**, with GCs in the Milky Way being distinguished from OCs by their age, which is typically greater than ~ 6 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_global_2013**). In total, differences in the fundamental definition of an OC between works should have a small impact on the inclusion of OCs in this work when compared to others.

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_gaia_2021**). Poor photometry and the bound or unbound status hence do not impact our recoverability of clusters in Fig. **??**.

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 3σ. 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. **??**). 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 be 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 0.68σ, which counted as a non-detection; however, using *Gaia* DR3, we clearly detect Teutsch 156 with an S/N of 16.3σ. 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 3σ are likely to be statistical fluctuations and not real clusters, especially for S/Ns below 1σ. This can be approximately estimated using the histogram of detected cluster S/Ns in Fig. **??**. Since the distribution of literature cluster S/Ns is roughly flat for S/Ns below 10σ, assuming that this trend continues for S/Ns below 3σ, we may have missed approximately ~ 300 crossmatches to clusters reported before *Gaia* DR3 and an additional ~ 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.

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_207_2019**, which was re-detected by **cantat-gaudin_clusters_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. **??** 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.

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

While our crossmatching procedure as outlined in Sect. **??** 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_new_2022** 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_new_2022** with 24 member stars, but appears to be a small offshoot of the recently reported new cluster OCSN 82 from **qin_hunting_2023**, which has an overall position different by around 3° 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 (**he_new_2022**). In Paper 1, we commented that while DBSCAN has an excellent sensitivity and low false positive rate (depending strongly how the ϵ 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.

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_global_2013** was compiled before *Gaia* and used infrared data from 2MASS (**skrutskie_two_2006**). **cantat-gaudin_clusters_2020** are unable to recover a majority of the clusters from **kharchenko_global_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_global_2013** may still be missing from *Gaia*-based catalogues due to extinction limits.

However, Fig. **??** shows that extinction does not appear to play a major role in the non-detection of many clusters from **kharchenko_global_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 $f_{\text{undetected}}$; 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_gaia_2021**), and given that our chance of detecting a cluster reported in **kharchenko_global_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.

3.8.2 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 **??** 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. **??** 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. **??**, 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. **??**. 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 to 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_young_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. **??**, despite all of them other than **kharchenko_global_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_clusters_2020**, clusters older than an age of around 1 Gyr ($\log t > 9$) are much less likely to be redetected. **zucker_disconnecting_dots_2022** have recently investigated the nature of the groups reported in **kounkel_untangling_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_untangling_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_clusters_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. **??**.) 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_global_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_hunting_2020**; **castro-ginard_hunting_2022** and between $6 < \log t < 7$ in **hao_newly_2022** remain unclear. These works are entirely compiled

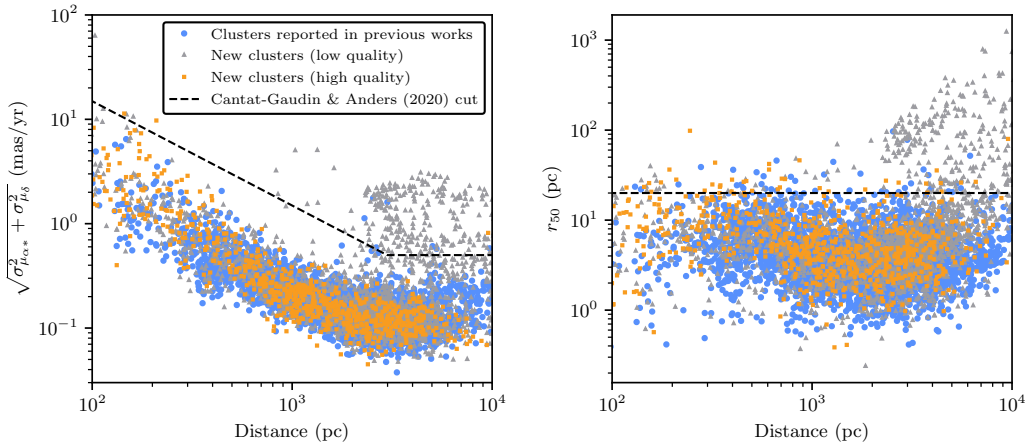

Fig. 3.16: Geometric mean of the proper motion dispersion (left) and radius containing 50% of members (right) for the clusters reported in this work, as a function of distance. Clusters are split between those detected in previous works (blue circles) and those newly reported in this work, divided between the high quality (orange squares) and low quality (grey triangles) samples defined in Sect. **??**. The cuts on cluster parameters to distinguish between bound OCs and unbound moving groups or associations proposed in **cantat-gaudin_clusters_2020** are shown as a dashed black line.

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. **??**, we discuss that A_V values of greater than ~ 5 appear to reduce the chance of a cluster being recovered in *Gaia* data. The increasing trend in $f_{\text{undetected}}$ for **cantat-gaudin_clusters_2020** as a function of A_V appears to entirely be due to our lower chance of detecting clusters with $d > 10$ kpc, since distant clusters also often have a high A_V. No other clear correlations exist for other works in Fig. **??** with respect to extinction, other than for a few dozen pre-*Gaia* clusters from the infra-red catalogue of **kharchenko_global_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

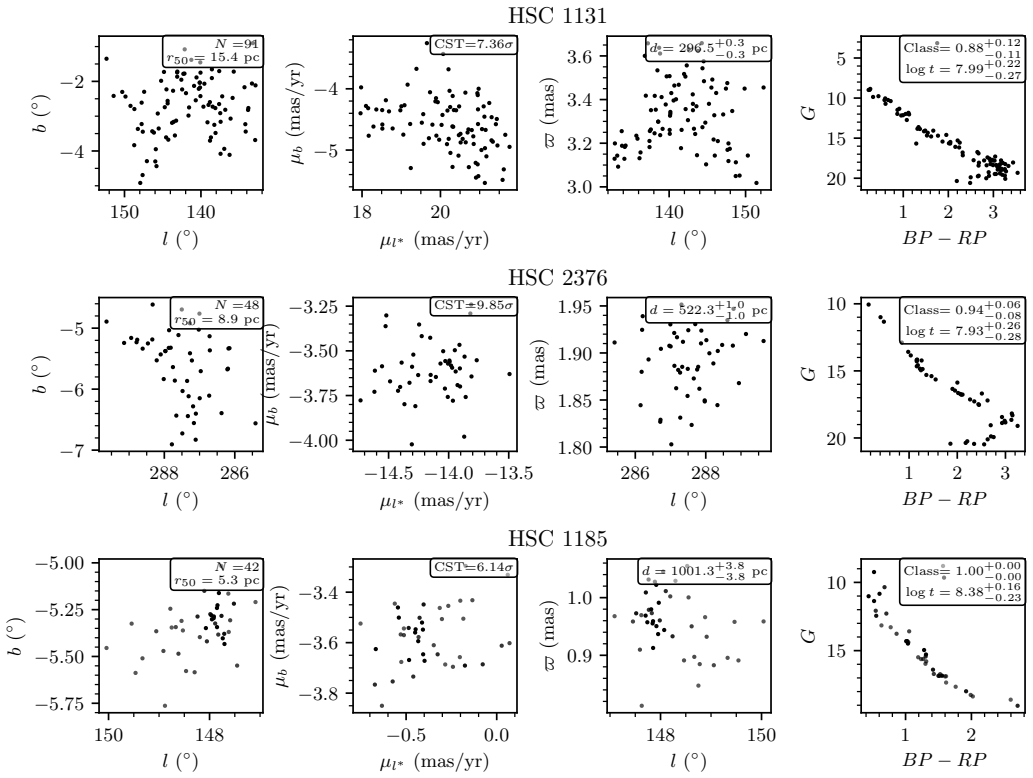

Fig. 3.17: Three newly reported clusters randomly selected from the cluster catalogue and ordered by increasing distance, with member stars plotted as a function of their astrometric and photometric data as in Fig. **??**. All clusters pass the cuts proposed in **cantat-gaudin_clusters_2020**, have good-quality CMDs passing the cuts from Sect. **??**, and have astrometric significances of greater than 5σ, meaning they are almost certainly real overdensities in *Gaia* data.

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 **??** and at the CDS.

3.9 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 2387 new objects reported in this work – 739 of which have a median CMD class above 0.5 and a CST of greater than 5σ, and are hence the most reliable new objects that we report.

3.9.1 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_structure_star_1962**) or Plummer-like (**plummer_problem_1911**) profiles. Figure **??** shows three clusters randomly selected from the 739 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_clusters_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 ~ 5 kms^{-1}) and $r_{50} < 20pc$ removed objects highly unlikely to be OCs from their sample.

However, Fig. **??** 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 ~ 3 kpc), most new clusters detected in this work are compatible with OCs given the tolerant cuts in **cantat-gaudin_clusters_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 profile, 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.

3.9.2 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, $\sigma_{\rm vir}$, based on its mass and radius. This can be compared to the cluster's measured 1D velocity dispersion σ_{1D}, which should equal $\sigma_{\rm vir}$ if the cluster is bound:

$$\sigma_{\rm vir} = \sqrt{\frac{GM}{\eta r_{\rm hm}}} \approx \sigma_{1D} \text{ for a bound cluster,} \tag{3.1}$$

where $r_{\rm hm}$ is the cluster's half-mass radius, M is the cluster's mass, G is the gravitational constant and η is a constant depending on the cluster's density profile that is usually set to 10 (**portegies_zwart_young_2010**). In the case when $\sigma_{1D} >> \sigma_{\rm vir}$, the cluster is likely to be unbound. This relationship has been used by works such as **bravi_gaia-eso_2018**, **kuhn_kinematics_2019**, and **pang_3d_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_clusters_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.*).

3.10 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 7167 clusters, of which 4105 clusters form the most reliable sub-sample of objects with median CMD classifications greater than 0.5 and S/Ns greater than 5σ.

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 ~ 1 kpc. Membership lists for all of our clusters are also available as a part of the catalogue (see Appendix **??** 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 (**cantat-gaudin_painting_2020**; **castro-ginard_milky_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 2387 new objects as yet unreported in the literature, 739 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 5σ. 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_clusters_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 ∼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_gaia_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 (**ratzenbock_significance_2022-1**) 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.

4 The masses and dynamics of star clusters in the Milky Way

> *Things are only impossible until they're not.*
>
> — **Jean-Luc Picard**
> (2364)

Details of authorship. *The results presented in this chapter will be published in Hunt and Reffert (*in prep.*). All calculations, figures, and writing in this chapter were conducted by myself.*

TODO: dynamics section

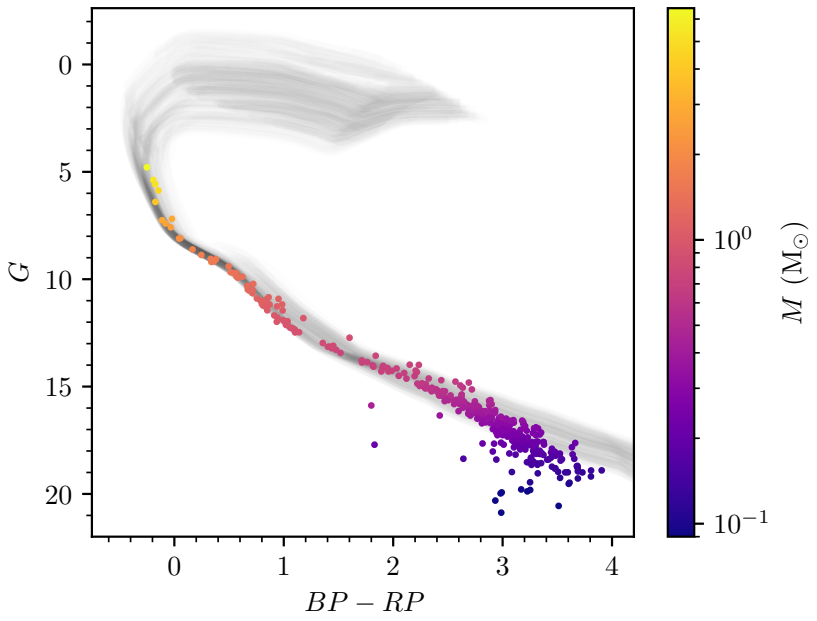

Fig. 4.1: TODO

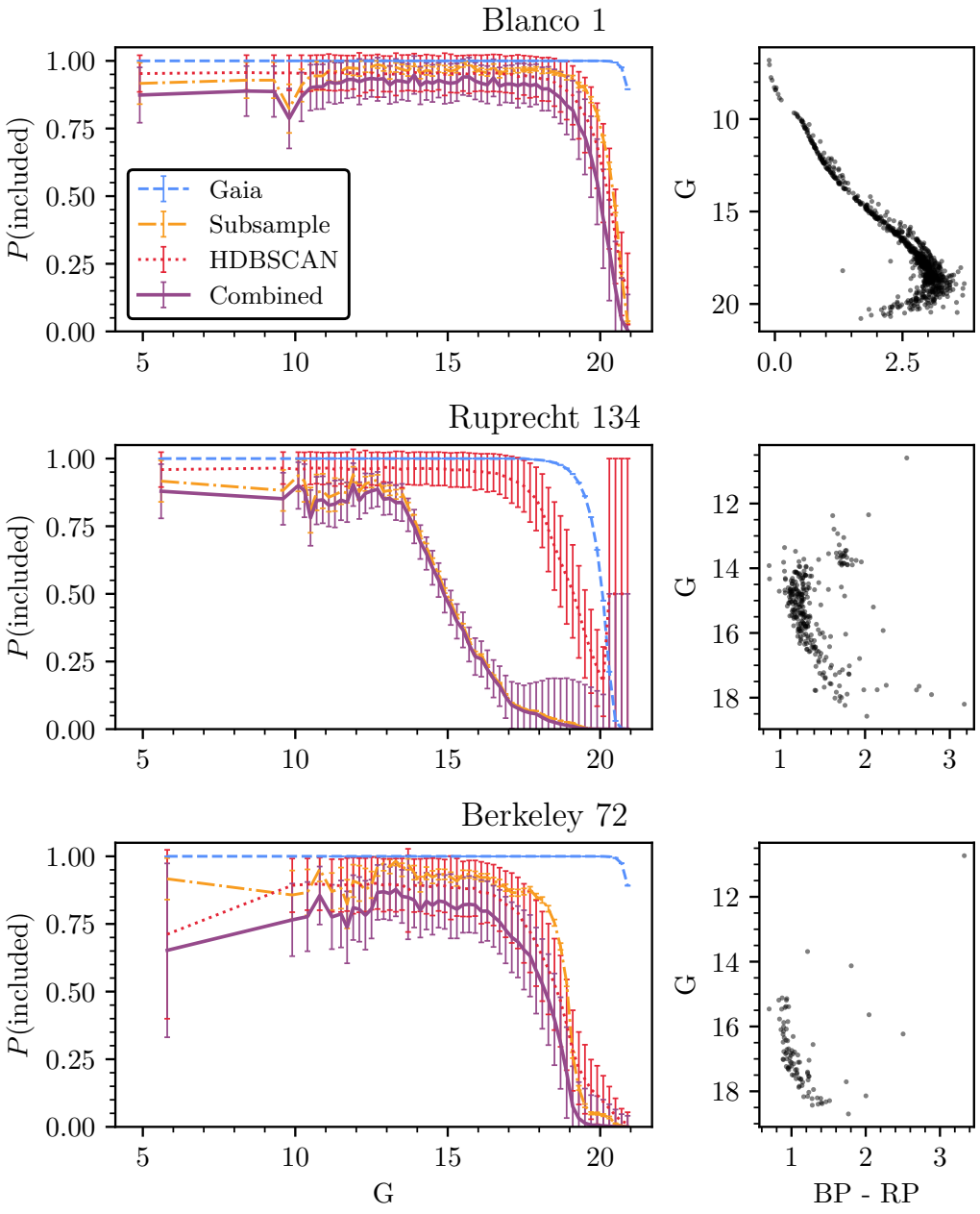
Blanco 1
Ruprecht 134
Berkeley 72
P(included)
G
BP - RP
Gaia
Subsample
HDBSCAN
Combined

Fig. 4.2: TODO

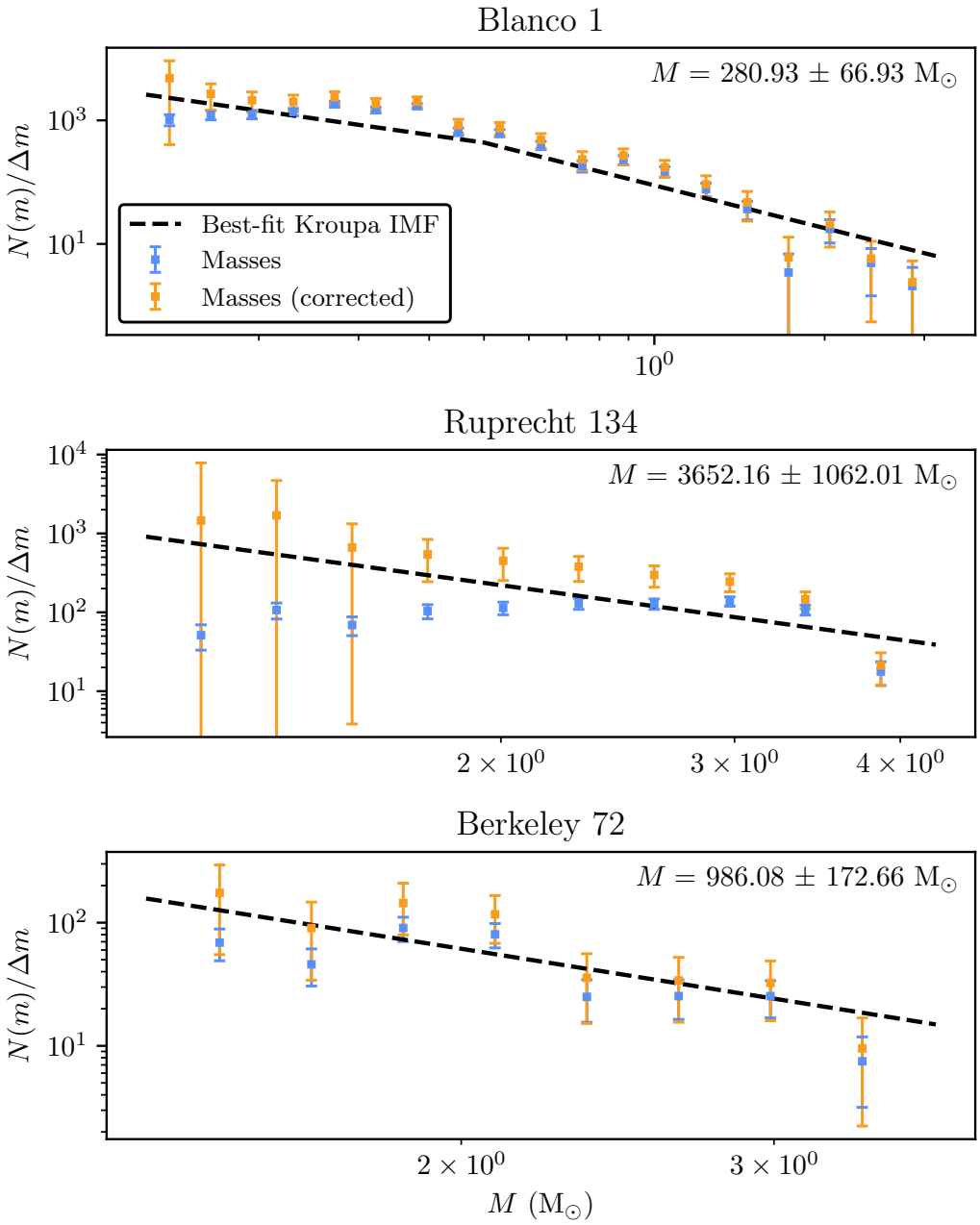
Blanco 1
$M = 280.93 \pm 66.93$ M$_\odot$
Best-fit Kroupa IMF
Masses
Masses (corrected)
$N(m)/\Delta m$
10^3
10^1
10^0
Ruprecht 134
$M = 3652.16 \pm 1062.01$ M$_\odot$
10^4
10^3
10^2
10^1
2×10^0
3×10^0
4×10^0
Berkeley 72
$M = 986.08 \pm 172.66$ M$_\odot$
10^2
10^1
2×10^0
3×10^0
M (M$_\odot$)

Fig. 4.3: TODO

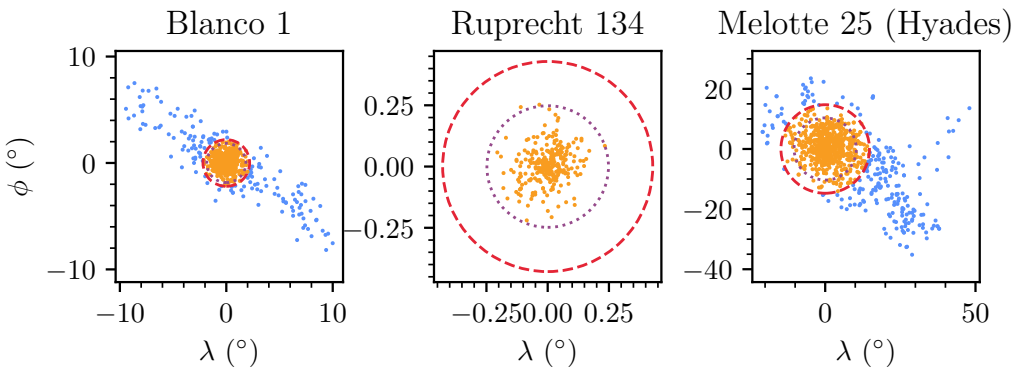

Fig. 4.4: TODO

4.1 Introduction

4.2 Mass calculations

4.2.1 Inference of stellar primary masses

4.2.2 Correction for selection effects

4.2.3 Correction for binaries

4.2.4 Mass function fits

4.2.5 Jacobi radius inference

4.3 Velocity dispersion inference

4.3.1 Gaussian velocity dispersion model

4.3.2 Coordinate frame and radial velocity corrections

4.3.3 Binary star contamination

4.4 Results

4.4.1 Masses

4.4.2 Jacobi radii

4.4.3 Virial ratios

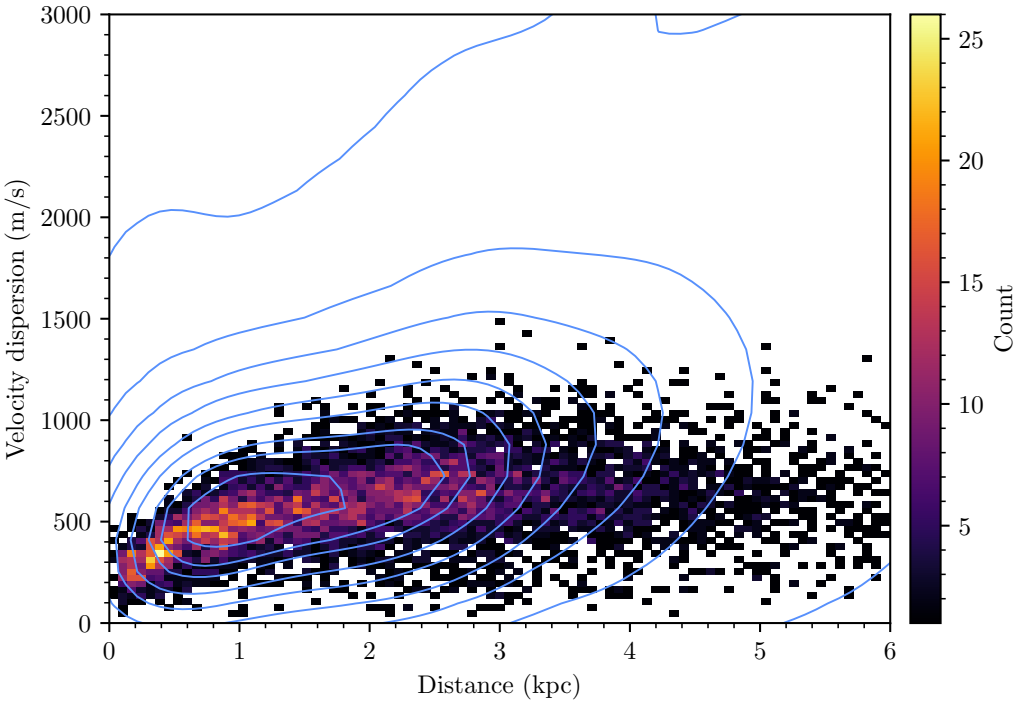

Fig. 4.5: TODO

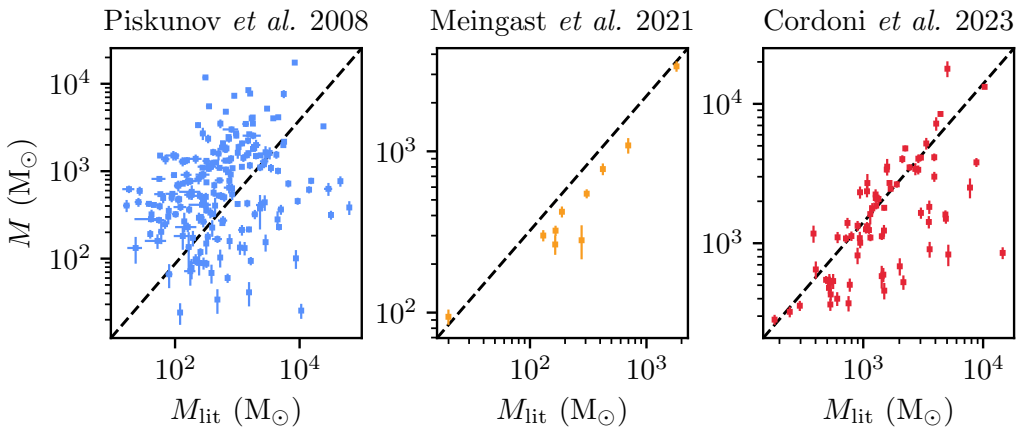

Fig. 4.6: TODO

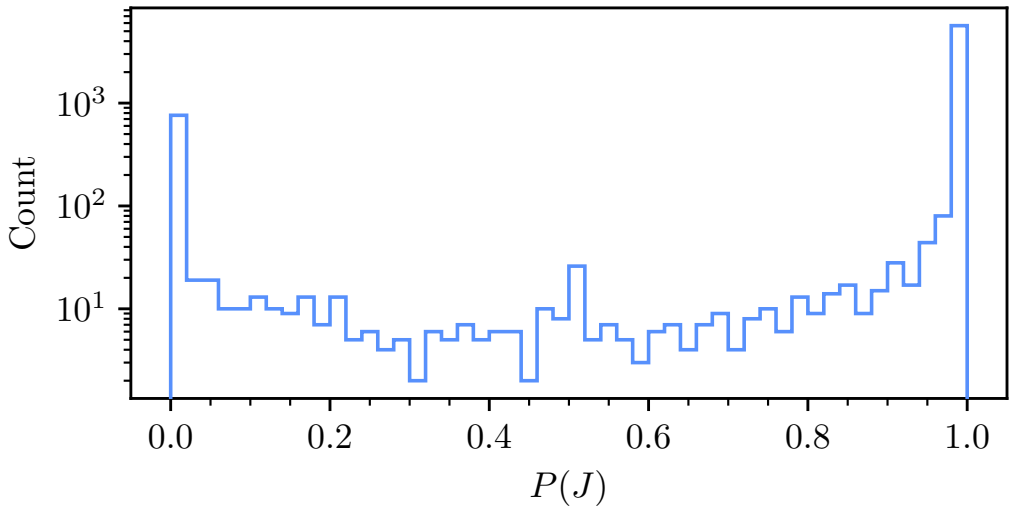

Fig. 4.7: TODO

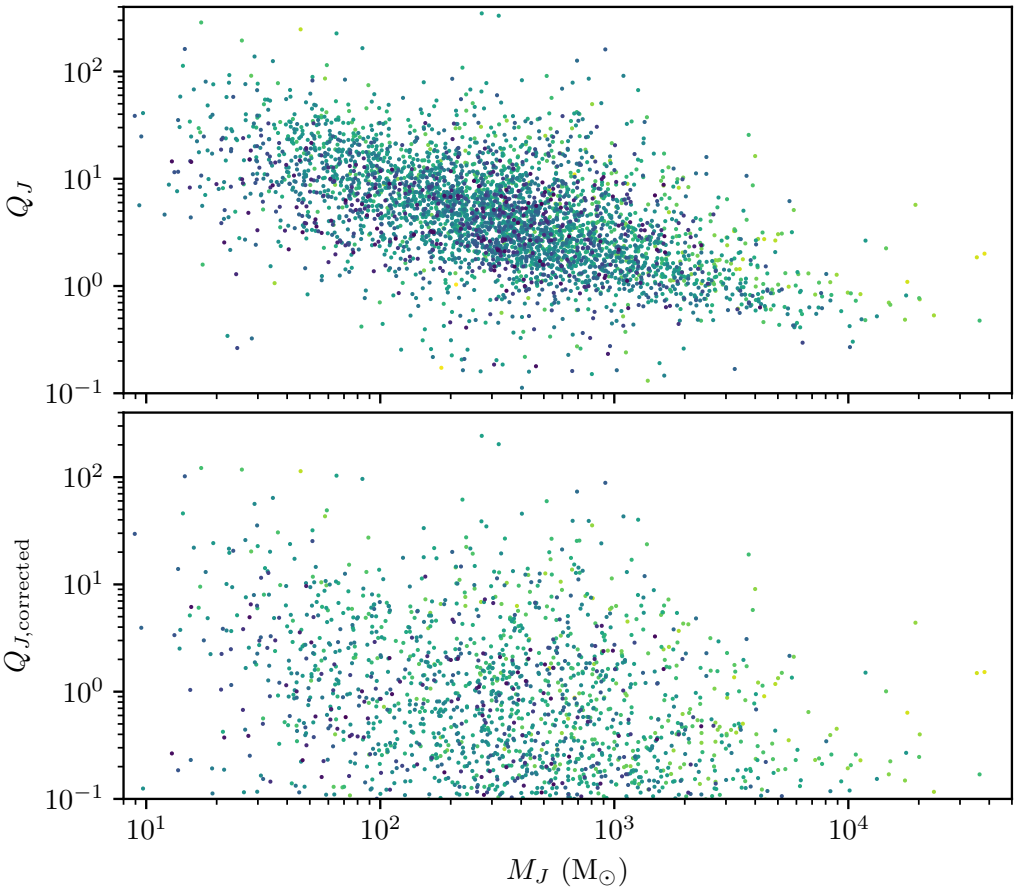

Fig. 4.8: TODO

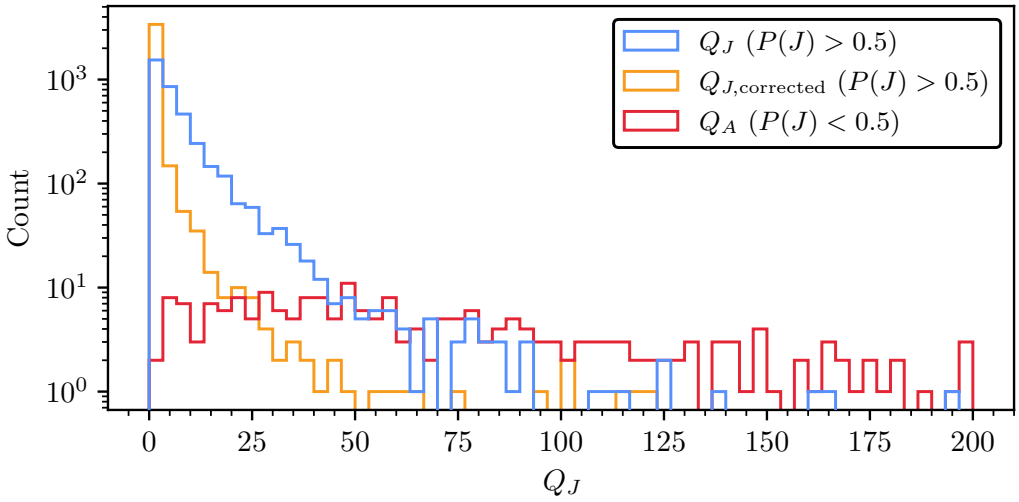

Fig. 4.9: TODO

Tab. 4.1: TODO

Type	Identifier	Count
OC	`o`	TODO
- bound OC	`o`	TODO
- unbound OC	`ou`	TODO
MG	`m`	TODO
- $P(r_J) < 0.5$	`m`	TODO
- $P(r_J) > 0.5$, $M_J < 50\ \mathrm{M}_\odot$	`mj`	TODO
GC	`g`	TODO

Conclusion

5

> *No observational problem will not be solved by more data.*
>
> — **Vera Rubin**

TODO: conclusion

List of Figures

List of Tables

Appendix A

A.1 Appendices for Chapter ??

A.1.1 ADQL query used to download data

Gaia DR2 data for this work was downloaded with the following ADQL query. `{start_number}` should be replaced with the first possible `source_id` of the desired pixel using Eqn. **??**. `{end_number}` should be replaced with the first possible `source_id` of the next integer pixel.

```
SELECT
-- Gaia astrometry
g.source_id, g.l, g.b,
g.ra, g.ra_error, g.dec, g.dec_error,
g.parallax, g.parallax_error,
g.parallax_over_error,
g.pmra, g.pmra_error, g.pmdec, g.pmdec_error,
g.astrometric_params_solved,

-- Gaia photometry
g.phot_g_mean_mag, g.phot_g_mean_flux,
g.phot_g_mean_flux_error,
g.phot_bp_mean_mag, g.phot_bp_mean_flux,
g.phot_bp_mean_flux_error,
g.phot_rp_mean_mag, g.phot_rp_mean_flux,
g.phot_rp_mean_flux_error,
g.phot_bp_rp_excess_factor,

-- Calculate HEALPix level 5 index
GAIA_HEALPIX_INDEX(5, g.source_id)
  AS gaia_healpix_5,
```

```
-- RUWE statistics
r.ruwe,

-- CBJ+2018 distances
d.r_est, d.r_lo, d.r_hi,
d.r_len, d.result_flag

-- Inner join the tables
FROM gaiadr2.gaia_source AS g
INNER JOIN
  gaiadr2.ruwe
  AS r
  ON g.source_id = r.source_id
INNER JOIN
  external.gaiadr2_geometric_distance
  AS d
  ON g.source_id = d.source_id

-- Select only valid points
WHERE g.source_id >= {start_number}
AND g.source_id < {end_number}
AND g.astrometric_params_solved=31
AND g.phot_bp_mean_mag IS NOT NULL
AND g.phot_rp_mean_mag IS NOT NULL
```

A.1.2 Comparison with other OC catalogues

We present brief comparisons with the results of other OC catalogues, in lieu of best practices proposed in **cantat-gaudin_clusters_2020** and as a part of efforts towards generally improving the quality of the OC census, reporting on both positive and negative detections. In future works, we hope to expand comparisons such as this across the entire OC census, offering another viewpoint on the existence of many literature OCs.

cantat-gaudin_clusters_2020

Of the 537 objects listed in **cantat-gaudin_clusters_2020** and in the fields in this study, we are able to detect 86.4% of them with at least one algorithm or parameter combination, many of which are clear overdensities with well-resolved parameters.

We single out Auner 1, Berkeley 91 and Patchick 75 from Sect. **??** as objects that should be detectable but are not found by any algorithm. In addition, FSR 1460 and FSR 1509 are also undetected. If real, these objects are distant and difficult to detect in *Gaia* data, although these objects also have heavily polluted CMDs in the membership lists of **cantat-gaudin_clusters_2020** and hence may simply be associations. Future *Gaia* data releases with better astrometric precision will shed more light on the status of these edge-case objects.

MWSC

We concur with the results of **cantat-gaudin_gaia_2018** and **cantat-gaudin_clusters_2020** that a majority of the objects in MWSC are undetectable in *Gaia* data. Some of these objects may simply not be visible in *Gaia* data due to reddening or large distances, although many are also likely to not be real. Future studies will have to quantify this for all OCs on a case-by-case basis. Of our 100 main OCs that were randomly selected from the MWSC catalogue, we detected OCs corresponding to 35 of them, suggesting that ≈35% of the total MWSC catalogue is visible in *Gaia* data.

However, our results show that a number of MWSC objects appear to have been missed by works such as **cantat-gaudin_clusters_2020**. In our larger crossmatching effort, we recovered candidates corresponding to 193 of the 607 objects listed in MWSC (31.8%) but that are undetected in **cantat-gaudin_clusters_2020**. Some of these objects may be new OCs that happen to have similar parameters to old objects, although some others are new detections of MWSC OCs in *Gaia* data.

The best examples of re-detected OCs were Collinder 347, FSR 0124, FSR 0270 and FSR 1406, which were were clearly crossmatched and are clearly visible by eye in *Gaia* data. In addition, Collinder 347 has also been well detected by **piatti_extended_2019** in *Gaia* DR2 data and recently by **claria_ccd_2019** in visual spectrum photometric data. The sparse OCs Sgr OB6, Sgr OB7 and ASCC 100 were also detected, the latter of which has few members but is nearby with a parallax of 2.75 mas, suggesting that some OCs are yet to be recovered in *Gaia* data even at small distances. In all seven cases, the crossmatched objects were clearly compatible in positional and distance space with MWSC values. They are also compatible in

proper motion space, although at large distances the PPMXL proper motions in MWSC provide very little constraint.

While the catalogue of **cantat-gaudin_clusters_2020** is the most complete and homogeneous OC catalogue to date, it still appears to lack some OCs from the literature and contains a handful of OCs that are somewhat putative. Ongoing comparisons with the results of multiple different clustering algorithms and methodologies will help to confirm, question or deny the existence of more OCs in the literature.

castro-ginard_hunting_2020 and liu_catalog_2019

castro-ginard_hunting_2020 and **liu_catalog_2019** have recently reported a combined total of over 600 new OCs in *Gaia* DR2 data respectively. Of the 209 objects from **castro-ginard_hunting_2020** in the fields in this study, we detected OCs compatible with 135 of them (64.6%), representing a sizable fraction of their catalogue of new OCs that has been detected independently in *Gaia* data for the first time. We note that the undetected OC UBC 638 is very close to UBC 637 (which is detected) - their reported centres are within 0.05° of one another, their proper motions 0.07 mas yr^{-1} and parallaxes to within 0.1 mas, so they may be the same object.

We are able to detect OCs compatible with 24 of the 32 OCs from the catalogue of **liu_catalog_2019** that are included in this study. The reasons for non-detections of OCs from both of these works remain unclear, and would need to be investigated in a future study.

A.1.3 Tables of detected clusters and members

Four supplementary tables are available in online-only material at the CDS[1]. For literature clusters, all detections by all algorithms are listed following the same format as Table **??**.1. Any one cluster may have up to 12 different entries from detections by different algorithm and parameter combinations. When no detections were made of a literature cluster, a single blank row is given with only columns one and 26 filled. The 41 new objects have their mean parameters listed in a separate table following the format of Table **??**.1 except with column 26 omitted. For both literature and new OCs, members are listed in tables following the format of Table **??**.2.

[1]https://vizier.u-strasbg.fr/

Tab. A.1: Description of the tables of detected OCs.

Col.	Label	Unit	Description
1	Name	–	Designation
2	Internal ID	–	Internal designation
3	Algorithm	–	Algorithm for detection
4	Parameters	–	Algorithm parameters
5-7[a]	α	deg	Right ascension
8-10[a]	δ	deg	Declination
11	l	deg	Galactic longitude
12	b	deg	Galactic latitude
13-15[a]	μ_{α^*}	mas yr^{-1}	Prop. motion in $\alpha \cdot \cos\delta$
16-18[a]	μ_δ	mas yr^{-1}	Prop. motion in δ
19-21[a]	ϖ	mas	Parallax
22	r_{50}	deg	Radius containing 50% of members
23	r_t	deg	Estimated tidal radius[b]
24	n	–	Number of members
25	σ_{CST}	–	CST score
26	Source	–	Source catalogue

Notes. [(a)] Where marked, three columns are provided: the mean value, standard deviation σ, and standard error $\sigma/\sqrt{n}$. [(b)] Estimated using the maximum distance between the centre of the cluster and an identified member star.

A.1.4 Plots of newly detected OCs

See Figs. **??**, **??**, **??**, **??**, **??**, **??**, and **??**.

A.1.5 List of fields used in this study

See Table **??**.

Tab. A.2: Description of the membership tables for detected OCs.

Col.	Label	Unit	Description
1	Name	–	Designation
2	Internal ID	–	Internal designation
3	Algorithm	–	Algorithm for detection
4	Parameters	–	Algorithm parameters
5	Source ID	–	*Gaia* DR2 source ID
6-7[a]	α	deg	Right ascension
8-9[a]	δ	deg	Declination
10	l	deg	Galactic longitude
11	b	deg	Galactic latitude
12-13[a]	μ_{α^*}	mas yr^{-1}	Prop. motion in $\alpha \cdot \cos\delta$
14-15[a]	μ_δ	mas yr^{-1}	Prop. motion in δ
16-17[a]	ϖ	mas	Parallax
18	Gmag	mag	G-band magnitude
19	BPmag	mag	BP-band magnitude
20	RPmag	mag	RP-band magnitude
21-22[a]	G flux	e^{-1} s^{-1}	G-band flux
23-24[a]	BP flux	e^{-1} s^{-1}	BP-band flux
25-26[a]	RP flux	e^{-1} s^{-1}	RP-band flux
27[b]	p	–	Membership probability

Notes. [(a)] Where marked, two columns are provided: the mean value and the standard error. [(b)] Always equal to one for DBSCAN as it does not produce membership probabilities for individual stars.

Tab. A.3: Sky locations and HEALPix indices of the central pixels included in this study.

Number	α (°)	δ (°)	Pixel[a]	Number	α (°)	δ (°)	Pixel[a]
0	313.6	-12.0	12238	50	318.1	46.6	3844
1	104.1	18.2	5976	51	91.4	30.0	6106
2	128.0	-41.8	9817	52	136.9	-54.3	9434
3	343.9	69.4	3953	53	357.6	61.9	3575
4	90.0	6.0	5900	54	48.1	46.6	772
5	281.2	-3.6	7564	55	76.5	45.0	365
6	99.8	9.6	5909	56	120.9	-31.4	9940
7	92.8	-6.0	5364	57	278.4	-23.3	7243
8	98.4	6.0	5563	58	293.9	22.0	3585
9	281.2	-25.9	7235	59	143.7	-51.3	9437
10	113.9	-30.0	9946	60	34.4	64.9	915
11	61.9	40.2	403	61	246.1	-27.3	10738
12	225.0	-64.9	10432	62	274.2	-17.0	7278
13	106.9	-8.4	5420	63	169.3	-58.9	9484
14	281.2	-8.4	7552	64	84.4	31.4	6124
15	73.1	31.4	284	65	168.2	-61.9	9480
16	286.9	13.2	7663	66	325.8	52.8	3860
17	80.2	41.8	346	67	246.1	-32.8	10702
18	78.8	48.1	378	68	321.1	57.4	3870
19	268.6	-30.0	7205	69	302.3	35.7	3657
20	303.8	34.2	3651	70	116.7	-17.0	10158
21	152.1	-58.9	9340	71	88.6	27.3	6094
22	120.9	-10.8	5396	72	272.8	-31.4	7192
23	316.6	48.1	3846	73	85.8	30.0	6118
24	295.3	23.3	3588	74	203.6	-58.9	10427
25	319.2	35.7	3317	75	48.6	52.8	792
26	357.0	67.9	3927	76	315.0	46.6	3843
27	112.5	-20.7	9983	77	274.2	30.0	8153
28	143.4	-31.4	10004	78	112.5	-18.2	5376
29	272.8	-18.2	7275	79	299.5	30.0	3606
30	299.5	35.7	3658	80	289.7	10.8	7655
31	136.6	-48.1	9462	81	307.8	52.8	3875
32	157.5	-57.4	9506	82	98.4	23.3	6008
33	257.3	-38.7	10610	83	111.1	-14.5	5385
34	36.6	66.4	918	84	285.5	32.8	3630
35	261.6	-37.2	10612	85	255.9	-37.2	10616
36	204.8	-60.4	10425	86	272.8	-8.4	7387
37	245.0	-49.7	10543	87	300.9	34.2	3656
38	258.8	-37.2	10611	88	272.8	-20.7	7272
39	310.8	12.0	3117	89	23.8	64.9	911
40	277.0	-14.5	7290	90	102.7	0.0	5530
41	122.3	-22.0	10126	91	317.8	40.2	3496
42	278.4	23.3	8056	92	109.7	-31.4	9956
43	105.5	-19.5	5209	93	165.0	-58.9	9483
44	146.7	-55.9	9428	94	95.6	-10.8	5331
45	91.4	19.5	5994	95	71.7	41.8	361
46	61.7	49.7	439	96	253.1	-34.2	10704
47	319.2	38.7	3490	97	282.7	0.0	7578
48	193.1	-43.4	10904	98	95.6	-23.3	5216
49	97.0	7.2	5905	99	119.5	-24.6	10122

Notes. [a] Index of the level 5 HEALPix pixel of the field. To reproduce each full field, the eight nearest neighbour HEALPix level 5 pixels must also be selected.

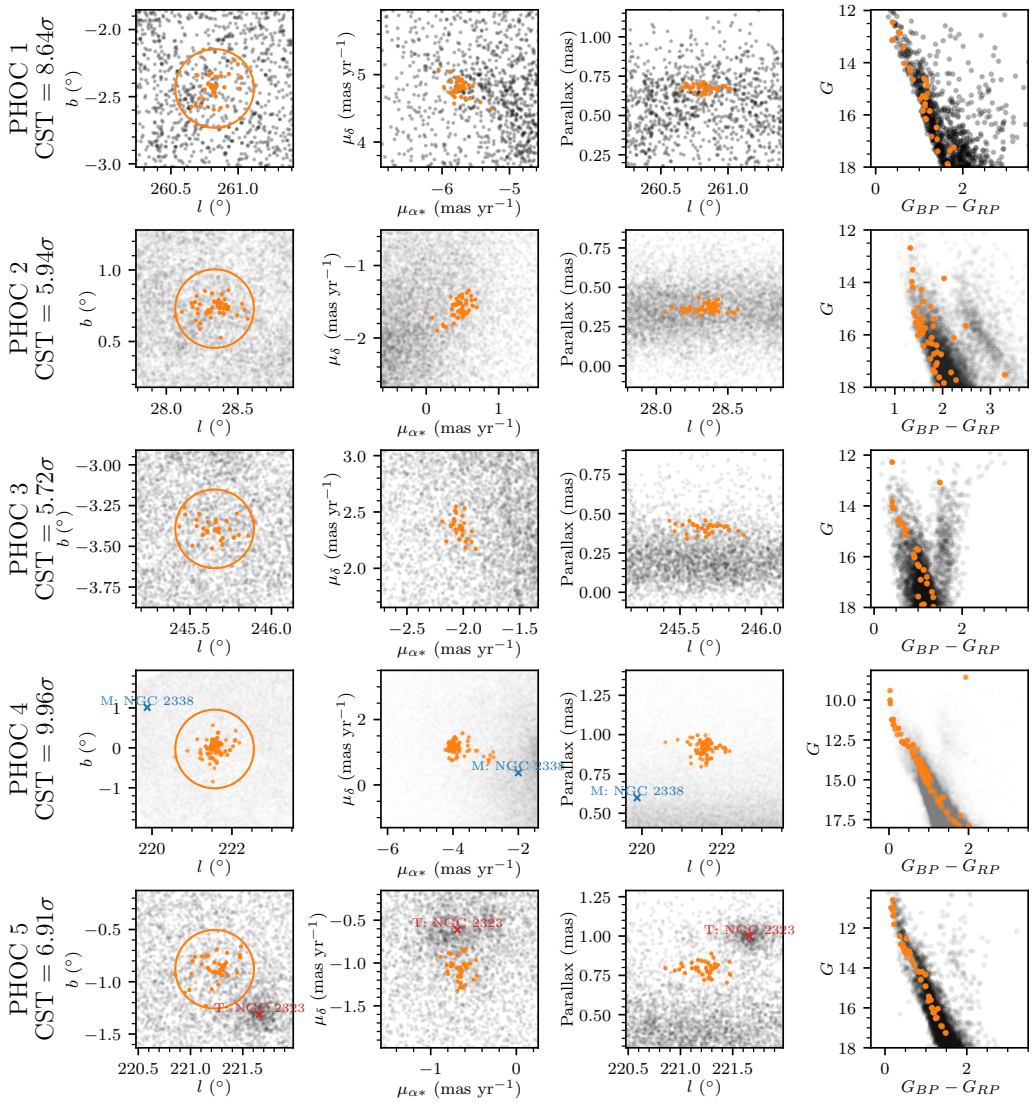

Fig. A.1: Astrometric and photometric plots of the first five new OCs from Sect. **??**. Identified member stars are shown in orange, with background stars in black. Only members with a membership probability of greater than 50% are plotted. The estimated tidal radius for the OCs is depicted with a circle in the l vs. b plots in the first column. CST scores for each object are shown with its name on the left. Nearby OCs from literature catalogues are marked when visible. T (in red text) denotes sources from **cantat-gaudin_clusters_2020**, while M (blue) and S (purple) denote sources from MWSC and **sim_207_2019** respectively that were not detected by **cantat-gaudin_clusters_2020**. A (brown) denotes new OCs detected recently by **castro-ginard_hunting_2020**.

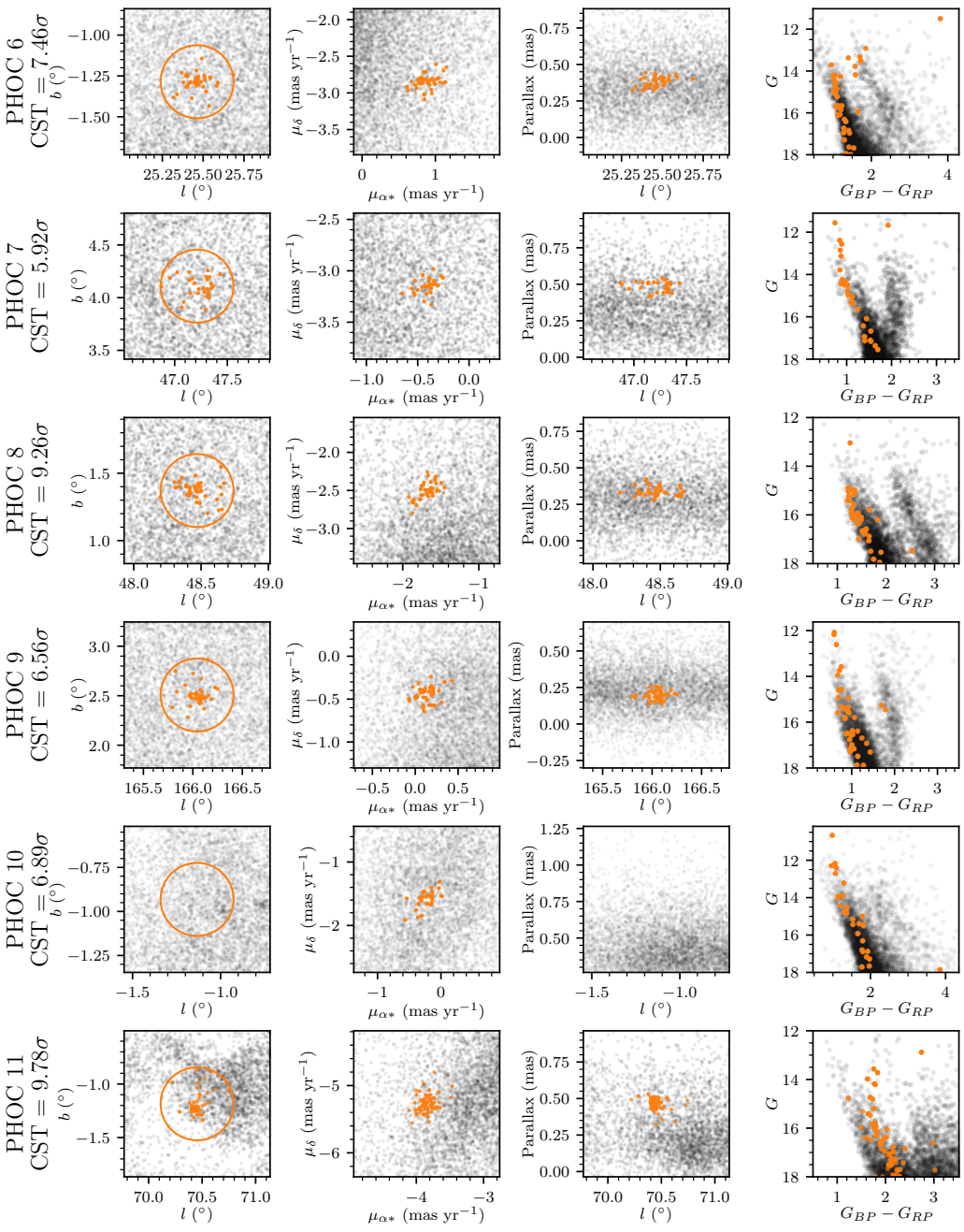

Fig. A.2: Plots of the new OCs PHOC 6 to 11, plotted in the same style as Fig. **??**.

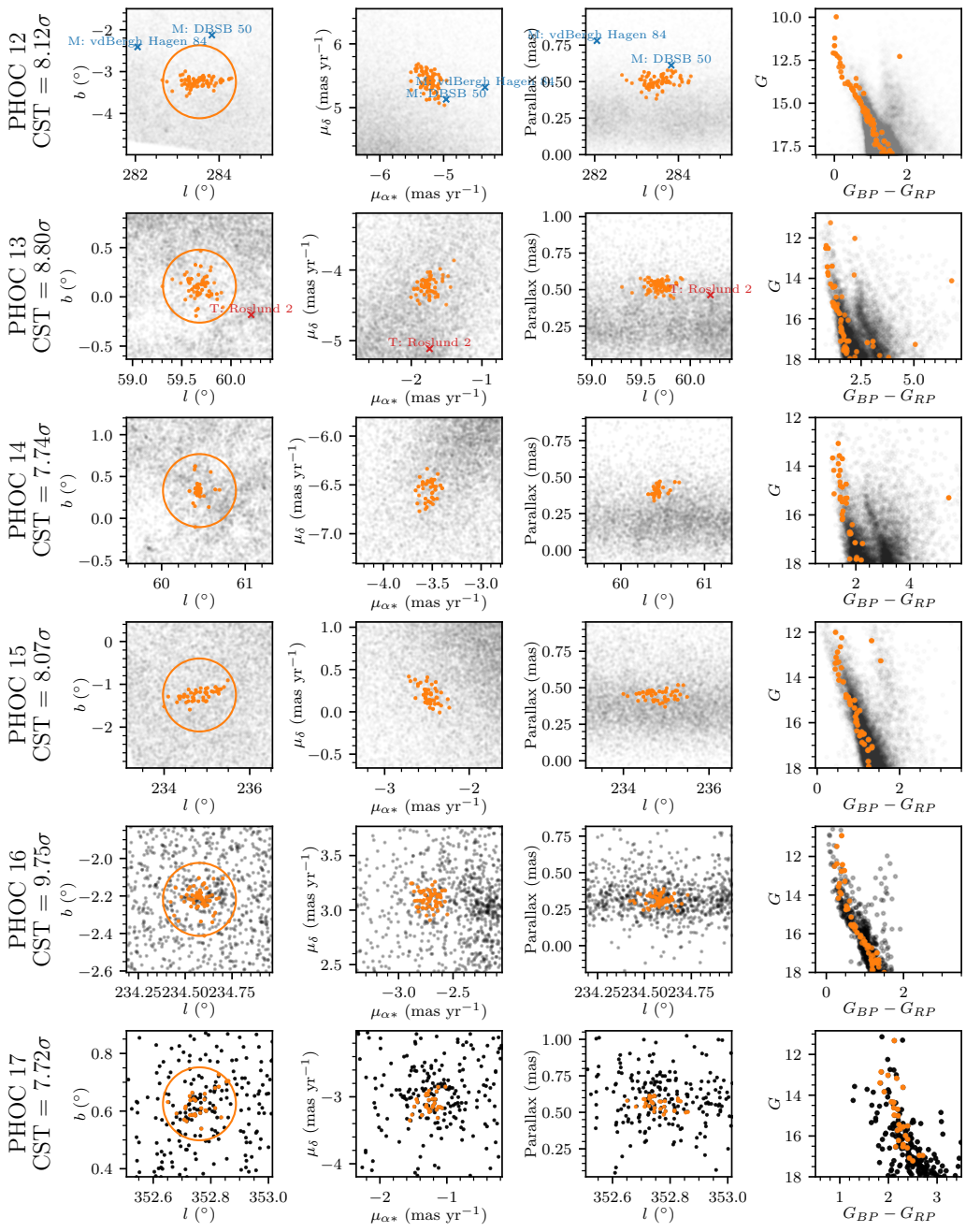

Fig. A.3: Plots of the new OCs PHOC 12 to 17, plotted in the same style as Fig. ??.

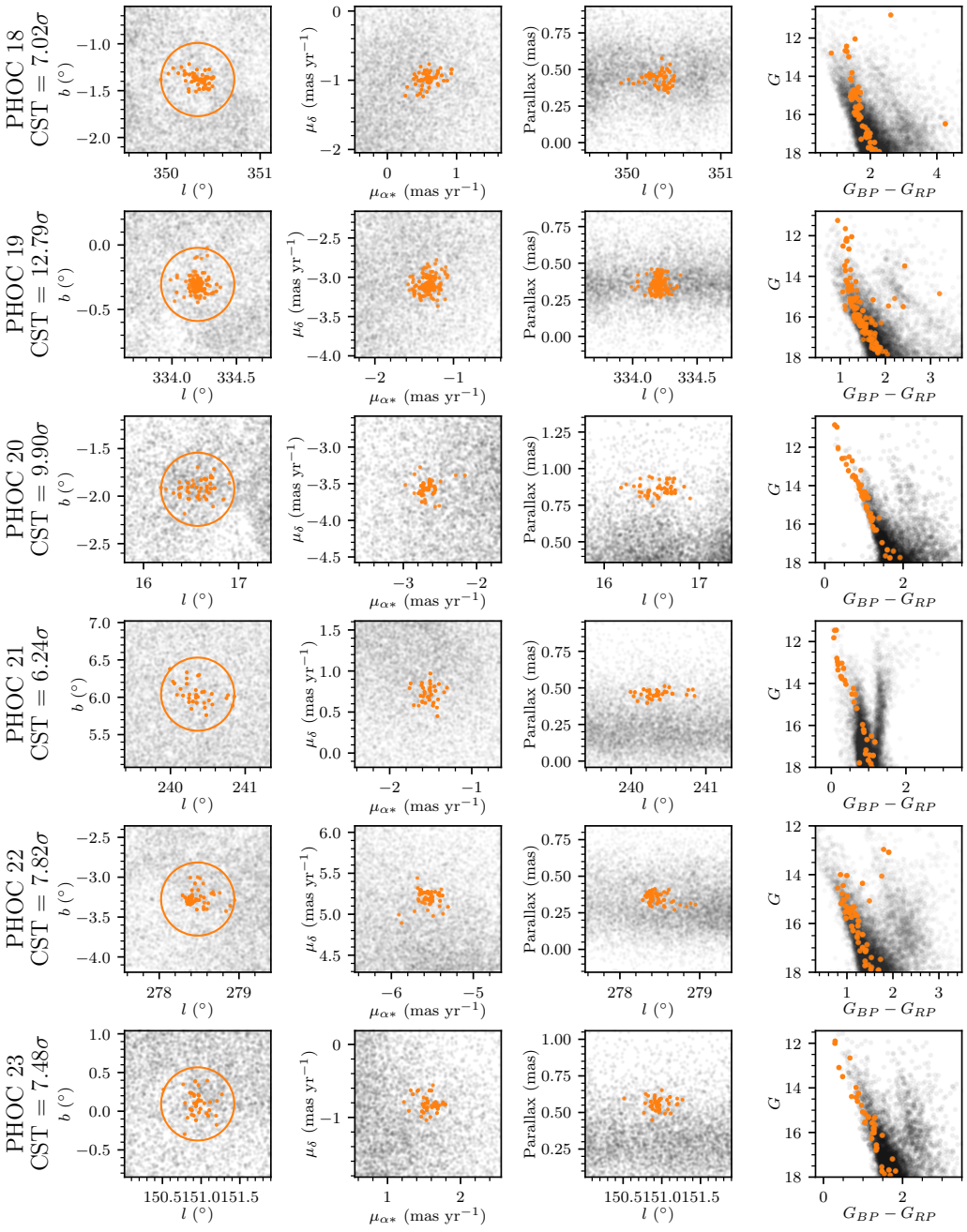

Fig. A.4: Plots of the new OCs PHOC 18 to 23, plotted in the same style as Fig. **??**.

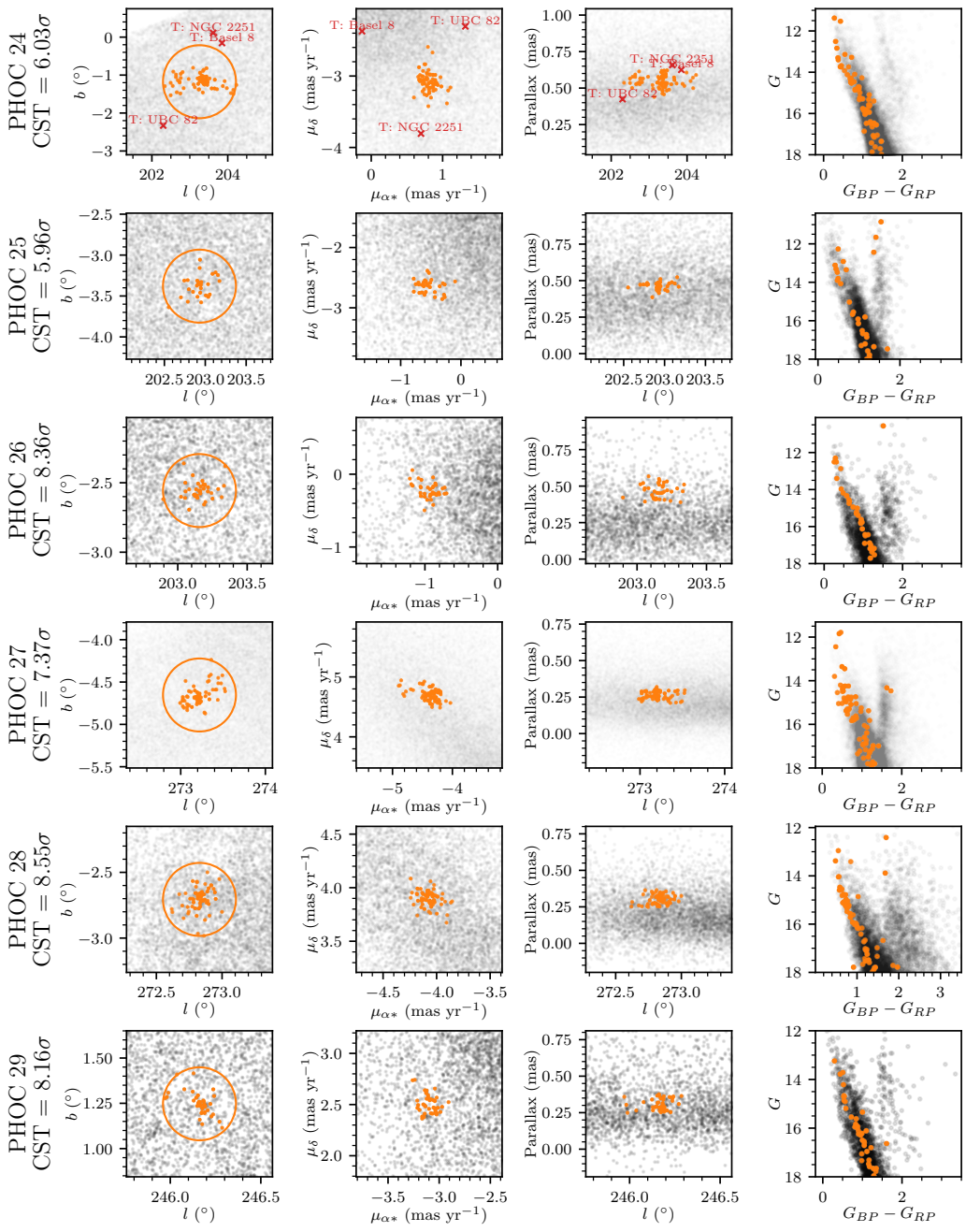

Fig. A.5: Plots of the new OCs PHOC 24 to 29, plotted in the same style as Fig. **??**.

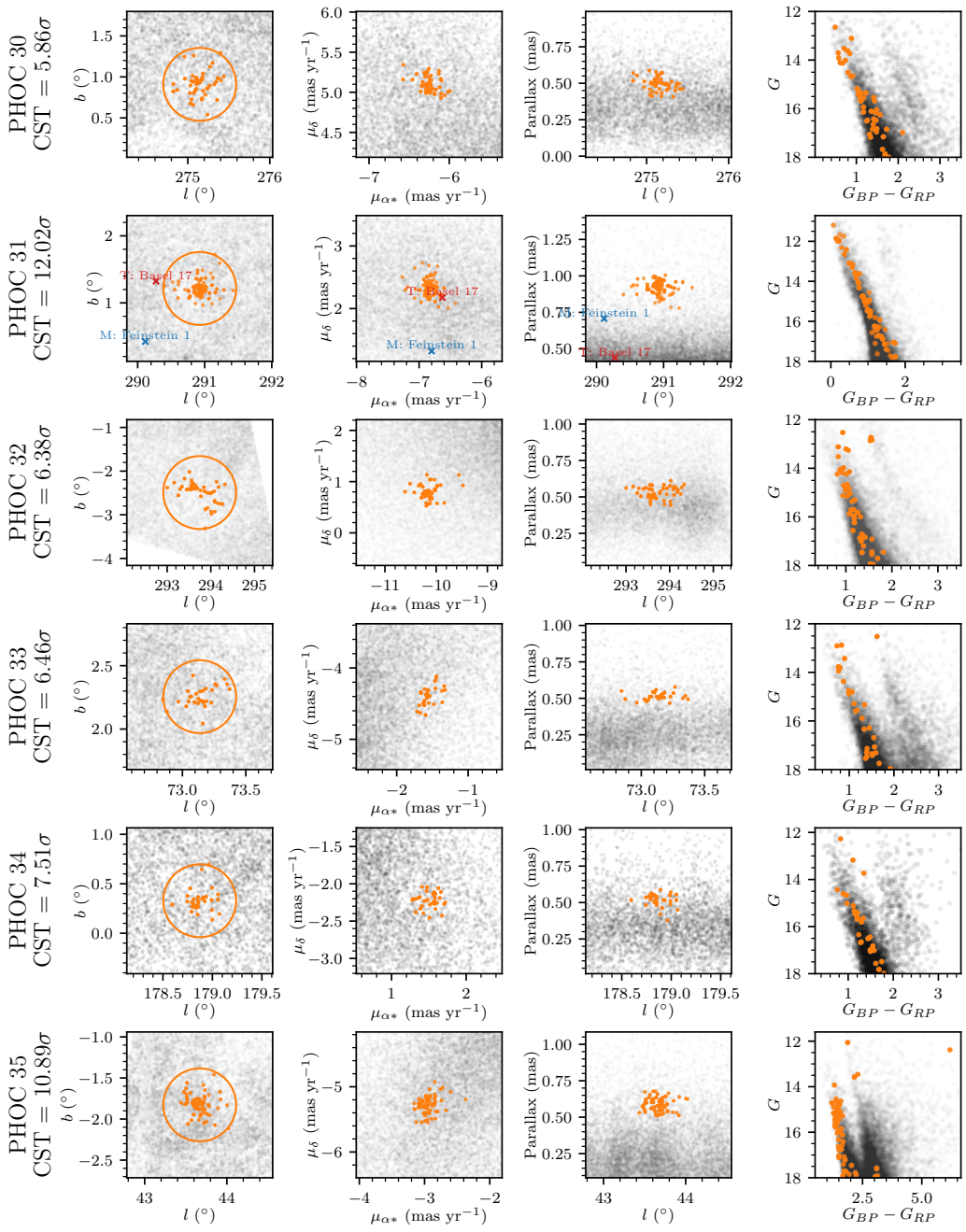

Fig. A.6: Plots of the new OCs PHOC 30 to 35, plotted in the same style as Fig. **??**.

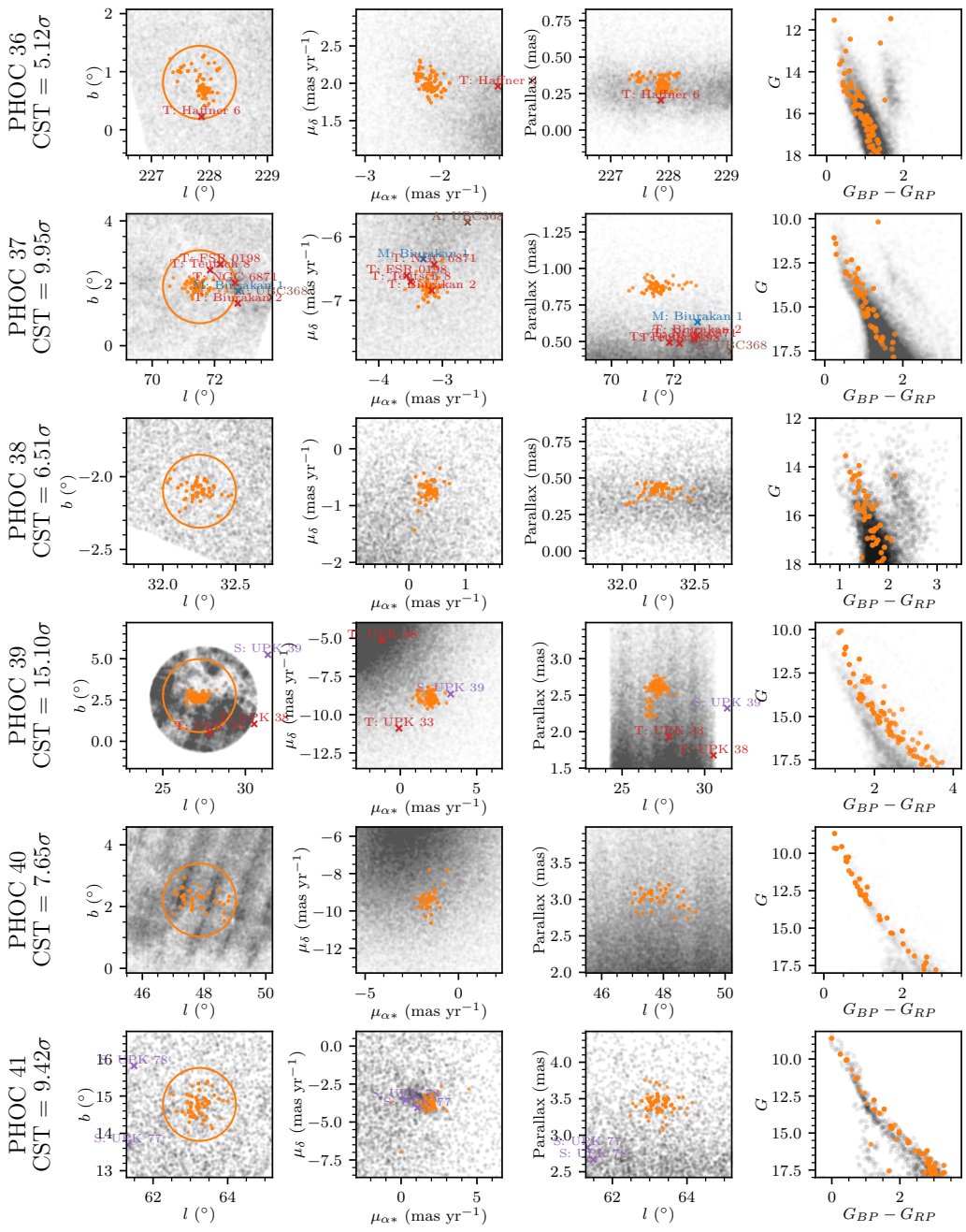

Fig. A.7: Plots of the new OCs PHOC 36 to 41, plotted in the same style as Fig. **??**.

A.2 Appendices for Chapter ??

A.2.1 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 **??**. Tables of members follow the same columns and column naming scheme as in *Gaia* DR3 (**gaia_collaboration_gaia_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.

A.2.2 Table of crossmatch results

Here we provide a table of all crossmatches to all literature clusters that meet our adopted crossmatch criteria from Sect. **??** in Table **??**. 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 **??** 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.

A.2.3 Bayesian neural networks

Given that Bayesian neural networks (BNNs) are only just beginning to see use in the astronomical literature (**huertas-company_hubble_2019**), here we provide a brief background overview of the advantages and caveats of the approximate BNN methodology we adopted in Sect. **??** and Sect. **??**.

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

Tab. A.4: 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[c]
5	n_{stars}	–	Num. of member stars
6	S/N	–	Astrometric S/N
7	$n_{\text{stars}}\|_{r_t}$	–	n_{stars} within r_t
8	$\text{S/N}\|_{r_t}$	–	S/N within r_t
9-10	α, δ	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[a]	μ_{α^*}, μ_δ	mas yr^{-1}	ICRS proper motions
27-29[a]	ϖ	mas	Parallax
30-32[b]	d	pc	Distance
33	n_d	pc	n_{stars} for distance calc.
34	ϖ_0 type	–	Parallax offset type[d]
35-37	X, Y, Z	pc	Galactocentric coords.
38-40[a]	RV	km s^{-1}	Radial velocity[e]
41	n_{RV}	–	n_{stars} with RVs
42-46[b]	CMD class	–	CMD class quantiles[f]
47	Human class	–	(where available)[f]
48-50[b]	$\log t$	$\log$ [yr]	Cluster age
51-53[b]	A_V	mag	V-band extinction
54-56[b]	ΔA_V	mag	Differential A_V
57-59[b]	$m - M$	mag	Photometric dist. mod.
60	m_{clSize}	–	HDBSCAN parameter
61	`merged`	–	Flag if merged[g]
62	`is_gmm`	–	Flag if GMM used[h]
63	$n_{\text{crossmatches}}$	–	Num. crossmatches
64	Xmatch type	–	Type of crossmatch[i]

Notes. The full version is available at the CDS. [(a)] Mean value, standard deviation σ, and standard error $\sigma/\sqrt{n}$ are given. [(b)] Median value and various confidence intervals are given. [(c)] `g` for objects in the **vasiliev_gaia_2021** GC catalogue, otherwise `o` (OC) or `m` (moving group) for clusters according to the empirical cuts in **cantat-gaudin_clusters_2020**. [(d)] Flag indicating six clusters for which parallax bias correction using the method of **lindegren_gaia_2021** was not possible, and a global offset was used instead (see Sect. **??**). [(e)] Corrected using cluster distances to be relative to cluster centre. [(f)] Cluster CMD classes derived using the neural network in Sect. **??**. [(g)] Indicates 25 clusters merged by hand (see Sect. **??**). [(h)] Indicates nine clusters with members from an additional Gaussian mixture model clustering step. [(i)] Method used to assign name to cluster (see Sect. **??**.)

Tab. A.5: All cluster crossmatches, including literature clusters that have no match.

ID	Name	Source	Type	θ (°)	θ_r [a]	$s_{\mu_{\alpha*}}$ (mas yr^{-1})	$\sigma_{\mu_{\alpha*}}$	s_{μ_δ} (mas yr^{-1})	$\sigma_{\mu_{\delta*}}$	s_ϖ (mas)	σ_ϖ
				...							
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	-	-	-	-	-	-
				...							

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 θ may be listed. For works with astrometry, separations s with respect to $\mu_{\alpha*}$, μ_δ, and ϖ are shown, in addition to separations σ 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. **??**, for a valid crossmatch, we require $\theta_r < 1$, and additionally, when crossmatching to a work with full five parameter astrometry, all σ values to be less than two. [(a)] The separation between cluster centres in terms of the largest cluster radius available, $\theta_r = \theta / \max(r_t,\, r_{t,\,\mathrm{lit}})$

a wide range of answers that demonstrate its uncertainty in such difficult cases (**goan_bayesian_2020**, **jospin_hands-bayesian_2022**).

In practice, there is currently no perfect BNN architecture, with all approaches having some flaws (**goan_bayesian_2020**, **jospin_hands-bayesian_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_bayesian_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_dropout_2015-1**), having also been used within astronomy (e.g. **huertas-company_hubble_2019**, **leung_deep_2019**). However, this approximation is not inherently Bayesian (**hron_variational_2017**), and may be improved 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_weight_2015**). This approximation can hold relatively well for (simple) neural networks, which often have normally distributed weights, but may cause underfitting on more complicated problems (**goan_bayesian_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_flipout_2018**). Similar approaches using Bayes by backpropagation and Flipout have seen some use in the astronomy literature (**lin_detection_2021**). We use the implementations of `DenseFlipout` and `Convolution2DFlipout` layers in TensorFlow Probability (**dillon_tensorflow_2017**), minimising the evidence lower bound (ELBO) loss (**blundell_weight_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_bayesian_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.

Colophon

This thesis was typeset with LaTeX 2_ε. It uses the *Clean Thesis* style developed by Ricardo Langner. The design of the *Clean Thesis* style is inspired by user guide documents from Apple Inc.

Download the *Clean Thesis* style at `http://cleanthesis.der-ric.de/`.